CHAPTER THREE

Mitochondrial DNA and Human Evolution

The mitochondria were first identified and isolated more than 60 years ago and have since contributed to answering challenging questions related to anthropology, disease, evolution and biogenesis. The existence of DNA within the human mitochondria was discovered in the 1960's (Nass *et al.*, 1964) and this discovery was followed by the publication of the first mtDNA sequence in 1981 (Anderson *et al.*, 1981) and the discovery of the first pathogenic mutation in 1988 (Holt *et al*., 1988; Wallace *et al.*, 1988), which made scientists aware of the important role of the mitochondria in studies of human disease. Upon the discovery that human mtDNA consisted of circular double-stranded molecules and harboured specific characteristics that are valuable in the investigation of disease and evolution, a period of in-depth investigation of mitochondrial genome variation followed in the 1990's, which was widely regarded as the decade of the mitochondria (Scheffler, 2001). It was during this period that the power of the genetic diversity of this uniparentally inherited marker was discovered in terms of the investigation of human evolution based on its unique features, such as high copy number, uniparental mode of inheritance, lack of recombination and high mutation rate (Cann *et al.*, 1984; Ingman *et al*., 2000).

3.1 HISTORY AND DEVELOPMENT

Mitochondria came into existence when, according to the serial endosymbiont theory, a proteobacterium was encapsulated by endocytosis by a protoeukaryotic cell about two billion years ago (Wallace, 2007). It is speculated that the aforementioned proteobacterium originated from the rickettsia subdivision of the α-proteobacteria, *Rickettsia, Anaplasma* and *Ehrlichia,* and that it was respiration-competent. The protoeukaryotic cell, which was the precursor to eukaryotic cells and originated from the *Archeozoa* amitochondriate eukaryotes, existed without mitochondria and was therefore an anaerobic archeobacteria host to the respiration-competent symbiont. The protoeukaryotic nucleus-cytosol had limited energy, which it obtained from the symbiosis with the probacterium (Gray *et al.*, 1999). The dependency of the nucleus-cytosol on this energy resulted in the regulation of nuclear replication and gene expression according to the availability of energy, and also to the regulation of growth and replication of the probacterium (Wallace, 2007).

After a symbiotic relationship between the host and symbiont had been established and maintained for 1,2 billion years, redundant genes were lost and some genes transferred to the nucleus of the host (Gray *et al*., 1999). The nucleus specialised in specifying structure in the cell and the mitochondria specialised in energy production (Wallace and Fan, 2010). There are different hypotheses about why not all of the genes of the mitochondria were transferred to the nucleus. One possible reason is that the mitochondrial genes that remained in the mitochondrial genome became divergent and could not be interpreted by the nuclear-cytosolic system. Another possibility is that some of the proteins of the OXPHOS needed to remain in the mitochondria in order to ensure that the OXPHOS complexes were constructed on the inner membrane of the mitochondria. Relocation of the OXPHOS system to other parts of the cell would have resulted in exposure to reactive oxygen species (ROS) and would have destroyed the cell. The third hypothesis is that the remaining mitochondrially encoded proteins were too hydrophobic to be translated to the cytosol and were thus directly incorporated into the inner membrane of the mitochondria (Wallace, 2007).

3.2 MITOCHONDRIAL STRUCTURE AND MORPHOLOGY

Mitochondria are essential for the survival of the cell and are therefore crucial to life. Eukaryotic cells generally contain many mitochondria that move, fuse and divide and can occupy as much as 25% of the cytoplasm of the cell (Bereiter-Hahn and Vöth, 1994). The basic morphology of the mitochondria consists of an inner membrane that encloses a matrix and an outer membrane, with an inter-membrane space between the two membranes. The existence of two membranes creates an inter-membrane compartment between the outer membrane and the inner membrane and another inner compartment that lies within the inner membrane. The inner membrane consists of folds (cristae) that increase the surface of the membrane and that contain an intracristal space within the cristae, which is in contact with the inter-membrane compartment through broad openings (Palde, 1953; Frey *et al.*, 2002). Mitochondria display dynamic behaviour in the cell, which is attributed to the interactions between components of the cytoskeleton such as actin filaments, microtubules and intermediate filaments with outer membrane proteins (Frey *et al*., 2002), as well as displaying a large matrix *in situ*, which presses the inner membrane against the outer membrane and thereby decreases the inter-membrane space. The same is observed of the cristae membranes that are pushed together to allow a small intracristal space (Frey *et al.*, 2002). The five enzyme complexes of the OXPHOS system are embedded in the inner membrane of the mitochondrion. The mitochondrion

further contains its own genome, which is located in the matrix and is present in several identical copies in each mitochondrion.

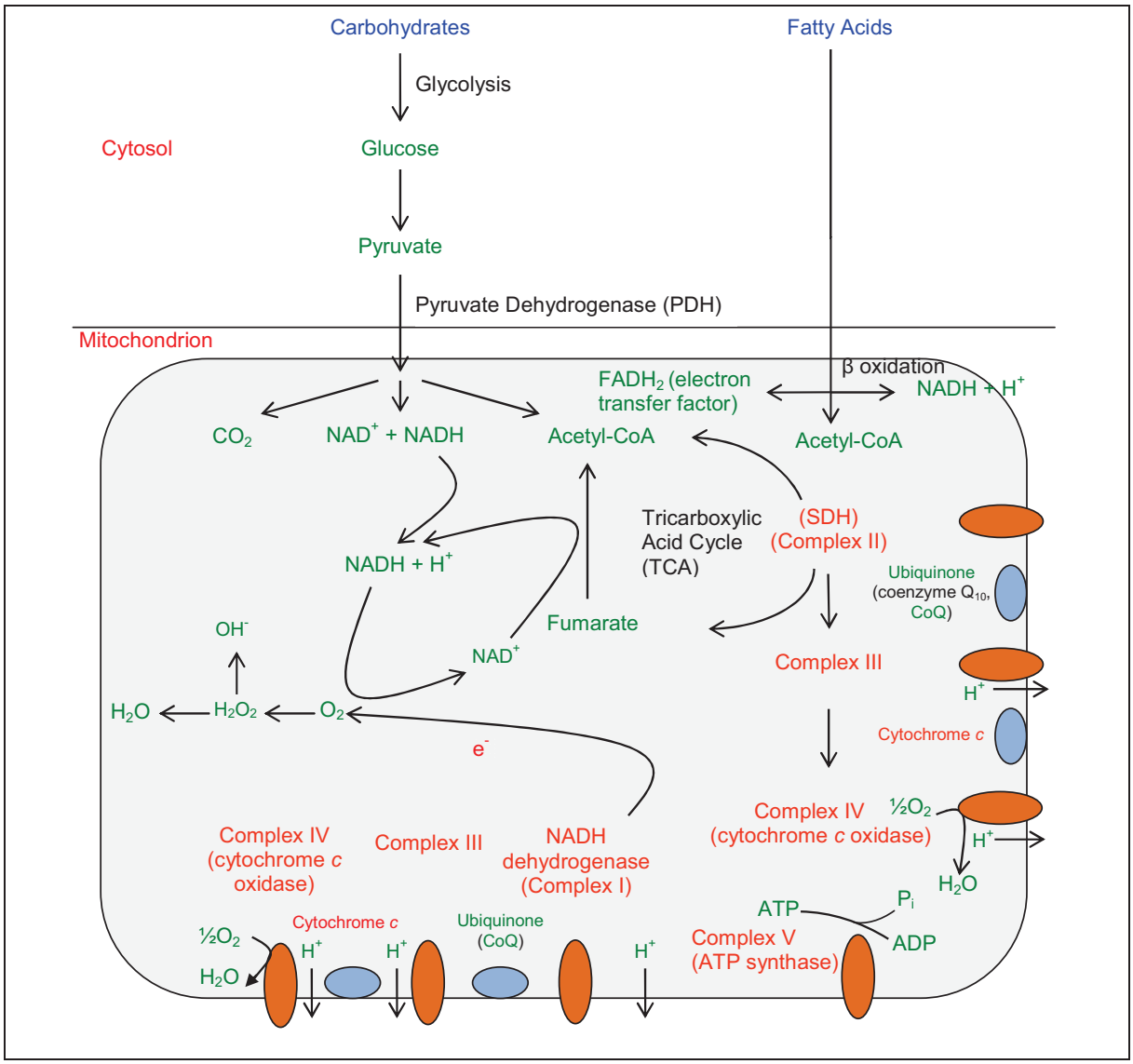
3.3 MITOCHONDRIAL FUNCTION

The mitochondrion is fundamentally involved in cellular energy metabolism because of the role it plays in ATP production through the respiratory chain, fatty acid β oxidation and the urea cycle. Mitochondria are further involved in maintaining a constant cell environment and play a role in intracellular signalling and cell death, which make them an important organelle in terms of developmental biology and cancer aetiology. They also play a role in the metabolism of amino acids, lipids, cholesterol, steroids and nucleotides (Chinnery and Schon, 2003).

The most important function of the mitochondria is, however, their involvement with ATP production. Mitochondria provide energy in the form of ATP to most of the cells through the OXPHOS system, which transfers electrons obtained from the calories of the diet down the electron transport chain, which consists of a number of redox enzyme complexes that are present in the inner membrane of the mitochondria to produce ATP. ATP is, in turn, used for work or transformed to generate heat for thermal regulation (Wallace, 2007).

The OXPHOS system consists of two subsystems i.e. the electron transport chain (ETC) and the ATP synthase complex, that are driven by five enzyme complexes from the inner membrane of the mitochondria. The ETC consists of complexes I-IV, while ATP synthase consists of complex V. Hydrogen from carbohydrates and fats are oxidised by oxygen and water is formed as a by-product in the ETC reaction, as presented in Figure 3.1. The tricarboxylic acid cycle collects hydrogen from carbohydrates and organic acids or from fats through the β–oxidation pathway and transfers it to a carrier, oxidised nicotinamide adenine dinucleotide (NAD^+) molecule, which donates the electrons to complex I (NADH dehydrogenase) to form reduced nicotinamide adenine dinucleotide (NADH) and a hydrogen ion (H^+). Succinate is formed, which is used to pump protons over the membrane and by this mechanism generate ATP. This is achieved by driving complex V (ATP synthase) by the stored energy, to convert adenosine diphosphate (ADP) and orthophosphate (P_i) to form ATP (Hägerhäll, 1997; Wallace *et al.*, 1999; Wallace, 2007).

Figure 3.1 OXPHOS system

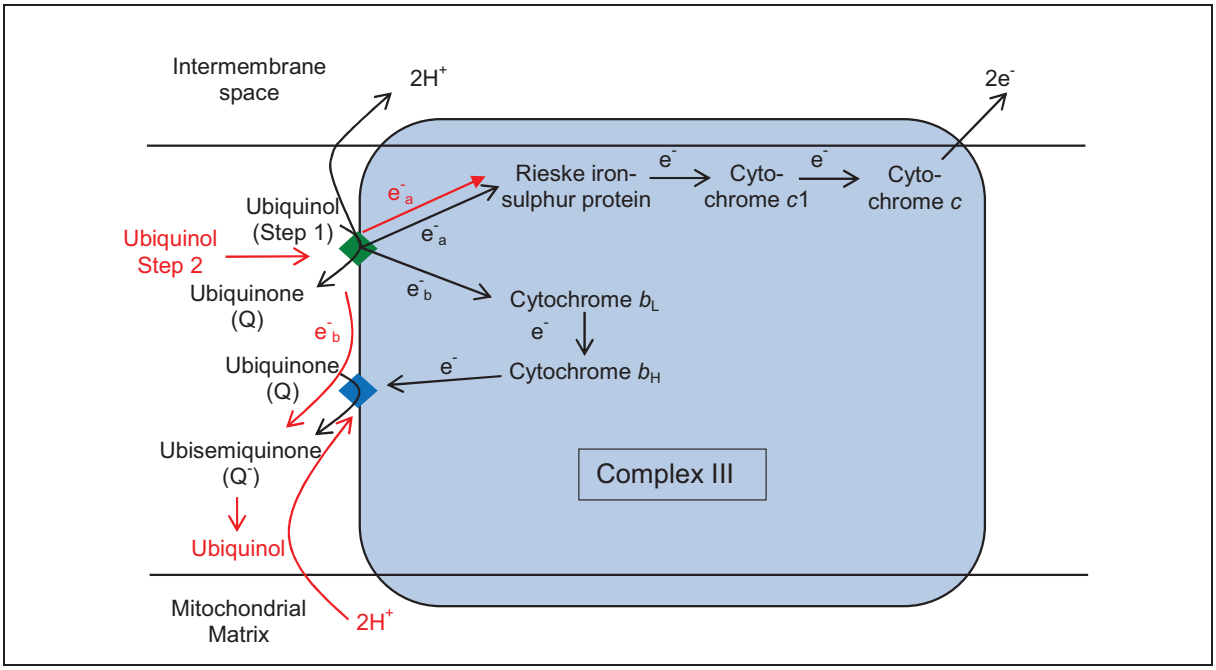


Oxidative phosphorylation (OXPHOS) consisting of complex I to V that constitutes two (2) subsystems, the ETC and the ATP synthase. Complex I consists of 45 polypeptides of which NADH dehydrogenase subunit (ND)1, -2, -3, -4L, -4, -5, and -6 are encoded by the mtDNA. Complex II consists of four (4) nDNA encoded polypeptides. Complex III consists of 11 polypeptides of which cytochrome b is encoded by the mtDNA. Complex IV consists of 13 polypeptides of which cytochrome c oxidase subunits (CO)I, COII and COIII are encoded by the mtDNA. Complex V consists of 16 polypeptides of which ATP synthase F0 subunit 6 (ATP6) and ATP synthase F0 subunit 8 (ATP8) are encoded by mtDNA. Abbreviations: Acetyl-CoA = acetyl-coenzyme A; ADP or ATP = adenosine di- or triphosphate; CO_2 = carbon dioxide; cytc = Cytochrome c; $FADH_2$ = Flavin Adenine Dinucleotide; H^+ = hydrogen ion; H_2O = water; NAD^+ = nicotinamide adenine dinucleotide; NADH = reduced nicotinamide adenine dinucleotide; SDH = succinate dehydrogenase; e^- = electron. From Knaff, 1993; Hägerhäll, 1997; Wallace *et al.*, 1999; Scheffler, 2001; Wallace, 2007.

