

Michelle Chantel Livesey
3339400

December 2019

The geographical analysis of the mitochondrial DNA
control region found in southern African
population's by using a bioinformatics approach

Michelle Chantel Livesey



**UNIVERSITY of the
WESTERN CAPE**

*A thesis submitted in partial fulfilment of the requirements for
the degree of Magister Scientiæ in the Department of
Biotechnology, University of the Western Cape.*

Supervisor: Prof. Maria Eugenia D'Amato
Co-Supervisor: Dr. Peter Gustav Ristow
Forensic DNA Laboratory, Department of Biotechnology

The geographical analysis of the mitochondrial DNA control region found in southern African population's by using a bioinformatics approach.

Keywords (6)

Mitochondrial DNA

Control region

Haplogroups

EMPOP

Ancestry

South Africa



ABSTRACT

South Africa's demographic complexity has been historically shaped by interethnic admixture between the native KhoiSan inhabitants and the Bantu expansion that started 3000 - 5000 YBP and reached South Africa about 1000 years ago. This, followed by the arrival of the European settlers in the 16th century who brought in slaves from their Asian colonies, mainly from Malaysia and the east- and western Africa. By the late 19th- and early 20th century colonial India (South Asia) also migrated to South Africa.

It has been hypothesized that the history of South African has influence ethnic group distribution in South African, to the degree that it is not homogeneous. This study further theorized that the history contributed to the gene flow at a regional level and not restricted to the ethnic distribution of genetic variation. Therefore, this study focuses on the geographical and ethnic dispersal of maternal lineages by investigating the mitochondrial DNA (mtDNA). It is predicted that the maternal lineages could be limited or appear in high frequencies in specific ethnic and/or geographical region in South Africa. This, in turn, can aid mitochondrial DNA forensic human identification applications when a specific haplotype is questioned.

This study produced a complete mitochondrial DNA control region (nucleotide position 16 024-576) sequences generated with Sanger and next-generation sequencing for 246 individuals residing in the Western Cape, Northern Cape and KwaZulu-Natal provinces of South Africa. The control region is of particular interest due to the vast majority of rapidly evolving sites that are of relevance in both forensic genetics and population studies. The haplotypes were inferred against the revised Cambridge Reference Sequences and haplogroups were determined by online tool HaploGrep 2.

The project aimed to generate high-quality forensic mtDNA haplotypes to enlarge the forensic-quality population reference database EMPOP with matrilineal lineages of self-declared Coloured, Griqua, Bantu and Nama populations. The project further aimed to assess both the phylogenetic and phylogeographical analysis of the South African populations.

The dataset showed a total of 104 haplotypes characterized by 136 polymorphic sites, of which 68 haplotypes were unique. A high frequency of maternal haplogroup L0 were observed, indicating that the most influential gene flow within the South African population is of the KhoiSan. The Coloured and Griqua showed high levels of admixture, while the native Bantu and Nama only displayed Sub-Saharan African lineages. However, $n=1$ individual in the Nama population illustrated Asian-ancestry. Further, at a population point of view, only one shared haplotype linked to haplogroup L0d2a1 were identified between all four population groups.


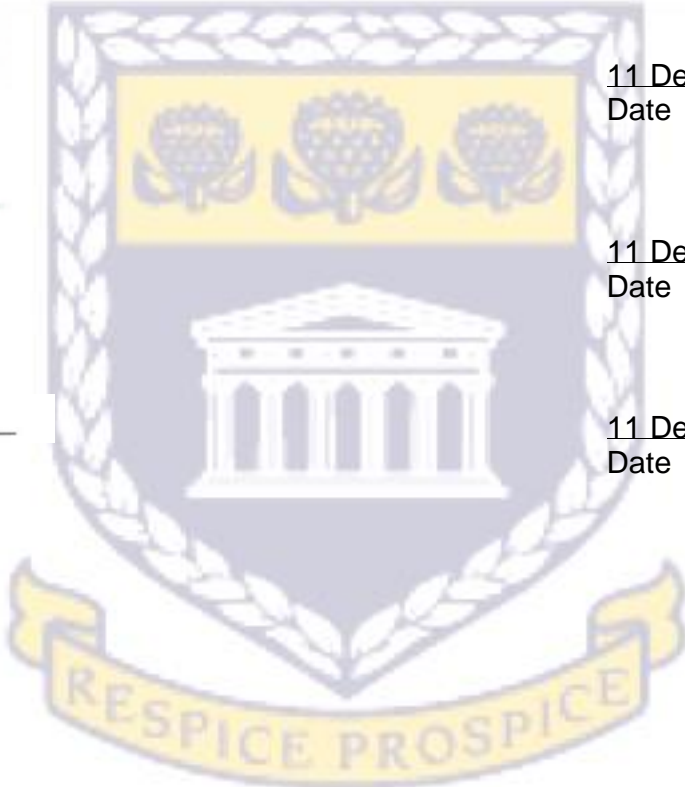

Hence, the Coloured, Griqua, Bantu and Nama experienced gene flow. While, the phylogeographical analysis showed four haplotypes linked to haplogroup L0d2a1, L0d1b2b and L0d1b2b2b1 were shared amongst the three provinces.

The availability of the data and assessments in this research project has the potential to aid in the assessment of rare mitochondrial DNA haplotype and haplogroup frequencies that is significant in forensic applications. Therefore, the mtDNA data presented can be utilized as a valuable tool in routine forensic casework.



DECLARATION

I, **Michelle Chantel Livesey** declare that the work presented here, **“The geographical analysis of the mitochondrial DNA control region found in southern African population’s by using a bioinformatics approach”** is my own work, that it has not been submitted for any degree or examination in any other university, and that all the sources I have used or quoted have been indicated and acknowledged by complete references.

 ----- Supervisor Prof ME D'Amato		<u>11 December 2019</u> Date
 ----- Co-supervisor Dr. PG Ristow		<u>11 December 2019</u> Date
 ----- MSc. Student Ms. MC Livesey		<u>11 December 2019</u> Date

UNIVERSITY of the
WESTERN CAPE

ACKNOWLEDGEMENTS

First, I would like to acknowledge our Heavenly Father who gave me the strength to persevere and see this project through.

I would like to extend a special word of thanks to my supervisor, **Prof. ME D'Amato**, for allowing me the opportunity to pursue an MSc. I appreciate the guidance and support you offered to help me grow as a research student. Your deep insight helped me at various stages of my research.

To my co-supervisor, **Dr. PG Ristow**, you are a true inspiration. Thank you for your continuous guidance and conscientious suggestions throughout this study.

To the rest of the laboratory members and technical staff, thank you for all the laughter shared between us. I sincerely appreciate that you were always there with a listening ear and encouraging words.

I wish to thank my friends and family for keeping me grounded, motivated and inspired. My parents, **JJ Livesey and LJ Livesey**, thank you for your endless love and support through the years. I consider myself the luckiest in the world to have such a supportive family, standing behind me.

I also want to thank the **National Research Foundation** of South Africa and **Ernst & Ethel Eriksen Trust** for providing financial support to undertake this research.

UNIVERSITY of the
WESTERN CAPE

LIST OF ABBREVIATIONS

3'	-	3 prime
5'	-	5 prime
°C	-	Degrees Celsius
B.C	-	Before Christ
bp	-	Base pair
C	-	Cytosine
C.E.	-	Common Era
CODIS	-	Combined DNA Index System
CRS	-	Cambridge Reference Sequence
D-Loop	-	Displacement loop
dATP	-	Deoxyadenosine triphosphate
DC	-	Discrimination capacity
dCTP	-	Deoxycytidine triphosphate
ddATP	-	Di-deoxyadenosine triphosphate
ddCTP	-	Di-deoxyadenosine triphosphate
ddGTP	-	Di-deoxyguanosine triphosphate
ddNTPs	-	Di-deoxynucleotidetriphosphate
ddTTP	-	Di-deoxythymidine triphosphate
dGTP	-	Deoxyguanosine triphosphate
DNA	-	Deoxyribonucleic acid
dTTP	-	Deoxythymidine triphosphate
Excl.	-	Excluding
FBI	-	Federal Bureau of Investigation
G	-	Guanine
H-strand	-	Heavy-strand
Hg	-	Haplogroup
HV(R)	-	Hypervariable (region)
IUPAC	-	International Union of Pure and Applied Chemistry
kya	-	Thousand Years Ago
KZN	-	KwaZulu-Natal
L-strand	-	Light-strand
mRNA	-	Messenger RNA
mtDNA	-	Mitochondrial DNA
<i>n</i>	-	Number/total
NC	-	Northern Cape
NFDD	-	National Forensic DNA Database of South Africa
No.	-	Number
O _H	-	Origin of replication
PCR	-	Polymerase Chain Reaction
rCRS	-	revised Cambridge Reference Sequence
RFLP	-	Restriction Fragment Length Polymorphism
rRNA	-	Ribosomal ribonucleic acid
SAPS	-	South African Police Service
SNP(s)	-	Single Nucleotide Polymorphism(s)
T	-	Thymine
tRNA	-	Transfer RNA

- U - Uracil
- VOC - Vereenigde Oostindische Compagnie



LIST OF FIGURES

Figure 1.1. Bantu expansion migration routes in Africa. Two routes (purple and red arrow) show the direction traveled from the original homeland.	2
Figure 1.2: Migration pattern of descendants from the Dutch East Indian Company and British settlers in southern Africa. The map of South Africa pre-1994 shows two immigrant population groups invading the territory of indigenous southern African populations.	4
Figure 1.3: Map illustrating the relocation of the Griqua population throughout South Africa from the late 18 th to the 20 th century (Heynes, 2015).	5
Figure 1.4: Nine provinces of South Africa as established in 1994 after the abolishment of the apartheid era. The figure right below shows the geographical location of South Africa on the World Map (Created in Tableau Public: https://public.tableau.com/en-us/s/).	7
Figure 1.5: Inheritance pattern of lineage markers; Y-chromosome and mitochondrial, along with autosomal DNA. The lineage markers present a strict paternal and maternal inheritance (Butler, 2012).	10
Figure 1.6: Single Nucleotide Polymorphism. Two variant alleles are displayed at one position by an arrow, the last (sixth) position in sequence 1 is a cytosine while in sequence 2, the same position consists of a Guanine.	11
Figure 1.7: Circular structure of the human mitochondrial DNA. The mtDNA includes the coding and non-coding region which holds hypervariable regions I, -II and -III.	12
Figure 1.8: Mitochondrial DNA phylogenetic tree build 17 (18 Feb 2016), divided into 25 subtrees available via links on phylotree.org/tree/index.htm . The mitochondrial most recent common ancestor (mt-MRCA) is used to root the tree. Accumulation of polymorphisms divides individuals into different haplogroups over time (van Oven & Kayser, 2009).	14
Figure 1.9: Map displaying the early diversification of modern human. Beginning with mitochondrial Eve and the movement throughout Africa (Hg L0; L1-6) and MacroHg M and N (Macauley <i>et al.</i> 2005).	15
Figure 1.10: Four steps of Sanger sequencing to obtain a DNA sequence of interest. The four ddNTPs are detected by four distinct coloured fluorescent dyes attached to each ddNTP. The ddTTP (thymine), ddCTP (cytosine), ddATP (adenine) and ddGTP (guanine) are labeled with red, blue, green and yellow, respectively. However, for easy visualization, the ddGTP is generally displayed in black (Murphy <i>et al.</i> , 2005).	17
Figure 1.11: Electropherogram generated by Sanger sequencing of the mitochondrial DNA. Each respective colour fluorescent dye (i.e. C = blue, A = green, T = red and G = black) indicates the respective nucleotide in the sequence.	17
Figure 1.12: Visual representation of the revised Cambridge Reference Sequence (rCRS) at site 73 (in HVII), compared to that of four individuals' sequences (Sample_01 - 04). The rCRS can be observed in the first line, showing an A (adenine) at site 73. The same is seen for the following two individuals (Sample_01 - _02), while Sample_03 - _04 holds a guanine.	20
Figure 1.13: Good quality sequence compared to that of a poor-quality sequencing that results in the C-stretch. Mitochondrial DNA sequence (A) is absent of the HVR I C-stretch (consists of 16 189T) and (B) illustrates a C-stretch. It can be noted that the effect of the transition on sequencing results downstream of the C-stretch and the quality rapidly declines after the series of cytosine residues (Butler, 2012).	22
Figure 1.14: Electropherogram of mtDNA positions 16 086 – 16 101 of two individuals to compare the presence and absence of heteroplasmy. An electropherogram of sequence heteroplasmy detected at a sole position 16 093, (A) occupying both nucleotides C and T compared to the same region of a different sample (B) consisting of a T. The heteroplasmy site 16 093 retaining both C and T nucleotides (Butler, 2012). (IUPAC codes label this heteroplasmy by a 'Y' as seen in Table 1.2).	23
Figure 3.1: Good quality chromatogram of Sanger sequenced mitochondrial DNA. A specifically amplified product, that shows good and clean peak heights (Heynes, 2015).	33

Figure 3.2: Fragment position(s) on the mitochondrial genome that was subjected to resolve the discrimination capacity. The various region(s) examined in the mitochondrial genome includes HVR I, HVR I and -II, HV I to -II and the whole control region which includes HVR III.....34

Figure 3.3: The discrimination capacity of the various region(s) of the mitochondrial DNA for human identification in forensic applications. The number of haplotypes (*n*) identified in the fragment(s) of the mtgenome.35

Figure 3.4: Three geographical locations in which a total of *n*=246 biological samples were collected. This includes individuals residing in the town of Kokstad, KwaZulu-Natal (-30.5096°S, 29.4063°E), Sanddrift, Northern Cape (-28.41°S, 16.77491°E) and Vredendal in the Western Cape (31.6391°S, 18.5285°E) province of South Africa (Created in Tableau Public:<https://public.tableau.com/en-us/s/>).41

Figure 3.5: Bar graph depicting the mitochondrial haplogroup composition of the Western Cape, Northern Cape and KwaZulu-Natal province. This is an indication of the various haplogroups in different frequencies found in all three studied provinces of South Africa.47

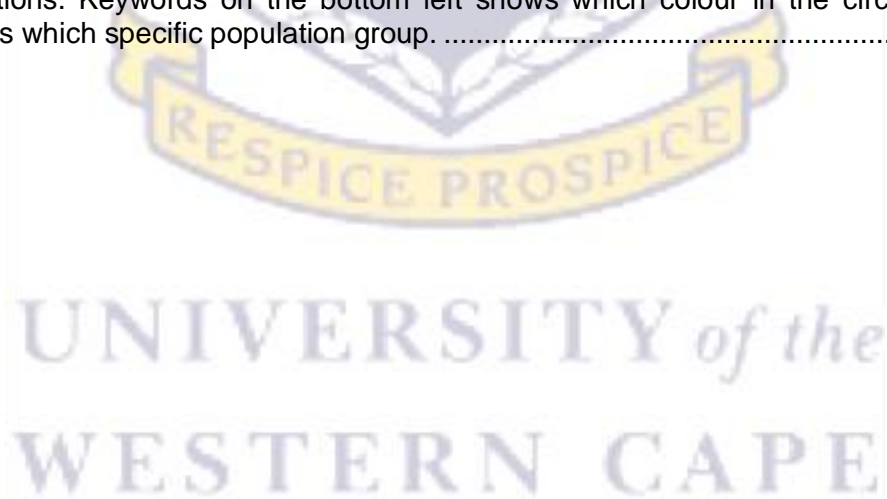
Figure 3.6: Bar graph depicting the mitochondrial haplogroup composition of all the South African population groups (*n*=246). This is an indication of the various haplogroups present in the Coloured, Griqua, Bantu and Nama population at different frequencies.52

Figure 3.7: Venn Diagram illustrating the shared and unique haplotypes among the Coloured, Griqua, Bantu and Nama populations. (Created with: <http://bioinformatics.psb.ugent.be/webtools/Venn/>).55

59

Figure 3.8: Gradient map illustrating the mtDNA distribution of macro-haplogroup L0 in South Africa representing 177 individuals. A colour scale at the right side of the figure defined the percentage of individuals observed in the area (Created with: MapViewer® Golden Software, LLC).59

Figure 3.9. Median-joining network analysis of the major L0 maternal haplogroup in the four populations. Keywords on the bottom left shows which colour in the circle diagrams displays which specific population group.61



LIST OF TABLES

Table 1.1: Contrast of the human mitochondrial - and nuclear DNA (Butler, 2012).....	9
Table 1.2: IUPAC codes for base calling when a site has more than one nucleotide i.e. heteroplasmy. R should signify a mixture of G and A, while Y should denote a mixture of T and C. The code uses capital letters, as small letters describe mixtures of deleted/undelated and inserted/non-inserted bases (Parson <i>et al.</i> , 2014; Stothard, 2000).	21
Table 2.1: Population groups selected for the study includes the Coloured, Griqua, Nama and Bantu. The Bantu in this study was made up of individuals who classified themselves as Black, Pedi, Tswana and Xhosa on the consent form.	27
Table 2.2: Analysis methods employed to produce the mtDNA population data.	28
Table 3.1: MtDNA haplogroup frequencies obtained in this project.	37
Table 3.2: MtDNA haplogroup frequencies of Coloured individuals in the Western Cape. The data set of this study was limited to the 24 self-declared Coloured individuals residing in the Western Cape to compare to a study conducted by Quintana-Murci <i>et al.</i> (2010) on the identical population and geographical region.....	38
Table 3.3: MtDNA haplogroup frequencies of Coloured individuals in the Northern Cape. The data set of this study was limited to the 73 self-declared Coloured individuals residing in the Northern Cape to compare to a study conducted by Schlebusch <i>et al.</i> (2011) on the identical population and geographical region.....	38
Table 3.4: Random match probability of the South African ethnic groups.	39
Table 3.5: Shared haplotypes identified in the human mtDNA amongst the 3 geographical locations. The frequency column shows the frequency observed in the Western Cape (WC), Northern Cape (NC) and KwaZulu-Natal (KZN) province as obtained in the program Arlequin v. 3.5.2.2.	44
Table 3.6: Haplotypes that were only identified at $n \geq 3$ in one of the three provinces of South Africa. The last column shows the province (i.e. WC = Western Cape, NC = Northern Cape and KZN = KwaZulu-Natal) in which the haplotype was identified as obtained in the program Arlequin v. 3.5.2.2.	45
Table 3.7: An in-depth look at the out of Africa haplogroups found in the Western Cape province.....	48
Table 3.8: An in-depth look at the out of Africa haplogroups found in the Northern Cape province.....	49
Table 3.9: An in-depth look at the out of Africa haplogroups found in the KwaZulu-Natal province.....	50
Table 3.10: Non-hierarchical AMOVA analysis of the four population groups of South Africa. Each group: Coloured, Griqua, Bantu and Nama individuals are classified as a population on its own. The results were obtained by Arlequin v.3.5.2.2.	54
Table 3.11: A descriptive statistic for four populations from South Africa. Polymorphism overview of each population group using the individuals control region sequences.	56
Table 3.12: Neutrality test of Tajima's D and Fu & Li's F according to each population group. The data presented in this table was inferred by DNAsp v5.10.01. Each statistical value is accompanied by its P-value illustrated directly to the right of the respective neutrality test.....	58

SUPPLEMENTARY DATA

Table A1: Shared haplotypes amongst the three geographical regions in this study with the number (n) of individuals as per ethnic group identified in the respective provinces. The haplotype profiles are in the same order as shown in Table 4.1. The ethnic groups are represented by the Coloured (COL), Griqua (GRI) and Bantu (B).73

Table A2: All 246 mtDNA samples classified into haplogroups according to each respective sample's SNPs. The first column shows the samples' name as given by the Forensic DNA Laboratory. This is followed by the haplogroup of each sample inferred by HaploGrep 2, which assign a haplogroup according to the SNPs as shown in the last column.....74

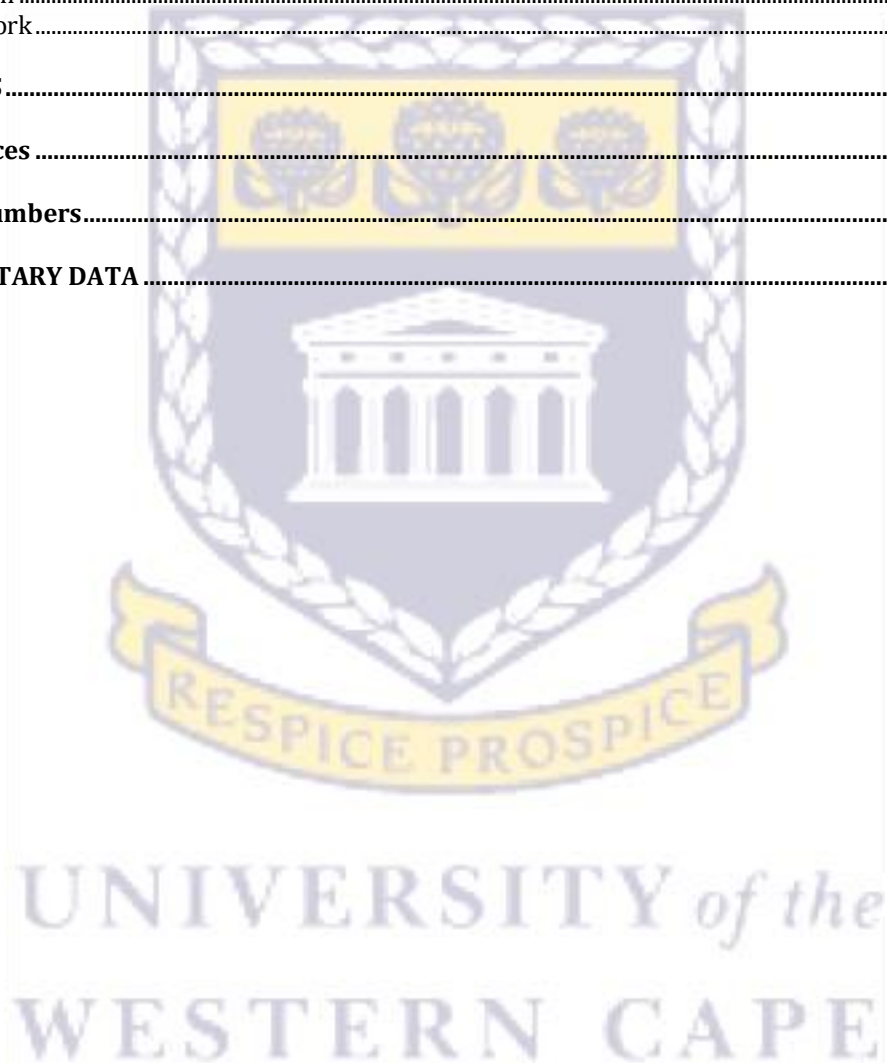


TABLE OF CONTENT

ABSTRACT.....	iii
DECLARATION.....	v
ACKNOWLEDGEMENTS.....	vi
LIST OF ABBREVIATIONS.....	vii
LIST OF FIGURES.....	ix
SUPPLEMENTARY DATA.....	xii
TABLE OF CONTENT.....	ix
CHAPTER 1: LITERATURE REVIEW.....	1
1.1. THE CREATION OF SOUTH AFRICA.....	1
1.1.1. Southern Africa’s first people: KhoiSan.....	1
1.1.2. Bantu expansion.....	2
1.1.3. Early colonization.....	3
1.1.4. Admixture: The effect of KhoiSan and immigrated populations.....	4
1.1.4.1. Griqua.....	4
1.1.4.2. Coloured.....	6
1.1.5. The Republic of South Africa.....	6
1.2. THE AIMS OF THE PROJECT.....	7
1.3. GENETIC ANCESTRY ANALYSES USING UNIPARENTAL INHERITANCE.....	8
1.3.1. Characteristics of mitochondrial DNA.....	8
1.3.1.1. Haploid transmission of mitochondria cells.....	9
1.3.2. Lineage markers.....	10
1.3.2.1. Single nucleotide polymorphism.....	10
1.4. MITOCHONDRIAL DNA STRUCTURE.....	11
1.4.1. Control region.....	11
1.4.2. Mitochondrial DNA coding region.....	12
1.5. MITOCHONDRIAL DNA HAPLOGROUPS.....	13
1.5.1 Haplogroup structure.....	13
1.5.2. Classification of haplogroup.....	13
1.5.2.1. Geographical origins based on haplogroups.....	14
1.5.3 Southern African population haplogroups.....	16
1.6. STANDARD MTDNA SEQUENCING IN FORENSIC CASEWORK.....	16
1.6.1. DNA sequencing.....	16
1.6.1.1. Sanger sequencing.....	16
1.6.1.2. Next-generation sequencing.....	18
1.7. REPORTING MITOCHONDRIAL DNA VARIATION.....	18
1.7.1. History of revised Cambridge Reference Sequence.....	19
1.8. MITOCHONDRIAL DNA TYPING.....	19
1.8.1. Data assessment and editing.....	19
1.8.1.1. Nomenclature.....	20
1.8.2. Issues impacting interpretation.....	21
1.8.2.1 Sequencing beyond polymeric C-stretches.....	21
1.8.2.2. Heteroplasmy.....	22
1.9. LOCAL CRIMINAL DATABASE.....	23
1.10. SIGNIFICANCE OF THIS STUDY.....	24
1.10.1. Population study for the participating individuals.....	24
1.10.1.1. Symbolism and belonging.....	24
1.10.2. MtDNA in Forensic applications.....	25

