Mitochondrial genome consensus sequence for the South African Khoi-San population

BY

CHRISTA MOUTON, B.Sc.(Hons)

Dissertation submitted for the degree Magister Scientiae in Biochemistry at the Potchefstroomse Universiteit vir Christelike Hoër Onderwys

SUPERVISOR: Professor Antonel Olckers
Centre for Genome Research,
Potchefstroom University for Christian Higher Education

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

December 2003

Mitochondriale genoomkonsensus-volgorde vir die Suid-Afrikaanse Khoi-San populasie

DEUR

CHRISTA MOUTON, B.Sc.(Hons)

Verhandeling ingedien vir die graad Magister Scientiae in Biochemie by die Potchefstroomse Universiteit vir Christelike Hoër Onderwys

STUDIELEIER: Professor Antonel Olckers
Sentrum vir Genomiese Navorsing,
Potchefstroomse Universiteit vir Christelike Hoër Onderwys

MEDESTUDIELEIER: Dokter Izelle Smuts
Departement Pediatrie, Fakulteit Gesondheidswetenskappe, Universiteit van Pretoria

Desember 2003

TO MY PARENTS

ABSTRACT

Maternal inheritance and the absence of recombination have contributed to mitochondrial deoxyribonucleic acid (mtDNA) being utilised to study human evolution. This, together with an increased mutation rate in mtDNA, provides information about the most recent common ancestor of modern humans.

Previous studies suggested that Africa harbours the highest mtDNA diversity, supporting an out-of-Africa hypothesis for modern human evolution. From subsequent studies it was suggested that the Khoi-San population, in particularly the !Kung, cluster at the deepest root of the global phylogenetic tree.

The Cambridge reference sequence is used worldwide in mitochondrial studies as a reference. However, recent studies have observed discrepancies from this sequence, which were confirmed by reanalysis.

During this investigation the complete mitochondrial sequences of 13 !Kung individuals were determined. From phylogenetic analyses their clustering in the African L0-lineage was revealed. The evolutionary rate of the derived sequences was investigated through statistical analysis and the hypothesis of neutral evolution was rejected. Pairwise nucleotide distribution suggested that sequences representing haplogroups L0, L1 and L2 are examples of populations that were of stable population size for a long time. However, L3 was suggested to have been subjected to population expansion, in support of the out-of-Africa theory of evolution.

From the comparative analysis of the 13 !Kung sequences with an L0-specific haplogroup tree it was observed that the 13 individuals clustered in two main groups. Ten individuals were added to one branch of the phylogenetic tree, revealing further branching, while three individuals were added to the terminal branches of another tree topology. A consensus sequence was derived from the 13 Khoi-San sequences, which was 99.25% similar to each of the sequences. This sequence could be utilised to investigate evolution of the mitochondrial genome over time as well as to evaluate the pathogenicity of mutations in patients.

OPSOMMING

Moederlike oorerwing asook die afwesigheid van rekombinasie het daartoe bygedra dat mitochondriale deoksieribonukleïensuur (mtDNS) gebruik word om menslike evolusie te bestudeer. Dit, tesame met 'n verhoogde mutasietempo in mtDNS, verskaf inligting aangaande die mees onlangse gemeenskaplike voorouer van die moderne mens.

Uit vorige studies spruit die veronderstelling dat Afrika die hoogste mtDNS-diversiteit huisves. Dit ondersteun die vanuit-Afrika teorie vir die evolusie van die moderne mens. Verdere studies het daarop gedui dat die Khoi-San populasie, in besonder die !Kung, in die diepste wortel van die globale filogenetiese boom groepeer.

Die Cambridge-volgorde word wêreldwyd as verwysing in mitochondriale studies gebruik. Teenstrydighede met die verwysingsvolgorde is egter waargeneem en bevestig deur herhaalde analises.

Die volledige mtDNS-volgordes van 13 !Kung-individue is in hierdie studie bepaal. Filogeneties groepeer dit in die Afrika-L0-stamboom. Die tempo van evolusie is statisties bereken vir die afgeleide volgordes, wat daarop gedui het dat die hipotese van neutrale evolusie verwerp word. Gepaarde nukleotiedverspreiding dui daarop dat volgordes wat verteenwoordigend is van haplogroepe L0, L1 en L2 voorbeelde is van populasies wat stabiele populasiegroottes gehandhaaf het oor 'n lang tydperk. Dit blyk dat L3 blootgestel was aan populasieuitbreiding en dit ondersteun die vanuit-Afrika teorie.

Die 13 !Kung-individue groepeer in twee groepe indien die volgordes met 'n L0-spesifieke haplogroepboom vergelyk word. Tien individue is by een tak van die filogenetiese boom gevoeg, wat verdere vertakking teweeg gebring het, terwyl drie individue tot die eindpunte van 'n ander boom-topologie gevoeg is. 'n Konsensusvolgorde is bepaal vanaf 13 Khoi-San-volgordes en stem 99.25% ooreen met die onderskeie volgordes. Hierdie volgorde kan gebruik word om die evolusie van die mitochondriale genoom oor tyd te bestudeer asook om die patogenisiteit van mutasies in pasiënte te evalueer.

TABLE OF CONTENTS

CHAPTER THREE

CHAPTER FOUR

CHAPTER FIVE

REFERENCES

APPENDIX A

APPENDIX B

APPENDIX C

LIST OF ABBREVIATIONS

Abbreviations and symbols are listed in alphabetical order.

α	alpha
A and a	adenine (in DNA sequence)
A	alanine (in nucleotide sequence)
Alu I	restriction endonuclease isolated from *Arthrobacter luteus*, with recognition site 5'-AG′CT-3'
ADP	adenosine diphosphate
ATP	adenosine triphosphate
ATPase	ATP synthase
ATPase 6 / ATP 6	ATPase subunit 6
ATPase 8 / ATP 8	ATPase subunit 8
Ava II	restriction endonuclease isolated from *Anabaena variabilis*, with recognition site 5'-G′G(A/T)CC-3'
β	beta
Bam HI	restriction endonuclease isolated from *Bacillus amyloliquefaciens* H, with recognition site 5'- G′GATCC-3'
bp	base pairs
Bst NI	restriction endonuclease isolated from *Bacillus stearothermophilus* N, with recognition site 5'- CC′(A/T)GG-3'
C	molar concentration of oligonuclotide (in Equation 2)
°C	degrees Celsius
C and c	cytosine (in DNA sequence)
ca.	circa: approximately
CCA	characteristic of all tRNA 3'-ends which is either transcribed from the DNA or added after transcription
CGR	Centre for Genome Research
CI	conservation index
Clustal X	Clustering analysis for multiple sequence and profile alignments
CoA	coenzyme A
CO_2	carbon dioxide
CO I-III	cytochrome *c* oxidase subunits
CR	control region
CRS	Cambridge reference sequence
Cs	cesium
CSB	conserved sequence blocks
Cu	copper
cyt	cytochrome
cyt b	cytochrome b
Da	daltons
Dde I	restriction endonuclease isolated from *Desulfovibrio desulfuricans*, with recognition site 5'-C′TNAG-3'

Del	deletion
D-loop	displacement loop
DMSO	dimethyl sulphoxide: C_2H_6SO
DNA	deoxyribonucleic acid
DnaSP	DNA Sequence Polymorphism
dNTP	deoxynucleotide triphosphate
ddNTP	dideoxynucleotide triphosphate
η	eta
e^-	electron
EDTA	ethylenediamine tetra-acetic acid: $C_{10}H_{16}N_2O_8$
EF	elongation factors
e.g.	*exempli gratia:* Latin abbreviation for "for example"
et al.	*et altera*: Latin abbreviation for "and others"
etc.	*et cetera*: Latin abbreviation for "and so forth"
EtBr	ethidium bromide: $C_{21}H_{20}BrN_3$
EtOH	ethanol
ETC	electron transport chain
FAD	flavin adenine dinucleotide
$FADH_2$	reduced flavin adenine dinucleotide
Fe-S	iron-sulphur protein
FMN	flavin mononucleotide
Fnu DII	restriction endonuclease isolated from *Fusobacterium nucleatum*, with recognition site 5'-CG′CG-3'
γ	gamma
g	grams
G and g	guanine (in DNA sequence)
GDP	guanosine diphosphate
Genbank	Genbank®[1] : United States repository of DNA sequence information
GTP	guanosine triphosphate
gDNA	genomic DNA
H^+	proton
Hae II	restriction endonuclease isolated from *Haemophilus aegyptius*, with recognition site 5'-RGCGC′Y-3'
Hae III	restriction endonuclease isolated from *Haemophilus aegyptius*, with recognition site 5'-GG′CC-3'
Heme	consists of protoporphyrin 1X and an iron atom
Heme a	derived from haeme and contains a 15-carbon isoprenoid chain on a modified vinyl group and a formyl group in place of one of the methyls
Heme a_3	the other haeme centre of cytochrome c oxidase
Heme b_L	a haeme derivative with a low (L) reduction potential of -0.100 V and a wavelength of maximum absorbance at 566 nm
Heme b_H	a haeme derivative with a high (H) reduction potential of +0.050 V and a wavelength of maximum absorbance at 562 nm
Heme c_1	a haeme group with covalently attached cysteine residues
HCl	hydrochloric acid

[1] Genbank® is a registered trademark of the National Institutes of Health, Bethesda, MD, U.S.A.

H. erectus	*Homo erectus*
HeLa	cervical cancer cells from Henrietta Lacks that are utilised in research labs
Hha I	restriction endonuclease isolated from *Haemophilus haemolyticus*, with recognition site 5'-C′GCG-3'
Hinc II	restriction endonuclease isolated from *Haemophilus influenzae*, with recognition site 5'-GTY′RAC-3'
Hinf I	restriction endonuclease isolated from *Haemophilus influenzae* Rf., with recognition site 5'-G′ANTC-3'
Hpa I	restriction endonuclease isolated from *Haemophilus parainfluenzae*, with recognition site 5'-GTT′AAC-3'
Hpa II	restriction endonuclease isolated from *Haemophilus parainfluenzae*, with recognition site 5'-C′CGG-3'
H. sapiens	*Homo sapiens*
H-strand	heavy strand
HSP	H-strand promoter
H_2O	water
IMM	inner mitochondrial membrane
Ins	insertion
IT_{H1}	upstream transcription initiation site of the H strand
IT_{H2}	downstream transcription initiation site of the H-strand
IT_L	transcription initiation site of the L-strand
k	considers the frequency of the substitutions
k	kilo: 10^3
k*a*	non-synonymous nucleotide substitution
kb	kilobase pairs
KCl	potassium chloride
kDa	kilodalton
k*s*	synonymous nucleotide substitution
L^{CUN}	leucine encoded by cytosine, uracil and any nucleotide
L^{UUR}	leucine encoded by two uracils and a purine
Leu	leucine
LHON	Leber's hereditary optic neuropathy
L-strand	light strand
LSP	L-strand promoter
Lys	lysine
µ	micro: 10^{-6}
µg	micrograms
µl	microlitres
µm	micrometres
µM	micromolar
M	ionic strength in mole per litre (in Equation 2)
MAMMAG	Centre for Molecular and Mitochondrial Medicine and Genetics
Mbo I	restriction endonuclease isolated from *Moraxella bovis*, with recognition site 5'-C′CGG-3'
MEGA	Molecular Evolutionary Genetics Analysis software
m	milli: 10^{-3}
ml	millilitres

mM	millimolar
$MgCl_2$	magnesium chloride
MP	maximum parsimony
MRCA	most recent common ancestor
mRNA	messenger RNA
Msp I	restriction endonuclease isolated from *Moraxella* species, with recognition site 5'- `GATC`-3'
mtDNA	mitochondrial DNA
mtEF-G, -Ts, -Tu	mitochondrial elongation factors G, -Ts and -Tu
mtIF-2	mitochondrial initiation factor 2
mtTERM	mitochondrial termination factor
mtTFA	mitochondrial transcription factor A
`N` and `N`	nucleotide: any of the four bases in DNA sequence
n	length of the primer (in Equation 2)
Na_2EDTA	disodium EDTA: $C_{10}H_{14}N_2Na_2O_8.2H_2O$
n	nano: 10^{-9}
NAD^+	nicotinamide adenine dinucleotide
NADH	reduced nicotinamide adenine dinucleotide
NADH-UQ reductase	NADH Coenzyme Q reductase
ND 1-6	NADH dehydrogenase subunits
ND4L	NADH dehydrogenase subunit 4, located on the L-strand
nDNA	nuclear DNA
ng	nanograms
NJ	neighbour-joining
nm	nanometres
O_2	oxygen
O_H	origin of H-strand synthesis
O_L	origin of L-strand synthesis
OMM	outer mitochondrial membrane
OXPHOS	oxidative phosphorylation
PCR	polymerase chain reaction
pH	defined as the negative logarithm of the hydrogen ion concentration: $pH = - \log_{10} [H^+]$
Pi	inorganic phosphate
%	percentage
Π	pi
p	pico: 10^{-12}
pmol	picomole
polyA	poly adenine
P value	probability value that determines the percentage of chance that the statistical result might be obtained randomly
ρ	rho
R	`A` or `G`: purine
RCRS	revised CRS
RE	restriction enzyme
RF	replacement mutation frequency
RFLP	restriction fragment length polymorphism

R-loop	precursor RNA primer that exists as a RNA-DNA hybrid and is involved in mitochondrial replication
RNA	ribonucleic acid
rRNA	ribosomal RNA
Rsa I	restriction endonuclease isolated from *Rhodopseudomonas sphaeroides*, with recognition site 5'-GT′AC-3'
S	number of segregating sites
S (AGY)	serine encoded by arginine, guanine and a pyrimidine
SD	standard deviation
S^{UCN}	serine encoded by uracil, cytosine and any nucleotide
θ	theta
T and t	thymine (in DNA sequence)
T_a	annealing temperature
Taq I	restriction endonuclease isolated from *Thermus aquaticus* YTI, with recognition site 5'-T′CGA-3'
Taq DNA polymerase	deoxynucleosidetriphosphate: DNA deoxynucleotidyltransferase, EC2.7.7.7, thermostable enzyme isolated from *Thermus aquaticus* BM, recombinant (*E. coli*)
TAS	termination associated sequence
TBE	Tris®[1]-borate EDTA buffer: 89.15 mM Tris® (pH 8.0), 88.95 mM boric acid, 2.498 mM Na_2EDTA
TFAM	mitochondrial transcription factor A
TIM	translocase of the IMM
T_m	melting temperature
TOM	translocase of the OMM
Tris®	tris(hydroxymethyl)aminomethan: 2-amino-2-(hydroxymethyl)-1,3-propanediol: $C_4H_{11}NO_3$
Tris-HCl	2-amino-2-(hydroxymethyl)-1,3-propanediol hydrochloride: $C_4H_{11}NO_3.H_2O$
Triton®[2] X-100	octylphenolpoly(ethylene-glycolether)$_n$: $C_{34}H_{62}O_{11}$, for n = 10
tRNA	transfer RNA
$tRNA^{Leu(UUR)}$	transfer RNA coding for leucine with anticodon UUR
$tRNA^{Lys}$	transfer RNA coding for lysine
$tRNA^{Phe}$	transfer RNA coding for phenylalanine
$tRNA^{Pro}$	transfer RNA coding for proline
$tRNA^{Thr}$	transfer RNA coding for threonine
UQ	coenzyme Q
UQH_2	reduced coenzyme Q or ubiquinol
U.S.A.	United States of America
UUR	anticodon encoded by two uracils and a purine
vs	versus
w/v	weight per volume
x *g*	times gravitational force
Y	C or T: pyrimidine
YBP	years before present

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

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

LIST OF GRAPHS

ACKNOWLEDGEMENTS

This achievement was made possible by the input of numerous people and institutions. I would therefore like to express my sincere gratitude to:

The **Khoi-San people** who participated in a previous study (Chen *et al.*, 2000), from whom a subset was included in the present study. To my supervisor, **Prof. Antonel Olckers**, for giving me the opportunity to work on these valuable samples and for being more than a supervisor, also an inspiration. Without her encouragement, patience and leadership I would not have been able to reach my highest level of potential and performance. Also for creating opportunities that no, or very few, students ever have, which were life- and career-changing. My co-supervisor, **Dr. Izelle Smuts**, for her encouragement, willingness to help and clinical expertise. **Prof. Doug Wallace** for the unforgettable opportunity, funding and support to work in his laboratory at the University of California at Irvine. The entire MAMMAG team, in particular **Arsen Akopyan, Pinar Coskun, Grant MacGregor, Diana Moise, Sam Schriner, Vaidya Subramaniam,** as well as **Don Cole, Cheri Seifert** and **Nadja Dvorkin** for embracing me with their friendship and for making my stay in the United States one of the best experiences ever. **Katrina Waymire** for helping and teaching me the skills of cell culturing. To **Dan Mishmar** and **Eduardo Ruiz-Pesini** for their invaluable assistance and advice and for taking me, a foreigner, under their wing. Thank you for always making me smile even when there seemed to be no light at the end of the tunnel.

The Centre for Genome Research for providing us with an environment where we could practice science at the highest possible level and for financial support by means of a post-graduate bursary. Equipment and financial resources were made available by **DNAbiotec (Pty) Ltd**, without which this study would not have succeeded. **Potchefstroom University for Christian Higher Education** for creating this unique environment with the commercial world, working closely together with academia and for allowing us to participate in both.

My sincere gratitude goes to the members of the Centre for Genome Research for all their support, friendship and encouragement during the year. To **Annelize van der Merwe** and **Marco Alessandrini** for their outstanding mentorship, helpful comments and patience.

Wayne Towers, **Tumi Semete**, **Desire Hart** and **Jake Darby** for lending a helpful hand wherever they could. To my M.Sc. colleagues, **Tharina van Brummelen** and **Madeleine Wessels**, for all your support, encouragement and good spirits, even when times were stressed and difficult.

William, for being an invaluable friend and for always being supportive and interested even during the months that we were oceans apart. To my **parents**, **sister** and **brother**, for always believing in me. Without their support I would not have been able to achieve this highlight of my life. To the Lord who has given me strength to exceed even my own expectations and whose blessings carried me through.

CHAPTER ONE

INTRODUCTION

Mitochondria play an essential role in the energy production and cellular metabolism of a cell (Borst, 1977). This organelle also houses various biochemically integrated pathways that oxidise carbohydrates, fats and proteins to carbon dioxide (CO_2) and water (H_2O). The released energy is transferred to adenosine triphosphate (ATP), which serves as a readily available source of energy in the cell (Scholte, 1988; Garrett and Grisham, 1999). It is therefore necessary that organs and tissue with high energy demands, such as the brain, skeletal muscle and the heart, contain high amounts of mitochondria.

The mitochondrion bears resemblance to the genetic system of prokaryotes (Borst, 1977). This leads to the endosymbiotic hypothesis, which suggests that the ancestors of mitochondria were free-living bacteria that developed an obligatory symbiotic life with primitive eukaryotic cells (Borst, 1977). Evidence supporting this hypothesis includes the circular structure of the mitochondrial genome, its extranuclear location and the absence of chromosomal organisation (Gray, 1993).

The mitochondrial genome is 16,569 base pairs in length. It contains 37 genes, which encodes two ribosomal ribonucleic acids (rRNAs), 22 transfer RNAs (tRNAs) and 13 polypeptides. The mitochondrion has certain unique characteristics. According to Giles *et al.* (1980) these include its exclusive inheritance via the female lineage, implying that a single female's offspring could have similar mitochondrial deoxyribonucleic acid (mtDNA). In addition, mitochondria have a repair mechanism that is not as sophisticated as that of nuclear DNA (nDNA) as described by Bauer *et al.* (1999). This, together with exposure to oxygen radicals, released from the respiratory chain, contributes to an increased mutation rate of mtDNA that is up to 20 times faster (Wallace *et al.*, 1987) than that of nDNA. The uniqueness of this organelle is discussed further in Chapter two.

The mtDNA mutation rate correlates to the origin and dispersion of human populations. Mitochondrial DNA polymorphisms have accumulated over time as women migrated out of Africa to other continents (Wallace, 1995). This resulted in the subsequent accumulation of neutral, or near neutral mtDNA polymorphisms, which are continent-specific. These

polymorphisms define specific haplotypes and haplogroups, which are discussed in more depth in Chapter two. The variation of mtDNA correlates with the ethnicity and geographic origin of individuals (Denaro *et al.*, 1981), with Africa showing the greatest variation and deepest root of origin for human mtDNA (Cann *et al.*, 1987). Chen *et al.* (2000) suggested that representatives of the Khoi-San population are some of the most ancient and distinct African populations. In correlation with the mtDNA, the Y-chromosome also harbours polymorphisms that allow for its evolutionary reconstruction (Hammer *et al.*, 1998). The most ancestral Y-chromosome haplotype is represented in Sudanese and Ethiopians from east Africa, together with southern African Khoi-San-speaking populations (Semino *et al.*, 2002).

The mitochondrial Cambridge reference sequence (CRS) is utilised as the premier mtDNA reference worldwide. However, discrepancies in this reference sequence have been observed, which led to the erroneous identification of alterations in comparative analyses (Howell *et al.*, 1992). In addition, it is a concern that the CRS was derived from one European individual, together with bovine and HeLa cell mtDNA sequences (Andrews *et al.,* 1999). During reanalysis of the CRS a revised CRS (RCRS) was established. It was observed that certain nucleotides from the CRS were erroneous or represented rare polymorphisms. It is, however, essential that a reference sequence does not contain rare polymorphic alleles.

This investigation is one of the first to study the complete mtDNA sequence of Southern African Khoi-San individuals. Various methods were utilised, including automated sequencing that is presented in Chapter three, from which the whole mitochondrial genome sequences were derived. The derived sequences were compared to one another and the RCRS, as discussed in Chapter four and presented in Appendix A.

A consensus sequence was derived from this ancient lineage. The derived sequences were also subjected to phylogenetic analysis to investigate their clustering in the African phylogenetic tree. Nucleotide alterations, presented in Appendix B, which are characteristic to certain haplogroups, were utilised to exclude the 13 derived sequences from the other global haplogroups. Statistical tests were performed to compare the sequences in order to test for neutral evolution and to investigate their distribution of nucleotide differences. The 13 Khoi-San sequences were added to 12 L0-specific sequences, from which an L0 phylogenetic tree was previously constructed, as illustrated

in Chapter four and Appendix C respectively. This investigation thus represents a pilot study to investigate the genetic variability of this ancient lineage.

CHAPTER TWO

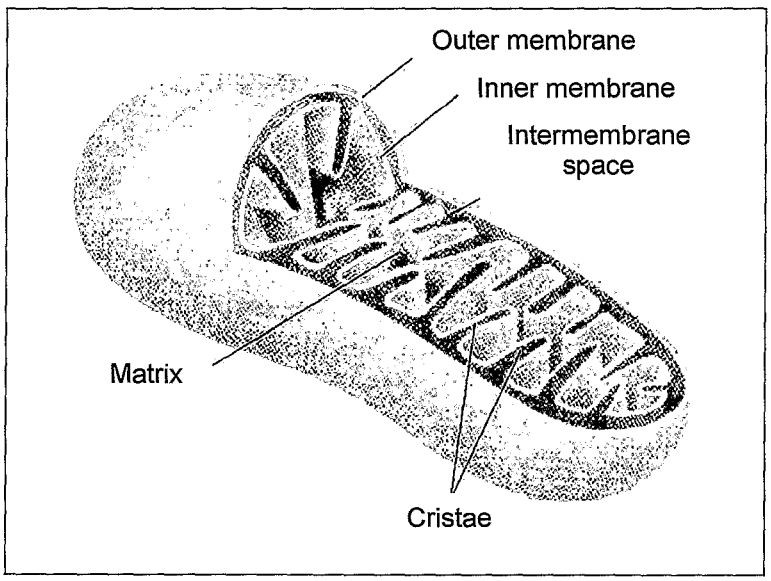
MOLECULAR, GENETIC AND EVOLUTIONARY CHARACTERISTICS OF mtDNA

Mitochondria are the primary energy-producing organelles of the cell. Up to 10 copies of mtDNA molecules are present in one mitochondrion with ca. 1,000 mitochondria per cell (Clayton, 1982). These genomes are favourable to study since they have an increased mutation rate when compared to the nucleus, do not undergo recombination and are inherited through the maternal lineage. This has led to the use of mtDNA in phylogenetic studies.

2.1 MITOCHONDRIAL STRUCTURE

Mitochondria were first described by Benda (1898) and obtained their name from the Greek words "mitos" and "chondrion", meaning "threads" and "granule" respectively. This organelle consists of an outer mitochondrial membrane (OMM) and a folded inner mitochondrial membrane (IMM) as illustrated in Figure 2.1 (Borst, 1977).

Figure 2.1: Schematic representation of the structure of the mitochondria



Adapted from Fairbanks and Andersen (1999).

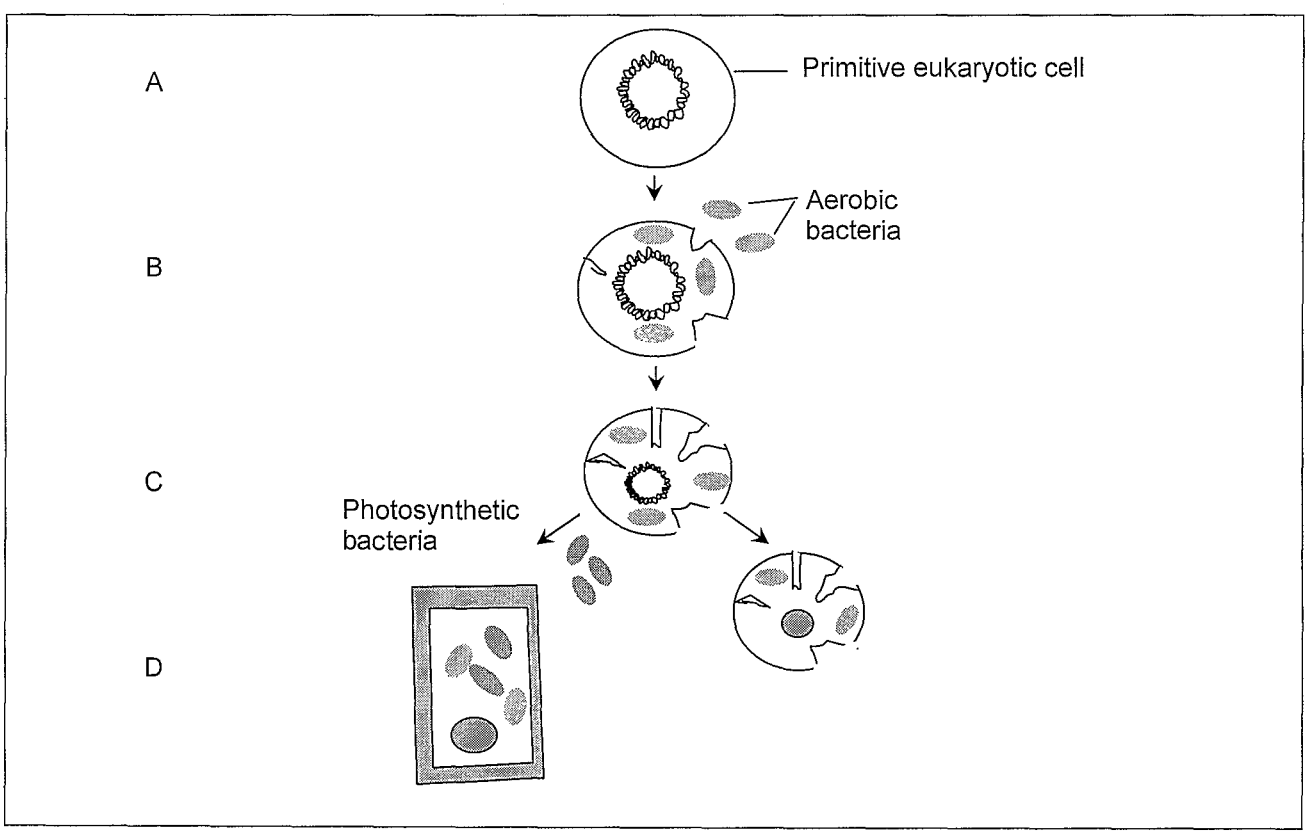
The molecular weight of a mitochondrion is 10^7 daltons (Da) and it is 5 micrometres (μm) in length (Giles *et al.*, 1980). The OMM has a smooth appearance and consists of 60-70% proteins and 30-40% lipids (Garrett and Grisham, 1999). It is suggested that the function of the OMM is to maintain the mitochondrion's shape (Garrett and Grisham, 1999).

Several channels are located in the OMM permitting differential transport of specific molecules into the organelle. The IMM is folded into flattened structures known as cristae that enlarge the IMM's surface area, as depicted in Figure 2.1. The IMM divides the mitochondria into two compartments, namely the intermembrane space, located between the OMM and the IMM, and the matrix, which is enclosed by the IMM (Borst, 1977). Most enzymes of the Krebs cycle and the fatty acid oxidation pathway are located in the matrix, along with the mtDNA molecules, ribosomes and enzymes required for mtDNA replication and protein synthesis (Garrett and Grisham, 1999). The IMM is more protein-rich than the OMM and is almost impermeable to molecules and ions. Carrier proteins are embedded in the IMM to regulate the exchange of substrates across the membrane (Garrett and Grisham, 1999).

2.2 MITOCHONDRIAL ORIGIN

This organelle bears resemblance to the genetic system of prokaryotes (Borst, 1977). This has led to the endosymbiotic hypothesis, which suggests that the ancestors of mitochondria were free-living bacteria that developed an obligatory symbiotic life with primitive eukaryotic cells (Borst, 1977), as illustrated in Figure 2.2.

Figure 2.2: Schematic representation of a model for the origin of a complex cell



Adapted from Westphal (2003).

In essence, aerobic bacteria were engulfed by an ancestor of eukaryotic cells (A in Figure 2.2) and developed a symbiotic relationship, as illustrated in B of Figure 2.2. Invaginations in the cell developed into the cell membrane (Figure 2.2, C). Over time, the symbiotic bacteria evolved into mitochondria whilst the invaginations in the cell evolved into the endoplasmic reticulum and nuclear membrane. Complex eukaryotic cells gave rise to fungi and animals, whereas eukaryotic cells that engulfed photosynthetic bacteria developed into plants whilst the bacteria formed chloroplasts (D of Figure 2.2). Evidence supporting this hypothesis includes the circular structure of the mitochondrial genome, its extra-nuclear location and the absence of chromosomal organisation within this genome (Gray, 1993).

The endosymbiotic hypothesis suggested that eukaryotic cells originated via two steps. Initially the nucleus originated from an Archaebacterium (Margulis, 1970). This was followed by the development of a symbiotic relationship with the modern mitochondrial eubacteria precursors. The hypothesis suggests that eukaryotes lacking mitochondria, known as Archaezoa, would be at the basis of the ancestral eukaryotic tree (Margulis, 1970).

An alternative theory, known as the hydrogen hypothesis (Martin and Müller, 1998), suggests the simultaneous development of the eukaryotic nucleus and the mitochondria. This occurred via the fusion of a host, a hydrogen requiring methanogenic Archaebacterium, with a hydrogen-producing Alpha-Proteobacterium symbiont. Support for this theory includes the observation that some genes of the eukaryotic nucleus are of Archaebacterial origin and others from that of eubacteria. The common ancestry of hydrogenosomes and mitochondria, such as a genome coding for proteins similar to those of the mitochondria, provides further support for this hypothesis.

Both theories suggest that a great amount of the proto-eubacterial genetic material was transferred to the nucleus, which resulted in a well defined interrelationship. The genome sizes of animal mitochondria are up to 100-fold less, compared to that of free-living bacteria, whereas hydrogenosomes, the modified mitochondria of anaerobic eubacteria, have lost their plastid genomes entirely (Embley *et al.*, 1997). According to Berg and Kurland (2000) there are two possible modes of genetic loss. The first involves the loss of nonessential coding sequences that became dispensable, such as those required for motility or cell wall building (Selosse *et al.*, 2001). A second mode of genetic loss includes transferring important mitochondrial genes to the nucleus. This is suggested to occur in

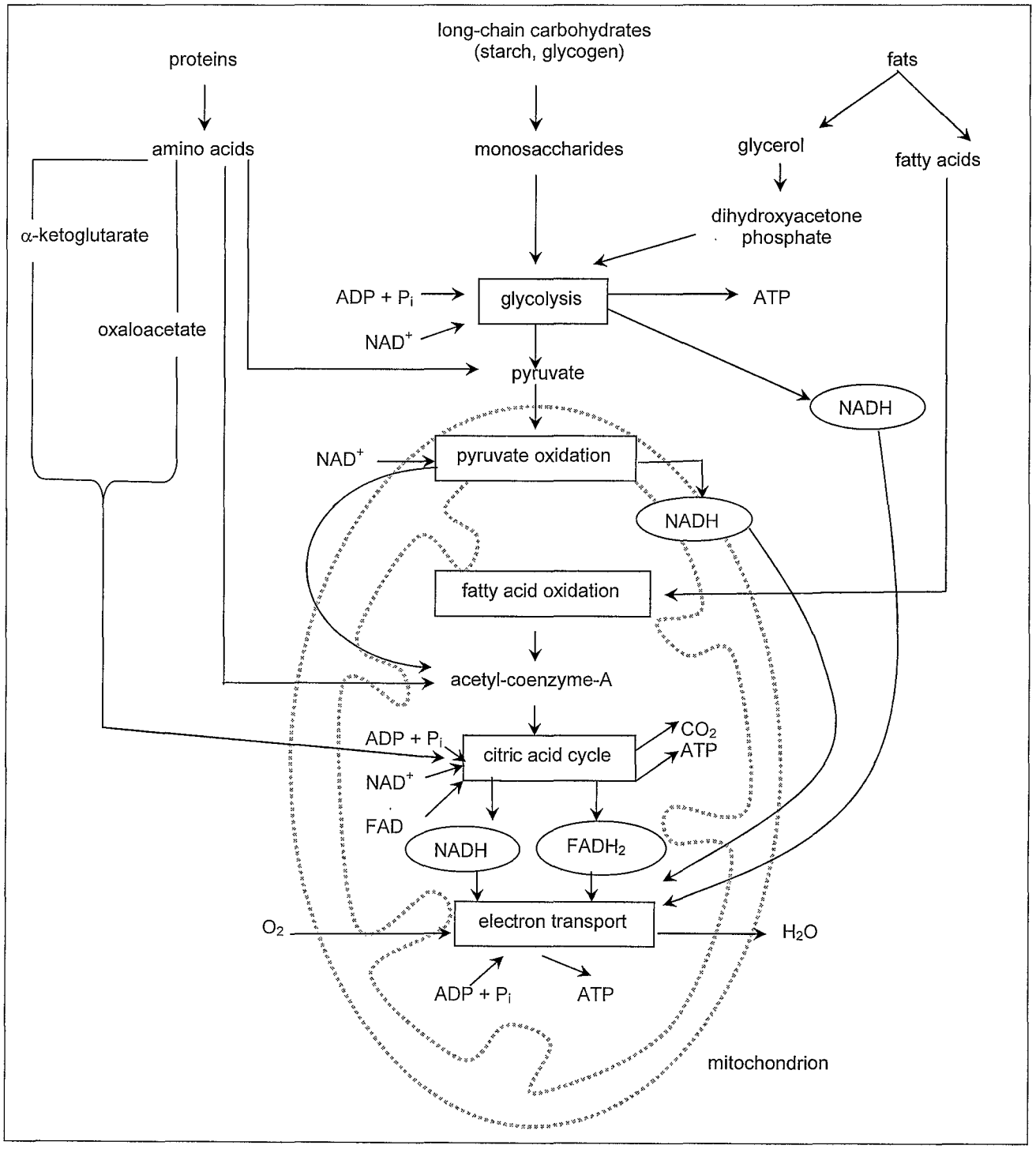
three steps. Firstly, the organelle gene is copied and integrated as a pseudogene in the nucleus (Blanchard and Lynch, 2000). This transferred sequence is transformed over time into an active nuclear gene by acquiring a promoter and a pre-sequence encoding a transit peptide, involved in the targeting of the protein product to the organelle, and by adapting to the nuclear code. The loss of the organelle gene copy due to redundancy completes the transfer process (Selosse *et al.*, 2001). Various hypotheses have been proposed to explain the erosion of organelle genomes. The unidirectional transfer hypothesis argues that it would be easier for genes to move from the organelle to the nucleus than *vice versa*, because one of the three transferring steps, described above, prevalently occurs in one direction, namely towards the nucleus (Selosse *et al.*, 2001). Other theories suggest that the properties of the organelle selectively favour the transfer to the nucleus. These properties include the higher mutation rate in the organelle and the production of free radicals, causing DNA mutation (Allen and Raven, 1996). The effects of Muller's ratchet have also been proposed to favour the loss of the organelle copy of the gene (Muller, 1964). This suggests that irreversible, deleterious mutations are more likely to occur in small genomes, such as those of organelles. Accordingly, once in the nucleus the gene escapes the ratchet and lineages containing the nuclear copy would be fitter than lineages containing the organelle's copy. However, the exact cause for genome erosion remains unclear.