The functions of the polypeptide subunits of complex I are not fully known. Disease-causing mutations have been observed in the subunits, indicating that the functions of these subunits are indeed critical to proper mitochondrial function. It is speculated that the subunits contribute to a Q cycle within complex I that is involved with pumping protons from the inner membrane to the inter-membrane space (Wallace, 2007). In contrast, complex II is not involved with proton pumping and is the simplest of all the complexes (Scheffler, 2001).

Complex III has been resolved and it is known that this complex is involved with the pumping of protons to and from the inter-membrane space. One polypeptide out of the 11 polypeptides of Complex III (Knaff, 1993), namely cytochrome b (Cytb), is encoded for by a mitochondrial gene. Protons are pumped from the inner membrane to the inter-membrane space via the inner membrane Q cycle as presented in Figure 3.2. In the Q cycle, the reduced ubiquinol binds to Cytb at the coenzyme Q binding site and transfers one electron via the Rieske iron-sulphur protein to cytochrome *c*. A second electron is transferred to *Cytb* where it reduces ubiqinone, which is bound to the coenzyme Q_{10} binding site and reduces the ubiquinone to ubisemiquinone. Coenzyme Q loses two electrons and consequently releases two protons into the inter-membrane space. The ubiquinone is replaced with another ubiquinol and passes one electron to the Rieske iron-sulphur protein and one electron to the ubisemiquinone, reducing it to ubiquinol. Because of the negative charges on the coenzyme Q, two protons are absorbed from the matrix (Mitchell, 1975; Wallace, 2007).

Figure 3.2 Q cycle



Proton pumping via the Q cycle and Complex III, here presented in blue. The Q cycle takes place in two (2) steps. Step 1 is denoted in black and step 2 is denoted in red. Reduced ubiquinol binds to cytochrome *b* at the coenzyme Q binding sites: Coenzyme Q-binding site on the outside of the inner membrane adjacent to the inter-membrane space = ◆; CoenzymeQ_{10} binding site on the inside of the inner membrane adjacent to the matrix = ◆; Ubiquinol Step 1 = the ubiquinol that takes part in the first step of the Q cycle; e^-_a = first electron transferred; electron transferred via the Rieske iron-sulphur protein to Cytochrome *c;* e^-_b = second electron transferred to cytochrome *b* thus reducing the ubiqinone bound to the coenzyme Q_{10} binding site to ubisemiquinone; electrons and protons released into the inter-membrane space; Step 2 = ubiquinone at ◆ replaced with another ubiquinol; e^-_a= first electron transferred on to the Rieske iron-sulphur protein; e^-_b = second electron passed on to the ubisemiquinone; reduced to ubiquinol and taking up two protons from the mitochondrial matrix. Adapted from Nicholls and Ferguson, 2002.

Complex IV has also been resolved. Cytochrome *c* oxidase subunit I (COI), Cytochrome *c* oxidase subunit II (COII) and Cytochrome *c* oxidase subunit III (COIII) polypeptides, which

link the ETC with proton pumping, are all encoded by the mtDNA. Cytochromes *a* and a_3 and the copper B centre (Cu_B) are nested in the COI protein where they form a trinuclear reaction centre. Oxygen binds to this centre and is reduced to form H_2O. Electrons are transferred from cytochrome *c* to COII to COI. The function of COII is not understood precisely, but it is believed that it forms an aqueous channel that allows protons to move through the membrane (Wallace, 2007).

The F_1 subunit of the adenosine tri-phosphate synthase (F_1 ATPase) of complex V protrudes into the matrix with a membranous base i.e. the F_0 component. A proton channel is formed by the mitochondrially encoded ATP synthase F0 subunit 6 (ATP6) polypeptide coupling the proton gradient to ATP synthase. This proton channel is formed by a complex interaction of ATP synthase F0 subunit 9 (ATP9) "spokes" forming a rotor-like wheel structure attached to an axis consisting of epsilon- (ε) and gamma- (Ɣ) subunits, which project into the stalk-like ATP F_1 structure that is linked to the ATP6 globular structure that contains two half-proton channels (Elston *et al.*, 1998). The ATP9 "spokes" contain negatively charged amino acid groups, which interact with the half-proton channel that is open to the inter-membrane space to pick up a proton. The "spoke" rotates and returns to interact with the other half-proton channel that is open to the matrix where it gives off the acquired proton. This rotation makes the ε, Ɣ-axis spin in the F_1 barrel, inducing the condensation of ADP + P_i to ATP (Abrahams *et al.*, 1994; Elston *et al.*, 1998; Stock *et al.*, 1999; Wallace, 2007).

The mitochondria have the ability to generate ROS. Superoxide anions (O_2^-) are generated because of an electron from either complex I or III being transferred to an O_2 molecule. Mn superoxide dismutase generates hydrogen peroxide (H_2O_2) from two O_2^- molecules, which diffuses out of the mitochondria where it is degraded in the cytosol. Highly reactive hydroxyl radicals are, however, produced if H_2O_2 acquires another electron and will damage mitochondrial proteins and lipids. When the damage to the mitochondria reaches the point where energy production is significantly affected, the permeability transition pore is activated and cell death is triggered (Wallace, 2007).

The OXPHOS system in the mitochondrion is therefore primarily involved in the production of energy that is critical to the survival of the cell and ultimately the organism. The OXPHOS converts dietary calories into energy through an oxidation process to pump protons across the inner membrane of the mitochondrion through the complexes I, II and IV to create a trans-membrane electrochemical gradient. This process is initiated and

conducted by the ETC and complexes I-IV of the mitochondria. The purposes of the trans-membrane electrochemical gradient are to: 1.) create a source of potential energy to synthesise ATP via the ATP synthase complex V, 2.) generate heat, 3.) transport proteins or ions for the production of ROS and 4.) regulate cell growth and death.

3.4 MITOCHONDRIAL DNA

Human mitochondria contain an autonomously replicating DNA genome, which consists of 16,569 nucleotide pairs (np) in a closed double stranded circular structure (Anderson *et al.*, 1981). The genome consists of a large component of DNA that encodes for mitochondrial proteins and a 1,121 bp non-coding D-loop or control region. The circular structure consists of a light strand (L), that is rich in C nucleotides and is generally regarded as the main coding strand, and a heavy strand (H) that is regarded as the anti-sense strand, which is rich in G nucleotides and also the strand from which the RNAs are transcribed. Replication and transcription start at an origin of replication within the heavy and the light strand (O_H and O_L) and use a replication promotor for the heavy strand (P_H) and replication promotor for the light strand (P_L). The P_H and P_L are located near the O_H in the non-coding region of the mitochondrial DNA (Anderson *et al.*, 1981; Wallace, 1994).

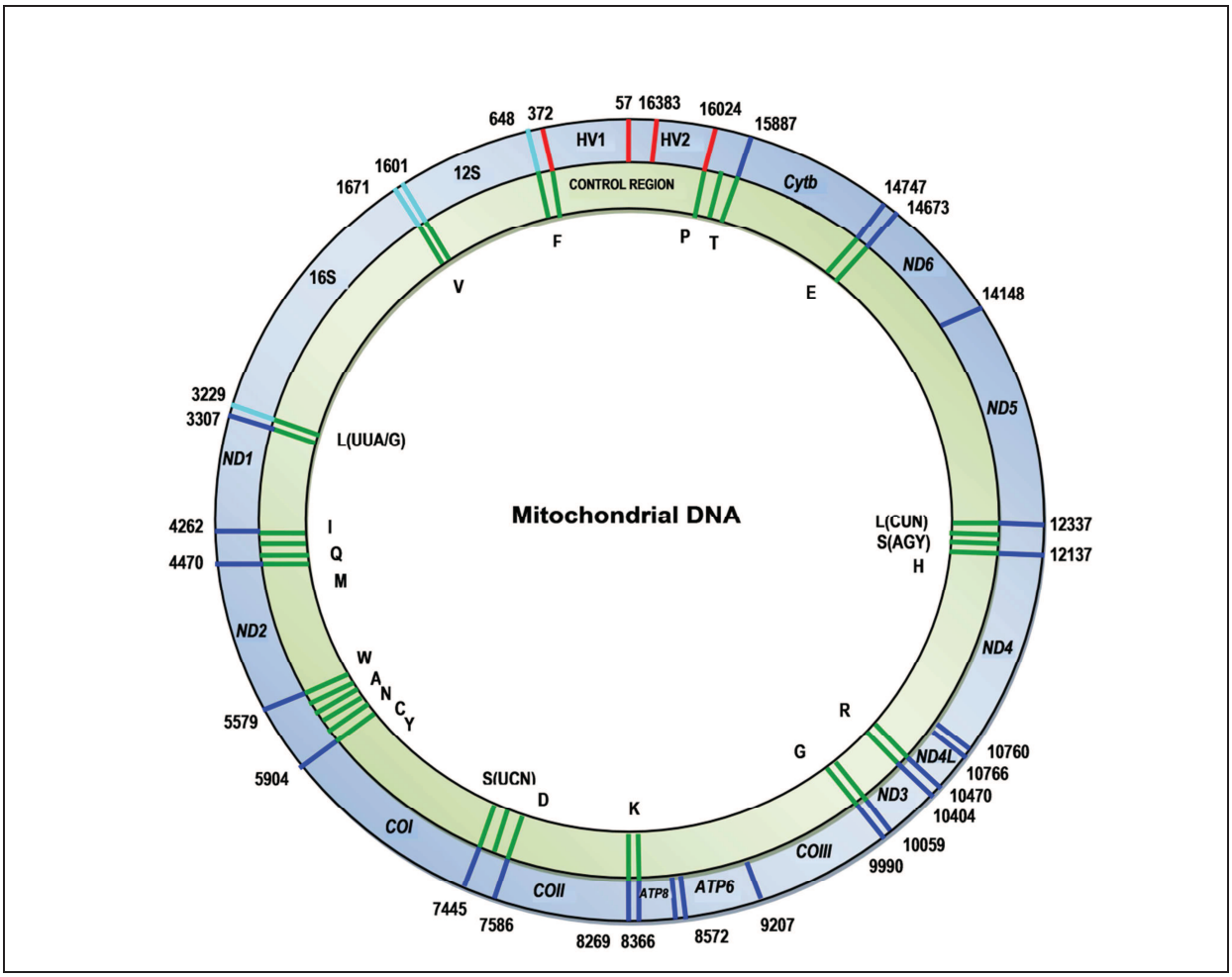
The control region consists of the P_L, P_H, O_H and the mitochondrial transcription factor A binding sites, as well as three (3) conserved blocks of sequence and the termination-associated sequences. The control region is also referred to as the D-loop because of an extra newly formed DNA fragment at the H strand origin of replication. The P_H transcribes all the genes except NADH dehydrogenase subunit 6 gene (*ND6*) and some of the tRNAs, which are transcribed by P_L. Replication starts at the 3'OH cleaved L strand in the control region and replicates to form a new H strand more than two-thirds of the way round until it exposes the O_L. The L strand replication starts at that point and moves back around the displaced H strand (Clayton, 1991).

It is believed that a strong evolutionary force was responsible for driving the transfer of mitochondrial genes from the mitochondrion to the nuclear genome (Wallace, 2007). The mitochondrial proteins present in the mitochondrion are therefore encoded by both mtDNA and nuclear DNA (nDNA) and synthesised by a separate mitochondrial translation system. The expression of the 1,500 nDNA genes involved in the regulation of mitochondrial

functions are regulated by the needs and availability of calories in the cell (Wallace *et al.*, 1999).

MtDNA contains the genes that encode for 13 polypeptides that define the efficiency of the mitochondrial energy-generating OXPHOS system and form part of the five protein complexes of the OXPHOS system that are associated with the mitochondrial inner membrane. These polypeptides include seven of the polypeptides of OXPHOS complex I i.e. NADH dehydrogenase subunit 1 (*ND1*), NADH dehydrogenase subunit 2 (*ND2*), NADH dehydrogenase subunit 3 (*ND3*), NADH dehydrogenase subunit 4 (*ND4*), NADH dehydrogenase subunit 4L (*ND4L*), NADH dehydrogenase subunit 5 (*ND5*) and NADH dehydrogenase subunit 6 (*ND6*), one of the polypeptides of OXPHOS complex III i.e. *Cytb*, three of the polypeptides of OXPHOS complex IV i.e. cytochrome *c* oxidase subunit 1 (*COI*), cytochrome *c* oxidase subunit 2 (*COII*) and cytochrome *c* oxidase subunit 3 (*COIII*) and two polypeptides of the OXPHOS complex V i.e. *ATP6* and *ATP8*. In addition to the polypeptide coding genes, the mtDNA also encodes two ribosomal RNAs (rRNAs) and 22 transfer RNAs (tRNAs); the tRNAs are used to translate the 13 polypeptides that are encoded by the mtDNA (Anderson *et al.*, 1981; Wallace, 1995). The organisation of the human mitochondrial genome is presented in Figure 3.3.