1.10.2.1. EDNAP mitochondrial DNA population database.....	26
1.10.2.1.1. EMPOP problem statement.....	26
CHAPTER 2: METHODS.....	27
2.1. ETHICS STATEMENT.....	27
2.2. STUDY AREA.....	27
2.3. BIOLOGICAL SAMPLE COLLECTION.....	27
2.4. METADATA AND SELECTION OF PARTICIPANTS.....	27
2.5. DATA ASSEMBLY.....	28
2.5.1. MtDNA consensus sequences generated.....	28
2.5.2. SNP haplotype analysis.....	29
2.5.3. Haplogroup assignment.....	29
2.6. QUALITY CONTROL WORKFLOW.....	29
2.6.1. Data submission to EMPOP.....	29
2.6.1.1. EMPOP Quality Control (QC).....	29
2.7. FORENSIC PARAMETERS ANALYSIS.....	30
2.7.1. Discrimination capacity of mtDNA fragments.....	30
2.7.2. Random Match Probability.....	30
2.8. POPULATION GENETIC ANALYSIS.....	30
2.8.1. Geographical population structure.....	30
2.8.2. Population diversity.....	31
2.8.3. Degree of differentiation in South African sample-set.....	31
2.8.4. Geographic distribution of indigenous maternal macro-haplogroup L0.....	31
2.8.5. Phylogenetic network.....	31
CHAPTER 3: RESULTS AND DISCUSSION.....	33
3.1. FORENSIC MTDNA CONTROL REGION.....	33
3.1.1. Forensic parameters associated with forensic mtDNA.....	33
3.1.1.1 Haplotype analysis for forensic applications.....	33
3.1.1.1.1. Haplotype analysis of Hypervariable I.....	35
3.1.1.1.2 Haplotype analysis of Hypervariable I and II.....	35
3.1.1.1.3. Haplotype analysis of Hypervariable I to II.....	35
3.1.1.1.4. Haplotype analysis of the whole control region.....	36
3.1.2. Frequency estimates of a database.....	36
3.1.3. Random match probability.....	39
3.2. POPULATION GENETICS OF MTDNA CONTROL REGION.....	41
3.2.1. GEOGRAPHICAL STRUCTURE OF MTDNA.....	41
3.2.1.1. Sample collection coordinates.....	41
3.2.1.1.2. Comparative structure analysis of 3 geographical locations.....	42
3.2.1.1.2.1. Mitochondrial DNA haplotype identification.....	42
3.2.1.2. The Maternal haplogroup composition of South African provinces.....	46
3.2.1.2.1. Haplogroup composition of the Western Cape.....	46
3.2.1.2.2. Haplogroup composition of the Northern Cape.....	48
3.2.1.2.3. Haplogroup composition of KwaZulu-Natal.....	49
3.2.2. COMPARATIVE STRUCTURE ANALYSIS AT AN INTRA-POPULATION LEVEL.....	51
3.2.2.1. The maternal haplogroup composition.....	51
3.2.2.1.1. Admixed: Coloured and Griqua.....	51
3.2.2.1.2. Native: Bantu and Nama.....	53
3.2.2.2. South Africa population differentiation.....	53
3.2.2.2.1. Shared and unique haplotypes in South African populations.....	55

3.2.2.3. Haplotype and nucleotide diversity	56
3.2.2.3.1. Polymorphic overview of the Coloured	56
3.2.2.3.2. Polymorphic overview of the Griqua	57
3.2.2.3.3. Polymorphic overview of the Bantu	57
3.2.2.3.4. Polymorphic overview of the Nama	57
3.2.2.4. Neutrality tests	57
3.2.3. STRUCTURE BY MATERNAL HAPLOGROUP L0.....	58
3.2.3.1. Distribution of the indigenous people.....	58
3.2.3.2. A network of indigenous haplogroup L0.....	60
CHAPTER 4: CONCLUSION AND FUTURE WORK.....	62
4.1. Conclusion	62
4.2. Future work.....	63
REFERENCES	64
Web Resources	72
Accession Numbers.....	72
SUPPLEMENTARY DATA.....	73



CHAPTER 1: LITERATURE REVIEW

1.1. THE CREATION OF SOUTH AFRICA

1.1.1. Southern Africa's first people: KhoiSan

Southern Africa consist of populations with significant human genomic variations (Ingman *et al.*, 2000). DNA variation patterns place modern human ancestries within Africa, with the most diverse existing lineages established in the native KhoiSan inhabitants of southern Africa. The term KhoiSan is used to refer to two major ethnic groups, the Khoi (pastoralists) and San (hunter-gatherers) in southern Africa. The KhoiKhoi ethnic group is largely made up of the Nama people who originated in the Northern Cape province of South Africa. (Barnard, 1992).

The Nama population are one of the oldest indigenous groups in Namibia. Historically they are known as the "Namaqua", while the early colonialists referred to the individuals as Hottentots. During 1904 – 1908, the Nama population endured a Namaqua genocide under the German Empire in which the population experience a large loss of life. The individuals were further displaced from their agricultural land by the colonialists. In 1991, a section of Namaqualand was named Richtersveld National Park. Only recently, December 2002, were the ancestral lands returned to the community, and today it is one of the few places where the native Nama traditions survive (Barnard, 1992).

This tribe, with several sub-tribes and the San people, are collectively known as the KhoiSan. For thousands of years, the KhoiSan of South Africa and southern Namibia maintained a nomadic life. The KhoiSan people are the indigenous population of southern Africa as it predates the Bantu expansion. Numerous other names exist for either one or both of these groups, such as the Khoi, Khoe, Khoi-San, Khoe-San, Bushmen and Hottentots (Barnard, 1992; Behar *et al.*, 2008).

The KhoiSan is further characterized by their heavy use of click consonants in their languages (Güldemann, 1997). The population use to represent a wide geographical distribution that reached Africa's utmost southern tip. Nowadays, the modern KhoiSan is mainly restricted to the greater Kalahari regions of Namibia and Botswana. The individuals of this population currently represents a group of isolated subpopulations with declining numbers and subpopulation extinctions (Barnard, 1992; Mitchell, 2002).

1.1.2. Bantu expansion

The African genetic and cultural landscape was immensely influenced by an event known as the Bantu expansion. It was proposed that the Bantu languages and Bantu-speaking people dispersed from their ancestral homeland in the Grassfield region (between southeast Nigeria and western Cameroon) at about 1000 B.C, into southern Africa. Two main routes were taken from the following starting points: a western and eastern route illustrated in Figure 1.1.

The figure shows the western route (red arrow), migrated through the west coast of Africa, arriving in Angola, Botswana and South Africa around 3.5 thousand years ago (kya). The eastern route (purple arrow), travelled towards the Great Lakes, reaching the region of Uganda almost 2.5 kya, and further expanded into the south, reaching Mozambique by approximately 1.8 kya (Newman, 1995; Phillipson, 2005).



Figure 1.1. Bantu expansion migration routes in Africa. Two routes (purple and red arrow) show the direction traveled from the original homeland.

Individuals of dark-skin in comparison to other population groups, or termed black people are currently known as Bantu in South Africa. This includes several groups of Bantu speakers. The two main groups that emerged during the southward migration into South Africa are the Nguni (Xhosa, Zulu, Ndebele and Swazi), who belonged to the eastern branch, and the Sotho-Tswana, residing in the interior plateau. The western branch of

Bantu-speakers reached the north of Namibia, however, their spread further south was halted by the Khoer herders (Ehret and Posnansky, 1982).

1.1.3. Early colonization

In the 1600s, the indigenous communities of southern Africa were invaded by the arrival of the European settlers. By 1652 the Dutch established the East India Company, better known as VOC (short for Vereenigde Oostindische Compagnie) on the land of the KhoiKhoi. A decade after their arrival the Dutch further brought in slaves from their Asian colonies, mainly from Malaysia and from east- and western Africa. By 1795 the British took over the Cape Colony and became a British colony. However, by 1820 the British settlers lived mainly in the eastern Cape (Thompson, 2006).

The first European settlers moved from the coastal area into the interior over the years, and by the late 1800s they had managed to control all the land that had previously belonged to the African people (Thompson, 2006). Meanwhile, due to a more deliberate emigration to escape British colonialism, the descendants from the VOC also moved from the Cape Colony into the interior by 1836 as seen in Figure 1.2. This event in which the Boer pastoralists known as the Voortrekkers (an Afrikaans and Dutch word for pioneers) migrated eastward, is referred to as the Great Trek, which resulted in the establishment of two republics, the Orange Free State and the Transvaal (Thompson, 2006).

On the border of the Orange Free State, the Boer acquired land from the Griqua population (discussed under *section 1.1.4.1*) to establish the first town, known as Griquatown, north of the Orange River (Thompson, 2006).

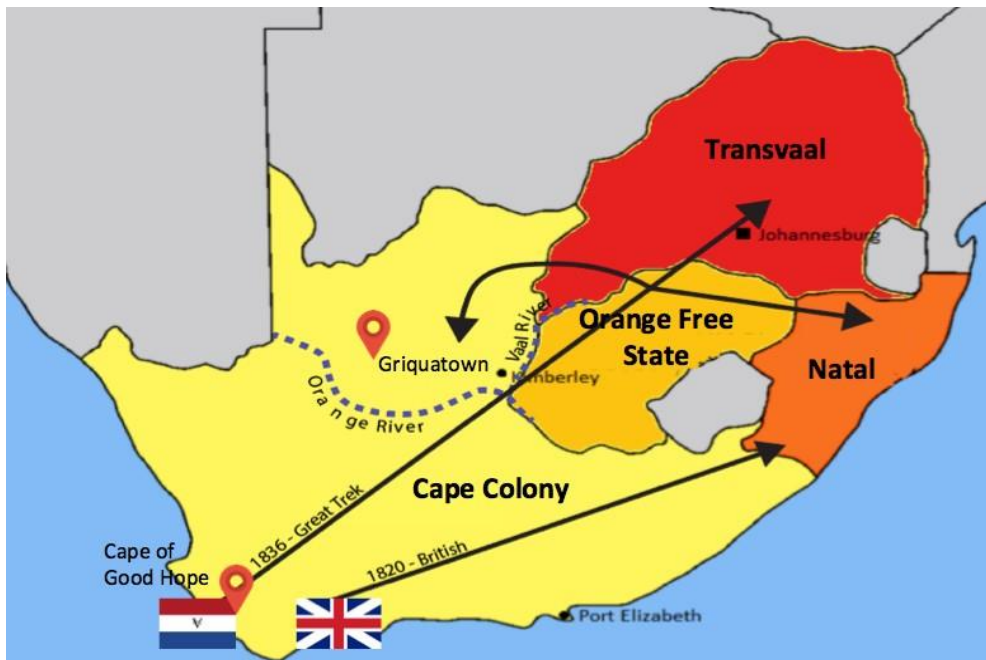


Figure 1.2: Migration pattern of descendants from the Dutch East Indian Company and British settlers in southern Africa. The map of South Africa pre-1994 shows two immigrant population groups invading the territory of indigenous southern African populations.

The discovery of diamonds in 1867 near Griquatown, led to the arrival of European prospectors, mainly from Britain and attracted southern African migrant workers. Determined to extend their control to the new Boer republics of the Dutch settlers, the British ultimately engaged the Dutch in the Anglo-Boer War. The battle lasted from 1899 - 1902 and ended with the British being victorious on the battlefield. Nonetheless, peace was achieved, which ensured the unity of the Europeans and the indigenous southern Africans (Thompson, 2006).

In 1909, an agreement was reached between the British and the Boers, which established a single country by combining the territories controlled by each group into one nation. Thus by 1910, the Union of South Africa was established, known today as South Africa (Thompson, 2006).

1.1.4. Admixture: The effect of KhoiSan and immigrated populations

1.1.4.1. Griqua

The Griqua people are of mixed white and KhoiKhoi ancestry, with a unique origin in the early history of the Cape Colony. Originally, known as the Bastards or Basters to the colonialists who forcefully removed them from their homeland (Saunders & Southey, 2001; Balson, 2007). The first Captain of this population, Adam Kok I (a former slave), lead the

Griqua's north from the interior to escape the discrimination. Ultimately, this population was led beyond the Cape Colony, near the Orange River, just west of the Orange Free State, and on the southern skirts of the Transvaal (Saunders, 1983).

By 1801, the group moved further north to an area called Griqualand West and later renamed to Griquatown ("Griquatown"), a region in the Northern Cape province of South Africa (Waldman, 2007). Here, the Griqua joined indigenous populations such as the Korana and other KhoiSan descendants (Besten, 2006). Griquatown continued to grow, and by the year 1823 also consisted of Sotho and Tswana heritage. These multiple ethnic groups allowed for diverse social-cultural and linguistic exchanges.

The Griqua population further encountered various ethnic groups throughout South Africa as the population relocated. Figure 1.3, shows the Griqua trek through Griquatown (1800), Campbell (1805), Philippolis (1823), Kokstad in the KwaZulu-Natal province (1870) and finally settling in Kranshoek located in the Western Cape province (1930) (Heynes, 2015). Due to the admixture that was contributed to the Griqua population during the trek, the people were marginalized and merged as Coloured during the apartheid era.

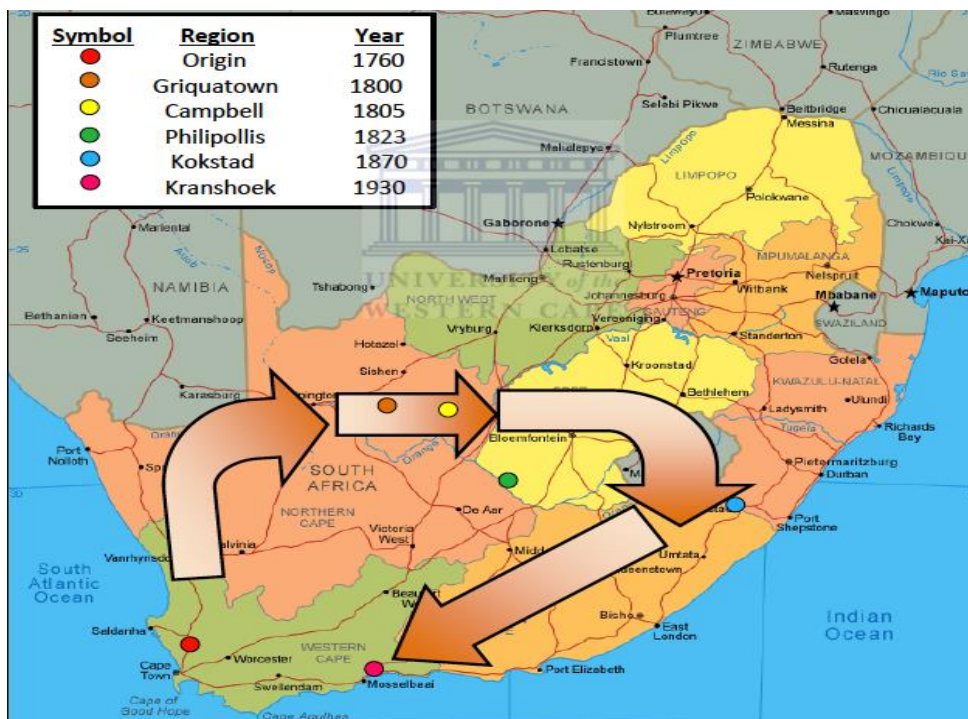


Figure 1.3: Map illustrating the relocation of the Griqua population throughout South Africa from the late 18th to the 20th century (Heynes, 2015).

1.1.4.2. Coloured

The Coloured population is native to Southern Africa, entailing more than one of the numerous population groups inhabiting the region. The complex genomic admixture in the Coloured population is a result of the social and demographical events that provoke population dispersal and isolation between populations in South Africa (Petersen *et al.*, 2013). The multi-ethnic heritage of the Coloured includes KhoiSan, Bantu, Afrikaner (descended predominantly from Dutch settlers), English (originating mainly from the British settlers) and East - or South-Asian.

Historically, the origins of this population can be traced to Table Bay shores, an area currently known as Cape Town in the Western Cape province. This occurrence was dated shortly after the VOC was established. Therefore, the Coloured population is formally known as the Cape Coloured and forms a minority group within South Africa, with the Western Cape province remaining home to most of the population (Petersen *et al.*, 2013; Statistics South Africa, 2012).

A related subgroup of the Coloured is termed “Cape Malay”. The Cape Malay is currently a large community and generally marry amongst themselves for religious purposes. The Muslim religion is the one main legacy kept by the individuals from their East-Asian origins.

1.1.5. The Republic of South Africa

South Africa is situated at the southernmost tip of the African continent, measuring at 1.2 million square kilometres. The geographically strategic location of South Africa has been the object of battles fought between European invaders and the indigenous Africans for centuries, along with the exploitation of human resources through enslavement and the exportation of African people. This gave rise to the rich diversity of people, languages and cultures found in the present day of South Africa (Thompson, 2006).

The country is a parliamentary representative democratic republic, which was radically transformed in 1994 when the previous apartheid system of racist segregation was formally abolished. Moreover, the aforementioned four provinces were divided into the nine currently existing provinces as shown in Figure 1.4.



Figure 1.4: Nine provinces of South Africa as established in 1994 after the abolishment of the apartheid era. The figure right below shows the geographical location of South Africa on the World Map (Created in Tableau Public: <https://public.tableau.com/en-us/s/>).

No reliable information was available regarding the country before the establishment of democracy on the 27th of April 1994. Between the 10th and 31st of October 1996, the post-apartheid government conducted its first population census, with the most recent, third, census in 2011. The statistics indicated that the South Africa population was made up of 51.8 million people at the time. This was further categorized according to race, where Black African constituted 76.4%, white (referring to a person of European-ancestry) at 9.1%, Coloured (persons of admixture) at 8.9%, Asians at 2.5% and others or unspecified at 0.5% (Statistics South Africa, 2012).

1.2. THE AIMS OF THE PROJECT

The present study was aimed at characterizing the maternal genetic ancestry of the current-day South African populations, by analysing the mtDNA control region. For this, a total of 246 individuals belonging to four population groups in three provinces of South Africa: native Nama and Bantu from the Northern Cape province and admixed Coloured and Griqua from the Northern Cape, Western Cape and KwaZulu-Natal provinces was investigated. The study further analysed the haplogroup distribution in both a phylogenetic and phylogeographical analysis to establish the perspective of maternal genetic lineages. To our knowledge, the study presents for the first-time mtDNA sequences for forensic applications in South Africa.

1.3. GENETIC ANCESTRY ANALYSES USING UNIPARENTAL INHERITANCE

The inheritance that consist of the transmission of DNA from one parental type to the progeny is referred to as uniparental inheritance. In other words, the DNA sequence is transmitted from the mother or the father without recombination. This is observed in Y-chromosome as well as mitochondrial DNA.

The mitochondrial organelle located in the cytoplasm is responsible for the energy production within cells of the human body through oxidative phosphorylation (Bär *et al.*, 2000; Budowle *et al.*, 2003). The mitochondrial DNA is characterized as a small 16.6 kb double-stranded (ds) circular genome establish within the mitochondria. This study relies on several characteristics of the mitochondrial genome as discussed below.

1.3.1. Characteristics of mitochondrial DNA

The mitochondrial genome has a few unique properties that are useful in both forensic and evolutionary applications in anthropology. A key feature involves a high number of copies of the mitochondrial DNA in cells. This enables mitochondrial DNA molecules to be easily recovered in forensic casework where samples fail to yield successful nuclear DNA. Hence, mitochondrial DNA has a higher likelihood of survival (Butler, 2012).

Studies of evolutionary anthropology (in non-pathological conditions) rely on the mitochondrial DNA strict maternal inheritance, lack of recombination and high mutation rate (Budowle *et al.*, 2003; Butler, 2012; Kivisild, 2015; Pakendorf & Stoneking, 2005). Table 1.1 shows the contrast of several key features of mitochondrial DNA and nuclear DNA (Butler, 2012).

Table 1.1: Contrast of the human mitochondrial - and nuclear DNA (Butler, 2012).

Features	Mitochondrial DNA	Nuclear DNA
Genome size	16,569 base pairs	3.2 billion base pairs
Copies per cell	Can occur > 1000	2 (one allele from each parent)
Percent of total DNA	0.25% per cell	99.75%
Organization	Circular	Linear
Inherited	Mother	Mother and Father
Chromosomal pairing	Haploid	Diploid
Generational recombination	No	Yes
Replication repair	No	Yes
Unique to individual	No	Yes (excl. identical twins)
Mutation rate	5-10 times nuclear DNA	Low
Reference	Anderson and co-workers	Human Genome Project

1.3.1.1. Haploid transmission of mitochondria cells

Mitochondrial DNA is transferred in the oocyte cytoplasm from one generation to the next. The sperm mitochondria enter the egg during fertilization, but in early embryogenesis they appear to be lost (Wallace *et al.*, 1999). Consequently, the mitochondrial genome is inherited solely along the maternal line of descent (Bandelt *et al.*, 2006). A few cases have been presented where paternal mitochondrial DNA could be passed to the offspring. However, this transmission of the paternal mitochondria or mitochondrial DNA has not been convincingly demonstrated in humans (Luo *et al.*, 2018).

Luo, et al. (2018) further suggested that the central dogma of maternal inheritance of the mitochondrial DNA remains acceptable. Although, the rejection of the non-recombination characteristic of the human mitochondrial DNA was concluded (Lunt and Hyman, 1997), and more recently by Perera et al. (2018). However, at this point, the influence on the evolution of population genetics is insignificant.

This study, therefore, relies on the fact that mtDNA is not reshuffled from generation to generation and, excluding mutations, a mother passes her mtDNA type to all her offspring and thus siblings and maternal relatives share the same mtDNA (Tully and Levin, 2000).

1.3.2. Lineage markers

The characterized feature of passing markers from generation to generation, with a lack of recombination lead to a lineage marker. Two markers, Y-chromosome and mitochondrial markers represent lineage markers as seen in Figure 1.5. Paternal and maternal lineages is traced with Y-chromosome and mitochondrial DNA markers, respectively. Meanwhile, autosomal DNA markers represents makers that are shuffled with each generation from both paternal and maternal inheritance (Butler, 2012).

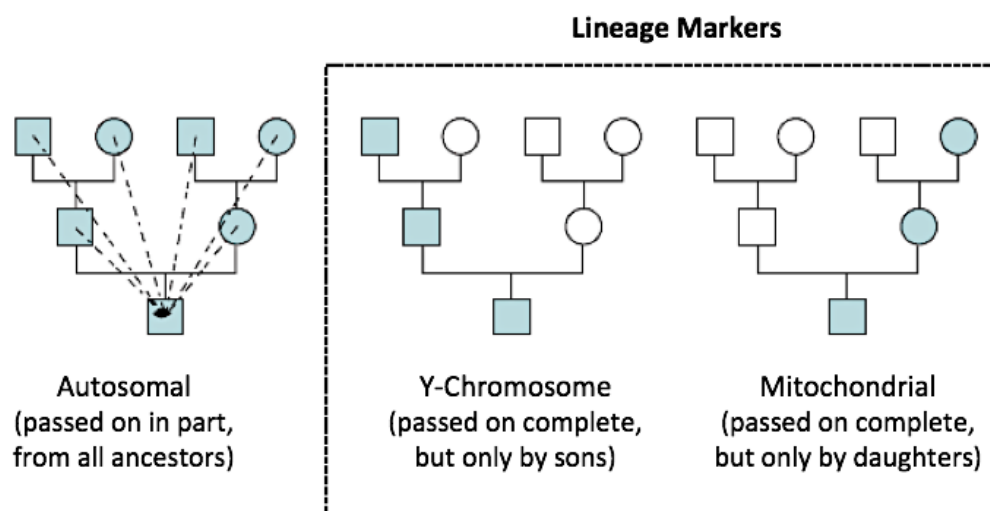


Figure 1.5: Inheritance pattern of lineage markers; Y-chromosome and mitochondrial, along with autosomal DNA. The lineage markers present a strict paternal and maternal inheritance (Butler, 2012).

1.3.2.1. Single nucleotide polymorphism

The mitochondrial DNA is a circular arrangement of four nucleotides, when comparing two sequences, different nucleotides may arise at the same position, resulting in a single nucleotide polymorphism (SNP), as shown in Figure 1.6 (Stoneking, 2001). These SNPs act as the aforementioned biological lineage markers. This event occurs as a result of errors during DNA replication in meiosis and their frequency varies across the mitochondrial genome regions (Goodwin *et al.*, 2007).

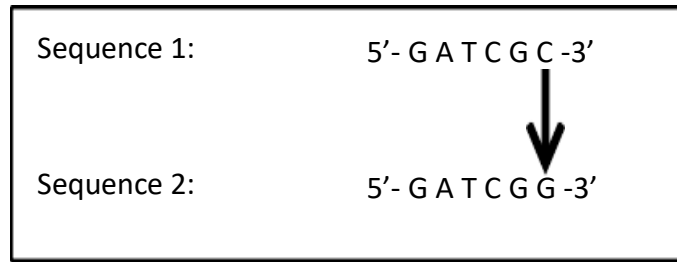


Figure 1.6: Single Nucleotide Polymorphism. Two variant alleles are displayed at one position by an arrow, the last (sixth) position in sequence 1 is a cytosine while in sequence 2, the same position consists of a Guanine.

In general, single nucleotide polymorphisms represent previous mutations that were (although not exclusively) unique events. The reported SNPs of the mitochondrial DNA sequence defines an individuals' haplotype. In turn, a collection of similar haplotypes defined by the combination of specific SNPs in mitochondrial inherited from a common maternal ancestor, is further categorized into a haplogroup. This was formed as a result of sequential mutation accumulation through maternal lineage (Amorim *et al.*, 2019; Mitchell *et al.*, 2015).

Therefore, it can be said that two individuals sharing a variant allele are marked by an evolutionary heritage. Hence, human genes have ancestors, and the study of shared patterns of single nucleotide polymorphisms can identify one's origins (Stoneking, 2001).

