A comparative ancestry analysis of Y-chromosome DNA haplogroups using high resolution melting

UNIVERSITY *of the*
WESTERN CAPE

Adria Michelle Burrows

November 2018

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

Supervisor: Prof. Maria Eugenia D'Amato; Co-Supervisor: Dr Peter Ristow

Keywords

Y-Chromosome

Single nucleotide polymorphism

Deoxyribonucleic acid

High resolution melting

Haplogroups

Ancestry

Multiplex

Abstract

The objective of this study is to deduce paternal ancestry using ancestry informative single nucleotide polymorphisms (SNPs) by means of High Resolution Melting (HRM). This was completed by producing a multiplex system that was designed in a hierarchical manner according to the YSNP tree. This project mainly focused on African ancestry and was used to infer paternal ancestral lineages on the Johannesburg Coloured population.

South Africa has a diverse population that has ancestral history from across the globe. The South African Coloured population is the most admixed population as it is derived from at least five different population groups: these being Khoisan, Bantu, Europeans, Indians and Southeast Asians. There have been studies done on the Western Cape/ Cape Town Coloured populations before but this study focused on the Johannesburg Coloured population.

The first step was to design the multiplex system. This was done by using in-house SNPs. A total of seven multiplexes were designed and optimised, each consisting of two, three or four different SNPs respectively.

A total of 143 saliva and buccal samples were collected from male Johannesburg Coloureds. DNA was extracted from the saliva samples using an optimised organic method. DNA was extracted from the buccal samples using an optimised salting out method. DNA was successfully extracted from 77 of the male samples.

A total of 69 samples were screened using Multiplex 1; of the 69 samples 56 samples were successfully screened to infer the paternal lineage of the samples.

The results show that the most frequent haplogroup of the Johannesburg male samples was haplogroup CF (39%). The second most frequent haplogroup was haplogroup DE (38%). Under further analysis of haplogroup DE it was seen that 37% of those samples were derived for the haplogroup E1b1b.

Upon further statistical analysis the relative contribution of African, European and Asian ancestry between the two population groups was tested. It was observed that the Western Cape Coloured population and the Johannesburg Coloured population had a significant statistical difference between them (Chi-square test, p= 0.00025).

When comparing the Western Cape Coloured paternal lineages of haplogroup E to the Johannesburg Coloured paternal lineages of haplogroup E clear differences were seen. The Western Cape Coloured samples most frequent E haplogroup was haplogroup E1b1a1 (24.12%) compared to the Johannesburg Coloured samples most frequent E haplogroup was haplogroup E1b1b (37%).

This indicates that even though the population classifications are the same the locations of the population groups play a huge role in the ancestral history of the population group.

Declaration

I, Adria Michelle Burrows, declare that (*A comparative ancestry analysis of Y-chromosome DNA haplogroups using high resolution melting*) is my own work, that it has not been submitted for any degree or examination in any other university, and that all the sources I have used or quoted have been indicated and acknowledged by complete references.

Adria Michelle Burrows

Date: 10 August 2018

Signed

Acknowledgements

To God, thank you

Thank you God for the grace you have shown me over these past three years.

To my exceptional parents thank you, thank you, and thank you! There are not enough words in this world to describe how grateful I am to have you as parents. Thank you for being my pillars of strength. Thank you for the countless amounts of opportunities that you have given me in my short life on this earth. I am truly blessed.

A special thank you to my mother; without her constant insistence that I sit down and write I do not know if I would have finished this thesis.

To my loving family, thank you for your constant support and motivation. Thank you for the countless hours you have motivated me to be the best version of myself. Thank you for encouraging me to aim high and reach for my dreams. Without your constant support I do not know where I would be today.

Thank you for listening to me explain my project even if you did not understand everything I was doing. Knowing that you are excited that I am enjoying what I am doing is more than enough.

To Sergei, these have possibly been the hardest three years we have been together. Thank you for staying at my side when I wanted to give up because science was just not working that day. Thank you for constantly telling me I can do this and that I would finish. Thank you for being one of my biggest motivations to getting done this year. You truly are one amazing human being. I love you.