Figure 3.3 Functional organisation of the human mitochondrial DNA



Blue = Coding region. Light blue = rRNA. Green = tRNA. Circular green band represents the L strand, blue band the H strand. Numbers refer to base pair positions relative to the rCRS (Anderson *et al*., 1981). HV2: hypervariable segment 2; 12S: 12S ribosomal RNA; 16S: 16S ribosomal RNA; *ND1*: NADH dehydrogenase subunit 1 gene; *COI*: cytochrome c oxidase subunit I gene; *COII*: cytochrome c oxidase subunit II gene; *ATP8*: ATP synthase F0 subunit 8 gene; *ATP6*: ATP synthase F0 subunit 6 gene; *COIII*: Cytochrome c oxidase subunit III gene; *ND2*: NADH dehydrogenase subunit 2 gene; *ND3*: NADH dehydrogenase subunit 3 gene; *ND4L*: NADH dehydrogenase subunit 4L gene; *ND4*: NADH dehydrogenase subunit 4 gene; *ND5*: NADH dehydrogenase subunit 5 gene; *ND6*: NADH dehydrogenase subunit 6 gene; *Cytb*: cytochrome b gene; Control region, including displacement loop; HV1: hypervariable segment 1; F: tRNA phenylalanine; V: tRNA valine; L(UUA/G): tRNA leucine 1; I: tRNA isoleucine; Q: tRNA glutamine; M: tRNA methionine; W: tRNA tryptophan; A: tRNA alanine; N: tRNA asparagine; C: tRNA cysteine; Y: tRNA tyrosine; S: tRNA serine 1; D: tRNA aspartic acid; K: tRNA lysine; G: tRNA glycine; R: tRNA arginine; H: tRNA histidine; S(UCN): tRNA serine 2; L(CUN): tRNA leucine 2; E: tRNA glutamic acid; T: tRNA threonine; P: tRNA proline. Adapted from MITOMAP: A Human Mitochondrial Genome Database. http://www.mitomap.org, 2011. Accessed 16 Feb 2011.

Anderson *et al*. (1981) published the first complete sequence of the human mitochondrion in which the location of the different genes were identified and numbered according to a standardised numbering system. Andrews *et al.* (1999) published a revision of this mtDNA sequence, referred to as the rCRS, in which some corrections were made and which was subsequently used by scientists as a standard reference sequence of the human mitochondrial DNA. The nucleotide positions of the 13 protein-coding genes, 2 rRNA genes and 22 tRNA genes are presented in Table 3.1 according to the rCRS location and numbering system.

Table 3.1 Functional organisation of human mitochondrial DNA

Locus code	Locus name	Sequence position
CR/ D loop	Control region / D-loop	16024 – 576
HV1	Hypervariable segment 1	16024 – 16383
HV2	Hypervariable segment 2	57 – 372
F	tRNA phenylalanine	577 – 647
12S	12S ribosomal RNA	648 – 1601
V	tRNA valine	1602 – 1670
16S	16S ribosomal RNA	1671 – 3229
L(UUA/G)	tRNA leucine 1	3230 – 3304
ND1	NADH dehydrogenase subunit 1	3307 – 4262
I	tRNA isoleucine	4263 – 4331
Q	tRNA glutamine	4329 – 4400
M	tRNA methionine	4402 - 4469
ND2	NADH dehydrogenase subunit 2	4470 – 5579
W	tRNA tryptophan	5512 – 5579
A	tRNA alanine	5587 – 5655
N	tRNA asparagine	5657 – 5729
C	tRNA cysteine	5761 – 5826
Y	tRNA tyrosine	5826 – 5891
COI	Cytochrome c oxidase subunit 1	5904 – 7445
S(UCN)	tRNA serine 1	7446 – 7514
D	tRNA aspartic acid	7518 – 7585
COII	Cytochrome c oxidase subunit 2	7586 – 8269
K	tRNA lysine	8259 – 8364
ATP8	ATP synthase F0 subunit 8	8366 – 8572
ATP6	ATP synthase F0 subunit 6	8527 – 9207
COIII	Cytochrome c oxidase subunit 3	9207 – 9990
G	tRNA glycine	9991 – 10058
ND3	NADH dehydrogenase subunit 3	10059 – 10404
R	tRNA arginine	10405 – 10469
ND4L	NADH dehydrogenase subunit 4L	10470 – 10766
ND4	NADH dehydrogenase subunit 4	10760 – 12137
H	tRNA histidine	12138 – 12206
S(AGY)	tRNA serine 2	12207 – 12265
L(CUN)	tRNA leucine 2	12266 – 12336
ND5	NADH dehydrogenase subunit 5	12337 – 14148
ND6	NADH dehydrogenase subunit 6	14149 – 14673
E	tRNA glutamic acid	14674 – 14742
Cytb	Cytochrome b	14747 – 15887
T	tRNA threonine	15888 – 15953
P	tRNA proline	15956 – 16023

CR = Control region / D-loop here refers to the non-coding region between positions 16024 – 576. Locus codes and names are the same as used in Figure 3.3 as reported in the MITOMAP database, www.mitomap.org ; 12S: 12S ribosomal RNA; 16S: 16S ribosomal RNA; *ND1*: NADH dehydrogenase subunit 1; *COI*: Cytochrome c oxidase subunit I; COII: Cytochrome c oxidase subunit II; *ATP8*: ATP synthase F0 subunit 8; *ATP6*: ATP synthase F0 subunit 6; *COIII*: Cytochrome c oxidase subunit III; *ND2*: NADH dehydrogenase subunit 2; *ND3*: NADH dehydrogenase subunit 3; *ND4L*: NADH dehydrogenase subunit 4L; *ND4*: NADH dehydrogenase subunit 4; *ND5*: NADH dehydrogenase subunit 5; *ND6*: NADH dehydrogenase subunit 6; *Cytb*: Cytochrome b; Control region, including displacement loop; HV1: Hypervariable segment 1; F: tRNA phenylalanine; V: tRNA valine; L(UUA/G): tRNA leucine 1; I: tRNA isoleucine; Q: tRNA glutamine; M: tRNA methionine; W: tRNA tryptophan; A: tRNA alanine; N: tRNA asparagine; C: tRNA cysteine; Y: tRNA tyrosine; S: tRNA serine 1; D: tRNA aspartic acid; K: tRNA lysine; G: tRNA glycine; R: tRNA arginine; H: tRNA histidine; S(AGY): tRNA serine2; L(CUN): tRNA leucine 2; E: tRNA glutamic acid; T: tRNA threonine; P: tRNA proline.. Sequence positions correspond to the rCRS positions (Andrews *et al*., 1999).

Mitochondrial DNA has the further virtue of being omnipresent in animals and humans and being mostly uniform in gene content. A study of multicellular organisms and protozoans demonstrated that the mtDNA composition was similar across species, making the mitochondrial genome an ideal marker for genetic studies between species (Wilson *et al.*, 1985).

3.4.1 Mitonuclear interactions

Stern and Lonsdale (1982) reported that genetic information has been transferred between organelles and the nucleus since the initial development of species and that this is also true for the transfer of genes between the mitochondria and the nucleus. A review of the genomes of different species revealed that strong selective pressure exists for the transfer of mtDNA genes to the nDNA, which stops when the mtDNA contains a full set of inner membrane polypeptide genes, as in the case of the human mtDNA that contains genes for 13 polypeptides of the inner membrane OXPHOS complexes. It is further also believed that the genes that were retained in the mtDNA were involved in coding for proteins that played a role in inner membrane proton translocations (Wallace, 2007). More than 1,500 genes from the mtDNA have been transferred to the nDNA with the added implication that the nuclear encoded mitochondrial proteins have to be reimported into the mitochondrion for purposes of mitochondrial metabolism. These proteins consist of a group of proteins that influence the structure of the ETC, a group of proteins that influence the copy number of the mtDNA and a group that is involved with the respiratory complexes of the mitochondria. Studies have provided evidence that mitochondrial evolution proceeds in parallel with nuclear evolution and that the mitonuclear interactions entail complex interaction between proteins encoded by the mitochondrial genome and proteins encoded by the nuclear genome, which are both critical for the efficient functioning of the OXPHOS system (Ballard and Rand, 2005).

Some of the gene transitions from the mitochondrion to the nucleus consist of nuclear pseudogenes of mitochondrial origin (numts), which represent an integration of functionless mitochondrial DNA segments into the nDNA. This process has also been going on for thousands of years (Mourier *et al.*, 2001). Although the numts are not recognised by the nucleus and cannot be deciphered by it, it can become problematic when PCR primers designed to amplify mitochondrial genes recognise the PCR binding sites of the numts and erroneously amplify the incorrect DNA fragments (Ruiz-Pesini *et al.,* 2007). The numts are exposed to a much lower rate of mutation and can therefore be

regarded as the fossils of the extant mitochondria and are consequently not always an accurate representation of the real mitochondrial sequence. For this reason the detection of numts has led to inaccurate phylogenetic and evolutionary conclusions (Pakendorf and Stoneking, 2005).

3.5 UNIQUE CHARACTERISTICS OF HUMAN MITOCHONDRIAL DNA

Avise *et al*. (1987) observed that there was a deficiency of communication and inter-relatedness between the fields of population genetics and systematics, which was referred to as a gap between microevolution and macroevolution. They realised the importance of the connection between the fields of phylogenetics and its implied systematics outcomes and the more detailed fields of population genetics, which covered aspects such as genetic drift and natural selection in populations over time. Both fields were driven by evolution and could therefore be regarded as complementary to each other. With the development of technology in the field of molecular biology over the last few decades, it became possible to sequence whole genomes. As a result, it also became possible to use these high-resolution DNA markers to form a connection between phylogenetics and population genetics. DNA markers that would be ideal for use in such endeavours would have to be easy to isolate and assay, be omnipresent in a wide variety of organisms, be deficient of complicated genetic structures, be transmitted in a simple manner from one generation to the next, possess character states that demonstrate phylogenetic relationships and display a high mutation rate that would be sufficient to result in high levels of genetic variability that could be used to distinguish between and within species (Avise *et al.*, 1987). It was subsequently demonstrated that the human mitochondrial genome displays unique properties that address most of these needs.

3.5.1 Copy number of mitochondrial DNA

One of the most important characteristics of mitochondrial DNA is the fact that it is present in high numbers within the cell and has an advantage in this regard over the two copies of nuclear genes per somatic cell. Its extranuclear location in the cell also makes it readily available for use. Thus, there is much more DNA readily available to work with, which is why mitochondrial DNA is preferred when investigating ancient remains and why disciplines such as forensic science are using mitochondrial DNA to identify victims of mass disasters (Pakendorf and Stoneking, 2005).

3.5.2 Mutation rate of mitochondrial DNA

The environment can change rapidly and species need to adapt to changes quickly and efficiently in order to ensure survival. In humans, this is achieved by the presence of a broad range of energy-producing genetic solutions in the form of different mitochondrial genotypes. In the absence of recombination to introduce genetic variability, the mitochondria have adapted by displaying a high rate of mutation, ensuring high genetic variability and a better chance of survival (Wallace, 1994). A high mutation rate also implies a high rate of evolution according to the basic equation of evolution, which describes evolution as the product of the rate of mutation and the rate of fixation of mutations in the germline (Wilson *et al.*, 1985). The high mutation rate of the mitochondrial genome has resulted in human populations that contain high levels of population-specific polymorphisms, which are well suited to the study of human evolution.

The mutation rate of the human mitochondrial genome is higher than the mutation rate of human nuclear DNA and this has been ascribed to the mitochondria's inability to repair DNA replication errors and DNA damage effectively (Brown *et al.*, 1979). The mitochondrial genome has demonstrated a high level of tolerance for inaccuracy in the process of DNA replication, most probably because it does not encode proteins that are directly involved in the translation and transcription of its own DNA replication process (Cann *et al.*, 1984). The high mutation rate is further attributed to the lack of protective histones and the high number of oxygen radicals generated by the OXPHOS system. In addition to the high mutation rate, the mutations are also easily fixed owing to the maternal germ line sorting of the mutant molecules followed by rapid genetic drift. The sorting takes place through replicative segregation when the mutant and the normal mitochondrial DNA is sorted into daughter cells to shift the distribution of the sequence variant to homoplasmy through either losing the mutation from the germ line population or through fixing the mutation (Jenuth *et al.*, 1997; Wallace, 2008).

