Kirstie Heynes
3270818

November 2015

A Dual Analysis of the South African Griqua Population using Ancestry Informative Mitochondrial DNA and Discriminatory Short Tandem Repeats on the Y Chromosome.

A thesis submitted in fulfilment of the requirements of Magister Scientiae in the Department of Biotechnology, University of the Western Cape.

Supervisor: Associate Prof. Maria Eugenia D'Amato
Co-supervisor: Prof. Sean Davison
Forensic DNA Laboratory, Department of Biotechnology, University of the Western Cape.

A Dual Analysis of the South African Griqua Population Using Ancestry Informative Mitochondrial DNA and Discriminatory Short Tandem Repeats on the Y Chromosome.
Kirstie Heynes

KEYWORDS (6)

Griqua population

Y chromosome Short Tandem Repeats (Y-STRs)

Discrimination capacity (DC)

Mitochondrial DNA control region (mtDNA)

Single Nucleotide Polymorphisms (SNPs)

Haplogroup

Abstract

The primary objective of this Masters project was to investigate the maternal ancient substructure of the Griqua population in South Africa. Genetic ancestry was determined by investigating ancestry informative single nucleotide polymorphisms. These are located in the control region of the mitochondrial genome. The auxiliary aim was to test the validity of the UWC 10plex system in relation to a sample group of Griqua males. This short tandem repeat multiplex targets specific mutations confined to paternal lineages.

The Khoi Khoi or Hottentots were the first inhabitants in the Cape. Indigenous Khoi Khoi female slaves had offspring with the European settlers in the 1800s which resulted in the Griqua population group. The incorporated European paternal ancestry is what set the Griqua apart from the native population groups at that time. Colonisation events from the mid-17th to 19th Century and the apartheid regime resulted in land dispossession of the native population and an extensively mixed gene pool in South Africa.

One hundred and seventy six (N=176) male and female Griqua people were collectively sampled in Kokstad (2012), Vredendal (2012 and 2013) and at the Griqua National Conference in Ratelgat (2013). All 176 samples were analysed using mtDNA control region Sanger sequencing. The sample group (N=176) was separated based on birthplace (Origin sample group and post-colonial sample group). The origin sample group consists of individuals whose ancestors were not part of the Griqua Trek to Northern regions of South Africa and were less likely to be exposed to colonial influences.

Mutations within the hypervariable segments of the mtDNA control region were used to infer haplogroups with geographic-specific population data. In this way one can plot the extent of ancient Khoisan (L0d) and Bantu influences (L1-L5) as well as the influence of East (M, A, B, E) and West (N, R, J, H) Eurasian haplogroups in the maternal ancestry of the Griqua population group.

The origin sample group showed 91% African ancestry (76.8% L0d) while the post-colonial group had 78% African ancestry (60% L0d). The origin sample group had 2% East Eurasian and 7% West Eurasian ancestry, while the post-colonial group contained 20% Eurasian ancestry. There is greater admixture in the post-colonial group which can be attributed to the integration of surrounding populations during settlement periods in parts of the Northern Cape and KwaZulu-Natal.

The UWC 10plex STR kit was tested to see if it could discriminate between male individuals of this admixed sample group (N=91 males). The markers for this multiplex were selected according to their ability to differentiate between individuals of African descent. It proved to be a viable Y chromosome short tandem repeat testing tool, displaying a statistically significant discrimination capacity value of 0.966 and only having 3 shared haplotypes in the sample group of 91 Griqua males.

November 2015

University of the Western Cape

Private Bag X17, Bellville 7535, South Africa

Email: **3270818@myuwc.ac.za**

FACULTY OF NATURAL SCIENCE

I declare that the MSc thesis titled "A dual analysis of the South African Griqua Population using Ancestry Informative Mitochondrial DNA and Discriminatory Short Tandem Repeats on the Y Chromosome." is my own work.

Name: Kirstie Heynes

Student number: 3270818

I hereby declare that I know what plagiarism entails, namely to use another's work and to present it as my own without attributing the sources in the correct way.

1. I know that plagiarism is a punishable offence because it constitutes theft.
2. I understand the plagiarism policy of the Faculty of Natural Science of the University of the Western Cape.
3. I know what the consequences will be if I plagiarise in any of the assignments for my course.
4. I declare therefore that all work presented by me for every aspect of my course, will be my own, and where I have made use of another's work, I will attribute the source in the correct way.

Signature	Date
--------------------------------	**November 2015**

ACKNOWLEDGEMENTS

This Masters project was a huge building block in my life that required resilience and tested me at times. Firstly, I would like to thank the NRF for the financial help I received over the course of my studies. I was presented with an amazing opportunity by my supervisor to study abroad for two months during the first year of my MSc. I would like to extend my gratitude to the IMBICE lab in Argentina for their generosity while I was there and the lab personnel for welcoming me so warmly.

I am eternally grateful to my supervisor, who nurtured me during my postgraduate years. I learnt some very difficult lessons and have gained what feels like an endless store of practical skills when it comes to Forensic DNA analysis. I continue to learn day to day in my work environment, but the facilities and opportunities afforded to me by UWC broadened my perspective on what was possible and certainly extended my thinking.

The FDL lab personnel provided much needed encouragement and peer counselling during problematic steps during my lab work. We bonded over sampling trips and shared more than just a lab space, the FDL really is an amazing team to be a part of at UWC. This project would not have been possible without the support of the Griqua National Council who openly welcomed us into their annual cultural event. Finally I would like to thank my family, friends and boyfriend who have supported me unconditionally through periods of difficulty during my thesis writing.

One piece of advice that I can offer up to anyone wishing to complete a MSC degree is that it requires patience, resilience, a strong support group and a little bit of luck when it comes to lab work. So work hard, but also work smart.

LIST OF ABBREVIATIONS

3'-OH	-	3 prime Hydroxyl group
A	-	Adenine
ABI	-	Applied Biosystems
APLP	-	Amplified Product Length Polymorphism
BAC	-	Bacterial artificial chromosome
bp	-	Base pair
C	-	Cytosine
CEPH	-	Centre d'Etude du Polymorphisme Humain
d	-	Deletion mutation
dATP	-	Deoxyadenosine triphosphate
DC	-	Discrimination capacity
dCTP	-	Deoxycytodine triphosphate
ddATP	-	Di-deoxyadenosine triphosphate
ddCTP	-	Di-deoxycytodine triphosphate
ddGTP	-	Di-deoxyguanosine triphosphate
ddNTP	-	Di-deoxynucleoside triphosphate
ddTTP	-	Di-deoxythymadine triphosphate
dGTP	-	Deoxyguanosine triphosphate
D-Loop	-	Displacement Loop
DNA	-	Deoxyribonucleic Acid
dsDNA	-	Double stranded DNA
dTTP	-	deoxythymadine triphosphate
DYS	-	Y chromosome DNA STR
EDNAP	-	European DNA Profiling Group
EDTA	-	Ethylene di-amine tetra-acetic acid
EH	-	Extended haplotype
EMPOP	-	EDNAP mtDNA Population Database
G	-	Guanine
GD	-	Gene diversity
GNC	-	Griqua National Conference
GRI_	-	Griqua sample collected in Kokstad
H	-	Heavy strand of mtDNA control region
HD	-	Haplotype diversity
HGDP	-	Human Genome Diversity Panel
HiDi	-	Highly deionized
HRM	-	High Resolution Melt
HV	-	Hypervariable
IDX	-	Human Identification software version 10
ISFG	-	International Society for Forensic Genetics
Kya	-	Thousand years ago
L	-	Light strand of mtDNA control region
LD	-	Linkage disequilibrium

LMS	-	London Missionary Society
MH	-	Minimal Haplotype
N	-	Number of individuals/samples
NaCl	-	Sodium Chloride
NGS	-	Next Generation Sequencing
NRY	-	Non-recombining segment of the Y chromosome
O_H	-	Origin of replication
PAR	-	Pseudo autosomal regions
PCR	-	Polymerase chain reaction
RCRS	-	Revised Cambridge reference sequence
RFLP	-	Restriction Fragment Length Polymorphism
rfu	-	Relative fluorescent units
RM Y-STR	-	Rapidly mutating Y short tandem repeats
RSRS	-	Reconstructed Sapiens Reference Sequence
SAPS	-	South African Police Services
SDS	-	Sodium dodecyl sulphate
SNP	-	Single Nucleotide Polymorphism
ssDNA	-	Single stranded DNA
STR	-	Short Tandem Repeat
STRbase	-	Short Tandem Repeat database
T	-	Thymidine
TBE	-	Tris-Borate-EDTA
TE	-	Tris-EDTA
Tm	-	Melting temperature
UV	-	Ultra violet
V_	-	Griqua samples collected in Vredendal
VGRI_	-	Griqua samples collected at the 2013 GNC
Y-STR	-	Y chromosome short tandem repeat

Table of contents

Chapter 1: Introduction

Section 1: Review of literature

Chapter 2: Methods and materials

Chapter 3: Results and Discussion

List of figures:

List of tables:

Chapter 1, Section 1: Review of literature

1.1. Population background

The Griqua can be categorized as pastoralists of Khoekhoe and mixed descent. They were initially known as Bastaards or Basters to the colonialists who forcefully removed them from their grazing lands in the Western Cape (Saunders, 2001; Balson, 2007). The term “bastaards” was commonly used in the 18th century for the offspring of mixed unions between whites and people of colour such as the Khoekhoe.

Most of the Griqua population moved out of the Cape under the leadership of Adam Kok I (1710 – 1795) in the late 18th century. They trekked to a region north of the Orange River uninhabited by the colonialists. Here, they set up a thriving community which was unsettled by the London Missionary Society (LMS). The LMS made their first attempts to evangelize the Kok clan in 1801 (Balson, 2007). They encouraged the group to move further north to Klaarwater (Griquatown) where the Griqua’s joined a greater autochthonous population of Bastaards, Korana, San and other Khoe-San descendants (Besten, 2006). The Griqua identity was diffused by extraneous cultural and heritage influences from neighbouring tribes.

On a visit to Klaarwater in 1813, Reverend John Campbell indicated that the term “bastaard” was offensive as an ethnic identity and the people as a whole resolved to be called Griqua’s. The re-adoption of the Griqua name in 1813 resulted in the onset of a stronger Griqua heritage, renaming of the land (Griquatown) and a move towards traditionalism by the Griqua people under leadership of the Kok family (Schoeman, 1997).

Adam Kok I’s successor, Adam Kok II (1771 - 1835) was not as pliable as the missionaries had hoped in their evangelizing attempts in the newly formed Griquatown. This resulted in Andries Waterboer, a leader and army chief, being elected by the missionaries and colonial government to lead the people of Griquatown into a new era of westernized subservience. The election of Waterboer and his strict rule in turn generated animosity against him by traditionalists who preferred a pastoral and raiding subsistence strategy (Philip, 1828).

Under leadership of Waterboer, Griquatown grew to 1600 inhabitants by the year 1823. This population consisted not only of the new Griqua, but Khoekhoe, San, Korana, Sotho and Tswana heritage. These multiple ethnic heritages allowed for diverse socio-cultural and linguistic exchanges. Many who were not Griqua would have been incorporated first as Griqua-dependents and later as full members due to the sanctuary provided from cattle-raiders and colonialists (Philip, 1828). Contemporary Griqua's of Griquatown still recognise the Waterboer clan as their true chief, while Griqua families from other parts of South Africa follow the descendent line of Adam Kok I.

Cornelius Kok I (1740 – 1822) along with his sons Cornelius Kok II and Adam Kok II (1771 – 1835) moved to Campbell with the Griqua's after Adam's exit as the chief of Griquatown in December 1820 (Schoeman, 1997). This group did not settle for an extended period in Campbell, they moved to Philippolis in 1825 and later Kokstad (1872) under the leadership of Adam Kok II and his sons Abraham and Adam Kok III (1811 – 1875). Adam Kok III took up the role of chief in the Natal province (now KwaZulu-Natal) midland town of Kokstad for three years. The colonialists took advantage of the leaderless Kokstad after Adam Kok III's death on the 30th of December 1875. There was no recognisable successor and strife between Kok's heirs led to an ungovernable town by 1878 (Balson, 2007).

As a result Andrew Abraham Stockenström le Fleur I (1867 – 1941) positioned himself with a measure of success as heir of Griqua Chief Adam Kok III. By 1920 he established a strong Griqua identity amongst the people of Kokstad and Campbell as well as newly established Griqua settlement areas along the East Coast of South Africa on Farmland in Kranshoek near Knysna. He initiated an annual meeting called the Griqua National Conference (GNC) which consisted of representatives from various groups such as Griquas, Namaquas, Outeniquas, Hottentots and mixed Coloured people to form a Griqua 'tribe'. The GNC run by Le Fleur and his descendants played a prominent role within the Griqua socio-political landscape prior to 1994 (Besten, 2006).

One would suspect a diluted cultural identity in present day Griqua's along the inland and East coast areas of South Africa, due to exposure of various tribes during the Griqua trek around South Africa. This trek led the Griqua people through Griquatown (1800), Campbell (1805), Philippolis (1823), Kokstad (1870) with a final Griqua settlement in Kranshoek (1930). To this day Kokstad and Kranshoek remain areas of Griqua concentration. Descendants of Waterboer have populated parts of Griquatown and the Kok's and Le Fleur's are scattered throughout the Northern Cape in Campbell, Philippolis, Kimberly and Koffiefontein as well as areas of the Western Cape in Cape Town, Vredendal and Kranshoek.

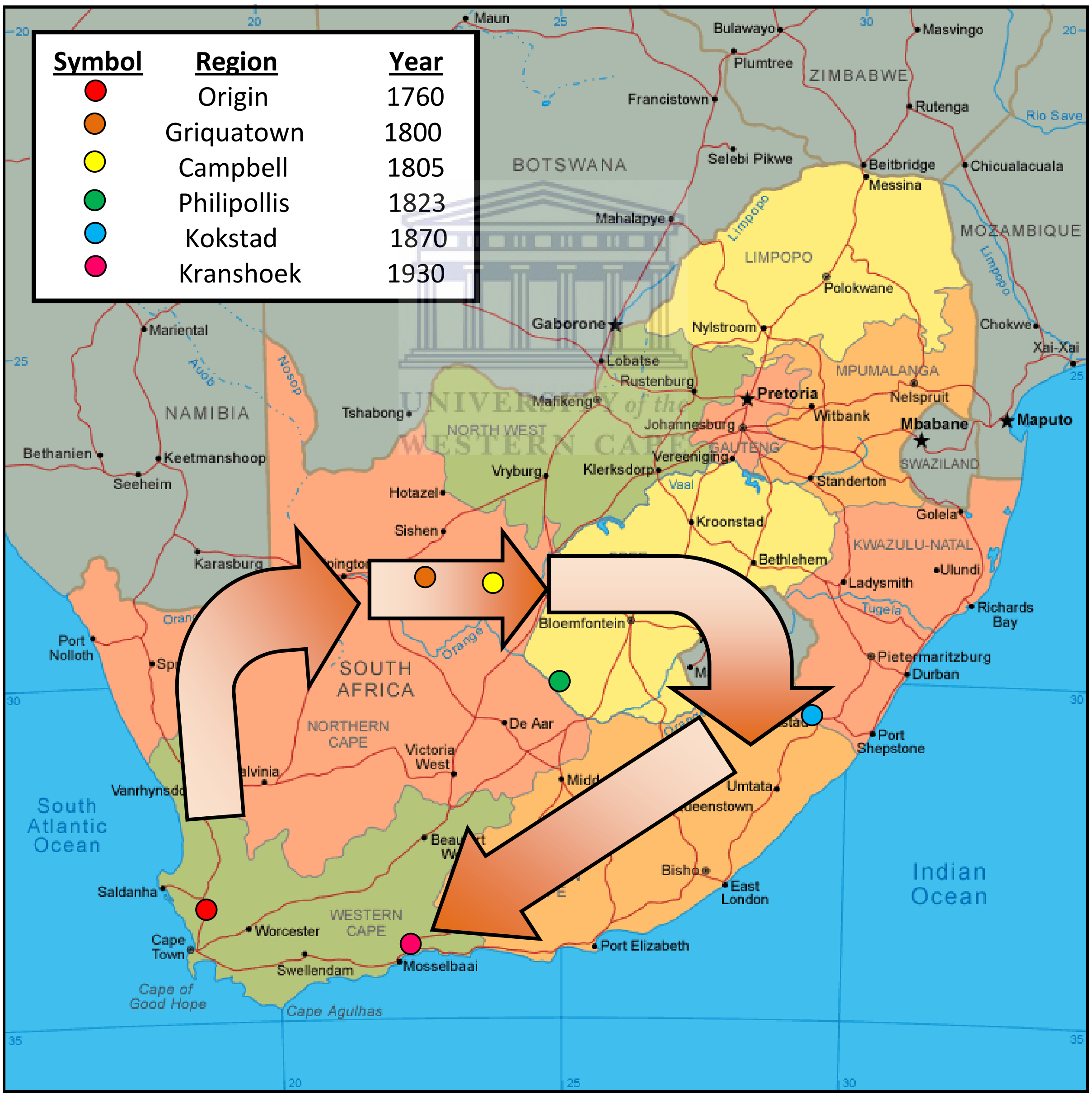


Figure 1.1 Map showing the relocation of the Griqua population throughout South Africa from the late 18th to the 20th Century with their final settlement in Kranshoek, Western Cape under Le Fleur leadership.

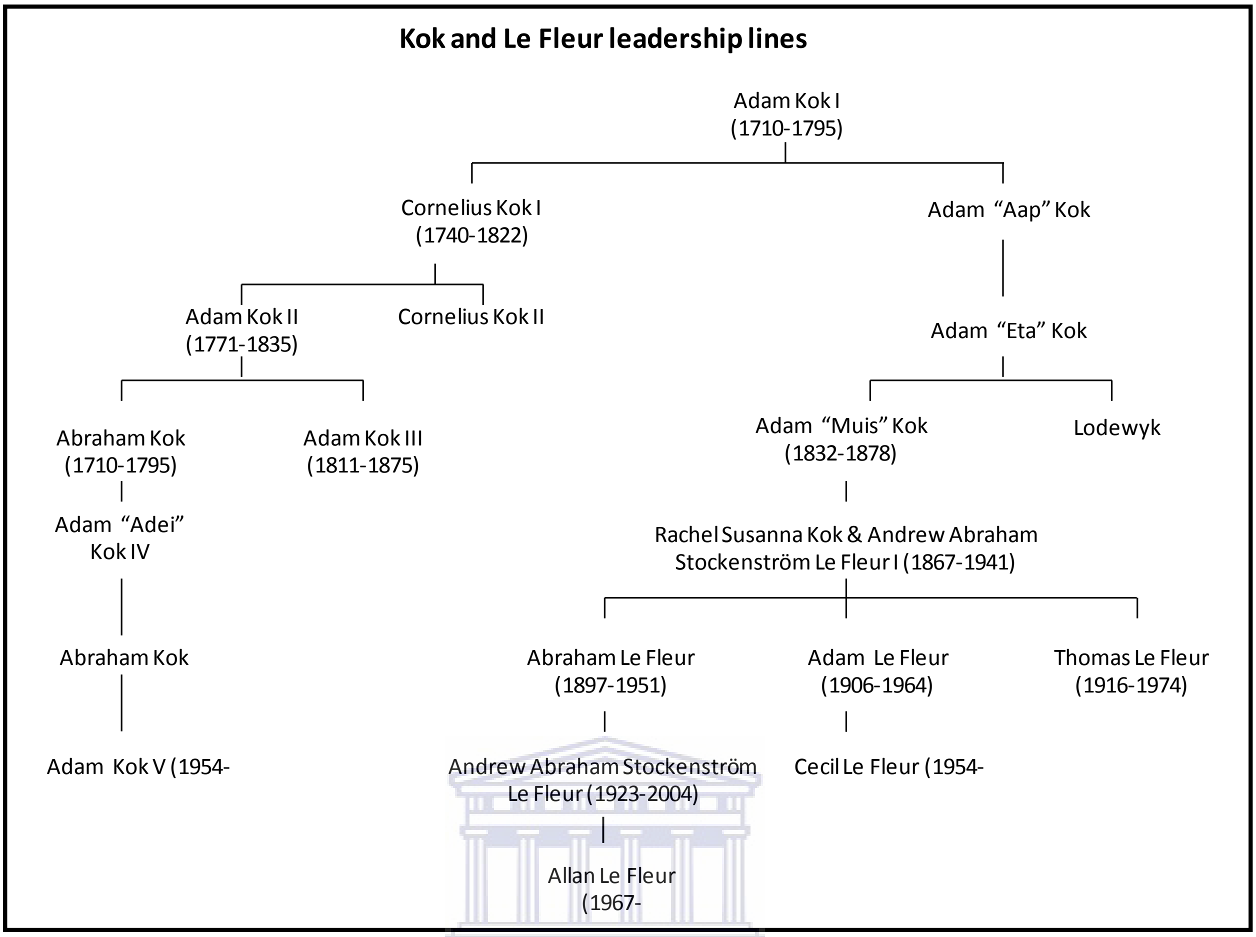

Kok and Le Fleur leadership lines
Adam Kok I (1710-1795)
Cornelius Kok I (1740-1822)
Adam "Aap" Kok
Adam Kok II (1771-1835)
Cornelius Kok II
Adam "Eta" Kok
Abraham Kok (1710-1795)
Adam Kok III (1811-1875)
Adam "Muis" Kok (1832-1878)
Lodewyk
Adam "Adei" Kok IV
Rachel Susanna Kok & Andrew Abraham Stockenström Le Fleur I (1867-1941)
Abraham Kok
Abraham Le Fleur (1897-1951)
Adam Le Fleur (1906-1964)
Thomas Le Fleur (1916-1974)
Adam Kok V (1954-
Andrew Abraham Stockenström Le Fleur (1923-2004)
Cecil Le Fleur (1954-
Allan Le Fleur (1967-

Figure 1.2 Family tree displaying the Griqua leadership lines starting with Adam Kok I in the 18th century.

1.2. Genetic data

1.2.1. Sources of genetic data

Both length and sequence genetic variation exists in human populations. Length variation, in the form of short tandem repeat (STR) markers, has become the primary means of forensic DNA profiling over the past decade (Butler and Coble 2007). These forms of variation enable forensic DNA testing because many different alleles can exist in the non-coding regions of a genome. STR profiles are generated based on a core set of STR markers which can be found at various sites on a chromosome. Information from several of these linked genetic markers can be combined to achieve statistically significant information on a male individual's DNA profile (Butler and Coble 2007). This technique, in the form of the UWC 10plex Y-STR kit (D'Amato *et al.* 2011), is used to discriminate between males of the Griqua sample group.

1.2.2. Defining genetic data: Haplotypes vs. haplogroups

Chromosomes contain a variety of highly polymorphic short tandem repeats. Several have been typed on uni-parental markers such as the Y chromosome. Because of their high mutation rate they can be used for human identification studies and discriminate between individuals from the same population group. The allelic states of Y-STR markers are confined as haplotypes within lineages. When these lineages are themselves confined within populations as a consequence of their histories, then population diversity can be very low (Jobling *et al.* 1997).

The haplotype frequency values for a multiplex Y-STR assay should reflect the distribution of male lineages within a population much more realistically than single Y-STR frequency data (Kayser *et al*. 1996). Because many Y-STRs can be genotyped in multiplex assays, it was originally thought that typing appropriate groups of STRs could represent a cost-effective method for assigning haplogroups to populations (Schlecht *et al.* 2008). However finding the correct set of STR markers which can be interchangeable with ancestry informative SNPs has proven to be quite a challenge (Wang *et al.* 2013).

An individual's ancestry can be traced back to a geographical location in relation to a group of single nucleotide polymorphic mutations. These single nucleotide polymorphisms are arranged into haplogroups on a common phylogenetic tree (Ravid-Amir *et al.* 2009). Maternal and paternal haplogroups can be assigned according to mtDNA and Y chromosome SNP data respectively (Goedbloed *et al.* 2009). These mutations on the mtDNA genome of an individual relate back to a geographical location but also provide a genetic timestamp as to when their ancestors diverged from the common mtDNA ancestor as the SNPs have a very slow mutation rate.

1.2.3. Importance of genetic data

Polymorphisms found in uni-parentally inherited chromosomes such as the Y chromosome and mitochondrial DNA have proven invaluable in the fields of anthropology and forensic casework. Each type of polymorphism has a specific use either at an individual or population level depending on the field being studied (Goodwin *et al.* 2007).

It has become highly significant for individuals in power (Kings or chiefs) with indigenous ancestry to undergo genetic testing to prove linkage to their forefathers. This is due to South Africa's political turmoil and land relocation policies (Natives Land Act: 27 of 1913) over the past century. This allows indigenous people to claim the land of their ancestors back and allow for equitable land distribution, in accordance with the Restitution of Land Rights Act: 22 of 1994. The land belonged to indigenous people prior to the colonisation events of the 18th century.

Griqua individuals could possibly have indigenous (Khoi or San) maternal ancestry. For this project, Griqua individuals have been tested using ancestry informative mtDNA control region markers. It has become apparent that aiding members of indigenous or admixed populations within South Africa is of remarkable importance in terms of the individual's cultural classification.

Returning the results of an ancestry study proves to be a pinnacle moment in the self-identification process for some individuals who may have been lacklustre in terms of their cultural identity. It can be said that mtDNA control region SNP testing shows notable anthropological value in the cultural structure of admixed population groups within South Africa.

Genetic data can also have forensic applications when a specific combination of STR alleles on a single Y chromosome defines a Y-STR haplotype. This means that the Y chromosome is analysed as a forensic marker and is a desirable tool in the fight against crime. For the purpose of this project the Y chromosome DNA of Griqua males from throughout South Africa was analysed using a STR kit (UWC 10plex). The robustness of this kit was tested with the most important parameter tested being the ability to tell individuals apart at the allelic level.

This property of the kit is especially important for possible future implementation as an identification tool in mixture profiles from a crime sample. By developing robust DNA based human identity testing systems, the field of biotechnology can assist in the identification of perpetrators of violent sexual crimes such as rape (Leat *et al.* 2004). The frequency of sexual violence in South Africa is unacceptably high with a number of factors leading to its high incidence. Male DNA evidence can be differentiated from female victim DNA collected at crime scenes by targeting the Y chromosome. A universal typing method can be utilised in order to make comparisons with results from different laboratories (Hammer and Redd 2006ii). Several areas in the Western Cape are riddled with drug-related crimes and a large proportion of the population falls within a low socio-economic bracket which coincides with gang activity. Some of the worst affected areas of violent and senseless drug-related crimes are Mitchell's Plein and Khayelitsha. There have been 62 649 reported sexual assault cases in South Africa from April to March 2013/2014 with the Western Cape having a total of 8062 reported cases (http://www.saps.gov.za/resource_centre/publications/statistics/crimestats/2014/crime).

1.3. Mitochondrial DNA (mtDNA)

The human mitochondrial DNA (mtDNA) is a closed double stranded circular molecule that is 16 569 base pairs (bp) in length. It is located within the mitochondrion in the cytoplasm of the cell. The mtDNA is strictly maternally inherited as it is transmitted from one generation to the next in the oocyte cytoplasm. Sperm mitochondria enter the egg during fertilization but they appear to be lost in early embryogenesis (Wallace *et al.* 1999). The mitogenome contains a high level of population-specific mtDNA polymorphisms (Wallace *et al.* 1999).

The mtDNA has two strands, a guanine-rich heavy (H) strand and a cytosine rich light (L) strand. The only non-coding segment of the mtDNA is the control region, or displacement loop (D-loop) a region of 1121bp that contains the origin of replication of the H-strand (O_H) and the H and L-strand transcriptional promoters.

1.3.1 Mitochondrial DNA control region

The D-loop or control region is the most variable region in the mitochondrial genome, with the most polymorphic nucleotide sites being concentrated in three ‘hypervariable’ segments HV-I, -II and -III in that order. Population specific variations have been found through sequencing and restriction fragment length polymorphism (RFLP) experiments on the hypervariable segments of the D-loop (Stoneking 2000). Initially the large majority of mtDNA sequence data published was limited to HVI (Avise *et al.* 1989; Wakeley 1993; Finilla 2000). But sequencing of the entire control region has now become more common practice (Lutz *et al.* 1997; Bini *et al.* 2003).

Figure 1.3 displays the circular structure of the mtDNA genome with the clearly labelled D-loop. This region provides a gold mine of ancestry informative polymorphic mutations that are conveniently located in regions of high variability. Mutations can be traced back several thousands of years to the geographical location of origin for an individual's ancestors (Schlebusch *et al.* 2011).

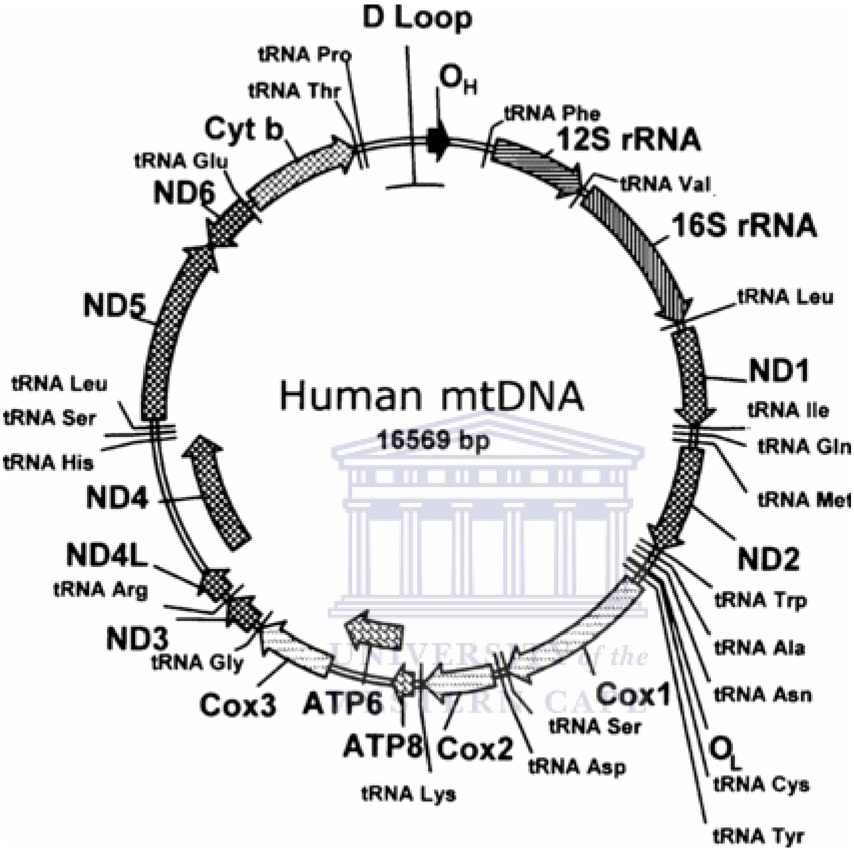


Figure 1.3 Circular structure of human Mitochondrial DNA (mtDNA) with the highly polymorphic D-loop which contains the hypervariable segments I, II and III (Wallace *et al.* 1999).

1.3.2. Mitochondrial DNA coding region

The coding region of the mitochondrion (bases 577-16023) contains highly conserved sequences which are important for the overall metabolic energy production of the cell. The average mutation rate in the 1.1 kb non-coding control region (bases 16024-576) is about 10 times higher than that of the 15.5 kb coding region (Howell *et al.* 2007; Pakendorf and Stoneking 2005). This can be explained by purifying selection filtering out the often deleterious functional mutations from the population.

Due to the higher overall mutation rate, the control region is relatively enriched in sequence variation, and therefore researchers typically sequence part of this region (HVI and II) for preliminary haplogroup assignment. Most haplogroups cannot be confidently assigned based on control region data only *e.g.* H4, V (van Oven and Kayser 2009). Therefore it is relevant to confirm predicted haplogroups based on control region sequences by use of informative SNPs from the coding region.

A common pattern of mutations places an individual in a haplogroup. More ancient haplogroups are found in the Khoisan of Southern Africa (L0d and L0a) and population groups of East Africa (L0d3). These groups branched off early in human history and have remained relatively genetically isolated since then. Haplogroups L1, L2 and L3 are largely confined to Africa in the Bantu speaking clans (Macaulay *et al.* 2005). There was an expansion ±63 000 years ago out of the African haplogroup L3 (±84 000 years old) which resulted in Macrohaplogroups N (Asian ancestry) and M (European ancestry) displayed in Figure 1.4 (Macaulay *et al.* 2005).

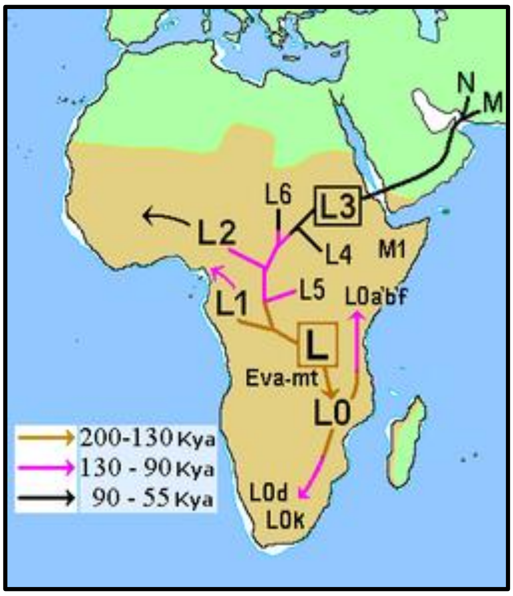


Figure 1.4 Map showing the early diversification of modern humans beginning with mitochondrial eve and moving through Africa (Hg L0; L1-6) and then Europe (MacroHg M and N) (Macaulay *et al.* 2005).

Initial mtDNA haplogroup assignment should be done using coding region information either solely or combined with control region data. The High Resolution Melt (HRM) system described in section 3.4.5 provides a highly efficient tool for targeting individual SNP sites in the coding region while the Next Generation Sequencing (NGS) platform (section 3.4.2) is better for sequencing of the whole mitochondrial genome.

1.3.3. Single Nucleotide Polymorphisms (SNPs)

Single nucleotide polymorphisms (SNPs) occur as a result of errors during DNA replication in meiosis and their frequency varies throughout the regions of the genome (Goodwin *et al.* 2007). Each SNP usually has two possible allele states, for example a person may have a point mutation occurring at position 16129 in the control region of the mtDNA that may result in either a guanine (G) or an adenine (A) base being present. Haplogroups can be assigned once data on several SNPs has been collected from a sequencing reaction. There are databases with extensive knowledge regarding the geographic origins of Y chromosome SNPs (Y-SNPs) (ISOGG Tree version 10.61) and mtDNA SNPs (Phylotree Tree build 16) based on studies of global populations. Because of the high geographic specificity of these SNPs, their haplogroups can be used to directly measure admixture among diverse populations (Hammer *et el.* 2006i). Knowing the origins of specific haplogroups is important when trying to characterize populations that have been influenced by colonial and settlement events, such as in the Griqua population in South Africa.

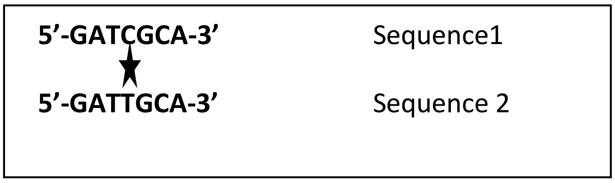


Figure 1.5 A Single Nucleotide Polymorphism (SNP). Two variant alleles are shown at one position, indicated by the star: the fourth position in sequence 1 is a cytosine while in sequence 2 it is a thymine. In most cases, the mutation event which creates a SNP at a specific locus is unique and only two different allele states (biallelic) are normally found.

SNPs exhibit a low rate of mutation and are therefore exclusively used for classification into related sets of lineages known as haplogroups. Performing the large number of SNP genotyping tests needed to properly infer haplogroup status is however expensive and time-consuming (Schlecht *et al.* 2008). Although SNPs have proven to be efficient in ancestry studies, several significant disadvantages exist with SNP markers when considered for use as a forensic marker. Since SNPs are not as polymorphic as STRs, more SNPs are required to reach equivalent powers of discrimination (Butler and Coble 2007). It has been suggested that 40-60 SNPs are required to reach the same power of discrimination as 13-15 STR loci (Butler and Coble 2007). There is also an inability to decipher mixtures when using SNPs which can be a huge hindrance in the successful completion of casework. It is therefore most commonly used to trace ancestry in anthropological studies.

1.3.4. Common mtDNA typing techniques

SNPs are the preferred marker used for maternal haplogroup assignment because of their slow mutation rate in comparison to STRs (Goodwin *et al.* 2007). This means that they show ancient changes (up to 170kya) that have occurred in the mtDNA genome. There are several SNP assay techniques available to suit the aims of any project. If the SNP site is known, an allele-specific hybridization technique such as SNaPshot or HRM can be used. If an entire region needs to be analyzed for SNPs, a sequencing reaction using fluorescent labelling can be used such as Sanger sequencing whereas if an entire genome needs to be sequenced, Next Generation Sequencing is the most efficient option (Kwok, 2001).

1.3.4.1. Sanger sequencing

This classical chain-termination method conceived in 1977 consisted of four separate sequencing reactions needed to synthesise a copy of the desired sequence. The reaction consisted of a single-stranded DNA template, primer, DNA polymerase and one of four dideoxynucleotides (ddATP, ddGTP, ddCTP, or ddTTP) is incorporated into each reaction along with three complimentary deoxynucleotidetriphosphates (dATP, dGTP, dCTP, dTTP). The modified di-deoxynucleotidetriphosphates (ddNTPs) terminate DNA strand elongation.

These chain-terminating nucleotides lack a 3'-OH group required for the formation of a phosphodiester bond between two nucleotides, causing DNA polymerase to cease extension of DNA when a modified ddNTP is incorporated. The ddNTPs may be radioactively or fluorescently labelled for detection in automated sequencing machines rather than the X-ray or gel methods initially utilised.