To Amber, 11 years later and you are still my number one fan. Thank you for putting up with my non-existence these past three years. Thank you for listening to me explaining what I was doing in the laboratory and you just nodding like you understood. Your friendship truly means a lot to me.

To Shelby, thank you for being a huge motivation for me finishing this degree. Thank you for being my travel buddie for the past two years. You have been missed this year. Thank you for constantly checking up on how far I am with my writing and for all the motivation you keep giving me. You are truly my sister in science.

To the FDL members without whom I do not think I would have enjoyed this degree as much as I did, thank you for always being helpful and encouraging. Thank you for the morning conversations and all the laughter that was shared amongst us all.

To my co-supervisor Dr Peter Ristow, a huge thank you to you sir! Without your help and supervision I think I might have lost my way a long time ago. Thank you for teaching me and helping me in the lab while you were a part of the FDL. You did make lab life enjoyable. Thank you for being an amazing role model for the whole FDL.

Thank you to my supervisor, Prof. D'Amato for being helpful and so approachable. Thank you for being a wonderful travel companion last year. You made my first extended trip away from my family enjoyable and an amazing experience.

This degree is dedicated to my ma, Bini Erasmus, whose kitchen table was used to write this thesis. Gone too soon but never forgotten.

To science, thank you for being a pain but a joy at the same time.

The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at are those of the author and are not necessarily to be attributed to the NRF.

List of abbreviations

°C- degrees celsius

μM- micromolar

bp- base pair

dH_2O- distilled water

DNA- Deoxyribonucleic acid

dsDNA- Double stranded DNA

HRM – High resolution melting

$MgCl_2$- Magnesium Chloride

mM- millimolar

NYR- Non-recombining portion of the Y-chromosome

PAR- Pseudoautosomal regions

PCR- Polymerase chain reaction

SNP- Single nucleotide polymorphism

STR- Short tandem repeat

Tm- Melting temperature

List of Figures

List of Tables

Table of Contents

Chapter 1: Introduction

1.1 Population genetics

In this day and age it has become increasingly popular to know your ancestral genetic roots. This is due to the countless ancestral DNA kits that have become readily available to the general public. DNA kits such as AncestryDNA™, MyHeritage DNA™, Family Tree DNA, 23andMe and Living DNA all provide one with being able to obtain one's genetic ancestral history.

But why has population genetics become such a huge sensation that has taken the world by storm? What do we gain by knowing where we come from or where certain population groups originated from?

Population genetics is the study of genetic variation that occurs within population groups by observing the variation in alleles and genotypes. Autosomal markers, mitochondrial DNA markers or Y-chromosome DNA markers are used to obtain one's ancestral information. They all contain alleles that allow the study of genetic variation in a population group or an individual. This project focuses on Y-chromosome DNA markers and is discussed further on in this chapter.

1.1.1 Southern Africa population groups

The countries that fall under Southern Africa are: Angola, Zambia, Namibia, Botswana, Zimbabwe, Mozambique, South Africa, Lesotho and Swaziland. This can be seen in Figure 1.1.