1.4. MITOCHONDRIAL DNA STRUCTURE

1.4.1. Control region

The control region or displacement loop (D-loop) is a 1.1 kb non-coding segment of the mtDNA, extending from position 16 024 to 576 (Butler, 2012). This region contains the origin of replication and does not code for gene products, therefore constraints are fewer for nucleotide variability leading to a mutation rate of 6 to 9 times higher in comparison to the coding region of the mitochondrial genome (Bär *et al.*, 2000 and Budowle *et al.*, 2003). Sigurğardóttir, et al. (2000) estimated the mutation rate in the human mitochondrial DNA control region to be 0.32×10^{-6} /site/year.

Most of the sequence variation between individuals is concentrated in three short hypervariable regions (HVR) of the control region, namely: HV-I, -II and -III. Figure 1.7 illustrates HV-I extending over base pairs 16 024 - 16 365 and HV-II and HV-III is observed at nucleotide position 73 - 340 and 438 - 576, respectively (Butler, 2012).

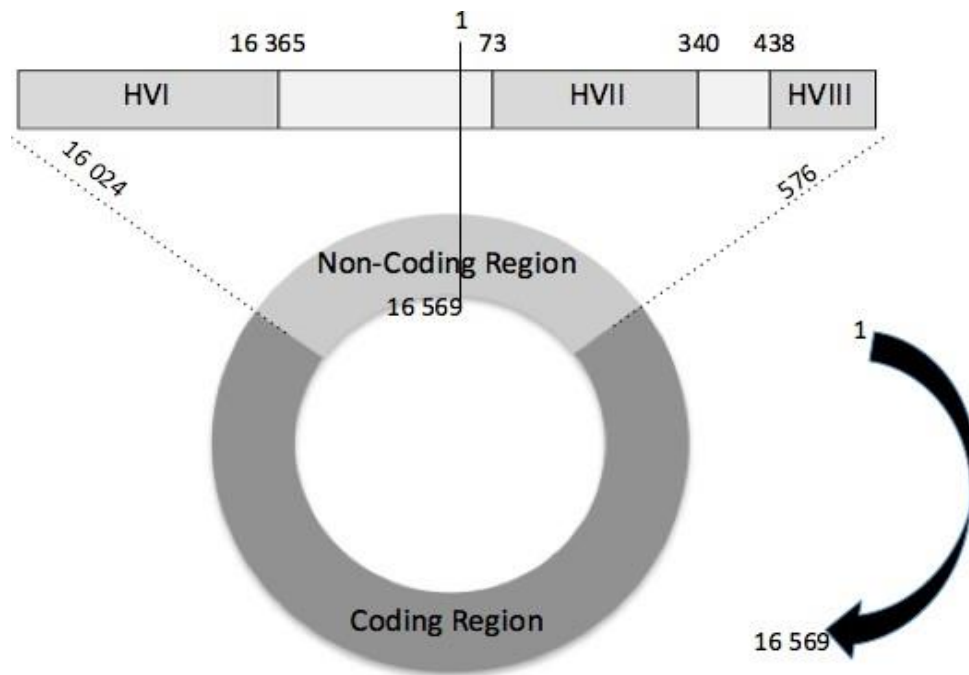


Figure 1.7: Circular structure of the human mitochondrial DNA. The mtDNA includes the coding and non-coding region which holds hypervariable regions I, -II and -III.

The non-coding regions of high variability present an abundant source of ancestral informative polymorphic mutations. Therefore, in population studies, these mutations can be utilized to trace an individual back several thousands of years to the geographical location of that individual's ancestors (Schlebusch *et al.*, 2011).

1.4.2. Mitochondrial DNA coding region

The mitochondrion coding region (bases 577 – 16 023) is highly conserved. It codes for 13 essential proteins of the respiratory chain. Two ribosomal RNA (rRNA) genes (12S and 16S rRNA), and 22 transfer RNAs (tRNA) are interspaced between the protein-coding genes. Due to the highly conserved characteristic of this region, the mutation rate is much lower than the control region (Butler, 2012; Chen *et al.*, 2011).

Previous studies have shown various situations in which the mtDNA coding region data can be beneficial. This includes, resolving numerous casualty cases where additional reference family shared the same mitochondrial DNA control region haplotype (Just *et al.*, 2009; Sturk *et al.*, 2008), sorting and re-association of commingled remains (Just *et al.*, 2009), increasing the statistical support when exclusionary references are unattainable (Irwin *et al.*, 2007), mitochondrial haplogroup typing for rapid screening of casework

samples (Álvarez-Iglesias *et al.*, 2007; Brandstätter *et al.*, 2003; Parson and Dür, 2007), and evaluating maternal bio-geographic ancestry (Köhnemann and Pfeiffer, 2011).

1.5. MITOCHONDRIAL DNA HAPLOGROUPS

The International Society of Genetic Genealogy defines the term haplogroup (Hg) as “a genetic population group of people who share a common ancestor on the patrilineal or matrilineal line”. Numerous mtDNA types can be separated from diverse populations globally by applying a classification technique introduced by Torroni *et al.* (1993).

1.5.1 Haplogroup structure

The nomenclature of mitochondrial DNA haplogroups was introduced by Torroni *et al.* (1993) based on a data set obtained from a restricted number of mutations in Native Americans. These mutations defined four primary branches in the phylogenetic tree, named A, B, C and D. Thereafter, the haplogroup structures of other continental populations were characterized (Torroni *et al.*, 1994i, ii), while Richards *et al.* (1998), overtly established the cladistics rules for the hierarchical ordering of haplogroups and sub-haplogroups.

The sub-haplogroup nomenclature moves from root to tip, with the gain of an extra identifying symbol (either a number or a letter, in alternating order) attached to the label at each branching point as follows: L>L0>L0d>L0d1>L0d1a. Each branch is characterized by one or more mutations, resulting in a combination of which it is marked for direct genotyping within a population. Generally, the mutations are reported as variants or deviations from the revised Cambridge Reference Sequence.

1.5.2. Classification of haplogroup

An individual can be assigned a haplogroup once the annotation is completed, following the recommended nomenclature (*section 1.8.1.1*) and a combination of single nucleotide polymorphisms has been allocated to an individual's mitochondrial DNA sequence. This haplogroup classification of mtDNA sequences has been significantly simplified with the establishment of PhyloTree (<http://www.phylotree.org/tree/index.htm>) as shown in Figure 1.8, which is broadly accepted as the “mitochondrial haplogroup dictionary” (Röck *et al.*, 2013; van Oven & Kayser, 2009).

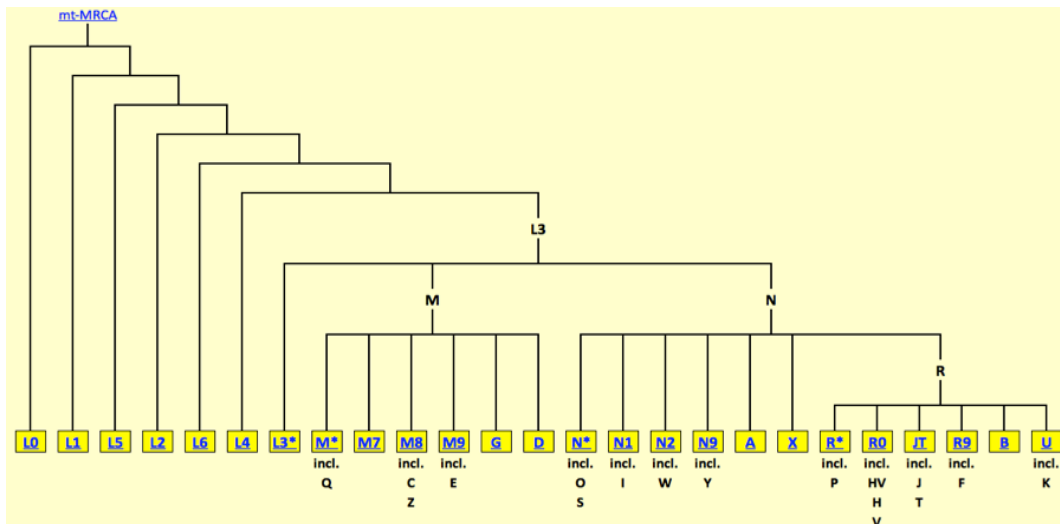


Figure 1.8: Mitochondrial DNA phylogenetic tree build 17 (18 Feb 2016), divided into 25 subtrees available via links on phyloree.org/tree/index.htm. The mitochondrial most recent common ancestor (mt-MRCA) is used to root the tree. Accumulation of polymorphisms divides individuals into different haplogroups over time (van Oven & Kayser, 2009).

The haplogroup-defining mutations described in PhyloTree allows for manual haplogroup assignment and serves as a source for several automatic mtDNA haplogroup classification software (e.g. HaploGrep 2, Haplofind, etc.). However, none of the automated solutions available generate reliable and unbiased haplogroup assessments. The main limitation of existing tools is that the assignment of the haplogroups is based solely on defined diagnostic mutation patterns. Nevertheless, the remaining “private” mutations of a sequence can be additionally informative for the haplogroup status. As a consequence, the assignment of haplogroups are often incorrect (Röck *et al.*, 2013).

1.5.2.1. Geographical origins based on haplogroups

The mtDNA macro-haplogroup L has been identified as the most ancestral mitochondrial lineage of humankind (Macaulay *et al.*, 2005). Therefore, this haplogroup is placed at the root of the mtDNA phylogeny and geographically restricted to sub-Saharan Africa (Gonder *et al.*, 2007). The root branches off into L0, the first daughter branch of the human mtDNA phylogeny. Haplogroup L0 represents the Pygmies, KhoiKhoi and San groups, and reaches its highest frequency in the KhoiSan population (Behar *et al.*, 2008; Chan *et al.*, 2019; Stoneking, 2008).

The branch further divides into sub-haplogroups L0a, L0d, L0f and L0k (Kivisild *et al.*, 2006; Mishmar *et al.*, 2003; Salas *et al.*, 2002; Salas *et al.*, 2004). Sub-haplogroup L0a probably originated in eastern Africa and is generally found in eastern, central, and south-

eastern Africa (Salas *et al.*, 2002). Sub-haplogroup L0d and L0k are exclusively established among the southern African KhoiSan speakers (Salas *et al.*, 2002).

Previous studies have displayed the significant phylogeographical structure in Africa, such as the aforementioned confinement of L0d and L0k to the KhoiSan people. The two haplogroups are known as the deepest mtDNA clades among modern humans, accounting for over 60% of the contemporary mitochondrial DNA gene pool (Behar *et al.*, 2008). Furthermore, genetic studies have observed that L0k and L0d exists at the root of the human mitochondrial DNA tree (Chen *et al.*, 1995; Henn *et al.*, 2011; Ingman *et al.*, 2000; Kivisild *et al.*, 2006; Mishmar *et al.*, 2003; Ruiz-Pesini *et al.*, 2004).

The second daughter branch of the human mitochondrial DNA phylogeny is haplogroups L1'2'3'4'5'6 (L1'6) (Behar *et al.*, 2008; Chan *et al.*, 2019; Torroni *et al.*, 2006). The L1'6 branch is widespread, as seen below in Figure 1.9, and has given rise to almost every mitochondrial DNA lineage found today.

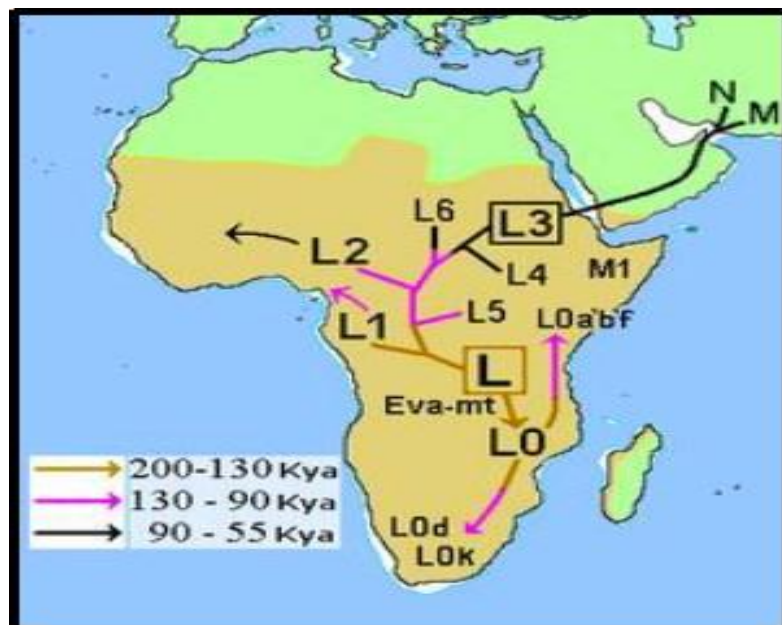


Figure 1.9: Map displaying the early diversification of modern human. Beginning with mitochondrial Eve and the movement throughout Africa (Hg L0; L1-6) and MacroHg M and N (Macauley *et al.* 2005).

The L1 - L6 branch was introduced during the recent Bantu expansion (Kivisild, 2015; Macauley *et al.*, 2005). The expansion of maternal haplogroup L3 can be singled out from the spectrum of African mitochondrial DNA haplogroups as the sole carrier of branches

covering mitochondrial DNA variation during the prehistoric times into the rest of the world (Bandelt *et al.*, 2006). This event is also known as the out-of-Africa dispersal and resulted in the occurrence of two clades on the L3 branch namely: MacroHg M and N.

1.5.3 Southern African population haplogroups

The population groups studied in this project resides in South Africa. Therefore, it can be hypothesized that the majority of the individuals from this study will fall into African ancestral haplogroups (L0 – L5). However, previous studies on similar southern African population groups by Schlebusch *et al.* (2011) and Quintana-Murci *et al.* (2010) presented the integration of European and Asian maternal lineages, although, the degree of which will be observed in this study is unknown.

1.6. STANDARD MTDNA SEQUENCING IN FORENSIC CASEWORK

Traditional methodologies used to identify single nucleotide polymorphisms include polymerase chain reaction (PCR) amplification, accompanied by sequencing or restriction fragment length polymorphism (RFLP) mapping. In the context of this thesis, where a region needs to be analysed for single nucleotide polymorphisms, Sanger sequencing is preferred. While the most efficient option for the whole genome is next-generation sequencing (Kwok, 2001).

1.6.1. DNA sequencing

1.6.1.1. Sanger sequencing

The first DNA sequencing method, known as the Sanger sequencing method, was developed in 1977 by Frederick Sanger and colleagues. This method won a Nobel Prize in chemistry in 1980. During a Sanger sequencing procedure, the DNA sequence of interest, forward and/or reverse primers, DNA polymerase, nucleotides (dATP, dTTP, dGTP and dCTP) and labelled di-deoxynucleotides (ddATP, ddTTP, ddGTP and ddCTP,) is combined in a tube (Chain & Heather, 2016).

The di-deoxynucleotidetriphosphates (ddNTPs) terminate the elongation of the DNA strand due to the lack of a 3'-OH group on the chain-terminating nucleotides (Chain & Heather, 2016). This results in the formation of extension products of various lengths terminated with modified di-deoxynucleotides as illustrated in Figure 1.10.

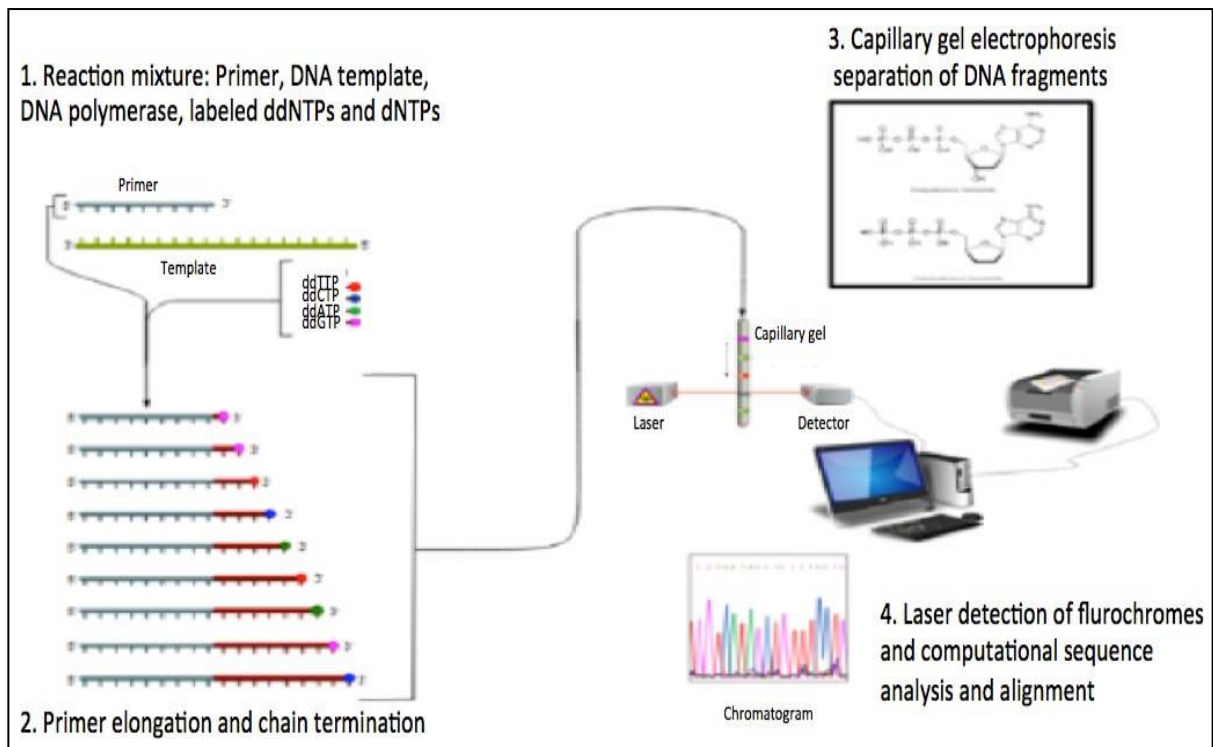


Figure 1.10: Four steps of Sanger sequencing to obtain a DNA sequence of interest. The four ddNTPs are detected by four distinct coloured fluorescent dyes attached to each ddNTP. The ddTTP (thymine), ddCTP (cytosine), ddATP (adenine) and ddGTP (guanine) are labeled with red, blue, green and yellow, respectively. However, for easy visualization, the ddGTP is generally displayed in black (Murphy *et al.*, 2005).

The third step in the sequencing process requires the DNA sequences to be separated by length, using capillary gel electrophoresis. Thereafter, a laser excites the fluorescent label on the nucleotide at the end of each sequence. The detector determined the emission of the distinct coloured fluorescent dye to generate a chromatogram, displaying the fluorescent peak of each nucleotide as shown in Figure 1.11. The advantage of this technique is that it is less time consuming and less labour intensive. Currently, this dye-terminator sequencing technique is predominant used in automated sequencing due to its greater efficiency (Franc *et al.*, 2002).

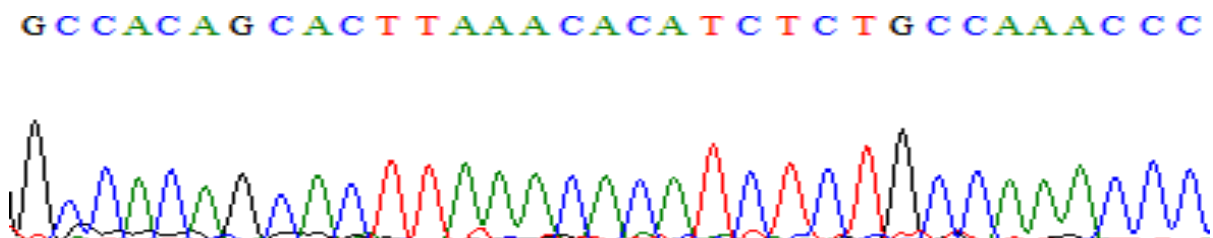


Figure 1.11: Electropherogram generated by Sanger sequencing of the mitochondrial DNA. Each respective colour fluorescent dye (i.e. C = blue, A = green, T = red and G = black) indicates the respective nucleotide in the sequence.

1.6.1.2. Next-generation sequencing

Over the past decade, next-generation sequencing (NGS) technologies have evolved with the advantage of generating read lengths as long as some whole genomes. NGS has the potential to produce a large amount of data, however the related error rates (~0.1–15%) are higher and the read lengths are typically shorter (35–700 bp for short-read approaches) than the aforementioned Sanger sequencing platforms and the results require thorough examination (Goodwin *et al.*, 2016). Although long-read sequencing exceeds the length limitation, it remains significantly more expensive and has a lower throughput (Goodwin *et al.*, 2016; Liu *et al.*, 2012, Mardis, 2008).

Roche's 454 pyrosequencing (first instrument to achieve commercial production in 2004), along with multiple sequencing platforms has the potential for NGS in evaluating mtDNA including; Illumina's -genome analyser, -HiSeq, -MiSeq and the use of the Ion Torrent's Personal Genome Machine (Metzker, 2010, McElhoe *et al.*, 2014). Currently, the Illumina platform is the most successful sequencing technology on the market (Goodwin *et al.*, 2016). The platform is characterized by its technique of sequencing-by-synthesis combined with PCR bridge amplification on the surface of a flow cell. It facilitates the generation of the clonally enriched template DNA for sequencing, known as cluster generation. This technology utilizes removable fluorescently labelled chain-terminating nucleotides and is thus more cost-effective. Furthermore, this method is less time consuming (Goodwin *et al.*, 2016).

1.7. REPORTING MITOCHONDRIAL DNA VARIATION

A reference sequence is the foundation on which mtDNA variation is recorded. It is used for the purpose of communicating the variation of a lineage in a consistent form by only reporting the variants of a new lineage relative to the reference sequence selected.

Numerous mitochondrial DNA reference sequences such as the Yoruban (African) sequence, Reconstructed Sapiens Reference Sequence (RSRS), GRCh37/ UCSC Hg19 or older GRCh36/UCSC Hg18 has been made available through the years however, the revised Cambridge Reference Sequence (rCRS) has been used as a reference to annotate mitochondrial DNA variation in molecular anthropology and forensic applications. Medical geneticists often misunderstood this reference sequence as a wild-type or consensus sequence (Bandelt *et al.*, 2014).

1.7.1. History of revised Cambridge Reference Sequence

The human mtDNA was initially sequenced in the year 1981, in Frederick Sanger's laboratory in Cambridge, England (Anderson *et al.*, 1981). The sequence was established from a woman of European descent, with haplogroup H2a2a1. However, some HeLa and bovine sequences were utilized to fill the gaps created by early rudimentary DNA sequencing techniques (Anderson *et al.* 1981; Butler, 2012; Parson *et al.*, 2014). The sequence was named after the initial author, "Anderson" or the Cambridge Reference Sequence. The Anderson sequence (GenBank accession number: M63933) has been used as a reference for numerous years to which new sequences were compared.

By 1999 DNA sequencing technology had improved and led to the re-sequencing of the original placental material used in Frederick Sanger's laboratory (Andrews *et al.*, 1999). During the reanalysis effort, 11 exceptions were identified to that of the original 1981 sequence. This included a loss of a single cytosine nucleotide at position 3107. Subsequently, the reference mitochondrial genome suggests 16 568bp rather than the traditional 16 569bp. However, excluding the additional position would have created confusion and an inability to easily correlate former work (Butler, 2012). The Sanger research group thus suggested that the original numbering must be retained with a placeholder, "N" at site 3107.

The original errors of the CRS were corrected for the robust use of the revised Cambridge Reference Sequence (rCRS). The rCRS is currently the accepted standard for assessment and can be obtained with NCBI reference sequence NC_012920.1 (Butler, 2012).

1.8. MITOCHONDRIAL DNA TYPING

1.8.1. Data assessment and editing

Some quality control steps are in place throughout the sequencing process of the mitochondrial DNA. Mitochondrial DNA sequencing is carried out in both the forward and in the reverse direction to evaluate the quality by comparing the complementary strands to one another (Butler, 2012). This in turn, facilitate alignments as multiple sequences of the same sample over a region are generated. In addition, the base calls for respective nucleotides provided by a software require manual review and potential editing.

1.8.1.1. Nomenclature

All nomenclature recommendations are intended to be compatible with IUPAC (International Union of Pure and Applied Chemistry) codes. The whole human mtDNA sequence defined by Anderson et al. (1981) is used as a reference standard to promote nomenclature of mitochondrial DNA types. When the mtDNA sequence of interest obtained from an individual varies from the revised Cambridge Reference Sequence, that difference is recorded.

The guidelines set out by the DNA commission of the International Society for Forensic Genetics states that only the specific position (designated by a number) and nucleotide that contradicts the reference standard are noted. For instance, in the Anderson sequence, position 73 (HVR II) holds an adenine, whereas most of the population carries a guanine as shown in Figure 1.12. An individual representing such a sequence of mitochondrial DNA is recorded as 73G. If no other sites are reported, only then is it assumed that the mtDNA sequence is identical to the Anderson sequence, except as recorded at position 73 (Carracedo *et al.*, 2000; Parson *et al.*, 2014).

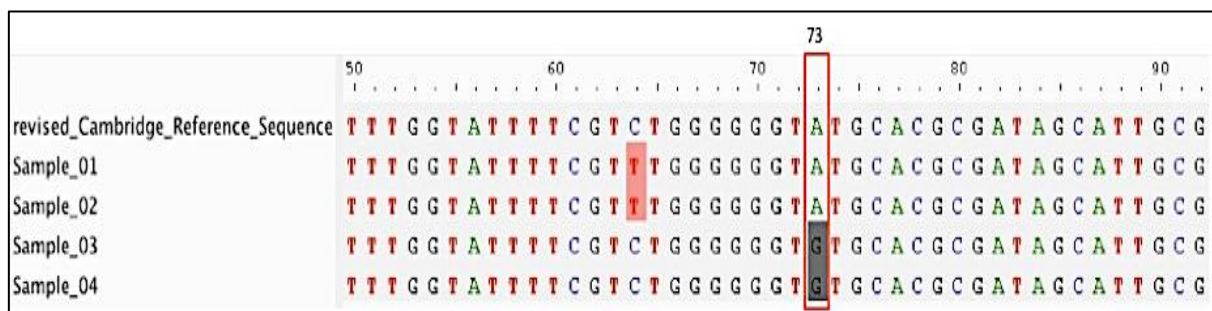


Figure 1.12: Visual representation of the revised Cambridge Reference Sequence (rCRS) at site 73 (in HVII), compared to that of four individuals' sequences (Sample_01 - 04). The rCRS can be observed in the first line, showing an A (adenine) at site 73. The same is seen for the following two individuals (Sample_01 - _02), while Sample_03 - _04 holds a guanine.

