

Construction of a mitochondrial consensus sequence for the Khoi-San population of Southern Africa

BY

MICHELLE KOEKEMOER, B.Sc. (Agric.), M.Sc. (Agric.)

Thesis submitted for the degree Philosophiae Doctor (Ph.D.)
in Biochemistry at the North-West University

PROMOTOR: Professor Antonel Olckers
Centre for Genome Research, North-West University (Potchefstroom Campus)

CO-PROMOTOR: Doctor Gordon Wayne Towers
Centre for Genome Research, North-West University (Potchefstroom Campus)

August 2010

Samestelling van 'n mitokondriale konsensusvolgorde vir die Khoi-San bevolking van Suidelike Afrika

DEUR

MICHELLE KOEKEMOER, B.Sc. (Agric.), M.Sc. (Agric.)

Proefskrif voorgelê vir die graad Philosophiae Doctor (Ph.D.)
in Biochemie aan die Noordwes-Universiteit

PROMOTOR: Professor Antonel Olckers
Sentrum vir Genomiese Navorsing, Noordwes-Universiteit (Potchefstroom Kampus)

MEDEPROMOTOR: Doktor Gordon Wayne Towers
Sentrum vir Genomiese Navorsing, Noordwes-Universiteit (Potchefstroom Kampus)

Augustus 2010

This thesis is dedicated to my late grandmother,
Hester Margaret Freeman,
and to my husband,
Theunis Johannes Koekemoer

*To know where we are going, we have to know where we are;
to know that we have to know where we came from*

Filipino version of an Oceanic proverb

Stephen Oppenheimer, 2003

ABSTRACT

The revised Cambridge Reference Sequence (rCRS) is used as a standard for the human mitochondrial DNA (mtDNA) sequence in studies of human evolution and the identification of disease-causing mutations. Due to the large number of differences observed between the rCRS and mitochondrial sequences obtained from individuals of African descent, it is frequently difficult to differentiate between alterations that are population-specific or have possible pathological significance. To address this problem, two human consensus sequences consisting of mitochondrial sequences from different continents and different haplogroups have been constructed. However, combining data from different continents and haplogroups led to a loss of variation in the human mitochondrial consensus sequences. This can be countered by using an African reference sequence, of which two sequences are currently available, namely NC_001807 (L3a1) and D38112 (L0c2). However, these sequences are not representative of the most ancient African populations (i.e. hunter-gatherers) or haplogroups (i.e. L0).

In the current investigation, the complete mitochondrial genome sequences of 30 Khoi-San individuals from southern Africa were determined. Twenty-two of these Khoi-San sequences, which belong to the L0a and L0b sub-haplogroups, were combined with 13 L0 Khoi-San sequences generated previously to compile a consensus sequence for the Southern African Khoi-San population. This Khoi-San consensus sequence will represent the first example of an African population-specific consensus sequence.

The results presented in the current investigation provide support for previous findings regarding the existence of a high level of genetic variation in the mtDNA of the Khoi-San population, as well as the ancient character of the Khoi-San population. In addition, it offers novel insights into the complexity of the L0 haplogroup within the Khoi-San population and the African population as a whole. The high level of genetic variation and increased age of mtDNA lineages in the Southern African Khoi-San population compared to other populations, support the use of the Khoi-San consensus sequence, alone or in conjunction with the rCRS, as a standard in studies of human evolution and mitochondrial disease.

OPSOMMING

Die hersiene Cambridge Verwysingsvolgorde (rCRS) word as 'n standaard vir die menslike mitokondriale DNS volgorde aangewend in studies van menslike evolusie en in die identifikasie van siekteveroorsakende mutasies. As gevolg van die groot aantal verskille wat waargeneem word tussen die rCRS en mitokondriale volgordes van individue van Afrika-oorsprong, is dit moeilik om te onderskei tussen veranderinge wat bevolkingsspesifiek is, of moontlike patologiese betekenis het. Om die rede is twee menslike konsensusvolgordes bestaande uit mitokondriale volgordes van verskillende kontinente en verskeie haplogroepe saamgestel. Die samevoeging van data van verskillende kontinente en haplogroepe het egter gelei tot 'n verlies in variasie in die menslike konsensusvolgordes. Dit kan teëgewerk word deur die gebruik van 'n Afrika verwysingsvolgorde en daar is tans twee volgordes beskikbaar, naamlik NC_001807 (L3a1) en D38112 (L0c2). Hierdie twee volgordes is egter nie verteenwoordigend van die oudste Afrika bevolkings (d.i. jagter-versamelaars) of haplogroepe (d.i. L0) nie.

In die huidige ondersoek is die volledige mitokondriale genoomvolgordes van 30 Khoi-San individue bepaal. Twee-en-twintig van hierdie Khoi-San volgordes, wat aan die L0a en L0b sub-haplogroepe behoort het, is met 13 L0 volgordes wat voorheen bepaal is, gekombineer om 'n konsensusvolgorde vir die Suidelike-Afrika Khoi-San bevolking saam te stel. Hierdie Khoi-San konsensusvolgorde verteenwoordig die eerste voorbeeld van 'n Afrika bevolkingsspesifieke konsensusvolgorde.

Die resultate wat aangebied word in die huidige ondersoek, ondersteun vorige bevindinge rakende die bestaan van 'n hoë vlak van genetiese variasie in die mitokondriale DNS van die Khoi-San bevolking, asook die antieke aard van die Khoi-San bevolking. Dit bied verder nuwe insigte in die kompleksiteit van die Khoi-San bevolking as 'n geheel. Die hoë vlak van genetiese variasie en die verhoogde ouderdom van die mtDNS lyne in die Suidelike-Afrika Khoi-San bevolking in vergelyking met ander bevolkings, ondersteun die gebruik van die Khoi-San konsensusvolgorde, alleen of saam met die rCRS, as 'n standaard in studies van menslike evolusie en mitokondriale afwykings.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS AND SYMBOLS.....	i
LIST OF EQUATIONS.....	xi
LIST OF FIGURES.....	xiii
LIST OF TABLES.....	xvii
ACKNOWLEDGEMENTS.....	xxv
 CHAPTER ONE	
INTRODUCTION.....	1
 CHAPTER TWO	
BIOCHEMICAL AND GENETIC ASPECTS OF THE MITOCHONDRION.....	7
2.1 ORIGIN OF THE MITOCHONDRION.....	7
2.2 STRUCTURE AND FUNCTION OF THE MITOCHONDRION.....	8
2.3 THE MITOCHONDRIAL ELECTRON TRANSPORT CHAIN.....	10
2.4 THE MITOCHONDRIAL GENOME.....	12
2.4.1 Inheritance of the mitochondrial genome.....	12
2.4.2 Genetic organisation of the mitochondrial genome.....	13
2.4.3 Replication of the mitochondrial genome.....	18
2.4.4 Transcription of the mitochondrial genome.....	21
2.4.5 Translation of the mitochondrial genome.....	24
2.5 GENETIC VARIATION AND THE MITOCHONDRIAL DNA GENOME.....	26
2.5.1 Mutation rate of the mitochondrial genome.....	27
2.5.2 Heteroplasmy and the mitochondrial genome.....	31
2.5.3 Recombination and the mitochondrial genome.....	33
2.5.4 Other evolutionary forces and population events that influence genetic variation in the mitochondrial genome.....	35
2.5.5 Selection and the mitochondrial genome.....	38
2.5.5.1 Statistical tests of selection.....	39
2.5.5.2 Adaptive selection and mitochondrial genome variation.....	41
2.6 MITOCHONDRIAL DNA VARIATION AND HUMAN DISEASE.....	44
 CHAPTER THREE	
MITOCHONDRIAL DNA VARIATION AND HUMAN ORIGINS.....	51
3.1 HUMAN ORIGINS AND MIGRATIONS.....	51
3.1.1 Models of modern human origins.....	51
3.1.2 Support for the Recent African Origin model.....	54

TABLE OF CONTENTS

3.1.3	The Bantu expansions in Africa.....	59
3.1.4	The African Diaspora.....	63
3.2	MITOCHONDRIAL PHYLOGENIES.....	68
3.2.1	Methods used to assess variation.....	68
3.2.1.1	Restriction fragment length polymorphism analysis.....	69
3.2.1.2	Sequencing analysis.....	70
3.2.1.3	Errors in mitochondrial sequence data.....	71
3.2.2	Construction of phylogenetic trees.....	72
3.2.2.1	Evolutionary models of nucleotide substitution.....	73
3.2.2.2	Phylogenetic tree construction methods.....	75
3.2.2.2.1	Distance methods.....	76
3.2.2.2.2	Discrete character method.....	76
3.2.2.2.3	Choice of phylogenetic tree construction method.....	77
3.2.2.3	Rooting of phylogenetic trees.....	79
3.2.2.4	Confidence limits and phylogenies.....	80
3.2.2.5	Consensus phylogenetic trees.....	81
3.3	MITOCHONDRIAL HAPLOGROUPS.....	82
3.3.1	African mitochondrial haplogroups.....	83
3.3.1.1	Classification of African mitochondrial haplogroups.....	85
3.3.1.2	Geographic distribution of African mitochondrial haplogroups.....	87
3.3.1.3	Population-specific African mitochondrial haplogroups.....	89
3.3.1.4	TMRCA of African mitochondrial haplogroups.....	90
3.3.2	Asian mitochondrial haplogroups.....	95
3.3.3	Native American mitochondrial haplogroups.....	97
3.3.4	Oceanic mitochondrial haplogroups.....	101
3.3.5	European mitochondrial haplogroups.....	106
3.4	THE ROLE OF MITOCHONDRIAL HAPLOGROUPS IN CLINICAL CONDITIONS AND PHENOTYPIC VARIATION.....	108
3.5	THE REVISED CAMBRIDGE REFERENCE SEQUENCE.....	111
3.6	A CONSENSUS SEQUENCE FOR HUMAN MITOCHONDRIAL DNA VARIATION.....	113
3.7	THE SOUTHERN AFRICAN KHOI-SAN POPULATION.....	115
3.8	OBJECTIVES OF THE RESEARCH PROGRAMME.....	120
3.8.1	Specific aims of the proposed project.....	121
CHAPTER FOUR		
MATERIALS AND METHODS.....		123
4.1	ETHICAL APPROVAL.....	123
4.2	SUBJECTS.....	123
4.3	ISOLATION OF GENOMIC DNA.....	124
4.4	POLYMERASE CHAIN REACTION.....	125
4.4.1	Polymerase chain reaction primer pairs.....	125

4.4.2	Polymerase chain reaction conditions.....	126
4.4.3	Polymerase chain reaction product purification.....	127
4.5	AGAROSE GEL ELECTROPHORESIS.....	129
4.6	DETERMINATION OF DNA CONCENTRATION.....	129
4.7	AUTOMATED DNA SEQUENCING.....	130
4.7.1	Cycle sequencing.....	130
4.7.2	Sodium dodecyl sulphate/heat treatment of extension products.....	133
4.7.3	Precipitation of extension products.....	134
4.7.4	Electrophoresis of extension products.....	135
4.7.5	Analysis of sequence data.....	136
4.8	MITOCHONDRIAL DNA GENOME SEQUENCES USED IN ANALYSES.....	137
4.9	DETERMINATION OF MITOCHONDRIAL HAPLOGROUPS.....	138
4.10	COMPILATION OF MITOCHONDRIAL SEQUENCE DATASETS USED IN ANALYSES.....	141
4.10.1	Compilation of datasets used for phylogenetic analyses.....	141
4.10.2	Compilation of datasets used for statistical analyses.....	143
4.10.3	Compilation of the dataset used for construction of the global L0-specific haplogroup network.....	144
4.10.4	Sequence datasets used in the current investigation.....	145
4.11	SIGNIFICANT FIGURES.....	146
4.12	PHYLOGENETIC ANALYSIS.....	147
4.12.1	Sequence alignment.....	147
4.12.2	Selection of random number seed.....	148
4.12.3	Estimation of transition/transversion ratio.....	148
4.12.4	Estimation of gamma shape parameter.....	149
4.12.5	Neighbour-joining method.....	151
4.12.6	Maximum parsimony method.....	152
4.13	CONSTRUCTION OF L0-SPECIFIC HAPLOGROUP NETWORKS.....	153
4.13.1	Southern African Khoi-San L0-specific haplogroup network.....	153
4.13.2	Global L0-specific haplogroup network.....	154
4.14	STATISTICAL ANALYSIS.....	154
4.14.1	Basic sequence statistics.....	155
4.14.2	Codon usage.....	155
4.14.2.1	Relative synonymous codon usage.....	155
4.14.2.2	The effective number of codons.....	156
4.14.2.3	Codon bias index.....	156
4.14.2.4	Scaled chi square.....	156
4.14.3	Genetic diversity measures.....	157
4.14.3.1	The number of segregating sites.....	157
4.14.3.2	The average number of nucleotide differences.....	157
4.14.3.3	Nucleotide diversity.....	157
4.14.4	Deviation from the neutral theory of molecular evolution.....	158

TABLE OF CONTENTS

4.14.4.1	Tajima's <i>D</i> test.....	159
4.14.4.2	Fu and Li's <i>D*</i> and <i>F*</i> tests.....	160
4.14.5	Population size changes.....	161
4.14.5.1	Frequency distribution of pairwise sequence differences.....	161
4.14.5.2	Statistical parameters to detect population expansion.....	162
4.14.6	Determination of the effect of selection on mitochondrial DNA variation...	164
4.14.6.1	Analysis of synonymous and non-synonymous substitutions.....	165
4.14.6.2	Conservation index.....	166
4.14.7	Estimation of coalescent dates.....	168
4.15	CONSTRUCTION OF A REVISED CLASSIFICATION SCHEME FOR THE L0 HAPLOGROUP.....	169
4.16	CONSTRUCTION OF A MITOCHONDRIAL CONSENSUS SEQUENCE.....	171
4.17	CALCULATION OF PERCENTAGE SIMILARITY AND PAIRWISE NUMBER OF DIFFERENCES.....	172
 CHAPTER FIVE		
RESULTS AND DISCUSSION.....		173
5.1	STUDY DESIGN.....	174
5.1.1	Selection of Khoi-San individuals.....	174
5.2	ISOLATION OF GENOMIC DNA.....	175
5.3	POLYMERASE CHAIN REACTION.....	176
5.3.1	Polymerase chain reaction primers.....	176
5.3.2	Polymerase chain reaction optimisation.....	177
5.3.3	Artefacts observed in PCR amplified samples.....	179
5.3.3.1	Amplification efficiency.....	180
5.3.3.2	Background smear.....	182
5.3.3.3	Secondary amplification.....	183
5.3.3.4	Primer-dimers.....	184
5.4	AGAROSE GEL ELECTROPHORESIS.....	185
5.4.1	Artefacts observed on agarose gels.....	186
5.4.1.1	Artefacts in the gel matrix.....	186
5.4.1.2	Distortion of sample fragments.....	187
5.4.1.3	Slanted fragments.....	188
5.5	PCR PRODUCT PURIFICATION.....	189
5.6	AUTOMATED DNA SEQUENCING.....	190
5.6.1	Primer design.....	190
5.6.2	DNA cycle sequencing optimisation.....	191
5.6.3	Precipitation of extension products.....	192
5.6.4	Artefacts observed on electropherograms.....	193
5.6.4.1	Background noise.....	193
5.6.4.2	Low signal strength.....	195

5.6.4.3	Poor mobility correction.....	196
5.6.4.4	Insertion and deletion of nucleotides.....	197
5.6.4.5	Excess dye peaks and dye blobs.....	198
5.6.4.6	Slippage in homopolymer regions.....	200
5.6.5	Template quality.....	200
5.6.6	Errors in mitochondrial sequence data.....	201
5.7	SEQUENCING RESULTS.....	202
5.7.1	PCR region 1 of the mitochondrial genome.....	205
5.7.1.1	Displacement loop.....	208
5.7.1.2	tRNA phenylalanine.....	211
5.7.1.3	12S ribosomal RNA.....	213
5.7.2	PCR region 2 of the mitochondrial genome.....	214
5.7.2.1	16S ribosomal RNA.....	217
5.7.2.2	NADH dehydrogenase subunit 1.....	218
5.7.3	PCR region 3 of the mitochondrial genome.....	220
5.7.3.1	tRNA genes.....	222
5.7.3.2	NADH dehydrogenase subunit 2.....	224
5.7.4	PCR region 4 of the mitochondrial genome.....	226
5.7.4.1	tRNA tyrosine.....	229
5.7.4.2	Cytochrome c oxidase subunit I.....	232
5.7.4.3	tRNA serine (UCN).....	234
5.7.5	PCR region 5 of the mitochondrial genome.....	235
5.7.5.1	Cytochrome c oxidase subunit II.....	237
5.7.5.2	ATP synthase F ₀ subunit 8.....	238
5.7.5.3	ATP synthase F ₀ subunit 6.....	239
5.7.5.4	Cytochrome c oxidase subunit III.....	241
5.7.5.5	NADH dehydrogenase subunit 3.....	242
5.7.6	PCR region 6 of the mitochondrial genome.....	244
5.7.6.1	NADH dehydrogenase subunit 4L.....	246
5.7.6.2	NADH dehydrogenase subunit 4.....	247
5.7.7	PCR region 7 of the mitochondrial genome.....	248
5.7.7.1	tRNA genes.....	251
5.7.7.2	NADH dehydrogenase subunit 5.....	251
5.7.7.3	NADH dehydrogenase subunit 6.....	256
5.7.8	PCR region 8 of the mitochondrial genome.....	258
5.7.8.1	tRNA genes.....	260
5.7.8.2	Cytochrome <i>b</i>	261
5.8	MITOCHONDRIAL HAPLOGROUP ANALYSIS.....	263
5.8.1	Khoi-San mitochondrial genome sequences.....	264
5.8.2	Other L-specific mitochondrial genome sequences.....	269

5.9	PHYLOGENETIC ANALYSIS.....	272
5.9.1	Construction of phylogenetic trees.....	273
5.9.2	Neighbour-joining trees.....	277
5.9.2.1	Global neighbour-joining tree.....	278
5.9.2.2	Neighbour-joining tree of African individuals.....	295
5.9.2.3	Neighbour-joining tree of Khoi-San individuals.....	304
5.9.3	Maximum parsimony trees.....	310
5.9.3.1	Global maximum parsimony tree.....	310
5.9.3.2	Maximum parsimony tree of African individuals.....	316
5.9.3.3	Maximum parsimony tree of Khoi-San individuals.....	320
5.9.4	Geographic distribution of African mitochondrial haplogroups.....	325
5.10	L0-SPECIFIC HAPLOGROUP NETWORK.....	331
5.10.1	Southern African Khoi-San L0-specific haplogroup network.....	332
5.10.2	Global L0-specific haplogroup network.....	334
5.10.2.1	Observed branching in global L0-specific haplogroup network.....	338
5.10.2.2	Geographic distribution of L0 sub-haplogroups and lineages.....	349
5.10.2.3	Recurrent alterations and reversions observed in the global L0-specific haplogroup network.....	353
5.11	STATISTICAL ANALYSIS.....	360
5.11.1	Basic sequence statistics.....	360
5.11.1.1	Nucleotide composition and G + C content.....	361
5.11.1.2	Codon usage.....	363
5.11.2	Genetic diversity measures.....	370
5.11.3	Deviation from the neutral theory of molecular evolution.....	373
5.11.4	Population size changes.....	376
5.11.4.1	Frequency distributions of pairwise differences between mitochondrial DNA sequences.....	377
5.11.4.2	Statistical parameters to detect population expansion.....	388
5.11.5	The effect of selection on human mitochondrial DNA variation.....	394
5.11.5.1	Analysis of synonymous and non-synonymous substitutions.....	394
5.11.5.2	Conservation index.....	396
5.11.6	Coalescent date estimates.....	403
5.12	REVISED CLASSIFICATION SCHEME FOR THE L0 HAPLOGROUP.....	417
5.13	KHOI-SAN CONSENSUS SEQUENCE.....	427
5.14	SUMMARY OF RESULTS GENERATED IN THE CURRENT INVESTIGATION.....	445
5.14.1	Sequencing results.....	445
5.14.2	Mitochondrial haplogroup analysis.....	448
5.14.3	Phylogenetic analysis.....	452
5.14.4	Global L0-specific haplogroup network.....	456
5.14.5	Geographic distribution of African mitochondrial haplogroups.....	461
5.14.6	Statistical analysis.....	463

5.14.6.1	Basic sequence statistics.....	463
5.14.6.2	Genetic diversity measures.....	464
5.14.6.3	Deviation from the neutral theory of molecular evolution and factors to which this could be attributed.....	465
5.14.6.4	Coalescent date estimates.....	468
5.14.7	Revised classification scheme for the L0 haplogroup.....	471
5.14.8	Khoi-San consensus sequence.....	475
 CHAPTER SIX		
CONCLUSIONS.....		481
6.1	PROPOSED MODEL FOR GENETIC VARIATION OBSERVED IN HUMAN MITOCHONDRIAL GENOMES.....	481
6.2	EVIDENCE PROVIDING SUPPORT FOR THE PROPOSED MODEL FOR GENETIC VARIATION OBSERVED IN HUMAN MITOCHONDRIAL GENOMES.....	483
6.2.1	Factors that introduce genetic variation in mitochondrial genomes.....	484
6.2.2	Factors that shape and govern genetic variation in mitochondrial genomes.....	485
6.2.3	Measurable outcomes that are affected by the level of genetic variation present in mitochondrial genomes.....	489
6.2.4	Standards used to investigate genetic variation present in mitochondrial genomes.....	493
6.2.5	Update of proposed model for genetic variation observed in human mitochondrial genomes.....	494
6.3	IMPLICATIONS OF THE PROPOSED MODEL FOR GENETIC VARIATION OBSERVED IN HUMAN MITOCHONDRIAL GENOMES.....	496
6.3.1	The place of origin in Africa.....	496
6.3.2	The estimated TMRCA of the most ancient African mitochondrial haplogroup.....	498
6.3.3	The migration of the L3 haplogroup from Africa to the rest of the world.....	499
6.3.4	The role of the mitochondrial genetic background.....	501
6.3.5	Practical considerations of the use of the Khoi-San consensus sequence.....	501
6.4	FUTURE DIRECTIONS IN THE STUDY OF GENETIC VARIATION OBSERVED IN HUMAN MITOCHONDRIAL GENOMES.....	504
6.4.1	Sequencing of additional Khoi-San individuals.....	504
6.4.2	The global human mtDNA phylogeny.....	506
6.4.3	Standard nomenclature for mitochondrial haplogroups.....	507
6.4.4	Statistical analyses.....	509
6.4.5	The role of natural selection in adaptation.....	510
6.4.6	The history and evolution of the Khoi-San population.....	510
6.4.7	Classification of human social groups.....	511

CHAPTER SEVEN

REFERENCES.....	511
7.1 GENERAL REFERENCES.....	511
7.2 ELECTRONIC REFERENCES.....	536

APPENDIX A

SEQUENCE ALTERATIONS OBSERVED IN COMPLETE KHOI-SAN SEQUENCES COMPARED TO THE RCRS.....	539

APPENDIX B

SYNOPSIS OF MOLECULAR INFORMATION FOR SEQUENCE ALTERATIONS OBSERVED BETWEEN KHOI-SAN SEQUENCES AND THE RCRS.....	555

APPENDIX C

A SCHEMATIC REPRESENTATION OF A GLOBAL PHYLOGENETIC TREE OF MITOCHONDRIAL DNA HAPLOGROUPS.....	565

APPENDIX D

EXCLUSION CRITERIA FOR L-SPECIFIC MITOCHONDRIAL DNA SEQUENCES	567

APPENDIX E

HAPLOGROUP ANALYSIS OF COMPLETE KHOI-SAN MITOCHONDRIAL DNA GENOME SEQUENCES.....	569

APPENDIX F

COMPARISON OF HAPLOGROUP CLASSIFICATION SCHEMES.....	571

APPENDIX G

MITOCHONDRIAL GENOME SEQUENCES USED IN ANALYSES.....	573

APPENDIX H

LIST OF MITOCHONDRIAL DNA SEQUENCES EXCLUDED DURING THE COMPILATION OF DIFFERENT DATASETS.....	587

APPENDIX I

COUNTRIES LOCATED IN THE DIFFERENT GEOGRAPHIC REGIONS OF AFRICA	597

APPENDIX J

GLOBAL AND AFRICA PHYLOGENETIC TREES.....	599

APPENDIX K	
L0 HAPLOGROUP NETWORK OBTAINED FROM MITOMAP.....	609
APPENDIX L	
GLOBAL L0-SPECIFIC HAPLOGROUP NETWORK.....	611
APPENDIX M	
SYNONYMOUS CODON USAGE OF GENES IN 43 KHOI-SAN MTDNA GENOME SEQUENCES.....	615
APPENDIX N	
MITOCHONDRIAL CONSENSUS SEQUENCE FOR THE KHOI-SAN POPULATION OF SOUTHERN AFRICA.....	621
APPENDIX O	
PERCENTAGE SIMILARITY OBSERVED BETWEEN KHOI-SAN MITOCHONDRIAL SEQUENCES AND THE KHOI-SAN CONSENSUS SEQUENCE.....	629

LIST OF ABBREVIATIONS AND SYMBOLS

Symbols and abbreviations are listed in alphabetical order:

LIST OF SYMBOLS

α	alpha: denoting gamma shape parameter
Δ	delta: denoting change
$^{\circ}\text{C}$	degrees centigrade
η	eta: denoting total number of mutations
γ	gamma
μ	micro: 10^{-6}
m	milli: 10^{-3}
μ	mu: denoting mutation rate
n	nano: 10^{-9}
p	pico: 10^{-12}
%	percent
π	pi: denoting nucleotide diversity or the average number of pairwise differences between nucleotide sequences
Π_n	average number of nucleotide differences between two sequences
Ψ	psi: denoting pseudouridylate
ρ	rho: denoting average number of sites differing between a set of sequences and a specified common ancestor
Σ	sigma: denoting summation
σ^2	sigma squared: variance
$\sqrt{\quad}$	square root
θ	theta: denoting the expected pairwise nucleotide site differences or nucleotide diversity parameter
θ_{π}	the mean number of nucleotide differences between two sequences
®	registered trademark
™	trademark
=	equal to
≥	greater than or equal to
≡	identical to
≤	less than or equal to
+	plus
±	plus-minus
\$	dollar: denoting incompletely classified sequences
-	minus or gap (in a nucleotide sequence alignment)
?	missing character (in a nucleotide sequence alignment)
■	green square: outgroup (in phylogenetic trees)
★	red star: Khoi-San sequences generated in the current investigation (in phylogenetic trees)
●	turquoise octagon: rCRS (in phylogenetic trees)
▲	violet triangle: African reference sequence (in phylogenetic trees)

LIST OF ABBREVIATIONS

3'	3 prime
5'	5 prime
5S	5S ribosomal RNA
7S DNA	small, noncovalently bound DNA found in the D-loop of mtDNA
12	12S rRNA (in L0-specific haplogroup networks)

LIST OF ABBREVIATIONS AND SYMBOLS

12S	12 Svedberg units
12S rRNA	12S ribosomal RNA
16	16S rRNA (in L0-specific haplogroup networks)
16S	16 Svedberg units
16S rRNA	16S ribosomal RNA
A	alanine (in amino acid sequence)
A, A or a	adenine (in DNA sequence)
A.D.	Anno Domini: of the Christian era
A ₂₆₀	absorbance of samples at 260 nm
A ₂₈₀	absorbance of samples at 280 nm
A ₂₆₀ /A ₂₈₀	ratio of absorbency measured at 260 nm and 280 nm
ADP	adenosine diphosphate
AF	African mtDNA haplogroups
.aln	CLUSTAL format file extension
Ala	alanine
<i>A</i> <i>lu</i> I	restriction endonuclease isolated from <i>Arthrobacter luteus</i> with recognition site 5'- AG↓CT -3'
AMP	adenosine monophosphate
Arg	arginine
AS	Asian mtDNA haplogroups
Asn	asparagine
Asp	aspartic acid
ATP	adenosine-5-triphosphate
ATP6	ATP synthase F ₀ subunit 6 (in L0-specific haplogroup networks)
ATP8	ATP synthase F ₀ subunit 8 (in L0-specific haplogroup networks)
<i>atp6</i>	gene encoding ATPase6
<i>atp8</i>	gene encoding ATPase8
ATPase6	adenosine triphosphatase subunit 6 or ATP synthase F ₀ subunit 6
ATPase8	adenosine triphosphatase subunit 8 or ATP synthase F ₀ subunit 8
ATT	membrane attachment site
<i>A</i> <i>va</i> II	restriction endonuclease isolated from <i>Anabaena variabilis</i> with recognition site 5'- G↓G(A/T)CC -3'
b	branch length
<i>B</i> <i>am</i> HI	restriction endonuclease isolated from <i>Bacillus amyloliquefaciens</i> H with recognition site 5'- G↓GATCC -3'
boric acid	boracic acid: H ₃ BO ₃
bp	base pair(s)
<i>B</i> <i>st</i> <i>N</i> I	restriction endonuclease isolated from <i>Bacillus stearothermophilus</i> N with recognition site 5'- CC↓(A/T)GG -3'
C	cysteine (in amino acid sequence)
C, C or c	cytosine (in DNA sequence)
CA	California
(CA) _n	cytosine and adenine nucleotide repeat stretch
ca.	circa: approximately
Ca ²⁺	calcium (II) ion
CAP	CONTIG ASSEMBLY PROGRAM (included in BioEdit version 5.09)
CAR	Central African Republic
CBI	codon bias index
CGR	Centre for Genome Research
CI	conservation index
CM	cardiomyopathy
CNS	central nervous system
CO	cytochrome c oxidase subunit
CO ₂	carbon dioxide
COI	cytochrome c oxidase subunit I
COII	cytochrome c oxidase subunit II
COIII	cytochrome c oxidase subunit III
CoQ	coenzyme Q or ubiquinone
CoQH ₂	reduced coenzyme Q
Cov	covariance
CPEO	chronic progressive external ophthalmoplegia
CR	control region
CRS	Cambridge Reference Sequence
CSB1	conserved sequence block 1

CSB2	conserved sequence block 2
CSB3	conserved sequence block 3
CT	Connecticut
Cu _A	copper atom
Cu _B	copper atom
Cys	cysteine (in amino acid sequence)
Cytb	cytochrome <i>b</i>
cytb	cytochrome <i>b</i> (in L0-specific haplogroup network)
Cytc	cytochrome <i>c</i>
Δ-mtDNA	deleted mitochondrial DNA
<i>d</i>	maximum number of differences
<i>D</i>	Tajima's <i>D</i> test statistic
D	adenine, guanine or thymine (in DNA sequence)
D	aspartic acid
d	deletion
<i>D</i> *	Fu and Li's <i>D</i> * test statistic
dATP	2'-deoxyadenosine-5'-triphosphate
DC	District of Columbia
dCTP	2'-deoxycytidine-5'-triphosphate
ddATP	2',3'-dideoxyadenosine-5'-triphosphate
<i>Dde</i> I	restriction endonuclease isolated from <i>Desulfovibrio desulfuricans</i> with recognition site 5'- C↓TNAG -3'
ddH ₂ O	double distilled water
ddNTP(s)	2',3'-dideoxynucleotide triphosphate(s)
DEL or del	deletion
DGGE	denaturing gradient-gel electrophoresis
dGTP	2'-deoxyguanosine-5'-triphosphate
<i>d_{ij}</i> and <i>d_{kl}</i>	estimate of the number of nucleotide substitutions per site between DNA sequences <i>i</i> (<i>k</i>) and <i>j</i> (<i>l</i>)
D-loop	displacement loop
DNA	deoxyribonucleic acid
DnaSP	DNA Sequence Polymorphism
dNTP(s)	deoxynucleotide triphosphate(s)
dsDNA	double stranded DNA
DTT	dithiothreitol-1,4-dimercapto-2,3-butanediol: C ₄ H ₁₀ O ₂ S ₂
dTTP	2'-deoxythymidine-5'-triphosphate
E	glutamic acid
e ⁻	electron
e.g.	exempli gratia: for example
ECM	encephalomyopathy
EDTA	ethylenediamine tetra-acetic acid: C ₁₀ H ₁₆ N ₂ O ₈
.emf	extended (enhanced) Windows metafile format file extension (to produce phylogenetic tree images)
emPCR	emulsion-based clonal amplification
ENT	ear-nose-throat
<i>et al.</i>	et alia: and other people/things
EtBr	ethidium bromide: 2,7-diamino-10-ethyl-9-phenyl-phenanthridinium bromide: C ₂₁ H ₂₀ BrN ₃
EtOH	ethanol: CH ₃ CH ₂ OH
EU	European mtDNA haplogroups
F	forward
F	phenylalanine
<i>F</i> *	Fu and Li's <i>F</i> * test statistic
F ₀	component of ATP synthase, oligomeric enzyme complex located in integral membrane
F ₁	component of ATP synthase, water-soluble oligomeric enzyme complex bound to F ₀
.fas	FASTA format file extension
FAD	flavin adenine dinucleotide (oxidised form)
FADH ₂	flavin adenine dinucleotide (reduced form)
FBSN	familial bilateral striatal necrosis
Fe-S	iron-sulphur protein
<i>F_i(t)</i>	the probability that two random neutral genes will differ at exactly <i>i</i> nucleotides in generation <i>t</i>
fMet	<i>N</i> -formyl-methionine

LIST OF ABBREVIATIONS AND SYMBOLS

FMN	reduced flavin mononucleotide
formamide	carbamide: CH ₃ NO
F _{ST}	fixation index (a measure of population subdivision)
G	glycine
<i>g</i>	gravitational force
G + C	G + C content: refers to the composition of nucleotide sequences, specifically to the number of cytosine and guanine nucleotides
G + C ₂	G + C content at second codon positions
G + C _{3s}	G + C content at third (synonymous) codon positions
G + C _c	G + C content at coding positions
G + C _n	G + C content at non-coding positions
G, G or g	guanine (in DNA sequence)
g.cm ⁻³	gram per cubic centimetre
gDNA	genomic DNA
GenBank ^{®1}	NIH genetic sequence database, an annotated collection of all publicly available DNA sequences
GI	gastrointestinal system
GI number	GenInfo Identifier sequence identification number
Gln	glutamine
Glu	glutamate or glutamic acid
Gly	glycine
GTP	guanosine-5'-triphosphate
GTPases	a large family of hydrolase enzymes that can bind and hydrolyze GTP
H	histidine
h	hour(s)
H ⁺	proton
H ₂ O	water
<i>Hae</i> II	restriction endonuclease isolated from a recombinant <i>E. coli</i> strain with recognition site 5'- (A/G)GCGC↓(T/C) -3'
<i>Hae</i> III	restriction endonuclease isolated from a recombinant <i>Haemophilus aegyptius</i> with recognition site 5'- GG↓CC -3'
Haeme <i>a</i>	<i>a</i> -type haeme
Haeme <i>a</i> ₃	<i>a</i> ₃ -type haeme
Haeme <i>b</i> _H	high potential <i>b</i> -type haeme
Haeme <i>b</i> _L	low potential <i>b</i> -type haeme
Haeme <i>c</i> ₁	<i>c</i> -type haeme
HeLa	human cervical carcinoma
<i>Hha</i> I	restriction endonuclease isolated from <i>Haemophilus haemolyticus</i> with recognition site 5'- GCG↓C -3'
<i>Hin</i> fl	restriction endonuclease isolated from <i>Haemophilus influenzae</i> Rf with recognition site 5'- G↓ANTC -3'
<i>Hinc</i> II	restriction endonuclease isolated from <i>Haemophilus influenzae</i> Rc with recognition site 5'- GT(T/C)↓(A/G)AC -3'
His	histidine
HMG	high mobility group
h-mtTF-1	mitochondrial transcription factor, involved in binding to enhancer elements
h-mtTFA	mitochondrial transcription factor, involved in binding to enhancer elements
h-mtTFB	mitochondrial transcription factor, involved in binding to enhancer elements, also referred to as TFB2M
H ₀	null hypothesis
<i>Hpa</i> I	restriction endonuclease isolated from <i>Haemophilus parainfluenzae</i> with recognition site 5'- GTT↓AAC -3'
HR	high-resolution
HSP	H-strand promoter
HSP1	major H-strand promoter
HSP2	minor H-strand promoter
H-strand	heavy strand of the mitochondrial DNA molecule
HVS1	hypervariable segment 1
HVS2	hypervariable segment 2
HVS3	hypervariable segment 3
i	insertion

¹ GenBank[®] is a registered trademark of the U.S. Department of Health and Human Services, Independence Avenue, S.W., Washington, DC, USA.

I	internal branch (in L0-specific haplogroup network)
I	isoleucine
i.e.	id est: that is to say
I/T	RF _I /RF _T
IgE	immunoglobulin E
Ile	isoleucine
IMM	inner mitochondrial membrane
IN	Indiana
INS	insertion
IT _{H1}	major initiation site for H-strand transcription
IT _{H2}	minor initiation site for H-strand transcription
IT _L	initiation site for L-strand transcription
<i>k</i>	average number of pairwise nucleotide differences or number of alleles
K	lysine (in amino acid sequence)
K, K or k	guanine or thymine (in DNA sequence)
K, K or k	guanine or uracil (in RNA sequence)
kb	kilobase pair
kDa	kilodalton
<i>k_{ij}</i>	number of nucleotide differences between the <i>i</i> th and <i>j</i> th sequences
KS_CGR_###	three digit sample number given to each Khoi-San individual
KSC	Khoi-San consensus sequence
KSS	Kearns-Sayre syndrome
L	leucine
L0aA	L0a (branch A)
L0aA1	L0a (branch A, group 1)
L0aA2	L0a (branch A, group 2)
L0aB	L0a (branch B)
L0aB1	L0a (branch B, group 1)
L0aB2	L0a (branch B, group 2)
L0aC	L0a (branch C)
L0bA	L0b (branch A)
L0bA1	L0b (branch A, group 1)
L0bA2	L0b (branch A, group 2)
L0bB	L0b (branch B)
L0bB1	L0b (branch B, group 1)
L0bB1a	L0b (branch B, group 1, part a)
L0bB1b	L0b (branch B, group 1, part b)
L0bB1b1	L0b (branch B, group 1, part b, section 1)
L0bB1b2	L0b (branch B, group 1, part b, section 2)
L0bB1b3	L0b (branch B, group 1, part b, section 3)
L0bB2	L0b (branch B, group 2)
L0bB2a	L0b (branch B, group 2, part a)
L0bB2a1	L0b (branch B, group 2, part a, section 1)
L0bB2a2	L0b (branch B, group 2, part a, section 2)
L0bB2a3	L0b (branch B, group 2, part a, section 3)
L0bB2b	L0b (branch B, group 2, part b)
L0bB3	L0b (branch B, group 3)
L0bB3a	L0b (branch B, group 3, part a)
L0bB3b	L0b (branch B, group 3, part b)
L0c1A	L0c1 (branch A)
L0c1B	L0c1 (branch B)
L0c2A	L0c2 (branch A)
L0c2ABCE	L0c2 (branch ABCE)
L0c2ACE	L0c2 (branch ACE)
L0c2B	L0c2 (branch B)
L0c2B1	L0c2 (branch B, group 1)
L0c2B2	L0c2 (branch B, group 2)
L0c2C	L0c2 (branch C)
L0c2D	L0c2 (branch D)
L0c2D1	L0c2 (branch D, group 1)
L0c2D2	L0c2 (branch D, group 2)
L0c2E	L0c2 (branch E)
Leu	leucine
LHON	Leber's hereditary optic neuropathy

LIST OF ABBREVIATIONS AND SYMBOLS

LR	low-resolution
LS	Leigh syndrome
LSP	L-strand promoter
L-strand	light strand of the mitochondrial DNA molecule
Ltd.	limited
Lys	lysine
M	methionine
M	molar: moles per litre
M	myopathy
<i>m</i>	number of nucleotides examined per sequence or rate of migration
M, M or m	adenine or cytosine (in DNA sequence)
µg	microgram
µL	microlitre
µm	micrometre
µM	micromolar
MA	Massachusetts
MALDI-TOF MS	matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry
Max.	maximum
<i>Mbo</i> I	restriction endonuclease isolated from <i>Moraxella bovis</i> ATCC 10900 with recognition site 5'- ↓GATC -3'
MD	Maryland
ME	Multiregional Evolution
.meg	MEGA alignment file
MEGA	Molecular Evolutionary Genetics Analysis
MELAS	mitochondrial encephalomyopathy, lactic acidosis, and stroke-like episodes
MERRF	myoclonic epilepsy with ragged-red fibres
Met	methionine
Mfold	multiple fold program
mg	milligram
MgCl ₂	magnesium chloride
[MgCl ₂]	magnesium chloride concentration
MILS	maternally inherited Leigh syndrome
min	minute(s)
Min.	minimum
mL	millilitre
mm	millimetre
mM	millimolar
MM	molecular marker: FastRuler™ DNA Ladder, High Range, ready to use
MN	Minnesota
<i>Mnl</i> I	restriction endonuclease isolated from an <i>E. coli</i> strain that carries the <i>Mnl</i> I gene from <i>Moraxella nonliquefaciens</i> (ATCC 17953) with recognition site 5'- CCTC(N) ₇ ↓ -3'
MO	Missouri
mol	mole: unit describing the amount of a particular chemical species; the amount being equal to one Avogadro's number (6.02 x 10 ²³) of atoms, ions, molecules, or electrons
MP	maximum parsimony
MRCA	most recent common ancestor
mRNA	messenger RNA
<i>Msp</i> I	restriction endonuclease isolated from a <i>Moraxella</i> species with recognition site 5'- C↓CGG -3'
mt3H	mt3 H-strand control element
mt3L	L-strand control element
mt4H	mt4 H-strand control element
Mt5	control element
MT-CO1	cytochrome c oxidase subunit I
MT-CO2	cytochrome c oxidase subunit II
MT-CO3	cytochrome c oxidase subunit III
MT-CYTB	cytochrome <i>b</i>
MT-DLOOP	control region, including displacement loop
mtDNA	mitochondrial deoxyribonucleic acid
mtEF-G	mitochondrial elongation factor
mtEF-Ts	mitochondrial elongation factor
mtEF-Tu	mitochondrial elongation factor
mTERF	mitochondrial termination factor

mtETC	mitochondrial electron transport chain
mtIF-2	mitochondrial initiation factor
mt-MRCA	mitochondrial most recent common ancestor
MT-ND1	NADH dehydrogenase subunit 1
MT-ND2	NADH dehydrogenase subunit 2
MT-ND3	NADH dehydrogenase subunit 3
MT-ND4	NADH dehydrogenase subunit 4
MT-ND4L	NADH dehydrogenase subunit 4L
MT-ND5	NADH dehydrogenase subunit 5
MT-ND6	NADH dehydrogenase subunit 6
mtRNase P	mitochondrial RNase P
MT-RNR1	12S ribosomal RNA
MT-RNR2	16S ribosomal RNA
mtSSB	mitochondrial single-stranded binding proteins
MT-TC	tRNA cysteine
MT-TE	tRNA glutamic acid
mtTERM	mitochondrial transcription terminator
MT-TF	tRNA phenylalanine
mtTF1	mitochondrial transcription factor
MT-TG	tRNA glycine
MT-TH	tRNA histidine
MT-TI	tRNA isoleucine
MT-TL1	tRNA leucine (UUR)
MT-TL2	tRNA leucine (CUN)
MT-TQ	tRNA glutamine
MT-TS1	tRNA serine (UCN)
MT-TS2	tRNA serine (AGY)
MT-TT	tRNA threonine
MT-TV	tRNA valine
MT-TW	tRNA tryptophan
MT-TY	tRNA tyrosine
Myr	million years
N	asparagine
η_s	number of singletons (mutations appearing only once among the nucleotide sequences)
N, N or n	adenine, cytosine, guanine or thymine (in DNA sequence)
N, N or n	adenine, cytosine, guanine or uracil (in RNA sequence)
N, <i>N</i> , <i>n</i> or <i>n</i>	sample size
N/A	not available
NA	Native American mtDNA haplogroups
Na ₂ EDTA	di-sodium ethylenediamine tetra-acetic acid: C ₁₀ H ₁₄ N ₂ Na ₂ O ₈ ·2H ₂ O
NaCl	sodium chloride
NAD ⁺	oxidised nicotinamide adenine dinucleotide
NADH	reduced nicotinamide adenine dinucleotide
NaOAc	sodium acetate
NARP	neuropathy, ataxia, and retinitis pigmentosa
N _c	effective number of codons
NC	non-coding
NC	North Carolina
NC1	non-coding nucleotide 1
NC2	non-coding nucleotide 2
NC3	non-coding nucleotide 3
NC4	non-coding nucleotide 4
NC5	non-coding nucleotide 5
NC6	non-coding nucleotide 6
NC7	non-coding nucleotide 7
NC8	non-coding nucleotide 8
NC9	non-coding nucleotide 9
NC10	non-coding nucleotide 10
NCBI	National Center for Biotechnology Information
ND	NADH dehydrogenase subunit
ND1	NADH dehydrogenase subunit 1
ND2	NADH dehydrogenase subunit 2
ND3	NADH dehydrogenase subunit 3

LIST OF ABBREVIATIONS AND SYMBOLS

ND4	NADH dehydrogenase subunit 4
ND4L	NADH dehydrogenase subunit 4L
ND5	NADH dehydrogenase subunit 5
ND6	NADH dehydrogenase subunit 6
<i>ndh4</i>	gene encoding ND4
<i>ndh4L</i>	gene encoding ND4L
nDNA	nuclear DNA
N_e	effective population size
NEG	negative control
ng	nanogram
NIH	National Institutes of Health, USA
NJ	New Jersey
N-J	neighbour-joining
<i>Nla</i> III	restriction endonuclease isolated from an <i>E. coli</i> strain that carries the <i>Nla</i> III gene from <i>Neisseria lactamica</i> (NRCC 2118) with recognition site 5' -GATG↓ -3'
nm	nanometre
nM	nanomolar
No.	number
NON-SYN	non-synonymous (describing a nucleotide substitution)
NRV	non-recombining portion of the Y chromosome
NS	non-synonymous (describing a nucleotide substitution)
Nucl #	nucleotide number
O.D.	optical density
O ₂	molecular oxygen
OC	Oceanic mtDNA haplogroups
O _H	H-strand origin of replication
O _L	L-strand origin of replication
OMM	outer mitochondrial membrane
Orange G	dye used in preparation of loading dye, 7-hydroxy-8-phenylazo-1,3-naphthalenedisulfonic acid: C ₁₆ H ₁₀ N ₂ O ₇ S ₂ Na ₂
OTUs	operational taxonomic units
OXPHOS	oxidative phosphorylation
P	proline
<i>P</i>	<i>P</i> -value
$p(k \theta)$	probability of having <i>k</i> alleles in a sample of <i>n</i> sequences
PA	Pennsylvania
PCR	polymerase chain reaction
PDF	portable document format
PEO	progressive external ophthalmoplegia
<i>Pfu</i> DNA polymerase	deoxynucleoside-triphosphate: DNA deoxynucleotidyltransferase from <i>Pyrococcus furiosus</i> : EC 2.7.7.7
pH	a measure of acidity: numerically equal to the negative logarithm of H ⁺ concentration expressed in molarity
.phy	PHYLP format file extension
P _{H1}	major H-strand promoter
P _{H2}	minor H-strand promoter
Phe	phenylalanine
PHYLP	Phylogeny Inference Package
P _i	inorganic phosphate.
.pir	NBRF/PIR format file extension
P _L	L-strand promoter
pmol	picomole
PNS	peripheral nervous system
POS	positive control
PPK	palmoplantar keratoderma
Pro	proline
PS	Pearson syndrome
Pty.	Proprietary
<i>P</i> -value	probability value, indicates statistical significance
Q	glutamine
R	arginine
r	average radius of the rotor in millimetre
R	reverse
R, R or r	adenine or guanine (in DNA sequence)

R_2	Ramos-Onsins and Rozas statistic
RAO	Recent African Origin
rCRS	revised Cambridge Reference Sequence
rDNA	ribosomal DNA
RE	restriction enzyme
RefSeq	NCBI reference sequence
RF	replacement mutation frequency
RF_i	frequency of replacement mutations in internal branches
RFLP(s)	restriction fragment length polymorphism(s)
RF_T	frequency of replacement mutations in terminal branches
rg	raggedness statistic
RNA	ribonucleic acid
RNase	ribonuclease
RNase MRP	mitochondrial RNA processing endonuclease
ROS	reactive oxygen species
rpm	revolutions per minute
rRNA(s)	ribosomal RNA(s)
RSCU	relative synonymous codon usage
S	number of segregating (polymorphic) sites
S	serine
S	Svedberg units
S	synonymous (describing nucleotide substitution)
S, S or s	cytosine or guanine (in DNA sequence)
S.W.	South West
S'	probability of having no fewer than k_0 alleles in a sample provided that $\theta = \pi$
SAK	Southern African Khoi-San
$S\chi^2$	scaled chi square
SD	standard deviation
SDS	sodium dodecyl sulphate: $C_{12}H_{25}NaSO_4$
sec	second(s)
Ser	serine
SIDS	sudden infant death syndrome
S_k	coefficient of θ^k_π in S_n
SNP(s)	single nucleotide polymorphism(s)
SSCP	single stranded conformational polymorphism
ssDNA	single stranded DNA
STR	short, tandemly repeated
$\sum(y - y)^2$	sum of squared deviations
SYN	synonymous (describing a nucleotide substitution)
t	time in generations
T	terminal branch (in LO-specific haplogroup network)
T	threonine
T Ψ C loop	the loop region of the tRNA molecule containing pseudouridine (a modified uracil nucleotide in a UUCG sequence)
T, T or t	thymine (in DNA sequence)
T_a	annealing temperature
<i>Taq</i> polymerase	deoxynucleosidetriphosphate: DNA deoxynucleotidyltransferase from <i>Thermus aquaticus</i> : EC 2.7.7.7
TAS	termination-associated sequence
TBE	Tris [®] borate-EDTA buffer
Ter	termination
Tfam	mitochondrial transcription factor, involved in binding to enhancer elements (previously known as h-mtTF-1 or h-mtTFA)
TFB1M	mitochondrial transcription specificity factor
TFB2M	mitochondrial transcription specificity factor, also referred to as h-mtTFB
TFX	mtTF1 binding site
TFY	mtTF1 binding site
Thr	threonine
T_m	melting temperature
TMRCA(s)	time(s) to most recent common ancestor
Tris ^{®1}	tris(hydroxymethyl)aminomethan: 2-amino-2-(hydroxymethyl)-1,3-propanediol: $C_4H_{11}NO_3$

¹ Tris[®] is a registered trademark of Rohm and Haas Company, Philadelphia, PA, USA.

LIST OF ABBREVIATIONS AND SYMBOLS

Tris [®] -HCl	2-amino-2(hydroxymethyl)-1,3-propanediol hydrochloride: C ₄ H ₁₁ NO ₃ .H ₂ O
tRNA(s)	transfer RNA(s)
tRNA ^{Ala}	tRNA alanine
tRNA ^{Arg}	tRNA arginine
tRNA ^{Asn}	tRNA asparagine
tRNA ^{Asp}	tRNA aspartic acid
tRNA ^{Cys}	tRNA cysteine
tRNA ^{Gln}	tRNA glutamine
tRNA ^{Glu}	tRNA glutamic acid
tRNA ^{Gly}	tRNA glycine
tRNA ^{His}	tRNA histidine
tRNA ^{Ile}	tRNA isoleucine
tRNA ^{Leu}	tRNA leucine
tRNA ^{Leu(CUN)}	tRNA leucine (CUN)
tRNA ^{Leu(UUR)}	tRNA leucine (UUR)
tRNA ^{Lys}	tRNA lysine
tRNA ^{Met}	tRNA methionine
tRNA ^{Phe}	tRNA phenylalanine
tRNA ^{Pro}	tRNA proline
tRNA ^{Ser}	tRNA serine
tRNA ^{Ser(AGY)}	tRNA serine (AGY)
tRNA ^{Ser(UCN)}	tRNA serine (UCN)
tRNA ^{Thr}	tRNA threonine
tRNA ^{Trp}	tRNA tryptophan
tRNA ^{Tyr}	tRNA tyrosine
tRNA ^{Val}	tRNA valine
Trp	tryptophan
TS	transition
TV	transversion
Tyr	tyrosine
U	unit(s)
U, U or u	uracil (in RNA sequence)
U.S.	United States
<i>U_i</i>	number of singleton mutations in sequence <i>i</i>
UK	United Kingdom
UPGMA	unweighted pair group method with arithmetic mean
UQ	coenzyme Q or ubiquinone
USA	United States of America
UV	ultraviolet
V	valine
V	volt
$V(\pi)$	sample variance of nucleotide diversity
V.cm ⁻¹	volts per centimetre
v/v	volume per volume
Val	valine
vs.	versus: against
W	tryptophan
W, W or w	adenine or thymine (in DNA sequence)
W, W or w	adenine or uracil (in RNA sequence)
w/v	weight per volume
WI	Wisconsin
.wmf	Windows metafile format file extension (to produce phylogenetic tree images)
x	mismatch distribution
y	value of observation
Y	tyrosine
Y, Y or y	cytosine or thymine (in DNA sequence)
Y, Y or y	cytosine or uracil (in RNA sequence)
YBP	years before present
Z	left-handed helical conformation (when referring to the structure of DNA)

LIST OF EQUATIONS

Equation No.	Title of Equation	Page
4.1	Conversion of gravitational force to revolutions per minute.....	123
4.2	Calculation of DNA concentration from the absorbance at 260 nm.....	130
4.3	Calculation of arithmetic mean and standard deviation.....	156
4.4	Average number of nucleotide differences.....	157
4.5	Calculation of nucleotide diversity, sampling variance and standard deviation.....	158
4.6	Tajima's D statistic.....	159
4.7	Fu and Li's D^* and F^* test statistics.....	161
4.8	Calculation of the pairwise number of sequence differences under the assumption of constant population size.....	162
4.9	Raggedness statistic.....	163
4.10	The Ramos-Onsins and Rozas statistic.....	163
4.11	Fu's F_S statistic.....	164
4.12	Calculation of the replacement mutation frequency.....	166
4.13	Calculation of the average number of sites differing between a set of sequences and a specified common ancestor.....	168
4.14	Calculation of the TMRCA.....	169
4.15	Calculation of standard deviation of the TMRCA.....	169

LIST OF FIGURES

Figure No.	Title of Figure	Page
2.1	The structure of the mitochondrion.....	8
2.2	Schematic representation of the mitochondrial electron transport chain.....	12
2.3	Diagrammatic representation of the mitochondrial genome.....	16
2.4	Diagrammatic representation of the replication cycle of mitochondrial DNA ..	19
2.5	Morbidity map of the human mitochondrial genome.....	47
3.1	Different models of human evolution.....	53
3.2	Human migration pattern based on regional distribution of mitochondrial haplogroups.....	56
3.3	Spread of Bantu-speakers in Africa during the first movement of the Bantu expansion during the Early Iron Age.....	61
3.4	Spread of Bantu-speakers in Africa during the second movement of the Bantu expansion.....	62
3.5	A schematic representation of a global phylogenetic tree of mtDNA haplogroups.....	83
3.6	A schematic representation of the mtDNA phylogeny of African mitochondrial haplogroups.....	84
3.7	Classification network for African mtDNA haplogroups.....	86
3.8	The Khoi-San language classification system.....	117
4.1	Schematic representation of the single nucleotide polymorphisms used to characterise African mtDNA L haplogroups.....	139
5.1	Location of the collection site and the geographic origin of the Khoi-San individuals studied in the current investigation.....	175
5.2	Photographic representation of the variation in amplification efficiency observed between amplified PCR products for different samples.....	180
5.3	Photographic representation of the variation in amplification efficiency observed between different amplified PCR regions.....	181
5.4	Photographic representation of the background smear observed for amplified PCR products.....	183
5.5	Photographic representation of secondary amplification observed for amplified PCR products.....	184
5.6	Photographic representation of primer-dimers observed for amplified PCR products.....	185
5.7	Photographic representation of an artefact observed in the gel matrix.....	186
5.8	Photographic representation of the distortion of amplified fragments.....	187
5.9	Photographic representation of slanted fragments.....	188
5.10	Representative electropherogram indicating background noise.....	194

LIST OF FIGURES

5.11	Representative electropherogram depicting low signal strength.....	196
5.12	Representative electropherogram illustrating poor mobility correction.....	196
5.13	Representative electropherogram illustrating the insertion and deletion of nucleotides.....	197
5.14	Representative electropherogram illustrating excess dye peaks.....	198
5.15	Representative electropherogram depicting dye blobs.....	199
5.16	Representative electropherogram depicting slippage in a homopolymer region.....	200
5.17	Photographic representation of amplified products of PCR region 1.....	205
5.18	Representative electropherograms of PCR region 1 generated using forward primers F32 to F3.....	206
5.19	Representative electropherograms of PCR region 1 generated using reverse primers R32 to R3.....	207
5.20	Proposed secondary structure and sequence alignment of tRNA ^{Phe} containing the A647G alteration.....	212
5.21	Photographic representation of amplified products of PCR region 2.....	215
5.22	Representative electropherograms of PCR region 2 generated using forward primers F4* to F7.....	216
5.23	A representative electropherogram of PCR region 2 generated using reverse primer R4.....	217
5.24	Photographic representation of amplified products of PCR region 3.....	220
5.25	Representative electropherograms of PCR region 3 generated using forward primers F8* to F11.....	221
5.26	A representative electropherogram of PCR region 3 generated using reverse primer R8.....	222
5.27	Proposed secondary structure and sequence alignment of tRNA ^{Trp} containing the A5515G alteration.....	223
5.28	Photographic representation of amplified products of PCR region 4.....	227
5.29	Representative electropherograms of PCR region 4 generated using forward primers F12* to F15.....	228
5.30	Representative electropherogram of PCR region 4 generated using the reverse primer R15.....	230
5.31	Proposed secondary structures and sequence alignment of tRNA ^{Tyr} containing the T5865A alteration.....	230
5.32	Photographic representation of amplified products of PCR region 5.....	235
5.33	Representative electropherograms of PCR region 5 generated using forward primers F16* to F19.....	236
5.34	A representative electropherogram of PCR region 5 generated using reverse primer R16.....	237
5.35	Photographic representation of amplified products of PCR region 6.....	244
5.36	Representative electropherograms of PCR region 6 generated using forward primers F20 to F23.....	245
5.37	Photographic representation of amplified products of PCR region 7.....	248
5.38	Representative electropherograms of PCR region 7 generated using forward primers F24* to F27.....	249

5.39	Representative electropherograms of PCR region 7 generated using reverse primers R24 to R27.....	250
5.40	Photographic representation of amplified products of PCR region 8.....	258
5.41	Representative electropherograms of PCR region 8 generated using forward primers F28* to F31.....	259
5.42	Global neighbour-joining tree constructed using the All L Sequences dataset (dataset 1a).....	279
5.43	Africa neighbour-joining tree constructed using the All Africa L Sequences dataset (dataset 2a).....	298
5.44	Khoi-San neighbour-joining tree constructed using the All Khoi-San L Sequences dataset (dataset 3a).....	307
5.45	Global maximum parsimony tree constructed using the All L Sequences dataset (dataset 1a).....	311
5.46	Africa maximum parsimony tree constructed using the All Africa L Sequences dataset (dataset 2a).....	317
5.47	Khoi-San maximum parsimony tree constructed using the All Khoi-San L Sequences dataset (dataset 3a).....	321
5.48	Global maximum parsimony tree adapted to indicate the geographic origin of the mitochondrial genome sequences.....	326
5.49	L0-specific haplogroup network based on the 22 Khoi-San L0 sequences generated in the current investigation.....	333
5.50	L0a section of the global L0-specific haplogroup network after addition of 46 L0 sequences.....	337
5.51	L0b section of the global L0-specific haplogroup network after addition of 46 L0 sequences.....	340
5.52	L0c1 section of the global L0-specific haplogroup network after addition of 46 L0 sequences.....	342
5.53	L0c2 section of the global L0-specific haplogroup network after addition of 46 L0 sequences.....	344
5.54	Graphical representation of pairwise differences for datasets 1b, 2b and 5 which exhibited deviation from neutrality and displayed characteristics of population expansion.....	376
5.55	Graphical representation of pairwise differences for dataset 6 which exhibited deviation from neutrality but displayed no evidence of population expansion.....	380
5.56	Graphical representation of pairwise differences for dataset 7 which did not exhibit deviation from neutrality and displayed characteristics of constant population size.....	381
5.57	Graphical representation of pairwise differences for datasets 3a and 4b which did not exhibit deviation from neutrality and displayed characteristics of constant population size.....	382
5.58	Graphical representation of pairwise differences for dataset 8a which did not exhibit deviation from neutrality and displayed characteristics of constant population size.....	383
5.59	Geographic distribution of L0 sub-haplogroups and lineages within sub-haplogroups in Africa.....	413

LIST OF FIGURES

5.60	Reduced global maximum parsimony tree based on the MP tree constructed using the All L Sequences dataset (dataset 1a).....	416
5.61	Revised classification scheme for the L0 haplogroup.....	421
5.62	Consensus sequence generated for the Khoi-San of southern Africa presented in the form of a list of SNPs compared to the rCRS.....	429
5.63	Schematic representation of the global L0-specific haplogroup network indicating new branches after the addition of 46 L0 sequences.....	454
6.1	Model indicating the different factors that influence the genetic variation observed in human mitochondrial genome sequences.....	478
6.2	Updated model indicating the contribution made by the unique population characteristics of the Khoi-San from southern Africa.....	491
C.1	Schematic representation of a global phylogenetic tree of mtDNA haplogroups.....	565
I.1	Geographic regions in Africa.....	597
J.1	Global and Africa phylogenetic trees.....	600
K.1	L0 haplogroup network obtained from MITOMAP.....	610
L.1	Global L0-specific haplogroup network.....	612
N.1	Mitochondrial consensus sequence for the Khoi-San population of southern Africa.....	621

LIST OF TABLES

Table No.	Title of Table	Page
2.1	The multi-subunit complexes which play a role in the electron transport chain.....	10
2.2	Subunits of the electron transport system encoded by mitochondrial genes.....	11
2.3	Functional elements located in the mitochondrial DNA genome.....	14
2.4	The mitochondrial genetic code.....	17
2.5	Genes encoded by the H-strand and the L-strand of the mitochondrial DNA genome.....	21
2.6	Differences between the genetic code of mammalian mitochondria and the universal genetic code.....	25
2.7	Evidence in support or rejection of recombination events within mitochondria.....	33
2.8	Adaptive mitochondrial mutations associated with specific haplogroups.....	42
2.9	Clinical features in mitochondrial diseases associated with mtDNA mutations.....	44
2.10	The genetic classification of human mitochondrial disorders.....	46
2.11	A selection of previously identified pathogenic mtDNA mutations.....	48
3.1	Regions in Africa that contributed slaves to other parts of the world.....	66
3.2	Strategies that can be employed to avoid the inclusion of errors in mitochondrial sequence datasets.....	72
3.3	Evolutionary models that describe the nucleotide substitution process.....	74
3.4	Estimates of the α parameter of the gamma distribution of rate variation for mitochondrial regions or genes.....	75
3.5	Aspects to take into consideration when deciding on the most appropriate method for use in constructing phylogenetic trees.....	78
3.6	Restriction enzyme sites defining continent-specific African mitochondrial haplogroups.....	85
3.7	Geographic distribution of major African mitochondrial haplogroups.....	88
3.8	Population-specific mitochondrial macrohaplogroup L lineages.....	89
3.9	Sequence divergence times for major African mitochondrial haplogroups.....	90
3.10	Restriction enzyme sites used to define continent-specific Asian mtDNA haplogroups.....	96
3.11	Distribution of haplogroups A, B, C, D, E, F and G.....	97
3.12	Divergence times of the major Asian mitochondrial haplogroups.....	97
3.13	Restriction enzyme sites used to define continent-specific Native American mtDNA haplogroups.....	98

LIST OF TABLES

3.14	Single nucleotide polymorphisms used to define continent-specific Native American mtDNA haplogroups.....	99
3.15	Sequence divergence times for Native American mitochondrial haplogroups.....	100
3.16	Single nucleotide polymorphisms used to define continent-specific Oceanic mtDNA haplogroups.....	102
3.17	Divergence times of some of the major Oceanic mitochondrial haplogroups.....	103
3.18	Restriction enzyme sites used to define continent-specific European mtDNA haplogroups.....	105
3.19	Divergence times of the major European mitochondrial haplogroups.....	105
3.20	Classification of haplogroup H into sub-haplogroups	106
3.21	Age estimates determined for sub-haplogroups of haplogroup H.....	107
3.22	Classification of haplogroup U into sub-haplogroups	107
3.23	Age estimates determined for selected sub-haplogroups of haplogroup U.....	108
3.24	Classification of haplogroup J into two sub-haplogroups	108
3.25	Classification of haplogroup T into two sub-haplogroups	108
3.26	mtDNA haplogroups which are associated with specific phenotypes and susceptibility to specific disorders.....	109
3.27	Characteristics of the Marzuki human mitochondrial consensus sequence.....	114
3.28	Characteristics of the Carter human mitochondrial consensus sequence.....	115
4.1	Primer pairs used for amplification of the complete mitochondrial genome.....	125
4.2	Polymerase chain reaction thermal cycling conditions.....	127
4.3	Estimation of the quantity of the purified PCR product to be used in a cycle sequencing reaction based on the size of the PCR product.....	131
4.4	Thermal cycling conditions to be used for sequence determination of segments of the mitochondrial genome.....	131
4.5	Primers used for sequencing of the complete mitochondrial genome.....	132
4.6	Thermal cycling conditions used during SDS/heat treatment of cycle sequencing reactions.....	134
4.7	Keywords and phrases used to recover published mitochondrial genome sequences belonging to macrohaplogroup L.....	138
4.8	Assignment of haplogroups to incompletely classified sequences included in the current investigation.....	140
4.9	Sequences obtained from GenBank® that were excluded from phylogenetic analyses.....	142
4.10	Description of all datasets used in the current investigation for analyses.....	145
4.11	Species used to calculate the conservation index.....	167
5.1	Primer pairs and PCR conditions used for amplification of the complete mitochondrial genome.....	177
5.2	The likely causes for the distortion of sample fragments and ways in which it can be prevented.....	187
5.3	The possible causes of slanted sample fragments and ways in which it can be prevented.....	189

5.4	The amount of purified PCR product template used for sequencing per region.....	192
5.5	Factors affecting template quality.....	208
5.6	Sequence alterations observed between the complete Khoi-San mitochondrial DNA genome sequences and the rCRS in the D-loop.....	208
5.7	Sequence alterations observed between the complete Khoi-San mitochondrial DNA genome sequences and the rCRS in the tRNA ^{Phe} gene.....	211
5.8	Sequence alterations observed between the complete Khoi-San mitochondrial DNA genome sequences and the rCRS in the 12S rRNA gene.....	213
5.9	Sequence alterations observed between the complete Khoi-San mitochondrial DNA genome sequences and the rCRS in the 16S rRNA gene.....	217
5.10	Sequence alterations observed between the complete Khoi-San mitochondrial DNA genome sequences and the rCRS in the ND1 gene.....	219
5.11	Sequence alterations observed between the complete Khoi-San mitochondrial DNA genome sequences and the rCRS in the tRNA genes.....	222
5.12	Sequence alterations observed between the complete Khoi-San mitochondrial DNA genome sequences and the rCRS in the ND2 gene.....	224
5.13	Sequence alteration observed between the complete Khoi-San mitochondrial DNA genome sequences and the rCRS in the tRNA ^{Tyr} gene.....	229
5.14	Sequence alterations observed between the complete Khoi-San mitochondrial DNA genome sequences and the rCRS in the COI gene.....	232
5.15	Sequence alteration observed between the complete Khoi-San mitochondrial DNA genome sequences and the rCRS in the tRNA ^{Ser(UCN)} gene.....	237
5.16	Sequence alterations observed between the complete Khoi-San mitochondrial DNA genome sequences and the rCRS in the COII gene.....	237
5.17	Sequence alterations observed between the complete Khoi-San mitochondrial DNA genome sequences and the rCRS in the ATPase8 gene.....	239
5.18	Sequence alterations observed between the complete Khoi-San mitochondrial DNA genome sequences and the rCRS in the ATPase6 gene.....	239
5.19	Sequence alterations observed between the complete Khoi-San mitochondrial DNA genome sequences and the rCRS in the COIII gene.....	242
5.20	Sequence alterations observed between the complete Khoi-San mitochondrial DNA genome sequences and the rCRS in the ND3 gene.....	243
5.21	Sequence alterations observed between the complete Khoi-San mitochondrial DNA genome sequences and the rCRS in the ND4L gene.....	246
5.22	Sequence alterations observed between the complete Khoi-San mitochondrial DNA genome sequences and the rCRS in the ND4 gene.....	247
5.23	Sequence alterations observed between the complete Khoi-San mitochondrial DNA genome sequences and the rCRS in tRNA genes.....	251

LIST OF TABLES

5.24	Sequence alterations observed between the complete Khoi-San mitochondrial DNA genome sequences and the rCRS in the ND5 gene.....	251
5.25	Sequence alterations observed between the complete Khoi-San mitochondrial DNA genome sequences and the rCRS in the ND6 gene.....	256
5.26	Sequence alterations observed between the complete Khoi-San mitochondrial DNA genome sequences and the rCRS in tRNA genes.....	260
5.27	Sequence alterations observed between the complete Khoi-San mitochondrial DNA genome sequences and the rCRS in the Cytb gene.....	261
5.28	Khoi-San genome sequences deviating from exclusion criteria for L-specific sequences.....	264
5.29	Assignment of haplogroups to Khoi-San individuals included in the current investigation.....	267
5.30	Possible reasons for the classification of Khoi-San individuals in L-specific haplogroups other than L0.....	267
5.31	Sequences obtained from GenBank® deviating from exclusion criteria for L-specific sequences.....	270
5.32	Parameters calculated for datasets used in phylogenetic analyses.....	277
5.33	Location of 30 Khoi-San sequences generated in the current investigation in the global N-J phylogenetic tree.....	280
5.34	Incompletely classified sequences and sequences that exhibit a discrepancy between their haplogroup classification and position in the global N-J phylogenetic tree.....	281
5.35	SNP profiles of the five sequences that exhibited a discrepancy between their haplogroup classification and the location of the sequence in the global N-J phylogenetic tree.....	282
5.36	African country of origin and ethnic affiliation of sequences included in dataset 2a.....	295
5.37	Geographic distribution of African mitochondrial haplogroups.....	327
5.38	Location of Khoi-San sequences in Khoi-San L0 network.....	334
5.39	Nucleotide position and location of recurrent alterations in the global L0-specific haplogroup network based on the mtDNA coding region.....	351
5.40	Number of recurrent alterations per region/gene in the global L0-specific haplogroup network based on the mtDNA coding region.....	353
5.41	Number of recurrent alterations per alteration type in the global L0-specific haplogroup network based on the mtDNA coding region.....	354
5.42	Nucleotide position and location of reversions in the global L0-specific haplogroup network based on the mtDNA coding region.....	356
5.43	Number of reversions per region/gene in the global L0-specific haplogroup network based on the mtDNA coding region.....	357
5.44	Number of reversions per alteration type in the global L0-specific haplogroup network based on the mtDNA coding region.....	357
5.45	Nucleotide composition and G + C content of the 43 Khoi-San mtDNA genome sequences in dataset 4a.....	359
5.46	Codon usage for the 43 coding region only Khoi-San mtDNA genome sequences in dataset 4b.....	362

5.47	Codon usage for the ND5 gene of the 43 coding region only Khoi-San mtDNA genome sequences in dataset 4b.....	364
5.48	Global codon usage bias measures for the 43 coding region only Khoi-San mtDNA genome sequences in dataset 4b.....	366
5.49	Summary of statistical parameters of human mtDNA genomes in different datasets.....	368
5.50	Datasets sorted according to decreasing nucleotide diversity.....	369
5.51	Tests evaluating the departure from neutrality in different datasets.....	372
5.52	Statistical parameters to evaluate population size changes in different datasets.....	387
5.53	Selection parameters for the L-haplogroups.....	393
5.54	Non-synonymous substitutions identified within the constructed global L0-specific haplogroup network based on the coding region of 100 mtDNA genome sequences.....	395
5.55	TMRCA estimated for haplogroups, sub-haplogroups and lineages within L0.....	404
5.56	Estimates of TMRCA of haplogroups and sub-haplogroups in other investigations.....	408
5.57	Nucleotide positions in the mitochondrial coding region at which recurrent alterations were observed, or which have been classified as hypervariable.....	419
5.58	Nucleotide positions selected for inclusion in the revised L0 haplogroup classification scheme.....	420
5.59	Differences in the revised L0 classification scheme compared to the classification scheme used in the current investigation.....	422
5.60	Comparison of haplogroup and sub-haplogroup designations between three classification schemes.....	423
5.61	Assignment of haplogroups and sub-haplogroups to L sequences that previously exhibited a discrepancy between classification and location in global N-J and MP trees.....	424
5.62	Khoi-San sequences used to compile the Khoi-San consensus sequence.....	427
5.63	Distribution of alterations observed throughout the mitochondrial genome in the Khoi-San consensus sequence compared to the rCRS.....	430
5.64	Minimum and maximum percentage similarity and number of pairwise differences between Khoi-San sequences, and between the Khoi-San sequences and the Khoi-San consensus sequence.....	433
5.65	Differences between the Khoi-San consensus sequence and 43 Khoi-San sequences included in the current investigation.....	435
5.66	Alterations contained in the Khoi-San consensus sequence that are likely to be population-specific.....	436
5.67	Comparison between the Khoi-San consensus sequence and two human consensus sequences constructed in previous investigations.....	438
5.68	Number of nucleotide alterations observed in comparisons between the rCRS, the Khoi-San consensus sequence and two African reference sequences.....	440

LIST OF TABLES

5.69	Summary of alterations observed in 30 complete Khoi-San mitochondrial genome sequences.....	441
5.70	Novel alterations observed in 30 Khoi-San mitochondrial genome sequences.....	443
5.71	Exclusion criteria for L-specific sequences from which Khoi-San genome sequences and sequences obtained from GenBank® deviated.....	445
5.72	Assignment of haplogroups to mitochondrial genome sequences included in the current investigation.....	447
5.73	Location of 30 Khoi-San sequences generated in the current investigation in the phylogenetic trees.....	449
5.74	Differences observed in comparisons of phylogenetic trees constructed using different datasets and methods.....	450
5.75	Location of Khoi-San sequences in global L0-specific haplogroup network.....	453
5.76	Summary of the locations of new internal and terminal branches observed in the global L0-specific haplogroup network.....	453
5.77	Summary of information relating to recurrent alterations and reversions.....	455
5.78	Summary of information relating to different character states observed at two nucleotide positions.....	456
5.79	Geographic distribution of African mitochondrial haplogroups and sub-haplogroups.....	457
5.80	Summary of results for tests of neutrality, population expansion and selection.....	462
5.81	Summary of proposed explanations for deviation from neutrality.....	463
5.82	Estimated TMRCA for different haplogroups, sub-haplogroups and lineages within sub-haplogroups.....	465
5.83	Differences in the revised L0 classification scheme compared to the classification scheme used in the current investigation.....	468
5.84	Revised haplogroup classifications of incompletely classified L sequences and L sequences that exhibited a discrepancy between classification and location in global N-J and MP trees.....	469
5.85	Comparison of haplogroup and sub-haplogroup designations between three classification schemes and their alignment with the global L0-specific haplogroup network.....	470
A.1	Sequence alterations observed between the 30 complete Khoi-San mitochondrial genome sequences determined and the rCRS.....	539
B.1	Information regarding sequence alterations observed between the 30 complete Khoi-San mitochondrial genome sequences determined and the rCRS.....	555
D.1	A list of polymorphisms used to exclude L-specific sequences from other mtDNA haplogroups.....	567
E.1	Haplogroup analyses of the 30 complete Khoi-San mitochondrial DNA genome sequences.....	569
F.1	Comparison of haplogroup and sub-haplogroup designations between the classification scheme used in the current investigation and the classification scheme generally used in the literature.....	571
G.1	List of mitochondrial genome sequences used in all analyses.....	573

G.2	Probable country or geographic region of origin for ethnic groups for which no geographic information were provided.....	585
H.1	List of mitochondrial genome sequences excluded from (or included in) datasets compiled for use in various analyses.....	587
M.1	Synonymous codon usage for ATPase, CO and Cytb genes in 43 Khoi-San mtDNA genome sequences.....	616
M.2	Synonymous codon usage for ND genes in 43 Khoi-San mtDNA genome sequences.....	617
O.1	Matrix containing percentage similarity and pairwise number of differences between 43 Khoi-San sequences and the Khoi-San consensus sequence.....	630

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude and appreciation to the following individuals and institutions:

The **Khoi-San individuals** who made this investigation possible by generously donating their DNA. As one of the last remaining hunter-gatherer populations, their contribution was invaluable to the current investigation, but also to the research field as a whole. The information contained in their mtDNA genomes will assist in elucidating our maternal ancestry, and will increase our understanding of the mtDNA diversity present at the dawn of the human race.

My supervisor, **Prof. Antonel Olckers**, for providing me with the opportunity to complete a Ph.D. under her supervision, and to participate in a truly unique project. For teaching me the value of being organised, planning ahead, to think strategically, and for reminding me that if it is worth doing, it is worth doing well. I have learnt important lessons and gained valuable skills that will guide me in my scientific career wherever I may find myself in the future.

My co-supervisor, **Dr. Gordon Wayne Towers**, for always being available and willing to help. For teaching me about the management of people and scientific projects, for being a sounding board for ideas, for helpful suggestions, and for also being a friend who offered support when experiments did not progress as planned.

Prof. Doug Wallace (Center for Molecular and Mitochondrial Medicine and Genetics, University of California, Irvine), for providing access to laboratory facilities for the preparation of lymphoblastoid cell lines, as well as the use of computing facilities to perform phylogenetic analyses. **Mrs. Christa Swanepoel** (née Mouton), for culturing of the immortal lymphoblastoid cell lines, and isolating the genomic DNA of the Khoi-San individuals included in the current investigation. **Dr. Gordon Wayne Towers** and **Dr. Jason C. Poole** (Center for Molecular and Mitochondrial Medicine and Genetics, University of California, Irvine) for their assistance with the phylogenetic analyses performed for datasets 1a and 2a.

The **North-West University (Potchefstroom Campus)**, the **Centre for Genome Research** and **DNAbiotec (Pty.) Ltd.**, for providing financial assistance and excellent research facilities. To **Mrs. Monique van Deventer** at Academic Administration (Potchefstroom Campus), for her assistance with a daunting intercontinental registration and thesis submission process. To the staff at the **Ferdinand Postma Library**, for ensuring that I had access to the journal articles I needed.

Past and present students of the Centre for Genome Research: **Mr. Jake Darby**, **Mr. Dan Isabirye**, **Dr. Desiré Dalton**, **Dr. Tumi Semete**, **Dr. Marco Alessandrini**, **Dr. Estie Kotze** and **Ms. Delia Tanner** for their assistance, camaraderie, and friendly encouragement. We shared an experience that will forever bind us together.

Past and present employees of DNAbiotec (Pty.) Ltd.: **Ms. Desiré Hart**, **Ms. Chricentia Khumalo**, **Ms. Tshireletso Mataboge**, **Mr. Kenneth Nkadimeng**, **Mr. Leonard Mdluli**, **Ms. Anri Raath** and **Mrs. Scheán Babst**, for their assistance, support and friendship through the course of my investigation. **Dr. Annelize van der Merwe**, for her support, expertise, and for ensuring that the day to day laboratory activities ran smoothly. **Ms. Martha Sebogoli**, for maintaining a clean and neat working environment.

My colleagues at the **Kolling Institute for Medical Research, Perinatal Research Group**, for their continued interest and encouragement. **Dr. Julia Warning**, for proofreading of the thesis and her valuable suggestions.

To my friends, old and new, in South Africa and Australia, for providing support and encouragement, and for understanding why I often did not have time to spend with them, or attend the events they organised. I know that the friendships still remaining are the true friendships that will last a lifetime. A special mention and thank you to the members of **Generation X** (2004 to 2005), **Hartklop** (2005 to 2008) and **Growing Together** (2009) for their support.

To my family: my parents, my father **Willie Freeman** and my mother **Riekie Freeman**. Thank you for your love and encouragement. Thank you for providing me with a quiet place and a comfortable environment, which made it possible for me to focus only on my writing. My brother, **André Freeman**, his wife **Lehandrie** and their son, **Luke**. Thank you for your interest and encouragement. Luke's arrival coincided with a new phase in my life, and as he grows I am reminded of all the possibilities of a new beginning. The Koekemoer

family, **Theuns Koekemoer**, **Hannetjie Koekemoer**, **Robert Koekemoer** and **Reinhardt Koekemoer**. Thank you for your interest, encouragement and support.

My husband, **Theunis Koekemoer**. Thank you for believing in me and for being a willing partner in this journey. It has been a long, hard road, full of unexpected twists and turns, but you stood by me through it all. Thank you for all the many sacrifices made, and for putting your life on hold so that I could pursue my dream of obtaining a Ph.D. Thank you for your love, understanding, unique methods of encouragement, incredible patience, unfailing support and the firm belief, often contrary to mine, that the thesis would one day be finished. I am grateful for you, and I love you very much.

CHAPTER ONE

Introduction

Mitochondria are organelles present in the eukaryotic cell, that contain their own genetic material in the form of mitochondrial DNA (mtDNA), and which are derived from ancient bacteria that were acquired through a series of endosymbiotic events (Gillham, 1994). Human mitochondrial DNA is a closed circular molecule consisting of 16,569 base pairs (Anderson *et al.*, 1981). It encodes small (12S) and large (16S) ribosomal RNA (rRNA) molecules, 22 transfer RNA (tRNA) molecules and 13 polypeptides. The polypeptides function as subunits in five multisubunit enzyme complexes that are positioned within the mitochondrial inner membrane and are responsible for the generation of energy through the process of oxidative phosphorylation (Schon, 1993). Mitochondria are strictly maternally inherited (Giles *et al.*, 1980), are present at high copy numbers in human cells (Satoh and Kuroiwa, 1991) and are subject to a nucleotide substitution rate several orders of magnitude higher than that detected for the nuclear genome (Brown *et al.*, 1979). Due to these characteristics, human mitochondrial DNA genomes have, since last sharing a common ancestor ca. 150,000 to 200,000 years before present, sequentially accumulated alterations on diverging maternal lineages (Ruiz-Pesini *et al.*, 2007).

Analysis of the mtDNA variation observed in mitochondrial lineages present in indigenous populations in diverse geographic regions globally, revealed that human mtDNA variation is extensive, as well as adaptive (Wallace, 2007). The genetic past of human mitochondrial DNA genomes can be represented in the form of a sequential mutational tree that diverges from a shared ancestor (Ruiz-Pesini *et al.*, 2007). This human mtDNA tree consists of distinct branches with every branch composed of a set of related mitochondrial DNA sequences or haplotypes, collectively referred to as a haplogroup. Further investigation of the human mtDNA tree revealed that the largest mtDNA variation, as well as the root of the tree, was located in Africa, and that there was a correlation between haplogroups and the geographic origin of native human populations (Wallace *et al.*, 1999). Based on the human mtDNA tree, the most diverse and therefore the most ancient mtDNA sequences are observed in Africa (Wallace, 2007).

As a rule, each mitochondrial haplogroup is established by one or more functionally important variants. Several of these functionally important variants have however arisen more than once in diverse populations. This indicates a role for convergent evolution and adaptive selection in human mtDNA variation. Thus, although the correlation between haplogroup and geographic origin was initially ascribed to random genetic drift, substantial proof has subsequently been presented that supports the conclusion that the observed pattern of mtDNA variation is the consequence of natural selection (Wallace *et al.*, 1999; Mishmar *et al.*, 2003; Ruiz-Pesini *et al.*, 2004; Ruiz-Pesini and Wallace 2006). The adaptation of mtDNA genes primarily involved with the regulation of energy production, together with the adaptation of the nuclear DNA genes mainly involved in structural roles, afford humans with a multidimensional capability to adapt to new environments (Wallace, 2008).

Although extensive adaptive mtDNA variation has been generated through the increased mtDNA mutation rate, it remains much more likely that random mutations will be deleterious than advantageous. This is confirmed by the high incidence of mitochondrial diseases which is currently estimated at 1.65/10,000 (Schaefer *et al.*, 2004; Schaefer *et al.*, 2008). In addition, numerous associations between mtDNA haplogroups and a variety of clinical conditions have been identified, for which additional support is currently being gathered through evidence obtained of physiological differences between haplogroups (Ruiz-Pesini *et al.*, 1998; Niemi and Majamaa, 2005; Montiel-Sosa *et al.*, 2006). It is therefore evident that the mtDNA variation observed in human populations affects human health (Wallace, 2008).

In recent years, as molecular techniques improved, the study of human mitochondrial DNA variation has become an invaluable tool in the investigation of our genetic past as revealed through evolutionary studies of human origins and population history, as well as in the study of mitochondrial diseases. Mitochondrial DNA sequence data is used in a number of different ways to elucidate human evolution and population history. Firstly, it is used to interpret socio-cultural effects that could have played a role in human evolution such as polygyny (Salem *et al.*, 1996; Dupanloup *et al.*, 2003; Kayser *et al.*, 2003; Wilder *et al.*, 2004b), the effects of matrilocality vs. patrilocality (Oota *et al.*, 2001; Wilder *et al.*, 2004b) or the social stratification brought about by the caste system (Bamshad *et al.*, 1998; Wooding *et al.*, 2004). Secondly, the high copy number of the mtDNA allows for the use of mtDNA in studies of ancient DNA (Stone and Stoneking, 1993; Stone and Stoneking, 1998; Keyser-Tracqui *et al.*, 2003; Vernesi *et al.*, 2004) and in some forensic applications,

if maternal relatives are available for comparison, such as the identification of otherwise unidentifiable victims (Gill *et al.*, 1994; Stone *et al.*, 2001) or victims of war or terror (Holland *et al.*, 1993; Ladika, 2001; Lawler, 2001), as well as in the fight against injustice as a result of the wrongful imprisonment of individuals (Innocence Project, established in 1992). Thirdly, mtDNA is increasingly used in determining individual genealogies (Shriver and Kittles, 2004). This application is however often problematic as the average client has the expectation of obtaining information regarding the client's specific place of origin, while the use of mtDNA haplogroups generally do not permit the identification of specific localities due to a lack of geographic specificity. Furthermore, clients fail to distinguish between the origin of their maternal lineage and that of their complete biological ancestry, which could have a profound effect on their self-image and how they perceive their own identity (Pakendorf and Stoneking, 2005).

One of the biggest questions that remain in human history revolves around human origins and evolution, i.e. where do modern humans come from? Where did *Homo sapiens sapiens* originate and how did they succeed in populating the whole world? With the data provided by studies of human mtDNA variation, complemented by Y chromosome and nuclear DNA data, it has become possible to assemble the pieces of the puzzle provided by data from other disciplines such as linguistics, archaeology, palaeo-climatology and anthropology in a way that tells the story of our origins.

At no time in human history has there been more public interest in our past and genetic ancestry. This is demonstrated by the increase in web-based genealogy research (Ancestry.com, <http://www.ancestry.com>; MyHeritage.com, <http://www.myheritage.com>), the emergence of private companies providing individualised genotyping services (Oxford Ancestors Ltd., <http://www.oxfordancestors.com>; Family Tree DNA - Genealogy by Genetics Ltd., <http://www.familytreedna.com>), articles in leading magazines (Newsweek Magazine, Scientists Explore a Controversial Theory About Man's Origins, 11 January 1988), books (Bryan Sykes, *The Seven Daughters of Eve*, 2001; Stephen Oppenheimer, *Out of Africa's Eden: The Peopling of the World*, 2003), documentaries (Discovery Channel, *The Real Eve: Modern Man's Journey Out of Africa*, 2002) and public participation in worldwide population studies (Human Genome Diversity Project, initiated in 1993; International HapMap Project, initiated in 2002; The Genographic Project, initiated in 2003). Today, communities are acutely aware that the past retains clues not only of human origins but also of the future of the human race. Contemporary human genetic profiles,

shaped by the course of genetic history, are the focus of numerous studies regarding health, specifically risk for disease, both congenital and acquired through lifestyle choices.

It would be a bold statement, although not inaccurate, to say that our genetic past holds the key to future health. For this reason it is of the utmost importance to study the most ancient populations accessible in order to compare the genetic past with the genetic present. Hunter-gatherers, such as the Southern African Khoi-San, are carriers of some of the oldest mtDNA lineages observed in extant human populations (Vigilant *et al.*, 1991; Chen *et al.*, 2000; Salas *et al.*, 2002; Knight *et al.*, 2003; Gonder *et al.*, 2007; Tishkoff *et al.*, 2007).

The broad aim of the mitochondrial research programme is to investigate the role of the mitochondrial genome in human migration and health. In the current investigation, the mtDNA genomes of 30 Khoi-San individuals were studied and a mitochondrial consensus sequence was constructed for the Southern African Khoi-San. In Chapter Four, the Khoi-San cohort, as well as the methods used to achieve the specific aims listed in Chapter Three, are described. The results obtained from the current investigation are presented in Chapter Five. The conclusions drawn from the results presented in the current investigation are described in Chapter Six, as well as a model proposed to describe the genetic variation observed in human mtDNA genomes.

The Khoi-San consensus sequence compiled in the current investigation does not only provide insight into human origins and evolution, but it also provides a baseline with which to compare and evaluate identified mitochondrial alterations in diseased individuals as either candidates for disease-causing mutations or harmless ancient polymorphisms. In Africa, researchers are faced with the unique problem that the indigenous African populations contain high amounts of variation within both their nuclear and their mtDNA genomes, and it is therefore an enormous challenge to identify the causative mutation in a background containing such high levels of polymorphism. Adding to that, the only reference sequence currently available is of European origin (Anderson *et al.*, 1981; Andrews *et al.*, 1999) and contains less variation than the average African mtDNA sequence. It is anticipated that the Khoi-San mtDNA consensus sequence determined in the current investigation will provide researchers with a tool with which to determine if an observed alteration is due to normal variation or represents a possible causative mutation. Since the mitochondrion generates the bulk of the cellular energy, it is imperative that an

understanding is gained of mitochondrial biology and genetics in order to fully understand human health and disease (Wallace, 2007).

CHAPTER TWO

Biochemical and Genetic Aspects of the Mitochondrion

Mitochondria are organelles present in the eukaryotic cell that play a central role in energy metabolism. Mitochondria are uniquely adapted to fulfil this role by being the centre where oxidative phosphorylation takes place. The genome of mitochondria has a unique organisation and has evolved to be compact and to selectively encode the protein subunits that form part of the electron transport chain.

2.1 ORIGIN OF THE MITOCHONDRION

The origin of the mitochondrion within eukaryotic cells can be explained by two competing hypotheses. The mitochondrion could have arisen autogenously in the process of cellular differentiation during the course of evolution, or the mitochondrion could have been acquired by the protoeukaryotic cell through a series of endosymbiotic events (Gillham, 1994). The endosymbiont theory was first posited in 1922 when Wallin proposed that mitochondria were derived from cyanobacterial endosymbionts. The hypothesis was disregarded for many years until Margulis revived it in 1970 when she proposed that the organelles observed in eukaryotic cells had been acquired through the ingestion of prokaryotic cells (Margulis, 1970).

Evidence for this endosymbiosis theory of origin was based on several observations that indicated similarities to prokaryotes. Mitochondria are of comparable size to many prokaryotes. Their genomes are circular and of similar size and complexity as the genomes observed in prokaryotes. Patterns of antibiotic resistance are similar to that of prokaryotes, and they contain a double phospholipid bilayer. Ribosomal RNA sequences were used to evaluate the phylogenetic ancestry of the mitochondrion and another organelle referred to as a plastid. It was determined that mitochondria and plastids have distinctly different phylogenetic ancestries and could not have originated autogenously within the same host cell. A sequential origin, in which mitochondria and then plastids (in that order) were acquired, was proposed. Currently, the generally accepted view of the origin of the mitochondrion is therefore through endosymbiotic events and not

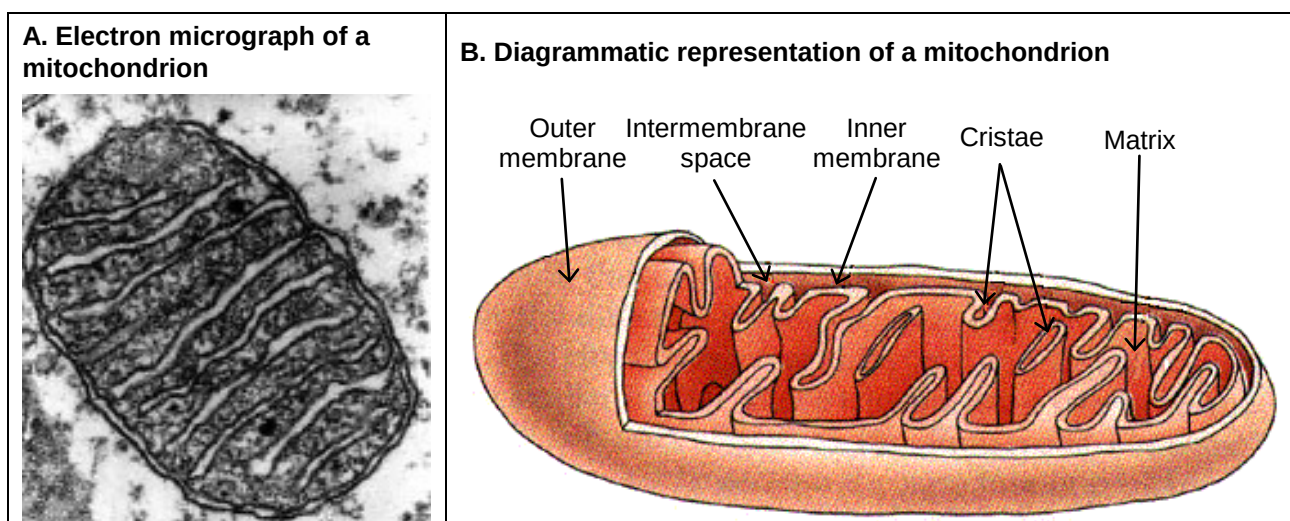
autogenously acquired via cellular differentiation through the course of evolution (Searcy, 2003).

Other features that mitochondria share with bacteria include the fact that mitochondria reproduce by cell division, that mitochondria are separate entities within a cell that do not fuse with other membranes within the cellular cytoplasm, and that mitochondria possess their own genetic material in the form of deoxyribonucleic acid (DNA), as well as protein machinery which is similar to that observed in bacteria and distinct from that observed in the nucleocytoplasm. The exact nature of the initial symbiotic metabolic relationship leading to the origin of the mitochondrion in the eukaryotic cell is however still unknown (Searcy, 2003).

2.2 STRUCTURE AND FUNCTION OF THE MITOCHONDRION

Mitochondria vary in size and shape, depending on their source and metabolic state. They are typically ellipsoid shaped organelles of ca. 0.5 micrometre (μm) in diameter and 1.0 μm in length with a molecular weight of 10^7 dalton (Giles *et al.*, 1980). An electron micrograph and a diagrammatic representation of a mitochondrion are presented in Figure 2.1. The mitochondrion is the primary location for metabolic functions such as the tricarboxylic acid cycle, fatty acid oxidation, oxidative phosphorylation and electron transport (Voet and Voet, 1995).

Figure 2.1 The structure of the mitochondrion



Electron micrograph adapted from D.S. Friend, Brigham and Women's Hospital (<http://publications.nigms.nih.gov/insidethecell/chapter1.html>, 2005). Diagrammatic representation adapted from a figure presented on the website of the Brooklyn College, The City University of New York, (<http://academic.brooklyn.cuny.edu/biology/bio4fv/page/mito.htm>, 2004).

The mitochondrion is enclosed by a smooth outer membrane and contains an extensively invaginated inner membrane (Borst, 1977). The invaginations in the inner membrane form structures referred to as cristae, which increase the inner mitochondrial membrane surface area and encloses the inner mitochondrial compartment, which consists of a gel-like substance referred to as the mitochondrial matrix. The mitochondrial matrix contains high concentrations of soluble enzymes that perform a function in oxidative metabolism, as well as the mitochondrial protein-synthesising system which consists of proteins that play a role in DNA replication and RNA translation. The mitochondrial matrix provides an environment that allows for the interaction of these enzymes, or other proteins, with the necessary substrates, nucleotide co-factors and inorganic ions contained therein (Voet and Voet, 1995).

The outer and inner membranes exhibit differences in membrane density due to the presence of different total amounts of lipids and proteins. The outer mitochondrial membrane (OMM) has a smooth appearance without obvious folds or protuberances and has a density of 1.13 gram per cubic centimetre (g.cm^{-3}). It consists of 60 to 70 percent (%) protein, 30 to 40% lipids (mainly phosphatidylinositol and cholesterol) and contains porin, a protein that forms non-specific pores that permits free diffusion of up to 10 kilodalton (kDa) size molecules (Gillham, 1994; Voet and Voet, 1995). The inner mitochondrial membrane (IMM) has a density of 1.21 g.cm^{-3} and is considerably more protein-rich (ca. 80%) with the lipid component (ca. 20%) consisting mainly of cardiolipin (Gillham, 1994; Lodish *et al.*, 1995). The inner membrane is freely permeable to carbon dioxide (CO_2), oxygen (O_2), and water (H_2O) exclusively. Comparison of the molecular weights of proteins present within the OMM and IMM by sodium dodecyl sulphate (SDS) gel electrophoresis, revealed little correspondence between the protein complement of the two membranes (Gillham, 1994).

The inner mitochondrial membrane is a highly specialised structure that contains the proteins that form part of the electron transport chain, as well as transport proteins that direct the movement of metabolites such as adenosine triphosphate (ATP), adenosine diphosphate (ADP), pyruvate, calcium (II) ions (Ca^{2+}) and phosphate (Voet and Voet, 1995). This controlled impermeability of the IMM to the majority of metabolites, ions and low molecular weight compounds, allows for the creation of ionic gradients across the membrane, which leads to the compartmentalisation of metabolic functions between the cytosol and mitochondria, as well as the formation of ATP (Voet and Voet, 1995).

2.3 THE MITOCHONDRIAL ELECTRON TRANSPORT CHAIN

The mitochondrial electron transport chain (mtETC) is localised in the inner mitochondrial membrane and the mitochondrial matrix. The free energy of electron transfer from the oxidised electron carriers nicotinamide adenine dinucleotide (NAD⁺) and flavin adenine dinucleotide (FAD) to their reduced forms (NADH and FADH₂) and subsequently to O₂, the final acceptor of electrons, is coupled with ATP synthesis (Lodish *et al.*, 1995). This process occurs via five multi-subunit protein complexes, indicated in Table 2.1, which are localised in the inner mitochondrial membrane (Schon, 1993).

Table 2.1 The multi-subunit complexes which play a role in the electron transport chain

No.	Complex	Mass (kDa)	Subunits	Prosthetic group	Binding site for
	Name				
I	NADH-ubiquinone oxidoreductase	980	45	FMN Fe-S	NADH (matrix side)
II	Succinate-ubiquinone oxidoreductase	140	4	FAD Fe-S	Succinate (matrix side) UQ (lipid core)
III	Ubiquinol-cytochrome c oxidoreductase	248	11	Haeme <i>b_L</i> Haeme <i>b_H</i> Haeme <i>c₁</i> Fe-S	Cytochrome <i>c</i> (intermembrane space side)
IV	Cytochrome c oxidase	162	13	Haeme <i>a</i> Haeme <i>a₃</i> Cu _A Cu _B	Cytochrome <i>c</i> (intermembrane space side)
V	F ₀ F ₁ -ATP synthase	~530	16	---	ADP and P _i (matrix side of the inner mitochondrial membrane)

No. = number; kDa = kilodalton; NADH = nicotinamide adenine dinucleotide; FMN = flavin mononucleotide; Fe-S = iron-sulphur protein; FAD = flavin adenine dinucleotide; UQ = coenzyme Q or ubiquinone; Haeme *b_L* = low potential *b*-type haeme; Haeme *b_H* = high potential *b*-type haeme; Haeme *c₁* = *c*-type haeme; Haeme *a* and Haeme *a₃* = *a*-type haeme; Cu_A = copper atom; Cu_B = copper atom; F₀ = component of ATP synthase, oligomeric enzyme complex located in integral membrane; F₁ = component of ATP synthase, water-soluble oligomeric enzyme complex bound to F₀; ATP = adenosine triphosphate; ADP = adenosine diphosphate; P_i = inorganic phosphate. Adapted from Wallace (1994), Wallace (1995), Lodish *et al.* (1995), Carroll *et al.* (2002) and Senior *et al.* (2002).

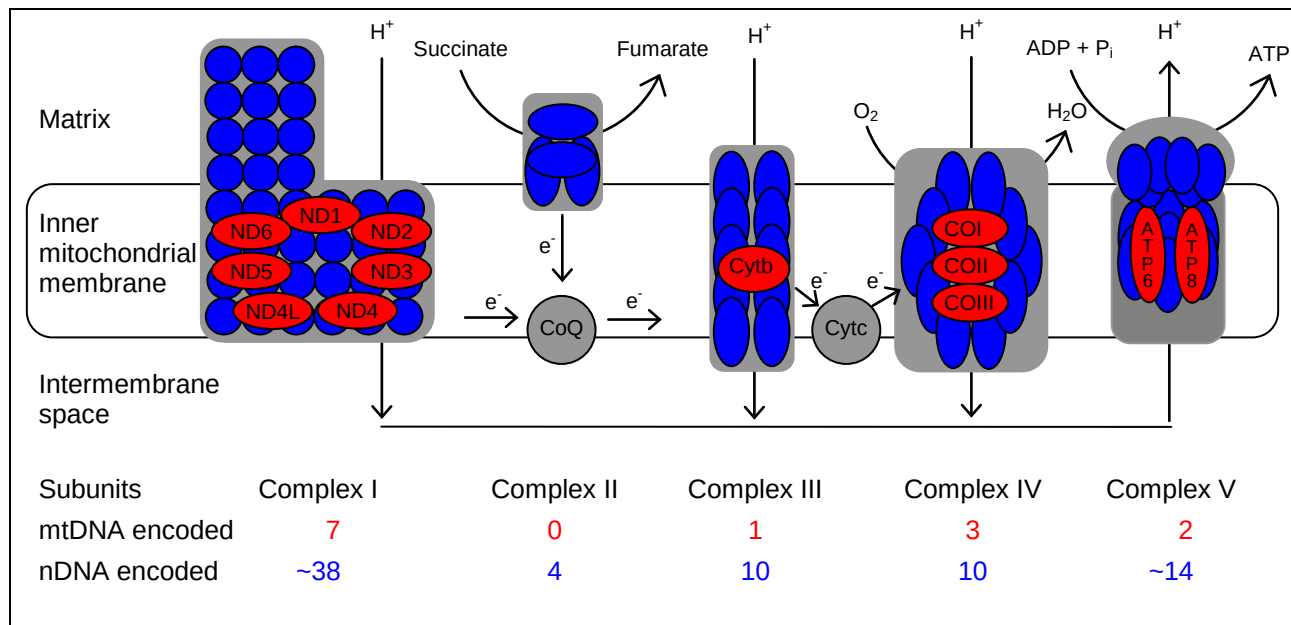
The mtETC consists of proteins encoded by both the nuclear and mitochondrial genomes, as well as a number of different prosthetic groups (refer to Table 2.1). The reason for this segregation of genes between the two genomes is unknown, it has however been observed that the mitochondrial gene products tend to be the most hydrophobic proteins in each individual complex (Gillham, 1994). The subunits of the electron transport system encoded by mitochondrial genes are indicated in Table 2.2.

Table 2.2 Subunits of the electron transport system encoded by mitochondrial genes

Complex number	Complex name	Mitochondrial encoded subunits	Number of subunits encoded by nDNA
I	NADH-ubiquinone oxidoreductase	ND1, ND2, ND3, ND4, ND4L, ND5, ND6	38
II	Succinate-ubiquinone oxidoreductase	None	4
III	Ubiquinol-cytochrome <i>c</i> oxidoreductase	Cytb	10
IV	Cytochrome <i>c</i> oxidase	COI, COII, COIII	10
V	F ₀ F ₁ -ATP synthase	ATPase6, ATPase8	14

nDNA = nuclear DNA; NADH = nicotinamide adenine dinucleotide; ND1 to ND4, ND4L, ND5 and ND6 = NADH dehydrogenase subunits; COI, COII and COIII = cytochrome *c* oxidase subunits; Cytb = cytochrome *b*; F₀ = component of ATP synthase, oligomeric enzyme complex located in integral membrane; F₁ = component of ATP synthase, water-soluble oligomeric enzyme complex bound to F₀; ATP = adenosine triphosphate; ATPase6 and ATPase8 = adenosine triphosphatase subunits. Adapted from Schon (1993), Wallace (1994), Carroll *et al.* (2002) and Senior *et al.* (2002).

Complex I transfers electrons from NADH to coenzyme Q (CoQ) and is the entry point for electrons travelling from NADH into the mitochondrial electron transport chain. NADH is oxidised to NAD⁺, releasing two electrons which are carried by FMN and Fe-S, which in turn reduces CoQ to CoQH₂. A total of four protons are transported across the IMM per molecule of NADH oxidised. Complex II transfers electrons from succinate to CoQ. Succinate is oxidised to fumarate, releasing two electrons which are transferred first to FAD, then to Fe-S and finally to CoQ to form reduced CoQH₂. No protons are translocated across the IMM by complex II. Complex III transports electrons from reduced CoQ to cytochrome *c* which regenerates oxidised CoQ. The two released electrons are transferred to Fe-S, to the two *b*-type cytochromes, and then to cytochrome *c*₁. The two electrons are finally transferred to oxidised cytochrome *c* to produce reduced cytochrome *c*. A total of four protons are transported from the matrix across the IMM per pair of electrons transported. Complex IV catalyses the oxidation of reduced cytochrome *c* and the reduction of O₂. Within this complex, the electrons are transferred one at a time to a copper ion, to cytochrome *a*, to a complex formed by a second copper ion and cytochrome *a*₃ and are finally accepted by O₂ to produce H₂O. Two protons are transported through the IMM per pair of electrons transported. Complex V harnesses the energy accumulated in the electrochemical proton gradient created across the IMM by coupling the movement of protons from the intermembrane space to the matrix with the production of ATP (Gillham, 1994; Voet and Voet, 1995; Lodish *et al.*, 1995). The electron transport chain and the above mentioned processes are indicated in Figure 2.2.

Figure 2.2 Schematic representation of the mitochondrial electron transport chain

mtDNA encoded subunits are indicated in red and nDNA encoded subunits are indicated in blue. Biochemical reactions are not indicated in full, i.e. electron transfer is indicated, although not the exact number of electrons transferred. e^- = electrons, H^+ = protons; nDNA = nuclear DNA; mtDNA = mitochondrial DNA; ND1 to ND4, ND4L, ND5 and ND6 = NADH dehydrogenase subunits; CoQ = coenzyme Q or ubiquinone; Cytb = cytochrome b; Cytc = cytochrome c; O_2 = oxygen; H_2O = water; ADP = adenosine diphosphate; P_i = inorganic phosphate; COI, COII and COIII = cytochrome c oxidase subunits; ATP = adenosine triphosphate; ATP6 and ATP8 = adenosine triphosphatase subunits. Adapted from Lodish *et al.* (1995) and DiMauro and Davidzon (2005).

2.4 THE MITOCHONDRIAL GENOME

The mitochondrion contains its own genome consisting of mitochondrial DNA. This genome has a number of unique features pertaining to its mode of inheritance, the genetic organisation of its genome, and characteristics of its replication, transcription and translation processes. These features differ from that observed for the nuclear genome and will be discussed in the sections that follow.

2.4.1 Inheritance of the mitochondrial genome

Mitochondria are maternally inherited (Giles *et al.*, 1980) and this mode of inheritance implies that there is either no paternal input or that there are ways of preventing the paternal contribution of mitochondria. There are various ways in which the paternal transmission of mitochondria can be prevented (Birky, 1995). In mammals as many as 100 paternal mitochondria enter the egg as part of the sperm cell, although they are destroyed within hours after fertilisation has occurred (Manfredi *et al.*, 1997; Shitara *et al.*, 1998). After entering the egg cytoplasm, the paternal mitochondria are selectively destroyed due to the presence of a universal proteolytic marker, namely ubiquitin (Sutovsky *et al.*, 2000) and detection of paternal mtDNA is dependent on phylogenetic relatedness (Kaneda *et al.*, 1995). Although these preventative mechanisms are retained

to ensure that only maternal mitochondria are present within the zygote, the maternal inheritance of mtDNA is frequently a quantitative phenomenon, as the ratio of sperm to egg mitochondria in a single cell zygote is ca. 1:103 (Birky, 2001).

In some instances the recognition mechanism of paternal mitochondria can however fail, leading to the transmission of paternal mitochondria, a phenomenon referred to as paternal leakage. An instance of paternal inheritance of mtDNA in humans has recently been identified (Schwartz and Vissing, 2002). This report described a man suffering from severe exercise intolerance and it was subsequently determined that his muscular mtDNA was primarily of paternal origin. This case created uncertainty about the generally accepted assumption of strict maternal mtDNA inheritance. Incidences of paternal leakage such as this reported case are however rare. In instances where paternal leakage is observed, it is usually due to failure of the normal detection and eradication processes of paternal mitochondria, and such occurrences are identified due to the fact that the individual is affected by a disorder that can be linked to the presence of the paternal mitochondria. In addition, numerous comparisons between maternal parent and offspring have been unsuccessful in providing evidence in support of paternal inheritance (Giles *et al.*, 1980; Howell *et al.*, 2003). Currently, the maternal inheritance of mtDNA in humans can therefore still be considered as the rule (Schwartz and Vissing, 2003).

2.4.2 Genetic organisation of the mitochondrial genome

Mitochondria contain a closed-circular, double-stranded DNA genome referred to as mitochondrial DNA, and each mitochondrion contains ca. five to 15 copies of mtDNA (Sato and Kuroiwa, 1991) which are organised into a structure referred to as a nucleoid (Nass *et al.*, 1965). A complete human mitochondrial genome sequence was first determined by Anderson *et al.* in 1981 and the complete mitochondrial genome has a length of 16,569 base pairs (bp). The locations of various functional entities observed in the mitochondrial genome are listed in Table 2.3.

Table 2.3 Functional elements located in the mitochondrial DNA genome

Locus	Map position	Abbreviation	Description
MT-HV2	57 - 372	HVS2	Hypervariable segment 2
MT-OHR	110 - 441	O _H	H-strand origin
MT-CSB1	213 - 235	CSB1	Conserved sequence block 1
MT-TFX	233 - 260	TFX	mtTF1 binding site
MT-TFY	276 - 303	TFY	mtTF1 binding site
MT-CSB2	299 - 315	CSB2	Conserved sequence block 2
MT-HPR	317 - 321	---	Replication primer
MT-CSB3	346 - 363	CSB3	Conserved sequence block 3
MT-4H	371 - 379	mt4H	mt4 H-strand control element
MT-3H	384 - 391	mt3H	mt3 H-strand control element
MT-LSP	392 - 445	P _L	L-strand promoter
MT-TFL	418 - 445	---	mtTF1 binding site
MT-TFH	523 - 550	---	mtTF1 binding site
MT-HSP1	545 - 567	P _{H1}	Major H-strand promoter
MT-TF	577 - 647	F	tRNA phenylalanine
MT-HSP2	645 - 645	P _{H2}	Minor H-strand promoter
MT-RNR1	648 - 1,601	12S	12S ribosomal RNA
MT-TV	1,602 - 1,670	V	tRNA valine
MT-RNR2	1,671 - 3,228	16S	16S ribosomal RNA
MT-RNR3	3,206 - 3,228	---	5S-like sequence
MT-TER	3,229 - 3,256	---	Transcription terminator
MT-TL1	3,230 - 3,304	L ^(UUR)	tRNA leucine recognising the UUR codon
MT-NC1	3,305 - 3,306	NC1	Non-coding nucleotide
MT-ND1	3,307 - 4,262	ND1	NADH dehydrogenase subunit 1
MT-TI	4,263 - 4,331	I	tRNA isoleucine
MT-TQ	4,329 - 4,400	Q	tRNA glutamine
MT-NC2	4,401 - 4,401	NC2	Non-coding nucleotide
MT-TM	4,402 - 4,469	M	tRNA methionine
MT-ND2	4,470 - 5,511	ND2	NADH dehydrogenase subunit 2
MT-TW	5,512 - 5,579	W	tRNA tryptophan
MT-NC3	5,580 - 5,586	NC3	Non-coding nucleotide
MT-TA	5,587 - 5,655	A	tRNA alanine
MT-NC4	5,656 - 5,656	NC4	Non-coding nucleotide
MT-TN	5,657 - 5,729	N	tRNA asparagine
MT-OLR	5,721 - 5,798	O _L	L-strand origin
MT-TC	5,761 - 5,826	C	tRNA cysteine
MT-TY	5,826 - 5,891	Y	tRNA tyrosine
MT-NC5	5,892 - 5,903	NC5	Non-coding nucleotide
MT-CO1	5,904 - 7,445	COI	Cytochrome c oxidase subunit I
MT-TS1	7,446 - 7,516	S ^(UCN)	tRNA serine recognising the UCN codon
MT-NC6	7,517 - 7,517	NC6	Non-coding nucleotide
MT-TD	7,518 - 7,585	D	tRNA aspartic acid
MT-CO2	7,586 - 8,269	COII	Cytochrome c oxidase subunit II
MT-NC7	8,270 - 8,294	NC7	Non-coding nucleotide

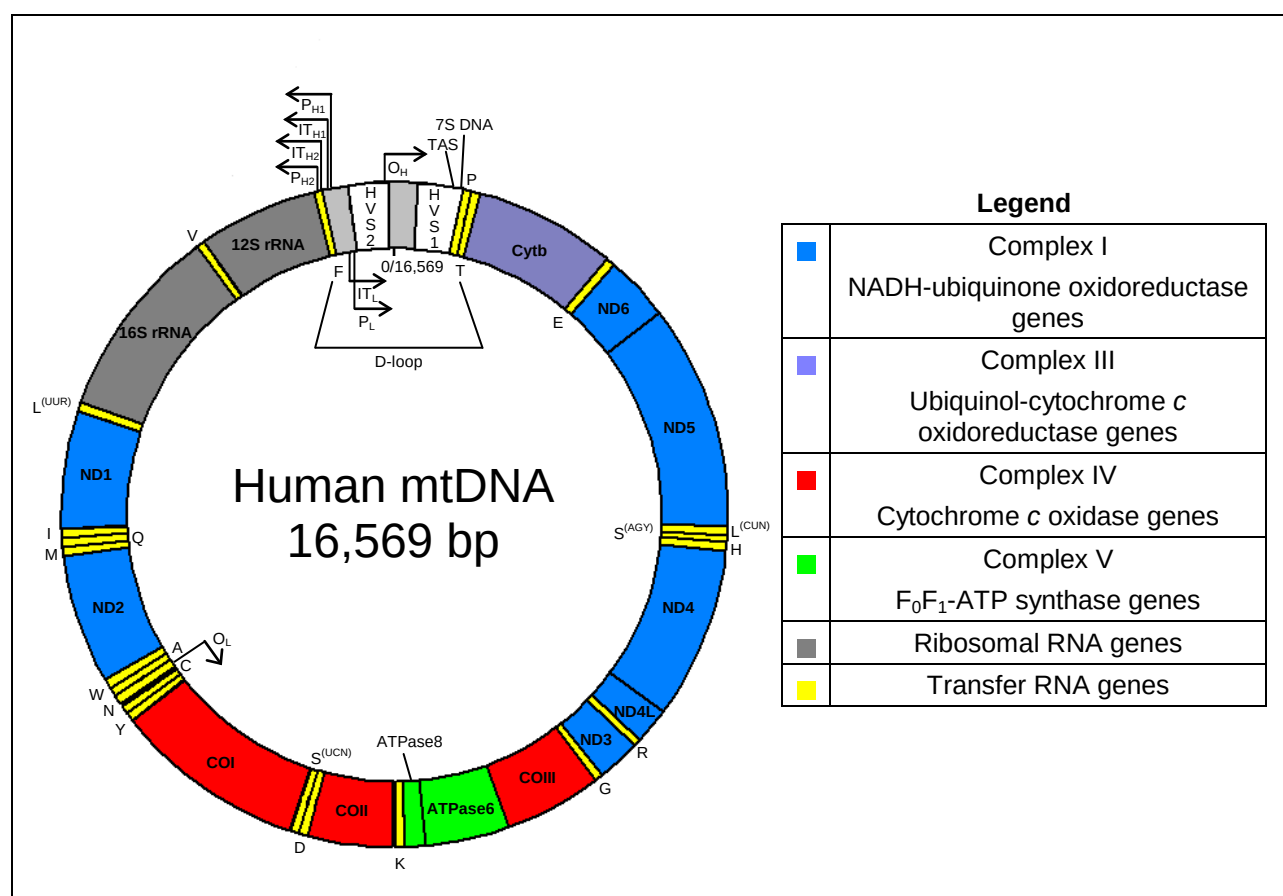
Table 2.3 Continued...

Locus	Map position	Abbreviation	Description
MT-TK	8,295 - 8,364	K	tRNA lysine
MT-NC8	8,365 - 8,365	NC8	Non-coding nucleotide
MT-ATP8	8,366 - 8,572	ATPase8	ATP synthase F ₀ subunit 8
MT-ATP6	8,527 - 9,207	ATPase6	ATP synthase F ₀ subunit 6
MT-CO3	9,207 - 9,990	COIII	Cytochrome c oxidase subunit III
MT-TG	9,991 - 10,058	G	tRNA glycine
MT-ND3	10,059 - 10,404	ND3	NADH dehydrogenase subunit 3
MT-TR	10,405 - 10,469	R	tRNA arginine
MT-ND4L	10,470 - 10,766	ND4L	NADH dehydrogenase subunit 4L
MT-ND4	10,760 - 12,137	ND4	NADH dehydrogenase subunit 4
MT-TH	12,138 - 12,206	H	tRNA histidine
MT-TS2	12,207 - 12,265	S ^(AGY)	tRNA serine recognising the AGY codon
MT-TL2	12,266 - 12,336	L ^(CUN)	tRNA leucine recognising the CUN codon
MT-ND5	12,337 - 14,148	ND5	NADH dehydrogenase subunit 5
MT-ND6	14,149 - 14,673	ND6	NADH dehydrogenase subunit 6
MT-TE	14,674 - 14,742	E	tRNA glutamic acid
MT-NC9	14,743 - 14,746	NC9	Non-coding nucleotide
MT-CYB	14,747 - 15,887	Cytb	Cytochrome <i>b</i>
MT-TT	15,888 - 15,953	T	tRNA threonine
MT-ATT	15,925 - 499	ATT	Membrane attachment site
MT-NC10	15,954 - 15,954	NC10	Non-coding nucleotide
MT-TP	15,955 - 16,023	P	tRNA proline
MT-DLOOP	16,024 - 576	CR/D-loop	Control Region, including displacement loop
MT-HV1	16,024 - 16,383	HVS1	Hypervariable segment 1
MT-7SDNA	16,106 - 191	7S DNA	7S DNA
MT-TAS	16,157 - 16,172	TAS	Termination-associated sequence
MT-5	16,194 - 16,208	Mt5	Control element
MT-3L	16,499 - 16,506	mt3L	L-strand control element

Locus names are the official designations delineated by the given nucleotide numbers. The map positions correspond to the nucleotide pair numbers determined from the DNA sequence. D-loop refers to the non-coding region between proline and phenylalanine (16,024 to 576). H-strand origin of replication positions have been identified at nucleotide positions 110, 147, 169, 191, 219, 310 and 441, L-strand promoter positions have been identified at nucleotide positions 407 and 392 to 435, H-strand promoter positions have been identified at nucleotide positions 559 to 561 and L-strand origin of replication positions have been identified at nucleotide positions 5,721 to 5,781, 5,761 and 5,799. mtTF1 = mitochondrial transcription factor; A = adenine; C = cytosine; G = guanine; U = uracil; R = A or G; Y = C or U; N = A, C, G or U. Adapted from MITOMAP (2007).

The mitochondrial genome encodes the 12S (Svedberg units) and 16S ribosomal RNAs (rRNAs) and 22 transfer RNAs (tRNAs) that are involved in protein synthesis, as well as 13 polypeptides that are components of the mitochondrial respiratory chain or oxidative phosphorylation (OXPHOS) pathways (Anderson *et al.*, 1981). It has been determined that the mtDNA molecule consists of a light (L) and heavy (H) strand which are distinguished based on their nucleotide composition. The H-strand is purine rich (mostly G and A nucleotides), while the L-strand is pyrimidine rich (mainly C and T nucleotides), the different nucleotide compositions of the two strands result in different buoyant densities in

Figure 2.3 Diagrammatic representation of the mitochondrial genome



The human mtDNA encompasses 16,569 bp with the numbering commencing at 0/16,569 and proceeding counter clockwise around the circular map. The function of each gene is identified by the use of colour according to the legend. Gene abbreviations: 12S rRNA = 12S ribosomal RNA; 16S rRNA = 16S ribosomal RNA; ATPase = ATP synthase F₀ subunit; CO = cytochrome *c* oxidase subunit; Cytb = cytochrome *b*; ND = NADH dehydrogenase subunit. The tRNA genes are indicated by the one letter code of the encoded amino acid. Nucleotide abbreviations: A = adenine; C = cytosine; G = guanine; U = uracil; R = A or G; Y = C or U; N = A, C, G or U. Amino acid abbreviations: A = alanine; C = cysteine; D = aspartic acid; E = glutamic acid; F = phenylalanine; G = glycine; H = histidine; I = isoleucine; K = lysine; L = leucine; M = methionine; N = asparagine; P = proline; Q = glutamine; R = arginine; S = serine; T = threonine; V = valine; W = tryptophan; Y = tyrosine. Other abbreviations: D-loop = displacement loop; HV1 = hypervariable segment 1; HV2 = hypervariable segment 2; IT_{H1} = major initiation site for H-strand transcription; IT_{H2} = minor initiation site for H-strand transcription; IT_L = initiation site for L-strand transcription; O_H = H-strand origin of replication; O_L = L-strand origin of replication; P_{H1} = HSP1, major H-strand promoter; P_{H2} = HSP2, minor H-strand promoter; P_L = LSP, L-strand promoter; TAS = termination-associated sequence. Adapted from Wallace (1992) and MITOMAP (2007).

The mitochondrial genome displays some features that are unique in comparison with the nuclear genome. For transcription of codons, the mitochondrial genome employs a genetic code that contains a number of differences compared to the universal genetic code (Barrell *et al.*, 1980, refer to Table 2.4). Certain codons that are used as stop codons in nuclear DNA (nDNA) encode amino acids in the mitochondrion, and some codons that encode a specific amino acid in nDNA will encode a different amino acid in the mitochondrion (refer to Section 2.4.5, Table 2.6). In addition, most genes possess incomplete termination codons with a T or TA following the last sense codon. The termination codon, TAA, is then completed post-transcriptionally in the messenger RNA (mRNA) through polyadenylation (Gillham, 1994).

Table 2.4 The mitochondrial genetic code

Amino acid	Codon	Amino acid	Codon	Amino acid	Codon
Ala A	GCA	His H	CAC	Ser(UCN) S	UCA
	GCC		CAU		UCC
	GCG	Ile I	AUC		UCG
	GCU		AUU		UCU
Arg R	CGA	Leu(CUN) L	CUA	Thr T	ACA
	CGC		CUC		ACC
	CGG		CUG		ACG
	CGU		CUU		ACU
Asn N	AAC	Leu(UUR) L	UUA	Trp W	UGA
	AAU		UUG		UGG
Asp D	GAC	Lys K	AAA	Tyr Y	UAC
	GAU		AAG		UAU
Cys C	UGC	Met M	AUA	Val V	GUA
	UGU		AUG		GUC
Gln Q	CAA	Phe F	UUC		GUG
	CAG		UUU		GUU
Glu E	GAA	Pro P	CCA	Ter *	AGA
	GAG		CCC		AGG
Gly G	GGA		CCG		UAA
	GGC		CCU		UAG
	GGG	Ser(AGY) S	AGC		
	GGU		AGU		

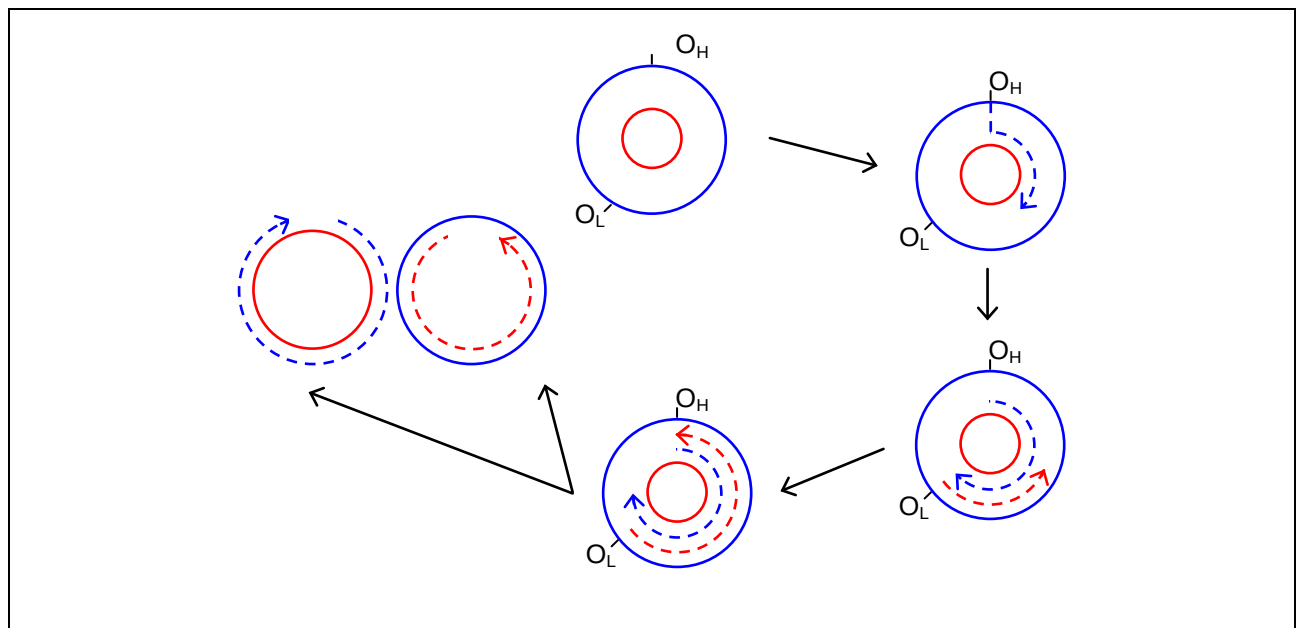
Nucleotide abbreviations: A = adenine; C = cytosine; G = guanine; U = uracil. Three letter amino acid abbreviations: Ala = alanine; Arg = arginine; Asn = asparagine; Asp = aspartic acid; Cys = cysteine; Gln = glutamine; Glu = glutamic acid; Gly = glycine; His = histidine; Ile = isoleucine; Leu = leucine; Lys = lysine; Met = methionine; Phe = phenylalanine; Pro = proline; Ser = serine; Thr = threonine; Trp = tryptophan; Tyr = tyrosine; Val = valine; Ter = termination. One letter amino acid abbreviations: A = alanine; C = cysteine; D = aspartic acid; E = glutamic acid; F = phenylalanine; G = glycine; H = histidine; I = isoleucine; K = lysine; L = leucine; M = methionine; N = asparagine; P = proline; Q = glutamine; R = arginine; S = serine; T = threonine; V = valine; W = tryptophan; Y = tyrosine; * = termination. In contrast to the universal genetic code, AUA codes for methionine instead of isoleucine, UGA codes for tryptophan instead of termination, and AGA and AGG are terminators instead of coding for arginine (refer to Table 2.6). During elongation, AUU encodes isoleucine, although it can also encode methionine for initiation. Adapted from MITOMAP (2008).

Mitochondria also make use of a simplified translation system (refer to Section 2.4.5). The translation of all codons can be performed by using less than the 32 different tRNA molecules used for the universal genetic code in nDNA. This decrease in the number of tRNAs employed is made possible through the use of a uracil (U) nucleotide in the first anticodon position of a single tRNA which recognises all the codons that form part of a specific four-codon family (Barrell *et al.*, 1980). In addition, extreme economy is observed regarding gene organisation within the mitochondrial genome. The mtDNA genome lacks introns (except for a single regulatory region), intergenic regions are either absent or limited to a few nucleotides, and some protein genes overlap (Montoya *et al.*, 1983).

The overlapping protein genes are positioned on the H-strand. The 3' end of *atp6* (encoding ATPase6) overlaps by 46 base pairs with the 5' end of *atp8* (encoding ATPase8). Similarly *ndh4L* (encoding ND4L) and *ndh4* (encoding ND4) overlap out of frame by seven base pairs (Anderson *et al.*, 1981). Another characteristic trait of the mitochondrial genome of mammals is the manner in which the tRNA genes are distributed. These genes are scattered throughout the mitochondrial genome and separate most of the rRNA and protein-coding genes. It has been suggested that the tRNA structures in the newly synthesised primary mRNA transcripts act as signals for RNA processing (Ojala *et al.*, 1981; Gillham, 1994).

2.4.3 Replication of the mitochondrial genome

Replication of human mitochondrial DNA is under relaxed control and is independent of the cell cycle phase (Clayton, 1982). Replication of mtDNA involves two strand-specific origins of replication that are spatially and temporally separate. The origin of replication for the H-strand (O_H) is situated downstream of the L-strand promoter (LSP) in the displacement loop (D-loop) region between the phenylalanine tRNA ($tRNA^{Phe}$) and the proline tRNA ($tRNA^{Pro}$), while the L-strand origin of replication (O_L) is flanked by five tRNA genes and is situated two thirds of the way around the genome from O_H in a non-coding region of ca. 30 base pairs. A diagrammatic representation of the replication process is presented in Figure 2.4.

Figure 2.4 Diagrammatic representation of the replication cycle of mitochondrial DNA

Blue lines = parental H-strands; red lines = parental L-strands; blue dashed lines = daughter H-strands; red dashed lines = daughter L-strands; O_H = H-strand origin of replication; O_L = L-strand origin of replication. Adapted from Clayton (1982).

The process of replication involves initiation at O_H which generates a triple-stranded structure, which consists of two DNA strands and a RNA primer of ca. 680 bp synthesised on the L-strand template (Anderson *et al.*, 1981). The newly synthesised daughter H-strand, referred to as 7S DNA, displaces the single-stranded parental H-strand to generate the D-loop. The H-strand replication commences at the 7S DNA and advances along the L-strand until the second origin of replication (O_L) has been passed and exposed by the replication fork. L-strand replication is subsequently initiated at O_L and advances in the opposite direction compared to that observed for H-strand replication (Clayton, 1982).

The replication of the H-strand is mediated through the formation of short mitochondrial transcripts originating at the initiation site for L-strand transcription (IT_L) which serve as primers for the initiation of replication. As there are no recognised differences between the initiation of L-strand transcription and RNA primer formation for mtDNA replication, it has been suggested that mammalian mtDNA replication and transcription is closely linked (Clayton, 1991). The transition from RNA to DNA synthesis takes place at a number of discrete sites that together comprise O_H in a region that contains three short evolutionary conserved sequence blocks, namely CSB I, II and III (Walberg and Clayton, 1981). It has been speculated that, due to their location, CSB I, II and III are involved in directing the precise enzymatic cleavage of the precursor RNA primer to generate the appropriate mature RNA primer species (Clayton, 1991). The enzymes involved in the processing of

the precursor RNA primers to yield the required mature 3' ends, are the mitochondrial RNA processing endonuclease (RNase MRP) and endonuclease G (Côté and Ruiz-Carrillo, 1993; Lee and Clayton, 1997).

Replication is subsequently initiated at O_L . This process involves the displacement of the parental H-strand, in the form of a single strand, by the growing daughter H-strand, after which it appears that O_L adopts a characteristic stem-loop structure (Clayton, 1991). This folded structure appears to serve as a recognition signal for a mitochondrial DNA primase which initiates L-strand synthesis by generating a short RNA primer (Wong and Clayton, 1985; Hixson *et al.*, 1986). The RNA primer begins at the T-rich section of the predicted stem-loop structure, and the transition from RNA to DNA synthesis subsequently occurs at a specific site located close to an essential GC-rich element positioned at the base of the stem-loop (Wong and Clayton, 1985; Hixson *et al.*, 1986).

The majority of H-strand synthesis events in vertebrates are suspended soon after initiation. The discontinued H-strands remain bound to their template L-strands to generate the triple-stranded D-loop structure (Clayton, 1982). H-strand synthesis must therefore be regulated to maintain a balance between D-loop formation, i.e. arrested H-strand elongation, and H-strand replication, i.e. H-strand elongation along the entire length of the genome (Taanman, 1999). This is achieved by the binding of unknown protein factors to a conserved sequence element (Shadel and Clayton, 1997) referred to as the termination-associated sequence (TAS).

The abovementioned process of replication requires a number of trans-acting factors. These factors include mtDNA polymerase γ (gamma) which is responsible for mtDNA synthesis (Bolden *et al.*, 1977), mitochondrial single-stranded binding proteins (mtSSB) which bind to exposed single-stranded mtDNA, preserve the integrity of these sections and stimulate the rate of DNA synthesis by mtDNA polymerase γ (Tiranti *et al.*, 1993), as well as additional enzymatic activities that are deemed critical for mtDNA replication. These additional enzymes include an ATP-dependent helicase responsible for catalysing the unwinding of duplex DNA to provide single-stranded DNA templates for mtDNA polymerase γ by disrupting the hydrogen bonds that hold the two strands together (Matson and Kaiser-Rogers, 1990), a type I topoisomerase which catalyses the relaxation of supercoiled DNA (Castora *et al.*, 1985), and a type II topoisomerase which is responsible for the introduction of supercoils (Castora *et al.*, 1985). Topoisomerases alter DNA

topology by temporarily disrupting the backbone bonds through the use of a mechanism which involves either single strand (type I) or double strand (type II) cleavage (Wang, 1996).

2.4.4 Transcription of the mitochondrial genome

A unique mode of DNA transcription has evolved to match the extremely compact and economical gene organisation of the mammalian mitochondrion (Attardi, 1986). Mitochondrial genes are transcribed from both strands, starting at promoters situated in the D-loop region, to produce polycistronic RNA molecules that are subsequently processed to yield the mRNA for specific polypeptides or tRNA molecules. The polypeptides, rRNA and tRNA molecules encoded by each strand are indicated in Table 2.5.

Table 2.5 Genes encoded by the H-strand and the L-strand of the mitochondrial DNA genome

Strand	rRNA	tRNA			Polypeptides		
H-strand	12S rRNA 16S rRNA	tRNA ^{Arg}	tRNA ^{Ile}	tRNA ^{Phe}	ND1		
		tRNA ^{Asp}	tRNA ^{Leu(CUN)}	tRNA ^{Ser(AGY)}	ND2		
		tRNA ^{Gly}	tRNA ^{Leu(UUR)}	tRNA ^{Thr}	ATPase6	COI	ND3
		tRNA ^{His}	tRNA ^{Lys}	tRNA ^{Trp}	ATPase8	COII	ND4
			tRNA ^{Met}	tRNA ^{Val}	Cytb	COIII	ND4L
L-strand	None	tRNA ^{Ala}	tRNA ^{Gln}	tRNA ^{Ser(UCN)}	ND6		
		tRNA ^{Asn}	tRNA ^{Glu}	tRNA ^{Tyr}			
		tRNA ^{Cys}	tRNA ^{Pro}				

S = Svedberg units; rRNA = ribosomal RNA; tRNA = transfer RNA; Ala = alanine; Asn = asparagine; Arg = arginine; Asp = aspartic acid; Cys = cysteine; Gln = glutamine; Glu = glutamic acid; Gly = glycine; His = histidine; Ile = isoleucine; Leu = leucine; Lys = lysine; Met = methionine; Phe = phenylalanine; Pro = proline; Ser = serine; Thr = threonine; Trp = tryptophan; Tyr = tyrosine; Val = valine; ATPase = adenosine triphosphatase subunit; Cytb = cytochrome *b*; CO = cytochrome *c* oxidase subunit; ND = NADH dehydrogenase subunit; A = adenine; C = cytosine; G = guanine; U = uracil; R = A or G; Y = C or U; N = A, C, G or U. Adapted from Attardi (1993).

There is a single initiation site for L-strand transcription (IT_L), and two closely located initiation sites for H-strand transcription (IT_{H1} and IT_{H2}). All the initiation sites are positioned within the D-loop region (Attardi, 1986). The major transcription initiation sites, IT_L and IT_{H1}, are surrounded by a promoter element with a 15 base pair consensus sequence motif, 5'-CANACC(G)CC(A)A¹AAGAYA-3', while the promoter region of the minor transcription initiation site, IT_{H2}, exhibits only limited similarity with the consensus sequence (Chang and Clayton, 1984; Hixson and Clayton, 1985). Transcription of the

¹ The transcription initiation site is indicated by underlining the relevant nucleotide.

L-strand and the H-strand takes place simultaneously, resulting in the complete symmetrical transcription of the mtDNA (Murphy *et al.*, 1975).

Transcription of the H-strand is initiated at IT_{H1} situated at nucleotide position 561 which is present within the H-strand promoter (HSP) and directly adjoining to the tRNA^{Phe} gene. Transcription of the L-strand is initiated at IT_L located at nucleotide position 407 which is present within the L-strand promoter (LSP). A second putative H-strand initiation site, IT_{H2} , is located within the tRNA^{Phe} gene at nucleotide position 638 immediately adjacent to the 12S rRNA gene (Montoya *et al.*, 1982). Although HSP and LSP are located in close proximity, it has been demonstrated that these two elements function independently (Moraes *et al.*, 1991).

Mitochondrial DNA transcription requires the involvement of two proteins, a core mitochondrial RNA polymerase which is relatively non-selective, and a dissociable transcription factor which provides the polymerase with promoter selectivity (Fisher and Clayton, 1985). It is anticipated that the RNA polymerase will interact with the HSP and LSP (Topper and Clayton, 1989), while the transcription factor, Tfam (previously referred to as h-mtTF-1 or h-mtTFA), is involved in binding to enhancer elements which are located upstream from the promoter regions and are required for optimal transcription.

Tfam is a 25 kDa mitochondrial protein consisting of two high mobility group (HMG) domains divided by an amino acid linker containing 27 residues, and followed by an amino acid C-terminal tail consisting of 25 basic residues (Fisher and Clayton, 1988; Parisi and Clayton, 1991; Gleyzer *et al.*, 2005). It is believed that the HMG domains are engaged in the binding of DNA (Grosschedl *et al.*, 1994), while the C-terminal tail is critical in specific recognition of DNA, and is required for high levels of specific initiation of transcription (Dairaghi *et al.*, 1995). The binding of Tfam to regions ten to 40 base pairs upstream from each start site within the HSP and LSP is required for specific initiation of transcription (Fisher *et al.*, 1987; Fisher *et al.*, 1991). Tfam binding induces changes in the mtDNA topology which may be necessary to ensure that the RNA polymerase can gain access to the DNA template in order to initiate the process of transcription (Taanman, 1999).

The affinity of Tfam for its binding site upstream of LSP is fairly strong, while its affinity for its binding site upstream of HSP is comparatively weaker. In addition, Tfam stimulates H-strand transcription moderately, while it enhances L-strand transcription markedly (Fisher and Clayton, 1988; Fisher *et al.*, 1991; Ghivizzani *et al.*, 1994). This suggests the

existence of additional factors that may be necessary for effective initiation of transcription (Taanman, 1999). Two isoforms, TFB1M and TFB2M (also referred to as h-mtTFB), of human mitochondrial transcription specificity factor have been identified in mammalian cells (McCulloch *et al.*, 2002). This protein functions in conjunction with mitochondrial RNA polymerase and Tfam to regulate the correct transcription initiation from HSP and LSP and markedly enhances mtDNA transcription.

After initiation at the LSP, the L-strand is transcribed to yield a single polycistronic precursor RNA which contains most, or all, of the genetic information encoded on the strand (Murphy *et al.*, 1975). In contrast, a dual transcription model has been proposed for the H-strand (Montoya *et al.*, 1983). This model postulates that transcription is frequently initiated at IT_{H1} and terminated at the downstream end of the 16S rRNA gene, thus leading to the synthesis of nearly all of the two rRNA molecules. In comparison, transcription is initiated less frequently at IT_{H2} resulting in a polycistronic molecule that corresponds to practically the complete H-strand and which consists of the majority of the tRNAs and all the mRNAs encoded by the H-strand. This dual transcription model is employed to explain the observed higher rate of synthesis of rRNAs compared to single mRNAs encoded on the H-strand in HeLa cells (Gelfand and Attardi, 1981).

The difference in synthesis rates of rRNA and mRNA can also be attributed to an attenuation event at the boundary between the 16S rRNA and tRNA^{Leu} genes. The early termination of the polycistronic H-strand transcript is mediated by a transcription factor referred to as mTERF (Kruse *et al.*, 1989) or mtTERM (Hess *et al.*, 1991). This factor induces curving of the DNA helix (Shang and Clayton, 1994), and binds to a conserved tridecamer sequence block, involved in directing termination (Kruse *et al.*, 1989), present within the tRNA^{Leu(UUR)} gene located at nucleotide position 3,237 to 3,249 of the human mtDNA. mtTERM functions bidirectionally when attached to its mtDNA target site and exhibits an increase in termination efficiency in the reverse orientation relative to the position of the promoter (Christianson and Clayton, 1986). Therefore, mtTERM may arrest the transcription of the L-strand at a position where no L-strand encoded genes are detected downstream, in addition to its attenuating role in H-strand transcription.

Binding of mtTERM inhibits transcription elongation by acting as a physical barrier instead of a specific interaction with the mitochondrial RNA polymerase (Shang and Clayton, 1994). The structure of mtTERM has been investigated by isolating the predominant polypeptide from a mitochondrial lysate fraction associated with mtTERM functionality,

which contained polypeptides with a molecular weight of around 34 kDa (Hess *et al.*, 1991; Daga *et al.*, 1993). The polypeptide has been characterised and consists of two widely separated basic regions and three leucine zipper motifs which are deemed essential for its specific DNA-binding capacity (Fernandez-Silva *et al.*, 1997). This polypeptide however did not perform all the functions associated with mtTERM and it has been suggested that termination activity requires an additional component of the 34 kDa polypeptide fraction (Taanman, 1999).

The transcription of the H-strand and L-strand in the form of polycistronic molecules necessitates processing of these primary transcripts to yield the three different molecules, i.e. rRNAs, mRNAs and tRNAs, encoded on each strand. Since, vertebrate mtDNA contains no intron sequences and intergenic sequences are limited (Anderson *et al.*, 1981), the formation of the different molecules from the polycistronic strands appears to be a fairly uncomplicated process that requires only a few enzymes.

The two rRNA genes and nearly all the protein-coding genes are flanked by tRNA genes. This unique genetic organisation facilitates the production of correctly processed rRNAs and mRNAs through the accurate endonucleolytic excision of tRNAs from the polycistronic primary transcripts (Montoya *et al.*, 1981; Ojala *et al.*, 1981). Two types of enzymatic activity are required for this process, an enzymatic activity that will cleave the transcripts precisely on the 5' side of the tRNA provided by mitochondrial RNase P (mtRNase P), and another activity, provided by an uncharacterised 3' endonuclease, which will cleave the transcript precisely on the 3' side of the tRNA (Attardi, 1993). The excised tRNAs are subsequently modified by the addition of a CCA sequence at their 3'-ends through an enzymatic reaction catalysed by adenosine triphosphate (cytidine triphosphate): tRNA nucleotidyl transferase (Attardi, 1993; Rossmannith *et al.*, 1995). In addition, both the rRNA and mRNA species resulting from the cleavage process are modified by the addition of adenine residues at their 3'-ends through an enzymatic reaction catalysed by poly(A) polymerase (Attardi, 1993).

2.4.5 Translation of the mitochondrial genome

The human mitochondrial translation apparatus consists of mitochondria-specific 55S mitochondrial ribosomes (Attardi and Ojala, 1971), 22 mtDNA-encoded tRNA species (Anderson *et al.*, 1981) and a set of specific nuclear-encoded aminoacyl tRNA synthetases, initiation and elongation factors. The 55S mitochondrial ribosomes consist of

a 39S major subunit and a 28S minor subunit, containing, respectively, the mtDNA-encoded 16S rRNA and 12S rRNA species (Attardi and Ojala, 1971), as well as distinctive sets of nuclear-encoded ribosomal proteins (Matthews *et al.*, 1982).

The translation process is initiated at the start or in close proximity to the 5'-end of the mitochondrial mRNA with the codon coding for *N*-formyl-methionine (Montoya *et al.*, 1981). The small 28S ribosomal subunit binds mRNA in a sequence-independent manner and without the requirement for secondary initiation factors or initiator tRNA (Liao and Spremulli, 1989). After attachment of the small 28S ribosomal subunit to the mRNA, the subunit moves to the 5'-end of the mRNA assisted by unspecified secondary initiation factors (Denslow *et al.*, 1989; Liao and Spremulli, 1990a). To date, mtIF-2 is the only initiation factor identified in mammalian mitochondria (Liao and Spremulli, 1990b; Liao and Spremulli, 1991). It is a monomeric protein belonging to the GTPases family and promotes fMet-tRNA attachment to the 28S ribosomal subunit in the presence of guanosine-5'-triphosphate (GTP) and a template (Ma and Spremulli, 1995). The hydrolysis of GTP mediates the release of mtIF-2, as well as the formation of the 55S initiation complex through the accompanying joining of the large 39S ribosomal subunit with the small 28S ribosomal subunit (Liao and Spremulli, 1991). Subunit joining is however not strictly dependent on GTP hydrolysis. Three mitochondrial elongation factors, i.e. mtEF-Tu, mtEF-G and mtEF-Ts, have been identified. Mammalian mtEF-Ts and mtEF-Tu form a tightly associated complex that requires the presence of both GTP and charged tRNAs for dissociation (Schwartzbach and Spremulli, 1989; Worliax *et al.*, 1997).

The number and specificities of the tRNAs encoded in human mtDNA reflect the unusual features of the genetic code and the pattern of codon recognition in mammalian mitochondria (Attardi, 1985). The differences between the genetic code of mammalian mitochondria and the universal genetic code are indicated in Table 2.6.

Table 2.6 Differences between the genetic code of mammalian mitochondria and the universal genetic code

Codon	Mammalian mitochondrial code	Universal code
UGA	Tryptophan	STOP
AUA	Methionine	Isoleucine
AGA	STOP	Arginine
AGG	STOP	Arginine

Adapted from Attardi (1993).

Another feature of the mitochondrial genetic system is the unusual codon recognition pattern. The translation of all codons can be performed by using only 22 tRNA molecules, which is less than the 32 different tRNA molecules used for the universal genetic code in nDNA (Attardi, 1993). This decrease in the number of tRNAs used, is due to the use of a uracil (U) nucleotide in the first anticodon position of a single tRNA which recognises all the codons contained within a four-codon family (Barrell *et al.*, 1980), the use of AGA and/or AGG as stop codons rather than as arginine codons, and the corresponding absence of an anticodon for these codons in the mtDNA sequence (Attardi, 1993). Furthermore, only one tRNA^{Met} gene occurs in the mammalian mitochondrial genome, and it is believed that this gene can specify both the methionine tRNA and the *N*-formyl-methionine tRNA through secondary modifications of the primary transcript (Attardi, 1993).

2.5 GENETIC VARIATION AND THE MITOCHONDRIAL DNA GENOME

When discussing genetic variation and the mitochondrial DNA genome, it must be considered that mitochondrial genetics is unique in the sense that the principles of genetics as appropriate for individuals (a single mtDNA genome) and populations (multiple mtDNA genomes observed within a single mitochondrion and mtDNA genomes observed within different individuals belonging to the human population) are applicable. Variation in the mitochondrial DNA genome can be introduced in the form of nucleotide substitutions (synonymous and non-synonymous substitutions), length differences (insertions and deletions), and the processes of rearrangement and possibly recombination (refer to Section 2.5.3). Mitochondrial DNA genomes are subsequently replicated and transmitted via the germ-line in the maternal lineage to offspring (refer to Section 2.4.1). In oogenesis, at a point in time between the primordial germ cell stage and the formation of the primary oocyte, the population of mitochondria pass through a bottleneck. This event can lead to an outcome whereby an mtDNA molecule that represented a small minority of the molecules in the mother's somatic cells can become typical in her offspring with proportions ranging from zero to a large majority. This phenomenon is referred to as cytoplasmic segregation, and can provide an explanation for the extensive variation observed in the severity of mtDNA-associated diseases between subsequent generations (refer to Section 2.6). Variation on a population level is first introduced via the high mutation rate of mtDNA genomes (refer to Section 2.5.1), which makes it possible that more than one type of mtDNA molecule can be present within a mitochondrion at different concentrations at any given time (refer to Section 2.5.2). The variation introduced into the

mitochondrial DNA genome through these mutational events and processes such as cytoplasmic segregation is further governed and shaped on a population level by other evolutionary forces and population events such as random genetic drift, natural selection, migration, effective population size, population substructure, population expansions and population reductions. All of these forces and events contribute to shaping the pattern of variation across mtDNA genomes and human populations.

The analysis of mtDNA variation has generated a number of important scientific breakthroughs. The observed variation in the mitochondrial DNA sequence allows for the reconstruction of the evolutionary history of populations by corroborating the “Recent out of Africa” hypothesis of human origins (refer to Section 3.1.2) and provides further insight in the role played by migrations in shaping human populations (Pakendorf and Stoneking, 2005). In addition, variation in the mitochondrial DNA sequence also allows for the study of disease linked to genetic change which has significant implications for the fields of pharmaceutical development and medicine. Due to the increased age of African populations compared to other human populations, the investigation of African mtDNA diversity is particularly important for the reconstruction of human evolutionary history, understanding the genetic basis of phenotypic variation and disease, as well as the mapping of genetic disorders that are common in individuals of recent African origin (Tishkoff and Williams, 2002).

2.5.1 Mutation rate of the mitochondrial genome

The greatest source of genetic variation is mutation which provides the raw material for evolution to act upon (Hartl, 1988). Variation in mtDNA genomes can be introduced through insertion and deletion events. Although mammalian mtDNAs are relatively uniform in size, minor length variants have been observed. These differences usually occur in non-coding sequences and are common in the control region (Aquadro and Greenberg, 1983; Cann and Wilson; 1983; Hauswirth and Clayton, 1985). Many of the smaller length differences, involving the insertion of one or two nucleotides, occur in homopolymer tracts such as the poly(C) tracts in the control region starting at nucleotide positions 303 and 311 (Aquadro and Greenberg, 1983). Differences in length can also arise from differences in copy number of short, tandemly repeated (STR) sequences such as the intergenic COII/tRNA^{Lys} nine base pair deletion which gives rise to a change in the copy number of a nine base pair tandem repeat motif, namely 5'-CCCCCTCTA-3' starting at nucleotide position 8,272 (Wrischnik *et al.*, 1987). Larger size differences can be caused by

duplication or deletion of stretches of sequence (Upholt and David, 1977; Wong *et al.*, 1983; Wrischnik *et al.*, 1987). Both small and large size differences can be caused by replication slippage, however larger size differences require mediation by short dispersed repeats (Albertini *et al.*, 1982, Levinson and Gutman, 1987).

Variation in mtDNA genomes are however primarily introduced through single nucleotide changes and the rate at which variation is introduced into the mtDNA genome is determined by the nucleotide substitution rate. The nucleotide substitution rate of primate mtDNA is several orders of magnitude higher (five- to tenfold) than that observed for the nuclear genome (Brown *et al.*, 1979). Factors that could be involved in generating the higher nucleotide substitution rate in mitochondrial genomes compared to nuclear genomes include the lack of protective histones, greater possibility of oxidative damage due to the continuous exposure to the mutagenic effects of free radicals generated by OXPHOS, a more error-prone system of replication, a less efficient editing or repair function, a higher turnover rate and the phenomenon of relaxed selection (Brown *et al.*, 1982; Wallace, 1992). Relaxed selection entails decreasing the effect of natural selection through a mechanism that prevents selection from acting on the genome, thereby leading to an increase in the number of nucleotide substitutions observed. Relaxed selection operates when deleterious alterations accumulate due to the slow elimination of these alterations which is caused by the fact that the alterations are no longer deleterious in an improved environment (Takahata, 1993).

The nucleotide substitution rate of primate mtDNA is characterised by three factors, namely the high frequency of synonymous substitutions, a high transition/transversion ratio (ca. 21.5) and a strong bias in the L-strand towards C to T transitions (Gillham, 1994; Kivisild *et al.*, 2006). A bias towards transitions ($A \rightleftharpoons G$, $C \rightleftharpoons T$), instead of transversions ($A \rightleftharpoons C$, $A \rightleftharpoons T$, $G \rightleftharpoons C$, $G \rightleftharpoons T$), has been observed for point mutations in the mtDNA genome of higher order primates, including humans, for both tRNA and protein-coding genes (Brown *et al.*, 1982; Greenberg *et al.*, 1983). This has been confirmed in another study (Crespillo *et al.*, 2004), in which it was observed that the majority (88%) of deviations between mitochondrial sequences were made up of transitions, whereas only 8% were attributed to transversions. The observation of transition bias has been ascribed to a bias in the mutation process, rather than the effect of selection on the mutation (Brown *et al.*, 1982). The percentage of transitions decreases when more distantly related organisms are compared, and this decrease has been attributed to the occurrence of

multiple substitutions at the same nucleotide site. This demonstrates the importance of comparing closely related species when inferring evolutionary relationships.

Multiple substitutions at the same nucleotide site can give rise to recurrent or parallel alterations, as well as reversion events which have been observed in phylogenetic networks of mtDNA genome sequences (Herrnstadt *et al.*, 2002a). These recurrent alterations can be classified as examples of parallel evolution, i.e. independent evolution of the same feature from the same ancestral condition, convergent evolution i.e. independent evolution of the same feature from different ancestral conditions, or secondary loss, i.e. reversion to an ancestral condition (Page and Holmes, 1998). These recurrent alterations and reversion events usually occur at mutational hot spots. Several mutational hot spots have been identified in both the control region (e.g. Bandelt *et al.*, 2001a) and the coding region (e.g. Herrnstadt *et al.*, 2002a; Galtier *et al.*, 2006; Kivisild *et al.*, 2006) of the mitochondrial DNA genome. The identification of recurrent alterations and reversion events within phylogenetic networks of mtDNA genome sequences are of importance as the presence of these events can influence the estimate of the sequence divergence by leading to underestimates of the true value (Jobling *et al.*, 2004).

Different rates of nucleotide substitution have been observed in the control and coding regions of the mtDNA genome with a relatively high rate in the control region compared to a relatively low rate in the coding region (Jobling *et al.*, 2004). The nucleotide substitution rate of the mitochondrial genome has been calculated to be 0.025 to 0.26×10^{-6} per site per year for the control region (Parsons *et al.*, 1997), or 0.017×10^{-6} substitutions per site per year for the complete genome excluding the control region, i.e. the coding region of the genome (Ingman *et al.*, 2000). The nucleotide substitution rate for the two hypervariable segments (HVS1 and HVS2) of the non-coding control region have been observed to be higher than that of the coding region of the mitochondrial genome, the exact rate is, however, a matter of debate. Phylogenetic comparisons yield estimates of 0.075 to 0.165×10^{-6} substitutions per site per year (Ward *et al.*, 1991; Stoneking *et al.*, 1992; Hasegawa *et al.*, 1993; Tamura and Nei, 1993), while direct observations of mtDNA mutations in deep-rooting pedigrees or families yield estimates ranging from 0.0 to 1.46×10^{-6} substitutions per site per year with an overall average of 0.47×10^{-6} substitutions per site per year (Howell *et al.*, 2003).

The estimate produced by pedigree analysis is significantly higher than that yielded by phylogenetic analysis, leading to confusion as to which rate is correct and should be used

in studies of population history (Pääbo, 1996; Howell and Mackey, 1997; Macaulay *et al.*, 1997; Parsons *et al.*, 1997; Jazin *et al.*, 1998; Sigurðardóttir *et al.*, 2000; Heyer *et al.*, 2001; Howell *et al.*, 2003). This discrepancy can be attributed to the extremely heterogeneous mutation rate, as well as the existence of phylogenetically defined mutational hot spots within the control region. These sites mutate four or five times faster than the average site within the control region (Aquadro and Greenberg, 1983; Greenberg *et al.*, 1983; Hasegawa *et al.*, 1993; Wakeley, 1993; Excoffier and Yang, 1999; Meyer *et al.*, 1999; Heyer *et al.*, 2001), and recent mutations arise preferentially at these sites (Stoneking, 2000; Heyer *et al.*, 2001). The hypermutability of these sites can be linked to an increased tendency to damage, as it has been demonstrated that in the control region of ancient human DNA samples, phylogenetically defined mutational hot spots are preferentially affected by DNA damage (Gilbert *et al.*, 2003).

The observed difference between pedigree-based and phylogenetically based estimates of the human mtDNA mutation rate could be due to the fact that mutations at fast-evolving sites are preferentially detected in pedigree studies, whereas phylogenetic studies also detect mutations at slowly evolving sites (Pääbo, 1996; Macaulay *et al.*, 1997; Jazin *et al.*, 1998; Heyer *et al.*, 2001). The discrepancy can also be explained by taking into account other factors that could play a role. These factors include genetic drift, the presence of mutational hot spots, the role of negative selection, and insufficient detection of recurrent mutations in phylogenetic studies.

The choice of which estimate to use in population history studies, will depend on the specific purpose of the investigation. The phylogenetically based estimate may be preferred for investigations of deep history since it takes into account mutations that have reached an appreciable frequency in the population, while the pedigree-based estimate will be preferred for studies of recent history as it takes into account mutations that have arisen recently and have not had time to become fixed within the population (Pääbo, 1996; Macaulay *et al.*, 1997). Alternatively, evolutionary models that permit the inclusion of different evolutionary rates for different classes of sites should be incorporated in studies of population history (Hasegawa *et al.*, 1993; Tamura and Nei, 1993; Heyer *et al.*, 2001). In addition, it is important to consider that the actual rate of human mtDNA mutation is not reflected by average estimates, and these estimates should therefore be employed as simplistic tools for phylogenetic studies (Pakendorf and Stoneking, 2005).

Mitochondrial DNA is used extensively as a marker in studies of evolution with the implicit assumption that it is a neutral marker. This assumption is important in a variety of applications such as the estimation of population size (Wilson *et al.*, 1985), the measurement of gene flow (Slatkin, 1985), the detection of subdivisions in populations (Avice *et al.*, 1987), and in the dating of events through the use of a molecular clock (Cann *et al.*, 1987). Evidence in support of the violation of clock-like mtDNA evolution in Africa has however been presented (Torroni *et al.*, 2001; Howell *et al.*, 2004). This implies that during the use of mtDNA to date phylogenetic events, care must be taken as the fundamental assumption of neutral clock-like mutation may not be valid (Pakendorf and Stoneking, 2005). It is therefore advisable to perform a neutrality test before assuming that the neutral model of evolution applies, i.e. that mtDNA will exhibit clock-like behaviour (Nachman *et al.*, 1996).

A number of sequence evolution models have been proposed to describe the nucleotide substitution process, and these models also attempt to take into account aspects such as the differences in substitution rate at different nucleotide sites, as well as the bias towards transitions observed in mtDNA. The most commonly used models and the parameters addressed by each model, as well as the assumptions under which each model was designed to function, are discussed in Section 3.2.2.1.

2.5.2 Heteroplasmy and the mitochondrial genome

Mitochondrial DNA is present at high copy numbers in human cells. A single cell contains ca. 107 mitochondria which each contains one to 15 copies of mtDNA (Satoh and Kuroiwa, 1991). It has been observed that due to mutational events (refer to Section 2.5.1), the copies of the mtDNA genome observed within a single individual are not always identical. A mutation can be present in only a fraction of the mtDNA copies and this fraction of mutated mtDNA may vary significantly between different tissues or between cells of the same tissue, a situation referred to as heteroplasmy. In some cases, a mutation can also be present in all of the copies of mtDNA, a situation referred to as homoplasmy. Subsequently, during mitotic and meiotic replication the different mtDNA types present within an individual are randomly allocated to daughter cells and this process is referred to as replicative segregation (Wallace, 1994). In somatic cells, mitotic replicative segregation requires a number of cell divisions for a mutation to reach homoplasmy. In contrast, in germ-line cells, meiotic replicative segregation can alter the

proportion of the mutant mtDNA from predominantly normal mtDNA to mostly mutant mtDNA within a generation or two (Wallace, 1992).

The observed frequency of heteroplasmy in humans is dependent on the measurement method. Initial estimates, which indicated an absence of heteroplasmy or very low levels of heteroplasmy, were based on results obtained through the use of relatively insensitive methods such as automated sequencing (Bendall *et al.*, 1996) and single stranded conformational polymorphism (SSCP) analysis (Gocke *et al.*, 1998). Using a denaturing gradient-gel electrophoresis (DGGE) assay, it was estimated that mtDNA genome molecules of ca. 14% of humans consist of more than one type, with the second mtDNA type observed at a frequency of ca. 1% (Tully *et al.*, 2000).

In most cases, the term heteroplasmy is used when referring to the presence of a mutation within the mtDNA genome which causes a mitochondrial disease (refer to Section 2.6). It has been observed that heteroplasmic mutations are frequently pathogenic, while homoplasmic mutations are generally neutral polymorphisms (Shanske *et al.*, 2001). For dysfunction and expression of the clinical phenotype to occur when heteroplasmy is observed, a critical number of mutated mtDNA must be present. This phenomenon is referred to as the threshold effect (Wallace, 1992).

In the case of mitochondrial disease, it is essential to estimate the risk of maternal transmission of pathogenic heteroplasmic mtDNA mutations from an affected mother to her offspring. Two mechanisms may be involved in the inheritance of heteroplasmy, the first is selection and the second is random genetic drift (Chinnery *et al.*, 2000). If selection is the predominant mechanism, positive selection could lead to an increase in mutation load in subsequent generations with the accompanying earlier onset of a more severe phenotype, while negative selection could decrease the mutation load and could cause the deleterious mutation to be lost without clinical effects. Random genetic drift requires a genetic “bottleneck” where the number of mtDNA molecules is reduced within the cell during oogenesis, followed by mitotic segregation which leads to the unequal distribution of the available mtDNA molecules in daughter cells. This could lead to the loss of a deleterious mutation, or to a mutation rapidly becoming homoplasmic within an individual and causing a severe clinical phenotype. The transmission of heteroplasmy has been studied, and based on the available data it was concluded that the predominant mechanism involved is random genetic drift (Chinnery *et al.*, 2000).

2.5.3 Recombination and the mitochondrial genome

The occurrence of mitochondrial DNA recombination could have grave consequences for the research of molecular evolution. Phylogenetic analysis performed on datasets consisting of sequence data that has undergone recombination, could produce phylogenetic trees that resemble those obtained from the analysis of sequence data of a population experiencing exponential growth (Schierup and Hein, 2000). The failure to detect recombination could lead to an underestimation of the time to the most recent common ancestor (TMRCA) of the sequences, as well as the incorrect rejection of the molecular clock (Schierup and Hein, 2000). Evidence for the support or rejection of recombination events in mitochondria is presented in Table 2.7.

Table 2.7 Evidence in support or rejection of recombination events within mitochondria

Evidence in support of recombination events	Reference
Incidences of paternal leakage can occur.	Schwartz and Vissing, 2002
Copies of mtDNA sequences are present in the nuclear genome.	Bensasson <i>et al.</i> , 2001
Existing methods for the detection of recombination are not very powerful.	Posada and Crandall, 2001 Wiuf <i>et al.</i> , 2001
Linkage disequilibrium analyses indicate the presence of recombination in five out of six hominid datasets.	Awadalla <i>et al.</i> , 1999
In certain animal mitochondrial genomes in which heteroplasmy is the rule, recombination does occur.	Ladoukakis and Zouros, 2001
Biochemical investigations imply that recombination is an essential component of the mtDNA replication and repair machinery and that the majority of animal genomes possess the enzymes required for recombination of mtDNA.	Thyagarajan <i>et al.</i> , 1996
Evidence leading to rejection of recombination events	Reference
Heteroplasmy is required to produce new haplotypes, heteroplasmy is however rare.	Rokas <i>et al.</i> , 2003
Failure to observe clear cases of recombinant haplotypes in natural populations.	Rokas <i>et al.</i> , 2003
Mitochondria in somatic cell hybrids do not harbour recombinant haplotypes.	Zuckerman <i>et al.</i> , 1984
mtDNA molecules become sequestered into clusters called nucleoids which prevent physical contact between unrelated mtDNA molecules.	Satoh and Kuroiwa, 1991
The increased mtDNA mutation rate is interpreted as an indication of the absence of recombination.	Howell, 1997
mtDNA recombination in human mitochondria is not generally supported by statistical analyses.	Eyre-Walker, 2000 Ingman <i>et al.</i> , 2000

It has always been accepted as an established fact that mtDNA recombination does not occur (Stoneking and Soodyall, 1996; Wallace *et al.*, 1999). However, this belief was challenged in 1999 and 2000 when four papers claimed evidence for human mtDNA

recombination (Awadalla *et al.*, 1999; Eyre-Walker *et al.*, 1999; Hagelberg *et al.*, 1999; Awadalla *et al.*, 2000). Three of the four studies were based on statistical and phylogenetic analyses of mtDNA sequences. The argument was made that evidence for recombination was provided by the observation of an excess of homoplasmic sites in phylogenetic trees based on mtDNA sequences (Eyre-Walker *et al.*, 1999), and the association between linkage disequilibrium and distance across the mtDNA genome (Awadalla *et al.*, 1999; Awadalla *et al.*, 2000). The fourth study investigated mtDNA diversity in a Melanesian population and presented sequence data that was interpreted as direct evidence for recombination (Hagelberg *et al.*, 1999).

It was however subsequently demonstrated that the first three studies made use of erroneous data and/or questionable statistical methods and significant results were not obtained following reanalyses (Arctander, 1999; Macaulay *et al.*, 1999a; Merriweather and Kaestle, 1999; Jorde and Bamshad, 2000; Kumar *et al.*, 2000; Parsons and Irwin, 2000). The claim made in the fourth study was retracted as it was based on an error introduced into the sequence alignment (Hagelberg *et al.*, 2000). However, recombination has been observed in human mtDNA in the only reported individual possessing mtDNA molecules from both maternal and paternal origin (Schwartz and Vissing, 2002; Kraytsberg *et al.*, 2004). Within 0.7% of the total mtDNA molecules present in this individual's muscle tissue, recombination occurred between maternal and paternal mtDNA. This finding emphasises that recombination remains possible due to the fact that mitochondria possess a functional recombinase (Thyagarajan *et al.*, 1996).

In the context of the current investigation, the occurrence of recombination could influence the sequence data generated, the phylogenies constructed and estimates of TMRCA and it is therefore an important consideration. Based on the discussion in the preceding paragraphs and the evidence presented in Table 2.7, a few important conclusions can be drawn. The occurrence of recombination events within an individual is possible. For a recombination event to have a measurable effect or outcome, i.e. to generate new haplotypes, it requires the presence of different types of mtDNA molecules within a single individual. Recombination events could of course occur between identical molecules, in which case it would not be possible to detect the event or an outcome caused by this event. Paternal leakage leading to heteroplasmy and heteroplasmy due to mutation events within the maternal lineage could function as sources of different mtDNA molecules, although both paternal leakage and heteroplasmy are rare phenomena (refer to Sections 2.4.1 and 2.5.2). It must therefore be concluded that due to the rarity of

heteroplasmy, recombination, whether it in fact occurs or not, will generally not influence the results obtained when using mitochondrial DNA in evolutionary studies.

2.5.4 Other evolutionary forces and population events that influence genetic variation in the mitochondrial genome

The variation created by mutational events (refer to Section 2.5.1) is influenced by a number of other evolutionary forces and population events such as random genetic drift, migration, effective population size, population size changes (expansions or reductions), and population substructure. Random genetic drift, defined as the fluctuation in allele frequencies between subsequent generations due to the random sampling of gametes (Hartl, 1988), lead to the loss or fixation of specific mtDNA haplotypes within a population. The effect of random genetic drift on genetic variation is determined by the effective population size (N_e) which is the number of individuals in each generation that are successful in producing offspring. Large effective populations can sustain high levels of genetic variation and are not severely affected by random genetic drift, while small effective populations are easily influenced by random genetic drift and contain low levels of genetic variation. The effects of random genetic drift are especially visible in populations that have experienced a population bottleneck event (an extreme decrease in population size), or in population founding events (a smaller group is separated from a larger group and subsequently colonises a new area). Mitochondrial DNA haplotypes that were prevalent prior to the bottleneck event may be lost and replaced by less common haplotypes, or rare haplotypes may displace widespread haplotypes. Likewise, haplotypes that were widespread in the larger population from which the founding population derived may be lost or decrease in frequency, while haplotypes that were uncommon in the parental population may increase in frequency in the founder population (Tishkoff and Verrelli, 2003).

The effect of random genetic drift on genetic variation is also influenced by migration. Migration, which is the movement of individuals among sub-populations, results in gene flow which can limit the amount of genetic divergence between sub-populations (Hartl, 1988). Using the effective population size and the rate of migration (m), which is assumed to be constant per generation, the reduction in random genetic drift due to migration can be estimated (Cavalli-Sforza *et al.*, 1994). This reduction in random genetic drift is proportional to $N_e m$. If $N_e m$ is large, genetic drift does not play a significant role in altering genetic diversity and the frequencies of different mtDNA haplotypes remain

relatively stable (Tishkoff and Verrelli, 2003). If $N_e m$ is small, random genetic drift plays a bigger part in shaping genetic diversity and the frequencies of different mtDNA haplotypes are subject to large fluctuations (Tishkoff and Verrelli, 2003). Although migration can limit genetic divergence between populations, it can also increase genetic divergence between populations. Mass migrations of groups into new geographic regions can over time lead to genetic divergence between the groups and the population from which the groups originated. In addition, such mass migrations are often followed by rapid population expansion events. The most important migration event which affected the patterns of genetic variation in the human population was undoubtedly the migration of anatomically modern humans out of Africa. This event is estimated to have occurred between 44,000 and 200,000 years before present (Goldstein *et al.*, 1995; Tishkoff *et al.*, 1996; Stoneking *et al.*, 1997; Harris and Hey, 1999; Ingman *et al.*, 2000; Underhill *et al.*, 2000). This migration from Africa is discussed in more detail in Section 3.1.

Effective population size is an important population factor and as already mentioned, can influence genetic variation through interaction with random genetic drift and migration. The effective population size is a measure that corrects for age-related differences in reproductive rate, unequal numbers of males and females and unequal family sizes within populations (Hartl, 1988). N_e is estimated using the mutation rate (μ) and nucleotide diversity parameter (θ) which is usually determined using two different measures, namely S which is the number of segregating sites in a sample of nucleotide sequences (Watterson, 1975) or π which is the average number of pairwise differences between nucleotide sequences (Slatkin and Hudson, 1991). For haploid genomes such as the mtDNA genome, $N_e = \theta / 2\mu$. Estimates of N_e in contemporary populations reveal the size of the female component of the human population prior to population expansion. Previous investigations of mtDNA genetic diversity have indicated an effective population size of ca. 5,000 female individuals (Jorde *et al.*, 1998). It is however still uncertain what the effective population size of the human population has been over the long term, although evidence has been presented that human populations dispersed over a large geographic area and experienced a rapid increase in size within the past 10,000 to 50,000 years (Rogers, 1995). Within the past 10,000 years, the human population expanded more rapidly due to the development and spread of agriculture (Cavalli-Sforza *et al.*, 1994).

Slatkin and Hudson (1991) demonstrated that a population expansion event produces a star-like phylogenetic tree. This star-like pattern is reproduced as a unimodal distribution

of the pairwise differences between nucleotide sequences (Rogers and Harpending, 1992). In a population exhibiting rapid expansion, the pairwise differences generate a smooth, wave-shaped distribution, whereas in a population of constant size, the distribution is ragged (Rogers and Harpending, 1992; Rogers and Jorde, 1995; Harpending and Rogers, 2000). The shape and position of the wave generated by pairwise differences (also referred to as a mismatch distribution) can supply information regarding the timing and scale of past population expansion events (Rogers and Jorde, 1995; Harpending and Rogers, 2000). Positive selection can however generate a similar distribution, and it is therefore difficult to distinguish between population expansion events, and the effect of selection based on mismatch distributions alone (Slatkin and Hudson, 1991). Analyses of pairwise mtDNA differences in globally diverse populations have generated estimated dates for population expansion events of 23,000 years before present (YBP) in Asia, 52,000 YBP in Europe, and 99,000 YBP in Africa (Harpending *et al.*, 1993; Sherry *et al.*, 1994; Rogers and Jorde, 1995; Jorde *et al.*, 1998; Harpending and Rogers, 2000; Jorde *et al.*, 2001).

Population substructure is another aspect that affects genetic variation. In terms of population substructure, there are three levels of complexity, namely the individual, the sub-populations, and the total population (Hartl, 1988). Phylogenetic analyses of mtDNA haplotypes indicate that African populations contain highly divergent lineages that imply a long history of population subdivision (Chen *et al.*, 1995; Watson *et al.*, 1997; Chen *et al.*, 2000). Several studies of mtDNA variation, using the F_{ST} measure, indicate higher levels of subdivision in populations from Africa than in populations from other geographic regions (Jorde *et al.*, 2000). In contrast, investigations of African populations, based on nuclear microsatellites, nuclear restriction fragment length polymorphisms (RFLPs), or protein polymorphisms do not consistently indicate high levels of subdivision (Jorde *et al.*, 2000). Overall, the pattern of genetic diversity observed within Africa indicates that populations have maintained a large and subdivided population structure over an extended period of time (Tishkoff and Williams, 2002). This subdivision is likely the result of linguistic and ethnic barriers, in addition to climatic, ecological and geographic factors, which could have played a role in population expansions, reductions, fragmentations and extinctions (Tishkoff and Williams, 2002).

2.5.5 Selection and the mitochondrial genome

Mitochondrial variation implies the simultaneous existence of different mtDNA types within a population. These different mtDNA genome types can be subdivided into groups, referred to as haplogroups, based on a characteristic pattern of variation which occurs within the specific haplogroup (refer to Section 3.3). The observed frequency of these different mtDNA types is expected to be similar due to the fact that under the assumption of neutral selection, the survival of each type of mtDNA should be equal.

The neutral theory of molecular evolution (Kimura, 1968; King and Jukes, 1969) assumes that the majority of mutations that can contribute to genetic variation between and within populations accumulate as a result of neutral processes, i.e. high levels of genetic variation may be attributed to mutations that have no effect on the fitness of the population due to the fact that these mutations are selectively neutral (Kimura, 1968). Under this neutral model of molecular evolution it is expected that mutations are lost or fixed in populations through the process of random genetic drift and that the evolution rate is dependent on the mutation rate. Thus, over time, populations may become phenotypically and genetically differentiated due to the process of random genetic drift coupled with geographic isolation. If the neutral theory is applicable to mtDNA variation, the observed polymorphisms or different mitochondrial haplotypes may have no particular significance in adaptation of humans to their present environment, and these selectively neutral polymorphisms can be viewed as “evolutionary noise” that provide no information about adaptive genetic changes in evolution (Hartl, 1988).

Extensive global population studies have however indicated that human mtDNA genomes exhibit striking regional variation, i.e. there are differences in the nature of the mtDNAs observed in different geographic regions (Wallace *et al.*, 1999). In addition, mitochondrial DNA frequencies have been studied in different populations and it was demonstrated that several Asian and Caucasian populations were highly homogenous, while African populations were more diverse and conformed to the neutral model of populations at equilibrium (Excoffier, 1990). The observed frequency distribution of mtDNA types or haplogroups could be due to founder effects where new geographic regions are colonised by only a few individuals that contribute a limited number of mtDNA types (Mishmar *et al.*, 2003), it could be the consequence of genetic drift (Wallace, 1995), or it could be attributed to the presence of unknown factors such as the effect of selection (Excoffier, 1990).

The process of selection provides an elegant explanation for the characteristic polymorphisms associated with the specific haplogroups (refer to Section 3.3). Two types of natural selection play a role in the evolution of mitochondrial genomes. Negative or purifying selection removes deleterious mutations from populations, and positive or adaptive selection increases the frequency of advantageous mutations within populations (Meiklejohn *et al.*, 2007). The influence of selection on any polymorphism within the mtDNA genome can not be limited to a single polymorphism, since in the absence of recombination, selection on a single polymorphism will bring about the concomitant fixation of all variants within that genome through the process of genetic hitchhiking (Smith and Haigh, 1974). This most likely occurs if an advantage is conferred by a non-synonymous mutation in a protein-coding gene. Alternatively, selection against deleterious mutations could reduce variation within the mtDNA genome (Charlesworth *et al.*, 1993; Charlesworth, 1994; Charlesworth *et al.*, 1995). The fact that the mitochondrial encoded polypeptides evolve much slower than the rate predicted by the high mutation rate (Saccone *et al.*, 1999), indicates that purifying selection is an important component of mitochondrial evolution.

2.5.5.1 Statistical tests of selection

Several statistical tests have been used to identify the effect of selection on genetic variation within a sample of nucleotide sequences (refer to Section 4.14.4). Tajima's D test statistic (Tajima, 1989) and Fu and Li's D^* and F^* test statistics (Fu and Li, 1993) test for a distortion in the frequency distribution of polymorphisms in datasets consisting of nucleotide sequences. The methods are performed under the assumption of neutrality, as well as the absence of recombination, and require only DNA polymorphism data, i.e. the genetic variation within a population at the DNA level.

Tajima's D test is based on the difference between the number of segregating sites (S) and the average number of nucleotide differences (k). When evolution is neutral, S should be equal to k and the value of D therefore should equal zero. A negative D value indicates the presence of positive selection, while a positive D value suggests the influence of balancing selection. The values determined for D can be influenced by factors such as sample size, mutation rate and effective population size.

Fu and Li's D^* and F^* tests are based on the distribution of mutations in the genealogy of a random sample of nucleotide sequences from a population. Fu and Li's D^* test is based

on the difference between the number of mutations appearing only once among the nucleotide sequences (η_s) and the total number of mutations (η), while Fu and Li's F^* test is based on the difference between the number of mutations appearing only once among the nucleotide sequences (η_s) and the average number of pairwise nucleotide differences (k) between the nucleotide sequences. A negative D^* value indicates an excess of mutations in external branches of the genealogy, thereby suggesting the presence of positive selection or negative selection, while a positive D^* value signifies a deficiency of mutations in external branches which suggests the influence of balancing selection (Fu and Li, 1993).

Significant deviations from neutrality as determined using the above statistical tests may indicate selection as discussed. However, it is possible that the positive or negative values may also be the result of a population subdivision or population expansion event, respectively (Excoffier, 2002). In addition, it is important to note that in the event that statistical tests lack sufficient power, it is possible to reject the hypothesis of selection in favour of the null hypothesis of neutrality due to the effect on variation caused by selection and other evolutionary processes (Braverman *et al.*, 1995; Simonsen *et al.*, 1995).

The effect of selection on mitochondrial DNA sequences in phylogenetic networks can be determined as described in Section 4.14.6 by calculating the ratio of non-synonymous substitutions compared to synonymous substitutions for internal and terminal branches of the phylogenetic network (Ruiz-Pesini *et al.*, 2004). Neutral alterations, such as synonymous substitutions, are expected to be uniformly distributed throughout the mtDNA network. Non-synonymous substitutions however can occur according to two possible scenarios. If non-synonymous substitutions cause deleterious amino acid replacements within the encoded mitochondrial proteins, the non-synonymous substitutions will be more frequent in terminal branches than in internal branches as this type of alteration is excluded from the population over time by the process of purifying or negative selection. If the non-synonymous substitutions are however advantageous and were retained through the process of adaptation to a specific environmental influence, it is predicted that these non-synonymous substitutions will be more common in internal branches compared to terminal branches. The frequency of non-synonymous alterations in different sections of the mtDNA network can therefore be normalised for differing time intervals by dividing by synonymous substitutions. This allows for a comparison to be made between the frequency of replacement mutations for internal branches (RF_i) and terminal branches

(RF_T). The ratio of non-synonymous/synonymous substitutions of internal compared to terminal branches can therefore be used as a measure of adaptive compared to purifying selection within a population of mtDNA sequences (Ruiz-Pesini *et al.*, 2004).

2.5.5.2 Adaptive selection and mitochondrial genome variation

Mitochondrial genes are subject to different functional constraints (Pesole *et al.*, 1999; Tourasse and Li, 2000) and it is therefore expected that the genes will evolve at different rates. However, within all lineages of a phylogeny, the rank order of these rates should remain the same (Rand, 2001). Studies investigating variation in the mtDNA genome have proposed that natural selection has influenced the course of human mtDNA evolution (Cann *et al.*, 1984; Nachman *et al.*, 1996; Ingman and Gyllensten, 2001; Mishmar *et al.*, 2003; Moilanen and Majamaa, 2003; Moilanen *et al.*, 2003; Elson *et al.*, 2004; Ruiz-Pesini *et al.*, 2004). Disagreement however exists between the different studies concerning whether the distribution of particular human mtDNA haplogroups is due to the effect of selection on specific lineages during the spread of humans from Africa and subsequent adaptation to different climates, or if their distribution is a function of random genetic drift in conjunction with the removal of non-synonymous changes by purifying selection.

A study by Mishmar and colleagues (2003) suggested that the observed regional human mitochondrial DNA variation is likely to have been shaped by the influence of natural selection with climate as one of the selective influences. By investigating the amino acid sequence variation of the mtDNA genes involved in OXPHOS, it was determined that the adenosine triphosphatase (ATPase) 6 gene was the most variable in mitochondrial DNA genomes from arctic zones, the cytochrome *b* (Cytb) gene was the most variable in mitochondrial DNA genomes from temperate zones, while the cytochrome *c* oxidase I (COI) gene was the most variable in mitochondrial DNA genomes from tropical zones. The correlation between increased amino acid sequence variation and climatic zone in these particular mtDNA genes, implied that environmental factors enriched for certain mitochondrial lineages as humans moved to the more northern latitudes, thereby shaping regional human mitochondrial DNA variation. It was suggested that the variants of the ATPase6 gene in the arctic zone could reduce coupling efficiency which would decrease ATP production and increase heat production. Partial uncoupling might also play a role in the increased basal metabolic rate observed in human populations in arctic regions.

In another study by Ruiz-Pesini and colleagues (2004), the observations supported the hypothesis that specific ancient mtDNA variants allowed humans to adapt to colder climates, which resulted in the enrichment of certain mtDNA lineages or haplogroups in different regions. Ruiz-Pesini *et al.*, (2004) identified specific adaptive mutations associated with the haplogroups present in arctic regions. The mutations that were identified are indicated in Table 2.8.

Table 2.8 Adaptive mitochondrial mutations associated with specific haplogroups

Haplogroup	Mutation	Gene	Haplogroup	Mutation	Gene
A	A4824G	ND2	J1	T14798C	Cytb
A	C8794T	ATPase6	J2	G15257A	Cytb
C	G11969A	ND4	K	T14798C	Cytb
C	T15240C	Cytb	T	A4917G	ND2

ND2 = NADH dehydrogenase subunit 2; ATPase6 = adenosine triphosphatase subunit 6; ND4 = NADH dehydrogenase subunit 4; Cytb = cytochrome *b*. A = adenine; C = cytosine; G = guanine; T = thymine. Adapted from Ruiz-Pesini *et al.*, (2004).

The cytochrome *b* mutations present in haplogroups J1, J2 and K affect the CoQ binding sites within cytochrome *b*. These sites are hypothesised to be critical for the complex III Q cycle and could therefore decrease proton pumping and thus efficiency of coupling (Ruiz-Pesini *et al.*, 2004). Haplogroups J and K provide protection from aging and neurodegenerative diseases, although they present with increased propensity to energy deficiency diseases (Ruiz-Pesini *et al.*, 2004). This discrepancy is however in agreement with the expectations for mtDNA mutations affecting coupling efficiency (Ruiz-Pesini *et al.*, 2004). As people migrated out of Africa into temperate and arctic Eurasia, adaptive selection therefore played an increasingly important role to generate more heat in colder climates by uncoupling the heat production from ATP production during mitochondrial oxidative phosphorylation.

However, the above claims regarding adaptation to climate is based mainly on statistical analyses of the distribution of non-synonymous substitutions compared to synonymous substitutions in a phylogenetic tree based on complete mtDNA sequences from the various haplogroups. To determine if there truly has been some functional change in oxidative phosphorylation in individuals containing these haplogroups, it is necessary to perform detailed biochemical analyses (Pakendorf and Stoneking, 2005).

In contrast, Elson and colleagues (2004) were not able to reproduce the results obtained by Mishmar and colleagues in 2003. Elson and colleagues (2004) observed the influence

of mainly negative selection on the 13 protein-coding genes of the mtDNA genome. ATPase6 was however an exception as it exhibited signs of balancing or positive selection in Europeans and Asians, an observation that is in agreement with the findings of Mishmar and colleagues in 2003. The observation made for ATPase6 could however also be explained by the relaxation of negative selection in European and Asian populations (Elson *et al.*, 2004).

At present, the general consensus is that purifying selection (negative selection) is the main force acting on human mtDNA variation within the coding region of the genome (Elson *et al.*, 2004; Pakendorf and Stoneking, 2005), while it is assumed that the evolution of the control region has not been influenced by selection (Elson *et al.*, 2004). However, it must also be taken into account that other processes such as recent population expansions or population bottlenecks could contribute to the effect of selection and it is therefore difficult to separate the contribution of these processes from the role of selection (Elson *et al.*, 2004).

It has been suggested that in different lineages, particular regions of mtDNA have experienced different selection histories, i.e. the influence of selection varies between the different genes, functional domains and nucleotide sites (Moilanen and Majamaa, 2003). This could possibly be due to lineage-specific interactions between gene products, or the effect of environmental factors such as climate, which led to differences that are haplogroup-specific (Moilanen *et al.*, 2003). It has also been observed that there are differences between sub-haplogroup-lineages that belong to the same lineage. Torroni *et al.*, (2001) investigated a phylogeny of complete L2 sequences and demonstrated that mtDNAs from L2b and L2d appeared to be disproportionately derived, compared with that of L2a and L2c. They concluded that this result did not agree with a simple model of neutral evolution with a uniform molecular clock, and that the observed pattern of non-synonymous versus synonymous substitutions suggested the influence of selection. The clinical implication of these differences between lineages and sub-haplogroup-lineages is significant since the pathogenicity of a mutation may depend greatly on the presence of other interacting mutations (Moilanen *et al.*, 2003), i.e. the haplogroup/sub-haplogroup-sub-haplogroup background on which it is presented (refer to Section 3.4).

If the neutral model of evolution does not apply to human mtDNA, as is suggested by numerous investigations, data on human evolution and population movements will have to

be re-evaluated using a more appropriate non-neutral model (Elson *et al.*, 2004). This model will have to take into account the potential role of selection in the fixation of non-synonymous nucleotide substitutions at different sites within the mtDNA genome (Kivisild *et al.*, 2006). The phylogenetic approach for analysing mtDNA sequence data within a species however remains viable due to the fact that the construction of the major branches is sound and the surplus of non-synonymous substitutions, indicating the role of selection, for the most part affects the terminal branches of the phylogenetic tree (Kivisild *et al.*, 2006). In addition, the fundamental observations of mtDNA studies regarding the global dispersal of modern humans subsequent to a recent initial migration out of Africa, are not affected by non-neutral evolution and are also supported by data generated for other loci (Excoffier, 2002; Cavalli-Sforza and Feldman, 2003).

2.6 MITOCHONDRIAL DNA VARIATION AND HUMAN DISEASE

Mitochondrial diseases are extremely diverse in their manifestation, affecting many different organs, and presenting with numerous symptoms of variable severity (Scheffler, 2001). Mitochondrial diseases, which are among the most common genetic disorders, are a major burden to society (DiMauro and Davidzon, 2005). Mitochondrial disorders can be caused by disruptions in nuclear genes encoding mitochondrial proteins (nDNA), mitochondrial protein-coding genes (mtDNA), or a combination of mutations in genes from both distinct genetic systems (Chinnery and Schon, 2003). Based on the combined data from investigations of children and adults, and when mutations in both nuclear and mitochondrial DNA are considered, the minimum prevalence of mitochondrial disorders is at least 1 in 5,000 (Schaefer *et al.*, 2004). A summary of symptoms and signs described in the six most common mitochondrial diseases is presented in Table 2.9.

Table 2.9 Clinical features in mitochondrial diseases associated with mtDNA mutations

Tissue	Symptom/sign	Δ -mtDNA		tRNA		ATPase	
		KSS	PS	MERRF	MELAS	NARP	MILS
Blood	Sideroblastic anaemia	±	+	-	-	-	-
Cardiac	Conduction block	+	-	-	±	-	-
	Cardiomyopathy	±	-	-	±	-	±

Table 2.9 Continued...

Tissue	Symptom/sign	Δ -mtDNA		tRNA		ATPase	
		KSS	PS	MERRF	MELAS	NARP	MILS
CNS	Seizures	-	-	+	+	-	+
	Ataxia	+	-	+	+	+	±
	Myoclonus	-	-	+	±	-	-
	Psychomotor retardation	-	-	-	-	-	+
	Psychomotor regression	+	-	±	+	-	-
	Hemiparesis/hemianopia	-	-	-	+	-	-
	Cortical blindness	-	-	-	+	-	-
	Migraine-like headaches	-	-	-	+	-	-
	Dystonia	-	-	-	+	-	+
Endocrine	Diabetes mellitus	±	-	-	±	-	-
	Short stature	+	-	+	+	-	-
	Hypoparathyroidism	±	-	-	-	-	-
ENT	Sensorineural hearing loss	-	-	+	+	±	-
Eye	Pigmentary retinopathy	+	-	-	-	+	±
	Optic atrophy	-	-	-	-	±	±
	Cataracts	-	-	-	-	-	-
GI	Exocrine pancreatic dysfunction	±	+	-	-	-	-
	Intestinal pseudo-obstruction	-	-	-	-	-	-
Muscle	Weakness	+	-	+	+	+	+
	Ophthalmoplegia	+	±	-	-	-	-
	Ptosis	+	-	-	-	-	-
PNS	Peripheral neuropathy	±	-	±	±	+	-
Renal	Fanconi's syndrome	±	±	-	±	-	-
Inheritance	Maternal	-	-	+	+	+	+
	Sporadic	+	+	-	-	-	-
Laboratory	Lactic acidosis	+	+	+	+	-	±
	Muscle biopsy: RRF	+	±	+	+	-	-

Red boxes (■) indicate typical features of specific syndromes (excluding Leigh syndrome, which is defined by the neuropathological or neuroradiological alterations). Δ -mtDNA = deleted mitochondrial DNA; tRNA = transfer RNA; ATPase = adenosine triphosphatase subunit; KSS = Kearns-Sayre syndrome; PS = Pearson syndrome; MERRF = myoclonus epilepsy with ragged-red fibres; MELAS = mitochondrial encephalomyopathy, lactic acidosis, and stroke-like episodes; NARP = neuropathy, ataxia, and retinitis pigmentosa; MILS = maternally inherited Leigh syndrome; CNS = central nervous system; ENT = ear-nose-throat; GI = gastrointestinal system; PNS = peripheral nervous system; RRF = ragged-red fibres. Adapted from DiMauro and Davidzon (2005).

Mitochondrial diseases can be caused by mitochondrial DNA rearrangements, such as in Pearson syndrome (PS), Kearns-Sayre syndrome (KSS) and progressive external ophthalmoplegia (PEO), by point mutations in tRNA or rRNA genes involved in protein synthesis, namely myoclonus epilepsy with ragged-red fibres (MERRF), and mitochondrial encephalomyopathy, lactic acidosis, and stroke-like episodes (MELAS), or by point mutations in protein-coding genes, namely neuropathy, ataxia, and retinitis pigmentosa (NARP), and maternally inherited Leigh syndrome (MILS). Disorders caused by mtDNA deletions such as KSS, PEO and PS are usually sporadic, i.e. mtDNA deletions are not

maternally inherited or transmitted to progeny. The genetic classification of human mitochondrial disorders caused by the disruption of mitochondrial genes is presented in Table 2.10.

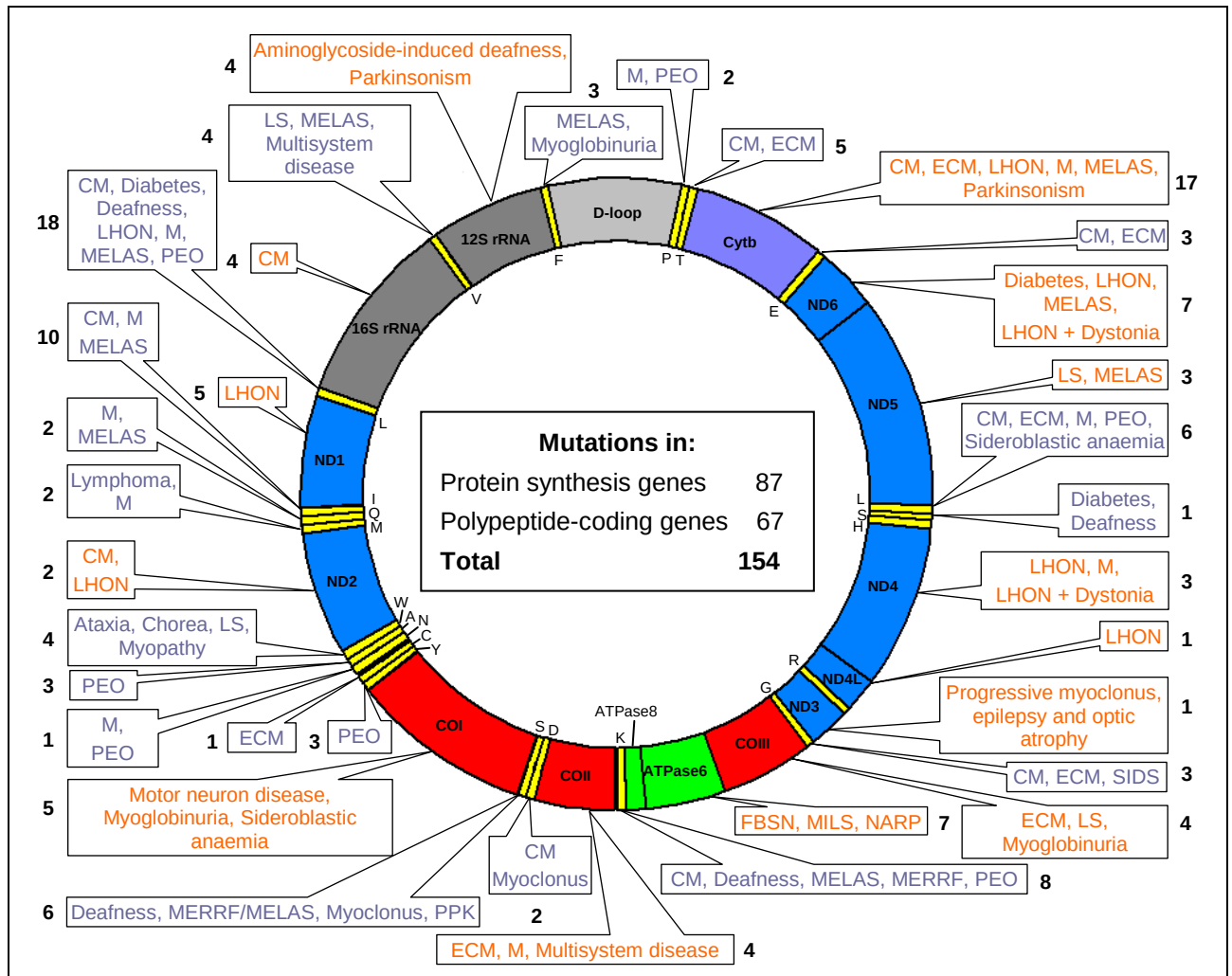
Table 2.10 The genetic classification of human mitochondrial disorders

<i>Rearrangements (large scale partial deletions and duplications)</i>		
Type of disorder	Mutation	Pattern of inheritance
CPEO	---	Sporadic
Diabetes and deafness	---	Sporadic
KSS	---	Sporadic
Pearson marrow-pancreas syndrome	---	Sporadic
Sporadic tubulopathy	---	Sporadic
<i>Point mutations</i>		
Type of disorder	Mutation	Pattern of inheritance
Protein-coding genes		
LHON	G3460A, G11778A, T14484C, G14459A, G15257A	Maternal
NARP/Leigh syndrome	T8993G/C	Maternal
tRNA genes		
Cardiomyopathy	A3243G, A4269G, A4300G	Maternal
CPEO	A3243G, T4274C	Maternal
Diabetes and deafness	A3243G, C12258A	Maternal
Encephalomyopathy	G1606A, T10010C	Maternal
MELAS	A3243G, T3271C, A3251G	Maternal
MERRF	C3256T, A8344G, T8356C	Maternal
Myopathy	T14709C, A12320G	Maternal
rRNA genes		
Aminoglycoside induced non-syndromic deafness	A1555G	Maternal
Non-syndromic sensorineural deafness	A7445G	Maternal

CPEO = chronic progressive external ophthalmoplegia; KSS = Kearns-Sayre syndrome; LHON = Leber's hereditary optic neuropathy; MELAS = mitochondrial encephalomyopathy, lactic acidosis, and stroke-like episodes; MERRF = myoclonus epilepsy with ragged-red fibres; NARP = neuropathy, ataxia, and retinitis pigmentosa; rRNA = ribosomal RNA; tRNA = transfer RNA. Adapted from Wallace (1995) and Chinnery and Schon (2003).

The association between mutations in mtDNA and human disease was first recognised in 1988 (Holt *et al.*, 1988; Wallace *et al.*, 1988). Since then, more than 150 different pathogenic point mutations and in excess of 100 mtDNA rearrangements, i.e. deletions and duplications, have been identified within the mitochondrial DNA genome (Wallace, 1995; Chinnery and Schon, 2003; DiMauro and Davidzon, 2005). These mutations are summarised in Figure 2.5.

Figure 2.5 Morbidity map of the human mitochondrial genome



The function of each gene is identified by the use of colour, the seven subunits of complex I in blue (■), the three subunits of complex III in red (■), the one subunit of complex IV in purple (■), the two subunits of complex V in green (■), the two ribosomal RNAs in grey (■) and the 22 transfer RNAs in yellow (■). Diseases caused by mutations that impair mitochondrial protein synthesis are indicated in blue-grey, and diseases due to mutations in protein-coding genes are designated in orange. The numbers in **bold** text indicate the number of mutations identified in the gene. Gene abbreviations: 12S rRNA = 12S ribosomal RNA; 16S rRNA = 16S ribosomal RNA; ATPase = ATP synthase F₀ subunit; CO = cytochrome c oxidase subunit; Cytb = cytochrome b; ND = NADH dehydrogenase subunit. The tRNA genes are indicated by the one letter code of the encoded amino acid. Amino acid abbreviations: A = alanine; C = cysteine; D = aspartic acid; E = glutamic acid; F = phenylalanine; G = glycine; H = histidine; I = isoleucine; K = lysine; L = leucine; M = methionine; N = asparagine; P = proline, Q = glutamine; R = arginine; S = serine; T = threonine; V = valine; W = tryptophan; Y = tyrosine. Other abbreviations: CM = cardiomyopathy; D-loop = displacement loop; ECM = encephalomyopathy; FBSN = familial bilateral striatal necrosis; KSS = Kearns-Sayre syndrome; LHON = Leber's hereditary optic neuropathy; LS = Leigh syndrome; M = myopathy; MELAS = mitochondrial encephalomyopathy, lactic acidosis, and stroke-like episodes; MERRF = myoclonic epilepsy with ragged-red fibres; MILS = maternally inherited Leigh syndrome; NARP = neuropathy, ataxia, and retinitis pigmentosa; PEO = progressive external ophthalmoplegia; PPK = palmoplantar keratoderma; SIDS = sudden infant death syndrome. Adapted from DiMauro and Davidzon (2005).

Mutations may become fixed in a population through providing some selective advantage to the individuals carrying the specific mutation. If the specific mutation is neutral, it may still become fixed through the process of genetic drift. Mutations may however also be deleterious, as discussed in this section, and result in genetic disorders. Mildly deleterious mutations that affect the phenotype and fitness of an individual due to the late onset of a disorder may become fixed within the population as a polymorphism, since clinical symptoms usually only appear after reproductive age. In contrast, deleterious mutations that affect the phenotype and fitness of an individual early in life will be rapidly removed by

selection. Early onset deleterious mutations are therefore transient, thereby indicating that observed deleterious mutations have a recent origin (Wallace, 1995).

A novel mutation is only classified as pathogenic if it complies with a list of criteria (Chinnery and Schon, 2003; DiMauro and Davidzon, 2005). Firstly, the mutation must not be a recognised polymorphism, i.e. it should not be observed in unaffected individuals of the same ethnic group. Secondly, the mutation should change a site that is evolutionarily conserved, i.e. a functionally important site, and the mutation must be located within a region that is recognised to have functional importance e.g. within a tRNA gene, an rRNA gene or a protein-coding gene. Thirdly, pathogenic mutations are usually heteroplasmic which suggests that the mutation arose recently and sufficient time has not yet passed for it to have become fixed in the lineage, or that selection acting on the organism level has prevented fixation (refer to previous paragraph). Fourthly, a correlation should be present between the level of heteroplasmy and the clinical severity, as well as the cell pathology. Fifthly, a correlation should be observed between the presence of the mutation and the biochemical profile. The mutation should cause single or multiple respiratory chain deficiencies in affected tissues, or defects of mitochondrial protein synthesis and respiration that can be confirmed through the use of cybrid cell lines. A selection of previously identified pathogenic mtDNA mutations are listed in Table 2.11.

Table 2.11 A selection of previously identified pathogenic mtDNA mutations

Mutation	Syndrome	Reference
3,902 to 3,908: 7 bp inversion	Exercise intolerance	Musumeci <i>et al.</i> , 2000
5132delAA	Exercise intolerance	Schwartz and Vissing, 2002
G5920A	Exercise intolerance	Taivassalo <i>et al.</i> , 2001
T7671A	Exercise intolerance	Rahman <i>et al.</i> , 1999
9,487 to 9,501: 15 bp deletion	Exercise intolerance	Keightley <i>et al.</i> , 1996
G11832A	Exercise intolerance	Andreu <i>et al.</i> , 1999c
G12147A	Headache, seizures, severe cerebral oedema	Melone <i>et al.</i> , 2004
A12770G	MELAS	Liolitsa <i>et al.</i> , 2003
G13042A	MELAS/MERRF	Naini <i>et al.</i> , 2005
A13045C	MELAS/LHON/LS	Liolitsa <i>et al.</i> , 2003
A13084T	MELAS/LS	Mitchell <i>et al.</i> , 2006
G13513A	MELAS	Santorelli <i>et al.</i> , 1997
	MELAS/LHON	Pulkes <i>et al.</i> , 1999
	LS	Kirby <i>et al.</i> , 2003
A13514G	MELAS	Corona <i>et al.</i> , 2001
G14846A	Exercise intolerance	Andreu <i>et al.</i> , 1999b

Table 2.11 Continued...

Mutation	Syndrome	Reference
G15059A	Exercise intolerance	Andreu <i>et al.</i> , 1999a Andreu <i>et al.</i> , 1999b
G15084A	Exercise intolerance	Andreu <i>et al.</i> , 1999b
G15150A	Exercise intolerance	Legros <i>et al.</i> , 2001
G15168A	Exercise intolerance	Andreu <i>et al.</i> , 1999b
T15197C	Exercise intolerance	Legros <i>et al.</i> , 2001
G15242A	Exercise intolerance	Keightley <i>et al.</i> , 2000
15,498 to 15,521: 24 bp deletion	Exercise intolerance	Andreu <i>et al.</i> , 1999b
G15615A	Exercise intolerance	Dumoulin <i>et al.</i> , 1996
G15723A	Exercise intolerance	Andreu <i>et al.</i> , 1999b
G15761A	Exercise intolerance	Mancuso <i>et al.</i> , 2003
C15800T	Exercise intolerance	Lamantea <i>et al.</i> , 2002

A = adenine; C = cytosine; G = guanine; T = thymine; bp = base pairs; del = deletion; LHON = Leber's hereditary optic neuropathy; LS = Leigh syndrome; MELAS = mitochondrial encephalomyopathy, lactic acidosis, and stroke-like episodes; MERRF = myoclonus epilepsy with ragged-red fibres. Adapted from DiMauro and Davidzon (2005).

Principles relating to mitochondrial genetics, namely heteroplasmy and the threshold effect (refer to Section 2.5.2), maternal inheritance (refer to Section 2.4.1), relaxed replication (refer to Section 2.4.3), mitotic segregation, the effect of the mtDNA background (refer to Section 3.4), interaction with nuclear genes, and the environment becomes important in the discussion of mitochondrial DNA variation and human disease (Chinnery and Schon, 2003; DiMauro and Davidzon, 2005). When heteroplasmy is present, the fraction of mutated mtDNA genomes detected within a cell may vary significantly between different tissues or between cells of the same tissue. A minimum critical number of mutated mtDNA genomes must however be present before mitochondrial dysfunction is caused within a specific organ or tissue, and expression of the clinical phenotype characteristic of a mitochondrial disease occurs. This is referred to as the threshold effect. Different tissues have different threshold requirements for OXPHOS. Tissues or organs with high OXPHOS requirements such as the brain, eye, heart and skeletal muscle have relatively low thresholds and are therefore particularly vulnerable to mtDNA mutations (Wallace, 1992; DiMauro and Davidzon, 2005).

Relaxed replication refers to the fact that mtDNA replication is independent of the cell cycle. In a cell containing mutant and wild type mtDNA genomes, it is possible that these two types of mtDNA genomes replicate at different rates. This could bring about changes in the proportion of mutant mtDNA genomes present within a cell, which in turn could provide an explanation for the late onset and progression of several mtDNA disorders (Chinnery and Schon, 2003). Maternal inheritance implies that a mutation contained within

the mtDNA genome will be passed on to both female and male offspring of an affected mother, however, only daughters will in turn transmit the mtDNA mutation to their progeny. Mitotic segregation involves the random distribution of mtDNA genome copies to daughter cells during cell division. The proportion of mutant mtDNA genomes within a daughter cell could therefore increase or decrease compared to the proportion of mutant mtDNA genomes originally contained within the parent cell. This could, over time as the patient ages, lead to a shift from one clinical phenotype to another clinical phenotype (DiMauro and Davidzon, 2005).

The mtDNA background on which a mutation is expressed is described by referring to the mitochondrial haplogroup to which the mtDNA genome type belongs. Subdivision of mtDNA genome types into genetic groups, referred to as haplogroups, is based on a characteristic pattern of polymorphisms which occurs within the specific haplogroup as observed in an apparently healthy population (refer to Section 3.3). It has been observed that mtDNA mutations are preferentially associated with specific haplogroups, thereby indicating that disease expression can be influenced by the mitochondrial genetic background (refer to Section 3.4). In addition, the observed phenotype can also be attributed to nuclear-mitochondrial interactions, as varied clinical phenotypes for a number of mitochondrial diseases can not be explained by the primary identified defect (Chinnery and Schon, 2003). It has also been observed in heteroplasmic mice that particular mtDNA genomes were favoured in different tissues (Jenuth *et al.*, 1997). Three specific nuclear genetic loci have been identified that influence this process by controlling the selection of mtDNA genomes at the level of the DNA molecule (Battersby *et al.*, 2003). It is proposed that in humans equivalent nuclear genes may influence the level of heteroplasmy in different organs and tissues, consequently modulating the clinical phenotype (Chinnery and Schon, 2003).

The diagnosis and the specific cause for the symptoms observed in mitochondrial diseases become more conclusive as more information becomes available regarding the causative mitochondrial mutations. This enables the formulation of improved predictions concerning the progression and management of the specific mitochondrial disease, and will permit genetic counsellors to offer more reliable assessments of risk in numerous cases. Eventually, it may become possible, through pharmacological interventions, to remedy the diseases (Scheffler, 2001).

CHAPTER THREE

Mitochondrial DNA Variation and Human Evolution

The origin of the human species has fascinated man for many years. In order to unravel the evolutionary path of humans, a repertoire of different approaches has been followed of which the most recent addition is the use of molecular techniques (Pakendorf and Stoneking, 2005). The mitochondrion is a relatively small organelle within the eukaryotic cell where oxidative metabolism takes place (Voet and Voet, 1995). It contains its own genome, which has become the focus of various molecular evolutionary studies. When comparing the DNA sequences generated for this genome between different population groups, variation is observed which allows us to elucidate human population history (Wallace, 1995).

3.1 HUMAN ORIGINS AND MIGRATIONS

The location of the birthplace of modern humans and the subsequent peopling of the world has been studied extensively for many years through the use of archaeological methods, and more recently, through the use of molecular genetic approaches (Oppenheimer, 2003). A number of evolutionary models have also been proposed to clarify the processes involved in the emergence of modern man and the colonisation of the different continents (Excoffier, 2002). The two main models, as well as the different migrations of humans from and within Africa, are discussed below.

3.1.1 Models of modern human origins

Two contrasting models have been proposed in order to explain modern human origins and the subsequent replacement of archaic humans by anatomically modern humans, namely the Recent African Origin (RAO) model (Cann *et al.*, 1987) and the Multiregional Evolution (ME) model (Wolpoff *et al.*, 2000). These models have been adapted and expanded to give rise to other models which incorporate the possibility of varying amounts of gene flow between archaic human populations.

The RAO model, also referred to as the “Out of Africa” or replacement model, suggests that all non-African populations are derived from *Homo sapiens* ancestors that evolved in Africa ca. 100,000 to 200,000 years before present (YBP). The descendants of these early ancestors subsequently spread all over the world and replaced archaic *Homo* populations as they were encountered. Based on this model, it is expected that all genetic lineages will be derived from a recent common African ancestor and that the genetic variation present in non-African populations will be a subset of that observed in modern African populations (Cann *et al.*, 1987). The RAO model is supported by the fossil record, as well as by archaeological data. The earliest modern human fossils, dating to 90,000 to 120,000 YBP, were discovered in Africa and the Middle East (Leakey, 1994; Lahr and Foley, 1998), while the earliest evidence for modern human behaviour in eastern and southern Africa is dated to 70,000 to 90,000 YBP (McBrearty and Brooks, 2000; Henshilwood *et al.*, 2002).

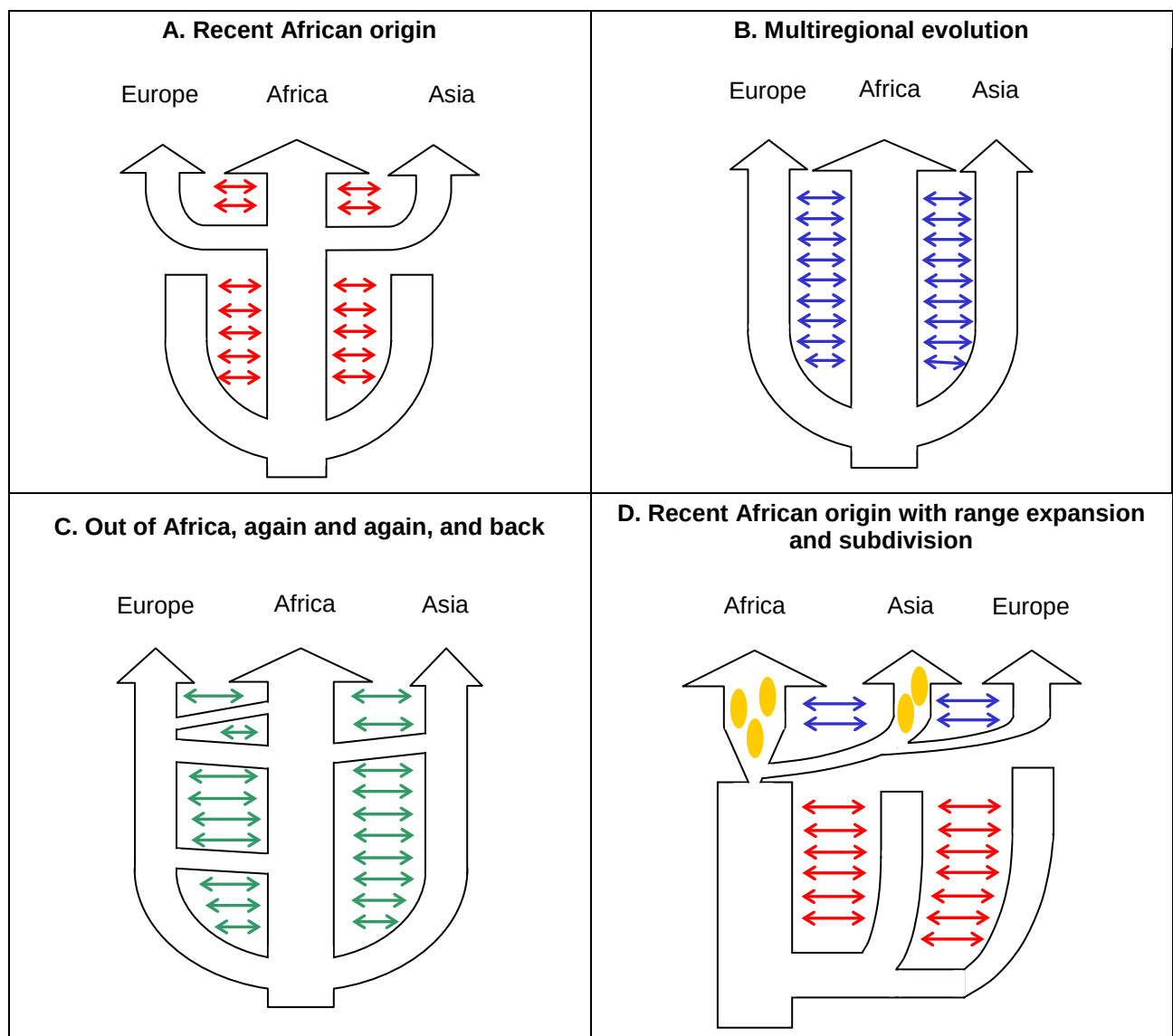
Alternatively, the ME model proposes that there was no single geographical origin for modern humans. Following the radiation of *Homo erectus* from Africa into Europe and Asia ca. 800,000 to 1.8 million YBP, regional populations independently evolved from *Homo erectus* to *Homo sapiens*. The concurrent evolution from *Homo erectus* to *Homo sapiens* in these scattered populations could have been accomplished through extensive gene flow between populations. A large effective population size is however required to maintain gene flow among such geographically diverse populations. The ME model is supported mainly by the continuity of certain morphological traits in the fossil record, which indicate that modern populations evolved over very long periods of time in the regions where they are currently located (Wolpoff *et al.*, 1988).

A modification of the ME model, the assimilation (or hybridisation) model, suggests that gene flow between the early human populations was not equal over time and space. This model allows for the occurrence of gene flow between modern humans that migrated from Africa and archaic populations outside Africa. This implies that the evolution of modern humans could have been due to the merging of modern traits derived from African populations with local features observed in archaic Eurasian populations. Based on this model, it is predicted that the modern gene pool is derived from the variable contribution of genes from both archaic African and non-African populations (Smith *et al.*, 1989).

A schematic representation of the different models of human evolution is presented in Figure 3.1. In Figure 3.1C, the RAO model has been modified to include a succession of large migrations out of Africa, a recent important migration from Asia back to Africa, and

ongoing but limited gene flow between continents. Another adaptation of the RAO model is presented in Figure 3.1D. Under this hypothesis, it is proposed that modern humans emerged from one of the existing subdivisions in Africa, that the spread to Asia occurred after a bottleneck event, and that western Asia was the source from which Europe was colonised. This hypothesis is different from the simple RAO model since it emphasises the importance of internal and dynamic subdivisions within continents, and the progressive colonisation of the world by a series of range expansions which commenced from locations that were geographically restricted (Excoffier, 2002).

Figure 3.1 Different models of human evolution



A = Recent African origin; B = Multiregional evolution; C = Out of Africa, again and again, and back; D = Recent African origin with range expansion and population subdivision within continents. The black open arrows indicate the direction of migration, while the size of the arrows indicates the relative size of the migrating population. Population bottlenecks are indicated by the smaller arrows leading to Europe and Asia in model A, while the arrows without arrowheads in models A and B indicate that populations in Asia and Europe did not migrate further and was subsequently replaced by populations migrating from Africa. Red arrows indicate putative gene flow that could have occurred between *Homo erectus* populations from different continents, blue arrows indicate ongoing gene flow between continents, green arrows represent hypothetical genetic exchanges between continents and yellow ovals indicate dynamic subdivisions within continents. Adapted from Excoffier (2002).

At present, based on both fossil and genetic evidence, the general consensus regarding models of modern human origin is that modern humans originated in Africa. Supporters of the ME model generally acknowledge the dominant role that Africa has played in modern human origins and in light of recent genetic evidence and the accumulation of fossil evidence in Africa which provides compelling support for an African origin, the debate has shifted focus to determining if the origin of modern humans was entirely African, i.e. did modern humans migrating from Africa completely replace archaic human populations without interbreeding, or mainly African, i.e. can a proportion of our genetic ancestry be traced to Neanderthals or other archaic human populations due to interbreeding of modern humans with resident archaic human populations (Templeton, 2002; Stoneking, 2006; Relethford, 2008).

3.1.2 Support for the Recent African Origin model

The RAO model is supported by the fossil record. The fossil record indicates that the ancestors of anatomically modern humans were present in Africa for ca. six million years (Relethford, 2008). Over time, due to factors such as fluctuations in climate, small populations of archaic humans ventured repeatedly out of Africa into eastern Europe and south-east Asia, establishing communities of archaic humans which survived until 27,000 to 54,000 YBP (Relethford, 2008). These archaic humans include the Neanderthals which are viewed by some anthropologists as a subspecies of *Homo sapiens* (*Homo sapiens neanderthalensis*), while others argue that they should be considered as a separate species (*Homo neanderthalensis*).

The majority of archaic humans however remained in Africa where they subsequently evolved to become anatomically modern humans. The exact location in Africa where anatomically modern humans first emerged is currently unknown, although the earliest fossils that exhibit modern characteristics were discovered in sub-Saharan Africa and the Middle East (Leakey, 1994; Lahr and Foley, 1998). At present, the majority of researchers favour a sub-Saharan origin for modern humans, likely in eastern or southern Africa based mainly on fossil evidence collected at a number of different sites in these regions (Leakey, 1994; White *et al.*, 2003; McDougall *et al.*, 2005), while the presence of anatomically modern human fossils in the Middle East is attributed to an initial failed exodus from Africa into Europe. It has been proposed that early modern humans travelled out of sub-Saharan Africa into North Africa and the Levant (a geographical term traditionally comprising modern Syria, Lebanon, Israel, Palestine, Jordan, i.e. the Near East, excluding Egypt).

Evidence for this initial exodus is provided by the finding of the earliest remains of modern humans outside Africa, dated between 90,000 to 120,000 YBP, in the Levant (Lahr and Foley, 1998). The current consensus view is that these first modern humans that ventured out of Africa into the Levant did not migrate further into Eurasia due to their extinction in this region as a result of the return of dry glacial conditions that caused extreme desert conditions in both the Levant and North Africa (Oppenheimer, 2003).

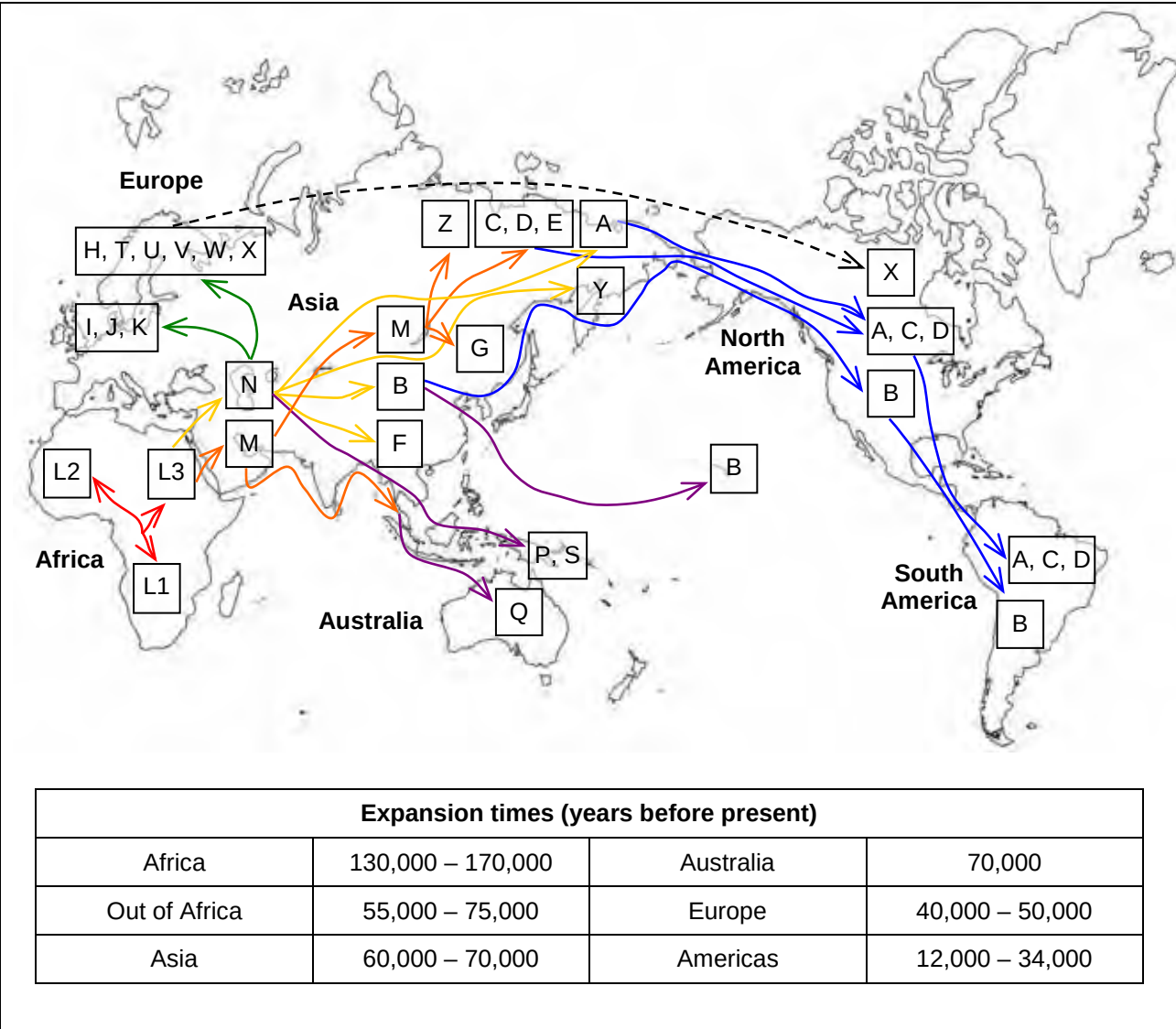
This initial failed exodus from Africa was followed by a successful migration that allowed anatomically modern humans to populate the rest of the world (Watson *et al.*, 1997; Oppenheimer, 2003). Two potential routes have been proposed for this exodus from Africa, namely a northern and a southern or coastal route. The route ultimately followed was dependent on the prevailing weather conditions, as this would have determined which route was the most accessible at the specific point in time. The northern route involved travelling via North Africa over the Sahara to the Levant and Europe, while the southern route entailed travelling from the Red Sea along the Indo-Pacific coast to Australia and beyond, including possible expansions into China, Japan and New Guinea (Lahr and Foley, 1998; Oppenheimer, 2003). At present, the southern route is favoured as it is supported by both genetic and archaeological data (Oppenheimer, 2003). Archaeological data indicate that subsequent to the first appearance of modern human fossils in the Middle East 90,000 to 130,000 YBP, modern humans did not re-emerge in that region or in Europe, until 40,000 to 50,000 YBP, while there is evidence of the presence of modern humans in Australia 60,000 YBP (Tishkoff and Verelli, 2003). Although the fossil record is undoubtedly useful for investigations of human migrations, it remains incomplete and it is therefore not advisable to rely on archaeological data only to confirm hypotheses.

The RAO model is also supported by genetic evidence. Genetic evidence that provides support for Africa as the point of origin for anatomically modern humans includes the fact that the highest mtDNA diversity is observed in Africa (Cann *et al.*, 1987; Vigilant *et al.*, 1991; Chen *et al.*, 1995), that African mtDNA sequences form the root in phylogenetic analyses of global mtDNA sequences (Penny *et al.*, 1995), and the results obtained in a variety of nuclear DNA studies which indicate that African populations contain the highest genetic diversity (Wainscoat *et al.*, 1986; Armour *et al.*, 1996; Tishkoff *et al.*, 1996).

Early human migrations and population expansion events can be deduced from the present day regional and continent-specific distribution of mtDNA haplogroups (Wallace *et al.*, 1999). These continent-specific mitochondrial haplogroups are defined by specific

restriction polymorphisms or by single nucleotide polymorphisms (SNPs) present within the non-coding and coding regions. The restriction enzyme sites defining continent-specific mtDNA haplogroups are indicated in the sections in which the haplogroups observed on the different continents are discussed (refer to Sections 3.3.1, 3.3.2, 3.3.3, 3.3.4 and 3.3.5) and the regional distribution and establishment of the major continent-specific mitochondrial haplogroups due to human migrations are illustrated in Figure 3.2.

Figure 3.2 Human migration pattern based on regional distribution of mitochondrial haplogroups



Coloured arrows are used to indicate the regional distribution of continent-specific haplogroups. Red arrows indicate African haplogroups, yellow and orange arrows indicate Asian haplogroups, green arrows indicate European haplogroups, blue arrows indicate Native American haplogroups, violet arrows indicate Oceanic haplogroups and the dotted arrow indicates the putative connection between haplogroup X observed in Europe and North America (refer to Section 3.3.3). Adapted from Wallace *et al.* (1999) and MITOMAP (2008).

Africa is home to the most diverse and therefore most ancient mtDNA haplotypes, and calculations based on the sequence divergence of African mitochondrial sequences, indicate that Africa was already occupied by 150,000 to 200,000 YBP (Wallace, 2008). Africans harbour the African-specific mtDNA haplogroups L0 (oldest), L1 and L2, as well as haplogroup L3 (youngest) which is also present in Asia (Excoffier and Langaney, 1989). Haplogroup L3 gave rise to macrohaplogroups M and N which arose in north-eastern Africa about 65,000 YBP, these two mtDNA lineages were the only lineages that successfully migrated from Africa to colonise Eurasia (Quintana-Murci *et al.*, 1999; Oppenheimer, 2003; Wallace, 2008). Macrohaplogroup M and its descendants are observed mainly in Asia, while it is generally absent from Europe and the Levant. In contrast, macrohaplogroup N and its descendants are observed in Europe, the Levant, Asia and Australia (Oppenheimer, 2003).

In Europe, macrohaplogroup N diverged to give rise to a large number of mtDNA haplogroups, namely haplogroups H, I, J, K, T, U, V, W and X. Haplogroups H, I, J, K, T, U, V and W are characteristic of individuals of European descent (Wallace *et al.*, 1999). Haplogroup X is present at low frequencies in North Africa, Europe, West Asia, Central Asia, and the New World, it is however not commonly observed in East Asia or Siberia (Reidla *et al.*, 2003).

In Asia, macrohaplogroups M and N contributed equally to the radiation of mtDNA with a large number of derivative mtDNA lineages being generated within south-eastern and central Asia. Macrohaplogroup M gave rise to haplogroups C, D, E, G, Q and Z, while macrohaplogroup N gave rise to haplogroups A, B, F, P, S and Y. Haplogroup B is found primarily along the Asian coast and out into the Pacific, while haplogroups P, Q and S are observed in Oceania. Only four of the southern Asian haplogroups, i.e. A, C, D and G are currently dominant in northern Siberia. From this group, only haplogroups A, C and D subsequently crossed the Bering land bridge to populate the Americas (Wallace *et al.*, 1999). Haplogroups G, Y and Z are observed primarily in Siberia (Schurr *et al.*, 1999). Based on the genetic evidence provided for mtDNA and Y chromosome analyses it has been suggested that modern humans colonised Asia first and subsequently colonised Europe and the Levant later from South Asia (Oppenheimer, 2003; Forster, 2004).

Studies based on human mtDNA variation have given rise to the “African Eve” hypothesis (Cann *et al.*, 1987; Vigilant *et al.*, 1991) which states that all extant mtDNA types are descended from a single ancestor living in Africa ca. 150,000 to 200,000 YBP

(Stoneking, 2006). Based on the analysis of mtDNA sequence data (Watson *et al.*, 1997; Quintana-Murci *et al.*, 1999; Chen *et al.*, 2000), Y chromosome diversity (Passarino *et al.*, 1998; Underhill *et al.*, 2000; Semino *et al.*, 2002) and nuclear markers (Tishkoff *et al.*, 1996; Tishkoff *et al.*, 2009), it has been suggested that migration out of Africa originated from East or north-east Africa. In addition, based on genetic evidence provided by the investigation of mitochondrial lineages, i.e. that one mitochondrial lineage, L3, gave rise to all non-African mitochondrial lineages, it is now generally accepted that a single exodus occurred from Africa (Watson *et al.*, 1997). This is also confirmed by the study of Y chromosome lineages where one African lineage, M168, gave rise to all non-African Y chromosome lineages (Underhill *et al.*, 2000), as well as other genetic markers passed down from both parents (Tishkoff *et al.*, 1996; Alonso and Armour, 2001).

