Comparison of complete mitochondrial genome sequence between different ethnic groups from Southern Africa

BY

JAKE DARBY, B.Sc. (Agric)

Dissertation submitted for the degree Magister Scientiae (M.Sc.)
in Biochemistry at the North-West University

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

CO-SUPERVISOR: Doctor Izelle Smuts
Department of Paediatrics, University of Pretoria

December 2004

Vergelyking van volledige mitochondriale-genoom-volgorde tussen verskillende etniese groepe van Suider-Afrika

DEUR

JAKE DARBY, B.Sc. (Agric)

Verhandeling voorgelê vir die graad Magister Scientiae (M.Sc.)
in Biochemie aan die Noordwes-Universiteit

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

MEDESTUDIELEIER: Dokter Izelle Smuts
Departement Pediatrie, Universiteit van Pretoria

Desember 2004

To my family

ABSTRACT

The human mitochondrial genome is a 16,569 base pair double-stranded deoxyribonucleic acid (DNA) molecule located in the mitochondrion. The mitochondrial genome sequence is conserved, maternally inherited and undergoes no recombination, making its evaluation ideal for evolutionary studies. Certain alterations in the genome are unique to specific human populations and haplogroups. The genetic background, or haplogroup of an individual, may act in concert with disease associated mutations. The ethnicity of an individual is often utilised as an indicator of haplogroup.

During this investigation the full mitochondrial sequence of 10 individuals belonging to three ethnic Southern African populations, namely three Xhosa, three Zulu and four Tswana individuals, was generated. The complete nucleotide sequences were compared to one another in order to determine the genetic relationship between individuals. Sequences were also evaluated against the 2001 revised Cambridge reference sequence (RCRS) to detect novel alterations as well as alterations present in all individuals analysed.

A total of 222 alterations (207 previously reported, 15 unreported) were detected relative to the RCRS. Five length alterations and 207 nucleotide substitutions were detected. Ninety-eight alterations were detected only once, and 115 were detected in at least two individuals. Haplogroup analysis clustered individuals into haplogroups L0, L2 and L3. No clear correlation between haplogroup assignment and ethnic origin could be observed. The distribution of shared alterations between individuals was in agreement with the haplogroup clustering.

Ethnicity can therefore not be utilised as an indicator of haplogroup in the context of the current study. To investigate the association between disease presentation and haplogroup, individuals will have to be randomly sampled and then haplogrouped. Reasons to substantiate the lack of association between ethnicity and haplogroups include the occurrence of gene flow between populations, inaccurate ethnic and incomplete haplogroup classification and populations not being sufficiently divergent.

OPSOMMING

Die menslike mitochondriale genoom is 'n 16,569 basispaar dubbeldraad-deoksieribonukléinsuur (DNS) molekule wat aangetref word in die mitochondrion. Die mitochondriale genoom is gekonserveerd, word maternaal oorgeërf en ondergaan geen rekombinasie nie. Hierdie aspekte maak die evaluasie daarvan ideaal vir evolusionêre studies. Sekere veranderinge in die genoom is uniek tot spesifieke menslike bevolkings en haplogroepe. Die genetiese agtergrond, of haplogroep van individue, kan in samewerking met siektegeassosieerde mutasies optree. Die etnisiteit van 'n individu word dikwels gebruik as 'n aanwyser van haplogroep.

Tydens hierdie ondersoek is die volledige mitochondriale volgorde van 10 individue uit drie verskillende etniese bevolkings in Suider-Afrika, naamlik drie Zulu-, drie Xhosa- en vier Tswana-individue, gegenereer. Die volledige nukleotiedvolgordes is met mekaar vergelyk om die genetiese verwantskap tussen individue te bepaal. Volgordes is ook vergelyk met die 2001 hersiene Cambridge verwysingsvolgorde (RCRS) om nuwe veranderinge, sowel as veranderinge wat voorkom in alle individue wat bestudeer is, te identifiseer.

'n Totaal van 222 veranderinge (207 voorheen gerapporteer, 15 ongerapporteer) is geïdentifiseer relatief tot die RCRS. Sewe lengteveranderinge en 207 nukleotiedvervangings is waargeneem. Agt-en-negentig veranderinge is slegs een keer waargeneem, en 115 is waargeneem tussen ten minste twee individue. Analise van die haplogroepe het die individue in haplogroepe L0, L2 en L3 gegroepeer. Geen duidelike korrelasie tussen die toedeling van haplogroepe en etniese oorsprong kon waargeneem word nie. Die verspreiding van gedeelde veranderinge tussen individue was in ooreenstemming met die groepering van die haplogroepe.

Teen die agtergrond van die huidige studie kan etnisiteit dus nie gebruik word as 'n aanwyser van haplogroep nie. Om assosiasies tussen fenotipe en haplogroep te bestudeer, sal individue ewekansig gekies en gehaplogroepeer moet word. Redes vir die gebrek aan assosiasie tussen etnisiteit en haplogroep sluit in geenvloei tussen bevolkings, onakkurate etniese en onvolledige haplogroepklassifikasie en dat bevolkings nie uiteenlopend genoeg is nie.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS AND SYMBOLS i
LIST OF EQUATION v
LIST OF FIGURES vi
LIST OF TABLES vii
ACKNOWLEDGEMENTS viii

CHAPTER ONE

INTRODUCTION 1

CHAPTER TWO

MOLECULAR AND EVOLUTIONARY ASPECTS OF MITOCHONDRIAL DNA 4

2.1 HUMAN ORIGINS AND EARLY MIGRATIONS 5
2.1.1 The emergence of man 6
2.2 THE MITOCHONDRION 7
2.2.1 Inheritance 7
2.2.2 Mitochondrial structure and function 8
2.3 MITOCHONDRIAL GENETICS 9
2.3.1 Mitochondrial replication and protein expression 10
2.4 CALIBRATION OF MOLECULAR PHYLOGENIES 11
2.5 CAMBRIDGE REFERENCE SEQUENCE 11
2.6 ORIGIN OF THE MITOCHONDRION 12
2.7 MITOCHONDRIAL GENETIC VARIATION 12
2.7.1 Genome evolution 12
2.7.2 Implications of variation 13
2.7.2.1 Adaptive variation 13
2.7.2.2 Mutations and human disorders 14
2.8 MITOCHONDRIAL PHYLOGENIES 15
2.8.1 Methods of assessment of variation 16
2.8.1.1 Restriction fragment length polymorphism analysis 16
2.8.1.2 Sequence analysis 18
2.8.2 Mitochondrial haplogroups 19
2.8.2.1 Global haplogroups 19
2.8.2.2 Haplogroup evolution 22
2.8.2.3 Deducing human migrations 24
2.9 AFRICAN EVOLUTION 25
2.10 LINGUISTIC INFLUENCE ON DIVERSITY 28
2.11 AIMS 30
2.11.1 Specific aims of this investigation 31

CHAPTER THREE

MATERIALS AND METHODS 32

3.1 SAMPLE POPULATION 32
3.2 DNA ISOLATION 33
3.3 AMPLIFICATION OF THE MITOCHONDRIAL GENOME 34
3.3.1 Polymerase chain reaction 34
3.3.1.1 Long PCR 35
3.3.1.2 Short PCR 37
3.3.2 Gel electrophoresis 40
3.4 CHAIN TERMINATION CYCLE SEQUENCING 40
3.4.1 Purification of amplified product 41
3.4.2 Chain termination sequencing 41
3.5 SEQUENCE ANALYSIS 43
3.5.1 Nucleotide differences 44
3.5.2 Haplogroup assignment 44
3.5.3 Phylogenetic analysis 45

CHAPTER FOUR

RESULTS AND DISCUSSION 46

4.1 OPTIMISATION OF EXPERIMENTAL PROTOCOLS 46
4.1.1 Polymerase chain reaction 46
4.1.1.1 Long PCR strategy 47
4.1.1.2 Short PCR strategy 48
4.1.2 PCR product electrophoresis and purification 48
4.1.3 Cycle sequencing 50
4.2 COMPLETE MITOCHONDRIAL GENOME SCREENING 51
4.2.1 The 12S ribosomal RNA gene 53
4.2.2 The 16S ribosomal RNA gene 54
4.2.2.1 The T2887C alteration in the 16S rRNA gene 55
4.2.2.2 The C3157T alteration in the 16S rRNA gene 56
4.2.3 The transfer RNA leucine 1 gene 57
4.2.4 NADH dehydrogenase subunit 1 gene 57
4.2.4.1 The T3618C alteration in the ND1 gene 58
4.2.5 The transfer RNA isoleucine gene 58
4.2.6 The transfer RNA glutamine gene 59
4.2.7 NADH dehydrogenase subunit 2 gene 59
4.2.8 The transfer RNA tryptophan gene 59
4.2.9 The transfer RNA alanine gene 60
4.2.10 The transfer RNA cysteine gene 60
4.2.11 Non-coding region between tRNA tyrosine and cytochrome *c* oxidase subunit I 60
4.2.12 The cytochrome *c* oxidase subunit I gene 61
4.2.12.1 The T6614C alteration in the COI gene 62
4.2.12.2 The A6806G alteration in the COI gene 62
4.2.13 The transfer RNA aspartic acid gene 63
4.2.14 The cytochrome *c* oxidase subunit II gene 63
4.2.15 The ATP synthase F0 subunit 8 gene 64
4.2.15.1 The T8503C alteration in the ATPase8 gene 64
4.2.16 The A8566G alteration in the ATPase8 and ATPase6 genes 65

4.2.17 The ATP synthase F0 subunit 6 gene 65
4.2.18 The cytochrome *c* oxidase subunit III gene 66
4.2.18.1 The A9350G alteration in the COIII gene 67
4.2.18.2 The T9637G alteration in the COIII gene 68
4.2.19 The NADH dehydrogenase subunit 3 gene 69
4.2.20 The NADH dehydrogenase subunit 4L gene 69
4.2.21 The NADH dehydrogenase subunit 4 gene 70
4.2.22 The transfer RNA histidine gene 71
4.2.23 The transfer RNA serine gene 71
4.2.24 The NADH dehydrogenase subunit 5 gene 71
4.2.24.1 The C12348T alteration in the ND5 gene 72
4.2.24.2 The C12436T alteration in the ND5 gene 73
4.2.24.3 The C12798T alteration in the ND5 gene 74
4.2.25 The NADH dehydrogenase subunit 6 gene 75
4.2.25.1 The A14176G alteration in the ND6 gene 75
4.2.25.2 The C14407T alteration in the ND6 gene 76
4.2.26 The cytochrome *b* gene 77
4.2.26.1 The A15692G alteration in the cyt *b* gene 78
4.2.27 The transfer RNA threonine gene 79
4.2.28 The D-loop region 79
4.2.29 SEQUENCE COMPARISON 84
4.2.29.1 Haplogroup analysis 88
4.2.30 CONSTRUCTION OF PHYLOGENETIC TREES 93
4.2.30.1 UPGMA tree 93
4.2.30.2 MP tree 94
4.2.30.3 Comparison of phylogenetic trees 96

CHAPTER FIVE

CONCLUSIONS 98

5.1 ALTERATIONS OBSERVED 99
5.2 HAPLOGROUP DETERMINATION 100
5.3 PHYLOGENETIC ANALYSIS 100
5.4 THE ROLE OF ETHNICITY IN GENETIC STUDIES 101

REFERENCES 104

6.1 GENERAL REFERENCES 104
6.2 ELECTRONIC REFERENCES 110

APPENDIX A 112

WHOLE MTDNA GENOME ALTERATIONS RELATIVE TO THE RCRS 112

APPENDIX B 118

DISTRIBUTION OF FRAGMENT TYPES UTILISED FOR SEQUENCING 118

APPENDIX C..119

ALTERATIONS CONTAINED IN SEQUENCE MOTIFS......................119

LIST OF ABBREVIATIONS AND SYMBOLS

LIST OF SYMBOLS

12S RNA	12S ribosomal RNA
16S RNA	16S ribosomal RNA
°C	degrees Celsius
μ	micro: 10^{-6}
%	percentage
+	addition
$\pm$	addition and subtraction
-	subtraction
I	Complex I
II	Complex II
III	Complex III
IV	Complex IV
V	Complex V

LIST OF ABBREVIATIONS

A	adenine (in DNA sequence)
A_{260}/A_{280}	ratio of absorbance measured at 260 nm and 280 nm
ADP	adenosine diphosphate
Ala	alanine
Alu I	restriction endonuclease isolated from *Arthrobacter luteus*, with recognition 5'-AG↓CT-3'
Arg	arginine
Asn	asparagine
ATP	adenosine triphosphate
ATPase6	ATP synthase F0 subunit 6
ATPase8	ATP synthase F0 subunit 8
Ava II	restriction endonuclease isolated from *Anabaena variabilis*, with recognition site 5'-GG(A/T)↓CC-3'
Bam HI	restriction endonuclease isolated from *Bacillus amyloliquefaciens* H., with recognition site 5'-G↓GATC-3'
bp	base pairs
*Bst*O I	restriction endonuclease isolated from *Bacillus stearothermophilus*, with recognition site 5'-CC↓(A/T)GG-3'
C	cytosine (in DNA sequence)
ca.	circa: approximately
Ca^{2+}	calcium cation
CHAPS	3-[(3-cholamidopropyl) dimethylammonio]-1-propanesulfonate
COI	cytochrome *c* oxidase subunit I
COII	cytochrome *c* oxidase subunit II
COIII	cytochrome *c* oxidase subunit III
CO_2	carbon dioxide
CoQ	coenzyme Q
CR	control region
CRS	Cambridge Reference Sequence
cyt *b*	cytochrome *b*
cyt *c*	cytochrome *c*
dATP	2'-deoxyadenosine-5'-triphosphate
dCTP	2'-deoxythymidine-5'-triphosphate
Dde I	restriction endonuclease isolated from *Desulfovibrio desulfuricans*, with recognition site 5'-C↓TNAG-3'
ddH_2O	double distilled water
del	deletion
dGTP	2'-deoxyguanosine-5'-triphosphate
D-loop	displacement loop
DNA	deoxyribonucleic acid

DTT	dithiothreitol
dTTP	2'-deoxythymidine-5'-triphosphate
EDTA	ethylenediamine tetra-acetic acid
et al.	*et alii*: Latin abbreviation for "and others"
EtBr	ethidium bromide
F	forward primer
$FADH_2$	reduced flavin adenine dinucleotide
Fe-S	iron sulphur
g	grams
G	guanine (in DNA sequence)
gDNA	genomic DNA
GenBank[1]	GenBank®: United States repository of DNA sequence information
Glu	glutamic acid
Gly	glycine
H_2O	water
Hae II	restriction endonuclease isolated from *Haemophilus aegyptius*, with recognition site 5'-(A/G)GC↓GC(T/G)-3'
Hae III	restriction endonuclease isolated from *Haemophilus aegyptius*, with recognition site 5'-GG↓C-3'
HeLa	cervical cancer cells from Henrietta Lacks
Hha I	restriction endonuclease isolated from *Haemophilus haemolyticus*, with recognition site 5'-GT(T/C)↓(A/G)AC-3'
Hinc II	restriction endonuclease isolated from *Haemophilus influenzae* C., with recognition site 5'-GT(T/C)↓(A/G)AC
Hinf I	restriction endonuclease isolated from *Haemophilus influenzae* Rf., with recognition site 5'G AN↓TC-3'
His	histidine
Hpa I	restriction endonuclease isolated from *Haemophilus parainfluenzae*, with recognition site 5'-C↓GG-3'
HR-RFLP	high resolution RFLP analysis
H-strand	heavy-strand
HVS-I	hypervariable segment I
HVS-II	hypervariable segment II
i.e.	that is to say
Ile	isoleucine
Inc.	incorporated
ins	insertion
IT_{H1}	first site of initiation for transcription of H-strand
IT_{H2}	second site of initiation for transcription of H-strand
IT_L	site of initiation for transcription of L-strand
kb	kilobase pairs
KCl	potassium chloride
KH_2PO_4	potassium phosphate monobasic
Leu	leucine
LHON	Leber's hereditary optic neuropathy
L-strand	light-strand
µg	microgram
$\mu g.ml^{-1}$	microgram per millilitre
µl	microlitre
µm	micrometre
µM	micromolar
M	molar
Mbo I	restriction endonucleases isolated from an *E. coli* strain that carries the cloned *Mbo* I gene from *Moraxella bovis*, with recognition site 5'-GA↓TC-3'
MBS	multiblock system
MEGA	Molecular Evolutionary Genetics Analysis
Met	methionine
mg	milligram
$MgCl_2$	magnesium chloride
$mg.ml^{-1}$	milligram per millilitre
$MgSO_4$	magnesium sulphate

[1] GenBank® is a registered trademark of the National Institute of Health and Human Services for the Genetic Sequence Data Bank, Bethesda, MD, USA.

min	minutes
ml	millilitre
mM	millimolar
MP	maximum parsimony
MRCA	most recent common ancestor
mRNA	messenger RNA
Msp I	restriction endonucleases isolated from *Moraxella* species, with recognition 5'-GA↓TC-3' and purified from *E. coli*
mtDNA	mitochondrial DNA
Myr	million years
n	nano: 10^{-9}
N.A.	not applicable
NaCl	sodium chloride
NADH	reduced nicotinamide adenine dinucleotide
Na_2HPO_4	disodium hydrogen phosphate
NaOAc	sodium acetate
NARP	neurogenic muscle weakness, ataxia and retinitis pigmentosa
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
nDNA	nuclear DNA
NEG	negative control
ng	nanograms
$ng.\mu l^{-1}$	nanograms per microlitre
$(NH_4)_2SO_4$	ammonium sulphate
NJ	neighbour-joining
Nla III	restriction endonuclease isolated from *Neisseria lactamica*, with recognition site 5'CATG↓-3'
nm	nanometre
np	nucleotide position
nt	nucleotide
O_2	oxygen
O_H	origin of H-strand synthesis
O_L	origin of L-strand synthesis
OXPHOS	oxidative phosphorylation
p	pico: 10^{-12}
PAGE	polyacrylamide gel electrophoresis
PBS	phosphate buffered saline
PCR	polymerase chain reaction
Pfu	*Pyrococcus furiosus*
pH	indicates acidity: numerically equal to the negative logarithm of H^+ concentration expressed in molarity
Phe	phenylalanine
pmol	picomol
POS	positive control
POWIRS	Profile of Obese Woman with Insulin Resistance Syndrome
Pro	proline
R	reverse primer
RCRS	Revised Cambridge Reference Sequence
rDNA	ribosomal DNA
RE	restriction enzyme
RFLP	restriction fragment length polymorphism
RNA	ribonucleic acid
rRNA	ribosomal RNA
Rsa I	restriction endonuclease isolated from *Rhodopseudomonas sphaeroides*, with recognition site 5'-GT↓AC-3'
RT	room temperature
S	Svedberg units
Ser	serine
SNP	single nucleotide polymorphism

T	thymine (in DNA sequence)
T_a	estimated annealing temperature
Taq	*Thermus aquaticus*
Taq I	restriction endonuclease isolated from *Thermus aquaticus* YTI., with recognition site 5'-T↓CGA-3'
TBE	89.15 mM Tris base [pH 8.0], 88.95 mM boric acid and 2.5 mM di-sodium ethylenediamine tetra-acetic acid
Thr	threonine
T_m	calculated annealing temperature
Tris[1]	Tris®: tris (hydroxymethyl)aminomethan:2-amino-2-(hydroxymethyl)-1,3 propanediol: $C_4H_{11}NO_3$
Tris-HCl	Tris®-hydrochloride
Triton[2] X-100	octylphenolpoly(ethylene-glycolether)$_n$, for n = 10
tRNA	transfer RNA
$tRNA^{Ala}$	tRNA alanine
$tRNA^{Asp}$	tRNA aspartic acid
$tRNA^{Cys}$	tRNA cysteine
$tRNA^{Gln}$	tRNA glutamine
$tRNA^{His}$	tRNA histidine
$tRNA^{Ile}$	tRNA isoleucine
$tRNA^{Leu}$	tRNA leucine
$tRNA^{Phe}$	tRNA phenylalanine
$tRNA^{Pro}$	tRNA proline
$tRNA^{Ser}$	tRNA serine
$tRNA^{Thr}$	tRNA threonine
$tRNA^{Trp}$	tRNA tryptophan
$tRNA^{Tyr}$	tRNA tyrosine
Trp	tryptophan
TS	transition
TV	transversion
U	uracil (in the context of one of the four base pairs found in RNA) or units
$U.\mu l^{-1}$	units per microlitre
UPGMA	unweighted pair-group method using arithmetic averages
USA	United States of America
v	version
Val	valine
$V.cm^{-1}$	volts per centimetre
w/v	weight per volume
YBP	years before present
x *g*	multiplied by gravitational force

[1] Tris® is a registered trademark of the United States Biochemical Corporation, Cleveland, OH, USA.
[2] Triton® is a registered trademark of Rohm & Haas Company, Philadelphia, PA, USA.

LIST OF EQUATION

Equation no. **Title of Equation** **Page**

3.1: Calculation of the estimated annealing temperature 36

LIST OF FIGURES

Figure no.	Title of Figure	Page
2.1:	Global mitochondrial phylogeny for human mtDNA	18
2.2:	Global migrations of human mtDNA	25
2.3:	African haplogroups	27
4.1:	Photographic representation of PCR fragments generated via long PCR	49
4.2:	Photographic representation of PCR fragments 1, 2, 3, 4, 5, 6 and 9 generated via short PCR	50
4.3:	Representative electropherogram of the T2887C alteration in the 16S rRNA gene	55
4.4:	Representative electropherogram of the C3157T alteration in the 16S rRNA gene	56
4.5:	Representative electropherogram of the T3618C alteration in the ND1 gene	58
4.6:	Representative electropherogram of the T6614C alteration in the COI gene	62
4.7:	Representative electropherogram of the A6806G alteration in the COI gene	63
4.8:	Representative electropherogram of the T8503C alteration in the ATPase8 gene	65
4.9:	Representative electropherogram of the A9350G alteration in the COIII gene	67
4.10:	Representative electropherogram of the T9637G alteration in the COIII gene	69
4.11:	Representative electropherogram of the C12348T alteration in the ND5 gene	72
4.12:	Representative electropherogram of the C12436T alteration in the ND5 gene	73
4.13:	Representative electropherogram of the C12798T alteration in the ND5 gene	74
4.14:	Representative electropherogram of the A14176G alteration in the ND6 gene	76
4.15:	Representative electropherogram of the C14407T alteration in the ND5 gene	77
4.16:	Representative electropherogram of the A15692G alteration in the cyt *b* gene	79
4.17:	Representative electropherogram of the 568 - 577insPolyC	83
4.18:	UPGMA tree of ten individuals of three ethnic origins	94
4.19:	MP tree of ten individuals from three ethnic origins	95

LIST OF TABLES

Table no.	Title of Table	Page
2.1:	Mitochondrial polymorphisms defining macro-haplogroup L	20
2.2:	Mitochondrial polymorphisms defining European haplogroups	21
2.3:	Mitochondrial polymorphisms defining Asian haplogroups	22
2.4:	Populations containing Asian mtDNA haplogroups	22
2.5:	Divergence times of Native American haplogroups	23
2.6:	Sequence divergence times of African macro-haplogroup L	26
3.1:	Primers utilised for amplification of the complete mitochondrial genome via a long PCR strategy	35
3.2:	Primer pairs utilised for amplification and sequencing of the complete mitochondrial genome in short fragments	37
4.1:	Annealing temperatures for amplification of fragments via the long PCR strategy	47
4.2:	Alterations detected in the 12S rRNA gene	54
4.3:	Alterations detected in the 16S rRNA gene	54
4.4:	Alterations detected in the ND1 gene	57
4.5:	Alterations detected in the ND2 gene	59
4.6:	Alterations detected in the COI gene	61
4.7:	Alterations detected in the COII gene	64
4.8:	Alterations detected in the ATPase8 gene	64
4.9:	Alterations detected in the ATPase6 gene	66
4.10:	Alterations detected in the COIII gene	66
4.11:	Alterations detected in the ND4L gene	70
4.12:	Alterations detected in the ND4 gene	70
4.13:	Alterations detected in the ND5 gene	71
4.14:	Alterations detected in the ND6 gene	75
4.15:	Alterations detected in the cyt b gene	78
4.16:	Alterations detected in the tRNAThr gene	79
4.17:	Nucleotide substitutions detected in the D-loop	80
4.18:	Insertions and deletions detected in the D-loop	82
4.19:	The division of nucleotide substitutions	85
4.20:	Observed number of nucleotide differences between individuals	87
4.21:	Combinations of alterations between individuals	88
4.22:	Haplogroup assignment of individuals of different ethnic origin	89
A.1	Sequence differences between individuals analysed and the RCRS	112
B.1:	Fragments of individuals sequenced utilising short and long PCR products	118
C.1:	Alterations in specific motifs	119

ACKNOWLEDGEMENTS

During the course of this investigation, certain individuals and institutions played an imperative role, whose input, dedication and patience, made the completion thereof possible. I would like to thank the following individuals and institutions with the greatest of appreciation.

To the patients who participated in this study, without whom the project could not have been initiated.

To my supervisor, **Prof. Antonel Olckers**, for giving me the opportunity to continue my studies under her guidance. The insight, knowledge and management skills she passed on will remain with me always. Whose time, dedication and above all patience, has allowed me to learn unsurpassed values. **Dr. Izelle Smuts**, my co-supervisor, whose knowledge, enthusiasm and dedication made the completion of the project possible. I would like to thank both my supervisor and co-supervisor for their time invested in me. **Dr. Alta Schutte**, for the organisation and recruitment of patients and all the collaborators making POWIRS possible.

Wayne Towers, my mentor, for his friendship, for passing on his superb writing skills and knowledge and most of all, for all his help, without which I could not have achieved my goals. To **Dr. Annelize van der Merwe**, my temporary mentor and guide, who has always been willing to provide me with the highest quality of knowledge. My fellow students **Marco Alessandrini, Desiré-Lee Dalton, Michelle Freeman, Dan Isabirye** and **Tumi Semete** for their support and assistance that has helped me through the year. The extended team **Desiré Hart** and **Martha Sebogoli** for making everything run smoothly.

North-West University, for excellent administrative and library services. The **Centre for Genome Research** and **DNAbiotec (Pty) Ltd**, for providing the resources which were indispensable for the completion of my studies.

My love, **Jantje**, for whom I thought I would be searching for most of my life. **My parents** and **brother**, for their love, friendship and support, which without I would have achieved nothing. An eternity of standing on a rock in the darkness would be more pleasurable than being without you. To **God**, our creator, for providing me with strength and allowing me to live.

CHAPTER ONE

INTRODUCTION

Mitochondrial function is essential for the maintenance of a cell (Borst, 1977) as it forms the primary location for metabolic functions such as the citric acid cycle, fatty acid oxidation, electron transport and oxidative phosphorylation (OXPHOS). The mitochondrial genome is a closed circular, double-stranded DNA molecule of 16,569 base pairs (bp) which encodes two ribosomal ribonucleic acids (rRNA), 22 transfer RNAs (tRNA) and 13 polypeptides (Anderson *et al.*, 1981). The mitochondrion is strictly maternally inherited (Giles *et al.*, 1980), its genome undergoes no recombination (Wallace, 1994) and has a mutation rate which is 10 times greater than that of the nuclear genome (Brown *et al.*, 1979). The mitochondrial genome is therefore ideal for genetic evolutionary studies.

Populations harbour unique and distinctive alterations that reflect their evolutionary history (Cann *et al.*, 1987). These alterations result in the formation of mitochondrial lineages, which are termed haplogroups (Denaro *et al.*, 1981). Haplogroups can therefore be utilised to identify groups that have similar evolutionary histories and determine the relationship between populations (Wallace *et al.*, 1999).

As energy metabolism is centred on the mitochondrion, alterations of its genome may cause a compromise in energy production, often resulting in dysfunction or disease (Wallace *et al.*, 1992). Alterations may also result in adaptations, such as climatic adaptations (Mishmar *et al.*, 2003), increased longevity (Ross *et al.*, 2001) and an adaptation to high altitudes (Torroni *et al.*, 1994c).). However, mitochondrial diseases are common, with a prevalence of 6.57 per 100,000 adults, as observed in the North East England population, which is comparable to Huntington's disease and Duchenne's muscular dystrophy (Chinnery *et al.*, 2000).

Several disease phenotypes have been observed to be prevalent in specific haplogroups (Shoffner *et al.*, 1993; Brown *et al.*, 2001). This is as a result of differential functionality between different haplogroups (Torroni, 2000). Differences in the genetic architecture of populations may cause different responses to environmental risks and disease-associated mutations (Tishkoff and Williams, 2002). These differences are manifested through the

increased penetrance of certain conditions in specific haplogroups, such as Leber's hereditary optic neuropathy (LHON) in haplogroup J (Brown et al., 2001) as well as the penetrance of specific conditions in haplogroup J individuals suffering from LHON, late-onset Alzheimer's disease in haplogroup H (Shoffner *et al.*, 1993) and asthenozoospermia in haplogroup H (Ruiz-Pesini *et al.*, 2000). Haplogroup-associated alterations, contained in the mitochondrial and nuclear genomes, may aid in the presentation of disease by acting in conjunction with disease-causing mutations and thereby lowering the disease threshold limit (Hofmann *et al.*, 1997).

The greatest differences in mtDNA sequence occur between the African population and the rest of the world. The African population is considered unique as it is genetically distinct (Johnson *et al.*, 1983), the most diverse (Cann *et al.*, 1987) and thus the oldest of all populations. African populations are highly subdivided, harbouring a great deal of genetic variation (Salas *et al.*, 2002). No single African population can therefore be representative of all African populations (Tishkoff and Williams, 2002).

Black Southern African populations appear to have distinct genetic aetiologies regarding mitochondrial disorders, not common to the rest of the world (Olckers *et al.*, 2001). The black South African population belongs to macro-haplogroup L (Chen *et al.*, 2000). Haplogroup L is divided into haplogroups L0, L1 and L2, which are limited to the African continent, and L3, which is detected in Africa and other parts of the world (Chen *et al.*, 2000).

A common language often represents a common origin, and language can thus be utilised to classify an ethnic group (Cavalli-Sforza *et al.*, 1988; Cavalli-Sforza *et al.*, 1994). An association between haplogroup and ethnic origin exists (Brown, 1980) to the extent that a correlation, in the global context, exists between the haplogroup and language of a population (Cavalli-Sforza *et al.*, 1988). Ethnic origin is therefore often utilised as an indicator to which haplogroup an individual belongs. However, limited or no correlation is often observed in certain populations (Rosser *et al.*, 2000). This investigation represents a pilot study aimed at determining whether the ethnic status of Southern African ethnic populations provides an indication of the haplogroup to which the population belongs.

The aims of the investigation are presented at the end of Chapter Two, and are preceded by a broad literature review. The methodology utilised to achieve these goals is described in Chapter Three, and the results and discussion thereof presented in Chapter Four.

Conclusions based on the results generated in this study are presented in Chapter Five. Supplementary information on the results presented in Chapter Four, are listed in appendices A to C.

CHAPTER TWO

MOLECULAR AND EVOLUTIONARY ASPECTS OF MITOCHONDRIAL DNA

The origin of the human species has intrigued humans for centuries. With the advent of modern molecular tools, it was thought that the mystery would soon be solved. However, with each answer produced, more questions arose. More tools were employed, including mitochondrial genetics in the last 20 years, to help answer these questions. The use of mitochondrial DNA (mtDNA) to infer phylogenetic evolution is favoured over several other molecular entities for many reasons, thus affording it increasing amounts of popularity in the investigation of human evolution. The utilisation of Y-chromosome and mitochondrial data often complement each other. However, differences in evolutionary histories between men and women result in differences between mitochondrial and Y-chromosomal information.

Evolutionary studies utilising autosomal regions may seem the most appropriate option due to the large size of the genome. However, the presence of recombination between chromosomal regions complicates genetic evolutionary histories. The autosomal genome can be regarded as a series of large blocks of low recombination, interrupted by small blocks of high recombination (recombination hotspots). Furthermore, these block boundaries are not always well defined, as recombination may occur within the large blocks and block characteristics may vary between populations (Stumpf, 2002; Tishkoff and Verrelli, 2003). The Y-chromosome may seem a more likely candidate as it undergoes no recombination and is inherited as a single unit (Jobling and Tyler-Smith, 2000). As the Y-chromosome is absent in females and only present in males as a single copy, it is considered to be in a haploid state with a coalescence time equal to one quarter of that of autosomes (Jobling and Tyler-Smith, 2000). These two aforementioned facts are advantageous, as mutations are not shuffled between maternal and paternal lineages and represent unique events in evolutionary history. However, the Y-chromosome is sensitive to introgression as admixture is often sex-biased (Jobling and King, 2004) and is particularly sensitive to drift due to its haploid nature (Rosser *et al.*, 2000).

Deducing human evolutionary histories utilising mitochondrial genetics has been most popular, as discussed in Paragraph 2.8. However, the genetic differentiation between mitochondria of different populations is decreased due to female-mediated recruitment during migration and the fact that women cross the cultural boundaries more often than men (Seielstad *et al.*, 1998).

2.1 HUMAN ORIGINS AND EARLY MIGRATIONS

Models have been proposed to explain the replacement of archaic humans by anatomically modern *Homo sapiens sapiens*. Africa has been proposed as the geographic location of emergence of anatomically modern humans (Cann *et al.*, 1987; Vigilant *et al.*, 1991). This forms the basis of the strongly supported replacement model in which humans evolved relatively recently from a small African population circa (ca.) 143,000 ± 18,000 years before present (YBP) and migrated out of Africa to colonise the entire globe, replacing archaic humans ca. 100,000 YBP (Horai *et al.*, 1995). The *Strong Garden of Eden* (Harpending *et al.*, 1993), the *African Eve* (Cann *et al.*, 1987) and the *Out of Africa* (Giles & Ambrose, 1986) models all describe this replacement model, which is also supported by fossil records (Horai *et al.*, 1995). The earliest fossils of anatomically modern humans found in Africa can be dated to between ca. 100,000 and 250,000 YBP. The earliest modern human fossil identified outside Africa was detected in the Levant (Israel, Syria and Lebanon) and dated back to ca. 100,000 YBP (Stringer & Andrews, 1988).

An alternative model for human origins was proposed by Wolpoff (1972) in which *Homo erectus* migrated out of Africa one million YBP and colonised regions of the New World. In this multiregional model, populations of *Homo erectus* in different regions of the New World evolved independently into modern humans. The occurrence of gene flow between continental populations would have prevented differential speciation so that the modern human precursor evolved concurrently. Variants of this model exist, such as the Hybridisation and Replacement hypothesis, which claims that Africans in Europe and Western Asia were assimilated through hybridisation by a large African contribution (Ambrose, 1998).

2.1.1 The emergence of man

The exact emergence of anatomically modern humans from their ancestor is not well defined. *Homo erectus* and its descendant *Homo heidelbergensis,* as well as *Homo neanderthalensis,* possibly had a direct role in modern human origins. Genetic evidence for the recent African origin model was provided by Cann *et al.* (1987). This is greatly supported by mtDNA evidence, which indicates that African populations possess greater mtDNA diversity than European and Asian populations (Cann *et al.*, 1987) and that all non-African populations can be traced back to a single African "mitochondrial Eve" that represents a small, homogeneous mitochondrial population which lived between ca. 140,000 and 200,000 YBP (Cann *et al.*, 1987, Vigilant *et al.*, 1991). Through sequence determination of ancient DNA, Krings *et al.* (1997) determined that *Homo neanderthalensis* had made no genetic contribution to modern humans. The lack of modern human genetic components in the Neanderthal mtDNA sequence may however be attributed to genetic drift. By calculating evolutionary rates in Caucasian haplogroups relative to known rates in other populations, Torroni *et al.* (1994a) were able to demonstrate that Caucasian-specific haplogroups could not have been derived from Neanderthal populations. This view is challenged by the finding of a human burial site dated to ca. 24,500 YBP in Portugal, which possibly represents millennia of hybridisation between resident Neanderthal populations and invading *Homo sapiens* (Duarte *et al.*, 1999). However, none of the samples considered as anatomically transitional between modern humans and Neanderthals have yet presented any evidence of mtDNA admixture between the two groups (Serre *et al.*, 2004).

A recent, more comprehensive study on ancient mtDNA involving 24 Neanderthal individuals could present no proof of admixture (Serre *et al.*, 2004). By including early modern human remains of individuals who lived closer in time to Neanderthals than contemporary humans into their study to minimise the effects of drift, a genetic contribution larger than 25 percent (%) could be statistically rejected. For a 10% Neanderthal mtDNA contribution to be excluded, it was concluded that an analysis of an additional 50 early human remains would have to be performed. It can thus be concluded that a Neanderthal mtDNA contribution to the modern human mtDNA pool cannot yet be excluded and the extinction of Neanderthal mitochondrial lineages may account for the absence thereof in contemporary humans (Serre *et al.*, 2004).

Ancient Aboriginal Australian remains of an anatomically modern human from the terminal Pleistocene/early Holocene periods may represent the oldest human mitochondrial ancestor (Adcock *et al.*, 2001). The findings represent fossils of anatomically modern humans of which one can be dated to ca. 60,000 YBP, making it older than most Neanderthal samples. This mtDNA is absent in living Australians and therefore represents an extinct mtDNA lineage. When comparing this ancient DNA to modern-day DNA, the finding demonstrates that one of the deepest mtDNA lineages are Australian. This does not imply that the geographic emergence of modern humans occurred in Australia, anymore than an African origin implies an African geographic origin. However, it does pose a challenge to the recent out of Africa model by implying that replacement of modern humans occurred in Australia (Adcock *et al.*, 2001). This also demonstrates that if an mtDNA lineage belonging to modern humans can become extinct, then the absence of Neanderthal mtDNA in modern humans cannot rule out the possibility of a genetic contribution.

2.2 THE MITOCHONDRION

Mitochondria are organellular entities in eukaryotic cells that form the location of cellular energy production. Mitochondria contain their own genomes consisting of mtDNA. Each mitochondrion contains ca. ten mtDNA genomes, which evolve faster than nuclear DNA (nDNA), are maternally inherited and do not undergo recombination (Brown *et al.*, 1982b; Wallace, 1994), as discussed in Paragraph 2.3. Mitochondria are abundant in cells, relatively small and easily isolated and have been the focus of many genome sequencing projects (Borst, 1977). The investigation of mtDNA is therefore ideal for the inference of genetic phylogenies of species (Wallace, 1994).

2.2.1 Inheritance

The major portion of a human's mitochondria is included in the egg cell, thus ensuring that mitochondria are inherited in a strictly maternal fashion. Giles et al. (1980) first demonstrated this but it would only be true if paternal mitochondria were present at less than 4%. Paternal mitochondrial contribution to offspring is considered negligible, as the mammalian egg cell contains ca. 1,000 times more mitochondria than a sperm cell, and most of the male mitochondrial contribution is destroyed upon reaching the oocyte cytoplasm (Kaneda et al., 1995).

2.2.2 Mitochondrial structure and function

The mitochondrion is usually an ellipsoid-shaped organelle, depending on the physiological state of the cell, approximately 0.5 micrometres (μm) in diameter and 1.0 μm in length. The organelle is made up of an outer smooth membrane and an inner invaginated membrane (Borst, 1977). The invaginations form structures known as cristae, which increase the membrane surface area, and surround the mitochondrial matrix. The mitochondrial shape has evolved in order to translocate protons across its semi-permeable membrane for the coupling of respiration to adenosine triphosphate (ATP) synthesis (Mitchell, 1961). The mitochondrion forms the primary location for metabolic functions such as the citric acid cycle, fatty acid oxidation, electron transport and OXPHOS. The mitochondrial matrix contains high concentrations of soluble enzymes that take part in oxidative metabolism, DNA replication, translation and interact with substrates, nucleotide cofactors and inorganic ions. The outer membrane contains porins, which allow non-specific diffusion of large molecules. The inner membrane contains considerably more proteins, which allow for increased permeability of oxygen (O_2), carbon dioxide (CO_2), water (H_2O), respiratory chain proteins and transport proteins, which control the movement of ATP, adenosine diphosphate (ADP), pyruvate, calcium cation (Ca^{2+}) and phosphate. It is this selective permeability of the inner membrane to most ions that allows for the generation of concentration gradients, which drive the production of ATP (Voet and Voet, 1995).

The electron transport system is localised in the mitochondrial matrix and inner membrane and couples the free energy of the electron transfer from nicotinamide adenine dinucleotide (NADH) and flavin adenine dinucleotide ($FADH_2$) to O_2, thus resulting in ATP synthesis. This process occurs via protein-bound redox centres localised in the inner mitochondrial membrane. Respiratory enzyme complexes located in the inner membrane "pump" protons from the matrix across the inner mitochondrial membrane whereafter the controlled re-entry of protons into the matrix drives ATP synthesis from ADP. The proteins in the inner membrane are grouped into four respiratory complexes. Complex I (NADH-coenzyme Q [CoQ] reductase) contains one molecule of flavin mononucleotide and several iron-sulphur (Fe-S) clusters and passes electrons from NADH to CoQ. Complex II (succinate-CoQ reductase) contains the citric acid cycle enzyme succinate dehydrogenase and three other small hydrophobic subunits and passes electrons from succinate to CoQ. Complex III (CoQ-cytochrome *c* [cyt *c*] reductase) consists of two *b*-cytochromes, one cytochrome c_1 and one Fe-S cluster and passes electrons from

reduced CoQ to cyt *c*. Complex IV (cyt *c* oxidase) catalyses the oxidation of reduced cyt *c* and the reduction of O_2. A fifth complex, complex v, the proton-translocating ATP synthase, is responsible for the creation of a proton gradient across the inner mitochondrial membrane and the production of ATP (Voet and Voet, 1995).

2.3 MITOCHONDRIAL GENETICS

A mammalian cell contains ca. 1,000 - 10,000 mitochondria (Clayton, 1982) and each mitochondrion contains ca. 10 mitochondrial genomes (Bogenhagen and Clayton, 1974). The mitochondrial genome is a closed circular, double-stranded DNA molecule of 16,569 bp. The genome contains the genes for the 12 Svedberg units (S) rRNA (12S RNA) and 16S rRNA (16S RNA) as well as 22 tRNA molecules that are essential for mitochondrial protein synthesis and 13 polypeptides that are integral to the enzymes of the OXPHOS pathway (Anderson *et al.*, 1981). The complete mitochondrial genome was sequenced by Anderson *et al.* in 1981. The two strands of the mitochondrial genome can be distinguished based on their guanine (G) + thymine (T) base composition. This results in differential buoyant densities in denaturing caesium chloride gradients and allows for the naming of heavy (H) and light (L) strands. The H-strand encodes two rRNA genes, 14 tRNAs and 12 polypeptides, which is more than the L-chain, which encodes only eight tRNAs and a single polypeptide. All 13 encoded polypeptides are components of the respiratory chain/OXPHOS system. These genes encode seven polypeptides of complex i, one polypeptide, namely cytochrome *b* (cyt *b*) of complex iii, three polypeptides (cyt *c* oxidase subunit I [COI], cyt *c* oxidase subunit II [COII], and cyt *c* oxidase subunit III [COIII]) of complex iv and two polypeptides, namely ATP synthase subunit 6 (ATPase6) and ATP synthase 8 (ATPase8) of complex v (Wallace, 1992; Taanman, 1999).

The mitochondrial genome displays unique features relative to the nuclear genome. The mitochondrial genome utilises a modified standard genetic code, which differs from the universal code (Barrel *et al.*, 1980). The codon UGA in the mitochondrial genetic code codes for tryptophan, whereas UGA in the standard genetic code represents a termination codon. AUA codes for methionine instead of isoleucine in the mitochondria, and AGA and AGG are termination codons instead of codons for arginine. The mitochondrion also has a simplified decoding system, which allows the translation of all codons with less than the 32 different tRNA molecules utilised in nDNA. The reduction in tRNAs is due to the use of a uracil (U) base in the first anticodon (wobble) of a single tRNA, which recognises all codons of a four-codon family (Barrel *et al.*, 1980). Mammalian mtDNA shows extreme

economy regarding gene organisation. The mitochondrial genome lacks introns (except for a single regulatory region), intergenetic regions are either absent or limited to a few bases, the rRNA and tRNA molecules are small (Wolstenholme, 1992) and some protein genes are overlapping (Montoya *et al.*, 1983).