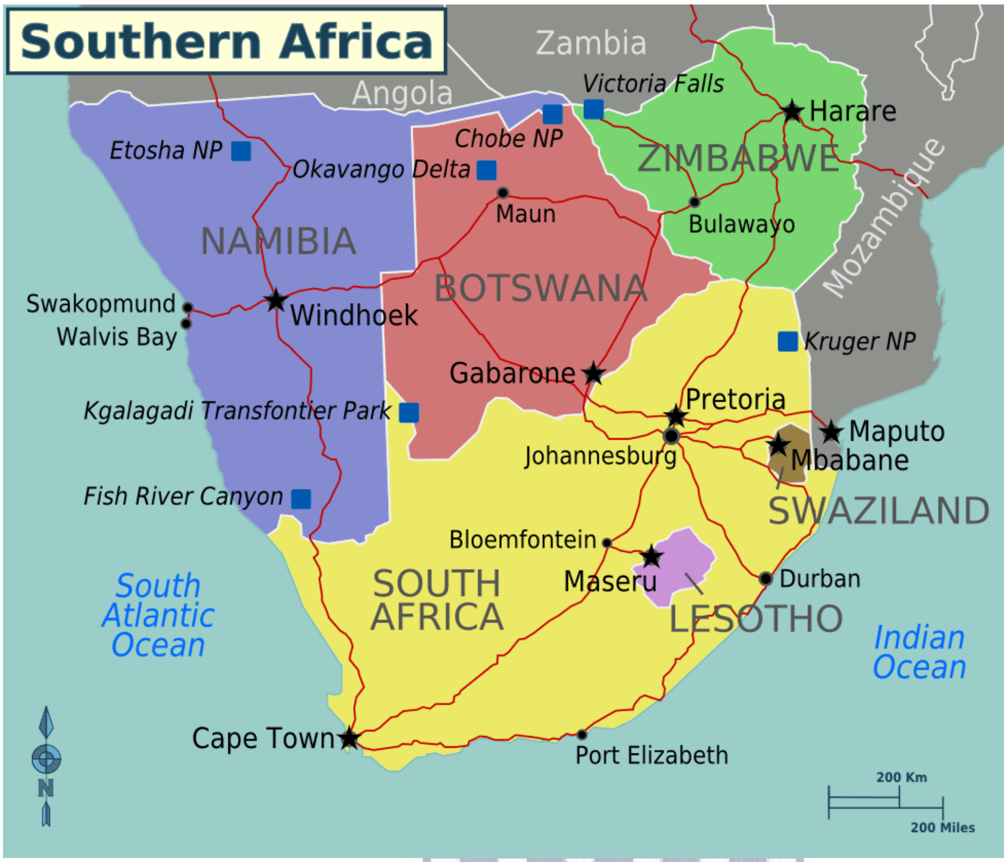


Figure 1.1 An illustration depicting which African countries belong to Southern Africa. Image taken from http://globalcitymap.com/africa/southern-africa-map.html

Southern Africa contains the highest human genomic variation within the population groups that inhabit the area (Petersen *et al.*, 2013). The Khoisan population is the most diverse and they are mainly found in the greater Kalahari regions of Namibia and Botswana (Li *et al.*, 2008; Petersen *et al.*, 2013).

Because of the vast genetic diversity that can be found in Southern Africa it has become a great area to conduct genetic studies.

1.1.1.1 South Africa

South Africa consists of a diverse population group with ancestral history from the rest of the world. After the 2011 census was completed it was deduced that 79.2% of the population defined themselves as African, 8.9% defined themselves as Coloured and 8.9% defined themselves as White. Asian/Indian population groups made up 2.5% and 0.5% were defined as other (Statistics South Africa, 2012).

Although the native African population is the largest in the country the Coloured population was the focus of this research project. The South African Coloured population is admixed as the population is derived from at least 5 different population groups. These being Khoisan, Bantu, Europeans, Indians and Southeast Asia (Quintana-Murci *et al.*, 2010). The heritage of the Coloured population can be dated back to 1652 when the Dutch East Indian Company established a refreshment station on the shores of Table Bay (Quintana-Murci *et al.*, 2010).

Quintana-Murci and team conducted a study evaluating the paternal and maternal genetic ancestry of 590 Coloured samples from the Western Cape. They tested 64 mitochondrial DNA binary markers and sequenced the Hyper Variable Segment I of the control region. Their results showed that 79% of the samples belonged to various Sub-Saharan Africa haplogroups. Of the 79%, 60% of the samples belonged to the haplogroup L0d. This haplogroup is specific to Khoisan people of South Africa (Tishkoff *et al.*, 2007; Quintana-Murci *et al.*, 2010). The other 19% belonged to African maternal lineages that could have been introduced to South Africa during the Bantu expansion period (Salas *et al.*, 2002; Quintana-Murci *et al.*, 2010). The study showed that 16.3% of the samples belonged to various South/ Southeast Asian haplogroups. The haplogroups B4 and B5a were observed within the Coloured population samples and are known to be haplogroups found in South/ Southeast Asia (Kong *et al.*, 2006; Quintana-Murci *et al.*, 2010). The last maternal lineages that were observed in the Coloured population were European haplogroups at 4.6% (Quintana-Murci *et al.*, 2010).