Two important questions remain regarding the exodus of anatomically modern humans from Africa, namely when it occurred and why. Estimated dates for the primary split between African and non-African populations, and therefore the exodus of anatomically modern humans from Africa, range from 44,000 to 200,000 YBP (Tishkoff *et al.*, 1996; Stoneking *et al.*, 1997; Harris and Hey, 1999; Ingman *et al.*, 2000; Underhill *et al.*, 2000). Estimates based on the approximate age of haplogroup L3, the precursor of modern European and Asian haplogroups, has been determined by different groups as 60,000 to 80,000 YBP (Watson *et al.*, 1997) and 70,000 to 95,000 YBP (Oppenheimer, 2003).

A number of different reasons have been proposed for the exodus of anatomically modern humans from Africa. Based on the investigation of mtDNA mismatch distributions (Harpending *et al.*, 1993; Sherry *et al.*, 1994; Harpending and Rogers, 2000) and the study of the mitochondrial haplogroups in modern African populations, it has been estimated that anatomically modern humans experienced a major population expansion event ca. 60,000 to 80,000 YBP from a small population probably located in eastern or southern Africa (Watson *et al.*, 1997; Salas *et al.*, 2002; Forster, 2004; Forster and Matsumura, 2005; Macaulay *et al.*, 2005; Kivisild *et al.*, 2006). This population expansion event is attributed to behavioural changes which facilitated an increase in the carrying capacity of the environment which subsequently allowed expansion in both population number and density. These changes include major technological advances (a greater variety of tools better suited to different forms of hunting, gathering and processing of available food resources), the emergence of the use of symbols as evidenced by abstract art and personal ornamentation, and improved social organisation as demonstrated by increased trade between different human groups (Mellars, 2006). These behavioural changes that

facilitated population growth could be attributed to a change in cognitive capacity of the population, or a major shift in the adaptive and selective pressures to which human populations were subjected (Mellars, 2006). It has been suggested that the use of language could be viewed as the major cognitive change that preceded the migration of anatomically modern humans from Africa (Cavalli-Sforza, 2001), although it is unclear at this stage if the evolution of language was due to a dramatic change or a gradual process (Dediu, 2008). Alternatively, climatic and environmental change could have influenced the adaptive and selective pressures acting on human populations. It has been proposed that during the last glacial cycle beginning at ca. 115,000 YBP, Africa experienced profound and sudden (over several decades) changes in climate that were associated with episodic occurrences of Heinrich events (Carto *et al.*, 2009). Heinrich events are freshwater fluxes that are linked with surges of ice sheets into the North Atlantic. These events disrupt the thermohaline circulation which plays a vital role in the cross-equatorial transport of heat from low latitudes to northern Polar Regions, thus causing increased aridity due to changes in rainfall patterns leading to the desiccation of biogeographic resources (Carto *et al.*, 2009). It is proposed that these events, in conjunction with progressively increasing aridity linked with global cooling and the onset of northern hemisphere glaciation, could have quickly made large parts of North, West and East Africa unsuitable for human habitation (Carto *et al.*, 2009). These changes would have presented anatomically modern humans with no other choice than to migrate into more hospitable regions.

3.1.3 The Bantu¹ expansions in Africa

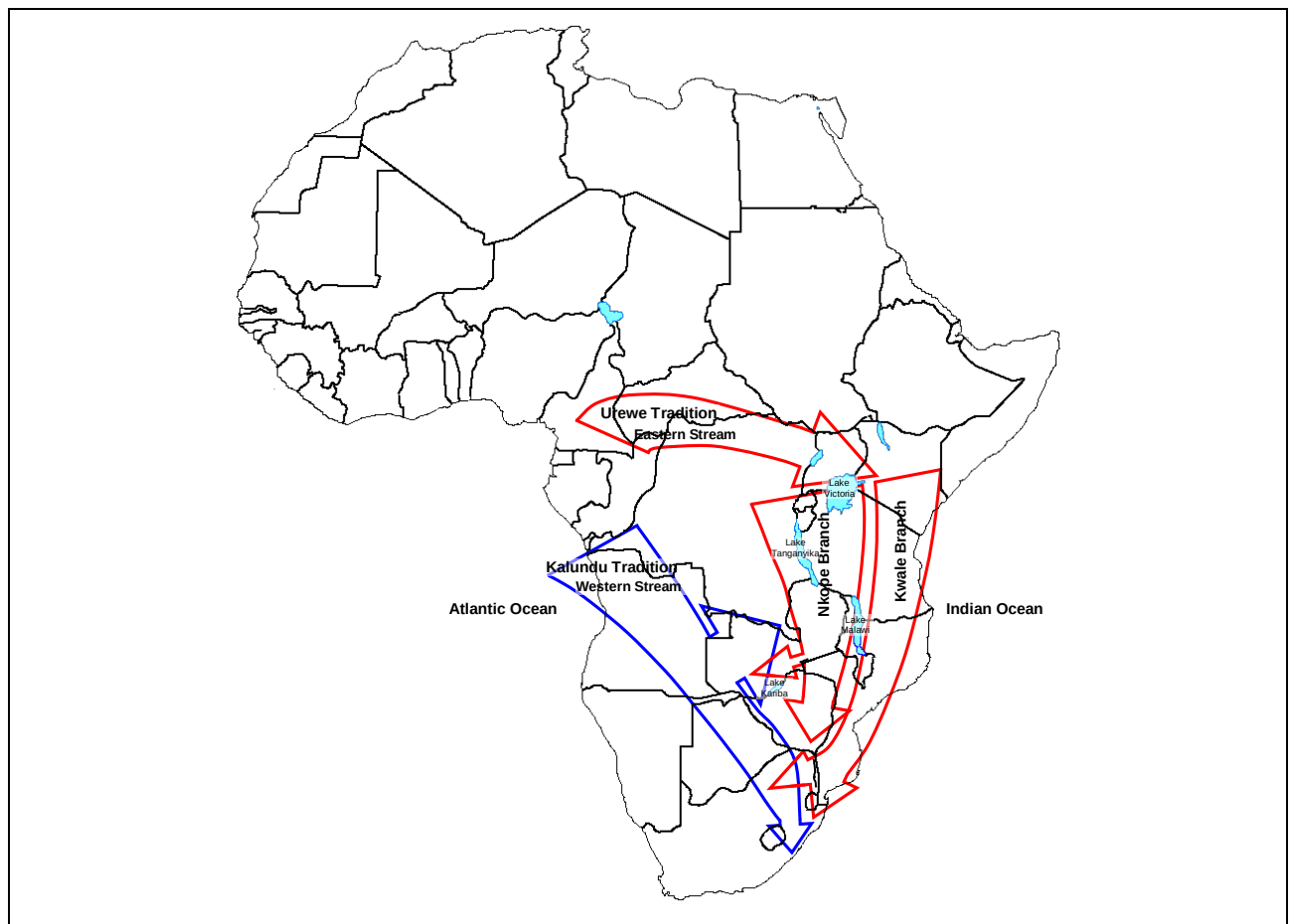
Archaeological, linguistic and genetic data indicate that the pattern of genetic variation in modern African populations has been shaped by a number of important long-range migration events in Africa within the past several thousand years. The most significant demographic, linguistic and technological event in sub-Saharan Africa is the spread of Bantu-speaking populations into the Great Lakes region located in eastern Africa, followed by expansion into southern Africa within the past 5,000 to 3,000 years from the so-called proto-Bantu homeland located around the border of present-day Nigeria and Cameroon in the Cross River Valley area of western-central Africa (Soodyall *et al.*, 1996, Cavalli-Sforza, 2001; Tishkoff and Williams, 2002; Huffman, 2006).

¹ Due to South Africa's history of Apartheid, the term Bantu has pejorative connotations. In the current investigation, this term is used purely to identify speakers of a specific branch of Niger-Congo languages.

Archaeological remains indicate that these Bantu-speaking populations were mixed agriculturists that herded cattle, sheep and goats and cultivated sorghum, millet and other crops (Huffman, 2006). In addition to agriculture, the Bantu-speakers also manufactured iron tools and copper ornaments and lived in settled villages (Huffman, 2006). It has been suggested that the Bantu expansions (also referred to as the Bantu migrations) were the result of population growth following the introduction of agriculture in western Africa in the region between Mali and Burkina Faso, 4,000 to 6,000 YBP, and were assisted by the adoption of an iron-based technology (Ammerman and Cavalli-Sforza 1984; Cavalli-Sforza, 2001; Plaza *et al.*, 2004). The expansion of Bantu-speaking populations is therefore credited with the dispersal of agriculture to southern and central Africa (Cavalli-Sforza, 2001; Pereira *et al.*, 2001).

The Bantu expansion toward the south divided into two major paths, namely the Western Stream that expanded to the south along the Atlantic coast, and the Eastern Stream, which passed by over the rainforest, to the area of the Great Lakes, after which the agropastoralists moved into the south (Plaza *et al.*, 2004). Based on evidence provided by ceramics, the two streams can be divided into three branches, namely the Kwale branch of the Urewe Tradition from the east, the Nkope branch of the Urewe Tradition from the north, and the Kalundu Tradition from the west (Huffman, 2006). This first movement of the Bantu expansion reached the Great Lakes region of Uganda and Kenya to form the Eastern Bantu core area and marked the beginning of the Early Iron Age (Cavalli-Sforza, 2001; Huffman, 2006). The routes for these movements within Africa are indicated in Figure 3.3.

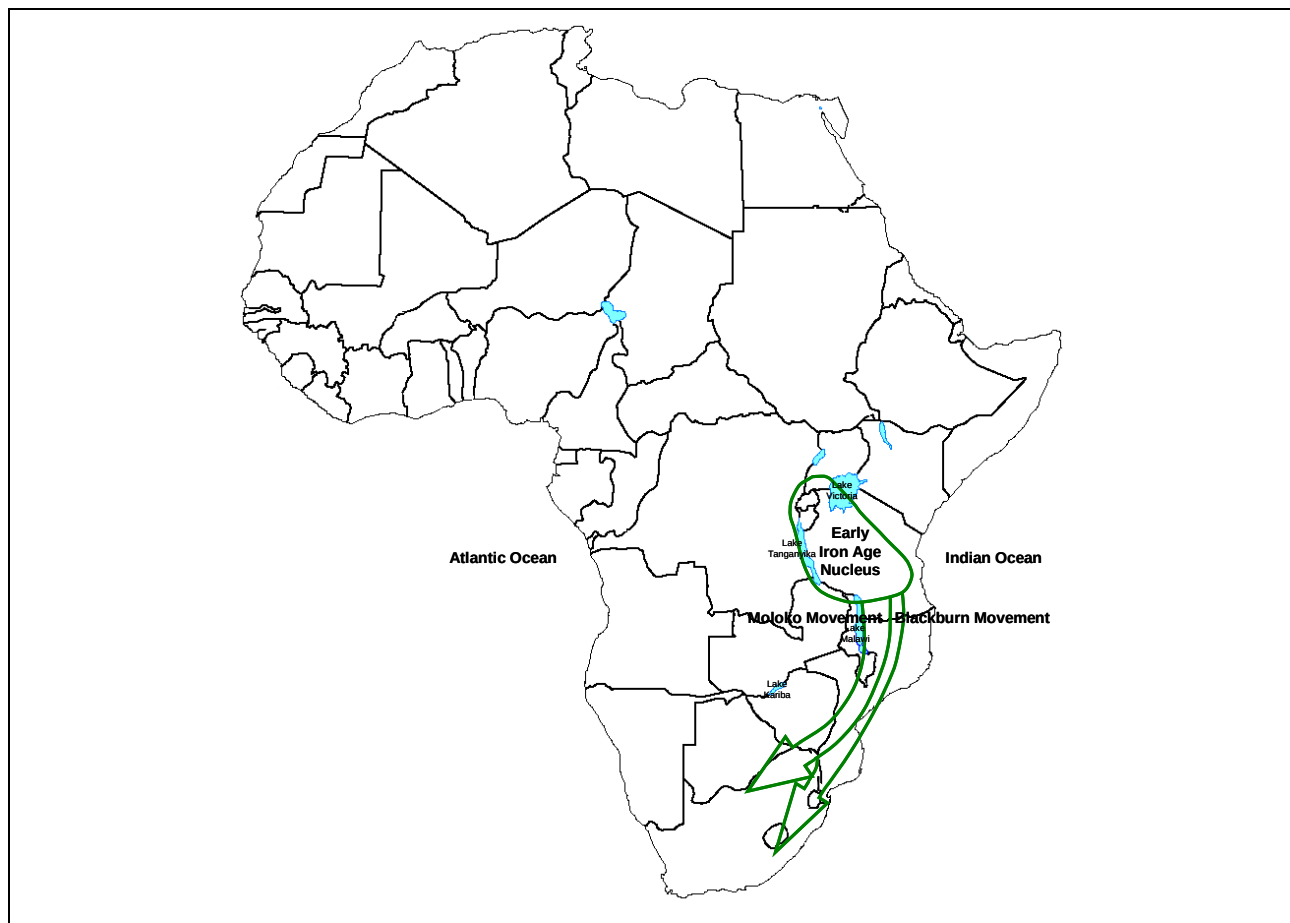
Figure 3.3 Spread of Bantu-speakers in Africa during the first movement of the Bantu expansion during the Early Iron Age



Early Iron Age movement of Bantu-speakers. Red arrows indicate the movements that form part of the Urewe Tradition, while the blue arrows indicate the movement associated with the Kalundu Tradition. Adapted from Huffman (2006).

Approximately 2,000 YBP, the ancestors of the Nguni-speakers left East Africa from an area west of Lake Victoria and moved down into South Africa using a route (Blackburn Movement) along the Ruvuma River along the coast of the Indian Ocean to reach present-day KwaZulu-Natal by the end of the third century A.D. (Pereira *et al.*, 2001). The ancestors of the Sotho-Tswana speakers also moved south slightly later using a further inland route (Moloko Movement) along the shores of Lake Malawi, through what is now eastern Zimbabwe, reaching present-day Limpopo approximately 500 A.D. (Pereira *et al.*, 2001). This second set of large scale movements marked the beginning of the Late Iron Age. The routes in Africa of the second set of movements and the different branches within each movement are indicated in Figure 3.4.

Figure 3.4 Spread of Bantu-speakers in Africa during the second movement of the Bantu expansion



Late Iron Age movement of Bantu-speakers. Green arrows indicate the Blackburn and Moloko movements. Adapted from Huffman (2006).

Mixing of Bantu-speaking populations occurred when the Western and Eastern Traditions (refer to Figure 3.3) eventually encountered each other in the south-central region of the African continent. In addition, after the occupation of central and southern Africa by Bantu-speakers, mixing also occurred with Nilotic-speakers in East Africa and Khoi-San speakers in southern Africa (Cavalli-Sforza, 2001).

It appears likely that indigenous Khoi-San speaking hunters and herders were gradually assimilated, or displaced towards the south and centre of the continent, due to contact with Bantu-speaking populations, where they currently inhabit parts of southern Africa, as well as southern Angola, mainly in harsh environments such as the Kalahari Desert (Plaza *et al.*, 2004). Evidence of contact leading to assimilation or displacement, is provided by the presence of characteristic Khoi-San clicks in several southern African Bantu languages, and the continued presence of Khoi-San speakers in southern regions which experience a winter-rainfall regime unsuitable for equatorial crops (Salas *et al.*, 2002). The Khoi-San thrived in these southern regions until the arrival of European colonists in the

15th century (Salas *et al.*, 2002), while the Bantu-speakers did not advance further south, remaining a few hundred kilometres from the Cape of Good Hope in South Africa at the time when the Dutch disembarked there in 1652 (Cavalli-Sforza, 2001).

Genetic data provides support for the Bantu expansions and genetic signatures of the Bantu expansions exist in mtDNA data (Soodyall *et al.*, 1996), Y chromosome data (Passarino *et al.*, 1998) and nuclear markers (Tishkoff *et al.*, 2009). Investigations of mtDNA identified a number of mtDNA haplogroups that are possibly associated with the Bantu expansions, and which can trace their ancestries to different geographic regions of Africa. These haplogroups include L0a (Bandelt *et al.*, 1995; Chen *et al.*, 1995; Soodyall *et al.*, 1996; Chen *et al.*, 2000; Pereira *et al.*, 2001), L1c (Rando *et al.*, 1998), L5a, previously known as L1e (Alves-Silva *et al.*, 2000; Bandelt *et al.*, 2001a), L2a (Pereira *et al.*, 2001; Salas *et al.*, 2002), L3b (Watson *et al.*, 1997; Pereira *et al.*, 2001), and L3e (Alves-Silva *et al.*, 2000; Bandelt *et al.*, 2001a; Pereira *et al.*, 2001; Rosa *et al.*, 2004). L0a is hypothesised to have been brought to south-eastern Africa by the Eastern Stream of the Bantu expansion having been acquired in East Africa, or brought directly from a region close to the source of the Bantu languages in western-central Africa. It may also have arrived from some intermediate position by the Western Stream through central Africa (Salas *et al.*, 2002).

A recent investigation of genetic structure in Africa based on nuclear markers provides additional evidence for an East African Bantu migration into southern Africa with the subsequent incorporation of Khoi-San ancestry into South-east Bantu populations ca. 1,500 to 1,000 YBP (Tishkoff *et al.*, 2009). In addition, a number of investigations has also indicated that the Sahel (the region stretching from the Atlantic Ocean in the west to the Red Sea in the east which includes present-day Senegal, Mauritania, Mali, Burkina Faso, Niger, Nigeria, Chad, Sudan, and Eritrea) has been employed as a passageway for bi-directional migration between western and eastern Africa (Černý *et al.*, 2007; Tishkoff *et al.*, 2009).

3.1.4 The African Diaspora

The African Diaspora refers to the voluntary and involuntary worldwide movement of Africans since antiquity (Harris *et al.*, 1996). In the majority of cases, the presence of individuals of recent African descent in countries outside of Africa is attributed to involuntary migration, i.e. the African Slave Trade (Harris *et al.*, 1996). Indigenous Africans

were transported from the African continent to other parts of the world via a number of different routes which involved travel across the Sahara, through the Red Sea, and from ports in the Indian and Atlantic Oceans (M'bokolo, 1998). Africans were enslaved for ten centuries (between the 9th and the 19th century) by predominantly Muslim countries (the Arab Slave Trade), which was followed by more than four centuries of enslavement (from the end of the 15th to the 19th century) by mainly Christian countries in Europe (the European Slave Trade) with the purpose of developing the Americas (M'bokolo, 1998).

The slave trade expanded from north-western Africa along the Atlantic coast, arriving in Angola by the 17th century and Mozambique by the 19th century. It is estimated that nine million slaves were transported along the trans-Saharan caravan route, four million slaves were exported via the Red Sea, four million through the Indian Ocean ports, and between eleven and twenty million across the Atlantic Ocean (M'bokolo, 1998). Based on historical records, it has been proposed that the different regions in Africa did not contribute equally to the slave trade with western Africa contributing ca. eight million people, west-central Africa (the area stretching from Cameroon to Angola) ca. four million, and south-east Africa (from Mozambique and Madagascar) ca. one million (Curtin, 1969; Fage, 1969; Salas *et al.*, 2002; Salas *et al.*, 2004a).

The majority of slaves were transported to colonies controlled at the time by the British, French, Spanish, Dutch, Danish and Portuguese and included the Americas (North America, Central America and South America) and the Caribbean which consists of a chain of islands surrounding the Caribbean Sea, which is bordered by the Straits of Florida and the Gulf of Mexico in the north, the Northern Atlantic Ocean in the east and north-east, and the coastline of the South American continent in the south (Curtin, 1969). The majority of slaves were used as labour in plantations to cultivate sugarcane, coffee and cotton, although some slaves also worked in gold and diamond mines or as servants (Curtin, 1969; Plaza *et al.*, 2004). Of the different routes used to transport slaves from Africa, the route across the Atlantic Ocean (the Atlantic Slave Trade), has received the most attention due to the magnitude of this movement and the fact that new and distinctive populations were created in the various colonies as a result of prejudice of European populations against African populations (Sans, 2000; Salas *et al.*, 2004a).

The genetic signature of these past population movements is evident in the current genetic composition of admixed populations in the countries to which the slaves were transported. In recent years, these admixed populations have been investigated to determine the extent

of the contribution made to these populations by different groups (Alves-Silva *et al.*, 2000; Bandelt *et al.*, 2001a; Pereira *et al.*, 2001; Brehm *et al.* in 2002; Salas *et al.*, 2002; Plaza *et al.*, 2004; Salas *et al.*, 2004a). These studies investigate the maternal ancestry of individuals in these populations by determining the mitochondrial haplogroups of individuals and subsequently using the information obtained to trace their origins by identifying the likely geographic location in Africa from which they originated. The geographic origin of an individual is frequently determined based on the current distribution of the specific mtDNA haplogroup in Africa (Salas *et al.*, 2004a).

Using this methodology, Alves-Silva *et al.* in 2000 investigated the recent maternal ancestry of individuals from Brazil. A considerable number of mtDNA lineages were traced to western Africa (Coast of Guinea), a few lineages represented south-eastern Africa (Mozambique), while almost 50% of the observed mtDNA lineages were assigned to haplogroups L1c and L3e which are uncommon in western Africa. The L1c and L3e lineages are however present at high frequencies in mtDNA datasets representing west-central Africa which suggested that the majority of the mtDNA lineages identified were of west-central African origin, an area which stretches from Cameroon to Angola. This corresponds to historical records, as both the Congo and Angola have been identified as major sources of Brazilian slaves (Curtin, 1969). Subsequently, Bandelt *et al.* in 2001a compared the L3e profiles of four regions namely Cabo Verde (a group of islands located off the western coast of Africa, opposite Mauritania and Senegal), Santo Domingo (located in the Dominican Republic in the Caribbean), São Tomé (located in the Gulf of Guinea, off the western equatorial coast of Africa) and Brazil. They determined that Cabo Verde was representative of the Senegambia region (present day Senegal and Gambia), Santo Domingo of non-Atlantic western Africa, São Tomé of central Africa and Brazil of all the areas with the exception of the Senegambia region. Additional work by Brehm *et al.* in 2002 confirmed the presence of mtDNA lineages originating from the western Atlantic Coast in the Cabo Verde islands from which slaves were further transported to the Caribbean and Brazil. In 2004, Plaza and colleagues investigated mtDNA variation in Angola and determined that the mtDNA lineages observed in Angola were similar to the African lineages identified in Brazil, thereby providing genetic support for the historical records indicating that Brazil was the major American destination for slaves from Angola.

The studies discussed in the above paragraph provide some insight on the origins in Africa of slaves transported from the Atlantic Coast. The contribution of south-eastern Africa to the slave trade has also been investigated, although to a lesser extent. Pereira *et al.* in

2001 studied the mtDNA haplogroups present in Mozambique. They detected a greater number of matches between sequences from Mozambique with sequences from the Americas than with sequences from Europe. This result corresponded with the historical documentation which indicated that slaves from East Africa were more likely to be transported to the Americas than to Europe. In addition, Salas *et al.* in 2002 attributed the presence of the L0a mtDNA lineage in the islands of São Tomé and Bioko (located in the Gulf of Guinea in the part of the Atlantic Ocean southwest of Africa), as well as in the Americas, to the transportation of slaves from Mozambique in East Africa to other parts of the world.

Salas and colleagues (2004a) attempted to calculate the contribution of the major sub-regions in Africa to the mtDNA lineages observed in the Americas and Eurasia. The different regions in Africa that contributed slaves to other parts of the world are summarised in Table 3.1.

Table 3.1 Regions in Africa that contributed slaves to other parts of the world

World region	Region in Africa from which slaves were transported	Estimated size of contribution to total slave numbers
Americas	Northern Africa	Minor
	Eastern Africa	
	Southern Africa	
North America	Western Africa	Major
	West-central Africa	Minor
Central America	Western Africa	Major
South America	West-central Africa	Major
	Western Africa	Substantial
	South-eastern Africa	Minor
Caribbean	West-central Africa	Major
Europe	Western Africa	Minor
	South-eastern Africa	
	South-western Africa	
Near East	Eastern Africa	Major

Adapted from Plaza *et al.* (2004) and Salas *et al.* (2004a).

The study of the distribution and geographic specificity of the most important mtDNA haplogroups, the geographic distribution of similar or closely related mitochondrial sequences and the diversity observed in clades of recent origin with the purpose of disentangling different settlement processes and routes of migration, are referred to as the phylogeography of mtDNA (Rando *et al.*, 1999). As our knowledge of the phylogeography of mtDNA expands and more mtDNA sequences become available to provide greater

resolution of the global mitochondrial landscape, it becomes possible to trace mtDNA haplotypes observed on other continents to broad geographical regions within Africa (Salas *et al.*, 2004a). However, there are many limitations associated with this approach due to characteristics linked to African mtDNA variation which will impede the process. These characteristics include (Salas *et al.*, 2004a):

- a. the limited sampling of some parts of the African continent,
- b. the wide distribution of major mtDNA haplogroups such as L2a within Africa during prehistoric times,
- c. the effect of the Bantu migrations which resulted in a small number of mtDNA haplotypes from different haplogroups being widely dispersed throughout subtropical Africa, and
- d. large-scale and widespread movement of people within Africa due to the slave trade.

Therefore, although mitochondrial sequence datasets continually increase in size due to increased sampling of regions in Africa and the availability of more complete mtDNA sequences, and it thus becomes feasible to perform more extensive phylogeographic analyses, the allocation of mtDNA haplogroups from individuals with African ancestry to specific geographic locations in Africa remains problematic as a result of the relatively recent widespread movements of people in Africa due to events such as the Bantu migrations and the slave trade (Salas *et al.*, 2002; Salas *et al.*, 2004a). This implies that although over time the broad pattern of human dispersal as revealed by mtDNA will gradually become clearer, tracing the mtDNA lineages of individuals will remain plagued by difficulty. The claims made that the ancestry of American or European mtDNA lineages can be traced to specific geographic regions or populations in Africa must therefore be treated with caution (Salas *et al.*, 2004a).

This conclusion was supported by a recent genetic investigation of the ancestry of African Americans. This study indicated that the majority of African Americans have high proportions of Bantu and non-Bantu Niger-Kordofanian ancestry, which corresponds with the geographic origins of movements to the Atlantic coast stretching from Senegambia in the west to Angola and South Africa in the south (Tishkoff *et al.*, 2009). This suggests that the ancestry of most African Americans is an amalgamation of the contributions from different populations from different regions of West Africa. The mixed ancestry of African Americans in combination with the substructure detected among Niger-Kordofanian speakers, make the discovery of ancestral linkages between African Americans and

specific ethnic groups in Africa extremely difficult, especially when a limited number of genetic markers are used (Tishkoff *et al.*, 2009).

3.2 MITOCHONDRIAL PHYLOGENIES

The mitochondrial genome has three favourable characteristics that have led to its use in phylogenetic studies. The mitochondrial genome has a tenfold increased mutation rate compared to the nuclear genome (Brown *et al.*, 1979), it does not undergo recombination (Zuckerman *et al.*, 1984), and it is only inherited through the maternal lineage (Giles *et al.*, 1980). Furthermore, mitochondria are abundant in cells, ca. 1,000 to 10,000 mitochondria are observed per cell (Clayton, 1982), and the mitochondrial genome is relatively small with a size of 16.5 kb (Anderson *et al.*, 1981). Mitochondria are therefore ideally suited to infer genetic phylogenies of species, including *Homo sapiens*.

A number of constraints must however be taken into account when considering the use of the mitochondrial genome. Control region mitochondrial DNA sequences, containing HVS1 and HVS2, used in isolation may not be informative enough for horizontal studies (i.e. individual or family typisation), and the possibility exists that identical mtDNA segments could have developed analogously or independently through multiple transition or transversion events (Pusch *et al.*, 2003).

3.2.1 Methods used to assess variation

When constructing a phylogeny, the nucleotide differences observed between mtDNA genomes is expressed as sequence variation. This variation can be studied using several methods. Early investigations of human mtDNA variation were based on RFLPs of either purified genomic DNA (gDNA) or purified mtDNA (Brown, 1980; Johnson *et al.*, 1983; Cann *et al.*, 1987). Subsequent to the invention of polymerase chain reaction (PCR) technology and the availability of methods for rapid sequence determination of PCR products, RFLP analysis of mtDNA PCR products and sequence analysis of HVS1 in the control region became the methods of choice (Chen *et al.*, 2000; Brehm *et al.*, 2002). In the last few years, following the development of high-throughput sequencing technology, it has become increasingly common to determine the sequence of complete mitochondrial genomes (Ingman *et al.*, 2000; Maca-Meyer *et al.*, 2001; Herrnstadt *et al.*, 2002a; Mishmar *et al.*, 2003; Gonder *et al.*, 2007). In the past, RFLP analysis was popular due to its

simplicity and robustness. Currently, sequencing has become more widely used due to this technique providing greater resolution and higher specificity (Vigilant *et al.*, 1991).

3.2.1.1 Restriction fragment length polymorphism analysis

RFLP analysis has revealed several mitochondrial haplotypes present within humans based on differential restriction patterns of restriction enzymes (Brown, 1980; Denaro *et al.*, 1981; Johnson *et al.*, 1983). Two types of RFLP analyses can be performed, low-resolution (LR) RFLP analysis using six rare-cutting restriction endonucleases, namely *Ava* II, *Bam* HI, *Hae* II, *Hinc* II, *Hpa* I and *Msp* I, or high-resolution (HR) RFLP analysis which uses the six rare-cutting restriction endonucleases as well as an additional eight endonucleases, namely *Alu* I, *Dde* I, *Hae* III, *Hha* I, *Hin* fI, *Hpa* II, *Mbo* I, *Rsa* I, and *Taq* I (Chen *et al.*, 1995). A number of studies have been performed employing these techniques.

In 1980, Brown analysed the mtDNA of individuals of diverse ethnic and geographic origin such as Caucasians from the United States of America (USA), the Philippines and Egypt, Mongolians from China, and African Americans from the USA using 18 restriction endonucleases. No differences between the individuals were detected for seven of the enzymes used. However, one or more differences were detected for the remaining 11 enzymes used. The observed differences were ascribed to the presence of mitochondrial alterations of which many were shared between two or more samples. It was thus concluded that the patterns of enzyme restriction were group-specific.

In 1981, Denaro and colleagues analysed the mtDNA of Caucasians from the USA and Europe, Asians from Taiwan, China and Japan, as well as the mtDNA of Pygmies, Khoi-San and Bantu-speaking Africans from Africa. Using the *Hpa* I restriction enzyme, six different cleavage patterns of varying frequency were discovered between these populations. Certain *Hpa* I sites were however absent in specific populations, and one haplotype was demonstrated to be present only in the African population in more than 90% of individuals tested. By using this information in conjunction with the frequency of the different haplotypes, it was possible to clearly distinguish between populations.

In 1983, Johnson and colleagues analysed Caucasians, Asians, Khoi-San and Bantu-speaking Africans. They extended a previous study (Denaro *et al.*, 1981) and used four additional enzymes *Ava* II, *Bam* HI, *Hae* II, and *Msp* I. Thirty-five distinct mtDNA types

were identified, and a high correlation between mtDNA and the ethnic origin of the individual was observed. Similar results were obtained in a subsequent study where RFLP analysis allowed for the classification of mtDNA genomes into groups, which correlated with the ethnic origin of the individual (Cann *et al.*, 1987).

3.2.1.2 Sequencing analysis

RFLP analysis is an indirect way of comparing mtDNA sequences. A more direct means by which sequence variation can be detected is by determining the mtDNA nucleotide sequence of selected regions. These selected regions can include parts of the non-coding region, e.g. HVS1 and HVS2, which are contained in the control region or D-loop. Alternatively, selected regions located within the coding region can be investigated. In some cases, sequencing of the complete mtDNA genome has also been performed (Maca-Meyer *et al.*, 2001). Sequence data can also be used to estimate substitution rates and divergence times more accurately than RFLP analyses when investigating molecular evolution.

In a study performed by Vigilant and colleagues (1991), the sequence of the control region, a 1,122 bp segment of non-coding DNA that contains HVS1 and HVS2, was determined. The choice of these regions was based on the fact that it is the most rapidly evolving and polymorphic region of the mtDNA genome (Vigilant *et al.*, 1989). The mtDNA sequences of the two hypervariable regions of 189 individuals from diverse geographic origins were determined (Vigilant *et al.*, 1991). The study indicated that the mtDNA types, as defined by the control region sequences, displayed strong geographic specificity. This observation was in agreement with other investigations of human mtDNA (Vigilant *et al.*, 1991).

Sequence data from the D-loop region alone is however insufficient for the accurate calculation of nucleotide substitution rates and divergence times of hominoid or hominid mtDNAs (Horai *et al.*, 1995). In order to estimate these substitution rates and divergence times, Horai and colleagues (1995) determined the complete mitochondrial DNA sequences of three humans and four non-human primates. Using this data they calculated a substitution rate of 7.00×10^{-8} per site per year for the D-loop region, and inferred the age of the last common ancestor for all human mtDNAs to be $143,000 \pm 18,000$ years (Horai *et al.*, 1995).

Both RFLP analysis and sequencing analysis generate mtDNA haplotype data. The relationship between the haplotypes determined by RFLP analysis and the control region sequences for African populations were addressed by Chen and colleagues (2000). They observed that the genetic relationships deduced from HR-RFLP analysis were completely consistent with, and sometimes more detailed than those obtained from control region sequencing alone. This finding implies a need for studies to combine the two methods of analysis in order to more fully understand the genetic relationships among human populations (Chen *et al.*, 2000), and further indicates that combined RFLP and sequencing data are more powerful than when used separately.

3.2.1.3 Errors in mitochondrial sequence data

The importance of the quality of mitochondrial sequence data is emphasised by the discovery in recent years that numerous mitochondrial datasets contain errors. These errors can be classified into five types which relate to the analysis, evaluation and recording of the sequence data (Bandelt *et al.*, 2001b; Salas *et al.*, 2005). The types are type I (base shift errors), type II (reference bias errors), type III (phantom mutations), type IV (base misscoring) and type V (artifactual recombination). Errors can be detected in datasets by the presence of specific indicators or warning signs (Salas *et al.*, 2005). These signs include the overrepresentation of transversions (transitions are much more frequent than transversions), an excess of insertion and deletion events (insertion and deletion events are generally rare), the observed alterations do not fit the motif for the haplogroup (expected alterations are absent), mismatches in database searches (inconsistent numbering of alterations can lead to false mismatches) and the occurrence of multiple heteroplasmy events (numerous heteroplasmies are uncommon and their presence will frequently allude to sequencing problems such as phantom mutations due to high background noise or contamination).

There are however a number of strategies that can be employed to avoid the inclusion of errors in mitochondrial sequence datasets (Bandelt *et al.*, 2002; Yao *et al.*, 2003; Salas *et al.*, 2005). These strategies are listed in Table 3.2.

Table 3.2 Strategies that can be employed to avoid the inclusion of errors in mitochondrial sequence datasets

a.	Prevent incorrect identification of samples by proper labelling of individual samples and by designing a sequencing strategy that ensures that there is sufficient overlap between sequences to prove identity.
b.	Reduce high levels of background in electropherograms by optimising the cycle sequencing reaction.
c.	Perform sequences using both a forward primer (L-strand) and reverse primer (H-strand) or perform two independent amplifications that are sequenced using different primers.
d.	Visually inspect the electropherogram at least twice, even when no ambiguity (non-determined nucleotides) is present. This should be performed independently by more than one person. Do not rely on the automatic electropherogram reading generated by the sequencing machine software.
e.	Screen both the control region and the coding region for alterations even if only the coding region will be used in analyses. The coding region information can be linked to motifs observed in the control region and the control region can be used to assess the quality of the sequences, i.e. the control region is more difficult to sequence and document, therefore if it has been sequenced and documented accurately, the data for the coding region will more likely be accurate.
f.	Compare aligned sequences with the correct version of the rCRS (Andrews <i>et al.</i> , 1999).
g.	Reduce the risk of documentation errors by nominating two different individuals to independently generate tables from raw data.
h.	Repeat sequences containing suspicious sites.
i.	Do not view the observed alterations in isolation, i.e. do the observed alterations for a haplogroup include the expected ones for that haplogroup according to the known phylogeny? Phylogenies are available which indicate the alterations associated with the basal branches of phylogenies.

rCRS = revised Cambridge Reference Sequence. Adapted from Bandelt *et al.* (2002), Yao *et al.* (2003) and Salas *et al.* (2005).

3.2.2 Construction of phylogenetic trees

Phylogenetic analysis can be defined as the methods used to convert information contained in sequences (or other forms of data such as RFLPs) into a phylogenetic tree. A phylogenetic tree can in turn be defined as a model of the evolutionary history of a group of sequences or organisms (Page and Holmes, 1998). A phylogenetic tree has a number of different uses (Page and Holmes, 1998):

- it provides a way in which to order data and contains a vast amount of evolutionary information within its branches,
- it can provide information regarding the relationships between sequences or organisms and the order of branches of lineages,
- it can link modern descendants to common ancestors,
- it can be used to estimate times of branching events or divergence, and
- it can provide information about the different evolutionary processes or forces (e.g. natural selection, migration and mutation) and demographic factors (e.g. varying rates of population growth) that shaped the structure of contemporary populations as these processes alter the branching structures within a tree.

A phylogenetic tree therefore serves as a visual representation of the evolutionary relationships that exists between the specific sequences or organisms under investigation, where evolution is defined as cumulative change in the genetic composition of an

individual or population that results in increased adaptation to the environment (Hartl, 1988).

When using mtDNA sequences in evolutionary studies, two basic approaches can be followed, namely the lineage-based approach and the population-based approach. The lineage-based approach is used to disentangle the history of mtDNA lineages (i.e. haplogroups). The population-based approach is used to investigate the prehistory of individual populations, geographical regions, or of population migrations through the use of population genetics methods (Forster *et al.*, 2001).

The lineage-based approach of investigating haplogroups has a number of limitations. It only reveals the history of the haplogroup under investigation, and does not offer direct insight into the history of the population in which the haplogroup is detected. Unfortunately, the tendency exists to equate the age of the haplogroup with the age of the population, and to assume that the distribution of each individual haplogroup is an indication of a separate migration event when in fact all the haplogroups contained within a migrating population are carried forth. The age of a haplogroup in a population experiencing a migration event, indicates when the point mutation which defines the specific haplogroup arose, not when a migration event occurred (Simoni *et al.*, 2000b).

The construction of phylogenetic trees involves a number of steps. These steps involve the alignment of the sequences to ensure comparison of homologous regions which are used to infer the common history of the sequences (Salemi and Vandamme, 2003), the choice of an evolutionary model to correct for unequal rates of transitions and transversions, as well as differences in substitution rates among nucleotide sites (refer to Section 3.2.2.1), and the choice of method which will depend on the specific investigation and aims thereof (refer to Section 3.2.2.2).

3.2.2.1 Evolutionary models of nucleotide substitution

A number of sequence evolution models have been proposed to imitate the nucleotide substitution process. The most commonly used models, the parameters addressed in these models and the assumptions made, are indicated in Table 3.3.

Table 3.3 Evolutionary models that describe the nucleotide substitution process

Evolutionary model		Description of model	
Full name	Short name	Parameters	Assumptions
Jukes-Cantor	JC	Equal nucleotide frequencies. All substitutions equally likely.	All nucleotide sites change independently. The substitution rate is constant over time and in different lineages. The nucleotide composition is at equilibrium. The conditional probabilities of nucleotide substitutions are the same for all sites and do not change over time.
Kimura two parameter	K2P	Equal nucleotide frequencies. Transitions and transversions have different substitution rates.	
Felsenstein (1981)	F81	Unequal nucleotide frequencies. Frequency of each nucleotide is ca. the same in all sequences compared. All substitutions equally likely.	
Hassegawa, Kishino, Yano (1985)	HKY85	Unequal nucleotide frequencies. Transitions and transversions have different substitution rates.	
General reversible	REV	Unequal nucleotide frequencies. Each possible substitution has a different substitution rate.	
---	F84	Unequal nucleotide frequencies. Transitions and transversions have different substitution rates. The substitution rate at different sites is not equal. All relevant sites are included in the sequence, not just those that have changed or those that are phylogenetically informative.	All nucleotide sites change independently. Different lineages evolve independently.

Adapted from Page and Holmes (1998).

The assumptions required for the JC, K2P, F81, HKY85 and REV models (refer to Table 3.3) allow the methods to be used with ease, although they are not always realistic. New models, e.g. the F84 model, or additions to models have been proposed to correct for these assumptions of which a very important one is the assumption that the substitution rate at all nucleotide positions in a sequence is similar. In reality, nucleotide sites exhibit different rates of substitution which depend on where in the genome the nucleotide position is located, e.g. in codon position one, two or three, or in the control region compared to the coding region. This variation in substitution rate between different nucleotide sites can be modelled using the gamma distribution. This distribution has a shape parameter, alpha (α), which specifies the rate variation among sites (Yang, 1996). Small values of α indicate extreme rate variation, while large values of α represent smaller variations in rates (Page and Holmes, 1998). Estimates of the gamma shape parameter for mitochondrial genes are indicated in Table 3.4.

Table 3.4 Estimates of the α parameter of the gamma distribution of rate variation for mitochondrial regions or genes

Mitochondrial region/gene	Gamma shape parameter (α)
12S rRNA	0.16
895 bp mtDNAs	0.43
D-loop region (HVS1 and HVS2)	0.17
First and second codon position of 13 genes ^a	0.13 to 0.95
First codon position of four genes	0.18
Second codon position of four genes	0.08
Third codon position of four genes	1.58

^a = The 13 protein-coding genes in the mitochondrial genome are analysed separately and the estimates of α are 0.49, 0.86, 0.13, 0.33, 0.23, 0.23, 0.27, 0.66, 0.39, 0.45, 0.87, 0.46 and 0.95 for ATPase6, ATPase8, COI, COII, COIII, Cytb, ND1, ND2, ND3, ND4, ND4L, ND5 and ND6, respectively. Adapted from Yang (1996).

3.2.2.2 Phylogenetic tree construction methods

Phylogenetic trees can be constructed using a number of different methods. These methods can be divided into groups (Page and Holmes, 1998) based on the way in which the method processes data (distance methods vs. discrete character methods) or the approach used when constructing trees (clustering methods vs. search methods). Distance methods, e.g. the neighbour-joining method and the unweighted pair group method with arithmetic means convert aligned sequences into pairwise distance matrixes, while discrete character methods, e.g. the maximum parsimony method consider each nucleotide site directly (Page and Holmes, 1998). The conversion of sequence data into distance data leads to a loss of information regarding which nucleotide site contributes to the length of a branch within the phylogenetic tree, while the discrete character method provides information regarding both the length of a tree branch and which nucleotide site contributed to the value. Cluster methods such as the neighbour-joining method follow an algorithm to construct a phylogenetic tree, while search methods such as the maximum parsimony method use optimality criteria to rank or score a tree thereby choosing the best tree among a set of all possible trees. Cluster methods are easy to implement, fast and usually produce a single tree, although these methods do not allow for the evaluation of competing hypotheses. Search or optimality methods require a function that relates the data and the tree, e.g. a model of sequence evolution, and allows the comparison of competing hypotheses, although they are computationally very expensive.

3.2.2.2.1 Distance methods

Neighbour-joining (N-J) is a distance method which constructs phylogenetic trees from sequence data that has been converted into distance data, and it is based on the principle of identifying pairs of operational taxonomic units (OTUs) that minimise the total branch length at each stage of clustering of OTUs when starting with a star-like tree (Saitou and Nei, 1987). This method has a number of advantages which include computational speed, the generation of a unique result (a single tree) and the ability to process large datasets (Page and Holmes, 1998; Salemi and Vandamme, 2003). In addition, there is no assumption of a molecular clock and it is the method most commonly used. The method also has some disadvantages which include the fact that the input order of the sequences influences the tree obtained due to the use of the clustering algorithm, the evaluation of competing hypotheses is not allowed and it is necessary to measure the fit between the tree and the data (Page and Holmes, 1998; Salemi and Vandamme, 2003).

Another commonly used distance method is the unweighted pair group method with arithmetic mean (UPGMA) which is similar to the N-J method. This method however assumes that the sequences evolve according to a molecular clock, i.e. a constant rate of evolution is present within all the sequences in the dataset (Page and Holmes, 1998). It has been speculated that the rate of the mtDNA clock may not have been constant throughout human history due to the influence of selection. Selection affects changes on a nucleotide level, and if selection has played a role in the radiation of different mtDNA lineages as well as in the shaping of human regional mtDNA variation, it therefore follows that different lineages may have evolved at different rates (Mishmar *et al.*, 2003). This has been supported by the claim that African mtDNA macrohaplogroup L sequences violate clock-like evolution (Howell *et al.*, 2004). Since UPGMA assumes an equal rate of evolution across different lineages within a population, it is therefore better to use the N-J method for mitochondrial DNA sequences as the assumption of a valid molecular clock is not required.

3.2.2.2.2 Discrete character method

The maximum parsimony (MP) method of tree generation is a discrete character method which considers each nucleotide site directly and provides information regarding both the length of a tree branch and which nucleotide site contributed to the value. The method requires a function that relates the data and the tree, e.g. a model of sequence evolution

and makes use of optimality criteria to rank a tree. It chooses the best tree(s) among a set of all possible trees, where the best tree(s) is defined as the tree containing the fewest possible evolutionary changes that explain the data. This method has a number of advantages which include being easy to understand, it makes few assumptions regarding the evolutionary process (except for the implicit assumption that evolutionary change is rare), it does not require the assumption of a molecular clock, it allows for the comparison of competing hypotheses, no information is lost due to the conversion of sequences to distances and powerful software programs are available (Page and Holmes, 1998; Salemi and Vandamme, 2003). The method however has some disadvantages which include the phenomenon of “long branch attraction”, which can be defined as the artificial clustering of unrelated long branches together with the outgroup at the base of the tree (Page and Holmes, 1998; Salemi and Vandamme, 2003). This phenomenon is usually observed when high levels of variation are present among sequences or when high levels of divergence exist between the sequences being analysed. Another disadvantage is that MP is computationally very expensive and therefore a time consuming process (Page and Holmes, 1998; Salemi and Vandamme, 2003).

3.2.2.2.3 Choice of phylogenetic tree construction method

Salemi and Vandamme (2003) reported that no evidence has been presented that one phylogenetic tree construction method is superior to another and it is suggested that more than one method should be employed for use with each dataset. While it is not unusual to obtain slightly dissimilar tree topologies when using different tree construction methods (Salemi and Vandamme, 2003), it is important to note that consistency among different methods is an inadequate measure of statistical significance (Felsenstein, 2006). The methods chosen should however always fit the dataset, i.e. the dataset must not violate assumptions made by the method. When choosing an appropriate method to construct a phylogenetic tree it is important to take a number of important aspects into consideration. These aspects are summarised in Table 3.5. When taking these aspects into consideration, the choice of method will therefore depend on the aim of the investigation, i.e. the problem statement, the size and complexity of the available dataset and practical considerations (Hall, 2004).

Table 3.5 Aspects to take into consideration when deciding on the most appropriate method for use in constructing phylogenetic trees

Aspect	Description
Efficiency	How fast is the method?
Power	How much data is needed for a reasonable result?
Consistency	Will the method converge on the correct answer provided with enough data?
Robustness	Will minor violations of the assumptions of the method result in poor estimates of phylogeny?
Falsifiability	Will the method indicate to the researcher if assumptions are being violated, i.e. that the wrong method is being used?

Adapted from Penny *et al.* (1992).

In order to reconstruct the true evolutionary relationships among a set of sequences, a number of factors must be considered when constructing a phylogenetic tree. These factors are potential sources of error (Page and Holmes, 1998) and include sampling error, i.e. the choice of sequences used to generate the tree, assumptions made regarding the dataset, e.g. the use of an incorrect model of sequence evolution (Salemi and Vandamme, 2003), and the fact that the structure of the tree obtained is a consequence of the evolutionary history, i.e. subsequent evolution removes traces of history. The effect of recombination can also lead to errors in a phylogeny, however, in the case of mtDNA sequences recombination does not play a role. It has been demonstrated that recombination is extremely rare and requires the presence of two different mitochondrial lineages originating from mitochondria contributed by both the maternal and the paternal parent, also taking into consideration that the paternal contribution is in most cases negligible (refer to Section 2.5.3).

The question however remains whether the phylogenetic tree obtained after the completed analysis is in fact an estimate of the true evolutionary relationships. The result obtained will depend on the sample of sequences used to generate the phylogenetic tree (Page and Holmes, 1998), as well as any assumptions made regarding the dataset, e.g. the choice of evolutionary model (Salemi and Vandamme, 2003). In the end, the phylogenetic analysis will be a compromise between the specific purpose of the analysis, the particular sequence dataset and the phylogenetic tools available. The phylogenetic tree obtained, can be evaluated by comparing the tree to actual data, i.e. the observed answer can be compared to the answer expected based on other reported information.

3.2.2.3 Rooting of phylogenetic trees

Phylogenetic trees are rooted to provide a reference point from which to determine the exact evolutionary distance and timescale difference between individual sequences (Nei and Kumar, 2000). Two methods are generally used to root phylogenetic trees, namely midpoint rooting and outgroup rooting. Midpoint rooting involves the rooting of a phylogenetic tree at the midpoint of a path joining the most dissimilar sequences, or at a mean point of the paths that join the most dissimilar sequences connected through a single branch. This method assumes a similar rate of evolution in different lineages (Salemi and Vandamme, 2003). Outgroup rooting involves the use of an outgroup sequence, i.e. the sequence that is the most distantly related of all the sequences in a tree, and equal rates of evolution in different lineages is not assumed when rooting the phylogenetic tree (Salemi and Vandamme, 2003).

The outgroup sequence is employed as an external point of reference to which other sequences can be compared to infer the degree of difference within a molecular phylogeny. The outgroup sequence determines the branching order of the tree, which in turn indicates the evolutionary relationships between single sequences or groups of sequences, as well as the branch length which indicates the proportional evolutionary difference (Page and Holmes, 1998).

To be used as an outgroup, a sequence must comply with specific criteria. The outgroup sequence should be related to the other sequences, although it must be the most distantly related, that is the difference between this sequence and any other sequence must be greater than the differences between any other combination of sequences. In terms of evolution, this would mean that the sequence representing the outgroup would be the most divergent and therefore the most ancient sequence within the group of sequences.

The outgroup sequence can however not be too distantly or closely related compared to the other sequences, as this could cause an incorrect branching order within the phylogenetic tree. It has been suggested that African apes may be too distantly related to humans to allow for the calibration of the mitochondrial D-loop clock (Horai *et al.*, 1995). Chimpanzee (*Pan troglodytes*) sequences are however often used as the outgroup sequence in human mitochondrial phylogenies. Several authors, using mtDNA sequences, have demonstrated that the human-chimpanzee clade is the most likely grouping, and that gorillas diverged from the human-chimpanzee clade at an earlier time (Horai *et al.*, 1995).

This implies that the use of a chimpanzee sequence instead of a gorilla sequence as an outgroup sequence is acceptable as it has been demonstrated that humans are more closely related to chimpanzees than gorillas (Horai *et al.*, 1992). The use of a modern human sequence such as the revised Cambridge Reference Sequence (rCRS) is however not supported as this sequence will be too closely related to the other sequences in the tree. The use of another humanoid sequence as an outgroup, such as a sequence obtained from the extinct Neanderthal, would however be feasible in terms of the criteria set for an outgroup.

3.2.2.4 Confidence limits and phylogenies

The bootstrap is a statistical method which is used to obtain a nonparametric estimate of error (Efron, 1979), and it involves the re-sampling of points, with replacement, from a dataset thereby creating a series of samples of the same size as the original. In 1985, Felsenstein was the first to employ this method to estimate phylogenies, and his approach is currently widely used. In the case of phylogenies, re-sampling takes place by keeping the original sequence names while randomly sampling nucleotide sites with replacement. A phylogenetic tree is constructed from each replicate of the data, and the bootstrap confidence limit or bootstrap *P*-value is equal to the frequency of appearance of particular phylogenetic groups. These phylogenetic groups consist of groups of sequences that are observed among all of the trees constructed by the re-sampling process. The bootstrap confidence limit of different branches within a phylogenetic tree can be used to determine the relative stability of the observed phylogenetic groups, or if applied strictly, to test their statistical significance at the 95% level (Hedges, 1992).

The use of bootstrapping in the estimation of phylogenies is a trade-off between the maximum number of replicates that can be completed in a reasonable amount of time and the minimum number of replicates required for accurate estimation of the bootstrap confidence limit. The bootstrap confidence limit is the proportion of phylogenetic trees that contain a specific phylogenetic group. It therefore follows the binomial distribution, which has a variance of $\sigma^2 = [P (1 - P) / n]$, where *P* is the bootstrap *P*-value and *n* is the number of replicates. The variance, and hence the accuracy, of the bootstrap confidence limit is thus a function only of *P* and *n*. It has been determined that to obtain a bootstrap confidence limit of 95%, 2,000 bootstrap replicates must be performed, unless the computational time is prohibitive (Hedges, 1992). In this case, additional replicates are unnecessary, while fewer replicates may sacrifice statistical accuracy (Hedges, 1992).

This implies that 2,000 replicates are required for a bootstrap confidence limit to be statistically significant at the 95% level.

The bootstrap method assumes that the nucleotide positions are independent and identically distributed, i.e. there is a single distribution rate of evolutionary change across all positions. Cummings and colleagues in 1995 investigated the use of bootstrapping in phylogenies and their results implied the existence of location-dependent aspects of DNA evolution, i.e. that neighbouring nucleotides did not evolve independently. This implied that the fundamental assumptions of the bootstrap were not valid for contiguous mitochondrial DNA sequences. It is unknown how robust phylogenetic bootstrapping is to the violation of these assumptions and how it affects the accuracy of the analysis.

This limitation of bootstrapping can however be countered by modifying the sampling method, for example by sampling less than the total number of characters randomly, with replacement, from the total number of characters (Felsenstein, 1985). Another limitation of bootstrapping is that the bootstrap confidence limit associated with a branch is a reflection of the particular sequence dataset investigated and clustering method used. High levels of sequence divergence or substantial variability in nucleotide substitution rates among lineages may cause statistical inconsistency when using the majority of tree construction methods (Felsenstein, 1988). In these instances, bootstrapping could indicate statistically significant support for an incorrect topology (Nei, 1991). However, if these limitations are considered, the bootstrap method can be an uncomplicated and effective way of assessing the results generated by phylogenetic analysis. At present, bootstrapping is still routinely employed as a test of the reliability of the inferred phylogenetic trees, and it is recommended that between 200 and 2,000 replicates are performed (Hedges, 1992), while bootstrap confidence limits, or bootstrap values, of 70% and upwards are viewed as providing acceptable support for observed branching (Salemi and Vandamme, 2003).

3.2.2.5 Consensus phylogenetic trees

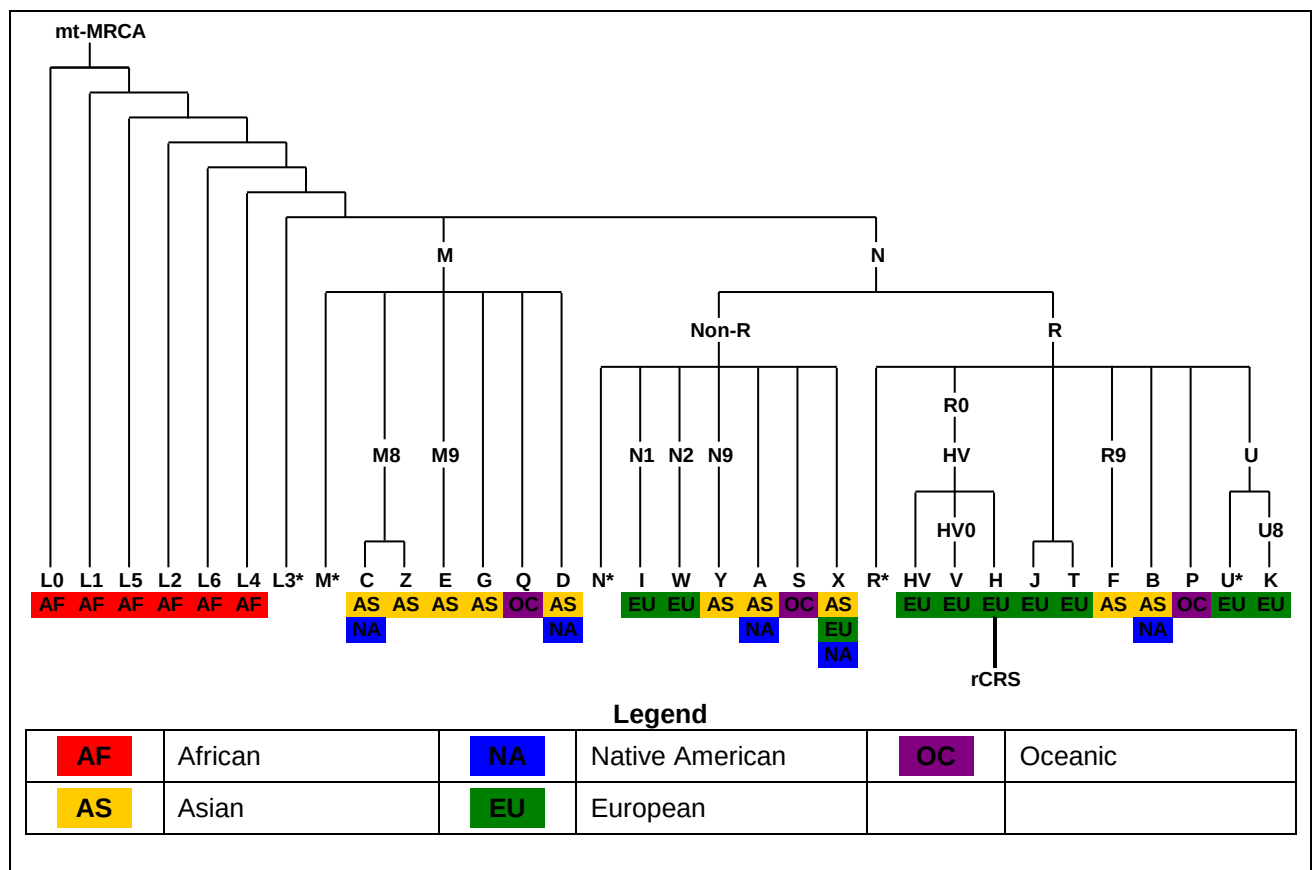
Consensus trees are usually compiled after the completion of bootstrapping for each of the methods employed in phylogenetic tree construction (refer to Section 3.2.2.2). A consensus tree can be defined as a summary of the information contained in two or more trees (Page and Holmes, 1998). There are different types of consensus trees (Felsenstein, 2006). The strict consensus tree includes only branching that occurs in all the trees considered, the 50% majority rule consensus tree retains branching observed in more than

50% of the trees, and the 50% majority rule consensus tree with extension conserves branching observed in more than 50% of the trees with the additional feature that other sets of sequences which are compatible with the branching order observed in the tree, are added to the tree in order of the frequency with which they have appeared until the tree is fully resolved.

3.3 MITOCHONDRIAL HAPLOGROUPS

Mitochondrial haplogroups represent groups of related mitochondrial sequences or haplotypes that descend from the same ancestral mtDNA molecule, are defined by shared distinguishing mutational motifs and are inclined to exhibit regional specificity (Pakendorf and Stoneking, 2005; Torroni *et al.*, 2006). Through the use of phylogenetic analyses, it has been observed that the clustering of mtDNA haplogroups are in agreement with geographic origin, i.e. there are haplogroups specific to African (Chen *et al.*, 1995; Watson *et al.*, 1997), Asian (Ballinger *et al.*, 1992; Torroni *et al.*, 1994b; Kivisild *et al.*, 2002), Native American (Torroni *et al.*, 1993a; Torroni *et al.*, 1993b), Oceanic (Forster *et al.*, 2001), and European populations (Torroni *et al.*, 1996; Macaulay *et al.*, 1999b). Based on current data, it is generally accepted that all the continent-specific mitochondrial haplogroups descended from a mitochondrial lineage that originated in Africa between 150,000 to 200,000 YBP (refer to Section 3.1.2). A schematic representation of the global phylogenetic tree of mitochondrial DNA haplogroups is presented in Figure 3.5 to illustrate the evolutionary relationships between the different continent-specific mitochondrial haplogroups.

Since mitochondrial haplogroups exhibit shared mtDNA ancestry, they are useful for estimating admixture proportions in populations originating from different geographical regions (Alves-Silva *et al.*, 2000; Pereira *et al.*, 2001; Salas *et al.*, 2004a), as well as in populations residing in areas that are recognised routes of migration (Comas *et al.*, 1998). The information obtained from different studies performing RFLP, sequencing and phylogenetic analyses have been refined and combined over time to generate a set of criteria used to define mitochondrial haplogroups. The different mitochondrial haplogroups found on the African, European and American continents, as well as in Oceania, will be discussed in the following sections (refer to Sections 3.3.1, 3.3.2, 3.3.3, 3.3.4 and 3.3.5).

Figure 3.5 A schematic representation of a global phylogenetic tree of mtDNA haplogroups

mt-MRCA = mitochondrial most recent common ancestor; rCRS = revised Cambridge Reference Sequence; AF = African mtDNA haplogroups; AS = Asian mtDNA haplogroups; NA = Native American mtDNA haplogroups; EU = European mtDNA haplogroups; OC = Oceanic mtDNA haplogroups. The asterisks placed next to several haplogroups (L3, M, N, R and U) indicate that the specific haplogroup consists of a number of sub-haplogroups. The sub-haplogroups of these haplogroups are however not indicated here for the sake of simplicity. Adapted from Olivieri *et al.* (2006), Ruiz-Pesini *et al.* (2007), Tamm *et al.* (2007), Gonder *et al.* (2007) and Van Oven and Kayser (2009).

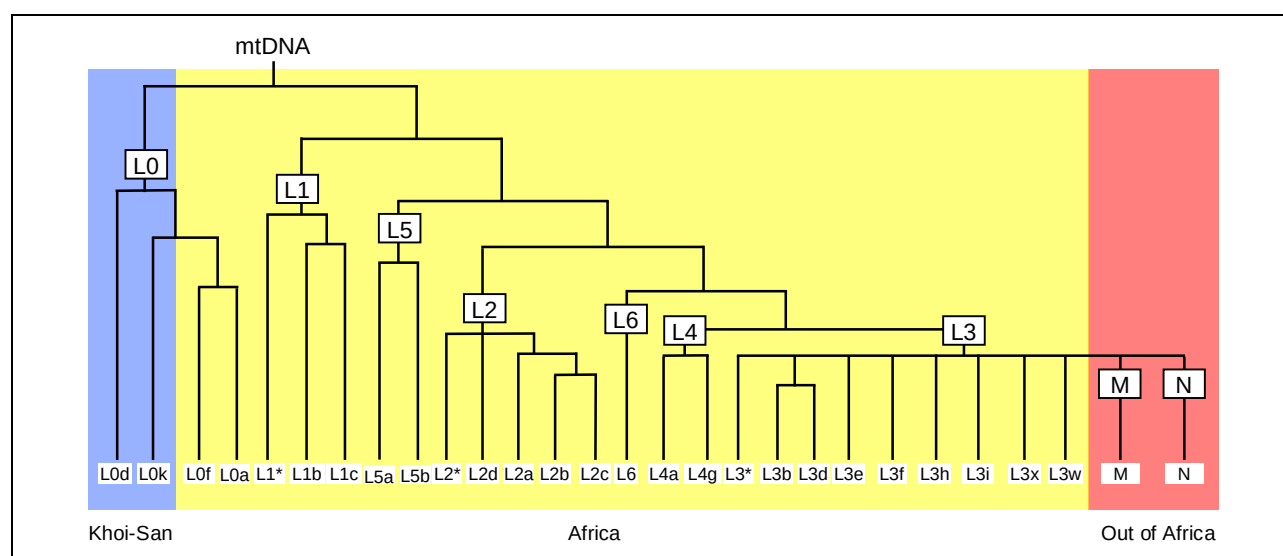
3.3.1 African mitochondrial haplogroups

Phylogenetic analysis of African mtDNA, which included individuals from Senegal, Western Pygmies from the Central African Republic, Eastern Pygmies from Zaire, and !Kung and Khwe individuals from southern Africa, revealed that 75 of the 105 haplotypes observed formed a single African-specific haplogroup referred to as macrohaplogroup L (Wallace *et al.*, 1999). This African-specific haplogroup is defined by an *Hpa* I site gain at nucleotide position 3,592 and a *Dde* I site gain at nucleotide position 10,394 (Wallace *et al.*, 1999). Macrohaplogroup L encompasses the mitochondrial lineages indigenous to Africa which are generally located at the base of the global mtDNA phylogeny (refer to Figure 3.5). Macrohaplogroup L can be subdivided into four major haplogroups, namely L0, L1, L2 and L3 (Chen *et al.*, 1995; Chen *et al.*, 2000; Mishmar *et al.*, 2003). An additional three haplogroups are also recognised, namely L4, L5 and L6, although these haplogroups are relatively rare (Gonder *et al.*, 2007).

Haplogroups L0, L1, L2, L3, L4 and L5 can be subdivided into a number of sub-haplogroups. Haplogroup L0 can be further subdivided into sub-haplogroups L0a, L0d, L0f and L0k (Salas *et al.*, 2002; Mishmar *et al.*, 2003; Salas *et al.*, 2004a). These haplogroups were previously referred to as L1a, L1d, L1f and L1k and were classified as members of haplogroup L1. Haplogroup L1 can be subdivided into two sub-haplogroups L1b and L1c, while haplogroup L2 can be further subdivided into four sub-haplogroups, namely L2a, L2b, L2c and L2d (Chen *et al.*, 2000; Torroni *et al.*, 2001; Salas *et al.*, 2002). Haplogroup L3 has also been subdivided into a number of sub-haplogroups namely L3b, L3d, L3e, L3f, L3h, L3i, L3x and L3w (Kivisild *et al.*, 2004). Haplogroup L4 contains two sub-haplogroups namely L4a and L4g (previously known as L3g) and haplogroup L5 consists of sub-haplogroups L5a (previously known as L1e) and L5b (Kivisild *et al.*, 2006). To date, no sub-haplogroups have been identified within haplogroup L6.

A high level of sequence divergence and rate variability is observed between lineages for African mitochondrial genome sequences (Wallace *et al.*, 1999; Torroni *et al.*, 2001). The evolutionary relationship between different sub-haplogroups is indicated in the schematic representation of the mtDNA phylogeny of African mitochondrial haplogroups presented in Figure 3.6. The classification of African mitochondrial haplotypes into the abovementioned haplogroups, the geographic distribution of the main African mitochondrial haplogroups and sub-haplogroups in Africa, as well as the estimated divergence times for the major haplogroups and sub-haplogroups are discussed in the following sections.

Figure 3.6 A schematic representation of the mtDNA phylogeny of African mitochondrial haplogroups



L0a, L0d, L0f and L0k were previously known as L1a, L1d, L1f and L1k, L5a was formerly identified as L1e, L4a has also been designated as L7 (Torroni *et al.*, 2006), and L4g was previously known as L3g (Salas *et al.*, 2004a; Kivisild *et al.*, 2004; Tishkoff *et al.*, 2007). Haplogroups observed mainly in Khoi-San populations are indicated in light blue. Haplogroups involved in the exodus from Africa is presented in light red. Adapted from Gonder *et al.* (2007).

3.3.1.1 Classification of African mitochondrial haplogroups

Originally, the classification of the African mtDNA haplotypes into haplogroups was performed through the use of a list of haplogroup defining restriction enzyme sites. The restriction enzyme sites defining the major haplogroups and sub-haplogroups within macrohaplogroup L are indicated in Table 3.6.

Table 3.6 Restriction enzyme sites defining continent-specific African mitochondrial haplogroups

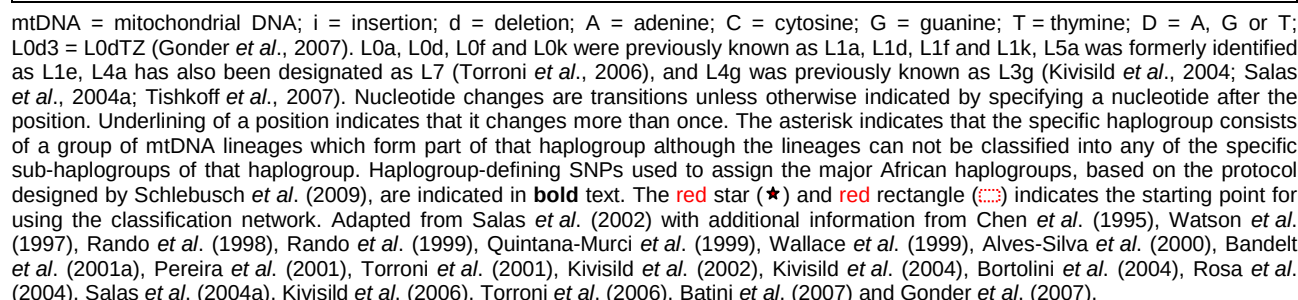
Population	Haplogroup	mtDNA variants	Reference
Africa	Macro-L	3,592 + <i>Hpa</i> I 10,394 + <i>Dde</i> I	Chen <i>et al.</i> , 1995 Chen <i>et al.</i> , 2000
	L1	2,758 - <i>Rsa</i> I 10,806 + <i>Hin</i> fl	Chen <i>et al.</i> , 1995 Chen <i>et al.</i> , 2000
	L2	16,389 + <i>Hin</i> fl 16,390 - <i>Ava</i> II	Chen <i>et al.</i> , 1995 Chen <i>et al.</i> , 2000
	L3	3,592 - <i>Hpa</i> I	Johnson <i>et al.</i> , 1983 Watson <i>et al.</i> , 1997

Adapted from Chen *et al.* (2000).

Additional haplogroups such as L0, L4, L5 and L6 were identified following the compilation of the classification scheme presented in Table 3.6. This classification scheme has subsequently been replaced by the single nucleotide polymorphism (SNP) based haplogroup classification scheme presented in Figure 3.7. These SNPs can be identified within the non-coding or the coding regions and are used to construct mtDNA networks or skeletons, which are used to assign haplogroups and to illustrate the relationships between different haplogroups and sub-haplogroups. The mtDNA skeleton of African haplogroups illustrated in Figure 3.7 was compiled using a combination of sequence data from HVS1, HVS2, the coding region, as well as selected restriction enzyme sites.

At present, two types of SNP based classification schemes are used to classify African mitochondrial genome sequences into haplogroups, i.e. a scheme based on a combination of SNPs observed in both the non-coding and the coding region (refer to Figure 3.7), and a scheme based on SNPs observed within the coding region only (refer to Section 4.9). Although differences are observed between the two classification schemes regarding the designations of sub-haplogroups within haplogroups, the major haplogroup classifications remain unchanged irrespective of the classification scheme used to assign the haplogroup designation. A comparison between the haplogroup classification schemes presented in

Figure 3.7 Classification network for African mtDNA haplogroups



The use of a classification scheme based on SNPs located only within the coding region is favoured as the coding region is viewed as being more informative compared to the non-coding region which can contain parallel or reverse mutations which can be interpreted as randomly occurring mutations (Howell *et al.*, 1996; Herrnstadt *et al.*, 2002a). The high mutation rate in the control region, in addition to the large time depth of several of the African haplogroups, give rise to the frequent occurrence of recurrent mutations at motif positions (Pereira *et al.*, 2001). This can cause the incorrect clustering of sequences, i.e. sequences that are not closely related cluster together, or closely related sequences cluster in different branches. An example is provided by the alteration at nucleotide position 16,319 in HVS1 which occurs in more than one lineage, it is observed in both L3i and in W from Eurasia (Kivisild *et al.*, 2006).

The fact that coding region SNPs are more informative in defining mitochondrial haplogroups than control region SNPs has been exploited in two recent investigations in which high-throughput methods for characterisation of coding region SNPs for haplogroup classification purposes were evaluated. Schlebusch *et al.* (2009) selected a small number of haplogroup-defining coding region SNPs (refer to Figure 3.7), and designed a SNaPshot minisequencing protocol to classify mtDNA sequences of African origin into the major African mitochondrial haplogroups. Cerezo *et al.* (2009) selected a panel of 230 SNPs and made use of matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry (MALDI-TOF MS) to classify African mtDNA sequences to both haplogroup and sub-haplogroup level. These methods provide a relatively simple, quick, cost-effective, high-throughput alternative to other methods, such as RFLP analysis and complete genome sequencing, when the genotyping of coding region SNPs is required for haplogroup classification in large sample groups.

3.3.1.2 Geographic distribution of African mitochondrial haplogroups

It has been observed that the geographic distribution of mtDNA haplogroups in Africa occur according to a distinct pattern (Rosa *et al.*, 2004). The distribution of the main haplogroups and sub-haplogroups in Africa based on geographic region is indicated in Table 3.7. The geographic distribution of the African mitochondrial haplogroups presented in Table 3.7 is based on the current distribution of haplogroups in Africa. However, due to demographic processes such as the Bantu expansions (refer to Section 3.1.3), distribution patterns of mitochondrial haplogroups present in Africa prior to these events may have been erased (Salas *et al.*, 2002).

Table 3.7 Geographic distribution of major African mitochondrial haplogroups

Haplogroup/ sub-haplogroup	Geographic distribution	Reference
L0a	Eastern, central and south-eastern Africa	Watson <i>et al.</i> , 1997 Salas <i>et al.</i> , 2002
L0a1	East and south-eastern Africa	Salas <i>et al.</i> , 2002
L0a2	Central and south-eastern Africa	Salas <i>et al.</i> , 2002
L0d	Southern, south-eastern and East Africa	Salas <i>et al.</i> , 2002
L0f	East Africa	Salas <i>et al.</i> , 2002
L0k	Southern Africa	Salas <i>et al.</i> , 2002
L1b	West, Central and North Africa	Watson <i>et al.</i> , 1997 Salas <i>et al.</i> , 2002 Rosa <i>et al.</i> , 2004 Barkhan and Soodyall, 2006
L1c	Central, North, West and East Africa	Salas <i>et al.</i> , 2002 Rosa <i>et al.</i> , 2004 Batini <i>et al.</i> , 2007
L2a	West, East, South and North Africa	Salas <i>et al.</i> , 2002 Rosa <i>et al.</i> , 2004
L2a1	South-eastern Africa	Salas <i>et al.</i> , 2002
L2b	West and western-central Africa	Torroni <i>et al.</i> , 2001 Salas <i>et al.</i> , 2002 Rosa <i>et al.</i> , 2004
L2c	West and western-central Africa	Torroni <i>et al.</i> , 2001 Salas <i>et al.</i> , 2002 Rosa <i>et al.</i> , 2004
L2d	West and western-central Africa	Torroni <i>et al.</i> , 2001 Salas <i>et al.</i> , 2002
L3b	West, North and south-eastern Africa	Salas <i>et al.</i> , 2002
L3d	West and south-eastern Africa	Rando <i>et al.</i> , 1998 Salas <i>et al.</i> , 2002 Rosa <i>et al.</i> , 2004
L3e	Widespread in Africa	Salas <i>et al.</i> , 2002
L3e1	South-eastern Africa	Salas <i>et al.</i> , 2002
L3e2	Central and West Africa	Salas <i>et al.</i> , 2002
L3e3	West Africa	Salas <i>et al.</i> , 2002
L3e4	East, Central and West Africa	Salas <i>et al.</i> , 2002
L3f	East, Central and south-eastern Africa	Salas <i>et al.</i> , 2002 Kivisild <i>et al.</i> , 2004 Rosa <i>et al.</i> , 2004
L3f1	West Africa	Salas <i>et al.</i> , 2002
L3h	West Africa	Rosa <i>et al.</i> , 2004
L3i	East Africa	Kivisild <i>et al.</i> , 2004
L3w	East Africa	Kivisild <i>et al.</i> , 2004
L3x	East Africa	Kivisild <i>et al.</i> , 2004
L4a	East and north-eastern Africa	Kivisild <i>et al.</i> , 2004
L4g	East, Central and south-eastern Africa	Salas <i>et al.</i> , 2002

Table 3.7 Continued...

Haplogroup/ sub-haplogroup	Geographic distribution	Reference
L5	East Africa	Gonder <i>et al.</i> , 2007
L5a	East, Central and south-eastern Africa	Salas <i>et al.</i> , 2002 Kivisild <i>et al.</i> , 2004
L5b	East Africa	Kivisild <i>et al.</i> , 2004
L6	East Africa	Kivisild <i>et al.</i> , 2004

L0a, L0d, L0f and L0k were previously known as L1a, L1d, L1f and L1k; L5a was formerly identified as L1e, L4a has also been designated as L7 (Torroni *et al.*, 2006), and L4g was previously known as L3g (Kivisild *et al.*, 2004; Salas *et al.*, 2004a; Tishkoff *et al.*, 2007).

3.3.1.3 Population-specific African mitochondrial haplogroups

There exists varying degrees of correlation between haplogroups classified within macrohaplogroup L and the ethnic status of African populations, and it appears that some mitochondrial lineages are population-specific. A summary of these population-specific mitochondrial lineages is presented in Table 3.8.

Table 3.8 Population-specific mitochondrial macrohaplogroup L lineages

Haplogroup/sub-haplogroup	Population ^a	Reference
L0d	Southern African Khoi-San (SAK)	Chen <i>et al.</i> , 1995 Chen <i>et al.</i> , 2000 Salas <i>et al.</i> , 2002 Tishkoff <i>et al.</i> , 2007
L0k	Southern African Khoi-San (SAK)	Chen <i>et al.</i> , 1995 Chen <i>et al.</i> , 2000 Salas <i>et al.</i> , 2002 Tishkoff <i>et al.</i> , 2007
L0a2	Pygmies: Biaka and Mbuti	Salas <i>et al.</i> , 2002
L1c lineages	Western Pygmies: Biaka and Mbenzele	Destro-Bisol <i>et al.</i> , 2004
L2a lineages	Eastern Pygmies: Mbuti	Chen <i>et al.</i> , 2000 Salas <i>et al.</i> , 2002
L5a	Eastern Pygmies: Mbuti	Destro-Bisol <i>et al.</i> , 2004

^a Two Pygmy populations are recognised, the Eastern Pygmies who settled in Zaire, and the Western Pygmies who settled in Cameroon, Congo, and Central African Republic. L0d and L0k were previously known as L1d and L1k, L5a was previously designated as L1e (Kivisild *et al.*, 2004; Salas *et al.*, 2004a).

It appears that the L0 and L1 haplogroup lineages presented in Table 3.8, which are some of the oldest lineages (refer to Section 3.3.1.4), are predominantly observed in populations such as the Khoi-San and Pygmies that are classified as hunter-gatherers, i.e. food gathering populations. Although population-specific lineages are observed almost exclusively in certain ethnic groups, lineages such as L0d, which is typical of Khoi-San populations, have been observed in other populations at low frequencies. In such cases, the presence of the population-specific lineage in the population under investigation is

usually attributed to recent admixture between populations (Pereira *et al.*, 2001). This observation is supported by recent investigations in which it was suggested that during the Bantu expansions (refer to Section 3.1.3) contacts between food producers and hunter-gatherers were characterised by unidirectional unions between hunter-gatherer females and food producer males (Pilkington *et al.*, 2008; Castri *et al.*, 2009). The investigation of the mitochondrial lineages observed in hunter-gatherer populations may therefore contribute to our understanding of early human history (Atkinson *et al.*, 2009).

3.3.1.4 TMRCAs of African mitochondrial haplogroups

The first attempt at calculating the age of the most recent common ancestor (MRCA) of all surviving human mtDNA lineages, was based on the assumption of a constant mitochondrial mutation rate of between two and four percent per million years (Myr) and thus in turn produced a date of between 140,000 and 290,000 YBP (Cann *et al.*, 1987). Subsequently, using a mutation rate of 2.2 to 2.9% per million years, the age of macrohaplogroup L was calculated to be 98,000 to 130,000 YBP, haplogroup L1 was 86,000 to 113,000 YBP, and haplogroup L2 was 59,000 to 78,000 YBP (Chen *et al.*, 1995).

As additional African mtDNA data became available, updated estimates of the divergence times for all the major African mitochondrial haplogroups and sub-haplogroups were determined by various investigators using different methods based on RFLP data (high resolution and low resolution), control region sequence data (HVS1 and/or HVS2) and coding region sequence data, or alternatively, through the use of a combination of the different data types. A summary of the sequence divergence times for the major African mitochondrial haplogroups and sub-haplogroups is presented in Table 3.9.

Table 3.9 Sequence divergence times for major African mitochondrial haplogroups

Haplogroup/ sub-haplogroup	Estimated TMRCA (YBP)	Type of data used	Reference
Human mtDNA	179,000 to 217,000	Coding region mtDNA sequence data	Mishmar <i>et al.</i> , 2003
	166,000 to 249,000	Control region sequence data	Vigilant <i>et al.</i> , 1991
	161,800 to 226,800	Coding region mtDNA sequence data	Gonder <i>et al.</i> , 2007
	156,000 to 169,000	Complete mtDNA sequence data	Maca-Meyer <i>et al.</i> , 2001
	125,000 to 161,000	Complete mtDNA sequence data	Horai <i>et al.</i> , 1995
	138,000 to 182,000	Coding region mtDNA sequence data	Kivisild <i>et al.</i> , 2006
Human mtDNA (SAK)	122,100 to 230,900	Khoi-San population, 25 individuals, COIII gene sequence data	Wilder <i>et al.</i> , 2004b

Table 3.9 Continued...

Haplogroup/ sub-haplogroup	Estimated TMRCA (YBP)	Type of data used	Reference
Human mtDNA	121,500 to 221,500	Coding region mtDNA sequence data	Ingman <i>et al.</i> , 2000
	111,000 to 148,000	Control region sequence data (16,090 - 16,365)	Watson <i>et al.</i> , 1997
African mtDNA	125,500 to 165,500	High resolution RFLP analysis and CR sequencing	Chen <i>et al.</i> , 2000
	101,000 to 133,000	High resolution RFLP analysis	Chen <i>et al.</i> , 1995
L	122,800 to 161,800	High resolution RFLP analysis and CR sequencing	Chen <i>et al.</i> , 2000
	98,000 to 130,000	High resolution RFLP analysis	Chen <i>et al.</i> , 1995
L0	128,345 to 138,825	Coding region mtDNA sequence data	Mouton, 2003
	125,000 to 159,000	Coding region mtDNA sequence data	Mishmar <i>et al.</i> , 2003
	124,000 to 172,000	Complete sequence data	Atkinson <i>et al.</i> , 2008a
	121,300 to 171,500	Coding region mtDNA sequence data	Gonder <i>et al.</i> , 2007
	92,000 to 124,000	Coding region mtDNA sequence data	Kivisild <i>et al.</i> , 2006
L0a	48,900 to 60,300	Coding region mtDNA sequence data	Gonder <i>et al.</i> , 2007
	48,000 to 56,000	Control region sequence data (16,090 - 16,365)	Watson <i>et al.</i> , 1997
	31,000 to 53,000	Coding region mtDNA sequence data	Kivisild <i>et al.</i> , 2006
	24,100 to 56,600	HVS1 sequence data, RFLPs	Salas <i>et al.</i> , 2002
L0a1	16,750 to 49,950	HVS1 sequence data, RFLPs	Salas <i>et al.</i> , 2002
	3,800 to 9,000	HVS1 and HVS2 sequence data, RFLPs	Rosa <i>et al.</i> , 2004
L0a2	23,000 to 41,000	Coding region mtDNA sequence data	Kivisild <i>et al.</i> , 2006
	4,650 to 11,950	HVS1 sequence data, RFLPs	Salas <i>et al.</i> , 2002
L0a2a	14,000 to 28,000	Coding region mtDNA sequence data	Kivisild <i>et al.</i> , 2006
L0a2b	1,000 to 5,000	Coding region mtDNA sequence data	Kivisild <i>et al.</i> , 2006
L0d	85,800 to 126,200	Coding region mtDNA sequence data	Gonder <i>et al.</i> , 2007
	51,000 to 73,000	Coding region mtDNA sequence data	Kivisild <i>et al.</i> , 2006
	36,150 to 63,050	HVS1 sequence data, RFLPs	Salas <i>et al.</i> , 2002
L0d (San)	71,500 to 109,300	Coding region mtDNA sequence data	Gonder <i>et al.</i> , 2007
L0d (SAK)	18,500 to 145,000	Control region sequence data (HVS1 and HVS2)	Tishkoff <i>et al.</i> , 2007
L0d (Tanzania and SAK)	33,700 to 90,100	Control region sequence data (HVS1 and HVS2)	Tishkoff <i>et al.</i> , 2007

Table 3.9 Continued...

Haplogroup/ sub-haplogroup	Estimated TMRCA (YBP)	Type of data used	Reference
L0d (Tanzania)	12,800 to 48,400	Coding region mtDNA sequence data	Gonder <i>et al.</i> , 2007
	9,400 to 50,700	Control region sequence data (HVS1 and HVS2)	Tishkoff <i>et al.</i> , 2007
L0d1	42,341 to 52,197	Coding region mtDNA sequence data	Mouton, 2003
L0f	85,500 to 104,300	Coding region mtDNA sequence data	Gonder <i>et al.</i> , 2007
L0k	48,000 to 56,000	Control region sequence data (16,090 - 16,365)	Watson <i>et al.</i> , 1997
	7,707 to 15,413	Coding region mtDNA sequence data	Mouton, 2003
L0k (SAK)	51,200 to 90,600	Coding region sequence data	Gonder <i>et al.</i> , 2007
L1	113,100 to 149,100	High resolution RFLP analysis and CR sequencing	Chen <i>et al.</i> , 2000
	107,300 to 173,400	Coding region mtDNA sequence data	Gonder <i>et al.</i> , 2007
	96,000 to 128,000	Coding region mtDNA sequence data	Kivisild <i>et al.</i> , 2006
	87,000 to 139,000	Complete sequence data	Atkinson <i>et al.</i> , 2008a
L1 (includes L0)	86,000 to 113,000	High resolution RFLP analysis	Chen <i>et al.</i> , 1995
L1b	16,000 to 22,000	Control region sequence data (16,090 - 16,365)	Watson <i>et al.</i> , 1997
	14,300 to 46,800	HVS1 sequence data, RFLPs	Salas <i>et al.</i> , 2002
	12,000 to 19,000	HVS1 and HVS2 sequence data	Rando <i>et al.</i> , 1998
L1b1	29,900 to 41,100	HVS1 and HVS2 sequence data, RFLPs	Rosa <i>et al.</i> , 2004
	8,000 to 22,000	Coding region mtDNA sequence data	Kivisild <i>et al.</i> , 2006
L1c	78,550 to 103,950	Control region sequence data	Batini <i>et al.</i> , 2007
	74,000 to 98,000	Coding region mtDNA sequence data	Kivisild <i>et al.</i> , 2006
	71,700 to 119,500	Coding region mtDNA sequence data	Gonder <i>et al.</i> , 2007
	47,850 to 71,450	HVS1 sequence data, RFLPs	Salas <i>et al.</i> , 2002
L1c1	71,020 to 107,380	HVS1 and HVS2 sequence data, RFLPs	Rosa <i>et al.</i> , 2004
	40,600 to 66,100	HVS1 sequence data, RFLPs	Salas <i>et al.</i> , 2002
L1c1a	44,000 to 70,000	Coding region mtDNA sequence data	Kivisild <i>et al.</i> , 2006
L1c2	33,450 to 54,750	HVS1 sequence data, RFLPs	Salas <i>et al.</i> , 2002
L1c3	13,500 to 25,000	HVS1 sequence data, RFLPs	Salas <i>et al.</i> , 2002
L2	90,000 to 99,000	Coding region mtDNA sequence data	Gonder <i>et al.</i> , 2007
	80,100 to 103,700	Coding region mtDNA sequence data	Mishmar <i>et al.</i> , 2003
	73,000 to 127,000	Complete sequence data	Atkinson <i>et al.</i> , 2008a

Table 3.9 Continued...

Haplogroup/ sub-haplogroup	Estimated TMRCA (YBP)	Type of data used	Reference
L2	60,000 to 88,000	Coding region mtDNA sequence data	Kivisild <i>et al.</i> , 2006
	59,000 to 78,000	High resolution RFLP analysis	Chen <i>et al.</i> , 1995
	59,000 to 77,700	High resolution RFLP analysis and CR sequencing	Chen <i>et al.</i> , 2000
	54,800 to 85,400	HVS1 sequence data, RFLPs	Salas <i>et al.</i> , 2002
	53,000 to 59,000	Control region sequence data (16,090 - 16,365)	Watson <i>et al.</i> , 1997
L2a	39,000 to 51,400	High resolution RFLP analysis and CR sequencing	Chen <i>et al.</i> , 2000
	35,800 to 74,500	HVS1 sequence data, RFLPs	Salas <i>et al.</i> , 2002
	24,000 to 40,000	Coding region mtDNA sequence data	Kivisild <i>et al.</i> , 2006
L2a*	21,950 to 48,050	HVS1 and HVS2 sequence data, RFLPs	Rosa <i>et al.</i> , 2004
L2a1	25,050 to 44,150	HVS1 and HVS2 sequence data, RFLPs	Rosa <i>et al.</i> , 2004
	20,300 to 47,100	HVS1 sequence data, RFLPs	Salas <i>et al.</i> , 2002
	9,000 to 15,000	Coding region mtDNA sequence data	Kivisild <i>et al.</i> , 2006
L2a2	22,000 to 40,000	Coding region mtDNA sequence data	Kivisild <i>et al.</i> , 2006
L2b	24,800 to 32,700	High resolution RFLP analysis and CR sequencing	Chen <i>et al.</i> , 2000
	23,000 to 41,000	Coding region mtDNA sequence data	Kivisild <i>et al.</i> , 2006
	20,400 to 42,800	HVS1 sequence data, RFLPs	Salas <i>et al.</i> , 2002
	17,350 to 59,250	HVS1 and HVS2 sequence data, RFLPs	Rosa <i>et al.</i> , 2004
L2b1	3,900 to 11,500	HVS1 and HVS2 sequence data, RFLPs	Rosa <i>et al.</i> , 2004
L2c	20,250 to 34,750	HVS1 sequence data, RFLPs	Salas <i>et al.</i> , 2002
	17,900 to 23,600	High resolution RFLP analysis and CR sequencing	Chen <i>et al.</i> , 2000
	13,400 to 23,000	HVS1 and HVS2 sequence data, RFLPs	Rosa <i>et al.</i> , 2004
	5,000 to 11,000	Coding region mtDNA sequence data	Kivisild <i>et al.</i> , 2006
L2d	87,700 to 156,100	HVS1 sequence data, RFLPs	Salas <i>et al.</i> , 2002
	62,200 to 116,600	HVS1 and HVS2 sequence data, RFLPs	Rosa <i>et al.</i> , 2004
L3	78,300 to 103,200	High resolution RFLP analysis and CR sequencing	Chen <i>et al.</i> , 2000
	74,600 to 79,400	Control region sequence data (16,090 - 16,365)	Watson <i>et al.</i> , 1997
	57,000 to 73,000	Coding region mtDNA sequence data	Kivisild <i>et al.</i> , 2006
	49,650 to 72,950	HVS1 sequence data, RFLPs	Salas <i>et al.</i> , 2002

Table 3.9 Continued...

Haplogroup/ sub-haplogroup	Estimated TMRCA (YBP)	Type of data used	Reference
L3b	41,000 to 47,000	Control region sequence data (16,090 - 16,365)	Watson <i>et al.</i> , 1997
	20,150 to 44,150	HVS1 and HVS2 sequence data, RFLPs	Rosa <i>et al.</i> , 2004
	16,000 to 32,000	Coding region mtDNA sequence data	Kivisild <i>et al.</i> , 2006
	14,750 to 28,450	HVS1 sequence data, RFLPs	Salas <i>et al.</i> , 2002
L3b1	7,100 to 26,500	HVS1 and HVS2 sequence data, RFLPs	Rosa <i>et al.</i> , 2004
L3d	32,000 to 46,000	Coding region mtDNA sequence data	Kivisild <i>et al.</i> , 2006
	31,500 to 52,700	HVS1 and HVS2 sequence data, RFLPs	Rosa <i>et al.</i> , 2004
	21,800 to 38,700	HVS1 sequence data, RFLPs	Salas <i>et al.</i> , 2002
	20,000 to 35,000	HVS1 and HVS2 sequence data	Rando <i>et al.</i> , 1998
L3d1	20,000 to 34,000	Coding region mtDNA sequence data	Kivisild <i>et al.</i> , 2006
L3d2	0 to 20,200	HVS1 sequence data	Kivisild <i>et al.</i> , 2004
L3d3	5,000 to 17,000	Coding region mtDNA sequence data	Kivisild <i>et al.</i> , 2006
L3e	37,500 to 61,000	HVS1 sequence data, RFLPs	Salas <i>et al.</i> , 2002
	31,500 to 52,700	HVS1 and HVS2 sequence data, RFLPs	Rosa <i>et al.</i> , 2004
	23,000 to 37,000	Coding region mtDNA sequence data	Kivisild <i>et al.</i> , 2006
L3e1	20,700 to 43,600	HVS1 sequence data, RFLPs	Salas <i>et al.</i> , 2002
L3e1a	14,750 to 38,750	HVS1 sequence data, RFLPs	Salas <i>et al.</i> , 2002
L3e2	19,050 to 55,750	HVS1 sequence data, RFLPs	Salas <i>et al.</i> , 2002
L3e2*	4,050 to 12,150	HVS1 and HVS2 sequence data, RFLPs	Rosa <i>et al.</i> , 2004
L3e2b	6,050 to 12,250	HVS1 sequence data, RFLPs	Salas <i>et al.</i> , 2002
	4,300 to 15,900	HVS1 and HVS2 sequence data, RFLPs	Rosa <i>et al.</i> , 2004
L3e3	9,650 to 18,650	HVS1 sequence data, RFLPs	Salas <i>et al.</i> , 2002
L3e4	13,800 to 34,600	HVS1 sequence data, RFLPs	Salas <i>et al.</i> , 2002
	5,800 to 16,200	HVS1 and HVS2 sequence data, RFLPs	Rosa <i>et al.</i> , 2004
L3f*	35,800 to 65,200	HVS1 sequence data	Kivisild <i>et al.</i> , 2004
L3f	23,500 to 49,100	HVS1 sequence data, RFLPs	Salas <i>et al.</i> , 2002
L3f1	33,150 to 65,550	HVS1 and HVS2 sequence data, RFLPs	Rosa <i>et al.</i> , 2004
	20,000 to 37,300	HVS1 sequence data, RFLPs	Salas <i>et al.</i> , 2002
	5,100 to 15,100	HVS1 sequence data	Kivisild <i>et al.</i> , 2004
L3h	5,600 to 22,400	HVS1 and HVS2 sequence data, RFLPs	Rosa <i>et al.</i> , 2004
L3w	1,700 to 9,900	HVS1 sequence data	Kivisild <i>et al.</i> , 2004
L3x	47,100 to 90,100	HVS1 sequence data	Kivisild <i>et al.</i> , 2004

Table 3.9 Continued...

Haplogroup/ sub-haplogroup	Estimated TMRCA (YBP)	Type of data used	Reference
L4	50,500 to 87,100	HVS1 sequence data	Kivisild <i>et al.</i> , 2004
L4a	34,000 to 68,400	HVS1 sequence data	Kivisild <i>et al.</i> , 2004
L4g	51,100 to 79,500	HVS1 sequence data	Salas <i>et al.</i> , 2004b
	32,600 to 57,600	HVS1 sequence data, RFLPs	Salas <i>et al.</i> , 2002
	27,800 to 53,000	HVS1 sequence data	Kivisild <i>et al.</i> , 2004
	11,200 to 47,600	Control region sequence data	Tishkoff <i>et al.</i> , 2007
L5	107,300 to 151,500	Coding region mtDNA sequence data	Gonder <i>et al.</i> , 2007
L5	69,700 to 122,100	HVS1 sequence data	Kivisild <i>et al.</i> , 2004
L5a	58,050 to 107,850	HVS1 sequence data, RFLPs	Salas <i>et al.</i> , 2002
L6	13,300 to 64,500	HVS1 sequence data	Kivisild <i>et al.</i> , 2004

L0a, L0d, L0f and L0k were previously known as L1a, L1d, L1f and L1k, L5a was formerly identified as L1e, L4a has also been designated as L7 (Torroni *et al.*, 2006), and L4g was previously designated as L3g (Kivisild *et al.*, 2004; Salas *et al.*, 2004a; Tishkoff *et al.*, 2007). The asterisks (*) included in certain haplogroup designations (L2a*, L3e2* and L3f*) are used to indicate that the specific haplogroup consists of a group of mtDNA lineages that form part of the particular haplogroup, although these lineages can not be classified into any of the specific sub-haplogroups of that haplogroup. No divergence times are provided for sub-haplogroups L3i and L5b as only a small number of individuals have been identified that belong to these lineages (Kivisild *et al.*, 2004). TMRCA = time to most recent common ancestor; YBP = years before present; mtDNA = mitochondrial DNA; COIII = cytochrome c oxidase subunit 3; RFLP = restriction fragment length polymorphism; HVS1 = hypervariable segment 1; HVS2 = hypervariable segment 2; CR = control region; SAK = Southern African Khoi-San.

Calculation of sequence diversity has revealed that macrohaplogroup L contains the highest sequence diversity of the continent-specific haplogroups and that Africa is the continent with the greatest diversity (Wallace *et al.*, 1999). The African origin of modern humans is supported by the estimated age of origin of macrohaplogroup L as it predates the expansion of *Homo sapiens sapiens* from Africa (Chen *et al.*, 1995).

3.3.2 Asian mitochondrial haplogroups

The Asian population can be divided into two major mitochondrial lineages based on the absence or presence of *Dde* I and *Alu* I recognition sites at nucleotide positions 10,394 and 10,397 respectively (Torroni *et al.*, 1994b). The major mitochondrial lineage defined by the presence of the aforementioned restriction enzyme sites has been designated macrohaplogroup M, while the mitochondrial lineage in which these restriction enzyme sites are absent is referred to as macrohaplogroup N. These two main lineages contributed equally to mtDNA radiation in Asia with macrohaplogroup M giving rise to Asian-specific haplogroups C, D, E, G, Z and macrohaplogroup N generating Asian-specific haplogroups A, B, F and Y (Kivisild *et al.*, 2002). Of these haplogroups, haplogroups A, B, C and D are also present in the Native American population (Torroni *et al.*, 1993a). The Asian mtDNA haplogroups are defined through the use of specific restriction enzymes as listed in Table 3.10.

Table 3.10 Restriction enzyme sites used to define continent-specific Asian mtDNA haplogroups

Population	Haplogroup	mtDNA variants	Continental mtDNA (%)	Reference
Asia	M	10,394 + <i>Dde</i> I 10,397 + <i>Alu</i> I	57	Wallace <i>et al.</i> , 1999
Asia	N	10,394 - <i>Dde</i> I 10,397 - <i>Alu</i> I	---	Wallace <i>et al.</i> , 1999
Asia	A	663 + <i>Hae</i> III 10,394 - <i>Dde</i> I 10,397 - <i>Alu</i> I	12	Torrioni <i>et al.</i> , 1992
Asia	B	8,271 - 8,281 9 bp deletion 10,394 - <i>Dde</i> I 10,397 - <i>Alu</i> I 16,517 + <i>Hae</i> III	4	Torrioni <i>et al.</i> , 1992
Asia	C	10,394 + <i>Dde</i> I 10,397 + <i>Alu</i> I 13,259 - <i>Hinc</i> II 13,262 + <i>Alu</i> I	21	Torrioni <i>et al.</i> , 1992
Asia	D	5,176 - <i>Alu</i> 10,394 + <i>Dde</i> I 10,397 + <i>Alu</i> I	14	Torrioni <i>et al.</i> , 1992
Asia	E	7,598 - <i>Hpa</i> I 10,394 + <i>Dde</i> I 10,397 + <i>Alu</i> I	---	Torrioni <i>et al.</i> , 1994b
Asia	F	10,394 - <i>Dde</i> I 10,397 - <i>Alu</i> I 12,406 - <i>Hpa</i> I/ <i>Hinc</i> II 16,517 + <i>Hae</i> III	4	Torrioni <i>et al.</i> , 1994b
Asia	G	4,830 + <i>Hae</i> III 4,831 + <i>Hpa</i> I 10,394 + <i>Dde</i> I 10,397 + <i>Alu</i> I	---	Torrioni <i>et al.</i> , 1994b
Asia/ Siberia	Y	7,933 + <i>Mbo</i> I 8,391 - <i>Hae</i> III 10,394 + <i>Dde</i> I 16,517 + <i>Hae</i> III	---	Schurr <i>et al.</i> , 1999
Asia/ Siberia	Z	10,394 - <i>Dde</i> I 10,397 + <i>Alu</i> I 11,074 + <i>Dde</i> I 16,517 + <i>Hae</i> III	---	Schurr <i>et al.</i> , 1999

bp = base pairs. mtDNA variant information was obtained from Wallace *et al.* (1999) and the references as provided, while the continental mtDNA percentages were obtained from MITOMAP (2006).

All of the Asian-specific mitochondrial haplogroups listed in Table 3.10 demonstrate distinct frequency variations in Asian, Native American and Siberian populations. Haplogroups A, C, D, E and G are absent in southern Asian populations such as the Malays, Vietnamese, Sabah Aborigines and Malay Aborigines, and are present at significant frequencies in northern Asian populations including Tibetans, Koreans, and Han

Chinese, while haplogroup F is prominent in southern Asian populations (Wallace *et al.*, 1999). This difference in haplogroup distribution between the north and the south supports the division of the Asian population into a northern and southern component. The distribution of the major Asian mitochondrial haplogroups is indicated in Table 3.11.

Table 3.11 Distribution of haplogroups A, B, C, D, E, F and G

Haplogroup	Population or geographic region
A	Amerinds (Northern, Central and Southern), Asians (Koreans, Taiwanese Han, Malaysian Han, Tibetans), Na-Dene (Northern and Southern), Siberians (Eskimos, Chukchi, Koryaks, Evenks, Naganasans)
B	Amerinds (Northern, Central and Southern), Asians (Koreans, Taiwanese Han, Vietnamese, Malays, Sabah Aborigines, Malay Aborigines, Tibetans), Melanesians (New Guineans), Na-Dene (Southern)
C	Amerinds (Northern, Central and Southern), Asians (Taiwanese Han, Tibetans), Na-Dene (Southern), Siberians (Chukchi, Koryaks, Yukagirs, Evens, Udegeys, Evenks, Naganasans, Sel'kups)
D	Amerinds (Northern, Central and Southern), Asians (Koreans, Taiwanese Han, Malaysian Han, Tibetans), Melanesians (New Guineans), Na-Dene (Northern and Southern), Siberians (Eskimos, Chukchi, Koryaks, Yukagirs, Evens, Nivkhs, Evenks, Naganasans)
E	Asians (Koreans, Tibetans)
F	Asians (Koreans, Taiwanese Han, Vietnamese, Malays, Sabah Aborigines, Malay Aborigines, Tibetans), Melanesians (New Guineans)
G	Asians (Koreans, Tibetans), Siberians (Nivkhs)

Adapted from Torroni *et al.* (1994b).

The estimated coalescent dates of the major Asian mtDNA haplogroups have been determined based on the sequence diversity present in each haplogroup. The approximate ages for the major Asian mitochondrial haplogroups are indicated in Table 3.12.

Table 3.12 Divergence times of the major Asian mitochondrial haplogroups

Haplogroup	Divergence time (YBP)
M	63,900 to 74,700
N	57,800 to 71,400
A	24,700 to 39,900
B	44,200 to 57,400
C	22,800 to 33,800
D	49,200 to 65,600
F	50,800 to 69,200

YBP = years before present. Adapted from Mishmar *et al.* (2003) and Kong *et al.* (2003).

3.3.3 Native American mitochondrial haplogroups

The mtDNAs of Native American populations in North, Central, and South America can be assigned almost exclusively to five mtDNA haplogroups, namely haplogroups A, B, C, D

and X (Wallace *et al.*, 1999). Four of these haplogroups, namely haplogroups A, B, C and D are shared with Asian populations, which indicate that the Native American mitochondrial haplogroups originated in Asia (Schurr *et al.*, 1990; Torroni *et al.*, 1992; Torroni *et al.*, 1993a; Torroni *et al.*, 1993b). The Native American mtDNA haplogroups A, B, C and D were originally defined through the use of the restriction enzymes listed in Table 3.13.

Table 3.13 Restriction enzyme sites used to define continent-specific Native American mtDNA haplogroups

Population	Haplogroup	mtDNA variants	Continental mtDNA (%)	Reference
America	A (Amerinds)	663 +Hae III 10,394 -Dde I 10,397 -Alu I	44	Torroni <i>et al.</i> , 1992
America	A (Na-Dene)	663 +Hae III 10,394 -Dde I 10,397 -Alu I 16,329 -Rsa I	29	Torroni <i>et al.</i> , 1992
America	B (Amerinds)	8,271 to 8,281 9 bp deletion 10,394 -Dde I 10,397 -Alu I 16,517 +Hae III	22	Torroni <i>et al.</i> , 1992
America	C (Amerinds)	10,394 +Dde I 10,397 +Alu I 13,259 -Hinc II 13,262 +Alu I	18	Torroni <i>et al.</i> , 1992
America	D (Amerinds)	5,176 -Alu 10,394 +Dde I 10,397 +Alu I	16	Torroni <i>et al.</i> , 1992
America	X (Amerinds)	1,715 -Dde I 10,394 -Dde I 14,465 +Ava II 16,517 +Hae III	---	Torroni <i>et al.</i> , 1996

bp = base pairs. mtDNA variant information was obtained from Wallace *et al.* (1999) and the references as provided, while the continental mtDNA percentages were obtained from MITOMAP (2006).

Haplogroups A, B, C, D and X in Native Americans have since been renamed to haplogroups A2, B2, C1, D1 and X2a (Bandelt *et al.*, 2003). Haplogroups A2, B2, C1 and D1 are observed in North, Central and South America, while the geographic distribution of X2a is limited to some Amerindian populations located in northern North America (Forster *et al.*, 1996; Brown *et al.*, 1998; Bandelt *et al.*, 2003; Reidla *et al.*, 2003). Subsequently, two more Native American specific haplogroups were identified namely D2a in the Aleuts and Eskimos (Derbeneva *et al.*, 2002; Tamm *et al.*, 2007) and D3 only in the Eskimos (Saillard *et al.*, 2000; Helgason *et al.*, 2006). Recently, an additional two uncommon Native American specific haplogroups have been identified, namely D4h3 and C4c (Kemp *et al.*,

2007; Tamm *et al.*, 2007) bringing the total number of Native American haplogroups to nine.

With the advent of frequent complete mitochondrial genome sequencing, the haplogroup classification scheme based on RFLPs presented in Table 3.13 has been replaced by the use of a combination of HVS1, HVS2 and coding region motifs consisting of informative nucleotide positions. The current list of defining substitutional motifs for Native American specific haplogroups is indicated in Table 3.14.

Table 3.14 Single nucleotide polymorphisms used to define continent-specific Native American mtDNA haplogroups

Haplogroup	HVS1 motif	HVS2 motif	Coding region motif
A	16223-16290-16319	73-235-263	663-1736-4248-4824-8794
A2	16111-16223-16290-16319-16362	64-73-146-153-235-263	8027-12007
A2a	16111-16223-16290-16319-16362	64-73-146-153-235-263	3330
B	16189	73-263	8281-8289d
B2	16189-16217	73-263	3547-4977-6473-9950-11177
B4b	16189-16217	73-263	499-4820-13590
B4bd	16189-16217	73-263	827-15535
C	16223-16298-16327	73-249d-263	3552A-9545-11914-13263-14318
C1	16223-16298-16325-16327	73-249d-263-290-291d	---
C1b	16223-16298-16325-16327	73-249d-263-290-291d	493
C1c	16223-16298-16325-16327	73-249d-263-290-291d	1888-15930
C1d	16223-16298-16325-16327	73-249d-263-290-291d	7697
C4	16223-16298-16327	73-249d-263	2232iA-6026-11969-15204
C4c	16223-16245-16298-16327	73-263	11440-13368-14433-15148
D	16223-16362	73-263	4883-5178A
D1	16223-16325-16362	73-263	2092
D2	16129-16223-16271-16362	73-263	3316-7493-8703-9536-11215
D2a	16129-16223-16271-16362	73-263	11959
D2b	16129-16223-16271-16362	73-263	9181
D3	16223-16319-16362	73-263	951-8020-10181-15440-15951
D4	16223-16362	73-263	3010-8414-14668
D4h3	16223-16241-16301-16342-16362	73-263	3336-3396-3644-5048-6285-8949-9458-13135
X	16189-16223-16278	73-153-263	6221-6371-13966-14470
X2a	16189-16213-16223-16278	73-153-195-200-263	1719-8913-12397-14502

HVS1 = hypervariable segment 1; HVS2 = hypervariable segment 2; i = insertion; d = deletion. Substitutions at nucleotide positions in the motifs represent transitions unless otherwise indicated. Adapted from Tamm *et al.* (2007).

Divergence times for the Native American haplogroups have been estimated based on coding region sequence data. The estimated coalescent times for each Native American haplogroup is indicated in Table 3.15.

Table 3.15 Sequence divergence times for Native American mitochondrial haplogroups

Haplogroup	Divergence time (YBP)
A2	15,500 to 18,900
B2	18,800 to 23,600
C	11,500 to 23,000
C1b	14,700 to 26,500
C1c	16,500 to 23,700
C1d	8,000 to 13,800
D1	16,300 to 20,900
D2	4,200 to 16,400
X2a	6,100 to 29,900

YBP = years before present. Adapted from Tamm *et al.* (2007), Achilli *et al.* (2008) and Fagundes *et al.* (2008).

The widespread distribution of haplogroups A to D in the New World, as well as Asia, coupled with the similarity in coalescence times for these haplogroups, allowed the suggestion of a single-migration model for the peopling of the New World (Forster *et al.*, 1996; Bonatto and Salzano, 1997; Stone and Stoneking, 1998). However, in some studies, differences were identified in the pattern of diversification as well as dispersion of haplogroup B. This included differences in the divergence times between haplogroup B and the average divergence times of haplogroups A, C and D, as well as the fact that haplogroup B is widespread throughout southern and central Asia, prominent in coastal Asian populations, virtually absent in Siberia, rare in northern North America, while reappearing in North, Central, and South American Native American populations (Stoneking *et al.*, 1990; Schurr *et al.*, 1990; Ballinger *et al.*, 1992; Torroni *et al.*, 1993a; Torroni *et al.*, 1993b). These differences suggest that Native American populations may have originated from two separate migrations, i.e. a multiple-migration model. The first migration originated from central Asia moving through Siberia carrying haplogroups A, C, and D, while the second migration occurred much later, bringing haplogroup B into the Americas through a different route which bypassed Siberia while moving along the Siberian and Alaskan coasts (Torroni *et al.*, 1994c; Starikovskaya *et al.*, 1998; Wallace *et al.*, 1999).

The unravelling of the colonisation history of the New World based on mitochondrial haplogroups is further complicated by the distribution and divergence of haplogroup X.

Haplogroup X is detected in Native Americans at a relatively low frequency, and only in North America, whereas it is nearly absent in Siberia and uncommon in West Eurasians (Smith *et al.*, 1999; Derenko *et al.*, 2001; Reidla *et al.*, 2003; Dornelles *et al.*, 2005). In addition, it has been proposed that haplogroup X in Native Americans is less diverse with a younger estimated coalescent date than haplogroups A to D (Brown *et al.*, 1998). These characteristics of haplogroup X have been used to support the suggestion that haplogroup X represents a separate migration to the Americas from Europe or Asia (Brown *et al.*, 1998). Specifically, this suggestion has been employed to lend support to the greatly contested “Solutrean hypothesis” which suggests that a link exists between the American Clovis and European Solutrean lithic technologies, and that North America was colonised by Europeans through the North Atlantic (Meltzer, 2004; Straus *et al.*, 2005). In a recent investigation, it was demonstrated that mitochondrial haplogroups A, B, C, D and X formed part of a single founding population which eventually resulted in the colonisation of the American continent, thereby disproving multiple-migration models such as the “Solutrean hypothesis” (Fagundes *et al.*, 2008). A complex model for the peopling of the New World was proposed in which the initial divergence from Asian populations concluded with a mild bottleneck in Beringia during the last glacial maximum (ca. 19,000 to 23,000 YBP) followed by a rapid population expansion along the Pacific coast toward the end of the last glacial maximum (15,000 to 18,000 YBP). These results provide support for a pre-Clovis habitation of the New World (Fagundes *et al.*, 2008).

3.3.4 Oceanic mitochondrial haplogroups

The mtDNAs of Oceanic populations which include populations residing in the Pacific Islands (Melanesia, Micronesia and Polynesia), Papua New Guinea, New Zealand and Australia can be allocated to two major region-specific mtDNA haplogroups, namely haplogroups P and Q (Redd and Stoneking, 1999; Forster *et al.*, 2001; Friedlaender *et al.*, 2005). It has been suggested that haplogroup P is the older of the two haplogroups as it is more widely distributed and is more heterogeneous than haplogroup Q (Friedlaender *et al.*, 2005). With the exception of one shared sub-haplogroup of P, different sub-haplogroups of P are identified in Australia and Papua New Guinea, while haplogroup Q is not present in Australia although it is extremely widespread in Papua New Guinea and Melanesia (Friedlaender *et al.*, 2005). This distribution pattern suggests that Australia has been isolated from Papua New Guinea/Melanesia since around the period of first settlement.

A third haplogroup, namely haplogroup S, has been identified only in Australians (Van Holst Pellekaan *et al.*, 2006). Haplogroups P, Q and S diverged from the two major out of Africa mtDNA haplogroups with haplogroup P descending from the R branch of macrohaplogroup N, haplogroup S arising from the Non-R branch of macrohaplogroup N, and haplogroup Q originating from macrohaplogroup M (Friedlaender *et al.*, 2005). The major Oceanic mtDNA haplogroups and their sub-haplogroups are defined through the use of non-coding and coding region motifs as listed in Table 3.16.

Table 3.16 Single nucleotide polymorphisms used to define continent-specific Oceanic mtDNA haplogroups

Haplogroup/ sub-haplogroup	HVS1 motif	HVS2 motif	Coding region motif	Continental mtDNA (%)
P	---	---	15607	11
P1	16266-16357	212	6077-10118	8
P2	---	---	3203-3882-4122-8859-14890	3
P3	---	127-128-573iCCC	3645-14338-15748-15937T	---
P4	---	---	1719-5460-10398	---
P5	<i>16192-16311</i>	<i>152</i>	<i>7419-7657-9509-12406-12831-13368-15412-15970</i>	---
P6	16311-16362	---	6719	---
P7	<i>16148-16234</i>	<i>152-374</i>	<i>1598-2283-6249-7325-8387-10373-10810-12026-12235-12361-14502-15613</i>	---
P9	<i>16241T-16261-16362</i>	<i>152</i>	<i>3391-5237-7805-8167-8614-11065-11353-11404-12414-12756-15110-15511-15521C</i>	---
Q	16129-16241	---	4117-5843-8790-12940-13500	24
Q1	16144-16148-16265C-16343	89-92-146	8964-14025	16
Q2	16066-16311	228T	11061	7
Q3	---	---	4335-15172	1
S	---	---	8404	---
S1	16075	---	14384C	---
S2	---	---	2380-3438-6167	---
S3	16140-16187-16189-16287-16311	63-64-66-151-152-195	710-5302-5492-8277-8279iCCC-9938-10398-10786-11151-15204-15317-15663-15664	---
S4	16172-16192-16319	152	1007-4181-5090-6528-7498-9866	---
S5	16243-16256	152-234-249d-508	1375-3169-3593-3996-8027-10188-11470-13770A-14121	---

HVS1 = hypervariable segment 1; HVS2 = hypervariable segment 2; i = insertion; d = deletion. Nucleotide substitutions are transitions unless indicated otherwise. Nucleotide positions presented in *italic* text indicate reversions to the nucleotide observed in the rCRS. Continental mtDNA percentages were determined based on the data published by Friedlaender *et al.* (2005). Adapted from Friedlaender *et al.* (2005), Pierson *et al.* (2006) and Van Oven and Kayser (2009).

Divergence times for the major Oceanic haplogroups have been calculated based on coding region sequence data. The estimated coalescent times for some of the major Oceanic haplogroups and sub-haplogroups are indicated in Table 3.17.

Table 3.17 Divergence times of some of the major Oceanic mitochondrial haplogroups

Haplogroup/sub-haplogroup	Divergence time (YBP)
P	28,600 to 51,200
P1	35,800 to 49,800
P2	13,700 to 18,900
P3	50,400 to 68,800
Q	37,400 to 51,600
Q1	21,700 to 32,100
Q2	32,400 to 56,600
Q3	24,300 to 32,900
S	10,500 to 32,400

YBP = years before present. Adapted from Van Holst Pellekaan *et al.* (2006) and Friedlaender *et al.* (2007).

Studies on ancient Aboriginal Australian remains of an anatomically modern human belonging to the terminal Pleistocene or early Holocene periods have suggested that the remains represent the oldest human mitochondrial ancestor (Adcock *et al.*, 2001). It was further suggested that the sequences obtained diverged before the MRCA of sequences in extant Aboriginal Australians, which form deeper roots in the mtDNA tree than African sequences. This observation challenged the Recent Out of Africa model (refer to Section 3.1.1) of human evolution as the results implied that Australia was inhabited by anatomically modern humans prior to the complete fixation of the mtDNA lineage currently present in all living people. This however does not necessarily indicate that humans did not originate from Africa. This observation could be explained by an early, unsuccessful migration out of Africa which gave rise to the Australian mitochondrial lineage observed by Adcock *et al.* (2001). This African lineage, however might not have survived within Africa, possibly due to a lack of surviving offspring or the production of only male offspring, and when a subsequent migration from Africa took place, another lineage gave rise to all the mtDNA lineages observed presently in all living people.

A recent investigation by Kumar *et al.* in 2009 using complete mtDNA sequence data attempted to elucidate the colonisation process of Australia by anatomically modern humans. The results generated in the study demonstrated that Australians and New Guineans exclusively belong to the out of Africa founder macrohaplogroups M and N (Kumar *et al.*, 2009). This indicates that Australians and New Guineans descend from the

same small group of African hunter-gatherers that left Africa ca. 50,000 to 70,000 YBP similar to all other Eurasians. In addition, a shared mtDNA lineage between Indians and Australian Aborigines was identified which provided direct genetic evidence that anatomically modern humans populated Australia through South Asia via the “Southern Route” (Kumar *et al.*, 2009). The divergence, based on coding region sequence data, observed between the Indian and Australian mitochondrial sequences provided evidence for an early colonisation of Australia ca. 50,000 to 60,000 YBP which is in agreement with archaeological evidence (O’Connell and Allen, 2004).

3.3.5 European mitochondrial haplogroups

Early investigations of the Caucasian population in Europe revealed that the population could be divided into two main mitochondrial lineages based on the presence or absence of a *Dde* I recognition site at nucleotide position 10,394 (Torroni *et al.*, 1994a). Of these European mtDNAs analysed, 64% belonged to four mtDNA haplogroups, namely haplogroups H, I, J, and K (Torroni *et al.*, 1994a). A further study of three diverse European populations (from Finland, Sweden and Tuscany) revealed that nearly all the mtDNA sequences grouped into ten mtDNA haplogroups, namely haplogroups H, I, J, K, M, T, U, V, W and X (Torroni *et al.*, 1996). Haplogroups H, T, U, V, W and X lack the *Dde* I recognition site at nucleotide position 10,394, while haplogroups I, J, and K retain the *Dde* I site.

The classification of European mitochondrial sequences into ten haplogroups as described in the previous paragraph, suggests that European mtDNAs belong to distinct haplogroups that originated following the migration of the ancestors of the Caucasians from Africa and Asia more than 80,000 YBP (Torroni *et al.*, 1996). In addition, it has been established that haplogroups H, I, J, K, T, U, V, W and X are essentially specific to European populations (Torroni *et al.*, 1996). These European mtDNA haplogroups are defined through the use of restriction enzymes as listed in Table 3.18.

Table 3.18 Restriction enzyme sites used to define continent-specific European mtDNA haplogroups

Haplogroup	mtDNA variants	Continental mtDNA (%)	Reference
H	7,025 - <i>Alu</i> I 10,349 - <i>Dde</i> I	44.8	Torrioni <i>et al.</i> , 1994a
I	1,715 - <i>Dde</i> I 8,249 + <i>Ava</i> II 10,028 + <i>Alu</i> I 10,394 + <i>Dde</i> I	2.5	Torrioni <i>et al.</i> , 1994a
J	10,394 + <i>Dde</i> I 13,704 - <i>Bst</i> NI	10.2	Torrioni <i>et al.</i> , 1994a
K	9,052 - <i>Hae</i> II 10,394 + <i>Dde</i> I 12,308 + <i>Hin</i> fI	6.0	Torrioni <i>et al.</i> , 1994a Torrioni <i>et al.</i> , 1996
T	10,394 - <i>Dde</i> I 13,366 + <i>Bam</i> HI 15,606 + <i>Alu</i> I	8.5	Torrioni <i>et al.</i> , 1996
U	10,394 - <i>Dde</i> I 12,308 + <i>Hin</i> fI	15.7	Torrioni <i>et al.</i> , 1996
V	4,577 - <i>Nla</i> III 10,394 - <i>Dde</i> I	6.2	Torrioni <i>et al.</i> , 1996
W	8,249 + <i>Ava</i> II 8,994 - <i>Hae</i> III 9,052 + <i>Hae</i> II 10,394 - <i>Dde</i> I	1.9	Torrioni <i>et al.</i> , 1996 Helgason <i>et al.</i> , 2001
X	1,715 - <i>Dde</i> I 10,394 - <i>Dde</i> I	1.6	Torrioni <i>et al.</i> , 1996

mtDNA variant information was obtained from Wallace *et al.* (1999) and continental mtDNA percentages were determined based on the data presented by Helgason *et al.* (2001).

The approximate ages of the major European mtDNA haplogroups have been calculated based on the sequence diversity present in each haplogroup and by using an estimated mtDNA sequence evolution rate of 2.2 - 2.9% per million years. The approximate ages for the major European mitochondrial haplogroups are presented in Table 3.19.

Table 3.19 Divergence times of the major European mitochondrial haplogroups

Haplogroup	Divergence time (YBP)
H	22,000 to 30,000
I	26,000 to 34,000
J	17,000 to 23,000
K	13,000 to 19,000
T	8,000 to 11,000
U	51,000 to 67,000
V	16,000 to 22,000

YBP = years before present. Adapted from Wallace (1995), Torrioni *et al.* (1996) and Wallace *et al.* (1999).

Continent-specific haplogroups are present in extremely low frequencies on other continents, e.g. haplogroup U has been identified at low frequencies in African populations (Torroni *et al.*, 1996). This implies that there is a limited overlap in variation between European and non-European mtDNAs. The observed mitochondrial haplogroups are also not uniformly distributed across Caucasian populations and there is a degree of enrichment for specific haplogroups in different areas of Europe (Torroni *et al.*, 1994a). This characteristic of European mitochondrial haplogroup distribution is exploited in studies to determine through which migratory routes, at what approximate times and from which refuge areas Europe was re-populated following successive glacial events (Torroni *et al.*, 2006).

Haplogroup H is defined by transitions at nucleotide positions 2,706 and 7,028 (Van Oven and Kayser, 2009). Although haplogroup H is the most frequently observed haplogroup in Europe (40 to 50%), it is one of the most inadequately, phylogenetically characterised haplogroups based on control region sequence data of HVS1 and HVS2 (Pereira *et al.*, 2005). Sequence data of the complete mitochondrial genome has however in recent years revealed a number of informative coding-region polymorphisms that allow for the identification of distinct sub-haplogroups within haplogroup H (Pereira *et al.*, 2005). The classification of haplogroup H into sub-haplogroups is presented in Table 3.20.

Table 3.20 Classification of haplogroup H into sub-haplogroups

Sub-haplogroup	Sequence motif	Sub-haplogroup	Sequence motif
H1	3010	H14	7645-10217
H2	1438-4769	H15	55-57-6253
H3	6776	H16	10394
H4	3992-5004-9123	H17	3915-16129
H5	456-16304	H18	13708-14364
H6	239-16362-16482	H19	6272-14869
H7	4793	H20	16218
H8	146-195-709-13101C-16288	H21	186-3822-8994-16192
H9	152-13020	H22	16145-16227
H10	14470A	H23	10211
H11	195-8448-13759-16311	H25	9620
H12	3936-14552-16287	H30	195-504-8200-8638-8817-14241-16192
H13	14872	---	---

Adapted from Roostalu *et al.* (2007), Ennafaa *et al.* (2009) and Van Oven and Kayser (2009).

Using phylogenetic analysis and other available data, age estimates were determined for some of the most common H sub-haplogroups and are presented in Table 3.21. Due to the

differences observed in the estimated ages of the H sub-haplogroups, they may be informative for investigation of different time scales in European demographic history.

Table 3.21 Age estimates determined for sub-haplogroups of haplogroup H

Sub-haplogroup	Age estimate (years before present)
H1	14,900 to 30,300
H2	11,100 to 18,300
H3	7,900 to 24,100
H4	5,700 to 17,300
H5	8,500 to 16,900
H6	8,800 to 19,200
H7	8,700 to 23,500
H8	5,300 to 22,900
H10	7,700 to 15,500
H11	25,400 to 62,400
H17	8,200 to 26,400

Adapted from Loogväli *et al.* (2004) and Pereira *et al.* (2004).

Haplogroup U is defined by transitions at nucleotide positions 11,467, 12,308 and 12,372 (Van Oven and Kayser, 2009). Haplogroup U can be further divided into nine sub-haplogroups as indicated in Table 3.22.

Table 3.22 Classification of haplogroup U into sub-haplogroups

Sub-haplogroup	Sequence motif
U1	285-12879-13104-14070-15148-15954C-16249
U2	1811-9055-12308-12372
U3	150-1811-14139-15454-16343
U4	195-499-1811-4646-5999-6047-11332-14620-15693-16356
U5	3197-9477-13617-16192-16270
U6	3348-16172
U7	152-980-1811-5360-10142-16318T
U8	1811-9698
U9	499-1811-3531-3834-5999-6386-14094

Adapted from Herrnstadt *et al.* (2002a) and Van Oven and Kayser (2009).

The ages of some of the sub-haplogroups of haplogroup U has been determined. The estimated ages are presented in Table 3.23.

Table 3.23 Age estimates determined for selected sub-haplogroups of haplogroup U

Sub-haplogroup	Age estimate (years before present)
U2	23,600 to 48,000
U3	11,900 to 26,800
U4	16,100 to 24,700
U5	45,100 to 52,800
U7	11,900 to 45,400

Adapted from Richards *et al.* (2000).

Haplogroup J is identified by alterations at nucleotide positions 295, 489, 10,398, 12,612, 13,708 and 16,069 (Van Oven and Kayser, 2009). Haplogroup J is currently subdivided into two sub-haplogroups as indicated in Table 3.24.

Table 3.24 Classification of haplogroup J into two sub-haplogroups

Sub-haplogroup	Sequence motif	Age estimate (years before present)
J1	462-3010	10,100 to 25,100
J2	150-152-7476-15257	14,500 to 35,000

Adapted from Palanichamy *et al.* (2004), Logan (2008) and Van Oven and Kayser (2009).

Haplogroup T is defined by the sequence motif 709-1888-4917-8697-10463-13368-14905-15607-15928-16294 (Van Oven and Kayser, 2009). Haplogroup T can also be subdivided into two sub-haplogroups based on mutations at different nucleotide positions (Herrnstadt *et al.*, 2002a). The division of haplogroup T into sub-haplogroups is indicated in Table 3.25.

Table 3.25 Classification of haplogroup T into two sub-haplogroups

Sub-haplogroup	Sequence motif	Age estimate (years before present)
T1	12633A-16163-16189	6,100 to 12,800
T2	11812-14233-16296	9,200 to 16,100

Adapted from Richards *et al.* (2000), Finnäll *et al.* (2001b) and Herrnstadt *et al.* (2002a).

3.4 THE ROLE OF MITOCHONDRIAL HAPLOGROUPS IN CLINICAL CONDITIONS AND PHENOTYPIC VARIATION

When an mtDNA mutation is present, there are at least four factors that can influence the expression of the observed phenotype or disorder (DiMauro and Davidzon, 2005). These factors include environmental aspects, tissue-specific expression of interacting genes, the nuclear DNA background and the mtDNA haplotype. Related mtDNA haplotypes are classified into specific mtDNA haplogroups (refer to Section 3.3), and in this section the

link between mitochondrial haplogroups and disorders or phenotypes will be discussed briefly.

It has been demonstrated that there is a correlation between the geographic distribution of mitochondrial haplogroups and indigenous populations, and by extension the environments which they inhabit (Wallace, 2008). Typically, each haplogroup is established by one or more functional variants (refer to Section 3.3). In addition, quite a few of these functional mtDNA variants have arisen several times in diverse populations which indicate the role of convergent evolution and confirm the occurrence of adaptive selection (Wallace *et al.*, 1999; Mishmar *et al.*, 2003; Ruiz-Pesini *et al.*, 2004; Ruiz-Pesini and Wallace 2006). The influence of mtDNA variation on human health is illustrated by the detection of numerous associations between a variety of clinical conditions and specific mtDNA haplogroups. In addition, specific mitochondrial haplogroups are associated with a variety of potentially advantageous phenotypes. Examples of associations between mtDNA haplogroups and specific disorders or phenotypes are indicated in Table 3.26.

Table 3.26 mtDNA haplogroups which are associated with specific phenotypes and susceptibility to specific disorders

Disorders associated with mtDNA haplotypes		
Haplogroup	Disorder	Reference
H	Late onset Alzheimer's disease	Shoffner <i>et al.</i> , 1993
	Increased risk of Parkinson's disease	Van der Walt <i>et al.</i> , 2003
J	Associated with increased drusen levels	Jones <i>et al.</i> , 2007
	Susceptibility to multiple sclerosis	Kalman <i>et al.</i> , 1996
	Increased penetrance of milder LHON pathogenic mutations	Brown <i>et al.</i> , 1997 Torroni <i>et al.</i> , 1997 Brown <i>et al.</i> , 2001 Brown <i>et al.</i> , 2002
J/J1	Increased risk of diabetes	Crispim <i>et al.</i> , 2006 Feder <i>et al.</i> , 2009
K	Susceptibility to multiple sclerosis	Kalman <i>et al.</i> , 1996
T	Reduced sperm motility in Europeans	Ruiz-Pesini <i>et al.</i> , 2000
U	Increased prevalence of retinal pigment abnormalities	Jones <i>et al.</i> , 2007
	Associated with elevated serum IgE levels	Raby <i>et al.</i> , 2007
	Differences in sperm motility of sub-haplogroups	Montiel-Sosa <i>et al.</i> , 2006

Table 3.26 Continued...

Physiological characteristics influenced by mtDNA haplotypes		
Haplogroup	Phenotype	Reference
H	Protection against sepsis	Baudouin <i>et al.</i> , 2005
	Reduced risk of age-related macular degeneration	Jones <i>et al.</i> , 2007
J	Protection against Parkinson's disease	Van der Walt <i>et al.</i> , 2003
	Longevity	De Benedictis <i>et al.</i> , 1999 Ross <i>et al.</i> , 2001 Niemi <i>et al.</i> , 2003
K	Protection against Parkinson's disease	Van der Walt <i>et al.</i> , 2003
	Longevity	Bonafè <i>et al.</i> , 2002
	Protection against Alzheimer's disease	Carrieri <i>et al.</i> , 2001
L0	Endurance	Scott <i>et al.</i> , 2009
M	Longevity	Tanaka <i>et al.</i> , 1998
T	Protection against Alzheimer's disease	Chagnon <i>et al.</i> , 1999
U	Protection against Alzheimer's disease	Carrieri <i>et al.</i> , 2001

LHON = Leber's hereditary optic neuropathy; IgE = immunoglobulin E. Adapted from Coskun *et al.* (2003), Ruiz-Pesini *et al.* (2004), Wallace (2005), Wallace (2007) and Wallace (2008).

It is evident from Table 3.26 that the same mtDNA haplogroups (haplogroups J and K) can be associated with protection against neurodegenerative diseases and increased lifespan, while at the same time exhibiting an increased predisposition to the development of other degenerative diseases. This discrepancy can possibly be attributed to the interaction between the generation of mitochondrial energy (ATP), the production of reactive oxygen species (ROS) and apoptosis. If it is assumed that mtDNA variants associated with haplogroups J and K partially uncouple OXPHOS, this could lead to the preferential generation of heat, therefore reducing the number of excess electrons available to generate ROS. It is generally accepted that neurodegenerative diseases and aging are the result of the accumulation of somatic mtDNA mutations due to an increase in mitochondrial damage as a consequence of ROS, and a reduction in ROS should therefore provide protection against neurodegenerative diseases and aging (Kujoth *et al.*, 2005; Schriener *et al.*, 2005; Wallace, 2005). In contrast, the same mtDNA variants that cause OXPHOS uncoupling, would decrease the efficiency of ATP production, thereby rendering the individuals susceptible to diseases that are the result of energy deficiency (Wallace, 2005).

The role of different mtDNA haplogroups in disease and phenotypic variation implies that mtDNA haplogroups are functionally different. An ancient mtDNA variant that confers an advantage to individuals harbouring it in a specific climatic and dietary environment may therefore become detrimental in a different environment. It is therefore possible that mtDNA variants which are adapted to specific geographic regions and are harboured by

individuals that have since relocated to new geographic regions with the simultaneous adoption of new lifestyles, may currently be contributing to a host of modern diseases such as cardiovascular disease, hypertension, obesity, diabetes and neurodegenerative diseases (Mishmar *et al.*, 2003).

Conflicting reports of association, or a lack of association, between mitochondrial haplogroups and diseases or phenotypes are however common. It is therefore imperative that investigations of this nature continue in order to better define the modulatory role of haplotypes/haplogroups in promoting disease and contributing to phenotypic variation.

3.5 THE REVISED CAMBRIDGE REFERENCE SEQUENCE

A reference sequence is compiled to serve as the foundation for functional, medical and diversity studies, as it provides a constant reference for the identification and characterisation of genes, the analysis of mutations, investigation of expression profiles, the discovery of polymorphisms, as well as comparative analyses. The construction of a reference sequence requires the sequence determination of a number of sequences and usually the observed variation between the sequences is low enough to assign nucleotide identities to specific nucleotide positions within the sequence with a high degree of certainty.

A reference sequence for the human mitochondrial genome was constructed in 1981 when the complete sequence and organisation of the human mitochondrial genome was determined for the first time (Anderson *et al.*, 1981). This sequence was compiled using sequence data from a human placental mtDNA preparation, as well as human cervical carcinoma (HeLa) cell mtDNA. Ambiguities in the sequence were resolved by assuming the nucleotide identity to be identical to that of the bovine mtDNA sequence. This sequence became known as the Cambridge Reference Sequence (CRS) and was used as a standard in studies of human evolution, population genetics and to identify alterations that could be involved in various mitochondrial diseases. The CRS was in general use for a number of years as a standard, i.e. other mtDNA sequences were compared to the CRS and differences between the newly generated sequences and the reference sequence were studied. The observed differences could be genetic polymorphisms or deleterious mutations, especially if the newly generated mtDNA sequence was obtained from an individual that was affected by a disorder that is in some way linked to energy metabolism.

A number of discrepancies were consistently observed between sequencing results generated by several independent studies, involving subjects from diverse geographic and ethnic backgrounds, and the CRS (Howell *et al.*, 1992). Due to the importance of the CRS in investigations of molecular evolution and mitochondrial disease, it became necessary to determine the identities of the nucleotides at the positions where discrepancies were observed (Howell *et al.*, 1992). Andrews *et al.* in 1999 repeated the sequencing of the original placental mtDNA sample and observed an error frequency of 0.07% in the original sequence data which equates to a total of 11 nucleotide positions at which the CRS contained incorrect data. The sequence was subsequently revised and a new reference sequence referred to as the revised Cambridge Reference Sequence (rCRS) was compiled (Andrews *et al.*, 1999). The rCRS has since been used as the mitochondrial standard in investigations of human evolution and mitochondrial diseases.

There however remain deficiencies within the rCRS. It contains rare polymorphic alleles at seven nucleotide positions, which include a repeat sequence at position 311 to 315 where only five cytosine residues are observed instead of the more frequently occurring six. In addition, an error was identified at nucleotide positions 3,106 to 3,107 which led to the removal of one cytosine residue from the rCRS. In order to prevent an unacceptable level of confusion, the original nucleotide numbering of the CRS was retained in the rCRS. On the basis of two corrections made at nucleotide positions 14,776 and 7,028, the rCRS was assigned to haplogroup H. Haplogroup H is one of the European continent-specific haplogroups that is only present in low concentrations on other continents. The rCRS is therefore not a globally representative reference sequence.

In addition, it has been observed that when complete mitochondrial genome sequences of individuals from African descent are compared to the reference sequence, a greater number of differences are identified than when complete mitochondrial genome sequences from European descent are compared to the reference sequence (Horai *et al.*, 1995). This makes it extremely difficult to differentiate between a polymorphism and a deleterious mutation when investigating mtDNA sequences of African origin.

3.6 A CONSENSUS SEQUENCE FOR HUMAN MITOCHONDRIAL DNA VARIATION

A consensus sequence can be defined as a sequence that summarises the common features observed among a set of sequences and consists of the nucleotide present in the majority of sequences at each site (Page and Holmes, 1998). A consensus sequence is compiled in the same way as a reference sequence (refer to Section 3.5) by determining the nucleotide sequence of a number of different sequences. In the case of a consensus sequence, the observed variation between the different sequences is however higher. This implies that there are differences between the individual sequences at a greater number of nucleotide positions and that the most frequently appearing nucleotide for that position is assigned to that position, thereby determining the consensus identity for the specific nucleotide position. Sequence consensus methods assume that all the individual sequences used to compile the consensus sequence are equally related (Page and Holmes, 1998).

A consensus sequence is not necessarily the same as an ancestral sequence, it is only in the case where a star phylogenetic tree is present that this is the case (Page and Holmes, 1998). Consensus sequences are not appropriate for reconstructing the evolutionary history of a set of sequences, unless all the individual sequences used to compile the consensus sequence are equally related and generate a star phylogeny. Consensus sequences are simply tools to summarise sequence variation (Page and Holmes, 1998) and as such should be used as a standard to which other sequences can be compared and evaluated in terms of the variation present.

To date, two human mitochondrial consensus sequences have been generated to describe the variation observed in human mitochondrial DNA genomes (Marzuki *et al.*, 1991; Carter, 2007). The sequence generated in each of these investigations is discussed in the following paragraphs.

The Marzuki human mitochondrial consensus sequence was compiled using 13 unrelated mitochondrial sequences and has been deposited in GenBank® under accession number X62996 (Marzuki *et al.*, 1991). The 13 sequences used to construct this consensus sequence consist of the coding region only (bp 576 to 16,023) and included three previously published sequences, eight sequences obtained from individuals affected by various neuromuscular disorders (MERRF, encephalomyopathy, MELAS, CPEO and myotonic dystrophy) and two sequences obtained from unaffected control individuals. The

inclusion of sequences obtained from individuals affected by mitochondrial disorders (MELAS and MERRF) in which mtDNA lesions were suspected, necessitated the use of a four-point strategy to identify normal variants. Normal mtDNA variants were defined as those that were not disease-specific, i.e. variants that were not observed in all patients with MERRF or MELAS, were observed in at least one of the non-MERRF or MELAS mtDNA samples, resulted in no or conservative amino acid changes, or were associated with non-conserved nucleotides. The characteristics of the Marzuki human mitochondrial sequence are presented in Table 3.27.

Table 3.27 Characteristics of the Marzuki human mitochondrial consensus sequence

Nucleotide position	Gene	Marzuki consensus sequence	rCRS	Amino acid
1,438	12S rRNA	G	A	---
4,769	ND2	G	A	Met
4,985	ND2	A	G	Gln
8,860	ATPase6	G	A	Thr → Ala
15,326	Cytb	G	A	Ala → Thr

The D-loop is excluded in the analysis presented in the above table. rCRS = revised Cambridge Reference Sequence. Gene abbreviations: 12S rRNA = 12S ribosomal RNA; ATPase6 = ATP synthase F₀ subunit 6; Cytb = cytochrome *b*; ND2 = NADH dehydrogenase subunit 2. Nucleotide abbreviations: A = adenine; G = guanine. Amino acid abbreviations: Ala = alanine; Gln = glutamine; Met = methionine; Thr = threonine. Adapted from Marzuki *et al.* (1991).

The Carter human mitochondrial consensus sequence was constructed using 827 unrelated mitochondrial sequences and is available as a text file as part of the supplementary information provided online (Carter, 2007). The 827 sequences used to construct this full length consensus sequence (16,569 bp) were obtained from GenBank[®] and consisted of published sequences from regions across the world, although sub-Saharan Africa was not well represented. The 827 sequences were selected based on the quality of the sequence data (sequences exhibiting problems relating to quality or validity were removed), the publication status of the sequences (sequences not published in peer reviewed journals or databases were excluded) and the origin of the sequences (sequences from patents, HeLa cell lines, and individuals with recognised clinical diseases were not included). The most important features of the Carter human consensus sequence is summarised in Table 3.28. Although sequences from sub-Saharan regions were underrepresented in the dataset used, it was suggested that this consensus sequence, with the exclusion of a small number of weakly ambiguous nucleotide positions, was a reliable representation of the founder human mitochondrial sequence (Carter, 2007).

Table 3.28 Characteristics of the Carter human mitochondrial consensus sequence

Nucleotide position	Gene	Carter consensus sequence	rCRS	Amino acid
73	D-loop	G	A	---
263	D-loop	G	A	---
315	D-loop	CC	C	---
750	12S rRNA	G	A	---
1,438	12S rRNA	G	A	---
2,706	16S rRNA	G	A	---
4,769	ND2	G	A	Met
7,028	COI	T	C	Ala
8,860	ATPase6	G	A	Thr → Ala
11,719	ND4	A	G	Gly
14,766	Cytb	T	C	Ile → Thr
15,326	Cytb	G	A	Thr → Ala
16,519	D-loop	C	T	---

Gene abbreviations: 12S rRNA = 12S ribosomal RNA; 16S rRNA = 16S ribosomal RNA; ATPase6 = ATP synthase F₀ subunit 6; COI = cytochrome c oxidase subunit I; Cytb = cytochrome *b*; D-loop = displacement loop; ND2 = NADH dehydrogenase subunit 2; ND4 = NADH dehydrogenase subunit 4. Nucleotide abbreviations: A = adenine; C = cytosine; G = guanine; T = thymine. Amino acid abbreviations: Ala = alanine; Gly = glycine; Ile = isoleucine; Met = methionine; Thr = threonine. Adapted from Carter (2007).

A consensus sequence of the human mitochondrial genome can be used as a tool in a variety of ways. It can be used to investigate the evolution of the mitochondrial genome over time, as a baseline to measure human mitochondrial variation in populations worldwide (Carter, 2007), or to provide a standard reference to study normal and disease-related nucleotide variants in mtDNA genomes (Marzuki *et al.*, 1991). It is essential to recognise non-pathogenic mtDNA variation so that a distinction can be made between normal mtDNA variants and pathogenic mutations (Wallace, 1995). In cases where a population-specific consensus sequence is constructed, this sequence can be used to distinguish between population-specific polymorphisms, i.e. non-pathogenic mtDNA variants, and possible pathogenic mutations in mitochondrial sequences obtained from patients with mitochondrial myopathies belonging to the same or a related population group.

3.7 THE SOUTHERN AFRICAN KHOI-SAN POPULATION

Four major macro-language families have been recognised in Africa namely Afro-Asiatic, Niger-Kordofanian, Nilo-Saharan and Khoi-San (Gordon, 2005). The term Khoi-San was first used to describe a group of physically, culturally, and linguistically distinctive peoples and is a combination of the names that the two distinct groups of people that comprise this group use to refer to themselves, namely Khoi and San (Greenberg, 1950). The Khoi include groups that speak the southern and central Southern African Khoi-San (SAK)

languages, while the San include groups that speak southern, central and northern SAK languages (Chen *et al.*, 2000; refer to Figure 3.8).

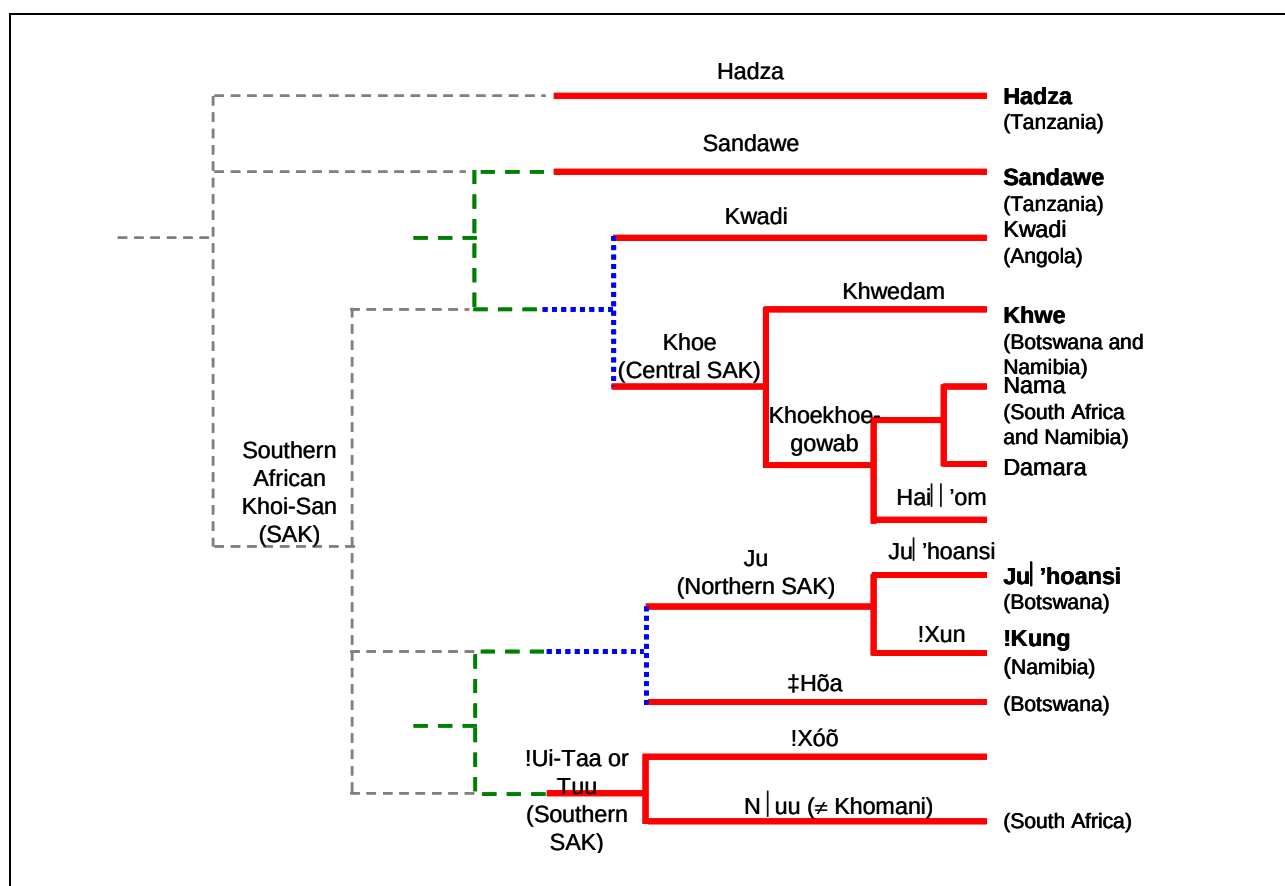
Culturally, the Khoi and San groups can be distinguished, as the Khoi are viewed as cattle-herders, while the San are viewed as hunter-gatherers (Greenberg, 1950). Since this division between the Khoi and San is based on lifestyle, it has been suggested that the creation of the Khoi ethnic group or identity is a colonial construct devised by early colonialists whereby groups with cattle were labelled as Khoi and groups without cattle as San (Smith, 2006). Based on the study and distribution of Khoi rock art, it has also been suggested that the Khoi might retain some Pygmy genetic markers, as well as evidence of recent genetic admixture with southern African San and Eastern Bantu-speaking farmers (Smith, 2006). Genetic studies conducted on Khoi and San populations should shed some light on genetic differences between these two groups who are currently grouped together to form the Khoi-San. At present however, it appears that in many investigations the Khoi-San are treated as a single population, and the Khoi-San population as a whole is classified as hunter-gatherers (Behar *et al.*, 2008), although only some groups such as the !Kung, Khwe, Hadza and Sandawe follow this lifestyle (Tishkoff *et al.*, 2009).

The Khoi-San language family classification was devised to refer to the language of the Khoi-San population which is predominated by click-sounds. The Khoi-San language classification system generally used is based on a system first proposed by Greenberg in 1950 and recognises three major groups i.e. the Hadza and the Sandawe (collectively referred to as the Eastern African Khoi-San), and the Southern African Khoi-San which is further divided into three groups (Güldemann and Vossen, 2000), namely northern Khoi-San (Ju), central Khoi-San (Khoe) and southern Khoi-San (!Ui-Taa or Tuu). The diffusion and differentiation of the Khoi-San language into three major language families can be attributed to one of two processes, namely the migration of proto-Khoi-San speakers from East Africa to southern Africa or the diffusion of the language itself (Crawhall, 2006). The inclusion of Hadza and Sandawe under the Khoi-San language family is however viewed as controversial by some investigators since these languages are divergent from each other and from the Southern African Khoi-San languages (Güldemann and Vossen, 2000). In a recent article, the classification structure within the Khoi-San language family and the issue of lumping together diverse groups under one language classification was discussed in detail, and it was suggested that the widespread view of the Khoi-San language classification, as shaped by Greenberg's work, was misguided and should yield to a more balanced view (Güldemann, 2008). A revision of the

Khoi-San language classification is therefore required to assist both linguists and researchers in other disciplines dealing with the individuals speaking the languages concerned (Güldemann, 2008).

The language classification system used for groups speaking click-languages is very important as it is used as a guide in determining relationships between populations and will ultimately affect the decision of which groups should be sampled to investigate genetic relationships among populations. It has also been suggested that genetic investigations could prove to be useful in determining if the groups speaking the three families of Khoi-San languages, namely Sandawe, Hadza or Southern African Khoi-San, share genetic ancestry or are genetically divided along language lines (Crawhall, 2006). A version of the Khoi-San language classification system currently used for the purpose of genetic investigations of Khoi-San speakers is presented in Figure 3.8. This classification system corresponds broadly to the Greenberg classification structure, although it also incorporates recent proposals made by linguists (Güldemann and Vossen, 2000; Güldemann, 2008).

Figure 3.8 The Khoi-San language classification system



In Figure 3.8, the geographic region in Africa where the specific Khoi-San language is spoken as well as the subsistence mode or lifestyle followed is indicated where the relevant information was available (Gordon, 2005). It is evident that the Khoi-San speakers are confined to sub-Saharan Africa. Although Khoi-San languages are currently confined to southern Africa with the exception of the Hadza and Sandawe languages which are present in eastern Africa (Tanzania), it has been suggested that prior to the Bantu migrations, hunter-gatherers occupied the majority of sub-Saharan Africa with individuals resembling the San populations of southern Africa being distributed over the whole of sub-Saharan Africa, reaching as far as Ethiopia and Sudan (Salas *et al.*, 2002; Smith, 2006). This conclusion is reached based on the presence of the Hadza and Sandawe in Tanzania which speak click-language isolates, as well as the presence of click consonants in some languages in Kenya and Ethiopia (Salas *et al.*, 2002).

The distribution of the San in sub-Saharan Africa is mirrored by the presence of San rock art across southern Africa. San rock art is observed in different provinces of South Africa (Free State, North-West, Western Cape, Eastern Cape, Northern Cape, Limpopo, and KwaZulu-Natal), as well as in Zimbabwe and Namibia, indicating that the ancestors of the San occupied a region in southern Africa which started in the south at the Drakensberg mountains in South Africa and ended in the north in line with the Zambezi River at the Namibia/Angola border area (Smith, 2006). In the case of the Sandawe and Hadza groups in East Africa, rock art is observed in a small part of central Tanzania centred on the Kondoa district with the southern boundary following the southern limits of the Kondoa and Singida districts and the northern boundary located at Lake Eyasi (Smith, 2006). Based on studies of mtDNA and Y chromosome data, it has been suggested that Khoi-San speaking populations may have originated in eastern Africa and migrated into southern Africa ca. 10,000 YBP (Cavalli-Sforza *et al.*, 1994; Scozzari *et al.*, 1999).

A recent investigation of the relationship between the three Khoi-San populations based on mtDNA data, concluded that it was not possible to distinguish whether the ancestral populations of contemporary Khoi-San speakers originated in East Africa or southern Africa (Tishkoff *et al.*, 2007). It was however suggested that the southern and eastern Khoi-San speaking populations became separated from each another between 15,000 to 60,000 YBP (Tishkoff *et al.*, 2007), possibly due to a notable dry period in southern Africa during the last glacial maximum 17,000 to 24,000 YBP (Lahr and Foley, 1998). The isolation between Khoi-San speaking populations was also likely furthered by other demographic events such as the Bantu migrations, as this could contribute to either the

assimilation or displacement of Khoi-San speaking populations in Africa (Salas *et al.*, 2002). High levels of genetic drift are observed in the L0d and L0k mtDNA lineages observed in Khoi-San groups (refer to Section 3.3.1.3) and this may be an indication that a number of L0 mtDNA lineages became extinct due to the fact that the territory of the Khoi-San were encroached upon by metal-using agriculturists (Salas *et al.*, 2002).

To date, genetic investigations of the mitochondrial DNA variation in the Southern African Khoi-San population has focussed mainly on the San and have involved LR-RFLP analysis (Denaro *et al.*, 1981; Johnson *et al.*, 1983), sequencing of the COIII gene (Wilder *et al.*, 2004b), sequencing of the control region only (Vigilant *et al.*, 1989; Vigilant *et al.*, 1991), HR-RFLP analysis combined with control region sequencing (Chen *et al.*, 2000), and complete mitochondrial genome sequencing (Gonder *et al.*, 2007; Tishkoff *et al.*, 2007). Some of the oldest mtDNA lineages observed in extant human populations have been observed in individuals presently identified as Khoi-San speakers. These lineages include L0d and L0k which appear to be localised to the Khoi-San population as they account for over 60% of the contemporary mtDNA gene pool (Vigilant *et al.*, 1991; Chen *et al.*, 2000; Salas *et al.*, 2002; Knight *et al.*, 2003; Gonder *et al.*, 2007; Tishkoff *et al.*, 2007).

In the different genetic studies mentioned in the above paragraph, the Khoi-San consistently cluster at the deepest nodes of phylogenies, exhibit the oldest TMRCA estimates, and display high levels of sequence divergence compared to other African populations. Genetic evidence for mtDNA as well as the Y chromosome indicates that the Hadza, Sandawe and the Southern African Khoi-San are the most genetically divergent of all the world populations (Knight *et al.*, 2003). Recently, based on nuclear markers, it has again been demonstrated by evaluating the genetic variation in Africa based on geographic, linguistic and subsistence classifications, that the highest level of population variation was exhibited in southern Africa, in hunter-gatherer populations and in Khoi-San language speakers (Tishkoff *et al.*, 2009). In a recent investigation (Schuster *et al.*, 2010), based on the complete nuclear genomes of Khoi-San and Bantu-speakers from southern Africa, it was determined that the Khoi-San contained higher levels of genetic diversity (i.e. more SNPs) than have been reported in other complete nuclear genomes of European or Asian origin. These characteristics support the suggestion that the Khoi-San (Hadza, Sandawe and Southern African Khoi-San) represent one of the most ancient populations in Africa.

3.8 **OBJECTIVES OF THE RESEARCH PROGRAMME**

A thorough understanding of the mitochondrial genetic diversity present in African populations is essential for the reconstruction of human evolutionary history, to determine the evolutionary relationships between populations, for the elucidation of the genetic basis of phenotypic variation and disease, and for mapping mitochondrial genetic diseases in individuals of recent African origin. The rCRS is currently used as a standard in studies of both human evolution and mitochondrial disease. However, as discussed in Section 3.5, this reference sequence has a number of deficiencies that decreases its effectiveness, specifically when used in studies involving African populations.

Historically Africa has been inadequately represented in studies of genetic diversity (Tishkoff and Williams, 2002), and there exists a general lack of mitochondrial genome sequence data from Africa, especially southern and eastern Africa. In spite of this, the highest mtDNA genetic diversity has consistently been observed in Africa (Cann *et al.*, 1987; Vigilant *et al.*, 1991; Chen *et al.*, 1995; Ingman *et al.*, 2000). Due to this general lack of mitochondrial genome sequence data from African populations, individuals of recent African descent such as African Americans have been investigated as representative of African populations and the genetics of Africans, even though these individuals are likely to be the product of admixture which will confound the investigation of genetic features confined to Africans (Herrnstadt *et al.*, 2002a). The current investigation will contribute to the number of available mitochondrial genome sequences from southern Africa, and it is therefore envisioned that the increased number of African mitochondrial genome sequences will in due course allow researchers to investigate the genetics of individuals of recent African descent by using genetic data of African populations that have not been involved in admixture with other populations.

When taking into consideration the high levels of mtDNA diversity in Africa, the frequent observation that mitochondrial sequences from individuals of continents other than Europe, especially individuals of African descent, exhibit a large number of differences compared to the rCRS, is not surprising. In investigations of mitochondrial disease, it therefore becomes difficult to distinguish between alterations associated with specific population groups and alterations with possible pathogenic significance. In the past, a recent disease-causing alteration was distinguished from background polymorphisms by determining the phylogenetic relationship between the patient's haplotype and the global mtDNA phylogeny (Wallace, 1994). This permitted the identification of the most closely

related “apparently healthy” haplotypes, thereby allowing the identification of variants that were present in the patient, while absent in the related controls (Wallace, 1994).

The main objective of this research programme is to construct a mitochondrial consensus sequence for the Khoi-San population by generating complete mtDNA sequences of Southern African Khoi-San individuals. The complete mtDNA sequences will aid in resolving the southern Africa genealogy (Salas *et al.*, 2002) and the Khoi-San consensus sequence will represent the first example of an African population-specific consensus sequence. Due to the fact that this sequence will be an African population-specific consensus sequence and will be based on one of the most divergent, and therefore ancient, populations in Africa, it is expected that the use of this consensus sequence will provide greater resolution of evolutionary histories and the relationships observed between populations. It is also envisioned that the construction of this mitochondrial consensus sequence will assist in the identification of differences between populations and will permit researchers to distinguish between normal African mtDNA variants and potential pathogenic mutations in individuals of recent African descent.

3.8.1 Specific aims of the proposed project

- a. To determine complete mitochondrial genome sequences of 30 Khoi-San individuals.
- b. To investigate mtDNA diversity in the Southern African Khoi-San population by:
 - i Identifying previously unreported alterations in the Khoi-San mitochondrial genome sequences.
 - ii Determining the haplogroup classification of the Khoi-San mitochondrial genome sequences.
- c. To further elucidate the phylogenetic architecture of haplogroup L0 by constructing a global L0-specific haplogroup network by means of the inclusion of L0 Khoi-San mitochondrial genome sequences in the L0 section of a previously constructed global mtDNA phylogenetic network (Ruiz-Pesini *et al.*, 2007).
- d. To investigate African mitochondrial DNA variation within the generated Khoi-San mitochondrial genome sequences, as well as a selection of macrohaplogroup L mitochondrial genome sequences from other sources through the use of phylogenetic analyses to determine the evolutionary relationships between the macrohaplogroup L sequences.

- e. To investigate African mitochondrial DNA variation within the generated Khoi-San mitochondrial genome sequences, as well as a selection of macrohaplogroup L mitochondrial genome sequences from other sources through the use of statistical analyses to:
 - i Determine basic sequence statistics.
 - ii Determine if deviations from the neutral theory of molecular evolution is observed for the different African haplogroups (L0, L1, L2 and L3) and if any observed occurrences of deviation can be attributed to changes in population size or the effect of selection.
 - iii Estimate the TMRCA for branches in the constructed global L0-specific haplogroup network.
- f. To use the global L0-specific haplogroup network to propose a revised classification scheme for the L0 haplogroup.
- g. To compile a mitochondrial consensus sequence for the Southern African Khoi-San population based on the population-specific lineages observed in the Khoi-San through the use of the 30 sequences in conjunction with 13 previously generated mitochondrial genome sequences (Mouton, 2003).

CHAPTER FOUR

Materials and Methods

All reagents used in the current investigation were, unless otherwise stated, analar grade products supplied by Promega^{®1}, Sigma^{®2} or Roche^{®3}. Protocols provided with the various kits used were followed and any deviations from these standard protocols have been reported. Where necessary, Equation 4.1 was used to convert gravitational force (*g*) to revolutions per minute (rpm) in the different protocols followed.

Equation 4.1 Conversion of gravitational force to revolutions per minute

$$g = 1.12 \times 10^{-6} \times r \times (\text{rpm})^2$$

g = gravitational force; *r* = average radius of the rotor in millimetre; rpm = revolutions per minute. Adapted from Sambrook and Russell (2001).

4.1 ETHICAL APPROVAL

The current investigation formed part of a research programme that has been approved by the Ethics Committee of the North-West University (Potchefstroom Campus). The research programme has been approved under the title of “The role of the mitochondrial genome in human migration and health”, and has been assigned the approval number NWU-00038-07-S8. Informed consent was obtained from all individuals, or their legal guardians, participating in the current investigation prior to the collection of samples.

4.2 SUBJECTS

The Khoi-San cohort studied in the current investigation consisted of 30 Vasikela !Kung individuals. These individuals formed a subset of a sample group previously collected in the north-western Kalahari Desert in southern Africa (Chen *et al.*, 2000). The total sample group collected contained 181 Khoi-San speaking individuals, of which 144 individuals were identified as Vasikela !Kung (Chen *et al.*, 2000). Individuals within the sample group were classified as !Kung according to the ethnicity listed on either their identity documents

¹ Promega[®] is a registered trademark of the Promega Corporation, Madison, WI, USA.

² Sigma[®] is a registered trademark of Sigma-Aldrich Corporation, St. Louis, MO, USA.

³ Roche[®] is a registered trademark of Roche Molecular Biochemicals, Indianapolis, IN, USA.

or birth certificates. In the few rare cases where such information was not available, verbal confirmation of ethnicity of a child was accepted after both parents were independently questioned about their own ancestry (Chen *et al.*, 2000).

4.3 ISOLATION OF GENOMIC DNA

Genomic DNA for each of the Khoi-San individuals used in the current investigation was isolated from cultured lymphoblastoid cells obtained from immortal lymphoblastoid cell lines by a previous student, Miss C. Mouton, of the Centre for Genome Research (CGR) at the North-West University (Potchefstroom Campus). The gDNA was extracted from the cultured lymphoblastoid cells using the PUREGENE^{®1} DNA Isolation Kit (Gentra Systems, Inc.) and the concentration of the isolated gDNA was determined by optical density (O.D.) analytical measurements. A Beckman DU-600 spectrophotometer was used for this purpose, and the gDNA samples were diluted 25-fold in double distilled water (ddH₂O) for O.D. readings. The isolated gDNA was stored at 2 to 8 degrees centigrade (°C) until use.

During storage of the isolated gDNA, the samples became dehydrated. The samples were rehydrated by the addition of 8 microlitres (μL) of hydration buffer FG3 from the FlexiGene DNA Isolation kit (QIAGEN[®]) followed by incubation overnight at room temperature or for one hour at 64°C in a water bath. The concentration and purity of the rehydrated gDNA in each of the samples was determined through O.D. analytical measurements using the Thermo Scientific NanoDrop^{™2} 1000 Spectrophotometer (Thermo Fisher Scientific, Inc.). The method employed was based on that described by Sambrook *et al.* in 1989. The spectrophotometer was calibrated with the appropriate buffer, and the O.D. of 1 μL undiluted sample was determined at 260, 280 and 320 nanometres (nm). The concentration of the gDNA was calculated by using the measurement at 260 nm and Equation 4.2 (refer to Section 4.6). The purity was estimated by determining the ratio between the readings recorded at 260 nm and 280 nm. For the purpose of the current investigation, a ratio of A_{260}/A_{280} greater than 1.75 was presumed to signify a DNA preparation suitable for further analyses. A 25 nanogram per microlitre (ng.μL⁻¹) DNA working solution was prepared for each individual and stored at 4°C, whilst the rehydrated stock solutions were stored at -20°C.

¹ PUREGENE[®] is a registered trademark of Gentra Systems, Inc., Minneapolis, MN, USA.

² NanoDrop[™] is a trademark of Thermo Fisher Scientific, Inc., Waltham, MA, USA.

4.4 POLYMERASE CHAIN REACTION

This method is based on a methodology first described by Mullis *et al.* in 1986 and involves the exponential amplification of a specific DNA sequence located between two regions of known DNA sequence. The complete DNA sequences of the mitochondrial genomes of the 30 Khoi-San individuals studied in the current investigation were determined by employing a PCR strategy that divided the mitochondrial genome into eight segments or PCR regions. PCR amplification, in segments, of the mitochondrial genomes under investigation was necessary in order to generate sufficient quantities of DNA for direct sequencing.

4.4.1 Polymerase chain reaction primer pairs

The mitochondrial genome of each Khoi-San individual was amplified by using eight overlapping primer pairs which represented eight different PCR regions. For some Khoi-San individuals, PCR region 3 and 5 were difficult to amplify. Subsequently, alternative primer sets were used for these two regions and the PCR regions amplified using the alternative primer sets are indicated by asterisks to distinguish between the original primer sets used for these regions and the alternative primer sets. All the primer pairs used for PCR amplification are listed in Table 4.1.

Table 4.1 Primer pairs used for amplification of the complete mitochondrial genome

Primer pair	Primer name	Primer sequence	T _m	Mean T _m	Product size (bp)
1	F32:mtL15996	5'-ctc cac cat tag cac cca aag c-3'	57	57	2,103
	R3:mtH1487	5'-gta tac ttg agg agg gtg acg g-3'	57		
2	F3:mtL923	5'-gtc aca cga tta acc caa gtc a-3'	53	52.5	2,789
	R7:mtH3670	5'-ggc gta gtt tga gtt tga tgc-3'	52		
3	F7:mtL3073	5'-aaa gtc cta cgt gat ctg agt tc-3'	53	53.5	2,801
	R11:mtH5832	5'-gac agg ggt tag gcc tct tt-3'	54		
3*	F8:mtL3644	5'-gcc acc tct agc cta gcc gt-3'	58	56	2,227
	R11:mtH5832	5'-gac agg ggt tag gcc tct tt-3'	54		
4	F11:mtL5278	5'-tgg gcc att atc gaa gaa tt-3'	48	51	2,679
	R15:mtH7918	5'-aga tta gtc cgc cgt agt cg-3'	54		
5	F15:mtL7379	5'-aga aga acc ctc cat aaa cct g-3'	53	50.5	3,124
	R20:mtH10462	5'-aat gag ggg cat ttg gta aa-3'	48		
5*	F16:mtL7882	5'-tcc ctc cct tac cat caa atc a-3'	53	52.5	2,089
	R19:mtH9928	5'-aac cac atc tac aaa atg cca gt-3'	52		

Table 4.1 Continued...

Primer pair	Primer name	Primer sequence	T _m	Mean T _m	Product size (bp)
6	F20:mtL9886	5'-tcc gcc aac taa tat ttc act t-3'	49	51.5	2,231
	R23:mtH12076	5'-gga gaa tgg ggg ata ggt gt-3'	54		
7	F23:mtL11486	5'-aaa act agg cgg cta tgg ta-3'	50	49.5	2,740
	R27:mtH14186	5'-tgg ttg aac att gtt tgt tgg-3'	49		
8	F27:mtL13612	5'-aag cgc cta tag cac tcg aa-3'	52	52	2,828
	R32:mtH16401	5'-tga ttt cac gga gga tgg tg-3'	52		

F = forward primer; R = reverse primer; T_m = melting temperature calculated for each primer; mean T_m = mean melting temperature for the primer set; product size (bp) = size of the amplified fragment in base pairs; 1, 2, 3, 4, 5, 6, 7 and 8 = primer pairs for PCR region 1, 2, 3, 4, 5, 6, 7 and 8; 3* and 5* = primer pairs for alternative PCR regions 3 and 5. The primer names indicate the nucleotide positions where the primers anneal (numbering of nucleotides is according to the rCRS) as well as the strand to which the primer anneals where L indicates the light strand and H indicates the heavy strand. Primer sequences were obtained from Maca-Meyer *et al.* (2001).

The PCR conditions for each primer set were optimised with regard to the annealing temperature (T_a) and magnesium chloride (MgCl₂) concentration. The T_a of the primer pairs was dependent on the mean melting temperature (T_m) of the individual primers in the primer pair. The T_m of the individual primers in a set was calculated by using the OligoCalc: Oligonucleotide Properties Calculator software version 3.07 (Kibbe, 2007). This program calculates T_m by taking into account the complex thermodynamic relationship between entropy, enthalpy, free energy and temperature. Optimisation of the T_a for a primer set was initiated at 2°C below the mean T_m of the two primers.

4.4.2 Polymerase chain reaction conditions

The reaction mixture that was used for the amplification of the specific mitochondrial segments consisted of 1 x PCR buffer [100 millimolar (mM) sodium chloride (NaCl), 20 mM 2-amino-2-(hydroxymethyl)-1,3-propanediol hydrochloride (Tris[®]-HCl, pH 8.0), 0.1 mM ethylenediamine tetra-acetic acid (EDTA), 1 mM dithiothreitol (DTT), stabilisers and 50% glycerol], 1.5 or 2.0 mM MgCl₂, 200 micromolar (μM) each of 2'-deoxyadenosine-5'-triphosphate (dATP), 2'-deoxycytidine-5'-triphosphate (dCTP), 2'-deoxyguanosine-5'-triphosphate (dGTP), 2'-deoxythymidine-5'-triphosphate (dTTP), 10 μM of both the forward (F) and reverse (R) primers, 0.25 or 0.5 units (U) of Super-Therm Polymerase^{®1} (Southern Cross Biotechnology), 25 nanogram (ng) of gDNA and ddH₂O was added to a final volume of 12.5 μL. The PCR reactions were prepared on ice and overlaid with an equal volume, i.e. 12.5 μL, of mineral oil to prevent evaporation.

¹ Super-Therm Polymerase[®] is a registered trademark of JMR Holdings, Sevenoaks, Kent, UK.

A master mixture of reagents was prepared for each reaction set and 11.5 μL was aliquoted into each PCR reaction microcentrifuge tube, to which 25 ng of gDNA was added. Two additional reaction mixtures were aliquoted for each amplification set, a negative control reaction in which the gDNA was replaced by 1 μL ddH₂O, and a positive control reaction to which 25 ng control gDNA was added. The controls underwent amplification together with the other samples in the reaction set and served to verify that the PCR reaction mixture had not been contaminated (negative control) and that the PCR amplification progressed as expected by yielding the correct result (positive control). Amplification was performed in a Thermo Hybaid^{®1} Multiblock System 0.5 Satellite (MBS 0.5S) thermocycler using the thermal cycling parameters indicated in Table 4.2. A standard PCR program or an extended PCR program was used depending on the mitochondrial segment amplified. After amplification, the success of the PCR was evaluated by using 0.9% agarose gel electrophoresis (refer to Section 4.5).

Table 4.2 Polymerase chain reaction thermal cycling conditions

No. of cycles	Cycle	Regular PCR		Extended PCR	
		Temperature	Duration	Temperature	Duration
1 cycle	Initial denaturation	94°C	10 min	94°C	10 min
30 cycles	Denaturation	94°C	30 sec	94°C	1 min
	Annealing	T _a	30 sec	T _a	1 min
	Elongation	72°C	30 sec	72°C	2 min
1 cycle	Final elongation	72°C	7 min	72°C	10 min
1 cycle	Hold	4°C	Indefinitely	4°C	Indefinitely

No. = number; T_a = the optimised annealing temperature; min = minute(s); sec = seconds; °C = degrees centigrade.

4.4.3 Polymerase chain reaction product purification

After amplification, PCR products were purified to remove unincorporated deoxynucleotide triphosphates (dNTPs), excess primer, remaining thermostable enzyme, mineral oil and salts. Purification was performed using one of two kits according to the manufacturer's instructions.

The first kit used was the QIAquick^{®2} PCR Purification Kit (Qiagen[®]). The protocol relies on the principle that nucleic acids specifically bind to the surface of glass fibres or silica materials in the presence of a chaotropic salt (Vogelstein and Gillespie, 1979). To achieve binding of the DNA to the silica matrix, 5 volumes of Buffer PB were added to 1 volume of

¹ Thermo Hybaid[®] is a registered trademark of Hybaid Ltd., Ashford, Middlesex, UK.

² QIAquick[®] is a registered trademark of QIAGEN Pty. Ltd., Australia.

PCR sample. The sample was applied to the QIAquick[®] spin column and centrifuged for 1 minute (min) at 13,000 rpm in an Eppendorf[®] 5810 centrifuge using a fixed angle rotor F-45-30-11. The flow-through was discarded and 750 μ L Buffer PE was added to remove contaminants, after which the sample was centrifuged for 1 min at 13,000 rpm. To remove residual ethanol (EtOH) from the column, the sample was again centrifuged for 1 min at 13,000 rpm. The column was inserted into a 1.5 millilitre (mL) microcentrifuge tube and the nucleic acids eluted by adding 30 to 50 μ L Buffer EB (10 mM Tris[®]-HCl, pH 8.5) to the centre of the column, incubating for 1 min at room temperature and centrifuging at 13,000 rpm for 1 min. The concentration and purity of the purified PCR product were estimated via spectrophotometry (refer to Section 4.6) and the remainder of the purified PCR product was stored at 4°C until further analysis.

The second kit used was the DNA Clean & Concentrator-5^{™1} kit (Zymo Research Corporation). The kit employs a single-buffer system that allows for the efficient adsorption of nucleic acids onto the matrix of a column to yield high quality purified DNA suitable for a variety of applications in a short time without the need for organic denaturants or chloroform. The first step of the protocol involved the addition of 2 volumes of DNA Binding Buffer to each volume of PCR sample. The solution was mixed briefly by pipetting and transferred to a Zymo-Spin^{™2} Column (Zymo Research Corporation). The column was centrifuged in an Eppendorf[®] 5810 centrifuge using a fixed angle rotor F-45-30-11 at 10,000 rpm for 30 seconds (sec). After the flow-through was discarded, contaminants were removed by addition of 200 μ L Wash Buffer to the column and centrifugation at 10,000 rpm for 30 sec. This wash step was repeated and the flow-through discarded. The column was placed in a 1.5 mL microcentrifuge tube and 15 to 20 μ L ddH₂O was added to the centre of the column matrix. The column was incubated for 1 min at room temperature to improve the yield of the purified PCR product, after which the column was centrifuged at 10,000 rpm for 30 sec to elute the DNA. The purified PCR product collected in the 1.5 mL microcentrifuge tube was transferred to a 0.2 mL microcentrifuge tube for storage. The concentration and purity of the purified PCR product were determined through the use of a spectrophotometer (refer to Section 4.6) and the remainder of the purified PCR product was stored at 4°C until use.

¹ DNA Clean & Concentrator-5[™] is a trademark of Zymo Research Corporation, Orange, CA, USA.

² Zymo-Spin[™] is a trademark of Zymo Research Corporation, Orange, CA, USA.

4.5 **AGAROSE GEL ELECTROPHORESIS**

PCR products were analysed by horizontal electrophoresis in 0.9% weight per volume (w/v) horizontal agarose slab gels using a Scigen agarose gel electrophoresis unit at 80 volts (V) for 45 min in 1 x TBE buffer [89.15 mM 2-amino-2-(hydroxymethyl)-1,3-propanediol (Tris[®], pH 8.1), 88.95 mM boric acid, 2.5 mM di-sodium ethylenediamine tetra-acetic acid (Na₂EDTA)]. Two different sized gels were used, a mini gel (40 mL final volume) and a midi gel (100 mL final volume). Gels were prepared as described in Sambrook *et al.* (1989). Briefly, an agarose solution was prepared as required by adding 10% volume per volume (v/v) 10 x TBE buffer to 0.9% (w/v) molecular grade agarose powder (Promega[®]) and adding ddH₂O to the desired final volume. The solution was brought to a boil, the volume adjusted with ddH₂O to compensate for any loss due to evaporation and the ethidium bromide (EtBr) solution (Sigma[®]) added to a final concentration of 0.5 µg EtBr/mL gel solution just before casting. The amplified PCR products (2.5 µL) were mixed with 2 µL of a 2 x loading buffer [0.04% w/v Orange G (Sigma[®]), 50% v/v glycerol] prior to loading. FastRuler™¹ DNA Ladder, High Range, ready-to-use marker (Fermentas International, Inc.) was loaded on each gel as a reference in order to estimate the fragment sizes of the PCR products. The fragments were visualised on a UVItec ultraviolet (UV) transilluminator, and electronic copies of the gel image were captured via a photo documentation system.

4.6 **DETERMINATION OF DNA CONCENTRATION**

The concentration of the purified PCR product was determined through optical density analytical measurements. The method that was employed is based on that described by Sambrook *et al.* in 1989. The DNA sample was diluted 10 times in ddH₂O, i.e. 5 µL DNA solution was added to 45 µL ddH₂O. The spectrophotometer (Eppendorf[®] BioPhotometer 6131) was calibrated with a solution containing ddH₂O, and the appropriate buffer, and the O.D. was determined at 260, 280 and 320 nm. The measurement recorded at 260 nm was used to calculate the concentration of DNA in the sample through the use of Equation 4.2. The absorbance at 280 nm was used to detect the presence of protein in the sample, and the purity of the sample was assessed by calculating the ratio between the measurements recorded at 260 nm and 280 nm. For the purpose of the current investigation, an A₂₆₀/A₂₈₀ ratio of 1.8 was accepted as indicative of a pure preparation of DNA (Sambrook *et al.*, 1989), and samples with ratios greater than 1.75 were used for analyses. Ratios below 1.8

¹ FastRuler™ is a trademark of Fermentas International, Inc., Ontario, Canada.

indicate the presence of protein contamination, while ratios greater than 1.8 signify the presence of RNA. The measurement at 320 nm was used to detect any light-scattering insoluble components within the sample and the value of this reading should be close to zero or very low to indicate that the solution is optically clear (Stephenson, 2003).

Equation 4.2 Calculation of DNA concentration from the absorbance at 260 nm

$$A_{260} \times \text{dilution factor} \times \text{extinction coefficient} = \text{concentration in } \mu\text{g.mL}^{-1}$$

A_{260} = absorbance of samples at 260 nm; dilution factor = 10; extinction coefficient for double stranded DNA (dsDNA) is 1 $A_{260} \equiv 50 \mu\text{g.mL}^{-1}$ and for single stranded DNA (ssDNA) is 1 $A_{260} \equiv 34.5 \mu\text{g.mL}^{-1}$; μg = microgram. Adapted from Sambrook and Russell (2001).

4.7 AUTOMATED DNA SEQUENCING

The sequencing method used is based on the dideoxynucleotide chain-termination method of Sanger *et al.* (1977). The ABI PRISM^{®1} BigDye^{™2} Terminator version 3.1 Ready Reaction Cycle Sequencing Kit (Applied Biosystems^{®3}) was used for automated sequencing.

In this protocol, detection is enabled by the addition of spectrally distinct fluorescent dye tags to each of the four types of chain-terminating nucleotides. The reaction mixture used in the cycle sequencing reaction further contains the DNA polymerase, AmpliTaq^{®4} DNA polymerase FS, which is a *Taq* polymerase specific for cycle sequencing.

4.7.1 Cycle sequencing

The standard protocol provided for cycle sequencing was followed with modifications where necessary. The cycle sequencing reaction consisted of 30 to 70 ng purified PCR product (refer to Table 4.3), 3.2 picomole (pmol) forward or reverse primer (refer to Table 4.5), 2 μL BigDye[™] Terminator Ready Reaction Mix [containing dNTPs, 2',3'-dideoxynucleotide triphosphates (ddNTPs), AmpliTaq[®] DNA polymerase FS, MgCl_2 and Tris[®]-HCl (pH 9.0)], 1/10th of the final volume BigDye[™] 5 x Sequencing Buffer and ddH₂O to a final volume of 10 μL .

¹ ABI Prism[®] is a registered trademark of Applied Biosystems Corporation, Foster City, CA, USA.

² BigDye[™] is a trademark of Applied Biosystems Corporation, Foster City, CA, USA.

³ Applied Biosystems[®] is a registered trademark of Applied Biosystems Corporation, Foster City, CA, USA.

⁴ AmpliTaq[®] is a registered trademark of Roche Molecular Systems, Inc., Pleasanton, CA, USA.

Table 4.3 Estimation of the quantity of the purified PCR product to be used in a cycle sequencing reaction based on the size of the PCR product

Size of PCR product	Quantity
100 – 200 bp	1 to 3 ng
200 – 500 bp	3 to 10 ng
500 – 1,000 bp	5 to 20 ng
1,000 – 2,000 bp	10 to 40 ng
> 2,000 bp	40 to 100 ng

bp = base pairs; ng = nanogram. Adapted from the Automated Sequencing Chemistry Guide (Applied Biosystems®, 2000).

The cycle sequencing reactions were performed using the Thermo Hybaid® Multiblock System 0.5 Satellite (MBS 0.5S) thermocycler. The cycle sequencing program that was used is presented in Table 4.4, and the primers that were used are presented in Table 4.5.

Table 4.4 Thermal cycling conditions to be used for sequence determination of segments of the mitochondrial genome

No. of cycles	Cycle	Temperature	Duration
25 cycles	Denaturation	96°C	10 sec
	Annealing	50°C	5 sec
	Elongation	60°C	4 min
1 cycle	Hold	4°C	Indefinitely

No. = number; min = minutes; sec = seconds; °C = degrees centigrade.

The mtDNA sequence of each of the Khoi-San individuals was determined by using the primers listed in Table 4.5. The majority of these primers are forward primers that anneal to the L-strand of the mitochondrial DNA genome, while a few are reverse primers that anneal to the H-strand. The primers were based on primers described by Maca-Meyer *et al.* in 2001, with modifications where necessary, in order to generate sufficient sequencing data. The eight PCR segments generated for each individual were sequenced in ca. 500 bp size fragments by using four forward primers and the reverse primers indicated for each PCR region as required. Reverse primers were used to generate sequence data in the event that an overlap was not obtained between contiguous sequences within or between PCR regions, to verify the identity of a nucleotide at a position that displayed ambiguity and to determine the number of nucleotides in homopolymer stretches. All the sequences generated were analysed as described in Section 4.7.5.

Table 4.5 Primers used for sequencing of the complete mitochondrial genome

PCR region	Primer name	Primer sequence	T _m	Product size (bp)
1	F32:mtL15996	5'-ctc cac cat tag cac cca aag c-3'	56	344
	R32:mtH16401	5'-tga ttt cac gga gga tgg tg-3'	52	240
	F1:mtL16340	5'-agc cat tta ccg tac ata gca ca-3'	55	611
	R1:mtH406	5'-tgt taa aag tgc ata ccg cca-3'	50	578
	F2:mtL382	5'-caa aga acc cta aca cca gcc-3'	53	541
	R2:mtH945	5'-ggg agg ggg tga tct aaa ac-3'	54	535
	F3:mtL923	5'-gtc aca cga tta acc caa gtc a-3'	52	449
	R3:mtH1487	5'-gta tac ttg agg agg gtg acg g-3'	57	544
2	F4*:mtL1372	5'-cta ccc cag aaa act acg-3'	48	653
	R4:mtH2053	5'-tta gag ggt tct gtg ggc aaa-3'	52	565
	F5:mtL2025	5'-gcc tgg tga tag ctg gtt gtc c-3'	57	534
	F6:mtL2559	5'-cac cgc ctg ccc agt gac aca t-3'	60	478
	F7:mtL3073	5'-aaa gtc cta cgt gat ctg agt tc-3'	53	467
3	F8*:mtL3540	5'-tct cac cat cgc tct tct ac-3'	52	670
	R8:mtH4227	5'-atg ctg gag att gta atg ggt-3'	50	557
	F9:mtL4210	5'-cca ctc acc cta gca tta ctt a-3'	52	540
	F10:mtL4750	5'-cca ata cta cca atc aat act c-3'	47	528
	F11:mtL5278	5'-tgg gcc att atc gaa gaa tt-3'	50	421
3*	F8:mtL3644	5'-gcc acc tct agc cta gcc gt-3'	58	623
	R8:mtH4227	5'-atg ctg gag att gta atg ggt-3'	50	557
	F9:mtL4210	5'-cca ctc acc cta gca tta ctt a-3'	52	540
	F10:mtL4750	5'-cca ata cta cca atc aat act c-3'	47	528
	F11:mtL5278	5'-tgg gcc att atc gaa gaa tt-3'	50	421
4	F12*:mtL5699	5'-cta agc acc cta atc aac tg -3'	50	638
	F13:mtL6337	5'-cct gga gcc tcc gta gac ct-3'	57	532
	F14:mtL6869	5'-ccg gcg tca aag tat tta gc-3'	53	510
	F15:mtL7379	5'-aga aga acc ctc cat aaa cct g-3'	52	415
	R15:mtH7918	5'-aga tta gtc cgc cgt agt cg-3'	54	510
5	F16*:mtL7794	5'-cca tca tcc tag tcc tca tc -3'	52	505
	R16:mtH8345	5'-ttt cac tgt aaa gag gtg ttg g-3'	51	429
	F17:mtL8299	5'-acc ccc tct aga gcc cac tg-3'	56	500
	F18:mtL8799	5'-ctc gga ctc ctg cct cac tca-3'	58	563
	F19:mtL9362	5'-ggc cta cta acc aac aca cta-3'	51	524
	R19:mtH9928	5'-aac cac atc tac aaa atg cca gt-3'	52	534
5*	F16:mtL7882	5'-tcc ctc cct tac cat caa atc a-3'	53	506
	R16:mtH8345	5'-ttt cac tgt aaa gag gtg ttg g-3'	51	429
	F17:mtL8299	5'-acc ccc tct aga gcc cac tg-3'	56	500
	F18:mtL8799	5'-ctc gga ctc ctg cct cac tca-3'	58	563
	F19:mtL9362	5'-ggc cta cta acc aac aca cta-3'	51	524
	R19:mtH9928	5'-aac cac atc tac aaa atg cca gt-3'	52	534

Table 4.5 Continued...

PCR region	Primer name	Primer sequence	T _m	Product size (bp)
6	F20:mtL9886	5'-tcc gcc aac taa tat ttc act t-3'	51	517
	F21:mtL10403	5'-aaa gga tta gac tga acc gaa-3'	50	546
	F21:mtL10403MF	5'-aaa gga tta gac tga gcc gaa-3'	50	546
	R21:mtH10975	5'-cca tga ttg tga ggg gta gg-3'	54	513
	F22:mtL10949	5'-ctc cga ccc cct aac aac cc-3'	54	537
	F23:mtL11486	5'-aaa act agg cgg cta tgg ta-3'	51	478
	R23:mtH12076	5'-gga gaa tgg ggg ata ggt gt-3'	54	549
7	F24*:mtL11964	5'-tca cag ccc tat act ccc-3'	50	608
	R24:mtH12603	5'-acg aac aat gct aca ggg atg-3'	52	528
	F25:mtL12572	5'-aca acc cag ctc tcc cta ag-3'	53	516
	R25:mtH13124	5'-att ttc tgc tag ggg gtg ga-3'	52	520
	F26:mtL13088	5'-agc cct act cca ctc aag cac-3'	56	524
	R26:mtH13666	5'-agg gtg ggg tta ttt tcg tt-3'	50	542
	F27:mtL13612	5'-aag cgc cta tag cac tcg aa-3'	56	443
	R27:mtH14186	5'-tgg ttg aac att gtt tgt tgg-3'	49	521
8	F28*:mtL14055	5'-ctc cat cat cac ctc aac c-3'	51	595
	F29:mtL14650	5'-ccc cat tac taa acc cac act c-3'	52	512
	F30:mtL15162	5'-ctc ccg tga ggc caa ata tc-3'	53	514
	F31:mtL15676	5'-tcc cca tcc tcc ata tat cc-3'	49	320

F = forward primer; R = reverse primer; T_m = melting temperature calculated for each primer; product size (bp) = size of the region between the current F or R primer and the next F or R primer in base pairs. The primer names indicate the nucleotide positions where the primers anneal (numbering of nucleotides is according to the rCRS) as well as the strand to which the primer anneals (light strand (L) or heavy strand (H)). Primers indicated by * were primers designed for use in the current investigation, the remainder of the primer sequences were obtained from Maca-Meyer *et al.* (2001). F21:mtL10403 is the L10403 primer as designed by Maca-Meyer *et al.* (2001), and is also referred to as unmodified F21 in the current investigation; F21:mtL10403MF is the L10403 primer as designed by Maca-Meyer *et al.* (2001) with a modification at position 10,399 (A10399G), this primer is also referred to as modified F21 in the current investigation. The choice of F21 primer for sequencing (unmodified or modified F21) was dependent on the specific Khoi-San individual investigated.

4.7.2 Sodium dodecyl sulphate/heat treatment of extension products

Prior to precipitation of extension products using 96-well spin plate purification (refer to Section 4.7.3), it was necessary to prepare the extension products to ensure the complete removal of excess dye terminators and to prevent the occurrence of dye blobs in the sequencing data. The standard protocol provided for sodium dodecyl sulphate (SDS)/heat treatment of extension products was followed with modifications where necessary.

A 2.2% (v/v) SDS solution was prepared using deionised water. The cycle sequencing reactions were removed from the thermocycler and an appropriate amount (1 µL) of the 2.2% SDS solution was added to the reactions to bring the final concentration of SDS to 0.2%. The cycle sequencing reactions were mixed briefly by vortexing, and centrifuged in

a Spectrafuge™¹ Mini Personal Centrifuge for 10 sec. The cycle sequencing reactions were placed into a Thermo Hybaid® Multiblock System 0.5 Satellite (MBS 0.5S) thermocycler and the reactions were subjected to the heating/cooling cycle described in Table 4.6.

Table 4.6 Thermal cycling conditions used during SDS/heat treatment of cycle sequencing reactions

No. of cycles	Cycle	Temperature	Duration
1 cycle	Heating	98°C	5 min
1 cycle	Cooling	25°C	10 min
1 cycle	Hold	4°C	Indefinitely

No. = number; min = minutes; °C = degrees centigrade. Adapted from the Applied Biosystems® 3730/3730x/ DNA Analyzers Sequencing Chemistry Guide (Applied Biosystems®, 2002).

After removal of the cycle sequencing reactions from the thermocycler, the reactions were centrifuged in a Spectrafuge™ Mini Personal Centrifuge for 10 sec. The reactions were subsequently stored at 4°C until precipitation of the extension products was performed.

4.7.3 Precipitation of extension products

After cycle sequencing, it was necessary to remove the unincorporated BigDye™ terminators remaining in the sequencing reaction in order to prevent interference with the signal produced during electrophoresis. This was achieved by precipitation of the DNA extension products which separates the DNA fragments from other reaction components remaining in the sequencing reaction. Precipitation of extension products was performed according to one of two protocols as required by the specific sequencing facility contracted to perform the electrophoresis of the sequencing reactions (refer to Section 4.7.4).

The standard protocol provided for the EtOH/sodium acetate precipitation method was followed with modifications where necessary. Double distilled water was added to the cycle sequencing reaction to a final volume of 20 µL. This was followed by the addition of 80 µL of an EtOH/sodium acetate (NaOAc) solution consisting of 3 µL of a 3 molar (M) NaOAc (pH 4.6) solution, 62.5 µL of non-denatured 99.8% EtOH and 14.5 µL deionised water. The reaction was vortexed briefly, incubated at room temperature for 15 min and centrifuged for 20 min at 10,600 x *g* in an Eppendorf® 5810R centrifuge. The supernatant was discarded and the pellet washed by adding 250 µL of a 70% EtOH solution. The sample was again centrifuged for 5 min at 10,600 x *g*, the supernatant removed, and the

¹ Spectrafuge™ is a registered trademark of Labnet International, Inc., Berkshire, UK.

pellet air-dried for 15 to 20 min. The precipitated product was stored in the dark at -20°C until further analysis.

Centri•Sep™¹ 96 well plates (Princeton Separations, Inc.) were also used to precipitate extension products prior to electrophoresis of sequencing reactions. The Centri•Sep™ 96 well plate protocol was not performed by the author, instead it was contracted to a second party who was responsible for the electrophoresis of the sequencing reactions. The standard protocol provided for the Centri•Sep™ 96 well plates was followed and involved the following steps. The Centri•Sep™ 96 well plate was allowed to equilibrate to room temperature. Thereafter, the adhesive foil sealing film was removed from both the top and the bottom of the Centri•Sep™ 96 well plate. The plate was stacked on top of a reusable deep well wash plate and centrifuged at 1,500 x *g* for 2 min. The eluant was discarded by shaking the wash plate until dry. Thereafter, the cycle sequencing reactions (total volume 10 µL) were transferred to the individual wells in the Centri•Sep™ 96 well plate, taking care to place the samples in the centre of the gel bed. A 96 well PCR sample collection plate was placed into the deep well wash plate and the Centri•Sep™ 96 well plate was stacked on top of the PCR plate. The extension products were eluted into the 96 well PCR collection plate by centrifugation at 1,500 x *g* for 2 min. The 96 well PCR plate was removed from the deep well wash plate and dried in a speed-vac equipped with the appropriate rotor. Alternatively, the 96 well PCR sample collection plate was sealed with ThermalSeal film for storage until use.

4.7.4 Electrophoresis of extension products

Extension products were analysed through capillary electrophoresis. Capillary electrophoresis has the advantage that quantitation is more accurate, electrophoresis run times are shorter and only a small amount of precipitated product is used which allows the repeat of electrophoresis of a specific product if necessary (Applied Biosystems®, 2002).

After precipitation of extension products through the use of the EtOH/sodium acetate precipitation method, the precipitates were resuspended in 6 µL Hi-Di™² deionised formamide and 3 µL of each product was electrophoresed in a SpectruMedix™³ (SCE2410) Genetic Analysis System Sequencer. The length of the electrophoretic run was dependent on the length of the sequenced product. Alternatively, after precipitation of

¹ Centri•Sep™ is a trademark of Princeton Separations, Inc., Adelphia, NJ, USA.

² Hi-Di™ is a trademark of Applied Biosystems, Foster City, CA, USA.

³ SpectruMedix™ is a trademark of SpectruMedix LLC., State College, PA, USA.

extension products through the use of Centri•Sep™ 96 well plates, the precipitate in each well of the 96 well PCR collection plate was resuspended in 10 to 30 µL Hi-Di™ deionised formamide and the samples were electrophoresed using an Applied Biosystems® 3130xl Genetic Analyzer.

4.7.5 Analysis of sequence data

After electrophoresis, the sequence data obtained was evaluated. Sequences were assigned to two classes, successful or failed, based on the appearance of the electropherogram. A successful sequence included two types of sequences, “nice” sequences and “usable” sequences. Nice sequences were sequences with well-defined peaks and a negligible amount of background noise, while usable sequences were sequences which were truncated, contained missing nucleotides and/or gaps (usually due to the presence of dye blobs), as well as unconfirmed sequence changes (the change occurred in one sequence of the region and not in another sequence of the same region). Although these sequences were not optimal and contained artefacts (refer to Section 5.6.4), the sequence data could still be used.

Failed sequences also consisted of two types of sequences, “failed” sequences and “unusable” sequences. Failed sequences were defined as sequences that had failed completely, i.e. there was a complete lack of signal due to insufficient template in the cycle sequencing reaction, or failure of the sequence precipitation step. Unusable sequences were defined as sequences that could not be used due to serious artefacts within the sequence, such as a high level of background noise that obscured the true peaks making it very difficult to distinguish between the true peaks and the background peaks. Failed sequences were repeated as described in Sections 4.7.1, 4.7.2, 4.7.3, and 4.7.4, while successful sequences were subjected to further analyses.

The sequences obtained for each individual were evaluated, analysed and alignment performed of the overlapping segments by using the freeware computer program BioEdit Sequence Alignment Editor version 5.0.9 (Hall, 1999). After alignment of all overlapping regions and removal of all primer sequences, the CONTIG ASSEMBLY PROGRAM (CAP), written by Huang (1992), contained within the BioEdit version 5.0.9 (Hall, 1999) program, was used to compile a contig representing the complete mtDNA sequence for each individual. The complete mtDNA sequence for each individual was aligned to and compared with the rCRS to identify sequence changes. These alterations

were compared to the polymorphism data available in MITOMAP: A Human Mitochondrial Genome Database (<http://www.mitomap.org>, 2008), as well as mtDB - Human Mitochondrial Genome Database (<http://www.genpat.uu.se/mtDB>, 2008) to determine if the observed discrepancies were previously reported or novel polymorphisms and whether the observed alteration resulted in a synonymous or a non-synonymous change. Novel alterations were defined as alterations that had not been reported in the scientific literature, or were not listed in either of the two mitochondrial genome databases.

4.8 MITOCHONDRIAL DNA GENOME SEQUENCES USED IN ANALYSES

A total of 390 mitochondrial genome sequences were retrieved from the GenBank® database (Benson *et al.*, 2008) for use in both the phylogenetic and statistical analyses (refer to Sections 4.12 and 4.14), as well as in the construction of a global L0-specific haplogroup network (refer to Section 4.13). Ninety-three complete mitochondrial genome sequences obtained from African American individuals that were published in a database owned by a commercial entity, MitoKor, Inc. (Herrnstadt *et al.*, 2002a; Howell *et al.*, 2004) were not included in the current investigation. These sequences were excluded due to the fact that the sequences were not published in an open access database and were therefore not subjected to a thorough examination by members of the scientific community. The mitochondrial genome sequences obtained from GenBank® included both complete (bp 1 to 16,569) and coding region only (bp 577 to 16,023) sequences. The sequences were obtained from the Nucleotide database in the period between January 2004 to June 2007 by performing searches using different combinations of the keywords and phrases listed in Table 4.7, or by searching for the first author of articles published prior to June 2007 in which African mitochondrial genome sequences were reported. Sequences released by GenBank®, or sequences reported in articles published after June 2007, are not included in the datasets compiled for use in the various analyses performed in the current investigation (refer to Section 4.10).

Table 4.7 Keywords and phrases used to recover published mitochondrial genome sequences belonging to macrohaplogroup L

Keyword	Phrase	Publications
African	Complete mitochondrial genome	Horai <i>et al.</i> , 1995
Kung/!Kung	African origin	Ingman <i>et al.</i> , 2000
San	L-haplogroup	Maca-Meyer <i>et al.</i> , 2001
Khoi-San	L ^{*a}	Herrnstadt <i>et al.</i> , 2002a
Khoisan	<i>Homo sapiens</i>	Mishmar <i>et al.</i> , 2003
---	---	Macaulay <i>et al.</i> , 2005
---	---	Kivisild <i>et al.</i> , 2006
---	---	Torroni <i>et al.</i> , 2006
---	---	Gonder <i>et al.</i> , 2007

^a = the asterisk denotes any character, and was used in the database search to indicate any character after L e.g. L1, L2 or L3.

An additional 13 Khoi-San mitochondrial genome sequences were also obtained from a previous investigation (Mouton, 2003). The rCRS and the complete mitochondrial genome sequence of the Common Chimpanzee (*Pan troglodytes*) were also retrieved from the GenBank[®] database for use in the evaluation of sequence data (refer to Section 4.7.5) and in phylogenetic analysis (refer to Section 4.12). Information regarding the GenBank[®] accession number, the GenInfo Identifier sequence identification number (GI number), sequence name or isolate, ethnicity or geographic origin of the individual from which the sequence was obtained, the haplogroup and sub-haplogroup classification, the length of the sequence in base pairs and the relevant references for all sequences used in the current investigation are provided in Appendix G. This Appendix also includes the relevant information for seven mitochondrial genome sequences that form part of the mtDNA L0 haplogroup network constructed by Ruiz-Pesini *et al.* (2007). The information for these sequences are provided as these sequences form part of the network that was used as the basis for the global L0-specific haplogroup network constructed in the current investigation (refer to Section 4.13).

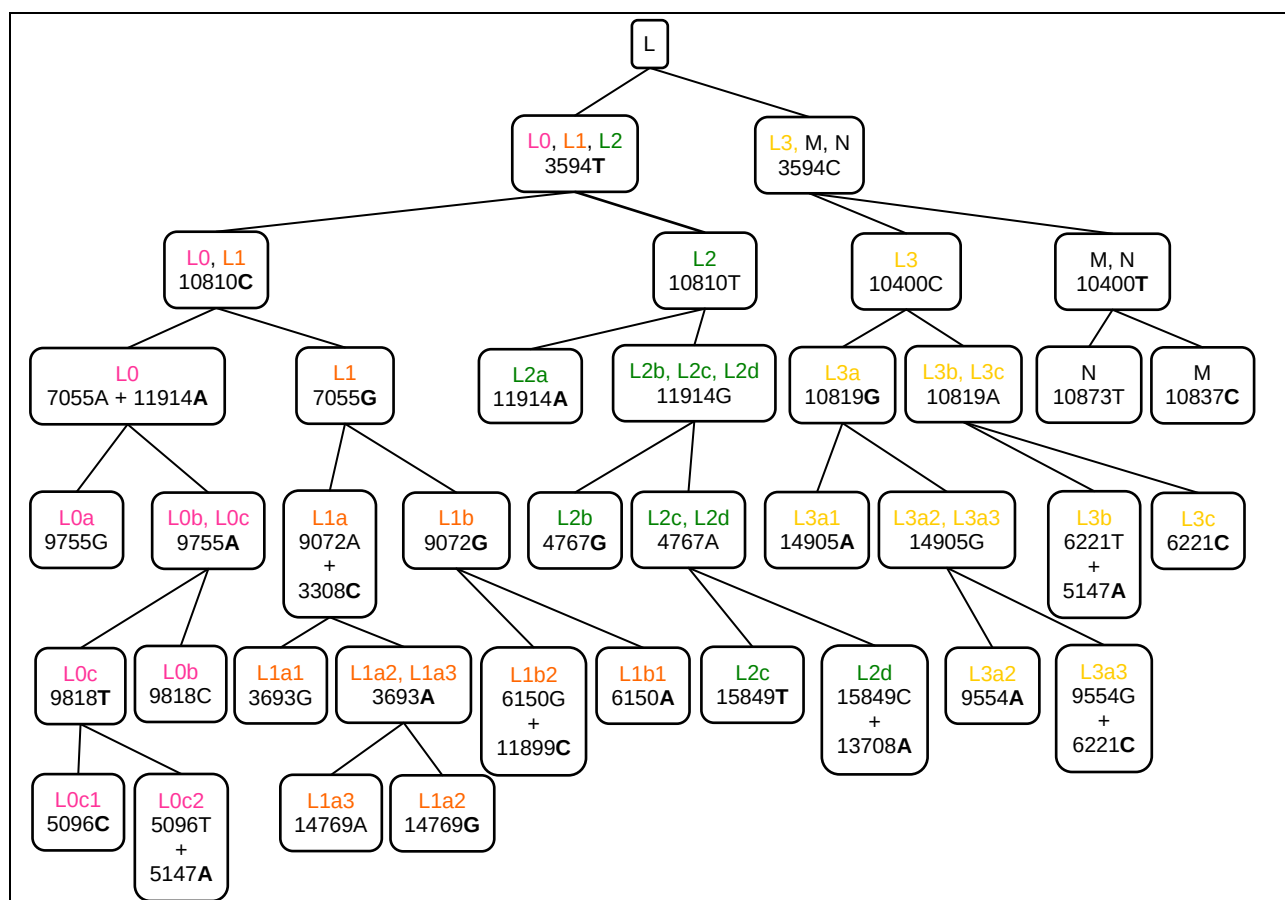
4.9 DETERMINATION OF MITOCHONDRIAL HAPLOGROUPS

Mitochondrial haplogroups and sub-haplogroups were assigned to the Khoi-San mitochondrial genome sequences generated in the current investigation, as well as to all other mitochondrial genome sequences used in analyses (refer to Section 4.8), based on the presence or absence of previously identified haplogroup or sub-haplogroup informative SNPs. The mtDNA sequences generated for each Khoi-San individual, as well as the mitochondrial sequences obtained from GenBank[®] and other sources (refer to Section 4.8), were analysed and the SNPs indicated in the classification scheme

presented in Figure 4.1 were used to classify sequences within the different African mitochondrial L haplogroups or sub-haplogroups.

The classification scheme presented in Figure 4.1 was adapted from Wallace (2004) and is part of a larger scheme based on the analysis of 103 SNPs in a number of sequences. All the SNPs used in the construction of the classification scheme are observed in the coding region of the mitochondrial genome and are located in polypeptide genes (ATPase6, COI, COII, Cytb, ND1, ND2, ND3, ND4 and ND5). Some of these SNPs are repeated in different clusters and the haplogroup is therefore defined by a combination of SNPs. The numbering of the nucleotide positions is based on the light strand of the rCRS. It is important to note that the nomenclature used for some sub-haplogroups are not universal and this should be taken into consideration when comparing sequences classified using this scheme with sequences generated and classified in other investigations (refer to Section 3.3.1.1 and Appendix F).

Figure 4.1 Schematic representation of the single nucleotide polymorphisms used to characterise African mtDNA L haplogroups



This is a classification scheme for macrohaplogroup L. Nucleotides present in the rCRS are indicated in regular text, while changes compared to the rCRS are indicated in **bold** text. Adapted from Wallace (2004).

4.10 COMPILATION OF MITOCHONDRIAL SEQUENCE DATASETS USED IN ANALYSES

Mitochondrial DNA haplogroups were assigned to a total of 433 genome sequences (refer to Section 4.9). The sequences consisted of 390 sequences retrieved from GenBank[®], 13 Khoi-San sequences obtained from a previous investigation (Mouton, 2003) and 30 Khoi-San sequences generated in the current investigation (refer to Section 4.8). After haplogroup classification, these sequences were used to compile a sequence database from which all the sequence datasets described in Table 4.10 were derived. The sizes of the datasets ranged from 43 sequences to 429 sequences and the datasets contained coding region only sequences, except for datasets 3b and 4a which contained complete sequences for the construction of phylogenetic trees (refer to Section 4.12) and the determination of basic sequence statistics (refer to Section 4.14).

The incompletely classified sequences (refer to Section 4.9) were included in datasets used in phylogenetic (refer to Section 4.12) and statistical analyses (refer to Section 4.14), with the exception of the sequences as listed in Sections 4.10.1, 4.10.2 and 4.10.3. All relevant information (GenBank[®] accession number, sequence name or isolate, ethnicity or geographic origin, and haplogroup or sub-haplogroup classification) regarding the sequences included or excluded in the various datasets are presented in Appendix G.

4.10.1 Compilation of datasets used for phylogenetic analyses

A number of mitochondrial sequences were excluded from the sequence database prior to the compilation of the individual sequence datasets used for phylogenetic analyses (refer to Section 4.12). These sequences and the reasons for their exclusion from the database are listed in Table 4.9.

Four sequences were excluded from the database, namely DQ112926 and DQ112933 from Ethiopia (Kivisild *et al.*, 2006), EF184599 from Tanzania (Gonder *et al.*, 2007) and EF184609 from South Africa (Gonder *et al.*, 2007). These sequences could not be classified within macrohaplogroup L, or although belonging to macrohaplogroup L, could not be classified completely.

Table 4.9 Sequences obtained from GenBank® that were excluded from phylogenetic analyses

GenBank® accession number	Haplogroup assignment	SNP profile	Reason for exclusion	Reference
DQ112926	M	3594C + 10400T + 10873C	Classified as macrohaplogroup M	Kivisild <i>et al.</i> , 2006
DQ112933	M	3594C + 10400T + 10873C	Classified as macrohaplogroup M	Kivisild <i>et al.</i> , 2006
EF184599	L0, L1, L2	3594T + 10810A	Could not be classified beyond L0, L1, L2 level, contained an A at nucleotide position 10,810	Gonder <i>et al.</i> , 2007
EF184609	Unknown	3594A	Could not be classified within macrohaplogroup L, contained an A at nucleotide position 3,594	Gonder <i>et al.</i> , 2007

Sequences were assigned haplogroup designations based on the classification scheme described in Section 4.9, Figure 4.1.

After the exclusion of the four sequences listed in Table 4.9, dataset 1a was compiled. This dataset consisted of 429 macrohaplogroup L sequences from individuals of African descent residing in African countries and in countries outside of Africa, i.e. “indigenous black Africans” and “blacks of African ancestry”. All other sequence datasets (refer to Sections 4.10.1, 4.10.2 and 4.10.3) are subsets of dataset 1a. Dataset 2a contains 242 sequences from individuals of African descent residing in African countries at the time of sample collection, i.e. “indigenous black Africans”. A total of 187 sequences obtained from individuals of African descent residing in countries outside of Africa were excluded from dataset 1a to compile dataset 2a (refer to Appendix H). The excluded sequences consisted of one sequence from Jordan (Maca-Meyer *et al.*, 2001), 42 sequences from the Dominican Republic (Kivisild *et al.*, 2006; Torroni *et al.*, 2006), 140 sequences from African Americans (Just *et al.*, 2008), and three sequences from Hispanics (Just *et al.*, 2008).

Dataset 3a consists of all available coding region Khoi-San L sequences and contained the 30 Khoi-San sequences generated in the current investigation, 13 Khoi-San sequences generated previously in the CGR (Mouton, 2003), 12 San sequences (Ingman *et al.*, 2000; Kivisild *et al.*, 2006; Gonder *et al.*, 2007) and three Khoi-San sequences (Kivisild *et al.*, 2006). Dataset 3b contained all available complete Khoi-San L sequences and consisted of the 43 Khoi-San sequences (Mouton, 2003; current investigation) and 10 San sequences (Ingman *et al.*, 2000; Gonder *et al.*, 2007). Information regarding datasets 1a, 2a, 3a and 3b are listed in Table 4.10.

For the purpose of the estimation of the transition/transversion ratio (refer to Section 4.12.3), representative subsets of datasets 1a and 2a, namely datasets 1c and 2c, were compiled. Dataset 1c was compiled as follows:

- a. Exclude a number of sequences from dataset 1a (refer to Appendix H):
 - i Sequences that could not be completely classified.
 - ii Sequences for which there was a discrepancy between the reported haplogroup and the haplogroup assigned in the current investigation.
 - iii Sequences which contain nucleotides that have been assigned ambiguity codes (N, R, Y, S, W, K or M). The inclusion of sequences containing ambiguity codes increases computing time for the DNAPARS program considerably.
- b. Determine the percentage of each haplogroup in dataset 1a after exclusion of the above sequences. Convert the percentage to a number of sequences by multiplying the percentage by the required size of the representative subset (i.e. 50) and rounding off to the nearest whole number.
- c. Divide dataset 1a into groups based on haplogroup designation. Select the number of sequences required from each haplogroup through simple random sampling performed using Statistica version 7.1 (StatSoft, Inc.).
- d. Compile dataset 1c by including the sequences selected in the previous step in a dataset with the rCRS and the outgroup.

Dataset 2c was compiled from dataset 2a in a similar fashion as that described above for dataset 1c. The same list of sequences were excluded (refer to Appendix H). However, 17 of these sequences were excluded previously when dataset 2a was compiled from dataset 1a (refer to Appendix H). Dataset 1c consisted of 49 sequences and dataset 2c contained 51 sequences (refer to Table 4.10).

4.10.2 Compilation of datasets used for statistical analyses

Dataset 1a was used for the compilation of the individual sequence datasets used for statistical analyses (refer to Section 4.14). Three mitochondrial sequences were excluded from this dataset prior to the compilation of the different datasets for statistical analyses. These three sequences were DQ112792 from Eritrea, DQ112932 from Ethiopia and DQ112962 from a Tuareg individual (Kivisild *et al.*, 2006). These sequences were excluded from dataset 1a based on a discrepancy between their haplogroup designation based on the classification scheme used in the current investigation (refer to Section 4.9)

and the haplogroup assignment suggested by the clustering of the sequences in the global phylogenetic trees (refer to Section 5.9). These sequences were classified as L3b\$ (refer to Appendix G), although their location in the global phylogenetic tree indicated that the sequences were closely related to the rCRS which belongs to haplogroup H. Further investigation of the sequences revealed that the sequences were classified as pre-HV1, U1a, and H1 respectively (Kivisild *et al.*, 2006).

Dataset 1b contains 426 macrohaplogroup L sequences from individuals of African descent residing in African countries and in countries outside of Africa. Dataset 2b consists of 239 sequences from individuals of African descent living in African countries at the time of sample collection. A total of 187 sequences obtained from individuals of African descent residing in countries outside of Africa were excluded from dataset 1b to compile dataset 2b (refer to Appendix H). Datasets 4 (4a and 4b), 5, 6, 7 and 8a consist of 43 Khoi-San, 128 L3, 113 L2, 63 L1 and 103 L0 sequences, respectively. A number of sequences were included or excluded from these datasets based on the phylogenetic analysis due to a discrepancy between the haplogroup classification and the location of the sequences in the global phylogenetic trees (refer to Section 5.9). These sequences are listed in Appendix H. Information regarding datasets 1b, 2b, 4a, 4b, 5, 6, 7 and 8a are listed in Table 4.10.

Prior to statistical analyses performed using DNA Sequence Polymorphism (DnaSP) software version 4.10.9 (Rozas *et al.*, 2003), the sequences contained in datasets 1b, 2b, 3a, 4b, 5, 6, 7 and 8a were edited to replace all ambiguity codes (R, Y, S, W, K and M) with N which represents any of the four nucleotides. This was done as DnaSP version 4.10.9 only recognises A, C, G, T and N as nucleotides.

4.10.3 Compilation of the dataset used for construction of the global L0-specific haplogroup network

Dataset 8a was used for the compilation of sequence dataset 8b which was used to construct the global L0-specific haplogroup network (refer to Section 4.13). A number of sequences were excluded from dataset 8a to compile dataset 8b. These sequences included one L0 sequence from Sudan (Kivisild *et al.*, 2006) and nine L0 sequences from South Africa and Tanzania (Gonder *et al.*, 2007) and were excluded due to the presence of nucleotide alterations which were inconsistent with the most probable haplogroup classification according to the classification scheme used in the current investigation (refer

to Section 4.9). A list of the excluded sequences is provided in Appendix H and information regarding dataset 8b is listed in Table 4.10. Dataset 8b consisted of a total of 100 L0 sequences which included the 54 L0 sequences used by Ruiz-Pesini *et al.* (2007) to construct their L0 haplogroup network, as well as an additional 46 L0 sequences consisting of 22 Southern African Khoi-San L0 mtDNA sequences generated in the current investigation, eight L0 African-American sequences (Just *et al.*, 2008), one L0 sequence from the Dominican Republic (Torroni *et al.*, 2006) and 15 L0 sequences from southern Africa and Tanzania (Gonder *et al.*, 2007).

4.10.4 Sequence datasets used in the current investigation

The sequence datasets used for all analyses are listed in Table 4.10. For the purpose of the current investigation, datasets will be referred to by using the dataset number and a short description e.g. dataset 1a (All L Sequences).

Table 4.10 Description of all datasets used in the current investigation for analyses

Dataset	Description	No. of sequences	Source
1a	All L sequences (global) Individuals of African descent residing in African countries and in countries outside of Africa, i.e. "indigenous black Africans" and "blacks of African ancestry"	429 ^a (coding region)	Current investigation; Mouton, 2003; GenBank [®]
1b	All L sequences (global) Individuals of African descent residing in African countries and in countries outside of Africa, i.e. "indigenous black Africans" and "blacks of African ancestry"	426 (coding region)	Current investigation; Mouton, 2003; GenBank [®]
1c	All L sequences (representative subset)	49 ^a (coding region)	Current investigation; Mouton, 2003; GenBank [®]
2a	All Africa L sequences Individuals of African descent residing in African countries, i.e. "indigenous black Africans"	242 ^a (coding region)	Current investigation; Mouton, 2003; GenBank [®]
2b	All Africa L sequences Individuals of African descent residing in African countries, i.e. "indigenous black Africans"	239 (coding region)	Current investigation; Mouton, 2003; GenBank [®]
2c	All Africa L sequences (representative subset)	51 ^a (coding region)	Current investigation; Mouton, 2003; GenBank [®]
3a	All Khoi-San L sequences L0, L1, L2 and L3 sequences	58 ^a (coding region)	Current investigation; Mouton, 2003; GenBank [®]
3b	All Khoi-San L sequences L0, L1, L2 and L3 sequences	53 ^a (complete sequence ^b)	Current investigation; Mouton, 2003; GenBank [®]

Table 4.10 Continued...

Dataset	Description	No. of sequences	Source
4a	Khoi-San L sequences !Kung ethnic group: L0, L1, L2 and L3 sequences	43 (complete sequence ^b)	Current investigation; Mouton, 2003
4b	Khoi-San L sequences !Kung ethnic group: L0, L1, L2 and L3 sequences	43 (coding region)	Current investigation; Mouton, 2003
5	L3 sequences (subset of All L sequences dataset)	128 (coding region)	Current investigation; Mouton, 2003; GenBank [®]
6	L2 sequences (subset of All L sequences dataset)	113 (coding region)	Current investigation; Mouton, 2003; GenBank [®]
7	L1 sequences (subset of All L sequences dataset)	63 (coding region)	Current investigation; Mouton, 2003; GenBank [®]
8a	L0 sequences (subset of All L sequences dataset)	103 (coding region)	Current investigation; Mouton, 2003; GenBank [®]
8b	L0 sequences (subset of L0 sequences dataset)	100 (coding region)	Current investigation; Mouton, 2003; GenBank [®]

GenBank[®] refers to mitochondrial DNA genome sequences available in and retrieved from the GenBank database up to June 2007. ^a = number of sequences excluding the rCRS and the outgroup sequence. ^b = complete sequence consists of the control region and the coding region. Colours: **yellow** = datasets used for phylogenetic analyses (refer to Section 4.12); **green** = datasets used for statistical analyses (refer to Section 4.14); **orange** = dataset used for construction of a global L0-specific haplogroup network (refer to Section 4.13).

4.11 SIGNIFICANT FIGURES

Significant figures are defined as the minimum number of digits required to write a given value in scientific notation without a loss of accuracy (Cann, 2003). The most significant figure is the digit furthest to the left as it is the most precise, while the least significant figure is the digit furthest to the right as it is the least precise.

Calculations involving complex computational algorithms performed using computers frequently provide values with large numbers of decimals which are not all significant, and are beyond the sensitivity with which the measurement could be made. For this reason, and for ease of performing subsequent calculations with the obtained values, the values are rounded off to the nearest significant number. Rounding up is generally avoided as it leads to inaccuracy and errors. Cann (2003) provides some general rules for rounding off:

- If the digit following the figure to be the last digit is less than five, drop the digit and all figures to the right of it.
- If the digit following the figure to be the last digit is more than five, increase the digit to be rounded by one.

- c. If the digit following the figure to be the last digit is equal to five, round the preceding figure off so that it is even.

In the current investigation values obtained using computer programs in phylogenetic and statistical analyses (refer to Sections 4.12 and 4.14) were rounded off to two decimals (e.g. nucleotide diversity and standard deviation), three decimals (e.g. transition/transversion ratio, Tajima's D value, Fu and Li's D^* and F^* values) or four decimals (raggedness statistic and Fu's F_s statistic). The rounding off for these values was according to the convention that is evident in the literature regarding the reporting of these specific values.

4.12 PHYLOGENETIC ANALYSIS

The phylogenetic relationships between the Khoi-San mtDNA sequences generated in the current investigation, mtDNA sequences obtained from other studies (refer to Section 4.8) and the rCRS were investigated through the construction of phylogenetic trees. All the sequences used in the phylogenetic analysis, as well as other relevant information regarding the sequences, are listed in Appendix G. Mitochondrial sequences consisting of the coding region only (bp 577 to 16,023) were used in the analyses (refer to Section 4.10.1).

4.12.1 Sequence alignment

The mitochondrial sequences in each dataset were aligned using ClustalX Multiple Sequence Alignment Program version 1.83 (Thompson *et al.*, 1997). The default settings of the pairwise alignment parameters and the multiple alignment parameters were used for all sequence alignments. Three output files, a CLUSTAL format file (.aln), a NBRF/PIR format file (.pir) and a PHYLIP format file (.phy) were generated for each alignment. The CLUSTAL format output file of all sequence alignments generated was evaluated through visual inspection and edited if deemed necessary through the use of BioEdit version 5.0.9 (Hall, 1999). If editing of an alignment file was performed, the file was saved as a FASTA format file (.fas) and the file was converted to the three formats previously mentioned through the use of ClustalX 1.83 (Thompson *et al.*, 1997). Gaps in alignments were indicated by "-", missing sequence data was denoted using "?" and unknown nucleotides were designated by "N" which represents any of the four nucleotides (A, C, G or T).

4.12.2 Selection of random number seed

A number of the programs contained in the Phylogeny Inference Package (PHYLIP) version 3.66 (Felsenstein, 1989) used for phylogenetic analysis requires the user to provide a random number seed. The random number seed is used to start the random number generator and should conform to a number of criteria (Felsenstein, 2006). The random number seed must be an integer (greater than zero and less than 32,767) and of the form $4n + 1$, i.e. a remainder of one is left when dividing by four. This can be judged by looking at the last two digits of the integer and any odd number can be used. Due to the apparent unpredictable nature of random numbers, it is not possible to distinguish statistically between the numbers produced from one seed and the numbers produced by another seed (Felsenstein, 2006).

Random number seeds were generated through the use of the Random Integer Generator (<http://www.random.org/integers/>, 2007). A list of random integers, arranged in columns, was obtained using this program. Random number seeds were selected from this list by reading the numbers in each row and selecting numbers conforming to the above criteria.

4.12.3 Estimation of transition/transversion ratio

Unequal rates of transitions and transversions were corrected for in each sequence dataset used in phylogenetic analysis by calculating a transition/transversion ratio through the use of the PHYLIP version 3.66 (Felsenstein, 1989). The method used was adapted from Tuimala (2005) and the DNAPARS program in the PHYLIP package was used for this purpose. If the number of sequences in the dataset exceeded 100, a representative subset of the original dataset was first compiled as described in Section 4.10.1. Sequence alignment was performed as described in Section 4.12.1 for all datasets, including the representative datasets, prior to the estimation of the transition/transversion ratio. The PHYLIP format alignment file of the dataset under investigation was used for the estimation of the transition/transversion ratio which involved two steps.

The DNAPARS program estimates phylogenies based on nucleic acid sequences by using the parsimony method, and also provides the number of alterations required to produce the tree. In the first step, the DNAPARS program was used to determine the number of changes, transitions as well as transversions, observed in the dataset. The jumble option (J) was selected to allow randomisation of the input order of the sequences. The jumble

process was repeated 100 times and a random number seed (refer to Section 4.12.2) was supplied for this process. The outgroup option (O) was selected as the estimated ratio is dependent on the specific outgroup used for the subsequent phylogenetic analyses. The outgroup sequence used was specified by providing the number representing the position of the outgroup sequence in the dataset. The number of changes required to produce the tree was recorded.

In the second step, the DNAPARS program was used to determine the number of transversions observed in the dataset. The same settings were used as in the first step, and the jumble and outgroup options were selected and specified as before. In addition, the transversion parsimony option (N) was selected and it was specified that only transversions must be counted. The number of changes required to generate the tree was noted. The transition/transversion ratio was subsequently calculated by dividing the number of changes (transitions and transversions) observed in step one by the number of changes (transversions only) observed in step two.

4.12.4 Estimation of gamma shape parameter

Analyses of reduced-median-networks of the coding region of mtDNA sequences have led to the finding of extensive homoplasy within the coding region which suggests that there is site variability in the rate of sequence evolution within the coding region (Herrnstadt *et al.*, 2002a). Differences in substitution rates among nucleotide sites were corrected for by using a gamma distribution model. The gamma shape parameter (α), i.e. the value that determines the shape of the gamma distribution, was calculated for each dataset used in the phylogenetic analysis by using the GZ-Gamma software package (Gu and Zhang, 1997). The gz-DNA program in the GZ-Gamma package was used for this purpose. Sequence alignment (refer to Section 4.12.1), as well as the estimation of the transition/transversion ratio (refer to Section 4.12.3) were performed for the dataset under investigation prior to the estimation of the gamma shape parameter. The PHYLIP and NBRF/PIR format alignment files of the dataset under investigation were used for the estimation of the gamma shape parameter which involved three steps.

In the first step, an unrooted phylogenetic tree in the Newick format (Felsenstein, 2006) was generated to obtain a tree topology for the sequences contained in the dataset. This tree was produced by performing sequential analyses using two programs contained in the PHYLIP version 3.66 package, namely DNADIST and NEIGHBOR. The DNADIST

program computes distances between nucleic acid sequences. The distance option (D) was selected and the F84 model (Kishino and Hasegawa, 1989; Felsenstein and Churchill, 1996) was chosen (refer to Section 3.2.2.1). The empirical base frequency option (F) was selected and the use of empirical base frequencies was specified. The transition/transversion ratio option (T) was chosen and the previously estimated transition/transversion ratio (refer to Section 4.12.3) was supplied. Subsequently, the NEIGHBOR program, a distance matrix method, was used to generate an unrooted phylogenetic tree without the assumption of a molecular clock. The tree method option (N) was chosen and the neighbour-joining method was selected. In addition, the jumble option (J) was used. It was specified that the jumble process must be repeated 100 times and a random number seed (refer to Section 4.12.3) was supplied. The phylogenetic tree obtained from these analyses was subsequently edited to remove all calculated distances and to replace the sequence names with numbers that correspond to the position of the sequence in the dataset.

The second step involved the editing of the NBRF/PIR format alignment file using BioEdit version 5.0.9 (Hall, 1999). All nucleotide positions from all sequences in the alignment that contained gaps ("-"), missing characters ("?") or ambiguity codes (N, R, Y, S, W, K, M) were removed and the alignment was saved as a FASTA format file. The alignment was edited further to remove all paragraph marks, except the marks between the sequence name and sequence data, and the sequence data and the next sequence name. The third step involved the compilation of a text format file containing the edited FASTA format file (sequence data) and the edited phylogenetic tree (topology of the sequences). This file began with two numbers, the number of sequences in the dataset and the number of nucleotide sites (sequence length). The second line contained the name of the first sequence, and the third line the sequence data of the first sequence, and so on. The last line of the file consisted of the tree topology of the sequences. This text file was used as input file for the gz-DNA program in the GZ-Gamma package, and the value of the gamma shape parameter (alpha) was estimated using option 3. The value of alpha provided in the outfile generated by the program was recorded. This value is used in the form of a coefficient of variation during tree construction via the neighbour-joining method (refer to Section 4.12.5) The coefficient of variation of substitution rate among sites was subsequently calculated by dividing one by the square root of the value of alpha.

4.12.5 Neighbour-joining method

The neighbour-joining method is based on a method proposed by Saitou and Nei in 1987. Neighbour-joining trees were generated for the different sequence datasets through the use of PHYLIP version 3.66 (Felsenstein, 1989). The sequence datasets were first aligned (refer to Section 4.12.1), thereafter the transition/transversion ratio (refer to Section 4.12.3) and the gamma shape parameter (refer to Section 4.12.4) were estimated for each dataset. The neighbour-joining tree for each dataset was constructed by using the PHYLIP format alignment file as input file. In order to perform bootstrap analyses to test the statistical support for the branching structure present within the generated neighbour-joining trees, sequential analyses were performed using four programs contained in the PHYLIP package, namely SEQBOOT, DNADIST, NEIGHBOR and CONSENSE. The resulting confidence scores were presented as percentages on the branches of the trees.

Firstly, the SEQBOOT program was used to produce multiple datasets from the original sequence dataset by bootstrap re-sampling. The type of data used for analysis was specified by selecting the data type option (D) and choosing molecular sequences. The type of algorithm to be employed for re-sampling of the dataset was selected by choosing option J and selecting bootstrap. The number of bootstrap replicates to be performed was assigned by selecting the replication option (R) and specifying the use of 1,000 replicates. An appropriate random number seed was also provided for the bootstrap re-sampling process (refer to Section 4.12.3). Secondly, the DNADIST program was used to compute distances between nucleic acid sequences. The multiple datasets option (M) was selected and the number of datasets to be analysed (1,000) was provided. The distance option (D) was selected and the F84 model (Kishino and Hasegawa, 1989; Felsenstein and Churchill, 1996) was chosen. This model takes into account different frequencies of the four nucleotides, different rates of transition and transversion and unequal rates of substitution at different sites (refer to Section 3.2.2.1). These parameters were defined through the selection of specific options in the DNADIST program prior to its use. The empirical base frequency option (F) was selected and the use of empirical base frequencies was specified. The transition/transversion ratio option (T) was chosen and the previously estimated transition/transversion ratio (refer to Section 4.12.3) was provided. The gamma distribution option (G) was selected and the estimated value of the coefficient of variation of substitution rate among sites (refer to Section 4.12.4) was supplied. Thereafter, the NEIGHBOR program, a distance matrix method, was used to produce an unrooted

phylogenetic tree without the assumption of a molecular clock. The tree method option (N) was selected and the neighbour-joining method was chosen. As before, the multiple datasets option (M) was selected and the number of datasets to be analysed (1,000) was provided. In addition, the jumble option (J) was selected and it was specified that the jumble process must be repeated 100 times using the provided random number seed (refer to Section 4.12.3). The last program used was CONSENSE which computes consensus trees and is used as the final step in tree construction when bootstrap analyses are performed. The type of consensus tree was selected by specifying the majority rule (extended) choice in the consensus type option (C).