2.3 BIOCHEMICAL ASPECTS OF THE MITOCHONDRIA

Oxidation of sugars, illustrated in Figure 2.3, starts at the conversion of these molecules to pyruvate during glycolysis, which occurs outside the mitochondrion (Garret and Grisham, 1999). Proteins are first broken down into amino acids. The deamination of amino acids results in α-keto acids, of which several are intermediates from the Krebs cycle, also known as the citric acid cycle, and enter the cycle directly. Oxidised amino acids are converted to pyruvate or the acetyl groups of acetyl coenzyme A (acetyl-CoA). However, pyruvate can also enter the mitochondrion where it is oxidised in the matrix to acetyl-CoA. Acetyl-CoA subsequently participates in the citric acid cycle, as depicted in Figure 2.3. Fats first undergo fatty acid oxidation, after which they enter the biochemical oxidation process to produce acetyl-CoA. The energy from the oxidised substrates (Garrett and Grisham, 1999) is transferred to flavoproteins and coenzymes to form reduced flavin adenine dinucleotide ($FADH_2$) and reduced nicotinamide adenine dinucleotide phosphate (NADH) respectively. The electron transport chain (ETC) accordingly oxidises these

substrates by removing electrons and channelling them to the final electron acceptor, oxygen (O_2), through a series of oxidation-reduction reactions, as illustrated in Figure 2.3.

Figure 2.3: Schematic representation of biochemical pathways associated with the mitochondria

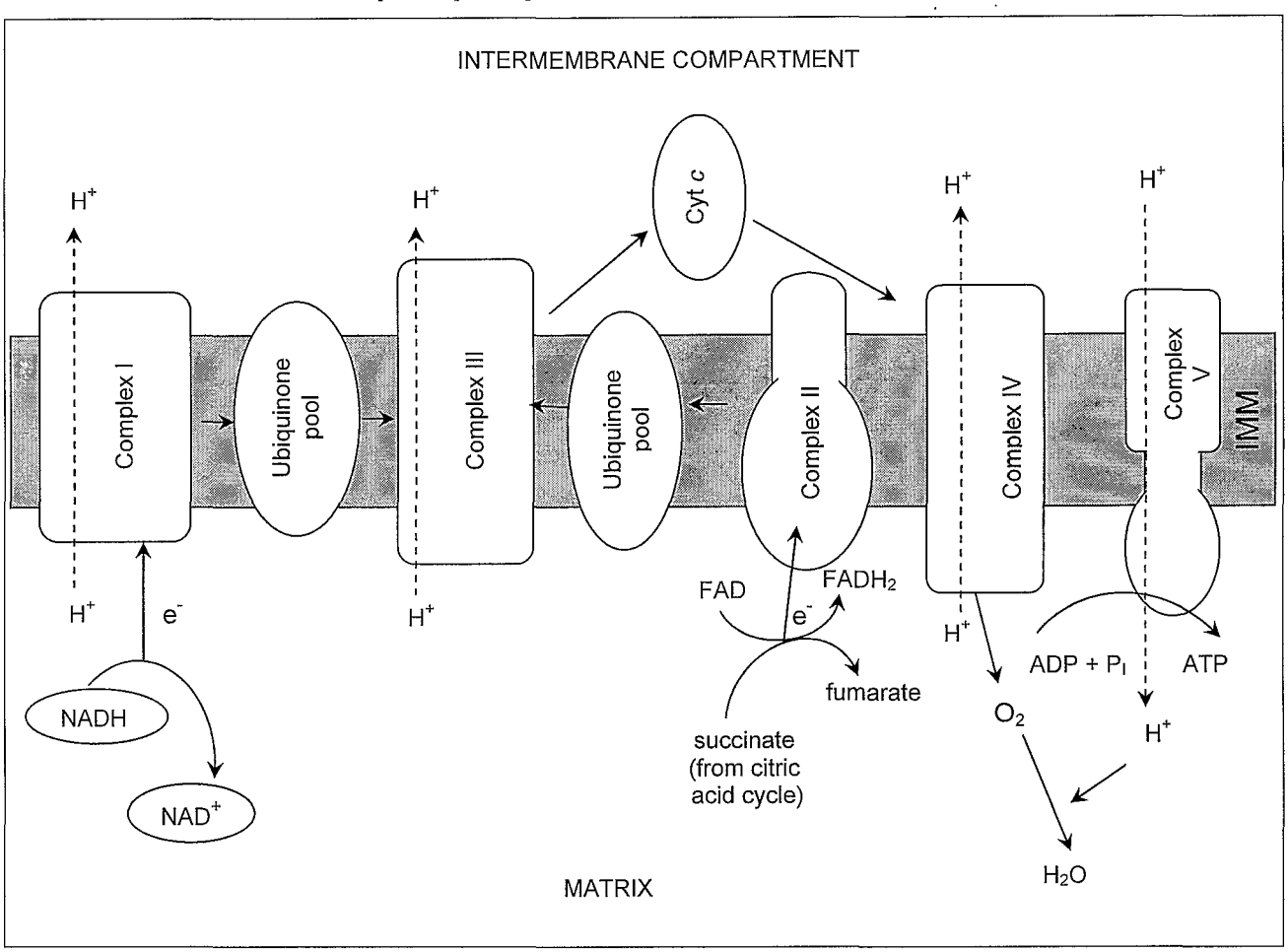


NADH = reduced nicotinamide adenine dinucleotide; NAD^+ = nicotinamide adenine dinucleotide; ADP = adenosine diphosphate; P_i = inorganic phosphate; ATP = adenosine triphosphate; FAD = flavin adenine dinucleotide; CO_2 = carbon dioxide. Adapted from Fairbanks and Andersen (1999).

2.3.1 The electron transport chain

The ETC consists of three complexes (I, III and IV) embedded in the IMM. These are coupled to a fifth complex, complex V (Figure 2.4), which couples oxidative phosphorylation (OXPHOS) to the respiratory chain (Fairbanks and Andersen, 1999).

Figure 2.4: Schematic representation of the electron transport chain and oxidative phosphorylation



NADH = Reduced nicotinamide adenine dinucleotide; NAD^+ = nicotinamide adenine dinucleotide; e^- = electron; H^+ = proton; cyt *c* = cytochrome *c*; ADP = adenosine diphosphate; P_i = inorganic phosphate; ATP = adenosine triphosphate; complex I = NADH-ubiquinone oxidoreductase; complex II = succinate-ubiquinone oxidoreductase; complex III = ubiquinone-cytochrome *c* oxidoreductase; complex IV = cytochrome *c*-O_2 oxidoreductase; complex V = ATP synthase; IMM = inner mitochondrial membrane. Adapted from Fairbanks and Andersen (1999).

Electrons are accepted from NADH by complex I, thus linking the ETC with the Krebs cycle and fatty acid oxidation (Garrett and Grisham, 1999). Complex II, the only integral IMM protein of the Krebs cycle, accepts electrons from $FADH_2$, which is reduced during the Krebs cycle, establishing a second link between the Krebs cycle and the ETC. Owing to its association to the mitochondrial inner membrane, complex II was previously considered to be part of the ETC. Since complex II is only involved with the oxidation of succinate to fumarate and the subsequent reduction of coenzyme Q and is not capable of conserving energy for ATP production, it is no longer considered to be an integral part of

the ETC (Wikström, 2003). Complex II utilises coenzyme Q as an electron acceptor, which forms an integral part of the inner membrane, thus accounting for the association of this complex to the inner membrane.

The product from both complexes I and II, reduced coenzyme Q, or ubiquinol (UQH_2), serves as a substrate for complex III (Garrett and Grisham, 1999). The oxidation of UQH_2 results in the reduction of cytochrome *c*, the substrate for complex IV, which in turn reduces O_2 to form H_2O. At the same time the energy from the ETC can be used by complex I, III and IV, to pump protons across the IMM resulting in a proton gradient across the membrane (Bauer *et al.*, 1999).

An ATP synthase (ATPase) complex, embedded in the IMM as complex V, consists of an F_1 unit, which catalyses ATP synthesis, and an integral membrane protein unit, F_0, that forms a channel through which protons move to drive ATP synthesis. Complex V uses the reverse flow of protons through both subunits to generate energy, in the form of ATP, via phosphorylation of adenine diphosphate and inorganic phosphate (Pi), as described by Campbell (1991). Certain characteristics of the respective complexes are presented in Table 2.1.

Table 2.1: Complexes of the electron transport chain

Complex		Mass (kDa)	Subunits	Prosthetic group	Binding site for
Number	Name				
I	NADH-UQ reductase	980	46	FMN	NADH (matrix side)
				Fe-S	UQ (lipid core)
II	Succinate-UQ reductase	140	4	FAD	Succinate (matrix side)
				Fe-S	UQ (lipid core)
III	UQ-Cyt c reductase	248	11	Heme b_L Heme b_H Heme c_1 Fe-S	Cyt *c* (intermembrane space side)
IV	Cytochrome c oxidase	162	>10	Heme *a* Heme a_3 Cu_A Cu_B	Cyt *c* (intermembrane space side)

NADH-UQ = NADH-Coenzyme Q reductase; UQ = Coenzyme Q or ubiquinone; FMN = reductase; flavin mononucleotide; Fe-S = iron-sulphur protein; FAD = flavin adenine dinucleotide; cyt = cytochrome; Cu = copper. Adapted from Garrett and Grisham (1999) and Carroll *et al.* (2002).

2.4 MITOCHONDRIAL GENETICS

The mitochondrion is, apart from the nucleus, the only cellular organelle that contains DNA (Borst, 1977). The double-stranded, circular mitochondrial genome is 16,569 base pair (bp) in length and its complete sequence was determined by Anderson *et al.* (1981).

Mitochondrial DNA, together with nDNA, encodes the complete respiratory chain (Borst, 1977). The mitochondrial genome is replicated within the organelle and encodes the essential transcripts for processing and expressing mitochondrial proteins (Clayton, 1984). Most mitochondrial proteins are encoded by nDNA and synthesised in the form of precursor proteins (as discussed in section 2.5), which are imported into the mitochondria via translocation systems that are located in the inner and outer mitochondrial membranes (Eilers *et al.,* 1988).

2.4.1 Mitochondrial encoded genes

The 16,569 bp genome encodes for 37 genes (Figure 2.5), namely 22 tRNAs, two rRNAs and 13 polypeptides (Anderson *et al.*, 1981). The different polypeptides encoding different subunits of respiratory chain complexes are summarised in Table 2.2.

Table 2.2: Mitochondrial and nuclear encoded subunits of the respiratory and oxidative phosphorylation chain

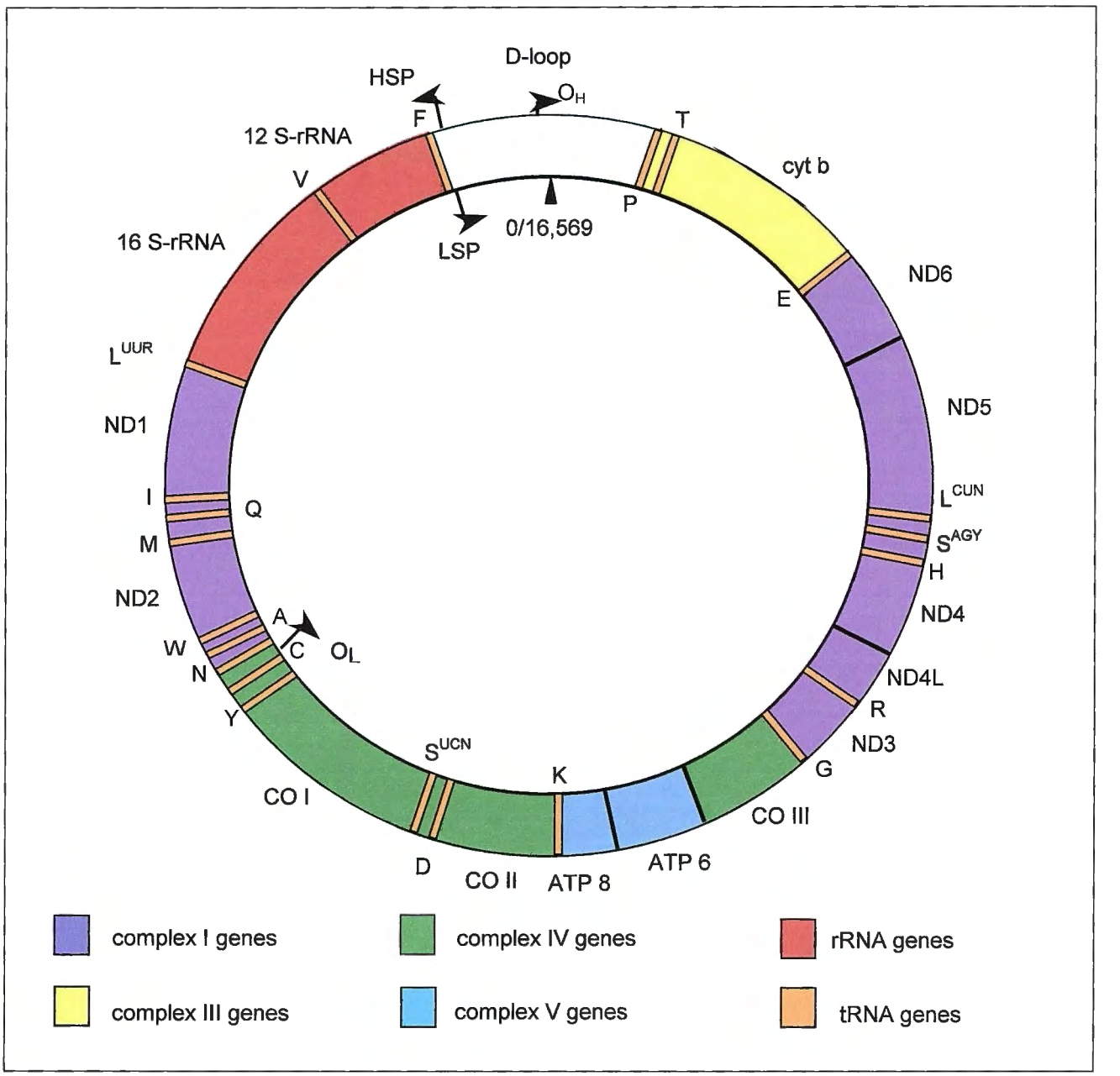
Complex number	Complex name	Mitochondrial encoded subunits	Number of subunits encoded by nDNA
I	NADH-UQ reductase	ND1 – ND4, ND4L, ND5 and ND6	36
II	Succinate-UQ reductase	None	4
III	UQ-Cyt c reductase	Cyt b	10
IV	Cytochrome c oxidase	COI, COII and COIII	10
V	ATPase	ATPase 6 and ATPase 8	14

NADH-UQ reductase = NADH-Coenzyme Q reductase; cyt = cytochrome; UQ = ubiquinone; ND1-ND4, ND4L, ND5 and ND6 = NADH dehydrogenase subunits; CO = cytochrome c oxidase subunits. Adapted from Schon (1993) and Wallace (1994).

The subunits of the respiratory chain complex II are all encoded by the nuclear genome (Anderson *et al.*, 1981). The base composition of one of the mitochondrial genome strands is purine (adenine [A] and guanine [G]) rich, while the complementary strand is pyrimidine (cytosine [C] and thymine [T]) rich (Anderson *et al.*, 1981). Because of this asymmetric composition the two strands have different buoyant densities and therefore separate differently in an alkaline cesium chloride gradient. The purine rich strand is thus known as the "heavy" strand (H-strand) and the complementary strand as the "light", or L-strand

(Anderson *et al.*, 1981). Twenty eight of the 37 mitochondrial encoded genes are encoded by the H-strand, including both the rRNA genes, 14 tRNA genes and 12 of the 13 polypeptide encoding genes, as illustrated in Figure 2.5.

Figure 2.5: Schematic representation of the mitochondrial genome



All tRNA genes are indicated by the single letter amino acid abbreviation. A = alanine; C = cysteine; D = aspartic acid; E = glutamic acid; F = phenylalanine; G = glycine; H = histidine; I = isoleucine; K = lysine; L = leucine; M = methionine; N = asparagine; P = proline; Q = glutamine; R = arginine; S = serine; T = threonine; V = valine; W = tryptophan; Y = tyrosine; L^{CUN} = leucine with anticodon CUN; L^{UUR} = leucine with anticodon UUR; S^{AGY} = serine with anticodon AGY; S^{UCN} serine with anticodon UCN; cyt b = cytochrome b; D-loop = displacement loop; ND1-6 = NADH dehydrogenase 1-6; CO I-III = cytochrome c oxidase I-III; ATP 6 = ATP synthetase subunit 6; ATP 8 = ATP synthetase subunit 8; O_H = heavy strand origin of replication; O_L = light strand origin of replication; HSP = heavy strand promoter; LSP = light strand promoter; 0/16569 = starting and ending point of the mtDNA. Adapted from MITOMAP (2003).

One polypeptide encoding gene is encoded by the L-strand together with eight tRNA genes (Clayton, 1984). Apart from the approximately 1 kilo bp (kb) noncoding sequences known as the control region (CR), containing the displacement loop (D-loop), mammalian

mtDNA is compact. Controversy exists on the exact size of the D-loop. According to Spelbrink (2003), the D-loop arises from O_H and is ca. 500 bp in length. However, Taanman (1999) defines the D-loop as being flanked by the $tRNA^{Phe}$ and $tRNA^{Pro}$, implying a D-loop length of 1,122 bp. This author also utilises CR and D-loop as synonyms. In this study the D-loop was defined as being 1,122 bp in length and synonymous with the CR, containing three conserved sequence boxes (CSB), hypervariable sequences as well as the termination associated sequences (TAS), which are discussed in section 2.4.2.

Nontranslated regions and complete termination codons are absent from almost all open reading frames (Anderson *et al.*, 1981). Termination codons are completed during posttranslational processing with polyadenylation of the polypeptide encoding genes. In addition, the genes lack introns (Anderson *et al.*, 1981). The compatibility of mtDNA, together with the observation that most structural genes are immediately flanked by tRNA genes, suggests the absence of multiple control regions responsible for the expression of genes, except if they were located within the gene (Anderson *et al.*, 1981). Trans-acting nuclear encoded factors are required for mitochondrial replication and transcription. The mitochondrial ribosomal proteins of vertebrates as well as enzymes of various mitochondrial located catabolytic pathways are synthesised outside the organelle (Taanman, 1999). Mitochondrial destined polypeptides encoded by the nucleus are usually synthesised with a cleavable presequence, which serves to target the polypeptide to the organelle.

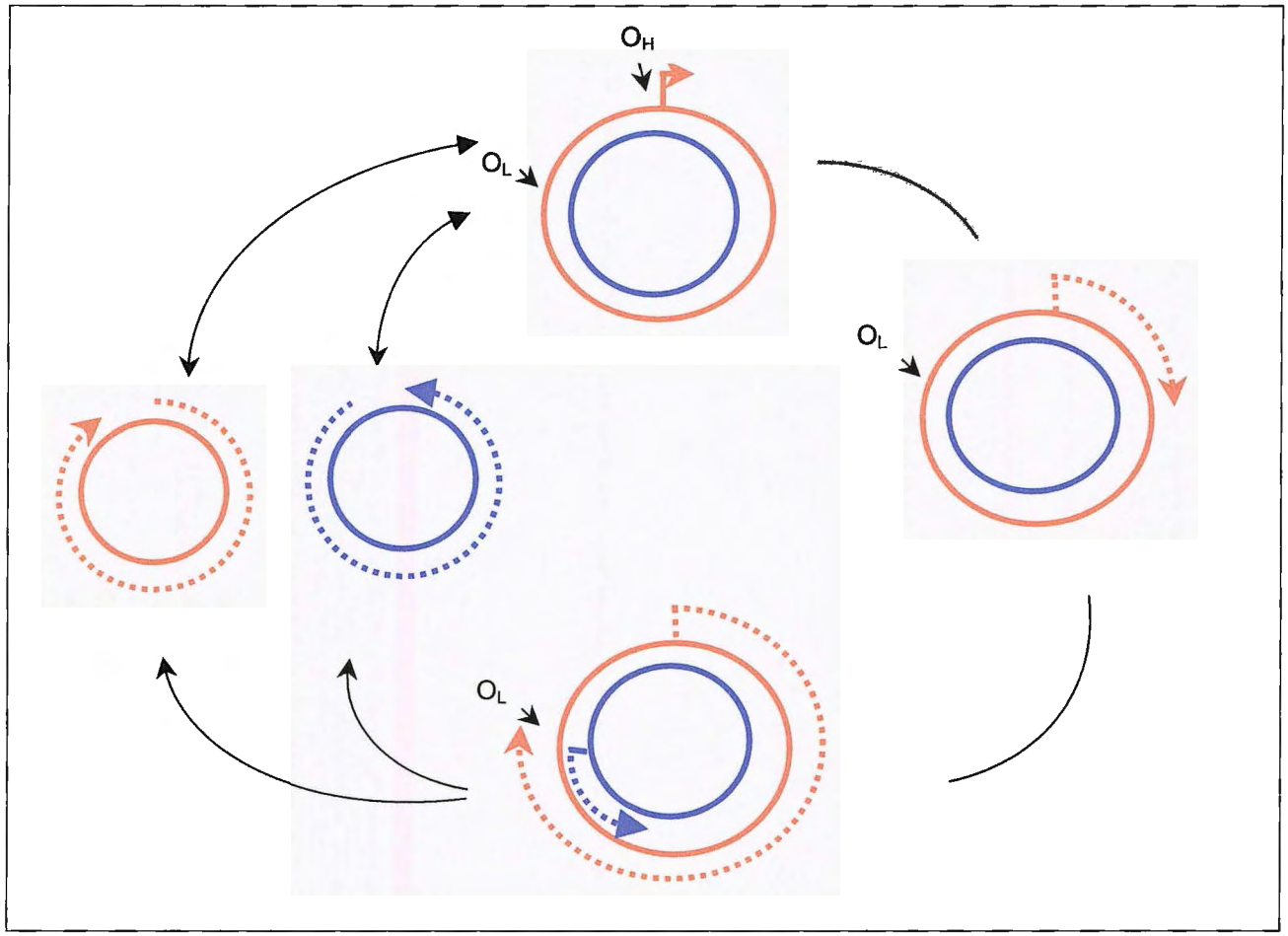
2.4.2 Replication of the mitochondrial genome

Replication of mtDNA is initiated at the H-strand origin (Gillum and Clayton, 1979). The D-loop is the main regulatory region of the mitochondrial genome since it contains the origin of H-strand synthesis as well as promoters of both the H- and the L-strands (Clayton, 1982). Other regions of mtDNA regulatory sequences include the origin of L-strand synthesis, which overlaps four L-strand transcribed tRNA genes, and the binding site for the mitochondrial transcription termination factor, mTERM, which has a role in termination of rRNA transcription (Attardi, 1993).

A characteristic of mtDNA replication is the presence of a short triplex region, of which the third strand is known as the D-loop. It is suggested that the D-loop represents intermediates of replication that were aborted (Spelbrink, 2003). This implies that only a few of the strands that are initiated for replication extend past the D-loop region.

The D-loop control region is displaced by a segment of RNA prior to initiation of replication, at which time the mitochondrial polymerase γ binds and starts synthesis of the complement of the H-strand, as depicted in Figure 2.6 (Anderson *et al.*, 1981).

Figure 2.6: Schematic representation of mitochondrial genome replication



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

DNA polymerase γ has 3'→5' exonuclease activity, apart from its 5'→3' polymerase activity, to ensure faithful mtDNA replication (Wang, 1991). Chang *et al.* (1985) suggested that short transcripts, which originate at the L-strand transcription initiation site, prime the initiation of H-strand synthesis, as illustrated in Figure 2.7. This suggests a link between mitochondrial transcription and replication. The precursor RNA primer is suggested to exist as an RNA-DNA hybrid, known as the R-loop (Moraes *et al.*, 1991a). Three CSB, known as CSB I, II and III, exist where transition from RNA to DNA synthesis occurs (Figure 2.7). The transition is suggested to involve the processing of the RNA in close approximation of O_H or by the replacement of the transcription machinery, near O_H, with that necessary for DNA synthesis.

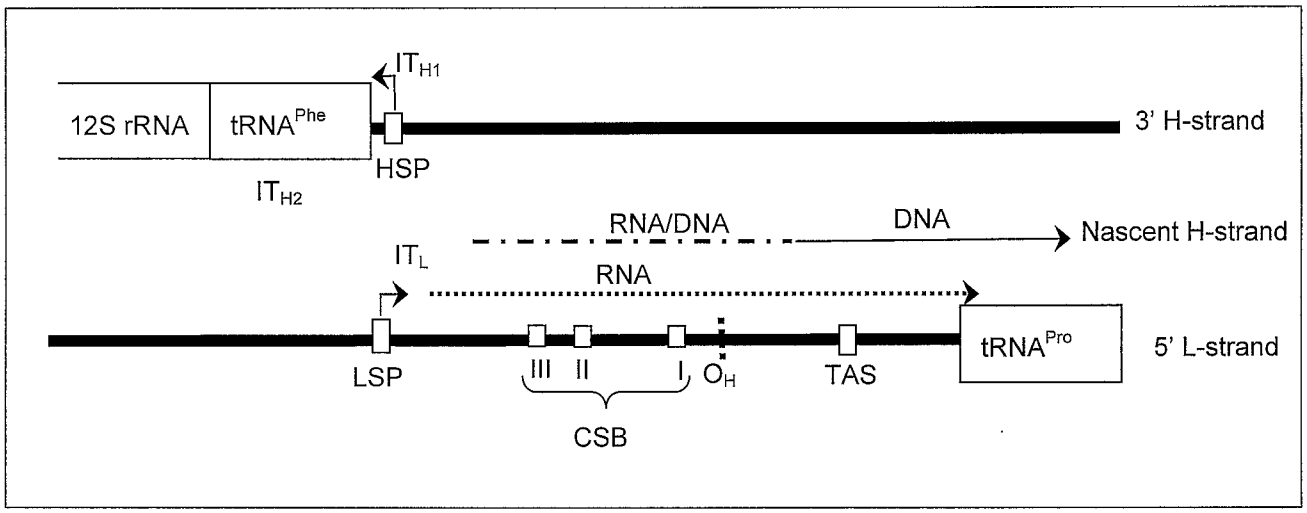
Heavy strand synthesis extends clockwise and initiates L-strand synthesis on the opposite strand, and in the opposite direction, when passing two thirds of the circular genome (Anderson *et al.*, 1981) as illustrated in Figure 2.6.

Synthesis is continued until the initial origins are reached. Daughter molecules then segregate resulting in the nascent H-strand existing on a daughter molecule with a single nick together with daughter molecules containing L-strands with gaps. These gaps are subsequently filled to produce replicated closed circle molecules. Synthesis of the H-strand in vertebrates is stalled shortly after initiation, ca. 50 nucleotides downstream of the TAS (Taanman, 1999). Termination of the H-strand downstream of the TAS element, or alternatively elongation to produce the complete H-strand, is determined by yet unknown mechanisms.

2.4.3 Mitochondrial transcription

According to Schon (1993), transcription of the mitochondrial genome starts at initiation sites (IT_{H1} and IT_L) for both the H- and the L-strands located in the promoter regions (H-strand promoter [HSP] and L-strand promoter [LSP]), as depicted in Figure 2.7.

Figure 2.7: Schematic representation of mitochondrial replication and transcription



IT_{H1} = upstream heavy strand initiation site, IT_{H2} = downstream heavy strand initiation site; IT_L = L-strand initiation site; HSP = H-strand promoter; LSP = L-strand promoter; O_H = origin of H-strand synthesis; O_L = origin of L-strand synthesis; CSB = conserved sequence blocks; TAS = termination associated sequence; bent arrows = transcription initiation directions; — · — · = transition from RNA to DNA occurs in the region around the CSB. Adapted from Taanman (1999).

These transcription initiation sites are located within 150 bp of one another in the D-loop. In support of its bacterial ancestry, the mitochondrial genome is transcribed in a polycistronic fashion, unlike the monocistronic replication of nuclear genes (Attardi, 1993). Additional enhancer elements, located upstream from the initiation regions, are essential

for sufficient transcription (Taanman, 1999). An example of such an element is mitochondrial transcription factor A, abbreviated as mtTFA or TFAM (Spelbrink, 2003). Binding of this element to regions upstream of the respective promoter sequences is required for transcription initiation.

Transcription of the L-strand is initiated from a single initiation site (IT_L) and yields one transcript that can be processed into the encoded mRNA and tRNAs (Attardi, 1993). Heavy strand synthesis commences from two initiation sites. The transcript from the upstream initiation site (IT_{H1}) produces a short product extending from the promoter to the end of the 16S rRNA gene and includes the two rRNAs and two tRNAs (Schon, 1993). Termination of this shorter transcript occurs when a certain protein factor, mtTERM, binds within the tRNA leucine gene with anticodon UUR (tRNA $^{Leu(UUR)}$) and causes the polymerase to fall off and terminate transcription (Christianson and Clayton, 1986).

Transcription from the downstream initiation site (IT_{H2}), located at the 5' end of the 12S rRNA gene, includes almost the entire mitochondrial genome and when processed produces the remaining mature RNA species (Attardi, 1993). The shorter transcript is transcribed at a higher rate than the full length transcript to ensure the availability of sufficient amounts of rRNAs for translation of the mRNAs (Attardi *et al.*, 1993).

2.4.3.1 Post-transcriptional processing

Processing of transcripts involves the exact excision of the different RNA genes from the polycistronic transcript. Transfer RNA processing would require precise 5' and 3' cleavage from the surrounding rRNAs and mRNAs. These RNAs are further modified by addition of CCA sequences at the 3'-end after transcription, since this is not encoded by the genome (Clayton, 1984).

The activity of polymerase A is necessary in adding adenine bases to rRNAs and mRNAs (Clayton, 1984). Polyadenylation creates stop codons for those mRNAs that do not have mitochondrial encoded termination codons (Clayton, 1984).

2.4.4 Mitochondrial translation

One of the distinct features of mammalian mtDNA is the difference from the universal genetic code when compared to nDNA (Table 2.3). The mitochondrial genome is also read by a unique set of mitochondrial tRNAs (Anderson *et al.*, 1981). Accordingly, 22 tRNA molecules are able to sufficiently translate all 13 mitochondrial proteins (Attardi, 1993).

Table 2.3: Comparisons between the nuclear and the mitochondrial genetic codes

Anticodon	Nuclear code	Mitochondrial code
UGA	STOP	Tryptophan
AUA	Isoleucine	Methionine
AGA	Arginine	STOP
AGG	Arginine	STOP
AUA	Isoleucine	Initiation
AUU	Isoleucine	Possible initiation
AUG	Initiation	Possible initiation

Adapted from Anderson *et al.* (1981).

The mitochondrial translation system is quite distinct from its cytosolic and bacterial counterparts and is not fully understood. The mitochondrial ribosomes, also referred to as mitoribosomes, which are located in the mitochondrial matrix, have a very low RNA content (Taanman, 1999). However, they have high protein content, resulting in a total mass similar to that of bacterial ribosomes.

Initiation of mitochondrial translation occurs differently from that of eukaryotic cells. The main difference is that eukaryotic initiation occurs through 5'-cap recognition and scanning, but mitochondrial mRNAs lack a 7-methylguanylate cap structure. Liao *et al.* (1989) have suggested that the 28S subunit of mitoribosomes is able to bind to mRNA. After binding, the 28S subunit is suggested to move to the mRNAs 5'-end with the aid of initiation factors (Liao *et al.*, 1990).

At present the only mammalian mitochondrial initiation factor known is mtIF-2, which binds to the small ribosomal subunit prior to binding to mRNA (Liao *et al.*, 1990). Schwartzbach and Spremulli (1989) have identified three mitochondrial elongation factors (mtEF), mtEF-Tu, mtEF-Ts and mtEF-G, which are very similar to those of prokaryotic factors. Prokaryotic elongation factor Tu (EF-Tu) binds aminoacyl-tRNA, which recognises the first

codon, and guanosine triphosphate (GTP). The GTP is accordingly hydrolysed to guanosine diphosphate and P_i, with the subsequent formation of an EF-Tu:guanosine diphosphate (GDP) complex (Schwartzbach and Spremulli, 1989). Elongation factor Ts (EF-Ts) promotes the recycling of EF-Tu by displacing GDP with GTP. The elongation factor G (EF-G) couples the energy from the hydrolysis of GTP to movement, thus promoting translocation of the ribosome along the mRNA. One difference between bacterial and mitochondrial EF-Tu is the inhibition to catalyze polymerization by the antibiotic kirromycin of the bacterial factor in contrast to the resistance of the mitochondrial factor. Mitochondrial translational elongation is therefore suggested to proceed similarly to that of prokaryotes.