In the event of an unresolved nucleotide (i.e. indecisive of A, C, T or G), the given site is recorded as an 'N'. An insertion is defined by recording the site 5' to the insertion, a decimal point and a '1', while a '2' signifies an additional insertion, and so on, followed by the base pair inserted. Deletions, on the other hand, were formerly noted with the missing site followed by a 'd'. However, in 2014, the revised and extended guidelines for mtDNA typing indicated the use of "DEL" or "del" (Parson *et al.*, 2014).

In view of the homopolymeric tracts discussed below, where the specific location of the insertion is uncertain, it should always be assumed that the insertion took place at the highest end of the homopolymeric region (Carracedo *et al.*, 2000; Bandelt & Parson, 2008). Lastly, the heteroplasmic sites, discussed under *section 1.8.2.2*, are reported with the nomenclature scheme provided by IUPAC as seen in Table 1.2.

Table 1.2: IUPAC codes for base calling when a site has more than one nucleotide i.e. heteroplasmy. R should signify a mixture of G and A, while Y should denote a mixture of T and C. The code uses capital letters, as small letters describe mixtures of deleted/undelated and inserted/non-inserted bases (Parson *et al.*, 2014; Stothard, 2000).

IUPAC nucleotide code	Base
A	Adenine
C	Cytosine
G	Guanine
T (or U)	Thymine (or Uracil)
R	A or G
Y	C or T
S	G or C
W	A or T
K	G or T
M	A or C
B	C or G or T
D	A or G or T
H	A or C or T
V	A or C or G
N	any base
. or -	gap

1.8.2. Issues impacting interpretation

1.8.2.1 Sequencing beyond polymeric C-stretches

Figure 1.13A illustrates a stretch of cytosine nucleotides in the dotted box as observed in the HVR I, generally referred to as a “C-stretch” (Butler, 2012). The C-stretch region in the control region can be detected at 2 locations: HVR I (nucleotide positions 16 184 - 16 193) and the other in HVR II (sites 303 - 315).

HVR I should have a T at site 16 189, however, some samples have a C, resulting in a stretch of 8 to 14 or more cytosines in a row (Bendall & Sykes 1995). The length heteroplasmy is likely due to replication slippage following a transition from T to C that occurs at site 16 189. Meanwhile, a similar situation occurs with the C-stretch located in HVR II, where the insertion of cytosines occurs in the area of 303 - 315 and a transition of T to C occurs at position 310, resulting in a homopolymeric C-stretch (Bendall & Sykes 1995; Butler, 2012). This mixture of length variants could either be generated in the sequencing reaction or present within the original DNA.

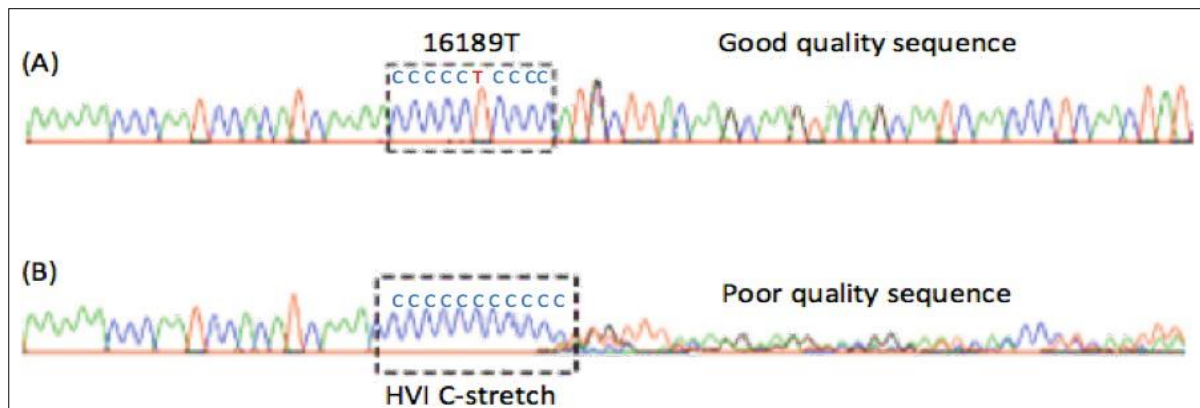


Figure 1.13: Good quality sequence compared to that of a poor-quality sequencing that results in the C-stretch. Mitochondrial DNA sequence (A) is absent of the HVR I C-stretch (consists of 16 189T) and (B) illustrates a C-stretch. It can be noted that the effect of the transition on sequencing results downstream of the C-stretch and the quality rapidly declines after the series of cytosine residues (Butler, 2012).

A quality examination of HVR I PCR products, with the use of additional heteroduplex peaks, is beneficial as it allows the C-stretch to be scanned easily before sequencing. If the C-stretch occurs in a sample, several sequencing primers could be used to generate valid mtDNA sequencing information downstream of the homopolymeric stretches (Butler, 2012; Rasmussen *et al.*, 2002).

1.8.2.2. Heteroplasmy

It has been reported that the regions of the mitochondrial genome evolve 6 to 17 times the rate of single-copy nuclear genes (Brown *et al.*, 1979; Wallace *et al.*, 1987; Tully, 1999). It is therefore unlikely for millions of mtDNA molecules distributed throughout an individual's cells to be entirely identical, thus resulting in heteroplasmy. Hence, two or more mtDNA populations can appear amongst cells in an individual or inside a single cell.

The presence of heteroplasmy is confirmed if two nucleotides can be found in the sequence from both DNA strands distinctly above the background level. This event is marked by the occurrence of two nucleotides (overlapping peaks) at a single position in a sequence electropherogram as seen in Figure 1.14 (Butler, 2012).

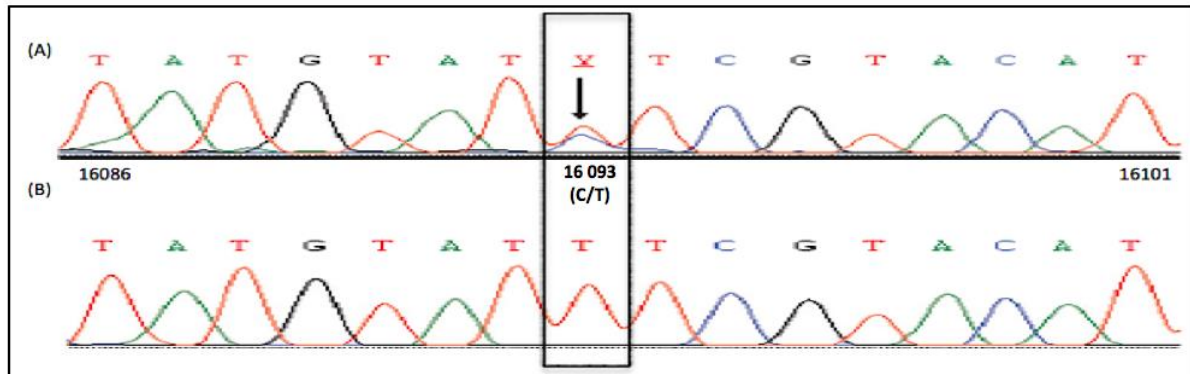


Figure 1.14: Electropherogram of mtDNA positions 16 086 – 16 101 of two individuals to compare the presence and absence of heteroplasmy. An electropherogram of sequence heteroplasmy detected at a sole position 16 093, (A) occupying both nucleotides C and T compared to the same region of a different sample (B) consisting of a T. The heteroplasmy site 16 093 retaining both C and T nucleotides (Butler, 2012). (IUPAC codes label this heteroplasmy by a 'Y' as seen in Table 1.2).

Single nucleotide (point heteroplasmy) and length heteroplasmy, are two distinct categories of heteroplasmy. Other hotspots previously reported for heteroplasmy includes the following positions in HVR I: 16093, 16129, 16153, 16189, 16192, 16293, 16309, and 16337 and 72, 152, 189, 207, and 279 in HVR II. The occurrence of heteroplasmy does not discredit the usage of mtDNA for forensic purposes. While, it may complicate the interpretation, it can improve the match probability in DNA identification (Butler, 2012).

1.9. LOCAL CRIMINAL DATABASE

A criminal database such as the National Forensic DNA Database of South Africa (NFDD) comprises of forensic DNA profiles categorized in six sub-divisions: a crime scene -, arrestee -, convicted offender -, investigative -, elimination -, missing persons and unidentified human remains indexes. The profiles in the South African database consist only of microsatellite profiles. This information enables the South African Police Service (SAPS) or criminal justice agencies and associated law enforcement to have access to crimes and criminals.

SAPS laboratories do not use mitochondrial DNA technology. Meanwhile, the United States national DNA database also referred to as the Combined DNA Index System

(CODIS), encompasses microsatellite profiles and allows for the upload of mtDNA information into the missing person indices. The Federal Bureau of Investigation (FBI) has compiled this mtDNA population database also known as CODIS_{mt}, to determine legally defensible frequency estimates (Monson *et al.*, 2002).

1.10. SIGNIFICANCE OF THIS STUDY

The human mtDNA has been utilized as a tool in investigating human variation and evolution since Rebecca Cann proposed the theory of a “mitochondrial Eve” in the eighties. The mt-Eve is the matrilineal most recent common ancestor of all living humans. Hence, the most recent woman from whom all living humans descent (Cann *et al.*, 1987). Thus, the study of mtDNA plays an important role in numerous parts for instance; investigating the socio-cultural impact on human genetic variation, ancient DNA, tracing personal genetic history and some forensic applications.

This thesis focuses on the study of the mitochondrial DNA in two applications: genetic anthropology and forensics. The importance of these respective applications, in the context of this thesis, is discussed in the following two sections below.

1.10.1. Population study for the participating individuals

From a scientific point of view, the South African populations studied in this thesis provided a rare opportunity to explore the genomic variability from some of the oldest human lineages to the implications of complex admixture patterns. However, for the donor individuals, testing for one’s ancestry has become very significant in South Africa for the following reasons.

1.10.1.1. Symbolism and belonging

Genealogy testing, such as 23andme, focuses on the geographical origin (ancestry) to which individual genes align with. It has been considered that this information is more relevant than a person’s self-identified race. However, in recent years, South Africa’s political recognition of human rights and land reallocation policies revolves around the identity of an individual.

A key component in categorizing the ethnicity of an individual is languages and culture. However, they can be changed by environmental influence and surrounding populations.

Therefore, the study of the geographical origin is parallel to a respective population group (Gabie, 2014).

Moreover, it can be noted that for centuries some southern African people such as the Coloureds, 'Boesmans' and Griquas have been made to feel ashamed of their identity, their loss of language and self-worth. Thus, the indigenous qualities that these individuals once had, has been erased and suppressed during the apartheid era (Gabie, 2014; Oomen, 2005). This study will contribute to the awareness of the once historically marginalized groups and empower the participating individuals to reclaim their indigenous status by identifying unknown indigenous maternal ancestry detected in their genome. Hence, population studies enable a person to embrace one's heritage and signifies one's notion of wanting to belong.

1.10.2. MtDNA in Forensic applications

Mitochondrial DNA profiles are presented in the events of unidentified remain cases as observed during catastrophic and mass disasters, together with searching of a missing person's investigation. To aid the aforementioned investigations, an estimation of the rarity of haplotypes and (or) haplogroups of mitochondrial DNA in question is essential to determine if a suspect's mitochondrial DNA sequence matches to an evidentiary sample. For this, a mitochondrial DNA database stands as the basis for frequency estimations of mitochondrial DNA sequences that became relevant in casework (Parson *et al.*, 2011).

A mitochondrial DNA database holds information on published mitochondrial DNA variation for inferring geographical origins of mitochondrial DNA lineages from anonymous donors. This type of database is generally accessible to the public through the World Wide Web and allows for both browsing and querying capabilities. It is informative on the frequency of mitochondrial DNA haplotypes and (or) haplogroups in a particular geographical location or population, along with the absence of that haplotype or haplogroup.

Various mtDNA hypervariable regions I and (or) -II databases and additional whole mtDNA genome databases exist such as Mitomap, Human Mitochondrial DataBase (HmtDB) and GenBank. However, Genbank is not curated as the EDNAP mitochondrial DNA Population Database, in short EMPOP. This worldwide forensic database was established by the

European DNA Profiling Group (EDNAP) and released in 2006. The database represents a comprehensive forensic-quality population reference database (Parson and Dür, 2007).

1.10.2.1. EDNAP mitochondrial DNA population database

EMPOP is a collaborative effort of several DNA laboratories under the lead of the DNA Laboratory of the Institute of Legal Medicine (GMI) in Innsbruck, Austria. EMPOP v4/R12 database is a repository of 42,839 mtDNA haplotypes and continues to evolve (<https://empop.online>). The main emphasis of the EMPOP database continues to be on providing highest quality mitochondrial DNA data. EMPOP also serves as a quality-control tool in forensic genetics and other disciplines. While several other reference population databases are available for forensic comparisons, EMPOP is the most comprehensive resource from a population perspective (Parson *et al.*, 2014).

As mentioned above the quality of a reference database is crucial in achieving the correct frequency estimations. Hence, mitochondrial DNA population data should be subjected to analytical software tools that assist as a data quality control step. Fortunately, EMPOP employed a comprehensive suite of quality control tools, such as “EMPcheck” and “Network”, a phylogenetic tool developed and designed by EMPOP, based of quasi-median network analysis that graphically represents the genetic structure of the lineages in a data set (Parson & Dür 2007; Zimmermann *et al.*, 2014).

1.10.2.1.1. EMPOP problem statement

The EMPOP mitochondrial DNA database v4/R12 (<https://empop.online>) consists of 33,447 mtDNA control region sequences submitted or used for variant calling, with a total data set of 42,839 mtDNA samples submitted from Africa, America, Asia and Europe continents. The African continent constituted 2,179 samples from only ten of the total fifty-four countries in Africa, none were contributed from South Africa.

This is a principle limitation as it lacks population groups, and will not provide an accurate estimation of the rarity of questioned mitochondrial DNA. Therefore, this project aims at improving the collection of mtDNA lineages from the diverse populations in South African over three geographical regions.

CHAPTER 2: METHODS

2.1. ETHICS STATEMENT

The University of the Western Cape Senate Research Committee granted ethical clearance 15-4-97. An informed and written consent was furnished as a prerequisite for a subject's participation. Further information regarding ethnic group, birthplace and home language was recorder.

2.2. STUDY AREA

Sample collection was conducted in South Africa, a country on the southernmost tip of the African continent. The samples were collected from individuals residing in; (i) Northern Cape, the largest province in South Africa, (ii) KwaZulu-Natal located in the southeast of the country and (iii) Western Cape regions. The samples were labelled as MR14- and N13 for the collection in the Northern Cape, while KwaZulu-Natal and Western Cape samples were labelled as GRI and VGRI, respectively.

2.3. BIOLOGICAL SAMPLE COLLECTION

A total of two hundred and forty-six (246) biological material was obtained in the form of buccal swabs. This was obtained by gently massaging a buccal swab for 30 seconds against the inner cheeks of the donor, after which it was placed and sealed into separate envelopes. The samples were put in a freezer box until -20°C storage could be accessed and held there until the DNA was extracted.

2.4. METADATA AND SELECTION OF PARTICIPANTS

Biological samples are categorized according to the self-declared ethnic population group that was provided by the donor on the consent form.

Table 2.1: Population groups selected for the study includes the Coloured, Griqua, Nama and Bantu. The Bantu in this study was made up of individuals who classified themselves as Black, Pedi, Tswana and Xhosa on the consent form.

Province	Participant self-declared ethnicity			
	Coloured [n = 138]	Griqua [n = 88]	Nama [n = 10]	Bantu [n = 10]
Western Cape	24	68	0	0
Northern Cape	73	2	10	10
KwaZulu-Natal	41	18	0	0

2.5. DATA ASSEMBLY

The sequencing of the mitochondrial DNA was performed as described by a previously published Magister Scientiae thesis by Heynes, K (2015) and Doctor of Philosophiae thesis by Ristow, P (2017). The specific details that are known to affect data interpretation are summarized in Table 2.2. It can be noted in the table that data for this investigation was acquired from two sequencing methods; (i) Sanger sequencing analysis, along with (ii) next-generation sequencing approach.

Table 2.2: Analysis methods employed to produce the mtDNA population data.

Sequencing type	Amplification primers				Sequencing chemistry	Sequencing instrument	
Sanger sequencing	L15969	H16509	L15	H185	AB Terminator v3.1	BigDye	ABI 3100
Next-generation sequencing	L15989	H16237	L16190	H389	MiSeq reagent 250 kit v2		MiSeq
	H16410	L34	H180	L155			
	L403	H639					

The fasta files acquired during Sanger sequencing were viewed in BioEdit v7.0.5 (Hall, 1999), while FreeBayes were used to generate a Variant Call Format (VCF) file from the next-generation sequenced Binary Alignment Map (BAM) files (Garrison and Marth, 2012). Finally, a fasta file was created for the NGS data with Samtools (Li *et al.*, 2009) in the software UGENE v1.31.0 (Okonechnikov *et al.*, 2012). This fasta file serves as a quality control step, used to compare to that of the VCF file. Additional manual revision was performed for final variant calling following a forensic nomenclature guidelines as described under 2.5.2. SNP haplotype analysis.

2.5.1. MtDNA consensus sequences generated

The Clustal W algorithm implemented in the aforementioned software packages (Larkin *et al.* 2007) was used to generate the 246 consensus sequences, by aligning the query sequence against the revised Cambridge Reference Sequence. The present study only focuses on the control region (16 024bp – 576bp positions) of all the mitochondrial DNA samples. Therefore, the sequences that fall outside of the control region were trimmed to only include the 1121bp region of interest for computational analysis purposes.

2.5.2. SNP haplotype analysis

Haplotypes were defined by identifying variants in the mtDNA control region against the revised Cambridge Reference Sequence (Andrews *et al.*, 1999). The data was recorded and used to produce a list of SNPs for each sample following the nomenclature guidelines set out by the DNA Commission of the International Society of Forensic Genetics (ISFG) concerning mtDNA typing (Carracedo *et al.*, 2000; Parson *et al.*, 2014). For the purpose of computational analysis, a haplotype was defined by the software package DnaSP v5.10.01 with a fasta file (Librado & Rozas, 2009). This software examines the control region of each mtDNA sample to identify identical patterns of genetic variation.

2.5.3. Haplogroup assignment

The mitochondrial DNA haplotypes in this study based on the phylogenetic informative SNPs in the control region were affiliated to (sub)-haplogroups using the web-based tool HaploGrep 2 (v2.1.19) (<http://haplogrep.uibk.ac.at/>), based on Phylotree build 17, 2016. (Kloss-Brandstätter *et al.*, 2010; van Oven and Kayser, 2009; van Oven, 2015).

2.6. QUALITY CONTROL WORKFLOW

2.6.1. Data submission to EMPOP

The mitochondrial DNA haplotype data generated in the South African population dataset in this study were submitted to the EMPOP group at the Institute of Legal Medicine, Innsbruck Medical University as a registered user on <http://www.empop.org>.

The following information was provided during submission; (i) general information such as the submitting institution, contact details and a provisional journal. (ii) Information regarding the sample-set included the type of specimen and number of donor individuals. Further, the geographical origin and metapopulation of the donor sample were provided. (iii) Information on laboratory and data analysis included the sequencing chemistry and instruments used. Lastly, the software utilized for the alignment was provided.

2.6.1.1. EMPOP Quality Control (QC)

The Forensic DNA Laboratory at The University of the Western Cape provided data in the form of single nucleotide polymorphisms to EMPOP for a careful revision process. The data were subjected to quality control and phylogenetic checks utilizing EMPCHECK, NETWORK along with in-house software and manual inspection of the data by the EMPOP team (Parson and Dür, 2007; Huber *et al.*, 2018; Zimmermann *et al.*, 2014).

After EMPOP completed the evaluation process, a few discrepancies were found in some samples and addressed by the Forensic DNA Laboratory. Hence, the raw data were requested and provided for a re-examination by both laboratories for the positions in question. The data has not yet been made available by EMPOP, however, the data presented in this study were rectified as requested during the initial quality control.

2.7. FORENSIC PARAMETERS ANALYSIS

2.7.1. Discrimination capacity of mtDNA fragments

The discrimination capacity (DC) of the various fragments of the mitochondrial DNA was calculated as follows; (N distinct haplotypes/ N individuals). The assessed regions of interest were hypervariable I, hypervariable I and II, hypervariable I to II and the whole control region.

2.7.2. Random Match Probability

The random match probability was calculated as the sum of squared haplotype frequencies based on the mitochondrial DNA control region sequences. The C-stretch length variants in HVR-I (base pair position 16184 - 16193) and HVR-II (position 303 - 315) were discarded in distinguishing haplotypes for calculation of random match probability.

2.8. POPULATION GENETIC ANALYSIS

The population study section (3.2) was divided into sub-sections. The first section explores the mtDNA genetic structure at a geographical point of view i.e. Western Cape, Northern Cape and KwaZulu-Natal. Thereafter, each respective population group i.e. Coloured, Griqua, Bantu and Nama were investigated. Lastly, the study concentrates on the indigenous population of South Africa (maternal haplogroup L0).

2.8.1. Geographical population structure

The definition of all 246 individuals' haplotypes was generated with DnaSP v5.10.01 (Librado & Rozas, 2009) using the sequence data to determine the shared and unique haplotypes that were identified between all three provinces by inferring a genetic structure based on the Western Cape, Northern Cape and KwaZulu-Natal samples.

2.8.2. Population diversity

Population diversity indices in particular the number of variable sites (S), haplotypes number (h), unique haplotypes and nucleotide diversity (π) within a population, were estimated in DnaSP v5.10.01 software (Librado & Rozas, 2009). The gene diversity were calculated with the following formula; $n(1-\sum x^2)/(n-1)$, where N represents the total number of samples and x the relative haplotype frequency (Nei, 1987).

Tajima D (1989) and Fu and Li's F_s (1997), with its statistical significance were employed to detect a deviation from neutrality by the program DnaSP v5.10.01 (Librado & Rozas, 2009). This program does not accept ambiguities and therefore heteroplasmic sites were replaced with an "N" (unknown) IUPAC code.

2.8.3. Degree of differentiation in South African sample-set

An analysis of molecular variance (AMOVA) was further calculated with Arlequin v. 3.5.2.2 (Excoffier and Lischer, 2010). The significance of the variance components between populations and their associated F_{ST} -statistics was evaluated by randomisation tests (10 000 permutations). The variation was tested for the four population groups; Coloured, Griqua, Bantu and Nama using a matrix of molecular distance.

2.8.4. Geographic distribution of indigenous maternal macro-haplogroup L0

Individuals identified with haplogroup L0a, L0d and L0g were distributed according to their birth location coordinates provided on the consent form. A gradient map was created in Mapviewer 8® Golden software, LLC. The gradient map was drawn according to the municipal district that defines the birth location of the individuals.

2.8.5. Phylogenetic network

Haplotype relationships of L0 haplogroups were inferred using a median-joining network constructed with Network version 5.0.0.0 (Fluxus-engineering.com, 2012) to compare the haplotype profiles in the four population groups (Bandelt *et al.*, 1995). The definition of all the individuals' haplotypes was externally provided by DnaSP v5.10.01 (Librado & Rozas, 2009). For the network analysis, the epsilon value was set to zero and the transversions were weighted 3x higher than transitions, as transversion occurs 20 times more infrequent.

For phylogenetic reconstruction, some polymorphic sites within the mitochondrial DNA considered to have a high rate of mutation are not used. Moreover, these sites are generally not an indication of a specific population group (Carracedo *et al.*, 2000). Thus, the following regions and nucleotide positions were excluded from the Network analysis as described by PhyloTree; two poly C-stretches at 303 - 315 and 16184 - 16193, an AC run at 515 - 525, A16182c, A16183c and a mutational hotspot that is a frequent variant that occurs world-wide, T16519C (van Oven & Kayser, 2009).

3.1. FORENSIC MTDNA CONTROL REGION

The two hundred and forty-six biological samples were extracted and amplified as set out in a previously published Magister Scientiae thesis by Heynes, K (2015) and Doctor of Philosophiae thesis by Ristow, P (2017). Once the amplified product was absent of non-specifically amplified DNA fragments, the individuals' samples were sequenced to obtain quality sequenced data. A high-quality mtDNA sequence will result in a chromatogram in which the DNA sequence peaks are clearly visualized as in Figure 3.1. The Sanger sequencing chromatogram produced two to four sequence reads in either direction to produce a consensus sequence. A consensus sequence was generated from the mtDNA samples of the each individual, covering the mitochondrial control region (16 024 - 576bp) for further computational analysis.