The neighbour-joining tree generated by CONSENSE was evaluated using TreeView version 1.6.6 (Page, 1996). This program allows the user to visualise, edit and save trees constructed through the use of the PHYLIP package in a desired image format (.emf or .wmf) or in the Newick format (.phy) for further manipulation or editing. The neighbour-joining trees were rooted using the *Pan troglodytes* sequence, GenBank[®] accession number NC_001643 (refer to Appendix G), as an outgroup. The rooted tree was saved in Newick format, retaining the branch length information and outgroup rooting, and edited further for presentation purposes by using the Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0 (Tamura *et al.*, 2007).

4.12.6 Maximum parsimony method

The maximum parsimony method is based on a method described by Fitch in 1971. Maximum parsimony trees were generated for the different sequence datasets through the use of MEGA software version 4.0 (Tamura *et al.*, 2007). The sequence datasets were aligned (refer to Section 4.12.1), and the PHYLIP format alignment file was converted to a MEGA alignment file (.meg) using the MEGA version 4 package.

The MEGA version 4 package allows the specification of a number of different analysis preferences. The type of data to be analysed was specified as nucleotide sequences and the sequence data was identified as protein encoding. The genetic code assigned to the sequences was that which was applicable to the mitochondria of vertebrates. Phylogeny reconstruction analysis was selected and the method chosen was maximum parsimony. Nucleotide positions within the sequence alignment that contained gaps and/or missing sequence data were excluded from all analyses. A bootstrap test of the phylogeny of the maximum parsimony tree was performed, 1,000 replicates were specified, and the random

number seed supplied by the program was used. The resulting confidence scores were presented as percentages on the branches of the produced maximum parsimony trees. The maximum parsimony tree generated by MEGA version 4 was outgroup rooted using the *Pan troglodytes* sequence, GenBank® accession number NC_001643 (refer to Appendix G) and edited further for presentation purposes.

4.13 CONSTRUCTION OF L0-SPECIFIC HAPLOGROUP NETWORKS

The branching structure present in haplogroup L0 was investigated through the construction of a Southern African Khoi-San L0-specific haplogroup network consisting of the Khoi-San mtDNA sequences generated in the current investigation, as well as a global L0-specific haplogroup network containing the L0 sequences generated in the current investigation in addition to L0 sequences obtained from other sources (refer to Section 4.8). All the sequences used in the construction of the L0-specific haplogroup networks, as well as other relevant information regarding the sequences, are listed in Appendix G. Mitochondrial sequences consisting of the coding region only (bp 577 to 16,023) were used in the construction of the L0-specific haplogroup networks (refer to Section 4.10.3).

4.13.1 Southern African Khoi-San L0-specific haplogroup network

An L0-specific haplogroup network was constructed using 22 of the 30 complete Khoi-San sequences generated in the current investigation. The network was constructed by using the rCRS (Andrews *et al.*, 1999) as a reference, and identifying alterations (synonymous substitutions, non-synonymous substitutions, insertions and deletions) within the coding region (bp 577 to 16,023) of each of the 22 L0 mtDNA sequences. Alterations identified within the control region of sequences were excluded in the analysis as the control region is prone to reverse nucleotide changes. A maximum parsimony tree was generated for the 22 L0 Khoi-San sequences (refer to Section 4.12.6) and was used as a guide for the topological structure of the network. The network was subsequently assembled by using the identified alterations to arrange the 22 mtDNA sequences into smaller and smaller groups based on shared alterations, until a point was reached where a sequence was on its own at the tip of a branch.

4.13.2 Global L0-specific haplogroup network

Each of the L0 mtDNA sequences included in dataset 8b (refer to Section 4.10.3) was compared to the rCRS (Andrews *et al.*, 1999) and a list was compiled of the alterations (synonymous substitutions, non-synonymous substitutions, insertions and deletions) observed in the coding region (bp 577 to 16,023). The control region was excluded for all sequences analysed due to the increased possibility of reverse mutations within this region (Mishmar *et al.*, 2003). The list compiled for each mtDNA sequence was compared to the L0 haplogroup network constructed by Ruiz-Pesini *et al.* (2007) and new branches were added to the existing network where new alterations were identified within a group of sequences. These groups of alterations were subsequently divided into smaller groups until the tip of a branch was reached where the remaining alteration(s) defined the mtDNA sequence of a specific individual or individuals. Branches present within each sub-haplogroup were named according to the estimated TMRCA for that branch (refer to Section 4.14.7), e.g. branch A is the oldest branch in the sub-haplogroup, while B3 is the youngest of the three groups belonging to branch B of the same sub-haplogroup.

4.14 STATISTICAL ANALYSIS

A number of statistical analyses were performed to obtain information regarding the Khoi-San mitochondrial sequences and the other L-specific sequences studied in the current investigation. Basic sequence statistics such as nucleotide composition and G + C content were determined for the complete Khoi-San mitochondrial genome sequences in dataset 4a (!Kung Sequences) as these sequences have not been described previously (refer to Section 4.14.1). In addition, codon usage was determined for the coding region only Khoi-San mitochondrial genome sequences in dataset 4b (!Kung Sequences) as codon usage for these sequences has not been investigated previously (refer to Section 4.14.2). Genetic diversity measures (refer to Section 4.14.3) were used to evaluate the genetic diversity within specific datasets. Tajima's *D* test (Tajima, 1989) and Fu and Li's *D** and *F** tests (Fu and Li, 1993) were performed on the same datasets to test the assumption that all mutations are neutral. Datasets, in which deviation from neutrality was observed, were investigated further to determine if it could be attributed to population size changes or the effect of selection. Population size changes were analysed by determining the frequency distribution of pairwise sequence differences (refer to Section 4.14.5.1) and calculating statistical parameters to detect population expansion (refer to Section 4.14.5.2). The effect of selection on the mitochondrial genome sequences

was investigated by analysing synonymous and non-synonymous substitutions (refer to Section 4.14.6.1) and determining conservation indices (refer to Section 4.14.6.2.)

4.14.1 Basic sequence statistics

The relative frequencies of the four nucleotides for all sequences in dataset 4a (!Kung Sequences) were determined through the use of MEGA software version 3.1 (Kumar *et al.*, 2004). The calculation of G + C content for all Khoi-San mitochondrial sequences in dataset 4a was performed using DnaSP software version 4.10.9 (Rozas *et al.*, 2003).

4.14.2 Codon usage

Codon usage for synonymous codons present in the coding region only Khoi-San mitochondrial sequences in dataset 4b was determined using four different statistical codon bias measures. These measures are based on the null hypothesis of the random use of different synonymous codons due to the degeneracy of the genetic code.

4.14.2.1 Relative synonymous codon usage

Relative synonymous codon usage or RSCU (Sharp *et al.*, 1986) describes the observed frequency of a codon divided by the frequency expected under the assumption of equal usage of synonymous codons. RSCU values and codon frequencies were calculated using MEGA software version 3.1 (Kumar *et al.*, 2004). A RSCU value of one indicates that the frequency of the specific codon is that expected for equal codon usage. Values less than one (or more than one) indicate that the specific codon is used less often (or more often) than expected. Extreme codon usage bias is indicated by very high or very low RSCU values. The cut-off values for these very low or very high values were determined by calculating the arithmetic mean for a range of RSCU values and determining the standard deviation (SD). The equations used to calculate the arithmetic mean and standard deviation is presented in Equation 4.3. A RSCU value was considered very low (or very high) if it fell outside of the range of the mean plus/minus one standard deviation, i.e. the cut-off value for very low values (≤ 0.3) was calculated by subtracting the value of the standard deviation from the mean, while the cut-off value for very high values (≥ 1.7) was computed by adding the value of the standard deviation to the mean.

Equation 4.3 Calculation of arithmetic mean and standard deviation**A. Arithmetic mean**

$$\text{Mean} = \sum y / n$$

B. Standard deviation

$$\text{SD} = \sqrt{\sum (y - \bar{y})^2 / n - 1}$$

Σ = summation; y = value of observation; n = sample size (number of sequences); SD = standard deviation; $\sum (y - \bar{y})^2$ = sum of squared deviations. Adapted from Samuels (1989).

4.14.2.2 The effective number of codons

The effective number of codons (N_C) represents the "effective" number of codons that are used (Wright, 1990). The value of N_C ranges from 20 (maximum codon bias, i.e. only one codon is used for each amino acid) to 61 (no codon bias, i.e. all synonymous codons for each amino acid are used equally). The value of N_C was determined for dataset 4b using DnaSP software version 4.10.9 (Rozas *et al.*, 2003).

4.14.2.3 Codon bias index

Codon bias index (CBI) is a measure of the deviation from the equal use of synonymous codons (Morton, 1993), and the values range from zero (uniform use of synonymous codons) to one (maximum codon bias). The value of CBI was calculated for dataset 4b using DnaSP software version 4.10.9 (Rozas *et al.*, 2003).

4.14.2.4 Scaled chi square

The scaled chi square ($S\text{Chi}^2$) is a measure of the departure from the equal use of synonymous codons estimated by a chi square statistic which is scaled by dividing by the total number of codons analysed, excluding those codons coding for a unique amino acid (Shields *et al.*, 1988). The $S\text{Chi}^2$ value is calculated using Yates' correction and the average G + C percentage. The application of the Yates' correction (continuity correction) to the scaled chi square method reduces the overestimation of the degree of codon bias that is obtained for short sequences without the correction for conditions of no or medium levels of bias among synonymous codons (Comeron and Aguade, 1998). The scaled chi square value therefore provides a simple measure of general synonymous codon usage bias which is independent of gene length. The higher the $S\text{Chi}^2$ value, the higher the degree of codon usage bias, while zero indicates perfectly uniform codon usage. The

value of $S\text{Chi}^2$ was determined for dataset 4b using DnaSP software version 4.10.9 (Rozas *et al.*, 2003).

4.14.3 Genetic diversity measures

The genetic diversity within datasets 1b, 2b, 3a, 4b, 5, 6, 7 and 8a, consisting of coding region only sequences (refer to Section 4.10.2), was evaluated by determining the number of segregating sites (S), the average number of nucleotide differences (k) and the nucleotide diversity (π). These genetic diversity measures were calculated using DnaSP software version 4.10.9 (Rozas *et al.*, 2003). Nucleotide positions within the datasets that contained alignment gaps or missing data were completely excluded from all analyses.

4.14.3.1 The number of segregating sites

A nucleotide position is considered segregating if there are two or more nucleotides at that position upon comparison of a number of sequences. The number of segregating (polymorphic) sites within each dataset was calculated as described (refer to Section 4.14.3) and subsequently used in the calculation of Tajima's D test (refer to Section 4.14.4.1).

4.14.3.2 The average number of nucleotide differences

The average number of nucleotide differences was determined as described (refer to Section 4.14.3) by using Equation 4.4. This equation calculates the average number of nucleotide differences per site between two or more nucleotide sequences.

Equation 4.4 Average number of nucleotide differences

$$k = \sum_{i < j} k_{ij} / (n / 2)$$

k = average number of pairwise nucleotide differences between the sequences; k_{ij} is the number of nucleotide differences between the i^{th} and j^{th} sequences; n = sample size (number of sequences). Adapted from Tajima (1983).

4.14.3.3 Nucleotide diversity

Nucleotide diversity is defined as the average number of nucleotide differences per position between two sequences. The standard deviation was determined by calculating the square root of the variance. The nucleotide diversity, its sampling variance, and the

standard deviation, were determined as described (refer to Section 4.14.3) by using Equation 4.5.

Equation 4.5 Calculation of nucleotide diversity, sampling variance and standard deviation

A. Nucleotide diversity

$$\pi = \sum_{i \neq j} d_{ij} / [n(n-1)]$$

B. Sampling variance

$$V(\pi) = 1 / [n(n-1)]^2 [\sum_{i \neq j} V(d_{ij}) + \sum_{i \neq j} \sum_{k \neq l} \text{Cov}(d_{ij}, d_{kl})]$$

C. Standard deviation

$$SD = \sqrt{V(\pi)}$$

Where:

$$V(d_{ij}) = 9p_{ij}(1-p_{ij}) / [(3-4p_{ij})^2 m]$$

$$\text{Cov}(d_{ij}, d_{kl}) = 9p_{ij}(1-p_{ij}) / [(3-4p_{ij})^2 m]$$

$$p_{ij} = (3/4)[1 - e^{-4b/3}]$$

π = nucleotide diversity; $\sum$ = summation for all pairwise comparisons; d_{ij} is an estimate of the number of nucleotide substitutions per site between DNA sequences i and j ; n is the number of DNA sequences examined; $V(\pi)$ = sample variance of the nucleotide diversity; $\text{Cov}(d_{ij}, d_{kl})$ = covariance; m = number of nucleotides examined per sequence; b = branch length. Adapted from Nei and Miller (1990).

4.14.4 Deviation from the neutral theory of molecular evolution

Under the neutral theory of molecular evolution (refer to Section 2.5.5), the majority of mutations are viewed as neutral or nearly neutral (Kimura, 1968). In order to determine if the genetic diversity present in the sequence datasets is due to neutral alterations, it was therefore necessary to test the neutrality of the observed alterations. Different neutrality tests were employed, namely Tajima's D test and Fu and Li's D^* and F^* tests. Calculations were performed on datasets 1b, 2b, 3a, 4b, 5, 6, 7 and 8a (refer to Section 4.10.2) consisting of the coding region of mtDNA genome sequences using DnaSP software version 4.10.9 (Rozas *et al.*, 2003). Statistical significance (P -value) was determined at a level of 5%, i.e. a 5% probability existed that the result obtained was generated by chance. The null hypothesis of the neutrality of mutations was rejected at significant P -values ($P < 0.05$).

4.14.4.1 Tajima's D test

Tajima's D test was used to test the hypothesis that all mutations are neutral. This method was developed according to the assumption of random mating within a population, the absence of both selection and recombination, and that the infinite site model applies, i.e. a very large number of independent sites are available for mutation (Kimura, 1969).

The method is based on the relationship between two estimates of genetic variation at the DNA sequence level, namely the number of segregating sites and the average number of nucleotide differences estimated from pairwise comparison. Differences observed between these two estimates of genetic variation allow for the detection of selection among sequences. The method has some limitations which include the fact that DNA sequences used must be from a random sample of the population under investigation, the population's equilibrium status must be taken into consideration, and if a selectively neutral site is linked to a site where natural selection is active, the value of D for the neutral site might be affected by the selected site.

The test was performed using DnaSP software version 4.10.9 (Rozas *et al.*, 2003). The mathematical equation that was used is indicated in Equation 4.6. Sites containing alignment gaps (or sites with missing data) were not used in the analysis. Statistical significance was determined through the use of confidence limits of D (two tailed test) which are published in "Table 2" in the article by Tajima in 1989. The confidence limits presented in this publication were obtained by assuming that D follows the beta distribution (Tajima, 1989).

Equation 4.6 Tajima's D statistic

$$D = \frac{k - S / a_1}{\sqrt{e_1 S + e_2 S(S - 1)}}$$

Where:

$$a_1 = \sum (1 / i) \text{ from } i = 1 \text{ to } n - 1$$

$$e_1 = c_1 / a_1$$

$$e_2 = c_2 / a_1^2 + a_2$$

$$c_1 = b_1 - 1 / a_1$$

$$c_2 = b_2 - n + 2 / a_1 n + a_2 / a_1^2$$

$$a_2 = \sum (1 / i^2) \text{ from } i = 1 \text{ to } n - 1$$

$$b_1 = n + 1 / 3(n - 1)$$

$$b_2 = 2(n^2 + n + 3) / 9n(n - 1)$$

k = average number of pairwise nucleotide differences between the sequences; S = number of segregating (polymorphic) sites; n = sample size (number of sequences). Adapted from Tajima (1989).

Tajima's D values less than zero (negative values) indicate the presence of population growth or the influence of positive selection, values equal to zero indicate that the mutations were neutral, while values greater than zero (positive values) indicate the presence of population subdivision or the effect of balancing selection (Jobling *et al.*, 2004).

4.14.4.2 Fu and Li's D^* and F^* tests

Fu and Li's D^* and F^* tests were used to test the hypothesis that all mutations are neutral. These methods were developed under the assumption of random sample collection from a randomly mating population of constant size in the absence of migration events and recombination, wherein mutations are selectively neutral. Fu and Li's D^* and F^* tests consider the distribution of the total number of mutations in a genealogy of a random sample of sequences (Fu and Li, 1993) and is based on the differences between the number of mutations appearing only once among the sequences, and the total number of mutations (Fu and Li's D^*), or the average number of nucleotide differences between pairs of sequences (Fu and Li's F^*). Selection is detected by comparing the observed numbers of mutations in internal branches (i.e. the older part of the genealogy containing "older" mutations) and external branches (i.e. the younger part of the genealogy containing "new" mutations) with the expected numbers of mutations under the assumption of selective neutrality.

The test was performed using DnaSP software version 4.10.9 (Rozas *et al.*, 2003). The mathematical equations that were used are indicated in Equation 4.7. Sites containing alignment gaps (or sites with missing data) were not used in the analysis. The critical values obtained by Fu and Li in 1993 (two tailed test, Tables 2 and 4) are used to determine the statistical significance of the D^* and F^* test statistics.

Fu and Li's D^* and F^* values less than zero (negative values) indicate the presence of population growth or the influence of positive selection, values equal to zero indicate that the mutations were neutral, while values greater than zero (positive values) indicate the influence of balancing selection (Fu and Li, 1993).

Equation 4.7 Fu and Li's D^* and F^* test statistics**A. Fu and Li's D^* test**

$$D^* = \frac{\left(\frac{n}{n-1} \right) \eta - a_n \eta_s}{\sqrt{u_{D^*} \eta + v_{D^*} \eta^2}}$$

Where:

$$a_n = \sum (1/k) \text{ from } k = 1 \text{ to } n - 1$$

$$u_{D^*} = (n/n-1) [a_n - (n/n-1) - v_{D^*}]$$

$$v_{D^*} = [(n/n-1)^2 b_n + a_n^2 d_n - 2(na_n(a_n+1))/(n-1)^2] / (a_n^2 + b_n)$$

$$b_n = \sum (1/k^2) \text{ from } k = 1 \text{ to } n - 1$$

$$d_n = [c_n + (n-2/(n-1)^2) + 2/n-1] [(3/2) - ((2a_{n+1}-3)/(n-2)) - 1/n]$$

$$c_n = 2 [na_n - 2(n-1)/(n-1)(n-2)]$$

B. Fu and Li's F^* test

$$F^* = \frac{\Pi_n - \left(\frac{n-1}{n} \right) \eta_s}{\sqrt{u_{F^*} \eta + v_{F^*} \eta^2}}$$

Where:

$$u_{F^*} = [((n/n-1) + (n+1/3(n-1)) - 2(2/n(n-1)) + 2(n+1/(n-1)2)(n-1)^2)(a_{n+1}-2n/n+1)] / a_n - v_{F^*}$$

$$a_n = \sum (1/k) \text{ from } k = 1 \text{ to } n - 1$$

$$v_{F^*} = [d_n + 2(n^2 + n + 3)/9n(n-1) - 2(1/n-1)(4b_n - 6 + 8/n)] (a_n^2 + b_n)$$

$$d_n = [c_n + (n-2/(n-1)^2) + 2/n-1] [(3/2) - ((2a_{n+1}-3)/(n-2)) - 1/n]$$

$$b_n = \sum (1/k^2) \text{ from } k = 1 \text{ to } n - 1$$

$$c_n = 2 [na_n - 2(n-1)/(n-1)(n-2)]$$

n = sample size (number of sequences); η_s = number of singletons; η = total number of mutations; Π_n = average number of nucleotide differences between two sequences. Adapted from Fu and Li (1993).

4.14.5 Population size changes

Deviation from neutrality as indicated by negative results obtained from Tajima's D test and Fu and Li's D^* and F^* tests can be attributed to population growth or the effect of selection (Simonsen *et al.*, 1995). The role of population growth can be investigated by studying the distribution produced by the nucleotide differences between pairs of sequences, referred to as the frequency distribution of pairwise differences between sequences, or alternatively, as the mismatch distribution (refer to Section 4.14.5.1). A number of statistical measures were also employed to describe the smoothness of the observed pairwise differences distributions and to detect population expansion processes (refer to Section 4.14.5.2).

4.14.5.1 Frequency distribution of pairwise sequence differences

The frequency distribution of pairwise sequence differences is used to study population size changes. Constant population size, as well as episodes of population growth and

decline can be inferred from the analysis of the frequency distribution of pairwise sequence differences. The analysis was performed using DnaSP software version 4.10.9 (Rozas *et al.*, 2003). The distribution of the observed pairwise nucleotide differences, and the expected values (at equilibrium for no recombination) in a stable population, i.e. with a constant population size, were calculated using the mathematical equation presented in Equation 4.8. The method requires the use of the neutral infinite sites model (Kimura, 1971) with the assumption that the population under investigation is not divided into sub-populations and that no recombination takes place (Slatkin and Hudson, 1991; Rogers and Harpending, 1992).

Equation 4.8 Calculation of the pairwise number of sequence differences under the assumption of constant population size

$$F(i) \approx \frac{\theta^i}{(\theta + 1)^{i+1}} = F_0(1 - F_0)^i$$

Where:

$$F_0(t+1) \approx 1/N + (1 - 2\mu - 1/N)F_0(t)$$

$$F_{i+1}(t+1) \approx (1 - 2\mu - 1/N)F_{i+1}(t) + 2\mu F_i(t)$$

$$\theta = 2N\mu$$

N = size of haploid population (N is the effective size of the female population for mtDNA in higher animals); μ = mutation rate per generation; $\theta = 2N\mu$ is the expected pairwise nucleotide difference; t = time in generations; $F_i(t)$ = the probability that two random neutral genes will differ at exactly i nucleotides in generation t . Adapted from Watterson (1975), Slatkin and Hudson (1991) and Rogers and Harpending (1992).

Nucleotide sites containing alignment gaps (or sites with missing data) were not used in the analyses. The results obtained for the observed and expected values are displayed graphically with the frequency of differences on the Y-axis and the number of pairwise differences on the X-axis. The graph contains two types of data points, a red line in the form of a smooth curve which represents the expected values calculated under the assumption of a constant population size and the absence of recombination, and the observed values which are indicated by a blue line.

4.14.5.2 Statistical parameters to detect population expansion

Three statistical parameters were employed to detect population expansion events, namely the raggedness (rg) statistic (Harpending *et al.*, 1993; Harpending, 1994), the Ramos-Onsins and Rozas (R_2) statistic (Ramos-Onsins and Rozas, 2002) and Fu's F_s statistic (Fu, 1997). The statistics were calculated using DnaSP software version 4.10.9 (Rozas *et al.*, 2003) and all sites containing alignment gaps (or sites with missing data) were excluded from analyses.

4.14.5.2.1 The raggedness statistic

The raggedness statistic is a measure of high frequency variation in a series of numbers and is not based on any theory or model. It quantifies the smoothness of the observed pairwise differences distribution. Low rg values obtained for a dataset indicate that the mismatch distribution is smooth, i.e. population growth has taken place, while high rg values indicate that the mismatch distribution is ragged, i.e. the population has remained constant in size (Harpending *et al.*, 1993; Harpending, 1994). The mathematical equation that was used to determine rg values is indicated in Equation 4.9.

Equation 4.9 Raggedness statistic

$$rg = \frac{\sum_{i=1}^{d-1} (x_i - x_{i+1})^2}{d-1}$$

rg = raggedness statistic; d = maximum number of differences; x = mismatch distribution. Adapted from Harpending (1994).

4.14.5.2.2 The Ramos-Onsins and Rozas statistic

The Ramos-Onsins and Rozas statistic is based on the difference between the number of singleton mutations and the average number of nucleotide differences (Ramos-Onsins and Rozas, 2002). Low values of R_2 in a dataset are obtained when a population expansion event has occurred. The R_2 statistic is defined in Equation 4.10.

Equation 4.10 The Ramos-Onsins and Rozas statistic

$$R_2 = \frac{\left[\sum_{i=1}^n \left(U_i - k/2 \right)^2 / n \right]^{1/2}}{S}$$

R_2 = Ramos-Onsins and Rozas statistic; n = sample size; S = total number of segregating sites; k = average number of nucleotide differences between two sequences; U_i = number of singleton mutations in sequence i . Adapted from Ramos-Onsins and Rozas (2002).

4.14.5.2.3 Fu's F_S statistic

The F_S test statistic (Fu, 1997) is based on the haplotype (gene) frequency distribution and is conditional on the value of theta (Ewens, 1972). The F_S test statistic is inclined to be negative when there is an excess of recent alterations, i.e. an excess of rare alleles, within a dataset. A large negative value for F_S is therefore interpreted as evidence against the neutrality of mutations and as support for a population expansion event. The mathematical

equation that was used to determine values for Fu's F_S statistic is indicated in Equation 4.11.

Equation 4.11 Fu's F_S statistic

$F_S = \ln[S' / (1 - S')]$

Where:

$$S' = p(k_0 \leq k \mid \theta = \theta_\pi) = \sum_{k=k_0} \frac{|S_k| \theta_\pi^k}{S_n(\theta_\pi)}$$

$$S_n(\theta_\pi) = \theta_\pi(\theta_\pi - 1)(\theta_\pi - n + 1)$$

n = sample size; k = number of alleles; $p(k \mid \theta)$ = probability of having k alleles in a sample of n sequences; θ_π = the mean number of nucleotide differences between two sequences; S' = probability of having no fewer than k_0 alleles in a sample provided that $\theta = \pi$; S_k = coefficient of θ_π^k in S_n . Adapted from Fu (1997).

4.14.5.2.4 Statistical significance

The 95% confidence intervals for the three statistical measures were calculated by computer simulations using the coalescent algorithm performed by DnaSP software version 4.10.9 (Rozas *et al.*, 2003). The computer simulations were based on the coalescent process for a neutral infinite-sites model and assumed a large constant population size (Hudson, 1990). The ran1 routine (Press *et al.*, 1992) was used as a source of uniform random deviates, i.e. random numbers uniformly distributed within a specified range. Under the assumption of no intragenic recombination, 1,000 replicates were performed, and the simulations were conditional on the number of segregating sites (S), i.e. the "fixed S " method was employed which involves the generation of random genealogies, followed by placing S mutations on each tree. This method is easy to simulate and appears to be appropriate for testing purposes as the value of theta must be estimated from the data, while S is observed (Rozas and Rozas, 1999; Wall and Hudson, 2001). The P -values generated by this process provide the probability of obtaining values for the different statistics that are equal or lower than the observed values under the hypothesis of constant population size. The hypothesis of constant population size was rejected for $P \leq 0.05$, i.e. at a 5% level of significance.

4.14.6 Determination of the effect of selection on mitochondrial DNA variation

As already discussed, deviation from neutrality can be attributed to population growth or the effect of selection (refer to Section 4.14.5). The effect of selection on mitochondrial sequences belonging to macrohaplogroup L was investigated by comparing the frequency

of replacement mutations in internal branches to the frequency of replacement mutations in terminal branches of phylogenetic networks of each haplogroup (refer to Section 4.14.6.1), or by determining the degree of conservation of every replacement or non-synonymous substitution observed (refer to Section 4.14.6.2).

4.14.6.1 Analysis of synonymous and non-synonymous substitutions

The effect of selection on mitochondrial genome sequences belonging to haplogroups L, L0, L1, L2 and L3 was estimated as described by Ruiz-Pesini *et al.* in 2004. The global L0-specific haplogroup network constructed in the current investigation (refer to Section 4.13) was used to investigate haplogroup L0, while the L1, L2 and L3 sections of the phylogenetic network available on MITOMAP (2008), which was constructed by Ruiz-Pesini *et al.* in 2007, were used to study haplogroups L1, L2 and L3.

The method used is based on the comparison of the frequency of replacement mutations within internal branches to the frequency of replacement mutations within terminal branches of a phylogenetic network. This ratio can be used as a measure of adaptive versus purifying selection in a population. The frequency of replacement mutations, i.e. non-synonymous (NS) substitutions, fluctuates in different parts of the phylogenetic network depending on if they are deleterious (common in terminal branches), or advantageous (common in internal branches), while neutral mutations, i.e. synonymous (S) substitutions, should be uniformly distributed. The frequency of non-synonymous substitutions in the different parts of the phylogenetic network can thus be normalised for different time intervals by dividing by the number of synonymous substitutions. This allows the frequency of replacement mutations (RF) to be compared for internal branches (RF_i) and terminal branches (RF_T). The frequency of replacement mutations was therefore calculated for internal and terminal branches of the phylogenetic network for each haplogroup by counting the total number of non-synonymous substitutions for every branch and dividing by the total number of synonymous substitutions. Only non-synonymous and synonymous substitutions located in the coding region of the mitochondrial genome sequences included in the phylogenetic networks were counted as the phylogenetic networks were constructed using mitochondrial genome sequences consisting of coding region sequence data. Non-synonymous or synonymous substitutions present in tRNA or rRNA genes, deletions, insertions, reversions, confirmed pathogenic mutations and substitutions affecting non-coding nucleotides were excluded. The

calculation employed to determine the replacement mutation frequency is indicated in Equation 4.12.

Equation 4.12 Calculation of the replacement mutation frequency

$$RF = NS / S$$

RF = replacement mutation frequency; NS = non-synonymous substitution; S = synonymous substitution. Adapted from Ruiz-Pesini *et al.* (2004).

The ratio, RF_i/RF_T (I/T), between the internal and the terminal branches for each phylogenetic network was determined. Lower RF_i values indicate the presence of fewer internal non-synonymous substitutions, which implies that most of these alterations have been removed by purifying selection, while higher RF_i values suggest that more non-synonymous substitutions have become established, which reflects the influence of adaptive selection. Therefore, a low I/T ratio indicates the influence of purifying selection, while a high I/T ratio indicates the increased importance of adaptive selection.

Statistical significance was determined by calculating one-sided *P*-values for the I/T ratios by using DnaSP version 4.10.9 software (Rozas *et al.*, 2003) to calculate Fisher's Exact Test (Fisher, 1922). The Fisher's Exact Test is a contingency test of neutrality and is based on the assumption that the data can be written in the form of a contingency table, consists of independent observations and was obtained through random sampling (Motulsky, 1995; Templeton, 1996). The counts of the non-synonymous and synonymous substitutions in the terminal and internal branches used to determine the RF_i and RF_T values can be represented as a 2 x 2 contingency table and the Fisher's Exact Test was subsequently used to test the hypothesis that the observed alterations were strictly neutral, i.e. occurring randomly and at the same rate in the different mtDNA sequences in the phylogenetic network, thereby implying that selection did not drive the loss or gain of alterations. This hypothesis was rejected if $P < 0.05$.

4.14.6.2 Conservation index

The conservation index (CI) describes the degree of conservation of every replacement or non-synonymous substitution observed and can be used to investigate the possible functional significance of variation within genes, as well as to analyse the influence of purifying compared to adaptive selection. The conservation of non-synonymous

substitutions was estimated by using 39 different species (Ruiz-Pesini *et al.*, 2004). The species are listed in Table 4.11.

Table 4.11 Species used to calculate the conservation index

Primates	<i>Cebus albifrons</i> <i>Gorilla gorilla</i> <i>Homo sapiens</i> <i>Hylobates lar</i> <i>Macaca sylvanus</i> <i>Nycticebus coucang</i> <i>Pan paniscus</i> <i>Pan troglodytes</i> <i>Papio hamadryas</i> <i>Pongo pygmaeus</i> <i>Pongo pygmaeus abelii</i> <i>Tarsius bancanus</i>	Other Mammals	<i>Balaenoptera musculus</i> <i>Balaenoptera physalus</i> <i>Bos taurus</i> <i>Canis familiaris</i> <i>Equus asinus</i> <i>Equus caballus</i> <i>Erinaceus europaeus</i> <i>Felis catus</i> <i>Hippopotamus amphibious</i> <i>Lama pacos</i> <i>Loxodonta africana</i> <i>Marcropus robustus</i> <i>Mus musculus</i> <i>Ornitorhynchus anatinus</i> <i>Oryctolagus cuniculus</i> <i>Ovis aries</i> <i>Phoca vitulina</i> <i>Rattus norvegicus</i> <i>Rhinocerus unicornis</i> <i>Sciurus vulgaris</i> <i>Sus scrofa</i> <i>Talpa europaea</i>
	Vertebrates		
	<i>Aligator mississippiensis</i> <i>Cyprinus carpio</i> <i>Gallus gallus</i> <i>Xenopus laevis</i>		
Invertebrates	<i>Drosophila melanogaster</i>		

Adapted from Ruiz-Pesini *et al.* (2004).

To determine the conservation index of the non-synonymous substitutions under investigation, the amino acid encoded at that position was investigated. Two classifications were used to describe amino acids. The wild type amino acid for each non-synonymous substitution under investigation was defined as the amino acid present in most of the global human phylogeny (i.e. the amino acid encoded when the human mitochondrial genome does not contain the non-synonymous substitution investigated), while the mutant amino acid was defined as the amino acid specific to the sequence, or the haplotype, or group of related haplotypes, in question. For the non-synonymous change under investigation, the occurrence of the wild type amino acid at the encoded amino acid position within a list of 39 species (12 primates, 22 other mammals, four non-mammalian vertebrates and one invertebrate) was counted and the conservation index was regarded as the total number of the occurrence of the wild type amino acid within the 39 species. The conservation index therefore ranged between a minimum of one and a maximum of 39.

4.14.7 Estimation of coalescent dates

The genetic divergence time from the MRCA was calculated for the L0 haplogroup, sub-haplogroups and lineages within sub-haplogroups as previously described (Forster *et al.* 1996; Saillard *et al.* 2000; Mishmar *et al.*, 2003). The global L0-specific haplogroup network (refer to Section 4.13) was used to calculate the TMRCA estimates. The number of nucleotide substitutions observed in the branches of the global L0-specific haplogroup network was counted for the L0 haplogroup, each individual sub-haplogroup, and for each lineage within a sub-haplogroup, up to the point of divergence from a common ancestor. Only nucleotide substitutions in the mtDNA coding region (bp 577 to 16,023) in genes other than the tRNA or rRNA genes were counted. The non-coding or control region (bp 16,024 to 576) was excluded due to the high probability of reverse mutations in this region (Mishmar *et al.*, 2003). Insertions, deletions, reversions, nucleotide substitutions affecting non-coding nucleotides present in the coding region, as well as possible pathogenic non-synonymous substitutions for which pathogenicity was confirmed by two independent studies, were all excluded (Kivisild *et al.*, 2006).

The average number of sites differing between a set of sequences and a specified common ancestor, ρ (Forster *et al.*, 1996), was determined by dividing the total number of substitutions observed in the complete phylogenetic network representing haplogroup L0, or a section of the network representing a specific sub-haplogroup or lineage within a sub-haplogroup, by the number of individuals that were used to construct the complete network or specific section of the phylogenetic network. The calculation used is indicated in Equation 4.13.

Equation 4.13 Calculation of the average number of sites differing between a set of sequences and a specified common ancestor

$$\rho = \text{number of substitutions} / \text{number of individuals}$$

ρ = average number of sites differing between a set of sequences and a specified common ancestor. Adapted from Morral *et al.* (1994).

The average number of sites differing between a set of sequences and a specified common ancestor was subsequently converted to time measured in years before present by using an average sequence evolution rate of $1.26 \pm 0.08 \times 10^{-8}$ nucleotide substitutions per nucleotide per year (Mishmar *et al.*, 2003). This rate was calculated using the HKY85 model (Hasegawa *et al.*, 1985). The calculation used for the conversion of ρ to time is indicated in Equation 4.14. The number of base pairs (15,447) is the number of

nucleotides contained within the coding region, as the average sequence evolution rate was calculated taking into account only the coding region.

Equation 4.14 Calculation of the TMRCA

$$\text{TMRCA} = \rho / (1.26 \times 10^{-8} \times 15,447)$$

TMRCA = time to most recent common ancestor in years before present (YBP); ρ = average number of sites differing between a set of sequences and a specified common ancestor. Adapted from Mishmar *et al.* (2003).

The relative time period during which the MRCA may have existed, was calculated by determining the standard deviation for the TMRCA. The calculation used is indicated in Equation 4.15. The calculated SD is subsequently converted to time (years before present) by dividing by $1.26 \times 10^{-8} \times 15,447$ which is the product of the substitution rate and the number of nucleotides in the coding region.

Equation 4.15 Calculation of standard deviation of the TMRCA

$$\text{SD} = \sqrt{\rho / n}$$

SD = standard deviation; ρ = average number of sites differing between a set of sequences and a specified common ancestor; n = population size. Adapted from Saillard *et al.* (2000).

4.15 CONSTRUCTION OF A REVISED CLASSIFICATION SCHEME FOR THE L0 HAPLOGROUP

A revised classification scheme for the L0 haplogroup was constructed using the classification scheme described in Section 4.9, as well as the global mtDNA mutational phylogeny available on MITOMAP (Ruiz-Pesini *et al.*, 2007) as a starting point. New branches, which were identified in the global L0-specific haplogroup network (refer to Section 4.13), were used to update the classification scheme used in the current investigation (refer to Section 4.9) by incorporating the information provided by these new branches in the revised classification scheme. A reduced version of the global maximum parsimony tree generated using dataset 1a (refer to Section 4.12.6), which displayed the branching structure present in the L0 haplogroup, was used as a guide to determine the order in which sub-haplogroups and lineages diverged from one another. The sub-haplogroups and lineages within sub-haplogroups in the revised classification scheme were named according to the system followed in the global L0-specific haplogroup network (refer to Section 4.13.2).

For the purpose of constructing a classification scheme two types of SNPs are used, namely haplogroup defining and haplogroup-specific polymorphisms. Haplogroup defining polymorphisms are alterations that occur in all mtDNA sequences within a haplogroup, although not in mtDNA sequences from other haplogroups, and are used to describe a specific haplogroup, while haplogroup-specific polymorphisms are alterations that have arisen only once during evolution, and occur in only a subset of mtDNA sequences that define a sub-cluster of the haplogroup (Herrnstadt *et al.*, 2002a). These two types of SNPs were selected for inclusion in the classification scheme by adhering to the specific criteria discussed in the next two paragraphs.

SNPs were selected for inclusion in the revised classification scheme based on their location in the mitochondrial genome. For the construction of a classification scheme, slowly evolving nucleotide positions are preferable as it is expected that these nucleotide positions will remain stable over a long period of time. Therefore, only SNPs observed at nucleotide positions located in the mitochondrial coding region (bp 577 to 16,023) were considered, while SNPs observed at nucleotide positions in the control region were excluded as the control region evolves at a different overall rate and contains fast evolving sites (Pesole *et al.*, 1999). A list of SNPs that were classified as recurrent alterations or reversions were compiled using data from the current investigation, as well as published information (Herrnstadt *et al.*, 2002a; Galtier *et al.*, 2006; Kivisild *et al.*, 2006) and these SNPs were excluded on the basis that the nucleotide positions at which these types of alterations occur are generally viewed as fast evolving. These SNPs were however not excluded if the SNPs were historically used to define macrohaplogroup L, or if no other haplogroup defining or haplogroup-specific SNP was available for classification of the sub-haplogroup or lineage within the sub-haplogroup.

SNPs were further selected for inclusion in the revised classification scheme based on their presence in all sequences assigned to a specific haplogroup, sub-haplogroup or lineage within a sub-haplogroup, the nature of the alteration, and the genes in which the SNP was located. A list of SNPs present in all sequences assigned to the specific haplogroup, sub-haplogroup or lineage within a sub-haplogroup was compiled using the global L0-specific haplogroup network, as well as the global mtDNA mutational phylogeny available on MITOMAP (Ruiz-Pesini *et al.*, 2007). Thereafter, SNPs involving synonymous substitutions in genes other than the tRNA or rRNA genes were selected in accordance with convention, i.e. substitutions in polypeptide genes were used in the construction of the classification scheme used in the current investigation (refer to Section 4.9). If more than

one SNP that defined the specific sub-haplogroup/lineage remained after this elimination process, the choice of which SNP was included in the revised classification scheme was arbitrary. If a suitable synonymous substitution was however not available for selection, a non-synonymous substitution with a high conservation index was selected, and if no non-synonymous substitutions were available for selection, a synonymous substitution in a tRNA or rRNA gene was selected.

4.16 CONSTRUCTION OF A MITOCHONDRIAL CONSENSUS SEQUENCE

A consensus sequence can be defined as a sequence that summarises the common features identified among a set of sequences (Page and Holmes, 1998). A consensus sequence is therefore comprised of the nucleotide detected in the majority of sequences at each nucleotide position (Page and Holmes, 1998).

A consensus sequence for the Southern African Khoi-San population was compiled by using the 22 complete mtDNA sequences generated in the current investigation, as well as 13 complete Khoi-San mitochondrial genome sequences generated in a previous investigation (Mouton, 2003). All the sequences used to construct the Khoi-San consensus sequence were assigned to the L0 haplogroup and belonged either to the L0a or L0b sub-haplogroups (refer to Section 4.9). Khoi-San sequences classified in other L-specific haplogroups such as L2 or L3 were excluded from the analysis.

The Khoi-San consensus sequence was constructed in two steps. The first step consisted of the construction of a consensus sequence for the L0a sequences, as well as a separate consensus sequence for the L0b sequences. The Khoi-San sequences representing each sub-haplogroup were aligned using ClustalX 1.83 (Thompson *et al.*, 1997) and a consensus sequence was compiled for each set of sequences representing a sub-haplogroup using BioEdit version 5.0.9 (Hall, 1999) where the consensus sequence for each set of sequences was determined by retaining the nucleotide most commonly observed at each position. Nucleotide positions for which more than two of the four possible nucleotides were present, and positions containing gaps or missing information within the sequence alignment, were however not assigned a consensus nucleotide by the software. These nucleotide positions were resolved manually after visual inspection of the sequence alignment.

The second step involved combining the two separate consensus sequences to compile a single consensus sequence. At positions where differences were observed between the L0a and L0b consensus sequences for which a consensus character state could not be determined, the nucleotide was resolved by defining it using an ambiguity code (R, Y, S or M). In the case of insertions and deletions, the consensus for that position was determined by using a consensus sequence constructed using all 35 L0 sequences.

4.17 CALCULATION OF PERCENTAGE SIMILARITY AND PAIRWISE NUMBER OF DIFFERENCES

The percentage similarity between 43 complete Khoi-San mitochondrial genome sequences (Mouton, 2003; current investigation) and the Khoi-San consensus sequence (refer to Section 4.16) was determined by performing sequence alignment as described in Section 4.12.1, followed by the compilation of a sequence similarity matrix using the DNADIST program in the PHYLIP version 3.66 suite of programs (Felsenstein, 1989). The values generated in the similarity matrix were reported in the form of a fraction of the number of nucleotide sites identical between the sequences, and these values were converted to a percentage for use in the current investigation. The similarity values were calculated by excluding nucleotide positions containing gaps, i.e. nucleotide positions which exhibited insertions or deletions within the sequence alignment used to generate the similarity matrix were not included in the calculation. In addition, the DNADIST program does not take into account the similarity of non-identical nucleotides, i.e. no credit is given for having different purines/pyrimidines at corresponding nucleotide positions, and the similarity values are not corrected for multiple changes.

The pairwise number of differences between the Khoi-San consensus sequence (refer to Section 4.16) and 43 complete Khoi-San mitochondrial genome sequences (Mouton, 2003; current investigation) was determined by performing sequence alignment as described in Section 4.12.1, followed by the construction of a matrix consisting of the number of pairwise nucleotide differences between sequences using MEGA version 4 (Tamura *et al.*, 2007). The number of pairwise nucleotide differences between sequences was calculated by excluding nucleotide positions containing gaps, i.e. nucleotide positions which contained insertions or deletions within the sequence alignment used to generate the pairwise differences matrix were not included in the calculation.

CHAPTER FIVE

Results and Discussion

Mitochondrial DNA sequence data has been widely used as a tool in the study of human evolution due to its unique characteristics which include a tenfold increased mutation rate compared to the nuclear genome (Brown *et al.*, 1979), a lack of recombination (Zuckerman *et al.*, 1984) and the fact that it is maternally inherited (Giles *et al.*, 1980). Previous human mtDNA studies have demonstrated that the highest levels of mtDNA diversity are observed within Africa (Cann *et al.*, 1987; Vigilant *et al.*, 1991; Chen *et al.*, 1995) which supports the “Out of Africa” theory of human evolution (refer to Section 3.1.2). This implies that the most ancient mitochondrial lineages will be observed within the African population.

It has been demonstrated that the Khoi-San population, studied in the current investigation, represents one of the most ancient human mitochondrial lineages (Chen *et al.*, 2000). The data obtained from the investigation of this Khoi-San population will therefore make an important contribution to the study of human evolution. The main aim of the current investigation was to compile a mitochondrial genome consensus sequence for the Southern African Khoi-San population. To this end, the complete mitochondrial DNA sequences of 30 Khoi-San individuals were determined. These Khoi-San mtDNA sequences were further analysed to determine the number of alterations present compared to the rCRS, as well as to determine the haplogroup and sub-haplogroup structure present within the Khoi-San population. This data was in turn used to investigate the phylogenetic relationships between the Khoi-San mitochondrial sequences and other macrohaplogroup L-specific sequences, in order to estimate the level of sequence diversity present within the population. Furthermore, the data allowed for the analysis of population size, the estimation of the ages of the Khoi-San mitochondrial lineages and determination of the effect of selection on the nucleotide diversity of the Khoi-San mitochondrial DNA sequences. It was envisaged that the Khoi-San population would contain high levels of sequence variation which would support an ancient origin of their mitochondrial lineage.

The results obtained for the Khoi-San individuals are reported and discussed in subsequent sections of this chapter. The methods used to obtain the results are described in the indicated sections of Chapter Four and any changes made to these protocols are discussed. The results obtained for the Khoi-San individuals studied in the current investigation will be summarised at the end of this chapter in Section 5.14 in order to provide a quick reference to the most important findings of the current investigation.

5.1 STUDY DESIGN

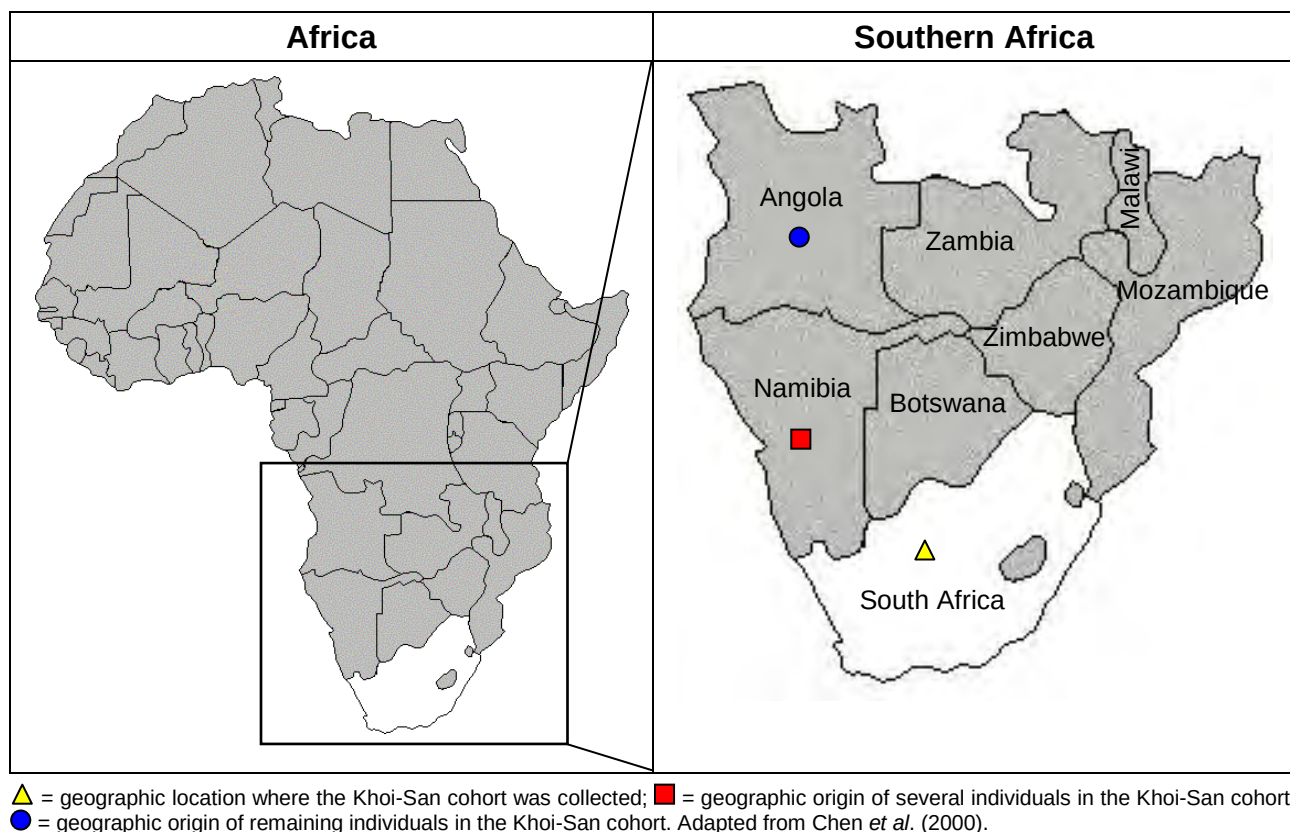
It has been suggested that the Khoi-San population harbours one of the most ancient mitochondrial lineages (Chen *et al.*, 2000). In the study of genetic relationships, through the use of phylogenetic methods, researchers are confronted with the problem of “hypothetical ancestors”, i.e. the ancestral state of a DNA sequence must be determined by studying sequence data of extant organisms, as no sequence data is available for extinct ancestral organisms. In the case of human origins and evolution, the Khoi-San, by harbouring the most ancient known mitochondrial lineage (Chen *et al.*, 2000), represents the closest approximation of the ancestral state for all living humans. This makes the Khoi-San a unique population on a genetic level and therefore the mitochondrial sequences and results generated in the current investigation are integral to future phylogenetic studies of human evolution and migration.

5.1.1 Selection of Khoi-San individuals

The Khoi-San individuals studied in the current investigation formed a subset of a larger sample previously collected in the north-western Kalahari Desert in southern Africa (Chen *et al.*, 2000). This group consisted of a total of 181 Khoi-San speaking individuals, of which 144 individuals were identified as Vasikela !Kung (Chen *et al.*, 2000). From this group of 144 individuals, 30 unrelated individuals were selected based on the concentration of the gDNA sample available for each individual. This was done due to the fact that previous investigations performed on individuals within this group (Mouton, 2003) had encountered experimental difficulties with PCR amplification and automated sequencing due to low DNA concentrations of the isolated Khoi-San gDNA. In addition, the availability of gDNA samples for Khoi-San individuals are limited, and by using the gDNA samples with the highest DNA concentration, the possibility of successful PCR and automated sequencing were increased, and the remaining gDNA samples were reserved for other investigations.

The location of the collection site and the geographic origin of the selected Khoi-San individuals are presented in Figure 5.1.

Figure 5.1 Location of the collection site and the geographic origin of the Khoi-San individuals studied in the current investigation



5.2 ISOLATION OF GENOMIC DNA

Genomic DNA for each of the Khoi-San individuals used in the current investigation was isolated from cultured lymphoblast cells obtained previously from lymphoblastoid cell lines (Mouton, 2003). The gDNA was extracted from cultured lymphoblastoid cells using the PUREGENE[®] DNA Isolation Kit (Gentra Systems, Inc.) and the samples were stored at 4°C until use.

During storage at 4°C, the samples became dehydrated and were subsequently rehydrated by the addition of 8 µL of rehydration Buffer FG3 from the FlexiGene DNA Isolation kit (QIAGEN[®]) followed by incubation overnight at room temperature or for one hour at 64°C in a water bath. The concentration and purity of the rehydrated gDNA in each of the samples was determined through O.D. analytical measurements using the NanoDrop[®] ND-1000 Spectrophotometer. The method used was based on that reported by Sambrook *et al.* in 1989, and is described in Section 4.6. The concentration of the

gDNA of the 30 selected samples ranged between $139 \text{ ng.}\mu\text{L}^{-1}$ and $802.15 \text{ ng.}\mu\text{L}^{-1}$ and the A_{260}/A_{280} ratio which indicated purity varied between 1.7 and 1.99. A $25 \text{ ng.}\mu\text{L}^{-1}$ DNA working solution was prepared for each sample and stored at 4°C , whilst the rehydrated stock solution was stored at -20°C . A working DNA solution with a concentration of $25 \text{ ng.}\mu\text{L}^{-1}$ was used for the samples in order to optimise use of these limited DNA resources.

5.3 POLYMERASE CHAIN REACTION

This method is based on the methodology first described by Mullis and colleagues (Mullis *et al.*, 1986; Mullis and Faloona, 1987). PCR was performed as discussed in Section 4.4. The success of the PCR reaction was judged according to the following criteria:

- a. The amplified PCR product was of the desired size as determined by comparison with a molecular weight marker containing an appropriate size range of fragments.
- b. The absence of amplification in the negative control, which indicated that the PCR reaction was not contaminated with foreign DNA from an unknown source.
- c. The presence of the expected PCR product in the positive control, which indicated that the PCR reaction progressed as expected.

A negative control and a positive control were included in every batch of PCR reactions performed. The complete mitochondrial DNA genomes of 30 Khoi-San individuals were subsequently amplified in eight overlapping PCR fragments, using eight overlapping primer pairs.

5.3.1 Polymerase chain reaction primers

The eight overlapping primer pairs used for PCR amplification are presented in Section 4.4.1, Table 4.1. These primer pairs were selected from a set of 32 forward and 32 reverse primers published by Maca-Meyer *et al.* in 2001, which were designed according to the rCRS. The primer pairs used were selected to amplify regions of between 2,000 and 3,000 bp in size. The primers were compared to sequence data of 13 Khoi-San mitochondrial genome sequences (Mouton, 2003) in order to exclude primers that contained nucleotide differences compared to the Khoi-San sequence data. The primer pairs were also selected to ensure a minimum overlap size of ca. 100 bp between PCR regions. This was done to ensure that when the DNA sequence of a PCR region was determined, there would be sufficient regions of overlap between the sequences

generated for each region, and for the next region, in order to assemble complete sequences for each Khoi-San individual.

5.3.2 Polymerase chain reaction optimisation

Optimisation of the different PCR reactions was required in order to achieve both specificity of amplification and a high yield of the desired PCR product (Moretti *et al.*, 1998). Through the optimisation process it was therefore attempted to satisfy the requirement for both specificity and high yield by determining the optimal balance between reaction components and cycling conditions.

The PCR conditions were as described in Section 4.4.2 and the conditions for each primer set used to amplify a specific PCR region were optimised with regard to the amount of gDNA added to the PCR reaction, the T_a , the $MgCl_2$ concentration, the amount of Super-Therm Polymerase[®] used per reaction, and the length of the PCR program. Table 5.1 contains a summary of the optimised reaction conditions for each primer pair and each PCR region. The primer sequences are listed in Table 4.1 and the PCR programs used are described in Table 4.2 (refer to Section 4.4).

Table 5.1 Primer pairs and PCR conditions used for amplification of the complete mitochondrial genome

PCR region	Primer pair	Amount of gDNA (ng)	T_a (°C)	[$MgCl_2$]	Amount of enzyme (U)	PCR program used	Product size (bp)
1	F32:mtL15996	25	60	1.5 mM	0.25	Regular program	2,103
	R3:mtH1487						
2	F3:mtL923	25	64	1.5 mM	0.5	Regular program	2,789
	R7:mtH3670						
3	F7:mtL3073	25	68	2.0 mM	0.5	Extended program	2,801
	R11:mtH5832						
3*	F8:mtL3644	25	63	1.5 mM	0.25	Regular program	2,227
	R11:mtH5832						
4	F11:mtL5278	25	62	1.5 mM	0.25	Regular program	2,679
	R15:mtH7918						
5	F15:mtL7379	25	65	2.0 mM	0.5	Extended program	3,124
	R20:mtH10462						
5*	F16:mtL7882	25	58	1.5 mM	0.25	Regular program	2,089
	R19:mtH9928						
6	F20:mtL9886	25	64	1.5 mM	0.25	Regular program	2,231
	R23:mtH12076						

Table 5.1 Continued...

PCR region	Primer pair	Amount of gDNA (ng)	T _a (°C)	[MgCl ₂]	Amount of enzyme (U)	PCR program used	Product size (bp)
7	F23:mtL11486	25	61	1.5 mM	0.5	Extended program	2,740
	R27:mtH14186						
8	F27:mtL13612	25	63	1.5 mM	0.5	Extended program	2,828
	R32:mtH16401						

T_a = optimised annealing temperature; [MgCl₂] = magnesium chloride concentration; ng = nanogram; °C = degrees Celsius; U = unit; bp = base pairs; mM = millimolar; 1, 2, 3, 4, 5, 6, 7, 8 = PCR regions; 3* and 5* = alternative PCR regions. The regular and extended PCR programs are described in Table 4.2.

Super-Therm Polymerase[®] is a thermostable DNA polymerase. It has both 5'-3' polymerase and 5'-3' exonuclease activities, however, no 3'-5' exonuclease activity is present. This means that the enzyme has no proofreading function. The error rate for this enzyme, in errors per base pair, has been calculated as 1.1×10^{-4} (Lundberg *et al.*, 1991) or 2×10^{-5} (Barnes, 1992). For a PCR product with a size of 3,000 bp, this equates to between 0.06 and 0.33 errors per product. The combination of this relatively low error rate and the high temperature under which the enzyme performs amplification, was expected to limit the inclusion of erroneous nucleotides in the amplified PCR products. If during analysis of the sequencing data, nucleotide positions were identified that were suspected to contain errors due to the PCR process, the option was available to add *Pfu* DNA polymerase to the PCR reaction as this enzyme has a proofreading function. This option was however not employed as other strategies were preferred to resolve any observed occurrences of ambiguity due to the PCR process or other errors at specific nucleotide sites. These strategies included the use of overlapping sequence data and repeating the sequencing reaction using a newly purified PCR product and a different sequencing primer. These alternative strategies are discussed in Section 5.6.

An amount of 25 ng gDNA, i.e. 1 µL of a 25 ng.µL⁻¹ working solution, was used per PCR reaction instead of 50 ng which is the standard amount added to PCR reactions performed at the CGR at the North-West University (Potchefstroom Campus). This was done in order to limit the amount of the available Khoi-San DNA samples used per PCR reaction. These samples are a very rare and limited resource and it was expected, due to the quality and generally low DNA concentration of the Khoi-San samples, that the experimental procedures, though optimised, would not always be performed successfully, necessitating repeat procedures.

Optimisation of PCR conditions was performed using four control DNA samples of Caucasian origin. Control samples of African origin, belonging to mitochondrial macrohaplogroup L, would have been better suited for the purpose of the current investigation. However, at the time of optimisation, control samples of Caucasian origin were the only available samples. To compensate for the possible effect this could have had on optimisation, DNA samples of three Khoi-San individuals previously investigated (Mouton, 2003) were used to test the final optimised conditions for each primer pair and PCR region. It was generally observed that if all four Caucasian controls generated the correct PCR product, similar results were achieved for the three Khoi-San control samples, although the amount of product was less than that observed for the Caucasian controls. This difference in yield was attributed to the quality of the gDNA. The Khoi-San gDNA samples were subjected to a period of storage at 4°C during which the samples became dehydrated (refer to Section 5.2) which possibly caused some degradation of the gDNA samples compared to the Caucasian samples. PCR amplification of the desired fragment was confirmed firstly by comparing the size of the generated fragment with the sizes of the fragments contained in the DNA molecular weight marker during agarose gel electrophoresis (refer to Section 4.5), and secondly by sequence analysis of the PCR product (refer to Section 4.7).

PCR reactions were performed in amplification sets consisting of the reactions of 10 Khoi-San individuals, one reaction representing a negative control in which the gDNA was replaced by 1 µL ddH₂O and one reaction representing a positive control in which 1 µL of one of the four control DNA samples used for optimisation was added. A regular PCR program was used for PCR regions 1, 2, 3*, 4, 5* and 6, whereas an extended PCR program was used for PCR regions 3, 5, 7 and 8 (refer to Section 4.4.2, Table 4.2).

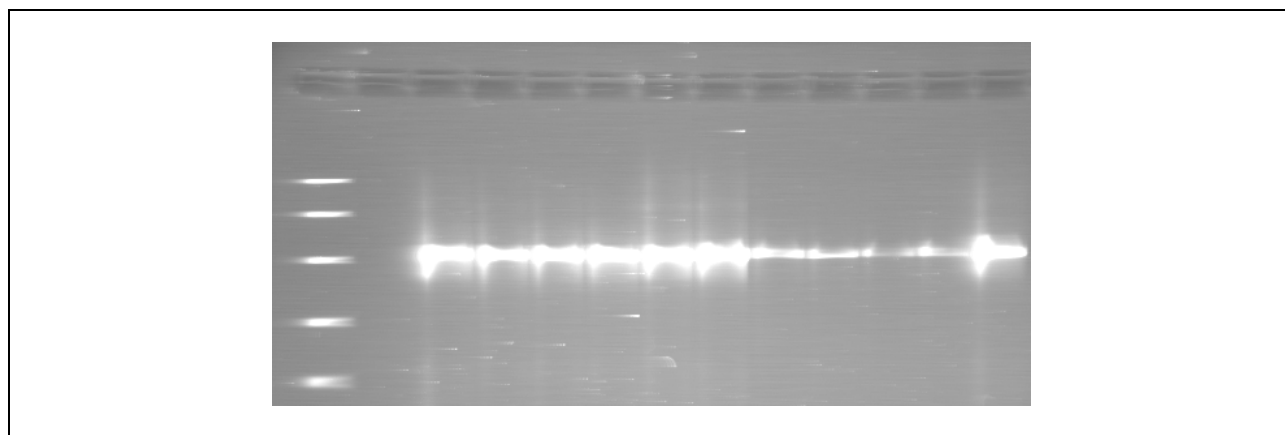
5.3.3 Artefacts observed in PCR amplified samples

A number of different artefacts were observed within the PCR reactions, even though precautions were taken to minimise or prevent the occurrence of such artefacts. The observed artefacts are described in the following sections. In all cases where PCR artefacts were observed, the artefact did not interfere with the PCR reaction or the subsequent sequencing of the purified PCR product.

5.3.3.1 Amplification efficiency

The amplification efficiency for all amplified samples within a specific amplification set was not equal. Although the same amount of gDNA template was used in all PCR reactions performed within an amplification set, it was observed that some Khoi-San samples consistently amplified better or worse than other samples. This was most likely due to some unknown property linked to the specific sample such as the quality of the isolated gDNA. An example of this variation in amplification efficiency between samples is indicated in Figure 5.2.

Figure 5.2 Photographic representation of the variation in amplification efficiency observed between amplified PCR products for different samples



Fragments were electrophoresed on a 0.9% agarose gel at 80 V for 45 min.

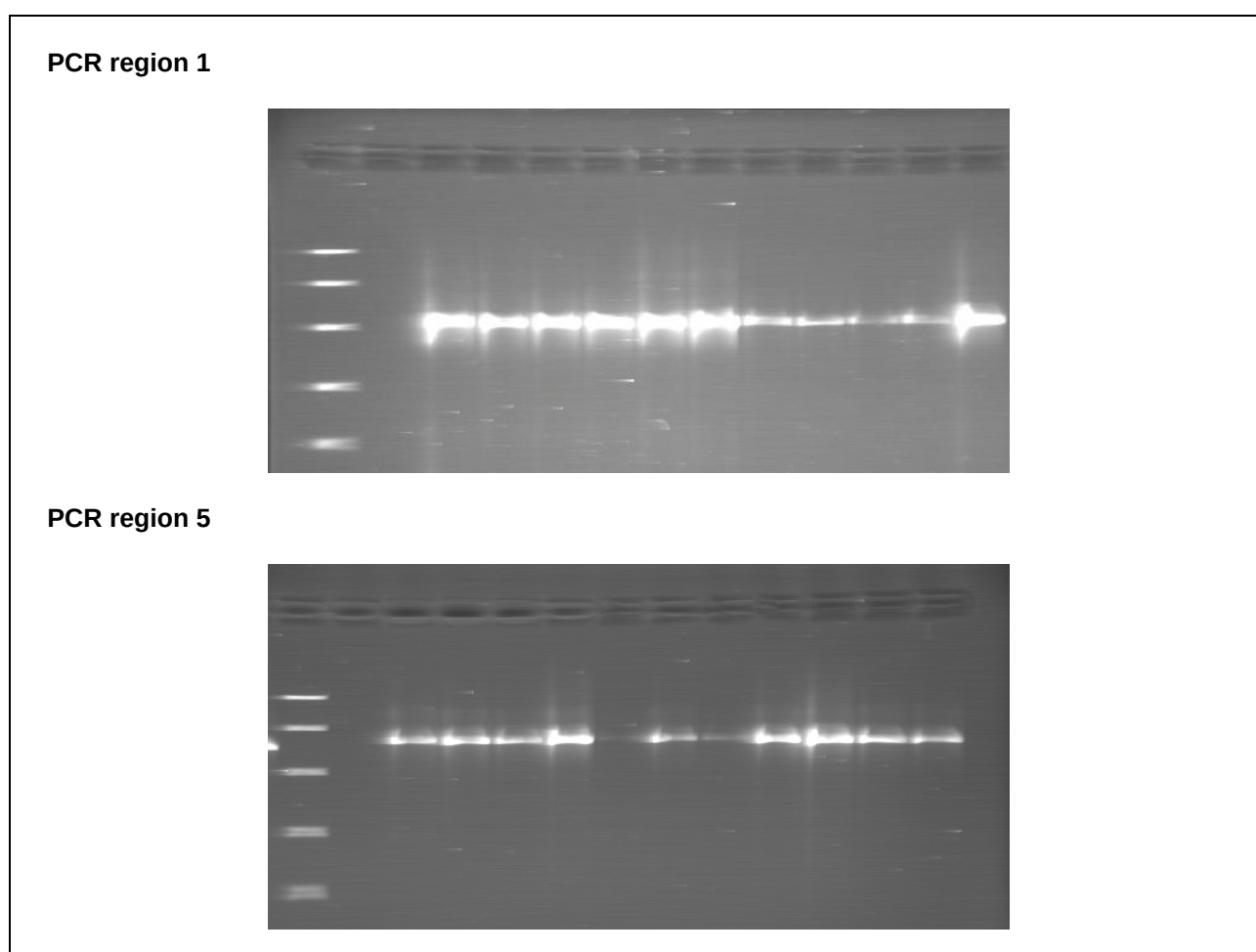
The problem of unequal amplification efficiency between samples could have been solved by repeating the genomic DNA isolation process for the samples presenting with low amplification efficiency. However, as discussed in Section 5.2, the gDNA of the samples used in the current investigation was isolated from cultured lymphoblast cells obtained from lymphoblastoid cell lines. A repeat gDNA extraction was therefore not logistically feasible at the time of analysis and an alternative strategy was employed for the specific samples.

In the cases wherein the amount of PCR product generated for a Khoi-San sample was low, the PCR reaction was repeated as described in Section 4.4 by replacing the 1 μ L volume of gDNA with 4 μ L of the first PCR reaction before purification of the PCR product. Sequences generated using these repeat PCR products were carefully analysed to identify sequence alterations that could be ascribed to the repeat PCR process. No such alterations were detected when comparing the sequence from a repeat PCR product to other sequences of the same region generated from first PCR products. In addition, in an

attempt to improve the yield of PCR reactions for these Khoi-San samples, the amount of starting gDNA template was increased from 25 ng to 50 ng. This however did not improve the yield for the PCR reactions in question and the repeat PCR process described above was instead used where necessary.

The amplification efficiency varied between samples and between PCR regions. An example of this variation in amplification efficiency between PCR regions is presented in Figure 5.3.

Figure 5.3 Photographic representation of the variation in amplification efficiency observed between different amplified PCR regions



Fragments were electrophoresed on a 0.9% agarose gel at 80 V for 45 min.

It was observed that amplification of certain PCR regions (region 3, 5 and 8) were less successful for all Khoi-San samples than for other regions. These three regions were the regions that were the most difficult to optimise and required the extended PCR program. As these regions amplified well when using the Caucasian control samples, the difficulties experienced with the Khoi-San samples could be ascribed to characteristics specific to the Khoi-San samples such as the quality of the isolated gDNA. It is however more likely that

the low amplification efficiency was caused by properties specific to these PCR regions, such as the size of the PCR product generated and the primer set used for amplification in the Khoi-San cohort.

The repeat PCR process described in Section 5.3.3.1 was again employed to increase the amplification efficiency in PCR regions 3, 5 and 8. However, for a few samples, the yield for PCR region 3 and 5 was still insufficient for sequencing purposes. In these cases, an alternative PCR reaction was performed for both PCR region 3 and PCR region 5. The alternative PCR reaction conditions for these two regions are indicated in Table 5.1 and the alternative PCR regions are indicated by an asterisk following the number of the PCR region. The alternative PCR regions are both smaller in size than the original PCR regions, and this difference in size could possibly explain the increased amplification efficiency and PCR product yield obtained for the samples involved when performing the alternative PCR reactions. The reduction in size of the PCR regions could increase amplification efficiency by excluding a section of the original PCR region which contained a region of sequence that interfered with amplification by forming secondary structures. Alternatively, the primer set used to amplify the original PCR regions possibly did not anneal correctly due to differences in the primer sequence and the target sequence, or the primers in the set interacted with each other to form primer-dimers, or the primers themselves formed secondary structures such as hair-pin loops which prevented proper annealing. The use of an alternative primer set that generated a smaller size PCR region could subsequently have overcome such sequence related obstacles.

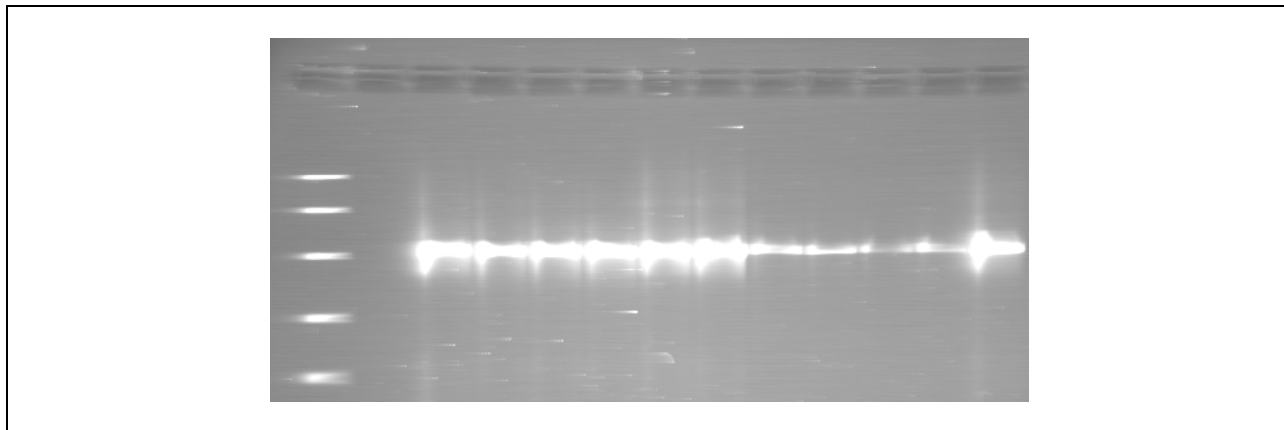
The observed variation in amplification efficiency between samples and between PCR regions however did not affect the outcome of the subsequent sequencing reactions. This is due to the fact that after PCR purification, the same amount of PCR product was used as template for each sequencing reaction of a specific PCR region.

5.3.3.2 Background smear

A background smear was observed for certain amplified samples. This smear pattern is formed by the migration of differently sized DNA fragments through the porous agarose gel matrix to different positions within the gel (Sambrook and Russell, 2001). Background smears may be present in all samples, it may however not always be observed to the same degree due to the resolving power of the detection system and differences in sample concentration. Agarose gels have a lower resolving power than polyacrylamide gels, which

can separate DNA fragments that differ in size by as little as 1 bp in length (Sambrook and Russell, 2001). An example of samples with background smears is indicated in Figure 5.4.

Figure 5.4 Photographic representation of the background smear observed for amplified PCR products



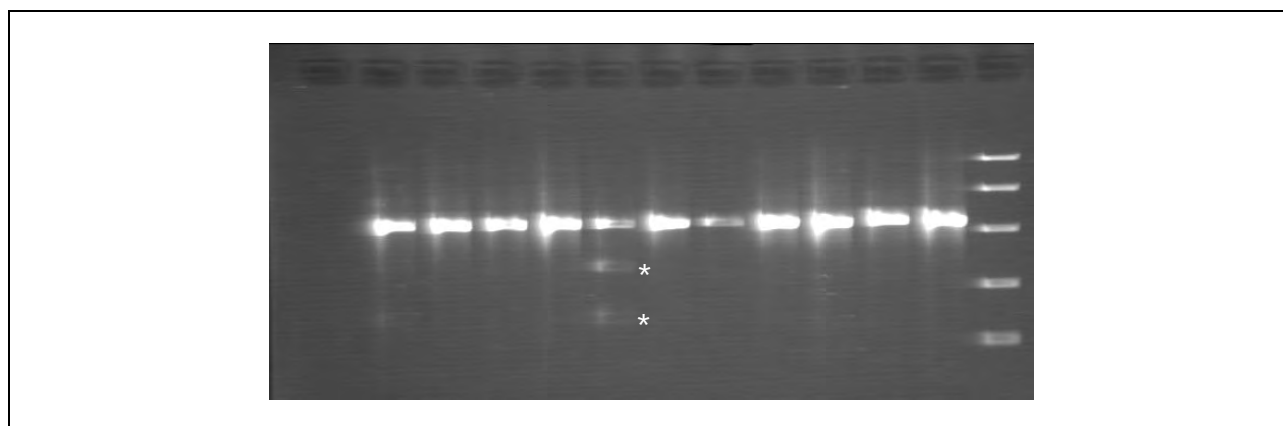
Fragments were electrophoresed on a 0.9% agarose gel at 80 V for 45 min.

The background smears observed for the amplified samples consist of two parts, fragments bigger than the PCR product and fragments smaller than the PCR product. The part of the background smear above the PCR product for certain amplified samples could be due to the presence of fragmented gDNA within the PCR reaction, or overloading of the agarose gel. Support for overloading of the gel as a possible cause for the presence of background smears is provided by the fact that a correlation exists between the amount of PCR product visible on the agarose gel and the intensity of the background smear. The part of the background smear below the PCR product for certain amplified samples could be caused by the presence of unspecific or incomplete PCR amplification products of different sizes. As the PCR reactions for these samples were optimised, it is likely that the observed smears are due to a combination of the overloading of samples and the incomplete extension of PCR products.

5.3.3.3 Secondary amplification

Secondary amplification was observed in some samples during PCR amplification. An example of a sample exhibiting secondary amplification is illustrated in Figure 5.5.

Figure 5.5 Photographic representation of secondary amplification observed for amplified PCR products



Fragments were electrophoresed on a 0.9% agarose gel at 80 V for 45 min. The white asterisks (*) indicate the artefacts observed due to secondary amplification.

Secondary amplification is usually the consequence of primers annealing at positions other than the intended annealing sites and occurs when reaction conditions are not optimal. It is however unlikely that the secondary amplification observed was due to inadequate optimisation of PCR conditions as only certain samples within an amplification set were affected by this artefact. It is possible that in some samples a nucleotide alteration (substitution, deletion or insertion) created a secondary annealing site for either the forward or reverse primer which resulted in the presence of secondary amplification.

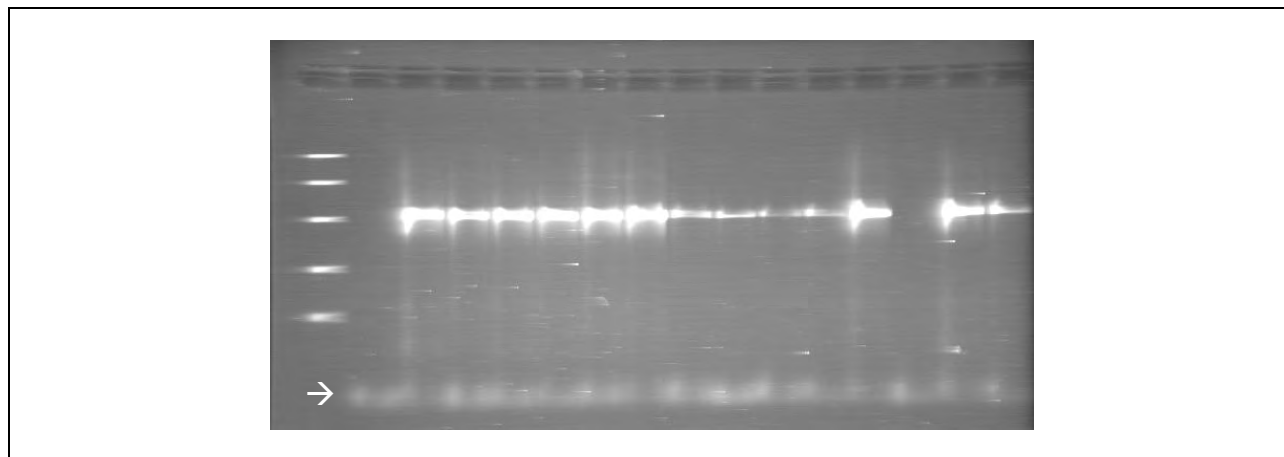
Secondary amplification could be prevented by increasing the specificity of the PCR reaction by re-designing the primers, or replacing the primers with other available primers from the set of 32 forward and 32 reverse primers (Maca-Meyer *et al.*, 2001), or by increasing the annealing temperature which would in turn decrease the already low yield of the PCR reaction for some samples. However, in all cases where secondary amplification was observed, the yield of the correctly amplified PCR product exceeded that of the secondary amplified product, and the presence of the secondary amplification product did not interfere with subsequent sequencing reactions.

5.3.3.4 Primer-dimers

For specific PCR regions and certain samples, primer-dimers were observed. The occurrence of primer-dimers is dependent on the efficiency of the PCR for each PCR region or sample amplified. Primer-dimers are formed when one primer is extended by the DNA polymerase using the other primer or itself as a template resulting in short undesired products (Newton and Graham, 1997). Within a specific PCR region this process could be

due to complementarity at the 3' end of the primer, within a single primer or between both primers in the primer set. An example of primer-dimers is indicated in Figure 5.6.

Figure 5.6 Photographic representation of primer-dimers observed for amplified PCR products



Fragments were electrophoresed on a 0.9% agarose gel at 80 V for 45 min. The white arrow (→) indicates the artefact observed due to primer-dimer formation.

Primer-dimer formation can be minimised by increasing the specificity and efficiency of PCR through the use of a hot start PCR protocol (Chou *et al.*, 1992) or a touchdown PCR protocol (Don *et al.*, 1991). In addition, primer-dimer formation can also be limited through the use of a different DNA polymerase such as AmpliTaq[®] Gold¹ (Brownie *et al.*, 1997). However, in all the cases where primer-dimers were observed in the current investigation, they did not interfere with PCR amplification and the subsequent sequencing of the purified PCR products.

5.4 AGAROSE GEL ELECTROPHORESIS

Agarose gel electrophoresis is a simple and rapid method that can separate and resolve DNA fragments of 50 to 20,000 bp in size (Sambrook and Russell, 2001). Agarose gel electrophoresis was performed as discussed in Section 4.5 and was used after PCR amplification to confirm and evaluate the success of the PCR reactions, according to the criteria listed in Section 5.3, prior to the purification of PCR products and subsequent sequencing of these products.

A 0.9% (w/v) agarose gel was used for all amplified PCR regions in order to visualise the separation of DNA fragments that ranged in size from 500 to 7,000 bp (Barker, 2005). The molecular weight of each PCR product was estimated by comparing its mobility to that of

¹ AmpliTaq Gold is a registered trademark of Roche Molecular Systems, Inc., Pleasanton, CA, USA.

the High Range FastRuler™ DNA molecular weight marker. Electronic copies of the gel image were captured via a photo documentation system which resulted in an unavoidable loss of resolution in the depicted gel pictures compared to the resolution of the original gel as viewed on the UV transilluminator.

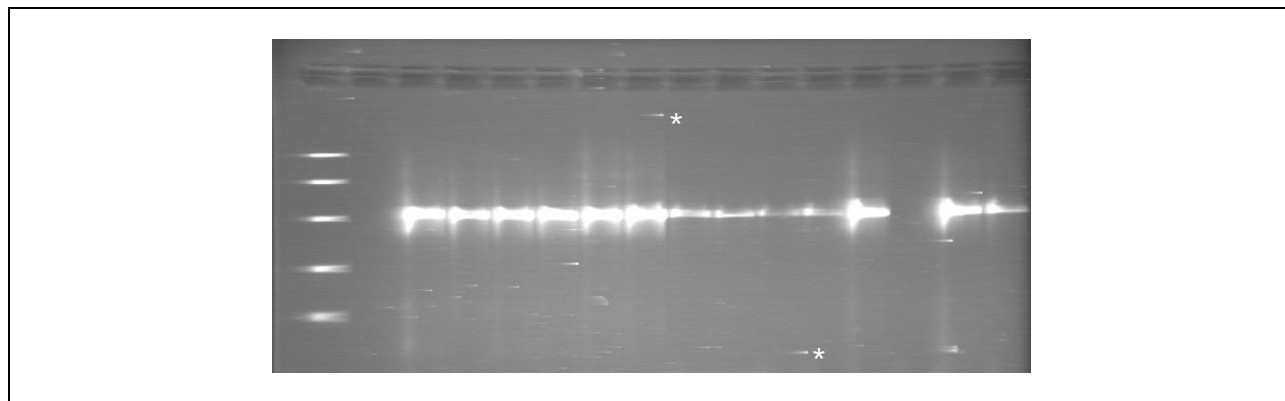
5.4.1 Artefacts observed on agarose gels

Even though precautions were taken to prevent the occurrence of artefacts, a number of different artefacts were observed within the agarose gels. The observed artefacts are described and discussed in the following sections. In all cases where agarose gel artefacts were observed, the artefact did not interfere with the confirmation of the success of the PCR reaction, or the subsequent sequencing of the purified PCR product.

5.4.1.1 Artefacts in the gel matrix

Artefacts in the gel matrix were observed throughout certain agarose gels. An example of an agarose gel containing gel matrix artefacts is indicated in Figure 5.7.

Figure 5.7 Photographic representation of an artefact observed in the gel matrix



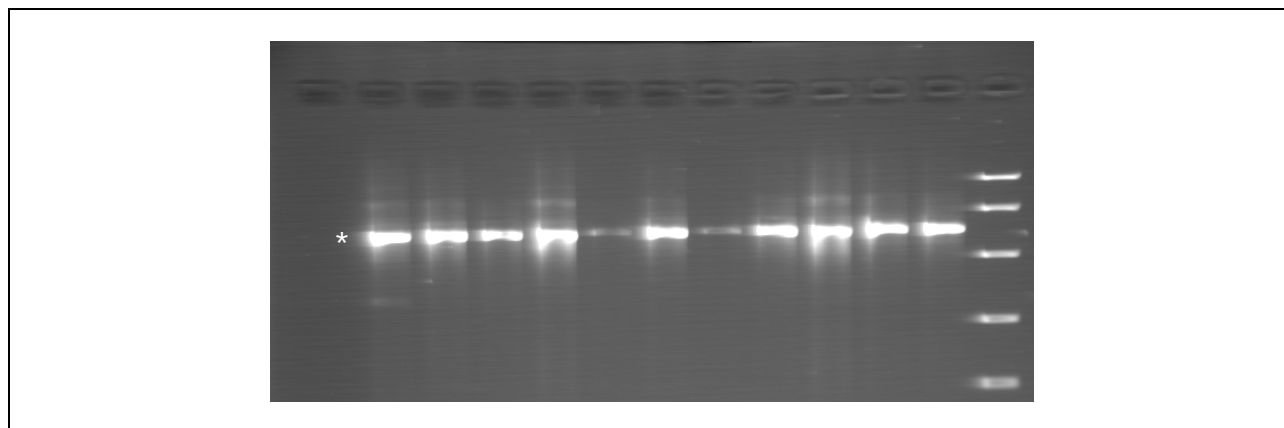
Fragments were electrophoresed on a 0.9% agarose gel at 80 V for 45 min. The white asterisks (*) indicate examples of the artefacts observed in the gel matrix.

This artefact in the gel matrix is probably due to the presence of lint. Paper towels were used to clean the electrophoresis trays and the UV transilluminator before and after use, leaving lint on these surfaces which would only become visible once the gel was viewed using a UV light source. A lint-free paper towel should be used to clean these surfaces in order to prevent the presence of external contaminants.

5.4.1.2 Distortion of sample fragments

Distortion of the amplified PCR fragments was observed for some samples. An example of the distortion of sample fragments is illustrated in Figure 5.8.

Figure 5.8 Photographic representation of the distortion of amplified fragments



Fragments were electrophoresed on a 0.9% agarose gel at 80 V for 45 min. The white asterisk (*) indicates an example of distortion of the sample fragment.

The sample fragments appear broader on the sides of the fragments, with the left side generally appearing broader than the right. This distortion of the fragments can be due to a number of factors. The factors and possible solutions are indicated in Table 5.2.

Table 5.2 The likely causes for the distortion of sample fragments and ways in which it can be prevented

Likely cause of distortion	Possible solution to prevent distortion
a. During casting of the agarose gel, the gel comb was not level.	i. Ensure that the gel comb is level during gel casting by using a small circular “bulls-eye” level to check the surface on which gels are being cast.
b. The wells contained residual agarose particles which remained in the wells after removal of the gel comb.	ii. Before loading the well with sample, rinse the wells with the appropriate electrophoresis buffer using a disposable pipette.
c. The wells contained air bubbles which were introduced into the wells during the removal of the gel comb, or when loading the wells with sample.	
d. Too much sample was loaded into each well (overloading).	iii. Load less of each sample into each well.
e. The way in which the wells were loaded with sample, i.e. the position of the pipette tip in the well during loading of the wells, may have led to an unequal distribution of sample in the well with the bulk of the sample concentrated in the sides of the wells.	iv. Position the pipette tip above the centre of the well during sample loading and load the sample slowly allowing it to spread evenly within the well.

Adapted from the HORIZON^{®1} 58 Horizontal Gel Electrophoresis Apparatus Instruction Manual (Life Technologies[®], 2000) and Sambrook and Russell (2001).

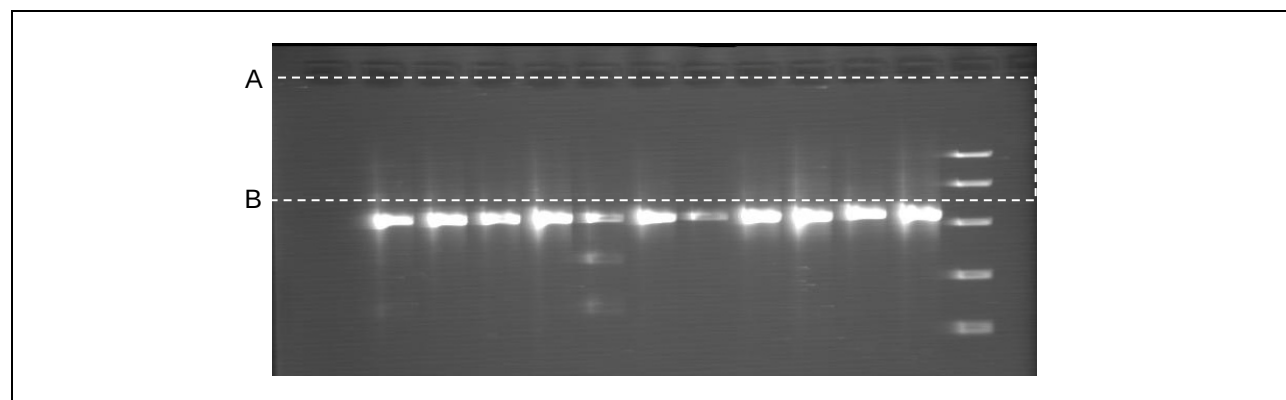
¹ HORIZON[®] is a registered trademark of Life Technologies, Inc., Gaithersburg, MD, USA.

The most likely explanation for the distortion of the fragments in the current investigation is the way in which the wells were loaded with sample. In order to keep the pipette steady and to thereby avoid the loss of the sample during loading, the pipette was held at an angle with the pipette tip pointing towards the left side of the well. If the pipette was held so that the pipette tip was positioned above the centre of the well before loading of the sample, and the sample was loaded slowly to allow it to spread evenly within the well, and to settle in the well prior to electrophoresis, the fragments appeared less distorted. The other likely causes for the observed distortion of the fragments (refer to Table 5.2) were countered prior to loading the wells with sample by using the solutions presented. Gels were cast on a level surface in a designated gel preparation area, the wells were rinsed with electrophoresis buffer prior to sample loading, and the same amount, i.e. 2.5 μL of the PCR reaction, was routinely loaded for every sample.

5.4.1.3 Slanted fragments

For some agarose gels, the usually straight line formed by the similarly sized amplified fragments when migrating the same distance from the top of the gel appears slanted, i.e. the fragments in the different lanes are not parallel to the wells or to the top edge of the gel. An example of slanted fragments is presented in Figure 5.9, with line A and line B being parallel to the wells.

Figure 5.9 Photographic representation of slanted fragments



Fragments were electrophoresed on a 0.9% agarose gel at 80 V for 45 min. Lines A and B are both parallel to the wells of the agarose gel. The slanted fragments can be observed below line B. The row of fragments is not parallel to line B, thereby indicating their degree of slant.

Slanted fragments are observed when there is a barrier that prevents the migration of a fragment in one or more lanes so that similar sized fragments do not migrate the same distance in the same time, and may be the consequence of a number of factors. The

possible causes and ways in which to prevent slanted fragments are indicated in Table 5.3.

Table 5.3 The possible causes of slanted sample fragments and ways in which it can be prevented

Possible cause of slanting	Possible solution to prevent slanting
a. The agarose gel was not cast or electrophoresed on a level surface.	i. Ensure that the gel tray is level during gel casting and that the gel box containing the gel tray is level during electrophoresis. Small circular “bulls-eye” levels are available to check the surface on which gels are being prepared or electrophoresed, and are often supplied with gel casting apparatus.
b. Residual agarose particles remained in the wells after removal of the gel comb.	ii. Before loading the well with sample, rinse the wells with the appropriate electrophoresis buffer using a disposable pipette.
c. The wells contained air bubbles which were introduced into the wells during the removal of the gel comb, or when loading the wells with sample.	
d. The agarose solution used to prepare the gel contained incompletely dissolved agarose particles.	iii. Verify that the agarose is completely dissolved before gel casting to form a homogenous solution without residual agarose particles.
e. The agarose solution used to prepare the gel contained particles from another source.	iv. Verify that the agarose solution is free from any particulate matter prior to casting the gel.

Adapted from the HORIZON® 58 Horizontal Gel Electrophoresis Apparatus Instruction Manual (Life Technologies®, 2000) and Sambrook and Russell (2001).

5.5 PCR PRODUCT PURIFICATION

PCR product purification was performed as described in Section 4.4.3. Two different kits were employed for purification, the QIAquick® PCR Purification Kit (QIAGEN®) and the DNA Clean & Concentrator-5™ kit (Zymo Research Corporation). The concentration and purity of the purified PCR products were determined as described in Section 4.6. The DNA concentrations of the successfully sequenced PCR products after use of the QIAquick® PCR Purification Kit (QIAGEN®) ranged from 6.50 ng.μL⁻¹ to 52.10 ng.μL⁻¹, while the A₂₆₀/A₂₈₀ ratios were between 1.14 and 2.16. The DNA concentrations of the successfully sequenced PCR products after use of the DNA Clean & Concentrator-5™ kit (Zymo Research Corporation) varied between 5.90 ng.μL⁻¹ and 47.90 ng.μL⁻¹, while the A₂₆₀/A₂₈₀ ratios ranged from 0.91 to 2.66. For the samples amplified in the current investigation, the DNA Clean & Concentrator-5™ kit provided, on average, slightly better yields (20.82 ng.μL⁻¹ vs. 17.02 ng.μL⁻¹) and ca. the same purity (1.65 vs. 1.66) as the QIAquick® PCR Purification Kit. The DNA Clean & Concentrator-5™ kit was however easier to use and the more economical choice.

5.6 AUTOMATED DNA SEQUENCING

All successfully amplified samples were subjected to automated sequencing according to the dideoxynucleotide chain-termination method (Sanger *et al.*, 1977). The ABI PRISM[®] BigDye[™] Terminator version 3.1 Ready Reaction Cycle Sequencing Kit (Applied Biosystems[®]) was used and cycle sequencing reactions were performed as described in Section 4.7.1. After cycle sequencing, the DNA extension products were precipitated as outlined in Section 4.7.3 in order to remove the unincorporated BigDye[™] terminators remaining in the sequencing reaction and to prevent interference with the signal produced during electrophoresis.

The success of the sequencing reaction was judged according to the following criteria:

- a. The presence of well-defined peaks with high amplitude.
- b. Even spacing of peaks.
- c. Minimal or no background noise which can be defined as the presence of background peaks or peaks under peaks.

The quality of sequence data was dependent on the purity and the amount of the PCR product template used, as well as the specific primer used. A number of problems were experienced with the sequencing of the samples (refer to Section 5.6.2) and it was necessary to repeat sequencing for certain samples. In cases where the level of background noise was too high to allow for the accurate reading of the electropherogram, the sequencing was repeated. Sequencing was also repeated when no overlap was obtained with sequences from previous or subsequent regions, or when different nucleotides were observed at the same nucleotide position in different sequences of the same sample. The observed ambiguity could be due to the presence of heteroplasmy, it was however more likely to be due to errors introduced by the sequencing process or PCR amplification. In these cases, the PCR amplification and sequencing were repeated for the PCR region/sequence affected by the observed ambiguity.

5.6.1 Primer design

The DNA sequence of the complete mitochondrial genome of each Khoi-San individual was determined using 32 forward primers (refer to Section 4.7, Table 4.5). Most of the primers used formed part of the set of 32 forward and 32 reverse primers published by Maca-Meyer *et al.* in 2001, however where necessary, new primers were designed. The

use of only forward primers, i.e. primers that anneal to the L-strand, to derive the mitochondrial DNA sequences implied that only the L-strand was sequenced. This was not ideal, although it was necessitated by both financial and time constraints. In the event that ambiguity was encountered at a specific nucleotide site, it was resolved by using overlapping sequence data for that region generated by another sequencing primer, or if overlapping data was not available, the sequence was repeated in the reverse direction using one of the primers indicated in Table 4.5 after repeating the PCR amplification and PCR purification of the relevant PCR region.

All the primers used were compared to the mitochondrial genome sequences of 13 Khoi-San individuals determined by Mouton (2003) in order to identify primers that contained nucleotide changes compared to the Khoi-San sequence data. A total of five primers were identified that contained nucleotide changes compared to the Khoi-San sequence data. It was however decided to use the primers containing nucleotide changes without modification as the nucleotide changes observed did not involve regions of the primer which could influence annealing to the PCR product thereby preventing sequencing reactions from proceeding as expected.

5.6.2 DNA cycle sequencing optimisation

The standard protocol provided for cycle sequencing was followed with modifications where necessary. For all PCR regions, an amount of 3.2 pmol primer was used. As discussed in the previous section, both forward and reverse primers were used as appropriate.

The amount of PCR product used for different reactions varied according to the PCR region sequenced as well as the primer used. The PCR products varied in length from 2,103 to 3,124 bp. The recommended amounts of sequencing template for PCR products in this size range are presented in Table 4.3 and vary between 40 ng and 100 ng. The amount of PCR product template used per region is indicated in Table 5.4. These amounts were based on both the recommended amounts of sequencing template, as well as experimental observations. Where necessary, as indicated by the sequencing results obtained, the amounts of sequencing template used were adjusted per region and per primer within a region.

Table 5.4 **The amount of purified PCR product template used for sequencing per region**

PCR region	Size of PCR product (bp)	Amount of PCR product to be used for sequencing (ng)
1	2,103	30 – 40
2	2,789	40 – 50
3	2,801	50
3*	2,227	40
4	2,679	40
5	3,124	60
5*	2,089	50
6	2,231	40
7	2,740	40
8	2,828	50

bp = base pairs; ng = nanogram; 1, 2, 3, 4, 5, 6, 7, 8 = PCR regions; 3* and 5* = alternative PCR regions.

It was observed that sequences were generated more easily for certain PCR regions and sequencing primers. This phenomenon seemed to be PCR region- and primer-specific. Thus it was not related to the size of the PCR region, i.e. the smallest regions were not the easiest to sequence. Difficulties in sequencing with specific primers located in particular PCR regions could be explained by the fact that some of the primers contained alterations compared to the Khoi-San sequence data (refer to Section 5.6.1). As the primers were based on the rCRS which falls in haplogroup H and is of European origin (Andrews *et al.*, 1999), and not macrohaplogroup L, which is of African origin, it was expected that some primers would be better suited to the task than others. However, primers containing no alterations, which were located in the same PCR region, also experienced sequencing difficulties and this led to the hypothesis that the sequencing problems experienced were more likely due to PCR region-specific factors than primer-specific factors.

5.6.3 Precipitation of extension products

Precipitation of extension products was performed prior to electrophoresis in order to remove unincorporated BigDye™ terminators remaining in the cycle sequencing reaction. Two different methods were employed for this purpose, i.e. the ethanol/sodium acetate precipitation method and the Centri•Sep™ 96 well clean-up kit (refer to Section 4.7.3). The different methods were used because two different commercial entities were contracted to perform the electrophoresis of the cycle sequencing reactions. The first entity requested the sequencing reactions to be precipitated using the ethanol/sodium acetate precipitation method, while the second entity performed precipitation of the extension products through the use of the Centri•Sep™ 96 well clean-up kit as part of the sequencing service

provided. After precipitation of the sequencing reactions, the products were re-suspended in deionised formamide. The products were electrophoresed on either a SpectruMedix™ (SCE2410) Genetic Analysis System Sequencer or on an Applied Biosystems® 3130x/ Genetic Analyzer. Sequence data obtained after electrophoresis was analysed as discussed in Section 4.7.5 by using BioEdit Sequence Alignment Editor (Hall, 1999).

5.6.4 Artefacts observed on electropherograms

Many of the sequences obtained for the Khoi-San individuals were not of a satisfactory quality and a number of different artefacts were observed on the electropherograms obtained after sequencing. These artefacts can be introduced during the different stages of the sequencing process, firstly, during template preparation, secondly, during the cycle sequencing reaction and thirdly, during the precipitation of the extension products. The observed artefacts are described in the following sections. Where possible, the occurrence of artefacts was addressed by optimisation of the PCR product purification protocol, optimisation of the cycle sequencing reactions (amount of PCR product used per reaction), and optimisation of the sequence precipitation protocol.

In spite of the presence of the observed artefacts, the use of the sequence data was not prohibited. Where necessary, regions affected by artefacts such as excess dye peaks or dye blobs, were excluded from the analysis, and data for the affected region of the sequence was obtained from another source. Data for regions affected by artefacts was obtained from analysis of another sequence spanning the same region, obtained after sequencing with a forward or reverse primer, or from an overlapping region provided by analysis of another sequence.

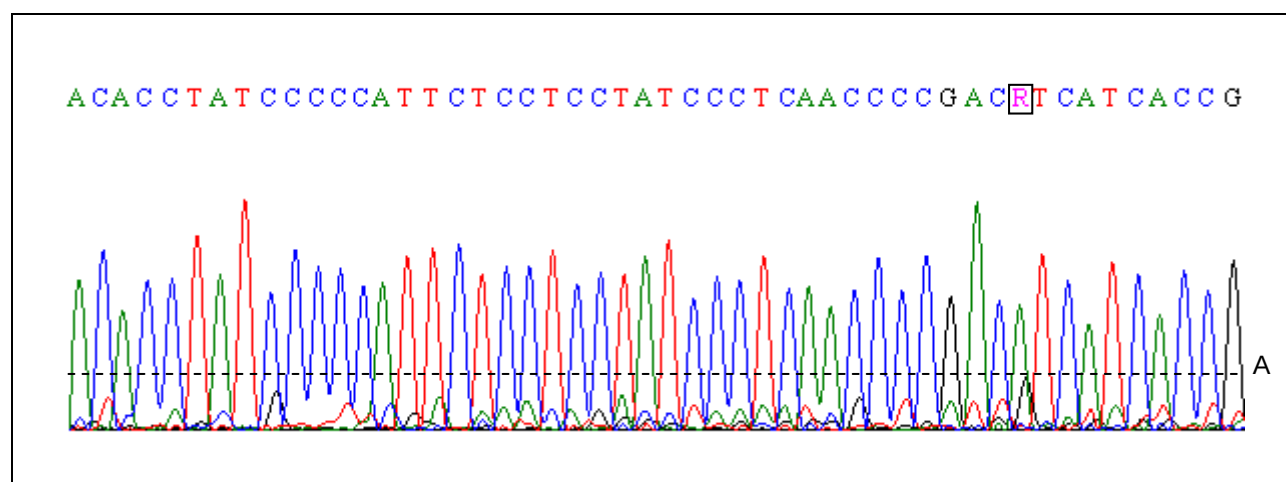
In order to permit direct comparisons between the different electropherograms representing different mitochondrial DNA sequences, the distance between the baseline and the ACGT text line was kept constant, i.e. the amplitude of the peaks in the electropherogram was not altered. This practice was applied to all of the electropherograms in all the figures in which sequence data is presented.

5.6.4.1 Background noise

Background noise is defined as secondary peaks, usually located under the peaks of interest, although background peaks can in some instances exceed the amplitude of the

true peaks. A varying amount of background noise was present throughout all the sequences analysed in the current investigation. It did not interfere with the analysis if the signal-to-noise ratio was high, i.e. the signal strength of the peaks of interest exceeded that of the background peaks, and therefore did not influence the base-calling of the sequence. An example of background noise is depicted in Figure 5.10.

Figure 5.10 Representative electropherogram indicating background noise



A = adenine; C = cytosine; G = guanine; T = thymine; R = adenine or guanine. The dashed line A indicates the maximum amplitude of the background peaks. The block (□) indicates an example of ambiguous base-calling due to background noise.

In this example, the background noise at the position indicated by a block is high enough to cause ambiguity in the automated base-calling process. The base-calling software which assigns a value to each peak of the electropherogram is generally effective. Manual editing is however required to improve the accuracy of base-calling, particularly towards the end of readable sequence where single guanine residues may have low signal strength and be missed, and in the case of homopolymer stretches where miscounting of the number of residues in the stretch can occur (Alphey, 1997). This ambiguity can therefore be resolved by visual inspection of the peak and manual editing of the base-call. In order to ensure the accuracy of base-calling in each individual sequence, all sequence data in the current investigation was evaluated via visual inspection of the electropherogram and where appropriate the base-call was altered manually.

Numerous factors may lead to the presence of background noise, including the amount of template used, the purity of the template, the presence of multiple templates and the presence of multiple primer binding sites. In the study presented here, it is most likely that the background noise was caused by the quality and the quantity of the PCR template used for sequencing (refer to Sections 5.5 and 5.6.2). The quality of the template is affected by residual salts or organic chemicals carried over after template preparation and

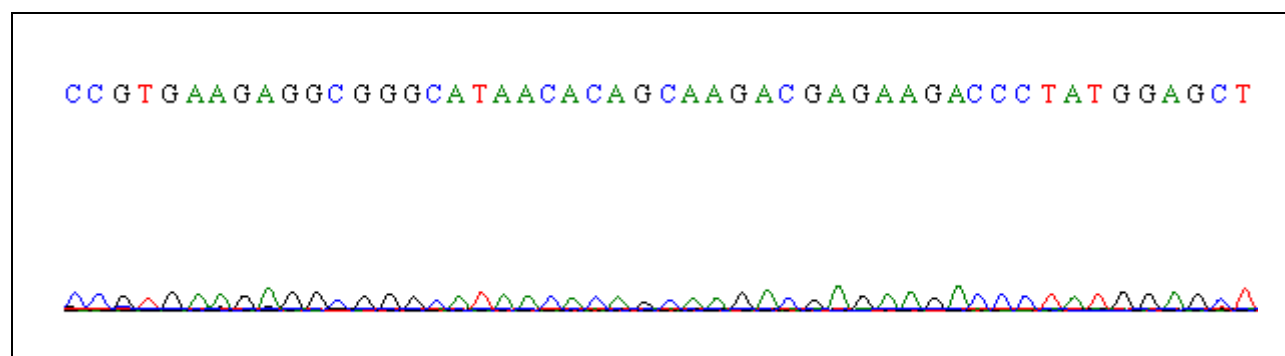
purification, degradation of the template during storage (e.g. repeated freeze-thaw cycles) and the presence of multiple templates.

The factors affecting the quality of the template used for cycle sequencing reactions were addressed by taking the following preventative measures. The PCR for each region was optimised prior to automated sequencing in order to prevent the presence of multiple templates. In cases where secondary amplification products were present, the PCR template was only used if the concentration of the correct fragment exceeded that of the secondary fragment (refer to Section 5.3.3.3). The PCR products were purified using one of two different kits in order to remove unincorporated dNTPs, excess primer, remaining thermostable enzyme, mineral oil and salts. After purification, the concentration of the PCR product was determined through optical density analytical measurements. The determined A_{260}/A_{280} ratios were used to select, where possible, purified PCR products with ratios of ca. 1.75 for cycle sequencing reactions, although a wide range of ratios were observed for successfully sequenced PCR products (refer to Section 5.5). The purified PCR products were stored at 4°C until use, in order to avoid repeated freeze-thaw cycles. It is therefore likely that the background noise observed throughout all the sequences analysed in the current investigation was a consequence of the quantity of PCR template used for sequencing. Relatively low concentrations were determined for the purified PCR products (refer to Section 5.5), and it is possible that the spectrophotometer readings were not accurate, as reliable readings are not always obtained when determining the concentration of small amounts of DNA. It can thus be speculated that in the majority of cycle sequencing reactions, too little DNA template was added as the concentration of the purified PCR product was overestimated.

5.6.4.2 Low signal strength

Low signal strength is characterised by low amplitude of the peaks. An example of low signal strength is presented in Figure 5.11.

Low signal strength is most commonly caused by the addition of insufficient template to the sequencing reaction (Applied Biosystems®, 2000). Other possible causes include the presence of an inhibitor within the sequencing reaction, insufficient primer, the loss of extension products during the precipitation process and incomplete re-suspension of precipitated extension products.

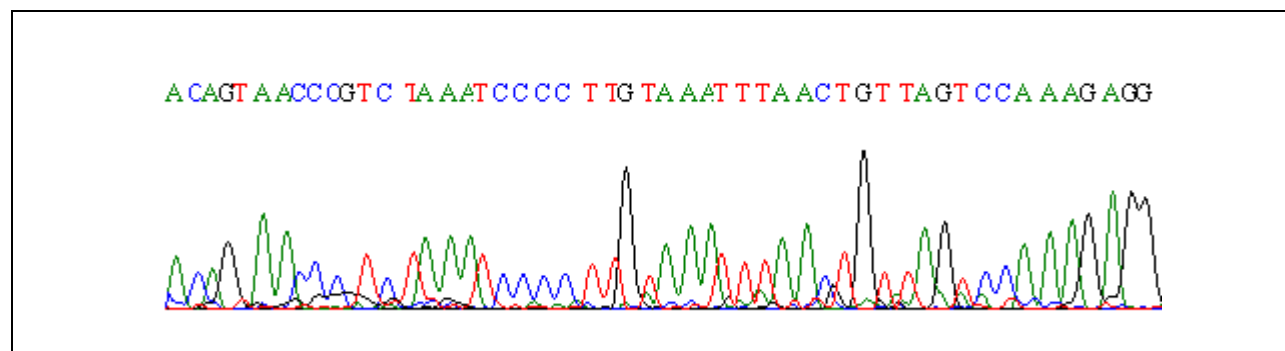
Figure 5.11 Representative electropherogram depicting low signal strength

A = adenine; C = cytosine; G = guanine; T = thymine.

In the current investigation, low signal strength was usually observed in the sequences of the samples that had the lowest PCR product concentration after PCR product purification. The problem was corrected by increasing the amount of template added to the sequencing reaction, or by repeating the PCR reaction to obtain purified PCR products with a higher concentration.

5.6.4.3 Poor mobility correction

Poor mobility correction is characterised by an uneven distribution of the peaks, i.e. some peaks are close together, even overlapping, while other peaks have large gaps between them. An example of poor mobility correction is presented in Figure 5.12.

Figure 5.12 Representative electropherogram illustrating poor mobility correction

A = adenine; C = cytosine; G = guanine; T = thymine.

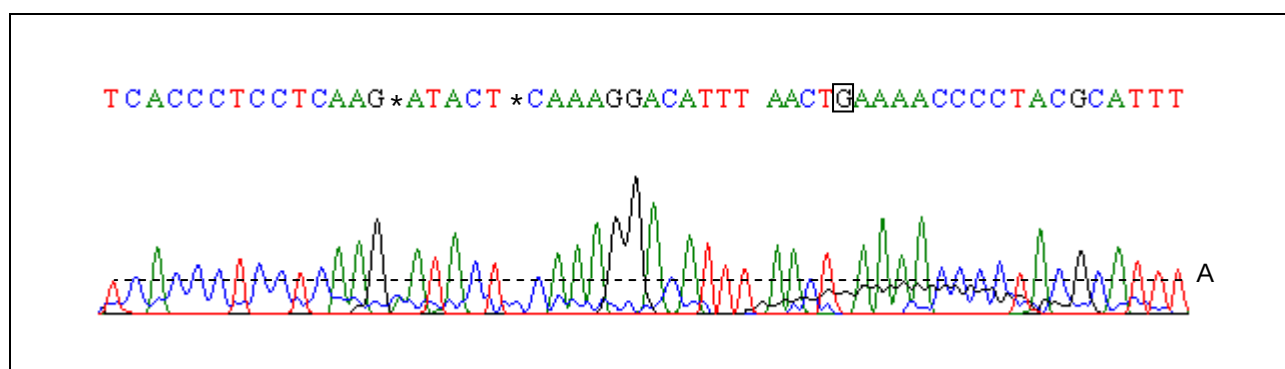
This artefact can be caused by choosing the incorrect dye set or primer (mobility) file, or by assigning the peak one location, which determines the first peak from which sequence data is analysed, incorrectly. The above refer to software settings used in the analysis of raw sequence data after electrophoresis of the sequencing samples on a sequencer. In the current investigation, cycle sequencing reactions of all samples were performed by the author (refer to Section 4.7.1). After precipitation of the sequencing samples (refer to

Section 4.7.3), electrophoresis was performed on contract by two external sequencing facilities (refer to Section 4.7.4). This artefact can be corrected by re-analysing the collected sequence data after adjusting the above software settings. In the event that poor mobility correction was observed, two strategies were followed in the current investigation. Depending on which sequence was affected, as well as the severity of the artefact, the sequence was either repeated, or re-analysis of the sequence data was requested from the sequencing facility that had performed electrophoresis of the sample exhibiting the artefact.

5.6.4.4 Insertion and deletion of nucleotides

Insertion and deletion of nucleotides in this section refers to spurious events due to artefacts in the sequencing process, and not to true insertions and deletions as caused by evolutionary processes. In the current investigation, the insertion of additional G nucleotides was observed, as well as the deletion of A, C and T nucleotides. Insertion and deletion events are usually present in electropherograms with low signal strength and are caused by incorrect base-calling as performed by the analysis software of the sequencer. An example of the insertion and deletion of nucleotides is depicted in Figure 5.13.

Figure 5.13 Representative electropherogram illustrating the insertion and deletion of nucleotides



A = adenine; C = cytosine; G = guanine; T = thymine. The asterisks (*) indicate missing T nucleotides, the block (□) indicates an inserted G nucleotide. The dashed line A indicates the maximum amplitude of the background noise.

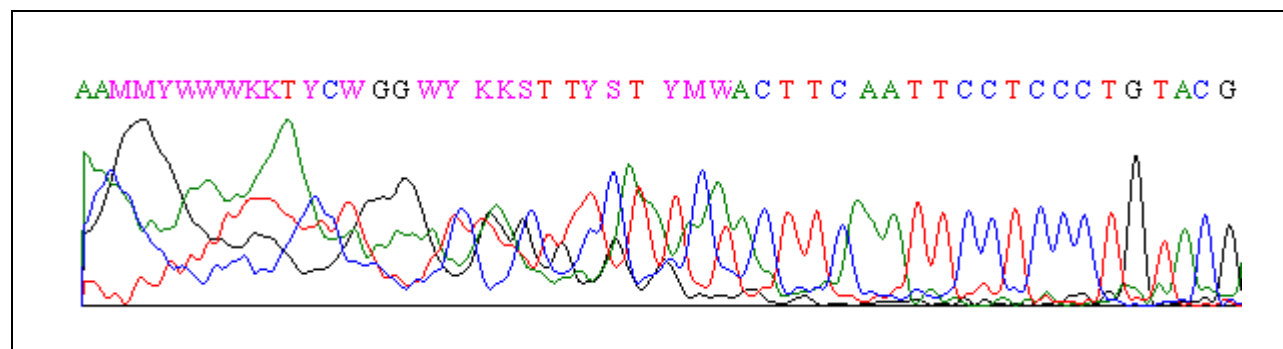
Insertion events occur in regions where there is uneven spacing of peaks leading to a slight space between two adjacent nucleotides. Within this space the base-caller incorrectly interprets a background peak as a valid peak. Deletions occur when the signal strength, i.e. fluorescence, for a specific type of nucleotide is lower than that of all the other nucleotides. During electrophoresis, this nucleotide is subsequently not detected with the other nucleotides, causing loss of peaks within the electropherogram. Both insertion

and deletion events can be prevented through optimisation of the sequencing reaction by increasing the amount of template used, thereby intensifying the signal strength and reducing background noise. Insertions can be identified through visual inspection of the electropherogram, or by comparison with a reference sequence, and can be rectified by manual editing of the sequence data. Deletions are usually identified by comparison with a reference sequence and can be corrected by repeating the sequencing reaction, or through use of sequence data from another source such as the sequence data from overlapping regions.

5.6.4.5 Excess dye peaks and dye blobs

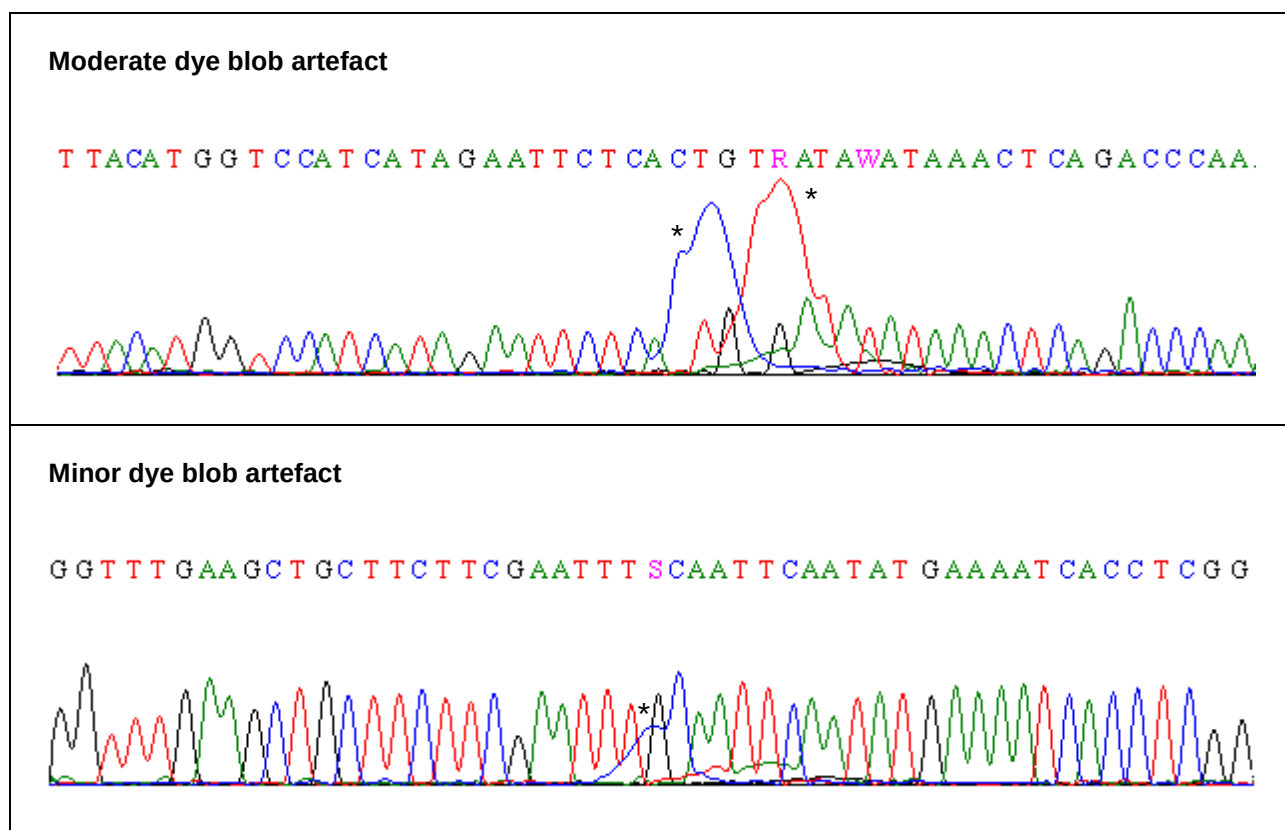
Excess dye peaks and dye blobs are common when using dye terminator cycle sequencing chemistries. Excess dye peaks are observed at the start of the sequence electropherogram leading to a loss of sequence data. An example of excess dye peaks is depicted in Figure 5.14.

Figure 5.14 Representative electropherogram illustrating excess dye peaks



A = adenine; C = cytosine; G = guanine; T = thymine; K = guanine or thymine; M = adenine or cytosine; S = cytosine or guanine; W = adenine or thymine; Y = cytosine or thymine.

Dye blobs are characterised by the presence of one or more large, broad peaks and are usually located between nucleotides 50 and 140 of the electropherogram. Dye blob artefacts can be moderate or minor, depending on the influence exerted on the base-calling software. Regions affected by moderate dye blobs were excluded from sequence analysis, while regions containing minor dye blobs were visually inspected and ambiguities resolved by manual editing. The differences between these two types of dye blobs are illustrated in Figure 5.15.

Figure 5.15 Representative electropherogram depicting dye blobs

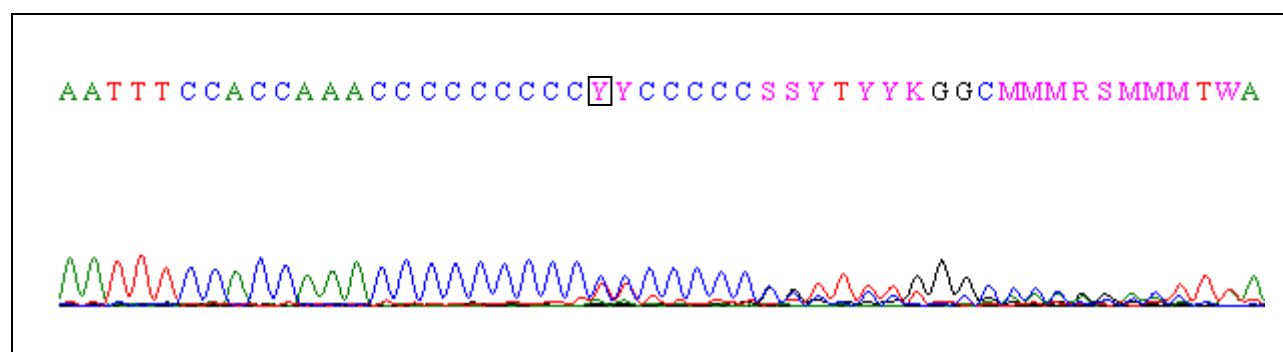
A = adenine; C = cytosine; G = guanine; T = thymine; R = adenine or guanine; S = cytosine or guanine; W = adenine or thymine. The asterisks (*) indicate the moderate and minor dye blob artefacts present in each of the electropherograms.

Excess dye peaks and dye blobs are caused by incomplete removal of unincorporated, fluorescently labelled ddNTPs during sequence precipitation. These artefacts can be avoided by choosing the appropriate precipitation method for the dye terminator cycle sequencing chemistry employed, and by using the correct concentration of non-denatured, room temperature alcohol (Applied Biosystems®, 2000). In the current investigation, two precipitation methods were used, namely ethanol/sodium acetate precipitation (Applied Biosystems®, 2000) and Centri•Sep™ 96 well clean-up plates (Princeton Separations). The use of the Centri•Sep™ 96 well plates for sequence precipitation replaced the ethanol/sodium acetate precipitation, and did not require the use of ethanol, which eliminated the possible problems associated with its use. In addition, an SDS/heat treatment preparation step was added after cycle sequencing prior to sequence precipitation, when using the Centri•Sep™ 96 well plates (Applied Biosystems®, 2002). This extra step promoted the elimination of unincorporated dye terminators from cycle sequencing reactions (refer to Section 4.7.2).

5.6.4.6 Slippage in homopolymer regions

Slippage in homopolymer regions occurred in stretches of cytosine nucleotides exceeding six nucleotides. An example of slippage in such a homopolymer region is presented in Figure 5.16.

Figure 5.16 Representative electropherogram depicting slippage in a homopolymer region



A = adenine; C = cytosine; G = guanine; T = thymine; K = guanine or thymine; M = adenine or cytosine; R = adenine or guanine; S = cytosine or guanine; W = adenine or thymine; Y = cytosine or thymine. The block (□) indicates the point at which slippage occurs in a homopolymer region.

The occurrence of slippage in homopolymer regions is dependent on the length of the affected region as short regions are generally not affected by this artefact. The exact mechanism of slippage in homopolymer regions is unknown, although it could be due to the presence of multiple PCR products with different numbers of nucleotides in the homopolymer region. The occurrence of slippage is usually more prevalent in homopolymer regions consisting of adenosine nucleotides (Applied Biosystems®, 2000). This artefact can be corrected by sequencing the complementary strand, i.e. through the use of a reverse primer, or by cloning the PCR fragment into a suitable vector and sequencing the vector. The latter option however requires more time and expense than the first option. The first option was employed in the current investigation. For all sequences affected by this artefact, the PCR reaction and PCR purification for the specific PCR region was repeated, and thereafter a cycle sequencing reaction was performed using the appropriate reverse primer (refer to Section 4.7).

5.6.5 Template quality

A number of the artefacts discussed in the previous section (Section 5.6.4) can be ascribed to the quality of the template used for automated sequencing. The quality of the template in the current investigation can be affected by a number of factors that are listed

in Table 5.5. The artefact caused by the template quality, as well as a possible solution to the problem is presented.

Table 5.5 Factors affecting template quality

Factor	Artefact	Solution
Incomplete removal of cellular components such as RNA, proteins and polysaccharides during DNA extraction from cell lines.	Background noise	Repeat DNA isolation.
Degradation of isolated gDNA during storage due to repeated freeze-thawing cycles.	Background noise	Repeat DNA isolation.
Residual salts or organic chemicals carried over after PCR amplification.	Low peak signal Background noise	Perform an ethanol precipitation to remove contaminants. Repeat PCR product purification.
Degradation of purified PCR products during storage.	Background noise	Repeat PCR reaction and PCR product purification.
Secondary amplification products present in purified PCR product.	Background noise	Re-optimize PCR conditions. Purify correct amplified fragment from agarose gel.

Adapted from the Automated Sequencing Chemistry Guide (Applied Biosystems®, 2000).

As discussed in Section 5.3.3.1, repeating the isolation of DNA was not an option due to a number of logistical constraints. Where necessary and to minimise the occurrence of background noise and increase the concentration of PCR products used for sequencing reactions, PCR reactions were re-optimised, repeated and PCR products purified.

5.6.6 Errors in mitochondrial sequence data

The importance of the quality of mitochondrial sequence data is emphasised by the discovery over the past few years that numerous mitochondrial datasets contain errors. These errors can be classified into types which relate to the analysis, evaluation and recording of the sequence data, and can be detected in datasets by the presence of specific indicators or warning signs. A number of strategies have also been devised to minimise the inclusion of errors in datasets (refer to Section 3.2.1.3).

In the current investigation, the inclusion of errors in the final sequencing data was minimised through complying with the following preventative measures. Sample mix up was prevented by the systematic labelling of each sequencing reaction with the sample number of the individual sequenced and the primer used for sequencing, as well as through the use of a sequencing strategy which allowed a minimum of 100 bp of overlap between contiguous sequences which assisted in the identification of individuals and the assembly of the complete contigs. Due to the fact that errors can be the consequence of a

high level of background (refer to Section 5.6.4), the amount of background noise was reduced through optimisation of the cycle sequencing reaction for each primer in each PCR region sequenced. The electropherograms generated by the sequencer was subjected to visual inspection. Every single nucleotide was checked and edited accordingly where mistakes were identified. The sequences determined for an individual were aligned with the version of the rCRS (Andrews *et al.*, 1999), available from MITOMAP in 2007 (J01415, GI:113200490). The sequences from a specific individual were aligned with the rCRS in one alignment file, i.e. different individuals were not included in the same alignment. Sequences that contained alterations which appeared questionable, i.e. alterations that did not occur in all sequences of the same region for a single individual, or did not fit the observed haplogroup profile, were repeated by generating a new purified PCR product and performing a cycle sequencing reaction using an appropriate reverse primer (refer to Section 4.7.1). The visual inspection, editing of sequences, and alignment of sequences with the rCRS to identify overlapping segments and alterations, were performed using BioEdit version 5.0.9 (Hall, 1999) as described in Section 4.7.5. The observed alterations were documented twice, once while assembling the sequences determined for an individual to create the completed contig, and a second time while aligning all the completed contigs for phylogenetic and statistical analyses (refer to Section 4.10). The two datasets were compared and any discrepancy regarding the presence of an alteration was resolved by referring back to the original raw sequence data, i.e. the electropherograms. This systematic approach also eliminated transcription errors that could have occurred during the documentation of alterations.

5.7 SEQUENCING RESULTS

The complete mtDNA sequences of 30 Khoi-San individuals were determined and analysed to investigate the nature and level of variation present within the Southern African Khoi-San population. Complete mitochondrial sequences were determined firstly because complete sequences were required to compile a consensus sequence, and secondly because nucleotide sequences provide the highest possible resolution in the examination of molecular evolution in populations (Vigilant *et al.*, 1989).

The sequencing results obtained for each of the eight PCR regions, that together constitute the complete mtDNA genome, are discussed separately in the following sections (Sections 5.7.1 to 5.7.8). It was expected that a large number of discrepancies would be observed between the Khoi-San mitochondrial DNA sequences and the rCRS due to the

fact that the Khoi-San sequences would most likely belong to macrohaplogroup L, while the rCRS is classified as haplogroup H (Andrews *et al.*, 1999). In addition, the Khoi-San is one of the most ancient populations (Chen *et al.*, 2000) and as such should exhibit a high level of variation.

The observed discrepancies or alterations were classified into two main groups, i.e. nucleotide substitutions which included transitions as well as transversions, and length alterations which included insertions and deletions (Page and Holmes, 1998). The observed nucleotide substitutions were further classified as a non-synonymous substitution if the alteration caused a change in the encoded amino acid or as a synonymous substitution if the alteration did not bring about a change in the encoded amino acid (Page and Holmes, 1998).

Alterations such as nucleotide substitutions that cause a change in a DNA sequence are often referred to as mutations. This can be confusing as the designation of mutation is often understood to imply a detrimental change associated with disease or an alteration of the phenotype. For the purpose of the current investigation, the designation of mutation is used to indicate causative mutations which are implicated in pathology. Any change in DNA sequence is referred to as an alteration, and an alteration that occurs at a frequency of between 0.01 and 0.99 in a population without being associated with a known disease phenotype, is referred to as a polymorphism (Falconer and Mackay, 1996). Polymorphisms are indicators of the level of genetic variation present within an individual. Genetic variation between individuals/organisms can be used to infer the ancestral history of a group of individuals/organisms, to investigate genetic relationships among sub-populations in a species or as genetic markers for disease (Hartl, 1988).

A total of 292 alterations were observed of which 20 were classified as previously unreported or novel. All the observed alterations are listed in Appendix A and the molecular information for each of the observed alterations is listed in Appendix B. Heteroplasmy was not observed for any of the previously reported or novel alterations, although it is possible that it was not detected due to the sequencing strategy employed (refer to Appendix B). In addition, no mitochondrial disorders or chronic disease were recorded for any of the individuals included in the original sample group collected in the north-western Kalahari Desert in southern Africa (Chen *et al.*, 2000). In the absence of heteroplasmy, and since the individuals in the Khoi-San cohort appeared to be healthy and not suffering from any diagnosed disease, it was therefore considered unlikely that the

30 Khoi-San individuals included in the current investigation would harbour any alterations that could be described as causative mitochondrial mutations. In addition, it was expected that most alterations would be located within the control (non-coding) region of the mitochondrial genome compared to the coding region, i.e. the rate of nucleotide substitution would be different for the two regions. This distribution pattern can be attributed to the mechanism of functional constraint (Page and Holmes, 1998).

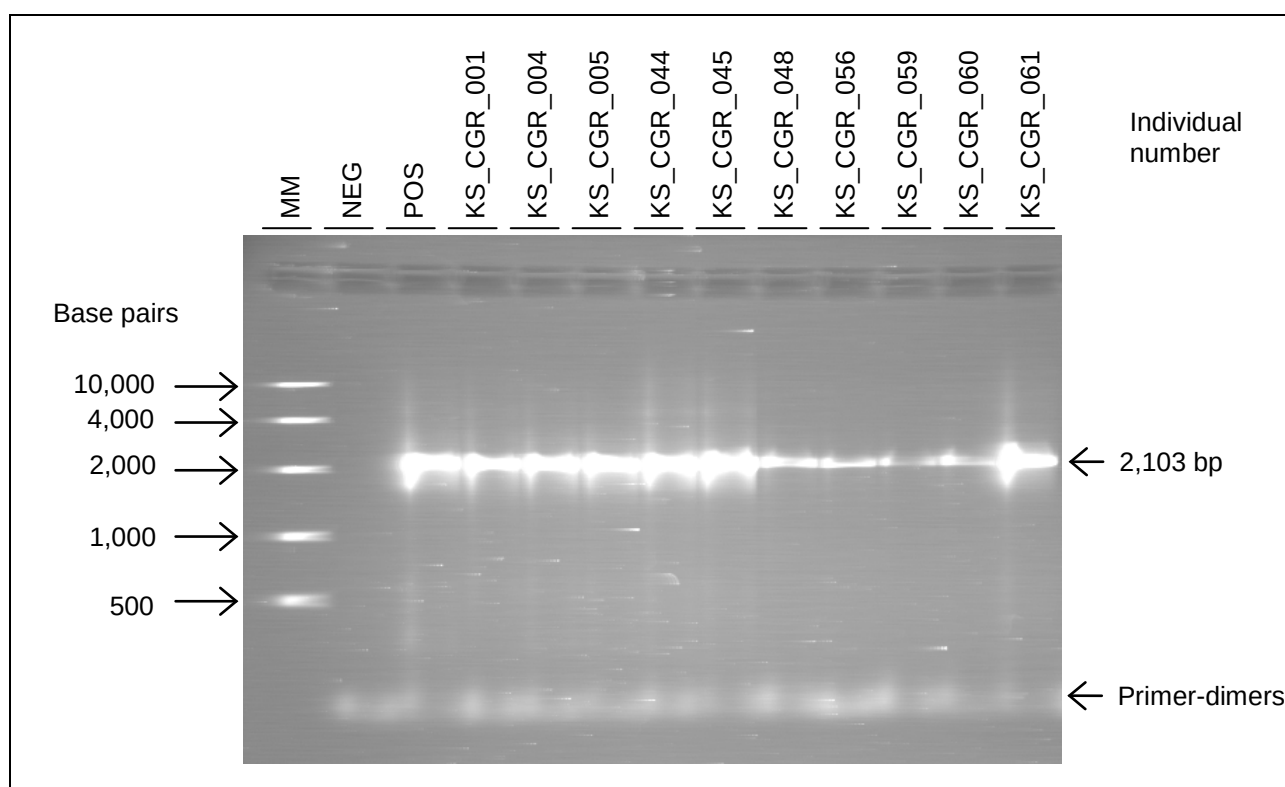
On a nucleotide level, functional constraint dictates that non-coding DNA evolves at a higher rate than coding DNA. Within the coding DNA, alterations occurring in different coding positions (first, second or third codon position) exhibit differences in the probability of altering an amino acid due to the degeneracy of the genetic code and the fact that the three nucleotide positions within the codon may differ in the level of functional constraint the positions are subjected to. When investigating mitochondrial genome sequences, the most variation is therefore expected within the control region (D-loop) and at the third codon position which is generally implicated in synonymous changes.

The identification and confirmation of alterations compared to the rCRS (Andrews *et al.*, 1999) is performed keeping in mind that errors can be made. The causes of errors in sequence databases, the inclusion of these errors in the databases and strategies employed to avoid these errors were discussed in Sections 3.2.1.3 and 5.6.6. Another source of sequencing error over which the researcher has little or no control is the nucleotide changes introduced into mitochondria that are present in cell lines that are subjected to repeated passaging. The passage number of a cell line refers to the number of times the cells in the culture have been subcultured. It has been observed for HeLa cells that repeated passaging introduces changes in the mitochondrial genome sequence and that these changes are observed as instances of heteroplasmy (Herrnstadt *et al.*, 2002b). The passage number for the lymphoblastoid cell lines used for gDNA isolation for individuals studied in the current investigation ranged between a minimum of 13 to a maximum of 18 (Mouton, 2003). These passage numbers are relatively low. Changes in the mitochondrial genome sequence are introduced over time periods of years, it is therefore expected that alterations observed in the mitochondrial genome sequence compared to the rCRS will be true differences.

5.7.1 PCR region 1 of the mitochondrial genome

PCR region 1, or the first segment of the mitochondrial genome, consists of a total of 2,103 bp and contains base pairs 15,975 to 1,508 according to the numbering system of the rCRS (Andrews *et al.*, 1999). This region contains a number of functional elements which include two tRNA genes, the 12S rRNA gene, the D-loop, HVS1, HVS2, 7S DNA, TAS, various control elements, origins of replication and promoter regions for both the L-strand and the H-strand, three conserved sequence blocks, as well as mtTF1 binding sites (refer to Section 2.4.2, Table 2.3). PCR was performed according to the conditions listed in Section 4.4, and the PCR products obtained were electrophoresed on an agarose gel as discussed in Section 4.5. A representative gel picture of the amplified products obtained for PCR region 1 is depicted in Figure 5.17.

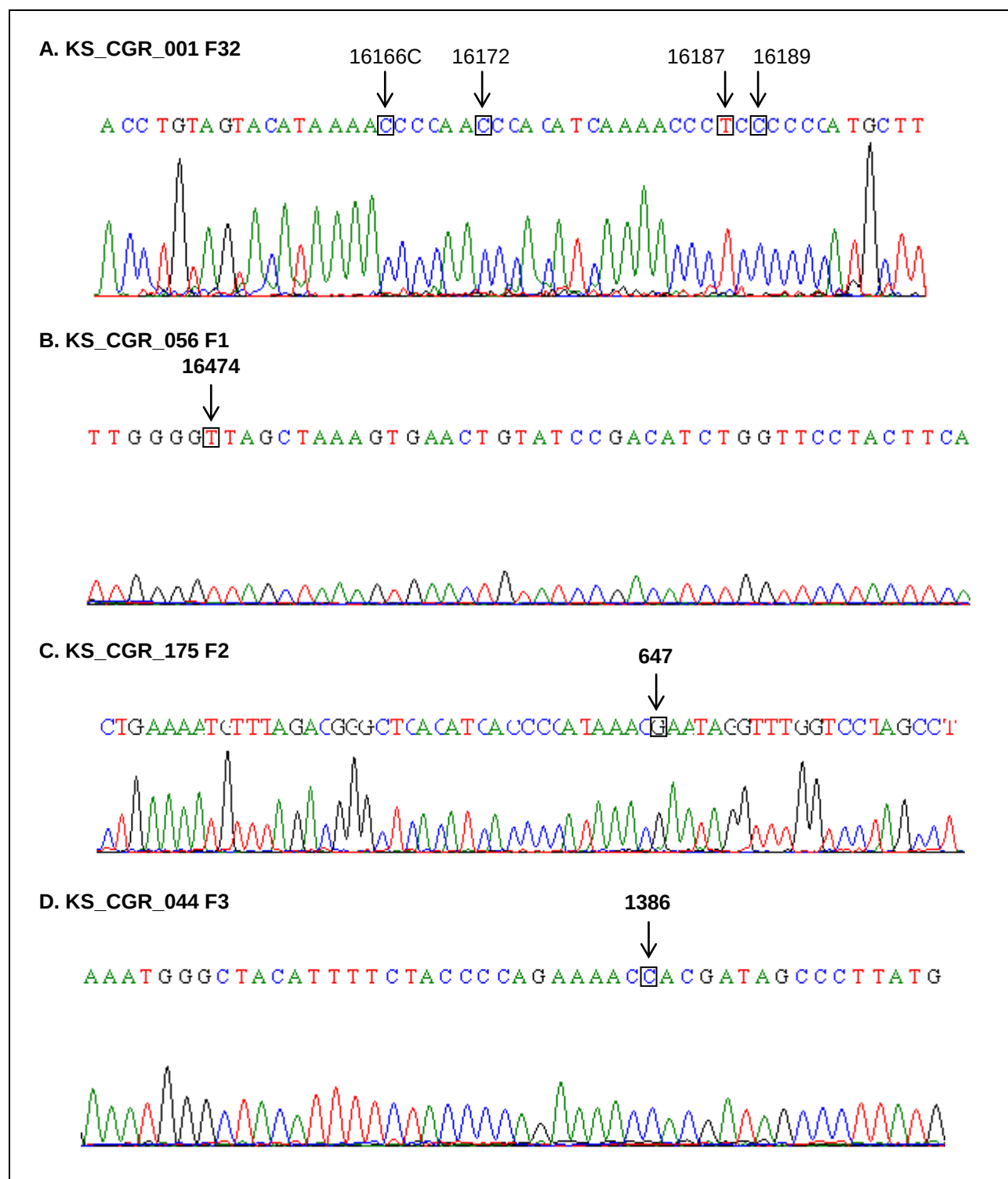
Figure 5.17 Photographic representation of amplified products of PCR region 1



Fragments were electrophoresed on a 0.9% agarose gel at 80 V for 45 min. MM = FastRuler™ DNA Ladder, High Range, ready to use, molecular marker (the sizes of the marker fragments are indicated to the left of the figure); NEG = negative control; POS = positive control; KS_CGR_001 to KS_CGR_061 = sample numbers given to each Khoi-San individual. The amplification efficiency was not equal for all samples, background smears are present in the sample lanes and primer-dimers are observed as discussed in Section 5.3.3. Artefacts are present in the gel matrix and the fragments are distorted as discussed in Section 5.4.1.

After gel electrophoresis, the PCR products were purified as discussed in Section 4.4.3 and sequencing of the PCR region was performed as described in Section 4.7. The sequencing results were analysed as discussed in Section 4.7.5. Representative electropherograms for each of the four forward primers used for the sequence determination of this region are presented in Figure 5.18.

Figure 5.18 Representative electropherograms of PCR region 1 generated using forward primers F32 to F3

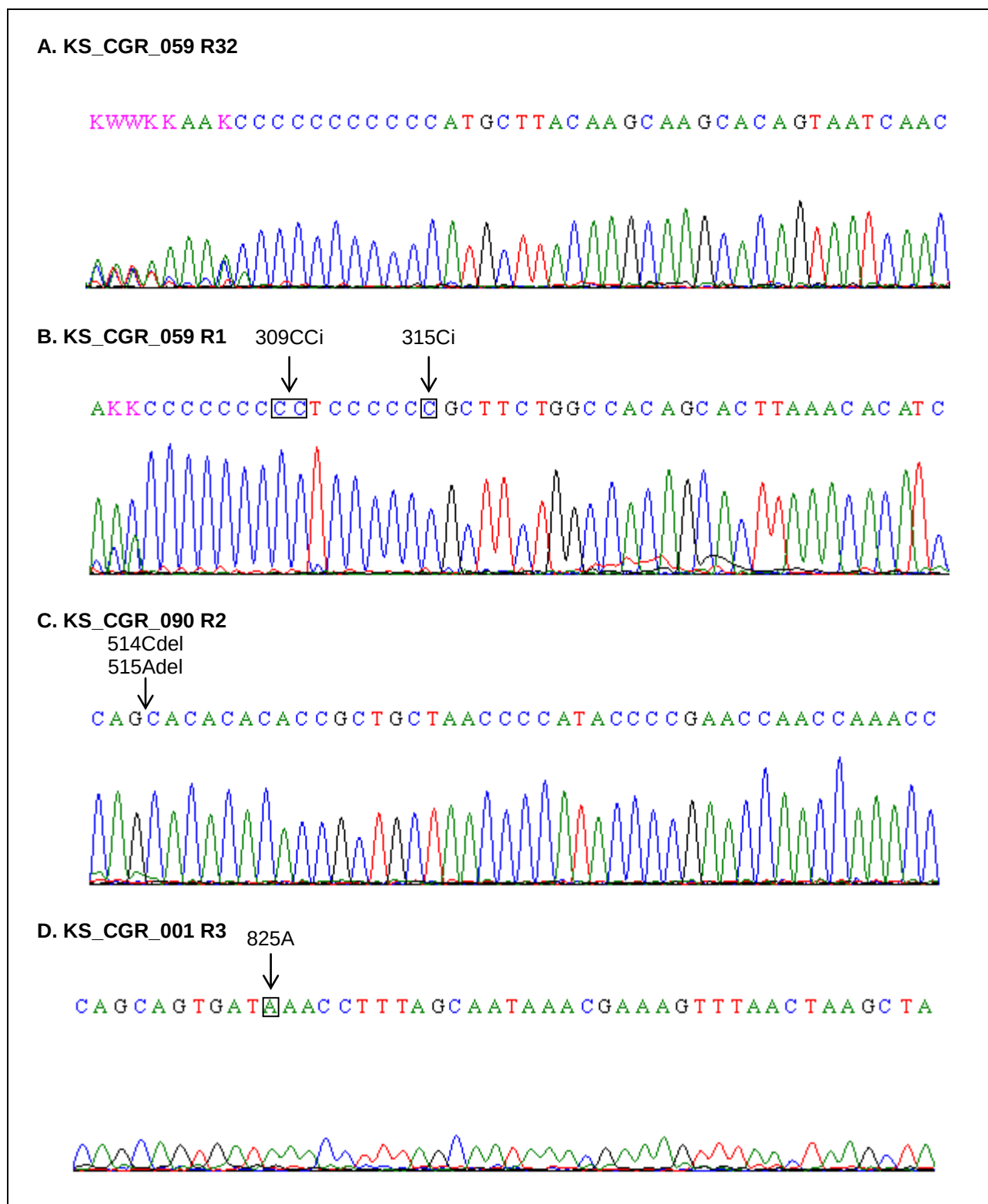


A = adenine; C = cytosine; G = guanine; T = thymine. The first part of the sequence name designates the sample, and the second part of the name indicates the primer used to generate the sequence. The nucleotide positions of alterations are indicated by blocks (□). Alterations are transitions unless specified otherwise. Reported alterations are indicated in regular black text and novel alterations are indicated in **bold** black text. Poor mobility correction, low signal strength and background noise can be observed in the electropherograms as discussed in Section 5.6.4.

In the event that electropherograms contained sequence artefacts (refer to Section 5.6.4), or when alterations were observed that appeared to be sequence errors (refer to Section 5.6.6), reverse primers were used to repeat the appropriate sequencing reactions.

Representative electropherograms for each of the four reverse primers used for sequencing PCR region 1 are presented in Figure 5.19.

Figure 5.19 Representative electropherograms of PCR region 1 generated using reverse primers R32 to R3



A = adenine; C = cytosine; G = guanine; T = thymine; K = guanine or thymine; W = adenine or thymine. The first part of the sequence name designates the sample, and the second part of the name indicates the primer used to generate the sequence. The sequence results are presented as the reverse complement. The nucleotide positions of alterations are indicated by blocks (□). Alterations are transitions unless specified otherwise. i = insertion; del = deletion. Reported alterations are indicated in regular black text. Homopolymer slippage, minor dye blobs, background noise, low signal strength and poor mobility correction can be observed in the electropherograms as discussed in Section 5.6.4.

The electropherograms representing sequence results obtained using the reverse primers are presented as the reverse complement. According to convention for mitochondrial sequence data, the sequence of the L-strand is usually presented in the 5' to 3' orientation. The reverse primers anneal to the H-strand, therefore in order to present the sequence data according to convention, the electropherogram representing the sequence data obtained using the reverse primer is converted to the reverse complement and presented in this manner. Presenting all electropherograms of sequence data obtained using the reverse primer in this manner also permitted direct comparison between sequences.

Comparison of PCR region 1 sequences with the section of the rCRS representing PCR region 1 (refer to Section 4.7.5), revealed a total of 85 discrepancies which consisted of 70 transitions, seven transversions, five insertions and three deletion events. Of the 85 discrepancies, four transitions represented novel polymorphisms. The sequence alterations present in all screened Khoi-San individuals are presented in Sections 5.7.1.1 to 5.7.1.3. PCR region 1 has been divided into functional units to facilitate discussion of the observed alterations.

5.7.1.1 Displacement loop

The D-loop region is the non-coding region located between base pairs 16,024 and 576, and contains HVS1 (bp 16,024 to bp 16,383) and HVS2 (bp 57 to bp 372). These regions are recognised sources of variation within the mitochondrial genome and it was anticipated that a large number of sequence differences would be observed within this region. The observed alterations are indicated in Table 5.6.

Table 5.6 Sequence alterations observed between the complete Khoi-San mitochondrial DNA genome sequences and the rCRS in the D-loop

Alteration		Locus	Alteration type	Codon position	Reference
16,069	C – T	MT-DLOOP	TS	non-coding	Di Rienzo and Wilson, 1991
16,093	T – C	MT-DLOOP	TS	non-coding	Di Rienzo and Wilson, 1991
16,114	C – A	MT-DLOOP	TV	non-coding	Horai and Hayasaka, 1990
16,124	T – C	MT-DLOOP	TS	non-coding	Greenberg <i>et al.</i> , 1983
16,126	T – C	MT-DLOOP	TS	non-coding	Cann and Wilson, 1983
16,129	G – A	MT-DLOOP	TS	non-coding	Aquadro and Greenberg, 1983
16,153	G – A	MT-DLOOP	TS	non-coding	Vigilant <i>et al.</i> , 1991
16,166	A – C	MT-DLOOP	TV	non-coding	Wilkinson-Herbots <i>et al.</i> , 1996
16,167	C – T	MT-DLOOP	TS	non-coding	Aquadro and Greenberg, 1983
16,169	C – T	MT-DLOOP	TS	non-coding	Horai and Hayasaka, 1990

Table 5.6 Continued...

Alteration		Locus	Alteration type	Codon position	Reference
16,172	T – C	MT-DLOOP	TS	non-coding	Greenberg <i>et al.</i> , 1983
16,182	A – C	MT-DLOOP	TV	non-coding	Lahermo <i>et al.</i> , 1996
16,183	A – C	MT-DLOOP	TV	non-coding	Ozawa <i>et al.</i> , 1995
16,187	C – T	MT-DLOOP	TS	non-coding	Aquadro and Greenberg, 1983
16,188	C – T	MT-DLOOP	TS	non-coding	Shields <i>et al.</i> , 1993
16,189	T – C	MT-DLOOP	TS	non-coding	Aquadro and Greenberg, 1983
16,193	INS C	MT-DLOOP	INS	non-coding	Brandstätter <i>et al.</i> , 2004
16,209	T – C	MT-DLOOP	TS	non-coding	Di Rienzo and Wilson, 1991
16,212	A – G	MT-DLOOP	TS	non-coding	Vigilant <i>et al.</i> , 1989
16,213	G – A	MT-DLOOP	TS	non-coding	Horai and Hayasaka, 1990
16,214	C – T	MT-DLOOP	TS	non-coding	Vigilant <i>et al.</i> , 1989
16,223	C – T	MT-DLOOP	TS	non-coding	Aquadro and Greenberg, 1983
16,230	A – G	MT-DLOOP	TS	non-coding	Aquadro and Greenberg, 1983
16,234	C – T	MT-DLOOP	TS	non-coding	Horai and Hayasaka, 1990
16,239	C – T	MT-DLOOP	TS	non-coding	Vigilant <i>et al.</i> , 1989
16,242	C – T	MT-DLOOP	TS	non-coding	Aquadro and Greenberg, 1983
16,243	T – C	MT-DLOOP	TS	non-coding	Aquadro and Greenberg, 1983
16,264	C – T	MT-DLOOP	TS	non-coding	Horai and Hayasaka, 1990
16,266	C – G	MT-DLOOP	TV	non-coding	Vigilant <i>et al.</i> , 1989
16,278	C – T	MT-DLOOP	TS	non-coding	Aquadro and Greenberg, 1983
16,291	C – T	MT-DLOOP	TS	non-coding	Horai and Hayasaka, 1990
16,291	C – G	MT-DLOOP	TV	non-coding	Ingman <i>et al.</i> , 2000
16,294	C – T	MT-DLOOP	TS	non-coding	Aquadro and Greenberg, 1983
16,304	T – C	MT-DLOOP	TS	non-coding	Aquadro and Greenberg, 1983
16,311	T – C	MT-DLOOP	TS	non-coding	Aquadro and Greenberg, 1983
16,320	C – T	MT-DLOOP	TS	non-coding	Aquadro and Greenberg, 1983
16,354	C – T	MT-DLOOP	TS	non-coding	Piercy <i>et al.</i> , 1993
16,362	T – C	MT-DLOOP	TS	non-coding	Aquadro and Greenberg, 1983
16,390	G – A	MT-DLOOP	TS	non-coding	Horai and Hayasaka, 1990
16,474	G – T	MT-DLOOP	TS	non-coding	Current investigation
16,497	A – G	MT-DLOOP	TS	non-coding	Horai and Hayasaka, 1990
16,519	T – C	MT-DLOOP	TS	non-coding	Aquadro and Greenberg, 1983
73	A – G	MT-DLOOP	TS	non-coding	Aquadro and Greenberg, 1983
146	T – C	MT-DLOOP	TS	non-coding	Aquadro and Greenberg, 1983
150	C – T	MT-DLOOP	TS	non-coding	Aquadro and Greenberg, 1983
152	T – C	MT-DLOOP	TS	non-coding	Aquadro and Greenberg, 1983
182	C – T	MT-DLOOP	TS	non-coding	Aquadro and Greenberg, 1983
183	A – G	MT-DLOOP	TS	non-coding	Liu <i>et al.</i> , 2001
189	A – G	MT-DLOOP	TS	non-coding	Aquadro and Greenberg, 1983
195	T – C	MT-DLOOP	TS	non-coding	Aquadro and Greenberg, 1983
198	C – T	MT-DLOOP	TS	non-coding	Vigilant <i>et al.</i> , 1989
199	T – C	MT-DLOOP	TS	non-coding	Vigilant <i>et al.</i> , 1989
204	T – C	MT-DLOOP	TS	non-coding	Graven <i>et al.</i> , 1995

Table 5.6 Continued...

Alteration		Locus	Alteration type	Codon position	Reference
207	G – A	MT-DLOOP	TS	non-coding	Ozawa <i>et al.</i> , 1991a
247	G – A	MT-DLOOP	TS	non-coding	Aquadro and Greenberg, 1983
263	A – G	MT-DLOOP	TS	non-coding	Aquadro and Greenberg, 1983
265	T – C	MT-DLOOP	TS	non-coding	Barbosa <i>et al.</i> , 2008
309	C – T	MT-DLOOP	TS	non-coding	Brandstätter <i>et al.</i> , 2004
309	INS C	MT-DLOOP	INS	non-coding	Vigilant <i>et al.</i> , 1989
309	INS CC	MT-DLOOP	INS	non-coding	Vigilant <i>et al.</i> , 1989
310	T – C	MT-DLOOP	TS	non-coding	Budowle <i>et al.</i> , 1999
315	INS C	MT-DLOOP	INS	non-coding	Andrews <i>et al.</i> , 1999
385	A – G	MT-DLOOP	TS	non-coding	Redd <i>et al.</i> , 1995
418	C – T	MT-DLOOP	TS	non-coding	Mishmar <i>et al.</i> , 2003
456	C – T	MT-DLOOP	TS	non-coding	Aquadro and Greenberg, 1983
494	DEL C	MT-DLOOP	DEL	non-coding	Mishmar <i>et al.</i> , 2003
514	DEL C	MT-DLOOP	DEL	non-coding	Ozawa <i>et al.</i> , 1991a
515	DEL A	MT-DLOOP	DEL	non-coding	Ozawa <i>et al.</i> , 1991a
573	INS CCCC	MT-DLOOP	INS	non-coding	Torroni <i>et al.</i> , 1994a

Nucleotides are represented by the capital letters, with the rCRS nucleotide indicated first, followed by the observed change. Nucleotide numbers are based on rCRS numbering (Andrews *et al.*, 1999). The numbers indicated in **red** represent previously reported polymorphisms. The numbers indicated in **blue** represent previously unreported (novel) polymorphisms. MT-DLOOP = control region, including displacement loop; A = adenine; C = cytosine; G = guanine; T = thymine; TV = transversion; TS = transition; DEL = deletion; INS = insertion.

The D-loop contains the bulk of the alterations observed for PCR region 1. A single novel alteration, G16474T, was observed in this region in a single Khoi-San individual (KS_CGR_056). Although the D-loop region consists of non-coding sequence, a number of control elements, binding sites and conserved sequence blocks are located within this region (refer to Section 2.4.2, Table 2.3). This alteration however does not occur in any of the known control elements, binding sites or conserved sequence blocks within the D-loop region. It is therefore expected that the observed alteration will have no functional effect. Since the alteration has only been observed in a single individual of sub-haplogroup L0b (refer to Appendix E) it is not likely to be haplogroup-specific or haplogroup-associated, and will probably only be useful for identification of the specific individual within the L0b sub-haplogroup.

The other observed alterations have all been reported previously. Most of the alterations were nucleotide substitutions which involved the change of a single nucleotide to another nucleotide. Some alterations however were length alterations and involved insertions and deletion events which added or deleted mono- or dinucleotide segments within repeating nucleotide stretches, leading to alterations in the total length of the mitochondrial genome sequence. Insertion events occurred at two cytosine homopolymer stretches located at

nucleotide position 303 to 309 and nucleotide position 311 to 315. Functional importance has been suggested for these homopolymer regions such as involvement in the initiation of replication (Sekiya *et al.*, 1980) and H-strand transcription (Cantatore and Attardi, 1980) due to their proximity to these loci.

Deletion events occurred at nucleotide position 494, 514 and 515. The deletions at nucleotide positions 514 and 515 removed a dinucleotide segment from the (CA)_n repeat motif. The (CA)_n repeating nucleotide stretch, due to its purine-pyrimidine character, has the ability to assume a left-handed (Z) helical conformation under certain physiological conditions (Klysik *et al.*, 1981; Nordheim *et al.*, 1981). This conformation, due to its location upstream of the 7S RNA transcriptional start site, may be associated with several physical and biological functions (Cantor, 1981). The variability in length of this region however does not compromise its potential functionality, as the conformation and integrity of the repeat is not affected by its length. The above mentioned length alterations can both be attributed to the primer-strand slippage model for frameshift mutagenesis (Streisinger *et al.*, 1966). In addition, it has been observed that mutational events that generate length changes are usually reversible (Drake and Baltz, 1976).

5.7.1.2 tRNA phenylalanine

The tRNA phenylalanine region consists of bp 577 to bp 647. The observed alterations are indicated in Table 5.7.

Table 5.7 Sequence alterations observed between the complete Khoi-San mitochondrial DNA genome sequences and the rCRS in the tRNA^{Phe} gene

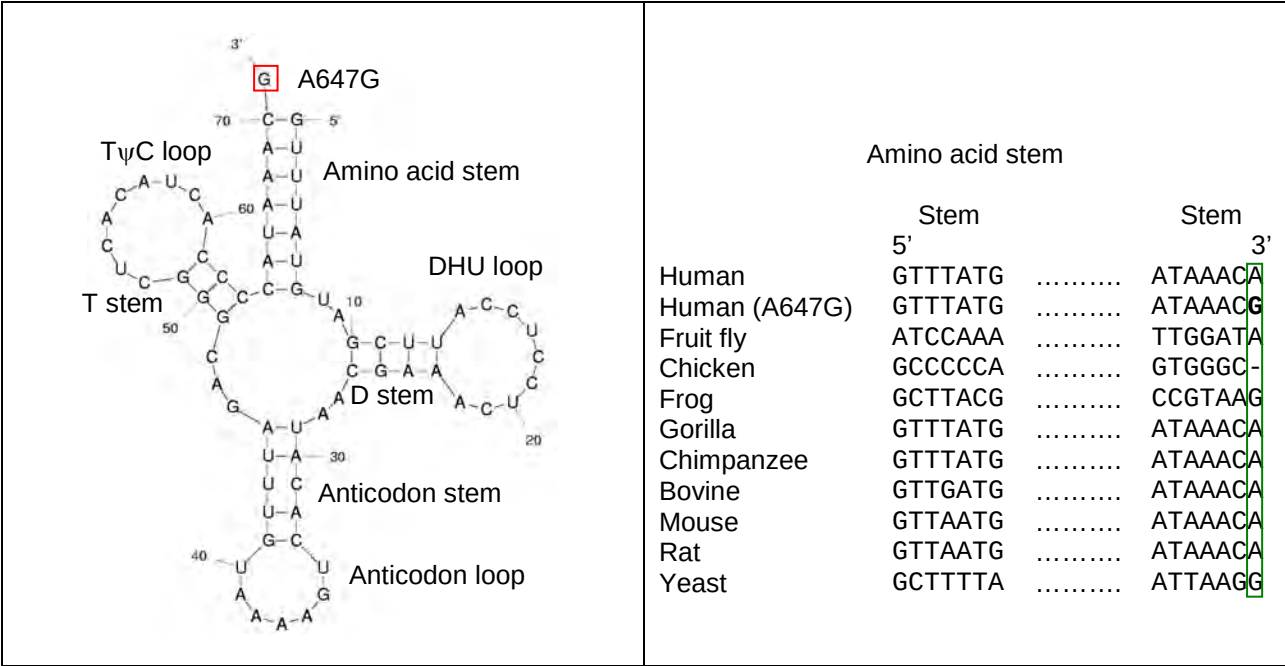
Alteration		Locus	Alteration type	Codon position	Reference
593	T – C	MT-TF	TS	non-coding	Herrnstadt <i>et al.</i> , 2002a
647	A – G	MT-TF	TS	non-coding	Current investigation

Nucleotides are represented by the capital letters, with the rCRS nucleotide indicated first, followed by the observed change. Nucleotide numbers are based on rCRS numbering (Andrews *et al.*, 1999). The number indicated in red represents a previously reported polymorphism, while the number indicated in blue represents a previously unreported (novel) polymorphism. MT-TF = tRNA phenylalanine; A = adenine; C = cytosine; G = guanine; T = thymine; TS = transition.

A single novel alteration, A647G, was observed in this region in all three Khoi-San individuals (KS_CGR_140, KS_CGR_143 and KS_CGR_175) of the L3a1 sub-haplogroup (refer to Appendix E). To determine the functional importance of this tRNA gene alteration, it is necessary to perform functional translation studies. This type of study did not form part of the current investigation, although it was possible to analyse the secondary structure of

the tRNA^{Phe} using the mfold web server version 3.2 (Mathews *et al.*, 1999; Zuker, 2003). This web server, through the use of a number of closely related software programs, predicts the secondary structure or folding of single stranded nucleic acids. The predicted structure of the tRNA^{Phe} containing the A647G alteration is presented in Figure 5.20.

Figure 5.20 Proposed secondary structure and sequence alignment of tRNA^{Phe} containing the A647G alteration



The proposed secondary structure of the tRNA^{Phe} was determined using the mfold web server version 3.2 (Mathews *et al.*, 1999; Zuker, 2003) and is depicted on the left. Hydrogen bonds between cytosine and guanine nucleotides are indicated in dark red. A sequence alignment based on tRNA^{Phe} sequences obtained from GenBank® (accession numbers indicated below), constructed as described in Section 4.12.1, is presented on the right to illustrate the evolutionary conservation of the region of the tRNA^{Phe} containing the A647G alteration. In the figure the nucleotide position of the A647G alteration is indicated by a block (□), and in the sequence alignment the A647G alteration is indicated in bold text, while the corresponding positions in the different sequences are indicated by a box (□). Human = *Homo sapiens* (J01415, GI:113200490); fruit fly = *Drosophila melanogaster* (NC_001709, GI:5835233); chicken = *Gallus gallus* (NC_001323, GI:5834843); frog = *Xenopus laevis* (NC_001573, GI:5834981); gorilla = *Gorilla gorilla* (NC_001645, GI:5835149); chimpanzee = *Pan paniscus* (NC_001644, GI:5835135); bovine = *Bos taurus* (NC_006853, GI:60101824); mouse = *Mus musculus* (NC_005089, GI:34538597); rat = *Rattus norvegicus* (NC_001665, GI:110189714); yeast = *Saccharomyces cerevisiae* (NC_001224, GI:6226515). A = adenine; C = cytosine; G = guanine; T = thymine; U = uracil.

The A647G alteration is located in the amino acid stem of the secondary structure of the tRNA^{Phe}. This nucleotide is however not involved in base pairing with any of the other nucleotides within the amino acid stem and a change in this nucleotide therefore does not disrupt formation of the amino acid stem. It is therefore likely that the A647G alteration is not functionally significant or pathogenic. This is further supported by a comparison of the tRNA^{Phe} gene in ten different organisms. From the alignment presented in Figure 5.20, it can be observed that an adenine nucleotide at this position is highly conserved in mammals (human, gorilla, chimpanzee, bovine, mouse and rat), whereas a guanine nucleotide is observed in the frog and yeast. Therefore, either an adenine or a guanine is observed at position 647. This suggests that either one of the two nucleotides can be present at this position for the correct folding of the tRNA^{Phe} molecule.

The lack of a functional effect for the alteration is further supported by the fact that the alteration was observed in individuals without any diagnosed form of mitochondrial disease, as well as the presence of this alteration in all individuals belonging to the L3a1 sub-haplogroup. As the A647G alteration was observed in all the individuals harbouring this sub-haplogroup, it is possible that this alteration is associated with sub-haplogroup L3a1. Although the A647G alteration has not previously been reported in haplogroup L3 or sub-haplogroup L3a, it is however possible that the alteration is characteristic of the L3a1 sub-haplogroup as observed in the Khoi-San population.

5.7.1.3 12S ribosomal RNA

The 12S ribosomal RNA region consists of bp 648 to bp 1,601. The observed alterations are indicated in Table 5.8.

Table 5.8 Sequence alterations observed between the complete Khoi-San mitochondrial DNA genome sequences and the rCRS in the 12S rRNA gene

Alteration		Locus	Alteration type	Codon position	Reference
709	G – A	MT-RNR1	TS	non-coding	Ozawa <i>et al.</i> , 1991a
719	G – A	MT-RNR1	TS	non-coding	Kivisild <i>et al.</i> , 2006
750	A – G	MT-RNR1	TS	non-coding	Andrews <i>et al.</i> , 1999
769	G – A	MT-RNR1	TS	non-coding	Greenberg <i>et al.</i> , 1983
825	T – A	MT-RNR1	TV	non-coding	Prezant <i>et al.</i> , 1993
850	T – C	MT-RNR1	TS	non-coding	Crimi <i>et al.</i> , 2002
921	T – C	MT-RNR1	TS	non-coding	Liu <i>et al.</i> , 2001
957	C – T	MT-RNR1	TS	non-coding	Current investigation
1,018	G – A	MT-RNR1	TS	non-coding	Prezant <i>et al.</i> , 1993
1,048	C – T	MT-RNR1	TS	non-coding	Moraes <i>et al.</i> , 1993
1,243	T – C	MT-RNR1	TS	non-coding	Jones <i>et al.</i> , 2001
1,386	T – C	MT-RNR1	TS	non-coding	Current investigation
1,438	A – G	MT-RNR1	TS	non-coding	Ozawa <i>et al.</i> , 1991b
1,442	G – A	MT-RNR1	TS	non-coding	Herrnstadt <i>et al.</i> , 2002a

Nucleotides are represented by the capital letters, with the rCRS nucleotide indicated first, followed by the observed change. Nucleotide numbers are based on rCRS numbering (Andrews *et al.*, 1999). The numbers indicated in **red** represent previously reported polymorphisms. The numbers indicated in **blue** represent previously unreported (novel) polymorphisms. MT-RNR1 = 12S ribosomal RNA; A = adenine; C = cytosine; G = guanine; T = thymine; TS = transition; TV = transversion.

Two novel alterations, C957T and T1386C, were observed in this region. C957T was detected in two individuals (KS_CGR_048 and KS_CGR_060) of the L0b sub-haplogroup, while T1386C was observed in one individual (KS_CGR_044) of the L0b sub-haplogroup (refer to Appendix E). The functional importance of these rRNA gene alterations can not be determined without performing functional translation studies. The secondary structure

of the 12S rRNA has however been predicted (Zwieb *et al.*, 1981) and using this information the likely functional consequences of the alterations can be envisioned.

The C957T alteration is located in a loop region between nucleotide positions 306 and 320 of the 12S rRNA and is therefore not involved in base pairing. Mouse mitochondrial 12S rRNA has a deletion of three C residues in this region at nucleotide position 315 to 317 compared to human mitochondrial 12S rRNA (Zwieb *et al.*, 1981). Insertions have also been reported in this region, such as an insertion of five C residues between nucleotide 956 and 965 detected in a single individual suffering from late-onset Alzheimer's disease and Parkinson's disease (Shoffner *et al.*, 1993). This insertion has however not been confirmed to be a causative mutation in either of these or other disorders. These deletion and insertion events suggest that the loop region can tolerate some variation. It is therefore likely that the C957T alteration will not have a functional effect on the 12S rRNA.

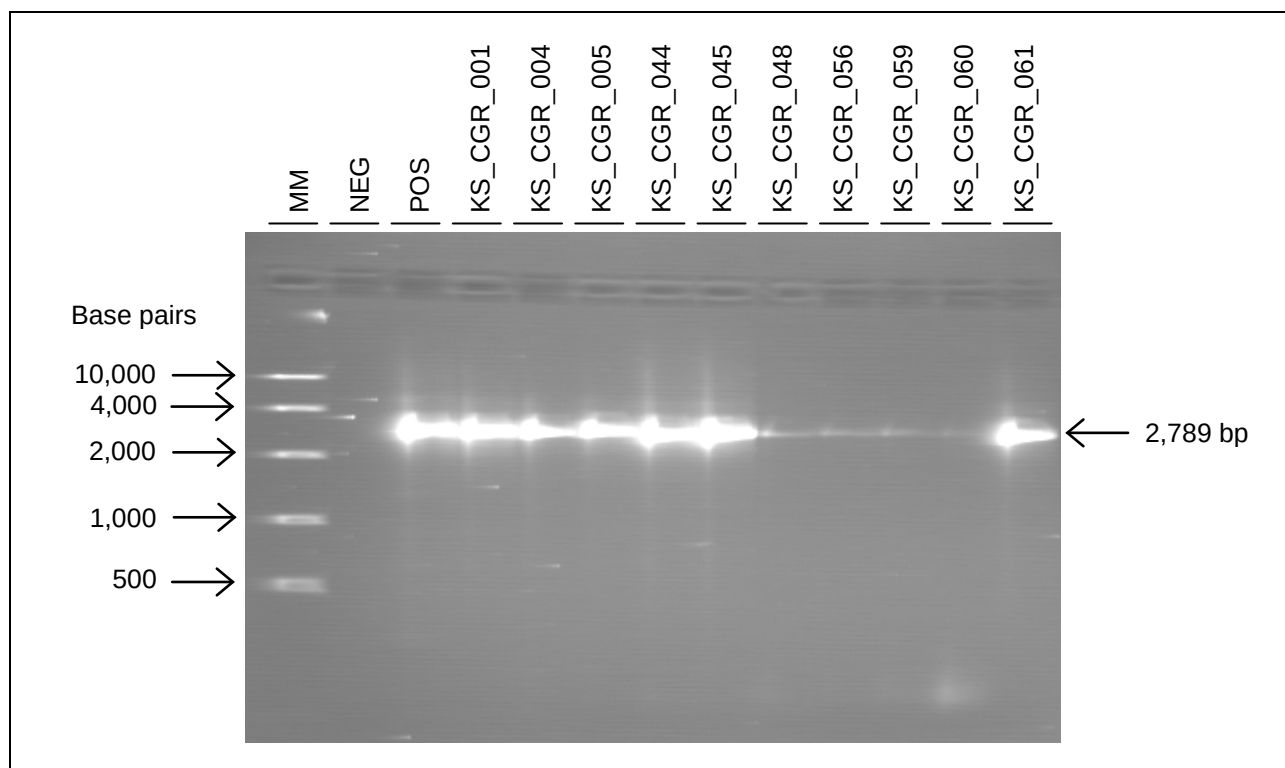
The T1386C alteration is located in a short stem region where base pairing occurs between four base pairs. This region exhibits some variability compared to the mouse mitochondrial 12S rRNA (Zwieb *et al.*, 1981), this includes a loop with modified base pairing and a deleted nucleotide preceding the four nucleotide stem, as well as different nucleotides following the four nucleotide stem. The variability in this section of the 12S rRNA between mouse and human indicates a low level of conservation between species, which suggests that this region is not subjected to functional constraint and can therefore tolerate some variation. It is therefore anticipated that the T1386C alteration will not have a functional effect on the 12S rRNA. In addition, since the alteration was only observed in a single individual of sub-haplogroup L0b, it is envisaged that it will be useful for identification of the specific individual within the L0b sub-haplogroup. The T1386C alteration is therefore not expected to be haplogroup-specific.

5.7.2 PCR region 2 of the mitochondrial genome

PCR region 2 comprises a total of 2,789 bp and includes base pairs 902 to 3,690 based on the numbering system of the rCRS (Andrews *et al.*, 1999). This region contains a number of functional elements which include two tRNA genes, the 16S rRNA gene, the 5S-like sequence, the transcription terminator, the ND1 gene and a few non-coding nucleotides (refer to Section 2.4.2, Table 2.3). PCR was performed as described in Section 4.4, and the PCR products obtained were electrophoresed on an agarose gel as recounted in

Section 4.5. A representative gel picture for the amplified products of PCR region 2 is presented in Figure 5.21.

Figure 5.21 Photographic representation of amplified products of PCR region 2

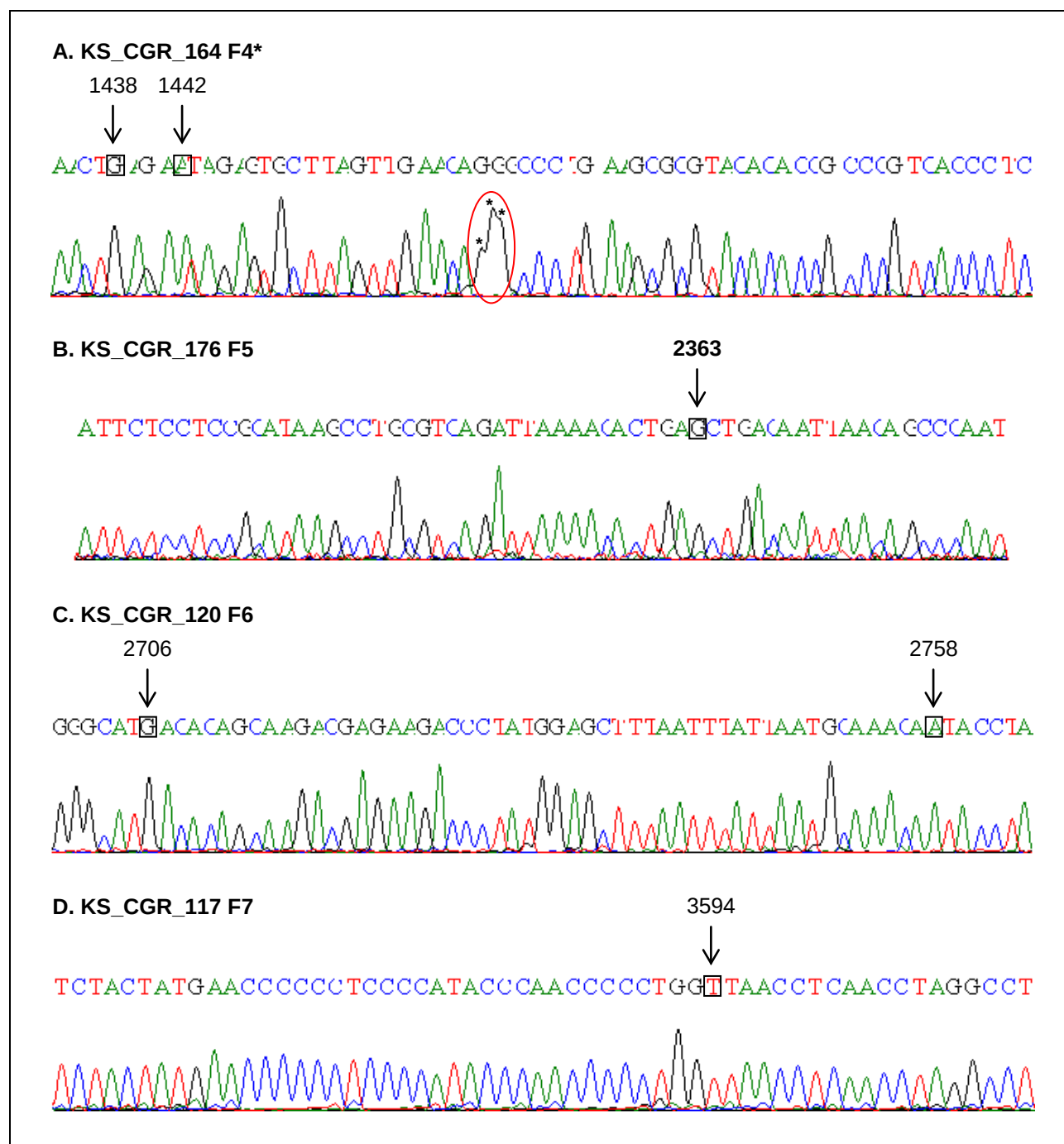


Fragments were electrophoresed on a 0.9% agarose gel at 80 V for 45 min. MM = FastRuler™ DNA Ladder, High Range, ready to use, molecular marker (the sizes of the marker fragments are indicated to the left of the figure); NEG = negative control; POS = positive control; KS_CGR_001 to KS_CGR_061 = sample numbers given to each Khoi-San individual. The amplification efficiency was not equal for all samples, background smears are present in the sample lanes and primer-dimers are observed as discussed in Section 5.3.3. Artefacts are present in the gel matrix and the fragments are distorted as discussed in Section 5.4.1.

After gel electrophoresis, PCR product purification was performed as described in Section 4.4.3 and the sequence of the PCR region was determined as discussed in Section 4.7. Sequencing results were analysed as reported in Section 4.7.5. Representative electropherograms for each of the four forward primers used to determine the sequence of PCR region 2 are depicted in Figure 5.22.

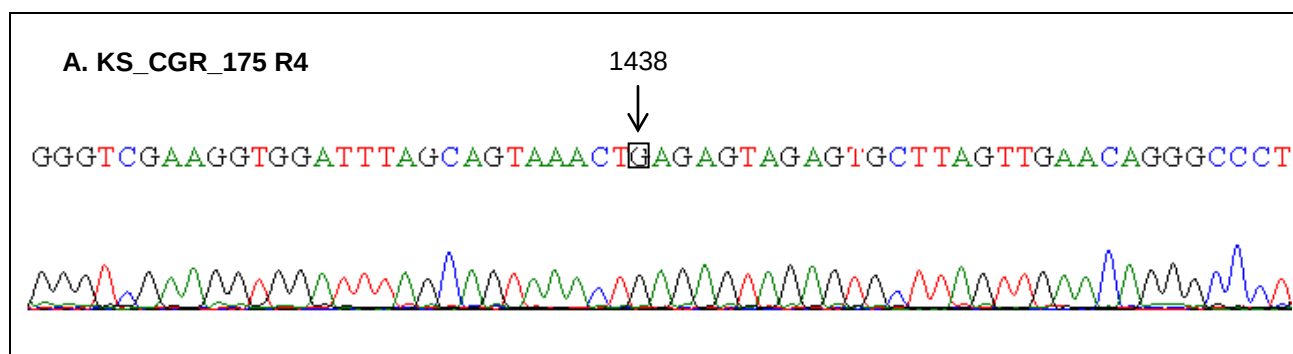
The section of the electropherogram for KS_CGR_164 F4* circled in red was interpreted as three peaks representing three G nucleotides. This interpretation was firstly based on the base-calling as performed by the analysis software of the sequencer, and was supported by the visual inspection of the peaks in the affected region which allowed the identification of three peaks (top of each peak is indicated by an asterisk). The presence of three G nucleotides in this sequence region was subsequently confirmed by comparing the sequence with the rCRS, which indicated that three G nucleotides were usually located within this region of the mitochondrial genome sequence.

Figure 5.22 Representative electropherograms of PCR region 2 generated using forward primers F4* to F7



A = adenine; C = cytosine; G = guanine; T = thymine. The first part of the sequence name indicates the sample, and the second part of the name designates the primer used to generate the sequence. The nucleotide positions of alterations are indicated by blocks (□). Alterations are transitions unless specified otherwise. Reported alterations are indicated in regular black text and novel alterations are indicated in **bold** black text. A poorly resolved sequence region is indicated by a red oval (○) and the separate peaks within this region are indicated by asterisks (*). Poor mobility correction, background noise and low signal strength can be observed in the electropherograms as discussed in Section 5.6.4.

A representative electropherogram for the reverse primer used to generate sequence data for PCR region 2 is depicted in Figure 5.23. As previously discussed, the electropherogram representing the sequence result obtained using a reverse primer is presented as the reverse complement (refer to Section 5.7.1).

Figure 5.23 A representative electropherogram of PCR region 2 generated using reverse primer R4

A = adenine; C = cytosine; G = guanine; T = thymine. The first part of the sequence name designates the sample, and the second part of the name indicates the primer used to generate the sequence. The sequence result is presented as the reverse complement. The nucleotide positions of alterations are indicated by blocks (□). Alterations are transitions unless specified otherwise. A reported alteration is indicated in regular black text. Background noise and low signal strength can be observed in the electropherogram as discussed in Section 5.6.4.

The sequence data generated for PCR region 2 was compared to the corresponding section of the rCRS (refer to Section 4.7.5). A total of 29 discrepancies were observed which consisted of 26 transitions and three transversions. Of the 29 discrepancies, one transition represents an alteration not previously reported. The sequence alterations observed in the screened Khoi-San individuals are presented in Sections 5.7.2.1 and 5.7.2.2. The observed alterations are discussed according to the functional elements present in PCR region 2.

5.7.2.1 16S ribosomal RNA

The 16S ribosomal RNA region consists of bp 1,671 to bp 3,228. The observed alterations are indicated in Table 5.9.

Table 5.9 Sequence alterations observed between the complete Khoi-San mitochondrial DNA genome sequences and the rCRS in the 16S rRNA gene

Alteration		Locus	Alteration type	Codon position	Reference
1,706	C – T	MT-RNR2	TS	non-coding	Herrnstadt <i>et al.</i> , 2002a
1,719	G – A	MT-RNR2	TS	non-coding	Noer <i>et al.</i> , 1991
2,332	C – T	MT-RNR2	TS	non-coding	Herrnstadt <i>et al.</i> , 2002a
2,352	T – C	MT-RNR2	TS	non-coding	Chen <i>et al.</i> , 1995
2,358	A – G	MT-RNR2	TS	non-coding	Herrnstadt <i>et al.</i> , 2002a
2,363	A – G	MT-RNR2	TS	non-coding	Current investigation
2,416	T – C	MT-RNR2	TS	non-coding	Herrnstadt <i>et al.</i> , 2002a
2,650	C – T	MT-RNR2	TS	non-coding	Tang <i>et al.</i> , 1997
2,706	A – G	MT-RNR2	TS	non-coding	Tanaka <i>et al.</i> , 1991
2,758	G – A	MT-RNR2	TS	non-coding	Moraes <i>et al.</i> , 1993

Table 5.9 Continued...

Alteration		Locus	Alteration type	Codon position	Reference
2,836	C – A	MT-RNR2	TV	non-coding	Ingman <i>et al.</i> , 2000
2,885	T – C	MT-RNR2	TS	non-coding	Moraes <i>et al.</i> , 1993

Nucleotides are represented by the capital letters, with the rCRS nucleotide indicated first, followed by the observed change. Nucleotide numbers are based on rCRS numbering (Andrews *et al.*, 1999). The numbers indicated in red represent previously reported polymorphisms. The number indicated in blue represents a previously unreported polymorphism. MT-RNR2 = 16S ribosomal RNA; A = adenine; C = cytosine; G = guanine; T = thymine; TS = transition; TV = transversion; DEL = deletion; INS = insertion.

The majority of the alterations observed for the 16S rRNA gene have been previously reported, with the exception of a transition at nucleotide position 2,363. This previously unreported alteration, A2363G, was observed within the 16S rRNA gene of one Khoi-San individual (KS_CGR_176) belonging to the L0a sub-haplogroup (refer to Appendix E). Functional translation studies are usually required to determine if an rRNA alteration will have a functional effect. The secondary structure of human mitochondrial 16S rRNA has however been predicted (Glotz *et al.*, 1981) and through using this structure it is possible to speculate about the possible functional effects of this alteration.

The A2363G alteration is observed in a loop region between nucleotide positions 679 and 694 of the 16S rRNA and is therefore not involved in base pairing. Two deletions have been reported in this region at nucleotide positions 685 (685delA) and 688 (688delC) for mouse mitochondrial 16S rRNA compared to human mitochondrial 16S rRNA (Glotz *et al.*, 1981). In addition, eight out of the 14 nucleotides (including the nucleotide at position 692 which is equivalent to nucleotide 2,363), differ in this loop region upon comparison between mouse and human mitochondrial 16S rRNA (Glotz *et al.*, 1981). The deletion events and variability suggest that the loop region can tolerate large levels of variation and it is therefore likely that the A2363G alteration will not have a functional effect on the 16S rRNA.

5.7.2.2 NADH dehydrogenase subunit 1

The NADH dehydrogenase subunit 1 region consists of bp 3,307 to bp 4,262. The observed alterations are indicated in Table 5.10.

Table 5.10 Sequence alterations observed between the complete Khoi-San mitochondrial DNA genome sequences and the rCRS in the ND1 gene

Alteration		Locus	Alteration type	Codon position	Amino acid change	Reference
3,438	G – A	MT-ND1	TS	coding: 3	SYN	Thomas <i>et al.</i> , 1996
3,498	C – G	MT-ND1	TV	coding: 3	SYN	Kivisild <i>et al.</i> , 2006
3,516	C – A	MT-ND1	TV	coding: 3	SYN	Herrnstadt <i>et al.</i> , 2002a
3,591	G – A	MT-ND1	TS	coding: 3	SYN	Schwartz and Vissing, 2002
3,594	C – T	MT-ND1	TS	coding: 3	SYN	Chen <i>et al.</i> , 1995
3,606	A – G	MT-ND1	TS	coding: 3	SYN	Herrnstadt <i>et al.</i> , 2002a
3,618	T – C	MT-ND1	TS	coding: 3	SYN	Herrnstadt <i>et al.</i> , 2002a
3,666	G – A	MT-ND1	TS	coding: 3	SYN	Herrnstadt <i>et al.</i> , 2002a
3,756	A – G	MT-ND1	TS	coding: 3	SYN	Mishmar <i>et al.</i> , 2003
3,981	A – G	MT-ND1	TS	coding: 3	SYN	Mishmar <i>et al.</i> , 2003
4,025	C – T	MT-ND1	TS	coding: 2	NON-SYN: Thr → Met	Huoponen <i>et al.</i> , 1993
4,044	A – G	MT-ND1	TS	coding: 3	SYN	Mishmar <i>et al.</i> , 2003
4,092	G – A	MT-ND1	TS	coding: 3	SYN	Finnilä <i>et al.</i> , 2001a
4,104	A – G	MT-ND1	TS	coding: 3	SYN	Kösel <i>et al.</i> , 1998
4,158	A – G	MT-ND1	TS	coding: 3	SYN	Chen <i>et al.</i> , 2000
4,197	C – T	MT-ND1	TS	coding: 3	SYN	Kivisild <i>et al.</i> , 2006
4,232	T – C	MT-ND1	TS	coding: 2	NON-SYN: Ile → Thr	Herrnstadt <i>et al.</i> , 2002a

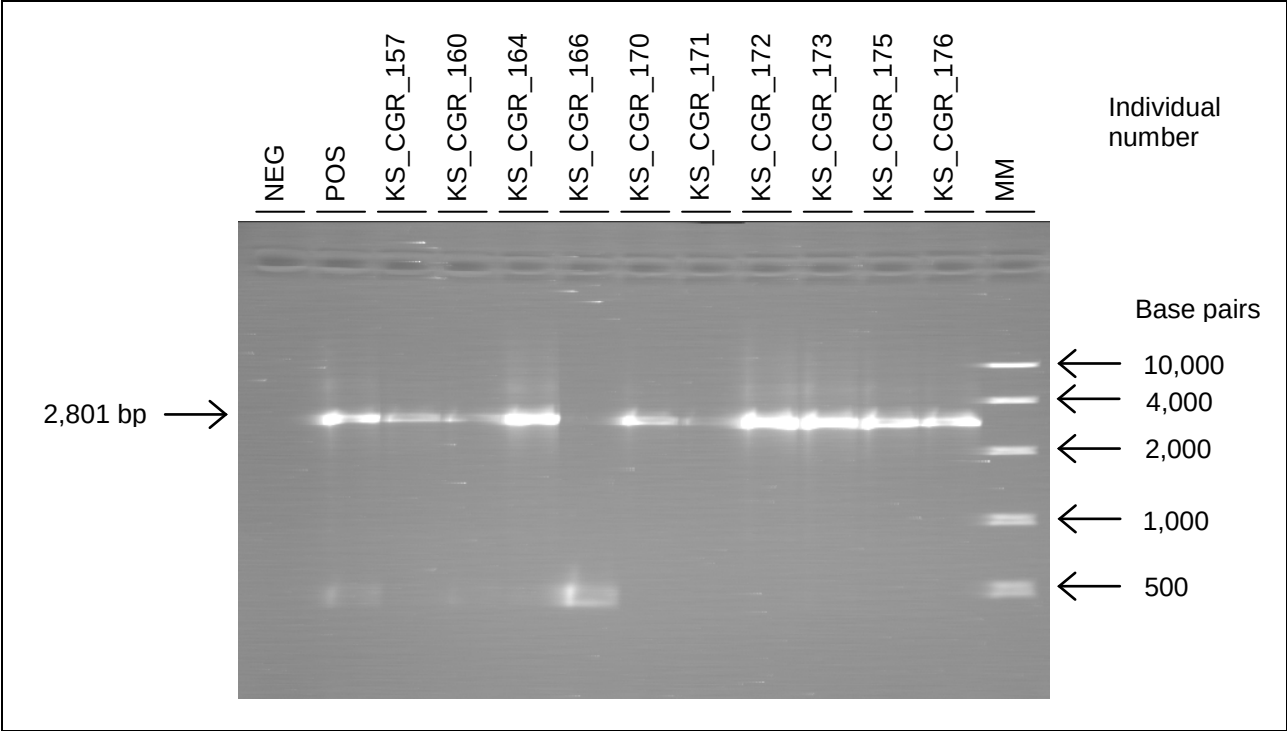
Nucleotides are represented by the capital letters, with the rCRS nucleotide indicated first, followed by the observed change. Nucleotide numbers are based on rCRS numbering (Andrews *et al.*, 1999). The numbers indicated in **red** represent previously reported polymorphisms. MT-ND1 = NADH dehydrogenase subunit 1; A = adenine; C = cytosine; G = guanine; T = thymine; TS = transition; TV = transversion; INS = insertion; DEL = deletion; SYN = synonymous substitution; NON-SYN = non-synonymous substitution; Ile = isoleucine; Met = methionine; Thr = threonine.

All of the alterations observed for the ND1 gene in the current investigation have been previously reported. Of the list of transitions and transversions presented, the majority involve synonymous changes and only two of the transitions give rise to non-synonymous changes. The alteration at nucleotide position 4,025 leads to an amino acid change from a polar threonine to a non-polar methionine (Campbell, 1995). Although the biochemical properties of these two amino acids are different, it is not expected that this alteration will be deleterious due to the fact that, to date, no disease has been reported in which this alteration is identified as the causative mutation (MITOMAP, 2008). The alteration at nucleotide 4,232 results in a non-polar isoleucine changing to a polar threonine (Campbell, 1995). Due to the fact that the biochemical properties of the two amino acids are different, it can be speculated that this change could have a potential functional effect. However, as discussed in Section 5.11.5.2, it is suggested that the alteration is retained due to the effect of random genetic drift and it is not expected that this alteration will have a deleterious functional effect.

5.7.3 PCR region 3 of the mitochondrial genome

PCR region 3 consists of a total of 2,801 bp and contains base pairs 3,051 to 5,851 according to the numbering system of the rCRS (Andrews *et al.*, 1999). This region contains a number of functional elements which include eight tRNA genes, the ND2 gene, the L-strand origin of replication and a number of non-coding nucleotides (refer to Section 2.4.2, Table 2.3). PCR was performed according to the conditions described in Section 4.4. The PCR products obtained were electrophoresed on an agarose gel as discussed in Section 4.5 and a representative gel picture of the amplified products obtained for PCR region 3 is illustrated in Figure 5.24.

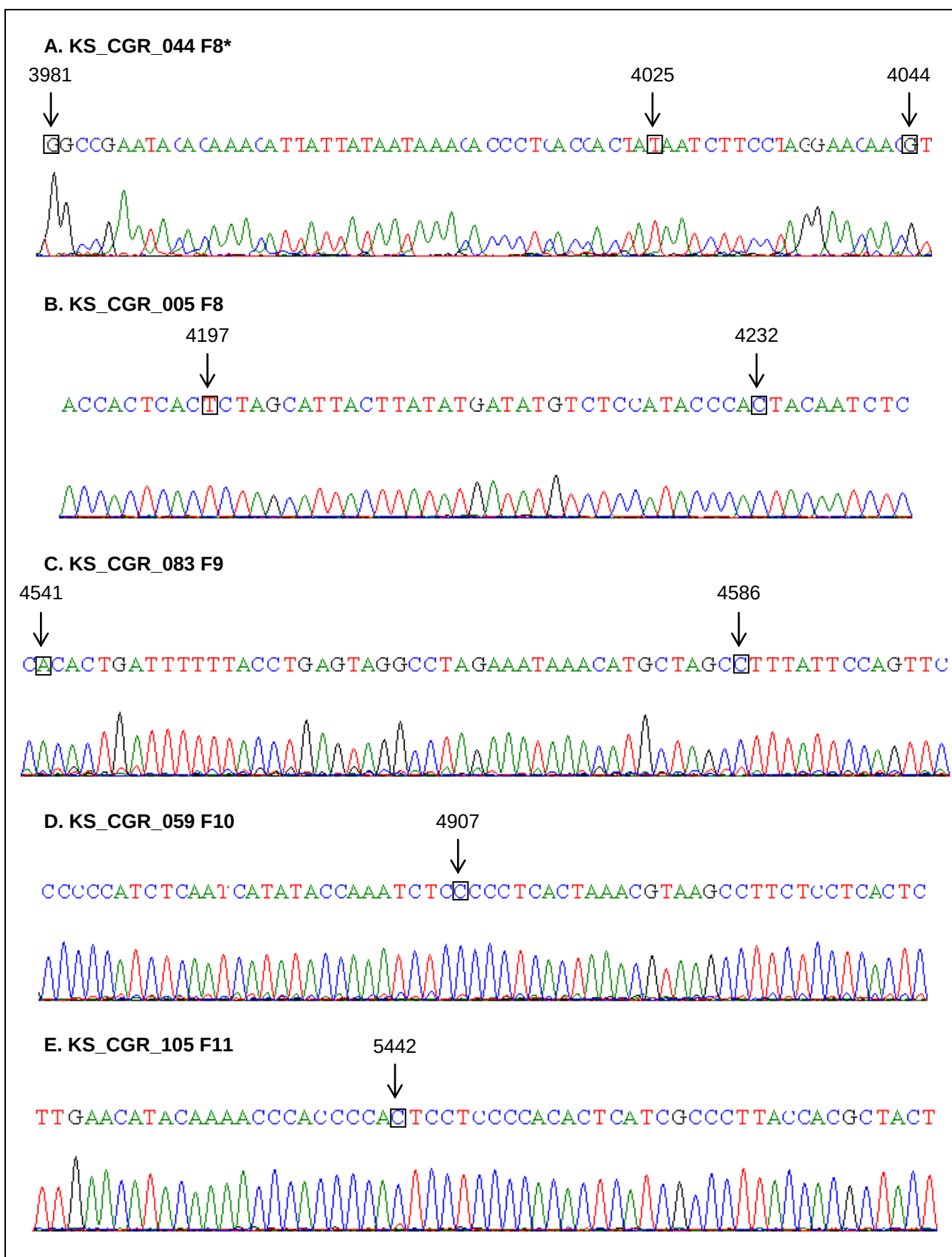
Figure 5.24 Photographic representation of amplified products of PCR region 3



Fragments were electrophoresed on a 0.9% agarose gel at 80 V for 45 min. MM = FastRuler™ DNA Ladder, High Range, ready to use, molecular marker (the sizes of the marker fragments are indicated to the right of the figure); NEG = negative control; POS = positive control; KS_CGR_157 to KS_CGR_176 = sample numbers given to each Khoi-San individual. The amplification efficiency was not equal for all samples, background smears are present in the sample lanes and secondary amplification products are observed as discussed in Section 5.3.3. Artefacts are present in the gel matrix and the fragments are distorted as discussed in Section 5.4.1.

Gel electrophoresis was followed by the purification of the PCR products as described in Section 4.4.3 and the sequence of the PCR region was obtained as discussed in Section 4.7. The sequence results obtained for this PCR region were analysed as described in Section 4.7.5. Representative electropherograms for each of the five forward primers used to generate sequence data for PCR region 3 are presented in Figure 5.25.

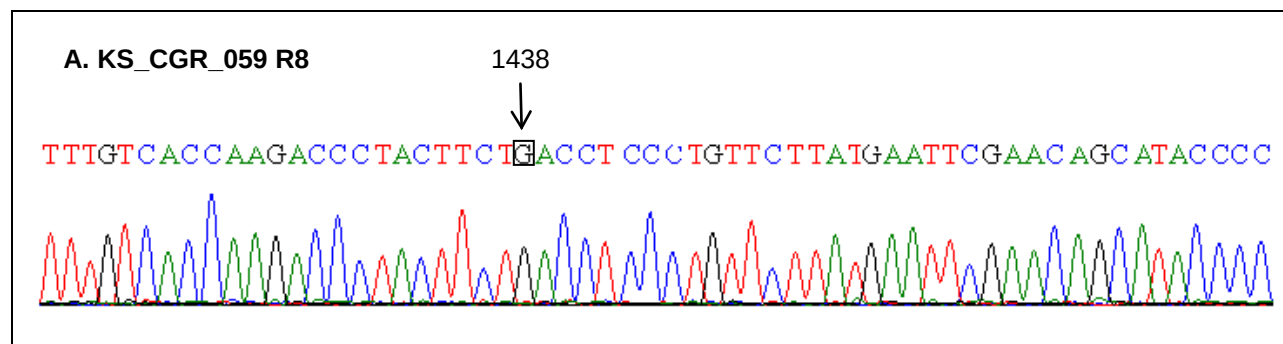
Figure 5.25 Representative electropherograms of PCR region 3 generated using forward primers F8* to F11



A = adenine; C = cytosine; G = guanine; T = thymine. The first part of the sequence name designates the sample, and the second part of the name indicates the primer used to generate the sequence. The nucleotide positions of alterations are indicated by blocks (□). Alterations are transitions unless specified otherwise. All the alterations observed in this region have been reported and are indicated in regular black text. Poor mobility correction, low signal strength and background noise can be observed in the electropherograms as discussed in Section 5.6.4.

A representative electropherogram for the reverse primer used to generate sequence data for PCR region 3 is illustrated in Figure 5.26. The sequence result obtained using the reverse primer is illustrated in the reverse complement (refer to Section 5.7.1).

Figure 5.26 A representative electropherogram of PCR region 3 generated using reverse primer R8



A = adenine; C = cytosine; G = guanine; T = thymine. The first part of the sequence name designates the sample, and the second part of the name indicates the primer used to generate the sequence. The sequence result is presented as the reverse complement. The nucleotide position of an alteration is indicated by a block (□). The observed alteration, a transition, has been reported and is indicated in regular black text. Background noise can be observed in the electropherogram as discussed in Section 5.6.4.

A total of 18 discrepancies were revealed when sequence data generated for PCR region 3 was compared to the corresponding section of the rCRS (refer to Section 4.7.5). The 18 discrepancies included 17 transitions and one transversion. A total of 16 alterations were previously reported polymorphisms, while two alterations were novel. The sequence alterations present in the screened Khoi-San individuals are presented in Sections 5.7.3.1 and 5.7.3.2. To facilitate the discussion of the observed alterations, PCR region 3 has been divided into functional units.

5.7.3.1 tRNA genes

The tRNA gene regions consist of bp 4,263 to bp 4,331 (tRNA isoleucine), bp 4,329 to bp 4,400 (tRNA glutamine), bp 5,512 to bp 5,579 (tRNA tryptophan) and bp 5,761 to bp 5,826 (tRNA cysteine). The alterations observed are indicated in Table 5.11.

Table 5.11 Sequence alterations observed between the complete Khoi-San mitochondrial DNA genome sequences and the rCRS in the tRNA genes

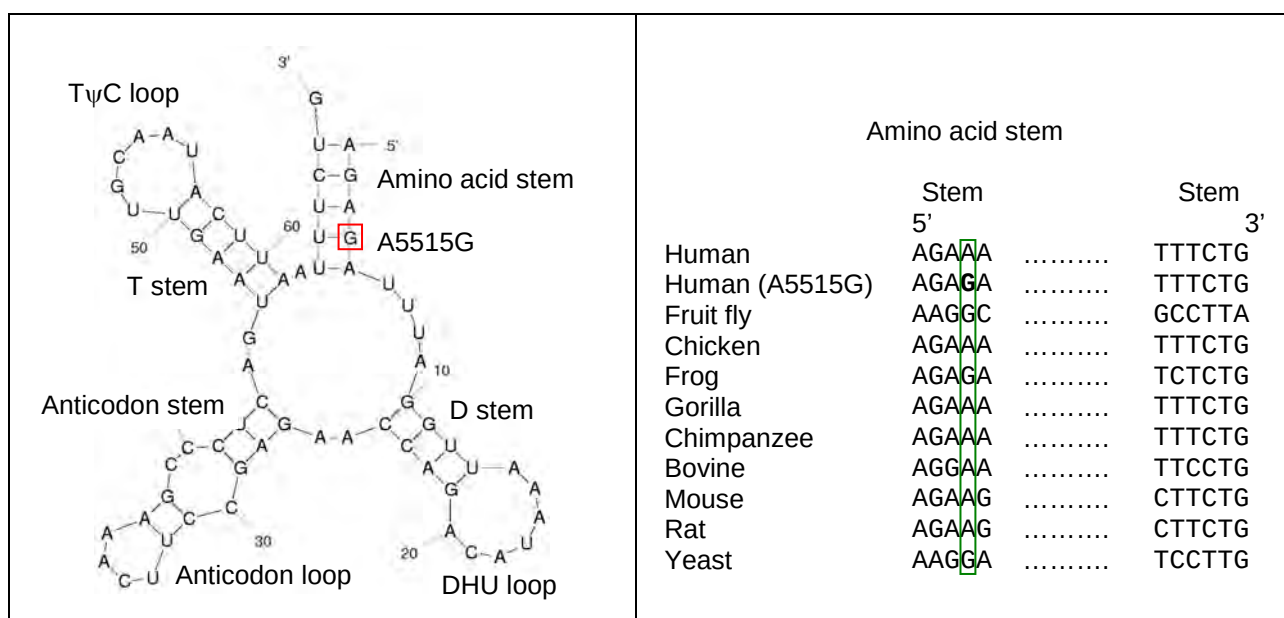
Alteration		Locus	Alteration type	Codon position	Reference
4,312	C – T	MT-TI	TS	non-coding	Moraes <i>et al.</i> , 1993
4,370	T – C	MT-TQ	TS	non-coding	Herrnstadt <i>et al.</i> , 2002a
5,515	A – G	MT-TW	TS	non-coding	Current investigation

Table 5.11 Continued...

Alteration		Locus	Alteration type	Codon position	Reference
5,811	A – G	MT-TC	TS	non-coding	Houshmand <i>et al.</i> , 1994
5,814	T – C	MT-TC	TS	non-coding	Larsson and Clayton, 1995

Nucleotides are represented by the capital letters, with the rCRS nucleotide indicated first, followed by the observed change. Nucleotide numbers are based on rCRS numbering (Andrews *et al.*, 1999). The numbers indicated in **red** represent previously reported polymorphisms. The number indicated in **blue** represents a previously unreported polymorphism. MT-TI = tRNA isoleucine; MT-TQ = tRNA glutamine; MT-TW = tRNA tryptophan; MT-TC = tRNA cysteine; A = adenine; C = cytosine; G = guanine; T = thymine; TS = transition.

All of the alterations observed for the tRNA genes were transitions with the majority of the alterations being previously reported. A single alteration, A5515G, observed in a single Khoi-San individual (KS_CGR_044) belonging to the L0b sub-haplogroup (refer to Appendix E), was present in the tRNA^{Trp} gene. In order to determine the functional importance of this tRNA gene alteration, it is necessary to perform functional translation studies which were beyond the scope of the current investigation. The secondary structure of tRNA^{Trp} was however determined through the use of the mfold web server version 3.2 (Mathews *et al.*, 1999; Zuker, 2003) and the predicted structure of the tRNA^{Trp} containing the A5515G alteration is depicted in Figure 5.27.

Figure 5.27 Proposed secondary structure and sequence alignment of tRNA^{Trp} containing the A5515G alteration

The proposed secondary structure of the tRNA^{Trp} was determined using the mfold web server version 3.2 (Mathews *et al.*, 1999; Zuker, 2003) and is depicted on the left. Hydrogen bonds between cytosine and guanine nucleotides are indicated in **dark red**. A sequence alignment based on tRNA^{Trp} sequences obtained from GenBank® (accession numbers indicated below), constructed as described in Section 4.12.1, is presented on the right to illustrate the evolutionary conservation of the region of the tRNA^{Trp} containing the A5515G alteration. In the figure the nucleotide position of the A5515G alteration is indicated by a block (■) and in the sequence alignment the A5515G alteration is indicated in **bold** text, while the corresponding positions in the different sequences are indicated by a box (□). Human = *Homo sapiens* (J01415, GI:113200490); fruit fly = *Drosophila melanogaster* (NC_001709, GI:5835233); chicken = *Gallus gallus* (NC_001323, GI:5834843); frog = *Xenopus laevis* (NC_001573, GI:5834981); gorilla = *Gorilla gorilla* (NC_001645, GI:5835149); chimpanzee = *Pan paniscus* (NC_001644, GI:5835135); bovine = *Bos taurus* (NC_006853, GI:60101824); mouse = *Mus musculus* (NC_005089, GI:34538597); rat = *Rattus norvegicus* (NC_001665, GI:110189714); yeast = *Saccharomyces cerevisiae* (NC_001224, GI:6226515). A = adenine; C = cytosine; G = guanine; T = thymine; U = uracil.

The A5515G alteration is located in the amino acid stem of the tRNA^{Trp}. This nucleotide at this position is involved in base pairing with another nucleotide within the amino acid stem, and it is possible that a change in this nucleotide could disrupt formation of the amino acid stem. The alteration however occurs within an internal section of the amino acid stem, and it is possible that the stem formation is not interrupted due to the fact that the base pairing of nucleotides located at the ends of the stem remains intact. Furthermore, the alteration interrupts an A-T pairing which is not as strong as a C-G pairing (two hydrogen bonds vs. three hydrogen bonds), suggesting that this position within the stem is not as important for stem stability as position 5,513 which contains a C-G pairing. In addition, this alteration was observed in an apparently healthy individual. Therefore, it is not expected that this alteration will have a detrimental functional effect.

This deduction is further supported by a comparison of the tRNA^{Trp} gene in ten different organisms. In the sequence alignment presented in Figure 5.27, it can be observed that either an adenine or a guanine is present at position 5,515 within all the different organisms investigated. It appears that the nucleotide at this position is highly conserved in mammals (human, gorilla, chimpanzee, bovine, mouse and rat), whereas a guanine is observed in the other species (fruit fly, frog and yeast). This suggests that the presence of a guanine at this position can be tolerated and that correct folding of the tRNA^{Trp} molecule remains possible in the presence of the A5515G alteration.

5.7.3.2 NADH dehydrogenase subunit 2

The NADH dehydrogenase subunit 2 region consists of bp 4,470 to bp 5,511. The observed alterations are indicated in Table 5.12.

Table 5.12 Sequence alterations observed between the complete Khoi-San mitochondrial DNA genome sequences and the rCRS in the ND2 gene

Alteration		Locus	Alteration type	Codon position	Amino acid change	Reference
4,541	G – A	MT-ND2	TS	coding: 3	SYN	Ingman <i>et al.</i> , 2000
4,586	T – C	MT-ND2	TS	coding: 3	SYN	Ingman <i>et al.</i> , 2000
4,688	T – C	MT-ND2	TS	coding: 3	SYN	Ingman <i>et al.</i> , 2000
4,767	A – G	MT-ND2	TS	coding: 1	NON-SYN: Met → Val	Herrnstadt <i>et al.</i> , 2002a
4,769	A – G	MT-ND2	TS	coding: 3	SYN	Cann <i>et al.</i> , 1987
4,771	C – T	MT-ND2	TS	coding: 2	NON-SYN: Ala → Val	Current investigation
4,907	T – C	MT-ND2	TS	coding: 3	SYN	Ingman <i>et al.</i> , 2000

Table 5.12 Continued...

Alteration		Locus	Alteration type	Codon position	Amino acid change	Reference
5,027	C – T	MT-ND2	TS	coding: 3	SYN	Herrnstadt <i>et al.</i> , 2002a
5,147	G – A	MT-ND2	TS	coding: 3	SYN	Kobayashi <i>et al.</i> , 1991
5,331	C – A	MT-ND2	TV	coding: 1	NON-SYN: Leu → Ile	Herrnstadt <i>et al.</i> , 2002a
5,442	T – C	MT-ND2	TS	coding: 1	NON-SYN: Phe → Leu	Ozawa <i>et al.</i> , 1991a
5,460	G – A	MT-ND2	TS	coding: 1	NON-SYN: Ala → Thr	Lin <i>et al.</i> , 1992
5,471	G – A	MT-ND2	TS	coding: 3	SYN	Howell <i>et al.</i> , 1995

Nucleotides are represented by the capital letters, with the rCRS nucleotide indicated first, followed by the observed change. Nucleotide numbers are based on rCRS numbering (Andrews *et al.*, 1999). The numbers indicated in **red** represent previously reported polymorphisms. The number indicated in **blue** represents a previously unreported polymorphism. MT-ND2 = NADH dehydrogenase subunit 2; A = adenine; C = cytosine; G = guanine; T = thymine; TS = transition; TV = transversion; SYN = synonymous substitution; NON-SYN = non-synonymous substitution; Ala = alanine; Ile = isoleucine; Leu = leucine; Met = methionine; Phe = phenylalanine; Val = valine; Thr = threonine.

The majority of the alterations observed in the ND2 gene in the current investigation have previously been reported. A single previously unreported alteration, C4771T, was observed in a single Khoi-San individual (KS_CGR_120) belonging to the L0a sub-haplogroup (refer to Appendix E).

The alteration at nucleotide position 4,771 is a transition that effects a non-synonymous amino acid change whereby a non-polar alanine changes to a non-polar valine (Campbell, 1995). Due to the fact that the biochemical properties of these two amino acids are similar, it is not expected that this alteration will be deleterious (refer to Section 5.11.5.2). The remainder of the alterations observed within the ND2 gene include non-synonymous changes at nucleotide positions 4,767, 5,331, 5,442 and 5,460.

A4767G causes a change from a non-polar methionine to a non-polar valine, and C5331A leads to a change from a non-polar leucine to a non-polar isoleucine (Campbell, 1995). Both these alterations involve the exchange of biochemically similar amino acids, and both occur in all three Khoi-San individuals belonging to the L2 haplogroup (refer to Section 5.8). These observations suggest that the two alterations are not likely to be deleterious, and could possibly be haplogroup-specific for the cohort studied in the current investigation.

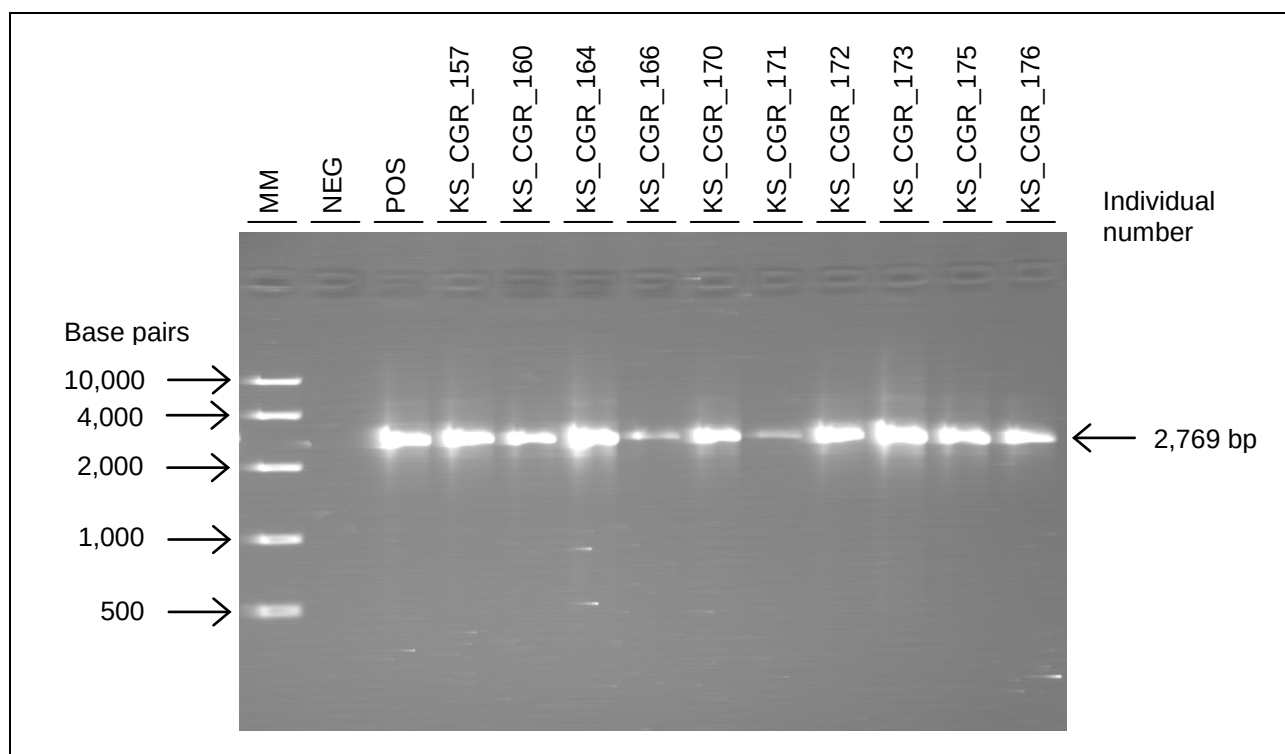
T5442C leads to an amino acid change from a non-polar phenylalanine to a non-polar leucine (Campbell, 1995). These amino acids share similar biochemical properties, and the alteration is observed in all the Khoi-San individuals belonging to the L0a and

L0b sub-haplogroups (refer to Appendix E). Based on these two facts, it can be suggested that this alteration is unlikely to be deleterious. Furthermore, based on the current investigation, it can be speculated that this alteration is possibly specific to these two sub-haplogroups within the Southern African Khoi-San population.

G5460A gives rise to an amino acid change from a non-polar alanine to a polar threonine (Campbell, 1995). The biochemical properties of these two amino acids are different which suggests a potential functional effect for the alteration. However, due to the fact that this alteration is observed in two apparently healthy Khoi-San individuals (KS_CGR_048 and KS_CGR_060), and that no disease has been reported in which this alteration is identified as the causative mutation (MITOMAP, 2008), it is likely that this alteration is an example of a polymorphism due to the normal processes involved in the creation of genetic variation (refer to Section 2.5).

5.7.4 PCR region 4 of the mitochondrial genome

PCR region 4 comprises a total of 2,679 bp and includes base pairs 5,259 to 7,937 based on the numbering system of the rCRS (Andrews *et al.*, 1999). This region contains a number of functional elements which include three tRNA genes, the COI gene, and a few non-coding nucleotides (refer to Section 2.4.2, Table 2.3). PCR was performed according to the conditions listed in Section 4.4, and the PCR products obtained were electrophoresed on an agarose gel as discussed in Section 4.5. A representative gel picture of the amplified products obtained for PCR region 4 is depicted in Figure 5.28.

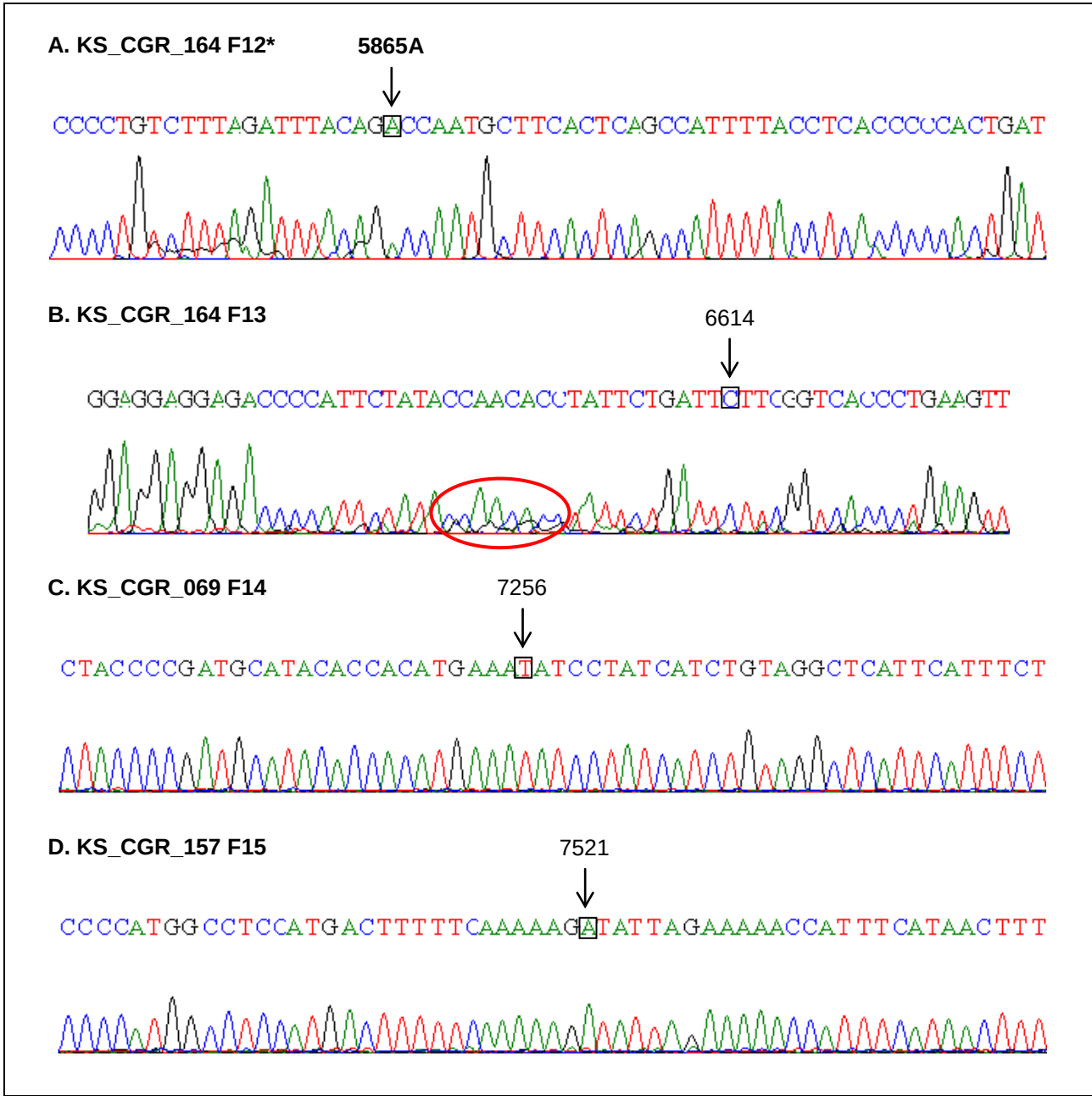
Figure 5.28 Photographic representation of amplified products of PCR region 4

Fragments were electrophoresed on a 0.9% agarose gel at 80 V for 45 min. MM = FastRuler™ DNA Ladder, High Range, ready to use, molecular marker (the sizes of the marker fragments are indicated to the left of the figure); NEG = negative control; POS = positive control; KS_CGR_157 to KS_CGR_176 = sample numbers given to each Khoi-San individual. The amplification efficiency was not equal for all samples and background smears are present in the sample lanes as discussed in Section 5.3.3. Artefacts are present in the gel matrix and the fragments are distorted as discussed in Section 5.4.1.

After gel electrophoresis, the PCR products were purified as discussed in Section 4.4.3. Sequencing of the PCR region was performed as described in Section 4.7. The sequencing results obtained for PCR region 4 were analysed as discussed in Section 4.7.5. Representative electropherograms for each of the four forward primers used to generate sequence data for PCR region 4 are illustrated in Figure 5.29.

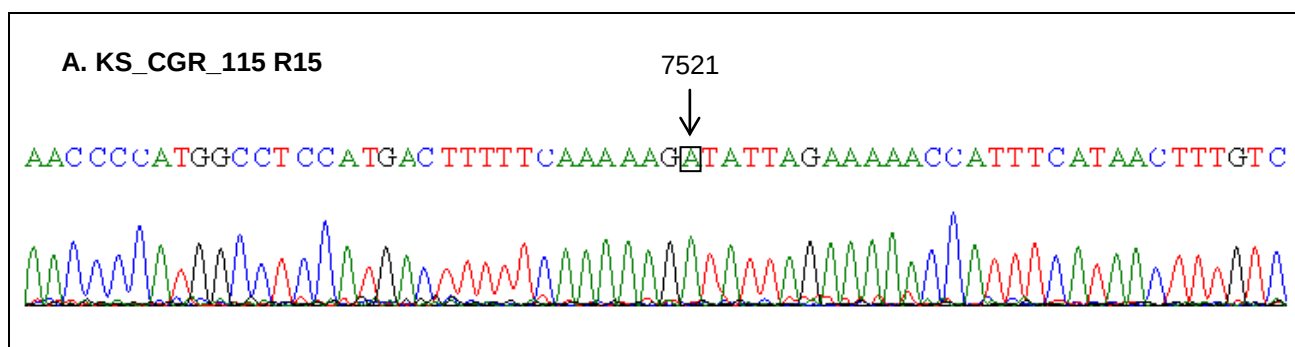
The section of the electropherogram for KS_CGR_164 F13 circled in **red** is affected by a number of artefacts, which include low signal strength, background noise and a minor dye blob. Due to the combination of artefacts present in this region, it appears that an additional G nucleotide could be present between the two C nucleotides observed at the start of the circled region. However, the base-calling performed by the analysis software of the sequencer does not support the insertion of a G nucleotide in this position. Furthermore, the absence of an additional G nucleotide between the two C nucleotides in this sequence region was supported by comparison of the sequence with the rCRS, which indicated that the mitochondrial genome sequence does not contain a G nucleotide within this region.

Figure 5.29 Representative electropherograms of PCR region 4 generated using forward primers F12* to F15



A = adenine; C = cytosine; G = guanine; T = thymine. The first part of the sequence name designates the sample, and the second part of the name indicates the primer used to generate the sequence. The nucleotide positions of alterations are indicated by blocks (□). Alterations are transitions unless specified otherwise. Reported alterations are indicated in regular black text, while novel alterations are indicated in **bold** black text. Minor dye blobs, poor mobility correction, low signal strength and background noise can be observed in the electropherograms as discussed in Section 5.6.4.

A representative electropherogram for the reverse primer used to obtain sequence data for PCR region 4 is depicted in Figure 5.30. As before, the sequence result obtained using the reverse primer is indicated in the reverse complement (refer to Section 5.7.1).

Figure 5.30 Representative electropherogram of PCR region 4 generated using the reverse primer R15

A = adenine; C = cytosine; G = guanine; T = thymine. The first part of the sequence name designates the sample, and the second part of the name indicates the primer used to generate the sequence. The sequence result is presented as the reverse complement. The nucleotide position of an alteration is indicated by a block (□). The observed alteration, a transition, has been reported and is indicated in regular black text. Low signal strength and background noise can be observed in the electropherogram as discussed in Section 5.6.4.

A comparison between sequence data for PCR region 4, and the section of the rCRS that represents this PCR region (refer to Section 4.7.5), revealed a total of 21 discrepancies. The 21 discrepancies include 19 transitions and two transversions. Of the 21 discrepancies, two transversions and one transition are previously unreported polymorphisms. The sequence alterations present in all screened Khoi-San individuals are presented in Sections 5.7.4.1 to 5.7.4.3. PCR region 4 has been divided into functional units to facilitate discussion of the observed alterations.

5.7.4.1 tRNA tyrosine

The tRNA tyrosine region consists of bp 5,826 to bp 5,891. The observed alteration is indicated in Table 5.13.

Table 5.13 Sequence alteration observed between the complete Khoi-San mitochondrial DNA genome sequences and the rCRS in the tRNA^{Tyr} gene

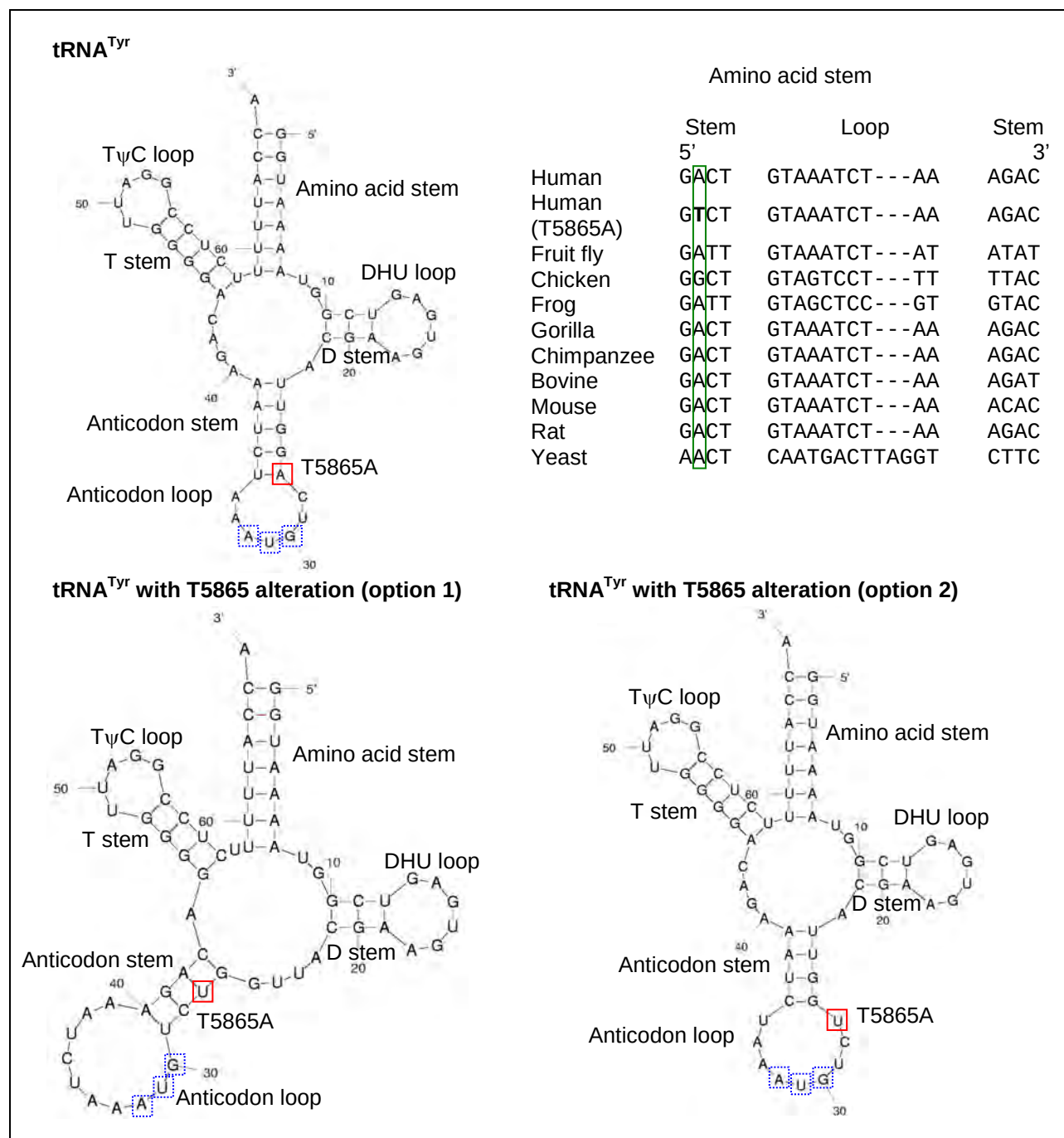
Alteration	Locus	Alteration type	Codon position	Reference
5,865 T – A	MT-TY	TV	non-coding	Current investigation

Nucleotides are represented by the capital letters, with the rCRS nucleotide indicated first, followed by the observed change. Nucleotide numbers are based on rCRS numbering (Andrews *et al.*, 1999). The number indicated in blue represents a novel polymorphism. MT-TY = tRNA tyrosine; A = adenine; T = thymine; TV = transversion.

A single novel alteration, T5865A, was observed in this region in one Khoi-San individual (KS_CGR_164) belonging to the L2b sub-haplogroup (refer to Appendix E). Functional translation studies will have to be performed in order to determine the functional importance of this tRNA gene alteration. However, such studies did not form part of the current investigation and the secondary structure of the tRNA^{Tyr} was instead analysed

using the mfold web server version 3.2 (Mathews *et al.*, 1999; Zuker, 2003). The predicted structures of the tRNA^{Tyr} containing the T5865A alteration are presented in Figure 5.31.

Figure 5.31 Proposed secondary structures and sequence alignment of tRNA^{Tyr} containing the T5865A alteration



The proposed secondary structures of the normal tRNA^{Tyr} and the altered tRNA^{Tyr} molecules as determined using the mfold web server version 3.2 (Mathews *et al.*, 1999; Zuker, 2003) are depicted in the top half (left hand corner) and bottom half of the figure. Hydrogen bonds between cytosine and guanine nucleotides are indicated in dark red. A sequence alignment based on tRNA^{Tyr} sequences obtained from GenBank[®] (accession numbers indicated below), constructed as described in Section 4.12.1, is presented in the top half (right hand corner) to illustrate the evolutionary conservation of the region of the tRNA^{Tyr} containing the T5865A alteration. In the tRNA^{Tyr} figures the nucleotide position of the T5865A alteration is indicated by a block (■), and the three nucleotides that form part of the anticodon is indicated by dotted blocks (⋯). In the sequence alignment the T5865A alteration is indicated in bold text, while the corresponding positions in the different sequences are indicated by a box (□). The sequence data presented in both the figures and the alignment is the reverse complement, as the tRNA^{Tyr} gene is transcribed from the L-strand (refer to Section 2.4.4, Table 2.5). Human = *Homo sapiens* (J01415, GI:113200490); fruit fly = *Drosophila melanogaster* (NC_001709, GI:5835233); chicken = *Gallus gallus* (NC_001323, GI:5834843); frog = *Xenopus laevis* (NC_001573, GI:5834981); gorilla = *Gorilla gorilla* (NC_001645, GI:5835149); chimpanzee = *Pan paniscus* (NC_001644, GI:5835135); bovine = *Bos taurus* (NC_006853, GI:60101824); mouse = *Mus musculus* (NC_005089, GI:34538597); rat = *Rattus norvegicus* (NC_001665, GI:110189714); yeast = *Saccharomyces cerevisiae* (NC_001224, GI:6226515). A = adenine; C = cytosine; G = guanine; T = thymine; U = uracil.