2.5 MITOCHONDRIAL PROTEIN IMPORT

Since most mitochondrial proteins are nuclear encoded, it is required that they be imported into the organelle. Importation into the mitochondria requires specific targeting information and import pathways since this organelle has various subcompartments, namely the OMM, IMM and the matrix. Therefore, proteins to be imported are synthesised as precursors containing additional N- or C-terminal presequences or internal targeting information (Duby and Boutry, 2002).

Once synthesised, cytosolic chaperones interact with the precursor proteins, usually in an ATP-dependent process. Chaperones mainly prevent the presequences from misfolding or aggregating and transport them to the mitochondria (Eilers *et al.,* 1988). The precursor accordingly crosses the outer membrane of the organelle through the help of translocase proteins, known as the translocase of the OMM (TOM) complex. The TOM complex is composed of different subunits that are embedded in the outer membrane (Duby and Boutry, 2002).

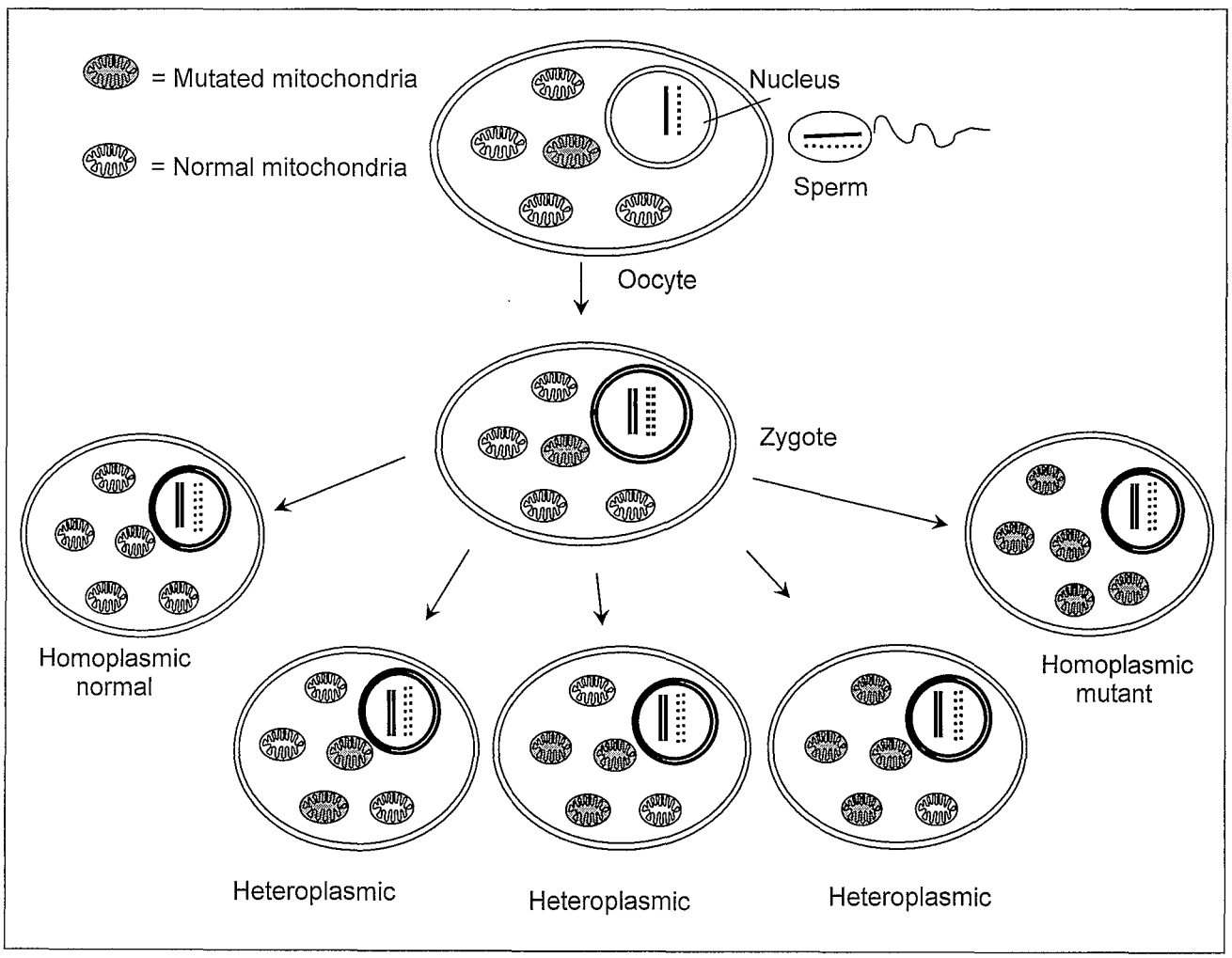
Subsequently the precursor is transported through the intermembrane space and inner membrane. This involves two complexes known as translocase of the IMM (TIM), in particular TIM22 and TIM23 (Sirrenberg *et al.*, 1996). The TIM23 complex imports proteins, with typical amino-terminal preseqeunces, that are targeted to the matrix, while TIM22 is responsible for the insertion of carrier proteins, which usually lack a targeting sequence, in the inner membrane. Precursors undergo maturation, through cleavage by the mitochondrial processing peptidase, during or after import into the mitochondria.

2.6 MITOCHONDRIAL INHERITANCE

The mtDNA is inherited through the maternal lineage (Figure 2.8) implying that a single female's offspring could have similar mtDNA (Giles *et al.*, 1980). Piko and Matsumoto (1976) suggested that spermatozoa contribute approximately 100 mitochondria at fertilisation, in comparison to the 10^5 to 10^8 mitochondria from the oocyte.

However, despite the entrance of the midpiece region of vertebrate sperm into the egg at fertilisation, paternal mtDNA is not detectable in the offspring (David and Blackler, 1972). It is possible that paternal mitochondria may be diluted below detectable levels or that the oocyte mediates the removal of the paternal mitochondria. The maternal inheritance of chloroplast DNA in *Chlamydomonas* and other higher plants is ensured by the methylation and subsequent protection from degradation. However, no such phenomenon was observed in mtDNA from sperm and oocytes (David and Blackler, 1972).

Figure 2.8: Schematic representation of maternal inheritance and replicative segregation



Adapted from DiMauro *et al.* (1990).

Shitara *et al.* (1998) investigated the presence of leaked paternal mtDNA in the hybrid offspring of fertilised mice eggs, into which sperm mtDNA was introduced. The unequal distribution of the paternal mtDNA in all hybrid offspring tissues and the failure to transmit to following generations confirms the exclusion of sperm mtDNA and strict maternal inheritance of mtDNA (Shitara *et al.*, 1998).

2.7 HETEROPLASMY

During early development all cells have identical copies of mtDNA, known as homoplasmy. Heteroplasmy is the occurrence of mixed populations of mtDNA, resulting from a mutation, in one cell or in different organelles (Wallace, 1994). The mutant and normal mtDNA will segregate randomly during mitosis and meiosis to the daughter cells. Over time and through replicative segregation (Wallace, 1986), the mtDNA in a cell could become purely mutant or normal or represent any state in between (Figure 2.8). Since replicative segregation can occur during both mitosis and meiosis, different proportions of mutant mtDNA can be present in a heteroplasmic individual or in a heteroplasmic mother's offspring. This implies that the time at which the mutation occurs, as well as the developmental goal of the specific cell, will determine the distribution of the heteroplasmic mtDNA (Wallace, 1994).

2.8 MUTATION RATE

Unlike nDNA, mtDNA lacks protective histones and has a repair mechanism that is not as sophisticated as that of the nucleus (Bauer *et al.*, 1999). This, together with exposure to oxygen radicals, which are released from the respiratory chain, contributes to an increased mtDNA mutation rate approximately 10 to 20 times faster than that of nDNA (Wallace *et al.,* 1987). The significance of each mutation depends on the time in the life cycle that it occurs and on the position in the genome.

Mitochondrial mutations observed in the germline are either neutral or deleterious and have accumulated in human lineages over time. Since mtDNA is only inherited maternally and no recombination occurs, the number of mitochondrial sequence differences between two individuals is directly proportional to the time they diverged from a common maternal ancestor (Wallace, 1995). Severely or even moderately pathogenic mutations are usually deleterious and relatively recent occurrences eliminated from future lineages by natural selection.

2.9 PATHOGENIC MITOCHONDRIAL MUTATIONS

Heteroplasmy of mtDNA in humans often has pathogenic consequences with the increased occurrence of this mutated mtDNA correlating with the severity of the disease (Wallace and Lott, 1992). In the case of deleterious mutations the capability to generate ATP decreases with the increase in mutant mtDNA, with the ultimate result being that the output of energy is not enough to sustain cell function (Wallace, 1992). As the proportion of mutant mtDNA varies, so does the type and severity of the clinical symptoms (Wallace, 1992).

2.9.1 Disorders caused by mtDNA mutations

Mitochondrial myopathies caused by mutations in mtDNA generally cause functional deficiencies in the respiratory chain. Included in this class of mutations are point mutations, rearrangements and mtDNA depletions.

2.9.1.1 Point mutations

Missense mutations occasionally affect the visual pathway and are associated with neurodegeneration. Leber's Hereditary Optic Neuropathy (LHON) is an mtDNA caused disease, leading to death of the optic nerve and resulting in adult onset blindness. Three point mutations alone are responsible for ca. 90% of worldwide LHON cases (Wallace *et al.*, 1988; Howell *et al.*, 1995). These include mutations in the ND1 gene at nucleotide 3,460, in the ND4 gene at nucleotide 11,778, and in the ND6 gene at nucleotide 14,484. Other point mutations have also been associated with the LHON phenotype. However, their exact pathological significance is still unclear (De Vivo, 1993; Bauer *et al.*, 1999).

Holt *et al.* (1990) described point mutations in the ATPase genes, associated with neuropathy, ataxia and retinitis pigmentosa. These mutations occur in the gene for subunit six and are associated with a combination of the above-mentioned symptoms together with developmental delay, dementia, seizures, proximal limb weakness and sensory neuropathy.

2.9.1.2 Rearrangements

Mitochondrial DNA deletions generally occur *de novo,* and during development. The first syndrome to be described in this class of mutations was Kearns-Sayre Syndrome, characterised by restricted eye movement and even heart conduction defects with childhood onset. Possible causes for this pathogenic disorder are mainly large-scale rearrangements (Zeviani *et al.*, 1988). These include mainly deletions and in a few cases insertions.

Molecular analysis of Pearson's Syndrome, which affects bone marrow and pancreatic function, has revealed large-scale rearrangements. However, patients with Pearson's Syndrome have also been observed to harbour single deletions and duplications (Wallace and Lott, 1992). Thus, the genetic aetiology of Pearson's Syndrome is variable.

Inoue *et al.* (2000) developed mice that harboured a 4,696 bp deletion. This deletion was transmitted, in a partially duplicated form as a combination with wild type mtDNA, for three generations. These mice expressed respiratory defects in various tissues and could thus be utilised as mouse models to study mitochondrial diseases.

2.9.1.3 Depletions

Another subset of mitochondrial mutations is mtDNA depletions. This is a rare but fatal syndrome. Moraes *et al.* (1991b) observed lethal infantile respiratory failure, lactic acidosis and muscle, liver or kidney failure to be associated with depleted mtDNA. These patients had extremely low levels of mtDNA.

2.9.2 Mitochondrial disorders caused by nDNA mutations

Since mitochondrial gene products comprise products encoded by both the mitochondrial and nuclear genome, mutations in the nuclear encoded products could also contribute to mitochondrial myopathies. Nuclear encoded mutations include substitutions in genes encoding structural proteins from the OXPHOS (Loeffen *et al.*, 1998), genes that indirectly affect the OXPHOS (Tiranti *et al.*, 1998) and genes not involved in mitochondrial pathways but linked to energy production (Koehler *et al.*, 1999). Mutated nDNA can also cause defects in the transport of substrates across the mitochondrial membrane and in the

utilisation of these substrates. A final subset of nDNA mutations are those that result in defects in intergenomic communication (DiMauro, 1996). This would result in altered control of the function between the mtDNA and the nDNA, having pathological consequences.

Examples of nuclear encoded mitochondrial disorders include Friedreich's ataxia and Wilson's disease, caused respectively by iron sensitivity together with mitochondrial instability and dysfunctional ATPase in the mitochondrial membrane (Lutsenko *et al.*, 1998). Further examples include mutations in the nuclear encoded subunit of succinate dehydrogenase, which is one of the causes of the severe neurodegenerative disease known as Leigh Syndrome (Bourgeron *et al.*, 1995)

2.9.2.1 Mutation in genes encoding mitochondrial enzymes

One of the consequences of all the above-mentioned mutations is protein dysfunction. Dysfunctions in protein synthesis are generally found to be associated with mutations in the mitochondrial encoded tRNA genes. The mitochondrial encephalomyopathy with lactic acidosis and stroke-like episodes disorder is mainly caused by a mutation in the $tRNA^{Leu(UUR)}$ gene at nucleotide 3,243 (Goto *et al.*, 1990). According to Majamaa *et al.* (1998) the prevalence of the A3243G mutation in the adult Finnish population was 16 in 100,000, excluding the children. Similarly, the pathological cause of myoclonus epilepsy with ragged red fibres occurs within the $tRNA^{Lys}$ gene at nucleotide 8,344 (Shoffner *et al.*, 1990). Diseases such as diabetes have also been linked with mitochondrial pathogenesis. In the study of Suzuki *et al.* (1997), the authors observed a mutation at nucleotide 3,264 in the $tRNA^{Leu(UUR)}$ gene that is associated with non-insulin-dependent diabetes mellitus.

2.10 MITOCHONDRIAL EFFECT ON SENESCENCE

Somatic mutations accumulate in postmitotic cells with age, causing the subsequent decrease in OXPHOS. This is responsible for the late onset and progress of mitochondrial diseases (Wallace, 1992). Corral-Debrinski *et al.* (1992) observed that the common 5 kb deletion, associated with Pearson marrow/pancreas syndrome and ocular myopathy, is usually not observed in normal hearts until 40 years of age. However, the deletion accumulates with an increase in age. It is suggested that a great portion of postmitotic cells' mtDNA may be mutated in elderly individuals. Thus, the accumulation of somatic mtDNA mutations with age may lead to the progressive reduction of OXPHOS and

subsequently increase the OXPHOS defect that is associated with inherited mtDNA mutations. If the combinations of somatic and inherited mutant mtDNA reduce a tissue's OXPHOS capacity below the bioenergetic threshold, it may have pathogenic consequences, as illustrated in Figure 2.9 (Wallace, 1994).

Figure 2.9: Schematic representation of age-related decline of OXPHOS and progression of disease

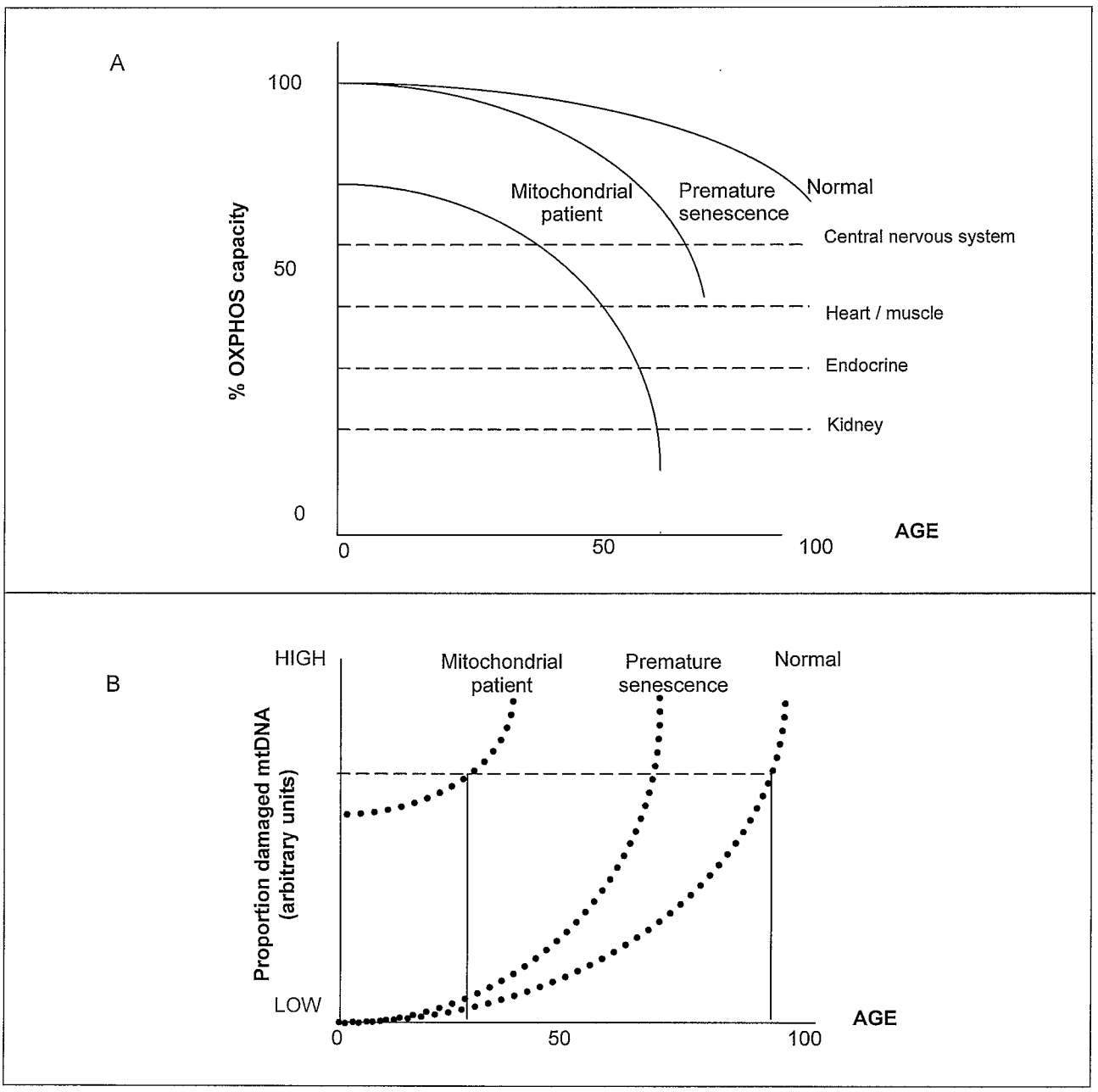


Figure A represents the age-related decrease of OXPHOS individuals with normal OXPHOS, mutant OXPHOS and with an increased mtDNA somatic mutation rate. Figure B represents the relative proportion of mutant mtDNA with age for each of the previously mentioned individuals. Dashed horizontal lines, in both figures, represent different tissue-specific expression thresholds. In Figure B damaged mtDNA is represented by •••••••• and OXPHOS capacity by ————. Adapted from Wallace (1995).

Individuals born with a normal OXPHOS mtDNA genome would have to accumulate a substantial amount of somatic mtDNA mutations to cross the threshold, consequently developing tissue and organ degeneration at a later age. In contrast to this are the

individuals with inherited mtDNA mutations who thus have an initial lower OXPHOS capacity and would need fewer mtDNA mutations to cross the threshold and develop degenerative symptoms (Wallace, 1992). The threshold effect explains how the concentration of mtDNA mutations have to cross a critical limit before that mutation is manifested in a phenotype (Wallace, 1992). Examples of such a degenerative disease associated with the accumulation of somatic mtDNA mutations with age and the subsequent erosion of OXPHOS, include Alzheimer's disease (Corral-Debrinski *et al.*, 1994). The prevalence of disease phenotypes varies between the developing and developed world. The variation in the genetic basis of these diseases was demonstrated to affect the susceptibility to certain phenotypes. An example includes haemaglobinopathies, prevalent in sub-Saharan Africa, which offer protection to those affected by malaria. In addition it is suggested that an increased urbanised Western lifestyle might be the cause of the more frequent observation of diseases such as diabetes, obesity and hypertension in the sub-Saharan African population (Colilla *et al.*, 2000). Thus, by comparing African populations to populations of African ancestry outside Africa, it may be possible to identify the role and interaction of environmental and genetic risk factors contributing to certain complex diseases (Tishkof and Williams, 2002).

2.11 MITOCHONDRIAL DNA VARIATION AND HUMAN ORIGINS

The evolutionary rate of mtDNA reflects the history of the maternal lineage. Mitochondrial mutations have accumulated in female lineages over the ca. 150,000 years of human evolution, since mtDNA sequences have diverged as women migrated from Africa into different continents (Wallace, 1995). This resulted in the subsequent accumulation of neutral, or near neutral mtDNA polymorphisms, which are continent-specific (Wallace, 1995). These polymorphisms are in association with specific haplotypes or haplogroups, and therefore related haplotypes (Wallace, 1995).

Knowledge of human evolution is utilised to understand patterns of variation and to identify alleles or genotypes that might be disease-causing. There are three main theories for modern human evolution, namely the multiregional hypothesis, the out-of-Africa hypothesis and thirdly, the assimilation model (Lahr and Foley, 1998; Tishkoff and Williams, 2002).

The multiregional model for human evolution states that there was no single geographical modern human origin. This model suggests that *Homo erectus* (*H. erectus*) radiated from

Africa into Eurasia ca. 800,000 to 1.8 million years before present (YBP), after which there were independent transitions in populations from *H. erectus* to *H. sapiens* (Tishkoff and Williams, 2002). The continuity of certain fossil records' morphological traits (for example robust cheekbones of *H. erectus* fossils from southeast Asia and in modern Australian aborigines) supports this model. It also suggests that modern populations evolved in the regions where they are found today. Evolution from *H. erectus* to *H. sapiens* that occurred simultaneously in dispersed populations occurred through gene flow between the geographically diverse populations, which would require a large effective population size (Tishkoff and Williams, 2002).

The second, out of Africa model, proposes that all non-African populations are descendants from an *H. sapiens* ancestor that evolved 100,000 to 200,000 YBP in Africa. This recent African origin model hypothesises that this ancestor migrated and spread throughout the world, leading to the replacement of archaic Homo populations, such as the Neanderthals (Templeton, 2002; Tishkoff and Williams, 2002). Support for this model includes the earliest modern human fossils dating to 90,000 to 120,000 YBP that were found in Africa and the Middle East (Lahr and Foley, 1998). According to this model all genetic lineages derive from a common African ancestor. Non-African populations accordingly carry a subset of the genetic variation that is present in modern African populations.

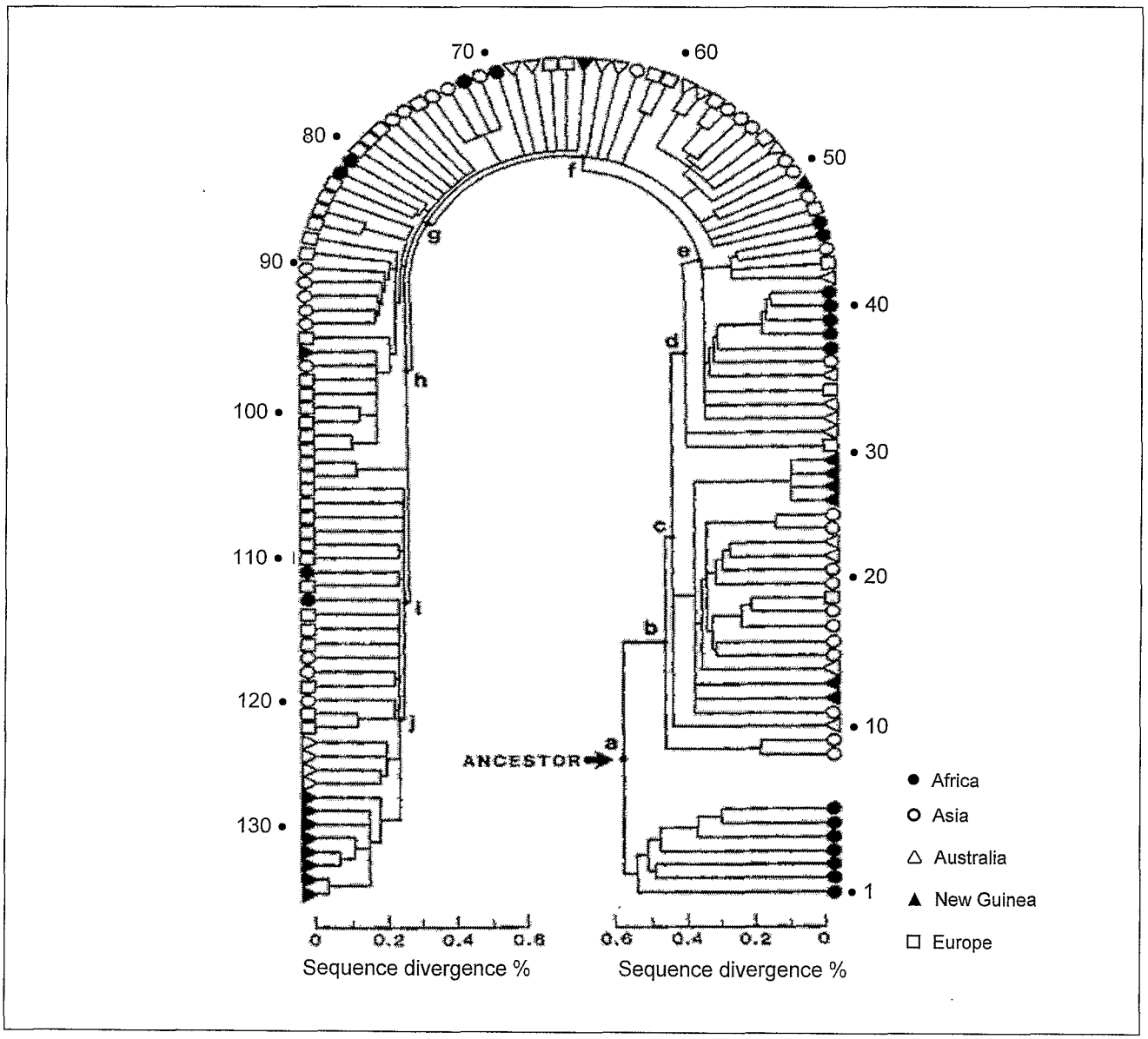
The assimilation or hybridisation model proposes that the evolution of modern humans could have occurred as a blending of African populations' modern characteristics with the archaic characteristics of Eurasian populations (Templeton, 2002). It also proposes unequal gene flow between early human populations over time and space. Thus, according to this model, the modern human gene pool is derived from variable contributions from archaic African and non-African populations' genes.

2.11.1 Global mtDNA phylogeny

In the study of Denaro *et al.* (1981) the authors observed that the variation of mtDNA correlated with the ethnicity and geographic origin of the individual. These authors investigated an *Hpa* I restriction fragment length polymorphism (RFLP) in mtDNA of African, Asian and European-American individuals. An *Hpa* I site was observed at nucleotide 3,592 in 96% of Pygmies, 93% of Khoi-San and 71% of Bantus, but was absent from Asian and European mtDNAs. The restriction site is generated by a C to T transition

at nucleotide 3,594. Denaro *et al.* (1981) also observed the absence of an *Hpa* I restriction site at nucleotide 12,406 in ca. 13% of Asians. Furthermore, mtDNA analysis using six (*Ava* II, *Bam* HI, *Hae* II, *Hinc* II, *Hpa* I and *Msp* I) and 12 (*Alu* I, *Ava* II, *Dde* I, *Fnu* DII, *Hae* III, *Hha* I, *Hinf* I, *Hpa* I, *Hpa* II, *Mbo* I, *Rsa* I, *Taq* I) restriction enzymes in the studies of Johnson *et al.* (1983) and Cann *et al.* (1987) respectively, supported the observation that variation in mtDNA correlated with the individuals' ethnicity and geographic origin. These authors suggested that all global mtDNA form one phylogenetic tree and that Africa shows the greatest variation and deepest root of this tree, which is in accordance with an African origin for humans, as depicted in Figure 2.10. Two distinct features can be observed in Figure 2.10. One branch consists of only African populations, while the other branch consists of all the other populations under investigation.

Figure 2.10: Genealogical tree for human mtDNA



The arrow indicates the root or common ancestor of the tree. Numbers correlate to mtDNA types, number 1 being from the !Kung cell line (GM3043), number 45 from HeLa cells and 110 from the Cambridge reference sequence. Percentage (%) sequence divergence is indicated on the scales at the bottom of the picture. Adapted from Cann *et al.* (1987).

Cann *et al.* (1987) used an estimated sequence evolution rate of 2-4% nucleotides per million years to determine the age of the human mtDNA tree of 200,000 years. This supports the mitochondrial Eve hypothesis that suggest the most recent common ancestor (MRCA) of all mitochondrial lineages was a woman. This does not imply that Eve was the only individual, but suggests that her mtDNA was the only surviving lineage (Cann *et al.*, 1987).

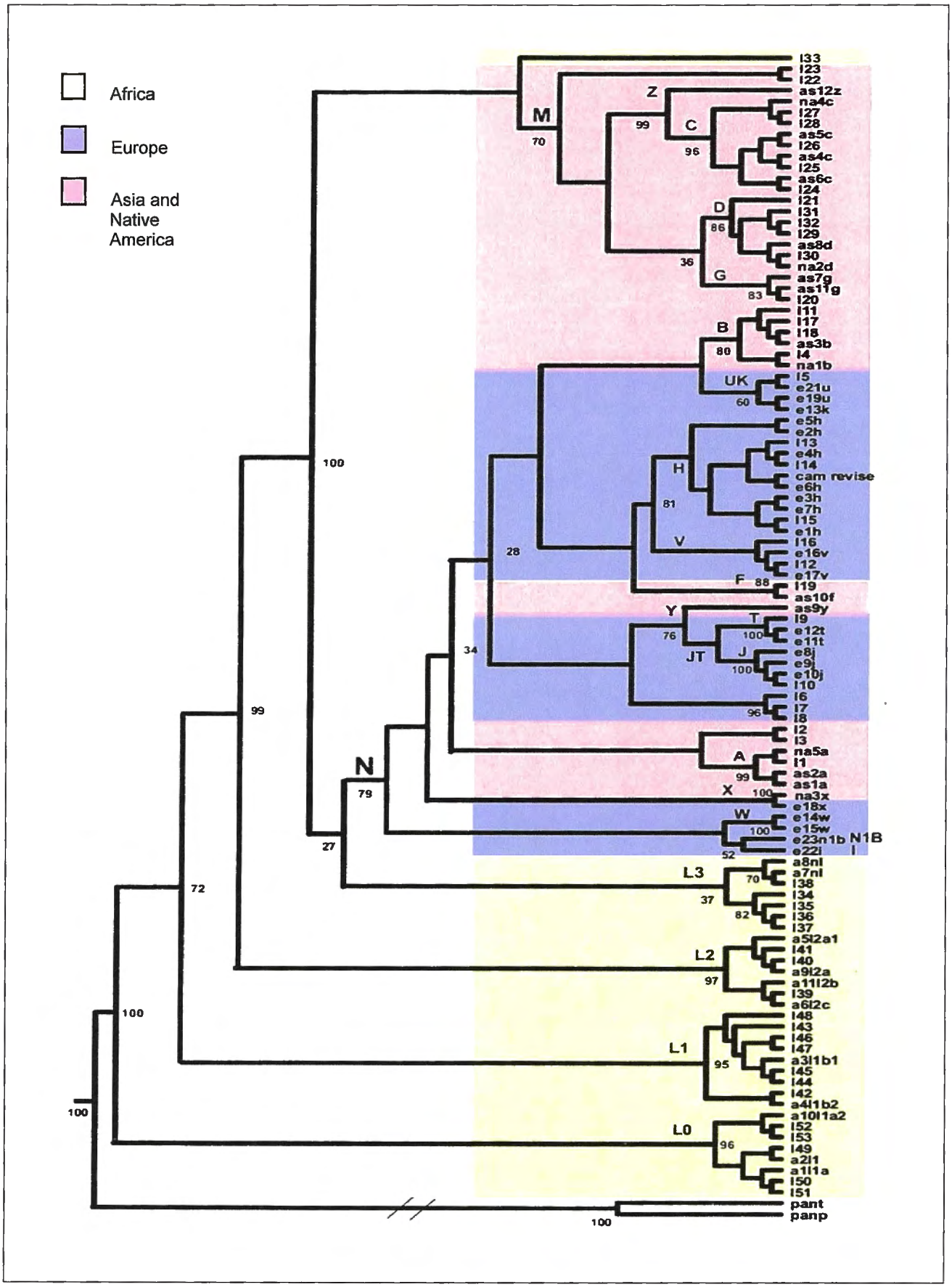
Sequence analysis of the mtDNA noncoding control region (Vigilant *et al.*, 1991), which is 1,122 bp in size and has a mutation rate that is three to four times faster than that of coding mtDNA, also confirmed the diversity in Africans to be the greatest. The time of coalescence of the mtDNA phylogenetic tree was estimated to be 166,000 to 249,000 YBP. Sequence analysis of the control region allowed for maximum resolution when attempting to distinguish very closely related mtDNAs. From the above-mentioned studies it was suggested that the origin of human mtDNA was in Africa and that as women migrated into new lands, mitochondrial mutations became fixed and resulted in mtDNA variation that is continent-specific (Wallace, 1995).

2.11.2 Variation in African mtDNA

Phylogenetic analysis of African mtDNA, which included individuals from Senegal, Eastern Pygmies from Zaire and Western Pygmies from the Central African Republic (Chen *et al.*, 2000), revealed that 55 of the 79 haplotypes formed the "L" African specific haplogroup that is represented in Figure 2.11. This haplogroup is defined by an *Hpa* I site, at nucleotide 3,592, as well as a *Dde* I site at nucleotide 10,394, which is caused by an A to G transition at bp 10,398. Macrohaplogroup L, redefined as L* by Chen *et al.* (2000), can be divided into two major haplogroups: L1 and L2 that encompass 39% and 61% of the L haplotypes respectively.