2.3.1 Mitochondrial replication and protein expression

The mitochondrial genome self-replicates and is utilised as a template for DNA expression. Mitochondria are dependent on nuclear-encoded products for their maintenance and propagation. Mitochondrial replication of mammalian mtDNA initiates unidirectionally from two separate strand-specific origins. Synthesis of the leading H-strand initiates at a non-coding region, termed the origin of H-strand synthesis (O_H), and proceeds two-thirds around the molecule displacing the original H-strand. H-strand replication exposes the replication initiation site of the lagging strand (O_L), allowing DNA synthesis of the L-strand to initiate, resulting in the formation of a displacement loop (D-loop). The newly synthesised H-strand of ca. 680 bp is flanked by tRNA phenylalanine ($tRNA^{Phe}$) and tRNA proline ($tRNA^{Pro}$) genes, and is known as 7S DNA. This is a short triplex region thought to represent aborted replication intermediates (Anderson *et al.*, 1981). A second, more conventional model of replication exists in which synthesis of the leading and lagging strands initiate simultaneously. Replication initiates at a single site, O_H, and proceeds unidirectionally with the formation of short Okazaki fragments on the lagging strand (Spelbrink, 2003).

Two major transcription initiation sites exist (IT_{H1} and IT_L) within 150 bp of each other in the major non-coding region. There are two independent promoters for H and L-strand transcription with a second initiation site for H-strand transcription (IT_{H2}) which is utilised less frequently than IT_{H1}. Once L-strand transcription is initiated, a single polycistronic precursor RNA encompassing almost all genetic information contained on the strand is synthesised. H-strand transcription is complicated by the presence of two promoters. Transcription occurs frequently at IT_{H1} and terminates at the end of the 16S RNA gene, resulting in elevated synthesis of the two rRNA genes. Transcription initiates less frequently at IT_{H2}, which produces a polycistronic molecule that is equivalent to almost the complete H-strand. Processing of these primary transcripts is relatively easy, given the lack of intervening sequences. Genes for tRNAs flanking the two rRNA genes and nearly every protein gene are proposed to form punctuation marks in reading mtDNA information through the secondary structure that tRNAs adopt (Taanman, 1999).

Translation occurs in the mitochondrial matrix through mitochondrial ribosomes that have a lower RNA and a higher protein content than their cytosolic counterparts. Ribosome binding to mammalian mitochondrial messenger RNA (mRNA) is not facilitated via upstream sequence, as it contains no upstream leader sequences. Translation is initiated at the 5'-end with the codon for initiation being *N*-formylmethionine. Ribosomes are not directed to the initiation sites via a recognition and scanning approach, as a 7-methylguanylate cap structure is absent from the 5'-termini of mitochondrial mRNA. The resulting decrease in efficiency of translation may explain the relative abundance of mRNA species in the mitochondrion (Taanman, 1999).

2.4 CALIBRATION OF MOLECULAR PHYLOGENIES

Hominid phylogenies based solely on morphological characteristics are not reliable, therefore the employment of molecular data is vital (Collard and Wood, 2000). Degrees of difference between clades can only be inferred if compared relative to a related clade that is an external point of reference. This clade is referred to as an outgroup and should be the oldest point in the tree. The outgroup determines the order of branching, which indicates the evolutionary relationships between groups as well as the branch length, which represents the proportional evolutionary difference. Although the outgroup is not a natural member of the group, a too distantly related outgroup may result in an outcast group (Baldauf, 2003). African apes are thus too distantly related to humans to allow for calibration of the mitochondrial D-loop clock (Horai *et al.*, 1995).

2.5 CAMBRIDGE REFERENCE SEQUENCE

A reference sequence of the entire mitochondrial genome was constructed when the genome was fully sequenced in 1981 by Anderson *et al.* (1981). This sequence was utilised to compare generated sequence data to a proposed correct reference sequence and was termed the Cambridge Reference Sequence (CRS). The sequence was generated from a single H haplogroup European individual and sequences from HeLa (Henrietta Lacks) cells and bovine mtDNA (Anderson *et al.*, 1981). Andrews *et al.* (1999) revised the CRS due to the finding of an error frequency of 0.07% in the original sequence when compared to other mitochondrial sequences. Once this revised CRS (RCRS) was determined, it was possible to distinguish between a polymorphism and a functional alteration for a given mtDNA type. By comparing entire mitochondrial genome sequences of African and European individuals to the RCRS, Van Brummelen (2003) identified

exceedingly greater differences between African mtDNA and the RCRS than European mtDNA and the RCRS. It was therefore extremely difficult to differentiate between a polymorphism and a functional alteration in African individuals. Although the generation of a reference sequence is not an objective of this study, a future aim may be to generate haplogroup-specific reference sequences in macro-haplogroup L.

2.6 ORIGIN OF THE MITOCHONDRION

General consensus on the origin of the mitochondrion is through endosymbiotic events and not autogenously through cellular differentiation during the course of evolution. Wallin (1922) first proposed mitochondria to be derived from cyanobacterial endosymbionts. This idea was later revived by Margulis (1970) who stated that eukaryotic organelles had been acquired via ingestion of prokaryotic cells. Evidence of this origin is based on several observations, namely that the mitochondrion is of a comparable size to many prokaryotes, its genome is circular and of a similar size and complexity as observed in prokaryotes, its pattern of antibiotic resistance is similar to that of prokaryotes and it has a double phospholipid bilayer. Based on ribosomal DNA (rDNA) sequences, plastids and mitochondria have distinctly different phylogenetic ancestries and could not have originated autogenously within the same host cell. A sequential origin, in which mitochondria and plastids (in that order) were acquired, was proposed by Margulis (1970).

2.7 MITOCHONDRIAL GENETIC VARIATION

Mitochondrial genomes are replicated and transmitted in their respective mitochondria to the offspring. The mitochondrial sequence is conserved from parent to offspring, but differences can occur through chance, population history or selection (Elson *et al.*, 2004). The variation generated is utilised to reconstruct the evolutionary history of populations and determine the implications of sequence change (Brown *et al.*, 1980).

2.7.1 Genome evolution

The evolutionary rate of the animal mitochondrial genome is greater than that of the nuclear genome. Using restriction map analysis to generate the percentage sequence differences between Guinea baboon, rhesus macaque, guenon and other higher order primates, Brown *et al.* (1979) allowed for the elucidation that mtDNA evolves five to ten times faster than single-copy nDNA. The mutation rate of mtDNA was calculated as 0.02 substitutions per bp per million years. Possible factors responsible for the higher

mutation rate in mtDNA compared to nDNA were greater exposure to oxidative damage, a more error-prone system of replication, less efficient editing or repair functions and a higher rate of turnover (Brown *et al.*, 1982b). The same authors noted that 92% of point-mutations in human, chimpanzee, gorilla, orang-utan and gibbon mtDNA were transitions. This was observed in both tRNA and protein-coding genes and was attributed to a bias in the mutation process rather than to selection on the mutation. It was also observed that the percentage of transitions decreases when more distantly related organisms are compared. This time-linked decrease may be due to multiple substitutions at the same nucleotide site and demonstrates the importance of comparing closely related species to infer evolutionary relationships.

2.7.2 Implications of variation

According to the infinite allele model, which applies to selectively neutral loci, the chance of survival of different mtDNA types is equal. The observed frequency distribution of mtDNA types is thus due to drift and frequency equilibrium is maintained through neutrality. The mitochondrial molecule was considered to be neutral (Moritz *et al.*, 1987), until Johnson *et al.* (1983) detected deviations from neutrality in mtDNA type frequency distributions. Excoffier (1990) studied mtDNA frequency distributions in 31 human populations and identified several Oriental and Caucasian populations as excessively homogeneous. All African samples investigated were more diverse and conformed to the neutral model of populations at equilibrium.

2.7.2.1 Adaptive variation

Mutations in the mitochondrial genome may bring about a selective advantage over the parental type. Torroni *et al.* (1994c) studied a possible link between mtDNA variation and adaptation to high altitudes in Tibetans. Haplogroups of high-altitude Tibetans were similar to those of low-altitude Tibetans as well as other Asians. This suggested that no major selective pressure had acted on high-altitude haplogroups. Based on the fact that only mtDNA haplogroups M and N left Africa to enrich Eurasia and that there is a five-fold enrichment of haplogroups A, C, D and G between central Asia and Siberia, Mishmar *et al.* (2003) suggested these enrichments to be due to a selective advantage on mitochondrial haplogroups as migrations took place into colder habitats. Analysis of amino acid substitution mutations versus neutral mutations in mtDNA protein coding genes of 104 complete mtDNA genomes from a vast global region revealed that the ATPase6 gene held

the greatest variation. The most variant ATPase6 gene was observed in populations from the arctic zones. Variations of the ATPase6 gene would reduce the coupling efficiency in ATP synthesis in the mitochondrial OXPHOS, thereby increasing the basal metabolic rate. The individual would require a higher caloric intake, which could be provided by a high-fat diet to produce excess heat energy in order to cope with a colder habitat. Mitochondrial mutations that result in differences in energy metabolism also result in altered mitochondrial oxidative damage, which affects human health and longevity (Coskun *et al.*, 2003; Ruiz-Pesini *et al.*, 2004).

2.7.2.2 Mutations and human disorders

Mutations in the mitochondrial genome may become fixed in a population through being selectively advantageous or through genetic drift if they are neutral. However, mutations may also be deleterious and result in genetic disorders. Mildly deleterious mutations that affect the phenotype and therefore fitness at a late onset will become fixed in the population as a polymorphism. However, deleterious mutations having an early onset will be rapidly removed by selection (Wallace, 1995). Therefore, early onset deleterious mutations are transient in nature, indicating that modern-day disorder-causing mutations have a recent origin (Wallace, 1994).

Pathologically deleterious mtDNA mutations are either missense mutations that alter polypeptide encoding genes or protein synthesis mutations that alter rRNA or tRNA genes (Wallace, 1992). LHON and neurogenic muscle weakness ataxia and retinitis pigmentosa (NARP), together with Leigh syndrome, represent the most investigated clinical phenotypes that are due to missense mutations. LHON is a form of acute or subacute blindness which leads to central scotoma. Nineteen point mutations have been identified as associated with the disorder, of which five are primarily causative (Wallace, 1995). All five mutations lead to the same phenotype but with different levels of severity, which are due to differences in heteroplasmic status and the ability to cause additional neurological symptoms. Two mutations observed in the ATPase6 gene are causative of NARP and highly pathogenic, which suggests that the mutations arose independently. The heteroplasmic nature of these two mutations result in neurological symptoms ranging from retinitis pigmentosa, mental retardation, olivopontocerebellar atrophy to Leigh syndrome, which represents the most severe phenotypic presentation of NARP syndrome (Wallace, 1992).

Nearly 30 mitochondrial tRNA and rRNA gene mutations have been associated with disorders, which range in severity from mild to lethal. Protein synthesis alterations that are moderate to severe can result in severe clinical symptoms in children or young adults, tend to reduce reproductive success and are usually heteroplasmic. Alterations of protein synthesis that are mild in nature may only result in the presentation of clinical symptoms once a reproduction age has been reached. A population can maintain these mutations at a low frequency as they are generally homoplasmic. Rearrangements in mtDNA can also be causative of disorders. More than 100 mtDNA rearrangements have been observed to be associated with degenerative disorders. These mutations are associated with three main clinical phenotypes, namely ocular myopathies such as Kearns-Sayre syndrome and chronic progressive external ophthalmoplegia, Pearson marrow/pancreas syndrome, and adult-onset diabetes mellitus with deafness (Wallace, 1995).

Certain clusters of mtDNA disease variants may be associated with specific haplogroups, thereby increasing the likelihood of disease. This is supported by several findings, such as an increased penetrance of LHON mutations in haplogroup J (Hofmann *et al.*, 1997), variants in haplogroup H that are associated with late-onset Alzheimer's disease (Shoffner *et al.*, 1993), an increased probability of becoming blind if an individual belonging to haplogroup J has Leber's hereditary optic neuropathy (Brown *et al.*, 2001) and positive and negative associations of haplogroups T and H with asthenozoospermia (Ruiz-Pesini *et al.*, 2000). Nucleotide variants identified as non-disease causing, may contribute to the disease by acting synergistically towards the expression thereof (Hofmann *et al.*, 1997). Differences in disease expression and prevalence exist, as well as variability in drug response, between certain genetically identified clusters (Tate and Goldstein, 2004), and are discussed further in Paragraph 2.10.

2.8 MITOCHONDRIAL PHYLOGENIES

The combination of parameters such as high mutation rate (Wilson *et al.*, 1987), strict maternal inheritance (Giles *et al.*, 1980) and no recombination has made the mitochondrial genome ideal for deducing evolutionary phylogenies. Regions of the Y-chromosome are subjected to similar parameters, however, phylogenies that complement mitochondrial histories have not always been matched.

The distribution and frequency of shared alterations between individuals is an indication of the degree of genetic relatedness between individuals (Cann *et al.*, 1987). The variation

observed within human populations is comparable to that observed between different populations (Cann *et al.*, 1987). Therefore, individuals of different populations share more alterations than individuals within a population (Cann *et al.*, 1987). Populations that have undergone the same history share unique alterations that other populations do not possess. It is these unique alterations that are utilised to identify populations (Brown, 1980).

2.8.1 Methods of assessment of variation

Variation between mitochondrial genomes is expressed as sequence variation and can be detected utilising several methods. These include restriction fragment length polymorphisms (RFLP) analysis, partial genome sequencing, full genome sequencing, denaturing high-performance liquid chromatography, and a combination of RFLP and sequence data (Graven, 1995). RFLP analysis has thus far received most attention because of its simplicity and robust results, but full genome sequencing has become popular of late because it provides greater resolution and specificity.

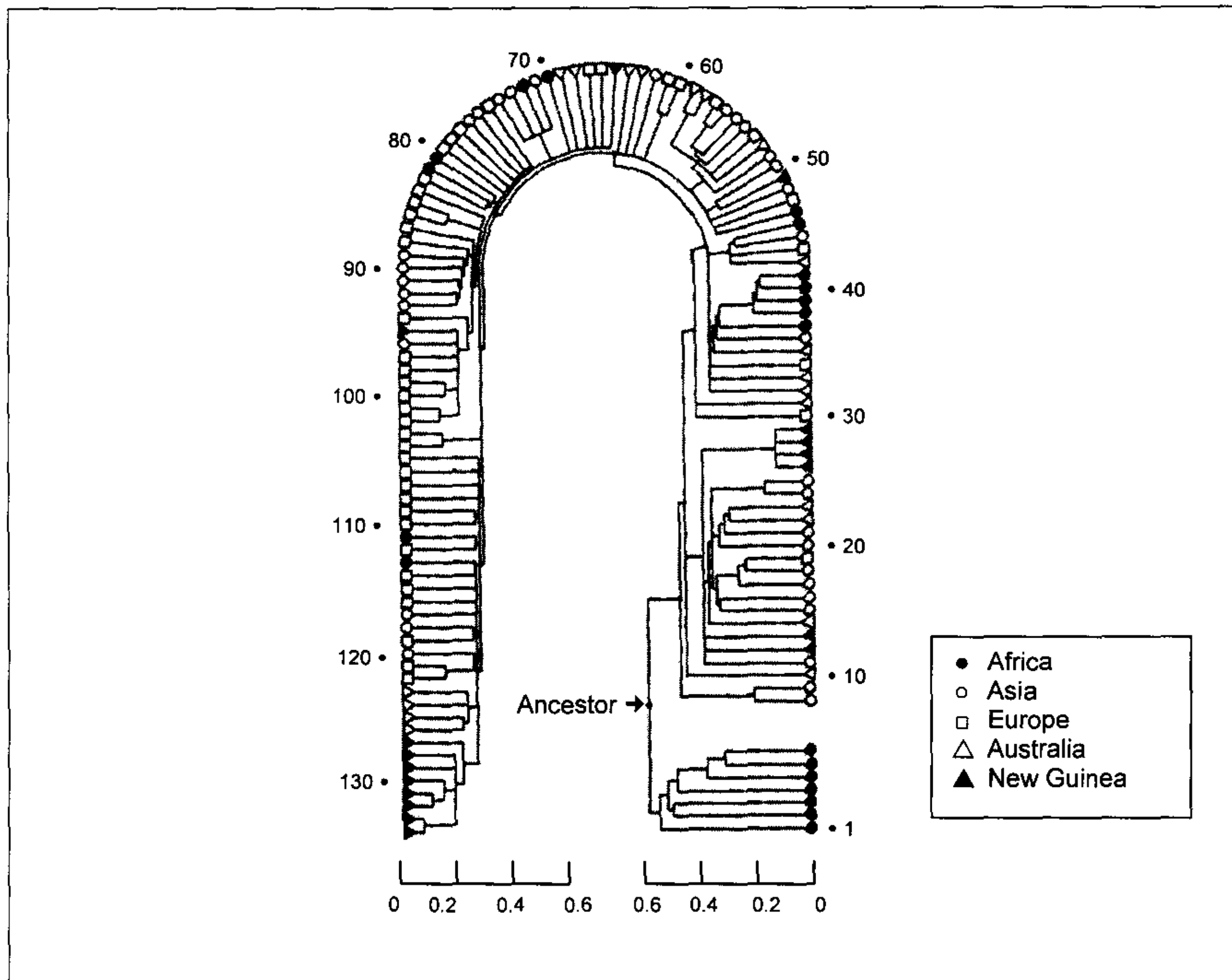
2.8.1.1 Restriction fragment length polymorphism analysis

RFLP analysis has revealed several mitochondrial haplogroups present in humans based on differential restriction patterns of restriction enzymes (RE). A haplogroup is assigned to each combination of restriction patterns produced. A change in restriction sites will be caused by a nucleotide change in the RE recognition sequence. The mitochondria of humans of mixed origin such as Caucasians from the United States of America (USA), the Philippines and Egypt, Mongolians from China, and African Americans from the USA were first analysed utilising 18 RE (Brown, 1980). No differences between individuals were detected for seven of the enzymes, but the remaining enzymes showed one or more differences. These differences were due to the presence of mitochondrial alterations, of which many were shared between two or more samples. It was established that the patterns of enzyme restriction were group-specific. Utilising the RE *Hpa* I, Denaro *et al.* (1981) discovered six different cleavage patterns in varying frequencies between Caucasians from the USA and Europe, Orientals from Taiwan, China and Japan, as well as Pygmies, Khoi-San and Bantu-speaking Africans from Africa. Some of the *Hpa* I sites were absent in the populations, and together with the frequency of the different morphs, it was possible to differentiate clearly between populations. Utilising four additional enzymes, *Bam* HI, *Hae* II, *Msp* I and *Ava* II, Johnson *et al.* (1983) detected 35 distinct mtDNA types

in Caucasians, Orientals, Khoi-San and Bantu-speaking Africans. A high correlation was identified between mtDNA type and the ethnic origin of an individual.
Most African populations form a distinct lineage. *Hpa* I RFLP analysis of individuals from different geographic origins revealed one morph to be present only in the African population in more than 90% of the individuals tested. A phylogeny constructed by Johnson *et al.* (1983) utilising RE cleavage patterns from five RE, namely *Hpa* I, *Bam* HI, *Hae* II, *Msp* I and *Ava* I, was centred on one specific mtDNA type. This mtDNA type had most branches radiating from it and was the most common type in all African samples. One of the branches formed a distinct African lineage, which had the highest frequency of this central mtDNA type. Africans were observed to be genetically the most diverse, with the variation in the African population being as great as that between Africans and any other ethnic group (Cann *et al.*, 1987).

Africa is considered the continent that gave rise to the present human mitochondrial gene pool. Cann *et al.* (1987), utilising high resolution RFLP analysis (HR-RFLP), constructed a global mitochondrial phylogeny (Figure 2.1). All minimum length trees constructed have two consistent features: (1) there are two primary clades, one consisting exclusively of Africans and the other of Africans and all other ethnic groups studied, and (2) the populations are not monophyletic but rather dispersed throughout the tree.

Figure 2.1: Global mitochondrial phylogeny for human mtDNA



The root of the most recent common ancestor (MRCA) is indicated by the arrow. Numbers indicate the mtDNA types, number one being !Kung cell line (GM3043), number 45 HeLA cells and number 110 from the CRS. Percentage sequence divergence is indicated on the scales at the bottom of the figure. Adapted from Cann *et al.* (1987).

2.8.1.2 Sequence analysis

Nucleotide sequences provide the greatest resolution possible when studying molecular evolution of populations. Various regions of the mtDNA genome have been sequenced, the most popular of which is the 1,121 bp non-coding control region (CR) or "D-loop" which includes both hypervariable segments (HVS-I and HVS-II). The popularity of the CR can be ascribed to it having a three to four times greater sequence diversity than that of coding regions of the mitochondrial genome. Vigilant *et al.* (1989) analysed CR sequences from 83 individuals from Africa, Asia, Europe and America, determined the sequence diversity between the populations and were able to imply a relationship between geographic origin and sequence differentiation preliminarily. However, CR sequence data alone are insufficient to calculate nucleotide substitution rates and divergence times. Horai *et al.* (1995) sequenced seven full mtDNA genomes of three humans and four nonhuman

hominids. Utilising these data, the authors were able to estimate the non-synonymous substitution rate of 0.35 x 10^{-8} substitutions per site per year and estimated the age of the most recent common ancestor (MRCA) to be ca. 143,000 ± 18,000 YBP.

In some instances, the same CR sequence may be associated with several RFLP haplogroups and *vice versa*. Combined RFLP and sequence data are more powerful than when utilised separately (Chen *et al.*, 2000). mtDNA analysis can be combined with Y-chromosome analysis to infer differential population histories between males and females (Richards *et al.*, 2003).

2.8.2 Mitochondrial haplogroups

Any combination of polymorphic markers along a non-recombining molecule constitutes a haplogroup. As the mitochondrial genome does not undergo recombination, polymorphic markers can be utilised to infer haplogroups. The status of a combination of biallelic markers is utilised to assign a haplogroup to an individual (Jobling and Tyler-Smith, 2000).

2.8.2.1 Global haplogroups

African populations can be divided into two main haplogroups based on RFLP analysis. Macro-haplogroup L is characterised by an *Hpa* I site gain at nucleotide position (np) 3592 which is specific for African populations (Chen *et al.*, 1995). Macro-haplogroup L is further subdivided into haplogroup L1 characterised by a *Hinf* I site gain at np 10806 and L2 characterised by a *Hinf* I site gain at np 16389. The remaining African populations are characterised by lacking the *Hpa* I site and are termed haplogroup L3 (Watson *et al.*, 1997). This haplogroup occurs in the Senegalese population (Chen *et al.*, 1995), the Bamileke from the Cameroon (Scozzari *et al.*, 1994), Khoi-San populations from Namibia (Soodyall *et al.*, 1996) and several Bantu-speaking populations from Southern Africa (Johnson *et al.*, 1983).

Macro-haplogroup L, redefined haplogroup L* (Chen *et al.*, 2000), haplogroup L2 and haplogroup L3 can be further subdivided (Watson *et al.*, 1997, Chen *et al.*, 2000). The subdivision is based on the absence or presence of restriction sites as presented in Table 2.1.

Table 2.1: Mitochondrial polymorphisms defining macro-haplogroup L

Haplogroup	Defining polymorphisms
L	+ 3592*Hpa* I; + 10394*Dde* I
L1	+ 10806*Hinf* I; - 2758*Rsa* I
L1a	+ 4310*Alu* I
$L1a_1$	- 4853 *Alu* I
$L1a_2$	- 8112*Msp* I; - 8150*Msp* I; + 8249*Ava* II; - 8250*Hae* III
L1b	+ 7055*AluI*
$L1b_1$	+ 2349*Mbo* I
$L1b_2$	+ 9070*Taq* I
L2	+ 16389 *Hinf* I;- 16390*Ava* II
L2a	+ 13803*Hae* III
L2b	+ 4157*Alu* I
L2c	- 322*Hae* III; - 679*Dde* I, - 13957*Hae* III
L3	- 3592*Hpa* I
L3a	+ 2349*Mbo* I
L3b	- 8616*Mbo* I
L3c	+ 10084*Taq* I
L3d	- 10349*Dde* I

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

There are varying degrees of correlation between haplogroups L* and the ethnic status of African populations. The Mbuti Pygmies (L2a) and Biaka Pygmies ($L1b_2$) grouped separately with a 100% association between ethnic group and haplogroup. Southern African !Kung and Khwe samples grouped together in the L1a haplogroup. $L1a_2$ consisted predominantly of !Kung and L1a was composed of !Kung and Khwe. L2b also included !Kung and Khwe, and L3a consisted predominantly of Khwe. The majority of Vasikela !Kung, Khwe and Senegalese groups belong to haplogroup L* but also have large proportions of L3 haplogroups, with L3a to L3d being nonuniformly distributed among the populations (Chen *et al.*, 2000). Haplogroup L3, which lacks the *Hpa* I recognition site at np 3592, is not restricted to African populations. The four sub-haplogroups of L3 share some degree of similarity with haplogroups detected in Europe and Asia (Chen *et al.*, 2000).

The Caucasian population of Europe can be divided into two main lineages based on the presence or absence of a *Dde* I recognition site at np 10394. The European population was one of the last major populations to undergo high resolution (HR) RFLP analysis. Torroni *et al.* (1994a) discovered that 64% of European mtDNAs analysed fell into four

ancient Caucasian-specific mtDNA haplogroups. The study of three diverse European populations (Finland, Sweden and Tuscany) revealed that nearly all the mtDNAs grouped into ten mtDNA haplogroups, namely H, I, J, K, M (Asians), T, U, V, W and X as depicted in Table 2.2. This pattern of grouping suggested that all European mtDNAs form distinct haplogroups similar to those observed in Native Americans (Torroni *et al.*, 1993a; Torroni *et al.*, 1996). Continent-specific haplogroups are also detected in extremely low frequencies on different continents, which imply that there is limited overlap in variation between European and non-European mtDNAs. Haplogroups H, I, J, K, T and W are essentially specific to European populations and most likely originated after Caucasians left Africa and Asia. Haplogroup U, defined by the presence of a *Hinf* I recognition site at np 12308, was also identified in low frequencies in African populations.

Native American mtDNAs form four distinct haplogroups, namely haplogroups A, B, C and D. Haplogroup A is defined by the presence of a *Hae* III recognition site at np 663, haplogroup B is defined by a 9-bp deletion between np 8272 and np 8289 which is always associated with the *Hae* III recognition site gain, haplogroup C is defined by an *Alu* I recognition site gain at np 13263 and a *Hinc* II recognition site loss at np 13259 and haplogroup D is defined by an *Alu* I recognition site loss at np 5178 (Torroni *et al.*, 1993a).

Table 2.2: Mitochondrial polymorphisms defining European haplogroups

Haplogroup	Polymorphic sites
H	- 7025*Alu* I
I	- 1715*Dde* I; + 8249*Ava* II; + 10028*Alu* I
J	- 13704*Bst* OI
K	- 9052*Hae* II; + 13308 *Hinf* I
T	+ 13366*Bam* HI; + 15606*Alu* I
U	+ 12308*Hinf* I
V	- 4577*Nla* III
W	+ 8249*Ava* II; + 8994*Hae* III
X	- 1715*Dde* I
M (Asian)	+ 10397*Alu* I

Adapted from Torroni *et al.* (1996).

Asian mtDNAs can be divided into two major lineages (Torroni *et al.*, 1994c) based on the absence or presence of *Dde* I and *Alu* I recognition sites at np 10394 and np 10397 respectively. These two macro-haplogroups are further subdivided into haplogroups A, B, C, D, E, F and G as presented in Table 2.3. Haplogroups A, B and F are defined by the

absence of the *Dde* I and *Alu* I recognition sites at np 10394 and np 10397 respectively, and haplogroups C, D, E and G are defined by the presence of these two sites (Torroni *et al.*, 1993a).

Table 2.3: Mitochondrial polymorphisms defining Asian haplogroups

Haplogroup	Defining polymorphisms
A	+ 663*Hae* III
B	8271-8281delCCCTCTA
C	- 13259*Hinc* II; + 13262*Alu* I
D	- 5176*Alu* I
E	- 7598*Hha* I
F	- 12406*Hinc* II/*Hpa* I
G	+ 4830*Hae* III; + 4831*Hha* I

Adapted from Torroni *et al.* (1993a).

Populations containing Asian mtDNA haplogroups are presented in Table 2.4. Haplogroups A, B, C and D are the four founder lineages which colonised America (Torroni *et al.*, 1993a). Twenty percent of the Ethiopian population possesses the - 10394*Dde* I; - 10397*Alu* I macro-haplogroup. The presence of this haplogroup in the Ethiopian population, as well as the fact that it is not detected in any other population, gives rise to the possibility that the Asian and Ethiopian - 10394*Dde* I; - 10397*Alu* I haplogroup may have a common origin (Passarino *et al.*, 1998).

Table 2.4: Populations containing Asian mtDNA haplogroups

Haplogroup	Population (s)	Reference (s)
A	Taiwanese Han, Native Americans	Ballinger *et al.* (1992); Torroni *et al.* (1993a)
B	Asians, Americans	Ballinger *et al.* (1992)
C	Taiwanese Han, Tibetans, Aboriginal Siberians and Native Americans	Ballinger *et al.* (1992)
D	Asians, Siberians and Native Americans	Ballinger *et al.* (1992); Torroni *et al.* (1993a)
E	Koreans	Ballinger *et al.* (1992)
F	Asians	Ballinger *et al.* (1992)
G	Korean, Nivkhs	Ballinger *et al.* (1992); Torroni *et al.* (1993b)

Populations in specific haplogroups are indicated.

2.8.2.2 Haplogroup evolution

Identification of a founding population is based on the fulfilment of certain criteria. The founding haplogroup must be widespread throughout the population to which it gave rise, it

must have a central branching position in the haplogroup phylogeny and it should still be present in the proposed geographic location of origin (Torroni *et al.*, 1993a).

Shared haplogroups between tribes were determined in the four Native American haplogroups. Although only haplogroup A had a haplogroup that was central to its radiation, haplogroup A, B, C and D each had a shared haplogroup in Asia, which is thus hypothesised to have been the origin of American haplogroups. Based on the degree of sequence divergence between haplogroups, divergence times of haplogroups were calculated and are presented in Table 2.5.

Table 2.5: Divergence times of Native American haplogroups

Haplogroup	Divergence time range (YBP)
A	22,750 - 45,500
B	6,000 - 12,000
C	24,000 - 48,000
D	13,250 - 26,500

YBP = years before present. Adapted from Torroni *et al.* (1993a).

The difference in divergence times between haplogroup B (ca. 6,000 - 12,000 YBP), the average of haplogroups A, C and D (ca. 25,500 - 41,000 YBP) and the fact that all these haplogroups except B are present in Siberia give rise to the possibility that haplogroup B entered America through a different route and at a different time than haplogroups A, C and D. Torroni *et al.* (1994c) proposed a dual early and late entry into America in which haplogroup B arrived ca. 11,724 - 15,456 YBP and the average arrival of haplogroups A, C and D was ca. 25,707 - 33,939 YBP. Although these dates of entry are not statistically compatible with the accepted early and late entry dates, the fact that haplogroup B entered America separately relative to haplogroups A, C and D is most likely. The finding of haplogroup X, defined by a C to T transition at the 16223 and np 16278 positions in Native American populations, which is related to the haplogroup X detected in Europe, makes it a candidate for a possible additional founding haplogroup with an arrival time of ca. 12,000 - 36,000 YBP. Since haplogroup X had not been observed in Asians, it was hypothesised that the Native American haplogroup X founding population was of Caucasian descent (Brown *et al.*, 1998) until haplogroup X was observed in Siberia, which could have been the origin of haplogroup X in Native Americans (Derenko *et al.*, 2001).

Based on the fact that haplogroups H, I, J and K are specific to Caucasian populations, it was proposed that these haplogroups formed after the separation of Caucasians from

Africa and Asia more than ca. 80,000 YBP. It was observed that these four haplogroups are not uniformly distributed across Caucasian populations, as there is a degree of enrichment for haplogroup I in Europe. It is haplogroup I that contributed the most to the North American mitochondrial gene pool (Torroni *et al.*, 1994a).

Torroni *et al.* (2001) rejected the neutral model of a uniform molecular clock acting on regions outside the control region. Utilising full mitochondrial sequence data, it was observed that sub-haplogroup L2d was the most divergent clade, with L2a first branching from L2d and L2b and L2c being the most closely related. This may imply that the different clades evolve at different rates due to different selective forces acting on clades in different geographic regions.

2.8.2.3 Deducing human migrations

Early human expansions and migrations can be deduced from the present-day localisation of mtDNA haplogroups. The first split in the oldest MRCA mtDNA lineage L1a occurred in Africa, generating L1b/c and L2 ca. 122,000 - 132,000 YBP and ca. 85,000 - 95,000 YBP respectively. The second split also occurred in Africa ca. 59,000 - 69,000 YBP and gave rise to L3 and haplogroups taking part in the first expansion out of Africa (Maca-Meyer *et al.*, 2001). Two possible routes for the first migration out of Africa have been proposed. The central Asian route through North Africa into the Levant was based on fossil evidence (Stringer *et al.*, 1989). This migration gave rise to three main lineages ca. 43,000 - 53,000 YBP, which are present in Asia, Europe and America. One lineage consists of haplogroups X and A, the second lineage of W, I and N1b and the third lineage gave rise to at least four clusters. The migration route into South Asia is from Africa through Eastern Africa/Western India, as indicated in Figure 2.2. The migrating group would have been haplogroup M, defined by a C to T transition at np 10400, which is widespread in Asia and is observed in 20% of Eastern Africans (Passarino *et al.*, 1998). The coalescence time for haplogroup M in Eastern Africa was calculated to be ca. 48,000 $\pm$ 15,000 YBP, which is compatible with that of the Indian haplogroup M, which is ca. 56,000 $\pm$ 7,000 YBP. This was the first genetic evidence of an exit route from Africa through Eastern Africa/Western India (Quintana-Murci *et al.*, 1999). Two independent migrations into South Asia have been proposed of which one was haplogroup M, which radiated ca. 30,000 - 57,600 YBP (Maca-Meyer *et al.*, 2001).

Figure 2.2: Global migrations of human mtDNA

Global migration of the oldest mtDNA haplogroup (macro-haplogroup L) out of Africa and into Asia, and subsequently into Europe and the Americas. Letters correspond to haplogroups. Numbers indicate date of event in YBP. Adapted from MITOMAP (2004).

Although a major human population expansion occurred between ca. 49,000 and 640,000 YBP in Africa, before the split between African and non-African populations, no such expansion signal has been detected outside Africa. If it is assumed that the expansion predates the split between African and non-African populations, then it is surprising to find no signal in non-African populations. This could be explained by a bottleneck occurring in the non-African population (Reich and Goldstein, 1998). The occurrence of a population bottleneck also explains the low diversity in modern humans relative to other hominids in African and non-African populations (Watson *et al.*, 1997; Ruvolo *et al.*, 1994).

2.9 AFRICAN EVOLUTION

Denaro *et al.* (1981) were among the first researchers to perform RFLP analysis on human mtDNA to infer human origins. Unfortunately, incorrect conclusions were drawn from that study, as Denaro *et al.* (1981) proposed Orientals from Asia to be the origin of the formation of human ethnic groups. This was based on restriction patterns generated by *Hpa* I. One of the patterns producing two fragments, of 4.21 kb and 12.30 kb, was identical to that present in all apes and was observed only in Orientals and Bantus. The frequency of this pattern is highest in Orientals and was only detected in African Bantus but not African Pygmies or Khoi-San.

Cann *et al.* (1987) assumed a constant mitochondrial mutation rate of between two and four percent per million years and calculated the age of the MRCA of all surviving human

mtDNAs to be between ca. 140,000 and 290,000 YBP. Utilising a mutation rate of 2.2% - 2.9% per million years (Myr), the age of macro-haplogroup L was calculated to be ca. 98,000 - 130,000 YBP, haplogroup L1 ca. 86,000 - 113,000 YBP, and haplogroup L2 ca. 59,000 - 78,000 YBP (Chen *et al.*, 1995). A more comprehensive list of divergence times for African haplogroups was constructed by Chen *et al.* (2000) placing the age of the MRCA at ca. 125,500 - 165,500 YBP, macro-haplogroup L at ca. 122,800 - 161,800 YBP, L1 at ca. 113,100 - 149,100 YBP and L2 at ca. 59,000 - 77,700 YBP, as presented in Table 2.6. This makes macro-haplogroup L the oldest human mitochondrial lineage having a sequence diversity of 0.356% and it is therefore considered to represent the oldest node on the present human mtDNA phylogeny. The Vasikela !Kung and the Biaka Pygmies represent the oldest of the human populations with a Vasikela !Kung sublineage having the deepest root in the African lineage relative to the chimpanzee outgroup haplogroup. Furthermore, the !Kung displayed a high intragroup diversity of 0.320%, which is only slightly lower than that of the Biaka Pygmies at 0.342% (Chen *et al.*, 2000).

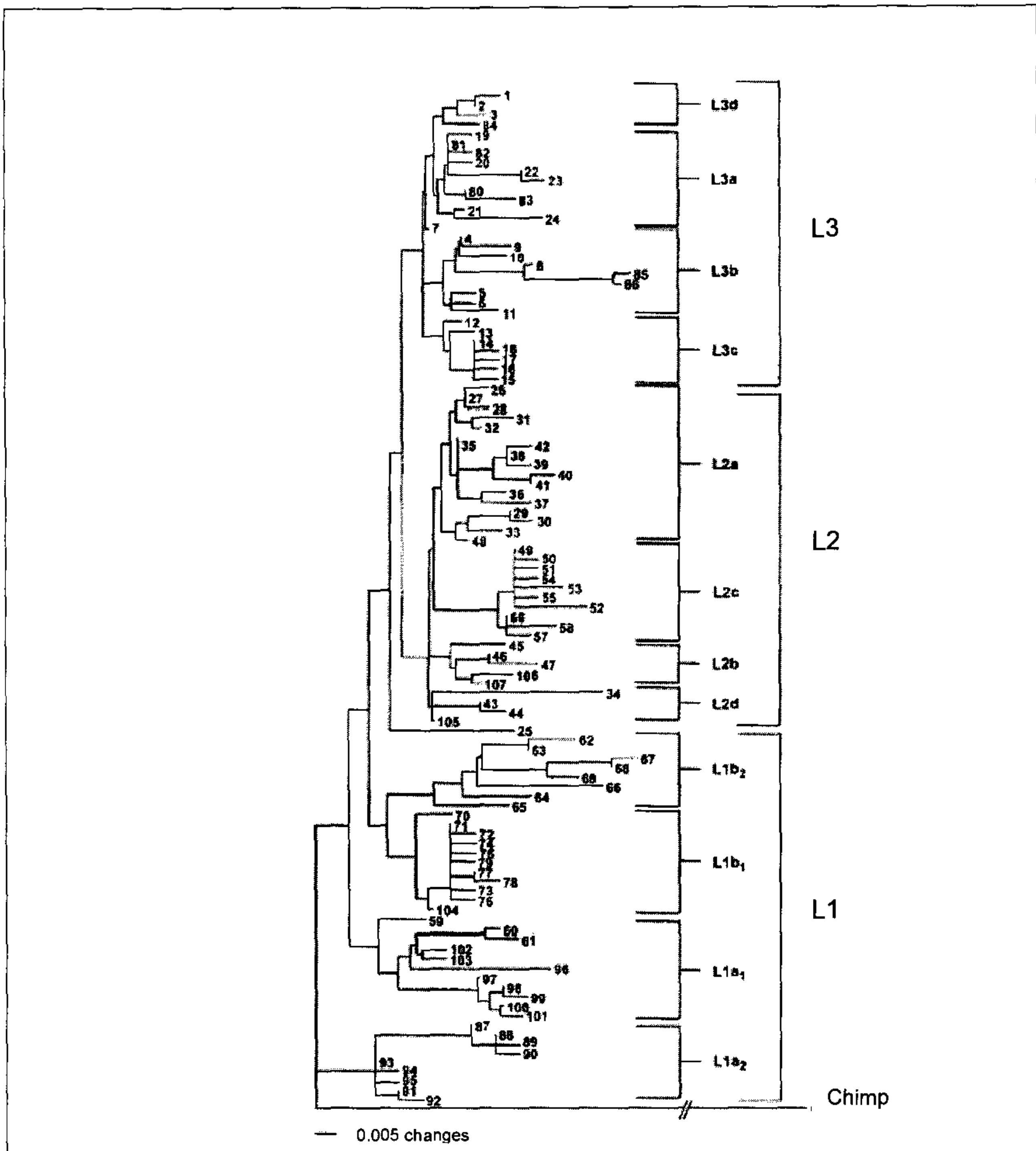
Table 2.6: Sequence divergence times of African macro-haplogroup L

Haplogroup	Divergence time (YBP)	Haplogroup	Divergence time (YBP)
All African	125,500 - 165,500	L2	59,000 - 77,700
L	122,800 - 161,800	L2a	39,000 - 51,400
L1	113,100 - 149,100	L2b	24,800 - 32,700
L1a	91,400 - 120,500	L2c	17,900 - 23,600
$L1a_1$	57,200 - 75,500	L3	78,300 - 103,200
$L1a_2$	41,000 - 18,600	L3a	41,400 - 54,500
L1b	73,800 - 97,300	L3b	77,600 - 102,300
$L1b_1$	14,100 - 18,600	L3c	28,300 - 37,300
$L1b_2$	84,800 - 111,800	L3d	17,600 - 23,200

YBP = years before present. Adapted from Chen *et al.* (2000).

Africa harbours the most complex genetic makeup of any continent by having the most diverse mtDNA lineages (Figure 2.3). Native Africans cluster within macro-haplogroup L, and can be further subdivided into haplogroups and sub-haplogroups (Chen *et al.*, 1995; Chen *et al.*, 2000; Salas *et al.*, 2002). Africa also harbours the most recent widespread demographic shift (Salas *et al.*, 2002). The first permanent settlements in Africa can be dated back to ca. 18,000 YBP in the Nile Valley, with the most recent being Bantu dispersals in West Africa ca. 3,000 - 4,000 YBP.

Figure 2.3: African haplogroups



A neighbour-joining (NJ) tree of Southern African haplogroups utilising RFLP data. Individuals belong exclusively to macro-haplogroup L, and can be divided into haplogroups and sub-haplogroups. Many of the sub-haplogroups are even further subdivided. Adapted from Chen *et al.* (2000).

The Bantu ethnic group had its origin in West-Central Africa ca. 5,000 YBP (Cavalli-Sforza *et al.*, 1994). The Bantu lineage migrated south in several expansions during the Bantu dispersals (Cavalli-Sforza *et al.*, 1994). The Xhosa, Zulu and Tswana ethnic groups, detected in Southern Africa, belonged to the Southeast Bantu wave (Cavalli-Sforza *et al.*, 1994), which reached Southeast Africa ca. 2,000 YBP (Salas *et al.*, 2002). Therefore, the Bantu ethnic groups shared a common ancestor ca. 5,000 YBP (Salas *et al.*, 2002).The

mtDNA of contemporary South-Eastern African populations is composed of two distinct components. These are ancestral types carried by resident populations, i.e. the Khoi-San, before the arrival of the Bantus, and the types introduced by the Bantus as a result of the Bantu dispersal (Salas *et al.*, 2002).

Linguistically related populations of Sub-Saharan Africa share a common pool of genes (Excoffier *et al.*, 1991). The population giving rise to the Bantu expansion could therefore have been heterogeneous regarding haplogroup composition. The Bantu ethnic groups thus separated from a heterogeneous population. It is plausible that each Bantu ethnic group could therefore be assumed to be composed of more than one haplogroup. A genetically heterogeneous population may consist of different haplogroups, which may each influence the expression of a specific disorder in a slightly different manner. The level of genetic differentiation between ethnic groups will affect the level at which sampling is executed during genetic studies.

African prehistory has been focused on linguistics because of a poor understanding of archaeology. However, controversy still remains when these tools are utilised and this has prompted the use of mtDNA studies. Although many mtDNA markers have been proposed to indicate signals of Bantu dispersals, most studies do not include any southern Bantu data (Salas *et al.*, 2002).

The South African population represents unique genetic aetiologies. Mutations identified as disorder-causing in non-South African populations are not associated with the same disorder symptoms in the South African population (Olckers *et al.*, 2001). The South African population also harbours unique polymorphisms not observed in other populations (Van Brummelen, 2003). The South African population is thus unique on a genetic level and is also considered culturally unique, in terms of the age of recent evolution and language (Salas *et al.*, 2002).