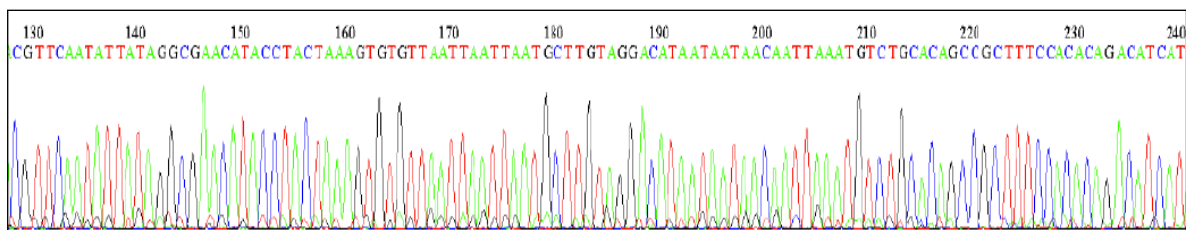


Figure 3.1: Good quality chromatogram of Sanger sequenced mitochondrial DNA. A specifically amplified product, that shows good and clean peak heights (Heynes, 2015).

The consensus sequence of each mtDNA control region was aligned to the revised Cambridge Reference Sequence to recorded a list of SNPs (Andrews *et al.*, 1999). The haplotypes were subjected to a quality control by the EMPOP team and examined with the online tool HaploGrep 2 to obtain the haplogroup of each sample (Kloss-Brandstätter *et al.*, 2010). The haplotypes of all 246 samples with the corresponding haplogroups is shown in Supplementary data, Table A1.

3.1.1. Forensic parameters associated with forensic mtDNA

3.1.1.1 Haplotype analysis for forensic applications

A principle limitation of forensic mitochondrial DNA testing is the low power of discrimination that is obtained from HVR I and HVR II. Most forensic analyses relied on the aforementioned fragments, and occasionally HVR III (Butler, 2012; Lutz *et al.*, 2000).

Each fragment (HVR I, -II and -III) in the mtDNA yields various polymorphic sites that result in a haplotype. The number of haplotypes in the mtDNA sample is thus separated by a definite number of polymorphic sites resulting in the discrimination capacity of the mitochondrial DNA sample. A high discrimination is required as forensic human identification concentrates on similarities and/or differences amongst individuals.

This study expresses the discrimination capacity as the ratio between the number of different haplotypes and the total number of haplotypes. The fragment analysed included: HVR I, HVR I and -II, HVR I to -II and the whole control region as shown in Figure 3.2. The total number of sites analysed increase with each additional fragment. HVR I encompasses 341bp, HVR I and -II includes 608bp, HVR I to -II holds 892bp and the whole control region is made up of 1121bp.

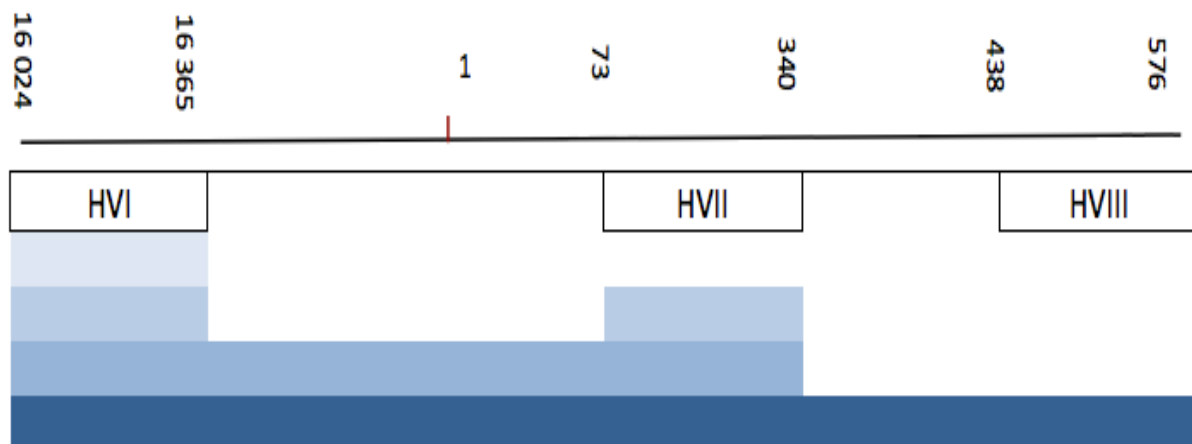


Figure 3.2: Fragment position(s) on the mitochondrial genome that was subjected to resolve the discrimination capacity. The various region(s) examined in the mitochondrial genome includes HVR I, HVR I and -II, HV I to -II and the whole control region which includes HVR III.

The complete South African data set of 246 mtDNA samples was analysed to determine the discrimination capacity of the aforementioned fragments in the mitochondrial genome as illustrated in Figure 3.3.

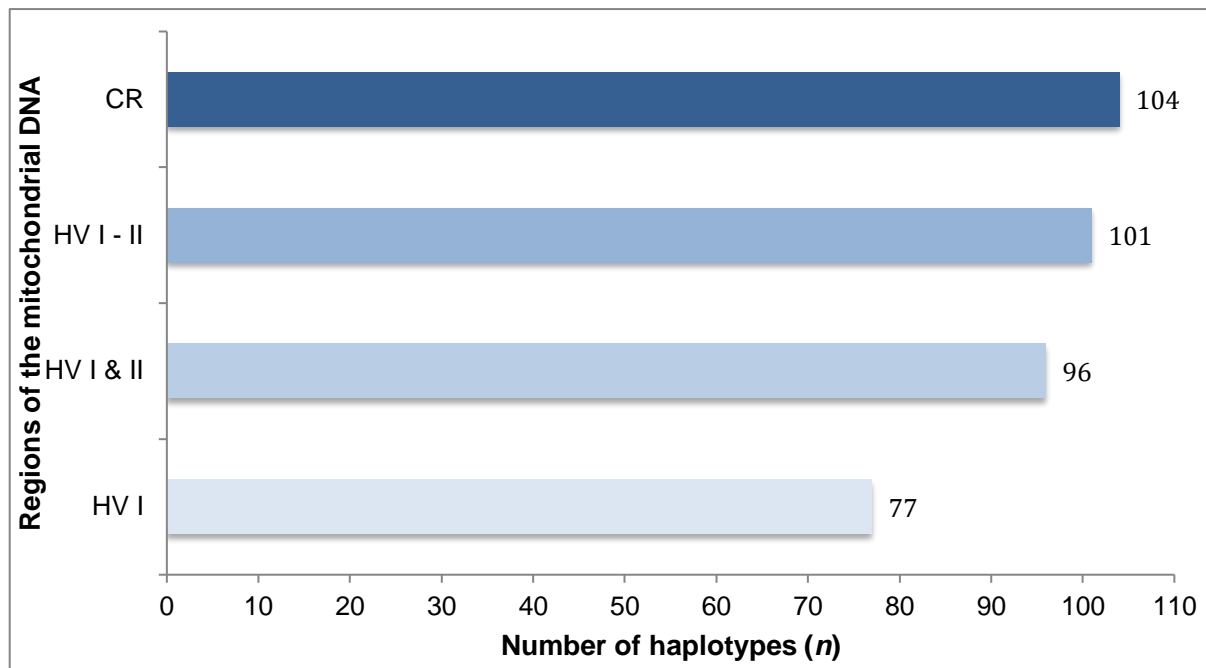


Figure 3.3: The discrimination capacity of the various region(s) of the mitochondrial DNA for human identification in forensic applications. The number of haplotypes (n) identified in the fragment(s) of the mtgenome.

3.1.1.1.1. Haplotype analysis of Hypervariable I

The analysis of Hypervariable I consisted of 69 variable sites that resulted in a total of 77 haplotypes and lead to a discriminatory capacity of 31%. This mitochondrial DNA fragment shows the lowest discrimination capacity amongst all the mtDNA samples. Therefore, it can be concluded that sequencing data acquired solely from hypervariable region I is insufficient in a forensic application aimed at human identification.

3.1.1.1.2 Haplotype analysis of Hypervariable I and II

The analysis of HV I and -II, gave rise to additional fragment analysis. The increased fragment size leads to additional polymorphic sites being analysed, thus an increased discrimination capacity is expected. Figure 3.3 illustrates that this remains true, as 109 variable sites were observed and 96 haplotypes were identified in both the hypervariable regions I, and II. This analysis led to an increase in discrimination capacity from 31% to 39%.

3.1.1.1.3. Haplotype analysis of Hypervariable I to II

The analysis of hypervariable I to -II was preform to evaluate if additional SNPs were identified outside of the hypervariable regions, which could potentially result in additional haplotypes. The region outside the hypervariable region showed 11 additional variable

positions. Hence, 120 variable sites were detected in the analysis of Hypervariable I to - II. Figure 3.3 further shows a total of 101 haplotypes identified, thus 5 additional haplotypes were observed outside the hypervariable regions.

3.1.1.1.4. Haplotype analysis of the whole control region

The mtDNA control region is made up of 1121 sites as showed in Figure 3.2. A total of 104 haplotypes were identified in the control region in the complete South African sample-set. This resulted in a discrimination capacity of 42%, which is a 11% increase in comparison to the analysis of 341 base pairs in HV I. This finding demonstrates the importance of considering all sites upon evaluating differences and similarities. This will allow for a higher discrimination capacity for human identification in forensic applications to differentiate between two individuals.

3.1.2. Frequency estimates of a database

A quality mtDNA haplotype is a fundamental building block in establishing a high-quality forensic database, followed by frequency estimations. It is a common and accepted practice to determine the frequency of the mtDNA in question in a relevant database. To determine a meaningful estimate of haplotype frequency, a large and diverse database is required.

Table 3.1 shows the haplogroups with their respective frequency obtained in this study. The Forensic DNA Laboratory aims at continuing with the collaboration of this project with the EMPOP team and will make the data publicly available on the EMPOP database. The haplogroup column shows sub-Saharan African haplogroups (L0 and L1'5), including the total Eurasian haplogroups.

Table 3.1: MtDNA haplogroup frequencies obtained in this project.

Haplogroup	Frequency % (n)
L0a	2,4 (6)
<i>L0a1b1</i>	0,8 (2)
<i>L0a2</i>	0,4 (1)
<i>L0a2a2</i>	1,2 (3)
L0d	69 (170)
<i>L0d1a</i>	13 (32)
<i>L0d1b</i>	25,6 (63)
<i>L0d1c</i>	1,2 (3)
<i>L0d2a</i>	17,5 (43)
<i>L0d2b</i>	0,8 (2)
<i>L0d2c</i>	6,5 (16)
<i>L0d2d</i>	0,4 (1)
<i>L0d3</i>	4 (10)
L0g₁	0,4 (1)
L1'5	15,9 (39)
<i>L1</i>	2 (5)
<i>L2</i>	3,7 (9)
<i>L3</i>	7 (17)
<i>L4</i>	1,6 (4)
<i>L5</i>	1,6 (4)
Eurasian	12,6 (31)
Total (n)	246

¹ New observation by Chan et al. (2015)

The sample-set in this study showed an exceptionally high frequency of mtDNA haplogroups L0d1 and L0d2a in the South African population. To identify if similar patterns of mitochondrial diversity was observed in similar studies, Table 3.2 and 3.3 shows restricted haplogroup frequencies to a specific population and geographical region for a comparative analysis.

Table 3.2: MtDNA haplogroup frequencies of Coloured individuals in the Western Cape. The data set of this study was limited to the 24 self-declared Coloured individuals residing in the Western Cape to compare to a study conducted by Quintana-Murci et al. (2010) on the identical population and geographical region.

Haplogroup	Western Cape Coloured of this study (%)	Quintana-Murci et al. (2010) (%)
L0a	-	4.44
L0d	70,7	60,03
<i>L0d1</i>	41,6	29,66
<i>L0d2</i>	-	0,36
<i>L0d2a</i>	25	22,20
<i>L0d2b</i>	-	1,07
<i>L0d2c</i>	-	3,37
<i>L0d2d</i>	4,1	-
<i>L0d3</i>	-	3,37
L1'5	16,5	14,8
<i>L1</i>	4,1	1,6
<i>L2</i>	4,1	4,8
<i>L3</i>	-	7,11
<i>L4</i>	-	0,89
<i>L5</i>	8,3	0,36
Eurasian	12,5	20,7
Total (n)	24	569

- Not found

Table 3.3: MtDNA haplogroup frequencies of Coloured individuals in the Northern Cape. The data set of this study was limited to the 73 self-declared Coloured individuals residing in the Northern Cape to compare to a study conducted by Schlebusch et al. (2011) on the identical population and geographical region.

Haplogroup	Northern Cape Coloured of this study (%)	Schlebusch et al. (2011) (%)
L0a	5,25	7,9
L0d		
<i>L0d1</i>	43,8	29
<i>L0d2a</i>	16,4	21,1
<i>L0d2b</i>	1,4	2,6
<i>L0d2c</i>	9,6	2,6
<i>L0d2d</i>	-	-
<i>L0d3</i>	2,7	9,2
<i>L0g1</i>	1,4	-
L1'5	-	
<i>L1</i>	1,4	1,3
<i>L2</i>	2,7	14,5
<i>L3</i>	11	2,6
<i>L4</i>	-	-
<i>L5</i>	-	-
Eurasian	5,4	7,9
Total (n)	73	76

- Not found

¹ New observation by Chan et al. (2015)

The study of Quintana-Murci et al. (2010) and Schlebusch et al. (2011) confirms the high frequencies observed for haplogroups L0d1 and L0d2a. Hence, a similar pattern of mitochondrial diversity was observed. Moreover, the overall high frequency of L0d in this study, as well as similar studies conducted on the South African population groups, showed the substantial maternal contribution of the KhoiSan people to the South African gene pool.

Once forensic scientists have generated the frequency estimate in a database using a counting method (a process that involves counting the number times that the shared profile matched a profile in the database), the analyst next estimates the rarity of the profile in a population group. This is based on the number of observations in each population group which is categorized by self-reported ancestry.

3.1.3. Random match probability

The random match probability (RMP) is an assessment that defines the probability of an unrelated person, chosen at random from a general population and matching the genotype from the evidence sample. This probability ranges from 0 to 1. This estimate is of value as it is influential evidence in a criminal jury trial. To our knowledge, no forensic statistical parameters of the mitochondrial DNA have been reported for South Africa. Therefore, this study, reports the random match probability below for the Coloured, Griqua, Bantu and Nama populations.

Table 3.4: Random match probability of the South African ethnic groups.

Forensic parameter	Coloured [<i>n</i> = 138]	Griqua [<i>n</i> = 88]	Bantu [<i>n</i> = 10]	Nama [<i>n</i> = 10]
Random match probability	0,0410	0,0423	0,04	0,04

Table 3.4 illustrates a random match probability of less than 0,05 in all the South African populations. A probability of 0,05 would suggest a randomly selected individual only has 5 in 100 chances of sharing a profile with the evidence profile reported. Therefore, it would be recommended that the mitochondrial DNA is more powerful for the purpose of exclusion in South African populations. It was further found, that a study conducted by Fendt et al. (2012) focused on KhoiSan lineages (a predominant lineage in this study) found a higher RMP of 0,09. Other lineages represented in this study were identified as European – and Asian lineages. A random match probability of 0,012 and 0,028 has been

reported for European and Asian population groups, respectively (Lan *et al.*, 2019; Vanecek *et al.*, 2004).

It is important to note that this frequency is more accurate in a geographically diverse and randomly selected sample-set. Therefore, the random match probability for the Bantu and Nama populations can be more accurately determined with a wider geographical sample collection, as the two populations only represents Northern Cape samples.

To further assess the rarity of a mitochondrial DNA profile, the forensic science community has to produce frequencies in a particular geographical and population group. The following section (3.2) of this study focusses on the extent of geographical and population related substructure in the distribution of mitochondrial DNA profiles of South Africa.

3.2. POPULATION GENETICS OF MTDNA CONTROL REGION

Population genetics is defined as a study of genetic variation within and between populations. This is also referred to as a phylogenetic analysis. However, in terms of a forensic database, this will only reflect the haplogroups that exist in a relevant population. To determine an accurate frequency of a profile, forensic scientists acquire further knowledge on the geographical distribution of the mitochondrial DNA haplotypes within the population. Therefore, section 3.2 aims at evaluating a comprehensive study on the phylogenetic and phylogeographical analysis of the South African populations with the same sample-set of 246 that were studied in section 3.1. This section presents knowledge on the spectrum and area-specificity of major haplotypes and haplogroups within South Africa.

3.2.1. GEOGRAPHICAL STRUCTURE OF MTDNA

3.2.1.1. Sample collection coordinates

Figure 3.4 illustrates the three points at which the samples were collected from donor individuals with the respective sample size in South Africa.

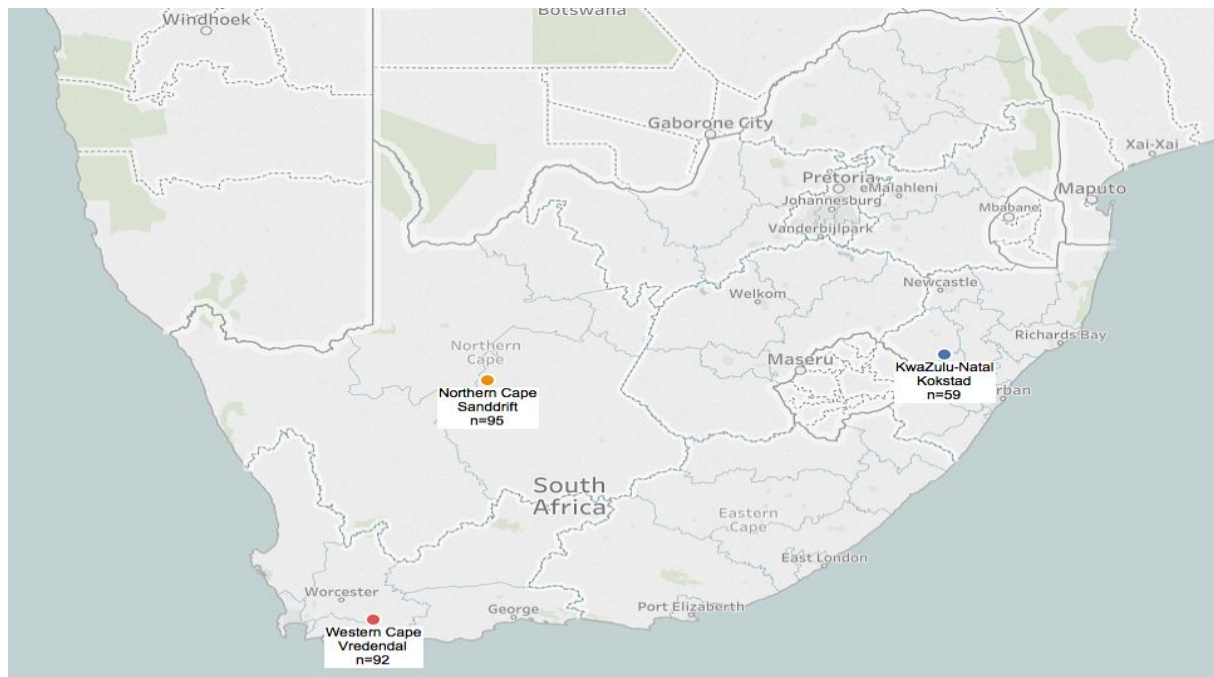


Figure 3.4: Three geographical locations in which a total of $n=246$ biological samples were collected. This includes individuals residing in the town of Kokstad, KwaZulu-Natal (-30.5096°S, 29.4063°E), Sanddrift, Northern Cape (-28.41°S, 16.77491°E) and Vredendal in the Western Cape (31.6391°S, 18.5285°E) province of South Africa (Created in Tableau Public:<https://public.tableau.com/en-us/s/>).

3.2.1.1.2. Comparative structure analysis of 3 geographical locations

3.2.1.1.2.1. Mitochondrial DNA haplotype identification

The 1121bp control region of all the individuals' sequences was examined to identify identical patterns of SNPs to determine a haplotype. The haplotype profiles were categorized according to the province in which they were born i.e. Western Cape, Northern Cape and KwaZulu-Natal, to infer the genetic structure.

Table 3.5 illustrates the haplotype profiles that were identified as shared across all three provinces, with the frequency at which it was observed. Hence, individuals that were born in three separate geographical locations share a common ancestor with identical sequence patterns. A total of four haplotype profiles were observed that shared profiles amongst the Western Cape, Northern Cape and KwaZulu-Natal province. The first two haplotypes represented haplogroup L0d2a1 (in which the haplotype were either absent of or presented a 309.1C insertion), at a high frequency of 10,9%. This frequency is a combined total of the two haplotypes, as for the purpose of evolutionary and forensic studies, the insertion does not constitute evidence for excluding two identical haplotypes as deriving from the same maternal lineage as outlined by Parson et al., 2014, recommendation 10. Meanwhile, the lowest frequency were 1,6% for haplogroup L0d1b2b.

The individuals that were identified in each of the shared haplotype profiles among the three geographical regions were further categorized by ethnicity. This were summarized in the Supplementary data, Table A1.

Meanwhile, Table 3.6 illustrates the three provinces in South Africa in which the maternal haplotype was observed in only one of the provinces, in a frequency of $n \geq 3$. Hence, the other two provinces were absent from the specific haplotype profile. Table 3.6 shows haplogroups L2a1a2, L0d3b and L0d1b2b with the respective haplotype in the Western Cape province at a relative frequency of 0,033, 0,065 and 0,065, respectively. The Northern Cape represents haplogroup L3e1b2 with the respective haplotype profile at a frequency of 0,032. Furthermore, haplogroup M3 and H15 were present in the KwaZulu-Natal province only at a frequency of 0,085 and 0,102, respectively.

This analysis shows that human mtDNA displays a regional distinction. The shaping of human regional mtDNA variation is traditionally attributed to genetic drift, which is the

effect of random changes in the gene pool. However, it can be hypothesized that the history of southern Africa had an influence on the distribution of mitochondrial DNA lineages in South Africa.

Table 3.5: Shared haplotypes identified in the human mtDNA amongst the 3 geographical locations. The frequency column shows the frequency observed in the Western Cape (WC), Northern Cape (NC) and KwaZulu-Natal (KZN) province as obtained in the program Arlequin v. 3.5.2.2.

Haplotype Profile	Haplogroup	Total Frequency	<i>n</i>		
			WC	NC	KZN
16129A 16187T 16189C 16212G 16223T 16230G 16243C 16311C 16390A 16519C 73G 146C 152C 195C 198T 247A 309.1C 315.1C 498DEL 522DEL 523DEL	L0d2a1	8.1%	4	10	6
16129A 16187T 16189C 16212G 16223T 16230G 16243C 16311C 16390A 16519C 73G 146C 152C 195C 198T 247A 315.1C 498DEL 522DEL 523DEL	L0d2a1	2.8%	4	1	2
16129A 16187T 16189C 16223T 16230G 16239T 16243C 16294T 16311C 16519C 73G 146C 152C 195C 247A 309.1C 315.1C 498DEL	L0d1b2b	1,6%	2	1	1
16129A 16187T 16189C 16218T 16223T 16239T 16243C 16294T 16311C 16519C 73G 146C 152C 195C 247A 315.1C 498DEL 523DEL 524DEL	L0d1b2b2b1	4,5%	7	2	2

Table 3.6: Haplotypes that were only identified at $n \geq 3$ in one of the three provinces of South Africa. The last column shows the province (i.e. WC = Western Cape, NC = Northern Cape and KZN = KwaZulu-Natal) in which the haplotype was identified as obtained in the program Arlequin v. 3.5.2.2.

Haplotype	Haplogroup	Province	Frequency (<i>n</i>)
16223T 16278T 16286T 16294T 16309G 16390A 16519C 73G 146C 152C 195C 263G 315.1C	L2a1a2	WC	0,033 (3)
16187T 16189C 16214T 16223T 16230G 16243C 16274A 16278T 16290T 16300G 16311C 16519C 73G 146C 150T 195C 247A 315.1C 316A	L0d3b	WC	0,065 (6)
16129A 16187T 16189C 16223T 16230G 16239T 16243C 16294T 16311C 16519C 73G 146C 152C 247A 315.1C 498DEL	L0d1b2b	WC	0,065 (6)
16223T 16239T 16325DEL 73G 150T 185A 189G 263G 309.1C 315.1C	L3e1b2	NC	0,032 (3)
16126C 16223T 16519C 73G 263G 309.1C 315.1C 482C 489C	M3	KZN	0,085 (5)
55C 57C 263G 309.1C 309.2C 315.1C	H15	KZN	0,102 (6)

3.2.1.2. The Maternal haplogroup composition of South African provinces

The maternal haplogroup frequency and distribution of the Western Cape, Northern Cape and KwaZulu-Natal province is presented in the following three sub-sections. The haplogroups are shown in Supplementary data, Table A2 in which VGRI- represents the Western Cape haplogroups, MR14- and N13 the Northern Cape collection and GRI- the KwaZulu-Natal provinces.

3.2.1.2.1. Haplogroup composition of the Western Cape

Figure 3.5 illustrates the haplogroup composition of the Western Cape province taking into account a sample size of 92 individuals. The Western Cape province displays the highest frequency of maternal haplogroup L0 in comparison to the other two provinces in South Africa.

Haplogroup L0 encompasses L0d and L0a in dark and light blue, respectively. Haplogroup L0d makes up 78,2% of the bar graph, indicative of KhoiSan-ancestry. This high frequency reflects the immense maternal influence of the KhoiSan population in the Western Cape province. This finding is further due to the province being home to the majority of the South African Coloured population, whose gene pool consists mainly of KhoiSan heritage (Quintana-Murci *et al.* 2010; Statistics South Africa, 2012).

Furthermore, the Western Cape includes haplogroups L1, L2, L3 and L5 clades at a frequency of 11%, indicative of Bantu-ancestry. It is believed that these maternal haplogroups were introduced to southern Africa during the Bantu expansion. However, presently, these lineages are also observed among the KhoiSan population. Therefore, the immense maternal presence of the KhoiSan population in the Western Cape could reflect that haplogroups L1, L2, L3 and L5 are either directly contributed by the Bantu people or an indirect Bantu contribution through admixture with the KhoiSan, who received these lineages through their previous admixture with Bantus (Quintana-Murci *et al.*, 2010).