Dye-terminator sequencing utilises labelling of the chain terminator ddNTPs, which permits sequencing in a single reaction, rather than four reactions as in the labelled-primer method. In dye-terminator sequencing, each of the four dideoxynucleotide chain terminators is labelled with fluorescent dyes, each of which emits light at different wavelengths. Owing to its greater efficiency, dye-terminator sequencing is now the predominant technique utilised in automated sequencing. Its limitations include dye effects due to differences in the incorporation of the dye-labelled chain terminators into the DNA fragment, resulting in unequal peak heights and shapes in the electronic DNA sequence chromatogram after capillary electrophoresis. This problem has been addressed with the use of modified DNA polymerase enzyme systems and dyes that minimize incorporation variability, as well as methods for eliminating "dye blobs". The dye-terminator sequencing method, along with automated high-throughput DNA sequence analyzers, is now being used for the vast majority of sequencing projects.

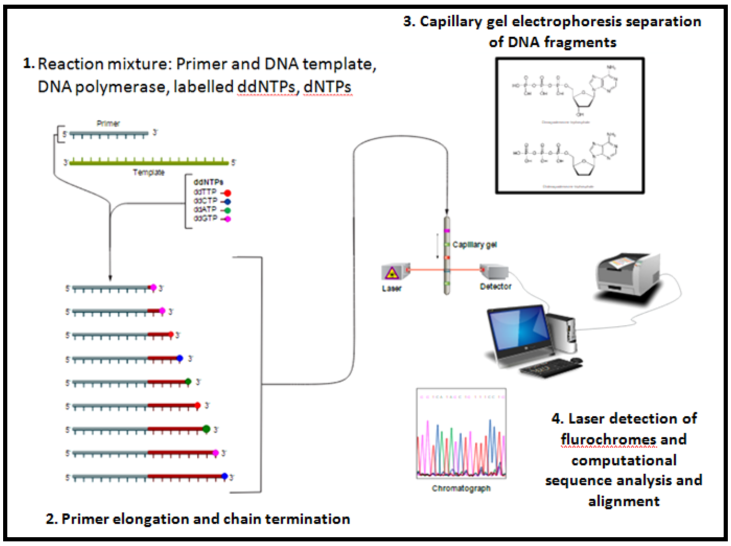

3. Capillary gel electrophoresis separation of DNA fragments
1. Reaction mixture: Primer and DNA template, DNA polymerase, labelled ddNTPs, dNTPs
Primer
5'
3'
3'
5'
Template
ddNTPs
ddTTP
ddCTP
ddATP
ddGTP
Capillary gel
Laser
Detector
Chromatograph
4. Laser detection of flurochromes and computational sequence analysis and alignment
2. Primer elongation and chain termination

Figure 1.6 Diagram displaying the Sanger (chain termination) method for DNA sequencing utilising four simple steps (Murphy *et al.* 2005).

1.3.4.2. Next Generation Sequencing (NGS)

Traditionally there was a reliance on large bacterial artificial chromosomes (BACs) to investigate the whole genome of an organism by incorporating 100kb fragments of a genome of interest and “growing” it up in a genomic library. Results were manually inspected prior to the sequencing reaction and a consensus sequence of several megabases was built. This method proved highly time-consuming and required a large amount of scientific expertise. There has been a great demand commercially for a revolutionary technology that delivers fast, inexpensive and accurate genome information (Metzker, 2010; Davey *et al.* 2011).

Next Generation sequencing provides an inexpensive method for production of large volumes of sequencing data using various platforms such as Roche/454 pyrosequencing, Illumina genome analyzer and Applied Biosystems SOLiD™ sequencer. The Roche/454 pyrosequencing instrument was first to achieve commercial production (2004) and has proved to be a reliable sequencing tool for non-model species, ancient DNA typing, metagenomic sequencing and single base resolution that is necessary for SNP typing (Metzker, 2010). It is based on a simple adhesion technique whereby the library fragments adhere to agarose beads. These beads then adhere to individual 454-specific adapters. Each of these fragment:bead complexes is isolated into individual oil/water pockets which contain PCR reactants; resulting in millions of copies of each of these fragments being present on the surface of each bead. These amplified single molecules are then sequenced en masse via a stepwise approach with an imaging step after incorporation of each nucleotide. Each of these platforms uses a similar adhesion method with various control techniques to ensure optimal sequencing results (Mardis, 2008).

There are various troubleshooting guides in place for the Sanger sequencing method since it has been in commercial use for nearly 50 years. Recently developed statistical methods have been used on Next Generation data to ensure quality control between the various platforms and in multiple sequencing runs. These statistical methods are used to improve and quantify the considerable uncertainty associated with genotype calling when using this technology (Nielson *et al.* 2011).

1.3.4.3. SNaPshot®

The SNaPshot® Multiplex Kit uses a single-tube reaction to interrogate SNPs at known locations. The chemistry is based on the dideoxy single-base extension of an unlabelled oligonucleotide primer (or primers). Each primer binds to a complementary template in the presence of fluorescently labelled dideoxynucleotide triphosphates (ddNTPs) and DNA polymerase. The polymerase extends the primer by one nucleotide, adding a single ddNTP to its 3´ end. The fluorescence colour readout reports which base was added (Life Technologies). Although it has primarily been utilised for marker detection, ideally the SNaPshot® assay represents an efficient means for locating coding region SNPs within the mtDNA coding region and to confirm haplogroup assignment (Salas *et al.* 2005).

The use of an assay such as SNaPshot in cancer research, genetic disease research on the human genome and mtDNA studies has proven successful (Quintanas *et al.* 2004).

1.3.4.4. Amplified Product Length Polymorphisms (APLP)

An experimental approach that targets ancestry informative SNPs in the mtDNA coding region by utilising two alternate allele specific (ancestral/derived) forward primers (ancestral/derived) with a shared reverse primer. The principle of this method is based on an attachment of a non-complementary sequence to the 5'-end of one of two allele specific primers. The result is two size-different amplification products that can be resolved using an acrylamide gel. Primers can be designed by isolating a SNP in the mitochondrial control region found of the desired haplogroup in silico, using the comprehensive mtDNA SNP database Phylotree (http://www.phylotree.org/).

A total of six polymorphic sites from the mtDNA genome of East Asian populations was analysed using a multiplex APLP technique in a 2001 study by Umetsu *et al.* These sites were important in the assignment of individuals into East Asian haplogroups. Five of them are SNPs of the coding regions (nt3010, nt4386, nt5178, nt8794, and nt10398) and the other, a 9-bp repeat variation in the noncoding cytochrome oxidase II/tRNALys intergenic region (9bp).

It proved to be a reliable tool with assigning individuals of this specific sample group into haplogroups. The mtDNA variations of 2471 individuals from 20 populations of Japanese, Korean, Chinese, and German were examined and classified into 18 haplotypes. Two of these haplotypes, B1 (estimated ancestral haplotype) and C1, were distributed among all populations tested. However, the haplotypes A1, A2, B2, B3, and C2 were mostly restricted to the Mongoloid populations, whereas haplotypes B5 and C5 appeared almost exclusively in the German population (Umetsu *et al.* 2001). This experimental haplogroup assignment technique using various combinations of alleles between six polymorphic sites is a huge stepping stone for efficient ancestral studies. A variety of APLP multiplexes can be tailor designed for various sample groups based on their expected heritage. For example, a multiplex separating L0 sub clades can be designed to initially type individuals from the Griqua population group.

1.3.4.5. High Resolution Melting (HRM)

The high resolution melt technique provides experimental return for minimal effort by using an inexpensive closed tube method which is homogenous, accurate and rapid (Reed *et al.* 2007; Vossen *et al.* 2009; Wittwer 2009). DNA denaturation or melting has been used for many years to study DNA structure and composition. HRM is a multi-purpose technology whereby a melt profile is generated based on the amount of saturated dye incorporated into the double stranded (ds) amplified DNA products in a reaction over a temperature gradient. As the temperature increases, the dsDNA is denatured into ssDNA and the amount of dye incorporated decreases (Reed *et al.* 2007).

The quantity of incorporated dye is measured over the temperature gradient to produce a melt curve. The PCR products amplified are designed to be of different lengths so as to produce a melt curve specific to each variant. This allows an individual to be SNP typed into either the ancestral of derived state based on the HRM melt curve produced (Reed *et al.* 2007).

Sensitivity and specificity when detecting variants in a DNA sequence were significantly improved through development of saturating DNA dyes and instrumentation to measure melting behaviour. Similarities in melt curves from different heterozygotes cannot be differentiated at the individual level and this poses a problem. Although this technique poses an efficient method for SNP typing in comparison to gel electrophoresis, several commercial labs are still relying on the tried and trusted Sanger sequencing technique (Wittwer, 2009). Figure 1.7 displays a typical melt curve when using a HRM Y-SNP multiplex system.

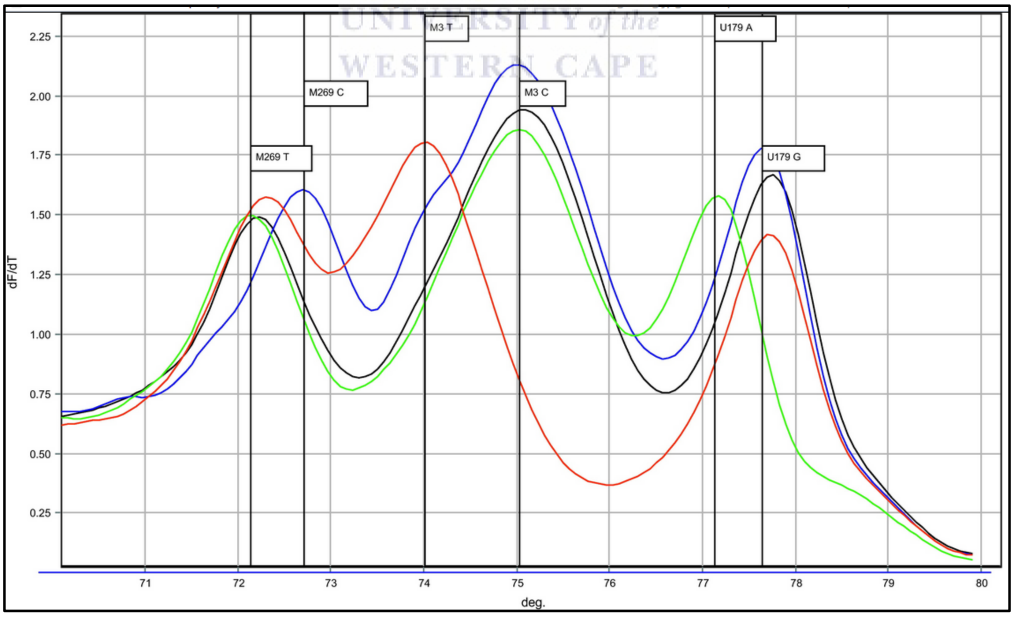


Figure 1.7 The first derivative melt curve when using an HRM Y-SNP triplex for haplogroup assignment. The samples in red belong to haplogroup Q1a3a, in blue to haplogroup R1b1b2 and in green to haplogroup I. Samples in black do not belong to either of the three haplogroups included in analysis (Zuccarelli *et al.* 2011).

1.3.5. Classification of individuals into mitochondrial haplogroups

The first entire human mtDNA light strand to be sequenced was described by Anderson *et al.* (1981) and this sequence is used as a reference standard in the assignment of haplogroups to individuals. mtDNA Sanger sequencing results are aligned with a revised version of this reference sequence, Revised Cambridge Reference Sequence (RCRS). When differences between the individual's sequence and the RCRS are observed, the numbered site and nucleotide differing from the reference standard are recorded. For example, at site 16519 (in the mtDNA control region), the Anderson sequence has a T; however, a large portion of the population carries a C at site 16519. Such an individual's mtDNA sequence is described as 16519C. If no other bases (or sites) are noted, then it is understood that the particular mtDNA sequence is identical to the Anderson sequence, except as noted at site 16519 (van Oven *et al*. 2009)

Not all SNP assignments are this simple and one needs to take into account ambiguities and insertion/deletion polymorphisms (INDELs) as well as homopolymeric tracts which are found in HVII and III of the control region. If an unresolved ambiguity is observed at any site, the base number for the site is listed followed by an 'N' (*e.g.*, 16125N). Insertions are described by first noting the site immediately 5' to the insertion followed by a decimal point and a '1' (for the first insertion), a '2' (if there is a second insertion), and so on, and then by the nucleotide that is inserted. In the case of homopolymeric tracts, where the exact position at which the insertion has occurred is unknown, the assumption is always made that the insertion has occurred at the highest numbered end of the homopolymeric region. For example, a homopolymeric region, at which insertions are common, occurs between nucleotide positions 311 and 315 (inclusive). The polymorphism, a C insertion, is assumed to occur after site 315, so the nomenclature used is 315.1C. Deletions are recorded by listing the missing site followed by a 'd' (*i.e.*, 220d). At confirmed heteroplasmic sites, IUPAC codes for base calling can be applied. A list of IUPAC codes for DNA and protein assignment is accessible at http://www.bioinformatics.org/sms/iupac.html. If heteroplasmy is suspected but not confirmed, an N can be used. For example, an A/G heteroplasmy can either be designated as N or as R.

Once a combination of SNPs has been assigned to an individual, they can be placed into a haplogroup that relates to the geographical location of their ancestors. The first mtDNA haplogroups, discovered in Native Americans, were called A, B, C, and D (Torroni *et al*., 1993). Subsequently detected haplogroups were designated using other letters of the alphabet. The L haplogroups represent the most deep-rooting lineages and are African specific indicating the African origin of modern humans as well as the out-of-Africa theory, based on other genetic as well as fossil data. Haplogroup L3 gave rise to Macrohaplogroups M, N and R (the latter itself a subclade of N), which encompass all variation observed outside Africa. Nomenclature evolved in such a way that letters C, D, E, G, Q, and Z designate lineages belonging to M; letters A, I, S, W, X, and Y lineages within N; and B, F, HV, H, J, K, P, T, U, and V lineages within R (van Oven *et al.* 2009).

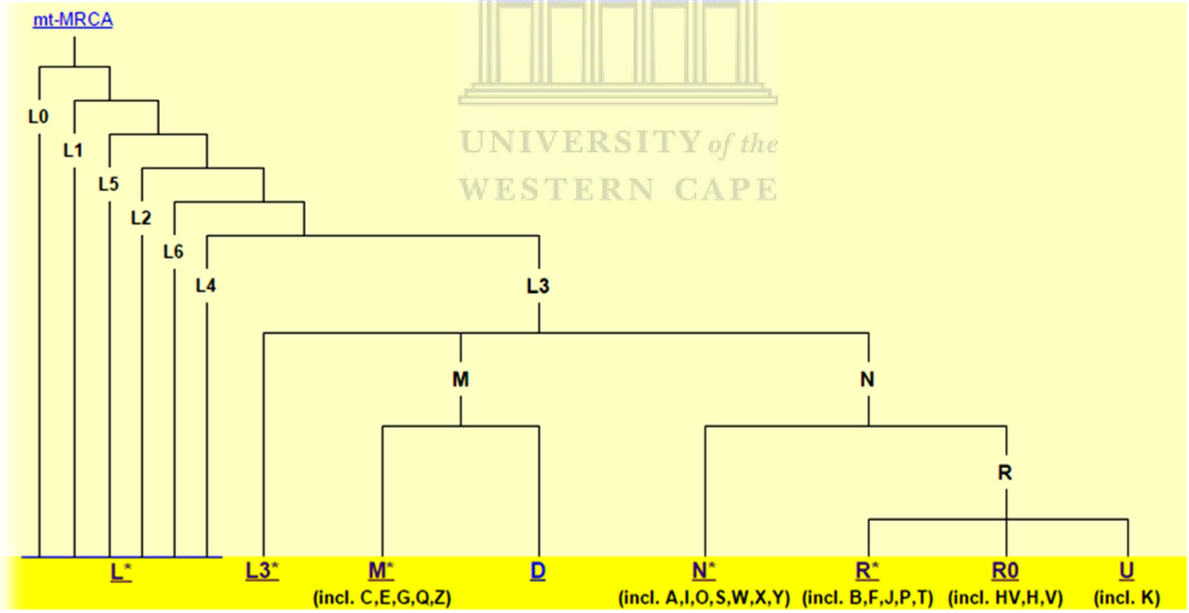


Figure 1.8 mtDNA phylogenetic tree build 16 which is divided into eight subtrees accessible through links on http://www.phylotree.org/tree/main.htm. The mitochondrial most recent common ancestor (mt-MRCA) is used to root the tree. Accumulation of polymorphisms over time separates individuals into various haplogroups.

1.3.6. Proposed nomenclature adjustments

The typing of mtDNA control region mutations into distinct haplogroups according to the rCRS has recently been disputed by Behar *et al.* (2012). This is because nomenclature refers to the recently coalescing European haplogroup H2a2a1. The use of the rCRS as a reference results in a number of practical problems such as the misidentification of ancestral versus derived states; and the count of non-synonymous mutations that map the path between the rCRS and the query sequences. Behar *et al.* (2012) have proposed a “Copernican” reassessment of the human mtDNA phylogeny by switching to a Reconstructed Sapiens Reference Sequence (RSRS) as the phylogenetically valid reference point. The RSRS is a mtDNA reference sequence that uses both global sampling of modern human samples and samples from ancient hominids. It is based on the likely modal haplotype of the common ancestor to both modern humans and such ancient groups as the Neanderthals. It shows an unbiased path back from any one modern mtDNA sequence to our distant common maternal ancestor.

The previously suggested root was also updated to incorporate the available mitogenomes from H. neanderthalensis (Behar *et al.,* 2012). In principle, ancient mtDNA from early human fossils might be informative but unreachable because of technical problems inherent to the analysis process. However, as the split between H. sapiens and H. neanderthalensis predates the appearance of the RSRS, a resolution of the deepest node might be achieved by rooting the human phylogeny with H. neanderthalensis complete mtDNA sequences. The sub-Saharan haplogroup L0 and African L1’2’3’4’5’6 are separated from each other by 14 coding and four control-region mutations. The human mtDNA root is allocated at this node.

The structure shown in figure 1.9 is explicitly for bifurcations that occurred 40,000 years before present (YBP) or earlier, and a graph showing the explosion of haplogroups since then. The y axis indicates the approximate number of haplogroups from each time layer that have survived to nowadays. The upper and lower x axes of the rooted tree are scaled according to the number of accumulated mutations since the RSRS and the corresponding coalescence ages, respectively.

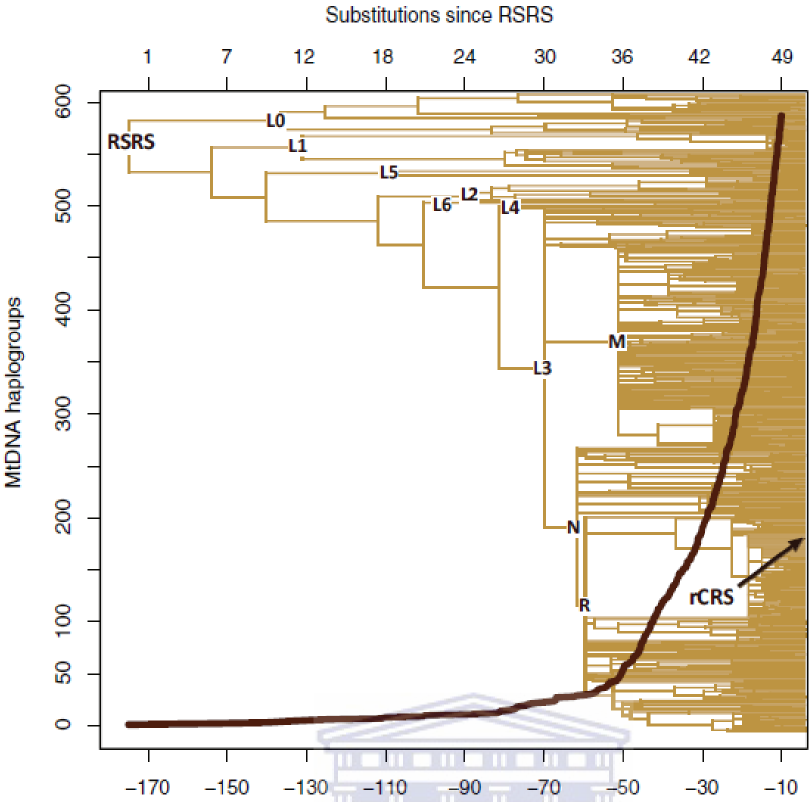


Figure 1.9 A schematic representation of the most parsimonious human mtDNA phylogeny as inferred from 18,843 complete mtDNA sequences by Behar *et al.* (2012).

The principal change, as suggested by Behar *et al.* (2012), means that an ancestral sequence (RSRS) serves as the epicentre of the human mtDNA reference system rather than a modern mitogenome from Europe (rCRS). The use of the rCRS for haplogroup assignment in this Masters project is due to current phylogenetic programs (Haplogrep; Kloss-Brandstaetter *et al.* 2010) and databases (Phylotree; van Oven & Kayser 2009) not being amended. Tens of thousands of additional complete mtDNA sequences are expected to be generated in accordance with the RSRS over the next few years.

1.3.7. Expected haplogroups and structure for the Griqua population

Based on historical records it can be hypothesised that the majority of the Griqua individuals from this study will fall into African ancestral haplogroups (L0 -5). However some admixture is expected, the degree of which is unknown. It is a common phenomenon that colonised areas show greater gender-biased gene flow. Previous studies on similar South African populations by Schlebusch (*et al.* 2011) and Quintana-Murci (*et al.* 2010) have shown an incorporation of European (macrohaplogroup N) and Asian (macrohaplogroup M) maternal haplogroups.

A strong maternal Khoisan contribution to the South African coloured population was found in a 2010 study by Quintana-Murci *et al.* The origins of the South African coloured population can be traced historically to the shores of Cape Town in the mid-17th century. The Dutch East India Company established a refreshment station here and the colonisers were encouraged to dwell amongst the natives or Khoisan. This resulted in a strong colonial (out of Africa haplogroups) paternal contribution in the South African coloured population.

The maternal ancestry of the South African coloured population is dominated by Khoisan haplogroups L0d with the second most prominent influence (L0a) also having African origins (Quintana-Murci *et al.* 2010). It is hypothesised that the presence of lineages of pan-African origins: L1 – L5 were introduced during the recent Bantu expansion. This is either due to direct gene flow from Bantu peoples or indirectly via prior admixture with the Khoisan.

Schlebusch's 2011 study on a population group consisting exclusively of Khoe and/or San ancestry, the Karretjie population of South Africa, showed that the neighbouring coloured populations had a more heterogeneous origin of mtDNA haplogroups. This provides an example of how geographically closely situated sample groups can have varying colonial influences present in their maternal ancestry. There were varying levels of Eurasian and Bantu-admixture in the coloured populations.

A study by Petersen *et al.* (2013) confirmed Schelbusch's findings that Khoisan assimilation with European settlement at the most southern tip of Africa resulted in significant ancestral Khoisan contributions to the Coloured (n = 25) and Baster (n = 30) populations. The latter populations were further impacted by 170 years of East Indian slave trade and intra-continental migrations resulting in a complex pattern of genetic variation (admixture). The populations of southern Africa provide a unique opportunity to investigate the genomic variability from some of the oldest human lineages to the implications of complex admixture patterns including ancient and recently diverged human lineages.

Barbieri *et al.* (2013) investigated the deepest roots of maternal ancestry, L0d and L0k, it was hypothesised that the early divergence of these lineages led to the ancient substructures found in Africa. Figure 1.10 displays a contour map of the distribution of these ancient lineages within population groups from South Africa, Botswana and Namibia.

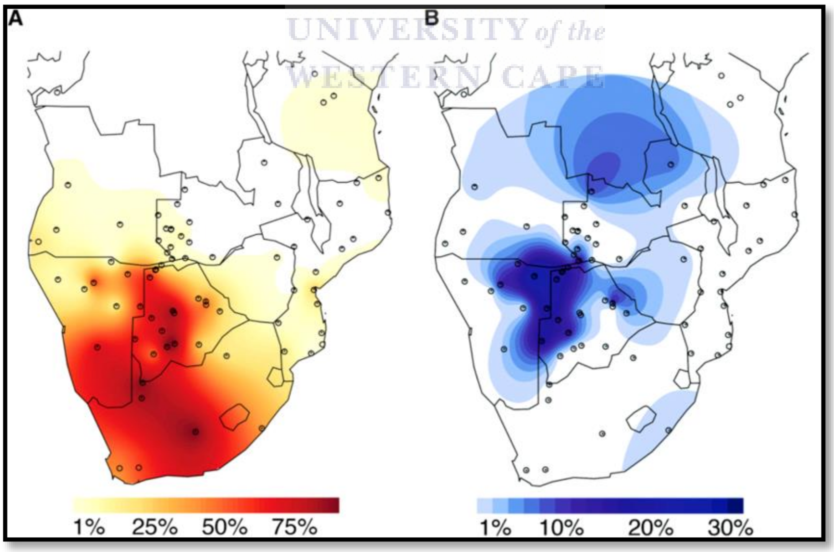


Figure 1.10 Surfer map displaying the spatial distribution of haplogroup L0d (a) and L0k (b) in populations from South Africa, Namibia and Botswana (Barbiera *et al.* 2013).

High levels of genetic differentiation within Khoisan populations was found in a follow up study by Barbieri *et al.* (2014) on 700 individuals from 26 population groups in South Africa. This genetic differentiation can be attributed to drifting and the multilocal residence pattern that is often seen in Khoisan populations. There is genetic evidence of contact between Khoisan populations over vast geographic distances in Central Kalahari which was not previously hypothesised. Genetic drift cannot always be accurately predicted for a population group. Various external factors can lead to vast admixture, even in native population groups that are geographically isolated.

A study by Chan *et al.* (2015) profiled the early mitochondrial lineages found in the southern African click-speaking forager peoples or Khoisan. Data generated by analysing mitochondrial genomes of 182 individuals from southern Africa was used to refine basal phylogenetic divergence, coalescence times and Khoisan prehistory. Results confirm L0d as the earliest diverged lineage (172 kya, 95%CI: 149–199 kya), followed by L0k (159 kya, 95%CI: 136 – 183 kya) and a new lineage classified as L0g (94kya, 95%CI: 72 – 116 kya).

Two new L0d1 subclades were classified L0d1d and L0d1c4/L0d1e, and L0d2 and L0d1 divergence was estimated at 93 kya (95%CI: 76 – 112 kya). Results concur with Morris *et al.* (2014) that the earliest emerging L0d1’2 sub-lineage L0d1b (49 kya, 95%CI: 37 – 58 kya) is widely distributed across southern Africa. Concomitantly, it has been found that the most recent sub-lineage L0d2a (17 kya, 95%CI: 10 – 27 kya) is equally common.

Chan *et al.* (2015) found that while the L0d1c and L0k1a lineages are restricted to contemporary inland Khoisan populations, the observed predominance of L0d2a and L0d1a in non-Khoisan populations suggests a once independent coastal Khoisan prehistory. The distribution of early-diverged human maternal lineages within contemporary southern Africans suggests a rich history of human existence prior to any archaeological evidence of migration into the region (Morris *et al.* 2014). This provides genetic-based evidence for significant modern human evolution in southern Africa at the time of the Last Glacial Maximum at between21–17 kya, coinciding with the emergence of the major lineages L0d1a, L0d2b, L0d2d and L0d2a.

An indirect incorporation of Bantu and Pan-African lineages (L1-5) is to be highly expected in the Griqua population found along inland and East Coast regions of Griqua trek through South Africa. The Griqua population samples collected from these areas can be referred to as the post-colonial sample group. While samples found along isolated regions of the West Coast of South Africa can be referred to as the origin sample group. It can be said that this origin group represents the ancestral state, since their forefathers were not part of the Griqua trek and colonisation events that affected the post-colonial sample group.

1.4. Y chromosome

Y chromosome DNA testing is important for a number of fields of study including: Human identification, paternity testing and genealogical studies (Decker *et al.* 2007; Butler and Coble 2007). The Y-chromosome is one of the smallest chromosomes in the human genome with an approximate size of ~60 mega bases (Mb) including approximately 24 Mb of euchromatin and 30 Mb of heterochromatin. At the tip of both the long (q) and short (p) arms are pseudo-autosomal regions (PARs) which are homologous to X-chromosome sequences and as such are responsible for correct pairing between the sex chromosomes during meiosis. Between these PARs are large interspersed tandemly repeated arrays which have been well characterized (Kayser *et al.* 1996; Jobling *et al.* 1997).

Duplications and deletions are known to occur on the Y chromosome and can be observed in both the father and the son since the sequences of the non-pseudo autosomal proportion of the Y-chromosome do not recombine (Kayser *et al.* 1996; Decker *et al.* 2008; van Oven *et al.* 2011). STR loci are spread throughout the genome in the 22 autosomes and the X and Y chromosomes. Over the past decade Y chromosome testing has grown in popularity with different Y-STR markers selected for various uses and for marker availability. The high variability of STRs makes them useful for forensic and paternity testing and are especially useful for inferring affinities among closely related populations (Schlecht *et al.* 2008). Analysis of tandem repeats on the Y chromosome is of particular significance in violent sexual crime cases which are prevalent in South Africa.

1.4.1. Short Tandem Repeats (STRs)

Microsatellites or Short Tandem Repeats are repetitive stretches of DNA made of short sequence motifs (2 – 6 bp) repeating 5 – 100 times. STRs are very common in eukaryotic genomes with the major part of the human Y chromosome consisting of polymorphic sequences. These sequences are organised into large interspersed tandemly repeated arrays which are mutated via a stepwise mutation mechanism (Kayser *et al.* 1996).

Mutation rates of STRs are expected to be small, about 1 mutation per 1000 generations per locus (Bacolla *et al.* 2008; Goedbloed *et al*. 2009). This results in allelic polymorphisms which are specific to individuals. It is an efficient tool for identification in forensics and paternity tests as well as for studying demographic history and population structure (Ravid-Amir *et al.* 2009).

Candidate Y-STR loci for forensic casework must meet a range of requirements and tetranucleotides (4bp repeats) usually serve as viable candidates because they are less likely to incur a high level of artefactual data. Forensic genetics relies heavily on the analysis of a number of loci simultaneously – *i.e*. multiplex amplification (Kimpton *et. al* 1993; Richards *et. al* 1993). There are a range of STRs starting with simple repeats which contain units of identical length and sequence. Compound repeats comprise of two or more adjacent simple repeats and lastly complex repeats which contain several repeat blocks of variable unit length as well as variable intervening sequences (Butler and Hill, 2012).

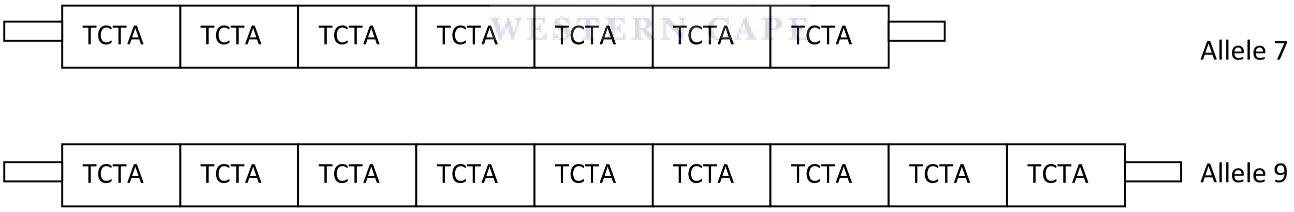


Figure 1.11 The structure of a simple short tandem repeat. The core repeat can range between 2 - 7 bp. This example shows a tetranucleotide repeat (4bp). The DNA on either side of the core repeats is called flanking DNA. The alleles are named according to the number of repeats that they contain.

Several commercial multiplex kits are available which contain sets of markers that together confer a greater discrimination capacity, or ability to discern between individuals in a sample group. It is determined by dividing the number of different haplotypes seen in a given population by the total number of samples in that population. The first set of universal markers that was established for use in forensic casework was the Minimal Haplotype (MH) which consisted of the markers: DYS391, DYS389I, DYS439, DYS389II, DYS438, DYS393, DYS385a/b, DYS19, DYS392 and DYS390 (Hammer and Redd 2006ii). After a few years there was a demand for a greater discrimination capacity to be established because individuals could not be easily discerned from one another. A further 6 markers were added to increase the discrimination capacity of these 11 markers. This was then termed the Extended Haplotype (EH) which consisted of the MH and the six additional markers: DYS456, DYS635, DYS437, DYS458, Y-GATA-H4 and DYS448.

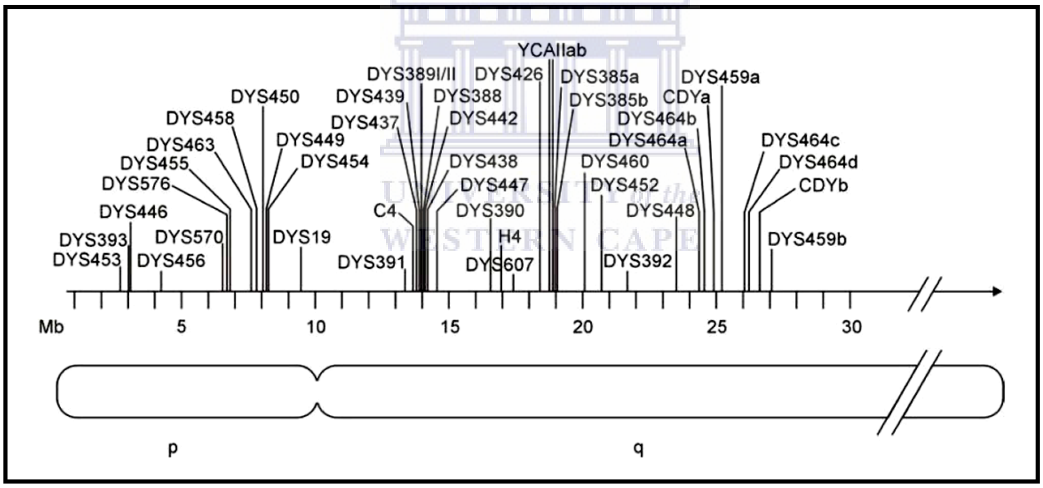


Figure 1.12 Y-chromosome map showing the locations of 43 Y-STR loci Typed in Michael Hammer's lab at the University of Arizona (Hammer and Redd 2006ii). The markers for the MH and EH are clearly allocated within the NRY.

One limitation of Y-STRs in forensic and paternity applications is the lack of independence of these markers on the NRY because they do not undergo recombination with the X chromosome during meiosis (Hammer and Redd 2006ii). This results in difficulty with distinguishing paternally related males in a population; which makes it increasingly challenging in highly inbred populations such as the Afrikaner population in South Africa.

One could speculate that if Y-STR kits with substantially higher mutation rates (as discussed in section 1.4.1.6.2) than the majority of the markers in the commercial kits, it may become possible to differentiate between male relatives at the individual level. This advancement would solve the current dilemma of Y chromosome applications in forensics and truly set this genetic marker apart from the rest (Ballantyne *et al.* 2010). Recent work by Ballantyne *et al.* (2012, 2014) and Purps *et al.* (2014) has led to the incorporation of rapidly mutating Y-STRs in commercial kits from Applied Biosystems and Promega to increase their discrimination capacity.

1.4.1.1. Y-STR Nomenclature assignment

The use of common nomenclature is crucial in the forensic and population genetic fields to allow communication and data comparison (Gusmão *et al.* 2005). The TAGA-repeat locus of DYS19 (DYS – DNA Y chromosome STR) was the first Y-chromosome STR marker to be discovered and amplified with the polymerase chain reaction (PCR) while the GATA-repeat locus DYS439 was observed several years later (Barni *et al.* 2007). The nomenclature of widely used Y-STRs should not be altered to avoid confusion. This is applied to the core Y-STRs: DYS19, DYS385, DYS389i, DYS389ii, DYS390, DYS391, DYS392, DYS393, DYS438 and DYS439, which are already included in well-known databases and widely-used commercial kits in the DNA forensic field (Gusmão *et al.* 2005). Ideally alleles should be designated according to the total number of repeats, simple or complex, for a specific individual.

1.4.1.2. Problems with STR markers

There are several artefacts that can affect allele assignment during STR analysis. These artefacts may occur as a result of mutations or errors during PCR amplification. Not all alleles for an STR locus contain complete repeat units. Even simple repeats can contain non-consensus alleles that fall between alleles with full repeat units. Microvariants are alleles that contain incomplete repeat units (Butler and Hill, 2012). An example would be an allele 30.2 microvariant rather than allele 31 at a locus. Instead of 31 repeat units (Allele 31) there are 30 repeat units and an incomplete repeat due to a mutation event (Allele 30.2).

A Split peak is the visualisation of two peaks (separated by 1bp) at a single locus marker in profiles. When the non-template "A" addition does not occur with several PCR products the result is one true peak and a second peak which is 1bp smaller. Stutter peaks are formed when slippage between the template and nascent DNA strands occurs during PCR. This leads to the copied strand containing one repeat less than the template strand. The stutter peaks are normally less than 15% of the true amplification product height and occur 4bp before the true peak (Goodwin *et al.* 2007).

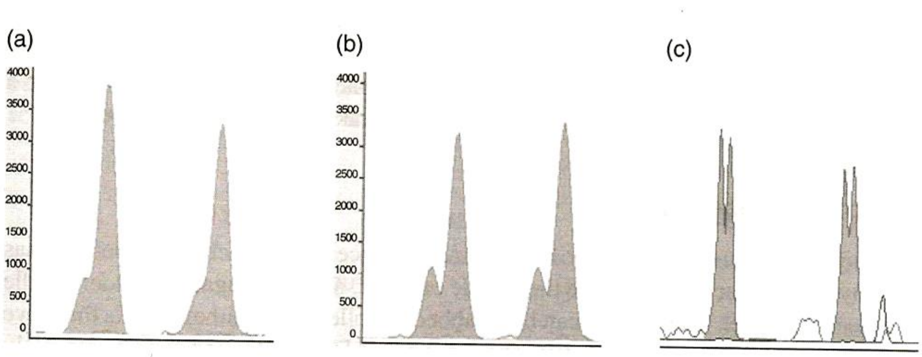


Figure 1.13 Split peaks: Three examples which show decreasing amounts of non-template "A" addition which results in split peak formation. Panel (a) shows an example where the vast majority of PCR products have the non-template addition through to panel (c), where 50% of the PCR product has non-template "A" addition (Goodwin *et al.* 2007).