2.10 LINGUISTIC INFLUENCE ON DIVERSITY

Populations diverge genetically due to genetic drift, selection, bottlenecks, founder effects and size fluctuations, but these differences are reduced by the exchange of individuals between populations, namely migrants. The number of migrants is influenced by geographic distances as well as barriers, such as mountains, seas and deserts. In addition, more elusive obstacles for migrants are cultural barriers, such as language,

politics and religion. Therefore, the observed levels of genetic differentiation between populations may be attributed to geographic as well as cultural factors, of which language is the most significant as it has the greatest time depth (Rosser *et al.*, 2000). However, although most genetic barriers correlate with linguistic boundaries, the presence of some languages is not a result of persistence but rather recent acquisition through elite dominance. Elite dominance results via language acquisition from invading populations and is not always linked to a high degree of genetic admixture (Rosser *et al.*, 2000).

In a global Y-chromosome haplogroup study, Poloni *et al.* (1997) observed that linguistic differentiation is strongly linked to geographic structure, although it was observed that genetic differences are more important in distinguishing between language groups than between populations within a language group. This was confirmed by Rosser *et al.* (2000) who detected clines of Y-chromosomal variation in Europe that did not have a high degree of correlation with linguistics. This can be due to the fact that the study of Poloni *et al.* (1997) was global and showed greater genetic and linguistic differentiation, whereas individuals included in the study by Rosser *et al.* (2000) were from a single continent and from a single language group, the Indo-European language group.

Rosser *et al.* (2000) stated that the finding of a homogeneous landscape as revealed by European mtDNA, in contrast to Y-chromosomal studies, may be a result of the depth of analysis, for example, at which sublineages are studied. This is also a result of a high degree of Y-chromosome differentiation, as a result of drift and a greater migration rate of women compared to men. The greater migration rate of women is due to the phenomenon of patrilocality (Seielstad *et al.* 1998), whereby women are more likely to move from their birthplace after marriage than men because of custom, and the fact that it is easier for women to cross the cultural barrier than men.

A large degree of genetic structure is observed in Africa due to the large amount of genetic differentiation between the Niger-Congo, the Afro-Asiatic and the Khoi-san groups. However, the Niger-Congo group, to which the Bantu group belongs, showed the lowest level of internal genetic diversity (Poloni *et al.*, 1997), which may be the result of a recent Niger-Congo population dispersal (Salas *et al.*, 2002). Similar results for the Niger-Congo sample were obtained by mtDNA analysis, although low resolution RFLPs were utilised.

The categorisation of genetic groups in the study presented, utilising linguistic criteria, has several limitations. Firstly, genetic differences are inversely related to the amount of gene

flow, which depends on geographic distance and on the presence of barriers (Barbujani, 1991), such as language differences. Migrants between linguistic groups would increase the genetic relatedness between them. Secondly, female migration occurs through patrilocality, as described above, as women tend to cross cultural barriers more easily than men (Seielstad *et al.*, 1998). The woman's children would adopt the language of the recipient population, but would possess the mother's mitochondria. Obscuring of evolutionary histories through patrilocality may be resolved via Y-chromosomal instead of mtDNA analysis. Thirdly, languages evolve more rapidly than genes (Cavalli-Sforza *et al.*, 1988) and may thus reflect evolutionary units not definable through genetic components. Finally, languages can undergo rapid replacement from invading populations without a significant amount of genetic admixture (Rosser *et al.*, 2000).

Genetic clusters that are formed by common disease expression and prevalence, as well as variability in drug response, as mentioned in Paragraph 2.7.2.2, correspond to major continental populations, classified according to self-identified ethnicity. There also exists the aforementioned variability between sub-populations in major global populations (Risch *et al.*, 2002). These sub-populations are not always identifiable by self-reported ethnicity, but are effectively characterised utilising genetic inferences (Risch *et al.*, 2002).

The effectiveness of self-identified ethnicity, as an indicator of the genetic make-up of individuals, is reliable on a global scale, as discussed in the preceding paragraph, and varies in sub-populations of continent-specific populations, as discussed at the beginning of Paragraph 2.10. Health discrepancies between populations exist not only as a result of genetic differences, but are also attributable to the following factors: culture, diet, socioeconomic status, access to health care, education, environmental exposure, social marginalisation, discrimination and stress (Collins, 2004). Ethnicity may serve as a proxy to clarify genetic variation between sub-populations, to elucidate the effects of environmental factors and as a predictor of common disease expression, prevalence and response to certain drugs. Self-defined ethnicity can therefore not be ignored in the study of genetic disorders (Tate and Goldstein, 2004).

2.11 AIMS

The broad aim of the extended research programme is the clarification of the genetic relationship between different Southern African ethnic groups. The genetic association between ethnic groups is an indication of whether or not ethnicity serves as a marker in

the identification of genetic groups in the Southern African population. This will assist in the sampling strategy employed to obtain a representative sample of the Southern African population.

2.11.1 Specific aims of this investigation

The current study included the following aims:

i) To generate the complete mitochondrial genome sequences of 10 individuals from three different ethnic groups.

ii) To detect and report all alterations between individuals and between individuals and the RCRC.

iii) To determine the genetic make-up of each ethnic group by detecting the following:

- The amount and distribution of sequence differences between individuals.
- The haplogroup assignment of each individual.
- Phylogenetic analysis of the complete mitochondrial sequence of all individuals analysed.

CHAPTER THREE

MATERIALS AND METHODS

Ethical approval was granted by the Ethics Committee of the North-West University in 2002 for a project entitled "Mitochondrial DNA (mtDNA) mutations in patients with a suspected mitochondrial disorder in the South African context" with approval number 02M02. Individals analsed during the current study were also enrolled in the project entitled "Molecular analysis of non-insulin dependent diabetes mellitus (NIDDM) in the South African population" with approval number 02M08. Participants in this investigation were thus enrolled separately in the two projects, and written informed consent was obtained from all individuals.

All reagents utilised in this investigation, unless stated otherwise, were analar quality products supplied by Promega®[1]. The manufacturer's specifications were followed for all kits utilised with modifications thereof indicated where applicable.

3.1 SAMPLE POPULATION

The samples utilised were previously collected during the Profile of Obese Women with Insulin Resistance Syndrome (POWIRS) project at the metabolic unit of North-West University. This project was a multi-disciplinary investigation by numerous departments from North-West University conducted in April and May 2003. The individuals were black African women from different ethnic origins. Ten individuals of different ethnic background were analysed, namely three Zulu, three Xhosa and four Tswana individuals. Blood samples of individuals representing the ethnic groups had previously been obtained in the POWIRS project and the DNA of each individual analysed had been isolated previously by the Centre for Genome Research. Individuals were considered as unlikely to harbour causative mitochondrial mutations, as no mitochondrial disorders were recorded in any of the individuals. The availability of DNA samples, as well as the absence of mitochondrial disorders, determined the criteria for inclusion of individuals. For the purpose of this study it was not necessary to obtain individuals belonging to distinct ethnic groups outside South Africa, as the groups chosen are well represented in South Africa and form distinct ethnic groups.

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

The ethnic origin of an individual was determined utilising language information that was provided by the participant. The ethnicity of the individual as well as that of both parents was utilised to ascertain the ethnic origin of the individual. Individuals having the same ethnicity as both parents were designated as representing that specific ethnic group for the purpose of this study.

3.2 DNA ISOLATION

DNA had previously been collected and isolated during the POWIRS project at the Metabolic Unit of the North-West University in April and May 2003, as discussed in Paragraph 3.1. It was therefore not necessary to re-isolate DNA. Blood had previously been collected and stored in ethylenediamine tetra-acetic acid (EDTA) blood collection tubes to prevent clotting. Genomic DNA (gDNA) had previously been isolated from whole blood utilising a modified NucleoSpin®[1] kit protocol developed by Machery Nagel. The isolation kit specifies that three millilitres (ml) of whole blood be added to three ml chilled buffer C1, to which 9 ml chilled distilled water was added. The tube containing the solution was inverted until translucence was observed and then centrifuged for 15 minutes (min) at 1,300 x gravitational acceleration (x *g*) at four degrees Celsius (°C). Following centrifugation the supernatant was discarded.

Four hundred microlitres (µl) phosphate buffered saline (PBS) solution [4.3 millimolar (mM) di-sodium hydrogen phosphate (Na_2HPO_4), 1.4 mM potassium phosphate monobasic (KH_2PO_4), 137 mM sodium chloride (NaCl), 2.7 mM potassium chloride (KCl)] was added to dissolve the remaining leukocyte pellet via vortexing the solution for one min. The leukocytes were subsequently lysed by the addition of 400 µl buffer B3 and 50 µl proteinase K solution. The tubes were vortexed to ensure the entire pellet was dissolved and in order to ensure complete lysis, it was incubated at 70°C for 10 min. The pellet was homogenised by the addition of 400 µl of ethanol and subsequent vortexing.

Six hundred µl of the above-mentioned solution was added to a NucleoSpin® column, which was placed into a 2 ml collection tube, and subsequently centrifuged at 10,000 x *g* in an Eppendorf®[2] 5810 R centrifuge for 2 min. The column was removed following centrifugation and the flow through discarded and this procedure repeated. Seven µl buffer

[1] NucleoSpin® is a registered trademark of Machery Nagel GmbH & Co., Duren, Germany.
[2] Eppendorf® is a registered trademark of Eppendorf AG, Hamburg, Germany.

B5 was added to wash the matrix bound DNA and subsequently centrifuged at 10,000 x *g* for one min. Following centrifugation, the supernatant was discarded and the DNA was again washed for two min. The addition of two hundred μl prewarmed buffer BE, and subsequent incubation for 15 min at 70°C was performed in order to elute the DNA. This was collected by centrifugation at 10,000 x *g* for two min. The aforementioned step was repeated and the eluted DNA solution amounted to 400 μl.

The light absorbance of each sample was determined spectrophotometrically by means of optical density determination utilising an Eppendorf® BioPhotometer. The DNA and contaminating protein concentrations were measured at 260 nanometres (nm) and 280 nm respectively and a ratio of A_{260}/A_{280} was calculated. DNA working dilutions were made from stock solutions to a final concentration of 50 nanograms (ng) per μl ($ng.\mu l^{-1}$), and stored at 4°C. DNA stock solutions were stored at -20°C. Isolation from three millilitres of whole blood yielded DNA with a concentration range between 132.5 - 1170.3 $ng.\mu l^{-1}$ and an A_{260}/A_{280} absorbance ratio range between 1.77 - 1.88. The lowest value for the DNA concentration was more than sufficient, as the DNA was diluted to make working dilutions to a final concentration of 50 $ng.\mu l^{-1}$. An absorbance ratio greater than 1.75 indicates that the DNA is sufficiently pure for use in PCR amplification. All isolated DNA had an A_{260}/A_{280} absorbance ratio greater than 1.75 and could therefore be utilised.

3.3 AMPLIFICATION OF THE MITOCHONDRIAL GENOME

The complete mitochondrial genome of each individual was amplified utilising a modified version of the standard polymerase chain reaction (PCR) as first described by Mullis *et al.* (1986). Two strategies of the PCR were followed in order to amplify the complete mitochondrial genome, namely a strategy producing small fragments (ca. 600 bp), and a strategy producing long fragments (ca. 2400 bp). The former strategy had previously been optimised, whereas the latter strategy was optimised during the current investigation. The success of each PCR reaction was evaluated via gel electrophoresis.

3.3.1 Polymerase chain reaction

The complete mitochondrial genome was initially amplified in 11 overlapping fragments via a modified long PCR approach utilising primers published by Maca-Meyer *et al.* (2001). The complete mitochondrial genome was amplified in 11 long fragments instead of 32 short fragments. However, not all long PCR fragments could be optimised and it was

therefore necessary to adopt a short PCR fragment strategy in order to amplify these regions.

3.3.1.1 Long PCR

Utilising this strategy, the complete mitochondrial genome was amplified in 11 overlapping fragments, utilising primers listed in Table 3.1. Fragment sizes ranged between 1341 and 2324 bp, with overlaps between successive fragments ranging between 598 to 680 bp, as indicated in Table 3.1, together with the nucleotide (nt) positions of the fragments in the mitochondrial genome. Primer lengths ranged between 20 - 23 bp, with the calculated annealing temperature (T_m) ranging from 47 - 57°C, and are also listed in Table 3.1. The T_m of each primer was calculated utilising the, Oligonucleotide Properties Calculator (Oligonucleotide Properties Calculator, 2004) which is based on the thermodynamic relationship between entropy, enthalpy, free energy and temperature. The thermodynamic parameters of each primer are utilised to predict the temperature-dependent behaviour, and therefore the annealing temperature, of the DNA duplex based on the primer sequence (Breslauer *et al.*, 1986). The annealing temperature utilised for initial optimisation was an estimated value (T_a), calculated utilising Equation 3.1. Complementarity between the chosen combinations of primers was evaluated, utilising Fast PCR version (v) 3.3.112 (Kalendar, 2004). Combinations of primers were chosen in order to prevent excessive self-binding between forward and reverse primers.

Table 3.1: Primers utilised for amplification of the complete mitochondrial genome via a long PCR strategy

Fragment	Primer	T_m	Mean T_m	T_a	Fragment size (bp)	Fragment range (nt)		Overlap (bp)
1	F1:mtL-16340	55	54	52	2,324	16318	2073	628
	R4:mtH-2053	53						
2	F4:mtL-1466	56	54.5	52.5	2,245	1445	3690	639
	R7:mtH-3670	53						
3	F7:mtL-3073	53	54	52	2,276	3051	5327	598
	R10:mtH-5306	55						
4	F10:mtL-4750	47	50	48	2,189	4729	6918	600
	R13:mtH-6899	53						
5	F13:mtL-6337	57	54	52	2,048	6318	8366	505
	R16:mtH-8345	51						
6	F16:mtL-7882	52	52.5	50.5	2,089	7861	9950	608
	R19:mtH-9928	53						

Table 3.1: Continued...

Fragment	Primer	T_m	Mean T_m	T_a	Fragment size (bp)	Fragment range (nt)		Overlap (bp)
7	F19:mtL-9362	51	51.5	49.5	2,204	9342	11546	616
	R22:mtH-11527	52						
8	F22:mtL-10949	54	53	51	2,213	10930	13143	590
	R25:mtH-13124	52						
9	F25:mtL-12572	53	52.5	50.5	2,152	12553	14705	601
	R28:mtH-14685	52						
10	F28:mtL-14125	52	51.5	49.5	2,076	14104	16180	523
	R31:mtH-16157	51						
11	F31:mtL-15676	49	51	49	1,341	15657	429	680
	R1:mtH-408	53						
1	F1:mtL-16340	55	54	52	2,324	16318	2073	N.A.
	R4:mtH-2053	53						

Annealing temperature = experimental obtained annealing temperature utilised for the primer set, F= forward primer and R = reverse primer. T_m = calculated annealing temperature. T_a = estimated annealing temperature. Primer overlap indicates the overlap between the fragment listed and the fragment below it. Primer sequences and names were obtained from Maca-Meyer *et al.* (2001). Nucleotide (nt) positions for fragment ranges are based on the numbering system of Andrews *et al.* (1999). N.A. = not applicable.

A modified standard PCR protocol was utilised for all PCR reactions, which consisted of 0.5 units (U) *Thermus aquaticus* (*Taq*) DNA polymerase, 1 X *Pyrococcus furiosus* (*Pfu*) buffer [20 mM Tris®- hydrochloride (Tris®-HCl) (pH 8.8 at 25°C), 10 mM KCl, 10 mM ammonium sulphate $(NH_4)_2SO_4$, 2 mM magnesium sulphate ($MgSO_4$), 0.1% Triton® X-100 and 0.1 milligrams (mg) per ml ($mg.ml^{-1}$) nuclease-free bovine serum albumin], 0.01 U *Pfu* enzyme in storage buffer [50 mM Tris®-HCl, (pH 8.2 at 25°C), 0.1 mM EDTA, 1 mM dithiothreitol (DTT), 0.05% 3-[(3-cholamidopropyl) dimethylammonio]-1-propanesulfonate (CHAPS) and 50% glycerol), 200 micromolar (μM) of each 2'-deoxynucleotide-triphosphate [2'-deoxyadenosine-5'-triphosphate (dATP), 2'-deoxythymidine-5'-triphosphate (dCTP), 2'-deoxyguanosine-5'-triphosphate (dGTP) and 2'-deoxythymidine-5'-triphosphate (dTTP)], one μl of a 5 picomol (pmol) per μl solution of the appropriate forward and reverse primers and 100 ng of DNA. Sterile double distilled water (ddH_2O) was added to the mixture for a final reaction volume of 12.5 μl. Negative controls were included in every batch of PCR reactions to test for the presence of contamination in the reactions.

Equation 3.1: Calculation of the estimated annealing temperature

$$T_a = \text{Mean } T_m - 2°C$$

T_a = estimated annealing temperature of primer set in °C. Mean T_m = the mean melting temperature of the forward and reverse primers in °C.

Amplification was performed utilising a Thermo Hybaid®[1] Multiblock System (MBS) 0.5S thermocycler through the completion of the following thermal parameters:

- Initial denaturation at 94°C
- 30 cycles of the following programme
  - denaturation at 94°C for two min
  - annealing at specific annealing temperature for two min
  - elongation at 72°C for four min
- Final extension at 72°C for 10 min
- An indefinite holding step at 4°C

Once the PCR cycling programme was completed the samples were evaluated utilising gel electrophoresis.

3.3.1.2 Short PCR

Short PCR refers to the strategy utilised by Maca-Meyer *et al.* (2001) to amplify the complete mitochondrial genome. Twenty eight forward and reverse primer pairs (see Appendix B) published by Maca-Meyer *et al.* (2001) were utilised for amplification, and are presented in Table 3.2, together with the respective primer sequences and annealing temperatures. The lengths of primers ranged between 20 - 23 bp. All fragments were amplified utilising an annealing temperature of 55°C and a magnesium chloride ($MgCl_2$) concentration of 1.5 mM. Negative controls were included in every batch of PCR reactions to verify the absence of contamination in the reactions.

Table 3.2: Primer pairs utilised for amplification and sequencing of the complete mitochondrial genome in short fragments

Primer	Primer Sequence	T_a (°C)	Fragment size (bp)	Fragment range (nt)		Overlap (bp)
F1:mtL16340	5′-agc cat tta ccg tac ata gca ca-3′	55	680	16318	429	67
R1:mtH408	5′-tgt taa aag tgc ata ccg cca-3′					
F2:mtL382	5′-caa aga acc cta aca cca gcc-3′	55	602	362	964	62
R2:mtH945	5′-ggg agg ggg tga tct aaa ac-3′					
F3:mtL923	5′-gtc aca cga tta acc caa gtc a-3′	55	606	902	1508	63
R3:mtH1487	5′-gta tac ttg agg agg gtg acg g-3′					

[1] Thermo Hybaid® is the registered trademark of Hybaid Limited, Ashford, Middlesex, United Kingdom.

Table 3.2: Continued...

Primer	Primer Sequence	T_a (°C)	Fragment size (bp)	Fragment range (nt)		Overlap (bp)
F4:mtL1466	5′-gag tgc tta gtt gaa cag ggc c-3′	55	628	1445	2073	69
R4:mtH2053	5′-tta gag ggt tct gtg ggc aaa-3′					
F5:mtL2025	5′-gcc tgg tga tag ctg gtt gtc c-3′	55	608	2004	2612	74
R5:mtH2591	5′-gga aca agt gat tat gct acc t-3′					
F6:mtL2559	5′-cac cgc ctg ccc agt gac aca t-3′	55	590	2538	3128	77
R6:mtH3108	5′-tcg tac agg gag gaa ttt gaa-3′					
F7:mtL3073	5′-aaa gtc cta cgt gat ctg agt tc-3′	55	639	3051	3690	65
R7:mtH3670	5′-ggc gta gtt tga gtt tga tgc-3′					
F8:mtL3644	5′-gcc acc tct agc cta gcc gt-3′	55	622	3625	4247	58
R8:mtH4227	5′-atg ctg gag att gta atg ggt-3′					
F9:mtL4210	5′-cca ctc acc cta gca tta ctt a-3′	55	624	4189	4813	84
R9:mtH4792	5′-act cag aag tga aag ggg gct a-3′					
F10:mtL4750	5′-cca ata cta cca atc aat act c-3′	55	598	4729	5327	68
R10:mtH5306	5′-ggt gat ggt ggc tat gat ggt g-3′					
F11:mtL5278	5′-tgg gcc att atc gaa gaa tt-3′	55	592	5259	5851	89
R11:mtH5832	5′-gac agg ggt tag gcc tct tt-3′					
F12:mtL5781	5′-agc ccc ggc agg ttt gaa gc-3′	55	625	5762	6387	69
R12:mtH6367	5′-tgg ccc cta aga tag agg aga-3′					
F13:mtL6337	5′-cct gga gcc tcc gta gac ct-3′	55	600	6318	6918	68
R13:mtH6899	5′-gca ctg cag cag atc att tc -3′					
F14:mtL6869	5′-ccg gcg tca aag tat tta gc-3′	55	577	6850	7427	69
R14:mtH7406	5′-ggg ttc ttc gaa tgt gtg gta g-3′					
F15:mtL7379	5′-aga aga acc ctc cat aaa cct g-3′	55	579	7358	7937	76
R15:mtH7918	5′-aga tta gtc cgc cgt agt cg-3′					
F16:mtL7882	5′-tcc ctc cct tac cat caa atc a-3′	55	505	7861	8366	86
R16:mtH8345	5′-ttt cac tgt aaa gag gtg ttg g-3′					
F17:mtL8299	5′-acc ccc tct aga gcc cac tg-3′	55	602	8280	8882	103
R17:mtH8861	5′-gag cga aag cct ata atc act g-3′					
F18:mtL8799	5′-ctc gga ctc ctg cct cac tca-3′	55	637	8779	9416	74
R18:mtH9397	5′-gtg gcc ttg gta tgt gct tt-3′					
F19:mtL9362	5′-ggc cta cta acc aac aca cta-3′	55	608	9342	9950	85
R19:mtH9928	5′-aac cac atc tac aaa atg cca gt-3′					
F20:mtL9886	5′-tcc gcc aac taa tat ttc act t-3′	55	616	9865	10481	98
R20:mtH10462	5′-aat gag ggg cat ttg gta aa-3′					
F21:mtL10403	5′-aaa gga tta gac tga acc gaa-3′	55	611	10383	10994	64
R21:mtH10975	5′-cca tga ttg tga ggg gta gg-3′					

Table 3.2: Continued...

Primer	Primer Sequence	T_a (°C)	Fragment size (bp)	Fragment range (nt)		Overlap (bp)
F22:mtL10949	5'-ctc cga ccc cct aac aac cc-3'	55	616	10930	11546	79
R22:mtH11527	5'-caa gga agg ggt agg cta tg-3'					
F23:mtL11486	5'-aaa act agg cgg cta tgg ta-3'	55	628	11467	12095	87
R23:mtH12076	5'-gga gaa tgg ggg ata ggt gt-3'					
F24:mtL12028	5'-ggc tca ctc acc cac cac att-3'	55	615	12008	12623	70
R24:mtH12603	5'-acg aac aat gct aca ggg atg-3'					
F25:mtL12572	5'-aca acc cag ctc tcc cta ag-3'	55	590	12553	13143	75
R25:mtH13124	5'-att ttc tgc tag gg g gtg ga-3'					
F26:mtL13088	5'-agc cct act cca ctc aag cac-3'	55	617	13068	13685	92
R26:mtH13666	5'-agg gtg ggg tta ttt tcg tt-3'					
F27:mtL13612	5'-aag cgc cta tag cac tcg aa-3'	55	613	13593	14206	102
R27:mtH14186	5'-tgg ttg aac att gtt tgt tgg-3'					
F28:mtL14125	5'-tct ttc ttc ttc cca ctc atc c-3'	55	601	14104	14705	76
R28:mtH14685	5'-cat tgg tcg tgg ttg tag tcc-3'					
F29:mtL14650	5'-ccc cat tac taa acc cac act c-3'	55	603	14629	15232	89
R29:mtH15211	5'-ttg aac tag gtc tgt ccc aat g-3'					
F30:mtL15162	5'-ctc ccg tga ggc caa ata tc-3'	55	596	15143	15739	82
R30:mtH15720	5'-gtc tgc ggc tag gag tca at-3'					
F31:mtL15676	5'-tcc cca tcc tcc ata tat cc-3'	55	523	15657	16180	205
R31:mtH16157	5'-tga tgt gga ttg ggt ttt tat gta-3					
F32:mtL15996	5'-ctc cac cat tag cac cca aag c-3'	55	445	15975	16420	102
R32:mtH16401	5'-tga ttt cac gga gga tgg tg-3'					
F1:mtL16340	5'-agc cat tta ccg tac ata gca ca-3'	55	680	16318	429	N.A.
R1:mtH408	5'-tgt taa aag tgc ata ccg cca-3'					

F = forward primer, R = reverse primer, L = light strand, H = heavy strand. Primer sequences were obtained from Maca-Meyer *et al.* (2001). bp = base pair. Primer overlap indicates the overlap between the fragment listed and the fragment below it. Nucleotide (nt) positions for fragment ranges are based on the numbering system of Andrews *et al.* (1999). Primer names and sequence formats were adapted from Van Brummelen (2003). N.A. = not applicable.

Each PCR reaction contained 1 X PCR buffer (10 mM Tris-HCl, 50 mM KCl and 0.1% Triton® X-100), 0.5 U *Taq* DNA polymerase, 1.5 mM $MgCl_2$, 200 μM of each 2'-deoxynucleotide-triphosphate, 5 pmol of the appropriate forward and reverse primer and 100 ng of DNA. Sterile ddH_2O was added to obtain a final reaction volume of 12.5 μl. Amplification was performed by utilisation of a Thermo Hybaid® MBS 0.5S thermocycler. The following thermal parameters were utilised:

- Initial denaturation at 94°C
- 30 cycles of the following programme
  - denaturation at 94°C for 30 seconds (s)
  - annealing at 55°C 30 s
  - elongation at 72°C for 30 s
- Final extension at 72°C for seven min
- An indefinite holding step at 4°C

Samples were evaluated utilising gel electrophoresis once the PCR programme was completed.

3.3.2 Gel electrophoresis

PCR amplified fragments were evaluated on agarose gels to detect the presence of specific amplification products. Long PCR fragments were evaluated via the utilisation of a 0.9% weight per volume (w/v) agarose gel, whilst short PCR fragments were evaluated utilising a 2.0% w/v agarose gel. Gels where either mini (30 ml) or midi (100 ml) sized and were made with 1 X TBE buffer (89.15 mM Tris base [pH 8.0], 88.95 mM boric acid and 2.5 mM di-sodium EDTA). Gel staining was achieved through the addition of ethidium bromide (EtBr), from Sigma®[1], to form a final concentration of 0.5 micrograms (μg) per ml (μg ml^{-1}). Two and a half μl of PCR product was mixed with 2.0 μl 1 X loading buffer (0.04% Orange G, 50% glycerol) and loaded into the wells. Gels were electrophoresed in 1 X TBE buffer, at 10 volts per centimetre ($V.cm^{-1}$), for 50 min and 30 min for long and short PCR fragments respectively. Amplified fragments were visualised under an ultraviolet transilluminator. Gels utilised to separate long PCR fragments were loaded with a 1 kilobase pair (kb) molecular weight marker, and gels utilised to separate short PCR fragments with a 100 bp molecular weight marker. Molecular weight markers served as a reference of known size to which amplified fragments could be compared and their size estimated.

3.4 CHAIN TERMINATION CYCLE SEQUENCING

The PCR products produced, as discussed in Paragraph 3.3.1, were subjected to cycle sequencing in order to obtain the nucleotide sequence of the amplified fragments. These PCR products were purified prior to being sequenced, as discussed in Paragraph 3.4.1.

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

Sequencing was performed utilising the automated cycle sequencing method with chain-termination inhibitors, as first described by Sanger *et al.* (1977). The generated sequence fragments were precipitated and analysed electrophoretically on a genetic analysis system sequencer.

3.4.1 Purification of amplified product

PCR products were purified to remove excess primers and impurities, such as salts, enzymes and unincorporated nucleotides, which may interfere with downstream automated sequencing. This was achieved by utilising the QIAquick®[1] silica-gel membrane based spin-column as per the manufactures instructions, which are described below.

Five volumes of Buffer PB was added to the PCR sample in order to alter the pH and salt content to allow binding of the DNA to the silica matrix. The mixture was applied to the column to allow binding of DNA and centrifuged at 10,000 x *g* for one min to allow contaminants to be eluted together with Buffer PB. The column-bound DNA was washed to remove primers and impurities through the addition of 0.75 ml ethanol-containing Buffer PE to the column and subsequent centrifugation at 10,000 x *g* for one min. Residual Buffer PE was removed by an additional centrifugation step of 10,000 x *g* for one min. The DNA was released from the column by increasing the pH. This was achieved through the addition of 30 μl of Buffer EB. The column was allowed to stand for one min to provide maximum yield. The DNA was eluted through centrifugation at 10,000 x *g* for one min. The concentration of the eluted DNA was determined spectrophotometrically utilising an Eppendorf® BioPhotometer. This allowed for the determination of the correct amount of purified PCR template to be utilised in sequencing reactions.

3.4.2 Chain termination sequencing

The complete mitochondrial genome was sequenced in 32 overlapping fragments, utilising primers listed in Table 3.2. Single fragments generated via long PCR, as presented in Table 3.1, were sequenced utilising several forward primers. Fragments 1 to 10 contained four different forward primer binding sites, of which the first three were utilised for sequencing. Fragment 11 contained three forward primer binding sites, of which the first two were utilised for sequencing. Short PCR generated fragments were sequenced utilising a single forward primer.

---

[1] QIAquick® is a registered trademark of QIAGEN Inc., Valencia, CA, USA.

The sequencing reaction was performed utilising the ABI PRISM®[1] BigDye™[2] Terminator v3.1 Ready Reaction Cycle Sequencing Kit with AmpliTaq®[3] DNA Polymerase FS. The sequencing enzyme is a variant of *Taq* DNA polymerase, as it contains two point mutations. A mutation in the active site results in less discrimination against dideoxynucleotides, therefore generating a signature with even peaks. A mutation in the amino terminal domain virtually eliminates the 5' → 3' exonuclease activity of AmpliTaq® DNA Polymerase FS.

This method utilises a chain termination sequencing approach by incorporation of four 2',3'-dideoxynucleotide triphosphates, which are labelled with different coloured fluorophores to allow for different bases to be distinguished and detected. The cycle sequencing reaction contained ca. 20 ng of purified PCR product, 3.2 pmols of the appropriate primer, 2.0 μl of the BigDye™ 5 X Sequencing Buffer and 2.0 μl BigDye™ Ready Reaction Premix, which contains BigDye™ terminators adenine (A), cytosine (C), G and T, deoxynucleoside triphosphates [dATP, dCTP, 2'-deoxyinosine-5'-triphosphate and 2'-deoxyuridine-5'-triphosphate]), deoxynucleotide triphosphates, AmpliTaq® DNA Polymerase FS, $MgCl_2$ and Tris-HCl buffer (pH 9.0). Sterile ddH_2O was added to obtain a final reaction volume of 10 μl.

Cycle sequencing was performed utilising a Thermo Hybaid® MBS 0.5S thermocycler through the completion of the following thermal parameters:

- Initial denaturation at 94°C
- 25 cycles of the following programme
  - rapid thermal ramp to 96°C
  - denaturation at 96°C for 10 s
  - rapid thermal ramp to 50°C
  - annealing at 50°C for 10 s
  - rapid thermal ramp to 60°C
  - elongation at 60°C for four min
- An indefinite holding step at 4°C

[1] ABIPrism® is a registered trademark of Applera Corporation, Foster City, CA, USA.
[2] BigDye™ is a trademark of Applera Corporation, Foster City, CA, USA.
[3] AmpliTaq® is a registered trademark of Roche Molecular Systems Inc., Pleasanton, CA, USA.

Once cycle sequencing was performed it was necessary to remove unincorporated dye terminators before samples could be analysed by electrophoresis, as excess dye terminators may interfere with base calling by obscuring the initial sequence data. Removal of unincorporated dyes was achieved through ethanol/sodium acetate (NaOAc) precipitation. PCR amplified products were precipitated by adding 80 μl of a three molar (M) NaOAc solution in 95% ethanol and allowing the mixture to stand at room temperature (RT) for 15 min. The precipitate was collected through centrifugation at 10,600 x *g* for 20 min. The supernatant was discarded and the precipitated products washed with 250 μl 70% ethanol and centrifuged at 10,600 x *g* for 10 min. The supernatant was discarded and the precipitated pellet allowed to air dry at RT for 15 min to allow excess ethanol to evaporate, which may interfere with base calling. The precipitate was resuspended in six μl Hi-Di™[1] deionised formamide and three μl was utilised for electrophoretic separation on a SpectruMedix™[2] (SCE2410) Genetic analysis system sequencer.

3.5 SEQUENCE ANALYSIS

Generated sequences were evaluated and manually aligned utilising BioEdit v5.0.9 (Hall, 1999). The overlapping sequences of the complete mitochondrial genome were aligned against the RCRS (NCBI, 2004a) in order to obtain the correct numbering system. Contiguous sequences of the complete mitochondrial genomes were therefore constructed. Primer sequence present at terminal ends of sequences as a result of primer inclusion during PCR was excluded. Regions of sequences that were generated utilising primers as templates were obtained from the adjacent overlapping fragment sequence. Generally sequences were generated through sequencing only in the forward direction. Sequencing in the reverse direction was performed in order to obtain the sequence of the 568-577insPolyC alteration, as discussed in Paragraph 4.2.27. The two novel transitions located in the 16S rRNA gene, as discussed in Paragraph 4.2.2, were not confirmed by sequencing in the reverse direction, as they were predicted not to have a functional effect. As non-synonymous alterations are unlikely to result in a functional effect, the novel non-synonymous alterations detected in this study were not confirmed by sequencing in the reverse direction. Although the novel non-synonymous C12436T alteration detected in the current investigation and discussed in Paragraph 4.2.24.2 is regarded as unlikely to have a functional effect, it is essential to confirm this alteration by sequencing in the reverse direction in future studies.

[1] Hi-Di™ is a trademark of Applera Corporation, Foster City, CA, USA.
[2] SpectruMedix™ is a trademark of SpectruMedix LLC., State College, PA, USA.

3.5.1 Nucleotide differences

Nucleotide alterations relative to the RCRS were categorised based on their nature and location. Alterations were divided into nucleotide substitutions, namely transitions (TS) and transversions (TV) and length alterations namely insertions (ins) and deletions (del). Alterations were classified according to their location in the mitochondrial genome. A distinction was made between alterations occurring in protein coding regions, rRNA-coding regions and non-coding regions (which included the D-loop and intergenic regions). Alterations occurring in protein-coding regions were divided further into synonymous and non-synonymous alterations.

The number of nucleotide differences between individuals was counted in several different ways, as listed below:

a. The number of nucleotide differences between all combinations of pairs of individuals was counted.
b. Alterations that occurred between more than two individuals were identified and counted.
c. Combinations of individuals that shared a specific alteration were identified and the combination was termed a motif.
d. The number of times a specific motif was detected was counted.
e. The number of unique shared alterations in identified haplogroups (as discussed in Paragraph 3.5.2) and ethnic groups, was counted.

The above-mentioned methods of investigating alterations were performed with the specific aim of determining the genetic relationships between all individuals. Individuals comprising a single haplogroup and individuals in each of the ethnic groups were also compared in the current investigation.

3.5.2 Haplogroup assignment

Haplogroups were assigned to individuals based on combinations of mtDNA polymorphisms. Haplogroups were identified utilising unpublished criteria generated by Wallace (2003). Individuals were haplogrouped through the investigation of single nucleotide polymorphisms (SNP) that are specific to haplogroups in macro-haplogroup L.

3.5.3 Phylogenetic analysis

Phylogenetic analysis was performed utilising the nucleotide sequence data generated in the study. The complete mitochondrial genome sequences, each comprising 16,569 bp, were utilised from all 10 individuals included in the current study. Sequences were analysed via the maximum parsimony (MP) and the unweighted pair-group method using arithmetic averages (UPGMA) via the Molecular Evolutionary Genetics Analysis (MEGA) v2.1 software programme (Kumar *et al.*, 2001). The full mitochondrial sequence of each individual was analysed as a single entity. Nucleotides were divided into non-coding and coding, as well as first, second and third codon positions.

CHAPTER FOUR

RESULTS AND DISCUSSION

The results generated from the complete mtDNA sequences of 10 individuals and subsequent discussions thereof are presented in this chapter. These individuals belong to three different ethnic groups, namely Xhosa, Zulu and Tswana. Sequences were compared with one another and nucleotide differences between individuals were investigated. The haplogroup of each individual was determined, and the sequence similarities between individuals in a haplogroup were resolved. The sequence similarities between individuals in an ethnic group were also investigated. Finally, it was determined whether the ethnicity of individuals, as defined in this study, reflects their mitochondrial genetic composition.

4.1 OPTIMISATION OF EXPERIMENTAL PROTOCOLS

Procedures that were performed utilising commercially available kits were followed strictly according to the manufacturer's guidelines and this yielded the expected results. It was therefore not necessary to optimise these procedures. However, for procedures not performed utilising commercially available kits, it was necessary to first optimise the procedure utilising control DNA samples.

4.1.1 Polymerase chain reaction

The complete mitochondrial genomes were amplified utilising primer sets that generated overlapping fragments. An initial strategy for amplification was adopted, namely that of long PCR, as described in Paragraph 3.3.1.1, in which the complete genome was amplified utilising 11 primer sets, creating 11 overlapping fragments. Each fragment contained a significant overlap of ca. 600 bp with the adjoining fragments. PCR amplification yielded the desired fragments, once the reactions were optimised, with no secondary amplification detected via agarose gel electrophoresis. However, the sequencing reactions of specific long PCR fragments of certain individuals failed even after extensive optimisation of the PCR and sequencing reactions. Results of the long PCR are described in Paragraph 4.1.1.1. A short PCR strategy that had previously been optimised was adopted to amplify the specific fragments that failed via long PCR, as

described in Paragraph 3.3.1.2. Results generated via the short PCR strategy, that amplified the mitochondrial genome in 32 overlapping fragments, are presented in Paragraph 4.1.1.2.

4.1.1.1 Long PCR strategy

Fragments generated via the long PCR strategy contained either three or four forward primer annealing sites that could be used to sequence the fragment, as discussed in Paragraph 4.1.3. The overlap generated in this strategy thus consisted of an entire fragment generated by a forward and reverse primer set. Primer pairs chosen for amplification are indicated in Table 3.1 (page 35). Annealing temperatures utilised in the long PCR strategy are listed in Table 4.1.

Table 4.1: Annealing temperatures for amplification of fragments via the long PCR strategy

Fragment	T_a (°C)	Fragment	T_a (°C)	Fragment	T_a (°C)
1	57	5	60	9	65
2	54	6	58	10	55
3	65	7	65	11	60
4	52	8	65	N.A.	N.A.

T_a = experimentally obtained annealing temperature utilised for amplification. N.A. = not applicable.

PCR reactions were performed according to a modified standard PCR protocol, as described in Paragraph 3.3.1.1. Fragments amplified were longer than those generated via the short PCR strategy, and therefore certain steps/stages of the standard short PCR protocol could not be utilised. Cycle parameters of the PCR programme and DNA polymerase composition were altered accordingly. The denaturing time at 94°C was increased to 2 min to allow for complete denaturation of the double-stranded DNA. The extension time at the respective annealing temperatures was increased to 4 min to allow for the complete extension of the long PCR products.

In long PCR, *Pfu* DNA polymerase was added to a final concentration of 0.004 U per μl ($U.\mu l^{-1}$). *Taq* DNA polymerase lacks 3' → 5' exonuclease (proofreading) activity, resulting in a mis-incorporation rate of 2.1×10^{-4} nucleotides per cycle (Keohavong and Thillly, 1989). As these fragments are relatively large, the chance of an error occurring in the extension product is increased. *Pfu* DNA polymerase is a high fidelity DNA polymerase exhibiting 3' → 5' exonuclease activity, therefore decreasing the probability of

mis-incorporation in the extension product. A very low concentration of *Pfu* DNA polymerase was sufficient and optimal for the removal of mismatches. Utilising *Pfu* DNA polymerase as the sole amplification enzyme would have resulted in degradation of the PCR primers. This occurs as a result of the 3' → 5' exonuclease property of *Pfu* DNA polymerase, especially during the long elongation periods of the PCR cycling programme (Barnes, 1994).

It was necessary to optimise the T_a of PCR reactions via the utilisation of control DNA. This allowed for reactions to be performed optimally without the depletion of sample DNA stocks. The estimated T_a (Table 3.1, page 35) was chosen for initial optimisation of each primer set. This increased the likelihood of obtaining product, as the T_a was lower than the T_m. The T_a was adjusted accordingly in order to generate a single specific fragment of desired intensity. After extensive efforts, not all PCR reactions producing long fragments were successful via optimisation of the T_a and $MgCl_2$ concentration of the reactions. A short PCR strategy that had previously been optimised was therefore adopted for these reactions, and is discussed in Paragraph 4.1.1.2. Only the T_a and $MgCl_2$ concentration parameters were optimised, as these could be altered easily and quickly. The optimisation of additional parameters would have been more time consuming than the utilisation of the short PCR strategy.

4.1.1.2 Short PCR strategy

Primers utilised for amplification of the short fragments are listed in Table 3.2 (page 37). All fragments were easily amplified utilising the strategy as discussed in Paragraph 3.3.1.2. A complete list of fragments, from the different individuals amplified via short PCR, is presented in Appendix B. All remaining fragments not listed in Appendix B were amplified and sequenced successfully via the long PCR strategy.

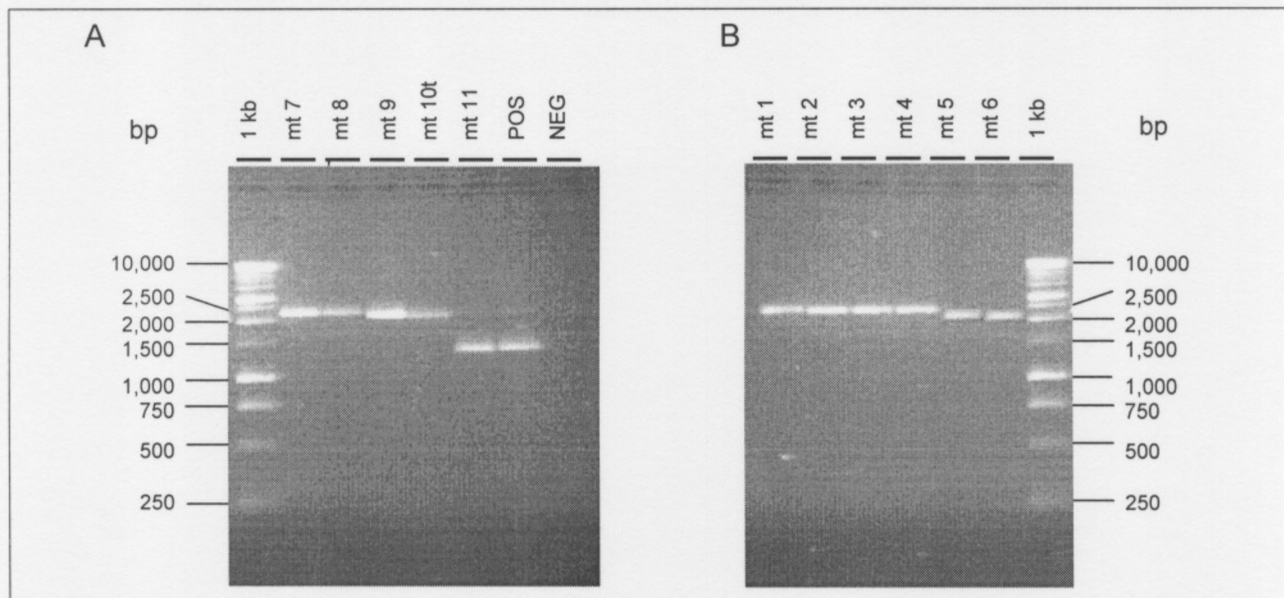
4.1.2 PCR product electrophoresis and purification

To evaluate the outcomes of the PCR reaction, the products were separated utilising agarose gel electrophoresis as discussed in Paragraph 3.3.2. Separation of PCR amplified products together with fragments of known size, namely a molecular weight marker that serves as a reference, allowed for the estimation of the amplified product size. Gel electrophoresis provides relatively low resolution visualisation and thus fragment sizes are

only an estimate of the true size. For this reason this technique was utilised mainly to confirm the presence of specific fragments.