The mutation rate of the coding region of the mitochondrial DNA in general is estimated at 0.017×10^{-6} substitutions per site per year (Ingman *et al.*, 2000). A molecular clock with an average rate of 1.26×10^{-8} nucleotide substitutions per site per year or in other terms, a rate of 5.138 mutations per year for the mitochondrial coding region, is also widely used in coalescence time estimates (Mishmar *et al.*, 2003; Behar *et al.*, 2008). The mutational rate of the control region of the mitochondria, however, is much higher than the mutational rate of the coding region. There is controversy over the mutational rate of this highly variable

region of the mitochondria and two different estimates are currently used in population studies, namely a value of 0.075-0.165 X 10^{-6} substitutions per site per year based on phylogenetic studies (Stoneking *et al.*, 1992; Hasegawa *et al.*, 1993; Tamura and Nei, 1993) or a value of 0.47 X 10^{-6} substitutions per site per year based on studies of pedigrees (Howell *et al.*, 2003). Although controversy still exists over which method to follow for the calculation of the mutation rate (Pakendorf and Stoneking, 2005), it holds true that the control region displays a much higher rate of mutation in comparison to the coding region of the mitochondrial genome, which is explained by the removal of deleterious mutations from the coding regions through the process of natural selection (Howell *et al*., 2003).

Furthermore, there is also some controversy about whether the mutation rate is variable between specific sites within the same mitochondrial genome. Some scientists have reported that the rate of mutation does not display a Poisson distribution as was expected under the assumption of a constant rate of mutation (Hasegawa *et al.*, 1993; Wakeley, 1993) and rather displays a gamma distribution indicative of a site-to-site variation in mutation rate (Aris-Brosou and Excoffier, 1996). This phenomenon results in regions or specific sites of the mitochondrial genome that display high mutation rates and others that are stable. The mutation hotspots, such as nucleotide positions 146, 150, 152, 195, 16189, 16311, 16362 and 16519 of the control region and nucleotide positions 709, 1719, 3010, 5460, 10398, 11914, 13105, 13708, 15884 of the coding regions, present with a greater probability of harbouring homoplasies i.e. where a mutational event occurs more than once at a specific position and therefore obscures the true number of mutations that occur at that position. Reasons for this hypervariability have not been resolved yet (Howell *et al.*, 2003; Kivisild *et al.*, 2006; Galtier *et al.*, 2008).

3.5.3 Maternal inheritance

The mammalian egg contains about 100,000 mitochondria in contrast to the sperm, which contains only about 100 mitochondria (Chen *et al.*, 1995b; Reynier *et al.*, 2001). During reproduction, the mitochondria of the sperm are destroyed by ubiquitination by the oocyte, making the inheritance of mitochondrial DNA maternal (Giles *et al.*, 1980; Sutovsky *et al.*, 2000). Through this mechanism of uniparental inheritance, a germ line bottleneck is introduced, which effectively limits the number of mtDNAs that are transmitted from one generation to the next. This results in the rapid removal of mtDNA mutations from the gene

pool through the process of genetic drift, which will act more strongly within the smaller population of inherited mtDNA molecules.

It is believed that the proteobacterium and protoeukaryotic symbiont form of the developing mitochondrion undergo selective pressure through nuclear mutations, which restricts the transmission of organelle genomes that contain deleterious mutations acquired by biparental inheritance. This results in the mtDNA being inherited uniparentally to limit this deleterious effect on the survival of the organism (Hoekstra, 2000). The genes of the mtDNA are all critical to the OXPHOS system and could therefore affect life and health when a deleterious mutation occurs. The mode of inheritance through a single parent therefore limits the adverse effects to the organism and is also the reason for the strong selective pressure that removes the critical gene functions of the early symbiont to the nuclear DNA to protect it from mutational decay through sexual reproduction and recombination (Felsenstein, 1974).

The proteins that are encoded by the mtDNA are all involved in the OXPHOS system either as electron or proton carriers, and interact in the mechanisms that underlie the electrochemical gradient established over the inner membrane of the mitochondrion. The genes of the other proteins involved in mitochondrial functioning are transferred to the nucleus and the genes that orchestrate the energy-producing circuit in the mitochondrion remain in the mtDNA. Mutations in the mtDNA genes would therefore affect the whole energy circuit and also the other polypeptides in the mitochondrion to create a new metabolic strategy over many generations to enable coping with environmental changes. The uniparental inheritance of the mitochondrion prevents the mixing of different mitochondrial lineages and thus different sets of polymorphisms. Should this happen, compatible genetic changes within each of the mitochondrial lineages would mix and most probably cause an incompatible combination of energy metabolism regulating polymorphisms, which could result in the death of the cell or ultimately the organism. It is further hypothesised that the mitochondrial DNA is inherited uniparentally to conserve the combinations of mitochondrial polymorphisms that enable an organism to adapt and survive in changing environmental conditions (Wallace, 2008).

The maternal inheritance of the mitochondrial genome makes it ideal for the study of maternal ancestry because of the direct inheritance of sequence variants from generation to generation without the confounding effects of recombination. Mitochondrial DNA is thus

widely used in phylogenetic studies of human evolution because the lineages can be traced back to a single maternal ancestor (Pakendorf and Stoneking, 2005).

3.5.4 Lack of recombination

Mitochondrial myopathies have demonstrated that paternal inheritance has occurred and this has raised some serious concerns about the validity of the hypothesis of strict maternal inheritance (Bromham *et al.*, 2003). Further studies, however, have reported that this phenomenon is highly unlikely and that the incidences of paternal mitochondrial inheritance were sufficiently rare to accept the theory of maternal inheritance of the human mitochondrial genome (Hazkani-Covo *et al*., 2010). The presumed maternal inheritance of the human mitochondrial genome includes the assumption that the inheritance of mtDNA is clonal and that mtDNA variation accumulates in the lineages upon divergence from a common maternal ancestor. It is widely accepted that the human mitochondrial genome displays a lack of recombination (Stoneking *et al*., 1992; Stoneking and Soodyall, 1996). This assumption was challenged when evidence of recombination in the mitochondrial genomes of yeast and some animal species was observed (Thyagarajan *et al.*, 1996) and was followed by more studies that argued that recombination occurred in the human mitochondrial genome (Kaneda *et al*., 1995; Howell *et al*., 2003; Awadalla *et al*., 1999; Eyre-Walker *et al.*, 1999; Hagelberg *et al*., 1999). The results of those studies were, however, refuted on the basis of erroneous published data (Elson *et al.*, 2004; Hagelberg *et al.*, 1999). A study by Kajander *et al.* (2001) claimed to have observed mtDNA recombination intermediates in human heart muscle and it was argued that heteroplasmy, exchanges between numts and mtDNA and low levels of leakage of paternal mtDNA because of the failure of the mechanisms to destroy paternal mtDNA could be responsible for recombination of human mtDNA. The possibility of recombination in the mitochondria was further explored on the basis that the mitochondria contained functional recombinase although there is still uncertainty about the issues of fusion and exchange of genetic information between paternal and maternal mtDNA (Legros *et al.*, 2002).

3.5.5 Homoplasmy and heteroplasmy

The mitochondrial DNA of a single individual is not identical within and between all cells and the presence of a new mutation in the mtDNA of a cell will initially present as a combination of normal mtDNA and mtDNA that contains the mutation. In the heteroplasmic state, the mutant mtDNAs will be shuffled through a process of replicative segregation

during cytokinesis (Wallace, 1988). The mtDNAs will be partitioned between the two newly formed cells on a random basis and in such a way that the mutant mtDNAs drift during each mitotic and meiotic division until a lineage with only wild type or only mutant type mtDNA is established (Wallace *et al.*, 1999). The generation that receives the heteroplasmic mutation would not necessarily all carry the same level of the new mutation and would also not display the same clinical features or symptoms in the case of a disease-associated mutation (Wallace, 1994; Jun *et al.*, 1994). Mutations linked to disease are usually heteroplasmic in the sense that the wild type is also present in the cells affected by the mutation. This heteroplasmic characteristic has an effect on the penetrance of the disease phenotype and can be linked with the level of heteroplasmy of the mutation. Inheritance of the heteroplasmic mtDNA mutation often shifts by large amounts between mother and offspring and makes the estimation of the recurrence risk of the disease within the offspring a complicated task (Wallace, 2008).

Furthermore, the mitochondrion has a unique manner of self-preservation, via a mitochondrial mutant selection system, which eliminates severely deleterious mutations before ovulation. The mtDNA within the oocytes will undergo about 20 mitotic cell divisions during which millions of proto-oocytes will be created, containing either mainly mutant or normal mtDNA, owing to the process of replicative segregation and drift (Jenuth *et al.*, 1997). This is then followed by the elimination of the oocytes harbouring the most deleterious mtDNA mutations, thereby protecting the offspring from extinction (Wallace, 2008).

3.5.6 Effective population size

The nuclear genome is inherited biparentally and therefore is diploid as opposed to the mtDNA, which is haploid. In terms of population genetics this has a dramatic impact on the effective population size when using DNA markers. The number of diploid nuclear DNA copies transmitted to the next generation would be twofold more than in the case of mtDNA copies and because of the diploid nature of humans, it means that the effective population size of the mtDNAs is half that of the nDNA. MtDNA mutations would therefore drift more rapidly to fixation during replicative segregation than for example in the case of nDNA. Effective population size is, however, also determined by the reproductive success of the species and in the case of nDNA would depend on the reproductive success of both male and female individuals. If one of the genders is high or low in reproductive success in comparison to the other, it could affect the effective population size of the nDNA

considerably and it should therefore not be assumed that the mitochondrial genome would always display a smaller effective population size, especially in populations where there are high levels of sexual selection (Ballard and Whitlock, 2004).

The effective population size of the mitochondrial genome is also affected by the lack of recombination and it therefore functions as a single locus. Under circumstances of selective sweeps in which a whole haplotype is selected because of one or more single advantageous mutations, it would have a more dramatic impact on the population size of the mitochondrial genome than in the case of nDNA under recombination (Kivisild *et al.*, 2006).

3.5.7 Neutrality versus selection

One of the reasons for the popularity of mtDNA as a marker of human evolutionary history was the belief that human mtDNA was a neutral marker and that the genealogy of the mitochondrial genome was therefore only shaped through mutation and genetic drift. The widespread acceptance of the selective neutrality of human mtDNA was based on the neutral theory developed by Kimura (1971), in which it was hypothesised that the majority of mutations that were fixed in a population were selectively neutral and were fixed through the random process of genetic drift rather than through selection. With the development of this theory, the concept that DNA sequences evolved in a clock-like manner came into play and it was proposed that evolution was governed by the stabilising-purifying selection that eliminated deleterious mutations as they occurred because of a constant mutation rate. They are then fixed or lost through the process of genetic drift (Kimura, 1991). The assumption of neutrality of the human mtDNA as a marker of evolution plays a critical role in measuring several of the population genetic events of the evolutionary history of a population, such as gene flow, effective population size, population subdivision and dating of the divergence events (Slatkin, 1985; Wilson *et al*., 1985; Avise *et al.*, 1987).

Evidence for selective neutrality of the mtDNA was based on the high rate of mutation that was observed in the mitochondrial genome and especially in the mitochondrially encoded rRNAs and tRNAs (Cann *et al*., 1984). It was argued that it could be expected that the translation apparatus of a small genome such as the mitochondrial genome would have to be under relaxed constraints and that the high levels of sequence variation were therefore evidence that fewer mutations were subjected to purifying selection (Ballard and Kreitman, 1995). In addition, it was assumed that the rate of evolution under the assumptions of

neutrality would equal the mutation rate of neutral alleles. It was therefore expected that neutrally functional nucleotide positions would exhibit lower mutation rates than the non-functional nucleotides, as was demonstrated by the higher rate of evolution of the first and second codon nucleotide positions as opposed to the lower rate of evolution of the third codon positions (Kimura, 1991).

A number of studies over the past years have, however, provided evidence that natural selection was at play in the mitochondrial genome and that the assumptions of the neutral theory were therefore inconsistent (Excoffier, 1990; Merriwether *et al.*, 1991; Nachman *et al.*, 1994). The first evidence of the effects of natural selection on mtDNA was reported by Whittam *et al.* (1986) in a study in which the high-frequency alleles of human mtDNA were observed more commonly, the intermediate-frequency alleles less commonly and the private mutations more commonly than expected under the assumptions of the neutral theory. Further studies followed with more evidence of the presence of selective constraints on the human mitochondrial genome. Nachman *et al*. (1994) demonstrated that the ratio of synonymous to nonsynonymous substitutions in humans and chimpanzees was higher than would be expected under the assumptions of neutrality. Merriwether *et al*. (1991) demonstrated that human mtDNA variation did not fit the mutation-drift equilibrium that would be expected under neutrality. Rogers and Harpending (1992) reported that the pairwise differences of human mtDNA from African populations displayed strong evidence of population expansions rather than conforming to the requirements of neutrality. The findings of these studies were not surprising, as it could be expected that the mitochondria, as the powerhouses of the cell, would have to be under selective constraints because of the lethal effects of deleterious mutations in the mtDNA. The ETC protein complexes formed by the proteins encoded by the mtDNA and the proteins encoded by the nDNA provided a further reason to expect selective constraints on the mtDNA. The lack of recombination in the mitochondrial genome will result in the whole genome acting as a single locus and therefore evolutionary forces would affect the whole genome. This means that if an advantageous mutation is fixed through selection, the other polymorphisms in the genome will also be fixed through a process of genetic hitchhiking and would therefore not be neutral (Smith, 1994; Ballard and Kreitman, 1995).