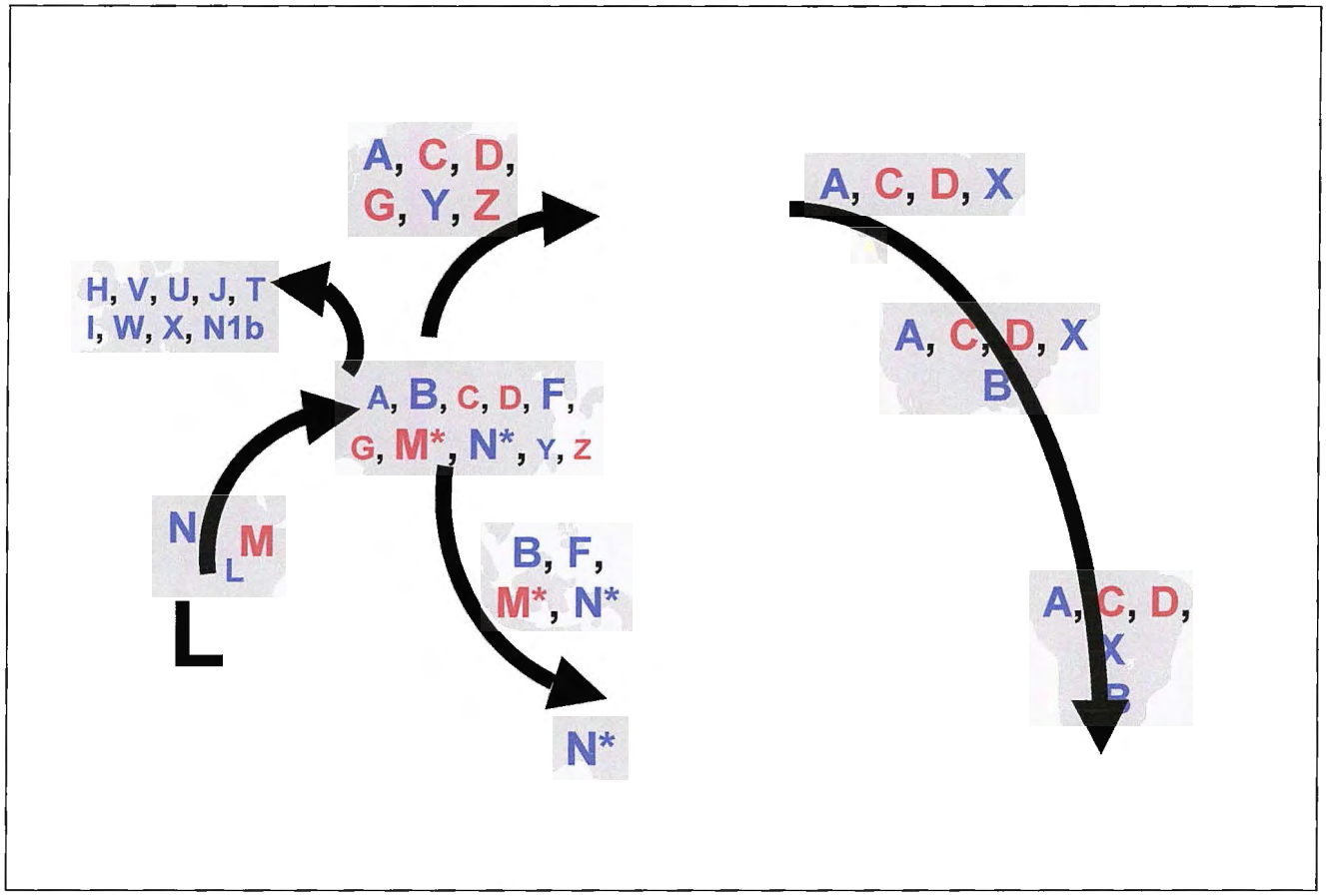
Haplogroup L1 is defined by a *Hinf* I site gain at bp 10,806 and L2 by a combined *Hinf* I site gain at bp 16,389 and an *Ava* II site deletion at bp 16,390 (Chen *et al.*, 1995). These authors confirmed the existence of a third L haplotype, L3, which lacks the *Hpa* I restriction enzyme (RE) site at nucleotide 3,592. It was suggested (Chen *et al.*, 2000) through phylogenetic analysis that the L3 haplogroup is the precursor of non-African haplogroups M and N of whom a subset migrated out of Africa, ca. 60,000 to 80,000 YBP, to establish the ancestors of modern Eurasia (Figure 2.12).

Figure 2.11: Consensus neighbour-joining tree of mtDNA representing the African-specific haplogroup L



Numbers = bootstrap values. Colours represent continental origins of individuals included in this analysis: yellow = Africa; purple = Europe; pink = Asia and Native America; pant = *Pan trodlodytes*; panp = *Pan paniscus*. Capital letters correspond to specific haplogroups. Adapted from Mishmar *et al.* (2003).

Figure 2.12: Schematic representation of world migrations of mtDNA haplogroups



Capital letters represent the corresponding haplogroups. Font sizes correspond to the relative haplogroup population size. Newly discovered haplogroups are represented by M* and N*. Haplogroups derived from L, M and N are illustrated in black, red and blue respectively. Adapted from Ruiz-Pesini *et al.* (2004).

Calculation of the sequence diversity revealed that macrohaplogroup L* has the highest sequence diversity of the continent-specific haplogroups and that Africa is the continent with the greatest diversity (Wallace, 1995). Chen *et al.* (1995) suggested the estimated age of the origin of haplogroup L predates the expansion of *Homo sapiens sapiens* from Africa, further supporting the African origin of modern humans.

Cann and Wilson (1983) observed two rearrangements in haplogroup L1: a 9 bp deletion in the COII/tRNA^{Lys} and a 10-12 bp insertion of Cs. These authors estimated this African-specific haplogroups' coalescence date to be ca. 126,000 to 165,000 YBP.

Haplogroup L1 is the oldest of the macrohaplogroup L* lineages and is mainly represented by the South African San populations together with the Biaka Pygmies from the Central African Republic. Apart from being the oldest human lineage, Chen *et al.* (2000) also suggested macrohaplogroup L* to be the root of human mtDNA from which the other haplogroups evolved. Soodyall *et al.* (1996) observed that the intergenic 9-bp deletion, present in central Africans, Pygmies and Bantu-speaking South African populations, was

absent from Khoi-San-speaking populations and was rare in populations from west and southwest Africa. This suggested the Khoi-San to be a more ancient population and that the deletion arose in central Africa and spread via Bantu migrations to the southern regions of Africa.

According to Chen *et al.* (1995) lineages L2 and L3 diverged from L1 ca. 59,000 to 78,000 YBP, on the basis of RFLP data, with the split of L3 from L2 soon after the split from L1. The split of L1 from macrohaplogroup L* is suggested to have occurred 86,000 to 113,000 YBP (Chen *et al.*, 1995). The Mbuti Pygmies from the Democratic Republic of the Congo and west African Bantu-speaking populations are representatives of the L2 lineage, while the L3 lineage is rarely represented in sub-Saharan Africa, but widely distributed throughout east Africa.

Haplogroup L1 can further be divided into subhaplogroups L1a and L1b, of which L1a is the most ancient (Chen *et al.*, 2000). Further subdivisions of these subhaplogroups include lineages $L1a_1$ and $L1a_2$ as well as lineages $L1b_1$ and $L1b_2$. The two most ancient and distinct African populations, the Vasikela !Kung and the Biaka Pygmies, are representatives of subhaplogroups $L1a_2$ and $L1b_2$ respectively (Chen *et al.*, 2000). This view is supported by the high intragroup sequence diversity of both these groups. This is in correlation with the study of Maca-Meyer *et al.* (2001) where rooting of the phylogenetic tree, utilising chimpanzee sequences, showed the oldest lineage of extant modern humans to be the L1a cluster.

Moreover, Chen *et al.* (2000) suggested the deepest root of the African phylogenetic tree to be represented by the Vasikela !Kung lineage, relative to the chimpanzee outgroup. The two subhaplogroups L1a and L1b have been redefined and are at present known as L0 and L1 respectively (Mishmar *et al.*, 2003). An example of an L0-specific phylogenetic tree is presented in Appendix C, with permission from MAMMAG (Centre for Molecular and Mitochondrial Medicine and Genetics). This tree was constructed from the coding regions of 12 individuals (Ruiz-Pesini *et al.*, 2004). The structure and variation of the internal sequences of the L0 haplogroup have not been studied in detail.

Similar to the L1 divisions, haplogroup L2 can be divided into four subhaplogroups, namely L2a, L2b, L2c and L2d (Chen *et al.,* 2000; Torroni *et al.*, 2001). The Mbuti Pygmies are representatives of the L2a subhaplogroups whereas the Bantu-speaking Senegalese population represents L2c. According to Torroni *et al.* (2001) the L2d clade, although not

frequently represented in the authors' sample population, is the most divergent L2 clade and branched earliest within the L2 haplogroup.

Haplogroup L3 can be divided into four subhaplogroups (Chen *et al.*, 1995) of which L3a and L3b are of a more ancient origin than L3c and L3d (Chen *et al.*, 2000). The divergence time for the major African mitochondrial haplogroups was estimated by Chen *et al.* (2000) and is shown in Table 2.4.

Table 2.4: Sequence divergence times for African mtDNA

Haplogroup	Sequence divergence frequency	Divergence time (YBP)
L*	0.364	125,500-165,500
L	0.356	122,800-161,800
L1	0.328	113,100-149,100
L1a	0.265	91,400-120,500
L1a1	0.166	57,200-75,500
L1a2	0.119	41,000-18,600
L1b	0.214	73,800-97,300
L1b1	0.041	14,100-18,600
L1b2	0.246	84,800-111,800
L2	0.171	59,000-77,700
L2a	0.113	39,000-51,400
L2b	0.072	24,800-32,700
L2c	0.052	17,900-23,600
L3	0.227	78,300-103,200
L3a	0.120	41,400-54,500
L3b	0.225	77,600-102,300
L3c	0.082	28,300-37,300
L3d	0.051	17,600-23,200

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

2.11.3 European mtDNA haplogroups

From the study of Torroni *et al.* (1994) European mtDNA was divided firstly into two groups, 25% bearing the *Dde* I restriction site at nucleotide 10,394 and 75% lacking this site. In Europe, macrohaplogroup N contributes mainly to the observed European

haplogroups. In addition to this division, 10 European mtDNA haplogroups, encompassing 99% of all European mtDNA, have been identified. These include haplogroups H, I, J, K, M, T, U, V, W and X (Torroni *et al.*, 1996). Fifty percent of Europeans harbour haplogroup H, characterised by the absence of the *Dde* I site at nucleotide 10,394 and an *Alu* I site at nucleotide 7,025. Haplogroups I, V, W and X are also observed in European populations, although they appear to be less frequent (Torroni *et al.*, 1996). Since haplogroups H, I, J, K, T and W are mainly confined to European populations it is suggested that these haplogroups originated when the Caucasian ancestors diverged from the modern Africans and Asians (Torroni *et al.*, 1994).

2.11.4 Asian and Native American haplogroups

Similar to European mtDNA, Asian mtDNA can also be divided into two main groups based on the presence or absence of the *Dde* I site at nucleotide 10,394. In addition, Asians harbouring the *Dde* I site also have an *Alu* I site at nucleotide 10,397 and are representative of haplogroup M (Torroni *et al.*, 1993). This association, which is not observed in Africans or Europeans, suggests that the *Alu* I mutation occurred shortly after the arrival of women in Asia. Two other haplogroups that are prominent in Asians are F and G.

The diversification of macrohaplogroups M and N contributed equally to the variation in mtDNA sequences in Asia. In the study of Mishmar *et al.* (2003) it was observed that 75% of Siberian mtDNAs were represented by haplogroups A, C, D, G, Z and Y in contrast to the 14% representation by haplogroups A, C, D and G in mtDNA from regions south of Tibet and Korea. Haplogroups Y and Z were rarely observed in these latter regions. Native American mtDNA is primarily represented by haplogroups A, B, C, D and X, although these haplogroups are occasionally observed in Asian populations (Torroni *et al.*, 1994). Table 2.5 lists RE characteristic of the specific haplogroups.

Table 2.5: Restriction enzyme sites defining continent-specific mtDNA haplogroups

Population	Haplogroup	mtDNA Variants
AFRICA	Macro-L	3,592+ *Hpa* I
EUROPE	I	10,394+ *Dde* I; 1,715- *Dde* I; 10,028+ *Alu* I; 8,249+ *Ava* II
EUROPE	K	10,394+ *Dde* I; 9,052- *Hae* II; 12,308+ *Hinf* I

Table 2.5: continued ...

EUROPE	H	7,025- *Alu* I
EUROPE	J	10,394+ *Dde* I; 13,704- *BstN* I;
EUROPE	T	10,394- *Dde* I; 13,366+ *Bam* HI; 15,606 +*Alu* I
EUROPE	U	10,394- *Dde* I; 12,308+ *Hinf* I
EUROPE	V	10,394- *Dde* I; 4,577- *Nla* III
EUROPE	W	10,394- *Dde* I; 8,249+ *Ava* II; 9,052+ *Hae* II
EUROPE	X	10,394- *Dde* I; 1,715- *Dde* I
ASIA	Macro-M	10,394+ *Dde* I; 10397+ *Alu* I
ASIA	A	663+ *Hae* III
ASIA	B	8271-8281 9bp deletion
ASIA	F	12406- *Hpa* I / *HInc* II; 16517+ *Hae* III
ASIA	C	13,259-/13262+ *Hinc* II / *Alu* I
ASIA	D	5176- *Alu* I
ASIA	E	7,598- *Hpa* I
ASIA	G	4,830+*Hae* III; 12,406- *Hpa* I / *Hinc* II
SIBERIA	Y	7,933+ *Mbo* I; 8,391- *Hae* III; 10,394+ *Dde* I; 16,517+ *Hae* III
SIBERIA	Z	11,074+ *Dde* I; 16,517+ *Hae* III

Adapted from MITOMAP (2003).

Although it was previously hypothesised that the differences in global mtDNAs from different geographic regions were caused by genetic drift, Mishmar *et al.* (2003) suggested that the variants might be attributable to natural selection. These authors also suggest that natural selection may be associated with certain polymorphisms observed in specific populations in different temperature regions. Support for the observation that different human mtDNA haplogroups are associated with different biological functioning includes reduced sperm motility associated with European haplogroup T (Ruiz-Pesini *et al.,* 2000). Further support includes the higher probability of blindness when the ND6 and ND4L missense mutations, at nucleotide positions 14,484 and 10,663 respectively, associated with LHON, are present on the European J haplogroup (Brown *et al.,* 2001).

The different mtDNA lineages that are associated with variable functionality might imply that advantageous variants in one population might be disadvantageous or even pathogenic in another population that is adapted to different climates and diets

(Mishmar *et al.*, 2003). This could be a possible explanation for the occurrence of disorders such as obesity, diabetes and neurodegenerative diseases, which might have been favourable ancient polymorphisms that became pathogenic over time, with the migration into new climatic regions and the adoption of new lifestyles.

2.12 WORLD MIGRATIONS

Genetic, archaeological and linguistic data suggest migrations occurred in Africa during the last several thousand years that culminated in the mtDNA distribution of modern African populations (Tishkoff and Williams, 2002). The Bantu-speaking populations, which consisted mainly of farmers, had the most significant migration out of Cameroon into southern Africa in the past 3,000 years (Guthrie, 1962). Similarly there have also been migrations of non-Africans, resulting in the admixture with native Africans. During the last few centuries one of the largest migration events occurred when Africans were brought to Europe and America as slaves. This also resulted in admixture with the native European and American populations, resulting in heterogeneous gene pools.

Watson *et al.* (1997) suggested that a small African sublineage might have acquired some advantage, for example in response to an environmental change such as the onset of the Last Glacial period. It could be speculated that 60,000 to 80,000 YBP a behavioural innovation appeared in an anatomically modern human subpopulation, which included the L3 ancestors, who previously implemented Middle Palaeolithic or Stone Age technology. A subset of this population migrated and expanded from Africa (Figure 2.12). Evidence for this hypothesis includes the increase in communicative activity associated with the Later Stone Age or an earlier period in Africa and the Upper Palaeolithic period in Eurasia. Although the exact cause of the expansion is unknown, it is suggested to have originated in Africa and was the decisive event in the modern human migration from Africa to Eurasia (Watson *et al.*, 1997).

Maca-Meyer *et al.* (2001) suggested that two independent lineages had spread out of Africa. One southern route radiated ca. 30,000 to 57,600 YBP and included all M haplogroup lineages with descendants prevalent in India and eastern Asia (Figure 2.12). Phylogenetic analyses revealed star-shaped trees that suggested that these radiations could have occurred in a short time period. There are also possibilities of a back migration from Asia returning to Africa (Quintana-Murci *et al.*, 1999). A second northern migration, ca. 43,000 to 53,000 YBP, from Africa is also suggested by Maca-Meyer *et al.* (2001). This

event gave rise to haplogroup A, which is widespread in Asia, and haplogroup X, which is prevalent in Europe. Support for this back and forth migration from Africa has also been suggested by Y-chromosome analysis (Hammer *et al.*, 1998).

Some Africans are susceptible to complex diseases, such as hypertension and prostate cancer, in which the role of the genetics is poorly understood in both Africans and non-Africans (Tishkoff and Williams, 2002). It is therefore necessary to analyse genetic variation in diverse African populations to elucidate these populations' genetic structures.

2.13 Y-CHROMOSOME HAPLOTYPE ANALYSIS

Similar to mtDNA, the Y-chromosome also harbours polymorphisms that allow for its evolutionary reconstruction (Hammer *et al.*, 1998). Ten haplotypes have been observed for Y-chromosome specific markers in the study of Hammer *et al.* (1998). The most ancestral Y-chromosome haplotype is represented in Sudanese and Ethiopians from east Africa together with southern African Khoi-San speaking populations (Semino *et al.*, 2002). This is in concordance with mtDNA data indicating that the most genetically diverse population in Africa is the Khoi-San. Moreover, it supports the suggestion that the origin of these Khoi-San-speaking individuals was in east Africa, after which they migrated into southern Africa (Semino *et al.*, 2002).

2.14 CAMBRIDGE REFERENCE SEQUENCE

The mitochondrial CRS is utilised as the premier mtDNA reference worldwide. Recently studies have revealed discrepancies in this reference sequence, implying either experimental errors or the presence of rare polymorphisms (Howell *et al.*, 1992). During reanalysis of the CRS it was observed that 11 nucleotides were incorrect and seven additional base pairs represent rare polymorphisms (Andrews *et al.*, 1999). This reanalysed sequence is known as the RCRS, with Genbank Accession number J01415. It is of great concern that this reference sequence was derived from a single European individual representing haplogroup H, together with bovine and HeLa cell mtDNA sequences (Anderson *et al.*, 1981). Comparative analysis utilising the CRS will more often than not lead to erroneous identification of alterations. It is essential that any reference sequence be a true reference, and thus not contain rare polymorphic alleles or errors.

2.15 AIMS OF THIS STUDY

The primary aim of this study was to generate an African-specific haplogroup consensus sequence. If this study revealed that discrepancies from the CRS are population-specific polymorphisms, it would imply that a mitochondrial reference, or consensus, sequence will have to be derived for each haplogroup. The African consensus sequence could also be utilised to investigate how the mitochondrial genome evolved over time. Possible disease genotypes could be compared to this reference to exclude population-specific polymorphisms and narrow down possible pathogenic candidate changes.

2.15.1 Specific aims

This study included the following specific aims:

a) Sequencing of the entire mitochondrial genome of 10 Khoi-San individuals.
b) Generating a Khoi-San-specific mitochondrial consensus sequence.
c) Investigating the mtDNA variation within these generated sequences via:
    i) Phylogenetic analyses to determine the evolutionary relationship between the sequences.
    ii) Statistical analyses to investigate the correlation of the derived sequences against the null hypothesis of neutral evolution.

CHAPTER THREE

MATERIALS AND METHODS

Chemicals utilised during this study were of analar quality and were obtained from Sigma®[1] or Roche®[2], unless stated otherwise. Protocols were performed according to the manufacturer's specifications and modifications are stated in the text. The Ethics Committee of the Potchefstroom University for Christian Higher Education granted ethical approval for this study, with approval number 02M02, and all subjects were included after obtaining informed consent.

3.1. SAMPLE POPULATION

The sample population was randomly selected and consisted of previously collected and extracted DNA from individuals from the Khoi-San, and specifically the Vasikela !Kung population (Chen *et al.*, 2000). Individuals were identified as !Kung when both parents belonged to this ethnic group or via verbal confirmation of their ethnicity, as stated in Chen *et al.* (2000). Furthermore, identity documents were utilised to confirm each individual's ethnic group.

3.2 DNA ISOLATION

Genomic DNA (gDNA) was previously isolated from buffy coats utilising the Puregene™[3] DNA Isolation Kit. Red Blood Cell Lysis Solution removed red blood cells and the subsequent addition of Cell Lysis Solution ensured the lysis of white blood cells. An RNase step was included to free the cell lysate solution from RNA. The subsequent isolated DNA was desalted and concentrated by isopropanol precipitation.

The exact concentration of the isolated DNA was determined via optical density using a Beckman DU-600 spectrophotometer, utilising 20-fold dilutions of the DNA in double distilled water. The concentrations ranged from a minimum of 600 $ng.\mu l^{-1}$ to a maximum of 3.65 $\mu g.\mu l^{-1}$, with an average of 1.55 $\mu g.\mu l^{-1}$. DNA stock solutions were stored at -80°C.

[1] Sigma® is a registered trademark of Sigma-Aldrich Corporation, St. Louis, Missouri, MO, U.S.A.
[2] Roche®is a registered trademark of Hoffmann-La Roche Ltd., Basel, Switzerland.
[3] Puregene™ is a trademark of Gentra Systems, Inc., Minneapolis, MN, U.S.A.

3.3 POLYMERASE CHAIN REACTION (PCR)

The mitochondrial genome of each individual was amplified using nine overlapping primer pairs, listed in Table 3.1. Conditions for the primers were optimised mainly with regard to the annealing temperature (T_a) and magnesium chloride ($MgCl_2$) concentration.

Table 3.1: Primer pairs utilised for amplification of the whole mitochondrial genome

Primer pair	5'-end*	3'-end*	Primer name	Sequence	T_m	Product size
1	131	150	131-F	5'-tctttgattcctgcctcatc-3'	53	2,328
	2,458	2,439	2,458-R	5'-acgagtattcctttccaatt-3'	49	
2	2,206	2,225	2,206-F	5'-caagctcaacacccactacc-3'	57	2,747
	4,952	4,934	4,952-R	5'-gtgagagagttagaatagg-3'	49	
3	4,500	4,519	4,500-F	5'-tctaccatctttgcaggcac-3'	55	694
	5,193	5,174	5,193-R	5'-gttcgattgtactgattgtg-3'	51	
4	5,146	5,165	5,146-F	5'-cgaccctactactatctcgc-3'	57	2,483
	7,628	7,610	7,628-R	5'-gatgttctgcgatgaaggg-3'	53	
5	7,367	7,384	7,367-F	5'-cctccataaacctggagtg-3'	53	2,494
	9,859	9,840	9,859-R	5'-agttgaaaggagtgatagac-3'	51	
6	9,384	9,403	9,384-F	5'-gatgtaacacgagaaagcac-3'	57	2,485
	11,868	11,851	11,868-R	5'-gcgattggagcggaatgg-3'	53	
7	11,141	11,158	11,141-F	5'-cccaccttggctatcatc-3'	51	2,523
	13,663	13,644	13,663-R	5'-gaaggggtgggaatgattgt-3'	55	
8	13,172	13,190	13,172-F	5'-gcttaggcgctatcaccac-3'	55	1,407
	14,578	14,559	14,578-R	5'-gctggtgtggcgattgttag-3'	57	
9	14,260	14,279	14,260-F	5'-atcctcccgaatgaaccctg-3'	57	2,739
	429	408	429-R	5'-accgccatacgtgaaaattgtc-3'	59	

* The 5'-and 3'-end positions of the primers are indicated by their nucleotide number. T_m represents the melting temperature that was calculated for each primer. Primer sequences were obtained from Wallace (1999).

The T_a of primer pairs is dependent on the calculation of their melting temperature, T_m. According to Thein and Wallace (1986), the T_m of 18 base primers was calculated via Equation 3.1.

Equation 3.1: Calculation of the melting temperature (T_m) of a single primer

$$T_m = 2(G+C) + 4(T+A)$$

From Thein and Wallace (1986).

Optimisation of a primer was generally started at two degrees below the mean T_m of the two primers. Overestimation of the T_m of primers longer than 18 nucleotides can be overcome by using an alternative equation, Equation 3.2, derived by Meinkoth and Wahl (1984).

Equation 3.2: Calculation of the melting temperature (T_m) of primers longer than 18 bases

$$T_m = 81.5\ °C + 16.6 \log M + 0.41\ (\%\ GC) - 0.61\ (\%\ formamide) - 500.n^{-1}$$

M = ionic strength in mole per litre, n = primer length in bases. From Meinkoth and Wahl (1984).

Where necessary, the DNA amplification was performed in duplicate to ensure accurate results and sufficient amounts of product. Negative controls were included in PCR reactions to detect if contamination occurred. PCR reactions were prepared on ice in a total volume of 25 µl. Each reaction consisted of 1X PCR buffer [50 mM KCl, 10 mM Tris®-HCl (pH 9.0) and 0.1% Triton® X-100], 2.0 mM $MgCl_2$, 200 µM of each deoxynucleotide triphosphate (dNTP), 10 picomoles (pmol) of each of the appropriate primers, one unit *Taq* polymerase and 30 ng of gDNA. The MT Research PTC-200 Peltier Thermal cycler contained a heated lid function that prevented evaporation of the reactions during amplification, and no mineral oil was thus required.

Amplifications were performed by first denaturing the samples at 94°C for 5 minutes, followed by 35 cycles at 94°C for 30 seconds, 1 minute at the appropriate Ta, an elongation step for 30 seconds at 72°C, followed by a final extension step at 72°C for 5 minutes. Choosing a T_a of 51°C allowed all nine the primer pairs to be amplified in the same block during a single PCR run. Results were analysed by agarose gel electrophoresis.

3.4 GEL ELECTROPHORESIS

Amplified PCR products were analysed on 2% agarose gels and electrophoresed at 16 volts per centimetre, for 20 minutes in 1X TBE buffer. Five microlitres of the amplified products were mixed with 2 µl of loading buffer [0.03% bromophenol blue, 0.03% xylene cyanol FF, 0.4% Orange G, 15% Ficoll®[1] 400, 10 mM Tris-HCl (pH 7.5), and 50 mM EDTA (pH 8.0)] prior to loading the gel.

[1] Ficoll® is a registered trademark of Pharmacia Biotech AB, Piscataway, NJ, U.S.A.

Molecular Weight Marker X (Roche®) was loaded on each gel, as a reference, from which the fragment sizes of the samples could be estimated. Agarose gels were visualised under ultraviolet light using ethidium bromide (EtBr) in a final concentration of 0.5 $\mu g.ml^{-1}$.

3.5 AUTOMATED SEQUENCING

Mitochondrial DNA of the extracted DNA from the Khoi-San individuals was sequenced with 48 sequencing primers (Wallace, 1999), listed in Table 3.2. The nine fragments, generated with primers listed in Table 3.1, were each sequenced in ca. 600 bp fragments. This ensured the correct and complete mitochondrial sequence of both the H- and the L-strands from each individual.

Table 3.2: Sequencing primers utilised for sequencing of the mitochondrial genome

PCR fragment	Primer name	Sequencing primer	5'-End*	3'-End*	T_m
1	131-F	5'-tctttgattcctgcctcatc-3'	131	150	53
	371-F	5'-ctaacaccagcctaaccaga-3'	371	390	55
	919-F	5'-agtcaatagaagccggcg-3'	919	936	18
	1,064-R	5'-cttgtgtgttatcgattctg-3'	1,064	1,045	51
	1,800-R	5'-catggcgttccctttctac-3'	1,800	1,782	53
	2,005-F	5'-cctggtgatagctggttg-3'	2,005	2,022	51
	2,079-R	5'-tgtcttgggagatttagggg-3'	2,079	2,060	55
2	2,772-F	5'-cacaggtcctaaactacc-3'	2,772	2,789	49
	2,900-R	5'-gatgatatgagttaactagg-3'	2,900	2,881	49
	3,108-F	5'-ttcaaattcctccctgtacg-3'	3,108	3,127	53
	3,388-F	5'-ctaggctatatacaactacgc-3'	3,388	3,408	55
	3,669-F	5'-gcatcaaactcaaactacgc-3'	3,670	3,689	57
	4,308-F	5'-ggagcttaaacccccctta-3'	4,308	4,325	49
	4,678-R	5'-cgttggcgtaggtattagga-3'	4,678	4,659	55
3	4,831-F	5'-gcacccctctgacatcc-3'	4,831	4,847	51
4	5,317-F	5'-ccaccatcaccctcctt-3'	5,317	5,333	49
	5,971-F	5'-gcgcatgagctggagtcc-3'	5,971	5,988	55
	6,149-F	5'-agttcccctaataatcggtg-3'	6,149	6,168	53
	6,449-R	5'-cctcttcgtctgatccg-3'	6,449	6,465	49
	6,638-R	5'-cagtgggacttcaaatataa-3'	6,638	6,619	49
	7,006-R	5'-agtgatctgtagcatgatgt-3'	7,006	6,987	51
	7,146-F	5'-actatcatattcatcggcgt-3'	7,146	7,165	51

Table 3.2: continued ...

PCR fragment	Primer name	Sequencing primer	5'-End*	3'-End*	T_m
5	7,131-R	5'-ggatctggtttggatgc-3'	7,131	7,115	49
	7,871-F	5'-accatcaaatcaattggcca-3'	7,871	7,890	51
	8,150-F	5'-ccgggggtatactacgg-3'	8,150	8,166	51
	8,557-F	5'-gcccccacaatcctagg-3'	8,557	8,573	51
	9,082-F	5'-cttccctctacacttatcatcttcacaat-3'	9,082	9,110	75
6	10,357-F	5'-ctaagtctggcctatgagtg-3'	10,356	10,375	53
	10,569-R	5'-taggagggatgatacggatc-3'	10,569	10,550	55
	10,718-F	5'-cacatatggcctagactacg-3'	10,718	10,737	55
	11,141-F	5'-cccaccttggctatcatc-3'	11,141	11,158	51
7	11,464-F	5'-cttaaaactaggcggctatgg-3'	11,464	11,484	57
	11,711-F	5'-gcccacgggcttacatc-3'	11,711	11,727	51
	12,292-F	5'-tggtcttaggccccaaaaat-3'	12,292	13,011	62
	12,376-R	5'-aagtcagggttagggtggtt-3'	12,376	12,357	60
	12,599-F	5'-tattcatccctgtagcattg-3'	12,599	12,618	51
	12,785-F	5'-tatccttcttgctcatcagt-3'	12,785	12,804	51
8	13,197-F	5'-cgcagcagtctgcgccc-3'	13,197	13,213	55
	13,595-F	5'-gcgcctatagcactcga-3'	13,595	13,611	49
	14,578-R	5'-gctggtgtggcgattgttag-3'	14,578	14,559	57
9	14,407-F	5'-caagacctcaaccctga-3'	14,407	14,424	51
	14,618-F	5'-aaccccacaaaccccattac-3'	14,618	14,637	55
	15,021-F	5'-tctgcctcttcctacacatc-3'	15,021	15,040	60
	15,409-F	5'-cccttactacacaatcaaag-3'	15,409	15,428	51
	15,838-F	5'-cctaataccaactatctccc-3'	15,838	15,857	53
	16,225-F	5'-caactatcacacatcaactg-3'	16,225	16,244	51
	16,421-R	5'-gtggtaggaggcactttagt-3'	16,420	16,401	55

* The 5'- and 3'- end positions of the primers are indicated by their nucleotide numbers. T_m represents the melting temperature that was calculated for each primer. Primer sequences obtained from Wallace (1999).

3.5.1 PCR purification

Contamination of PCR products with enzymes and other excess reagents can interfere with further analyses. Therefore, it was crucial to purify PCR products from unconsumed dNTPs, excess primers, salts and *Taq* polymerase before it was utilised for sequencing. ExoSAP-IT®[1] was utilised to achieve this goal.

[1] ExoSAP-IT® is a registered trademark of the USB Corporation, Cleveland, OH, U.S.A.

Two hydrolytic enzymes, Exonuclease I and Shrimp Alkaline Phosphatase, were used to remove the excess dNTPs and primers. Single stranded primers and DNA were degraded by the Exonuclease I enzyme, while unconsumed dNTPs were hydrolysed by Shrimp Alkaline Phosphatase. Two microlitres of the ExoSAP-IT enzymes were directly applied to 5 μl of the amplified product. The enzymes were activated via incubation at 37°C for 15 minutes. Heating the mixture to 80°C for 15 minutes ensured the inactivation of the enzymes.

3.5.2 Cycle sequencing

Automated sequencing using modifications of the dideoxynucleotide triphosphate (ddNTP) chain terminator reaction procedure from Sanger (1977) was performed. Cycle sequencing was performed utilising the ABI Prism®[1] BigDye™[2] Terminator version 3.0 Ready Reaction Cycle Sequencing Kit, based on fluorescent sequencing and containing the AmpliTaq®[3] DNA polymerase sequencing enzyme. Sequencing reactions were performed on ice in final volumes of 20 μl for each reaction. Each reaction contained up to 10 ng of purified PCR product, 1 pmol of the appropriate sequence primer, and 4 μl of the Ready Reaction mixture, containing dye terminators, deoxynucleoside triphosphates, the AmpliTaq DNA Polymerase, $MgCl_2$ and Tris®-HCl buffer (pH 9.0). The cycle sequencing was performed in a Peltier Thermal Cycler. Samples were denatured at 96°C for 1 minute. This was followed by 25 cycles of denaturation at 96°C for 10 seconds, primer annealing at 50°C for 5 seconds and elongation for 4 minutes at 60°C.

Prior to analysis of the samples via electrophoresis, it was essential to remove unincorporated dye terminators, since these substrates could interfere with the results. This was achieved by utilising Centri•Sep 96 Multi-well Filter Plates. The filter plates were firstly centrifuged at 1,500 gravitational acceleration (x *g*) for 2 minutes to separate the inserted columns from the hydrated matrix. The sequenced PCR reactions were accordingly loaded into the filter plates and subsequently precipitated via centrifugation for 2 minutes at 1,500 x *g* and air-dried utilising a vacuum pump.

[1] ABI Prism® is a registered trademark of Applera Corporation, Foster City, CA, U.S.A.
[2] BigDye™ is a trademark of Applera Corporation, Foster City, CA, U.S.A.
[3] AmpliTaq DNA polymerase FS® is a registered trademark of Roche Molecular Systems Inc., Pleasanton, CA, U.S.A.

The precipitated samples were resuspended in Hi-Di™[1] deionised formamide prior to electrophoretic analysis utilising the ABI 16 capillary DNA Sequencer (Model 3100). Results were analysed utilising the ABI 3100 Genetic Analyzer Data Collection Software, version 1.1.

3.5.3 Sequence analysis

Sequence data were compared and edited by utilising the SEQUENCHER™[2] 4.0.5 software package. The generated sequence information was compared to the RCRS and other sequence information in the MITOMAP (2003) databank. Discrepancies from this reference sequence were investigated to determine if they were previously reported or novel polymorphisms and whether they were specific to the Khoi-San population.