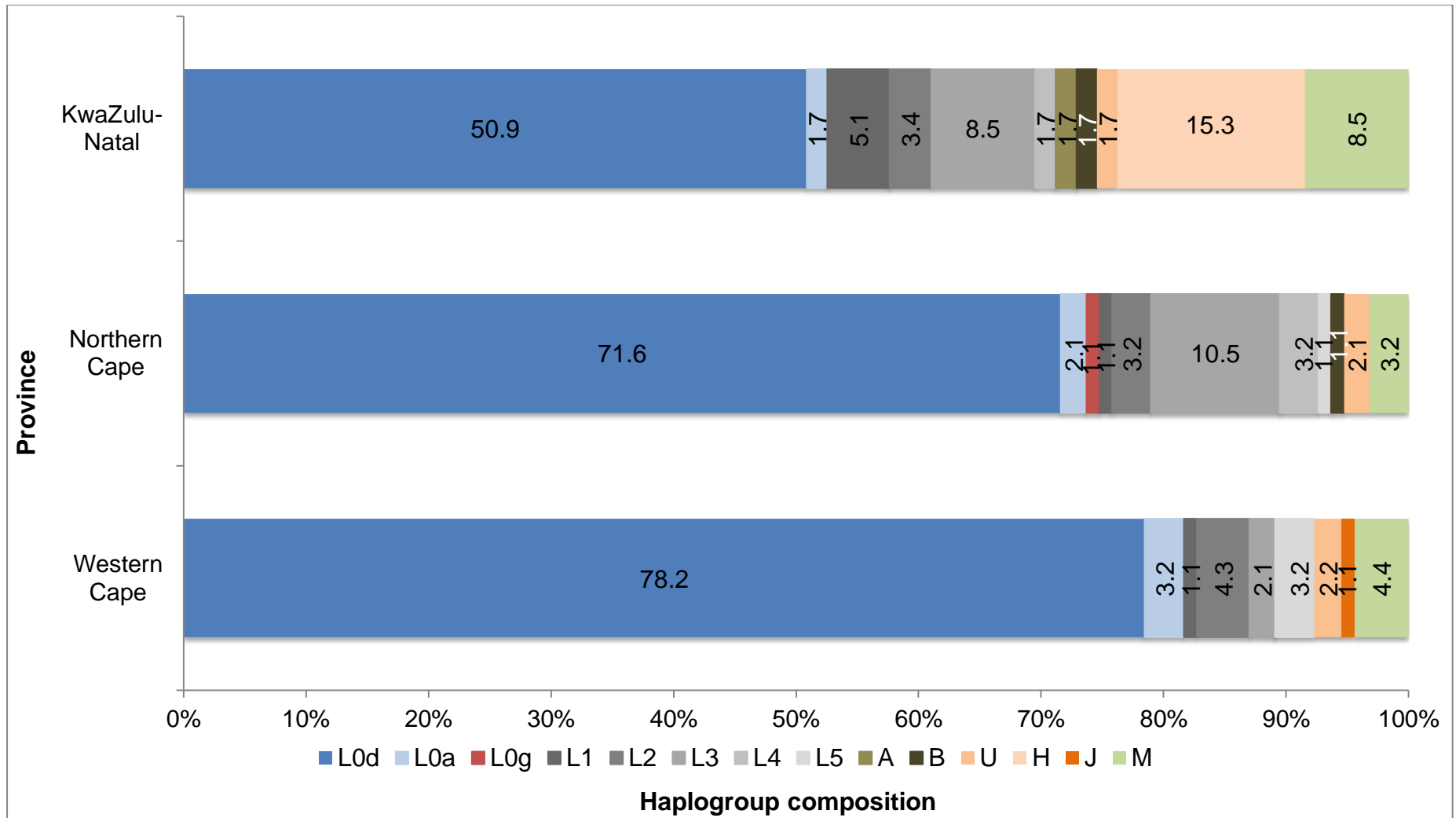


Figure 3.5: Bar graph depicting the mitochondrial haplogroup composition of the Western Cape, Northern Cape and KwaZulu-Natal province. This is an indication of the various haplogroups in different frequencies found in all three studied provinces of South Africa.

Table 3.7: An in-depth look at the out of Africa haplogroups found in the Western Cape province.

European ancestry (HG)	<i>n</i>	Asian ancestry (HG)	<i>n</i>
J1c1	1	U2c'd	2
		M53	1
		M41a1	1
		M26	1
		M51b1	1

The remaining portion of the bar graph represents European ($n=1$) and Asian ($n=6$) ancestries, as illustrated in Figure 3.5. Table 3.7 identifies the maternal haplogroups J, U and M. The European ancestry suggests admixture between the European settlers that arrived in the Cape of Good Hope located in the Western Cape province. It can be noted that maternal haplogroup J, is exclusively observed in the Western Cape during this investigation. Meanwhile, the Asian-ancestry (haplogroup U2c'd, M53, M41a1, M26 and M51b1) has been reported by previous publications to be of Indian descent (Chandrasekar *et al.*, 2009; Dubut *et al.*, 2009; Quintana-Murci *et al.*, 2004; Thangaraj *et al.*, 2006). This ancestry could be an indication of the slaves that were brought to the Cape Colony between 1653 and 1822 (Jonas, 2011).

3.2.1.2.2. Haplogroup composition of the Northern Cape

Figure 3.5 illustrates the haplogroup composition of the Northern Cape province taking into account a sample size of 95 individuals. The bar graph shows the second-highest frequency at 71,6% of maternal haplogroup L0. This is an indication of KhoiSan-ancestry. Moreover, the Northern Cape is the only province in which L0g was observed.

The maternal haplogroup L0g was initially identified by Chan, E.K.F *et al.* (2015), from an individual residing in Namibia. The paper speculated the haplogroup as a probably linguistically or culturally extinct independent KhoiSan-specific lineage that diverged from L0a approximately 93.8 kya.

The Northern Cape further comprises all the Bantu clades (maternal haplogroup L1'5) at a frequency of 19,1%. However, due to the immense KhoiSan influence in the Northern Cape the Bantu-ancestry haplogroups reflect either a direct gene flow from the Bantu or indirect from the KhoiSan individuals as observed in the Western Cape province.

Furthermore, the Northern Cape province displays a high frequency at 10,5% of maternal haplogroup L3. This haplogroup contains many sub-Saharan Africans lineages and also ancient non-African lineages. Hence, it has a strong association with the out-of-Africa migration of modern humans, however, it is also inherited by some populations in Africa (Behar *et al.*, 2008; Soares *et al.*, 2012). The bar diagram further presents out of Africa maternal haplogroups that are shown more descriptively in Table 3.8.

Table 3.8: An in-depth look at the out of Africa haplogroups found in the Northern Cape province.

European ancestry (HG)	<i>n</i>	Eurasian ancestry (HG)		Asian ancestry	<i>n</i>
U5a1c1a	1	B4c1b2a2	1	U2c'd	1
				M1a5	3

Maternal haplogroup U descends from a woman in the haplogroup R branch of the phylogenetic tree (van Oven & Kayser, 2009). It use to represent a dominant type of mtDNA in Europe. Today, the various subclades are found widely distributed. Haplogroup U2c'd in this study were first observed in the Western Cape province by 2 individuals (Indian descend) and presents a shared haplotype with an individual born in the KwaZulu-Natal province. Meanwhile, haplogroup U5 is spread widely although at low levels throughout Europe.

Furthermore, the Eurasian-ancestry (mixed Asian and European-ancestry) is indicated by haplogroup B4c1b2a2. The B4c1b2a2 haplogroup has been suggested to be of South Chinese origin (Brandão *et al.*, 2016).

Lastly, haplogroup M1a5 was found in three individuals in this study. Over the years controversy was raised regarding the origin of the M1 haplogroup. It has been suggested by Gonzalez *et al.* (2007) that the macro-haplogroup M1 originated in Asia and represents a backflow to Africa. However, other studies believe that the M1 haplogroup originated in Africa (Quintana-Murci *et al.*, 1999 Sun *et al.*, 2006). In this study, the two individuals are classified as Asian descent.

3.2.1.2.3. Haplogroup composition of KwaZulu-Natal

The KwaZulu-Natal province consisted of a sample size of 59 individuals, by self-declared Coloured and Griqua individuals (refer to Table 2.1). Therefore, a high frequency of indigenous ancestry was hypothesized. However, KwaZulu-Natal in comparison to the

aforementioned provinces illustrates the lowest indigenous KhoiSan-ancestry (haplogroup L0) at 50.9% in Figure 3.5.

The frequency of L0 in Kwazulu-Natal could suggest sex-linked migration rates. Alternatively, this could be due to the encounters with various tribes around South Africa as the early Coloured and Griqua individuals originated in the Western Cape. Therefore, the present-day Coloured and Griqua populations along the inland and East coast areas of South Africa could display a lower frequency of indigenous ancestry.

Further, the low frequency of KhoiSan-ancestry observed could also be due to the smaller sample size representing this province. Meanwhile, KwaZulu-Natal presents the highest frequency of Bantu-ancestry (maternal haplogroups L1, L2, L3 and L4), as pioneering groups of the Bantu expansion reached the modern KwaZulu-Natal by C.E. 300 (Newman, 1995).

Table 3.9: An in-depth look at the out of Africa haplogroups found in the KwaZulu-Natal province.

European ancestry (HG)	<i>n</i>	Asian ancestry (HG)	<i>n</i>
H15	7	A14	1
H7a1	1	B5b1c	1
H101	1	M3	5
U2e1	1		

Table 3.9 illustrates a frequency of $n=10$ European-ancestry and $n=7$ Asian-ancestry observed in KwaZulu-Natal. The European-ancestry shown in the first column is demonstrated by maternal haplogroup H and U, which is currently predominantly found in Europe. The second column displays haplogroups of Asian descent composed of haplogroup M3, A14 and B5b1c. Maternal haplogroup M3 was found to be an Indian-specific lineage (Maji *et al.*, 2009). While haplogroup A and B can be observed in American populations, however, the specific haplogroup A14 and B5b1c is derived from Asian populations (Noguera-Santamaría *et al.*, 2015).

3.2.2. COMPARATIVE STRUCTURE ANALYSIS AT AN INTRA-POPULATION LEVEL

3.2.2.1. The maternal haplogroup composition

The population groups are categorized according to the self-declared ethnic population that was provided on the consent form by each donor individual. The population groups investigated in this study were the admixed Coloured and Griqua and the native Bantu and Nama with sample sizes $n = 138, 88, 10$ and 10 , respectively. The maternal haplogroup frequency of each population group was analysed and illustrated in Figure 3.6.

3.2.2.1.1. Admixed: Coloured and Griqua

Figure 3.6 illustrates the immense influence of the indigenous KhoiSan maternal haplogroups L0d, L0a and L0g in the Coloured and Griqua population. Here, it is observed that the L0g haplogroup was presented in the Coloured population.

The Coloured population further illustrates all Bantu lineages (L1'5), while the Griqua population was absent of haplogroups L1 and L4. Moreover, both populations encompasses European-ancestry. It is noteworthy that the exclusive haplogroup J observed in the Western Cape (under *Geographical structure of mtDNA, Sub-section 3.2.1.2.1*) were acquired from an individual of self-declared Griqua ethnicity.

Lastly, the Coloured population displayed Asian maternal haplogroup A, B and M, while the Griqua population consisted only of haplogroup M. Overall, the Coloured population of South Africa illustrates the highest degree of admixture.

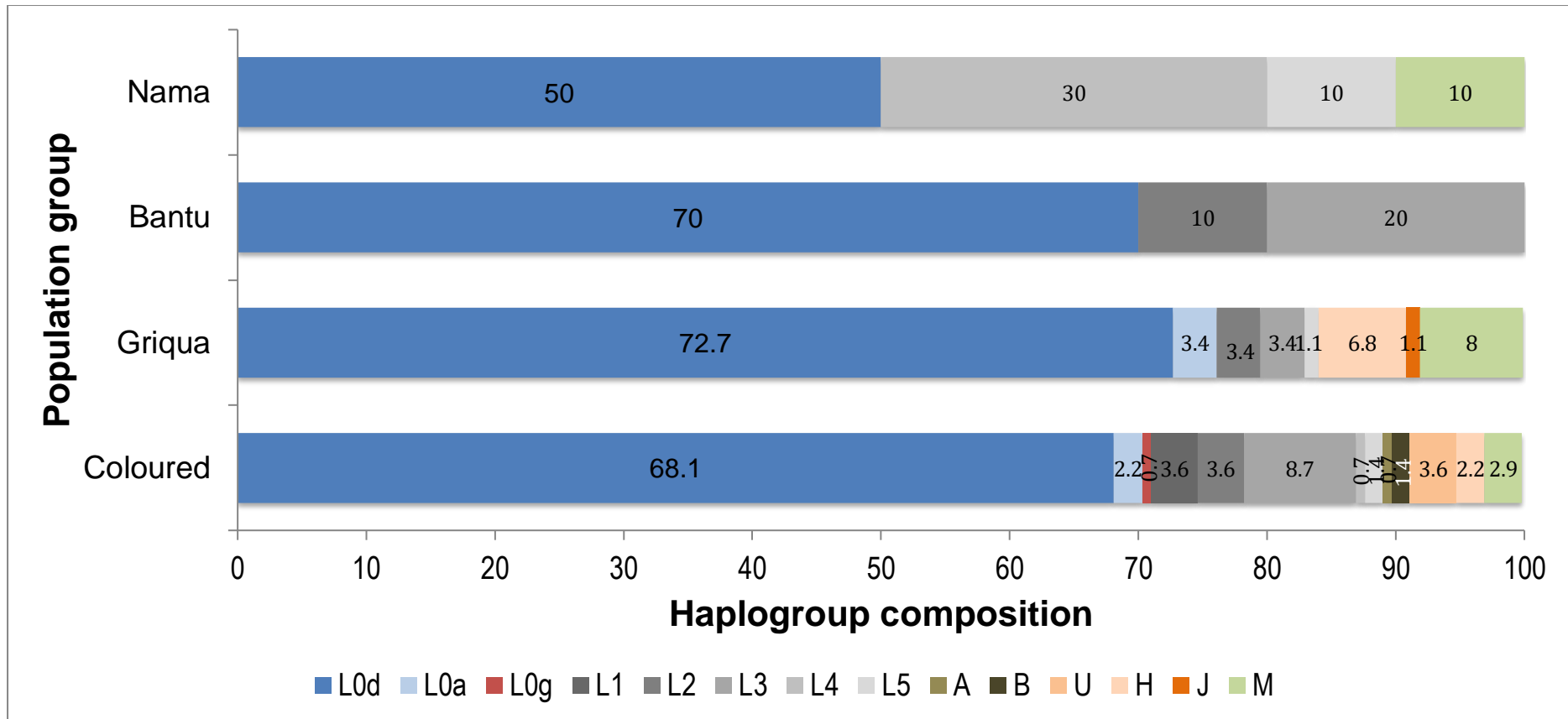


Figure 3.6: Bar graph depicting the mitochondrial haplogroup composition of all the South African population groups (n=246). This is an indication of the various haplogroups present in the Coloured, Griqua, Bantu and Nama population at different frequencies.

3.2.2.1.2. Native: Bantu and Nama

The Bantu and Nama populations both illustrated haplogroup L0d, with the Nama population representing the lowest frequency of the maternal haplogroup. It is further noticeable, that both population groups are absent of haplogroup L0a. Similarly, a paper by Schlebusch et al. (2013), had a Nama sample size of $n = 28$ and found no L0a maternal haplogroup contribution. However, L0a has been proposed as a marker of the Bantu expansion 3000-4000 years ago (Bandelt *et al.* 1995; Chen *et al.* 1995) and found in relatively high frequencies in the southern African Bantu-speakers (Barbieri *et al.*, 2014). The absence of this haplogroup could be due to the small sample size that were investigated for the Bantu population. Therefore, the Bantu sample-set could be absent from certain Bantu-speaking groups.

The investigation of Bantu-ancestry in the Bantu and Nama population shows haplogroups L2 & L3 and L4 & L5, respectively. It is noted that the Nama population comprises a high L4 haplogroup frequency (30%), while Schlebusch et al. (2013) found no evidence of this haplogroup in the Nama population.

No evidence was found for European lineages in the native population groups. The Bantu further showed no indication of Asian-ancestry, however Asian-ancestry were present in the Nama population by one individual. Meanwhile, the study of Schlebusch et al. (2013) illustrated that the Nama population was absent from Asian-ancestry and supports the absences of European-ancestry.

3.2.2.2. South Africa population differentiation

The above section shows that the frequencies of mtDNA ancestries have some variation in the various population groups of South Africa. Therefore, it is hypothesized that a heterogeneous nature will be observed in the population groups due to the influence of historical colonization processes and continues mixing in South Africa.

To confirm or reject this statement, an Analysis of Molecular Variance (AMOVA) method was performed in the program Arlequin v. 3.5.2.2. The haplotype profiles of the Coloured, Griqua, Bantu and Nama individuals defined the genetic structure. The method detects for population differentiation among the four populations using the number of mutations between haplotypes by inferring a distance matrix (Excoffier *et al.*, 1992).

Table 3.10 categorize the source of variations as follows: "among populations" and "within populations". The two tests directly compares the haplotype profiles of the population groups and the differences within elements of the same population, respectively. The analysis shows that 1,20% of the variation were distributed among the four populations, while a high of 98,80% of variation were detected within the populations.

Table 3.10: Non-hierarchical AMOVA analysis of the four population groups of South Africa. Each group: Coloured, Griqua, Bantu and Nama individuals are classified as a population on its own. The results were obtained by Arlequin v.3.5.2.2.

Source of variation	d.f.	Sum of squares	Variance components	Percentage of variation
Among populations	3	23,469	0,06113 V_a	1,20
Within populations	242	1221,067	5,04573 V_b	98,80
Total	245	1244,537	5,10687	

Fixation Index (F_{ST}): 0,01197

Significance tests

V_a and F_{ST} : $P(\text{rand. value} > \text{obs. value}) =$ 0,02743
 $P(\text{rand. value} = \text{obs. value}) =$ 0,00000
 P-value = 0,02743 + -0,00150

The average variability among the populations were compared to the average variability within the populations with a fixation index. The F_{ST} of 0,01197 (where $P < 0,05$) indicate that the Coloured, Griqua, Bantu and Nama populations are not significantly different from one another and may indicate shared ancestry. This could be due to the highly prevalent L0d haplogroup and influential gene flow from the Bantu lineages (L1'5) in all the population groups.

3.2.2.2.1. Shared and unique haplotypes in South African populations

The complete sample-set were further examined for shared and unique haplotypes among the four population groups. A total of 52, 25, 4 and 5 unique haplotypes were observed in the Coloured, Griqua, Bantu and Nama population, respectively in Figure 3.7.

The highest shared haplotype total of 12 were observed between the admixed Colored and Griqua. While, the native Bantu and Nama did not illustrate any shared haplotypes. However, it should be considered that this investigation could be biased due to the low sample size of individuals in the native population groups. Overall, the four populations share only 1 haplotype, linked to haplogroup L0d2a1.

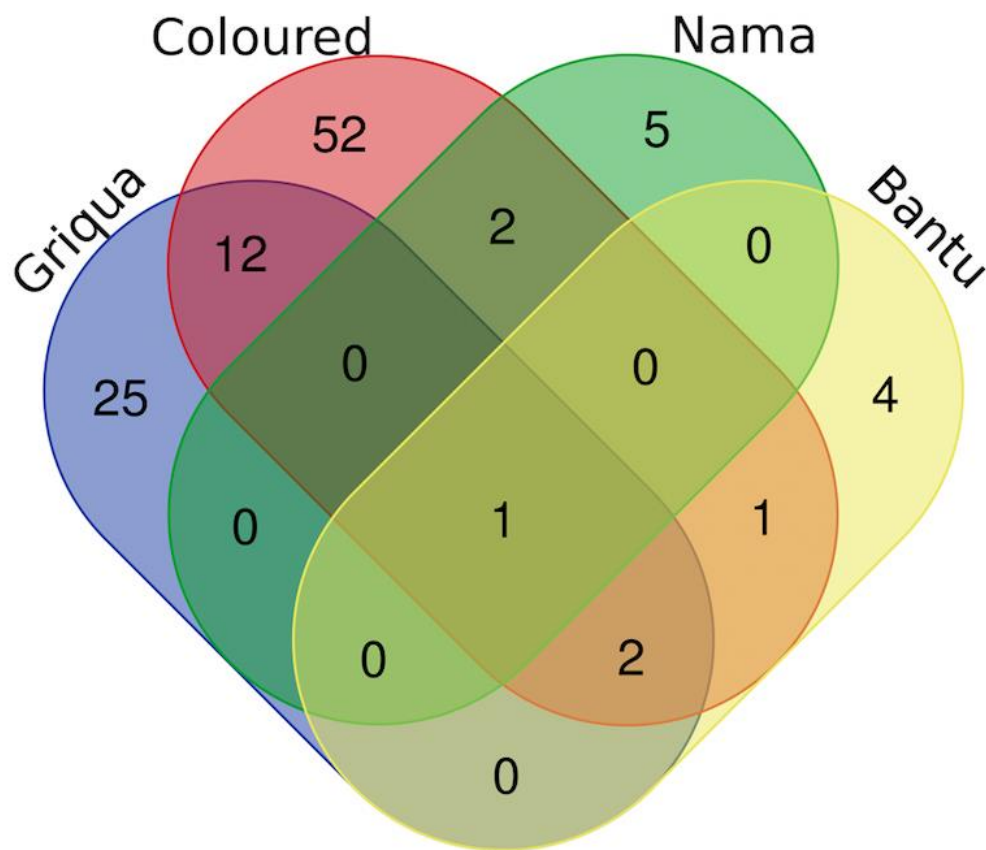


Figure 3.7: Venn Diagram illustrating the shared and unique haplotypes among the Coloured, Griqua, Bantu and Nama populations. (Created with: <http://bioinformatics.psb.ugent.be/webtools/Venn/>).

3.2.2.3. Haplotype and nucleotide diversity

Table 3.11 represents a population statistical overview of each respective population group in this study. Each population is accompanied by the sample size (n) that corresponds to the total number of individuals that samples were collected from. The statistical analysis of each population includes the following: the variable sites (S), otherwise known as segregating sites that denotes the number of polymorphic sites during a sequence alignment assessment. Additionally, the total number of haplotypes were calculated. The last two rows illustrate the genetic and nucleotide diversity referring to the degree of nucleotide polymorphisms and the variation in the amount of genetic information within each population.

Table 3.11: A descriptive statistic for four populations from South Africa. Polymorphism overview of each population group using the individuals control region sequences.

Population statistics	Coloured [$n = 138$]	Griqua [$n = 88$]	Bantu [$n = 10$]	Nama [$n = 10$]
Variable sites (S)	127	88	34	47
Haplotypes (h)	71	41	8	8
Gene diversity	0,95895	0,95687	0,95556	0,95556
Nucleotide diversity (π)	0,00982	0,01049	0,01125	0,01437

3.2.2.3.1. Polymorphic overview of the Coloured

In Table 3.11, the Coloured population illustrates a high percentage of different haplotypes ($71/138=51,45\%$) with the lowest nucleotide diversity of 0,00982, in comparison to the other population groups. This statistical evaluation reflects a history of recent demographic expansions from small population groups. This study therefore expands and supports the current knowledge from historical records, in that the modern South African Coloured is mainly the result of the early encounter between the European and African males with indigenous KhoiSan females of the Cape of Good Hope around 350 years ago (Quintana-Murci *et al.*, 2010).

Quintana-Murci *et al.* (2010) further reported a gene diversity of approximately 0,9 based on HVS-I haplotypes in the Coloured population, while this study observed a 0,95895 gene diversity index based on the whole control region (Quintana-Murci *et al.*, 2010).

3.2.2.3.2. Polymorphic overview of the Griqua

The Griqua population shows a high haplotype of 41 (percentage of $41/88=46,59\%$) and nucleotide diversity of 0,01049 illustrated in Table 3.11. Thus, presenting a large population with a stable effective size through their evolutionary history.

3.2.2.3.3. Polymorphic overview of the Bantu

It can be noted that the polymorphic overview of the Bantu and Nama (*this sub-section including sub-section 3.2.2.3.4 below*) may introduce biases due to the low sample size of individuals in these population groups. Therefore, the haplotype and nucleotide diversity could be falsely increased or decreased (Nei, 1987). Nevertheless, the Bantu population expresses a high haplotype percentage ($8/10=80\%$) and high nucleotide diversity, suggesting that the Bantu population presents a large population with a stable effective size through their evolutionary history.

3.2.2.3.4. Polymorphic overview of the Nama

The same polymorphic overview is presented in the Nama population in Table 3.11 as observed in the Griqua and Bantu population. The Nama population expresses a high haplotype percentage ($8/10=80\%$) and high nucleotide diversity of 0,01437. Therefore, the same concept is accepted in the Nama population as described for the Griqua and Bantu.

3.2.2.4. Neutrality tests

In population genetics it is not adequate to report a phenomenon, the data available has to be quantified and assign probabilities, and explore how probable the events are under certain scenarios, such as a population expansion or neutrality. To measure this, the study introduced two neutrality tests: (i) Tajima's D and (ii) Fu and Li's F (Fu and Li 1993; Tajima 1989). The two tests are mainly used to evaluate past demographic or selection scenarios and sensitive to recent population size expansion. (Fu and Li 1993; Tajima 1989).