Nucleotide position 5,865 is located in the anticodon stem of the secondary structure of the tRNA^{Tyr} where it is involved in base pairing. This position appears to be highly conserved across different species with a guanine reported at this position in only one species, i.e. the chicken. Two proposed secondary structures were generated for the tRNA^{Tyr} with the T5865A alteration.

In option 1, the T5865A alteration changes the secondary structure of the tRNA molecule by introducing a new base pairing. In the unaltered tRNA^{Tyr} the adenine at position 27 pairs with the uracil at position 35, while in the altered tRNA^{Tyr} the introduced uracil at position 27 pairs with an adenine in the anticodon stem at position 42. The alteration shifts the position of the start of the anticodon stem from the uracil at position 23, pairing with the adenine at position 39, to the guanine at position 26, pairing with the cytosine at position 43. This reorganisation of base pairing leads to reductions in the sizes of the anticodon stem and the adjacent T stem region (five base pairs vs. four base pairs). Although the alteration therefore does not disrupt base pairing, and instead introduces new base pairing in a different region of the molecule, it remains likely that the alteration could have a detrimental effect on the function of the tRNA^{Tyr} molecule.

The tRNA^{Tyr} molecule has two functions, to bind to tyrosine and to recognise the codon for tyrosine in a messenger RNA (mRNA) molecule so that an aminoacyl-tRNA synthetase can add the amino acid to the growing polypeptide chain (Lodish *et al.*, 1995). The anticodon for tyrosine, 5'-GUA-3' is located in the anticodon loop. In option 1, the T5865A alteration increases the size of the anticodon loop from seven nucleotides to ten nucleotides, the GUA sequence is however still present within the anticodon loop, although at a different position relative to its original position. It is possible that binding of the anticodon to the codon within the mRNA will be impeded by this shift. Since binding of the anticodon to codons is however still possible, it is suggested that this alteration will not be pathogenic. For option 1, an impact on the efficiency of protein translation in the presence of the T5865A alteration can however not be excluded in the absence of functional translation studies.

In option 2, the T5865A alteration changes the secondary structure of the tRNA molecule through the loss of a single base pairing in the anticodon loop between the adenine at position 27 and the uracil at position 35. This leads to a decrease in the size of the anticodon stem from five base pairings to four base pairings, and a concomitant increase in the anticodon loop size from seven nucleotides to nine nucleotides. The position of the

three nucleotides comprising the anticodon does not shift within the anticodon loop relative to the position in the unaltered tRNA^{Tyr}. It is thus expected that binding to the codon in the mRNA molecule will not be hindered and therefore a detrimental functional effect is not anticipated if option 2 accurately depicts the RNA folding pattern in the presence of the T5865A alteration.

The T5865A alteration occurred in an apparently healthy individual, and binding of codons to the anticodon loop remains possible regardless of which of the two options describe the correct folding of the tRNA^{Tyr} molecule. Although the position was highly conserved in the comparison of ten organisms, due to the observations made regarding RNA folding options 1 and 2, it is suggested that this alteration will not be pathogenic.

The T5865A alteration was observed in a single individual within the cohort and for this reason the alteration is viewed as a “personal polymorphism” within the cohort investigated, i.e. a variant that is useful to identify the specific individual. This alteration that presently appears unique to this individual may however not be so exclusive when more sequence data is generated for the tRNA^{Tyr} gene in other studies of individuals of African descent including more Khoi-San individuals.

5.7.4.2 Cytochrome c oxidase subunit I

The cytochrome c oxidase subunit I region consists of bp 5,904 to bp 7,445. The observed alterations are indicated in Table 5.14.

Table 5.14 Sequence alterations observed between the complete Khoi-San mitochondrial DNA genome sequences and the rCRS in the COI gene

Alteration		Locus	Alteration type	Codon position	Amino acid change	Reference
5,986	T – C	MT-CO1	TS	coding: 2	NON-SYN: Val → Ala	Current investigation
6,026	G – A	MT-CO1	TS	coding: 3	SYN	Ingman <i>et al.</i> , 2000
6,185	T – C	MT-CO1	TS	coding: 3	SYN	Herrnstadt <i>et al.</i> , 2002a
6,266	A – G	MT-CO1	TS	coding: 3	SYN	Herrnstadt <i>et al.</i> , 2002a
6,344	C – G	MT-CO1	TV	coding: 3	NON-SYN: Ile → Met	Current investigation
6,614	T – C	MT-CO1	TS	coding: 3	SYN	Tanaka <i>et al.</i> , 2004
6,629	A – G	MT-CO1	TS	coding: 3	SYN	Herrnstadt <i>et al.</i> , 2002a
6,692	A – G	MT-CO1	TS	coding: 3	SYN	Kivisild <i>et al.</i> , 2006
6,713	C – T	MT-CO1	TS	coding: 3	SYN	Herrnstadt <i>et al.</i> , 2002a
6,806	A – G	MT-CO1	TS	coding: 3	SYN	Rajkumar <i>et al.</i> , 2005

Table 5.14 Continued...

Alteration		Locus	Alteration type	Codon position	Amino acid change	Reference
6,815	T – C	MT-CO1	TS	coding: 3	SYN	Mishmar <i>et al.</i> , 2003
6,938	C – T	MT-CO1	TS	coding: 3	SYN	Finnilä <i>et al.</i> , 2001b
7,028	C – T	MT-CO1	TS	coding: 3	SYN	Cann <i>et al.</i> , 1987
7,146	A – G	MT-CO1	TS	coding: 1	NON-SYN: Thr → Ala	Herrnstadt <i>et al.</i> , 2002a
7,154	A – G	MT-CO1	TS	coding: 3	SYN	Mishmar <i>et al.</i> , 2003
7,256	C – T	MT-CO1	TS	coding: 3	SYN	Silva <i>et al.</i> , 2002
7,257	A – G	MT-CO1	TS	coding: 1	NON-SYN: Ile → Val	Ingman <i>et al.</i> , 2000
7,283	T – C	MT-CO1	TS	coding: 3	SYN	Moilanen <i>et al.</i> , 2003
7,424	A – G	MT-CO1	TS	coding: 3	SYN	Marzuki <i>et al.</i> , 1991

Nucleotides are represented by the capital letters, with the rCRS nucleotide indicated first, followed by the observed change. Nucleotide numbers are based on rCRS numbering (Andrews *et al.*, 1999). The numbers indicated in **red** represent previously reported polymorphisms. The numbers indicated in **blue** represent previously unreported polymorphisms. A = adenine; C = cytosine; G = guanine; T = thymine; MT-CO1 = cytochrome c oxidase subunit I; TS = transition; TV = transversion; SYN = synonymous substitution; NON-SYN = non-synonymous substitution; Ala = alanine; Ile = isoleucine; Met = methionine; Val = valine; Thr = threonine.

The majority of the alterations observed for the COI gene in the current investigation are transitions and have been reported previously. Two previously unreported alterations, T5986C and C6344G, were observed in a single Khoi-San individual (KS_CGR_164) belonging to the L2b sub-haplogroup (refer to Appendix E).

The T5986C transition leads to a non-synonymous change whereby a non-polar valine changes to a non-polar alanine, while the C6344G transversion gives rise to a non-synonymous amino acid change involving the replacement of a non-polar isoleucine with a non-polar methionine (Campbell, 1995). As the biochemical properties of the two amino acids are similar in both the non-synonymous changes effected by the two alterations, it is not expected that the alterations will have a deleterious effect.

The previously reported alterations include two non-synonymous transitions at nucleotide positions 7,146 and 7,257. The A7146G alteration gives rise to an amino acid change from a polar threonine to a non-polar alanine (Campbell, 1995). Due to the fact that the biochemical properties of the two amino acids are different, it can be suggested that this alteration should have a functional effect. However, this alteration is present in the majority of Khoi-San individuals studied in the current investigation (refer to Appendix A), and this alteration is also used to exclude haplogroups L0 and L1 from L2, L3 and M (refer to Appendix D). Therefore, it is not expected to have a detrimental effect, if it is so commonly observed in haplogroups L0 and L1. Instead, it may be an example of an adaptive alteration that was retained in African populations since it provided some advantage to

individuals carrying these haplogroups in specific environments (Wallace, 1994; Wallace, 2005). Functional studies will however be required to investigate the biochemical effect of this alteration on the L0 and L1 haplogroups.

The A7257G alteration leads to the replacement of a non-polar isoleucine with a non-polar valine (Campbell, 1995). The biochemical characteristics of these two amino acids are similar, so it is possible that the alteration will not have a functional effect.

5.7.4.3 tRNA serine (UCN)

The tRNA serine (UCN) region consists of bp 7,446 to bp 7,516. The observed alteration is indicated in Table 5.15.

Table 5.15 Sequence alteration observed between the complete Khoi-San mitochondrial DNA genome sequences and the rCRS in the tRNA^{Ser(UCN)} gene

Alteration		Locus	Alteration type	Codon position	Reference
7,521	G – A	MT-TS1	TS	non-coding	Prezant <i>et al.</i> , 1993

Nucleotides are represented by the capital letters, with the rCRS nucleotide indicated first, followed by the observed change. Nucleotide numbers are based on rCRS numbering (Andrews *et al.*, 1999). The number indicated in red represents a previously reported polymorphism. A = adenine; G = guanine; MT-TS1 = tRNA serine (UCN); TS = transition.

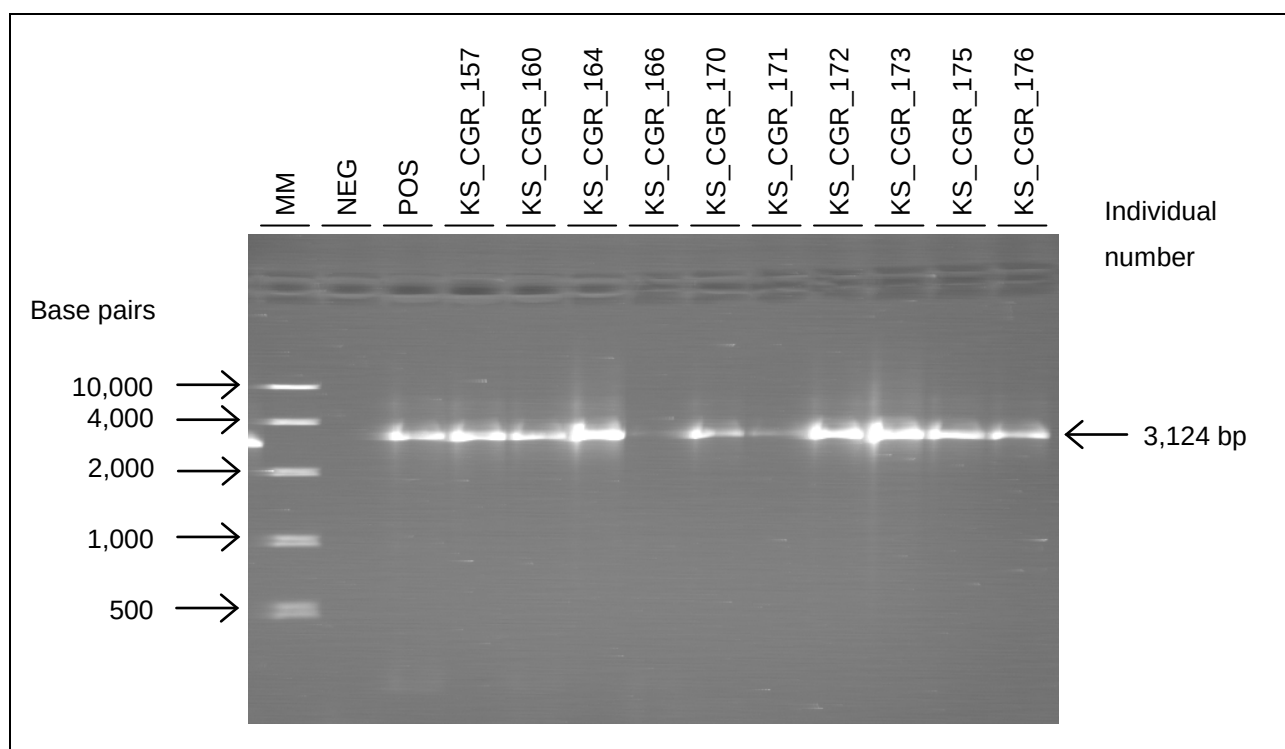
A single alteration was observed in the tRNA^{Ser(UCN)} gene in all except four Khoi-San individuals studied in the current investigation (refer to Appendix A). This alteration has been reported previously and is part of the criteria used to exclude L0, L1 and L2 haplogroups from other haplogroups (refer to Appendix D).

5.7.5 PCR region 5 of the mitochondrial genome

PCR region 5 consists of a total of 3,124 bp and contains base pairs 7,358 to 10,481 based on the numbering system of the rCRS (Andrews *et al.*, 1999). This region contains a number of functional elements which include five tRNA genes, two CO genes, both ATPase genes, ND3 and a few non-coding nucleotides (refer to Section 2.4.2, Table 2.3).

PCR was performed for PCR region 5 according to the conditions described in Section 4.4, and the PCR products generated were electrophoresed on an agarose gel as discussed in Section 4.5. A representative gel picture of the amplified products obtained for PCR region 5 is presented in Figure 5.32.

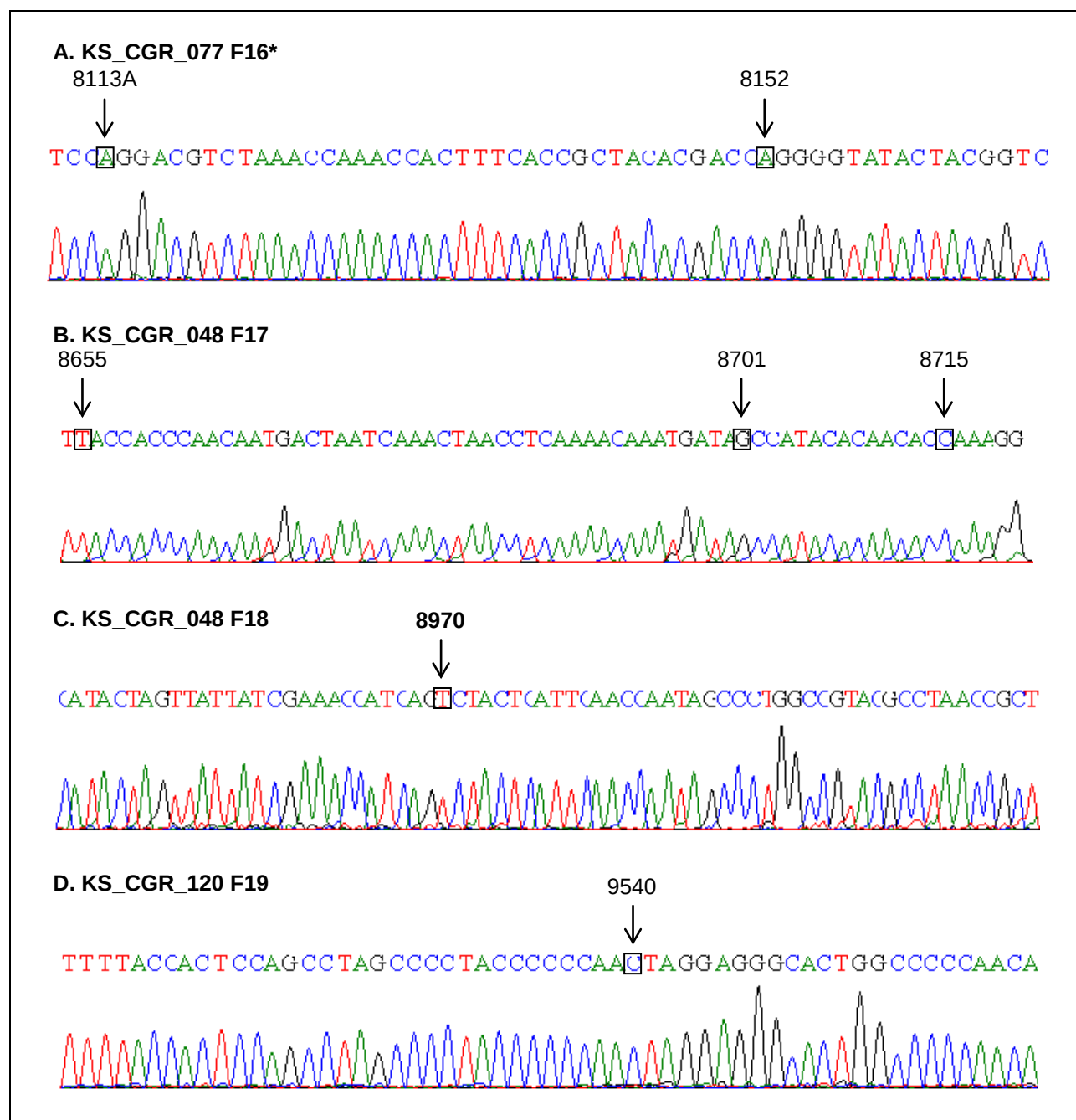
Figure 5.32 Photographic representation of amplified products of PCR region 5



Fragments were electrophoresed on a 0.9% agarose gel at 80 V for 45 min. MM = FastRuler™ DNA Ladder, High Range, ready to use, molecular marker (the sizes of the marker fragments are indicated to the left of the figure); NEG = negative control; POS = positive control; KS_CGR_157 to KS_CGR_176 = sample numbers given to each Khoi-San individual. The amplification efficiency was not equal for all samples and background smears are present in the sample lanes as discussed in Section 5.3.3. Artefacts are present in the gel matrix and the fragments are distorted as discussed in Section 5.4.1.

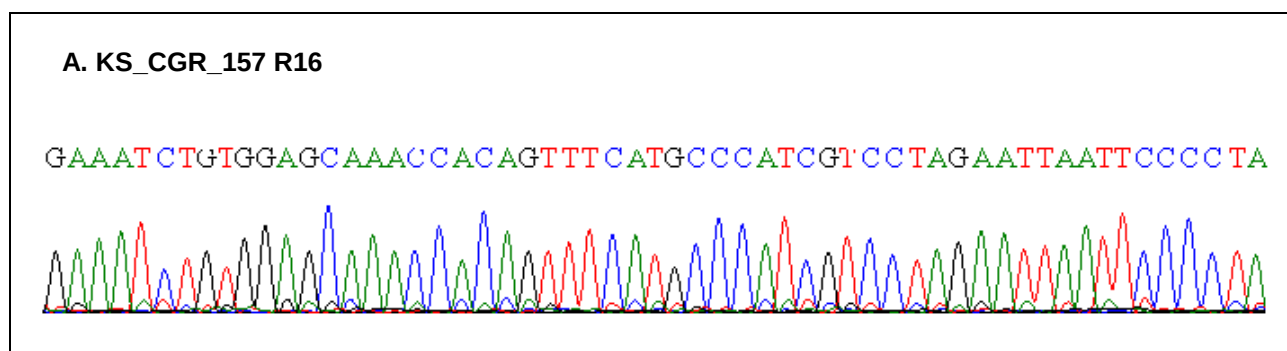
Gel electrophoresis was followed by purification of the PCR products as described in Section 4.4.3. The sequence of PCR region 5 was determined as discussed in Section 4.7 and the results were analysed as described in Section 4.7.5. Representative electropherograms for each of the four forward primers used to determine the sequence of PCR region 5 are presented in Figure 5.33.

Figure 5.33 Representative electropherograms of PCR region 5 generated using forward primers F16* to F19



A = adenine; C = cytosine; G = guanine; T = thymine. The first part of the sequence name designates the sample, and the second part of the name indicates the primer used to generate the sequence. The nucleotide positions of alterations are indicated by blocks (□). Alterations are transitions unless specified otherwise. Reported alterations are indicated in regular black text, and a previously unreported alteration is indicated in **bold** black text. Poor mobility correction, low signal strength and background noise can be observed in the electropherograms as discussed in Section 5.6.4.

A representative electropherogram for the reverse primer used to generate sequence data for PCR region 5 is depicted in Figure 5.34. As discussed previously, the sequence result is presented as the reverse complement (refer to Section 5.7.1).

Figure 5.34 A representative electropherogram of PCR region 5 generated using reverse primer R16

A = adenine; C = cytosine; G = guanine; T = thymine. The first part of the sequence name designates the sample, and the second part of the name indicates the primer used to generate the sequence. Low signal strength and background noise can be observed in the electropherogram as discussed in Section 5.6.4.

Comparison of the sequence data generated for PCR region 5 with the section of the rCRS representing this region (refer to Section 4.7.5), revealed a total of 46 discrepancies. The discrepancies observed included 41 transitions and five transversions. Of the 46 discrepancies, three transversions and two transitions represented previously unreported polymorphisms. The sequence alterations present in all screened Khoi-San individuals are presented in Sections 5.7.5.1 to 5.7.5.5. In order to discuss the observed alterations PCR region 5 has been divided into its functional units.

5.7.5.1 Cytochrome c oxidase subunit II

The cytochrome c oxidase subunit II region consists of bp 7,586 to bp 8,269. The observed alterations are indicated in Table 5.16.

Table 5.16 Sequence alterations observed between the complete Khoi-San mitochondrial DNA genome sequences and the rCRS in the COII gene

Alteration	Locus	Alteration type	Codon position	Amino acid change	Reference
7,624 T – A	MT-CO2	TV	coding: 3	SYN	Herrnstadt <i>et al.</i> , 2002a
7,673 A – G	MT-CO2	TS	coding: 1	NON-SYN: Ile → Val	Ozawa <i>et al.</i> , 1991a
7,828 A – G	MT-CO2	TS	coding: 3	SYN	Ruvolo <i>et al.</i> , 1993
7,961 T – C	MT-CO2	TS	coding: 1	SYN	Tanaka <i>et al.</i> , 2004
7,967 C – T	MT-CO2	TS	coding: 1	SYN	Kivisild <i>et al.</i> , 2006
8,080 C – T	MT-CO2	TS	coding: 3	SYN	Herrnstadt <i>et al.</i> , 2002a
8,113 C – A	MT-CO2	TV	coding: 3	SYN	Ruvolo <i>et al.</i> , 1993
8,152 G – A	MT-CO2	TS	coding: 3	SYN	Ruvolo <i>et al.</i> , 1993
8,206 G – A	MT-CO2	TS	coding: 3	SYN	Ozawa <i>et al.</i> , 1995
8,214 T – A	MT-CO2	TV	coding: 2	NON-SYN: Val → Asp	Current investigation
8,222 T – C	MT-CO2	TS	coding: 1	SYN	Ruppert <i>et al.</i> , 2004

Table 5.16 Continued...

8,223	T – G	MT-CO2	TV	coding: 2	NON-SYN: Leu → Trp	Current investigation
8,251	G – A	MT-CO2	TS	coding: 3	SYN	Ruvolo <i>et al.</i> , 1993

Nucleotides are represented by the capital letters, with the rCRS nucleotide indicated first, followed by the observed change. Nucleotide numbers are based on rCRS numbering (Andrews *et al.*, 1999). The numbers indicated in **red** represent previously reported polymorphisms, and the numbers indicated in **blue** represent previously unreported polymorphisms. A = adenine; C = cytosine; G = guanine; T = thymine; MT-CO2 = cytochrome c oxidase subunit II; TV = transversion; TS = transition; NON-SYN = non-synonymous substitution; SYN = synonymous substitution; Asp = aspartic acid; Ile = isoleucine; Leu = leucine; Trp = tryptophan; Val = valine.

In the current investigation, the majority of the alterations observed for the COII gene were transitions and have been reported previously. Two previously unreported transversions, T8214A and T8223G, were observed in a single Khoi-San individual (KS_CGR_105) belonging to the L0b sub-haplogroup (refer to Appendix E).

The T8214A alteration gives rise to a non-synonymous amino acid change involving the replacement of a non-polar valine with a polar aspartic acid (Campbell, 1995). Due to the fact that the biochemical properties of the two amino acids are different, it is possible that the alteration could have a functional effect. The potential functional effect of this alteration is discussed in Section 5.11.5.2.

The T8223G alteration leads to a non-synonymous amino acid change whereby a non-polar leucine changes to a non-polar tryptophan (Campbell, 1995). The biochemical properties of these two amino acids are similar and therefore it is not likely that the alteration will have a functional effect.

One additional previously reported non-synonymous alteration was observed at nucleotide position 7,673. The A7673G alteration was observed in three Khoi-San individuals (KS_CGR_059, KS_CGR_084 and KS_CGR_120) and all three individuals were classified as members of the L0a sub-haplogroup (refer to Appendix E). This alteration changes a non-polar isoleucine to a non-polar valine. Due to the fact that the biochemical characteristics of the two amino acids are similar, it is not expected that this alteration will have a functional effect (refer to Section 5.11.5.2).

5.7.5.2 ATP synthase F₀ subunit 8

The ATP synthase F₀ subunit 8 region consists of bp 8,366 to bp 8,572. The observed alterations are indicated in Table 5.17.

Table 5.17 Sequence alterations observed between the complete Khoi-San mitochondrial DNA genome sequences and the rCRS in the ATPase8 gene

Alteration		Locus	Alteration type	Codon position	Amino acid change	Reference
8,387	G – A	ATPase8	TS	coding: 1	NON-SYN: Val → Met	Keightley <i>et al.</i> , 1996
8,468	C – T	ATPase8	TS	coding: 1	SYN	Moraes <i>et al.</i> , 1993
8,503	T – C	ATPase8	TS	coding: 3	SYN	Merriwether <i>et al.</i> , 2005

Nucleotides are represented by the capital letters, with the rCRS nucleotide indicated first, followed by the observed change. Nucleotide numbers are based on rCRS numbering (Andrews *et al.*, 1999). The numbers indicated in red represent previously reported polymorphisms. A = adenine; C = cytosine; G = guanine; T = thymine; ATPase8 = ATP synthase F₀ subunit 8; TS = transition; NON-SYN = non-synonymous substitution; SYN = synonymous substitution; Met = methionine; Val = valine.

The alterations observed for the ATPase8 gene in the current investigation have all been previously reported. Two of the transitions observed in the ATPase8 gene result in synonymous changes, while the third causes a non-synonymous change. The G8387A alteration was observed in three Khoi-San individuals (KS_CGR_069, KS_CGR_164 and KS_CGR_173) belonging to the L2b sub-haplogroup (refer to Appendix E). This alteration results in an amino acid change from a non-polar valine to a non-polar methionine (Campbell, 1995). Due to the fact that the biochemical properties of these two amino acids are similar, as well as the observation that the alteration was present in all three Khoi-San individuals belonging to the L2 haplogroup, it is not expected that this alteration would have a functional effect.

5.7.5.3 ATP synthase F₀ subunit 6

The ATP synthase F₀ subunit 6 region consists of bp 8,527 to bp 9,207. The observed alterations are indicated in Table 5.18.

Table 5.18 Sequence alterations observed between the complete Khoi-San mitochondrial DNA genome sequences and the rCRS in the ATPase6 gene

Alteration		Locus	Alteration type	Codon position	Amino acid change	Reference
8,618	T – C	ATPase6	TS	coding: 2	NON-SYN: Ile → Thr	Poetsch <i>et al.</i> , 2003
8,655	C – T	ATPase6	TS	coding: 3	SYN	Ingman <i>et al.</i> , 2000
8,701	A – G	ATPase6	TS	coding: 1	NON-SYN: Thr → Ala	Wallace <i>et al.</i> , 1988
8,715	T – C	ATPase6	TS	coding: 3	SYN	Herrnstadt <i>et al.</i> , 2002a
8,790	G – A	ATPase6	TS	coding: 3	SYN	Taylor <i>et al.</i> , 2003
8,860	A – G	ATPase6	TS	coding: 1	NON-SYN: Thr → Ala	Brown <i>et al.</i> , 1992

Table 5.18 Continued...

Alteration		Locus	Alteration type	Codon position	Amino acid change	Reference
8,865	G – A	ATPase6	TS	coding: 3	SYN	Herrnstadt <i>et al.</i> , 2002a
8,911	T – C	ATPase6	TS	coding: 1	SYN	Ingman <i>et al.</i> , 2000
8,922	C – T	ATPase6	TS	coding: 3	SYN	Current investigation
8,970	C – T	ATPase6	TS	coding: 3	SYN	Current investigation
8,994	G – A	ATPase6	TS	coding: 3	SYN	Ingman <i>et al.</i> , 2000
9,042	C – T	ATPase6	TS	coding: 3	SYN	Ingman <i>et al.</i> , 2000
9,111	T – C	ATPase6	TS	coding: 3	SYN	Herrnstadt <i>et al.</i> , 2002a
9,136	A – G	ATPase6	TS	coding: 1	NON-SYN: Ile → Val	Ingman <i>et al.</i> , 2000
9,150	A – G	ATPase6	TS	coding: 3	SYN	Rieder <i>et al.</i> , 1998

Nucleotides are represented by the capital letters, with the rCRS nucleotide indicated first, followed by the observed change. Nucleotide numbers are based on rCRS numbering (Andrews *et al.*, 1999). The numbers indicated in **red** represent previously reported polymorphisms and the numbers indicated in **blue** represent novel polymorphisms. A = adenine; C = cytosine; G = guanine; T = thymine; ATPase6 = ATP synthase F₀ subunit 6; TS = transition; NON-SYN = non-synonymous substitution; SYN = synonymous substitution; Ala = alanine; Ile = isoleucine; Val = valine; Thr = threonine.

All of the alterations observed in the ATPase6 gene are transitions and the majority of the alterations have been reported previously. Two transitions at nucleotide positions 8,922 and 8,970 are previously unreported alterations.

The C8922T alteration was observed in a single Khoi-San individual (KS_CGR_001) belonging to the L0a sub-haplogroup (refer to Appendix E). This alteration does not lead to an amino acid change and therefore it is expected that this alteration will have no functional consequences. The presence of this alteration in a single individual belonging to the L0a sub-haplogroup, appears to indicate that this alteration is another example of a “personal polymorphism” that will be useful in identifying this specific Khoi-San individual within the group of 30 Khoi-San individuals studied in the current investigation (refer to Section 5.7.4.1).

The C8970T alteration was observed in two Khoi-San individuals (KS_CGR_048 and KS_CGR_060) belonging to the L0b sub-haplogroup (refer to Appendix E). This alteration gives rise to a synonymous change and therefore no functional effect is expected. However, within the L0b sub-haplogroup, these two individuals belong to the L0bB1b1 group (refer to Section 5.10.2.1, Figure 5.51) and it is possible that this alteration could be specific to this group. Sequence data from more Khoi-San individuals belonging to the L0bB1 group is however required to confirm this.

A number of previously reported non-synonymous alterations were observed in the ATPase6 gene. The T8618C alteration results in the replacement of a non-polar isoleucine with a polar threonine (Campbell, 1995). This alteration is present in two Khoi-San individuals (KS_CGR_090 and KS_CGR_072) classified as members of the L3b sub-haplogroup (refer to Appendix E). Although the alteration allows for the replacement of an amino acid with another amino acid with different biochemical properties, the presence of this alteration in two apparently healthy individuals, both belonging to the same sub-haplogroup, seems to suggest that the alteration will not have a detrimental functional effect. This is further supported by the fact that to date this alteration has not been identified as a causative mutation in any disease (MITOMAP, 2008).

The A8701G and the A8860G alterations both give rise to an amino acid change from a polar threonine to a non-polar alanine (Campbell, 1995). These alterations were observed in all the Khoi-San individuals studied in the current investigation (refer to Appendix A). These alterations are part of the exclusion criteria that are used to identify L-specific mitochondrial genome sequences (refer to Appendix D). Thus, although the alterations bring about an amino acid change, where the replacement amino acid has different biochemical characteristics compared to the original amino acid, the presence of these alterations in all the Khoi-San individuals studied in the current investigation and the use of these alterations as part of the exclusion criteria in the classification scheme used, appear to indicate that the alterations will not have a deleterious functional effect. This is further supported by the fact that the two alterations have not been identified as causative mutations in any disorders (MITOMAP, 2008).

The A9136G alteration led to an amino acid change from a non-polar isoleucine to a non-polar valine (Campbell, 1995). This alteration was detected in seven Khoi-San individuals (refer to Appendix A) belonging to the L0a sub-haplogroup (refer to Appendix E). This alteration involves the replacement of an amino acid with a biochemically similar amino acid and was observed in seven apparently healthy Khoi-San individuals belonging to the same sub-haplogroup. For these two reasons it is not anticipated that this alteration will have a detrimental functional effect.

5.7.5.4 Cytochrome c oxidase subunit III

The cytochrome c oxidase subunit III region consists of bp 9,207 to bp 9,990. The observed alterations are indicated in Table 5.19.

Table 5.19 Sequence alterations observed between the complete Khoi-San mitochondrial DNA genome sequences and the rCRS in the COIII gene

Alteration		Locus	Alteration type	Codon position	Amino acid change	Reference
9,221	A – G	MT-CO3	TS	coding: 3	SYN	Ingman <i>et al.</i> , 2000
9,347	A – G	MT-CO3	TS	coding: 3	SYN	Ingman <i>et al.</i> , 2000
9,350	A – G	MT-CO3	TS	coding: 3	SYN	Kivisild <i>et al.</i> , 2006
9,438	G – A	MT-CO3	TS	coding: 1	NON-SYN: Gly → Ser	Johns and Neufeld, 1993
9,540	T – C	MT-CO3	TS	coding: 1	SYN	Wallace <i>et al.</i> , 1988
9,755	G – A	MT-CO3	TS	coding: 3	SYN	Ingman <i>et al.</i> , 2000
9,818	C – T	MT-CO3	TS	coding: 3	SYN	Ingman <i>et al.</i> , 2000
9,887	T – C	MT-CO3	TS	coding: 3	SYN	Kivisild <i>et al.</i> , 2006
9,932	G – A	MT-CO3	TS	coding: 3	SYN	Ozawa <i>et al.</i> , 1991a
9,950	T – C	MT-CO3	TS	coding: 3	SYN	Tanaka and Ozawa, 1994

Nucleotides are represented by the capital letters, with the rCRS nucleotide indicated first, followed by the observed change. Nucleotide numbers are based on rCRS numbering (Andrews *et al.*, 1999). The numbers indicated in **red** represent previously reported polymorphisms. A = adenine; C = cytosine; G = guanine; T = thymine; MT-CO3 = cytochrome *c* oxidase subunit III; TS = transition; NON-SYN = non-synonymous substitution; SYN = synonymous substitution; Gly = glycine; Ser = serine.

The transitions observed in the COIII gene have all been previously reported. All of the transitions involved synonymous changes with the exception of the transition observed at nucleotide position 9,438 which gave rise to a non-synonymous change. The G9438A alteration led to an amino change from a polar glycine to a polar serine (Campbell, 1995). These two amino acids have similar biochemical characteristics and therefore a functional effect is not expected for this alteration. The G9438A alteration was observed in eight Khoi-San individuals (refer to Appendix A) belonging to the L0b sub-haplogroup (refer to Appendix E). This alteration appeared to be characteristic of the L0bB2a group within this sub-haplogroup (refer to Section 5.10.2.1, Figure 5.51) and was located at a branching point in the global L0-specific haplogroup network (refer to Section 5.10) where L0bB2a was divided into three groups. This alteration is further discussed in Section 5.11.5.2.

5.7.5.5 NADH dehydrogenase subunit 3

The NADH dehydrogenase subunit 3 region consists of bp 10,059 to bp 10,404. The observed alterations are indicated in Table 5.20.

Table 5.20 Sequence alterations observed between the complete Khoi-San mitochondrial DNA genome sequences and the rCRS in the ND3 gene

Alteration		Locus	Alteration type	Codon position	Amino acid change	Reference
10,084	T – C	MT-ND3	TS	coding: 2	NON-SYN: Ile → Thr	Clark <i>et al.</i> , 1999
10,098	G – T	MT-ND3	TV	coding: 1	NON-SYN: Ala → Ser	Current investigation
10,115	T – C	MT-ND3	TS	coding: 3	SYN	Kösel <i>et al.</i> , 1998
10,398	A – G	MT-ND3	TS	coding: 1	NON-SYN: Thr → Ala	Cann <i>et al.</i> , 1987

Nucleotides are represented by the capital letters, with the rCRS nucleotide indicated first, followed by the observed change. Nucleotide numbers are based on rCRS numbering (Andrews *et al.*, 1999). The numbers indicated in **red** represent previously reported polymorphisms and the number indicated in **blue** represents a novel polymorphism. A = adenine; C = cytosine; G = guanine; T = thymine; MT-ND3 = NADH dehydrogenase subunit 3; TS = transition; TV = transversion; NON-SYN = non-synonymous substitution; SYN = synonymous substitution; Ala = alanine; Ile = isoleucine; Ser = serine; Thr = threonine.

Four alterations were observed for the ND3 gene within the Khoi-San individuals studied in the current investigation. Of the four alterations, three transitions were previously reported and one transversion was previously unreported. This transversion was observed at nucleotide position 10,098 and gave rise to a non-synonymous change whereby a non-polar alanine was replaced with a polar serine (Campbell, 1995). The two amino acids have different biochemical properties and it can therefore be expected that the G10098T alteration should have a functional effect. However, this alteration was detected in three apparently healthy Khoi-San individuals (KS_CGR_140, KS_CGR_141 and KS_CGR_175), all belonging to the L3a1 sub-haplogroup (refer to Appendix E). To date this alteration has not been identified as the causative mutation in any mitochondrial disease (MITOMAP, 2008). It thus appears more likely that this alteration does not have a detrimental functional effect and it is possible that it is specific to the L3a1 sub-haplogroup as observed in the Southern African Khoi-San population. More sequence data from Khoi-San individuals belonging to the L3a1 sub-haplogroup is however required to confirm this observation.

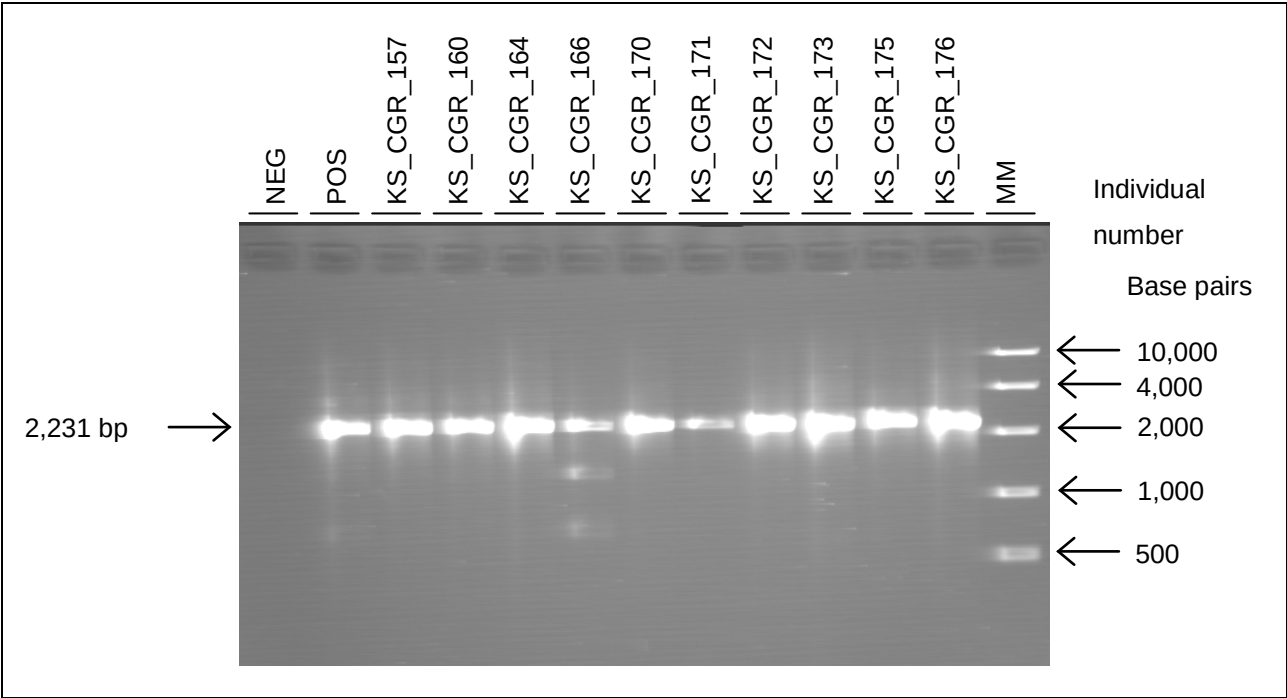
Two previously reported non-synonymous alterations were observed in the ND3 gene. The T10084C alteration resulted in the replacement of a non-polar isoleucine with a polar threonine (Campbell, 1995). This alteration was observed in a single Khoi-San individual (KS_CGR_044) classified as a member of the L0b sub-haplogroup (refer to Appendix E). The biochemical properties of the two amino acids involved in the non-synonymous change are different, however a detrimental functional effect is not expected due to the fact that the individual in whom the alteration was observed appeared to be healthy, and to date this alteration has also not been identified as a causative mutation in any disorders

(MITOMAP, 2008). The A10398G alteration caused an amino acid change from a polar threonine to a non-polar alanine (Campbell, 1995). This alteration was present in all the Khoi-San individuals studied in the current investigation (refer to Appendix A). This alteration involves the replacement of an amino acid with another amino acid with different biochemical characteristics. It is however not anticipated that this alteration will have a deleterious functional effect due to the fact that this alteration was present in all the studied Khoi-San individuals. In support of this deduction, the alteration also formed part of the exclusion criteria to identify L-specific mitochondrial genome sequences (refer to Appendix D).

5.7.6 PCR region 6 of the mitochondrial genome

PCR region 6 contains a total of 2,231 bp and includes base pairs 9,865 to 12,095 based on the numbering system of the rCRS (Andrews *et al.*, 1999). This region contains a number of functional elements namely ND4L and ND4 (refer to Section 2.4.2, Table 2.3). PCR was performed according to the conditions listed in Section 4.4. The amplified products generated for PCR region 6 were electrophoresed on an agarose gel as described in Section 4.5 and a representative gel picture of the amplified products obtained for PCR region 6 is depicted in Figure 5.35.

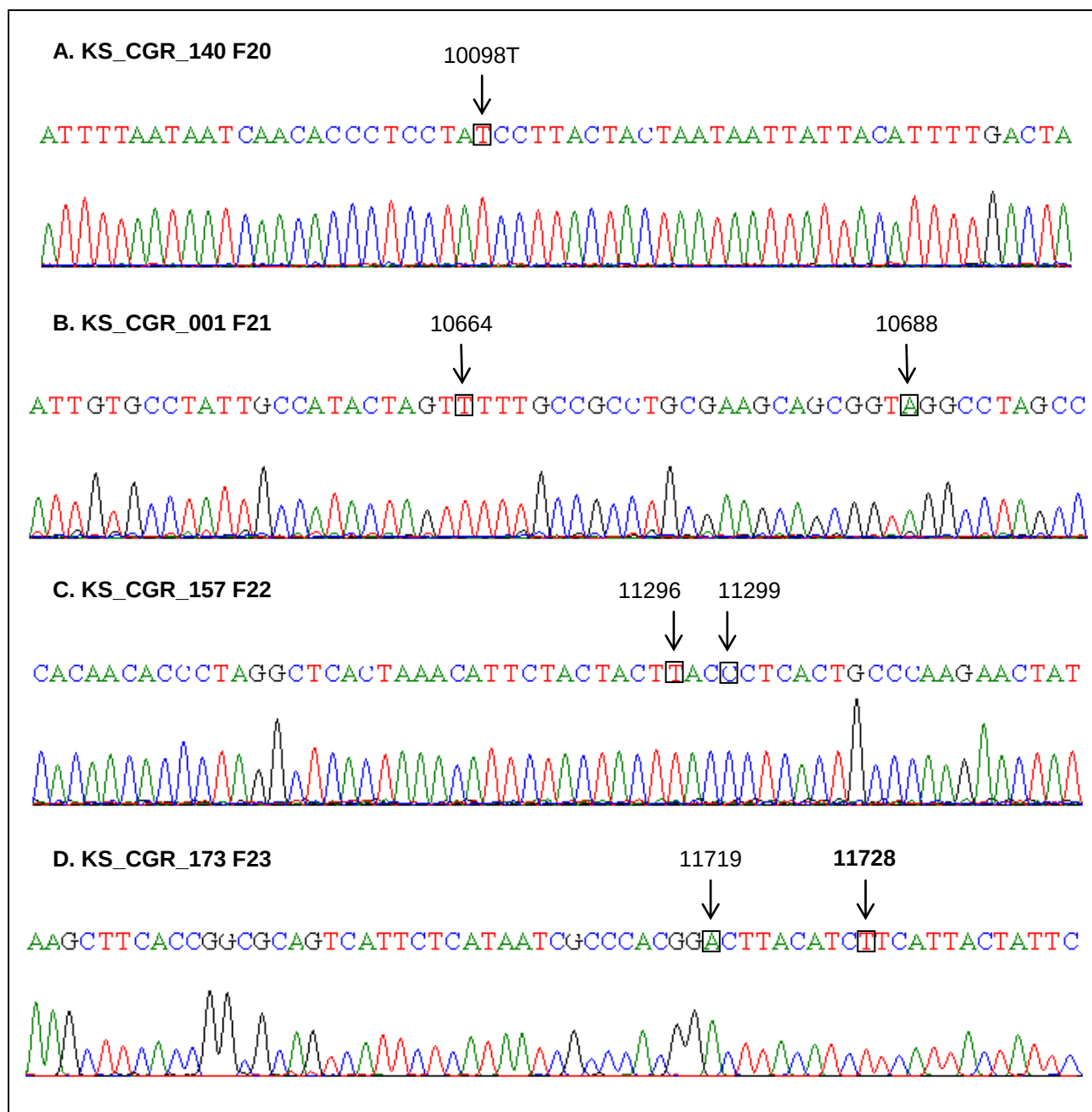
Figure 5.35 Photographic representation of amplified products of PCR region 6



Fragments were electrophoresed on a 0.9% agarose gel at 80 V for 45 min. MM = FastRuler™ DNA Ladder, High Range, ready to use, molecular marker (the sizes of the marker fragments are indicated to the right of the figure); NEG = negative control; POS = positive control; KS_CGR_157 to KS_CGR_176 = sample numbers given to each Khoi-San individual. The amplification efficiency was not equal for all samples, background smears are present in the sample lanes and secondary amplification is observed as discussed in Section 5.3.3. Artefacts are present in the gel matrix and the fragments are distorted as discussed in Section 5.4.1.

Gel electrophoresis of the amplified products was followed by PCR product purification as described in Section 4.4.3. The sequence of PCR region 6 was determined as discussed in Section 4.7 and the results analysed as described in Section 4.7.5. Sequence data for PCR region 6 was generated using four forward primers and representative electropherograms for each primer used are illustrated in Figure 5.36.

Figure 5.36 Representative electropherograms of PCR region 6 generated using forward primers F20 to F23



A = adenine; C = cytosine; G = guanine; T = thymine. The first part of the sequence name designates the sample, and the second part of the name indicates the primer used to generate the sequence. The nucleotide positions of alterations are indicated by blocks (□). Alterations are transitions unless specified otherwise. Reported alterations are indicated in regular black text, and a novel alteration is indicated in **bold** black text. Low signal strength and background noise can be observed in the electropherogram as discussed in Section 5.6.4.

The sequence data generated for PCR region 6 was compared with the corresponding section of the rCRS (refer to Section 4.7.5), and the comparison revealed a total of 27 discrepancies. All 27 discrepancies were transitions. A total of 26 discrepancies represent previously reported polymorphisms, while one discrepancy was novel. The sequence alterations observed in all screened Khoi-San individuals are presented in Sections 5.7.6.1 and 5.7.6.2. The alterations observed in PCR region 6 are discussed according to the functional unit in which the alteration was observed.

5.7.6.1 NADH dehydrogenase subunit 4L

The NADH dehydrogenase subunit 4L region consists of bp 10,470 to bp 10,766. The observed alterations are indicated in Table 5.21.

Table 5.21 Sequence alterations observed between the complete Khoi-San mitochondrial DNA genome sequences and the rCRS in the ND4L gene

Alteration	Locus	Alteration type	Codon position	Amino acid change	Reference
10,499 A – G	MT-ND4L	TS	coding: 3	SYN	Rieder <i>et al.</i> , 1998
10,589 G – A	MT-ND4L	TS	coding: 3	SYN	Howell <i>et al.</i> , 1995
10,664 C – T	MT-ND4L	TS	coding: 3	SYN	Ingman <i>et al.</i> , 2000
10,688 G – A	MT-ND4L	TS	coding: 3	SYN	Ingman <i>et al.</i> , 2000

Nucleotides are represented by the capital letters, with the rCRS nucleotide indicated first, followed by the observed change. Nucleotide numbers are based on rCRS numbering (Andrews *et al.*, 1999). The numbers indicated in red represent previously reported polymorphisms. A = adenine; C = cytosine; G = guanine; T = thymine; MT-ND4L = NADH dehydrogenase subunit 4L; TS = transition; SYN = synonymous substitution.

All of the alterations observed in the ND4L gene have been reported previously. In addition, the alterations are all transitions which bring about synonymous changes. The G10688A alteration forms part of the criteria that excludes haplogroups L0 and L1 from haplogroups L2, L3 and M (refer to Appendix D) and this alteration is present in all Khoi-San individuals belonging to the L0a and L0b sub-haplogroups (refer to Appendix E). The G10589A and C10664T alterations are also present in all Khoi-San individuals classified within the L0a or L0b sub-haplogroups (refer to Appendix E). The A10499G alteration is present in all Khoi-San individuals belonging to the L0a sub-haplogroup, as well as in a number of individuals classified within the L0b sub-haplogroup (refer to Appendix E).

5.7.6.2 NADH dehydrogenase subunit 4

The NADH dehydrogenase subunit 4 region consists of bp 10,760 to bp 12,137. The observed alterations are indicated in Table 5.22.

Table 5.22 Sequence alterations observed between the complete Khoi-San mitochondrial DNA genome sequences and the rCRS in the ND4 gene

Alteration	Locus	Alteration type	Codon position	Amino acid change	Reference
10,810 T – C	MT-ND4	TS	coding: 3	SYN	Ingman <i>et al.</i> , 2000
10,819 A – G	MT-ND4	TS	coding: 3	SYN	Wallace <i>et al.</i> , 1988
10,828 T – C	MT-ND4	TS	coding: 3	SYN	Mishmar <i>et al.</i> , 2003
10,873 T – C	MT-ND4	TS	coding: 3	SYN	Wallace <i>et al.</i> , 1988
10,876 A – G	MT-ND4	TS	coding: 3	SYN	Finnilä <i>et al.</i> , 2001b
10,899 A – G	MT-ND4	TS	coding: 2	NON-SYN: Asn → Ser	Herrnstadt <i>et al.</i> , 2002a
10,915 T – C	MT-ND4	TS	coding: 3	SYN	Yoneda <i>et al.</i> , 1990
10,920 C – T	MT-ND4	TS	coding: 2	NON-SYN: Pro → Leu	Ingman <i>et al.</i> , 2000
10,939 C – T	MT-ND4	TS	coding: 3	SYN	Ingman <i>et al.</i> , 2000
11,087 T – C	MT-ND4	TS	coding: 1	NON-SYN: Phe → Leu	Herrnstadt <i>et al.</i> , 2002a
11,296 C – T	MT-ND4	TS	coding: 3	SYN	Schwartz and Vissing, 2002
11,299 T – C	MT-ND4	TS	coding: 3	SYN	Johns, 1991
11,404 A – G	MT-ND4	TS	coding: 3	SYN	Tanaka <i>et al.</i> , 2004
11,437 T – C	MT-ND4	TS	coding: 3	SYN	Kong <i>et al.</i> , 2003
11,653 A – G	MT-ND4	TS	coding: 3	SYN	Horváth <i>et al.</i> , 2002
11,719 G – A	MT-ND4	TS	coding: 3	SYN	Wallace <i>et al.</i> , 1988
11,728 C – T	MT-ND4	TS	coding: 3	SYN	Current investigation
11,854 T – C	MT-ND4	TS	coding: 3	SYN	Mishmar <i>et al.</i> , 2003
11,914 G – A	MT-ND4	TS	coding: 3	SYN	Tanaka and Ozawa, 1994
11,944 T – C	MT-ND4	TS	coding: 3	SYN	Ozawa <i>et al.</i> , 1991a
12,007 G – A	MT-ND4	TS	coding: 3	SYN	Mackey and Howell, 1992
12,070 G – A	MT-ND4	TS	coding: 3	SYN	Brown <i>et al.</i> , 2001
12,121 T – C	MT-ND4	TS	coding: 3	SYN	Mishmar <i>et al.</i> , 2003

Nucleotides are represented by the capital letters, with the rCRS nucleotide indicated first, followed by the observed change. Nucleotide numbers are based on rCRS numbering (Andrews *et al.*, 1999). The numbers indicated in red represent previously reported polymorphisms and the number indicated in blue represents a novel polymorphism. A = adenine; C = cytosine; G = guanine; T = thymine; MT-ND4 = NADH dehydrogenase subunit 4; TS = transition; NON-SYN = non-synonymous substitution; SYN = synonymous substitution; Asn = asparagine; Leu = leucine; Phe = phenylalanine; Pro = proline; Ser = serine.

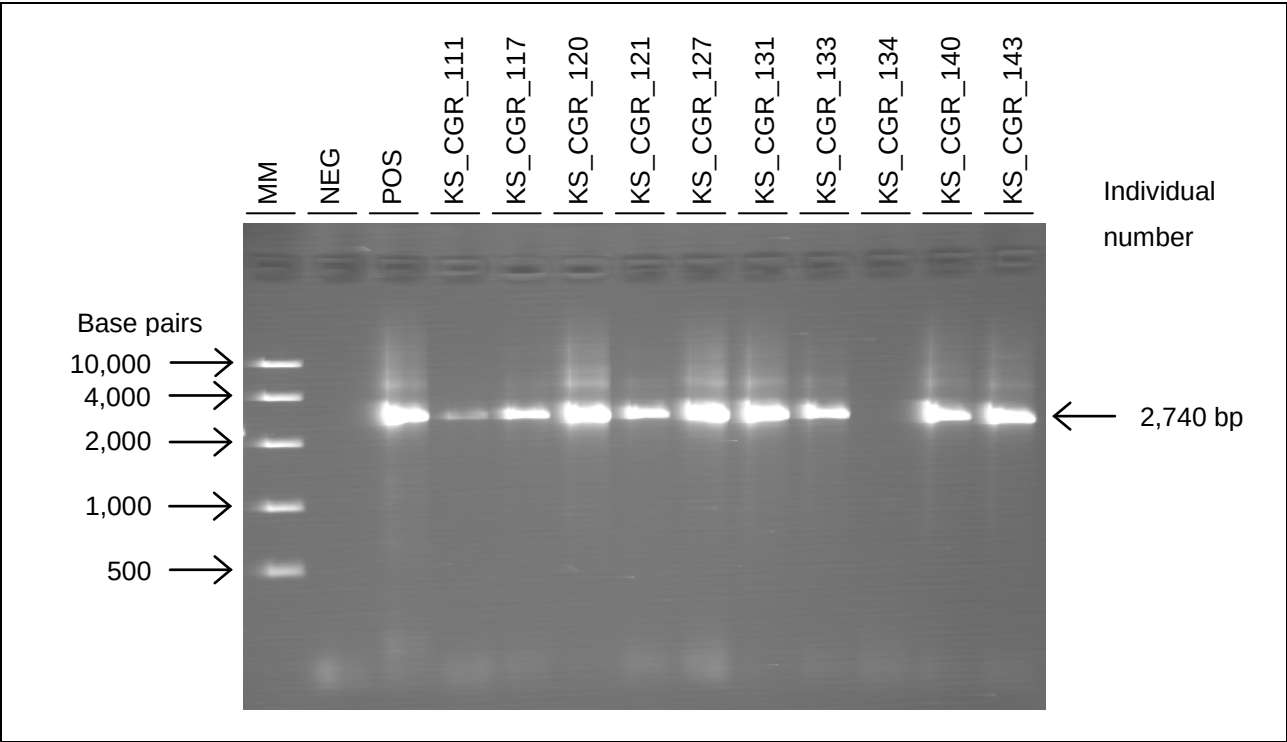
All of the alterations observed in the ND4 gene were transitions and the majority were previously reported polymorphisms. A single alteration, C11728T, observed in a single Khoi-San individual (KS_CGR_173) belonging to the L2b sub-haplogroup (refer to Appendix E) was previously unreported. This alteration gives rise to a synonymous change and therefore a functional effect is not expected for this alteration.

Three transitions were observed in the ND4 gene that led to a non-synonymous change. The A10899G alteration changes a polar asparagine to a polar serine, the C10920T alteration leads to the replacement of a non-polar proline with a non-polar leucine, and the T11087C alteration gives rise to a change from a non-polar phenylalanine to a non-polar leucine (Campbell, 1995). In the non-synonymous change brought about by these three alterations, the original amino acid and the replacement amino acid share similar biochemical properties. It is therefore not expected that these alterations will have a functional effect.

5.7.7 PCR region 7 of the mitochondrial genome

PCR region 7 consists of a total of 2,740 bp and contains base pairs 11,467 to 14,206 according to the numbering system of the rCRS (Andrews *et al.*, 1999). This region consists of a number of functional elements which include three tRNA genes, the ND5 and the ND6 genes (refer to Section 2.4.2, Table 2.3). PCR region 7 was amplified according to the conditions described in Section 4.4. The PCR products obtained were electrophoresed on an agarose gel as discussed in Section 4.5. A representative gel picture of the amplified products generated for PCR region 7 is depicted in Figure 5.37.

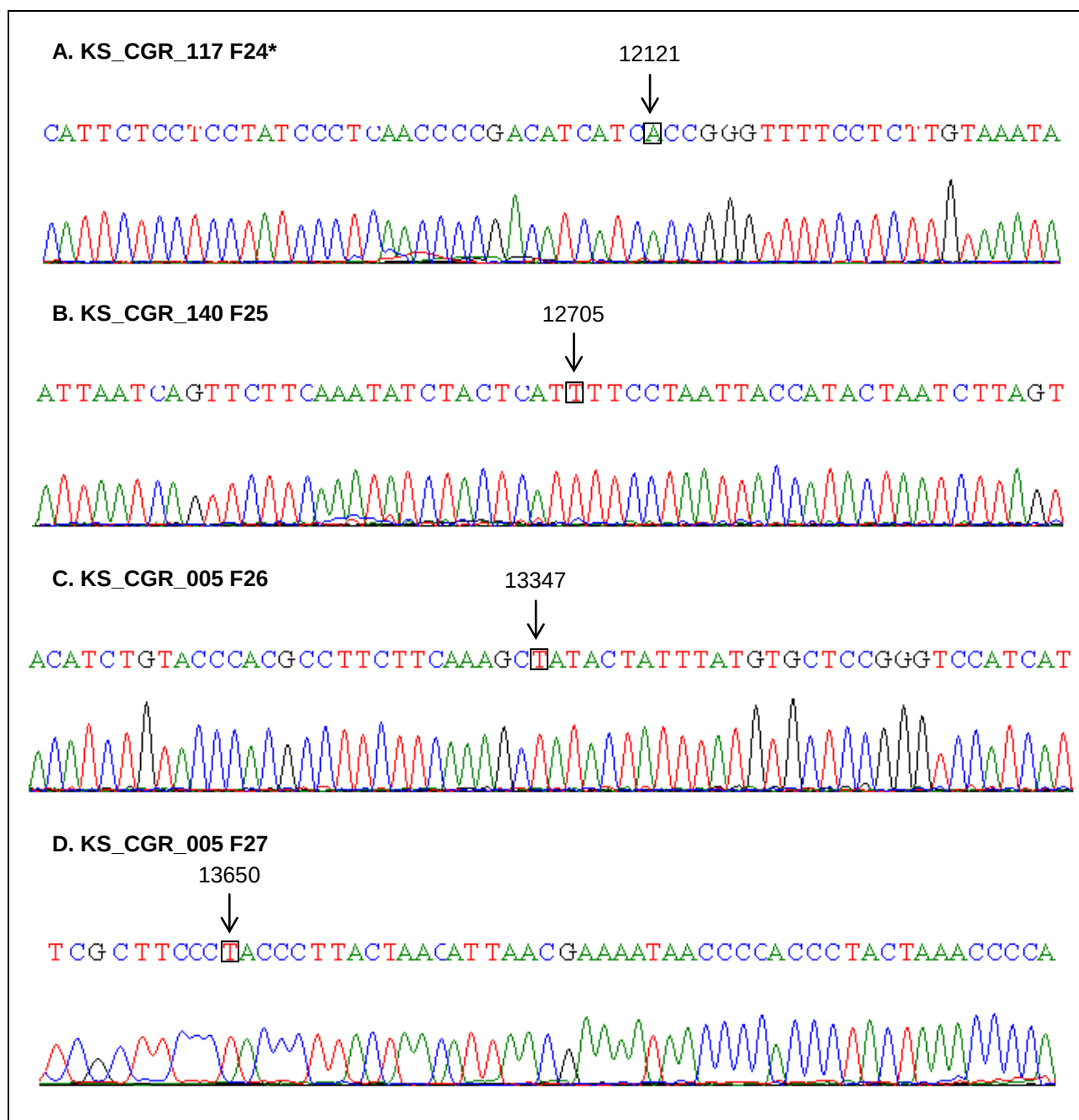
Figure 5.37 Photographic representation of amplified products of PCR region 7



Fragments were electrophoresed on a 0.9% agarose gel at 80 V for 45 min. MM = FastRuler™ DNA Ladder, High Range, ready to use, molecular marker (the sizes of the marker fragments are indicated to the left of the figure); NEG = negative control; POS = positive control; KS_CGR_111 to KS_CGR_143 = sample numbers given to each Khoi-San individual. The PCR for sample KS_CGR_134 failed, no PCR product is observed in the sample lane. The amplification efficiency was not equal for all samples, background smears are present in the sample lanes and primer-dimers are observed as discussed in Section 5.3.3. Artefacts are present in the gel matrix and the fragments are distorted as discussed in Section 5.4.1.

The PCR product purification was performed as described in Section 4.4.3. Sequencing of PCR region 7 was carried out as discussed in Section 4.7 and the results analysed as described in Section 4.7.5. A representative electropherogram for each of the four forward primers used to generate sequence data for PCR region 7 is illustrated in Figure 5.38.

Figure 5.38 Representative electropherograms of PCR region 7 generated using forward primers F24* to F27

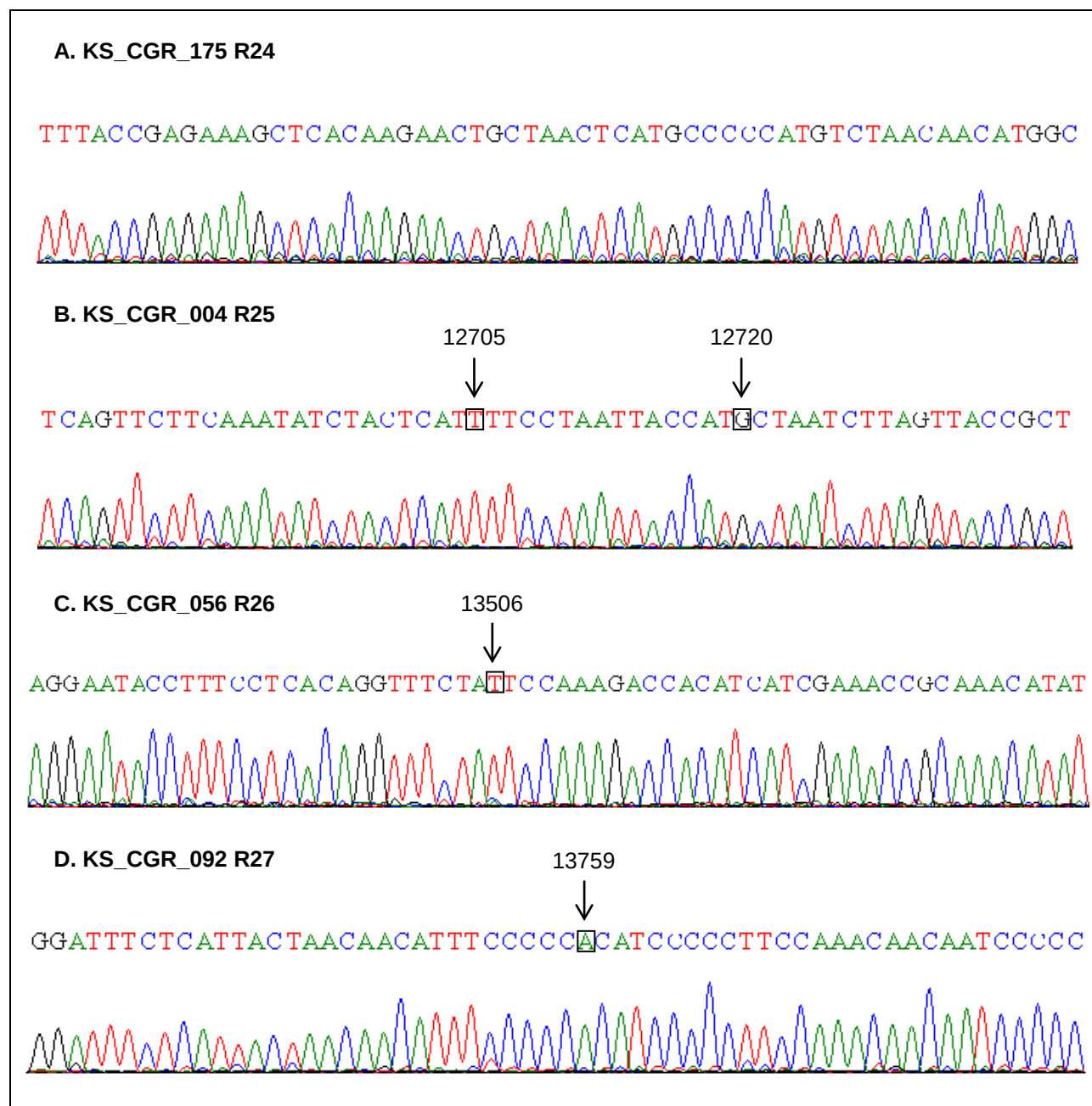


A = adenine; C = cytosine; G = guanine; T = thymine. The first part of the sequence name designates the sample, and the second part of the name indicates the primer used to generate the sequence. The nucleotide positions of alterations are indicated by blocks (□). Alterations are transitions unless specified otherwise. Reported alterations are indicated in regular black text. Minor dye blobs, low signal strength and background noise can be observed in the electropherogram as discussed in Section 5.6.4.

A representative electropherogram for the four reverse primers used to determine the sequence of PCR region 7 is presented in Figure 5.39. As before, the electropherograms

generated using the reverse primers are presented as the reverse complement (refer to Section 5.7.1).

Figure 5.39 Representative electropherograms of PCR region 7 generated using reverse primers R24 to R27



A = adenine; C = cytosine; G = guanine; T = thymine. The first part of the sequence name designates the sample, and the second part of the name indicates the primer used to generate the sequence. The nucleotide positions of alterations are indicated by blocks (□). Alterations are transitions unless specified otherwise. The observed alterations, all transitions, have been reported and are indicated in regular black text. Low signal strength and background noise can be observed in the electropherogram as discussed in Section 5.6.4.

When the sequence data generated for PCR region 7 was compared to the section of the rCRS representing this region (refer to Section 4.7.5), a total of 46 discrepancies were revealed which included 45 transitions and one transversion. The 46 discrepancies consisted of 42 previously reported and four previously unreported polymorphisms. The sequence alterations present in all screened Khoi-San individuals are presented in

Sections 5.7.7.1 to 5.7.7.3. PCR region 7 has been divided into functional units to facilitate the discussion of the observed alterations.

5.7.7.1 tRNA genes

The tRNA histidine region consists of bp 12,138 to bp 12,206, the tRNA serine (AGY) region includes bp 12,207 to bp 12,265 and the tRNA leucine (CUN) region comprises bp 12,266 to bp 12,336. The observed alterations are indicated in Table 5.23.

Table 5.23 Sequence alterations observed between the complete Khoi-San mitochondrial DNA genome sequences and the rCRS in tRNA genes

Alteration		Locus	Alteration type	Codon position	Reference
12,142	A – G	MT-TH	TS	non-coding	Kivisild <i>et al.</i> , 2006
12,235	T – C	MT-TS2	TS	non-coding	Ingman and Gyllensten, 2003
12,236	G – A	MT-TS2	TS	non-coding	Prezant <i>et al.</i> , 1993
12,303	C – T	MT-TL2	TS	non-coding	Kivisild <i>et al.</i> , 2006

Nucleotides are represented by the capital letters, with the rCRS nucleotide indicated first, followed by the observed change. Nucleotide numbers are based on rCRS numbering (Andrews *et al.*, 1999). The numbers indicated in red represent previously reported polymorphisms. A = adenine; C = cytosine; G = guanine; T = thymine; MT-TH = tRNA histidine; MT-TL2 = tRNA leucine (CUN); MT-TS2 = tRNA serine (AGY); TS = transition.

All of the alterations observed in the tRNA genes contained in PCR region 7 have been reported previously. In addition, all the alterations were transitions and no pathogenic effects have been reported for any of the alterations (MITOMAP, 2008). The Khoi-San individuals in which these alterations were observed are listed in Appendix A.

5.7.7.2 NADH dehydrogenase subunit 5

The NADH dehydrogenase subunit 5 region consists of bp 12,337 to bp 14,148. The observed alterations are indicated in Table 5.24.

Table 5.24 Sequence alterations observed between the complete Khoi-San mitochondrial DNA genome sequences and the rCRS in the ND5 gene

Alteration		Locus	Alteration type	Codon position	Amino acid change	Reference
12,469	A – G	MT-ND5	TS	coding: 1	NON-SYN: Ile → Val	Tanaka <i>et al.</i> , 2004
12,630	G – A	MT-ND5	TS	coding: 3	SYN	Finnilä <i>et al.</i> , 2001b
12,705	C – T	MT-ND5	TS	coding: 3	SYN	Wallace <i>et al.</i> , 1988
12,720	A – G	MT-ND5	TS	coding: 3	SYN	Brown <i>et al.</i> , 2001
12,798	C – T	MT-ND5	TS	coding: 3	SYN	Kivisild <i>et al.</i> , 2006
12,804	T – C	MT-ND5	TS	coding: 3	SYN	Tanaka <i>et al.</i> , 2004

Table 5.24 Continued...

Alteration		Locus	Alteration type	Codon position	Amino acid change	Reference
12,948	A – G	MT-ND5	TS	coding: 3	SYN	Herrnstadt <i>et al.</i> , 2002a
13,020	T – C	MT-ND5	TS	coding: 3	SYN	Finnilä <i>et al.</i> , 2001b
13,105	A – G	MT-ND5	TS	coding: 1	NON-SYN: Ile → Val	Prezant <i>et al.</i> , 1993
13,129	C – T	MT-ND5	TS	coding: 1	NON-SYN: Pro → Ser	Herrnstadt <i>et al.</i> , 2002a
13,260	T – C	MT-ND5	TS	coding: 3	SYN	Tanaka <i>et al.</i> , 2004
13,276	A – G	MT-ND5	TS	coding: 1	NON-SYN: Met → Val	Ingman <i>et al.</i> , 2000
13,347	C – T	MT-ND5	TS	coding: 3	SYN	Current investigation
13,386	T – C	MT-ND5	TS	coding: 3	SYN	Current investigation
13,466	G – C	MT-ND5	TS	coding: 2	NON-SYN: Ser → Thr	Current investigation
13,506	C – T	MT-ND5	TS	coding: 3	SYN	Ingman <i>et al.</i> , 2000
13,590	G – A	MT-ND5	TS	coding: 3	SYN	Tanaka and Ozawa, 1994
13,650	C – T	MT-ND5	TS	coding: 3	SYN	Liu <i>et al.</i> , 2001
13,722	A – G	MT-ND5	TS	coding: 3	SYN	Tanaka and Ozawa, 1994
13,759	G – A	MT-ND5	TS	coding: 1	NON-SYN: Ala → Thr	Rieder <i>et al.</i> , 1998
13,819	T – C	MT-ND5	TS	coding: 1	NON-SYN: Phe → Leu	Ingman <i>et al.</i> , 2000
13,855	C – T	MT-ND5	TS	coding: 1	SYN	Herrnstadt <i>et al.</i> , 2002a
13,886	T – C	MT-ND5	TS	coding: 2	NON-SYN: Leu → Pro	Maca-Meyer <i>et al.</i> , 2001
13,924	C – T	MT-ND5	TS	coding: 1	NON-SYN: Pro → Ser	Herrnstadt <i>et al.</i> , 2002a
13,928	G – C	MT-ND5	TV	coding: 2	NON-SYN: Ser → Thr	Finnilä <i>et al.</i> , 2001b
13,966	A – G	MT-ND5	TS	coding: 1	NON-SYN: Thr → Ala	Kösel <i>et al.</i> , 1998
13,967	C – T	MT-ND5	TS	coding: 2	NON-SYN: Thr → Met	Huoponen <i>et al.</i> , 1993
14,020	T – C	MT-ND5	TS	coding: 1	SYN	Kösel <i>et al.</i> , 1998
14,059	A – G	MT-ND5	TS	coding: 1	NON-SYN: Ile → Val	Herrnstadt <i>et al.</i> , 2002a

Nucleotides are represented by the capital letters, with the rCRS nucleotide indicated first, followed by the observed change. Nucleotide numbers are based on rCRS numbering (Andrews *et al.*, 1999). The numbers indicated in **red** represent previously reported polymorphisms and the numbers indicated in **blue** represent novel polymorphisms. A = adenine; C = cytosine; G = guanine; T = thymine; MT-ND5 = NADH dehydrogenase subunit 5; TS = transition; TV = transversion; NON-SYN = non-synonymous substitution; SYN = synonymous substitution; Ala = alanine; Ile = isoleucine; Leu = leucine; Met = methionine; Phe = phenylalanine; Pro = proline; Ser = serine; Thr = threonine; Val = valine.

Twenty-nine alterations, 28 transitions and one transversion, were observed in the ND5 gene located in PCR region 7. This number of alterations exceeds the number of alterations observed for the other protein-coding genes (refer to Section 5.14.1). Fifteen of the alterations also gave rise to non-synonymous amino acid changes. The large number

of alterations and the large number of non-synonymous changes caused by these alterations were however not surprising as it has previously been reported that increased amino acid sequence variation is observed for ND5 and COI in Africans (Mishmar *et al.*, 2003).

While the majority of the alterations observed in the ND5 gene have previously been reported, three transitions at nucleotide positions 13,347, 13,386 and 13,466 were novel to the current investigation. The C13347T alteration was observed in one Khoi-San individual (KS_CGR_005) and the T13386C alteration was also detected in a single Khoi-San individual (KS_CGR_044). Both these Khoi-San individuals were classified as members of the L0b sub-haplogroup (refer to Appendix E). Both the C13347T and the T13386C alterations produced synonymous changes and no functional effects are therefore expected for these alterations.

The G13466C alteration was observed in a single Khoi-San individual (KS_CGR_044) and gives rise to a non-synonymous amino acid change whereby a polar serine is replaced by a polar threonine (Campbell, 1995). The biochemical properties of these two amino acids are similar and a functional effect is therefore not expected for this alteration (refer to Section 5.11.5.2).

The presence of the C13347T alteration in only one Khoi-San individual, and the presence of both the T13386C and the G13466C alterations in a single Khoi-San individual, could indicate that these alterations are “personal polymorphisms”. Such polymorphisms will be useful to identify these specific Khoi-San individuals belonging to the L0b sub-haplogroup within the group of 30 Khoi-San individuals studied in the current investigation (refer to Section 5.7.4.1).

A number of previously reported non-synonymous alterations were observed in the ND5 gene. The A12469G alteration was observed in a single Khoi-San individual (KS_CGR_092) belonging to the L0b sub-haplogroup (refer to Appendix E). This alteration leads to a change from a non-polar isoleucine to a non-polar valine (Campbell, 1995). As the biochemical characteristics of the two amino acids are similar, a functional effect is not anticipated for this alteration.

The A13105G alteration leads to the replacement of a non-polar isoleucine with a non-polar valine (Campbell, 1995). The A13105G alteration forms part of the criteria used

to exclude haplogroups L0 and L1 from haplogroups L2, L3 and M (refer to Appendix D) and was observed in all Khoi-San individuals classified as members of haplogroup L0 (refer to Section 5.8). Due to the similar biochemical properties of the two amino acids and the presence of this alteration in all Khoi-San individuals belonging to the L0 haplogroup, a deleterious functional effect is not expected for this alteration.

The C13129T alteration results in an amino acid change from a non-polar proline to a polar serine (Campbell, 1995). Although the two amino acids do not have similar biochemical properties, a functional effect is not expected for this alteration (refer to Section 5.11.5.2). The A13276G alteration causes a change from a non-polar methionine to a non-polar valine (Campbell, 1995). These two amino acids have similar biochemical characteristics and a functional effect due to this alteration is not expected (refer to Section 5.11.5.2).

The G13759A alteration gives rise to a non-synonymous change from a non-polar alanine to a polar threonine (Campbell, 1995). This alteration was observed in all Khoi-San individuals belonging to the L0b sub-haplogroup, with the exception of KS_CGR_044 (refer to Appendix E). Although the biochemical properties of the original amino acid and the replacement amino acid are different, the presence of this alteration in 14 out of 15 of the Khoi-San individuals classified as members of the L0b sub-haplogroup appears to suggest that the alteration will not have a detrimental functional effect. This conclusion is supported by the fact that to date this alteration has not been identified as a causative mutation in any disease (MITOMAP, 2008).

The T13819C alteration results in an amino acid change from a non-polar phenylalanine to a non-polar leucine, and the G13928C alteration gives rise to an amino acid change from a polar serine to a polar threonine (Campbell, 1995). These alterations were detected in all Khoi-San individuals belonging to the L0a sub-haplogroup (refer to Appendix E). Due to the fact that the original amino acids share similar biochemical properties with the replacement amino acids, and the presence of the alterations in all Khoi-San individuals classified as members of the L0a sub-haplogroup, it is expected that the alterations will not have a functional effect.

The T13886C alteration gives rise to a non-synonymous change from a non-polar leucine to a non-polar proline (Campbell, 1995). This alteration was observed in two Khoi-San individuals (KS_CGR_090 and KS_CGR_172) belonging to the L3b sub-haplogroup (refer

to Appendix E). Due to the fact that the amino acid replacement involves amino acids with similar biochemical characteristics, it is not expected that this alteration will have a functional effect.

The C13924T alteration leads to the amino acid replacement of a non-polar proline with a polar serine (Campbell, 1995). This alteration was observed in two Khoi-San individuals (KS_CGR_069 and KS_CGR_173) belonging to the L2b sub-haplogroup (refer to Appendix E). Although the alteration involves the replacement of an amino acid with another amino acid with different biochemical characteristics, the alteration was observed in two apparently healthy individuals and for this reason it is expected that the alteration will not have a detrimental functional effect. This is further supported by the fact that the alteration has not been identified as a causative mutation in any disease (MITOMAP, 2008).

The A13966G alteration results in a non-synonymous change whereby a polar threonine is replaced by a non-polar alanine (Campbell, 1995). The alteration was observed in a single Khoi-San individual (KS_CGR_164) belonging to the L2b sub-haplogroup (refer to Appendix E). The alteration involves the replacement of one amino acid with another amino acid with different biochemical properties. It is however not expected that the alteration will have a functional effect due to the fact that the alteration was observed in an apparently healthy individual and the alteration has to date not been implicated as the cause of any disorder (MITOMAP, 2008).

The C13967T alteration causes a non-synonymous change and involves the replacement of a polar threonine with a non-polar methionine (Campbell, 1995). This alteration was observed in two Khoi-San individuals (KS_CGR_061 and KS_CGR_127) classified as members of the L0b sub-haplogroup (refer to Appendix E). Although the biochemical properties of the two amino acids are not the same, the alteration occurred in two apparently healthy individuals, which seems to indicate that the alteration will not have a detrimental functional effect. This is supported by the fact that the alteration has not been identified as the causative mutation in any disease (MITOMAP, 2008).

The A14059G alteration gives rise to a non-synonymous change whereby a non-polar isoleucine is replaced by a non-polar valine (Campbell, 1995). This alteration was detected in all three Khoi-San individuals belonging to the L2 haplogroup (refer to Section 5.8), and as the original amino acid and the replacement amino acid share similar biochemical

properties, and the alteration is observed in apparently healthy Khoi-San individuals, it is therefore not anticipated that the alteration will have a deleterious functional effect.

5.7.7.3 NADH dehydrogenase subunit 6

The NADH dehydrogenase subunit 6 region consists of bp 14,149 to bp 14,673. The observed alterations are indicated in Table 5.25.

Table 5.25 Sequence alterations observed between the complete Khoi-San mitochondrial DNA genome sequences and the rCRS in the ND6 gene

Alteration	Locus	Alteration type	Codon position	Amino acid change	Reference
14,182 T – C	MT-ND6	TS	coding: 3	SYN	Huoponen <i>et al.</i> , 1993
14,211 C – T	MT-ND6	TS	coding: 1	NON-SYN: Val → Ile	Ingman and Gyllensten, 2007a
14,212 T – C	MT-ND6	TS	coding: 3	SYN	Wallace <i>et al.</i> , 1988
14,280 A – G	MT-ND6	TS	coding: 1	NON-SYN: Ser → Pro	Herrnstadt <i>et al.</i> , 2002a
14,284 C – T	MT-ND6	TS	coding: 3	NON-SYN: Glu → Asp	Ingman <i>et al.</i> , 2000
14,302 T – C	MT-ND6	TS	coding: 3	SYN	Current investigation
14,315 C – T	MT-ND6	TS	coding: 2	NON-SYN: Ser → Asn	Kivisild <i>et al.</i> , 2006
14,371 T – C	MT-ND6	TS	coding: 3	SYN	Ingman <i>et al.</i> , 2000
14,374 T – C	MT-ND6	TS	coding: 3	SYN	Ingman <i>et al.</i> , 2000
14,407 C – T	MT-ND6	TS	coding: 3	SYN	Thangaraj <i>et al.</i> , 2005
14,560 G – A	MT-ND6	TS	coding: 3	SYN	Howell <i>et al.</i> , 1991
14,569 G – A	MT-ND6	TS	coding: 3	SYN	Ozawa <i>et al.</i> , 1991a
14,659 C – T	MT-ND6	TS	coding: 3	SYN	Kivisild <i>et al.</i> , 2006

Nucleotides are represented by the capital letters, with the rCRS nucleotide indicated first, followed by the observed change. Nucleotide numbers are based on rCRS numbering (Andrews *et al.*, 1999). The numbers indicated in red represent previously reported polymorphisms and the number indicated in blue represents a novel polymorphism. A = adenine; C = cytosine; G = guanine; T = thymine; MT-ND6 = NADH dehydrogenase subunit 6; TS = transition; NON-SYN = non-synonymous substitution; SYN = synonymous substitution; Asn = asparagine; Asp = aspartic acid; Glu = glutamic acid; Ile = isoleucine; Pro = proline; Ser = serine; Thr = threonine; Val = valine.

All of the alterations observed in the ND6 gene were transitions and the majority of the alterations have been reported previously. A single transition at nucleotide position 14,302 was previously unreported. The T14302C alteration was observed in a single Khoi-San individual (KS_CGR_044) classified as a member of the L0b sub-haplogroup (refer to Appendix E). This alteration gives rise to a synonymous change, and it is therefore not expected to have a functional effect. Due to the fact that this alteration occurs in a single Khoi-San individual, it is possible that this alteration is an example of a “personal polymorphism” which will be useful to distinguish the specific Khoi-San individual from

other Khoi-San individuals belonging to the same sub-haplogroup within the cohort under investigation (refer to Section 5.7.4.1).

A number of previously reported non-synonymous alterations were observed in the ND6 gene. The C14211T alteration was observed in two Khoi-San individuals (KS_CGR_048 and KS_CGR_060) belonging to the L0b sub-haplogroup (refer to Appendix E). This alteration gives rise to an amino acid change from a non-polar valine to a non-polar isoleucine (Campbell, 1995). The biochemical properties of these two amino acids are similar and a functional effect due to this alteration is therefore not anticipated.

The A14280G alteration was detected in a single Khoi-San individual (KS_CGR_056) belonging to the L0b sub-haplogroup (refer to Appendix E). This alteration causes the replacement of a polar serine with a non-polar proline (Campbell, 1995). Although the biochemical characteristics of the original amino acid and the replacement amino acid are different, a detrimental functional effect is not expected for this alteration, as it was observed in an apparently healthy individual and to date this alteration has not been reported as the causative mutation in any disease (MITOMAP, 2008).

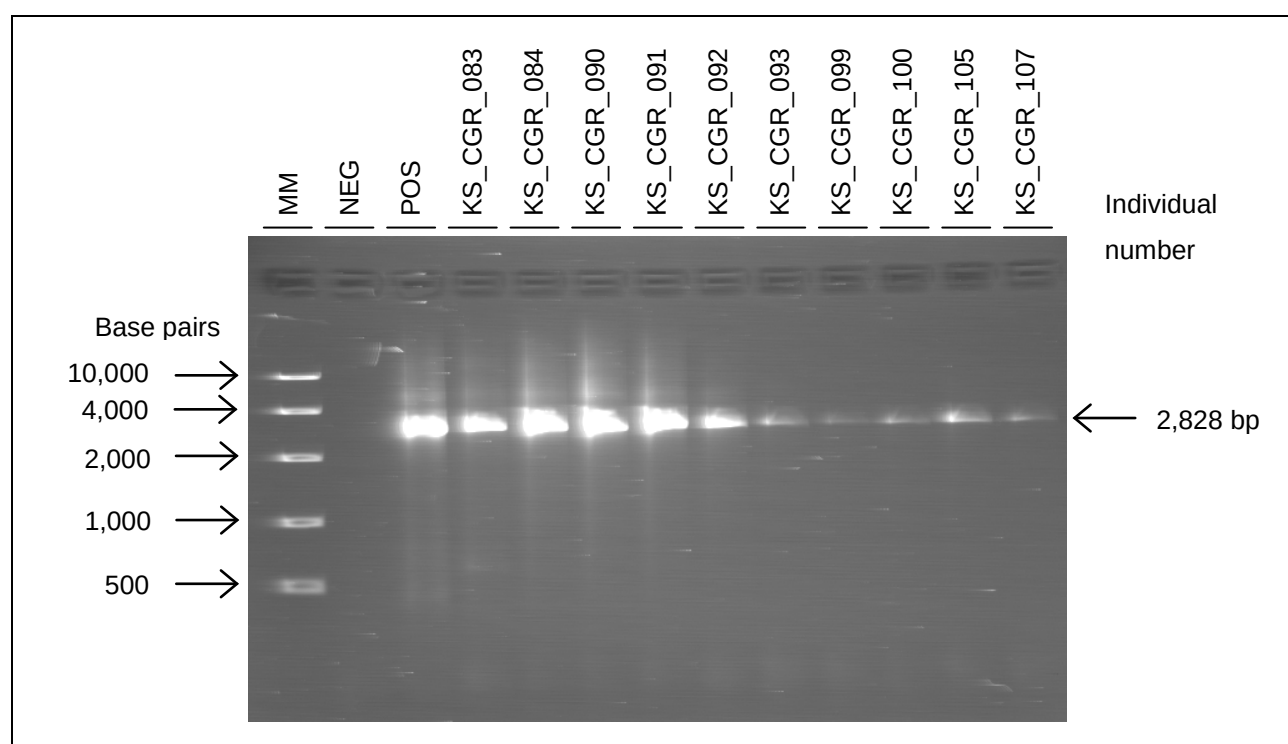
The C14284T alteration gives rise to a non-synonymous change from a glutamic acid to an aspartic acid. Both of these amino acids have acidic side chains (Campbell, 1995). This alteration was observed in two Khoi-San individuals (KS_CGR_090 and KS_CGR_172) that were classified as members of the L3b sub-haplogroup (refer to Appendix E). Due to the similar biochemical properties of the two amino acids, a functional effect is not expected for this alteration.

The C14315T alteration leads to an amino acid change from a polar serine to a polar asparagine (Campbell, 1995). This alteration was detected in three Khoi-San individuals (KS_CGR_048, KS_CGR_056 and KS_CGR_060). These individuals were classified as members of the L0b sub-haplogroup (refer to Appendix E). The amino acid replacement involves the exchange of amino acids with similar biochemical properties, and for this reason a functional effect due to the alteration is not expected.

5.7.8 PCR region 8 of the mitochondrial genome

PCR region 8 comprises a total of 2,828 bp and includes base pairs 13,593 to 16,420 based on the numbering system of the rCRS (Andrews *et al.*, 1999). This region contains a number of functional elements which include three tRNA genes, the cytochrome *b* gene and a number of non-coding nucleotides (refer to Section 2.4.2, Table 2.3). PCR was performed according to the conditions described in Section 4.4. Agarose gel electrophoresis was carried out as discussed in Section 4.5 and a representative gel picture of the amplified products obtained for PCR region 8 is presented in Figure 5.40.

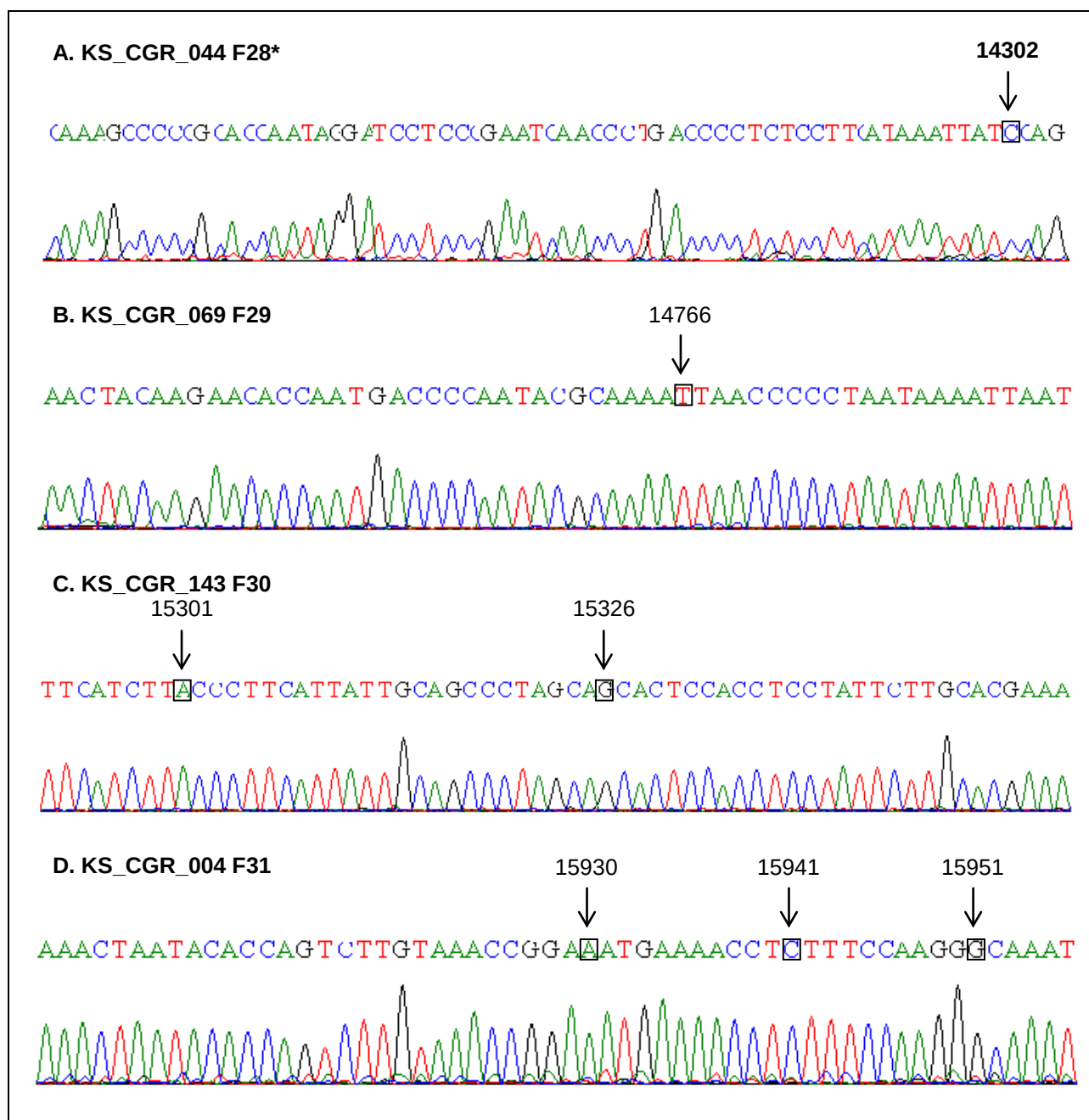
Figure 5.40 Photographic representation of amplified products of PCR region 8



Fragments were electrophoresed on a 0.9% agarose gel at 80 V for 45 min. MM = FastRuler™ DNA Ladder, High Range, ready to use, molecular marker (the sizes of the marker fragments are indicated to the left of the figure); NEG = negative control; POS = positive control; KS_CGR_083 to KS_CGR_107 = sample numbers given to each Khoi-San individual. The amplification efficiency was not equal for all samples and background smears are present in the sample lanes as discussed in Section 5.3.3. Artefacts are present in the gel matrix and the fragments are distorted as discussed in Section 5.4.1.

After gel electrophoresis, the PCR products were purified as discussed in Section 4.4.3. Sequencing of the PCR region was performed as discussed in Section 4.7 and the results were analysed as discussed in Section 4.7.5. A representative electropherogram for each of the four forward primers used to generate sequence data for PCR region 8 is presented in Figure 5.41.

Figure 5.41 Representative electropherograms of PCR region 8 generated using forward primers F28* to F31



A = adenine; C = cytosine; G = guanine; T = thymine. The first part of the sequence name designates the sample, and the second part of the name indicates the primer used to generate the sequence. The nucleotide positions of alterations are indicated by blocks (□). Alterations are transitions unless specified otherwise. Reported alterations are indicated in regular black text, while novel alterations are indicated in **bold** black text. Poor mobility correction, low signal strength and background noise can be observed in the electropherograms as discussed in Section 5.6.4.

A comparison between the sequence data generated for PCR region 8 and the section of the rCRS representing this region (refer to Section 4.7.5), revealed a total of 20 discrepancies. The 20 discrepancies observed were all transitions and were all previously reported polymorphisms. The sequence alterations detected in all screened Khoi-San individuals are presented in Sections 5.7.8.1 and 5.7.8.2. To facilitate discussion of the observed alterations, PCR region 8 was divided into functional units.

5.7.8.1 tRNA genes

The tRNA glutamic acid region consists of bp 14,674 to bp 14,742 and the tRNA threonine region contains bp 15,888 to bp 15,953. The observed alterations are indicated in Table 5.26.

Table 5.26 Sequence alterations observed between the complete Khoi-San mitochondrial DNA genome sequences and the rCRS in tRNA genes

Alteration		Locus	Alteration type	Codon position	Reference
14,693	A – G	MT-TE	TS	non-coding	Sternberg <i>et al.</i> , 2001
15,930	G – A	MT-TT	TS	non-coding	Herrnstadt <i>et al.</i> , 2002a
15,941	T – C	MT-TT	TS	non-coding	Mishmar <i>et al.</i> , 2003
15,942	T – C	MT-TT	TS	non-coding	Rose <i>et al.</i> , 2001
15,951	A – G	MT-TT	TS	non-coding	Ozawa <i>et al.</i> , 1991b

Nucleotides are represented by the capital letters, with the rCRS nucleotide indicated first, followed by the observed change. Nucleotide numbers are based on rCRS numbering (Andrews *et al.*, 1999). The numbers indicated in **red** represent previously reported polymorphisms. A = adenine; C = cytosine; G = guanine; T = thymine; MT-TE = tRNA glutamic acid; MT-TT = tRNA threonine; TS = transition.

All of the alterations detected in the tRNA genes contained in PCR region 8 have been reported previously. The observed alterations were all transitions, and the Khoi-San individuals in which these alterations were detected are listed in Appendix A. Detrimental functional effects due to any of the listed alterations are not expected, as to date none of the alterations have been confirmed as a causative mutation for a specific disorder (MITOMAP, 2008).

The A14693G alteration has however been reported in a patient affected by MELAS (Tzen *et al.*, 2003). The A14693G alteration occurs at a position within the TΨC loop of the secondary structure of the tRNA^{Glu} and affects a nucleotide that is not involved in base pairing and does not form part of the characteristic TΨC triplet located in this loop. The secondary structure of the tRNA^{Glu} is therefore not disrupted, and A14693G is regarded as a rare polymorphism until such time that it is reported in other cases with mitochondriopathy, or a pathogenic effect is confirmed by functional studies (Tzen *et al.*, 2003).

The A15951G alteration has been detected in a pedigree of a Han Chinese family suffering from LHON and it was proposed that this alteration may function as a potential modulator by increasing the penetrance and expressivity of the primary LHON-associated G11778A alteration (Li *et al.*, 2006). This alteration disrupts the base pairing in the amino

acid stem of tRNA^{Thr} and can impair the translation process through a failure to bind the amino acid properly. However, this alteration was also observed in controls used within the study, and it was concluded that the alteration on its own was not sufficient to cause a clinical phenotype (Li *et al.*, 2006).

5.7.8.2 Cytochrome *b*

The cytochrome *b* region consists of bp 14,747 to bp 15,887. The observed alterations are indicated in Table 5.27.

Table 5.27 Sequence alterations observed between the complete Khoi-San mitochondrial DNA genome sequences and the rCRS in the Cytb gene

Alteration		Locus	Alteration type	Codon position	Amino acid change	Reference
14,766	C – T	MT-CYTB	TS	coding: 2	NON-SYN: Thr → Ile	Marzuki <i>et al.</i> , 1991
14,770	C – T	MT-CYTB	TS	coding: 3	SYN	Schuelke <i>et al.</i> , 2002
14,905	G – A	MT-CYTB	TS	coding: 3	SYN	Howell <i>et al.</i> , 1995
15,061	A – G	MT-CYTB	TS	coding: 3	SYN	Maca-Meyer <i>et al.</i> , 2001
15,106	G – A	MT-CYTB	TS	coding: 3	SYN	Ozawa <i>et al.</i> , 1991b
15,110	G – A	MT-CYTB	TS	coding: 1	NON-SYN: Ala → Thr	Reid <i>et al.</i> , 1994
15,208	A – G	MT-CYTB	TS	coding: 3	SYN	Kivisild <i>et al.</i> , 2006
15,217	G – A	MT-CYTB	TS	coding: 3	SYN	Ozawa <i>et al.</i> , 1991b
15,301	G – A	MT-CYTB	TS	coding: 3	SYN	Ozawa <i>et al.</i> , 1991a
15,326	A – G	MT-CYTB	TS	coding: 1	NON-SYN: Thr → Ala	Ozawa <i>et al.</i> , 1991a
15,449	T – C	MT-CYTB	TS	coding: 1	NON-SYN: Phe → Leu	Kivisild <i>et al.</i> , 2006
15,466	G – A	MT-CYTB	TS	coding: 3	SYN	Herrnstadt <i>et al.</i> , 2002a
15,519	T – C	MT-CYTB	TS	coding: 2	NON-SYN: Leu → Pro	Mishmar <i>et al.</i> , 2003
15,550	C – T	MT-CYTB	TS	coding: 3	SYN	Tanaka <i>et al.</i> , 2004
15,766	A – G	MT-CYTB	TS	coding: 3	SYN	Andreu <i>et al.</i> , 1999d

Nucleotides are represented by the capital letters, with the rCRS nucleotide indicated first, followed by the observed change. Nucleotide numbers are based on rCRS numbering (Andrews *et al.*, 1999). The numbers indicated in **red** represent previously reported polymorphisms. A = adenine; C = cytosine; G = guanine; T = thymine; MT-CYTB = cytochrome *b*; TS = transition; NON-SYN = non-synonymous; SYN = synonymous; Ala = alanine; Ile = isoleucine; Leu = leucine; Phe = phenylalanine; Pro = proline; Thr = threonine.

All of the transitions detected in the Cytb gene have been reported previously. The majority of the transitions involve synonymous changes with the exception of the transitions observed at nucleotide positions 14,766, 15,110, 15,326, 15,449 and 15,519 which lead to non-synonymous changes.

The C14766T alteration gives rise to an amino change from a polar threonine to a non-polar isoleucine, while the A15326G alteration leads to an amino acid change from a polar threonine to a non-polar alanine (Campbell, 1995). In both cases, the original amino acids and the replacement amino acids have different biochemical properties, therefore it is possible that the non-synonymous change will have a functional effect. Both of the alterations were present in all the Khoi-San individuals studied in the current investigation (refer to Appendix A) and form part of the exclusion criteria used to classify mitochondrial genome sequences as macrohaplogroup L-specific (refer to Appendix D). Although the biochemical characteristics of the original amino acids and the replacement amino acids are different, the presence of these alterations in all the Khoi-San individuals suggests that these alterations will not have a detrimental functional effect. This conclusion is supported by the fact that neither of the alterations has been identified as causative in any disease to date (MITOMAP, 2008).

The G15110A alteration results in a non-synonymous change whereby a non-polar alanine is replaced by a polar threonine (Campbell, 1995). Due to the different biochemical characteristics of the two amino acids, it is possible that the alteration will have a functional effect. The alteration was observed in three Khoi-San individuals (KS_CGR_069, KS_CGR_164 and KS_CGR_173) belonging to the L2b sub-haplogroup (refer to Appendix E). The presence of this alteration in all three Khoi-San individuals classified as members of the L2 haplogroup seems to suggest that the alteration will not have a detrimental functional effect, in spite of the fact that the biochemical properties of the original and replacement amino acids differ. This is supported by the fact that to date this alteration has not been implicated as the cause of any disease (MITOMAP, 2008).

The T15449C alteration gives rise to a non-synonymous change of a non-polar phenylalanine to a non-polar leucine, and the T15519C alteration leads to an amino acid change from a non-polar leucine to a non-polar proline (Campbell, 1995). The T15449C alteration was detected in a single Khoi-San individual (KS_CGR_056) belonging to the L0b sub-haplogroup, and the T15519C alteration was present in three Khoi-San individuals (KS_CGR_005, KS_CGR_091 and KS_CGR_117) belonging to the L0b sub-haplogroup (refer to Appendix E). Due to the fact that both these alterations involve the replacement of an amino acid with a biochemically similar amino acid, it is not anticipated that these alterations will have a deleterious functional effect.

5.8 MITOCHONDRIAL HAPLOGROUP ANALYSIS

Haplogroup assignment is an important tool in understanding the way in which the genetic background of an individual exerts an influence on the expression of the phenotype caused by an alteration in the mitochondrial DNA genome (refer to Section 3.4). In addition, the determination of the haplogroup designation of a mitochondrial genome sequence is one of the first steps in defining the evolutionary relationship that exists between different mitochondrial genome sequences. In this section haplogroups were assigned to the mitochondrial genome sequences of the Khoi-San individuals studied in the current investigation, as well as to a number of L-specific sequences obtained from other sources and used in various analyses.

The classification scheme used to perform the haplogroup analysis was adapted from Wallace, 2004 (refer to Section 4.9) and is based on the presence or absence of previously identified informative SNPs within the coding region of the mitochondrial DNA genome. In the past, various classification schemes were based on combinations of SNPs present in the control region (mainly HVS1 as well as certain HVS2 sequence data) and specific SNPs in the coding region that were identified through the use of selected restriction enzymes (Salas *et al.*, 2002; Salas *et al.*, 2004a). The introduction of new technologies, which increase the rate and ease with which sequence data can be generated, has however led to the increased use of classification schemes that rely on the presence or absence of informative SNPs within the coding region (e.g. Kivisild *et al.*, 2006; Ruiz-Pesini *et al.*, 2007). At present, two types of SNP based classification schemes are used to classify mitochondrial genome sequences into haplogroups (refer to Section 3.3.1.1). The coding region is generally viewed as more informative than the non-coding region which is subject to frequent occurrences of recurrent or parallel mutations, as well as reversions which can be interpreted as random mutations and can cause the incorrect clustering of sequences (refer to Section 3.3.1.1). Due to these factors, it was decided that the classification scheme based on SNPs observed only within the coding region would be employed for haplogroup analyses, since coding region markers allow mtDNA haplogroups to be defined without ambiguity. The haplogroup analysis of the Khoi-San mitochondrial genome sequences undertaken in the current investigation is discussed in Section 5.8.1, while the haplogroup assignment of other L-specific mitochondrial sequences used in the phylogenetic and statistical analyses, as well as in the construction of a global L0-specific haplogroup network (refer to Section 5.10), is described in Section 5.8.2. For the purpose of discussion in the current investigation, the

haplogroup designation determined according to the classification scheme used in the current investigation is provided first, and the haplogroup classification as reported in the literature is given thereafter in brackets. A rudimentary comparison of haplogroup and sub-haplogroup designations between the classification scheme used in the current investigation (refer to Section 4.9) and the classification scheme generally used in the literature (refer to Section 3.3.1.1) is provided in Appendix O in order to relate the data generated in the current investigation with the data obtained in other investigations as described in the literature.

5.8.1 Khoi-San mitochondrial genome sequences

The complete Khoi-San mtDNA genome sequences were compared to the global mtDNA haplogroup tree (Ruiz-Pesini *et al.*, 2007; refer to Appendix K), as well as the exclusion criteria for L-specific sequences (refer to Appendix D). This was done to ensure that the Khoi-San sequences fit the criteria for macrohaplogroup L designation. Thereafter, based on the classification scheme described in Section 4.9, haplogroup and sub-haplogroup designations were assigned to the mitochondrial DNA genome sequence data generated for each Khoi-San individual included in the current investigation.

Certain Khoi-San sequences belonging to the L0b sub-haplogroup did not strictly adhere to the exclusion criteria for L-specific sequences. The criteria that were not adhered to, and the sample numbers of the Khoi-San individuals in which these deviations from the criteria were observed, are presented in Table 5.28.

Table 5.28 Khoi-San genome sequences deviating from exclusion criteria for L-specific sequences

Nucleotide substitution	Exclusion criteria	Khoi-San sample			
A1438G	From the sub-haplogroup containing the CRS	KS_CGR_004, KS_CGR_056, KS_CGR_091, KS_CGR_117,	KS_CGR_005, KS_CGR_060, KS_CGR_092, KS_CGR_121,	KS_CGR_044, KS_CGR_061, KS_CGR_105, KS_CGR_127	KS_CGR_048, KS_CGR_077, KS_CGR_111,
A4769G		KS_CGR_044			
A2706G	From haplogroup H	KS_CGR_004, KS_CGR_060, KS_CGR_092, KS_CGR_121,	KS_CGR_005, KS_CGR_061, KS_CGR_105, KS_CGR_127	KS_CGR_048, KS_CGR_077, KS_CGR_111,	KS_CGR_056, KS_CGR_091, KS_CGR_117,
T6185C	L0 from L1	KS_CGR_091			

Khoi-San mitochondrial genome sequences were compared to the exclusion criteria for L-specific sequences listed in Appendix D. KS_CGR numbers indicate the sample number given to each Khoi-San individual in the current investigation. A = adenine; C = cytosine; G = guanine; T = thymine.

The A1438G alteration, which excludes sequences from sub-haplogroup H2 containing the CRS, was absent in all 15 Khoi-San samples classified within the L0b sub-haplogroup. The A1438G alteration was however present in all Khoi-San individuals belonging to the L0a, L2, L3a1 and L3b haplogroups. The presence of an adenine at this nucleotide position has been identified as a rare polymorphism (Andrews *et al.*, 1999), while the presence of a guanine at this position is viewed as the human consensus nucleotide for this position (MITOMAP, 2008). It can be speculated that the absence of this alteration is due to a reversion event in an ancestor that gave rise to the L0b sub-haplogroup. This deduction is supported by the previous observation of a reversion event at this nucleotide position (Herrnstadt *et al.*, 2002a). The absence of the A1438G alteration in all Khoi-San individuals classified as belonging to the L0b sub-haplogroup (current investigation), as well as the absence of this alteration in the L0aB group (refer to Section 5.10.2.1, Figure 5.50) consisting of Khoi-San speakers from East Africa (refer to Section 5.10.2.2) suggest that this nucleotide position is not suitable for use in the list of exclusion criteria for L-specific sequences. It is thus proposed that the exclusion criteria should be revised to take into account the L0 sequence data generated recently for Khoi-San individuals (Gonder *et al.*, 2007; current investigation) as it appears that the A1438G alteration is not present in a large number of Khoi-San speaking individuals.

The A to G transition at nucleotide position 4,769, that excludes sequences from the sub-haplogroup containing the CRS, was absent in a single Khoi-San individual (KS_CGR_044) belonging to the L0b sub-haplogroup. This alteration was present in all other Khoi-San individuals studied in the current investigation. The Khoi-San individual (KS_CGR_044) in whom this alteration was absent, grouped with three other individuals (AY195777, DQ112830 and DQ112850) in the L0bA cluster (refer to Section 5.10.2.1, Figure 5.51). This seems to suggest that the loss of this alteration is characteristic of this group. However, this alteration was present in the other three sequences clustered in this group, as well as all other L0 sequences used to construct the global L0-specific haplogroup network (refer to Sections 5.8.2 and 5.10.2.1, Figure 5.51). In addition, the presence of an adenine at this nucleotide position has been observed previously and is a rare polymorphism (Andrews *et al.*, 1999), while a guanine observed at this position is the human consensus nucleotide for this position (MITOMAP, 2008). Therefore it is concluded that the absence of this alteration is an example of a rare reversion event and it is an isolated occurrence which does not necessitate the re-evaluation of the inclusion of this alteration in the current exclusion criteria for L-specific sequences.

The A to G transition at nucleotide position 2,706, which excludes sequences from haplogroup H, was absent in 14 out of the 15 samples from the L0b sub-haplogroup. All of the individuals within the L0b sub-haplogroup in which this alteration was absent, belonged to the L0bB group (refer to Section 5.10.2.1, Figure 5.51). In contrast, the A2706G alteration was present in all Khoi-San individuals classified as members of the L0a, L0bA, L2, L3a1 and L3b haplogroups. It can be speculated that the absence of this alteration is due to a reversion event within an ancestor of the L0bB group. The possibility of such an event is supported by the observation that the 2,706 nucleotide position is hypervariable and underwent recurrent or parallel alteration events three times within the global L0-specific haplogroup network (refer to Section 5.10.2.3, Table 5.39). Due to the hypervariability of this nucleotide position which indicates its instability, and the fact that it is absent in the L0bB group consisting of 35 sequences from mainly Khoi-San speaking individuals from southern Africa (refer to Section 5.10.2.1, Figure 5.51), it is suggested that this position is not suitable for use in the list of exclusion criteria for L-specific sequences. It is further proposed that the exclusion criteria should be revised to reflect the recently obtained L0 sequence data for Khoi-San individuals (Mouton, 2003; Kivisild *et al.*, 2006; Gonder *et al.*, 2007; current investigation).

The T to C transition at nucleotide position 6,185, which excludes sequences in haplogroup L0 from sequences in haplogroup L1, was absent in a single Khoi-San individual (KS_CGR_091) belonging to the L0b sub-haplogroup. This alteration was present in all other Khoi-San individuals studied in the current investigation. The Khoi-San individual harbouring this alteration grouped with ten other individuals in the L0bB2a3 group (refer to Section 5.10.2.1, Figure 5.51). This appears to suggest that the loss of this alteration is typical of this group. However, the alteration was present in the other sequences observed in this group, as well as all other L0 sequences used in the construction the global L0-specific haplogroup network (refer to Sections 5.8.2 and 5.10.2.1, Figure 5.51). It is therefore concluded that the absence of this alteration in KS_CGR_091 is the result of a reversion event, and due to the relative rarity of this occurrence, it is not necessary to re-assess the inclusion of this alteration in the current exclusion criteria for L-specific sequences.

Mitochondrial haplogroups were assigned to the 30 complete Khoi-San mitochondrial DNA genome sequences as described in Section 4.9 and the results obtained are indicated in Table 5.29. The haplogroup results are also summarised in Appendix E, where additional results regarding the sub-haplogroup of each individual is presented.

Table 5.29 Assignment of haplogroups to Khoi-San individuals included in the current investigation

Haplogroup assignment	Individuals investigated					
L0	KS_CGR_001, KS_CGR_056, KS_CGR_083, KS_CGR_111, KS_CGR_157,	KS_CGR_004, KS_CGR_059, KS_CGR_084, KS_CGR_117, KS_CGR_176	KS_CGR_005, KS_CGR_060, KS_CGR_091, KS_CGR_120,	KS_CGR_044, KS_CGR_061, KS_CGR_092, KS_CGR_121,	KS_CGR_048, KS_CGR_077, KS_CGR_105, KS_CGR_127,	
L2	KS_CGR_069, KS_CGR_164, KS_CGR_173					
L3	KS_CGR_090, KS_CGR_140, KS_CGR_143, KS_CGR_172, KS_CGR_175					

Khoi-San individuals were assigned haplogroup designations based on the classification scheme presented in Section 4.9. KS_CGR numbers indicate the sample number given to each Khoi-San individual.

All the Khoi-San individuals grouped within macrohaplogroup L, as was expected based on their recent African ancestry. As can be observed in Table 5.29, of the 30 Khoi-San individuals investigated, 22 individuals were classified as L0, three individuals as L2 and five individuals as L3. It was expected that all or most of the Khoi-San individuals would be classified as belonging to the L0 haplogroup. In a previous study undertaken by Chen *et al.* in 2000 with a sample of 43 !Kung individuals, it was observed that 34 individuals grouped in haplogroup L0, two individuals grouped in haplogroup L2 and seven individuals grouped in haplogroup L3. Possible reasons for the classification of Khoi-San individuals in L-specific haplogroups other than L0 are presented in Table 5.30.

Table 5.30 Possible reasons for the classification of Khoi-San individuals in L-specific haplogroups other than L0

Reason	Description
The occurrence of admixture	Instances of intermarriage have previously been reported between the Khoi-San population and neighbouring Bantu populations (Crawhall, 2006).
Self-reporting of ethnicity	Ethnic affinity is frequently decided by focussing more on language than physical characteristics. An individual therefore defines their ethnicity based on their language and/or culture which is an acquired characteristic determined by their environment, instead of their actual ancestry determined by their genetic heritage (Renfrew, 1992).
Availability of ethnic data restricted to two generations	The classification of the Khoi-San individuals as !Kung was based on information provided by identity documents or birth certificates, or verbal confirmation of the ethnicity of a child through the questioning of both parents regarding their ancestry (Chen <i>et al.</i> , 2000). This method of ethnic classification is not as accurate as a pedigree spanning over several generations. In addition, when verbal confirmation of ethnicity is accepted, self-reporting of ethnicity is also incorporated into the ethnic classification method employed.

In the current investigation the most likely explanation for the presence of L2 and L3 haplogroups within the cohort is a combination of the three possible reasons presented in Table 5.30. Bantu-speaking people moved into East and southern Africa ca. 2,000 years ago in three movements, from the east, from the north and from the west (Huffman, 2006). Historically, a certain proportion of Bantu men (agro-pastoralists migrating from central and

eastern Africa into South Africa), appear to have married Khoi-San hunter-gatherer women (Crawhall, 2006). If admixture is assumed to have occurred between the Bantu and Khoi-San populations, specifically between certain Bantu women and Khoi-San men, it is possible that haplogroups generally associated with Bantu populations could be incorporated into the Khoi-San population. As the contact between these populations occurred in the distant past, the genetic signature will remain, although the carrier of the haplogroup has accepted the culture and language of the population group of which he/she is part. Thus an individual proclaims him/herself as a member of the Khoi-San population.

The assumption of admixture is supported by other investigations which provide evidence that indicates gene flow between the Khoi-San population and neighbouring Bantu-speaking populations. Salas *et al.* in 2002 identified a type of L3 sub-haplogroup (L3d3¹) that occurred at a high frequency in both Khwe and !Kung individuals in southern Africa which they concluded most likely arose in the Bantu population and spread later into the Khoi-San speaking population. The presence of another L3 sub-haplogroup, L3a1 (L3e2) in the Khoi-San individuals studied in the current investigation is similarly explained by exchanges with neighbouring Bantu-speaking populations. L3a (L3e) is observed in East, Central and West Africa (Alves-Silva *et al.*, 2000; Salas *et al.*, 2002), with L3a1 (L3e2) mainly observed in Atlantic western Africa. It was proposed that the L3a (L3e) sub-haplogroup originated in Central and eastern Africa ca. 46,000 ± 14,000 YBP and was dispersed after the rise of agriculture and the introduction of iron smelting technology due to the Bantu migrations in Africa (Bandelt *et al.*, 2001a). In addition, Khoi-San lineages, i.e. sub-haplogroup L0b (L0d), comprise 5% of the total mtDNA lineages in Bantu speakers of Mozambique (Salas *et al.*, 2002), as well as 25% and 50% respectively of the mtDNA lineages in the southern African Xhosa and Zulu ethnic groups, while the assimilation of Bantu lineages into Khoi-San speakers is ca. 24% (Chen *et al.*, 2000; Salas *et al.*, 2002). Tishkoff *et al.* in 2007 also observed that Southern African Khoi-San speakers were likely to have experienced gene flow from neighbouring Bantu-speaking populations and proposed that the extent of genetic exchange with the neighbouring populations varied with more isolated populations experiencing a lower level of gene flow.

¹ The haplogroup classifications presented in brackets refer to the classifications according to the alternative classification scheme (refer to Section 3.3.1.1).

Another possibility is admixture between the !Kung and Khwe as both speak Khoi-San languages and it has been observed that these two groups often intermarry (Chen *et al.*, 2000). The Khwe have physical characteristics that appear similar to that of Bantu-speaking populations, and are mainly classified within haplogroups L1, L2 and L3 (Chen *et al.*, 2000). The morphological similarities between the Khwe and Bantu speakers (Barnard, 1992) suggest gene flow has occurred between these two groups. The haplogroups observed within the Khwe could however also be due to ancient admixture caused by Bantu migrations within Africa as discussed in the previous paragraph. As ethnic data was only confirmed for two generations and by self-reporting, it is highly likely that admixture has occurred in the past, however this can not be revealed by investigating only two generations. African populations are ancient and their history mostly consists of oral tradition, i.e. few written records are available. For the purpose of this type of investigation, it would be ideal to sample an isolated population which has not had contact with many other populations. In addition, a family history spanning at least four generations should be obtained when possible. It is also important to have a working knowledge of the language groups and the movement of people within Africa, as this type of background information will place results in context.

5.8.2 Other L-specific mitochondrial genome sequences

A number of different datasets (refer to Section 4.10.4, Table 4.10) consisting of L-specific mitochondrial genome sequences for use in phylogenetic and statistical analyses, as well as in the construction of a global L0-specific haplogroup network, were compiled as described in Section 4.10. The mtDNA genome sequences used in the global L0-specific haplogroup network (refer to Section 5.10), were compared to the exclusion criteria for L-specific sequences (refer to Appendix D) to ensure that the sequences fit the criteria for macrohaplogroup L designation.

It was observed that certain mtDNA sequences included in the global L0-specific haplogroup network did not adhere to all the criteria for L-specific sequences, and the criteria that were deviated from, as well as the GenBank® accession numbers or sample numbers of the sequences that exhibited the deviations, are listed in Table 5.31.

Table 5.31 Sequences obtained from GenBank® deviating from exclusion criteria for L-specific sequences

Nucleotide substitution	Exclusion criteria	GenBank® accession numbers
G769A	L0, L1, and L2 from other haplogroups	EF184610
A1438G	From sub-haplogroup of H containing CRS	AY195777 DQ112830, DQ112844, DQ112850, DQ112852, DQ112854, DQ112855, DQ112856, DQ112899 EF184585, EF184586, EF184587, EF184588, EF184590, EF184592, EF184593, EF184594 KS_CGR_010, KS_CGR_032, KS_CGR_040, KS_CGR_052, KS_CGR_072, KS_CGR_082, KS_CGR_094, KS_CGR_102, KS_CGR_122, KS_CGR_130
A2706G	From haplogroup H	DQ112844, DQ112852, DQ112854, DQ112855, DQ112856, DQ112899 EF184586, EF184590, EF184592, EF184593, EF184594 KS_CGR_010, KS_CGR_032, KS_CGR_040, KS_CGR_052, KS_CGR_072, KS_CGR_082, KS_CGR_094, KS_CGR_102, KS_CGR_122, KS_CGR_130
A7146G	L0 and L1 from L2, L3 and M	AF346998
A10398G	From haplogroup N	KS_CGR_052
G12007A	L0 from L1	EF184585

Mitochondrial genome sequences included in the global L0-specific haplogroup network were compared to the exclusion criteria for L-specific sequences listed in Appendix D. The GenBank® accession numbers of the sequences are provided. The KS_CGR numbers indicate the sample numbers given to Khoi-San individuals (Mouton, 2003) for which GenBank® accession numbers were not yet available. A = adenine; C = cytosine; G = guanine; T = thymine.

The A1438G alteration, which excludes sequences from the sub-haplogroup containing the CRS, was absent in all 27 of the L-specific mitochondrial genome sequences classified within the L0aB group in the L0a sub-haplogroup (refer to Section 5.10.2.1, Figure 5.50), as well as in sequences belonging to both the L0bA and L0bB groups within the L0b sub-haplogroup (refer to Section 5.10.2.1, Figure 5.51). This alteration was discussed in Section 5.8.1. The absence of this SNP in all the additional L-specific sequences belonging to the L0a and L0b sub-haplogroups, and groups within these sub-haplogroups, support the observation made in Section 5.8.1 that this nucleotide position is not suitable for use in the list of exclusion criteria for L-specific sequences. As before, it is suggested that the exclusion criteria listed in Appendix D be revised to reflect the data obtained from the L-specific mitochondrial genome sequences generated in the current and other recent investigations (Mishmar *et al.*, 2003; Mouton, 2003; Kivisild *et al.*, 2006; Gonder *et al.*, 2007).

The A to G transition at nucleotide position 2,706, which excludes sequences from haplogroup H, was absent in 21 L-specific mitochondrial genome sequences belonging to the L0b sub-haplogroup. All of these mitochondrial genome sequences clustered in the

L0bB group (refer to Section 5.10.2.1, Figure 5.51). This alteration was discussed in Section 5.8.1. The fact that this SNP is absent in so many sequences all belonging to the same group within the L0b sub-haplogroup, supports the suggestions already made in Section 5.8.1 that this position is not appropriate to be used in the list of exclusion criteria for L-specific sequences, and that the exclusion criteria should be revised to reflect this.

An additional four SNPs that form part of the exclusion criteria were absent in four different L-specific mitochondrial genome sequences. The A to G transition at nucleotide position 769, which excludes haplogroups L0, L1 and L2 from other haplogroups, was absent in a single mitochondrial sequence (EF184610) that belonged to the L0a sub-haplogroup and clustered within the L0aA group (refer to Section 5.10.2.1, Figure 5.50). The A to G transition at nucleotide position 7,146, that excludes haplogroups L0 and L1 from L2, L3 and M, was absent in a single mitochondrial sequence (AF346998) that was classified within the L0c2 sub-haplogroup, and belonged to the L0c2D group within this sub-haplogroup (refer to Section 5.10.2.1, Figure 5.53). The A to G transition at nucleotide position 10,398, that excludes sequences from haplogroup N, was absent in a single Khoi-San individual (KS_CGR_052) that was classified within the L0b sub-haplogroup, specifically in the L0bB group (refer to Section 5.10.2.1, Figure 5.51). The G to A transition at nucleotide position 12,007, which excludes haplogroup L0 from haplogroup L1, was absent in a single mitochondrial sequence (EF184585) belonging to the L0aB group in the L0a sub-haplogroup (refer to Section 5.10.2.1, Figure 5.50). All four of these alterations were present in all the Khoi-San individuals studied in the current investigation, as well as in all the L-specific mitochondrial genome sequences included in the global L0-specific haplogroup network. This appears to suggest that the loss of each of these alterations in a single sequence represents an example of a rare reversion event and is therefore viewed as an isolated occurrence. Due to the infrequency of the loss of these alterations, re-evaluation of the inclusion of these alterations in the current exclusion criteria for L-specific sequences is not required.

In Sections 5.8.1 and 5.8.2 it was determined that all the Khoi-San sequences generated in the current investigation, as well as all the other L-specific sequences obtained from GenBank® and included in the global L0-specific haplogroup network, do not strictly adhere to the exclusion criteria for L-specific sequences (refer to Appendix D). It was suggested that the criteria should be revised to take into account the sequence data generated in the current investigation, as well as sequence data obtained from other recent investigations. The implication of the absence of certain exclusion criteria SNPs on

haplogroup assignment is discussed further in Section 5.12. In this section a revised classification scheme is proposed for the L0 haplogroup taking into account the exclusion criteria from which the sequences deviated. The optimal list of exclusion criteria for L-specific sequences will include only those SNPs which are observed in all L-specific sequences, with the exception of the inclusion of SNPs such as G769A, A7146G, A10398G and G12007A which are only rarely absent.

Haplogroup and sub-haplogroup designations were assigned to all the mitochondrial DNA genome sequences obtained from GenBank® based on the classification scheme described in Section 4.9 prior to their use in the phylogenetic (refer to Section 5.9) and statistical analyses (refer to Section 5.11), as well as in the construction of a global L0-specific haplogroup network (refer to Section 5.10). The haplogroups and sub-haplogroups assigned to these sequences are indicated in Appendix G. A total of 70 sequences could not be classified completely using the classification scheme employed in the current investigation, i.e. the sequences could not be classified to either the haplogroup or sub-haplogroup level. This was due to the absence of certain haplogroup- or sub-haplogroup-defining SNPs. The incompletely classified sequences were indicated by a dollar sign (\$) preceding the putative haplogroup classification (refer to Appendix G), and descriptions of the SNP profiles of the incompletely classified haplogroups are given in Section 4.9, Table 4.8. The incompletely classified sequences, with the exception of certain sequences (refer to Sections 4.10.1, 4.10.2 and 4.10.3), were included in both the phylogenetic (refer to Section 5.9) and statistical analyses (refer to Section 5.11), as well as in the construction of the global L0-specific haplogroup network (refer to Section 5.10).

5.9 PHYLOGENETIC ANALYSIS

In the current investigation the lineage-based approach (refer to Section 3.2.2) was followed to gather information regarding African mtDNA haplogroups. The lineage-based approach of studying haplogroups elucidates the history of the haplogroups themselves, including the approximate ages of the haplogroups within populations, it however does not provide direct insights into the history of the individual populations in which they are present (Simoni *et al.*, 2000).

The aim of the current investigation was to generate a consensus sequence for the Khoi-San population of southern Africa. The Khoi-San sequences generated in the current

investigation are mainly classified as L0 which is the oldest known haplogroup of macrohaplogroup L (refer to Section 3.3.1, Figure 3.6). In order to ensure that the L0 sequences used to construct the consensus sequence did in fact belong to one of the most ancient lineages within macrohaplogroup L, the evolutionary relationships between the L0 sequences and the other sequences grouped within macrohaplogroup L were explored. This was done by constructing phylogenetic trees consisting of Khoi-San sequences only, African macrohaplogroup L sequences (sequences determined from individuals residing in countries on the African continent), and global macrohaplogroup L sequences (sequences obtained from individuals of African descent residing in countries on the African continent and countries outside of Africa). It was expected that the sequences generated for the Khoi-San individuals would group at the base of each of the trees, closest to the outgroup.

5.9.1 Construction of phylogenetic trees

Genome sequences consisting of the coding region only (bp 577 to 16,023) were used to generate the phylogenetic trees. A list of all the sequences generated in the current investigation, as well as the sequences obtained from GenBank[®] and other sources (Mouton, 2003) for use in the phylogenetic trees, is indicated in Appendix G. The sequences were arranged into different datasets in order to investigate evolutionary relationships in the different sequence subsets. The datasets used for the construction of the different phylogenetic trees are indicated in Section 4.10.4, Table 4.10. The sizes of the datasets ranged from 58 sequences to 429 sequences (rCRS and outgroup sequence excluded), and the datasets used for phylogenetic analysis contained coding region only sequences. The datasets were considered to be both large (more than 20 sequences within a set) and complex due to the levels of variation between the sequences in each set (refer to Section 5.11.2).

In the past, phylogenetic trees were constructed based on control region sequences, i.e. HVS1 only, or HVS1 and HVS2 (e.g. Chen *et al.*, 2000; Salas *et al.*, 2002). Complete genome sequences however provide better resolution than trees based only on HVS1 sequences (Ingman and Gyllensten, 2001). In recent years, the trend has been to use only the coding region of complete genome sequences as the control region is likely to contain recurrent alterations or reversions that can influence the clustering of sequences within the phylogenetic trees leading to differences in the topology of phylogenetic trees based on complete sequences and coding region only sequences (Gonder *et al.*, 2007). However, in

order to achieve the main aim of the current investigation, i.e. to generate a consensus sequence, complete genome sequences were required. As limited complete mtDNA sequence data is available for the Khoi-San population, complete sequences will therefore contribute to the body of knowledge for this unique population. However, as discussed earlier, only the coding region of the complete genome sequences will be used in the phylogenetic analyses as this section of the sequences is not subject to the problems associated with the control region. In addition, the haplogroup classification scheme used in the current investigation is based on coding region SNPs only (refer to Section 5.8) and the use of only the coding region of the sequences in the phylogenetic analyses will therefore assist in drawing inferences regarding the clustering of haplogroups within the phylogenetic trees.

Phylogenetic trees were constructed as described in Section 4.12. The first step in the construction of the phylogenetic trees involved the correct alignment of the mitochondrial sequences contained in each sequence dataset (refer to Section 4.12.1). The quality of all alignments obtained was ensured by visual inspection of the alignment. The second step involved correcting for unequal rates of transitions and transversions, i.e. the transition/transversion ratio (refer to Section 4.12.3) and differences in substitution rates among nucleotide sites, i.e. the gamma shape parameter (refer to Section 4.12.4). The third step involved the choice of method. Two methods were employed, namely neighbour-joining or N-J and maximum parsimony or MP (refer to Section 3.2.2.2). The two methods and the reasons for their use are briefly discussed in Sections 5.9.2 and 5.9.3. The choice of the two different methods allows for the comparison of phylogenetic trees constructed by sequence data converted into distance data (the N-J method) with that constructed by the consideration of each nucleotide site directly (the MP method). The N-J method generates a tree with specific branch lengths, while the MP method provides additional information regarding the nucleotide site(s) which contribute to the length of the specific branch (Page and Holmes, 1998). This information is useful for the determination of attributes of extinct ancestors and provides insights about molecular evolution (Page and Holmes, 1998).

The F84 model (refer to Section 3.2.2.1) was employed in the construction of the neighbour-joining trees. This model includes a number of parameters (refer to Section 3.2.2.1). Each parameter that is added to this model is estimated based on the specific sequence dataset, and the addition of each parameter therefore introduces more uncertainty (Page and Holmes, 1998). However, the choice of model was determined by

which model was deemed to be the most realistic for the datasets used in the current investigation, and the choice of this model was based on the minimum parameters that would yield a reasonable estimate.

In order to obtain information about ancestor-descendant relationships, times of divergence and the direction of evolution, it was decided that the trees be rooted (refer to Section 4.12). Two methods were available to root the phylogenetic trees, namely midpoint rooting and outgroup rooting (refer to Section 3.2.2.3). A similar rate of evolution can not be assumed in the case of African mtDNA sequences, as it has been observed for African mtDNA sequences belonging to haplogroup L2 that the sub-haplogroups evolve at different rates (Torroni *et al.*, 2001). In contrast with midpoint rooting, rooting with an outgroup does not require the assumption of equal rates of evolution in different lineages (Hall, 2004). Outgroup rooting was therefore selected. Given the specific criteria (refer to Section 3.2.2.3) that the outgroup sequence must comply with, a chimpanzee sequence (*Pan troglodytes*) was chosen as the outgroup to root the trees as it was not too distantly related or too closely related to the other sequences in the tree (Horai *et al.*, 1995).

The reliability of the inferred trees was tested by bootstrap analysis (refer to Section 3.2.2.4) and 1,000 bootstrap replicates (Felsenstein, 1989) were performed for each tree (refer to Section 4.12). The number of bootstrap replicates performed was dependent on the size of the dataset, the computing power available, the time required for the analysis and the desired accuracy of the bootstrap value (Hedges, 1992). The accuracy of bootstrap values are dependent on the frequency of the appearance of particular phylogenetic groups among all of the trees constructed using this re-sampling method, as well as the number of replicates performed, and it has previously been suggested that 2,000 replicates are required to achieve 95% confidence in the bootstrap values (Hedges, 1992). However, due to the size of the dataset and computational limitations, only 1,000 replicates were performed in the current investigation and the bootstrap values for each branch and/or grouping within the trees are indicated (refer to Appendix J). It was observed that the bootstrap values obtained for the trees were generally low. This can be attributed to the fact that the bootstrap value associated with each node is a reflection of the specific dataset investigated and the tree construction method employed (Hedges, 1992), and that bigger datasets containing more taxa are likely to have smaller bootstrap values (Sanderson and Wojciechowski, 2000). In addition, bootstrap values can be influenced by high levels of sequence divergence and rate variability among lineages (Felsenstein, 1988). It is therefore likely that the observed low

bootstrap values are due to the large datasets investigated, as well as the high sequence divergence and rate variability among lineages that is observed for African mitochondrial genome sequences (refer to Section 3.3.1). Another factor that is likely to play a role in the generation of low bootstrap values is recent divergence between different lineages (Soltis and Soltis, 2003). Low bootstrap values were observed for branches in the L3 group which were likely due to the recent divergence of sub-haplogroups. The same major groups were however observed in trees produced using the N-J method (refer to Section 5.9.2) and the MP method (refer to Section 5.9.3), this appeared to lend support to the observed groups even though the observed bootstrap values were lower than desired.

A consensus tree was compiled after bootstrap analysis for each of the methods employed (refer to Section 3.2.2.5). For the purpose of the current investigation, a 50% majority rule consensus tree was determined for the MP method, and a majority rule (extended) consensus tree was determined for the N-J method (refer to Section 3.2.2.5). The MP method produced numerous equally parsimonious trees, and the consensus tree was constructed to represent a composite tree in which conflicting branching patterns were resolved by selecting the pattern observed in more than 50% of the trees. Although the N-J method produces a single tree for each dataset (refer to Section 3.2.2.2.1), a set of trees is generated when using the PHYLIP package to perform bootstrap analysis due to the generation of multiple re-sampled datasets (refer to Section 4.12.5). The consensus tree was therefore produced to depict a combined tree. This tree contains any set of sequences that appears in more than 50% of the trees, with the additional feature that other sets of sequences which are compatible with the tree are included in order of the frequency with which they have appeared until the tree is fully resolved.

The parameters required for the construction of phylogenetic trees as discussed in the preceding paragraphs were calculated for the different sequence datasets used for phylogenetic analysis. The values obtained for these different parameters are summarised in Table 5.32.

Table 5.32 Parameters calculated for datasets used in phylogenetic analyses

Dataset	Short description	Number of sequences ^a	Transition/transversion ratio	Gamma shape parameter	Number of bootstrap replicates
1a	All L Sequences	429 (coding region)	10.911 ^c	0.152	1,000
2a	All Africa L Sequences	242 (coding region)	9.054 ^d	0.131	1,000
3a	All Khoi-San L Sequences	58 (coding region)	10.406	0.016	1,000
3b	All Khoi-San L Sequences	53 (complete ^b)	6.957	0.017	1,000

^a = number of sequences excluding the rCRS and the outgroup sequence. ^b = complete sequences consist of the control region and the coding region. ^c = value estimated using a randomly generated representative subset of dataset 1a consisting of 51 sequences (including the rCRS and the outgroup sequence). ^d = value estimated using a randomly generated representative subset of dataset 2a consisting of 53 sequences (including the rCRS and the outgroup sequence). Refer to Section 4.12.3 for the estimation of the transition/transversion ratio using a randomly generated representative subset. Refer to Section 4.10.4, Table 4.10 for a detailed description of the datasets. Refer to Appendix G for information regarding the sequences included in each dataset.

The phylogenetic trees constructed using the N-J and MP methods are presented in Sections 5.9.2 and 5.9.3. The major branches are labelled L0*, L1*, L2* and L3*. The asterisks included in the haplogroup designations are generally used to indicate that the specific haplogroup consists of a group of mtDNA lineages that form part of the particular haplogroup, although these lineages can not be classified into any of the specific sub-haplogroups of that haplogroup (refer to Section 3.1.1.1). In this chapter, the asterisks are used to indicate that incompletely classified sequences, as well as confirmed L0, L1, L2 and L3 sequences, are observed within the branches labelled L0*, L1*, L2* and L3*. Previously unnamed branches observed within the phylogenetic trees were assigned sub-haplogroup designations to facilitate discussion with the understanding that the addition of more L-specific sequences to the different phylogenetic trees could alter the observed branching pattern, which will necessitate the renaming of the newly designated sub-haplogroups.

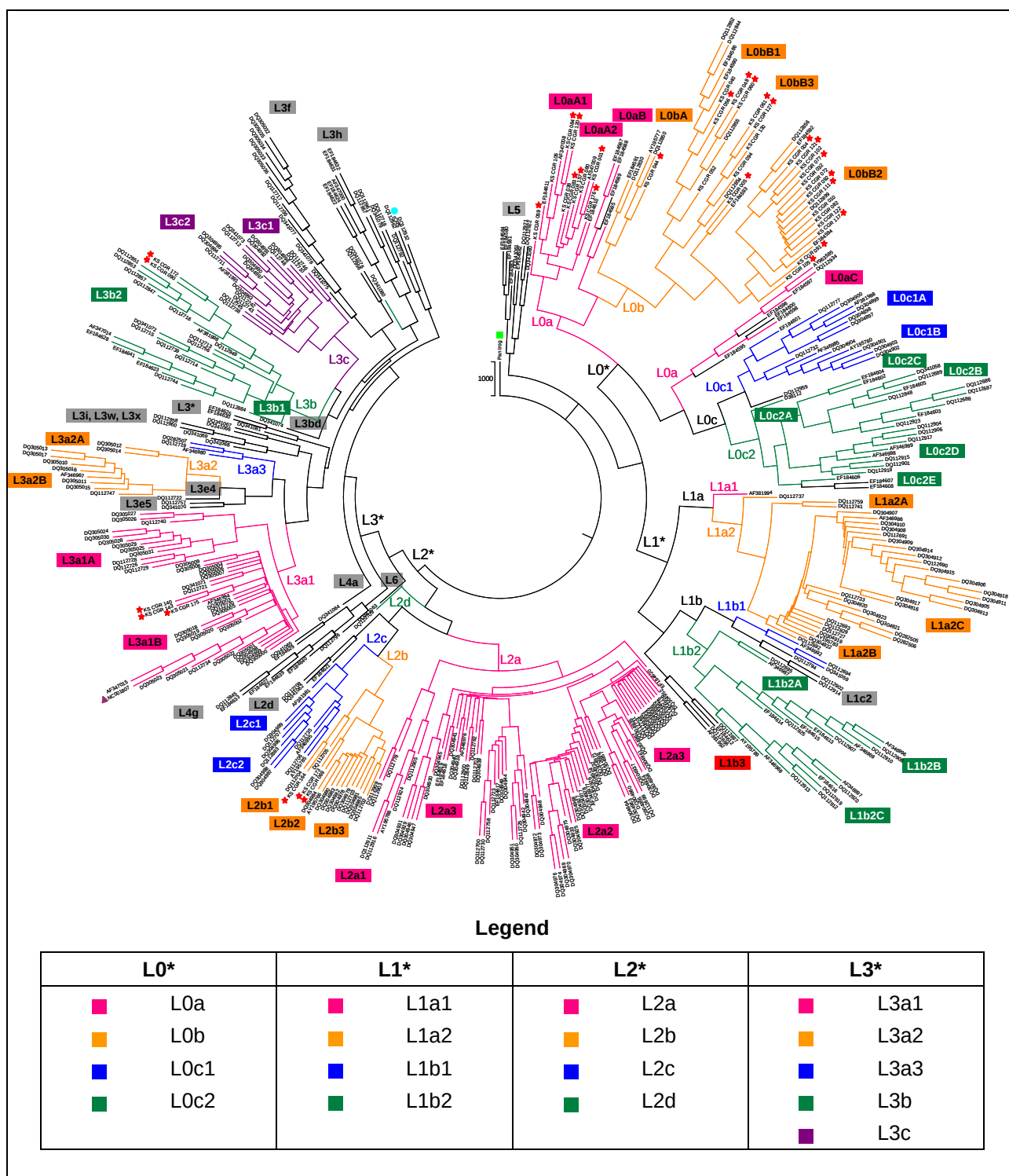
5.9.2 Neighbour-joining trees

N-J trees were constructed as discussed in Sections 4.12.5 and 5.9.1. The N-J method is the method most commonly used and it has the advantage that it is rapidly calculated, generates a unique result (a single tree for each dataset), is able to process large datasets, the assumption of a valid molecular clock is not required and a number of different evolutionary models are available (Page and Holmes, 1998; Salemi and Vandamme, 2003). This method was therefore chosen as it has a number of advantages suited to the datasets under investigation. The size and complexity of the datasets, notably datasets 1a and 2a, necessitated the use of a tree construction method that could handle large and complex datasets (refer to Section 3.2.2.2.1). In addition, a valid molecular clock, or a constant rate of evolution, can not be assumed for mitochondrial DNA

sequences as evidence has been gathered in support of the hypothesis that different lineages evolve at different rates due to the influence of selection (refer to Section 3.2.2.2.1). The choice of the F84 evolutionary model allowed for the incorporation of different estimated parameters, namely the different frequencies of the four nucleotides, different rates of transition and transversion and unequal rates of substitution at different nucleotide sites (refer to Section 3.2.2.1). The N-J method however also has a potential disadvantage, namely the fact that the input order of the sequences influence the tree obtained due to the use of the specific clustering algorithm. This was however countered during the tree construction process by the use of a “jumble option” which randomised the input order of the sequences.

5.9.2.1 Global neighbour-joining tree

Global phylogenetic trees consisting of all macrohaplogroup L sequences collected (refer to Section 4.10) were constructed to determine where in the tree the Khoi-San sequences would group when compared to a set of sequences in which all the main L haplogroups (L0, L1, L2 and L3) were represented. The N-J method was performed on dataset 1a (All L Sequences) which consisted of 429 macrohaplogroup L sequences, the rCRS and the chimpanzee outgroup sequence (refer to Section 4.10.4). The 429 sequences included sequences originating from countries located in Africa, as well as additional macrohaplogroup L sequences from the United States of America (African Americans and Hispanics) and the Dominican Republic (refer to Appendix G). Four sequences obtained from GenBank® were excluded due to the reasons discussed in Section 4.10.1. The dataset contained coding region only sequences. A transition/transversion ratio of 10.911 and a gamma shape parameter value of 0.152 were calculated for this dataset (refer to Table 5.32). PHYLIP version 3.66 (Felsenstein, 1989) was used to generate a phylogenetic tree as described in Section 4.12. The global N-J tree, a majority rule (extended) consensus tree, is presented in Figure 5.42.

Figure 5.42 Global neighbour-joining tree constructed using the All L Sequences dataset (dataset 1a)

This is a majority rule (extended) consensus tree constructed using PHYLIP 3.66 as described in Section 4.12. The tree consists of 429 mitochondrial DNA sequences, the rCRS and the *Pan troglodytes* sequence as outgroup. All the relevant information regarding the sequences used is listed in Appendix G. Statistical support for the observed phylogeny was determined by performing 1,000 bootstrap replicates, and the bootstrap values for the major branches and groups are indicated on an A3 size copy of this tree included in Appendix J. Khoi-San sequences generated in the current investigation are indicated by a red star (★), the rCRS is indicated by a turquoise octagon (●), the outgroup is indicated by a green square (■) and the African reference sequence is indicated by a violet triangle (▲). The major haplogroups are indicated on the figure at the relevant branching points. Sub-haplogroups are indicated at the relevant branching points, and by the use of different colours as listed in the legend. Sub-haplogroups indicated in grey boxes represent sub-haplogroups assigned based on the alternative classification scheme (refer to Section 3.3.1.1), sub-haplogroups indicated in red boxes are new sub-haplogroups which have not previously been named and sub-haplogroups presented in the other coloured boxes are subdivisions identified in previously described sub-haplogroups. Sequences that could not be classified completely are indicated by black lines. Sequences that exhibit a discrepancy between their location in the tree and their haplogroup classification are indicated by dotted lines.

In Figure 5.42 the rCRS is indicated by a **turquoise** octagon (●), the *Pan troglodytes* outgroup is represented by a **green** square (■), the African reference sequence is denoted by a **violet** triangle (▲), and the Khoi-San sequences generated in the current investigation are signified by **red** stars (★). The location of all Khoi-San sequences generated in the current investigation within the global N-J tree is summarised in Table 5.33.

Table 5.33 Location of 30 Khoi-San sequences generated in the current investigation in the global N-J phylogenetic tree

Sequence name	Haplogroup classification ^a	Sequence name	Haplogroup classification	Sequence name	Haplogroup classification
KS_CGR_001	L0a	KS_CGR_077	L0b	KS_CGR_121	L0b
KS_CGR_004	L0b	KS_CGR_083	L0a	KS_CGR_127	L0b
KS_CGR_005	L0b	KS_CGR_084	L0a	KS_CGR_140	L3a1
KS_CGR_044	L0b	KS_CGR_090	L3b	KS_CGR_143	L3a1
KS_CGR_048	L0b	KS_CGR_091	L0b	KS_CGR_157	L0a
KS_CGR_056	L0b	KS_CGR_092	L0b	KS_CGR_164	L2b
KS_CGR_059	L0a	KS_CGR_105	L0b	KS_CGR_172	L3b
KS_CGR_060	L0b	KS_CGR_111	L0b	KS_CGR_173	L2b
KS_CGR_061	L0b	KS_CGR_117	L0b	KS_CGR_175	L3a1
KS_CGR_069	L2b	KS_CGR_120	L0a	KS_CGR_176	L0a

^a = the haplogroup classification assigned to the Khoi-San sequence according to the classification scheme used in the current investigation (refer to Section 4.9). The colours in which the haplogroup/sub-haplogroup classifications are presented correspond to the colours used to indicate the different branches in the global N-J tree.

Due to the size of the dataset (429 sequences) and computational limitations, 1,000 bootstrap replicates were performed (refer to Section 4.12.5), and the bootstrap values obtained for the major branches and groups are indicated on an A3 size copy of this tree presented in Appendix J. The obtained bootstrap values were generally low, likely due to the large dataset investigated (429 sequences), as well as high sequence divergence and rate variability among lineages (refer to Section 5.9.1). The major haplogroups and sub-haplogroups are indicated on the figure at the relevant branching points, and sub-haplogroups are distinguished through the use of different colours as listed in the legend. Sequences indicated by black lines are sequences for which the haplogroup classification was incomplete (refer to Section 4.9), while sequences designated using dotted lines are sequences of which the haplogroup classification was incompatible with the location of the sequence in the phylogenetic tree. A list of these sequences is provided in Table 5.34.

Table 5.34 Incompletely classified sequences and sequences that exhibit a discrepancy between their haplogroup classification and position in the global N-J phylogenetic tree

GenBank [®] accession number	Haplogroup classification			GenBank [®] accession number	Haplogroup classification		
	(Wallace, 2004) ^a	N-J tree ^b	Alternative scheme ^c		(Wallace, 2004) ^a	N-J tree ^b	Alternative scheme ^c
AF381992	L0,L1	L1	L1c3	EF184633	L3b\$	L3	L4g
DQ112696	L0,L1	L1	L1c3	EF184639	L3b\$	L3	L4g
DQ112697	L0,L1	L1	L1c3	EF184640	L3b\$	L3	L4g
EF184612	L0,L1	L1	L1c3	DQ341081	L3b\$	L3	L3*
DQ112921	L0\$a	L5	L5a1	EF184625	L3b\$	L3	L3*
DQ112922	L0\$a	L5	L5a1	EF184630	L3b\$	L3	L3*
DQ341060	L0\$a	L5	L5a1	DQ341074	L3b\$	L3	L3bd
DQ341061	L0\$a	L5	L5a2	DQ112717	L3b\$	L3	L3f1
EF184582	L0\$a	L5	L5a2	DQ112736	L3b\$	L3	L3f1
EF184580	L0\$b	L5	L5b	DQ112956	L3b\$	L3	L3bd
EF184581	L0\$b	L5	L5b	DQ305032	L3b\$	L3	L3f1
EF184583	L0\$b	L5	L5b	DQ305033	L3b\$	L3	L3f1
EF184584	L0\$b	L5	L5b	DQ305034	L3b\$	L3	L3f1
EF184600	L0\$a	L0	L0f	DQ305035	L3b\$	L3	L3f1
DQ112959	L0c2\$	L0c1	L0a	DQ305036	L3b\$	L3	L3f1
DQ112794	L1b2\$	L1b1	L1c2	DQ341075	L3b\$	L3	L3f
DQ112902	L1b2\$	L1b1	L1c2	DQ341076	L3b\$	L3	L3f
DQ112914	L1b2\$	L1b1	L1c2	DQ341077	L3b\$	L3	L3f1
DQ112708	L2d\$	L2	L2d	DQ341078	L3b\$	L3	L3f1
DQ341062	L2d\$	L2	L2d	AF347000	L3b\$	L3	L3h
EF184621	L2d\$	L2	L2d	DQ112718	L3b\$	L3	L3h
DQ341063	L2d\$	L2	L6	DQ112957	L3b\$	L3	L3h
DQ112722	L3a3\$	L3a	L3e4	DQ112961	L3b\$	L3	L3h
DQ112757	L3a3\$	L3a	L3e5	DQ341079	L3b\$	L3	L3h
DQ341070	L3a3\$	L3a	L3e5	EF184622	L3b\$	L3	L3h
DQ112958	L3a3\$	L3	L3i	EF184624	L3b\$	L3	L3h
DQ112960	L3a3\$	L3	L3i	EF184631	L3b\$	L3	L3h
DQ341069	L3a3\$	L3	L3i	EF184632	L3b\$	L3	L3h
DQ341066	L3a3\$	L3	L3x2	DQ112792	L3b\$	H	L3*
DQ341067	L3a3\$	L3	L3x1	DQ112932	L3b\$	H	L3*
DQ341068	L3a3\$	L3	L3w	DQ112962	L3b\$	H	L3*
DQ341064	L3b\$	L3	L4a1	EF184589	L2a	L0	---
DQ112796	L3b\$	L3	L4g	EF184595	L2a	L0	---
DQ112845	L3b\$	L3	L4g	EF184596	L2a	L0	---
DQ341065	L3b\$	L3	L4g	EF184607	L2a	L0c2	---
EF184627	L3b\$	L3	L4g	EF184608	L2a	L0c2	---
EF184629	L3b\$	L3	L4g	---	---	---	---

Incompletely classified sequences are indicated by the presence of a dollar sign (\$) in the haplogroup/sub-haplogroup designation and sequences for which a discrepancy in haplogroup classification exists are indicated in **bold** text. ^a = refers to the haplogroup classification assigned to the sequence according to the classification scheme used in the current investigation (refer to Section 4.9). ^b = refers to the haplogroup classification expected for the sequence based on the clustering of the sequence in the phylogenetic tree. ^c = refers to the haplogroup classification assigned to the sequence based on the alternative classification scheme described in Section 3.3.1.1. The asterisk (*) included in the L3* haplogroup designation is used to indicate that the specific haplogroup consists of a group of mtDNA lineages that form part of the particular haplogroup, although these lineages can not be classified into any of the specific sub-haplogroups of that haplogroup (refer to Section 3.3.1.1). Refer to Appendix G for additional information regarding the sequences listed in this table.

The putative haplogroup classification of the incompletely classified sequences indicated in Table 5.34 was determined by their location in the phylogenetic trees constructed in the current investigation. The haplogroup classification according to the alternative classification scheme (refer to Section 3.3.1.1) presented in Table 5.34 is provided for use in the discussion of groups consisting of incompletely classified sequences in each of the major haplogroups.

Five sequences (EF184589, EF184595, EF184596, EF184607 and EF184608) exhibited a discrepancy between their haplogroup classification and their location within the global N-J tree. These five sequences were classified as L2a according to the classification scheme used in the current investigation (refer to Section 4.9). The SNP profile and the haplogroup classification determined based on the presence of these characteristic SNPs are indicated in Table 5.35.

Table 5.35 SNP profiles of the five sequences that exhibited a discrepancy between their haplogroup classification and the location of the sequence in the global N-J phylogenetic tree

Characteristic SNP	Haplogroup classification ^a
3594T	L0, L1, L2
10810T	L2
11914A	L2a

^a = refers to the haplogroup classification assigned to the sequence according to the classification scheme used in the current investigation (refer to Section 4.9).

The incompatibility between the haplogroup classification assigned and the location of the sequence in the global phylogenetic tree for these five sequences presented in Table 5.34 can be due to reverse mutations, or the presence of sequencing errors at haplogroup informative nucleotide positions (refer to Table 5.35). Sequencing errors can be introduced at different stages in the generation of the final sequence data, namely during PCR, reading of the sequence electropherogram, data capturing, or can be attributed to sequencing artefacts (refer to Section 3.2.1.3). However, due to the fact that all of the sequences were classified as L2a while in fact L0, it is suggested that the classification scheme used was lacking in some way by not making provision for the recently generated sequences (EF184589, EF184595, EF184596, EF184607 and EF184608) obtained from Tanzania in East Africa (Gonder *et al.*, 2007). This conclusion was supported by the subsequent re-classification of these sequences using a revised classification scheme for haplogroup L0 proposed in Section 5.12 which classified all of these sequences as belonging to the L0aB, L0aC or L0c2E sub-haplogroups.

There are several notable characteristics of the N-J tree presented in Figure 5.42. The most basal lineage of the global N-J tree consists of a group of sequences that were incompletely classified as L0\$a or L0\$b, hereafter referred to as L5 sequences (refer to Table 5.34), followed by a group consisting of L0/L1 sequences, and a group containing L2/L3 sequences. This is in contrast to previous studies in which L0 was reported as the most basal haplogroup of the human mtDNA tree (Ingman *et al.*, 2000; Maca-Meyer *et al.*, 2001; Mishmar *et al.*, 2003; Ruiz-Pesini *et al.*, 2004; Kivisild *et al.*, 2006; Gonder *et al.*, 2007; Ruiz-Pesini *et al.*, 2007). This discrepancy could be explained by the use of only African-specific L sequences in the current investigation compared to the use of sequences belonging to African and non-African haplogroups in the other investigations, the lack of L5 sequences in the datasets employed in other investigations, differences in the tree construction methods used, and different sequences used for rooting of the trees. Alternatively, this difference can be attributed to the characteristics of the specific sequences included in the dataset investigated. The second option is considered to be the most likely explanation in the current investigation, and this option is discussed in the next paragraph.

The L5 group observed in the current investigation consisted of nine sequences from East Africa (two from Ethiopia and seven from Tanzania), as well as two sequences from Pygmy individuals from Central Africa (Democratic Republic of the Congo). Using the alternative classification scheme (refer to Section 3.3.1.1), three sequences (DQ112921, DQ112922 and DQ341060) were classified as L5a1, two sequences (DQ341061 and EF184582) were designated L5a2 and four sequences (EF184580, EF184581, EF184583 and EF184584) were assigned to L5b. To date, only a few sequences have been identified that belong to the L5 haplogroup (Pereira *et al.*, 2001; Kivisild *et al.*, 2004; Stevanovitch *et al.*, 2004; Kivisild *et al.*, 2006; Torroni *et al.*, 2006; Gonder *et al.*, 2007) and all of these sequences for which coding region sequence data were available were included in the phylogenetic trees constructed in the current investigation. The L5 haplogroup appears to be rare and is observed at low frequencies in countries mainly in eastern Africa such as Mozambique (Pereira *et al.*, 2001), Ethiopia (Torroni *et al.*, 2006) and Tanzania (Gonder *et al.*, 2007), as well as in North Africa in Egypt (Stevanovitch *et al.*, 2004), and in Central Africa among the Mbuti Pygmies (Kivisild *et al.*, 2006). In a schematic diagram representing the evolutionary history of mtDNA haplogroups based on mtDNA D-loop and RFLP analysis, L5 is placed between L1 and L2 (Gonder *et al.*, 2007), indicating a closer relationship between L1 and L5 than between L0 and L5. To date, due to the rarity of sequences belonging to the L5 haplogroup and the paucity of sequence data from eastern

Africa, it is possible that the true evolutionary relationship of L5 with other L haplogroups has been obscured by this lack of sequence data, i.e. haplogroup L5 consists of lineages that are more closely related to L0 than to L1. The inclusion of all available L5 sequences as well as additional L0 sequences in the large dataset employed in the current investigation may support this previously unreported relationship. The addition of more sequences belonging to haplogroup L5, as well as more sequences from East Africa, to the phylogenetic tree should provide further elucidation of this close relationship between L0 and L5, and could lend further support to the replacement of L0 by L5 as the most basal group as observed in the current investigation. Due to the presence of haplogroup L5 in mainly East Africa, which has been proposed as the birth place for modern humans based on paleobiological and archaeological data (McBrearty and Brooks 2000; White *et al.*, 2003) and the location from which the earliest migrations of humans out of Africa took place (Tishkoff *et al.*, 1996; Quintana-Murci *et al.*, 1999; Kivisild *et al.*, 2004), it can be speculated that the observation of L5 as the most basal group in the current investigation instead of L0, can be attributed to the fact that this haplogroup represents a very ancient haplogroup, perhaps even more ancient than L0.

The L0* haplogroup consisted of two main groups, the first of which contained sequences classified as L0a and L0b, and the second consisted of sequences classified as L0a, L0c1 and L0c2. The presence of L0a sequences in both of these two groups was unexpected and presented a problem pertaining to the classification scheme, i.e. sequences classified within the same sub-haplogroup were expected to cluster together in the global N-J tree. This discrepancy between haplogroup assignment and location in the phylogenetic tree was addressed after the construction of a global L0-specific haplogroup network (refer to Section 5.10) by proposing the construction of a revised classification scheme for the L0 haplogroup (refer to Section 5.12). Three classifications were proposed, L0aA for the majority of L0a sequences (14 sequences), L0aB for the L0a sequences grouped closest to the L0b group (four sequences), and L0aC for the L0a sequences grouped with L0c1 and L0c2 (seven sequences).

Twenty-two of the 30 Khoi-San sequences generated in the current investigation grouped within the L0* group. Of the 22 sequences, seven (KS_CGR_001, KS_CGR_059, KS_CGR_083, KS_CGR_084, KS_CGR_120, KS_CGR_157 and KS_CGR_176) were classified as L0aA. The seven Khoi-San sequences in this group clustered with three additional Khoi-San sequences (KS_CGR_038, KS_CGR_080 and KS_CGR_106) and four San sequences (AF347008, AF347009, EF184610 and EF184611). The Khoi-San

sequences generated in the current investigation expanded this branch within L0* and it appears that two groups are present within the L0aA group, one group consisting of KS_CGR_059 and the other group containing the remaining 13 sequences. The second group appears to be divided into two groups, one containing two sequences and the other the remaining 11 sequences. These two groups are further discussed in Section 5.9.3.1.

Three L0a sequences (EF184585, EF184587 and EF184588) were classified as L0aB. A fourth sequence clustered within this group. This sequence, EF184589, was initially classified as L2a (refer to Table 5.34), and later re-classified as L0aB. The L0aB group consisted of sequences from Tanzania in East Africa, and was first described by Gonder *et al.* in 2007 which classified the sequences in this group as Tanzanian L0d (L0dTZ). Although the L0aB group is closely related to the L0aA group consisting of Khoi-San and San individuals from southern Africa, it appears to be more closely related to the L0b sub-haplogroup. It appears that the L0aB group may be divided into two groups, L0aB1 containing EF184587, EF184588 and EF184589, and L0aB2 consisting of EF184585. The L0aB group and its subdivisions are discussed in more detail in Section 5.10.2.1.

Four sequences (AY963585, DQ112934, EF184597 and EF184598) were classified as L0aC. Three additional sequences formed part of this group. EF184600 was incompletely classified as L0\$a (refer to Table 5.34) and its classification was revised based on the location of the sequence in the phylogenetic tree. EF184595 and EF184596 were initially classified as L2a (refer to Table 5.34), and were subsequently re-classified as L0aC. It appears that the L0aC group may be subdivided into two groups, one containing a single sequence (EF184595) and the other the remaining six sequences. The L0aC group is discussed further in Section 5.10.2.1.

The remaining 15 Khoi-San sequences were classified as L0b. The L0b group is divided into two groups, L0bA containing one Khoi-San sequence (KS_CGR_044), and L0bB consisting of the remaining 14 Khoi-San sequences. L0bB is further divided into another two groups, which are each subdivided into two more groups. The addition of the 15 Khoi-San sequences to the L0b sub-haplogroup has therefore increased the branching in this group. The L0bB group and its subdivisions (L0bB1, L0bB2, L0bB2a, L0bB2b and L0bB3) are discussed further in Section 5.10.2.1. L0bA and L0bB correspond to two sub-haplogroups (L0d2 and L0d1) that have been previously recognised within the L0b (L0d) group according to the alternative classification scheme (refer to Section 3.3.1.1

and Appendix F). The Khoi-San sequences generated in the current investigation gave rise to further branching within the L0bB group. It is likely that with the addition of more L0b sequences to this group, the subdivisions representing the four sub-haplogroups (L0bB1, L0bB2a, L0bB2b and L0bB3) observed in the current investigation will be confirmed.

None of the Khoi-San sequences generated in the current investigation were classified as either L0c1 or L0c2. The L0c1 group appeared to be divided into two groups, L0c1A which consisted of one sequence (DQ112959) and L0c1B that contained the remaining sequences. DQ112959 was incompletely classified as L0c2\$ based on its SNP profile (3594T, 10810C, 7055A + 11914A, 9755A, 9818T, 5096T + 5147G) in which a characteristic SNP for L0c2 classification (5147A) was absent (refer to Section 4.9, Table 4.8). However, based on its location in the tree, its haplogroup classification was revised to L0c1 (refer to Table 5.34). The L0c1A and L0c1B groups are discussed further in Section 5.9.3.1.

The L0c2 group appeared to be divided into two groups. The first group was composed of one sequence (D38112) and the second group consisted of 22 sequences. Further subdivisions were observed in the second group, and although bootstrap values for these groups were low, it is suggested that these groups represent four possible sub-haplogroups. Two sequences (EF184607 and EF184608) were previously classified as L2a (refer to Table 5.34). Both sequences were re-classified as L0c2E using the revised classification scheme (refer to Section 5.12). The L0c2 group and its subdivisions are discussed further in Section 5.9.3.1.

The focus of the current investigation regarding the constructed phylogenetic trees was the L0 haplogroup, as this was the haplogroup to which the majority of the Khoi-San sequences generated in the current investigation belonged. The other haplogroups and the branching patterns observed within these haplogroups are however also briefly discussed in the following paragraphs.

The L1* haplogroup consisted of two main groups, the first containing sequences classified as L1a1 and L1a2, and the second consisting of sequences classified as L1b1 and L1b2. Only one sequence (AF381994) from Mauritania in West Africa was classified as L1a1 (refer to Appendix G), and it appears that this sub-haplogroup is rare as it was observed only once in a dataset containing 429 sequences. The L1a2 group appeared to

be subdivided into three groups. The first group (L1a2A) contained two or three sequences, DQ112741 and DQ112759, with the inclusion or exclusion of DQ112737 in this group dependent on its specific grouping within the global N-J and MP trees, respectively (refer to Figure 5.45, Section 5.9.3.1). The second group (L1a2B) consisted of 13 sequences (AY195783, DQ112692, DQ112693, DQ112727, DQ112733, DQ112829, DQ282505, DQ282506, DQ304919, DQ304920, DQ304921, DQ304922 and DQ304923) and the third group (L1a2C) contained the remaining 17 sequences. These three groups represent possible sub-haplogroups within the L1a2 sub-haplogroup. The L1a2 sub-haplogroup corresponds to the L1b1 sub-haplogroup in the alternative haplogroup classification scheme (refer to Appendix F) and to date one sub-haplogroup (L1b1a) has been described for this sub-haplogroup (refer to Section 3.3.1.1).

The sequences within sub-haplogroup L1b1 grouped with three incompletely classified sequences (DQ112794, DQ112902 and DQ112914). The haplogroup classification of these sequences was revised according to their position in the phylogenetic tree (refer to Table 5.34). Based on the alternative classification scheme, these three sequences were classified as L1c2 (refer to Section 3.3.1.1). The L1b2 group consisted of 19 sequences and it appeared to be divided into three smaller groups. The first group (L1b2A) consisted of two sequences (AF346987 and DQ112695), the second group (L1b2B) contained nine sequences (AF346969, AF346996, DQ112907, DQ112908, DQ112910, DQ112925, EF184613, EF184614 and EF184615) and the third group (L1b2C) was composed of the remaining eight sequences. The L1b1 sub-haplogroup is equivalent to the L1c2 sub-haplogroup in the alternative classification scheme (refer to Appendix F) and to date only one sub-haplogroup has been described for this sub-haplogroup (refer to Section 3.3.1.1). The L1b2 sub-haplogroup corresponds to the L1c1 sub-haplogroup in the alternative classification scheme (refer to Appendix F) and to date three sub-haplogroups have been identified for this sub-haplogroup (refer to Section 3.3.1.1).

A group consisting of four sequences (AF381992, DQ112696, DQ112697 and EF184612) incompletely classified as L0, L1 was present in the L1* group. The haplogroup classification of these sequences was revised according to their position in the phylogenetic tree (refer to Table 5.34). These sequences were classified as L1c3 according to the alternative classification scheme (refer to Section 3.3.1.1). At present there is no corresponding haplogroup designation for this group in the classification scheme used in the current investigation, and the use of L1b3 is suggested to identify this group.

From the above discussion of the L1* haplogroup, it can be concluded that this section of the classification scheme used in the current investigation (refer to Section 4.9) should be revised to include new sub-haplogroups such as the three groups observed in the L1a2 group, and to provide a way to classify the groups consisting of unclassified sequences by incorporating the haplogroup assignments already provided by the alternative classification scheme.

The L2* haplogroup consisted of three main groups, the first containing sequences classified as L2a, the second consisting of sequences classified as L2b and L2c, and the third consisting of a single sequence designated L2d. The L2a group consisted of a total of 82 sequences and could be subdivided into three smaller groups consisting of eight, one and 73 sequences, respectively. It appears that the largest group consisting of 73 sequences may be further subdivided into three smaller groups containing 29, 19 and 25 sequences, respectively. A difference was observed between the global N-J and MP trees regarding the location of a single sequence (EF184620). In the global MP tree, this sequence groups within the group consisting of 73 sequences, and in the global MP tree four groups are therefore observed containing eight (L2a1), 19 (L2a2), 25 (L2a3) and 30 sequences (L2a3), respectively. These four groups may represent previously unnamed sub-haplogroups. The L2a sub-haplogroup corresponds to the L2a sub-haplogroup described in the alternative classification scheme (refer to Appendix F). To date only two sub-haplogroups have been described for L2a (refer to Section 3.3.1.1).

The L2b group contained 16 sequences and was subdivided into three smaller groups (L2b1, L2b2 and L2b3) which may represent additional sub-haplogroups. The L2b sub-haplogroup is equivalent to the L2b sub-haplogroup described in the alternative classification scheme (refer to Appendix F). To date only one sub-haplogroup has been described for L2b in the alternative classification scheme (refer to Section 3.3.1.1). Three Khoi-San sequences (KS_CGR_069, KS_CGR_164 and KS_CGR_173) were classified as L2b. One Khoi-San sequence (KS_CGR_164) grouped with two sequences (DQ112704 and DQ112705) that were obtained from individuals from the Dominican Republic. The remaining two Khoi-San sequences (KS_CGR_069 and KS_CGR_173) formed a separate group. It is possible that this Khoi-San only group may represent a sub-haplogroup of L2 that is present mainly in southern Africa (refer to Section 5.9.4). However, in the global MP tree (refer to Figure 5.45, Section 5.9.3.1), the two Khoi-San sequences do not form a separate group, as they cluster with four other sequences (DQ112702, DQ112703, DQ304981 and DQ304985).

The L2c group was composed of 11 sequences which were subdivided into two smaller groups. One group (L2c1) consisted of one sequence (AF381981) from Mauritania in West Africa. The other group (L2c2) consisted of the remaining ten sequences and was further subdivided to give rise to two groups, one consisting of four sequences and the other of six. These groups may represent sub-haplogroups of L2c. The L2c sub-haplogroup corresponds to the L2c sub-haplogroup identified in the alternative classification scheme (refer to Appendix F). Two sub-haplogroups have been described for L2c in the alternative classification scheme (refer to Section 3.3.1.1).

Three sequences (DQ112708, DQ341062 and EF184621) clustered within the group containing the L2b and L2c sequences. The three sequences were incompletely classified as L2d\$ (refer to Table 5.34) based on their SNP profile (3594T, 10810T, 11914G, 4767A, 15849C + 13708G) in which a characteristic SNP for L2d classification (13708A) was not present (refer to Section 4.9, Table 4.8). Based on the alternative classification scheme, the sequences in this group were classified as L2d (refer to Section 3.3.1.1).

The L2d group contained only one sequence (DQ112709). This sequence was obtained from an individual of the Dominican Republic (refer to Appendix G). It appears that this sub-haplogroup is rare as it was observed only once in a dataset consisting of 429 sequences. Based on the preceding discussion of the L2* haplogroup, it is suggested that this section of the classification scheme used in the current investigation (refer to Section 4.9) should be revised to include the new sub-haplogroups identified. This can be done by incorporating the haplogroup assignments for groups consisting of incompletely classified sequences already provided for by the alternative classification scheme.

A single sequence (DQ341063) was positioned adjacent to the L2* group. This sequence was incompletely classified as L2d\$ (refer to Table 5.34). However, when using the alternative classification scheme, this sequence was assigned to the L6 haplogroup. A group consisting of eight sequences (DQ112796, DQ112845, DQ341065, EF184627, EF184629, EF184633, EF184639 and EF184640) incompletely classified as L3b\$ (refer to Table 5.34) was observed next to the L3* group. These sequences were classified as L4g according to the alternative classification scheme (refer to Section 3.3.1.1). A single sequence (DQ341064) incompletely classified as L3b\$ (refer to Table 5.34) was located adjacent to the L4g group. This sequence was however designated as L4a according to the alternative classification scheme (refer to Section 3.3.1.1). As stated before, the classification scheme used in the current investigation (refer to Section 4.9) should be

revised to include the haplogroups assigned to the incompletely classified single sequences, or groups containing such sequences, as provided by the alternative classification scheme.

The L3* haplogroup was the biggest group observed in the global phylogenetic trees and consisted of a number of closely related groups. The bootstrap values in this region of the tree were generally lower than in other regions of the phylogenetic tree, indicating that this region of the tree was not completely resolved. This could be due to the rapid divergence of lineages from one another to form new sub-haplogroups. Such a pattern is consistent with a population expansion event and the evidence for such an event in this haplogroup will be discussed in Section 5.11.4.

The L3a1 group consisted of 40 sequences which included three Khoi-San sequences (KS_CGR_140, KS_CGR_143 and KS_CGR_175), as well as the African reference sequence (NC_001807). The L3a1 group was divided into two groups, one group (L3a1A) contained 12 sequences and the other group (L3a1B) was composed of the remaining 28 sequences. The L3a1 group corresponds to the L3e2 group in the alternative classification scheme (refer to Appendix F) for which two sub-haplogroups have been described (refer to Section 3.3.1.1). The three Khoi-San sequences increased the branching within the larger L3a1 subgroup by forming a Khoi-San only branch. It is possible that this Khoi-San group may represent a lineage of L3a1 that is localised to southern Africa (refer to Section 5.9.4). The African reference sequence has been compiled from archival data retrieved from GenBank® and is curated on an ongoing basis by collaborating groups and the National Center for Biotechnology Information (NCBI) staff to provide a consensus of the generated primary information. The aim is to provide a single sequence that represents a standard version of the mitochondrial genome for Africans to be used for mutation analysis and polymorphism discovery (Pruitt *et al.*, 2007). NC_001807 is classified within the L3a1 haplogroup (refer to Section 4.9). Although this sequence is undoubtedly useful as a reference for Africans in comparison to sequences from individuals belonging to non-African haplogroups, the information contained in more ancient lineages, such as L0, is not represented in this sequence. The consensus sequence compiled in the current investigation (refer to Section 5.13) is based on Khoi-San sequences which were classified as L0a or L0b and is therefore based on a more ancient African lineage (L0 vs. L3). This consensus sequence should be useful in comparisons with sequences from African haplogroups and non-African haplogroups as this consensus sequence should provide more information regarding the age (ancient or

recent) and population-specific nature (specific to African populations or not) of observed polymorphisms (refer to Section 5.13).

The L3a2 group consisted of ten sequences and was subdivided into two groups. The first group (L3a2A) contained two sequences (DQ305012 and DQ305014), and the second group (L3a2B) was composed of the remaining eight sequences. These groups may represent sub-haplogroups within the L3a2 sub-haplogroup. The L3a2 group corresponds to the L3e3 group in the alternative classification scheme (refer to Appendix F), however, no sub-haplogroups have been described for this group to date (refer to Section 3.3.1.1). A single sequence (DQ112722) was observed next to the L3a2 group. This sequence was incompletely classified as L3a3\$ (refer to Table 5.34). According to the alternative classification scheme, this sequence was classified as L3e4 (refer to Section 3.3.1.1). Another group composed of two sequences (DQ112757 and DQ341070) was observed adjacent to the single sequence. These sequences were also incompletely classified as L3a3\$ (refer to Table 5.34), and were classified as L3e5 according to the alternative classification scheme (refer to Section 3.3.1.1).

The L3a3 group was composed of three sequences (AF346980, DQ112719 and DQ282507). This group is equivalent to the L3e1 group described in the alternative classification scheme (refer to Appendix F). As can be observed in Figure 5.42, the L3a group consisted of three sub-haplogroups (L3a1, L3a2 and L3a3), as well as a single sequence and a group of two sequences which may represent additional sub-haplogroups. It therefore appears that there may be five sub-haplogroups within the L3a group. The L3a group corresponds to the L3e group in the alternative classification scheme (refer to Appendix F) for which five sub-haplogroups (L3e1, L3e2, L3e3, L3e4 and L3e5) have been described to date (refer to Section 3.3.1.1).

A group consisting of six incompletely classified sequences was located next to the L3a group. This group was subdivided into two groups. The first group contained two sequences (DQ341066 and DQ341067), while the second group consisted of four sequences (DQ112958, DQ112960, DQ341068 and DQ341069). The second group formed two groups, the first containing one sequence (DQ341068), and the second, the remaining three sequences. All six sequences were incompletely classified as L3a3\$ (refer to Table 5.34). According to the alternative classification scheme, two sequences (DQ341066 and DQ341067) belonged to the L3x group, three sequences (DQ112958, DQ112960 and DQ341069) were designated L3i and one sequence (DQ341068) was

assigned to the L3w group (refer to Section 3.3.1.1). The alternative classification corresponds with the grouping observed within this group of six incompletely classified sequences. The first group of two sequences (DQ341066 and DQ341067) were classified as L3x, while in the second group, the group of three sequences (DQ112958, DQ112960 and DQ341069) were classified as L3i and the remaining single sequence (DQ341068) as L3w.

A group composed of three incompletely classified sequences (DQ341081, EF184625 and EF184630) was observed next to the group containing six incompletely classified sequences. These sequences were incompletely classified as L3b\$ (refer to Table 5.34), and could not be classified beyond the L3* haplogroup when using the alternative classification scheme (refer to Section 3.3.1.1). All three of these sequences were obtained from individuals from East Africa (Ethiopia and Tanzania). In the global MP tree (refer to Figure 5.45), this group of three sequences is located at a different position within the tree and groups closely to a group containing the rCRS. The difficulties experienced regarding haplogroup classification of the three sequences, the different location of the three sequences in the two phylogenetic trees and the generally low bootstrap values observed in the L3* group in the two phylogenetic trees, suggest that the branching of these three sequences within the L3 haplogroup has not been resolved. Although it is possible that these three sequences represent a previously undescribed L3 sub-haplogroup, more sequences obtained from individuals from East Africa are required to resolve the branching within the L3* group and to confirm if these three sequences represent a new sub-haplogroup.

A single sequence (DQ341074) was located adjacent to the L3b group. This sequence was incompletely classified as L3b\$ (refer to Table 5.34) and was classified as L3bd according to the alternative classification scheme (refer to Section 3.3.1.1). In the alternative classification scheme L3bd is a precursor to L3b and L3d which are equivalent to L3c and L3b respectively in the classification scheme used in the current investigation (refer to Appendix F).

The L3b group consisted of 20 sequences and was subdivided into two groups. The first group (L3b1) contained one sequence (DQ112884) and the second group (L3b2) contained the remaining 19 sequences. The second group was divided into two groups. The first group consisted of nine sequences (AF381998, DQ112716, DQ112847, DQ112851, DQ112853, DQ112857, DQ112949, KS_CGR_172 and KS_CGR_090) and

the second group contained 11 sequences (AF347014, DQ112713, DQ112714, DQ112715, DQ112739, DQ112744, DQ112756, DQ341072, EF184623, EF184628 and EF184641). The location of DQ112949 in the L3b2 group is different in the global N-J tree compared to the global MP tree (refer to Figure 5.45). In the global MP tree the two groups consist of eight and 12 sequences, respectively, with DQ112949 located in the second group instead of the first group. The group of 11 (N-J) or 12 (MP) sequences was further divided into two groups, the first containing two sequences (DQ112713 and DQ112756) or three sequences (DQ112713, DQ112756 and DQ112949), and the second consisting of the remaining sequences. It therefore appears that the L3b group can be subdivided into two sub-haplogroups, of which one sub-haplogroup can be further subdivided into three more sub-haplogroups. The L3b group corresponds to the L3d group described in the alternative classification scheme (refer to Appendix F) for which three sub-haplogroups have been described to date (refer to Section 3.3.1.1). The two Khoi-San sequences observed in the L3b group increased the branching within this group by creating a Khoi-San only branch. It is possible that these sequences form part of a group representing an L3b lineage that is confined to southern Africa (refer to Section 5.9.4).

The L3c group was composed of 19 sequences and was divided into two smaller groups. The first group (L3c1) contained two sequences (DQ112710 and DQ112746) and the second group (L3c2) consisted of the remaining 17 sequences. It appears that the second group is further divided into two groups, one containing two sequences (DQ112738 and DQ112745) and the other composed of the remaining 15 sequences. The L3c group is equivalent to the L3b group described in the alternative classification scheme (refer to Appendix F). To date, two sub-haplogroups have been identified for this group (refer to Section 3.3.1.1).

A group consisting of 12 incompletely classified sequences was observed adjacent to the L3c group. This group was divided into two groups. The first group was composed of two sequences (DQ112956 and DQ341076), and the second group contained the remaining ten sequences. The sequences were incompletely classified as L3b\$ (refer to Table 5.34). Using the alternative classification scheme (refer to Section 3.3.1.1), the sequences were assigned to either the L3f or L3f1 sub-haplogroups. It is possible that this group may be divided into two smaller groups representing two sub-haplogroups. However, more sequences belonging to both the L3f and the L3f1 sub-haplogroups should be added to the phylogenetic tree to confirm this division.

A group composed of one L3b sequence (DQ341080) and nine incompletely classified sequences was observed next to the group of 12 incompletely classified sequences. The nine sequences were incompletely classified as L3b\$ (refer to Table 5.34). Based on the alternative classification scheme (refer to Section 3.3.1.1), these sequences were classified as L3h. The inclusion of the single L3b sequence in this group consisting of mainly L3b\$ sequences can be ascribed to the presence of two transitions, T7861C and G9575A, which define the L3h sub-haplogroup, i.e. the L3b sequence is more closely related to sequences classified as L3h according to the alternative classification scheme (refer to Section 3.3.1.1) than to other sequences designated as L3b based on the classification scheme employed in the current investigation (refer to Section 4.9).

The last group observed within the L3* group consisted of three incompletely classified sequences and the rCRS. This group was divided into two groups, the first contained one sequence (DQ112932) and the second group contained three sequences (rCRS, DQ112792 and DQ112962). All three incompletely classified sequences were classified as L3b\$ (refer to Table 5.34). Using the alternative classification scheme (refer to Section 3.3.1.1), the sequences were assigned to L3*. Both of these classifications were inconsistent with the haplogroup classification as suggested by the location of the sequences in the phylogenetic trees (refer to Table 5.34). Upon further investigation it was determined that DQ112932, DQ112792 and DQ112962 belonged to the U1a, pre-HV1 and H1 haplogroups respectively (Kivisild *et al.*, 2006). The initial classification of these three sequences as L3* can be explained by the fact that haplogroup L3 gave rise to macrohaplogroup N from which haplogroups U1a, pre-HV1 and H1 arose (refer to Section 3.3, Figure 3.5). This implies that some SNPs will be shared between sequences from haplogroup L3 and sequences from haplogroups that diverged from haplogroup L3.

The above discussion of the L3* haplogroup can be summarised by concluding that this section of the classification scheme used in the current investigation (refer to Section 4.9) should be revised to include any new sub-haplogroups observed in the L3 haplogroup or sub-haplogroups, and the designations provided by the alternative classification scheme for groups consisting of unclassified sequences should be incorporated in order to be able to classify these sequences which are currently incompletely classified.

5.9.2.2 Neighbour-joining tree of African individuals

Africa phylogenetic trees containing macrohaplogroup L sequences collected from individuals residing in Africa at the time of the sample collection were constructed to determine where in the trees the Khoi-San sequences would group when compared to a set of sequences in which all the main L haplogroups (L0, L1, L2 and L3) observed currently in Africa were represented. The construction of both global and Africa trees were performed to determine if there would be a difference in the tree topology if sequences obtained from individuals with African ancestry living outside of Africa were included, i.e. does the sequence data obtained from groups with African ancestry such as African Americans reflect the variation observed in Africa and can sequences from these groups be viewed as representative of African populations. This is an important consideration as population groups with African ancestry are often used as substitutes for African populations when samples can not be obtained from African countries due to reasons as diverse as insufficient funding or the presence of civil wars. The N-J method was performed on dataset 2a (All Africa L Sequences) which consisted of 242 sequences obtained from individuals of African origin reported to be residing in countries located in Africa, the rCRS and the chimpanzee outgroup sequence (refer to Section 4.10.4). The 242 sequences included sequences from a number of different countries and ethnic groups, of which the relevant information is listed in Table 5.36.

Table 5.36 African country of origin and ethnic affiliation of sequences included in dataset 2a

Country in Africa	Number of sequences	Ethnic/language group	Reference
Africa ^a	1	Ambo	Kivisild <i>et al.</i> , 2006
	1	Bamileke	Ingman <i>et al.</i> , 2000
	4	Biaka	Ingman <i>et al.</i> , 2000 Kivisild <i>et al.</i> , 2006
	2	Effik	Ingman <i>et al.</i> , 2000
	1	Ewondo	Ingman <i>et al.</i> , 2000
	1	Hausa	Ingman <i>et al.</i> , 2000
	3	Herero	Kivisild <i>et al.</i> , 2006
	2	Ibo	Ingman <i>et al.</i> , 2000
	3	Khoi-San	Kivisild <i>et al.</i> , 2006
	1	Kikuyu	Ingman <i>et al.</i> , 2000
	1	Lisongo ^c	Ingman <i>et al.</i> , 2000
	2	Lissongo ^c	Kivisild <i>et al.</i> , 2006
	3	Mandenka	Ingman <i>et al.</i> , 2000 Kivisild <i>et al.</i> , 2006

Table 5.36 Continued...

Country in Africa	Number of sequences	Ethnic/language group	Reference
Africa ^a	2	Mbenzele	Ingman <i>et al.</i> , 2000
	2	Mbuti	Ingman <i>et al.</i> , 2000
	1	Mkamba	Ingman <i>et al.</i> , 2000
	1	NaSan	Kivisild <i>et al.</i> , 2006
	1	Pedi	Kivisild <i>et al.</i> , 2006
	2	San	Ingman <i>et al.</i> , 2000
	1	Sotho	Kivisild <i>et al.</i> , 2006
	2	Tswana	Kivisild <i>et al.</i> , 2006
	1	Tuareg	Kivisild <i>et al.</i> , 2006
	3	Yoruba	Ingman <i>et al.</i> , 2000
	2	Zulu	Kivisild <i>et al.</i> , 2006
	8	Unreported	Mishmar <i>et al.</i> , 2003
Southern Africa ^b	1	!Kung	Current investigation
Angola	32	!Kung	Current investigation Mouton, 2003
Burkina Faso	10	Foulbe	Kivisild <i>et al.</i> , 2006
	10	Mossi	
	10	Rimaibe	
Cameroon	4	Bakola	Gonder <i>et al.</i> , 2007
Central African Republic	8	Pygmy	Kivisild <i>et al.</i> , 2006
Eritrea	1	Unreported	Kivisild <i>et al.</i> , 2006
Ethiopia	1	Berta	Kivisild <i>et al.</i> , 2006
	15	Unreported	Kivisild <i>et al.</i> , 2006 Torroni <i>et al.</i> , 2006
Ghana	1	Unreported	Kivisild <i>et al.</i> , 2006
Mauritania	4	Unreported	Maca-Meyer <i>et al.</i> , 2001
Morocco	1	Unreported	Maca-Meyer <i>et al.</i> , 2001
Namibia	9	!Kung	Current investigation Mouton, 2003
	2	San	Kivisild <i>et al.</i> , 2006
Nigeria	3	Unreported	Torroni <i>et al.</i> , 2006
South Africa	1	!Kung	Mouton, 2003
	7	San	Gonder <i>et al.</i> , 2007
Sudan	2	Nuba	Kivisild <i>et al.</i> , 2006
	7	Unreported	
Tanzania	1	Akie	Gonder <i>et al.</i> , 2007
	4	Burunge	
	4	Burunge	
	1	Gogo	
	6	Hadza	
	2	Iraqw	
	1	Masaai	
	2	Mbugwe	

Table 5.36 Continued...

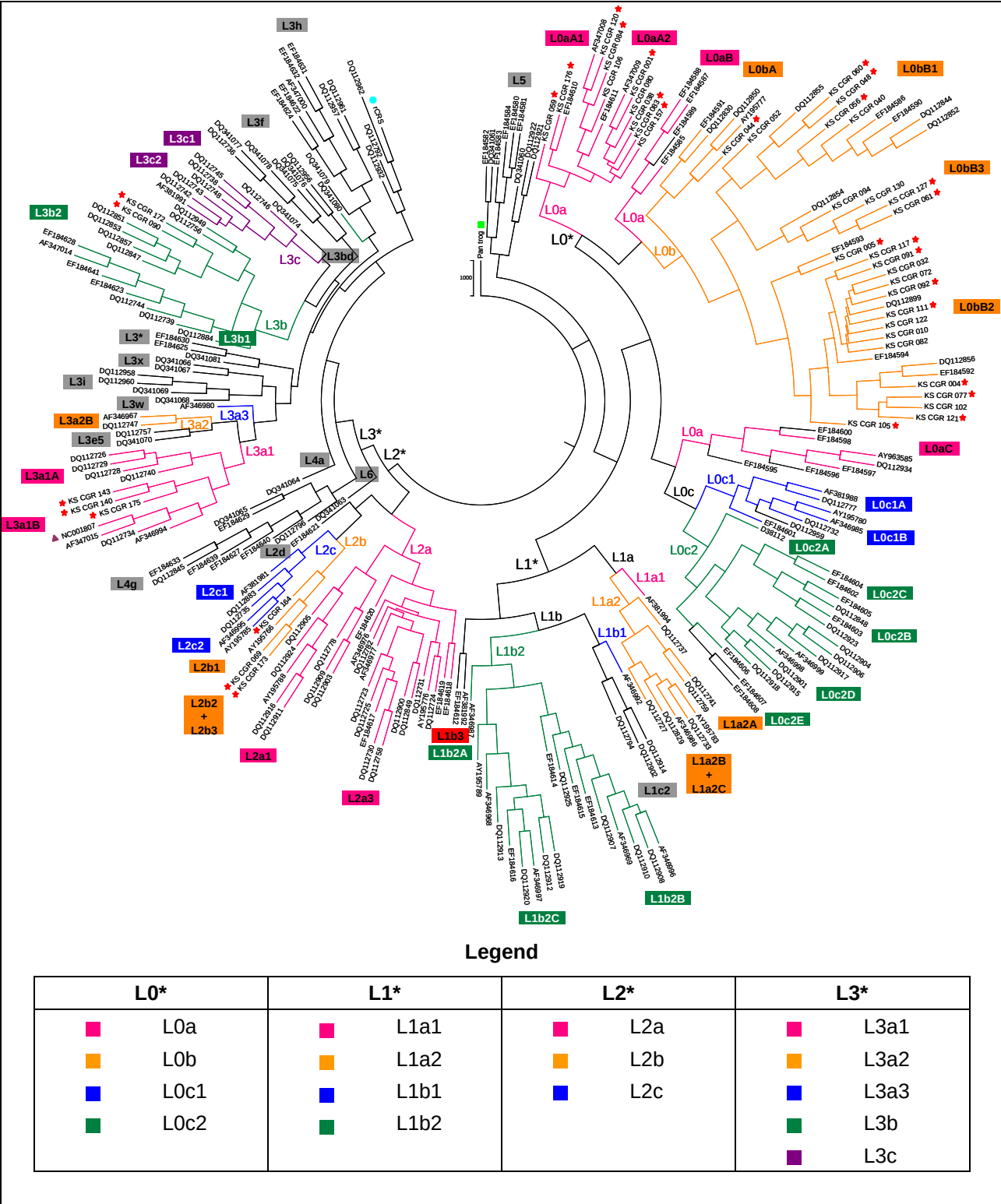
Country in Africa	Number of sequences	Ethnic/language group	Reference
Tanzania	1	Pare	Gonder <i>et al.</i> , 2007
	1	Rangi	
	13	Sandawe	
	8	Turu	
	1	Wafiome	
Uganda	2	Unreported	Horai <i>et al.</i> , 1995 Macaulay <i>et al.</i> , 2005
Zaire	15	Pygmy	Kivisild <i>et al.</i> , 2006

^a and ^b = information regarding country of origin was not available. ^c = the spelling of the ethnic group is indicated as it has been reported in the relevant reference. However, according to Gordon (2005), it appears that these different spellings describe the same ethnic group. Refer to Appendix G for additional information regarding the sequences.

Four sequences obtained from GenBank® were excluded due to the reasons summarised in Section 4.10.1. The dataset contained coding region only sequences. A transition/transversion ratio of 9.054 and a gamma shape parameter value of 0.131 were calculated for this dataset (refer to Table 5.32) and PHYLIP version 3.66 (Felsenstein, 1989) was used to construct a phylogenetic tree as described in Section 4.12. The Africa N-J tree, a majority rule (extended) consensus tree, is presented in Figure 5.43.

The Africa N-J tree presented in Figure 5.43 follows the same conventions as the global N-J tree depicted in Figure 5.42 regarding the way in which sequences of interest (Khoi-San sequences, rCRS, *Pan troglodytes* and the African reference sequence) are indicated, as well as the manner in which haplogroup and sub-haplogroups are identified through the use of different colours. The location of all Khoi-San sequences determined in the current investigation within the Africa N-J tree is as summarised in Table 5.33 and a list of incompletely classified sequences, including their haplogroup classifications, is provided in Table 5.34. As before, 1,000 bootstrap replicates were performed (refer to Section 4.12.5) and the obtained bootstrap values for the major branches and groups are indicated on an A3 size copy of this tree included in Appendix J. The observed bootstrap values were generally low, although comparable to the values observed in the global N-J tree, which is likely due to the size and complexity of the dataset (refer to Section 5.9.1). In the following paragraphs, the Africa N-J tree is discussed briefly as details regarding the branching observed within the major haplogroups, the identification of subdivisions within groups that represent potential sub-haplogroups, as well as the corresponding designations in the alternative classification scheme for any identified sub-haplogroups have already been discussed in Section 5.9.2.1.

Figure 5.43 Africa neighbour-joining tree constructed using the All Africa L Sequences dataset (dataset 2a)



This is a majority rule (extended) consensus tree constructed using PHYLIP 3.66 as described in Section 4.12. The tree consists of 242 mitochondrial DNA sequences, the rCRS and the *Pan troglodytes* sequence as outgroup. All the relevant information regarding the sequences used is listed in Appendix G. Statistical support for the observed phylogeny was determined by performing 1,000 bootstrap replicates, and the bootstrap values for the major branches and groups are indicated on an A3 size copy of this tree included in Appendix J. Khoi-San sequences generated in the current investigation are indicated by a red star (★), the rCRS is indicated by a turquoise octagon (●), the outgroup is indicated by a green square (■) and the African reference sequence is indicated by a violet triangle (▲). The major haplogroups are indicated on the figure at the relevant branching points. Sub-haplogroups are indicated at the relevant branching points, and by the use of different colours as listed in the legend. Sub-haplogroups indicated in grey boxes represent sub-haplogroups assigned based on the alternative classification scheme (refer to Section 3.3.1.1), sub-haplogroups indicated in red boxes are new sub-haplogroups which have not previously been named and sub-haplogroups presented in the other coloured boxes are subdivisions identified in previously described sub-haplogroups. Sequences that could not be classified completely are indicated by black lines. Sequences that exhibit a discrepancy between their location in the tree and their haplogroup classification are indicated by dotted lines.

As observed in the global N-J tree, the most basal lineage of the Africa N-J tree consisted of a group of L5 sequences. The observation of L5 sequences at this position in the phylogenetic tree was discussed in Section 5.9.2.1. The L5 group was subdivided into two groups. In the Africa N-J tree, the sequences in the L5a2 sub-haplogroup cluster with the L5b sequences, instead of the L5a1 sequences. This discrepancy may be attributed to differences in the composition of the two datasets, i.e. the two datasets contain different numbers of sequences from the different haplogroups and sub-haplogroups, or due to unresolved branching structure within the L5 haplogroup as suggested by the low bootstrap value for this branch. The addition of more L5 sequences to the datasets should provide better resolution.

The L0* haplogroup was composed of two main groups, the first consisting of L0a and L0b sequences, and the second containing L0a, L0c1 and L0c2 sequences. The presence of L0a sequences in both of these two groups was previously observed in the global N-J tree and was addressed in Section 5.9.2.1 by proposing three new haplogroup designations (L0aA, L0aB and L0aC).

The same branching structure was observed within the L0* section of the Africa N-J tree as in the global N-J tree (refer to Section 5.9.2.1). Seven of the Khoi-San sequences generated in the current investigation were classified as L0aA. The L0aA group was divided into two groups, one group was composed of KS_CGR_059 and the other group contained six Khoi-San sequences (KS_CGR_001, KS_CGR_083, KS_CGR_084, KS_CGR_120, KS_CGR_157 and KS_CGR_176). The second group appeared to be further subdivided into two groups with one group containing KS_CGR_176 and the other the five remaining Khoi-San sequences. The two groups are discussed further in Section 5.9.3.2. The L0aB group contained four sequences (EF184567, EF184585, EF184588 and EF184589) and appeared to be divided into two groups, L0aB1 and L0aB2. The L0aC group consisted of seven sequences with a possible further subdivision into two groups. Fifteen Khoi-San sequences generated in the current investigation were classified as L0b. The L0b group was divided into a number of sub-haplogroups (L0bB1, L0bB2, L0bB2a, L0bB2b and L0bB3). The L0c1 group appeared to be divided into two groups, the first of which consisted of one sequence (EF184601) and the second which contained the remaining sequences. A difference was observed between the global N-J tree and the Africa N-J tree. In the global tree, DQ112959 formed a group on its own, while in the Africa tree it was included in the second subgroup of L0c1 consisting of six sequences. This observation, coupled with the relatively low bootstrap values observed for branches in

L0c1, suggests that the branching structure within this sub-haplogroup of L0 has not been fully resolved. The L0c2 group was divided into two groups. The first group was composed of one sequence (D38112) and the second group consisted of 17 sequences for which further subdivisions were proposed as discussed in Section 5.9.2.1.

The Africa N-J tree exhibited the same branching structure within the L1* section of the tree as in the global N-J tree (refer to Section 5.9.2.1). The L1* haplogroup consisted of two main groups. The first group contained sequences classified as L1a1 and L1a2, and the second group consisted of sequences classified as L1b1 and L1b2. Only one sequence (AF381994) was classified as L1a1. The L1a2 group appeared to be subdivided into two groups. The first group (L1a2A) contained one sequence (DQ112737) and the second group (L1a2B + L1a2C) consisted of seven sequences. The L1b1 group consisted of a single sequence (AF346992). A group consisting of three incompletely classified sequences (DQ112794, DQ112902 and DQ112914) was observed adjacent to this single sequence. According to the alternative classification scheme, these sequences were classified as belonging to the L1c2 haplogroup (refer to Section 3.3.1.1). The L1b2 group was divided into three smaller groups. The first group (L1b2A) consisted of one sequence (AF346987), the second group (L1b2B) was composed of nine sequences (AF346969, AF346996, DQ112907, DQ112908, DQ112910, DQ112925, EF184613, EF184614 and EF184615) and the third group (L1b2C) contained the remaining eight sequences. A group consisting of two incompletely classified sequences (AF381992 and DQ112696) was located next to the L1b2 group. This group appears to represent a new sub-haplogroup of L1 and the use of L1b3 is suggested to identify this group.

The L2* section of the Africa N-J tree was similar to that of the global N-J tree (refer to Section 5.9.2.1). The L2* haplogroup contained two main groups. The first group consisted of sequences classified as L2a and the second group contained sequences classified as L2b and L2c. A third group composed of a single sequence designated L2d was observed in the global N-J tree, in contrast this group was absent in the Africa N-J tree. The L2a group was divided into three groups consisting of eight, one (EF184620) and 15 sequences, respectively. In the Africa MP tree (refer to Figure 5.46), the L2a group was divided into three groups containing eight, 11 and five sequences, respectively. This difference can be ascribed to the difference in location of EF184620 in the two trees. In the Africa N-J tree this sequence formed a group on its own, while in the MP tree it was located within the group consisting of 11 sequences. Three groups namely L2a1, L2a2 and L2a3, were observed in the global N-J tree. In the Africa N-J tree, the group containing

eight sequences corresponds to the L2a1 group, while the two groups consisting of 11 and five sequences each, together form a group corresponding to the L2a3 group. The L2b group consisted of four sequences which included three Khoi-San sequences (KS_CGR_069, KS_CGR_164 and KS_CGR_173). The L2c group contained five sequences and was divided into two groups (L2c1 and L2c2), one of which contained a single sequence (AF381981). A single incompletely classified sequence (EF184621) was observed within the group containing the L2b and L2c sequences. This sequence formed part of a group of incompletely classified sequences observed in the global N-J tree that were assigned to the L2d haplogroup based on the use of the alternative classification scheme (refer to Section 3.3.1.1). Another incompletely classified single sequence (DQ341063) was located next to the L2* group. This sequence was assigned to the L6 haplogroup according to the alternative classification scheme (refer to Section 3.3.1.1).

A group consisting of eight incompletely classified sequences was observed next to the L3* group. These sequences were classified as L4g based on the alternative classification scheme (refer to Section 3.3.1.1). A single sequence (DQ341064) was located next to the group consisting of L4g sequences and this sequence was assigned to the L4a haplogroup according to the alternative classification scheme (refer to Section 3.3.1.1).

The branching structure observed within the L3* section of the Africa N-J tree was similar to that observed in the corresponding section of the global N-J tree (refer to Section 5.9.2.1) and low bootstrap values for some branches suggested that this section of the tree was not fully resolved. The L3a1 group contained three Khoi-San sequences (KS_CGR_140, KS_CGR_143 and KS_CGR_175), as well as the African reference sequence (NC_001807). The L3a1 group was divided into two groups, the first group (L3a1A) was composed of four sequences and the other group (L3a1B) consisted of seven sequences. It appeared that the three Khoi-San sequences formed a Khoi-San only branch and it is possible that this Khoi-San group may represent a lineage of L3a1 that is localised to southern Africa (refer to Table 5.37). The L3a2 group consisted of two sequences and in contrast to the global N-J tree no further subdivision was observed within this group. A group containing two incompletely classified sequences (DQ112757 and DQ341070) was observed adjacent to the L3a2 group. Based on the alternative classification scheme (refer to Section 3.3.1.1), these sequences were classified as L3e5 and may represent a sub-haplogroup of L3a. The L3a3 group consisted of a single sequence (AF346980). A group consisting of six incompletely classified sequences was

observed next to the L3a group. This group was subdivided into two groups containing two sequences (DQ341066 and DQ341067) and four sequences, respectively. The two sequences (DQ341066 and DQ341067) were classified as L3x, and the four sequences were classified as L3i (DQ112958, DQ112960 and DQ341069) and L3w (DQ341068) according to the alternative classification scheme (refer to Section 3.3.1.1). A group composed of three incompletely classified sequences (DQ341081, EF184625 and EF184630) was observed adjacent to the group containing six sequences. In the Africa MP tree (refer to Figure 5.46), this group of three sequences are observed in close proximity of a group containing the rCRS and it was suggested that these three sequences represent a previously undescribed L3 sub-haplogroup (refer to Section 5.9.2.1).

The L3b group contained 15 sequences and was divided into two groups, the first group (L3b1) was composed of one sequence (DQ112884) and the other group (L3b2) contained the remaining 14 sequences. The second group was further subdivided into three groups. The first group consisted of two sequences (DQ112756 and DQ112949), the second contained six sequences, including two Khoi-San sequences (KS_CGR_090 and KS_CGR_172) and the third was composed of six sequences. The location of DQ112756 is different in the Africa N-J tree compared to the global N-J tree (refer to Figure 5.42). In the Africa N-J tree this sequence groups with DQ112949, close to the group containing the two Khoi-San sequences, while in the global N-J tree this sequence groups with DQ112713 to form a subgroup within the group which does not contain Khoi-San sequences. This discrepancy in the location of DQ112756 is due to the inclusion of DQ112713 in the global N-J tree as the two sequences appear to be closely related. It therefore appears that the L3b group can be subdivided into two sub-haplogroups, of which one sub-haplogroup can be further subdivided into three more sub-haplogroups.

The L3c group was composed of seven sequences and was divided into two smaller groups. The first group (L3c1) contained one sequence (DQ112746) and the second group (L3c2) consisted of the remaining six sequences. The second group was further subdivided into two groups, one containing two sequences (DQ112738 and DQ112745) and the other was composed of the remaining four sequences. A single incompletely classified sequence (DQ341074) was located adjacent to the L3c group. However, in the global N-J tree this sequence was positioned outside of both the L3b and L3c groups. According to the alternative classification scheme (refer to Section 3.3.1.1), this sequence was classified as L3bd. A group containing six incompletely classified sequences was observed next to the L3c group. This group was divided into two subgroups. The first

group contained two sequences (DQ112956 and DQ341076), and the second group consisted of the remaining four sequences. Based on the alternative classification scheme (refer to Section 3.3.1.1), the six sequences were classified as L3f or L3f1. A group containing one L3b and eight incompletely classified sequences was observed adjacent to the group consisting of six sequences. According to the alternative classification scheme (refer to Section 3.3.1.1), these sequences were assigned to the L3h haplogroup. The last group of sequences observed within the L3* group consisted of three incompletely classified sequences and the rCRS. This group was further divided into two groups, the first contained one sequence (DQ112932) and the second group contained three sequences (rCRS, DQ112792 and DQ112962).

From the above data it can be concluded that the topology of the Africa N-J tree was similar to that observed in the global N-J tree, i.e. the same branching order was observed in each of the sections of the trees representing the major haplogroups and sub-haplogroups. Potential subdivisions observed within sub-haplogroups in the Africa N-J tree were confirmed by observations made regarding the global N-J tree, i.e. the resolution within branches was improved by the addition of more sequences. It was observed that only one sub-haplogroup, L2d, was not represented in the Africa N-J tree. This can however be attributed to the rarity of this haplogroup (refer to Section 5.9.2.1). Upon comparison of the global and Africa N-J trees, it was also observed that the sequences obtained from individuals of African ancestry residing in countries outside of Africa (refer to Appendix G) were scattered throughout the global N-J tree. These sequences were not confined to specific haplogroups and sub-haplogroups, although they were more prevalent in younger haplogroups such as L2 and L3 and were absent in older haplogroups such as L5, L0a and L0b (refer to Appendix G). It is possible that the distribution of African mitochondrial haplogroups amongst the sequences contained in the Africa and global N-J trees could be due to the sampling procedure followed (random vs. a sample of convenience) or to historical events such as the slave trade which transported African individuals from specific parts of Africa, e.g. West Africa (refer to Section 3.1.4). Datasets consisting of sequences obtained from individuals of African ancestry in countries such as Brazil, the Dominican Republic or the United States of America can however still be employed to investigate the mitochondrial haplogroups and sub-haplogroups specific to Africa and to draw inferences about African mitochondrial haplogroups with the proviso that the data presented might not be representative of all known African mitochondrial haplogroups. The use of sequences obtained from individuals of African descent residing in Africa, as well as from individuals residing in countries outside of Africa is however

advocated as it provides the best resolution in macrohaplogroup L phylogenetic trees. This is especially applicable when older haplogroups such as L0 are investigated, or when the investigation requires the inclusion of rare haplogroups such as L2d (Dominican Republic) and L1a1 (Mauritania) which have to date only been observed outside or inside of Africa, respectively.

As already discussed, the groups consisting of incompletely classified sequences could be assigned a haplogroup designation based on the alternative classification scheme (refer to Section 5.9.2.1). This supports the revision of the haplogroup classification scheme used in the current investigation to include the designations provided by the alternative classification scheme for groups consisting of unclassified sequences, as well as new sub-haplogroups observed in any of the previously described sub-haplogroups.

5.9.2.3 Neighbour-joining tree of Khoi-San individuals

The relationship between the mitochondrial genome sequences generated for the Khoi-San individuals studied in the current investigation can be investigated using the presence or absence of certain alterations (refer to Sections 5.7 and 5.8.1). Based on such sequence alterations, a previous investigation using 13 individuals from the Khoi-San sample group described in Section 4.2 identified three groups within the Southern African Khoi-San population (Mouton, 2003). In the current investigation phylogenetic analyses were employed to investigate whether these three groups remain intact when more sequences from Khoi-San individuals are added to the phylogenetic tree, and to determine the specific clustering of Southern African Khoi-San individuals.

The N-J method was performed on dataset 3a and dataset 3b containing Khoi-San sequences only, the rCRS and the chimpanzee outgroup sequence (refer to Section 4.10.4). Dataset 3a contained 58 coding region only Khoi-San sequences which included 30 !Kung sequences generated in the current investigation, 13 !Kung sequences generated previously in the CGR (Mouton, 2003), 12 San sequences (Ingman *et al.*, 2000; Kivisild *et al.*, 2006; Gonder *et al.*, 2007) and three Khoi-San sequences (Kivisild *et al.*, 2006). Dataset 3b consisted of 55 complete Khoi-San sequences and included 43 !Kung sequences (Mouton, 2003; current investigation) and ten San sequences (Ingman *et al.*, 2000; Gonder *et al.*, 2007). A single San sequence, EF184609 (Gonder *et al.*, 2007), was excluded from both datasets as this sequence contained an A at nucleotide position 3,594 which prevented its classification as a macrohaplogroup L sequence (refer to

Section 4.10.1). A transition/transversion ratio of 10.406 and a gamma shape parameter value of 0.016 were calculated for dataset 3a, while a transition/transversion ratio of 6.957 and a gamma shape parameter value of 0.017 were calculated for dataset 3b (refer to Table 5.32). The differences in the two values for the two datasets can be ascribed to the inclusion of the non-coding or control region in dataset 3b. The control region contains high levels of homoplasy (Howell *et al.*, 2004) and a number of hypervariable sites, i.e. sites which are prone to repeated forward and reverse alterations (refer to Section 2.5.1). This could lead to an increased number of transversions compared to transitions due to a reduction in selective constraint on transversions in the non-coding region. This would decrease the transition/transversion ratio, and alter the gamma shape parameter which is calculated based on the transition/transversion ratio. PHYLIP version 3.66 (Felsenstein, 1989) was used to generate phylogenetic trees for datasets 3a and 3b as described in Sections 4.12 and 5.9.1.

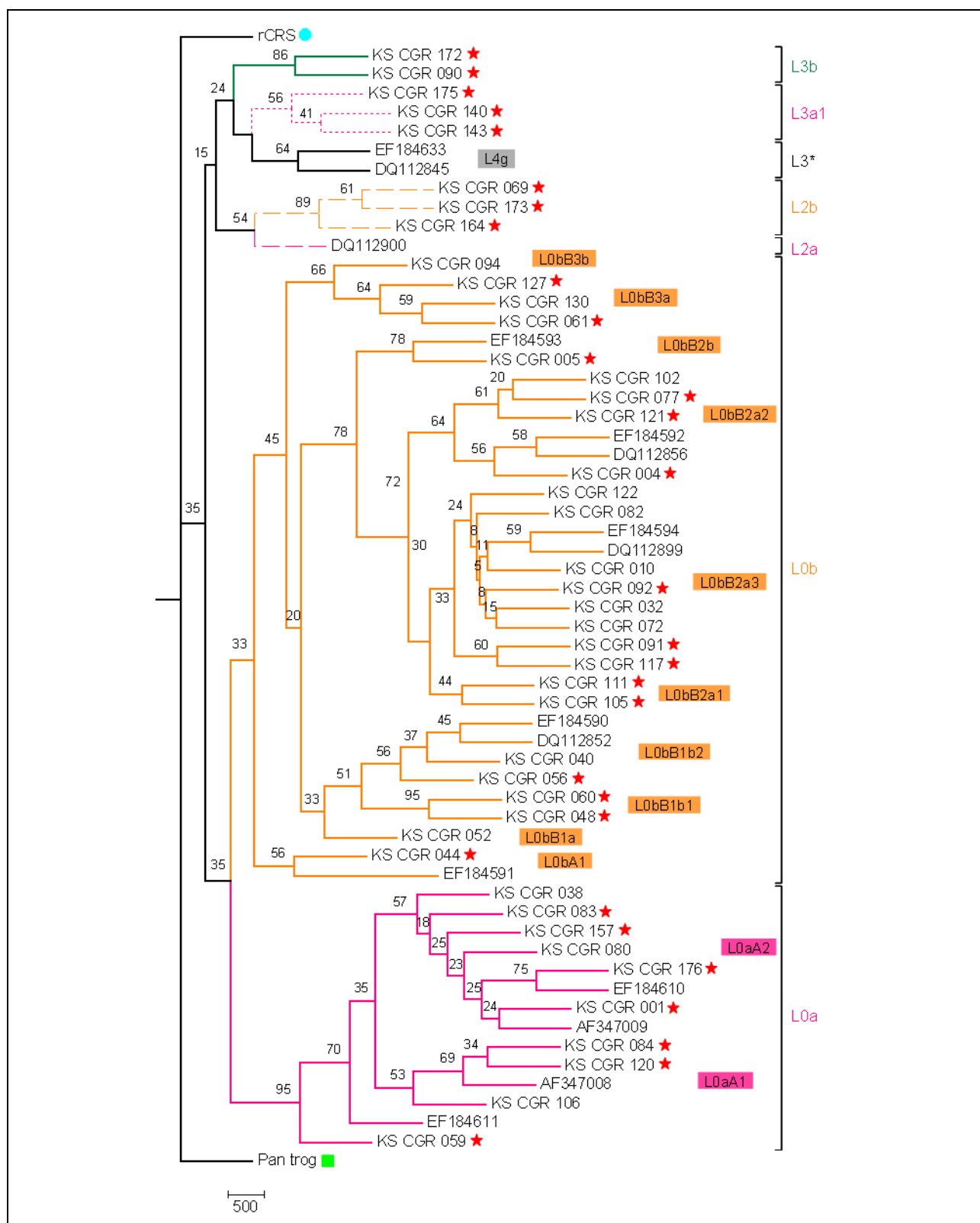
The N-J trees constructed from datasets 3a (coding region sequences) and 3b (complete sequences) were compared. The two trees exhibited the same general topology with comparable bootstrap values for the observed groups. Three differences were however observed between the two trees relating to the location of the rCRS, the subdivision of the L0aA group and the position of the EF184633 sequence. These differences between the two trees corresponded to the differences observed between the Khoi-San N-J and MP trees constructed using dataset 3a (coding region sequences) which are discussed in the following paragraphs.

The differences between the N-J trees constructed using datasets 3a and 3b can be attributed to the tree construction method employed. The sequences included in dataset 3a and 3b were all obtained from Southern African Khoi-San language speakers which included Khoi-San and San individuals, whom are therefore likely to be closely related. The N-J method converts character differences data into evolutionary distance data and does not take into account the actual nucleotide differences that separate the sequences from one another, which leads to a loss of information (Page and Holmes, 1998). In a dataset consisting of closely related sequences, it is likely that this loss of data will obscure some relationships leading to some sequences grouping where they are not expected. This was confirmed by a comparison of the N-J tree based on dataset 3b with the MP trees based on datasets 3a and 3b. The Khoi-San N-J tree based on complete sequences corresponds to both of the MP trees regarding the position of the rCRS, the grouping of sequences in the L0aA group and the location of the EF184633 sequence. For a dataset

consisting of Khoi-San and San sequences, it therefore appears that complete sequences provide better resolution of the branching structure when the N-J method is used due to the additional information contained in the control region. However, only the Khoi-San N-J tree based on dataset 3a (coding region sequences) is presented and discussed as this tree is comparable to the global and Africa N-J trees generated using coding region only sequence data.

The Khoi-San N-J tree presented in Figure 5.44 follows the same conventions as the global N-J tree depicted in Figure 5.42 regarding the way in which sequences of interest (Khoi-San sequences, rCRS and *Pan troglodytes*) are indicated, as well as the manner in which haplogroup and sub-haplogroups are identified through the use of different colours. The location of all Khoi-San sequences generated in the current investigation within the Khoi-San N-J tree is as presented in Table 5.33. Information regarding incompletely classified sequences, including their haplogroup classifications, is provided in Table 5.34. Bootstrap analysis (1,000 replicates) was performed (refer to Section 4.12.5) and the obtained bootstrap values for the major branches and groups are presented. The observed bootstrap values were generally low which suggested that the branching was not completely resolved in some groups due to either the close relationship of the sequences or a lack of sequences. This is supported by the fact that the bootstrap values obtained for the corresponding groups in the global and Africa N-J trees were generally higher. The majority of the Khoi-San sequences included in this tree belonged to the L0 haplogroup. Details regarding the branching structure within the L0 haplogroup and the potential sub-haplogroups identified within this haplogroup have already been discussed in Section 5.9.2.1.

Figure 5.44 Khoi-San neighbour-joining tree constructed using the All Khoi-San L Sequences dataset (dataset 3a)



This is a majority rule (extended) consensus tree constructed using PHYLIP 3.66 as described in Section 4.12. The tree consists of 58 mitochondrial DNA sequences, the rCRS and the *Pan troglodytes* sequence as the outgroup. All the relevant information regarding the sequences used is listed in Appendix G. Statistical support for the observed phylogeny was determined by performing 1,000 bootstrap replicates, and the bootstrap values for the major branches and groups are indicated. Khoi-San sequences generated in the current investigation are indicated by a red star (★), the rCRS is indicated by a turquoise octagon (●) and the outgroup is indicated by a green square (■). Sub-haplogroups are indicated by brackets on the right hand side of the figure and by the use of different colours and different types of lines (solid, dashed or dotted lines). The sub-haplogroups indicated in a grey box represents a sub-haplogroup assigned based on the alternative classification scheme (refer to Section 3.3.1.1), and sub-haplogroups presented in coloured boxes are subdivisions identified in previously described sub-haplogroups. Sequences that could not be classified completely are indicated by black lines.

There are several noteworthy features of the Khoi-San N-J tree presented in Figure 5.44. It appears that three groups can be distinguished, a group consisting of L0a and L0b sequences, a group containing L2 and L3 sequences as well as two incompletely classified sequences, and a group composed of the rCRS. This is in contrast to the Khoi-San MP tree (refer to Figure 5.47). In the Khoi-San MP tree, only two groups are observed that consist of L0a and L0b sequences, and L2 and L3 sequences as well as the rCRS, respectively. The difference in the location of the rCRS can be attributed to the difference between the two tree construction methods. In the N-J tree, evolutionary distances are used to determine the relationships between sequences. The majority of sequences in the Khoi-San N-J tree are more similar to each other than the rCRS, leading to the rCRS being positioned outside of the two groups. In the Khoi-San MP tree character differences are used to define the relationships between sequences, leading to the rCRS grouping within the L3 group as the L3 haplogroup shares certain SNPs with the H haplogroup to which the rCRS belongs. The grouping of the rCRS in the Khoi-San MP tree (refer to Figure 5.47) is also supported by observations made in the global and Africa N-J and MP trees where the rCRS groups closer to L3 sequences, than to the *Pan troglodytes* sequence or the L0 haplogroup.

Thirty Khoi-San sequences were generated in the current investigation. Twenty-two sequences grouped within the L0 haplogroup, three sequences were located in the L2 haplogroup and five sequences were positioned in the L3 haplogroup. The L0 haplogroup consisted of two main groups, the first containing sequences classified as L0a and the second consisting of sequences classified as L0b. The L0a group consisted of ten Khoi-San sequences (KS_CGR_001, KS_CGR_038, KS_CGR_059, KS_CGR_080, KS_CGR_083, KS_CGR_084, KS_CGR_106, KS_CGR_120, KS_CGR_157 and KS_CGR_176) and four San sequences (AF347008, AF347009, EF184610 and EF184611). All of these L0a sequences were classified as L0aA according to the revised classification scheme (refer to Section 5.12). The L0aA group was further subdivided into two groups, one consisting of one sequence (KS_CGR_059) and the other containing the remaining 13 sequences. The second group appeared to be subdivided into two groups, one containing EF184611 and the other the remaining 12 sequences. The L0aA group is further discussed in Section 5.9.3.3.

The L0b group consisted of 26 Khoi-San sequences (DQ112899, KS_CGR_004, KS_CGR_005, KS_CGR_010, KS_CGR_032, KS_CGR_040, KS_CGR_044, KS_CGR_048, KS_CGR_052, KS_CGR_056, KS_CGR_060, KS_CGR_061,

KS_CGR_072, KS_CGR_077, KS_CGR_082, KS_CGR_091, KS_CGR_092, KS_CGR_094, KS_CGR_102, KS_CGR_105, KS_CGR_111, KS_CGR_117, KS_CGR_121, KS_CGR_122, KS_CGR_127 and KS_CGR_130) and seven San sequences (DQ112852, DQ112856, EF184590, EF184591, EF184592, EF184593 and EF184594). Based on the revised classification scheme (refer to Section 5.12), the L0b group was subdivided into two groups. L0bA contained two sequences (EF184591 and KS_CGR_044) and L0bB consisted of the remaining 31 sequences. In the global and Africa N-J trees, the L0bA group was subdivided into two groups. This difference is due to the number of sequences in the group as only two sequences were present in the Khoi-San N-J tree compared to four sequences in the global and Africa N-J trees. L0bB was further divided into another three groups (L0bB1, L0bB2 and L0bB3), which were each subdivided into more groups. The L0bA and L0bB groups and their subdivisions are discussed further in Section 5.9.3.3.

The remaining Khoi-San sequences were observed in the group consisting of L2 and L3 sequences. Three Khoi-San sequences (KS_CGR_069, KS_CGR_164 and KS_CGR_173) were classified as L2b, one Khoi-San sequence (DQ112900) was assigned to the L2a haplogroup, three Khoi-San sequences (KS_CGR_140, KS_CGR_143 and KS_CGR_175) were designated L3a1 and two Khoi-San sequences (KS_CGR_090 and KS_CGR_172) belonged to the L3b sub-haplogroup. As already mentioned, it was expected that all the Khoi-San sequences generated in the current investigation would be classified within the L0 haplogroup, however, the likely reasons for the presence of haplogroups other than L0 in Khoi-San individuals have been discussed (refer to Section 5.8.1). Two sequences (DQ112845 and EF184633), incompletely classified as L3b\$ (refer to Table 5.34), were observed in the group consisting of L2 and L3 sequences. These two sequences were classified as L4g according to the alternative classification scheme (refer to Section 3.3.1.1). A difference was observed regarding this group of sequences between the Khoi-San N-J and MP trees (refer to Figure 5.47). In the N-J tree, these two sequences are positioned within the L2/L3 sequences group, and cluster close to the L3a1 sequences. In the MP tree, the L3a1 and L3b sequences group together, and the two sequences form a separate group. This difference can be ascribed to unresolved branching within the L2/L3 group as indicated by the low bootstrap values observed in the N-J tree compared to the MP tree.

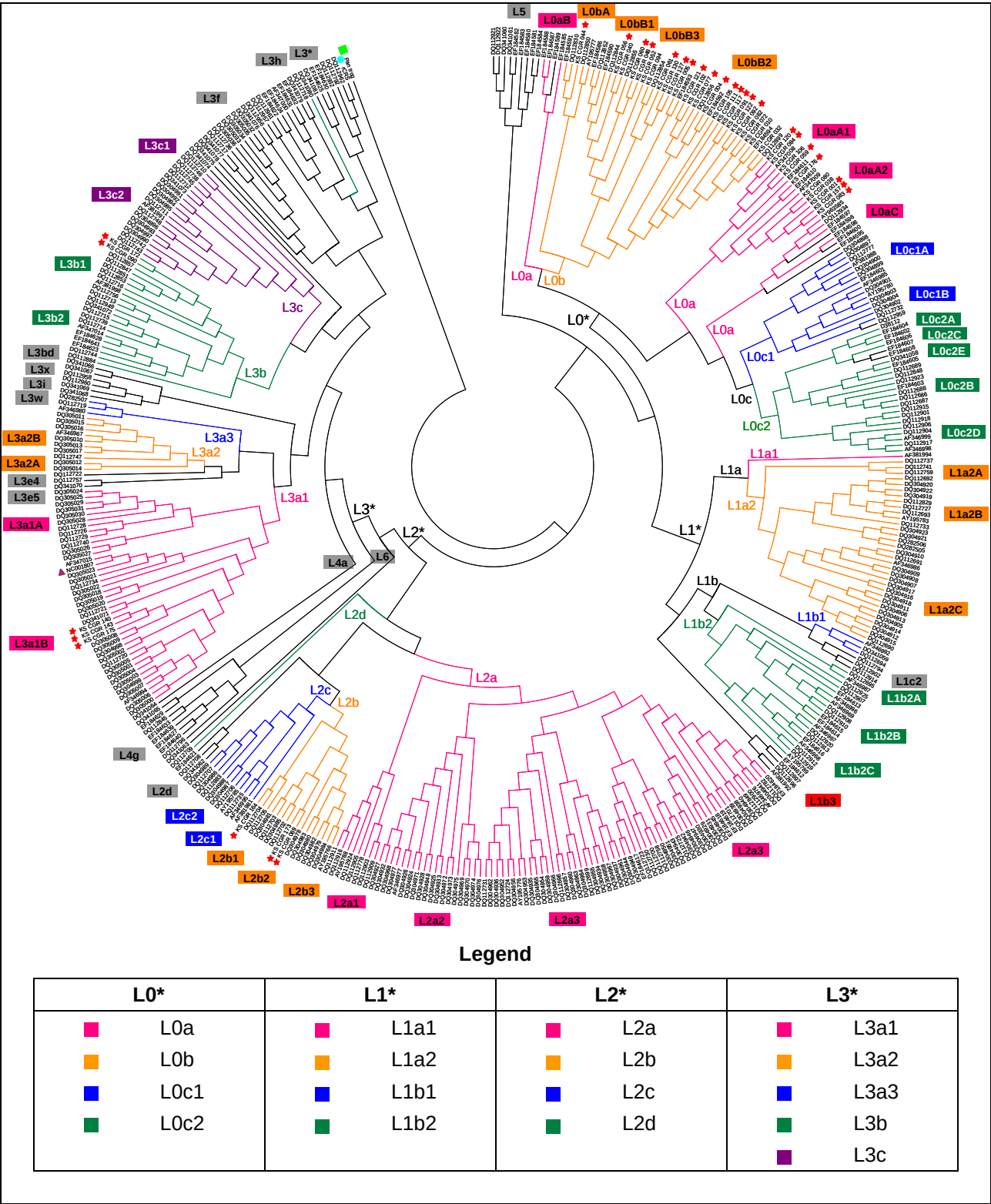
5.9.3 Maximum parsimony trees

MP trees were constructed as discussed in Sections 4.12.6 and 5.9.1. This method was chosen due to the fact that it has a number of advantages suited to the datasets under investigation (refer to Section 3.2.2.2.2). These advantages include the fact that few assumptions regarding the evolutionary process are required when the MP method is employed. A constant rate of evolution, i.e. a molecular clock (refer to Section 5.9.2) is not assumed, and information is not lost by converting sequence data to distance data, i.e. the observed groupings are due to character or nucleotide differences between sequences. The most significant disadvantages of this method are that, although powerful software programs are available, it remains computationally expensive, requiring prolonged computing periods and large amounts of processing power. Another disadvantage is the occurrence of long branch attraction which is usually due to the analysis of very divergent sequences. This was however countered by the fact that the sequences analysed all belonged to macrohaplogroup L, and although considerable variation is present within this group, it remains a single group that shares the same common African ancestry.

5.9.3.1 Global maximum parsimony tree

The MP method was performed on dataset 1a (All L Sequences) which contained 429 macrohaplogroup L sequences, the rCRS and the chimpanzee outgroup sequence (refer to Section 4.10.4). The 429 sequences, consisting of the coding region only, included sequences from individuals residing in countries located in Africa, as well as from individuals from the United States of America (African Americans and Hispanics) and the Dominican Republic (refer to Appendix G). Due to the reasons discussed in Section 4.10.1, four sequences obtained from GenBank® were excluded. MEGA software version 4 (Tamura *et al.*, 2007) was used to construct a phylogenetic tree as described in Section 4.12. The MP tree, which represents a 50% majority rule consensus tree, is presented in Figure 5.45.

Figure 5.45 Global maximum parsimony tree constructed using the All L Sequences dataset (dataset 1a)



This is a 50% majority rule consensus tree constructed using MEGA version 4 as described in Section 4.12. The tree consists of 429 mitochondrial DNA sequences, the rCRS and the *Pan troglodytes* sequence as outgroup. All the relevant information regarding the sequences used is listed in Appendix G. Statistical support for the observed phylogeny was determined by performing 1,000 bootstrap replicates, and the bootstrap values for the major branches and groups are indicated on an A3 size copy of this tree included in Appendix J. Khoi-San sequences generated in the current investigation are indicated by a red star (★), the rCRS is indicated by a turquoise octagon (◐), the outgroup is indicated by a green square (■) and the African reference sequence is indicated by a violet triangle (▲). The major haplogroups are indicated on the figure at the relevant branching points. Sub-haplogroups are indicated at the relevant branching points, and by the use of different colours as listed in the legend. Sub-haplogroups indicated in grey boxes represent sub-haplogroups assigned based on the alternative classification scheme (refer to Section 3.3.1.1), sub-haplogroups indicated in red boxes are new sub-haplogroups which have not previously been named and sub-haplogroups presented in the other coloured boxes are subdivisions identified in previously described sub-haplogroups. Sequences that could not be classified completely are indicated by black lines. Sequences that exhibit a discrepancy between their location in the tree and their haplogroup classification are indicated by dotted lines.

In Figure 5.45, the rCRS is signified by a **turquoise** octagon (●), the *Pan troglodytes* outgroup is designated by a **green** square (■) and the African reference sequence is represented by a **violet** triangle (▲). The Khoi-San sequences generated in the current investigation are indicated by **red** stars (★) and their location within the global MP tree is as summarised in Table 5.33. Due to the size of the dataset and computational limitations, 1,000 bootstrap replicates were performed (refer to Section 4.12.6). The bootstrap values obtained for the major branches and groups are indicated on an A3 size copy of this tree included in Appendix J. The obtained bootstrap values were generally low. This was ascribed to the size of the dataset investigated (429 sequences), as well as to high sequence divergence and rate variability among lineages (refer to Section 5.9.1). The major haplogroups and sub-haplogroups are indicated on the figure at the relevant branching points and different colours are used to distinguish between the different sub-haplogroups. Sequences designated using dotted lines are sequences of which the haplogroup classification was incompatible with the location of the sequence in the phylogenetic tree. Sequences indicated by black lines are sequences for which the haplogroup classification was incomplete (refer to Section 4.9). A list of the sequences indicated by dotted and black lines is provided in Table 5.34.

The MP tree presented in Figure 5.45 was similar to the N-J tree depicted in Figure 5.42. The most basal lineage of the global MP tree consisted of the group of incompletely classified sequences referred to as L5 (refer to Table 5.34), followed by a group composed of L0/L1 sequences, and a group containing L2/L3 sequences. The L5 group was subdivided into two groups. The first group contained four sequences (EF184581, EF184583, EF184584 and EF184590) and the second group was composed of five sequences (DQ112921, DQ112922, DQ341060, DQ341061 and EF184582). Using the alternative classification scheme (refer to Section 3.3.1.1), the four sequences in the first group were assigned to L5b and the five sequences in the second group were classified as either L5a1 or L5a2. The two groups observed within the L5 group therefore correspond to two sub-haplogroups already described in the alternative classification scheme.