3.6 PHYLOGENETIC ANALYSIS

The phylogenetic relationships between the mtDNA sequences of the individuals, compared to that of the RCRS, were investigated and a maximum parsimony (MP) and neighbour-joining (NJ) tree constructed (Saitou and Nei, 1987), with percentages resulting from 1,000 bootstrap replicates, by utilising the version 1.81 Clustal X (Thompson *et al.*, 1997) and MEGA, version 2.1 (Kumar *et al.*, 2002) statistical packages. Further NJ trees were constructed with the addition of 106 L-specific sequences, 12 representing L0, 23 representing L1, 37 representing L2 and 34 representing L3 lineages (Ruiz-Pesini *et al.*, 2004). Permission was granted by MAMMAG to include these 106 sequences in this study. All trees were rooted utilising *Pan trodlodytes*, Genbank Accession number GI 5835121, as an outgroup sequence.

3.7 STATISTICAL ANALYSIS

Sequences were also tested for the hypothesis that mutations are under neutral evolution and specifically subjected to the Tajima's *D* and Fu and Li *D** tests (Tajima, 1989; Fu and Li, 1993) utilising the DnaSP software (Rozas and Rozas, 1999). Tajima's *D* statistic is calculated based on Equation 3.3.

[1] Hi-Di™ is a trademark of Applera Corporation, Foster City, CA, U.S.A.
[2] SEQUENCHER™ is a trademark of Gene Codes Corporation, Ann Arbor, MI, U.S.A.

Equation 3.3: Estimation of Tajima's *D* statistic

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

With $a_1 = \sum (1 / i)$ from i = 1 to n -1

$e_1 = c_1 / a_1$

$e_2 = c_2 / (a^s_1 + a_2)$

$c_1 = b_1 - (1 / a_1)$

$c_2 = b_2 - [(n + 2) / a_1 n] + (a_2 / a_1^2)$

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

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

k = average number of nucleotide differences between pairs of sequences; n = number of DNA sequences; *S* = the total number of segregating sites. Adapted from Tajima (1989).

Fu and Li's *D** test is based on Equation 3.4 (Fu and Li, 1993). The statistical significance of each result from the statistical tests was accordingly estimated.

Equation 3.4: Estimation of Fu and Li's *D test**

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

With $a_n = \sum (1 / k)$ from k = 1 to n – 1

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

$v_{D^*} = [(n / n\text{-}1)^2 b_n + a^2_n d_n – 2 (n a_n(a_n + 1)) / (n – 1)^2] / (a^2_n + b_n)$

n = the number of nucleotide sequences; η = total number of mutations; ηs = total number of singletons. Adapted from Fu and Li (1993).

The frequency distribution of pairwise sequence differences was estimated via the DnaSP version 3.51 software (Rozas and Rozas, 1999). The expected values were estimated, utilising the above-mentioned software, via Equation 3.5 (Slatkin and Hudson, 1991). The results of the expected number of pairwise differences were graphically presented against the observed number of differences.

Equation 3.5: Calculation of the pairwise number of nucleotide differences between sequences

$$Q(i) = \frac{1}{1+\theta}\left(\frac{\theta}{1+\theta}\right)^{i}$$

With $\theta = 2N\mu$

Mean of *i* is θ

Variance of *i* is $\theta(1+\theta)$

2N = a haploid population size and μ representing the mutation rate per generation. Adapted from Slatkin and Hudson (1991) and Watterson (1975).

3.8 COALESCENT DATE ESTIMATES

The genetic divergence times from the MRCA were determined, as described by Mishmar *et al.* (2003). The mean number of substitutions per site (ρ) was estimated by dividing the total number of substitutions in the phylogenetic tree by the number of individuals that were utilised to construct the tree, as illustrated in Equation 3.6.

Equation 3.6: Calculation of the mean number of substitutions per site

$$\rho = \#\text{ substitutions} / \#\text{ individuals}$$

ρ = mean number of substitutions per site. Adapted from Morral *et al.* (1994).

The mean number of substitutions per site was converted to time by utilising a substitution rate of 1.26×10^{-8}, as illustrated in Equation 3.7 (Mishmar *et al.*, 2003). Substitution rates were calculated taking into account only the coding region (15,447 bp) from the sequences, since back and parallel mutations occur frequently in the control region (Equation 3.7).

Equation 3.7: Calculation of the MRCA

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

MRCA = most recent common ancestor; ρ = mean number of substitutions per site. Adapted from Mishmar *et al.* (2003).

Standard deviations (SD) for the MRCA were calculated via Equation 3.8. The SDs provide a relative time period during which the MRCA may have existed, thus the time estimate for an MRCA of 100,000 ± 1,000 YBP is 99,000 to 101,000 YBP.

Equation 3.8: Calculation of standard deviations of the MRCA

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

SD = standard deviation; ρ = mean number of substitutions per site; n = population size. Adapted from Saillard *et al.* (2000).

3.9 ANALYSIS OF NON-SYNONYMOUS AND SYNONYMOUS CHANGES

The influence of selection was estimated by comparing the ratios of non-synonymous (ka) amino acid replacement substitutions, divided by the number of synonymous (ks) silent mutations. The total number of non-synonymous mutations were counted for every branch constituting the phylogenetic tree and divided by the total number of synonymous mutations in the tree. This was estimated for the internal and terminal branches and compared for each of the L-specific lineages. The *P* values were calculated utilising the Fisher's Exact Test via the DnaSP software (Rozas and Rozas, 1999).

3.10 CONSERVATION INDEX

The degree of conservation was determined for every replacement or non-synonymous mutation. Thirty nine different species, listed in Table 3.3, were used to estimate the conservation of the non-synonymous changes (Ruiz-Pesini *et al.*, 2004). The wild type was considered to be the amino acid found in most of the global phylogeny, whereas the mutant was regarded as the amino acid change specific for the sequence or haplotype in question. For every non-synonymous change, the occurrence of a wild type allele was counted and the conservation number regarded as the total number of occurrences of the wild type in the 39 species.

Table 3.3: Species utilised to determine the conservation indices of the non-synonymous changes

Primates	Other mammals	Other vertebrates	Invertebrates
Homo sapiens	*Felix catus*	*Gallus gallus*	*Drosophila melanogaster*
Pan troglodytes	*Phoca vitulina*	*Aligator mississippiensis*	
Pan paniscus	*Rhinoceros unicornis*	*Xenopus laevis*	
Gorilla gorilla	*Equus asinus*	*Cyprinus carpio*	
Pongo pygmaeus	*Equus caballus*		
Macaca sylvanus	*Canis familiaris*		
Papio hamadryas	*Talpa europaea*		

Table 3.3: continued ...

Hylobates lar	*Balaenoptera musculus*		
Cebus albifrons	*Balaenoptera physalus*		
Tarsius bancanus	*Hippopotamus amphibious*		
Nycticebus	*Ovis aries*		
	Bos taurus		
	Sus scrofa		
	Lama pacos		
	Loxodonta africana		
	Oryctolagus cuniculus		
	Sciurus vulgaris		
	Mus musculus		
	Rattus norvegicus		
	Erinaceus europaeus		
	Macropus robustus		
	Ornitorhynchus anatinus		

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

CHAPTER FOUR

RESULTS AND DISCUSSION

The materials and methods stated in the previous chapter were utilised to obtain the results described in this chapter. The specific procedures are evaluated and the results obtained are discussed.

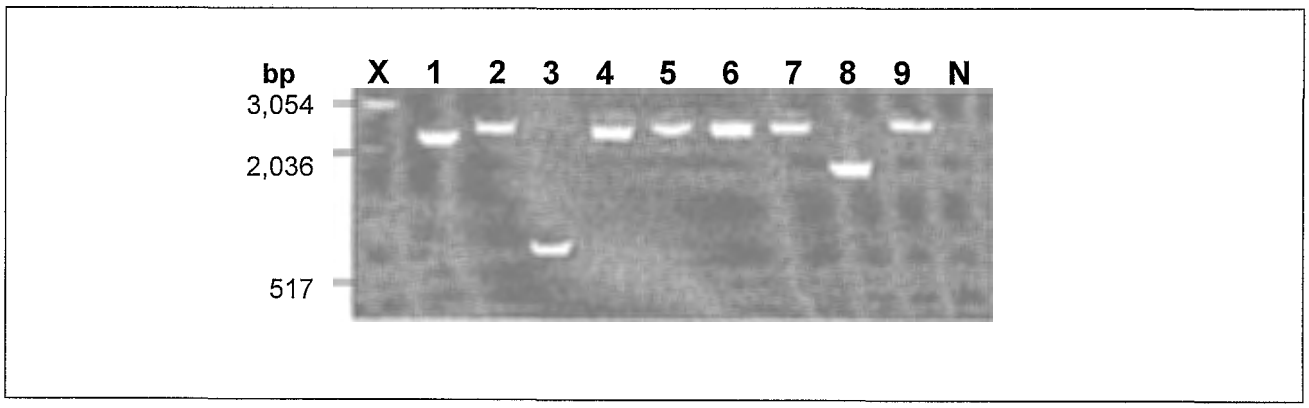
4.1 EVALUATION OF METHODS

The results generated from each of the procedures utilised in this investigation are discussed below. Difficulties that were encountered and the manner in which they were overcome are also presented.

4.1.1 PCR amplification

Whole mitochondrial genome amplification was performed utilising nine primer pairs, as described in section 3.3. Since the primer sets were optimised to amplify at a single T_m, all nine fragments could be amplified in the same thermocycler during a single PCR run. A photographic representation of fragments typically amplified throughout the study is depicted in Figure 4.1. Fragment sizes of each of the amplified products were previously listed in Table 3.1.

Figure 4.1: Photographic representation of the nine overlapping fragments, covering the full mitochondrial genome, amplified via PCR



X = molecular weight marker X; 1-9 = the amplified fragments of the mtDNA; N = negative. Fragment sizes of X are listed on the left.

Amplification of mtDNA is generally performed without much need for optimisation. The whole mitochondrial genome of an individual can be amplified from only 1 μg of DNA (Wallace, 2003). Occasionally few or no products were obtained, but usually the repetition of the experiment and a slight increase in DNA template resulted in well amplified products. A possible explanation for the unsuccessful results may be that the DNA utilised was isolated 10 years previously and was of a very low concentration, on average 1.55 $\mu g.\mu l^{-1}$. However, when no product was obtained after repeating the PCR, the 10 pmol primers were prepared again. This generally resulted in improved products. Another factor that could have contributed to unsatisfactory results is that certain primers were synthesised 11 years previously and had been thawed and frozen frequently. Resynthesis of these primers assured satisfactory results.

4.1.2 Sequence analysis

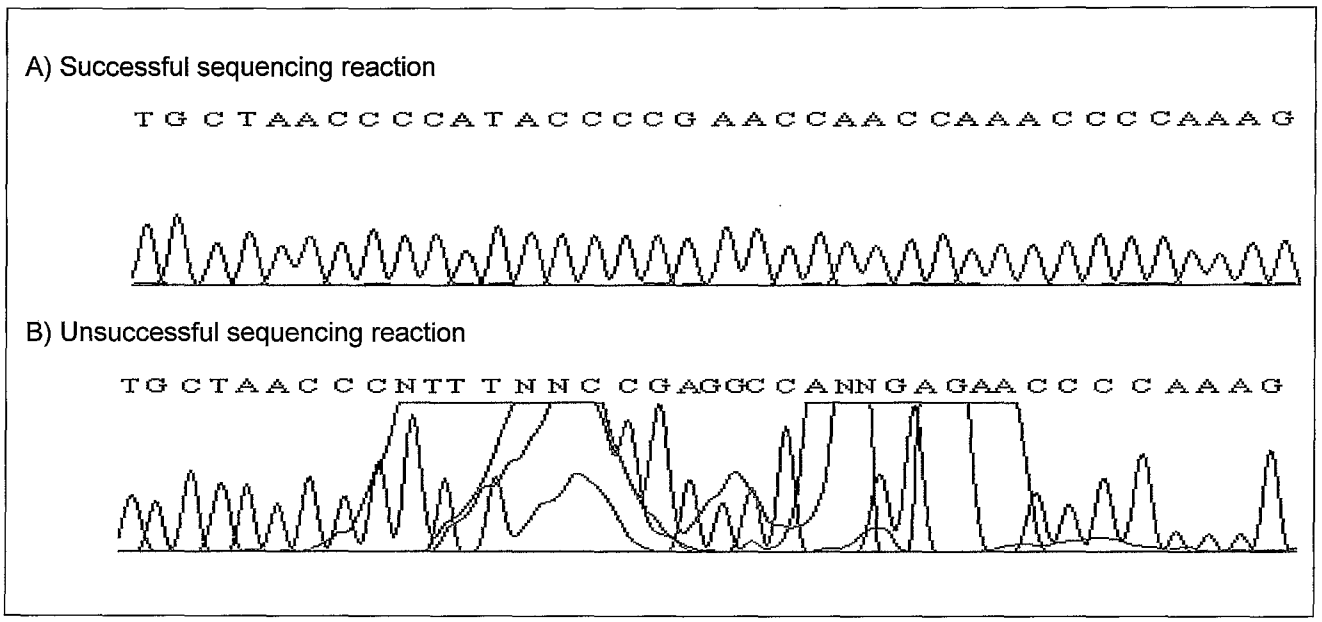
Prior to sequencing the amplified mtDNA fragments were purified utilising ExoSAP-IT. This purification protocol is easy to perform and less labour-intensive than most other procedures, since it could be programmed and performed in a PCR block. The enzymes required for purification could be added to the PCR products even when the PCR reactions were performed with the addition of mineral oil.

The study was performed utilising fluorescent sequencing, since it is safer and less labour-intensive than its radioactive counterpart. Automated cycle sequencing utilises four different fluorescently labelled dideoxy nucleotides, which can be distinguished from one another based on their respective fluorescent wavelengths. This has the advantage that sequencing reactions can be performed in one reaction tube, instead of four (Smith *et al.*, 1986).

In previous studies utilising RFLP analysis, only a small percentage of the mitochondrial genome was investigated. Other researchers performed sequence analysis of the CR. However, there are always possibilities of back and parallel mutations occurring in the hypervariable segments, included in the CR. From sequence analysis of the whole mitochondrial genome, the entire genome was investigated and provided sufficient data for comparative analyses. Thus whole mitochondrial genome sequencing will account for the highest level of molecular resolution (Vigilant *et al.*, 1991). The automated sequencing protocol worked efficiently and the detection of the sequence took ca. four hours per reaction.

In order to achieve full mitochondrial genome sequences, certain individuals had to be resequenced up to five times. The total number of sequencing reactions for 10 individuals was originally estimated to be 480, if all sequencing reactions were successful on the first attempt. Eventually 1,563 sequencing reactions were performed in order to derive full mitochondrial genome sequences for 13 individuals. The most frequent reason for resequencing was to fill gaps between sequenced fragments, generally by utilising different sequencing primers. Figure 4.2 illustrates a representative electrophorogram of a sequencing reaction that did not work well along with the resequenced final product. The primers were designed and synthesised based on the RCRS, which is representative of haplogroup H. It is therefore not surprising that their annealing to the Khoi-San DNA template, which is one of the most diverse populations, was problematic. Other parameters that could have influenced or contributed to unsatisfactory results include those mentioned for problematic PCR amplification, as discussed in section 4.1.1.

Figure 4.2: Representative electropherograms of successful and unsuccessful sequencing reactions



A = adenine; C = cytosine; G = guanine; T = thymine ; N = unknown nucleotide; green = A; blue = C; black = G; red = T; pink = N.

4.1.3 Phylogenetic and statistical analyses

All phylogenetic and statistical software was easily accessible. The analysed sequences were first converted into simplified and consistent formats, prior to importing them into the specific programs. Results from the analyses and reanalyses were obtained easily, with the longest time of analysis being four hours.

4.2 SEQUENCE ALIGNMENT AND COMPARISON

The complete consistent mtDNA genomes of 13 Khoi-San, specifically !Kung, individuals were derived and analysed to determine the nature of their variation. Comparison analysis with the RCRS revealed a total of 173 discrepancies within all individuals, as presented in Appendix A. All nucleotide substitutions are referred to as mutations since they cause a change in the nucleotide sequence of the DNA. Causative or pathogenic mutations are absent from the general population and are usually associated with disease. Polymorphic alleles are passed on to the offspring and coexist with the original gene. Substitutions are usually referred to as polymorphisms when they occur in the general population at a frequency of 0.01 (Passarge, 1995). Thus, the mutations reported in this study cannot be referred to as polymorphisms since the size of the group investigated was too small.

Of the 173 differences observed when compared to the RCRS, 99 represent previously reported polymorphisms whereas 74 substitutions have not been reported to date. A large number of differences with the RCRS was expected, since that reference is based on an individual representing haplogroup H. The !Kung individuals represent the most ancient and diverse population in the world. Thus, when comparative analyses are performed, Khoi-San sequences would be expected to differ most from other haplogroups. Future studies are therefore required in order to determine whether the unreported substitutions detected in this study represent polymorphisms or novel mutations.

The presence of the C to T transition at nucleotide 3,594 together with the A to G transition at nucleotide 10,398 implies the presence of an *Hpa* I and *Dde* I site gain at nucleotides 3,592 and 10,394 respectively. This confirms the macrolineage L ancestry of the individuals. The additional observation of the T to C transition at nucleotide 10,810, resulting in a *Hinf* I site gain, groups these individuals in haplogroup L1.

Investigation of the nucleotide changes, presented in Appendix A, led to the conclusion that individuals 38, 80 and 106 may group in the same or similar cluster, hereafter referred to as group I, as illustrated in Table 4.1. Individuals clustering in group I all share 39 nucleotide substitutions that differ from the RCRS and the other analysed individuals. In two of these, individuals 38 and 80, close relatedness was observed in that they only differ by one nucleotide substitution, namely a T to C transition in individual 80 at bp 13,572. Individual 106 was distinguishable from both 38 and 80 by the observation of an A to G

transition at nucleotide 7,673 and a G to A transition at nucleotide 8,251 in the coding region.

The remaining individuals were grouped into one main group based on the similarity of 17 nucleotide substitutions. This main group was subdivided into three smaller groups consisting of individuals 40 and 52 in one group and 94 and 130 in a second group, hereafter collectively known as group II. Group II is characterised by normal nucleotides when compared to the RCRS at nucleotides 198, 514 and 515, whereas the other individuals harboured a C to T transition, a C deletion and an A deletion at the respective sites. The other individuals, namely 10, 32, 72, 82, 102 and 122, clustered in group III, as illustrated in Table 4.1. Individuals 40 and 52, clustering in group II, were linked by ancestry on the basis of a T to C transition at bp 3,618, while individuals from group III had 11 shared substitutions.

The specific clustering of all the analysed sequences was determined by utilising phylogenetic analysis, as discussed in section 4.3. Clustering of the specific groups is also illustrated in Figure 4.6 (page 59) and Figure 4.7 (page 61).

Table 4.1 Identification of three groups into which the 13 derived Khoi-San sequences clustered

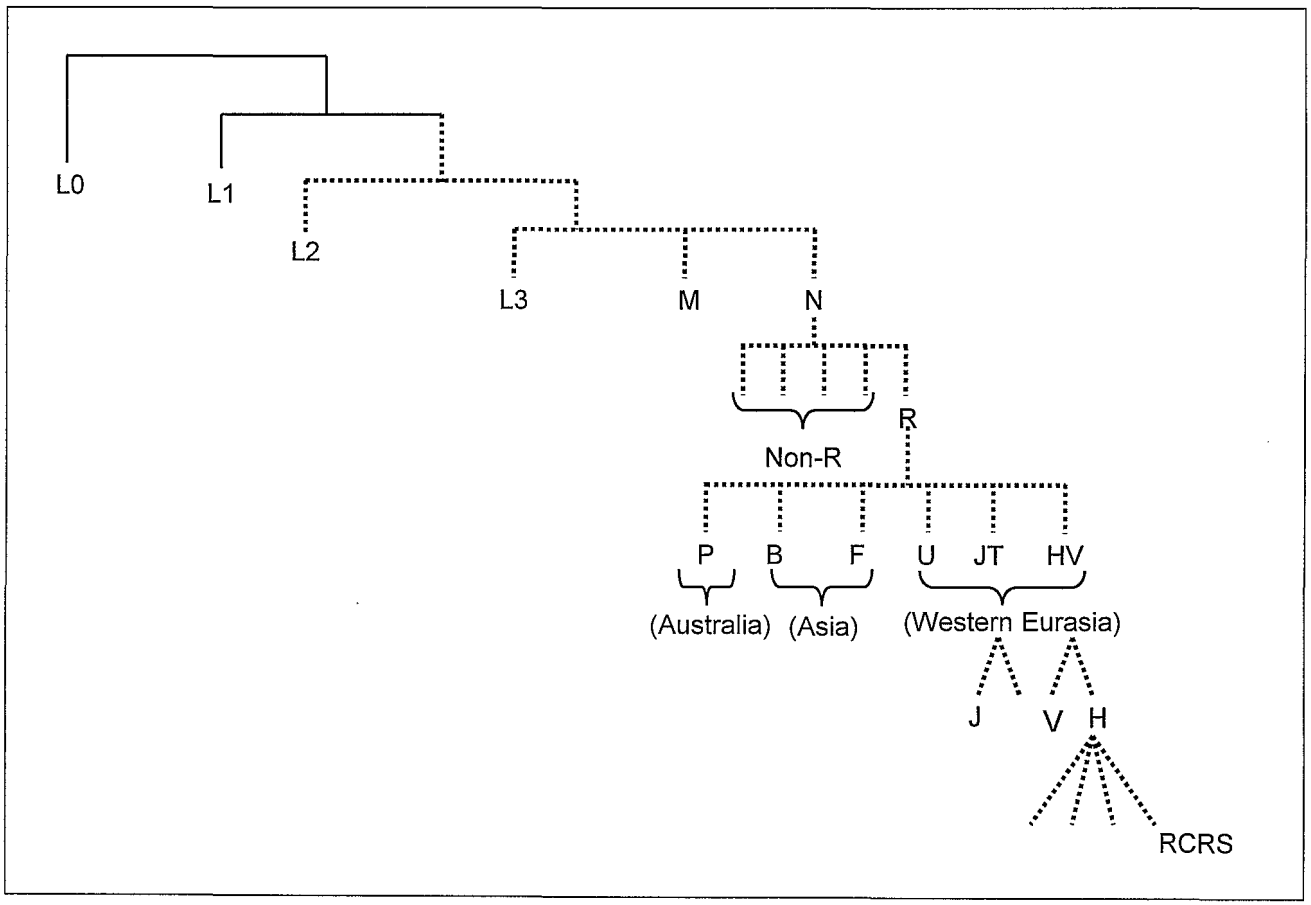
Group I	Group II	Group III
38	40	10
80	52	32
106	94	72
	130	82
		102
		122

Subsequent comparison of the derived sequences with the global mtDNA haplogroup tree agreed with the exclusion criteria for L-specific sequences, as described in Appendix B and depicted in Figure 4.3 (Ruiz-Pesini, 2003). Investigation of the nucleotide changes and phylogenetic tree, presented in Appendix B and Figure 4.3 respectively, suggested that the Khoi-San sequences generated in this study all cluster in either haplogroup L0 or L1.

In Table B.1, of Appendix B, the A to G transition at nucleotide 10,398, which excludes sequences from haplogroup N, was absent in individual 52. A possible explanation for this is homoplasy, which can be explained by two individuals having the same nucleotide base although it occurred independently and not through ancestry. Interestingly, the 1,438 and

2,706 transitions are only observed in individuals 38, 80 and 106. According to Ruiz-Pesini (2003) the 2,706 transition occurs only later in the global mtDNA haplogroup tree where it defines haplogroup H. The observation of the absence of this substitution in the 10 Khoi-San individuals might imply that it was a parallel mutation that occurred twice in the global haplogroup tree. The first might have occurred in the ancestor of individuals comprising groups II and III, whilst the second defined haplogroup H, implying that the 10 Khoi-San samples now have the same nucleotide than the RCRS. This could also be true for the 1,438 transition that is characteristic of the haplogroup that contains the sequences of the individual from whom the RCRS was generated. In the study of Tawata *et al.* (1997) it was suggested that the A to G substitution at nucleotide 1,438 was associated with type 2 diabetes in Japanese patients. However, the three Khoi-San individuals comprising group III did not have diabetes. This supports the fact that individuals' genetic background, such as their haplogroup, may have an influence on their phenotype. Similar effects have been reported in patients from different haplogroups with the same LHON mutation that displayed different phenotypes (Brown *et al.*, 2001).

Figure 4.3: Schematic representation of a global phylogenetic tree of mtDNA haplogroups



Dashed lines = branches and respective haplogroups that were excluded from the derived sequence data (Ruiz-Pesini, 2003).

4.3 CONSTRUCTION OF PHYLOGENETIC TREES

Phylogenetic trees relate organisms to one another and contain information on the order of the branches of lineages (Vigilant *et al.*, 1991). Via these trees, modern descendants can be linked to their common ancestors. Approximate times of the branching events, or divergence, can also be estimated.

All sequences utilised to construct phylogenetic trees were compared to the RCRS. Phylogenetic trees can either be rooted or unrooted. Rooted trees have a node that is identified as the root from which the other nodes descend (Page and Holmes, 1998). The root corresponds to evolutionary time. Hence, the closer a node is to the root, the older it is in evolutionary terms. Utilising rooted trees, ancestor-descendent relationships can be defined where the node closest to the root is known as the ancestor and the node further away as the descendant. Since no root exists in unrooted trees, no time correlation can be assigned to the tree, nor can the direction of evolution be determined. Thus, two adjacent sequences on an unrooted tree are not necessarily closely related in evolutionary terms.

Trees were rooted using the chimpanzee (*Pan trodlodytes*) coding region as an outgroup sequence. The outgroup method utilises the sequence of another species, such as that of the African ape, to root and place the common mitochondrial ancestor on the tree (Vigilant *et al.*, 1991). Neanderthal sequences were not utilised as an outgroup sequence since only sections of the CR sequence are available, unlike the whole mitochondrial sequence of the chimpanzee. The midpoint method for rooting an evolutionary tree is regarded as being less accurate than the outgroup method, since midpoint rooting assumes that the rate of evolution is similar in different lineages (Vigilant *et al.*, 1991).

Bootstrap analysis tests the robustness of the trees by reanalysing and redrawing the particular tree (Felsenstein, 1985). The question was asked whether the nodes in the original tree had the same cluster of sequences as the redrawn tree. Each interior branch of the original tree that is the same as the redrawn tree is assigned a value of 1, and those that are different a value of 0. The result is displayed as a percentage of the occasions when the same clusters are present among the resampled trees. A high value will lend confidence that the cluster of sequences actually do belong together. Generally, when an interior branch has a bootstrap value of 95% the topology, or branching pattern, is considered to be correct.

4.3.1 Neighbour-joining tree

Neighbour-joining trees are based on evolutionary distances. It reconstructs the evolutionary history based on the evolutionary distances between sequences, and particularly seeks the tree that results from the sum of branch lengths that is the minimum (Page and Holmes, 1998).

There are, however, some objections to distance trees, including loss of information when converted to distances and thus being unable to track the evolution of individual sites. Another problem is uninterruptible branch lengths calculated by some distance methods that resulted in 0.5 substitutions (Page and Holmes, 1998). However, a nucleotide can either be substituted or not. On a practical level, the value of 0.5 substitutions thus becomes problematic. Despite these objections, NJ trees are appealing in that they construct a single tree in a relatively short period of time.

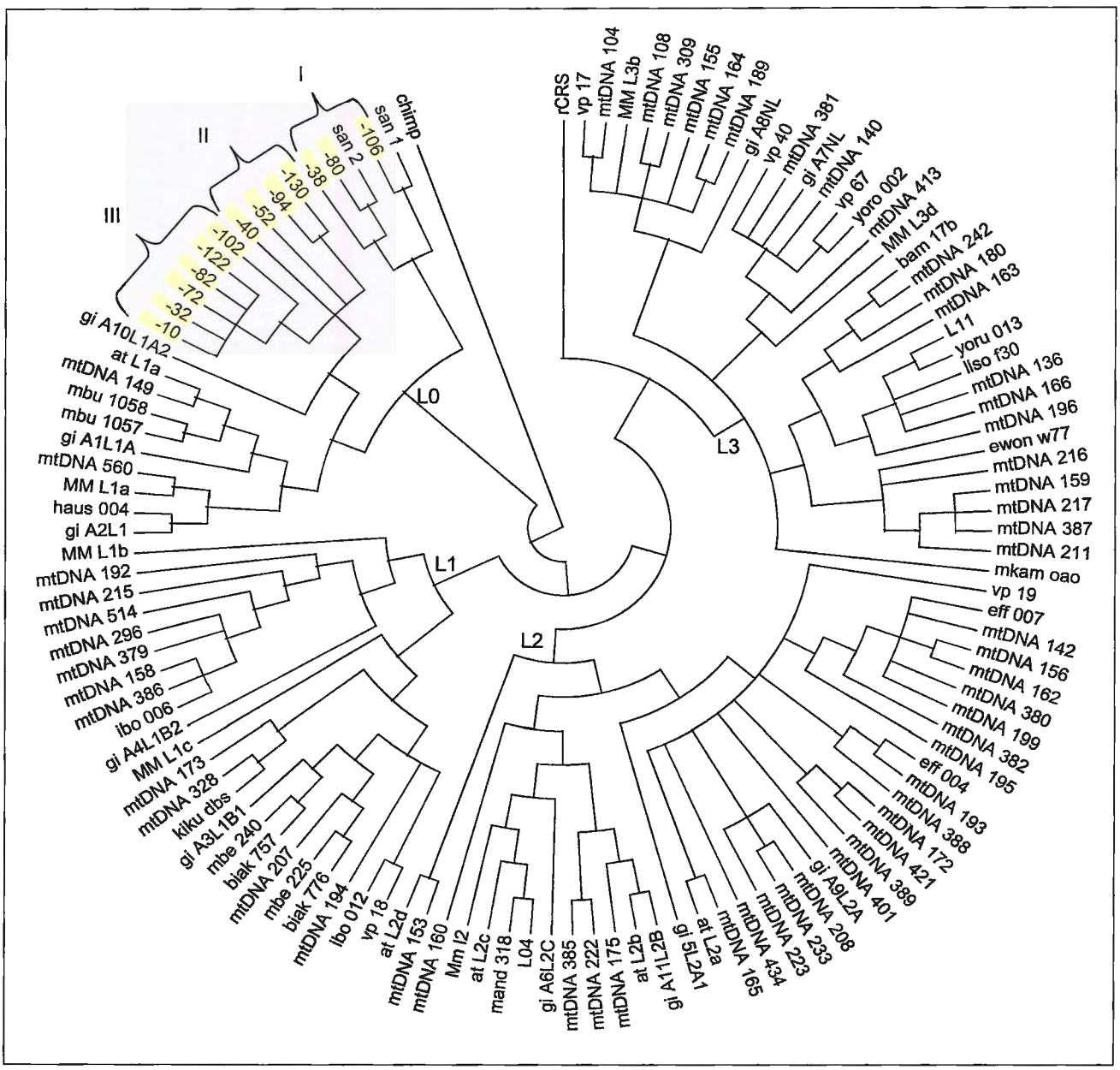
4.3.1.1 Global mtDNA Neighbour-joining tree

Permission was obtained from MAMMAG to have access to their database and include 106 L-specific sequences (Ruiz-Pesini *et al.*, 2004) in the present study. The main reason for the inclusion of the additional sequences was to derive an NJ tree in order to determine where, in the L-specific haplogroups, the derived 13 Khoi-San sequences would cluster. The NJ tree was constructed from the mitochondrial coding region of all 119 sequences and is presented in Figure 4.4. Previous studies that derived the L-specific sequences were performed utilising RFLP or sequence analysis of the CR, as opposed to the whole mitochondrial genome sequencing strategy utilised in this study.

In the L0 specific cluster, individual 106 clusters closely with the "san 1" sequence and individual 80 with "san 2". Although their sequences were very similar, closer investigation of the whole mitochondrial genomes of these sequences revealed that individual 106 differed from the "san 1" sequence at nucleotide 2,062, where the latter sequence harboured a G to A transition. The "san 2" sequence differed from individual 80 at nucleotides 2,652, where it harboured a T to C transition, and at nucleotide 8,294, where an A to G transition was observed. In Figure 4.4, which was constructed utilising 1,000 bootstrap repetitions, the clustering of the Khoi-San sequences in the L0 haplogroup is suggested. The three groups, into which the derived Khoi-San sequences clustered, as illustrated in Table 4.1, are indicated in brackets. The clustering of the Khoi-San

sequences into four similar groups, as suggested in section 4.2, is also suggested in this figure. However, in Figure 4.4 the subdivision of individual 102 from the group including individuals 10, 32, 72, 82 and 122 is illustrated.

Figure 4.4: Neighbour-joining tree constructed from the mitochondrial coding region sequences of 119 individuals, including the 13 Khoi-San sequences generated in this study



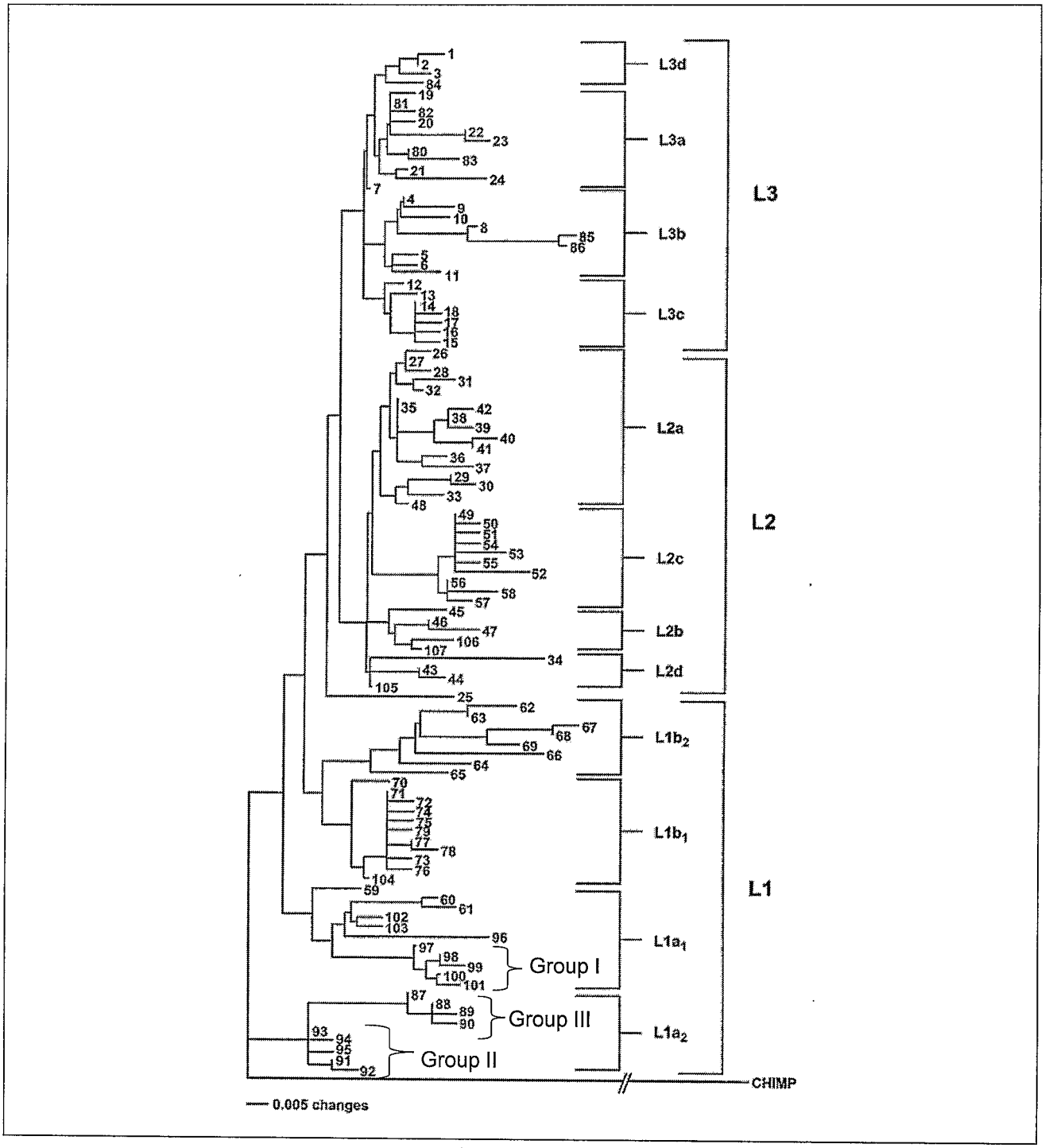
Numbers and letters indicate individuals representing the 106 L-specific sequences. The 13 Khoi-San sequences generated in this study are highlighted within yellow blocks. Specific haplogroups are indicated at the respective branches of divergence. Adapted from Ruiz-Pesini *et al.* (2004).