Quintana-Murci and team evaluated the paternal genetic ancestry by looking at 46 binary markers and 14 Y-STRs. Their results showed that 45.2% of the samples belonged to Sub-Saharan haplogroups. Of the 45.2%, 24.9% of the samples were derived for haplogroup E-M2. This haplogroup is restricted to the Sub-Saharan population and is found predominantly among Bantu speaking groups and varies within Khoisan groups (Sims *et al.*, 2007; Wiesner and Slavin, 2009; Quintana-Murci *et al.*, 2010). The study showed that the paternal ancestry of the Coloured population showed ancestry for Khoisan and Bantu speakers but the majority of the results were derived to the wider pan-African ancestral history (Quintana-Murci *et al.*, 2010). The study showed that 37.5% of the samples belonged to European haplogroups and 31.7% of the samples belonged to South/ Southeast Asian haplogroups (Quintana-Murci *et* *al.*, 2010). This study concluded that the Coloured population is a diverse admixed population but that the diversity is also gender biased.

The Johannesburg Coloured population is the focus of this project. Figure 1.2 illustrates where the samples were collected from. It will be interesting to discover what the genetic ancestry of the Johannesburg Coloured population will be; if it will be similar to the Cape Coloured as stated in Quintana-Murci *et al.* (2010) or if the genetic diversity will be vastly different.

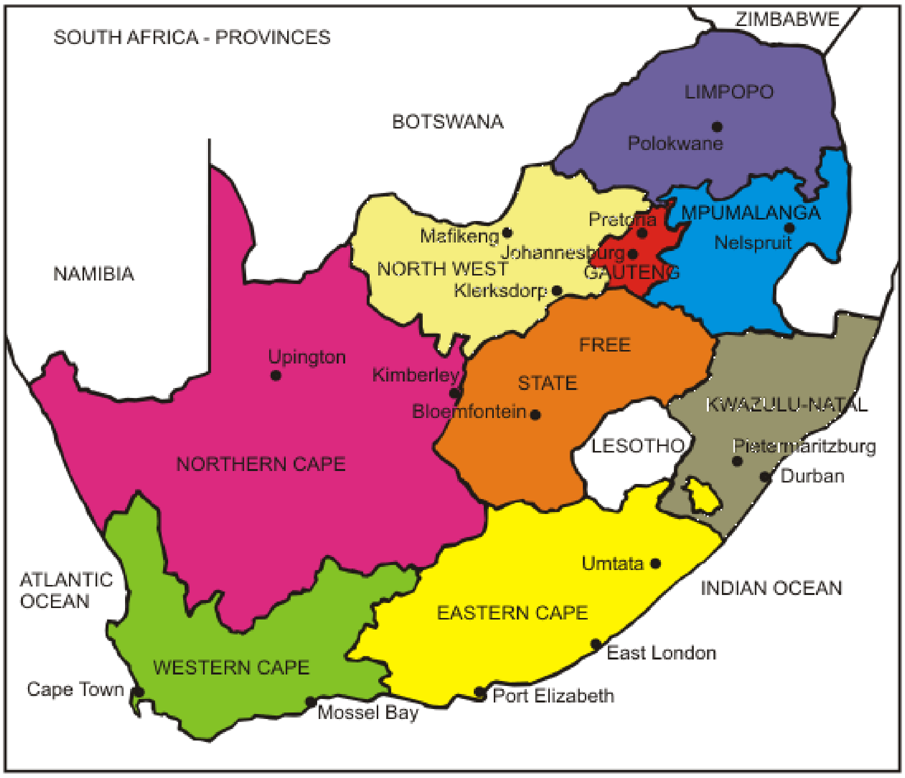


Figure 1.2 Illustrates the map of South Africa. The samples used for this project were collected in Johannesburg. The red coloured province is Gauteng where Johannesburg is located. Map taken from http://www.sataxguide.co.za/south-africa-political-map/

1.2 Y-chromosome

The Y-chromosome forms part of the 23 pairs of human chromosomes. It is one of the two sex chromosomes in humans, the X-chromosome being the other one. The Y-chromosome characterises 2% of the total DNA in cells and spans more than 59 million building blocks of DNA. Y-chromosomes are found in males and contain genes that are involved in male sex determination and development (Butler, 2012b).