The L0* haplogroup consisted of two main groups, the first containing sequences classified as L0a and L0b, and the second consisting of sequences classified as L0a, L0c1 and L0c2. A difference was observed between the global N-J tree and the global MP tree regarding the position of the group of L0a sequences referred to as the L0aA group. In the N-J tree, these 14 sequences group with the L0aB and the L0b sequences, while in the MP tree the sequences group with the L0aC, L0c1 and L0c2 sequences. This difference

could be explained by the use of the two different methods. The N-J method converted the observed pairwise character differences between sequences into distance data, while the MP method used the actual character differences to construct the tree. However, in both trees, this group is divided into two groups, L0aA1 and L0aA2. It is interesting to note that seven sequences in the L0aA group are Khoi-San sequences generated in the current investigation, three sequences are Khoi-San sequences obtained from another investigation (Mouton, 2003) and four sequences are San sequences (Ingman *et al.*, 2000; Gonder *et al.*, 2007). Based on the data generated in the current investigation, it can be speculated that this group is Khoi-San specific (refer to Section 5.10.2.2). The branching order of sequences within L0aA differs slightly compared to that observed in the phylogenetic tree constructed using the N-J method. The most important difference pertains to the location of KS_CGR_059. In the MP tree this sequence groups within the L0aA1 group. It appears that two groups, L0aA1 (containing KS_CGR_059, KS_CGR_084 and KS_CGR_120), and L0aA2 (containing KS_CGR_001, KS_CGR_083, KS_CGR_157 and KS_CGR_176) are present within L0aA and these two groups are further discussed in Section 5.10.2.1.

The L0aB group contained four sequences, which included one incompletely classified sequence, and appeared to be divided into two groups. L0aB1 contained three sequences and L0aB2 consisted of one sequence. This group and its subdivisions are further discussed in Section 5.10.2.1. The L0aC group was composed of seven sequences, including one incompletely classified sequence and two sequences which exhibited a discrepancy between their haplogroup classification and location in the phylogenetic tree. The L0aC group is further subdivided into two groups, the first group contained a single sequence (EF184595) and the other group consisted of the remaining six sequences. Both the L0aB and the L0aC groups are discussed further in Section 5.10.2.1.

The L0b group was subdivided into two groups, referred to as L0bA and L0bB, as previously observed in the global N-J tree. Three groups (L0bB1, L0bB2 and L0bB3) were observed in the global N-J tree. However, only two groups (L0bB1 and L0bB3 vs. L0bB2) are supported within the L0bB group in the MP tree. The L0bB2 group is however subdivided into two groups in both the global N-J and MP trees. The L0b group and its subdivisions are discussed further in Section 5.10.2.1.

The L0c1 group appeared to be divided into two groups, L0c1A which consisted of seven sequences (AF381988, DQ112777, DQ304897, DQ304898, DQ304899, DQ304900 and

EF184601), and L0c1B that contained eight sequences (AF346985, AY195780, DQ112732, DQ112959, DQ304901, DQ304902, DQ304903 and DQ304904). A difference regarding the location of DQ112959 was observed compared to the phylogenetic tree constructed using the N-J method. In the global MP tree, this sequence grouped within the L0c1B group. The L0c1A and L0c1B groups are discussed further in Section 5.10.2.1.

The L0c2 group appeared to be divided into two groups with further subdivisions in each group to generate at least three putative sub-haplogroups. A difference was observed between the global N-J and MP trees relating to the location of D38112 and the number of possible sub-haplogroups. In the MP tree D38112 grouped closer to EF184602 and EF184604, and five possible sub-haplogroups were identified. The L0c2 group and its subdivisions are discussed further in Section 5.10.2.1.

The majority of the 30 Khoi-San sequences generated in the current investigation were located within the L0* haplogroup section of the global MP tree. Fourteen Khoi-San sequences were observed in the L0b group in the same sub-haplogroups as in the N-J tree. Seven Khoi-San sequences were observed in the L0a sub-haplogroup. As already discussed, these seven sequences were classified as belonging to the L0aA group and in contrast to the global N-J tree, these sequences were observed closer to L0c1 and L0c2 than to L0b. The remaining eight Khoi-San sequences were located within the L2b group (KS_CGR_069, KS_CGR_164 and KS_CGR_173), the L3a1 group (KS_CGR_140, KS_CGR_143 and KS_CGR_175) and the L3b group (KS_CGR_090 and KS_CGR_172). In each of these sub-haplogroups, the Khoi-San sequences increased the resolution of the phylogenetic tree by increasing the number of branches within the group.

The branching structure within the L1*, L2* and L3* haplogroups, as well as the grouping of the Khoi-San sequences within these groups, have been discussed in Section 5.9.2.1. The differences observed between the global N-J and MP trees were also discussed where applicable. It appears that, with the exception of the differences already mentioned, the branching order in the main haplogroup and sub-haplogroup branches were similar. With the exception of haplogroup L3*, the bootstrap values were slightly higher in the global MP tree than in the global N-J tree and the same subdivisions within haplogroups or sub-haplogroups were supported in both trees.

As already stated for the global N-J tree (refer to Section 5.9.2.1), it can be concluded from the discussion of the L0*, L1*, L2* and L3* haplogroups in the global MP tree that the

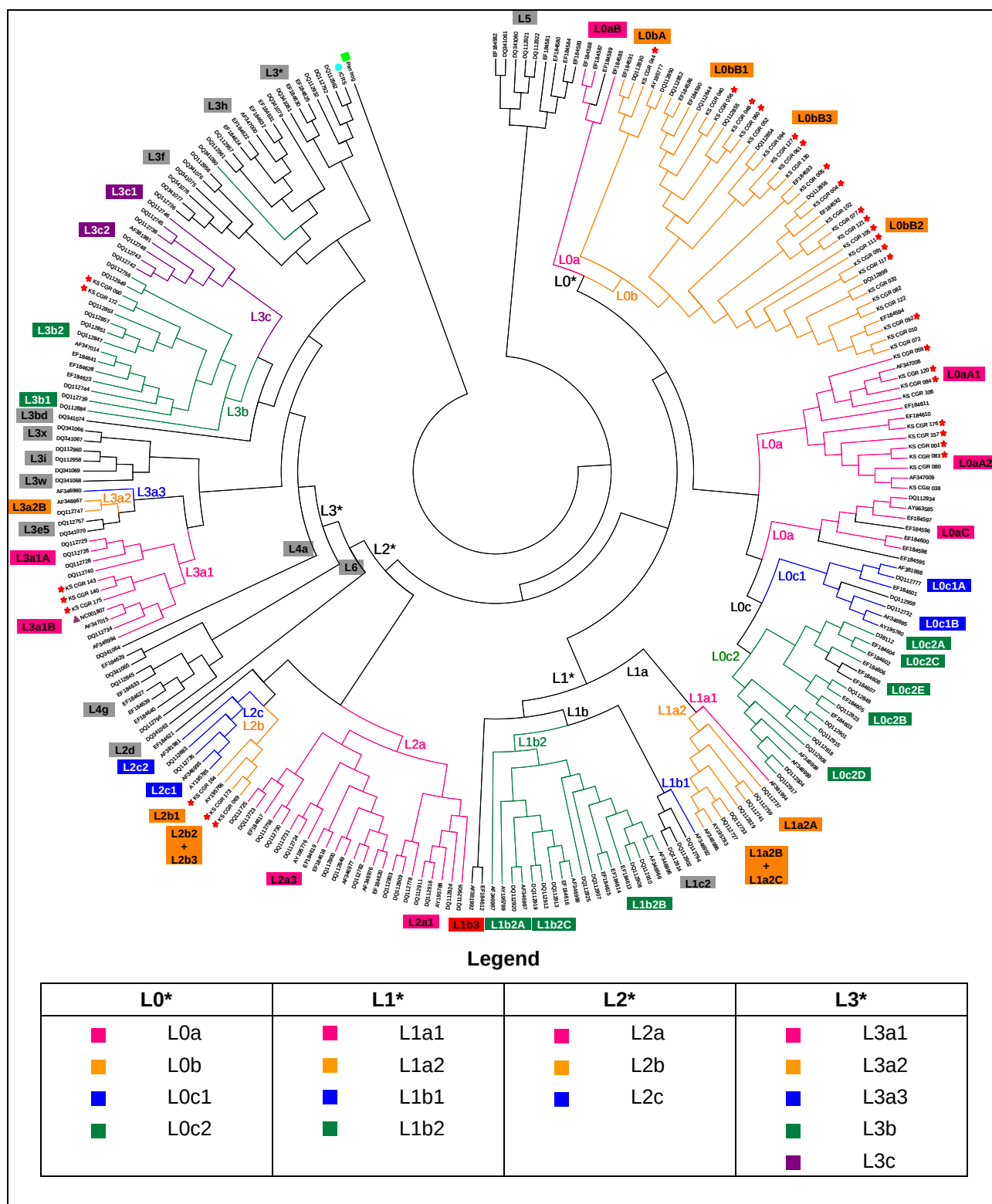
classification scheme used in the current investigation (refer to Section 4.9) does not permit the complete classification of all sequences into sub-haplogroups and does not adequately describe all divisions observed in sub-haplogroups. In contrast, the alternative classification scheme describes these incompletely classified single sequences and groups of sequences, as well as the subdivisions in groups by providing a haplogroup designation. The classification scheme used in the current investigation should therefore be revised to include the haplogroup designations provided by the alternative classification scheme in order to classify the sequences and groups which are currently incompletely classified.

The high level of diversity observed in African lineages, the rapid divergence between lineages to form new sub-haplogroups, and the large number of sequences in the datasets investigated may play a role in obscuring the true evolutionary relationships among sequences (Page and Holmes, 1998). In the current investigation, the use of both the N-J and the MP methods generated essentially the same phylogenetic trees in terms of the topology observed within these two trees. Although the generation of the same tree topology when using different tree construction methods is not a guarantee that the true topology has been revealed, the similar topology in both trees appear to support the observed branching orders within haplogroups, sub-haplogroups and divisions within sub-haplogroups, even though the bootstrap value for the grouping may be low. The MP tree however proved more useful in terms of the investigation of new sub-haplogroups and divisions within sub-haplogroups (refer to Section 5.12). This can be attributed to the difference between the two tree construction methods employed. The N-J method converts character differences to pairwise distances between sequences to determine which sequences group together, thereby losing the information regarding the actual nucleotide differences between sequences. In contrast, the MP method uses the character differences between sequences to determine the branching structure. The haplogroup classification scheme used in the current investigation (refer to Section 4.9) is based on the presence/absence of specific SNPs, i.e. character differences. As the MP tree and the haplogroup classification schemes are both based on nucleotide differences, the branching order in the MP tree should correspond to the actual haplogroup and sub-haplogroup divisions.

5.9.3.2 Maximum parsimony tree of African individuals

The MP method was performed on dataset 2a (All Africa L Sequences) which consisted of only the coding region of 242 mitochondrial DNA sequences obtained from African individuals from diverse ethnic backgrounds residing in various countries in Africa (refer to Table 5.36), as well as the rCRS and the chimpanzee outgroup sequence (refer to Section 4.10.4). As before, four sequences acquired from GenBank® were excluded due to the reasons stated in Section 4.10.1. MEGA version 4 (Tamura *et al.*, 2007) was used to construct a phylogenetic tree as described in Section 4.12. The MP tree, which represents a 50% majority rule consensus tree, is depicted in Figure 5.46.

The Africa MP tree presented in Figure 5.46 follows the same conventions as the global MP tree depicted in Figure 5.45. Sequences of interest (Khoi-San sequences, rCRS, *Pan troglodytes* and the African reference sequence) are indicated using the same symbols, and haplogroup and sub-haplogroups are identified through the use of different colours. The location of all Khoi-San sequences generated in the current investigation within the Africa MP tree is as presented in Table 5.33. A list of incompletely classified sequences, as well as their assigned haplogroup classifications, is provided in Table 5.34. Bootstrap analysis (1,000 replicates) was performed (refer to Section 4.12.6) and the bootstrap values obtained for the major branches and groups are indicated on an A3 size copy of this tree included in Appendix J. In general, the observed bootstrap values were low, the values were however comparable to the values observed in the Africa N-J tree. In the following section, the Africa MP tree is discussed briefly. The branching observed within the major haplogroups, the identification of potential sub-haplogroups as indicated by subdivisions within groups, as well as the corresponding designations in the alternative classification scheme for any identified sub-haplogroups have already been discussed in Section 5.9.2.1.

Figure 5.46 Africa maximum parsimony tree constructed using the All Africa L Sequences dataset (dataset 2a)

This is a 50% majority rule consensus tree constructed using MEGA version 4 as described in Section 4.12. The tree consists of 242 mitochondrial DNA sequences, the rCRS and the *Pan troglodytes* sequence as outgroup. All the relevant information regarding the sequences used is listed in Appendix G. Statistical support for the observed phylogeny was determined by performing 1,000 bootstrap replicates, and the bootstrap values for the major branches and groups are indicated on an A3 size copy of this tree included in Appendix J. Khoi-San sequences generated in the current investigation are indicated by a red star (★), the rCRS is indicated by a turquoise octagon (◐), the outgroup is indicated by a green square (■) and the African reference sequence is indicated by a violet triangle (▲). The major haplogroups are indicated on the figure at the relevant branching points. Sub-haplogroups are indicated at the relevant branching points, and by the use of different colours as listed in the legend. Sub-haplogroups indicated in grey boxes represent sub-haplogroups assigned based on the alternative classification scheme (refer to Section 3.3.1.1), sub-haplogroups indicated in red boxes are new sub-haplogroups which have not previously been named and sub-haplogroups presented in the other coloured boxes are subdivisions identified in previously described sub-haplogroups. Sequences that could not be classified completely are indicated by black lines. Sequences that exhibit a discrepancy between their location in the tree and their haplogroup classification are indicated by dotted lines.

The Africa MP tree presented in Figure 5.46 was similar to the Africa N-J tree depicted in Figure 5.43. A group of sequences classified as L5 (refer to Table 5.34) formed the most basal lineage of the Africa MP tree, followed by a group containing L0/L1 sequences, and a group consisting of L2/L3 sequences. The L5 group was divided into two groups, the first group contained five sequences (DQ112921, DQ112922, DQ341060, DQ341061 and EF184582) and the second group was composed of four sequences (EF184580, EF184581, EF184583 and EF184584). According to the alternative classification scheme (refer to Section 3.3.1.1), the five sequences in the first group were classified as either L5a1 or L5a2, and the four sequences in the second group were assigned to L5b. A difference was observed between the Africa MP tree and the Africa N-J tree regarding the subdivisions in the L5 group. In the Africa MP tree, the L5a2 sequences group with the L5a1 sequences, instead of the L5b sequences. This grouping was also observed in the global N-J tree, and it is speculated that the discrepancy between trees regarding the L5 group may be attributed to unresolved branching structure within this haplogroup as suggested by the low bootstrap value for this branch.

The L0* haplogroup was composed of two main groups. The first group contained sequences classified as L0a and L0b, and the second consisted of sequences assigned to the L0a, L0c1 and L0c2 sub-haplogroups. A difference was observed between the Africa N-J and MP trees regarding the position of the group of L0a sequences referred to as the L0aA group. In the MP tree these 14 sequences group with the L0aC, L0c1 and L0c2 sequences, while in the N-J tree the sequences group with the L0aB and the L0b sequences. As discussed previously, this difference is attributed to the use of the two different tree construction methods (refer to Section 5.9.3.1) and this difference is observed in comparisons of the global and Africa N-J trees with the corresponding MP trees. The L0aA group contained Khoi-San and San sequences (Ingman *et al.*, 2000; Mouton, 2003; Gonder *et al.*, 2007; current investigation) and was divided into two groups, L0aA1 and L0aA2. This group appears to be specific to the Khoi-San from southern Africa (refer to Section 5.10.2.2). As stated previously, the branching order of sequences within L0aA is different between the Africa N-J and MP trees with the most notable difference relating to the location of KS_CGR_059 (refer to Section 5.9.3.1). It appears that two groups, L0aA1 and L0aA2, are present within L0aA. The L0aB group contained four sequences, which included one incompletely classified sequence, and appeared to be divided into two groups. L0aB1 contained three sequences and L0aB2 consisted of one sequence. The L0aC group was composed of seven sequences, including one incompletely classified sequence and two sequences which exhibited a discrepancy

between their haplogroup classification and location in the phylogenetic tree. The L0aC group is further subdivided into two groups, the first group contained a single sequence (EF184595) and the other group consisted of the remaining six sequences. The L0aA, L0aB and L0aC groups, as well as their subdivisions, are discussed further in Section 5.10.2.1.

The L0b group was divided into two groups, L0bA and L0bB. In the Africa N-J tree, L0bB was subdivided further and three groups (L0bB1, L0bB2 and L0bB3) were observed. In contrast, in the Africa MP tree, only two groups (L0bB2 vs. L0bB1 and L0bB3) were supported for L0bB. In both the Africa N-J and MP trees, the L0bB2 group was subdivided into two groups. The L0b group and its subgroups are discussed further in Section 5.10.2.1.

The L0c1 group appeared to be divided into two groups. L0c1A contained three sequences (AF381988, DQ112777 and EF184601), and L0c1B was composed of four sequences (AF346985, AY195780, DQ112732 and DQ112959). A difference regarding the location of EF184601 was observed compared to the phylogenetic tree constructed using the N-J method. In the MP tree, this sequence grouped within the L0c1A group. Differences regarding groupings in the L0c1 group were also observed in comparisons between the global and Africa N-J trees (refer to Section 5.9.2.2). These differences, together with low bootstrap values, suggest that the L0c1 group has not been completely resolved. The L0c1 group and its subdivisions are discussed further in Section 5.10.2.1.

The L0c2 group appeared to be divided into three groups. Differences were observed between the Africa N-J and MP trees relating to the location of D38112 and the number of potential sub-haplogroups. In the MP tree, D38112 grouped with EF184602 and EF184604, and four or five subgroups were possible. The differences between the Africa N-J and MP trees relating to the location of D38112, the number of potential sub-haplogroups, and the generally low bootstrap values observed for branches in this group, suggest that the L0c2 group is unresolved and more sequences are required to clarify relationships in this section of the phylogenetic trees. The L0c2 group and its subdivisions are discussed further in Section 5.10.2.1.

The majority of the 30 Khoi-San sequences generated in the current investigation were located within the L0* haplogroup section of the Africa MP tree. The L0b group contained 14 Khoi-San sequences in the same sub-haplogroups as in the Africa N-J tree. The L0a

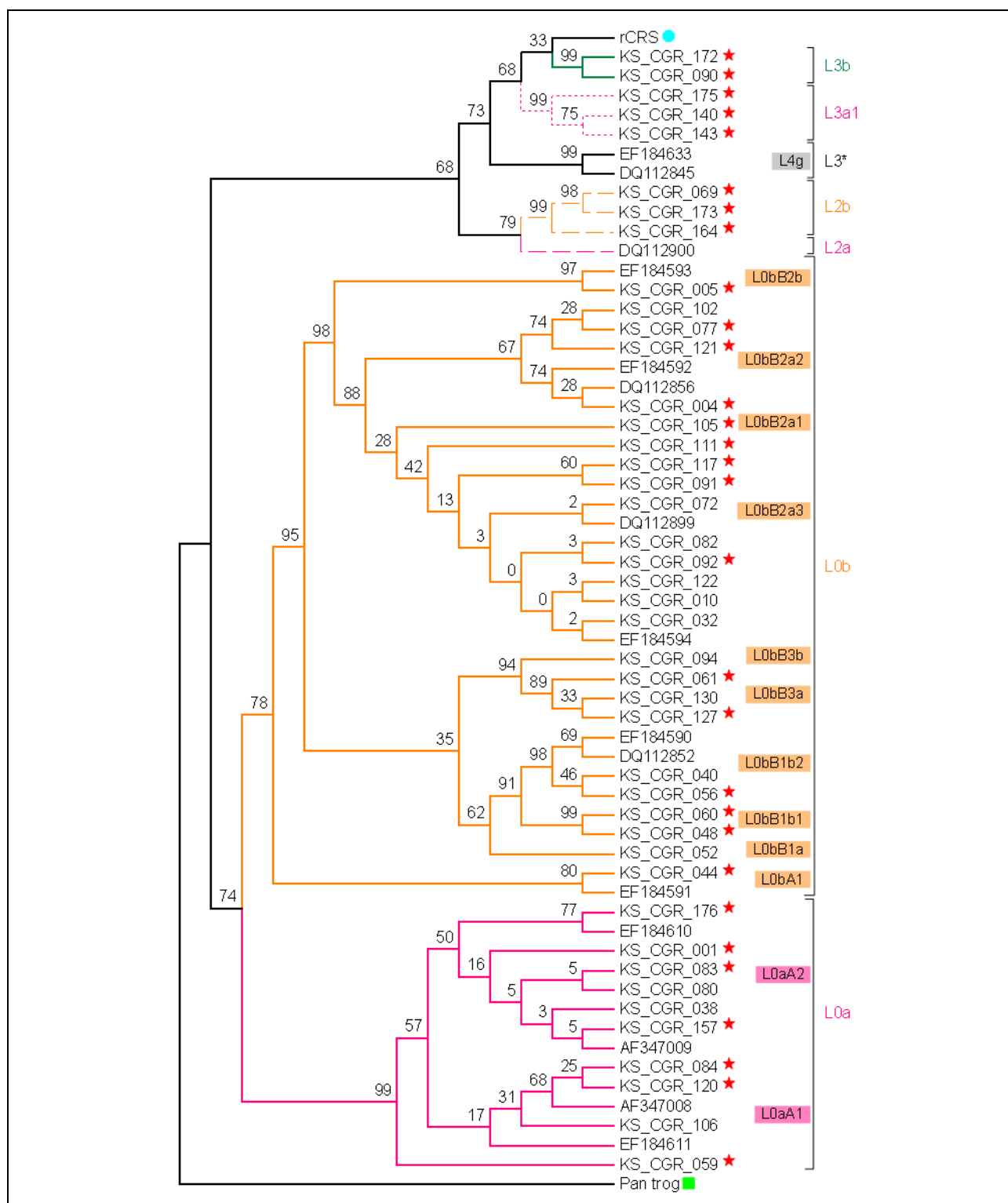
group, specifically the L0aA subgroup, contained seven Khoi-San sequences and was observed within the same group as the L0c1 and L0c2 sequences. This observation was in contrast to the branching order observed in the Africa N-J tree. The remaining eight Khoi-San sequences were scattered throughout the Africa MP tree and were observed within the L2b, L3a1 and L3b groups. In each of these sub-haplogroups, the Khoi-San sequences added new branches within the group thereby increasing the resolution of the phylogenetic tree.

The topology within the L1*, L2* and L3* haplogroups, as well as the location of the Khoi-San sequences within these groups, has been discussed in Section 5.9.2.2. Where necessary, the observed differences between the Africa N-J and MP trees were also discussed. With the exception of the differences already mentioned, it appears that the branching order within the main haplogroup and sub-haplogroups were similar. The same sequences were observed within the same groups in the two trees. However, slight differences were observed within groups in terms of the sequences that group closest to each other. The bootstrap values obtained were comparable and the same subdivisions within haplogroups or sub-haplogroups were supported in both the Africa N-J and MP trees. The use of both the N-J and MP methods therefore produced similar phylogenetic trees, and although this is not a guarantee that the true topology has been obtained, it appears to lend support to the branching structure observed. As discussed previously (refer to Section 5.9.3.1), the Africa MP tree is more conducive to the investigation of haplogroup and sub-haplogroup branching structures (refer to Section 5.12).

5.9.3.3 Maximum parsimony tree of Khoi-San individuals

The MP method was performed on datasets 3a and 3b which contained 58 or 55 Khoi-San sequences respectively, the rCRS and the chimpanzee outgroup sequence (refer to Section 4.10.4). The datasets are the same datasets that were used to construct the Khoi-San N-J trees (refer to Section 5.9.2.3) and contained coding region only sequences (dataset 3a) or complete sequences (dataset 3b). A single San sequence, EF184609 (Gonder *et al.*, 2007), was excluded from both datasets as previously described (refer to Section 5.9.2.3). MEGA version 4 (Tamura *et al.*, 2007) was used to construct a phylogenetic tree as described in Section 4.12. The MP tree based on dataset 3a, represented as a 50% majority rule consensus tree, is presented in Figure 5.47.

Figure 5.47 Khoi-San maximum parsimony tree constructed using the All Khoi-San L Sequences dataset (dataset 3a)



This is a 50% majority rule consensus tree constructed using MEGA version 4 as described in Section 4.12. The tree consists of 58 mitochondrial DNA sequences, the rCRS and the *Pan troglodytes* sequence as outgroup. All the relevant information regarding the sequences used is listed in Appendix G. Khoi-San sequences generated in the current investigation are indicated by a red star (*), the rCRS is indicated by a turquoise octagon (●) and the outgroup is indicated by a green square (■). Statistical support for the observed phylogeny was determined by performing 1,000 bootstrap replicates, and the bootstrap values for the major branches and groups are indicated. Sub-haplogroups are indicated by brackets on the right hand side of the figure and by the use of different colours and different types of lines (solid, dashed or dotted lines). The sub-haplogroup indicated in a grey box represents a sub-haplogroup assigned based on the alternative classification scheme (refer to Section 3.3.1.1) and sub-haplogroups presented in coloured boxes are subdivisions identified in previously described sub-haplogroups. Sequences that could not be classified completely are indicated by black lines.

The MP trees constructed from datasets 3a (coding region sequences) and 3b (complete sequences) were compared. The topology of the MP tree constructed using dataset 3a was remarkably similar to the MP tree generated using dataset 3b, and the same groups and subdivisions were supported in both trees with comparable bootstrap values for the observed groups. Only minor differences were observed between the two trees regarding the branching order of single sequences within groups. The Khoi-San MP tree based on coding region only Khoi-San sequences is presented and discussed as this tree is comparable to the global and Africa MP trees which were generated using coding region only sequence data.

The Khoi-San MP tree depicted in Figure 5.47 follows the same conventions as the global MP tree presented in Figure 5.45. Haplogroup and sub-haplogroups are identified through the use of different colours and sequences of interest (Khoi-San sequences, rCRS and *Pan troglodytes*) are indicated by coloured symbols. Table 5.33 provides the location of all Khoi-San sequences determined in the current investigation within the Khoi-San MP tree and Table 5.34 provides a list of incompletely classified sequences, including their haplogroup designations. Bootstrap analysis (1,000 replicates) was performed (refer to Section 4.12.6) and the bootstrap values obtained for the major branches and groups are presented. The observed bootstrap values were generally low, although higher than the values observed in the corresponding Khoi-San N-J tree. The higher bootstrap values in the Khoi-San MP tree, as well as the clearly distinguishable groups, suggest that the MP tree is better suited to the investigation of the relationships between the closely related Khoi-San sequences generated in the current investigation than the N-J tree. In addition, the MP method determines relationships based on character differences which can be directly correlated to the haplogroup and sub-haplogroup classifications assigned to the Khoi-San sequences as the classification scheme used is based on the presence/absence of specific alterations (refer to Section 4.9). The bulk of the Khoi-San sequences included in this tree were assigned to the L0 haplogroup. The branching structure within the L0 haplogroup, as well as potential sub-haplogroups identified within this haplogroup, has previously been discussed in Section 5.9.2.1.

Two main groups can be distinguished in the Khoi-San MP tree depicted in Figure 5.47. One group consisted of L0a and L0b sequences, and the other group contained the rCRS, two incompletely classified sequences, L2 sequences and L3 sequences. The L0 haplogroup was divided into two main groups, L0a and L0b. The L0a group consisted of ten Khoi-San sequences and four San sequences which were all classified as L0aA

based on the revised classification scheme (refer to Section 5.12). The L0aA group was divided into two groups, one group was composed of one sequence (KS_CGR_059) and the other group contained the remaining 13 sequences. The second group appeared to be divided into two more groups, one group containing five sequences (AF347008, EF184611, KS_CGR_084, KS_CGR_106 and KS_CGR_120) and the other group consisting of eight sequences (AF347009, EF184610, KS_CGR_001, KS_CGR_038, KS_CGR_080, KS_CGR_083, KS_CGR_157 and KS_CGR_176). This grouping was different to that observed in the Khoi-San N-J tree where the two groups consisted of one and 12 sequences. The difference in the group sizes can be attributed to the position of EF184611 within the L0aA group as it forms a group on its own in the Khoi-San N-J tree. According to the revised classification scheme proposed in Section 5.12, KS_CGR_059 and the group consisting of five sequences are all assigned to the L0aA1 group, while the group of six sequences are all classified as L0aA2. The grouping of KS_CGR_059 outside of the group containing the bulk of the L0aA sequences and the group consisting of L0aA1 sequences can be attributed to the presence of two reversion events at nucleotide positions 10,876 and 10,939 (refer to Section 5.10.2.1). Mouton in 2003 constructed a phylogenetic tree and identified a group, referred to as Group I, which contained three Khoi-San sequences (KS_CGR_038, KS_CGR_080 and KS_CGR_106). Group I is equivalent to L0aA and the Khoi-San sequences generated in the current investigation therefore expanded Group I by subdividing this group into two groups. The L0aA group and its subdivisions are discussed in more detail in Section 5.10.2.1.

In the global and Africa N-J and MP trees, two additional L0a lineages (L0aB and L0aC) were observed. Based on data available to date, the L0aB and L0aC lineages appear to be confined to individuals from eastern Africa (refer to Section 5.10.2.2) which includes the Sandawe (refer to Appendix G) who speak click languages which form part of the Khoi-San language family (refer to Section 3.7). In contrast, the L0aA lineage seems to be specific to the Khoi-San and San (refer to Section 5.10.2.2) which are Southern African Khoi-San language speakers (refer to Section 3.7). It was noted that none of the Khoi-San sequences generated in the current investigation, or obtained from GenBank® or other sources, were classified as L0c1 or L0c2. Based on this data, it appears that the L0a lineages (L0aA, L0aB and L0aC) may be specific to Khoi-San speaking individuals.

The L0b group contained 33 sequences and was divided into two groups which were named L0bA and L0bB based on the revised classification scheme (refer to Section 5.12). As stated previously (refer to Section 5.9.2.3), the L0bA group can be subdivided into two

groups (L0bA1 and L0bA2). This division was however not observed in the Khoi-San MP tree. The L0bB group was subdivided into three groups (L0bB1, L0bB2 and L0bB3), each of which were subdivided into two more groups (L0bB1a, L0bB1b, L0bB2a, L0bB2b, L0bB3a, L0bB3b). A difference was however observed in the branching structure for the L0bB group in the Khoi-San N-J tree compared to the Khoi-San MP tree. In the N-J tree, L0bB1 and L0bB2 group together, while in the MP tree, L0bB1 and L0bB3 are closely related. The grouping of L0bB1 with L0bB3 corresponds to the branching order observed in the global and Africa N-J and MP trees. This again indicated that the Khoi-San MP tree was better suited to the investigation of the evolutionary relationships between the Khoi-San sequences generated in the current investigation than the Khoi-San N-J tree. Subsequent to phylogenetic analyses, Mouton in 2003 identified a group (Group II) that contained four Khoi-San sequences (KS_CGR_040, KS_CGR_052, KS_CGR_094 and KS_CGR_130) and another group (Group III) that consisted of six Khoi-San sequences (KS_CGR_010, KS_CGR_032, KS_CGR_072, KS_CGR_082, KS_CGR_102 and KS_CGR_122). Two of these sequences were observed in the L0bB1 group (KS_CGR_040 and KS_CGR_052), two in the L0bB3 group (KS_CGR_094 and KS_CGR_130) and the remaining six sequences in the L0bB2 group. The addition of the 15 Khoi-San sequences generated in the current investigation therefore expanded both Groups II (L0bB1 and L0bB3) and III (L0bB2) by increasing the branching in these groups to generate additional sub-haplogroups (L0bB1a, L0bB1b, L0bB2a, L0bB2b, L0bB3a and L0bB3b). The L0bA and L0bB groups and their subdivisions are discussed further in Section 5.10.2.1.

The majority of Khoi-San sequences included in the Khoi-San N-J tree are classified as L0b and it appears that this lineage may be Khoi-San specific. However, in the global and Africa N-J trees, it can be observed that sequences from individuals from other populations are also classified as L0b. All except one of these individuals can trace their origin to southern Africa, and it is therefore more likely that this lineage is specific to southern Africa (refer to Section 5.9.4).

Twelve sequences were observed in the group consisting of the rCRS, L2 sequences and L3 sequences. One Khoi-San sequence was assigned to the L2a haplogroup, three Khoi-San sequences were classified as L2b, three Khoi-San sequences were designated L3a1, two Khoi-San sequences belonged to the L3b sub-haplogroup and two sequences (EF184633 and DQ112845) were incompletely classified as L3b\$ (refer to Table 5.34). The two incompletely classified sequences were re-classified as L4g according to the

alternative classification scheme (refer to Section 3.3.1.1). Differences were observed between the Khoi-San N-J and MP trees relating to the rCRS and the two incompletely classified sequences within this group (refer to Section 5.9.2.3). In contrast to the Khoi-San N-J tree, the branching structure observed within this group in the Khoi-San MP tree corresponds to that observed in the global and Africa N-J and MP trees.

5.9.4 **Geographic distribution of African mitochondrial haplogroups**

The geographic distribution of African mtDNA haplogroups was investigated based on the geographic origin of the mitochondrial sequences included in the global and Africa MP trees. The geographic origin of each of the mitochondrial sequences was determined based on information supplied regarding the country in which the sample was collected, or from the ethnic affiliation of the individual from which the sequence was obtained (refer to Appendix G). A map of Africa indicating the countries contained in each geographic region is presented in Appendix I. The aim was to determine if there was any correlation between the mitochondrial haplogroup and the distribution of the specific haplogroup within Africa. A copy of the global MP tree, altered to indicate the geographic origin of the mitochondrial sequences, is presented in Figure 5.48.

The global MP tree presented in Figure 5.48 follows the same conventions as the global MP tree presented in Figure 5.45. Sequences of interest (Khoi-San sequences, rCRS and *Pan troglodytes*) are indicated by coloured symbols. The haplogroup and sub-haplogroup designation of the branches in the phylogenetic tree are indicated. The classification according to the classification scheme used in the current investigation (refer to Section 4.9) is indicated in coloured text on the relevant branch. New sub-haplogroups/subdivisions identified in the haplogroups/sub-haplogroups in the classification scheme used in the current investigation are indicated in regular text in a coloured block, whereas the classification according to the alternative classification scheme (refer to Section 3.3.1.1) is presented in *italic* text in a coloured block, and the classification based on the revised classification scheme (refer to Section 5.12) is indicated in underlined text in a coloured block. The colours of the blocks correspond to the colour of the branch described by the haplogroup. The branches representing haplogroups or sub-haplogroups are indicated in colours according to the predominant geographic region of Africa from which the sequences contained in the branch was obtained, i.e. if the majority of the sequences in the branch was obtained from individuals from western Africa, the branch is presented in blue.

Legend

Region in Africa	Region outside Africa
■ Southern Africa	■ USA (African American)
■ East Africa	■ Dominican Republic
■ West Africa	■ USA (Hispanic)
■ North Africa	■ Africa
■ Central Africa	

326

The geographic distribution of each haplogroup or sub-haplogroup in Africa is presented in Table 5.37. The haplogroup classification provided for each branch is based on the classification scheme used throughout the current investigation (refer to Section 4.9). For branches that contained sequences that were incompletely classified, the corresponding haplogroup/sub-haplogroup designation according to the alternative classification scheme (refer to Section 3.3.1.1) is provided as previously used during the discussion of the global and Africa N-J and MP trees (refer to Sections 5.9.2 and 5.9.3).

Table 5.37 Geographic distribution of African mitochondrial haplogroups

Haplogroup/sub-haplogroup classification			Geographic distribution
Current investigation ^a	Alternative ^b	Proposed ^c	
L0\$a	L5a1	---	East and Central Africa
	L5a2	---	East Africa
L0\$b	L5b	---	East Africa
L0,L1	L1c3	L1b3	West and East Africa, Dominican Republic
L0a	---	L0aA1	Southern Africa
	---	L0aA2	Southern Africa
	---	L0aB	East Africa
	---	L0aC	East Africa
L0b	---	L0bA	Southern Africa
	---	L0bB1	Southern Africa
	---	L0bB2	Southern Africa
	---	L0bB3	Southern Africa
L0c1	---	L0c1A	North and East Africa, African Americans
	---	L0c1B	West and North Africa, African Americans
L0c2	---	L0c2A	East Africa
	---	L0c2B	East, southern and Central Africa, Dominican Republic
	---	L0c2C	East Africa
	---	L0c2D	Central Africa
	---	L0c2E	East Africa
L1a1	---	---	West Africa
L1a2	---	---	West Africa, African Americans, Dominican Republic, Hispanics
L1b1	---	---	East Africa, Dominican Republic
L1b2	---	---	Central Africa
L1b2\$	L1c2	---	Central Africa
L2a	---	L2a1	Central Africa
	---	L2a2	West Africa, African Americans
	---	L2a3	West, East and southern Africa, African Americans, Dominican Republic
L2b	---	---	Southern Africa, African Americans
L2c	---	---	West Africa, African Americans, Dominican Republic
L2d	---	---	Dominican Republic

Table 5.37 Continued...

Haplogroup/sub-haplogroup classification			Geographic distribution
Current investigation ^a	Alternative ^b	Proposed ^c	
L2d\$	L2d	---	East Africa, Dominican Republic
	L6	---	East Africa
L3a1	---	---	West Africa, African Americans, Dominican Republic
L3a2	---	---	West and Central Africa, African Americans
L3a3	---	---	Central Africa, Hispanics, Dominican Republic
L3a3\$	L3e4	---	Dominican Republic
	L3e5	---	West Africa
	L3i	---	North and East Africa
	L3w	---	East Africa
	L3x	---	East Africa
L3b	---	---	Southern, West, East and North Africa, Dominican Republic
L3b\$	L3*	---	East Africa
	L3bd	---	East Africa
	L3f	---	West, East and North Africa, African Americans
	L3h	---	East, North and Central Africa, Dominican Republic
	L4a	---	East Africa
	L4g	---	East, southern and Central Africa
L3c	---	---	West Africa, African Americans, Dominican Republic

Groups of incompletely classified sequences are indicated by the presence of a dollar sign (\$) in the haplogroup/sub-haplogroup designation. ^a = refers to the haplogroup classification assigned to a group according to the classification scheme used in the current investigation (refer to Section 4.9). ^b = refers to the haplogroup classification assigned to the group based on the alternative classification scheme described in Section 3.3.1.1. ^c = refers to the haplogroup/sub-haplogroup designation assigned to a group according to the revised classification scheme proposed in Section 5.12, or to a name proposed to describe a previously unnamed sub-haplogroup in haplogroups other than L0.

A number of observations can be made from the data presented in Figure 5.48 and Table 5.37. The L5 haplogroup was observed in East and Central Africa, with the L5a1 sub-haplogroup identified in Pygmies and the L5b sub-haplogroup observed mainly in individuals from the Sandawe population (Gonder *et al.*, 2007). It therefore appears that these two lineages are very old and may be confined to specific populations. However, more sequence data is needed from East and Central Africa to confirm this relationship.

The sub-haplogroups of L0* appeared to have a set geographic distribution. L0aA was mainly observed in southern Africa, L0aB and L0aC in East Africa, L0b in southern Africa, L0c1A in North Africa, L0c1B in West Africa, and L0c2 in East and Central Africa. It also appeared that some lineages could possibly be population-specific with L0aA and L0b observed mainly in Khoi-San, as well as San individuals, and L0aB observed in Sandawe individuals. The geographic distribution of the L0 haplogroup and its sub-haplogroups are discussed in more detail in Section 5.10.2.1.

The geographic distribution of the L1*, L2* and L3* haplogroups was determined by using a copy of the Africa MP tree (refer to Figure 5.43) which was altered to indicate the geographic origin of the mitochondrial sequences (not presented). The addition of sequences from African Americans, Hispanics and individuals from the Dominican Republic in order to generate the global MP tree (refer to Figure 5.48), expanded most branches observed within the L1*, L2* and L3* haplogroups of the Africa MP tree. The majority of these individuals can trace their African ancestry to the slave trade that transported millions of individuals from the African continent to new colonies outside of Africa (Salas *et al.*, 2004a). As the exact geographic origins of these sequences are unknown, the addition of these sequences to the geographic tree generally does not provide additional information about the geographic distribution of the haplogroup in question. However, the other sequences that branch with these sequences from African Americans, Hispanics and individuals from the Dominican Republic, may provide information as to the part of Africa from which these individuals may have originated.

The sub-haplogroups of L1* seemed to have a defined geographic distribution. L1a was mostly observed in West Africa and L1b in Central Africa. A new sub-haplogroup of L1b, namely L1b3, was identified and was observed in West and East Africa. The L1b2 lineage was observed mainly in Pygmies and it is possible that this lineage is population-specific. More sequences are however needed to confirm this observation.

The L2a sub-haplogroup was subdivided into three groups, L2a1 was identified mainly in Pygmies in Central Africa, L2a2 was observed in West Africa and L2a3 was present mainly in West Africa. L2b was observed in southern Africa and L2c in West Africa. A number of sequences were incompletely classified as L2d\$. Three of these sequences (DQ112708, DQ341062 and EF184621) formed a group that was present in East Africa and was assigned to the L2d group according to the alternative classification scheme, while the remaining sequence (DQ341063) was re-classified using the alternative classification scheme as L6 which was observed in East Africa (refer to Table 5.37).

Haplogroup L3* was divided into a number of sub-haplogroups. L3a1 and L3a2 were observed mainly in West Africa and L3a3 in Central Africa. The L3a1 sub-haplogroup contained a southern African lineage consisting of Khoi-San individuals only. It is possible that this lineage is a Khoi-San specific L3a1 lineage. More Khoi-San L3 sequences are however needed to confirm this result. A number of sequences were incompletely classified as L3a3\$. These sequences formed groups which were classified using the

alternative classification scheme as L3e4, L3e5, L3i, L3x and L3w (refer to Table 5.37). L3e4 was present in an individual from the Dominican Republic, L3e5 was observed in West Africa, L3i was present in North and East Africa, whereas L3x and L3w were observed in East Africa. The L3b sub-haplogroup was observed mainly in southern and West Africa, and the L3c sub-haplogroup was present in West Africa. The L3b sub-haplogroup contained a southern African lineage consisting of individuals from Angola or Namibia. It could be that this lineage is localised to these two countries, however, more sequences from other countries in southern Africa are required to test this hypothesis. A number of sequences were incompletely classified as L3b\$ and formed groups which were re-classified as L3*, L3bd, L3h, L3f, L4a and L4g using the alternative classification scheme (refer to Table 5.37). L3*, L3bd, L3h and L4a were observed in East Africa, L3f was mostly present in West Africa and L4g was observed mainly in East Africa.

The results presented in this section were based on the phylogenetic trees constructed using datasets 1a and 2a. As already described (refer to Section 4.10.1), these datasets were compiled from sequences generated in the current investigation, as well as sequences obtained from other investigations and GenBank[®]. The L-specific sequences contained in the datasets were assigned to geographic regions in Africa based on information regarding the country of sample collection or the ethnicity of the individual where no information was supplied regarding the African country of origin. Dataset 1a consisted of ten sequences from North Africa (Morocco and Sudan), 41 sequences from Central Africa (Cameroon, Central African Republic and Democratic Republic of the Congo), 50 sequences from West Africa (Nigeria, Burkina Faso, Ghana and Mauritania), 65 sequences from East Africa (Ethiopia, Tanzania and Uganda), and 68 sequences from southern Africa (Angola, Namibia and South Africa). The remaining sequences were from countries outside of Africa (Dominican Republic and United States of America), or were obtained from Africa, and no information relating to ethnicity or region in Africa was available to assign the sequence to a specific geographic region. The majority of sequences from southern Africa included in the dataset were from Khoi-San and San individuals, and the sequences from Central Africa were obtained mostly from Pygmies. From the above discussion of the dataset, it can be concluded that the majority of L haplogroup sequence data in the dataset has been obtained from only a few populations and countries in each region of Africa. These datasets therefore do not represent a sample obtained in a systematic manner to ensure sampling of sequences of all haplogroups and from all countries in Africa and the results presented are therefore specific to the datasets employed in the current investigation. It is however envisaged that the results presented in

this section can serve as a starting point for a complete analysis of haplogroup distribution within Africa. The geographic distribution of haplogroups and sub-haplogroups provided in Table 5.37 can aid in the design of a systematic sampling strategy whereby specific geographic regions in Africa are targeted in order to collect samples which represent specific haplogroups, or to select regions for which sequence data is lacking.

5.10 L0-SPECIFIC HAPLOGROUP NETWORK

Mitochondrial DNA alterations accumulate sequentially along maternal mtDNA lineages (Ruiz-Pesini *et al.*, 2007). A haplotype therefore evolves over time through the sequential addition of alterations to form groups of related haplotypes referred to as haplogroups. All mtDNA sequences derived from the original haplotype therefore share a set of common polymorphisms and these mtDNA sequences group together in specific branches of the mtDNA tree. These evolutionary relationships between mtDNA sequences in different haplogroups can be represented by haplogroup networks.

Haplogroup networks are constructed with the aim to evaluate variation identified in new mtDNA genome sequences relative to the rCRS (Andrews *et al.*, 1999). The variation is evaluated by taking into consideration the association of nucleotide variants in the new sequence to other variants on particular branches in the human mtDNA tree, i.e. the haplogroup (refer to Section 5.8), the time during the human mtDNA radiation that the variant appeared (refer to Section 5.11.6) and the position of the variant in polypeptide or RNA genes (refer to Section 5.7).

In this section, two L0-specific haplogroup networks are discussed. The first L0 network (Khoi-San L0 network) was constructed using the Khoi-San sequences generated in the current investigation. The second L0 network (Global L0 network) constructed is a global L0-specific haplogroup network based on a network generated by Ruiz-Pesini *et al.* (2007) to which an additional 46 published and unpublished L0 sequences were added (refer to Section 4.13). According to Ruiz-Pesini *et al.* (2007), a global L0-specific haplogroup network can be used to:

- a. define sub-haplogroups by indicating which polymorphisms are common in L0 sub-haplogroups,
- b. estimate the ages of sub-haplogroups and/or lineages within the L0 haplogroup,
- c. identify new branching and new lineages created by the addition of new L0 sequences to the network,

- d. determine if a branch is geographic and/or population-specific,
- e. differentiate between ancient and recent mutations, and
- f. assist in determining which polymorphisms are likely to be pathogenic.

Potential pathogenic alterations are usually recent, and the functional severity of a possible pathogenic alteration can be estimated through the use of inter-species conservation indices (refer to Section 5.11.5.2) or by the effect of the substitution on RNA secondary structure (refer to Section 5.7) for tRNA and rRNA genes involved in protein synthesis (Ruiz-Pesini *et al.*, 2007).

5.10.1 Southern African Khoi-San L0-specific haplogroup network

An L0-specific haplogroup network was generated as described in Section 4.13 for 22 of the 30 complete Khoi-San mitochondrial genome sequences generated in the current investigation, as only 22 sequences out of the 30 Khoi-San mtDNA sequences could be classified as L0 (refer to Section 5.8). A maximum parsimony tree was generated for the 22 L0 Khoi-San mitochondrial genome sequences as described in Section 4.12.6 (not presented) and was used as a guide for the topological structure of the network. The network obtained in this manner is indicated in Figure 5.49.

The network was constructed by using SNPs identified within the coding region (bp 577 to 16,023), as well as insertions and deletions. The control region was excluded due to the increased possibility of reverse mutations within this region. All the alterations contained in the network (synonymous substitutions, non-synonymous substitutions, insertions and deletions) were identified by comparing the 22 Khoi-San mtDNA sequences with the rCRS (Andrews *et al.*, 1999). The nucleotide substitutions, insertions and deletions presented in Figure 5.49 are listed in numerical order and do not indicate the order in which the specific alteration occurred over time.

L0

*** 750 (12)**
769 (12)
825A (12)
1018 (12)
1048 (12)
2758 (16)
2885 (16)
3516A (ND1)
3594 (ND1)
4104 (ND1)
4312 (I)
5442 (F325L / ND2) 7
7028 (COI)
7146 (T415A / COI) 13
7256 (COI)
7521 [S (UCN)]
8468 (ATP8)
8655 (ATP6)
8701 (T59A / ATP6) 17
8860 (T112A / ATP6) 28
9042 (ATP6)
9347 (COIII)
9540 (COIII)
10398 (T114A / ND3) 21
10589 (ND4L)
10664 (ND4L)
10688 (ND4L)
10810 (ND4)
10873 (ND4)
10915 (ND4)
11719 (ND4)
11914 (ND4)
12007 (ND4)
12705 (ND5)
13105 (I257V / ND5) 17
13276 (M314V / ND5) 31
13506 (ND5)
13650 (ND5)
14766 (T7I / cytb) 19
15326 (T194A / cytb) 8

L0a

850 (12)
1243 (12)
1438 (12)
2706 (16)
2836A (16)
4541 (ND2)
4586 (ND2)
4769 (ND2)
4907 (ND2)
5811 (C)
6185 (COI)
6938 (COI)
7257 (I452V / COI) 10
8911 (ATP6)
8994 (ATP6)
9136 (I204V / ATP6) 8
9818 (COIII)
10499 (ND4L)
10920 (P54L / ND4) 5
11296 (ND4)
11299 (ND4)
11653 (ND4)
12070 (ND4)
13020 (ND5)
13590 (ND5)
13819 (F495L / ND5) 3
13928C (S531T / ND5) 4
14020 (ND5)
14182 (ND6)
14371 (ND6)
14374 (ND6)
10876 (ND4)
10939 (ND4)
(KS_CGR_083)
2650 (16)
(KS_CGR_084)
4771 (A101V / ND2) 37
(KS_CGR_120)
8222 (COII)
8922 (ATP6)
(KS_CGR_001)
5471 (ND2)
(KS_CGR_157)
1386 (12)
2706 (16)
3981 (ND1)
4025 (T240M / ND1) 12
4044 (ND1)
5515 (W)
7154 (COI)
8865 (ATP6)
8994 (ATP6)
9932 (COIII)
10084 (I9T / ND3) 12
11854 (ND4)
13386 (ND5)
13466C (S377T / ND5) 30
14302 (ND6)
14693 (E)
14770 (cytb)
15766 (cytb)
15942 (T)
(KS_CGR_044)

L0b

3756 (ND1)
4232 (I309T / ND1) 31
6815 (COI)
9755 (COIII)
12121 (ND4)
12720 (ND5)
15466 (cytb)
15930 (T)
8113A (COII)
8152 (COII)
8251 (COII)
6185 (COI)
6266 (COI)
719 (12)
3438 (ND1)
4769 (ND2)
13759 (A475T / ND5) 1
15941 (T)
L0bA
L0bB
L0bB2
L0bB3
L0bB1

6185 (COI)
6266 (COI)
719 (12)
3438 (ND1)
4769 (ND2)
13759 (A475T / ND5) 1
15941 (T)
L0bA
L0bB
L0bB2
L0bB3
L0bB1

4197 (ND1)
9150 (ATP6)
11437 (ND4)
12235 [S (AGY)]
13129 (P265S / ND5) 36
15951 (T)
3591 (ND1)
9111 (ATP6)
13347 (ND5)
(KS_CGR_005)
9438 (G78S / COIII) 36
15550 (cytb)
7967 (COII)
15519 (L258P / cytb)
(KS_CGR_091)
(KS_CGR_117)
12469 (I45V / ND5)
(KS_CGR_092)
13260 (ND5)
(KS_CGR_111)
7828 (COII)
15106 (cytb)
(KS_CGR_061)
7828 (COII)
(KS_CGR_127)
3666 (ND1)
593 (F)
(KS_CGR_004)
7961 (COII)
(KS_CGR_077)
(KS_CGR_121)
8214A (V210D / COII)
8223G (L213W / COII)
(KS_CGR_105)
957 (12)
5460 (A331T / ND2) 2
8715 (ATP6)
8970 (ATP6)
9111 (ATP6)
10920 (P54L / ND4) 5
11087 (F110L / ND4) 39
12303 [L (CUN)]
14211 (V21I / ND6) 31
6692 (COI)
8790 (ATP6)
9887 (COIII)
14280 (S44P / ND6) 20
14560 (ND6)
15449 (F235L / cytb) 6
(KS_CGR_056)
10499 (ND4L)
10876 (ND4)
10939 (ND4)
(KS_CGR_060)
15106 (cytb)
(KS_CGR_048)

10499 (ND4L)
10876 (ND4)
10939 (ND4)
(KS_CGR_060)
15106 (cytb)
(KS_CGR_048)

333

Synonymous substitutions are indicated in **red** and represent transitions, except where indicated otherwise. An example is the transversion of a C to an A at position 825 indicated as 825A. The gene in which the substitution occurs is indicated in brackets. The non-synonymous substitutions are indicated in black, followed in brackets by the single letter abbreviations and the position of the amino acids involved in the non-synonymous change, and the gene in which the substitution is located. Recurrent alterations are indicated in **bold** text, and reversions are indicated in *italic* text. Conservation numbers were determined as described in Section 5.11.5.2 and indicated after each non-synonymous substitution. Substitutions in rRNA and tRNA molecules are indicated in **green**, with the specific rRNA (12S or 16S) or tRNA (single letter abbreviation for the amino acid accepted by the tRNA) indicated in brackets.

The 22 Khoi-San mitochondrial sequences grouped in two main branches. The branches consisted of individuals classified as L0a and L0b, respectively. Within the L0a branch, two groups, i.e. L0aA1 and L0aA2, were observed. Two groups were also present in the L0b branch, namely L0bA and L0bB. The L0bB group was further subdivided into an additional three groups, namely L0bB1, L0bB2 and L0bB3. The Khoi-San individuals observed in each group is listed in Table 5.38.

Table 5.38 Location of Khoi-San sequences in Khoi-San L0 network

Group	Khoi-San individual	Group	Khoi-San individual
L0aA1	KS_CGR_059, KS_CGR_084, KS_CGR_120	L0bB1	KS_CGR_048, KS_CGR_056, KS_CGR_060
L0aA2	KS_CGR_001, KS_CGR_083, KS_CGR_157, KS_CGR_176	L0bB2	KS_CGR_004, KS_CGR_005, KS_CGR_077, KS_CGR_091, KS_CGR_092, KS_CGR_105, KS_CGR_111, KS_CGR_117, KS_CGR_121
L0bA	KS_CGR_044	L0bB3	KS_CGR_061, KS_CGR_127

These branches and groups were similar to the results obtained through the phylogenetic analysis that is presented in Section 5.9. The observed branches and groups in the network presented in Figure 5.49 were further investigated by adding the 22 Khoi-San mitochondrial genome sequences generated in the current investigation to a global L0-specific haplogroup network.

5.10.2 Global L0-specific haplogroup network

A global L0-specific haplogroup network was constructed using the L0 haplogroup section of a sequential mutational phylogenetic tree available on MITOMAP (Ruiz-Pesini *et al.*,

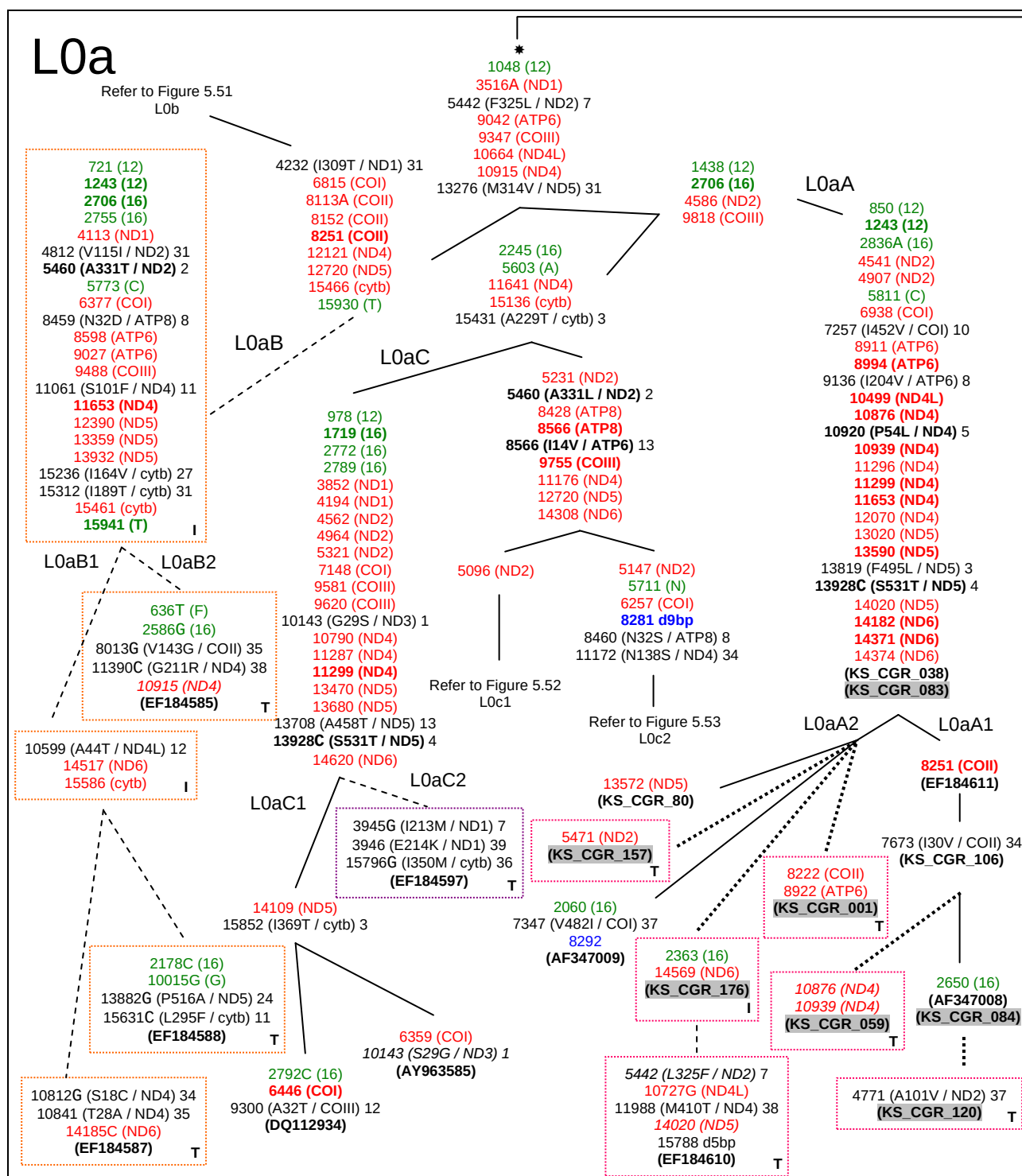
2007; refer to Appendix K). The L0 haplogroup tree obtained from MITOMAP was constructed as follows by Ruiz-Pesini *et al.* (2007). The rCRS was used as a reference, and SNPs were identified in the coding region of a total of 54 L0 mtDNA genome sequences obtained from the literature (Horai *et al.*, 1995; Ingman *et al.*, 2000; Maca-Meyer *et al.*, 2001; Torroni *et al.*, 2001; Herrnstadt *et al.*, 2002a; Mishmar *et al.*, 2003; Mouton, 2003, Howell *et al.*, 2004 and Kivisild *et al.*, 2006). Certain sequences obtained from the literature (AF381992, DQ112696, DQ112697, DQ112921 and DQ112922) were excluded due to the presence of multiple potential sequencing errors, as indicated by nucleotide alterations inconsistent with the most probable haplogroup classification. Ruiz-Pesini and colleagues determined the topology of the tree for the L0 haplogroup using the neighbour-joining method and MEGA version 2 software (Kumar *et al.*, 2001) with 1,000 bootstrap replicates. Thereafter, the positions of the individual nucleotide alterations were incorporated into the tree at the various branches and nodes by using sequence alignment and sequence comparison. New mtDNA sequences, such as the 22 Khoi-San mitochondrial genome sequences generated in the current investigation, can therefore be manually added into the existing tree without the need for additional computer analyses.

The L0 haplogroup network obtained from MITOMAP (2007), consisting of 54 L0 mtDNA genome sequences, was subsequently expanded manually as described in Section 4.13 through the addition of a total of 46 L0 sequences, so that the newly constructed global L0 network consisted of a total of 100 sequences. The 46 L0 sequences added included the 22 Southern African Khoi-San L0 mtDNA sequences generated in the current investigation, eight L0 African-American sequences (Just *et al.*, 2008), one L0 sequence from the Dominican Republic (Torroni *et al.*, 2006) and 15 L0 sequences from southern Africa and Tanzania (Gonder *et al.*, 2007). The set of 100 mitochondrial genome sequences used for the construction of the global L0-specific haplogroup network corresponds to dataset 8b which was compiled as described in Section 4.10.3. The global L0 haplogroup network with additional branching is divided into sections based on sub-haplogroup designation and the different parts of the network are depicted in Figures 5.50, 5.51, 5.52 and 5.53. A single figure consisting of all the different parts of the network presented in Figures 5.50, 5.51, 5.52 and 5.53 is displayed in Appendix L. The global L0 network follows the same conventions as the network depicted in Figure 5.49, regarding the colours by which polymorphisms are represented, the use of **bold** or *italic* text, and the descriptions of genes wherein these alterations are located. Sequences generated in the current investigation are highlighted in grey for ease of identification

within the network, and new branching within the network compared to the original L0 haplogroup network (Ruiz-Pesini *et al.*, 2007) is indicated by dashed lines. Branches within each sub-haplogroup are named according to the estimated TMRCA for that branch (refer to Section 5.11.6). This implies that branch A is the oldest branch in the sub-haplogroup, while B3 is the youngest of the three groups belonging to branch B of the same sub-haplogroup. As before, the nucleotide substitutions, insertions and deletions presented in Figures 5.50, 5.51, 5.52 and 5.53 are listed in numerical order and provide no information regarding the order in which the specific alterations occurred over time. Conservation numbers were calculated as described in Section 4.14.6.2 and indicated after each non-synonymous substitution. The possible functional effect and location of non-synonymous substitutions with high conservation numbers are discussed in Section 5.11.5.2.

5.10.2.1 Observed branching in global L0-specific haplogroup network

The L0a section of the global L0-specific haplogroup network is presented in Figure 5.50. Three branches are visible within the L0a sub-haplogroup, L0aA, L0aB and L0aC. The addition of seven Southern African Khoi-San sequences generated in the current investigation, as well as one South African San sequence, EF184610 (Gonder *et al.*, 2007), led to additional branching in the L0aA branch. The additional branches are indicated by pink boxes and include five terminal branches and one internal branch. The terminal branch consisting of KS_CGR_059 is composed of two SNPs which have reverted to the state present in the rCRS (10876A and 10939C). Further branching at this terminal branch is not expected as this sequence seems to be an exception regarding the SNPs present in sequences belonging to the L0aA branch as all other L0aA sequences contain 10876G and 10939T. The terminal branch consisting of KS_CGR_120 is an expansion of the branch containing KS_CGR_084 and KS_CGR_106, and led to further delineation of the branch by converting the branch consisting of KS_CGR_084 to an internal branch. The terminal branch consisting of EF184610 expanded the internal branch produced by KS_CGR_176, leading to further delineation of the branching structure in this section of the L0aA branch. The remaining two terminal branches consisting of KS_CGR_001 and KS_CGR_157 could, with the addition of a greater number of L0 sequences, expand in a similar manner to delineate the branching within L0aA further.

Figure 5.50 L0a section of the global L0-specific haplogroup network after addition of 46 L0 sequences

At present, it appears that two groups can be distinguished within the L0aA branch. These groups are named L0aA1 and L0aA2 and correspond to the two groups observed in the L0a branch in the L0-specific haplogroup network based on 22 Khoi-San sequences generated in the current investigation (refer to Figure 5.49). The L0aA1 group consists of six L0 sequences (AF347008, EF184611, KS_CGR_059, KS_CGR_084, KS_CGR_106 and KS_CGR_120) and the remaining L0 sequences in L0aA belong to the L0aA2 group. The L0aA2 group consists of a number of short branches separated by single SNPs. This indicates that the L0aA branch could be expanded by the addition of more L0 sequences and this could lead to further subdivision of L0aA into more groups which would represent additional lineages within the L0aA sub-haplogroup. The inclusion of the Khoi-San sequences generated in the current investigation has therefore increased the branching structure observed within the L0aA branch and with the addition of more L0aA sequences it is likely that these additional branches will be resolved further to give rise to distinct lineages.

The addition of three Sandawe L0 sequences from Tanzania, namely EF184585, EF184587 and EF184588 (Gonder *et al.*, 2007), to the global L0-specific haplogroup network produced an additional branch, L0aB, in the L0a sub-haplogroup. This branch contains two new internal branches and three new terminal branches. The new branches are indicated by orange boxes (refer to Figure 5.50). The L0aB branch corresponds to a branch identified within the global neighbour-joining and maximum parsimony phylogenetic trees constructed in Section 5.9 using dataset 1a (All L Sequences) that were presented in Figures 5.42 and 5.45. The L0 sequences located in this branch are classified as L0a, however, based on the location of the branch in the phylogenetic trees, it can be observed that these sequences lie outside another branch identified as L0a and are located closer to the L0b branch in the trees. In the L0 network, these sequences lie between L0a and L0b, and share a number of characteristic SNPs with L0b. The location of the L0aB branch within the phylogenetic trees and the global L0-specific haplogroup network requires the definition of a new L0 sub-haplogroup, as two branches with such different locations within the phylogenetic tree and the global L0-specific haplogroup network should not both be described as L0a (refer to Section 5.12).

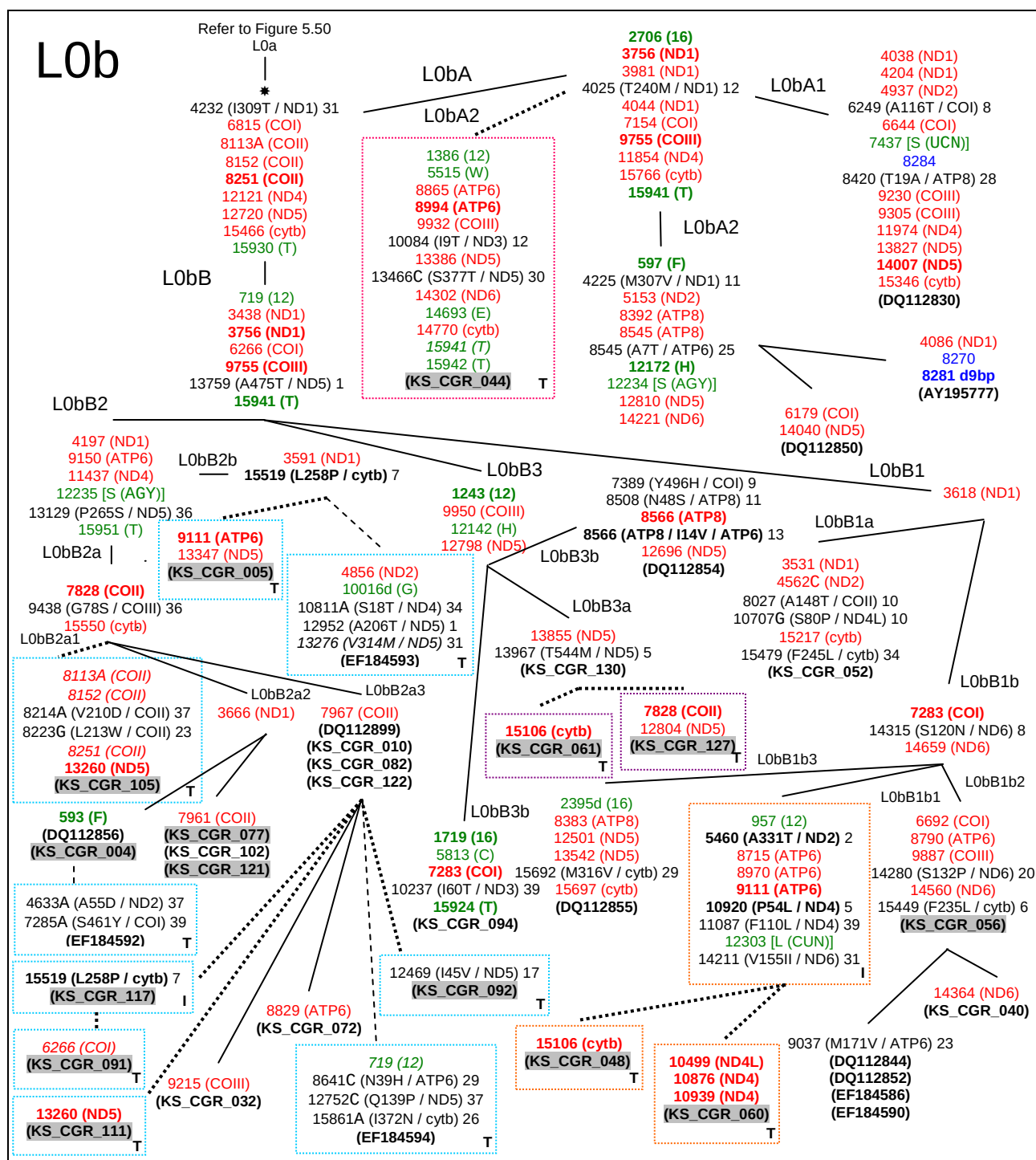
At present the L0aB branch contains two groups, L0aB1 and L0aB2. Both these groups contain L0 sequences from the Sandawe. Based on the global L0-specific haplogroup network, it can therefore be speculated that the L0aB branch is specific to the Sandawe population. However, in the global N-J and MP trees, the three Sandawe L0 sequences lie

in a group that also includes a sequence from an individual belonging to the Burunge ethnic group while L0 sequences from the Sandawe also group in other branches of the trees. More L0 sequences from East Africa and specifically the Sandawe population are required to determine if the L0aB branch represents a lineage that may be specific to a sector of the Sandawe population and to further delineate the two groups that are observed within the branch.

One additional L0 sequence from Tanzania, EF184597 (Gonder *et al.*, 2007), was added to the L0aC branch. One new terminal branch, indicated by a **violet** box (refer to Figure 5.50) is observed. The other sequences, AY963585 and DQ112934, present in the L0aC branch are of Ugandan and Ethiopian origin, respectively. To date a limited number of complete L0 sequences have been generated for individuals from East Africa (Horai *et al.*, 1995; Macaulay *et al.*, 2005; Kivisild *et al.*, 2006; Torroni *et al.*, 2006; Gonder *et al.*, 2007). The addition of more L0 sequences from East Africa to the global L0-specific haplogroup network can therefore be expected to lead to additional branching in the L0aC branch. It is possible that the new terminal branch consisting of EF184597 could expand to become a separate lineage within the L0aC branch.

The L0b section of the global L0-specific haplogroup network is presented in Figure 5.51. Two branches are visible within the L0b sub-haplogroup, namely L0bA and L0bB. This observation corresponds with the two groups observed in the L0b branch in the L0-specific haplogroup network based on 22 Khoi-San sequences generated in the current investigation (refer to Figure 5.49). The addition of one Southern African Khoi-San sequence, KS_CGR_044, to the L0bA branch led to the formation of a new terminal branch indicated by a **pink** box (refer to Figure 5.51). Due to the large number of SNPs contained in this terminal branch, it is expected that the addition of more L0 sequences will lead to further delineation of this branch. At present three groups are visible within the L0bA branch, however, based on branching order observed in the global phylogenetic trees (refer to Section 5.9), it is likely that only two groups will form within the L0bA branch after the addition of more L0 sequences.

Figure 5.51 LOb section of the global L0-specific haplogroup network after addition of 46 L0 sequences



Adapted from Ruiz-Pesini *et al.* (2007). The starting point of the network is indicated by *. Colour coding of polymorphisms: black = non-synonymous substitution (amino acid, position, amino acid, gene), red = synonymous substitution (gene), green = tRNA or rRNA (gene), blue = non-coding nucleotide in coding region. All alterations are transitions except where indicated otherwise. Sequences used to construct the network were obtained from the following sources: Horai *et al.* (1995), Ingman *et al.* (2000), Maca-Meyer *et al.* (2001), Torroni *et al.* (2001), Herrnstadt *et al.* (2002a), Mishmar *et al.* (2003), Mouton (2003), Howell *et al.* (2004), Macaulay *et al.* (2005), Kivisild *et al.* (2006), Torroni *et al.* (2006), Gonder *et al.* (2007), Just *et al.* (2008) and the current investigation. The names of sequences are indicated at the bottom of the branch in brackets and in bold text. GenBank® accession numbers are indicated where available. Sequences generated in the current investigation are highlighted in grey. New branches within the network are indicated by dashed lines, where big dashes (-----) represent branches added by published sequences and small dashes (-----) indicate branches added by sequences generated in the current investigation. Polymorphisms in bold text signify that the polymorphism occurs in multiple branches, and polymorphisms in italic text indicate that the polymorphism reverted to the nucleotide present in the rCRS. Conservation numbers are indicated after each non-synonymous substitution. Nucleotide abbreviations: A = adenine; C = cytosine; G = guanine; U = uracil; Y = C or U; N = A, C, G or U. Amino acid abbreviations: A = alanine; C = cysteine; D = aspartic acid; E = glutamic acid; F = phenylalanine; G = glycine; H = histidine; I = isoleucine; L = leucine; M = methionine; N = asparagine; P = proline; Q = glutamine; S = serine; T = threonine; V = valine; W = tryptophan; Y = tyrosine. Gene abbreviations: 12 = 12S rRNA; 16 = 16S rRNA; ATP = ATP synthase F₀ subunit; CO = cytochrome c oxidase subunit; cytb = cytochrome b; ND = NADH dehydrogenase subunit. Other abbreviations: I = internal branch; T = terminal branch; d = deletion; bp = base pairs.

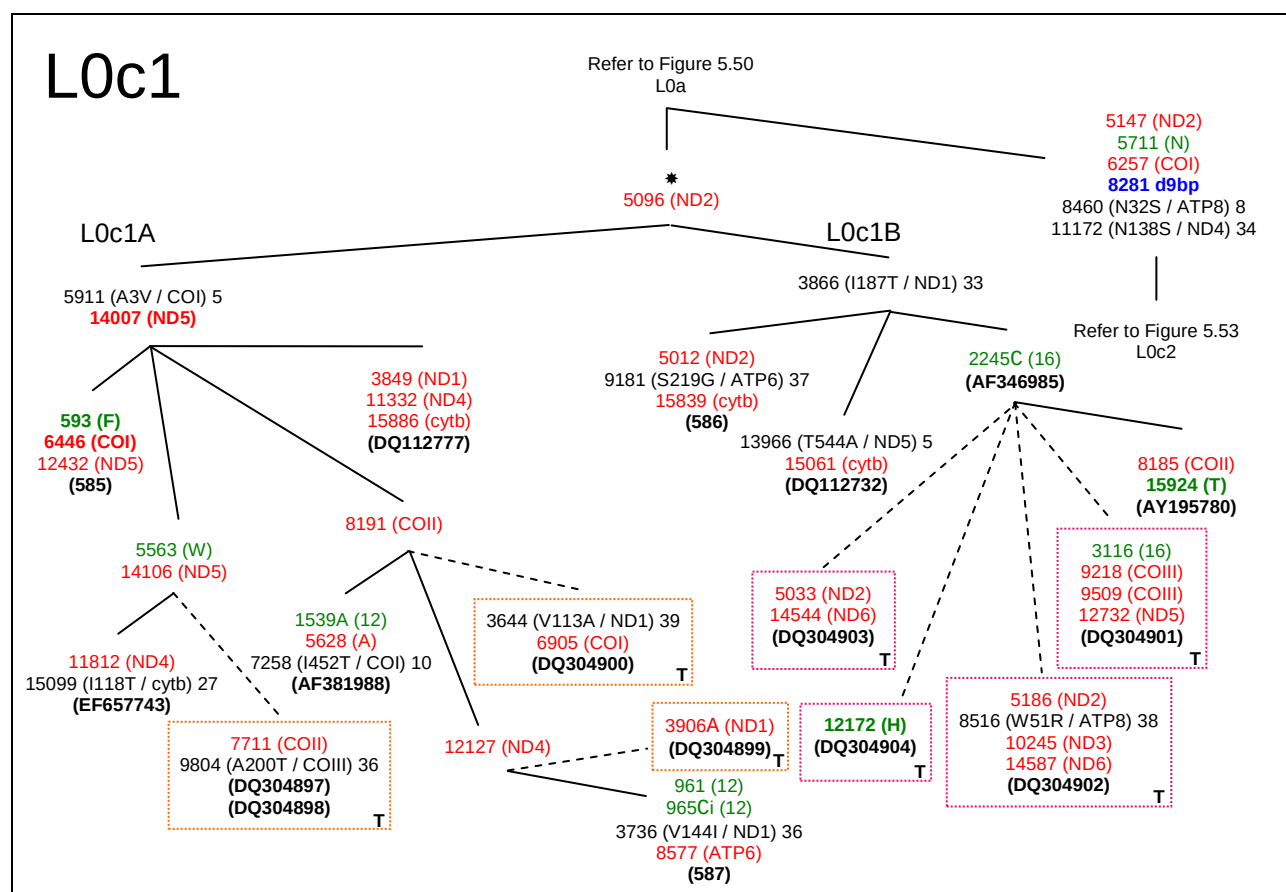
The L0bB branch contained three groups, i.e. L0bB1, L0bB2 and L0bB3, which were similar to the three groups observed in the L0bB branch in the L0-specific haplogroup network based on 22 Khoi-San sequences generated in the current investigation (refer to Figure 5.49). The L0bB1 branch was further delineated by the addition of four L0 sequences, which included two Southern African Khoi-San sequences, KS_CGR_048 and KS_CGR_060 (current investigation), one South African San sequence, EF184590 (Gonder *et al.*, 2007), and one Sandawe sequence from Tanzania, EF184586 (Gonder *et al.*, 2007). The additional sequences led to the formation of one new internal branch and two new terminal branches containing KS_CGR_048 and KS_CGR_060 indicated by orange boxes (refer to Figure 5.51). At present two groups can be distinguished within the L0bB1 branch, i.e. L0bB1a and L0bB1b and it is possible that this branch and the groups within will expand and become more clearly defined by the addition of more L0 sequences.

The addition of nine Southern African Khoi-San sequences (current investigation) and three South African San sequences (Gonder *et al.*, 2007) led to further delineation of the branching pattern observed in L0bB2. Eight new terminal branches and one new internal branch are observed and are indicated by light blue boxes (refer to Figure 5.51). The L0bB2 branch can be divided into two main lineages named L0bB2a and L0bB2b. The L0bB2a lineage is composed of three groups. The first group, L0bB2a1, consists of a new terminal branch generated by a single L0 sequence, KS_CGR_105. The second group, L0bB2a2, contains one new terminal branch formed by the addition of a single sequence, EF184592. The third group, L0bB2a3, consists of one new internal branch containing a single sequence, KS_CGR_117, and four new terminal branches, each due to the contribution of a single sequence (EF184594, KS_CGR_091, KS_GCR_092 and KS_CGR_111). Within the L0bB2a lineage, it is not expected that terminal branches consisting of two or less SNPs will expand after addition of more L0 sequences. The terminal branch containing KS_CGR_091 in lineage L0bB2a3 is formed due to the reversion of the SNP to the state observed in the rCRS and is an exception with regards to the SNPs expected to be present in the L0bB branch. It is therefore not expected that this terminal branch will be expanded further by the addition of more L0 sequences. Terminal branches consisting of more than two SNPs such as the branches generated by EF184594 and KS_CGR_105, are expected to be expanded by the addition of more L0 sequences thereby becoming internal branches. The L0bB2b lineage contains two new terminal branches, each containing one L0 sequence, EF184593 and KS_CGR_005. It is expected that the L0bB2b branch will be expanded by the addition of more L0 sequences.

The L0bB3 branch was expanded by the addition of two Southern African Khoi-San sequences generated in the current investigation. The two sequences, KS_CGR_061 and KS_CGR_127, gave rise to two new terminal branches, indicated by **violet** boxes (refer to Figure 5.51), and converted the branch consisting of KS_CGR_130 to an internal branch. The L0bB3 branch contains two lineages, L0bB3a and L0bB3b, and it is expected that the addition of more L0 sequences will delineate this branch further and define the lineages therein.

The L0c1 section of the global L0-specific haplogroup network is presented in Figure 5.52. Two branches are visible within the L0c1 sub-haplogroup, namely L0c1A and L0c1B.

Figure 5.52 L0c1 section of the global L0-specific haplogroup network after addition of 46 L0 sequences



Adapted from Ruiz-Pesini *et al.* (2007). The starting point of the network is indicated by *. Colour coding of polymorphisms: black = non-synonymous substitution (amino acid, position, amino acid, gene), red = synonymous substitution (gene), green = tRNA or rRNA (gene), blue = non-coding nucleotide in coding region. All alterations are transitions except where indicated otherwise. Sequences used to construct the network were obtained from the following sources: Horai *et al.* (1995), Ingman *et al.* (2000), Maca-Meyer *et al.* (2001), Torroni *et al.* (2001), Herrnstadt *et al.* (2002a), Mishmar *et al.* (2003), Mouton (2003), Howell *et al.* (2004), Macaulay *et al.* (2005), Kivisild *et al.* (2006), Torroni *et al.* (2006), Gonder *et al.* (2007), Just *et al.* (2008) and the current investigation. The names of sequences are indicated at the bottom of the branch in brackets and in **bold** text. GenBank[®] accession numbers are indicated where available, in other cases the sequence is named according to the name given by the author e.g. 585. New branches within the network due to the addition of published sequences are indicated by lines consisting of big dashes (---). Polymorphisms in **bold** text signify that the polymorphism occurs in multiple branches. Conservation numbers are indicated after each non-synonymous substitution. Nucleotide abbreviations: A = adenine; C = cytosine. Amino acid abbreviations: A = alanine; C = cysteine; F = phenylalanine; G = glycine; H = histidine; R = arginine; T = threonine; V = valine; W = tryptophan. Gene abbreviations: 12 = 12S rRNA; 16 = 16S rRNA; ATP = ATP synthase F₀ subunit; CO = cytochrome c oxidase subunit; cytb = cytochrome b; ND = NADH dehydrogenase subunit. Other abbreviations: I = internal branch; T = terminal branch; d = deletion; i = insertion; bp = base pairs.

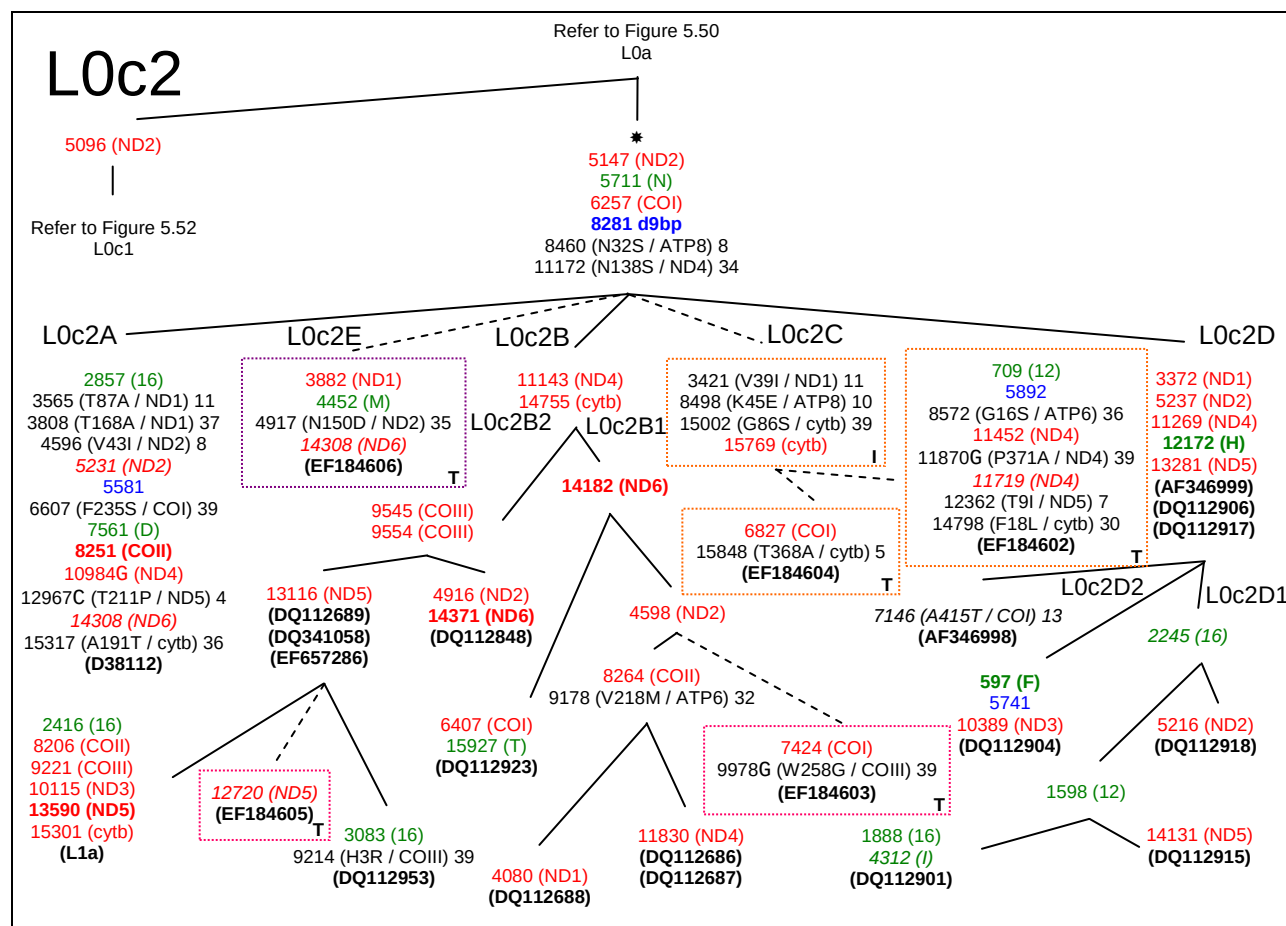
The L0c1A branch was further delineated by the addition of four African American L0 sequences (Just *et al.*, 2008) that gave rise to three new terminal branches indicated by orange boxes (refer to Figure 5.52). The L0c1B branch was expanded by the addition of four African American L0 sequences (Just *et al.*, 2008) that formed four new terminal branches indicated by pink boxes (refer to Figure 5.52). Both the L0c1A and L0c1B branches consist of a number of short branches separated by more than one SNP. This indicates that both branches could be further delineated by the addition of more L0 sequences. This could lead to further subdivision of both branches into more groups. It is likely, based on the branching order observed in the global MP tree (refer to Section 5.9.3.1), that three groups will form within the L0c1A branch and that two groups will form within the L0c1B branch. These groups would represent additional lineages within the L0c1 sub-haplogroup.

The L0c2 section of the global L0-specific haplogroup network is presented in Figure 5.53. Five branches are visible within the L0c2 sub-haplogroup namely L0c2A, L0c2B, L0c2C, L0c2D and L0c2E. No L0 sequences were added to the L0c2A branch and this branch currently consists of one sequence, i.e. D38112, from an individual from Uganda in East Africa. It is possible that sequences belonging to this branch are rare and have escaped sampling in the populations investigated to date. Alternatively, due to the paucity of sequences from East Africa, sequences that would group in this branch did not form part of the available sequence datasets due to random sampling (Horai *et al.*, 1995; Macaulay *et al.*, 2005; Kivisild *et al.*, 2006; Torroni *et al.*, 2006; Gonder *et al.*, 2007). The addition of more L0 sequences from East Africa should provide additional data to alleviate this problem.

Three sequences, i.e. one L0 sequence from the Dominican Republic, DQ341058 (Torroni *et al.*, 2006), one Pare L0 sequence from Tanzania, EF184603 (Gonder *et al.*, 2007) and one Turu L0 sequence from Tanzania, EF184605 (Gonder *et al.*, 2007), were added to the L0c2B branch. Two new terminal branches are observed and are indicated by pink boxes (refer to Figure 5.53). It is not expected that the terminal branch consisting of EF184605 will expand with the addition of more L0 sequences as this terminal branch contains one SNP that has reverted to the state observed in the rCRS. The other terminal branch consisting of EF184603 obtained from an individual belonging to the Pare ethnic group (Gonder *et al.*, 2007) could possibly expand if more L0 sequences from the same ethnic group were added to the global L0-specific haplogroup network. At present, two groups

can be observed within the L0c2B branch, namely L0c2B1 and L0c2B2, and the addition of more L0 sequences to this branch will assist in defining these two groups further.

Figure 5.53 L0c2 section of the global L0-specific haplogroup network after addition of 46 L0 sequences



Adapted from Ruiz-Pesini *et al.* (2007). The starting point of the network is indicated by *. Colour coding of polymorphisms: black = non-synonymous substitution (amino acid, position, amino acid, gene), red = synonymous substitution (gene), green = tRNA or rRNA (gene), blue = non-coding nucleotide in coding region. All alterations are transitions except where indicated otherwise. Sequences used to construct the network were obtained from the following sources: Horai *et al.* (1995), Ingman *et al.* (2000), Maca-Meyer *et al.* (2001), Torroni *et al.* (2001), Herrnstadt *et al.* (2002a), Mishmar *et al.* (2003), Mouton (2003), Howell *et al.* (2004), Macaulay *et al.* (2005), Kivisild *et al.* (2006), Torroni *et al.* (2006), Gonder *et al.* (2007), Just *et al.* (2008) and the current investigation. The names of sequences are indicated at the bottom of the branch in brackets and in **bold** text. GenBank® accession numbers are indicated where available, in other cases the sequence is named according to the name given by the author e.g. L1a. New branches within the network due to the addition of published sequences are indicated by dashed lines consisting of big dashes (---). Polymorphisms in **bold** text signify that the polymorphism occurs in multiple branches, and polymorphisms in *italic* text indicate that the polymorphism reverted to the nucleotide present in the rCRS. Conservation numbers are indicated after each non-synonymous substitution. Nucleotide abbreviations: C = cytosine; G = guanine. Amino acid abbreviations: A = alanine; D = aspartic acid; F = phenylalanine; G = glycine; H = histidine; I = isoleucine; L = leucine; M = methionine; N = asparagine; P = proline; R = arginine; S = serine; T = threonine; V = valine; W = tryptophan. Gene abbreviations: 12 = 12S rRNA; 16 = 16S rRNA; ATP = ATP synthase F₀ subunit; CO = cytochrome c oxidase subunit; cytb = cytochrome b; ND = NADH dehydrogenase subunit. Other abbreviations: I = internal branch; T = terminal branch; d = deletion; bp = base pairs.

Two sequences, EF184602 and EF184604, obtained from Sandawe individuals from Tanzania (Gonder *et al.*, 2007), were added to the L0c2C branch. This gave rise to the formation of one internal and two terminal branches indicated by orange boxes (refer to Figure 5.53). It is expected that the addition of more L0 sequences to this branch will lead to further delineation of the branching structure as both terminal branches contain more than one SNP.

No sequences were added to the L0c2D branch which can be divided into two groups, L0c2D1 and L0c2D2. It is possible that sequences belonging to this branch are rare, or are confined to a specific population. This branch consists of Pygmies from Zaire (now referred to as the Democratic Republic of the Congo). This population was not sampled in the current investigation and no new sequences from this population were available to add to the global L0-specific haplogroup network at the time when the network was constructed. Although, based on Figure 5.53, it appears that this branch may be confined to the Pygmy population. The possible population-specific nature of this branch, as well as the observed groups therein, can only be confirmed in the future by the addition of more L0 sequences from Pygmies originating from the Democratic Republic of the Congo.

The L0c2E branch represents a new branch in the L0c2 section of the global L0-specific haplogroup network and was generated by the addition of one L0 sequence, EF184606, from a Sandawe individual from Tanzania (Gonder *et al.*, 2007). The new branch is indicated by a **violet** box (refer to Figure 5.53). It can be proposed that this branch represents a separate lineage within the Sandawe group. This will however have to be confirmed by the addition of more L0 sequences from East Africa and specifically the Sandawe population.

The five branches within the L0c2 sub-haplogroup consist of a number of short branches. This indicates that the branches within the L0c2 sub-haplogroup could be expanded by the addition of more L0 sequences and this could lead to further subdivision of the branches in L0c2 which would represent additional lineages within the L0c2 sub-haplogroup.

The addition of 46 L0 sequences to the original L0 haplogroup network consisting of 54 L0 sequences (Ruiz-Pesini *et al.*, 2007) gave rise to new branches, internal branches and terminal branches, which represent new groups and lineages within each of the L0 sub-haplogroups discussed above. The 22 Khoi-San sequences generated in the current investigation were added to the L0a and L0b sub-haplogroups and expanded previously identified branches (L0aA1, L0aA2, L0bA2, L0bB2a2, L0bB2a3, L0bB2b and L0bB3a) and gave rise to two previously unidentified branches (L0bB1b1 and L0bB2a1). The expanded branches and new branches generated by the addition of the Khoi-San sequences represent potential new lineages in the L0 haplogroup and is an indication of the genetic diversity present in the Southern African Khoi-San population. To date, the sequence data collected for this population has been limited to control region sequence data, i.e. HVS1 and HVS2 (Vigilant *et al.*, 1989), RFLP data (Chen *et al.*, 2000), coding

region only sequence data for five individuals (Kivisild *et al.*, 2006) and 23 complete sequences (Ingman *et al.*, 2000; Mouton, 2003; Gonder *et al.*, 2007). The coding region sequence data from 22 Khoi-San individuals added to the global L0-specific haplogroup network in the current investigation, therefore increases our understanding of the genetic diversity present in the L0 haplogroup, and makes a substantial contribution to the “global base of knowledge” for Khoi-San sequence data.

The global L0-specific haplogroup network presented in this section can be used to determine the sub-haplogroup classification of a new L0 sequence by adding the sequence to the network and tracing nucleotide alterations in the new sequence along the branching order of the network (Ruiz-Pesini *et al.*, 2007). The network can also be used to estimate the time when an observed alteration arose, i.e. ancient alterations are present in internal branches, while recent alterations are usually located in terminal branches (Ruiz-Pesini *et al.*, 2007). The effect of sequence alterations on genes can also be investigated by using the global L0-specific haplogroup network (Ruiz-Pesini *et al.*, 2007). If an L0 sequence is obtained from an individual affected by a maternally inherited mitochondrial disease, and a potential functionally significant alteration is identified, i.e. an alteration that involves a non-synonymous substitution with a high conservation index, or alters a conserved nucleotide, or influences the secondary structure of tRNA or rRNA molecules, it is likely that the observed alteration could be a potential candidate for a pathogenic mutation, if the alteration is not present in the L0-specific haplogroup network which is based on apparently healthy individuals.

The identification of new branches within the L0-specific haplogroup network has a number of implications on our understanding of the Southern African Khoi-San and the L0 haplogroup as a whole. Firstly, a number of branches, e.g. L0aB, appear to be specific to either a region in Africa, or a specific population and these putative relationships between sub-haplogroups or lineages and their geographic distribution in Africa are discussed in Section 5.10.2.2. Secondly, the identification of new branches in the L0-specific haplogroup network necessitates a re-evaluation of the haplogroup classification scheme used currently for the L0 haplogroup (refer to Section 5.8). The new branches and groups observed in the global L0-specific haplogroup network are therefore discussed further in Section 5.12 where a revised classification scheme is proposed for the L0 haplogroup in order to include the new information presented in the global L0-specific haplogroup network. Thirdly, the addition of more sequences to the global L0-specific haplogroup network and the emergence of new branches affect the estimated TMRCA of

the L0 haplogroup, and estimates of the TMRCAs determined for the different L0 sub-haplogroups and lineages are discussed in Section 5.11.6.

5.10.2.2 Geographic distribution of L0 sub-haplogroups and lineages

The ethnicity and/or geographic origin of the L0 sequences contained in the global L0-specific haplogroup network are listed in Appendix G. A map, indicating the main geographic regions in Africa, according to the United Nations Cartographic Section, is presented in Appendix I. It is important to note that Angola is classified as belonging to Central Africa in the original map used to construct the map indicated in Appendix I. However, most studies of mtDNA to date classify Angola as part of southern Africa (Chen *et al.*, 2000). In the current investigation, Angola is therefore also classified as belonging to southern Africa.

Branch L0aA contains 14 L0 sequences originating from countries in southern Africa such as Angola, Namibia and South Africa. It appears that this branch is specific to the Khoi-San population and to the southern African region. The L0aB branch is composed of three L0 sequences from the Sandawe of Tanzania located in East Africa (Gonder *et al.*, 2007). As mentioned previously, this branch appears to be specific to both population, namely the Sandawe, and region, i.e. Tanzania in East Africa. However, only three sequences are currently assigned to this branch and more L0 sequences from East Africa are required to clarify this relationship. The L0aC branch contains three L0 sequences from Tanzania, Uganda and Ethiopia, which are also located in East Africa. This branch also appears to be specific to East Africa, but more L0 sequences from East Africa are needed to test this hypothesis. It thus appears that the L0a sub-haplogroup has a wide distribution across Africa as it is observed in both East Africa and southern Africa. Specific branches within the L0a sub-haplogroup are however restricted to specific Khoi-San speaking populations such as the Sandawe and the Southern African Khoi-San represented by the !Kung ethnic group in the current investigation.

The L0bA branch contains three L0 sequences from southern African countries such as Angola and it appears that this branch is specific to southern Africa. More L0 sequences will however be needed to confirm this observation. The L0bB1 branch consists of eight L0 sequences from southern Africa, including countries such as Angola, Namibia and South Africa, and one L0 sequence from Tanzania in East Africa. This branch consists mainly of Khoi-San speakers and appears to be Southern African Khoi-San specific. The

L0bB2 branch contains 19 L0 sequences from southern Africa, including countries such as Angola, Namibia and South Africa. This branch consists of Khoi-San speakers only and appears to be specific to Khoi-San speakers from southern Africa. For both L0bB1 and L0bB2 branches more sequences from Southern African Khoi-San speakers are required to confirm that these branches are specific to this group. The L0bB3 branch consists of five L0 sequences from southern Africa, including countries such as Angola and Namibia. This branch is composed of L0 sequences from mostly Southern African Khoi-San speakers and appears to be Khoi-San specific. The L0bB branch therefore appears to contain mainly L0 sequences from Khoi-San speakers from southern Africa. The L0b sub-haplogroup as a whole appears to be confined to southern Africa and is observed mainly in the Southern African Khoi-San population.

The L0c1A branch consists mainly of L0 sequences from African Americans or individuals of African descent living in the UK (United Kingdom) or USA (Herrnstadt *et al.*, 2002a), and no additional information is available for the L0 sequences of these individuals regarding their place of origin in Africa. One L0 sequence in this branch, DQ112777, was obtained from an individual from Sudan in North Africa and the remaining L0 sequence in this branch, namely AF381988, was acquired from an individual from Morocco in North Africa (Maca-Meyer *et al.*, 2001; Kivisild *et al.*, 2006). Based on the limited geographical data for this branch, it is postulated that L0 sequences belonging to this branch will most likely be observed in a large geographic area stretching across North Africa from Morocco in the west to Sudan in the east.

The L0c1B branch is composed mainly of L0 sequences from African Americans or individuals of African descent living in the USA or UK (Herrnstadt *et al.*, 2002a). No additional information regarding their place of origin in Africa is available for these individuals. One L0 sequence in this branch, DQ112732, was obtained from an individual from Burkina Faso in West Africa (Kivisild *et al.*, 2006) and another L0 sequence was from a Hausa individual who likely originated from Nigeria or Niger in West Africa (Ingman *et al.*, 2000). It is speculated that L0 sequences belonging to this branch will be detected mainly in West Africa.

From the above observations it appears that both the L0c1A and L0c1B branches are likely to be specific to West Africa, although L0c1A has been detected in Sudan. The majority of L0 sequences in these two branches are from African Americans or individuals with African ancestry for whom no information is available regarding their country of origin

in Africa. It has however been determined by using high-resolution analysis that West Africa contributed $60\% \pm 3\%$, south-western Africa contributed $30\% \pm 2\%$ and west-central Africa contributed $9\% \pm 3\%$ to the mtDNA lineages observed today in the African American population (Salas *et al.*, 2005). A large number of African American individuals can therefore trace their ancestry to West Africa. This therefore supports a West African distribution for the L0c1 sub-haplogroup.

The L0c2A branch contains a single L0 sequence from Uganda in East Africa and it can be speculated that this branch is likely to be specific to East Africa. More L0 sequences are however required to confirm this and to expand this branch further. The L0c2B branch contains L0 sequences from five individuals from the Dominican Republic (DQ112686, DQ112687, DQ112688, DQ112689 and DQ341058), two individuals from Tanzania (EF184603 and EF184605), one Pygmy individual from the Central African Republic, one individual of African descent and one Asian individual. The presence of individuals from the Dominican Republic in this branch can likely be ascribed to the Atlantic Slave Trade which transported individuals from north-western, west-central and south-eastern Africa along the Atlantic coast (Fage, 1969). The same argument applies to the individual of African descent for which no information is available other than that the individual is from the USA (Herrnstadt *et al.*, 2002a). The presence of an Asian individual in this branch can be explained by assuming that the individual's maternal lineage is of African ancestry (Kivisild *et al.*, 2006). Although the individual is classified as Asian based on phenotype and cultural identity, admixture at some point in the individual's pedigree introduced an African mtDNA lineage into the individual's genetic composition. This branch therefore appears to have a wide distribution across Africa, including East, Central and West Africa.

The L0c2C branch contains two L0 sequences from Sandawe individuals (EF184602 and EF184604) from Tanzania in East Africa. It appears that this branch is both population- and region-specific, although the addition of more L0 sequences from the Sandawe from East Africa are required to determine if this grouping will remain. The L0c2D branch is composed of eight L0 sequences. All eight L0 sequences were obtained from Pygmy individuals from the Democratic Republic of the Congo. This branch appears to be specific for the Pygmy population from the Democratic Republic of Congo in Central Africa. The L0c2E branch consists of a single sequence, i.e. EF184606, obtained from a Sandawe individual from Tanzania in East Africa. More L0 sequences from East Africa will be required to determine if this branch is indeed specific to East Africa. The L0c2 sub-haplogroup thus appears to be widely distributed across Africa with branches in

East, Central and West Africa and is well represented in individuals with African ancestry present in countries outside of Africa. It can be speculated that the current distribution of the L0c1 and L0c2 sub-haplogroups both within and outside of Africa is linked to dispersals within Africa and outside of Africa brought about by the Atlantic Slave Trade.

In the current investigation, the distribution of L0 sub-haplogroups and branches within can not be confined to specific countries in Africa, although it has previously been reported that mtDNA haplogroups present with distinct geographic patterns in Africa (Rosa *et al.*, 2004). The distribution of L0 sub-haplogroups and branches within can however be localised to geographic regions within Africa such as the L0aA sub-haplogroup which is detected in southern Africa. The association between L haplogroups and geographic regions in Africa has been investigated in a number of other studies (refer to Section 3.3.1.2). Even with datasets that increase in size, and with more extensive phylogeographic analyses, it remains a problem to allocate mtDNAs from individuals with an African ancestry, such as African Americans, to narrow geographic locations in Africa. This is due to the fact that many mtDNAs are very widely distributed throughout the African continent, most likely as a result of the Bantu dispersal (Salas *et al.*, 2002), and both earlier and more recent movements in Africa, including those due to the Atlantic Slave Trade (Salas *et al.*, 2004a).

The results regarding the geographic distribution in Africa of L0 sub-haplogroups presented in this section are based on the global L0-specific haplogroup network constructed using dataset 8b. This dataset contains L0 sequences generated in the current investigation, as well as L0 sequences obtained from other investigations and GenBank® (refer to Section 4.10.3). The L0 sequences contained in the dataset were assigned to geographic regions in Africa based on information regarding the country of sample collection or the ethnicity of the individual where no information was supplied regarding the African country of origin (refer to Appendix G). Dataset 8b consisted of two sequences from North Africa (Morocco and Sudan), nine sequences from Central Africa (Democratic Republic of the Congo and Central African Republic), two sequences from West Africa (Burkina Faso and Nigeria), 13 sequences from East Africa (Ethiopia, Tanzania and Uganda), and 51 sequences from southern Africa (Angola, Namibia and South Africa). The remaining sequences in the dataset were obtained from Africa, but no additional information regarding ethnicity or region in Africa was available to assign the sequence to a specific geographic region, or from countries outside of Africa (Dominican Republic and United States of America). The majority of the sequences included in the dataset from Central Africa were obtained from Pygmy individuals, while the bulk of the

sequences from southern Africa were from Khoi-San and San individuals. As discussed in Section 5.9.4 for datasets 1a and 2a, the majority of L0 haplogroup sequence data in dataset 8b has therefore been obtained from only a few populations and countries in each region of Africa. Dataset 8b therefore represents a sample of convenience obtained from the available body of knowledge, i.e. the sample was not obtained in a systematic manner to ensure sampling of sequences of all L0 sub-haplogroups and from all countries in Africa. The results presented are therefore specific to dataset 8b.

5.10.2.3 Recurrent alterations and reversions observed in the global L0-specific haplogroup network

During analysis of the global L0-specific haplogroup network, it was observed that some alterations occurred in more than one sub-haplogroup, or in multiple branches within the same sub-haplogroup. These recurrent alterations, or parallel alterations, have been observed in studies investigating the control region (Bandelt *et al.*, 2001a), as well as in analyses of coding region sequences (Herrnstadt *et al.*, 2002a). The recurrent alterations observed in the current investigation involved synonymous substitutions, non-synonymous substitutions, as well as deletions, and occurred in polypeptide, tRNA and rRNA genes. The nucleotide positions, as well as the areas in the coding region where the alterations occurred, are indicated in Table 5.39.

Table 5.39 Nucleotide position and location of recurrent alterations in the global L0-specific haplogroup network based on the mtDNA coding region

Nucleotide position	Number of events	Nucleotide change	Region/gene	Amino acid substitution
593	2	T → C	tRNA ^{Phe}	---
597	2	C → T	tRNA ^{Phe}	---
1,243	3	T → C	12S rRNA	---
1,719	2	G → A	16S rRNA	---
2,706	3	A → G	16S rRNA	---
3,756	2	A → G	ND1	SYN
5,460	3	G → A	ND2	NON-SYN: Ala → Leu
6,446	2	G → A	COI	SYN
7,283	2	T → C	COI	SYN
7,828	2	A → G	COII	SYN
8,251	3	G → A	COII	SYN
8,281	2	---	Non-coding nucleotide in coding region	Deletion: 9bp

Table 5.39 Continued...

Nucleotide position	Number of events	Nucleotide change	Region/gene	Amino acid substitution
8,566 ^a	2	A → G	ATPase6	NON-SYN: Ile → Val
8,566 ^a	2	A → G	ATPase8	SYN
8,994	2	G → A	ATPase6	SYN
9,111	2	T → C	ATPase6	SYN
9,755	3	G → A	COIII	SYN
10,499	2	A → G	ND4L	SYN
10,876	2	A → G	ND4	SYN
10,920	2	C → T	ND4	NON-SYN: Pro → Leu
10,939	2	C → T	ND4	SYN
11,299	2	T → C	ND4	SYN
11,653	2	A → G	ND4	SYN
12,172	3	A → G	tRNA ^{His}	---
12,720	2	A → G	ND5	SYN
13,260	2	T → C	ND5	SYN
13,590	2	G → A	ND5	SYN
13,928	2	G → C	ND5	NON-SYN: Ser → Thr
14,007	2	A → G	ND5	SYN
14,182	2	T → C	ND6	SYN
14,371	2	T → C	ND6	SYN
15,106	2	G → A	Cytb	SYN
15,519	2	T → C	Cytb	NON-SYN: Leu → Pro
15,924	2	A → G	tRNA ^{Thr}	---
15,941	3	T → C	tRNA ^{Thr}	---

^a = a single nucleotide alteration at position 8,566 affects both the ATPase6 (non-synonymous substitution) and ATPase8 (synonymous substitution) genes. Nucleotide abbreviations: A = adenine; C = cytosine; G = guanine; T = thymine. Amino acid abbreviations: Ala = alanine; His = histidine; Ile = isoleucine; Leu = leucine; Phe = phenylalanine; Pro = proline; Ser = serine; Thr = threonine; Val = valine. Gene abbreviations: ATP = ATP synthase F₀ subunit; CO = cytochrome c oxidase subunit; Cytb = cytochrome b; ND = NADH dehydrogenase subunit; rRNA = ribosomal RNA; tRNA = transfer RNA; SYN = synonymous substitution; NON-SYN = non-synonymous substitution.

A total of 75 recurrent alterations were observed at 34 nucleotide positions. Alterations at seven nucleotide positions (1,243, 2,706, 5,460, 8,251, 9,755, 12,172 and 15,941) recurred three times, and the alterations at the remaining positions recurred twice. A total of 43 recurrent alterations involved A → G or G → A transitions, 32 recurrent alterations entailed C → T or T → C transitions, and one recurrent alteration resulted in a G → C transversion. The prevalence of A → G or G → A transitions at positions where recurrent alterations occurred corresponded to that observed by Galtier *et al.* (2006) and confirmed a role for these positions in the mtDNA coding region as hypervariable sites. The recurrent

alterations were scattered throughout the coding region and their distribution is presented in Table 5.40.

Table 5.40 Number of recurrent alterations per region/gene in the global L0-specific haplogroup network based on the mtDNA coding region

Region/gene	Number of recurrent alterations	Region/gene	Number of recurrent alterations
12S rRNA	3	ND1	2
16S rRNA	5	ND2	3
ATPase6	6	ND4	10
ATPase8	2	ND4L	2
COI	4	ND5	10
COII	5	ND6	4
COIII	3	tRNA ^{His}	3
Cytb	4	tRNA ^{Phe}	4
Non-coding nucleotides	2	tRNA ^{Thr}	5

Region/genes in which the highest numbers of recurrent alterations are present are indicated in **red**. Amino acid abbreviations: His = histidine; Phe = phenylalanine; Thr = threonine. Gene abbreviations: ATP = ATP synthase F₀ subunit; CO = cytochrome c oxidase subunit; Cytb = cytochrome b; ND = NADH dehydrogenase subunit; rRNA = ribosomal RNA; tRNA = transfer RNA.

The regions/genes in the coding region that contained the largest number of recurrent alterations were the NADH dehydrogenase subunits 4 and 5 (ten recurrent alterations), followed by ATPase6 (six recurrent alterations). The high number of recurrent alterations in the ND4, ND5 and ATPase6 genes indicates the high level of variation present in these three genes and the transmission and maintenance of observed variation in mtDNA can be attributed to random genetic drift or the effect of natural selection (refer to Section 2.5). In a recent investigation, the pattern of polymorphisms in mitochondrial genes was attributed to relaxed purifying selection and the action of genetic drift (Ingman and Gyllensten, 2007b). In contrast, a number of studies have presented evidence for the effect of positive selection on the ND4, ND5 and ATPase6 genes (Mishmar *et al.*, 2003; Ruiz-Pesini *et al.*, 2004; Kivisild *et al.*, 2006; Mishmar *et al.*, 2006), while in other studies mtDNA variation has been attributed to the effect of purifying selection (Elson *et al.*, 2004; Stewart *et al.*, 2008). Although the studies disagree on the type of natural selection (positive vs. negative) that plays a role in the generation of polymorphisms in mtDNA genes, the acknowledgment of the effect of natural selection remains the common factor. In the current investigation, based on the presence of an excess of non-synonymous alterations in the terminal branches of the global L0-specific haplogroup network, it is expected that purifying selection will play a role in shaping mitochondrial DNA variation (Elson *et al.*, 2004) and the effect of selection on mitochondrial variation in macrohaplogroup L is addressed in Section 5.11.5.

The recurrent alterations observed in the current investigation were also classified according to the type of alteration. The results are presented in Table 5.41.

Table 5.41 Number of recurrent alterations per alteration type in the global L0-specific haplogroup network based on the mtDNA coding region

Alteration type	Number of recurrent alterations
Synonymous substitution	44 ^a
Non-synonymous substitution	11 ^a
Deletion	2

^a = a single nucleotide alteration at position 8,566 affects both the ATPase6 (non-synonymous substitution) and ATPase8 (synonymous substitution) genes.

The ratio of synonymous to non-synonymous substitutions in the global L0-specific haplogroup network based on the mtDNA coding region was ca. 2:1. The ratio of recurrent synonymous to recurrent non-synonymous substitutions was ca. 4:1. This increase in synonymous substitutions compared to non-synonymous substitutions for recurrent alterations can be attributed to the existence of mutation hot spots (Galtier *et al.*, 2006). Hypervariable sites exist in the coding region of mtDNA and recurrent alterations are highly likely to occur at these sites. In addition, the majority of alterations at hypervariable sites are transitions (Galtier *et al.*, 2006). Transitions usually give rise to synonymous substitutions (Maynard Smith, 1994) and are therefore not likely to have a functional effect. Thus, recurrent alterations are more likely to involve synonymous substitutions and this will lead to the increase in the ratio of recurrent synonymous substitutions to recurrent non-synonymous substitutions as observed in the current investigation.

The presence of recurrent alterations in different parts of the global L0-specific haplogroup network can be due to a number of different evolutionary processes. It has been proposed that the presence of recurrent alterations is due to independent evolution of the same feature resulting from adaptive selection acting on specific functional variants (Ruiz-Pesini *et al.*, 2004; Ruiz-Pesini and Wallace, 2006; Ruiz-Pesini *et al.*, 2007). In another study recurrent alterations, assumed to be hot spots of mutational activity, were identified at nucleotide positions 709, 3,010, 10,398 and 13,928 within a global mtDNA network, and it was proposed that these sites were subjected to site-specific positive selection (Kivisild *et al.*, 2006). It therefore appears that recurrent alteration events are likely to occur at nucleotide positions that experience evolutionary pressure, and that the alterations are most likely retained at these positions due to positive selection.

In addition to recurrent alterations, i.e. the same alteration at the same position in different parts of the global L0-specific haplogroup network, different alterations were also observed at the same nucleotide position. At nucleotide position 2,245 located in the 16S rRNA gene, three different character states were observed in different parts of the global L0-specific haplogroup network, namely 2245G, 2245C and 2245A. The 2245G alteration was observed in the L0aC, L0c1 and L0c2 branches, the 2245G alteration changed to 2245C and characterised a section of the L0c1B branch, and the 2245A alteration, which represented a reversion to the character state observed in the rCRS, defined the L0c2D1 branch. The 2245C alteration that characterised a section of the L0c1B branch represents a split between this part of the L0c1B branch which consists of L0 sequences obtained from African Americans, and the remaining L0 sequences in the L0c1B branch that were obtained from an individual from Burkina Faso (DQ112732) and an African American (586). Due to the fact that this alteration is observed mainly in African Americans (DQ304901, DQ304902, DQ304903 and DQ304904), it is tempting to speculate that the alteration occurred in the ancestors of the African Americans after their displacement to countries outside of Africa. However, the presence of this alteration in two L0 sequences obtained from Africa, AF346985 from a Hausa individual from Nigeria/Niger and AY195780, indicates that the alteration was already present in West Africa prior to the removal of African individuals to other countries due to the slave trade (refer to Appendix G).

Two character states were observed in different sections of the global L0-specific haplogroup network at nucleotide position 4,562 located in the ND2 gene, namely 4562G which defined L0aC and 4562C which characterised L0bB1a. The L0aC branch consists of L0 sequences (AY963585, DQ112934 and EF184597) obtained from individuals of East Africa, while L0bB1a contains a single L0 sequence (KS_CGR_052) from southern Africa (refer to Appendix G). It can be speculated that the 4562G alteration originated in East Africa, while 4562C has a southern African origin. However, many more sequences are needed to confirm the existence of a relationship between these alterations and their geographic origin.

Reversions of alterations to the status observed in the rCRS were also identified in certain L0 sequences contained within the global L0-specific haplogroup network, and the reversions in these sequences were indicated in terminal branches of selected lineages. The reversions generally represented exceptions to the rule, i.e. all L0 sequences present in a specific branch or lineage in the global L0-specific haplogroup network contained the

alteration except the specific L0 sequence in which the alteration reverted to the status observed in the rCRS. The positions at which reversions were observed are presented in Table 5.42.

Table 5.42 Nucleotide position and location of reversions in the global L0-specific haplogroup network based on the mtDNA coding region

Nucleotide position	Number of events	Nucleotide change	Region/gene	Amino acid substitution
719	1	A → G	12S rRNA	---
2,245	1	G → A	16S rRNA	---
4,312	1	T → C	tRNA ^{Ile}	---
5,231	1	A → G	ND2	SYN
5,442	1	C → T	ND2	NON-SYN: Leu → Phe
6,266	1	G → A	COI	SYN
7,146	1	G → A	COI	NON-SYN: Ala → Thr
8,113	1	A → C	COII	SYN
8,152	1	A → G	COII	SYN
8,251	1	A → G	COII	SYN
10,143	1	A → G	ND3	NON-SYN: Ser → Gly
10,876	1	G → A	ND4	SYN
10,915	1	C → T	ND4	SYN
10,939	1	T → C	ND4	SYN
11,719	1	A → G	ND4	SYN
12,720	1	G → A	ND5	SYN
13,276	1	G → A	ND5	NON-SYN: Val → Met
14,020	1	C → T	ND5	SYN
14,308	2	C → T	ND6	SYN
15,941	1	C → T	tRNA ^{Thr}	---

Amino acid abbreviations: Ala = alanine; Gly = glycine; Ile = isoleucine; Leu = leucine; Met = methionine; Phe = phenylalanine; Ser = serine; Thr = threonine; Val = valine. Gene abbreviations: CO = cytochrome c oxidase subunit; ND = NADH dehydrogenase subunit; rRNA = ribosomal RNA; tRNA = transfer RNA; SYN = synonymous substitution; NON-SYN = non-synonymous substitution.

A total of 21 reversions were observed at 20 nucleotide positions. Reversions at most nucleotide positions occurred only once, with the exception of nucleotide position 14,308 where a reversion occurred twice. A total of 12 reversions involved A → G or G → A transitions, eight reversions entailed C → T or T → C transitions, and one reversion resulted in an A → C transversion. As previously observed for recurrent alterations, there was a prevalence of A → G or G → A transitions at positions where reversions occurred, confirming a role for these positions in the mtDNA coding region as hypervariable sites. The distribution of the reversions within the mtDNA coding region is presented in Table 5.43.

Table 5.43 Number of reversions per region/gene in the global L0-specific haplogroup network based on the mtDNA coding region

Region/gene	Number of reversions	Region/gene	Number of reversions
12S rRNA	1	ND4	4
16S rRNA	1	ND5	3
COI	2	ND6	2
COII	3	tRNA ^{Ile}	1
ND2	2	tRNA ^{Thr}	1
ND3	1	---	---

Region/genes in which the highest numbers of reversions are present, are indicated in red. Amino acid abbreviations: Ile = isoleucine; Thr = threonine. Gene abbreviations: CO = cytochrome c oxidase subunit; ND = NADH dehydrogenase subunit; rRNA = ribosomal RNA; tRNA = transfer RNA.

The regions/genes in the coding region that contained the largest number of reversions were NADH dehydrogenase subunit 4 (four reversions), followed by the NADH dehydrogenase subunit 5 and cytochrome c oxidase subunit 2 (three reversions). As already mentioned, ND4 and ND5 are subject to the effects of natural selection. Deviation from neutrality has been observed for alterations in the COII gene and this can be attributed to the effect of negative selection in the portion of the gene that encodes the cytosolic section of the polypeptide (Templeton, 1996). The effect of selection on these genes therefore may have allowed the occurrence and preservation of a high number of reversions. The reversions were also classified according to the type of alteration and the results are presented in Table 5.44.

Table 5.44 Number of reversions per alteration type in the global L0-specific haplogroup network based on the mtDNA coding region

Alteration type	Number of reversions
Synonymous substitution	12
Non-synonymous substitution	4

The ratio of synonymous to non-synonymous reversion substitutions in the global L0-specific haplogroup network was ca. 3:1. This increase in synonymous substitutions compared to non-synonymous substitutions for reversions can be explained in a similar manner to the increased ratio observed for recurrent alterations. Reversions are likely to occur at hypervariable sites which exist in the coding region of mtDNA, and preferentially involve synonymous substitutions leading to the increase in the ratio of synonymous to non-synonymous substitutions for reversions as observed in the current investigation.

Alterations and the subsequent reversion of these alterations to the state observed in the rCRS have previously been observed in the mtDNA coding region. Chen *et al.* (1995)

observed that two restriction sites, 7055 *Alu* I and 13065 *Dde* I, were altered and later reverted within individual mtDNA lineages, while Herrnstadt *et al.* (2002a) observed reversions at nucleotide positions 750, 1,438 and 15,326.

The relatively high frequency of recurrent mutations and reversions confirm that detailed analyses are necessary to accurately determine the relationships between mtDNA genome sequences (Chen *et al.*, 1995). Both the control and the coding region of the mtDNA genome are subject to recurrent alterations and reversions, with the control region exhibiting overall higher variability than the coding region. In order to compensate for the uncertainty introduced by these processes, the complete coding region of mtDNA genome sequences should be used for analyses.

5.11 STATISTICAL ANALYSIS

Certain basic descriptive sequence statistics were calculated as described in Section 4.14 for datasets 4a and 4b (!Kung Sequences) which consisted of the Khoi-San mitochondrial genome sequences generated in the current investigation, as well as the Khoi-San mitochondrial genome sequences determined previously by Mouton (2003). These statistics included measures of nucleotide composition, G + C content, as well as codon usage, and the results are presented in Section 5.11.1. In addition, the number of segregating sites, the nucleotide diversity and the average number of nucleotide differences were calculated for all the different sequence datasets (datasets 1b, 2b, 3a, 4b, 5, 6, 7 and 8a) as defined in Section 4.10.2 and the results are presented in Section 5.11.2. These datasets were also investigated to determine if the mitochondrial genome sequences evolved according to the standard neutral theory of evolution (Section 2.5.5) and the results are presented in Section 5.11.3. If a dataset exhibited deviation from the neutral theory of molecular evolution, the frequency distributions of pairwise mtDNA differences were investigated and a number of statistical parameters calculated to determine if the observed deviation could be explained by the occurrence of a population expansion event (refer to Section 5.11.4) or the effect of selection (refer to Section 5.11.5).

5.11.1 Basic sequence statistics

Nucleotide composition and G + C content were calculated for dataset 4a which consisted of the complete mitochondrial DNA genome sequences of 43 Khoi-San individuals (refer to

Section 4.10.4). The dataset contained the 30 Khoi-San mitochondrial genome sequences generated in the current investigation, as well as 13 Khoi-San mitochondrial DNA genome sequences obtained from a previous investigation (Mouton, 2003). The 43 sequences were all generated from samples obtained from !Kung individuals (refer to Appendix G). Codon usage was calculated for dataset 4b which contained the coding region only of the 43 sequences contained in dataset 4a (refer to Section 4.10.4). Basic sequence statistics (nucleotide composition and G + C content) were determined for dataset 4a only as this dataset contained complete mitochondrial DNA sequences which allowed the comparison of results for the control and coding region where applicable.

5.11.1.1 Nucleotide composition and G + C content

The relative frequencies of the four nucleotides in the 43 Khoi-San complete mtDNA genome sequences in dataset 4a were determined as described in Section 4.14.1. The average frequency of each nucleotide was calculated as a percentage of the total number of nucleotides and the results are summarised in Table 5.45. The average G + C content of the mitochondrial genome sequences within the dataset was also calculated, and is also indicated in the table below.

Table 5.45 Nucleotide composition and G + C content of the 43 Khoi-San mtDNA genome sequences in dataset 4a

Nucleotide composition (%)				G + C content				
A	C	G	T	G + Cn	G + C2	G + C3s	G + Cc	G + C
30.9	31.2	13.0	24.9	0.481	0.504	0.460	0.466	0.444

A = adenine; C = cytosine; G = guanine; T = thymine; G + Cn = G + C content at non-coding positions; G + C2 = G + C content at second codon positions; G + C3s = G + C content at third (synonymous) codon positions; G + Cc = G + C content at coding positions; G + C = G + C content in the selected region i.e. the complete mtDNA genome. Nucleotide composition was calculated using MEGA version 3.1 (Kumar *et al.*, 2004) and DnaSP version 4.10.9 (Rozas *et al.*, 2003) was used for the calculation of G + C content. A detailed description of dataset 4a is provided in Section 4.10.4.

The observed average values for each nucleotide reported in Table 5.45 correspond to the previously reported nucleotide composition for the L-strand (Anderson *et al.*, 1981). It can also be observed that the average A + T content of the mtDNA genome sequences is 55.8% (30.9% A + 24.9% T), compared to an average G + C content of 44.2% (13.0% G + 31.2% C). The greater A + T content compared to the G + C content can be attributed to a bias toward A/T ending codons in the mtDNA of many species (Ballard and Whitlock, 2004), which may be a consequence of factors that impair rates of endogenous DNA damage (Martin, 1995). A difference was observed in the average G + C content value when using different programs to determine this value. MEGA version 3.1 (Kumar *et al.*, 2004) was used for the calculation of the nucleotide composition and a G + C

content value of 44.2% was determined by adding the individual values for G and C. DnaSP version 4.10.9 (Rozas *et al.*, 2003) was used for the direct calculation of the G + C content and a value of 44.4% was calculated. The difference in the two average G + C content values (44.2% vs. 44.4%) can be ascribed to the methods employed by the different programs, since DnaSP excludes all sites with alignment gaps in single DNA sequences.

The G + C content for the different positions within codons, and codons located in different genome regions are also indicated in Table 5.45. In the absence of selection and mutational bias, it is expected that the G + C content of the different positions within codons and the different regions within the mtDNA genome, i.e. the non-coding regions (G + C_n) and the coding regions (G + C_c), should be similar as it is assumed that the mutations occur randomly (Shields *et al.*, 1988). The G + C content values are however different (refer to next paragraph). The G + C content at the second codon position (G + C₂) is the highest and this can be ascribed to the fact that alterations at this position are always non-synonymous substitutions which are subject to purifying selection due to functional constraint (Nei and Kumar, 2000). The G + C content at silent sites (third position in the codon) is slightly lower as this position is affected by both synonymous and non-synonymous substitutions and it can tolerate both types of alterations. In the case of non-synonymous substitutions purifying selection can play a role in maintaining the amino acid encoded by the codon (Nei and Kumar, 2000).

The G + C content of the non-coding region (G + C_n) is higher than the G + C content of the coding region (G + C_c). This can be attributed to the difference between the forward and backward mutation rate within the highly variable D-loop region (Sueoka, 1962) leading to the conversion of A/T pairs to C/G pairs. Due to the high mutation rate observed in the D-loop (Jobling *et al.*, 2004), it is subject to multiple forward and reverse mutations. These forward and reverse mutations influence the conversion of A/T to C/G pairs and vice versa, i.e. the interaction of the two mutation rates determines the conversion rate (Sueoka, 1962). The observed difference in the G + C content between the non-coding and the coding regions can also be ascribed to the influence of selection, i.e. the two regions have different G + C contents due to the fact that selection acts differently on the two regions as it governs and shapes the variation observed in the protein coding region. This means that the heterogeneity of the nucleotide composition observed in the two regions of the mitochondrial DNA genome may be interpreted as the result of a resistance

of nucleotide composition to the levelling effect of mutation pressure by selection in the coding region (Sueoka, 1962).

5.11.1.2 Codon usage

Codon usage was calculated as described in Section 4.14.2 for the 43 coding region only Khoi-San mitochondrial sequences contained in dataset 4b. The positions of coding and non-coding regions within the sequences were assigned according to the regions first described by Anderson *et al.* in 1981 and listed on MITOMAP (2007). The vertebrate (MEGA version 3.1) or mammalian (DnaSP version 4.10.9) mitochondrial DNA code table was selected for the analyses. Sites or codons containing alignment gaps in single mtDNA sequences were excluded from all analyses (DnaSP version 4.10.9) and all the values determined were average values calculated for the entire dataset of mtDNA sequences.

Due to the degeneracy of the genetic code, a number of different codons can specify a particular amino acid and it is expected that all the codons for a specific amino acid will be used with equal frequency (Page and Holmes, 1998). It has however been observed in most unicellular and multicellular organisms studied to date that all the codons encoding a particular amino acid are not used with equal frequency. This non-random pattern of codon usage is referred to as codon usage bias (Page and Holmes, 1998).

The extent of the non-random usage of synonymous codons within dataset 4b was determined by calculating four different statistical codon bias measures namely RSCU (Sharp *et al.*, 1986), N_C (Wright, 1990), CBI (Morton, 1993) and SCh_i^2 (Shields *et al.*, 1988) as described in Section 4.14.2. The estimates of codon usage bias obtained by using the N_C , CBI and SCh_i^2 measures were compared in a previous investigation by Comeron and Aguade (1998) in which computer simulation was used to determine the effect of the coding region length on the different estimates of codon bias, and to evaluate the variability of the codon bias values. It was observed that when coding regions of different lengths were compared, only the N_C measure produced comparable values for all lengths and conditions of codon bias. In contrast, the SCh_i^2 and CBI measures generated distinctly different average values when coding regions with different numbers of codons were compared for a specific degree of codon bias. However, when a specific analysis was performed to compare values of codon bias among coding regions of comparable size, the N_C and CBI measures were preferred, as these measures displayed the lowest variability in codon bias values. Generally, the N_C measure exhibited the lowest variability

in codon bias values for the different codon bias conditions and region lengths. However, when medium or short coding regions were investigated, the CBI measure presented the lowest dispersion for regions with an overall high degree of codon bias. It is thus clear that each of these measures has advantages and disadvantages dependent on the length of the coding region under investigation and the specific codon bias conditions present within the coding region. Prior to the determination of the codon bias measure within the coding region under investigation, insufficient information regarding codon bias conditions is available to guide in the selection of the appropriate codon bias measure, and it is therefore not clear which method is the most applicable to mitochondrial DNA sequences prior to the performance of the calculations.

The RSCU values for dataset 4b were calculated as described in Section 4.14.2.1 using MEGA version 3.1 (Kumar *et al.*, 2004) and are indicated in Table 5.46. A RSCU value of 1.00 indicates that there is no codon bias detectable for a specific codon, i.e. its frequency is that which is expected under the assumption of equal codon usage. RSCU values less than or more than one indicate that codons are used less often or more often than expected. Since the majority of RSCU values are below or above 1.00, it can be suggested that the coding region of the mitochondrial genome is subject to codon usage bias.

Table 5.46 Codon usage for the 43 coding region only Khoi-San mtDNA genome sequences in dataset 4b

Amino acid	Codon	Frequency	RSCU	Amino acid	Codon	Frequency	RSCU
Ala	GCA	0.021	1.28	Lys	AAA	0.024	1.82
	GCC	0.033	1.98		AAG	0.002	0.18
	GCG	0.001	0.08	Met	AUA	0.046	1.70
	GCU	0.011	0.66		AUG	0.008	0.30
Arg	CGA	0.007	1.78	Phe	UUC	0.037	1.33
	CGC	0.008	1.84		UUU	0.018	0.67
	CGG	0.000	0.00	Pro	CCA	0.015	0.94
	CGU	0.002	0.38		CCC	0.035	2.26
Asn	AAC	0.037	1.54		CCG	0.002	0.14
	AAU	0.011	0.46		CCU	0.010	0.65
Asp	GAC	0.014	1.64	Ser	AGC	0.011	0.89
	GAU	0.003	0.36		AGU	0.003	0.28
Cys	UGC	0.005	1.57		UCA	0.022	1.79
	UGU	0.001	0.43		UCC	0.028	2.27
Gln	CAA	0.024	1.78		UCG	0.002	0.15
	CAG	0.003	0.22		UCU	0.008	0.62

Table 5.46 Continued...

Amino acid	Codon	Frequency	RSCU	Amino acid	Codon	Frequency	RSCU
Glu	GAA	0.017	1.62	Thr	ACA	0.035	1.48
	GAG	0.004	0.38		ACC	0.043	1.82
Gly	GGA	0.017	1.38		ACG	0.003	0.11
	GGC	0.023	1.89		ACU	0.014	0.58
	GGG	0.005	0.38	Trp	UGA	0.025	1.84
	GGU	0.004	0.35		UGG	0.002	0.16
His	CAC	0.023	1.53	Tyr	UAC	0.026	1.45
	CAU	0.007	0.47		UAU	0.010	0.55
Ile	AUC	0.052	1.25	Val	GUA	0.017	1.79
	AUU	0.031	0.75		GUC	0.012	1.29
Leu	CUA	0.072	2.59		GUG	0.002	0.21
	CUC	0.047	1.68		GUU	0.007	0.71
	CUG	0.011	0.41	*	AGA	0.001	0.80
	CUU	0.017	0.62		AGG	0.001	0.80
	UUA	0.017	0.61		UAA	0.003	2.01
	UUG	0.003	0.10		UAG	0.001	0.40

Frequency = frequency of codon usage, i.e. the number of times the specific codon was used in the total number of codons present; RSCU = relative synonymous codon usage, i.e. the observed frequency of the specific codon divided by that expected under the assumption of equal usage of synonymous codons. A-ending codons are indicated in **red**, and C-ending codons are indicated in **blue**. Nucleotide abbreviations: A = adenine; C = cytosine; G = guanine; U = uracil. Amino acid abbreviations: Ala = alanine; Arg = arginine; Asp = aspartic acid; Asn = asparagine; Cys = cysteine; Gln = glutamine; Glu = glutamic acid; Gly = glycine; His = histidine; Ile = isoleucine; Leu = leucine; Lys = lysine; Met = methionine; Phe = phenylalanine; Pro = proline; Ser = serine; Thr = threonine; Trp = tryptophan; Tyr = tyrosine; Val = valine; * = termination. A detailed description of dataset 4b is provided in Section 4.10.4.

It is of interest that RSCU values greater than one were determined for the majority of codons that end with an A or C nucleotide (refer to Table 5.46). This observation is consistent with the effect of strand-specific mutation bias (Xia, 2005). This process involves the spontaneous deamination of the H-strand during mtDNA replication, which results in A → G and C → U mutations in the H-strand. As a consequence, the H-strand becomes rich in G and T, while the L-strand becomes rich in A and C. As a result, the majority of codons in the polypeptide encoding genes, which are located on the H-strand and are therefore co-linear to the L-strand (refer to Section 2.4.4, Table 2.5), end with A or C nucleotides. NNY codon families are dominated by C-ending codons, while NNR and NNN codon families are dominated by A-ending codons (refer to Table 5.46). The observation that NNN codon families are dominated by NNA codons, instead of by both NNA and NNC codons, might have adaptive significance (Xia, 1996). This conclusion is based on the fact that, in the mitochondrial environment, the cellular concentration of ATP exceeds that of the other three ribonucleotides (Xia, 1996). This observation led to the transcription hypothesis of codon usage which states that due to the abundance of ATP, the efficiency of transcription can be increased through the use of A in the third codon position of polypeptide encoding genes (Xia, 1996).

The RSCU values calculated per polypeptide encoding gene supported the observation of codon usage bias for the coding region of the mitochondrial DNA genome (refer to Appendix M). The RSCU values calculated for the majority of the codons of individual polypeptide genes were in fact considerably lower than 1.00 or considerably higher than 1.00, i.e. some codons were used considerably less or considerably more than other codons encoding the same amino acid. A general pattern was observed for the individual polypeptide genes whereby one, two or three of the available codons were preferred where the amino acid was encoded by two, four or six codons. The preferred codon could however be different for the same encoded amino acid in the different polypeptide genes (refer to Appendix M). As an example the RSCU values determined for the ND5 gene are indicated in Table 5.47.

Table 5.47 Codon usage for the ND5 gene of the 43 coding region only Khoi-San mtDNA genome sequences in dataset 4b

Amino acid	Codon	Frequency	RSCU	Amino acid	Codon	Frequency	RSCU
Ala	GCA	0.032	1.79	Lys	AAA	0.033	1.90
	GCC	0.031	1.75		AAG	0.002	0.10
	GCG	0.000	0.00	Met	AUA	0.036	1.71
	GCU	0.008	0.46		AUG	0.006	0.29
Arg	CGA	0.005	1.50	Phe	UUC	0.054	1.74
	CGC	0.007	2.00		UUU	0.008	0.26
	CGG	0.000	0.00	Pro	CCA	0.015	1.14
	CGU	0.002	0.50		CCC	0.025	1.87
Asn	AAC	0.045	1.64		CCG	0.000	0.00
	AAU	0.010	0.36		CCU	0.013	0.99
Asp	GAC	0.013	1.45	Ser	AGC	0.016	1.20
	GAU	0.005	0.55		AGU	0.005	0.36
Cys	UGC	0.008	1.67		UCA	0.022	1.59
	UGU	0.002	0.33		UCC	0.034	2.48
Gln	CAA	0.028	1.70		UCG	0.002	0.12
	CAG	0.005	0.30		UCU	0.003	0.25
Glu	GAA	0.013	1.78	Thr	ACA	0.037	1.37
	GAG	0.002	0.22		ACC	0.049	1.78
Gly	GGA	0.017	1.54		ACG	0.001	0.06
	GGC	0.019	1.73		ACU	0.022	0.79
	GGG	0.002	0.15	Trp	UGA	0.018	1.84
	GGU	0.006	0.58		UGG	0.002	0.16
His	CAC	0.022	1.86	Tyr	UAC	0.015	1.15
	CAU	0.002	0.14		UAU	0.011	0.85

Table 5.47 Continued...

Amino acid	Codon	Frequency	RSCU	Amino acid	Codon	Frequency	RSCU
Ile	AUC	0.056	1.28	Val	GUA	0.010	1.39
	AUU	0.031	0.72		GUC	0.008	1.18
Leu	CUA	0.075	2.61		GUG	0.000	0.00
	CUC	0.052	1.79		GUU	0.010	1.44
	CUG	0.013	0.45	*	AGA	0.000	0.00
	CUU	0.018	0.64		AGG	0.000	0.00
	UUA	0.011	0.39		UAA	0.002	4.00
	UUG	0.003	0.12		UAG	0.000	0.00

Frequency = frequency of codon usage, i.e. the number of times the specific codon was used in the total number of codons present; RSCU = relative synonymous codon usage, i.e. the observed frequency of the specific codon divided by that expected under the assumption of equal usage of synonymous codons. A-ending codons are indicated in **red**, and C-ending codons are indicated in **blue**. Very high (≥ 1.7) or very low (≤ 0.3) RSCU values are indicated in **green**. Nucleotide abbreviations: A = adenine; C = cytosine; G = guanine; U = uracil. Amino acid abbreviations: Ala = alanine; Arg = arginine; Asp = aspartic acid; Asn = asparagine; Cys = cysteine; Gln = glutamine; Glu = glutamic acid; Gly = glycine; His = histidine; Ile = isoleucine; Leu = leucine; Lys = lysine; Met = methionine; Phe = phenylalanine; Pro = proline; Ser = serine; Thr = threonine; Trp = tryptophan; Tyr = tyrosine; Val = valine; * = termination. A detailed description of dataset 4b is provided in Section 4.10.4.

Inspection of the RSCU values for the ND5 gene revealed the presence of codon usage bias. Codons with very low (≤ 0.3) or very high (≥ 1.7) RSCU values are indicated in **green** text in Table 5.47. These cut-off values were determined as described in Section 4.14.2.1 based on the range of values obtained for the coding region of the mitochondrial genome (refer to Table 5.46). These codons exhibit extreme codon usage bias, i.e. the codons with low values are used rarely and the codons with high values are used very frequently. Extreme codon usage bias was observed for codons encoding alanine, arginine, glutamine, glutamic acid, glycine, histidine, leucine, lysine, methionine, phenylalanine, proline, serine, threonine and tryptophan. In the case of amino acids encoded by two codons (glutamine, glutamic acid, histidine, lysine, methionine, phenylalanine and tryptophan), one codon was preferred above the other possible codon. For amino acids encoded by four codons (alanine, arginine, glycine, proline and threonine), it was observed that two codons were usually preferred. Amino acids encoded by six codons (leucine and serine) exhibited a similar pattern with three or four of the possible six codons used more frequently. The codon usage bias observed in the polypeptide genes can be attributed to the influence of selection (refer to Appendix M).

The values calculated for the N_C , CBI and SCh^2 codon usage bias measures as described in Section 4.14.2 using DnaSP version 4.10.9 (Rozas *et al.*, 2003) are indicated in Table 5.48. The SCh^2 value was calculated using the G + C percentage of 44.4% (refer to Table 5.45) which was determined using DnaSP version 4.10.9 (Rozas *et al.*, 2003).

Table 5.48 Global codon usage bias measures for the 43 coding region only Khoi-San mtDNA genome sequences in dataset 4b

Codon bias measure	Calculated value	Implication of value
N_C	41.928	Codon bias (low)
CBI	0.488	Codon bias (low)
SCh^2	0.474	Codon bias (low)

N_C = effective number of codons; CBI = codon bias index; SCh^2 = scaled chi square. A detailed description of dataset 4b is provided in Section 4.10.4.

A N_C value of 41.928, a CBI value of 0.488 and a SCh^2 value of 0.474 were calculated for dataset 4b which indicates low codon bias was present within the mitochondrial DNA genome (refer to Section 4.14.2). The different codon usage bias measures together therefore indicate the presence of a generally low level of codon usage bias within the Khoi-San mitochondrial DNA genomes.

The origin and maintenance of codon usage bias has been attributed to either mutation bias or natural selection (Nei and Kumar, 2000; Yang and Nielsen, 2008), although other factors such as random genetic drift, tRNA availability, effective population size, evolutionary history, biased gene conversion, mRNA secondary structure and translational initiation may also play a role (Qin *et al.*, 2004). The presence of a low level of codon usage bias is therefore an indication that one or both of the above processes is at work in the Khoi-San mitochondrial DNA genomes. While mutation bias is certainly a factor in the human mitochondrial genome as it has previously been determined that a relative preference for G-to-A and T-to-C mutations are observed in the third codon position, which extends over the non-silent positions and is characteristic of the D-loop region (Aquadro and Greenberg, 1983; Kivisild *et al.*, 2006), the significance of selection in driving the evolution of codon usage is less certain. Evidence for the role of selection in codon usage bias is however provided by observations regarding expression levels of genes, the level of conservation of amino acids and protein folding (Yang and Nielsen, 2008).

The level of observed codon usage bias is linked to the expression level of genes carried in the genome (Page and Holmes, 1998). In unicellular organisms, such as the organism which gave rise to mitochondria (refer to Section 2.1), strong codon usage bias is observed in highly expressed genes, as codon usage correlates with the relative abundance of isoaccepting tRNAs in a cell (Ikemura, 1981; Ikemura, 1985). Highly expressed genes require translational efficiency, i.e. rapid production of a specific protein coupled with a low error rate (Page and Holmes, 1998) and this is facilitated by the use of codons corresponding to tRNAs that are abundant within the cell. Codons which do not

correspond to the abundant tRNAs are subsequently eliminated by purifying selection due to their inefficiency in the protein synthesis process. In genes with low levels of expression, codon usage bias is generally less because there is more flexibility regarding the use of rare tRNAs (Nei and Kumar, 2000). A recent investigation has however suggested that differences in codon usage in human mtDNA might be primarily a function of strand asymmetry rather than of differences in the availability of tRNAs (Kivisild *et al.*, 2006).

In vertebrate mitochondria, it has been observed that a smaller number of tRNAs is used in the translation of all codons than expected, due to the fact that a single tRNA molecule recognises a codon family consisting of synonymous codons (Barrell *et al.*, 1980). The flexibility of a single tRNA molecule in recognising two or four synonymous codons is essential for translation in the mitochondrion (Agris, 2004). This mitochondrial process, involving the translation of a synonymous codon family by a single tRNA molecule due to the recognition of more than one codon, has been termed the selection hypothesis of anticodon versatility (Agris, 2004). This hypothesis allows the prediction of the presence of an U or G nucleotide at the wobble site (first codon position) of the tRNA anticodon which implies the presence of an A or C nucleotide at the third codon position within the codons most frequently used. A recent investigation of vertebrate mitochondrial DNA has revealed codon usage bias and tRNA anticodon data consistent with the selection hypothesis of anticodon versatility (Xia, 2005). In addition, it was demonstrated that codon usage bias is largely maintained by strand-specific mutation pressure and directs the evolution of tRNA anticodons, i.e. in vertebrate mitochondria the tRNA anticodons evolve in response to the codon usage bias which is retained due to mutation pressure (Xia, 2005). In the current investigation, this can be interpreted to imply that only 22 tRNAs are encoded by the mitochondrial genome since the anticodons present in these tRNA molecules correspond to the codons exhibiting high codon usage bias.

Codon usage bias is also linked to the level of conservation of the particular amino acid, i.e. codon usage is more biased for conserved amino acids than for non-conserved amino acids (Akashi, 1994). This may be due to selection for translational accuracy because highly conserved amino acids are expected to be functionally important and less tolerant to the incorporation of erroneous amino acids. In addition, codon usage bias may also play a role in translation kinetics where the use of preferred codons may affect the rate at which a protein is translated. This is important in ensuring the correct folding of proteins through the separation of temporal folding events during protein synthesis (Yang and Nielsen, 2008).

The presence of codon usage bias in the Khoi-San mitochondrial DNA genomes can therefore also be attributed to selective constraints due to their role in important metabolic processes such as oxidative phosphorylation (refer to Section 2.3) which requires high levels of functionally active genes. The fact that the codon usage bias is not as strong as would be expected is likely due to the fact that although elevated expression of specific proteins is required for the important role of the mitochondrion in energy metabolism, the copy number of mitochondria in each individual cell (refer to Section 2.5.2) compensates for this effect, as it provides sufficient copies of the relevant proteins to perform the important functions. In addition, the level of codon usage bias observed for the Khoi-San mitochondrial DNA genome sequences is influenced by the point in time when the analyses are performed, i.e. the codon usage determined in the current investigation reflects the codon usage bias present in the genome at the current point in time. That is, stronger codon usage bias may have been present in the past, but the effects of subsequent mutation events, genetic drift and selection over evolutionary time may have altered the observed values.

5.11.2 Genetic diversity measures

The number of segregating sites (S), the average number of nucleotide differences (k) and the nucleotide diversity (π) were calculated for each dataset consisting of coding region only sequences as defined in Section 4.14.3 by completely excluding all sites with alignment gaps (or missing data). This was done to ensure that only sequence data from homologous positions within the sequences were used for analyses (Salemi and Vandamme, 2003). The results are summarised in Table 5.49.

Table 5.49 Summary of statistical parameters of human mtDNA genomes in different datasets

Dataset	Number of sequences (n)	Number of segregating sites (S)	Average number of nucleotide differences (k)	Nucleotide diversity $\times 10^{-3}$ (π)
1b	426	1,345	52.620	3.42 ± 0.06
2b	239	1,102	58.238	3.78 ± 0.05
3a	58	279	46.322	3.00 ± 0.20
4b	43	224	45.533	2.95 ± 0.25
5	128	464	22.294	1.44 ± 0.06
6	113	336	21.824	1.41 ± 0.09

Table 5.49 Continued...

Dataset	Number of sequences (n)	Number of segregating sites (S)	Average number of nucleotide differences (k)	Nucleotide diversity $\times 10^{-3}$ (π)
7	63	252	34.497	2.24 ± 0.13
8a	103	416	41.078	2.66 ± 0.08

n = number of sequences; S = number of segregating sites; k = average number of nucleotide differences; π = nucleotide diversity. Datasets: 1b = All L Sequences; 2b = All Africa L Sequences; 3a = All Khoi-San L Sequences; 4b = !Kung Sequences; 5 = L3 Sequences; 6 = L2 Sequences; 7 = L1 Sequences; 8a = L0 Sequences. Refer to Section 4.10.4 for detailed descriptions of the different datasets.

The nucleotide diversity observed within the different datasets in the current investigation was consistent with that observed in a previous study (Ingman *et al.*, 2000). Ingman *et al.* (2000) reported a value of 3.7×10^{-3} for a dataset consisting of 21 African sequences and ascribed the high diversity to a large effective population size or a long genetic history. Taking into account that the datasets studied in the current investigation consist of representatives from different African ethnic groups, the high nucleotide diversity values can also be explained by longer periods of separation between different African sub-populations (Cann *et al.*, 1987). The increased diversity observed in the datasets under investigation can also be explained by ancient population expansion events in Africa (Excoffier and Schneider, 1999). The different datasets studied in the current investigation can be sorted according to decreasing nucleotide diversity which is indicated in Table 5.50.

Table 5.50 Datasets sorted according to decreasing nucleotide diversity

Dataset	Short description	Number of sequences (n)	Nucleotide diversity $\times 10^{-3}$ (π)
2b	All Africa L Sequences	239	3.78 ± 0.05
1b	All L Sequences	426	3.42 ± 0.06
3a	All Khoi-San L Sequences	58	3.00 ± 0.20
4b	!Kung Sequences	43	2.95 ± 0.25
8a	L0 Sequences	103	2.66 ± 0.08
7	L1 Sequences	63	2.24 ± 0.13
5	L3 Sequences	128	1.44 ± 0.06
6	L2 Sequences	113	1.41 ± 0.09

n = number of sequences; π = nucleotide diversity. Refer to Section 4.10.4 for detailed descriptions of the different datasets.

The amount of mtDNA sequence diversity observed for datasets 1b, 2b, 3a and 4b was ca. double that observed for datasets 5 and 6. The nucleotide diversity values of datasets 2b, 1b, 3a and 4b are in alignment with the composition of these datasets. Dataset 2b consists of sequences from indigenous Africans, i.e. a large number of different haplogroups (L0,

L1, L2 and L3) were present within this single dataset which gave rise to large differences between the sequences and therefore high nucleotide diversity. Dataset 1b consists of sequences obtained from indigenous black Africans as well as blacks of African ancestry e.g. African Americans. The sequences obtained from blacks of African ancestry were mostly classified as L2 or L3 and the addition of a large number of similar L2 and L3 sequences to dataset 1b decreased the nucleotide diversity present in this dataset compared to dataset 2b. The same argument holds for datasets 3a and 4b. Dataset 4b was a subset of dataset 3a and consisted of Khoi-San sequences from the !Kung ethnic group only. These sequences were more similar to one another than the sequences in dataset 3a which were obtained from individuals from different ethnic groups that were all classified as Khoi-San speakers.

The order of the nucleotide diversity values for the datasets containing haplogroups L0, L1, L2 and L3 can be attributed to the way in which the different haplogroups are related to each other (refer to Section 5.9) and to the estimated ages of the haplogroups (refer to Sections 5.11.6 and 3.3.1.4). In the global N-J and MP phylogenetic trees, it was observed that the L0 and L1 haplogroups cluster close to the outgroup, and well-defined subgroups are observed within each of these groups, i.e. there are sufficient nucleotide differences between sequences to form separate lineages which are supported by the observed bootstrap values. The L2 and L3 haplogroups clustered further from the outgroup, and although defined subgroups are present, the haplogroups or subgroups within, appear to be incompletely resolved. That is, sequences within each haplogroup appear to be closely related and subgroups seem to have recently diverged from one another, especially in the case of L3. This is supported by lower bootstrap values for some groups e.g. L2a, L3a1 and other groupings within L3. A correlation also exists between the age of the haplogroup (TMRCA) and the nucleotide diversity within the haplogroup, i.e. the older a haplogroup, the more diverse. L0 is the oldest haplogroup and therefore contains the most nucleotide diversity, L1 is the next oldest, followed by L2 and lastly L3 (Chen *et al.*, 1995; Chen *et al.*, 2000; Salas *et al.*, 2002; Mishmar *et al.*, 2003; Kivisild *et al.*, 2006; Ruiz-Pesini *et al.*, 2007; refer to Appendix C). In the current investigation, the dataset consisting of L0 sequences displayed the highest nucleotide diversity, followed by L1, L3 and L2. The increased nucleotide diversity observed for L3 compared to L2 can be attributed to the overrepresentation of the L2a sub-haplogroup in the L2 dataset. The inclusion of a great number of closely related sequences from the same sub-haplogroup led to an overall loss of diversity within the dataset.

The high nucleotide diversity values observed for the different datasets indicate the high level of nucleotide diversity present in African mitochondrial haplogroups. The nucleotide diversity observed for African haplogroups is higher than that observed for other continent-specific mitochondrial haplogroups (Chen *et al.*, 1995; Wallace, 1995). The commonly used rCRS is based on a European mitochondrial genome sequence (Anderson *et al.*, 1981) and therefore does not represent all the genetic variation evident in the African mitochondrial haplogroups. The L0 haplogroup contains the highest nucleotide diversity and is therefore the most diverse of the major African haplogroups (L0, L1, L2 and L3). A consensus based on sequences obtained from Khoi-San individuals who belong to the L0 haplogroup will therefore contain a high level of diversity which should assist in distinguishing normal variation from potentially pathogenic alterations in both African sequences and sequences from other continent-specific haplogroups.

5.11.3 Deviation from the neutral theory of molecular evolution

Under the neutral theory of molecular evolution (Section 2.5.5) it is assumed that the observed polymorphisms within the mtDNA sequences are the consequence of forces such as migration and random genetic drift and that the polymorphisms play no role in human adaptation. This implies that the majority of mutations are neutral or nearly neutral (Kimura, 1968). In order to evaluate if this theory can be applied to the polymorphisms observed within the different datasets, it was therefore necessary to test the neutrality of the observed alterations. Different neutrality tests were employed, namely Tajima's *D* test and Fu and Li's *D** and *F** tests, and were performed as described in Section 4.14.4. The alternative tests make use of different types of information contained within the datasets investigated, and vary in their ability to detect the influence of different types of selection (Jobling *et al.*, 2004). Tajima's *D* test uses DNA polymorphism information, i.e. the actual genetic variation within the dataset, and compares the number of segregating sites per position with the nucleotide diversity (Tajima, 1989). Fu and Li's *D** and *F** tests consider the distribution of the total number of mutations in a genealogy of a random sample of sequences (Fu and Li, 1993). The three different tests were therefore performed due to the fact that no single neutrality test predominates in terms of performance (Jobling *et al.*, 2004). The test results and their corresponding *P*-values are indicated in Table 5.51.

Table 5.51 Tests evaluating the departure from neutrality in different datasets

Dataset	Tajima's <i>D</i> test		Fu and Li <i>D</i> * test		Fu and Li <i>F</i> * test	
	Value	<i>P</i> -value	Value	<i>P</i> -value	Value	<i>P</i> -value
1b	-2.275	<i>P</i> < 0.01	-6.977	<i>P</i> < 0.02	-5.053	<i>P</i> < 0.02
2b	-2.165	<i>P</i> < 0.01	-5.785	<i>P</i> < 0.02	-4.630	<i>P</i> < 0.02
3a	-0.825	<i>P</i> > 0.10	-0.850	<i>P</i> > 0.10	-1.007	<i>P</i> > 0.10
4b	-0.444	<i>P</i> > 0.10	-0.080	<i>P</i> > 0.10	-0.255	<i>P</i> > 0.10
5	-2.453	<i>P</i> < 0.01	-5.595	<i>P</i> < 0.02	-4.962	<i>P</i> < 0.02
6	-2.192	<i>P</i> < 0.01	-5.439	<i>P</i> < 0.02	-4.777	<i>P</i> < 0.02
7	-1.252	<i>P</i> > 0.10	-2.375	0.10 > <i>P</i> > 0.05	-2.316	0.1 > <i>P</i> > 0.05
8a	-1.641	0.1 > <i>P</i> > 0.05	-3.590	<i>P</i> < 0.02	-3.278	<i>P</i> < 0.02

P-value = probability that the neutrality test results could have been obtained randomly, referred to as statistical significance; *P* < 0.05 is statistically significant and is indicated in red text. *P* > 0.05 is not statistically significant. Datasets: 1b = All L Sequences; 2b = All Africa L Sequences; 3a = All Khoi-San L Sequences; 4b = !Kung Sequences; 5 = L3 Sequences; 6 = L2 Sequences; 7 = L1 Sequences; 8a = L0 Sequences. Refer to Section 4.10.4 for detailed descriptions of the different datasets.

Significant departures from neutrality, as measured by Tajima's *D* statistic and supported by the calculated *P*-values, were observed for datasets 1b, 2b, 5 and 6 which consist of all L sequences, Africa L sequences, L3 sequences and L2 sequences, respectively. This was not the case for datasets 3a, 4b, 7 and 8a which consist of Khoi-San sequences, !Kung sequences, L1 sequences and L0 sequences. Fu and Li's *D** and *F** statistics and the calculated *P*-values similarly revealed significant departures from neutrality in the same datasets as for Tajima's *D* statistic, as well as dataset 8a consisting of L0 sequences. The discrepancy observed in the results obtained for dataset 8a using Tajima's *D* test compared to Fu and Li's *D** and *F** tests can be ascribed to the fact that Tajima's *D* test is very conservative. The confidence limits of *D* are obtained assuming that *D* follows the beta distribution (Tajima, 1989), and when the beta distribution is used, rejection of the null hypothesis, i.e. that all mutations are neutral, becomes almost impossible (Simonsen *et al.*, 1995).

Other studies of complete mtDNA genome diversity in African individuals have produced contradictory results, with some studies not reporting significant deviations from neutrality (Mishmar *et al.*, 2003; Ruiz-Pesini *et al.*, 2004), while others have observed significant deviations (Torroni *et al.*, 2001; Kivisild *et al.*, 2006). The observed deviation from neutrality in the different datasets studied in the current investigation can be ascribed to a number of different factors. It could be due to the increased numbers of sequences in the current datasets which provide a more precise estimate of the genetic diversity within a population. Other factors that can cause a deviation from neutrality are the effect of selection, population expansion events, population subdivision, population bottlenecks, a selective sweep (reduction in genetic diversity due to recent and strong selection

pressure), genetic hitchhiking (a neutral alteration is fixed due to its association with another alteration), and an excess of variants due to artefacts of sampling from structured populations (Simonsen *et al.*, 1995; Ptak and Przeworski 2002; Ray *et al.* 2003; Hammer *et al.*, 2004).

The most likely explanations for deviation from neutrality in the datasets studied in the current investigation are the effect of selection or population expansion events. The significant negative values calculated for datasets 1b, 2b, 5 and 6 using Tajima's *D* statistic indicate the influence of positive selection or population growth (Jobling *et al.*, 2004) within the populations or haplogroups represented by these datasets. The significant negative values determined for datasets 1b, 2b, 5, 6 and 8a using Fu and Li's *D** and *F** statistics similarly indicate the effect of positive selection or an expanding population. It has been observed that when only Fu and Li's tests are significant, as in the case of dataset 8a, it is more likely that selection, and not population expansion, is responsible for the observed deviation from neutrality (Fu, 1997).

The possible presence of selection in the different datasets, notably datasets 1b and 2b, will influence the choice of method for phylogenetic tree construction. The neutral theory of molecular evolution supports the molecular clock hypothesis which is based on the assumption that the rate of nucleotide substitution within a gene remains constant across evolutionary time (Hartl, 1988). The effect of selection disrupts this constant rate of nucleotide substitution, and methods that do not assume the presence of a molecular clock should therefore be employed. The N-J and MP tree construction methods do not require the assumption of a molecular clock (refer to Sections 5.9.2 and 5.9.3) and were therefore used for phylogenetic analyses.

The genetic variation observed within a population is produced by mutation, but is organised, maintained, eliminated or dispersed among sub-populations according to a complex balance between selection, migration and random genetic drift (Hartl, 1988). If selection is present within the population, it changes the frequency of a specific alteration. Selection will therefore increase the frequency of a specific advantageous alteration in a population, with the reverse being true for deleterious alterations. Population growth can similarly affect the frequency of alterations when specific alterations are preferentially eliminated or propagated, when individuals within the population carrying specific alterations make varied contributions in terms of numbers of offspring involved in population growth.

To determine which of the two possibilities (selection or population expansion) are likely to be responsible for the deviation from neutrality observed in the different datasets, the occurrence of population expansion events was tested for by determining the distribution of the number of pairwise differences between mtDNA sequences, as well as the calculation of three different statistical measures, for each dataset. The results of the frequency distributions of pairwise differences and the statistical tests are presented in Section 5.11.4. The effect of selection in the different datasets was investigated and the results are discussed in Section 5.11.5.

5.11.4 Population size changes

Mathematical models have been proposed to evaluate human population history. In the current investigation, a relatively simplistic model (Slatkin and Hudson, 1991; Rogers and Harpending, 1992; Rogers, 1995) is employed to investigate population size changes, specifically population expansion events. In this context population expansion is used to describe a demographic expansion and not a range or spatial expansion.

This population expansion model is based on the distribution of nucleotide site differences between pairs of sequences which can have a characteristic signature that is indicative of population growth or decline events (Rogers and Harpending, 1992). In the current investigation, this distribution is referred to as the frequency distribution of pairwise differences between sequences, or alternatively, as the mismatch distribution. The distribution generated by continued exponential growth is impossible to differentiate from that of a sudden spurt of population growth (Rogers and Harpending, 1992). In addition, population bottlenecks also produce distributions very similar to that of a sudden expansion, but with elevated upper-tail probabilities (Rogers and Harpending, 1992). In contrast, Rogers and Harpending (1992) stated that a decrease in population size initially generates an L-shaped distribution with high probability of identity, but this distribution converges rapidly to a new equilibrium. The results generated for each of the datasets (refer to Section 4.10.4) are presented in the following section and the conclusions drawn for each dataset are based on the specific dataset and population growth model as used in the current investigation.

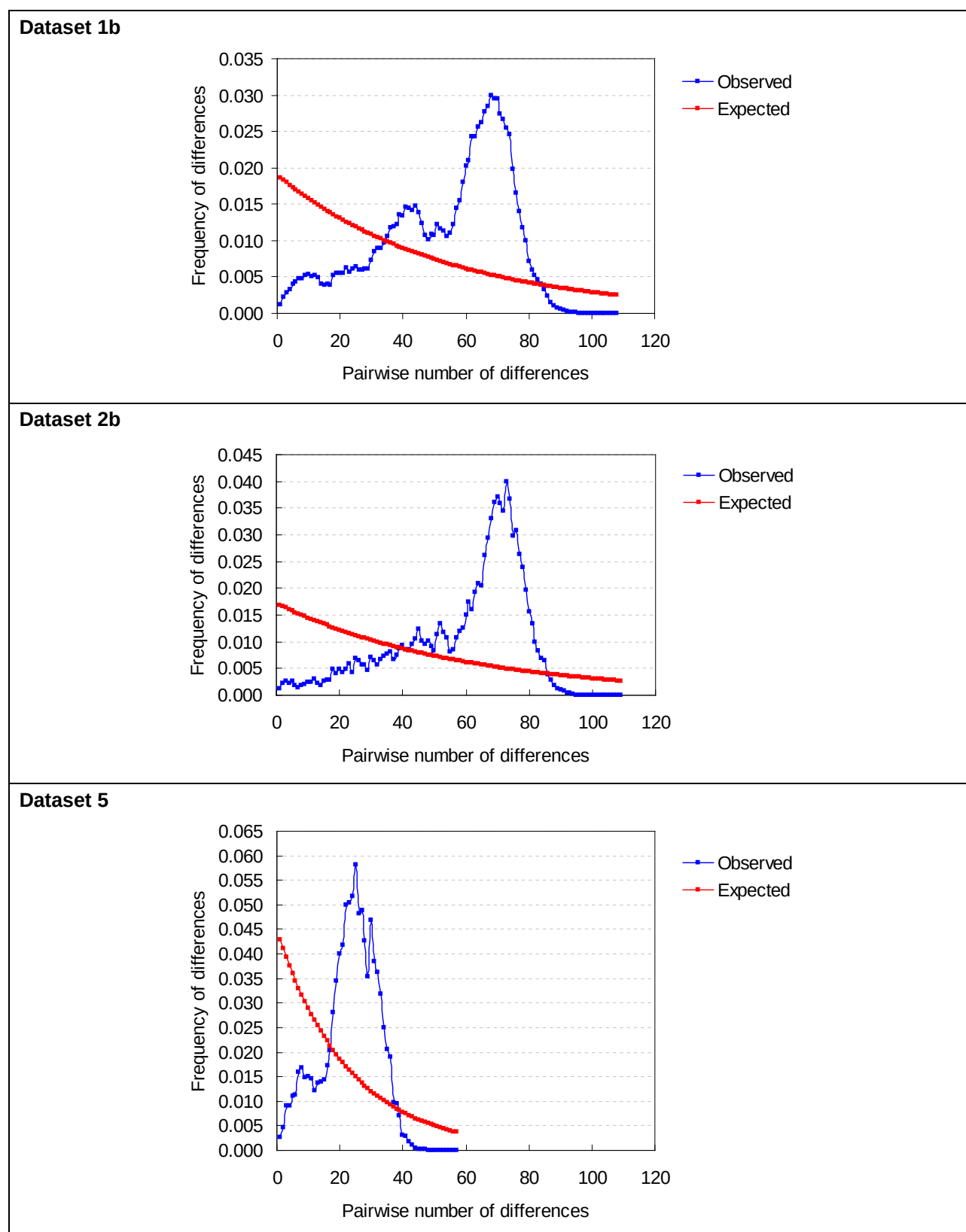
5.11.4.1 Frequency distributions of pairwise differences between mitochondrial DNA sequences

The frequency distributions of pairwise sequence differences were determined as described in Section 4.14.5.1 for each of the different datasets to determine if a population expansion event had occurred (Rogers and Harpending, 1992). The results are presented graphically in the form of mismatch distributions, which provide an indication of the history of the sequences and the resulting phylogenetic tree (Slatkin and Hudson, 1991). The results obtained for the different datasets are presented in Figures 5.54, 5.55, 5.56 and 5.57.

Analyses of datasets 1b, 2b and 5 (All L Sequences, All Africa L Sequences and L3 Sequences) generated curves that were ca. bell-shaped when compared to the smooth distribution expected for a constant population size. The results for datasets 1b, 2b and 5 are indicated in Figure 5.54.

The results presented in Figure 5.54 provide evidence for the occurrence of a rapid population expansion event or continued exponential growth for the African population, and the L3 mtDNA haplogroup (Rogers and Harpending, 1992; Page and Holmes, 1998). The presence of a second peak in the mismatch distribution calculated for dataset 1b can be attributed to the inclusion of mtDNA sequences that were obtained from countries out of Africa (United States of America and the Dominican Republic). These sequences represent individuals of African descent that can trace their African ancestry to individuals that were transported from Africa during the slave trade (refer to Section 3.1.4). To investigate the effect of the inclusion of mtDNA sequences obtained from individuals from countries outside Africa on dataset 1b, a dataset consisting of 187 L sequences was compiled. The sequences included in this dataset correspond to the 187 L sequences excluded from dataset 1b to generate dataset 2b (refer to Appendix H). A multimodal mismatch distribution was observed for this dataset consisting of mtDNA sequences from countries outside of Africa (not presented). The presence of the smaller second peak in the mismatch distribution determined for dataset 1b is therefore due to the inclusion of the mtDNA sequences obtained from individuals from countries outside Africa.

Figure 5.54 Graphical representation of pairwise differences for datasets 1b, 2b and 5 which exhibited deviation from neutrality and displayed characteristics of population expansion



Expected = expected pairwise nucleotide difference values under the assumption of a constant population size and the absence of recombination; Observed = observed pairwise nucleotide difference values. Expected values are represented by a smooth red line and the observed values are represented by a blue line. For comparison purposes, all the X-axes have a scale extending to 120 pairwise differences. Datasets: 1b = All L Sequences; 2b = All Africa L Sequences; 5 = L3 Sequences. Refer to Section 4.10.4 for detailed descriptions of datasets 1b, 2b and 5.

Other studies of African populations (Watson *et al.*, 1997; Harpending and Rogers, 2000) have reported similar results to that observed in the current investigation, i.e. smooth unimodal pairwise distributions indicating an ancient population explosion, for mtDNA sequences based on both control region sequences and RFLP analyses. Studies of African populations based on microsatellites in nuclear genes (Shriver *et al.*, 1997; Reich and Goldstein, 1998) similarly suggest that a population expansion event occurred in Africa. In the current investigation, due to information provided by these other investigations, it is thus more likely that the results presented in Figure 5.54 provide evidence for the occurrence of a rapid population expansion event.

The distribution generated by pairwise numbers of differences can however also be the consequence of natural selection or the separation of the human population into several relatively isolated sub-populations, rather than population expansion (Rogers, 1995). If an advantageous alteration occurred in a mitochondrion, the carriers of that alteration might increase in number until the alteration is fixed within the population. This process is similar to population growth as the number of female carriers is initially small, although it subsequently increases and starts to level off (Rogers, 1995). It has however been argued by Harpending *et al.* (1993) that the interpretation of selection is inconsistent with results obtained from investigations of between-population mismatch distributions, while the interpretation of isolated sub-populations may be possible. Due to incomplete information regarding the ethnic groups, geographic origin and populations from which the sequences contained in datasets 1b, 2b and 5 were obtained, sub-populations within the African population were not addressed in the current investigation. The effect of selection on human mitochondrial DNA variation is however discussed in Section 5.11.5.

Support for the observation of population growth in African populations is provided by the archaeological record (Rogers, 1995) and by investigations of human evolution which propose models that suggest population expansions that increase the geographical range of the human species (Lahr and Foley, 1998; Templeton, 2002). The observation of population growth in datasets 1b, 2b and 5 can therefore be explained by the evolutionary history of the African population. Datasets 1b and 2b consist of sequences representing L0, L1, L2 and L3 mitochondrial haplogroups, while dataset 5 contains only L3 sequences. As already stated, dataset 1b contains 187 sequences from individuals that are descendants of persons who were transported from Africa during the slave trade. A dataset consisting of these 187 sequences produced a multimodal mismatch distribution which indicates that the population remained constant in size (Di Rienzo and Wilson,

1991). This can be interpreted to mean that subsequent to the forced “migration” out of Africa due to the slave trade, the African population in countries outside of Africa remained constant in size. It is however more likely that the lack of population growth observed for this dataset can be ascribed to the fact that the sequences included in the dataset were mainly obtained from two different populations (African Americans vs. individuals from the Dominican Republic) and population growth prior to the migration is therefore masked in this dataset by the presence of sub-populations. Another possible explanation for the multimodal mismatch distribution is a loss of genetic diversity in the slave population due to the extinction of lineages when great numbers of people perished due to factors relating to the slave trade (seizure of slaves, marches to the coast, detention before sailing, the sea voyage, adjustment to new surroundings, labour on plantations) prior to and after reaching the New World (Klein *et al.*, 2001). The inclusion of these 187 sequences in dataset 1b however does not affect the overall result as population growth is still observed in the mismatch distribution generated using dataset 1b.

Watson *et al.* (1997) investigated expansion events within Africa and identified a pattern of demographic expansion. The distribution of L0b (L0d¹) and L0a (L0k) amongst the Khoi-San in southern Africa, L1b (L1c) in central and western Africans (particularly in the Biaka pygmies) and L5a and L0a (L0f) in East Africans are indicative of an early spread of L0 and L1 mtDNA haplogroups across Africa. This expansion is dated between the age of the MRCA and the onset of a major African re-expansion ca. 60,000 to 80,000 YBP. This second expansion repopulated Africa with L2 and L3 mtDNA haplogroups, whereas the original L0 and L1 haplogroups became a minority group except in the ancestors of the Khoi-San and Biaka pygmies. This re-expansion led to a successful migration of modern humans out of Africa and the rise of the M and N mtDNA haplogroups which were descended from the L3 haplogroup (Watson *et al.*, 1997).

The population growth observed in datasets 1b and 2b can therefore be explained by the dispersals and differentiation of mtDNA haplogroups within Africa, while the population expansion evident in dataset 5 can be attributed to an increase in carriers of the L3 haplogroup which migrated from Africa to populate the rest of the world (Watson *et al.*, 1997; Forster, 2004). Population growth within African populations could in fact have been

¹ The haplogroup classifications presented in brackets refer to the classifications according to the alternative classification scheme (refer to Section 3.3.1.1).

instrumental in driving the first migration from Africa, due to overpopulation and crowding leading to competition for available resources (Stringer, 2002; Cavalli-Sforza *et al.*, 1994) or due to the expansion of habitats in which humans could survive (Lahr and Foley, 1998). Due to climatic reasons, these habitat expansions occurred first in Africa during interglacial periods (Cavalli-Sforza *et al.*, 1994). A recent investigation by Atkinson *et al.* (2009) also observed population growth in the L3 haplogroup from 61,000 to 86,000 YBP. They argued that because L3 is the only haplogroup with descendants outside Africa and it exhibits a clear growth signal 8,000 to 12,000 YBP prior to the emergence of its non-African descendants (macrohaplogroups M and N), the L3 haplogroup did not simply spill over into Eurasia, but was driven there as part of an expansion that had begun in sub-Saharan Africa thousands of years earlier. They further attributed the observed pattern of population growth to an initially small group carrying the ancestral L3 haplotype which gained a cultural advantage that allowed its population to out-compete rival groups. On the basis of archaeological evidence, Mellars (2006) has argued that the African exodus was preceded by a cultural revolution involving new stone blade technologies, skin working tools, ornaments and imported red ochre.

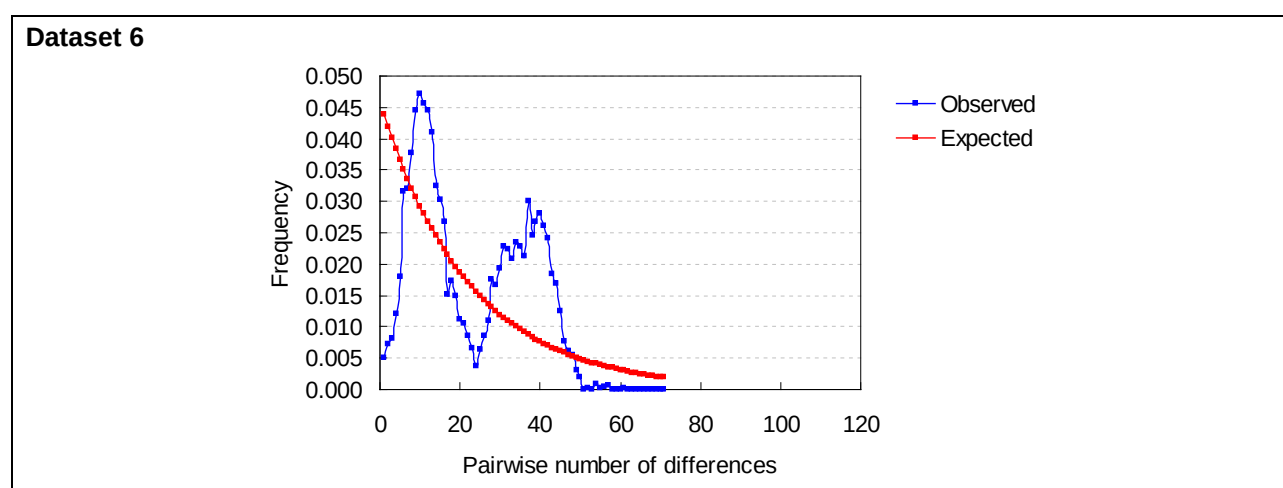
At present, three main models of modern human origins, namely African replacement, assimilation and multiregional evolution models (refer to Section 3.1.1), are employed to explain the origin of modern humans and their dispersal across all continents (Fagundes *et al.*, 2007). The observed population expansion in datasets 1b, 2b and 5 supports the “Recent African Origin” model as population growth within Africa is a requirement of this model. The “Multiregional Evolution” model however does not preclude population expansion (Rogers, 1995). The high nucleotide diversity values which were determined for datasets 1b, 2b and 5 however provide additional support for the “Recent African Origin” model compared to the “Multiregional Evolution” model.

Two studies (Ingman *et al.*, 2000; Mishmar *et al.*, 2003) which investigated complete mtDNA genome sequences have however reported contradictory results. Both these studies reported broad and ragged distributions for African mtDNA sequences belonging to L haplogroups, which is consistent with constant population size (Rogers and Harpending, 1992). This discrepancy with the data reported in the current investigation can be explained in the context of the size of the datasets used. Ingman *et al.* (2000) investigated a dataset consisting of only 21 African mtDNA sequences, while Mishmar *et al.* (2003) examined a total of 32 mtDNA sequences. In the current investigation the datasets consisted of 426 L sequences (dataset 1b), 239 African sequences (dataset 2b) or

128 L3 haplogroup sequences (dataset 5). The increased sizes of the datasets compared to that investigated by Ingman *et al.* (2000) and Mishmar *et al.* (2003) imply that the results obtained in the current investigation are less likely to be influenced by artefacts associated with small population sizes such as a loss in genetic diversity due to random genetic drift, therefore the results of the current study are more accurate.

Analysis of dataset 6 (L2 Sequences) generated a curve that was bimodal when compared to the smooth distribution expected for constant population size. The result for dataset 6 is indicated in Figure 5.55 and provides evidence for the absence of a rapid population expansion event (Page and Holmes, 1998), i.e. the population remained constant in size (Di Rienzo and Wilson, 1991). This result was unexpected as a population expansion in the L2 haplogroup has previously been suggested (Watson *et al.*, 1997).

Figure 5.55 Graphical representation of pairwise differences for dataset 6 which exhibited deviation from neutrality but displayed no evidence of population expansion



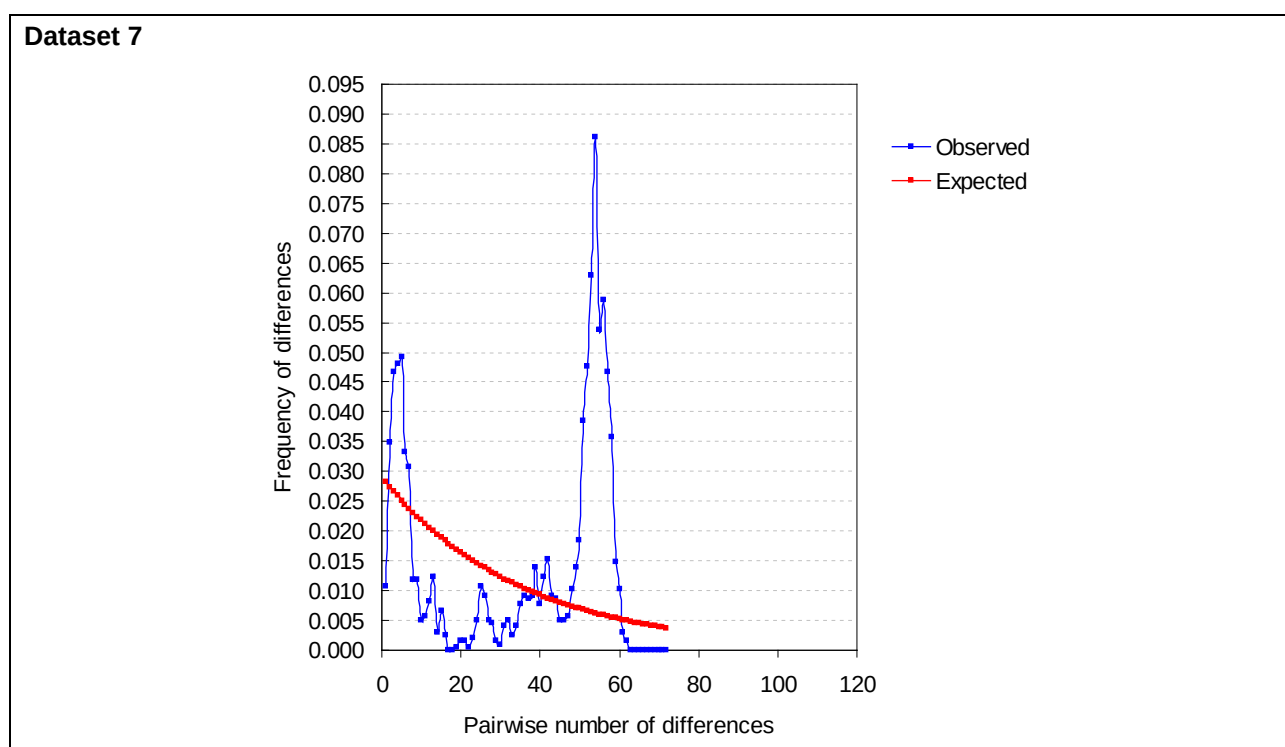
Expected = expected pairwise nucleotide difference values under the assumption of a constant population size and the absence of recombination; Observed = observed pairwise nucleotide difference values. Expected values are represented by a smooth red line and the observed values are represented by a blue line. For comparison purposes, the X-axis has a scale extending to 120 pairwise differences. Dataset 6 = L2 Sequences. Refer to Section 4.10.4 for a detailed description of dataset 6.

Torrioni *et al.* (2001) studied the phylogeny of complete sequences from the L2 haplogroup and observed that sequences from the L2b and L2d sub-haplogroups appeared to exhibit higher levels of sequence divergence compared to sequences from the L2a and L2c sub-haplogroups. It was suggested that the result could be explained by a succession of founder events and bottlenecks which could produce the extreme divergence patterns observed in the L2 lineages (Torrioni *et al.*, 2001). This could explain the result obtained in Figure 5.55, if it is assumed that population expansion of the L2 haplogroup, as reported by Watson *et al.* in 1997, has been obscured by subsequent bottleneck and founder events. It has also been suggested that different selective pressures acted on the different

L2 lineages giving rise to the divergence patterns observed in the L2 phylogeny (Torrioni *et al.*, 2001). Selection is one of the two explanations suggested for deviation from neutrality in dataset 6 (refer to Section 5.11.3) and is discussed in Section 5.11.5.

Analyses of datasets 3a, 4b, 7 and 8a (All Khoi-San L Sequences, !Kung Sequences, L1 Sequences and L0 Sequences) generated distributions that were ragged and erratic, with more than one peak being present in addition to the peaks being widely spaced. The results for datasets 3a, 4b, 7 and 8a are presented in Figure 5.56, 5.57 and 5.58. These results provide evidence for the absence of a rapid population expansion event (Page and Holmes, 1998) and instead provide support that the population remained constant in size (Di Rienzo and Wilson, 1991).

Figure 5.56 Graphical representation of pairwise differences for dataset 7 which did not exhibit deviation from neutrality and displayed characteristics of constant population size

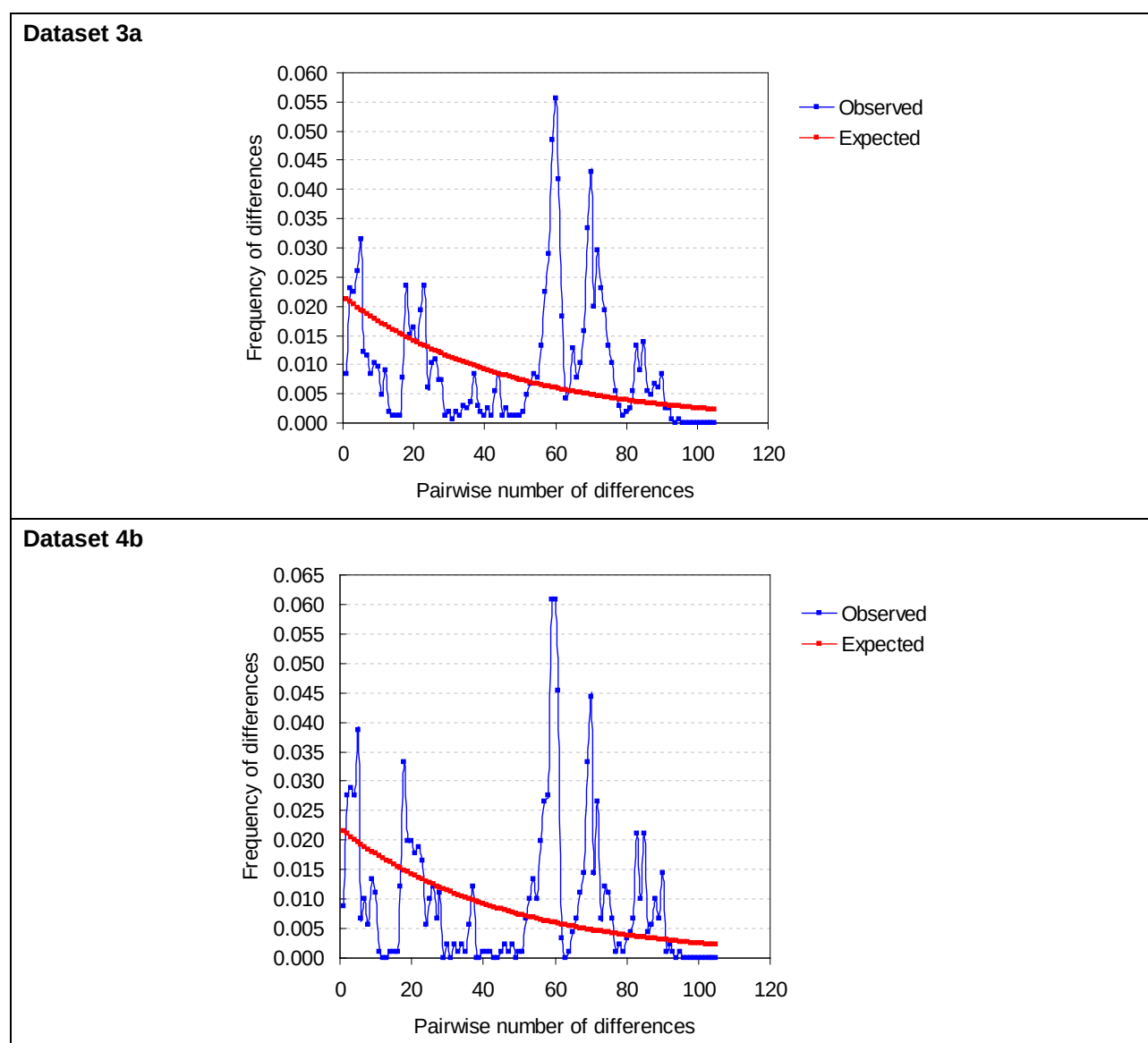


Expected = expected pairwise nucleotide difference values under the assumption of a constant population size and the absence of recombination; Observed = observed pairwise nucleotide difference values. Expected values are represented by a smooth red line and the observed values are represented by a blue line. For comparison purposes, the X-axis has a scale extending to 120 pairwise differences. Dataset 7 = L1 Sequences. Refer to Section 4.10.4 for a detailed description of dataset 7.

Dataset 7 consists of L1 sequences. The result obtained, which indicated a constant population size over time, was unexpected compared to a previous investigation (Watson *et al.*, 1997) which suggested population expansion for the L1 mtDNA haplogroup based on control region sequence data (bp 16,090 to 16,365). This discrepancy can be explained if it is assumed that the signal of a population expansion event is obscured by subsequent

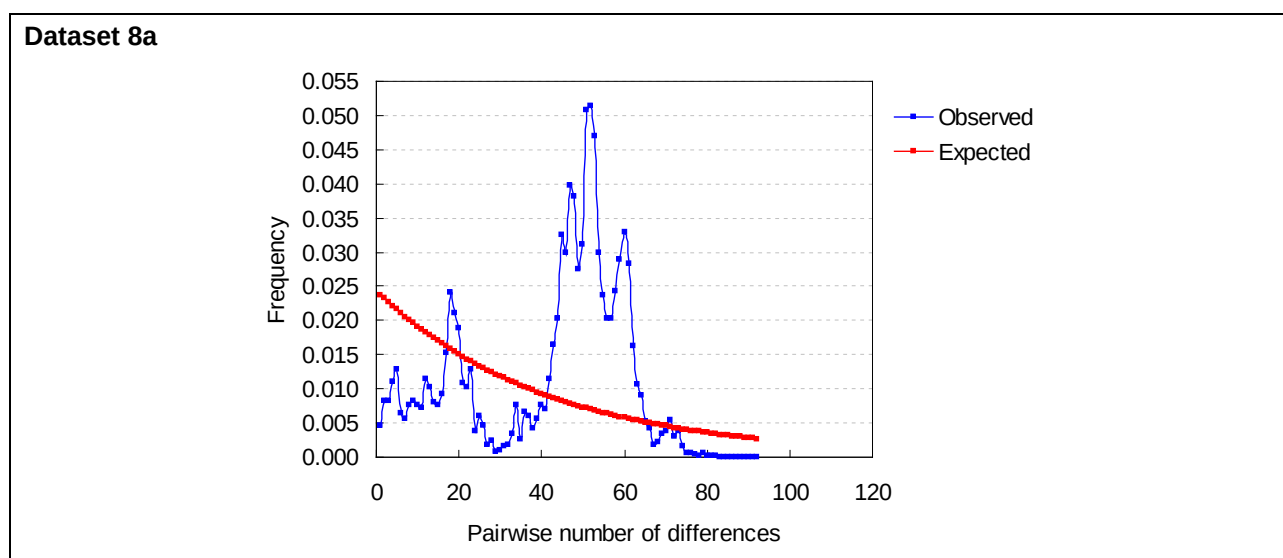
population events such as a population bottleneck. This is possible when the age of the L1 haplogroup (refer to Section 3.3.1.4) and the hypothesis that the L1 lineage was replaced by the L2 and L3 lineages during population dispersion and repopulation within Africa (Watson *et al.*, 1997) is considered. It is also possible that the ragged mismatch distribution generated for the dataset consisting of L1 mtDNA haplogroup sequences is due to the grouping together of sequences from different ethnic groups or populations. In a very diverse gene pool, such as that of Africa, population fusion events can easily generate ragged mismatch distributions in expanding populations (Bandelt and Forster, 1997).

Figure 5.57 Graphical representation of pairwise differences for datasets 3a and 4b which did not exhibit deviation from neutrality and displayed characteristics of constant population size



Expected = expected pairwise nucleotide difference values under the assumption of a constant population size and the absence of recombination; Observed = observed pairwise nucleotide difference values. Expected values are represented by a smooth red line and the observed values are represented by a blue line. For comparison purposes, all the X-axes have a scale extending to 120 pairwise differences. Datasets: 3a = All Khoi-San L Sequences; 4b = !Kung Sequences. Refer to Section 4.10.4 for detailed descriptions of datasets 3a and 4b.

Figure 5.58 Graphical representation of pairwise differences for dataset 8a which did not exhibit deviation from neutrality and displayed characteristics of constant population size



Expected = expected pairwise nucleotide difference values under the assumption of a constant population size and the absence of recombination; Observed = observed pairwise nucleotide difference values. Expected values are represented by a smooth red line and the observed values are represented by a blue line. For comparison purposes, the X-axis has a scale extending to 120 pairwise differences. Dataset 8a = L0 Sequences. Refer to Section 4.10.4 for a detailed description of dataset 8a.

Dataset 3a contains sequences obtained from Khoi-San individuals, dataset 4b is composed of sequences obtained from individuals belonging to the !Kung ethnic group which are part of the greater Khoi-San language family and dataset 8a consists of sequences classified as belonging to the L0 mtDNA haplogroup. The Khoi-San and the !Kung are classified as hunter-gatherers (refer to Section 3.7) and the L0 haplogroup is predominantly observed in populations that are classified as hunter-gatherers, i.e. food gathering populations (refer to Section 3.3.1.3). Most human populations present with unimodal mismatch distributions, which imply demographic expansions during the Pleistocene (Slatkin and Hudson, 1991; Rogers and Harpending, 1992; Rogers, 1995). Differences have however been observed in the mismatch distributions of mtDNA sequences for populations classified as food gatherers, consisting predominantly of the L0 mtDNA haplogroup, and populations such as the Khoi-San and !Kung, compared to food producers, consisting of predominantly L2 and L3 mtDNA haplogroups (Ray *et al.*, 2003). The mismatch distributions observed for most present or former hunter-gatherer populations (food gatherers) are ragged and multimodal, indicating a constant population size, in contrast to the mismatch distributions observed for post-Neolithic populations (food producers) which are bell-shaped, suggesting population expansion (Excoffier and Schneider, 1999).

Other studies on African populations classified as hunter-gatherers have reported similar results compared to that obtained in the current investigation. Di Rienzo and Wilson (1991) investigated a segment of the mtDNA control region for two African populations consisting of 26 !Kung individuals and 37 Pygmy individuals. Their investigation revealed multimodal distributions with significant disagreement with a Poisson distribution ($P < 0.0001$), whereas an approximately Poisson distribution is expected for growing populations. Watson *et al.* (1996) studied HVS1 sequences of nine ethnic groups which included three groups consisting of 19 !Kung individuals, 17 Biaka (Western Pygmy) individuals and 11 Mbuti (Eastern Pygmy) individuals. They observed ragged distributions indicative of a constant population size over a long time (Slatkin and Hudson, 1991; Rogers and Harpending, 1992) and explained their results compared to other African populations in context of the lifestyles of the populations. The !Kung, Mbuti and Biaka are food gatherers and follow a lifestyle consisting of hunting and food gathering, while other African populations are food producers that practice agriculture and animal husbandry (Cavalli-Sforza *et al.*, 1994). The difference in mismatch distributions is due to the transition from an economy based on hunting and gathering to one based on food production, which results in an increase in population size due to the potential for redistribution of food resources throughout the year thereby mitigating seasonal food scarcity which was likely an important cause of mortality in hunter-gatherer cultures (Ammerman and Cavalli-Sforza, 1971). Wilder *et al.* (2004a) studied the variation within 782 bp of the mitochondrial cytochrome c oxidase subunit 3 gene as well as 26.5 kb of non-coding DNA from the non-recombining portion of the Y chromosome (NRY) in a sample of 25 Khoi-San individuals. They concluded that data from both loci were best described by a model of constant population size. Pilkington *et al.* (2008) investigated 780 bp of the cytochrome c oxidase subunit 3 gene in five populations consisting of 25 Khoi-San individuals from Namibia, 24 Bakola individuals from Cameroon, 49 Dogon individuals from Mali, 23 Dinka individuals from Sudan and 50 Southeast Bantu-speakers (Zulu, Ndebele, Xhosa, Sotho, Swazi, Tswana, Pedi and Tsonga) from southern Africa. Based on the results generated in their investigation, they concluded that the two food-gatherer populations (Khoi-San and Bakola) better fit a model of constant population size, while the three food-producer populations (Dogon, Dinka and Southeast Bantu-speakers) better fit a model of population expansion. Based on the studies discussed above, it is evident that a correlation exists between lifestyle and population size, i.e. the population size of hunter-gatherer populations (Khoi-San, !Kung, Biaka, Bakola and Mbuti) have remained constant over time, while food producing populations have experienced population expansion.

Alternative explanations for the observed differences in mismatch distributions for hunter-gatherer populations compared to food producing populations have been proposed. The absence of traces of Pleistocene expansions in present day hunter-gatherers can be explained, in the context of unsubdivided populations, by the occurrence of a recent bottleneck in the history of the remaining hunter-gatherer populations (Excoffier and Schneider, 1999). A bottleneck effect, or the reduction of the effective population size, can be achieved through a drastic reduction in absolute census size, or by fragmentation of the environment, leading to a metapopulation phase with much smaller effective size (Excoffier and Schneider, 1999). It is likely that the rise of competing Neolithic farmers caused the Holocene hunter-gatherer populations to enter such a metapopulation phase by pushing the hunter-gatherers into refuge areas not easily exploited by farmers or pastoralists (Excoffier and Schneider, 1999). Population expansion events during the Pleistocene in hunter-gatherer populations are thus subsequently erased in populations that did not go through the Neolithic transition, i.e. food gathering populations that did not become food-producing populations. This corresponds with the pattern of population expansion in Africa suggested by Watson *et al.* (1996) where subsequent repopulation by L2 and L3 haplogroups (food producers) reduced the frequency of L1 and L0 haplogroups (food gatherers) so that they are presently predominantly observed in hunter-gatherer populations such as the !Kung and Pygmies.

Ray *et al.* (2003) suggested another possible explanation for the difference observed in mismatch distribution between hunter-gatherer and post-Neolithic populations. They assumed that the present distribution of human populations resulted from some spatial range expansion and that the contrast in mismatch distributions between food producers and food gatherers could therefore simply result from the higher population density of Neolithic populations compared to hunter-gatherer populations which resulted in larger gene flow values for Neolithic populations than for hunter-gatherer populations (Ray *et al.*, 2003). In addition, it is possible that food producing populations could have experienced both a range expansion following a climate change, e.g. after the last glacial maximum, or a local demographic increase, perhaps driven by some local adaptation or cultural innovation, e.g. the Neolithic transition (Excoffier, 2004) to generate the unimodal mismatch distributions observed in various investigations. In the current investigation it is therefore suggested that the ragged mismatch distributions observed for datasets 3a, 4b and 8a are the result of a subsequent population bottleneck caused by contact with food producing populations leading to a loss of the expected population expansion signal in the mismatch distribution.

The model employed to investigate population size changes in the current investigation was relatively simple and assumed that populations, and by implication the datasets under investigation, were not subdivided (Slatkin and Hudson, 1991; Rogers and Harpending, 1992; Rogers, 1995; Excoffier and Schneider, 1999). Population expansion, as measured by mismatch distributions, are however influenced by population structure (subdivided populations), migration between sub-populations and range expansions (Ray *et al.*, 2003; Excoffier, 2004). These aspects were not included in the population growth model employed and it is therefore not possible to make observations regarding these factors. In addition, information regarding population substructure, migration and range expansions were not available for the datasets under investigation and a limitation of the current study is thus that these aspects are therefore not addressed in the current investigation. It has however been observed that a pure demographic expansion and a range expansion in subdivided populations both generate unimodal mismatch distributions within demes (Excoffier, 2004). Discrepancies regarding population expansions observed in different populations can be due to differences in demographic histories among regions, and among ethnic groups (food gatherers vs. food producers), differences between loci, ascertainment bias in the choice of markers, a lack of resolution of some markers, or from making inferences based on erroneous models of population history (Ray *et al.*, 2003). In the current investigation it must be concluded that the observed mismatch distributions for datasets 1b, 2b and 5 which exhibit characteristics of population expansion, and datasets 3a, 4b, 6, 7 and 8a which exhibit characteristics of constant population size could have been influenced by other factors such as population subdivision, migration and range expansion which could have contributed to population size changes, or bottlenecks and fragmentation of habitats which could have led to removal of the population expansion signal in mismatch distributions. The results presented for the mismatch distributions in this section are therefore applicable to the specific datasets and are based on the specific rapid population growth model employed in the current investigation (refer to Section 4.14.5.1).

5.11.4.2 Statistical parameters to detect population expansion

In addition to the graphical representation of the pairwise number of nucleotide differences, a number of statistical parameters can be employed to detect population expansion processes and to describe the smoothness of the observed pairwise differences distributions. These statistical parameters provide support for the observations made

pertaining to the visual differences between mismatch distributions. The statistical parameters are listed in Table 5.52.

Table 5.52 Statistical parameters to evaluate population size changes in different datasets

Dataset	Number of sequences (n)	Raggedness statistic (rg)	P -value	Ramos-Onsins and Rozas statistic (R_2)	P -value	Fu's F_s statistic	P -value
1b	426	0.0001	0.000	0.0186	0.000	-358.509	0.000
2b	239	0.0004	0.190	0.0247	0.000	-34.088	0.007
3a	58	0.0047	0.647	0.0786	0.228	-6.350	0.053
4b	43	0.0094	0.816	0.0969	0.362	-4.275	0.100
5	128	0.0009	0.055	0.0206	0.000	-98.408	0.000
6	113	0.0009	0.008	0.0287	0.000	-67.854	0.000
7	63	0.0049	0.648	0.0628	0.055	-7.133	0.061
8a	103	0.0020	0.489	0.0455	0.010	-31.906	0.000

N = number of sequences; rg = raggedness statistic; R_2 = Ramos-Onsins and Rozas statistic. P -value = probability of obtaining values equal or lower than the observed value. $P < 0.05$ is statistically significant and is indicated in red text. $P > 0.05$ is not statistically significant. Datasets: 1b = All L Sequences; 2b = All Africa L Sequences; 3a = All Khoi-San L Sequences; 4b = !Kung Sequences; 5 = L3 Sequences; 6 = L2 Sequences; 7 = L1 Sequences; 8a = L0 Sequences. Refer to Section 4.10.4 for detailed descriptions of the different datasets.

The rg statistic quantifies the smoothness of the frequency of pairwise differences distributions (Harpending *et al.*, 1993; Harpending, 1994). The values obtained for datasets 1b, 2b, 5, 6 and 8a are all quite low compared to the values obtained for the other datasets and appear to provide support for a rapid population expansion event in these datasets. Population expansion is however only supported by the P -values calculated for datasets 1b and 6. The R_2 statistic is based on the difference between the number of singleton alterations (alteration present in only one sequence in the dataset) and the average number of nucleotide differences (Ramos-Onsins and Rozas, 2002). The values obtained for datasets 1b, 2b, 5, 6 and 8a are all low compared to the values obtained for the other datasets and seem to support a rapid population expansion event. The calculated P -values for these datasets support the presence of population expansion events in each of these datasets. The F_s test statistic (Fu, 1997) is based on the haplotype (gene) frequency distribution and is dependent on the value of theta (Ewens, 1972). The large negative values obtained for datasets 1b, 2b, 5, 6 and 8a compared to the other datasets provide evidence for a rapid population expansion event in these datasets which is supported by the calculated P -values.

The different results obtained for the datasets when using the rg , R_2 and Fu's F_s statistics can be attributed to differences in statistical power between the different statistics. The rg statistic, due to it being based on the mismatch distribution, uses little of the information

accumulated within the data (Felsenstein, 1992) and it is therefore very conservative. The *rg* statistic therefore has low statistical power for detecting population expansion, although it has the advantage that it is not affected by population subdivision. The fact that population expansion is not supported for datasets 2b, 5 and 8a when using the *rg* statistic compared to the R_2 and F_S statistics, can be attributed to this lack of statistical power. Therefore, more powerful statistical measures such as the R_2 (Ramos-Onsins and Rozas, 2002) and Fu's F_S test statistic (Fu, 1997) were employed. These statistics have the greatest power to detect population expansion for non-recombining regions of the genome under a variety of different circumstances (Ramos-Onsins and Rozas, 2002). The R_2 test has superior performance in the case of small sample sizes (and a small number of segregating sites), whereas Fu's F_S performs better for large sample sizes. This would imply that the results obtained for datasets 3a, 4b and 7 using the R_2 statistic will be more reliable than that obtained for the same datasets using Fu's F_S statistic, while the results obtained for the remaining datasets using Fu's F_S statistic will be more reliable than that observed using the R_2 statistic. In the current investigation, the same results were however obtained for all datasets using both statistical measures.

The results obtained for the different tests of neutrality, the pairwise comparisons test and the statistical tests of population expansion can be summarised as follows:

- a. Datasets 1b, 2b and 5 consisting of all L sequences, all Africa L sequences and L3 sequences displayed deviation from neutrality and this was probably due to a rapid population expansion event.
- b. Datasets 3a, 4b and 7 containing Khoi-San, !Kung and L1 sequences, respectively, did not exhibit deviation from neutrality and no evidence of a population expansion event was observed in these datasets.
- c. Dataset 6 composed of L2 sequences displayed deviation from neutrality and the other tests evaluating population size supported a population expansion event (refer to Table 5.52), although the distribution generated by the pairwise differences test was ragged indicating constant population size.
- d. Dataset 8a consisting of L0 sequences presented with deviation from neutrality based on the Fu and Li D^* and F^* tests only, and displayed evidence of a population expansion event based on the results obtained for the R_2 and Fu F_S statistics. The results generated for the different datasets are discussed in the paragraphs that follow.

The rapid population expansion event observed for datasets 1b, 2b and 5 was not surprising. Datasets 1b and 2b represent the African population as a whole and consists of L0, L1, L2 and L3 sequences, while dataset 5 represents the L3 haplogroup. Previous investigations of African populations have provided evidence for the presence of population expansion events in African populations (Sherry *et al.*, 1994; Reich and Goldstein, 1998), while the L3 mtDNA haplogroup has been recognised as the precursor of modern European and Asian mtDNA haplogroups (Chen *et al.*, 1995; Watson *et al.*, 1997). A geographic expansion is expected to be preceded by an accompanying population expansion event (Watson *et al.*, 1997). As a result of the geographic expansion event and exposure to new environments where adaptation is required, distortions occur in the distribution of mtDNA due to arising nucleotide alterations. The results obtained for datasets 1b, 2b and 5 therefore support the “Out of Africa” model of human evolution (refer to Section 3.1.1).

The results obtained for datasets 3a and 4b were also not surprising. As already discussed, datasets 3a and 4b represent hunter-gatherer populations that are classified as food gatherers which usually do not display evidence of population expansion (refer to Section 5.11.4.1). In addition, a recent investigation performed by Pilkington *et al.* (2008) on the cytochrome c subunit 3 gene of 25 Khoi-San individuals reported similar R_2 values (0.096; $P > 0.05$) which, in conjunction with their Khoi-San mismatch distribution, was interpreted as being consistent with a constant population size.

Dataset 7 consists of L1 sequences, and the result obtained for this group was unexpected when compared to a previous investigation performed by Watson *et al.* in 1997 (refer to Section 5.11.4.1). However, dataset 7 contains sequences from Bakola, Biaka and Pygmy individuals (refer to Appendix G) that belong to hunter-gatherer populations. This implies that the arguments presented for datasets 3a and 4b can also apply to dataset 7, i.e. the signature of constant population size is expected, although it is possible that a population expansion event has been obscured by a subsequent population bottleneck (refer to Section 5.11.4.1). It is important to note that of all the tests used to investigate population expansion in the current investigation, the mismatch distribution test has the weakest power to detect growth (Ramos-Onsins and Rozas, 2002), due to the fact that information is lost when studying pairwise differences data rather than the sequences themselves (Marjoram and Donnelly, 1994), although all tests used may be sensitive to unknown structure within populations (Ptak and Przeworski, 2002; Hammer *et al.*, 2003). In addition, the statistical tests used in the current investigation to detect the presence of a population

expansion event are unlikely to detect population expansion that occurred very recently or is very ancient (Simonsen *et al.*, 1995; Fu, 1997; Ramos-Onsins and Rozas, 2002). Furthermore, large datasets are required to detect small population growth events (Ramos-Onsins and Rozas, 2002). These three facts support the argument that population growth could have occurred, but was not detected, in the populations or haplogroups represented by datasets 3a, 4b and 7. This could be attributed to a lack of sensitivity of the methods used to detect population expansion, or alternatively no evidence of such events remained due to the time at which the population growth occurred, or the size of the dataset investigated was insufficient.

A recent study performed by Atkinson *et al.* (2009) investigated population growth in African populations by determining Bayesian skyline plots which are based on a Bayesian coalescent inference algorithm together with a Markov chain Monte Carlo sampling algorithm. This method allows the estimation of population demographic parameters through time with increased accuracy, without assuming a simple parametric growth curve and with an explicit framework for quantifying uncertainty in population parameter estimates (Shapiro *et al.*, 2004; Atkinson *et al.*, 2008). In addition, this method avoids the loss of information that is associated with distance-based methods, because it is applied directly to a set of sequence data rather than to pairwise sequence differences (Marjoram and Donnelly, 1994). Using this method, Atkinson and colleagues observed slow constant growth over 100,000 to 200,000 YBP for their datasets consisting of L0 and L1 sequences. This result supports the argument for population growth in datasets 3a and 4b presented in the previous paragraph, and therefore population growth most likely occurred, although it was not observed in the current investigation due to limitations of the methods used, including a lack of sensitivity, to detect ancient population growth obscured by subsequent events.

The discrepancies in the results observed for datasets 6 and 8a can be explained as follows. The performance of different statistical tests of neutrality under different models, such as population growth, genetic hitchhiking or background selection, is not equivalent. These models share the characteristic that DNA polymorphisms tend to exhibit excesses of rare alleles or young mutations (Fu, 1997). The differences in performance of the various statistical tests can be an indication of the underlying mechanism that is responsible for the observed polymorphism, i.e. deviation from the neutral model of evolution (Fu, 1997). In a comparison of different statistical tests, Fu and Li's D^* and F^* tests performed better at identifying background selection than Tajima's D test and Fu's F_s

test, while Fu's F_s test performed better in the detection of population growth and genetic hitchhiking (Fu, 1997). Therefore, if only Fu and Li's D^* and F^* tests are significant, this suggests that background selection is the more likely cause of the observed deviation from neutrality (Fu, 1997). This suggests that in the case of dataset 8a, the observed deviation from neutrality could be due to the effect of selection, although a rapid population expansion event can not be ruled out (refer to preceding paragraphs and Section 5.11.4.1). The effect of selection as a possible explanation for the results obtained for L0 sequences will be investigated in Section 5.11.5.

In the case of dataset 6, all the tests suggest population expansion except for the pairwise comparison test. The pairwise comparison test is viewed as very conservative due to the fact that the method employed uses little of the information accumulated in the data (Felsenstein, 1992). This implies that the method falls short in identifying some instances of population expansion events, especially in populations that approach the threshold values of the test. This interpretation is supported by results previously obtained. Watson *et al.* (1997) used a phylogenetic approach to detect sequences that could reveal recent population expansion events, and the sequences were subsequently subjected to pairwise comparison tests. L2 sequences displayed a smooth unimodal distribution indicative of population expansion, and in addition to this the diverse geographical spread of the sampled sequences hinted at an associated geographic expansion. According to Watson *et al.* (1997), one or a small set of closely related lineages expanded in Africa, giving rise to a L2/L3 cluster. Of this cluster an eastern African subset subsequently spread into Eurasia. This implies that a rapid population expansion for L2 sequences is feasible if it is assumed that a population expansion occurred within Africa in order to generate the L2/L3 population from which African L3 mtDNA haplogroups spread to Eurasia. A recent study by Atkinson *et al.* (2009) provides further support for population expansion in dataset 6. They observed population growth for the L2 haplogroup beginning 12,000 to 20,000 YBP and identified the L2a sub-haplogroup as the main source of the population expansion within the L2 haplogroup. They attributed population growth in the L2 haplogroup to the spread of L2 lineages due to environmental changes associated with the last glacial maximum (Salas *et al.*, 2002) which may have allowed an expansion of human populations into new territories, particularly in Central Africa.

In the current investigation, approximate dates for population expansion events and female population sizes in datasets that exhibited evidence for population expansion (datasets 1b, 2b, 5 and 6) were not estimated due to the limitations of the simple rapid population

expansion model used. Values determined under this pure demographic expansion model are based on the assumption of an undivided population and an implicit migration rate of zero (Rogers and Harpending, 1992). The model therefore does not take into account population subdivision and the subsequent migration events between subpopulations or range expansions in addition to demographic expansion (Excoffier, 2004). Estimation of dates or population sizes under this model was therefore not performed as it is likely that unrealistic population expansion dates or sizes would be obtained due to the fact that no correction is made for the abovementioned factors when using this simple model (Wilder *et al.*, 2004b).

5.11.5 The effect of selection on human mitochondrial DNA variation

Two methods, namely the analysis of synonymous and non-synonymous substitutions (refer to Section 4.14.6.1) and the calculation of the conservation index (refer to Section 4.14.6.2) were employed to evaluate the effect of selection on mitochondrial DNA sequences belonging to macrohaplogroup L. Both these methods attempt to determine the importance of selection in shaping human mtDNA variation.

5.11.5.1 Analysis of synonymous and non-synonymous substitutions

The effect of selection on mitochondrial DNA sequences was determined as described in Section 4.14.6.1 by calculating the ratio of non-synonymous substitutions versus synonymous substitutions for both internal and terminal branches of the phylogenetic networks representing the coding region of the L0, L1, L2 and L3 haplogroups. As discussed in Section 2.5.5, the ratio of non-synonymous/synonymous substitutions in the internal branches (RF_I) compared to the ratio of non-synonymous/synonymous substitutions in the terminal branches (RF_T) can be used as a measure of adaptive selection compared to purifying selection within a population of mtDNA sequences, i.e. within an mtDNA haplogroup.

The L0 phylogenetic network constructed in Section 5.10 was used to calculate the replacement mutation frequency (RF) values for the L0 haplogroup, while the phylogenetic networks published by Ruiz-Pesini *et al.* (2007) were used to calculate RF values for the L1, L2 and L3 haplogroups. The RF values calculated for both the internal and terminal branches for the different haplogroups are indicated in Table 5.53.

A test for the effect of selection was performed using the one-tailed Fisher's Exact test (refer to Section 4.14.6.1). The null hypothesis of selective neutrality, i.e. RF_I is equal to RF_T , was rejected when the one-tailed P -value was less than 0.05.

Table 5.53 Selection parameters for the L-haplogroups

Haplogroup	N	Internal branches		Terminal branches		$H_0: RF_I = RF_T$	I/T
		RF_I	NS/S	RF_T	NS/S	P -value	
L ^a	330	0.353	147/417	0.479	217/453	0.009	0.737
L0	100	0.373	47/126	0.630	58/92	0.019	0.592
L1	59	0.356	32/90	0.483	43/89	0.177	0.737
L2	87	0.376	41/109	0.458	55/120	0.247	0.821
L3	84	0.293	27/92	0.401	61/152	0.147	0.731

^a = the RF_I and RF_T values for macrohaplogroup L are determined from the sum of the individual NS and S of the internal and terminal branches for each haplogroup. N = population size (number of mtDNA sequences); RF_I = replacement mutation frequency within internal branches; RF_T = replacement mutation frequency within terminal branches; NS = non-synonymous substitution; S = synonymous substitution; P -value = probability value; I/T = RF_I/RF_T . $H_0: RF_I = RF_T$ = null hypothesis that there is no difference between RF_I and RF_T . Significant P -values are indicated in red text.

The RF_I is less than the RF_T for all the L haplogroups indicating that more non-synonymous substitutions were present in the terminal branches than the internal branches of the network. The presence of more non-synonymous substitutions in the terminal branches indicate that the majority of non-synonymous substitutions are recent and are being removed by purifying selection. As expected, the calculated RF_I/RF_T (I/T) values were therefore low, with the lowest I/T value calculated for L0, while higher I/T values were determined for L1 to L3, with L2 being the highest. A low I/T (I/T < 1.00) predicts the influence of purifying selection. A significant difference between the RF for internal and terminal branches was observed for L0, hence the null hypothesis of neutral evolution was rejected for L0 ($P < 0.05$). This indicates that purifying selection has played a role in shaping the mtDNA diversity observed in the L0 haplogroup. This result was not surprising. As discussed previously, deviation from neutrality was observed for the L0 haplogroup which was attributed either to the possible effect of selection or a population growth event (refer to Section 5.11.3). The result generated in this section therefore appears to confirm the role of selection, in addition to population expansion (refer to Section 5.11.4), in shaping the evolution of the L0 haplogroup.

A significant difference between RF_I and RF_T was also observed for macrohaplogroup L. This result suggests that macrohaplogroup L has been subjected to purifying selection. This was not surprising as it has previously been reported by Ruiz-Pesini *et al.* in 2004, based on the analysis of a global human mitochondrial DNA sequence dataset containing L-specific sequences, that the predominant factor influencing human mtDNA variation has

been purifying selection due to the essential function of the mtDNA proteins. It has however been demonstrated in the current investigation that population growth events have also occurred in macrohaplogroup L (refer to Section 5.11.4). The deviation from neutrality observed for macrohaplogroup L (refer to Section 5.11.3) is therefore due to two factors, namely population growth and purifying selection.

The differences in RF values between internal and terminal branches for L1 to L3 were not significant ($P > 0.05$), leading to the hypothesis of neutral selection not being rejected. Therefore, although low I/T values that suggest purifying selection were observed for the L1, L2 and L3 haplogroups, a role for selection could not be confirmed in these haplogroups. This result was as expected for the L1 haplogroup, as this haplogroup had not exhibited deviation from neutrality (refer to Section 5.11.3). The results observed for the L2 and L3 haplogroups can likely be attributed to population growth, as it was demonstrated in the current investigation that a population growth event was likely for both these haplogroups (refer to Section 5.11.4). The results presented in this section for the L2 and L3 haplogroups appear to confirm that the deviation from neutrality observed for these two haplogroups are due to population growth events.

5.11.5.2 Conservation index

The influence of purifying versus adaptive selection was further analysed as described in Section 4.14.6.2 by determining the conservation index (CI) of each non-synonymous substitution in the global L0-specific haplogroup network (refer to Section 5.10.2) through comparison between 39 different species. In the current investigation, conservation indices were defined as follows by indicating the number of different species that contained the wild type amino acid: very low (0 to 10), low (10 to 20), intermediate (20 to 30) and high (30 to 39). A non-synonymous substitution with a high conservation index indicates that the majority, i.e. 30 to 39, of the 39 species to which the amino acid at a specific position was compared, contained the wild type amino acid. A high conservation index therefore also implies that the amino acid at that position is of functional importance, as most species contain the same amino acid at that specific position. Adaptive or deleterious non-synonymous substitutions frequently alter functionally important amino acids, while functionally unimportant amino acids are often affected by synonymous substitutions (Ruiz-Pesini *et al.*, 2007). Adaptive or deleterious alterations should therefore occur at sites which have a high conservation index, while neutral alterations generally occur at sites with a low conservation index (Ruiz-Pesini *et al.*, 2004). The conservation index of

each site at which a non-synonymous substitution was identified within the global L0-specific haplogroup network is indicated as a number after the non-synonymous substitution (refer to Section 5.10.2). A list of the nucleotide positions affected by non-synonymous substitutions, as well as the conservation index for the position, is provided in Table 5.54.

Table 5.54 Non-synonymous substitutions identified within the constructed global L0-specific haplogroup network based on the coding region of 100 mtDNA genome sequences

Nucleotide position	Amino acid change	Gene	Conservation index	Position in global L0-specific haplogroup network
3,421	V39I	ND1	11	---
3,565	T87A	ND1	11	---
3,644	V113A	ND1	39	Terminal branch
3,736	V144I	ND1	36	Terminal branch
3,808	T168A	ND1	37	Terminal branch
3,866	I187T	ND1	33	Branching point at L0c1B
3,945G	I213M	ND1	7	---
3,946	E214K	ND1	39	Terminal branch
4,025	T240M	ND1	12	---
4,225	M307V	ND1	11	---
4,232	I309T	ND1	31	Branching point between L0b and L0aB
4,596	V43I	ND2	8	---
4,633A	A55D	ND2	37	Terminal branch
4,771	A101V	ND2	37	Terminal branch
4,812	V115I	ND2	31	Branching point at L0aB
4,917	N150D	ND2	35	Terminal branch
5,442	F325L	ND2	7	---
5,460	A331T	ND2	2	---
5,911	A3V	COI	5	---
6,249	A116T	COI	8	---
6,607	F235S	COI	39	Terminal branch
7,257	I452V	COI	10	---
7,258	I452T	COI	10	---
7,285A	S461Y	COI	39	Terminal branch
7,347	V482I	COI	37	Terminal branch
7,389	Y496H	COI	9	---
7,673	I30V	COII	34	Branching point within L0aA
8,013G	V143G	COII	35	Terminal branch
8,027	A148T	COII	10	---
8,214A	V210D	COII	37	Terminal branch
8,223G	L213W	COII	23	---
8,420	T19A	ATPase8	28	---
8,459	N32D	ATPase8	8	---

Table 5.54 Continued...

Nucleotide position	Amino acid change	Gene	Conservation index	Position in global L0-specific haplogroup network
8,460	N32S	ATPase8	8	---
8,498	K45E	ATPase8	10	---
8,508	N48S	ATPase8	11	---
8,516	W51R	ATPase8	38	Terminal branch
8,545	A7T	ATPase8/6	25	---
8,566	I14V	ATPase8/6	13	---
8,572	G16S	ATPase6	36	Terminal branch
8,641C	N39H	ATPase6	29	---
9,037	M171V	ATPase6	23	---
9,136	I204V	ATPase6	8	---
9,178	V218M	ATPase6	32	Branching point within L0c2B
9,181	S219G	ATPase6	37	Terminal branch
9,214	H3R	COIII	39	Terminal branch
9,300	A32T	COIII	12	---
9,438	G78S	COIII	36	Branching point within L0bB2
9,804	A200T	COIII	36	Terminal branch
9,978G	W258G	COIII	39	Terminal branch
10,084	I9T	ND3	12	---
10,143	G29S	ND3	1	---
10,237	I60T	ND3	39	Terminal branch
10,599	A44T	ND4L	12	---
10,707G	S80P	ND4L	10	---
10,811A	S18T	ND4	34	Terminal branch
10,812G	S18C	ND4	34	Terminal branch
10,841	T28A	ND4	35	Terminal branch
10,920	P54L	ND4	5	---
11,061	S101F	ND4	11	---
11,087	F110L	ND4	39	Branching point within L0bB1
11,172	N138S	ND4	34	Branching point at L0c2
11,390C	G211R	ND4	38	Terminal branch
11,870G	P371A	ND4	39	Terminal branch
11,988	M410T	ND4	38	Terminal branch
12,362	T9I	ND5	7	---
12,469	I45V	ND5	17	---
12,752C	Q139P	ND5	37	Terminal branch
12,952	A206T	ND5	1	---
12,967C	T211P	ND5	4	---
13,129	P265S	ND5	36	Branching point within L0bB2
13,276	M314V	ND5	31	Branching point between L0a and L0b
13,466C	S377T	ND5	30	Terminal branch
13,708	A458T	ND5	13	---
13,759	A475T	ND5	1	---

Table 5.54 Continued...

Nucleotide position	Amino acid change	Gene	Conservation index	Position in global L0-specific haplogroup network
13,819	F495L	ND5	3	---
13,882G	P516A	ND5	24	---
13,928C	S531T	ND5	4	---
13,966	T544A	ND5	5	---
13,967	T544M	ND5	5	---
14,211	V155I	ND6	31	Branching point within L0bB1
14,280	S132P	ND6	20	---
14,315	S120N	ND6	8	---
14,798	F18L	Cytb	30	Terminal branch
15,002	G86S	Cytb	39	Branching point at L0c2C
15,099	I118T	Cytb	27	---
15,236	I164V	Cytb	27	---
15,312	I189T	Cytb	31	Branching point at L0aB
15,317	A191T	Cytb	36	Terminal branch
15,431	A229T	Cytb	3	---
15,479	F245L	Cytb	34	Terminal branch
15,519	L258P	Cytb	7	---
15,631C	L295F	Cytb	11	---
15,692	M316V	Cytb	29	---
15,796G	I350M	Cytb	36	Terminal branch
15,848	T368A	Cytb	5	---
15,852	I369T	Cytb	3	---
15,861A	I372N	Cytb	26	---

Nucleotide numbers are based on rCRS numbering (Andrews *et al.*, 1999). Non-synonymous substitutions classified as transversions are indicated by listing the appropriate nucleotide change after the nucleotide position. Non-synonymous substitutions with high conservation numbers (> 30) are indicated in **bold** text. Position in global L0-specific haplogroup network is indicated for non-synonymous substitutions with high conservation numbers only. The names of sub-haplogroups and groupings within sub-haplogroups correspond to the branches indicated in Figures 5.50, 5.51, 5.52 and 5.53 (refer to Section 5.10.2). Nucleotide abbreviations: A = adenine; C = cytosine; G = guanine. Amino acid abbreviations: A = alanine; C = cysteine; D = aspartic acid; E = glutamic acid; F = phenylalanine; G = glycine; H = histidine; I = isoleucine; K = lysine; L = leucine; M = methionine; N = asparagine; P = proline; Q = glutamine; R = arginine; S = serine; T = threonine; V = valine; W = tryptophan; Y = tyrosine. Other abbreviations: ATPase6 = ATP synthase F₀ subunit 6; ATPase8 = ATP synthase F₀ subunit 8; COI = cytochrome *c* oxidase subunit I; COII = cytochrome *c* oxidase subunit II; COIII = cytochrome *c* oxidase subunit III; Cytb = cytochrome *b*; ND1 = NADH dehydrogenase subunit 1; ND2 = NADH dehydrogenase subunit 2; ND3 = NADH dehydrogenase subunit 3; ND4 = NADH dehydrogenase subunit 4; ND4L = NADH dehydrogenase subunit 4L; ND5 = NADH dehydrogenase subunit 5; ND6 = NADH dehydrogenase subunit 6.

The majority of the non-synonymous substitutions were located in the ND genes, i.e. 51 substitutions (ND1 = 11, ND2 = 7, ND3 = 3, ND4 = 10, ND4L = 2, ND5 = 15 and ND6 = 3) compared to 18 in the CO genes (COI = 8, COII = 5 and COIII = 5) and 15 in the Cytb gene. The ATPase6 and ATPase8 genes both contained six non-synonymous substitutions, and shared an additional two non-synonymous substitutions. This supports previous observations by Mouton in 2003 and Kivisild *et al.* in 2006 who also observed the majority of non-synonymous substitutions in ND genes.

The non-synonymous substitutions at sites with the highest conservation, i.e. conservation indices between 30 and 39, were observed in both the internal branches and terminal branches of the global L0-specific haplogroup network. The majority of the non-synonymous substitutions, 31/44, were located in the terminal branches of the global L0-specific haplogroup network. This indicates that these non-synonymous substitutions are possibly slightly deleterious, are recent, are not yet fixed within the population, and will likely be removed by purifying selection (Ruiz-Pesini *et al.*, 2004; Ruiz-Pesini *et al.*, 2007). The remaining 13 non-synonymous substitutions were positioned at branching points within the network. These branching points were located close to terminal branches within defined sub-haplogroup branches (e.g. 7,673, 9,178, 9,438, 11,087, 13,129 and 14,211), at positions which divided sub-haplogroups into smaller groups (e.g. 3,866, 4,812, 11,172, 15,002 and 15,312) and at positions where sub-haplogroups segregated from one another (e.g. 4,232 and 13,276).

Of the highly conserved non-synonymous substitutions located in the terminal branches, only three were observed in mtDNA sequences generated in the current investigation:

- a. The first is a non-synonymous transition at nucleotide position 4,771 in the ND2 gene, observed in individual KS_CGR_120, where a non-polar alanine was changed to a non-polar valine (Campbell, 1995), the former amino acid being conserved in 37 different species (refer to Section 4.14.6.2, Table 4.7). Although this change occurred at a highly conserved amino acid position, it was not expected to have a functional implication as the biochemical properties of the two amino acids are similar.
- b. The second is a non-synonymous transversion and was identified within individual KS_CGR_105 at position 8,214 in the COII gene. It was conserved in 37 species (refer to Section 4.14.6.2, Table 4.7) and was a change of a non-polar valine to a polar aspartic acid (Campbell, 1995). Aspartic acid contains an additional carboxyl group within its side chain which is negatively charged at neutral pH (Campbell, 1995). This non-synonymous substitution involves the replacement of one amino acid with another with different biochemical properties (Campbell, 1995). The COII protein is a subunit of cytochrome c oxidase which is involved in the mitochondrial electron transport chain which produces energy in the form of ATP (refer to Section 2.3). This alteration therefore has the potential to be deleterious, leading to problems with the secondary structure or folding of the mature mitochondrial protein, or advantageous, providing the means to adapt when subjected to selective

pressure. Functional studies will however need to be performed to determine if the alteration will have a detrimental or advantageous effect.

- c. The third is a non-synonymous transversion which was observed within individual KS_CGR_044 at position 13,466 in the ND5 gene. It was conserved in 30 species (refer to Section 4.14.6.2, Table 4.7) and represents a change from a serine to a threonine. Given the neutral side chains present in both amino acids (Campbell, 1995), it was not expected that the change would have functional implications.

The remaining highly conserved non-synonymous substitutions were observed at branching points within the global L0-specific haplogroup network. In the ND5 gene at nucleotide 13,276 where L0 segregates from L1 (Ruiz-Pesini *et al.*, 2007), a transition was observed. The amino acid encoded by the nucleotide located at this position was conserved in 31 species (refer to Section 4.14.6.2, Table 4.7) and the alteration gave rise to a change from a non-polar methionine to a non-polar valine (Campbell, 1995). It was assumed that this alteration did not have functional consequences as the two amino acids share the same biochemical properties (Campbell, 1995). In addition, this alteration was also one of the substitutions used to define the L0 haplogroup (Ruiz-Pesini *et al.*, 2007). The alteration identified in the ND1 gene at nucleotide 4,232 resulted in a non-polar isoleucine changing to a polar threonine (Campbell, 1995). Due to the fact that the amino acid at this position was conserved in 31 species (refer to Section 4.14.6.2, Table 4.7), and the biochemical properties of the two amino acids are different (Campbell, 1995), it is possible that this alteration could be deleterious. This alteration was located at the split between L0b and L0aB and its presence at this position in the network could be due to the occurrence of a neutral mutation that was fixed by the effect of random genetic drift. It is therefore expected that the alteration will not have detrimental functional implications. Although not likely, it can be speculated that an adaptive function exists for the alteration as it was retained and is subsequently observed in individuals from two related haplogroups. In support of this, it has been argued by Wallace (2005) that ancient non-synonymous substitutions affecting highly conserved amino acids, due to their retention in populations over tens of thousands of years, must have been adaptive and conferred some advantage to the population in the environment in which the population currently resides.

A non-synonymous transition at position 3,866 in the ND1 gene divided L0c1B into three groups. The amino acid encoded by the nucleotide at this position was conserved in 33 species (refer to Section 4.14.6.2, Table 4.7), and the transition resulted in a change

from a non-polar isoleucine to a polar threonine (Campbell, 1995). It has been suggested that this alteration contributes to the phenotypic expression of the G11778A mutation in a single Chinese family affected by LHON (Qian *et al.*, 2005). However, this alteration has been reported in a number of individuals of African descent (Ingman *et al.*, 2000; Maca-Meyer *et al.*, 2003; Mishmar *et al.*, 2003; Kivisild *et al.*, 2006) and European descent (Herrnstadt *et al.*, 2002a) suggesting that it is not likely to be pathogenic. Non-synonymous transitions at position 4,812 in the ND2 gene and position 15,312 in the Cytb gene identified the B branch of L0a. Both amino acids encoded by nucleotides at these positions had a conservation index of 31 (refer to Section 4.14.6.2, Table 4.7). The alteration at the first position changed a non-polar valine to a non-polar isoleucine. As these two amino acids are similar (Campbell, 1995), no functional effects are expected. The result of the alteration at the second position is a change from a non-polar isoleucine to a polar threonine (Campbell, 1995). The biochemical properties of the two amino acids are different, and it is therefore possible that this alteration could be deleterious. However, due to the position of the two alterations in the global L0-specific haplogroup network and their presence in individuals of East African origin, these alterations may identify a new sub-haplogroup located between L0a and L0b (refer to Sections 5.9 and 5.10) as they were both associated with a specific group of individuals, i.e. the Sandawe ethnic group from Tanzania (Gonder *et al.*, 2007).