The 11 fragments, generated via the long PCR strategy discussed in Paragraph 3.3.1.1, encompassing the complete mitochondrial genome, ranged in size from 1,341 to 2,324 bp, with an average length of ca. 2,105 bp. The fragments were separated on a 0.9 % (w/v) agarose gel, which allowed an optimum migration rate for the relatively large fragments. A photographic representation of the 11 fragments is presented in Figure 4.1. Not all the fragments could be distinguished from one another according to size, as the gel system utilised had insufficient resolution power to separate fragments, except for fragment 11 which was 1,341 bp in length.

Figure 4.1: Photographic representation of PCR fragments generated via long PCR



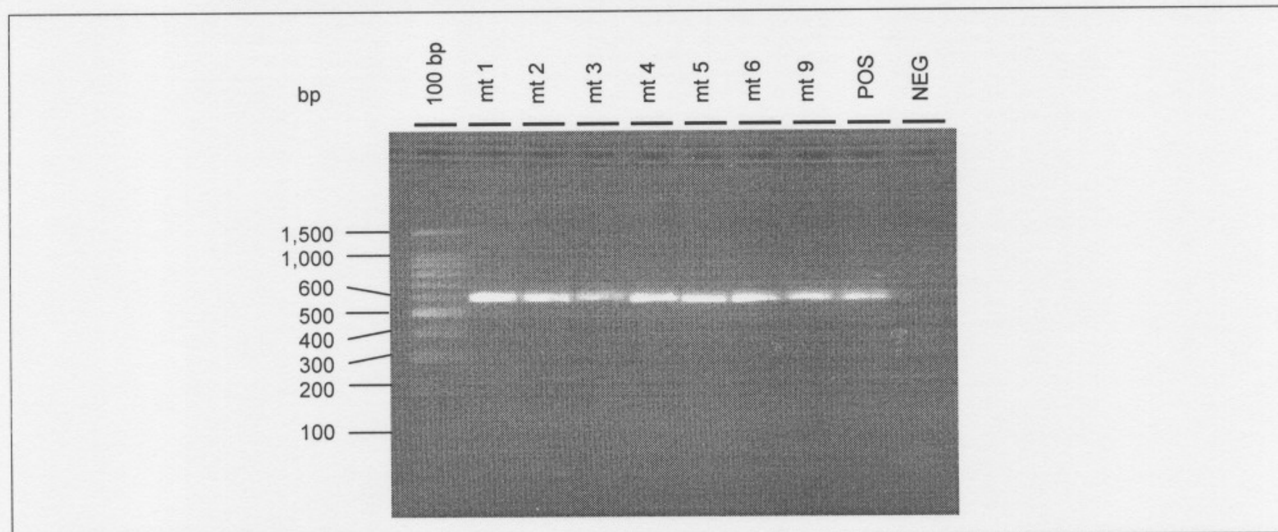
0.9% agarose gel electrophoresed at 10 V.cm^{-1} for 50 min in 1 X TBE buffer. 1 kb = 1 kilobase pair molecular weight marker. NEG = negative control. POS = positive control. mt 1 - mt 11 = fragments 1 - 11 generated via the long PCR strategy. The POS and NEG controls are of fragment 11, and represent the same results obtained for all the other fragment controls. Sizes of fragments are indicated in base pairs (bp). Figures 4.1 A and B represent a single photographic representation that has been split in two in order to include the molecular weight marker at the side of each representation.

Fragments generated via the short PCR strategy, as described in Paragraph 3.3.1.2, ranged in size between 445 to 680 bp, with the majority of fragments having an average length of ca. 600 bp. The fragments were indistinguishable with respect to size when evaluated via agarose gel electrophoresis. A photographic representation of selected fragments is presented in Figure 4.2.

Successfully amplified products were purified in order to remove impurities such as primers, salts, enzymes and unincorporated nucleotides, which may interfere with the

sequencing reaction. This was achieved through utilising the QIAquick® PCR purification kit, as discussed in Paragraph 3.4.1. The protocol was effective in terms of time and the quality of purified product generated. Concentrations observed for different regions and samples varied between 10 and 25 ng.μl^{-1}.

Figure 4.2: Photographic representation of PCR fragments 1, 2, 3, 4, 5, 6 and 9 generated via short PCR



2.0% agarose gel electrophoresed at 10 V.cm^{-1} for 30 min in 1 X TBE buffer. 100 bp = 100 base pair molecular weight marker. NEG = negative control. POS = positive control. Numbers indicate fragment numbers. mt = mitochondrial DNA fragments 1 – 6 and 9 of 28 fragments. Fragments 1 – 6 and 9 are indicated as representative of all fragments generated via the short PCR strategy. The POS and NEG controls are of fragment 9, and represent the same results obtained for all the other fragment controls. Sizes of fragments are indicated in base pairs (bp).

4.1.3 Cycle sequencing

The sequencing strategy discussed in Paragraph 3.4.2 was adopted for fragments generated via the long and short PCR strategies. Fragments were sequenced in the forward direction, utilising forward primers listed in Table 3.2 (page 37). The utilisation of 20 ng of template DNA for all reactions, and precipitating the sequencing products, as described in Paragraph 3.4.2, provided the most successful results. However, certain fragments generated via the long PCR strategy could not be sequenced successfully. The electropherogram illustrating the sequence was unreadable, as the true sequence peaks seemed fused with a mass of background peaks of equal intensity. Extensive optimisation of template quantity, the precipitation procedure and the annealing temperature of the PCR reaction that generated the specific fragment failed to generate the desired results.

All fragments generated via the long PCR strategy that could not be successfully sequenced were amplified via the short PCR strategy. Optimal sequencing quality was obtained when the short fragments were utilised as templates in the sequencing reaction.

A possible reason to substantiate the lack of success of the sequencing reactions when utilising long PCR fragments as templates, could have been the non-specific annealing of the forward primer in the region amplified by the respective forward and reverse primers. Possible annealing of forward primers to secondary sites in their respective long PCR fragments was evaluated utilising Fast PCR v3.3.112 (Kalendar, 2004) and BLAST (Altschul *et al.*, 1997; NCBI, 2004b). No significant secondary annealing sites were detected in the mitochondrial or nuclear genomes.

Non-specific annealing results in secondary amplification in the PCR reaction. The secondary product generated via long PCR would be greater than that observed in the short PCR reaction performed to generate the corresponding region. This is due to the secondary annealing site being present in the primary product after each amplification step in the PCR reaction. The sequencing reaction was therefore performed utilising mixed template, which resulted in an electropherogram that was unreadable due to the presence of overlapping peaks.

The distribution of sequencing failures was not random among individuals, indicating possible sequence alterations in primer annealing sites. Only three alterations were detected in primer annealing regions, namely the novel A9350G alteration in the primer F19 annealing site, T15670C in the primer F31 annealing site and T16327C in the primer F1 annealing site, and are discussed in Paragraphs 4.2.18.1, 4.2.26 and 4.2.28 respectively.

All these alterations were detected in either individual 934 or 939. However, alterations in primer annealing sites were predicted as unlikely to contribute to the failure of sequences in all individuals, via this PCR strategy. This is because the corresponding regions were successfully sequenced via utilisation of the short PCR templates. Fragments of the different individuals that were amplified via the short PCR strategy are listed in Appendix B.

4.2 COMPLETE MITOCHONDRIAL GENOME SCREENING

Differences detected in the generated sequences relative to the RCRS are presented in this section. Alterations are presented per locus region in order of occurrence, starting at nucleotide 577 (Andrews et al., 1999). Haplogroups were determined utilising haplogroup informative alterations according to the haplogroup criteria as described by Wallace

(2003). All other alterations presented in this section, unless otherwise stated, can be assumed to be uninformative for macro-haplogroup L. The complete distribution of alterations detected in all individuals analysed in this study is presented in Appendix A. Only the alterations that were considered to be novel are illustrated with a representative electropherogram. Although essential for future studies, it was not in the scope of the current study to determine the frequency of the alterations in reported populations, nor was it an objective to determine in which populations the alterations occurred.

The strategy of full mitochondrial genome sequencing was pursued for several reasons. Firstly, this allowed for the comparison of the maximum number of mitochondrial nucleotides between individuals (Salas *et al.*, 2002). Secondly, the analysis of both fast and slow evolving sites allowed for the maximum resolution of sequence differences between closely related mtDNAs (Vigilant *et al.*, 1989). Thirdly, it allowed for the elucidation of all mitochondrial alterations present between the individuals analysed. Finally, potential ambiguities could be resolved through the inclusion of information from many mtDNA segments in a given genome (Torroni, 2000). Therefore, to allow for the highest possible resolution of nucleotide evolution in populations, the study of rapidly evolving, non-recombining sequences, such as mtDNAs, are ideal (Vigilant *et al.*, 1989).

It was essential to confirm the novel status of an alteration utilising the methods described in Paragraph 3.5. Alterations not detected in databases that were searchable by nucleotide position alone are occasionally considered as novel by various authors. However, several alterations detected in the current investigation, and initially considered as novel, were later discovered to have been reported previously when the alterations were searched for via the utilisation of the actual sequence data. For example, the A3756G alteration was not detected after an extensive database and literature search. A BLAST search (Altschul *et al.*, 1997; NCBI, 2004b) utilising sequence data of the alteration, however, listed the alteration as reported by Mishmar *et al.* (2003), a reference source that had previously been investigated. The alteration was not detected in this paper, as no specific alterations were listed. The paper did reference the GenBank® accession number of the sequence, which contained the alteration that was detected via the BLAST search. However, the publication did not state that any specific novel alterations had been detected. It would therefore be impractical to scan every GenBank® sequence listed in a publication for possible novel alterations. It is therefore essential to utilise the BLAST method described above when assigning novel status to an alteration.

Furthermore, it was important to verify every alteration that was detected in order to eliminate errors in the final data. This was achieved through analysis of the electropherogram representing each alteration and reanalysis of processed data. Phantom mutations, i.e. systematic artefacts generated in the course of the sequencing process itself, are not detectable through analysing electropherograms (Bandelt *et al.*, 2002). However, most errors in published mtDNA sequences are the result of poor documentation (Dennis, 2003) and can be corrected by systematic editing. It has been documented that errors in published documents have resulted in erroneous conclusions and subsequent rejection of research papers (Cann *et al.*, 1987; Hagelberg *et al.*, 1999).

Alterations detected during this investigation were divided into two main classes, namely nucleotide substitutions, which consisted of transitions and transversions, and length alterations, namely insertions and deletions. The synonymous substitutions in protein-coding regions that did not result in an amino acid replacement were distinguished from the non-synonymous alterations that resulted in amino acid replacements (Nei and Kumar, 2000). In the given context, polymorphisms refer to alterations that fail to evoke a disease phenotype, and not to the general definition of a polymorphism being an alteration in which the most common allele has a frequency of less than 0.99 (Falconer and Mackay, 1996). All novel alterations are defined with the term alteration and not polymorphism. Novel alterations predicted to have a functional effect may evoke a disorder in an individual even though the functional effect is not phenotypically visible. Novel alterations have been detected in individuals in the current study only. Therefore, no information on the population frequency of these alterations is available.

4.2.1 The 12S ribosomal RNA gene

The 12S rRNA gene spans the 648 - 1601 position (MITOMAP, 2004) of the human mitochondrial genome and encodes an mRNA product that is not translated (Wallace, 1992). Nine different alterations were detected in the 12S rRNA gene when compared to the RCRS. Table 4.2 lists all alterations observed in the 12S rRNA gene together with the distribution of alterations between individuals. All nine of the alterations have previously been reported as polymorphisms, of which the T1243C polymorphism (Ozawa *et al.*, 1991a) has been observed to render Japanese individuals susceptible to developing type 2 diabetes (Taylor *et al.*, 2003). The A at position 1438 is present in the RCRS as a rare polymorphism, whereas the G is the more common nucleotide (Andrews *et al.*, 1999). The

A at nucleotide 1438 is maintained in order for the RCRS to represent a reference sequence rather than a consensus sequence.

Table 4.2: Alterations detected in the 12S rRNA gene

Alteration[1]	Individuals	Alteration Type	Reference
G709A	939	Transition	Ozawa *et al.* (1991b)
G719A	941, 962, 972	Transition	Maca-Meyer *et al.* (2001)
G769A	926, 939, 941, 962, 964, 972 976	Transition	Moraes *et al.* (1993)
T825A	941, 964, 972	Transition	Prezant *et al.* (1993)
G1018A	926, 995, 939, 941, 962, 964,972, 976	Transition	Prezant *et al.* (1993)
C1048T	995, 941, 962, 964, 972	Transition	Moraes *et al.* (1993)
T1243C	964	Transition	Taylor *et al.* (2001)
A1438G	921, 926, 995, 934, 939, 964, 972	Transition	Ozawa *et al.* (1991b); Andrews *et al.* (2001)
G1442A	939	Transition	Palanichamy *et al.* (2004)

1 = Alteration format includes initial nucleotide present in the RCRS, followed by the nucleotide position and then by the nucleotide detected. The RCRS was utilised as the reference sequence (Andrews *et al.*, 1999).

4.2.2 The 16S ribosomal RNA gene

The 16S rRNA gene is localised to position 1671 - 3228 of the mitochondrial genome and is a 1,557 bp gene that encodes a functional mRNA that is not translated (Wallace *et al.*, 1992). A total of 15 alterations were detected in this gene and are presented in Table 4.3. Fourteen of the alterations detected have previously been described as polymorphisms. After extensive literature searches, references to the T2887C and the C3157T transitions could not be found. They are thus considered novel in the current investigation and are discussed in Paragraphs 4.2.2.1 and 4.2.2.2 respectively.

Table 4.3: Alterations detected in the 16S rRNA gene

Alteration[1]	Individuals	Alteration Type	Reference
C1706T	939	Transition	Herrnstadt *et al.* (2002)
G1709A	926	Transition	Zhoa *et al.* (2004)
A2245G	995, 962	Transition	Moraes *et al.* (1993)
C2332T	939	Transition	Herrnstadt *et al.* (2002)
T2352C	921, 995, 934	Transition	Chen *et al.* (1995)
A2358G	939	Transition	Herrnstadt *et al.* (2002)
T2416C	926, 939, 976	Transition	Herrnstadt *et al.* (2002)
T2483C	921	Transition	Herrnstadt *et al.* (2002)
A2706G	921, 926, 995, 934, 939, 962, 976	Transition	Ozawa *et al.* (1991a)
G2758A	941, 962, 964, 972	Transition	Moraes *et al.* (1993)

Table 4.3: Continued...

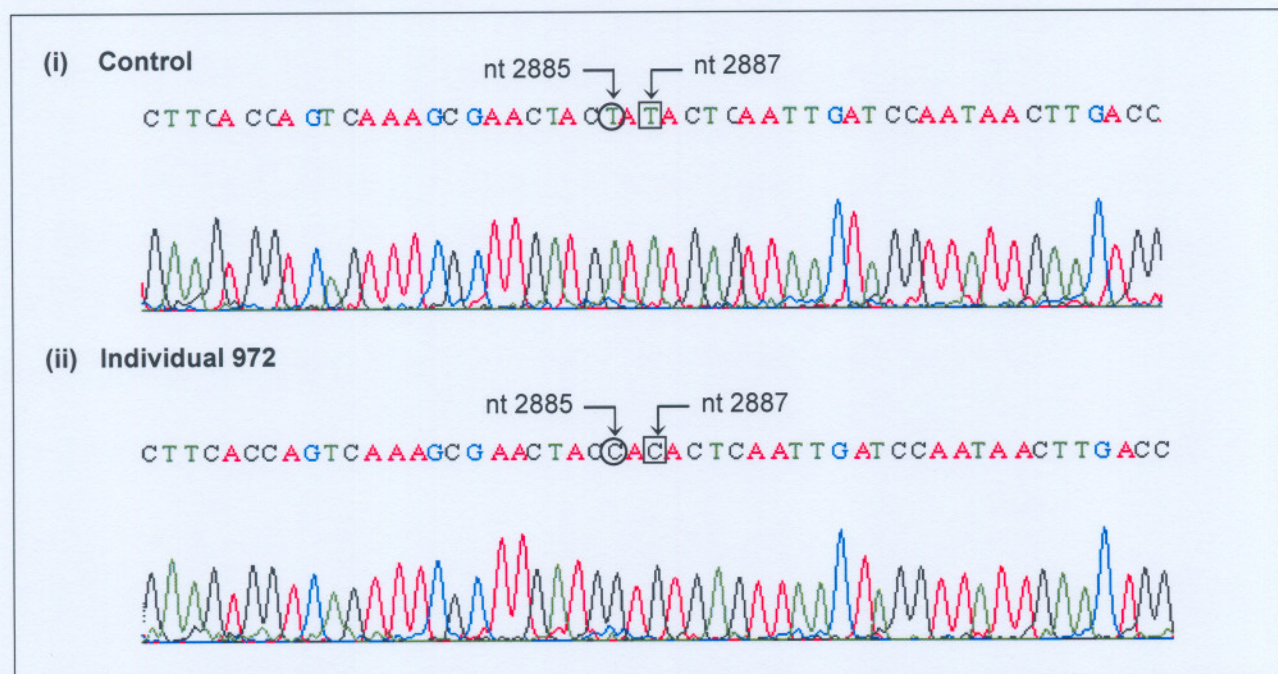
Alteration[1]	Individuals	Alteration Type	Reference
C2789T	926, 976	Transition	Rose *et al.* (2001)
T2885C	941, 962, 964, 972	Transition	Moraes *et al.* (1993)
T2887C	972	Transition	Not yet reported
G3010A	976	Transition	Marzuki *et al.* (1991)
C3157T	941	Transition	Not yet reported

1 = Alteration format includes initial nucleotide present in the RCRS, followed by the nucleotide position and then by the nucleotide detected. The RCRS was utilised as the reference sequence (Andrews *et al.*, 1999).

4.2.2.1 The T2887C alteration in the 16S rRNA gene

An electropherogram illustrating the T2887C alteration, which was detected only in individual 972, is presented in Figure 4.3. This alteration is reported for the first time in the current investigation, subsequent to an extensive literature search during which no reference to this alteration could be identified. The T2887C transition is located in the 16S rRNA gene. This gene is not translated, therefore alterations occurring in this gene will alter only the rRNA sequence.

Figure 4.3: Representative electropherogram of the T2887C alteration in the 16S rRNA gene



Control = representative electropherogram of an individual with identical sequence to the RCRS. Nucleotides enclosed in a box (□) represent novel alterations, and those enclosed in a circle (○) represent reported changes. Arrows indicate the nucleotide (nt) positions at which alterations occur.

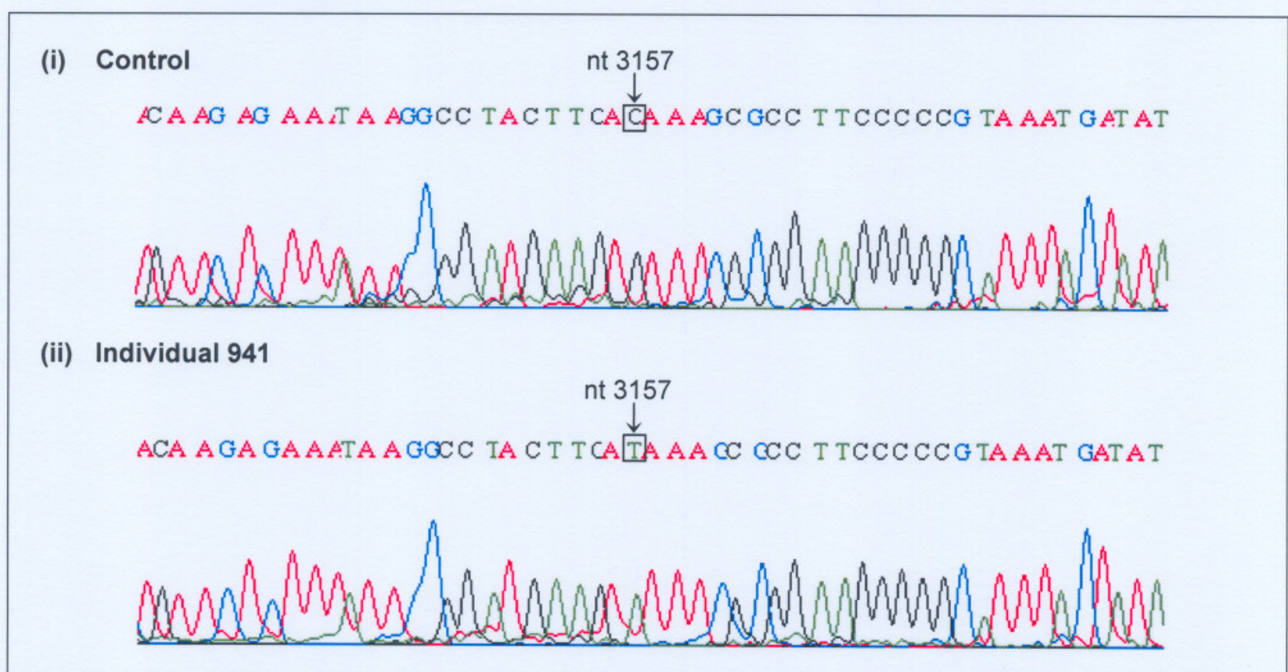
The T2887C transition is located in a loop of the central region of the 16S rRNA secondary structure and therefore does not bind to another nucleotide in order to maintain the secondary structure of 16S rRNA molecule (Glotz *et al.*, 1981). The overall secondary

structure of the 16S rRNA is conserved. However, T2887C is not located in a protein binding site (Glotz *et al.*, 1981), and is not considered as an evolutionary conserved site (Gravel *et al.*, 1987; Moazed and Noller, 1987). Included in the electropherogram presented in Figure 4.3 is the T2885C transition listed in Table 4.3, which represents a polymorphism that was detected in individuals 941, 962, 964 and 972 during this investigation. The high level of background observed in the electropherogram illustrating the sequence of individual 972 does not interfere with the sequence results. The intensity of the true sequence is far greater than that generated by the background.

4.2.2.2 The C3157T alteration in the 16S rRNA gene

The C3157T alteration, detected in individual 941, is located in the 16S rRNA gene. This transition is reported for the first time in the current investigation after an extensive literature search failed to detect this change. An electropherogram illustrating the transition is presented in Figure 4.4.

Figure 4.4: Representative electropherogram of the C3157T alteration in the 16S rRNA gene



Control = representative electropherogram of an individual with identical sequence to the RCRS. Nucleotides enclosed in a box (□) represent novel alterations. Arrows indicate the nucleotide (nt) positions at which alterations occur.

The C3157T transition is located in a loop structure in the 3'-end of the 16S rRNA secondary structure and does not bind to another nucleotide. Therefore, the transition is unlikely to result in an alteration in the secondary structure as the nucleotide does not contribute to the preservation of the secondary structure of the 16S rRNA molecule (Glotz

et al., 1981). Considering the aforementioned and the fact that the 3516 nucleotide site is not evolutionary conserved (Gravel *et al.*, 1987; Moazed and Noller, 1987), it is unlikely that the C3157T transition is functionally significant. The significance of this alteration can be confirmed by performing functional studies as a result of the amino acid replacement.

4.2.3 The transfer RNA leucine 1 gene

The tRNA leucine ($tRNA^{Leu}$) gene is located in the 3230 - 3304 position (MITOMAP, 2004) and encodes a functional tRNA product that is not translated (Anderson *et al.*, 1981). A single alteration was detected in this gene, namely the G3277A transition, which has previously been reported as a polymorphism (Sternberg *et al.*, 2001). This alteration was detected only in a single individual, 921, during this study.

4.2.4 NADH dehydrogenase subunit 1 gene

The NADH dehydrogenase subunit 1 (ND1) gene encodes an mRNA product which is translated into a functional polypeptide, namely ND1 (Wallace, 1992), which is a component of the large hydrophobic domain of complex i (Walker, 1992). The gene is located in nucleotides 3307 - 4262 of the human mitochondrial genome. Eight alterations were detected in this gene when its sequence was compared to that of the RCRS. Seven of the alterations have previously been described as synonymous polymorphisms. The nucleotide position 3594 is a haplogroup-determining site. Individuals having a T nucleotide at this position belong to haplogroup L0, L1 or L2, and individuals with a C nucleotide belong to haplogroup L3, M or N (Wallace, 2003). A single alteration, namely T3618C, is reported for the first time, as the alteration was not detected after an extensive search of the literature, and is discussed in Paragraph 4.2.4.1. The distribution of alterations between individuals is indicated in Table 4.4.

Table 4.4: Alterations detected in the ND1 gene

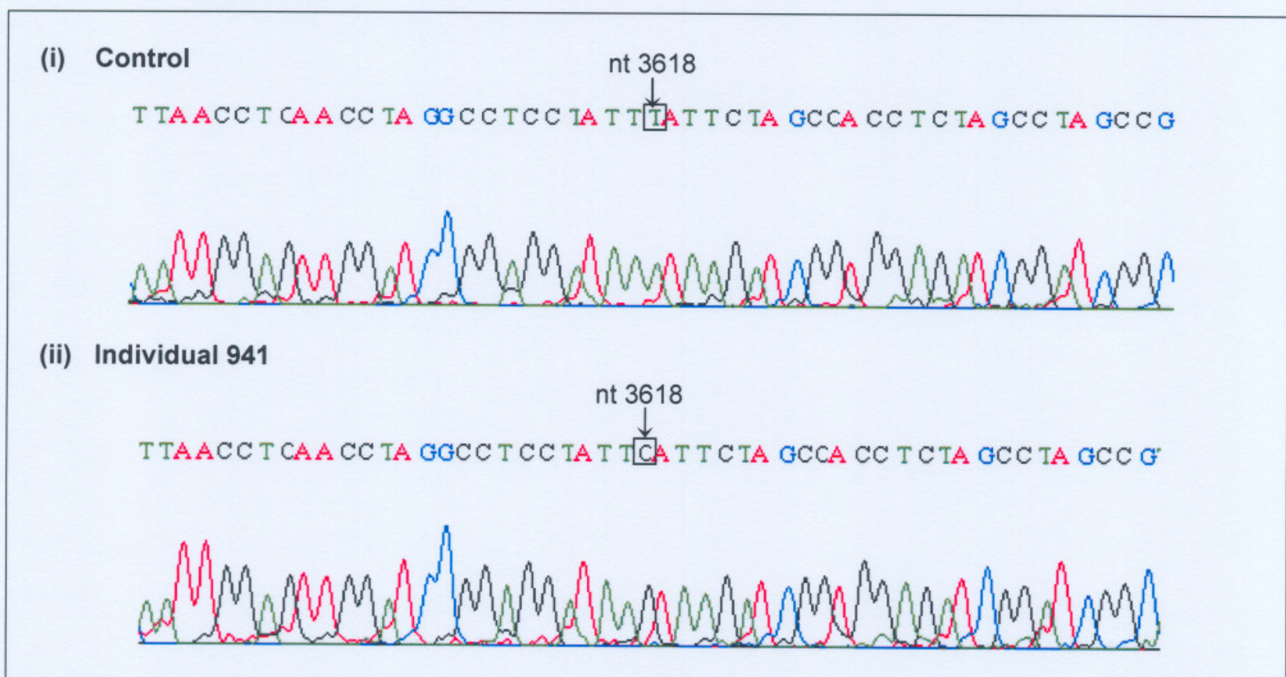
Alteration[1]	Individuals	Alteration Type	Reference
G3438A	941, 964, 972	Synonymous	Thomas *et al.* (1996)
C3516A	995, 941, 962, 964, 972	Synonymous	Herrnstadt *et al.* (2002)
C3594T	926, 939, 941, 962, 964, 972, 976	Synonymous	Chen *et al.* (1995)
T3618C	941, 972	Synonymous	Not yet reported
A3756G	941, 964, 972	Synonymous	Mishmar *et al.* (2003)
A4104G	926, 995, 939, 941, 962, 964, 972, 976	Synonymous	Ruppert *et al.* (2004)
A4158G	939	Synonymous	Herrnstadt *et al.* (2002)
T4232C	941, 964, 972	Ile → Thr	Mishmar *et al.* (2003)

1 = Alteration format includes initial nucleotide present in the RCRS, followed by the nucleotide position and then by the nucleotide detected. The RCRS was utilised as the reference sequence (Andrews *et al.*, 1999). The non-synonymous alteration is indicated by a three-letter amino acid notation. Ile = isoleucine, Thr = threonine. The arrow indicates the direction of amino acid change.

4.2.4.1 The T3618C alteration in the ND1 gene

The T3618C alteration is a transitional change as one purine, T nucleotide, is replaced by another purine, C nucleotide. The electropherogram presented in Figure 4.5 illustrates the T3618C alteration. This alteration occurs in a gene region that is translated, therefore a codon is altered. However, the alteration occurs in the third codon position, representing a synonymous change as both codons translate leucine due to the degeneracy of the mitochondrial genetic code (Barrel *et al.*, 1980). Since the codon alteration does not alter the amino acid, no functional effect is expected.

Figure 4.5: Representative electropherogram of the T3618C alteration in the ND1 gene



Control = representative electropherogram of an individual with identical sequence to the RCRS. Nucleotides enclosed in a box (□) represent novel alterations. Arrows indicate the nucleotide (nt) positions at which alterations occur.

4.2.5 The transfer RNA isoleucine gene

A single alteration, namely the C4312T transition, was detected in the tRNA isoleucine ($tRNA^{Ile}$) gene in individuals 941, 962, 964, 972. The $tRNA^{Ile}$ gene of the human mitochondrial genome includes nucleotide positions 4263 - 4331 (MITOMAP, 2004). This alteration represents a polymorphism that was first reported by Moraes *et al.* (1993). The frequency of this alteration is yet to be determined in the global population.

4.2.6 The transfer RNA glutamine gene

A single alteration was detected in this region, namely T4370C. This polymorphism was detected in individual 939. Herrnstadt *et al.* (2002) first reported this alteration and considered it to represent a polymorphism. The tRNA glutamine ($tRNA^{Gln}$) gene is 71 bp in length and is observed in the 4329 - 4400 position of the human mitochondrial genome (Anderson *et al.*, 1981).

4.2.7 NADH dehydrogenase subunit 2 gene

A total of nine alterations where detected when the NADH dehydrogenase subunit 2 (ND2) sequences were compared to the RCRS. The ND2 gene is 1041 bp in length, and is translated into a functional polypeptide (Wallace, 1992). The gene is located in the 4470 - 5511 position of the mitochondrial genome (MITOMAP, 2004). All alterations detected in the current investigation were previously described as polymorphisms and are presented in Table 4.5. The A4769G polymorphism (Andrews *et al.*, 1999) represents the most common allele observed at position 4769 during the reanalysis and revision of the CRS (Andrews *et al.*, 1999). The A at position 4769 detected in the RCRS represents a rare polymorphism.

Table 4.5: Alterations detected in the ND2 gene

Alteration[1]	Individuals	Alteration Type	Reference
T4586C	995, 962	Synonymous	Ingman *et al.* (2000)
A4767G	939	Trp → Val	Herrnstadt *et al.* (2002)
A4769G	921, 926, 995, 934, 939, 941, 962, 964, 972, 976	Synonymous	Andrews *et al.* (1999)
C5027T	939	Synonymous	Herrnstadt *et al.* (2002)
T5096C	962	Synonymous	Ingman *et al.* (2000)
G5231A	962	Synonymous	Ozawa *et al.* (1991a)
C5331A	939	Leu → Ile	Herrnstadt *et al.* (2002)
T5442C	941, 962, 964, 972	Phe → Leu	Ozawa *et al.* (1991a)
G5460A	962	Ala → Thr	Kosel *et al.* (1994)

1 = Alteration format includes initial nucleotide present in the RCRS, followed by the nucleotide position and then by the nucleotide detected. The RCRS was utilised as the reference sequence (Andrews *et al.*, 1999). Non-synonymous alterations are indicated by three-letter amino acid notations. Trp = tryptophan, Val = valine, Leu = leucine, Ile = isoleucine, Phe = phenylalanine, Ala = alanine and Thr = threonine. Arrows indicate the direction of amino acid change.

4.2.8 The transfer RNA tryptophan gene

The tRNA tryptophan ($tRNA^{Trp}$) gene is a 67 bp tRNA encoding gene positioned in nucleotides 5512 - 5579. A single alteration was detected in this region, namely T5553C.

This transitional alteration, detected in individual 964, represents a previously reported polymorphism (Herrnstadt *et al.*, 2002).

4.2.9 The transfer RNA alanine gene

The tRNA alanine (tRNA^{Ala}) gene is a 68 bp gene which spans the 5587 - 5655 region of the human mitochondrial genome. It encodes a functional tRNA molecule, which is not translated. A single alteration was detected in this region, namely C5603T, which is a reported transitional polymorphism (Moraes *et al.,* 1993). This was detected in a single individual, namely individual 962.

4.2.10 The transfer RNA cysteine gene

The tRNA cysteine (tRNA^{Cys}) gene is located in nucleotide positions 5761 - 5826. The gene encodes a functional tRNA which is not translated into an amino acid polypeptide. A single transversion in this region, namely T5814C, was detected in individual 939. The T5814C transition has previously been described as a polymorphism (Herrnstadt *et al.*, 2002).

4.2.11 Non-coding region between tRNA tyrosine and cytochrome *c* oxidase subunit I

The region between tRNA tyrosine (tRNA^{Tyr}) and COI, observed in nucleotide position 5892 - 5903, is an 11 bp intergenic non-coding region of the human mitochondrial genome (MITOMAP, 2004). Intergenic regions of the mitochondria do not code for protein products, and no direct functions have been assigned to these regions (Anderson *et al.*, 1981). It was suggested that these regions may contain control signals for transcription of the mitochondrial genome (Anderson *et al.*, 1981). However, alterations in these regions may have a functional effect if the changes affect upstream or downstream elements, affecting transcription or replication of the mitochondrial genome. These sites are generally more variable than coding regions, as they are not under direct selection. In the current investigation a single alteration was detected in this intergenic region. The A5894G transition was observed in individual 964 and represents a previously reported polymorphism (Herrnstadt *et al.*, 2002).

The prediction of a functional change as a result of this alteration is difficult. To date only one other alteration in this region, relative to the RCRS, has been described. The C

insertion in a pentanucleotide C repeat region located at position 5895 - 5899 was reported by Sternberg *et al.* (1998).

4.2.12 The cytochrome *c* oxidase subunit I gene

The COI gene encodes a subunit of the complex iv OXPHOS enzyme (Wallace, 1992). The COI gene is observed in the 5904 - 7445 region of the human mitochondrial genome. A total of 17 alterations were detected in this gene during this study, of which 15 have been reported as polymorphisms in the past. Two alterations are reported in the current investigation as novel, after an extensive literature search, in which no references to these alterations were detected. These changes, T6614C and A6806G, are presented in Paragraphs 4.2.12.1 and 4.2.12.2 respectively. All the alterations detected in this region are listed in Table 4.6.

Table 4.6: Alterations detected in the COI gene

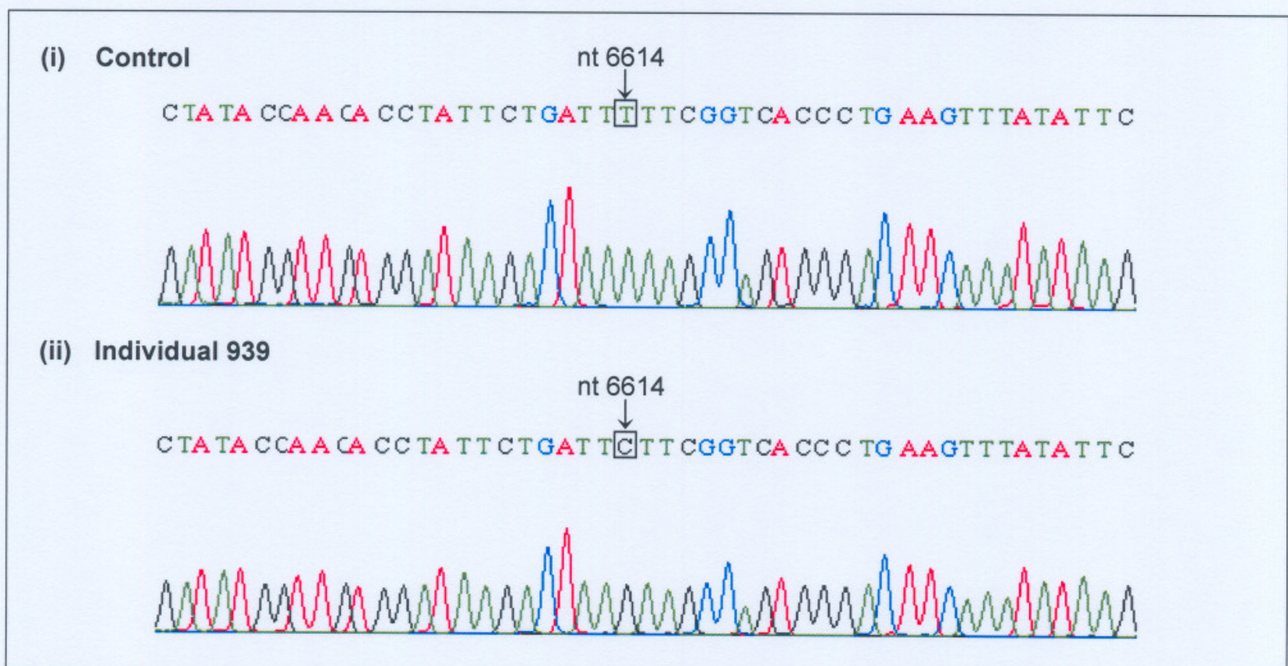
Alteration[1]	Individuals	Alteration Type	Reference
C5911T	962	Ala → Val	Ingman *et al.* (2000)
T6185C	995, 941, 962, 964, 972	Synonymous	Ingman *et al.* (2000)
T6221C	995, 934,	Synonymous	Marzuki *et al.* (1991)
A6266G	941, 964, 972	Synonymous	Herrnstadt *et al.* (2002)
C6587T	934, 941	Synonymous	Marzuki *et al.* (1991)
T6614C	939	Synonymous	Not yet reported
A6663G	976	Ile → Val	Scaglia *et al.* (2003)
C6713T	939	Synonymous	Herrnstadt *et al.* (2002)
A6806G	939	Synonymous	Not yet reported
T6815C	995, 964, 972	Synonymous	Mishmar *et al.* (2003)
A6891G	964	Ser → Gly	Hirakawa *et al.* (2002)
C7028T	921, 926, 995, 934, 939, 941, 962, 964, 972, 976	Synonymous	Macaulay *et al.* (1999)
A7146G	995, 941, 962, 964, 972	Synonymous	Polyak *et al.* (1999)
T7175C	926, 976	Synonymous	Silva *et al.* (2002)
C7256T	926, 939, 941, 962, 964, 972, 976	Synonymous	Silva *et al.* (2002)
C7274T	926, 976	Synonymous	Silva *et al.* (2002)
T7283C	941, 972	Synonymous	Finnila *et al.* (2001)

1 = Alteration format includes initial nucleotide present in the RCRS, followed by the nucleotide position and then by the nucleotide detected. The RCRS was utilised as the reference sequence (Andrews *et al.*, 1999). Non-synonymous alterations are indicated by three-letter amino acid notations. Ala = alanine, Val = valine and Ile = isoleucine, Ser = serine, Gly = glycine. Arrows indicate the direction of amino acid change.

4.2.12.1 The T6614C alteration in the COI gene

An electropherogram illustrating the T6614C alteration is presented in Figure 4.6. This represents a transitional nucleotide substitution located in the CO1 gene. This nucleotide is observed in the third codon site and alters the codon UUU, which codes for phenylalanine, to UUC, which also codes for phenylalanine (MITOMAP, 2004). There is thus no amino acid replacement. The T6614C transition is a synonymous alteration that has no functional effect on the resulting protein as no amino acid replacement occurs.

Figure 4.6: Representative electropherogram of the T6614C alteration in the COI gene

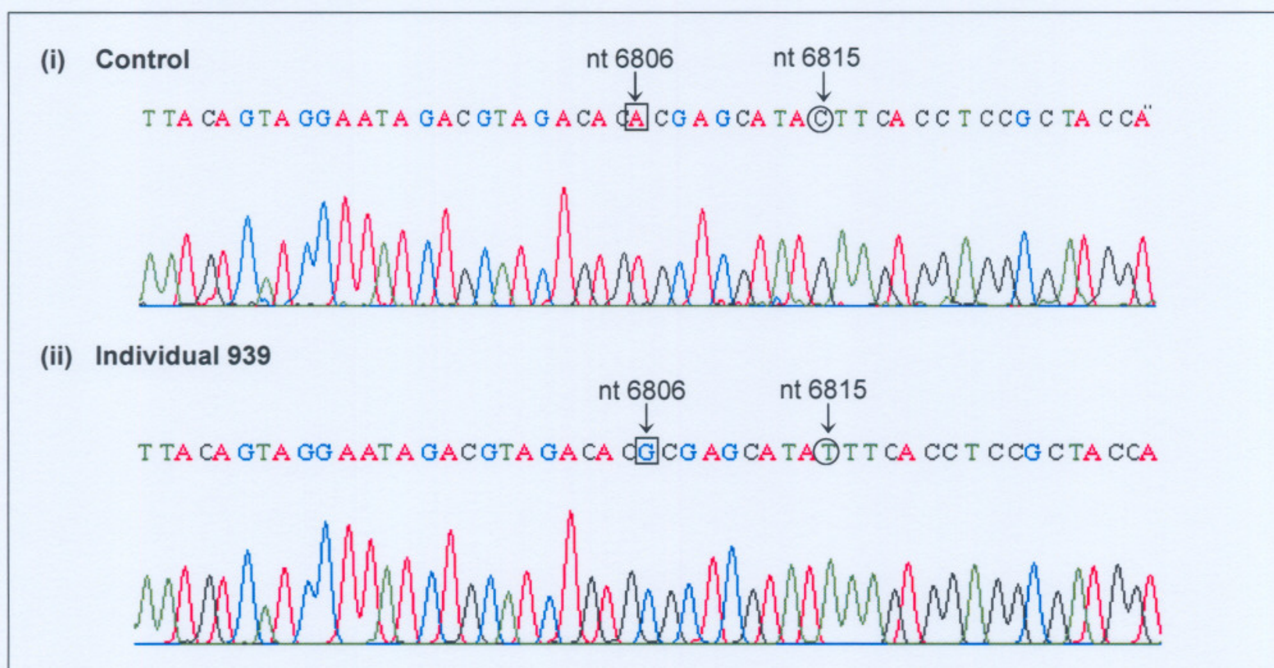


Control = representative electropherogram of an individual with identical sequence to the RCRS. Nucleotides enclosed in a box (□) represent novel alterations. Arrows indicate the nucleotide (nt) positions at which alterations occur.

4.2.12.2 The A6806G alteration in the COI gene

A second alteration, the A6806G transition, was detected in the COI gene. An electropherogram illustrating the alteration is presented in Figure 4.7, together with an illustration of the C6815T transition, previously reported by Mishmar *et al.* (2003). The A6806G alteration occurs in the third codon position (MITOMAP, 2004), altering the codon from ACA to ACG. However, this alteration is synonymous, as both codons code for threonine. Due to the absence of an amino acid replacement, no functional effect is predicted.

Figure 4.7: Representative electropherogram of the A6806G alteration in the COI gene



Control = representative electropherogram of an individual with identical sequence to the RCRS. Nucleotides enclosed in a box (□) represent novel alterations, and those enclosed in a circle (○) represent reported changes. Arrows indicate the nucleotide (nt) positions at which alterations occur.

4.2.13 The transfer RNA aspartic acid gene

A single transition, in the tRNA aspartic acid ($tRNA^{Asp}$) gene, namely G7521A, was detected in individuals 926, 939, 941, 962, 964, 972 and 976. This gene is 67 bp in length, and is located in the 7518 - 7585 position of the human mitochondrial genome (MITOMAP, 2004). This alteration has previously been reported by Silva *et al.* (2002) as a polymorphism.