Currently, the presence of natural selection as a shaping agent of human mitochondrial variation is widely accepted. It is also generally believed that selection plays a role mainly through purifying selection that removes the deleterious mutations from the genome and thereby protects the fitness of individuals. The shaping of mtDNA variation through

processes of natural selection made it necessary to investigate the effect of positive selection on the mitochondrial variation among the major populations of the world and to consider the possibility that selection contributed to the adaption of humans to changing environmental conditions such as climate change and changes in diet. The findings of early investigations of the complete mitochondrial genomes of populations that resided in tropical Africa and populations that resided in the more temperate northern continents were interpreted as evidence that climate was a strong selective force that shaped the mitochondrial variation between the populations of the different continents (Mishmar *et al.*, 2003). Other studies, however, reported that climate had no influence on the mitochondrial sequence variation of individuals based on phylogenetic analyses and a bioenergetics approach (Moilanen and Majamaa, 2003; Elson *et al.*, 2004). Kivisild *et al*. (2006) demonstrated that diet could have been a selective force that drove mitochondrial variation because of the deficiency of the essential amino acids, threonine and valine, in most grains. The question of whether the distribution of mitochondrial variation between geographically diverse populations of the world was driven by positive selection in a quest to adapt to changing environmental conditions or whether it was driven by genetic drift assisted by purifying selection has elicited great controversy and is still under investigation.

3.6 MITOCHONDRIAL DNA VARIATION

Genetic mitochondrial variation is mainly introduced in humans through mutation and carried to the next generation without recombination owing to the maternal inheritance of the mitochondrial genome and is further shaped by natural selection and by genetic drift. The mitochondria have a high evolutionary rate due to the high mutation and high mutation fixation rate of the mitochondrial DNA (Wilson *et al.*, 1985). Mutations occur in the germ line cells or in somatic cells of humans and are subjected to natural selection, which removes deleterious mutations from the mitochondrial genomes and retains the advantageous mutations that assist the population to adapt to changing environments. Mitochondrial genetic variation is further shaped by the demographics of a population over time as it migrates and admixes or is isolated through a complex interplay of environmental factors and genetic survival. Mitochondrial variation is therefore critical to the study of the evolutionary history of human populations, as well as in the investigation of disease aetiology and treatment.

3.6.1 The nature of human mitochondrial DNA variation

Mitochondrial mutations occur either in the germ line cells or in somatic cells. During fertilisation, the mitochondrial DNA only starts to replicate at the blastocyst phase. This means that while the oocyte is dividing and replicating during the initial phase after fertilisation, the mitochondria are sorted to daughter cells in a way that causes a sampling error in terms of the number of mutant and normal mitochondria within each cell. This sorting can lead to a high amount of mutant mitochondrial DNA being sorted into a single germ-line daughter cell that will carry the mutation to the offspring. The mutant mitochondrial DNA can rapidly become homoplasmic within only a few generations (Giles *et al.*, 1980; Wallace, 1994; Sutovsky *et al.*, 2000). Mitochondrial variation is thus transmitted from mother to child through the process of germ line sampling and because there is no recombination involved, the mtDNA variation that is observed in individuals is radiated along maternal lineages.

Somatic mutations accumulate with age and consist of point mutations and deletions that present at the highest level in the basal ganglia and the cortical regions of the brain, the skeletal muscle and the heart (Cortopassi *et al.*, 1992; Wallace, 1994). Research has indicated that these types of mutations are most probably caused by oxygen radical damage and accumulate over time with age. Individuals who have inherited low levels of mitochondrial defects would need a high number of somatic mitochondrial mutations before they display symptoms of disease and it could take years before their organs are affected by somatic mitochondrial mutations (Wallace, 1994).

The patterns of mitochondrial variation between indigenous populations from different geographical regions have demonstrated that mtDNA variation is both extensive and adaptive. Most of the mitochondrial variation observed in the mitochondrial genomes of humans consists of neutral mutations at third codon positions or mutations in the non-coding regions of the mitochondrial genome. The mutations that occur in the functional protein-coding genes are under strong selective pressure and would be rapidly removed when deleterious to prevent compromising the fitness of the individual. Neutral mutations were fixed in radiating maternal lineages as the early humans migrated from their region of origin to populate the African continent and eventually the rest of the world (Cann *et al*., 1984; Merriwether *et al*., 1991, Ingman *et al.*, 2000). This phenomenon is displayed in phylogenetic trees as clusters or branches of similar mtDNA sequence variations that group together because of a shared common ancestor. The mtDNA

haplotypes observed in a population can therefore be grouped according to their ancestral origin and can thus be assigned to specific haplogroups that have been observed to be located within populations of the same geographical origin (Merriwether *et al*., 1991; Wallace, 1994; Ruiz-Pesini *et al.*, 2007). Human mitochondrial polymorphisms are therefore grouped into geographical regions according to the mitochondrial variation that occurred in the ancestral maternal lineages and because of the migration of population groups into different geographical regions of the world, where the mitochondrial variation was further shaped by evolutionary forces such as genetic drift and selection over a long period of time. The study of patterns of mitochondrial variation under the assumptions of different evolutionary scenarios forms the basis of using mtDNA in the study of the evolutionary past of human populations.

Mutations that occur in the human mitochondria will not always affect individuals in the same manner and therefore will not be subjected to the same evolutionary forces. The deleterious nature of pathogenic mutations, for example, eventually destroys the energy-producing function of the mitochondria and these alterations are usually eliminated by natural selection. For this reason it can be assumed that observed pathogenic mutations would be recent mutations that have not been removed from the mitochondrial genome by purifying selection yet. The study of these deleterious mutations is, however, important because they cause disease and are often linked to specific geographical regions. Adaptive mutations are the opposite of pathogenic mutations in that they affect the conserved sequence of the mtDNA in an advantageous way and are not removed by natural selection but rather retained. They consist of single substitutions and are present in ancient DNA at polymorphic frequencies. In contrast to the heteroplasmic nature of the pathogenic mutations, the adaptive mutations lead to homoplasmy (Ruiz-Pesini and Wallace, 2006).

The mutations that are observed in the mitochondrial genomes of individuals of different populations can be classified as recent or ancient based on the position of the nucleotide polymorphisms in a phylogenetic tree or network. Some mtDNA polymorphisms will be present at high frequencies and located in deep phylogenetic branches and other mutations will be at low frequencies and located at the tips of phylogenetic branches (Fu and Li, 1993). Recent mutations will either be removed or fixed through the evolutionary processes of natural selection or genetic drift. Recent advantageous mutations will similarly be fixed through the process of positive selection to become a polymorphism over time (Nachman *et al*., 1994). Most polymorphisms are therefore observed in populations

that are ancient or in lineages that have been present in a population for a long time (Fu and Li, 1993). For this reason, the mitochondrial variation observed in individuals can be described as recent and deleterious, ancient and neutral or ancient and adaptive (Wallace, 1994).

In the case of ancient adaptive mutations, it has been observed that despite the fact that these types of mutations affect highly conserved amino acids, they have persisted and expanded in populations over time and are often restricted to a geographically constrained branch of the mitochondrial tree (Wallace *et al*., 1999; Mishmar *et al.*, 2003; Ruiz-Pesini *et al.*, 2004). Adaptive changes can occur as many different missense mutations that alter a mildly conserved amino acid, or as only a few missense mutations changing a highly conserved amino acid. This phenomenon is demonstrated in the haplogroups J1 and J2 of European origin, which contain missense mutations at positions that affect highly conserved amino acids that are involved with the Q cycle and therefore affect the proton pump. The affected proton pump results in lower ATP production and more heat generation. Although these mutations affect highly conserved amino acids, the physiological effect on the human body was good adaptation to the cold of the European climate into which these populations migrated. In thermogenesis, mutations alter the energy allocation from ATP production to heat generation, which leads to an adaptive change. The survival of a species therefore depends on finding the fine balance between maintaining a number of adaptive mutations that prepare the body for extreme environmental changes and raise the chances of survival, and suppressing a number of mutations that can be potentially deleterious and lead to extinction (Wallace, 2007).

3.6.2 Mitochondrial DNA variation in studies of human evolution

Human evolution can be studied by using mitochondrial DNA in two distinct ways. The first is to study the history of mitochondrial DNA lineages by using haplogroups to identify how related two or more mtDNA sequences are to each other. The disadvantage of this method is that the haplogroups do not necessarily represent the history of the populations under investigation. The origin and therefore the age of a haplogroup does not indicate the origin or age of a population and the demographic movement of haplogroups does not represent the movement of one population only but that of a whole population with many different haplogroups.

The second method to study human evolution is by using the population-based approach. This approach uses the application of population-genetics methods to population groups to study population phenomena such as population expansion, migration and admixture. These generally consist of statistical methods that use a model or assumption against which a set of observed data is tested for validity or can consist of algorithms based on certain evolutionary models and assumptions against which the phylogenetic relationships between mtDNA sequences are modelled (Parkendorf and Stoneking, 2005).

Both approaches to the study of human evolution by using mtDNA require the identification or observation of changes to nucleotide substitutions in the mtDNA sequences of individuals of a population or between populations. Conventionally this is achieved by comparing the sequences of the mtDNA under investigation to a standard reference mtDNA sequence. The Cambridge Reference Sequence (CRS) was first published in 1981 by Anderson *et al.* (1981) and since then has been updated by Andrews *et al*. (1999) and is publicly available for these purposes.

Technological developments over the past 30 years have changed the methodology with regard to mtDNA analysis rather drastically. Initially, studies of mitochondrial DNA variation were based on the study of RFLPs, which involved the cleavage of mtDNA at five or six restriction enzyme sites (Merriwether *et al*., 1991; Salas *et al.*, 2002). This method was followed by a method for high-resolution RFLP analysis, which involved the cutting of the mitochondrial genome at 12 to 14 restriction enzyme sites (Macaulay *et al.*, 1999; Torroni *et al*., 2001). The early PCR-RFLP studies were performed on the coding region of the human mitochondrial genome to avoid homoplasy (Soodyall *et al.*, 1996; Torroni *et al*., 1992) and were subsequently expanded to include the hyper-variable segment I (HVS-I) of the non-coding control region of the mitochondrial genome to increase the resolution (Torroni *et al*., 1996; Chen *et al*., 2000; Kivisild *et al.*, 2002). The rate of mutation in the control region is higher than in the coding region and is reflected in the abundance of mutations that are present in the control regions. The problem with the high mutation rate is that the incidence of homoplasy is also much higher in this region of the mitochondrial genome and results in an obscured genetic signal of evolution (Tamura and Nei, 1993). In addition, the control region consists of 1,121 bp, which is about 7% of the total number of nucleotides of the human mitochondrial genome and therefore it was not surprising that the evolutionary signal from this region alone was not sufficient to distinguish between important ancient phylogenetic branches (Maddison *et al.*, 1992). In contrast to RFLPs, DNA sequencing strategies initially focused on the control regions only or the HVS-I or

HVS-II only. It is only since the late 1990's that sequencing of the full mitochondrial genome has become possible (Richards *et al.*, 2001). Some studies focused on studying mitochondrial genome variation by performing RFLPs on the whole mitochondrial genome (Cann *et al.*, 1984) or only on the HVSs I and II (Vigilant *et al.*, 1989, Chen *et al.*, 1995a, Watson *et al.*, 1996). More recently, the focus has shifted to studying mitochondrial variation by sequencing the complete mtDNA genome to provide mitochondrial sequence data of the highest resolution (Ingman *et al*., 2000; Finnila *et al*., 2001; Maca-Meyer *et al.*, 2001; Kong *et al*., 2003; Coble *et al*., 2004; Behar *et al.*, 2008).

3.7 MITOCHONDRIAL DNA HAPLOGROUPS

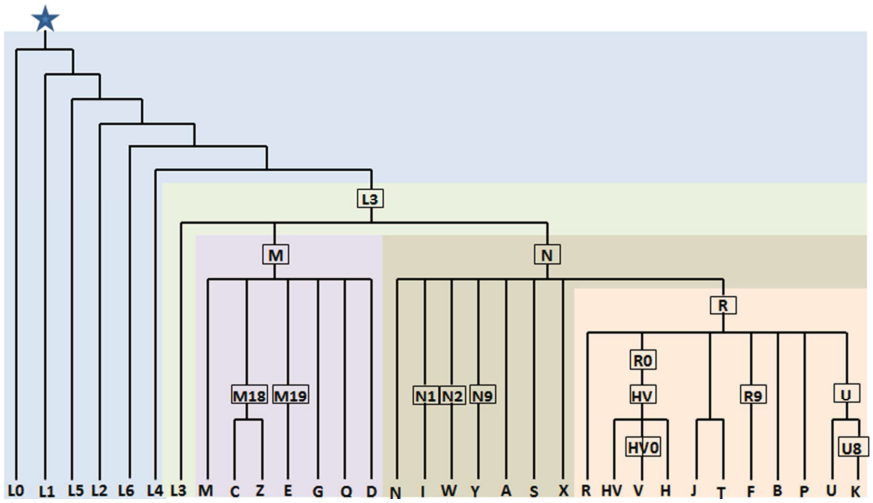
Haplogroups are defined by shared mutation profiles in the human mitochondrial genome, which are acquired through the sequential accumulation of mutations of maternal lineages that have developed over time. Mutations are added to a founder haplotype to eventually form an established motif of sequence variants within a group of humans that share a common ancestry. These molecular differentiations are rapid and have occurred during and after the dispersal of humans to different regions and continents of the world and thus are generally restricted to particular geographical locations (Torroni *et al.*, 2006).

Haplogroups are suitable for the study of evolution because they consist of arrays of alleles that are closely linked, which show very little, if any, recombination. They also present with ancient origins. MtDNA haplogroups provide a molecular record of the genealogical history of a population as well as a molecular trace of migration patterns of humans over time. The radiating maternal lineages also provide good evidence of the extent of genetic subdivision among populations of regions or continents (Torroni *et al.,* 2006). Haplogroups are therefore valuable in the investigation of evolution and population behaviour and are used extensively in the construction of pylogenetic trees and networks.