A non-synonymous transition at position 11,172 in the ND4 gene identified the L0c2 sub-haplogroup. The amino acid encoded by the nucleotide at this position was conserved in 34 species (refer to Section 4.14.6.2, Table 4.7), and the alteration resulted in a change from an asparagine to a serine, both of which are polar (Campbell, 1995). The biochemical properties of these two amino acids are similar, and as this alteration identified a sub-haplogroup, it is not considered to be detrimental. Another non-synonymous transition was observed at position 15,002 in the Cytb gene and was located in branch C of the L0c2 sub-haplogroup. This substitution, conserved in 39 species (refer to Section 4.14.6.2, Table 4.7), resulted in the replacement of a polar glycine with a polar serine (Campbell, 1995). Although the amino acid encoded by the nucleotide at the position under discussion is highly conserved, the change is not expected to have functional implications as the two amino acids have similar biochemical properties.

The last group of non-synonymous substitutions affecting branching points were substitutions located at points close to or adjacent to terminal branches within defined sub-haplogroup branches. A non-synonymous transition was observed in individual

KS_CGR_106 in the COII gene at position 7,673. The amino acid encoded by the nucleotide at this position was conserved in 34 species (refer to Section 4.14.6.2, Table 4.7) and the alteration resulted in a change from a non-polar isoleucine to a non-polar valine (Campbell, 1995). The biochemical character of the two amino acids is similar and therefore no functional effect is expected to be associated with this change. A non-synonymous transition was observed in the ATPase6 gene at position 9,178. The conservation index for the wild type amino acid encoded by this nucleotide was 32 (refer to Section 4.14.6.2, Table 4.7), and the alteration resulted in a change from a non-polar valine to a non-polar methionine (Campbell, 1995). Due to the similarities between the two amino acids, no functional effect is expected. A non-synonymous transition was observed in individuals KS_CGR_004, KS_CGR_077, KS_CGR_105 and KS_CGR_121 in the COIII gene at nucleotide 9,438. This transition was located in branch B, group 2 of the L0b sub-haplogroup and was the branching point that gave rise to three groupings within group 2. The conservation index for this amino acid site affected by this non-synonymous transition was 36 (refer to Section 4.14.6.2, Table 4.7) and this alteration effected a change from a polar glycine to a polar serine (Campbell, 1995). It has been suggested that this alteration is of pathogenic importance in LHON (Johns and Neufeld, 1993). However, this alteration has been observed at significant frequencies in apparently healthy African populations and apparently healthy Cuban populations of mixed ancestry (Brown *et al.*, 1994). In contrast, this alteration has rarely been observed in apparently healthy Caucasian populations (Johns, 1994). This suggests that the alteration could be associated with specific African haplotypes and that the alteration will have no pathogenic consequences for individuals carrying these African haplotypes. The possibility that the alteration could be pathogenic in Caucasian individuals harbouring specific mtDNA haplotypes of European origin is however not excluded due to the rarity of the alteration in Caucasian populations. It can thus be argued that the mitochondrial haplogroup on which the alteration is presented, i.e. African versus European, will play a role in determining the pathogenicity of the alteration. A non-pathogenic role for this alteration in African haplogroups, i.e. no detrimental functional effect, is further supported by the biochemical similarities between the two amino acids, as well as by the presence of this alteration in the Khoi-San consensus sequence (refer to Section 5.13, Figure 5.61).

Two non-synonymous transitions were observed in individuals KS_CGR_048 and KS_CGR_060 in branch B, group 1 of the L0b sub-haplogroup. These transitions were detected at position 11,087 in the ND4 gene, and at nucleotide 14,211 in the ND6 gene. The first position at 11,087 encodes an amino acid with a conservation index of 39 (refer to

Section 4.14.6.2, Table 4.7) and the alteration causes a change from a non-polar phenylalanine to a non-polar leucine (Campbell, 1995). The amino acid encoded by the nucleotide at position 14,211 was conserved in 31 species (refer to Section 4.14.6.2, Table 4.7) and the alteration at this nucleotide resulted in a change from a non-polar valine to a non-polar isoleucine (Campbell, 1995). At both positions, the alterations observed are not expected to have functional implications, as both the wild type amino acid and the mutant amino acid share biochemical similarities. It is however possible, especially when the conservation index is very high (39), that the amino acid replacement results in various effects on other factors such as protein folding, and in the current investigation these effects can not be excluded as these factors were not studied. The last non-synonymous transition under discussion was located in branch B, group 2 of L0b. This transition was detected at nucleotide 13,129 in the ND5 gene and the amino acid encoded by the nucleotide at this position had a conservation index of 36 (refer to Section 4.14.6.2, Table 4.7). This alteration resulted in a change from a non-polar proline to a polar serine (Campbell, 1995). Although the conservation index of the amino acid encoded by this position is high, and the amino acid change does not involve amino acids of similar biochemical profiles, it is not anticipated that the change will be deleterious as it is one of the alterations used to define branch B, group 2 of sub-haplogroup L0b. It is more likely that the presence of this alteration in the branch is the result of a neutral mutation that became fixed due to random genetic drift.

For the non-synonymous substitutions within internal branches that resulted in an amino acid being replaced by another amino acid with different biochemical properties, it can be suggested that these non-synonymous substitutions have arisen through the influence of adaptive selection (Ruiz-Pesini *et al.*, 2004). This becomes especially likely in the current investigation where the non-synonymous substitutions are observed in specific African haplogroups or sub-haplogroups, thereby suggesting that divergence between haplogroups observed on different continents may have occurred in order to adapt to specific environmental challenges. As mentioned previously, the retention of non-synonymous substitutions that became established in lineages through genetic drift and thus affected highly conserved amino acid positions, likely does not substantially reduce reproductive fitness and possibly confers some adaptive advantage to populations in specific environmental conditions (Wallace, 1994; Wallace, 2005). Functional studies for all these non-synonymous substitutions will however have to be undertaken in the future to determine their relevance in mitochondrial activity.

5.11.6 Coalescent date estimates

The coalescence times, i.e. the time at which two lineages last shared a common ancestor (Page and Holmes, 1998), of the L0 haplogroup, sub-haplogroups and lineages within sub-haplogroups were estimated as described in Section 4.14.7 by using the global L0-specific haplogroup network constructed in Section 5.10.2. The number of nucleotide substitutions in the branches of the global L0-specific haplogroup network was calculated for the L0 haplogroup, each individual sub-haplogroup, and for each lineage within a sub-haplogroup, up to the point of divergence from a common ancestor. Only nucleotide substitutions in the coding region (bp 577 to 16,023) in genes other than the tRNA or rRNA genes were counted. The coding region was used to determine TMRCA and not the control region due to the high mutation rate in the control region which is subject to saturation due to excess homoplasy (Tamura and Nei, 1993), the high probability of reversions in the control region (Mishmar *et al.*, 2003), the rate heterogeneity present in the control region, i.e. alterations not being randomly distributed across the length of the control region (Excoffier and Yang, 1999; Meyer *et al.*, 1999), and the fact that several equally likely gene trees can be inferred from control region sequences, especially when large numbers of sequences are analysed (Maddison *et al.*, 1992; Ballard and Whitlock, 2004). The number of nucleotide substitutions was divided by the number of individuals used in the construction of the section of the network representing the specific sub-haplogroup or lineage within a sub-haplogroup. The result was subsequently converted to time and standard deviations were calculated as described in Section 4.14.7 to obtain a time frame for the age of the MRCA. Coalescent times were calculated for all branches representing sub-haplogroups and lineages within sub-haplogroups in the global L0-specific haplogroup network (refer to Section 5.10.2). The estimated TMRCA's are however only presented for those branches representing sub-haplogroups and lineages within sub-haplogroups that were present in both the global L0-specific haplogroup network and the reduced global maximum parsimony tree (refer to Section 5.12, Figure 5.60). In the case of the L0aC branch, two groups are present in this branch in the global L0-specific haplogroup network. However, in the reduced global maximum parsimony tree, these two branches, L0aC1 and L0aC2, are only subdivisions within one of two groups (refer to Section 5.12, Figure 5.60). The TMRCA's calculated for these two branches are therefore not presented. The addition of further L0aC sequences to the global L0-specific haplogroup network is necessary before the TMRCA can be determined for lineages within the L0aC branch.

The estimated TMRCAs for the L0 haplogroup, L0 sub-haplogroups and L0 lineages are presented in Table 5.55. TMRCAs indicated in green for a number of sub-haplogroups or lineages within sub-haplogroups are estimates based on four or fewer individuals.

Table 5.55 TMRCAs estimated for haplogroups, sub-haplogroups and lineages within L0

Haplogroup designation	Description of location in global L0-specific haplogroup network	Estimated TMRCA (YBP)	Number of individuals
L0	L0	126,495 ± 2,549	100
L0a	L0a	124,777 ± 6,767	14
L0aA	L0a (branch A)	124,777 ± 6,767	14
L0aA1	L0a (branch A, group 1)	10,276 ± 2,966	6
L0aA2	L0a (branch A, group 2)	7,707 ± 2,569	6
L0aB	L0a (branch B)	104,471 ± 13,376	3
L0aB1	L0a (branch B, group 1)	10,276 ± 7,266	2
L0aB2	L0a (branch B, group 2)	28,258 ± 8,520	1
L0aC	L0a (branch C)	102,758 ± 13,266	3
L0b	L0b	111,848 ± 3,839	39
L0bA	L0b (branch A)	86,060 ± 10,514	4
L0bA1	L0b (branch A, group 1)	51,379 ± 11,489	2
L0bA2	L0b (branch A, group 2)	38,534 ± 9,949	2
L0bB	L0b (branch B)	69,435 ± 3,193	35
L0bB1	L0b (branch B, group 1)	54,975 ± 5,315	10
L0bB1a	L0b (branch B, group 1, part a)	30,827 ± 12,585	1
L0bB1b	L0b (branch B, group 1, part b)	51,950 ± 5,446	9
L0bB1b1	L0b (branch B, group 1, part b, section 1)	46,241 ± 10,899	2
L0bB1b2	L0b (branch B, group 1, part b, section 2)	35,109 ± 5,483	6
L0bB1b3	L0b (branch B, group 1, part b, section 3)	25,689 ± 11,489	1
L0bB2	L0b (branch B, group 2)	42,645 ± 3,310	20
L0bB2a	L0b (branch B, group 2, part a)	25,119 ± 2,678	18
L0bB2a1	L0b (branch B, group 2, part a, section 1)	15,414 ± 8,899	1
L0bB2a2	L0b (branch B, group 2, part a, section 2)	9,419 ± 2,840	6
L0bB2a3	L0b (branch B, group 2, part a, section 3)	9,342 ± 2,089	11
L0bB2b	L0b (branch B, group 2, part b)	23,121 ± 7,707	2
L0bB3	L0b (branch B, group 3)	25,689 ± 5,138	5
L0bB3a	L0b (branch B, group 3, part a)	15,414 ± 5,138	3
L0bB3b	L0b (branch B, group 3, part b)	15,414 ± 6,293	2
L0c	L0c	79,825 ± 3,163	41
L0c1	L0c1	25,387 ± 2,770	17
L0c1A	L0c1 (branch A)	25,119 ± 3,787	9
L0c1B	L0c1 (branch B)	14,771 ± 3,398	8

Table 5.55 Continued...

Haplogroup designation	Description of location in global L0-specific haplogroup network	Estimated TMRCA (YBP)	Number of individuals
L0c2	L0c2	49,024 ± 3,240	24
L0c2ABCE	L0c2 (branch ABCE)	28,901 ± 3,046	16
L0c2ACE	L0c2 (branch ACE)	32,112 ± 6,422	4
L0c2A	L0c2 (branch A)	41,103 ± 14,532	1
L0c2B	L0c2 (branch B)	31,256 ± 3,658	12
L0c2B1	L0c2 (branch B, group 1)	21,579 ± 4,709	5
L0c2B2	L0c2 (branch B, group 2)	20,552 ± 3,884	7
L0c2C	L0c2 (branch C)	38,534 ± 9,949	2
L0c2D	L0c2 (branch D)	22,478 ± 3,800	8
L0c2D1	L0c2 (branch D, group 1)	3,425 ± 2,422	3
L0c2D2	L0c2 (branch D, group 2)	2,569 ± 2,569	2
L0c2E	L0c2 (branch E)	10,276 ± 7,266	1

TMRCA = time to most recent common ancestor; YBP = years before present. Sub-haplogroups and lineages within sub-haplogroups are as indicated in Figures 5.50, 5.51, 5.52 and 5.53. Sub-haplogroups or lineages within sub-haplogroups where TMRCA were calculated using four or fewer individuals, are indicated in green.

Generally, the TMRCA values were as expected, i.e. the ages of sub-groups were younger than that of the group from which they diverged. This was however not the case for L0c2ABCE and L0c2ACE. The fact that L0c2ACE appears more ancient than L0c2ABCE, can be ascribed to the use of only four sequences to calculate the TMRCA for L0c2ACE. This TMRCA estimate is likely not precise. This observation is further supported by the larger standard deviation determined for L0c2ACE compared to L0c2ABCE. The TMRCA determined for L0c2C was also older than that determined for L0c2ACE and again this is likely due to the small sample sizes used in the calculations.

The TMRCA of L0a is not equal to the combined estimate of the TMRCA's calculated for L0aA, L0aB and L0aC. The reason for this is that these three lineages do not form a monophyletic clade, but are three separate branches within the L0 branch of the global phylogenetic trees (refer to Section 5.12, Figure 5.60). These three branches were classified as L0a in the current investigation based on shared characteristics as described by the classification scheme used (refer to Section 4.9). To avoid confusion, this nomenclature was retained throughout the current investigation, i.e. reference is made to L0aA, L0aB and L0aC even after constructing the revised classification scheme in which L0aB and L0aC are re-classified as new sub-haplogroups named L0d and L0e (refer to Section 5.12, Table 5.60).

As mentioned previously, L0aB represents a new branch located between L0a and L0b within the global L0-specific haplogroup network. This branch shares characteristics of both the L0a and the L0b haplogroups. This branch was first described by Gonder *et al.* in 2007 and named L0dTZ as it belongs to the L0d haplogroup according to the alternative classification scheme (refer to Section 3.3.1.1). The L0dTZ branch is specific to East African Khoi-San speakers and consists of Sandawe individuals from Tanzania. In the current investigation, when using the alternative classification scheme (refer to Section 3.3.1.1), the L0dTZ branch is referred to as L0d3 to distinguish it from L0d1 and L0d2, since at present no designation has been given to this branch in the alternative classification scheme (refer to Section 3.3.1.1).

As stated previously (refer to Section 5.9), two new branches were identified in the L0-specific haplogroup network after the addition of 46 L0 sequences. These two branches, namely L0bB1b1 and L0bB2a1, represent two previously unidentified lineages in the Southern African Khoi-San population. The TMRCA of these two lineages were determined as $46,241 \pm 10,899$ YBP for L0bB1b1, and $15,414 \pm 8,899$ YBP for L0bB2a1. The TMRCA calculated for these two lineages were based on less than four individuals, and is therefore not likely to be precise. This lack of resolution in the calculation of the TMRCA for these two previously unidentified lineages can be attributed to small sample sizes. However, the identification of two previously undescribed lineages with a difference in age of ca. 30,000 years provides evidence for the prolonged presence of Khoi-San speaking populations in southern Africa, and suggests that the extent of the genetic diversity present in the Khoi-San has not been completely revealed.

The estimates of TMRCA for the different sub-haplogroups and lineages within sub-haplogroups are influenced by the way in which the sequences used to calculate the TMRCA are sampled. A small sample of sequences might capture only some of the variation and while it is likely that the most common sub-haplogroups will be included, it is possible that rare sub-haplogroups will be excluded (Torrioni *et al.*, 2001). As mentioned previously, the estimated TMRCA for a number of sub-haplogroups, or lineages within sub-haplogroups, were calculated based on very small sample sizes ranging from one to four individuals. The remaining TMRCA were determined based on between five and 100 individuals. It is therefore likely that the TMRCA for a number of sub-haplogroups or lineages in sub-haplogroups may have been underestimated due to the small sample size (Mishmar *et al.*, 2003). Another factor that may cause the underestimation of divergence

times is the presence of recurrent alterations in the global L0-specific haplogroup network as it is unlikely that all recurrent alterations can be reconstructed at the time depths observed in African populations (Salas *et al.*, 2002). This is due to the high nucleotide substitution rate which causes saturation of nucleotide sites over time.

The estimates of TMRCA is further affected by subdivision within populations, purifying selection and sex-specific demographic processes such as the type of mating system and sex-bias in rates of migration (Wilder *et al.*, 2004a). All of these factors could play a role in the results generated in the current investigation. The global L0-specific haplogroup network was constructed using L0 sequences from various ethnic groups and different populations, and therefore does not represent a single ethnic group or population. In Section 5.11.5, the effect of selection was investigated and it was demonstrated that the null hypothesis of neutral evolution was rejected for the L0 haplogroup and that purifying selection has played a role in shaping the mtDNA diversity observed in haplogroup L0. The influence of adaptive selection can however not be excluded as it was demonstrated in Section 5.11.3 that dataset 8a consisting of L0 sequences exhibited a significant departure from neutrality when using Fu and Li's D^* and F^* statistics which could be explained by either a population expansion event or adaptive selection. However, based on results presented in Section 5.11.4, it is more likely that the observed deviation from neutrality in this instance is the result of population expansion and not adaptive selection. The maternal inheritance of mtDNA (Giles *et al.*, 1980) and cultural practises such as patrilocality, where a woman after marriage moves from her birthplace to that of her husband (Fox, 1983), influence the level of mtDNA diversity in and between populations and thus also affect the estimated TMRCA. The effects of recent selection, population expansion or population structure on estimates of TMRCA for populations can however be contained in the large evolutionary variances associated with mutation and random genetic drift over an extended period of time (Harding *et al.*, 1997).

The TMRCA estimates presented in Table 5.55 were compared to TMRCA's determined in other investigations. A list of TMRCA's determined in other investigations is presented in Table 5.56.

Table 5.56 Estimates of TMRCA of haplogroups and sub-haplogroups in other investigations

Haplogroup (Wallace, 2004)	Alternative haplogroup	Estimated TMRCA (YBP)	Type of data used	Reference
Human mtDNA	Human mtDNA	179,000 to 217,000	Coding region mtDNA sequence data	Mishmar <i>et al.</i> , 2003
	Human mtDNA	161,800 to 226,800	Coding region mtDNA sequence data	Gonder <i>et al.</i> , 2007
	Human mtDNA	143,000	Complete mtDNA sequence data	Horai <i>et al.</i> , 1995
	Human mtDNA	138,000 to 182,000	Coding region mtDNA sequence data	Kivisild <i>et al.</i> , 2006
	Human mtDNA	122,100 to 230,900	Khoi-San population, 25 individuals, COIII gene sequence data	Wilder <i>et al.</i> , 2004a
	Human mtDNA	121,500 to 221,500	Coding region mtDNA sequence data	Ingman <i>et al.</i> , 2000
	Human mtDNA	111,000 to 148,000	Control region sequence data (16,090 - 16,365)	Watson <i>et al.</i> , 1997
African mtDNA	African mtDNA	125,500 to 165,500	High resolution RFLP analysis and CR sequencing	Chen <i>et al.</i> , 2000
	African mtDNA	101,000 to 133,000	High resolution RFLP analysis	Chen <i>et al.</i> , 1995
L	L	122,800 to 161,800	High resolution RFLP analysis and CR sequencing	Chen <i>et al.</i> , 2000
	L	98,000 to 130,000	High resolution RFLP analysis	Chen <i>et al.</i> , 1995
L0	L0	128,345 to 138,825	Coding region mtDNA sequence data	Mouton, 2003
	L0	125,000 to 159,000	Coding region mtDNA sequence data	Mishmar <i>et al.</i> , 2003
	L0	121,300 to 171,500	Coding region mtDNA sequence data	Gonder <i>et al.</i> , 2007
	L0	92,000 to 124,000	Coding region mtDNA sequence data	Kivisild <i>et al.</i> , 2006
L0a	L0k	48,000 to 56,000	Control region sequence data (16,090 - 16,365)	Watson <i>et al.</i> , 1997
L0aA	L0k (SAK)	51,200 to 90,600	Coding region sequence data	Gonder <i>et al.</i> , 2007
	L0k	7,707 to 15,413	Coding region mtDNA sequence data	Mouton, 2003
L0aB	L0d (Tanzania)	12,800 to 48,400	Coding region mtDNA sequence data	Gonder <i>et al.</i> , 2007
	L0d (Tanzania)	9,400 to 50,700	Control region sequence data (HVS1 + HVS2)	Tishkoff <i>et al.</i> , 2007
L0aB + L0b	L0d (Tanzanian + SAK)	33,700 to 90,100	Control region sequence data (HVS1 + HVS2)	Tishkoff <i>et al.</i> , 2007
L0aC	L0f	85,500 to 104,300	Coding region mtDNA sequence data	Gonder <i>et al.</i> , 2007

Table 5.56 Continued...

Haplogroup (Wallace, 2004)	Alternative haplogroup	Estimated TMRCA (YBP)	Type of data used	Reference
L0b	L0d	85,800 to 126,200	Coding region mtDNA sequence data	Gonder <i>et al.</i> , 2007
	L0d San	71,500 to 109,300	Coding region mtDNA sequence data	Gonder <i>et al.</i> , 2007
	L0d	51,000 to 73,000	Coding region mtDNA sequence data	Kivisild <i>et al.</i> , 2006
	L0d	36,150 to 63,050	HVS1 sequence data, RFLPs	Salas <i>et al.</i> , 2002
	L0d (SAK)	18,500 to 145,000	Control region sequence data (HVS1 + HVS2)	Tishkoff <i>et al.</i> , 2007
L0bB	L0d1	42,341 to 52,197	Coding region mtDNA sequence data	Mouton, 2003
L0c	L0a	48,900 to 60,300	Coding region mtDNA sequence data	Gonder <i>et al.</i> , 2007
	L0a	31,000 to 53,000	Coding region mtDNA sequence data	Kivisild <i>et al.</i> , 2006
	L0a	24,100 to 56,600	HVS1 sequence data, RFLPs	Salas <i>et al.</i> , 2002
L0c1	L0a1	16,750 to 49,950	HVS1 sequence data, RFLPs	Salas <i>et al.</i> , 2002
	L0a1 (Gambia)	3,800 to 9,000	HVS1 and HVS2 sequence data, RFLPs	Rosa <i>et al.</i> , 2004
L0c1B	L0a1a	9,400 to 45,300	HVS1 sequence data, RFLPs	Salas <i>et al.</i> , 2002
L0c2	L0a2	23,000 to 41,000	Coding region mtDNA sequence data	Kivisild <i>et al.</i> , 2006
	L0a2	4,650 to 11,950	HVS1 sequence data, RFLPs	Salas <i>et al.</i> , 2002
L0c2B	L0a2a	14,000 to 28,000	Coding region mtDNA sequence data	Kivisild <i>et al.</i> , 2006
L0c2D	L0a2b	1,000 to 5,000	Coding region mtDNA sequence data	Kivisild <i>et al.</i> , 2006

mtDNA = mitochondrial DNA; TMRCA = time to most recent common ancestor; YBP = years before present; SAK = Southern African Khoi-San; COIII = cytochrome c oxidase subunit 3; RFLP = restriction fragment length polymorphism; HVS1 = hypervariable segment 1; HVS2 = hypervariable segment 2. Haplogroup (Wallace, 2004) is the haplogroup classification according to the classification scheme presented in Section 4.9 and alternative haplogroup is the haplogroup designation according to the alternative classification scheme presented in Section 3.3.1.1 which corresponds to the published haplogroup assignment reported in the accompanying reference. TMRCA estimates indicated in red are similar to the TMRCA estimates determined in the current investigation (refer to Table 5.55).

The estimated TMRCA for the L0 haplogroup (123,946 to 129,044 YBP) determined in the current investigation is within the range of a minimum of 92,000 to a maximum of 171,500 YBP previously reported in other genetic investigations of mtDNA (refer to Table 5.56). The TMRCA for the L0 haplogroup determined in the current investigation compares favourably with an early estimate of the paleontological date of transition from archaic *Homo sapiens* to anatomically modern *Homo sapiens* which occurred at 120,000 to 130,000 YBP (Watson *et al.*, 1997). The estimated TMRCA for the L0 haplogroup is however more recent than the current estimates of dates determined radioisotopically for fossil evidence representing the emergence of modern *Homo sapiens* in Ethiopia in East Africa, i.e. 154,000 to 160,000 YBP for the Middle Awash (Clark *et al.*, 2003; White *et al.*,

2003) and 190,000 to 200,000 YBP for the Kibish (McDougall *et al.*, 2005). Keeping these estimates of the age of modern humans in mind, it can be speculated that the L0 haplogroup likely arose after the emergence of anatomically modern *Homo sapiens* and if the mtDNA of archaic *Homo sapiens* could be sampled and investigated, it would be an ancestral sequence to the L0 mtDNA sequences observed today. In addition, it is proposed that the ancestral sequence will be genetically similar to the L5 haplogroup which is observed closer to the outgroup than the L0 haplogroup branch in the global neighbour-joining and maximum parsimony trees (refer to Section 5.9, Figures 5.42 and 5.45). In addition, the congruity of the estimated TMRCA for the L0 haplogroup with the reported paleontological dates supports the hypothesis that modern *Homo sapiens* populations share a common and recent origin from an ancestral population from Africa, with data from the current investigation lending support to the “Out of Africa” theory of modern human population expansion and continental colonisation.

In comparisons of the estimated TMRCA in the current investigation with TMRCA calculated in other investigations, the TMRCA determined in the current investigation for L0b, L0c1 and L0c1B were broadly consistent with the previously published TMRCA (refer to Table 5.56). In the remaining comparisons of the estimated TMRCA with those calculated in other investigations, the TMRCA determined in the current investigation were however more ancient. This was the case for L0a, L0aA, L0aB, L0aC, L0bB, L0c, L0c2, L0c2B and L0c2D.

The increased TMRCA values for L0aA, L0bB, L0c, L0c2, L0c2B and L0c2D is due to the use of a greater number of L0 sequences to calculate the TMRCA for each of these sub-haplogroups or lineages within sub-haplogroups compared to previous investigations which also made use of coding region data (refer to Table 5.56). The use of more L0 sequences in the estimation of TMRCA values for the L0aA, L0bB, L0c, L0c2, L0c2B and L0c2D groups also facilitated the calculation of reduced standard deviation values which in turn generated a narrower range for the TMRCA, thus making the estimate more precise. In the case of L0aB and L0aC, smaller sample sizes were used to calculate the TMRCA values than in previous investigations (refer to Table 5.56). The use of small sample sizes to calculate the TMRCA values, led to increased standard deviation values for these smaller samples, which extends the range of the TMRCA thereby making the estimate less precise. The more ancient values observed for L0aB and L0aC are therefore not likely to be precise and should be revised after the addition of more sequences to the L0aB and L0aC branches. The different types of data used to determine the TMRCA also

influence the estimates obtained. In the current investigation, coding region sequence data was used. In other investigations such as those performed by Watson *et al.* (1997), Salas *et al.* (2002), Rosa *et al.* (2004) and Tishkoff *et al.* (2007), the control region which includes HVS1 was used. It has been demonstrated that TMRCA estimates based on HVS1 data are decreased due to the presence of recurrent alterations in this region (Salas *et al.*, 2002). In the case of L0a, the increased value of the TMRCA can therefore be attributed to the use of coding region data (current investigation) compared to control region data (refer to Table 5.56) as the presence of recurrent alterations and reversions in the control region can lead to an underestimation of the TMRCA.

The various methods employed to estimate the TMRCA also affect the estimates determined. In the current investigation, all nucleotide substitutions in genes other than the tRNA and rRNA genes in the coding region were used in calculations. In the investigation performed by Kivisild *et al.* (2006), the TMRCA estimates were based on the average number of synonymous substitutions to the root of the clade, i.e. they did not use all the available sites in the coding region to calculate the TMRCA. In contrast, Tishkoff *et al.* (2007) used a Markov chain coalescent simulation with the assumption of constant population size, while Gonder *et al.* (2007) employed a penalised likelihood model to estimate TMRCA.

The method used by Kivisild *et al.* (2006) is a variation of the method used in the current investigation, and is used with the aim to reduce bias by not using all sites in the coding region, due to the possibility that slightly deleterious non-synonymous mutations may be overrepresented in the terminal branches of the network (Elson *et al.*, 2004). The method used in the current investigation, as well as that used by Kivisild *et al.* (2006), are conceptually simplistic and easy to interpret, and are therefore widely used. The use of coalescent modelling as employed by Tishkoff *et al.* (2007) however permits the incorporation of a number of population parameters in the model, but has the disadvantage that a constant population size is assumed although in practice a population is expected to encounter demographic pressures, such as bottlenecks and expansions, resulting in the violation of this assumption (Basu and Majumder, 2003). In contrast, the penalised likelihood model used by Gonder *et al.* (2007) incorporated the observation that rates of molecular evolution vary widely between lineages (Sanderson, 2002).

The TMRCA values obtained by Kivisild *et al.* (2006) for comparable sub-haplogroups or lineages were consistently more recent than the values obtained in other investigations

using other methods, thereby suggesting that their method, due to the exclusion of some sites, generated imprecise estimates. The TMRCA values obtained by Tishkoff *et al.* (2007) and Gonder *et al.* (2007) for L0aB and L0b are similar, although the TMRCA values calculated by Tishkoff *et al.* (2007) have larger standard deviation values (refer to Table 5.56). Both Tishkoff *et al.* (2007) and Gonder *et al.* (2007) calculated values for L0b that were similar to the TMRCA value determined in the current investigation (refer to Table 5.56). Therefore, although the use of “coalescent of likelihood” models are likely to generate more precise TMRCA values, the method employed in the current investigation appears to provide sufficient estimates and computationally expensive models were thus not employed.

The TMRCA of the different sub-haplogroups and lineages within sub-haplogroups, in conjunction with the current geographic distribution of these groups and lineages (refer to Section 5.10.2.2), have the potential to provide information regarding the movement of L0 sub-haplogroups and lineages within sub-haplogroups across the African continent. The L0 sub-haplogroups, as well as the TMRCA for each sub-haplogroup, are indicated in Figure 5.59 according to geographic region in Africa.

The oldest sub-haplogroups and lineages (L0a, L0aA, L0aB, L0aC and L0b) are located in southern and eastern Africa (refer to Figure 5.59). The L0aA (L0k¹) and L0b (L0d) sub-haplogroups are observed in Southern African Khoi-San speakers (Salas *et al.*, 2002), which include the !Kung, while the L0aB (L0d3) sub-haplogroup is present in eastern African Khoi-San speakers, which include the Sandawe (Gonder *et al.*, 2007). Tishkoff *et al.* (2007) investigated three Khoi-San speaking populations, two Eastern Khoi-San speaking populations, namely the Sandawe and the Hadza, and one Southern Khoi-San speaking mixed population containing !Kung and Khwe individuals. They determined that the Southern and Eastern Khoi-San speakers shared ancient mtDNA and Y chromosome haplogroups and proposed three scenarios of population history for the three Khoi-San speaking populations.

The three scenarios all assumed shared descent of mtDNA haplogroups from an ancestral Khoi-San speaking population for the Southern African Khoi-San speakers and the

¹ The haplogroup classifications presented in brackets refer to the classifications according to the alternative classification scheme (refer to Section 3.3.1.1).

Sandawe, i.e. L0aB (L0d3) and L0b (L0d1 and L0d2) diverged from a shared mtDNA haplogroup. This establishes a genetic link between the Khoi-San populations in East and southern Africa, although it does not provide information on whether the ancestral population originated in East Africa or in southern Africa.

Figure 5.59 Geographic distribution of L0 sub-haplogroups and lineages within sub-haplogroups in Africa



Sub-haplogroups and lineages within sub-haplogroups with the appropriate TMRCA values are indicated on the map of Africa according to distribution in major geographic regions in Africa. The geographic regions are indicated using colour as stipulated in the legend. The map of Africa was constructed by adapting map number 4045, revision 4 (January 2004) available from the United Nations website (<http://www.un.org/Depts/Cartographic/english/htmain.htm>, 2008).

In the current investigation, it is suggested that the ancestral population that gave rise to Khoi-San speakers originated in East Africa and members of this population subsequently dispersed to southern Africa and became genetically isolated from one another possibly due to a marked dry period in southern Africa at the height of the last glacial maximum, ca. 17,000 to 24,000 YBP, which caused the contraction of available geographic ranges (Lahr and Foley, 1998). This argument is supported by the suggestion that Khoi-San speakers were once widely distributed across Africa reaching as far as present day Ethiopia and Sudan, based on the presence of Khoi-San speakers in Tanzania (Semino *et al.*, 2002),

and an East African origin for the ancestral population is favoured due to the rich fossil record and the high levels of mtDNA diversity present in East Africa. The earliest remains of transitional modern humans, dated as early as 196,000 YBP, have been discovered in Ethiopia (Clark *et al.*, 2003; White *et al.*, 2003; McDougall *et al.*, 2005) and the earliest artefacts associated with modern humans have also been detected in eastern Africa (Foley, 1998). In addition, Stone Age technology was established in several regions in eastern Africa prior to 40,000 YBP, but was not observed in southern Africa until 22,000 YBP (Lahr, 1996; Foley, 1998). Studies of genetic diversity in Africa indicate that high levels of mtDNA diversity are detected in populations from East Africa (Watson *et al.*, 1997; Chen *et al.*, 2000; Kivisild *et al.*, 2004; Gonder *et al.*, 2007).

The observation that both of the two oldest lineages, namely L0aA and L0b are present in southern Africa, while L0b is absent in East Africa appears to contradict the argument presented in the previous paragraph. However, the argument remains plausible if the time when divergence between the two Khoi-San speaking populations occurred, is taken into consideration. Tishkoff *et al.* (2007) estimated that this split occurred between 35,000 to 50,000 YBP. Therefore, if the two populations became genetically isolated as suggested in the previous paragraph and sufficient time had elapsed since the split between populations occurred, it is possible that the L0b (L0d1 and L0d2) lineage remained in southern Africa, while becoming extinct in East Africa due to a number of possible reasons which include factors such as random genetic drift, population bottlenecks and interaction with neighbouring populations. Additional L0 sequences, obtained by extensive sampling in East Africa, are however required to test this hypothesis.

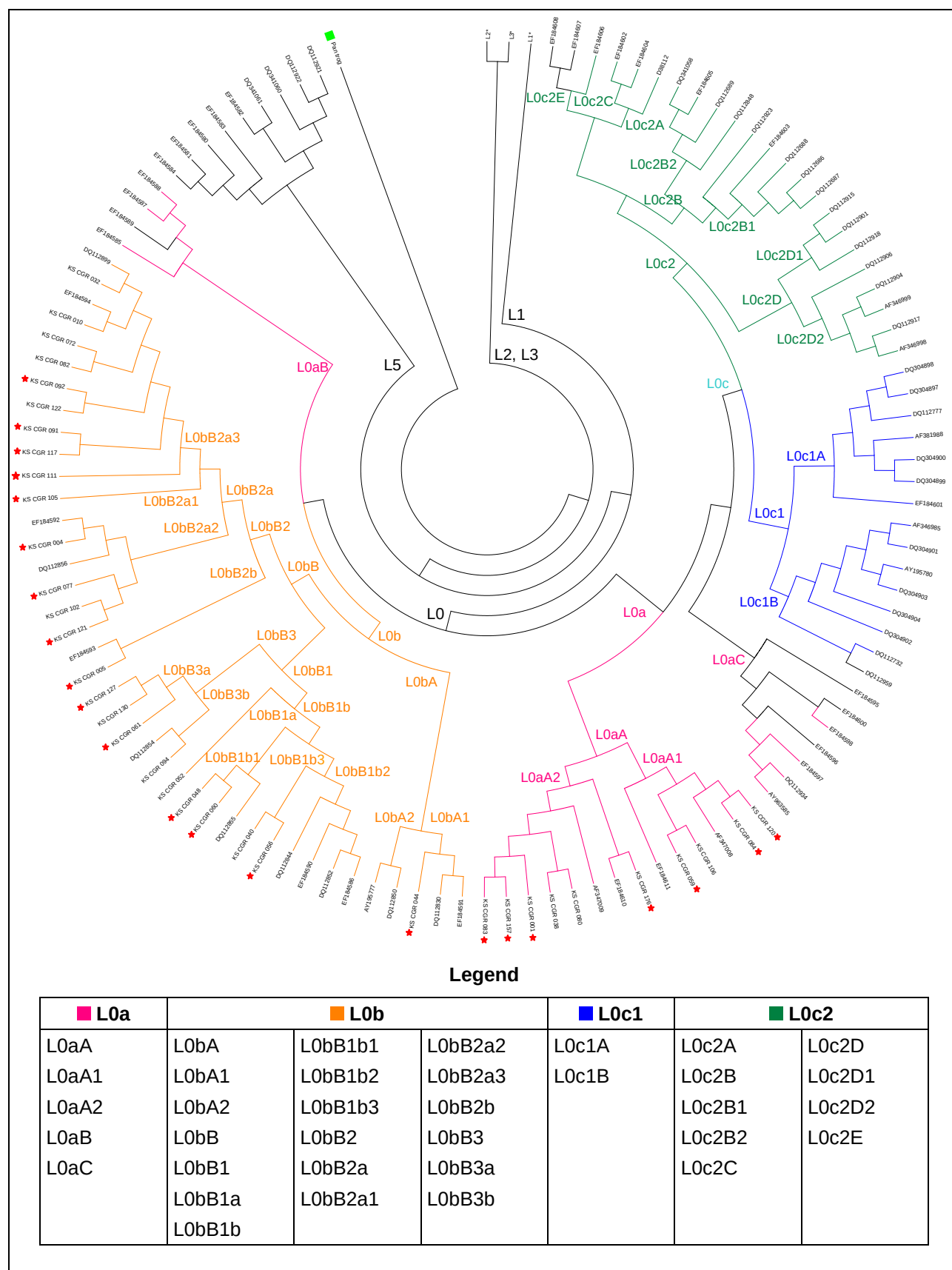
The L0c2 sub-haplogroup is located in East, Central and West Africa, while the L0c1 sub-haplogroup is localised in West Africa (refer to Figure 5.59). These two sub-haplogroups are observed in different African populations (González *et al.*, 2006), with the exception of L0c2D which appears to be confined to Pygmies in Central Africa (refer to Section 5.10.2.2). An East African origin for L0c was also previously suggested by Salas *et al.* (2002). Based on the data presented in Figure 5.59, it appears that the L0c haplogroup originated in East Africa from where it dispersed to Central and West Africa. This is supported by the observation that L0c2, the older of the two L0c sub-haplogroups, is observed in East Africa, while L0c1, the younger of the two sub-haplogroups, is observed in West Africa only, as well as the presence of L0c2B in all three regions which suggests that L0c2B dispersed from East Africa to West Africa via Central Africa.

The apparent lack of L0 lineages in North Africa is also of interest. It can be speculated that the dispersal of L0 haplogroup lineages in Africa occurred via two routes, a “southern” route whereby the L0 lineages observed in Khoi-San speaking populations dispersed from East Africa to southern Africa, and a “western” route whereby L0 lineages observed in populations other than the Khoi-San dispersed from East to West Africa via Central Africa. The ages of the L0 lineages indicate that the “southern” dispersal took place prior to the “western” dispersal and the specific routes taken within Africa for each of these two dispersal events were likely linked to climatic changes within Africa which allowed populations carrying these lineages to migrate into previously uninhabited territories, or to the expansion of other populations into East Africa, which forced the L0 haplogroup carriers into the rest of Africa. In addition, it is likely that the relatively recent expansion of Bantu-speaking populations into eastern Africa and then into southern Africa (refer to Section 3.1.3), led to the displacement or gradual assimilation of remaining Khoi-San populations carrying the L0 haplogroup (Salas *et al.*, 2002) with the result that only small, isolated Khoi-San speaking groups remained within eastern and southern Africa. An in depth analysis of climate change and population movements in Africa, correlated with accurate TMRCA dates based on additional L0 sequences for all sub-haplogroups, especially L0bA, L0aC, L0aB, L0c2C, L0c2E and L0c2A, is however required to accurately infer the dispersal of L0 sub-haplogroups and lineages within Africa.

5.12 REVISED CLASSIFICATION SCHEME FOR THE L0 HAPLOGROUP

In Section 5.9, three branches were observed in the global neighbour-joining and maximum parsimony trees that were classified as belonging to the L0a sub-haplogroup. As discussed previously (refer to Section 5.10), a revised classification scheme is required to compensate for the observation that more than one branch, in different parts of the global N-J and MP trees, is referred to with the same designation. The branching observed in the global MP tree generated using dataset 1a (refer to Section 5.9.3.1, Figure 5.45) was compared to that observed in the global L0-specific haplogroup network (refer to Section 5.10.2.1) and the new branches identified within the global L0-specific haplogroup network were superimposed upon a reduced version of the global MP tree displaying only the detail of the branching observed in the L0 haplogroup. The branches observed in the reduced global MP tree were named according to the system followed in the global L0-specific haplogroup network, i.e. branches within each sub-haplogroup were named according to the estimated TMRCA for that branch (refer to Section 5.11.6). The results of this analysis are presented in Figure 5.60.

Figure 5.60 Reduced global maximum parsimony tree based on the MP tree constructed using the All L Sequences dataset (dataset 1a)



This is a reduced version of the 50% majority rule consensus tree presented in Figure 5.45 in Section 5.9.3. The branches containing L1, L2 and L3 sequences have been reduced to focus on the L0 branch. Khoi-San sequences generated in the current investigation are indicated by a red star (★) and the outgroup is indicated by a green square (■). The major haplogroups are indicated on the figure at the relevant branching points. Sub-haplogroups and new branches within the sub-haplogroups are indicated at the relevant branching points, and by the use of different colours as listed in the legend. Sequences that could not be classified completely are indicated by black lines. Sequences that exhibit a discrepancy between their location in the tree and their haplogroup classification are indicated by dashed lines.

As previously mentioned, the sections of the global L0-specific haplogroup network indicated in Figures 5.50, 5.51, 5.52 and 5.53, as well as the complete global L0-specific haplogroup network depicted in Appendix J, represent networks and are not phylogenetic trees. Within these networks, the nucleotide alterations are listed in numerical order and the order in which the specific alterations occurred over time are not indicated. These networks are however useful for the classification of new mtDNA sequences as part of the L0 haplogroup, and to determine the sub-haplogroup designation of the new L0 sequences. As discussed in Section 5.10.2.1, new internal and terminal branches were observed after the addition of 54 L0 sequences to the global L0-specific haplogroup network. These branches represent new groups and lineages within each of the L0 sub-haplogroups and will form the basis of a revised classification scheme for the L0 haplogroup proposed in this section.

When constructing a classification scheme using nucleotide alterations, there are three types of SNPs to consider. Haplogroup defining polymorphisms occur in all mtDNA sequences within a haplogroup, but not in mtDNA sequences from other haplogroups (Herrnstadt *et al.*, 2002a). In contrast, haplogroup-specific polymorphisms have arisen only once during evolution, and occur in only a subset of mtDNA sequences that define a sub-cluster of the haplogroup, and haplogroup-associated polymorphisms occur in at least two members from one branch of the same haplogroup (Herrnstadt *et al.*, 2002a). For the purpose of a classification scheme, haplogroup defining polymorphisms and haplogroup-specific polymorphisms are the most useful, and in the current investigation it was attempted to identify and use only these polymorphisms in the scheme by adhering to specific criteria.

The revised L0 classification scheme was subsequently constructed according to specific criteria as discussed in Section 4.15, and by using the reduced global MP tree (Figure 5.60) as a guide to determine the order in which sub-haplogroups and lineages diverged from one another. The classification scheme (Wallace, 2004) presented in Section 4.9, and the global mtDNA mutational phylogeny available on MITOMAP (Ruiz-Pesini *et al.*, 2007) were used as a starting point, and the classification scheme was revised using the global L0-specific haplogroup network constructed in Section 5.10.2.

The global mtDNA mutational phylogeny (Ruiz-Pesini *et al.*, 2007) and the global L0-specific haplogroup network (current investigation) were used to identify a number of SNPs at specific nucleotide positions which defined or were specific to each

sub-haplogroup and lineage within that sub-haplogroup. All of these nucleotide positions were located in the mitochondrial coding region (bp 577 to 16,023) as this region was used to construct the global mtDNA mutational phylogeny (Ruiz-Pesini *et al.*, 2007) and the global L0-specific haplogroup network (current investigation). For the construction of a classification scheme, slowly evolving nucleotide positions are preferable, given their expected stability, over a long period of time. The revised L0 classification scheme proposed in this section is therefore based on slowly evolving nucleotide positions in the mitochondrial coding region only, as the control region evolves at a different overall rate and contains fast evolving sites (Pesole *et al.*, 1999). Nucleotide positions at which recurrent alterations or reversions were observed, were viewed as fast evolving and were therefore excluded where possible. Nucleotide positions at which recurrent alterations or reversions were observed were however included if the nucleotide position has historically been used to define macrohaplogroup L, or if no other haplogroup defining or haplogroup-specific SNP was available for classification of the sub-haplogroup or lineage within the sub-haplogroup (refer to Table 5.58).

The excluded nucleotide positions consisted of the previously identified positions where recurrent alterations or reversions occurred within the global L0-specific haplogroup network (refer to Section 5.10.2.3, Tables 5.39 and 5.42). An additional list was compiled of nucleotide positions identified in human mtDNA networks in other investigations (Herrnstadt *et al.*, 2002a; Galtier *et al.*, 2006; Kivisild *et al.*, 2006), where recurrent alterations occurred, or where the same alterations were located at the same nucleotide positions in different human mtDNA haplogroups, or where nucleotide positions were classified as hypervariable. This additional list was also used to exclude nucleotide positions and is presented in Table 5.57. However, if no other SNPs were available in the list of SNPs used to define a branch or lineage within a branch, or if a SNP located at a nucleotide position where recurrent alterations or reversions occurred was used in conjunction with another SNP, the nucleotide positions were retained in the classification scheme even if the nucleotide position was subjected to recurrent alterations or reversion events.

Table 5.57 Nucleotide positions in the mitochondrial coding region at which recurrent alterations were observed, or which have been classified as hypervariable

Nucleotide position									
593	3,434	4,793	6,182	8,251	9,824	11,467	13,106	14,470	15,327
709	3,438	4,820	6,221	8,269	9,932	11,641	13,263	14,471	15,452
750	3,547	4,824	6,253	8,392	9,947	11,719	13,359	14,560	15,514
769	3,552	4,917	6,260	8,473	9,950	11,800	13,368	14,569	15,670
930	3,591	5,004	6,261	8,572	10,084	11,812	13,435	14,570	15,758
1,018	3,594	5,046	6,413	8,616	10,373	11,914	13,590	14,687	15,775
1,438	3,666	5,147	6,446	8,618	10,398	11,944	13,650	14,766	15,784
1,503	3,693	5,231	6,671	8,697	10,400	12,007	13,708	14,767	15,785
1,598	3,705	5,237	6,680	8,701	10,454	12,172	13,709	14,769	15,812
1,719	3,796	5,319	6,719	8,790	10,586	12,236	13,734	14,783	15,884
1,811	3,915	5,417	6,734	8,860	10,589	12,308	13,759	14,798	15,885
1,888	3,918	5,426	6,752	8,964	10,598	12,346	13,886	14,905	15,924
2,158	4,025	5,442	6,755	8,994	10,685	12,354	13,928	15,043	15,925
2,217	4,048	5,460	6,827	9,055	10,790	12,372	13,934	15,099	15,927
2,332	4,104	5,471	7,028	9,117	10,810	12,414	13,966	15,110	15,930
2,352	4,185	5,581	7,055	9,254	10,819	12,477	13,980	15,110	---
2,706	4,216	5,585	7,076	9,266	10,873	12,501	14,007	15,115	---
2,768	4,561	5,656	7,256	9,449	10,907	12,630	14,034	15,217	---
3,010	4,562	5,773	7,388	9,509	10,915	12,642	14,110	15,218	---
3,027	4,580	5,821	7,424	9,540	11,002	12,705	14,148	15,236	---
3,197	4,646	5,913	7,521	9,545	11,176	12,810	14,182	15,244	---
3,308	4,688	5,999	8,020	9,548	11,251	12,811	14,212	15,257	---
3,316	4,703	6,023	8,027	9,554	11,260	12,950	14,233	15,301	---
3,394	4,767	6,150	8,152	9,667	11,353	13,020	14,308	15,302	---

The nucleotide positions are listed according to the rCRS numbering (Andrews *et al.*, 1999). The columns are read from top to bottom, starting with the most left hand column, and then repeating the process in the next column to the right. The list of nucleotide positions was compiled using data from investigations performed by Herrnstadt *et al.* (2002a), Galtier *et al.* (2006) and Kivisild *et al.* (2006).

After exclusion of nucleotide positions where recurrent alterations or reversion events transpired, nucleotide positions were selected for inclusion in the network based on their presence in all sequences located in the branch described by the specific haplogroup, sub-haplogroup or lineage within a sub-haplogroup, the nature of the alteration, and the genes in which the SNP was located. Selected SNPs were investigated in order to confirm that the SNP was present in all sequences located in a specific branch in the reduced global MP tree. Synonymous substitutions in genes other than the tRNA or rRNA genes were preferred. This is in accordance with convention as substitutions in polypeptide genes such as the ATPase6, COI, COII, Cytb, ND1, ND2, ND3, ND4 and ND5 genes were used in the construction of the classification scheme used in the current investigation (refer to Section 4.9). The nucleotide positions selected for inclusion in the revised L0 classification scheme are listed in Table 5.58.

Table 5.58 Nucleotide positions selected for inclusion in the revised L0 haplogroup classification scheme

Nucleotide position	Gene	Amino acid change	Nucleotide position	Gene	Amino acid change
2,245	16S rRNA	---	7,967	COII	SYN
2,758	16S rRNA	---	8,251	COII	SYN
3,372	ND1	SYN	8,428	ATPase8	SYN
3,423	ND1	SYN	9,545	COIII	SYN
3,438	ND1	SYN	10,589	ND4L	SYN
3,531	ND1	SYN	11,087	ND4	NON-SYN
3,594	ND1	SYN	11,143	ND4	SYN
3,618	ND1	SYN	12,121	ND4	SYN
3,666	ND1	SYN	12,798	ND1	SYN
3,866	ND1	NON-SYN	13,855	ND5	SYN
3,882	ND1	SYN	14,178	ND6	NON-SYN
3,981	ND1	SYN	15,136	Cytb	SYN
4,197	ND1	SYN	15,312	Cytb	NON-SYN
5,153	ND2	SYN	15,550	Cytb	SYN
6,257	COI	SYN	15,586	Cytb	SYN
6,692	COI	SYN	15,769	Cytb	SYN
7,148	COI	SYN	---	---	---

Nucleotide numbers are based on the rCRS numbering (Andrews *et al.*, 1999). Nucleotide positions indicated in **bold** text represent positions where recurrent alterations or reversions have been observed. These positions were however included if the nucleotide position has always been used to define macrohaplogroup L (3,594), or if no other haplogroup defining or haplogroup-specific SNP was available for classification of the sub-haplogroup or lineage within the sub-haplogroup. 16S rRNA = 16S ribosomal RNA; ATPase8 = ATP synthase F_0 subunit 8; COI = cytochrome *c* oxidase subunit I; COII = cytochrome *c* oxidase subunit II; COIII = cytochrome *c* oxidase subunit III; Cytb = cytochrome *b*; ND1 = NADH dehydrogenase subunit 1; ND2 = NADH dehydrogenase subunit 2; ND4 = NADH dehydrogenase subunit 4; ND4L = NADH dehydrogenase subunit 4L; ND5 = NADH dehydrogenase subunit 5; ND6 = NADH dehydrogenase subunit 6; SYN = synonymous substitution; NON-SYN = non-synonymous substitution.

Synonymous substitutions occurred at 27 of the 33 nucleotide positions selected for construction of the revised classification scheme. Of these nucleotide positions, 19 of the 33 were located in NADH dehydrogenase subunit genes, 6 of the 33 in cytochrome *c* oxidase subunit genes, 5 of the 33 in the cytochrome *b* gene, and 1 of the 33 in the ATP synthase F_0 subunit 8 gene. Only 2 of the 33 nucleotide positions selected were located in the 16S rRNA gene, and in both cases, no other SNP was available to define the lineage within the sub-haplogroup. Nucleotide positions indicated in **bold** text represent positions that were used to define macrohaplogroup L, or were subjected to recurrent alterations or reversion events, but were included as no other SNPs were available to define the specific sub-haplogroup or lineage within a sub-haplogroup. Detailed descriptions of the alterations at the nucleotide positions used in the revised L0 classification scheme were presented previously in Section 5.7. The revised L0 classification scheme is presented in Figure 5.61.

[illegible]

The colours used to indicate the different haplogroups and sub-haplogroups correspond to that used for the classification scheme presented in Section 4.9 and the alternative classification scheme illustrated in Section 3.3.1.1. The nucleotide present in the rCRS is indicated in regular text, and the nucleotide alteration is indicated in **bold** text. Haplogroup L5 describes the branch observed in the reduced global MP tree (refer to Figure 5.60), and corresponds to L5 described in the alternative classification scheme (refer to Section 3.3.1.1).

The revised L0 classification scheme exhibited a number of differences compared to the classification scheme presented in Section 4.9. These differences are listed in Table 5.59.

Table 5.59 Differences in the revised L0 classification scheme compared to the classification scheme used in the current investigation

Branching point in revised L0 classification scheme	Classification scheme (Wallace, 2004)	Revised L0 classification scheme	Reason for replacement or inclusion of nucleotide position
L0, L1, L2 vs. L5	---	3,423	Separates L0, L1, L2 from L5 Inclusion of L5 to classify group of sequences (DQ341060, DQ341061, EF184580, EF184581, EF184582, EF184583, EF184584)
L0, L1 vs. L2	10,810	2,758	10810C not present in all L1 and L0 sequences (absent in EF184589, EF184595, EF184596, EF184607, EF184608, EF184614)
L0 vs. L1	7,055	14,178	Recurrent alteration in human mtDNA network (Kivisild <i>et al.</i> , 2006)
L0 vs. L1	11,914	10,589	Mutation hot spot (Galtier <i>et al.</i> , 2006) Recurrent alteration (Kivisild <i>et al.</i> , 2006)
L0c1 vs. L0c2	5,096	6,257	5096C not present in all L0c1 sequences (absent in DQ112959)

The majority of the differences between the classification scheme presented in Section 4.9 and the revised L0 classification scheme can be ascribed to the inclusion of new L0 sequences in the global L0-specific haplogroup network. These new L0 sequences consisted of the 22 Southern African Khoi-San L0 mtDNA sequences generated in the current investigation, eight L0 African-American sequences (Just *et al.*, 2008), one L0 sequence from the Dominican Republic (Torrioni *et al.*, 2006) and 15 L0 sequences from southern Africa and Tanzania (Gonder *et al.*, 2007). These recently generated L0 sequences did not conform to the classification scheme used in the current investigation, and the alterations identified within these new L0 sequences necessitated an update of the nucleotide positions and the SNPs identified at these positions used to classify L0 sequences.

The sub-haplogroup designations and the assignment of lineages within the sub-haplogroups were compared between the classification scheme used in the current investigation (refer to Section 4.9, Figure 4.1), the alternative classification scheme (refer to Section 3.3.1.1, Figure 3.7), and the revised L0 classification scheme constructed in the current investigation (refer to Figure 5.61). The comparison of the haplogroup and sub-haplogroup designation between the three classification schemes is presented in Table 5.60. Haplogroups, sub-haplogroups and lineages within sub-haplogroups in the revised L0 classification scheme were named as in the global L0-specific haplogroup network. New designations were however proposed for the new branches representing sub-haplogroups and lineages within sub-haplogroups identified in the global L0-specific haplogroup network and the global N-J and MP trees. The proposed new designations were based on the estimated ages of the sub-haplogroups or lineages, as determined in Section 5.11.6. Thereby, L0d was expected to be an older sub-haplogroup than L0e. In the current investigation, using the revised L0 classification scheme, it is suggested that L0aB is renamed to L0d, L0aC to L0e, L0bA to L0b1, and L0bB to L0b2. From the information listed in Table 5.60, it can also be deduced that there is an urgent need for a standardised classification scheme for mtDNA haplogroups, as is currently available for the Y chromosome (The Y Chromosome Consortium, 2002), to prevent confusion and to facilitate the direct comparison of data generated for mtDNA haplogroups in different human populations.

Table 5.60 Comparison of haplogroup and sub-haplogroup designations between three classification schemes

Classification scheme (Wallace, 2004) ^a	Global L0-specific haplogroup network ^b	Revised L0 classification scheme ^c	Alternative haplogroup classification scheme ^d
L0a	L0aA	L0a	L0k
---	L0aB	L0d	L0d3 ^e
---	L0aC	L0e	L0f
L0b	L0b	L0b	L0d
---	L0bA	L0b1	L0d2
---	L0bB	L0b2	L0d1
L0c	L0c	L0c	L0a
L0c1	L0c1	L0c1	L0a1
L0c2	L0c2	L0c2	L0a2

^a = classification scheme used in current investigation (refer to Section 4.9). ^b = global L0-specific haplogroup network constructed in current investigation (refer to Section 5.10.2). ^c = revised classification scheme for L0 haplogroup proposed in the current investigation (refer to Section 5.12). ^d = haplogroup classification scheme used in other investigations described in the literature (refer to Section 3.3.1.1). ^e = new branch in L0d, specific to East African Khoi-San speakers (Gonder *et al.*, 2007), named L0d3 in the current investigation to distinguish it from L0d1 and L0d2. This branch was first described by Gonder *et al.* (2007) and named L0dTZ.

The revised L0 classification scheme is used to determine the sub-haplogroup or lineage designation of a new L0 mtDNA genome sequence. This is accomplished by first identifying the nucleotide alterations within the new sequence, and then tracking these nucleotide alterations in the new sequence along the branching order of the revised L0 classification scheme. The revised L0 classification scheme was tested by re-classifying mtDNA sequences that presented with a discrepancy between their classifications according to the classification scheme presented in Section 4.9 and their location within the global N-J and MP trees (refer to Section 5.9). The haplogroup designations of the re-classified L0 sequences are indicated in Table 5.61.

Table 5.61 Assignment of haplogroups and sub-haplogroups to L sequences that previously exhibited a discrepancy between classification and location in global N-J and MP trees

GenBank [®] accession number	Haplogroup classification		
	Wallace, 2004 ^a	Global phylogenetic trees ^b	Revised L0 classification scheme ^c
AF381992	L0,L1	L1	L1
DQ112696	L0,L1	L1	L1
DQ112697	L0,L1	L1	L1
DQ112921	L0a\$	L5	L5
DQ112922	L0a\$	L5	L5
DQ112959	L0c2\$	L0c1	L0c1B
DQ341060	L0a\$	L5	L5
DQ341061	L0a\$	L5	L5
EF184580	L0b\$	L5	L5
EF184581	L0b\$	L5	L5
EF184582	L0a\$	L5	L5
EF184583	L0b\$	L5	L5
EF184584	L0b\$	L5	L5
EF184589	L2a	L0	L0aB1
EF184595	L2a	L0	L0aC
EF184596	L2a	L0	L0aC
EF184600	L0a\$	L0	L0aC
EF184607	L2a	L0c2	L0c2E
EF184608	L2a	L0c2	L0c2E
EF184612	L0,L1	L1	L1

^a = haplogroup according to classification scheme used in the current investigation (refer to Section 4.9). ^b = haplogroup classification according to global phylogenetic trees constructed in current investigation (refer to Section 5.9). ^c = haplogroup classification according to the revised classification scheme for haplogroup L0 proposed in the current investigation (refer to Section 5.12). The \$ sign indicates that these L sequences could not be completely classified using the classification scheme presented in Section 4.9.

The revised L0 classification scheme resolved the discrepancies previously observed between the haplogroup classification and location of specific L sequences in the global N-J and MP trees. This was expected, as the revised scheme was based on

100 sequences from the L0 haplogroup, which included the complete coding region sequence information for each individual to identify all SNPs in each L0 sequence, and was used to determine which SNPs were specific to which sub-haplogroups or lineages by using the global L0-specific haplogroup network as a guide. In contrast, the classification scheme used in the current investigation was based on the identification of 103 previously identified SNPs in a group of sequences, after which a set of exclusion criteria was compiled to separate sequences into the major L haplogroups, namely L0, L1, L2 and L3 (refer to Section 4.9). Revision of the L1, L2 and L3 haplogroup classification schemes are also required as a large number of coding region mtDNA sequences are presently available to be used for this purpose. The revision of these classification schemes can be performed in the same manner as for the L0 haplogroup by using the global mtDNA mutational phylogeny (Ruiz-Pesini *et al.*, 2007) as a starting point to construct haplogroup-specific networks for each individual haplogroup.

5.13 KHOI-SAN CONSENSUS SEQUENCE

As previously mentioned, African populations are generally older and exhibit higher levels of variation compared to populations on other continents. The increased age of African populations compared to other populations is indicated by the position of the mitochondrial genome sequences assigned to African mitochondrial haplogroups in global phylogenetic trees based on mitochondrial sequences collected from all continents (refer to Section 3.3), as well as the estimated TMRCAs determined for African mtDNA haplogroups (refer to Section 3.3.1.4). The Khoi-San sequences determined in the current investigation shared this characteristic regarding age as indicated by the location of the Khoi-San genome sequences in the global phylogenetic trees (refer to Section 5.9) and the estimated TMRCAs determined for the mitochondrial haplogroups present in the Khoi-San cohort of 118,010 to 131,544 YBP for L0a and 108,009 to 115,687 YBP for L0b (refer to Section 5.11.6). The high levels of variation observed in African populations are evident from the high nucleotide diversity determined for the African population as a whole, the Khoi-San population and individual African mitochondrial haplogroups (refer to Section 5.11.2).

In the current investigation it is proposed that a consensus sequence based on the Southern African Khoi-San population should be constructed in order to summarise the nucleotide variation present in African mitochondrial genome sequences. As discussed in this section, such a consensus sequence should prove invaluable in the study of the

evolution of the human mitochondrial genome as it is based on the genome sequences of Khoi-San individuals who have been identified as carriers of some of the most ancient human mitochondrial lineages.

The construction of a consensus sequence for the human mitochondrial DNA genome is not an unprecedented concept and is necessary for the study of human mtDNA diversity (refer to Section 3.6). Marzuki *et al.* (1991) and Carter (2007) each attempted to construct a consensus sequence that would represent all human mtDNA sequences. These consensus sequences were constructed using mtDNA sequences from different human mtDNA haplogroups and were referred to as human mitochondrial consensus sequences. To study human mitochondrial genome diversity, the ideal would be to generate a consensus sequence for every population-specific haplogroup in order to obtain a haplogroup-specific consensus sequence. A distinction must however be made between a haplogroup-specific consensus sequence and a population-specific consensus sequence. A haplogroup-specific consensus sequence consists of sequences of only one haplogroup, while a population-specific consensus sequence can contain contributions from different haplogroups as long as those haplogroups are observed within the study population. In the current investigation the aim was to construct a consensus sequence for a specific population group, the Khoi-San of southern Africa, which harbours two of the oldest known mtDNA L0 sub-haplogroups, L0a and L0b, as well as low frequencies of other L-specific haplogroups. The consensus sequence constructed in the current investigation will be specific for the Khoi-San population of southern Africa, but as discussed later in this section, it must be stressed that the Khoi-San consensus sequence is not a haplogroup-specific consensus sequence.

High levels of variation are generally observed in African populations as supported by the variation exhibited in African mtDNA sequences compared to the current mtDNA genome standard, the rCRS. Of these African populations, the Khoi-San is one of the most ancient and one of the most divergent human populations (Chen *et al.*, 2000). Therefore, a consensus sequence, which provides a summary of the observed sequence variation (Page and Holmes, 1998), was constructed instead of a reference sequence (refer to Sections 3.5 and 3.6). The consensus sequence constructed in the current investigation can be used to investigate the evolution of the mitochondrial genome, and thus the manner in which the different mtDNA haplogroups adapted over time. In addition, the consensus sequence can be used as a baseline for all other mitochondrial sequences. This will permit the exclusion of population-specific alterations from mtDNA sequences

obtained from individuals with mitochondrial disorders thereby facilitating the identification of possible pathogenic alterations. This is especially relevant for individuals of African descent, as comparison with the rCRS generates a large number of alterations, some of which are specific to African populations. Comparison with the Khoi-San consensus sequence will allow differentiation between alterations that are specific to African populations, and alterations that are likely to be disease-related.

A consensus sequence was constructed for the Khoi-San cohort as described in Section 4.16 and a total of 35 Khoi-San mtDNA sequences were used to construct the Khoi-San consensus sequence. Thirteen of these sequences were determined in a previous investigation (Mouton, 2003), while the remaining 22 sequences were generated during the course of the current investigation. All the sequences used to construct the Khoi-San consensus sequence were assigned to the L0 haplogroup and belonged either to the L0a or L0b sub-haplogroups (refer to Section 5.8.1). Both the L0a and L0b sub-haplogroups were included as these two sub-haplogroups represent the two most basal branches in global phylogenetic trees (Gonder *et al.*, 2007). Khoi-San sequences classified in other L-specific haplogroups (refer to Section 5.8.1) such as L2 or L3 were excluded from the analysis. The Khoi-San mitochondrial genome sequences used for the construction of a consensus sequence for the Khoi-San population of southern Africa is listed in Table 5.62.

Table 5.62 Khoi-San sequences used to compile the Khoi-San consensus sequence

Sequence name	Haplogroup/ sub-haplogroup classification	Sequence name	Haplogroup/ sub-haplogroup classification	Sequence name	Haplogroup/ sub-haplogroup classification
KS_CGR_001	L0a	KS_CGR_060	L0b	KS_CGR_105	L0b
KS_CGR_004	L0b	KS_CGR_061	L0b	KS_CGR_106	L0a
KS_CGR_005	L0b	KS_CGR_072	L0b	KS_CGR_111	L0b
KS_CGR_010	L0b	KS_CGR_077	L0b	KS_CGR_117	L0b
KS_CGR_032	L0b	KS_CGR_080	L0a	KS_CGR_120	L0a
KS_CGR_038	L0a	KS_CGR_082	L0b	KS_CGR_121	L0b
KS_CGR_040	L0b	KS_CGR_083	L0a	KS_CGR_122	L0b
KS_CGR_044	L0b	KS_CGR_084	L0a	KS_CGR_127	L0b
KS_CGR_048	L0b	KS_CGR_091	L0b	KS_CGR_130	L0b
KS_CGR_052	L0b	KS_CGR_092	L0b	KS_CGR_157	L0a
KS_CGR_056	L0b	KS_CGR_094	L0b	KS_CGR_176	L0a
KS_CGR_059	L0a	KS_CGR_102	L0b	---	---

Haplogroup or sub-haplogroup classification was assigned according to the coding region only classification scheme described in Section 4.9. The Khoi-San mitochondrial sequences obtained from a previous investigation (Mouton, 2003) are indicated in **bold** text. The remaining Khoi-San sequences listed in the table were generated in the current investigation.

Only Khoi-San sequences classified as belonging to the L0 haplogroup were used to construct the consensus sequence for this population due to the fact that the L0 haplogroup was the most prevalent haplogroup in the Khoi-San cohort. The cohort consisted of 73.33% L0 sequences, and 26.67% other L-specific sequences (10% L2 and 16.67% L3), which was similar to the 79.07% L0 and 20.93% other L-specific sequences (4.65% L2 and 16.28% L3) observed by Chen *et al.* in 2000 (refer to Section 5.8.1). The L0 sequences observed in the Khoi-San cohort belong to the L0aA (L0k¹) and L0b (L0d) sub-haplogroups (refer to Appendix G), and as previously mentioned, these two sub-haplogroups, based on the datasets investigated in the current investigation, appear to be specific to the Southern African Khoi-San population (refer to Section 5.10.2.2). It can therefore be suggested that the L0aA and L0b mtDNA haplogroups are a defining feature of this population. This observation is supported by data from other investigations and it has previously been suggested that the L0aA (L0k) and L0b (L0d) sub-haplogroups are population-specific (refer to Section 3.3.1.3). In addition, the presence of other L-specific mtDNA haplogroups within the cohort can most likely be attributed to the introduction of the L2 and L3 haplogroups into the Khoi-San population due to the occurrence of admixture, incomplete data regarding ethnicity of previous generations in pedigrees, and self-reporting of ethnicity (refer to Section 5.8.1). As previously discussed, the presence of other L mtDNA haplogroups within the cohort can therefore be attributed to genetic influences from neighbouring African populations (refer to Section 5.8.1). It is however important to note that the Khoi-San consensus sequence is not an L0 haplogroup consensus sequence, as L0 sequences belonging to the L0c1 and L0c2 sub-haplogroups were not present in the Khoi-San cohort, and were therefore not considered in the construction of the consensus sequence.

The Khoi-San consensus sequence was constructed in two steps. The first step entailed the construction of a consensus sequence for the ten L0a sequences and a separate consensus sequence for the 25 L0b sequences where the consensus sequence for each set of sequences was determined by retaining the nucleotide most commonly observed at each position. The second step involved combining the two separate consensus sequences to compile a single consensus sequence. At positions where differences were observed between the L0a and L0b consensus sequences for which a consensus

¹ The haplogroup classifications presented in brackets refer to the classifications according to the alternative classification scheme (refer to Section 3.3.1.1).

character state could not be determined, the nucleotide was resolved by defining it using an ambiguity code (R, Y, S or M). In the case of insertions and deletions, the consensus for that position was determined by using a consensus sequence constructed using all 35 L0 sequences. An example of such an instance was the retention of the deletion of a C at position 498 which occurred in the L0b consensus sequence, but not in the L0a consensus sequence.

The complete Khoi-San consensus sequence consisted of a total of 16,567 bp and is presented in Appendix N. The N at position 3,107 in the rCRS is retained in the Khoi-San consensus sequence in order to adhere to the numbering system used in the rCRS (Andrews *et al.*, 1999). In the current investigation, the Khoi-San consensus sequence is presented as a list of SNPs compared to the rCRS. The list of SNPs is presented in Figure 5.62.

Figure 5.62 Consensus sequence generated for the Khoi-San of southern Africa presented in the form of a list of SNPs compared to the rCRS

73G	146C	152C	189R	195C	198T	207R	247A
315iC	456Y	498dC	514dC	515dA	719R	750G	769A
825A	850Y	1018A	1048T	1243Y	1438R	2706R	2758A
2836M	2885C	3438R	3516A	3594T	3756R	4104G	4197Y
4232Y	4312T	4541R	4586Y	4769G	4907Y	5442C	5811R
6185C	6266R	6815Y	6938Y	7028T	7146G	7256T	7257R
7521A	7828R	8113M	8152R	8251R	8468T	8655T	8701G
8860G	8911Y	8994R	9042T	9136R	9150R	9347G	9438R
9540C	9755R	9818Y	10398G	10499R	10589A	10664T	10688A
10810C	10873C	10876R	10915C	10920Y	10939Y	11296Y	11299Y
11437Y	11653R	11719A	11914A	12007A	12070R	12121Y	12235Y
12705T	12720R	13020Y	13105G	13129Y	13276G	13506T	13590R
13650T	13759R	13819Y	13928S	14020Y	14182Y	14371Y	14374Y
14766T	15326G	15466R	15550Y	15930R	15941Y	15951R	16166M
16167Y	16172Y	16187T	16189C	16209Y	16214Y	16219S	16223T
16230G	16234Y	16242Y	16243Y	16278Y	16311C	16519C	

The numbering of the SNPs follows the convention used for the rCRS (Andrews *et al.*, 1999). The nucleotide number is followed by the nucleotide present at the position. SNPs presented in red text are located in the control region (nucleotide 16,024 to 576), while SNPs presented in black text are situated in the coding region (nucleotide 577 to 16,023). Alterations indicated in blue text refer to SNPs used as exclusion criteria (refer to Appendix D) and SNPs presented in green text represent the SNPs used to exclude L0 from L1 (refer to Appendix D). SNP = single nucleotide polymorphism; rCRS = revised Cambridge Reference Sequence; A = adenine; C = cytosine; G = guanine; T = thymine; R = A or G; Y = C or T; S = G or C; M = A or C; i = insertion; d = deletion.

The Khoi-San consensus sequence exhibited a total of 127 alterations which consisted of 123 nucleotide changes, three deletions and one insertion compared to the rCRS. The alterations were distributed throughout the consensus sequence with 22.83% (29/127) in the control region and 77.17% (98/127) in the coding region. The distribution of observed

alterations throughout the mitochondrial genome in the Khoi-San consensus sequence compared to the rCRS is presented in Table 5.63.

Table 5.63 Distribution of alterations observed throughout the mitochondrial genome in the Khoi-San consensus sequence compared to the rCRS

Region	Locus	Alterations per region	
		Number	Percentage
Control region	HVS1	15	11.81
	HVS2	9	7.09
	HVS3	4	3.15
	Other	1	0.79
	Sub-Total	29	22.83
Coding region	12S ribosomal RNA	9	7.09
	16S ribosomal RNA	4	3.15
	ATPase6	8	6.30
	ATPase8	1	0.79
	ATP synthase subunits	9	7.09
	Cytochrome <i>b</i>	4	3.15
	COI	8	6.30
	COII	4	3.15
	COIII	5	3.94
	Cytochrome c oxidase subunits	17	13.39
	ND1	7	5.51
	ND2	5	3.94
	ND3	1	0.79
	ND4	15	11.81
	ND5	13	10.24
	ND6	3	2.36
	All NADH dehydrogenase subunits	48	37.80
	Transfer RNA genes	7	5.51
	Sub-Total	98	77.17
Total		127	100.00

rCRS = revised Cambridge Reference Sequence; HVS = hypervariable segment; ATPase = ATP synthase F₀ subunit; CO = cytochrome c oxidase subunit; ND = NADH dehydrogenase subunit.

The majority of alterations in the control region, namely 11.81% (15/29) were observed in the HVS1, while the bulk of alterations in the coding region, namely 37.80% (48/98) were present in the NADH dehydrogenase subunit genes with 11.81% (15/48) in ND4 and 10.24% (13/48) in ND5. This was not surprising as it has previously been demonstrated that ND4 and ND5 are subject to recurrent alterations (refer to Section 5.10.2.3) which is indicative of the generally high levels of variability in mtDNA genes in African populations (Mishmar *et al.*, 2003). Previous investigations have attributed the presence of high levels of polymorphisms in mtDNA genes to the effects of natural selection (Mishmar *et al.*, 2003;

Ruiz-Pesini *et al.*, 2004; Mishmar *et al.*, 2006; Ingman and Gyllensten, 2007b). As expected, the alterations in the coding region included all the SNPs that were listed as exclusion criteria, with the exception of 15,301 which excludes haplogroups from macrohaplogroup N (refer to Appendix D).

As indicated in the list of SNPs presented in Figure 5.62, the Khoi-San consensus sequence consisted of 54.33% (69/127) ambiguous positions which were located throughout the consensus sequence with a total of 44.83% (13/29) in the control region (position 16,024 to 576) and 57.14% (56/98) in the coding region (position 577 to 16,023). The use of ambiguity codes in the Khoi-San consensus sequence permitted the retention of polymorphism information contributed by both the L0a and L0b sub-haplogroups observed within the Khoi-San population. The use of ambiguity codes in the consensus sequence would also provide a way to deal with the occurrence of heteroplasmy at specific positions. A comprehensive and sensitive assessment of heteroplasmy was beyond the technical boundaries of the sequencing approach employed in the current investigation (refer to Appendix B) and sequence electropherograms were visually examined to ensure correct base-calling (refer to Section 5.6.6). Heteroplasmy was however not observed in the Khoi-San sequences generated in the current investigation. This can be attributed to the fact that the Khoi-San individuals included in the current investigation were all apparently healthy, since no mitochondrial disorders were reported for any of the individuals included in the original sample group collected by Chen *et al.* in 2000.

Four ambiguity codes were used in the Khoi-San consensus sequence, Y which identified a C or T (35/69), R which indicated the presence of an A or G (29/69), M which defined a G or C (3/69) and S which denoted an A or C (2/69) at a specific position. In the majority of the ambiguous nucleotide positions (64/69) the class of nucleotide at the specific position remained the same, i.e. a purine or pyrimidine was replaced with another purine or pyrimidine ($A \rightleftharpoons G$ or $T \rightleftharpoons C$). In addition, in the case of every ambiguous nucleotide position, one of the options represented the nucleotide observed in one of the two L0 sub-haplogroups, while the other represented the nucleotide observed in the rCRS. In the case of these ambiguous nucleotide positions, it implies that when the nucleotide is identical to that observed in the rCRS, the nucleotide present in one of the two sub-haplogroups was identical to the nucleotide observed in the rCRS. The use of ambiguity codes in the Khoi-San consensus sequence should decrease the number of observed alterations in the comparisons between mtDNA sequences and the consensus

sequence, thereby reducing the number of alterations identified that could be responsible for any observed pathogenicity.

The presence of 10/14 (71.43%) ambiguous nucleotide positions in HVS1 (nucleotide 16,024 to 16,383) of the control region of the Khoi-San consensus sequence provided information regarding the normal variants at the positions under investigation. The high number of ambiguous nucleotide positions in this region is an indication of the high level of variation present and supports the use of a classification scheme that is not based on HVS1 information alone, but which also includes additional information from HVS2 and the coding region. This observation therefore provided further support for the use of a coding region only haplogroup classification scheme (refer to Section 5.12). A classification scheme based on coding region SNPs only would thus provide a greater number of defined alterations to be employed for the unambiguous classification of sequences.

The percentage similarity and the pairwise number of differences between each of the 43 individual Khoi-San sequences included in the current investigation, as well as between the 43 Khoi-San sequences and the Khoi-San consensus sequence were determined using PHYLIP version 3.66 (Felsenstein, 1989) and MEGA version 4 (Tamura *et al.*, 2007) as described in Section 4.17. A matrix containing the percentage similarity and the pairwise number of nucleotide differences between the 43 Khoi-San sequences, as well as between the 43 Khoi-San sequences and the Khoi-San consensus sequence is presented in Appendix O. These values were calculated by excluding positions containing gaps, i.e. positions with insertions and deletions within the sequence alignment used to generate the matrix. The minimum and maximum percentage similarity and number of pairwise differences between Khoi-San sequences, and between the Khoi-San sequences and the Khoi-San consensus sequence presented in Appendix O is summarised in Table 5.64.

Table 5.64 Minimum and maximum percentage similarity and number of pairwise differences between Khoi-San sequences, and between the Khoi-San sequences and the Khoi-San consensus sequence

Comparison of 43 Khoi-San sequences				
Value	Min.	Sequence name	Max.	Sequence name
Percentage similarity (%)	99.30	KS_CGR_044 vs. KS_CGR_164	100.00	KS_CGR_038 vs. KS_CGR_083 KS_CGR_077 vs. KS_CGR_102 KS_CGR_077 vs. KS_CGR_121 KS_CGR_082 vs. KS_CGR_122 KS_CGR_102 vs. KS_CGR_121 KS_CGR_140 vs. KS_CGR_143
No. of pairwise differences (nucleotides)	0	KS_CGR_038 vs. KS_CGR_083 KS_CGR_059 vs. KS_CGR_106 KS_CGR_077 vs. KS_CGR_102 KS_CGR_077 vs. KS_CGR_121 KS_CGR_082 vs. KS_CGR_122 KS_CGR_102 vs. KS_CGR_121 KS_CGR_140 vs. KS_CGR_143	99	KS_CGR_044 vs. KS_CGR_164
Comparison of 43 Khoi-San sequences with Khoi-San consensus sequence				
Value	Min.	Sequence name	Max.	Sequence name
Percentage similarity (%)	99.55	KS_CGR_164	100.00	KS_CGR_038 KS_CGR_083
No. of pairwise differences (nucleotides)	0	KS_CGR_038 KS_CGR_083	74	KS_CGR_164

Min. = minimum; Max. = maximum; No. = number; vs. = versus.

In the comparison between the 43 Khoi-San sequences, the minimum percentage similarity and the maximum number of pairwise differences were observed between KS_CGR_044 and KS_CGR_164. This can be attributed to differences between L haplogroups as KS_CGR_044 is classified within the L0b haplogroup, while KS_CGR_164 is assigned to the L2b haplogroup (refer to Section 5.8.1 and Appendix E). The maximum percentage similarity and the minimum number of pairwise differences were observed between seven pairs of Khoi-San sequences (refer to Table 5.64). One pair, KS_CGR_059 vs. KS_CGR_106 displayed a minimum number of zero pairwise differences, but did not present with 100% similarity as observed for the other six pairs. A percentage similarity of 99.99% was calculated for this pair. This discrepancy can be ascribed to the use of two different methods to calculate the two values as PHYLIP version 3.66 was used to calculate percentage similarity, while MEGA version 4 was employed to determine the number of pairwise differences. From the data presented in Table 5.64, it can be deduced that KS_CGR_038 is identical to KS_CGR_083, KS_CGR_059 to KS_CGR_106, KS_CGR_077 to KS_CGR_102 and KS_CGR_121, KS_CGR_082 to KS_CGR_122, and KS_CGR_140 to KS_CGR_143. A closer inspection

of the individual Khoi-San sequences confirmed that KS_CGR_082 was identical to KS_CGR_122, and KS_CGR_140 to KS_CGR_143. Comparisons of the sequences in the remaining pairs however uncovered a number of differences. KS_CGR_038 contained a deletion of a C at nucleotide position 309 in comparison with KS_CGR_083, and KS_CGR_059 exhibited a CC insertion at nucleotide position 309 and a C insertion at nucleotide position 16,193, while KS_CGR_106 contained a C deletion at nucleotide position 309. KS_CGR_077 was identical to KS_CGR_121, but KS_CGR_102 did not contain an insertion of a C at nucleotide position 309 which was observed in both KS_CGR_077 and KS_CGR_121. The exclusion of insertions and deletions therefore influenced the results obtained in comparisons between the 43 Khoi-San sequences.

In the comparison of the consensus sequence with the 43 Khoi-San sequences, the minimum percentage similarity and maximum number of pairwise differences were observed in KS_CGR_164. This is likely due to differences between L haplogroups as the Khoi-San consensus sequence was compiled using sequences classified as L0a and L0b, while KS_CGR_164 is classified within the L2b haplogroup (refer to Section 5.8.1 and Appendix E). The maximum percentage similarity and the minimum number of pairwise differences were observed between the Khoi-San consensus sequence and two Khoi-San sequences, namely KS_CGR_038 and KS_CGR_083. As already observed in the comparison of the 43 Khoi-San sequences, the exclusion of insertions and deletions are likely to influence the results obtained upon comparison of the 43 Khoi-San sequences to the Khoi-San consensus sequence. The number of alterations observed between the Khoi-San consensus sequence and the Khoi-San sequences generated in the current investigation, as well as 13 Khoi-San sequences obtained from a previous investigation (Mouton, 2003), were therefore also determined by taking insertions and deletions into account and these values are presented in Table 5.65.

Table 5.65 Differences between the Khoi-San consensus sequence and 43 Khoi-San sequences included in the current investigation

Haplogroup/ sub-haplogroup	Sequence name	No. of differences	Haplogroup/ sub-haplogroup	Sequence name	No. of differences
L0a	KS_CGR_084	7	L0b	KS_CGR_082	4
	KS_CGR_120	7		KS_CGR_092	4
	KS_CGR_059	6		KS_CGR_117	4
	KS_CGR_106	5		KS_CGR_032	3
	KS_CGR_176	5		KS_CGR_072	3
	KS_CGR_001	3		KS_CGR_102	3
	KS_CGR_080	3		KS_CGR_105	3
	KS_CGR_038	2		KS_CGR_122	3
	KS_CGR_157	2		KS_CGR_004	2
	KS_CGR_083	1		KS_CGR_010	2
L0b	KS_CGR_044	32	L0b	KS_CGR_077	2
	KS_CGR_048	19		KS_CGR_111	2
	KS_CGR_060	18		KS_CGR_121	2
	KS_CGR_040	17	L2b	KS_CGR_164	77
	KS_CGR_056	17		KS_CGR_173	74
	KS_CGR_094	17		KS_CGR_069	72
	KS_CGR_052	15	L3a1	KS_CGR_140	52
	KS_CGR_061	15		KS_CGR_143	52
	KS_CGR_127	14		KS_CGR_175	51
	KS_CGR_130	14	L3b	KS_CGR_172	56
	KS_CGR_005	7		KS_CGR_090	55
	KS_CGR_091	5		---	---

Haplogroup or sub-haplogroup classification was assigned according to the coding region only classification scheme described in Section 4.9. The Khoi-San mitochondrial sequences obtained from a previous investigation (Mouton, 2003) are indicated in **bold** text. The remaining Khoi-San sequences listed in the table were generated in the current investigation.

The largest number of differences between the Khoi-San consensus sequence and the 43 Khoi-San sequences included in the current investigation were observed between the Khoi-San sequences classified within the L2b haplogroup, followed by the Khoi-San sequences classified within the L3b and L3a1 sub-haplogroups. This could be due to the times of divergence of the different haplogroups. Based on RFLP and control region sequence data, Chen *et al.* in 2000 estimated that L3 diverged between 78,300 and 103,200 YBP, while L2 diverged between 59,000 and 77,700 YBP. According to their data, the L3 haplogroup is older than the L2 haplogroup. It is therefore expected that the L3 haplogroup will be more similar to the Khoi-San consensus sequence which is based on L0 sequences which are older than both haplogroups, while the L2 haplogroup which is younger should display more differences compared to the Khoi-San consensus sequence. From this observation it is suggested that the haplogroups observed outside of Africa, i.e. the more recently diverged haplogroups, should have more differences compared to the

Khoi-San consensus sequence than haplogroups observed within Africa. This trend is the opposite to that observed for the rCRS where sequences from haplogroups observed within Africa exhibit more differences compared to the rCRS than sequences from haplogroups present outside of Africa.

As previously mentioned in this section, it is envisaged that comparison of mtDNA sequences with the Khoi-San consensus sequence will permit differentiation between alterations that are specific to African populations, and alterations that are likely to be disease-related. A list of putative population-specific alterations for the Southern African Khoi-San population was compiled in the current investigation using the Khoi-San consensus sequence which was based on the alterations observed in 43 Khoi-San sequences from southern African. The alterations were selected based on their presence in the coding region of all 43 Khoi-San sequences studied in the current investigation, and nucleotide positions exhibiting ambiguity were excluded. The list of potential population-specific alterations is provided in Table 5.66, and includes alterations that form part of the exclusion criteria listed for L-specific sequences (refer to Appendix D), alterations that cause synonymous changes, as well as non-synonymous alterations that cause amino acid changes. As the list of alterations is based on the Khoi-San consensus sequence which was constructed using 43 Khoi-San mitochondrial genome sequences obtained from apparently healthy Khoi-San individuals, none of the alterations are expected to be disease-related.

Table 5.66 Alterations contained in the Khoi-San consensus sequence that are likely to be population-specific

Nucleotide	Alteration	Type of alteration	Gene	Amino acid change	Conservation index
750	A – G	TS	Non-coding	---	---
769	G – A	TS	Non-coding	---	---
825	T – A	TV	Non-coding	---	---
1,018	G – A	TS	Non-coding	---	---
1,048	C – T	TS	Non-coding	---	---
2,758	G – A	TS	Non-coding	---	---
2,885	T – C	TS	Non-coding	---	---
3,516	C – A	TV	ND1	SYN	---
3,594	C – T	TS	ND1	SYN	---
4,104	A – G	TS	ND1	SYN	---
4,312	C – T	TS	Non-coding	---	---
4,769	A – G	TS	ND2	SYN	---
5,442	T – C	TS	ND2	F325L	7

Table 5.66 Continued...

Nucleotide	Alteration	Type of alteration	Gene	Amino acid change	Conservation index
6,185	T – C	TS	COI	SYN	---
7,028	C – T	TS	COI	SYN	---
7,146	A – G	TS	COI	T415A	13
7,256	C – T	TS	COI	SYN	---
7,521	G – A	TS	Non-coding	---	---
8,468	C – T	TS	ATPase8	SYN	---
8,655	C – T	TS	ATPase6	SYN	---
8,701	A – G	TS	ATPase6	T59A	17
8,860	A – G	TS	ATPase6	T112A	28
9,042	C – T	TS	ATPase6	SYN	---
9,347	A – G	TS	COIII	SYN	---
9,540	T – C	TS	COIII	SYN	---
10,398	A – G	TS	ND3	T114A	21
10,589	G – A	TS	ND4L	SYN	---
10,664	C – T	TS	ND4L	SYN	---
10,688	G – A	TS	ND4L	SYN	---
10,810	T – C	TS	ND4	SYN	---
10,873	T – C	TS	ND4	SYN	---
10,915	T – C	TS	ND4	SYN	---
11,719	G – A	TS	ND4	SYN	---
11,914	G – A	TS	ND4	SYN	---
12,007	G – A	TS	ND4	SYN	---
12,705	C – T	TS	ND5	SYN	---
13,105	A – G	TS	ND5	I257V	17
13,129	C – Y	TS	ND5	P265S	36
13,276	A – G	TS	ND5	M314V	31
13,506	C – T	TS	ND5	SYN	---
13,650	C – T	TS	ND5	SYN	---
14,766	C – T	TS	Cytb	T7I	19
15,326	A – G	TS	Cytb	T194A	8

Nucleotides are represented by capital letters, with the rCRS nucleotide indicated first, followed by the observed change. Nucleotide numbers are based on rCRS numbering (Andrews *et al.*, 1999). Nucleotide positions indicated in **blue** text refer to positions used as exclusion criteria (refer to Appendix D) and nucleotide positions presented in **green** text represent the positions used to exclude L0 from L1 (refer to Appendix D). A = adenine; C = cytosine; G = guanine; T = thymine; TS = transition; TV = transversion; ATPase = ATP synthase F₀ subunit; CO = cytochrome c oxidase subunit; Cytb = cytochrome b; ND = NADH dehydrogenase subunit; SYN = synonymous substitution; A = alanine; F = phenylalanine; I = isoleucine; L = leucine; M = methionine; P = proline; S = serine; T = threonine; V = valine.

Certain alterations such as the transitions at positions 8,860, 13,129 and 13,276 occur at amino acid residues with high conservation indices (28, 36 and 31) which suggest a possible detrimental effect. However, as discussed in Section 5.11.5.2, the alterations are not expected to have detrimental functional effects due to their retention in mitochondrial lineages at branching points in the global L0-specific haplogroup network between sub-haplogroups and within branches of sub-haplogroups. This observation is confirmed

by the presence of these specific alterations in the Khoi-San consensus sequence (refer to Figure 5.62), and instead it can be suggested that these alterations are either neutral variants within the population, or confer some adaptive function to the specific haplogroups and sub-haplogroups in which they are observed (Wallace, 1994; Wallace, 2005). In addition, the inclusion of all the SNPs that form part of the exclusion criteria, which are used for classification of L-specific sequences, in the list of potential population-specific alterations, indicates that the Khoi-San consensus sequence is a good approximation of the variation observed in African L mitochondrial haplogroups and can therefore be viewed as representative of the variation observed in the African population.

As previously stated in this section, two human mitochondrial consensus sequences have been constructed by Marzuki *et al.* in 1991 and Carter in 2007. The Khoi-San consensus sequence was compared to both of these human consensus sequences. The differences observed between these two human consensus sequences and the Khoi-San consensus sequence are indicated in Table 5.67.

Table 5.67 Comparison between the Khoi-San consensus sequence and two human consensus sequences constructed in previous investigations

Khoi-San consensus sequence	Marzuki consensus sequence ^a	Carter consensus sequence ^b
1048T	1048C	1048C
3423T	3423G	3423T
3516A	3516C	3516C
5442C	5442T	5442T
9042T	9042C	9042C
9347G	9347A	9347A
9559C	9559G	9559C
10664T	10664C	10664C
10915C	10915T	10915T
11335C	11335T	11335C
11447G	11447C	11447G
13276G	13276A	13276A
13702C	13702G	13702C
14199T	14199G	14199T
14272C	14272G	14272C
14365C	14365G	14365C
14368C	14368G	14368C

The numbering of the nucleotide positions follows the convention used for the rCRS (Andrews *et al.*, 1999). Nucleotide positions presented in blue indicate differences in the human consensus sequences compared to the Khoi-San consensus sequence. Nucleotide positions indicated in red identify nucleotide positions that are unique to the Khoi-San consensus sequence compared to the two human consensus sequences. ^a = human consensus sequence constructed by Marzuki *et al.* (1991). ^b = human consensus sequence compiled by Carter (2007). A = adenine; C = cytosine; G = guanine; T = thymine.

A total of 17 differences were observed between the Marzuki consensus sequence and the Khoi-San consensus sequence, while eight differences were observed upon comparison with the Carter consensus sequence. The 17 differences in the Marzuki consensus sequence can be ascribed to the use of only 13 sequences to compile the consensus sequence, of which the majority of sequences were obtained from individuals diagnosed with various neuromuscular disorders (Marzuki *et al.*, 1991). The eight differences in the Carter consensus sequence can be attributed to the presence of individuals from European and Asian ancestry in the sequence dataset used to compile the Carter human consensus sequence (Carter, 2007). It is interesting to note, that in the case of all eight differences observed in the Carter consensus sequence, the nucleotide at the position is identical to the nucleotide present in the rCRS which was obtained from a European individual (results not presented). This implies that although the Carter consensus sequence may be a good summary of the variation present in human mitochondrial genomes, in comparisons with mitochondrial sequences from individuals with African ancestry, alterations at these eight positions will be scored as potential candidates for pathogenicity even though the nucleotide present at the position may represent a normal variant. The Khoi-San consensus sequence is therefore preferred over the Carter consensus sequence for use in comparisons with mitochondrial sequences obtained from African individuals or persons with African ancestry. The small number of differences between the Khoi-San consensus sequence constructed from 43 Khoi-San individuals and the Carter consensus sequence compiled using a dataset consisting of 827 mitochondrial sequences obtained globally, also indicates that the Khoi-San sequence, although based on a small number of sequences from an ancient population in Africa, provides a good approximation of the variation present in human mitochondrial genomes worldwide.

Due to the characteristics of the Khoi-San consensus sequence discussed in the above paragraphs, it is suggested that the Khoi-San consensus sequence is used as an African reference sequence in the study of mitochondrial evolution and disease. At present, a single sequence from a Yoruban individual (NC_001807), classified in the current investigation within the L3a1 haplogroup (refer to Appendix G), is available from the NCBI reference sequence (RefSeq) database to be used as an African reference sequence. Another sequence from a Ugandan individual (D38112), classified in the current investigation as belonging to the L0c2 haplogroup (refer to Appendix G), can also be used as an African reference sequence according to MITOMAP (2008). As already discussed in Section 5.9.2.1, the African reference sequence (NC_001807) does not include the polymorphism data present in more ancient lineages such as L0, while the consensus

sequence compiled in this section is based on Khoi-San sequences which were classified as L0a or L0b. In order to determine which sequence would be more useful in comparisons of mitochondrial sequences to an African reference sequence, the number of differences between the rCRS, the Khoi-San consensus sequence and the two African reference sequences were determined and is presented in Table 5.68. The rCRS was used in this comparison as it is the current standard used for identification of polymorphisms, and thus provides a reliable standard to which all three sequences could be compared.

Table 5.68 Number of nucleotide alterations observed in comparisons between the rCRS, the Khoi-San consensus sequence and two African reference sequences

Mitochondrial sequence	Haplogroup/ sub-haplogroup designation	GenBank® accession number	Number of nucleotide differences compared to rCRS
African reference sequence (Yoruba)	L3a1	NC_001807	41
African reference sequence (Uganda)	L0c	D38112	105
Khoi-San consensus sequence	L0	---	127

Haplogroup or sub-haplogroup classification was assigned according to the coding region only classification scheme described in Section 4.9.

From Table 5.68, it can be concluded that the greatest number of differences were observed between the Khoi-San consensus sequence and the rCRS. It can thus be argued, compared to the two African reference sequences, the Khoi-San consensus sequence contains the highest level of variation. In addition, the Khoi-San consensus sequence can be viewed as a summary of the variation observed in the Khoi-San population specifically, and African populations in general. This implies that the Khoi-San consensus sequence will be the most useful standard in comparisons of mitochondrial genome sequences obtained from African individuals as it contains a large number of polymorphisms which are usually observed in African populations. The presence of a large number of normal variants in the Khoi-San consensus sequence therefore provides a template of a typical African mitochondrial background and will thus allow the identification of potential disease-related polymorphisms. It is therefore proposed that the Khoi-San consensus sequence is used as a tool to identify potential pathogenic alterations in mitochondrial genome sequences obtained from indigenous Africans, i.e. African individuals born and currently living in Africa, as well as individuals of African descent, i.e. individuals with African ancestry displaced to countries outside of Africa during the African Diaspora.

5.14 SUMMARY OF RESULTS GENERATED IN THE CURRENT INVESTIGATION

This section provides a summary of the most important results generated in the current investigation. The most important findings are summarised in sections named to correspond with the sections in which the results were presented.

5.14.1 Sequencing results

Complete mitochondrial genome sequences of 30 Khoi-San individuals collected from southern Africa (refer to Section 5.1) were determined after genomic DNA isolation (refer to Section 5.2) using the polymerase chain reaction (refer to Section 5.3), agarose gel electrophoresis (refer to Section 5.4), PCR product purification (refer to Section 5.5) and automated DNA sequencing (refer to Section 5.6). Optimised PCR conditions were described in Section 5.3.2, optimised automated sequencing conditions are listed in Section 5.6.2, and observed artefacts related to PCR (refer to Section 5.3.3), gel electrophoresis (refer to Section 5.4.1) and automated sequencing (refer to Section 5.6.4), as well as ways in which these artefacts were countered are discussed. None of the observed artefacts interfered with the final sequence data obtained.

A total of 292 alterations were observed in the 30 Khoi-San mitochondrial genome sequences upon comparison with the rCRS of which 272 were previously reported and 20 were novel. The distribution of the observed alterations in the 30 Khoi-San sequences are listed in Appendix A, and the molecular information for each of the alterations is listed in Appendix B. A summary of the number, type and report status of the alterations observed per locus and PCR region in the 30 complete Khoi-San mitochondrial genome sequences is provided in Table 5.69. A large number of alterations were expected as the rCRS is based on sequence data from a European individual belonging to haplogroup H, while the Khoi-San mtDNA sequences formed part of macrohaplogroup L.

Table 5.69 Summary of alterations observed in 30 complete Khoi-San mitochondrial genome sequences

PCR region	Region/ Gene	No. of alterations			Type of alteration				Amino acid change		
		Reported	Novel	Total	DEL	INS	TS	TV	NC	SYN	NON-SYN
1	D-loop	68	1	69	3	5	55	6	69	---	---
	tRNA ^{Phe}	1	1	2	0	0	2	0	2	---	---
	12S rRNA	12	2	14	0	0	13	1	14	---	---
	Sub-Total	81	4	85	3	5	70	7	85	---	---

Table 5.69 Continued...

PCR region	Region/ Gene	No. of alterations			Type of alteration				Amino acid change		
		Reported	Novel	Total	DEL	INS	TS	TV	NC	SYN	NON-SYN
2	16S rRNA	11	1	12	0	0	11	1	12	---	---
	ND1	17	0	17	0	0	15	2	---	15	2
	Sub-Total	28	1	29	0	0	26	3	12	15	2
3	tRNA ^{Cys}	2	0	2	0	0	2	0	2	---	---
	tRNA ^{Trp}	0	1	1	0	0	1	0	1	---	---
	tRNA ^{Gln}	1	0	1	0	0	1	0	1	---	---
	tRNA ^{Ile}	1	0	1	0	0	1	0	1	---	---
	ND2	12	1	13	0	0	12	1	---	8	5
	Sub-Total	16	2	18	0	0	17	1	5	8	5
4	tRNA ^{Tyr}	0	1	1	0	0	0	1	1	---	---
	COI	17	2	19	0	0	18	1	---	15	4
	tRNA ^{Ser(UCN)}	1	0	1	0	0	1	0	1	---	---
	Sub-Total	18	3	21	0	0	19	2	2	15	4
5	COII	11	2	13	0	0	9	4	---	10	3
	ATPase8	3	0	3	0	0	3	0	---	2	1
	ATPase6	14	2	16	0	0	16	0	---	12	4
	COIII	10	0	10	0	0	10	0	---	9	1
	ND3	3	1	4	0	0	3	1	---	1	3
	Sub-Total	41	5	46	0	0	41	5	0	34	12
6	ND4L	4	0	4	0	0	4	0	---	4	0
	ND4	22	1	23	0	0	23	0	---	20	3
	Sub-Total	26	1	27	0	0	27	0	0	24	3
7	tRNA ^{His}	1	0	1	0	0	1	0	1	---	---
	tRNA ^{Ser(AGY)}	2	0	2	0	0	2	0	2	---	---
	tRNA ^{Leu(CUN)}	1	0	1	0	0	1	0	1	---	---
	ND5	26	3	29	0	0	28	1	---	16	13
	ND6	12	1	13	0	0	13	0	---	9	4
	Sub-Total	42	4	46	0	0	45	1	4	25	17
8	tRNA ^{Glu}	1	0	1	0	0	1	0	1	---	---
	tRNA ^{Thr}	4	0	4	0	0	4	0	4	---	---
	Cytb	15	0	15	0	0	15	0	---	10	5
	Sub-Total	20	0	20	0	0	20	0	5	10	5
Total		272	20	292	3	5	265	19	113	131	48

Nucleotide abbreviations: A = adenine; C = cytosine; G = guanine; U = uracil; Y = C or U; N = A, C, G or U. Gene abbreviations: 12S rRNA = 12S ribosomal RNA; 16 rRNA = 16S ribosomal RNA; ATPase6 = ATP synthase F₀ subunit 6; ATPase8 = ATP synthase F₀ subunit 8; COI = cytochrome c oxidase subunit I; COII = cytochrome c oxidase subunit II; COIII = cytochrome c oxidase subunit III; Cytb = cytochrome b; D-loop = control region, including displacement loop; ND = NADH dehydrogenase subunit 1; ND2 = NADH dehydrogenase subunit 2; ND3 = NADH dehydrogenase subunit 3; ND4 = NADH dehydrogenase subunit 4; ND4L = NADH dehydrogenase subunit 4L; ND5 = NADH dehydrogenase subunit 5; ND6 = NADH dehydrogenase subunit 6; tRNA = transfer RNA; tRNA^{Cys} = tRNA cysteine; tRNA^{Gln} = tRNA glutamine; tRNA^{Glu} = tRNA glutamic acid; tRNA^{His} = tRNA histidine; tRNA^{Ile} = tRNA isoleucine; tRNA^{Leu(CUN)} = tRNA leucine (CUN); tRNA^{Phe} = tRNA phenylalanine; tRNA^{Ser(AGY)} = tRNA serine (AGY); tRNA^{Ser(UCN)} = tRNA serine (UCN); tRNA^{Thr} = tRNA threonine; tRNA^{Trp} = tRNA tryptophan; tRNA^{Tyr} = tRNA tyrosine. Other abbreviations: No. = number; DEL = deletion; INS = insertion; TS = transition; TV = transversion; NC = non-coding; SYN = synonymous substitution; NON-SYN = non-synonymous substitution.

A number of the reported alterations observed in the current investigation were previously classified in other investigations as possible pathogenic alterations such as G9438A (refer

to Section 5.7.5.4), A10398G (refer to Section 5.7.5.5), A14693G and A15951G (refer to Section 5.7.8.1). In the current investigation, these alterations were however observed in an “apparently healthy” Khoi-San cohort thereby indicating that these alterations are not likely to be pathogenic, but are more likely to be polymorphisms present in ancient African populations.

As previously mentioned, a total of 20 previously unreported or novel alterations were observed in the Khoi-San mitochondrial genome sequences. The novel alterations are listed in Table 5.70. The possible functional implication of each novel alteration is discussed in the section indicated. It appeared that the majority of novel alterations were not detrimental or pathogenic, and a functional effect is therefore not anticipated. An exception was the novel alteration present at nucleotide position 8,214 in COII, as a possible detrimental effect could not be excluded for this alteration (refer to Section 5.11.5.2).

Table 5.70 Novel alterations observed in 30 Khoi-San mitochondrial genome sequences

Alteration	Khoi-San individual	Haplogroup classification	Locus	Alteration type	Amino acid change	Functional effect
A647G	KS_CGR_140 KS_CGR_143 KS_CGR_175	L3a1	MT-TF	TS	Non-coding	Section 5.7.1.2 None
C957T	KS_CGR_048 KS_CGR_060	L0b	MT-RNR1	TS	Non-coding	Section 5.7.1.3 None
T1386C	KS_CGR_044	L0b	MT-RNR1	TS	Non-coding	Section 5.7.1.3 None
A2363G	KS_CGR_176	L0a	MT-RNR2	TS	Non-coding	Section 5.7.2.1 None
C4771T	KS_CGR_120	L0a	MT-ND2	TS	NON-SYN: Ala → Val	Section 5.7.3.2 None
A5515G	KS_CGR_044	L0b	MT-TW	TS	Non-coding	Section 5.7.3.1 None
T5865A	KS_CGR_164	L2	MT-TY	TV	Non-coding	Section 5.7.4.1 None
T5986C	KS_CGR_164	L2	MT-CO1	TS	NON-SYN: Val → Ala	Section 5.7.4.2 None
C6344G	KS_CGR_164	L2	MT-CO1	TV	NON-SYN: Ile → Met	Section 5.7.4.2 None
T8214A	KS_CGR_105	L0b	MT-CO2	TV	NON-SYN: Val → Asp	Section 5.11.5.2 Possible
T8223G	KS_CGR_105	L0b	MT-CO2	TV	NON-SYN: Leu → Trp	Section 5.7.5.1 None

Table 5.70 Continued...

Alteration	Khoi-San individual	Haplogroup classification	Locus	Alteration type	Amino acid change	Functional effect
C8922T	KS_CGR_001	L0a	ATPase6	TS	SYN	Section 5.7.5.3 None
C8970T	KS_CGR_048 KS_CGR_060	L0b	ATPase6	TS	SYN	Section 5.7.5.3 None
G10098T	KS_CGR_140 KS_CGR_141 KS_CGR_175	L3a1	MT-ND3	TV	NON-SYN: Ala → Ser	Section 5.7.5.5 None
C11728T	KS_CGR_173	L2	MT-ND4	TS	SYN	Section 5.7.6.2 None
C13347T	KS_CGR_005	L0b	MT-ND5	TS	SYN	Section 5.7.7.2 None
T13386C	KS_CGR_044	L0b	MT-ND5	TS	SYN	Section 5.7.7.2 None
G13466C	KS_CGR_044	L0b	MT-ND5	TS	NON-SYN: Ser → Thr	Section 5.7.7.2 None
T14302C	KS_CGR_044	L0b	MT-ND6	TS	SYN	Section 5.7.7.3 None
G16474T	KS_CGR_056	L0b	MT-DLOOP	TS	Non-coding	Section 5.7.1.1 None

Nucleotides are represented by the capital letters, with the rCRS nucleotide indicated first, followed by the observed change. Nucleotide numbers are based on rCRS numbering (Andrews *et al.*, 1999). A = adenine; C = cytosine; G = guanine; T = thymine; ATPase6 = ATP synthase F₀ subunit 6; MT-CO1 = cytochrome *c* oxidase subunit I; MT-CO2 = cytochrome *c* oxidase subunit II; MT-DLOOP = control region, including displacement loop; MT-ND2 = NADH dehydrogenase subunit 2; MT-ND3 = NADH dehydrogenase subunit 3; MT-ND4 = NADH dehydrogenase subunit 4; MT-ND5 = NADH dehydrogenase subunit 5; MT-ND6 = NADH dehydrogenase subunit 6; MT-RNR1 = 12S ribosomal RNA; MT-RNR2 = 16S ribosomal RNA; MT-TF = tRNA phenylalanine; MT-TW = tRNA tryptophan; MT-TY = tRNA tyrosine; TV = transversion; TS = transition; NON-SYN = non-synonymous substitution; SYN = synonymous substitution; Ala = alanine; Asp = aspartic acid; Ile = isoleucine; Leu = leucine; Met = methionine; Ser = serine; Thr = threonine; Trp = tryptophan; Val = valine.

5.14.2 Mitochondrial haplogroup analysis

Mitochondrial DNA sequences were classified as L0, L2 or L3 using a classification scheme based on SNPs identified in the coding region (refer to Section 4.9). Haplogroups were assigned to the mitochondrial genome sequences of the 30 Khoi-San individuals studied in the current investigation (refer to Section 5.8.1), as well as to a number of L-specific sequences obtained from other sources and used in various analyses (refer to Section 5.8.2). In addition, the 30 complete Khoi-San mtDNA genome sequences, as well as the mtDNA genome sequences used in the construction of the global L0-specific haplogroup network, were compared to a previously constructed global mtDNA haplogroup tree available from MITOMAP (Ruiz-Pesini *et al.*, 2007), as well as the exclusion criteria for L-specific sequences (refer to Section 5.8). This was done to ensure that the sequences fit the criteria for L-specific haplogroup designation.

It was observed that a number of the 30 Khoi-San sequences, as well as some of the mtDNA sequences included in the global L0-specific haplogroup network, did not strictly adhere to the exclusion criteria for L-specific sequences described in Appendix D (refer to Section 5.8). The criteria that were deviated from, as well as the number of sequences in which the specific deviation was observed, are presented in Table 5.71.

Table 5.71 Exclusion criteria for L-specific sequences from which Khoi-San genome sequences and sequences obtained from GenBank® deviated

Nucleotide substitution	Exclusion criteria	Number of sequences
G769A	L0, L1, and L2 from other haplogroups	Absent in a single L-specific sequence classified in the L0aA group within the L0a sub-haplogroup: EF184610
A1438G	From sub-haplogroup H containing the CRS	Absent in 15 Khoi-San individuals classified within the L0b sub-haplogroup: KS_CGR_004 , KS_CGR_005 , KS_CGR_044 , KS_CGR_048 , KS_CGR_056 , KS_CGR_060 , KS_CGR_061 , KS_CGR_077 , KS_CGR_091 , KS_CGR_092 , KS_CGR_105 , KS_CGR_111 , KS_CGR_117 , KS_CGR_121 , KS_CGR_127 Absent in all 27 of the L-specific mitochondrial genome sequences classified within the L0aB group in the L0a sub-haplogroup and the L0bA and L0bB groups within the L0b sub-haplogroup: AY195777 , DQ112830 , DQ112844 , DQ112850 , DQ112852 , DQ112854 , DQ112855 , DQ112856 , DQ112899 , EF184585 , EF184586 , EF184587 , EF184588 , EF184590 , EF184592 , EF184593 , EF184594 , KS_CGR_010 , KS_CGR_032 , KS_CGR_040 , KS_CGR_052 , KS_CGR_072 , KS_CGR_082 , KS_CGR_094 , KS_CGR_102 , KS_CGR_122 , KS_CGR_130
A2706G	From haplogroup H	Absent in 14 Khoi-San individuals classified within the L0b sub-haplogroup: KS_CGR_004 , KS_CGR_005 , KS_CGR_048 , KS_CGR_056 , KS_CGR_060 , KS_CGR_061 , KS_CGR_077 , KS_CGR_091 , KS_CGR_092 , KS_CGR_105 , KS_CGR_111 , KS_CGR_117 , KS_CGR_121 , KS_CGR_127 Absent in 21 L-specific mitochondrial genome sequences classified within the L0bB group within the L0b sub-haplogroup: DQ112844 , DQ112852 , DQ112854 , DQ112855 , DQ112856 , DQ112899 , EF184586 , EF184590 , EF184592 , EF184593 , EF184594 , KS_CGR_010 , KS_CGR_032 , KS_CGR_040 , KS_CGR_052 , KS_CGR_072 , KS_CGR_082 , KS_CGR_094 , KS_CGR_102 , KS_CGR_122 , KS_CGR_130
A4769G	From sub-haplogroup H containing the CRS	Absent in a single Khoi-San individual classified within the L0b sub-haplogroup: KS_CGR_044
T6185C	L0 from L1	Absent in a single Khoi-San individual classified within the L0b sub-haplogroup: KS_CGR_091
A7146G	L0 and L1 from L2, L3 and M	Absent in a single L-specific sequence classified in the L0c2D group within the L0c2 sub-haplogroup: AF346998
A10398G	From haplogroup N	Absent in a single Khoi-San individual classified within the L0bB group in the L0b sub-haplogroup: KS_CGR_052
G12007A	L0 from L1	Absent in a single L-specific sequence classified in the L0aB group within the L0a sub-haplogroup: EF184585

Nucleotide numbers are based on rCRS numbering (Andrews *et al.*, 1999). A = adenine; C = cytosine; G = guanine; T = thymine; CRS = Cambridge Reference Sequence. Khoi-San (indicated in [blue](#)) and other L-specific mitochondrial genome sequences (indicated in [red](#)) included in the global L0-specific haplogroup network were compared to the exclusion criteria for L-specific sequences listed in Appendix D.

A number of the Khoi-San sequences generated in the current investigation, as well as a number of the other L-specific sequences obtained from GenBank® and included in the global L0-specific haplogroup network, did not strictly adhere to the exclusion criteria for L-specific sequences. The exclusion criteria should therefore be revised to take these discrepancies into account. The list of exclusion criteria for L-specific sequences should only include those SNPs which are observed in all L-specific sequences. This implies that A2706G and A1438G should be excluded from the list as these SNPs are absent in many sequences classified within the L0a and L0b sub-haplogroups. SNPs such as G769A, A4769G, T6185C, A7146G, A10398G and G12007A can however be included as to date these SNPs are absent in single sequences only. The absence of these SNPs can be attributed to rare reversion events (refer to Section 5.8).

Haplogroup and sub-haplogroup designations were assigned to the 30 complete Khoi-San mitochondrial DNA genome sequences (refer to Section 5.8.1), as well as an additional 403 mitochondrial genome sequences which consisted of 390 sequences retrieved from GenBank® and 13 Khoi-San sequences obtained from a previous investigation (refer to Section 5.8.2). A summary of the number of sequences assigned to each of the haplogroups and sub-haplogroups is presented in Table 5.72.

A total of 96 sequences were assigned to L0, 56 sequences to L1, 115 sequences to L2, 94 sequences to L3, two sequences to a haplogroup other than L (macrohaplogroup M), and 70 sequences were classified incompletely. The two sequences assigned to macrohaplogroup M were excluded from subsequent phylogenetic and statistical analyses (refer to Section 4.10), while some sequences belonging to haplogroups L0a\$, L0b\$, L0c2\$ and L2a were later re-classified using a revised classification scheme owing to discrepancies observed between the haplogroup classification and the location of the sequences in the phylogenetic trees (refer to Section 5.12).

Table 5.72 Assignment of haplogroups to mitochondrial genome sequences included in the current investigation

Haplogroup/ sub-haplogroup classification	Number of sequences	Haplogroup/ sub-haplogroup classification	Number of sequences	Haplogroup/ sub-haplogroup classification	Number of sequences
Other^a	2	L2	115	Incompletely classified^b	70
M	2	L2a	87	L0\$a	6
L0	96	L2b	16	L0\$b	4
L0a	21	L2c	11	L0,1,2	1
L0b	40	L2d	1	L0,L1	4
L0c1	14	L3	94	L0c2\$	1
L0c2	21	L3a1	40	L1b2\$	3
L1	56	L3a2	10	L2d\$	4
L1a1	1	L3a3	3	L3a3\$	9
L1a2	33	L3b	22	L3b\$	37
L1b1	3	L3c	19	Not assigned^c	1
L1b2	19	---	---	EF184609	1

^a = haplogroups other than macrohaplogroup L. ^b = incompletely classified sequences were indicated by a dollar sign (\$) in the putative haplogroup classification (refer to Section 5.8). ^c = a haplogroup classification was not assigned to this sequence due to the presence of an A nucleotide at position 3,594 (refer to Section 4.10.1).

The mitochondrial haplogroups assigned to the 30 complete Khoi-San mitochondrial DNA genome sequences are presented in Section 5.8.1. The haplogroup results for the 30 Khoi-San individuals are also summarised in Appendix E, where additional results regarding the sub-haplogroup of each individual are presented. As was expected based on their recent African ancestry, all 30 Khoi-San individuals grouped within macrohaplogroup L. Twenty-two individuals were classified as L0 (L0a and L0b), three individuals as L2 (L2b) and five individuals as L3 (L3a1 and L3b). The presence of haplogroups other than L0 in the Khoi-San cohort was attributed to the combination of three factors, namely the occurrence of admixture, the availability of limited data regarding ethnicity of previous generations, as well as self-reporting of ethnicity (refer to Section 5.8.1). A recent investigation of the maternal lineages present in the Khoi-San provided support for the suggestion that the presence of L2 and L3 haplogroups in the Khoi-San cohort was the result of recent gene flow (Behar *et al.*, 2008).

A total of 70 sequences could not be classified completely using the classification scheme employed in the current investigation, i.e. the sequences could not be classified to either the haplogroup or sub-haplogroup level. This was due to the absence of certain haplogroup or sub-haplogroup defining SNPs. The incompletely classified sequences, with the exception of certain sequences (refer to Appendix H), were included in both the phylogenetic and statistical analyses, as well as in the construction of the global

L0-specific haplogroup network. The reasons for the inclusion or exclusion of these sequences from the sequence datasets used are listed in the relevant sections (refer to Sections 4.10 and 5.9).

5.14.3 Phylogenetic analysis

Phylogenetic analysis was performed as described in Section 5.9 with the aim to determine the evolutionary relationships of the 30 Khoi-San sequences generated in the current investigation with each other, as well as with additional sequences belonging to macrohaplogroup L. Genome sequences consisting of the coding region only (bp 577 to 16,023) were used to generate the phylogenetic trees (refer to Section 5.9.1), and the databases used to construct the various phylogenetic trees were described in Section 4.10, while the information regarding the individual sequences is provided in Appendix G. The major findings regarding the phylogeny of the 30 Khoi-San sequences are presented in this section according to the results obtained from the global phylogenetic trees.

The most basal lineage of the global phylogenetic trees consisted of a group of L5 sequences, followed by a group consisting of L0/L1 sequences, and a group containing L2/L3 sequences. This finding in the current investigation is in contrast to previous studies in which L0 was reported as the most basal haplogroup of the human mtDNA tree. This discrepancy could be attributed to a number of factors (refer to Section 5.9.2.1). Due to the rarity of sequences belonging to the L5 haplogroup and the paucity of sequence data from eastern Africa, the true evolutionary relationship of L5 with other L haplogroups has been obscured, and the results obtained in the current investigation may therefore support a previously unreported close relationship between L0 and L5 (refer to Section 5.9.2.1). In addition, the observation of the replacement of L0 by L5 as the most basal group in the current investigation could be due to the fact that the L5 haplogroup is a very ancient haplogroup. The addition of more sequences belonging to haplogroup L5, as well as more sequences from East Africa, to the human mtDNA phylogenetic tree should however provide further elucidation of the relationship between L0 and L5.

The L0* haplogroup consisted of two main groups, the first contained sequences classified as L0a and L0b, and the second consisted of sequences classified as L0a, L0c1 and L0c2. The L0a group was further divided into L0aA and L0aB groups, while the L0b group was divided into two groups, L0bA and L0bB. Twenty-two of the 30 Khoi-San sequences

generated in the current investigation grouped within the L0* group. Seven of the 22 Khoi-San sequences were classified as L0aA, while the remaining 15 Khoi-San sequences were classified as L0b. None of the Khoi-San sequences generated in the current investigation were classified as either L0c1 or L0c2. The addition of the Khoi-San sequences generated in the current investigation to the global phylogenetic trees led to the expansion of the L0aA branch by generating additional subgroups namely, L0aA1 and L0aA2, as well as the expansion of the L0bB branch to generate subdivisions, namely L0bB1, L0bB2, L0bB2a, L0bB2b and L0bB3 (refer to Section 5.9.2.1).

Three Khoi-San sequences were classified as L2b, three Khoi-San sequences were assigned to L3a1, and two Khoi-San sequences were classified within the L3b group. In each of these sub-haplogroups, the Khoi-San sequences increased the resolution of the phylogenetic trees by increasing the number of branches within the group. In addition, the Khoi-San sequences formed a branch consisting of Khoi-San sequences only in each of these sub-haplogroups, and based on the cohort investigated in the current investigation, these branches may represent lineages or sub-lineages of L2b, L3a1 and L3b that are confined to southern Africa (refer to Section 5.9.4).

The location of each of the 30 Khoi-San sequences generated in the current investigation in the phylogenetic trees is determined according to haplogroup classification. The location of all 30 Khoi-San sequences is summarised in Table 5.73.

Table 5.73 Location of 30 Khoi-San sequences generated in the current investigation in the phylogenetic trees

Sequence name	Haplogroup classification ^a	Sequence name	Haplogroup classification	Sequence name	Haplogroup classification
KS_CGR_001	L0aA	KS_CGR_077	L0b	KS_CGR_121	L0b
KS_CGR_004	L0b	KS_CGR_083	L0aA	KS_CGR_127	L0b
KS_CGR_005	L0b	KS_CGR_084	L0aA	KS_CGR_140	L3a1
KS_CGR_044	L0b	KS_CGR_090	L3b	KS_CGR_143	L3a1
KS_CGR_048	L0b	KS_CGR_091	L0b	KS_CGR_157	L0aA
KS_CGR_056	L0b	KS_CGR_092	L0b	KS_CGR_164	L2b
KS_CGR_059	L0aA	KS_CGR_105	L0b	KS_CGR_172	L3b
KS_CGR_060	L0b	KS_CGR_111	L0b	KS_CGR_173	L2b
KS_CGR_061	L0b	KS_CGR_117	L0b	KS_CGR_175	L3a1
KS_CGR_069	L2b	KS_CGR_120	L0aA	KS_CGR_176	L0aA

^a = the haplogroup classification assigned to the Khoi-San sequence according to the classification scheme used in the current investigation (refer to Section 4.9). The colours in which the haplogroup/sub-haplogroup classifications are presented correspond to the colours used to indicate the different branches in the phylogenetic trees.

Although the focus of the current investigation was on the evolutionary relationship of the 30 Khoi-San sequences with one another and other L-specific sequences, and the majority of the Khoi-San sequences were grouped within the L0 branch of the phylogenetic trees, the L1, L2 and L3 haplogroups were also discussed (refer to Section 5.9.2.1). A number of branches were observed in the global phylogenetic trees within these haplogroups for which no designation is available when using the haplogroup classification scheme employed in the current investigation (refer to Section 4.9), or when using the alternative classification scheme (refer to Section 3.3.1.1). For the purpose of the discussion of the global phylogenetic trees in the current investigation these branches were assigned names as necessary, namely L1a2A, L1a2B, L1a2C, L1b3, L2a1, L2a2, L2a3, L2b1, L2b2 and L2b3 (refer to Section 5.9.2.1). In addition, the discussion of the L1, L2 and L3 haplogroups led to the conclusion that these sections of the classification scheme used in the current investigation (refer to Section 4.9) should be revised to include any new sub-haplogroups observed in the haplogroups or sub-haplogroups, and to incorporate the designations provided by the alternative classification scheme for groups consisting of unclassified sequences in order to completely classify all the sequences included in the phylogenetic analyses (refer to Section 5.9.2.1).

The overall branching structure of the phylogenetic trees were similar, irrespective of the dataset used (dataset 1a, 2a, 3a or 3b), or method employed to construct the tree (N-J or MP method). A few differences were however observed between the trees constructed using the different datasets and phylogenetic methods (refer to Sections 5.9.2 and 5.9.3). These differences were mainly related to the grouping of individual sequences within the same haplogroup or sub-haplogroup and are summarised in Table 5.74.

Table 5.74 Differences observed in comparisons of phylogenetic trees constructed using different datasets and methods

Phylogenetic tree	Tree compared to	Haplogroup	Difference
Global N-J	Africa N-J	L0c1	DQ112959 formed a group on its own vs. DQ112959 included in L0c1B.
		L2a	L2a2 group present vs. L2a2 group absent.
		L2d	L2d group present vs. L2d group absent.
		L3	DQ341074 present outside L3b and L3c groups, forming its own group vs. DQ341074 adjacent to the L3c group.
		L3a	L3a2A and L3a2B (subdivision present in L3a2) vs. L3a2 (no subdivision in group).
		L3b	DQ112756 groups with DQ112713 vs. DQ112756 groups with DQ112949.
		L5	L5a1 groups with L5a2 vs. L5a1 groups with L5b.

Table 5.74 Continued...

Phylogenetic tree	Tree compared to	Haplogroup	Difference
Global N-J	Global MP	L0	L0aA groups with L0aB and L0b vs. L0aA groups with L0aC, L0c1 and L0c2.
		L0a	KS_CGR_059 forms own group outside L0aA1 vs. KS_CGR_059 included in L0aA1.
		L0b	Three groups present, L0bB1, L0bB2 and L0bB3 vs. two groups present, L0bB1 plus L0bB3, and L0bB2.
		L0c1	DQ112959 forms own group vs. DQ112959 included in L0c1B group.
		L0c2	D38112 forms own group vs. D38112 groups with EF184602 and EF184604.
		L1a	DQ112737 forms own group vs. DQ112737 groups with DQ112741 and DQ112759.
		L2a	EF184620 forms own group vs. EF184620 groups with other sequences in L2a3 group.
		L2b	KS_CGR_173 and KS_CGR_069 form a separate group vs. KS_CGR_173 and KS_CGR_069 group within a group with four other sequences.
		L3	DQ341081, EF184625 and EF184630 group adjacent to six incompletely classified sequences vs. DQ341081, EF184625 and EF184630 group adjacent to group containing rCRS.
		L3b	DQ112949 groups in L3b1 vs. DQ112949 groups in L3b2.
	Khoi-San N-J	L0b	L0bA and L0bB present vs. L0bA group only.
Africa N-J	Africa MP	L0	L0aA groups with L0aB and L0b vs. L0aA groups with L0aC, L0c1 and L0c2.
		L0a	KS_CGR_059 forms own group outside L0aA1 vs. KS_CGR_059 included in L0aA1.
		L0b	Three groups present, L0bB1, L0bB2 and L0bB3 vs. two groups present, L0bB1 plus L0bB3 and L0bB2.
		L0c1	EF184601 forms own group vs. EF184601 included in L0c1A.
		L0c2	D38112 forms own group vs. D38112 groups with EF184602 and EF184604.
		L5	L5a1 groups with L5b vs. L5a1 groups with L5a2.
Khoi-San N-J	Khoi-San MP	L0a	EF184611 forms own group vs. EF184611 included in L0aA1.
		L0b	L0bB1 and L0bB2 group together vs. L0bB1 and L0bB3 group together.

Global N-J tree = phylogenetic tree constructed using the N-J method and dataset 1a (refer to Section 5.9.2.1), Global MP tree = phylogenetic tree constructed using the MP method and dataset 1a (refer to Section 5.9.3.1), Africa N-J tree = phylogenetic tree constructed using the N-J method and dataset 2a (refer to Section 5.9.2.2), Africa MP tree = phylogenetic tree constructed using the MP method and dataset 2a (refer to Section 5.9.3.2), Khoi-San N-J tree = phylogenetic tree constructed using the N-J method and dataset 3a (refer to Section 5.9.2.3), Khoi-San MP tree = phylogenetic tree constructed using the MP method and dataset 3a (refer to Section 5.9.3.3). The colours in which the haplogroup/sub-haplogroup classifications are presented correspond to the colours used to indicate the different branches in the phylogenetic trees.

The differences observed between the phylogenetic trees can be attributed to differences in the composition of the sequence datasets used for tree construction (sequences included in dataset 1a or 2a which were excluded from dataset 2a or 3a), or to the

unresolved branching structures within the specific haplogroups/sub-haplogroups as suggested by low bootstrap values (refer Section 5.9.2.2). The addition of more sequences from the specific haplogroup/sub-haplogroup, or from the specific region in Africa from which the other sequences in the affected group were obtained, should provide better resolution (refer Section 5.9.2.2).

Overall, the MP trees proved to be better suited to investigating the evolutionary relationships between the Khoi-San sequences and the different haplogroups and sub-haplogroups. This was due to the fact that both the MP method and the haplogroup classification scheme used in the current investigation (refer to Section 4.9) are based on nucleotide differences. Therefore, the branching order in the MP trees corresponded closely to the actual haplogroup and sub-haplogroup divisions (refer to Section 5.9.3.1). MP trees were thus employed in subsequent analyses relating to haplogroup classification such as the geographic distribution of African mitochondrial haplogroups (refer to Section 5.9.4), the construction of the L0-specific haplogroup network (refer to Section 5.10) and the revision of the classification scheme for the L0 haplogroup (refer to Section 5.12).

5.14.4 Global L0-specific haplogroup network

A global L0-specific haplogroup network was constructed using the L0 haplogroup section of a sequential mutational phylogenetic tree available on MITOMAP (Ruiz-Pesini *et al.*, 2007), which consisted of 54 L0 mtDNA genome sequences, and expanding it manually through the addition of a total of 46 L0 sequences so that the newly constructed global L0 network consisted of a total of 100 sequences (refer to Section 5.10). The 46 L0 sequences added to the global L0 network included the 22 Southern African Khoi-San L0 mtDNA sequences generated in the current investigation, as well as another 24 recently published L0 sequences (Torroni *et al.*, 2006; Gonder *et al.*, 2007; Just *et al.*, 2008). The global L0-specific haplogroup network was constructed with the aim to define sub-haplogroups by indicating which polymorphisms were common in L0 sub-haplogroups, to estimate the ages of sub-haplogroups and/or lineages within the L0 haplogroup, to identify new branching and new lineages created by the addition of new L0 sequences to the network, to determine if a branch is geographic and/or population-specific, to differentiate between ancient and recent mutations, and to assist in determining which polymorphisms were likely to be pathogenic (refer to Section 5.10).

The 22 Khoi-San mitochondrial sequences grouped in the L0a and L0b branches of the global L0-specific haplogroup network. The location of the individual Khoi-San sequences in the global L0-specific haplogroup network is listed in Table 5.75.

Table 5.75 Location of Khoi-San sequences in global L0-specific haplogroup network

Branch/group	Khoi-San individual
L0aA1	KS_CGR_059, KS_CGR_084, KS_CGR_120
L0aA2	KS_CGR_001, KS_CGR_083, KS_CGR_157, KS_CGR_176
L0bA	KS_CGR_044
L0bB1	KS_CGR_048, KS_CGR_056, KS_CGR_060
L0bB2	KS_CGR_004, KS_CGR_005, KS_CGR_077, KS_CGR_091, KS_CGR_092, KS_CGR_105, KS_CGR_111, KS_CGR_117, KS_CGR_121
L0bB3	KS_CGR_061, KS_CGR_127

A number of new internal and terminal branches were observed in the newly constructed global L0-specific haplogroup network (refer to Section 5.10.2). A summary of the locations of the new branches and the sequences present in the new branches is presented in Table 5.76.

Table 5.76 Summary of the locations of new internal and terminal branches observed in the global L0-specific haplogroup network

Branch/ group	No. of new branches		Colour ^a	Sequence names
	I	T		
L0aA	1	5	Pink	EF184610 KS_CGR_001, KS_CGR_059, KS_CGR_083, KS_CGR_084, KS_CGR_157, KS_CGR_120, KS_CGR_176
L0aB	2	3	Orange	EF184585, EF184587, EF184588
L0aC	0	1	Violet	EF184597
L0bA	0	1	Pink	KS_CGR_044
L0bB1	1	2	Orange	EF184586, EF184590 KS_CGR_048, KS_CGR_060
L0bB2	1	8	Light blue	EF184592, EF184593, EF184594 KS_CGR_004, KS_CGR_005, KS_CGR_077, KS_CGR_091, KS_CGR_092, KS_CGR_105, KS_CGR_111, KS_CGR_117, KS_CGR_121
L0bB2a1	0	1	Light blue	KS_CGR_105
L0bB2a2	0	1	Light blue	EF184592
L0bB2a3	1	4	Light blue	EF184594 KS_CGR_117, KS_CGR_091, KS_GCR_092, KS_CGR_111
L0bB2b	0	2	Light blue	EF184593, KS_CGR_005
L0bB3	1	2	Violet	KS_CGR_061, KS_CGR_127

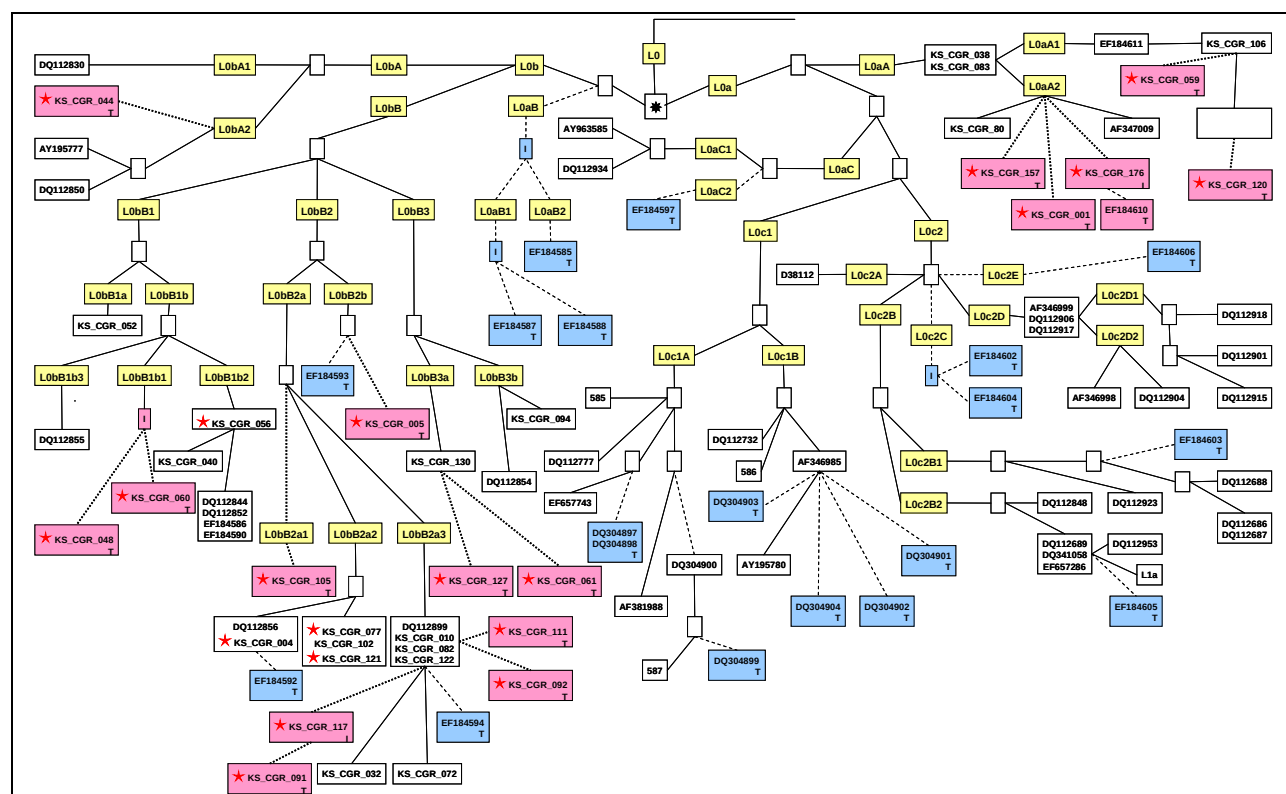
Table 5.76 Continued...

Branch/ group	No. of new branches		Colour ^a	Sequence names
	I	T		
L0c1A	0	3	Orange	DQ304897, DQ304898, DQ304899, DQ304900
L0c1B	0	4	Pink	DQ304901, DQ304902, DQ304903, DQ304904
L0c2B	0	2	Pink	DQ341058 EF184603, EF184605
L0c2C	1	2	Orange	EF184602, EF184604
L0c2E	0	1	Violet	EF184606

^a = colour used to indicate new branch in Figures 5.50, 5.51, 5.52 and 5.53 (refer to Section 5.10.2). I = internal branch; T = terminal branch.

A schematic representation of the global L0-specific haplogroup network indicating the locations of new internal and terminal branches is presented in Figure 5.63. The new internal and terminal branches present in the global L0-specific haplogroup network represent new sub-haplogroups or lineages within current sub-haplogroups.

Figure 5.63 Schematic representation of the global L0-specific haplogroup network indicating new branches after the addition of 46 L0 sequences



The starting point of the network is indicated by *. The names of sequences are indicated at the bottom of the branch in **bold** text. GenBank® accession numbers are indicated where available. Sequences generated in the current investigation are indicated by red stars (★). New internal (I) and terminal branches (T) within the network are indicated by dashed lines, where big dashes (---) represent branches added by published sequences (blue boxes) and small dashes (.....) indicate branches added by sequences generated in the current investigation (pink boxes). The haplogroup/sub-haplogroup designation of the branches is indicated in yellow boxes, while empty boxes represent internal branches to which no sequences were assigned.

The branches and groups observed in the global L0-specific haplogroup network correspond to the haplogroups/sub-haplogroups identified in the L0 section of the phylogenetic trees (refer to Section 5.9). The addition of the 22 Khoi-San sequences generated in the current investigation to the L0a and L0b branches expanded previously identified branches (L0aA1, L0aA2, L0bA2, L0bB2a1, L0bB2a3, L0bB2b and L0bB3a) and gave rise to two previously unidentified branches (L0bB1b1 and L0bB2a1). The expanded branches and new branches generated by the addition of the Khoi-San sequences represent potential new lineages in the L0 haplogroup and are an indication of the genetic diversity present in the Southern African Khoi-San population. The addition of the sequence data generated for these 22 Khoi-San individuals to the global L0-specific haplogroup network therefore increases our understanding of the genetic diversity present in the L0 haplogroup (refer to Section 5.10.2).

The global L0-specific haplogroup network was used in the construction of a revised classification scheme for the L0 haplogroup (refer to Section 5.12). The network was evaluated to identify alterations that were specific to, or associated with, particular haplogroups or sub-haplogroups, as well as to exclude alterations that were not suitable for inclusion in the classification scheme. Alterations that were not suitable for use in the construction of the revised classification scheme included recurrent alterations (alterations which occurred in more than one sub-haplogroup, or in multiple branches within the same sub-haplogroup), reversions (alterations that underwent reversion to the character state present in the rCRS which did not conform to the state observed in all other L0 sequences contained in the same branch or lineage in the global L0-specific haplogroup network), and alterations that involved more than one character state at the same nucleotide position in different parts of the global L0-specific haplogroup network. The information regarding the number of recurrent alterations and reversions present in the global L0-specific haplogroup network are summarised in Table 5.77.

Table 5.77 Summary of information relating to recurrent alterations and reversions

	Recurrent alterations	Reversions
Type of alteration		
Transition (A↔G)	43	12
Transition (C↔T)	28	8
Transversion	2	1
Deletion	2	0
Total	75	21

Table 5.77 Continued...

	Recurrent alterations	Reversions
Location of alteration		
Polypeptide genes		
Synonymous substitution	44 ^a	13
Non-synonymous substitution	11 ^a	4
Non-coding regions		
Non-coding nucleotides	2	0
rRNA and tRNA genes	20	4
Total	75	21

^a = a single nucleotide alteration at position 8,566 affects both the ATPase6 (non-synonymous substitution) and ATPase8 (synonymous substitution) genes and the corrected total, after subtracting one synonymous substitution and one non-synonymous substitution, is therefore 75.

The 75 recurrent alterations were observed at 34 nucleotide positions, while the 21 reversions were observed at 20 nucleotide positions. Recurrent alterations and reversions included synonymous substitutions, non-synonymous substitutions and deletions, and occurred in polypeptide, tRNA and rRNA genes (refer to Section 5.10.2.3). A prevalence of A ⇌ G transitions were observed at positions where recurrent alterations or reversions occurred, and this was likely due to these positions in the mtDNA coding region performing a role as hypervariable sites (refer to Section 5.10.2.3). The largest number of recurrent alterations was observed in ND4, ND5 and ATPase6, while the largest number of reversions was present in ND4, ND5 and COII. The presence of a high number of recurrent alterations or reversions in these genes could be attributed to the effects of natural selection (refer to Section 5.10.2.3).

Different character states for the same nucleotide position were observed for two nucleotide positions, 2,245 and 4,562, in different sections of the global L0-specific haplogroup network. Nucleotide position 2,245 located in the 16S rRNA gene, presented with three different character states (2245G, 2245C and 2245A), while two character states (4562G and 4562C) were observed at nucleotide position 4,562 located in the ND2 gene. The information regarding these two nucleotide positions is summarised in Table 5.78.

Table 5.78 Summary of information relating to different character states observed at two nucleotide positions

Alteration	Location in Global L0 network	Alteration	Location in Global L0 network
2245G	L0aC, L0c1 and L0c2 branches	4562G	L0aC branch
2245C	L0c1B branch	4562C	L0bB1a branch
2245A	L0c2D1 branch	---	---

Nucleotide numbers are based on rCRS numbering (Andrews *et al.*, 1999). A = adenine; C = cytosine; G = guanine; T = thymine.

Based on the information presented in Table 5.78, it can be suggested that the different character states are associated with specific sub-haplogroups or lineages within haplogroup L0 (refer to Section 5.10.2.3).

5.14.5 Geographic distribution of African mitochondrial haplogroups

The geographic distribution in Africa of the major African mitochondrial haplogroups and sub-haplogroups was determined by using data regarding ethnicity and geographic origin that was available for the individual sequences (refer to Appendix G), the global MP tree (refer to Section 5.9.4), as well as the global L0-specific haplogroup network (refer to Section 5.10.2.2). The geographic distribution of each haplogroup was subsequently assigned according to the major geographic regions in Africa as indicated on the map of Africa presented in Appendix I. Two L0 sub-haplogroups, namely L0aA and L0b, were observed mainly in southern Africa in the Khoi-San population. The major findings regarding geographic distribution and ethnicity are summarised in Table 5.79.

Table 5.79 Geographic distribution of African mitochondrial haplogroups and sub-haplogroups

Haplogroup classification ^a	Sub-haplogroup classification ^a	Geographic distribution	Ethnic/language group
L0	L0a	East and southern Africa	Khoi-San speakers
	L0aA ^b	Southern Africa	Khoi-San
	L0aB ^b	East Africa	Sandawe
	L0aC ^b	East Africa	---
	L0b	Southern Africa	Southern African Khoi-San speakers
	L0bA ^b	Southern Africa	---
	L0bB1 ^b	Southern Africa	Khoi-San
	L0bB2 ^b	Southern Africa	Khoi-San
	L0bB3 ^b	Southern Africa	Khoi-San
	L0c1	West Africa	---
	L0c1A ^b	North Africa	---
	L0c1B ^b	West Africa	---
	L0c2	East and Central Africa	---
	L0c2A ^b	East Africa	---
	L0c2B ^b	East, Central and West Africa	---
	L0c2C ^b	East Africa	---
	L0c2D ^b	Central Africa	Pygmies
	L0c2E ^b	East Africa	---

Table 5.79 Continued...

Haplogroup classification ^a	Sub-haplogroup classification ^a	Geographic distribution	Ethnic/language group
L1	L1a	West Africa	---
	L1b	Central Africa	---
	L1b2	Central Africa	Pygmies
	L1c2 ^c	Central Africa	Pygmies
	L1b3 ^d	West and East Africa	---
L2	L2a1	Central Africa	Pygmies
	L2a2	West Africa	---
	L2a3	West Africa	---
	L2b	Southern Africa	---
	L2c	West Africa	---
	L2d	Dominican	---
	L2d ^c	East Africa	---
L3	L3* ^c	East Africa	---
	L3a1	West Africa	---
	L3a2	West Africa	---
	L3a3	Central Africa	---
	L3b	Southern and West Africa	---
	L3bd ^c	East Africa	---
	L3c	West Africa	---
	L3e5 ^c	West Africa	---
	L3f ^c	West Africa	---
	L3h ^c	East Africa	---
	L3i ^c	North and East Africa	---
	L3w ^c	East Africa	---
	L3x ^c	East Africa	---
L4	L4a ^c	East Africa	---
	L4g ^c	East Africa	---
L5	L5a1 ^c	East and Central Africa	Pygmies
	L5a2 ^c	East and Central Africa	---
	L5b ^c	East and Central Africa	Sandawe
L6	L6 ^c	East Africa	---

The asterisk (*) in the L3* haplogroup classification indicates that this haplogroup consists of a group of mtDNA lineages, which form part of this haplogroup, although the lineages can not be classified into any of the specific sub-haplogroups of this haplogroup (refer to Section 3.1.1.1). ^a = refers to the haplogroup classification assigned to a group according to the classification scheme used in the current investigation (refer to Section 4.9). ^b = refers to the haplogroup/sub-haplogroup designation assigned to a group according to the revised classification scheme proposed in Section 5.12. ^c = refers to the haplogroup classification assigned to the group based on the alternative classification scheme described in Section 3.3.1.1. ^d = a name proposed to describe a previously unnamed sub-haplogroup in haplogroups other than L0. The colours in which the haplogroup/sub-haplogroup classifications are presented correspond to the colours used to indicate the different branches in the phylogenetic trees. The geographic distribution of haplogroup L0 in Africa was illustrated in Figure 5.59.

These findings are influenced by characteristics of the specific sequence datasets employed in the analyses such as the number and haplogroup classification of sequences in the specific dataset, as well as the sampling method used to compile the sequence dataset (refer to Sections 5.9.4 and 5.10.2.2). The results therefore need to be confirmed

by the addition of more mitochondrial sequences to both the global phylogenetic trees and the global L0-specific haplogroup network. These additional mitochondrial sequences need to be obtained from all geographic regions in Africa, and through the use of a systematic sampling strategy in order to collect samples of all known haplogroups/sub-haplogroups present in all regions.

5.14.6 Statistical analysis

The presence of high levels of genetic variation in the Khoi-San population and the African population as a whole was clearly evident after evaluation of the sequence data generated for the 30 Khoi-San individuals, the construction of phylogenetic trees and the compilation of the global L0-specific haplogroup network. A number of different statistical analyses were therefore performed on different sequence datasets (refer to Section 4.10) in order to investigate the role of different processes in the generation of the observed genetic diversity (refer to Section 5.11). These statistics included measures of nucleotide composition and codon usage (refer to Section 5.11.1). In addition, the number of segregating sites, the nucleotide diversity and the average number of nucleotide differences were calculated for a number of different sequence datasets (refer to Section 5.11.2). These datasets were also investigated to determine if the mitochondrial genome sequences evolved according to the standard neutral theory of evolution (refer to Section 5.11.3). If deviation from the neutral theory of molecular evolution was observed, population size changes (refer to Section 5.11.4) and the effect of selection (refer to Section 5.11.5) were subsequently investigated to determine which of these two processes would provide a possible explanation for the observed deviation from neutrality.

5.14.6.1 Basic sequence statistics

Certain basic descriptive sequence statistics were calculated for datasets 4a and 4b which consisted of the 30 complete or coding region only Khoi-San mitochondrial genome sequences generated in the current investigation, as well as 13 complete or coding region only Khoi-San mitochondrial genome sequences determined previously. These statistics included measures of nucleotide composition, G + C content and codon usage (refer to Section 5.11.1).

The relative frequencies of the four nucleotides in the 43 Khoi-San complete mtDNA genome sequences in dataset 4a were determined as 30.9% A, 31.2% C, 13.0% G and

24.9% T and corresponded to the nucleotide composition for the L-strand reported by Anderson *et al.* in 1981. The average A + T content of the mtDNA genome sequences was 55.8%, compared to an average G + C content of 44.2%. The increase in A + T content was attributed to a bias toward A/T ending codons in the mtDNA of many species (refer to Section 5.11.1.1).

The extent of the codon usage bias within dataset 4b was determined by calculating four different statistical codon bias measures namely RSCU, N_C , CBI and SCh_i^2 . The different codon usage bias measures together indicated the presence of a generally low level of codon usage bias within the Khoi-San mitochondrial DNA genomes which was attributed to mutation bias or natural selection, where evidence for the role of selection was supported by observations regarding expression levels of genes, the level of conservation of amino acids and protein folding (refer to Section 5.11.1.2).

5.14.6.2 Genetic diversity measures

Three different measures describing genetic diversity, namely the number of segregating sites, nucleotide diversity and the average number of nucleotide differences were calculated for sequence datasets 1b (All L Sequences), 2b (All Africa L Sequences), 3a (All Khoi-San L Sequences), 4b (!Kung Sequences), 5 (L3 Sequences), 6 (L2 Sequences), 7 (L1 Sequences) and 8a (L0 Sequences).

The nucleotide diversity observed within the different datasets in the current investigation was consistent with a previous study (refer to Section 5.11.2). High nucleotide diversity values were observed for all datasets, with the highest value, $3.78 \pm 0.05 \times 10^{-3}$, observed in dataset 2b and the lowest value, $1.41 \pm 0.09 \times 10^{-3}$, in dataset 6. In comparisons of datasets that represent specific populations, dataset 2b, consisting of sequences obtained from individuals currently living in Africa, displayed the highest nucleotide diversity, followed by dataset 1b (All L Sequences), dataset 3a (All Khoi-San L Sequences) and dataset 4b (!Kung Sequences). In the comparison of datasets consisting of sequences from specific mitochondrial haplogroups, dataset 8a consisting of L0 sequences displayed the highest nucleotide diversity, followed by datasets 7, 5 and 6 containing L1, L3 and L2 sequences, respectively. The high nucleotide diversity values determined for the different sequence datasets could be attributed to factors such as a large effective population size, a long genetic history, long periods of separation between different African sub-populations, or ancient population expansion events in Africa and is indicative of the

high level of diversity present in African populations and African mitochondrial haplogroups (refer to Section 5.11.2).

5.14.6.3 Deviation from the neutral theory of molecular evolution and factors to which this could be attributed

The high levels of nucleotide diversity observed in datasets 1b, 2b, 3a, 4b, 5, 6, 7 and 8a indicate the presence of a high number of polymorphisms within the individual mtDNA sequences. Under the neutral theory of molecular evolution (refer to Section 2.5.5), it is assumed that these polymorphisms are the consequence of forces such as migration and random genetic drift, and that the polymorphisms play no role in human adaptation which implies that the majority of mutations are neutral or nearly neutral (Kimura, 1968). Sequence datasets 1b, 2b, 3a, 4b, 5, 6, 7 and 8a were thus investigated to determine if the mitochondrial genome sequences in these datasets evolved according to the standard neutral theory of evolution (refer to Section 5.11.3). The neutrality of the observed alterations in the sequence datasets were evaluated using three different neutrality tests, namely Tajima's D test, as well as Fu and Li's D^* and F^* tests. Deviation from neutrality in any of the datasets could likely be attributed to one of two factors, i.e. population expansion events or the effect of selection. The role of population expansion events was investigated using frequency distributions of pairwise mtDNA differences (mismatch distributions), as well as a number of statistical parameters that can be used to describe the smoothness of observed mismatch distributions, namely the raggedness statistic, or those that can be employed to detect population expansion processes, i.e. the Ramos-Onsins and Rozas statistic and F_s test statistic (refer to Section 5.11.4). The effect of selection on the major L haplogroups (L0, L1, L2 and L3) was investigated through the analysis of synonymous and non-synonymous substitutions, as well as the calculation of conservation indices using haplogroup-specific phylogenetic networks (refer to Section 5.11.5). The results obtained after performing tests of neutrality, population expansion and selection are summarised in Table 5.80.

Table 5.80 Summary of results for tests of neutrality, population expansion and selection

Dataset ^a	Short description	Deviation from neutrality			Population expansion				Selection
		D	D^*	F^*	Mismatch distribution	rg	R_2	F_s	$RF_I - RF_T$
1b	All L Sequences	✓	✓	✓	✓	✓	✓	✓	✓
2b	All Africa L Sequences	✓	✓	✓	✓	x	✓	✓	x
3a	All Khoi-San L Sequences	x	x	x	x	x	x	x	x
4b	!Kung Sequences	x	x	x	x	x	x	x	x
5	L3 Sequences	✓	✓	✓	✓	x	✓	✓	x
6	L2 Sequences	✓	✓	✓	x	✓	✓	✓	x
7	L1 Sequences	x	x	x	x	x	x	x	x
8a	L0 Sequences	x	✓	✓	x	x	✓	✓	✓

^a = corresponding phylogenetic networks were used to investigate the effect of selection in datasets 5, 6, 7 and 8a. The L0 network was constructed in the current investigation (refer to Section 5.10), while the L1, L2 and L3 networks were obtained from MITOMAP (Ruiz-Pesini *et al.*, 2007). D = Tajima's D test; D^* = Fu and Li's D^* test; F^* = Fu and Li's F^* test; rg = raggedness statistic; R_2 = Ramos-Onsins and Rozas statistic; F_s = F_s test statistic; RF_I = replacement mutation frequency within internal branches; RF_T = replacement mutation frequency within terminal branches. ✓ = significant departure from neutrality (neutrality test), or support for occurrence of rapid population expansion event (population expansion), or significant difference observed between the RF for internal and terminal branches (selection); x = no significant departure from neutrality (neutrality test), or no evidence present for rapid population expansion event, i.e. population remained constant in size (population expansion), or no significant difference observed between the RF for internal and terminal branches (selection).

When using the three neutrality tests, significant negative values were calculated for datasets 1b, 2b, 5 and 6 which suggested the influence of positive selection or population growth within the populations or haplogroups represented by these datasets. A significant negative value was also determined for dataset 8a, but only when using Fu and Li's D^* and F^* statistics. When a significant value is obtained for a dataset using Fu and Li's D^* and F^* statistics, but a significant result is not obtained for the same dataset when using Tajima's D test, it is likely that selection, and not population expansion, is responsible for the observed deviation from neutrality (refer to Section 5.11.3).

The mismatch distributions generated for datasets 1b, 2b and 5 displayed characteristics of population expansion, while datasets 3a, 4b, 6, 7 and 8a exhibited characteristics of constant population size. It is however possible that the mismatch distributions were influenced by factors that could have contributed to suggest population size changes, or other factors which could have led to the removal of the population expansion signal in the mismatch distributions. In addition, the pairwise comparison test is conservative, which implies that the method could fail to identify population expansion events in some cases. Additional measures of population expansion (raggedness statistic, Ramos-Onsins and Rozas statistic and the F_s test statistic) were therefore employed to detect population expansion processes. Using these measures, population expansion events were supported in datasets 1b, 2b, 5, 6 and 8a, while datasets 3a, 4b and 7 represented

populations that had remained constant in size over a long period of time and therefore did not undergo population expansion events (refer to Section 5.11.4). This was in agreement with previous observations for hunter-gatherer populations to which the Khoi-San and !Kung belong (refer to Section 5.11.4).

The effect of selection on mitochondrial L haplogroups was evaluated by determining the ratio of non-synonymous substitutions versus synonymous substitutions for both internal and terminal branches of phylogenetic networks of the L0, L1, L2 and L3 haplogroups. A significant difference between the replacement mutation frequency values for internal and terminal branches was observed for haplogroups L0 and L, which suggested that purifying selection has played a role in shaping the mtDNA diversity observed in macrohaplogroup L, as well as the L0 haplogroup (refer to Section 5.11.5.1).

The most likely explanations for the observed deviation from neutrality in the different datasets, based on the results obtained for the different tests of neutrality, the pairwise comparisons test, the statistical tests of population expansion, and selection are summarised in Table 5.81. The most likely reasons for the observed deviation from neutrality, namely population expansion, selection, or a combination of both processes, are indicated for each dataset with the proviso that the result obtained for each dataset was specific to the dataset in question, and dependent on the specific characteristics of the dataset as outlined in Sections 5.11.3, 5.11.4 and 5.11.5.

Table 5.81 Summary of proposed explanations for deviation from neutrality

Dataset ^a	Short description	Reason for deviation from neutrality
1b	All L Sequences	Population expansion event, coupled with the effect of purifying selection
2b	All Africa L Sequences	Population expansion event
5	L3 Sequences	Population expansion event
6	L2 Sequences	Population expansion event
8a	L0 Sequences	Population expansion event, coupled with the effect of purifying selection

^a = corresponding phylogenetic networks were used to investigate the effect of selection in datasets 5, 6 and 8a. The L0 network was constructed in the current investigation (refer to Section 5.10), while the L1, L2 and L3 networks were obtained from MITOMAP (Ruiz-Pesini *et al.*, 2007).

The population growth observed in datasets 1b, 2b, 6 and 8a can be attributed to dispersals and differentiation of mtDNA haplogroups within Africa due to interactions between different populations or environmental factors, while the population expansion evident in dataset 5 can be due to an increase in carriers of the L3 haplogroup which migrated from Africa to populate the rest of the world. The observed population expansion

in datasets 1b, 2b, 5 and 6 therefore supports the “Recent African Origin” model as population growth within Africa is a requirement of this model (refer to Section 5.11.3). The effect of purifying selection observed for haplogroups L and L0 confirms the role of selection in shaping mitochondrial variation in these haplogroups. Purifying selection has previously been described as the predominant factor influencing human mtDNA variation due to the essential function of the mtDNA proteins (refer to Section 5.11.5.1).

Datasets 3a, 4b and 7 containing Khoi-San, !Kung and L1 sequences respectively, did not exhibit deviation from neutrality, and no evidence of a population expansion event or effect of selection were observed in these datasets. These three datasets consist mainly of sequences obtained from individuals that form part of hunter-gatherer populations which do not usually display evidence of population expansion as these populations have remained constant in size over long periods of time. However, in a recent investigation (Atkinson *et al.*, 2009) slow constant growth was observed in datasets consisting of L0 and L1 sequences. Population growth could have similarly occurred in the populations or haplogroups represented by datasets 3a, 4b and 7. However, population expansion was not detected in the current investigation due to low sensitivity of the methods employed, the small sample sizes of the datasets investigated, or alternatively no evidence of such events remain due to subsequent population events such as bottlenecks, or the amount of time that has passed since the population growth event occurred (refer to Section 5.11.4).

5.14.6.4 Coalescent date estimates

The coalescence times of the L0 haplogroup, sub-haplogroups and lineages within sub-haplogroups, were estimated by using the global L0-specific haplogroup network. The estimates of TMRCA for the different sub-haplogroups and lineages within sub-haplogroups presented in the current investigation could be influenced by a number of factors which could cause the TMRCA to be underestimated. These factors included the sampling method used to obtain sequences, sample size, and the presence of recurrent alterations in the phylogenetic network used for calculations. The TMRCA estimates are further affected by subdivision in populations, purifying selection and sex-specific demographic processes. All of these factors could play a role in the results generated in the current investigation (refer to Section 5.11.6). The TMRCA for some sub-haplogroups and lineages within the L0 haplogroup is presented in Table 5.82.

Table 5.82 Estimated TMRCA for different haplogroups, sub-haplogroups and lineages within sub-haplogroups

Haplogroup/ sub-haplogroup classification	Estimated TMRCA (YBP) based on coding region sequence data		
	Current investigation	Other investigations	
L0	123,946 to 129,044	92,000 to 124,000 121,300 to 171,500 125,000 to 159,000 128,345 to 138,825	Kivisild <i>et al.</i> , 2006 Gonder <i>et al.</i> , 2007 Mishmar <i>et al.</i> , 2003 Mouton, 2003
L0a	118,010 to 131,544	48,000 to 56,000 ^a	Watson <i>et al.</i> , 1997
L0aA	118,010 to 131,544	7,707 to 15,413 51,200 to 90,600	Mouton, 2003 Gonder <i>et al.</i> , 2007
L0aB	91,095 to 117,847	12,800 to 48,400	Gonder <i>et al.</i> , 2007
L0aC	89,492 to 116,024	85,500 to 104,300	Gonder <i>et al.</i> , 2007
L0b	108,009 to 115,687	51,000 to 73,000 71,500 to 109,300 85,800 to 126,200	Kivisild <i>et al.</i> , 2006 Gonder <i>et al.</i> , 2007 Gonder <i>et al.</i> , 2007
L0bB	66,242 to 72,628	42,341 to 52,197	Mouton, 2003
L0bB1b1	35,342 to 57,140	---	---
L0bB2a1	6,515 to 24,313	---	---
L0c	76,662 to 82,988	31,000 to 53,000 48,900 to 60,300	Kivisild <i>et al.</i> , 2006 Gonder <i>et al.</i> , 2007
L0c1	22,617 to 28,157	16,750 to 49,950 ^a	Salas <i>et al.</i> , 2002
L0c1B	11,373 to 18,169	9,400 to 45,300 ^a	Salas <i>et al.</i> , 2002
L0c2	45,784 to 52,264	23,000 to 41,000	Kivisild <i>et al.</i> , 2006
L0c2B	27,598 to 34,914	14,000 to 28,000	Kivisild <i>et al.</i> , 2006
L0c2D	18,678 to 26,278	1,000 to 5,000	Kivisild <i>et al.</i> , 2006

TMRCA = time to most recent common ancestor; YBP = years before present. Sub-haplogroups and lineages within sub-haplogroups are as indicated in Figures 5.50, 5.51, 5.52 and 5.53 (refer to Section 5.10). ^a = estimated TMRCA of L0a is based on control region sequence data (16,090 to 16,365), and estimated TMRCA of L0c1 and L0c1B is based on HVS1 sequence data, as well as RFLPs. The colours in which the haplogroup/sub-haplogroup classifications are presented correspond to the colours used to indicate the different branches in the phylogenetic trees.

A complete list of the TMRCA for all sub-haplogroups and lineages within the L0 haplogroup is presented in Section 5.11.6. In the majority of TMRCA estimates determined for sub-haplogroups presented in this list, i.e. TMRCA estimates based on more than four individuals, the TMRCA of sub-haplogroups were equal to the combined TMRCA estimates of the lineages that form part of the sub-haplogroup. This was however not the case for L0a as the TMRCA of this sub-haplogroup was not equal to the combined estimates of the TMRCA calculated for L0aA, L0aB and L0aC. This is attributed to the fact that the three lineages do not form a monophyletic clade, but are three separate branches within the L0 branch of the global phylogenetic trees (refer to Section 5.9). These three branches were however classified as L0a in the current investigation based on shared characteristics as described by the classification scheme used (refer to Section 4.9) and for the sake of clarity, reference is made to L0aA, L0aB and L0aC

throughout the current investigation, even when new designations are proposed for these sub-haplogroups in the revised classification scheme (refer to Section 5.12).

Three new branches were identified in the L0-specific haplogroup network after the addition of 46 L0 sequences. L0aB represents a new branch located between L0a and L0b which shares characteristics of both the L0a and the L0b haplogroups. Further investigation revealed that this branch was first described by Gonder *et al.* in 2007 and appears to be specific to East African Khoi-San speakers as it consists of Sandawe individuals from Tanzania. The remaining two new branches, namely L0bB1b1 and L0bB2a1, represent two previously unidentified lineages in the Southern African Khoi-San population. The TMRCA of these two lineages were determined as $46,241 \pm 10,899$ YBP for L0bB1b1, and $15,414 \pm 8,899$ YBP for L0bB2a1. Although the TMRCA calculated for these two lineages were based on small sample sizes, and is therefore not likely to be precise, the identification of two previously undescribed lineages with a difference in age of ca. 30,000 years provides evidence for the prolonged presence of Khoi-San speaking populations in southern Africa. This implies that the extent of the genetic diversity present in the Khoi-San has to date not been described completely, alternatively the difference in TMRCA of the two lineages may suggest prolonged genetic isolation coupled with loss of lineages due to random genetic drift.

The TMRCA estimates determined in the current investigation were compared to estimates calculated in previous investigations. The estimated TMRCA for the L0 haplogroup determined in the current investigation is within the range of a minimum of 92,000 to a maximum of 138,825 YBP previously reported in other genetic investigations of mtDNA (refer to Table 5.82). The agreement between the estimated TMRCA for the L0 haplogroup and previously reported paleontological dates supports the hypothesis that modern *Homo sapiens* populations share a common and recent origin from an ancestral population that was present in Africa, and data from the current investigation therefore provides support for the “Out of Africa” theory of modern human population expansion and colonisation of the world. TMRCA determined in the current investigation for L0b, L0c1 and L0c1B were also broadly consistent with previously published TMRCA (refer to Table 5.82). TMRCA determined in the current investigation for L0a, L0aA, L0aB, L0aC, L0bB, L0c, L0c2, L0c2B, and L0c2D were however more ancient (refer to Section 5.11.6).

The increased TMRCA values for L0aA, L0bB, L0c, L0c2, L0c2B and L0c2D can be attributed to the use of a greater number of L0 sequences to determine the TMRCA for

each of these sub-haplogroups or lineages within sub-haplogroups compared to previous investigations. The use of a greater number of L0 sequences in the determination of TMRCA values for the L0aA, L0bB, L0c, L0c2, L0c2B and L0c2D groups also permitted the calculation of smaller standard deviation values which made the estimate more precise by generating a narrower range for the TMRCA. In the case of L0aB and L0aC, smaller sample sizes were used to calculate the TMRCA values than in previous investigations. The use of smaller samples to determine TMRCA values gave rise to the calculation of increased standard deviation values which extended the range of the TMRCA thereby making the estimate less precise. The more ancient values observed for L0aB and L0aC should be revised after the addition of more sequences to the L0aB and L0aC branches (refer to Section 5.11.6).

The TMRCA of the different sub-haplogroups and lineages within sub-haplogroups, used in conjunction with the current geographic distribution of these groups and lineages, have the potential to provide information regarding the dispersal of L0 sub-haplogroups and lineages within sub-haplogroups across the African continent. The oldest sub-haplogroups and lineages, namely L0a, L0aA, L0aB, L0aC and L0b, are located in southern and eastern Africa. The L0aA (L0k¹) and L0b (L0d) sub-haplogroups are observed in Southern African Khoi-San speakers, while the L0aB (L0d3) sub-haplogroup is present in eastern African Khoi-San speakers. If it is assumed that L0aB (L0d3) and L0b (L0d1 and L0d2) diverged from a shared mtDNA haplogroup observed in an ancestral Khoi-San speaking population, the ancestral population that gave rise to Khoi-San speakers originated in East Africa and members of this population subsequently dispersed to southern Africa. The two populations then became genetically isolated from one another probably due to a marked dry period in southern Africa which caused the contraction of available geographic ranges. It also appears that the L0c haplogroup originated in East Africa from where it dispersed to Central and West Africa (refer to Section 5.11.6, Figure 5.59).

5.14.7 Revised classification scheme for the L0 haplogroup

The branching pattern observed within the global phylogenetic trees whereby three branches in different parts of the global N-J and MP trees were assigned the same sub-haplogroup designation (refer to Section 5.9.2), as well as the branching pattern in the L0-specific haplogroup network, highlighted the fact that the classification scheme used in

¹ The haplogroup classifications presented in brackets refer to the classifications according to the alternative classification scheme (refer to Section 3.3.1.1).

the current investigation (refer to Section 4.9) did not make adequate provision for the variation observed in the L0 haplogroup. This was further supported by the fact that some L-specific sequences could not be classified when using the classification scheme (refer to Section 5.8). A revised classification scheme for the L0 haplogroup was thus proposed (refer to Section 5.12).

A revised classification scheme for the L0 haplogroup was proposed based on the new internal and terminal branches observed in the global L0-specific haplogroup network which represented new groups and lineages within each of the L0 sub-haplogroups. Using the global L0 network, haplogroup defining and haplogroup-specific polymorphisms were identified using specific criteria in order to select slowly evolving nucleotide positions in the mitochondrial coding region. Slowly evolving nucleotide positions, mainly synonymous substitutions, were selected in genes, other than the tRNA or rRNA genes, located in the coding region. Nucleotide positions, at which recurrent alterations or reversions were present, were viewed as fast evolving and were excluded where possible. The revised L0 classification scheme is presented in Section 5.12.

The revised L0 classification scheme exhibited a number of differences compared to the classification scheme used throughout the current investigation (refer to Section 4.9). These differences are listed in Table 5.83.

Table 5.83 Differences in the revised L0 classification scheme compared to the classification scheme used in the current investigation

Branching point in revised L0 classification scheme	Classification scheme used in current investigation ^a (Wallace, 2004)	Revised L0 classification scheme ^b (current investigation)
L0, L1, L2 vs. L5	---	3,423
L0, L1 vs. L2	10,810	2,758
L0 vs. L1	7,055	14,178
L0 vs. L1	11,914	10,589
L0c1 vs. L0c2	5,096	6,257

^a = haplogroup classification according to classification scheme used in the current investigation (refer to Section 4.9). ^b = haplogroup classification according to the revised classification scheme for the L0 haplogroup proposed in the current investigation (refer to Section 5.12).

The majority of the differences between the classification scheme used throughout the current investigation and the revised L0 classification scheme can be attributed to the addition of the 46 new L0 sequences, which include 22 Southern African Khoi-San L0 mtDNA sequences generated in the current investigation, to the global L0-specific

haplogroup network. These new L0 sequences did not conform completely to the classification scheme used in the current investigation, and the alterations identified within these new L0 sequences were employed to revise the nucleotide positions and the SNPs identified at these positions used to classify L0 sequences (refer to Section 5.12).

The revised L0 classification scheme was tested by re-classifying mtDNA sequences that were incompletely classified or exhibited a discrepancy between their classifications according to the classification scheme used in the current investigation and their locations within the global N-J and MP trees. The haplogroup designations of the re-classified sequences are indicated in Table 5.84.

Table 5.84 Revised haplogroup classifications of incompletely classified L sequences and L sequences that exhibited a discrepancy between classification and location in global N-J and MP trees

Haplogroup classification ^a (Wallace, 2004)	Revised L0 classification scheme ^b (current investigation)
L0a\$	L0aC, L1 or L5
L0b\$	L5
L0c2\$	L0c1B
L2a	L0aB1, L0aC or L0c2E

^a = haplogroup classification according to the classification scheme used in the current investigation (refer to Section 4.9).

^b = haplogroup classification according to the revised classification scheme for the L0 haplogroup proposed in the current investigation (refer to Section 5.12). The \$ sign indicates that these L sequences could not be completely classified using the classification scheme presented in Section 4.9.

The revised L0 classification scheme resolved the classification of incompletely classified L0 sequences, as well as the discrepancies previously observed between the haplogroup classification and location of specific L sequences in the global N-J and MP trees. That is, the designation conferred by the revised L0 classification scheme corresponded to the observed location of the specific sequence in the phylogenetic trees.

Haplogroups, sub-haplogroups and lineages within sub-haplogroups in the revised L0 classification scheme were named as in the global L0-specific haplogroup network. New designations were however proposed for the new branches representing sub-haplogroups and lineages within sub-haplogroups identified in the global L0-specific haplogroup network and the global N-J and MP trees. The proposed new designations were based on the estimated ages of the sub-haplogroups or lineages (refer to Section 5.11.6). It was thus suggested that L0aB be renamed to L0d, L0aC to L0e, L0bA to L0b1, and L0bB to L0b2.

The sub-haplogroup designations and the assignment of lineages within the sub-haplogroups in the proposed revised L0 classification scheme (refer to Section 5.12) were compared to the classification scheme used in the current investigation (refer to Section 4.9), as well as the alternative classification scheme (refer to Section 3.3.1.1). The comparison indicated that an urgent need exists for a standardised classification scheme for mtDNA haplogroups to clarify and facilitate the direct comparison of data generated for mtDNA haplogroups in different human populations and investigations. The comparison of the haplogroup and sub-haplogroup designation between the three classification schemes is presented in Table 5.85.

Table 5.85 Comparison of haplogroup and sub-haplogroup designations between three classification schemes and their alignment with the global L0-specific haplogroup network

Classification scheme (Wallace, 2004) ^a	Global L0 specific haplogroup network ^b	Revised L0 classification scheme ^c	Alternative haplogroup classification scheme ^d
L0a	L0aA	L0a	L0k
---	L0aB	L0d	L0d3 ^e
---	L0aC	L0e	L0f
L0b	L0b	L0b	L0d
---	L0bA	L0b1	L0d2
---	L0bB	L0b2	L0d1
L0c	L0c	L0c	L0a
L0c1	L0c1	L0c1	L0a1
L0c2	L0c2	L0c2	L0a2

^a = classification scheme used in current investigation (refer to Section 4.9). ^b = global L0-specific haplogroup network constructed in current investigation (refer to Section 5.10). ^c = revised classification scheme for the L0 haplogroup proposed in the current investigation (refer to Section 5.12). ^d = haplogroup classification scheme used in other investigations described in the literature (refer to Section 3.3.1.1). ^e = new branch in L0d, specific to East African Khoi-San speakers (Gonder *et al.*, 2007), named L0d3 in the current investigation to distinguish it from L0d1 and L0d2. This branch was first described by Gonder *et al.* (2007) and named L0dTZ.

The revised L0 classification scheme incorporated mitochondrial sequence data recently obtained from individuals classified as belonging to the L0 haplogroup (Gonder *et al.*, 2007; current investigation). The revised classification scheme therefore represents the most recent version of the classification scheme for haplogroup L0 which accommodates the genetic diversity present in the L0 haplogroup and provides a way in which to describe and classify the observed diversity.

The revised L0 haplogroup classification scheme, in conjunction with similarly updated classification schemes for the other mtDNA haplogroups, can be used to devise a standard haplogroup classification scheme for mtDNA genome sequences, similar to the standardised classification scheme currently used for Y chromosome haplogroups.

5.14.8 Khoi-San consensus sequence

The construction of a consensus sequence for the human mitochondrial DNA genome is necessary for the study of human mtDNA diversity. To date, two human mitochondrial consensus sequences, based on mtDNA sequences from different human mtDNA haplogroups, have been constructed by Marzuki *et al.* (1991) and Carter (2007).

However, if the evolution of the human mitochondrial genome, and mitochondrial disease, are to be studied, a consensus sequence that is based on an ancient human population and which summarises the variation present in human populations prior to their exodus from their point of origin, i.e. Africa, is essential. Such a consensus sequence will be an approximation of the “ancestral” human mitochondrial genome sequence prior to differentiation into separate haplogroup lineages that evolved as humans migrated into new environments. This consensus sequence must therefore be constructed taking into account the high levels of variation observed in African populations compared to populations on other continents, and the increased age of African populations compared to other populations, as indicated by the position of African mitochondrial genome sequences in phylogenetic trees based on mitochondrial sequences collected worldwide, as well as the estimated TMRCA determined for African mtDNA haplogroups (refer to Section 3.3.1.4). At present, the sequence used for the investigation of human evolution and mitochondrial disease, is the rCRS which does not incorporate these two factors. However, a consensus sequence based on the Southern African Khoi-San population conforms to these requirements regarding diversity and age. The high levels of variation observed in the Southern African Khoi-San population is evident from the high nucleotide diversity determined for the Khoi-San population (refer to Section 5.11.2), and the ancient nature of the Khoi-San population is confirmed by the location of the Khoi-San genome sequences in the global phylogenetic trees (refer to Sections 5.9.2 and 5.9.3), as well as the estimated TMRCA determined for the mitochondrial haplogroups present in the Khoi-San cohort (refer to Section 5.11.6).

A consensus sequence was constructed for the Southern African Khoi-San population using 35 Khoi-San mtDNA sequences that were classified within the L0a or L0b sub-haplogroups. Thirteen of these sequences were obtained from a previous investigation (Mouton, 2003), while the remaining 22 sequences were generated during the course of the current investigation. The complete Khoi-San consensus sequence consisted of a total of 16,567 bp and is presented in Appendix N. The Khoi-San consensus

sequence was also presented in the form of a list of SNPs compared to the rCRS (refer to Section 5.13).

The Khoi-San consensus sequence presented with a total of 127 alterations (123 nucleotide changes, three deletions and one insertion) compared to the rCRS. The alterations were distributed throughout the consensus sequence with 22.83% (29/127) in the control region and 77.17% (98/127) in the coding region. The alterations observed in the coding region included all the SNPs that were listed as exclusion criteria, with the exception of 15,301 which excludes haplogroups from macrohaplogroup N (refer to Appendix D). In order to retain the polymorphism data contributed by both the L0a and L0b sub-haplogroups, ambiguity codes were included in the Khoi-San consensus sequence. The Khoi-San consensus sequence contained 69 ambiguous positions which were located throughout the consensus sequence with a total of 13 in the control region (nucleotide position 16,023 to 576) and 56 in the coding region (nucleotide position 577 to 16,024).

The use of ambiguity codes in the Khoi-San consensus sequence is advocated as it should decrease the number of observed alterations in comparisons between mtDNA sequences and the consensus sequence, thereby reducing the number of alterations identified as possible candidates for the observed pathogenicity. The presence of a high number of ambiguous nucleotide positions in the control region indicates the high level of sequence diversity in this region, and provides support for the use of classification schemes based on coding region data as coding region SNPs would provide distinct alterations for the unambiguous classification of sequences.

The use of the Khoi-San consensus sequence in comparison with the 43 Khoi-San sequences included in the current investigation, generated a minimum number of nucleotide differences of one nucleotide with KS_CGR_083 (L0a sub-haplogroup), and a maximum number of nucleotide differences of 77 with KS_CGR_164 (L2b sub-haplogroup). Therefore, a greater number of differences should be observed between the Khoi-San consensus sequence and younger haplogroups than between the consensus sequence and older haplogroups. This is in contrast to the trend observed in comparisons with the rCRS where sequences from haplogroups observed within Africa exhibit more differences compared to the rCRS than sequences from haplogroups detected outside of Africa.

The Khoi-San consensus sequence was used to compile a list of putative population-specific alterations for the Southern African Khoi-San population. This list consisted of a total of 43 SNPs and included alterations that form part of the exclusion criteria listed for L-specific sequences (refer to Appendix D). This list of potential population-specific alterations can be used to distinguish between alterations that are normal variants versus potentially pathogenic alterations. In addition, the inclusion of all the SNPs used as exclusion criteria for L-specific sequences in the list of potential population-specific alterations, implies that the Khoi-San consensus sequence can be viewed as representative of the normal variation observed in the African population.

As previously stated in this section, two human mitochondrial consensus sequences have been constructed. The Khoi-San consensus sequence was compared to both of these human consensus sequences, and it was determined that for use in comparisons with mitochondrial sequences obtained from individuals of African descent, the Khoi-San consensus sequence was more suitable than the Carter consensus sequence. In addition, the small number of differences between the Khoi-San consensus sequence and the Carter consensus sequence constructed using a global dataset consisting of 827 mitochondrial sequences, also indicated that the Khoi-San sequence, although based on a small number of sequences, was representative of the variation present in human mitochondrial genomes worldwide (refer to Section 5.13).

To date, two sequences, NC_001807 from Yoruba and D38112 from Uganda, have been proposed for use as African reference sequences. In order to determine which sequence, NC_001807, D38112 or the Khoi-San consensus sequence, would be more useful in comparisons with mitochondrial sequences when used as an African reference sequence, the number of differences between the rCRS, and the Khoi-San consensus sequence or the two African reference sequences were determined. The greatest number of differences was observed between the Khoi-San consensus sequence and the rCRS, which indicated that, compared to the two African reference sequences, the Khoi-San consensus sequence contained the highest level of variation. The presence of a large number of polymorphisms in the Khoi-San consensus sequence therefore provides a template of a typical African mitochondrial background. Thus, the Khoi-San consensus sequence should be used as the African reference to identify putative pathogenic variants in African mitochondrial genome sequences.

The findings presented here represent an in-depth investigation of the mtDNA diversity present in the Southern African Khoi-San population. The sequence information obtained for the 30 Khoi-San individuals was used to investigate sequence variation due to polymorphisms and to determine the nucleotide diversity present in the Khoi-San population. In addition, phylogenetic relationships of these sequences with other L-specific sequences were investigated, the internal topology of the L0 haplogroup was elucidated and aspects that shape variation such as population growth and the effect of selection on the L0 haplogroup were studied. Furthermore, the current classification scheme for the L0 haplogroup was re-evaluated and updated and a mitochondrial consensus sequence was constructed for the Khoi-San population of southern Africa. The common factor in the analyses is the investigation of the variation observed in human mtDNA, and it is evident from the results presented in the current investigation that the Southern African Khoi-San population contains high levels of genetic variation in their mitochondrial genomes. Through the use of the correct technology, insight into the genetic variation present in ancient human populations such as the Southern African Khoi-San, will vastly improve understanding of variation within the human mitochondrial genome, which will greatly impact on the way that human evolution and mitochondrial disease are studied.

CHAPTER SIX

Conclusions

In studies of human evolution, population genetics and the identification of alterations that could be involved in various mitochondrial diseases, the rCRS is used as a standard for the human mitochondrial DNA sequence. Due to the large number of differences observed between the rCRS and mitochondrial sequences obtained from individuals of African descent, it often becomes difficult to distinguish between alterations associated with specific population groups and alterations with possible pathological significance. In the past, a recent disease-causing alteration was distinguished from background polymorphisms by determining the phylogenetic relationship between the patient's haplotype and the global mtDNA phylogeny (Wallace, 1994). Investigators have also attempted to address the problem of distinguishing between polymorphisms and potential pathogenic alterations by constructing consensus sequences using human mitochondrial sequences from different continents and different haplogroups (Marzuki *et al.*, 1991; Carter, 2007; refer to Section 3.6). The downfall of these consensus sequences is however the loss of sequence information through the amalgamation of different haplogroups and different populations, and the combination of ancient population data with recent population data. When using such a consensus sequence, population-specific or haplogroup-specific alterations can not be identified, the age of an alteration can not be estimated (i.e. ancient or recent) and the evolution of the human mitochondrial genome can therefore not be studied.

The loss of variation in the human mitochondrial consensus sequences discussed in the previous paragraph, can be countered by using an African reference sequence. An African reference sequence provides a good approximation of the variation present in the human mitochondrial genome prior to the divergence of all haplogroups other than macrohaplogroup L. At present, two sequences are available for use as African reference sequences. The first sequence, NC_001807 (Ingman *et al.*, 2000), is a sequence from a Yoruban individual which was classified in the current investigation as belonging to the L3a1 haplogroup. The second sequence, D38112 (Horai *et al.*, 1995), is a sequence from a Ugandan individual that was assigned to the L0c2 haplogroup in the current

investigation. The use of the first sequence as an African reference sequence is advocated by NCBI and it is contained in their RefSeq database, while the use of the second sequence is suggested by MITOMAP (refer to Section 5.13). NC_001807 however belongs to the L3 haplogroup which gave rise to macrohaplogroups M and N of which the descendant haplogroups populated the rest of the world (Quintana-Murci *et al.*, 1999; Wallace *et al.*, 1999), while D38112 belongs to the L0c haplogroup. These sequences do not represent the most ancient African populations or ancient haplogroups, i.e. the Khoi-San population or the L0a (L0k¹) and L0b (L0d) haplogroups (Chen *et al.*, 2000; Gonder *et al.*, 2007), and are therefore more useful in the study of the comparatively more recent European and Asian sequences, than for comparison with more ancient African sequences.

In 2003, Mouton determined the complete mtDNA sequences for 13 Khoi-San individuals from southern Africa. In the current investigation the complete mitochondrial genome sequences of another 30 Khoi-San individuals from southern Africa were determined. To date, a total of 54 complete or coding region only Khoi-San mitochondrial genome sequences have been published (Ingman *et al.*, 2000; Mouton, 2003; Kivisild *et al.*, 2006; Gonder *et al.*, 2007; Behar *et al.*, 2008). The results generated in the current investigation increases the number of published Khoi-San mitochondrial DNA genome sequences by 42.25% to a total of 71 sequences and adds to the diversity of the global human mitochondrial sequence dataset.

The 13 individual sequences generated by Mouton in 2003 and the 30 individual sequences obtained in the current investigation were combined to compile a consensus sequence for the Southern African Khoi-San population. This was done as discussed in Section 5.13 because these two groups of individuals were obtained from the sample group previously collected in the north-western Kalahari Desert in southern Africa (Chen *et al.*, 2000).

It was envisaged that the compilation of a consensus sequence of one of the most ancient lineages (Chen *et al.*, 2000) would provide a means to investigate the evolution of the human mitochondrial genome, and the mitochondrial variation present in humans. The consensus sequence could be used as a standard to which mtDNA sequences could be

¹ The haplogroup classifications presented in brackets refer to the classifications according to the alternative classification scheme (refer to Section 3.3.1.1).

compared in order to differentiate between alterations that are likely to be associated with African populations and alterations that are potential disease-causing mutations.

Based on the current investigation, the use of the constructed Southern African Khoi-San consensus sequence, alone or in conjunction with the rCRS, is advocated as a tool in the investigation of human mitochondrial evolution and disease. The results presented in the current investigation, as summarised in Section 5.14, lend support to previous findings of the presence of a high level of genetic variation in the mtDNA of the Khoi-San population, the ancient character of the Khoi-San population and in addition provide new insights into the complexity of the L0 haplogroup within the Khoi-San and African population as a whole. The high level of genetic variation and increased age, compared to other populations, of the Southern African Khoi-San support the use of the Khoi-San consensus sequence as a standard in studies of human evolution and mitochondrial disease.

The observations made in the current investigation can be placed in context with other investigations by evaluating the results using a model describing the genetic variation observed in human mitochondrial genomes. Such a model is introduced in the next section and incorporates a number of aspects relating to genetic variation in human mitochondrial genomes.

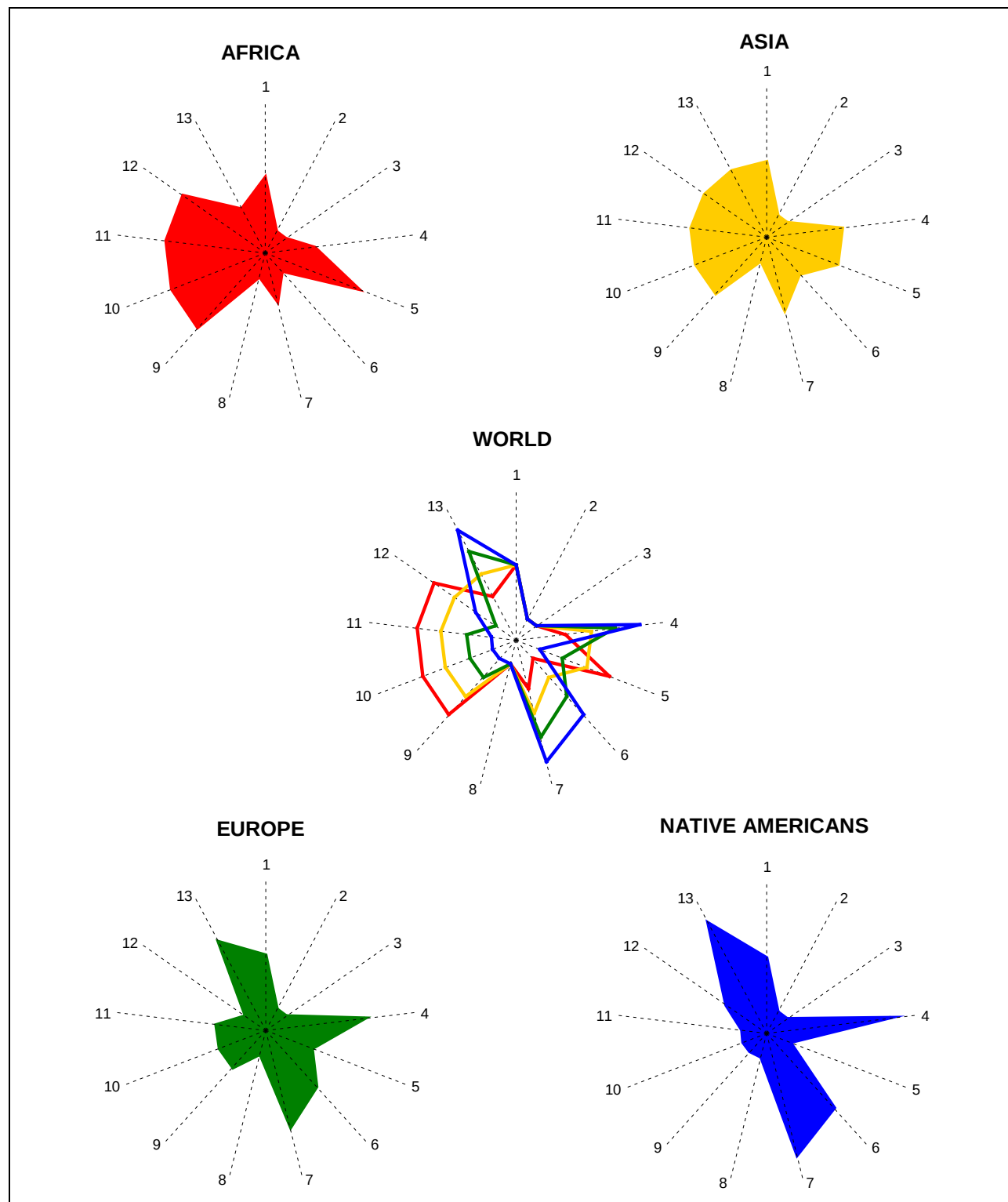
6.1 PROPOSED MODEL FOR GENETIC VARIATION OBSERVED IN HUMAN MITOCHONDRIAL GENOMES

The genetic variation observed in human mitochondrial DNA can be the consequence of a number of different factors or events. In order to identify and explain the different factors, a model is presented in Figure 6.1 which incorporates the most important aspects giving rise to genetic variation in mitochondrial DNA genomes. The different aspects are further discussed in Section 6.2.

Genetic variation in mitochondrial genomes may be due to both internal and external factors. Internal factors include aspects unique to the genetics of the mitochondrion such as a high mutation rate, while external factors involve aspects such as selection which are due to the influence of environmental factors. In order to evaluate the genetic variation present in different human mitochondrial genomes, a standard is required against which mitochondrial genomes can be compared. The rCRS provides such a standard and is

therefore included in the proposed model. The proposed model for genetic variation observed in human mitochondrial genomes is presented in Figure 6.1.

Figure 6.1 Model indicating the different factors that influence the genetic variation observed in human mitochondrial genome sequences



The genetic variation profile of the different populations are indicated with shapes or lines using different colours where ■ = Africa, ■ = Asia, ■ = Europe and ■ = Native Americans. The different factors that are involved in generating, shaping and evaluating genetic variation in the mitochondrial genomes of the different populations are indicated using numbers where 1 = mutation rate (substitutions per site per year), 2 = cytoplasmic segregation, 3 = random genetic drift, 4 = natural selection, 5 = effective population size (number of breeding females), 6 = population size changes, 7 = human migration, 8 = population substructure, 9 = intra-population nucleotide diversity, 10 = inter-population nucleotide diversity, 11 = estimated age of population (years before present), 12 = reference sequence, 13 = consensus sequence.

The model presented in Figure 6.1 indicates that genetic variation in human mitochondrial genomes in different populations can be attributed to a number of factors, and the effect of these different factors on genetic variation can be measured by estimating a number of values. These factors and values include:

- a. Factors that introduce genetic variation in mitochondrial genomes (refer to Section 6.2.1):
 - i. High mutation rate (1).
 - ii. Cytoplasmic segregation (2).
- b. Factors that shape and govern genetic variation in mitochondrial genomes (refer to Section 6.2.2):
 - i. Random genetic drift (3).
 - ii. Natural selection (4).
 - iii. Effective population size (5).
 - iv. Population size changes, i.e. expansions or reductions (6).
 - v. Human migration (7).
 - vi. Population substructure (8).
- c. Measurable outcomes that are affected by the level of genetic variation present in mitochondrial genomes (refer to Section 6.2.3):
 - i. Intra-population genetic diversity (9).
 - ii. Inter-population genetic diversity (10).
 - iii. The estimated ages of populations (11).
- d. Standards used to investigate genetic variation present in mitochondrial genomes (refer to Section 6.2.4):
 - i. Revised Cambridge Reference Sequence (12).
 - ii. Consensus sequence (13).

The evidence providing support for the proposed model indicated in Figure 6.1 is discussed in the next section. Each aspect contributing to the observed genetic variation is discussed separately and the way in which the results generated in the current investigation contributes to or differs from the model is addressed.

6.2 EVIDENCE PROVIDING SUPPORT FOR THE PROPOSED MODEL FOR GENETIC VARIATION OBSERVED IN HUMAN MITOCHONDRIAL GENOMES

Evidence for the proposed model, presented in Figure 6.1, is discussed in this section of the chapter and includes the following:

- a. Factors that introduce genetic variation in mitochondrial genomes.
- b. Factors that shape and govern genetic variation in mitochondrial genomes.
- c. Measurable outcomes that are affected by the level of genetic variation present in mitochondrial genomes.
- d. Standards used to investigate genetic variation present in mitochondrial genomes.

6.2.1 Factors that introduce genetic variation in mitochondrial genomes

The main source of genetic variation observed in mitochondrial genomes is mutation, although other processes such as cytoplasmic segregation also contribute to the generation of variation (refer to Section 2.5). The mutation rate of primate mtDNA is several orders of magnitude higher (five- to tenfold) than that observed for the nuclear genome (Brown *et al.*, 1979) and mutations can occur in the form of substitutions, deletions and insertions.

In the current investigation, evidence of the effect of the high mutation rate was observed through the identification of sequence alterations compared to the rCRS. A total of 292 alterations were reported for the 30 Khoi-San mitochondrial genome sequences of which 272 were previously reported, while 20 alterations were novel. The majority of the alterations observed were transitions (265/292), followed by transversions (19/292), insertions (5/292) and deletions (3/292). This was not surprising as a bias towards transitions has previously been reported in the mtDNA genome of humans (Brown *et al.*, 1982; Greenberg *et al.*, 1983; Crespillo *et al.*, 2004). Although generally rare, the identification of insertion and deletion events at nucleotide positions in the control region of the 30 Khoi-San sequences generated in the current investigation was also not surprising. The insertion of one or two nucleotides has previously been detected at poly(C) tracts starting at nucleotide positions 303 and 311 (Aquadro and Greenberg, 1983). In the current investigation, the insertion of one or two C nucleotides was observed at nucleotide position 309, while the insertion of one C nucleotide was observed at nucleotide position 315. Insertions of one and four C nucleotides were also observed at nucleotide positions 573 and 16,193, respectively. A deletion of a C nucleotide was detected at nucleotide position 494, while the deletion of a two nucleotide repeat element, CA, was observed at nucleotide positions 514 and 515. Again, this was not surprising as it has previously been demonstrated that length differences can arise from differences in the copy number of STR sequences (Wrischnik *et al.*, 1987).

Further evidence in support of a high mutation rate in mitochondrial genomes was discovered during the construction of the global L0-specific haplogroup network (refer to Section 5.10.2.3). A number of recurrent alterations or reversions were observed in the global L0-specific haplogroup network. The largest number of recurrent alterations was observed in ND4, ND5 and ATPase6, while the largest number of reversions was present in ND4, ND5 and COII. Although these types of alterations can be ascribed to the high mutation rate and the presence of mutational hot spots in the mitochondrial genome, the presence of a high number of recurrent alterations or reversions in these genes could be attributed to the effects of natural selection (refer to Section 5.10.2.3).

When genetic variation is introduced through mutational processes in the mtDNA located in germ-line cells, the variation is transferred to the descendants of the individual. The subsequent accumulation of alterations along radiating maternal lineages leads to the introduction of additional variation (Wallace, 1994). The genetic variation introduced to the mitochondrial lineage is further influenced and shaped by evolutionary factors, as well as demographic factors. Through the influence of these different factors, mitochondrial genome variation was shaped to give rise to different mitochondrial lineages. These different lineages are continent-specific and therefore gave rise to differences regarding the mitochondrial genome variation observed between African, Asian, European and Native American populations.

6.2.2 Factors that shape and govern genetic variation in mitochondrial genomes

The genetic variation introduced into the mitochondrial DNA genome through mutational events and processes such as cytoplasmic segregation is governed and shaped on a population level by evolutionary forces and population events such as random genetic drift, natural selection, migration, effective population size, population substructure, population expansions and population reductions. All of these forces and events contribute to shaping the pattern of variation across mtDNA genomes and human populations (refer to Section 2.5).

The fixation or loss of specific mtDNA lineages or haplotypes in human populations is mainly attributed to the influence of random genetic drift (refer to Section 2.5.4). The occurrence of continent-specific lineages can therefore be viewed as evidence of the influence of random genetic drift on mitochondrial genetic variation and the different mitochondrial haplogroups observed on the African, European, Asian and American

continents were discussed in Section 3.3. In the current investigation, it was observed that the 30 Khoi-San individuals carried macrohaplogroup L lineages (L0, L2 and L3) which are predominantly observed in individuals of African descent (Chen *et al.*, 2000). This further supported the occurrence of continent-specific lineages as the Khoi-San are a population confined to Africa. Another feature related to the occurrence of continent-specific haplogroups is the exclusion criteria which consist of a list of SNPs that are used to distinguish mitochondrial haplogroups from one another (refer to Appendix D). In the current investigation the exclusion criteria were used to ensure that the mitochondrial sequences studied were indeed L-specific sequences. Based on the 30 Khoi-San sequences, as well as other L0 sequences added to the global L0-specific haplogroup network, the exclusion criteria currently used should be updated to reflect the information obtained from these recently generated L0 sequences, specifically regarding the use of A1438G and A2706G as exclusion criteria for L-specific sequences. These two SNPs should be excluded from the list of exclusion criteria as these SNPs are absent in many of the mitochondrial sequences classified within the L0a and L0b sub-haplogroups (refer to Section 5.8).

The effect of random genetic drift on genetic variation is determined by the effective population size, i.e. the number of successful breeding individuals (refer to Section 2.5.4). The effective population sizes for African and non-African populations have been estimated using autosomal, Y chromosome and mtDNA data and the estimates obtained have varied. The differences in the effective population sizes calculated can be attributed to the use of different methods, and the fact that the estimates were determined for different points in time (Tenesa *et al.*, 2007). Based on mtDNA, it has been estimated that the effective population size of African populations was in the range of 1,000 to 6,000 females (Rogers and Harpending 1992; Harpending *et al.*, 1993; Sherry *et al.*, 1994), for a population ca. 200,000 YBP (ca. 10,000 generations ago). In the current investigation, the effective population size was not calculated due to the relatively simple population expansion model employed (refer to Section 5.11.4). A recent investigation by Tenesa *et al.* in 2007, employing a relatively small sample of individuals combined with high-density genome-wide marker genotyping of human chromosomes, inferred ancestral population sizes based on the observed amounts of linkage disequilibrium, and provided estimates of effective population sizes of ca. 3,100 for non-African populations and ca. 7,500 for the African population.

The effect of random genetic drift on genetic variation is also influenced by migration (refer to Section 2.5.4). Human migration can influence mitochondrial DNA variation in two ways. Firstly by separating groups, thereby leading to increased divergence between populations, or secondly by integrating groups and thus decreasing the divergence between populations (Cavalli-Sforza and Feldman, 2003). Human migrations can involve a few individuals or larger groups of individuals. The migration of groups into new geographic regions can over time lead to genetic divergence between the groups and the population from which the groups originated. Migrations of a group of individuals are often followed by rapid population expansion events. In the current investigation, population expansion events were observed in datasets representing the African population (datasets 1b and 2b), the L3 haplogroup (dataset 5), the L2 haplogroup (dataset 6), and the L0 haplogroup (dataset 8a), while the datasets representing hunter-gatherer populations (dataset 3a and 4b) remained constant in size over a long period of time and did not exhibit evidence of population expansion events (refer to Section 5.11.4). The population growth observed in datasets 1b, 2b, 6 and 8a could be attributed to dispersals and differentiation of mtDNA haplogroups within Africa, while the population expansion observed in dataset 5 could be due to an increase in numbers of L3 haplogroup carriers which subsequently migrated from Africa to colonise the rest of the world. This small group of individuals was subjected to an initial population bottleneck event, followed by subsequent population expansion (Forster *et al.*, 2001; Macaulay *et al.*, 2005). The observed population expansion in datasets 1b, 2b, 5 and 6 therefore supports the “Recent African Origin” model, while the lack of evidence for population expansion in datasets 3a and 4b which contained Khoi-San and !Kung sequences is in agreement with previous observations for hunter-gatherer populations (refer to Section 5.11.4).

Further evidence for human migration in the current investigation is provided by the evaluation of the current geographic distribution of macrohaplogroup L in Africa (refer to Sections 5.9.4 and 5.10.2.2). Although the geographic distribution presented in the current investigation is dependent on the dataset used, as well as the specific ethnic and geographic information available for each individual sequence included in the dataset, and any conclusions drawn should be confirmed by the use of additional sequence data, it is evident that some sub-haplogroups exhibit specific geographic distributions. It was observed that two L0 sub-haplogroups, namely L0aA and L0b, were observed mainly in southern Africa in the Khoi-San population. The information generated about the distribution of macrohaplogroup L lineages, and specifically the distribution of the

L0 haplogroup, in Africa will be useful in the design of sampling strategies for future investigations of mitochondrial variation in African populations.

Population substructure is another aspect that affects genetic variation (refer to Section 2.5.4). Phylogenetic analyses of mtDNA haplotypes indicate that African populations contain highly divergent lineages that imply a long history of population subdivision (Chen *et al.*, 1995; Watson *et al.*, 1997; Chen *et al.*, 2000). In the current investigation, evidence of the presence of population substructure in African populations can be observed when evaluating the ethnic and geographic origin of mitochondrial sequences included in the phylogenetic trees and the global L0-specific haplogroup network (refer to Appendix G). When studying the information contained in Appendix G, it is evident that the mitochondrial sequences were obtained from a variety of different ethnic or language groups, as well as from different geographic regions. In addition, the global L0-specific haplogroup network contained a number of different lineages, notably L0aA, L0aB, L0aC and L0b, which contained Khoi-San individuals. The presence of a number of closely related, yet different lineages in both the Southern and Eastern African Khoi-San populations also suggests the presence of population substructure within these populations. At present, there is however insufficient information to support a subdivision of the Southern and Eastern African Khoi-San populations into sub-populations based on L0 sub-haplogroup designation.

Natural selection is another factor that influences genetic variation. Two types of natural selection play a role in the evolution of mitochondrial genomes (Meiklejohn *et al.*, 2007), namely negative or purifying selection and positive or adaptive selection (refer to Section 2.5.5). Purifying selection has been identified as the predominant factor influencing human mtDNA variation due to the essential function of the mtDNA proteins (refer to Section 2.5.5). In the current investigation, the determination of codon usage bias indicated the presence of a generally low level of bias within the Khoi-San mitochondrial DNA genomes, which suggests a role for natural selection (refer to Section 5.11.1.2). Further support for the influence of natural selection in shaping mitochondrial variation is the effect of purifying selection observed for haplogroups L and L0 (refer to Section 5.11.5).

Based on a number of investigations, it has also been proposed that an estimated 25% of all mtDNA variation can be attributed to adaptation which has allowed humans to adjust to varying caloric availability and differences in the demand for mitochondrial energy

generation via ATP production compared to heat generation (refer to Section 2.5.5). This implies that after the initial migration from Africa, humans adapted to their new environment outside of Africa regarding aspects such as food availability and climate so that people of recent African descent and people of recent European or Asian descent may present with different susceptibilities to the same diseases. Young *et al.* in 2005 demonstrated that differences in susceptibility to hypertension could be due to discrepancies in the level of exposure to selection pressure during the expansion out of Africa, and that climate was the most important source of selection pressure. This implies that in the modern world, when ancient, adapted genotypes move from the environment they are adapted to, into a new environment for which they are genetically unprepared, a myriad of potential health challenges can come to the fore as alleles that were adaptive in their ancestral environment are maladaptive in their current environment (Young *et al.*, 2005). The information presented in the current investigation is therefore vital in providing information about the genetic background on which diseases present, and could provide additional information about the ancestral mitochondrial genotypes. This information would be of value in studies of adaptation and the health challenges currently faced by the global population.

It must be concluded that the genetic variation observed in the human mitochondrial genome in different populations is determined by the interaction of the different processes discussed in this section. At present, with the evolutionary models currently available, it is difficult to distinguish between the relative contributions of each of the different factors. The additional data presented in the current investigation can however contribute to the formulation of more complex human evolutionary models as the Khoi-San represent an ancient population and can thus be used as a starting point in the design of such models.

6.2.3 Measurable outcomes that are affected by the level of genetic variation present in mitochondrial genomes

The genetic variation present in mitochondrial genomes in specific populations can be measured and compared by determining genetic diversity measures such as nucleotide diversity, or by calculating the times when the most recent common ancestors diverged to form the separate mitochondrial haplogroups currently observed in different populations in different regions of the world. Nucleotide diversity and TMRCA estimates are therefore included in the model proposed in Figure 6.1.

Mitochondrial DNA sequence data indicates that African populations have the highest genetic diversity, followed by Asian, European and Native American populations (Jorde *et al.*, 2000). Non-African populations harbour a subset of the genetic diversity present in Africa (Tishkoff and Williams, 2002). In the current investigation, it was observed that the nucleotide diversity values determined for datasets 1b, 2b, 3a and 4b were similar to that observed in other African populations (refer to Section 5.11.2). The high genetic diversity generally observed in African populations, as well as the high diversity values observed in the current investigation, support the conclusion that modern humans originated in Africa (refer to Section 3.1).

Additional support for a high level of genetic variation in mitochondrial genomes in the African population is provided by phylogenetic analyses of mitochondrial haplogroups. In the current investigation, the analysis of the global phylogenetic trees consisting of macrohaplogroup L sequences involved a number of important observations. A branch consisting of L5 sequences clustered closer to the chimpanzee outgroup than the L0 haplogroup branch (refer to Section 5.9). This observation is in contrast to a number of other investigations that observed the L0 branch to be closest to the outgroup (refer to Section 5.9). It was concluded that the true evolutionary relationship of L5 with other L haplogroups has been obscured due to the rarity of sequences belonging to the L5 haplogroup and the current dearth of sequence data from eastern Africa, and that the results obtained in the current investigation may therefore support a previously unreported close relationship between L0 and L5 (refer to Section 5.9.2.1). This will however only be confirmed through the addition of more L5 sequences, as well as additional sequences from East Africa, to the phylogeny of macrohaplogroup L.

As already mentioned, the addition of the 30 Khoi-San sequences generated in the current investigation to the global phylogenetic trees delineated the phylogeny of macrohaplogroup L further by expanding previously identified branches through the creation of new subgroups. In the L0 branch, subdivisions namely L0aA1, L0aA2, L0bB1, L0bB2, L0bB2a, L0bB2b and L0bB3 were observed, while in the L2 and L3 branches the Khoi-San sequences increased the number of branches within the L2b, L3a1 and L3b sub-haplogroups (refer to Section 5.9).

Further support for a high level of genetic variation in mitochondrial genomes in the African population is provided by the analysis of the global L0-specific haplogroup network. In the current investigation, 22 of the Khoi-San sequences were added to the global L0-specific

haplogroup network. These sequences were present in the L0a and L0b branches and expanded previously identified branches (L0aA1, L0aA2, L0bA2, L0bB2a1, L0bB2a3, L0bB2b and L0bB3a), as well as giving rise to two previously unidentified branches (L0bB1b1 and L0bB2a1).

The new subdivisions and branches generated by the addition of the Khoi-San sequences to the global phylogenetic trees, and the global L0-specific haplogroup network, represent potential new lineages in the L0 haplogroup, as well as the L2b, L3a and L3b sub-haplogroups, and is an indication of the genetic diversity present in the Southern African Khoi-San population. The addition of the sequence data generated for the 30 Khoi-San individuals to the global phylogenetic trees and the global L0-specific haplogroup network therefore increases our understanding of the genetic diversity present in macrohaplogroup L, specifically the L0 haplogroup, and sheds light on the contribution made by the Southern African Khoi-San population to the genetic diversity of the African population as a whole (refer to Sections 5.9 and 5.10). In addition, previously unknown lineages identified in the L0, L2 and L3 haplogroups, due to the inclusion of the 30 Khoi-San sequences in phylogenetic or network analyses, have implications for the study of the genetic diversity in the African population, as well as the investigation of these haplogroups, and it is therefore necessary to include these lineages when designing sampling strategies for future investigations.

It is important to distinguish between the age of a population and the age of a haplogroup observed within that population as the two estimates do not describe the same value and are therefore not equal. The TMRCA estimates of haplogroups within a population can however be used to compare different populations. The ages of the major haplogroups observed in African, Asian, European and Native American populations have been determined and it was observed that the African haplogroups are the oldest, thereby implying that African populations are the oldest (refer to Section 3.3). In the current investigation the TMRCA estimates for the L0 haplogroup, sub-haplogroups and lineages within sub-haplogroups, were estimated by using the global L0-specific haplogroup network (refer to Section 5.11.6).

A TMRCA estimate of 123,946 to 129,044 YBP was determined for the L0 haplogroup which is in agreement with estimates previously reported in other genetic investigations of mtDNA, while the TMRCA estimates of the two previously unreported lineages, namely L0bB1b1 and L0bB2a1, were determined as $46,241 \pm 10,899$ and $15,414 \pm 8,899$ YBP,

respectively (refer to Section 5.11.6). TMRCA estimates determined for L0b, L0c1 and L0c1B were also broadly consistent with previously published TMRCA, while the TMRCA estimates determined for L0a, L0aA, L0aB, L0aC, L0bB, L0c, L0c2, L0c2B, and L0c2D were more ancient which was attributed to the use of different sample sizes in the calculation process compared to previous investigations (refer to Section 5.11.6). The presence of two new lineages, as well as their estimated ages, provides evidence for the prolonged presence of Khoi-San speaking populations in southern Africa, and indicates the genetic diversity present in this population. The TMRCA for the L0 haplogroup corresponds to previously reported paleontological dates for the emergence of modern *Homo sapiens* and therefore provides support for the “Out of Africa” theory of modern human population expansion and colonisation of the world.

The identification of new lineages, as well as subdivisions in reported lineages which have generated new subgroups, necessitates the construction of a classification scheme that can adequately summarise and organise the genetic diversity evident in African mitochondrial haplogroups. The classification scheme used throughout the current investigation fell short of this expectation as was demonstrated by the number of sequences that could not be completely classified (refer to Section 5.8). Another classification scheme, referred to as the alternative classification scheme in the current investigation, is generally used (refer to Section 3.3.1.1). This scheme is based mainly on SNPs in the control region, as well as SNPs identified in the coding region through the use of RFLPs. The use of control region SNPs for classification is not optimal, as many of the nucleotide positions where the SNPs occur are mutational hot spots and are subjected to repeated forward and back mutations. Based on the global L0-specific haplogroup network constructed in the current investigation, a revised classification scheme was proposed for the L0 haplogroup which was based on coding region SNPs only (refer to Section 5.12). The revised classification scheme for the L0 haplogroup should be incorporated into a standardised classification scheme for all the mitochondrial haplogroups (refer to Section 6.4.3).

In the previous paragraphs in this section, the genetic variation present in populations, i.e. intra-population variation, has been discussed. The genetic variation observed between different populations is however also of interest as it provides information regarding aspects such as human migration and the effect of natural selection. The genetic variation observed between populations is less than the genetic variation observed within a population (Jorde *et al.*, 2000). If the African continent is viewed as the starting point, it is

observed that the inter-population genetic variation increases between populations as the distance from Africa increases (Oppenheimer, 2003), i.e. the inter-population genetic variation between African populations and Native American populations will be greater than that observed between African and Asian populations.

6.2.4 Standards used to investigate genetic variation present in mitochondrial genomes

The rCRS provides a standard in the investigation of molecular evolution and mitochondrial disease to which all human mitochondrial DNA sequences can be compared. The rCRS is however of European origin and has been classified as belonging to haplogroup H. It is therefore not a globally representative reference sequence (refer to Section 3.5). In addition, it has been observed that when complete mitochondrial genome sequences of individuals from African descent are compared to the rCRS, a greater number of differences are identified than when complete mitochondrial genome sequences from European descent are compared to the rCRS (Horai *et al.*, 1995). A large number of differences are therefore expected for comparisons between this sequence and sequences from African populations. This was supported in the current investigation by the identification of a large number of alterations when comparing the 30 Khoi-San sequences to the rCRS (refer to Section 5.7). The large number of differences identified in comparisons of African mtDNA sequences with the rCRS therefore makes it extremely difficult to differentiate between a polymorphism and a deleterious mutation when investigating mtDNA sequences of African origin.

In order to distinguish between potential pathogenic alterations and normal variation, it is necessary to identify population-specific polymorphisms. This can be done by compiling a consensus sequence for the Southern African Khoi-San population studied in the current investigation as a consensus sequence will provide a summary of the observed sequence variation in this population (Page and Holmes, 1998). As the Khoi-San are one of the most ancient African populations (refer to Section 3.7), this consensus sequence can be used in the study of all African populations to discriminate between harmless polymorphisms and potential pathogenic mutations.

A consensus sequence can also be used in the study of global human mtDNA diversity and for this purpose two human mitochondrial consensus sequences have been constructed by Marzuki *et al.* in 1991 and Carter in 2007. A consensus sequence based on

the Southern African Khoi-San population however has the advantage that it satisfies the two requirements regarding age and genetic diversity necessary for the study of the evolution of the human mitochondrial genome and mitochondrial disease, as it is based on an ancient human population and summarises the genetic variation present in the human mitochondrial genome prior to migration into new environments and the divergence of mitochondrial haplogroups (refer to Section 5.13).

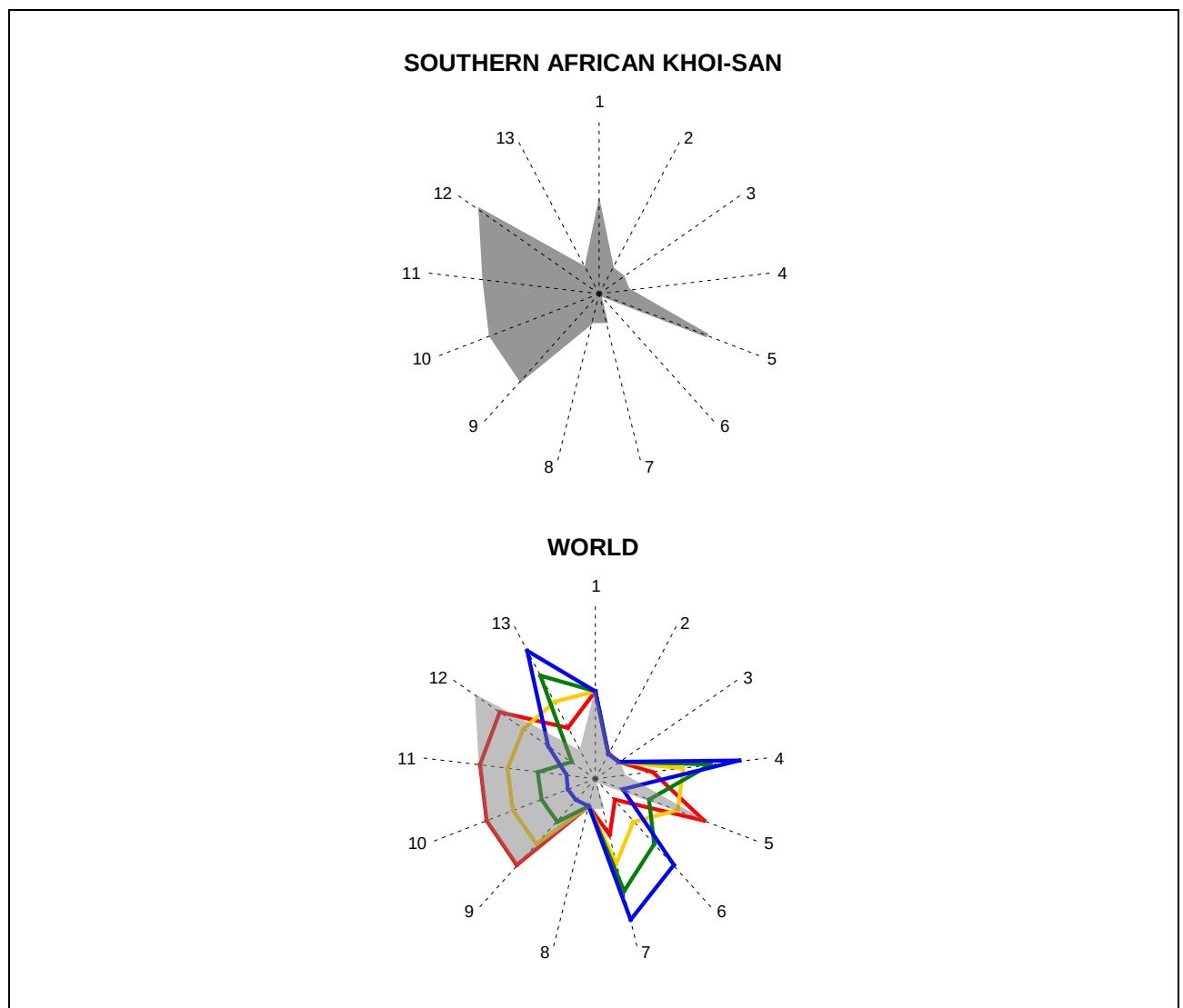
In comparisons of mitochondrial haplogroups observed outside of Africa, i.e. the more recently diverged haplogroups, with the Khoi-San consensus sequence, more differences should be observed than in comparisons of haplogroups observed within Africa with the Khoi-San consensus sequence. This trend is the opposite to that observed for the rCRS where sequences from haplogroups observed within Africa exhibit more differences compared to the rCRS than sequences from haplogroups located outside of Africa (refer to Section 5.13).

The Khoi-San consensus sequence can therefore be used as a tool to investigate the evolution of the mitochondrial genome, and thus the manner in which the different mtDNA haplogroups adapted over time, thus allowing for the exclusion of population-specific alterations during the identification of possible pathogenic alterations. In addition, the Khoi-San consensus sequence appears to be a good approximation of the variation present in the African population and thus provides an alternative to the currently used African reference sequences, namely NC_001807 and D38112 (refer to Section 5.13).

6.2.5 Update of proposed model for genetic variation observed in human mitochondrial genomes

The model proposed in Figure 6.1 was updated to highlight the unique contribution made by the current investigation. The implications of this contribution are discussed below (refer to Section 6.3). The updated model is presented in Figure 6.2.

Figure 6.2 Updated model indicating the contribution made by the unique population characteristics of the Khoi-San from southern Africa



The genetic variation profiles of the different populations are indicated with shapes or lines using different colours where ■ = Khoi-San, ■ = Africa, ■ = Asia, ■ = Europe and ■ = Native Americans. The different factors that are involved in generating, shaping and evaluating genetic variation in the mitochondrial genomes of the different populations are indicated using numbers where 1 = mutation rate (substitutions per site per year), 2 = cytoplasmic segregation, 3 = random genetic drift, 4 = natural selection, 5 = effective population size (number of breeding females), 6 = population size changes, 7 = human migration, 8 = population substructure, 9 = intra-population nucleotide diversity, 10 = inter-population nucleotide diversity, 11 = estimated age of population (years before present), 12 = reference sequence, 13 = consensus sequence.

The following aspects in the current investigation represent unique contributions to the existing body of knowledge in the field:

- Addition of 30 Khoi-San mtDNA sequences representative of one of the most ancient populations in Africa.
- Addition of more African mtDNA sequences to the global human mtDNA sequence dataset.
- Elucidation of branching in the L0 haplogroup of the global macrohaplogroup L mtDNA phylogenetic tree.
- Update of the global L0-specific haplogroup network.

- e. Proposed revised classification scheme for the L0 haplogroup based on the new branches in the global L0-specific haplogroup network and the global macrohaplogroup L phylogenetic trees.
- f. Compilation of a consensus sequence for the Khoi-San population of southern Africa. This consensus sequence represents an ancient African population, and can be employed as an alternative to the use of the rCRS in studies of mitochondrial disease in African populations. The consensus sequence provides a baseline to discriminate between population-specific polymorphisms and possible disease-causing mutations.

6.3 IMPLICATIONS OF THE PROPOSED MODEL FOR GENETIC VARIATION OBSERVED IN HUMAN MITOCHONDRIAL GENOMES

The proposed model for genetic variation observed in human mitochondrial genomes provides a reference for the evaluation of human mitochondrial genetic variation. A number of questions however remain unanswered when investigating the genetic variation in African populations. These questions include the following:

- a. The place of origin in Africa of the most ancient African mitochondrial haplogroup and population (refer to Section 6.3.1).
- b. The estimated TMRCA of the most ancient African mitochondrial haplogroup (refer to Section 6.3.2).
- c. The events within Africa leading to the L3 haplogroup being the haplogroup present in the small population of modern humans that migrated from Africa and populated the rest of the world (refer to Section 6.3.3).
- d. The role of the mitochondrial genetic background of an individual in the presentation and progression of disease (refer to Section 6.3.4).
- e. Practical considerations of the use of the Khoi-San consensus sequence in conjunction with, or as replacement for, the rCRS in order to identify polymorphisms and disease-causing alterations (refer to Section 6.3.5).

6.3.1 The place of origin in Africa

The general consensus is that modern humans originated in Africa (refer to Section 3.1), and the results generated in the current investigation provides additional support for this interpretation (refer to Section 6.2). The precise location or region in Africa where anatomically modern humans first appeared is however still a matter of debate as the

archaeological evidence supports the presence of ancient populations in both eastern and southern Africa (Forster, 2004; Mellars, 2006).

In the current investigation, a Khoi-San cohort consisting of 30 Southern African Khoi-San individuals was studied. Investigation of the individual mitochondrial genome sequences revealed that the majority of Khoi-San individuals in the cohort carried the L0aA (L0k¹) and L0b (L0d) mitochondrial haplogroups. These mitochondrial haplogroups are observed primarily in the Khoi-San and belong to the deepest clades recognised among modern humans (Chen *et al.*, 2000; Tishkoff *et al.*, 2007; Behar *et al.*, 2008). The Khoi-San is considered to be a unique relic of the hunter-gatherer lifestyle and is one of the most ancient populations in Africa. The investigation of the Khoi-San population should therefore shed light on the timing and location of the emergence of modern man in Africa.

In the current investigation, the TMRCA estimates, as well as the current geographic distribution of different L0 sub-haplogroups and lineages within sub-haplogroups observed in the Khoi-San, were used to suggest patterns for the dispersal of these sub-haplogroups and lineages within sub-haplogroups across the African continent (refer to Section 5.11.6). The oldest sub-haplogroups and lineages, namely L0a, L0aA, L0aB, L0aC and L0b, are located in southern and eastern Africa. The L0aA (L0k) and L0b (L0d) sub-haplogroups are observed in Southern African Khoi-San speakers, while the L0aB (L0d3) sub-haplogroup is present in Eastern African Khoi-San speakers. If it is assumed that L0aB (L0d3) and L0b (L0d1 and L0d2) diverged from a shared mtDNA haplogroup observed in an ancestral Khoi-San speaking population, the ancestral population that gave rise to Khoi-San speakers may have originated in East Africa, and members of this population subsequently dispersed to southern Africa. The two populations then became genetically isolated from one another, most likely due to a marked dry period in southern Africa which caused the contraction of available geographic ranges.

The argument presented in the previous paragraph is supported by the hypothesis presented by Behar *et al.* in 2008. This hypothesis assumes the existence of a single ancestral Middle Stone Age *Homo sapiens* population in eastern or southern Africa in which the different L haplogroups co-evolved and the localisation of the L0aA (L0k) and L0b (L0d) sub-haplogroups observed in the Khoi-San of southern Africa is the result of a migration event followed by isolation and genetic drift which led to the enrichment of

¹ The haplogroup classifications presented in brackets refer to the classifications according to the alternative classification scheme (refer to Section 3.3.1.1).

certain lineages in the Khoi-San and the extinction of these lineages in non-Khoi-San populations (Behar *et al.*, 2008).

Further support for East Africa as the location where anatomically modern man evolved from an ancestral population is provided by the high mtDNA diversity observed in this region. Recent investigations have identified a number of different and rare mitochondrial haplogroups in eastern Africa such as L0d, L0f, L4a, L4g, L5 and L6 (Torroni *et al.*, 2006; Gonder *et al.*, 2007; Tishkoff *et al.*, 2007). In the current investigation, L5 replaced L0 as the most basal group in the phylogenetic trees and this could be attributed to the fact that L5 is a very ancient haplogroup (refer to Section 5.11.6). The location of the L5 group, consisting of nine sequences, in the phylogenetic trees, as well as the current geographic distribution of this haplogroup, suggest that as more mitochondrial sequences become available from countries located in eastern Africa, it is possible that the oldest mitochondrial haplogroups will be observed in East Africa, which will provide additional support for East Africa as the location where anatomically modern man emerged.

6.3.2 The estimated TMRCA of the most ancient African mitochondrial haplogroup

At present, the L0 haplogroup is viewed as the most ancient mitochondrial haplogroup as it is observed at the root of the human mtDNA tree (Chen *et al.*, 1995; Ingman *et al.*, 2000; Mishmar *et al.*, 2003; Ruiz-Pesini *et al.*, 2004; Gonder *et al.*, 2007). TMRCA estimates for haplogroup L0 based on coding region mtDNA sequence data ranges between a minimum of 92,000 to a maximum of 171,500 YBP (refer to Section 5.11.6). In the current investigation, the TMRCA for the L0 haplogroup was estimated between 123,946 and 129,044 YBP based on the global L0-specific haplogroup network (refer to Section 5.11.6). This TMRCA estimate was aligned to previous estimates, although the range of the estimate was narrower than that of previous estimates, and it is therefore more precise.

As previously mentioned, the L5 group was located at the base of the phylogenetic trees constructed in the current investigation (refer to Section 5.9). L5 is an ancient haplogroup and is mainly observed in eastern Africa, similarly, other ancient mitochondrial haplogroups may be present in eastern African which have not been identified to date due to the paucity of mitochondrial sequence data from eastern Africa (refer to Section 6.3.1). A recent investigation determined a TMRCA estimate for the L5 haplogroup of between 107,300 and 151,500 YBP based on coding region data of five mitochondrial sequences of which one sequence from an Mbugwe individual was highly divergent compared to the

other sequences obtained from Sandawe individuals (Gonder *et al.*, 2007). When this TMRCA estimate is compared to the estimates available for the L0 haplogroup, it is evident that this estimate is either more ancient or similar to the TMRCA estimates for the L0 haplogroup (refer to Section 5.11.6). It is however likely that this observation can be attributed to differences in the sizes of samples used for calculation and different methods used to perform the calculations.

The identification and the determination of the TMRCA estimate of the most ancient African mitochondrial haplogroup are therefore dependent on two factors. Firstly, adequate sampling must be performed in all geographic regions of Africa in order to include all mitochondrial haplogroups and to obtain sufficient numbers of sequences representing each mitochondrial haplogroup/sub-haplogroup for analyses, as incomplete sampling and insufficient sequence data will obstruct the process of identifying and dating ancient mitochondrial haplogroups. Secondly, methods should be used that take into account population and demographic factors so that the determined TMRCA values are as close to the true values as possible (refer to Section 6.4). The 30 Khoi-San mitochondrial genome sequences generated in the current investigation assist in the estimation of the most ancient mitochondrial haplogroup by adding to the total number of L0a (L0k) and L0b (L0d) sequences generated to date.

6.3.3 The migration of the L3 haplogroup from Africa to the rest of the world

Watson *et al.* in 1997 investigated population expansion events in Africa and concluded that ca. 60,000 to 80,000 YBP, one or a small set of closely related mitochondrial lineages expanded in Africa and gave rise to the L2 and L3 haplogroups. This re-expansion led to the first successful modern human migration from Africa between 46,000 to 62,000 YBP (Forster *et al.* 2001) and two possible routes, a northern and a southern route, have been proposed for this migration (Forster, 2004). All non-Africans are descended from an L3 sub-haplogroup, namely L3a, which is specific to eastern Africa (Watson *et al.*, 1997). This implies that humans carrying this L3 sub-haplogroup migrated from eastern Africa and subsequently spread into Eurasia where the L3 haplogroup was the originator of macrohaplogroups M and N, which are the progenitors of all recognised non-African mitochondrial haplogroups.

The observation of many diverse, phylogenetically isolated mitochondrial lineages scattered across the African continent, begs the question as to why only one mitochondrial

haplogroup survived the migration from Africa into Eurasia to give rise to all non-African mitochondrial haplogroups. The survival of a single lineage has been attributed to the small size of the human population that migrated from Africa, the effect of random genetic drift within this population, and the possibility that a sub-population within the population of anatomically modern humans carrying this haplogroup acquired an advantage which allowed this subgroup to expand (Watson *et al.*, 1997). This advantage could include a coping mechanism in response to an environmental change, or behavioural innovations such as the ability to communicate effectively, the design and use of new technologies and cognitive development (Cavalli-Sforza *et al.*, 1994; Watson *et al.*, 1997; Forster, 2004; Mellars, 2006). Other factors that may also have contributed to the migration of humans from Africa include competition between groups within the population of anatomically modern humans for available resources, and climate change that forced humans to migrate into previously uninhabited territories (Forster, 2004; Mellars, 2006; Carto *et al.*, 2009).

The findings presented in the current investigation assist in providing data about the maternal lineages present at the dawn of the mitochondrial phylogeny and the distribution of these mitochondrial haplogroups across the African continent. It therefore reveals some aspects regarding the shape of the mitochondrial landscape prior to population expansions that gave rise to the human migration from Africa. This data is invaluable in the study of African mitochondrial haplogroups, the geographic distribution of haplogroups, early population expansion events and the identification of population substructure within the African population as a whole. In addition, the data presented in the current investigation can assist in providing answers as to why the L3 haplogroup was carried to other continents instead of haplogroups such as the L0 haplogroup. The findings presented in the current investigation provide further evidence that individuals carrying L0a and L0b sub-haplogroups migrated to southern Africa where these haplogroups became localised and are predominantly observed in Khoi-San speaking populations, while individuals harbouring the L0c sub-haplogroup dispersed from east to west Africa (refer to Section 5.11.6). The migration of L0a and L0b sub-haplogroups to southern Africa can be attributed to factors such as climate fluctuations that facilitated the movement into new environments and the contact of hunter-gatherer populations carrying these lineages with food producing populations.