1.2.1 Y-chromosome structure

The Y-chromosome is the third smallest human chromosome at 50Mb. The pseudoautosomal regions (PAR), tips of the Y-chromosome, recombine with their sister sex X-chromosome homologous regions (Butler, 2012b). The rest of the Y-chromosome is known as the non-recombining portion of the Y-chromosome (NRY). This part of the Y-chromosome is the part that is passed down from father to son and remains the same throughout generations (Figure 1.3). The NRY is also known as the male-specific region (MSY) (Skaletsky *et al.*, 2003).

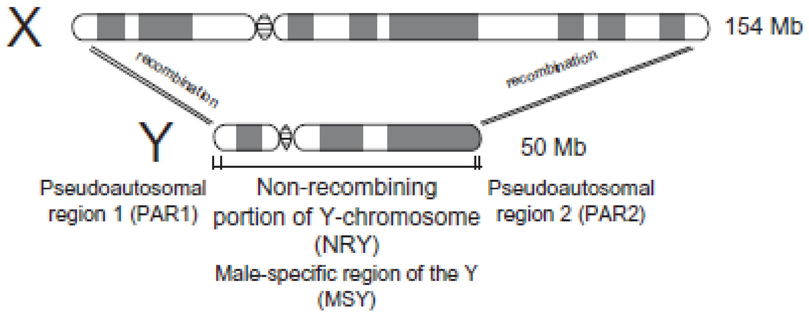

X
154 Mb
recombination
recombination
Y
50 Mb
Pseudoautosomal region 1 (PAR1)
Non-recombining portion of Y-chromosome (NRY)
Male-specific region of the Y (MSY)
Pseudoautosomal region 2 (PAR2)

Figure 1.3 A schematic of the X and Y chromosomes taken from Butler (2012b). PAR1 and PAR2 recombine with the tips of the X chromosome. The NRY or MSY makes up 95% of the Y-chromosome.

1.2.2 Lineage Markers

Y-chromosome and mitochondrial markers represent lineage markers. These markers are passed from generation to generation without having significant changes except mutational events. Paternal lineage is traced with Y-chromosome markers and maternal lineage is traced with mitochondrial DNA makers. Lineage markers are different to autosomal DNA markers as autosomal DNA markers recombined with each generation as half the genetic information is obtained from paternal inheritance and the other half maternal inheritance (Butler, 2012b). This can be seen Figure 1.4.

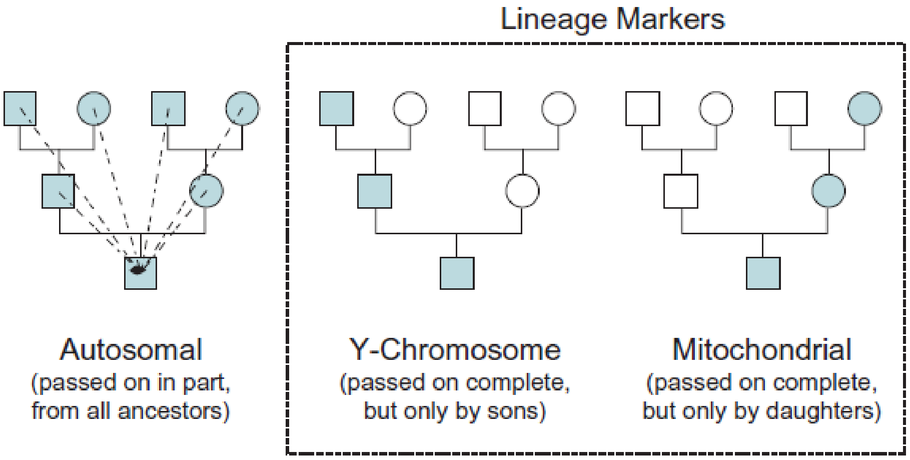

Lineage Markers
Autosomal
(passed on in part, from all ancestors)
Y-Chromosome
(passed on complete, but only by sons)
Mitochondrial
(passed on complete, but only by daughters)