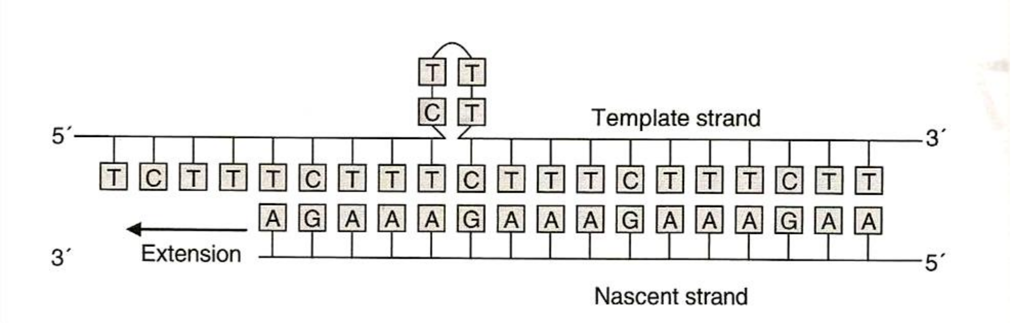

T
T
C
T
5´
Template strand
3´
T C T T T C T T T C T T T C T T T C T T
A G A A A G A A A G A A A G A A
3´
Extension
5´
Nascent strand

Figure 1.14 Stutter peaks: During PCR, slippage between the template and nascent DNA strands leads to the copied strand containing one repeat less than the template strand (Goodwin *et al.* 2007).

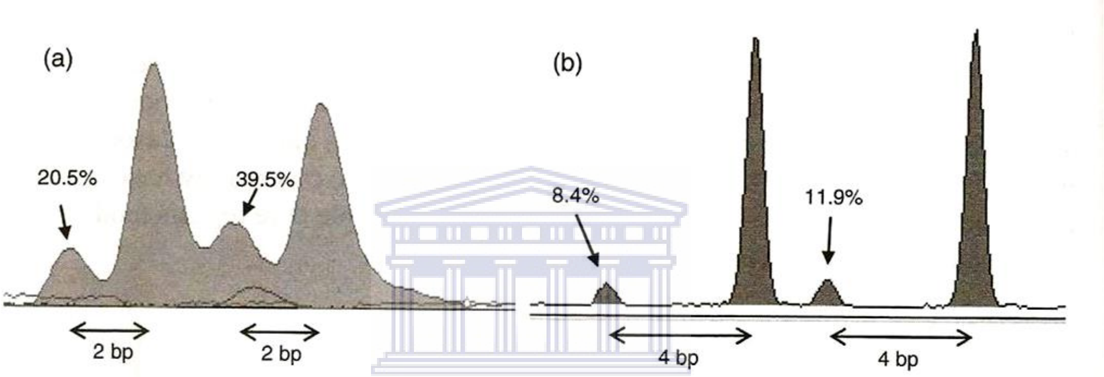


Figure 1.15 Some characteristic stutter peaks. Panel (a) shows a dinucleotide repeat, which is prone to high levels of slippage, the stutter peaks are indicated by the arrow and their size relative to the main peak is shown (based on peak area). Panel (b) is a tetranucleotide repeat, which displays lower levels of stutter (Goodwin *et al.* 2007).

1.4.1.3. Choice of Y-STR loci for use in forensic casework

The potential to distinguish between relatives belonging to the same paternal lineage will be increased due to the accumulation of Y-STR mutations from generation to generation. In identity testing involving male relatives, it is necessary to take mutation rates into account (Gusmão *et al*. 2005). Studies of Y-STR mutation rates have been carried out on a number of markers and have been compiled in various databases such as Y-STR haplotype reference database - YHRD (www.yhrd.org).

1.4.1.4. Constructing a Y short tandem repeat kit

There are several parameters that one must keep in mind when selecting markers for a PCR multiplex. The choice between equally polymorphic simple and complex Y-STRs follows a simple rule. Preference should be given to simple Y-STRs since there is a lower occurrence of artefactual data (Gusmão *et al.* 2005).

Complex Y-STRs can be found in the form of multi-copy loci. An example of a multi-copy locus is the DYS385 marker which shows two male-specific PCR products after amplification. It has been suggested that the repeated sequences are duplicated on the Y chromosome with identical flanking sites allowing duplex amplification of length variable alleles from two independent loci. Because of overlapping sizes of the length variants the origin of the alleles from either of the two loci cannot be unambiguously determined. The frequencies in a population refer to allelic classes and not to single alleles (Kayser *et al*. 1996).

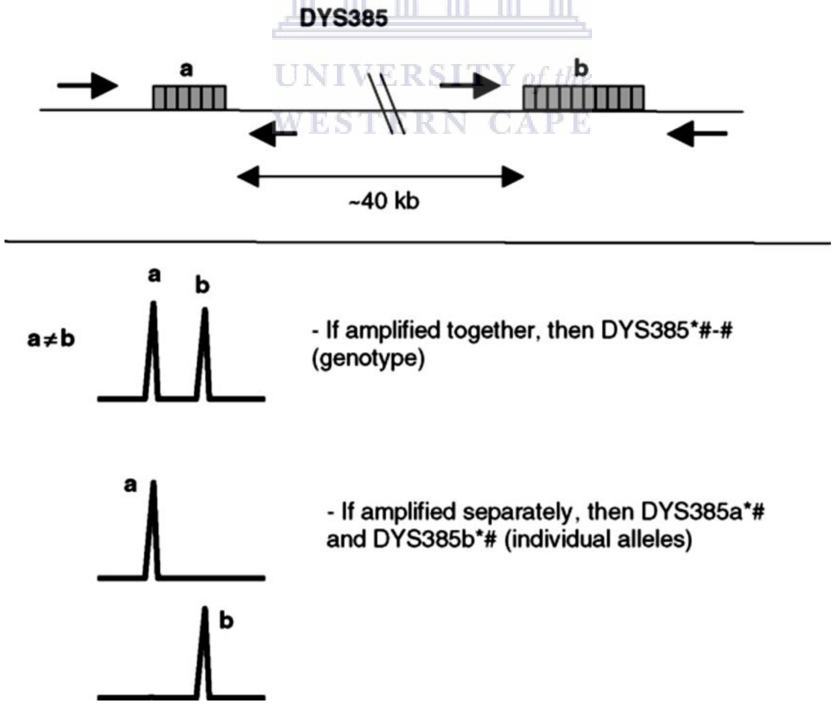

DYS385
a
b
~40 kb
a
b
a≠b
- If amplified together, then DYS385*#-#
(genotype)
a
- If amplified separately, then DYS385a*#
and DYS385b*# (individual alleles)
b

Figure 1.16 Illustration of the multi-copy marker DYS385, which occurs in two inverted regions of the Y-chromosome separated by about 40 kilobases (kb). These regions are typically amplified together because PCR primers anneal to both regions simultaneously due to the presence of identical sequences immediately surrounding the two DYS385 copies. The observed fragments are treated as genotypes and the alleles are recorded by being separated by a hyphen "DYS385 11-14" (Gusmão *et al.* 2005).

Two products of different sizes can also be amplified at the locus DYS389. Sequence analysis shows that the priming site of the forward primer has been duplicated. Thus, the larger (DYS389ii) product includes CTGT/CTAT repeat stretches; whilst the smaller (DYS389i) product includes just one of the two. There is about a 100bp difference in length between the products which allows unambiguous assignment of alleles to either of the two systems (DYS389i or DYS389ii) (Kayser *et al.* 1996).

It is of paramount importance to note that the inclusion of additional Y-STRs to a commercial PCR multiplex can increase its power of discrimination. One must however keep in mind that males who are not closely related based on autosomal DNA evidence may be difficult to differentiate by Y-chromosomal DNA-analysis, even when large numbers of Y-STRs are used (Vermeulen *et al.* 2009). This phenomenon is common in samples from indigenous populations of usually small population size. Additional effects that could cause this observed Y chromosome sharing may include strong male bottlenecks, preferential mating (endogamy) and polygyny, patrilocality and strongly biased male expansion due to male occupation history and privilege (Vermeulen *et al.* 2009).

1.4.1.5. Commercially available Y STR kits

1.4.1.5.1. Minimal haplotype

A basic set of Y-chromosome STR (Y-STR) loci has been widely used in laboratories worldwide for human identity testing and genetic genealogy (Butler, 2003). The minimal haplotype multiplex (MH) was the first set of commercially compiled loci in 1997 (Kayser *et al*. 1997). The MH includes DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, and the polymorphic, multi-copy marker DYS385. This multiplex has a high discriminatory capacity between male individuals in various populations (Kayser *et al.* 1997; Schoske *et al.* 2004; Roewer *et al*. 2009).There are certain limitations to commercial marker sets which leads to the fact that the Minimal Haplotype, traditionally used in forensic and anthropological studies, requires supplementation with additional Y-STRs (Vermeulen *et al.* 2009).

1.4.1.5.2. Extended haplotype

The extended haplotype was initially constituted simply by including the duplicated dinucleotide locus YCA II to the loci of the minimal haplotype. Upon recommendations by the Y-chromosome subcommittee of the Scientific Working Group on DNA Analysis Methods (SWGDAM), in 2003, two additional Y-STRs named DYS438 and DYS439 replaced YCA II in the U.S. minimal haplotype forming the extended haplotype marker (Butler, 2003). Though locus YCA II was highly polymorphic (Kayser *et al*. 1997) it was replaced because it is a polymorphic dinucleotide marker that suffers from 'stutter' products during PCR process due to polymerase spillage products (Hammer *et al.* 2006i). Stutter products which are pronounced dinucleotide repeats provide additional resolution in the sample mixture of a multiple-male DNA profile (Butler and Coble 2007) Some of the stutter products are as high as 50 % of the height of the true allele and they complicate the exact interpretation of results (Butler, 2003).

1.4.1.6. Recently designed Y-STR multiplexes

If a 'new' Y-STR is considered for addition to an existing set of Y-STRs, the additional information the extra Y-STR will add to the information obtained by the original set of Y-STRs needs to be investigated. Due to the lack of recombination between Y-specific loci, the whole haplotype is transmitted as a single marker, and haplotype diversity defined by a set of STRs must be established by frequency estimates of the whole haplotype. The haplotype diversity cannot be predicted by combining the average diversity at each single locus. There are two main factors that contribute to single-locus diversity, namely the presence of distinct lineages (several lineages in different combinations may be population-specific) and the variation accumulated within each lineage by mutation. Accrued mutations will contribute to the decrease in association between alleles at different loci and therefore be reflected in the Y-STR diversity at the haplotype level (Gusmão *et al.* 2005). Y-STR diversity is studied in relation Y-SNP defined haplogroups rather than in specific populations in order to choose the best markers to increase Y-STR haplotype discrimination capacity in forensic genetics.

When designing a multiplex it is important to note that programs for STR detection vary significantly in repeat definition, search algorithm and filtering method. Whole genome searches for the presence of microsatellites is less time consuming for simple tandem arrays as opposed to more complex ones (Merkel *et al.* 2008). This reinforces the fact that the majority of commercial multiplex kits contain simple tandem repeats.

1.4.1.6.1. The UWC 10plex

A recent multiplex kit was developed based on population data obtained from individuals born in Sub-Saharan Africa (D'Amato *et al.* 2011). This multiplex kit, UWC 10plex, consists of ten Y-STR loci that successfully discriminate between individuals of this area. It is currently in the time-consuming process of being validated before it can be considered for use in the commercial market. The primary objective of the research behind this multiplex kit is to obtain a genotyping system of higher discrimination capacity on male individuals in South Africa. The four major population groups according to the South African government are Caucasian, Indian, Coloured and Black African. Samples with Indian, European, African (Xhosa, Zulu) and Cape Coloured ancestry were incorporated during developmental validation of this kit. Data generated using this kit will aid in developing a male specific genetic typing system that can be used in genealogical and evolutionary studies. The UWC 10plex shows potential for application in human identification studies on forensic casework from South Africa.

In a case study by Leat et al. (2004) it was found that there were low levels of polymorphism at two loci of the MH, DYS391 and DYS392, in the Xhosa population from the Cape Town metropolitan area. Similar results were also observed for an indigenous population sampled in Maputo, Mozambique with 82% of the sample sharing the same allele. This data suggests that the DYS391 and DYS392 loci are not ideally suited to forensic casework in Sub-Saharan Africa populations (Leat *et al.* 2004).

There is a lengthy process involved with developing a novel PCR multiplex kit and the gene diversity values, the number of alleles identified and the average stutter product needs to be determined for each locus. This process results in the identification of several highly polymorphic loci that are specific to the sub populations involved. DYS385 is the only shared marker between the UWC 10plex and the EH.

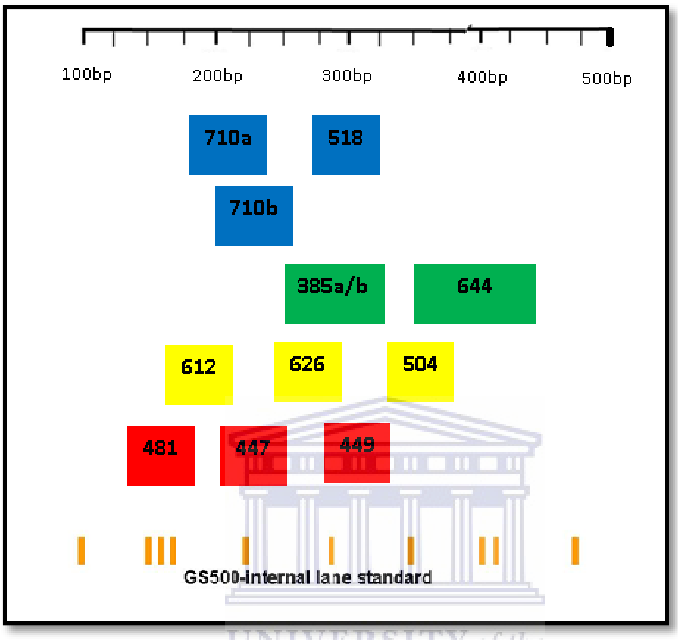


Figure 1.17 Diagram showing the fluorescent dye labels and size range for the UWC 10plex according to D'Amato *et al*. (2011).

1.4.1.6.2. Rapidly mutating Y-STRs

The widely used commercial panels of between 12 loci (Promega PPY12) and 17 loci (Y-filer) have adequate resolution of different paternal lineages in many human populations. They do however have a lower power of resolution in separating out paternal lineages from populations expressing low Y-chromosome diversity. A global study on 19,630 Y chromosomes from 129 worldwide populations using the PPY23 (PowerPlex Y23, Promega Corporation) kit showed substantially stronger discriminatory power than other available kits but at the same time revealed the same general patterns of population structure (Purps *et al.* 2014). In this study, a comparison of haplotype diversity (HD) was done using five commercial kits with an increasing amount of markers.

The results showed the highest HD value for the PPY23 kit at 92.9%, the Y-filer (17 markers) had 77.8% and the PPY12, SWGDAM marker set and MH all had diversity values below 50%. Increasing the number of markers was found to be associated with an almost linear increase of all forensic parameters used to discriminate among individuals, including the number of different and unique haplotypes (Purps *et al.* 2014). There is however a failure to differentiate between related males who belong to the same paternal lineage which is undesirable for forensic interpretations (Ballantyne, 2012). A new panel of thirteen rapidly mutating (RM) Y-STRs with mutation rates above $1x10^{-2}$ (1 mutation/100 generations/ locus) has been identified to get around this problem. Previously used forensic markers had mutation rates in the order of $1x10^{-3}$ or lower. The incorporation of only a few RM Y-STRs into a Y-filer plus kit was found to greatly increase its discrimination capacity (Thermo Fischer).

A study by Ballantyne *et al.* (2012) on 604 unrelated male samples from 51 worldwide populations demonstrates that RM Y-STRs provide substantially higher haplotype diversity and haplotype discrimination capacity than currently used commercial Y-STR kits. The afore-mentioned RM Y-STR panel is extremely successful at differentiating between closely and distantly related males. It was found that 66% of the three hundred and five related males tested with this panel could be differentiated from each other. This is a rare achievement, considering the reliability and widespread use of the Y-filer multiplex, which only showed a 15% ability to differentiate between related males in the same test group.

The RM Y-STR panel is potentially powerful enough to separate closely related males in father-son pairs as well as brothers. Overcoming this hurdle proves a vital step in pushing past the current limitations of Y chromosome analysis in forensic genetics. It provides a good basis for the implementation of a RM Y-STR kit for use in the forensic setting and can be easily incorporated into a much need DNA database here in South Africa.

Table 1.1 Table incorporating mutation rates of markers in the Applied Biosystems AmpFLSTR® YFiler™ kit as well as the RM Y-STRs used in the Ballantyne *et al.* 2012 study.

AmpFℓSTR® YFiler™		**RM Y-STRs**	
Marker	**Mutation Rate**	**Marker**	**Mutation Rate**
DYS19	2.7 X10^{-3}	DYF387S1	1.59X10^{-2}
DYS385a/b	2.1 X10^{-3}	DYS399S1	7.73 X10^{-2}

DYS389I/II	I= 2.8 X10^{-3}
	II= 3.1 X10^{-3}
		DYS403S1a/b	A= 3.10 X10^{-2}
			B= 1.19 X10^{-2}

DYS390	2.2 X10^{-3}	DYS404S1	1.25 X10^{-2}
DYS391	3 X10^{-3}	DYS449	1.22 X10^{-2}
DYS392	6 X10^{-4}	DYS518	1.84 X10^{-2}
DYS393	1 X10^{-3}	DYS526a/b	1.25 X10^{-2}
DYS438	6 X10^{-4}	DYS547	2.36 X10^{-2}
DYS635	3.9 X10^{-3}	DYS570	1.24 X10^{-2}
DYS437	1.3 X10^{-3}	DYS576	1.43 X10^{-2}
DYS448	6 X10^{-4}	DYS612	1.45 X10^{-2}
DYS456	4.4 X10^{-3}	DYS626	1.22 X10^{-2}
DYS458	7.1 X10^{-3}	DYS627	1.23 X10^{-2}
Y GATA H4	3.1 X10^{-3}		
DYS439	5.6 X10^{-3}		
Average mutation rate	**2.4 X10^{-3}**		**2.01X10^{-2}**

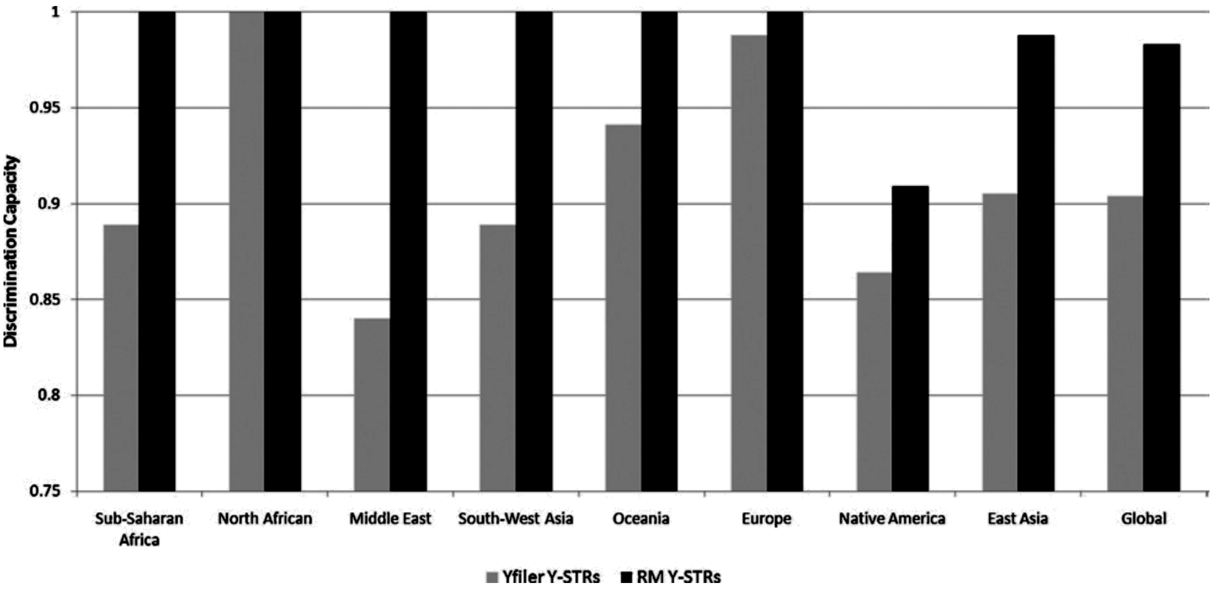


Figure 1.18 Discrimination capacity (DC) achieved with RM Y-STR and Yfiler Y-STR sets in eight geographic regions, and globally from the Human Genome Diversity Panel provided by the Centre d'Etude du Polymorphisme Humain (HGDP-CEPH). The RM Y-STR set displayed significantly greater discrimination power in 7 of the 8 regions, and an 8% increase globally, relative to the Y-filer set (Ballantyne *et al. 2012*).

1.4.2. Human identification applications of Y chromosome polymorphisms

Regions of the Y chromosome are commonly isolated for purposes of identification, to infer population histories, discover genealogical relationships and identify males for criminal justice purposes (Ballantyne *et al.* 2010). In forensic analyses, Y-STRs are often used to determine the number of individuals contributing to a mixture of DNA in a biological sample or stain. They are useful for characterizing male DNA in biological material from sexual assault cases. With Y chromosome haplotype data being of particular use in cases where it is challenging to separate the perpetrators DNA from the victim's (Hammer *et al.* 2006i).

Knowledge of Y-STR mutation rates is a key requirement for statistical analysis and application of its data. There are a number of factors that affect Y-STR locus mutatbility including repeat complexity (simple or compound), the length in base pairs of the repeated motif, the father's age and most strongly the total repeat number of the Y-STR. In this case the larger the numbers of repeats, the more likely mutations are incurred (Ballantyne *et al.* 2010).

1.4.2.1. Paternity testing

DNA sequence comparison between individuals can show whether one of them was derived from the other. Specific sequences are analysed to see whether they have been copied directly from one individual's genome to another. If this is the case, then the genetic material of the older individual indicates a parental role over the younger individual. In testing the paternity of a male child, comparison of the Y chromosome can be used since it is passed directly from father to son (Kayser *et al.* 2001).

Accurate knowledge of the mutability of the applied Y-STRs is needed to obtain reliable paternity probabilities (Ballantyne *et al.* 2010). The Y chromosome can be used confidently for exclusions of an alleged father. The unusual properties of the Y chromosome means that inclusions will be more difficult to make: haplotypes are confined within lineages because of the lack of recombination in the Y chromosome, so population sub-structuring is a major problem (Jobling *et al*. 1997). It is of paramount importance in paternity cases that the correct outcome is achieved so that parental custody can be given.

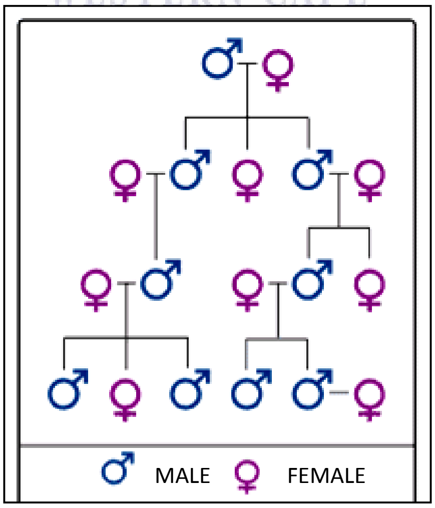


Figure 1.19 A family tree showing the transmission of the Y chromosome from father to son. Every male in this line will have the same Y-STR profile. A son's haplotype can therefore be compared to his alleged father's haplotype. A match indicates kinship as all males in the same line will have the same STR profile. Autosomal markers need to be compared to confirm the match (www.ibdna.com).

1.4.2.2. The importance of databases

Databases can be separated into reference and criminal platforms. In America during 1997 the National Institute of Standards and Technology (NIST) compiled a Short Tandem Repeat DNA Internet Database (http://www.cstl.nist.gov/biotech/strbase/) commonly referred to as STRBase. This database is an information resource for the forensic DNA typing community which contains details on commonly used STR markers. Observed alleles and the annotated sequence for each STR locus are described along with analysis techniques. Commercial STR multiplex kits are also included *i.e.* AmpFℓSTR Identifiler along with published PCR primer sequences (Ruitberg *et al.* 2001).

The Y-STR Haplotype Reference Database (YHRD) is a searchable worldwide database of Y-STR haplotypes from various populations around the world. It contains five databases generated using various kits: Minimal Haplotype (YHRD core loci), Promega Powerplex® Y, Applied Biosystems AmpFℓSTR® YFiler™, Promega Powerplex® Y23 and the Applied Biosystems AmpFℓSTR® YFiler™ plus. Input data from a study using one of the kits can be compared to the database to get necessary population and ancestry information. It is possible for the phylogenetic branch position of the target sample to be retrieved using the combined STR/SNP database and further statistical tools can be used to further refine the results. This includes Analysis of Molecular Variance (AMOVA) Kinship Analysis and Mixture Analysis.

The UK National DNA Database (NDNAD) was the first criminal database to be established shortly after the advent of the extended haplotype multiplex. DNA databases that store STR profiles from convicted felons have emerged as a powerful tool in the investigation of crime. The effective use of this DNA database acted as a catalyst for the establishment and expansion of criminal DNA databases in other countries. There is also a need for a reference database which compiles any suspects and previously convicted felons for comparison with all evidentiary samples. Forensic behavioural psychologists have found an escalating pattern of violence with repeat offenders (Pistorius, 2005). Therefore it is of paramount importance to keep a database of previous offenders for comparisons at new crime scenes.

The Combined DNA Index System (CODIS) is a set of local, state and national databases of DNA profiles in the USA which began in 1990. The CODIS system consists of three types of databases categorised by the type of information they contain. The first database contains DNA profiles that are obtained from crime scenes (forensic database). In most cases the source of this DNA is not known. The second database consists of profiles of criminal offenders and sometimes even those arrested for felonies and misdemeanours. Different states have different criteria for what DNA types will be contributed. If a crime occurs, CODIS may be searched to see if the offender's DNA is on file or if the DNA recovered from the scene is also found at another scene, indicating that a serial criminal is probably at large. The third and most recent database is of missing persons (Houck *et al.* 2010).

In South Africa the non-profit organisation, The DNA Project (www.dnaproject.org.za) aided in promulgating legislation necessary to form a DNA database. It will be used by the South African Police Services (SAPS) to aid in apprehending perpetrators. This project is of paramount importance and the resulting reference index database is of paramount importance in catching repeat offenders. The Criminal Law (Forensic Procedures) Amendment Act 37 of 2013 or DNA Act was passed into law on the 27th of January 2014 (http://www.justice.gov.za/legislation/acts/2013-037.pdf). The passing of this Act will lead to the establishment and maintenance of a National Forensic DNA Database (NFDD) in South Africa.

1.4.3. DNA typing methodology

Current forensic DNA typing is conducted in approximately 8-10 hours with steps including DNA extraction, quantification, polymerase chain reaction (PCR) amplification of multiple STR loci, capillary electrophoresis separation with fluorescence detection, data analysis and DNA profile interpretation (Vallone *et al*. 2008).

1.4.3.1 DNA isolation

STRs can be isolated from a variety of biological materials including: seminal fluids from sexual assault cases; epithelial cells from dandruff, faeces and more recently touch DNA; hairs with the root attached; saliva from cigarettes, drinks or bite marks and white blood cells from blood evidence. Blood is the primary biological material found at crime scenes and is easier to visualize than other biological fluids such as saliva (Goodwood *et al.* 2007). In the research/ laboratory setting cheek swabs can be used to collect DNA contained in buccal cells. These swabs are then put through a simple DNA extraction method to lyse the cells and expose the high purity DNA. The extraction success rate is determined by the amount of DNA extracted and the successful removal of PCR inhibitors.

The choice of extraction method is dependent on the sample type and quality. DNA extraction methods have been designed to work in a step-wise manner by first disrupting the cellular membrane, then denaturing any proteins present using a proteinase (*e.g*. Proteinase K) and finally separating the DNA from any proteins or cell debris present. A few common extraction methods include Chelex 100 ©, silica based extraction, phenol-chloroform extraction and ethanol precipitation.

Chelex 100 © resin based extraction which results in the isolation of ssDNA from polyvalent metal ions. The silica based extraction method binds DNA in the presence of a chaotrpoic salt which is subsequently released in a low salt concentration medium. Phenol-chloroform extraction involves the denaturation of proteins in the cell lysate. This method is no longer commonly used because phenol is toxic and the method involves multiple tube changes which can easily lead to contamination (Goodwood *et al.* 2007). An ethanol precipitation method has proven to be cost effective and relatively efficient with concentration yields around 20ng.ul^{-1} for buccal swabs. Samples need to be quantified once they have been successfully extracted. This is done to ensure that the extracted DNA is pure and so that correct concentration of DNA can be utilised in the subsequent PCR multiplex reaction. The incorporation of an incorrect concentration of DNA in PCR reaction can lead to a failed amplification.

1.4.3.2. Polymerase Chain Reaction (PCR) and multiplex assays

The polymerase chain reaction amplifies specific regions of target DNA with the use of primer pairs. The power of this amplification technique means that, in theory a single molecule can be amplified one billion-fold by 30 cycles of amplification. In practice the PCR is not 100% efficient but it does still produce tens of millions of copies of the target sequence (Goodwin *et al.* 2007). The amplification of DNA occurs in the cycling phase of PCR, which consists of three stages displayed in figure 1.20. During denaturation the dsDNA is melted at a high temperature to form two ssDNA molecules. The temperature is then lowered and the primers anneal to the template strand. After the primers have annealed, the temperature is then increased to the optimum working temperature for *Taq* polymerase. The enzyme catalyzes the addition of nucleotides to the 3' end of the primers using the original DNA strand as a template.

The PCR requires tightly controlled thermal conditions and these are achieved using a thermocycler. This consists of a conducting metal block with heating/cooling elements and wells that accommodate the plastic reaction tubes (Goodwin *et al.* 2007).The PCR amplification portion of the workflow typically takes approximately 3 hours with standard thermal cycling protocols. The use of commercially available polymerases such as PyroStart (Fermentas ©) or SpeedSTAR (Takara Bio USA ©) and a rapid thermocycler can be used in order to speed up this process at a costly fee. For the purpose of population studies the 3 hour cycling period is acceptable because the results do not necessarily need to be analysed within a short time period as with casework (Vallone *et al.* 2008). A standard *Taq* polymerase that is supplied in the majority of the commercial kits can be utilised, such as the AmpliTaq Gold polymerase.

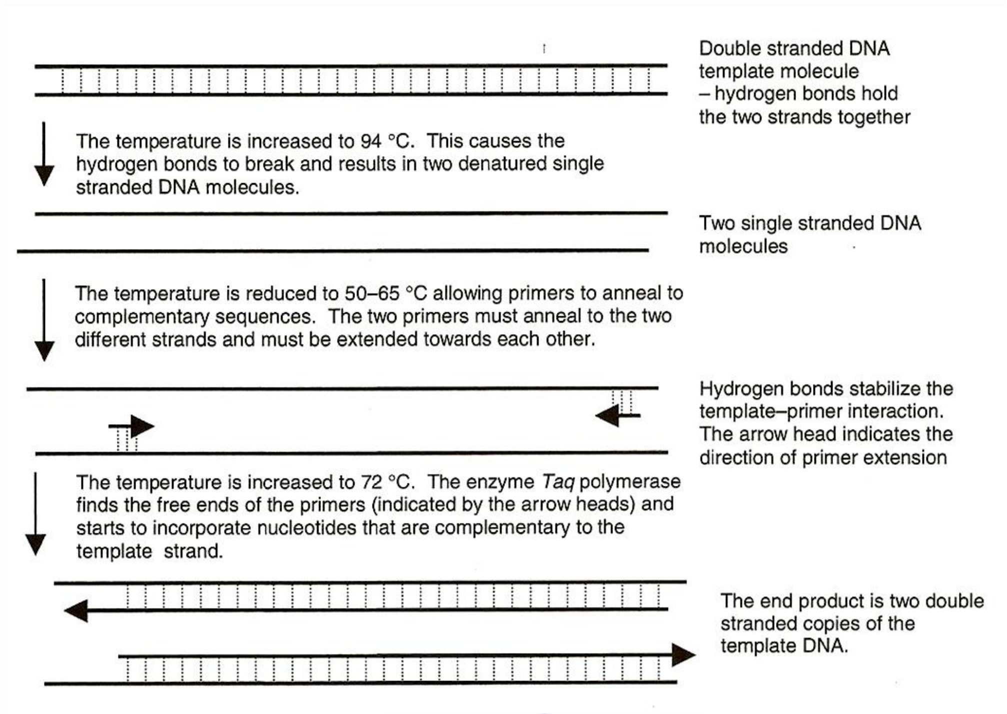

Double stranded DNA template molecule – hydrogen bonds hold the two strands together
The temperature is increased to 94 °C. This causes the hydrogen bonds to break and results in two denatured single stranded DNA molecules.
Two single stranded DNA molecules
The temperature is reduced to 50–65 °C allowing primers to anneal to complementary sequences. The two primers must anneal to the two different strands and must be extended towards each other.
Hydrogen bonds stabilize the template–primer interaction. The arrow head indicates the direction of primer extension
The temperature is increased to 72 °C. The enzyme *Taq* polymerase finds the free ends of the primers (indicated by the arrow heads) and starts to incorporate nucleotides that are complementary to the template strand.
The end product is two double stranded copies of the template DNA.

Figure 1.20 The PCR process – each PCR cycle consists of three phases: denaturing, annealing and extension. A mere thirty cycles can result in tens of millions of copies of the template DNA (Goodwin *et al.* 2007).

Since DNA extraction does not produce pure DNA, some chemicals will co-purify and inhibit the action of the *Taq* polymerase. High concentrations of ions such as calcium and magnesium can act as potent inhibitors. When it is not possible to remove all the possible inhibitors from a DNA extract, the addition of the protein bovine serum albumin (BSA) to the PCR can prevent or reduce the inhibition of the *Taq* polymerase (Goodwin *et al*. 2007). Due to the high sensitivity of the PCR contaminants such as extraneous DNA can easily be co-amplified. Therefore preparation of PCR samples needs to be done in a sterile environment and as such, the laboratory is clearly divided into preparation and PCR rooms.

1.4.3.3. Electrophoresis and results interpretation

After Y-STR polymorphisms have been amplified using PCR, the length of the products must be precisely measured. Gel electrophoresis of the PCR products through denaturing polyacrylamide gels can be used to separate DNA molecules between 20-500 nucleotides long. Early systems utilised slab gels but there has been a movement to capillary electrophoresis (Butler and Hill, 2012). A series of covalently attached fluorescent dyes can be used to detect STR markers within the same size ranges. Up to six different dyes can be used in a single multiplex analysis on the Applied Biosystems 3500xl genetic analyzer; which allows for considerable overlap of products (Goodwin *et al.* 2007). In capillary electrophoresis, narrow glass tubes are filled with an entangled polymer solution to separate the DNA molecules by size. The reference marker, or internal-lane size standard, is added to the post-PCR product to correctly assign repeat number, along with formamide which denatures the dsDNA.

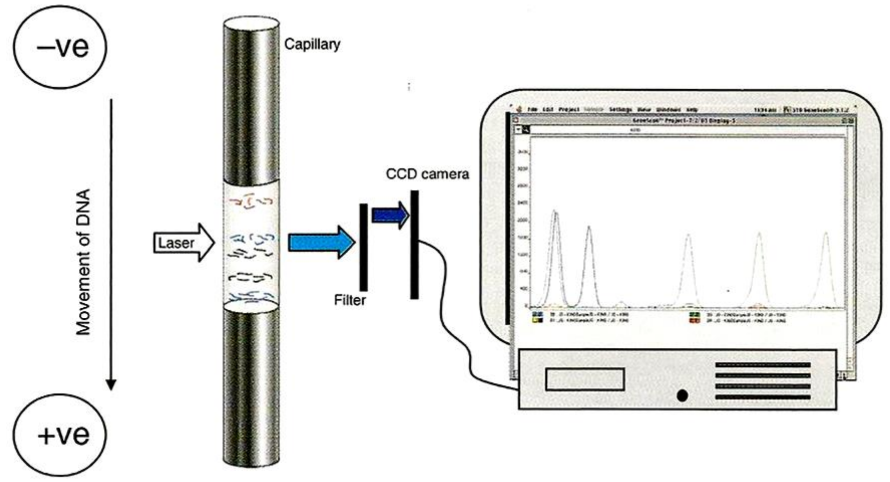


Figure 1.21 Diagram showing the DNA detection process during capillary electrophoresis. DNA moves from cathode to anode in the capillary where it is detected by a CCD camera through a detection window. The physical data is transformed into electronic data which can be interpreted by an analyst (Applied Biosystems).

During electrophoresis an argon/ solid state laser is shone through the detection window in a capillary. The labelled PCR products are separated according to size, shape and charge as they migrate through the polymer towards the anode. When the laser hits the fluorescent label on the PCR products, the label is excited and emits fluorescent light that passes through a filter to remove any background noise. They then go on to a charged coupled device camera that detects the wavelength of the light emission and sends the information to a computer where software records the profile. There is a conversion of physical fluorescently labelled DNA products into electronic data in the form of an electropherogram.