6.3.4 The role of the mitochondrial genetic background

It has been demonstrated in a number of mitochondrial diseases that the mitochondrial haplogroup on which the pathogenic mutation is presented, plays a role in the progression and severity of the disease (refer to Section 3.4). This is attributed to the fact that as humans diversified within Africa and then migrated from Africa into new environments around the world, alterations accumulated in the human mitochondrial genome along radiating maternal lineages to give rise to region-specific mitochondrial haplogroups. These region-specific haplogroups exhibit ancient alterations which permit the carriers of these haplogroups to adapt to their respective environments (Mishmar *et al.*, 2003; Ruiz-Pesini *et al.*, 2004).

These region-specific haplogroups will therefore exhibit functional differences and it is likely that a variant that confers an advantage in a specific climatic and dietary environment may be maladaptive when exposed to a different environment. This implies that ancient regionally beneficial mtDNA variants could contribute to the occurrence of modern bioenergetic disorders such as obesity, diabetes, hypertension, cardiovascular disease, and neurodegenerative diseases when individuals migrate to new geographic regions and adopt new lifestyles (Mishmar *et al.*, 2003; Wallace, 2005).

The information presented in the current investigation, obtained from one of the most ancient human populations in Africa, is therefore essential in providing information about the ancestral human mitochondrial genotype. Knowledge about the ancestral mitochondrial haplotype from which all other mitochondrial haplogroups evolved, could assist in studies of the adaptation of the human mitochondrial genome to different environments, as well as in the study of health challenges currently faced by the global population as it provides additional information about the genetic background on which these diseases present.

6.3.5 Practical considerations of the use of the Khoi-San consensus sequence

The rCRS is currently used as a standard in the majority of investigations of the evolution of the human mitochondrial genome, and mitochondrial disease. Under ideal conditions, the standard reference used in these types of investigations would be based on an ancient population and would summarise the variation present in human populations prior to their migration from Africa. The rCRS is therefore lacking as it is based on mitochondrial

haplogroup H which is observed in comparatively young European populations, and does not take into account the high mtDNA diversity generally observed in African populations.

In contrast, the Khoi-San consensus sequence constructed in the current investigation provides a summary of the variation present in the mitochondrial genome sequences of Southern African Khoi-San individuals which belong to one of the most ancient human populations in Africa (refer to Section 3.7). It can therefore be suggested that the Khoi-San consensus sequence provides the closest approximation of the ancestral mitochondrial genome sequence currently available. The Khoi-San consensus sequence, based on such an ancient population, provides a glimpse into our genetic past, allowing the identification of potentially pathogenic variants and prediction of which individuals are likely to respond to therapy based on knowledge of the underlying cause, thereby focusing the use of limited therapeutic resources.

The Khoi-San consensus sequence constructed in the current investigation has a number of attributes that facilitate its potential uses (refer to Section 5.13). These attributes include the following:

- a. The Khoi-San consensus sequence is a population-specific consensus sequence, not a haplogroup-specific consensus sequence. Although it is based on L0 sequences belonging to the L0a and L0b sub-haplogroups, sequences from other L0 sub-haplogroups such as L0c1 and L0c2 were not included in the set of mitochondrial sequences used to compile the consensus sequence and the consensus sequence therefore does not represent the L0 haplogroup as a whole.
- b. The inclusion of only L0a and L0b sub-haplogroup sequences in the Khoi-San consensus sequence with the concomitant exclusion of L2 and L3 haplogroup sequences. A recent investigation of the maternal lineages present in the Khoi-San population indicated that, apart from extinct lineages, the mtDNA gene pool of the Middle Stone Age Khoi-San ancestor was likely limited to L0a and L0b lineages (Behar *et al.*, 2008).
- c. Presence of ambiguity codes in the Khoi-San consensus sequence in order to retain polymorphism data contributed by both the L0a and L0b sub-haplogroups present in the Southern African Khoi-San population. The use of ambiguity codes in the Khoi-San consensus sequence should reduce the number of alterations identified as possible pathogenic mutations.
- d. Observation of a greater number of differences between the Khoi-San consensus sequence and younger mitochondrial haplogroups observed in non-African

populations than between the consensus sequence and older mitochondrial haplogroups detected in African populations. This is in contrast to the trend observed in comparisons of mitochondrial haplogroup sequences with the rCRS. A smaller number of alterations that are possibly pathogenic are therefore observed when using the Khoi-San consensus sequence instead of the rCRS in comparisons with mitochondrial sequences of African origin.

- e. Provides a list of putative population-specific alterations for the Southern African Khoi-San population. This list of potential population-specific alterations can be used to distinguish between alterations that are polymorphic or potentially pathogenic.
- f. In comparison with other available human consensus sequences, the Khoi-San consensus sequence was preferred over the Carter consensus sequence for use in comparisons with mitochondrial sequences obtained from individuals of African descent. The small number of differences between the Khoi-San consensus sequence and the Carter consensus sequence, based on a global dataset consisting of 827 mitochondrial sequences, also indicated that the Khoi-San sequence included most of the variation present in human mitochondrial genomes worldwide.

The attributes mentioned in the previous paragraph promote the use of the Khoi-San consensus sequence in a number of different applications (refer to Section 5.13). A list of the potential uses for the Khoi-San consensus sequence, on its own or in conjunction with the rCRS, is outlined below:

- a. Used in the study of the evolution of the human mitochondrial genome:
 - i. To investigate the sequential accumulation of alterations that gave rise to the different mitochondrial haplogroups present in the different regions of the world.
 - ii. To investigate the effect of purifying and adaptive selection over time as humans migrated into new environments and adapted to the challenges faced in these environments.
 - iii. To assign ancestral character states in investigations of coalescent times and effective female population size.
- b. Used to investigate mitochondrial diseases:
 - i. Provides a baseline for all extant mitochondrial sequences as the Khoi-San consensus sequence is the best approximation currently available of the ancestral mitochondrial genome sequence.

- ii. To exclude alterations specific to African populations.
 - iii. To distinguish between normal variation and potentially pathogenic alterations in mitochondrial sequences from African and non-African individuals.
- c. Used as an African reference sequence in comparisons with mitochondrial genome sequences obtained from indigenous Africans and individuals of African descent:
 - i. Provides a template of a typical African mitochondrial background.
 - ii. To replace NC_001807 and D38112 which are currently used as African reference sequences.
 - iii. To identify potentially pathogenic mutations and to discover polymorphisms.
- d. Used as a source of informative SNPs for haplogroup classification in the African population:
 - i. The Khoi-San consensus sequence provides a list of SNPs.

6.4 FUTURE DIRECTIONS IN THE STUDY OF GENETIC VARIATION OBSERVED IN HUMAN MITOCHONDRIAL GENOMES

The results generated in the current investigation highlight the numerous factors that together contribute to genetic diversity in human mitochondrial genomes. Several aspects were identified in the current investigation that need to be addressed in the future in order to further elucidate the role of specific factors in the generation of genetic diversity.

6.4.1 Sequencing of additional Khoi-San individuals

In the current investigation, 30 complete Khoi-San mitochondrial genome sequences were generated and partial sequences were determined for an additional 20 Khoi-San individuals (refer to Section 5.1). The results generated for the 20 Khoi-San individuals were not included in the current investigation as the 20 individuals were only partially sequenced due to difficulties encountered with PCR and sequencing of the Khoi-San samples (refer to Sections 5.3.3 and 5.6.4). The difficulties included low PCR amplification efficiency, as well as the presence of background noise, low signal strength, the insertion and deletion of nucleotides, excess dye peaks and dye blobs, and slippage in homopolymer regions observed during automated sequencing. There is a number of ways in which these difficulties can be overcome in the future.

The first approach would be to repeat the extraction of gDNA for the 20 Khoi-San samples, as some of the difficulties encountered with PCR and automated sequencing could be due

to the quality of the isolated gDNA (refer to Sections 5.3.3.1 and 5.6.5). This would involve culturing the lymphoblastoid cell lines of each of the 20 Khoi-San individuals which is a costly and laborious process. As an alternative approach, existing samples can be subjected to new technological processes. New sequencing technology has become available which will greatly decrease the time and costs needed to obtain complete mitochondrial genome sequences for these 20 individuals.

The Genome Sequencer FLX System™¹ is a fast, accurate and cost-effective high-throughput sequencing system (Margulies *et al.*, 2005; Roche Diagnostics GmbH, 2006a). The system consists of three steps, DNA library preparation and titration, emulsion-based clonal amplification (emPCR) and sequencing. This technology has the advantage that either gDNA or PCR products representing segments of the mitochondrial genome can be used as starting material at different stages in the three step process. This method however has the disadvantage that a large amount of starting material is required. An amount of 15 µg gDNA is needed, if gDNA is used as starting material (Margulies *et al.*, 2005) and as already mentioned, the amount of isolated gDNA available for the individual Khoi-San samples was limited (refer to Section 5.1.1). A better course of action would therefore be to start with the available gDNA and to generate PCR products representing segments of the mitochondrial genome. This can be done by following a combination of the methodologies of Torroni *et al.* (2001) and Sogin *et al.* (2006) with adaptations where necessary. Torroni and colleagues in 2001 used a PCR strategy that involved the generation of 11 overlapping PCR fragments which could be produced simultaneously by using the same annealing temperature for the different reactions. Following the methodologies employed by Torroni *et al.* in 2001 and Sogin *et al.* in 2006 will thus necessitate the design of a new PCR strategy which will involve the re-design of the PCR primers to contain the necessary sequencing adapters added to the 5' end of the forward primers and the 3' end of the reverse primers, as well as scaling up reaction volumes, and re-optimisation of the PCR to increase yield. The purified PCR products can subsequently be used as starting material in step two of the Genome Sequencer FLX System™. Step two, the emPCR step, permits simultaneous sequencing in both the forward and reverse directions which will contribute to ensuring the accuracy of the obtained sequence data. The process involves parallel sequencing from single molecules (clonal sequencing), which allows the accurate identification of alterations at the level of a single molecule in a rapid, cost-effective and highly sensitive manner (Roche Diagnostics GmbH, 2006b). This allows the detection of heteroplasmy which is difficult to observe when using the Sanger

¹ The Genome Sequencer FLX System is a trademark of 454 Life Sciences Corporation, Branford, CT, USA.

sequencing method unless the level of heteroplasmy exceeds a specific threshold value (refer to Appendix B).

The additional 20 Khoi-San mitochondrial genome sequences generated should be classified into haplogroups (refer to Section 4.9), added to the global phylogenetic trees (refer to Section 5.9) and the L0 sequences incorporated into the global L0-specific haplogroup network (refer to Section 5.10). This will allow for the confirmation of the branching patterns currently observed in the phylogenetic trees and mutational network, as well as further delineation of new branches. The addition of L0 sequences to the global L0-specific haplogroup network also allows the identification of nucleotide positions with unexpected alterations which are alterations that do not fit the SNP profile of the haplogroups, sub-haplogroups or lineages within the haplogroups. These alterations could be potential sequencing errors. In the current investigation such unexpected alterations were identified in KS_CGR_059 at nucleotide position 10,876 and 10,939, in KS_CGR_044 at nucleotide position 1,438, 2,706, and 4,769, and in KS_CGR_091 at nucleotide position 6,185. These alterations were however identified as exceptions to the general rule and were not sequencing errors (refer to Sections 5.8 and 5.10.2.3).

6.4.2 The global human mtDNA phylogeny

Complete human mitochondrial genome sequences belonging to African and non-African mitochondrial haplogroups are being generated at an increasing rate by researchers, as well as commercial companies involved in uncovering the genetic history of individuals. A number of L-specific sequences have been submitted to GenBank® after June 2007 which was stipulated as the cut-off date for inclusion in the current investigation (refer to Section 4.8). These sequences include 56 L sequences from the United States of America and the United Kingdom (Herrnstadt *et al.*, 2002a), 309 L sequences from East, North, South, West and sub-Saharan Africa, Europe, the Mediterranean, the Near East, and South West Asia (Behar *et al.*, 2008), 27 sequences from Central Africa (Quintana-Murci *et al.*, 2008) and 16 sequences from West-Central Africa (Černý *et al.*, 2009). In the future, after the addition of these L sequences to the global macrohaplogroup L dataset (refer to Section 4.10), the phylogenetic and network analyses presented in the current investigation (refer to Sections 5.9 and 5.10) should be repeated.

The traditional phylogenetic methods employed in the current investigation are increasingly being replaced or complimented by more sophisticated methods such as

Bayesian phylogenetic inference that take into account more variables in order to model the complex reality of sequence evolution (Huelsenbeck *et al.*, 2001; Holder and Lewis, 2003), while phylogenetic networks are increasingly constructed using computer software packages such as Network provided by Fluxus Technology Ltd (Gonder *et al.*, 2007). These methods are however computationally expensive, and although the Bayesian phylogenetic inference method permits the modelling of many parameters to simulate complex evolutionary models, a number of investigations have demonstrated that traditional phylogenetic methods may be better suited to determine the phylogeny for large datasets as these methods are faster and easier to implement (Holder and Lewis, 2003). A number of different methods are thus currently available for phylogenetic analysis, and the choice of method will therefore depend on the aims of the specific investigation, as well as other factors such as the available computer facilities, time available for completion of analyses and financial costs incurred in using the different methods.

Recently, a complete mitochondrial genome sequence was generated for a Neanderthal individual (Green *et al.*, 2008). It was determined, based on the mitochondrial variation present in this sequence that the Neanderthal mtDNA fell outside of the variation of the mtDNA of extant humans. This result was further supported by the construction of a phylogenetic tree in which 54 extant humans formed a monophyletic group to the exclusion of the Neanderthal. It is therefore proposed that the NCBI RefSeq based on the Neanderthal sequence AM948965 submitted by Green *et al.* in 2008 (NC_011137), is used as outgroup in future phylogenetic analyses instead of the chimpanzee sequence (NC_001643) used in the current investigation.

6.4.3 Standard nomenclature for mitochondrial haplogroups

At present, human mitochondrial genome sequences are classified into haplogroups according to a classification scheme based on a combination of control region and coding region sequence data (refer to Section 3.3.1.1). In the current investigation, a classification scheme based on only coding region data was employed for this purpose (refer to Section 4.9). From the results presented in the current investigation, it was evident that the use of different classification schemes can lead to confusion regarding the classification of sequences, and that both schemes are lacking in some way, i.e. the alternative classification scheme includes SNPs located in the control region which are not as stable as SNPs in the coding region, while the scheme used in the current investigation does not

make provision for all the sub-haplogroups identified to date (refer to Sections 5.8 and 5.9).

The proposed revised classification scheme for the L0 haplogroup (refer to Section 5.12) contains differently named as well as a different number of sub-haplogroups and lineages compared to the classification scheme used in the current investigation (refer to Section 4.9) and the alternative classification scheme (refer to Section 3.3.1.1). Additional Khoi-San L0 sequences, as well as L0 sequences from other populations, can be used to confirm the proposed L0 sub-haplogroups and lineages and to further test and refine the revised classification scheme proposed in the current investigation (refer to Section 5.12). The revised classification scheme, based on the mitochondrial coding region, should also be extended to address all the possible subgroups currently identified within the L0 haplogroup by incorporating sequence results presented in other recent investigations such as the study performed by Behar *et al.* in 2008. Revised classification schemes based on phylogenetic networks constructed using coding region only sequence data should also be generated for all the other L haplogroups which include the major haplogroups L1, L2, L3, as well as the rare L4, L5 and L6 haplogroups. These revised classification schemes should also incorporate the exclusion criteria currently used to classify L-specific sequences (refer to Appendix D).

The proposed revised classification scheme for the L0 haplogroup, as well as the revised classification schemes for the other L haplogroups, can be used in conjunction with classification schemes for the mitochondrial haplogroups observed in non-African populations to generate a standard nomenclature for mitochondrial haplogroups. To date the naming of mitochondrial haplogroups and sub-haplogroups has progressed in a haphazard way with haplogroups and sub-haplogroups being named as they were identified. A consistent nomenclature, as has been compiled for Y chromosome haplogroups, is urgently needed (Pakendorf and Stoneking, 2005). The new haplogroup designations proposed for L0 sub-haplogroups in the current investigation (refer to Section 5.12) can be used to name L0 sub-haplogroups, or it can be adapted in light of new information as needed.

A standard nomenclature for all mitochondrial haplogroups based on coding region data will provide a standard list of SNPs that can be assayed and used to assign new mitochondrial sequences to specific haplogroups. These informative SNPs can be assayed to assign individuals to haplogroups using RFLP analysis, multiplex SNaPshot

The search criteria used to obtain mitochondrial genome sequences from GenBank[®] were designed to retrieve previously published mitochondrial sequences that were classified within macrohaplogroup L (refer to Section 4.8). As already mentioned, the classification scheme used in the current investigation is different to the scheme generally used in the literature (refer to Section 3.3.1.1 and Appendix F). When using the classification scheme depicted in Figure 4.1, a number of sequences could not be classified completely, i.e. the sequences could not be classified to either the haplogroup or sub-haplogroup level, due to the absence of certain haplogroup or sub-haplogroup defining SNPs. The incompletely classified sequences were indicated by a dollar sign (\$) preceding the putative haplogroup classification (refer to Appendix G), and descriptions of the SNP profiles of the incompletely classified haplogroups are given in Table 4.8. The incompletely classified sequences were included in the datasets used for analyses, with the exception of the sequences listed in Section 4.10.

Table 4.8 Assignment of haplogroups to incompletely classified sequences included in the current investigation

Putative haplogroup assignment	SNP profile	Haplogroup assignment
L ^a	3594A	Unclassified
L0, L1, L2	3594T + 10810A	L0, L1, L2
L0, L1	3594T + 10810C + (7055A + 11914G)	L0, L1
L0\$a	3594T + 10810C + (7055A + 11914G)	L0, L1
L0\$b	3594T + 10810C + (7055A + 11914G)	L0, L1
L0c2\$	3594T + 10810C + (7055A + 11914A) + 9755A + 9818T + (5096T + 5147G)	L0c
L1b2\$	3594T + 10180C + 7055G + 9072G + (6150G + 11899T)	L1b
L2d\$	3594T + 10810T + 11914G + 4767A + (15849C + 13708G)	L2
L3a3\$	3594C + 10400C + 10810G + 14905G + (9554G + 6221T)	L3a
L3b\$	3594C + 10400C + 10819A + (6221T + 5147G)	L3

^a = putative haplogroup assignment is based on information provided in the publication in which this sequence was described (Gonder *et al.*, 2007). \$ = incompletely classified sequences. Sequences were assigned haplogroup designations based on the classification scheme described in Figure 4.1. The SNPs deviating from the expected SNP as presented in the classification scheme are indicated in **bold** text. Brackets indicate a combination of SNPs that have to be present for classification.

The mtDNA sequences generated in the current investigation were also compared to the global mtDNA haplogroup tree (refer to Appendix C) and the exclusion criteria for L-specific sequences (refer to Appendix D). This was done to ensure that the Khoi-San sequences fit the criteria to be grouped within macrohaplogroup L.

assays (Brandstätter *et al.*, 2003), or mtDNA sequencing microarrays which have been successfully tested for clinical purposes (Maitra *et al.*, 2004). The identification of informative SNPs and the use of new technology, in conjunction with a standard classification scheme, should make rapid and accurate haplogroup assignment possible.

6.4.4 Statistical analyses

In the current investigation the TMRCA for the L0 haplogroup and sub-haplogroups were estimated as described in Section 4.14.7. The method used is based on the number of substitutions observed in the branches of the global L0-specific haplogroup network and is a fairly simple method which does not take into account the possible variation in mutation rate at different nucleotide positions. The presence of recurrent alterations in the global L0-specific haplogroup network used to calculate the divergence times can also lead to the underestimation of divergence times as it is unlikely that all recurrent alterations can be reconstructed at the time depths observed in African populations due to the high nucleotide substitution rate which causes saturation of nucleotide sites over time (Salas *et al.*, 2002). After the addition of new L0 sequences to the global L0-specific haplogroup network, it is therefore suggested that the identified recurrent alterations are excluded prior to the re-calculation of TMRCA estimates.

Improved computational and analytical methods have also recently become available for the determination of TMRCA estimates such as the coalescent simulation. Tishkoff *et al.* in 2007 based their TMRCA calculations on HVS1 sequence data. They used Genetree 9.0 (<http://www.stats.ox.ac.uk/~griff/software.html>) and employed Markov chain coalescent simulations, assuming a constant population size and using 10 million step chains to estimate the likelihood surface. This method however has the disadvantage that it requires a large amount of computing power to run the simulations.

Current statistical procedures available to estimate the effects of different processes on mitochondrial variation are lacking in sensitivity, as indicated by the fact that the effect of selection is not always detected by Tajima's D and Fu and Li D^* and F^* tests, as well as the ability to distinguish between the different factors playing a role in shaping variation. New computational and analytical methods are therefore required in order to increase the confidence in determination of ages of mutations and TMRCA. An integral requirement for these new methods will be the capacity to differentiate between the role of selection and demographic effects (Cavalli-Sforza and Feldman, 2003).

In the current investigation, approximate dates for population expansion events and female population sizes in datasets that exhibited evidence of population expansion was not calculated due to the use of a simple population growth model which did not take into account all the relevant parameters (refer to Section 5.11.4.2). It will however be of interest to determine how the sequence data generated for the 30 Khoi-San individuals influence recent estimates of effective population size and population expansion dates calculated in other investigations such as the study by Atkinson *et al.* (2008a). This group used Bayesian coalescent inference with a dataset containing 224 complete human mitochondrial DNA sequences to estimate effective population size through time for each of the four major African mtDNA haplogroups.

6.4.5 The role of natural selection in adaptation

In the current investigation, a number of non-synonymous alterations were observed in highly conserved amino acids in certain branches of the global L0-specific haplogroup network (refer to Section 5.11.5). It will be of interest to investigate the functional significance of these alterations, especially in those observed in internal branches, by determining through protein expression studies if the protein structure and/or function are affected by the observed alteration. If an alteration has given rise to a structural and/or functional change, it will also be of interest to compare the gene/protein in which the alteration was identified between populations in different climatic or geographic regions. This can be done by measuring the activity levels of enzymes composed of altered protein subunits in populations in different climatic/geographic areas (Mishmar *et al.*, 2003). This should provide a possible reason for the observed structural and/or functional change and could provide clues regarding adaptive variation due to environmental influences.

6.4.6 The history and evolution of the Khoi-San population

The current investigation provides information regarding the mitochondrial genetic variation present in the Southern African Khoi-San population. This information was used in both phylogenetic and statistical analyses (refer to Sections 5.9 and 5.11) to draw inferences about the history and evolution of the Khoi-San population. The information provided by mitochondrial DNA however only reflects the maternal history of an individual or population. Due to the influences of genetic drift and/or selection acting on mtDNA, the history of mtDNA on its own may not accurately reflect the history of an individual or population (Pakendorf and Stoneking, 2005). Studies of mtDNA variation need to be

complemented with data regarding the male lineage, as provided by studies on Y chromosome variation, as well as with autosomal data where possible (Bamshad *et al.*, 2003; Basu *et al.*, 2003; Nasidze *et al.*, 2004; Shen *et al.*, 2004; Tambets *et al.*, 2004; Wooding *et al.*, 2004). For this reason, it is necessary to investigate both the Y chromosome and autosomal variation observed for the !Kung individuals used in the current investigation, in order to determine if similar conclusions regarding the history and evolution of the Southern African Khoi-San population can be drawn based on Y chromosome and autosomal data. A recent investigation of complete nuclear genome sequences of four Khoi-San individuals support the model proposed in the current investigation (refer to Figure 6.1 and Section 6.2.3) that in terms of intra-population genetic diversity the Khoi-San are more different from each other than for instance a European and an Asian (Schuster *et al.*, 2010). The use of mtDNA alone in studies of human evolution will probably decrease in the future (Pakendorf and Stoneking, 2005).

In the current investigation, the !Kung ethnic group was investigated as representative of the Southern African Khoi-San population. The Southern African Khoi-San speakers however also consist of other groups, including the Khwe ethnic group. In order to obtain more information regarding the mitochondrial variation present in Khoi-San speaking populations, it will be necessary to determine mitochondrial genome sequences for the Khwe. The sample group previously collected in the north-western Kalahari desert in southern Africa (Chen *et al.*, 2000) from which the !Kung individuals studied in the current investigation were obtained, also contains a number of Khwe individuals. The Khwe appear to share similarities with Bantu populations regarding physical appearance (Chen *et al.*, 2000). It will be of interest to determine mitochondrial DNA genome sequences for this group in order to determine if the sequences are more closely related to sequences observed in Khoi-San populations or Bantu populations, i.e. does this group consist of mainly L0 sequences as observed for the !Kung (Mouton, 2003; current investigation) or mostly L2 and L3 sequences as observed in other Bantu-speaking populations (Chen *et al.*, 2000).

6.4.7 Classification of human social groups

According to Maslow's hierarchy of needs one of the most basic human needs is to belong (Maslow, 1943). For this reason, humans identify with others by classifying people into groups based on perceived similarities. Classification of human social groups can be based on indigenous continental affiliation, skin colour, membership of a group defined on

religious-historical grounds, or current geography and it has been suggested that the most suitable default method of classification is to use geographical information, rather than national, cultural or phenotypic labels (Jobling *et al.*, 2004).

In the current investigation, and in other investigations of Khoi-San individuals (Chen *et al.*, 2000; Gonder *et al.*, 2007; Tishkoff *et al.*, 2007; Behar *et al.*, 2008), classification was based on a combination of language, geography and economy. This has the potential to lead to confusion as individuals that share the same genetic ancestry can be classified into different groups based on aspects such as language and economy. In addition, different derivatives of the same names are used to describe the same ethnic group. The !Kung is also referred to as the !Xun, !Xung, Vasekela, Vasikela or San (Chen *et al.*, 2000; Gonder *et al.*, 2007; Tishkoff *et al.*, 2007; Behar *et al.*, 2008). In order to avoid confusion, it is necessary to develop a standardised system of classification for the Khoi-San whereby the results of different genetic investigations can be compared. Although language is not a good indicator of genetic history, since not all gene flow is accompanied by the movement of languages (Jobling *et al.*, 2004), the use of language in the classification of Khoi-San individuals is firmly entrenched in genetic investigations of both mtDNA and Y chromosome data. It is therefore suggested that the use of language for classification is retained, but that a standardised scheme is constructed that includes the use of geographical information.

It is tempting to suggest the use of genetic information for the basis of the classification of individuals into social groups. The regional distribution of mitochondrial haplogroups is already used to distinguish between African, European, Asian and Native American populations. The results presented in the current investigation, as well as other recent investigations demonstrate that the L0a (L0k¹) and L0b (L0d) mitochondrial lineages are predominantly observed in Southern African Khoi-San populations, and that the presence of these lineages in other populations is the result of recent gene flow (Gonder *et al.*, 2007; Tishkoff *et al.*, 2007; Behar *et al.*, 2008). The localisation of these two lineages to the Khoi-San population in southern Africa could provide the basis for a classification scheme based on genetic information. However, such a scheme will only take into account the maternal genetic heritage of an individual and will provide an incomplete picture of the individual's genetic past as autosomal and Y chromosome data is not included. In addition, any classification scheme based on genetic information will inevitably be very

¹ The haplogroup classifications presented in brackets refer to the classifications according to the alternative classification scheme (refer to Section 3.3.1.1).

controversial, and it is therefore prudent to instead use these lineages as biomarkers for individuals with Khoi-San genetic heritage in the development of prevention and treatment strategies in personalised genetic medicine.

The current investigation provides a view of the mitochondrial genetic variation present in one of the most ancient African populations and investigates one of the most ancient mitochondrial haplogroups, namely the L0 haplogroup. The consensus sequence generated in the current investigation provides a baseline for comparison of African populations, as well as other populations on other continents, in order to distinguish between polymorphisms and potential disease-causing mutations. *Homo sapiens* originated in Africa with a specific mitochondrial genetic background adapted to the unique challenges faced in Africa, challenges which included the climate, diet, parasites, and diseases. This background profile migrated with modern man into new environments to face new challenges. All humans alive today therefore carry variants of the original mitochondrial genome. The different variants were ordered into mitochondrial DNA haplogroups which affect the way in which an individual responds to challenges regarding the environment and disease. The consensus sequence provides a starting point from which to track the way in which the mitochondrial genome evolved and adapted to new environments, diets, parasites and diseases.

In this era of personalised medicine when the sequence of an individual's nuclear genome can be completely determined in a matter of months (Wheeler *et al.*, 2008), the impact of the mitochondrial genome should not be underestimated as it holds the key to the understanding of our maternal ancestry, energy metabolism and our ability to adapt to an ever-changing environment.

CHAPTER SEVEN

References

7.1 GENERAL REFERENCES

- Achilli A., Perego U.A., Bravi C.M., Coble M.D., Kong Q.-P., Woodward S.R., Salas A., Torroni A. and Bandelt H.-J. The phylogeny of the four pan-American mtDNA haplogroups: implications for evolutionary and disease studies. *PLoS ONE*, **3**(3), e1764, 2008.
- Adcock G.J., Dennis E.S., Easteal S., Huttley G.A., Jermini L.S., Peacock W.J. and Thorne A. Mitochondrial DNA sequences in ancient Australians: implications for modern human origins. *Proc. Natl. Acad. Sci. U.S.A.*, **98**(2), 537-542, 2001.
- Agris P.F. Decoding the genome: a modified view. *Nucleic Acids Res.*, **32**(1), 223-238, 2004.
- Akashi H. Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. *Genetics*, **136**(3), 927-935, 1994.
- Albertini A.M., Hofer M., Calos M.P. and Miller J.H. On the formation of spontaneous deletions: the importance of short sequence homologies in the generation of large deletions. *Cell*, **29**(2), 319-328, 1982.
- Alonso S. and Armour J.A. A highly variable segment of human subterminal 16p reveals a history of population growth for modern humans outside Africa. *Proc. Natl. Acad. Sci. U.S.A.*, **98**(3), 864-869, 2001.
- Alphey L. Introduction to biotechniques. DNA sequencing. From experimental methods to bioinformatics. Springer-Verlag, New York, USA, 1997.
- Alves-Silva J., Da Silva Santos M., Guimaraes P.E., Ferreira A.C., Bandelt H.-J., Pena S.D. and Prado V.F. The ancestry of Brazilian mtDNA lineages. *Am. J. Hum. Genet.*, **67**(2), 444-461, 2000.
- Ammerman A.J. and Cavalli-Sforza L.L. Measuring the rate of spread of early farming in Europe. *Man (N.S.)*, **6**(4), 674-688, 1971.
- Ammerman A.J. and Cavalli-Sforza L.L. The Neolithic transition and the genetics of populations in Europe. Princeton University Press, Princeton, New Jersey, USA, 1984.
- Anderson S., Bankier A.T., Barrell B.G., De Bruijn M.H., Coulson A.R., Drouin J., Eperon I.C., Nierlich D.P., Roe B.A., Sanger F., Schreier P.H., Smith A.J., Staden R. and Young I.G. Sequence and organization of the human mitochondrial genome. *Nature*, **290**(5806), 457-465, 1981.
- Andreu A.L., Bruno C., Dunne T.C., Tanji K., Shanske S., Sue C.M., Krishna S., Hadjigeorgiou G.M., Shtilbans A., Bonilla E. and DiMauro S. A nonsense mutation (G15059A) in the cytochrome *b* gene in a patient with exercise intolerance and myoglobinuria. *Ann. Neurol.*, **45**(1), 127-130, 1999a.
- Andreu A.L., Bruno C., Hadjigeorgiou G.M., Shanske S. and DiMauro S. Polymorphic variants in the human mitochondrial cytochrome *b* gene. *Mol. Genet. Metab.*, **67**(1), 49-52, 1999d.
- Andreu A.L., Hanna M.G., Reichmann H., Bruno C., Penn A.S., Tanji K., Pallotti F., Iwata S., Bonilla E., Lach B., Morgan-Hughes J. and DiMauro S. Exercise intolerance due to mutations in the cytochrome *b* gene of mitochondrial DNA. *N. Engl. J. Med.*, **341**(14), 1037-1044, 1999b.
- Andreu A.L., Tanji K., Bruno C., Hadjigeorgiou G.M., Sue C.M., Jay C., Ohnishi T., Shanske S., Bonilla E. and DiMauro S. Exercise intolerance due to a nonsense mutation in the mtDNA ND4 gene. *Ann. Neurol.*, **45**(6), 820-823, 1999c.
- Andrews R.M., Kubacka I., Chinnery P.F., Lightowlers R.N., Turnbull D.M. and Howell N. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat. Genet.*, **23**, 147, 1999.
- Applied Biosystems®. Applied Biosystems® 3730/3730xl DNA Analyzers Sequencing Chemistry Guide, USA, 2002.
- Applied Biosystems®. Automated DNA Sequencing Chemistry Guide, USA, 2000.
- Aquadro C.F. and Greenberg B.D. Human mitochondrial DNA variation and evolution: analysis of nucleotide sequences from seven individuals. *Genetics*, **103**(2), 287-312, 1983.
- Arctander P. Mitochondrial recombination? *Science*, **284**(5423), 2090-2091, 1999.
- Armour J.A., Anttinen T., May C.A., Vega E.E., Sajantila A., Kidd J.R., Kidd K.K., Bertranpetit J., Pääbo S. and Jeffreys A.J. Minisatellite diversity supports a recent African origin for modern humans. *Nat. Genet.*, **13**(2), 154-160, 1996.

- Atkinson Q.D., Gray R.D. and Drummond A.J. Bayesian coalescent inference of major human mitochondrial DNA haplogroup expansions in Africa. *Proc Biol Sci.*, **276**(1655), 367-373, 2009.
- Atkinson Q.D., Gray R.D. and Drummond A.J. mtDNA variation predicts population size in humans and reveals a major Southern Asian chapter in human prehistory. *Mol. Biol. Evol.*, **25**(2), 468-474, 2008.
- Attardi G. and Ojala D. Mitochondrial ribosomes in HeLa cells. *Nature New Biol.*, **229**, 133-136, 1971.
- Attardi G. Animal mitochondrial DNA: an extreme example of genetic economy. *Int. Rev. Cytol.*, **93**, 93-145, 1985.
- Attardi G. The elucidation of the human mitochondrial genome: a historical perspective. *BioEssays*, **5**, 34-39, 1986.
- Attardi G. The human mitochondrial genetic system. In: Mitochondrial DNA in human pathology. DiMauro S. and Wallace D.C. (editors). Raven Press Ltd., New York, USA, 9-25, 1993.
- Avise J.C., Arnold J., Ball R.M., Bermingham E., Lamb T., Neigel J.E., Reeb C.A. and Saunders N.C. Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. *Annu Rev Ecol Syst*, **18**, 489-522, 1987.
- Awadalla P., Eyre-Walker A. and Maynard Smith J. Linkage disequilibrium and recombination in hominid mitochondrial DNA. *Science*, **286**, 2524-2525, 1999.
- Awadalla P., Eyre-Walker A. and Maynard Smith J. Questioning evidence for recombination in human mitochondrial DNA. *Science*, **288**(5473), 1931a, 2000.
- Ballard J.W. and Whitlock M.C. The incomplete natural history of mitochondria. *Mol. Ecol.*, **13**(4), 729-744, 2004.
- Ballinger S.W., Schurr T.G., Torroni A., Gan Y.Y., Hodge J.A., Hassan K., Chen K.-H. and Wallace D.C. Southeast Asian mitochondrial DNA analysis reveals genetic continuity of ancient mongoloid migrations. *Genetics*, **130**, 139-152, 1992.
- Bamshad M.J., Watkins W.S., Dixon M.E., Jorde L.B., Rao B.B., Naidu J.M., Prasad B.V., Rasanayagam A. and Hammer M.F. Female gene flow stratifies Hindu castes. *Nature*, **395**(6703), 651-652, 1998.
- Bamshad M.J., Wooding S., Watkins W.S., Ostler C.T., Batzer M.A. and Jorde L.B. Human population genetic structure and inference of group membership. *Am. J. Hum. Genet.*, **72**(3), 578-589, 2003.
- Bandelt H.-J. and Forster P. The myth of bumpy hunter-gatherer mismatch distributions. *Am. J. Hum. Genet.*, **61**(4), 980-983, 1997.
- Bandelt H.-J., Alves-Silva J., Guimarães P.E.M., Santos M.S., Brehm A., Pereira L., Coppa A., Larruga J.M., Rengo C., Scozzari R., Torroni A., Prata M.J., Amorim A., Prado V.F. and Pena S.D.J. Phylogeography of the human mitochondrial haplogroup L3e: a snapshot of African prehistory and Atlantic slave trade. *Ann. Hum. Genet.*, **65**(6), 549-563, 2001a.
- Bandelt H.-J., Forster P., Sykes B.C. and Richards M.B. Mitochondrial portraits of human populations using median networks. *Genetics*, **141**(2), 743-753, 1995.
- Bandelt H.-J., Herrnstadt C., Yao Y.-G., Kong Q.-P., Kivisild T., Rengo C., Scozzari R., Richards M., Villems R., Macaulay V., Howell N., Torroni A. and Zhang Y.-P. Identification of Native American founder mtDNAs through the analysis of complete mtDNA sequences: some caveats. *Ann. Hum. Genet.*, **67**(6), 512-524, 2003.
- Bandelt H.-J., Lahermo P., Richards M. and Macaulay V. Detecting errors in mtDNA data by phylogenetic analysis. *Int. J. Legal Med.*, **115**(2), 64-69, 2001b.
- Bandelt H.-J., Quintana-Murci L., Salas A. and Macaulay V. The fingerprint of phantom mutations in mitochondrial DNA data. *Am. J. Hum. Genet.*, **71**(5), 1150-1160, 2002.
- Barbosa A.B., Da Silva L.A., Azevedo D.A., Balbino V.Q. and Mauricio-da-Silva L. Mitochondrial DNA control region polymorphism in the population of Alagoas state, north-eastern Brazil. *J. Forensic Sci.*, **53**(1), 142-146, 2008.
- Barker K. At the bench. A laboratory navigator. Updated edition. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, USA, 2005.
- Barkhan D. and Soodyall H. Mitochondrial DNA (mtDNA) and Y chromosome DNA variation in African populations. In: The prehistory of Africa. Tracing the lineage of modern man. Soodyall H. (editor). Jonathan Ball Publishers, Johannesburg, South Africa, 109-124, 2006.
- Barnard A. Hunters and herders of southern Africa: a comparative ethnography of the Khoisan peoples. Cambridge University Press, Cambridge, New York, 1992.
- Barnes W.M. The fidelity of *Taq* polymerase catalyzing PCR is improved by an N-terminal deletion. *Gene*, **112**(1), 29-35, 1992.
- Barrell B.G., Anderson S., Bankier A.T., De Bruijn M.H.L., Chen E., Coulson T.A.R., Drouin J., Eperon F.I.C., Nierlich D.P., Roe B.A., Sanger F.T., Schreier P.H., Smith A.J.H., Staden R. and Young I.G. Different pattern of codon recognition by mammalian mitochondrial tRNAs. *Proc. Natl. Acad. Sci. U.S.A.*, **77**(6), 3164-3166, 1980.
- Basu A. and Majumder P.P. A comparison of two popular statistical methods for estimating the time to most recent common ancestor (TMRCA) from a sample of DNA sequences. *J. Genet.*, **82**(1-2), 7-12, 2003.

- Basu A., Mukherjee N., Roy S., Sengupta S., Banerjee S., Chakraborty M., Dey B., Roy M., Roy B., Bhattacharyya N.P., Roychoudhury S. and Majumder P.P. Ethnic India: a genomic view, with special reference to peopling and structure. *Genome Res.*, **13**(10), 2277-2290, 2003.
- Batini C., Coia V., Battaglia C., Rocha J., Pilkington M.M., Spedini G., Comas D., Destro-Bisol G. and Calafell F. Phylogeography of the human mitochondrial L1c haplogroup: genetic signatures of the prehistory of Central Africa. *Mol. Phylogenet. Evol.*, **43**(2), 635-644, 2007.
- Battersby B.J., Loredó-Osti J.C. and Shoubridge E.A. Nuclear genetic control of mitochondrial DNA segregation. *Nat. Genet.*, **33**(2), 183-186, 2003.
- Baudouin S.V., Saunders D., Tiangyou W., Elson J.L., Poynter J., Pyle A., Keers S., Turnbull D.M., Howell N. and Chinnery P.F. Mitochondrial DNA and survival after sepsis: a prospective study. *Lancet*, **366**(9503), 2118-2121, 2005.
- Behar D.M., Villems R., Soodyall H., Blue-Smith J., Pereira L., Metspalu E., Scozzari R., Makkan H., Tzur S., Comas D., Bertranpetit J., Quintana-Murci L., Tyler-Smith C., Wells R.S., Rosset S. and The Genographic Consortium. The dawn of human matrilineal diversity. *Am. J. Hum. Genet.*, **82**(5), 1130-1140, 2008.
- Bendall K.E., Macaulay V.A., Baker J.R. and Sykes B.C. Heteroplasmic point mutations in the human mtDNA control region. *Am. J. Hum. Genet.*, **59**, 1276-1287, 1996.
- Bensasson D., Zhang D.-X., Hartl D.L. and Hewitt G.M. Mitochondrial pseudogenes: evolution's misplaced witnesses. *Trends Ecol. Evol. (Amst.)*, **16**(60), 314-321, 2001.
- Benson D.A., Karsch-Mizrachi I., Lipman D.J., Ostell J. and Wheeler D.L. GenBank. *Nucleic Acids Res.*, **36**(Database issue), D25-30, 2008.
- Birky C.W. Jr. The inheritance of genes in mitochondria and chloroplasts: laws, mechanisms, and models. *Annu. Rev. Genet.*, **35**, 125-148, 2001.
- Birky C.W. Jr. Uniparental inheritance of mitochondrial and chloroplast genes: mechanisms and evolution. *Proc. Natl. Acad. Sci. U.S.A.*, **92**, 11331-11338, 1995.
- Bolden A., Noy G.P. and Weissbach A. DNA polymerase of mitochondria is a γ -polymerase. *J. Biol. Chem.*, **252**(10), 3351-3356, 1977.
- Bonafè M., Barbi C., Olivieri F., Yashin A., Andreev K.F., Vaupel J.W., De Benedictis G., Rose G., Carrieri G., Jazwinski S.M. and Franceschi C. An allele of *HRAS1* 3' variable number of tandem repeats is a frailty allele: implication for an evolutionarily-conserved pathway involved in longevity. *Gene*, **286**(1), 121-126, 2002.
- Bonato S.L. and Salzano F.M. A single and early migration for the peopling of the Americas supported by mitochondrial DNA sequence data. *Proc. Natl. Acad. Sci. U.S.A.*, **94**(5), 1866-1871, 1997.
- Borst P. Structure and function of mitochondrial DNA. *Trends Biochem. Sci.*, **2**(2), 31-34, 1977.
- Bortolini M.C., Da Silva W.A. Jr., Zago M.A., Elion J., Krishnamoorthy R., Gonçalves V.F. and Pena S.D. The phylogeography of mitochondrial DNA haplogroup L3g in Africa and the Atlantic slave trade. *Am. J. Hum. Genet.*, **75**(3), 522-524, 2004.
- Brandstätter A., Parsons T.J. and Parson W. MtDNA coding region SNPs for rapid screening and haplogroups identification of forensic samples. *Int. Congr. Ser.*, **1261**, 422-424, 2004.
- Brandstätter A., Parsons T.J. and Parson W. Rapid screening of mtDNA coding region SNPs for the identification of west European Caucasian haplogroups. *Int. J. Legal Med.*, **117**(5), 291-298, 2003.
- Braverman J.M., Hudson R.R., Kaplan N.L., Langley C.H. and Stephan W. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics*, **140**, 783-796, 1995.
- Brehm A., Pereira L., Bandelt H.-J., Prata M.J. and Amorim A. Mitochondrial portrait of the Cabo Verde archipelago: the Senegambian outpost of Atlantic slave trade. *Ann. Hum. Genet.*, **66**(1), 49-60, 2002.
- Brown M.D., Hosseini S.H., Torroni A., Bandelt H.-J., Allen J.C., Schurr T.G., Scozzari R., Cruciani F. and Wallace D.C. MtDNA haplogroup X: an ancient link between Europe/Western Asia and North America? *Am. J. Hum. Genet.*, **63**(6), 1852-1861, 1998.
- Brown M.D., Starikovskaya E., Derbeneva O., Hosseini S., Allen J.C., Mikhailovskaya I.E., Sukernik R.I. and Wallace D.C. The role of mtDNA background in disease expression: a new primary LHON mutation associated with Western Eurasian haplogroup J. *Hum. Genet.*, **110**(2), 130-138, 2002.
- Brown M.D., Sun F. and Wallace D.C. Clustering of Caucasian Leber hereditary optic neuropathy patients containing the 11778 or 14484 mutations on an mtDNA lineage. *Am. J. Hum. Genet.*, **60**(2), 381-387, 1997.
- Brown M.D., Torroni A., Huoponen K., Chen Y.-S., Lott M.T. and Wallace D.C. Pathological significance of the mtDNA COX III mutation at nucleotide pair 9438 in Leber hereditary optic neuropathy. *Am. J. Hum. Genet.*, **55**(2), 410-412, 1994.
- Brown M.D., Torroni A., Shoffner J.M. and Wallace D.C. Mitochondrial tRNA^{Thr} mutations and lethal infantile mitochondrial myopathy. *Am. J. Hum. Genet.*, **51**(2), 446-447, 1992.
- Brown M.D., Zhadanov S., Allen J.C., Hosseini S., Newman N.J., Atamonov V.V., Mikhailovskaya I.E., Sukernik R.I. and Wallace D.C. Novel mtDNA mutations and oxidative phosphorylation dysfunction in Russian LHON families. *Hum. Genet.*, **109**(1), 33-39, 2001.

- Brown W.M. Polymorphism in mitochondrial DNA of humans as revealed by restriction endonuclease analysis. *Proc. Natl. Acad. Sci. U.S.A.*, **77**(6), 3605-3609, 1980.
- Brown W.M., George M. Jr. and Wilson A.C. Rapid evolution of animal mitochondrial DNA. *Proc. Natl. Acad. Sci. U.S.A.*, **76**(4), 1967-1971, 1979.
- Brown W.M., Prager E.M., Wang A. and Wilson A.C. Mitochondrial DNA sequences of primates: tempo and mode of evolution. *J. Mol. Evol.*, **18**(4), 225-239, 1982.
- Brownie J., Shawcross S., Theaker J., Whitcombe D., Ferrie R., Newton C. and Little S. The elimination of primer-dimer accumulation in PCR. *Nucleic Acids Res.*, **25**(16), 3235-3241, 1997.
- Budowle B., Wilson M.R., DiZinno J.A., Stauffer C., Fasano M.A., Holland M.M. and Monson K.L. Mitochondrial DNA regions HVI and HVII population data. *Forensic Sci. Int.*, **103**(1), 23-35, 1999.
- Campbell M.K. Biochemistry. Second edition. Saunders College Publishing, Philadelphia, USA, 1995.
- Cann A.J. Maths from scratch for biologists. John Wiley and Sons Ltd., West Sussex, England, 2003.
- Cann R.L. and Wilson A.C. Length mutations in human mitochondrial DNA. *Genetics*, **104**(4), 699-711, 1983.
- Cann R.L., Brown W.M. and Wilson A.C. Polymorphic sites and the mechanism of evolution in human mitochondrial DNA. *Genetics*, **106**(3), 479-499, 1984.
- Cann R.L., Stoneking M. and Wilson A.C. Mitochondrial DNA and human evolution. *Nature*, **325**, 31-36, 1987.
- Cantatore P. and Attardi G. Mapping of nascent light and heavy strand transcripts on the physical map of HeLa cell mitochondrial DNA. *Nucleic Acids Res.*, **8**(12), 2605-2625, 1980.
- Cantor C.R. DNA choreography. *Cell*, **25**(2), 293-295, 1981.
- Carrieri G., Bonafè M., De Luca M., Rose G., Varcasia O., Bruni A., Maletta R., Nacmias B., Sorbi S., Corsonello F., Feraco E., Andreev K.F., Yashin A.I., Franceschi C. and De Benedictis G. Mitochondrial DNA haplogroups and APOE4 allele are non-independent variables in sporadic Alzheimer's disease. *Hum. Genet.*, **108**(3), 194-198, 2001.
- Carroll J., Shannon R.J., Fearnley I.M., Walker J.E. and Hirst J. Definition of the nuclear encoded protein composition of bovine heart mitochondrial complex I. Identification of two new subunits. *J. Biol. Chem.*, **277**(52), 50311-50317, 2002.
- Carter R.W. Mitochondrial diversity within modern human populations. *Nucleic Acids Res.*, **35**(9), 3039-3045, 2007.
- Carto S.L., Weaver A.J., Hetherington R., Lam Y. and Wiebe E.C. Out of Africa and into an ice age: on the role of global climate change in the late Pleistocene migration of early modern humans out of Africa. *J. Hum. Evol.*, **56**(2), 139-151, 2009.
- Castora F.J., Lazarus G.M. and Kunes D. The presence of two mitochondrial DNA topoisomerases in human acute leukemia cells. *Biochem. Biophys. Res. Commun.*, **130**(2), 854-866, 1985.
- Castrì L., Tofanelli S., Garagnani P., Bini C., Fosella X., Pelotti S., Paoli G., Pettener D. and Luiselli D. mtDNA variability in two Bantu-speaking populations (Shona and Hutu) from Eastern Africa: implications for peopling and migration patterns in sub-Saharan Africa. *Am. J. Phys. Anthropol.*, **140**(2), 302-311, 2009.
- Cavalli-Sforza L.L. and Feldman M.W. The application of molecular genetic approaches to the study of human evolution. *Nat. Genet.*, **33**(Suppl), 266-275, 2003.
- Cavalli-Sforza L.L. Genes, peoples and languages. Penguin Books Ltd., London, England, 2001.
- Cavalli-Sforza L.L., Piazza A. and Menozzi P. History and geography of human genes. Princeton University Press, Princeton, New Jersey, USA, 1994.
- Cerezo M., Černý V., Carracedo A. and Salas A. Applications of MALDI-TOF MS to large-scale human mtDNA population-based studies. *Electrophoresis*, **30**(21), 3665-3673, 2009.
- Černý V., Fernandes V., Costa M.D., Hájek M., Mulligan C.J. and Pereira L. Migration of Chad speaking pastoralists within Africa based on population structure of Chad Basin and phylogeography of mitochondrial L3f haplogroup. *BMC Evol. Biol.*, **9**(1), 63, 2009.
- Černý V., Salas A., Hájek M., Zaloudková M. and Brdicka R. A bidirectional corridor in the Sahel-Sudan belt and the distinctive features of the Chad Basin populations: a history revealed by the mitochondrial DNA genome. *Ann. Hum. Genet.*, **71**(4), 433-452, 2007.
- Chagnon P., Gee M., Filion M., Robitaille Y., Belouchi M. and Gauvreau D. Phylogenetic analysis of the mitochondrial genome indicates significant differences between patients with Alzheimer disease and controls in a French-Canadian founder population. *Am. J. Med. Genet.*, **85**(1), 20-30, 1999.
- Chang D.D. and Clayton D.A. Precise identification of individual promoters for transcription of each strand of human mitochondrial DNA. *Cell*, **36**(3), 635-643, 1984.
- Charlesworth B. The effect of background selection against deleterious mutations on weakly selected, linked variants. *Genet. Res.*, **63**(3), 213-227, 1994.
- Charlesworth B., Morgan M.T. and Charlesworth D. The effect of deleterious mutations on neutral molecular variation. *Genetics*, **134**(4), 1289-1303, 1993.
- Charlesworth D., Charlesworth B. and Morgan M.T. The pattern of neutral molecular variation under the background selection model. *Genetics*, **141**(4), 1619-1632, 1995.

- Chen Y.-S., Olckers A., Schurr T.G., Kogelnik A.M., Huoponen K. and Wallace D.C. mtDNA variation in the South African Kung and Khwe – and their genetic relationships to other African populations. *Am. J. Hum. Genet.*, **66**, 1362-1383, 2000.
- Chen Y.-S., Torroni A., Excoffier L., Santachiara-Benerecetti A.S. and Wallace D.C. Analysis of mtDNA variation in African populations reveals the most ancient of all human continent-specific haplogroups. *Am. J. Hum. Genet.*, **57**(1), 133-149, 1995.
- Chinnery P.F. and Schon E.A. Mitochondria. *J. Neurol. Neurosurg. Psychiatr.*, **74**(9), 1188-1199, 2003.
- Chinnery P.F., Thorburn D.R., Samuels D.C., White S.L., Dahl H.M., Turnbull D.M., Lightowlers R.N. and Howell N. The inheritance of mitochondrial DNA heteroplasmy: random drift, selection or both? *Trends Genet.*, **16**(11), 500-505, 2000.
- Chou Q., Russell M., Birch D.E., Raymond J. and Bloch W. Prevention of pre-PCR mis-priming and primer dimerization improves low-copy-number amplifications. *Nucleic Acids Res.*, **20**(7), 1717-1723, 1992.
- Christianson T.W. and Clayton D.A. *In vitro* transcription of human mitochondrial DNA: accurate termination requires a region of DNA sequence that can function bidirectionally. *Proc. Natl. Acad. Sci. U.S.A.*, **83**(17), 6277-6281, 1986.
- Clark J.D., Beyene Y., WoldeGabriel G., Hart W.K., Renne P.R., Gilbert H., Defleur A., Suwa G., Katoh S., Ludwig K.R., Boissarie J.R., Asfaw B. and White T.D. Stratigraphic, chronological and behavioural contexts of Pleistocene *Homo sapiens* from Middle Awash, Ethiopia. *Nature*, **423**(6941), 747-752, 2003.
- Clark K.M., Taylor R.W., Johnson M.A., Chinnery P.F., Chrzanowska-Lightowlers Z.M., Andrews R.M., Nelson I.P., Wood N.W., Lamont P.J., Hanna M.G., Lightowlers R.N. and Turnbull D.M. An mtDNA mutation in the initiation codon of the cytochrome C oxidase subunit II gene results in lower levels of the protein and a mitochondrial encephalomyopathy. *Am. J. Hum. Genet.*, **64**(5), 1330-1339, 1999.
- Clayton D.A. Replication and transcription of vertebrate mitochondrial DNA. *Annu. Rev. Cell Biol.*, **7**, 453-478, 1991.
- Clayton D.A. Replication of animal mitochondrial DNA. *Cell*, **28**, 693-705, 1982.
- Comas D., Calafell F., Mateu E., Pérez-Lezaun A., Bosch E., Martínez-Arias R., Clarimon J., Facchini F., Fiori G., Luiselli D., Pettener D. and Bertranpetit J. Trading genes along the silk road: mtDNA sequences and the origin of central Asian populations. *Am. J. Hum. Genet.*, **63**(6), 1824-1838, 1998.
- Cameron J.M. and Aguadé M. An evaluation of measures of synonymous codon usage bias. *J. Mol. Evol.*, **47**(3), 268-274, 1998.
- Corona P., Antozzi C., Carrara F., D'Incerti L., Lamantea E., Tiranti V. and Zeviani M. A novel mtDNA mutation in the ND5 subunit of complex I in two MELAS patients. *Ann. Neurol.*, **49**(1), 106-110, 2001.
- Coskun P.E., Ruiz-Pesini E. and Wallace D.C. Control region mtDNA variants: longevity, climatic adaptation, and a forensic conundrum. *Proc. Natl. Acad. Sci. U.S.A.*, **100**(5), 2174-2176, 2003.
- Côté J. and Ruiz-Carrillo A. Primers for mitochondrial DNA replication generated by endonuclease G. *Science*, **261**(5122), 765-769, 1993.
- Crawhall N. Languages, genetics and archaeology: problems and the possibilities in Africa. In: The prehistory of Africa. Tracing the lineage of modern man. Soodyall H. (editor). Jonathan Ball Publishers, Johannesburg, South Africa, 109-124, 2006.
- Crespillo M., Paredes M., Luque J.A., Luque G. and Valverde J.L. Population variation of human mitochondrial DNA (HV1 and HV2) in Spanish unrelated individuals (Northeast Spain). *Int. Congr. Ser.*, **1261**, 407-409, 2004.
- Crimi M., Sciacco M., Galbiati S., Bordoni A., Malferrari G., Del Bo R., Biunno I., Bresolin N. and Comi G.P. A collection of 33 novel human mtDNA homoplasmic variants. *Hum. Mutat.*, **20**(5), 409, 2002.
- Crispim D., Canani L.H., Gross J.L., Tschiedel B., Souto K.E. and Roisenberg I. The European-specific mitochondrial cluster J/T could confer an increased risk of insulin-resistance and type 2 diabetes: an analysis of the m.4216T > C and m.4917A > G variants. *Ann. Hum. Genet.*, **70**(4), 488-495, 2006.
- Cummings M.P., Otto S.P. and Wakeley J. Sampling properties of DNA sequence data in phylogenetic analysis. *Mol. Biol. Evol.*, **12**(5), 814-822, 1995.
- Curtin P.D. The Atlantic slave trade. A census. University of Wisconsin Press, Madison, Wisconsin, USA, 1969.
- Daga A., Micol V., Hess D., Aebersold R. and Attardi G. Molecular characterization of the transcription termination factor from human mitochondria. *J. Biol. Chem.*, **268**(11), 8123-8130, 1993.
- Dairaghi D.J., Shadel G.S. and Clayton D.A. Addition of a 29 residue carboxyl-terminal tail converts a simple HMG box-containing protein into a transcriptional activator. *J. Mol. Biol.*, **249**(1), 11-28, 1995.

- De Benedictis G., Rose G., Carrieri G., De Luca M., Falcone E., Passarino G., Bonafè M., Monti D., Baggio G., Bertolini S., Mari D., Mattace R. and Franceschi C. Mitochondrial DNA inherited variants are associated with successful aging and longevity in humans. *FASEB J.*, **13**(12), 1532-1536, 1999.
- Dediu D. Causal correlations between genes and linguistic features – the mechanism of gradual language evolution. In: The evolution of language: proceedings of the 7th international conference (EVOLANG7). Smith A.D.M., Smith K. and Ferrer-i-Cancho R. (editors). World Scientific Publishing, Singapore, 83-90, 2008.
- Denaro M., Blanc H., Johnson M.J., Chen K.H., Wilmsen E., Cavalli-Sforza L.L. and Wallace D.C. Ethnic variation in *Hpa* I endonuclease cleavage patterns of human mitochondrial DNA. *Proc. Natl. Acad. Sci. U.S.A.*, **78**(9), 5768-5772, 1981.
- Denslow N.D., Michaels G.S., Montoya J., Attardi G. and O'Brien T.W. Mechanism of mRNA binding to bovine mitochondrial ribosomes. *J. Biol. Chem.*, **264**(14), 8328-8338, 1989.
- Derbeneva O.A., Sukernik R.I., Volodko N.V., Hosseini S.H., Lott M.T. and Wallace D.C. Analysis of mitochondrial DNA diversity in the Aleuts of the Commander islands and its implications for the genetic history of Beringia. *Am. J. Hum. Genet.*, **71**(2), 415-421, 2002.
- Derenko M.V., Grzybowski T., Malyarchuk B.A., Czarny J., Miścicka-Sliwka D. and Zakharov I.A. The presence of mitochondrial haplogroup X in Altaians from South Siberia. *Am. J. Hum. Genet.*, **69**, 237-241, 2001.
- Destro-Bisol G., Coia V., Boschi I., Verginelli F., Caglià A., Pascali V., Spedini G. and Calafell F. The analysis of variation of mtDNA hypervariable region 1 suggests that Eastern and Western Pygmies diverged before the Bantu expansion. *Am. Nat.*, **163**(2), 212-226, 2004.
- Di Rienzo A. and Wilson A.C. Branching pattern in the evolutionary tree for human mitochondrial DNA. *Proc. Natl. Acad. Sci. U.S.A.*, **88**(5), 1597-1601, 1991.
- DiMauro S. and Davidzon G. Mitochondrial DNA and disease. *Ann. Med.*, **37**(3), 222-232, 2005.
- Don R.H., Cox P.T., Wainwright B.J., Baker K. and Mattick J.S. 'Touchdown' PCR to circumvent spurious priming during gene amplification. *Nucleic Acids Res.*, **19**(14), 4008, 1991.
- Dornelles C.L., Bonatto S.L., Freitas L.B. and Salzano F.M. Is haplogroup X present in extant South American Indians? *Am. J. Phys. Anthropol.*, **127**(4), 439-448, 2005.
- Drake J.W. and Baltz R.H. The biochemistry of mutagenesis. *Annu. Rev. Biochem.*, **45**, 11-37, 1976.
- Dumoulin R., Sagnol I., Ferlin T., Bozon D., Stepien G. and Mousson B. A novel gly290asp mitochondrial cytochrome *b* mutation linked to a complex III deficiency in progressive exercise intolerance. *Mol. Cell. Probes*, **10**(5), 389-391, 1996.
- Dupanloup I., Pereira L., Bertorelle G., Calafell F., Prata M.J., Amorim A. and Barbujani G. A recent shift from polygyny to monogamy in humans is suggested by the analysis of worldwide Y-chromosome diversity. *J. Mol. Evol.*, **57**, 85-97, 2003.
- Efron B. Bootstrap methods: another look at the jackknife. *Ann. Statist.*, **7**(1), 1-26, 1979.
- Elson J.L., Turnbull D.M. and Howell N. Comparative genomics and the evolution of human mitochondrial DNA: assessing the effects of selection. *Am. J. Hum. Genet.*, **74**(2), 229-238, 2004.
- Ennafaa H., Cabrera V.M., Abu-Amero K.K., González A.M., Amor M.B., Bouhaha R., Dzimiri N., Elgaaïed A.B. and Larruga J.M. Mitochondrial DNA haplogroup H structure in North Africa. *BMC Genet.*, **10**, 8, 2009.
- Ewens W.J. The sampling theory of selectively neutral alleles. *Theor Popul Biol*, **3**, 87-112, 1972.
- Excoffier L. and Langaney A. Origin and differentiation of human mitochondrial DNA. *Am. J. Hum. Genet.*, **44**, 73-85, 1989.
- Excoffier L. and Schneider S. Why hunter-gatherer populations do not show signs of pleistocene demographic expansions. *Proc. Natl. Acad. Sci. U.S.A.*, **96**(19), 10597-10602, 1999.
- Excoffier L. and Yang Z. Substitution rate variation among sites in mitochondrial hypervariable region I of humans and chimpanzees. *Mol. Biol. Evol.*, **16**(10), 1357-1368, 1999.
- Excoffier L. Evolution of human mitochondrial DNA: evidence for departure from a pure neutral model of populations at equilibrium. *J. Mol. Evol.*, **30**, 125-139, 1990.
- Excoffier L. Human demographic history: refining the recent African origin model. *Curr. Opin. Genet. Dev.*, **12**, 675-682, 2002.
- Excoffier L. Patterns of DNA sequence diversity and genetic structure after a range expansion: lessons from the infinite-island model. *Mol. Ecol.*, **13**(4), 853-864, 2004.
- Eyre-Walker A. Do mitochondria recombine in humans? *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, **355**, 1573-1580, 2000.
- Eyre-Walker A., Smith N.H. and Maynard Smith J. How clonal are human mitochondria? *Proc Biol Sci.*, **266**(1418), 477-483, 1999.
- Fage J.D. Slavery and the slave trade in the context of West African history. *J Afr Hist*, **10**(3), 393-404, 1969.
- Fagundes N., Ray N., Beaumont M., Neuenschwander S., Salzano F.M., Bonatto S.L. and Excoffier L. Statistical evaluation of alternative models of human evolution. *Proc. Natl. Acad. Sci. U.S.A.*, **104**(45), 17614-17619, 2007.

- Fagundes N.J., Kanitz R., Eckert R., Valls A.C., Bogo M.R., Salzano F.M., Smith D.G., Silva W.A. Jr., Zago M.A., Ribeiro-dos-Santos A.K., Santos S.E., Petzl-Erler M.L. and Bonatto S.L. Mitochondrial population genomics supports a single pre-Clovis origin with a coastal route for the peopling of the Americas. *Am. J. Hum. Genet.*, **82**(3), 583-592, 2008.
- Falconer D.S. and Mackay T.F.C. Introduction to quantitative genetics. Fourth edition. Pearson Education Ltd., Essex, England, 1996.
- Feder J., Ovadia O., Blech I., Cohen J., Wainstein J., Harman-Boehm I., Glaser B. and Mishmar D. Parental diabetes status reveals association of mitochondrial DNA haplogroup J1 with type 2 diabetes. *BMC Med. Genet.*, **10**, 60, 2009.
- Felsenstein J. and Churchill G.A. A hidden Markov model approach to variation among sites in rate of evolution. *Mol. Biol. Evol.*, **13**, 93-104, 1996.
- Felsenstein J. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, **39**(4), 783-791, 1985.
- Felsenstein J. Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. *Genet. Res.*, **59**(2), 139-147, 1992.
- Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, **17**(6), 368-376, 1981.
- Felsenstein J. PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics*, **5**, 164-166, 1989.
- Felsenstein J. PHYLIP (Phylogeny Inference Package) version 3.66. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle, USA, 2006.
- Felsenstein J. Phylogenies from molecular sequences: inference and reliability. *Annu. Rev. Genet.*, **22**, 521-565, 1988.
- Fernandez-Silva P., Martinez-Azorin F., Micol V. and Attardi G. The human mitochondrial transcription termination factor (mTERF) is a multizipper protein but binds to DNA as a monomer, with evidence pointing to intramolecular leucine zipper interactions. *EMBO J.*, **16**(5), 1066-1079, 1997.
- Finnilä S., Hassinen I.E. and Majamaa K. Phylogenetic analysis of mitochondrial DNA in patients with an occipital stroke. Evaluation of mutations by using sequence data on the entire coding region. *Mutat. Res.*, **458**(1-2), 31-39, 2001b.
- Finnilä S., Lehtonen M.S. and Majamaa K. Phylogenetic network for European mtDNA. *Am. J. Hum. Genet.*, **68**(6), 1475-1484, 2001a.
- Fisher R.A. On the interpretation of χ^2 from contingency tables, and the calculation of P. *J R Stat Soc*, **85**(1), 87-94, 1922.
- Fisher R.P. and Clayton D.A. A transcription factor required for promoter recognition by human mitochondrial RNA polymerase. Accurate initiation at the heavy- and light-strand promoters dissected and reconstituted *in vitro*. *J. Biol. Chem.*, **260**(20), 11330-11338, 1985.
- Fisher R.P. and Clayton D.A. Purification and characterization of human mitochondrial transcription factor 1. *Mol. Cell. Biol.*, **8**(8), 3496-3509, 1988.
- Fisher R.P., Lisowsky T., Breen G.A. and Clayton D.A. A rapid, efficient method for purifying DNA-binding proteins. Denaturation-renaturation chromatography of human and yeast mitochondrial extracts. *J. Biol. Chem.*, **266**(14), 9153-9160, 1991.
- Fisher R.P., Topper J.N. and Clayton D.A. Promoter selection in human mitochondria involves binding of a transcription factor to orientation-independent upstream regulatory elements. *Cell*, **50**(2), 247-258, 1987.
- Fitch W.M. Toward defining the course of evolution: minimum change for a specific tree topology. *Syst. Zool.*, **20**, 406-416, 1971.
- Foley R. The context of human genetic evolution. *Genome Res.*, **8**(4), 339-347, 1998.
- Forster P. and Matsumura S. Evolution. Did early humans go north or south? *Science*, **308**(5724), 965-966, 2005.
- Forster P. Ice Ages and the mitochondrial DNA chronology of human dispersals: a review. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, **359**(1442), 255-264, 2004.
- Forster P., Harding R., Torroni A. and Bandelt H.-J. Origin and evolution of Native American mtDNA variation: a reappraisal. *Am. J. Hum. Genet.*, **59**(4), 935-945, 1996.
- Forster P., Torroni A., Renfrew C. and Rohl A. Phylogenetic star contraction applied to Asian and Papuan mtDNA evolution. *Mol. Biol. Evol.*, **18**(10), 1864-1881, 2001.
- Fox R. Kinship and marriage: an anthropological perspective. Cambridge University Press, Cambridge, United Kingdom, 1983.
- Friedlaender J., Schurr T., Gentz F., Koki G., Friedlaender F., Horvat G., Babb P., Cerchio S., Kaestle F., Schanfield M., Deka R., Yanagihara R. and Merriwether D.A. Expanding Southwest Pacific mitochondrial haplogroups P and Q. *Mol. Biol. Evol.*, **22**(6), 1506-1517, 2005.
- Friedlaender J.S., Friedlaender F.R., Hodgson J.A., Stoltz M., Koki G., Horvat G., Zhadanov S., Schurr T.G. and Merriwether D.A. Melanesian mtDNA complexity. *PLoS ONE*, **2**(2), e248, 2007.
- Fu Y.-X. and Li W.-H. Statistical tests of neutrality of mutations. *Genetics*, **133**, 693-709, 1993.
- Fu Y.-X. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics*, **147**(2), 915-925, 1997.

- Galtier N., Enard D., Radondy Y., Bazin E. and Belkhir K. Mutation hot spots in mammalian mitochondrial DNA. *Genome Res.*, **16**(2), 215-222, 2006.
- Gelfand R. and Attardi G. Synthesis and turnover of mitochondrial ribonucleic acid in HeLa cells: the mature ribosomal and messenger ribonucleic acid species are metabolically unstable. *Mol. Cell. Biol.*, **1**(6), 497-511, 1981.
- Ghivizzani S.C., Madsen C.S., Nelen M.R., Ammini C.V. and Hauswirth W.W. In organello footprint analysis of human mitochondrial DNA: human mitochondrial transcription factor A interactions at the origin of replication. *Mol. Cell. Biol.*, **14**(12), 7717-7730, 1994.
- Gilbert M.T., Willerslev E., Hansen A.J., Barnes I., Rudbeck L., Lynnerup N. and Cooper A. Distribution patterns of postmortem damage in human mitochondrial DNA. *Am. J. Hum. Genet.*, **72**(1), 32-47, 2003.
- Giles R.E., Blanc H., Cann H.M. and Wallace D.C. Maternal inheritance of human mitochondrial DNA. *Proc. Natl. Acad. Sci. U.S.A.*, **77**(11), 6715-6719, 1980.
- Gill P., Ivanov P.L., Kimpton C., Piercy R., Benson N., Tully G., Evett I., Hagelberg E. and Sullivan K. Identification of the remains of the Romanov family by DNA analysis. *Nat. Genet.*, **6**, 130-135, 1994.
- Gillham N.W. Organelle genes and genomes. First edition. Oxford University Press, Inc., New York, USA, 1994.
- Gleyzer N., Vercauteren K. and Scarpulla R.C. Control of mitochondrial transcription specificity factors (TFB1M and TFB2M) by nuclear respiratory factors (NRF-1 and NRF-2) and PGC-1 family coactivators. *Mol. Cell. Biol.*, **25**(4), 1354-1366, 2005.
- Glitz C., Zwieb C., Brimacombe R., Edwards K. and Kössel H. Secondary structure of the large subunit ribosomal RNA from *Escherichia coli*, *Zea mays* chloroplast, and human and mouse mitochondrial ribosomes. *Nucleic Acids Res.*, **9**(14), 3287-3306, 1981.
- Gocke C.D., Benko F.A. and Rogan P.K. Transmission of mitochondrial DNA heteroplasmy in normal pedigrees. *Hum. Genet.*, **102**, 182-186, 1998.
- Goldstein D.B., Ruiz Linares A., Cavalli-Sforza L.L. and Feldman M.W. Genetic absolute dating based on microsatellites and the origin of modern humans. *Proc. Natl. Acad. Sci. U.S.A.*, **92**, 6723-6727, 1995.
- Gonder M.K., Mortensen H.M., Reed F.A., De Sousa A. and Tishkoff S.A. Whole-mtDNA genome sequence analysis of ancient African lineages. *Mol. Biol. Evol.*, **24**(3), 757-768, 2007.
- González A.M., Cabrera V.M., Larruga J.M., Tounkara A., Nouns G., Thomas B.N. and Moulds J.M. Mitochondrial DNA variation in Mauritania and Mali and their genetic relationship to other Western Africa populations. *Ann. Hum. Genet.*, **70**(5), 631-657, 2006.
- Gordon R.G. Ethnologue: languages of the world. Fifteenth edition. SIL International, Dallas, Texas, USA, 2005.
- Graven L., Passarino G., Semino O., Bowsot P., Santachiara-Benerecetti S., Langaney A. and Excoffier L. Evolutionary correlation between control region sequence and restriction polymorphisms in the mitochondrial genome of a large Senegalese Mandenka sample. *Mol. Biol. Evol.*, **12**(2), 334-345, 1995.
- Green R.E., Malaspina A.S., Krause J., Briggs A.W., Johnson P.L., Uhler C., Meyer M., Good J.M., Maricic T., Stenzel U., Prüfer K., Siebauer M., Burbano H.A., Ronan M., Rothberg J.M., Egholm M., Rudan P., Brajković D., Kučan Z., Gušić I., Wikström M., Laakkonen L., Kelso J., Slatkin M. and Pääbo S. A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell*, **134**(3), 416-426, 2008.
- Greenberg B.D., Newbold J.E. and Sugino A. Intraspecific nucleotide sequence variability surrounding the origin of replication in human mitochondrial DNA. *Gene*, **21**(1-2), 33-49, 1983.
- Greenberg J.H. Studies in African linguistic classification: VI. The click languages. *Southwest J Anthropol*, **6**, 223-237, 1950.
- Grosschedl R., Giese K. and Pagel J. HMG domain proteins: architectural elements in the assembly of nucleoprotein structures. *Trends Genet.*, **10**(3), 94-100, 1994.
- Gu X. and Zhang J. A simple method for estimating the parameter of substitution rate variation among sites. *Mol. Biol. Evol.*, **14**, 1106-1113, 1997.
- Güldemann T. and Vossen R. Khoisan. In: African languages: an introduction. Heine B. and Nurse D. (editors). Cambridge University Press, Cambridge, 99-122, 2000.
- Güldemann T. Greenberg's "case" for Khoisan: the morphological evidence. In: Problems of linguistic-historical reconstruction in Africa. Ibrizimow D. (editor). Sprache und Geschichte in Afrika. Rüdiger Köppe, Köln, **19**, 2008.
- Hagelberg E., Goldman N., Lio P., Whelan S., Schiefenhövel W., Clegg J.B. and Bowden D.K. Evidence for mitochondrial DNA recombination in a human population of island Melanesia. *Proc Biol Sci.*, **266**(1418), 485-492, 1999.
- Hagelberg E., Goldman N., Lió P., Whelan S., Schiefenhövel W., Clegg J.B. and Bowden D.K. Evidence for mitochondrial DNA recombination in a human population of island Melanesia: correction. *Proc. R. Soc. Lond., B, Biol. Sci.*, **267**(1452), 1595-1596, 2000.

- Hall B.G. *Phylogenetic trees made easy: a how-to manual*. Second edition. Sinauer Associates, Inc., Sunderland, Massachusetts, USA, 2004.
- Hall T.A. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp. Ser.*, **41**, 95-98, 1999.
- Hammer M.F., Blackmer F., Garrigan D., Nachman M.W. and Wilder J.A. Human population structure and its effects on sampling Y chromosome sequence variation. *Genetics*, **164**(4), 1495-1509, 2003.
- Hammer M.F., Garrigan D., Wood E., Wilder J.A., Mobasher Z., Bigham A., Krenz J.G. and Nachman M.W. Heterogeneous patterns of variation among multiple human X-linked loci: the possible role of diversity-reducing selection in non-Africans. *Genetics*, **167**(4), 1841-1853, 2004.
- Harding R.M., Fullerton S.M., Griffiths R.C., Bond J., Cox M.J., Schneider J.A., Moulin D.S. and Clegg J.B. Archaic African and Asian lineages in the genetic ancestry of modern humans. *Am. J. Hum. Genet.*, **60**(4), 772-789, 1997.
- Harpending H.C. and Rogers A. Genetic perspectives on human origins and differentiation. *Annu Rev Genomics Hum Genet*, **1**, 361-385, 2000.
- Harpending H.C. Signature of ancient population growth in a low-resolution mitochondrial DNA mismatch distribution. *Hum. Biol.*, **66**(4), 591-600, 1994.
- Harpending H.C., Sherry S.T., Rogers A.R. and Stoneking M. Genetic structure of ancient human populations. *Curr. Anthropol.*, **34**, 483-496, 1993.
- Harris E.E. and Hey J. X chromosome evidence for ancient human histories. *Proc. Natl. Acad. Sci. U.S.A.*, **96**, 3320-3324, 1999.
- Harris J.E., Jalloh A., Inikori J.E., Palmer C.A., Chambers D.B. and Grader D.T. *The African diaspora*. Texas A&M University Press, Texas, USA, 1996.
- Hartl D.L. *A primer of population genetics*. Second edition. Sinauer Associates, Inc., Sunderland, Massachusetts, USA, 1988.
- Hasegawa M., Di Rienzo A., Kocher T.D. and Wilson A.C. Toward a more accurate time scale for the human mitochondrial DNA tree. *J. Mol. Evol.*, **37**(4), 347-354, 1993.
- Hasegawa M., Kishino H. and Yano T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.*, **22**(2), 160-174, 1985.
- Hauswirth W.W. and Clayton D.A. Length heterogeneity of a conserved displacement-loop sequence in human mitochondrial DNA. *Nucleic Acids Res.*, **13**(22), 8093-8104, 1985.
- Hedges S.B. The number of replications needed for accurate estimation of the bootstrap *P* Value in phylogenetic studies. *Mol. Biol. Evol.*, **9**(2), 366-369, 1992.
- Helgason A., Hickey E., Goodacre S., Bosnes V., Stefánsson K., Ward R. and Sykes B. mtDNA and the islands of the North Atlantic: estimating the proportions of Norse and Gaelic ancestry. *Am. J. Hum. Genet.*, **68**(3), 723-737, 2001.
- Helgason A., Pálsson G., Pedersen H.S., Angulalik E., Gunnarsdóttir E.D., Yngvadóttir B. and Stefánsson K. mtDNA variation in Inuit populations of Greenland and Canada: migration history and population structure. *Am. J. Phys. Anthropol.*, **130**(1), 123-134, 2006.
- Henshilwood C.S., d'Errico F., Yates R., Jacobs Z., Tribolo C., Duller G.A., Mercier N., Sealy J.C., Valladas H., Watts I. and Wintle A.G. Emergence of modern human behavior: Middle Stone Age engravings from South Africa. *Science*, **295**(5558), 1278-1280, 2002.
- Herrnstadt C., Elson J.L., Fahy E., Preston G., Turnbull D.M., Anderson C., Ghosh S.S., Olefsky J.M., Beal M.F., Davis R.E. and Howell N. Reduced-median-network analysis of complete mitochondrial DNA coding-region sequences for the major African, Asian, and European haplogroups. *Am. J. Hum. Genet.*, **70**(5), 1152-1171, 2002a.
- Herrnstadt C., Preston G., Andrews R., Chinnery P., Lightowlers R.N., Turnbull D.M., Kubacka I. and Howell N. A high frequency of mtDNA polymorphisms in HeLa cell sublines. *Mutat. Res.*, **501**(1-2), 19-28, 2002b.
- Hess J.F., Parisi M.A., Bennett J.L. and Clayton D.A. Impairment of mitochondrial transcription termination by a point mutation associated with the MELAS subgroup of mitochondrial encephalomyopathies. *Nature*, **351**(6323), 236-239, 1991.
- Heyer E., Zietkiewicz E., Rochowski A., Yotova V., Puymirat J. and Labuda D. Phylogenetic and familial estimates of mitochondrial substitution rates: study of control region mutations in deep-rooting pedigrees. *Am. J. Hum. Genet.*, **69**(5), 1113-1126, 2001.
- Hixson J.E. and Clayton D.A. Initiation of transcription from each of the two human mitochondrial promoters requires unique nucleotides at the transcriptional start sites. *Proc. Natl. Acad. Sci. U.S.A.*, **82**(9), 2660-2664, 1985.
- Hixson J.E., Wong T.W. and Clayton D.A. Both the conserved stem-loop and divergent 5'-flanking sequences are required for initiation at the human mitochondrial origin of light-strand DNA replication. *J. Biol. Chem.*, **261**(5), 2384-2390, 1986.
- Holder M. and Lewis P.O. Phylogeny estimation: traditional and Bayesian approaches. *Nat. Rev. Genet.*, **4**(4), 275-284, 2003.

- Holland M.M., Fisher D.L., Mitchell L.G., Rodriguez W.C., Canik J.J., Merrill C.R. and Weedn V.W. Mitochondrial DNA sequence analysis of human skeletal remains: identification of remains from the Vietnam War. *J. Forensic Sci.*, **38**, 542-553, 1993.
- Holt I.J., Harding A.E. and Morgan-Hughes J.A. Deletions of muscle mitochondrial DNA in patients with mitochondrial myopathies. *Nature*, **331**(6158), 717-719, 1988.
- Horai S. and Hayasaka K. Intraspecific nucleotide sequence differences in the major noncoding region of human mitochondrial DNA. *Am. J. Hum. Genet.*, **46**(4), 828-842, 1990.
- Horai S., Hayasaka K., Kondo R., Tsugane K. and Takahata N. Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. *Proc. Natl. Acad. Sci. U.S.A.*, **92**, 532-536, 1995.
- Horai S., Satta Y., Hayasaka K., Kondo R., Inoue T., Ishida T., Hayashi S. and Takahata N. Man's place in Hominoidea revealed by mitochondrial DNA genealogy. *J. Mol. Evol.*, **35**(1), 32-43, 1992.
- Horváth R., Scharfe C., Hoeltzenbein M., Do B.H., Schröder C., Warzok R., Vogelgesang S., Lochmüller H., Müller-Höcker J., Gerbitz K.D., Oefner P.J. and Jaksch M. Childhood onset mitochondrial myopathy and lactic acidosis caused by a stop mutation in the mitochondrial cytochrome c oxidase III gene. *J. Med. Genet.*, **39**(11), 812-816, 2002.
- Houshmand M., Larsson N.-G., Holme E., Oldfors A., Tulinius M.H. and Andersen O. Automatic sequencing of mitochondrial tRNA genes in patients with mitochondrial encephalomyopathy. *Biochim. Biophys. Acta*, **1226**(1), 49-55, 1994.
- Howell N. and Mackey D. Reply to Macauley et al. *Am. J. Hum. Genet.*, **61**, 986-990, 1997.
- Howell N. mtDNA recombination: what do in vitro data mean? *Am. J. Hum. Genet.*, **61**, 19-22, 1997.
- Howell N., Bindoff L.A., McCullough D.A., Kubacka I., Poulton J., Mackey D., Taylor L. and Turnbull D.M. Leber hereditary optic neuropathy: identification of the same mitochondrial ND1 mutation in six pedigrees. *Am. J. Hum. Genet.*, **49**(5), 939-950, 1991.
- Howell N., Elson J.L., Turnbull D.M. and Herrnstadt C. African haplogroup L mtDNA sequences show violations of clock-like evolution. *Mol. Biol. Evol.*, **21**(10), 1843-1854, 2004.
- Howell N., Kubacka I. and Mackey D.A. How rapidly does the human mitochondrial genome evolve? *Am. J. Hum. Genet.*, **59**(3), 501-509, 1996.
- Howell N., Kubacka I., Halvorson S., Howell B., McCullough D.A. and Mackey D. Phylogenetic analysis of the mitochondrial genomes from Leber hereditary optic neuropathy pedigrees. *Genetics*, **140**(1), 285-302, 1995.
- Howell N., McCullough D.A., Kubacka I., Halvorson S. and Mackey D. The sequence of human mtDNA: the question of errors versus polymorphisms. *Am. J. Hum. Genet.*, **50**(6), 1333-1340, 1992.
- Howell N., Smejkal C.B., Mackey D.A., Chinnery P.F., Turnbull D.M. and Herrnstadt C. The pedigree rate of sequence divergence in the human mitochondrial genome: there is a difference between phylogenetic and pedigree rates. *Am. J. Hum. Genet.*, **72**(3), 659-670, 2003.
- Huang X. A contig assembly program based on sensitive detection of fragment overlaps. *Genomics*, **14**(1), 18-25, 1992.
- Hudson R.R. Gene genealogies and the coalescent process. In: Oxford surveys in evolutionary biology. Volume 7. Futuyma D.J. and Antonovics J. (editors). Oxford University Press, Oxford, England, 1-44, 1990.
- Huelsenbeck J.P., Ronquist F., Nielsen R. and Bollback J.P. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, **294**(5550), 2310-2314, 2001.
- Huffman T.N. Bantu migrations in southern Africa. In: The prehistory of Africa. Tracing the lineage of modern man. Soodyall H. (editor). Jonathan Ball Publishers, Johannesburg, South Africa, 97-108, 2006.
- Huoponen K., Lamminen T., Juvonen V., Aula P., Nikoskelainen E. and Savontaus M.-L. The spectrum of mitochondrial DNA mutations in families with Leber hereditary optic neuroretinopathy. *Hum. Genet.*, **92**(4), 379-384, 1993.
- Ikemura T. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.*, **2**(1), 13-34, 1985.
- Ikemura T. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *J. Mol. Biol.*, **151**(3), 389-409, 1981.
- Ingman M. and Gyllenstein U. Analysis of the complete human mtDNA genome: methodology and inferences for human evolution. *J. Hered.*, **92**, 454-461, 2001.
- Ingman M. and Gyllenstein U. Mitochondrial genome variation and evolutionary history of Australian and New Guinean aborigines. *Genome Res.*, **13**, 1600-1606, 2003.
- Ingman M. and Gyllenstein U. Rate variation between mitochondrial domains and adaptive evolution in humans. *Hum. Mol. Genet.*, **16**(19), 2281-2287, 2007b.
- Ingman M. and Gyllenstein U.A. Recent genetic link between Sami and the Volga-Ural region of Russia. *Eur. J. Hum. Genet.*, **15**(1), 115-120, 2007a.
- Ingman M., Kaessmann H., Pääbo S. and Gyllenstein U. Mitochondrial genome variation and the origin of modern humans. *Nature*, **408**, 708-713, 2000.

- Jazin E., Soodyall H., Jalonen P., Lindholm E., Stoneking M. and Gyllensten U. Mitochondrial mutation rate revisited: hot spots and polymorphism. *Nat. Genet.*, **18**, 109-110, 1998.
- Jenuth J.P., Peterson A.C. and Shoubridge E.A. Tissue-specific selection for different mtDNA genotypes in heteroplasmic mice. *Nat. Genet.*, **16**(1), 93-95, 1997.
- Jobling M.A., Hurles M.E. and Tyler-Smith C. Human evolutionary genetics. Origins, peoples and disease. Garland Science, New York, USA, 2004.
- Johns D. Allelic mutations of the fourth subunit of NADH dehydrogenase are not pathogenetically important in 11778-negative Leber Hereditary Optic Neuropathy. *Am. J. Hum. Genet.*, **48**(6), 1209-1212, 1991.
- Johns D.R. and Neufeld M.J. Cytochrome c oxidase mutations in Leber hereditary optic neuropathy. *Biochem. Biophys. Res. Commun.*, **196**(2), 810-815, 1993.
- Johns D.R. Reply to Brown et al. *Am. J. Hum. Genet.*, **55**(2), 410-412, 1994.
- Johnson M.J., Wallace D.C., Ferris S.D., Rattazzi M.C. and Cavalli-Sforza L.L. Radiation of human mitochondria DNA types analyzed by restriction endonuclease cleavage patterns. *J. Mol. Evol.*, **19**, 255-271, 1983.
- Jones J.B., Song J.J., Hempen P.M., Parmigiani G., Hruban R.H. and Kern S.E. Detection of mitochondrial DNA mutations in pancreatic cancer offers a "mass"-ive advantage over detection of nuclear DNA mutations. *Cancer Res.*, **61**(4), 1299-1304, 2001.
- Jones M.M., Manwaring N., Wang J.J., Rochtchina E., Mitchell P. and Sue C.M. Mitochondrial DNA haplogroups and age-related maculopathy. *Arch. Ophthalmol.*, **125**(9), 1235-1240, 2007.
- Jorde L.B. and Bamshad M. Questioning evidence for recombination in human mitochondrial DNA. *Science*, **288**, 1931a, 2000.
- Jorde L.B., Bamshad M. and Rogers A.R. Using mitochondrial and nuclear DNA markers to reconstruct human evolution. *Bioessays*, **20**, 126-136, 1998.
- Jorde L.B., Watkins W.S. and Bamshad M.J. Population genomics: a bridge from evolutionary history to genetic medicine. *Hum. Mol. Genet.*, **10**, 2199-2207, 2001.
- Jorde L.B., Watkins W.S., Bamshad M.J., Dixon M.E., Ricker C.E., Seielstad M.T. and Batzer M.A. The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data. *Am. J. Hum. Genet.*, **66**(3), 979-988, 2000.
- Just R.S., Diegoli T.M., Saunier J.L., Irwin J.A. and Parsons T.J. Complete mitochondrial genome sequences for 265 African American and U.S. "Hispanic" individuals. *Forensic Sci Int Genet*, **2**(3), e45-48, 2008.
- Kalman B., Lublin F.D. and Alder H. Characterization of the mitochondrial DNA in patients with multiple sclerosis. *J. Neurol. Sci.*, **140**(1-2), 75-84, 1996.
- Kaneda H., Hayashi J., Takahama S., Taya C., Lindahl K.F. and Yonekawa H. Elimination of paternal mitochondrial DNA in intraspecific crosses during early mouse embryogenesis. *Proc. Natl. Acad. Sci. U.S.A.*, **92**(10), 4542-4546, 1995.
- Kayser M., Brauer S., Weiss G., Schiefenhövel W., Underhill P., Shen P., Oefner P., Tommaseo-Ponzetta M. and Stoneking M. Reduced Y-chromosome, but not mitochondrial DNA, diversity in human populations from West New Guinea. *Am. J. Hum. Genet.*, **72**, 281-302, 2003.
- Keightley J.A., Anitori R., Burton M.D., Quan F., Buist N.R. and Kennaway N.G. Mitochondrial encephalomyopathy and complex III deficiency associated with a stop-codon mutation in the cytochrome b gene. *Am. J. Hum. Genet.*, **67**(6), 1400-1410, 2000.
- Keightley J.A., Hoffbuhr K.C., Burton M.D., Salas V.M., Johnston W.S., Penn A.M., Buist N.R. and Kennaway N.G. A microdeletion in cytochrome c oxidase (COX) subunit III associated with COX deficiency and recurrent myoglobinuria. *Nat. Genet.*, **12**(4), 410-416, 1996.
- Kemp B.M., Malhi R.S., McDonough J., Bolnick D.A., Eshleman J.A., Rickards O., Martinez-Labarga C., Johnson J.R., Lorenz J.G., Dixon E.J., Fifield T.E., Heaton T.H., Worl R. and Smith D.G. Genetic analysis of early Holocene skeletal remains from Alaska and its implications for the settlement of the Americas. *Am. J. Phys. Anthropol.*, **132**(4), 605-621, 2007.
- Keyser-Tracqui C., Crubézy E. and Ludes B. Nuclear and mitochondrial DNA analysis of a 2,000-year-old necropolis in the Egyin Gol Valley of Mongolia. *Am. J. Hum. Genet.*, **73**, 247-260, 2003.
- Kibbe W.A. OligoCalc: an online oligonucleotide properties calculator. *Nucleic Acids Res.*, **35**(Web Server issue), W43-46, 2007.
- Kimura M. Evolutionary rate at the molecular level. *Nature*, **217**(5129), 624-626, 1968.
- Kimura M. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, **61**(4), 893-903, 1969.
- Kimura M. Theoretical foundation of population genetics at the molecular level. *Theor Popul Biol*, **2**(2), 174-208, 1971.
- King J.L. and Jukes T.H. Non-Darwinian evolution. *Science*, **164**(881), 788-798, 1969.
- Kirby D.M., Boneh A., Chow C.W., Ohtake A., Ryan M.T., Thyagarajan D. and Thorburn D.R. Low mutant load of mitochondrial DNA G13513A mutation can cause Leigh's disease. *Ann. Neurol.*, **54**(4), 473-478, 2003.

- Kishino H. and Hasegawa M. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. *J. Mol. Evol.*, **29**, 170-179, 1989.
- Kivisild T., Reidla M., Metspalu E., Rosa A., Brehm A., Pennarun E., Panik J., Geberhiwot T., Usanga E. and Villems R. Ethiopian mitochondrial DNA heritage: tracking gene flow across and around the gate of tears. *Am. J. Hum. Genet.*, **75**, 752-770, 2004.
- Kivisild T., Shen P., Wall D.P., Do B., Sung R., Davis K.K., Passarino G., Underhill P.A., Scharfe C., Torroni A., Scozzari R., Modiano D., Coppa A., De Knijff P., Feldman M.W., Cavalli-Sforza L.L. and Oefner P.J. The role of selection in the evolution of human mitochondrial genomes. *Genetics*, **172**(1), 373-387, 2006.
- Kivisild T., Tolk H.-V., Parik J., Wang Y., Papiha S.S., Bandelt H.-J. and Villems R. The emerging limbs and twigs of the East Asian mtDNA tree. *Mol. Biol. Evol.*, **19**(10), 1737-1751, 2002.
- Klein H.S., Engerman S.L., Haines R. and Shlomowitz R. Transoceanic mortality: the slave trade in comparative perspective. *William Mary Q.*, **58**(1), 93-117, 2001.
- Klysik J., Stirdivant S.M., Larson J.E., Hart P.A. and Wells R.D. Left-handed DNA in restriction fragments and a recombinant plasmid. *Nature*, **290**(5808), 672-677, 1981.
- Knight A., Underhill P.A., Mortensen H.M., Zhivotovsky L.A., Lin A.A., Henn B.M., Louis D., Ruhlen M. and Mountain J.L. African Y chromosome and mtDNA divergence provides insight into the history of click languages. *Curr. Biol.*, **13**(6), 464-473, 2003.
- Kobayashi Y., Momoi M.Y., Tominaga K., Shimoizumi H., Nihei K., Yanagisawa M., Kagawa Y. and Ohta S. Respiration-deficient cells are caused by a single point mutation in the mitochondrial tRNA-Leu (UUR) gene in mitochondrial myopathy, encephalopathy, lactic acidosis, and stroke-like episodes (MELAS). *Am. J. Hum. Genet.*, **49**(3), 590-599, 1991.
- Kong Q.-P., Yao Y.-G., Sun C., Bandelt H.-J., Zhu C.-L. and Zhang Y.-P. Phylogeny of east Asian mitochondrial DNA lineages inferred from complete sequences. *Am. J. Hum. Genet.*, **73**, 671-676, 2003.
- Kösel S., Grasbon-Frodl E.M., Mautsch U., Egensperger R., Von Eitzen U., Frishman D., Hofmann S., Gerbitz K.D., Mehraein P. and Graeber M.B. Novel mutations of mitochondrial complex I in pathologically proven Parkinson disease. *Neurogenetics*, **1**(3), 197-204, 1998.
- Kraysberg Y., Schwartz M., Brown T.A., Ebralidse K., Kunz W.S., Clayton D.A., Vissing J. and Khrapko K. Recombination of human mitochondrial DNA. *Science*, **304**, 981, 2004.
- Kruse B., Narasimhan N. and Attardi G. Termination of transcription in human mitochondria: identification and purification of a DNA binding protein factor that promotes termination. *Cell*, **58**(2), 391-397, 1989.
- Kujoth G.C., Hiona A., Pugh T.D., Someya S., Panzer K., Wohlgemuth S.E., Hofer T., Seo A.Y., Sullivan R., Jobling W.A., Morrow J.D., Van Remmen H., Sedivy J.M., Yamasoba T., Tanokura M., Weindrich R., Leeuwenburgh C. and Prolla T.A. Mitochondrial DNA mutations, oxidative stress, and apoptosis in mammalian aging. *Science*, **309**(5733), 481-484, 2005.
- Kumar S., Hedrick P., Dowling T. and Stoneking M. Questioning evidence for recombination in human mitochondrial DNA. *Science*, **288**, 1931a, 2000.
- Kumar S., Ravuri R.R., Koneru P., Urade B.P., Sarkar B.N., Chandrasekar A. and Rao V.R. Reconstructing Indian-Australian phylogenetic link. *BMC Evol. Biol.*, **9**, 173, 2009.
- Kumar S., Tamura K. and Nei M. MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief. Bioinformatics*, **5**, 150-163, 2004.
- Ladika S. DNA forensics. Laying ghosts to rest in Bosnia. *Science*, **293**, 1422-1423, 2001.
- Ladoukakis E.D. and Zouros E. Direct evidence for homologous recombination in mussel (*Mytilus galloprovincialis*) mitochondrial DNA. *Mol. Biol. Evol.*, **18**, 1168-1175, 2001.
- Lahermo P., Sajantila A., Sistonen P., Lukka M., Aula P., Peltonen L. and Savontaus M.-L. The genetic relationship between the Finns and the Finnish Saami (Lapps): analysis of nuclear DNA and mtDNA. *Am. J. Hum. Genet.*, **58**(6), 1309-1322, 1996.
- Lahr M.M. and Foley R.A. Towards a theory of modern human origins: geography, demography, and diversity in recent human evolution. *Am. J. Phys. Anthropol.*, **107**(S27), 137-176, 1998.
- Lahr M.M. The evolution of modern human diversity: a study of cranial variation. Cambridge University Press, New York, USA, 1996.
- Lamantea E., Carrara F., Mariotti C., Morandi L., Tiranti V. and Zeviani M. A novel nonsense mutation (Q352X) in the mitochondrial cytochrome *b* gene associated with a combined deficiency of complexes I and III. *Neuromuscul. Disord.*, **12**(1), 49-52, 2002.
- Larsson N.-G. and Clayton D.A. Molecular genetic aspects of human mitochondrial disorders. *Annu. Rev. Genet.*, **29**, 151-178, 1995.
- Lawler A. Terrorism. Massive DNA identification effort gets under way. *Science*, **294**, 278-279, 2001.
- Leahey R. The origin of humankind. Phoenix, London, 1994.
- Lee D.Y. and Clayton D.A. RNase mitochondrial RNA processing correctly cleaves a novel R loop at the mitochondrial DNA leading-strand origin of replication. *Genes Dev.*, **11**(5), 582-592, 1997.

- Legros F., Chatzoglou E., Frachon P., Ogier De Baulny H., Laforêt P., Jardel C., Godinot C. and Lombès A. Functional characterization of novel mutations in the human cytochrome b gene. *Eur. J. Hum. Genet.*, **9**(7), 510-518, 2001.
- Levinson G. and Gutman G.A. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol. Biol. Evol.*, **4**(3), 203-221, 1987.
- Li R., Qu J., Zhou X., Tong Y., Hu Y., Qian Y., Lu F., Mo J.Q., West C.E. and Guan M.X. The mitochondrial tRNA^{Thr} A15951G mutation may influence the phenotypic expression of the LHON-associated ND4 G11778A mutation in a Chinese family. *Gene*, **376**(1), 79-86, 2006.
- Liao H.X. and Spremulli L.L. Effects of length and mRNA secondary structure on the interaction of bovine mitochondrial ribosomes with messenger RNA. *J. Biol. Chem.*, **265**(20), 11761-11765, 1990a.
- Liao H.X. and Spremulli L.L. Identification and initial characterization of translational initiation factor 2 from bovine mitochondria. *J. Biol. Chem.*, **265**(23), 13618-13622, 1990b.
- Liao H.X. and Spremulli L.L. Initiation of protein synthesis in animal mitochondria. Purification and characterization of translational initiation factor 2. *J. Biol. Chem.*, **266**(31), 20714-20719, 1991.
- Liao H.X. and Spremulli L.L. Interaction of bovine mitochondrial ribosomes with messenger RNA. *J. Biol. Chem.*, **264**(13), 7518-7522, 1989.
- Life Technologies®. Instruction Manual. HORIZON® 58 Horizontal Gel Electrophoresis Apparatus, USA, 2000.
- Lin F., Lin R., Wisniewski H.M., Hwang Y., Grundke-Iqbal I., Healy-Louie G. and Iqbal K. Detection of point mutations in codon 331 of mitochondrial NADH dehydrogenase subunit 2 in Alzheimer's brains. *Biochem. Biophys. Res. Commun.*, **182**(1), 238-246, 1992.
- Liolitsa D., Rahman S., Benton S., Carr L.J. and Hanna M.G. Is the mitochondrial complex I ND5 gene a hot-spot for MELAS causing mutations? *Ann. Neurol.*, **53**(1), 128-132, 2003.
- Liu V.W., Shi H.H., Cheung A.N., Chiu P.M., Leung T.W., Nagley P., Wong L.C. and Ngan H.Y. High incidence of somatic mitochondrial DNA mutations in human ovarian carcinomas. *Cancer Res.*, **61**(16), 5998-6001, 2001.
- Lodish H., Baltimore D., Berk A., Zipursky S.L., Matsudaira P. and Darnell J. Molecular cell biology. Third edition. Scientific American Books, Inc., New York, USA, 1995.
- Logan J.J. A comprehensive analysis of mtDNA haplogroup J. *J Genet Geneol.*, **4**, 104-124, 2008.
- Loogväli E.L., Roostalu U., Malyarchuk B.A., Derenko M.V., Kivisild T., Metspalu E., Tambets K., Reidla M., Tolk H.-V., Parik J., Pennarun E., Laos S., Lunkina A., Golubenko M., Barac L., Pericic M., Balanovsky O.P., Gusar V., Khusnutdinova E.K., Stepanov V., Puzyrev V., Rudan P., Balanovska E.V., Grechanina E., Richard C., Moisan J.P., Chaventré A., Anagnou N.P., Pappa K.I., Michalodimitrakis E.N., Claustres M., Gölge M., Mikerezi I., Usanga E. and Villems R. Disuniting uniformity: a pied cladistic canvas of mtDNA haplogroup H in Eurasia. *Mol. Biol. Evol.*, **21**(11), 2012-2021, 2004.
- Lundberg K.S., Shoemaker D.D., Adams M.W., Short J.M., Sorge J.A. and Mathur E.J. High-fidelity amplification using a thermostable DNA polymerase isolated from *Pyrococcus furiosus*. *Gene*, **108**(1), 1-6, 1991.
- Ma L. and Spremulli L.L. Cloning and sequence analysis of the human mitochondrial translational initiation factor 2 cDNA. *J. Biol. Chem.*, **270**(4), 1859-1865, 1995.
- Maca-Meyer N., González A.M., Larruga J.M., Flores C. and Cabrera V.M. Major genomic mitochondrial lineages delineate early human expansions. *BMC Genet.*, **2**, 13, 2001.
- Maca-Meyer N., González A.M., Pestano J., Flores C., Larruga J.M. and Cabrera V.M. Mitochondrial DNA transit between West Asia and North Africa inferred from U6 phylogeography. *BMC Genet.*, **4**, 15, 2003.
- Macaulay V., Hill C., Achilli A., Rengo C., Clarke D., Meehan W., Blackburn J., Semino O., Scozzari R., Cruciani F., Taha A., Shaari N.K., Raja J.M., Ismail P., Zainuddin Z., Goodwin W., Bulbeck D., Bandelt H.-J., Oppenheimer S., Torroni A. and Richards M. Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. *Science*, **308**(5724), 1034-1036, 2005.
- Macaulay V., Richards M. and Sykes B. Mitochondrial DNA recombination – no need to panic. *Proc. R. Soc. Lond., B, Biol. Sci.*, **266**, 2037-2039, 1999a.
- Macaulay V., Richards M., Hickey E., Vega E., Cruciani F., Guida V., Scozzari R., Bonnè-Tamir B., Sykes B. and Torroni A. The emerging tree of West Eurasian mtDNAs: a synthesis of control-region sequences and RFLPs. *Am. J. Hum. Genet.*, **64**(1), 232-249, 1999b.
- Macaulay V.A., Richards M.B., Forster P., Bendall K.E., Watson E., Sykes B. and Bandelt H.-J. mtDNA mutation rates - no need to panic. *Am. J. Hum. Genet.*, **61**, 983-986, 1997.
- Mackey D. and Howell N. A variant of Leber hereditary optic neuropathy characterized by recovery of vision and by an unusual mitochondrial genetic etiology. *Am. J. Hum. Genet.*, **51**(6), 1218-1228, 1992.
- Maddison D.R., Ruvolo M. and Swofford D.L. Geographic origins of human mitochondrial DNA: phylogenetic evidence from control region sequences. *Syst. Biol.*, **41**(1), 111-124, 1992.
- Maitra A., Cohen Y., Gillespie S.E., Mambo E., Fukushima N., Hoque M.O., Shah N., Goggins M., Califano J., Sidransky D. and Chakravarti A. The human MitoChip: a high-throughput sequencing microarray for mitochondrial mutation detection. *Genome Res.*, **14**(5), 812-819, 2004.

- Mancuso M., Filosto M., Stevens J.C., Patterson M., Shanske S., Krishna S. and DiMauro S. Mitochondrial myopathy and complex III deficiency in a patient with a new stop-codon mutation (G339X) in the cytochrome *b* gene. *J. Neurol. Sci.*, **209**(1-2), 61-63, 2003.
- Manfredi G., Thyagarajan D., Papadopoulou L.C., Pallotti F. and Schon E.A. The fate of human sperm-derived mtDNA in somatic cells. *Am. J. Hum. Genet.*, **61**(4), 953-960, 1997.
- Margulies M., Egholm M., Altman W.E., Attiya S., Bader J.S., Bemben L.A., Berka J., Braverman M.S., Chen Y.J., Chen Z., Dewell S.B., Du L., Fierro J.M., Gomes X.V., Godwin B.C., He W., Helgesen S., Ho C.H., Irzyk G.P., Jando S.C., Alenquer M.L., Jarvie T.P., Jirage K.B., Kim J.B., Knight J.R., Lanza J.R., Leamon J.H., Lefkowitz S.M., Lei M., Li J., Lohman K.L., Lu H., Makhijani V.B., McDade K.E., McKenna M.P., Myers E.W., Nickerson E., Nobile J.R., Plant R., Puc B.P., Ronan M.T., Roth G.T., Sarkis G.J., Simons J.F., Simpson J.W., Srinivasan M., Tartaro K.R., Tomasz A., Vogt K.A., Volkmer G.A., Wang S.H., Wang Y., Weiner M.P., Yu P., Begley R.F. and Rothberg J.M. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**(7057), 376-380, 2005.
- Margulis L. Origin of eukaryotic cells: evidence and research implications for a theory of the origin and evolution of microbial, plant, and animal cells on the Precambrian earth. Yale University Press, New Haven, USA, 1970.
- Marjoram P. and Donnelly P. Pairwise comparisons of mitochondrial DNA sequences in subdivided populations and implications for early human evolution. *Genetics*, **136**(2), 673-683, 1994.
- Martin A.P. Metabolic rate and directional nucleotide substitution in animal mitochondrial DNA. *Mol. Biol. Evol.*, **12**(6), 1124-1131, 1995.
- Marzuki S., Noer A.S., Letrit P., Thyagarajan D., Kapsa R., Utthanaphol P. and Byrne E. Normal variants of human mitochondrial DNA and translation products: the building of a reference data base. *Hum. Genet.*, **88**(2), 139-145, 1991.
- Maslow A.H. A theory of human motivation. *Psychol Rev*, **50**, 370-396, 1943.
- Mathews D.H., Sabina J., Zuker M. and Turner D.H. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, **288**(5), 911-940, 1999.
- Matson S.W. and Kaiser-Rogers K.A. DNA helicases. *Annu. Rev. Biochem.*, **59**, 289-329, 1990.
- Matthews D.E., Hessler R.A., Denslow N.D., Edwards J.S. and O'Brien T.W. Protein composition of the bovine mitochondrial ribosome. *J. Biol. Chem.*, **257**, 8788-8794, 1982.
- Maynard Smith J. and Haigh J. The hitch-hiking effect of a favourable gene. *Genet. Res.*, **23**(1), 23-35, 1974.
- Maynard Smith J. Estimating selection by comparing synonymous and substitutional changes. *J. Mol. Evol.*, **39**(2), 123-128, 1994.
- McBrearty S. and Brooks A.S. The revolution that wasn't: a new interpretation of the origin of modern human behavior. *J. Hum. Evol.*, **39**(5), 453-563, 2000.
- McCulloch V., Seidel-Rogol B.L. and Shadel G.S. A human mitochondrial transcription factor is related to RNA adenine methyltransferases and binds S-adenosylmethionine. *Mol. Cell. Biol.*, **22**(4), 1116-1125, 2002.
- McDougall I., Brown F.H. and Fleagle J.G. Stratigraphic placement and age of modern humans from Kibish, Ethiopia. *Nature*, **433**(7027), 733-736, 2005.
- Meiklejohn C.D., Montooth K.L. and Rand D.M. Positive and negative selection on the mitochondrial genome. *Trends Genet.*, **23**(6), 259-263, 2007.
- Mellars P. Why did modern human populations disperse from Africa ca. 60,000 years ago? A new model. *Proc. Natl. Acad. Sci. U.S.A.*, **103**(25), 9381-9386, 2006.
- Melone M.A., Tessa A., Petrini S., Lus G., Sampaolo S., Di Fede G., Santorelli F.M. and Cotrufo R. Revelation of a new mitochondrial DNA mutation (G12147A) in a MELAS/MERFF phenotype. *Arch. Neurol.*, **61**(2), 269-272, 2004.
- Meltzer D.J. Peopling of North America. In: The quaternary period in the United States. Gillespie A., Porter S.C. and Atwater B. (editors). Elsevier Science, New York, USA, 539-563, 2004.
- Merriweather D.A. and Kaestle F.A. Mitochondrial recombination? (continued). *Science*, **285**(5429), 837, 1999.
- Merriwether D.A., Hodgson J.A., Friedlaender F.R., Allaby R., Cerchio S., Koki G. and Friedlaender J.S. Ancient mitochondrial M haplogroups identified in the Southwest Pacific. *Proc. Natl. Acad. Sci. U.S.A.*, **102**(37), 13034-13039, 2005.
- Meyer S., Weiss G. and Von Haeseler A. Pattern of nucleotide substitution and rate heterogeneity in the hypervariable regions I and II of human mtDNA. *Genetics*, **152**(3), 1103-1110, 1999.
- Mishmar D., Ruiz-Pesini E., Golik P., Macaulay V., Clark A.G., Hosseini S., Brandon M., Easley K., Chen E., Brown M.D., Sukernik R.I., Olckers A. and Wallace D.C. Natural selection shaped regional mtDNA variation in humans. *Proc. Natl. Acad. Sci. U.S.A.*, **100**(1), 171-176, 2003.
- Mishmar D., Ruiz-Pesini E., Mondragon-Palomino M., Procaccio V., Gaut B. and Wallace D.C. Adaptive selection of mitochondrial complex I subunits during primate radiation. *Gene*, **378**, 11-18, 2006.
- Mitchell A.L., Elson J.L., Howell N., Taylor R.W. and Turnbull D.M. Sequence variation in mitochondrial complex I genes: mutation or polymorphism? *J. Med. Genet.*, **43**(2), 175-179, 2006.

- Moilanen J.S. and Majamaa K. Phylogenetic network and physicochemical properties of nonsynonymous mutations in the protein-coding genes of human mitochondrial DNA. *Mol. Biol. Evol.*, **20**(8), 1195-1210, 2003.
- Moilanen J.S., Finnillä S. and Majamaa K. Lineage-specific selection in human mtDNA: lack of polymorphisms in a segment of MTND5 gene in haplogroup J. *Mol. Biol. Evol.*, **20**(12), 2132-2142, 2003.
- Montiel-Sosa F., Ruiz-Pesini E., Enríquez J.A., Marcuello A., Díez-Sánchez C., Montoya J., Wallace D.C. and López-Pérez M.J. Differences of sperm motility in mitochondrial DNA haplogroup U sublineages. *Gene*, **368**, 21-27, 2006.
- Montoya J., Christianson T., Levens D., Rabinowitz M. and Attardi G. Identification of initiation sites for heavy-strand and light-strand transcription in human mitochondrial DNA. *Proc. Natl. Acad. Sci. U.S.A.*, **79**(23), 7195-7199, 1982.
- Montoya J., Gaines G. and Attardi G. The pattern of transcription of the human mitochondrial rRNA genes reveals two overlapping transcription units. *Cell*, **34**, 151-159, 1983.
- Montoya J., Ojala D. and Attardi G. Distinctive features of the 5'-terminal sequences of the human mitochondrial mRNAs. *Nature*, **290**(5806), 465-470, 1981.
- Moraes C.T., Andreetta F., Bonilla E., Shanske S., DiMauro S. and Schon E.A. Replication-competent human mitochondrial DNA lacking the heavy-strand promoter region. *Mol. Cell. Biol.*, **11**(3), 1631-1637, 1991.
- Moraes C.T., Ciacci F., Bonilla E., Jansen C., Hirano M., Rao N., Lovelace R.E., Rowland L.P., Schon E.A. and DiMauro S. Two novel pathogenic mitochondrial DNA mutations affecting organelle number and protein synthesis. Is the tRNA^{Leu(UUR)} gene an etiologic hot spot? *J. Clin. Invest.*, **92**(6), 2906-2915, 1993.
- Moretti T., Koons B. and Budowle B. Enhancement of PCR amplification yield and specificity using AmpliTaq Gold™ DNA polymerase. *Biotechniques*, **25**(4), 716-722, 1998.
- Morral N., Bertranpetit J., Estivill X., Nunes V., Casals T., Giménez J., Reis A., Varon-Mateeva R., Macek M. Jr., Kalaydjieva L., Angelicheva D., Dancheva R., Romeo G., Russo M.P., Garnerone S., Restagno G., Ferrari M., Magnani C., Claustres M., Desgeorges M., Schwartz M., Schwarz M., Dallapiccola B., Novelli G., Ferec C., De Arce M., Nemeti M., Kere J., Anvret M., Dahl N. and Ladas L. The origin of the major cystic fibrosis mutation ($\Delta F508$) in European populations. *Nat. Genet.*, **7**, 169-175, 1994.
- Morton B.R. Chloroplast DNA codon use: evidence for selection at the *psb A* locus based on tRNA availability. *J. Mol. Evol.*, **37**(3), 273-280, 1993.
- Motulsky H.J. Intuitive biostatistics. Oxford University Press, Inc., New York, USA, 1995.
- Mouton C. Mitochondrial genome consensus sequence for the South African Khoi-San population. M.Sc. Thesis, North-West University, South Africa, 2003.
- Mullis K., Faloona F., Scharf S., Saiki R., Horn G. and Erlich H. Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harb. Symp. Quant. Biol.*, **51**(1), 263-273, 1986.
- Mullis K.B. and Faloona F.A. Specific synthesis of DNA *in vitro* via a polymerase-catalyzed chain reaction. *Meth. Enzymol.*, **155**, 335-350, 1987.
- Murphy W.I., Attardi B., Tu C. and Attardi G. Evidence for complete symmetrical transcription *in vivo* of mitochondrial DNA in HeLa cells. *J. Mol. Biol.*, **99**, 809-814, 1975.
- Musumeci O., Andreu A.L., Shanske S., Bresolin N., Comi G.P., Rothstein R., Schon E.A. and DiMauro S. Intragenic inversion of mtDNA: a new type of pathogenic mutation in a patient with mitochondrial myopathy. *Am. J. Hum. Genet.*, **66**(6), 1900-1904, 2000.
- Nachman M.W., Brown W.M., Stoneking M. and Aquadro C.F. Nonneutral mitochondrial DNA variation in humans and chimpanzees. *Genetics*, **142**, 953-963, 1996.
- Naini A.B., Lu J., Kaufmann P., Bernstein R.A., Mancuso M., Bonilla E., Hirano M. and DiMauro S. Novel mitochondrial DNA ND5 mutation in a patient with clinical features of MELAS and MERRF. *Arch. Neurol.*, **62**(3), 473-476, 2005.
- Nasidze I., Ling E.Y., Quinque D., Dupanloup I., Cordaux R., Rychkov S., Naumova O., Zhukova O., Sarraf-Zadegan N., Naderi G.A., Asgary S., Sardas S., Farhud D.D., Sarkisian T., Asadov C., Kerimov A. and Stoneking M. Mitochondrial DNA and Y-chromosome variation in the Caucasus. *Ann. Hum. Genet.*, **68**(3), 205-221, 2004.
- Nass M.M., Nass S. and Afzelius B.A. The general occurrence of mitochondrial DNA. *Exp. Cell Res.*, **37**, 516-539, 1965.
- Nei M. and Kumar S. Molecular evolution and phylogenetics. Oxford University Press, New York, USA, 2000.
- Nei M. and Miller J.C. A simple method for estimating average number of nucleotide substitutions within and between populations from restriction data. *Genetics*, **125**(4), 873-879, 1990.
- Nei M. Relative efficiencies of different tree-making methods for molecular data. In: Phylogenetic analysis of DNA sequences. Miyamoto M.M. and Cracraft J. (editors). Oxford University Press, New York, USA, 90-128, 1991.
- Newton C.R. and Graham A. Introduction to biotechniques. PCR. Second edition. BIOS Scientific Publishers Ltd., Oxford, UK, 1997.

- Niemi A.K. and Majamaa K. Mitochondrial DNA and ACTN3 genotypes in Finnish elite endurance and sprint athletes. *Eur. J. Hum. Genet.*, **13**(8), 965-969, 2005.
- Niemi A.K., Hervonen A., Hurme M., Karhunen P.J., Jylhä M. and Majamaa K. Mitochondrial DNA polymorphisms associated with longevity in a Finnish population. *Hum. Genet.*, **112**(1), 29-33, 2003.
- Noer A.S., Sudoya H., Lertrit P., Thyagarajan D., Utthanaphol P., Kapsa R., Byrne E. and Marzuki S. A tRNA^{Lys} mutation in the mtDNA is the causal genetic lesion underlying myoclonic epilepsy and ragged-red fiber (MERRF) syndrome. *Am. J. Hum. Genet.*, **49**(4), 715-722, 1991.
- Nordheim A., Pardue M.L., Lafer E.M., Möller A., Stollar B.D. and Rich A. Antibodies to left-handed Z-DNA bind to interband regions of *Drosophila* polytene chromosomes. *Nature*, **294**(5840), 417-422, 1981.
- O'Connell J.F. and Allen J. Dating the colonization of Sahul (Pleistocene Australia-New Guinea): a review of recent research. *J. Archaeol. Sci.*, **31**, 835-853, 2004.
- Ojala D., Montoya J. and Attardi G. tRNA punctuation model of RNA processing in human mitochondria. *Nature*, **290**(5806), 470-474, 1981.
- Olivieri A., Achilli A., Pala M., Battaglia V., Fornarino S., Al-Zahery N., Scozzari R., Cruciani F., Behar D.M., Dugoujon J.M., Coudray C., Santachiara-Benerecetti A.S., Semino O., Bandelt H.-J. and Torroni A. The mtDNA legacy of the Levantine early Upper Palaeolithic in Africa. *Science*, **314**(5806), 1767-1770, 2006.
- Oota H., Settheetham-Ishida W., Tiwawech D., Ishida T. and Stoneking M. Human mtDNA and Y-chromosome variation is correlated with matrilineal versus patrilineal residence. *Nat. Genet.*, **29**, 20-21, 2001.
- Oppenheimer S. Out of Africa's Eden. The peopling of the world. Jonathan Ball Publishers, Johannesburg, 2003.
- Ozawa T., Katsumata K., Hayakawa M., Tanaka M., Sugiyama S., Tanaka T., Itoyama S., Nunoda S. and Sekiguchi M. Genotype and phenotype of severe mitochondrial cardiomyopathy: a recipient of heart transplantation and the genetic control. *Biochem. Biophys. Res. Commun.*, **207**(2), 613-620, 1995.
- Ozawa T., Tanaka M., Ino H., Ohno K., Sano T., Wada Y., Yoneda M., Tanno Y., Miyatake T., Tanaka T., Itoyama S., Ikebe S., Hattori N. and Mizuno Y. Distinct clustering of point mutations in mitochondrial DNA among patients with mitochondrial encephalomyopathies and with Parkinson's disease. *Biochem. Biophys. Res. Commun.*, **176**(2), 938-946, 1991a.
- Ozawa T., Tanaka M., Sugiyama S., Ino H., Ohno K., Hattori N., Ohbayashi T., Ito T., Deguchi H., Kawamura K., Nakane Y. and Hashiba K. Patients with idiopathic cardiomyopathy belong to the same mitochondrial DNA gene family of Parkinson's disease and mitochondrial encephalomyopathy. *Biochem. Biophys. Res. Commun.*, **177**(1), 518-525, 1991b.
- Pääbo S. Mutational hot spots in the mitochondrial microcosm. *Am. J. Hum. Genet.*, **59**(3), 493-496, 1996.
- Page R.D.M. and Holmes E.C. Molecular evolution: a phylogenetic approach. Blackwell Science Ltd., Oxford, UK, 1998.
- Page R.D.M. TreeView: an application to display phylogenetic trees on personal computers. *Comput. Appl. Biosci.*, **12**(4), 357-358, 1996.
- Pakendorf B. and Stoneking M. Mitochondrial DNA and human evolution. *Annu Rev Genomics Hum Genet.*, **6**, 165-183, 2005.
- Palanichamy M.G., Sun C., Agrawal S., Bandelt H.-J., Kong Q.-P., Khan F., Wang C.-Y., Chaudhuri T.K., Palla V. and Zhang Y.-P. Phylogeny of mitochondrial DNA macrohaplogroup N in India, based on complete sequencing: implications for the peopling of South Asia. *Am. J. Hum. Genet.*, **75**(6), 966-978, 2004.
- Parisi M.A. and Clayton D.A. Similarity of human mitochondrial transcription factor 1 to high mobility group proteins. *Science*, **252**(5008), 965-969, 1991.
- Parsons T.J. and Irwin J.A. Questioning evidence for recombination in human mitochondrial DNA. *Science*, **288**(5473), 1931a, 2000.
- Parsons T.J., Muniec D.S., Sullivan K., Woodyatt N., Alliston-Greiner R., Wilson M.R., Berry D.L., Holland K.A., Weedn V.W., Gill P. and Holland M.M. A high observed substitution rate in the human mitochondrial DNA control region. *Nat. Genet.*, **15**(4), 363-368, 1997.
- Passarino G., Semino O., Quintana-Murci L., Excoffier L., Hammer M. and Santachiara-Benerecetti A.S. Different genetic components in the Ethiopian population, identified by mtDNA and Y-chromosome polymorphisms. *Am. J. Hum. Genet.*, **62**(2), 420-434, 1998.
- Penny D., Hendy M.D. and Steel M.A. Progress with methods for constructing evolutionary trees. *Trends Ecol. Evol. (Amst.)*, **7**(3), 73-79, 1992.
- Penny D., Steel M., Waddell P.J. and Hendy M.D. Improved analyses of human mtDNA sequences support a recent African origin for *Homo sapiens*. *Mol. Biol. Evol.*, **12**(5), 863-882, 1995.
- Pereira L., Macaulay V., Torroni A., Scozzari R., Prata M.J. and Amorim A. Prehistoric and historic traces in the mtDNA of Mozambique: insights into the Bantu expansions and the slave trade. *Ann. Hum. Genet.*, **65**(5), 439-458, 2001.

- Pereira L., Richards M., Alonso A., Albarrán C., Garcia O., Macaulay V. and Amorin A. Subdividing mtDNA haplogroup H based on coding-region polymorphisms – a study in Iberia. *Int. Congr. Ser.*, **1261**, 416-418, 2004.
- Pereira L., Richards M., Goios A., Alonso A., Albarrán C., Garcia O., Behar D.M., Gölge M., Hatina J., Al-Gazali L., Bradley D.G., Macaulay V. and Amorim A. High-resolution mtDNA evidence for the late-glacial resettlement of Europe from an Iberian refugium. *Genome Res.*, **15**(1), 19-24, 2005.
- Pesole G., Gissi C., De Chirico A. and Saccone C. Nucleotide substitution rate of mammalian mitochondrial genomes. *J. Mol. Evol.*, **48**(4), 427-434, 1999.
- Piercy R., Sullivan K.M., Benson N. and Gill P. The application of mitochondrial DNA typing to the study of white Caucasian genetic identification. *Int. J. Legal Med.*, **106**(2), 85-90, 1993.
- Pierson M.J., Martinez-Arias R., Holland B.R., Gemmell N.J., Hurler M.E. and Penny D. Deciphering past human population movements in Oceania: Provably optimal trees of 127 mtDNA genomes. *Mol. Biol. Evol.*, **23**(10), 1966-1975, 2006.
- Pilkington M.M., Wilder J.A., Mendez F.L., Cox M.P., Woerner A., Angui T., Kingan S., Mobasher Z., Batini C., Destro-Bisol G., Soodyall H., Strassmann B.I. and Hammer M.F. Contrasting signatures of population growth for mitochondrial DNA and Y chromosomes among human populations in Africa. *Mol. Biol. Evol.*, **25**(3), 517-525, 2008.
- Plaza S., Salas A., Calafell F., Corte-Real F., Bertranpetit J., Carracedo A. and Comas D. Insights into the western Bantu dispersal: mtDNA lineage analysis in Angola. *Hum. Genet.*, **115**(5), 439-447, 2004.
- Poetsch M., Wittig H., Krause D. and Lignitz E. The impact of mtDNA analysis between positions nt8306 and nt9021 for forensic casework. *Mitochondrion*, **3**(3), 133-137, 2003.
- Posada D. and Crandall K.A. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 13757-13762, 2001.
- Press W.H., Teukolsky S.A., Vetterling W.T. and Flannery B.P. Numerical recipes in C. The art of scientific computing. Cambridge University Press, Cambridge, USA, 1992.
- Prezant T.R., Agapian J.V., Bohlman M.C., Bu X., Oztas S., Qiu W.Q., Arnos K.S., Cortopassi G.A., Jaber L., Rotter J.I., Shohat M. and Fischel-Ghodsian N. Mitochondrial ribosomal RNA mutation associated with both antibiotic-induced and non-syndromic deafness. *Nat. Genet.*, **4**(3), 289-294, 1993.
- Pruitt K.D., Tatusova T. and Maglott D.R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**(Database issue), D61-65, 2007.
- Ptak S.E. and Przeworski M. Evidence for population growth in humans is confounded by fine-scale population structure. *Trends Genet.*, **18**(11), 559-563, 2002.
- Pulkes T., Eunson L., Patterson V., Siddiqui A., Wood N.W., Nelson I.P., Morgan-Hughes J.A. and Hanna M.G. The mitochondrial DNA G13513A transition in ND5 is associated with a LHON/MELAS overlap syndrome and may be a frequent cause of MELAS. *Ann. Neurol.*, **46**(6), 916-919, 1999.
- Pusch C.M., Broghammer M. and Blin N. Molecular phylogenetics employing modern and ancient DNA. *J. Appl. Genet.*, **44**(3), 269-290, 2003.
- Qian Y., Zhou X., Hu Y., Tong Y., Li R., Lu F., Yang H., Mo J.Q., Qu J. and Guan M.X. Clinical evaluation and mitochondrial DNA sequence analysis in three Chinese families with Leber's hereditary optic neuropathy. *Biochem. Biophys. Res. Commun.*, **332**(2), 614-621, 2005.
- Qin H., Wu W.B., Comeron J.M., Kreitman M. and Li W.-H. Intragenic spatial patterns of codon usage bias in prokaryotic and eukaryotic genomes. *Genetics*, **168**(4), 2245-2260, 2004.
- Quintana-Murci L., Quach H., Harmant C., Luca F., Massonnet B., Patin E., Sica L., Mouguiama-Daouda P., Comas D., Tzur S., Balanovsky O., Kidd K.K., Kidd J.R., Van der Veen L., Hombert J.M., Gessain A., Verdu P., Froment A., Bahuchet S., Heyer E., Dausset J., Salas A. and Behar D.M. Maternal traces of deep common ancestry and asymmetric gene flow between Pygmy hunter-gatherers and Bantu-speaking farmers. *Proc. Natl. Acad. Sci. U.S.A.*, **105**(5), 1596-1601, 2008.
- Quintana-Murci L., Semino O., Bandelt H.-J., Passarino G., McElreavey K. and Santachiara-Benerecetti A.S. Genetic evidence of an early exit of *Homo sapiens sapiens* from Africa through eastern Africa. *Nat. Genet.*, **23**(4), 437-441, 1999.
- Raby B.A., Klanderma B., Murphy A., Mazza S., Camargo C.A. Jr., Silverman E.K. and Weiss S.T. A common mitochondrial haplogroup is associated with elevated total serum IgE levels. *J. Allergy Clin. Immunol.*, **120**(2), 351-358, 2007.
- Rahman S., Taanman J.W., Cooper J.M., Nelson I., Hargreaves I., Meunier B., Hanna M.G., García J.J., Capaldi R.A., Lake B.D., Leonard J.V. and Schapira A.H. A missense mutation of cytochrome oxidase subunit II causes defective assembly and myopathy. *Am. J. Hum. Genet.*, **65**(4), 1030-1039, 1999.
- Rajkumar R., Banerjee J., Gunturi H.B., Trivedi R. and Kashyap V.K. Phylogeny and antiquity of M macrohaplogroup inferred from complete mtDNA sequence of Indian specific lineages. *BMC Evol. Biol.*, **5**(1), 26, 2005.
- Ramos-Onsins S.E. and Rozas J. Statistical properties of new neutrality tests against population growth. *Mol. Biol. Evol.*, **19**(12), 2092-2100, 2002.

- Rand D. Mitochondrial genomics flies high. *Trends Ecol. Evol. (Amst.)*, **16**(1), 2-4, 2001.
- Rando J.C., Cabrera V.M., Larruga J.M., Hernández M., González A.M., Pinto F. and Bandelt H.-J. Phylogeographic patterns of mtDNA reflecting the colonization of the Canary Islands. *Ann. Hum. Genet.*, **63**(5), 413-428, 1999.
- Rando J.C., Pinto F., González A.M., Hernández M., Larruga J.M., Cabrera V.M. and Bandelt H.-J. Mitochondrial DNA analysis of northwest African populations reveals genetic exchanges with European, near-eastern, and sub-Saharan populations. *Ann. Hum. Genet.*, **62**(6), 531-550, 1998.
- Ray N., Currat M. and Excoffier L. Intra-deme molecular diversity in spatially expanding populations. *Mol. Biol. Evol.*, **20**(1), 76-86, 2003.
- Redd A.J. and Stoneking M. Peopling of Sahul: mtDNA variation in aboriginal Australian and Papua New Guinean populations. *Am. J. Hum. Genet.*, **65**(3), 808-828, 1999.
- Redd A.J., Takezaki N., Sherry S.T., McGarvey S.T., Sofro A.S. and Stoneking M. Evolutionary history of the COII/tRNA^{Lys} intergenic 9 base pair deletion in human mitochondrial DNAs from the Pacific. *Mol. Biol. Evol.*, **12**(4), 604-615, 1995.
- Reich D.E. and Goldstein D.B. Genetic evidence for a Paleolithic human population expansion in Africa. *Proc. Natl. Acad. Sci. U.S.A.*, **95**(14), 8119-8123, 1998.
- Reid F.M., Vernham G.A. and Jacobs H.T. Complete mtDNA sequence of a patient in a maternal pedigree with sensorineural deafness. *Hum. Mol. Genet.*, **3**(8), 1435-1436, 1994.
- Reidla M., Kivisild T., Metspalu E., Kaldma K., Tambets K., Tolk H.-V., Parik J., Loogväli E.L., Derenko M., Malyarchuk B., Bermisheva M., Zhadanov S., Pennarun E., Gubina M., Golubenko M., Damba L., Fedorova S., Gusar V., Grechanina E., Mikerezi I., Moisan J.P., Chaventré A., Khusnutdinova E., Osipova L., Stepanov V., Voevoda M., Achilli A., Rengo C., Rickards O., De Stefano G.F., Papiha S., Beckman L., Janicijevic B., Rudan P., Anagnou N., Michalodimitrakis E., Koziel S., Usanga E., Geberhiwot T., Herrnstadt C., Howell N., Torroni A. and VILLEMS R. Origin and diffusion of mtDNA haplogroup X. *Am. J. Hum. Genet.*, **73**(5), 1178-1190, 2003.
- Relethford J.H. Genetic evidence and the modern human origins debate. *Heredity*, **100**(6), 555-563, 2008.
- Renfrew C. Archaeology, genetics and linguistic diversity. *Man (N.S.)*, **27**(3), 445-478, 1992.
- Richards M., Macaulay V., Hickey E., Vega E., Sykes B., Guida V., Rengo C., Sellitto D., Cruciani F., Kivisild T., Villems R., Thomas M., Rychkov S., Rychkov O., Rychkov Y., Gölge M., Dimitrov D., Hill E., Bradley D., Romano V., Calì F., Vona G., Demaine A., Papiha S., Triantaphyllidis C., Stefanescu G., Hatina J., Belledi M., Di Rienzo A., Novelletto A., Oppenheim A., Nørby S., Al-Zaheri N., Santachiara-Benerecetti S., Scozari R., Torroni A. and Bandelt H.-J. Tracing European founder lineages in the near Eastern mtDNA pool. *Am. J. Hum. Genet.*, **67**, 1251-1276, 2000.
- Rieder M.J., Taylor S.L., Tobe V.O. and Nickerson D.A. Automating the identification of DNA variations using quality-based fluorescence re-sequencing: analysis of the human mitochondrial genome. *Nucleic Acids Res.*, **26**(4), 967-973, 1998.
- Roche Diagnostics GmbH. Genome Sequencer 20 System^{TM1} First to the Finish Technical Manual, Mannheim, Germany, 2006b.
- Roche Diagnostics GmbH. Genome Sequencer FLX SystemTM First to the Finish Technical Manual, Mannheim, Germany, 2006a.
- Rogers A.R. and Harpending H. Population growth makes waves in the distribution of pairwise genetic differences. *Mol. Biol. Evol.*, **9**(3), 552-569, 1992.
- Rogers A.R. and Jorde L.B. Genetic evidence on modern human origins. *Hum. Biol.*, **67**, 1-36, 1995.
- Rogers A.R. Genetic evidence for a Pleistocene population explosion. *Evolution*, **49**(4), 608-615, 1995.
- Rokas A., Ladoukakis E. and Zouros E. Animal mitochondrial DNA recombination revisited. *Trends Ecol. Evol. (Amst.)*, **18**(8), 411-417, 2003.
- Roostalu U., Kutuev I., Loogväli E.L., Metspalu E., Tambets K., Reidla M., Khusnutdinova E.K., Usanga E., Kivisild T. and Villems R. Origin and expansion of haplogroup H, the dominant human mitochondrial DNA lineage in West Eurasia: the Near Eastern and Caucasian perspective. *Mol. Biol. Evol.*, **24**(2), 436-448, 2007.
- Rosa A., Brehm A., Kivisild T., Metspalu E. and Villems R. mtDNA profile of West Africa Guineans: towards a better understanding of the Senegambia region. *Ann. Hum. Genet.*, **68**(4), 340-352, 2004.
- Rose G., Passarino G., Carrieri G., Altomare K., Greco V., Bertolini S., Bonafè M., Franceschi C. and De Benedictis G. Paradoxes in longevity: sequence analysis of mtDNA haplogroup J in centenarians. *Eur. J. Hum. Genet.*, **9**(9), 701-707, 2001.
- Ross O.A., McCormack R., Curran M.D., Duguid R.A., Barnett Y.A., Rea I.M. and Middleton D. Mitochondrial DNA polymorphism: its role in longevity of the Irish population. *Exp. Gerontol.*, **36**(7), 1161-1178, 2001.
- Rossmann W., Tullo A., Potuschak T., Karwan R. and Sbisà E. Human mitochondrial tRNA processing. *J. Biol. Chem.*, **270**(21), 12885-12891, 1995.

¹ The Genome Sequencer 20 System is a trademark of 454 Life Sciences Corporation, Branford, CT, USA.

- Rozas J. and Rozas R. DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics*, **15**(2), 174-175, 1999.
- Rozas J., Sánchez-DelBarrio J.C., Messeguer X. and Rozas R. DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics*, **19**, 2496-2497, 2003.
- Ruiz-Pesini E. and Wallace D.C. Evidence for adaptive selection acting on the tRNA and rRNA genes of human mitochondrial DNA. *Hum. Mutat.*, **27**(11), 1072-1081, 2006.
- Ruiz-Pesini E. Center for Molecular and Mitochondrial Medicine and Genetics, University of California, Irvine, USA. Personal communication, 2003.
- Ruiz-Pesini E., Díez C., Lapeña A.C., Pérez-Martos A., Montoya J., Alvarez E., Arenas J. and López-Pérez M.J. Correlation of sperm motility with mitochondrial enzymatic activities. *Clin. Chem.*, **44**(8), 1616-1620, 1998.
- Ruiz-Pesini E., Lapeña A.C., Díez-Sánchez C., Pérez-Martos A., Montoya J., Alvarez E., Díaz M., Urriés A., Montoro L., López-Pérez M.J. and Enríquez J.A. Human mtDNA haplogroups associated with high or reduced spermatozoa motility. *Am. J. Hum. Genet.*, **67**(3), 682-696, 2000.
- Ruiz-Pesini E., Lott M.T., Procaccio V., Poole J.C., Brandon M.C., Mishmar D., Yi C., Kreuziger J., Baldi P. and Wallace D.C. An enhanced MITOMAP with a global mtDNA mutational phylogeny. *Nucleic Acids Res.*, **35**(Database issue), D823-828, 2007.
- Ruiz-Pesini E., Mishmar D., Brandon M., Procaccio V. and Wallace D.C. Effects of purifying and adaptive selection on regional variation in human mtDNA. *Science*, **303**(5655), 223-226, 2004.
- Ruppert V., Nolte D., Aschenbrenner T., Pankuweit S., Funck R. and Maisch B. Novel point mutations in the mitochondrial DNA detected in patients with dilated cardiomyopathy by screening the whole mitochondrial genome. *Biochem. Biophys. Res. Commun.*, **318**(2), 535-543, 2004.
- Ruvolo M., Zehr S., Von Dornum M., Pan D., Chang B. and Lin J. Mitochondrial COII sequences and modern human origins. *Mol. Biol. Evol.*, **10**, 1115-1135, 1993.
- Saccone C., De Giorgi C., Gissi C., Pesole G. and Reyes A. Evolutionary genomics in Metazoa: the mitochondrial DNA as a model system. *Gene*, **238**, 195-209, 1999.
- Saillard J., Forster P., Lynnerup N., Bandelt H.-J. and Nørby S. mtDNA variation among Greenland Eskimos: the edge of the Beringian expansion. *Am. J. Hum. Genet.*, **67**(3), 718-726, 2000.
- Saitou N. and Nei M. The neighbour-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406-425, 1987.
- Salas A., Carracedo A., Macaulay V., Richards M. and Bandelt H.-J. A practical guide to mitochondrial DNA error prevention in clinical, forensic, and population genetics. *Biochem. Biophys. Res. Commun.*, **335**, 891-899, 2005.
- Salas A., Richards M., De la Fe T., Lareu M.V., Sobrino B., Sánchez-Diz P., Macaulay V. and Carracedo A. The making of the African mtDNA landscape. *Am. J. Hum. Genet.*, **71**(5), 1082-1111, 2002.
- Salas A., Richards M., Lareu M.V., Scozzari R., Coppa A., Torroni A., Macaulay V. and Carracedo A. The African diaspora: mitochondrial DNA and the Atlantic slave trade. *Am. J. Hum. Genet.*, **74**(3), 454-465, 2004a.
- Salas A., Torroni A., Richards M., Quintana-Murci L., Hill C., Macaulay V. and Carracedo A. Letter to the Editor. Reply to Bortolini et al. *Am. J. Hum. Genet.*, **75**(3), 524-526, 2004b.
- Salem A.H., Badr F.M., Gaballah M.F. and Pääbo S. The genetics of traditional living: Y-chromosomal and mitochondrial lineages in the Sinai Peninsula. *Am. J. Hum. Genet.*, **59**(3), 741-743, 1996.
- Salemi M. and Vandamme A.-M. The phylogenetic handbook: a practical approach to DNA and protein phylogeny. Cambridge University Press, Cambridge, UK, 2003.
- Sambrook J. and Russell D.W. Molecular cloning. A laboratory manual. Third edition. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, USA, 2001.
- Sambrook J., Fritsch E.F. and Maniatis T. Molecular cloning. A laboratory manual. Second Edition. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, USA, 1989.
- Samuels M.L. Statistics for the life sciences. Dellen Publishing Company, San Francisco, USA, 1989.
- Sanderson M.J. and Wojciechowski M.F. Improved bootstrap confidence limits in large-scale phylogenies, with an example from Neo-Astragalus (Leguminosae). *Syst. Biol.*, **49**(4), 671-685, 2000.
- Sanderson M.J. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. *Mol. Biol. Evol.*, **19**(1), 101-109, 2002.
- Sanger F., Nicklen S. and Coulson A.R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.*, **74**, 5463-5467, 1977.
- Sans M. Admixture studies in Latin America: from the 20th to the 21st century. *Hum. Biol.*, **72**(1), 155-177, 2000.
- Santorelli F.M., Tanji K., Kulikova R., Shanske S., Vilarinho L., Hays A.P. and DiMauro S. Identification of a novel mutation in the mtDNA ND5 gene associated with MELAS. *Biochem. Biophys. Res. Commun.*, **238**(2), 326-328, 1997.
- Satoh M. and Kuroiwa T. Organization of multiple nucleoids and DNA molecules in mitochondria of a human cell. *Exp. Cell Res.*, **196**, 137-140, 1991.
- Schaefer A.M., McFarland R., Blakely E.L., He L., Whittaker R.G., Taylor R.W., Chinnery P.F. and Turnbull D.M. Prevalence of mitochondrial DNA disease in adults. *Ann. Neurol.*, **63**(1), 35-39, 2008.

- Schaefer A.M., Taylor R.W., Turnbull D.M. and Chinnery P.F. The epidemiology of mitochondrial disorders - past, present and future. *Biochim. Biophys. Acta*, **1659**(2-3), 115-120, 2004.
- Scheffler I.E. A century of mitochondrial research: achievements and perspectives. *Mitochondrion*, **1**, 3-31, 2001.
- Schierup M.H. and Hein J. Consequences of recombination on traditional phylogenetic analysis. *Genetics*, **156**, 879-891, 2000.
- Schlebusch C.M., Naidoo T. and Soodyall H. SNaPshot minisequencing to resolve mitochondrial macro-haplogroups found in Africa. *Electrophoresis*, **30**(21), 3657-3664, 2009.
- Schon E.A. Mitochondria. In: Mitochondrial DNA in human pathology. DiMauro S. and Wallace D.C. (editors). Raven Press Ltd., New York, USA, 1-7, 1993.
- Schriner S.E., Linford N.J., Martin G.M., Treuting P., Ogburn C.E., Emond M., Coskun P.E., Ladiges W., Wolf N., Van Remmen H., Wallace D.C. and Rabinovitch P.S. Extension of murine life span by overexpression of catalase targeted to mitochondria. *Science*, **308**(5730), 1909-1911, 2005.
- Schuelke M., Krude H., Finckh B., Mayatepek E., Janssen A., Schmelz M., Trefz F., Trijbels F. and Smeitink J. Septo-optic dysplasia associated with a new mitochondrial *cytochrome b* mutation. *Ann. Neurol.*, **51**(3), 388-392, 2002.
- Schurr T.G., Ballinger S.W., Gan Y.Y., Hodge J.A., Merriwether D.A., Lawrence D.N., Knowler W.C., Weiss K.M. and Wallace D.C. Amerindian mitochondrial DNAs have rare Asian mutations at high frequencies, suggesting they derived from four primary maternal lineages. *Am. J. Hum. Genet.*, **46**(3), 613-623, 1990.
- Schurr T.G., Sukernik R.I., Starikovskaya Y.B. and Wallace D.C. Mitochondrial DNA variation in Koryaks and Itel'men: population replacement in the Okhotsk Sea-Bering Sea region during the Neolithic. *Am. J. Phys. Anthropol.*, **108**(1), 1-39, 1999.
- Schuster S.C., Miller W., Ratan A., Tomsho L.P., Giardine B., Kasson L.R., Harris R.S., Petersen D.C., Zhao F., Qi J., Alkan C., Kidd J.M., Sun Y., Drautz D.I., Bouffard P., Muzny D.M., Reid J.G., Nazareth L.V., Wang Q., Burhans R., Riemer C., Wittekindt N.E., Moorjani P., Tindall E.A., Danko C.G., Teo W.S., Buboltz A.M., Zhang Z., Ma Q., Oosthuysen A., Steenkamp A.W., Oosthuisen H., Venter P., Gajewski J., Zhang Y., Pugh B.F., Makova K.D., Nekrutenko A., Mardis E.R., Patterson N., Pringle T.H., Chiaromonte F., Mullikin J.C., Eichler E.E., Hardison R.C., Gibbs R.A., Harkins T.T. and Hayes V.M. Complete Khoisan and Bantu genomes from southern Africa. *Nature*, **463**(7283), 943-947, 2010.
- Schwartz M. and Vissing J. New patterns of inheritance in mitochondrial disease. *Biochem. Biophys. Res. Commun.*, **310**(2), 247-251, 2003.
- Schwartz M. and Vissing J. Paternal inheritance of mitochondrial DNA. *N. Engl. J. Med.*, **347**, 576-580, 2002.
- Schwartzbach C.J. and Spemulli L.L. Bovine mitochondrial protein synthesis elongation factors. Identification and initial characterization of an elongation factor Tu-elongation factor Ts complex. *J. Biol. Chem.*, **264**(32), 19125-19131, 1989.
- Scott R.A., Fuku N., Onywera V.O., Boit M., Wilson R.H., Tanaka M., Goodwin W.H. and Pitsiladis Y.P. Mitochondrial haplogroups associated with elite Kenyan athlete status. *Med Sci Sports Exerc*, **41**(1), 123-128, 2009.
- Scozzari R., Cruciani F., Santolamazza P., Malaspina P., Torroni A., Sellitto D., Arredi B., Destro-Bisol G., De Stefano G., Rickards O., Martinez-Labarga C., Modiano D., Biondi G., Moral P., Olckers A., Wallace D.C. and Novelletto A. Combined use of biallelic and microsatellite Y-chromosome polymorphisms to infer affinities among African populations. *Am. J. Hum. Genet.*, **65**(3), 829-846, 1999.
- Searcy D.G. Metabolic integration during the evolutionary origin of mitochondria. *Cell Res.*, **13**(4), 229-238, 2003.
- Sekiya T., Kobayashi M., Seki T. and Koike K. Nucleotide sequence of a cloned fragment of rat mitochondrial DNA containing the replication origin. *Gene*, **11**(1-2), 53-62, 1980.
- Semino O., Santachiara-Benerecetti A.S., Falaschi F., Cavalli-Sforza L.L. and Underhill P.A. Ethiopians and Khoisan share the deepest clades of the human Y-chromosome phylogeny. *Am. J. Hum. Genet.*, **70**, 265-268, 2002.
- Senior A.E., Nadanaciva S. and Weber J. The molecular mechanism of ATP synthesis by F₁F₀-ATP synthase. *Biochim. Biophys. Acta*, **1553**, 188-211, 2002.
- Shadel G.S. and Clayton D.A. Mitochondrial DNA maintenance in vertebrates. *Annu. Rev. Biochem.*, **66**, 409-435, 1997.
- Shang J. and Clayton D.A. Human mitochondrial transcription termination exhibits RNA polymerase independence and biased bipolarity *in vitro*. *J. Biol. Chem.*, **269**(46), 29112-29120, 1994.
- Shanske A.L., Shanske S. and DiMauro S. The other human genome. *Arch Pediatr Adolesc Med*, **155**, 1210-1216, 2001.

- Shapiro B., Drummond A.J., Rambaut A., Wilson M.C., Matheus P.E., Sher A.V., Pybus O.G., Gilbert M.T., Barnes I., Binladen J., Willerslev E., Hansen A.J., Baryshnikov G.F., Burns J.A., Davydov S., Driver J.C., Froese D.G., Harington C.R., Keddle G., Kosintsev P., Kunz M.L., Martin L.D., Stephenson R.O., Storer J., Tedford R., Zimov S. and Cooper A. Rise and fall of the Beringian steppe bison. *Science*, **306**(5701), 1561-1565, 2004.
- Sharp P.M., Tuohy T.M.F. and Mosurski K.R. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res.*, **14**(13), 5125-5143, 1986.
- Shen P., Lavi T., Kivisild T., Chou V., Sengun D., Gefel D., Shpirer I., Woolf E., Hillel J., Feldman M.W. and Oefner P.J. Reconstruction of patrilineages and matrilineages of Samaritans and other Israeli populations from Y-chromosome and mitochondrial DNA sequence variation. *Hum. Mutat.*, **24**(3), 248-260, 2004.
- Sherry S.T., Rogers A.R., Harpending H., Soodyall H., Jenkins T. and Stoneking M. Mismatch distributions of mtDNA reveal recent human population expansions. *Hum. Biol.*, **66**(5), 761-775, 1994.
- Shields D.C., Sharp P.M., Higgins D.G. and Wright F. "Silent" sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. *Mol. Biol. Evol.*, **5**, 704-716, 1988.
- Shields G.F., Schmiechen A.M., Frazier B.L., Redd A., Voevoda M.I., Reed J.K. and Ward R.H. mtDNA sequences suggest a recent evolutionary divergence for Beringian and northern North American populations. *Am. J. Hum. Genet.*, **53**(3), 549-562, 1993.
- Shitara H., Hayashi J.I., Takahama S., Kaneda H. and Yonekawa H. Maternal inheritance of mouse mtDNA in interspecific hybrids: segregation of the leaked paternal mtDNA followed by the prevention of subsequent paternal leakage. *Genetics*, **148**(2), 851-857, 1998.
- Shoffner J.M., Brown M.D., Torroni A., Lott M.T., Cabell M.F., Mirra S.S., Beal M.F., Yang C.C., Gearing M., Salvo R., Watts R.L., Juncos J.L., Hansen L.A., Crain B.J., Fayad M., Reckord C.L. and Wallace D.C. Mitochondrial DNA variants observed in Alzheimer disease and Parkinson disease patients. *Genomics*, **17**(1), 171-184, 1993.
- Shriver M.D. and Kittles R.A. Genetic ancestry and the search for personalized genetic histories. *Nat. Rev. Genet.*, **5**(8), 611-618, 2004.
- Shriver M.D., Jin L., Ferrell R.E. and Deka R. Microsatellite data support an early population expansion in Africa. *Genome Res.*, **7**(6), 586-591, 1997.
- Sigurðardóttir S., Helgason A., Gulcher J.R., Stefánsson K. and Donnelly P. The mutation rate in the human mtDNA control region. *Am. J. Hum. Genet.*, **66**(5), 1599-1609, 2000.
- Silva W.A. Jr., Bonatto S.L., Holanda A.J., Ribeiro-dos-Santos A.K., Páixao B.M., Goldman G.H., Abe-Sandes K., Rodriguez-Delfin L., Barbosa M., Paço-Larson M.L., Petzl-Erler M.L., Valente V., Santos S.E. and Zago M.A. Mitochondrial genome diversity of Native Americans supports a single early entry of founder populations into America. *Am. J. Hum. Genet.*, **71**(1), 187-192, 2002.
- Simoni L., Calafell F., Pettener D., Bertranpetit J. and Barbujani G. Reconstruction of prehistory on the basis of genetic data. *Am. J. Hum. Genet.*, **66**(3), 1177-1179, 2000.
- Simonsen K.L., Churchill G.A. and Aquadro C.F. Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics*, **141**, 413-429, 1995.
- Slatkin M. and Hudson R.R. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics*, **129**, 555-562, 1991.
- Slatkin M. Gene flow in natural populations. *Annu Rev Ecol Syst*, **16**, 393-430, 1985.
- Smith B. Reading rock art and writing genetic history: regionalism, ethnicity and the rock art of southern Africa. In: The prehistory of Africa. Tracing the lineage of modern man. Soodyall H. (editor). Jonathan Ball Publishers, Johannesburg, South Africa, 76-96, 2006.
- Smith D.G., Malhi R.S., Eshleman J., Lorenz J.G. and Kaestle F.A. Distribution of mtDNA haplogroup X among Native North Americans. *Am. J. Phys. Anthropol.*, **110**(3), 271-284, 1999.
- Smith F.H., Falsetti A.B. and Donnelly S.M. Modern human origins. *Am. J. Phys. Anthropol.*, **32**(S10), 35-68, 1989.
- Sogin M.L., Morrison H.G., Huber J.A., Mark Welch D., Huse S.M., Neal P.R., Arrieta J.M. and Herndl G.J. Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc. Natl. Acad. Sci. U.S.A.*, **103**(32), 12115-12120, 2006.
- Soltis P.S. and Soltis D.E. Applying the bootstrap in phylogeny reconstruction. *Stat Sci*, **18**(2), 256-267, 2003.
- Soodyall H., Vigilant L., Hill A.V., Stoneking M. and Jenkins T. mtDNA control-region sequence variation suggests multiple independent origins of an "Asian-specific" 9-bp deletion in Sub-Saharan Africans. *Am. J. Hum. Genet.*, **58**, 595-608, 1996.
- Starikovskaya Y.B., Sukernik R.I., Schurr T.G., Kogelnik A.M. and Wallace D.C. mtDNA diversity in Chukchi and Siberian Eskimos: implications for the genetic history of ancient Beringia and the peopling of the New World. *Am. J. Hum. Genet.*, **63**(5), 1473-1491, 1998.
- Stephenson F.H. Calculations for molecular biology and biotechnology: a guide to mathematics in the laboratory. Academic Press, San Diego, California, USA, 2003.

- Sternberg D., Chatzoglou E., Laforêt P., Fayet G., Jardel C., Blondy P., Fardeau M., Amselem S., Eymard B. and Lombès A. Mitochondrial DNA transfer RNA gene sequence variations in patients with mitochondrial disorders. *Brain*, **124**(5), 984-994, 2001.
- Stevanovitch A., Gilles A., Bouzaid E., Kefi R., Paris F., Gayraud R.P., Spadoni J.L., El-Chenawi F. and Béraud-Colomb E. Mitochondrial DNA sequence diversity in a sedentary population from Egypt. *Ann. Hum. Genet.*, **68**(1), 23-39, 2004.
- Stewart J.B., Freyer C., Elson J.L., Wredenberg A., Cansu Z., Trifunovic A., Larsson N.-G. Strong purifying selection in transmission of mammalian mitochondrial DNA. *PLoS Biol.*, **6**(1), e10, 2008.
- Stone A.C. and Stoneking M. Ancient DNA from a pre-Columbian Amerindian population. *Am. J. Phys. Anthropol.*, **92**, 463-471, 1993.
- Stone A.C. and Stoneking M. mtDNA analysis of a prehistoric Oneota population: implications for the peopling of the New World. *Am. J. Hum. Genet.*, **62**(5), 1153-1170, 1998.
- Stone A.C., Starrs J.E. and Stoneking M. Mitochondrial DNA analysis of the presumptive remains of Jesse James. *J. Forensic Sci.*, **46**, 173-176, 2001.
- Stoneking M. and Soodyall H. Human evolution and the mitochondrial genome. *Curr. Opin. Genet. Dev.*, **6**(6), 731-736, 1996.
- Stoneking M. Genetic evidence for our recent African ancestry. In: The prehistory of Africa. Soodyall H. (editor). Jonathan Ball Publishers, Johannesburg, South Africa, 21-30, 2006.
- Stoneking M. Hypervariable sites in the mtDNA control region are mutational hotspots. *Am. J. Hum. Genet.*, **67**(4), 1029-1032, 2000.
- Stoneking M., Fontius J.J., Clifford S.L., Soodyall H., Arcot S.S., Saha N., Jenkins T., Tahir M.A., Deininger P.L. and Batzer M.A. *Alu* insertion polymorphisms and human evolution: evidence for a larger population size in Africa. *Genome Res.*, **7**(11), 1061-1071, 1997.
- Stoneking M., Jorde L.B., Bhatia K. and Wilson A.C. Geographic variation in human mitochondrial DNA from Papua New Guinea. *Genetics*, **124**(3), 717-733, 1990.
- Stoneking M., Sherry S.T., Redd A.J. and Vigilant L. New approaches to dating suggest a recent age for the human mtDNA ancestor. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, **337**(1280), 167-175, 1992.
- Straus L.G., Meltzer D.J. and Goebel T. Ice age Atlantis? Exploring the Solutrean-Clovis 'connection'. *World Archaeol.*, **37**(4), 507-532, 2005.
- Streisinger G., Okada Y., Emrich J., Newton J., Tsugita A., Terzaghi E. and Inouye M. Frameshift mutations and the genetic code. This paper is dedicated to Professor Theodosius Dobzhansky on the occasion of his 66th birthday. *Cold Spring Harb. Symp. Quant. Biol.*, **31**, 77-84, 1966.
- Stringer C. Modern human origins: progress and prospects. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, **357**, 563-579, 2002.
- Sueoka N. On the genetic basis of variation and heterogeneity of DNA base composition. *Proc. Natl. Acad. Sci. U.S.A.*, **48**, 582-592, 1962.
- Sutovsky P., Moreno R.D., Ramalho-Santos J., Dominko T., Simerly C. and Schatten G. Ubiquitinated sperm mitochondria, selective proteolysis, and the regulation of mitochondrial inheritance in mammalian embryos. *Biol. Reprod.*, **63**(2), 582-590, 2000.
- Sykes B. The seven daughters of Eve. Transworld Publishers, London, Great Britain, 2001.
- Taanman J.W. The mitochondrial genome: structure, transcription, translation and replication. *Biochim. Biophys. Acta*, **1410**(2), 103-123, 1999.
- Taivassalo T., Shoubridge E.A., Chen J., Kennaway N.G., DiMauro S., Arnold D.L. and Haller R.G. Aerobic conditioning in patients with mitochondrial myopathies: physiological, biochemical, and genetic effects. *Ann. Neurol.*, **50**(2), 133-141, 2001.
- Tajima F. Evolutionary relationship of DNA sequences in finite populations. *Genetics*, **105**(2), 437-460, 1983.
- Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585-595, 1989.
- Takahata N. Relaxed natural selection in human populations during the Pleistocene. *Jpn. J. Genet.*, **68**(6), 539-547, 1993.
- Tambets K., Rootsi S., Kivisild T., Help H., Serk P., Loogväli E.L., Tolk H.-V., Reidla M., Metspalu E., Pliss L., Balanovsky O., Pshenichnov A., Balanovska E., Gubina M., Zhadanov S., Osipova L., Damba L., Voevoda M., Kutuev I., Bermisheva M., Khusnutdinova E., Gusar V., Grechanina E., Parik J., Pennarun E., Richard C., Chaventre A., Moisan J.P., Barac L., Perićić M., Rudan P., Terzić R., Mikerezi I., Krumina A., Baumanis V., Koziel S., Rickards O., De Stefano G.F., Anagnou N., Pappa K.I., Michalodimitrakis E., Ferák V., Füredi S., Komel R., Beckman L. and Villems R. The western and eastern roots of the Saami - the story of genetic "outliers" told by mitochondrial DNA and Y chromosomes. *Am. J. Hum. Genet.*, **74**(4), 661-682, 2004.
- Tamm E., Kivisild T., Reidla M., Metspalu M., Smith D.G., Mulligan C.J., Bravi C.M., Rickards O., Martinez-Labarga C., Khusnutdinova E.K., Fedorova S.A., Golubenkov M.V., Stepanov V.A., Gubina M.A., Zhadanov S.I., Ossipova L.P., Damba L., Voevoda M.I., Dipierri J.E., Villems R. and Malhi R.S. Beringian standstill and spread of Native American founders. *PLoS ONE*, **2**, e829, 2007.
- Tamura K. and Nei M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.*, **10**(3), 512-526, 1993.

- Tamura K., Dudley J., Nei M. and Kumar S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol. Biol. Evol.*, **24**, 1596-1599, 2007.
- Tanaka M. and Ozawa T. Strand asymmetry in human mitochondrial DNA mutations. *Genomics*, **22**(2), 327-335, 1994.
- Tanaka M., Cabrera V.M., González A.M., Larruga J.M., Takeyasu T., Fuku N., Guo L.J., Hirose R., Fujita Y., Kurata M., Shinoda K., Umetsu K., Yamada Y., Oshida Y., Sato Y., Hattori N., Mizuno Y., Arai Y., Hirose N., Ohta S., Ogawa O., Tanaka Y., Kawamori R., Shamoto-Nagai M., Maruyama W., Shimokata H., Suzuki R. and Shimodaira H. Mitochondrial genome variation in eastern Asia and the peopling of Japan. *Genome Res.*, **14**(10a), 1832-1850, 2004.
- Tanaka M., Gong J.S., Zhang J., Yoneda M. and Yagi K. Mitochondrial genotype associated with longevity. *Lancet*, **351**(9097), 185-186, 1998.
- Tanaka M., Ino H., Ohno K., Ohbayashi T., Ikebe S., Sano T., Ichiki T., Kobayashi M., Wada Y. and Ozawa T. Mitochondrial DNA mutations in mitochondrial myopathy, encephalopathy, lactic acidosis, and stroke-like episodes (MELAS). *Biochem. Biophys. Res. Commun.*, **174**(2), 861-868, 1991.
- Tang J., Qi Y., Bao X.-H. and Wu X.-R. Mutational analysis of mitochondrial DNA of children with Rett syndrome. *Pediatr. Neurol.*, **17**(4), 327-330, 1997.
- Taylor R.W., Giordano C., Davidson M.M., d'Amati G., Bain H., Hayes C.M., Leonard H., Barron M.J., Casali C., Santorelli F.M., Hirano M., Lightowlers R.N., DiMauro S. and Turnbull D.M. A homoplasmic mitochondrial transfer ribonucleic acid mutation as a cause of maternally inherited hypertrophic cardiomyopathy. *J. Am. Coll. Cardiol.*, **41**(10), 1786-1796, 2003.
- Templeton A.R. Contingency tests of neutrality using intra/interspecific gene trees: the rejection of neutrality for the evolution of the mitochondrial cytochrome oxidase II gene in the hominoid primates. *Genetics*, **144**(3), 1263-1270, 1996.
- Templeton A.R. Out of Africa again and again. *Nature*, **416**(6876), 45-51, 2002.
- Tenesa A., Navarro P., Hayes B.J., Duffy D.L., Clarke G.M., Goddard M.E. and Visscher P.M. Recent human effective population size estimated from linkage disequilibrium. *Genome Res.*, **17**(4), 520-526, 2007.
- Thangaraj K., Chaubey G., Kivisild T., Reddy A.G., Singh V.K., Rasalkar A.A. and Singh L. Reconstructing the origin of Andaman Islanders. *Science*, **308**(5724), 996, 2005.
- Thomas A.W., Edwards A., Sherratt E.J., Majid A., Gagg J. and Alcolado J.C. Molecular scanning of candidate mitochondrial tRNA genes in type 2 (non-insulin dependent) diabetes mellitus. *J. Med. Genet.*, **33**(3), 253-256, 1996.
- Thompson J.D., Gibson T.J., Plewniak F., Jeanmougin F. and Higgins D.G. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.*, **25**(24), 4876-4882, 1997.
- Thyagarajan B., Padua R.A. and Campbell C. Mammalian mitochondria possess homologous DNA recombination activity. *J. Biol. Chem.*, **271**, 27536, 1996.
- Tiranti V., Rocchi M., DiDonato S. and Zeviani M. Cloning of human and rat cDNAs encoding the mitochondrial single-stranded DNA-binding protein (SSB). *Gene*, **126**(2), 219-225, 1993.
- Tishkoff S.A. and Verrelli B.C. Patterns of human genetic diversity: implications for human evolutionary history and disease. *Annu Rev Genomics Hum Genet*, **4**, 293-340, 2003.
- Tishkoff S.A. and Williams S.M. Genetic analysis of African populations: human evolution and complex disease. *Nat. Rev. Genet.*, **3**, 611-621, 2002.
- Tishkoff S.A., Dietzsch E., Speed W., Pakstis A.J., Kidd J.R., Cheung K., Bonn -Tamir B., Santachiara-Benerecetti A.S., Moral P., Krings M., P  abo S., Watson E., Risch N., Jenkins T. and Kidd K.K. Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science*, **271**(5254), 1380-1387, 1996.
- Tishkoff S.A., Gonder M.K., Henn B.M., Mortensen H., Knight A., Gignoux C., Fernandopulle N., Lema G., Nyambo T.B., Ramakrishnan U., Reed F.A. and Mountain J.L. History of click-speaking populations of Africa inferred from mtDNA and Y chromosome genetic variation. *Mol. Biol. Evol.*, **24**(10), 2180-2195, 2007.
- Tishkoff S.A., Reed F.A., Friedlaender F.R., Ehret C., Ranciaro A., Froment A., Hirbo J.B., Awomoyi A.A., Bodo J.M., Doumbo O., Ibrahim M., Juma A.T., Kotze M.J., Lema G., Moore J.H., Mortensen H., Nyambo T.B., Omar S.A., Powell K., Pretorius G.S., Smith M.W., Thera M.A., Wambebe C., Weber J.L. and Williams S.M. The genetic structure and history of Africans and African Americans. *Science*, **324**(5930), 1035-1044, 2009.
- Topper J.N. and Clayton D.A. Identification of transcriptional regulatory elements in human mitochondrial DNA by linker substitution analysis. *Mol. Cell. Biol.*, **9**(3), 1200-1211, 1989.
- Torroni A., Achilli A., Macaulay V., Richards M. and Bandelt H.-J. Harvesting the fruit of the human mtDNA tree. *Trends Genet.*, **22**(6), 339-345, 2006.
- Torroni A., Huoponen K., Francalacci P., Petrozzi M., Morelli L., Scozzari R., Obinu D., Savontaus M.-L. and Wallace D.C. Classification of European mtDNAs from an analysis of three European populations. *Genetics*, **144**, 1835-1850, 1996.

- Torroni A., Lott M.T., Cabell M.F., Chen Y.-S., Lavergne L. and Wallace D.C. mtDNA and the origin of Caucasians: identification of ancient Caucasian-specific haplogroups, one of which is prone to a recurrent somatic duplication in the D-loop region. *Am. J. Hum. Genet.*, **55**, 760-776, 1994a.
- Torroni A., Miller J.A., Moore L.G., Zamudio S., Zhuang J., Droma T. and Wallace D.C. Mitochondrial analysis in Tibet: implications for the origin of the Tibetan population and its adaptation to high altitude. *Am. J. Phys. Anthropol.*, **93**, 189-199, 1994b.
- Torroni A., Neel J.V., Barrantes R., Schurr T.G. and Wallace D.C. Mitochondrial DNA "clock" for the Amerinds and its implications for timing their entry into North America. *Proc. Natl. Acad. Sci. U.S.A.*, **91**, 1158-1162, 1994c.
- Torroni A., Petrozzi M., D'Urbano L., Sellitto D., Zeviani M., Carrara F., Carducci C., Leuzzi V., Carelli V., Barboni P., De Negri A. and Scozzari R. Haplotype and phylogenetic analyses suggest that one European-specific mtDNA background plays a role in the expression of Leber hereditary optic neuropathy by increasing the penetrance of the primary mutations 11778 and 14484. *Am. J. Hum. Genet.*, **60**(5), 1107-1121, 1997.
- Torroni A., Rengo C., Guida V., Cruciani F., Sellitto D., Coppa A., Calderon F.L., Simionati B., Valle G., Richards M., Macaulay V. and Scozzari, R. Do the four clades of the mtDNA haplogroup L2 evolve at different rates? *Am. J. Hum. Genet.*, **69**, 1348-1356, 2001.
- Torroni A., Schurr T.G., Cabell M.F., Brown M.D., Neel J.V., Larsen M., Smith D.G., Vullo C.M. and Wallace D.C. Asian affinities and continental radiation of the four founding Native American mtDNAs. *Am. J. Hum. Genet.*, **53**(3), 563-590, 1993a.
- Torroni A., Schurr T.G., Yang C.C., Szathmary E.J.E., Williams R.C., Schanfield M.S., Troup G.A., Knowler W.C., Lawrence D.N., Weiss K.M. and Wallace D.C. Native American mitochondrial DNA analysis indicates that the Amerind and the Nadene populations were founded by two independent migrations. *Genetics*, **130**(1), 153-162, 1992.
- Torroni A., Sukernik R.I., Schurr T.G., Starikovskaya Y.B., Cabell M.F., Crawford M.H., Comuzzie A.G. and Wallace D.C. mtDNA variation of aboriginal Siberians reveals distinct genetic affinities with Native Americans. *Am. J. Hum. Genet.*, **53**(3), 591-608, 1993b.
- Tourasse N.J. and Li W.-H. Selective constraints, amino acid composition, and the rate of protein evolution. *Mol. Biol. Evol.*, **17**(4), 656-664, 2000.
- Tuimala J. A primer to phylogenetic analysis using Phylip package. Fourth edition. Center for Scientific Computing, Espoo, Finland, 2005.
- Tully L.A., Parsons T.J., Steighner R.J., Holland M.M., Marino M.A. and Prenger V.L. A sensitive denaturing gradient-gel electrophoresis assay reveals a high frequency of heteroplasmy in hypervariable region 1 of the human mtDNA control region. *Am. J. Hum. Genet.*, **67**(2), 432-443, 2000.
- Tzen C.-Y., Thajeb P., Wu T.-Y. and Chen S.-C. MELAS with point mutations involving tRNA^{Leu} (A3243G) and tRNA^{Glu} (A14693g). *Muscle Nerve*, **28**(5), 575-581, 2003.
- Underhill P.A., Shen P., Lin A.A., Jin L., Passarino G., Yang W.H., Kauffman E., Bonn -Tamir B., Bertranpetit J., Francalacci P., Ibrahim M., Jenkins T., Kidd J.R., Mehdi S.Q., Seielstad M.T., Wells R.S., Piazza A., Davis R.W., Feldman M.W., Cavalli-Sforza L.L. and Oefner P.J. Y chromosome sequence variation and the history of human populations. *Nat. Genet.*, **26**(3), 358-361, 2000.
- Upholt W.B. and Dawid I.B. Mapping of mitochondrial DNA of individual sheep and goats: rapid evolution in the D loop region. *Cell*, **11**(3), 571-583, 1977.
- Van der Walt J.M., Nicodemus K.K., Martin E.R., Scott W.K., Nance M.A., Watts R.L., Hubble J.P., Haines J.L., Koller W.C., Lyons K., Pahwa R., Stern M.B., Colcher A., Hiner B.C., Jankovic J., Ondo W.G., Allen F.H. Jr., Goetz C.G., Small G.W., Mastaglia F., Stajich J.M., McLaurin A.C., Middleton L.T., Scott B.L., Schmechel D.E., Pericak-Vance M.A. and Vance J.M. Mitochondrial polymorphisms significantly reduce the risk of Parkinson disease. *Am. J. Hum. Genet.*, **72**(4), 804-811, 2003.
- Van Holst Pellekaan S.M., Ingman M., Roberts-Thomson J. and Harding R.M. Mitochondrial genomics identifies major haplogroups in Aboriginal Australians. *Am. J. Phys. Anthropol.*, **131**(2), 282-294, 2006.
- Van Oven M. and Kayser M. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum. Mutat.*, **30**(2), E386-E394, 2009.
- Vernesi C., Caramelli D., Dupanloup I., Bertorelle G., Lari M., Cappellini E., Moggi-Cecchi J., Chiarelli B., Castri L., Casoli A., Mallegni F., Lalueza-Fox C. and Barbujani G. The Etruscans: a population-genetic study. *Am. J. Hum. Genet.*, **74**(4), 694-704, 2004.
- Vigilant L., Pennington R., Harpending H., Kocher T.D. and Wilson A.C. Mitochondrial DNA sequences in single hairs from a southern African population. *Proc. Natl. Acad. Sci. U.S.A.*, **86**, 9350-9354, 1989.
- Vigilant L., Stoneking M., Harpending H., Hawkes K. and Wilson A.C. African populations and the evolution of human mitochondrial DNA. *Science*, **253**, 1503-1507, 1991.
- Voet D. and Voet J.G. Biochemistry. Second Edition. John Wiley and Sons, Inc., Somerset, New Jersey, USA., 1995.

- Vogelstein B. and Gillespie D. Preparative and analytical purification of DNA from agarose. *Proc. Natl. Acad. Sci. U.S.A.*, **76**(2), 615-619, 1979.
- Wainscoat J.S., Hill A.V., Boyce A.L., Flint J., Hernandez M., Thein S.L., Old J.M., Lynch J.R., Falusi A.G., Weatherall D.J. and Clegg J.B. Evolutionary relationships of human populations from an analysis of nuclear DNA polymorphisms. *Nature*, **319**(6053), 491-493, 1986.
- Wakeley J. Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA. *J. Mol. Evol.*, **37**(6), 613-623, 1993.
- Walberg M.W. and Clayton D.A. Sequence and properties of the human KB cell and mouse L cell D-loop regions of mitochondrial DNA. *Nucleic Acids Res.*, **9**(20), 5411-5421, 1981.
- Wall J.D. and Hudson R.R. Coalescent simulations and statistical tests of neutrality. *Mol. Biol. Evol.*, **18**(6), 1134-1135, 2001.
- Wallace D.C. Center for Molecular and Mitochondrial Medicine and Genetics, University of California, Irvine, USA. Personal communication, 2004.
- Wallace D.C. Diseases of the mitochondrial DNA. *Annu. Rev. Biochem.*, **61**, 1175-1212, 1992.
- Wallace D.C. Mitochondria as chi. *Genetics*, **179**(2), 727-735, 2008.
- Wallace D.C. Mitochondrial DNA sequence variation in human evolution and disease. *Proc. Natl. Acad. Sci. U.S.A.*, **91**, 8739-8746, 1994.
- Wallace D.C. Mitochondrial DNA variation in human evolution, degenerative disease, and aging. *Am. J. Hum. Genet.*, **57**, 201-223, 1995.
- Wallace D.C. The mitochondrial genome in human adaptive radiation and disease: on the road to therapeutics and performance enhancement. *Gene*, **354**, 169-180, 2005.
- Wallace D.C. Why do we still have a maternally inherited mitochondrial DNA? Insights from evolutionary medicine. *Annu Rev Biochem*, **76**, 781-821, 2007.
- Wallace D.C., Brown M.D. and Lott M.T. Mitochondrial DNA variation in human evolution and disease. *Gene*, **238**, 211-230, 1999.
- Wallace D.C., Singh G., Lott M.T., Hodge J.A., Schurr T.G., Lezza A.M. Elsas L.J. II and Nikoskelainen E.K. Mitochondrial DNA mutation associated with Leber's hereditary optic neuropathy. *Science*, **242**(4884), 1427-1430, 1988.
- Wallin I.E. On the nature of mitochondria. I. Observations on mitochondria staining methods applied to bacteria. II. Reactions of bacteria to chemical treatment. III. The demonstration of mitochondria by bacteriological methods. IV. A comparative study of the morphogenesis of root-nodule bacteria and chloroplasts. *Am. J. Anat.*, **30**, 203-229, 451-471, 1922.
- Wang J.C. DNA topoisomerases. *Annu. Rev. Biochem.*, **65**, 635-692, 1996.
- Ward R.H., Frazier B.L., Dew-Jager K. and Pääbo S. Extensive mitochondrial diversity within a single Amerindian tribe. *Proc. Natl. Acad. Sci. U.S.A.*, **88**(19), 8720-8724, 1991.
- Watson E., Bauer K., Aman R., Weiss G., Von Haeseler A. and Pääbo S. mtDNA sequence diversity in Africa. *Am. J. Hum. Genet.*, **59**(2), 437-444, 1996.
- Watson E., Forster P., Richards M. and Bandelt H.-J. Mitochondrial footprints of human expansions in Africa. *Am. J. Hum. Genet.*, **61**(3), 691-704, 1997.
- Watterson G.A. On the number of segregating sites in genetical models without recombination. *Theor Popul Biol*, **7**, 256-276, 1975.
- Wheeler D.A., Srinivasan M., Egholm M., Shen Y., Chen L., McGuire A., He W., Chen Y.-J., Makhijani V., Roth G.T., Gomes X., Tartaro K., Niazi F., Turcotte C.L., Irzyk G.P., Lupski J.R., Chinault C., Song X.-Z., Liu Y., Yuan Y., Nazareth L., Qin X., Muzny D.M., Margulies M., Weinstock G.M., Gibbs R.A. and Rothberg J.M. The complete genome of an individual by massively parallel DNA sequencing. *Nature*, **452**(7189), 872-876, 2008.
- White T.D., Asfaw B., DeGusta D., Gilbert H., Richards G.D., Suwa G. and Howell F.C. Pleistocene *Homo sapiens* from Middle Awash, Ethiopia. *Nature*, **423**(6941), 742-747, 2003.
- Wilder J.A., Kingan S.B., Mobasher Z., Pilkington M.M. and Hammer M.F. Global patterns of human mitochondrial DNA and Y-chromosome structure are not influenced by higher migration rates of females versus males. *Nat. Genet.*, **36**(10), 1122-1125, 2004a.
- Wilder J.A., Mobasher Z. and Hammer M.F. Genetic evidence for unequal effective population sizes of human females and males. *Mol. Biol. Evol.*, **21**, 2047-2057, 2004b.
- Wilkinson-Herbots H.M., Richards M.B., Forster P. and Sykes B.C. Site 73 in hypervariable region II of the human mitochondrial genome and the origin of European populations. *Ann. Hum. Genet.*, **60**(6), 499-508, 1996.
- Wilson A.C., Cann R.L., Carr S.M., George M., Gyllenstein U.B., Helm-Bychowski K.M., Higuchi R.G., Palumbi S.R., Prager E.M., Sage R.D. and Stoneking M. Mitochondrial DNA and two perspectives on evolutionary genetics. *Biol. J. Linn. Soc. Lond.*, **26**, 375-400, 1985.
- Wiuf C., Christensen T. and Hein J. A simulation study of the reliability of recombination detection methods. *Mol. Biol. Evol.*, **18**, 1929-1939, 2001.
- Wolpoff M.H., Hawks J. and Caspari R. Multiregional, not multiple origins. *Am. J. Phys. Anthropol.*, **112**, 129-136, 2000.

- Wolpoff M.H., Spuhler J.N., Smith F.H., Radovcic J., Pope G., Frayer D.W., Eckhardt R. and Clark G. Modern human origins. *Science*, **241**(4867), 772-774, 1988.
- Wong J.F., Ma D.P., Wilson R.K. and Roe B.A. DNA sequence of the *Xenopus laevis* mitochondrial heavy and light strand replication origins and flanking tRNA genes. *Nucleic Acids Res.*, **11**(14), 4977-4995, 1983.
- Wong T.W. and Clayton D.A. In vitro replication of human mitochondrial DNA: accurate initiation at the origin of light-strand synthesis. *Cell*, **42**(3), 951-958, 1985.
- Wooding S., Ostler C., Prasad B.V., Watkins W.S., Sung S., Bamshad M. and Jorde L.B. Directional migration in the Hindu castes: inferences from mitochondrial, autosomal and Y-chromosomal data. *Hum. Genet.*, **115**(3), 221-229, 2004.
- Woriat V.L., Bullard J.M., Ma L., Yokogawa T. and Spremulli L.L. Mechanistic studies of the translational elongation cycle in mammalian mitochondria. *Biochim. Biophys. Acta*, **1352**(1), 91-101, 1997.
- Wright F. The 'effective number of codons' used in a gene. *Gene*, **87**(1), 23-29, 1990.
- Wrischnik L.A., Higuchi R.G., Stoneking M., Erlich H.A., Arnheim N. and Wilson A.C. Length mutations in human mitochondrial DNA: direct sequencing of enzymatically amplified DNA. *Nucleic Acids Res.*, **15**(2), 529-542, 1987.
- Xia X. Maximizing transcription efficiency causes codon usage bias. *Genetics*, **144**(3), 1309-1320, 1996.
- Xia X. Mutation and selection on the anticodon of tRNA genes in vertebrate mitochondrial genomes. *Gene*, **345**(1), 13-20, 2005.
- Yakan M.Z. Almanac of African peoples and nations. Transaction Publishers, New Brunswick, New Jersey, USA, 1999.
- Yang Z. Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol. Evol. (Amst.)*, **11**(9), 367-372, 1996.
- Yang Z. and Nielsen R. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Mol. Biol. Evol.*, **25**(3), 568-579, 2008.
- Yao Y.-G., Macaulay V., Kivisild T., Zhang Y.-P. and Bandelt H.-J. To trust or not to trust an idiosyncratic mitochondrial data set. *Am. J. Hum. Genet.*, **72**, 1341-1346, 2003.
- Yoneda M., Tanno Y., Horai S., Ozawa T., Miyatake T. and Tsuji S. A common mitochondrial DNA mutation in the t-RNA^{Lys} of patients with myoclonus epilepsy associated with ragged-red fibers. *Biochem. Int.*, **21**(5), 789-796, 1990.
- Young J.H., Chang Y.-P.C., Kim J.D., Chretien J.P., Klag M.J., Levine M.A., Ruff C.B., Wang N.Y. and Chakravarti A. Differential susceptibility to hypertension is due to selection during the out-of-Africa expansion. *PLoS Genet.*, **1**(6), e82, 2005.
- Zuckerman S.H., Solus J.F., Gillespie F.P. and Eisenstadt J.M. Retention of both parental mitochondrial DNA species in mouse-Chinese hamster somatic cell hybrids. *Somat. Cell Mol. Genet.*, **10**(1), 85-92, 1984.
- Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**(13), 3406-3415, 2003.
- Zwieb C., Glotz C. and Brimacombe R. Secondary structure comparisons between small subunit ribosomal RNA molecules from six different species. *Nucleic Acids Res.*, **9**(15), 3621-3640, 1981.

7.2 ELECTRONIC REFERENCES

- Ancestry.com, 2010
<http://www.ancestry.com>
- Bibliography for MITOMAP's phylogenetic tree, 2007
<http://www.mitomap.org/treebib.html>
- Diagrammatic representation of mitochondrion: Brooklyn College, The City University of New York, 2004
<http://academic.brooklyn.cuny.edu/biology/bio4fv/page/mito.htm>
- Electron micrograph of mitochondrion: D.S. Friend, Brigham and Women's Hospital, 2005
<http://publications.nigms.nih.gov/insidethecell/chapter1.html>
- Family Tree DNA - Genealogy by Genetics Ltd., 2010
<http://www.familytreedna.com>
- GenBank®, 2007
<http://www.ncbi.nlm.nih.gov/Genbank/>
- Genetree software version 9.0, 2007
<http://www.stats.ox.ac.uk/~griff/software.html>
- Human Genome Diversity Project, 2010
<http://www.stanford.edu/group/morrinst/hgdp/faq.html>
- Innocence Project, 2010
<http://www.innocenceproject.org>

- International HapMap Project, 2010
<http://hapmap.ncbi.nlm.nih.gov>
- M'bokolo E. A hundred and fifty years after France abolished slavery. The impact of the slave trade on Africa. *Le Monde diplomatique*, April, 1998.
<http://mondediplo.com/1998/04/02africa>
- Mfold web server, 2008
<http://www.bioinfo.rpi.edu/applications/mfold>
- MITOMAP: A Human Mitochondrial Genome Database, 2006
<http://www.mitomap.org>
- MITOMAP: A Human Mitochondrial Genome Database, 2007
<http://www.mitomap.org>
- MITOMAP: A Human Mitochondrial Genome Database, 2008
<http://www.mitomap.org>
- mtDB - Human Mitochondrial Genome Database, 2008
<http://www.genpat.uu.se/mtDB>
- MyHeritage.com, 2010
<http://www.myheritage.com>
- NCBI's reference sequence (RefSeq) database, 2008
<http://www.ncbi.nlm.nih.gov/RefSeq>
- OligoCalc: Oligonucleotide Properties Calculator, 2005
<http://www.basic.northwestern.edu/biotools/OligoCalc.html>
- Oxford Ancestors Ltd., 2010
<http://www.oxfordancestors.com>
- Random Integer Generator, 2007
<http://www.random.org/integers/>
- Sequential mutational phylogenetic tree (Large mtDNA Tree), 2007
<http://www.mitomap.org/MitomapFigures.html>
- STATISTICA (data analysis software system) version 7.1, 2006
www.statsoft.com
- The Genographic Project, 2010
<https://genographic.nationalgeographic.com/genographic/index.html>
- United Nations Cartographic Section website, map number 4045, revision 4, January 2004
<http://www.un.org/Depts/Cartographic/map/profile/africa.pdf>

APPENDIX A

Sequence alterations observed in complete Khoi-San mitochondrial sequences compared to the rCRS

This appendix contains a list of the sequence alterations observed in a comparison between the 30 complete Khoi-San mitochondrial genome sequences and the rCRS. The nucleotide alterations observed in each of the 30 Khoi-San individuals are listed in the tables below, as follows:

- Table A.1, section A (KS_CGR_001 to KS_CGR_069),
- Table A.1, section B (KS_CGR_077 to KS_CGR_120), and
- Table A.1, section C (KS_CGR_121 to KS_CGR_176).

Nucleotide positions indicated in **red** represent previously reported polymorphisms and nucleotide positions presented in **blue** indicate novel polymorphisms. Detailed information for each alteration is provided in Appendix B.

Table A.1 Sequence alterations observed between the 30 complete Khoi-San mitochondrial genome sequences determined and the rCRS

Section A. Sequence alterations observed between KS_CGR_001 to KS_CGR_069 and the rCRS										
Nucl #	Khoi-San individual sample number (KS_CGR_###)									
	001	004	005	044	048	056	059	060	061	069
73	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G
146	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C
150										C - T
152	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C
182										C - T
183										A - G
189							A - G			
195	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C
198	C - T	C - T					C - T			C - T
199									T - C	
204							T - C			T - C
207							G - A			
247	G - A	G - A	G - A	G - A	G - A	G - A	G - A	G - A	G - A	
263										A - G
265			T - C	T - C						
309				C - T			INS CC			INS C

Table A.1 Continued...

Nucl #	Khoi-San individual sample number (KS_CGR_###)									
	001	004	005	044	048	056	059	060	061	069
310				T - C						
315	INS C	INS C	INS C		INS C	INS C	INS C	INS C	INS C	INS C
385										A - G
418										C - T
456		C - T	C - T							
494		DEL C	DEL C	DEL C	DEL C	DEL C		DEL C	DEL C	
514	DEL C	DEL C	DEL C	DEL C			DEL C			DEL C
515	DEL A	DEL A	DEL A	DEL A			DEL A			DEL A
573				INS CCCC						
593		T - C								
719		G - A	G - A		G - A	G - A		G - A	G - A	
750	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G
769	G - A	G - A	G - A	G - A	G - A	G - A	G - A	G - A	G - A	G - A
825	T - A	T - A	T - A	T - A	T - A	T - A	T - A	T - A	T - A	
850	T - C						T - C			
957					C - T			C - T		
1,018	G - A	G - A	G - A	G - A	G - A	G - A	G - A	G - A	G - A	G - A
1,048	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	
1,243	T - C						T - C		T - C	
1,386				T - C						
1,438	A - G						A - G			A - G
1,442										G - A
1,706										C - T
2,332										C - T
2,358										A - G
2,416										T - C
2,706	A - G			A - G			A - G			A - G
2,758	G - A	G - A	G - A	G - A	G - A	G - A	G - A	G - A	G - A	
2,836	C - A						C - A			
2,885	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	
3,438		G - A	G - A		G - A	G - A		G - A	G - A	
3,516	C - A	C - A	C - A	C - A	C - A	C - A	C - A	C - A	C - A	
3,594	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T
3,618					T - C	T - C		T - C		
3,666		G - A								
3,756		A - G	A - G	A - G	A - G	A - G		A - G	A - G	
3,981				A - G						
4,025				C - T						
4,044				A - G						
4,104	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G
4,158										A - G
4,197		C - T	C - T							

Table A.1 Continued...

Nucl #	Khoi-San individual sample number (KS_CGR_###)									
	001	004	005	044	048	056	059	060	061	069
4,232		T - C	T - C	T - C	T - C	T - C		T - C	T - C	
4,312	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	
4,370										T - C
4,541	G - A						G - A			
4,586	T - C						T - C			
4,767										A - G
4,769	A - G	A - G	A - G		A - G	A - G	A - G	A - G	A - G	A - G
4,907	T - C						T - C			
5,027										C - T
5,331										C - A
5,442	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	
5,460					G - A			G - A		
4,767										A - G
5,515				A - G						
5,811	A - G						A - G			
5,814										T - C
6,026										G - A
6,185	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	
6,266		A - G	A - G		A - G	A - G		A - G	A - G	
6,629										A - G
6,692						A - G				
6,713										C - T
6,815		T - C	T - C	T - C	T - C	T - C		T - C	T - C	
6,938	C - T						C - T			
7,028	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T
7,146	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	
7,154				A - G						
7,256	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T
7,257	A - G						A - G			
7,283					T - C	T - C		T - C		
7,521	G - A	G - A	G - A	G - A	G - A	G - A	G - A	G - A	G - A	G - A
7,624										T - A
7,673							A - G			
7,828		A - G								
8,080										C - T
8,113		C - A	C - A	C - A	C - A	C - A		C - A	C - A	
8,152		G - A	G - A	G - A	G - A	G - A		G - A	G - A	
8,206										G - A
8,222	T - C									
8,251		G - A	G - A	G - A	G - A	G - A	G - A	G - A	G - A	
8,387										G - A
8,468	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	

Table A.1 Continued...

Nucl #	Khoi-San individual sample number (KS_CGR_###)									
	001	004	005	044	048	056	059	060	061	069
8,655	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	
8,701	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G
8,715					T - C			T - C		
8,790						G - A				
8,860	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G
8,865				G - A						
8,911	T - C						T - C			
8,922	C - T									
8,970					C - T			C - T		
8,994	G - A			G - A			G - A			
9,042	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	
9,111			T - C		T - C			T - C		
9,136	A - G						A - G			
9,150		A - G	A - G							
9,221										A - G
9,347	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	
9,438		G - A								
9,540	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C
9,755		G - A	G - A	G - A	G - A	G - A		G - A	G - A	
9,818	C - T						C - T			
9,887						T - C				
9,932				G - A						
9,950									T - C	
10,084				T - C						
10,115										T - C
10,398	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G
10,499	A - G						A - G	A - G		
10,589	G - A	G - A	G - A	G - A	G - A	G - A	G - A	G - A	G - A	
10,664	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	
10,688	G - A	G - A	G - A	G - A	G - A	G - A	G - A	G - A	G - A	
10,810	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	
10,828										T - C
10,873	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C
10,876	A - G							A - G		
10,915	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	
10,920	C - T				C - T		C - T	C - T		
10,939	C - T							C - T		
11,087					T - C			T - C		
11,296	C - T						C - T			
11,299	T - C						T - C			
11,437		T - C	T - C							
11,653	A - G						A - G			

Table A.1 Continued...

Nucl #	Khoi-San individual sample number (KS_CGR_###)									
	001	004	005	044	048	056	059	060	061	069
11,719	G - A	G - A	G - A	G - A	G - A	G - A	G - A	G - A	G - A	G - A
11,854				T - C						
11,914	G - A	G - A	G - A	G - A	G - A	G - A	G - A	G - A	G - A	
11,944										T - C
12,007	G - A	G - A	G - A	G - A	G - A	G - A	G - A	G - A	G - A	
12,070	G - A						G - A			
12,121		T - C	T - C	T - C	T - C	T - C		T - C	T - C	
12,142									A - G	
12,235		T - C	T - C							
12,236										G - A
12,303					C - T			C - T		
12,705	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T
12,720		A - G	A - G	A - G	A - G	A - G		A - G	A - G	
12,798									C - T	
12,948										A - G
13,020	T - C						T - C			
13,105	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	
13,129		C - T	C - T							
13,276	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	
13,347			C - T							
13,386				T - C						
13,466				G - C						
13,506	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	
13,590	G - A						G - A			
13,650	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T
13,722										A - G
13,759		G - A	G - A		G - A	G - A		G - A	G - A	
13,819	T - C						T - C			
13,855									C - T	
13,924										C - T
13,928	G - C						G - C			
13,967									C - T	
14,020	T - C						T - C			
14,059										A - G
14,182	T - C						T - C			
14,211					C - T			C - T		
14,280						A - G				
14,302				T - C						
14,315					C - T	C - T		C - T		
14,371	T - C						T - C			
14,374	T - C						T - C			
14,560						G - A				

Table A.1 Continued...

Nucl #	Khoi-San individual sample number (KS_CGR_###)									
	001	004	005	044	048	056	059	060	061	069
14,659					C - T	C - T		C - T		
14,693				A - G						
14,766	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T
14,770				C - T						
15,106					G - A				G - A	
15,110										G - A
15,217										G - A
15,301										G - A
15,326	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G
15,449						T - C				
15,466		G - A	G - A	G - A	G - A	G - A		G - A	G - A	
15,519			T - C							
15,550		C - T								
15,766				A - G						
15,930		G - A	G - A	G - A	G - A	G - A		G - A	G - A	
15,941		T - C	T - C		T - C	T - C		T - C	T - C	
15,942				T - C						
15,951		A - G	A - G							
16,069				C - T						
16,093			T - C							
16,114										C - A
16,126				T - C						
16,129				G - A	G - A	G - A		G - A	G - A	G - A
16,153						G - A				G - A
16,166	A - C						A - C			
16,167		C - T								
16,169				C - T						
16,172	T - C						T - C			
16,182				A - C						
16,183				A - C						
16,187	C - T	C - T	C - T		C - T	C - T		C - T	C - T	
16,188				C - T						
16,189	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	
16,193							INS C			
16,209	T - C						T - C			
16,212				A - G						
16,213										G - A
16,214	C - T						C - T			
16,223	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T		C - T
16,230	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	
16,234		C - T	C - T						C - T	
16,239					C - T			C - T		

Table A.1 Continued...

Nucl #	Khoi-San individual sample number (KS_CGR_###)									
	001	004	005	044	048	056	059	060	061	069
16,242		C - T								
16,243		T - C	T - C	T - C	T - C	T - C		T - C	T - C	
16,264									C - T	
16,266									C - G	
16,278	C - T						C - T			C - T
16,294					C - T	C - T		C - T		
16,311	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C
16,362										T - C
16,390										G - A
16,474						G - T				
16,291	C - G			C - T			C - G			
16,519	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	
Section B. Sequence alterations observed between KS_CGR_077 to KS_CGR_120 and the rCRS										
Nucl #	Khoi-San individual sample number (KS_CGR_###)									
	077	083	084	090	091	092	105	111	117	120
73	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G
146	T - C	T - C	T - C		T - C	T - C	T - C	T - C	T - C	T - C
152	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C
189		A - G	A - G							A - G
195	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C
198	C - T	C - T	C - T		C - T	C - T	C - T	C - T	C - T	C - T
204			T - C							T - C
207		G - A	G - A							G - A
247	G - A	G - A	G - A		G - A	G - A	G - A	G - A	G - A	G - A
263				A - G						
309			INS C	INS C	INS C	INS C			INS C	INS C
315	INS C	INS C	INS C	INS C	INS C	INS C	INS C	INS C	INS C	INS C
456	C - T				C - T	C - T	C - T	C - T	C - T	
494	DEL C				DEL C	DEL C	DEL C	DEL C	DEL C	
514	DEL C	DEL C	DEL C	DEL C	DEL C	DEL C	DEL C	DEL C	DEL C	DEL C
515	DEL A	DEL A	DEL A	DEL A	DEL A	DEL A	DEL A	DEL A	DEL A	DEL A
719	G - A				G - A	G - A	G - A	G - A	G - A	
750	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G
769	G - A	G - A	G - A		G - A	G - A	G - A	G - A	G - A	G - A
825	T - A	T - A	T - A		T - A	T - A	T - A	T - A	T - A	T - A
850		T - C	T - C							T - C
921				T - C						
1,018	G - A	G - A	G - A		G - A	G - A	G - A	G - A	G - A	G - A
1,048	C - T	C - T	C - T		C - T	C - T	C - T	C - T	C - T	C - T
1,243		T - C	T - C							T - C
1,438		A - G	A - G	A - G						A - G
1,719				G - A						

Table A.1 Continued...

Nucl #	Khoi-San individual sample number (KS_CGR_###)									
	077	083	084	090	091	092	105	111	117	120
2,650			C - T							C - T
2,706		A - G	A - G	A - G						A - G
2,758	G - A	G - A	G - A		G - A	G - A	G - A	G - A	G - A	G - A
2,836		C - A	C - A							C - A
2,885	T - C	T - C	T - C		T - C	T - C	T - C	T - C	T - C	T - C
3,438	G - A				G - A	G - A	G - A	G - A	G - A	
3,498				C - G						
3,516	C - A	C - A	C - A		C - A	C - A	C - A	C - A	C - A	C - A
3,594	C - T	C - T	C - T		C - T	C - T	C - T	C - T	C - T	C - T
3,666	G - A									
3,756	A - G				A - G	A - G	A - G	A - G	A - G	
4,104	A - G	A - G	A - G		A - G	A - G	A - G	A - G	A - G	A - G
4,197	C - T				C - T	C - T	C - T	C - T	C - T	
4,232	T - C				T - C	T - C	T - C	T - C	T - C	
4,312	C - T	C - T	C - T		C - T	C - T	C - T	C - T	C - T	C - T
4,541		G - A	G - A							G - A
4,586		T - C	T - C							T - C
4,688				T - C						
4,769	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G
4,771										C - T
4,907		T - C	T - C							T - C
5,147				G - A						
5,442	T - C	T - C	T - C		T - C	T - C	T - C	T - C	T - C	T - C
5,811		A - G	A - G							A - G
6,185	T - C	T - C	T - C	T - C		T - C	T - C	T - C	T - C	T - C
6,266	A - G			A - G		A - G	A - G	A - G	A - G	
6,815	T - C				T - C	T - C	T - C	T - C	T - C	
6,938		C - T	C - T							C - T
7,028	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T
7,146	A - G	A - G	A - G		A - G	A - G	A - G	A - G	A - G	A - G
7,256	C - T	C - T	C - T		C - T	C - T	C - T	C - T	C - T	C - T
7,257		A - G	A - G							A - G
7,424				A - G						
7,521	G - A	G - A	G - A		G - A	G - A	G - A	G - A	G - A	G - A
7,673			A - G							A - G
7,828	A - G				A - G	A - G	A - G	A - G	A - G	
7,961	T - C									
7,967					C - T	C - T		C - T	C - T	
8,113	C - A				C - A	C - A		C - A	C - A	
8,152	G - A				G - A	G - A		G - A	G - A	
8,214							T - A			
8,223							T - G			

Table A.1 Continued...

Nucl #	Khoi-San individual sample number (KS_CGR_###)									
	077	083	084	090	091	092	105	111	117	120
8,251	G - A		G - A	G - A	G - A	G - A		G - A	G - A	G - A
8,468	C - T	C - T	C - T		C - T	C - T	C - T	C - T	C - T	C - T
8,618				T - C						
8,655	C - T	C - T	C - T		C - T	C - T	C - T	C - T	C - T	C - T
8,701	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G
8,860	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G
8,911		T - C	T - C							T - C
8,994		G - A	G - A							G - A
9,042	C - T	C - T	C - T		C - T	C - T	C - T	C - T	C - T	C - T
9,136		A - G	A - G							A - G
9,150	A - G				A - G	A - G	A - G	A - G	A - G	
9,347	A - G	A - G	A - G		A - G	A - G	A - G	A - G	A - G	A - G
9,438	G - A				G - A	G - A	G - A	G - A	G - A	
9,540	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C
9,755	G - A				G - A	G - A	G - A	G - A	G - A	
9,818		C - T	C - T							C - T
10,398	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G
10,499		A - G	A - G							A - G
10,589	G - A	G - A	G - A		G - A	G - A	G - A	G - A	G - A	G - A
10,664	C - T	C - T	C - T		C - T	C - T	C - T	C - T	C - T	C - T
10,688	G - A	G - A	G - A		G - A	G - A	G - A	G - A	G - A	G - A
10,810	T - C	T - C	T - C		T - C	T - C	T - C	T - C	T - C	T - C
10,873	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C
10,876		A - G	A - G							A - G
10,899				A - G						
10,915	T - C	T - C	T - C		T - C	T - C	T - C	T - C	T - C	T - C
10,920		C - T	C - T							C - T
10,939		C - T	C - T							C - T
11,296		C - T	C - T							C - T
11,299		T - C	T - C							T - C
11,404				A - G						
11,437	T - C				T - C	T - C	T - C	T - C	T - C	
11,653		A - G	A - G							A - G
11,719	G - A	G - A	G - A	G - A	G - A	G - A	G - A	G - A	G - A	G - A
11,914	G - A	G - A	G - A		G - A	G - A	G - A	G - A	G - A	G - A
12,007	G - A	G - A	G - A		G - A	G - A	G - A	G - A	G - A	G - A
12,070		G - A	G - A							G - A
12,121	T - C				T - C	T - C	T - C	T - C	T - C	
12,235	T - C				T - C	T - C	T - C	T - C	T - C	
12,469						A - G				
12,705	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T
12,720	A - G				A - G	A - G	A - G	A - G	A - G	

Table A.1 Continued...

Nucl #	Khoi-San individual sample number (KS_CGR_###)									
	077	083	084	090	091	092	105	111	117	120
13,020		T - C	T - C							T - C
13,105	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G
13,129	C - T				C - T	C - T	C - T	C - T	C - T	
13,260							T - C	T - C		
13,276	A - G	A - G	A - G		A - G	A - G	A - G	A - G	A - G	A - G
13,506	C - T	C - T	C - T		C - T	C - T	C - T	C - T	C - T	C - T
13,590		G - A	G - A							G - A
13,650	C - T	C - T	C - T		C - T	C - T	C - T	C - T	C - T	C - T
13,759	G - A				G - A	G - A	G - A	G - A	G - A	
13,819		T - C	T - C							T - C
13,886				T - C						
13,928		G - C	G - C							G - C
14,020		T - C	T - C							T - C
14,182		T - C	T - C							T - C
14,284				C - T						
14,371		T - C	T - C							T - C
14,374		T - C	T - C							T - C
14,766	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T
15,061				A - G						
15,208				A - G						
15,301				G - A						
15,326	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G
15,466	G - A				G - A	G - A	G - A	G - A	G - A	
15,519					T - C				T - C	
15,550	C - T				C - T	C - T	C - T	C - T	C - T	
15,930	G - A				G - A	G - A	G - A	G - A	G - A	
15,941	T - C				T - C	T - C	T - C	T - C	T - C	
15,951	A - G				A - G	A - G	A - G	A - G	A - G	
16,124				T - C						
16,166		A - C	A - C							A - C
16,167	C - T				C - T	C - T	C - T	C - T	C - T	
16,172		T - C	T - C							T - C
16,183				A - C						
16,187	C - T	C - T			C - T	C - T	C - T	C - T	C - T	
16,189	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C
16,193			INS C							
16,209		T - C	T - C							T - C
16,214		C - T	C - T							C - T
16,223	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T
16,230	A - G	A - G	A - G		A - G	A - G	A - G	A - G	A - G	A - G
16,234	C - T				C - T	C - T	C - T	C - T	C - T	
16,242	C - T				C - T	C - T	C - T	C - T	C - T	

Table A.1 Continued...

Nucl #	Khoi-San individual sample number (KS_CGR_###)									
	077	083	084	090	091	092	105	111	117	120
16,243	T - C				T - C	T - C	T - C	T - C	T - C	
16,278		C - T	C - T	C - T						C - T
16,291		C - G	C - G							C - G
16,304				T - C						
16,311	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C
16,497					A - G	A - G			A - G	
16,519	T - C	T - C	T - C		T - C	T - C	T - C	T - C	T - C	T - C
Section C. Sequence alterations observed between KS_CGR_121 to KS_CGR_176 and the rCRS										
Nucl #	Khoi-San individual sample number (KS_CGR_###)									
	121	127	140	143	157	164	172	173	175	176
73	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G
146	T - C	T - C			T - C	T - C		T - C		T - C
150			C - T	C - T		C - T		C - T	C - T	
152	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C
182						C - T		C - T		
183								A - G		
189					A - G					A - G
195	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C
198	C - T				C - T	C - T		C - T		C - T
199		T - C								
204						T - C		T - C		
207					G - A					G - A
247	G - A	G - A			G - A					G - A
263			A - G	A - G		A - G	A - G	A - G	A - G	
309							INS C	INS C		INS C
315	INS C	INS C	INS C	INS C	INS C	INS C	INS C	INS C	INS C	INS C
385								A - G		
418								C - T		
456	C - T									
494	DEL C	DEL C								
514	DEL C				DEL C		DEL C	DEL C		DEL C
515	DEL A				DEL A		DEL A	DEL A		DEL A
647			A - G	A - G					A - G	
709						G - A				
719	G - A	G - A								
750	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G
769	G - A	G - A			G - A	G - A		G - A		G - A
825	T - A	T - A			T - A					T - A
850					T - C					T - C
921							T - C			
1,018	G - A	G - A			G - A	G - A		G - A		G - A
1,048	C - T	C - T			C - T					C - T

Table A.1 Continued...

Nucl #	Khoi-San individual sample number (KS_CGR_###)									
	121	127	140	143	157	164	172	173	175	176
1,243		T - C			T - C					T - C
1,438			A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G
1,442						G - A		G - A		
1,706						C - T		C - T		
1,719							G - A			
2,332						C - T		C - T		
2,352			T - C	T - C					T - C	
2,358						A - G		A - G		
2,363										A - G
2,416						T - C		T - C		
2,706			A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G
2,758	G - A	G - A			G - A					G - A
2,836					C - A					C - A
2,885	T - C	T - C			T - C					T - C
3,438	G - A	G - A								
3,498							C - G			
3,516	C - A	C - A			C - A					C - A
3,594	C - T	C - T			C - T	C - T		C - T		C - T
3,606			A - G	A - G					A - G	
3,666	G - A									
3,756	A - G	A - G								
4,092			G - A	G - A					G - A	
4,104	A - G	A - G			A - G	A - G		A - G		A - G
4,158						A - G		A - G		
4,197	C - T									
4,232	T - C	T - C								
4,312	C - T	C - T			C - T					C - T
4,370						T - C		T - C		
4,541					G - A					G - A
4,586					T - C					T - C
4,688							T - C			
4,767						A - G		A - G		
4,769	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G
4,907					T - C					T - C
5,027						C - T		C - T		
5,147							G - A			
5,331						C - A		C - A		
5,442	T - C	T - C			T - C					T - C
5,471					G - A					
5,811					A - G					A - G
5,814						T - C		T - C		
5,865						T - A				

Table A.1 Continued...

Nucl #	Khoi-San individual sample number (KS_CGR_###)									
	121	127	140	143	157	164	172	173	175	176
5,986						T - C				
6,026								G - A		
6,185	T - C	T - C			T - C					T - C
6,266	A - G	A - G								
6,344						C - G				
6,614						T - C				
6,629								A - G		
6,713						C - T		C - T		
6,806						A - G				
6,815	T - C	T - C								
6,938					C - T					C - T
7,028	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T
7,146	A - G	A - G			A - G					A - G
7,256	C - T	C - T			C - T	C - T		C - T		C - T
7,257					A - G					A - G
7,424							A - G			
7,521	G - A	G - A			G - A	G - A		G - A		G - A
7,624						T - A		T - A		
7,828	A - G	A - G								
7,961	T - C									
8,080						C - T		C - T		
8,113	C - A	C - A								
8,152	G - A	G - A								
8,994					G - A					G - A
9,042	C - T	C - T			C - T					C - T
9,136					A - G					A - G
9,150	A - G									
9,221						A - G		A - G		
9,347	A - G	A - G			A - G					A - G
9,350						A - G				
9,438	G - A									
9,540	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C
9,755	G - A	G - A								
9,818					C - T					C - T
9,950		T - C								
10,098			G - T	G - T					G - T	
10,115						T - C		T - C		
10,398	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G
10,499					A - G					A - G
10,589	G - A	G - A			G - A					G - A
10,664	C - T	C - T			C - T					C - T
10,688	G - A	G - A			G - A					G - A

Table A.1 Continued...

Nucl #	Khoi-San individual sample number (KS_CGR_###)									
	121	127	140	143	157	164	172	173	175	176
10,810	T - C	T - C			T - C					T - C
10,819			A - G	A - G					A - G	
10,828								T - C		
10,873	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C
10,876					A - G					A - G
10,899							A - G			
10,915	T - C	T - C			T - C					T - C
10,920					C - T					C - T
10,939					C - T					C - T
11,296					C - T					C - T
11,299					T - C					T - C
11,404							A - G			
11,437	T - C									
11,653					A - G					A - G
11,719	G - A	G - A	G - A	G - A	G - A	G - A	G - A	G - A	G - A	G - A
11,728								C - T		
11,914	G - A	G - A			G - A					G - A
11,944						T - C		T - C		
12,007	G - A	G - A			G - A					G - A
12,070					G - A					G - A
12,121	T - C	T - C								
12,142		A - G								
12,235	T - C									
12,236						G - A		G - A		
12,630			G - A	G - A						
12,705	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T
12,720	A - G	A - G								
12,798		C - T								
12,804		T - C								
12,948						A - G		A - G		
13,020					T - C					T - C
13,105	A - G	A - G			A - G		A - G			A - G
13,129	C - T									
13,276	A - G	A - G			A - G					A - G
13,506	C - T	C - T			C - T					C - T
13,590					G - A	G - A		G - A		G - A
13,650	C - T	C - T			C - T	C - T		C - T		C - T
13,722								A - G		
13,759	G - A	G - A								
13,819					T - C					T - C
13,855		C - T								
13,886							T - C			

Table A.1 Continued...

Nucl #	Khoi-San individual sample number (KS_CGR_###)									
	121	127	140	143	157	164	172	173	175	176
13,924								C - T		
13,928					G - C					G - C
13,966						A - G				
13,967		C - T								
14,020					T - C					T - C
14,059						A - G		A - G		
14,182					T - C					T - C
14,212			T - C	T - C					T - C	
14,284							C - T			
14,371					T - C					T - C
14,374					T - C					T - C
14,407						C - T				
14,569										G - A
14,766	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T
14,905			G - A	G - A					G - A	
15,061							A - G			
15,110						G - A		G - A		
15,208							A - G			
15,217						G - A		G - A		
15,301			G - A	G - A		G - A	G - A	G - A	G - A	
15,326	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G
15,466	G - A	G - A								
15,550	C - T									
15,930	G - A	G - A								
15,941	T - C	T - C								
15,951	A - G									
16,093		T - C								
16,114						C - A		C - A		
16,124							T - C			
16,129		G - A				G - A		G - A		
16,153								G - A		
16,166					A - C					A - C
16,167	C - T									
16,172			T - C	T - C	T - C				T - C	T - C
16,183							A - C			
16,187	C - T	C - T			C - T					C - T
16,189	T - C	T - C	T - C	T - C	T - C		T - C		T - C	T - C
16,209					T - C					T - C
16,213						G - A		G - A		
16,214					C - T					C - T
16,223	C - T		C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T
16,230	A - G	A - G			A - G					A - G

Table A.1 Continued...

Nucl #	Khoi-San individual sample number (KS_CGR_###)									
	121	127	140	143	157	164	172	173	175	176
16,234	C - T	C - T								
16,242	C - T									
16,243	T - C	T - C								
16,266		C - G								
16,278					C - T	C - T	C - T	C - T		C - T
16,291					C - G					C - G
16,304							T - C			
16,311	T - C	T - C			T - C		T - C	T - C		T - C
16,320			C - T	C - T					C - T	
16,354						C - T				
16,362								T - C		
16,390						G - A		G - A		
16,519	T - C	T - C	T - C	T - C	T - C				T - C	T - C

Nucleotides are represented by the capital letters, with the rCRS nucleotide indicated first, followed by the observed change. Nucleotide numbers are based on rCRS numbering (Andrews *et al.*, 1999). The numbers indicated in **red** represent previously reported polymorphisms and the numbers presented in **blue** indicate previously unreported polymorphisms. A = adenine; C = cytosine; G = guanine; T = thymine; INS = insertion; DEL = deletion; Nucl # = nucleotide number; KS_CGR_### = three digit sample number given to each Khoi-San individual.

APPENDIX B

Synopsis of molecular information for sequence alterations observed between Khoi-San sequences and the rCRS

This appendix contains information regarding the locus, alteration type, codon position, amino acid change, and report status of the sequence alterations observed in comparisons between the 30 complete Khoi-San mitochondrial genome sequences determined in the current investigation and the rCRS which were listed in Appendix A. The information available regarding the nucleotide alterations observed in each of the 30 Khoi-San individuals is listed in Table B.1. The nucleotide positions presented in **red** indicate alterations that have been reported previously, while nucleotide positions illustrated in **blue** represent alterations that are novel.

Table B.1 Information regarding sequence alterations observed between the 30 complete Khoi-San mitochondrial genome sequences determined and the rCRS

Alteration		Locus	Alteration type	Codon position	Amino acid change	Reference
73	A – G	MT-DLOOP	TS	non-coding	---	Aquadro and Greenberg, 1983
146	T – C	MT-DLOOP	TS	non-coding	---	Aquadro and Greenberg, 1983
150	C – T	MT-DLOOP	TS	non-coding	---	Aquadro and Greenberg, 1983
152	T – C	MT-DLOOP	TS	non-coding	---	Aquadro and Greenberg, 1983
182	C – T	MT-DLOOP	TS	non-coding	---	Aquadro and Greenberg, 1983
183	A – G	MT-DLOOP	TS	non-coding	---	Liu <i>et al.</i> , 2001
189	A – G	MT-DLOOP	TS	non-coding	---	Aquadro and Greenberg, 1983
195	T – C	MT-DLOOP	TS	non-coding	---	Aquadro and Greenberg, 1983
198	C – T	MT-DLOOP	TS	non-coding	---	Vigilant <i>et al.</i> , 1989
199	T – C	MT-DLOOP	TS	non-coding	---	Vigilant <i>et al.</i> , 1989
204	T – C	MT-DLOOP	TS	non-coding	---	Graven <i>et al.</i> , 1995
207	G – A	MT-DLOOP	TS	non-coding	---	Ozawa <i>et al.</i> , 1991a
247	G – A	MT-DLOOP	TS	non-coding	---	Aquadro and Greenberg, 1983
263	A – G	MT-DLOOP	TS	non-coding	---	Aquadro and Greenberg, 1983
265	T – C	MT-DLOOP	TS	non-coding	---	Barbosa <i>et al.</i> , 2008
309	C – T	MT-DLOOP	TS	non-coding	---	Brandstätter <i>et al.</i> , 2004
309	INS C	MT-DLOOP	INS	non-coding	---	Vigilant <i>et al.</i> , 1989
309	INS CC	MT-DLOOP	INS	non-coding	---	Vigilant <i>et al.</i> , 1989
310	T – C	MT-DLOOP	TS	non-coding	---	Budowle <i>et al.</i> , 1999
315	INS C	MT-DLOOP	INS	non-coding	---	Andrews <i>et al.</i> , 1999

Table B.1 Continued...

Alteration		Locus	Alteration type	Codon position	Amino acid change	Reference
385	A – G	MT-DLOOP	TS	non-coding	---	Redd <i>et al.</i> , 1995
418	C – T	MT-DLOOP	TS	non-coding	---	Mishmar <i>et al.</i> , 2003
456	C – T	MT-DLOOP	TS	non-coding	---	Aquadro and Greenberg, 1983
494	DEL C	MT-DLOOP	DEL	non-coding	---	Mishmar <i>et al.</i> , 2003
514	DEL C	MT-DLOOP	DEL	non-coding	---	Ozawa <i>et al.</i> , 1991a
515	DEL A	MT-DLOOP	DEL	non-coding	---	Ozawa <i>et al.</i> , 1991a
573	INS CCCC	MT-DLOOP	INS	non-coding	---	Torroni <i>et al.</i> , 1994a
593	T – C	MT-TF	TS	non-coding	---	Herrnstadt <i>et al.</i> , 2002a
647	A – G	MT-TF	TS	non-coding	---	Current investigation
709	G – A	MT-RNR1	TS	non-coding	---	Ozawa <i>et al.</i> , 1991a
719	G – A	MT-RNR1	TS	non-coding	---	Kivisild <i>et al.</i> , 2006
750	A – G	MT-RNR1	TS	non-coding	---	Andrews <i>et al.</i> , 1999
769	G – A	MT-RNR1	TS	non-coding	---	Greenberg <i>et al.</i> , 1983
825	T – A	MT-RNR1	TV	non-coding	---	Prezant <i>et al.</i> , 1993
850	T – C	MT-RNR1	TS	non-coding	---	Crimi <i>et al.</i> , 2002
921	T – C	MT-RNR1	TS	non-coding	---	Liu <i>et al.</i> , 2001
957	C – T	MT-RNR1	TS	non-coding	---	Current investigation
1,018	G – A	MT-RNR1	TS	non-coding	---	Prezant <i>et al.</i> , 1993
1,048	C – T	MT-RNR1	TS	non-coding	---	Moraes <i>et al.</i> , 1993
1,243	T – C	MT-RNR1	TS	non-coding	---	Jones <i>et al.</i> , 2001
1,386	T – C	MT-RNR1	TS	non-coding	---	Current investigation
1,438	A – G	MT-RNR1	TS	non-coding	---	Ozawa <i>et al.</i> , 1991b
1,442	G – A	MT-RNR1	TS	non-coding	---	Herrnstadt <i>et al.</i> , 2002a
1,706	C – T	MT-RNR2	TS	non-coding	---	Herrnstadt <i>et al.</i> , 2002a
1,719	G – A	MT-RNR2	TS	non-coding	---	Noer <i>et al.</i> , 1991
2,332	C – T	MT-RNR2	TS	non-coding	---	Herrnstadt <i>et al.</i> , 2002a
2,352	T – C	MT-RNR2	TS	non-coding	---	Chen <i>et al.</i> , 1995
2,358	A – G	MT-RNR2	TS	non-coding	---	Herrnstadt <i>et al.</i> , 2002a
2,363	A – G	MT-RNR2	TS	non-coding	---	Current investigation
2,416	T – C	MT-RNR2	TS	non-coding	---	Herrnstadt <i>et al.</i> , 2002a
2,650	C – T	MT-RNR2	TS	non-coding	---	Tang <i>et al.</i> , 1997
2,706	A – G	MT-RNR2	TS	non-coding	---	Tanaka <i>et al.</i> , 1991
2,758	G – A	MT-RNR2	TS	non-coding	---	Moraes <i>et al.</i> , 1993
2,836	C – A	MT-RNR2	TV	non-coding	---	Ingman <i>et al.</i> , 2000
2,885	T – C	MT-RNR2	TS	non-coding	---	Moraes <i>et al.</i> , 1993
3,438	G – A	MT-ND1	TS	coding: 3	SYN	Thomas <i>et al.</i> , 1996
3,498	C – G	MT-ND1	TV	coding: 3	SYN	Kivisild <i>et al.</i> , 2006
3,516	C – A	MT-ND1	TV	coding: 3	SYN	Herrnstadt <i>et al.</i> , 2002a
3,591	G – A	MT-ND1	TS	coding: 3	SYN	Schwartz and Vissing, 2002
3,594	C – T	MT-ND1	TS	coding: 3	SYN	Chen <i>et al.</i> , 1995
3,606	A – G	MT-ND1	TS	coding: 3	SYN	Herrnstadt <i>et al.</i> , 2002a
3,618	T – C	MT-ND1	TS	coding: 3	SYN	Herrnstadt <i>et al.</i> , 2002a

Table B.1 Continued...

Alteration		Locus	Alteration type	Codon position	Amino acid change	Reference
3,666	G – A	MT-ND1	TS	coding: 3	SYN	Herrnstadt <i>et al.</i> , 2002a
3,756	A – G	MT-ND1	TS	coding: 3	SYN	Mishmar <i>et al.</i> , 2003
3,981	A – G	MT-ND1	TS	coding: 3	SYN	Mishmar <i>et al.</i> , 2003
4,025	C – T	MT-ND1	TS	coding: 2	NON-SYN: Thr → Met	Huoponen <i>et al.</i> , 1993
4,044	A – G	MT-ND1	TS	coding: 3	SYN	Mishmar <i>et al.</i> , 2003
4,092	G – A	MT-ND1	TS	coding: 3	SYN	Finnilä <i>et al.</i> , 2001a
4,104	A – G	MT-ND1	TS	coding: 3	SYN	Kösel <i>et al.</i> , 1998
4,158	A – G	MT-ND1	TS	coding: 3	SYN	Chen <i>et al.</i> , 2000
4,197	C – T	MT-ND1	TS	coding: 3	SYN	Kivisild <i>et al.</i> , 2006
4,232	T – C	MT-ND1	TS	coding: 2	NON-SYN: Ile → Thr	Herrnstadt <i>et al.</i> , 2002a
4,312	C – T	MT-TI	TS	non-coding	---	Moraes <i>et al.</i> , 1993
4,370	T – C	MT-TQ	TS	non-coding	---	Herrnstadt <i>et al.</i> , 2002a
4,541	G – A	MT-ND2	TS	coding: 3	SYN	Ingman <i>et al.</i> , 2000
4,586	T – C	MT-ND2	TS	coding: 3	SYN	Ingman <i>et al.</i> , 2000
4,688	T – C	MT-ND2	TS	coding: 3	SYN	Ingman <i>et al.</i> , 2000
4,767	A – G	MT-ND2	TS	coding: 1	NON-SYN: Met → Val	Herrnstadt <i>et al.</i> , 2002a
4,769	A – G	MT-ND2	TS	coding: 3	SYN	Cann <i>et al.</i> , 1987
4,771	C – T	MT-ND2	TS	coding: 2	NON-SYN: Ala → Val	Current investigation
4,907	T – C	MT-ND2	TS	coding: 3	SYN	Ingman <i>et al.</i> , 2000
5,027	C – T	MT-ND2	TS	coding: 3	SYN	Herrnstadt <i>et al.</i> , 2002a
5,147	G – A	MT-ND2	TS	coding: 3	SYN	Kobayashi <i>et al.</i> , 1991
5,331	C – A	MT-ND2	TV	coding: 1	NON-SYN: Leu → Ile	Herrnstadt <i>et al.</i> , 2002a
5,442	T – C	MT-ND2	TS	coding: 1	NON-SYN: Phe → Leu	Ozawa <i>et al.</i> , 1991a
5,460	G – A	MT-ND2	TS	coding: 1	NON-SYN: Ala → Thr	Lin <i>et al.</i> , 1992
5,471	G – A	MT-ND2	TS	coding: 3	SYN	Howell <i>et al.</i> , 1995
5,515	A – G	MT-TW	TS	non-coding	---	Current investigation
5,811	A – G	MT-TC	TS	non-coding	---	Houshmand <i>et al.</i> , 1994
5,814	T – C	MT-TC	TS	non-coding	---	Larsson and Clayton, 1995
5,865	T – A	MT-TY	TV	non-coding	---	Current investigation
5,986	T – C	MT-CO1	TS	coding: 2	NON-SYN: Val → Ala	Current investigation
6,026	G – A	MT-CO1	TS	coding: 3	SYN	Ingman <i>et al.</i> , 2000
6,185	T – C	MT-CO1	TS	coding: 3	SYN	Herrnstadt <i>et al.</i> , 2002a
6,266	A – G	MT-CO1	TS	coding: 3	SYN	Herrnstadt <i>et al.</i> , 2002a
6,344	C – G	MT-CO1	TV	coding: 3	NON-SYN: Ile → Met	Current investigation
6,614	T – C	MT-CO1	TS	coding: 3	SYN	Tanaka <i>et al.</i> , 2004

Table B.1 Continued...

Alteration		Locus	Alteration type	Codon position	Amino acid change	Reference
6,629	A – G	MT-CO1	TS	coding: 3	SYN	Herrnstadt <i>et al.</i> , 2002a
6,692	A – G	MT-CO1	TS	coding: 3	SYN	Kivisild <i>et al.</i> , 2006
6,713	C – T	MT-CO1	TS	coding: 3	SYN	Herrnstadt <i>et al.</i> , 2002a
6,806	A – G	MT-CO1	TS	coding: 3	SYN	Rajkumar <i>et al.</i> , 2005
6,815	T – C	MT-CO1	TS	coding: 3	SYN	Mishmar <i>et al.</i> , 2003
6,938	C – T	MT-CO1	TS	coding: 3	SYN	Finnilä <i>et al.</i> , 2001b
7,028	C – T	MT-CO1	TS	coding: 3	SYN	Cann <i>et al.</i> , 1987
7,146	A – G	MT-CO1	TS	coding: 1	NON-SYN: Thr → Ala	Herrnstadt <i>et al.</i> , 2002a
7,154	A – G	MT-CO1	TS	coding: 3	SYN	Mishmar <i>et al.</i> , 2003
7,256	C – T	MT-CO1	TS	coding: 3	SYN	Silva <i>et al.</i> , 2002
7,257	A – G	MT-CO1	TS	coding: 1	NON-SYN: Ile → Val	Ingman <i>et al.</i> , 2000
7,283	T – C	MT-CO1	TS	coding: 3	SYN	Moilanen <i>et al.</i> , 2003
7,424	A – G	MT-CO1	TS	coding: 3	SYN	Marzuki <i>et al.</i> , 1991
7,521	G – A	MT-TS1	TS	non-coding	---	Prezant <i>et al.</i> , 1993
7,624	T – A	MT-CO2	TV	coding: 3	SYN	Herrnstadt <i>et al.</i> , 2002a
7,673	A – G	MT-CO2	TS	coding: 1	NON-SYN: Ile → Val	Ozawa <i>et al.</i> , 1991a
7,828	A – G	MT-CO2	TS	coding: 3	SYN	Ruvolo <i>et al.</i> , 1993
7,961	T – C	MT-CO2	TS	coding: 1	SYN	Tanaka <i>et al.</i> , 2004
7,967	C – T	MT-CO2	TS	coding: 1	SYN	Kivisild <i>et al.</i> , 2006
8,080	C – T	MT-CO2	TS	coding: 3	SYN	Herrnstadt <i>et al.</i> , 2002a
8,113	C – A	MT-CO2	TV	coding: 3	SYN	Ruvolo <i>et al.</i> , 1993
8,152	G – A	MT-CO2	TS	coding: 3	SYN	Ruvolo <i>et al.</i> , 1993
8,206	G – A	MT-CO2	TS	coding: 3	SYN	Ozawa <i>et al.</i> , 1995
8,214	T – A	MT-CO2	TV	coding: 2	NON-SYN: Val → Asp	Current investigation
8,222	T – C	MT-CO2	TS	coding: 1	SYN	Ruppert <i>et al.</i> , 2004
8,223	T – G	MT-CO2	TV	coding: 2	NON-SYN: Leu → Trp	Current investigation
8,251	G – A	MT-CO2	TS	coding: 3	SYN	Ruvolo <i>et al.</i> , 1993
8,387	G – A	ATPase8	TS	coding: 1	NON-SYN: Val → Met	Keightley <i>et al.</i> , 1996
8,468	C – T	ATPase8	TS	coding: 1	SYN	Moraes <i>et al.</i> , 1993
8,503	T – C	ATPase8	TS	coding: 3	SYN	Merriwether <i>et al.</i> , 2005
8,618	T – C	ATPase6	TS	coding: 2	NON-SYN: Ile → Thr	Poetsch <i>et al.</i> , 2003
8,655	C – T	ATPase6	TS	coding: 3	SYN	Ingman <i>et al.</i> , 2000
8,701	A – G	ATPase6	TS	coding: 1	NON-SYN: Thr → Ala	Wallace <i>et al.</i> , 1988
8,715	T – C	ATPase6	TS	coding: 3	SYN	Herrnstadt <i>et al.</i> , 2002a
8,790	G – A	ATPase6	TS	coding: 3	SYN	Taylor <i>et al.</i> , 2003

Table B.1 Continued...