A concern, when utilising the bootstrap method of analysis, is the assumption that nucleotides are independent and identically distributed, thus assuming a single rate of evolution across all the sites (Page and Holmes, 1998). This implies that when analysing mtDNA, the genome would be treated as a homogeneous, single entity. In the study of

Cummings *et al.* (1995) it was observed that the independent and identically distributed assumption is not valid for mtDNA. It is, however, not known how robust bootstrap analysis is when this assumption is violated.

The derived sequences were compared to the NJ tree constructed by Chen *et al.* (2000) as illustrated in Figure 4.5. This tree was constructed utilising RFLP data for L1, L2 and L3 sequences.

Figure 4.5: Neighbour-joining tree previously constructed utilising RFLP data



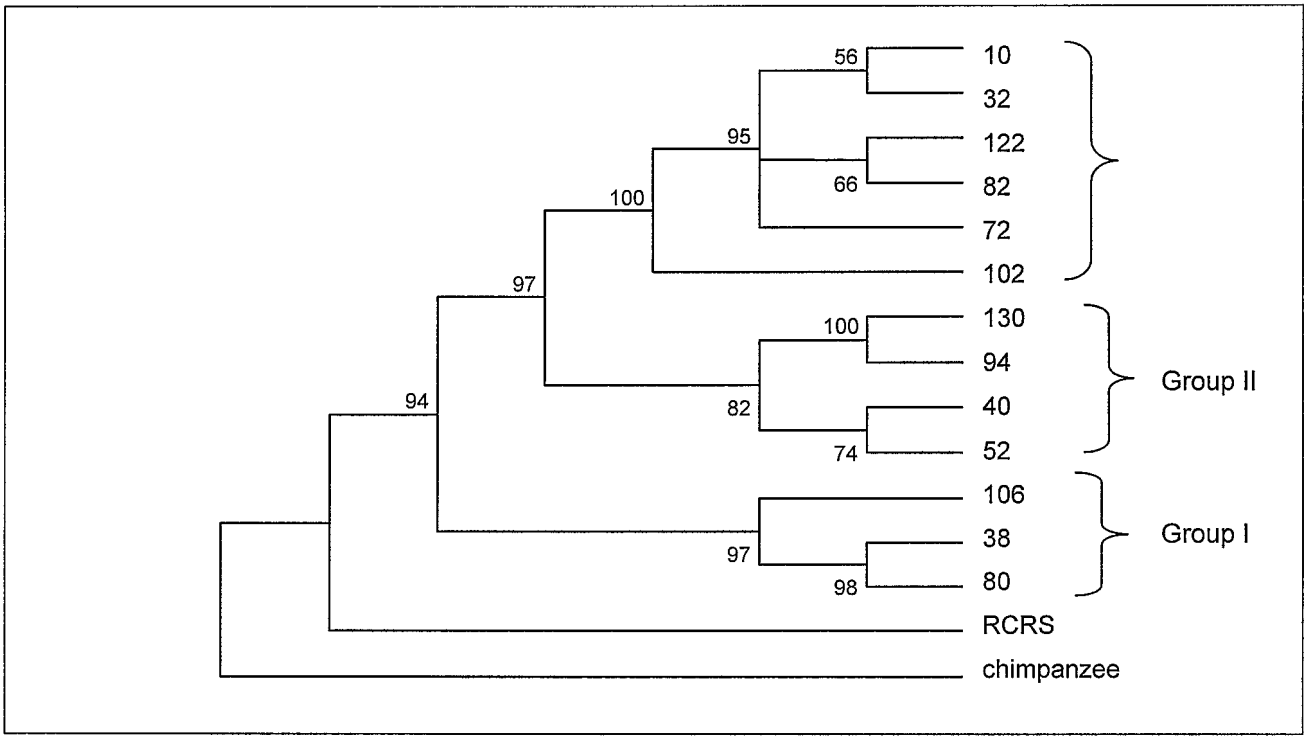
Haplogroups are indicated at the base of evolutionary branches. Groups of the derived 13 Khoi-San sequences that are similar to specific NJ tree clusters are indicated in brackets. Adapted from Chen *et al.* (2000).

Comparative analysis of the derived 13 Khoi-San sequences with the RFLP data could only identify certain groups of similar sequences. It could not be determined from the RFLP data which of the 13 Khoi-San individuals clustered with specific sequences in the previously constructed NJ tree (Chen *et al.,* 2000). Identification of these specific sequences would require comparison with complete mitochondrial genome sequences. Thus, it could only be determined that the Khoi-San individuals from group I clustered in a group representing the $L1a_1$ lineage and that individuals from group II and III clustered into two groups representing the $L1a_2$ lineage. Lineage L1a, which includes $L1a_1$ and $L1a_2$, was recently redefined as haplogroup L0 (Mishmar *et al.*, 2003).

4.3.1.2 Neighbour-joining tree of the 13 Khoi-San sequences

To define the specific clustering of the Khoi-San sequences an NJ tree was constructed with the application of 1,000 bootstrap repetitions and an outgroup sequence, as illustrated in Figure 4.6. All sequences were compared to the RCRS.

Figure 4.6: Neighbour-joining tree of the 13 generated Khoi-San sequences



The numbers on the far right represent the !Kung individuals that were analysed. Numbers at the nodes = bootstrap values. Bootstrap values for the clusters were calculated by 1,000 replications. Chimpanzee = *Pan trodlodytes,* utilised as an outgroup sequence. RCRS = revised Cambridge Reference Sequence.

The clustering of the 13 Khoi-San sequences are further defined in Figure 4.6 and supports the division into three main groups, as suggested previously in sections 4.2 (page 52) and 4.3.1.1 (page 55). Group I has a bootstrap value of 97% and group II a

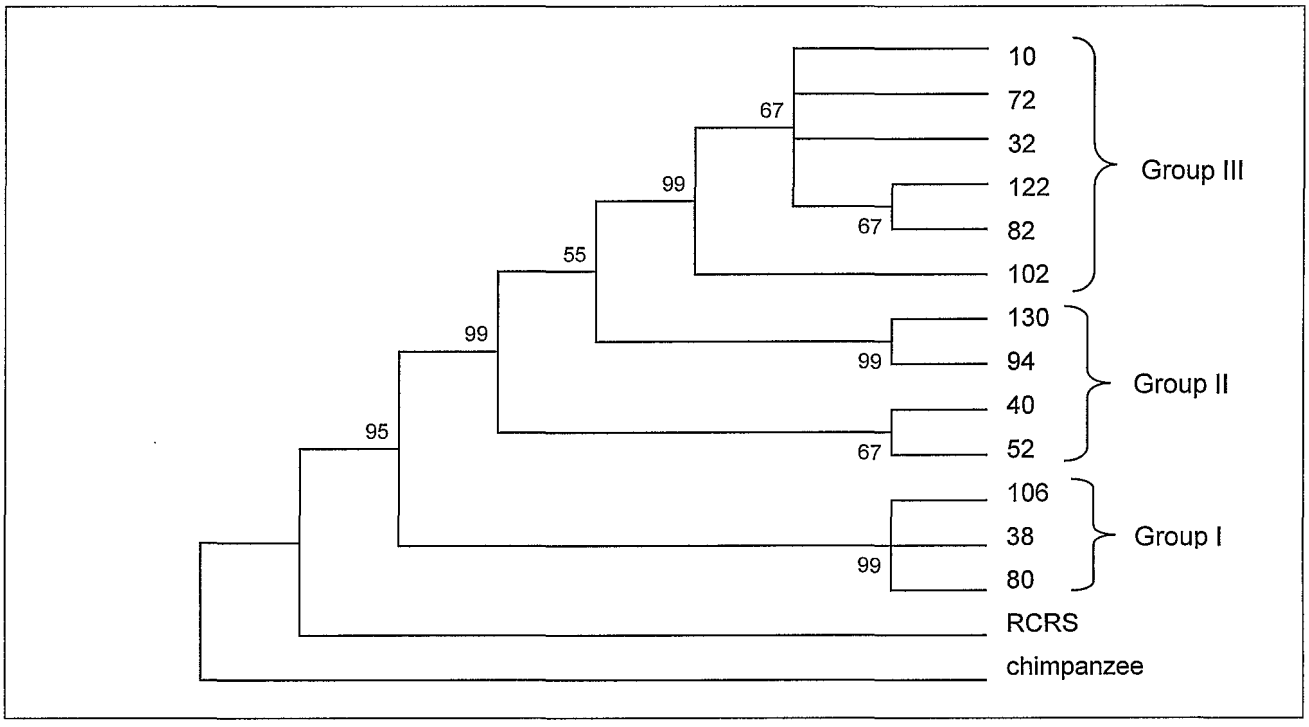
bootstrap value of 82%. A third group had a 100% bootstrap value and separated 102 from 10, 32, 72, 82 and 122, with the latter group having a bootstrap value of 95%. Group I could be further subdivided into two groups where numbers 38 and 80 group together, with a bootstrap value of 98%. Group II subdivided into two groups consisting of numbers 52 and 40, with a 74% bootstrap value, and a second group including numbers 94 and 130, which has a bootstrap value of 100%. In group III numbers 122 and 82 seem to cluster together, with a 66% bootstrap value, as well as 10 and 32, with a bootstrap value of 56%.

4.3.2 Maximum parsimony tree

An MP tree was constructed, to allow further definition of the mitochondrial sequences, as depicted in Figure 4.7. Maximum parsimony analyses the specific sequences, or functions derived from it, and is thus a discrete method of phylogenetic tree drawing. This is in contrast to the distance method of NJ, which operates on pairwise distances (Page and Holmes, 1998).

Utilising MP trees avoids the possible loss of information when sequences are converted to distances. Essentially maximum parsimony chooses the tree derived from the fewest evolutionary changes. It is often time-consuming to find the MP tree for more than 15 sequences. However, the main objection to parsimony methods is that under certain conditions the wrong tree can be constructed. An example of this occurrence is variable evolution rates between sequences or quite diverse sequences. This results in long branches, where the same substitution occurs in two different branches, but which are joined by the tree building method (Page and Holmes, 1998). Inclusion of the chimpanzee and RCRS with the 13 Khoi-San sequences was still acceptable to derive an MP tree without it being too time-consuming. The constructed tree is depicted in Figure 4.7.

Figure 4.7: Maximum parsimony tree of the 13 Khoi-San sequences



The numbers on the far right represent the !Kung individuals that were analysed. Numbers at the nodes = bootstrap values. Bootstrap values for the clusters were calculated by 1,000 replications. Chimpanzee = *Pan trodlodytes,* utilised as an outgroup sequence. RCRS = revised Cambridge Reference Sequence.

The constructed MP tree resulted in a virtually identical tree structure to that of the NJ tree (Figure 4.6) with the only exceptions being the topologies of individuals 72 and 106. In the MP tree, individuals 10 and 32 do not have a common ancestor as in the NJ tree, but share the same ancestor with individual 72. In group I of the NJ tree the definite split between individual 106 and the cluster including 38 and 80 can no longer be observed in the MP tree. Hence, these three individuals seem to share a common ancestor based on the MP tree. A possible explanation for these discrepancies is that the NJ tree is based on the evolutionary distances between the sequences compared to the MP tree, which is based on the specific sequences. Comparative analyses of Appendix A and Figure 4.7 suggest that individual 106 differs from the other two members of group I by two nucleotide substitutions, one at nucleotide 7,673 and the other at 8,251. Individual 72 harbours a transition at nucleotide 8,829, which distinguishes it from other members of group III. Based upon Figure 4.7 individuals 32 and 72 should be closely related. Thus, the choice between the NJ and the MP tree analysis depends upon the question that is asked and the purpose of the analysis, for example evolutionary distances compared to parsimony analysis. The bootstrap values for the topologies of individuals 94 and 130 as well as for and 38, 80 and 106 can be assumed to be correct, since they all have bootstrap values of 95% or more.

When Figures 4.6 and 4.7 are compared, the topologies of groups I and III, as well as the subgroup including individuals 94 and 130, have high bootstrap values, which suggest that the branching is correct. The observation of the similar trees suggests, together with the observation of high bootstrap values for the higher branches, that the phylogenies shown in Figure 4.6 and Figure 4.7 are reasonable representations of the phylogenetic relationships between the derived sequences.

4.3.3 Statistical analysis

In order to determine if the derived sequences deviated from the standard neutral model of evolution, Tajima's *D* and Fu and Li *D** tests were performed. The neutral theory of molecular evolution (Kimura, 1968) proposes that high levels of genetic variation may be explained by mutations in the population that have no effect on the populations' fitness (Page and Holmes, 1998). These neutral mutations are usually lost, or occasionally fixed, through the process of genetic drift. This implies that quickly evolving genes will have more variation within and between species than more slowly evolving genes. The frequency of the distribution of pairwise mtDNA sequence differences was calculated to test for rapid population expansion.

4.3.3.1 Tajima's *D* and Fu and Li *D** tests

The Tajima's *D,* as well as Fu and Li's *D** statistics test the levels of variation within and between species, under the assumption of neutrality and absence of recombination, utilising sequence data from a single species (Tajima, 1989). This author proposed a test for the assumption of neutrality of mutations by considering *S* and *k* in a sample size of n sequences from a population. The number of segregating sites (*S*) is affected by the presence of deleterious alleles, which exist in low frequencies. However, *S* does not take the frequency of nucleotide substitutions into account. On the contrary, *k* considers the frequency of the substitutions and is, therefore, not much affected by the existence of deleterious alleles. Hence if some of the sequences from the sample population are subjected to selection, the estimate of theta (θ), which is equal to $2N\mu$, based on *S*, will be different from the estimate based on *k*. 2N denotes the effective population size for haploid individuals, and thus for mtDNA, and μ the mutation rate per DNA sequence per generation. The difference (*D*) between *S* and *k* is, therefore, suggested to detect selection. When evolution is neutral the amount of genetic variation should be equal to the number of times that the substitutions occur, thus $S = k$, which implies that $D = 0$.

Fu and Li (1993) further investigated the numbers of mutations in the internal and terminal branches of a phylogenetic tree. The internal topologies are seen as being the older part of the phylogenetic tree, and the terminal branches as the younger part. An excess of mutations is expected in the terminal branches in the presence of purifying or negative selection or when advantageous mutations have recently become fixed. Balanced selection may, however, cause a deficiency of mutations in the terminal branches since some alleles might be old. Therefore, the main difference between Fu and Li's D^* and Tajima's D statistics (Equations 3.3 and 3.4) is that the former computes D^* utilising the total number of changes, eta (η), whereas Tajima utilises the total number of segregating sites, S. Note that Fu and Li utilizes Πn to indicate k (Fu and Li, 1993).

A positive D value suggests either a low occurrence of single mutations or high frequency of substitutions and may be explained by balanced selection. In contrast, a negative value suggests the accumulation of a high number of mutations, which may suggest population expansion or selection, be it positive or negative.

Thus, in essence, both Tajima's D and Fu and Li's D* statistical tests estimate the number of single nucleotide changes and compare it to the frequency with which the changes occur. Although these two tests utilises different algorithms, they both determine the number of singletons, mutations that appear only once among the sequences.

4.3.3.1.1 Statistical significance

The statistical significance of the neutrality tests were determined by utilising the probability or P value. This determines the probability that the test result could have been obtained randomly, and would thus be false. A P value of 0.05 implies that there is a 5% probability that the result could have been generated by chance. A significant P value suggests that the null hypothesis, in this case neutral evolution, is rejected. The D statistic results, together with the P values, were computed utilising the DnaSP version 3.51 software (Rozas and Rozas, 1999). These results are illustrated in Table 4.2.

Table 4.2: Statistical and significance tests for haplogroups L0 – L3

Haplogroups	Tajima's *D*	Fu and Li's *D**	Significance		
			Tajima's *D*	Fu and Li's *D**	Result
L0	-0.484	-1.057	$P > 0.10$	$P > 0.10$	Not significant
L1	-0.942	-1.58	$P > 0.10$	$P > 0.10$	Not significant
L2	-1.915	-2.951	$P < 0.05$	$P < 0.05$	Significant
L3	-1.916	-3.328	$P < 0.05$	$P < 0.02$	Significant

L0 – L3 = the respective haplogroups; *P* = probability.

Analysis of Table 4.2 suggests that the genetic variation of haplogroups L0 – L3 might be explained by either population expansion, or positive or negative selection. This can be observed from the negative values obtained for all four haplogroups.

The *D* statistic results for L0 and L1 did not differ significantly from the neutral model. Values obtained for L2 and L3 were significantly different from neutrality, implying that the neutral model for evolution is rejected. In order to determine whether the negative values obtained suggest population expansion or selection, the pairwise differences were compared for each of the L-specific haplogroups, as described in section 4.3.2.2.

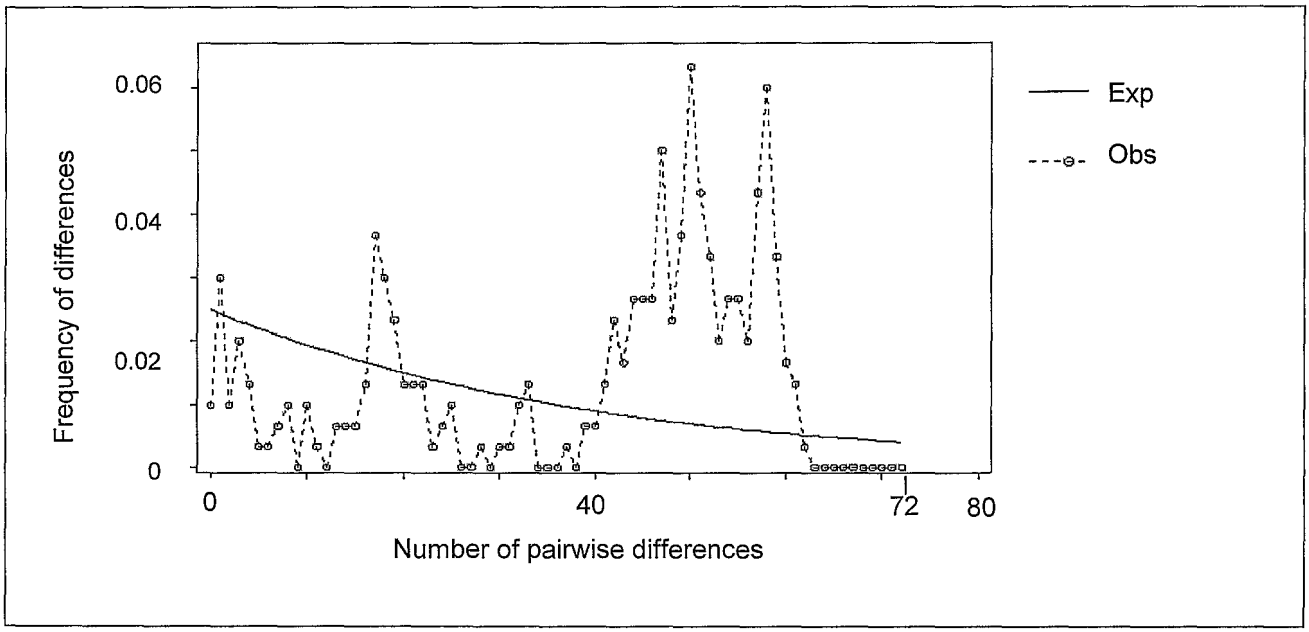
4.3.3.2 Pairwise comparisons

In order to test for rapid population expansion, the frequency distribution of pairwise sequence differences were calculated for all the 106 L-specific sequences together with those the of the 13 Khoi-San sequences, utilising the DnaSP version 3.51 software (Rozas and Rozas, 1999). The resulting graph for each respective L-haplogroup is presented in Graphs 4.1 to 4.4.

The pairwise number of difference test compares sequences and measures the number of nucleotide differences. The expected values for a population of constant size, under the assumption that recombination is absent, were graphically illustrated against the observed values, as illustrated in Graphs 4.1 to 4.4. Expected values were estimated utilising Equation 3.5. According to Slatkin and Hudson (1991) plotting the pairwise frequency distributions of sequences provides an indication of the history of the sequences and the resulting phylogenetic tree.

Graph 4.1 was constructed from the 13 derived Khoi-San sequences together with 12 L0-specific sequences, which formed part of the group of 106 L-specific sequences obtained from MAMMAG. The solid line in the graph represents the expected values of the pairwise number of differences for a population of constant size, in the absence of recombination. A ragged distribution (indicated by the dashed line) was observed for the L0-specific sequences, which implies that these sequences were not subjected to a rapid population expansion (Page and Holmes, 1998).

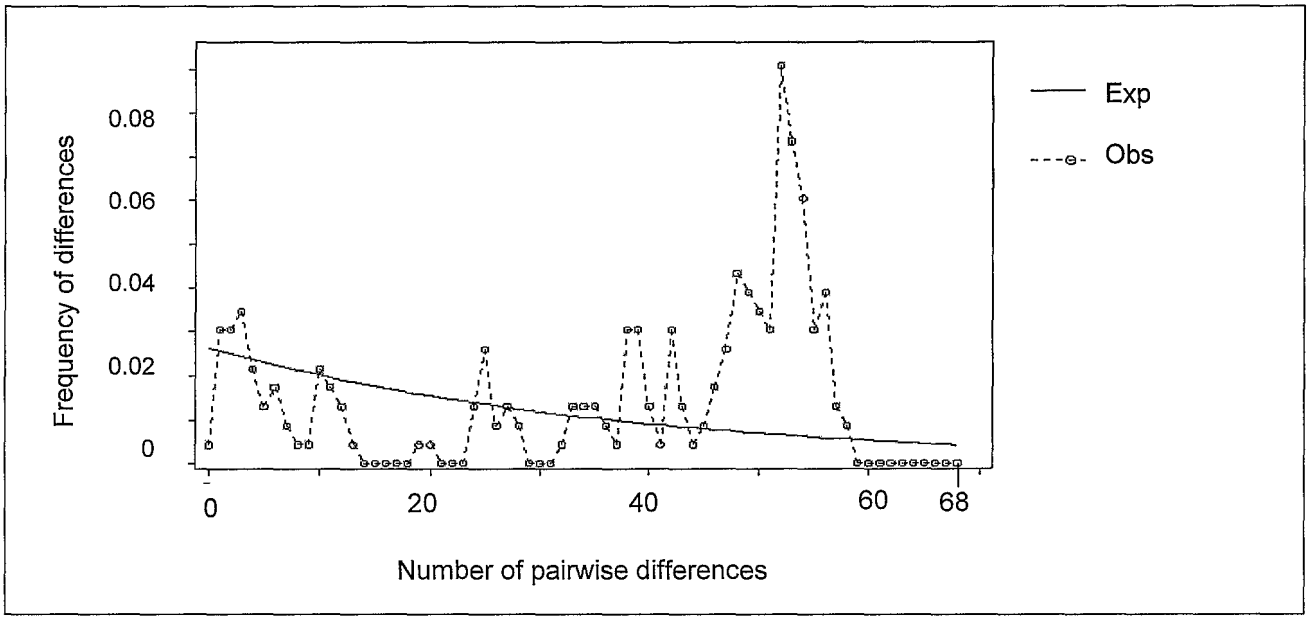
Graph 4.1: Graphical representation of pairwise comparisons for haplogroup L0



Exp = expected values for a constant population size with the absence of recombination; Obs = observed pairwise nucleotide differences.

Similar to the results from Graph 4.1, a ragged distribution was observed for the sequences in Graph 4.2, compared to the solid line representing a constant population size. Graph 4.2 was constructed from the 23 L1-specific sequences, included in the 106 L-specific sequences.

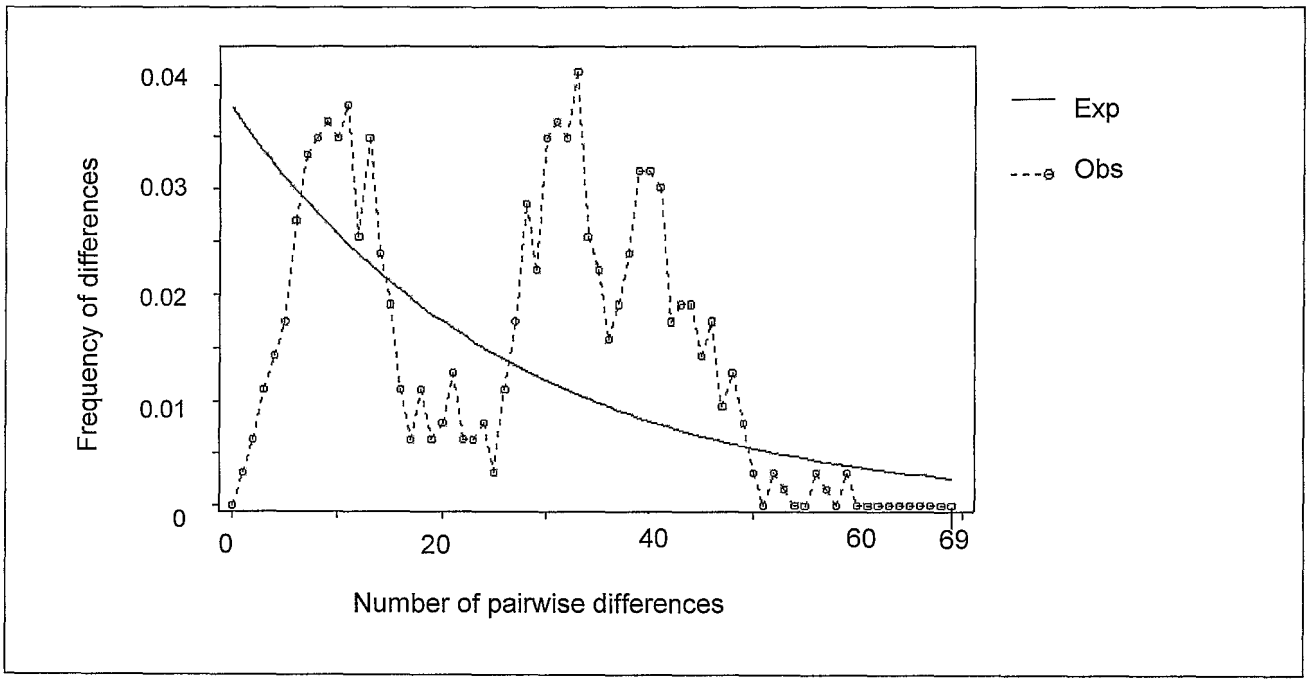
Graph 4.2: Graphical representation of pairwise comparisons for haplogroup L1



Exp = expected values for a constant population size with the absence of recombination; Obs = observed pairwise nucleotide differences.

Graph 4.3 was derived from the 37 L2-specific sequences, which were included in the group of 106 L-specific sequences. According to the results presented in Graph 4.3 it can be deduced that these sequences were not subjected to a rapid population expansion.

Graph 4.3: Graphical representation of pairwise comparisons for haplogroup L2

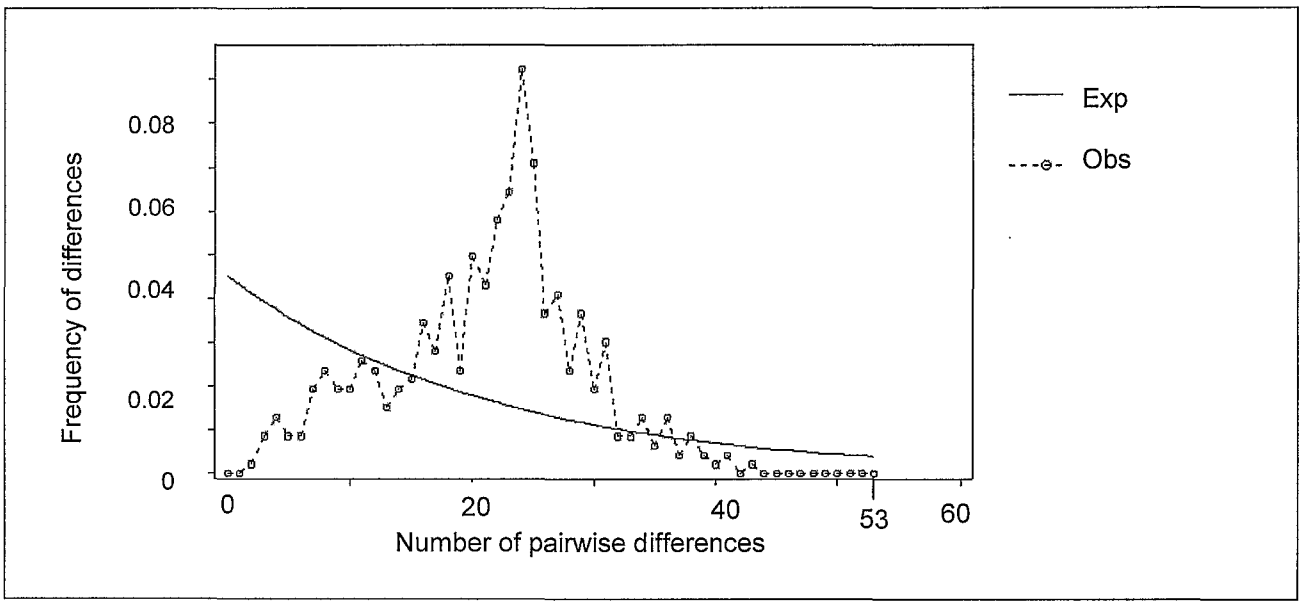


Exp = expected values for a constant population size with the absence of recombination; Obs = observed pairwise nucleotide differences.

When comparing Graphs 4.1, 4.2 and 4.3 broad and ragged distributions were observed, which suggested the absence of a rapid population expansion for haplogroups L0, L1 and L2. From the observation of the ragged distributions, it was also suggested that the sequences that were utilised may represent populations that have been relatively stable for a long period. Although a ragged nucleotide difference distribution does not necessarily exclude a population expansion, it was certainly not a rapid event.

In contrast, the distribution of the observed pairwise sequences of haplogroup L3 was bell-shaped when compared to the smooth distribution expected for a constant population size, as illustrated in Graph 4.4. This bell-shaped graph is characteristic of a population expansion event. This expansion accounted for distortions of the distributions of mtDNA, since numerous nucleotide alterations occurred as a result of this event when people were subjected to new environments and adaptations. Although the L3 haplogroup is not very well defined, since it consists of all that are not L0, L1 or L2, results from Graph 4.4 support the out-of-Africa theory.

Graph 4.4: Graphical representation of pairwise comparisons for haplogroup L3



Exp = expected values for a constant population size with the absence of recombination; Obs = observed pairwise nucleotide differences.

Thus, from the *D* statistical tests and the pairwise comparisons, it is suggested that haplogroups L0 – L3 deviated from the null hypothesis of neutral evolution and were not subjected to balanced selection. Pairwise comparisons suggested that haplogroups L0 – L2 did not undergo a rapid population expansion, but had maintained a relatively stable population size over a long time. Haplogroup L3, however, was subjected to a population expansion.

4.4 L0-SPECIFIC HAPLOGROUP TREE

The derived Khoi-San sequences were compared to an L0-specific haplogroup tree, as illustrated in Appendix C, and used here with permission from MAMMAG (Ruiz-Pesini *et al.*, 2004). Closer investigation and comparison with this L0-specific tree revealed that the derived sequences could all be assigned to haplogroup L0, as depicted in Figure 4.8.