Alphabetical letters were assigned to the haplogroups as they were reported, starting with a study by Torroni *et al.* (1992) on Native Americans in which four fundamental clusters of different sequence variant motifs were described and named with the first four letters of the alphabet, i.e. A, B, C and D. Subsequent studies have resulted in the classification of many more haplogroups by using all the letters of the alphabet, with the exception of O. The classification of haplogroups is a constant process and new haplogroups are constantly discovered as more mitochondrial sequence data become available. As more mitochondrial sequence motifs were reported through subsequent studies, it was

necessary to establish rules for the hierarchical ordering of the haplogroups. It was decided that the alphabetical letter that constituted the major haplogroup that the mitochondrial sequence belonged to, be followed by alternating numbers and letters that would assign it to a hierarchical level (Richards and Macaulay, 2001; Kivisild *et al.*, 2006). Although standard rules for the nomenclature existed, there were cases where different sequence motifs were assigned to the same haplogroup name such as haplogroup M12 that was assigned to mtDNA sequences of Kong *et al*. (2006) and Tanaka *et al.* (2004). Cases where the same mtDNA sequence motifs were assigned to different haplogroup names also occurred, such as the mtDNA sequences of Achilli *et al*. (2008) that were assigned to C4 and the same sequence motifs in Volodka *et al.* (2008) that were assigned to haplogroup C2. To overcome these miscommunications, Van Oven and Kayser (2009) constructed a global human phylogenetic tree that contained the complete mitochondrial genomes of 55 published studies and constituted the most updated and recent human phylogenetic tree at the time of publishing in 2009. The PhyloTree, as it is known, is publicly available and regularly updated to ensure that it is always current and updated with the latest mitochondrial sequence motifs as they become available (Van Oven and Kayser, 2009). A basic outline of the phylogenetic hierarchical structure of the human mitochondrial haplogroups of the world as presented in PhyloTree (Van Oven and Kayser, 2009) is presented in Figure 3.4.

Figure 3.4 Global mitochondrial haplogroup hierarchy



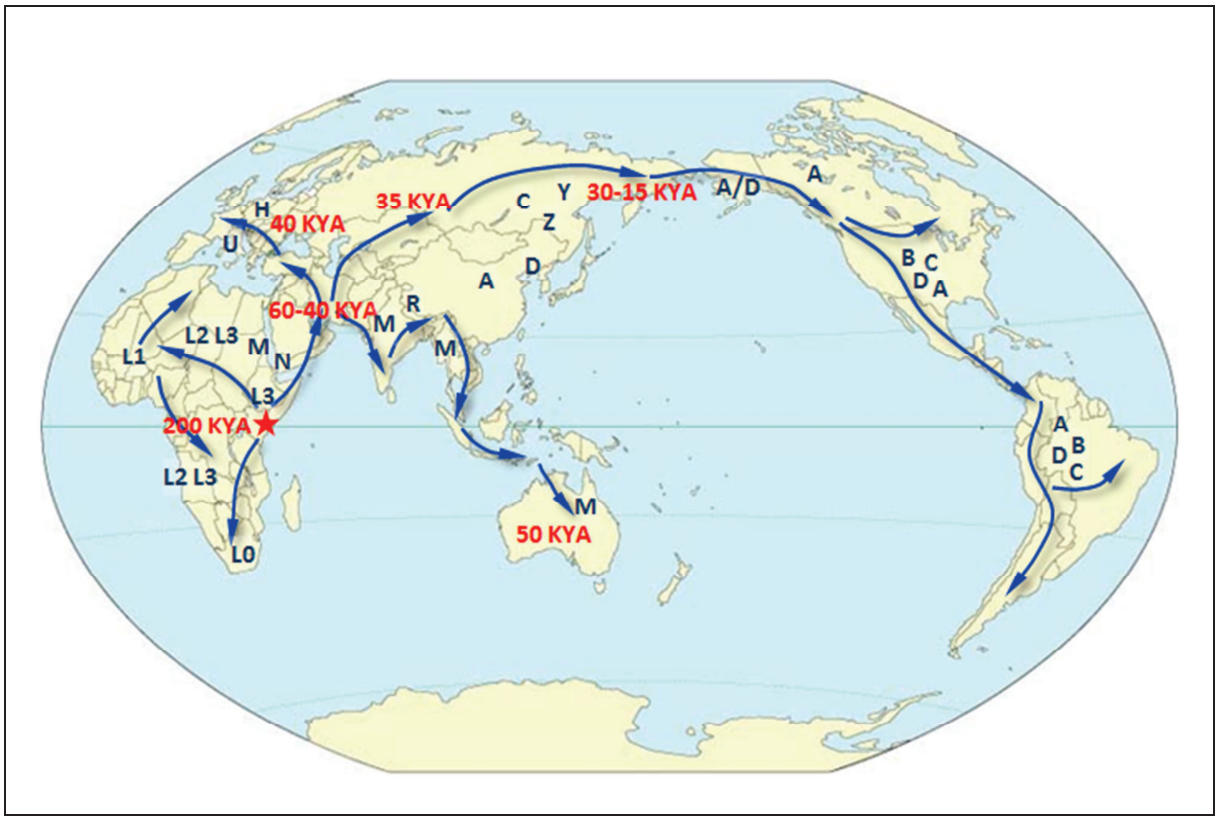
Basic phylogeny of global mtDNA haplogroups. The root of the phylogeny is indicated by a blue star, which represents the single matrilineal ancestor of all humans. Haplogroup L is specific to the African continent, indicating that the origin of modern humans was in Africa. Haplogroup L3 gave rise to all the other major global haplogroups. Major haplogroup M gave rise to lineages C, D, E, G, Q, and Z; major haplogroup N gave rise to lineages A, I, S, W, X, and Y; major haplogroup R gave rise to lineages B, F, HV, H, J, K, P, T, U, and V. Adapted from Van Oven and Kayser (2009).

The major haplogroup L lineages constitute the deepest phylogenetic branches, suggesting that the African lineages are the most ancient lineages of all haplogroup lineages of the world. This finding supports the Out of Africa hypothesis, which states that the first anatomically modern humans had their origin in Africa from where they populated the rest of the world. Haplogroup L3 forms the ancestral branch to the major haplogroups M, N and R, suggesting that an early human population that belonged to haplogroup L3 was the first human population to migrate from Africa to regions outside Africa where it gave rise to the rest of the global haplogroups. Major haplogroup M gave rise to haplogroups C, D, E, G, Q and Z. Major haplogroup N gave rise to haplogroups A, I, S, W, X and Y. Major haplogroup R gave rise to haplogroups B, F, HV, H, J, K, P, T, U and V (Van Oven and Kayser, 2009).

3.7.1 Mitochondrial haplogroup dispersal in the world

The discovery that the mitochondrial genome with its unique characteristics, such as maternal inheritance, high mutation rate, lack of recombination and high copy number, was well suited for the study of population histories and evolution, unlocked a new field in science often referred to as archaeogenetics. This term was coined by Amorim (Torroni *et al.*, 2006) and referred to the application of genetics to the study of population history. Since the first studies, aimed at predicting the ancient evolutionary history of early migrations and the development of anatomically modern humans by combining genetics and archaeology, many others have followed. Much has been learnt from the large amounts of mitochondrial DNA sequence data that were generated and from the large number of new haplogroups that was assigned to populations from all over the globe. This information has enabled scientists to track the migrations of AMHs since their origin in east Africa, across the continent of Africa, into Europe, southward to Asia and Oceania and eventually through Siberia to the Americas. The distribution and migration routes of the early modern humans are summarised in Figure 3.5.

Figure 3.5 MtDNA and the migration of world populations



Routes of world migrations of early AMH populations ; origin of AMHs indicated by **red star**; **blue arrows** show directions of migrations; date estimates of the major migration events are indicated in red; prevalent haplogroups that developed in the different world regions in **dark blue**. Adapted from Forster (2004).World map adapted from http://www.worldmapsphotos.com accessed on 8 November 2011.

Studies of human mitochondrial genomes supported a late Pleistocene expansion of modern humans from Africa (Cann, 1987; Vigilant *et al*., 1991; Watson *et al.*, 1997; Ingman *et al.*, 2000) across the Red Sea into Arabia from where the AMHs migrated southward towards southern Asia to Eurasia (Macaulay *et al.*, 2005), India and eventually into the Sahul, which later became New Guinea and Australia (Ingman and Gyllensten, 2003; Friedlaender *et al.*, 2005; Van Holst Pellekaan *et al.*, 2006). A later migration occurred from northern Africa to Europe, followed by the population of northern Europe, moving into Siberia and from there over the Beringia land bridge to America (Merriwether *et al.*, 2005).

3.7.1.1 Origin of anatomically modern humans

The theory of a mainly single origin for AMHs is supported by a large body of mitochondrial and other sequencing evidence (Cann, 1987; Underhill *et al.,* 2000). Two major branches, the L0 and L1'2'3'4'5'6 (L1'6) daughter lineages (Torroni *et al.*, 2006), coalesce into a single maternal point of origin in the global human phylogenetic tree, providing strong support for an African origin of AMHs (Wallace *et al.*, 1999; Mishmar *et al.*, 2003; Kivisild

et al., 2006). Coalescent dating of the daughter branches at the root of the human phylogenetic tree sets this event at about 200 kya (Ingman *et al.*, 2000; Mishmar *et al.*, 2003; Torroni *et al.*, 2006; Gonder *et al.*, 2007; Behar *et al.*, 2008). It is currently widely accepted that the L1'6 lineages gave rise to most of the extant African haplogroups through an early expansion of modern humans from their place of origin, which is most probably in east Africa, to the western and northern regions of Africa. The distribution of haplogroups L2 and L3 provides reason to believe that the initial migration was followed by an overwhelming wave of L2 and L3 lineages around 100 kya, after which a small population of L3 lineages split from the other populations and migrated out of Africa to populate the rest of the world between 60 kya and 40 kya (Watson *et al.*, 1997; Forster, 2004).

3.7.1.2 Out of Africa

The human phylogenetic tree displays the major haplogroup L3 at the root of all the other haplogroups observed in non-African individuals and based on other evidence such as archaeology and coalescent dating of the L3 branch of the human phylogenetic tree, it has been concluded that this haplogroup most probably represented the genetic makeup of a small human population that migrated out of Africa through the Horn of Africa between 60 kya and 40 kya (Salas *et al.*, 2002; Macaulay *et al.*, 2005; Kivisild *et al*., 2006; Torroni *et al.*, 2006). Evidence of climate changes towards the beginning of the glacial interstadial phase, followed by cultural changes in human behaviour in response to the changing environment, gave this small population a competitive advantage over other populations, which triggered the migration. The high genetic diversity of haplogroup L3 in current Ethiopians provides reason to believe that the migration took place through the Horn of Africa (Kivisild *et al.*, 2004).

Haplogroups M and N are the two main haplogroups that branched from haplogroup L3, suggesting that the individuals that belonged to these haplogroups were the ancestors of all the non-African populations that followed (Macaulay *et al.*, 2005). Haplogroup R diverged from haplogroup N and it is hypothesised that these haplogroups were the founders of the Eurasian settlement between 60 kya and 65 kya (Kong *et al.*, 2003; Macaulay *et al.*, 2005). Individuals who belong to these three (3) haplogroups reside over a large geographical area that stretches from Eurasia to Australasia and Oceania through to the Americas. It is the southern regions of Eurasia, however, that display the largest number of haplogroup M, N and R lineages; providing evidence that the founder

populations colonised and expanded along the southern coast of Asia to India and southeast Asia. The coalescent time estimates indicate that this migration and settlement took place at about 60 kya as opposed to the colonisation of the northern Eurasian regions that only occurred at about 45 kya (Macaulay *et al.*, 2005).

3.7.1.3 Migration to Oceania and Australia

Genetic evidence suggests that the AMHs dispersed from Africa via southern Asia to Oceania and Australia about 45 kya. This is also referred to as the southern route hypothesis, which speculates that the early humans migrated out of Africa through the Horn of Africa across the Red Sea into Arabia and from there migrated in a southern direction towards southern Asia (Kivisild *et al.*, 2006; Mellars, 2006; Torroni *et al.*, 2006). Based on genetic and other evidence (of which the phylogenetic relatedness between Indians and Australian Aborigines was the most powerful (Kumar *et al.*, 2009)), it was further believed that the humans then migrated via the coast of southern Asia to Indonesia and eventually into Australia where the first modern humans populated Australia between 60 kya and 55 kya (Stringer, 2002; Macaulay *et al.*, 2005). During that time the sea levels were low and the southern region of Asia consisted of the single land mass, Sunda. The humans migrated over the sea straits from Sunda to the Sahul, which constituted the modern-day New Guinean and Australian continents. It is hypothesised that the modern humans preferred warmer climates and therefore migrated to the southern regions prior to their migration to Europe. Studies of the mtDNA sequences of Australians and New Guineans demonstrated that they belonged to founder haplogroup M and N lineages as the Eurasians did, and that the Australian and New Guinean early populations were descended from the same founders as the Europeans (Ingman *et al*., 2000; Kivisild *et al.*, 2006; Kumar *et al*., 2009).