Alteration		Locus	Alteration type	Codon position	Amino acid change	Reference
8,860	A – G	ATPase6	TS	coding: 1	NON-SYN: Thr → Ala	Brown <i>et al.</i> , 1992
8,865	G – A	ATPase6	TS	coding: 3	SYN	Herrnstadt <i>et al.</i> , 2002a
8,911	T – C	ATPase6	TS	coding: 1	SYN	Ingman <i>et al.</i> , 2000
8,922	C – T	ATPase6	TS	coding: 3	SYN	Current investigation
8,970	C – T	ATPase6	TS	coding: 3	SYN	Current investigation
8,994	G – A	ATPase6	TS	coding: 3	SYN	Ingman <i>et al.</i> , 2000
9,042	C – T	ATPase6	TS	coding: 3	SYN	Ingman <i>et al.</i> , 2000
9,111	T – C	ATPase6	TS	coding: 3	SYN	Herrnstadt <i>et al.</i> , 2002a
9,136	A – G	ATPase6	TS	coding: 1	NON-SYN: Ile → Val	Ingman <i>et al.</i> , 2000
9,150	A – G	ATPase6	TS	coding: 3	SYN	Rieder <i>et al.</i> , 1998
9,221	A – G	MT-CO3	TS	coding: 3	SYN	Ingman <i>et al.</i> , 2000
9,347	A – G	MT-CO3	TS	coding: 3	SYN	Ingman <i>et al.</i> , 2000
9,350	A – G	MT-CO3	TS	coding: 3	SYN	Kivisild <i>et al.</i> , 2006
9,438	G – A	MT-CO3	TS	coding: 1	NON-SYN: Gly → Ser	Johns and Neufeld, 1993
9,540	T – C	MT-CO3	TS	coding: 1	SYN	Wallace <i>et al.</i> , 1988
9,755	G – A	MT-CO3	TS	coding: 3	SYN	Ingman <i>et al.</i> , 2000
9,818	C – T	MT-CO3	TS	coding: 3	SYN	Ingman <i>et al.</i> , 2000
9,887	T – C	MT-CO3	TS	coding: 3	SYN	Kivisild <i>et al.</i> , 2006
9,932	G – A	MT-CO3	TS	coding: 3	SYN	Ozawa <i>et al.</i> , 1991a
9,950	T – C	MT-CO3	TS	coding: 3	SYN	Tanaka and Ozawa, 1994
10,084	T – C	MT-ND3	TS	coding: 2	NON-SYN: Ile → Thr	Clark <i>et al.</i> , 1999
10,098	G – T	MT-ND3	TV	coding: 1	NON-SYN: Ala → Ser	Current investigation
10,115	T – C	MT-ND3	TS	coding: 3	SYN	Kösel <i>et al.</i> , 1998
10,398	A – G	MT-ND3	TS	coding: 1	NON-SYN: Thr → Ala	Cann <i>et al.</i> , 1987
10,499	A – G	MT-ND4L	TS	coding: 3	SYN	Rieder <i>et al.</i> , 1998
10,589	G – A	MT-ND4L	TS	coding: 3	SYN	Howell <i>et al.</i> , 1995
10,664	C – T	MT-ND4L	TS	coding: 3	SYN	Ingman <i>et al.</i> , 2000
10,688	G – A	MT-ND4L	TS	coding: 3	SYN	Ingman <i>et al.</i> , 2000
10,810	T – C	MT-ND4	TS	coding: 3	SYN	Ingman <i>et al.</i> , 2000
10,819	A – G	MT-ND4	TS	coding: 3	SYN	Wallace <i>et al.</i> , 1988
10,828	T – C	MT-ND4	TS	coding: 3	SYN	Mishmar <i>et al.</i> , 2003
10,873	T – C	MT-ND4	TS	coding: 3	SYN	Wallace <i>et al.</i> , 1988
10,876	A – G	MT-ND4	TS	coding: 3	SYN	Finnilä <i>et al.</i> , 2001b
10,899	A – G	MT-ND4	TS	coding: 2	NON-SYN: Asn → Ser	Herrnstadt <i>et al.</i> , 2002a
10,915	T – C	MT-ND4	TS	coding: 3	SYN	Yoneda <i>et al.</i> , 1990
10,920	C – T	MT-ND4	TS	coding: 2	NON-SYN: Pro → Leu	Ingman <i>et al.</i> , 2000

Table B.1 Continued...

Alteration		Locus	Alteration type	Codon position	Amino acid change	Reference
10,939	C – T	MT-ND4	TS	coding: 3	SYN	Ingman <i>et al.</i> , 2000
11,087	T – C	MT-ND4	TS	coding: 1	NON-SYN: Phe → Leu	Herrnstadt <i>et al.</i> , 2002a
11,296	C – T	MT-ND4	TS	coding: 3	SYN	Schwartz and Vissing, 2002
11,299	T – C	MT-ND4	TS	coding: 3	SYN	Johns, 1991
11,404	A – G	MT-ND4	TS	coding: 3	SYN	Tanaka <i>et al.</i> , 2004
11,437	T – C	MT-ND4	TS	coding: 3	SYN	Kong <i>et al.</i> , 2003
11,653	A – G	MT-ND4	TS	coding: 3	SYN	Horváth <i>et al.</i> , 2002
11,719	G – A	MT-ND4	TS	coding: 3	SYN	Wallace <i>et al.</i> , 1988
11,728	C – T	MT-ND4	TS	coding: 3	SYN	Current investigation
11,854	T – C	MT-ND4	TS	coding: 3	SYN	Mishmar <i>et al.</i> , 2003
11,914	G – A	MT-ND4	TS	coding: 3	SYN	Tanaka and Ozawa, 1994
11,944	T – C	MT-ND4	TS	coding: 3	SYN	Ozawa <i>et al.</i> , 1991a
12,007	G – A	MT-ND4	TS	coding: 3	SYN	Mackey and Howell, 1992
12,070	G – A	MT-ND4	TS	coding: 3	SYN	Brown <i>et al.</i> , 2001
12,121	T – C	MT-ND4	TS	coding: 3	SYN	Mishmar <i>et al.</i> , 2003
12,142	A – G	MT-TH	TS	non-coding	---	Kivisild <i>et al.</i> , 2006
12,235	T – C	MT-TS2	TS	non-coding	---	Ingman and Gyllensten, 2003
12,236	G – A	MT-TS2	TS	non-coding	---	Prezant <i>et al.</i> , 1993
12,303	C – T	MT-TL2	TS	non-coding	---	Kivisild <i>et al.</i> , 2006
12,469	A – G	MT-ND5	TS	coding: 1	NON-SYN: Ile → Val	Tanaka <i>et al.</i> , 2004
12,630	G – A	MT-ND5	TS	coding: 3	SYN	Finnilä <i>et al.</i> , 2001b
12,705	C – T	MT-ND5	TS	coding: 3	SYN	Wallace <i>et al.</i> , 1988
12,720	A – G	MT-ND5	TS	coding: 3	SYN	Brown <i>et al.</i> , 2001
12,798	C – T	MT-ND5	TS	coding: 3	SYN	Kivisild <i>et al.</i> , 2006
12,804	T – C	MT-ND5	TS	coding: 3	SYN	Tanaka <i>et al.</i> , 2004
12,948	A – G	MT-ND5	TS	coding: 3	SYN	Herrnstadt <i>et al.</i> , 2002a
13,020	T – C	MT-ND5	TS	coding: 3	SYN	Finnilä <i>et al.</i> , 2001b
13,105	A – G	MT-ND5	TS	coding: 1	NON-SYN: Ile → Val	Prezant <i>et al.</i> , 1993
13,129	C – T	MT-ND5	TS	coding: 1	NON-SYN: Pro → Ser	Herrnstadt <i>et al.</i> , 2002a
13,260	T – C	MT-ND5	TS	coding: 3	SYN	Tanaka <i>et al.</i> , 2004
13,276	A – G	MT-ND5	TS	coding: 1	NON-SYN: Met → Val	Ingman <i>et al.</i> , 2000
13,347	C – T	MT-ND5	TS	coding: 3	SYN	Current investigation
13,386	T – C	MT-ND5	TS	coding: 3	SYN	Current investigation
13,466	G – C	MT-ND5	TS	coding: 2	NON-SYN: Ser → Thr	Current investigation
13,506	C – T	MT-ND5	TS	coding: 3	SYN	Ingman <i>et al.</i> , 2000
13,590	G – A	MT-ND5	TS	coding: 3	SYN	Tanaka and Ozawa, 1994
13,650	C – T	MT-ND5	TS	coding: 3	SYN	Liu <i>et al.</i> , 2001

Table B.1 Continued...

Alteration		Locus	Alteration type	Codon position	Amino acid change	Reference
13,722	A – G	MT-ND5	TS	coding: 3	SYN	Tanaka and Ozawa, 1994
13,759	G – A	MT-ND5	TS	coding: 1	NON-SYN: Ala → Thr	Rieder <i>et al.</i> , 1998
13,819	T – C	MT-ND5	TS	coding: 1	NON-SYN: Phe → Leu	Ingman <i>et al.</i> , 2000
13,855	C – T	MT-ND5	TS	coding: 1	SYN	Herrnstadt <i>et al.</i> , 2002a
13,886	T – C	MT-ND5	TS	coding: 2	NON-SYN: Leu → Pro	Maca-Meyer <i>et al.</i> , 2001
13,924	C – T	MT-ND5	TS	coding: 1	NON-SYN: Pro → Ser	Herrnstadt <i>et al.</i> , 2002a
13,928	G – C	MT-ND5	TV	coding: 2	NON-SYN: Ser → Thr	Finnilä <i>et al.</i> , 2001b
13,966	A – G	MT-ND5	TS	coding: 1	NON-SYN: Thr → Ala	Kösel <i>et al.</i> , 1998
13,967	C – T	MT-ND5	TS	coding: 2	NON-SYN: Thr → Met	Huoponen <i>et al.</i> , 1993
14,020	T – C	MT-ND5	TS	coding: 1	SYN	Kösel <i>et al.</i> , 1998
14,059	A – G	MT-ND5	TS	coding: 1	NON-SYN: Ile → Val	Herrnstadt <i>et al.</i> , 2002a
14,182	T – C	MT-ND6	TS	coding: 3	SYN	Huoponen <i>et al.</i> , 1993
14,211	C – T	MT-ND6	TS	coding: 1	NON-SYN: Val → Ile	Ingman and Gyllensten, 2007a
14,212	T – C	MT-ND6	TS	coding: 3	SYN	Wallace <i>et al.</i> , 1988
14,280	A – G	MT-ND6	TS	coding: 1	NON-SYN: Ser → Pro	Herrnstadt <i>et al.</i> , 2002a
14,284	C – T	MT-ND6	TS	coding: 3	NON-SYN: Glu → Asp	Ingman <i>et al.</i> , 2000
14,302	T – C	MT-ND6	TS	coding: 3	SYN	Current investigation
14,315	C – T	MT-ND6	TS	coding: 2	NON-SYN: Ser → Asn	Kivisild <i>et al.</i> , 2006
14,371	T – C	MT-ND6	TS	coding: 3	SYN	Ingman <i>et al.</i> , 2000
14,374	T – C	MT-ND6	TS	coding: 3	SYN	Ingman <i>et al.</i> , 2000
14,407	C – T	MT-ND6	TS	coding: 3	SYN	Thangaraj <i>et al.</i> , 2005
14,560	G – A	MT-ND6	TS	coding: 3	SYN	Howell <i>et al.</i> , 1991
14,569	G – A	MT-ND6	TS	coding: 3	SYN	Ozawa <i>et al.</i> , 1991a
14,659	C – T	MT-ND6	TS	coding: 3	SYN	Kivisild <i>et al.</i> , 2006
14,693	A – G	MT-TE	TS	non-coding	---	Sternberg <i>et al.</i> , 2001
14,766	C – T	MT-CYTB	TS	coding: 2	NON-SYN: Thr → Ile	Marzuki <i>et al.</i> , 1991
14,770	C – T	MT-CYTB	TS	coding: 3	SYN	Schuelke <i>et al.</i> , 2002
14,905	G – A	MT-CYTB	TS	coding: 3	SYN	Howell <i>et al.</i> , 1995
15,061	A – G	MT-CYTB	TS	coding: 3	SYN	Maca-Meyer <i>et al.</i> , 2001
15,106	G – A	MT-CYTB	TS	coding: 3	SYN	Ozawa <i>et al.</i> , 1991b
15,110	G – A	MT-CYTB	TS	coding: 1	NON-SYN: Ala → Thr	Reid <i>et al.</i> , 1994

Table B.1 Continued...

Alteration		Locus	Alteration type	Codon position	Amino acid change	Reference
15,208	A – G	MT-CYTB	TS	coding: 3	SYN	Kivisild <i>et al.</i> , 2006
15,217	G – A	MT-CYTB	TS	coding: 3	SYN	Ozawa <i>et al.</i> , 1991b
15,301	G – A	MT-CYTB	TS	coding: 3	SYN	Ozawa <i>et al.</i> , 1991a
15,326	A – G	MT-CYTB	TS	coding: 1	NON-SYN: Thr → Ala	Ozawa <i>et al.</i> , 1991a
15,449	T – C	MT-CYTB	TS	coding: 1	NON-SYN: Phe → Leu	Kivisild <i>et al.</i> , 2006
15,466	G – A	MT-CYTB	TS	coding: 3	SYN	Herrnstadt <i>et al.</i> , 2002a
15,519	T – C	MT-CYTB	TS	coding: 2	NON-SYN: Leu → Pro	Mishmar <i>et al.</i> , 2003
15,550	C – T	MT-CYTB	TS	coding: 3	SYN	Tanaka <i>et al.</i> , 2004
15,766	A – G	MT-CYTB	TS	coding: 3	SYN	Andreu <i>et al.</i> , 1999d
15,930	G – A	MT-TT	TS	non-coding	---	Herrnstadt <i>et al.</i> , 2002a
15,941	T – C	MT-TT	TS	non-coding	---	Mishmar <i>et al.</i> , 2003
15,942	T – C	MT-TT	TS	non-coding	---	Rose <i>et al.</i> , 2001
15,951	A – G	MT-TT	TS	non-coding	---	Ozawa <i>et al.</i> , 1991b
16,069	C – T	MT-DLOOP	TS	non-coding	---	Di Rienzo and Wilson, 1991
16,093	T – C	MT-DLOOP	TS	non-coding	---	Di Rienzo and Wilson, 1991
16,114	C – A	MT-DLOOP	TV	non-coding	---	Horai and Hayasaka, 1990
16,124	T – C	MT-DLOOP	TS	non-coding	---	Greenberg <i>et al.</i> , 1983
16,126	T – C	MT-DLOOP	TS	non-coding	---	Cann and Wilson, 1983
16,129	G – A	MT-DLOOP	TS	non-coding	---	Aquadro and Greenberg, 1983
16,153	G – A	MT-DLOOP	TS	non-coding	---	Vigilant <i>et al.</i> , 1991
16,166	A – C	MT-DLOOP	TV	non-coding	---	Wilkinson-Herbots <i>et al.</i> , 1996
16,167	C – T	MT-DLOOP	TS	non-coding	---	Aquadro and Greenberg, 1983
16,169	C – T	MT-DLOOP	TS	non-coding	---	Horai and Hayasaka, 1990
16,172	T – C	MT-DLOOP	TS	non-coding	---	Greenberg <i>et al.</i> , 1983
16,182	A – C	MT-DLOOP	TV	non-coding	---	Lahermo <i>et al.</i> , 1996
16,183	A – C	MT-DLOOP	TV	non-coding	---	Ozawa <i>et al.</i> , 1995
16,187	C – T	MT-DLOOP	TS	non-coding	---	Aquadro and Greenberg, 1983
16,188	C – T	MT-DLOOP	TS	non-coding	---	Shields <i>et al.</i> , 1993
16,189	T – C	MT-DLOOP	TS	non-coding	---	Aquadro and Greenberg, 1983
16,193	INS C	MT-DLOOP	INS	non-coding	---	Brandstätter <i>et al.</i> , 2004
16,209	T – C	MT-DLOOP	TS	non-coding	---	Di Rienzo and Wilson, 1991
16,212	A – G	MT-DLOOP	TS	non-coding	---	Vigilant <i>et al.</i> , 1989
16,213	G – A	MT-DLOOP	TS	non-coding	---	Horai and Hayasaka, 1990
16,214	C – T	MT-DLOOP	TS	non-coding	---	Vigilant <i>et al.</i> , 1989
16,223	C – T	MT-DLOOP	TS	non-coding	---	Aquadro and Greenberg, 1983
16,230	A – G	MT-DLOOP	TS	non-coding	---	Aquadro and Greenberg, 1983
16,234	C – T	MT-DLOOP	TS	non-coding	---	Horai and Hayasaka, 1990
16,239	C – T	MT-DLOOP	TS	non-coding	---	Vigilant <i>et al.</i> , 1989
16,242	C – T	MT-DLOOP	TS	non-coding	---	Aquadro and Greenberg, 1983
16,243	T – C	MT-DLOOP	TS	non-coding	---	Aquadro and Greenberg, 1983

Table B.1 Continued...

Alteration		Locus	Alteration type	Codon position	Amino acid change	Reference
16,264	C – T	MT-DLOOP	TS	non-coding	---	Horai and Hayasaka, 1990
16,266	C – G	MT-DLOOP	TV	non-coding	---	Vigilant <i>et al.</i> , 1989
16,278	C – T	MT-DLOOP	TS	non-coding	---	Aquadro and Greenberg, 1983
16,291	C – T	MT-DLOOP	TS	non-coding	---	Horai and Hayasaka, 1990
16,291	C – G	MT-DLOOP	TV	non-coding	---	Ingman <i>et al.</i> , 2000
16,294	C – T	MT-DLOOP	TS	non-coding	---	Aquadro and Greenberg, 1983
16,304	T – C	MT-DLOOP	TS	non-coding	---	Aquadro and Greenberg, 1983
16,311	T – C	MT-DLOOP	TS	non-coding	---	Aquadro and Greenberg, 1983
16,320	C – T	MT-DLOOP	TS	non-coding	---	Aquadro and Greenberg, 1983
16,354	C – T	MT-DLOOP	TS	non-coding	---	Piercy <i>et al.</i> , 1993
16,362	T – C	MT-DLOOP	TS	non-coding	---	Aquadro and Greenberg, 1983
16,390	G – A	MT-DLOOP	TS	non-coding	---	Horai and Hayasaka, 1990
16,474	G – T	MT-DLOOP	TS	non-coding	---	Current investigation
16,497	A – G	MT-DLOOP	TS	non-coding	---	Horai and Hayasaka, 1990
16,519	T – C	MT-DLOOP	TS	non-coding	---	Aquadro and Greenberg, 1983

Nucleotides are represented by capital letters, with the rCRS nucleotide indicated first, followed by the observed change. Nucleotide numbers are based on rCRS numbering (Andrews *et al.*, 1999). The numbers indicated in **red** represent previously reported polymorphisms and the numbers presented in **blue** indicate previously unreported polymorphisms. A = adenine; C = cytosine; G = guanine; T = thymine; ATPase6 = ATP synthase F₀ subunit 6; ATPase8 = ATP synthase F₀ subunit 8; MT-CO1 = cytochrome c oxidase subunit I; MT-CO2 = cytochrome c oxidase subunit II; MT-CO3 = cytochrome c oxidase subunit III; MT-CYTB = cytochrome b; MT-DLOOP = control region, including displacement loop; MT-ND1 = NADH dehydrogenase subunit 1; MT-ND2 = NADH dehydrogenase subunit 2; MT-ND3 = NADH dehydrogenase subunit 3; MT-ND4 = NADH dehydrogenase subunit 4; MT-ND4L = NADH dehydrogenase subunit 4L; MT-ND5 = NADH dehydrogenase subunit 5; MT-ND6 = NADH dehydrogenase subunit 6; MT-RNR1 = 12S ribosomal RNA; MT-RNR2 = 16S ribosomal RNA; MT-TC = tRNA cysteine; MT-TE = tRNA glutamic acid; MT-TG = tRNA glycine; MT-TH = tRNA histidine; MT-TI = tRNA isoleucine; MT-TL1 = tRNA leucine (UUR); MT-TL2 = tRNA leucine (CUN); MT-TQ = tRNA glutamine; MT-TS1 = tRNA serine (UCN); MT-TS2 = tRNA serine (AGY); MT-TT = tRNA threonine; MT-TV = tRNA valine; MT-TW = tRNA tryptophan; MT-TY = tRNA tyrosine; INS = insertion; DEL = deletion; TV = transversion; TS = transition; NON-SYN = non-synonymous substitution; SYN = synonymous substitution; Ala = alanine; Asn = asparagine; Asp = aspartic acid; Gly = glycine; Glu = glutamic acid; Ile = isoleucine; Leu = leucine; Met = methionine; Phe = phenylalanine; Pro = proline; Ser = serine; Thr = threonine; Trp = tryptophan; Val = valine.

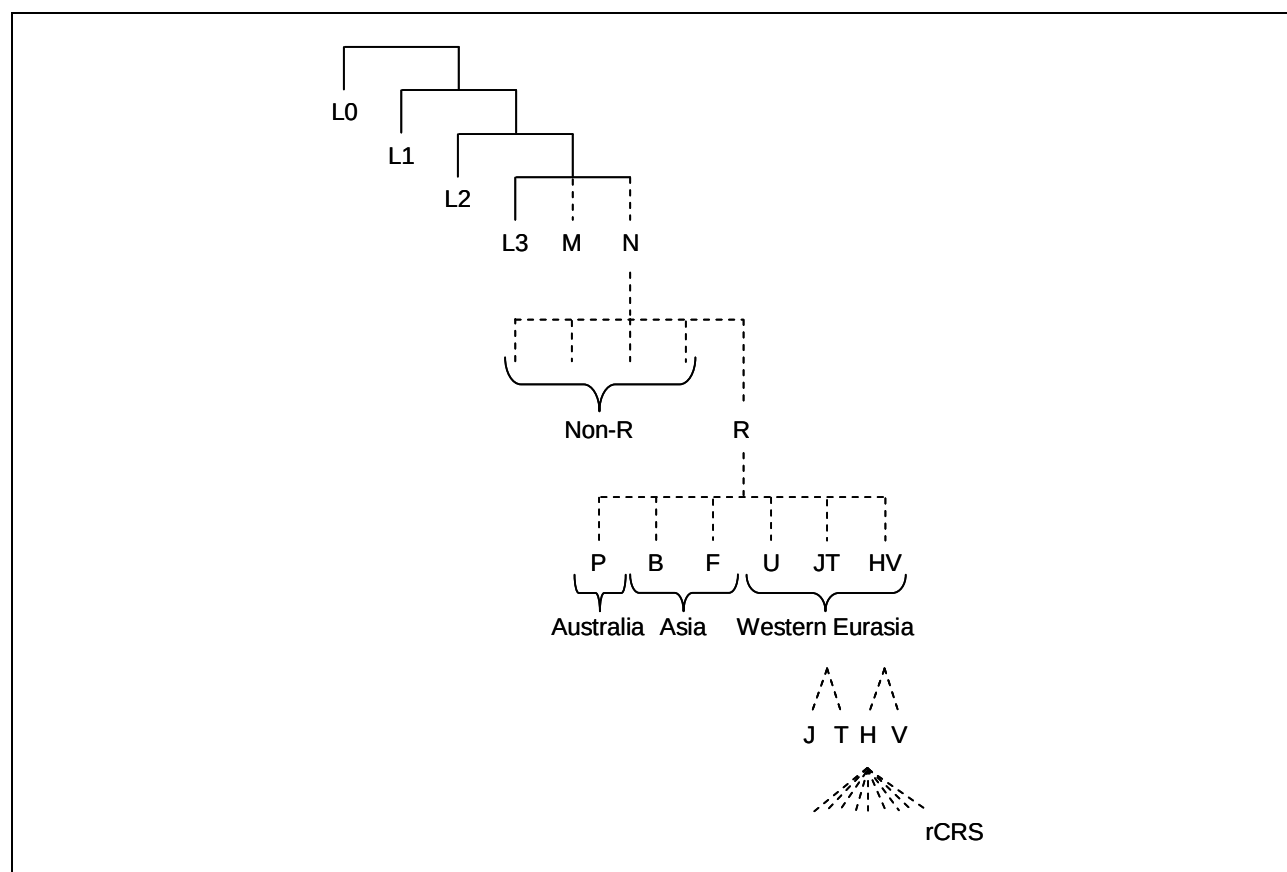
Heteroplasmy was not detected for any of the alterations presented here. Heteroplasmy is rare (Stoneking and Soodyall, 1996), especially when investigating apparently healthy individuals as is the case in the current investigation. A few ambiguous base-calls were observed during visual examination of the electropherograms. These ambiguous base-calls were attributed to the presence of sequence artefacts such as low signal strength and high background noise (refer to Section 5.6.4) and were resolved as discussed in Section 5.6.6 by visual examination of the electropherograms, or via the use of sequence data from another source (overlapping sequence data), or through the generation of a new purified PCR product and the use of a reverse primer to generate sequence data. It must however be noted that a comprehensive and sensitive assessment of heteroplasmy was beyond the technical boundaries of the sequencing approach employed in the current investigation.

APPENDIX C

A schematic representation of a global phylogenetic tree of mitochondrial DNA haplogroups

This appendix contains a schematic representation of the global phylogenetic tree of mitochondrial DNA haplogroups as constructed by Ruiz-Pesini *et al.* in 2007. A schematic representation of the tree is presented in Figure C.1, while the actual mutational mitochondrial DNA phylogenetic tree consisting of 3,000 mtDNA coding region sequences is available to download as a searchable PDF file (Large mtDNA tree) under the heading “Illustrations” on the MITOMAP website (<http://www.mitomap.org>, 2007). The 30 complete Khoi-San mitochondrial DNA sequences generated in the current investigation were expected to be L-specific sequences and therefore to group within the L haplogroup branches observed in the tree.

Figure C.1 Schematic representation of a global phylogenetic tree of mtDNA haplogroups



Dashed lines represent branches and respective haplogroups that were excluded from the sequence data studied in the current investigation. rCRS = revised Cambridge Reference Sequence. Adapted from Ruiz-Pesini *et al.* (2007).

APPENDIX D

Exclusion criteria for L-specific mitochondrial DNA sequences

This appendix contains information regarding the nucleotide substitutions that identify and exclude L-specific sequences from other mitochondrial DNA haplogroup sequences. The nucleotide substitutions are listed in Table D.1. The 30 complete Khoi-San mitochondrial genome sequences were expected to be L-specific sequences and therefore to contain the specific substitutions used for exclusion.

Table D.1 A list of polymorphisms used to exclude L-specific sequences from other mtDNA haplogroups

Criteria excluding haplogroups from other haplogroups	Nucleotide position	Substitution	Number
L0 from L1	4,312	C - T	38
	6,185	T - C	37
	10,589	G - A	36
	11,914	G - A	35
	12,007	G - A	34
L0 and L1 from L2, L3 and M	825	T - A	33
	2,758	G - A	32
	2,885	T - C	31
	7,146	A - G	30
	8,468	C - T	29
	8,655	C - T	28
	10,688	G - A	27
	10,810	T - C	26
	13,105	A - G	25
	13,506	C - T	24
	15,301	G - A	23
L0, L1 and L2 from other haplogroups	769	G - A	22
	1,018	G - A	21
	3,594	C - T	20
	4,104	A - G	19
	7,256	C - T	18
	7,521	G - A	17
	13,650	C - T	16

Table D.1 Continued...

Criteria excluding haplogroups from other haplogroups	Nucleotide position	Substitution	Number
From haplogroup N	8,701	A - G	15
	9,540	T - C	14
	10,398	A - G	13
	10,873	T - C	12
	15,301	G - A	11
From haplogroup R	12,705	C - T	10
From haplogroup HV	11,719	G - A	9
	14,766	C - T	8
From haplogroup H	2,706	A - G	7
	7,028	C - T	6
From sub-haplogroup of H containing CRS	750	A - G	5
	1,438	A - G	4
	4,769	A - G	3
From sub-haplogroup of H containing rCRS	8,860	A - G	2
	15,326	A - G	1

Nucleotides are represented by capital letters, with the rCRS nucleotide indicated first, followed by the observed change. Nucleotide numbers are based on rCRS numbering (Andrews *et al.*, 1999). A = adenine; C = cytosine; G = guanine; T = thymine; CRS = Cambridge Reference Sequence; rCRS = revised Cambridge Reference Sequence. Adapted from Ruiz-Pesini *et al.* (2007).

The exclusion criteria listed in Table D.1 are based on a global mtDNA mutational phylogeny constructed by Ruiz-Pesini *et al.* (2007) and are applied by starting at number one in the bottom right hand corner and working upwards in the table until number 38 is reached. An earlier version of the exclusion criteria (Ruiz-Pesini, 2003) could not distinguish between haplogroup L0 and L1, i.e. criteria were listed only up to number 33 which excludes haplogroups L0 and L1 from haplogroups L2, L3 and M. In the current investigation, this early version of the exclusion criteria was updated using the global mtDNA mutational phylogeny constructed by Ruiz-Pesini *et al.* (2007), as well as the global L0-specific haplogroup network presented in Section 5.10. The updated exclusion criteria presented in Table D.1 permits the exclusion of haplogroup L0 from L1. For a mitochondrial genome sequence to be classified in the L0 haplogroup, all 38 polymorphisms listed in Table D.1 should therefore be present. Exceptions to the criteria were observed in the current investigation (refer to Section 5.8), such exceptions are however rare. The exclusion criteria can not classify the mitochondrial sequence beyond the major haplogroup (L0, L1, L2 or L3), and the haplogroup classification scheme illustrated in Figure 4.1 (refer to Section 4.9) was therefore employed for this purpose.

APPENDIX E

Haplogroup analysis of complete Khoi-San mitochondrial DNA genome sequences

This appendix contains a list of the haplogroups assigned to the 30 complete Khoi-San mitochondrial genome sequences. The haplogroup for each individual, as well as the sub-haplogroup to which the individual belongs, are indicated in Table E.1. The complete Khoi-San mitochondrial genome sequences were identified as L-specific sequences through the use of the haplogroup exclusion criteria listed in Appendix D, and haplogroup analyses were performed by using the coding region only haplogroup classification scheme described in Section 4.9. After the compilation of the global L0-specific haplogroup network (refer to Section 5.10), and the construction of a revised classification scheme for the L0 haplogroup (refer to Section 5.12), the 30 Khoi-San mitochondrial genome sequences were re-classified using the revised classification scheme as indicated in Table E.1.

Table E.1 Haplogroup analyses of the 30 complete Khoi-San mitochondrial DNA genome sequences

Khoi-San individual	Ethnic group ^a	Geographic origin ^b	Haplogroup/sub-haplogroup classification	
			Wallace, 2004 ^c	Current investigation ^d
KS_CGR_001	!Kung	Southern Africa	L0a	L0aA2
KS_CGR_004	!Kung	Namibia	L0b	L0bB2a2
KS_CGR_005	!Kung	Namibia	L0b	L0bB2b
KS_CGR_044	!Kung	Angola	L0b	L0bA1
KS_CGR_048	!Kung	Angola	L0b	L0bB1b1
KS_CGR_056	!Kung	Angola	L0b	L0bB1b2
KS_CGR_059	!Kung	Angola	L0a	L0aA1
KS_CGR_060	!Kung	Angola	L0b	L0bB1b1
KS_CGR_061	!Kung	Angola	L0b	L0bB3a
KS_CGR_069	!Kung	Angola	L2b	---
KS_CGR_077	!Kung	Angola	L0b	L0bB2a2
KS_CGR_083	!Kung	Angola	L0a	L0aA2
KS_CGR_084	!Kung	Angola	L0a	L0aA1
KS_CGR_090	!Kung	Angola	L3b	---
KS_CGR_091	!Kung	Angola	L0b	L0bB2a3

Table E.1 Continued...

Khoi-San individual	Ethnic group ^a	Geographic origin ^b	Haplogroup/sub-haplogroup classification	
			Wallace, 2004 ^c	Current investigation ^d
KS_CGR_092	!Kung	Namibia	L0b	L0bB2a3
KS_CGR_105	!Kung	Angola	L0b	L0bB2a1
KS_CGR_111	!Kung	Angola	L0b	L0bB2a3
KS_CGR_117	!Kung	Angola	L0b	L0bB2a3
KS_CGR_120	!Kung	Angola	L0a	L0aA1
KS_CGR_121	!Kung	Angola	L0b	L0bB2a2
KS_CGR_127	!Kung	Namibia	L0b	L0bB3a
KS_CGR_140	!Kung	Angola	L3a1	---
KS_CGR_143	!Kung	Namibia	L3a1	---
KS_CGR_157	!Kung	Angola	L0a	L0aA2
KS_CGR_164	!Kung	Namibia	L2b	---
KS_CGR_172	!Kung	Angola	L3b	---
KS_CGR_173	!Kung	Angola	L2b	---
KS_CGR_175	!Kung	Namibia	L3a1	---
KS_CGR_176	!Kung	Angola	L0a	L0aA2

^a = the Khoi-San cohort studied in the current investigation consisted of 30 Vasikela !Kung individuals (Chen *et al.*, 2000). Individuals were classified as !Kung according to the ethnicity listed on either their identity documents or birth certificates, or through verbal confirmation of ethnicity in a few cases where such information was not available (Chen *et al.*, 2000). ^b = the Khoi-San samples studied in the current investigation were previously collected in the north-western Kalahari Desert in southern Africa (Chen *et al.*, 2000). Geographic origin was however not based on the sample collection location, but was assigned according to the country in which the individual was born. ^c = haplogroup classification according to the coding region only classification scheme described in Section 4.9. ^d = haplogroup classification according to the revised L0 classification scheme proposed in Section 5.12.

APPENDIX F

Comparison of haplogroup classification schemes

This appendix contains Table F.1 in which a comparison is drawn between the haplogroup classification scheme used in the current investigation (refer to Section 4.9) and the alternative classification scheme generally used in the literature (refer to Section 3.3.1.1). This haplogroup comparison was compiled using data from the current investigation as well as data generated by Salas *et al.* (2002), Salas *et al.* (2004a), Kivisild *et al.* (2004), Kivisild *et al.* (2006), Gonder *et al.* (2007) and Ruiz-Pesini *et al.* (2007).

Table F.1 Comparison of haplogroup and sub-haplogroup designations between the classification scheme used in the current investigation and the classification scheme generally used in the literature

Haplogroup	Haplogroup classification scheme		Haplogroup	Haplogroup classification scheme	
	Wallace, 2004 ^a	Alternative ^b		Wallace, 2004 ^a	Alternative ^b
L0	L0a	L0k	L1	L1b2	L1c1
	---	L0d3 ^c		L1b3 ^d	L1c3
	---	L0f	L2	L2	L2
	L0b	L0d		L2a	L2a
	---	L0d1		L2b	L2b
	---	L0d2		L2c	L2c
	L0c	L0a		L2d	L2d
	L0c1	L0a1	L3	L3	L3
	L0c2	L0a2		L3a	L3e
L1	L1	L1		L3a1	L3e2
	L1a	L1b		L3a2	L3e3
	L1a1	L1b1		L3a3	L3e1
	L1a2	L1b1		L3b	L3d
	L1b	L1c		L3c	L3b
	L1b1	L1c2		---	---

^a = classification scheme used in current investigation (refer to Section 4.9). ^b = haplogroup classification scheme used in other investigations described in the literature (refer to Section 3.3.1.1). ^c = a new branch in L0d, specific to East African Khoi-San speakers (Gonder *et al.*, 2007), named L0d3 in the current investigation to distinguish it from L0d1 and L0d2. The L0d3 branch was first described by Gonder *et al.* (2007) and named L0dTZ. ^d = a new sub-haplogroup to be included in classification schemes (refer to Section 5.9.2.1). This haplogroup comparison was compiled using data from the current investigation as well as data generated by Salas *et al.* (2002), Salas *et al.* (2004a), Kivisild *et al.* (2004), Kivisild *et al.* (2006), Gonder *et al.* (2007) and Ruiz-Pesini *et al.* (2007).

The comparison presented in the above table was used to relate the data generated in the current investigation with the data obtained in other investigations as described in the literature. This comparison is not a comprehensive list and does not include all

sub-haplogroups identified to date, as it was compiled to serve as a guide in the current investigation with the specific aims of the current investigation in mind. It can be observed that although the designations of sub-haplogroups differ within a haplogroup, the major haplogroup classification remains the same irrespective of the classification scheme used to assign the haplogroup designation.

APPENDIX G

Mitochondrial genome sequences used in analyses

This appendix contains a list of all the mitochondrial DNA genome sequences used in the construction of the phylogenetic trees (refer to Section 5.9), the assembly of the global L0-specific haplogroup network (refer to Section 5.10), and the performance of the statistical analyses (refer to Section 5.11). The assigned mitochondrial DNA haplogroups, as well as the relevant ethnic and geographic information available for each genome sequence are also presented. The mitochondrial genome sequences used for the different analyses included the rCRS (Andrews *et al.*, 1999), a mitochondrial genome sequence representing the chimpanzee (*Pan troglodytes*), 30 mtDNA genome sequences representing the 30 Khoi-San individuals studied in the current investigation, 13 Khoi-San mtDNA genome sequences obtained from a previous investigation (Mouton, 2003), as well as a number of L-specific mtDNA sequences obtained from GenBank® (Benson *et al.*, 2008) in the period between January 2004 and June 2007. The relevant information for each individual mitochondrial genome sequence is presented in Table G.1.

Table G.1 List of mitochondrial genome sequences used in all analyses

GenBank® accession number	GI number	Sequence name/ isolate	Ethnicity/ geographic origin ^a	Haplogroup/sub-haplogroup classification			No. of base pairs	Reference
				Published	Wallace, 2004	Current investigation		
---	---	KS_CGR_001/1800	!Kung/Southern Africa	---	L0a	L0aA2	16,567	Current investigation
---	---	KS_CGR_004/1803	!Kung/Namibia	---	L0b	L0bB2a2	16,566	Current investigation
---	---	KS_CGR_005/1804	!Kung/Namibia	---	L0b	L0bB2b	16,566	Current investigation
---	---	KS_CGR_044/1843	!Kung/Angola	---	L0b	L0bA1	16,569	Current investigation
---	---	KS_CGR_048/1847	!Kung/Angola	---	L0b	L0bB1b1	16,567	Current investigation
---	---	KS_CGR_056/1855	!Kung/Angola	---	L0b	L0bB1b2	16,568	Current investigation
---	---	KS_CGR_059/1858	!Kung/Angola	---	L0a	L0aA1	16,570	Current investigation
---	---	KS_CGR_060/1859	!Kung/Angola	---	L0b	L0bB1b1	16,568	Current investigation
---	---	KS_CGR_061/1860	!Kung/Angola	---	L0b	L0bB3a	16,568	Current investigation

Table G.1 Continued...

GenBank [®] accession number	GI number	Sequence name/ isolate	Ethnicity/ geographic origin ^a	Haplogroup/sub-haplogroup classification			No. of base pairs	Reference
				Published	Wallace, 2004	Current investigation		
---	---	KS_CGR_069/ 1868	!Kung/Angola	---	L2b	---	16,568	Current investigation
---	---	KS_CGR_077/ 1876	!Kung/Angola	---	L0b	L0bB2a2	16,566	Current investigation
---	---	KS_CGR_083/ 1882	!Kung/Angola	---	L0a	L0aA2	16,567	Current investigation
---	---	KS_CGR_084/ 1883	!Kung/Angola	---	L0a	L0aA1	16,569	Current investigation
---	---	KS_CGR_090/ 1889	!Kung/Angola	---	L3b	---	16,568	Current investigation
---	---	KS_CGR_091/ 1890	!Kung/Angola	---	L0b	L0bB2a3	16,567	Current investigation
---	---	KS_CGR_092/ 1891	!Kung/Namibia	---	L0b	L0bB2a3	16,571	Current investigation
---	---	KS_CGR_105/ 1904	!Kung/Angola	---	L0b	L0bB2a1	16,566	Current investigation
---	---	KS_CGR_111/ 1910	!Kung/Angola	---	L0b	L0bB2a3	16,566	Current investigation
---	---	KS_CGR_117/ 1916	!Kung/Angola	---	L0b	L0bB2a3	16,567	Current investigation
---	---	KS_CGR_120/ 1919	!Kung/Angola	---	L0a	L0aA1	16,568	Current investigation
---	---	KS_CGR_121/ 1920	!Kung/Angola	---	L0b	L0bB2a2	16,566	Current investigation
---	---	KS_CGR_127/ 1926	!Kung/Namibia	---	L0b	L0bB3a	16,568	Current investigation
---	---	KS_CGR_140/ 1939	!Kung/Angola	---	L3a1	---	16,569	Current investigation
---	---	KS_CGR_143/ 1942	!Kung/Namibia	---	L3a1	---	16,569	Current investigation
---	---	KS_CGR_157/ 1956	!Kung/Angola	---	L0a	L0aA2	16,567	Current investigation
---	---	KS_CGR_164/ 1963	!Kung/Namibia	---	L2b	---	16,569	Current investigation
---	---	KS_CGR_172/ 1971	!Kung/Angola	---	L3b	---	16,568	Current investigation
---	---	KS_CGR_173/ 1972	!Kung/Angola	---	L2b	---	16,568	Current investigation
---	---	KS_CGR_175/ 1974	!Kung/Namibia	---	L3a1	---	16,568	Current investigation
---	---	KS_CGR_176/ 1975	!Kung/Angola	---	L0a	L0aA2	16,568	Current investigation
---	---	KS_CGR_010/ SA93-10	!Kung/Namibia	---	L0b	L0bB2a3	16,567	Mouton, 2003
---	---	KS_CGR_032/ SA93-32	!Kung/South Africa	---	L0b	L0bB2a3	16,567	Mouton, 2003
---	---	KS_CGR_038/ SA93-38	!Kung/Namibia	---	L0a	L0aA2	16,568	Mouton, 2003

Table G.1 Continued...

GenBank [®] accession number	GI number	Sequence name/ isolate	Ethnicity/ geographic origin ^a	Haplogroup/sub-haplogroup classification			No. of base pairs	Reference
				Published	Wallace, 2004	Current investigation		
---	---	KS_CGR_040/ SA93-40	!Kung/Angola	---	L0b	L0bB1b2	16,569	Mouton, 2003
---	---	KS_CGR_052/ SA93-52	!Kung/Angola	---	L0b	L0bB1a	16,569	Mouton, 2003
---	---	KS_CGR_072/ SA93-72	!Kung/Angola	---	L0b	L0bB2a3	16,567	Mouton, 2003
---	---	KS_CGR_080/ SA93-80	!Kung/Angola	---	L0a	L0aA2	16,568	Mouton, 2003
---	---	KS_CGR_082/ SA93-82	!Kung/Angola	---	L0b	L0bB2a3	16,567	Mouton, 2003
---	---	KS_CGR_094/ SA93-94	!Kung/Angola	---	L0b	L0bB3b	16,569	Mouton, 2003
---	---	KS_CGR_102/ SA93-102	!Kung/Angola	---	L0b	L0bB2a2	16,567	Mouton, 2003
---	---	KS_CGR_106/ SA93-106	!Kung/Angola	---	L0a	L0aA1	16,568	Mouton, 2003
---	---	KS_CGR_122/ SA93-122	!Kung/Angola	---	L0b	L0bB2a3	16,567	Mouton, 2003
---	---	KS_CGR_130/ SA93-130	!Kung/Angola	---	L0b	L0bB3a	16,569	Mouton, 2003
AF346967	13272612	bam17b	Bamileke/Africa	---	L3a2	---	16,567	Ingman <i>et al.</i> , 2000
AF346968	13272626	biak757	Biaka/Africa	---	L1b2	---	16,567	Ingman <i>et al.</i> , 2000
AF346969	13272640	biak776	Biaka/Africa	---	L1b2	---	16,567	Ingman <i>et al.</i> , 2000
AF346976	13272738	eff 004	Effik/Africa	---	L2a	---	16,570	Ingman <i>et al.</i> , 2000
AF346977	13272752	eff007	Effik/Africa	---	L2a	---	16,569	Ingman <i>et al.</i> , 2000
AF346980	13272794	ewon w77	Ewondo/Africa	---	L3a3	---	16,570	Ingman <i>et al.</i> , 2000
AF346985	13272864	haus 004	Hausa/Africa	---	L0c1	L0c1B	16,567	Ingman <i>et al.</i> , 2000
AF346986	13272878	ibo 006	Ibo/Africa	---	L1a2	---	16,567	Ingman <i>et al.</i> , 2000
AF346987	13272892	ibo 12	Ibo/Africa	---	L1b2	---	16,566	Ingman <i>et al.</i> , 2000
AF346992	13272962	kiku dbs	Kikuyu/Africa	---	L1b1	---	16,571	Ingman <i>et al.</i> , 2000
AF346994	13272990	Liso f30	Lisongo ^e /Africa	---	L3a1	---	16,569	Ingman <i>et al.</i> , 2000
AF346995	13273004	mand 318	Mandenka/Africa	---	L2c	---	16,568	Ingman <i>et al.</i> , 2000
AF346996	13273018	mbe 225	Mbenzele/Africa	---	L1b2	---	16,566	Ingman <i>et al.</i> , 2000
AF346997	13273032	mbe 240	Mbenzele/Africa	---	L1b2	---	16,567	Ingman <i>et al.</i> , 2000
AF346998	13273046	mbu 1057	Mbuti/Africa	---	L0c2	L0c2D2	16,561	Ingman <i>et al.</i> , 2000
AF346999	13273060	mbu 1058	Mbuti/Africa	---	L0c2	L0c2D2	16,562	Ingman <i>et al.</i> , 2000
AF347000	13273074	mkam oao	Mkamba/Africa	---	L3b\$	---	16,569	Ingman <i>et al.</i> , 2000
AF347008	13273186	san 1	San/Africa	---	L0a	L0aA1	16,569	Ingman <i>et al.</i> , 2000
AF347009	13273200	san 2	San/Africa	---	L0a	L0aA2	16,567	Ingman <i>et al.</i> , 2000
AF347014	13273270	yoru 002	Yoruba/Africa	---	L3b	---	16,567	Ingman <i>et al.</i> , 2000
AF347015	13273284	yoru 013	Yoruba/Africa	---	L3a1	---	16,571	Ingman <i>et al.</i> , 2000
AF381981	17985375	441	Mauritania	L2	L2c	---	16,568	Maca-Meyer <i>et al.</i> , 2001
AF381988	17985473	271	Morocco	L1a	L0c1	L0c1A	16,567	Maca-Meyer <i>et al.</i> , 2001
AF381991	17985515	430	Mauritania	L3b	L3c	---	16,567	Maca-Meyer <i>et al.</i> , 2001
AF381992	29294531	432	Mauritania	L1c	L0,L1	L1	16,566	Maca-Meyer <i>et al.</i> , 2001
AF381994	29294532	451	Mauritania	L1b	L1a1	---	16,567	Maca-Meyer <i>et al.</i> , 2001
AF381998	17985613	800	Jordan	L3d	L3b	---	16,567	Maca-Meyer <i>et al.</i> , 2001

Table G.1 Continued...

GenBank® accession number	GI number	Sequence name/ isolate	Ethnicity/ geographic origin ^a	Haplogroup/sub-haplogroup classification			No. of base pairs	Reference
				Published	Wallace, 2004	Current investigation		
AY195766	29690686	A11L2b	Africa	L2b	L2b	---	16,567	Mishmar <i>et al.</i> , 2003
AY195776	48596191	A9L2a	Africa	L2a	L2a	---	16,569	Mishmar <i>et al.</i> , 2003
AY195777	29690840	A10L1A2	Africa	L1a2	L0b	L0bA2	16,557	Mishmar <i>et al.</i> , 2003
AY195780	29690882	A2L1	Africa	L1a	L0c1	L0c1B	16,567	Mishmar <i>et al.</i> , 2003
AY195783	48596203	A4L1B2	Africa	L1b2	L1a2	---	16,567	Mishmar <i>et al.</i> , 2003
AY195785	48596208	A6L2C	Africa	L2c	L2c	---	16,567	Mishmar <i>et al.</i> , 2003
AY195788	48596215	A5L2A1	Africa	L2a1	L2a	---	16,570	Mishmar <i>et al.</i> , 2003
AY195789	48596217	A3L1B1	Africa	L1b1	L1b2	---	16,567	Mishmar <i>et al.</i> , 2003
AY963585	75905883	Tor65	Uganda	L0f	L0a	L0aC	16,569	Macaulay <i>et al.</i> , 2005
D38112 ^b	644480	SB17	Uganda	---	L0c2	L0c2A	16,559	Horai <i>et al.</i> , 1995
DQ112686	70955706	AD1	Dominican Republic	L0a2a	L0c2	L0c2B1	15,574	Kivisild <i>et al.</i> , 2006
DQ112687	70955708	AD2	Dominican Republic	L0a2a	L0c2	L0c2B1	15,576	Kivisild <i>et al.</i> , 2006
DQ112688	70955710	AD3	Dominican Republic	L0a2a	L0c2	L0c2B1	15,574	Kivisild <i>et al.</i> , 2006
DQ112689	70955712	AD4	Dominican Republic	L0a2a	L0c2	L0c2B2	15,574	Kivisild <i>et al.</i> , 2006
DQ112690	70955714	AD5	Dominican Republic	L1b1	L1a2	---	15,583	Kivisild <i>et al.</i> , 2006
DQ112691	70955716	AD6	Dominican Republic	L1b1	L1a2	---	15,583	Kivisild <i>et al.</i> , 2006
DQ112692	70955718	AD7	Dominican Republic	L1b1	L1a2	---	15,583	Kivisild <i>et al.</i> , 2006
DQ112693	70955720	AD8	Dominican Republic	L1b1	L1a2	---	15,583	Kivisild <i>et al.</i> , 2006
DQ112694	70955722	AD9	Dominican Republic	L1c2	L1b1	---	15,584	Kivisild <i>et al.</i> , 2006
DQ112695	70955724	AD10	Dominican Republic	L1c1	L1b2	---	15,574	Kivisild <i>et al.</i> , 2006
DQ112696	70955726	AD11	Dominican Republic	L1c3	L0,L1	L1	15,582	Kivisild <i>et al.</i> , 2006
DQ112697	70955728	AD12	Dominican Republic	L1c3	L0,L1	L1	15,582	Kivisild <i>et al.</i> , 2006
DQ112698	70955730	AD13	Dominican Republic	L2a1	L2a	---	15,585	Kivisild <i>et al.</i> , 2006
DQ112699	70955732	AD14	Dominican Republic	L2a1	L2a	---	15,585	Kivisild <i>et al.</i> , 2006
DQ112700	70955734	AD15	Dominican Republic	L2a1	L2a	---	15,585	Kivisild <i>et al.</i> , 2006
DQ112701	70955736	AD16	Dominican Republic	L2a1	L2a	---	15,585	Kivisild <i>et al.</i> , 2006
DQ112702	70955738	AD17	Dominican Republic	L2b	L2b	---	15,583	Kivisild <i>et al.</i> , 2006
DQ112703	70955740	AD18	Dominican Republic	L2b	L2b	---	15,583	Kivisild <i>et al.</i> , 2006
DQ112704	70955742	AD19	Dominican Republic	L2b	L2b	---	15,585	Kivisild <i>et al.</i> , 2006
DQ112705	70955744	AD20	Dominican Republic	L2b	L2b	---	15,585	Kivisild <i>et al.</i> , 2006
DQ112706	70955746	AD21	Dominican Republic	L2c	L2c	---	15,583	Kivisild <i>et al.</i> , 2006
DQ112707	70955748	AD22	Dominican Republic	L2c	L2c	---	15,583	Kivisild <i>et al.</i> , 2006
DQ112708	70955750	AD23	Dominican Republic	L2d	L2d\$	---	15,585	Kivisild <i>et al.</i> , 2006
DQ112709	70955752	AD24	Dominican Republic	L2d	L2d	---	15,591	Kivisild <i>et al.</i> , 2006
DQ112710	70955754	AD25	Dominican Republic	L3b	L3c	---	15,582	Kivisild <i>et al.</i> , 2006
DQ112711	70955756	AD26	Dominican Republic	L3b	L3c	---	15,582	Kivisild <i>et al.</i> , 2006
DQ112712	70955758	AD27	Dominican Republic	L3b	L3c	---	15,582	Kivisild <i>et al.</i> , 2006
DQ112713	70955760	AD28	Dominican Republic	L3d	L3b	---	15,583	Kivisild <i>et al.</i> , 2006
DQ112714	70955762	AD29	Dominican Republic	L3d1	L3b	---	15,583	Kivisild <i>et al.</i> , 2006
DQ112715	70955764	AD30	Dominican Republic	L3d1	L3b	---	15,583	Kivisild <i>et al.</i> , 2006
DQ112716	70955766	AD31	Dominican Republic	L3d3	L3b	---	15,583	Kivisild <i>et al.</i> , 2006
DQ112717	70955768	AD32	Dominican Republic	L3f1	L3b\$	---	15,582	Kivisild <i>et al.</i> , 2006
DQ112718	70955770	AD33	Dominican Republic	L3h	L3b\$	---	15,583	Kivisild <i>et al.</i> , 2006
DQ112719	70955772	AD34	Dominican Republic	L3e1	L3a3	---	15,585	Kivisild <i>et al.</i> , 2006
DQ112720	70955774	AD35	Dominican Republic	L3e2	L3a1	---	15,585	Kivisild <i>et al.</i> , 2006

Table G.1 Continued...

GenBank® accession number	GI number	Sequence name/ isolate	Ethnicity/ geographic origin ^a	Haplogroup/sub-haplogroup classification			No. of base pairs	Reference
				Published	Wallace, 2004	Current investigation		
DQ112721	70955776	AD36	Dominican Republic	L3e2	L3a1	---	15,585	Kivisild <i>et al.</i> , 2006
DQ112722	70955778	AD37	Dominican Republic	L3e4	L3a3\$	---	15,583	Kivisild <i>et al.</i> , 2006
DQ112723	70955780	AF83	Mossi/Burkina Faso	L2a1	L2a	---	15,583	Kivisild <i>et al.</i> , 2006
DQ112724	70955782	AF84	Mossi/Burkina Faso	L2a1	L2a	---	15,585	Kivisild <i>et al.</i> , 2006
DQ112725	70955784	AF82	Mossi/Burkina Faso	L2a1	L2a	---	15,583	Kivisild <i>et al.</i> , 2006
DQ112726	70955786	AF65	Rimaibe/Burkina Faso	L3e2	L3a1	---	15,585	Kivisild <i>et al.</i> , 2006
DQ112727	70955788	AF91	Rimaibe/Burkina Faso	L1b1	L1a2	---	15,583	Kivisild <i>et al.</i> , 2006
DQ112728	70955790	AF66	Rimaibe/Burkina Faso	L3e2	L3a1	---	15,585	Kivisild <i>et al.</i> , 2006
DQ112729	70955792	AF63	Rimaibe/Burkina Faso	L3e2	L3a1	---	15,585	Kivisild <i>et al.</i> , 2006
DQ112730	70955794	AF79	Rimaibe/Burkina Faso	L2a1	L2a	---	15,585	Kivisild <i>et al.</i> , 2006
DQ112731	70955796	AF80	Rimaibe/Burkina Faso	L2a1	L2a	---	15,585	Kivisild <i>et al.</i> , 2006
DQ112732	70955798	AF92	Rimaibe/Burkina Faso	L0a1	L0c1	L0c1B	15,583	Kivisild <i>et al.</i> , 2006
DQ112733	70955800	AF90	Mossi/Burkina Faso	L1b1	L1a2	---	15,583	Kivisild <i>et al.</i> , 2006
DQ112734	70955802	AF67	Rimaibe/Burkina Faso	L3e2	L3a1	---	15,585	Kivisild <i>et al.</i> , 2006
DQ112735	70955804	AF81	Rimaibe/Burkina Faso	L2c	L2c	---	15,583	Kivisild <i>et al.</i> , 2006
DQ112736	70955806	AF86	Rimaibe/Burkina Faso	L3f1	L3b\$	---	15,582	Kivisild <i>et al.</i> , 2006
DQ112737	70955808	AF87	Foulbe/Burkina Faso	L1b	L1a2	---	15,583	Kivisild <i>et al.</i> , 2006
DQ112738	70955810	AF68	Foulbe/Burkina Faso	L3b	L3c	---	15,582	Kivisild <i>et al.</i> , 2006
DQ112739	70955812	AF69	Foulbe/Burkina Faso	L3d1	L3b	---	15,583	Kivisild <i>et al.</i> , 2006
DQ112740	70955814	AF64	Foulbe/Burkina Faso	L3e2	L3a1	---	15,585	Kivisild <i>et al.</i> , 2006
DQ112741	70955816	AF88	Foulbe/Burkina Faso	L1b1	L1a2	---	15,583	Kivisild <i>et al.</i> , 2006
DQ112742	70955818	AF70	Foulbe/Burkina Faso	L3b	L3c	---	15,582	Kivisild <i>et al.</i> , 2006
DQ112743	70955820	AF71	Foulbe/Burkina Faso	L3b	L3c	---	15,582	Kivisild <i>et al.</i> , 2006
DQ112744	70955822	AF72	Foulbe/Burkina Faso	L3d1	L3b	---	15,583	Kivisild <i>et al.</i> , 2006
DQ112745	70955824	AF73	Foulbe/Burkina Faso	L3b	L3c	---	15,582	Kivisild <i>et al.</i> , 2006
DQ112746	70955826	AF74	Mossi/Burkina Faso	L3b	L3c	---	15,584	Kivisild <i>et al.</i> , 2006
DQ112747	70955828	AF75	Mossi/Burkina Faso	L3e3	L3a2	---	15,585	Kivisild <i>et al.</i> , 2006
DQ112748	70955830	AF76	Mossi/Burkina Faso	L3b	L3c	---	15,582	Kivisild <i>et al.</i> , 2006
DQ112756	70955846	AF77	Mossi/Burkina Faso	L3d	L3b	---	15,583	Kivisild <i>et al.</i> , 2006
DQ112757	70955848	AF78	Mossi/Burkina Faso	L3e	L3a3\$	---	15,583	Kivisild <i>et al.</i> , 2006
DQ112758	70955850	AF85	Mossi/Burkina Faso	L2a1	L2a	---	15,582	Kivisild <i>et al.</i> , 2006
DQ112759	70955852	AF89	Foulbe/Burkina Faso	L1b1	L1a2	---	15,583	Kivisild <i>et al.</i> , 2006
DQ112777	70955888	AF16	Nuba/Sudan	L0a1	L0c1	L0c1A	15,583	Kivisild <i>et al.</i> , 2006
DQ112778	70955890	AF14	Nuba/Sudan	L2a2	L2a	---	15,585	Kivisild <i>et al.</i> , 2006
DQ112782	70955898	AF10	Berta/Ethiopia	L2a1	L2a	---	15,585	Kivisild <i>et al.</i> , 2006
DQ112792	70955918	AF41	Eritrea	pre-HV1	L3b\$	---	15,585	Kivisild <i>et al.</i> , 2006
DQ112794	70955922	AF06	Lissongo ^e /Africa	L1c2	L1b2\$	---	15,589	Kivisild <i>et al.</i> , 2006
DQ112796	70955926	AF07	Lissongo ^e /Africa	L4g	L3b\$	---	15,585	Kivisild <i>et al.</i> , 2006
DQ112829	70955992	AF53	Ghana	L1b1	L1a2	---	15,583	Kivisild <i>et al.</i> , 2006
DQ112830	70955994	AF54	Zulu/Africa	L0d2	L0b	L0bA1	15,582	Kivisild <i>et al.</i> , 2006
DQ112844	70956022	AF17	NaSan/Africa	L0d1	L0b	L0bB1b2	15,584	Kivisild <i>et al.</i> , 2006
DQ112845	70956024	AF18	Khoisan ^f /Africa	L4g	L3b\$	---	15,585	Kivisild <i>et al.</i> , 2006
DQ112847	70956028	AF19	Herero/Africa	L3d3a	L3b	---	15,583	Kivisild <i>et al.</i> , 2006
DQ112848	70956030	AF20	Sotho/Africa	L0a2a	L0c2	L0c2B2	15,574	Kivisild <i>et al.</i> , 2006
DQ112849	70956032	AF21	Pedi/Africa	L2a1	L2a	---	15,585	Kivisild <i>et al.</i> , 2006

Table G.1 Continued...

GenBank® accession number	GI number	Sequence name/ isolate	Ethnicity/ geographic origin ^a	Haplogroup/sub-haplogroup classification			No. of base pairs	Reference
				Published	Wallace, 2004	Current investigation		
DQ112850	70956034	AF22	Tswana/Africa	L0d2	L0b	L0bA2	15,584	Kivisild <i>et al.</i> , 2006
DQ112851	70956036	AF23	Ambo/Africa	L3d3a	L3b	---	15,583	Kivisild <i>et al.</i> , 2006
DQ112852	70956038	AF24	San/Namibia	L0d1	L0b	L0bB1b2	15,584	Kivisild <i>et al.</i> , 2006
DQ112853	70956040	AF26	Herero/Africa	L3d3a	L3b	---	15,583	Kivisild <i>et al.</i> , 2006
DQ112854	70956042	AF25	Zulu/Africa	L0d1	L0b	L0bB3b	15,584	Kivisild <i>et al.</i> , 2006
DQ112855	70956044	AF27	Tswana/Africa	L0d1	L0b	L0bB1b3	15,583	Kivisild <i>et al.</i> , 2006
DQ112856	70956046	AF55	San/Namibia	L0d1	L0b	L0bB2a2	15,582	Kivisild <i>et al.</i> , 2006
DQ112857	70956048	AF28	Herero/Africa	L3d3a	L3b	---	15,583	Kivisild <i>et al.</i> , 2006
DQ112883	70956100	AF29	Mandenka/Africa	L2c	L2c	---	15,583	Kivisild <i>et al.</i> , 2006
DQ112884	70956102	AF30	Mandenka/Africa	L3d	L3b	---	15,583	Kivisild <i>et al.</i> , 2006
DQ112899	70956132	AF09	Khoisan/Africa	L0d1	L0b	L0bB2a3	15,582	Kivisild <i>et al.</i> , 2006
DQ112900	70956134	AF08	Khoisan/Africa	L2a1	L2a	---	15,585	Kivisild <i>et al.</i> , 2006
DQ112901	70956136	AF102	Pygmy/Zaire	L0a2b	L0c2	L0c2D1	15,576	Kivisild <i>et al.</i> , 2006
DQ112902	70956138	AF02	Biaka Pygmy/Africa	L1c2	L1b2\$	---	15,587	Kivisild <i>et al.</i> , 2006
DQ112903	70956140	AF31	Pygmy/Zaire	L2a2	L2a	---	15,585	Kivisild <i>et al.</i> , 2006
DQ112904	70956142	AF05	Pygmy/Zaire	L0a2b	L0c2	L0c2D2	15,574	Kivisild <i>et al.</i> , 2006
DQ112905	70956144	AF32	Pygmy/Zaire	L2a2	L2a	---	15,585	Kivisild <i>et al.</i> , 2006
DQ112906	70956146	AF33	Pygmy/Zaire	L0a2b	L0c2	L0c2D2	15,576	Kivisild <i>et al.</i> , 2006
DQ112907	70956148	AF34	Pygmy/CAR	L1c1a	L1b2	---	15,582	Kivisild <i>et al.</i> , 2006
DQ112908	70956150	AF35	Pygmy/CAR	L1c1a	L1b2	---	15,582	Kivisild <i>et al.</i> , 2006
DQ112909	70956152	AF04	Pygmy/Zaire	L2a2	L2a	---	15,585	Kivisild <i>et al.</i> , 2006
DQ112910	70956154	AF93	Pygmy/CAR	L1c1a	L1b2	---	15,582	Kivisild <i>et al.</i> , 2006
DQ112911	70956156	AF52	Pygmy/Zaire	L2a2	L2a	---	15,585	Kivisild <i>et al.</i> , 2006
DQ112912	70956158	AF94	Pygmy/CAR	L1c1a	L1b2	---	15,582	Kivisild <i>et al.</i> , 2006
DQ112913	70956160	AF95	Pygmy/CAR	L1c1a	L1b2	---	15,582	Kivisild <i>et al.</i> , 2006
DQ112914	70956162	AF01	Biaka Pygmy/Africa	L1c2	L1b2\$	---	15,587	Kivisild <i>et al.</i> , 2006
DQ112915	70956164	AF97	Pygmy/Zaire	L0a2b	L0c2	L0c2D1	15,576	Kivisild <i>et al.</i> , 2006
DQ112916	70956166	AF36	Pygmy/Zaire	L2a2	L2a	---	15,585	Kivisild <i>et al.</i> , 2006
DQ112917	70956168	AF37	Pygmy/Zaire	L0a2b	L0c2	L0c2D2	15,576	Kivisild <i>et al.</i> , 2006
DQ112918	70956170	AF98	Pygmy/Zaire	L0a2b	L0c2	L0c2D1	15,576	Kivisild <i>et al.</i> , 2006
DQ112919	70956172	AF38	Pygmy/CAR	L1c1a	L1b2	---	15,582	Kivisild <i>et al.</i> , 2006
DQ112920	70956174	AF39	Pygmy/CAR	L1c1a	L1b2	---	15,582	Kivisild <i>et al.</i> , 2006
DQ112921	70956176	AF99	Pygmy/Zaire	L5a	L0\$a	L5	15,586	Kivisild <i>et al.</i> , 2006
DQ112922	70956178	AF100	Pygmy/Zaire	L5a	L0\$a	L5	15,586	Kivisild <i>et al.</i> , 2006
DQ112923	70956180	AF40	Pygmy/CAR	L0a2a	L0c2	L0c2B1	15,574	Kivisild <i>et al.</i> , 2006
DQ112924	70956182	AF101	Pygmy/Zaire	L2a2	L2a	---	15,585	Kivisild <i>et al.</i> , 2006
DQ112925	70956184	AF96	Pygmy/Zaire	L1c1a	L1b2	---	15,582	Kivisild <i>et al.</i> , 2006
DQ112926	70956186	AF15	Ethiopia	M1	M	---	15,585	Kivisild <i>et al.</i> , 2006
DQ112932	70956198	AF42	Ethiopia	U1a	L3b\$	---	15,584	Kivisild <i>et al.</i> , 2006
DQ112933	70956200	AF43	Ethiopia	M1	M	---	15,585	Kivisild <i>et al.</i> , 2006
DQ112934	70956202	AF44	Ethiopia	L0f	L0a	L0aC	15,585	Kivisild <i>et al.</i> , 2006
DQ112949	70956232	AF50	Sudan	L3d	L3b	---	15,585	Kivisild <i>et al.</i> , 2006
DQ112956	70956246	AF45	Sudan	L3	L3b\$	---	15,583	Kivisild <i>et al.</i> , 2006
DQ112957	70956248	AF46	Sudan	L3h	L3b\$	---	15,585	Kivisild <i>et al.</i> , 2006
DQ112958	70956250	AF47	Sudan	L3i	L3a3\$	---	15,585	Kivisild <i>et al.</i> , 2006

Table G.1 Continued...

GenBank® accession number	GI number	Sequence name/ isolate	Ethnicity/ geographic origin ^a	Haplogroup/sub-haplogroup classification			No. of base pairs	Reference
				Published	Wallace, 2004	Current investigation		
DQ112959	70956252	AF48	Sudan	L0a1	L0c2\$	L0c1B	15,583	Kivisild <i>et al.</i> , 2006
DQ112960	70956254	AF49	Sudan	L3i	L3a3\$	---	15,585	Kivisild <i>et al.</i> , 2006
DQ112961	70956256	AF51	Sudan	L3h	L3b\$	---	15,585	Kivisild <i>et al.</i> , 2006
DQ112962	70956258	AF13	Tuareg/Africa	H1	L3b\$	---	15,585	Kivisild <i>et al.</i> , 2006
DQ282505	82494118	L1b-3-01	Hispanic	L1b	L1a2	---	16,567	Just <i>et al.</i> , 2008
DQ282506	82494132	L1b-3-02	Hispanic	L1b	L1a2	---	16,567	Just <i>et al.</i> , 2008
DQ282507	82494146	L3e-5-01	Hispanic	L3e	L3a3	---	16,563	Just <i>et al.</i> , 2008
DQ304897	83265615	L0a-1-01	African American	L0a	L0c1	L0c1A	16,567	Just <i>et al.</i> , 2008
DQ304898	83265629	L0a-1-02	African American	L0a	L0c1	L0c1A	16,567	Just <i>et al.</i> , 2008
DQ304899	83265643	L0a-1-03	African American	L0a	L0c1	L0c1A	16,567	Just <i>et al.</i> , 2008
DQ304900	83265657	L0a-1-04	African American	L0a	L0c1	L0c1A	16,567	Just <i>et al.</i> , 2008
DQ304901	83265671	L0a-2-01	African American	L0a	L0c1	L0c1B	16,568	Just <i>et al.</i> , 2008
DQ304902	83265685	L0a-2-02	African American	L0a	L0c1	L0c1B	16,567	Just <i>et al.</i> , 2008
DQ304903	83265699	L0a-2-03	African American	L0a	L0c1	L0c1B	16,567	Just <i>et al.</i> , 2008
DQ304904	83265713	L0a-2-04	African American	L0a	L0c1	L0c1B	16,567	Just <i>et al.</i> , 2008
DQ304905	83265727	L1b-1-01	African American	L1b	L1a2	---	16,569	Just <i>et al.</i> , 2008
DQ304906	83265741	L1b-1-02	African American	L1b	L1a2	---	16,567	Just <i>et al.</i> , 2008
DQ304907	83265755	L1b-1-03	African American	L1b	L1a2	---	16,567	Just <i>et al.</i> , 2008
DQ304908	83265769	L1b-1-04	African American	L1b	L1a2	---	16,568	Just <i>et al.</i> , 2008
DQ304909	83265783	L1b-1-05	African American	L1b	L1a2	---	16,568	Just <i>et al.</i> , 2008
DQ304910	83265797	L1b-1-06	African American	L1b	L1a2	---	16,568	Just <i>et al.</i> , 2008
DQ304911	83265811	L1b-1-07	African American	L1b	L1a2	---	16,567	Just <i>et al.</i> , 2008
DQ304912	83265825	L1b-1-08	African American	L1b	L1a2	---	16,568	Just <i>et al.</i> , 2008
DQ304913	83265839	L1b-1-09	African American	L1b	L1a2	---	16,569	Just <i>et al.</i> , 2008
DQ304914	83265853	L1b-1-10	African American	L1b	L1a2	---	16,568	Just <i>et al.</i> , 2008
DQ304915	83265867	L1b-1-11	African American	L1b	L1a2	---	16,567	Just <i>et al.</i> , 2008
DQ304916	83265881	L1b-1-12	African American	L1b	L1a2	---	16,568	Just <i>et al.</i> , 2008
DQ304917	83265895	L1b-1-13	African American	L1b	L1a2	---	16,568	Just <i>et al.</i> , 2008
DQ304918	83265909	L1b-1-14	African American	L1b	L1a2	---	16,567	Just <i>et al.</i> , 2008
DQ304919	83265923	L1b-2-01	African American	L1b	L1a2	---	16,567	Just <i>et al.</i> , 2008
DQ304920	83265937	L1b-2-02	African American	L1b	L1a2	---	16,567	Just <i>et al.</i> , 2008
DQ304921	83265951	L1b-2-03	African American	L1b	L1a2	---	16,572	Just <i>et al.</i> , 2008
DQ304922	83265965	L1b-2-04	African American	L1b	L1a2	---	16,567	Just <i>et al.</i> , 2008
DQ304923	83265979	L1b-2-05	African American	L1b	L1a2	---	16,567	Just <i>et al.</i> , 2008
DQ304924	83265993	L2a-1-01	African American	L2a	L2a	---	16,570	Just <i>et al.</i> , 2008
DQ304925	83266007	L2a-1-02	African American	L2a	L2a	---	16,574	Just <i>et al.</i> , 2008
DQ304926	83266021	L2a-1-03	African American	L2a	L2a	---	16,571	Just <i>et al.</i> , 2008
DQ304927	83266035	L2a-1-04	African American	L2a	L2a	---	16,569	Just <i>et al.</i> , 2008
DQ304928	83266049	L2a-1-05	African American	L2a	L2a	---	16,568	Just <i>et al.</i> , 2008
DQ304929	83266063	L2a-1-06	African American	L2a	L2a	---	16,567	Just <i>et al.</i> , 2008
DQ304930	83266077	L2a-1-07	African American	L2a	L2a	---	16,570	Just <i>et al.</i> , 2008
DQ304931	83266091	L2a-1-08	African American	L2a	L2a	---	16,567	Just <i>et al.</i> , 2008
DQ304932	83266105	L2a-1-09	African American	L2a	L2a	---	16,569	Just <i>et al.</i> , 2008
DQ304933	83266119	L2a-1-10	African American	L2a	L2a	---	16,571	Just <i>et al.</i> , 2008
DQ304934	83266133	L2a-2-01	African American	L2a	L2a	---	16,573	Just <i>et al.</i> , 2008

Table G.1 Continued...

GenBank® accession number	GI number	Sequence name/ isolate	Ethnicity/ geographic origin ^a	Haplogroup/sub-haplogroup classification			No. of base pairs	Reference
				Published	Wallace, 2004	Current investigation		
DQ304935	83266147	L2a-2-02	African American	L2a	L2a	---	16,571	Just <i>et al.</i> , 2008
DQ304936	83266161	L2a-2-03	African American	L2a	L2a	---	16,570	Just <i>et al.</i> , 2008
DQ304937	83266175	L2a-2-04	African American	L2a	L2a	---	16,572	Just <i>et al.</i> , 2008
DQ304938	83266189	L2a-3-01	African American	L2a	L2a	---	16,560	Just <i>et al.</i> , 2008
DQ304939	83266203	L2a-3-02	African American	L2a	L2a	---	16,570	Just <i>et al.</i> , 2008
DQ304940	83266217	L2a-3-03	African American	L2a	L2a	---	16,559	Just <i>et al.</i> , 2008
DQ304941	83266231	L2a-3-04	African American	L2a	L2a	---	16,569	Just <i>et al.</i> , 2008
DQ304942	83266245	L2a-4-01	African American	L2a	L2a	---	16,568	Just <i>et al.</i> , 2008
DQ304943	83266259	L2a-4-02	African American	L2a	L2a	---	16,568	Just <i>et al.</i> , 2008
DQ304944	83266273	L2a-4-03	African American	L2a	L2a	---	16,569	Just <i>et al.</i> , 2008
DQ304945	83266287	L2a-4-04	African American	L2a	L2a	---	16,570	Just <i>et al.</i> , 2008
DQ304946	83266301	L2a-4-05	African American	L2a	L2a	---	16,570	Just <i>et al.</i> , 2008
DQ304947	83266315	L2a-4-06	African American	L2a	L2a	---	16,570	Just <i>et al.</i> , 2008
DQ304948	83266329	L2a-4-07	African American	L2a	L2a	---	16,569	Just <i>et al.</i> , 2008
DQ304949	83266343	L2a-4-09	African American	L2a	L2a	---	16,568	Just <i>et al.</i> , 2008
DQ304950	83266357	L2a-4-10	African American	L2a	L2a	---	16,567	Just <i>et al.</i> , 2008
DQ304951	83266371	L2a-4-11	African American	L2a	L2a	---	16,568	Just <i>et al.</i> , 2008
DQ304952	83266385	L2a-5-01	African American	L2a	L2a	---	16,570	Just <i>et al.</i> , 2008
DQ304953	83266399	L2a-5-02	African American	L2a	L2a	---	16,571	Just <i>et al.</i> , 2008
DQ304954	83266413	L2a-5-03	African American	L2a	L2a	---	16,571	Just <i>et al.</i> , 2008
DQ304955	83266427	L2a-5-04	African American	L2a	L2a	---	16,570	Just <i>et al.</i> , 2008
DQ304956	83266441	L2a-5-05	African American	L2a	L2a	---	16,570	Just <i>et al.</i> , 2008
DQ304957	83266455	L2a-5-06	African American	L2a	L2a	---	16,568	Just <i>et al.</i> , 2008
DQ304958	83266469	L2a-5-07	African American	L2a	L2a	---	16,570	Just <i>et al.</i> , 2008
DQ304959	83266483	L2a-5-08	African American	L2a	L2a	---	16,571	Just <i>et al.</i> , 2008
DQ304960	83266497	L2a-5-09	African American	L2a	L2a	---	16,570	Just <i>et al.</i> , 2008
DQ304961	83266511	L2a-5-10	African American	L2a	L2a	---	16,571	Just <i>et al.</i> , 2008
DQ304962	83266525	L2a-5-11	African American	L2a	L2a	---	16,571	Just <i>et al.</i> , 2008
DQ304963	83266539	L2a-5-12	African American	L2a	L2a	---	16,570	Just <i>et al.</i> , 2008
DQ304964	83266553	L2a-5-13	African American	L2a	L2a	---	16,570	Just <i>et al.</i> , 2008
DQ304965	83266567	L2a-5-14	African American	L2a	L2a	---	16,568	Just <i>et al.</i> , 2008
DQ304966	83266581	L2a-5-15	African American	L2a	L2a	---	16,570	Just <i>et al.</i> , 2008
DQ304967	83266595	L2a-5-16	African American	L2a	L2a	---	16,571	Just <i>et al.</i> , 2008
DQ304968	83266609	L2a-6-01	African American	L2a	L2a	---	16,570	Just <i>et al.</i> , 2008
DQ304969	83266623	L2a-6-02	African American	L2a	L2a	---	16,569	Just <i>et al.</i> , 2008
DQ304970	83266637	L2a-6-03	African American	L2a	L2a	---	16,569	Just <i>et al.</i> , 2008
DQ304971	83266651	L2a-6-04	African American	L2a	L2a	---	16,571	Just <i>et al.</i> , 2008
DQ304972	83266665	L2a-6-05	African American	L2a	L2a	---	16,571	Just <i>et al.</i> , 2008
DQ304973	83266679	L2a-6-06	African American	L2a	L2a	---	16,570	Just <i>et al.</i> , 2008
DQ304974	83266693	L2a-6-07	African American	L2a	L2a	---	16,569	Just <i>et al.</i> , 2008
DQ304975	83266707	L2a-6-08	African American	L2a	L2a	---	16,570	Just <i>et al.</i> , 2008
DQ304976	83266721	L2a-6-09	African American	L2a	L2a	---	16,569	Just <i>et al.</i> , 2008
DQ304977	83266735	L2a-6-12	African American	L2a	L2a	---	16,569	Just <i>et al.</i> , 2008
DQ304978	83266749	L2b-1-01	African American	L2b	L2b	---	16,568	Just <i>et al.</i> , 2008

Table G.1 Continued...

GenBank® accession number	GI number	Sequence name/ isolate	Ethnicity/ geographic origin ^a	Haplogroup/sub-haplogroup classification			No. of base pairs	Reference
				Published	Wallace, 2004	Current investigation		
DQ304979	83266763	L2b-1-02	African American	L2b	L2b	---	16,567	Just <i>et al.</i> , 2008
DQ304980	83266777	L2b-1-03	African American	L2b	L2b	---	16,567	Just <i>et al.</i> , 2008
DQ304981	83266791	L2b-1-04	African American	L2b	L2b	---	16,567	Just <i>et al.</i> , 2008
DQ304982	83266805	L2b-1-05	African American	L2b	L2b	---	16,567	Just <i>et al.</i> , 2008
DQ304983	83266819	L2b-1-07	African American	L2b	L2b	---	16,567	Just <i>et al.</i> , 2008
DQ304984	83266833	L2b-1-08	African American	L2b	L2b	---	16,567	Just <i>et al.</i> , 2008
DQ304985	83266847	L2b-1-09	African American	L2b	L2b	---	16,567	Just <i>et al.</i> , 2008
DQ304986	83266861	L2c-1-01	African American	L2c	L2c	---	16,567	Just <i>et al.</i> , 2008
DQ304987	83266875	L2c-1-02	African American	L2c	L2c	---	16,568	Just <i>et al.</i> , 2008
DQ304988	83266889	L2c-1-03	African American	L2c	L2c	---	16,568	Just <i>et al.</i> , 2008
DQ304989	83266903	L2c-1-04	African American	L2c	L2c	---	16,567	Just <i>et al.</i> , 2008
DQ304990	83266917	L3b-1-01	African American	L3b	L3c	---	16,566	Just <i>et al.</i> , 2008
DQ304991	83266931	L3b-1-02	African American	L3b	L3c	---	16,566	Just <i>et al.</i> , 2008
DQ304992	83266945	L3b-1-03	African American	L3b	L3c	---	16,566	Just <i>et al.</i> , 2008
DQ304993	83266959	L3b-2-01	African American	L3b	L3c	---	16,566	Just <i>et al.</i> , 2008
DQ304994	83266973	L3b-2-02	African American	L3b	L3c	---	16,566	Just <i>et al.</i> , 2008
DQ304995	83266987	L3b-2-03	African American	L3b	L3c	---	16,566	Just <i>et al.</i> , 2008
DQ304996	83267001	L3b-2-04	African American	L3b	L3c	---	16,566	Just <i>et al.</i> , 2008
DQ304997	83267015	L3b-2-05	African American	L3b	L3c	---	16,566	Just <i>et al.</i> , 2008
DQ304998	83267029	L3e-1-01	African American	L3e	L3a1	---	16,570	Just <i>et al.</i> , 2008
DQ304999	83267043	L3e-1-02	African American	L3e	L3a1	---	16,572	Just <i>et al.</i> , 2008
DQ305000	83267057	L3e-1-03	African American	L3e	L3a1	---	16,571	Just <i>et al.</i> , 2008
DQ305001	83267071	L3e-1-04	African American	L3e	L3a1	---	16,571	Just <i>et al.</i> , 2008
DQ305002	83267085	L3e-1-05	African American	L3e	L3a1	---	16,571	Just <i>et al.</i> , 2008
DQ305003	83267099	L3e-1-06	African American	L3e	L3a1	---	16,570	Just <i>et al.</i> , 2008
DQ305004	83267113	L3e-1-07	African American	L3e	L3a1	---	16,570	Just <i>et al.</i> , 2008
DQ305005	83267127	L3e-1-08	African American	L3e	L3a1	---	16,571	Just <i>et al.</i> , 2008
DQ305006	83267141	L3e-1-09	African American	L3e	L3a1	---	16,570	Just <i>et al.</i> , 2008
DQ305007	83267155	L3e-1-10	African American	L3e	L3a1	---	16,572	Just <i>et al.</i> , 2008
DQ305008	83267169	L3e-1-11	African American	L3e	L3a1	---	16,572	Just <i>et al.</i> , 2008
DQ305009	83267183	L3e-1-12	African American	L3e	L3a1	---	16,571	Just <i>et al.</i> , 2008
DQ305010	83267197	L3e-2-01	African American	L3e	L3a2	---	16,567	Just <i>et al.</i> , 2008
DQ305011	83267211	L3e-2-02	African American	L3e	L3a2	---	16,567	Just <i>et al.</i> , 2008
DQ305012	83267225	L3e-2-03	African American	L3e	L3a2	---	16,572	Just <i>et al.</i> , 2008
DQ305013	83267239	L3e-2-04	African American	L3e	L3a2	---	16,568	Just <i>et al.</i> , 2008
DQ305014	83267253	L3e-2-05	African American	L3e	L3a2	---	16,572	Just <i>et al.</i> , 2008
DQ305015	83267267	L3e-2-06	African American	L3e	L3a2	---	16,567	Just <i>et al.</i> , 2008
DQ305016	83267281	L3e-2-07	African American	L3e	L3a2	---	16,568	Just <i>et al.</i> , 2008
DQ305017	83267295	L3e-2-08	African American	L3e	L3a2	---	16,567	Just <i>et al.</i> , 2008
DQ305018	83267309	L3e-3-01	African American	L3e	L3a1	---	16,572	Just <i>et al.</i> , 2008
DQ305019	83267323	L3e-3-02	African American	L3e	L3a1	---	16,571	Just <i>et al.</i> , 2008
DQ305020	83267337	L3e-3-03	African American	L3e	L3a1	---	16,572	Just <i>et al.</i> , 2008
DQ305021	83267351	L3e-4-01	African American	L3e	L3a1	---	16,570	Just <i>et al.</i> , 2008
DQ305022	83267365	L3e-4-02	African American	L3e	L3a1	---	16,572	Just <i>et al.</i> , 2008

Table G.1 Continued...

GenBank® accession number	GI number	Sequence name/ isolate	Ethnicity/ geographic origin ^a	Haplogroup/sub-haplogroup classification			No. of base pairs	Reference
				Published	Wallace, 2004	Current investigation		
DQ305023	83267379	L3e-4-03	African American	L3e	L3a1	---	16,568	Just <i>et al.</i> , 2008
DQ305024	83267393	L3e-6-01	African American	L3e	L3a1	---	16,570	Just <i>et al.</i> , 2008
DQ305025	83267407	L3e-6-02	African American	L3e	L3a1	---	16,570	Just <i>et al.</i> , 2008
DQ305026	83267421	L3e-6-03	African American	L3e	L3a1	---	16,570	Just <i>et al.</i> , 2008
DQ305027	83267435	L3e-6-04	African American	L3e	L3a1	---	16,570	Just <i>et al.</i> , 2008
DQ305028	83267449	L3e-6-05	African American	L3e	L3a1	---	16,569	Just <i>et al.</i> , 2008
DQ305029	83267463	L3e-6-06	African American	L3e	L3a1	---	16,569	Just <i>et al.</i> , 2008
DQ305030	83267477	L3e-6-07	African American	L3e	L3a1	---	16,569	Just <i>et al.</i> , 2008
DQ305031	83267491	L3e-6-08	African American	L3e	L3a1	---	16,569	Just <i>et al.</i> , 2008
DQ305032	83267505	L3f-1-01	African American	L3f	L3b\$	---	16,568	Just <i>et al.</i> , 2008
DQ305033	83267519	L3f-1-02	African American	L3f	L3b\$	---	16,569	Just <i>et al.</i> , 2008
DQ305034	83267533	L3f-1-03	African American	L3f	L3b\$	---	16,569	Just <i>et al.</i> , 2008
DQ305035	83267547	L3f-1-04	African American	L3f	L3b\$	---	16,569	Just <i>et al.</i> , 2008
DQ305036	83267561	L3f-1-05	African American	L3f	L3b\$	---	16,568	Just <i>et al.</i> , 2008
DQ341058	84682334	Tor29	Dominican Republic	L0a2a	L0c2	L0c2B2	16,558	Torrioni <i>et al.</i> , 2006
DQ341059	84682348	Tor7	Dominican Republic	L1c2a	L1b1	---	16,584	Torrioni <i>et al.</i> , 2006
DQ341060	84682362	Tor74	Ethiopia	L5	L0\$a	L5	16,570	Torrioni <i>et al.</i> , 2006
DQ341061	84682376	Tor68	Ethiopia	L5	L0\$a	L5	16,571	Torrioni <i>et al.</i> , 2006
DQ341062	84682390	Tor38	Dominican Republic	L2d	L2d\$	---	16,569	Torrioni <i>et al.</i> , 2006
DQ341063	85541074	Tor39	Ethiopia	L6	L2d\$	---	16,572	Torrioni <i>et al.</i> , 2006
DQ341064	84682418	Tor66	Ethiopia	L7	L3b\$	---	16,569	Torrioni <i>et al.</i> , 2006
DQ341065	84682432	Tor71	Ethiopia	L4	L3b\$	---	16,567	Torrioni <i>et al.</i> , 2006
DQ341066	84682446	Tor72	Ethiopia	L3x	L3a3\$	---	16,574	Torrioni <i>et al.</i> , 2006
DQ341067	84682460	Tor82	Ethiopia	L3x	L3a3\$	---	16,570	Torrioni <i>et al.</i> , 2006
DQ341068	84682474	Tor70	Ethiopia	L3i	L3a3\$	---	16,566	Torrioni <i>et al.</i> , 2006
DQ341069	84682488	Tor69	Ethiopia	L3i	L3a3\$	---	16,571	Torrioni <i>et al.</i> , 2006
DQ341070	84682502	Tor79	Nigeria	L3e5	L3a3\$	---	16,567	Torrioni <i>et al.</i> , 2006
DQ341071	84682516	Tor5	Dominican Republic	L3e	L3a1	---	16,570	Torrioni <i>et al.</i> , 2006
DQ341072	84682530	Tor4	Dominican Republic	L3d1	L3b	---	16,567	Torrioni <i>et al.</i> , 2006
DQ341073	84682544	Tor6	Dominican Republic	L3b	L3c	---	16,567	Torrioni <i>et al.</i> , 2006
DQ341074	84682558	Tor84	Ethiopia	L3c	L3b\$	---	16,568	Torrioni <i>et al.</i> , 2006
DQ341075	84682572	Tor80	Nigeria	L3f	L3b\$	---	16,568	Torrioni <i>et al.</i> , 2006
DQ341076	84682586	Tor83	Ethiopia	L3f	L3b\$	---	16,569	Torrioni <i>et al.</i> , 2006
DQ341077	84682600	Tor75	Ethiopia	L3f1	L3b\$	---	16,569	Torrioni <i>et al.</i> , 2006
DQ341078	84682614	Tor81	Nigeria	L3f1	L3b\$	---	16,568	Torrioni <i>et al.</i> , 2006
DQ341079	84682628	Tor31	Ethiopia	L3h	L3b\$	---	16,568	Torrioni <i>et al.</i> , 2006
DQ341080	84682642	Tor67	Ethiopia	L3h	L3b	---	16,567	Torrioni <i>et al.</i> , 2006
DQ341081	84682656	Tor73	Ethiopia	L3a	L3b\$	---	16,567	Torrioni <i>et al.</i> , 2006
EF184580	133854527	TZSW053_L5	Sandawe/Tanzania	L5	L0\$b	L5	16,559	Gonder <i>et al.</i> , 2007
EF184581	133854541	TZSW087_L5	Sandawe/Tanzania	L5	L0\$b	L5	16,560	Gonder <i>et al.</i> , 2007
EF184582	133854555	TZMW009_L5	Mbugwe/Tanzania	L5	L0\$a	L5	16,570	Gonder <i>et al.</i> , 2007

Table G.1 Continued...

GenBank® accession number	GI number	Sequence name/ isolate	Ethnicity/ geographic origin ^a	Haplogroup/sub-haplogroup classification			No. of base pairs	Reference
				Published	Wallace, 2004	Current investigation		
EF184583	133854569	TZSW086_L5	Sandawe/Tanzania	L5	L0\$b	L5	16,559	Gonder <i>et al.</i> , 2007
EF184584	133854583	TZSW055_L5	Sandawe/Tanzania	L5	L0\$b	L5	16,560	Gonder <i>et al.</i> , 2007
EF184585	133854597	Tzsw109_L0d	Sandawe/Tanzania	L0d	L0a	L0aB	16,567	Gonder <i>et al.</i> , 2007
EF184586	133854611	Tzsw084_L0d	Sandawe/Tanzania	L0d	L0b	L0bB1b2	16,568	Gonder <i>et al.</i> , 2007
EF184587	133854625	Tzsw068_L0d	Sandawe/Tanzania	L0d	L0a	L0aB	16,569	Gonder <i>et al.</i> , 2007
EF184588	133854639	Tzsw010_L0d	Sandawe/Tanzania	L0d	L0a	L0aB	16,569	Gonder <i>et al.</i> , 2007
EF184589	133854653	Tzbg031_L0d	Burunge/Tanzania	L0d	L2a	L0aB	16,569	Gonder <i>et al.</i> , 2007
EF184590	133854667	sanC5_L0d	San/South Africa	L0d	L0b	L0bB1b2	16,568	Gonder <i>et al.</i> , 2007
EF184591	133854681	San_54_L0d	San/South Africa	L0d	L0b	L0bA1	16,568	Gonder <i>et al.</i> , 2007
EF184592	133854695	sanC6_L0d2	San/South Africa	L0d2	L0b	L0bB2a2	16,566	Gonder <i>et al.</i> , 2007
EF184593	133854709	sanC2_L0d2	San/South Africa	L0d2	L0b	L0bB2b	16,566	Gonder <i>et al.</i> , 2007
EF184594	133854723	San_107_L0d2	San/South Africa	L0d2	L0b	L0bB2a3	16,571	Gonder <i>et al.</i> , 2007
EF184595	133854737	TZWF021_L0f	Wafiome/Tanzania	L0f	L2a	L0aC	16,568	Gonder <i>et al.</i> , 2007
EF184596	133854751	TZAK002_L0f	Akie/Tanzania	L0f	L2a	L0aC	16,568	Gonder <i>et al.</i> , 2007
EF184597	133854765	TZTR029_L0f	Turu/Tanzania	L0f	L0a	L0aC	16,570	Gonder <i>et al.</i> , 2007
EF184598	133854779	TZBG034_L0f	Burunge/Tanzania	L0f	L0a	L0aC	16,566	Gonder <i>et al.</i> , 2007
EF184599	133854793	TZDT068_L0f	Datog/Tanzania	L0f	L0,1,2	---	16,570	Gonder <i>et al.</i> , 2007
EF184600	133854807	TZGG017_L0f	Gogo/Tanzania	L0f	L0\$a	L0aC	16,568	Gonder <i>et al.</i> , 2007
EF184601	133854821	TZDT067_L0a	Datog/Tanzania	L0a	L0c1	L0c1A	16,568	Gonder <i>et al.</i> , 2007
EF184602	133854835	TZSW038_L0a	Sandawe/Tanzania	L0a	L0c2	L0c2C	16,559	Gonder <i>et al.</i> , 2007
EF184603	133854849	TZPR030_L0a	Pare/Tanzania	L0a	L0c2	L0c2B1	16,559	Gonder <i>et al.</i> , 2007
EF184604	133854863	Tzsw050_L0a	Sandawe/Tanzania	L0a	L0c2	L0c2C	16,557	Gonder <i>et al.</i> , 2007
EF184605	133854877	tztr002_L0a	Turu/Tanzania	L0a	L0c2	L0c2B2	16,559	Gonder <i>et al.</i> , 2007
EF184606	133854891	Tzsw005_L0a	Sandawe/Tanzania	L0a	L0c2	L0c2E	16,559	Gonder <i>et al.</i> , 2007
EF184607	133854905	tzsw015_L0a	Sandawe/Tanzania	L0a	L2a	L0c2E	16,558	Gonder <i>et al.</i> , 2007
EF184608	133854919	TZTR001_L0a	Turu/Tanzania	L0a	L2a	L0c2E	16,560	Gonder <i>et al.</i> , 2007
EF184609	133854933	San_102_L0k	San/South Africa	L0k	N/A	N/A	16,568	Gonder <i>et al.</i> , 2007
EF184610	133854946	San_11_L0k	San/South Africa	L0k	L0a	L0aA2	16,564	Gonder <i>et al.</i> , 2007
EF184611	133854959	sanC3_L0k	San/South Africa	L0k	L0a	L0aA1	16,568	Gonder <i>et al.</i> , 2007
EF184612	133854973	TZTR031_L1c	Turu/Tanzania	L1c	L0,L1	L1	16,567	Gonder <i>et al.</i> , 2007
EF184613	133854987	CAPL081_L1c	Bakola/Cameroon	L1c	L1b2	---	16,567	Gonder <i>et al.</i> , 2007
EF184614	133855001	CAPL079_L1c	Bakola/Cameroon	L1c	L1b2	---	16,567	Gonder <i>et al.</i> , 2007
EF184615	133855015	CAPL029_L1c	Bakola/Cameroon	L1c	L1b2	---	16,567	Gonder <i>et al.</i> , 2007
EF184616	133855029	CAPL006_L1c	Bakola/Cameroon	L1c	L1b2	---	16,567	Gonder <i>et al.</i> , 2007
EF184617	133855043	Tzbg037_L2a	Burunge/Tanzania	L2a	L2a	---	16,570	Gonder <i>et al.</i> , 2007
EF184618	133855057	Tzhz037_L2a	Hadza/Tanzania	L2a	L2a	---	16,570	Gonder <i>et al.</i> , 2007
EF184619	133855070	Tzhz077_L2a	Hadza/Tanzania	L2a	L2a	---	16,570	Gonder <i>et al.</i> , 2007
EF184620	133855083	tztr037_L2a	Turu/Tanzania	L2a	L2a	---	16,571	Gonder <i>et al.</i> , 2007
EF184621	133855097	tzms014_L2	Maasai/Tanzania	L2	L2d\$	---	16,569	Gonder <i>et al.</i> , 2007
EF184622	133855111	TZHZ251_L3h	Hadza/Tanzania	L3h	L3b\$	---	16,570	Gonder <i>et al.</i> , 2007
EF184623	133855125	Tzbg004_L3	Burunge/Tanzania	L3	L3b	---	16,568	Gonder <i>et al.</i> , 2007
EF184624	133855139	Tzhz032_L3f	Hadza/Tanzania	L3f	L3b\$	---	16,570	Gonder <i>et al.</i> , 2007
EF184625	133855153	Tzhz061_L3f	Hadza/Tanzania	L3f	L3b\$	---	16,572	Gonder <i>et al.</i> , 2007
EF184627	133855181	Tzsw011_L3g	Sandawe/Tanzania	L3g	L3b\$	---	16,570	Gonder <i>et al.</i> , 2007

Table G.1 Continued...

GenBank® accession number	GI number	Sequence name/ isolate	Ethnicity/ geographic origin ^a	Haplogroup/sub-haplogroup classification			No. of base pairs	Reference
				Published	Wallace, 2004	Current investigation		
EF184628	133855195	tztr007_L3d	Turu/Tanzania	L3d	L3b	---	16,568	Gonder <i>et al.</i> , 2007
EF184629	133855209	tztr023_L3g	Turu/Tanzania	L3g	L3b\$	---	16,570	Gonder <i>et al.</i> , 2007
EF184630	133855223	Tztr024_L3	Turu/Tanzania	L3	L3b\$	---	16,568	Gonder <i>et al.</i> , 2007
EF184631	133855237	TZRG044_L3	Rangi/Tanzania	L3	L3b\$	---	16,569	Gonder <i>et al.</i> , 2007
EF184632	133855251	TZIQ006_L3	Iraqw/Tanzania	L3	L3b\$	---	16,569	Gonder <i>et al.</i> , 2007
EF184633	133855265	sanC1_L3	San/South Africa	L3	L3b\$	---	16,570	Gonder <i>et al.</i> , 2007
EF184639	133855349	TZIQ053_L3	Iraqw/Tanzania	L3	L3b\$	---	16,570	Gonder <i>et al.</i> , 2007
EF184640	133855363	TZHZ081_L3	Hadza/Tanzania	L3	L3b\$	---	16,569	Gonder <i>et al.</i> , 2007
EF184641	133855376	TZMW031_L3	Mbugwe/ Tanzania	L3	L3b	---	16,568	Gonder <i>et al.</i> , 2007
NC_001807 ^c	17981852	---	Yoruba/Africa	---	L3a1	---	16,571	Ingman <i>et al.</i> , 2000
J01415	113200490	rCRS	European	H	---	---	16,569	Andrews <i>et al.</i> , 1999
NC_001643 ^d	5835121	Pan trog	<i>Pan troglodytes</i>	---	---	---	16,554	Horai <i>et al.</i> , 1995
---	---	585	African American	L0a1a	---	---	16,568	Howell <i>et al.</i> , 2004
---	---	586	African American	L0a1	---	---	16,568	Howell <i>et al.</i> , 2004
---	---	587	African American	L0a1a	---	---	16,569	Howell <i>et al.</i> , 2004
---	---	L1a	Dominican Republic	L1a	---	---	---	Torroni <i>et al.</i> , 2001
DQ112953	---	AS46	Asian	L0a2a	---	---	15,574	Kivisild <i>et al.</i> , 2006
EF657286	151328456	MTDNA149	African descent	L1a	---	---	15,437	Herrnstadt <i>et al.</i> , 2002a
EF657743	151334843	MTDNA560	African descent	L1a	---	---	15,446	Herrnstadt <i>et al.</i> , 2002a

^a = Zaire was the name given to the present Democratic Republic of the Congo between 27 October 1971 and 17 May 1997. ^b = this sequence, D38112, is listed on GenBank® as one of a list of other reference human mtDNAs and represents mtDNA from an African individual from Uganda. ^c = this sequence, NC_001807, is an NCBI RefSeq representing mtDNA from an African individual from Yoruba and was derived from AF347015. ^d = this sequence, NC_001643, is an NCBI RefSeq representing mtDNA from a chimpanzee, *Pan troglodytes*, and was derived from D38113. ^e = alternative names for the same ethnic group. ^f = although the term Khoi-San has been used in the current investigation, Khoisan is used here to reflect the spelling of the ethnic group as it was reported in the relevant reference. The geographic origin of the individual mitochondrial sequences is indicated using colour: ■ = Africa, ■ = Northern Africa, ■ = Western Africa, ■ = Central Africa, ■ = Eastern Africa, ■ = Southern Africa, ■ = USA (African American), ■ = Dominican Republic and ■ = USA (Hispanic). Haplogroup/sub-haplogroup classification: published = classification provided in the accompanying reference, this classification is based on the haplogroup classification scheme described in the literature (refer to Section 3.3.1.1); Wallace, 2004 = classification according to the classification scheme used in current investigation (refer to Section 4.9), the \$ sign used in this column indicates incompletely classified sequences (refer to Section 4.9, Table 4.8); current investigation = classification based on the revised classification scheme for the L0 haplogroup proposed in the current investigation (refer to Section 5.12). In the alternative classification scheme (refer to Section 3.3.1.1), L0a, L0d, L0f and L0k were previously designated as L1a, L1d, L1f and L1k, L5a was formerly identified as L1e, L4a has been described as L7 (Torroni *et al.*, 2006), and L4g was previously designated as L3g (Salas *et al.*, 2004a; Kivisild *et al.*, 2004; Tishkoff *et al.*, 2007). The sequences indicated in red represent sequences included in the global L0-specific haplogroup network (refer to Section 5.10). These sequences were included in the L0 section of the global mtDNA mutational phylogeny (Ruiz-Pesini *et al.*, 2007) from which the global L0-specific haplogroup network was constructed. The information provided for these sequences in the current investigation are therefore incomplete as the sequences were not obtained from GenBank®. Abbreviations: GI = GenInfo Identifier sequence identification number; No. = number; CAR = Central African Republic; N/A = classification not available, EF184609 could not be classified within macrohaplogroup L due to the presence of an A at nucleotide position 3,594 (refer to Section 4.9, Table 4.8).

The GenBank® accession number is a unique identifier for a sequence record which does not change even if any information within the record, such as the sequence data, is altered. The GI number, or GenInfo Identifier sequence identification number, is a number used to identify a specific nucleotide sequence. This number changes if the sequence data within a specific sequence record is updated or changed in any way. Sequence name or isolate is the name used to indicate specific sequences in phylogenetic trees, in the global

L0-specific haplogroup network, or in statistical analyses when a GenBank® accession number was not available for unique identification of the relevant sequences. Information regarding ethnicity or geographic origin is provided where available. African Americans refer to individuals of African descent born in the USA (Herrnstadt *et al.*, 2002a; Just *et al.*, 2008). These individuals likely originated from West Africa as it has been estimated previously that West Africa contributed 60% $\pm$ 3%, south-western Africa 30% $\pm$ 2% and west-central Africa 9% $\pm$ 3% to the mtDNA lineages observed today in the African American population (Salas *et al.*, 2005). Individuals from the Dominican Republic included in the current investigation are of African descent and are likely to have originated from north-western, west-central and south-eastern Africa (Fage, 1969). Information regarding the country of origin in Africa was not provided in the references for a number of ethnic groups, and their probable African country/region of origin is indicated in Table G.2.

Table G.2 Probable country or geographic region of origin for ethnic groups for which no geographic information were provided

Ethnic group	Probable country of origin in Africa	Geographic region in Africa
Ambo	Angola	Southern Africa
Bamileke	Cameroon	Central Africa
Biaka	Central African Republic	Central Africa
Effik	Nigeria	Western Africa
Ewondo	Cameroon	Central Africa
Hausa	Nigeria	Western Africa
Herero	Namibia	Southern Africa
Ibo	Nigeria	Western Africa
Khoi-San	Angola, Namibia, Botswana, South Africa	Southern Africa
Kikuyu	Kenya	Eastern Africa
Lisongo/Lissongo	Central African Republic	Central Africa
Mandenka	Senegal	Western Africa
Mbenzele	Democratic Republic of the Congo, Central African Republic, Cameroon, Gabon	Central Africa
Mbuti	Democratic Republic of the Congo	Central Africa
Mkamba	N/A	Central Africa
NaSan	N/A	N/A
Pedi	Botswana, Lesotho, South Africa	Southern Africa
San	Angola, Namibia, Botswana, South Africa	Southern Africa
Sotho	Botswana, Lesotho, South Africa	Southern Africa
Tswana	Botswana, Lesotho, South Africa	Southern Africa
Tuareg	Niger	Western Africa
Yoruba	Nigeria	Western Africa
Zulu	South Africa	Southern Africa

Information regarding probable country of origin in Africa was compiled from Yakan (1999) and Gordon (2005). N/A = information not available.

Haplogroup and sub-haplogroup classifications indicated in Table G.1 were based on three classification schemes. The published classification represents the haplogroup designation according to the alternative classification scheme (refer to Section 3.3.1.1), the Wallace classification (Wallace, 2004) represents the haplogroup assigned according to the coding region only scheme described in Section 4.9, and the current investigation classification represents the haplogroup designation according to the revised L0 haplogroup classification scheme developed in Section 5.12.

APPENDIX H

List of mitochondrial DNA sequences excluded during the compilation of different datasets

This appendix contains a list of all the mitochondrial DNA genome sequences excluded or included during the compilation of the various datasets used in the construction of the phylogenetic trees (refer to Section 5.9), the compilation of the global L0-specific haplogroup network (refer to Section 5.10) and the performance of the statistical analyses (refer to Section 5.11). The assigned mitochondrial DNA haplogroups, the relevant ethnic and geographic information available for each sequence, as well as the reason for exclusion from or inclusion in the dataset in question are presented in Table H.1.

Table H.1 List of mitochondrial genome sequences excluded from (or included in) datasets compiled for use in various analyses

GenBank® accession number/ sequence name	Ethnicity/ geographic origin	Haplogroup/ sub-haplogroup classification		Reason for exclusion/inclusion	Reference
		Published	Wallace, 2004 ^b		
A. Sequences excluded from dataset 1a and 1b during the compilation of datasets 2a and 2b					
AF381998	Jordan	L3d	L3b	Obtained from country outside of Africa	Maca-Meyer <i>et al.</i> , 2001
DQ112686	Dominican Republic	L0a2a	L0c2	Obtained from country outside of Africa	Kivisild <i>et al.</i> , 2006
DQ112687	Dominican Republic	L0a2a	L0c2	Obtained from country outside of Africa	Kivisild <i>et al.</i> , 2006
DQ112688	Dominican Republic	L0a2a	L0c2	Obtained from country outside of Africa	Kivisild <i>et al.</i> , 2006
DQ112689	Dominican Republic	L0a2a	L0c2	Obtained from country outside of Africa	Kivisild <i>et al.</i> , 2006
DQ112690	Dominican Republic	L1b1	L1a2	Obtained from country outside of Africa	Kivisild <i>et al.</i> , 2006
DQ112691	Dominican Republic	L1b1	L1a2	Obtained from country outside of Africa	Kivisild <i>et al.</i> , 2006
DQ112692	Dominican Republic	L1b1	L1a2	Obtained from country outside of Africa	Kivisild <i>et al.</i> , 2006
DQ112693	Dominican Republic	L1b1	L1a2	Obtained from country outside of Africa	Kivisild <i>et al.</i> , 2006
DQ112695	Dominican Republic	L1c1	L1b2	Obtained from country outside of Africa	Kivisild <i>et al.</i> , 2006
DQ112696	Dominican Republic	L1c3	L0,L1	Obtained from country outside of Africa	Kivisild <i>et al.</i> , 2006
DQ112697	Dominican Republic	L1c3	L0,L1	Obtained from country outside of Africa	Kivisild <i>et al.</i> , 2006
DQ112698	Dominican Republic	L2a1	L2a	Obtained from country outside of Africa	Kivisild <i>et al.</i> , 2006
DQ112699	Dominican Republic	L2a1	L2a	Obtained from country outside of Africa	Kivisild <i>et al.</i> , 2006
DQ112700	Dominican Republic	L2a1	L2a	Obtained from country outside of Africa	Kivisild <i>et al.</i> , 2006
DQ112701	Dominican Republic	L2a1	L2a	Obtained from country outside of Africa	Kivisild <i>et al.</i> , 2006
DQ112702	Dominican Republic	L2b	L2b	Obtained from country outside of Africa	Kivisild <i>et al.</i> , 2006
DQ112703	Dominican Republic	L2b	L2b	Obtained from country outside of Africa	Kivisild <i>et al.</i> , 2006

Table H.1 Continued...

GenBank® accession number/ sequence name	Ethnicity/ geographic origin	Haplogroup/ sub-haplogroup classification		Reason for exclusion/inclusion	Reference
		Published	Wallace, 2004 ^b		
DQ112704	Dominican Republic	L2b	L2b	Obtained from country outside of Africa	Kivisild <i>et al.</i> , 2006
DQ112705	Dominican Republic	L2b	L2b	Obtained from country outside of Africa	Kivisild <i>et al.</i> , 2006
DQ112706	Dominican Republic	L2c	L2c	Obtained from country outside of Africa	Kivisild <i>et al.</i> , 2006
DQ112707	Dominican Republic	L2c	L2c	Obtained from country outside of Africa	Kivisild <i>et al.</i> , 2006
DQ112708	Dominican Republic	L2d	L2d\$	Obtained from country outside of Africa	Kivisild <i>et al.</i> , 2006
DQ112709	Dominican Republic	L2d	L2d	Obtained from country outside of Africa	Kivisild <i>et al.</i> , 2006
DQ112710	Dominican Republic	L3b	L3c	Obtained from country outside of Africa	Kivisild <i>et al.</i> , 2006
DQ112711	Dominican Republic	L3b	L3c	Obtained from country outside of Africa	Kivisild <i>et al.</i> , 2006
DQ112712	Dominican Republic	L3b	L3c	Obtained from country outside of Africa	Kivisild <i>et al.</i> , 2006
DQ112713	Dominican Republic	L3d	L3b	Obtained from country outside of Africa	Kivisild <i>et al.</i> , 2006
DQ112714	Dominican Republic	L3d1	L3b	Obtained from country outside of Africa	Kivisild <i>et al.</i> , 2006
DQ112715	Dominican Republic	L3d1	L3b	Obtained from country outside of Africa	Kivisild <i>et al.</i> , 2006
DQ112716	Dominican Republic	L3d3	L3b	Obtained from country outside of Africa	Kivisild <i>et al.</i> , 2006
DQ112717	Dominican Republic	L3f1	L3b\$	Obtained from country outside of Africa	Kivisild <i>et al.</i> , 2006
DQ112718	Dominican Republic	L3h	L3b\$	Obtained from country outside of Africa	Kivisild <i>et al.</i> , 2006
DQ112719	Dominican Republic	L3e1	L3a3	Obtained from country outside of Africa	Kivisild <i>et al.</i> , 2006
DQ112720	Dominican Republic	L3e2	L3a1	Obtained from country outside of Africa	Kivisild <i>et al.</i> , 2006
DQ112721	Dominican Republic	L3e2	L3a1	Obtained from country outside of Africa	Kivisild <i>et al.</i> , 2006
DQ112722	Dominican Republic	L3e4	L3a3\$	Obtained from country outside of Africa	Kivisild <i>et al.</i> , 2006
DQ282505	Hispanic	L1b	L1a2	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ282506	Hispanic	L1b	L1a2	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ282507	Hispanic	L3e	L3a3	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304897	African American	L0a	L0c1	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304898	African American	L0a	L0c1	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304899	African American	L0a	L0c1	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304900	African American	L0a	L0c1	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304901	African American	L0a	L0c1	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304902	African American	L0a	L0c1	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304903	African American	L0a	L0c1	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304904	African American	L0a	L0c1	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304905	African American	L1b	L1a2	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304906	African American	L1b	L1a2	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304907	African American	L1b	L1a2	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304908	African American	L1b	L1a2	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304909	African American	L1b	L1a2	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304910	African American	L1b	L1a2	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304911	African American	L1b	L1a2	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304912	African American	L1b	L1a2	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304913	African American	L1b	L1a2	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304914	African American	L1b	L1a2	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304915	African American	L1b	L1a2	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304916	African American	L1b	L1a2	Obtained from country outside of Africa	Just <i>et al.</i> , 2008

Table H.1 Continued...

GenBank® accession number/ sequence name	Ethnicity/ geographic origin	Haplogroup/ sub-haplogroup classification		Reason for exclusion/inclusion	Reference
		Published	Wallace, 2004 ^b		
DQ304917	African American	L1b	L1a2	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304918	African American	L1b	L1a2	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304919	African American	L1b	L1a2	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304920	African American	L1b	L1a2	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304921	African American	L1b	L1a2	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304922	African American	L1b	L1a2	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304923	African American	L1b	L1a2	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304924	African American	L2a	L2a	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304925	African American	L2a	L2a	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304926	African American	L2a	L2a	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304927	African American	L2a	L2a	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304928	African American	L2a	L2a	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304929	African American	L2a	L2a	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304930	African American	L2a	L2a	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304931	African American	L2a	L2a	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304932	African American	L2a	L2a	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304933	African American	L2a	L2a	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304934	African American	L2a	L2a	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304935	African American	L2a	L2a	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304936	African American	L2a	L2a	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304937	African American	L2a	L2a	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304938	African American	L2a	L2a	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304939	African American	L2a	L2a	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304940	African American	L2a	L2a	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304941	African American	L2a	L2a	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304942	African American	L2a	L2a	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304943	African American	L2a	L2a	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304944	African American	L2a	L2a	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304945	African American	L2a	L2a	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304946	African American	L2a	L2a	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304947	African American	L2a	L2a	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304948	African American	L2a	L2a	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304949	African American	L2a	L2a	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304950	African American	L2a	L2a	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304951	African American	L2a	L2a	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304952	African American	L2a	L2a	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304953	African American	L2a	L2a	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304954	African American	L2a	L2a	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304955	African American	L2a	L2a	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304956	African American	L2a	L2a	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304957	African American	L2a	L2a	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304958	African American	L2a	L2a	Obtained from country outside of Africa	Just <i>et al.</i> , 2008

Table H.1 Continued...

GenBank® accession number/ sequence name	Ethnicity/ geographic origin	Haplogroup/ sub-haplogroup classification		Reason for exclusion/inclusion	Reference
		Published	Wallace, 2004 ^b		
DQ304959	African American	L2a	L2a	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304960	African American	L2a	L2a	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304961	African American	L2a	L2a	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304962	African American	L2a	L2a	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304963	African American	L2a	L2a	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304964	African American	L2a	L2a	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304965	African American	L2a	L2a	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304966	African American	L2a	L2a	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304967	African American	L2a	L2a	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304968	African American	L2a	L2a	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304969	African American	L2a	L2a	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304970	African American	L2a	L2a	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304971	African American	L2a	L2a	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304972	African American	L2a	L2a	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304973	African American	L2a	L2a	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304974	African American	L2a	L2a	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304975	African American	L2a	L2a	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304976	African American	L2a	L2a	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304977	African American	L2a	L2a	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304978	African American	L2b	L2b	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304979	African American	L2b	L2b	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304980	African American	L2b	L2b	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304981	African American	L2b	L2b	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304982	African American	L2b	L2b	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304983	African American	L2b	L2b	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304984	African American	L2b	L2b	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304985	African American	L2b	L2b	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304986	African American	L2c	L2c	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304987	African American	L2c	L2c	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304988	African American	L2c	L2c	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304989	African American	L2c	L2c	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304990	African American	L3b	L3c	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304991	African American	L3b	L3c	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304992	African American	L3b	L3c	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304993	African American	L3b	L3c	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304994	African American	L3b	L3c	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304995	African American	L3b	L3c	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304996	African American	L3b	L3c	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304997	African American	L3b	L3c	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304998	African American	L3e	L3a1	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ304999	African American	L3e	L3a1	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ305000	African American	L3e	L3a1	Obtained from country outside of Africa	Just <i>et al.</i> , 2008

Table H.1 Continued...

GenBank® accession number/ sequence name	Ethnicity/ geographic origin	Haplogroup/ sub-haplogroup classification		Reason for exclusion/inclusion	Reference
		Published	Wallace, 2004 ^b		
DQ305001	African American	L3e	L3a1	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ305002	African American	L3e	L3a1	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ305003	African American	L3e	L3a1	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ305004	African American	L3e	L3a1	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ305005	African American	L3e	L3a1	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ305006	African American	L3e	L3a1	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ305007	African American	L3e	L3a1	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ305008	African American	L3e	L3a1	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ305009	African American	L3e	L3a1	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ305010	African American	L3e	L3a2	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ305011	African American	L3e	L3a2	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ305012	African American	L3e	L3a2	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ305013	African American	L3e	L3a2	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ305014	African American	L3e	L3a2	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ305015	African American	L3e	L3a2	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ305016	African American	L3e	L3a2	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ305017	African American	L3e	L3a2	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ305018	African American	L3e	L3a1	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ305019	African American	L3e	L3a1	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ305020	African American	L3e	L3a1	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ305021	African American	L3e	L3a1	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ305022	African American	L3e	L3a1	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ305023	African American	L3e	L3a1	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ305024	African American	L3e	L3a1	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ305025	African American	L3e	L3a1	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ305026	African American	L3e	L3a1	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ305027	African American	L3e	L3a1	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ305028	African American	L3e	L3a1	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ305029	African American	L3e	L3a1	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ305030	African American	L3e	L3a1	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ305031	African American	L3e	L3a1	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ305032	African American	L3f	L3b\$	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ305033	African American	L3f	L3b\$	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ305034	African American	L3f	L3b\$	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ305035	African American	L3f	L3b\$	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ305036	African American	L3f	L3b\$	Obtained from country outside of Africa	Just <i>et al.</i> , 2008
DQ341058	Dominican Republic	L0a2a	L0c2	Obtained from country outside of Africa	Torrioni <i>et al.</i> , 2006
DQ341059	Dominican Republic	L1c2a	L1b1	Obtained from country outside of Africa	Torrioni <i>et al.</i> , 2006
DQ341062	Dominican Republic	L2d	L2d\$	Obtained from country outside of Africa	Torrioni <i>et al.</i> , 2006
DQ341071	Dominican Republic	L3e	L3a1	Obtained from country outside of Africa	Torrioni <i>et al.</i> , 2006
DQ341072	Dominican Republic	L3d1	L3b	Obtained from country outside of Africa	Torrioni <i>et al.</i> , 2006
DQ341073	Dominican Republic	L3b	L3c	Obtained from country outside of Africa	Torrioni <i>et al.</i> , 2006

Table H.1 Continued...

GenBank® accession number/sequence name	Ethnicity/ geographic origin	Haplogroup/ sub-haplogroup classification		Reason for exclusion/inclusion	Reference
		Published	Wallace, 2004 ^b		
B. Sequences excluded from datasets 1a and 2a during the compilation of datasets 1c and 2c					
AF381992	Mauritania	L1c	L0,L1	Incomplete classification	Maca-Meyer <i>et al.</i> , 2001
DQ112696	Dominican Republic	L1c3	L0,L1	Incomplete classification	Kivisild <i>et al.</i> , 2006
DQ112697	Dominican Republic	L1c3	L0,L1	Incomplete classification	Kivisild <i>et al.</i> , 2006
DQ112921	Pygmy/Zaire ^a	L5a	L0\$a	Incomplete classification	Kivisild <i>et al.</i> , 2006
DQ112922	Pygmy/Zaire ^a	L5a	L0\$a	Incomplete classification	Kivisild <i>et al.</i> , 2006
DQ341060	Ethiopia	L5	L0\$a	Incomplete classification	Torroni <i>et al.</i> , 2006
DQ341061	Ethiopia	L5	L0\$a	Incomplete classification	Torroni <i>et al.</i> , 2006
EF184580	Sandawe/Tanzania	L5	L0\$b	Incomplete classification	Gonder <i>et al.</i> , 2007
EF184581	Sandawe/Tanzania	L5	L0\$b	Incomplete classification	Gonder <i>et al.</i> , 2007
EF184582	Mbugwe/Tanzania	L5	L0\$a	Incomplete classification	Gonder <i>et al.</i> , 2007
EF184583	Sandawe/Tanzania	L5	L0\$b	Incomplete classification	Gonder <i>et al.</i> , 2007
EF184584	Sandawe/Tanzania	L5	L0\$b	Incomplete classification	Gonder <i>et al.</i> , 2007
EF184600	Gogo/Tanzania	L0f	L0\$a	Incomplete classification	Gonder <i>et al.</i> , 2007
EF184612	Turu/Tanzania	L1c	L0,L1	Incomplete classification	Gonder <i>et al.</i> , 2007
EF184589	Burunge/Tanzania	L0d	L2a	Discrepancy between reported haplogroup and assigned haplogroup ^c	Gonder <i>et al.</i> , 2007
EF184595	Wafiome/Tanzania	L0f	L2a	Discrepancy between reported haplogroup and assigned haplogroup ^c	Gonder <i>et al.</i> , 2007
EF184596	Akie/Tanzania	L0f	L2a	Discrepancy between reported haplogroup and assigned haplogroup ^c	Gonder <i>et al.</i> , 2007
EF184607	Sandawe/Tanzania	L0a	L2a	Discrepancy between reported haplogroup and assigned haplogroup ^c	Gonder <i>et al.</i> , 2007
EF184608	Turu/Tanzania	L0a	L2a	Discrepancy between reported haplogroup and assigned haplogroup ^c	Gonder <i>et al.</i> , 2007
EF184614	Bakola/Cameroon	L1c	L2d\$	Discrepancy between reported haplogroup and assigned haplogroup ^d	Gonder <i>et al.</i> , 2007
DQ304934	African American	L2a	L2a	Sequence contains N nucleotides	Just <i>et al.</i> , 2008
DQ305013	African American	L3e	L3a2	Sequence contains N nucleotides	Just <i>et al.</i> , 2008
DQ305022	African American	L3e	L3a1	Sequence contains N nucleotides	Just <i>et al.</i> , 2008
EF184591	San/South Africa	L0d	L0b	Sequence contains N nucleotides	Gonder <i>et al.</i> , 2007
EF184593	San/South Africa	L0d2	L0b	Sequence contains N nucleotides	Gonder <i>et al.</i> , 2007
EF184594	San/South Africa	L0d2	L0b	Sequence contains N nucleotides	Gonder <i>et al.</i> , 2007
EF184597	Turu/Tanzania	L0f	L0a	Sequence contains N nucleotides	Gonder <i>et al.</i> , 2007
EF184598	Burunge/Tanzania	L0f	L0a	Sequence contains N nucleotides	Gonder <i>et al.</i> , 2007
EF184601	Datog/Tanzania	L0a	L0c1	Sequence contains N nucleotides	Gonder <i>et al.</i> , 2007
EF184602	Sandawe/Tanzania	L0a	L0c2	Sequence contains N nucleotides	Gonder <i>et al.</i> , 2007
EF184603	Pare/Tanzania	L0a	L0c2	Sequence contains N nucleotides	Gonder <i>et al.</i> , 2007
EF184605	Turu/Tanzania	L0a	L0c2	Sequence contains N nucleotides	Gonder <i>et al.</i> , 2007
EF184606	Sandawe/Tanzania	L0a	L0c2	Sequence contains N nucleotides	Gonder <i>et al.</i> , 2007
EF184610	San/South Africa	L0k	L0a	Sequence contains N nucleotides	Gonder <i>et al.</i> , 2007
EF184613	Bakola/Cameroon	L1c	L1b2	Sequence contains N nucleotides	Gonder <i>et al.</i> , 2007
EF184615	Bakola/Cameroon	L1c	L1b2	Sequence contains N nucleotides	Gonder <i>et al.</i> , 2007

Table H.1 Continued...

GenBank® accession number/ sequence name	Ethnicity/ geographic origin	Haplogroup/ sub-haplogroup classification		Reason for exclusion/inclusion	Reference
		Published	Wallace, 2004 ^b		
EF184616	Bakola/Cameroon	L1c	L1b2	Sequence contains N nucleotides	Gonder <i>et al.</i> , 2007
EF184618	Hadza/Tanzania	L2a	L2a	Sequence contains N nucleotides	Gonder <i>et al.</i> , 2007
EF184619	Hadza/Tanzania	L2a	L2a	Sequence contains N nucleotides	Gonder <i>et al.</i> , 2007
EF184620	Turu/Tanzania	L2a	L2a	Sequence contains N nucleotides	Gonder <i>et al.</i> , 2007
EF184621	Maasai/Tanzania	L2	L2d\$	Sequence contains N nucleotides	Gonder <i>et al.</i> , 2007
EF184622	Hadza/Tanzania	L3h	L3b\$	Sequence contains N nucleotides	Gonder <i>et al.</i> , 2007
EF184623	Burunge/Tanzania	L3	L3b	Sequence contains N nucleotides	Gonder <i>et al.</i> , 2007
EF184624	Hadza/Tanzania	L3f	L3b\$	Sequence contains N nucleotides	Gonder <i>et al.</i> , 2007
EF184625	Hadza/Tanzania	L3f	L3b\$	Sequence contains N nucleotides	Gonder <i>et al.</i> , 2007
EF184627	Sandawe/Tanzania	L3g ^h	L3b\$	Sequence contains N nucleotides	Gonder <i>et al.</i> , 2007
EF184629	Turu/Tanzania	L3g ^h	L3b\$	Sequence contains N nucleotides	Gonder <i>et al.</i> , 2007
EF184630	Turu/Tanzania	L3	L3b\$	Sequence contains N nucleotides	Gonder <i>et al.</i> , 2007
EF184639	Iraqw/Tanzania	L3	L3b\$	Sequence contains N nucleotides	Gonder <i>et al.</i> , 2007
EF184640	Hadza/Tanzania	L3	L3b\$	Sequence contains N nucleotides	Gonder <i>et al.</i> , 2007
EF184641	Mbugwe/Tanzania	L3	L3b	Sequence contains N nucleotides	Gonder <i>et al.</i> , 2007
DQ112740	Foulbe/Burkina Faso	L3e2	L3a1	Sequence contains R nucleotides	Kivisild <i>et al.</i> , 2006
DQ112883	Mandenka/Africa	L2c	L2c	Sequence contains R nucleotides	Kivisild <i>et al.</i> , 2006
DQ304931	African American	L2a	L2a	Sequence contains R nucleotides	Just <i>et al.</i> , 2008
DQ304973	African American	L2a	L2a	Sequence contains R nucleotides	Just <i>et al.</i> , 2008
DQ304994	African American	L3b	L3c	Sequence contains R nucleotides	Just <i>et al.</i> , 2008
DQ304996	African American	L3b	L3c	Sequence contains R nucleotides	Just <i>et al.</i> , 2008
DQ305005	African American	L3e	L3a1	Sequence contains R nucleotides	Just <i>et al.</i> , 2008
DQ305006	African American	L3e	L3a1	Sequence contains R nucleotides	Just <i>et al.</i> , 2008
DQ305016	African American	L3e	L3a2	Sequence contains R nucleotides	Just <i>et al.</i> , 2008
DQ112748	Mossi/Burkina Faso	L3b	L3c	Sequence contains Y nucleotides	Kivisild <i>et al.</i> , 2006
DQ304912	African American	L1b	L1a2	Sequence contains Y nucleotides	Just <i>et al.</i> , 2008
DQ304939	African American	L2a	L2a	Sequence contains Y nucleotides	Just <i>et al.</i> , 2008
DQ305018	African American	L3e	L3a1	Sequence contains Y nucleotides	Just <i>et al.</i> , 2008
DQ305033	African American	L3f	L3b\$	Sequence contains Y nucleotides	Just <i>et al.</i> , 2008
DQ304901	African American	L0a	L0c1	Sequence contains M nucleotides	Just <i>et al.</i> , 2008

Table H.1 Continued...

GenBank® accession number/ sequence name	Ethnicity/ geographic origin	Haplogroup/ sub-haplogroup classification		Reason for exclusion/inclusion	Reference
		Published	Wallace, 2004 ^b		
C. Sequences assigned to the L3 haplogroup which were excluded from dataset 5					
DQ112796	Lissongo/Africa	L4g	L3b\$	Discrepancy between reported haplogroup and assigned haplogroup	Kivisild <i>et al.</i> , 2006
DQ112845	Khoisan ⁱ /Africa	L4g	L3b\$	Discrepancy between reported haplogroup and assigned haplogroup	Kivisild <i>et al.</i> , 2006
DQ341064	Ethiopia	L7 ^h	L3b\$	Discrepancy between reported haplogroup and assigned haplogroup	Torroni <i>et al.</i> , 2006
DQ341065	Ethiopia	L4	L3b\$	Discrepancy between reported haplogroup and assigned haplogroup	Torroni <i>et al.</i> , 2006
EF184627	Sandawe/Tanzania	L3g ^h	L3b\$	Discrepancy between reported haplogroup and assigned haplogroup ^e	Gonder <i>et al.</i> , 2007
EF184629	Turu/Tanzania	L3g ^h	L3b\$	Discrepancy between reported haplogroup and assigned haplogroup ^e	Gonder <i>et al.</i> , 2007
EF184633	San/South Africa	L3	L3b\$	Discrepancy between reported haplogroup and location of sequence in global phylogenetic trees ^f	Gonder <i>et al.</i> , 2007
EF184639	Iraqw/Tanzania	L3	L3b\$	Discrepancy between reported haplogroup and location of sequence in global phylogenetic trees ^f	Gonder <i>et al.</i> , 2007
EF184640	Hadza/Tanzania	L3	L3b\$	Discrepancy between reported haplogroup and location of sequence in global phylogenetic trees ^f	Gonder <i>et al.</i> , 2007
D. Sequences assigned to the L2 haplogroup which were excluded from dataset 6					
EF184595	Wafiome/Tanzania	L0f	L2a	Discrepancy between reported haplogroup and assigned haplogroup ^c	Gonder <i>et al.</i> , 2007
EF184596	Akie/Tanzania	L0f	L2a	Discrepancy between reported haplogroup and assigned haplogroup ^c	Gonder <i>et al.</i> , 2007
EF184607	Sandawe/Tanzania	L0a	L2a	Discrepancy between reported haplogroup and assigned haplogroup ^c	Gonder <i>et al.</i> , 2007
EF184608	Turu/Tanzania	L0a	L2a	Discrepancy between reported haplogroup and assigned haplogroup ^c	Gonder <i>et al.</i> , 2007
EF184589	Burunge/Tanzania	L0d	L2a	Discrepancy between reported haplogroup and assigned haplogroup ^c	Gonder <i>et al.</i> , 2007
DQ341063	Ethiopia	L6	L2d\$	Discrepancy between reported haplogroup and assigned haplogroup	Torroni <i>et al.</i> , 2006
E. Sequences assigned to other haplogroups which were included in dataset 7					
AF381992	Mauritania	L1c	L0,L1	Location of sequence in global phylogenetic trees	Maca-Meyer <i>et al.</i> , 2001
DQ112696	Dominican Republic	L1c3	L0,L1	Location of sequence in global phylogenetic trees	Kivisild <i>et al.</i> , 2006
DQ112697	Dominican Republic	L1c3	L0,L1	Location of sequence in global phylogenetic trees	Kivisild <i>et al.</i> , 2006
EF184612	Turu/Tanzania	L1c	L0,L1	Location of sequence in global phylogenetic trees	Gonder <i>et al.</i> , 2007

Table H.1 Continued...

GenBank® accession number/ sequence name	Ethnicity/ geographic origin	Haplogroup/ sub-haplogroup classification		Reason for exclusion/inclusion	Reference
		Published	Wallace, 2004 ^b		
F. Sequences assigned to the L0 haplogroup which were excluded from dataset 8a					
DQ112921	Pygmy/Zaire ^a	L5a	L0\$a	Discrepancy between reported haplogroup and assigned haplogroup	Kivisild <i>et al.</i> , 2006
DQ112922	Pygmy/Zaire ^a	L5a	L0\$a	Discrepancy between reported haplogroup and assigned haplogroup	Kivisild <i>et al.</i> , 2006
DQ341060	Ethiopia	L5	L0\$a	Discrepancy between reported haplogroup and assigned haplogroup	Torrioni <i>et al.</i> , 2006
DQ341061	Ethiopia	L5	L0\$a	Discrepancy between reported haplogroup and assigned haplogroup	Torrioni <i>et al.</i> , 2006
EF184580	Sandawe/Tanzania	L5	L0\$b	Discrepancy between reported haplogroup and assigned haplogroup	Gonder <i>et al.</i> , 2007
EF184581	Sandawe/Tanzania	L5	L0\$b	Discrepancy between reported haplogroup and assigned haplogroup	Gonder <i>et al.</i> , 2007
EF184582	Mbugwe/Tanzania	L5	L0\$a	Discrepancy between reported haplogroup and assigned haplogroup	Gonder <i>et al.</i> , 2007
EF184583	Sandawe/Tanzania	L5	L0\$b	Discrepancy between reported haplogroup and assigned haplogroup	Gonder <i>et al.</i> , 2007
EF184584	Sandawe/Tanzania	L5	L0\$b	Discrepancy between reported haplogroup and assigned haplogroup	Gonder <i>et al.</i> , 2007
G. Sequences assigned to other haplogroups which were included in dataset 8a					
EF184589	Burunge/Tanzania	L0d	L2a	Location of sequence in global phylogenetic trees ^c	Gonder <i>et al.</i> , 2007
EF184595	Wafiome/Tanzania	L0f	L2a	Location of sequence in global phylogenetic trees ^c	Gonder <i>et al.</i> , 2007
EF184596	Akie/Tanzania	L0f	L2a	Location of sequence in global phylogenetic trees ^c	Gonder <i>et al.</i> , 2007
EF184607	Sandawe/Tanzania	L0a	L2a	Location of sequence in global phylogenetic trees ^c	Gonder <i>et al.</i> , 2007
EF184608	Turu/Tanzania	L0a	L2a	Location of sequence in global phylogenetic trees ^c	Gonder <i>et al.</i> , 2007
H. Sequences excluded from dataset 8a during the compilation of dataset 8b					
DQ112959	Sudan	L0a1	L0c2\$	Inconsistency between observed alterations and haplogroup classification	Kivisild <i>et al.</i> , 2006
EF184589	Burunge/Tanzania	L0d	L2a	Inconsistency between observed alterations and haplogroup classification	Gonder <i>et al.</i> , 2007
EF184591	San/South Africa	L0d	L0b	Inconsistency between observed alterations and haplogroup classification	Gonder <i>et al.</i> , 2007
EF184595	Wafiome/Tanzania	L0f	L2a	Inconsistency between observed alterations and haplogroup classification	Gonder <i>et al.</i> , 2007
EF184596	Akie/Tanzania	L0f	L2a	Inconsistency between observed alterations and haplogroup classification	Gonder <i>et al.</i> , 2007
EF184598	Burunge/Tanzania	L0f	L0a	Inconsistency between observed alterations and haplogroup classification	Gonder <i>et al.</i> , 2007
EF184600	Gogo/Tanzania	L0f	L0\$a	Inconsistency between observed alterations and haplogroup classification	Gonder <i>et al.</i> , 2007
EF184601	Datog/Tanzania	L0a	L0c1	Inconsistency between observed alterations and haplogroup classification	Gonder <i>et al.</i> , 2007
EF184607	Sandawe/Tanzania	L0a	L2a	Inconsistency between observed alterations and haplogroup classification	Gonder <i>et al.</i> , 2007

Table H.1 Continued...

GenBank® accession number/sequence name	Ethnicity/geographic origin	Haplogroup/sub-haplogroup classification		Reason for exclusion/inclusion	Reference
		Published	Wallace, 2004 ^b		
EF184608	Turu/Tanzania	L0a	L2a	Inconsistency between observed alterations and haplogroup classification	Gonder <i>et al.</i> , 2007
I. Sequences present in L0 haplogroup network (Ruiz-Pesini <i>et al.</i>, 2007) which were included in dataset 8b					
585	African American	L0a1a	---	Present in L0 haplogroup network ^g	Howell <i>et al.</i> , 2004
586	African American	L0a1	---	Present in L0 haplogroup network ^g	Howell <i>et al.</i> , 2004
587	African American	L0a1a	---	Present in L0 haplogroup network ^g	Howell <i>et al.</i> , 2004
L1a	Dominican Republic	L1a ^h	---	Present in L0 haplogroup network ^g	Torroni <i>et al.</i> , 2001
DQ112953	Asian	L0a2a	---	Present in L0 haplogroup network ^g	Kivisild <i>et al.</i> , 2006
EF657286	African descent	L1a ^h	---	Present in L0 haplogroup network ^g	Herrnstadt <i>et al.</i> , 2002a
EF657743	African descent	L1a ^h	---	Present in L0 haplogroup network ^g	Herrnstadt <i>et al.</i> , 2002a

^a = Zaire was the name given to the present Democratic Republic of the Congo between 27 October 1971 and 17 May 1997.

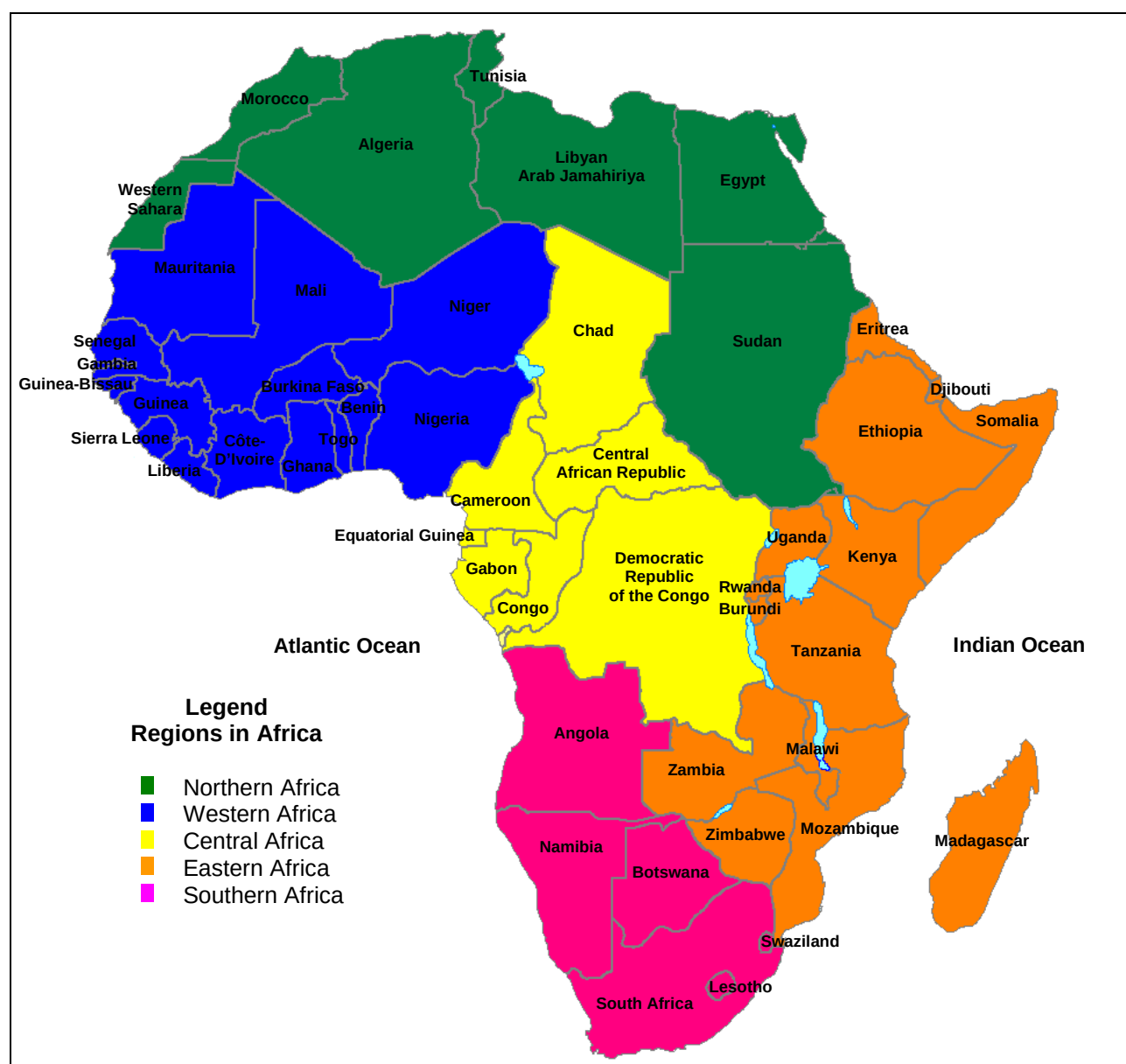
^b = sequences were assigned haplogroup designations based on the classification scheme described in Section 4.9, the \$ sign used in the Wallace column of the haplogroup/sub-haplogroup classification indicates incompletely classified sequences (refer to Section 4.9, Table 4.8). ^c = the sequences were classified as L2a according to the haplogroup classification scheme used in the current investigation (refer to Section 4.9), although the location of the sequences in the global phylogenetic trees indicated that these sequences should be classified as L0, and the classification of these L2a sequences as L0 was subsequently confirmed during re-classification of the sequences using the revised classification scheme presented in Section 5.12. ^d = EF184614 was initially erroneously classified as L2d\$, but was later re-classified as L1b2 (refer to Appendix G). ^e = L3g has been renamed L4g which forms part of the L4 haplogroup (Kivisild *et al.*, 2004). ^f = these sequences grouped with L4g sequences in the global phylogenetic trees (refer to Section 5.9) and were therefore not viewed as belonging to the L3 haplogroup. ^g = L0 haplogroup section of the mutational phylogenetic tree available on MITOMAP (Ruiz-Pesini *et al.*, 2007; refer to Appendix K). ^h = in the classification scheme generally used in the literature (refer to Section 3.3.1.1), L0a, L0d, L0f and L0k were previously designated as L1a, L1d, L1f and L1k, L5a was formerly identified as L1e, L4a has been described as L7 (Torroni *et al.*, 2006), and L4g was previously designated as L3g (Salas *et al.*, 2004a; Kivisild *et al.*, 2004; Tishkoff *et al.*, 2007). ⁱ = although the term Khoi-San has been used in the current investigation, Khoisan is used here to reflect the spelling of the ethnic group as it was reported in the relevant reference. Nucleotide abbreviations: N = A, C, G or T; R = A or G; Y = C or T; M = A or C.

APPENDIX I

Countries located in the different geographic regions of Africa

This appendix contains a map which indicates the main geographic regions in Africa according to the United Nations Cartographic Section. The map is presented in Figure I.1.

Figure I.1 Geographic regions in Africa



The five different geographic regions in Africa are indicated using colour as stipulated in the legend and the colours correspond to the colours used in the discussion of the geographic distribution of African mitochondrial haplogroups (refer to Section 5.9.4). The map of Africa was constructed by adapting United Nations, map number 4045, revision 4, January 2004 (africa.pdf). This map was compiled by the United Nations Cartographic Section and was obtained from the United Nations website (<http://www.un.org/Depts/Cartographic/english/htmain.htm>, 2008).

According to the United Nations, Africa is divided into five geographic regions, namely Northern Africa, Western Africa, Central Africa, Eastern Africa and Southern Africa. The different geographic regions are indicated in different colours as described in the legend. It is important to note that Angola is classified as belonging to Central Africa in United Nations map number 4045 from which the map indicated in Figure I.1 was derived. In contrast, most studies of mtDNA to date classify Angola as part of Southern Africa (Chen *et al.*, 2000; Plaza *et al.*, 2004). In the current investigation, Angola is therefore also classified as belonging to Southern Africa.

APPENDIX J

Global and Africa phylogenetic trees

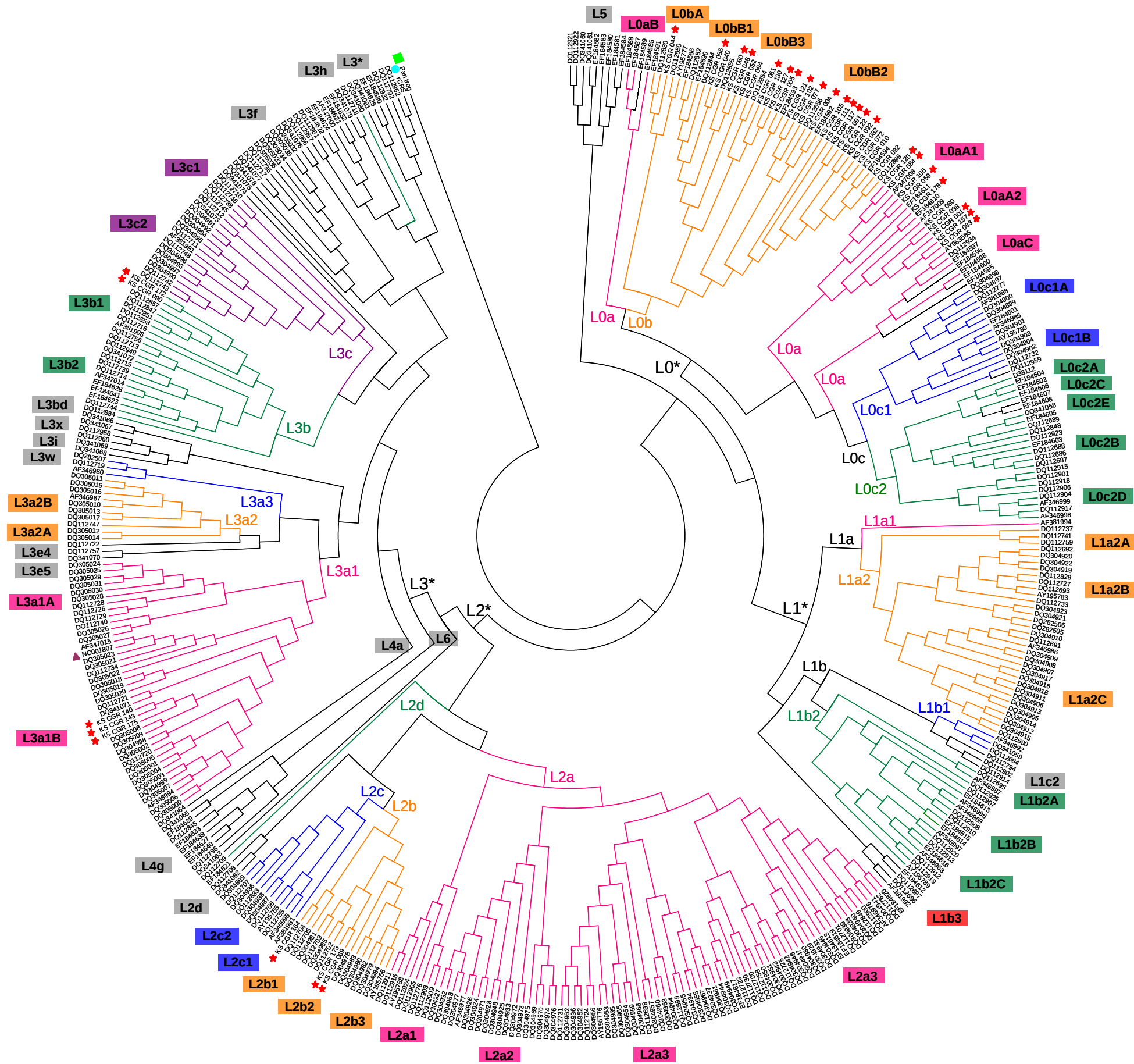
This appendix contains A3 size copies of the global and Africa phylogenetic trees constructed in the current investigation using PHYLIP version 3.66 or MEGA version 4 (refer to Section 5.9). The global phylogenetic trees were constructed by performing the neighbour-joining or maximum parsimony method on dataset 1a (All L Sequences) which consisted of 429 macrohaplogroup L sequences, the rCRS and the *Pan troglodytes* outgroup sequence. The Africa phylogenetic trees were constructed performing the neighbour-joining or maximum parsimony methods on dataset 2a (All Africa L Sequences) which contained 242 sequences obtained from individuals from African origin reported to be residing in countries located in Africa, the rCRS and the *Pan troglodytes* outgroup sequence. The global and Africa phylogenetic trees are presented in Figure J.1 (see figure legend on next page).

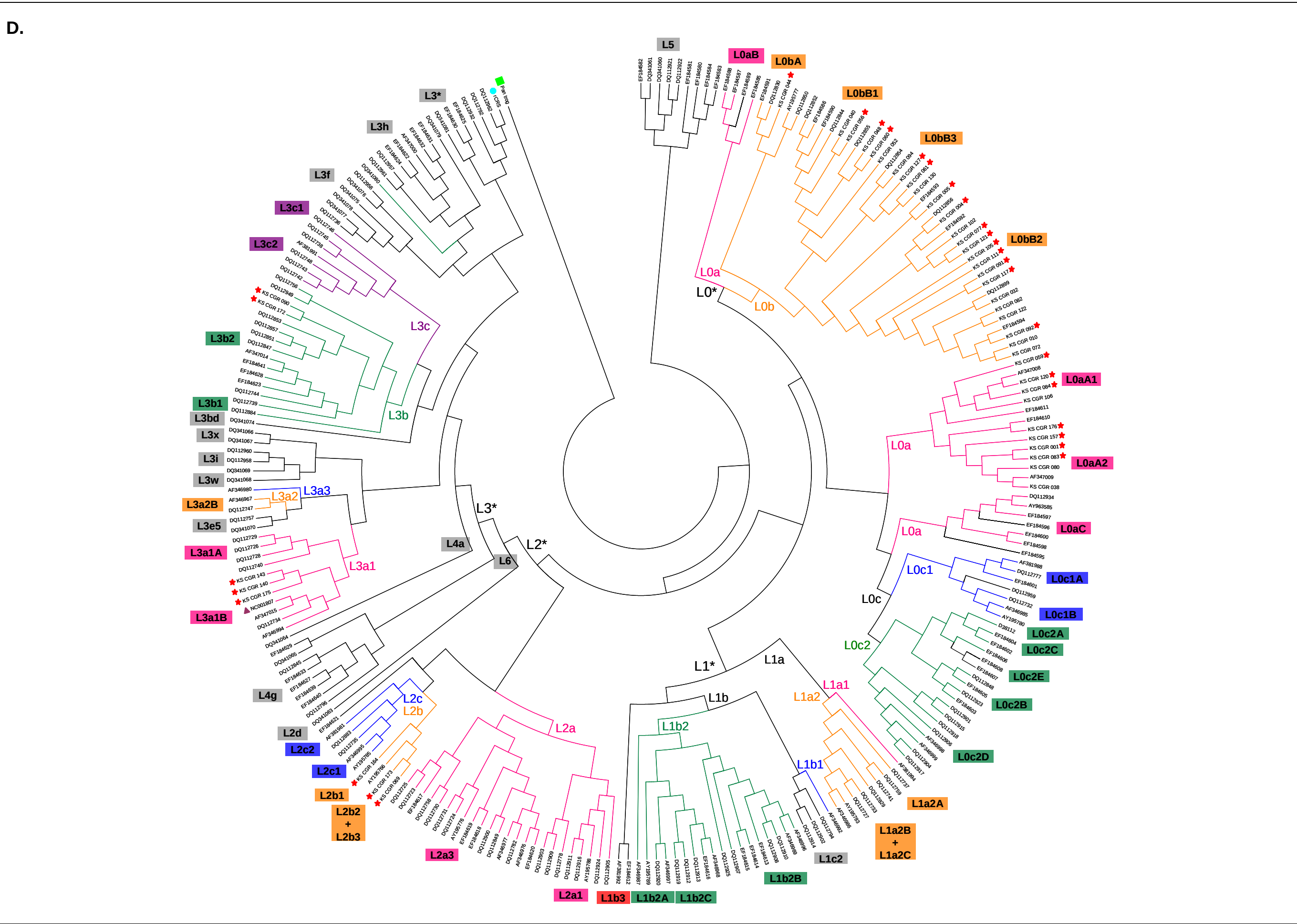
Figure J.1 Global and Africa phylogenetic trees

- A. Global neighbour-joining tree of 429 mitochondrial genome sequences.
- B. Global maximum parsimony tree of 429 mitochondrial genome sequences.
- C. Africa neighbour-joining tree of 242 mitochondrial genome sequences.
- D. Africa maximum parsimony tree of 242 mitochondrial genome sequences.

All the relevant information regarding the mitochondrial sequences used in the construction of the phylogenetic trees is listed in Appendix G. Khoi-San sequences generated in the current investigation are indicated by red stars (★), the rCRS is represented by a turquoise octagon (●), the *Pan troglodytes* outgroup sequence is denoted by a green square (■) and the African reference sequence is indicated by a violet triangle (▲). Statistical support for the observed phylogeny was determined by performing 1,000 bootstrap replicates, and the bootstrap values for the major branches and groups are indicated. Haplogroup classification was performed as described in Section 4.9. The major haplogroups and sub-haplogroups are indicated at the relevant branching points, and sub-haplogroups are distinguished through the use of different colours (■ = L0a, L1a1, L2a, L3a1; ■ = L0b, L1a2, L2b, L3a2; ■ = L0c1, L1b1, L2c, L3a3; ■ = L0c2, L1b2, L2d, L3b; ■ = L3c). Sub-haplogroups presented in coloured boxes are subdivisions identified in sub-haplogroups described in the classification scheme used in the current investigation (refer to Section 4.9), sub-haplogroups indicated in grey boxes represent sub-haplogroups assigned based on the alternative classification scheme (refer to Section 3.3.1.1), and sub-haplogroups indicated in red boxes are new sub-haplogroups which have not previously been assigned designations in the classification scheme used in the current investigation. Sequences that could not be classified completely are indicated by black lines. Sequences that show a discrepancy between their location in the phylogenetic tree and their haplogroup classification are indicated by dotted lines.

B.





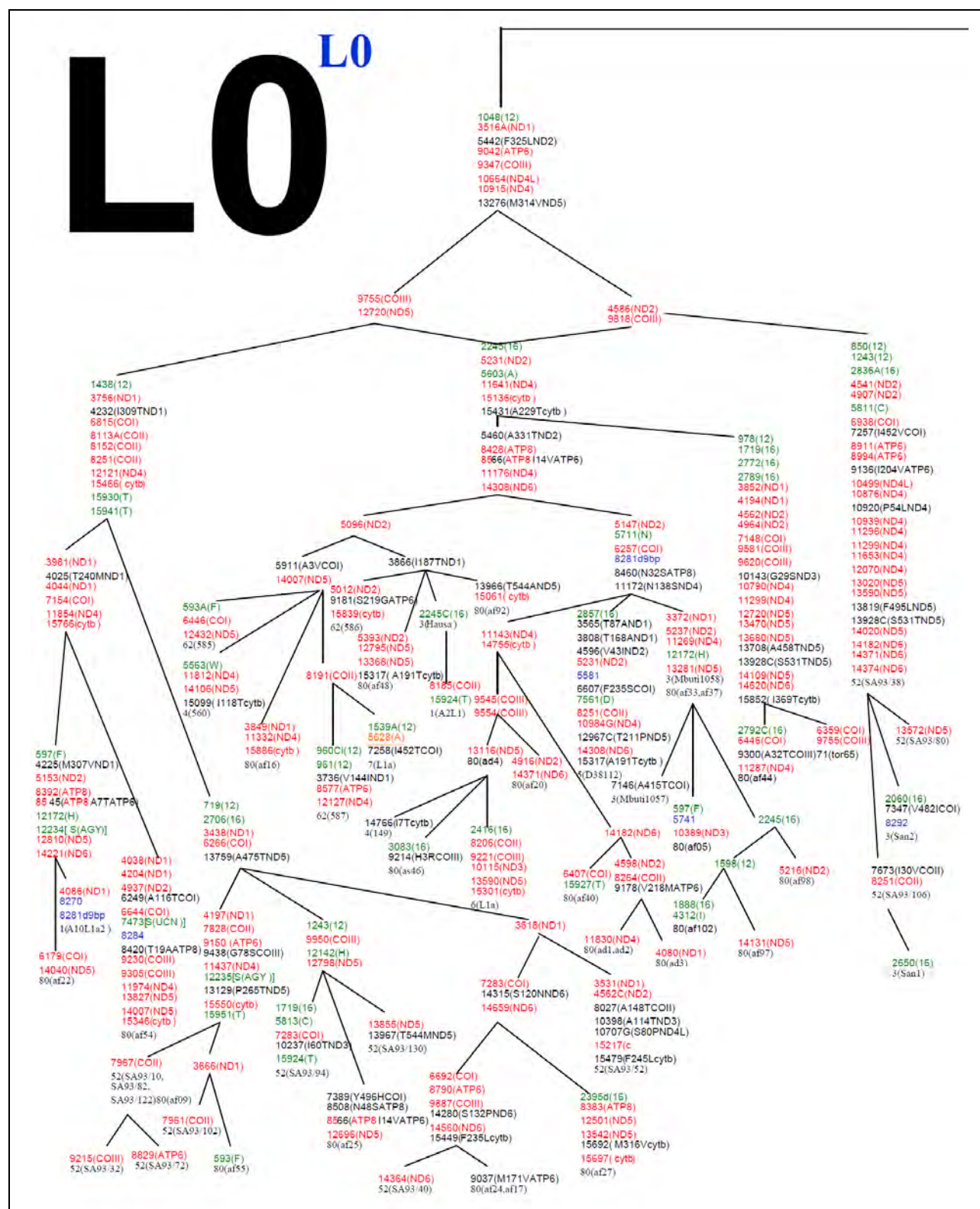
APPENDIX K

L0 haplogroup network obtained from MITOMAP

This appendix contains the L0 haplogroup section of a sequential mutational phylogenetic tree (Large mtDNA Tree) constructed by Ruiz-Pesini *et al.* (2007) and made available on MITOMAP (<http://www.mitomap.org/MitomapFigures.html>, 2007). The L0 haplogroup network is presented in Figure K.1.

The nucleotide substitutions, insertions and deletions presented in Figure K.1 are listed in numerical order and do not indicate the order in which the specific alteration occurred over time. Synonymous substitutions are indicated in **red** and are transitions, except where indicated otherwise. The gene in which the substitution occurs is presented in brackets. The non-synonymous substitutions are indicated in black, followed in brackets by the single letter abbreviations and position of the amino acids involved in the non-synonymous change, and the gene in which the substitution is located. Substitutions in rRNA and tRNA molecules are indicated in **green**, with the specific rRNA (12S or 16S) or tRNA (single letter abbreviation for the amino acid accepted by the tRNA) indicated in brackets.

Figure K.1 L0 haplogroup network obtained from MITOMAP



Adapted from Ruiz-Pesini *et al.* (2007). Colour coding of polymorphisms: black = non-synonymous substitution (amino acid, position, amino acid, gene), red = synonymous substitution (gene), green = tRNA or rRNA substitution (gene), orange = pathogenic substitution, blue = substitution affecting non-coding nucleotides observed in the coding region. All alterations are transitions except where indicated otherwise. The names of sequences are indicated at the bottom of the branch using a numeric code that refers to the bibliographic reference in which the sequence was described and the number/name of the sequence in the relevant article. Refer to the Bibliography for MITOMAP's Phylogenetic tree (<http://www.mitomap.org/treebib.html>, 2007) for more information. Nucleotide abbreviations: A = adenine; C = cytosine; G = guanine; U = uracil; Y = C or U; N = A, C, G or U. Amino acid abbreviations: A = alanine; C = cysteine; D = aspartic acid; E = glutamic acid; F = phenylalanine; G = glycine; H = histidine; I = isoleucine; L = leucine; M = methionine; N = asparagine; P = proline; S = serine; T = threonine; V = valine; W = tryptophan. Gene abbreviations: 12 = 12S rRNA; 16 = 16S rRNA; ATP = ATP synthase F_0 subunit; CO = cytochrome c oxidase subunit; cytb = cytochrome b; ND = NADH dehydrogenase subunit.

APPENDIX L

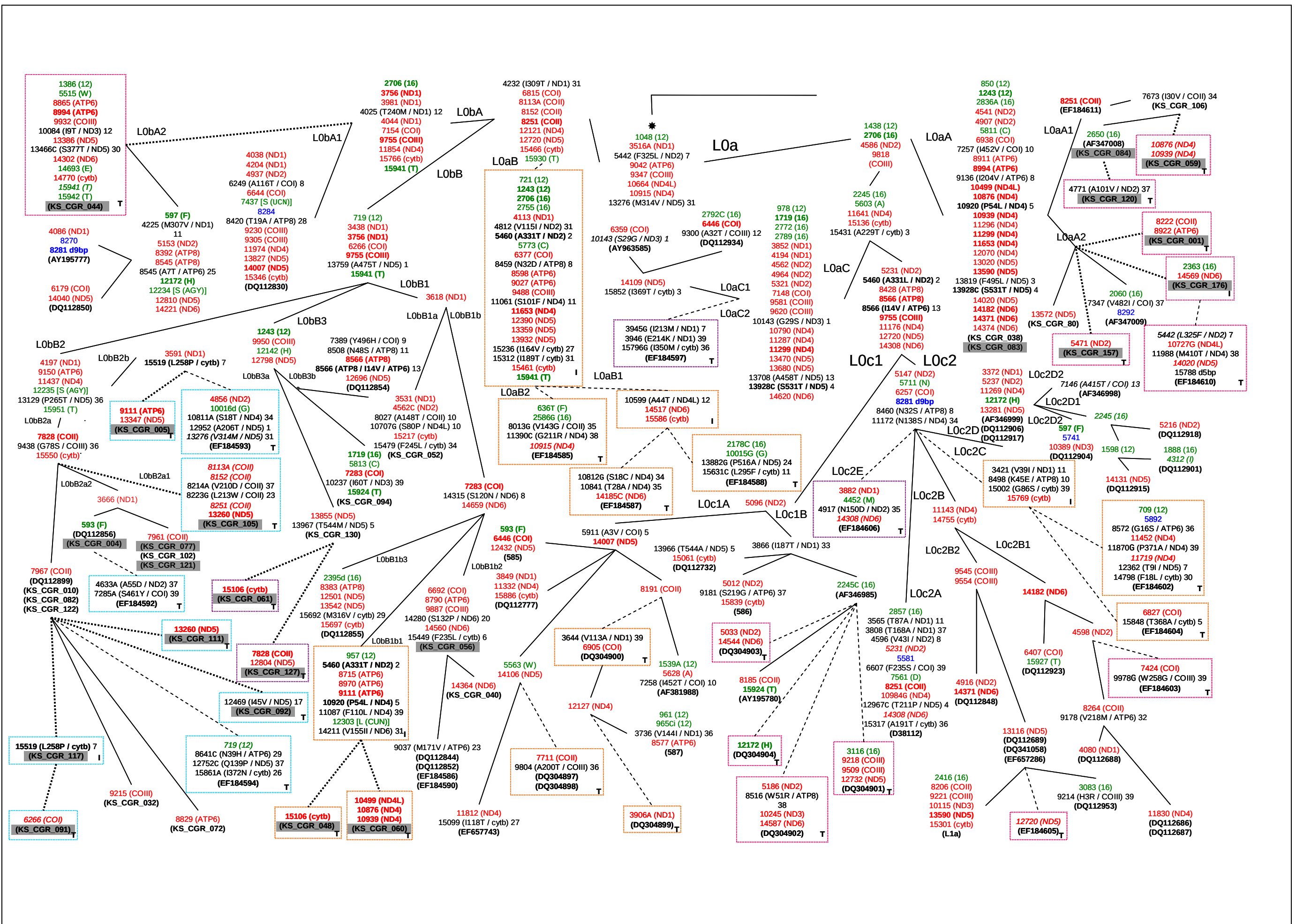
Global L0-specific haplogroup network

This appendix contains an A3 size copy of the global L0-specific haplogroup network constructed in the current investigation (refer to Section 5.10). The global L0-specific haplogroup network was constructed as described in Sections 4.13 and 5.10 by expanding the L0 haplogroup network obtained from MITOMAP in 2007 (refer to Appendix K) through the addition of a total of 46 L0 sequences so that the global L0-specific haplogroup network presented in this section consists of a total of 100 sequences.

The global L0-specific haplogroup network presented in this section follows the same conventions as the networks depicted in Section 5.10. Synonymous substitutions are indicated in **red** and are transitions, except where indicated otherwise. The gene in which the substitution occurs is indicated in brackets. The non-synonymous substitutions are indicated in black, followed in brackets by the single letter abbreviations and the position of the amino acids involved in the non-synonymous change, and the gene in which the substitution is located. Conservation numbers are indicated after each non-synonymous substitution. Substitutions in rRNA and tRNA genes are indicated in **green**, with the specific rRNA (12S or 16S) or tRNA (single letter abbreviation for the amino acid accepted by the tRNA) indicated in brackets. Substitutions in non-coding nucleotides present in the coding region are indicated in **blue**. Recurrent alterations are indicated in **bold** text, and reversions are indicated in *italic* text. Dashed lines were used to indicate new branching within the network compared to the original L0 haplogroup network (Ruiz-Pesini *et al.*, 2007), where big dashes (----) represent branches added by published sequences and small dashes (.....) indicate branches added by sequences generated in the current investigation. Coloured boxes (**pink**, **orange**, **light blue** and **violet**) were used to indicate new internal and terminal branches and **grey** highlighting was used to identify sequences generated in the current investigation. The global L0-specific haplogroup network is presented in Figure L.1.

Figure L.1 Global L0-specific haplogroup network

The global L0-specific haplogroup network was constructed by adapting the L0 network published by Ruiz-Pesini *et al.* (2007), and the branches within each sub-haplogroup were named according to the estimated TMRCA for the branch (refer to Section 5.11.6). The starting point of the network is indicated by *. Colour coding of polymorphisms: black = non-synonymous substitution (amino acid, position, amino acid, gene); red = synonymous substitution (gene); green = tRNA or rRNA (gene); blue = non-coding nucleotide in coding region. All alterations are transitions except where indicated otherwise. Polymorphisms in **bold** text signify that the polymorphism occurs in multiple branches, and polymorphisms in *italic* text indicate that the polymorphism reverted to the nucleotide present in the rCRS. Conservation numbers are indicated after each non-synonymous substitution. All the relevant information regarding the mitochondrial sequences used in the construction of the global L0-specific haplogroup network is listed in Appendix G. The names of sequences are indicated at the bottom of the branch in brackets and in **bold** text. GenBank® accession numbers are indicated where available. Sequences generated in the current investigation are highlighted in grey. New branches within the network are indicated by dashed lines, where big dashes (---) represent branches added by published sequences and small dashes (.....) indicate branches added by sequences generated in the current investigation. New internal and terminal branches are indicated by coloured boxes (pink, orange, light blue and violet). Nucleotide abbreviations: A = adenine; C = cytosine; G = guanine; T = thymine; U = uracil; Y = C or U; N = A, C, G or U. Amino acid abbreviations: A = alanine; C = cysteine; D = aspartic acid; F = phenylalanine; E = glutamic acid; G = glycine; H = histidine; I = isoleucine; L = leucine; M = methionine; N = asparagine; P = proline; Q = glutamine; R = arginine; S = serine; T = threonine; V = valine; W = tryptophan; Y = tyrosine. Gene abbreviations: 12 = 12S rRNA; 16 = 16S rRNA; ATP = ATP synthase F₀ subunit; CO = cytochrome c oxidase subunit; cytb = cytochrome *b*; ND = NADH dehydrogenase subunit. Other abbreviations: I = internal branch; T = terminal branch; d = deletion; bp = base pairs; i = insertion.



APPENDIX M

Synonymous codon usage of genes in 43 Khoi-San mtDNA genome sequences

The extent of the non-random usage of synonymous codons in the different polypeptide genes located in the mitochondrial genome was determined by calculating RSCU values for the 43 Khoi-San mitochondrial genome sequences contained in dataset 4b using MEGA version 3.1 (Kumar *et al.*, 2004) as described in Section 4.14.2.1. The positions of coding and non-coding regions within the genome sequences were assigned according to the regions first described by Anderson *et al.* in 1981 and listed on MITOMAP (2007). The vertebrate mitochondrial DNA code table was selected for analyses and all the values determined were average values calculated for all the mtDNA sequences in the dataset.

The RSCU values for the individual genes are presented in Table M.1 and Table M.2. The RSCU value of 1.00 observed for some codons indicates that there is no detectable codon bias for this specific codon and that its frequency is that which is expected under the assumption of equal codon usage. The remainder of the codons have RSCU values less than or more than 1.00 indicating that these codons are used less often or more often than expected. Some of these RSCU values were very low or very high indicating extreme codon usage bias, i.e. the codons with very low values are used rarely and the codons with very high values are used very frequently. The cut-off values for very low (≤ 0.3) or very high (≥ 1.7) RSCU values were determined as described in Section 4.14.2.1 based on the range of RSCU values obtained for the mitochondrial genome (refer to Section 5.11.1.2).

The RSCU values for the ATPase6, ATPase8, COI, COII, COIII and Cytb polypeptide genes are presented in Table M.1. The majority of RSCU values observed for the codons encoding amino acids for these polypeptide genes were below or above 1.00.

Table M.1 **Synonymous codon usage for ATPase, CO and Cytb genes in 43 Khoi-San mtDNA genome sequences**

Amino acid	Codon	ATPase6		ATPase8		COI		COII		COIII		Cytb	
		Freq	RSCU	Freq	RSCU	Freq	RSCU	Freq	RSCU	Freq	RSCU	Freq	RSCU
A	GCA	0.019	0.84	0.000	0.00	0.025	1.27	0.018	1.14	0.011	0.80	0.023	1.38
	GCC	0.057	2.53	0.000	0.00	0.027	1.37	0.022	1.42	0.042	2.93	0.045	2.62
	GCG	0.005	0.21	0.000	0.00	0.002	0.10	0.004	0.29	0.000	0.00	0.000	0.00
	GCU	0.010	0.42	0.000	0.00	0.025	1.26	0.018	1.14	0.004	0.27	0.000	0.00
C	UGC	0.000	0.00	0.000	0.00	0.002	2.00	0.009	1.33	0.004	2.00	0.005	2.00
	UGU	0.000	0.00	0.014	2.00	0.000	0.00	0.004	0.67	0.000	0.00	0.000	0.00
D	GAC	0.005	2.00	0.000	0.00	0.025	1.73	0.035	1.46	0.004	0.67	0.026	1.82
	GAU	0.000	0.00	0.000	0.00	0.004	0.27	0.013	0.54	0.008	1.33	0.003	0.18
E	GAA	0.010	2.00	0.014	2.00	0.013	1.38	0.035	1.45	0.021	1.59	0.011	2.00
	GAG	0.000	0.00	0.000	0.00	0.006	0.62	0.013	0.55	0.005	0.41	0.000	0.00
F	UUC	0.014	0.86	0.014	2.00	0.058	1.46	0.031	1.40	0.038	0.87	0.044	1.41
	UUU	0.019	1.14	0.000	0.00	0.022	0.54	0.013	0.60	0.050	1.13	0.018	0.59
G	GGA	0.014	1.71	0.000	0.00	0.036	1.60	0.016	1.34	0.015	0.86	0.024	1.50
	GGC	0.019	2.27	0.000	0.00	0.033	1.48	0.018	1.45	0.037	2.07	0.032	2.00
	GGG	0.000	0.00	0.000	0.00	0.007	0.31	0.006	0.47	0.015	0.86	0.008	0.50
	GGU	0.000	0.01	0.000	0.00	0.014	0.61	0.009	0.73	0.004	0.21	0.000	0.00
H	CAC	0.025	1.73	0.014	1.00	0.027	1.56	0.018	1.33	0.045	1.38	0.026	1.67
	CAU	0.004	0.27	0.014	1.00	0.008	0.44	0.009	0.67	0.020	0.62	0.005	0.33
I	AUC	0.067	1.05	0.029	1.33	0.050	1.36	0.065	1.36	0.027	1.00	0.073	1.42
	AUU	0.060	0.95	0.014	0.67	0.024	0.64	0.031	0.64	0.027	1.00	0.030	0.58
K	AAA	0.024	1.67	0.087	1.71	0.018	1.80	0.018	2.00	0.011	2.00	0.021	1.78
	AAG	0.005	0.33	0.014	0.29	0.002	0.20	0.000	0.00	0.000	0.00	0.003	0.22
L	CUA	0.098	2.94	0.046	1.91	0.064	3.20	0.050	2.10	0.062	2.85	0.066	2.34
	CUC	0.033	1.00	0.029	1.20	0.025	1.26	0.026	1.09	0.034	1.59	0.056	1.98
	CUG	0.022	0.67	0.000	0.00	0.008	0.38	0.014	0.61	0.015	0.68	0.005	0.18
	CUU	0.024	0.71	0.029	1.20	0.010	0.48	0.031	1.27	0.011	0.53	0.018	0.66
	UUA	0.016	0.49	0.026	1.09	0.014	0.68	0.018	0.74	0.008	0.35	0.019	0.67
	UUG	0.006	0.19	0.014	0.60	0.000	0.00	0.004	0.18	0.000	0.00	0.005	0.17
M	AUA	0.038	1.45	0.074	1.67	0.049	1.56	0.036	1.61	0.031	1.45	0.033	1.69
	AUG	0.014	0.55	0.014	0.33	0.014	0.44	0.008	0.39	0.011	0.55	0.006	0.31
N	AAC	0.038	1.77	0.043	1.21	0.024	1.43	0.022	1.43	0.019	1.67	0.032	1.60
	AAU	0.005	0.23	0.029	0.79	0.010	0.57	0.009	0.57	0.004	0.33	0.008	0.40
P	CCA	0.014	0.92	0.043	1.09	0.012	0.83	0.018	1.10	0.015	1.23	0.024	1.56
	CCC	0.033	2.15	0.101	2.55	0.035	2.48	0.037	2.25	0.023	1.85	0.024	1.56
	CCG	0.000	0.00	0.000	0.00	0.000	0.00	0.006	0.38	0.004	0.31	0.003	0.18
	CCU	0.014	0.92	0.014	0.36	0.010	0.69	0.004	0.27	0.008	0.62	0.011	0.69
Q	CAA	0.033	2.00	0.043	2.00	0.010	1.67	0.026	1.71	0.034	2.00	0.021	2.00
	CAG	0.000	0.00	0.000	0.00	0.002	0.33	0.004	0.29	0.000	0.00	0.000	0.00
R	CGA	0.010	2.00	0.000	0.00	0.008	2.00	0.013	2.00	0.011	2.40	0.011	2.29
	CGC	0.010	2.00	0.000	0.00	0.004	1.00	0.004	0.67	0.008	1.60	0.008	1.71
	CGG	0.000	0.00	0.000	0.00	0.000	0.00	0.000	0.00	0.000	0.00	0.000	0.00
	CGU	0.000	0.00	0.000	0.00	0.004	1.00	0.009	1.33	0.000	0.00	0.000	0.00

Table M.1 Continued...

Amino acid	Codon	ATPase6		ATPase8		COI		COII		COIII		Cytb	
		Freq	RSCU	Freq	RSCU	Freq	RSCU	Freq	RSCU	Freq	RSCU	Freq	RSCU
S	AGC	0.014	1.48	0.000	0.00	0.008	0.75	0.000	0.00	0.013	0.94	0.008	0.62
	AGU	0.000	0.02	0.000	0.00	0.000	0.00	0.004	0.60	0.004	0.28	0.003	0.21
	UCA	0.014	1.50	0.029	3.00	0.014	1.31	0.018	2.40	0.030	2.23	0.037	2.90
	UCC	0.010	1.00	0.014	1.50	0.019	1.88	0.018	2.40	0.027	1.97	0.029	2.28
	UCG	0.000	0.00	0.014	1.50	0.004	0.38	0.000	0.00	0.000	0.02	0.000	0.00
	UCU	0.019	2.00	0.000	0.00	0.018	1.69	0.004	0.60	0.008	0.56	0.000	0.00
T	ACA	0.038	1.39	0.029	1.00	0.031	1.93	0.036	1.54	0.042	1.83	0.032	1.66
	ACC	0.053	1.93	0.058	2.00	0.021	1.33	0.035	1.53	0.038	1.67	0.034	1.79
	ACG	0.000	0.00	0.014	0.50	0.004	0.24	0.000	0.00	0.004	0.17	0.003	0.14
	ACU	0.019	0.69	0.014	0.50	0.008	0.50	0.021	0.94	0.008	0.33	0.008	0.41
V	GUA	0.019	1.95	0.013	4.00	0.033	1.88	0.022	1.53	0.023	1.85	0.016	2.40
	GUC	0.006	0.60	0.000	0.00	0.024	1.35	0.031	2.14	0.016	1.26	0.008	1.20
	GUG	0.005	0.47	0.000	0.00	0.006	0.33	0.000	0.00	0.004	0.31	0.000	0.00
	GUU	0.010	0.97	0.000	0.00	0.008	0.44	0.005	0.33	0.007	0.59	0.003	0.40
W	UGA	0.014	2.00	0.029	1.33	0.031	2.00	0.013	1.50	0.034	1.50	0.029	1.83
	UGG	0.000	0.00	0.014	0.67	0.000	0.00	0.004	0.50	0.011	0.50	0.003	0.17
Y	UAC	0.010	1.33	0.014	1.00	0.036	1.68	0.035	1.78	0.027	1.27	0.029	1.29
	UAU	0.005	0.67	0.014	1.00	0.007	0.32	0.004	0.22	0.015	0.73	0.016	0.71
*	AGA	0.000	0.00	0.000	0.00	0.002	4.00	0.000	0.00	0.000	0.00	0.000	0.00
	AGG	0.000	0.00	0.000	0.00	0.000	0.00	0.000	0.00	0.000	0.00	0.000	0.00
	UAA	0.000	0.00	0.000	0.00	0.000	0.00	0.000	0.00	0.000	0.00	0.000	0.00
	UAG	0.000	0.00	0.014	4.00	0.000	0.00	0.004	4.00	0.000	0.00	0.000	0.00

Frequency = frequency of codon usage, i.e. the number of times the specific codon was used; RSCU = relative synonymous codon usage, i.e. the observed frequency of the specific codon divided by that expected under the assumption of equal usage of synonymous codons. RSCU values equal to **1.00** are indicated in **red** and very high or very low RSCU values are indicated in **green**. ATPase = ATP synthase F₀ subunit; CO = cytochrome c oxidase subunit; Cytb = cytochrome b. Nucleotide abbreviations: A = adenine; C = cytosine; G = guanine; U = uracil. Amino acid abbreviations: A = alanine; C = cysteine; D = aspartic acid; E = glutamic acid; F = phenylalanine; G = glycine; H = histidine; I = isoleucine; K = lysine; L = leucine; M = methionine; N = asparagine; P = proline; Q = glutamine; R = arginine; S = serine; T = threonine; V = valine; W = tryptophan; Y = tyrosine; * = termination.

The RSCU values for the ND1, ND2, ND3, ND4, ND4L, ND5 and ND6 genes are presented in Table M.2. Most of the RSCU values determined for the codons encoding amino acids for these polypeptide genes were above or below 1.00.

Table M.2 Synonymous codon usage for ND genes in 43 Khoi-San mtDNA genome sequences

Amino acid	Codon	ND1		ND2		ND3		ND4		ND4L		ND5		ND6	
		Freq	RSCU	Freq	RSCU	Freq	RSCU	Freq	RSCU	Freq	RSCU	Freq	RSCU	Freq	RSCU
A	GCA	0.031	1.48	0.023	1.61	0.000	0.00	0.020	0.90	0.018	1.23	0.032	1.79	0.000	0.00
	GCC	0.044	2.07	0.021	1.45	0.069	3.55	0.051	2.22	0.027	1.90	0.031	1.75	0.000	0.00
	GCG	0.000	0.01	0.003	0.20	0.000	0.00	0.010	0.44	0.000	0.00	0.000	0.00	0.000	0.00
	GCU	0.009	0.44	0.011	0.74	0.009	0.45	0.010	0.44	0.012	0.87	0.008	0.46	0.011	4.00
C	UGC	0.000	0.00	0.000	0.00	0.009	2.00	0.030	2.00	0.006	1.88	0.008	1.67	0.006	0.67
	UGU	0.000	0.00	0.000	0.00	0.000	0.00	0.000	0.00	0.000	0.12	0.002	0.33	0.011	1.33

Table M.2 Continued...

Amino acid	Codon	ND1		ND2		ND3		ND4		ND4L		ND5		ND6	
		Freq	RSCU	Freq	RSCU	Freq	RSCU	Freq	RSCU	Freq	RSCU	Freq	RSCU	Freq	RSCU
D	GAC	0.013	2.00	0.000	0.00	0.017	1.33	0.010	2.00	0.007	2.00	0.013	1.45	0.017	2.00
	GAU	0.000	0.00	0.000	0.00	0.009	0.67	0.000	0.00	0.000	0.00	0.005	0.55	0.000	0.00
E	GAA	0.025	1.45	0.014	1.67	0.026	1.20	0.020	2.00	0.020	2.00	0.013	1.78	0.000	0.00
	GAG	0.009	0.55	0.003	0.33	0.017	0.80	0.000	0.00	0.000	0.00	0.002	0.22	0.000	0.00
F	UUC	0.038	1.51	0.018	0.87	0.052	1.50	0.010	0.67	0.024	1.10	0.054	1.74	0.006	1.02
	UUU	0.012	0.49	0.023	1.13	0.017	0.50	0.020	1.33	0.020	0.90	0.008	0.26	0.006	0.98
G	GGA	0.005	0.55	0.014	1.54	0.009	1.33	0.020	2.00	0.011	1.18	0.017	1.54	0.000	0.00
	GGC	0.025	2.67	0.020	2.15	0.017	2.67	0.020	2.00	0.020	2.12	0.019	1.73	0.000	0.00
	GGG	0.007	0.78	0.000	0.00	0.000	0.00	0.000	0.00	0.004	0.47	0.002	0.15	0.000	0.00
	GGU	0.000	0.00	0.003	0.31	0.000	0.00	0.000	0.00	0.002	0.24	0.006	0.58	0.000	0.00
H	CAC	0.006	2.00	0.008	1.47	0.000	0.00	0.020	1.33	0.026	1.85	0.022	1.86	0.060	1.14
	CAU	0.000	0.00	0.003	0.53	0.000	0.00	0.010	0.67	0.002	0.15	0.002	0.14	0.046	0.86
I	AUC	0.041	1.16	0.069	1.50	0.035	0.90	0.020	0.57	0.052	1.21	0.056	1.28	0.017	0.98
	AUU	0.030	0.84	0.023	0.50	0.043	1.10	0.051	1.43	0.034	0.79	0.031	0.72	0.018	1.02
K	AAA	0.019	1.73	0.029	1.67	0.026	2.00	0.000	0.00	0.020	1.79	0.033	1.90	0.046	2.00
	AAG	0.003	0.27	0.006	0.33	0.000	0.00	0.000	0.00	0.002	0.21	0.002	0.10	0.000	0.00
L	CUA	0.083	2.50	0.075	2.41	0.096	2.36	0.147	3.80	0.090	2.58	0.075	2.61	0.000	0.00
	CUC	0.057	1.73	0.054	1.74	0.035	0.86	0.061	1.57	0.070	2.00	0.052	1.79	0.051	5.40
	CUG	0.020	0.62	0.012	0.37	0.017	0.43	0.004	0.12	0.009	0.27	0.013	0.45	0.000	0.00
	CUU	0.019	0.57	0.020	0.64	0.009	0.21	0.010	0.26	0.020	0.59	0.018	0.64	0.006	0.60
	UUA	0.016	0.48	0.023	0.74	0.087	2.14	0.010	0.26	0.018	0.50	0.011	0.39	0.000	0.00
	UUG	0.003	0.10	0.003	0.09	0.000	0.00	0.000	0.00	0.002	0.06	0.003	0.12	0.000	0.00
M	AUA	0.041	1.62	0.063	1.84	0.061	1.75	0.091	1.80	0.053	1.86	0.036	1.71	0.063	2.00
	AUG	0.009	0.38	0.005	0.16	0.009	0.25	0.010	0.20	0.004	0.14	0.006	0.29	0.000	0.00
N	AAC	0.041	2.00	0.040	1.40	0.035	2.00	0.040	1.33	0.048	1.91	0.045	1.64	0.051	0.86
	AAU	0.000	0.00	0.017	0.60	0.000	0.00	0.020	0.67	0.002	0.09	0.010	0.36	0.069	1.14
P	CCA	0.006	0.37	0.012	0.70	0.026	1.50	0.000	0.00	0.013	1.01	0.015	1.14	0.011	0.40
	CCC	0.053	3.09	0.035	2.09	0.017	1.00	0.010	2.00	0.032	2.60	0.025	1.87	0.087	3.01
	CCG	0.003	0.18	0.006	0.35	0.000	0.00	0.000	0.00	0.000	0.00	0.000	0.00	0.011	0.39
	CCU	0.006	0.36	0.014	0.87	0.026	1.50	0.010	2.00	0.005	0.39	0.013	0.99	0.006	0.20
Q	CAA	0.019	1.98	0.023	1.60	0.026	2.00	0.010	2.00	0.020	1.80	0.028	1.70	0.045	1.45
	CAG	0.000	0.02	0.006	0.40	0.000	0.00	0.000	0.00	0.002	0.20	0.005	0.30	0.017	0.55
R	CGA	0.016	2.86	0.000	0.00	0.000	0.00	0.000	0.00	0.009	1.78	0.005	1.50	0.000	0.00
	CGC	0.006	1.14	0.009	3.00	0.009	4.00	0.010	4.00	0.011	2.22	0.007	2.00	0.017	4.00
	CGG	0.000	0.00	0.000	0.00	0.000	0.00	0.000	0.00	0.000	0.00	0.000	0.00	0.000	0.00
	CGU	0.000	0.00	0.003	1.00	0.000	0.00	0.000	0.00	0.000	0.00	0.002	0.50	0.000	0.00
S	AGC	0.006	0.57	0.012	0.86	0.000	0.00	0.000	0.00	0.018	1.12	0.016	1.20	0.023	1.63
	AGU	0.003	0.29	0.003	0.21	0.009	0.99	0.000	0.00	0.004	0.28	0.005	0.36	0.011	0.81
	UCA	0.025	2.29	0.021	1.55	0.009	0.99	0.030	2.24	0.022	1.39	0.022	1.59	0.017	1.21
	UCC	0.025	2.29	0.032	2.41	0.036	4.02	0.040	3.01	0.037	2.37	0.034	2.48	0.033	2.35
	UCG	0.000	0.00	0.002	0.16	0.000	0.00	0.010	0.75	0.002	0.14	0.002	0.12	0.000	0.00
	UCU	0.006	0.57	0.011	0.81	0.000	0.00	0.000	0.00	0.011	0.70	0.003	0.25	0.000	0.00
T	ACA	0.044	1.57	0.032	1.03	0.017	1.32	0.010	0.80	0.050	1.90	0.037	1.37	0.006	0.31
	ACC	0.062	2.21	0.069	2.23	0.026	2.01	0.030	2.40	0.038	1.44	0.049	1.78	0.046	2.47
	ACG	0.000	0.00	0.008	0.27	0.000	0.00	0.000	0.00	0.000	0.02	0.001	0.06	0.006	0.31
	ACU	0.006	0.22	0.014	0.46	0.009	0.67	0.010	0.80	0.017	0.65	0.022	0.79	0.017	0.91

Table M.2 Continued...

Amino acid	Codon	ND1		ND2		ND3		ND4		ND4L		ND5		ND6	
		Freq	RSCU	Freq	RSCU	Freq	RSCU	Freq	RSCU	Freq	RSCU	Freq	RSCU	Freq	RSCU
V	GUA	0.009	1.20	0.009	1.47	0.009	1.33	0.028	1.88	0.017	2.39	0.010	1.39	0.000	0.00
	GUC	0.013	1.65	0.006	1.00	0.009	1.33	0.012	0.79	0.009	1.23	0.008	1.18	0.000	0.00
	GUG	0.000	0.00	0.000	0.03	0.000	0.00	0.012	0.79	0.003	0.38	0.000	0.00	0.000	0.00
	GUU	0.009	1.15	0.009	1.49	0.009	1.33	0.008	0.54	0.000	0.00	0.010	1.44	0.006	4.00
W	UGA	0.028	2.00	0.029	1.82	0.035	2.00	0.000	0.00	0.028	1.96	0.018	1.84	0.011	1.95
	UGG	0.000	0.00	0.003	0.18	0.000	0.00	0.000	0.00	0.000	0.04	0.002	0.16	0.000	0.05
Y	UAC	0.028	1.29	0.023	1.60	0.017	1.33	0.020	1.00	0.024	1.69	0.015	1.15	0.049	1.62
	UAU	0.016	0.71	0.006	0.40	0.009	0.67	0.020	1.00	0.004	0.31	0.011	0.85	0.011	0.38
*	AGA	0.000	0.00	0.000	0.00	0.000	0.00	0.000	0.00	0.000	0.00	0.000	0.00	0.017	0.79
	AGG	0.000	0.00	0.000	0.00	0.000	0.00	0.000	0.00	0.000	0.00	0.000	0.00	0.023	1.06
	UAA	0.000	0.00	0.000	0.00	0.000	0.00	0.010	4.00	0.000	0.00	0.002	4.00	0.046	2.15
	UAG	0.000	0.00	0.000	0.00	0.000	0.00	0.000	0.00	0.000	0.00	0.000	0.00	0.000	0.00

Frequency = frequency of codon usage, i.e. the number of times the specific codon was used; RSCU = relative synonymous codon usage, i.e. the observed frequency of the specific codon divided by that expected under the assumption of equal usage of synonymous codons. RSCU values equal to **1.00** are indicated in **red** and very high or very low RSCU values are indicated in **green**. ND = NADH dehydrogenase subunit. Nucleotide abbreviations: A = adenine; C = cytosine; G = guanine; U = uracil. Amino acid abbreviations: A = alanine; C = cysteine; D = aspartic acid; E = glutamic acid; F = phenylalanine; G = glycine; H = histidine; I = isoleucine; K = lysine; L = leucine; M = methionine; N = asparagine; P = proline; Q = glutamine; R = arginine; S = serine; T = threonine; V = valine; W = tryptophan; Y = tyrosine; * = termination.

The majority of RSCU values observed for the polypeptide genes were below or above 1.00, and it can therefore be suggested that the different genes were subjected to codon usage bias. A general pattern was observed for the individual polypeptide genes in that the majority of amino acids that were encoded by two, four or six codons, in turn presented with one, two or three of the available codons being preferred. The preferred codon could however be different for the same encoded amino acid in different genes.

Codon usage bias in the polypeptide genes can be attributed to selection as it has been demonstrated that these genes are influenced by positive selection (Mishmar *et al.*, 2003; Ruiz-Pesini *et al.*, 2004; Mishmar *et al.*, 2006; Ruiz-Pesini and Wallace, 2006). It has been proposed that an estimated 25% of all mtDNA variation is due to adaptation which has allowed humans to adjust to varying caloric availability and differences in the demand for mitochondrial energy production via ATP production compared to heat generation (Mishmar *et al.*, 2003; Ruiz-Pesini *et al.*, 2004). The effect of selection on human mitochondrial variation was discussed in Section 5.11.5.

APPENDIX N

Mitochondrial consensus sequence for the Khoi-San population of southern Africa

This appendix contains the complete Khoi-San consensus sequence for the Khoi-San population of southern Africa. The Khoi-San consensus sequence was constructed using 35 Southern African Khoi-San L0 sequences (Mouton, 2003; current investigation), and involved the construction of separate L0a and L0b consensus sequences which were subsequently combined to generate a single consensus sequence (refer to Sections 4.16 and 5.14.8). At nucleotide positions for which a consensus character state could not be determined due to differences observed between the L0a and L0b consensus sequences, the nucleotide was resolved by defining it using an ambiguity code (R, Y, S or M). In the case of insertions and deletions, the consensus for the nucleotide position was determined by using a consensus sequence constructed using all 35 L0 sequences.

The Khoi-San consensus sequence consisted of a total of 16,567 bp. The N at nucleotide position 3,107 in the rCRS was retained in the Khoi-San consensus sequence in order to adhere to the numbering system as used in the rCRS (Andrews *et al.*, 1999). The Khoi-San consensus sequence exhibited a total of 127 alterations compared to the rCRS and the types of alterations and their distribution throughout the mitochondrial genome was discussed in Section 5.14.8. The complete Khoi-San consensus sequence is presented in Figure N.1.

Figure N.1 Mitochondrial consensus sequence for the Khoi-San population of southern Africa

1	gatcacaggt	ctatcaccct	attaaccact	cacgggagct	ctccatgcat	ttggtathtt
61	cgtctggggg	gtgtgcacgc	gatagcattg	cgagacgctg	gagccggagc	accctatgtc
121	gcagtatctg	tctttgattc	ctgcccatc	cattatttta	tgcacctac	gttcaatatt
181	acaggcgar	atacctatta	aagtgrtta	attaattaat	gcttgttaga	cataataata
241	acaattaat	gtctgcacag	ccactttcca	cacagacatc	ataacaaaa	atttccacca
301	aacccccct	cccccgctt	ctggccacag	cacttaaaca	catctctgcc	aaaccccaaa
361	aacaaagaac	cctaacacca	gcctaaccag	atttcaaatt	ttatcttttg	gcggtatgca
421	cttttaacag	tcacccccca	actaacacat	tatttttccc	tcccactccc	atactactaa
481	tctcatcaat	acaacccgc	ccatcctacc	cagcacacac	accgctgcta	accccatacc
541	ccgaaccaac	caaaccccaa	agacaccccc	cacagtttat	gtagcttacc	tcctcaaagc
601	aatacactga	aatgttttag	acgggctcac	atcaccccat	aaacaaatag	gtttggctct

Figure N.1 Continued...

661	agcctttcta	ttagctctta	gtaagattac	acatgcaagc	atccccgttc	cagtgar ttc
721	accctctaaa	tcaccacgat	caaaagg g ac	aagcatcaag	cacgca a caa	tgcagctcaa
781	aacgcttagc	ctagccacac	ccccacggga	aacagcagtg	at a aaccttt	agcaataaac
841	gaaagtt y aa	ctaagctata	ctaacc c cag	ggttgggtcaa	tttcgtgcc	gccaccgcgg
901	tcacacgatt	aacccaagtc	aatagaagcc	ggcgtaaaga	gtgttttaga	tcacccctc
961	cccaataaag	ctaaaactca	cctgagttgt	aaaaaactcc	agttgacaca	aaata a acta
1021	cgaaagtggc	tttaacatat	ctgaat t acac	aatagctaag	acccaaactg	ggattagata
1081	ccccactatg	cttagcccta	aacctcaaca	gttaaataca	caaaactgct	cgccagaaca
1141	ctacgagcca	cagcttaaaa	ctcaaaggac	ctggcggtgc	ttcatatccc	tctagaggag
1201	cctgttctgt	aatcgataaa	ccccgatcaa	cctcaccacc	y cttgctcag	cctatatacc
1261	gccatcttca	gcaaaccctg	atgaaggcta	caaagtaagc	gcaagtaccc	acgtaaagac
1321	gttaggtcaa	gggtgtagccc	atgaggtggc	aagaaatggg	ctacattttc	tacccagaa
1381	aactacgata	gcccttatga	aacttaaggg	tcgaaggtgg	atttagcagt	aaact r agag
1441	tagagtgttt	agttgaacag	ggccctgaag	cgcgtaacaca	ccgcccgtca	ccctcctcaa
1501	gtatacttca	aaggacattt	aactaaaacc	cctacgcatt	tatatagagg	agacaagtcg
1561	taacatggta	agtgtactgg	aaagtgcact	tggacgaacc	agagtgtagc	ttaacacaaa
1621	gcaccaact	tacacttagg	agatttcaac	ttaacttgac	cgctctgagc	taaacctagc
1681	cccaaacc	ctccacctta	ctaccagaca	accttagcca	aaccatttac	ccaaataaag
1741	tataggcgat	agaaattgaa	acctggcgca	atagatatag	taccgcaagg	gaaagatgaa
1801	aaattataac	caagcataat	atagcaagga	ctaacc c cta	taccttctgc	ataatgaatt
1861	aactagaaat	aactttgcaa	ggagagccaa	agctaagacc	cccgaacca	gacgagctac
1921	ctaagaacag	ctaaaagagc	acacc c gtct	atgtagcaaa	atagtgggaa	gatttatagg
1981	tagaggcgac	aaacctaccg	agcctgggtga	tagctgggtg	tccaagatag	aatcttagtt
2041	caacttttaa	tttgcccaca	gaaccctcta	aatcc c cttg	taaatttaac	tgttagtcca
2101	aagaggaaca	gctctttgga	cactaggaaa	aaaccttgta	gagagagtaa	aaaatttaac
2161	acccatagta	ggcctaaaag	cagccacca	ttaagaaagc	gttcaagctc	aacacc c act
2221	acctaataaa	tcccaaacat	ataactgaac	tcctcacacc	caattggacc	aatctatcac
2281	cctatagaag	aactaatggt	agtataagta	acatgaaaac	attctcctcc	gcataagcct
2341	gcgtcagatt	aaaacactga	actgacaatt	aacagcccaa	tatctacaat	caaccaacaa
2401	gtcattatta	ccctcactgt	caacc c aaca	caggcatgct	cataaggaaa	ggttaaaaaa
2461	agtaaaagga	actcggcaaa	tcttacc c cg	cctgtttacc	aaaacatca	cctctagcat
2521	caccagtatt	agaggcaccg	cctgcccagt	gacacatggt	taacggccgc	ggtacc c taa
2581	ccgtgcaaag	gtagcataat	cacttg t tcc	ttaaataggg	acctgtatga	atggctccac
2641	gagggttcag	ctgtctctta	cttttaacca	gtgaaattga	cctgcccgtg	aagaggcggg
2701	cat r acacag	caagacgaga	agaccctatg	gagctttaat	ttattaatgc	aaaca a tacc
2761	taacaaaccc	acaggtccta	aactacc a aa	cctgcattaa	aaatttcggt	tggggcgacc
2821	tcggagcaga	acc m aacctc	cgagcagtac	atgctaagac	ttcaccagtc	aaagcgaact
2881	ac c atactca	attgatccaa	taacttgacc	aacggaacaa	gttaccctag	ggataacagc
2941	gcaatcctat	tctagagtcc	atatcaacaa	tagggttttac	gacctcgatg	ttggatcagg
3001	acatcccgat	gggtgcagccg	ctattaaagg	ttcgtttggt	caacgattaa	agtcctacgt
3061	gatctgagtt	cagaccggag	taatccaggt	cggtttctat	ctac m ttcaa	attcctccct
3121	gtacgaaagg	acaagagaaa	taaggcctac	ttcacaaagc	gccttcccc	gtaaatgata
3181	tcatctcaac	ttagtattat	accacaccc	acccaagaac	agggtttggt	aagatggcag
3241	agcccggtaa	tcgcataaaa	cttaaaactt	tacagtcaga	ggttcaattc	ctcttcttaa
3301	caacataccc	atggccaacc	tcctactcct	cattgtaccc	attctaatacg	caatggcatt
3361	cctaattgctt	accgaacgaa	aaattctagg	ctatatacaa	ctacgcaaag	gccccaacgt
3421	tgtaggcccc	tacgg r ctac	tacaaccctt	cgctgacgcc	ataaaactct	tcaccaaaga
3481	gcccctaaaa	cccgccacat	ctaccatcac	cct a atcac	accgccccga	ccttagctct
3541	caccatcgct	cttctactat	gaacccccct	ccccataccc	aaccc c ctgg	t a aacctcaa
3601	cctaggcctc	ctattttatc	tagccacctc	tagcctagcc	gtttactcaa	tcctctgac
3661	aggggtgagca	tcaaactcaa	actacgcct	gatcggcgca	ctgcgagcag	tagcccaaac
3721	aatctcatat	gaagtcaccc	tagccatcat	tct r ctatca	acattactaa	taagtggctc

Figure N.1 Continued...

3781	ctttaacctc	tccaccctta	tcacaacaca	agaacacctc	tgattactcc	tgccatcatg
3841	acccttggcc	ataatatgat	ttatctccac	actagcagag	accaaccgaa	cccccttcga
3901	ccttgccgaa	ggggagtcgg	aactagtctc	aggcttcaac	atcgaatacg	ccgcaggccc
3961	cttcgcccta	ttcttcatag	ccgaatacac	aaacattatt	ataataaaca	ccctcaccac
4021	tacaatcttc	ctaggaacaa	catatgacgc	actctccctt	gaactctaca	caacatattt
4081	tgtcaccaag	accctacttc	t ^g acctccct	gttcttatga	attcgaacag	cataccccg
4141	attccgctac	gaccaactca	tacacctcct	atgaaaaaac	ttcctaccac	tcac ^y ctagc
4201	attacttata	tgatatgtct	ccataccca ^y	tacaatctcc	agcattcccc	ctcaaacctt
4261	agaaatatgt	ctgataaaag	agttactttg	atagagtaaa	taataggag ^t	ttaaaccccc
4321	ttattttctag	gactatgaga	atcgaaccca	tccctgagaa	tccaaaattc	tccgtgccac
4381	ctatcacacc	ccatcctaaa	gtaaggctcag	ctaaataagc	tatcggggccc	ataccccgaa
4441	aatgttgggt	atacccttcc	cgtactaatt	aatcccctgg	cccaaccctg	catctactct
4501	accatctttg	caggcacact	catcacagcg	ctaagctc ^r	actgattttt	tacctgagta
4561	ggcctagaaa	taaacatgct	agc ^y ttttatt	ccagttctaa	ccaaaaaaat	aaaccctcgt
4621	tccacagaag	ctgccatcaa	gtatttccct	acgcaagcaa	ccgcatccat	aatccttcta
4681	atagctatcc	tcttcaacaa	tatactctcc	ggacaatgaa	ccataaccaa	tactaccaat
4741	caatactcat	cattaataat	cataat ^g gct	atagcaataa	aactaggaat	agcccccttt
4801	cacttctgag	tcccagaggt	tacccaaggc	accctcttga	catccggcct	gcttcttctc
4861	acatgacaaa	aactagcccc	catctcaatc	atataccaaa	tctc ^y ccctc	actaaacgta
4921	agccttctcc	tcactctctc	aatcttatcc	atcatagcag	gcagttgagg	tggattaaac
4981	caaaccagc	tacgcaaaat	cttagcatac	tcctcaatta	cccacatagg	atgaataata
5041	gcagttctac	cgtacaaccc	taacataacc	attcttaatt	taactattta	tattatccta
5101	actactaccg	cattcctact	actcaactta	aactccagca	ccacgaccct	actactatct
5161	cgcacctgaa	acaagctaac	atgactaaca	cccttaattc	catccaccct	cctctcccta
5221	ggaggcctgc	ccccgctaac	cggctttttg	cccaaattgg	ccattatcga	agaattcaca
5281	aaaaacaata	gcctcatcat	cccaccatc	atagccacca	tcaccctcct	taacctctac
5341	ttctacctac	gcctaatact	ctccacctca	atcacactac	tccccatata	taacaacgta
5401	aaaataaaaat	gacagtttga	acatacaaaa	cccaccca ^c	tcttccccac	actcatcgcc
5461	cttaccacgc	tactcctacc	tatctccctt	tttatactaa	taatcttata	gaaatttagg
5521	ttaaatacag	accaagagcc	ttcaaagccc	tcagtaagtt	gcaatactta	atttctgtaa
5581	cagctaagga	ctgcaaaacc	ccactctgca	tcaactgaac	gcaaatacgc	cactttaatt
5641	aagctaagcc	cttactagac	caatgggact	taaaccacac	aacacttagt	taacagctaa
5701	gcaccctaata	caactggctt	caatctactt	ctcccgccgc	cgggaaaaaa	ggcggggagaa
5761	gccccggcag	gtttgaagct	gcttcttcga	atttgcaatt	caatatga ^r	atcacctcgg
5821	agctggtaaa	aagaggccta	accctgtctt	ttagattttac	agtccaatgc	ttcactcagc
5881	cattttacct	cacccccact	gatgttcgcc	gaccgttgac	tattctctac	aaaccacaaa
5941	gacattggaa	cactatacct	attattcggc	gcatgagctg	gagtcctagg	cacagctcta
6001	agcctcctta	ttcgagccga	gctgggccag	ccaggcaacc	ttctaggtaa	cgaccacatc
6061	tacaacgtta	tcgtcacagc	ccatgcattt	gtaataatct	tcttcatagt	aatacccatc
6121	ataatcggag	gctttggcaa	ctgactagtt	cccctaataa	tcggtgcccc	cgatatggcg
6181	tt ^c ccccgca	taaacaacat	aagcttctga	ctcttacctc	cctctctcct	actcctgctc
6241	gcatctgcta	tagtggaggc	cgg ^r gcagga	acaggttgaa	cagtctaccc	tcccttagca
6301	gggaactact	cccaccctgg	agcctccgta	gacctaacca	tcttctcctt	acacctagca
6361	ggtgtctcct	ctatcttagg	ggccatcaat	ttcatcacaa	caattatcaa	tataaaaccc
6421	cctgccataa	ccaataacca	aacgcccctc	ttcgtctgat	ccgtcctaata	cacagcagtc
6481	ctacttctcc	tatctctccc	agtccctagc	gctggcatca	ctatactact	aacagaccgc
6541	aacctcaaca	ccaccttctt	cgaccccgcc	ggaggaggag	acccattctt	ataccaacac
6601	ctattctgat	ttttcgggtca	ccctgaagtt	tatattctta	tcctaccagg	cttcggaata
6661	atctcccata	ttgtaactta	ctactccgga	aaaaaagaac	catttggtata	cataggtatg
6721	gtctgagcta	tgatatcaat	tggcttccta	gggtttatcg	tgtgagcaca	ccatatattt
6781	acagtaggaa	tagacgtaga	cacacgagca	ta ^y ttcacct	ccgtaccat	aatcatcgct
6841	atccccaccg	gcgtcaaagt	atttagctga	ctcgccacac	tccacggaag	caatatgaaa

Figure N.1 Continued...

6901	tgatctgctg	cagtgtctctg	agccctagga	ttcatytttc	ttttcaccgt	aggtggcctg
6961	actggcattg	tattagcaaa	ctcatcacta	gacatcgtac	tacacgacac	gtactacgtt
7021	gtagctcact	tccactatgt	cctatcaata	ggagctgtat	ttgccatcat	aggaggcttc
7081	attcactgat	ttcccctatt	ctcaggctac	accctagacc	aaacctacgc	caaaatccat
7141	ttcgctatca	tattcatcgg	cgtaaactta	actttcttcc	cacaacactt	tctcggccta
7201	tccggaatgc	cccgacgtta	ctcggactac	cccgatgcac	acaccacatg	aaatrtccta
7261	tcatctgtag	gctcattcat	ttctctaaca	gcagtaatat	taataatttt	catgatttga
7321	gaagccttcg	cttcgaagcg	aaaagtccta	atagtagaag	aacctcccat	aaacctggag
7381	tgactatatg	gatgcccccc	accctaccac	acattcgaag	aaccggtata	cataaaatct
7441	agacaaaaaa	ggaaggaatc	gaacccccca	aagctgggtt	caagccaacc	ccatggcctc
7501	catgactttt	tcaaaaagat	attagaaaaa	ccatttcata	actttgtcaa	agttaaatta
7561	taggctaaat	cctatatatc	ttaatggcac	atgcagcgca	agtaggtcta	caagacgcta
7621	cttcccctat	catagaagag	cttatcacct	ttcatgatca	cgccctcata	atcattttcc
7681	ttatctgctt	cctagtctctg	tatgcccttt	tcctaacact	cacaacaaaa	ctaactaata
7741	ctaacatctc	agacgctcag	gaaatagaaa	ccgtctgaac	tatcctgccc	gccatcatcc
7801	tagtcctcat	cgccctccca	tccctrcgca	tcctttacat	aacagacgag	gtcaacgac
7861	cctcccttac	catcaaatca	attggccacc	aatggtactg	aacctacgag	tacaccgact
7921	acggcggact	aatcttcaac	tcctacatac	ttccccatt	attcctagaa	ccaggcgacc
7981	tgcgactcct	tgacgttgac	aatcgagtag	tactcccgat	tgaagcccc	attcgtataa
8041	taattacatc	acaagacgtc	ttgcactcat	gagctgtccc	cacattaggc	ttaaaaacag
8101	atgcaattcc	mggacgtcta	aaccaaacca	ctttcaccgc	tacacgaccr	gggggtatact
8161	acgggtcaatg	ctctgaaatc	tgtggagcaa	accacagttt	catgcccac	gtcctagaat
8221	taattcccct	aaaaatcttt	gaaataggrc	ccgtatttac	cctatagcac	cccctctacc
8281	ccctctagag	cccactgtaa	agctaactta	gcattaacct	tttaagttaa	agattaagag
8341	aaccaacacc	tctttacagt	gaaatgcccc	aactaaatac	taccgtatgg	cccaccataa
8401	ttacccccat	actccttaca	ctattcctca	tcacccaact	aaaaatatta	aacacaaact
8461	accacttacc	tccctcacca	aagcccataa	aaataaaaaa	ttataacaaa	ccctgagaac
8521	caaaatgaac	gaaaatctgt	tcgcttcatt	cattgcccc	acaatcctag	gcctaccgc
8581	cgcagtactg	atcattctat	ttccccctct	attgatcccc	acctccaaat	atctcatcaa
8641	caaccgacta	attaccaccc	aacaatgact	aatcaaacta	acctcaaaac	aatgatagc
8701	catacacaac	actaaaggac	gaacctgac	tcttatacta	gtatccttaa	tcatttttat
8761	tgccacaact	aacctcctcg	gactcctgcc	tcactcattt	acaccaacca	cccaactatc
8821	tataaaccta	gccatggcca	tccccctatg	agcgggca	gtgattatag	gctttcgctc
8881	taagattaaa	aatgccctag	cccacttcty	accacaaggc	acacctacac	cccttatccc
8941	catactagtt	attatcgaaa	ccatcagcct	actcattcaa	ccaatagccc	trggcgtacg
9001	cctaaccgct	aacattactg	caggccacct	actcatgcat	ctaattggaa	gcgccaccct
9061	agcaatatca	accattaacc	ttccctctac	acttatcatc	ttcacaattc	taattctact
9121	gactatccta	gaartcgctg	tcgcctttrac	ccaagcctac	gttttcacac	ttctagtaag
9181	cctctacctg	cacgacaaca	cataatgacc	caccaatcac	atgcctatca	tatagtaaaa
9241	cccagcccat	gaccctaacc	agggggccctc	tcagccctcc	taatgacctc	cggcctagcc
9301	atgtgatttc	acttccactc	cataacgctc	ctcatactag	gcctgctaac	caacacacta
9361	accatatacc	aatgatggcg	cgatgtaaca	cgagaaagca	cataccaagg	ccaccacaca
9421	ccacctgtcc	aaaaargcct	tcgatacggg	ataatcctat	ttattacctc	agaagttttt
9481	ttcttcgcag	gattttttctg	agccttttac	cactccagcc	tagcccctac	cccccaaata
9541	ggagggcact	ggcccccaac	aggcatcacc	ccgctaaatc	ccctagaagt	cccactccta
9601	aacacatccg	tattactcgc	atcaggagta	tcaatcacct	gagctcacca	tagtctaata
9661	gaaaacaacc	gaaaccaa	aattcaagca	ctgcttatta	caattttact	gggtctctat
9721	tttaccctcc	tacaagcctc	agagtacttc	garctctccct	tcaccatttc	cgacggcatc
9781	tacgggtcaa	cattttttgt	agccacaggc	ttccayggac	ttcacgtcat	tattgggtca
9841	actttcctca	ctatctgctt	catccgcca	ctaataatttc	actttacatc	caaacatcac
9901	tttggcttcg	aagccgccc	ctgatactgg	cattttgtag	atgtgggttg	actatttctg
9961	tatgtctcca	tctattgatg	agggctcttac	tcttttagta	taaatagtag	cgttaacttc

Figure N.1 Continued...

10021	caattaacta	gttttgacaa	cattcaaaaa	agagtaataa	acttcgcctt	aattttaata
10081	atcaacaccc	tcctagcctt	actactaata	attattacat	tttgactacc	acaactcaac
10141	ggctacatag	aaaaatccac	cccttacgag	tgcggcttcg	accctatatc	ccccgcccgc
10201	gtccctttct	ccataaaatt	cttcttagta	gctattacct	tcttattatt	tgatctagaa
10261	attgccctcc	ttttaccctt	accatgagcc	ctacaaacaa	ctaacctgcc	actaatagtt
10321	atgtcatccc	tcttattaat	catcatccta	gccctaagtc	tggcctatga	gtgactacaa
10381	aaaggattag	actga g ccga	attggtatat	agtttaaaca	aaacgaatga	tttcgactca
10441	ttaaattatg	ataatcatat	ttaccaaatg	cccctcattt	acataaatat	tatact r gca
10501	tttaccatct	cacttctagg	aatactagta	tatcgctcac	acctcatatc	ctccctacta
10561	tgcctagaag	gaataatact	atcgct a ttc	attatagcta	ctctcataac	cctcaacacc
10621	cactccctct	tagccaatat	tgtgcctatt	gccatactag	t ttttgccgc	ctgcgaagca
10681	gcggt a ggcc	tagccctact	agtctcaatc	tccaacacat	atggcctaga	ctacgtacat
10741	aacctaaacc	tactccaatg	ctaaaactaa	tcgtcccaac	aattatatta	ctaccactga
10801	catgact c tc	caaaaaacac	ataatttgaa	tcaacacaa	cacccacagc	ctaattatta
10861	gcatcatccc	c ct r ctat t	tttaaccaa	tcaacaacaa	cctattttagc	tg c tccc y aa
10921	ccttttcttc	cgaccc y cta	acaaccccc	tcctaatact	aactacctga	ctcctacccc
10981	tcacaatcat	ggcaagccaa	cgccacttat	ccagtgaacc	actatcacga	aaaaaactct
11041	acctctctat	actaatctcc	ctacaaatct	ccttaattat	aacattcaca	gccacagaac
11101	taatcatatt	ttatatcttc	ttcgaaacca	cacttatccc	caccttggct	atcatcacc
11161	gatgaggcaa	ccagccagaa	cgcctgaacg	caggcacata	cttcctattc	tacaccctag
11221	taggctccct	ttccctactc	atcgcaacta	tttactactca	caacacccta	ggctcactaa
11281	acattctact	act y ac y ctc	actgcccaag	aactatcaaa	ctcctgagcc	aacaacttaa
11341	tatgactagc	ttacacaata	gctttttatag	taaagatacc	tctttacgga	ctccacttat
11401	gactccctaa	agcccatgtc	gaagcccca	tcg y ggggtc	aatagtactt	gccgcagtac
11461	tcttaaaact	aggcggctat	ggtataatac	gcctcacact	cattctcaac	cccctgacaa
11521	aacacatagc	ctacccttct	cttgtactat	ccctatgagg	cataattata	acaagctcca
11581	tctgcctacg	acaaacagac	ctaaaatcgc	tcattgcata	ctcttcaatc	agccacatag
11641	ccctcgtagt	r acagccatt	ctcatccaaa	ccccctgaag	cttcaccggc	gcagtcattc
11701	tcataatcgc	ccacgg a ctt	acatcctcat	tactattctg	cctagcaaac	tcaaactacg
11761	aacgcactca	cagtcgcatc	ataatcctct	ctcaaggact	tcaaactcta	ctcccactaa
11821	tagctttttg	atgacttcta	gcaagcctcg	ctaacctcgc	cttaccctcc	actattaacc
11881	tactgggaga	actctctgtg	ctagtaacca	c attctcctg	atcaaatatc	actctcctac
11941	ttacaggact	caacatacta	gtcacagccc	tatactccct	ctacatattt	accacaacac
12001	aatg a ggctc	actcacccac	cacattaaca	acataaaaacc	ctcattcaca	cgagaaaaca
12061	ccctcat r tt	catacaccta	tccccattc	tcctcctatc	cctcaacccc	gacatcat y a
12121	ccgggttttc	ctcttgtaaa	tatagtttaa	ccaaaacatc	agattgtgaa	tctgacaaca
12181	gaggcttacg	acccttattt	taccgagaaa	gtcacacaaga	actgctaact	ca y gccccca
12241	tgtctaacaa	catggctttc	tcaactttta	aaggataaca	gctatccatt	ggtcttaggc
12301	cccaaaaatt	ttggtgcaac	tccaaataaa	agtaataacc	atgcacacta	ctataaccac
12361	cctaaccctg	acttccttaa	ttccccccat	ccttaccacc	ctcgtttaacc	ctaacaaaaa
12421	aaactcatac	ccccattatg	taaaatccat	tgtcgcatcc	acctttatta	tcagtctctt
12481	ccccacaaca	atattcatgt	gcctagacca	agaagttatt	atctcgaact	gacactgagc
12541	cacaacccaa	acaaccagc	tctccctaag	cttcaaacta	gactacttct	ccataatatt
12601	catccctgta	gcattgttcg	ttacatggtc	catcatagaa	ttctcactgt	gatataataa
12661	ctcagaccca	aacattaatc	agttcttcaa	atatctactc	at t ttcctaa	ttaccat r ct
12721	aatcttagtt	accgctaaca	acctattcca	actgttcac	ggctgagagg	gcgtaggaat
12781	tatatccttc	ttgtcatca	gttgatgata	cgcccagaca	gatgccacaa	cagcagccat
12841	tcaagcaatc	ctatacaacc	gtatcggcga	tatcggtttc	atcctcgcct	tagcatgatt
12901	tatcctacac	tccaactcat	gagaccacaa	acaaatagcc	cttctaaacg	ctaataccaag
12961	cctcacccca	ctactaggcc	tcctcctagc	agcagcaggc	aaatcagccc	aattaggy c t
13021	ccaccctga	ctccctcag	ccatagaagg	ccccaccca	gtctcagccc	tactccactc
13081	aagcactata	gttgtagcag	ga g tcttctt	actcatccgc	ttccac y ccc	tagcagaaaa

Figure N.1 Continued...

13141	tagcccacta	atccaaactc	taacactatg	cttaggcgct	atcaccactc	tgttcgcagc
13201	agtctgcgcc	cttacacaaa	atgacatcaa	aaaaatcgta	gcctttctcca	cttcaagtca
13261	actaggactc	atagtagtta	caatcggcat	caaccaacca	cacctagcat	tcctgcacat
13321	ctgtaccac	gcctttcttca	aagccatact	atttatgtgc	tccgggtcca	tcattccacaa
13381	ccttaacaat	gaacaagata	ttcgaaaaat	aggaggacta	ctcaaaaacca	tacctctcac
13441	ttcaacctcc	ctcaccattg	gcagcctagc	attagcagga	atacctttcc	tcacaggttt
13501	cta ^t ttccaaa	gaccacatca	tcgaaaccgc	aaacatatca	tacacaaacg	cctgagccct
13561	atctattact	ctcatcgcta	cctccct ^r ac	aagcgcctat	agcactcgaa	taattcttct
13621	caccctaaca	ggtcaacctc	gcttccct ^t ac	ccttactaac	attaacgaaa	ataaccccac
13681	cctactaaac	cccattaaac	gcctggcagc	cggaagccta	ttcgcaggat	ttctcattac
13741	taacaacatt	tcccc ^r cat	cccccttcca	aacaacaatc	cccctctacc	taaaactcac
13801	agccctcgct	gtcact ^y tcc	taggacttct	aacagcccta	gacctcaact	acctaaccac
13861	caaacttaaa	ataaaatccc	cactatgcac	attttatttc	tccaacatac	tcggattcta
13921	cccta ^s catc	acacaccgca	caatccccta	tctaggcctt	cttacgagcc	aaaacctgcc
13981	cctactcctc	ctagacctaa	cctgactaga	aaagcta ^y ta	cctaaaacaa	tttcacagca
14041	ccaaatctcc	acctccatca	tcacctcaac	ccaaaaaggc	ataattaaac	tttacttcct
14101	ctctttcttc	ttcccactca	tcctaaccct	actcctaate	acataaccta	ttcccccgag
14161	caatctcaat	tacaatat ^y	acaccaacaa	acaatgttca	accagtaact	actactaatc
14221	aacgcccata	atcatacaaa	gccccgcac	caataggatc	ctcccgaatc	aacctgacc
14281	cctctccttc	ataaattatt	cagcttccta	cactattaaa	gtttaccaca	accaccacc
14341	catcatactc	tttcaccac	agcacca ^y c	^y acctccat	cgctaacccc	actaaaacac
14401	tcaccaagac	ctcaaccct	gaccccatg	cctcaggata	ctcctcaata	gccatcgctg
14461	tagtatatcc	aaagacaacc	atcattcccc	ctaaataaat	taaaaaaact	attaaccaca
14521	tataacctcc	cccaaaattc	agaataataa	cacaccgcac	cacaccgcta	acaatcaata
14581	ctaaaccccc	ataaatagga	gaaggcttag	aagaaaaccc	cacaaacccc	attactaaac
14641	ccacactcaa	cagaaacaaa	gcatacatca	ttattctcgc	acggactaca	accacgacca
14701	atgatatgaa	aaaccatcgt	tgtatttcaa	ctacaagaac	accaatgacc	ccaatacgca
14761	aaa ^t taaccc	cctaataaaa	ttaattaaac	actcattcat	cgacctcccc	accccatcca
14821	acatctccgc	atgatgaaac	ttcggctcac	tccttggcgc	ctgcctgatc	ctccaaatca
14881	ccacaggact	attcctagcc	atgcactact	caccagacgc	ctcaaccgcc	ttttcatcaa
14941	tcgcccacat	cactcgagac	gtaaattatg	gctgaatcat	ccgctacctt	cacgccaatg
15001	gcgcctcaat	attctttatc	tgctctttcc	tacacatcgg	gcgaggccta	tattacggat
15061	catttctcta	ctcagaaacc	tgaaacatcg	gcattatcct	cctgcttgca	actatagcaa
15121	cagccttcat	aggctatgtc	ctcccgtgag	gccaaatatc	attctgaggg	gccacagtaa
15181	ttacaaactt	actatccgcc	atcccataca	ttgggacaga	cctagttcaa	tgaatctgag
15241	gaggctactc	agtagacagt	cccaccctca	cacgattctt	tacctttcac	ttcatcttgc
15301	ccttcattat	tgcagcccta	gcag ^r cactcc	acctcctatt	cttgcacgaa	acgggatcaa
15361	acaaccccc	aggaatcacc	tcccattccg	ataaaatcac	cttccaccct	tactacacaa
15421	tcaaagacgc	cctcggctta	cttctcttcc	ttctctcctt	aat ^r acatta	acactattct
15481	caccagacct	cctaggcgac	ccagacaatt	ataccctagc	caacccctta	aacacccctc
15541	cccacat ^y aa	gcccgaatga	tatttcctat	tcgcctacac	aattctccga	tccgtcccta
15601	acaaactagg	aggcgtcctt	gccctattac	tatccatcct	catcctagca	ataatcccca
15661	tcctccatat	atccaaacaa	caaagcataa	tatttcgccc	actaagccaa	tcactttatt
15721	gactcctagc	cgcagacctc	ctcattctaa	cctgaatcgg	aggacaacca	gtaagctacc
15781	cttttaccat	cattggacaa	gtagcatccg	tactatactt	cacaacaatc	ctaataccta
15841	taccaactat	ctccctaatt	gaaaacaaaa	tactcaaatg	ggcctgtcct	tgtagtataa
15901	actaatacac	cagtcttgta	aaccggar ^r at	gaaaacct ^y t	ttccaagg ^r c	aaatcagaga
15961	aaaagtcttt	aactccacca	ttagcaccca	aagctaagat	tctaatttaa	actattctct
16021	gttctttcat	ggggaagcag	atttggttac	cacccaagta	ttgactcacc	catcaacaac
16081	cgctatgtat	ttcgtaacat	actgccagcc	accatgaata	ttgtacggta	ccataaatac
16141	ttgaccacct	gtagtacata	aaamycca ^y	ccacatcaaa	acct ^t cccc	catgcttaca
16201	agcaag ^y aca	gyaatcaacc	^t tcaact ^g tc	ayacatca ^y	y ^y gcaactcca	aagccacccc

Figure N.1 Continued...

16261	tcacccacta	ggatay	caac	aaacctac	sc	acccttaaca	gtacatag	ca	cataaagcca
16321	tttaccgtac	atagcacatt	acagtcaa	at	cccttctcgt	cccatggat	gacccccctc		
16381	agataggggt	cccttgacca	ccatcctccg	tgaaatcaat	atcccgcaca	agagtgtac			
16441	tctcctcgct	ccgggcccat	aacacttggg	ggtagctaaa	gtgaactgta	tccgacatct			
16501	ggttcctact	tcaggg	ccat	aaagcctaaa	tagcccacac	gttcccctta	aataagacat		
16561	cacgatg								

Nucleotide numbering is based on the rCRS (Andrews *et al.*, 1999) and for this reason an N (indicated in pink) is retained at nucleotide position 3,107. The nucleotides indicated in red represent nucleotides that are different compared to the rCRS, and the nucleotides presented in blue indicate ambiguous nucleotide positions, i.e. one possible nucleotide is similar to the rCRS, while the other possible nucleotide differs from the rCRS (refer to Section 5.14.8). Insertions compared to the rCRS are indicated in green. a = adenine; c = cytosine; g = guanine; t = thymine; r = a or g; y = c or t; s = g or c; m = a or c; n = a, c, g or t.

APPENDIX O

Percentage similarity and pairwise number of differences observed between 43 Khoi-San mitochondrial sequences and the Khoi-San consensus sequence

This appendix contains a matrix of the percentage similarity and the pairwise number of nucleotide differences between 43 Khoi-San mitochondrial sequences and the Khoi-San consensus sequence constructed in the current investigation. The percentage similarity and the pairwise number of nucleotide differences are presented in the distance matrix presented in Table O.1.

The percentage similarity and the pairwise number of differences between the Khoi-San consensus sequence and each of the 43 individual Khoi-San sequences included in the current investigation were determined using PHYLIP version 3.66 (Felsenstein, 1989) and MEGA version 4 (Tamura *et al.*, 2007) as described in Section 4.17. These values were calculated by excluding positions containing gaps, therefore insertions and deletions were not taken into account in the determination of the percentage similarity or the number of pairwise differences between sequences.

QUICK REFERENCE GUIDE

Construction of a mitochondrial consensus sequence for the Khoi-San population of Southern Africa

M Koekemoer, Ph.D., August 2010

Datasets used for analyses

Dataset	Short description	No. of sequences	Sequence region
1a	All L Sequences	429 ^a	Coding region
1b	All L Sequences	426	Coding region
1c	All L Sequences (representative subset)	49 ^a	Coding region
2a	All Africa L Sequences	242 ^a	Coding region
2b	All Africa L Sequences	239	Coding region
2c	All Africa L Sequences (representative subset)	51 ^a	Coding region
3a	All Khoi-San L Sequences	58 ^a	Coding region
3b	All Khoi-San L Sequences	53 ^a	Complete sequence ^b
4a	!Kung Sequences	43	Complete sequence ^b
4b	!Kung Sequences	43	Coding region
5	L3 Sequences	128	Coding region
6	L2 Sequences	113	Coding region
7	L1 Sequences	63	Coding region
8a	L0 Sequences	103	Coding region
8b	L0 Sequences	100	Coding region

^a = number of sequences excluding the rCRS and the outgroup sequence.

^b = complete sequence consists of the control region and the coding region.

yellow = datasets used for phylogenetic analyses.

green = datasets used for statistical analyses.

orange = dataset used for construction of global L0-specific haplogroup network.

Refer to Section 4.10.4, Table 4.10 for a detailed description of the datasets.

Guide to symbols and colours used in phylogenetic analyses

Appendix J See page 599

Global and Africa phylogenetic trees of mitochondrial genome sequences

A. Global neighbour-joining tree

C. Africa neighbour-joining tree

B. Global maximum parsimony tree

D. Africa maximum parsimony tree

★ Khoi-San sequences generated in current investigation - **red** stars

● rCRS - **turquoise** octagon

■ *Pan troglodytes* outgroup sequence - **green** square

▲ African reference sequence - **violet** triangle

Sub-haplogroups

■ L0a, L1a1, L2a, L3a1

■ L0b, L1a2, L2b, L3a2

■ L0c1, L1b1, L2c, L3a3

■ L0c2, L1b2, L2d, L3b

■ L3c

Coloured boxes - subdivisions identified in sub-haplogroups described in the classification scheme used in the current investigation (refer to Section 4.9)

grey boxes - sub-haplogroups assigned based on the alternative classification scheme (refer to Section 3.3.1.1)

red boxes - new sub-haplogroups, designation not previously assigned in the classification scheme used in the current investigation

— black lines - sequences could not be classified completely

--- dotted lines - sequences exhibit a discrepancy between their location in the phylogenetic tree and their haplogroup classification

Bootstrap analysis (1,000 replicates) was performed and the bootstrap values for the major branches and groups are indicated. Haplogroup classification was performed as described in Section 4.9. The major haplogroups and sub-haplogroups are indicated at the relevant branching points. All the relevant information regarding the mitochondrial sequences used in the construction of the phylogenetic trees is listed in Appendix G.

Guide to symbols and colours used in global L0-specific haplogroup network

Appendix L See page 611

Global L0-specific haplogroup network

* = starting point of the network

Polymorphisms

black = non-synonymous substitution (amino acid, position, amino acid, gene)

Conservation numbers are indicated after each non-synonymous substitution

red = synonymous substitution (gene)

green = tRNA or rRNA (gene)

blue = non-coding nucleotide in coding region

bold = polymorphism occurs in multiple branches

italics = polymorphism reverted to the nucleotide present in the rCRS

Sequence names

brackets and **bold** = sequence name or GenBank® accession number

grey = sequences generated in the current investigation

Branching

dashed lines = new branches within the L0 network

big dashes (---) = branches added by published sequences

small dashes (.....) = branches added by sequences generated in the current investigation

coloured boxes (pink, orange, light blue, violet) = new internal and terminal branches

I = internal branches

T = terminal branches

Guide to colours used to indicate geographic origin/distribution of mitochondrial genome sequences

Region in Africa		Region outside Africa	
	Africa		USA (African American)
	Central Africa		USA (Hispanic)
	Eastern Africa		Dominican Republic
	Northern Africa		
	Southern Africa		
	Western Africa		

Summary of results from statistical analysis

Dataset	Deviation from neutrality	Reason
1b	Yes	Population expansion, coupled with the effect of purifying selection
2b, 5		Population expansion
3a, 4b, 7	No	No rapid population expansion, or effect of purifying selection
6	Yes	Population expansion event
8a	Yes	Population expansion event, coupled with the effect of purifying selection

Refer to Section 5.14.6 for a more detailed summary of the results obtained for the datasets.

Comparison of Wallace, 2004 and alternative classification schemes for haplogroup L0

Haplogroup	Haplogroup classification scheme	
	Wallace, 2004 ^a	Alternative ^b
L0	L0a	L0k
	---	L0d3 ^c
	---	L0f
	L0b	L0d
	---	L0d1
	---	L0d2
	L0c	L0a
	L0c1	L0a1
	L0c2	L0a2

^a = classification scheme used in current investigation (refer to Section 4.9).

^b = classification scheme used in other investigations (refer to Section 3.3.1.1).

^c = new branch first described by Gonder *et al.* (2007).

Refer to Appendix F for a more detailed comparison of the two classification schemes.