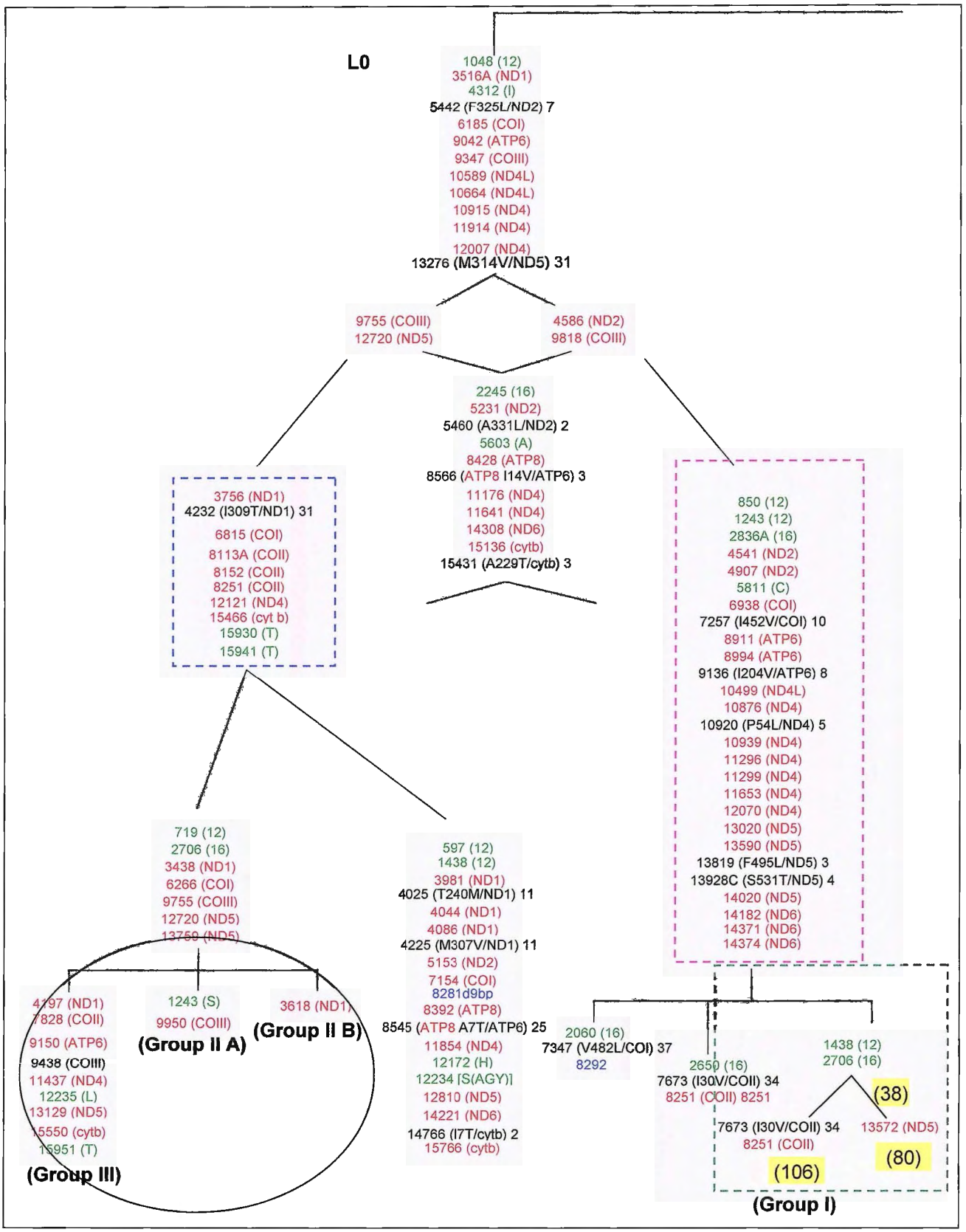
All synonymous substitutions are transitions, except where an amino acid abbreviation is indicated, for example 8113A, representing a transversion at nucleotide 8113 to an alanine. Conservation numbers are indicated after each non-synonymous substitution. The region where the substitution occurred is shown in brackets. Single letter abbreviations of the amino acids that are involved in the non-synonymous changes are also indicated in brackets.

Nucleotide substitutions in Appendix C and Figure 4.8 appear in numerical order and do not indicate the order in which the specific substitutions occurred through time. Appendix C and both Figures 4.8 and 4.9 were constructed from the mitochondrial DNA coding regions (nucleotides 577 – 16,023) of each of the 13 Khoi-San sequences as well as the 12 L0-specific individuals, from which the tree was constructed, to avoid the possibility of reverse mutations that might occur in the CR.

All 13 Khoi-San sequences grouped into two main branches. One consisted of individuals from group I, indicated in the dark green box in Figure 4.8, and the other consisted of the individuals from groups II and III, illustrated as groups II A, II B and III in Figure 4.9. These clusters were similar to results obtained from sequencing and phylogenetic analyses, as discussed in sections 4.2 (page 52) and 4.3 (page 54). The 13 Khoi-San individuals from which the sequences were derived are highlighted in yellow.

The original tree to which the Khoi-San sequences were compared was derived from 12 L0-specific sequences, which were included in the 106 L-specific sequences, utilised with permission from MAMMAG. The topology comprising the right branch of the original tree was derived from two individuals, the centre branch from nine individuals and the left topology was derived from one individual as illustrated in Appendix C.

Figure 4.8: Schematic representation of an L0-specific tree with the addition of the 13 Khoi-San sequences



Red = synonymous substitutions; black = non-synonymous substitutions; blue = substitutions in the non-coding region of the coding region; green = substitutions in RNA genes. Numbers at the bottom of specific branches represent the individuals that cluster in that branch. The Khoi-San sequences generated in this study are highlighted in yellow blocks. Topologies encircled are presented in more detail in Figure 4.9 Adapted from Ruiz-Pesini (2003).

The three !Kung individuals that grouped in the topology on the right (Figure 4.8, group I) harboured all 27 nucleotide substitutions, as indicated in the purple box in Figure 4.8, and were only added to the terminal branches of the original topology. These exact 27 substitutions were also observed in the original tree where they clustered to comprise the same internal branch as that observed in Figure 4.8. Thus, even with the addition of the three Khoi-San sequences to the original tree, no branching was revealed in that internal branch, indicated in the purple box in Figure 4.8. The sequence from individual 38 was added to the nodal or internal branch, while that of individuals 80 and 106 resulted in the creation of new terminal branches.

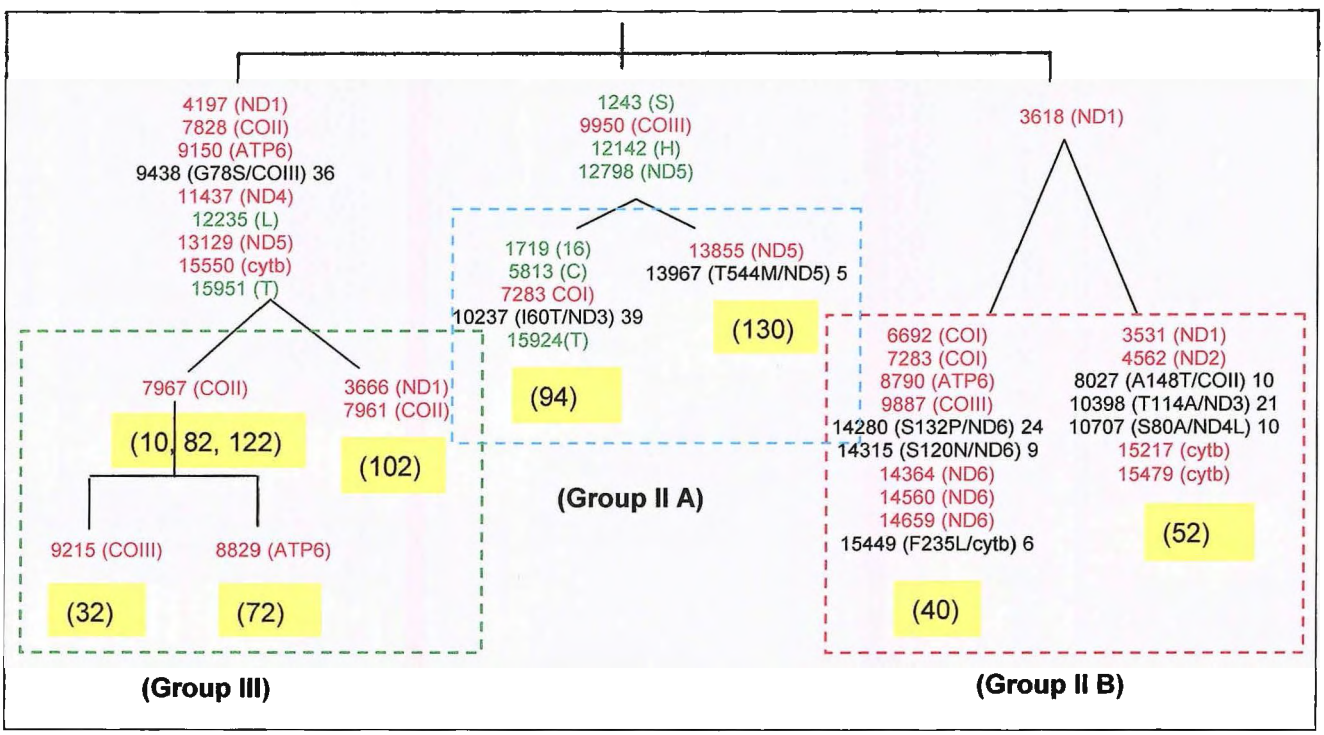
The addition of the Khoi-San sequences to the topology on the left in Figure 4.8 enriched the branch and revealed further delineation of the branching structure. This resulted in the clustering of the group of nucleotide substitutions, indicated in the blue box, into an internal branch of that topology. However, it previously comprised a terminal branch in the original tree, as illustrated in Appendix C.

The 10 individuals that were added to that specific branch grouped into three main groups, similar to results presented in sections 4.2 (page 52) and 4.3 (page 54). As observed from sequence and phylogenetic results, individuals 94 and 130 clustered together in one group, indicated as group IIA, whilst 40 and 52 clustered in group IIB in Figure 4.9. The substitution observed in individual 94 in the $tRNA^{Thr}$ gene at nucleotide 15,924 has been associated with a lethal respiratory chain defect in infants (Yoon *et al.*, 1991). This mutation, which occurred in the last base pair of the anticodon stem, contributed to the notion that mutations affecting the anticodon loop structure might cause lethal mitochondrial diseases in infants. However, this mutation has also been demonstrated to be polymorphic (Brown *et al.*, 1992). Since individual 94 had no obvious myopathy symptoms and because of this mutation's polymorphic nature, it was suggested that this substitution at nucleotide 15,924 is not a primary cause of lethal infantile myopathy.

From Figure 4.9 it was observed that individual 102 formed a distinct subgroup in group III. Figure 4.9 represents further delineation of the encircled topologies in Figure 4.8 and it was suggested that individuals 10, 82 and 122 clustered together and that individuals 32 and 72 clustered separately from this group, which was accounted for by a single nucleotide. The clustering of individuals in group III in Figure 4.9 is different from that illustrated in Figures 4.6 and 4.7, specifically concerning the clustering of individuals 82, 122 and 10. These individuals were grouped together in Figure 4.9, whilst individual 10

was grouped separately in Figures 4.6 and 4.7. A possible explanation for this observation is that Figures 4.6 and 4.7 were derived utilising the whole mitochondrial sequence of each individual. In contrast only the coding region was used to generate the results presented in Figure 4.11. Thus, substitutions in the CR could account for further definition and subdivision of the individuals. However, these substitutions may represent reverse or parallel mutations. Hence, in order to avoid any misinterpretations the L0 tree derived from the coding regions of each individual should be considered the most informative. This is due to the fact that data from the variable CR may result in errors if parallel or reverse mutations were interpreted as randomly occurring mutations. This may result in individuals who are not closely related clustering together or, alternatively, closely related individuals clustering in different topologies.

Figure 4.9: Schematic representation of the branching order of the terminal branches of the topology on the left in Figure 4.8



Red = synonymous substitutions; black = non-synonymous substitutions; blue = substitutions in the non-coding region of the coding region; green = substitutions in RNA genes. Numbers at the bottom of specific branches represent the individuals that cluster in that branch.

The two different branching orders may represent two different histories. From the topology in the purple box (Figure 4.8) it may be suggested that those three individuals accumulated a great number of mutations before any branching occurred, or that any branching that might have occurred did not survive. In the branch on the left, fewer mutations accumulated over time before branching occurred. Alternatively, delineation events that occurred in the past did survive to the present. However, the topology on the

right was constructed from a total of five individuals, compared to the 11 included in the left branch. An increase in the sample population may result in different branching structures and may reveal further delineation on the right.

4.5 COALESCENT DATES

The coalescence time, or time when the sequences shared a common ancestor, was estimated by utilising Equations 3.6 and 3.7. This included counting the number of nucleotide substitutions in the branches of a phylogenetic tree up to the point of divergence from the common ancestor. This number was then divided by the number of individuals from whom the tree was derived. The resulting ratio was converted to time as described in section 3.8. Standard deviations were calculated utilising Equation 3.8.

The MRCA for the L0 haplogroup tree was estimated at 133,585 ± 5,240 YBP, and the MRCA for the terminal branches of the topology on the left was estimated to be 47,269 ± 4,928 YBP. The MRCA of the right branch was calculated to be 11,560 ± 3,853 YBP. The coalescence date of the newly derived nodal branch of the topology on the left was estimated at 76,134 ± 5,963 YBP. The calculated L0 coalescence dates are similar to the estimate of 142,000 ± 17,000 YBP that was calculated for L0-specific sequences by Mishmar *et al.* (2003).

4.6 ANALYSIS OF NON-SYNONYMOUS AND SYNONYMOUS SUBSTITUTIONS

The effect of selection on the sequences was investigated by determining the ratio of non-synonymous to synonymous substitutions for internal and terminal branches. Neutral mutations, such as silent or synonymous mutations, were expected to occur randomly throughout the mitochondrial genome. The number of non-synonymous mutations can be normalised for time by division with the number of synonymous mutations. The aim of this study was not to investigate intergenic alteration between different haplogroups, therefore this avenue will not be discussed further. Determination of the ratio of non-synonymous to synonymous mutations, for internal as well as terminal branches, was sufficient to form an idea about the selection of the tree topologies.

It is expected that non-synonymous mutations, causing deleterious replacements of amino acids, would be excluded by purifying or negative selection from the population over time.

It is thus expected that non-synonymous mutations should be more frequent in terminal branches than internal branches. However, non-synonymous advantageous mutations obtained, for example through adaptation, would be expected to be prevalent in the internal branches compared to terminal branches (Ruiz-Pesini *et al.*, 2004). Hence the ratio of k*a* to k*s* of internal versus (vs) terminal branches would be a measure of the adaptive vs purifying selection in a population (Ruiz-Pesini *et al.*, 2004).

Low values for internal vs terminal branches would suggest purifying selection, while high values would suggest adaptive selection. The replacement mutation frequency (RF) for internal and terminal branches was determined for haplogroups L0 to L3, as presented in Table 4.3.

A simplified test of positive selection was performed utilising one-tailed Fisher's Exact Test. The null hypothesis of strictly neutral and purifying selection was rejected when the *P*-value was less than 0.05.

Table 4.3: Selective parameters for the L-haplogroups

Haplogroup	N	Internal branches		Terminal branches		*P*	I/T
		RF_I	NS/S	RF_T	NS/S	$RF_I - RF_T$	
L0	25	0.24	13/54	0.53	23/43	0.0348	0.45
L1	23	0.32	18/57	0.49	27/55	0.1448	0.65
L2	37	0.34	16/47	0.46	41/90	0.2509	0.74
L3	34	0.30	11/37	0.3	26/77	0.4628	0.88

N = population size; RF_I = replacement mutation frequency within internal branches; RF_T = replacement mutation frequency within terminal branches; *P* = probability values; I/T = RF_I / RF_T; NS/S = non-synonymous / synonymous mutations.

As expected, less k*a* was observed in the internal branches compared to the terminal branches. A significant difference between the ratios of k*a* vs k*s* was observed in the internal compared to the terminal branches of L0. Hence, the null hypothesis of neutral evolution was rejected for L0 ($P < 0.05$). Although differences in the actual numbers were observed, non-significant values ($P > 0.05$) were obtained for haplogroups L1 to L3.

Investigation of Table 4.3 suggested an increase in the values of I/T from L0 to L3. This may have suggested the occurrence of positive selection from L0 to the populations that migrated from Africa to other continents. From the lower I/T value that was observed for L0, it may be suggested that negative selection might have played a role to configure African lineages from L0 to L3.

4.6.1 Conservation index

The influence of purifying compared to adaptive selection was further analysed by determining the conservation of each non-synonymous mutation. Adaptive or deleterious mutations would often alter amino acids that were functionally important, while neutral mutations would frequently change functionally unimportant amino acids.

The conservation of each k*a* was determined by comparison to the 39 different species listed in Table 3.3. The k*a* for each of the derived sequences was compared to the wild type, which is the most abundant amino acid in the specific position of interest. An amino acid change with a high conservation index implies that the majority of the species to which it was compared harboured the wild type, and thus a different amino acid to that of the specific mutant amino acid under investigation. This suggests that the specific amino acid of interest could have an important function, since most species have the same amino acid.

Figures 4.8 and 4.9 illustrate six highly conserved non-synonymous substitutions, of which three are located in terminal branches. It should be noted that only five of these conserved substitutions were observed in the 13 derived sequences. The substitution at nucleotide 7,347, located in a terminal branch of the topology on the right (Figure 4.8), was observed in an individual that was included in the original tree that is presented in Appendix C.

In individual 106 a conservative non-synonymous transition was observed where a nonpolar isoleucine was changed to nonpolar valine at nucleotide 7,673, which was conserved in 34 species. The second highly conserved transition involved a non-conservative change from a nonpolar isoleucine to a polar threonine at nucleotide 10,237, specifically in individual 94. This substitution was conserved in 39 species, implying that this mutation occurred in an important and highly conserved gene.

At nucleotide 9,438, which forms part of an internal branch of group III (Figure 4.9), a transversion was observed. This alteration was conserved in 36 species and resulted in a non-conservative change from a non-polar glycine to a polar serine. This specific substitution has been demonstrated to be associated with the mitochondrial myopathy LHON (Johns and Neufeld, 1993). However, the 9,438 mutation was observed in significant frequencies in African and Cuban mtDNA (Brown *et al.*, 1994). This suggested that the 9,438 mutation had no pathogenic consequences for most individuals

(Brown *et al.*, 1994). With the addition of the 10 !Kung sequences to the topology on the left of the L0-haplogroup tree, as illustrated in Figure 4.8, the substitution at nucleotide 4,232 became a change in an internal branch. This specific transition resulted in a nonconservative change from a nonpolar isoleucine to a polar threonine that is conserved in 31 species. The internal branch substitution at nucleotide 13,276 is conserved in 31 species and involves a conservative change from nonpolar methionine to a nonpolar valine.

The internal branch k*a* changes could have arisen through adaptive selection and might involve amino acids that are functionally important (Ruiz-Pesini, 2003). The majority of the conserved non-synonymous mutations observed in the L0 haplogroup were observed in ND genes. This was in contrast to the majority observed in *cyt b* genes of Asian individuals that represented temperate climatic zones (Mishmar *et al.*, 2003).

4.7 CONSENSUS SEQUENCE

The Khoi-San, and in particular the !Kung, population have been shown to be representative of one of the most divergent, and accordingly most ancient, populations in the world. Therefore it was only possible to derive a consensus, rather than a reference, sequence. When compared to all 13 individuals, the consensus sequences showed 99.25% similarity to each derived sequence. When compared to the RCRS, the sequence generated in this study was the more accurate African-specific mitochondrial reference sequence. The RCRS, apart from containing erroneous sites, was also derived from sequences that are not even of human origin.

Comparative analysis of any haplogroup with the most ancient mitochondrial consensus sequences would enable the researcher to investigate the evolution of the mitochondrial genome over time. By utilising this consensus sequence for future comparative analyses, population-specific polymorphisms could be excluded from the mtDNA sequences of patients with mitochondrial myopathies, and would narrow down possible pathogenic candidate substitutions. This avoids the erroneous identification of alterations when comparisons are made with the RCRS.

From the variation that was observed between the Khoi-San-specific consensus sequence and that of the RCRS it was concluded that consensus sequences should be generated for every population-specific haplogroup in order to have a haplogroup-specific reference. In a

phylogenetic context, the consensus sequence generated in this study could be utilised to examine the evolution of the mitochondrial genome of any haplogroup. This may assist in understanding how haplogroups adapted over time. The derived consensus sequence forms a baseline for all other mitochondrial sequences, thus anchoring the current group of global haplogroup polymorphisms.

CHAPTER FIVE

CONCLUSION

This was the first study in which the whole mitochondrial genome sequence within the Southern African Khoi-San population was analysed. The Khoi-San population represents one of the most diverse and ancient lineages in the world (Chen *et al.*, 2000) and was expected to cluster in the L0 haplogroup (Mishmar *et al.*, 2003), of which the phylogenetic internal topologies were not well known. A consensus sequence of the most ancient African population would allow investigation of the evolution of mitochondrial variation.

The complete, consistent, mitochondrial genomes of 13 Khoi-San individuals were derived. From comparative analysis with the RCRS a total of 173 discrepancies were revealed for all the individuals that were studied, of which 74 substitutions had not been reported previously. These discrepancies were expected since the RCRS is based on an individual from haplogroup H, in contrast to the L-haplogroup ancestry of the Khoi-San population. Further investigation is needed to determine whether the 74 unreported substitutions are polymorphic or not. This underlines the shortcomings of the RCRS as a true reference for phylogenetic purposes.

Via investigation of the nucleotide substitutions, particularly those associated with RE sites, it was observed that the derived sequences all clustered in the L* macrolineage, and possibly haplogroup L1. Specific haplogroups are defined or associated with certain nucleotide substitutions, as illustrated in Appendix B and Figure 4.3 (page 54). Based on this data, the 13 derived Khoi-San sequences clustered either in haplogroup L0 or L1. By constructing an NJ tree, consisting of 106 L-specific sequences, previously determined, as well as these 13 Khoi-San sequences, the !Kung individuals were clustered in haplogroup L0. Three substitutions at nucleotides 1,438, 9,438 and 15,924 were observed and had previously been suggested to be associated with mitochondrial myopathies (Brown *et al.*, 1992; Tawata *et al.*, 1997; Yoon *et al.*, 1991). However, these mutations were also observed at polymorphic frequencies in different populations. Not one of the 13 !Kung individuals that were analysed had any myopathic symptoms. This supports the notion that the genetic background, such as their haplogroup, on which a mutation is expressed, is important. An example of such an occurrence is where the same mutation was observed in

different haplogroups, but was associated with LHON only in individuals that represented haplogroup J (Brown *et al.*, 2001), as discussed in section 4.2 (page 54).

From phylogenetic analysis, in particular NJ and MP analyses, it was suggested that the Khoi-San sequences clustered into three main groups, as illustrated in Table 4.1, as well as Figures 4.6 and 4.7 (pages 59 and 61) respectively, with further subdivision of groups II and III (Figures 4.6 and 4.7). Bootstrap values obtained for the main branches were high, suggesting a high probability of clustering into these specific groups.

The !Kung individuals, with the inclusion of the 106 L-specific sequences, were also subjected to statistical analyses of Tajima and Fu and Li, which suggested the rejection of the hypothesis of neutral evolution for haplogroups L0-L3. Values obtained from both these tests suggested that variation in all four L-specific lineages could be explained by either positive or negative selection or population expansion. Pairwise comparisons suggested that the L0 population, including the 13 !Kung individuals, as well as the L1 and L2 populations, did not undergo rapid population expansions, as observed from the ragged distribution of the resulting graphic representations indicated in Graphs 4.1 to 4.4. These results are supportive of the fact that these African-specific haplogroups consisted of a stable mtDNA population size for a great number of years. In contrast, a bell-shaped graph was observed for the L3 population, suggesting a population expansion that distorted the mtDNA distribution. This is in accordance with the out-of-Africa theory and suggests that a subset of people migrated from Africa to the other continents (Chen *et al.,* 2000). Thus, the subset of people that migrated from Africa to establish the ancestors of modern Eurasia was subjected to a population expansion in Africa, prior to their subsequent migration.

The 13 !Kung sequences were compared to the sequences of 12 L0-specific individuals, from which an L0-specific tree was constructed. Based on this tree, the derived sequences grouped into two main topologies, as illustrated in Figures 4.10 and 4.11. Three sequences, representing group I as illustrated in Table 4.1, clustered in the branch on the right, while the other 10 grouped in the left branch. The three sequences that were added to the topology on the right were only added to the terminal branches. This raises the question whether these sequences were subjected to an adaptive force, which prevented branching events or alternatively to a purifying force that resulted in the elimination of branches that did occur (Mishmar *et al.,* 2003).

Estimation of the MRCA of the L0 haplogroup, at 133,585 ± 5,240 YBP, correlated with the MRCA estimated by RFLP and CR data, which was calculated to be 125,500 to 165,500 YBP (Chen *et al.*, 1995). It was also in accordance with whole mtDNA data, estimated to be 143,000 ± 18,000 YBP (Horai *et al.,* 1995). Thus, consistent with previous results (Chen *et a*l., 2000), this study confirmed that the !Kung are positioned at the deepest root of the African phylogenetic tree.

Subsequent to comparison of the ratio of the non-synonymous to synonymous mutations in the terminal and internal branches it was suggested that terminal branches of the L0 tree harboured more non-synonymous mutations than internal branches. However, a number of these mutations may still be excluded from the tree through evolution in the future. Alternatively, future investigation may reveal that it has been excluded in the past in, as yet, unidentified individuals. Although significant *P* values were not observed in the internal and terminal branches for haplogroups L1 to L3, the actual numbers of the internal compared to terminal branches do show a tendency to differ. By increasing the sample size it may be possible to determine the significance of the current findings.

From the significant difference (P = 0.0348) in the ratio of non-synonymous to synonymous mutations in the internal and terminal branches it was suggested that the L0-haplogroup may have been subjected to negative selection. This was suggested from the observation of the low ratio for the internal vs the terminal branches, as discussed in section 4.6. Given the slight increase in the ratio of non-synonymous to synonymous changes in the internal and terminal branches it was suggested that the populations that descended from L0, namely L1 to L3, have been subjected to positive or adaptive selection.

Six highly conserved non-synonymous mutations were observed in the L0 tree, with the addition of the 13 !Kung sequences. Of these six, three were observed in internal branches. This raises the question whether these mutations had an adaptive function, since these mutations were not excluded through time but maintained their positions in the internal branches. It seems as if non-synonymous mutations are primarily observed in ND genes (Mishmar *et al.,* 2003). It could be hypothesised that these ND mutations were acquired to adapt to the climatic zone and dietary intake of these ancient African populations (Mishmar *et al.*, 2003).

The derived Khoi-San-specific mitochondrial sequence represented the consensus of the most ancient lineage in the world. From a phylogenetic point of view, comparative analysis of global mitochondrial haplogroups with this sequence would enable investigation of the evolution of the mitochondrial genome. Utilising the derived consensus sequence would exclude population-specific polymorphisms from the sequence of especially African patients and would assist in directing attention to possible pathogenic mutations.

Future prospects of this ongoing study include the investigation of each mutation of the 13 derived !Kung sequences, in particular the non-synonymous changes that are highly conserved. These changes should be investigated to determine whether they result in altered protein structure, or function or perhaps have an epistatic effect. These nucleotide substitutions should also be investigated for adaptive function, particularly those occurring in the ND genes. Functional studies could be performed to achieve this goal by investigating the activity levels of, for example cyt *b*, in populations adapted to different climatic regions (Mishmar *et al.*, 2003). Adaptability could also be investigated on the genetic level where a mitochondrial protein could be investigated for genetic changes in populations from different climatic zones.

A further aim for the continuation of this study would be to increase the sample population size in order to derive more !Kung mitochondrial genome sequences and to verify if any of the newly derived sequences may be added to the central branch of the original L0 haplogroup tree that is illustrated in Appendix C. These sequences should be compared to the RFLP and CR data analysed by Chen *et al.* (2000), in particular to those Khoi-San individuals that were suggested to cluster in haplogroups L2 and L3, and their clustering investigated based on their full mitochondrial sequences. Comparative analysis of the !Kung sequences with those from the Khwe population should be performed in order to broaden the knowledge of mitochondrial variation of the southern African Khoi-San population. Furthermore, the southern African Khoi-San sequences should be compared to Khoi-San populations from different geographical regions, more specifically to the Khoi-San population that was analysed by Tishkoff and Williams (2002). This particular population was suggested to represent the most ancient lineage located in north Africa. These Khoi-San sequences will enrich the African-specific database. The mitochondrial, together with Y-chromosome data from these individuals would shed light on the genetic variation of the most ancient population in the world.

This sequence, derived from the most ancient lineage, could be utilised to study the adaptation of the mitochondrial genome over time. Furthermore, it could aid in understanding how different populations, represented by different haplogroups, adapted to specific climates and available food resources. Thus, the genetic analysis of the Khoi-San population will contribute to our understanding of the ancestors of modern humans, how they evolved and offer possible explanations of why certain phenotypes are associated with specific haplogroups.

Analysis of mtDNA has been an important tool to examine human evolution, owing to its maternal inheritance, high mutation rate and absence of recombination. From previous studies it was observed that Africa hosts the most diverse haplogroup populations and that the Khoi-San populations are located at the deepest root of the global mtDNA phylogenetic tree (Chen *et al.*, 2000). Despite the fact that Africa is hypothesised to be the ancestral home of modern humans, African populations have been understudied in comparison to other non-African populations.

Mitochondrial sequence comparisons of especially African patients with the RCRS reveal discrepancies. As a result, possible pathogenic mutations are difficult to identify. The consensus sequence derived in this study would be a more accurate mtDNA reference sequence when utilised in comparative analysis. Although this consensus would not aid in excluding population-specific polymorphisms, it would provide a useful tool for future comparative analyses in both the medical and phylogenetic spheres of science.

REFERENCES

6.1 GENERAL REFERENCES

Allen J.F. and Raven J.A. Free-radical-induced mutation vs redox regulation: costs and benefits of genes in organelles. *J. Mol. Biol.*, **42**, 482-492, 1996.

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.

Attardi G. The human mitochondrial genetic system, In: Mitochondrial DNA in Human Pathology. DiMauro S. and Wallace D.C., editors. New York, Raven Press, 9-25, 1993.

Bauer F.M., Gempel K., Hofmann S., Jaksch M., Philbrook C. and Gerbitz K. Mitochondrial Disorders. A diagnostic challenge in clinical chemistry. *Clin. Chem. Lab. Med.*, **37**, 855-875, 1999.

Benda C. Ueber die Spermatogenese der Vertebraten und höherer Evertebraten. II. Theil: Die Histiogenese der Spermien (The spermatogenesis of vertebrates and higher invertebrates. The histology of sperm). *Arch. Anal. Physiol.,* (Physiol. Abt.), 393-398, 1898.

Berg O.G. and Kurland C.G. Why Mitochondrial genes are Most Often Found in Nuclei. *Mol. Biol. Evol.*, **17**, 951-961, 2000.

Blanchard J.L. and Lynch M. Organellar genes. Why do they end up in the nucleus? *Trends Genet.*, **16**, 315-320, 2000.

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

Bourgeron T., Rustin P., Chretirn D., Birch-Machin M., Bourgeois M., Viegas-Péquignot., Munnich A. and Rötig. Mutation of a nuclear succinate dehydrogenase gene results in mitochondrial respiratory chain deficiency. *Nat. Genet.*, **11**, 144-149, 1995.

Brown M.D., Torroni A., Shoffner J.M. and Wallace D.C. Mitochondrial $tRNA^{Thr}$ mutations and lethal infantile mitochondrial myopathy. *Am. J. Hum. Genet.*, **51**, 446-447, 1992.

Brown M.D., Torroni A., Huoponen K., Chen Y., Lott M. and Wallace D.C. Pathological significance of the mtDNA COXIII mutation at nucleotide pair 9438 in Leber Hereditary Optic Neuropathy. *Am J. Hum. Genet.*, **55**, 410, 1994.

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.

Campbell M.K. Biochemistry, Second Edition, Saunders College Publishing, 391-416, 1991.

Cann R.L. and Wilson A.C. Length mutations in human mitochondrial DNA. *Genetics*, **104**, 699-711, 1983.

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

Carroll J., Shannon R.J., Fearnley I.M., Walker J.E. and Hirst J. Definition of the nuclear encoded protein composition of bovine heart mitochondrial complex I. *J. Biol. Chem.,* **277**, 50311-50317, 2002.

Chang D.D. and Clayton D.A. Priming of human mitochondrial DNA replication occurs at the light-strand promoter. *Proc. Natl. Acad. Sci. U.S.A.*, **82**, 351-355, 1985.

Chen Y.S., Torroni A., Excoffier L., Santachiara-Benerecetti A.S. and Wallace D.C. Analysis of mtDNA variation in African populations reveals the most ancient of all human continent-specific haplogroups. *Am. J. Hum. Genet.*, **57**, 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 Khwe – and their genetic relationships to other African populations. *Am. J. Hum. Genet.*, **66**, 1362-1383, 2000.

Christianson T.W. and Clayton D.A. In vitro transcription of human mitochondrial DNA: accurate termination requires a region of DNA sequence that can function bidirectionally. *Proc. Natl. Acad. Sci. U.S.A.,* **83**, 6277-6281, 1986.

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

Clayton D.A. Transcription of the mammalian mitochondrial genome. *Annu. Rev. Biochem.*, **53**, 573-594, 1984.

Colilla S., Rotimi C., Cooper R., Goldberg J. and Cox N. Genetic inheritance of Body Mass Index in African-American and African families. *Genet. Epidemiol.*, **18**, 360-376, 2000.

Corral-Debrinski M., Horton T., Lott M.T., Shoffner J.M., Beal M.F. and Wallace D.C. Mitochondrial DNA deletions in human brain: regional variability and increase with advanced age. *Nat. Genet.*, **2**, 324-329, 1992.

Corral-Debrinski M., Horton T., Lott M.T., Shoffner J.M., McKee A.C., Beal M.F., Graham B.H. and Wallace D.C. Marked changes in mitochondrial DNA deletion levels in Alzheimer brains. *Genomics*, **23**, 471-476, 1994.

Cummings M.P., Otto S.P. and Wakely J. Sampling properties of DNA sequence data in phylogenetic analysis. *Mol. Biol. Evol.*, **12**, 814-822, 1995.

David I.B. and Blackler W. Maternal and cytoplasmic inheritance of mitochondrial DNA in *Xenopus*. *Dev. Biol.*, **29**, 152-161, 1972.

Denaro M., Blanc H., Johnson M.J., Chen K.H., Wilmsen E. and Cavalli-Sforza L.L. Ethnic variation in *Hpa* I endonuclease cleavage patterns of human mitochondrial DNA. *Proc. Natl. Acad. Sci. U.S.A.*, **78**, 5768-5722, 1981.

De Vivo D.C. Mitochondrial DNA defects: clinical features. In: Mitochondrial DNA in human pathology. DiMauro s. and Wallace D.C., editors. New York, Raven Press, 39-52, 1993.

DiMauro S., Bonilla E., Lombes A., Shanske S., Minetti C. and Moraes C.T. Mitochondrial encephalomyopathies. *Neurol. Clin.*, **8**, 483-506, 1990.

DiMauro S. Mitochondrial encephalomyopathies: what next? *J. Inherit. Metab. Dis.*, **19**, 489-503, 1996.

Duby G. and Boutry M. Mitochondrial protein import machinery and targeting information. *Plant Science*, **162**, 477-490, 2002.

Eilers M., Hwang S. and Schatz G. Unfolding and refolding of a purified precursor protein during import into isolated mitochondria. *E.M.B.O. J.*, **7**, 1139-1145, 1988.

Embley T.M., Horner D.A. and Hirt R.P. Anaerobic eukaryote evolution: hydrogenosomes as biochemically modified mitochondria? TREE, **12**, 437-441, 1997.

Fairbanks D.J. and Andersen W.R. Extranuclear inheritance. In: Genetics, the continuity of life. U.S.A:, Brooks/Cole Publishing Company, 549-570, 1999.

Felsenstein J. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, **39**, 783-791, 1985.