The Tajima's D statistic is often used for detecting population size expansion and/or positive or negative selection. The equation has the advantage of having little affect by the sample size or the number of segregating sites in the critical region of the formula (Tajima, 1989). Additionally, a neutrality test derived by Fu and Li shares much information with Tajima's D. However, population genetic scenarios for selective sweeps (tested by Tajima's D), tend to generate excess of singletons (segregating sites at which the rare

SNPs is only represented once), in this event the Fu and Li is more sensitive than Tajima's D.

Table 3.12: Neutrality test of Tajima's D and Fu & Li's F according to each population group. The data presented in this table was inferred by DNAsp v5.10.01. Each statistical value is accompanied by its P-value illustrated directly to the right of the respective neutrality test.

Population group	Neutrality Test				
	Tajima's D	P-value	Fu & Li's F	P-value	Statistically Significant
Coloured	-1,82802	P<0,05	-2,65369	P<0,05	Y
Griqua	-1,20209	P>0,10	-2,07624	0,10> P >0,05	N
Bantu	0,21567	P>0,10	0,40224	P>0,10	N
Nama	-0,17374	P>0,10	-0,38773	P>0,10	N

Table 3.12 illustrates the Coloured population as the only population to have obtained a significant Tajima's D ($P < 0,05$) and Fu & Li's F value ($P < 0,05$). The negative Tajima's D value in the Coloured population could be an indication of excess low-frequency polymorphism, which correlates with the low-nucleotide diversity of 0,00982 in Table 3.11. This signifies an expansion of the Coloured population and therefore supports the finding in *sub-section 3.2.2.3.1*.

3.2.3. STRUCTURE BY MATERNAL HAPLOGROUP L0

3.2.3.1. Distribution of the indigenous people

The most pronounced finding during the population genetics investigation is the high frequency of the deep rooting clade, L0 macro-haplogroup of the human mtDNA phylogeny in South Africa. Therefore, a gradient map was designed to show the distribution of L0 in South Africa. This includes the most prevalent maternal haplogroup L0d, along with L0a and L0g. The distribution of haplogroup L0 in South Africa encompasses a total of $n=177$ individuals identified in this study. The gradient map was created utilizing the birth location coordinates that were provided on the consent form. The gradient map was drawn according to the municipal district that defines the birth location of the 177 individuals.

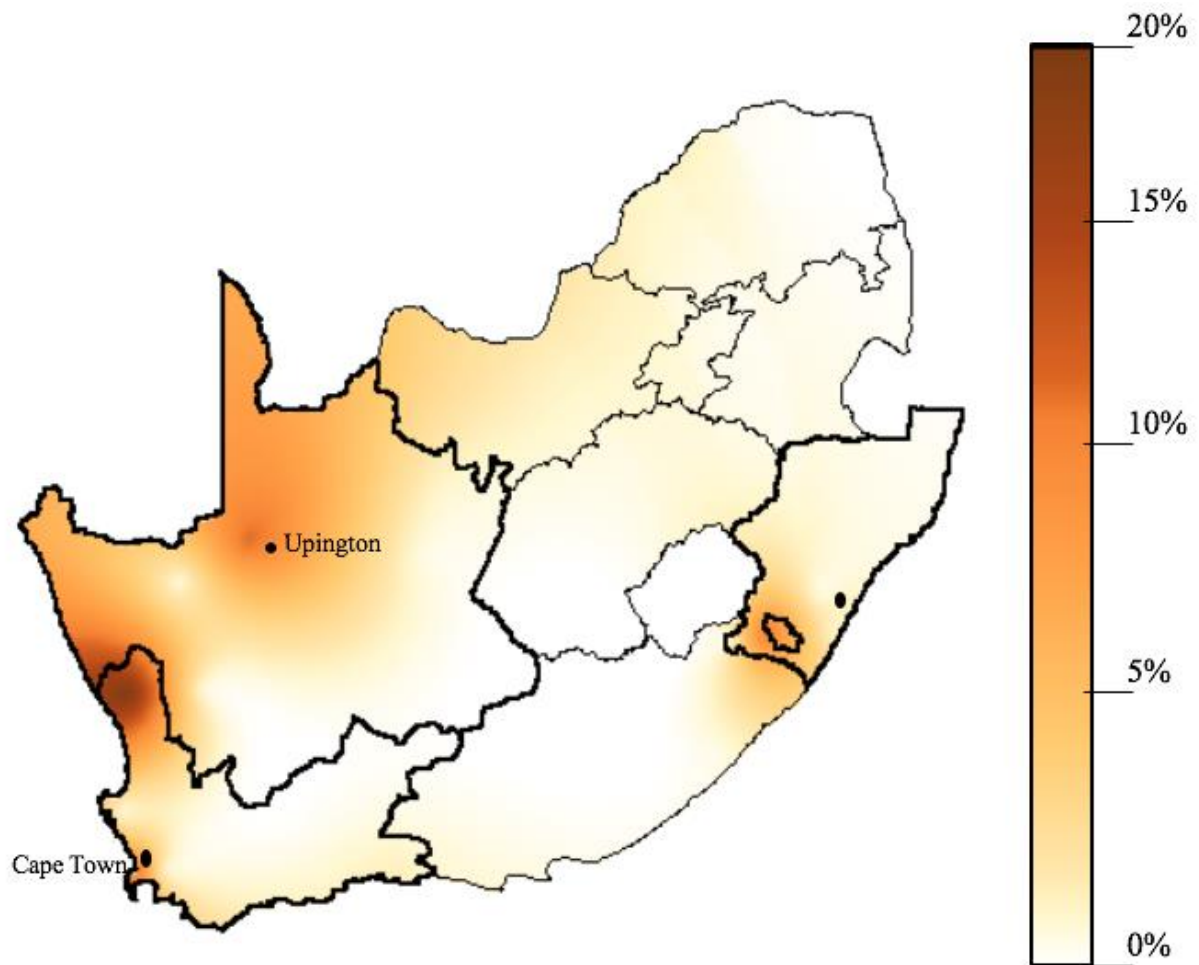


Figure 3.8: Gradient map illustrating the mtDNA distribution of macro-haplogroup L0 in South Africa representing 177 individuals. A colour scale at the right side of the figure defined the percentage of individuals observed in the area (Created with: MapViewer® Golden Software, LLC).

The map in Figure 3.8 illustrates two noteworthy findings. First, a high degree of maternal haplogroup L0 is observed in the Western Cape province. This is displayed on the border of the Northern Cape on account of the sample collection that was conducted in the Western Cape and Northern Cape region. This high indigenous ancestry was anticipated due to the Coloured and Griqua individuals that originated in the Western Cape and the Northern Cape that encompasses samples of native Bantu and Nama.

Additionally, it is significant that a high level of macro-haplogroup L0 was observed near the border of Namibia and Botswana. This finding is in correspondence with the territories in which the KhoiSan indigenous communities presently resides in southern Africa.

The gradient map lastly illustrates the low level of indigenous people around and near KwaZulu-Natal province as noted in *section: 3.2.1.2.3. Haplogroup composition of KwaZulu-Natal*. A genetic structure analysis of the Eastern Cape province would be of interest in the future to further extend the knowledge of indigenous ancestry around the East coast.

3.2.3.2. A network of indigenous haplogroup L0

Figure 3.9 shows a network analysis of haplogroup L0 representing all four of the population groups. Each population group is represented by a distinct colour shown in a key on the left corner of the figure. The portion of each colour in the circle diagrams is related to the frequency observed in that population. This major haplogroup on the Network consisted of 52 haplotypes.

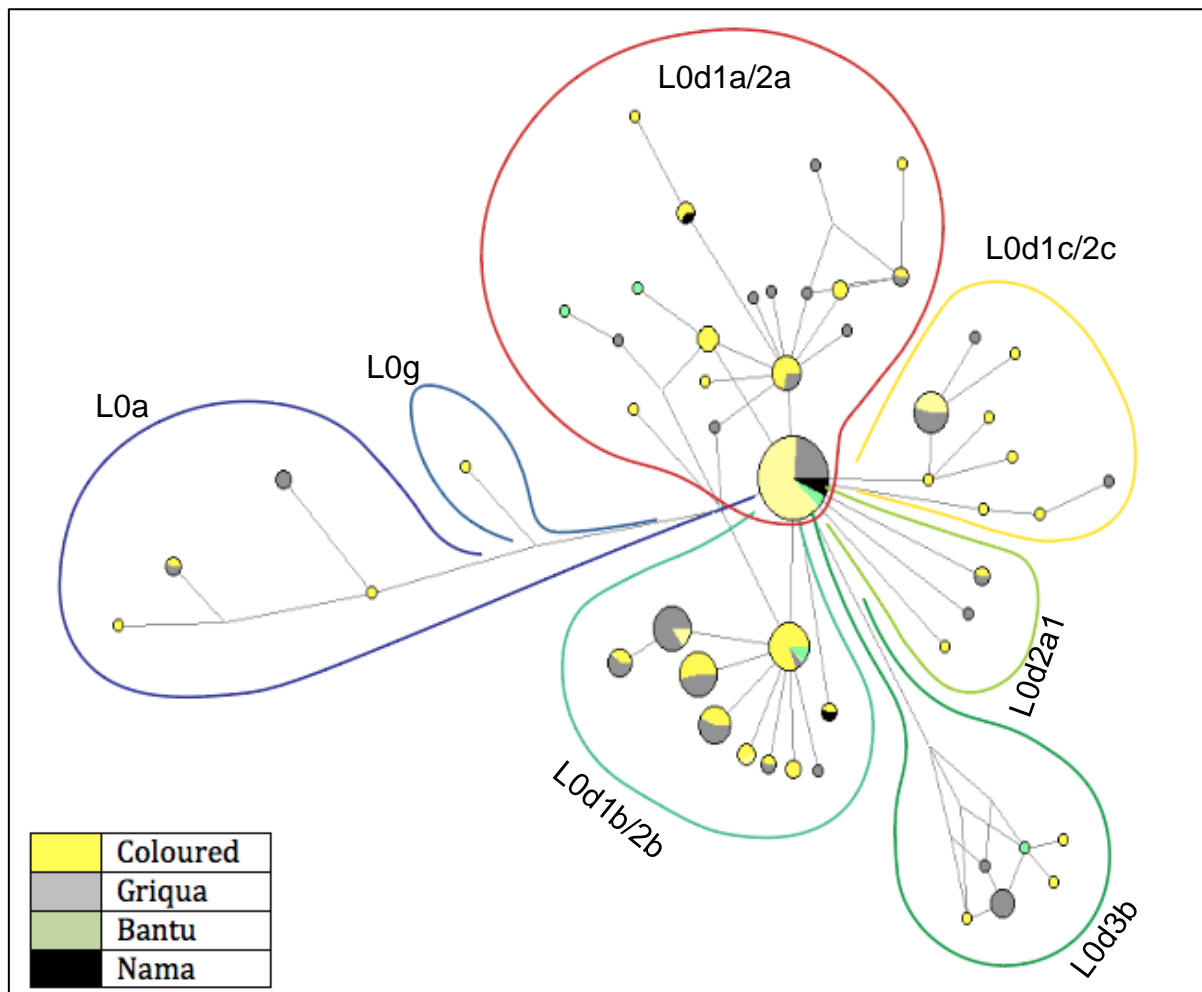


Figure 3.9. Median-joining network analysis of the major L0 maternal haplogroup in the four populations. Keywords on the bottom left shows which colour in the circle diagrams displays which specific population group.

Figure 3.9 illustrates a good unity in haplogroup L0a. This followed by the branching of the L0g genome. The network analysis agrees with Chan, E.K.F et al. (2015) finding in which haplogroup L0g presents an L0a sister branch. Moreover, the figure shows a cluster of mtDNA haplogroup L0d.

Haplogroup L0d1a and L0d2a shows a star phylogeny, which represents a recent and rapid expansion of the mitochondrial DNA type, this is in concordance with Quintana-Murci et al. (2010) that observed a star phylogeny in haplogroup L0d2a. Lastly, the network analysis highlights that several samples share families between the four population groups. Hence, the Coloured, Griqua, Bantu and Nama experienced gene flow.

4.1. CONCLUSION

To conclude, the mitochondrial DNA control region as a single marker is not adequate in forensic applications in South Africa, due to the poor random match probability. Therefore, the study recommends that the mtDNA be used for the purpose of exclusion, or that the mtDNA analysis can be complementary to other markers (such as autosomal and/or Y-chromosome) during a forensic investigation. Moreover, the use of the whole mtDNA as opposed to the control region can yield a higher discrimination capacity as additional fragments or sites are analysed.

The population study shows that distinctive mitochondrial DNA distribution are not just limited to rare or ancient populations. Today, various geographical regions demonstrate strikingly different mitochondrial DNA patterns. These mtDNA patterns can be useful for a haplotype analysis in familial reconstruction and /or identification processes. However, this analysis will be depended on the type of haplotype profile obtained, and the population frequency thereof. This was demonstrated by certain haplogroups and/or haplotypes of mitochondrial DNA sequences that were concentrated or absent within a specific population and further varied geographically. This distinct mitochondrial DNA distribution in South Africa further shows the potential for matrilineal tracking.

The hypothesis proves true, as the study demonstrated that the history of South Africa has influenced the maternal lineage distribution in South Africa and contributed to the admixture observed in the Coloured and Griqua. Therefore, the study demonstrated that mitochondrial DNA sequences were not randomly distributed.

It was, however, remarkable to observe the native Bantu and Nama illustrated only Sub-Saharan African lineages, with the exception of the one Nama individual presenting a Asian lineage. It is noteworthy that the haplogroup, M1a5, observed in the Nama individual remains under debate on whether the haplogroup presents an Asian lineage that represents a backflow to Africa or originated in Africa.

Overall, the Sub-Saharan African lineages demonstrates the most influential gene flow in all the geographical regions and between the Coloured, Griqua, Bantu and Nama

populations. However, this high frequency of Sub-Saharan African lineages implicates the identification of individuals within a forensic context.

4.2. FUTURE WORK

The study focused on the influence that history had on the South Africa population. However, an investigation centered around the genetic clustering that resulted from subtle linguistic, cultural, religion and/or economic forces could further the knowledge on the distribution of mitochondrial DNA sequences.

To provide more accurate estimates in the future, a larger sample collection will have to be conducted for the Bantu and Nama populations. It would be of further interest to obtain the aforementioned populations samples from various region to observe if non-African lineages can be observed in the native populations.

In view of the geographical regions, it would be of high value to investigate the remaining provinces of South Africa. The additional regions will contribute to the map illustrating the distribution of macro-haplogroup L0. While, a larger sample-set of the population group together with the regional study will aid in forensic applications where rare mtDNA haplotypes are questioned in missing persons or mass disaster cases.

REFERENCES

- Álvarez-Iglesias, V; Jaime, JC; Carracedo, A and Salas A. (2007) Coding region mitochondrial DNA SNPs: targeting East Asian and Native American haplogroups. *Forensic Science International: Genetics* **1**: 44–55. DOI: 10.1016/j.fsigen.2006.09.001.
- Amorim, A; Fernandes, T and Taveira (2019). Mitochondrial DNA in human identification: a review *PeerJ* **7**: e7314 DOI: [10.7717/peerj.7314](https://doi.org/10.7717/peerj.7314).
- Anderson, S; Bankier, A; Barrell, B; de Bruijn, H; Coulson, A; *et al.*, (5 co-authors). (1981). Sequence and organization of the human mitochondrial genome. *Nature* **290**: 457-465.
- Andrews, R; Kubacka, I; Chinnery, P; Lightowlers, R; Turnbull, D and Howell, N. (1999) Reanalysis and revision of the Cambridge Reference Sequence for human mitochondrial DNA. *Nature Genetics* **23**: 147.
- Balson, S (2007). Children of the mist: The Lost Tribe of South Africa. Brisbane: Interactive Publications Pty. Ltd.
- Bandelt, H-J and Parson, W (2008) Consistent treatment of length variants in the human mtDNA control region: a reappraisal. *International Journal of Legal Medicine* **122**: 11–21.
- Bandelt, H-J; Forster, P; Sykes, B.C and Richards M.B (1995) Mitochondrial portraits of human populations using median networks. *Genetics* **141**:743-753.
- Bandelt, H-J; Kloss-Brandstätter, A; Richards, M.B Yao, Y-G and Logan, I (2014). The case for the continuing use of the revised Cambridge Reference Sequence (rCRS) and the standardization of notation in human mitochondrial DNA studies. *Journal of Human Genetics* **59**: 66-77 DOI:10.1038/jhg.2013.120.
- Bandelt, H-J; Macaulay, V; & Richards, M. (2006). Human mitochondrial DNA and the evolution of homo sapiens. Berlin, New York: Springer.
- Bär, W; Brinkmann, B; Budowle, B; Carracedo, A; *et al.*, (9 co-authors) (2000) DNA commission of the international society for forensic genetics: Guidelines for mitochondrial DNA typing. *International journal of legal medicine* **113**: 193-196.
- Barbieri, C; Vicente, M; Oliveira, S; Bostoen, K; Rocha, J; Stoneking, M. and Pakendorf, B. (2014). Migration and Interaction in a Contact Zone: mtDNA Variation among Bantu-Speakers in Southern Africa. *PLoS ONE* **9**(6), e99117. DOI:10.1371/journal.pone.0099117.
- Barnard, A. (1992) Hunters and Herders of Southern Africa: A Comparative Ethnography of the Khoisan Peoples. Cambridge University Press, Cambridge, UK.
- Behar, D.M; Villems, R; Soodyall, H; Blue-Smith, J; Pereira, L; *et al.*, (11 co-authors). (2008) The dawn of human matrilineal diversity. *American Journal of Human Genetics* **82**(5): 1130-40. DOI: 10.1016/j.ajhg.2008.04.002.
- Bendall, K and Sykes, B. (1995) Length heteroplasmy in the first hypervariable segment of the human mitochondrial DNA control region. *American Journal of Human Genetics* **57**: 246- 256.
- Besten, M (2006). Transformation and Reconstitution of Khoe-San Identities: AAS Le Fleur I, Griqua Identities and Post-Apartheid Khoe-San Revivalism (1894-2004).

- Brandão, A; Eng, K; Rito, T; Cavadas, B; Bulbeck, D; *et al.*, (12 co-authors). (2016) Quantifying the legacy of the Chinese Neolithic on the maternal genetic heritage of Taiwan and Island Southeast Asia. *Human Genetics* **135**(4): 373-376.
- Brandstätter, A; Parsons, T.J; Niederstätter, H and Parson W. (2003) Rapid screening of mtDNA coding region SNPs for the identification of Caucasian haplogroups. *International Journal of Legal Medicine* **117**:291-298.
- Brown, W. M; George, M Jr; Wilson. (1979) Rapid evolution of animal mitochondrial DNA. *Proceedings of the National Academy of Science of the United States of America* **76**:1967-1971.
- Budowle, B; Allard, M; Wilson, M and Chakraborty, R. (2003) Forensics and mitochondrial DNA: applications, debates, and foundations. *Annual Review of Genomics and Human Genetics* **4**: 119–141.
- Butler, J. (2005) *Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers*. San Diego, CA: Elsevier Academic Press.
- Butler, J. (2012). *Advanced topics in forensic DNA typing*. San Diego, CA: Elsevier Academic Press.
- Cann, R.L; Stoneking, M and Wilson, A.C. (1987) Mitochondrial DNA and human evolution. *Nature* **325**, 31–36. DOI.org/10.1038/325031a0.
- Carracedo, A; Bar, W; Lincoln, P; Mayr, W; Morling, N; *et al.*, (8 co-authors). (2000) DNA Commission of the International Society for Forensic Genetics: guidelines for mitochondrial DNA Typing. *Forensic Science International* **110**: 79-85.
- Chain, B and Heather, J. (2016) The sequence of sequencers: The history of sequencing DNA. *Genomics* **107**: 1-8.
- Chan, E.K.F; Hardie, R-A; Petersen, D.C; Beeson, K; *et al.*, (3 co-authors). (2015). Revised Timeline and Distribution of the Earliest Diverged Human Maternal Lineages in Southern Africa. *PLoS ONE* **10**(3), e0121223. DOI:10.1371/journal.pone.0121223.
- Chan, E.K.F; Timmermann, A; Baldil, B.F; *et al.*, (9 co-authors) (2019). Human origins in a southern African palaeo-wetland and first migrations. *Nature* **575**: 185-189
- Chandrasekar, A; Kumar, S; Sreenath, J; Sarkar, B; Urade, B.P; *et al.*, (16 co-authors). (2009) Updating Phylogeny of Mitochondrial DNA Macrohaplogroup M in India: Dispersal of Modern Human in South Asian Corridor. *PLoS ONE* **4**(10): e7447. DOI:10.1371/journal.pone.0007447.
- Chen, T; He, J; Huang, Y and Zhao, W. (2011) The generation of mitochondrial DNA large-scale deletions in human cells. *Journal of Human Genetics* **56**: 689–694.
- Chen, Y.S; Torroni, A; Excoffier, L; Santachiara-Benerecetti and Wallace, D.C. (1995) Analysis of mtDNA variation in African populations reveals the most ancient of all human continent-specific haplogroups. *American Journal of Human Genetics* **57**(1): 133–149.
- Dubut, V; Murail, P; Pech, N; Thionville, M and Cartault, F. (2009) Inter- and Extra-Indian Admixture and Genetic Diversity in Reunion Island Revealed by Analysis of Mitochondrial DNA. *Annals of human genetics* DOI.org/10.1111/j.1469-1809.2009.00519.x.

- Ehret, C & Posnansky, M (1982) The archaeological and linguistic reconstruction of African history. Berkeley, CA: University of California Press
- Excoffier L. and Lischer H. (2010) Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. *Molecular ecology resources* **10**: 564-567.
- Excoffier L., Smouse P. and Quattro J. (1992) Analysis of Molecular Variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* **131**: 479-491.
- Fendt, L; Huber, G; Röck, A.W; Zimmermann, B; Bodner, M; Delport, R; Schidt, K and Parson, W (2012) Mitochondrial DNA control region data from indigenous Angolan Khoe-San lineages. *Forensic Science International Journal* **6**(5):662-3. DOI: 10.1016/j.fsigen.2012.02.010.
- Franc, L; Carrilho, E and Kist, T. (2002) A review of DNA sequencing techniques. *Quarterly Reviews of Biophysics* **35**: 169-200.
- Fu Y.X (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* **14**: 915–925.
- Fu, Y.X and Li, W.H (1993). Statistical tests of neutrality of mutations. *Genetics* **133** (3): 693-709.
- Gabie, S (2014) Khoisan ancestry and Coloured identity: A study of the Korana Royal House under Chief Josiah Kats. Unpublished Master's thesis. Johannesburg: University of the Witwatersrand.
- Garrison, E & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. arXiv Preprint arxiv. 1207:3907. Q-bio.GN
- Gonder, M.K; Mortensen, H.M; Reed, F.A; de Sousa A and Tishkoff, S.A. (2007) Whole-mtDNA genome sequence analysis of ancient African lineages. *Molecular Biology and Evolution*. **24**(3): 757-68.
- Gonzalez, A.M; Larruga, J.M; Abu-Amero, K.K; Shi, Y; Pestano, J and Cabrera V.M. (2007) Mitochondrial lineage M1 traces an early human backflow to Africa. *BMC Genomics* **8**(1): 223. DOI: 10.1186/1471-2164-8-223.
- Goodwin, S; McPherson, J and McCombie, R. (2016). Coming of age: ten years of next-generation sequencing technology. *Nature Reviews Genetics* **17**: 333-351.
- Goodwin, W; Linacre, A; Hadi, S. (2007). An Introduction to Forensic Genetics. John Wiley & Sons Ltd.
- Güldemann, T (1997) The Kalahari Basin as an Object of Areal Typology: A First Approach, in Schladt, M (Ed.). Language, Identity, and Conceptualization among the Khoisan p. 137-169.
- Hall T. (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic acids symposium series* **41**: 95-98.
- Henn, B.M; Gignoux, C.R; Jobin, M; Granka, J.M; *et al.*, (15 co-authors). (2011) Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. *PNAS* **108**(13) 5154-5162. DOI.org/10.1073/pnas.1017511108.

Heynes, K (2015) A Dual Analysis of the South African Griqua Population using Ancestry Informative Mitochondrial DNA and Discriminatory Short Tandem Repeats on the Y Chromosome. Published Magister Scientiae thesis. Western Cape: University of the Western Cape.

Huber, N; Parson, W; Dür, A. (2018) Next-generation database search algorithm for forensic mitogenome analyses. *Forensic Science International: Genetics* **37**: 204-214. DOI: 10.1016/j.fsigen.2018.09.001.

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

Irwin, J.A; Saunier, J.L; Strouss, K.M; Sturk, K.A; Diegoli, T.M; *et al.*, (4 co-authors). (2007) Development and expansion of high-quality control region databases to improve forensic mtDNA evidence interpretation. *Forensic Science International: Genetics* **1**: 154-157.

Jonas (2011). History of slavery and early colonization in South Africa. South Africa History Online. Available from: <https://www.sahistory.org.za/article/history-slavery-and-early-colonisation-south-africa> (accessed 10.15.18).

Just, R.S; Leney, M.D; Barritt, S.M; Los, C.W; Smith, B.C; Holland, T.D and Parsons, T.J (2009). The Use of Mitochondrial DNA Single Nucleotide Polymorphisms to Assist in the Resolution of Three Challenging Forensic Cases. *Journal of Forensic Sciences* **54(4)**, 887-891. DOI:10.1111/j.1556-4029.2009.01069.x

Kivisild, T; Shen, P; Wall, D.P; Do, B; Sung, R; Davis, K; *et al.*, (11 co-authors). (2006) The Role of Selection in the Evolution of Human Mitochondrial Genomes. *Genetics* **172(1)**: 373-387.

Kivisild, T. (2015) Maternal ancestry and population history from whole mitochondrial genomes. *Investigating Genetics* **6**: 3.