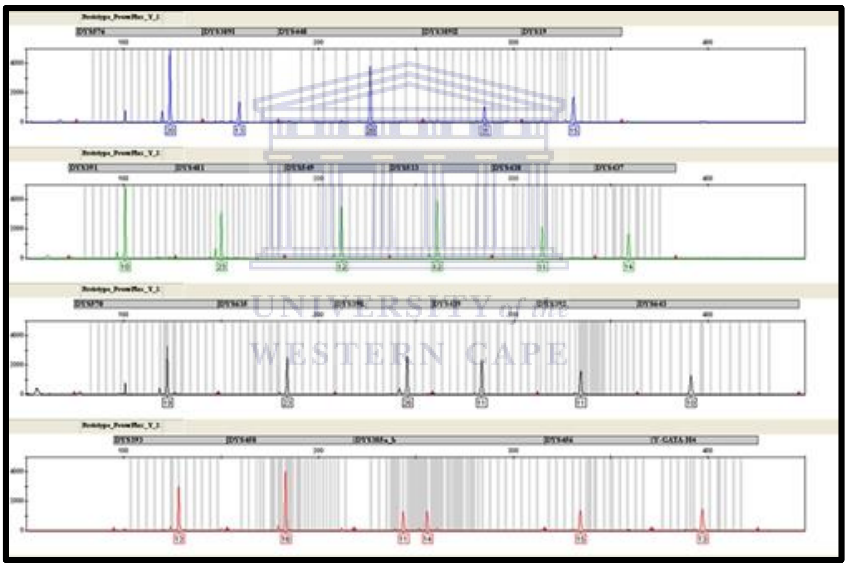


Figure 1.22 A typical electropherogram showing PCR products using four fluorescent dyes on the Applied Biosystems 3130 Genetic Analyzer. Each peak represents an allele and the amount of PCR product detected is measured as relative fluorescent units (RFU’s). The higher peaks indicate more fluorescence (Applied Biosystems).

Chapter 1, Section 2: Aims and objectives

Indigenous peoples of South Africa are currently making various land claims for ancestral grounds that may have been dispossessed during this tumultuous time period. The results of ancestry informative studies provide clarity for admixed “indigenous” population groups such as the Griqua. They have a complex heritage structure and their cultural identity is based on the struggle that their ancestors endured during the colonisation events that shaped South Africa.

In May 2013 the Forensic DNA Lab from UWC collected DNA samples at the 94th annual Griqua National Conference (GNC) of South Africa under leadership of Alan A. Le Fleur. Participants from various parts of South Africa attended this 3 day cultural festival to pay homage to their ancestors. The Griqua traditions are upheld by the people and can be preserved as they are passed down from generation to generation at these heritage festivals. Representation of the Griqua people at the GNC of 2013 was primarily from the elderly leaders and adults, with young adults being reluctant to categorize themselves as Griqua. Genetic classification into an African haplogroup may help younger individuals come to terms with their cultural roots and maternal ancestry. Self-identification as a Griqua with ancestors in the population group was sufficient to qualify as a participant for this project.

The primary objective of this Masters project was to investigate the maternal ancient substructure of the Griqua population in South Africa. Genetic ancestry was determined by investigating ancestry informative single nucleotide polymorphisms (SNPs). These SNPs are located in the control region of the mitochondrial genome. The auxiliary aim was to test the validity of the UWC 10plex system in relation to a sample group of Griqua males. This short tandem repeat (STR) multiplex targets specific mutations confined to paternal lineages.

Chapter 2: Methods and materials

2.1. Sample collection and population structure

Ethical clearance 10-3-39 was obtained for the project through the University of the Western Cape and each individual had to sign a consent form pertaining to the use of their DNA post-sampling. All individuals that took part in sampling were interviewed by lab personnel regarding their ancestry and place of birth. Self-classification and either maternal or paternal Griqua ancestry up to a third generation was sufficient to classify an individual as Griqua. All sampling was organised through Griqua community and clan leaders from various regions around South Africa. DNA samples were collected in three separate sampling trips: one in 2011 and two in 2013. Sixty three buccal swabs were collected in Kokstad (2011), thirty one buccal swabs were collected in Vredendal (January 2013) and eighty two buccal and saliva samples were collected from unrelated Griqua individuals that attended the 95th annual Griqua National Conference (GNC) held at the Griqua heritage site Ratelgat (March 2013).

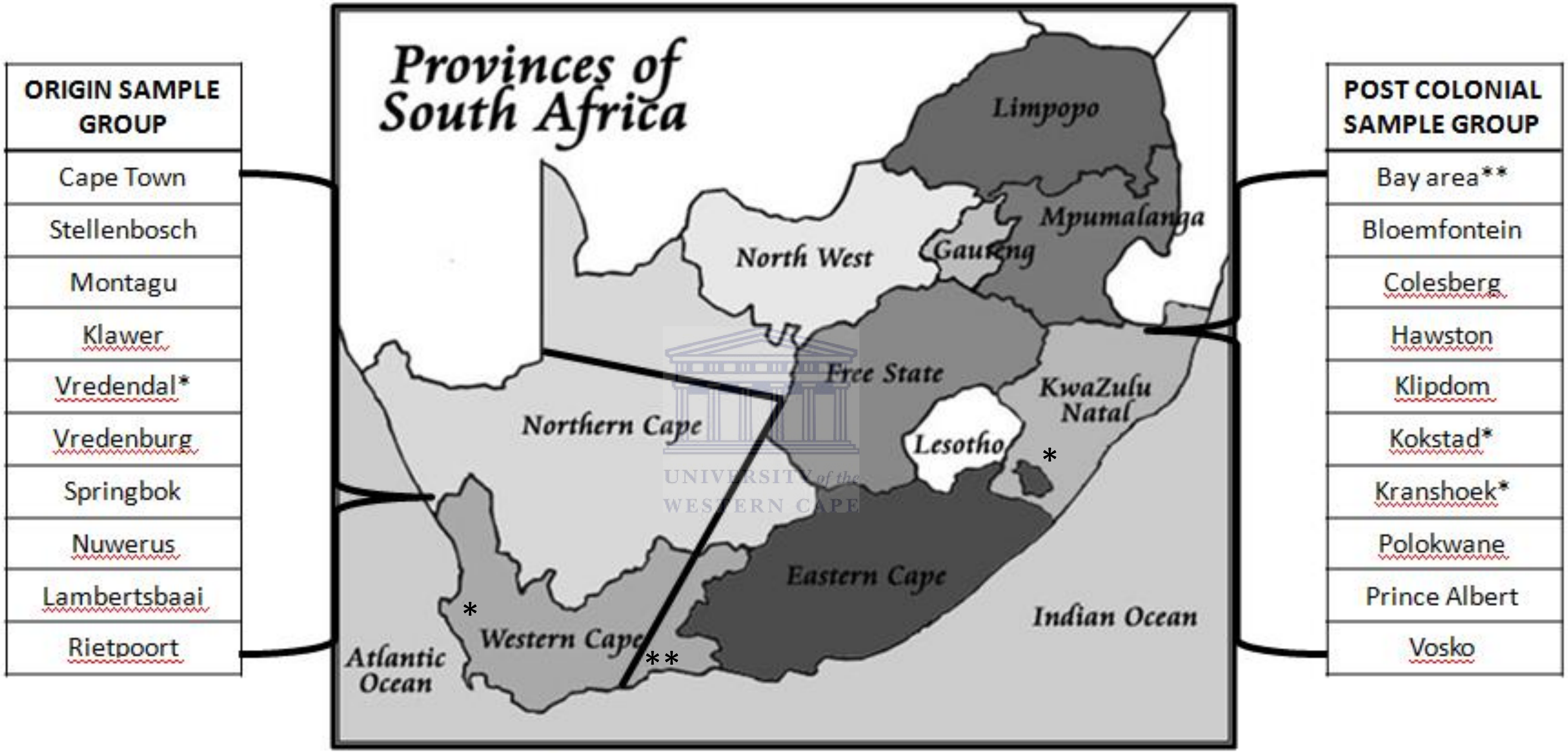
The buccal swabs were gently massaged against the inner cheeks of the volunteers for 30 seconds before being placed into a marked Eppendorf tube. Two millilitre saliva samples were also obtained from individuals that attended the GNC. Saliva samples were collected to ensure ample DNA for future work on the population. The samples were placed into a freezer box until -20°C storage could be accessed and kept there until the DNA could be extracted.

One hundred and four unrelated Griqua male samples were used from the three sampling trips in 2011 and 2013 for Y chromosome short tandem repeat analysis. The validity of the UWC 10plex was tested on this South African population group by assessing statistical values, including discrimination capacity and haplotype diversity.

For mitochondrial DNA haplogroup assignment, all samples collected (N=176) were divided into two sample groups:

One that is representative of the original ancestral region of the Griqua's along the West Coast of South Africa. This included individuals from Cape Town, Klawer, Lambertsbaai, Montagu, Nuwerus, Rietport, Springbok, Stellenbosch, Vredenburg and Vredendal. It was termed the ORIGIN SAMPLE GROUP (N=82).

The second sample group is representative of a portion of the Griqua population that has been exposed to colonisation events along the Griqua Trek route (Northern Cape, KwaZulu Natal and Western Cape). This means exposure to other indigenous population groups as well as colonial influences. Samples were from Bloemfontein, Brets, Colesburg, Griekwastad, Hawston, Hermanus, Klipdom, Knysna, Kranshoek, Mosselbaai, Plettenberg Bay, Polokwane, Prince Albert, Springfontein, Tsitsikama and Vosko along with the set of samples collected by a colleague from Kokstad in 2011 (N=63). It was termed the POST COLONIAL SAMPLE GROUP (N=94).



ORIGIN SAMPLE GROUP
Cape Town
Stellenbosch
Montagu
Klawer
Vredendal*
Vredenburg
Springbok
Nuwerus
Lambertsbaai
Rietpoort

POST COLONIAL SAMPLE GROUP
Bay area**
Bloemfontein
Colesberg
Hawston
Klipdom
Kokstad*
Kranshoek*
Polokwane
Prince Albert
Vosko

* Concentrated Griqua population
**Bay area – Hermanus, Mosselbaai, Knysna, Tsitsikama Plettenbergbaai

Figure 2.1 Map displaying the two sample groups that will be compared. The origin sample group (N=82) is one that is representative of the core Griqua population prior to the Griqua Trek and as such the maternal ancestry is expected to have more African (Khoisan) influences than the post-colonial sample group (N=94).

2.2. DNA extraction from buccal swabs

DNA extraction from buccal swabs was performed according to Medrano (1990) using the salting out extraction protocol which contains a salt lysis extraction solution and Proteinase K. The surface of the swab was cut off using a pair of sterile scissors. The swab surface was then placed into a labelled Eppendorf tube containing 600µl of lysis buffer and 3µl proteinase K. The tube was vortexed (Vortex mixer model SA3) for 30 seconds and incubated, with shaking, overnight at 56°C (Vortemp 56). The lysis buffer with biological material still trapped in the swab was recovered by perforating the end of a 0.5ml tube with a needle (21-22 gauges). The piece of swab was placed into this 0.5 ml tube which was placed inside a 1.5ml Eppendorf tube. The tubes were then centrifuged (Eppendorf Centrifuge 5414 D, rotor radius = 7.3cm) until the swabs were dry (about 1 minute). The collected volume was added to the previously separated lysis material.

2.2.1 DNA precipitation and purification

The precipitation was done by adding 200µl of 5.5M NaCl into the tube of lysis material. The tube was shaken vigorously for 15 seconds. It was then centrifuged for 15 minutes at 5000rpm, and the supernatant with DNA was transferred to another clean labelled tube. 800µl of cold isopropanol (-20°C) was added to the tube and it was left at -150°C for 15 minutes in the freezer (Nuaire Ultralow Temperature Freezer). The DNA pellet was recovered by centrifugation at 14000rpm for 30 minutes and the supernatant gently removed. The pellet was washed with 100µl of cold 70% ethanol to remove the salts. The tube was then centrifuged at 14000rpm for 15 minutes. The ethanol was discarded without disrupting the pellet, which was dried in a Speedy Vac briefly. The DNA pellet should not become too dry as this makes it challenging to dissolve. The DNA pellet was dissolved in 30µl of SABAX water and stored at -20°C. The DNA concentration was measured with the Nanodrop ND1000 Spectrophotometer. Working dilutions of 4ng/µl were prepared using the Nanodrop readings and stored at -20°C.

2.2.2 DNA extraction from saliva samples

Saliva extraction was initially done according to the protocol outlined by Kayser *et al*. (2006) which utilises a salt lysis extraction buffer and Proteinase K. Later modifications included a phenol-chloroform cleaning protocol to allow for a cleaner DNA yield. Two millilitres of saliva was mixed with an equal volume of saliva lysis buffer (50mM Tris pH 8.0, 50mM EDTA, 50mM sucrose, 100mM NaCl, 1% SDS). 700µl of saliva in buffer was then added to an equal volume of Phenol:Chloroform:Isoamyl (25:24:1) in a separate 2ml Eppendorf tube. The mixture was vortexed (Vortex mixer model SA3) for 10 seconds and then centrifuged at 10 000rpm for 10 minutes (Centrifuge 5414 D).

An equal volume Phenol:Chloroform:Isoamyl was then added to the recovered supernatant and it was centrifuged for a further 10 minutes at 10 000rpm. An equal volume Chloroform was then added to the recovered supernatant followed by a centrifuge step for 10 minutes at 10 000rpm. The supernatant was recovered once more and 1.4mL of 100% Ethanol was added. The 2mL Eppendorf tubes were then placed in a - 80°C freezer overnight (Nuaire Ultralow Temperature Freezer).

The DNA was then pelleted by centrifuging at 13 000rpm for 30 minutes. The supernatant was discarded and the pellet washed with 70% Ethanol followed by another centrifuge step of 30 minutes at 13 000rpm. The pellets were then dried for approximately 30 minutes at room temperature and then resuspended in 50µL 1xTE buffer. The DNA concentration was measured with the Nanodrop ND1000 Spectrophotometer. The working dilutions of 4ng/µl were prepared from the Nanodrop readings and stored at -20°C.

2.4. Mitochondrial DNA protocols

Sequence variation in the form of Single Nucleotide Polymorphisms (SNPs) was tested on the mitochondrial deoxyribonucleic acid (mtDNA) samples of unrelated males and females from the Griqua population in South Africa. The combination of alleles at multiple SNPs within a highly polymorphic region of the mitochondrial genome defines a mtDNA haplogroup (Hammer and Redd 2006ii). These haplogroups are representative of time-specific polymorphisms that are present in a group of individuals at a distinct geographical location. This means that the mitochondrial DNA is utilised for its ancestry informative properties.

2.4.1. DNA amplification of the mitochondrial genome's control region

Two primers amplifying the 1121bp control region were utilised in the PCR reaction outlined in table 2.1 using the cycling conditions are in table 2.2. Amplification was done to a final volume of 130µL to ensure enough amplicons for downstream processing.

Table 2.1 Reagents and concentrations used for PCR-amplification of the 1121bp control region fragment within the mitochondrial genome. Primers designed by D'Amato (unpublished).

Reagent	[Stock]	Volume (µl)	[Final]
Sabax water	-	84	-
Buffer	10x	13	1x
dNTPs	2mM	13	0.2mM
Primer L15969	10µM	10	0.77µM
Primer H658	10µM	10	0.77µM
Taq	5U	0.2	0.007U
Total	-	130.2	-

Table 2.2 Cycling conditions used for amplification of the 1121bp mitochondrial DNA control region

Stage	Temperature	Time (mins)	Cycles
Initial denaturation	94°C	5	x1
Denaturation	94°C	0.75	x38
Annealing	55°C	0.75	
Elongation	72°C	2	
Final elongation	72°C	2	x1
Hold	4°C	-	-

2.4.2 Verification of amplified DNA

After the PCR of the samples was completed, a 2% (w/v) agarose gel was prepared with 1x TBE and 3x GelRed™ to validate whether or not there was specific DNA amplification. 30% loading buffer was used to load 1µL of the amplified DNA sample, which was run alongside 1µL 1000bp Hyperladder IV ladder. Electrophoresis was run using a 1x TBE running buffer at 100 volts for approximately 35 minutes. The gel was then removed from the electrophoresis tank and placed onto a UV-illuminator and a photograph of the fluorescence was taken using an Olympus camedia C-5060 wide zoom camera and analysed using the software package Alphadigidoc RT.

2.4.3 Sanger sequencing of the PCR-amplified DNA fragments

Six primers were selected to sequence the ±1,2kb fragment of the control region (Table 2.3 and Figure 2.2), the TM was determined using the nearest neighbour method on Oligo v7.37 (Rychlik, 2007). The 130µL samples were aliquoted into 5 equal volume (26µL) aliquots for further purification and sequencing using BigDye terminator v3.1. The sequenced DNA was then run on an ABI 3100 genetic analyzer (Macrogen Inc.) with expected 2x coverage from a forward and reverse sequence read between 16024bp -576bp (HV I, II and II).

Table 2.3 Tm and sequences of primers used in Sanger sequencing of the control region in the mitochondrial genome. Primers designed by D'Amato (unpublished).

Primer name	Primer sequence	TM Nearest neighbour
L15969	CTT TAA CTC CAC CAT TAG CAC C	62.9
H16509	AGG AAC CAG ATG TCG GAT ACA G	65.9
L15	CAC CCT ATT AAC CAC TCA CGG	63.9
H185	CCT GTA ATA TTG AAC GTA GGT GC	63.1
H658	CCC CAT AAA CAA ATA GGT TTG G	63.2

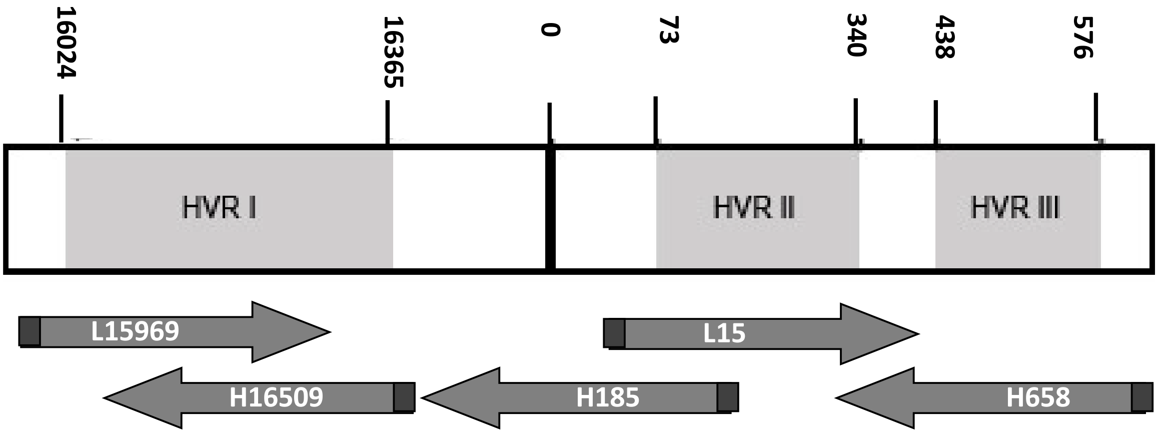


Figure 2.2 shows a section of the mitochondrial DNA genome from position 15969 - 658 which contains the control region. Within the fragment are the hypervariable regions I, II and III (shaded in light grey) with their respective start/end positions indicated above. The arrows below the fragment show the positions (dark base) and direction of the sequencing primers.

2.4.3.1. Troubleshooting of sequencing primers

The original terminal reverse primer (H658) utilised for amplification and sequencing of the mtDNA control region yielded undesirable sequencing results. The forward primer L15 produces a clean read until 600bp, but the reverse primer H658 produces a truncated sequence of 100bp or less and in some cases there is no observable sequencing data. Forward and reverse sequence data is required in order for a sample to be included in haplogroup assignment. Possible SNP polymorphisms for haplogroup L0a occur at C411G, T413G, 498d, 522d, 523d and therefore coverage of a reverse primer from nucleotide site 400-600 of the control region is required to confirm haplogroup assignment.

Alignment of possible alternative reverse primers was done using Bioedit V7.2.5 (Ibis Biosciences 2013) and compared with sequence alignments of confirmed L0 haplogroups (Genbank). The average coverage of the forward primers L15 and L15969 was also taken into consideration. Four alternate end primers (table 2.4) were tested and optimised for compatibility with the forward primers L15 and L15969 in amplification as well as sequencing reactions. The products were then run at the Central Analytical Facility (CAF) Stellenbosch University.

Table 2.4 Tm and sequences for the alternate primers used in Sanger sequencing of the control region (400-600bp area) in the mitochondrial genome. Reverse Primers in red. H605 (unpublished) H599 and H612 (based on literature).

	Primer name	Primer sequence	TM Nearest neighbour
New primers	H599	TAT GTA GCT TAC CTC CTC AA	54.7
	H605	GTA GCT TAC CTC CTC AAA GCA AT	61.9
	H612	ACC TCC TCA AAG CAA TAC ACT	59.8
Old primers	H658	CCC CAT AAA CAA ATA GGT TTG G	63.2
	L15	CAC CCT ATT AAC CAC TCA CGG	63.9
	L15969	CTT TAA CTC CAC CAT TAG CAC C	62.9

2.4.4. mtDNA analysis methods

Sequences were received in .ab1, .seq and pdf format so that the sequencing chromatogram could be checked for success and any ambiguities dealt with using Chromas Lite V2.1.1 (Technelysium Pty 2013). The .ab1 files were then aligned with the revised Cambridge reference sequence (RCRS) using the ClustalW algorithm (Larkin *et al.* 2007) on BioEdit V7.2.5 (Ibis Biosciences 2013). Alignment of the two reverse (H185 and H16509) and two forward (L15 and L15969) sequences were combined to form a consensus sequence for each individual in the population. Nucleotide changes (mutations) between the query sequence and reference were recorded according to guidelines set out by the DNA commission of the International Society for Forensic Genetics (ISFG, 2000). The recorded mutations were then analysed using Haplogrep (http://haplogrep.uibk.ac.at/ based on Phylotree build 16, 2014), a web-based haplogroup assignment program. A comparison was done on the haplogroup constituent of each of the sample groups: origin and post-colonial. Evaluation of the haplotype structure for the population was done using AMOVA statistical analysis of the sequence data Arlequin v3.5.1.2 © software (Excoffier *et al*. 2004). Conventional F-statistics of standard AMOVA computations was done using 1000 permutations. A Network 4.6.1.0 (www.fluxus-engineering.com/ 2014) analysis of the L0 haplogroups for the entire population was then done. For the Network analysis, the epsilon value was set to zero, transversions were weighted 3x higher than transitions, the hypervariable indels at position 16189, 305.1C and 315.1C were excluded following guidelines from Phylotree (www.phylotree.org/ 2014).

2.5. Y chromosome protocols

2.5.1 Y-STR multiplex design

The UWC 10plex is a Y chromosome STR multiplex that was designed to discriminate between male individuals with African ancestry. The validity of the multiplex was tested on ninety one Griqua males. The primers for the UWC 10-plex were designed by D'Amato *et al*. (2011) based on previous studies. D'Amato *et al.* (2011) have optimised PCR cycling conditions for the UWC-10-plex. Loci are organised in the multiplex based on allele size range and dye colour. Table 2.5 displays the 10 Y-STR markers along with their fluorescent dye colour, allele range, repeat sequence and size range.

Table 2.5 Allele ranges and PCR product sizes for 10 Y-STR markers with information gained from D'Amato *et al.* (2011).

Dye colour	STR locus	Allele range	Repeat motif	Size range (bp)
PET	DYS481	18-28	CTT	110-137
	DYS447	19-29	TAATA	192-242
	DYS449	24-36	$(TTTC)_nN_{50}(TTTC)_n$	270-322
NED	DYS612	19-33	$(TCT)_4(CCT)_1(TCT)_4(CTT)_1(TCT)_n$	151-192
	DYS626	12-23	$(GAAA)_n$	232-276
	DYS504	11-18	$(CTTC)_n$	304-332
6-FAM	DYS710	28.2-38.2	$(AAAG)_n(AG)_n(AAAG)_n$	174-212
	DYS518	25-36	$(AAAG)_nGGAG(AAAG)_3N_6(AAAG)_n$	262-300
VIC	DYS385	07-24	$(GAAA)_n$	237-303
	DYS644	12-26.4	$(TTTTA)_nTTTTTA(TTTTA)_1$	316-391

2.5.2. Polymerase Chain Reaction (PCR)

The multiplex was performed to a final volume of 10µl; with the UWC 10plex consisting of 2ng of male genomic DNA, 10x Super-Therm PCR buffer (with 15mM $MgCl_2$) (*Southern Cross Biotechnology*), 2mM dNTPs (*Roche*) and 5U/µl Super-Therm Gold DNA Polymerase (*Southern Cross Biotechnology*). The primer sequences with corresponding size ranges, fluorescent dye labels and final concentration are illustrated in table 2.6.

Table 2.6 The primer sequences for the UWC 10plex with corresponding size ranges, fluorescent dye labels and the final primer concentrations.

Y-STR	Primer sequences	Primer concentrations
DYS481	F: GAA TGT GGC TAA CGC TGT TC R: **PET** - TCA CCA GAA GGT TGC AAG AC	0.3µM
DYS447	F: GGG CTT GCT TTG CGT TAT CTC T R: **PET** - GGT CAC AGC ATG GCT TGG TT	0.72µM
DYS449	F: **PET** - GAA TAT TTT CCC TTA ACT TGT GTG R: CAC TCT AGG TTG GAC AAC AAG AG	0.72µM
DYS612	F: GAA GTT TCA CAC AGG TTC AGA GG R: **NED** - AAA AAG GGA ACT GAG GGA AGG	0.102µM
DYS626	F: GCA AGA CCC CAT AGC AAA AG R: **NED** - AAG AAG AAT TTT GGG ACA TGT TT	0.228µM
DYS504	F: **NED** - CTA AGC TGC AAG AAA AAG TCC R: GAA TCA CTT GAA CCC AAG ATG	0.168µM
DYS710	F: ACT TTT CTG AAT CCT GGA CAA GTG R: **6-FAM** - TTC CTC ATA CTC TCT CCC TCC C	0.3µM
DYS518	F: **6-FAM** - CAC AAG TGA AAC TGC TTC TCG R: CAT CTT CAG CTC TTA CCA TGG	0.192µM
DYS385	F: **VIC** - AGC ATG GGT GAC AGA GCT A R: GCC AAT TAC ATA GTC CTC CTT TC	0.24µM
DYS644	F: **VIC** - CAG GAG ACT GAG GCA GAA AGT C R: GGA AGA AGC TGA TTT CAA TCT CC	0.145µM

Amplification of the multiplex was performed in the Veriti 96 well thermal cycler (*Applied Biosystems*). The UWC 10plex is split into eight stages with stages 2-5 consisting of 2 cycles and stage seven of 24 cycles. The cycling parameters are as follows: denaturation for 10 minutes at 95°C; followed by 2 cycles of 94°C for 30 seconds, 66°C for 1 minute, 72°C for 1 minute; 2 cycles of 94°C for 30 seconds, 65.5°C for 1 minute, 72°C for 1 minute; 2 cycles of 95°C for 30 seconds, 65°C for 1 minute, 72°C for 1 minute; 2 cycles of 95°C for 30 seconds, 64.5°C for 1 minute, 72°C for 1 minute; then 95°C for 30 seconds, 64°C for 1 minute, 72°C for 1 minute; then 24 cycles of 95°C for 30 seconds, 62°C for 1 minute, 72°C for 1 minute; adenylation step for 75 minutes at 68°C and a holding temperature of 4°C.

2.5.3. Genotyping

The amplified DNA fragments of all 10 markers were analyzed using an ABI3500 Genetic Analyzer (*Applied Biosystems*). A loading mixture was prepared containing 8.7µl of Hi-Di™ formamide (*Applied Biosystems*) and 0.3µl of GS500 LIZ size standard (*Applied Biosystems*). This was added to each well of the plate along with 1µl of PCR product. Allelic ladder was added to each run to account for migrational variation between runs. Two male positive controls was included in each run to measure success of the amplification. The plate was closed with a septa, briefly spun down (Centrifuge 5414 D) and denatured on a 2720 Thermal Cycler (*Applied Biosystems*) for 5 minutes at 95°C. After denaturation it was snap-cooled on ice and loaded onto the ABI3500 Genetic Analyzer (*Applied Biosystems*). It was separated by the capillary array as per the manufacturer's specifications and the DNA sequences collected with the Genemapper-IDX software (*Applied Biosystems*). The size standard was used to size the DNA fragments for visualisation on an electropherogram. All the samples were designated alleles according to the allelic ladder. Markers are organised according to PCR product size range and dye colour as seen in figure 2.3.

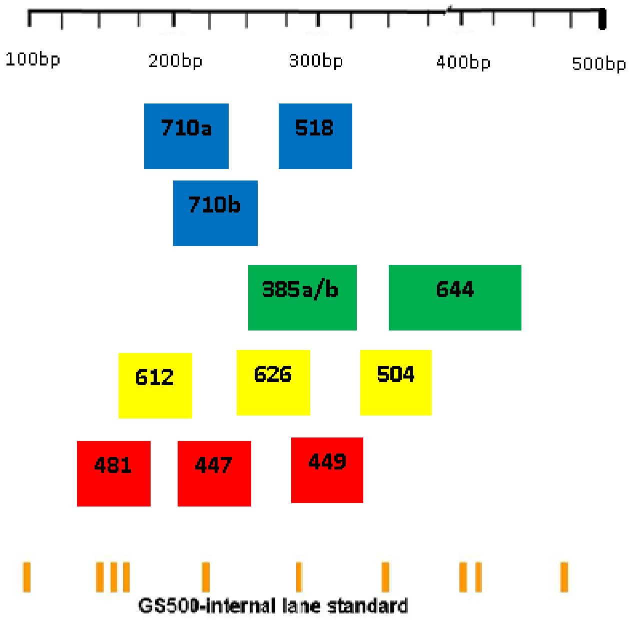


Figure 2.3 Schematic outline of the UWC10 plex allele size ranges for each locus and the corresponding dyes used to label the primers. The bar on top indicates the size range in base pairs of PCR products. Locus DYS710 produces two products: DYS710a and DYS710b, which differ in size by 22 bp. Locus DYS385 produces two different products with allele sizes within the range.

2.5.4. Data analysis and interpretation

The Genemapper-IDX data collection software (*Applied Biosystems*) uses the multi-component matrix DS-33 for dye set G5 to automatically analyze the five different coloured fluorescent dye-labelled samples in a single capillary. Each marker range was assigned according to the macros with a reference allele or allelic ladder being run with each plate to account for any shift that might occur due to changes in running conditions (temperature effects). The ladder used to construct the bin set and panel for the UWC 10plex is displayed in table 2.7. Four controls need to be manually inspected before a run can be approved. First the internal lane standard (GS500 LIZ, *Applied Biosystems*) needs to conform to the guidelines outlined by the manufacturer in terms of migration rate and morphology of the fragments. Then the allelic ladder needs to be assessed to confirm success of the capillary electrophoresis run and to allow for accurate allele designation. The PCR positive controls are then analysed to assess the quality of the amplification reagents and instrument. Finally the PCR negative control provides confidence in the process and allows detection of any cross-contamination.

Table 2.7 Allelic ladder size ranges and repeat range for the 10 markers in the UWC 10plex.

Marker	Dye colour	Allele size range (bp)	Allele repeat range
DYS481	PET	109-139	19-30
DYS447	PET	191-241	19-30
DYS449	PET	270-318	24-36
DYS612	NED	153-195	20-34
DYS626	NED	229-273	22-32
DYS504	NED	301-329	11-18
DYS710	6-FAM	179-221	28.2-39
DYS518	6-FAM	247-291	31-40
DYS385a/b	VIC	236-292	7-21
DYS644	VIC	318-396	12-26.4

2.5.5. Statistical analysis

Genepop v4.1.4 © software (Rousset *et al*. 1995) and Arlequin v3.5.1.2 © software (Excoffier *et al*. 2004) were used to perform statistical analysis of the data. Genepop v4.1.4 © software (Rousset *et al*. 1995) was used to gain a basic description of the 10 loci and the allele frequencies of each locus. Arlequin v3.5.1.2 © (Excoffier *et al*. 2004) was used to calculate haplotype diversity and the population structure differentiation. Haplotype frequency for the sample group was calculated by simple counting of observed typing results. A unique haplotype is defined as one that occurs only once in a given population. STR genetic diversity or the probability that two alleles chosen at random are different, was calculated using the formula: GD= $(N/N\text{-}1)(1\text{-}\sum p_i^2)$ where N is the sample size and p_i is the relative allelic frequency. Haplotype diversity (HD) was computed using the same equation by substituting haplotype frequencies for allelic frequencies. The Discrimination Capacity (DC) was determined by dividing the number of different haplotypes seen in a given population by the total number of samples in that population. Arlequin software v3.5.1.2 (Excoffier *et al*. 2004) was used to calculate haplotype frequencies and perform analysis of molecular variance (AMOVA) tests of population heterogeneity among the two populations excluding the DYS385 marker.

Chapter 3: Results and discussion

3.1. Maternal ancestry analysis

3.1.1. DNA quantitation results

Common DNA concentration values of DNA extracted from buccal swabs are around 20ng/µl (Feigelson *et al.* 2001). The average concentration of the DNA extracted from the total 104 Griqua saliva samples was 74.4ng/µl in 50µl of 1xTE buffer. Although this is a favourable mean, the values ranged from 2.13ng/µl to a substantial 820.6ng/µl. The absorbance values had to be checked to ensure the purity of the samples. It was found that even though some samples had seemingly high concentrations of DNA extracted, they were impure and yielded bad results after PCR amplification. Nucleic acids and proteins have absorbance maxima at 260nm and 280nm respectively. The ratio of absorbances at these wavelengths is used as a measure of purity after DNA extraction. A 260/280 ratio of ~ 1.8 is generally accepted as pure for DNA. Expected A260/A230 ratios fall within the 2.0 -2.2 range and absorbance at 230nm is indicative of the presence of contaminants. A positive indicator of low contamination levels is the A260/A230 (±2.0) levels being higher than the A260/A280 (±1.8).

A few samples proved challenging to amplify (GRI-04, -15, -20, -28, -29, -34, -44, V14, VGRI-43, -50, -59, -156 and -160). Some of the samples had A260/A280 ratios above 2.0 and A260/A230 levels below 2.0 which are indicative of the presence of RNA and NaCl contamination respectively. These samples may have been degraded over time or contained PCR inhibitors that made them very challenging to amplify. Re-extraction from saliva samples using an extra phenol:chloroform clean-up step did not yield a higher concentration of DNA without inhibitors. A troubleshooting step of adding more DNA in the reactions (4ng instead of 2ng) did not resolve the problem of amplification of the degraded samples. As a result the 13 degraded samples were excluded from the final statistical analysis step along with any partial profiles generated. For the 176 samples analysed, the GRI and V lab codes designate samples from Kokstad and Vredendal respectively and VGRI designates samples collected from the GNC festival in 2013.

3.1.2. Amplification, verification and sequencing of the control region

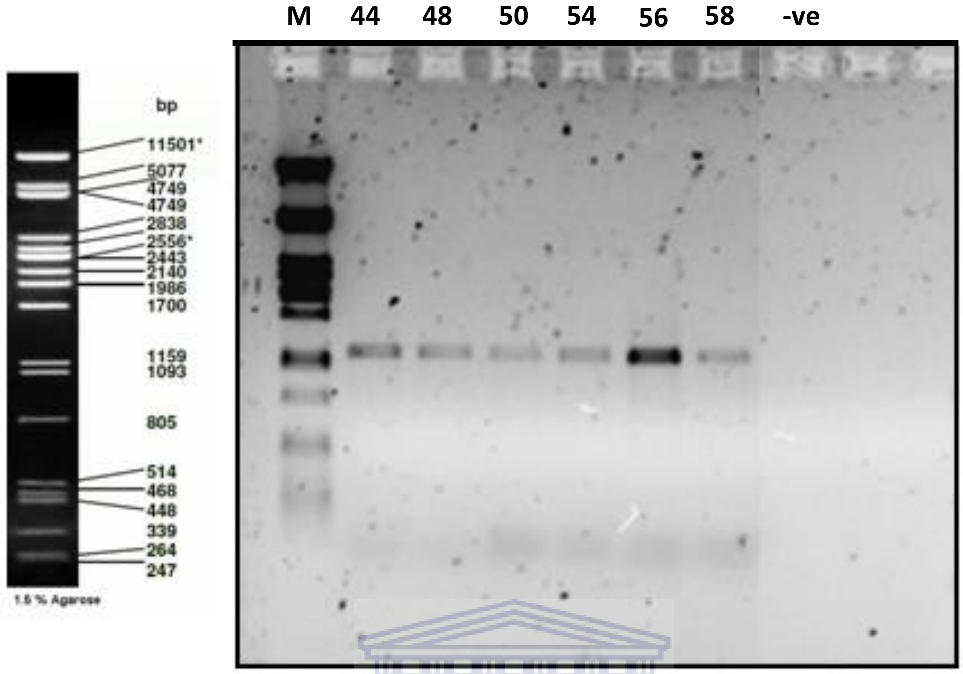


Figure 3.1 Photograph of amplified mtDNA from position 15969 to 658 (1258 bases). The fragments were separated on a 2% (w/v) agarose gel in a 1xTBE buffer. The lane marked M represents the phage λ marker with a range of 247 – 11501bp. Lanes 2-7 represent various samples and lane 8 is the negative control.

The DNA fragments of 176 unrelated Griqua individuals sampled were successfully amplified and sequenced (3x coverage) from positions 16024 – 400 of the mitochondrial control region. Polymorphisms were identified by comparing each of the samples to the RCRS. The haplogroup of each sample was inferred by using the web-based program HaploGrep (Kloss-Brandstaetter *et al.,* 2010). HVI and HVII are commonly utilised to infer ancestry using mtDNA and the focus is on mutations that are informative of the geographic origin of a population (Schlebusch *et al.* 2012; Quintana-Murci *et al.* 2010). The discussion below will deal with the problems surrounding sequencing of HVIII and analysis of haplogroup assignment within the two sample groups (Origin and Post-colonial).