4.2.14 The cytochrome *c* oxidase subunit II gene

The COII gene is located at position 7586 - 8269 of the human mitochondrial genome (Anderson *et al.*, 1981). This gene encodes a subunit of the OXPHOS complex iv enzyme (Wallace, 1992). A total of nine alterations were detected in this gene and are presented in Table 4.7. All changes detected in this region have previously been described as polymorphic.

Table 4.7: Alterations detected in the COII gene

Alteration[1]	Individuals	Alteration Type	Reference
T7624A	939	Synonymous	Herrnstadt *et al.* (2002)
T7741C	934	Synonymous	Coble *et al.* (2004)
A7771G	926, 976	Synonymous	Coble *et al.* (2004)
C8080T	939	Synonymous	Herrnstadt *et al.* (2002)
G8113A	941, 964, 972	Gly → Glu	Ruvolo *et al.* (1993)
G8152A	941, 964, 972	Synonymous	Ruvolo *et al.* (1993)
A8191G	962	Synonymous	Maca-Meyer *et al.* (2001)
G8206A	926, 939, 976	Synonymous	Ozawa *et al.* (1995)
G8251A	941, 964, 972	Synonymous	Ruvolo *et al.* (1993)

1 = Alteration format includes initial nucleotide present in the RCRS, followed by the nucleotide position and then by the nucleotide detected. The RCRS was utilised as the reference sequence (Andrews *et al.*, 1999). The non-synonymous alteration is indicated by a three-letter amino acid notation. Gly = glycine and Glu = glutamic acid. The arrow indicates the direction of amino acid change.

4.2.15 The ATP synthase F0 subunit 8 gene

The ATPase8 gene of the mitochondrial genome is located in the 8366 - 8572 position (Anderson *et al.*, 1981). ATPase8 encodes a subunit of the complex v enzyme of the OXPHOS system (Wallace, 1992). In total, seven alterations were detected in this gene and are presented in Table 4.8. All but one of the detected alterations are previously reported polymorphisms. The C8503T alteration, as discussed in Paragraph 4.2.15.1, is considered novel, as no reference to the alteration was found during an extensive search of the literature.

Table 4.8: Alterations detected in the ATPase8 gene

Alteration[1]	Individuals	Alteration Type	Reference
T8383C	972	Synonymous	Herrnstadt *et al.* (2002)
G8387A	939	Synonymous	Herrnstadt *et al.* (2002)
C8428T	962	Synonymous	Moraes *et al.*(1993)
A8460G	995	Asn → Ser	Ingman *et al.* (2000)
C8468T	995, 941, 962, 964, 972	Synonymous	Moraes *et al.*(1993)
T8503C	939, 964	Synonymous	Not yet reported
A8566G	962	Synonymous	Ingman *et al.* (2000)

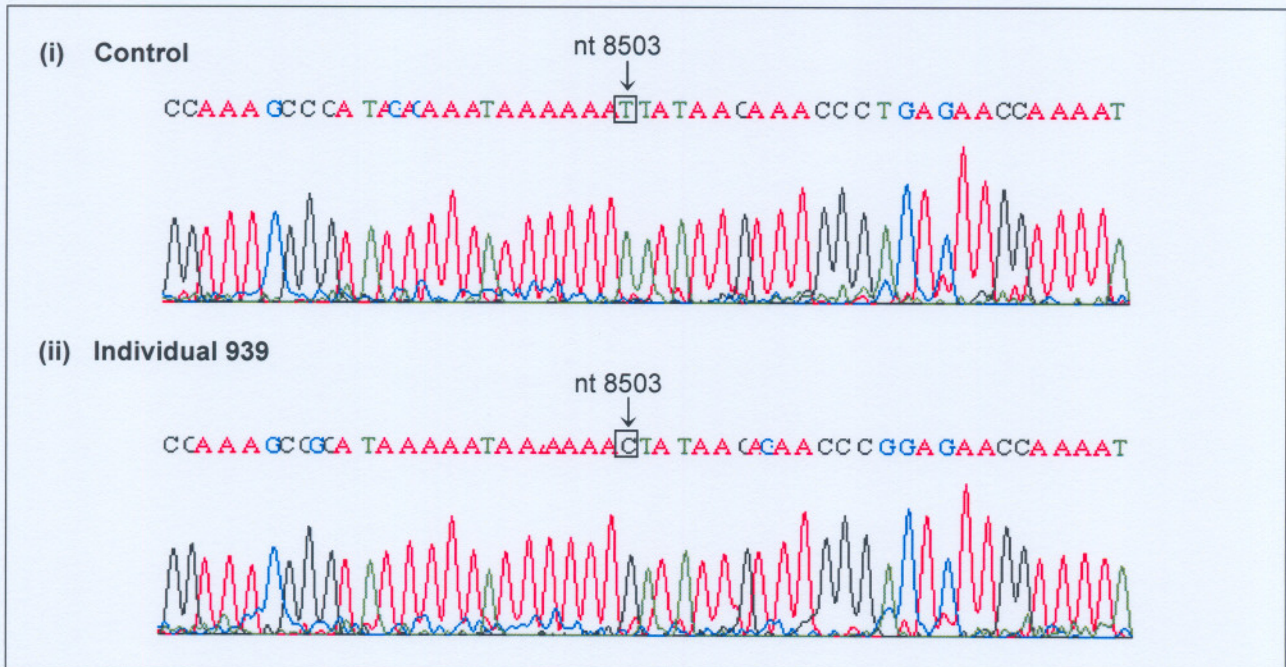
1 = Alteration format includes initial nucleotide present in the RCRS, followed by the nucleotide position and then by the nucleotide detected. The RCRS was utilised as the reference sequence (Andrews *et al.*, 1999). Non-synonymous alterations are indicated by three letter amino acid notations. Asn = asparagine and Ser = serine. Arrows indicate the direction of amino acid change.

4.2.15.1 The T8503C alteration in the ATPase8 gene

The novel T8503C transition is illustrated with an electropherogram presented in Figure 4.8. This transition is located in a third codon position, and the codon is altered from

AAU to AAC. The T8503C transition is a synonymous alteration, as both these codons code for asparagine. No functional alteration is therefore predicted to occur in the polypeptide products of this gene, due to this transition.

Figure 4.8: Representative electropherogram of the T8503C alteration in the ATPase8 gene



Control = representative electropherogram of an individual with identical sequence to the RCRS. Nucleotides enclosed in a box (☐) represent novel alterations. Arrows indicate the nucleotide (nt) positions at which alterations occur.

4.2.16 The A8566G alteration in the ATPase8 and ATPase6 genes

The A8566G alteration reported by Ingman *et al.* (2000) is located in the ATPase8 and ATPase6 genes, and was observed in individual 962. The genes have an overlap of 45 bp from position 8527 to 8572, and are translated in distinct reading frames (Anderson *et al.*, 1981). The A8566G alteration is located in this overlap, therefore resulting in the alteration being detected in both genes. This change results in a synonymous alteration in the ATPase8 gene and a non-synonymous Ile → Val alteration in the ATPase6 gene, as presented in Table 4.9.

4.2.17 The ATP synthase F0 subunit 6 gene

The 680 bp ATPase6 gene is located in the 8527 - 9207 position of the human mitochondrial genome (Anderson *et al.*, 1981). The gene is translated into a polypeptide that forms part of the complex v enzyme of the OXPHOS system (Wallace, 1992). In the current study 10 previously reported polymorphisms were detected in this region and are

presented in Table 4.9. The A8860G alteration, detected in all individuals analysed represents the most common alteration at that site, as an A at that site represents a rare polymorphism in the RCRS. The A8566G alteration is located in the ATPase8 and ATPase6 genes and is discussed in Paragraph 4.2.16.

Table 4.9: Alterations detected in the ATPase6 gene

Alteration[1]	Individuals	Alteration Type	Reference
A8566G	962	Ile → Val	Ingman *et al.* (2000)
A8577G	995	Synonymous	Abu-Amera *et al.* (2004)
C8655T	941, 962, 964, 972	Synonymous	Ingman *et al.* (2000)
A8701G	921, 926, 995, 934, 939, 941, 962, 964, 972, 976	Thr → Arg	Ozawa *et al.* (1991a)
C8703T	934	Synonymous	Ingman *et al.* (2000)
G8790A	939	Synonymous	Taylor *et al.* (2003)
T8793C	964	Synonymous	Ozawa *et al.* (1991b)
A8860G	921, 926, 995, 934, 939, 941, 962, 964, 972, 976	Thr → Arg	Andrews *et al.* (1999)
C9042T	995, 941, 962, 964, 972	Synonymous	Ingman *et al.* (2000)
T9111C	941	Synonymous	Herrnstadt *et al.* (2002)

1 = Alteration format includes initial nucleotide present in the RCRS, followed by the nucleotide position and then by the nucleotide detected. The RCRS was utilised as the reference sequence (Andrews *et al.*, 1999). Non-synonymous alterations are indicated by three-letter amino acid notations. Ile = isoleucine, Val = valine, Thr = threonine and Arg = arginine. Arrows indicate the direction of amino acid change.

4.2.18 The cytochrome *c* oxidase subunit III gene

A total of 12 alterations were detected in the COIII gene, as listed in Table 4.10 The human mitochondrial COIII gene is a 783 bp polypeptide encoding gene that is located in the 9207 - 9990 region (Anderson *et al.*, 1981). Ten alterations are previously reported polymorphisms and two, namely A9350G and T9637G, have to date not been described and are presented in Paragraph 4.2.18.1 and 4.2.18.2.

Table 4.10: Alterations detected in the COIII gene

Alteration[1]	Individuals	Alteration Type	Reference
A9221G	926, 939, 976	Synonymous	Silva *et al.* (2002)
G9300A	934	Synonymous	Snejina *et al.* (2003)
A9327G	934	Synonymous	Coble *et al.* (2004)
A9347G	995, 941, 962, 964, 972	Synonymous	Herrnstadt *et al.* (2002)
A9350G	939	Synonymous	Not yet reported
A9377G	921	Synonymous	Ozawa *et al.* (1991b)

Table 4.10: Continued...

Alteration[1]	Individuals	Alteration Type	Reference
T9540C	921, 926, 995, 934, 939, 941, 962, 964, 972, 976	Synonymous	Marzuki *et al.* (1991)
T9637G	962, 964, 972	Ile → Ser	Not yet reported
G9755A	941, 962, 964, 972	Synonymous	Herrnstadt *et al.* (2002)
C9818T	962	Synonymous	Herrnstadt *et al.* (2002)
T9833C	964	Synonymous	Kong *et al.* (2003)
T9950C	964	Synonymous	Silva *et al.* (2003)

1 = Alteration format includes initial nucleotide present in the RCRS, followed by the nucleotide position and then by the nucleotide detected. The RCRS was utilised as the reference sequence (Andrews *et al.*, 1999). Non-synonymous alterations are indicated by three-letter amino acid notations. Ile = isoleucine, Ser = serine. Arrows indicate the direction of amino acid change.

4.2.18.1 The A9350G alteration in the COIII gene

The A9350G transition is presented for the first time in the current investigation, as it was not detected subsequent to a broad literature search. The alteration occurs in the third codon position (MITOMAP, 2004) of the COIII gene and alters the codon sequence from CUA to CUG. However, both codons result in the translation of the amino acid leucine, and A9350G therefore results in a synonymous alteration. The alteration was detected in individual 939 and an electropherogram illustrating the alteration is presented in Figure 4.9.

Figure 4.9: Representative electropherogram of the A9350G alteration in the COIII gene

(i) Control

(ii) Individual 939

Control = representative electropherogram of an individual with identical sequence to the RCRS. Nucleotides enclosed in a box (□) represent novel alterations, and those enclosed in a circle (○) represent reported changes. Arrows indicate the nucleotide (nt) positions at which alterations occur.

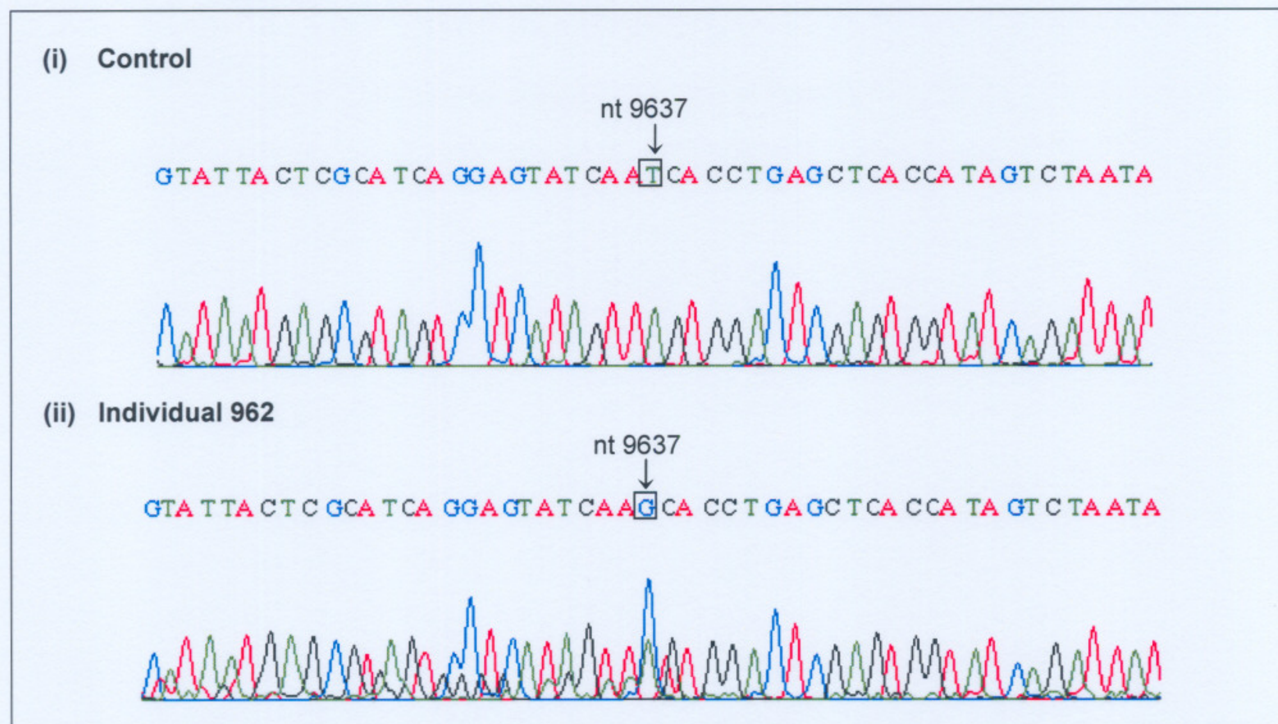
No functional effect is predicted as a result of the alteration, as no amino acid replacement occurs. The G9327A alteration is included in the illustration presented in Figure 4.9. The high level of background observed in the electropherogram illustrating individual 939, relative to the control individual, does not affect the final result, as discussed in Paragraph 4.2.2.1.

4.2.18.2 The T9637G alteration in the COIII gene

The T9637G alteration, detected in individuals 962, 964 and 972, is reported for the first time in the current study, as it was not detected in the literature. However, upon investigation of the electropherogram, which is illustrated in Figure 4.10, it seemed as if a level of heteroplasmy was present, due to a second, smaller overlapping peak that represents the nucleotide represented by the RCRS. Although this alteration has not yet been reported, it would be essential to confirm heteroplasmy, as described in Paragraph 4.2.28, before novel status is awarded to this alteration. Detection of heteroplasmy would then exclude the possibility of the observed double peaks being a result of an artefact produced during the sequencing reaction. It is also essential to sequence this alteration in the reverse direction in future studies in order to confirm the results.

This alteration occurred in the second codon position, and is altered from AUC, which codes for isoleucine, to AGC, coding for serine, and is therefore a non-synonymous change. The prediction of a functional effect as a result of this alteration is difficult. Isoleucine has a non-polar side chain and serine has a polar, neutral side chain (Campbell, 1993). These two amino acids differ in chemical properties, and may therefore result in the stereochemical disruption of the resulting protein, which may have a functional effect. A non-synonymous alteration located 18 nucleotides upstream of the A9637G alteration detected in the current study, namely G9655A, occurs in a conserved region and is predicted as unlikely to have a functional effect (Clark *et al.*, 1999). The A9637G alteration of the current study occurs in close proximity of the reported G9655A change and may therefore occur in the same protein domain. It is hypothesised that non-synonymous alterations in the COIII gene may be a result of selection in order to adapt to specific environments (Mishmar *et al.*, 2003), as discussed in Paragraph 4.2.26.1.

Figure 4.10: Representative electropherogram of the T9637G alteration in the COIII gene



Control = representative electropherogram of an individual with identical sequence to the RCRS. Nucleotides enclosed in a box (□) represent novel alterations. Arrows indicate the nucleotide (nt) positions at which alterations occur.

4.2.19 The NADH dehydrogenase subunit 3 gene

The NADH dehydrogenase subunit 3 (ND3) gene is located in the 10059 - 10404 position of the mitochondrial genome (Anderson *et al.*, 1981). In the current investigation, three alterations were detected in this gene when sequences were compared to the RCRS. The T10115C transition was detected in individuals 926, 939, 941 and 976, and was previously reported as a synonymous polymorphism (Silva *et al.*, 2003). The second alteration, namely G10143A, was detected in individual 926 and was described by Prezant *et al.* (1993) as a non-synonymous polymorphism resulting in an amino acid alteration from glycine to serine. The overall frequency of this alteration is still unknown. The A10398G non-synonymous alteration from methionine to threonine, was detected in all 10 individuals analysed in the current study, and has been reported previously as a polymorphism by Herrnstadt *et al.* (2002).

4.2.20 The NADH dehydrogenase subunit 4L gene

The NADH dehydrogenase subunit 4L (ND4L) gene is translated into a functional polypeptide that forms part of the complex i subunit of the OXPHOS system (Wallace, 1992). This gene is located at the 10470 - 10766 position of the human mitochondrial

genome. Three previously reported polymorphisms were detected in this region and are presented in Table 4.11.

Table 4.11: Alterations detected in the ND4L gene

Alteration[1]	Individuals	Alteration Type	Reference
G10589A	941, 962, 964, 972	Synonymous	Howell *et al.* (1995)
C10664T	995, 941, 962, 964, 972	Synonymous	Herrnstadt *et al.* (2002)
G10688C	941, 962, 964, 972	Synonymous	Ingman *et al.* (2000)

1 = Alteration format includes initial nucleotide present in the RCRS, followed by the nucleotide position and then by the nucleotide detected. The RCRS was utilised as the reference sequence (Andrews *et al.*, 1999).

4.2.21 The NADH dehydrogenase subunit 4 gene

The human mitochondrial NADH dehydrogenase subunit 4 (ND4) gene is located in the 10760 - 12137 position, and encodes a polypeptide of the OXPHOS complex i subunit (Wallace, 1992). A total of 14 alterations, presented in Table 4.12, were detected in this region. All alterations are considered as polymorphisms and have previously been reported.

Table 4.12: Alterations detected in the ND4 gene

Alteration[1]	Individuals	Alteration Type	Reference
T10810C	941, 962, 964, 972	Synonymous	Finnila *et al.* (2001)
A10819G	921, 995, 934	Synonymous	Marzuki *et al.* (1991)
T10873C	921, 926, 995, 934, 939, 941, 962, 964, 972, 976	Synonymous	Marzuki *et al.* (1991)
T10915C	995, 941, 962, 964, 972	Synonymous	Ozawa *et al.* (1991a)
C10920T	941	Pro → Leu	Ingman *et al.* (2000)
G11176A	995, 941, 962	Synonymous	Herrnstadt *et al.* (2002)
T11260C	964	Synonymous	Herrnstadt *et al.* (2002)
A11641G	995, 962	Synonymous	Ingman *et al.* (2000)
G11719A	921, 926, 995, 934, 939, 941, 962, 964, 972, 976	Synonymous	Silva *et al.* (2001)
G11914A	921, 941, 962, 964, 972, 976	Synonymous	Silva *et al.* (2001)
T11944C	926, 939, 976	Synonymous	Ingman *et al.* (2000)
G12007A	941, 962, 964, 972	Synonymous	Moraes *et al.* (1993)
T12121C	941, 964, 972	Synonymous	Mishmar *et al., al.* (2003)
G12127A	962	Synonymous	Herrnstadt *et al.* (2002)

1 = Alteration format includes initial nucleotide present in the RCRS, followed by the nucleotide position and then by the nucleotide detected. The RCRS was utilised as the reference sequence (Andrews *et al.*, 1999). Non-synonymous alterations are indicated by three-letter amino acid notations. Pro = praline and leu = leucine. Arrows indicate the direction of amino acid change.

4.2.22 The transfer RNA histidine gene

The transfer RNA histidine ($tRNA^{His}$) gene codes for a 68 bp functional tRNA molecule and is located in the nucleotide position the 12138 - 12206 position of the human mitochondrial genome (Anderson *et al.*, 1981). A single alteration, A12142G, was detected in this gene in individual 964. The A12142G transition has previously been reported by Mouton (2003).

4.2.23 The transfer RNA serine gene

The tRNA serine ($tRNA^{Ser}$) gene occurs in the 12266 - 12336 position of the mitochondrial genome (Anderson *et al.*, 1981). A single alteration, namely G12236A, was observed in this region and represents a previously reported polymorphism (Prezant *et al.*, 1993). This transition was detected in a single individual, namely 939.

4.2.24 The NADH dehydrogenase subunit 5 gene

The NADH dehydrogenase subunit 5 (ND5) gene is 1,811 bp in size (Anderson *et al.*, 1981) and encodes a complex i subunit of the OXPHOS system (Wallace *et al.*, 1992). The ND5 gene is located in the 12337 - 14148 region of the human mitochondrial genome (Anderson *et al.*, 1981). In total, 20 alterations were detected in this region and are presented in Table 4.13. Seventeen of these alterations are previously reported polymorphisms. Three alterations are considered to be novel, namely C12348T, C12436T and C12798T, and are presented in Paragraphs 4.2.24.1 to 4.2.24.3.

Table 4.13: Alterations detected in the ND5 gene

Alteration[1]	Individuals	Alteration Type	Reference
C12348T	964	Synonymous	Not yet reported
G12406A	921	Val → Ile	Herrnstadt *et al.* (2002)
C12436T	941	His → Asn	Not yet reported
A12693G	926, 976	Synonymous	Silva *et al.* (2002)
C12705T	921, 926, 995, 934, 939, 941, 962, 964, 972, 976	Synonymous	Marzuki *et al.* (1991)
A12720G	995, 962, 964, 972	Synonymous	Brown *et al.* (2001)
T12738C	934	Synonymous	Mishmar *et al.* (2003)
C12798T	962	Synonymous	Not yet reported
A12948G	939	Synonymous	Herrnstadt *et al.* (2002)
A13105G	995, 941, 962, 964, 972	Ile → Val	Prezant *et al.* (1993)
A13276G	941, 962, 964, 972	Met → Val	Ingman *et al.* (2000)
C13506T	995, 941, 962, 964, 972	Synonymous	Herrnstadt *et al.* (2002)

Table 4.13: Continued...

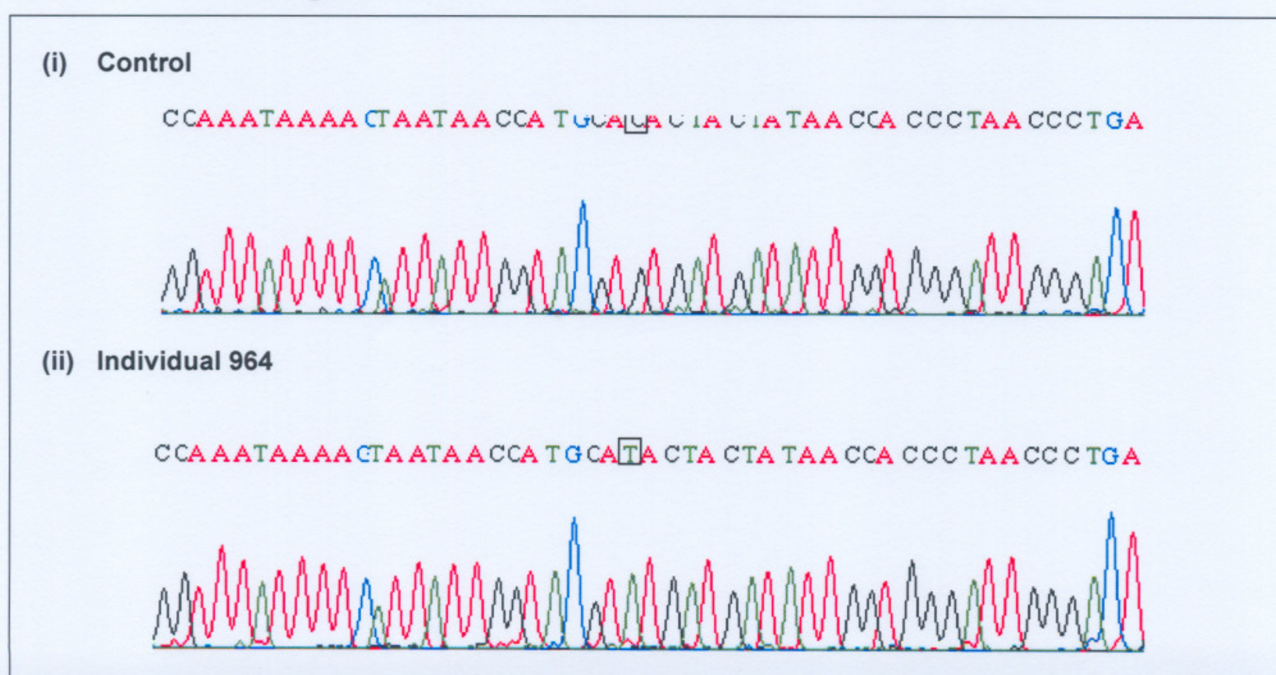
Alteration[1]	Individuals	Alteration Type	Reference
G13590A	926, 939, 976	Synonymous	Ozawa *et al.* (1995)
C13650T	926, 995, 939, 941, 962, 964, 972, 976	Synonymous	Silva *et al.* (2002)
G13708A	976	Ala → Thr	Torroni *et al.* (1996)
G13759A	941, 964, 972	Ala → Thr	Rieder *et al.* (1998)
A13803G	926, 976	Synonymous	Silva *et al.* (2002)
A13966G	939	Thr → Ala	Finnila *et al.* (2001)
A14007G	962	Synonymous	Ingman *et al.* (2000)
A14059G	939	Ile →Val	Mishmar *et al.* (2003)

1 = Alteration format includes initial nucleotide present in the RCRS, followed by the nucleotide position and then by the nucleotide detected. The RCRS was utilised as the reference sequence (Andrews *et al.*, 1999). Non-synonymous alterations are indicated by three-letter amino acid notations. His = histidine, Asn = asparagines, Ile = isoleucine, Val = valine, Met = methionine, Ala = alanine and Thr = threonine. Arrows indicate the direction of amino acid change.

4.2.24.1 The C12348T alteration in the ND5 gene

The C12348T alteration is a transition located in the ND5 gene and was detected in individual 964. An electropherogram illustrating this alteration is presented in Figure 4.11. No report of this transition could be found, and the C12348T transition was thus considered to be novel. The ND5 gene is translated into an amino acid polypeptide (Wallace, 1992). The alteration occurs in the third codon position, altering the codon from CAC to CAU (MITOMAP, 2004). However, both codons code for histidine, and the transition therefore represents a synonymous alteration.

Figure 4.11: Representative electropherogram of the C12348T alteration in the ND5 gene

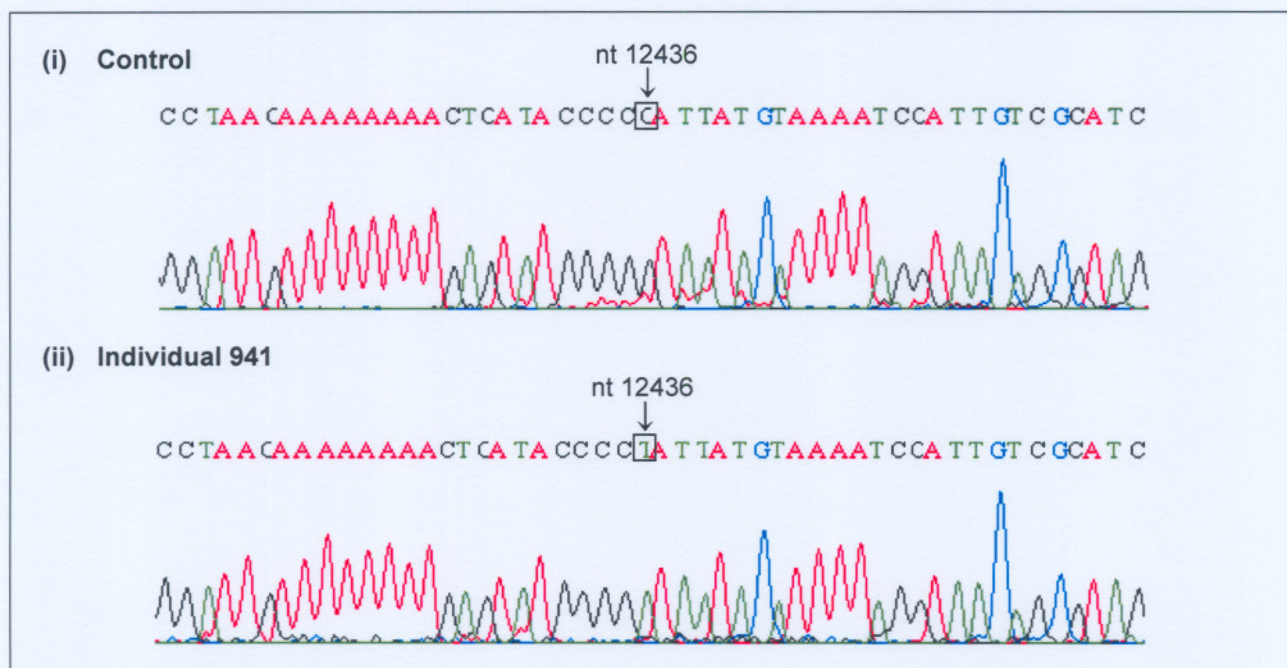


Control = representative electropherogram of an individual with identical sequence to the RCRS. Nucleotides enclosed in a box (□) represent novel alterations. Arrows indicate the nucleotide (nt) positions at which alterations occur.

4.2.24.2 The C12436T alteration in the ND5 gene

The C12436T transition was detected in the ND5 gene and Figure 4.12 presents an illustration of this alteration. This transition has, to date, not been reported. The ND5 gene is translated, therefore the codon is altered as a result of this alteration (Wallace, 1992). The C12436T transition, observed in individual 941, is located in the first codon position and alters the codon from CAU, which codes for histidine, to AAU, which codes for asparagine (MITOMAP, 2004). Histidine possesses a bulky and rigid ring structure side chain with basic properties, whereas asparagine possesses a polar uncharged and unbranched amino acid. The alteration in amino acid may affect the structure of the resulting protein and may result in a functional effect (Campbell, 1995).

Figure 4.12: Representative electropherogram of the C12436T alteration in the ND5 gene



Control = representative electropherogram of an individual with identical sequence to the RCRS. Nucleotides enclosed in a box (□) represent novel alterations. Arrows indicate the nucleotide (nt) positions at which alterations occur.

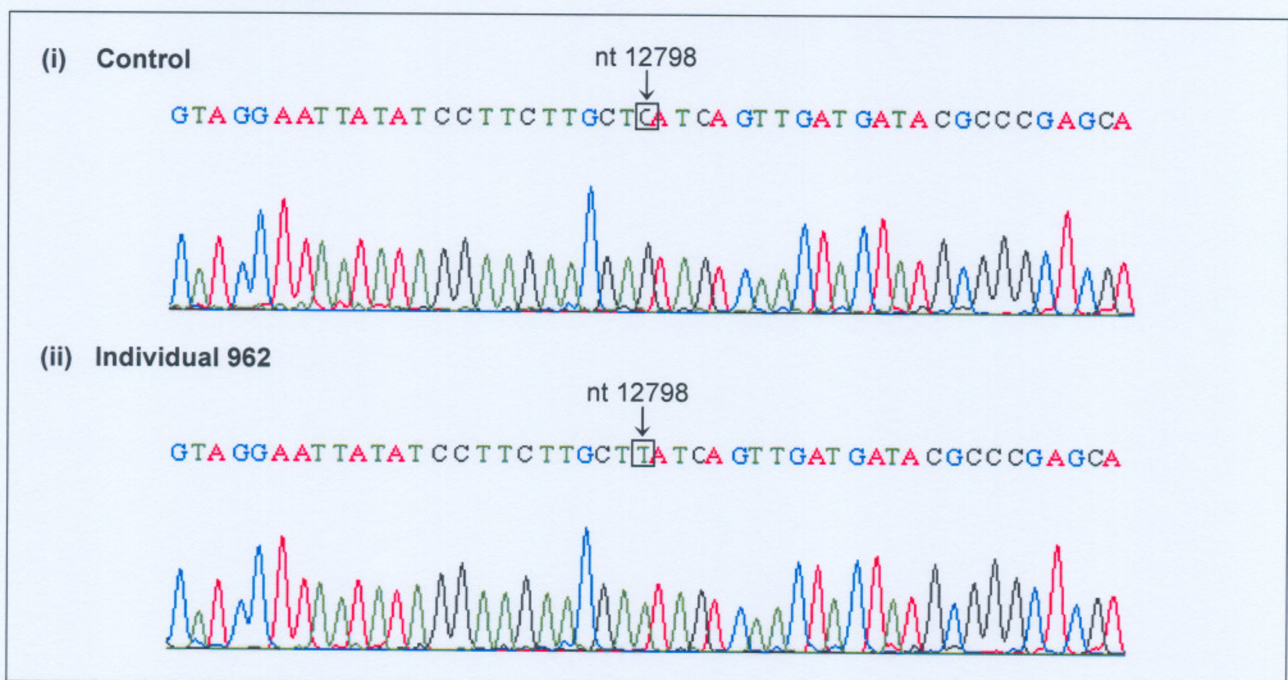
The prediction of a functional effect can be achieved through comparing the functional effect of a similar alteration in close proximity to the reported change. These reported alterations may affect similar domains of the protein, and may result in similar functional changes. An A12418C non-synonymous alteration (Silva *et al.*, 2003) has previously been reported in an individual and is located 18 nucleotides downstream of the C12436T alteration. In the A12418C alteration, lysine, a positively charged basic amino acid, is replaced with glutamine, a polar, uncharged amino acid. These two amino acids differ in stereo-chemical properties. The alteration may therefore affect the resulting protein.

However, the individuals in which this alteration was observed were not reported as harbouring any mitochondrial disorders (Silva *et al.*, 2002). The alteration is also present in the 12300 - 12900 region, which corresponds to the N-terminal region of the ND5 subunit. This region contains a significantly larger number of variants, indicating a lack of functional constraint (Marzuki *et al.*, 1991). Therefore, the prediction of a lack of functional effect associated with the C12436T alteration is supported.

4.2.24.3 The C12798T alteration in the ND5 gene

A transitional change, namely C12798T, was detected in the ND5 gene and is presented in Figure 4.13. This transition occurs in the third codon position of CUC (MITOMAP, 2004) which codes for leucine. The codon is changed from CUC to CUA, which codes for leucine as well. The transition thus results in a synonymous alteration, not affecting the amino acid composition of the polypeptide. This alteration was detected only in individual 962 and is presented for the first time in the current investigation, as no reference to the alteration could be detected.

Figure 4.13: Representative electropherogram of the C12798T alteration in the ND5 gene



Control = representative electropherogram of an individual with identical sequence to the RCRS. Nucleotides enclosed in a box (□) represent novel alterations. Arrows indicate the nucleotide (nt) positions at which alterations occur.

4.2.25 The NADH dehydrogenase subunit 6 gene

The NADH dehydrogenase subunit 6 (ND6) gene of the human mitochondrial genome is located in the 14149 - 14673 nucleotide position. The gene is translated into a subunit utilised by complex i of the OXPHOS system (Wallace, 1992). Nine alterations were detected in this gene and are presented in Table 4.14. Seven of the alterations are previously described polymorphisms. Two alterations are described as novel, as no references to these alterations were identified during a broad literature search. The novel A14176G and C14407T transitions are discussed in Paragraphs 4.2.25.1 and 4.2.25.2 respectively.

Table 4.14: Alterations detected in the ND6 gene

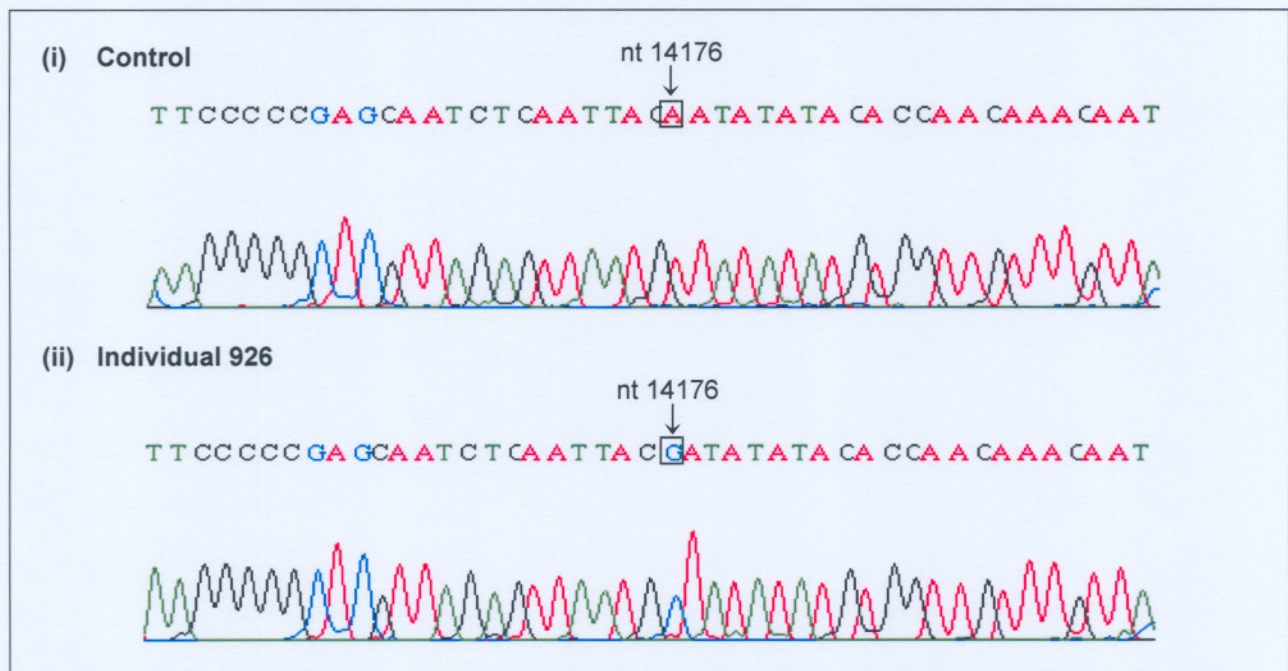
Alteration[1]	Individuals	Alteration Type	Reference
A14152G	995, 934	Synonymous	Marzuki *et al.* (1991)
A14176G	926	Synonymous	Not yet reported
T14182C	995	Synonymous	Huoponen *et al.* (1993)
T14212C	921, 995, 934	Synonymous	Marzuki *et al.* (1991)
T14308C	995, 962	Synonymous	Sudoyo *et al.* (2002)
C14315T	941, 972	Thr →Ile	Mouton (2003)
C14407T	939	Synonymous	Not yet reported
A14566G	926, 976	Synonymous	Silva *et al.* (2003)
C14659T	941, 972	Leu → Phe	Mouton (2003)

1 = Alteration format includes initial nucleotide present in the RCRS, followed by the nucleotide position and then by the nucleotide detected. The RCRS was utilised as the reference sequence (Andrews *et al.*, 1999). The non-synonymous alterations are indicated by three-letter amino acid notations. Thr = threonine, Ile = isoleucine, Leu = leucine, Phe = phenylalanine. The arrows indicate the direction of amino acid change.

4.2.25.1 The A14176G alteration in the ND6 gene

The A14176G alteration represents a transition located in the ND6 gene (Anderson *et al.*, 1981). This alteration has to date not been described and is illustrated in Figure 4.14. The ND6 gene is transcribed from the reverse strand of the mitochondrial genome. The transition thus occurs in the third codon position, altering the codon from AUU to AUC (MITOMAP, 2004). This alteration is synonymous, as both codons code for isoleuine, and was detected in individual 926.

Figure 4.14: Representative electropherogram of the A14176G alteration in the ND6 gene

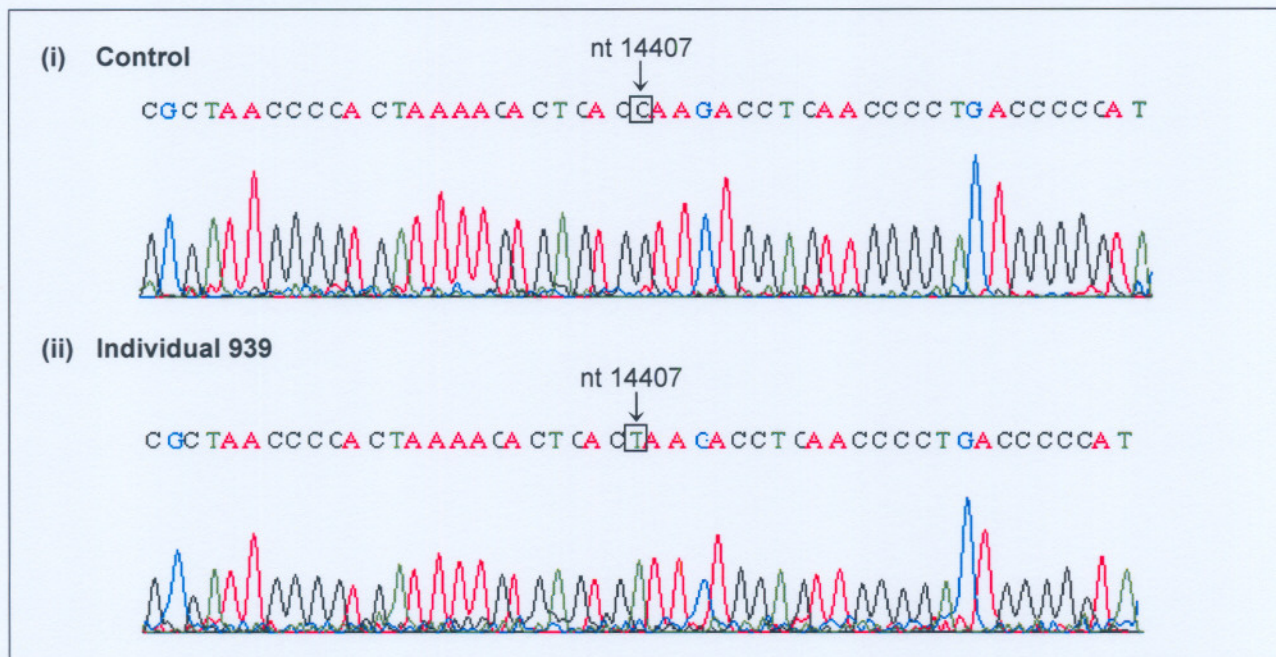


Control = representative electropherogram of an individual with identical sequence to the RCRS. Nucleotides enclosed in a box (□) represent novel alterations. Arrows indicate the nucleotide (nt) positions at which alterations occur.

4.2.25.2 The C14407T alteration in the ND6 gene

The C14407T alteration, detected in individual 939, is reported as novel in the current investigation and an illustration thereof is presented in Figure 4.15. The alteration occurs in a protein-coding region, which is transcribed from the reverse strand. The alteration represents a G to A transition occurring in a third codon position of the reverse strand. The codon is altered from UUG to UUA. As both codons code for the amino acid leucine, the transition is considered synonymous.

Figure 4.15: Representative electropherogram of the C14407T alteration in the ND5 gene



Control = representative electropherogram of an individual with identical sequence to the RCRS. Nucleotides enclosed in a box (□) represent novel alterations. Arrows indicate the nucleotide (nt) positions at which alterations occur.

4.2.26 The cytochrome *b* gene

The cyt *b* gene occurs in the 14747 to 15887 position of the mitochondrial genome (Anderson *et al.*, 1981). A total of 16 alterations were detected in this region when the individual's sequences were compared to the RCRS, and are presented in Table 4.15. Fifteen of the alterations have previously been described as polymorphisms. The A15692G alteration was detected in individual 972, and is presented for the first time in Paragraph 4.2.26.1. This alteration is considered novel, as the alteration was not detected in the literature.