Figure 1.4 An illustration of inheritance patterns for autosomal, Y-chromosome and mitochondrial DNA. Image taken from Butler (2012b)

Haplotypes are a set of markers from a single chromosome that are inherited together. Haplotypes can refer to a combination of alleles or a set of SNPs. Y-chromosome markers are not as informative as autosomal markers are when it comes to differentiating between two individuals. All chromosomes recombine with their homologous pair whereas the Y-chromosome does not recombine as the Y-chromosome markers are linked on the same chromosome. Genotypes from autosomal markers are unlinked and segregate separately from generation to generation. Y-chromosome and mitochondrial markers do, however, play an important role in forensic investigations as well as other human identification applications (Butler, 2012b).

1.2.3 Different Y-chromosomes markers

There have been two different types of DNA markers to investigate the diversity of the Y-chromosome. The first group is bi-allelic loci that display two possible alleles. Results from typing bi-allelic markers are classified into haplogroups (de Knijff, 2000). Single nucleotide polymorphisms (SNPs) fall into the bi-allelic loci category as they have low mutation rates (~ 10^{-8} to 10^{-9} per generation). However, not all bi-allelic markers are ancestry informative.

The second group is multi-allelic loci. The results from typing these markers are classified into haplotypes (de Knijff, 2000). Several hundred short tandem repeat (STR) markers fall into this category (Kayser *et al.*, 2004). They can differentiate between Y-chromosome haplotypes as they have a higher mutation rate than SNPs. Y-STRs are also used in paternity disputes of male offspring and other types of paternal kinship testing (Kayser, 2017).

1.2.4 Single nucleotide polymorphisms (SNPs)

Single nucleotide polymorphisms (SNPs) are defined as being a single-based sequence variation between individuals at a certain position in the genome (Butler, 2012a).

Firstly, a polymerase chain reaction (PCR) can be performed and produces product sizes smaller than 100bp in size. This means that these SNPs can recover information from degraded DNA samplers better than STRs. SNPs can be multiplexed better than STRs as detection methods used for SNPs does not have the same constraints as STR detection methods do. Sample processing and data analysis can be fully automated, depending on the method that is being used, because size separation is not needed. When it comes to analysis of the alleles using SNPs is easier than STRs as there is no stutter artefact associated with each allele. This helps simplify interpretation of the results. Lastly, one is able to predict ethnic origin and certain physical traits.

SNPs are categorised into four different groups. They are individual identification or identity testing SNPs, lineage-informative SNPs, ancestry-informative SNPs and phenotype-informative SNPs. Identity SNPs are chosen so that they give a low probability of two individuals having the same multi-locus genotype. Lineage SNPs are multi-allelic markers that can identify relatives better than bi-allelic SNPs. Ancestry SNPs are chosen so that they give a high probability of an individual's ancestry. Phenotype SNPs give a high probability of an individual having particular phenotypes like hair colour, eye colour and skin colour.

1.2.4.1 Y-SNPs and haplogroups

Y-SNPs play an important role in human migration studies because an effective evaluation of major differences between population groups can be conducted (Wang *et al.*, 2010). Y-SNP alleles are either designated as ancestral or derived according to the nucleotide that is present at the specific SNP site. This allows them to be recorded in simple binary format of 0 or 1 for ancestral and derived respectively (Butler, 2012b).

The first Y-chromosome phylogeny tree was first published by Y Chromosome Consortium Tree in 2002 and consisted of 153 binary haplogroups (Consortium, 2002). Updates of the phylogeny tree have been done by Jobling and Tyler-Smith (2003) and Karafet *et al.* (2008). The most recent update of the Y-chromosome phylogeny tree that PhyloTree is using was done by Van Oven *et al.* (2014). The latest Y-chromosome tree (Figure 1.5) consists of 417 primary defining markers. Phylotree Y is a website that provides a minimal reference phylogeny for the Y-chromosome as described in Van Oven *et al.* (2014).