3.1.2.1. Troubleshooting of sequencing primers

Initial sequencing experiments with primers L15969, H16509, H185, L15 and H658 resulted in four of the five primers providing adequate coverage of Hypervariable segments (HV) I and II of the mtDNA control region. Adequate coverage means that there is a forward and reverse primer that has produced a sequence read in both directions, these overlapping sequences can be combined to produce a consensus sequence. HVI extends from 16024bp to 16365bp and HVII from 73bp to 340bp in the mtDNA control region. It is acceptable practice to sequence only polymorphisms in these two hypervariable segments in order to assign a haplogroup to an individual (Wakeley, 1993; Stoneking, 2000 and Chong, 2004). In the initial sequencing results of this project it was found that primer L15 produces long, clean reads up until the end of the control region. This primer results in readable sequences from nucleotide base 15 of the control region up until about 650 bases. In order to cover any polymorphisms that might fall within the range of HVIII (438 – 576bp) a sequence read from a reverse primer is required.

Haplogroup assignment is possible using HVI and HVII polymorphisms, although incorporation of coding region polymorphisms increases the confidence in results (Salas *et al.* 2005). Common polymorphic tracts in HVIII such as the CA deletion at nucleotide location 522-523 are disregarded when assigning haplogroups and assembling phylogenetic trees. Preliminary sequencing results using HVI and II showed a large contingent of L0a and L0d haplogroups in the Griqua population. A read in the reverse direction from the end of the control region to nucleotide position 400 on the RCRS provides sufficient coverage of HVI & II polymorphisms. A graphical representation of this is provided in figures 3.2 where the problematic amplification zone covers two transversion mutation points C411G and T413G as well as three deletions 498d, 522d, 523d in HVIII which are not diagnostic for any haplogroups. The position of three alternative reverse end sequencing primers H599, H605 and H612 are indicated in figure 3.2.

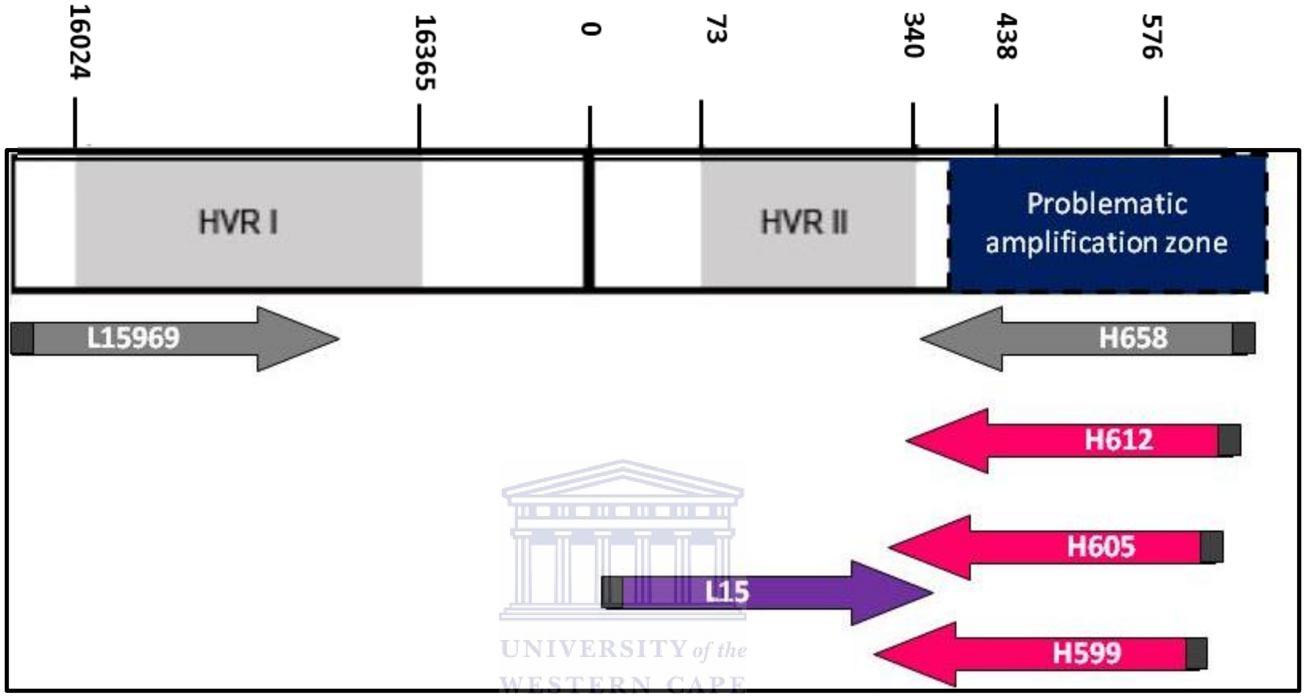


Figure 3.2 shows a section of the mitochondrial DNA genome from position 15969 – 658 which contains the control region. A problematic amplification zone was identified between 400 – 600bp following initial sequencing using the grey primers. Three alternate reverse primers were tested (pink) along with the forward primer L15 (purple) in order to resolve the problem.

3.1.2.2. Possible explanation for poor sequencing results

Initial sequencing experiments using L15969, H16509, L15, H185 and H658 to sequence the mtDNA control region yielded good reads from three of the four primers. Figure 3.3 below displays a successful sequencing read using forward primer L15 for sample VGRI_44 while figure 3.4 displays the unsuccessful sequencing read when using primer H658 for the same sample.

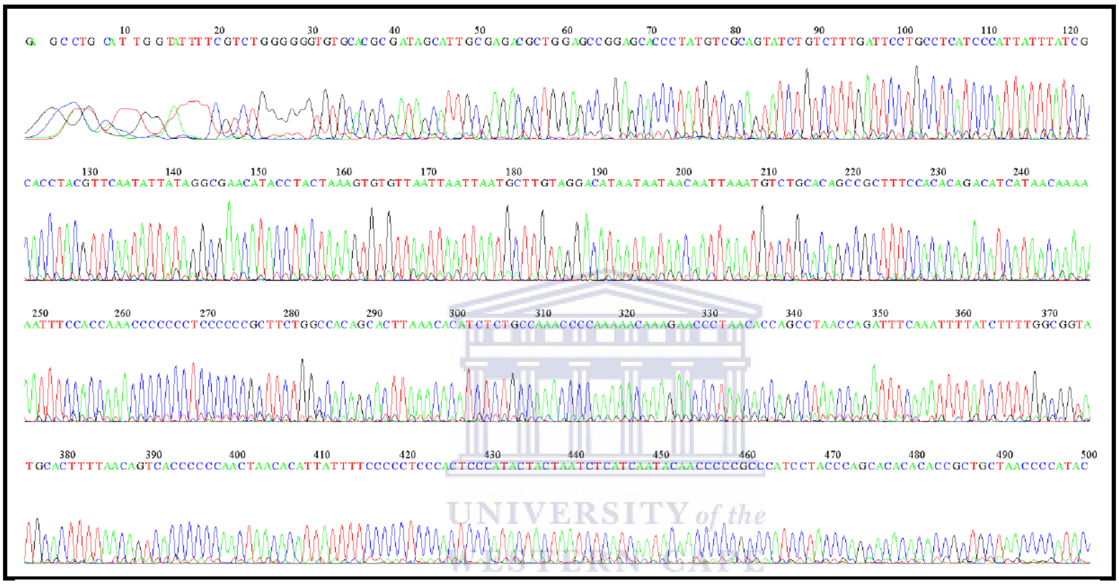


Figure 3.3 Chromatogram displaying a successful sequencing read using forward primer L15 with good peak heights and a read which covers the homopolymeric tract at 260 bases.

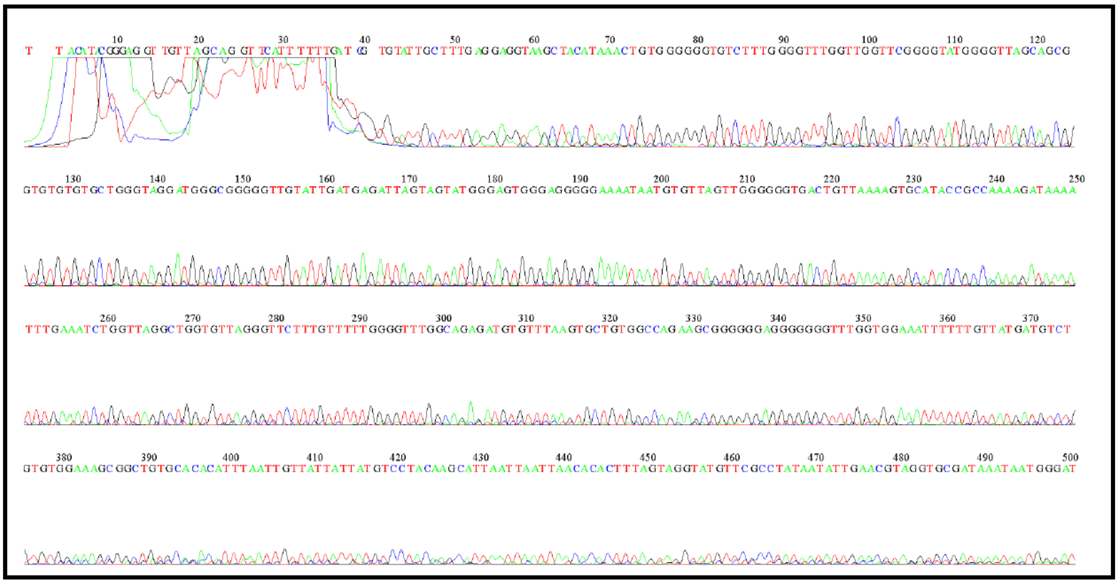


Figure 3.4 Chromatogram displaying an unsuccessful sequencing read using reverse primer H658. With low peak height and no readable sequencing results to compare with the forward sequence

When using reverse primer H658 (figure 3.4) there is a low peak height, below 100 relative fluorescent units (RFU's), and poor signal-to-noise ratio causing unreadable sequence data. There are various causes for this pattern of peaks in a sequencing read. It could be due to too little DNA template being present, insufficient primer-template binding or even bad purification techniques prior to sequencing (Kieleczawa and Mazaika, 2010). Since the the same amplified product was used for each of the five sequencing primers, the issue probably lies with the primer rather than the process. The next viable step was to try alternate reverse end sequencing primers to overcome this poor seqeuncing quality (Carracedo, 2000). Figure 3.5 shows three seperate sequencing reactions utilising forward primer L15 with three different reverse primers.

Panel 3.5a shows a retesting of primer H658. Although the sequencing signal remains strong, there are multiple overlaying sequences present. It is caused by multiple DNA fragments of different size migrating on top of each other during electrophoresis. This phenomenon is thought to be caused by regions of secondary structure within the template DNA, which are often found in regions of high G/C or A/T content (Reuter and Mathews, 2005). The best solution is to choose a new primer close to the compression site which can help avoid the effects of the secondary structure.

Panel 3.5b shows a failed sequencing reaction using commercial reverse primer H599 (EDNAP) and forward primer L15. There is a lack of sequence data and the sequencing trace is non-uniform. Unincorporated Dye Terminator peaks at the start of the sequence are out of scale and basecalls contain many ambiguous bases 'N'. This is caused by the primer not annealing to the template or the absence of DNA template in the sequencing reaction. Panel 3.5c shows a seqeuncing run with a secondary structure present when using reverse primer H612 (in-house design) and forward primer L15. There is a sharp drop or premature termination of sequencing signal during the sequencing run. Secondary structure in the template can cause various anomolies that result in inefficient chain elongation after the region of the secondary structure. Many of the same techniques used to eliminate compression can be helpful in obtaining good sequence data past regions of secondary structure.

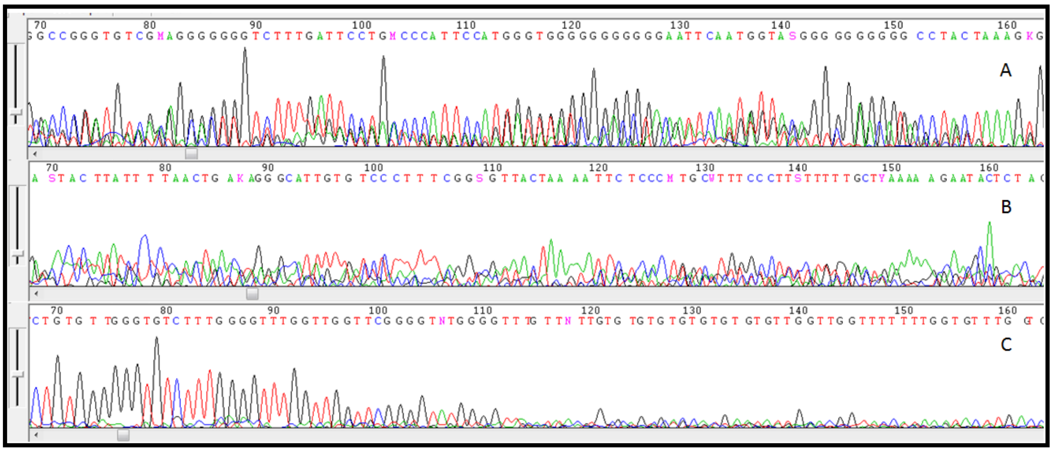


Figure 3.5 Problematic electropherograms arising when sequencing with end primers. Panel A, B and C show results from sequencing reactions using H658, H599 and H612 respectively.

In silico secondary structure prediction was done using a web-based structure program by Reuter and Mathews (2010) accesssed at http://rna.urmc.rochester.edu/RNAstructureWeb/Predict1.html. This was done due to the prevelance of secondary structure related sequencing problems encountered (compression and drop off). All troubleshooting that follows in this section is thus based on the hypothesis that secondary dtructure formation has affected the efficiecny of the sequencing reaction. The secondary structure prediction software package utilises the most recent set of nearest neighbour parameters in its thermodynamic algorithms for secondary structure prediction.

Figure 3.6 shows all possible hairpin and stem-and-loop secondary structures that can form at 60°C between 550 – 659bp on the mtDNA control region which also encompases the end of HVIII (576bp). The homopolymeric C stretch and the CA stretch at 520bp have proved a tricky region to resolve when sequencing in the reverse direction as they fall near the start of the sequencing read. Primers need to be ideally located so as to avoid the majority of the hairpin loops (6 present in this region) and dimers that can form under sub-optimal annealing conditions. Of course zero or low self-dimerization and primer dimer formation is expected from any viable primer. Such parameters can be checked on the Oligo 7 primer design program. A web-based multiplex oligo calculator that utilises modified nearest-neighbour method (Breslauer *et al.* 1986) also proves to be a useful tool for designing multiplex PCR cycling conditions. (http://www.thermoscientificbio.com/webtools/multipleprimer/)

If all of these primer paramaters are taken into account, the most viable option as a reverse primer for this sequencing reaction is H605.

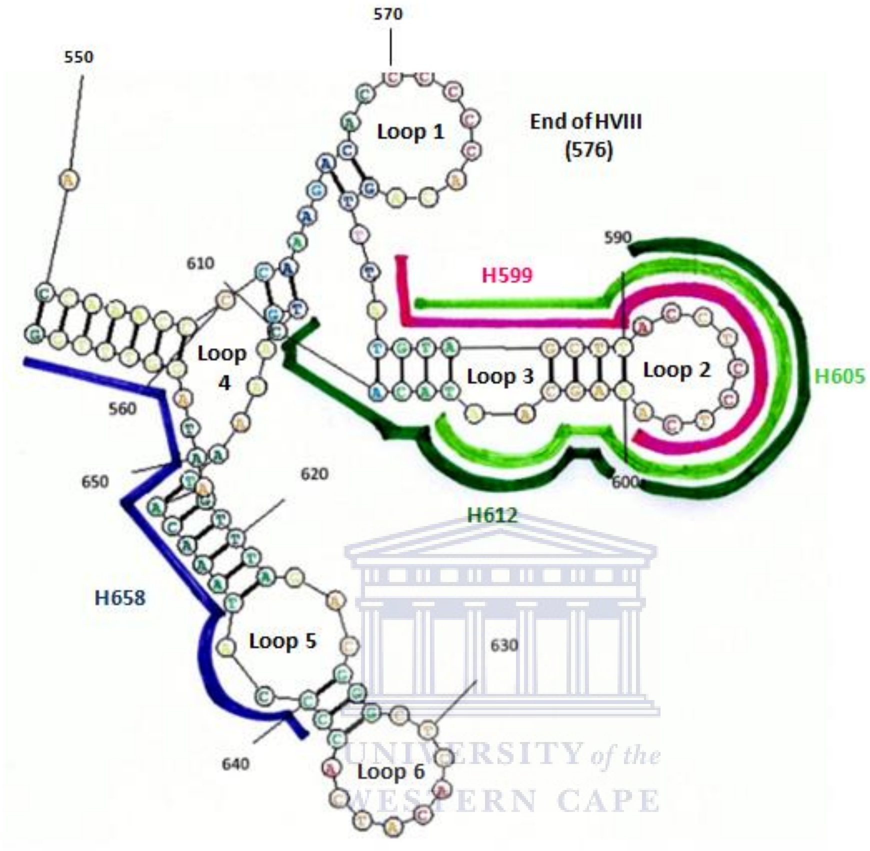


Figure 3.6 A figure generated using a secondary structure prediction programme accessible at (http://rna.urmc.rochester.edu/RNAstructureWeb/Servers/Predict1/Predict1.html) showing possible hairpin loop formation that may occur during amplification using the four alternate reverse primers.

The use of a new reverse primer site that avoids the majority of possible hairpin and stem-loop formation sites improves the sequencing read, as seen in figure 3.7. Amplification and sequencing using a combination of primer H605 and L15 resulted in higher RFU's than previous sequencing runs, there is also a fairly consistant read with no signal dropoff due to secondary structure formation. Future tests with this primer combination are necessary in order for it to be accepted as the standard, end, reverse primer for control region amplification in the Forensic DNA Lab at UWC.

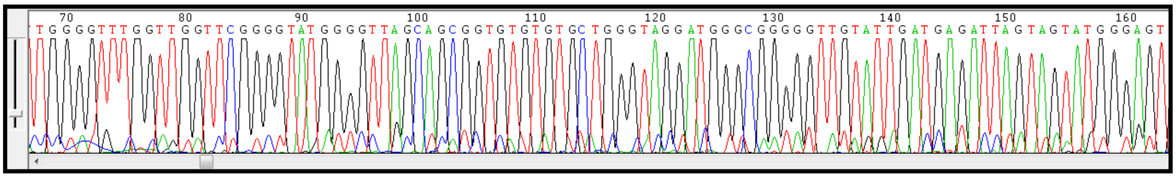

70
80
90
100
110
120
130
140
150
160
TGGGGTTTGGTTGGTTCGGGGTATGGGGTTAGCAGCGGTGTGTGTGCTGGGTAGGATGGGCGGGGGTTGTATTGATGAGATTAGTAGTATGGGAGT

Figure 3.7. Final sequencing read using H605 reverse primer. High RFU's and low signal-to-noise ratio are shown.

3.1.2.3. Final coverage obtained during sequencing

The final coverage of all 176 unrelated Griqua samples is outlined in table 3.1. The samples with the GRI_ and VGRI001 – VGRI031 assigned lab numbers were sequenced using the primers L15969, H16509, L15, H185 and H605. Full coverage of HVI (16024 – 16365bp), II (73 – 340bp) and III (438 – 576bp) were obtained by using this combination of forward and reverse primers. Full coverage means that the consensus sequence covers the entire area. A consensus sequence is constructed using two or more overlapping sequence reads (2x) in either direction. The average coverage of a consensus sequence for these samples is from 16024bp – 600bp. All polymorphisms up to 600bp were included when assigning haplogroups using the Haplogrep software (http://haplogrep.uibk.ac/ based on Phylotree build 16, 2014).

The samples with VGRI assigned lab numbers were sequenced using primers L15969, H16509, L15, H185 and H658. Full coverage of HVI was obtained from both the forward and reverse primers. Due to the failure of the reverse sequencing primer H658 there is only single coverage from 185bp. This is in the forward direction from primer L15 and confirmation from a reverse read primer is required, as previously discussed. Due to time constraints this could not be completed and the polymorphisms recorded from the L15 primer have been taken into account for haplogroup assignment using the Haplogrep software (http://haplogrep.uibk.ac/ based on Phylotree build 16, 2014). The mutations that occur after 185bp have been recorded, but require further consensus using reverse primer H605. They are identified in table 3.1 as "consensus required" samples. The average coverage of a consensus sequence (3x) for these samples is 16024bp – 150bp although single coverage (1x) from L15 extends to 600bp. All 176 Griqua samples were included in further data analysis.

3.1.3. Population genetics

3.1.3.1. Haplogroup assignment

Sequences were received in .ab1, .seq and .pdf format so that the sequencing chromatogram could be checked for success and any ambiguities dealt with using Chromas Lite V2.1.1 (Technelysium Pty 2013). The .ab1 files were then aligned with the revised Cambridge reference sequence (RCRS) using the ClustalW algorithm (Larkin *et al.* 2007) on BioEdit V7.2.5 (Ibis Biosciences 2013). Alignment of the three reverse (H185, H605 and H16509) and two forward (L15 and L15969) sequence reads were combined to form a consensus sequence for each individual in the population. Nucleotide changes (mutations) between the query sequence and reference were recorded according to guidelines set out by the DNA commission of the International Society for Forensic Genetics (ISFG; 2000). The recorded mutations were then analysed using Haplogrep (http://haplogrep.uibk.ac.at/ based on Phylotree build 16, 2014), a web-based haplogroup assignment program. A comparison was done on the haplogroup constituent of each of the sample groups: origin and post-colonial shown in figures 3.8a and b.

3.1.3.2 Analysis of Griqua population dynamics

The population is divided into an origin sample group and post-colonial sample group based on the place of birth of each individual. Geographically isolated populations along the west coast, have less exposure to early colonisation events that have affected the more inland populations. As described in the population background (chapter1 section 1), the Griqua population was exposed to various native (Bantu, Zulu and Xhosa) and colonial (European and Asian) genetic influences from the 18th century up until the apartheid regime.

A larger incorporation of the Khoisan haplogroup, L0d, was expected in the origin sample group which includes individuals whose ancestors were not part of the trek events and have been isolated along areas of the West Coast and Northern Cape. The lack of colonisation events affecting this origin sample group means that it represents the ancestral state for the Griqua population. In South Africa there are several small towns which have been isolated from the general population over extended periods. These isolated towns provide the most realistic way of tracing maternal ancestry back to the origin populations of South Africa, back to a time when the land was solely inhabited by the pastoralist (Khoe) and foraging (San) indigenous populations (Schlebusch *et al.* 2013).

A greater incorporation of other African maternal lineages (L0a, L1, L2, L3, L4 and L5) is expected in the post-colonial sample group. These clades (L1, L2 and L3) are thought to have been introduced to South Africa during the recent Bantu expansion (1000 – 500YBP). Since these lineages are also observed in the Khoisan (~40%) they should be seen as lineages of pan-African origin (Quintana-Murci *et al.* 2010). Out of Africa due to colonial influences (M and N sub groups) was expected in the post-colonial sample group. This group includes individuals whose ancestors participated in the Griqua trek in South Africa.

Table 3.2 mtDNA haplogroup assignment for the Griqua population of South Africa.

	Haplogroup	**Origin sample group (N = 82)**	**Post-colonial sample group (N = 94)**
Khoisan	**L0d**	63	52
	L0a	3	3
Pan - African	**L1-5**	8	18
Outside Africa	**M**	2	6
	E	0	1
	H	1	8
	N	1	3
	J	1	0
	U	3	0
	A	0	1
	B	0	2

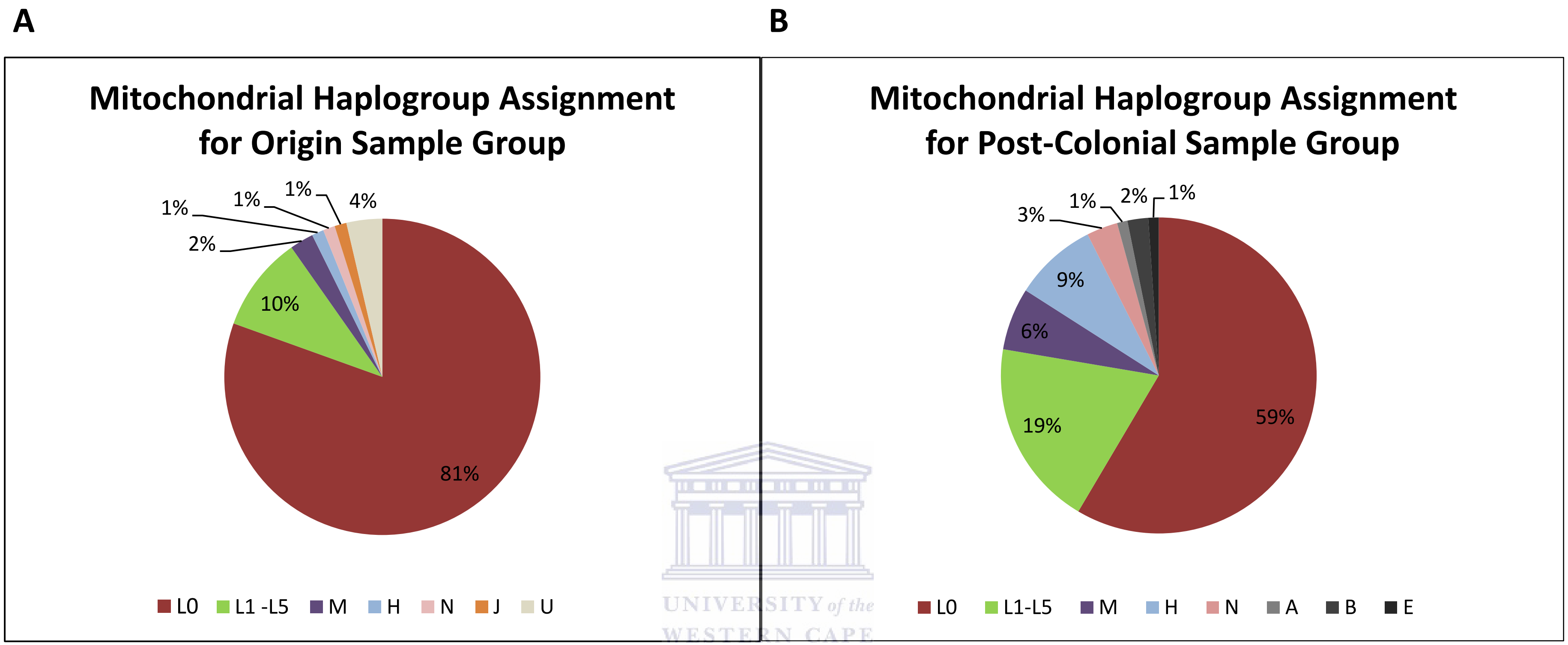


Figure 3.8 Two pie charts labelled A and B comparing the origin sample group and the post-colonial sample group respectively. The origin sample group contains 91% African ancestry (L0-L5) with 76.83% L0d Khoisan ancestry. Whereas the post-colonial sample group is made up of 78% African ancestry (L0-L5) and less than 60% L0d Khoisan ancestry.

Population statistics for each of these sample groups support the initial hypothesis of a greater component of ancient African ancestry in the origin sample group. Haplogroup L0d is found at low frequencies in Southern-African Bantu speakers, and is a useful marker to estimate the extent of maternal gene flow from the Khoisan (Schlebusch *et al.* 2011). The most predominant haplogroup in this sample group is L0d1b (33% of sample group) and the least prominent, a haplogroup usually found in Northern San populations (!Xun and Khwe) is L0d1c (1%). The incorporation of ancient African haplogroups other than L0d was limited to L0a from South East African populations (3.7% of sample group). The origin sample group consists of 2% macro-haplogroup M and 5% macro-haplogroup N and its sub-clades (H, J, N and U). Haplogroups J and H originated in East Eurasia. Haplogroup H is the most common mtDNA haplogroup found in approximately 41% of native Europeans (Achilli *et al.* 2004; Quintana-Murci *et al.* 2004). Haplogroup U is the oldest maternal haplogroup found in native Europeans with sub-haplogroups of U being widely distributed across Western Eurasia, North Africa and South Asia. The overall frequency of U in South Asia is largely accounted for by the haplogroup U2 in India (Metspalu *et al.* 2004). Possible incorporation of this haplogroup into Southern African population groups might be through the incorporation of slaves from India since the Dutch East Indian trading company used the Cape of Good Hope as one of its ports along the spice route.

The influence of African haplogroups related to the Bantu expansion was most predominant in haplogroup L2a1a2. There are two L2a clusters that are well represented in south-eastern Africans, L2a1a and L2a1b, both defined by transitions at quite stable HVI positions (Coelho *et al.* 2009). Subclade L2a1a is further refined by coding region substitutions at 3918, 5285, 15244, and 15629 (www.phylotree.org/ 2014). Both of these sub clades have an origin in West Africa or North West Africa and appear to have undergone dramatic expansion either in South East Africa or in an ancestral population to south-eastern Africans (Torroni *et al.* 2001). The presence of these lineages in the Griqua population reflects either direct gene flow from Bantu peoples or indirect Bantu contribution via admixture with the Khoisan. The Khoisan received these lineages through their prior admixture with Bantus (Quintana-Murci *et al.* 2010).

In comparison the post-colonial sample group displayed a more heterogeneous origin of mtDNA types. The most predominant haplogroup in this sample group is L0d2a (21.3%) and out of Africa ancestry (H, M, N, B, A and E) contributes 22.3% of the haplogroups in this sample group. Macrohaplogroup N includes sub haplogroups A and B, while Macrohaplogroup M includes sub haplogroup E (van Oven and Kayser 2009). Haplogroup A4 has Eastern Asian ancestry (Tanaka *et al.* 2004) and haplogroup B is frequently found in South-eastern Asia (Mona *et al.* 2009). There is no incorporation of these haplogroups in the origin sample set. Only these three individuals have this Eastern Asian ancestry in the entire Griqua population group. There is also only one individual with haplogroup E1 maternal ancestry which may have been incorporated due to the Malay influences in South Africa during the early 1800s.

It is of importance to note that Haplogroup H15 (T55C, T57C, A263G, 309.1C and 315.1C) is shared between seven individuals from this sample group. mtDNA haplogroup H15 is a very rare, highly dispersed clade ranging from Western Europe to the Middle East and from Central Asia to India. Genbank submissions of samples falling within this subclade (*e.g*. KC911292, H15) exhibit the following core mutations from the Cambridge Reference Sequence (CRS) T55C, T57C, A263G, A750G, A 1438G, A4769G, T6253C, A8860G, A15326G and various insertions at 309. and 315. The presence of the same control region mutations places these seven individuals into the same haplogroup, H15.

One could hypothesise that these individuals have a single female founder that migrated to South Africa between the 17th and 19th century. This ancestor might have settled in a region around the Kwa-Zulu Natal midlands near the Griqua population in Kokstad. Whole mitochondrial genome sequencing is necessary as coding region SNPs can further refine the haplogroups and confirm that these seven individuals do indeed share the same ancestor. There is a single additional SNP in one of the seven individuals (369A), but this low level of internal variation in the most variable regions of the genome suggest a short time period elapsing between the founder event and the present day population (Quintana-Murci *et al.* 2010).

3.1.3.3. Comparative data discussion on the dispersion of L0 Haplogroups throughout the Griqua population

A study by Quintana-Murci *et al.* (2010) assigned individuals into various L0d haplogroups using 11 diagnostic markers found in the mtDNA coding region along with polymorphisms found in HVI. Coverage of the mtDNA control region, from 16024 – 400bp, is sufficient for haplogroup assignment as mutations from HVI (16024 – 16365bp) and II (73 – 340) are taken in to account (Brandstätter *et al.* 2004; Chong *et al.* 2004). HVIII covers 438 – 576bp and mutations from this region have not been included in haplogroup assignment for the Griqua population. There are a few mutations between 450bp and 600bp that can further refine an individual into sub haplogroups of L0. Examples of these include the AC indel at position 523 (L0d2c), and the two transversion mutations at 456 (L0d2a) and 593 (L0d1c1) (Behar *et al.* 2012). This does not pose a problem as the indels at 515 – 522 are not included on Phylotree since they are not considered for phylogenetic reconstruction (www.phylotree.org/ 2014) and there are other control region mutations that provide evidence for L0d2a and L0d1c1 haplogroup assignment. What follows in this chapter will be an explanation of the L0 haplogroup statistics for the Griqua population as well as data comparisons with previous studies involving Khoe and/or San ancestry.

An initial study on the South African Coloured (SAC) population was done by Tishkoff *et al.* (2009). Through analysis of 1,327 nuclear microsatellite and insertion/deletion (indel) markers it was found that the SAC clusters in intermediate positions between African and non-African populations. The SAC showed an equal distribution of four maternal ancestries on the basis of STRUCTURE analysis. Namely southern African Khoisan, Bantu speaking, Indian and European. In a second study comparing a small sample of Coloured individuals with other worldwide populations it was concluded that the SAC resulted from a complex admixture process involving Bantu-speaking populations from South Africa, Europeans, south Asians and Indonesians.

This mixture of haplogroups strongly resembles the data collected on the Griqua population. A data comparison between results from Quintana-Murci *et al.* (2010) and the Griqua samples showed a shared pattern of L0 haplogroups in the post-colonial and Cape Coloured populations (Figure 3.9). The origin sample group showed major deviations with a large incorporation of haplogroup L0d2d and the absence of haplogroup L0d2a.

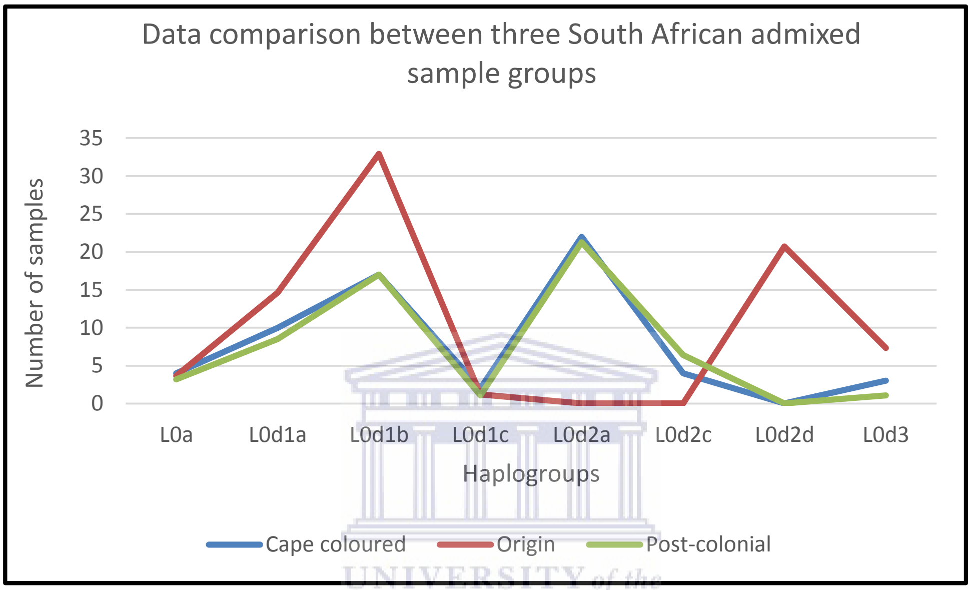


Figure 3.9. Data comparison between a previously researched Cape Coloured sample group (Quinata-Murci *et al.* 2010) and the two Griqua sample groups discussed in this thesis.

The isolated incorporation of haplogroups L0d2a and L0d2c in the post-colonial group and not the origin group confirms the initial hypothesis that the Griqua population was segregated between the late 18th century and present day (Figure 3.10). There is most likely a shared maternal ancestor for each of these haplogroups which could have been incorporated at one of the settlement areas along the Griqua Trek route (*e.g*. Griquastad and Kokstad). Isolated incorporation of haplogroup L0d2d in the origin group serves as further confirmation of this point. Clear differences between the two groups can also be observed by the incorporation of colonial ancestry (M and N) as well as indigenous populations (L1-6) as previously discussed.

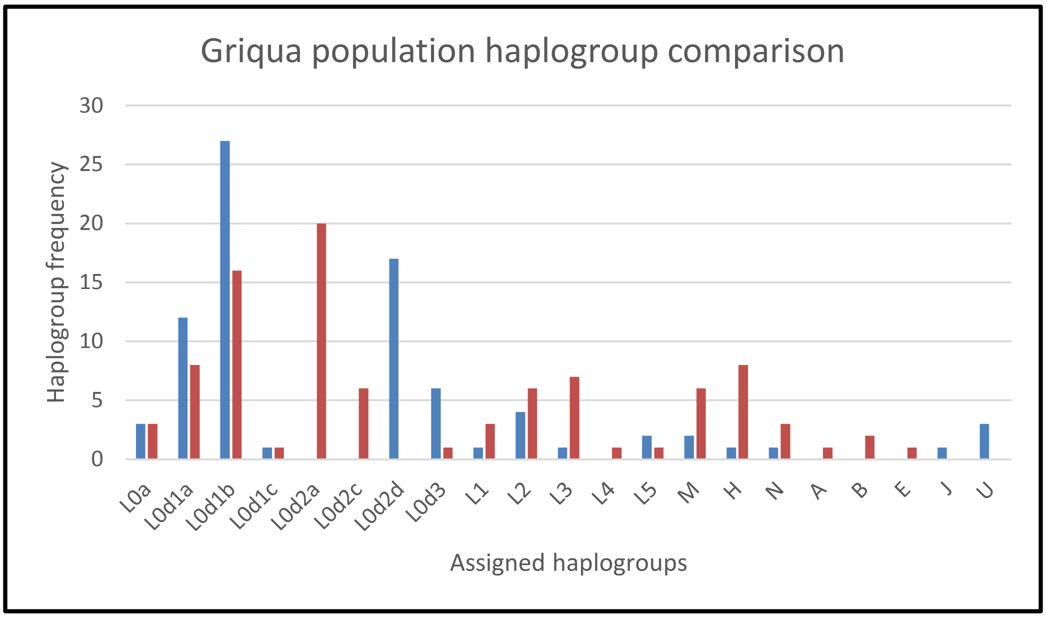


Figure 3.10. Haplogroup comparison between the origin (blue) and post-colonial (red) sample groups showing a large incorporation of L0d1b haplogroups in both groups.