The alterations C14766T and A15326G were both detected in all individuals analysed and represent an error and a rare polymorphism respectively in the RCRS. At nucleotide position 14766 the RCRS (Andrews *et al.*, 1999) has a C nucleotide, whereas the CRS (Anderson *et al.*, 1981) contained a T nucleotide. The CRS was altered as it represents the sequence of HeLa cells, which are derived from an individual of African origin (Anderson *et al.*, 1981). Therefore, it is not surprising that the C14766T alteration was detected, as the individuals of the current study are also of African origin. An A at nucleotide position 15326 represents a rare polymorphism in the RCRS, and the A15326G alteration represents the most common allele observed at that position.

Table 4.15: Alterations detected in the cyt *b* gene

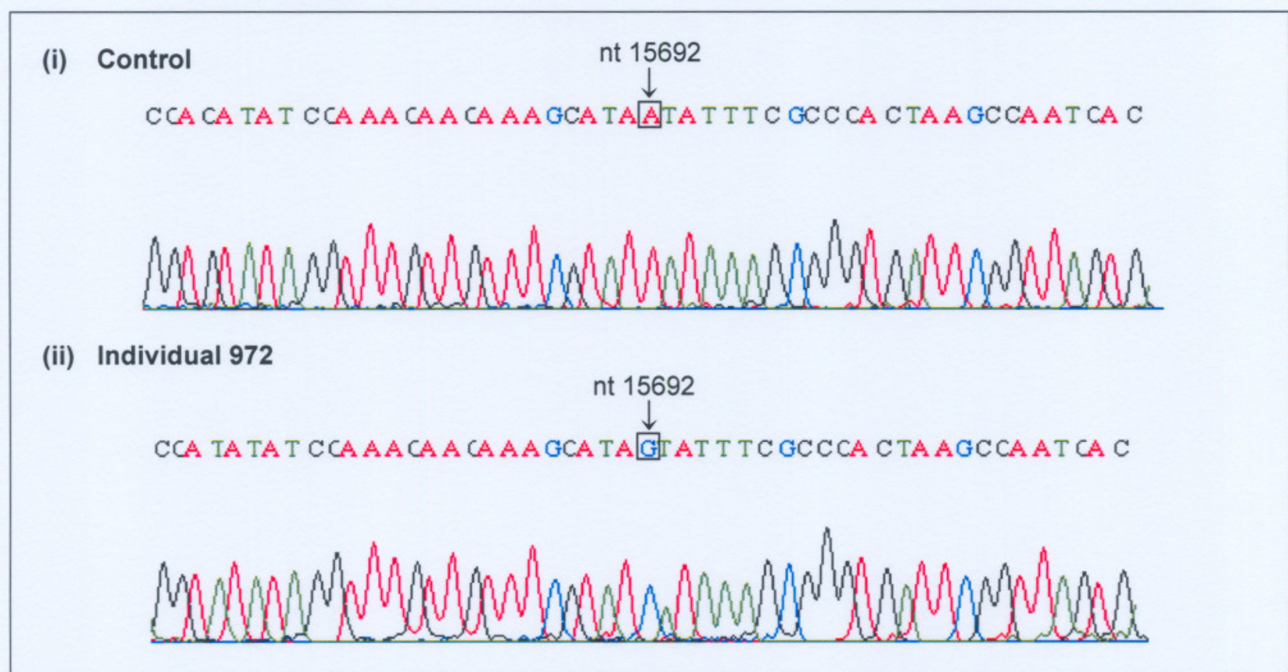
Alteration[1]	Individuals	Alteration Type	Reference
A14755G	995	Synonymous	Herrnstadt *et al.* (2002)
C14766T	921, 926, 995, 934, 939, 941, 962, 964, 972, 976	Synonymous	Andrews *et al.* (1999)
G14905A	921	Synonymous	Howell *et al.* (1995)
A14926G	995	Synonymous	Crimi *et al.* (2002)
G15110A	939	Ala → Thr	Legros *et al.* (2001)
C15136T	962	Synonymous	Ingman *et al.* (2000)
G15301A	921, 926, 934, 939, 976	Synonymous	Ozawa *et al.* (1991a)
A15326G	921, 926, 995, 934, 939, 941, 962, 964, 972, 976	Thr → Ala	Andrews *et al.* (1999)
C15337T	941	Synonymous	Finnila *et al.* (2001)
G15431A	962	Ala → Thr	Ingman *et al.* (2000)
G15466A	941, 964, 972	Synonymous	Mishmar *et al.* (2003)
T15670C	934	Synonymous	Valnot *et al.* (1999)
A15692G	972	Synonymous	Not yet reported
C15735T	926	Ala → Val	Coble *et al.* (2004)
T15784C	926, 976	Synonymous	Herrnstadt *et al.* (2002)
T15792C	976	Ile → Thr	Finnila *et al.* (2001)

1 = Alteration format includes initial nucleotide present in the RCRS, followed by the nucleotide position and then by the nucleotide detected. The RCRS was utilised as the reference sequence (Andrews *et al.*, 1999). Non-synonymous alterations are indicated by three-letter amino acid notations. Ala = alanine, Thr = threonine and Val = valine, Ile = isoleucine. Arrows indicate the direction of amino acid change.

4.2.26.1 The A15692G alteration in the cyt *b* gene

The A15692G transition, located in the cyt *b* gene, is reported as novel in the current investigation and is illustrated with an electropherogram presented in Figure 4.16. The A15692G transition alters the first codon position of the codon for methionine (AUA) and results in a non-synonymous alteration to valine (GUA). Both amino acids have non-polar side chains, thus resulting in minimal disruption in the protein (Campbell, 1995). Previous reports of non-synonymous alterations in the cyt *b* gene have been implicated in affecting the efficiency of OXPHOS ATP production by decreasing the coupling efficiency, and thus increasing heat production. Alterations of this nature have been suggested to allow adaptations to specific climatic environments (Mishmar *et al.*, 2003).

Figure 4.16: Representative electropherogram of the A15692G alteration in the cyt *b* gene



Control = representative electropherogram of an individual with identical sequence to the RCRS. Nucleotides enclosed in a box (□) represent novel alterations. Arrows indicate the nucleotide (nt) positions at which alterations occur.

4.2.27 The transfer RNA threonine gene

The tRNA threonine ($tRNA^{Thr}$) gene encodes a functional tRNA molecule and is located in the 15888 to 15953 position of the mitochondrial genome (Anderson *et al.*, 1981). Three alterations were detected in this region when generated sequences were compared to the RCRS, and are presented in Table 4.16. All alterations detected are previously reported polymorphisms.

Table 4.16: Alterations detected in the $tRNA^{Thr}$ gene

Alteration[1]	Individuals	Alteration Type	Reference
G15930A	941, 964, 972	Synonymous	Mishmar *et al.* (2003)
T15941C	941, 964, 972	Synonymous	Mishmar *et al.* (2003)
T15942C	934	Synonymous	Rose *et al.* (2001)

1 = Alteration format includes initial nucleotide present in the RCRS, followed by the nucleotide position and then by the nucleotide detected. The RCRS was utilised as the reference sequence (Andrews *et al.*, 1999).

4.2.28 The D-loop region

The D-loop of the human mitochondrial genome refers to the non-coding region between $tRNA^{Pro}$ and $tRNA^{Phe}$ (16024 - 576). It represents the least conserved region of the mitochondrial genome, varying in sequence content as well as length between different species (Anderson *et al.*, 1981). No function was initially assigned to this region, but it

contains blocks of nucleotide sequences which exhibit sequence homology between human, bovine and rat mitochondrial genomes, indicating a level of conservation (Anderson *et al.*, 1981). These are conserved regions that correspond to signal elements responsible for transcription and replication of the genome (Clayton, 1991). The D-loop contains unconserved regions corresponding to the HVS1 and HVS2 segments, located in the 16024 - 16383 and 57 - 372 positions respectively (MITOMAP, 2004). Apart from the hypervariable segments, the rest of the D-loop region evolves at a similar rate to that of the coding region (Greenberg *et al.*, 1983).

During the course of the current investigation, a total of 43 nucleotide substitutions, presented in Table 4.17, and five sequence length alterations, presented in Table 4.18, were detected in the D-loop. All nucleotide substitutions detected in the D-loop have previously been described as polymorphisms. Four of the sequence length alterations are previously described polymorphisms and one possibly represents a previously unreported insertion of an unknown number of C residues in the 568 - 577 position, and is discussed prior to Figure 4.17.

Table 4.17: Nucleotide substitutions detected in the D-loop

Alteration[1]	Individuals	Alteration Type	Reference
A73G	921, 926, 995, 934, 939, 941, 964, 972, 976	Transition	Aquadro and Greenberg (1983)
A93G	962	Transition	Graven *et al.* (1995)
A95C	962	Transversion	Graven *et al.* (1995)
G143A	976	Transition	Graven *et al.* (1995)
T146C	926, 939, 941, 964, 972, 976	Transition	Aquadro and Greenberg (1983)
C150T	921, 934, 939	Transition	Aquadro and Greenberg (1983)
T152C	926, 995, 934, 939, 941, 962, 972, 976	Transition	Aquadro and Greenberg (1983)
C182T	939	Transition	Aquadro and Greenberg (1983)
G185A	962	Transition	Aquadro and Greenberg (1983)
A189G	995, 934, 962	Transition	Aquadro and Greenberg (1983)
T195C	921, 926, 939, 941, 964, 972, 976	Transition	Aquadro and Greenberg (1983)
C198T	926, 939	Transition	Graven *et al.* (1995)
T199C	964	Transition	Ozawa *et al.* (1991b)
A200G	934, 976	Transition	Aquadro and Greenberg (1983)
T204C	939	Transition	Graven *et al.* (1995)
T236C	962	Transition	Aquadro and Greenberg (1983)
G247A	995, 941, 962, 964, 972	Transition	Aquadro and Greenberg (1983)

Table 4.17: Continued...

Alteration[1]	Individuals	Alteration Type	Reference
A263G	921, 926, 995, 934, 939, 962, 976	Transition	Andrews *et al.* (1999)
G16129A	939, 941, 962, 964, 972	Transition	Finnila *et al.* (2001)
C16148T	962	Transition	Aquadro and Greenberg (1983)
C16168T	962	Transition	Di Rienzo and Wilson (1991)
T16172C	921, 962	Transition	Greenberg *et al.* (1983)
C16176T	934	Transition	Maca-Meyer *et al.* (2001)
C16187T	995, 941, 962, 964, 972	Transition	Horai and Hayasaka (1990)
C16188G	962	Transversion	Horai and Hayasaka (1990)
C16188A	995	Transition	Horai and Hayasaka (1990)
T16189C	921, 926, 995, 941, 962, 964, 972	Transition	Horai and Hayasaka (1990)
G16213A	939	Transition	Horai and Hayasaka (1990)
C16223T	921, 926, 995, 934, 939, 941, 962, 964, 972, 976	Transition	Horai and Hayasaka (1990)
A16230G	995, 941, 962, 964	Transition	Aquadro and Greenberg (1983)
C16234T	964	Transition	Di Rienzo and Wilson (1991)
C16239T	995, 941, 972	Transition	Maca-Meyer *et al.* (2001)
T16243C	941, 964, 972	Transition	Horai and Hayasaka (1990)
C16266G	964	Transversion	Horai and Hayasaka (1990)
C16278T	926, 939, 962, 976	Transition	Horai and Hayasaka (1990)
C16294T	926, 941, 972, 976	Transition	Di Rienzo and Wilson (1991)
A16309G	926, 976	Transition	Horai and Hayasaka (1990)
T16311C	939, 941, 962, 964, 972	Transition	Horai and Hayasaka (1990)
C16320T	921, 941, 962	Transition	Horai and Hayasaka (1990)
T16327C	934	Transition	Horai *et al.* (1993)
C16354T	939	Transition	Richards *et al.* (2000)
G16390A	926, 939, 976	Transition	Mishmar *et al.* (2003)
T16519C	921, 995, 941, 964, 972	Transition	Aquadro and Greenberg (1983)

1 = Alteration format includes initial nucleotide present in the RCRS, followed by the nucleotide position and then by the nucleotide detected. The RCRS was utilised as the reference sequence (Andrews *et al.*, 1999).

All the sequence length variations, except for the insertion at 568 - 577, consisted of one or two nucleotide insertions or deletions and were generally located in multimeric repeats. The 311 - 315insC polymorphism, detected in all individuals analysed, represents the most common state at this site, as the sequence of the RCRS represents a rare polymorphism.

Table 4.18: Insertions and deletions detected in the D-loop

Alteration[1]	Individuals	Reference
311 - 315insC	921, 926, 995, 934, 939, 941, 962, 964, 972, 976	Graven *et al.* (1995); Andrews *et al.* (1999)
498delC	941, 964, 972	Vigilant *et al.* (1991)
522-523delCA	972, 976	Kleinle *et al.* (1998)
568 - 577insPolyC	972	Not yet reported
16184 - 16194insCC	921	Horai and Hayasaka (1990)

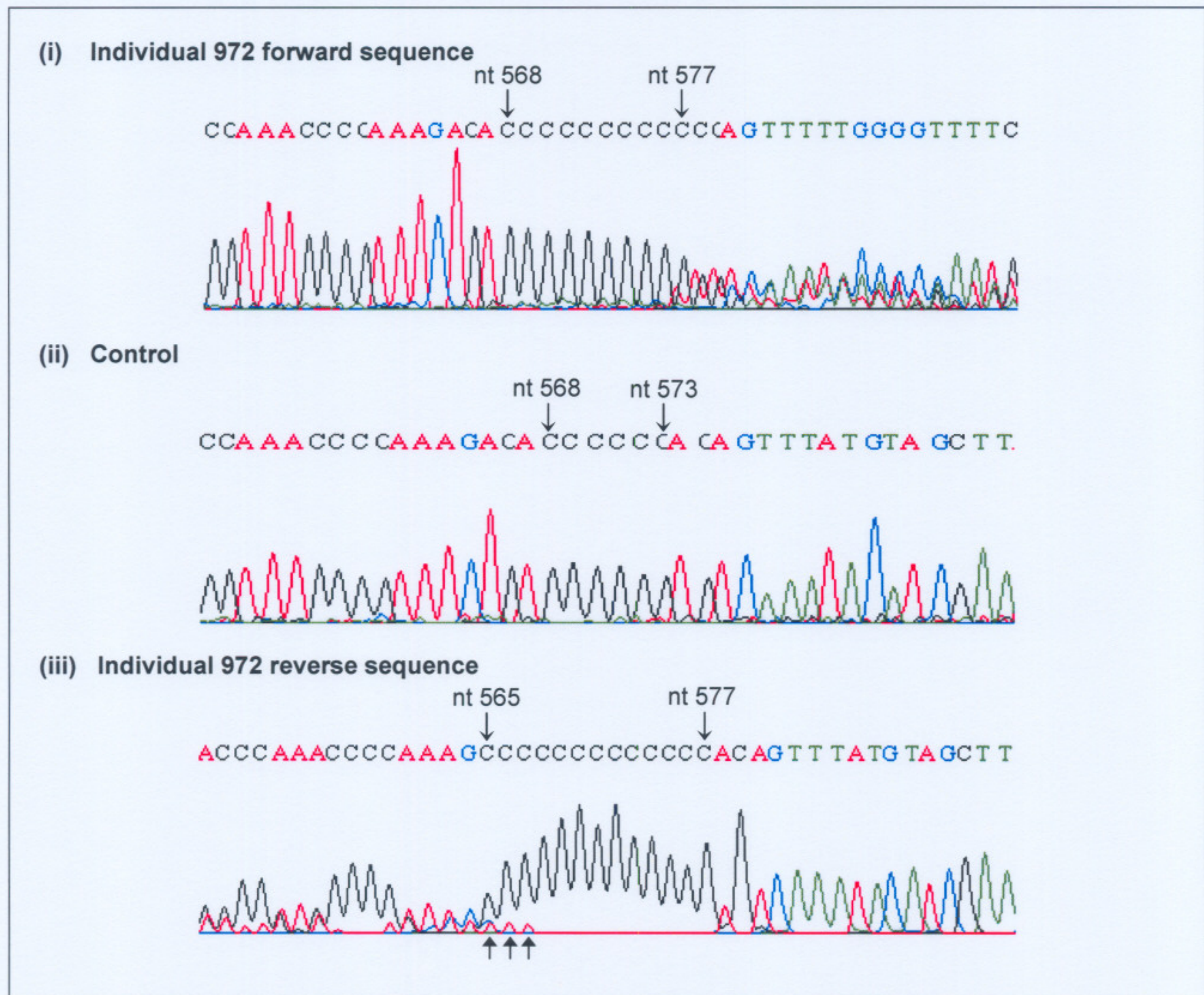
1 = Alteration format includes nucleotide position, followed by an insertion or deletion event, followed by the nucleotide (s) inserted or deleted. ins = insertion. del = deletion. The RCRS was utilised as the reference sequence (Andrews *et al.*, 1999).

An unusual sequence length alteration, 568 - 577insPolyC, was detected in individual 972, and is illustrated with an electropherogram in Figure 4.17. Initially, when the forward primer sequence was obtained, it was envisioned that more than one sequence template present in the sequencing reaction was responsible for the image of mixed template after a certain point in the electropherogram, as illustrated in Figure 4.17 (i). The PCR reaction producing the specific fragment was therefore optimised further, to decrease secondary amplification. However, after the sequencing reaction was repeated, an identical result was obtained when the optimised PCR product was utilised. The presence of mixed template, as a result of an unoptimised PCR reaction, was therefore unlikely to be responsible for the sequence failure.

It was noted that Van Brummelen (2003) obtained a similar result, which the author attributed to a heteroplasmic insertion. The region under investigation was therefore sequenced in the reverse direction, and a similar result as that observed in the forward sequence electropherogram was obtained, as presented in Figure 4.17 (iii).

The long run of Cs detected in the 568 - 577 region may be assigned to a slipped mispairing insertion, formed during replication of the genome (Wrischnik *et al.*, 1987) of four or more Cs in this region. When the forward and reverse sequence electropherograms were analysed, there was a discrepancy in the number of Cs observed in the 568 - 577 region. The electropherogram illustrating the forward sequence is presented in Figure 4.17 (i), and indicates the insertion of four Cs, whilst the reverse sequence electropherogram, illustrated in Figure 4.17 (iii), indicates seven Cs inserted. The difference observed in the number of inserted Cs may be due to slippage of the sequencing enzyme (Greenberg *et al.*, 1983) or may be due to heteroplasmy.

Figure 4.17: Representative electropherogram of the 568 - 577insPolyC



Control = representative electropherogram of an individual with identical sequence to the RCRS. Downward arrows (↓) indicate the nucleotide (nt) positions between which the length alteration occurs. Upward arrows (↑) indicate the positions of overlapping peaks.

An explanation linked to the heteroplasmic status for the 568 - 577insPolyC alteration is favoured over the possibility of slippage by the sequencing enzyme, due to two observations. Firstly, an insertion of four Cs was observed when the forward sequencing reaction was repeated, and an insertion of seven Cs was observed when the reverse sequencing reaction was repeated. It is unlikely that slippage of the sequencing enzyme could account for the same number of Cs inserted when the forward sequencing reaction was repeated. This situation is also true for the reverse sequencing reaction. Therefore, the option of slippage of the sequencing enzyme is less likely to have generated the observed results. When the forward sequence is critically analysed, the 568 - 577insCCCC generated the strongest signal in the electropherogram and when the reverse sequence is critically analysed, the 568 - 577insCCCCCCC generated the strongest signal in the electropherogram. This is confirmed by the presence of simultaneous or overlapping

peaks in the reverse sequence electropherogram at the 565 - 567 position, illustrated in Figure 4.17 (iii), which may represent the insertion of three Cs.

Secondly, the appearance of overlapping peaks downstream of the polyC insertion in the forward sequence and upstream of the polyC insertion in the reverse sequence indicates the presence of mixed templates. Thus, not all the mitochondrial genomes are equal with regard to the length of the polyC insertion. The simultaneous sequencing of both templates will result in amplicons of different lengths, which will produce electropherograms of a superimposed nature. No report describing this particular insertion could be identified. However, novel status was not assigned to the alteration, as the exact length of the polyC insertion could not be confirmed with complete confidence.

The heteroplasmic status of this alteration can be verified through several procedures. The level of heteroplasmy may differ between different tissues in an individual (Larsson and Clayton, 1995). Isolating DNA from various tissues of the individual, and subsequent sequencing thereof, will aid in confirming the absence or presence of heteroplasmy in various tissues. If heteroplasmy is detected in all tissues, RFLP analysis can be utilised to confirm the heteroplasmic status. The recognition of RE sites upstream and downstream of the insertion would result in two fragments of unequal size, that can be detected via polyacrylamide gel electrophoresis (PAGE), thus confirming the heteroplasmic status.

4.2.29 SEQUENCE COMPARISON

The complete mitochondrial genome sequences of 10 individuals of Zulu, Xhosa and Tswana ethnic origin were determined to investigate the relationship between different Southern African ethnic groups. The comparison of sequences relative to the RCRS revealed a total of 222 alterations, of which 207 were previously reported polymorphisms. Fifteen have to date not been reported, and are presented for the first time in the current investigation.

The large number of differences between the generated sequences and the RCRS can be attributed to the fact that the RCRS is derived from an individual with European haplogroup H. The RCRS therefore exhibits alterations that are European-specific, or more prevalent in the European population relative to the African population, and are thus not often detected in African populations.

All alterations detected are presented in Appendix A, and could be divided into nucleotide substitutions and length alterations. In total, six length alterations were detected and were all located in the D-loop region (Table 4.18, page 82), therefore not resulting in any frameshift mutations. Two of the length alterations detected did not occur in more than one individual, and four of the length alterations were shared between at least two individuals. A total of 216 nucleotide substitutions were detected. As expected, the number of nucleotide substitutions was greater than the number of length alterations, since the likelihood of occurrence as well as fixation is smaller for length variations than for nucleotide substitutions (Greenberg *et al.*, 1983).

Nucleotide substitutions that were detected could be divided into TS and TV and were located in either non-coding, tRNA/rRNA-coding or protein-coding regions. A further division between substitutions occurring once and being shared between at least two individuals was made. A summary of the above-mentioned classes is presented in Table 4.19. Transitions located in the protein-coding regions were most common, with shared transitions occurring twice as frequently as single transitions. It is noted that out of a total of 216 substitutions, 208 represent transitions. This translates to 96.3% of substitutions being classified as transitions, which is in general agreement with similar previous studies (Aquadro and Greenberg, 1983). The degree of transitional bias correlates with the phylogenetic relatedness between entities, where a high degree of transitional bias indicates a close relationship. The high degree of transitions is mainly due to a C deamination bias and multiple substitutions obscuring the transition record rather than to selection of specific gene products (Brown *et al.*, 1982a; Brown *et al.*, 1982b).

Table 4.19: The division of nucleotide substitutions

Mitochondrial Region	Transitions (TS)			Transversions (TV)			Total of TS+TV
	Single	Shared	Total	Single	Shared	Total	
Non-coding	39	24	63	3	0	3	66
tRNA/rRNA-coding	21	18	39	0	1	1	40
Protein-coding	35	71	106	2	2	4	110
Total	95	113	208	5	3	8	216
Percentage (%)	43.7	52.6	96.3	2.3	1.4	3.7	N.A.

Single = nucleotide substitutions detected in a single individual only. Shared = nucleotide substitutions shared between at least two individuals. A total of each class is indicated at the end of each row and column. TS = transition, TV = transversion. N.A. = not applicable.

Nucleotide substitutions in protein-coding regions were divided further into synonymous and non-synonymous alterations. A total of 108 synonymous and 32 non-synonymous

alterations were detected. Of the 108 synonymous alterations, 45 were detected once and 63 were detected in at least two individuals. Thirteen of the non-synonymous alterations were detected only once whilst 19 were shared between at least two individuals.

The degree of divergence between individuals can be estimated by the number of sequence differences between them (Brown *et al.*, 1982b). The total number of nucleotide substitutions between each combination of two individuals was calculated and is presented in Table 4.20. In the current study 101 alterations were detected only once, whilst 120 were detected in more than one individual. Of the 206 previously reported polymorphisms, 88 were detected once and 118 were detected in more than one individual. Eleven of the 13 novel alterations were detected only once, whilst two were shared alterations. The 568 - 577insPolyC alteration that was not assigned novel status, as discussed in Paragraph 4.2.28, was detected only in a single individual. The T9637A alteration occurred in more than one individual and was not given novel status, as discussed in Paragraph 4.2.18.1. Therefore, a total of 15 alterations were detected that had previously not been reported in the literature. The larger number of shared alterations between individuals is an indication of greater relatedness (Cann *et al.*, 1987), as discussed in Paragraph 2.8.

Through analysing Table 4.20, it is observed that some groups of individuals share more alterations than other combinations of individuals. Most nucleotide differences were shared between individuals 941 (Zulu), 962 (Tswana), 964 (Tswana) and 972 (Tswana). A large number of alterations were shared between individuals 926 (Xhosa), 976 (Tswana) and 939 (Zulu), as well as between individuals 995 (Xhosa), 962, 964 and 972 (Tswana) and 941 (Zulu). The number of shared alterations between individuals in an ethnic group is in most cases low compared to individuals in different ethnic groups. Thus, based on the number of shared alterations, more alterations seem to be shared between individuals belonging to different ethnic groups than between individuals in an ethnic group.

Table 4.20: Observed number of nucleotide differences between individuals

Individual		Xhosa			Zulu			Tswana			
		921	926	995	934	939	941	962	964	972	976
Xhosa	921										
	926	20									
	995	22	22								
Zulu	934	22	19	24							
	939	20	37	21	20						
	941	18	28	35	16	27					
Tswana	962	19	26	41	17	25	51				
	964	17	25	35	14	26	69	51			
	972	18	28	37	16	27	69	51	72		
	976	18	45	20	19	34	27	23	24	28	

Observed number of nucleotide differences between two individuals is presented at the intercept of the row of one and the column of the other individual. Xhosa individuals = 921, 926 and 995. Zulu individuals = 934, 939 and 941. Tswana individuals = 962, 964, 972 and 976. Grey coloured cells contain no data.

Changes that were detected in at least two individuals were called "shared alterations". These were unique in that they were only present in some individuals. Combinations of individuals were formed based on the presence of these shared alterations. Each combination was termed a motif. Some motifs occurred more than once, i.e. some combinations of individuals shared more than one unique alteration. The number of occurrences in which a specific motif was detected was an indication of the relatedness between individuals in the motif. The number of occurrences was calculated and is presented in Table 4.21. A complete list of alterations from which the motifs were compiled, is listed in Appendix C. Motifs with only one alteration were excluded in this particular analysis. The highest number of shared unique alterations was detected between individuals 941, 964 and 972, as indicated by motif xiv in Table 4.21.

The shared alterations indicated in Table 4.21 can be viewed in another context, namely that of nucleotide differences. Individuals comprising a motif have alterations in common, but have the same number of nucleotide differences relative to individuals not included in the motif. Therefore, the data presented in Table 4.21, and discussed in Paragraph 4.2.29.1, are also indicative of the nucleotide differences between individuals.

Table 4.21: Combinations of alterations between individuals

Motif name	Number of alterations in a motif	Xhosa			Zulu			Tswana			
		921	926	995	934	939	941	962	964	972	976
i	2	0	0	1	1	0	0	0	0	0	0
ii	2	0	0	0	0	1	1	1	1	1	0
iii	2	1	1	1	1	1	0	1	0	0	1
iv	3	1	0	1	1	0	0	0	0	0	0
v	3	0	1	1	0	1	1	1	1	1	1
vi	4	0	0	0	0	0	1	0	0	1	0
vii	4	0	0	1	0	0	0	1	0	0	0
viii	4	0	1	0	0	1	1	1	1	1	1
ix	6	0	1	0	0	1	0	0	0	0	1
x	9	0	1	0	0	0	0	0	0	0	1
xi	10	0	0	0	0	0	1	1	1	1	0
xii	13	0	0	1	0	0	1	1	1	1	0
xiii	13	1	1	1	1	1	1	1	1	1	1
xiv	15	0	0	0	0	0	1	0	1	1	0

A "1" indicates the presence and a "0" the absence of an individual in the motif. Alterations utilised to construct each motif are indicated in Appendix C.

4.2.29.1 Haplogroup analysis

Mitochondrial sequence alterations reflect the evolutionary history of an individual. Mutations arise in the female germ line and are transmitted to the next generation (Wallace, 1994). Thus, various mtDNA lineages arise that are diverse as a result of different evolutionary histories (Wallace *et al.*, 1999). Populations that have undergone differences in evolutionary histories can therefore be distinguished based on their sequence composition. For the purpose of the current study, haplogrouping was initiated successfully utilising macro-haplogroup L specific SNPs. This is due to the fact that all individuals investigated were of African origin, increasing the likelihood of an individual clustering in macro-haplogroup L.

Haplogroup definitive criteria as stated by Chen *et al.* (1995), Watson *et al.* (1997) and Torroni *et al.* (2001), and employed by most authors to identify African-specific haplogroups, were unsuccessful in assigning haplogroups to all individuals analysed. This was due to the fact that the populations investigated in the current study were not represented in the samples utilised by the aforementioned authors to generate the first generation of haplogroup-informative sites. It was therefore necessary to employ the haplogroup criteria of Wallace (2003), where additional data were included when defining haplogroup sites, thereby representing the populations in the current investigation.

The Sub-Saharan African-specific macro-haplogroup L is divided into four haplogroups, namely L0, L1 L2 and L3. The African macro-haplogroup L expanded and formed the African-specific L3 haplogroup as well as the Eurasian macro-haplogroups M and N. Haplogroups M and N left Africa to colonise Europe and Asia (Mishmar *et al.*, 2003).

Haplogroups L0, L1 and L2 are defined by a T nucleotide at position 3594 and haplogroups L3, M and N by a C nucleotide at this position (Wallace, 1994; Chen *et al.*, 2000). These haplogroups are further subdivided by other polymorphisms, as discussed in Paragraph 3.5.2. The haplogroup of each individual is presented in Table 4.22, with the time to the MRCA of haplogroups indicated. The age of haplogroups L0b, L0c1, L3a1 and L3a3 is not known. In cases where the age of a sub-haplogroup is not known, the age of the haplogroup from which it branched is stated. The haplogroups represented by the individuals in the current study therefore have a time to the MRCA several orders of magnitude larger than that specified by the ethnic group time to the MRCA, as discussed in Paragraph 2.9.

Table 4.22: Haplogroup assignment of individuals of different ethnic origin

Haplogroup	Individuals	Age of haplogroup (YBP)
L0b	941 (Z), 964 (T), 972 (T)	L0 = 125,000 - 159,000
L0c1	962 (T)	
L2a	926 (X), 976 (T)	39,000 - 51,400
L2b	939 (Z)	24,800 - 32,700
L3a1	921 (X)	L3a = 78,300 - 103,200
L3a3	995 (X), 934 (Z)	

Z = Zulu ethnic origin, T = Tswana ethnic origin and X = Xhosa ethnic origin. YBP = years before present. The age of haplogroups is indicated here, as stated by Mishmar *et al.* (2003) and Chen *et al.* (2000).

Upon haplogroup assignment, the individuals analysed were grouped into three haplogroups, namely L0, L1 and L3 (see Table 4.22). However, the three haplogroups identified did not correlate to the three ethnic groups, as originally anticipated due to the possibility of the ethnic groups sharing a common ancestor, as discussed in Paragraph 2.9. Haplogroup L0 consisted of three Tswana individuals and one Zulu individual, L2 consisted of one Xhosa, one Zulu and one Tswana individual, and L3 consisted of two Xhosa individuals and one Zulu individual. No ethnic group was definable by a single haplogroup. Therefore, there was no apparent association between the ethnic status and haplogroup of individuals analysed in the current investigation. There are, however, possible reasons to substantiate the lack of correlation between ethnic groups and haplogroups observed in the current investigation. Firstly, the haplogroup defining

polymorphisms utilised in the current investigation may only differentiate between older evolutionary units. The ethnic groups analysed may therefore represent recently founded sublineages that have not yet developed lineage-specific polymorphisms.

Secondly, a limitation of ethnic classification, that the ethnic status of an individual may be incorrectly stated by the individual under investigation, may result in the incorrect grouping of individuals. The ethnic origin of an individual was determined based on self-report by the subject, as well as the ethnicity of both parents, as provided by the individual, as described in Paragraph 3.1. This method of classification is limited in that it relies on the information given by the individual as being correct. The possibility that certain individuals may be incorrectly classified according to ethnicity can thus not be ignored. An example of this may be represented by individual 941, who is of Zulu ethnic origin and belongs to haplogroup L0. Three of the four Tswana individuals clustered in haplogroup L0. Therefore, individual 941 may be of Tswana ethnic origin and may be incorrectly classified as Zulu. It will thus be necessary in future studies to sample more individuals, and verify the ethnic status of their parents again in order to elucidate the relationship between language and mtDNA haplogroups. A similar scenario may be true for individual 976, who is of self-reported Tswana ethnic origin, and clusters in haplogroup L2, together with Xhosa and Zulu individuals. Therefore, incorrect self-reported classification of ethnicity may be responsible for an ethnic group not being definable by a single haplogroup.

Haplogroup L1 was not detected in any samples analysed, even though this haplogroup is specific to the African continent (Chen *et al.*, 2000) and has been reported in Southern African individuals, namely Zulu individuals from Johannesburg, South Africa (Watson *et al.*, 1996; Watson *et al.*, 1997). The absence of haplogroup L1 can be explained by the following: firstly, haplogroup L1 may be present in the ethnic groups analysed, but by chance is not represented in the individuals included in the current study. This can be attributed to the fact that the individuals investigated in this study are not a representative of their respective ethnic group, due to the small sample size of each ethnic group. Secondly, haplogroup L1 is concentrated in West Africa, and reported to be rare in Southern Africa (Salas *et al.*, 2002). Haplogroup L1 may therefore be rare in the ethnic groups investigated in the current study, or may not be represented in the samples of the ethnic groups analysed because of the small sample size.

Haplogroups L0, L2a, L2b and L3, detected in the current investigation, have previously been reported in the Bantu ethnic group (Johnson *et al.*, 1983; Salas *et al.*, 2002; Watson

et al., 1996; Watson *et al.*, 1997). However, the sub-haplogroups of L0, namely L0b and L0c1, as well as those of L3, namely L3a1 and L3a3, were not discussed in these reports. It is therefore not known whether the haplogroups described in the above papers included the specific sub-haplogroups detected in the current study. Haplogroup L2 contributes ca. 36% of the south-eastern Bantu population (Salas *et al.*, 2002). Haplogroup L2a is the most frequent and prevalent in Africa (Salas *et al.*, 2002), and haplogroup L2b is absent in Eastern Africans, rare in Southern Africans and mainly restricted to Western Africa (Torroni *et al.*, 2001), suggesting a western, western-central origin (Salas *et al.*, 2002). Haplogroups L3a and L3b arose in Sub-Saharan Africa and moved north to ultimately yield Asian mtDNAs belonging to haplogroup M (Chen *et al.*, 2000). Haplogroup L0 represents the most ancient of all haplogroups, from which haplogroups L1 and L2 diverged (Chen *et al.*, 2000), and is present at high frequencies in Southern African Khoi-San lineages (Chen *et al.*, 2000).

The haplogroups represented by individuals in the current study may be the result of admixture. It is noteworthy that three of the Tswana individuals group in haplogroup L0 lineages, which was also observed in Southern African Khoi-San populations, as discussed in the previous paragraph. This haplogroup may have been acquired by the Tswana Bantu group as a result of admixture with resident Khoi-San populations (Soodyall *et al.*, 1996), during the Bantu expansion, as discussed in Paragraph 2.9. Haplogroup L3, detected in the current study, is observed in non-African populations as well (Chen *et al.*, 2000) and may also have been acquired by admixture.

There are several shared alterations between individuals in haplogroups. Individuals in haplogroup L0 share a total of 46 alterations, of which 10 are unique to these individuals, as indicated in motif xi, Table 4.21. Of the unique changes, two are haplogroup informative. Individuals belonging to haplogroup L2 have a T nucleotide at position 10810, and individuals in haplogroup L0 and L1 have a C nucleotide. Haplogroup L0a contains a G nucleotide at position 9755, whereas haplogroups L0b and L0c have an A nucleotide. Thirty-five alterations, six of which are unique, are shared between individuals in haplogroup L2, as presented in motif ix in Table 4.21. Individuals belonging to haplogroup L3 share 18 alterations, of which three are unique, and are indicated in motif iv in Table 4.21. One of the unique alterations, namely A10819G, is detected in haplogroup L3a, and excludes individuals from haplogroups L3b and L3c.

Individuals in the Xhosa, Zulu and Tswana ethnic groups share a total of 13 unique alterations, as indicated in motif xii of Table 4.21. However, these shared alterations are detected in all individuals analysed. There are no unique alterations between individuals comprising a single ethnic group. Therefore, the grouping of individuals in an ethnic group is not supported by shared alterations. In contrast, the grouping of individuals via haplogroup informative sites is supported by unique alterations detected between individuals in this study.

The haplogroup assignment of individuals as well as the distribution of shared alterations suggests that the individual ethnic groups under investigation could not each be assigned to a single haplogroup via full mitochondrial DNA sequence analysis. However, several groups can be distinguished that have a partial correlation to ethnicity. Three of the four individuals of Tswana ethnic origin cluster in haplogroup L0. The fourth, namely 941, may represent an outlier whose ethnicity should be re-confirmed during further analysis and interviews with the individuals under investigation.

From the data generated in this study, the Xhosa and Zulu individuals seem to form a heterogeneous group, with five individuals clustering in either haplogroups L3a or L2 and one in haplogroup L0b, which may represent an outlier, as discussed above. Possible reasons to substantiate the lack of differentiation between these two ethnic groups may be assigned to the fact that they may originally have shared a common ancestor, or alternatively to admixture between them. This could have occurred during the Bantu expansion into Southern Africa, or after the ethnic groups had settled in Southern Africa. The sample size of each ethnic group analysed in the current study is small and therefore not necessarily representative of each ethnic population in Southern Africa. Xhosa and Zulu individuals may therefore not be limited to the haplogroups detected in the current study. Individual 976, of self-reported Tswana ethnic origin, may represent another outlier similar to the scenario discussed in the previous paragraph for individual 941.

The sample sizes of each ethnic group are small, and therefore may not accurately reflect the haplogroup composition of their respective populations. By analysing the above presented data, the ethnic groups are observed to be enriched for selected haplogroups, and certain haplogroups were not detected in specific ethnic groups. In the current study, each ethnic group is composed of more than one haplogroup. The self-reported ethnic origin of an individual in the current study can therefore not be utilised as a proxy to correctly envisage the haplogroup of an individual. This investigation indicates the

possibility that these ethnic groups may be composed of evolutionary components that merged at the inception of the ethnic group, or that the diverse haplogroup composition may be a result of recent admixture between the investigated ethnic groups and/or foreign ethnic groups.

However, the distribution of ethnic groups between haplogroups detected in the current study may not be an accurate reflection, as certain individuals may represent outliers, incorrectly classified by the self-reported ethnic origin, as described in Paragraph 3.1. These outliers are therefore grouped incorrectly and represent the false positive or negative presence or absence of haplogroups in an ethnic group. It may, however, be possible to start haplogrouping of individuals of Tswana ethnic origin via haplogroup L0 specific sites. Similarly, haplogroup analysis of individuals of Xhosa and Zulu origins may be initiated utilising haplogroup L2 and L3a specific sites.

4.2.30 CONSTRUCTION OF PHYLOGENETIC TREES

Molecular phylogenies convert information contained in sequences into an evolutionary tree. Phylogenetic trees can provide information on the evolutionary differences, patterns, rate and time estimates between organisms (Nei and Kumar, 2000). The full mitochondrial sequences generated in the current study were utilised to infer the evolutionary relationships between the 10 individuals analysed. A UPGMA tree, presented in Paragraph 4.2.30.1, and a MP tree, presented in Paragraph 4.2.30.2, was constructed utilising MEGA v2.1 (Kumar *et al.*, 2001). It was not necessary to root the phylogenetic trees, as the inclusion of a root, representing an outgroup sequence, serves only as a reference to determine the exact evolutionary distance and timescale difference between individuals (Nei and Kumar, 2000). The relative timescale between individuals is not indicated, as it was not an aim of the current investigation. The branch lengths of the trees presented in this study therefore represent the relative nucleotide differences between individuals.

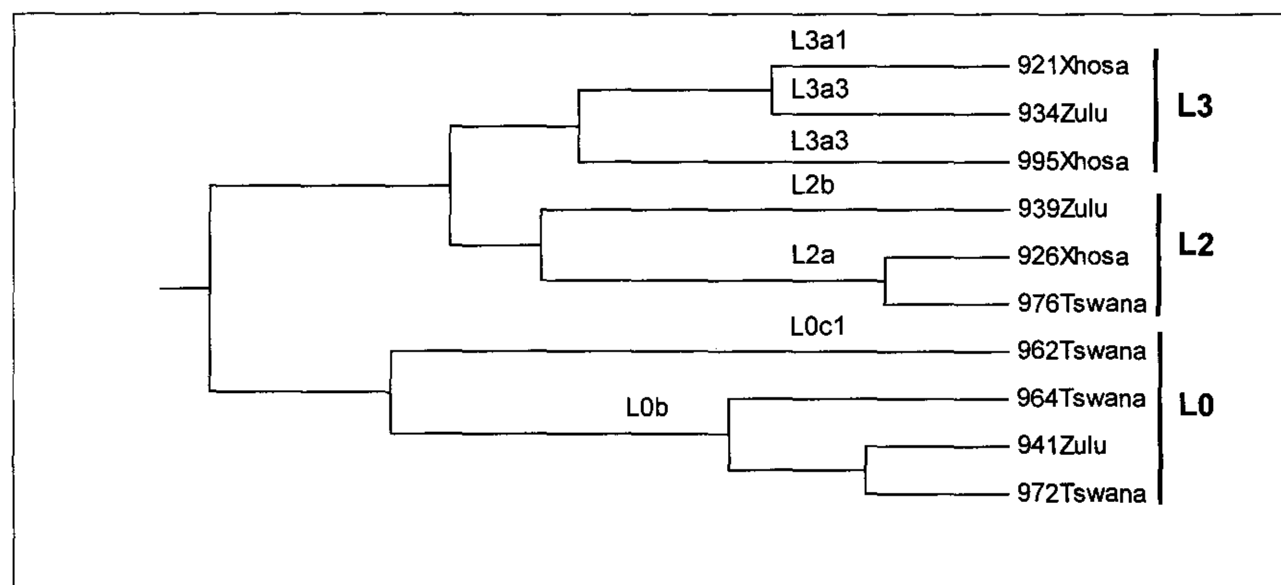
4.2.30.1 UPGMA tree

A UPGMA phylogenetic tree was constructed utilising full mitochondrial sequence, as discussed in Paragraph 3.5.3. With the UPGMA distance method, the sequence data were converted into evolutionary distances. This resulted in loss of information, as the nature of alterations was not taken into account (Page and Holmes, 1998). These distances were

computed for all pairs of individuals. The pairs having the smallest evolutionary distance were clustered together and the distances re-calculated. This process was repeated several times until all clusters were identified. These clusters were utilised to determine the branching order of the phylogenetic tree, and also allowed the inference of both the branch length and topology of the tree (Nei and Kumar, 2000).

UPGMA trees are ideally suited for constructing phylogenies if it can be assumed that the rate of nucleotide substitution is constant (Nei and Kumar, 2000). Although a neutral model of evolution cannot be assigned to human mitochondrial DNA (Excoffier *et al.*, 1990), specifics regarding selection (Mishmar *et al.*, 2003; Ruiz-Pesini *et al.*, 2004) and recent population history in the African population are currently unknown. The clustering of individuals by UPGMA analysis, as illustrated in Figure 4.18, is in agreement with the haplogroup assignment presented in Paragraph 4.2.29.1. Clusters also corresponded to sub-haplogrouping, with the exception of individual 995 (Xhosa, haplogroup L3a3), which, according to the haplogroup, should cluster with individual 934 (Zulu, haplogroup L3a3).

Figure 4.18: UPGMA tree of 10 individuals of three ethnic origins



UPGMA tree of full mitochondrial sequence of 10 individuals belonging to three different ethnic groups. Haplogroups are indicated on the right of each individual. Sub-haplogroups are indicated above each branch. Branch lengths indicate relative nucleotide differences between individuals.