Kloss-Brandstätter, A; Pacher, D; Schönherr, S; Weissensteiner, H; Binna, R; Specht, G. and Kronenberg, F. (2010) HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. *Human mutations* **32**: 25–32.

Köhnemann, S and Pfeiffer, H. (2011) Application of mtDNA SNP analysis in forensic casework. *Forensic Science International: Genetics* **5(3)**: 216-21. DOI: 10.1016/j.fsigen.2010.01.015.

Kwok, P. (2001). Methods for Genotyping Single Nucleotide Polymorphisms. *Annual Review of Genomics: Human Genetics* **2**: 235-258.

Lan, Q; Xie, T; Jin, X; Fang, Y; Mei, S; Yang, G and Zhu, B (2019) MtDNA polymorphism analyses in the Chinese Mongolian group: Efficiency evaluation and further matrilineal genetic structure exploration. *Molecular Genetics and Genomics Medicine* **7**:e934. DOI:10.1002/mgg3.934.

Larkin M; Blackshields G; Brown N; Chenna R; McGettigan P; *et al.*, (8 co-authors). (2007) *ClustalW and ClustalX Version 2*. *Bioinformatics* **23** pp. 2947 – 2948.

Li, H; Handsaker, B; Wysoker, A; Fennell, T; Ruan, J; Homer, N, Marth, G; Abecasis, G; Durbin, R and 100 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25(16)** 2078–2079. DOI:10.1093/bioinformatics/btp352.

- Librado P. and Rozas J. (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**: 1451-1452.
- Liu, L; Li, Y; Li, S; Hu, N; He, Y; Pong, R; Lin, D; Lu, L and Law, M. (2012) Comparison of next-generation sequencing systems. *Journal of Biomedicine and Biotechnology* 251364.
- Lunt, D.H and Hyman, B.C (1997). Animal mitochondrial DNA recombination. *Nature* 387(6630), 247-247. DOI: 10.1038/387247a0
- Luo, S; Valencia, A; Zhang, J and Lee, N (2018) Biparental Inheritance of Mitochondrial DNA in Humans. *Proceedings of the National Academy of Sciences* 115(51): 201810946
- Lutz, S; Wittig, H; Weisser, H.J; *et al.*, (8 co-authors) (2000) Is it possible to differentiate mtDNA by means of HVIII in samples that cannot be distinguished by sequencing the HVI and HVII regions? *Forensic Science International* **113**:97-101.
- Macaulay, V; Hill, C; Achilli, A; Rengo, C; Clarke, D; Meehan, W; Blackburn, J; Semino, O and Scozazari, R. (2005). Single Rapid Coastal Settlement of Asia Revealed by Analysis of Complete Mitochondrial Genomes. *Science* **308**: 1034-1036.
- Maji, S; Krithika, S and Vasulu, T. (2009) Phylogeographic distribution of mitochondrial DNA macrohaplogroup M in India. *Journal of Genetics* 88(1): 127-139.
- Mardis, E. (2008) Next-generation DNA Sequencing Methods. *Annual Review Genomics and Human Genetics* **9**: 387-402.
- McElhoe, J; Holland, M; Makova, K; Su, M; Pual, I; Baler, C; Faith, S and Young, B. (2014) Development and assessment of an optimized next-generation DNA sequencing approach for the mtgenome using the Illumina MiSeq. *Forensic Science International: Genetics* **13**: 20-29.
- Metzker, M. (2010) Sequencing Technologies- The Next-Generation. *Nature Reviews Genetics* **11**: 31-46.
- Mishmar, D; Ruiz-Pesini, E; Golik, P; Macaulay, V; Clark, A.G; *et al.*, (8 co-authors). (2003) Natural selection shaped regional mtDNA variation in humans. *Proceedings of the National Academy of Sciences* 100(1): 171-176.
- Mitchell S.L; Goodloe, R; Brown-gentry, K; Pendergrass, S.A; Murdock, D.G and Crawford, D.C (2015). Characterization of mitochondrial haplogroups in a large population-based sample from the United States. *Human Genetics* **133**(7):861-868. DOI: 10.1007/s00439-014-1421-9.
- Mitchell, P. (2002) *The Archaeology of Southern Africa*. Cambridge: Cambridge University Press.
- Monson, K.L; Miller, K.W.P; Wilson, M.R; DiZinno, J.A and Budowle, B (2002). The mtDNA population database: An integrated software and database resource. *Forensic Science Communications* **4**(2). Available at: <http://www2.fbi.gov/hq/lab/fsc/backissu/april2002/miller1.htm>.
- Murphy, K; Berg, K and Eshleman, J. (2005) Sequencing of Genomic DNA by Combined Amplification and Cycle Sequencing Reaction. *Clinical chemistry* **51**: 35-39.
- Nei, M. (1987). *Molecular Evolutionary Genetics*. Columbia University Press, New York, NY.

- Newman, J.L. (1995) *The Peopling of Africa: A Geographic Interpretation*. Yale University Press.
- Noguera-Santamaría, M; Anderson, C; Uricoechea, D; Durán, C; Briceño-Balcázar, I; Villegas, J (2015) Mitochondrial DNA analysis suggests a Chibchan migration into Colombia. *Universitas Scientiarum* 20(2): 261-278.
- Okonechnikov, K; Golosova, O; Fursov, M; UGENE team (2012) Unipro UGENE: A unified bioinformatics toolkit. *Bioinformatics* 28(8) 1166–1167. DOI: 10.1093/bioinformatics/bts091.
- Oomen, B. (2005). *Chiefs in South Africa: Law, Power & Culture in the Post-Apartheid Era*. New York: Palgrave.
- Pakendorf, B and Stoneking, M (2005) Mitochondrial D.N.A human evolution. *Annual Review of Genomics and Human Genetics* 6:165-183.
- Parson, W and Dür, A (2007) EMPOP-a forensic mtDNA database. *Forensic Science International: Genetics* 1: 88–92.
- Parson, W; Coble, M; Just, R and Irwin, J. (2011) mtGenome reference population databases and the future of forensic mtDNA analysis. *Forensic Science International: Genetics* 5: 222-225.
- Parson, W; Gusmão, L; Hares, D; Irwin, J; *et al.*, (7 co-authors). (2014) DNA Commission of the International Society for Forensic Genetics: Revised and extended guidelines for mitochondrial DNA typing. *Forensic Science International: Genetics* 13: 134-142.
- Perera, S; Ramos, A; Alvarez, L; Jurado, D; Guardiola, M; Lima, M; Aluja, M.P and Santos, C. (2018) Reappraising the human mitochondrial DNA recombination dogma. *BioRxiv* 304535. DOI.org/10.1101/304535.
- Petersen, D; Libiger, O; Tindall, E; Hardie, R; Hannick L; *et al.*, (7 co-authors). (2013) Complex Patterns of Genomic Admixture within Southern Africa. *PLoS Genetics* 9.
- Phillipson, D.W. (2005). *African Archaeology* (Cambridge:Cambridge University Press).
- Quintana-Murci, L; Chaix, R; Wells, R.S *et al.*, (14 co-authors) (2004). Where West Meets East: The Complex mtDNA Landscape of the Southwest and Central Asian Corridor. *The American Journal of Human Genetics* 74:827-845.
- Quintana-Murci, L; Harmant, C; Quach, H; Balanovsky, O; *et al.*, (5 co-authors). (2010) Strong Maternal Khoisan Contribution to the South African Coloured Population: A Case of Gender-Biased Admixture. *The American Journal of Human Genetics* 86: 611-620.
- Quintana-Murci, L; Semino, O; Bandelt, H-J; Passarino, G; McElreavey, K. and Santachiara-Benerecetti, A.S. (1999) Genetic evidence of an early exit of Homo sapiens sapiens from Africa through eastern Africa. *Nature Genetics* 23(4): 437-441.
- R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>.
- Rasmussen, E; Sorensen, E; Eriksen, B; Larsen, H and Morling, N. (2002) Sequencing strategy of mitochondrial HV1 and HV2 DNA with length heteroplasmy. *Forensic Science International* 129: 209-213.

- Richards, M; Macaulay, V; Bandelt, H-J and Sykes, B. (1998) Phylogeography of mitochondrial DNA in western Europe. *Annals of Human Genetics* **62**, 241-260.
- Ristow, P (2017) Comprehensive analyses of the genetic variation between forensic markers in South Africa: Mitogenomes, Autosomal STRs, and Retrotransposons. Published Doctor of Philosophiae thesis. Western Cape: University of the Western Cape.
- Röck, A; Dur, A; van Oven, M and Parson, W. (2013) Concept for estimating mitochondrial DNA haplogroups using a maximum likelihood approach (EMMA). *Forensic Science International Genetics* **7**(6): 601-609.
- Ruiz-Pesini, E; Mishmar, D; Brandon, M; Procaccio, V and Wallace, D.C. (2004) Effects of purifying and adaptive selection on regional variation in human mtDNA. *Science* **303**(5655): 223-6.
- Salas, A; Richards, M; Lareu, M.V; Scozzari, R; Coppa, A; Torroni, A; Macaulay, V and Carracedo A. (2004) The African diaspora: mitochondrial DNA and the Atlantic slave trade, *American Journal of Human Genetics* **74**: 454-465.
- Salas, A; Richards, M; De la Fe, T; Lareu, M.V; Sobrino, B; Sánchez-Diz, P; Macaulay, V and Carracedo, A. (2002). The making of the African mtDNA landscape. *American Journal of Human Genetics*. **71**: 1082-1111.
- Saunders, C and Southey, N (2001). A dictionary of South African History. David Philip publishers Cape Town pp. 81-82
- Saunders, C. (1983). Historical dictionary of South Africa. Metuchen, N.J: The Scarecrow Press.
- Schlebusch, C; de Jongh, M and Soodyall, H. (2011) Different Contributions of Ancient Mitochondrial and Y-Chromosomal Lineages in 'Karretjie People 'of the Great Karoo in South Africa. *Journal of human genetics* **56**: 623-630.
- Schlebusch, C; Lombard, M. and Soodyall H. (2013) MtDNA control region variation affirms diversity and deep sub-structure in populations from southern Africa. *BMC Evolutionary Biology* **13**:56.
- Sigurðardóttir, S; Helgason, A; Gulcher, J.R; Stefansson, K and Donnelly, P. (2000) The mutation rate in the human mtDNA control region. *American Journal of Human Genetics* **66**:1599-1609. DOI: 10.1086/302902.
- Soares, P.; Alshamali, F.; Pereira, J. B; *et al.*, (9 co-authors) (2012) The Expansion of mtDNA Haplogroup L3 within and out of Africa. *Molecular Biology and Evolution* **29** (3): 915–927.
- Statistics South Africa. (2012) *Census 2011 - Census in brief, World Wide Web*. DOI: ISBN 978- 0-621-41388-5.
- Stoneking M. (2008) Human origins. *EMBO reports* **9**: S46–S50.
- Stoneking, M. (2001) Single nucleotide polymorphisms: From the evolutionary past. *Nature* **409**: 821-822.
- Stothard, P. (2000) The Sequence Manipulation Suite: JavaScript programs for analyzing and formatting protein and DNA sequences. *Biotechniques* **28**: 1102-1104.

Sturk, K.A; Coble, M.D; Barritt, S.M; Parsons, T.J and Just, R.S. (2008) The application of mtDNA SNPs to a forensic case. *Forensic Science International: Genetics* 1(1) 295-297. DOI: <https://doi.org/10.1016/j.fsigs.2007.10.148>.

Sun, C; Kong, Q.P; Palanichamy, M.G; Agrawal, S; Bandelt, H-J; *et al.*, (6 co-authors). (2006) The dazzling array of basal branches in the mtDNA macrohaplogroup M from India as inferred from complete genomes. *Molecular Biology Evolution* 23(3): 683-690.

Tajima F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123: 585-595.

Thangaraj, K; Chaubey, G; Singh, V; *et al.*, (4 co-authors). (2006) In situ origin of deep rooting lineages of mitochondrial Macrohaplogroup 'M' in India. *BMC Genomics* 7: 151.

Thompson, L. (2006). A history of South Africa (3rd ed.). Johannesburg: Jonathan Ball.

Torrioni, A; Achilli, A; Macaulay, V; Richards, M and Bandelt, H-J. (2006). Harvesting the fruit of the human mtDNA tree. *Trends in Genetics* 22(6), 339-345. DOI:10.1016/j.tig.2006.04.001.

Torrioni, A; Lott, M.T; Cabell, MF; Chen, Y.S; Lavergne, L and Wallace, D.C (1994i) mtDNA and the origin of Caucasians: identification of ancient Caucasianspecific haplogroups, one of which is prone to a recurrent somatic duplication in the D-loop region. *The American Journal of Human Genetics* 55:760-776.

Torrioni, A; Miller, J.A; Moore, L.G; Zamudio, S; Zhuang, J; Droma, T and Wallace, D.C. (1994ii) Mitochondrial DNA analysis in Tibet: implications for the origin of the Tibetan population and its adaptation to high altitude. *American Journal of Physical Anthropology* 93:189-199

Torrioni, A; Schurr, T.G; Cabell, M.F; Brown, M.D; *et al.*, (5 co-authors). (1993) Asian affinities and continental radiation of the four founding Native American mtDNAs. *American Journal of Human Genetics*. 53: 563-590.

Tully, G. (1999). Mitochondrial DNA: A small but valuable genome. First international conference on forensic human identification. Forensic Science Service.

Tully, L and Levin, B. (2000) Human Mitochondrial Genetics. *Biotechnology and Genetic Engineering Reviews* 17: 147-178.

van Oven M and Kayser M. (2009) Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Human Mutation* 30(2): 386-394.

van Oven, M (2015) PhyloTree Build 17: Growing the human mitochondrial DNA tree. *Forensic Science International: Genetics Supplement Series* 5 e392–e394.

Vanecek, T; Vorel, F and Sip, M (2004) Mitochondrial DNA D-loop hypervariable regions: Czech population data. *International Journal of Legal Medicine* 118(1): 14-18.

Waldman, L., (2007). *The Griqua Conundrum: Political and Socio-cultural Identity in the Northern Cape, South Africa*. Peter Lang.

Wallace D; Brown M; and Lott M. (1999) Mitochondrial DNA Variation in Human Evolution and Disease. *Gene* 238: 211-230.

Wallace, D.C; Ye J.H; Neckelmann S.N; *et al.*, (3 co-authors). (1987) Sequence analysis of cDNAs for the human and bovine ATP synthase beta subunit: Mitochondrial DNA genes sustain seventeen times more mutations. *Current Genetics* **12**: 81-90.

Zimmermann, B; Röck, A; Dür, A and Parson, W. (2014) Improved visibility of character conflicts in quasi-median networks with the EMPOP NETWORK software. *Croatian Medical Journal* **55**: 115-120.

Web Resources

EMPOP, <https://empop.online>

HaploGrep 2, <http://haplogrep.uibk.ac.at/>

Network, Fluxus-engineering.com

PhyloTree, <http://www.phyloTree.org/>

Tableau Software, <https://public.tableau.com/en-us/s/>

Venn Diagram, <http://bioinformatics.psb.ugent.be/webtools/Venn/>

Accession Numbers

Cambridge Reference Sequence, GenBank accession number: M63933.

Revised Cambridge Reference Sequence, NCBI: NC_012920.1.

SUPPLEMENTARY DATA

Table A1: Shared haplotypes amongst the three geographical regions in this study with the number (*n*) of individuals as per ethnic group identified in the respective provinces. The haplotype profiles are in the same order as shown in Table 4.1. The ethnic groups are represented by the Coloured (COL), Griqua (GRI) and Bantu (B).

Haplotype Profile	WC		NC		KZN	
	COL	GRI	COL	B	COL	GRI
16129A 16187T 16189C 16212G 16223T 16230G 16243C 16311C 16390A 16519C 73G 146C 152C 195C 198T 247A 309.1C 315.1C 498DEL 522DEL 523DEL	2	2	9	1	3	3
16129A 16187T 16189C 16212G 16223T 16230G 16243C 16311C 16390A 16519C 73G 146C 152C 195C 198T 247A 315.1C 498DEL 522DEL 523DEL	2	2	0	1	2	0
16129A 16187T 16189C 16223T 16230G 16239T 16243C 16294T 16311C 16519C 73G 146C 152C 195C 247A 309.1C 315.1C 498DEL	2	0	1	0	1	0
16129A 16187T 16189C 16218T 16223T 16239T 16243C 16294T 16311C 16519C 73G 146C 152C 195C 247A 315.1C 498DEL 523DEL 524DEL	0	7	2	0	0	2

Sample_name	Haplogroup	Haplotype
MR14_217*	L0d2c2	16129A 16187T 16189C 16223T 16230G 16243C 16311C 16519C 73G 94A 146C 195C 247A 294A 315.1C 498DEL 523DEL 524DEL
MR14_236*	L0d1a'd	16129A 16187T 16189C 16230G 16234T 16243C 16266A 16311C 16519C 73G 146C 195C 199C 228T 247A 309.1C 315.1C 498DEL
MR14_268*	L0d3b	16187T 16189C 16223T 16230G 16243C 16266T 16274A 16278T 16290T 16300G 16311C 16519C 73G 146C 150T 195C 247A 309.1C 315.1C 316A
MR14_281*	L0d1b2b2	16129A 16187T 16189C 16192T 16223T 16239T 16271C 16294T 16311C 16519C 73G 146C 152C 195C 247A 309.1C 315.1C 498DEL 523DEL 524DEL
MR14_286*	L0d1b2b	16129A 16187T 16189C 16223T 16230G 16239T 16243C 16294T 16311C 16325C 16519C 73G 146C 152C 195C 247A 309.1C 315.1C 498DEL
MR14_293*	L0d2a1	16129A 16187T 16189C 16212G 16223T 16230G 16243C 16311C 16390A 16519C 73G 146C 152C 195C 198T 247A 309.1C 315.1C 498DEL 523DEL 524DEL
MR14_299*	L0d1a'd	16129A 16187T 16189C 16230G 16234T 16243C 16266A 16311C 16519C 73G 146C 188G 195C 199C 247A 309.1C 315.1C 498DEL
MR14_300*	L0d1b2b	16075C 16129A 16187T 16189C 16223T 16230G 16239T 16243C 16294T 16311C 16519C 73G 146C 152C 195C 247A 315.1C 498DEL
MR14_302*	L0d1b2b2	16129A 16140C 16187T 16189C 16223T 16239T 16243C 16294T 16311C 16519C 73G 146C 152C 195C 247A 315.1C 498DEL 523DEL 524DEL
MR14_303*	L0d1a'd	16129A 16148T 16187T 16189C 16209C 16230G 16234T 16243C 16266A 16311C 16519C 73G 146C 153G 195C 199C 247A 315.1C 498DEL
MR14_310*	L0a2	16093C 16148T 16172C 16187T 16188G 16189C 16223T 16230G 16311C 16320T 16519C 64T 93G 95C 152C 189G 236C 247A 263G 309.1C 315.1C 523DEL 524DEL
MR14_311*	L0d2c2	16129A 16187T 16189C 16223T 16230G 16243C 16311C 16519C 73G 94A 146C 195C 247A 294A 315.1C 498DEL 523DEL 524DEL
MR14_314*	L0d1b2b2	16129A 16187T 16189C 16192T 16223T 16239T 16271C 16294T 16311C 16519C 73G 146C 152C 195C 247A 315.1C 498DEL 523DEL 524DEL
MR14_315*	L0d2c2	16037G 16129A 16187T 16189C 16223T 16230G 16243C 16311C 16519C 73G 94A 140T 146C 195C 247A 294A 315.1C 498DEL 523DEL 524DEL
MR14_318*	L0d1a'd	16129A 16187T 16189C 16230G 16234T 16243C 16266A 16311C 16519C 73G 146C 195C 199C 247A 309.1C 315.1C 498DEL
VGRI-054*	L0d1b2b2b1	16129A 16187T 16189C 16218T 16223T 16239T 16243C 16294T 16311C 16519C 73G 146C 152C 195C 247A 315.1C 498DEL 523DEL 524DEL
VGRI-058*	L0d1b2b2b1	16129A 16187T 16189C 16218T 16223T 16239T 16243C 16294T 16311C 16519C 73G 146C 152C 195C 247A 309.1C 315.1C 498DEL 523DEL 524DEL
VGRI-074*	L0d2a1	16129A 16187T 16189C 16212G 16223T 16230G 16243C 16311C 16390A 16519C 73G 146C 152C 195C 198T 247A 309.1C 315.1C 498DEL 523DEL 524DEL
VGRI-075*	L0d2a1	16129A 16187T 16188G 16189C 16212G 16223T 16230G 16243C 16311C 16390A 16519C 73G 146C 152C 195C 198T 247A 309.1C 315.1C 498DEL
VGRI-076*	L0a2a2	16124Y 16148T 16172C 16187T 16188G 16189C 16223T 16230G 16311C 16320T 16519C 64T 93G 150T 152C 189G 204C 207A 236C 247A 263G 315.1C 523DEL 524DEL
VGRI-092*	L0d1b2b2b1	16129A 16187T 16189C 16218T 16223T 16239T 16243C 16294T 16311C 16519C 73G 146C 152C 195C 247A 315.1C 498DEL
VGRI-110*	L0d1b2b	16129A 16187T 16189C 16223T 16230G 16239T 16243C 16294T 16311C 16519C 73G 146C 152C 195C 247A 309.1C 315.1C 498DEL
VGRI-115*	L0d1b2b	16129A 16187T 16189C 16223T 16230G 16239T 16243C 16294T 16311C 16519C 73G 146C 152C 247A 315.1C 498DEL
VGRI-125*	L0d2a1	16129A 16187T 16189C 16212G 16223T 16230G 16243C 16311C 16390A 16519C 73G 146C 152C 195C 198T 247A 309.1C 315.1C 498DEL 523DEL 524DEL
VGRI-138*	L0d2a1	16129A 16187T 16189C 16212G 16223T 16230G 16243C 16311C 16390A 16519C 73G 146C 152C 195C 198T 204Y 247A 291.1A 309.1C 315.1C 498DEL 523DEL 524DEL
VGRI-139*	L0d2a1	16129A 16187T 16189C 16212G 16223T 16230G 16243C 16311C 16390A 16519C 73G 146C 152C 195C 198T 247A 315.1C 498DEL
VGRI-154*	L0d1a'd	16129A 16187T 16189C 16230G 16234T 16243C 16266A 16311C 16519C 73G 146C 188G 195C 199C 247A 309.1C 315.1C 498DEL 523DEL 524DEL
VGRI-166*	L0d2c2	16129A 16187T 16189C 16223T 16230G 16243C 16311C 16519C 73G 94A 146C 195C 247A 294A 315.1C 498DEL 523DEL 524DEL
VGRI_001	L0d3b	16187T 16189C 16214T 16223T 16230G 16243C 16274A 16278T 16290T 16300G 16311C 16519C 73G 146C 150T 195C 247A 315.1C 316A
VGRI_003	L0d1b2b2b1	16129A 16187T 16189C 16218T 16223T 16239T 16243C 16294T 16311C 16519C 73G 146C 152C 195C 247A 315.1C 498DEL 522DEL 523DEL
VGRI_005	L0d2c2	16129A 16187T 16189C 16223T 16230G 16243C 16311C 16519C 73G 94A 146C 195C 247A 294A 315.1C 498DEL 522DEL 523DEL
VGRI_007	L0d1a'd	16129A 16187T 16189C 16230G 16234T 16243C 16266A 16311C 16519C 16524G 73G 146C 195C 199C 247A 315.1C 318C 498DEL
VGRI_008	L0d1b2b2b1	16129A 16187T 16189C 16218T 16223T 16239T 16243C 16294T 16311C 16519C 73G 146C 152C 195C 247A 315.1C 498DEL 522DEL 523DEL
VGRI_010	L0d1c1a1	16093C 16167T 16187T 16189C 16223T 16230G 16234T 16242T 16243C 16311C 73G 146C 152C 195C 198T 247A 315.1C 456T 498DEL 522DEL 523DEL
VGRI_011	L0d1b2b2b1	16129A 16187T 16189C 16218T 16223T 16239T 16243C 16294T 16311C 16519C 73G 146C 152C 195C 247A 315.1C 498DEL 522DEL 523DEL
VGRI_013	L0d2c2	16129A 16187T 16189C 16223T 16230G 16243C 16311C 16519C 73G 94A 146C 195C 247A 294A 315.1C 498DEL 522DEL 523DEL
VGRI_014	L0d1b2b2	16129A 16140C 16187T 16189C 16223T 16239T 16243C 16294T 16311C 16519C 73G 146C 152C 195C 247A 315.1C 498DEL 522DEL 523DEL
VGRI_015	L0d1a'd	16129A 16187T 16189C 16230G 16234T 16243C 16266A 16311C 16519C 73G 146C 195C 199C 247A 309.1C 315.1C 498DEL
VGRI_016	L0d2a1	16129A 16187T 16189C 16212G 16223T 16230G 16243C 16311C 16390A 16519C 73G 146C 152C 195C 198T 247A 315.1C 498DEL 522DEL 523DEL
VGRI_017	L0a1b1	16129A 16148T 16168T 16172C 16187T 16188G 16189C 16223T 16230G 16278T 16293G 16311C 16320T 93G 95C 185A 189G 236C 247A 263G 315.1C 522DEL 523DEL
VGRI_018	L3e1f1	16189C 16223T 16311C 16327T 73G 114A 150T 189G 200G 204C 263G 315.1C
VGRI_019	L0d1b2b2	16129A 16140C 16187T 16189C 16223T 16239T 16243C 16294T 16311C 16519C 73G 146C 152C 195C 247A 315.1C 498DEL 522DEL 523DEL
VGRI_020	L0d1b2b	16129A 16187T 16189C 16223T 16230G 16239T 16243C 16294T 16311C 16519C 73G 146C 152C 247A 315.1C 498DEL