One of the least prominent haplogroups (L0d3, <1%) in the post-colonial sample group belongs to the only sub-haplogroup of L0 which has geographic distribution outside of Southern African and occurs in Tanzania. Whole mtDNA genome sequence data places L0d3 as the oldest branch (64 000 ± 23 000 years before present) within the L0d clade (Schlebusch *et al.* 2013). A possible explanation for the appearance of L0d3 in populations of east and south African origins, is that L0d occurred over a wide geographic region that spanned these areas. The younger branches of L0d (L0d1 and L0d2) only evolved in the ancestors of present day San, while intermediate branches of L0d3 were lost because of drift and external factors such as the Bantu expansion (Schlebusch *et al.* 2012). Another explanation is that L0d3 evolved separately in East Africa, while L0d1 and L0d2 remained in the south. Any subsequent incorporation of haplogroup L0d3 in southern populations is as a result of migration (Chen *et al.* 1995). These statistics show a prominent Khoisan basis in the origin sample group and more gene flow being present in the post-colonial group as it contains more admixture and haplogroups from outside of Africa. Table 3.3 displays the variation within haplogroup L0 from both populations.

Table 3.3. Variation within haplogroup L0 for both Griqua sample groups.

L0 Haplogroups	Frequency in origin population (%)	Frequency in post-colonial population (%)
L0a	3.66	3.19
L0d1a	14.63	8.51
L0d1b	32.93	17.02
L0d1c	1.22	1.06
L0d2a	0	21.28
L0d2c	0	6.38
L0d2d	20.73	0
L0d3	7.32	1.06

Both populations have a large contingent of haplogroup L0d1b with it being the most frequent in the ancestral population (~41%). It is separated into four sub haplogroups (L0d1b, L0d1b2b, L0d1b2b2, L0d1b2b2b1) with an accumulation of various SNP mutations separating individuals. Haplogroup L0d2a was most frequent in the post-colonial population suggesting a shared maternal ancestor along a region of the Griqua trek route.

A Network 4.6.1.0 (www.fluxus-engineering.com/ 2014) analysis of the L0 haplotypes for the entire population was then done. For the Network analysis, the epsilon value was set to zero, transversions were weighted 3x higher than transitions, the hypervariable indels at position 16189, 305.1C and 315.1C were excluded following guidelines from Phylotree (www.phylotree.org/ 2014). The divergence of the Neanderthal mtDNA from the line leading to the contemporary human mtDNA gene pool is almost 3-fold older than the deepest divergence among contemporary human mtDNAs. This shows that the Neanderthal mtDNA and the human ancestral mtDNA gene pool have evolved as separate entities for a substantial period of time and gives no support to the notion that Neanderthals should have contributed mtDNA to the modern gene pool (Krings *et al.* 1999). The *Homo sapiens neandethalensis* genome therefore provides a suitable root for creating a median-joining network of human mtDNA for any population group.

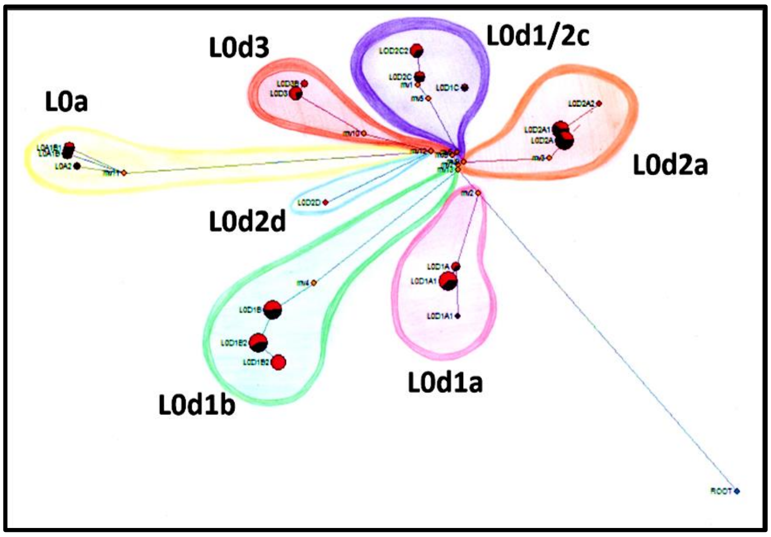


Figure 3.11. A median joining network analysis of the L0 haplotypes with *Homo sapiens neanderthalensis* (gi196123578) root with the origin group in red and the post-colonial group in black.

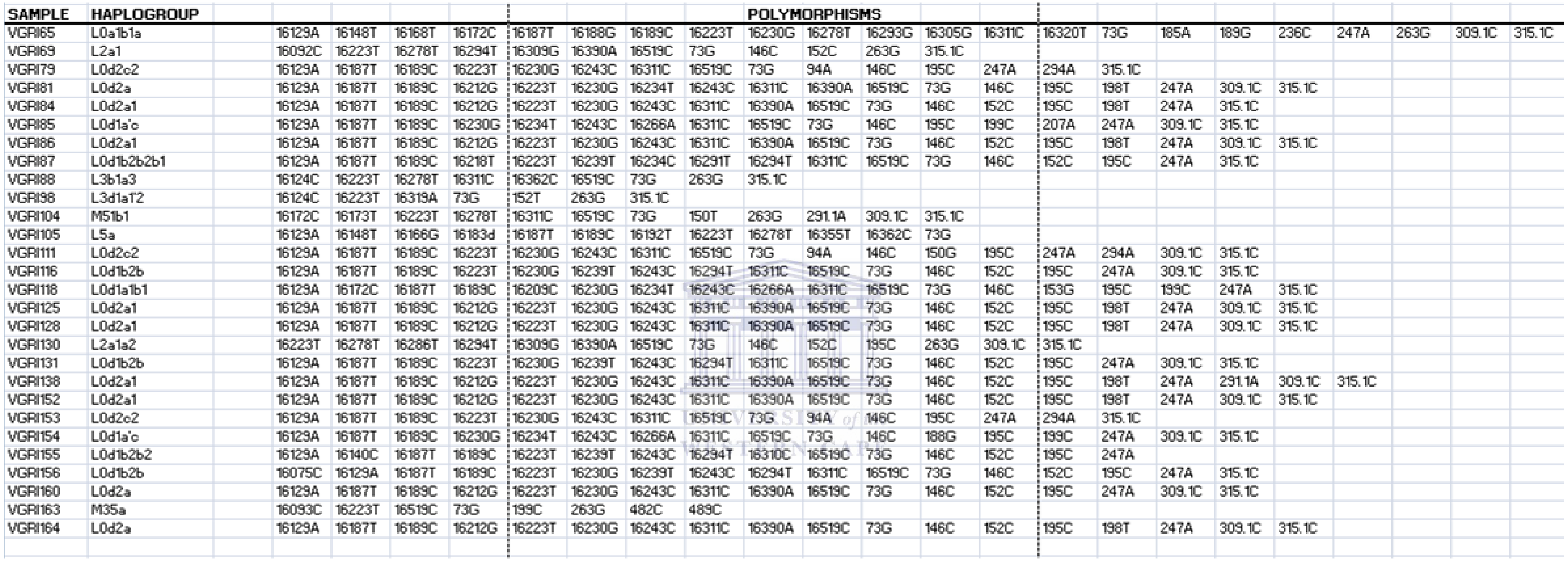
The network analysis shows the distribution of the L0 haplotypes and indicates shared ancestry between the two population groups. There are several individuals that share maternal family history between the two population groups, which indicates that they do share a common ancestry. This is to be expected since the Griqua Trek began in the Western Cape where the origin of the population lies.

Table 3.4 (pages 80 – 81) shows all the haplogroups assigned to individuals from the post-colonial sample group (N = 94) and table 3.5 (page 82 – 83) for the origin population group (N = 82).

Table 3.4. Haplogroup assignment for the post-colonial sample group (N = 94)

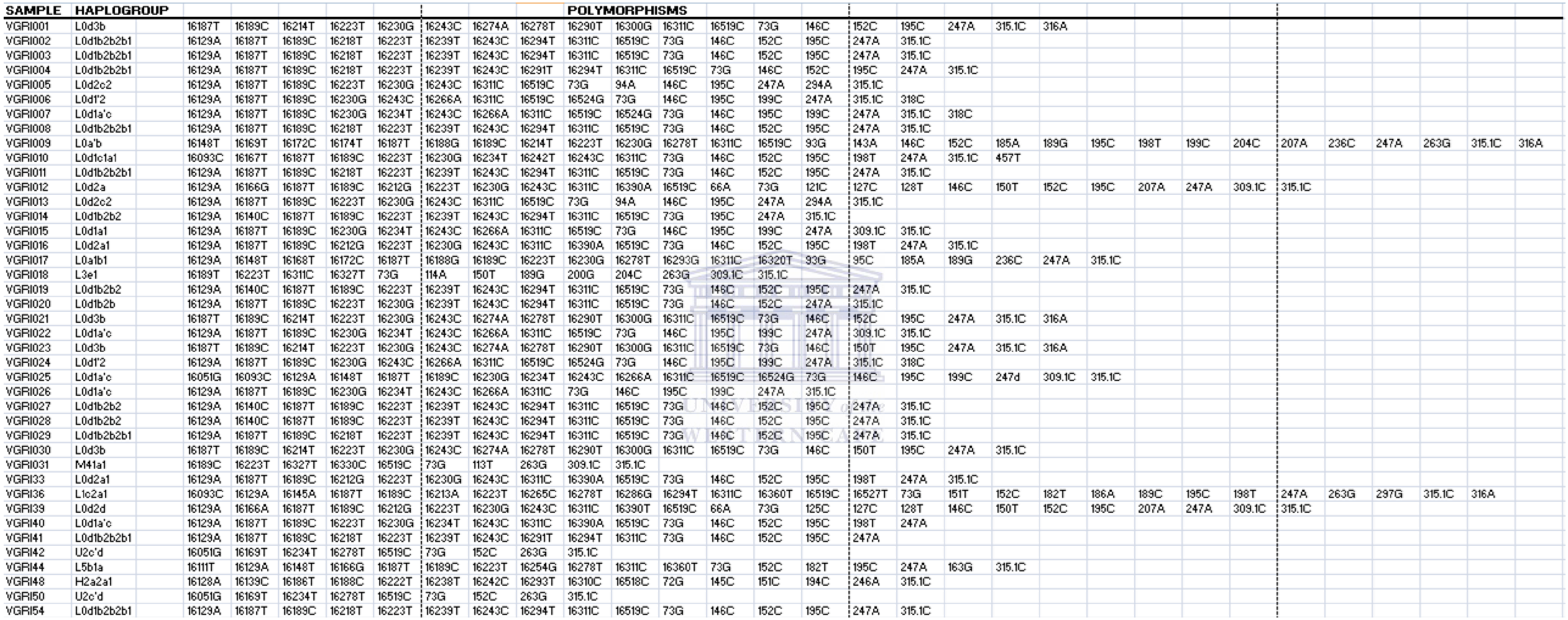
SAMPLE	HAPLOGROUP									POLYMORPHISMS																
GRI001	L3e2b	16172C	16183C	16189C	16223T	16245T	16320T	16519C	73G	150T	195C	263G	315.1G													
GRI002	L0d3b	16187T	16189C	16214T	16223T	16230G	16243C	16274A	16278T	16290T	16311C	16519C	73G	146C	150T	195C	247A	309.1C	315.1C	316A						
GRI003	B5b1c	16140C	16182C	16183C	16189C	16243C	16519C	73G	103A	152C	204C	263G	309.1CC	315.1C												
GRI004	L1c2a3	16172C	16184T	16187T	16189C	16223T	16265C	16278T	16286G	16294T	16311C	16360T	16519C	16527T	73G	151T	152C	182T	186A	189C	195C	198T	247A	263G	297G	309.1C
GRI005	L0d2a	16129A	16187T	16189C	16212G	16223T	16230G	16243C	16311C	16390A	16519C	73G	146C	152C	195C	247A	309.1C	315.1C								
GRI006	L0d1b2b2	16129A	16187T	16189C	16223T	16239T	16243C	16294T	16311C	16519C	73G	146C	152C	195C	247A	309.1C	315.1C									
GRI007	L1c1	16129A	16172C	16173T	16188A	16189C	16223T	16256T	16278T	16293G	16294T	16311C	16360T	16370C	16519C	73G	151T	152C	182T	186A	189G	195C	198T	247d	263G	297G
GRI008	L0d1a1b1	16129A	16187T	16189C	16209C	16230G	16234T	16243C	16266A	16311C	16519C	73G	146C	153G	195C	199C	247A	315.1C								
GRI009	L0d1a	16129A	16187T	16189C	16192T	16230G	16234T	16243C	16266G	16311C	16519C	73G	146C	195C	198T	199C	247A	309.1C	315.1C							
GRI010	L0d1b2b	16129A	16187T	16189C	16223T	16230G	16239T	16243C	16294T	16311C	16519C	73G	146C	152C	195C	247A	315.1C									
GRI011	L0d1b2b	16129A	16187T	16189C	16223T	16230G	16239T	16243C	16294T	16311C	16519C	73G	146C	152C	195C	247A	309.1C	315.1C	366A							
GRI012	L0d2c	16129A	16187T	16189C	16230G	16234T	16223T	16242T	16243C	16311C	16519C	73G	146C	152C	195C	247A	294A	315.1C								
GRI013	A14	16223T	16290T	16319A	16362C	73G	151T	152C	200G	235G	263G	315.1C														
GRI014	L0d2a1	16129A	16187T	16189C	16212G	16223T	16230G	16243C	16311C	16390A	16519C	73G	146C	152C	195C	198T	247A	309.1C	315.1C							
GRI015	L0d2a1	16129A	16187T	16189C	16212G	16223T	16230G	16243C	16311C	16390A	16519C	73G	146C	152C	195C	198T	247A	309.1C	315.1C							
GRI016	L0d1a1c	16129A	16187T	16189C	16230G	16234T	16243C	16266A	16311C	16519C	73G	146C	152C	199C	247A	315.1C										
GRI017	U2e1'2'3	16362C	16519C	16051A	16093C	16129C	16189C	73G	152C	217C	263G	309.1C	315.1C													
GRI018	L0d1a1	16129A	16182A	16183A	16189C	16230G	16234T	16243C	16266A	16311C	16519C	73G	146C	195C	199C	247A	309.1C	315.1C								
GRI019	M3	16126C	16223T	16519C	73G	263G	309.1C	315.1C	482C	489C																
GRI020	L0d1b2b2	16129A	16187T	16189C	161218T	16223T	16239T	16243C	16294T	16311C	16519C	73G	146C	152C	195C	247A	315.1C									
GRI021	L0d2a1	16129A	16187T	16189C	16212G	16223T	16230G	16243C	16311C	16390C	16519C	73G	146C	152C	195C	198T	247A	309.1C	315.1C							
GRI022	H7a1	16261T	16519C	263G	309.1CC	315.1C																				
GRI023	L0d2a1	16129A	16187T	16189C	16212G	16223T	16230G	16243C	16311C	16390A	16519C	73G	146C	152C	195C	198T	247A	315.1C								
GRI024	L2c2b1b	16223T	16264T	16265G	16278T	16311C	16390A	73G	93G	146C	150T	152C	182T	183G	195C	198T	263G	315.1C								
GRI025	L0d1b2b2	16129A	16187T	16189C	16223T	16239T	16243C	16294T	16311C	16519C	73G	146C	152C	195C	247A	315.1C										
GRI026	H15	55C	57C	263G	309.1C	315.1C																				
GRI027	L0d1b2b2	16129A	16187T	16189C	16192T	16223T	16239T	16243C	16294T	16311C	16519C	73G	146C	152C	195C	247A	315.1C									
GRI028	M7b2	16184T	16626.1C	56.1C	264G	309.1CC	315.1C																			
GRI029	H15	55C	57C	263G	309.1CC	315.1C																				
GRI030	H15	55C	57C	263G	309.1CC	315.1C																				
GRI031	H15	55C	57C	263G	305.1CC	315.1C																				
GRI032	N	16126C	16223T	16519C	73G	263G	309.1C	315.1C																		
GRI033	N	16126C	16223T	16519C	73G	263G	309.1C	315.1C																		
GRI034	M3	16126C	16223T	16519C	73G	263G	309.1C	315.1C																		
GRI035	M3	16126C	16223T	16519C	73G	263G	309.1C	315.1C																		
GRI040	H15	55C	57C	263G	309.1CC	315.1C	369A																			
GRI041	L2a1	16198T	16223T	16278T	16290T	16294T	16309G	16390A	73G	146C	152C	195C	263G	315.1CG												
GRI042	L2a1b1a	16182C	16183C	16189C	16223T	16278T	16290T	16294T	16309G	16390A	16519C	73G	146C	152C	195C	263G	315.1C									
GRI043	L0d2a1	16093C	16129A	16187T	16189C	16212G	16223T	16230G	16243C	16311C	16390A	16519C	73G	146C	152C	195C	198T	247A	315.1C							
GRI044	L2a1d	16093C	16223T	16278T	16294T	16311C	16390C	16519C	73G	143A	146C	152C	182T	195C	263G	315.1C										
GRI045	H15	55C	57C	263G	309.1CC	315.1C																				
GRI046	L3e2b+152	16189C	16266T	16172C	16223T	16320T	16519C	73G	150T	152C	195C	263G	309.1C	315.1C												
GRI047	L0d2a1	16187T	16189C	16223T	16230G	16243C	16311C	16390A	16519C	73G	146C	152C	195C	198T	204C	247A	309.1C	315.1C								
GRI048	L0a1b1	16129A	16148T	16168T	16172C	16187T	16188G	16189C	16223T	16230G	16278T	16293G	16311C	16320T	93G	95C	185A	189G	236C	247A	309.1C	315.1C				
GRI049	L0d2a1	16129A	16187T	16189C	16212G	16223T	16230G	16243C	16311C	16390A	16519C	73G	146C	152C	195C	198T	247A	257A	309.1C	315.1C						
GRI050	L0d1b2b2b1	16129A	16187T	16189C	16218T	16223T	16239T	16243C	16294T	16311C	16519C	73G	146C	152C	195C	247A	315.1C									
GRI051	L4b2b1	16051G	16114T	16189C	16192T	16223T	16293T	16311C	16316G	16355T	16362C	16399G	16519C	73G	146C	152C	195C	244G	263G	315.1C						
GRI052	L0d2a1	16129A	16187T	16189C	16212G	16223T	16230G	16243C	16311C	16390C	16519C	73G	146C	152C	195C	198T	247A	309.1C	315.1C							
GRI053	L3e1a	16185T	16223T	16327T	16519C	73G	150T	189G	200G	263G	315.1C															
GRI054	L0d1b2b	16129A	16187T	16189C	16223T	16230G	16239T	16243C	16294T	16311C	16519C	73G	146C	152C	195C	247A	315.1C									
GRI055	L0d1b2b2	16129A	16187T	16189C	16223T	16239T	16243C	16294T	16311C	16519C	73G	146C	152C	195C	247A	315.1C										
GRI056	L0d2c	16129A	16187T	16189C	16223T	16230G	16243C	16311C	16519C	73G	146C	152C	195C	247A	294A	315.1C										
GRI057	L3e3	16223T	16265T	16509C	73G	150T	195C	263G	315.1CG																	
GRI058	L0d1a1b1	16129A	16187T	16189C	16209C	16230G	16234T	16243C	16266A	16311C	16519C	73G	146C	153G	195C	199C	247A	315.1A								
GRI059	L0d1b2b	16129A	16187T	16189C	16223T	16230G	16239T	16243C	16294T	16311C	16519C	73G	146C	152C	195C	247A	315.1C									
GRI060	N	16219A	16223T	16291T	16519C	73G	263G	309.1CC	315.1C																	
GRI061	L0d1c	16189C	16223T	16230G	16234T	16243C	16249C	16311C	73G	146C	152C	195C	247A	315.1C	456T											
GRI062	L1c1	16129A	16172C	16173T	16188A	16189C	16223T	16256T	16278T	16293G	16294T	16311C	16360T	16368C	16519C	73G	151T	152C	182T	186A	189G	195C	198T	247d	268G	297G
GRI063	L0a2a2	16148T	16172C	16187T	16188G	16189C	16223T	16230G	16240G	16311C	16320T	16519C	64T	93G	152C	189G	207A	236C	247A	263G	309.1C	315.1C				

Table 3.4. Continued haplogroup assignment for the post-colonial sample group (N = 94)



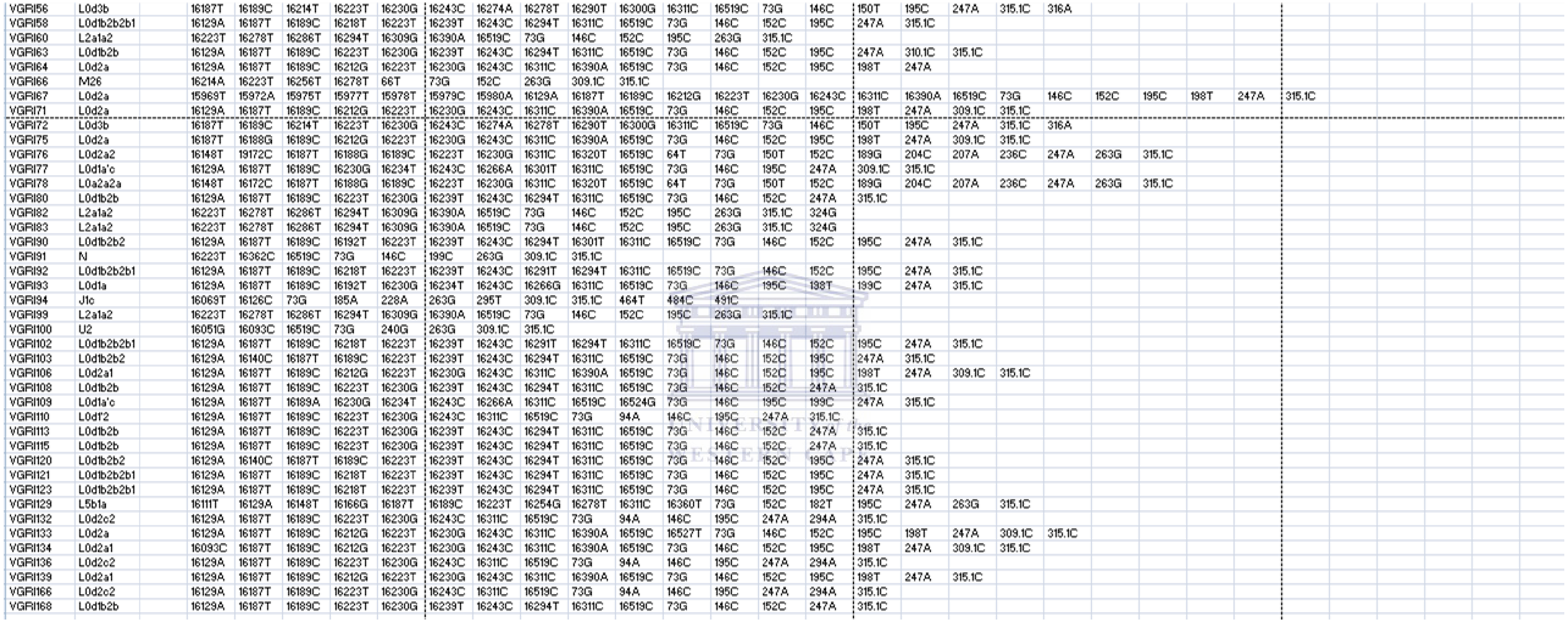
SAMPLE	HAPLOGROUP										POLYMORPHISMS													
VGRI65	L0a1b1a		16129A	16148T	16168T	16172C	16187T	16188G	16189C	16223T	16230G	16278T	16293G	16305G	16311C	16320T	73G	185A	189G	236C	247A	263G	309.1C	315.1C
VGRI69	L2a1		16092C	16223T	16278T	16294T	16309G	16390A	16519C	73G	146C	152C	263G	315.1C										
VGRI79	L0d2c2		16129A	16187T	16189C	16223T	16230G	16243C	16311C	16519C	73G	94A	146C	195C	247A	294A	315.1C							
VGRI81	L0d2a		16129A	16187T	16189C	16212G	16223T	16230G	16234T	16243C	16311C	16390A	16519C	73G	146C	195C	198T	247A	309.1C	315.1C				
VGRI84	L0d2a1		16129A	16187T	16189C	16212G	16223T	16230G	16243C	16311C	16390A	16519C	73G	146C	152C	195C	198T	247A	315.1C					
VGRI85	L0d1a'c		16129A	16187T	16189C	16230G	16234T	16243C	16266A	16311C	16519C	73G	146C	195C	199C	207A	247A	309.1C	315.1C					
VGRI86	L0d2a1		16129A	16187T	16189C	16212G	16223T	16230G	16243C	16311C	16390A	16519C	73G	146C	152C	195C	198T	247A	309.1C	315.1C				
VGRI87	L0d1b2b2b1		16129A	16187T	16189C	16218T	16223T	16239T	16234C	16291T	16294T	16311C	16519C	73G	146C	152C	195C	247A	315.1C					
VGRI88	L3b1a3		16124C	16223T	16278T	16311C	16362C	16519C	73G	263G	315.1C													
VGRI98	L3d1a1'2		16124C	16223T	16319A	73G	152T	263G	315.1C															
VGRI104	M51b1		16172C	16173T	16223T	16278T	16311C	16519C	73G	150T	263G	291.1A	309.1C	315.1C										
VGRI105	L5a		16129A	16148T	16166G	16183d	16187T	16189C	16192T	16223T	16278T	16355T	16362C	73G										
VGRI111	L0d2c2		16129A	16187T	16189C	16223T	16230G	16243C	16311C	16519C	73G	94A	146C	150G	195C	247A	294A	309.1C	315.1C					
VGRI116	L0d1b2b		16129A	16187T	16189C	16223T	16230G	16239T	16243C	16294T	16311C	16519C	73G	146C	152C	195C	247A	309.1C	315.1C					
VGRI118	L0d1a1b1		16129A	16172C	16187T	16189C	16209C	16230G	16234T	16243C	16266A	16311C	16519C	73G	146C	153G	195C	199C	247A	315.1C				
VGRI125	L0d2a1		16129A	16187T	16189C	16212G	16223T	16230G	16243C	16311C	16390A	16519C	73G	146C	152C	195C	198T	247A	309.1C	315.1C				
VGRI128	L0d2a1		16129A	16187T	16189C	16212G	16223T	16230G	16243C	16311C	16390A	16519C	73G	146C	152C	195C	198T	247A	309.1C	315.1C				
VGRI130	L2a1a2		16223T	16278T	16286T	16294T	16309G	16390A	16519C	73G	146C	152C	195C	263G	309.1C	315.1C								
VGRI131	L0d1b2b		16129A	16187T	16189C	16223T	16230G	16239T	16243C	16294T	16311C	16519C	73G	146C	152C	195C	247A	309.1C	315.1C					
VGRI138	L0d2a1		16129A	16187T	16189C	16212G	16223T	16230G	16243C	16311C	16390A	16519C	73G	146C	152C	195C	198T	247A	291.1A	309.1C	315.1C			
VGRI152	L0d2a1		16129A	16187T	16189C	16212G	16223T	16230G	16243C	16311C	16390A	16519C	73G	146C	152C	195C	198T	247A	309.1C	315.1C				
VGRI153	L0d2c2		16129A	16187T	16189C	16223T	16230G	16243C	16311C	16519C	73G	94A	146C	195C	247A	294A	315.1C							
VGRI154	L0d1a'c		16129A	16187T	16189C	16230G	16234T	16243C	16266A	16311C	16519C	73G	146C	188G	195C	199C	247A	309.1C	315.1C					
VGRI155	L0d1b2b2		16129A	16140C	16187T	16189C	16223T	16239T	16243C	16294T	16310C	16519C	73G	146C	152C	195C	247A							
VGRI156	L0d1b2b		16075C	16129A	16187T	16189C	16223T	16230G	16239T	16243C	16294T	16311C	16519C	73G	146C	152C	195C	247A	315.1C					
VGRI160	L0d2a		16129A	16187T	16189C	16212G	16223T	16230G	16243C	16311C	16390A	16519C	73G	146C	152C	195C	247A	309.1C	315.1C					
VGRI163	M35a		16093C	16223T	16519C	73G	199C	263G	482C	489C														
VGRI164	L0d2a		16129A	16187T	16189C	16212G	16223T	16230G	16243C	16311C	16390A	16519C	73G	146C	152C	195C	198T	247A	309.1C	315.1C				

Table 3.5. Haplogroup assignment for the origin sample group (N = 82)



SAMPLE	HAPLOGROUP									POLYMORPHISMS																				
VGRI001	L0d3b	16187T	16189C	16214T	16223T	16230G	16243C	16274A	16278T	16290T	16300G	16311C	16519C	73G	146C	152C	195C	247A	315.1C	316A										
VGRI002	L0d1b2b2b1	16129A	16187T	16189C	16218T	16223T	16239T	16243C	16294T	16311C	16519C	73G	146C	152C	195C	247A	315.1C													
VGRI003	L0d1b2b2b1	16129A	16187T	16189C	16218T	16223T	16239T	16243C	16294T	16311C	16519C	73G	146C	152C	195C	247A	315.1C													
VGRI004	L0d1b2b2b1	16129A	16187T	16189C	16218T	16223T	16239T	16243C	16291T	16294T	16311C	16519C	73G	146C	152C	195C	247A	315.1C												
VGRI005	L0d2c2	16129A	16187T	16189C	16223T	16230G	16243C	16311C	16519C	73G	94A	146C	195C	247A	294A	315.1C														
VGRI006	L0d1'2	16129A	16187T	16189C	16230G	16243C	16266A	16311C	16519C	16524G	73G	146C	195C	199C	247A	315.1C	318C													
VGRI007	L0d1a'c	16129A	16187T	16189C	16230G	16234T	16243C	16266A	16311C	16519C	16524G	73G	146C	195C	199C	247A	315.1C	318C												
VGRI008	L0d1b2b2b1	16129A	16187T	16189C	16218T	16223T	16239T	16243C	16294T	16311C	16519C	73G	146C	152C	195C	247A	315.1C													
VGRI009	L0a'b	16148T	16169T	16172C	16174T	16187T	16188G	16189C	16214T	16223T	16230G	16278T	16311C	16519C	93G	143A	146C	152C	185A	189G	195C	198T	199C	204C	207A	236C	247A	263G	315.1C	316A
VGRI010	L0d1c1a1	16093C	16167T	16187T	16189C	16223T	16230G	16234T	16242T	16243C	16311C	73G	146C	152C	195C	198T	247A	315.1C	457T											
VGRI011	L0d1b2b2b1	16129A	16187T	16189C	16218T	16223T	16239T	16243C	16294T	16311C	16519C	73G	146C	152C	195C	247A	315.1C													
VGRI012	L0d2a	16129A	16166G	16187T	16189C	16212G	16223T	16230G	16243C	16311C	16390A	16519C	66A	73G	121C	127C	128T	146C	150T	152C	195C	207A	247A	309.1C	315.1C					
VGRI013	L0d2c2	16129A	16187T	16189C	16223T	16230G	16243C	16311C	16519C	73G	94A	146C	195C	247A	294A	315.1C														
VGRI014	L0d1b2b2	16129A	16140C	16187T	16189C	16223T	16239T	16243C	16294T	16311C	16519C	73G	195C	247A	315.1C															
VGRI015	L0d1a1	16129A	16187T	16189C	16230G	16234T	16243C	16266A	16311C	16519C	73G	146C	195C	199C	247A	309.1C	315.1C													
VGRI016	L0d2a1	16129A	16187T	16189C	16212G	16223T	16230G	16243C	16311C	16390A	16519C	73G	146C	152C	195C	198T	247A	315.1C												
VGRI017	L0a1b1	16129A	16148T	16168T	16172C	16187T	16188G	16189C	16223T	16230G	16278T	16293G	16311C	16320T	93G	95C	185A	189G	236C	247A	315.1C									
VGRI018	L3e1	16189T	16223T	16311C	16327T	73G	114A	150T	189G	200G	204C	263G	309.1C	315.1C																
VGRI019	L0d1b2b2	16129A	16140C	16187T	16189C	16223T	16239T	16243C	16294T	16311C	16519C	73G	146C	152C	195C	247A	315.1C													
VGRI020	L0d1b2b	16129A	16187T	16189C	16223T	16230G	16239T	16243C	16294T	16311C	16519C	73G	146C	152C	247A	315.1C														
VGRI021	L0d3b	16187T	16189C	16214T	16223T	16230G	16243C	16274A	16278T	16290T	16300G	16311C	16519C	73G	146C	152C	195C	247A	315.1C	316A										
VGRI022	L0d1a'c	16129A	16187T	16189C	16230G	16234T	16243C	16266A	16311C	16519C	73G	146C	195C	199C	247A	309.1C	315.1C													
VGRI023	L0d3b	16187T	16189C	16214T	16223T	16230G	16243C	16274A	16278T	16290T	16300G	16311C	16519C	73G	146C	150T	195C	247A	315.1C	316A										
VGRI024	L0d1'2	16129A	16187T	16189C	16230G	16243C	16266A	16311C	16519C	16524G	73G	146C	195C	199C	247A	315.1C	318C													
VGRI025	L0d1a'c	16051G	16093C	16129A	16148T	16187T	16189C	16230G	16234T	16243C	16266A	16311C	16519C	16524G	73G	146C	195C	199C	247d	309.1C	315.1C									
VGRI026	L0d1a'c	16129A	16187T	16189C	16230G	16234T	16243C	16266A	16311C	73G	146C	195C	199C	247A	315.1C															
VGRI027	L0d1b2b2	16129A	16140C	16187T	16189C	16223T	16239T	16243C	16294T	16311C	16519C	73G	146C	152C	195C	247A	315.1C													
VGRI028	L0d1b2b2	16129A	16140C	16187T	16189C	16223T	16239T	16243C	16294T	16311C	16519C	73G	146C	152C	195C	247A	315.1C													
VGRI029	L0d1b2b2b1	16129A	16187T	16189C	16218T	16223T	16239T	16243C	16294T	16311C	16519C	73G	146C	152C	195C	247A	315.1C													
VGRI030	L0d3b	16187T	16189C	16214T	16223T	16230G	16243C	16274A	16278T	16290T	16300G	16311C	16519C	73G	146C	150T	195C	247A	315.1C											
VGRI031	M41a1	16189C	16223T	16327T	16330C	16519C	73G	113T	263G	309.1C	315.1C																			
VGRI33	L0d2a1	16129A	16187T	16189C	16212G	16223T	16230G	16243C	16311C	16390A	16519C	73G	146C	152C	195C	198T	247A	315.1C												
VGRI36	L1c2a1	16093C	16129A	16145A	16187T	16189C	16213A	16223T	16265C	16278T	16286G	16294T	16311C	16360T	16519C	16527T	73G	151T	152C	182T	186A	189C	195C	198T	247A	263G	297G	315.1C	316A	
VGRI39	L0d2d	16129A	16166A	16187T	16189C	16212G	16223T	16230G	16243C	16311C	16390T	16519C	66A	73G	125C	127C	128T	146C	150T	152C	195C	207A	247A	309.1C	315.1C					
VGRI40	L0d1a'c	16129A	16187T	16189C	16223T	16230G	16234T	16243C	16311C	16390A	16519C	73G	146C	152C	195C	198T	247A													
VGRI41	L0d1b2b2b1	16129A	16187T	16189C	16218T	16223T	16239T	16243C	16291T	16294T	16311C	73G	146C	152C	195C	247A														
VGRI42	U2c'd	16051G	16169T	16234T	16278T	16519C	73G	152C	263G	315.1C																				
VGRI44	L5b1a	16111T	16129A	16148T	16166G	16187T	16189C	16223T	16254G	16278T	16311C	16360T	73G	152C	182T	195C	247A	163G	315.1C											
VGRI48	H2a2a1	16128A	16139C	16186T	16188C	16222T	16238T	16242C	16293T	16310C	16518C	72G	145C	151C	194C	246A	315.1C													
VGRI50	U2c'd	16051G	16169T	16234T	16278T	16519C	73G	152C	263G	315.1C																				
VGRI54	L0d1b2b2b1	16129A	16187T	16189C	16218T	16223T	16239T	16243C	16294T	16311C	16519C	73G	146C	152C	195C	247A	315.1C													

Table 3.5. Continued haplogroup assignment for the origin sample group (N = 82)