Fu Y. and Li W. Statistical tests of neutrality of mutations. *Genetics*, **133**, 693-709, 1993.

Garrett R.H. and Grisham C.M. Electron transport and oxidative phosphorylation. In: Biochemistry, 2nd Edition. U.S.A: Saunders College Publishing, 673-703, 1999.

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

Gillum A.M. and Clayton D.A. Mechanism of mitochondrial DNA replication in mouse L-cells: RNA priming during the initiation of heavy-strand synthesis. *J. Mol. Biol.*, **135**, 353-368, 1979.

Goto Y., Nonaka I. and Horai S. A mutation in the $tRNA^{Leu(UUR)}$ gene associated with the MELAS subgroup of mitochondrial enephalomyopathies. *Nature*, **348**, 651-653, 1990.

Gray M.W. Evolution of organellar genomes. *Curr. Opin. Genet. Dev.*, **9**, 678-687, 1993.

Guthrie M. Some developments in the prehistory of the Bantu origins. *J. Afr. Hist.*, **3**, 273-282, 1962.

Hammer M.F., Karafet T., Rasanayagam A., Wood E.T., Altheide T.K., Jenkins T. and Griffiths R.C. Out of Africa and back again: nested cladistic analysis of human Y chromosome variation. *Mol. Biol. Evol.*, **15**, 427-441, 1998.

Holt I.J., Harding A.E., Petty R.K.H. and Morgan-Hughes J.A. A new mitochondrial disease associated with mitochondrial DNA heteroplasmy. *Am. J. Hum. Genet.*, **46**, 428-433, 1990.

Horai S. Recent African origin of modern humans revealed by complete sequences of homonoid mitochondrial DNAs. *Proc. Natl. Acad. Sci. U.S.A.*, **92**, 532-536, 1995.

Howell N., McCullough D.A., Kubacka I., Halvorson S. and Mackey D. The sequence of human mtDNA: The question of errors versus polymorphisms. *Am. J. Hum. Genet.*, **50**, 1333-1337, 1992.

Howell N., Kubacka I., Havorson 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.

Inoue K., Nakada K., Ogura A., Isobe K., Goto Y., Nonaka I. and Hayashi J. Generation of mice with mitochondrial dysfunction by introducing mouse mtDNA carrying a deletion into zygotes. *Nat. Genet.*, **26**, 176-181, 2000.

Johns D.R. and Neufeld M.J. Cytochrome c oxidase mutations in Leber hereditary optic neuropathy. *Biochem. Biophys. Res. Commun.*, **196**, 810-815, 1993.

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

Kimura M. Evolutionary rate at the molecular level. *Nature*, **217**, 624-626, 1968.

Koehler C.M., Leuenberger D., Merchant S., Renold A., Junne T. and Schatz G. Human deafness dystonia syndrome is a mitochondrial disease. *Proc. Natl. Acad. Sci. U.S.A.*, **96**, 2141-2146, 1999.

Lahr M.M. and Foley R.A. Towards a theory of modern human origins: geography, demography and diversity on recent human evolution. *Yb. Phys. Anthropol.*, **41**, 137-176, 1998.

Liao H. and Spremulli L.L. Interaction of bovine mitochondrial ribosomes with messenger RNA. *J. Biol. Chem.*, **264**, 7518-7522, 1989.

Liao H. and Spremulli L.L. Identification and initial characterization of translational initiation factor 2 from bovine mitochondria. *J. Biol. Chem.*, **265**, 13618-13622, 1990.

Loeffen J., Smeitink J., Triepels R., Smeets R., Schuelke M., Sengers R., Trijbels F., Hamel B., Mullaart R. and Van den Huevel L. The first nuclear-encoded complex I mutation in a patient with Leigh syndrome. *Am. J. Hum. Genet.*, **63**, 1598-1609, 1998.

Lutsenko S. and Cooper M.J. Localization of the Wilson's disease protein product to mitochondria. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 6004-6009, 1998.

Majamaa K., Moilanen J.S., Uimonen S., Remes A.M., Salmela P.I., Kärppä M., Majamaa-Voltti K.A.M., Sorri M., Peuhkurinen K.J. and Hassinen I.E. Epidemiology of A3243G, the mutation for mitochondrial encephalomyopathy, lactic acidosis, and strokelike episodes: prevalence of the mutation in an adult population. *Am. J. Hum. Genet.*, **63**, 447-454, 1998.

MAMMAG, The Center for Molecular and Mitochondrial Medicine and Genetics, University of California, Irvine, CA 92697-3940, U.S.A.

Margulis L. The symbiotic theory. In: Origin of eukaryotic cells. New Haven and London, Yale University Press, 45-68, 1970.

Martin W. and Müller M. The hydrogen hypothesis for the first eukaryote. *Nature*, **392**, 37-41, 1998.

Meinkoth J. and Wahl G. Hybridisation of nucleic acids immobilized on solid supports. *Anal. Biochem.*, **138**, 267-284, 1984.

Mishmar D., Ruitz-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. U.S.A.*, **100**, 171-176, 2003.

Montoya J., Ojala D. and Attardi G. Distinctive features of the 5' - terminal sequence of the human mitochondrial mRNAs. *Nature*, **290**, 465-470, 1981.

Moraes C.T., Andreetta F., Bonilla E., Shanske S., Dimauro S. and Schon E.A. Replication-competent human mitochondrial DNA lacking the heavy-strand promoter region. *Mol. Cell Biol.*, **11**, 1631-1637, 1991a.

Moraes C. T., Shanske S., Tritschler H. J., Aprille J. R., Andreetta F., Bonilla E., Schon E. A. and DiMauro, S. MtDNA depletion with variable tissue expression: a novel genetic abnormality in mitochondrial diseases. *Am. J. Hum. Genet.*, **48**, 492-501, 1991b.

Morral N., Bertranpetit J., Estivill X., Nunes V., Casals T., Giménez J., Reis A., Varon-Mateeva R., Macek Jr M., Kalaydjieva L., Angelicheva D., Dancheva R., Romeo G., Russo M.P., Garnerone S., Restagno G., Ferrari M., Magnani C., Claustres M., Desgeorges M., Schwartz M., Schwarz M., Dallapiccola B., Novelli G., Ferec C., de Arce M., Nemeti M., Kere J., Anvret M., Dahl N. and Kadasi L. The origin of the major cystic fibrosis mutation (ΔF508) in European populations. *Nat. Genet.*, **7**, 169-175, 1994.

Muller H.J. The relation of recombination to mutational advance. *Mut. Res.*, **1**, 2-9, 1964.

Page R.D.M. and Holmes E.C. Inferring molecular phylogeny. In: Molecular Evolution. A Phylogenetic Approach. United Kingdom, Blackwell Science Ltd., 172-279, 1998.

Passarge E. Polymorphism. In: Color Atlas of Genetics. New York, Thieme Medical Publishers, Inc., 156, 1995.

Piko L. and Matsumoto L. Number of mitochondria and some properties of mitochondrial DNA in the mouse egg. *Dev. Biol.*, **49**, 1-10, 1976.

Quintana-Murci L., Ornella S., Bandelt H., Passarino G., McElreavey K. and Santachiara-Benerecetti S. Genetic evidence of an early exit of *Homo sapiens sapiens* from Africa through eastern Africa. *Nat. Genet.*, **23**, 437-441, 1999.

Rozas J. and Rozas R. DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics*, **15**, 174-175, 1999.

Ruiz-Pesini E. Centre for Molecular and Mitochondrial Medicine and Genetics, University of California at Irvine, Irvine, U.S.A. Personal communication, 2003.

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

Ruiz-Pesini E., Mishmar D., Brandon M., Procaccio V. and Wallace D.C. Distinguishing the effect of purifying and adaptive selection on regional variation in human mtDNA. *Science*, **303**, 223-226, 2004.

Saillard J., Forster P., Lynnerup N., Bandelt H. and Nørby S. mtDNA Variation among Greenland Eskimos: the edge of the Beringian expansion. *Am. J. Hum. Genet.*, **67**, 718-726, 2000.

Saitou N. and Nei M. The neighbour-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406-425, 1987.

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

Scholte H.R. The biochemical basis of mitochondrial diseases. *J. Bioener. Biomem.*, **20**, 161-191, 1988.

Schon E.A. Mitochondria. In: Mitochondrial DNA in Human Pathology. DiMauro S. and Wallace D.C., editors. New York, Raven Press, 1-7, 1993.

Schwartzbach C.J. and Spremulli L.J. Bovine mitochondrial protein synthesis elongation factors. *J. Biol. Chem.*, **264**, 19125-19131, 1989.

Selosse M., Albert B. and Godelle B. Reducing the genome size of organelles favours gene transfer to the nucleus. *Trends Ecol. Evol.*, **16**,135-141, 2001.

Semino O., Santachiara-Benerecetti A.S., Falaschi F., Cavalli-Sforza L.L. and Underhill P.A. Ethiopians and Khoisan share the deepest clades of the human Y-chromosome phylogeny. *Am. J. Hum. Genet.*, **70**, 265-268, 2002.

Shitara H., Hayashi J.I., Takahama S., Kaneda H. and Yonekawa H. Maternal inheritance of mouse mtDNA in interspecific hybrids: segregation of the leaked paternal mtDNA followed by the prevention of subsequent paternal leakage. *Genetics*, **148**, 851-857, 1998.

Shoffner J.M., Lott M.T., Lezza A.M.S., Seibel P., Ballinger S.W. and Wallace D.C. Myoclonic epilepsy and ragged-red fiber disease (MERRF) is associated with a mitochondrial DNA $tRNA^{Lys}$ mutation. *Cell*, **61**, 931-937, 1990.

Sirrenberg C., Bauer M.F., Guiard B., Neupert W. and Brunner M. Import of carrier proteins into the mitochondrial inner membrane mediated by Tim22. *Nature*, **384**, 582-585, 1996.

Slatkin M. and Hudson R.R. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics*, **129**, 555-562, 1991.

Smith L.M., Sanders J.Z., Kaiser R.J., Hughes P., Dodd C., Connell C.R., Heiner C., Kent S.B.H. and Hodd L.E. Fluorescence detection in automated DNA sequence analysis. *Nature*, **321**, 674-679, 1986.

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 S., editor. Oxford, Oxford University Press, 3-25, 2003.

Suzuki Y., Suzuki S., Hinokio Y., Chiba M., Atsumi Y., Hosokawa K., Shimada A., Asahina T. and Matsuoka K. Diabetes associated with a novel 3264 mitochondrial $tRNA^{Leu(UUR)}$ mutation. *Diabetes Care*, **20**, 1138-1140, 1997.

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

Tajima F. Statistial method for testing the neutral mutation hypothesis by DNS polymorphism. *Genetics*, **123**, 585-595, 1989.

Tawata M., Ohtaka M., Iwase E., Ikegishi Y., Aida K. and Onaya T. New mitochondrial DNA homoplasmic mutations associated with Japanese patients with type 2 diabetes. *Diabetes*, **47**, 276-277, 1998.

Templeton A.R. Out of Africa again and again. *Nature,* **416**, 45-51, 2002.

Thein S.L. and Wallace R.B. The use of synthetic oligonucleotides as specific hybridization probes in the diagnosis of genetic disorders. In: Human genetic diseases: a practical approach. Davies K.E., editor. Oxford, IRL Press, 33-50, 1986.

Thompson J.D., Gibson T.J., Plewniak F., Jeanmougin F. and Higgins D.G. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.*, **24**, 4876-4882, 1997.

Tiranti V., Hoertnagel K., Carrozzo R., Galimberti C., Munaro M., Granatiero M., Zelante L., Gasparini P., Marzella R., Rocchi M., Bayona-Bafaluy M.P., Enriquez J., Uziel G., Bertini E., Dionisi-Vici C., Frnco B., Meitinger T. and Zeviani M. Mutations of SURF-1 in Leigh Disease associated with cytochrome *c* oxidase deficiency. *Am. J. Hum. Genet.*, **63**, 1609-1621, 1998.

Tishkoff S.A. and Williams S.M. Genetic analysis of African populations: Human evolution and complex disease. *Nat. Rev. Genet.*, **3**, 611-619, 2002.

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, 1993.

Torroni A., Lott M.T., Cabell M.F., Chen Y.S., Lavergne L. and Wallace D. 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, 1994.

Torroni A., Huoponen K., Francalacci P., Petrozzi M., Morelli L., Scozzari R., Obinu D., Savontaus M.L. and Wallace D. Classification of European mtDNAs from an analysis of three European populations. *Genetics*, **144**, 1835-1850, 1996.

Torroni A., Rengo C., Guida V., Cruciani F., Sellitto D., Coppa A., Calderon L.F., 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.

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.

Wallace D.C. Mitotic segregation of mitochondrial DNAs in human cell hybrids and expression of chloramphenicol resistance. *Somat. Cell. Mol. Genet.*, **12,** 41- 49, 1986.

Wallace D.C. Mitochondrial DNA sequence variation in human evolution and disease. *Proc. Natl. Acad. Sci. U.S.A.*, **91**, 8739-8746, 1994.

Wallace D.C. Mitochondrial DNA variation in human evolution, degenerative disease, and aging. *Am. J. Hum. Genet.*, **57**, 201-223, 1995.

Wallace D.C. Centre for Molecular Medicine, Emory University School of Medicine, Atlanta, U.S.A. Personal communication, 1999.

Wallace D.C. Centre for Molecular and Mitochondrial Medicine and Genetics, University of California at Irvine, Irvine, U.S.A. Personal communication, 2003.

Wallace D.C. and Lott M.T. Maternally Inherited Diseases, In: Mitochondrial DNA in human pathology. DiMauro S. and Wallace D.C., editors. New York, Raven Press, 63-83, 1992.

Wallace D.C., Ye J.H., Neckelmann S.N., Singh G., Webster K.A. and Greenberg, B.D. Sequence analysis of cDNAs for the human and bovine ATP synthase β-subunit: mitochondrial DNA genes sustain seventeen times more mutations. *Curr. Genet.*, **12**, 81-90, 1987.

Wallace D.C., Singh G., Lott M.T., Hodge J.A., Schurr T.G., Lezza A.M.S., Elsas L.J. and Nikoskelainen E.A. Mitochondrial DNA mutation associated with Leber's Hereditary Optic Neuropathy. *Science*, **242**, 1427-1430, 1988.

Wang T.S. Eukaryotic DNA polymerases. *Annu. Rev. Biochem.*, **60**, 513-552, 1991.

Watson E., Forster P., Richards M. and Bandelt H. Mitochondrial Footprints of Human Expansions in Africa. *Am. J. Hum. Genet.*, **61**, 691-704, 1997.

Watterson G.A. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.*, **7**, 256-276, 1975.

Westphal S.P. Rewrite the textbooks. *NewScientist*, **28 June**, 22, 2003.

Wikström M. Oxidative phosphorylation: an overview, In: Genetics of mitochondrial diseases. Holt S., editor. Oxford, Oxford University Press, 69-109, 2003.

Yoon K.L., Aprille J.R. and Ernst S.G. Mitochondrial $tRNA^{Thr}$ mutation in fatal infantile respiratory enzyme deficiency. *Biochem. Biophys. Res. Commun.*, **176**, 1112-1115, 1991.

Zeviani M., Moraes C.T., DiMauro S., Nakase H., Bonilla E., Schon E.A. and Rowland L.P. Deletions of mitochondrial DNA in Kearns-Sayre syndrome. *Neurology*, **38**, 1339-1346, 1988.

6.2 ELECTRONIC REFERENCES

Clustal X, A windows interface and clustering analysis for multiple sequence and profile alignments. http://inn-prot.weizmann.ac.il/software/ClustalX.html.

DnaSP, DNA Sequence Polymorphism version 3.51. http://www.bio.ub.es/~julio/DnaSP.html.

Kumar S., Tamura K., Jakobsen I.B. and Nei M. MEGA2: Molecular Evolutionary Genetics Analysis software, Arizona State University, Tempe, Arizona, U.S.A., 2002.

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

MITOMAP, A Human mitochondrial Genome Database, Centre for Molecular Medicine, Emory University, Atlanta, GA, U.S.A. http://www.mitomap.org.

MEGA, Molecular Evolutionary Genetics Analysis software version 2, Arizona State University, Tempe, Arizona, U.S.A. http://www.megasoftware.net.

APPENDIX A

SEQUENCE COMPARISONS BETWEEN THE KHOI-SAN AND RCRS

Base pair substitutions that were observed when the derived Khoi-San sequences were compared to the revised Cambridge reference sequence (RCRS) are listed in Table A.1.

Table A.1: Sequence comparisons between the derived Khoi-San sequences and the RCRS

bp #	Individual number												
	10	32	38	40	52	72	80	82	94	102	106	122	130
73	A-G	A-G	A-G	A-G	A-G	A-G	A-G	A-G	A-G	A-G	A-G	A-G	A-G
146	T-C	T-C	T-C	T-C	T-C	T-C	T-C	T-C	T-C	T-C	T-C	T-C	T-C
152	T-C	T-C	T-C	T-C	T-C	T-C	T-C	T-C	T-C	T-C	T-C	T-C	T-C
188					A-G								
189			A-G				A-G				A-G		
195	T-C	T-C	T-C	T-C	T-C	T-C	T-C	T-C		T-C	T-C	T-C	T-C
198	C-T	C-T	C-T			C-T	C-T	C-T		C-T	C-T	C-T	
199									T-C				T-C
204											T-C		
207			G-A				G-A				G-A		
247	G-A	G-A	G-A	G-A	G-A	G-A	G-A	G-A	G-A	G-A	G-A	G-A	G-A
303	Ins CC	Ins CC				Ins CC		Ins CC			Ins CC	Ins CC	
309	C-T	C-T	C-T	C-T	C-T	C-T	C-T	C-T	C-T	C-T	C-T	C-T	C-T
310	T-C	T-C	T-C	T-C	T-C	T-C	T-C	T-C	T-C	T-C	T-C	T-C	T-C
456	T-C	T-C				T-C		T-C		T-C		T-C	
494	Del C	Del C		Del C	Del C	Del C		Del C	Del C	Del C		Del C	Del C
514	Del C	Del C	Del C			Del C	Del C	Del C		Del C	Del C	Del C	
515	Del A	Del A	Del A			Del A	Del A	Del A		Del A	Del A	Del A	
719	G-A	G-A		G-A	G-A	G-A		G-A	G-A	G-A		G-A	G-A
750	A-G	A-G	A-G	A-G	A-G	A-G	A-G	A-G	A-G	A-G	A-G	A-G	A-G
769	G-A	G-A	G-A	G-A	G-A	G-A	G-A	G-A	G-A	G-A	G-A	G-A	G-A
825	T-A	T-A	T-A	T-A	T-A	T-A	T-A	T-A	T-A	T-A	T-A	T-A	T-A
850			T-C				T-C				T-C		
1,018	G-A	G-A	G-A	G-A	G-A	G-A	G-A	G-A	G-A	G-A	G-A	G-A	G-A
1,048	C-T	C-T	C-T	C-T	C-T	C-T	C-T	C-T	C-T	C-T	C-T	C-T	C-T
1,243			T-C				T-C		T-C		T-C		T-C
1,438			A-G				A-G				A-G		
1,719									G-A				
2,706			A-G				A-G				A-G		
2,758	G-A	G-A	G-A	G-A	G-A	G-A	G-A	G-A	G-A	G-A	G-A	G-A	G-A
2,836			C-A				C-A				C-A		
2,885	T-C	T-C	T-C	T-C	T-C	T-C	T-C	T-C	T-C	T-C	T-C	T-C	T-C

Table A.1: continued ...

bp #	Individual number												
	10	32	38	40	52	72	80	82	94	102	106	122	130
3,107	N	N	N	N	N	N	N	N	N	N	N	N	N
3,438	G - A	G - A		G - A	G - A	G - A		G - A	G - A	G - A		G - A	G - A
3,516	C - A	C - A	C - A	C - A	C - A	C - A	C - A	C - A	C - A	C - A	C - A	C - A	C - A
3,531					G - A								
3,594	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T
3,618				T - C	T - C								
3,666										G - A			
3,756	A - G	A - G		A - G	A - G	A - G		A - G	A - G	A - G		A - G	A - G
4,104	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G
4,197	C - T	C - T				C - T		C - T		C - T		C - T	
4,232	T - C	T - C		T - C	T - C	T - C		T - C	T - C	T - C		T - C	T - C
4,312	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T
4,541			G - A				G - A				G - A		
4,562					A - C								
4,586			T - C				T - C				T - C		
4,769	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G
4,907			T - C				T - C				T - C		
5,442	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C
5,811			A - G				A - G				A - G		
5,813									A - G				
6,185	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C
6,266	A - G	A - G		A - G	A - G	A - G		A - G	A - G	A - G		A - G	A - G
6,692				A - G									
6,815	T - C	T - C		T - C	T - C	T - C		T - C	T - C	T - C		T - C	T - C
6,938			C - T				C - T				C - T		
7,028	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T
7,146	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G
7,256	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T
7,257			A - G				A - G				A - G		
7,283				T - C					T - C				
7,521	G - A	G - A	G - A	G - A	G - A	G - A	G - A	G - A	G - A	G - A	G - A	G - A	G - A
7,673											A - G		
7,828	A - G	A - G				A - G		A - G		A - G		A - G	
7,961										T - C			
7,967	C - T	C - T				C - T		C - T				C - T	
8,027					G - A								
8,113	C - A	C - A		C - A	C - A	C - A		C - A	C - A	C - A		C - A	C - A
8,152	G - A	G - A		G - A	G - A	G - A		G - A	G - A	G - A		G - A	G - A
8,251	G - A	G - A		G - A	G - A	G - A		G - A	G - A	G - A	G - A	G - A	G - A
8,468	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T
8,655	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T
8,701	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G
8,790				G - A									
8,829						C - T							
8,860	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G
8,911			T - C				T - C				T - C		
8,994			G - A				G - A				G - A		
9,042	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T
9,136			A - G				A - G				A - G		
9,150	A - G	A - G				A - G		A - G		A - G		A - G	
9,215		C - T											

Table A.1: continued ...

bp #	Individual number												
	10	32	38	40	52	72	80	82	94	102	106	122	130
9,347	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G
9,438	G - A	G - A				G - A		G - A		G - A		G - A	
9,540	T – C	T – C	T – C	T – C	T – C	T – C	T – C	T – C	T – C	T – C	T – C	T – C	T – C
9,755	G – A	G – A		G – A	G – A	G – A		G – A	G – A	G – A		G – A	G – A
9,818			C – T				C – T				C – T		
9,887				T – C									
9,950									T – C				T – C
10,237									T – C				
10,398	A – G	A – G	A – G	A – G		A – G	A – G	A – G	A – G	A – G	A – G	A – G	A – G
10,499			A – G				A – G				A – G		
10,589	G – A	G – A	G – A	G – A	G – A	G – A	G – A	G – A	G – A	G – A	G – A	G – A	G – A
10,664	C – T	C – T	C – T	C – T	C – T	C – T	C – T	C – T	C – T	C – T	C – T	C – T	C – T
10,688	G – A	G – A	G – A	G – A	G – A	G – A	G – A	G – A	G – A	G – A	G – A	G – A	G – A
10,707					T – G								
10,810	T – C	T – C	T – C	T – C	T – C	T – C	T – C	T – C	T – C	T – C	T – C	T – C	T – C
10,873	T – C	T – C	T – C	T – C	T – C	T – C	T – C	T – C	T – C	T – C	T – C	T – C	T – C
10,876			A – G				A – G				A – G		
10,915	T – C	T – C	T – C	T – C	T – C	T – C	T – C	T – C	T – C	T – C	T – C	T – C	T – C
10,920			C – T				C – T				C – T		
10,939			C – T				C – T				C – T		
11,296			C – T				C – T				C – T		
11,299			T – C				T – C				T – C		
11,437	T – C	T – C				T – C		T – C		T – C		T – C	
11,653			A – G				A – G				A – G		
11,719	G – A	G – A	G – A	G – A	G – A	G – A	G – A	G – A	G – A	G – A	G – A	G – A	G – A
11,914	G – A	G – A	G – A	G – A	G – A	G – A	G – A	G – A	G – A	G – A	G – A	G – A	G – A
12,007	G – A	G – A	G – A	G – A	G – A	G – A	G – A	G – A	G – A	G – A	G – A	G – A	G – A
12,070			G – A				G – A				G – A		
12,121	T – C	T – C		T – C	T – C	T – C		T – C	T – C	T – C		T – C	T – C
12,142									A – G				A – G
12,235	T – C	T – C				T – C		T – C		T – C		T – C	
12,705	C – T	C – T	C – T	C – T	C – T	C – T	C – T	C – T	C – T	C – T	C – T	C – T	C – T
12,720	A – G	A – G		A – G	A – G	A – G		A – G	A – G	A – G		A – G	A – G
12,798									C – T				C – T
13,020			T – C				T – C				T – C		
13,105	A – G	A – G	A – G	A – G	A – G	A – G	A – G	A – G	A – G	A – G	A – G	A – G	A – G
13,129	C – T	C – T				C – T		C – T		C – T		C – T	
13,276	A – G	A – G	A – G	A – G	A – G	A – G	A – G	A – G	A – G	A – G	A – G	A – G	A – G
13,506	C – T	C – T	C – T	C – T	C – T	C – T	C – T	C – T	C – T	C – T	C – T	C – T	C – T
13,572							T – C						
13,590			G – A				G – A				G – A		
13,650	C – T	C – T	C – T	C – T	C – T	C – T	C – T	C – T	C – T	C – T	C – T	C – T	C – T
13,759	G – A	G – A		G – A	G – A	G – A		G – A	G – A	G – A		G – A	G – A
13,819			T – C				T – C				T – C		
13,855													C – T
13,928			G – C				G – C				G – C		
13,967													C – T
14,020			T – C				T – C				T – C		
14,182			T – C				T – C				T – C		
14,280				A – G									
14,315				C – T									
14,364				G – A									

Table A.1: continued ...

bp #	Individual number												
	10	32	38	40	52	72	80	82	94	102	106	122	130
14,371			T – C				T – C				T – C		
14,374			T – C				T – C				T – C		
14,560				G - A									
14,659				C - T									
14,766	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T
15,217					G - A								
15,326	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G
15,449				T - C									
15,466	G - A	G - A		G - A	G - A	G - A		G - A	G - A	G - A		G - A	G - A
15,479					T - C								
15,550	C - T	C - T				C - T		C - T		C - T		C - T	
15,924									A - G				
15,930	G - A	G - A		G - A	G - A	G - A		G - A	G - A	G - A		G - A	G - A
15,941	T - C	T - C		T - C	T - C	T - C		T - C	T - C	T - C		T - C	T - C
15,951	A - G	A - G				A - G		A - G		A - G		A - G	
16,129					G - A				G - A				G - A
16,153				G - A									
16,166			A - C				A - C				A - C		
16,167	C - T	C - T				C - T		C - T		C - T		C - T	
16,172			T - C				T - C				T - C		
16,179					C - T								
16,183											Ins C		
16,187	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T		C - T	C - T
16,189	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C
16,209			T - C				T - C				T - C		
16,214			C - T				C - T				C - T		
16,223	C - T	C - T	C - T	C - T	C - T	C - T	C - T	C - T		C - T	C - T	C - T	
16,230	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G	A - G
16,234	C - T	C - T				C - T		C - T	C - T	C - T		C - T	C - T
16,242	C - T	C - T				C - T		C - T		C - T		C - T	
16,243	T - C	T - C		T - C	T - C	T - C		T - C	T - C	T - C		T - C	T - C
16,264													C - T
16,266									C - A				C - G
16,278			C - T				C - T				C - T		
16,291			C - G				C - G				C - G		
16,294				C - T									
16,311	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C
16,497								A - G				A - G	
16,519	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C	T - C

Numbers representing the individuals that were sequenced are listed in the table heading. Nucleotides are represented by the capital letters, with the RCRS nucleotide shown first, followed by the observed change, e.g. T – C where T refers to the RCRS; A = adenine; C = cytosine; G = guanine; T = thymine; Ins = insertion; Del = deletion; bp # = base pair number of the mitochondrial genome. Adapted from MITOMAP (2003).

Numbers indicated in red represent previously reported polymorphisms, as listed in MITOMAP (2003). The original CRS harboured a CC at nucleotide numbers 3,106 and 3,107 (Anderson *et al.*, 1981). However, during reanalysis only a single C was observed at nucleotide 3,106 (Andrews *et al.,* 1999). Since the previous numbering system had been

utilised extensively, an N was indicated in the RCRS at nucleotide 3,107 in order to avoid re-numbering of the sequence nucleotides.

APPENDIX B

EXCLUSION CRITERIA FOR L-SPECIFIC SEQUENCES

Base pair substitutions that identify and exclude L-specific sequences from other mitochondrial haplogroups are indicated in Table B.1. The 13 Khoi-San sequences were expected to contain the specific substitutions through which they could be classified as being L-specific sequences (Ruiz-Pesini, 2003).

Table B.1: Polymorphisms excluding L-specific sequences from other mtDNA haplogroups

Exclusion criteria	Base pair number	Substitution
L0 and L1 from other haplogroups	825	T - A
	2,758	G - A
	2,885	T - C
	7,146	A - G
	8,468	C - T
	8,655	C - T
	10,688	G - A
	10,810	T - C
	13,105	A - G
	13,506	C - T
L0, L1 and L2 from other haplogroups	769	G - A
	1,018	G - A
	3,594	C - T
	4,104	A - G
	7,256	C - T
	7,521	G - A
	13,650	C - T
From N	8,701	A - G
	9,540	T - C
	10,398	A - G
	10,873	T - C
From R	12,705	C - T
From HV	11,719	G - A
	14,766	C - T
From H	2,706	A - G
	7,028	C - T

Table B.1: continued ...

Exclusion criteria	Base pair number	Substitution
From the subhaplogroup containing the CRS	750	A - G
	1,438	A - G
	4,769	A - G
From the RCRS	8,860	A - G
	15,326	A - G

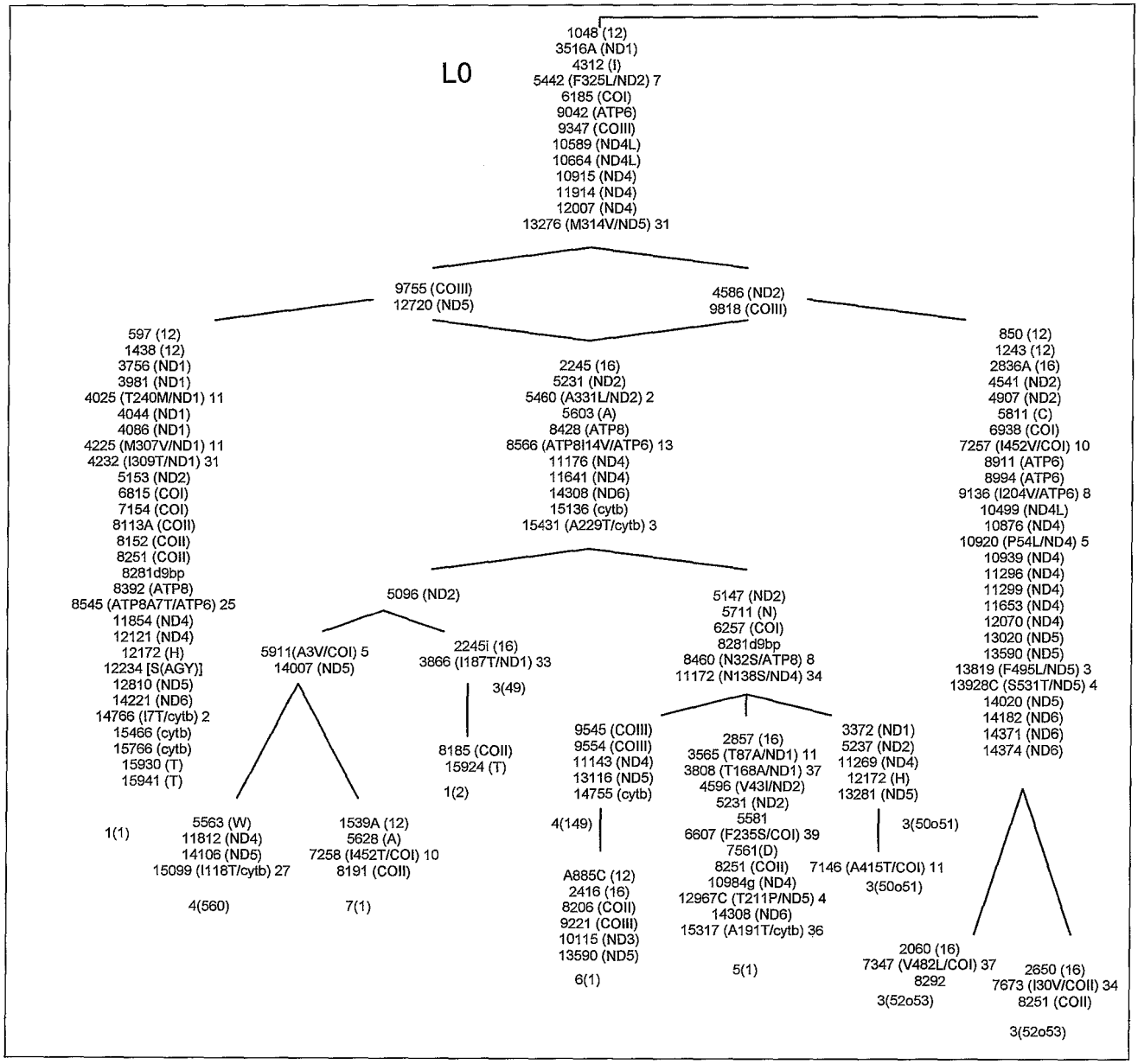
Capital letters in the right column indicate abbreviations of nucleotides, with the RCRS nucleotide shown first, followed by the observed change, e.g. T – C where T refers to the RCRS; A = adenine, C = cytosine, G = guanine, T = thymine. Base pair numbers are according to the numbering system of the RCRS. Capital letters in the left column represent the specific haplogroups. Adapted from Ruiz-Pesini (2003).

APPENDIX C

L0-SPECIFIC HAPLOGROUP TREE

An L0-haplogroup tree derived from 12 individuals is illustrated in Figure C.1. Permission to utilise the tree in comparative analysis was granted by MAMMAG.

Figure C.1: Schematic representation of an L0-haplogroup tree



Red = synonymous substitutions; black = non-synonymous substitutions; blue = substitutions in the non-coding region of the coding region; green = substitutions in RNA genes. Adapted from Ruiz-Pesini (2003).

All synonymous substitutions are transitions, except where an amino acid abbreviation is indicated, for example, 8113A, which represents a transversion. Conservation numbers

are indicated after each non-synonymous substitution. The region where the substitution occurred is indicated in brackets. Single letter abbreviations of the amino acids that are involved in the non-synonymous changes are also indicated in brackets. Numbers at the bottom of specific branches represent the individuals that cluster in that branch.