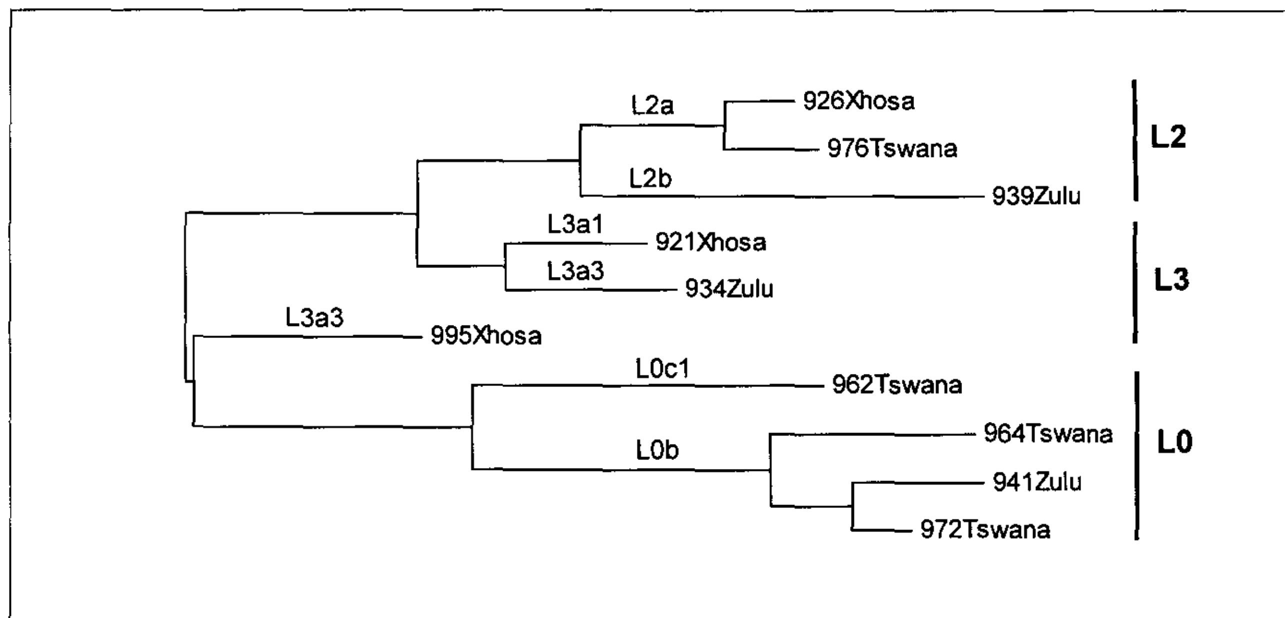
4.2.30.2 MP tree

Maximum parsimony is a discrete method of inferring phylogenies, which utilises direct sequence data, rather than pairwise distances (see Paragraph 3.5.3). The tree topology was constructed on the basis of the smallest number of nucleotide substitutions that resulted in the evolutionary process represented by the tree. All possibly correct topologies were computed and the topology that required the least number of substitutions was

chosen as the best tree. The minimum number of substitutions for each site and for each topology, as well as the sum of all substitutions for all topologies, was computed. This sum was called the tree length and the tree topology with the minimum tree length was called the MP tree. However, it is possible that more than one topology was constructed with the same number of minimum substitutions. Therefore, each MP tree with a different topology was considered as potentially representing the true topology (Nei and Kumar, 2000).

This method of tree construction is not suited for the current study for several reasons. The most correct topology is only generated if there are no backward or parallel substitutions (homoplasy). This cannot be assumed, as the high mtDNA mutation rate may result in homoplasy (Horai *et al.*, 1995). Only parsimony-informative sites i.e. nucleotide sites containing an alteration detected in more than two individuals, are considered. Therefore, not all alterations are utilised in constructing the tree. The clustering of individuals via MP analysis, as illustrated in Figure 4.19, is in agreement with that observed in the UPGMA tree. However, individual 995 is placed outside of the haplogroup L3a cluster. This implies that individual 995 was placed incorrectly due to the MP tree construction method. This may be due to the MP method not being the best suited to reflect the evolutionary relationships between the individuals in the current study, or due to individual 995 containing unique, non-haplogroup informative alterations that result in the branching pattern observed, as discussed in Paragraph 4.2.30.3.

Figure 4.19: MP tree of 10 individuals from three ethnic origins



MP tree of full mitochondrial sequence of 10 individuals belonging to three different ethnic groups. Haplogroups are indicated on the right of each individual. Sub-haplogroups are indicated above each branch. Branch lengths indicate relative nucleotide differences between individuals.

4.2.30.3 Comparison of phylogenetic trees

There are several reasons to substantiate the branching pattern observed, as well as the differences between the two trees. The trees that were constructed do not necessarily reflect the true evolutionary relationships between the individuals analysed (Nei and Kumar, 2000). This may be a result of the chosen tree-building methods not being ideally suited to reflect the evolutionary relationships between the individuals analysed. The observed topology may also be a result of the sequences not accurately reflecting the evolutionary history of individuals (Nei and Kumar, 2000), i.e. some individuals may harbour unique alterations not observed in the rest of the population.

In analysing both trees that were constructed, it was observed that each haplogroup formed a monophyletic group, with the exception of individual 995 in the MP tree. The disagreement between haplogroup and the placing of individual 995 in the UPGMA and MP tree may be assigned to an incorrect topology due to the number of differences in the third codon position. These differences reach a saturation level, as they accumulate faster than alterations in the first and second codon position, and introduce "noise" in phylogenetic construction (Russo *et al.*, 1996). The fact that individual 995 does not cluster together with the other L3a3 individual, namely 934, may be due to the "noise" mentioned above. Although these two individuals contain the same haplogroup informative sites, differences at other sites cause dispersed clustering. Although resulting in the loss of information when sequences are analysed, the UPGMA distance method may reflect a more accurate topology. This is due to the fact that all alterations, even parsimony-uninformative sites, are considered, which contribute to the topology of the constructed tree. However, the MP tree topology cannot be disregarded, as parsimony-informative sites reflect and emphasise the genetic relationship between individuals (Nei and Kumar, 2000).

The clustering and dispersal of individuals via shared alterations, haplogroup assignment and phylogenetic analysis are not random. The majority of Zulu and Xhosa individuals cluster together in haplogroups L2 and L3, and most Tswana individuals in haplogroup L0. Haplogroup L0 contains no Xhosa individuals and Haplogroup L3 contains no Tswana individuals. The absence of specific haplogroups in the ethnic groups analysed in the current study may be due to the ethnic groups not being representative of their specific populations due to the small sample size of the current investigation.

The black Southern African individuals analysed in the current investigation are not homogeneous on a genetic level. The individuals are diverse in this regard, as three of the four haplogroups in macro-haplogroup L were detected. There is a non-random clustering of individuals that needs to be investigated further by analysing more individuals in each ethnic group, as well as more ethnic groups.

Phylogenetic analyses of the full mitochondrial sequences, generated in the current study, as well as haplogroup assignment and the distribution of shared alterations, indicate no apparent correlation between ethnicity and the genetic make-up of an individual. Another possibility for the observed results is incorrect ethnic classification, as described in Paragraph 3.1. The outcomes of the analyses are critically reliant on the correctness of the self report assessment of ethnicity. It is therefore of utmost importance to reconfirm the ethnicity of each individual through a repeated interview.

The haplogroup and sub-haplogroup assignment of individuals, via haplogroup informative sites, is in agreement with full mitochondrial sequence data of individuals, except for distinguishing between sub-haplogroups of L3a, as observed in Figure 4.18. Analysing only haplogroup informative sites may therefore be sufficient to deduce the evolutionary relationships between individuals in the current investigation.

CHAPTER FIVE

CONCLUSIONS

An association between haplogroup and ethnic origin exists to the extent that most populations harbour a unique and distinctive mitochondrial sequence (Brown, 1980). Specific mutations may be prevalent in certain populations and haplogroups, resulting in differential functionality. This differential functionality includes adaptation to colder climates, increased longevity and possible adaptation to high altitudes, as discussed in Paragraph 2.7.2.1. In addition, the haplogroup of an individual has often been recognised to play a decisive role in the expression of certain disease phenotypes (Shoffner *et al.*, 1993), as discussed in Paragraph 2.7.2.2. There are also differences in disease prevalence and drug response between populations, as discussed in Paragraph 2.7.2.2.

Among populations there is also a relationship between the level of genetic differentiation and linguistics. On a global scale, populations can be distinguished from one another based on the linguistic phyla to which they belong (Cavalli-Sforza *et al.*, 1988). Therefore, in continental populations, there is correlation between haplogroup and linguistics, as described in Paragraph 2.10.

Although there is a high degree of genetic differentiation among Sub-Saharan African populations, it is not quantifiable through linguistic or geographic differentiation (Excoffier *et al.*, 1991). However, these authors only analysed blood groups, which are not as informative as mitochondrial analysis (Wallace, 1994). The correlation between language and haplogroup therefore still needs to be determined between Bantu ethnic groups in Southern Africa. This will have an impact upon sampling strategies employed in genetic studies in the Southern African context. If a linguistic-haplogroup association exists, then it would be sufficient to sample on an ethnic level and subsequently determine the expression of a specific disease in each ethnic group. If no association between ethnic group and haplogroup exists, then a broader sampling strategy should be employed that results in the representation of each haplogroup in the entire population.

In view of limited information on the black South African population the full mitochondrial genome sequences of 10 individuals belonging to three distinct Southern African ethnic

groups were generated. The individuals analysed were three Xhosa, three Zulu and four Tswana individuals. This was performed with the specific aim to determine whether the investigated individuals, belonging to different ethnic groups, could be distinguished from one another on a genetic level, and whether the genetic make-up of individuals corresponded to their self-identified ethnicity.

5.1 ALTERATIONS OBSERVED

Investigation of the full mitochondrial sequences generated revealed a total of 221 alterations, which constituted 206 previously reported polymorphisms, 13 novel alterations and two unreported alterations that need to be confirmed prior to assigning novel status to them, as discussed in Paragraph 4.2.29. The alterations that could be assigned novel status were few, relative to similar previous studies (Mouton, 2003; Van Brummelen, 2003). The number of novel changes detected in the current study represents an accurate representation of the true number of novel alterations, due to the effective method of assigning novel status to an alteration, as described in the introduction of Paragraph 4.2.

Ten of the novel alterations were synonymous, which could have no functional effect because of the nature of the alterations. Two novel changes were transitions in rRNA coding regions and one non-synonymous, which could possibly have a functional effect, as discussed in Paragraphs 4.2.2.1, 4.2.2.2 and 4.2.24.2. However, the individuals in whom the alterations were detected presented with no clinical symptoms, as described in Paragraph 3.1.

Except for two novel alterations that were detected twice, as presented in Paragraphs 4.2.4.1 and 4.2.15.1, novel alterations were detected only once. It is unlikely that either of the T3618C or T8503C alterations arose independently in two unrelated individuals. These alterations may therefore have arisen in a common ancestor to both individuals, giving rise to the possibility that more black African individuals harbour these alterations. The findings of future studies will determine whether the novel alterations detected in the current study are present in other individuals as well. Although all previously reported alterations are listed as polymorphisms in the current text, they may actually represent signatures of selection (Elson *et al.*, 2004) or demographic history (Excoffier *et al.*, 2002) that are lineage-specific, which may also be true of the novel alterations detected in the current investigation.

As discussed in Paragraph 2.8, the genetic relationships between individuals can be determined by analysis of the shared alterations between them. It was evident upon evaluation of the shared alterations between individuals included in the current study, and presented in Paragraph 4.2.27, that there are no unique shared alterations between individuals in an ethnic group. In contrast, several unique alterations are shared between individuals that belong to different ethnic groups.

5.2 HAPLOGROUP DETERMINATION

When genetic disorders are investigated in a population, it is essential to take into consideration the specific genetic background, or haplogroup, of the individuals under investigation, as discussed in Paragraph 2.7.2.2. In order to investigate the role genetic background plays in the expression and prevalence of a disease phenotype and drug response, it is necessary to investigate this correlation in each haplogroup. It is therefore prudent to haplogroup the individuals investigated in disorder studies.

Upon haplogroup determination it was detected that haplogroup L0 contained Tswana and Zulu individuals, haplogroup L2 contained Tswana, Zulu and Xhosa, and haplogroup L3a contained Zulu and Xhosa. Therefore, not one of the ethnic groups in the current study was definable by a single haplogroup, as discussed in Paragraph 4.2.29.1. The number of unique alterations, presented in Paragraph 4.2.29.1, between individuals belonging to a single haplogroup, relative to individuals within an ethnic group, is in sharp contrast with one another. No unique alterations were detected that were specific to an ethnic group and shared between all individuals in the ethnic groups of the current study. However, several unique alterations were shared between individuals comprising a single haplogroup. Therefore, the distribution of shared alterations in the current study is supportive of a haplogroup clustering rather than clustering based on self-identified ethnicity.

5.3 PHYLOGENETIC ANALYSIS

The clustering of individuals in their respective haplogroups is further supported by phylogenetic analysis of the full mitochondrial sequence, as presented in Paragraph 4.2.30. By analysing the UPGMA tree constructed in the current study, it was also possible to delineate sub-haplogroups within haplogroups L0 and L2. The topology of the MP tree was in agreement with the UPGMA tree, except for the placement of individual 995. This individual harbours a unique array of alterations, which may result in the difference in

clustering due to the difference in the tree-building methods of UPGMA and MP, as discussed in Paragraph 4.2.30.2. This needs to be verified in future studies. The UPGMA method may reflect a more accurate topology, as all alterations were considered, despite the loss of information when evolutionary data are converted to evolutionary distance. In contrast, only parsimony informative sites were considered in the MP method, thus not accurately reflecting the evolutionary relationships between the individuals analysed.

Haplogroup assignment of individuals is in agreement with the UPGMA phylogenetic analysis of the full mitochondrial genome. All sub-haplogroups of the haplogroups L0, L2 and L3 were delineated by phylogenetic analysis. There were anomalies in the clustering of individuals by haplogroup assignment relative to the phylogenetic analysis of full mitochondrial sequence, as observed in the clustering of individuals in sub-haplogroup L3a, discussed in Paragraph 4.2.30.3. However, full mitochondrial sequence analysis is more accurate than haplogroup analysis in describing evolutionary relationships, as more sites that have possibly undergone evolutionary change are analysed. These anomalies are unlikely to influence the findings when only haplogroup analyses are performed, as all haplogroup levels were correctly delineated. Haplogroup analysis may therefore be sufficient for describing the evolutionary relationships between individuals in the current study.

5.4 THE ROLE OF ETHNICITY IN GENETIC STUDIES

In the investigation presented, it was detected that ethnicity, based on language, is only effective as a partial marker for identifying genetically unique groups. The current investigation represents a pilot study that was aimed at generating results in support of the above statement. Haplogroup analysis indicated that most of the Tswana individuals are grouped in, but are not limited, to haplogroup L0. Phylogenetic analysis, as well as shared alterations of the full mitochondrial genome, is supportive of this grouping. Although the individuals analysed in the current study are not necessarily representative of their respective ethnic groups due to the small sample size, future haplogrouping studies may regard ethnicity as an indicator of which haplogroup informative sites to investigate first. For example, the haplogrouping procedure of Tswana individuals may be initiated by investigating L0 specific SNPs. The importance of future studies that include more individuals is emphasised by the possibility of the haplogroup clustering being due to chance.

A similar argument to the above may be valid regarding the Xhosa and Zulu ethnic groups, if they are considered as sharing a common ancestor, as the ethnic groups are not distinguishable by mtDNA investigation. The findings of the current study indicate that most of the Zulu and Xhosa individuals cluster together, as presented in Figure 4.18, and belong to either haplogroups L2a/L2b or L3a. Only a single Zulu individual, namely 941, belongs to haplogroup L0, as discussed in Paragraph 4.2.29.1. The haplogroup determination process of Xhosa and Zulu individuals can therefore be initiated utilising haplogroup L2a/L2b or L3a specific sites. An ethnic status of either Xhosa or Zulu may serve as an indicator of which haplogroup informative sites to investigate first.

Based on the analysis of the current data, it would seem as if ethnicity is not a definitive indicator of the haplogroup of an individual. Genetic studies of the association between disease expression and genetic background require representatives of each haplogroup in a given population. However, viewed in the context of the fact that the current investigation comprises 10 individuals from three different ethnic groups, the data from this pilot study may serve as an indicator of the genetic relatedness of individuals from these three ethnic groups in the Southern African context.

Ethnicity is an imperfect proxy for establishing ancestral geographic origin, which in itself is an indicator of the genetic make-up of individuals (Collins, 2004). Ethnicity is defined on the grounds of social construct, and cannot indicate whether ancestral admixture occurred (Keita *et al.*, 2004). Even if genetic groups were formed that corresponded to ethnicity, each group would be internally structured on a genetic level (Tate and Goldstein, 2004) and this correlation could only be sustained if the groups were endogamous (Risch *et al.*, 2002), which becomes less likely if the groups are situated in the same geographic vicinity. If no correlation between ethnicity and genetic inference exists, the verification of self-defined ethnicity is essential, as this serves as an indicator of similar environmental exposure, culture, and other factors, as discussed in Paragraph 2.10, which plays a key role in disease expression, and provides insight into the contribution of genetics to disease expression (Tate and Goldstein, 2004).

Certain haplogroups, of the presented study, seem to be present in specific ethnic groups and absent in the rest. Self-identified ethnicity, although not an ultimate proxy for the genetic make-up of an individual, may serve as a general indicator of the haplogroup to which an individual belongs. The term "general indicator" in this context refers to self-identified ethnicity correlating correctly to genetic background in the majority of

individuals analysed. Although no ethnic group in the current investigation is definable by a haplogroup, the analysis of additional individuals will clarify the haplogroup composition of each ethnic group. This will clarify to what degree ethnic status can be utilised as an indicator of the genetic background of an individual in the Southern African context.

REFERENCES

6.1 GENERAL REFERENCES

Adcock G.J., Dennis E.S., Easteal S., Huttley G.A., Jermiin L.S., Peacock W.J. and Thorne A. Mitochondrial DNA sequences in ancient Australians: implications for modern human origins. *Proc. Natl. Acad. Sci. USA.*, **98**, 537-542, 2001.

Altschul S.F., Madden T.L., Shaffer A.A., Zhang J., Zhang Z., Miller W. and Lipman D.J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389-3402, 1997.

Ambrose S.H. Late Pleistocene human population bottlenecks, volcanic winter, and differentiation of modern humans. *J. Hum Evol.*, **34**, 623-651, 1998.

Anderson S., Bankier A.T., Barrell B.G., De Bruijn M.H.L., Coulson A.R., Drouin J., Eperon I.C., Nierlich D.P., Roe B.A., Sanger F., Schreier P.H., Smith A.J.H., Staden R. and Young I.G. Sequence and organization of the human mitochondrial genome. *Nature*, **290**, 457-465, 1981.

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.

Aquadro C.F. and Greenberg B.D. Human mitochondrial DNA variation and evolution: analysis of nucleotide sequences from seven individuals. *Genetics*, **103**, 287-312, 1983.

Baldauf S.L. Phylogeny for the faint of heart: a tutorial. *Trends Genet.*, **19**, 345-351, 2003.

Ballinger S.W. Southeast Asian mitochondrial DNA analysis reveals genetic continuity of ancient mongoloid migrations. *Genetics*, **130**, 139-152, 1992.

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**, 1150-1160, 2002.

Barbujani G. What do languages tell us about human microevolution? *Trends Ecol. Evol.*, **6**, 151-156, 1991.

Barnes W.M. PCR amplification of up to 35-kb DNA with high fidelity and high yield from λ bateriophage templates. *Proc. Natl. Acad.Sci. USA.*, **91**, 2216-2220, 1994.

Barrell B.G., Anderson S., Bankier A.T., De Bruijn M.H.L., Chen E., Coulson A.R., Drouin J., Eperon I.C., Nierlich D.P., Roe B.A., Sanger F., 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. USA.*, **77**, 3164-3166, 1980.

Bogenhagen D. and Clayton D.A. The number of mitochondrial deoxyribonucleic acid genomes in mouse L and Human HeLa cells. *J. Biol. Chem.*, **249**, 7991-7995, 1974.

Borst P. Structure and function of mitochondrial DNA. *Trends in Biochem. Sci.*, **2**, 31-34, 1977.

Breslauer K.J., Frank R., Blocker H. and Marky L.A. Predicting DNA duplex stability from the base sequence. *Proc. Natl. Acad. Sci. USA.*, **83**, 3746-3750, 1986.

Brown G.G. and Simpson M.V. Novel features of animal mtDNA evolution as shown by sequences of two rat cytochrome oxidase subunit II genes. *Proc. Natl. Acad. Sci. USA.*, **79**, 3246-3250, 1982a.

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**, 1852-1861, 1998.

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**, 33-39, 2001.

Brown W.M., George M. and Wilson A.C. Rapid evolution of animal mitochondrial DNA. *Proc. Natl. Acad. Sci. USA.*, **76**, 1967-1971, 1979.

Brown W.M. Polymorphism in mitochondrial DNA of humans as revealed by restriction endonuclease analysis. *Proc. Natl. Acad. Sci. USA.*, **77**, 3605-3609, 1980.

Brown W.M., Prager E.M., Wanc A. and Wilson A.C. Mitochondrial DNA sequences of primates: tempo and mode of evolution. *J. Mol. Evol.*, **18**, 225-239, 1982b.

Campbell P.N. and Smith A.D. Biochemistry illustrated, 3rd ed. London: Chruchill Livingstone, 69-85, 1995.

Cann R.L., Stoneking M. and Wilson A.C. Mitochondrial DNA and human evolution. *Nature*, **325**, 31-36, 1987.

Cavalli-Sforza L.L., Piazza A., Menozzi P. and Mountain J. Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data. *Proc. Natl. Acad. Sci. USA.*, **85**, 6002-6006, 1988.

Cavalli-Sforza L.L., Menozzi P. and Piazza A. The history and geography of human genes. New Jersey, USA.,Princeton University Press, 159-194, 1994.

Chen Y.S., Torroni A., Exoffier 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**, 133-149, 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 Kwe – and their genetic relationships to other African populations. *Am. J. Hum. Genet.*, **66**, 1362-1383, 2000.

Chinnery P.F., Johnson M.A., Wardell T.M., Singh-Kler R., Hayes C., Brown D.T., Taylor R.W., Bindoff L. A. and Turnbull D.M. The epidemiology of pathogenic mitochondrial DNA mutations. *Ann. Neurol.*, **48**, 188-193, 2000.

Clark K.M., Taylor R.W., Margaret A.J., Chinnery P.F., Chrzanowska-Lightowlers Z.M.A., 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**, 1330-1339, 1999.

Clayton D.A. Replication of animal mitochondrial DNA. *Cell*, **28**, 693-705, 1982.

Clayton D.A. Replication and transcription of vertebrate mitochondrial DNA. *Annu. Rev. Cell. Biol.*, **7**, 453-478, 1991.

Coble M.D., Just R.S., O'Callaghan J.E., Letmanyi I.H., Peterson C.T., Irwin J.A. and Parsons T.J. Single nucleotide polymorphisms over the entire mtDNA genome that increases the power of forensic testing in Caucasians. *Int. J. Legal Med.*, **118**, 137-146, 2004.

Collard M. and Wood B. How reliable are human phylogenetic hypotheses? *Proc. Natl. Acad. Sci. USA.*, **97**, 5003-5006, 2000.

Collins F.S. What we do and don't know about "race", "ethnicity", genetics and health at the dawn of the genome era. *Nat. Genet.*, **36**, s13-s15, 2004.

Coskun P.E. Control region mtDNA variants: longevity, climatic adaptation, and a forensic conundrum. *Proc. Natl. Acad. Sci. USA.*, **100**, 2174-2176, 2003.

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**, 409-413, 2002.

Denaro M., Blank H., Hohnson 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. USA.*, **78**, 5768-5772, 1981.

Dennis C. Error reports threaten to unravel databases of mitochondrial DNA. *Nature*, **421**, 773-774, 2003.

Derenko M.V., Grzybowski T., Malyarchuk B.A., Czarny J., Miscicka-Sliwka D. and Zakharov L.A. The presence of mitochondrial haplogroup X in Altaians from South Siberia. *Am. J. Hum. Genet.*, **69**, 237-241, 2001.

Di Rienzo A. and Wilson A.C. Branching pattern in the evolutionary tree for human mitochondrial DNA. *Proc. Natl. Acad. Sci. USA.*, **88**, 1597-1601, 1991.

Duarte C., Mauricio J., Pettitt P.B., Souto P., Trinkaus E., Van Der Plicht H. and Zilhao J. The early upper Paleolithic human skeleton from the Abrigo do Lagar Velho (Portugal) and modern human emergence in Siberia. *Proc. Natl. Acad. Sci. USA.*, **96**, 7604-7609, 1999.

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**, 229-238, 2004.

Excoffiier 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., Harding R.M., Robert R.S., Pellegrini B. and Sanchez-Mazas A. Spatial differentiation of RH and GM haplotype frequencies in Sub-Saharan Africa and its relation to linguistic affinities. *Hum. Biol.*, **63**, 273-307, 1991.

Excoffier L. Human demographic history: refining the recent African origin model. *Curr. Opin. Genet. Dev.*, **12**, 675-682, 2002.

Falconer D.S. and Mackay T.F.C. Introduction to quantitative genetics, 4th ed. Essex: Pearson education limited, 23-47, 1996.

Finnila S., Lehtonen M.S. and Majamaa K. Phylogenetic network for European mtDNA. *Am. J. Hum. Genet.*, **68**, 1475-1484, 2001.

Giles R.E., Blanc H., Cann H.M. and Wallace D.C. Maternal inheritance of human mitochondrial DNA. *Proc. Natl. Acad. Sci. USA.*, **77**, 6715-6719, 1980.

Giles R.E. and Ambrose S.H. Are we all out of Africa? *Nature*, **322**, 21-21, 1986.

Glotz C., Zwieb C. and Brimacombe R. Secondary structure of the large subunit ribosomal RNA from *Escheriichia coli, Zea mays* chloroplast, and human and mouse mitochondrial ribosomes. *Nucleic Acids Res.*, **14**, 3287-3306, 1981.

Gravel M., Melancon P. and Brakier-Gingras L. Cross-linking of Streptomycin to the 16S ribosomal RNA of *Escherihia coli. Biochemistry*, **26**, 6227-6232, 1987.

Graven L., Passarino G., Semino O., Boursot 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**, 334-345, 1995.

Greenberg B.D., Newbold J.E. and Sugino A. Intraspecific nucleotide sequence variability surrounding the origin of replication in human mitochondrial DNA. *Gene*, **21**, 33-49, 1983.

Hagelberg E., Goldman N., Lio P., Whelan S., Schiefenhovel W., Clegg J.B. and Bowden D.K. Evidence for mitochondrial DNA recombination in a human population of island Melanesia. *Proc. R. Soc. Lond. B.*, **266**, 485-492, 1999.

Harpending H.C., Sherry S.T., Rogers A.R. and Stoneking M. The genetic structure of ancient human populations. *Current Anthropology*, **34**, 483-496, 1993.

Herrnstadt C., Elson J.L., Fahy E., Preston G., Turnbull D.M., Anderson C., Ghosh S.S., Olefsky J.M., Beal 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**, 1152-1171, 2002.

Hirakawa M., Tanaka T., Hashimoto Y., Kuroda M., Takagi T. and Nakamura Y. JSNP: a database of common gene variations in the Japanese population. *Nucleic Acids Res.*, **30**, 158-162, 2002.

Hofmann S., Bezold R., Jaksch M., Obermaier-Kusser B., Mertens S., Kaufhold P., Rabl W., Kecker W. and Gerbitz K-D. Wolfram (DIDMOAD) syndrome and Leber Heteditary Optic Neuropathy (LHON) are associated with distinct mitochondrial DNA haplotypes. *Genomics*, **39**, 8-18, 1997.

Horai S. and Hayasaka K. Intraspecific nucleotide sequence differences in the major noncoding region of human mitochondrial DNA. *Am. J. Hum. Genet.*, **46**, 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. USA.*, **92**, 532-536, 1995.

Horai S., Kondo R., Nakagawa-Hattori Y., Hayashi S., Sonoda S. and Jajima K. Peopling of the Americas, founded by four major lineages of mitochondrial DNA. *Mol. Biol. Evol.*, **10**, 23-47, 1993.

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**, 285-302, 1995.

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**, 379-384, 1993.

Ingman M., Kaessmann H., Paabo S. and Gyllensten U. Mitochondrial genome variation and the origin of modern humans. *Nature*, **408**, 708-713, 2000.

Ingman M. and Gyllensten U. Mitochondrial genome variation and evolutionary history of Australian and New Guinean Aborigines. *Genome Res.*, **13**, 1600-1606, 2003.

Jobling M.A. and Tyler-Smith C. New uses for new haplotypes. *Trends. Genet.*, **16**, 356-362, 2000.

Jobling M.A. and King T.E. The distribution of Y-chromosomal haplotypes: forensic implications. *International Congress Series*, **1261**, 70-72, 2004.

Johnson M.J., Wallace D.C., Ferris S.D., Rattazzi M.C. and Cavalli-Sforza L.L. Radiation of human mitochondrial DNA types analyzed by restriction endonucleases cleavage patterns. *J. Mol. Evol.*, **19**, 255-271, 1983.

Kaneda H., Hayashi J.I., Takahama S., Taya C., Lindahl K.F. and Yonekawa H. Elimination of paternal mitochondrial DNA in intraspecifiec crosses during early mouse embryogenesis. *Proc. Natl. Acad. Sci. USA.*, **92**, 4542-4546, 1995.

Keohavong P. and Thilly W.G. Fidelity of DNA polymerases in DNA amplification. *Proc. Natl. Acad. Sci. USA.*, **86**, 9253-9257, 1989.

Kleinle S., Schneider V., Moosmann P., Brandner S., Krahenbuhl S. and Liechti-Gallati S. A novel mitochondrial $tRNA^{Phe}$ mutation inhibiting anticodon stem formation associated with a muscle disease. *Biochem. Biophys. Res. Commun.*, **247**, 112-115, 1998.

Keita S.O.Y., Kitles R.A., Royal C.D.M., Bonney G.E., Furbert-Harris P., Dunston G.M. and Rotimi C.N. Conceptualizing human variation. *Nat. Genet.*, **36**, s17-s20, 2004.

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.

Kosel S., Egensperger R., Mehraein P. and Graeber M.B. No association of mutations at nucleotide 5460 of mitochondrial NADH dehydrogenase with Alzheimer's disease. *Biochem. Biophys. Res. Commun.*, **203**, 745-749, 1994.

Krings M., Stone A., Schmitz R.W., Krainitzki H., Stoneking M. and Paabo S. Neandertal DNA sequences and the origin of modern humans. *Cell*, **90**, 19-30, 1997.

Kumar S., Tamura K., Jakobsen I.B. and Nei M. MEGA2: molecular evolutionary genetics analysis software. *Bioinformatics*, **17**, 1244-1245, 2001.

Larsson N-G. and Clayton D.A. Molecular genetic aspects of human mitochondrial disorders. *Annu. Rev. Genet.*, **29**, 151-178, 1995.

Legros F., Chatzoglou E., Frachon P., Ogier de Baulny H., Laforet P., Jardel C., Godinot C. and Lombes A. *Eur. J. Hum. Genet.*, **9**, 510-518, 2001.

Macaulay V., Richards M., Hickey E., Vega E., Cruciani F., Guida V., Scozzari R., Bonne-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**, 232-249, 1999.

Margulis L. Origin of eukaryotic cells. New Haven: Yale University Press, 1970.

Marzuki S., Noer A.S., Lertrit 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**, 139-145, 1991.

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 mitochondrial DNA variation in humans. *Proc. Natl. Acad. Sci. USA.*, **100**, 171-176, 2003.

Mitchell P. Coupling of phosphorylation to electron and hydrogen transfer by a chemi-osmotic type of mechanism. *Nature*, **191**, 144-148, 1961.

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.

Moraes C.T., Ciaci 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. *J. Clin. Invest.*, **92**, 2906-2915, 1993.

Moritz C., Dowling T.E. and Brown W.M. Evolution of animal mitochondrial DNA: relevance for population biology and systematics. *Ann. Rev. Ecol Syst.*, **18**, 269-292, 1987.

Moazed D. and Noller H.F. Interaction of antibiotics with functional sites in 16S ribosomal RNA. *Nature*, **327**, 389-394, 1987.

Mouton C. Mitochondrial genome consensus sequence for the South African Khoi-San population [dissertation]. Pretoria: North-West University, 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 Harbour Symp. Quant. Biol.*, **L1**, 263-273, 1986.

Nei M. and Kumar S. Evolutionary change of DNA sequences. In: Molecular evolution and phylogenetics. New York, Oxford University Press Inc., 33-164, 2000.

Olckers A., Prosser D., Wallace D.C., Brown M.D. and Smuts I. Mitochondrial myopathies: a South African perspective. 51st Annual meeting of the American Society of Human Genetics, San Diego, USA., Oct., 2001.

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**, 938-946, 1991a.

Ozawa T., Tanaka M., Sugiyama S., Ino H., Ohno K., Hattori K., 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**, 518-525, 1991b.

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**, 613-620, 1995.

Page R.D.M. and Holmes E.E. Inferring molecular phylogeny. In: Molecular Evolution, A Phylogenetic Approach. United Kingdom, Blackwell Science Ltd., 172-227, 1998.

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 completed sequencing: implications for the peopling of South Asia. *Am. J. Hum. Genet.*, **75**, 966-978, 2004.

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**, 420-434, 1998.

Poloni E.S., Semino O., Passarino G., Santachiara-Benerecetti A.S., Dupanloup I., Langaney A. and Excoffier L. Human genetic affinities for Y-chromosome P49a,f/*TaqI* haplotypes show strong correspondence with linguistics. *Am. J. Hum. Genet.*, **61**, 1015-1035, 1997.

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**, 289-294, 1993.

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**, 437-441, 1999.

Reich D.E. and Goldstein D.B. Genetic evidence for a Palaeolithic human population expansion in Africa. *Proc. Natl. Acad. Sci. USA.*, **95**, 8119-8123, 1998.

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., Golge M., Dimitrov D., Hill E., Bradley D., Romano V., Cali F., Vonda G., Demaine A., Papiha S., Triantaphyllidis C., Stefanescu G., Hatina J., Belledi M., Di Rienzo A., Novelletto A., Oppenheim A., Norby S., Al-Zaheri N., Santachiara-Benerecetti S., Scozzari 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.

Richards M., Rengo C., Cruciani F., Gratrix F., Wilson J.F., Scozzari R., Macaulay V. and Torroni A. Extensive female-mediated gene flow from Sub-Saharan Africa into near eastern Arab populations. *Am. J. Hum. Genet.* **72**, 1058-1064, 2003.

Rieder M.J., Yaylor S.L., Tobe B.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**, 967-973, 1998.

Risch N., Burchard E., Ziv E. and Tang H. Categorization of humans in biomedical research: genes, race and disease. *Genome Biol.*, **3**, 2007.1-2007.12, 2002.

Rose G., Passarino G., Carrieri G., Altomare K., Greco V., Bertolini S., Bonafe M., Franceschi C. and De Benedictis G. Paradoxes in longevity: sequence analysis of mtDNA haplogroup J in centenarians. *Eur. J. Hum. Genet.*, **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**, 1161-1178, 2001.

Rosser Z.H., Xerjal T., Hurles M.E., Adojaan M., Alavantic D., Amorim A., Amos W., Armenteros M., Arroyo E., Barbuani G., Beckman G., Beckman L., Bertranpetit J., Bosch E., Bradley D.G., Brede G., Cooper G., Corte-Real H.B.S.M., de Knijff P., Decorte R., Dubrova Y.E., Evgrafov O., Gilissen A., Glisic S., Golge M., Hill E.W., Jeziorowska A., Kalaydjieva L., Kayser M., Kivisild T., Kravhenko S.A., Krumina A., Kucinskas V., Lavinha J., Livshits L.A., Malaspina P., Maria S., McElreavey K., Meitinger T.A., Mikelsaar A-V., Mitchell R.J., Nafa K., Nicholson J., Norby S., Pandya A., Parik J., Patsalis P.C., Pereira L., Peterlin B., Pielberg G., Prata M.J., Previdere C., Roewer L., Rootsi S., Rubinsztein D.C., Saillard J., Santos F.R., Stefanescu G., Sykes B.C., Tolun A., Villems R., Tyler-Smith C. and Jobling M.A. Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. *Am. J. Hum. Genet.*, **67**, 1526-1543, 2000.

Ruiz-Pesini E., Lapena A-C., Diez-Sanchez C., Perez-Martos A., Montoya J., Alvarez E., Diaz M., Uries A., Montoro L., Lopez-Perez M.J. and Enriquez J.A. Human mtDNA haplogroups associated with high or reduced spermatozoa motility. *Am. J. Hum. Genet.*, **67**, 682-696, 2000.

Ruiz-Pesini E., Mishmar D., Brandon M., Procacio V. and Wallace D.C. Effects of purifying and adaptive selection on regional variation in human mtDNA. *Science*, **303**, 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**, 535-543, 2004.

Russo C.A.M., Takezaki N. and Nei M. Efficiencies of different genes and different tree-building methods in recovering a known vertebrate phylogeny. *Mol. Biol. Evol.*, **13**, 525-536, 1996.

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.

Ruvolo M., Pan D., Zehr S., Goldberg T., Disotell T.R. and Von Dornum M. Gene trees and hominoid phylogeny. *Proc. Natl. Acad. Sci. USA.*, **91**, 8900-8904, 1994.

Salas A., Richards M., De la Fe T., Lareu M-V., Sobrino B., Sanchez-Diz P., Macaulay V. and Carracedo A. The making of the African mtDNA landscape. *Am. J. Hum. Genet.*, **71**, 1082-111, 2002.

Sanger F., Nicklen S. and Coulson A.R. DNA sequencing with chain-termination inhibitors. *Proc. Natl. Acad. Sci. USA.*, **74**, 5463-5467, 1977.

Scaglia F., Vogel H., Hawkins E.P., Vladutiu G.D., Liu L-L. and Wong L-J.C. Novel homoplasmic mutation in the mitochondrial $tRNA^{Tyr}$ gene associated with atypical mitochondrial cytopathy presenting with focal segmental glomerulosclerosis. *Am. J. Med. Genet.*, **123**, 172-178, 2003.

Scozzari R., Torroni A., Semino O., Cruciani F., Spendini G. and Santachiara Benerecetti S.A. Genetic studies in Cameroon: mitochondrial DNA polymorphisms in Bamileke. *Hum. Biol.*, **66**, 1-12, 1994.

Seielstad M.T., Minch E. and Cavalli-Sforza L.L. Genetic evidence for a higher female migration rate in humans. *Nat. Genet.*, **20**, 278-280, 1998.

Serre D., Langaney A., Chech M., Teschler-Nicola M., Paunovic M., Mennecier P., Hofreiter M., Possnert G. and Paabo S. No evidence of Neandertal mtDNA contribution to early modern humans. *Public Library of Science Biology*, **2**, 313-317, 2004.

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**, 171-184, 1993.

Silva W.A., Bonatto S.L., Holanda A.J., Ribeiro-dos-Santos A.K., Paizao B.M., Goldman G.H., Abe-Sandes K., Rodriguez-Delfin L., Barbosa M., Paco-Larson M.L., Petzl-Erler M.L., Valente V., Santos S.E.B.

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**, 187-192, 2002.

Silva W.A., Bonatto S.L., Holanda A.J., Ribeiro-dos-Santos A.K., Paizao B.M., Goldman G.H., Abe-Sandes K., Rodriguez-Delfin L., Barbosa M., Paco-Larson M.L., Petzl-Erler M.L., Valente V., Santos S.E.B. and Zago M.A. Correction: mitochondrial DNA variation in Amerindians. *Am. J. Hum. Genet.*, **72**, 1346-1348, 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.

Spelbrink J.N. Replication, repair, and recombination of mitochondrial DNA. In: Genetics of Mitochondrial Diseases. Holt I.J., editor. Oxford: Oxford University Press, 3-25, 2003.

Sternberg D., Danan C., Lombes A., Laforet P., Girodon E., Goossens M. and Amselem S. Exhaustive scanning approach to screen all the mitochondrial tRNA genes for mutations and its application to the investigation of 35 independent patients with mitochondrial disorders. *Hum. Mol. Genet.*, **7**, 33-42, 1998.

Sternberg D., Chatzoglou E., Laforet P., Fayat G., Jardel C., Blondy P., Fardeau M., Amselem S., Eymard B. and Lombes A. Mitochondrial DNA transfer RNA gene sequence variations in patients with mitochondrial disorders. *Brain*, **124**, 984-994, 2001.

Stringer C.B. and Andrews P. Genetic and fossil evidence for the origin of modern humans. *Science*, **239**, 1263-1268, 1988.

Stringer C.B., Grun R., Schwarcz H.P. and Goldberg P. ESR dates for the hominid burial site of Es Skhul in Israel. *Nature*, **338**, 756-758, 1989.

Stumpf M.P.H. Haplotype diversity and the block structure of linkage disequilibrium. *Trends Genet.*, **18**, 226-228, 2002.

Sudoyo H., Suryadi H., Lertrit P., Pramoonjago P., Lyrawati D. and Marzuki S. Asian-specific mtDNA backgrounds associated with the primary G1178A mutation of Leber's hereditary optic neuropathy. *Hum. Genet.*, **47**, 594-604, 2002.

Taanman J-W. The mitochondrial genome: structure, transcription, translation and replication. *Biochem. Biophys. Acta.*, **1410**, 103-123, 1999.

Tate S.K. and Goldstein D.B. Will tomorrow's medicines work for everyone? *Nat. Genet.*, **36**, s34-s42, 2004.

Taylor R.W., Taylor G.A., Durham S.E. and Turnbull D.M. The determination of complete human mitochondrial DNA sequences in single cells: implications for the study of somatic mitochondrial DNA point mutations. *Nucleic Acids Res.*, **29**, e74, 2001.

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**, 1786-1796, 2003.

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**, 253-255, 1996.

Tishkoff S.A. and Williams S.M. Genetic analysis of African populations: human evolution and complex disease. *Genetics*, **3**, 611-621, 2002.

Tishkoff S.A. and Verrelli B.C. Role of evolutionary history on haplotype block structure in the human genome: implications for disease mapping. *Curr. Opin. Genet. Dev.*, **13**, 569-575, 2003.

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**, 563-590, 1993a.

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**, 591-608, 1993b.

Torroni A., Lott M.T., Cabell M.F., Chen Y-C., 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 DNA analsis 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., Varrantes 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. USA.*, **91**, 1158-1162, 1994c.

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. mtDNA haplogroups in human populations and disease studies. *J. Cult. Heritage*, **1**, 33-34, 2000.

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.

Valnot I., Kassis J., Chretien D., de Lonlay P., Parfait B., Nunnich A., Kachaner J., Rustin P. and Rotig A. A mitochondrial cytochrome *b* mutation but no mutations of nuclearly encoded subunits in ubiquinol cytochrome *c* reductase (complex III) deficiency. *Hum. Genet.*, **104**, 460-466, 1999.

Van Brummelen A.C. Molecular analysis of the mitochondrial genome in South African patients with suspected mitochondrial disorders [dissertation]. Pretoria: University of the North-West, 45-46, 2003.

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. USA.*, **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. Fundamentals of Biochemistry. 2nd ed. New York: John Wiley and Sons Inc., 563-598, 1995.

Walker J.E. The NADH: ubiquinone oxidoreductase (complex I) of respiratory chains. *Q. Rev. Biophys.*, **25**, 253-324, 1992.

Wallin E. On the nature of mitochondria. *American Journal of Anatomy*, **30**, 203-229, 1922.

Wallace D.C. Diseases of the mitochondrial DNA. *Annu. Rev. Biochem.*, **61**, 1175-1212, 1992.

Wallace D.C. Mitochondrial DNA sequence variation in human evolution and disease. *Proc. Natl. Acad. Sci. USA.*, **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., Brown M.D. and Lott M.T. Mitochondrial DNA variation in human evolution and disease. *Gene*, **238**, 211-230, 1999.

Wallace D.C. Centre for Molecular and Mitochondrial Medicine and Genetics, University of California, Irvine, USA., personal communication, 2003.