VGRI56	L0d3b	16187T	16189C	16214T	16223T	16230G	16243C	16274A	16278T	16290T	16300G	16311C	16519C	73G	146C	150T	195C	247A	315.1C	316A					
VGRI58	L0d1b2b2b1	16129A	16187T	16189C	16218T	16223T	16239T	16243C	16294T	16311C	16519C	73G	146C	152C	195C	247A	315.1C								
VGRI60	L2a1a2	16223T	16278T	16286T	16294T	16309G	16390A	16519C	73G	146C	152C	195C	263G	315.1C											
VGRI63	L0d1b2b	16129A	16187T	16189C	16223T	16230G	16239T	16243C	16294T	16311C	16519C	73G	146C	152C	195C	247A	310.1C	315.1C							
VGRI64	L0d2a	16129A	16187T	16189C	16212G	16223T	16230G	16243C	16311C	16390A	16519C	73G	146C	152C	195C	198T	247A								
VGRI66	M26	16214A	16223T	16256T	16278T	66T	73G	152C	263G	309.1C	315.1C														
VGRI67	L0d2a	15969T	15972A	15975T	15977T	15978T	15979C	15980A	16129A	16187T	16189C	16212G	16223T	16230G	16243C	16311C	16390A	16519C	73G	146C	152C	195C	198T	247A	315.1C
VGRI71	L0d2a	16129A	16187T	16189C	16212G	16223T	16230G	16243C	16311C	16390A	16519C	73G	146C	152C	195C	198T	247A	309.1C	315.1C						
VGRI72	L0d3b	16187T	16189C	16214T	16223T	16230G	16243C	16274A	16278T	16290T	16300G	16311C	16519C	73G	146C	150T	195C	247A	315.1C	316A					
VGRI75	L0d2a	16187T	16188G	16189C	16212G	16223T	16230G	16243C	16311C	16390A	16519C	73G	146C	152C	195C	198T	247A	309.1C	315.1C						
VGRI76	L0d2a2	16148T	19172C	16187T	16188G	16189C	16223T	16230G	16311C	16320T	16519C	64T	73G	150T	152C	189G	204C	207A	236C	247A	263G	315.1C			
VGRI77	L0d1a'c	16129A	16187T	16189C	16230G	16234T	16243C	16266A	16301T	16311C	16519C	73G	146C	195C	247A	309.1C	315.1C								
VGRI78	L0a2a2a	16148T	16172C	16187T	16188G	16189C	16223T	16230G	16311C	16320T	16519C	64T	73G	150T	152C	189G	204C	207A	236C	247A	263G	315.1C			
VGRI80	L0d1b2b	16129A	16187T	16189C	16223T	16230G	16239T	16243C	16294T	16311C	16519C	73G	146C	152C	247A	315.1C									
VGRI82	L2a1a2	16223T	16278T	16286T	16294T	16309G	16390A	16519C	73G	146C	152C	195C	263G	315.1C	324G										
VGRI83	L2a1a2	16223T	16278T	16286T	16294T	16309G	16390A	16519C	73G	146C	152C	195C	263G	315.1C	324G										
VGRI90	L0d1b2b2	16129A	16187T	16189C	16192T	16223T	16239T	16243C	16294T	16301T	16311C	16519C	73G	146C	152C	195C	247A	315.1C							
VGRI91	N	16223T	16362C	16519C	73G	146C	199C	263G	309.1C	315.1C															
VGRI92	L0d1b2b2b1	16129A	16187T	16189C	16218T	16223T	16239T	16243C	16291T	16294T	16311C	16519C	73G	146C	152C	195C	247A	315.1C							
VGRI93	L0d1a	16129A	16187T	16189C	16192T	16230G	16234T	16243C	16266G	16311C	16519C	73G	146C	195C	198T	199C	247A	315.1C							
VGRI94	J1c	16069T	16126C	73G	185A	228A	263G	295T	309.1C	315.1C	464T	484C	491C												
VGRI99	L2a1a2	16223T	16278T	16286T	16294T	16309G	16390A	16519C	73G	146C	152C	195C	263G	315.1C											
VGRI100	U2	16051G	16093C	16519C	73G	240G	263G	309.1C	315.1C																
VGRI102	L0d1b2b2b1	16129A	16187T	16189C	16218T	16223T	16239T	16243C	16291T	16294T	16311C	16519C	73G	146C	152C	195C	247A	315.1C							
VGRI103	L0d1b2b2	16129A	16140C	16187T	16189C	16223T	16239T	16243C	16294T	16311C	16519C	73G	146C	152C	195C	247A	315.1C								
VGRI106	L0d2a1	16129A	16187T	16189C	16212G	16223T	16230G	16243C	16311C	16390A	16519C	73G	146C	152C	195C	198T	247A	309.1C	315.1C						
VGRI108	L0d1b2b	16129A	16187T	16189C	16223T	16230G	16239T	16243C	16294T	16311C	16519C	73G	146C	152C	247A	315.1C									
VGRI109	L0d1a'c	16129A	16187T	16189A	16230G	16234T	16243C	16266A	16311C	16519C	16524G	73G	146C	195C	199C	247A	315.1C								
VGRI110	L0d1'2	16129A	16187T	16189C	16223T	16230G	16243C	16311C	16519C	73G	94A	146C	195C	247A	315.1C										
VGRI113	L0d1b2b	16129A	16187T	16189C	16223T	16230G	16239T	16243C	16294T	16311C	16519C	73G	146C	152C	247A	315.1C									
VGRI115	L0d1b2b	16129A	16187T	16189C	16223T	16230G	16239T	16243C	16294T	16311C	16519C	73G	146C	152C	247A	315.1C									
VGRI120	L0d1b2b2	16129A	16140C	16187T	16189C	16223T	16239T	16243C	16294T	16311C	16519C	73G	146C	152C	195C	247A	315.1C								
VGRI121	L0d1b2b2b1	16129A	16187T	16189C	16218T	16223T	16239T	16243C	16294T	16311C	16519C	73G	146C	152C	195C	247A	315.1C								
VGRI123	L0d1b2b2b1	16129A	16187T	16189C	16218T	16223T	16239T	16243C	16294T	16311C	16519C	73G	146C	152C	195C	247A	315.1C								
VGRI129	L5b1a	16111T	16129A	16148T	16166G	16187T	16189C	16223T	16254G	16278T	16311C	16360T	73G	152C	182T	195C	247A	263G	315.1C						
VGRI132	L0d2c2	16129A	16187T	16189C	16223T	16230G	16243C	16311C	16519C	73G	94A	146C	195C	247A	294A	315.1C									
VGRI133	L0d2a	16129A	16187T	16189C	16212G	16223T	16230G	16243C	16311C	16390A	16519C	16527T	73G	146C	152C	195C	198T	247A	309.1C	315.1C					
VGRI134	L0d2a1	16093C	16187T	16189C	16212G	16223T	16230G	16243C	16311C	16390A	16519C	73G	146C	152C	195C	198T	247A	309.1C	315.1C						
VGRI136	L0d2c2	16129A	16187T	16189C	16223T	16230G	16243C	16311C	16519C	73G	94A	146C	195C	247A	294A	315.1C									
VGRI139	L0d2a1	16129A	16187T	16189C	16212G	16223T	16230G	16243C	16311C	16390A	16519C	73G	146C	152C	195C	198T	247A	315.1C							
VGRI166	L0d2c2	16129A	16187T	16189C	16223T	16230G	16243C	16311C	16519C	73G	94A	146C	195C	247A	294A	315.1C									
VGRI168	L0d1b2b	16129A	16187T	16189C	16223T	16230G	16239T	16243C	16294T	16311C	16519C	73G	146C	152C	247A	315.1C									

3.1.3.4. A forensic evaluation of the mtDNA haplotypes from the Griqua population

Haplotypes incorporating sequence data from HVI and II (16024 – 352bp) of all the samples (N = 176) were taken into consideration for this evaluation. AMOVA statistical analysis was done using Arlequin V3.5.1.2 (Excoffier *et al.* 2010) with the results shown in table 3.6. In total 148 haplotypes were found in the population group. With 63 unique haplotypes in the origin sample group and 85 in the post-colonial sample group. There was insignificant statistical diversity between the sample groups. Rather all of the variation occurred within the two sample groups (Origin and Post-colonial) with each group having a similar dispersion of haplotypes. This supports historical records of shared ancestors in the initial Griqua population group (pre-17th century). The most prevalent shared haplotype was between 7 individuals from the origin sample group and 2 from the post-colonial sample group, they had the assigned haplogroup L0d1b2b2b1. There were also 4, 3 and 2 origin sample group individuals with the assigned haplogroups L0d1b2b2, L0d1b2b and L0d2a respectively. The largest shared haplotype found solely in the Post-colonial group belonged to 2 individuals from haplogroup M3. Since the post-colonial sample group had 15% more unique haplotypes than the origin group, it can be said that these individuals showed more sequence variation in HVI and II. Which aligns with the original hypothesis that the post-colonial group contains greater genetic diversity than the origin group due to Pan-African and Eurasian (colonial) influences.

Table 3.6 Analysis of molecular variance (AMOVA) of the mtDNA data for the two Griqua sample groups.

Source of variation	d.f.	Sum of squares	Variance components	Percentage of variation
Among populations	1	0.500	0.00000Va	0.00
Within populations	174	87.000	0.50000Vb	100.00
Total	175	87.500	0.50000	

Fixation Index (FST): 0.00000 P-value = 1.00000+-0.00000

3.2. Analysis of the UWC 10plex

3.2.1. Multiplex performance

After Y-STR polymorphisms have been amplified using PCR, the length of the products must be precisely measured. In capillary electrophoresis, narrow glass tubes are filled with an entangled polymer solution to separate the DNA molecules by size, shape and charge. The reference marker, or internal lane size standard, is added to the post-PCR product along with formamide which denatures the dsDNA.

Data is collected from capillary electrophoresis using a program such as Genemapper-IDX (*Applied Biosystems*). Different fluorescent dyes are used to differentiate between markers in the same size range. A matrix file in the program which contains information on the amount of overlap in the spectra, will produce peaks within the profile that are composed of only one colour. The height of the peaks is measured in relative fluorescent units (rfu) with the height being proportional to the amount of PCR product that is detected.

The internal-lane size standard contains fragments of DNA of known lengths that are labelled with a fluorescent dye (LIZ). These fragments are detected along with the amplified PCR products during capillary electrophoresis and are used to size the unknown fragments based on migration speed.

After analysing the raw data with the software, the end result is an electropherogram with a series of peaks that represent different alleles. The size, peak height and peak area is also measured to assess morphology according to software specified quality checks. The final stage of generating a STR profile is to assign specific alleles to the amplified PCR products with one peak assigned at each marker as it is a hemizygous chromosome. Each peak above the threshold rfu (50 rfu's) in the profile is given a number. This number is a description of the structure of the allele (Goodwin *et al.* 2007).

The incorporation of an allelic ladder that contains all the common alleles found in a population at each locus has become common practice in the forensic community. Allelic ladders allow the analyst to assign alleles with more precision as slight variations can occur between runs due to ambient temperature fluctuations. The allelic ladder for the UWC 10plex was constructed according to Hill *et al.* (2008).

Complex Y-STRs can be found in the form of multi-copy loci which is found in the UWC 10plex kit. An example of a multi-copy locus is the DYS385 marker which shows two male-specific PCR products after amplification. It occurs in two inverted regions of the Y-chromosome separated by about 40 kilobases (kb). It has been suggested that the repeated sequences are duplicated on the Y chromosome with identical conserved flanking sites. Because of overlapping sizes of the length variants, the origin of the alleles from either of the two loci cannot be unambiguously determined (Kayser *et al*. 1996). The observed fragments are treated as genotypes and the alleles are recorded by being separated by a hyphen "DYS385 11-14" (Gusmão *et al.* 2005).

DYS710 is not a multi copy locus although it does at times yield a second peak 22bp larger than the first peak (D'Amato *et al.* 2011). The second product can be used as a secondary reference to the allele present in the individual profiles and aid in the correct allele designation. The structure of the DYS710 locus is complex; it is composed of two variable tetranucleotide repeats with an intervening stretch of dinucleotide repeats $(AAAG)_n(AG)_n(AAAG)_n$. The alleles at this locus present sequence repeat variation both at the tetra and dinucleotide stretches. This feature could confer difficulty in the allele identification process. If two peaks are present, with the 2nd peak 22bp apart and accounted for, then the first peak is designated as the correct allele (D'Amato *et al.* 2011).

Out of the 104 male samples there were 91 full profiles (*i.e*. a peak at each marker) obtained for the 10plex. Allele designations for these full profiles are represented in table 3.7. In the allele designation table only one allele number is reflected under the DYS710 marker and the homo- or heterozygous DYS385 products are separated into two columns. Three samples were excluded from further statistical analysis based on the presence of artefacts as explained in section 3.2.1.2.

Table 3.7. Allele designation for the full profiles in the UWC 10plex (N = 91)

	DYS710	DYS518	DYS385	DYS385	DYS644	DYS612	DYS626	DYS504	DYS481	DYS447	DYS449
GRI_01	30.2	37	14	11	14	31	27	15	24	24	30
GRI_02	32.2	40	11	11	14	33	28	15	22	25	30
GRI_03	34	35	11	14	16	35	25	16	20	25	28
GRI_05	30	36	15	15	16	32	24	14	27	26	27
GRI_06	36	34	11	14	17	34	24	18	24	25	28
GRI_07	29	35	15	16	16	26	24	15	25	24	33
GRI_08	28	34	14	14	16	27	26	15	26	24	37
GRI_09	32	34	15	16	24.4	32	20	16	25	27	30
GRI_10	35	36	16	16	24.4	31	21	13	26	26	32
GRI_11	36	42	16	20	25.4	22	23	14	25	21	30
GRI_13	32.2	35	16	17	22.4	29	25	13	28	25	27
GRI_16	33.2	34	11	14	17	29	24	17	21	25	28
GRI_22	30	35	15	15	16	32	24	14	27	21	27
GRI_30	30.2	36	15	15	16	25	28	17	25	22.4	30
GRI_31	34.2	34	11	14	17	33	23	17	22	25	29
GRI_36	35.2	36	12	13	16	34	24	18	20	25	27
GRI_39	37.2	30	14	14	16	27	30	17	19	24	24
GRI_40	30.2	37	19	19	22.4	29	25	13	28	25	30
GRI_41	34	38	15	15	22.4	31	26	12	24	25	28
GRI_45	34.2	34	11	14	17	27	23	17	22	25	29
GRI_46	37.2	36	14	15	17	33	29	13	19	22	29
GRI_52	31.2	33	12	13	22.4	26	16	13	23	26	28
GRI_53	36	36	12	13	16	34	24	18	22	25	27
GR_55	30.2	37	14	14	22.4	29	25	13	28	25	30
GRI_56	31.2	36	16	16	17	30	27	17	25	23	27
GRI_57	33.2	36	16	17	22.4	29	24	13	28	26	27

	DYS710	DYS518	DYS385	DYS385	DYS644	DYS612	DYS626	DYS504	DYS481	DYS447	DYS449
GRI_59	35	34	11	14	17	32	24	15	22	25	29
V2	33.2	36	16	17	22.4	29	25	13	25	24.4	27
V4	34	38	16	16	22.4	31	26	12	22	22.4	28
5V	31.2	32	14	14	16	28	29	15	21	22	31
V10	31.2	32	14	14	16	28	29	15	23	21	31
V13	31.2	31	14	14	16	26	31	15	20	22	32
V15	35	35	14	19	15	29	28	15	19	21	26
V17	31.2	32	14	14	16	28	29	15	21	22	31
V21	31.2	35	15	15	16	32	24	15	22	21	28
V23	33.2	35	15	18	15	31	29	15	25	23	32
VGRI_33	33	35	13	14	17	30	29	17	25	20	28
VGRI_036	40	33	13	15	15	31	29	16	26	25	29
VGRI_038	35.2	33	12	15	17	30	27	18	24	23	29
VGRI_040	35.2	33	11	13	17	30	26	17	23	25	29
VGRI_042	33	34	16	17	24.4	32	22	16	24	27	32
VGRI_048	30.2	35	10	17	15	35	26	14	23	26	28
VGRI_053	33.2	36	17	19	21.4	28	29	12	27	21	36
VGRI_055	32	33	14	15	15	28	29	17	25	26	29
VGRI_056	34	37	16	16	22.4	30	28	12	24	25	33
VGRI_057	33	37	17	18	21.4	28	28	11	24	27	32
VGRI_060	34.2	33	12	16	14	34	27	18	24	23	30
VGRI_061	30.2	35	15	19	22.4	28	27	13	28	25	31
VGRI_062	33.2	36	15	18	21.4	29	28	13	26	26	32
VGRI_063	30	31	14	14	16	29	29	15	22	23	33
VGRI_064	33	33	11	14	17	31	27	17	22	25	29
VGRI_065	30.2	36	15	18	22.4	28	27	13	28	25	29
VGRI_066	35.2	34	13	16	13	31	27	15	26	26	25
VGRI_067	34.2	34	11	14	17	28	26	17	22	25	29

	DYS710	DYS518	DYS385	DYS385	DYS644	DYS612	DYS626	DYS504	DYS481	DYS447	DYS449
VGRI_068	31	33	13	17	15	29	27	15	24	26	30
VGRI_073	31	35	14	15	18	31	29	16	22	22	32
VGRI_074	36	36	11	15	17	31	25	17	22	23	30
VGRI_076	38	32	14	16	15	33	28	16	27	23	29
VGRI_078	30.2	35	**15,16**	**18**	22.4	26	26	13	30	25	31
VGRI_082	36	33	11	14	17	31	27	16	22	25	30
VGRI_084	32	38	11	12	23.4	33	29	12	25	26	28
VGRI_088	32	37	11	11	23.4	30	29	12	24	26	34
VGRI_089	34	37	16	16	22.4	30	28	12	24	25	29
VGRI_090	35	34	11	14	17	31	26	16	22	25	31
VGRI_093	32	36	11	12	23.4	32	30	12	25	26	35
VGRI_095	32	33	11	14	16	32	27	19	22	25	29
VGRI_098	38.2	35	12	15	24.4	24	30	13	25	28	30
VGRI_101	32.2	35	15	17	22.4	27	26	13	28	25	26
VGRI_105	34.2	33	11	14	17	33	25	17	22	26	36
VGRI_106	32	38	11	11	23.4	34	29	12	26	25	31
VGRI_107	34	37	16	16	22.4	30	28	12	24	25	29
VGRI_112	32.2	33	11	13	16	31	26	15	25	25	30
VGRI_114	33	34	15	16	24.4	29	22	16	25	27	28
VGRI_115	35	31	11	14	17	30	26	20	23	25	29
VGRI_116	33.2	33	11	14	17	29	26	17	24	25	30
VGRI_117	34.2	32	15	15	15	29	25	16	26	26	31
VGRI_118	35	35	**9,11**	**15**	17	33	26	16	21	24	29
VGRI_121	32.2	30	15	16	22.4	29	27	18	26	24	29
VGRI_125	34	34	11	14	17	27	26	18	22	25	29
VGRI_127	34.2	34	11	14	16	31	26	17	22	24	31
VGRI_132	28	36	14	14	14	31	28	17	21	23	31
VGRI_133	34.2	36	15	16	22.4	26	26	13	28	25	29

	DYS710	DYS518	DYS385	DYS385	DYS644	DYS612	DYS626	DYS504	DYS481	DYS447	DYS449
VGRI_134	33.2	37	16	17	22.4	31	26	13	27	28	27
VGRI_139	30.2	35	16	16	22.4	28	27	13	27	25	32
VGRI_152	33	31	15	15	16	31	29	18	21	22	29
VGRI_153	33	31	15	15	16	31	29	18	21	22	29
VGRI_154	34.2	35	12	13	17	32	27	16	25	24	37
VGRI_155	35.2	34	15	18	15	30	27	13	23	26	29
VGRI_165	32	32	12	15	16	33	27	16	22	24	29
VGRI_166	35	33	12	14	16	29	27	17	22	24	29
VGRI_171	35	33	**9,11**	**14**	17	33	26	17	22	24	28

Table 3.8. Shared haplotypes for the Griqua male population using the UWC 10plex

	DYS710	DYS518	DYS385	DYS644	DYS612	DYS626	DYS504	DYS481	DYS447	DYS449
V5 and V17	31.2	32	14	16	28	29	15	21	22	31
VGRI_89 and VGRI_107	34	37	16	22.4	30	28	12	24	25	29
VGRI_152 and VGRI_153	33	31	15	16	31	29	18	21	22	29

3.2.1.1. Shared haplotypes

Some haplotype sharing is to be expected in any population group even though the commercial kit design is done in such a way that the combination of markers used can discriminate between closely related individuals. Commercial companies look at the dispersal of alleles in various population groups before assigning markers for their kit. There are several population databases that are accessible to the general public which assess the validity of a commercial kit on various population groups from around the world. Certain commercial markers display poor results in African population groups, which is why the UWC 10plex was designed. Ideally each individual should have a unique haplotype or pattern of alleles for the ten loci in the UWC 10plex.

There were three shared haplotypes in the Griqua male population using the UWC10plex. Samples V5 and V17 are two unrelated males from Vredendal. Samples VGRI_89 and VGRI_107 are paternal cousins whose forefathers originate from the Kranshoek population group. Samples VGRI_152 and VGRI_153 are paternal cousins from Johannesburg. It can be said that the UWC 10plex has an acceptable discrimination capacity, with only 3 shared haplotypes *i.e*. 85 unique haplotypes in a sample group of 88 males. Based on these statistics, the probability of randomly matching two males' haplotypes within the population is 3/88 (3.4%). Future studies comparing haplotypes generated using the UWC 10plex on male relatives (*e.g*. paternal cousins, father-son pairs and brothers) might serve as a useful tool to test the robustness of the multiplex.

3.2.1.2. Analysis of tri-allelic profiles

Three profiles (VGRI_78, VGRI_118 and VGRI_171) contained a tri- allele profile at DYS385 (highlighted in pink on table 4.1). The presence of three or more prominent peaks at one or more loci is used as an indicator for a mixture. At a single locus, a sample containing DNA from two sources can exhibit three peaks due to one of these genotype combinations: heterozygote + heterozygote (one overlapping allele), heterozygote + homozygote (no overlapping alleles - genotypes are unique).

According to literature, re-extraction and re-amplification is necessary to resolve these tri-alleles and rule out contamination in a single source sample (Butler, 2005). These three unrelated male samples continued to display the tri-allele haplotype after these steps were taken indicating that they are due to a mutational event. These three haplotypes were excluded from any further statistical analysis using Genepop and Arlequin. A comparison of common tri-allelic patterns and ways to resolve them can be found at http://www.cstl.nist.gov/strbase/tri_tab.htm. The profile of 2 peaks of similar height (180 rfu's) and one larger (250 rfu's) depicted in figure 3.12a is an electropherogram showing tri-alleles for sample VGRI_118. Figure 3.12b displays a single peak and Figure 3.12c displays two peaks at the multi-copy locus DYS385.

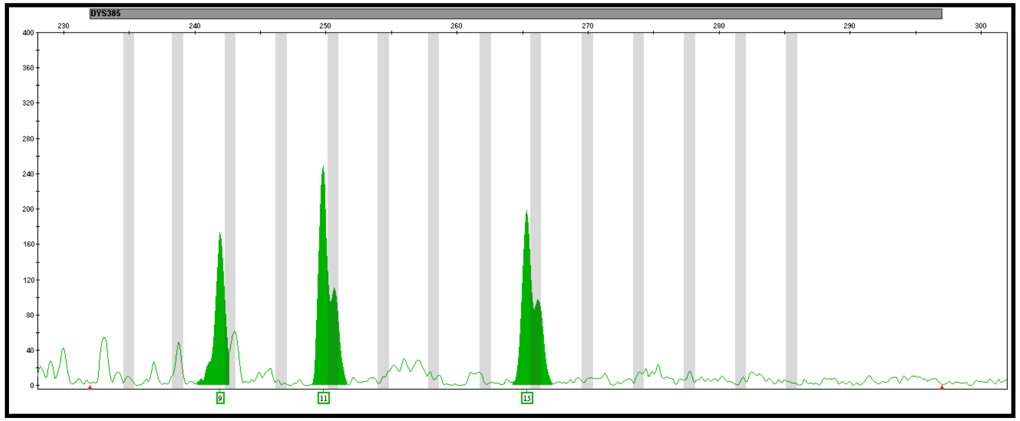


Figure 3.12a An electropherogram of marker DYS385 for sample VGRI_118 depicting a tri-allele profile. The presence of split peaks indicates incomplete adenylation which could be due to a DNA concentration that is too high, therefore requiring more PCR components (SuperTherm gold DNA polymerase; dNTPs) and time to complete adenylation of all three products.

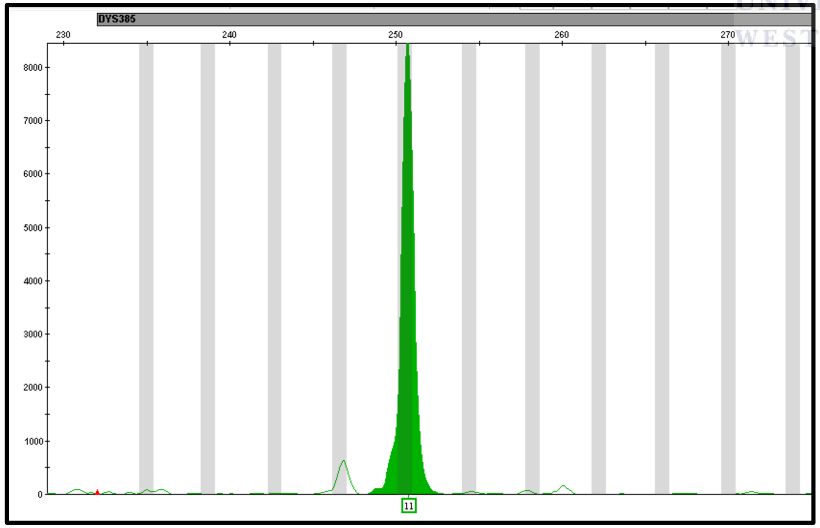


Figure 3.12.b An electropherogram of marker DYS385 for sample VGRI_106 depicting a single peak. One clear peak is recorded with a stutter product 4bp away with a peak height less than 15% of the true peak.

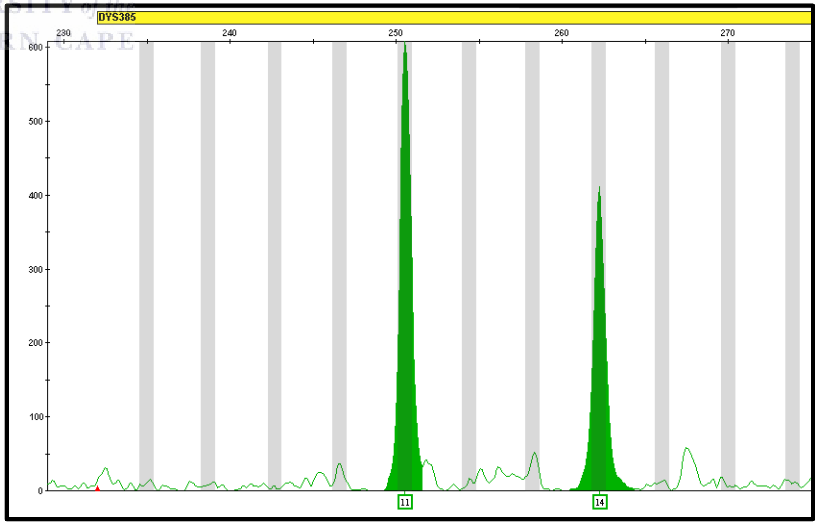


Figure 3.12c An electropherogram of marker DYS385 for sample VGRI_64 depicting two peaks. The imbalanced peak height with preferential amplification of the smaller allele is acceptable because both alleles are above the threshold height of 50 rfu's and they are within a 50% peak height ratio.

3.2.2. Statistical analysis of Y-STR data

The Y-STR mutations that occur in individuals of African descent are not accommodated for in the previously discussed commercial kits. The Griqua's are an indigenous population group of South Africa therefore a Y-STR multiplex that successfully differentiates between male individuals of this population is necessary. The UWC 10plex has already shown a high discrimination capacity on individuals from the Cape Coloured, Xhosa and Zulu population groups of South Africa (D'Amato *et al.* 2011) and a preliminary study (Heynes 2012) using the UWC 10plex on a small component of the Griqua sample group yielded favourable results.

Statistical analysis of the Y-STR data using various mathematical formulae is necessary in order to compare results between studies. Analysis of allelic frequency, haplotype diversity, discrimination capacity and STR genetic diversity are discussed below. Allelic frequency is the percentage occurrence of an allele within a specific locus of the multiplex. A higher allelic frequency indicates more of a probability for an individual from that population having that specific allele. The highest allelic frequency for the UWC 10plex is at DYS447 with a frequency of 37.5% (0.375) This means that nearly 40% of the population has allele 25 at the DYS447 locus. On its own, this locus would serve as a poor indication of male population diversity, but it is acceptable when allelic frequencies from multiple unlinked loci are taken into consideration. One can predict the most likely haplotype using the highest allele frequency for each marker in a multiplex. Table 3.10 lists the allelic frequencies for this population group.

Data generated in a preliminary study on a portion of the Griqua population confirmed developmental validation findings that the UWC 10plex is better suited to African populations. In the preliminary study (Heynes 2012) a set of commercial markers, the Y filer (extended haplotype and 6-Y-Plex: 17 markers), was tested on Kokstad and Vredendal Griqua samples alongside the UWC 10plex. The results showed more shared haplotypes between individuals when using the commercial marker set in comparison with the UWC 10plex.

The gene diversity for the Y-filer commercial kit (0.654) was significantly lower than the UWC 10plex (0.838). This indicates several shared alleles and difficulty in discriminating between individuals when using this multiplex. The commercial multiplex was designed based on discriminatory Y-STR mutations in European and American populations. The inclusion of Y-STR loci that successfully discriminate between individuals in an African population would increase the effectiveness of this kit on the Griqua population. There is a growing necessity for rapidly mutating Y-STRs to be incorporated into commercial kits in order for them to have a wider application basis and higher success rate.

The multiplex performance values for the UWC 10plex on the Griqua population are in Table 3.9. Three profiles were excluded based on their tri-allelic pattern at DYS385, as explained in section 3.2.1.2, which is why the number of haplotypes is lower than the number of profiles generated.

Table 3.9 Forensic values for the UWC 10plex on Griqua male samples (N=91)

Multiplex	No. of full profiles	No. of Haplotypes	No. of unique haplotypes	Discrimination capacity	Haplotype diversity
UWC 10plex	91	88	85	0.966	0.987

Due to the lack of recombination between Y-specific loci, the whole haplotype is transmitted as a single marker, and haplotype diversity defined by a set of STRs must be established by frequency estimates of the whole haplotype. The haplotype diversity cannot be predicted by combining the average diversity at each single locus. There are two main factors that contribute to single-locus diversity, namely the presence of distinct lineages (several lineages in different combinations may be population-specific) and the variation accumulated within each lineage by mutation. Accrued mutations will contribute to the decrease in association between alleles at different loci and therefore be reflected in the Y-STR diversity at the haplotype level. (Gusmão *et al.* 2005)

Haplotype diversity (HD) across the population was computed using a formula that takes into consideration the sample size and the relative haplotype frequencies. The haplotype diversity value of 0.9873 is significant enough for the UWC 10plex to be used when analysing the Griqua population group.

The Discrimination Capacity (DC) is the ability to differentiate between two individuals in a population. It was determined by dividing the number of unique haplotypes seen in a given population by the total number of samples from that population. An ideal DC value of 1 indicates a unique haplotype for each individual in the population. The DC value for the UWC 10plex in the Griqua population (0.9659) is accepted in the forensic community, when compared to current commercial forensic kits with an average DC value above ±0.85 (Ballantyne *et al.* 2012).

STR genetic diversity or the probability that two alleles chosen at random are different, was then calculated. Analysis of the gene diversities for the UWC 10plex indicates that the highest gene diversity of 0.93887 was at DYS710 and the lowest gene diversity value was 0.80643 at DYS447. For the DYS447 marker, allele 25 was shared between more than a quarter of the individuals in the population. The range of alleles within the 10 markers was between 9 and 19, with 19 different alleles present at DYS710 for the Griqua population group. This results in high genetic diversity for this marker which is a strong indicator of its value in distinguishing between individuals of this population group. Table 3.10 shows the allelic frequency and gene diversity values (at the base of each allele) for the multiplex as well as the average gene diversity of 0.86 ± 0.04. This indicates that the allelic frequencies over this small marker range and substantial population size (N = 88 males) shows a considerate amount of gene diversity and variation. The UWC 10plex kit can be utilised for effective individual identification of male individuals in the Griqua population.

Table 3.10. Allele frequency and gene diversity values for UWC 10plex samples (N=88 males)

	10plex									
	DYS710	**DYS518**	**DYS385**	**DYS644**	**DYS612**	**DYS626**	**DYS504**	**DYS481**	**DYS447**	**DYS449**
10			0.011							
11			0.273				0.011			
12			0.102				0.114			
13			0.034	0.011			0.193			
14			0.170	0.045			0.045			
15			0.227	0.102			0.182			
16			0.148	0.250		0.011	0.136			
17			0.023	0.227			0.193			
18				0.011			0.102			
19			0.011				0.011	0.034		
20						0.011	0.011	0.034	0.011	
21						0.011		0.068	0.068	
21.4				0.034						
22					0.011	0.023		0.227	0.080	
22.4				0.205					0.023	
23						0.034		0.068	0.091	
23.4				0.045						
24					0.011	0.114		0.148	0.102	0.011
24.4				0.057					0.011	
25					0.011	0.091		0.170	0.375	0.011
25.4				0.011						
26					0.045	0.170		0.091	0.170	0.023
27					0.057	0.193		0.068	0.045	0.102
28	0.023				0.114	0.114		0.091	0.023	0.125
29	0.011				0.170	0.182				0.273

					10plex					
	DYS710	**DYS518**	**DYS385**	**DYS644**	**DYS612**	**DYS626**	**DYS504**	**DYS481**	**DYS447**	**DYS449**
30	0.034	0.023			0.114	0.034				0.159
30.2	0.091									
31	0.023	0.057			0.205	0.011				0.102
31.2	0.080									
32	0.091	0.068			0.102					0.091
32.2	0.057									
33	0.080	0.159			0.080					0.034
33.2	0.091									
34	0.080	0.170			0.057					0.011
34.2	0.102									
35	0.068	0.170			0.023					0.011
35.2	0.057									
36	0.057	0.182								0.023
37		0.102								0.023
37.2	0.023									
38	0.011	0.045								
38.2	0.011									
39										
40	0.011	0.011								
41										
42		0.011								
Gene Diversity	0.93887	0.87226	0.82001	0.83412	0.88715	0.87304	0.85763	0.87461	0.80643	0.86207

*Average gene diversity 0.86± 0.04

3.2.3. Population comparison using the UWC 10plex

The 88 male Griqua samples were separated into an Origin and Post-colonial group to compare forensic parameters and overall multiplex performance. The origin sample group includes samples from the Western and Northern Cape, while the post-colonial sample group included individuals whose ancestors participated in the Griqua Trek (Griquastad, Kokstad and Kranshoek samples). The origin sample group (N = 49) consisted of 47 unique haplotypes which resulted in a discrimination capacity of 0.959. While the post-colonial sample group (N = 39) only had one shared haplotype between a set of paternal cousins. This resulted in a discrimination capacity of 0.974.

From these results it is clear to see that individual haplotype assignment to the majority of males in both sample groups was possible. AMOVA statistical analysis using Arlequin V3.5.1.2 (Excoffier *et al.* 2010) shows that the two sample groups share a similar population composition. Through haplotype assessment it seems that there is more allelic variation present within the post-colonial group.

Table 3.11. Analysis of molecular variance (AMOVA) of the Y-STR data for the two Griqua sample groups.

Source of variation	d.f.	Sum of squares	Variance components	Percentage of variation
Among populations	1	0.547	0.00110Va	0.22
Within populations	86	42.908	0.49893Vb	99.78
Total	87	43.455	0.50003	

Fixation index (FST) : 0.00220 P-value = 1.00000+-0.00000

The Haplotype diversity value for the origin group was 0.957152 and 0.971797 for the post-colonial group. While the average gene diversity when using the UWC 10plex was 0.821 and 0.852 for the origin and post-colonial groups respectively. There was low allelic diversity at a number of the markers (DYS644, DYS481 and DYS447) with nearly 40% of both sample groups possessing 25 repeats of allele DYS447 (figure 4.2). A multiplex consisting of a combination of these markers with low mutation rates will result in a large number of shared haplotypes for the sample group. The inclusion of the complex DYS710 marker and DYS626 however introduce enough variety to ensure acceptable forensic parameters when using this multiplex on the Griqua population.

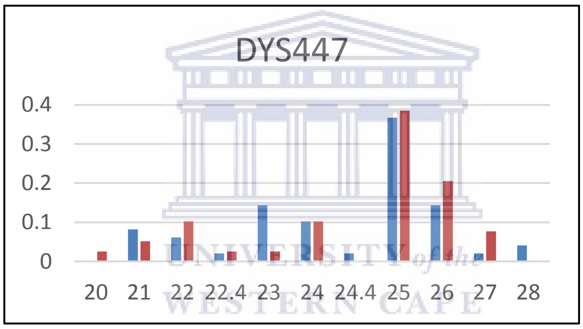
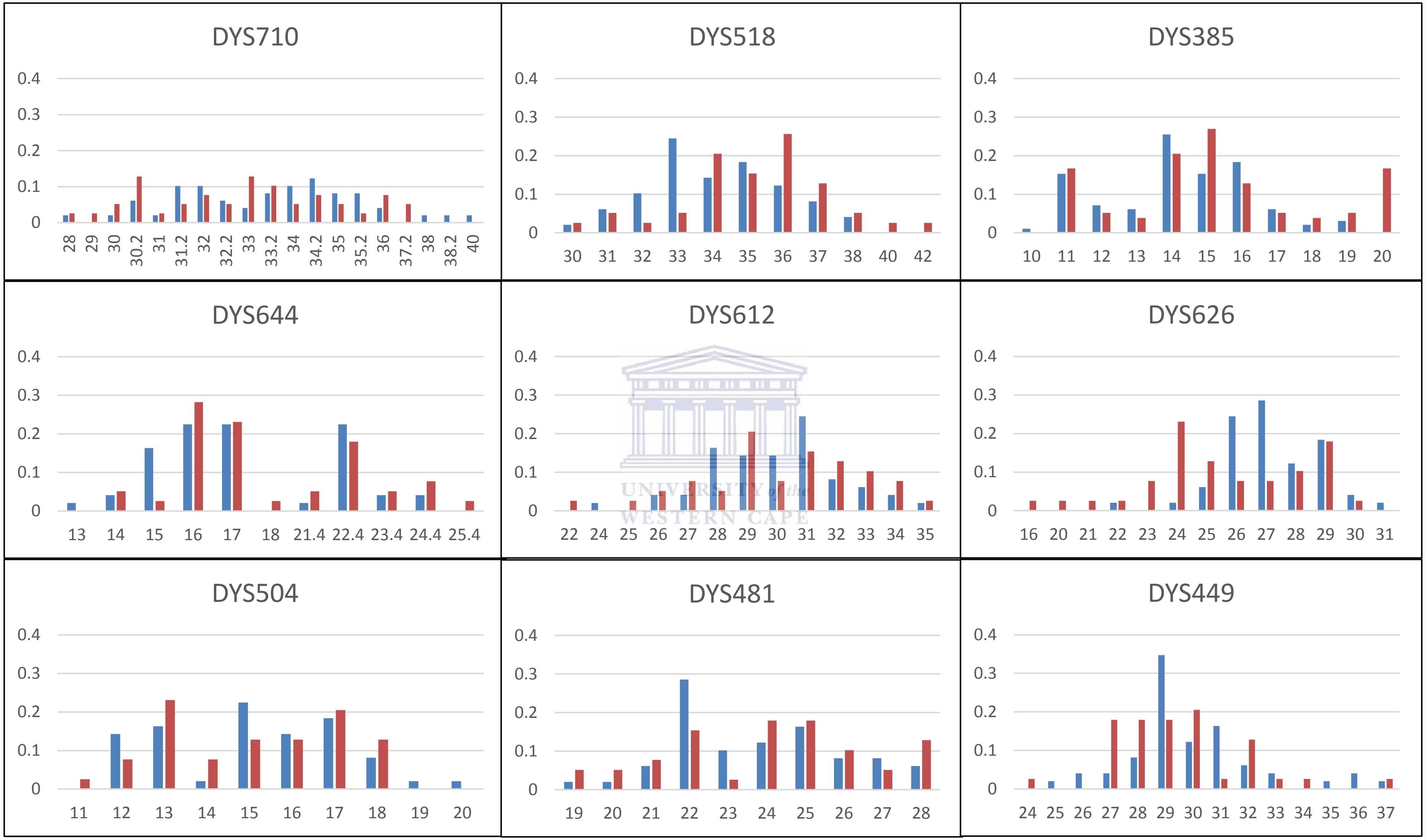


Figure 3.13. Allelic frequency comparison between the origin and post-colonial group for the DYS447 marker. Blue = Origin and Red = Post-colonial sample group.