3.7.1.4 Migrations into Europe

Europeans harbour haplogroups that are mostly all related to or the same as the eastern Asians and therefore most probably developed from a single phylogenetic ancestral branch and differentiated into many different sub-haplogroups. Studies have demonstrated that these haplogroups were absent from sub-Saharan and east African populations (Torroni *et al.*, 1996). It was further demonstrated that the European lineages were far more limited than what was observed in the south Asian populations and that most of the European haplogroups were clades of only three (3) haplogroup R branches, namely the R

(previously known as the pre-HV lineage), the JT and the U lineages. The only other haplogroups that have been observed in Europeans are three (3) minor clades of N i.e. N1 and N2, X and R i.e. R1 and R2 (Macaulay *et al.*, 2005). These lineages have been dated to between 50 kya and 40 kya and most probably diverged during the glacial interstadial phase (Kong *et al.*, 2003). Most of the current Europeans display haplogroups that originated directly from autochthonous Mesolithic or Palaeolithic ancestors, who most probably migrated from the southern regions of Europe into the rest of Europe during the last glacial phase (Torroni *et al.*, 1998; 2001). It is believed that these populations harboured large components of haplogroups H and U5 and that the sub-haplogroups H1, H3, V and U5b subsequently developed as the northern and western regions of Europe were populated (Achilli *et al.*, 2004; Pereira *et al.*, 2005). The divergence of these haplogroups was estimated to have occurred between 50 kya and 15 kya. Other lineages were introduced to Europe at a later stage during the Neolithic and post-Neolithic periods and consisted of immigrants that harboured haplogroups J1b1, J2a, T1a, R1, R2 and N1a (Richards *et al.*, 2000).

3.7.1.5 Migration to the Americas

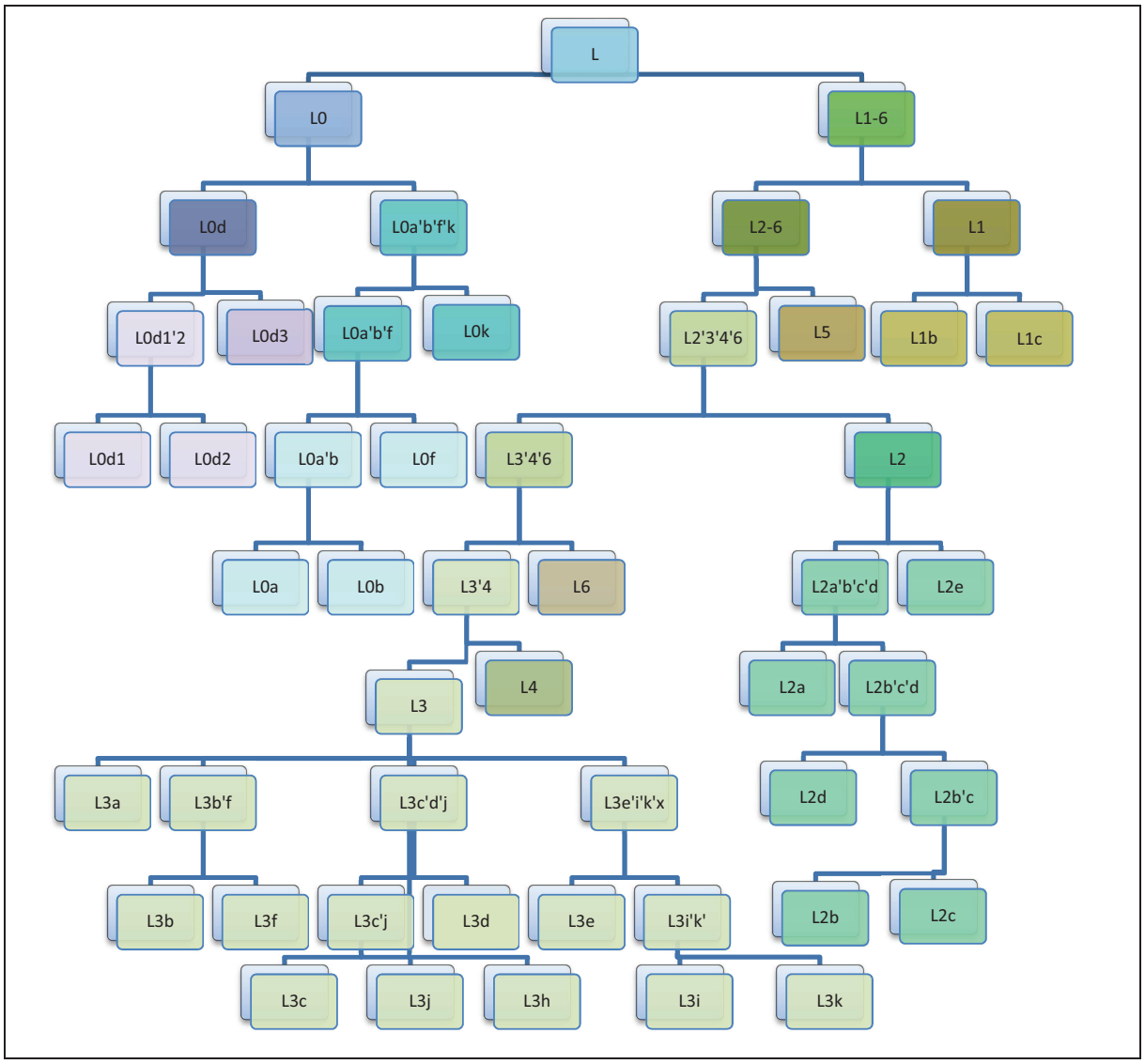
The major human expansion around 30 kya to 25 kya (due to climate changes during the last Ice Age that transitioned from phase 3 to phase 2), most probably provided the trigger for a migration of human populations that were located in Asia and northern Siberia to America. America and Asia were connected at that time by a wide land bridge i.e. Beringia, owing to the low sea level (Zazula *et al.*, 2003). During the period between 20 kya and 15 kya the glacial maximum was reached and the human populations were forced to migrate southward where refuge was sought in the regions of Beringia and Iberia. During that period the genetic diversity of the surviving populations was narrowed down to A/D and H/V respectively. Two major population groups survived the migration to the American continent. The first group consisted of populations that gave rise to the modern-day Eskimos, Aleut and Dene-speaking populations and who lost haplogroups B, C and X from their genealogy because of extinction. A larger group of Amerinds settled in Meadowcroft and in Monte Verde between 18 kya and 14 kya, from where they slowly migrated south towards the current Chile (Torroni *et al.*, 2001; Forster, 2004). These Asian populations from northern Siberia were characterised by haplogroups A, B, C, and D, which are widely observed in current America (Merriwether *et al.*, 2005).

3.7.2 Mitochondrial haplogroup dispersal in Africa

Macro-haplogroup L is regarded as unique to African populations and 70% to 100% of individuals of African descent belong to this haplogroup (Torroni *et al.*, 1996). Haplogroup L forms a deep basal split with the other lineages of the global human phylogenetic tree (Van Oven and Kayser, 2009) and this therefore supports the origin and development of early AMHs in Africa. African lineages further demonstrate the highest level of genetic diversity in the world because of their ancient nature (Watson *et al.*, 1997; Ingman *et al*., 2000; Mishmar *et al.*, 2003).

The phylogeny of haplogroup L consists of two major daughter branches, the L0 lineage and the L1, L2, L3, L4, L5 and L6 (L1'6) lineages (Salas *et al.*, 2002; 2004; Mishmar *et al*., 2003; Kivisild *et al.*, 2004). These lineages radiated into several daughter lineages that were shaped by periods of population migration and isolation, expansion and contraction and admixture driven by climatic and environmental conditions over many thousands of years. A broad outline of the haplogroup L hierarchy is presented in Figure 3.6 (Van Oven and Kayser, 2009).

Figure 3.6 Macrohaplogroup L hierarchy



Mitochondrial macrohaplogroup L hierarchy according to the PhyloTree (Van Oven and Kayser., 2009). The ancestral root indicated in blue as "L" followed by the two sister branches, L0 and L1'2'3'4'5'6 from where the lineages split into sub-groups. Adapted from PhyloTree, //www.phylotree.org accessed on 30 November 2011.

3.7.2.1 Mitochondrial haplogroup L0

Haplogroup L0 was the first branch of the mtDNA phylogeny between 160 kya and 140 kya (Mishmar *et al.*, 2003; Torroni *et al.*, 2006; Behar *et al.*, 2008). The L0 haplogroup consists of sub-haplogroups L0a', L0b', L0d, L0f' and L0k (Salas *et al.*, 2004; Gonder *et al.*, 2007, Behar *et al.*, 2008).

The ancestral root of haplogroup L0 branched off into the L0d lineage between 144 kya and 100 kya and a root lineage for the other sub-haplogroups i.e. L0abfk (Gonder *et al.*, 2007; Behar *et al.*, 2008). The L0k haplogroup diverged from the L0a'b'f early on and was, in addition to haplogroup L0d, the most prevalent haplogroup to be observed in the ancient

hunter-gatherer Khoi-San speaking populations of southern Africa. The only other African population that harboured these haplogroups was the ancient Tanzanian Sandawe population (Tishkoff *et al.*, 2007), which demonstrated a shared mitochondrial ancestry between the two population groups, which suggested that both populations were derived from a common early proto-Khoi-San population that most probably migrated from eastern Africa southward through the current Tanzania to the southern regions of Africa (Pereira *et al*., 2001; Salas *et al.*, 2002; Tishkoff *et al.*, 2007). The L0d haplogroup currently consists of three (3) sub-haplogroup lineages i.e. haplogroup L0d1, L0d2 and L0d3 (Van Oven and Kayser, 2009).

Some controversy exists about the route of migration and place of divergence of the haplogroups L0d and L0k from haplogroup L0a'b'f. One hypothesis argues that the L0 and L1'6 lineages coexisted in eastern Africa until a small population split from the parent L0 group to migrate to the southern regions of Africa, whereby the mitochondrial variation became limited to haplogroups L0d and L0k due to the isolated existence of these populations and genetic drift over time (Forster, 2004; Kivisild *et al.*, 2004; Mellars, 2006). An alternative hypothesis argues that the sister branches of the phylogeny of haplogroup L consisted of the L1'6 lineages and L0 lineages and that a small population containing L0 lineages migrated to the southern regions of Africa from where a back migration to eastern Africa occurred, containing lineages that developed into the root L0a'b'f lineage and that settled with the L1'6 lineages in eastern Africa (Behar *et al*., 2008). Based on the high prevalence of L0a in African populations, it is generally believed that the L0a'b'f lineage migrated with the L1'6 lineage to populate the rest of Africa (Watson *et al.,* 1997; Salas *et al*., 2002; Gonder *et al.*, 2007; Coelho *et al*., 2009).

The L0a lineage currently consists of sub-lineages L0a1, L0a2, L0a3 and L0a4 (Van Oven and Kayser, 2009). Haplogroup L0a1 is mainly observed in eastern and south-eastern Africa (Salas *et al*., 2002; Behar *et al*., 2008) and to a lesser extent in western Africa (Rando *et al.*, 1998) and diverged about 30 kya. Haplogroup L0a2 is widely regarded as a marker of the Bantu dispersal between 3,000 and 4,000 years ago (Soodyall *et al.*, 1996).

3.7.2.2 Mitochondrial haplogroup L1

The major haplogroup L1 diverged from the L1'6 root lineage between 150 kya and 140 kya (Torroni *et al.*, 2006; Behar *et al*., 2008). It consists of two major sub-haplogroup lineages, L1b and L1c (Van Oven and Kayser, 2009). Haplogroup L1b is prevalent in the

western and central African regions and is an example of a lineage that underwent a bottleneck, which would explain its extant low level of mitochondrial variation (Kivisild *et al.*, 2004; Behar *et al.*, 2008). This haplogroup is also present in African Americans because of the Atlantic slave trade (Salas *et al.*, 2002). Haplogroup L1c has six clearly defined sub-clades, namely L1c1, L1c2, L1c3, L1c4, L1c5 and L1c6, and high internal diversity (Van Oven and Kayser, 2009). Evidence points to the haplogroup L1c originating in central Africa (Rando *et al.*, 1998) and dispersing via the western Bantu expansion to the western central regions of Angola and the Congo Delta (Coelho *et al.*, 2009). It also constitutes about 70% of the maternal ancestry of the ancient hunter-gatherer Pygmy populations of central Africa, which displays only a single sub-lineage i.e., the haplogroup L1c1a, of all the sub-lineages of haplogroup L1c (Quintana-Murci *et al.*, 2008).

3.7.2.3 Mitochondrial haplogroup L5

Haplogroup L5 was initially assigned as haplogroup L1e (Salas *et al.*, 2002). Salas *et al.* (2002) reported that individuals who belonged to haplogroup L1e were mainly restricted to eastern Africa and this haplogroup was observed at low frequencies in Mbuti Pygmies of central Africa and Bantu-speaking individuals from south-eastern Africa. Later studies re-assigned it to haplogroup L5 at an intermediary position between haplogroups L1 and L2'6 (Kivisild *et al.*, 2004; Gonder *et al.*, 2007; Torroni *et al.*, 2006; Behar *et al.*, 2008). Haplogroup L5 displayed an ancient nature based on its divergence from haplogroup L1'6 between 140 kya and 120 kya. The ancient nature of this haplogroup was further indicated by its presence in the ancient hunter-gatherer Pygmy population of central Africa. It is regarded by scientists as one of the most primitive haplogroups present in the Pygmy populations over a long period of time (Gonder *et al*., 2007; Quintana-Murci *et al.*, 2008).