Watson E., Bauer K., Aman R., Weiss G., Von Haeseler A. and Paabo S. mtDNA sequence diversity in Africa. *Am. J. Hum. Genet.*, **59**, 437-444, 1996.

Watson E., Porster P., Richards M. and Bandelt H-J. Mitochondrial footprints of human expansions in Africa. *Am. J. Hum. Genet.*, **61**, 691-704, 1997.

Wilson A.C., Ochman H. and Prager E.M. Molecular time scale for evolution. *Trends Genet.*, **3**, 241-247, 1987.

Wolpoff M.H. Multiregional evolution: the fossil alternative to Eden. Mellars P. and Stringer C. The human revolution, Edinburgh: Edinburgh University Press, 62-108, 1972.

Wolstenholme D.R. Animal mitochondrial DNA: structure and evolution. *Int. Rev. Cytol.*, **141**, 179-216, 1992.

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**, 529-542, 1987.

Zhao H., Li R., Wang Q., Yan Q., Deng J-H., Han D., Bai Y., Young W-Y. and Guan M-X. Maternally inherited aminoglycoside-induced and nonsyndromic deafness is associated with the novel C1494T mutation in the mitochondrial 12S rRNA gene in a large Chinese family. *Am. J. Hum. Genet.*, **74**, 139-152, 2004.

6.2 ELECTRONIC REFERENCES

Abu-Amero K.K. and Bosley T. MITOMAP mtDNA Sequence Data, http://www.mitomap.org/cgi-bin/mitomap/tbl15gen.pl#20040210013, 2004.

Hall T. A. BioEdit, Biological Sequence Alignment Editor, http://www.mbio.ncsu.edu/BioEdit/bioedit/html, 1999.

Kalendar R. Fast PCR v.3.3.112. "FastPCR, PCR primer design, DNA and protein tools, repeats and own database searches program", Institute of Biotechnology, University of Helsinki, Finland, http://www.biocenter.helsinki.fi/bi/bare-1_html/fastpcr.html, 2004.

Maca-Meyer N., Gonzalez A.M., Larruga J.M., Flores C. and Cabrera V.M. Major genomic mitochondrial lineages delineate early human expansions. BMC genetics, **2**, http://www.biomedcentral.com/1471-2156/2/13, 2001.

MITOMAP, A Human Mitochondrial Genome Database, Center for Molecular Medicine, Emory University, Atlanta, GA, USA., http://www.gen.emory.edu/mitomap.html, 2004.

NCBI, National Centre of Biotechnology Information, National Institutes of Health, Genbank accession number NC001807.3, http://www.nbi.nlm.nih.gov, 2004a.

NCBI, National Centre of Biotechnology Information, National Institutes of Health, http://www.nbi.nlm.nih.gov, 2004b.

OLIGONUCLEOTIDE PROPERTIES CALCULATOR, http://www.basic.nwu.edu/biotools/oligoalc.html, 2004.

Polyak K. and Vogelstein B. MITOMAP mtDNA Sequence Data, http://www.mitomap.org/cgi-bin/mitomap/tbl15gen.pl#19990316025, 1999.

Snejina M. and Naumova E. MITOMAP mtDNA Sequence Data, http://www.mitomap.org/cgi-bin/mitomap/tbl15gen.pl#20030925001, 2003.

APPENDIX A

WHOLE mtDNA GENOME ALTERATIONS RELATIVE TO THE RCRS

The full mitochondrial genome sequence of 10 individuals from three ethnic groups, namely Xhosa individuals (921, 926, 995), Zulu individuals (934, 939, 941) and Tswana individuals (962, 964, 972 and 976) were screened for alterations relative to the RCRS. A complete list of alterations detected within all individuals analysed is presented in Table A.1. Alterations detected within protein coding regions are listed as either synonymous or non-synonymous. Non-synonymous alterations are indicated with three letter amino acid notations and an arrow indicating the direction of change. Alterations occurring within rRNA, tRNA or non-coding regions, are listed as transitions or transversions.

Table A.1: Sequence differences between individuals analysed and the RCRS

Alteration	Individuals										Locus	Alteration Type	Reference
	921	926	995	934	939	941	962	964	972	976			
A73G	1	1	1	1	1	1	0	1	1	1	D-loop	Transition	Aquadro and Greenberg (1983)
A93G	0	0	0	0	0	0	1	0	0	0	D-loop	Transition	Graven *et al.* (1995)
A95C	0	0	0	0	0	0	1	0	0	0	D-loop	Transversion	Graven *et al.* (1995)
G143A	0	0	0	0	0	0	0	0	0	1	D-loop	Transition	Graven *et al.* (1995)
T146C	0	1	0	0	1	1	0	1	1	1	D-loop	Transition	Aquadro and Greenberg (1983)
C150T	1	0	0	1	1	0	0	0	0	0	D-loop	Transition	Aquadro and Greenberg (1983)
T152C	0	1	1	1	1	1	0	0	1	1	D-loop	Transition	Aquadro and Greenberg (1983)
C182T	0	0	0	0	1	0	0	0	0	0	D-loop	Transition	Aquadro and Greenberg (1983)
G185A	0	0	0	0	0	0	1	0	0	0	D-loop	Transition	Aquadro and Greenberg (1983)
A189G	0	0	1	1	0	0	1	0	0	0	D-loop	Transition	Aquadro and Greenberg (1983)
T195C	1	1	0	0	1	1	0	1	1	1	D-loop	Transition	Aquadro and Greenberg (1983)
C198T	0	1	0	0	1	0	0	0	0	0	D-loop	Transition	Graven *et al.* (1995)
T199C	0	0	0	0	0	0	0	1	0	0	D-loop	Transition	Ozawa *et al.* (1991b)
A200G	0	0	0	1	0	0	0	0	0	1	D-loop	Transition	Aquadro and Greenberg (1983)
T204C	0	0	0	0	1	0	0	0	0	0	D-loop	Transition	Graven *et al.* (1995)
T236C	0	0	0	0	0	0	1	0	0	0	D-loop	Transition	Aquadro and Greenberg (1983)
G247A	0	0	1	0	0	1	1	1	1	0	D-loop	Transition	Aquadro and Greenberg (1983)
A263G	1	1	1	1	1	0	1	0	0	1	D-loop	Transition	Andrews *et al.* (1999)
311-315insC	1	1	1	1	1	1	1	1	1	1	D-loop	Insertion	Graven *et al.* (1995); Andrews *et al.* (1999)
498delC	0	0	0	0	0	1	0	1	1	0	D-loop	Deletion	Vigilant *et al.* (1991)
522-523delCA	0	0	0	0	0	0	0	0	1	1	D-loop	Insertion	Kleinle *et al.* (1998)

Table A.1: Continued...

Alteration	Individuals										Locus	Alteration Type	Reference
	921	926	995	934	939	941	962	964	972	976			
568-577insPolyC	0	0	0	0	0	0	0	0	1	0	D-loop	Insertion	Not yet reported
G709A	0	0	0	0	1	0	0	0	0	0	12S rRNA	Transition	Ozawa *et al.* (1991)
G719A	0	0	0	0	0	1	0	1	1	0	12S rRNA	Transition	Maca-Meyer *et al.* (2001)
G769A	0	1	0	0	1	1	1	1	1	1	12S rRNA	Transition	Moraes *et al.* (1993)
T825A	0	0	0	0	0	1	0	1	1	0	12S rRNA	Transition	Prezant *et al.* (1993)
G1018A	0	1	1	0	1	1	1	1	1	1	12S rRNA	Transition	Prezant *et al.* (1993)
C1048T	0	0	1	0	0	1	1	1	1	0	12S rRNA	Transition	Moraes *et al.* (1993)
T1243C	0	0	0	0	0	0	0	1	0	0	12S rRNA	Transition	Taylor *et al.* (2001)
A1438G	1	1	1	1	1	0	1	0	1	0	12S rRNA	Transition	Ozawa *et al.* (1991); Andrews *et al.* (1999)
G1442A	0	0	0	0	1	0	0	0	0	0	12S rRNA	Transition	Palanichamy *et al.* (2004)
C1706T	0	0	0	0	1	0	0	0	0	0	16S rRNA	Transition	Herrnstadt *et al.* (2002)
G1709A	0	1	0	0	0	0	0	0	0	0	16S rRNA	Transition	Zhoa *et al.* (2004)
A2245G	0	0	1	0	0	0	1	0	0	0	16S rRNA	Transition	Moraes *et al.* (1993)
C2332T	0	0	0	0	1	0	0	0	0	0	16S rRNA	Transition	Herrnstadt *et al.* (2002)
T2352C	1	0	1	1	0	0	0	0	0	0	16S rRNA	Transition	Chen *et al.* (1995)
A2358G	0	0	0	0	1	0	0	0	0	0	16S rRNA	Transition	Herrnstadt *et al.* (2002)
T2416C	0	1	0	0	1	0	0	0	0	1	16S rRNA	Transition	Herrnstadt *et al.* (2002)
T2483C	1	0	0	0	0	0	0	0	0	0	16S rRNA	Transition	Herrnstadt *et al.* (2002)
A2706G	1	1	1	1	1	0	1	0	0	1	16S rRNA	Transition	Ozawa *et al.* (1991b)
G2758A	0	0	0	0	0	1	1	1	1	0	16S rRNA	Transition	Moraes *et al.* (1993)
C2789T	0	1	0	0	0	0	0	0	0	1	16S rRNA	Transition	Rose *et al.* (2001)
T2885C	0	0	0	0	0	1	1	1	1	0	16S rRNA	Transition	Moraes *et al.* (1993)
T2887C	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**0**	**16S rRNA**	**Transition**	**Not yet reported**
G3010A	0	0	0	0	0	0	0	0	0	1	16S rRNA	Transition	Marzuki *et al.* (1991)
C3157T	**0**	**0**	**0**	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**16S rRNA**	**Transition**	**Not yet reported**
G3277A	1	0	0	0	0	0	0	0	0	0	$tRNA^{Leu}$	Transition	Sternberg *et al.* (2001)
G3438A	0	0	0	0	0	1	0	1	1	0	ND1	Synonymous	Thomas *et al.* (1996)
C3516A	0	0	1	0	0	1	1	1	1	0	ND1	Synonymous	Herrnstadt *et al.* (2002)
C3594T	0	1	0	0	1	1	1	1	1	1	ND1	Synonymous	Chen *et al.* (1995)
T3618C	**0**	**0**	**0**	**0**	**0**	**1**	**0**	**0**	**1**	**0**	**ND1**	**Synonymous**	**Not yet reported**
A3756G	0	0	0	0	0	1	0	1	1	0	ND1	Synonymous	Mishmar *et al.* (2002)
A4104G	0	1	1	0	1	1	1	1	1	1	ND1	Synonymous	Ruppert *et al.* (2004)
A4158G	0	0	0	0	1	0	0	0	0	0	ND1	Synonymous	Herrnstadt *et al.* (2002)
T4232C	0	0	0	0	0	1	0	1	1	0	ND1	Ile → Thr	Mishmar *et al.* (2002)
C4312T	0	0	0	0	0	1	1	1	1	0	$tRNA^{Ile}$	Transition	Moraes *et al.* (1993)
T4370C	0	0	0	0	1	0	0	0	0	0	$tRNA^{Glu}$	Transition	Anderson *et al.* (1981)
T4586C	0	0	1	0	0	0	1	0	0	0	ND2	Synonymous	Ingman *et al.* (2000)
A4767G	0	0	0	0	1	0	0	0	0	0	ND2	Trp → Val	Herrnstadt *et al.* (2002)
A4769G	1	1	1	1	1	1	1	1	1	1	ND2	Synonymous	Andrews *et al.* (1999)
C5027T	0	0	0	0	1	0	0	0	0	0	ND2	Synonymous	Herrnstadt *et al.* (2002)
T5096C	0	0	0	0	0	0	1	0	0	0	ND2	Synonymous	Ingman *et al.* (2000)
G5231A	0	0	0	0	0	0	1	0	0	0	ND2	Synonymous	Ozawa *et al.* (1991a)
C5331A	0	0	0	0	1	0	0	0	0	0	ND2	Leu → Ile	Herrnstadt *et al.* (2002)
T5442C	0	0	0	0	0	1	1	1	1	0	ND2	Phe → Leu	Ozawa *et al.* (1991a)
G5460A	0	0	0	0	0	0	1	0	0	0	ND2	Ala → Thr	Kosel *et al.* (1994)
T5553C	0	0	0	0	0	0	0	1	0	0	$tRNA^{Trp}$	Transition	Herrnstadt *et al.* (2002)
C5603T	0	0	0	0	0	0	1	0	0	0	$tRNA^{Ala}$	Transition	Moraes *et al.* (1993)

Table A.1: Continued...

Alteration	Individuals										Locus	Alteration Type	Reference
	921	926	995	934	939	941	962	964	972	976			
T5814C	0	0	0	0	1	0	0	0	0	0	$tRNA^{Cys}$	Transition	Herrnstadt *et al.* (2002)
A5894G	0	0	0	0	0	0	0	1	0	0	NC	Transition	Herrnstadt *et al.* (2002)
C5911T	0	0	0	0	0	0	1	0	0	0	COI	Ala → Val	Ingman *et al.* (2000)
T6185C	0	0	1	0	0	1	1	1	1	0	COI	Synonymous	Ingman *et al.* (2000)
T6221C	0	0	1	1	0	0	0	0	0	0	COI	Synonymous	Marzuki *et al.* (1991)
A6266G	0	0	0	0	0	1	0	1	1	0	COI	Synonymous	Herrnstadt *et al.* (2002)
C6587T	0	0	0	1	0	1	0	0	0	0	COI	Synonymous	Marzuki *et al.* (1991)
T6614C	0	0	0	0	1	0	0	0	0	0	COI	**Synonymous**	**Not yet reported**
A6663G	0	0	0	0	0	0	0	0	0	1	COI	Ile → Val	Scaglia *et al.* (2003)
C6713T	0	0	0	0	1	0	0	0	0	0	COI	Synonymous	Herrnstadt *et al.* (2002)
A6806G	0	0	0	0	1	0	0	0	0	0	COI	**Synonymous**	**Not yet reported**
T6815C	0	0	1	0	0	0	0	1	1	0	COI	Synonymous	Mishmar *et al.* (2002)
A6891G	0	0	0	0	0	0	0	1	0	0	COI	Ser → Gly	Tanaka *et al.* (2004)
C7028T	1	1	1	1	1	1	1	1	1	1	COI	Synonymous	Macaulay *et al.* (1999)
A7146G	0	0	1	0	0	1	1	1	1	0	COI	Synonymous	Polyak *et al.* (1999)
T7175C	0	1	0	0	0	0	0	0	0	1	COI	Synonymous	Silva *et al.* (2002)
C7256T	0	1	0	0	1	1	1	1	1	1	COI	Synonymous	Silva *et al.* (2002)
C7274T	0	1	0	0	0	0	0	0	0	1	COI	Synonymous	Silva *et al.* (2002)
T7283C	0	0	0	0	0	1	0	0	1	0	COI	Synonymous	Finnila *et al.* (2001)
G7521A	0	1	0	0	1	1	1	1	1	1	$tRNA^{Asp}$	Transition	Silva *et al.* (2002)
T7624A	0	0	0	0	1	0	0	0	0	0	COII	Synonymous	Herrnstadt *et al.* (2002)
T7741C	0	0	0	1	0	0	0	0	0	0	COII	Synonymous	Coble *et al.* (2004)
A7771G	0	1	0	0	0	0	0	0	0	1	COII	Synonymous	Coble *et al.* (2004)
C8080T	0	0	0	0	1	0	0	0	0	0	COII	Synonymous	Herrnstadt *et al.* (2002)
G8113A	0	0	0	0	0	1	0	1	1	0	COII	Gly → Glu	Ruvolo *et al.* (1993)
G8152A	0	0	0	0	0	1	0	1	1	0	COII	Synonymous	Ruvolo *et al.* (1993)
A8191G	0	0	0	0	0	0	1	0	0	0	COII	Synonymous	Maca-Meyer *et al.* (2001)
G8206A	0	1	0	0	1	0	0	0	0	1	COII	Synonymous	Ozawa *et al.* (1995)
G8251A	0	0	0	0	0	1	0	1	1	0	COII	Synonymous	Ruvolo *et al.* (1993)
T8383C	0	0	0	0	0	0	0	0	1	0	ATPase8	Synonymous	Herrnstadt *et al.* (2002)
G8387A	0	0	0	0	1	0	0	0	0	0	ATPase8	Synonymous	Herrnstadt *et al.* (2002)
C8428T	0	0	0	0	0	0	1	0	0	0	ATPase8	Synonymous	Moraes *et al.*(1993)
A8460G	0	0	1	0	0	0	0	0	0	0	ATPase8	Asn → Ser	Ingman *et al.* (2000)
C8468T	0	0	1	0	0	1	1	1	1	0	ATPase8	Synonymous	Moraes *et al.*(1993)
T8503C	0	0	0	0	1	0	0	1	0	0	ATPase8	**Synonymous**	**Not yet reported**
A8566G	0	0	0	0	0	0	1	0	0	0	ATPase8	Synonymous	Ingman *et al.* (2000)
A8566G	0	0	0	0	0	0	1	0	0	0	ATPase6	Ile → Val	Ingman *et al.* (2000)
A8577G	0	0	1	0	0	0	0	0	0	0	ATPase6	Synonymous	Abu-Amera *et al.* (2004)
C8655T	0	0	0	0	0	1	1	1	1	0	ATPase6	Synonymous	Ingman *et al.* (2000)
A8701G	1	1	1	1	1	1	1	1	1	1	ATPase6	Thr → Arg	Ozawa *et al.* (1991a)
C8703T	0	0	0	1	0	0	0	0	0	0	ATPase6	Synonymous	Ingman *et al.* (2000)
G8790A	0	0	0	0	1	0	0	0	0	0	ATPase6	Synonymous	Taylor *et al.* (2003)
T8793C	0	0	0	0	0	0	0	1	0	0	ATPase6	Synonymous	Ozawa *et al.* (1991b)
A8860G	1	1	1	1	1	1	1	1	1	1	ATPase6	Thr → Arg	Andrews *et al.* (1999)
C9042T	0	0	1	0	0	1	1	1	1	0	ATPase6	Synonymous	Ingman *et al.* (2000)
T9111C	0	0	0	0	0	1	0	0	0	0	ATPase6	Synonymous	Herrnstadt *et al.* (2002)
A9221G	0	1	0	0	1	0	0	0	0	1	COIII	Synonymous	Silva *et al.* (2002)
G9300A	0	0	0	1	0	0	0	0	0	0	COIII	Synonymous	Snejina *et al.* (2003)

Table A.1: Continued...

Alteration	Individuals										Locus	Alteration Type	Reference
	921	926	995	934	939	941	962	964	972	976			
A9327G	0	0	0	1	0	0	0	0	0	0	COIII	Synonymous	Coble *et al.* (2004)
A9347G	0	0	1	0	0	1	1	1	1	0	COIII	Synonymous	Herrnstadt *et al.* (2002)
A9350G	0	0	0	0	1	0	0	0	0	0	COIII	Synonymous	**Not yet reported**
A9377G	1	0	0	0	0	0	0	0	0	0	COIII	Synonymous	Ozawa *et al.* (1991)
T9540C	1	1	1	1	1	1	1	1	1	1	COIII	Synonymous	Marzuki *et al.* (1991)
T9637G	0	0	0	0	0	0	1	1	1	0	COIII	Ile → Ser	Not yet reported
G9755A	0	0	0	0	0	1	1	1	1	0	COIII	Synonymous	Herrnstadt *et al.* (2002)
C9818T	0	0	0	0	0	0	1	0	0	0	COIII	Synonymous	Herrnstadt *et al.* (2002)
T9833C	0	0	0	0	0	0	0	1	0	0	COIII	Synonymous	Kong *et al.* (2003)
T9950C	0	0	0	0	0	0	0	1	0	0	COIII	Synonymous	Silva *et al.* (2003)
T10115C	0	1	0	0	1	1	0	0	0	1	ND3	Synonymous	Silva *et al.* (2003)
G10143A	0	1	0	0	0	0	0	0	0	0	ND3	Gly → Ser	Prezant *et al.* (1993)
A10398G	1	1	1	1	1	1	1	1	1	1	ND3	Thr → Ala	Herrnstadt *et al.* (2002)
G10589A	0	0	0	0	0	1	1	1	1	0	ND4L	Synonymous	Howell *et al.* (1995)
C10664T	0	0	1	0	0	1	1	1	1	0	ND4L	Synonymous	Herrnstadt *et al.* (2002)
G10688C	0	0	0	0	0	1	1	1	1	0	ND4L	Synonymous	Ingman *et al.* (2000)
T10810C	0	0	0	0	0	1	1	1	1	0	ND4	Synonymous	Finnila *et al.* (2001)
A10819G	1	0	1	1	0	0	0	0	0	0	ND4	Synonymous	Marzuki *et al.* (1991)
T10873C	1	1	1	1	1	1	1	1	1	1	ND4	Synonymous	Marzuki *et al.* (1991)
T10915C	0	0	1	0	0	1	1	1	1	0	ND4	Synonymous	Ozawa *et al.* (1991a)
C10920T	0	0	0	0	0	1	0	0	0	0	ND4	Pro → Leu	Ingman *et al.* (2000)
G11176A	0	0	1	0	0	1	1	0	0	0	ND4	Synonymous	Herrnstadt *et al.* (2002)
T11260C	0	0	0	0	0	0	0	1	0	0	ND4	Synonymous	Herrnstadt *et al.* (2002)
A11641G	0	0	1	0	0	0	1	0	0	0	ND4	Synonymous	Ingman *et al.* (2000)
G11719A	1	1	1	1	1	1	1	1	1	1	ND4	Synonymous	Silva *et al.* (2001)
G11914A	0	1	0	0	0	1	1	1	1	1	ND4	Synonymous	Silva *et al.* (2001)
T11944C	0	1	0	0	1	0	0	0	0	1	ND4	Synonymous	Ingman *et al.* (2000)
G12007A	0	0	0	0	0	1	1	1	1	0	ND4	Synonymous	Moraes *et al.* (1993)
T12121C	0	0	0	0	0	1	0	1	1	0	ND4	Synonymous	Mishmar *et al.* (2002)
G12127A	0	0	0	0	0	0	1	0	0	0	ND4	Synonymous	Herrnstadt *et al.* (2002)
A12142G	0	0	0	0	0	0	0	1	0	0	$tRNA^{His}$	Transition	Mouton (2003)
G12236A	0	0	0	0	1	0	0	0	0	0	$tRNA^{Ser}$	Transition	Prezant *et al.* (1993)
C12348T	0	0	0	0	0	0	0	1	0	0	ND5	Synonymous	**Not yet reported**
G12406A	1	0	0	0	0	0	0	0	0	0	ND5	Val → Ile	Herrnstadt *et al.* (2002)
C12436T	0	0	0	0	0	1	0	0	0	0	ND5	His → Asn	**Not yet reported**
A12693G	0	1	0	0	0	0	0	0	0	1	ND5	Synonymous	Silva *et al.* (2002)
C12705T	1	1	1	1	1	1	1	1	1	1	ND5	Synonymous	Marzuki *et al.* (1991)
A12720G	0	0	1	0	0	0	1	1	1	0	ND5	Synonymous	Brown *et al.* (2001)
T12738C	0	0	0	1	0	0	0	0	0	0	ND5	Synonymous	Mishmar *et al.* (2002)
C12798T	0	0	0	0	0	0	1	0	0	0	ND5	Synonymous	**Not yet reported**
A12948G	0	0	0	0	1	0	0	0	0	0	ND5	Synonymous	Herrnstadt *et al.* (2002)
A13105G	0	0	1	0	0	1	1	1	1	0	ND5	Ile → Val	Prezant *et al.* (1993)
A13276G	0	0	0	0	0	1	1	1	1	0	ND5	Met → Val	Ingman *et al.* (2000)
C13506T	0	0	1	0	0	1	1	1	1	0	ND5	Synonymous	Herrnstadt *et al.* (2002)
G13590A	0	1	0	0	1	0	0	0	0	1	ND5	Synonymous	Ozawa *et al.* (1995)
C13650T	0	1	1	0	1	1	1	1	1	1	ND5	Synonymous	Silva *et al.* (2002)
G13708A	0	0	0	0	0	0	0	0	0	1	ND5	Ala → Thr	Torroni *et al.* (1996)
G13759A	0	0	0	0	0	1	0	1	1	0	ND5	Ala → Thr	Rieder *et al.* (1998)

Table A.1: Continued...

Alteration	Individuals										Locus	Alteration Type	Reference
	921	926	995	934	939	941	962	964	972	976			
A13803G	0	1	0	0	0	0	0	0	0	1	ND5	Synonymous	Silva *et al.* (2002)
A13966G	0	0	0	0	1	0	0	0	0	0	ND5	Thr → Ala	Finnila *et al.* (2001)
A14007G	0	0	0	0	0	0	1	0	0	0	ND5	Synonymous	Ingman *et al.* (2000)
A14059G	0	0	0	0	1	0	0	0	0	0	ND5	Ile →Val	Mishmar *et al.* (2002)
A14152G	0	0	1	1	0	0	0	0	0	0	ND6	Synonymous	Marzuki *et al.* (1991)
A14176G	**0**	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**ND6**	**Synonymous**	**Not yet reported**
T14182C	0	0	1	0	0	0	0	0	0	0	ND6	Synonymous	Huoponen *et al.* (1993)
T14212C	1	0	1	1	0	0	0	0	0	0	ND6	Synonymous	Marzuki *et al.* (1991)
T14308C	0	0	1	0	0	0	1	0	0	0	ND6	Synonymous	Sudoyo *et al.* (2002)
C14315T	0	0	0	0	0	1	0	0	1	0	ND6	Thr → Ile	Mouton (2003)
C14407T	**0**	**0**	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**0**	**ND6**	**Synonymous**	**Not yet reported**
A14566G	0	1	0	0	0	0	0	0	0	1	ND6	Synonymous	Silva *et al.* (2003)
C14659T	0	0	0	0	0	1	0	0	1	0	ND6	Leu → Phe	Mouton (2003)
A14755G	0	0	1	0	0	0	0	0	0	0	cyt *b*	Synonymous	Herrnstadt *et al.* (2002)
C14766T	1	1	1	1	1	1	1	1	1	1	cyt *b*	Synonymous	Andrews *et al.* (1999)
G14905A	1	0	0	0	0	0	0	0	0	0	cyt *b*	Synonymous	Howell *et al.* (1995)
A14926G	0	0	1	0	0	0	0	0	0	0	cyt *b*	Synonymous	Crimi *et al.* (2002)
G15110A	0	0	0	0	1	0	0	0	0	0	cyt *b*	Ala → Thr	Legros *et al.* (2001)
C15136T	0	0	0	0	0	0	1	0	0	0	cyt *b*	Synonymous	Ingman *et al.* (2000)
G15301A	1	1	0	1	1	0	0	0	0	1	cyt *b*	Synonymous	Ozawa *et al.* (1991a)
A15326G	1	1	1	1	1	1	1	1	1	1	cyt *b*	Thr → Ala	Andrews *et al.* (1999)
C15337T	0	0	0	0	0	1	0	0	0	0	cyt *b*	Synonymous	Finnila *et al.* (2001)
G15431A	0	0	0	0	0	0	1	0	0	0	cyt *b*	Ala → Thr	Ingman *et al.* (2000)
G15466A	0	0	0	0	0	1	0	1	1	0	cyt *b*	Synonymous	Mishmar *et al.* (2002)
T15670C	0	0	0	1	0	0	0	0	0	0	cyt *b*	Synonymous	Valnot *et al.* (1999)
A15692G	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**0**	**cyt *b***	**Synonymous**	**Not yet reported**
C15735T	0	1	0	0	0	0	0	0	0	0	cyt *b*	Ala → Val	Coble *et al.* (2004)
T15784C	0	1	0	0	0	0	0	0	0	1	cyt *b*	Synonymous	Herrnstadt *et al.* (2002)
T15792C	0	0	0	0	0	0	0	0	0	1	cyt *b*	Ile → Thr	Finnila *et al.* (2001)
G15930A	0	0	0	0	0	1	0	1	1	0	$tRNA^{Thr}$	Synonymous	Mishmar *et al.* (2002)
T15941C	0	0	0	0	0	1	0	1	1	0	$tRNA^{Thr}$	Synonymous	Mishmar *et al.* (2002)
T15942C	0	0	0	1	0	0	0	0	0	0	$tRNA^{Thr}$	Synonymous	Rose *et al.* (2001)
G16129A	0	0	0	0	1	1	1	1	1	0	D-loop	Transition	Finnila *et al.* (2001)
C16148T	0	0	0	0	0	0	1	0	0	0	D-loop	Transition	Aquadro and Greenberg (1983)
C16168T	0	0	0	0	0	0	1	0	0	0	D-loop	Transition	Di Rienzo and Wilson (1991)
T16172C	1	0	0	0	0	0	1	0	0	0	D-loop	Transition	Greenberg *et al.* (1983)
C16176T	0	0	0	1	0	0	0	0	0	0	D-loop	Transition	Maca-Meyer *et al.* (2001)
16184-16194insCC	1	0	0	0	0	0	0	0	0	0	D-loop	Insertion	Horai and Hayasaka (1990)
C16187T	0	0	1	0	0	1	1	1	1	0	D-loop	Transition	Horai and Hayasaka (1990)
C16188G	0	0	0	0	0	0	1	0	0	0	D-loop	Transversion	Horai and Hayasaka (1990)
C16188A	0	0	1	0	0	0	0	0	0	0	D-loop	Transition	Horai and Hayasaka (1990)
T16189C	1	1	1	0	0	1	1	1	1	0	D-loop	Transition	Horai and Hayasaka (1990)
G16213A	0	0	0	0	1	0	0	0	0	0	D-loop	Transition	Horai and Hayasaka (1990)
C16223T	1	1	1	1	1	1	1	1	1	1	D-loop	Transition	Horai and Hayasaka (1990)
A16230G	0	0	1	0	0	1	1	1	0	0	D-loop	Transition	Aquadro and Greenberg (1983)
C16234T	0	0	0	0	0	0	0	1	0	0	D-loop	Transition	Di Rienzo and Wilson (1991)
C16239T	0	0	1	0	0	1	0	0	1	0	D-loop	Transition	Maca-Meyer *et al.* (2001)

Table A.1: Continued...

Alteration	Individuals										Locus	Alteration Type	Reference
	921	926	995	934	939	941	962	964	972	976			
T16243C	0	0	0	0	0	1	0	1	1	0	D-loop	Transition	Horai and Hayasaka (1990)
C16266G	0	0	0	0	0	0	0	1	0	0	D-loop	Transversion	Horai and Hayasaka (1990)
C16278T	0	1	0	0	1	0	1	0	0	1	D-loop	Transition	Horai and Hayasaka (1990)
C16294T	0	1	0	0	0	1	0	0	1	1	D-loop	Transition	Di Rienzo and Wilson (1991)
A16309G	0	1	0	0	0	0	0	0	0	1	D-loop	Transition	Horai and Hayasaka (1990)
T16311C	0	0	0	0	1	1	1	1	1	0	D-loop	Transition	Horai and Hayasaka (1990)
C16320T	1	0	0	0	0	1	1	0	0	0	D-loop	Transition	Horai and Hayasaka (1990)
T16327C	0	0	0	1	0	0	0	0	0	0	D-loop	Transition	Horai *et al.* (1993)
C16354T	0	0	0	0	1	0	0	0	0	0	D-loop	Transition	Richards *et al.* (2000)
G16390A	0	1	0	0	1	0	0	0	0	1	D-loop	Transition	Mishmar *et al.* (2002)
T16519C	1	0	1	0	0	1	0	1	1	0	D-loop	Transition	Aquadro and Greenberg (1983)

Nucleotide substitution format includes nucleotide present in the RCRS, followed by the nucleotide position and then by the nucleotide detected. 1 = presence of alteration, 0 = absence of alteration. The format of insertions and deletions include the nucleotide position, followed by an insertion or deletion event, followed by the nucleotide (s) inserted or deleted. ins = insertion. del = deletion. The RCRS was utilised as the reference sequence (Andrews *et al.*, 1999). Alterations in red = novel, alterations in blue = utilised for haplogroup analysis. NC = the non-coding region between the tRNA tyrosine and cytochrome *c* oxidase subunit genes. tRNA = transfer RNA. $tRNA^{Leu}$, $tRNA^{Ile}$, $tRNA^{Glu}$, $tRNA^{Trp}$, $tRNA^{Ala}$, $tRNA^{Cys}$, $tRNA^{Asp}$, $tRNA^{His}$, $tRNA^{Ser}$ and $tRNA^{Thr}$ = genes encoding tRNA leucine, isoleucine, glutamic acid, tryptophan, alanine, cysteine, aspartic acid, histidine, serine and threonine. rRNA = ribosomal RNA. ND1, ND2, ND3, ND4L, ND4, ND5 and ND6 = genes encoding subunits 1, 2, 3, 4L, 4, 5 and 6 of the NADH dehydrogenase. COI, COII and COIII = genes encoding subunits I, II and III of cytochrome *c* oxidase. ATPase6 and ATPase8 = genes encoding subunits 6 and 8 of ATP synthase. cyt *b* = gene encoding cytochrome *b*. Ile = isoleucine, Thr = threonine, Trp = tryptophan, Val = valine, Phe = phenylalanine, Ala = alanine, Gly = glycine, Glu = glutamic acid, Ser = serine, Arg = arginine, Pro = proline, Met = methionine.

APPENDIX B

DISTRIBUTION OF FRAGMENT TYPES UTILISED FOR SEQUENCING

Fragments generated via the short PCR strategy were utilised as templates in specific sequencing reactions. Regions of the mitochondrial genome that were not successfully sequenced utilising long fragments, were sequenced via the short PCR strategy. The type of fragment that was sequenced per individual is listed in Table B.1.

Table B.1: Fragments of individuals sequenced utilising short and long PCR products

Fragment	Individuals									
	921	926	995	934	939	941	962	964	972	976
1	1	1	0	1	0	0	1	1	1	1
2	1	1	1	1	1	1	1	1	0	1
3	1	1	1	1	1	1	0	1	1	1
4	0	1	0	0	0	0	0	0	00	0
5	1	1	0	0	1	0	1	1	1	1
6	1	1	1	1	1	0	1	0	1	1
7	1	1	1	1	1	1	1	1	1	1
8	1	1	1	1	1	1	1	1	1	1
9	1	1	1	0	1	1	1	1	1	1
10	0	1	0	0	0	0	1	1	0	1
11	1	1	1	1	1	1	1	1	0	1
12	1	1	0	1	1	1	1	1	0	1
13	0	1	0	0	0	0	1	1	1	1
14	0	1	1	0	0	0	1	1	1	1
15	0	1	1	0	0	0	1	1	0	1
16	1	1	0	1	1	0	1	1	1	1
17	1	1	1	1	1	0	1	1	1	1
18	1	1	1	1	1	0	1	1	1	1
19	0	1	1	0	0	0	1	1	1	1
20	0	1	1	0	0	0	1	1	1	1
21	0	0	0	0	0	0	0	0	0	0
22	0	1	0	0	0	0	1	1	1	1
23	0	0	0	0	0	0	0	0	0	0
24	0	0	0	0	0	0	1	1	0	0
25	0	0	0	0	0	0	0	0	0	0
26	1	1	1	1	1	1	1	1	1	1
27	1	1	1	1	1	1	1	1	1	1
28	0	1	1	0	0	0	1	1	1	1
29	0	0	1	1	0	1	1	1	1	1
30	0	1	1	1	1	1	1	1	1	1
31	1	1	1	1	1	0	1	1	1	1
32	0	1	0	0	0	0	0	0	0	1

Fragment numbers correspond to the forward and reverse primer numbers utilised for amplification. 1 indicates that the fragment of a specific individual was amplified via the long PCR strategy. Grey coloured blocks indicate fragments sequenced utilising templates generated via the short PCR strategy.

APPENDIX C

ALTERATIONS CONTAINED IN SEQUENCE MOTIFS

Each combination of individuals that shared the same alteration was termed a motif. Sequence motifs were compiled from alterations that were detected in more than one individual in the current study. Only motifs that occurred more than once were considered. The number of times a specific motif occurred, after all alterations were considered, was counted, and is presented in Table C.1.

Table C.1: Alterations in specific motifs

Alteration	[1]Individuals										Motif name	[2]Individuals harbouring motif (#)	Number of alterations in motif
	921	926	995	934	939	941	962	964	972	976			
	IND #1	IND #2	IND #3	IND #4	IND #5	IND #6	IND #7	IND #8	IND #9	IND #0			
T6221C	0	0	1	1	0	0	0	0	0	0	i	34	2
A14152G	0	0	1	1	0	0	0	0	0	0		34	
G16129A	0	0	0	0	1	1	1	1	1	0	ii	56789	2
T16311C	0	0	0	0	1	1	1	1	1	0		56789	
A263G	1	1	1	1	1	0	1	0	0	1	iii	1234570	2
A2706G	1	1	1	1	1	0	1	0	0	1		1234570	
T2352C	1	0	1	1	0	0	0	0	0	0	iv	134	3
A10819G	1	0	1	1	0	0	0	0	0	0		134	
T14212C	1	0	1	1	0	0	0	0	0	0		134	
G1018A	0	1	1	0	1	1	1	1	1	1	v	23567890	3
A4104G	0	1	1	0	1	1	1	1	1	1		23567890	
C13650T	0	1	1	0	1	1	1	1	1	1		23567890	
T3618C	0	0	0	0	0	1	0	0	1	0	vi	69	4
T7283C	0	0	0	0	0	1	0	0	1	0		69	
C14315T	0	0	0	0	0	1	0	0	1	0		69	
C14659T	0	0	0	0	0	1	0	0	1	0		69	
A2245G	0	0	1	0	0	0	1	0	0	0	vii	37	4
T4586C	0	0	1	0	0	0	1	0	0	0		37	
A11640G	0	0	1	0	0	0	1	0	0	0		37	
T14308C	0	0	1	0	0	0	1	0	0	0		37	
G769A	0	1	0	0	1	1	1	1	1	1	viii	2567890	4
C3594T	0	1	0	0	1	1	1	1	1	1		2567890	
C7256T	0	1	0	0	1	1	1	1	1	1		2567890	
G7521A	0	1	0	0	1	1	1	1	1	1		2567890	
T2416C	0	1	0	0	1	0	0	0	0	1	ix	250	6
G8206A	0	1	0	0	1	0	0	0	0	1		250	
A9221G	0	1	0	0	1	0	0	0	0	1		250	
T11944C	0	1	0	0	1	0	0	0	0	1		250	
G13590A	0	1	0	0	1	0	0	0	0	1		250	
G16390A	0	1	0	0	1	0	0	0	0	1		250	

Table C.1: Continued...

Alteration	[1]Individuals										Motif name	[2]Individuals harbouring motif (#)	Number of alterations in motif
	921	926	995	934	939	941	962	964	972	976			
	IND #1	IND #2	IND #3	IND #4	IND #5	IND #6	IND #7	IND #8	IND #9	IND #0			
C2789T	0	1	0	0	0	0	0	0	0	1		20	
T7175C	0	1	0	0	0	0	0	0	0	1		20	
C7274C	0	1	0	0	0	0	0	0	0	1		20	
A7771G	0	1	0	0	0	0	0	0	0	1		20	9
A12693G	0	1	0	0	0	0	0	0	0	1	x	20	
A13803G	0	1	0	0	0	0	0	0	0	1		20	
A14566G	0	1	0	0	0	0	0	0	0	1		20	
T15784C	0	1	0	0	0	0	0	0	0	1		20	
A16309G	0	1	0	0	0	0	0	0	0	1		20	
G2758A	0	0	0	0	0	1	1	1	1	0		6789	
T2885C	0	0	0	0	0	1	1	1	1	0		6789	
C4312T	0	0	0	0	0	1	1	1	1	0		6789	
C8655T	0	0	0	0	0	1	1	1	1	0		6789	
G9755A	0	0	0	0	0	1	1	1	1	0	xi	6789	10
G10589A	0	0	0	0	0	1	1	1	1	0		6789	
G10688C	0	0	0	0	0	1	1	1	1	0		6789	
T10810C	0	0	0	0	0	1	1	1	1	0		6789	
G12007A	0	0	0	0	0	1	1	1	1	0		6789	
A13276G	0	0	0	0	0	1	1	1	1	0		6789	
G247A	0	0	1	0	0	1	1	1	1	0		36789	
C1048T	0	0	1	0	0	1	1	1	1	0		36789	
C3516A	0	0	1	0	0	1	1	1	1	0		36789	
T6185C	0	0	1	0	0	1	1	1	1	0		36789	
A7146G	0	0	1	0	0	1	1	1	1	0		36789	
C8468T	0	0	1	0	0	1	1	1	1	0		36789	
C9042T	0	0	1	0	0	1	1	1	1	0	xii	36789	13
A9347G	0	0	1	0	0	1	1	1	1	0		36789	
C10664T	0	0	1	0	0	1	1	1	1	0		36789	
T10915C	0	0	1	0	0	1	1	1	1	0		36789	
A13105G	0	0	1	0	0	1	1	1	1	0		36789	
C13506T	0	0	1	0	0	1	1	1	1	0		36789	
C16187T	0	0	0	0	0	1	1	1	1	0		36789	
311-315insC	1	1	1	1	1	1	1	1	1	1		1234567890	
A4769G	1	1	1	1	1	1	1	1	1	1		1234567890	
C7028C	1	1	1	1	1	1	1	1	1	1		1234567890	
A8701G	1	1	1	1	1	1	1	1	1	1		1234567890	
A8860G	1	1	1	1	1	1	1	1	1	1		1234567890	
T9540C	1	1	1	1	1	1	1	1	1	1		1234567890	
A10398G	1	1	1	1	1	1	1	1	1	1	xiii	1234567890	13
T10873C	1	1	1	1	1	1	1	1	1	1		1234567890	
G11719A	1	1	1	1	1	1	1	1	1	1		1234567890	
C12705T	1	1	1	1	1	1	1	1	1	1		1234567890	
C14766T	1	1	1	1	1	1	1	1	1	1		1234567890	
A15326G	1	1	1	1	1	1	1	1	1	1		1234567890	
C16223T	1	1	1	1	1	1	1	1	1	1		1234567890	

Table C.1: Continued...

Alteration	[1]Individuals										Motif name	[2]Individuals harbouring motif (#)	Number of alterations in motif
	921	926	995	934	939	941	962	964	972	976			
	IND #1	IND #2	IND #3	IND #4	IND #5	IND #6	IND #7	IND #8	IND #9	IND #0			
498delC	0	0	0	0	0	1	0	1	1	0		689	
G719A	0	0	0	0	0	1	0	1	1	0		689	
T825A	0	0	0	0	0	1	0	1	1	0		689	
A3756G	0	0	0	0	0	1	0	1	1	0		689	
T4232C	0	0	0	0	0	1	0	1	1	0		689	
A6266G	0	0	0	0	0	1	0	1	1	0		689	
G8113A	0	0	0	0	0	1	0	1	1	0		689	
G8152A	0	0	0	0	0	1	0	1	1	0	xiv	689	15
G8251G	0	0	0	0	0	1	0	1	1	0		689	
T12121C	0	0	0	0	0	1	0	1	1	0		689	
G13759A	0	0	0	0	0	1	0	1	1	0		689	
G15466A	0	0	0	0	0	1	0	1	1	0		689	
G15930A	0	0	0	0	0	1	0	1	1	0		689	
T15941C	0	0	0	0	0	1	0	1	1	0		689	
T16243C	0	0	0	0	0	1	0	1	1	0		689	

Nucleotide substitution format includes nucleotide present in the RCRS, followed by the nucleotide position and then by the nucleotide detected. The format of insertions and deletions include the nucleotide position, followed by an insertion or deletion event, followed by the nucleotide (s) inserted or deleted. ins = insertion. del = deletion. The RCRS was utilised as the reference sequence (Andrews *et al.*, 1999). 1 = distribution of alterations between individuals, with 1 indicating the presence of the specific alteration, and 0 indicating the absence of the alteration. 2 = individuals contained within a motif, with IND #1 = 921, IND #2 = 326, IND #3 = 995, IND #4 = 934, IND #5 = 939, IND #6 = 941, IND #7 = 962, IND #8 = 964, IND #9 = 972 and IND #0 = 976. The 522insC and 523insA were excluded, as both insertions together represent a single mutational event.