Figure 3.14. Allelic frequency comparison between the origin (blue) and post-colonial (red) sample groups for 9 of the 10 UWC 10plex markers.



3.3. Future work

Whole mitochondrial genome sequencing using a NGS platform is the next step in maternal haplogroup assignment using mtDNA. There is also a recently published alternate semi-automated analysis software for haplogroup assignment – EMMA. The availability of PhyloTree (www.phylotree.org/ 2014), a widely accepted phylogenetic tree of human mitochondrial DNA lineages, led to the development of several (semi-)automated software solutions for haplogrouping *e.g*. HaploGrep (Kloss-Brandstätter *et al.* 2011) MitoTool (Fan *et al.* 2011) HmtDB (Rubino *et al.* 2012) and mtDNAoffice (Soares *et al.* 2012).

These haplogrouping tools provide inconsistent and at times inaccurate data interpretation results from control region sequences as they only make use of haplogroup-defining mutations. A new concept for estimating mitochondrial DNA haplogroups using a maximum likelihood approach - EMMA (Röck *et al.* 2013) contributes substantial value for quality control in population genetics. It uses PhyloTree (www.phylotree.org/ 2014), along with a quality-controlled database consisting of 14,990 full mtGenomes downloaded from GenBank as a reference for haplogroup estimates. A maximum likelihood estimation of the stability of mutations based on 19,171 full control region haplotypes is also incorporated. EMMA utilises an algorithm combining the reference databases with empirically determined fluctuation rates to estimate the haplogroup of a mtDNA sequence. This tool provides precise haplogroup assignments without the need to incorporate coding region mutations.

The UWC 10plex Y chromosome STR kit produced viable results in terms of its ability to discriminate between male Griqua individuals. There were however problems encountered with sample extraction and quantification. A few suggestions are permissible, to avoid future problematic encounters. Investing in a genomic DNA clean-up step after extractions should produce a higher concentration of DNA free from PCR inhibitors. The use of an alternative quantification method would aid in accurate Y-STR genotyping by more accurately measuring the DNA concentration as well as its purity. The Quantifiler Quant DUO kit on the 7500 RT-

PCR instrument uses fluorescent tags on specific fragments of male human DNA to assess its concentration and quality.

The Griqua population data generated can be compiled into a population database that provides Y-STR information on Griqua individuals from throughout South Africa, including direct descendants of the Le Fleur and Kok clans. Data comparisons with other indigenous populations (Xhosa, Zulu, Cape Coloured *etc*.) using the UWC 10plex might yield interesting results in terms of distribution patterns of Y-STRs, marker robustness and allelic population indicators. Intra- and inter-population statistics on samples from throughout South Africa can be calculated using AMOVA. This data presents an interesting line of analysis to assess diversity in South Africa and how populations can be linked in various regions.

The incorporation of RM Y-STRs into the UWC 10plex kit increased its ability to discriminate between individuals and gave it a performance advantage over commercial Y-STR kits at the time. A comparative study with current RM Y-STR commercial kits, such as the PPY23 from Promega and AmpFℓSTR® Yfiler Plus™ from Applied Biosystems, represents itself as a viable option to truly test the discriminatory power of the UWC 10plex kit.

The usage of this male specific kit in comparison to commercial autosomal kits for HID studies in South Africa can also be assessed in a follow up study. The currently used autosomal commercial kits in South Africa contain between 10 (AmpFℓSTR® Profiler Plus) and 24 (Globalfiler®) markers from Thermo Fisher.

A follow up study for the Griqua population samples could involve Y chromosome haplogroup assignment using HRM technology. A comparison of the Y chromosome haplogroup assigned to the Griqua males with their corresponding mtDNA haplogroup can provide vast insight into the migration patterns of their ancestors.

4. Conclusion

The Griqua population of South Africa has a strong cultural heritage which has been passed downed from generation to generation even though they are not a politically recognised population group in South Africa. The story of their ancestors struggle for recognition throughout the centuries is not so different from other indigenous or native population groups. Genetic classification in this modern era is the method that the clan leaders have turned to in order to claim back their ancestral land. This Masters project presented itself as an anthropological venture; a way to explore the ancient maternal substructure of the Griqua population in South Africa. The results do not carry any political weight. The large collection of samples meant that all the male participants could also be included in a validation study using the UWC 10plex – a discriminatory Y-chromosome STR multiplex.

Individuals who identified themselves as Griqua due to their ancestry were included in the study. Each individual that was sampled was interviewed by lab personnel to verify Griqua heritage. The types of polymorphisms used to analyse the Griqua population of South Africa provided favourable genetic data in terms of fulfilling objectives. The ancestry informative mtDNA SNPs with their low mutation rates provided a genetic confirmation of the historical Griqua trek. While the pattern of Y-STRs found in Griqua males made individuals discernible using the UWC 10plex markers.

STR mutations are highly polymorphic tandem repeats and represent fairly recent changes in genetic code. The implication thereof is that alleles can differ between two related males separated by a few generations. This is highly dependent on the mutation rate of the marker in question and the multiplex in which it is included. Y-STR multiplex testing was done to assess the validity of the UWC 10plex kit on samples with African heritage. Commercial kits do not incorporate population data from South African populations during marker selection and development of their allelic ladders. The UWC 10plex addresses this issue and presents a viable HID test for males with South African heritage which include colonial as well as Bantu influences.

Studying polymorphisms in the control region of the mitochondrial genome provides a means for studying ancient geographical roots for a population. This study provided an apt assessment of the possible L0d sub-haplogroup structure present within an indigenous (origin) versus admixed (post-colonial) population group. It provides a strong case for ancestral changes that may be present in closely located sample groups such as the Griqua population in South Africa.

The mtDNA control region data displayed a clear pattern when comparing the origin and post-colonial haplogroups. A larger incorporation of colonial and African L1 – L5 haplogroups in the post-colonial or trek population group supports the hypothesis of greater genetic drift. While the L0 ancestry component in both groups confirms their shared maternal origins. Movement and mixture in the Griqua population is more prevalent in the post-colonial sample group with the isolated incorporation of a number of maternal lineages. While the individuals from the origin group display a less diverse gene pool and very little incorporation from Eurasian haplogroups.

After interviewing several Griqua individuals at the GNC it became apparent that there is an element of rural living which keeps exposure to other populations minimal and the gene pool remains small. Older members of the Griqua society choose to remain in their place of birth, surrounded by a strong Griqua cultural identity. Their forefathers were the pioneers in areas such as Griquastad, Kokstad and Kranshoek which is why they chose to remain in these areas. Several colonies were set up throughout South Africa based on land availability and work and they are still functional in modern society. We are however moving towards a more connected society due to the expansion of technology and the new generation of Griqua's will most likely become genetically integrated with the greater South African population. SNP mutations of the mtDNA control region occurred thousands of years ago and are related to the geographic movement of ancient populations. Just as there were population expansions in the past due to trade routes it is now more common for individuals to migrate to other countries and there is an explosion of mixed cultures and societies. One could even say that there is a global movement towards admixed populations.

Reference list

Achilli A; Rengo C; Magri C and Torroni A (2004). *The Molecular Dissection of mtDNA Haplogroup H Confirms that the Franco-Cantabrian Glacial Refuge was a Major Source for the European Gene Pool.* American Journal of Human Genetics 75 pp. 910 – 918.

Anderson S; Bankier A; Barrell B; De Bruijn M; Coulson A; Drouin J; Eperon I; Nierlich D; Roe B; Sanger S; Schreier A; Smith A; Staden R and Young I (1981). *Sequence and Organization of the Human Mitochondrial Genome.* Nature 290 pp. 457 – 465.

Avise J; Bowen B and Lamb T (1989). *DNA Fingerprints from Hypervariable Mitochondrial Genotypes.* Molecular biology and evolution 6 pp. 258 – 269.

Bacolla A; Larson J; Collins J; Li J; Milosavljevic A; Stenson P; Cooper D and Wells R (2008). *Abundance and Length of Simple Repeats in Vertebrate Genomes are Determined by Their Structural Properties.* Genome Research 18 pp. 1545 – 1553.

Ballantyne K; Goedbloed M; Fang R; Schaap O; Lao O; Wollstein A; Choi Y; Duijn K; Vermeulen M; Brauer S; Decorte R; Poetsch M; Wurmb-Schwark N; de Knijff P; Labuda D; Vézina H; Knoblauch H; Lessig R; Roewer L; Ploski R; Dobosz T; Henke L; Henke J; Furtado M and Kayser M (2010). *Mutability of Y-Chromosomal Microsatellites: Rates, Characteristics, Molecular Bases, and Forensic Implications.* The American Journal of Human Genetics 87 pp. 341 – 353.

Ballantyne K; Keerl V; Wollstein A; Choi Y; Zuniga S; Ralf A; Vermeulen M; de Knijf P & Kayser M (2012). *A new future of forensic Y-chromosome analysis: Rapidly mutating Y-STRs for differentiating male relatives and paternal lineages.* Forensic Science International: Genetics 6 pp. 208 – 218s.

Ballantyne K; Arwin R; Aboukhalid R *et al.* The International RM Y-STR study group (2014). *Toward Male Individualization with Rapidly mutating Y-chromosomal Short Tandem Repeats.* Human mutation pp. 1021 – 1032.

Balson, S (2007). *Children of the mist: The Lost Tribe of South Africa.* Brisbane: Interactive Publications Pty. Ltd.

Bandelt H-J; Forster P and Röhl A (1999) *Median-Joining Networks for Inferring Intraspecific Phylogenies.* Molecular Biology Evolution 16 pp. 37 – 48

Barbieri C; Vicente M; Rocha J; Mpoloka S; Stoneking M; and Pakendorf B (2013). *Ancient Substructure in Early mtDNA Lineages of Southern Africa.* The American Journal of Human Genetics 92 pp. 285 – 292.

Barbieri C; Güldemann T; Naumann C; Gerlach L; Berthold F; Nakagawa H and Pakendorf B (2014). *Unravelling the Complex Maternal History of Southern African Khoisan Populations.* American journal of physical anthropology 153 pp. 435 – 448.

Barni F; Salata E; Rapone C; Kline M; Lago G; Butler J and Berti A (2007). *Characterisation of Novel Alleles and Duplication Events in the STR DYS19, DYS439, DYS389II and DYS385 loci.* Promega Identity available at www.promega.com/geneticidentity/

Behar D; van Oven M; Rosset S; Metspalu M; Logvali E-L; Silva N; Kivisild T; Torroni A; and Villems R (2012). *A "Copernican" Reassement of the Human Mitochondrial DNA Tree from its Root.* The American Journal of Human Genetics 90 pp. 675 – 684.

Besten, M (2006). *Transformation and Reconstitution of Khoe-San Identities: AAS Le Fleur I, Griqua Identities and Post-Apartheid Khoe-San Revivalism (1894-2004)*.

Bini C; Ceccardi S; Luiselli D; Ferri G; Pelotti S; Colalongo C; Falconi M and Pappalardo G (2003). *Different Informativeness of the Three Hypervariable Mitochondrial DNA Regions in the Population of Bologna (Italy)*. Forensic science international 135 pp. 48 – 52.

Budowle B; Bieber F and Eisenberg A (2005). *Forensic Aspects of Mass Disasters: Strategic Considerations for DNA-based Human Identification.* Journal of Legal Medicine pp. 230 – 243.

Butler J (2003).*Recent Developments in Y-Single Tandem Repeat and Y-Single Nucleotide Polymorphism Analysis*. Forensic Science Reviews 15 pg 91

Butler J and Coble M (2007). *STRs vs. SNPs: Thoughts on the Future of Forensic DNA Testing.* Journal of Forensic Science and Medical Pathology 3 pp. 200 – 205.

Butler J; Kline M and Decker A. (2008) *Addressing Y-Chromosome Short Tandem Repeat Allele Nomenclature.* Journal of Genetic Genealogy.

Butler J and Hill C (2012) *Biology and Genetics of New Autosomal STR Loci Useful for Forensic DNA Analysis.* National Institute of Standards and Technology.

Carracedo A; Bär W; Lincoln P; Mayr W; Morling N; Olaisen B; Schneider P; Budowle B; Brinkmann B; Gill P; Holland M; Tully G and Wilson M (2000). *DNA Commission of the International Society for Forensic Genetics: guidelines for mitochondrial DNA typing.* Forensic science international 110 pp. 79 – 85.

Chan E; Hardie R; Petersen D; Beeson K; Bornman R; Smith A and Hayes V (2015). *Revised Timeline and Distribution of the Earliest Diverged Human Maternal Lineages in Southern Africa.* PLoS ONE 10 pp. 1 – 17.

Chen Y; Torroni A; Excoffier L; Santachiara-Benerecetti A and Wallace D. (1995). *Analysis Of mtDNA Variation in African Populations Reveals the Most Ancient of All Human Continent-Specific Haplogroups.* American journal of human genetics 57 pg 133.

Chong M; Calloway C; Klein S; Orrego C and Buoncristiani M (2004). *Optimization of a Duplex Amplification and Squencing Strategy for the HVI/HVII regions of Human Mitochondrial DNA for Forensic Casework.* Forensic Science International.

Coelho M; Sequeira F; Luiselli D; Beleza S and Rocha J (2009). *On the Edge of Bantu Expansions: MtDNA, Y chromosome and lactase persistence genetic variation in south-western Angola.* BMC Evolutionary Biology 80 pp. 1 – 18.

D'Amato ME, Bajic Vladimir B, Benjeddou M and Davison S. (2011) *Design and Validation of a Highly Discriminatory 10-locus Y-chromosome STR Multiplex System.* Forensic Science International.

Davey J; Hohenlohe P; Etter P; Boone J; Catchen J and Blaxter M (2011). *Genome-Wide Genetic Marker Discovery and Genotyping Using Next-Generation Sequencing.* Nature Reviews Genetics 12 pp. 499 – 510.

Decker A; Kline M; Vallone P and Butler J (2007). *The Impact of Additional Y-STR Loci on Resolving Common Haplotypes and Closely Related Individuals.* Forensic Science International: Genetics pp. 215 – 217.

Decker A; Kline M; Redman J; Reid T and Butler J (2008). *Analysis of Mutations in Father-Son Pairs with 17 Y-STR Loci.* Forensic Science International: Genetics 2 e31-e35.

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 pp. 479 – 491.

Excoffier L and Lischer H (2010). *Arlequin Suite version 3.5: A New Series of Programs to Perform Population Genetics Analyses under Linux and Windows.* Molecular Ecology Resources 10 pp. 564 – 567.

Fan L and Yao Y (2011). *Mitotool: A Web Server for the Analysis and Retrieval of Human Mitochondrial DNA Sequence Variations*. Mitochondrion 11 pp. 351 – 356.

Feigelson H; Rodriguez C; Robertson A; Jacobs E; Calle E; reid Y and Thun M (2001). *Determinants of DNA Yield and Quality from Buccal Cell Samples Collected with Mouthwash.* Cancer Epidemiology, Biomarkers and Prevention 10 pp. 1005 – 1008.

Finnilä S; Lehtonen M and Majamaa K (2001). *Phylogenetic Network for European mtDNA*. The American Journal of Human Genetics 68 pp. 1475 – 1484.

Goedbloed M; Vermeulen M; Fang R; Lembring M; Wollstein A; Ballantyne K; Lao O; Brauer S; Krüger C; Roewer L; Lessig R; Ploski R; Dobosz T; Henke L; Henke J; Furtado M and Kayser M (2009). *Comprehensive Mutation Analysis of 17 Y-Chromosomal Short Tandem Repeat Polymorphisms Included in the Ampflstr® Yfiler® PCR Amplification Kit.* International Journal of Legal Medicine 123 pp. 471 – 482

Goodwin W; Linacre A and Hadi S (2007). *An Introduction to Forensic Genetics.* John Wiley & Sons Ltd.

Gusmão L; Butler J; Carracedo A; Gill P; Kayser M; Mayr W; Morling N; Prinz L; Roewer L; Tyler-Smith C and Schneider P (2005). *DNA Commission of the International Society of Forensic Genetics (ISFG): An Update of the Recommendations on the Use of Y-STRs in Forensic Analysis.* Forensic Science International.

Hammer M; Chamberlain V; Kearney V; Stover D; Zhang G; Karafet T; Walsh and Redd A (2006i). *Population Structure of Y Chromosome SNP Haplogroups in the United States and Forensic Implications for constructing Y Chromosome STR Databases.* Forensic Science International 164 pp. 45 – 55.

Hammer M and Redd A (2006ii). *Forensic Applications of Y-chromosome STRs and SNPs.* U.S Department of Justice, Cumulative Technical Report 2000-IJ-CX-K006.

Herrnstadt C; Elson J; Fahy E; Preston G; Turnbull D; Anderson C; Soumitra S; Jerrold M; Loefsky M; Beal F and Howell, N. (2002). *Reduced-Median-Network Analysis of Complete Mitochondrial DNA Coding-Region Sequences for the Major African, Asian, and European Haplogroups.* The American Journal of Human Genetics 70 1152 – 1171.

Heynes K (2012). *Characterizing the Short Tandem Repeats of the Griqua Population in South Africa.* Unpublished thesis completed during UWC Biotechnology Honours.

History of the Griqua nation and Nomansland. Retrieved from the Griqua Royal house website: http://www.gwb.com.au/gwb/strachan/griqua.html/

Houck M and Siegal J (2010). *Fundamentals of Forensic Science.* Elsevier Ltd Burlington, MA second edition.

Jobling M; Pandaya A and Tyler-Smith C (1997). *The Y Chromosome in Forensic Analysis and Paternity Testing.* International Journal of Legal Medicine 110 pp. 118 – 124.

Kayser M; Caglià A; Corach D; Fretwell N; Gehrig C; Graziosi G; Heidorn F; Herrmann S; Herzog B; Hidding M; Honda K; Jobling M; Krawczak M; Leim K; Meuser S; Meyer E; Oesterreich W; Pandya A; Parson W; Penacino G; Perez-Lezaun A; Piccinini A; Prinz M; Schmitt C; Schneider P; Szibor P; Teifel-Greding J; Weichhold G; de Knijff P and Roewer L (1996). *Evaluation of Y-chromosomal STRs: a Multicenter Study.* International Journal of Legal Medicine 110 pp. 125 – 133.

Kieleczawa J and Mazaika E (2010). *Optimization of Protocol for Sequencing of Difficult Templates.* Journal of Biomolecular Techniques 21 pp. 97 – 102.

Kimpton C; Gill P; Walton A; Urquhart A; Millican E and Adams M. (1993). *Automated DNA Profiling Employing Multiplex Amplification of Short Tandem Repeat Loci*. Genome Research 3 pp. 13 – 22.

Kloss- Brandstätter A; Peterson C; Irwin J; Mpoke S; Koech D; Parson W and Parsons T (2004). *Mitochondrial DNA Control Region Sequences from Nairobi (Kenya): Inferring Phylogenetic Parameters for the Establishment of a Forensic Database.* International Journal of Legal Medicine 118 pp. 294 – 306.

Kloss-Brandstätter A; Pacher D; Schönherr S; Weissensteiner H; Binna R; Specht G and Kronenberg F (2011). *Haplogrep: A Fast and Reliable Algorithm for Automatic Classification of Mitochondrial DNA Haplogroups*. Human Mutations 32 pp. 25 – 32

Krings M; Geisert H; Schmitz R; Krainitzki H and Pääbo S. (1999). *DNA Sequence of the Mitochondrial Hypervariable Region II from the Neandertal Type Specimen.* Proceedings of the National Academy of Sciences, 96 pp. 5581 – 5585.

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

Lansing J; Watkins J; Hallmark B; Cox M; Karafet T; Sudoyo and Hammer M (2008). *Male Dominance Rarely Skews the Frequency Distribution of Y chromosome Haplotypes in Human Populations.* Proceedings of the National Academy of Sciences (PNAS) 105 pp. 11645 – 11650.

Larkin M; Blackshields G; Brown N; Chenna R; McGettigan P; McWilliam H; Valentin F; Wallace I; Wilm A; Lopez R; Thompson J; Gibson T and Higgins D (2007). *ClustalW and ClustalX Version 2*. Bionformatics 23 pp. 2947 – 2948.

Leat N; Ehrenreich L; Benjeddou M and Davison S (2004). *Developments in the Use of Y-chromosome Markers in Forensic Genetics.* African Journal of Biotechnology 3 pp. 637 – 642.

Lutz S; Weisser H; Heizmann J and Pollak S (1997). *A Third Hypervariable Region in the Human Mitochondrial D-Loop.* Human genetics 101 pp. 384 – 394.

Macaulay V; Hill C; Achilli A;Rengo C; Clarke D; Meehan W; Blackburn J; Semino O; Scozazari R (2005). *Single Rapid Coastal Settlement of Asia Revealed by Analysis of Complete Mitochondrial Genomes.* Science 308 pp. 1034 – 1036.

Mardis E. (2008). *Next-generation DNA Sequencing Methods.* Annual Review Genomics and Human Genetics 9 387 – 402.

Medrano J; Aasen E; Sharrow L (1990). *DNA Extraction from Nucleated Red Blood Cells.* Biotechniques 8 pg 43.

Merkel A and Gemmel N (2008). *Detecting Short Tandem Repeats from Genome Data: Opening the Software Black Box.* Briefings in Bioinformatics 9 pp. 335 – 366.

Metspalu M; Kivisild T; Metspalu E; Parik J; Hudjashov G; Kaldma K; Serk P; Karmin M; Behar D; Thomas M; Gilbert P; Endicott P; Mastana S; Papiha S; Skorecki K; Torroni A and Villems R (2004). *Most of the Extant mtDNA Boundaries in South and Southwest Asia were likely Shaped during the Initial Settlement of Eurasia by Anatomically Modern Humans.* BMC Genetics 26 pp. 1 – 25.

Metzker M. (2009). *Sequencing Technologies—the Next Generation.* Nature Reviews Genetics 11 pp. 31 – 46.

Mona S; Grunz K and Brauer S (2009). *Genetic Admixture History of Eastern Indonesia as Revealed by Y-chromosome and Mitochondrial DNA Analysis.* Molecular Biology Evolution 26 pp. 1865 – 1877.

Morris A; Heinze A; Chan E; Smith and Hayes V (2014). *First Ancient Mitochondrial Human Genome from a Pre-pastoralist Southern African.* Genome Biology and Evolution 6 pp. 2647 – 2653.

Murphy K; Berg K; Eshleman J. (2005). *Sequencing of Genomic DNA by Combined Amplification and Cycle Sequencing Reaction*. Clinical chemistry 51 pp. 35 – 39

Nielsen R; Paul J; Albrechtsen A and Song Y. (2011). *Genotype and SNP Calling from Next-Generation Sequencing Data*. Nature Reviews Genetics 12 pp. 443 – 451.

Petersen D; Libiger O; Tindall E; Hardie R; Hannick L; Glasshoff R; Mukerji M; Indian Genome Consortium; Fernandez P; Haacke W; Schork N and Hayes V (2013). *Complex Patterns of Genomic Admixture within Southern Africa.* PLoS Genetics 9.

Pistorius M. (2005) *Profiling Serial Killers and Other Crimes in South Africa.* Penguin books South Africa.

Purps J; Siegert S and Willuweit S *et al.* (2014). *A Global Analysis of Y-chromosomal Haplotype Diversity for 23 STR Loci.* Forensic Science International: Genetics 12 pp. 12 – 23.

Quintana-Murci L; Chaix R; Wells S; Behar D; Sayar H; Scozzari R; Rengo C; Al-Zahery N; Semino O; Santachiara-Benerecetti S; Coppa A; Ayub Q; Mohyuddin A; Tyler-Smith C; Mehdi Q; Torroni A and McElreavey K (2004). *Where West meets East: the Complex mtDNA Landscape of the south-west and Central Asian Corridor.* American Journal of Human Genetics 74 pp. 827 – 845.

Quintana-Murci L; Harmant C; Quach H; Balanovsky O; Zaporozhchenko V; Bormans C; van Helden P; Hoal E and Behar D (2010). *Strong Maternal Khoisan Contribution to the South African Coloured Population: A Case of Gender-Biased Admixture.* The American Journal of Human Genetics 86 pp. 611 – 620.

Quintas B; Alwarez-Iglesias V; Salas A; Phillips C; Lareu M and Carracedo A (2004). *Typing of Mitochondrial DNA Coding Region SNPs of Forensic and Anthropological Interest Using SNaPshot Minisequencing.* Forensic Science International 10 pp. 251 – 257.

Ravid-Amir O and Rosset S (2009). *Maximum Likelihood of Locus-specific Mutation Rates in Y-Chromosome Short Tandem Repeats.* Bioinformatics vol. 26 pp. i440 – i445.

Reed G; Kent J and Wittwer C. (2007). *High-Resolution DNA Melting Analysis for Simple and Efficient Molecular Diagnostics.* Pharmcogenomics vol. 8 pp 597 – 608.

Reuter J and Mathews D (2010). *RNAstructure: Software for RNA secondary Structure Prediction and Analysis.* BMC Bioinformatics 11 pg 129.

Rev Philip J (1828). *Researches in South Africa; Illustrating the Civil, Moral and Religious Conditions of the Native Tribes: Including Journals of the Author's Travels in the Interior; Together with Detailed Accounts for the Progress of the Christian Missions, Exhibiting the Influence of Christianity in Promoting Civilization*. James Duncan, volume 1.

Richards B; Skoletsky J; Shuber A; Balfour R; Stern R; Dorkin H; Parad R; Witt D and Klinger K. (1993). *Multiplex PCR Amplification from the CFTR Gene using DNA Prepared from Buccal Brushes/Swabs.* Human Molecular Genetics 2 159 – 163.

Röck A, Dür 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 pp. 601 – 609.

Roewer L (2009). *Y Chromosome STR Typing in Casework.* Forensic Science and Medical Pathology 5 pp. 77 – 84.

Rubino F; Piredda R; Calabrese F; Simone D; Lang M; Calabrese C; Petruzzella V; Tommaseo-Ponzetta M; Gasparre G and Attimonelli M (2012). *HmtDB, a Genomic Resource for Mitochondrion-Based Human Variability Studies.* Nucleic Acids Resources 40 D1150 – D1159.

Ruitberg C; Reeder D and Butler J (2001). *STRBase: a Short Tandem Repeat Database for the Human Identity Testing Community.* Nucleic Acids Research 29 pp. 320 – 322.

Salas A; Quintanas B; Alvarez-Iglesias V (2005). *SNaPshot Typing of Mitochondrial Coding Region Variants.* Methods in Molecular Biology 297 pp. 197 – 208.

Saunders, C and Southey, N (2001). *A dictionary of South African History*. David Philip publishers Cape Town pp. 81 – 82

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 pp. 623 – 630.

Schlebusch C and Soodyall H. (2012). *Extensive Population Structure in San, Khoe, and Mixed Ancestry Populations from Southern Africa Revealed by 44 Short 5-SNP Haplotypes.* Human biology pp. 695 – 724.

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 pg 56.

Schlect J; Kaplan M; Barnard K; Karafet T; Hammer M and Merchant N (2008) *Machine-Learning Approaches for Classifying Haplogroup from Y Chromosome STR Data.* PLoS Computational Biology.

Schoeman K (1997). *The Mission at Griquatown (1801-1821).* National Book Printers

Schoske R; Vallone P; Kline M; Redman J and Butler J. (2004). *High-Throughput Y-STR Typing of US Populations with 27 Regions of the Y Chromosome Using Two Multiplex PCR Assays.* Forensic science international 139 pp. 107 – 121.

Soares I; Amorim A and Goios A (2012). *Mtdnaoffice: A Software to Assign Human mtDNA Macro Haplogroups through Automated Analysis of the Protein Coding Region.* Mitochondrion 12 pp. 666 – 668

Stoneking M. (2000). *Hypervariable Sites in the mtDNA Control Region are Mutational Hotspots.* The American Journal of Human Genetics 67 pp. 1029 – 1032.

Tanaka M; Cabrera V; Gonzalez A; Laruga J; Takeyasu T; Fuku N; Guo L; Hirose R; Fujita Y; Kurata M; Shinoda K; Umetsu K; Yamada Y; Oshida Y; Hirose N; Ohta S; Ogawa O; Tanaka Y; Kawamori R and Shimodaira H (2004). *Mitochondrial Genome Variation in Eastern Asia and the Peopling of Japan.* Genome Research 14 pp. 1832 – 1850.

The International HapMap Consortium (2003). *The International HapMap Project.* Nature 426 pp. 789 – 796.

Tishkoff S; Reed F; Friedlander F; Ehret C; Ranciaro A; Froment A; Hirbo J; Awomovi A; Bodo J; Doumbo O (2009). *The Genetic Structure and History of Africans and African Americans.* Science 324 pp. 1035 – 1044.

Torroni A; Rengo C; Guida V; Cruciani F; Sellitto D; Coppa A; Calderon F; Simionati B; Valle G; Richards M; Macauly V and Scozzari, R. (2001). *Do the Four Clades of the mtDNA Haplogroup L2 Evolve at Different Rates?* The American Journal of Human Genetics 69 pp. 1348 – 1356.

Traditional leadership report 2012 presentation by Mr. Cecil Le Fleur (Chairperson: Council of Heads for the Griqua National Conference of South Africa) Retrieved from the Wolpe trust website: http://www.wolpetrust.org.za/dialogue/

Umetsu K; Tanaka M; Yuasa I; Saitou N; Takeyasu T; Fuku N; Naito E; Kazutoshi A; Nakayayashiki N; Miyoshi A; Kashimura S; Watanabe G and Osawa, M. (2001). *Multiplex Amplified Product-Length Polymorphism Analysis for Rapid Detection of Human Mitochondrial DNA Variations.* Electrophoresis 22 pp. 3533 – 3538.

Underhill P and Kivisild T (2007) *Use of Y chromosome and Mitochondrial DNA Population Structure in Tracing Human Migrations*. Annual Review in Genetics 4 pp. 539 – 564.

Vallone P; Hill C and Butler J (2008). *Demonstration of Rapid Multiplex PCR Amplification Involving 16 Genetic Loci.* Forensic Science International: Genetics pp. 42 – 45.

Van Oven M and Kayser M. (2009). *Updated Comprehensive Phylogenetic Tree of Global Human Mitochondrial DNA Variation.* Human mutation 30 E386 – E394. http://www.phylotree.org. doi:10.1002/humu.20921

Van Oven M; Ralf A and Kayser M (2011). *An Efficient Multiplex Genotyping Approach for Detecting the Major Worldwide Human Y-Chromosome Haplogroups.* International Journal of Legal Medicine 125 pp. 879 – 885.

Vermeulen M; Wollstein A; van der Gaag K; Lao O; Xue Y; Wang Q; Roewer L; Knoblauch H; Tyler-Smith C; de Knijff P and Kayser M (2009). *Improving Global and Regional Resolution of Male Lineage Differentiation by Simple Single-Copy Y-Chromosomal Short Tandem Repeat Polymorphisms.* Forensic Science International: Genetics 3 pp. 205 – 213.

Vossen R; Aten E; Roos A and den Dunnen J. (2009). *High-Resolution Melting Analysis (HRMA)—More Than Just Sequence Variant Screening.* Human mutation 30 pp. 860 – 866.

Wakeley J (1993). *Substitution Rate Variation among Sites in Hypervariable Region 1 Of Human Mitochondrial DNA*. Journal of Molecular Evolution 37 pp. 613 – 623.

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

Wang C; Wang L; Shrestha R; Wen S; Zhang M; Tong X; Jin L and Li H (2013). *Convergence of Y chromosome STR Haplotypes from Different SNP Haplogroups Compromises Accuracy of Haplogroup Prediction.* Population and Evolution pp. 1 – 13.

Wittwer C. (2009). *High-Resolution DNA Melting Analysis: Advancements and Limitations.* Human mutation 30 pp. 857 – 859.

Zuccarelli G; Alechine E; Caputo M; Bobillo C; Corach D and Sala A (2011). *Rapid Screening for Native American Mitochondrial and Y-chromosome haplogroups Detection in Routine DNA Analysis.* Forensic Science International: Genetics 5 pp. 105 – 108.

Appendix

1. DNA extraction; Ethanol/NaCl precipitation; Quantification

1.1 Medrano (1990) DNA extraction protocol

1.1.1. Reagents

Lysis buffer (50ml)

2M NaCl (*Merck Laboratory Supplies*)	10ml
2M Tris-HCl pH8 (*Merck Laboratory Supplies*)	0.5ml
0.5M EDTA (*Merck Laboratory Supplies*)	0.2ml
1% SDS (*Merck Laboratory Supplies*)	0.5g
SABAX Water (*Adcock Ingram*)	39.3ml
20mg/ml Proteinase K	3µl

1.1.2. Procedure

1. The tip of the swab was cut off with sterile scissors and placed in a 1.5ml eppendorf tube.
2. 600µl of lysis buffer and 3µl of Proteinase K were added to the tube.
3. The tube was vortexed for 30 seconds (Vortex mixer model S A 3) and incubated at
   56°C (Vortemp 56) overnight.
4. All the volume was transferred to a clean tube.
5. The tip of the swab still had lysis buffer and biological material trapped inside. This liquid
   was released by placing it into a 0.5ml tube with a perforated bottom.

6. This tube was then placed into a 1.5ml Eppendorf tube and spun down (Centrifuge 5414

   D) at 5000rpm until the swab was dry. The collected volume was added to the previously separated lysis material.

1.2 Ethanol/NaCl precipitation procedure

1.2.1 Reagents

99% Ethanol or isopropanol	800µl
70% Ethanol	100µl
5.5M NaCl (50ml)	
NaCl (*Merck Laboratory Supplies*)	32.142g
Sabax Water (*Adcock Ingram*)	50ml

1.2.2 Procedure

1. Precipitation was done by adding 200µl of 99% Ethanol and shaking the tube vigorously
   for 15 seconds.
2. The tube was centrifuged for 15 minutes at 5000rpm (Centrifuge 5414 D) and the supernatant containing DNA was transferred to another clean tube.
3. 800µl of cold isopropanol (-20°C) was added and the eppendorf tube was left at -150°C
   for 15 minutes.
4. The DNA pellet was collected by centrifugation at 14000rpm for 30 minutes (Centrifuge 5414 D)
5. The pellet was washed with 100µl of 70% Ethanol to remove the salts; the eppendorf was centrifuged at 14000rpm for 15 minutes (Centrifuge 5414 D).

6. The DNA pellet was dried for ten minutes in the Speedy Vac (Memmert) to allow ethanol evaporation.
7. The DNA was dissolved in 30µl of SABAX water and stored at -20°C.

1.3 DNA Quantification and Working Stock Solutions

1. 1µl of DNA was used for quantification, using the Nanodrop ND1000 spectrophotometer.
2. The DNA concentration was recorded from the Nanodrop readings.
3. The working stock dilutions of 2ng/ml were prepared from all the DNA samples.
4. The original samples and working stocks were stored at -20°C.

2. Electronic Supplementary Resources

STR base website	http://www.cstl.nist.gov/biotech/strbase/
South Africa Police Service rape statistics database	http://www.saps.gov.za/statistics/reports/crimestats/2014/downloads/crime_statistics_presentation.pdf
Summary list of Y-chromosome STR loci	http://www.cstl.nist.gov/biotech/strbase/y_strs.htm.
STR fact sheets for individual markers	http://www.cstl.nist.gov/strbase/str_fact.htm.
Duplication events in DYS19, DYS385a/b, DYS439 and DYS389II	www.promega.com/geneticidentity.
Y chromosome haplotype reference database (YHRD)	http://www.yhrd.org
SWGDAM recommended loci	http://www.fbi.gov/hq/lab/fsc/backissu/july2004/standards2004_03_standards02.html
BioEdit sequence alignment software ©1997-2013 Thomas Hall, Ibis Biosciences, Carlsbad, CA 92008	http://www.mbio.ncsu.edu/bioedit/bioedit.html
Chromas Lite v2.1.1	http://www.technelysium.com.au/chromas.html

Network software v4.6.1.0. Fluxus Engineering	http://www.fluxus-engineering.com/sharenet.html
Multiple primer Tm calculator	http://www.thermoscientificbio.com/webtools/multipleprimer/
RNAstructure: software for secondary structure prediction and analysis	http://rna.urmc.rochester.edu/RNAstructureWeb/Predict1.html