3.7.2.4 Mitochondrial haplogroup L2

Haplogroup L2 is one of the most common haplogroups observed in sub-Saharan African populations (Torroni *et al.*, 2001) and currently contains sub-haplogroups L2a, L2b, L2c, L2d and L2e (Van Oven and Kayser, 2009). This lineage diverged from the L1'6 root lineage between 120 kya and 100 kya in western or central Africa (Salas *et al.*, 2002).

The L2a clade is the most common sub-group of haplogroup L2 and is widespread over the African continent, thus obscuring its exact place of origin (Torroni *et al.,* 2001; Salas *et al*., 2002; Kivisild *et al.*, 2004; Behar *et al.*, 2008). Evidence indicates that haplogroup

L2a could have originated in central Africa between 55 kya and 45 kya and dispersed from there in western and eastern directions into Africa along the Sahel corridor after the LGM (Salas *et al.*, 2002). The presence of sub-haplogroups L2a1a and L2a2 in south-eastern Bantu-speaking individuals suggested that these later sub-clades originated during the Bantu dispersal to the southern regions of Africa and could be regarded as markers of the Bantu dispersal (Pereira *et al.*, 2001; Salas *et al.*, 2002; Atkinson *et al.*, 2009).

Haplogroup L2b diverged later, between 30 kya and 25 kya, into the western or west-central regions of Africa (Salas *et al*., 2002; Rosa *et al*., 2004; Behar *et al*., 2008). Haplogroup L2d is the most ancient of the haplogroup L2 sub-clades and diverged between 120 kya and 100 kya also in western or west-central Africa (Salas *et al.*, 2002; Behar *et al*., 2008). The L2c clade is also common in the western African Senegalese populations and is rarely observed in the eastern and southern parts of Africa (Torroni *et al.,* 2001).

3.7.2.5 Mitochondrial haplogroup L3

Haplogroup L3 originated in eastern Africa between 75 kya and 60 kya (Salas *et al.*, 2002, Macaulay *et al*., 2005; Kivisild *et al.*, 2004; Behar *et al.*, 2008). It currently contains sub-haplogroups L3a, L3b, L3c, L3d, L3e, L3f, L3h, L3i, L3j, L3k and L3x (Van Oven and Kayser, 2009), which are distributed widely in high numbers in Africa (Salas *et al*., 2002). Haplogroup L3 was also the founding maternal ancestor of the major haplogroups M and N, which are associated with the migration out of Africa between 60 kya and 40 kya (Salas *et al.*, 2002).

Haplogroup L3b and L3d are prevalent in west Africa and in African Americans owing to the Atlantic slave trade (Rando *et al.*, 1998; Salas *et al.*, 2002). Haplogroup L3b is common in Hutu individuals from Rwanda (Castri *et al.*, 2009) as well as in the southern African !Kung Khoi-San speaking populations (Chen *et al.*, 2000). Sub-clade L3d3 is also well represented in the !Kung and Khwe Khoi-San speaking populations of southern Africa and Angola and is associated with the recent Bantu dispersals to the southern regions of Africa (Watson *et al.*, 1997; Coelho *et al.*, 2009). Haplogroup L3d has also been reported in individuals from Tanzania (Tishkoff *et al.*, 2007).

L3e is the most widespread and ancient of the L3 clades and is common in the sub-Saharan populations. It is believed that it originated between 50 kya and 40 kya in

central Africa (Salas *et al.*, 2002; Rosa *et al.*, 2004; Torroni *et al.*, 2006; Behar *et al.*, 2008). Haplogroup L3e has also been observed in the mtDNA sequences of Brazilian individuals, which suggests that this sub-clade must be present in Angola and that it could have been carried to Brazil through the slave trade from Angola (Pereira *et al.*, 2001). Haplogroup L3e1 is most common in sub-Saharan Africa and especially in south-eastern African populations and is believed to have originated about 16 kya in central Africa (Pereira *et al*., 2001; Salas *et al.*, 2002). Haplogroup L3e2 diverged prior to haplogroup L3e1 between 35 kya and 25 kya and is mostly present in central and western Africa (Salas *et al*., 2002; Rosa *et al*., 2004). The L3e2 lineages, however, continued diverging until nine kya, which suggests that this clade was dispersed throughout the Sahara during the migrations of the Great Wet Phase of the Holocene (Bandelt *et al.*, 2001b). Haplogroups L3e3 and L3e4 are prevalent in western Africa and are associated with population expansion events during a period of increased food production and iron smelting in western Africa (Bandelt *et al*., 2001b; Rosa *et al.*, 2004). Haplogroup L3e5 most probably originated in the Chad Basin from where it dispersed to the northern regions of Africa (Cerný *et al*., 2007).

3.7.2.6 Mitochondrial haplogroup L4

Haplogroup L4 is commonly observed in eastern and north-eastern Africa and is a sister clade of haplogroup L3 (Kivisild *et al*., 2004; Tishkoff *et al*., 2007). It currently contains sub-haplogroups L4a and L4b (Van Oven and Kayser, 2009).

Evidence suggests that haplogroup L4 diverged from haplogroup L3 during the late Pleistocene in a period just prior to the migration of the haplogroup L3 population from Africa toward Eurasia (Kivisild *et al*., 2004). Sub-haplogroup L4a is commonly observed in Ethiopia and Sudan and sub-haplogroup L4b in Tanzania and Ethiopia and they have been estimated to have coalesced between 90 kya and 55 kya in eastern Africa (Kivisild *et al.*, 2004; Behar *et al.*, 2008).

3.7.2.7 Mitochondrial haplogroup L6

Haplogroup L6 has not been reported widely in literature and therefore much of its evolutionary history is still unknown. Divergence from L3’4 has been estimated at about 110 kya (Torroni *et al.*, 2006). In contrast, most of the sub-clades of haplogroup L6 display recent coalescent dates of about 22 kya (Behar *et al.*, 2008). This could be explained by

an extinction of past variation, lack of expansion or the lack of mtDNA samples from the place of origin.

3.8 MITOCHONDRIA AND DISEASE

MtDNA variation plays an important part in understanding age-related, metabolic and neurodegenerative diseases, as well as the process of ageing and cancer. These diseases are defined by a lack of cellular energy and are often classified as complex diseases, as they occur in families although they are not inherited according to Mendelian principles and seem to be affected by environmental factors. Epigenetic changes in response to environmental changes influence the regulation of nDNA gene expression, and they are therefore also partly responsible for these diseases (Wallace, 1994). It is estimated that mtDNA diseases have an incidence of 1.65/10,000 and have been linked to a wide range of clinical problems (Wallace, 2007).

One of the most important factors affecting an organism is its supply of and need for energy and whether the organism is able to use the energy provided by the environment. This mainly happens through the mitochondria where compounds such as fats and carbohydrates are converted to ATP, and glycolysis of energy-rich compounds, such as acetyl-Coenzyme A (acetyl-CoA), S-adenosyl-methionine (SAM) and NADH takes place, providing the cell with usable energy to replicate and grow. The replication and growth of the nDNA is regulated *inter alia* by the modification of histones that allow the opening of the chromatin through phosphorylation by ATP, acetylation by acetyl-CoA, deacetylation by NAD^+ and methylation by SAM, which are all energy products of the mitochondria. The expression of mitochondrial genes is regulated by the requirements of the nDNA, through inter-chromosomal coordinate transcriptional regulation. Through these complicated interactions, it seems that the mitochondrion plays a role in the diseases of the epigenome and *vice versa* (Wallace and Fan, 2010).

The high mutation rate of the mitochondria is an advantage in terms of genetic variation but a disadvantage in terms of disease association. Mutations in either mitochondrial DNA or nDNA will most probably cause deficient energy production, which affects the tissues that demand most energy, such as the heart, muscle, brain, renal and endocrine systems. It also affects the metabolic systems that result in diseases such as diabetes and obesity (Wallace, 1994). The onset of mitochondrial disease is determined by the threshold effect, which is the level of mutant mtDNA in a cell that would be necessary to cause biochemical

defects and dysfunction (Wong, 2007). The number of mutant mtDNAs increases over time with replicative segregation and as the level of deleterious mutations increase, the cellular energy output will decrease. The tissues and organs depend on different levels of energy and therefore will display symptoms of disease at different times because of the presence of the mutant mtDNAs (Wallace *et al.*, 1988; Wallace, 1995).

The level of mutant mtDNA that is transmitted to the next generation is mainly determined by a genetic bottleneck during replicative segregation. Although the existence of such an mtDNA bottleneck has been widely reported (Holt *et al.*, 1988; Vilkki *et al.*, 1990; Larsson *et al.*, 1992), the mechanism that underlies the bottleneck is not fully understood. One hypothesis is that it is caused by a drastic reduction of mtDNA copy number in the germ line (Jenuth *et al.*, 1997), while another states that it is caused by the preferential replication of a selected group of mtDNAs during oogenesis (Cao *et al*., 2007). Recently it has been suggested that the bottleneck takes place during postnatal folliculogenesis and not during oogenesis (Wai *et al.*, 2008).

The disease-associated mutations that are transmitted from one generation to the next are either recent and deleterious or ancient and adaptive, affecting individuals by predisposing them to disease through exposure to environmental factors. Mitochondrial mutations can also be age-related through the accumulation of somatic mtDNA mutations over time, which leads to delayed onset of disease. Mitochondrial disease is either caused by different mutations in the same gene or by single heteroplasmic mutations. (Wallace *et al.*, 1988; Wallace, 1995).

Mutations are caused by substitution or by rearrangement, deletion or insertion events. Point mutations are usually transmitted from mother to child and can occur either in the protein-coding regions of the mtDNA genome or in tRNA or rRNA genes. Most of the mtDNA-related diseases are, however, caused by point mutations on the mitochondrial tRNA genes because these mutations will affect the overall mitochondrial translation system and therefore disrupt the functioning of the OXPHOS system (Wallace, 1994). Examples of diseases of substitution mutations are LHON, Leigh syndrome, myoclonic epilepsy and ragged red fibre disease, mitochondrial encephalomyopathy, lactic acidosis and stroke-like episodes, mitochondrial myopathy and exercise intolerance, chronic progressive external ophthalmoplegia (CPEO), Kearns-Sayre syndrome (KSS), gastrointestinal syndrome, dystonia, diabetes, deafness, cardiomyopathy, renal failure, Alzheimer’s disease and Parkinson’s Disease.

Most of the rearrangement mutations consist of large deletions that can span many genes and in general will occur between the origins of replication. Diseases in this category include maternally inherited diabetes and deafness, CPEO, KSS and Pearson marrow/pancreas syndrome (Wallace and Fan, 2010).

Mutations in the nDNA affecting the bioenergetics of the cell will also affect the mitochondria by altering the production of ATP, acetyl-CoA, SAM and NADH. Disease can also be caused when the coordination between the bioenergetics genes (nDNA and mtDNA) are compromised through deficient *cis* and *trans* regulation. Mutations accumulate with age in the cells of the body and are the reason why mitochondrial mutations are linked to many age-related diseases and even to some cancers (Ruiz-Pesini *et al.*, 2007). Mutations in the nDNA that affect the OXPHOS complex lead to multisystem disorders through inactivation of structural and assembly OXPHOS genes, unstable mtDNA and inefficient mitochondrial fusion and fission (Wallace and Fan, 2010).

Many of the epigenetic diseases are also linked to mitochondrial dysfunction. Cancer, for example, is linked to a modulated mitochondrial OXPHOS due to modifications of the chromatin and the methylation patterns of certain cancer cells. Abnormalities in mitochondrial energy metabolism are also likely to play a role in causing neurological disorders such as Fragile X syndrome and Rett syndrome, while some forms of autism and genetic errors in imprinted loci are linked to diseases such as Prader-Willi and Angelman syndromes (Wallace and Fan, 2010).

3.8.1 Mitochondrial haplogroups and diseases

Several connections have been made between mitochondrial haplogroups and various clinical conditions. The LHON pathogenic mutations have displayed a high penetrance within individuals that belong to haplogroup J (Brown *et al.*, 1994; Torroni *et al.*, 1997). Haplogroup J further displays an association with longevity in Europeans (De Benedictus *et al.*, 1999; Rose *et al.*, 2001), a decreased risk of developing Parkinson's disease (Van der Walt *et al.*, 2003) and an increased risk of developing diabetes in individuals of European descent (Crispim *et al*., 2006). In contrast to haplogroup J, haplogroup H is reported to be associated with a decreased risk of developing Parkinson's disease but an increase in the risk of the development of Alzheimer's disease. Alzheimer's disease has, however, displayed a lower risk of development in individuals who belong to haplogroups U and T (Chagnon *et al.*, 1999; Carrieri *et al.*, 2001). Haplogroup H has further also been

associated with a reduced risk of macular degeneration and protection against sepsis (Baudouin *et al.*, 2005; Jones *et al.*, 2007). Associations between mtDNA gene mutations and certain types of cancers have also been established (Wallace, 1995). Mitochondrial haplogroups have also been connected to physiological differences between individuals, as demonstrated by the different haplogroup distributions between Finnish sprinters and long-distance runners (Niemi *et al*., 2005) and sperm motility based on the mitochondrial energy output in the sperm mid-piece (Ruiz-Pesini *et al*., 1998).