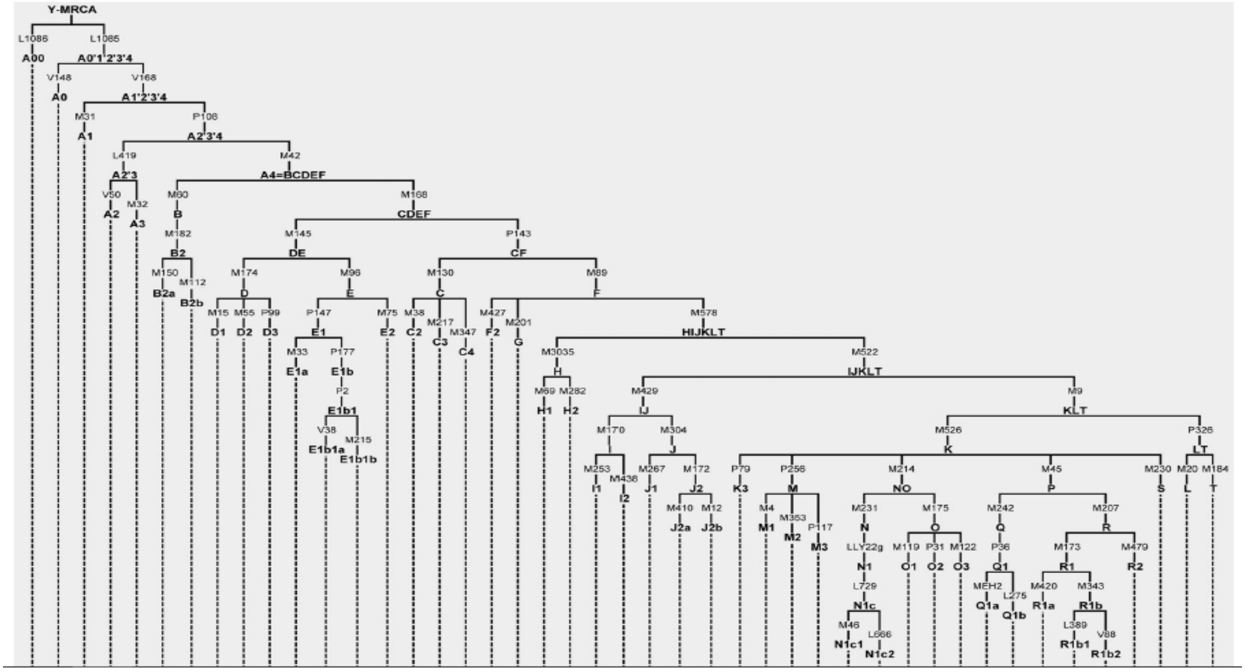


Figure 1.5 Skeleton of the Y-chromosome phylogeny. Taken from Van Oven *et al.* (2014)

1.2.5 SNPs chosen for this project

The haplogroups that were focused on for this study are mainly found in Africa and Southern Africa.

SNPs were chosen from publications (Selelstad *et al.*, 1994; Underhill *et al.*, 2001; Cruciani *et al.*, 2002; Sims *et al.*, 2007; Karafet *et al.*, 2008; Mendez *et al.*, 2013) then compared to the Y-SNP tree (http://www.phylotree.org/Y/tree/index.htm).

1.2.5.1 Haplogroup A

This haplogroup is mainly found in Africa. According to Van Oven *et al.* (2014) haplogroup A is defined by the following SNPs: L1085, V148 (rs181335666), M31 (rs369315948), V50 (rs189205028) and M32. These SNPs can be seen in Figure 1.5. The SNPs seen on Phylotree (http://www.phylotree.org/Y/tree/index.htm) (accessed 1/09/2017) and the SNPs seen in Figure 1.5 correspond to one another. This shows that the SNP names that define haplogroup A have not changed.

It can be seen in Figure 1.5 that SNP L1085 separates haplogroup A00 from the rest of the Y-chromosome tree. If samples are ancestral for SNP L1085 one can assume that they are derived for haplogroup A00. Haplogroup A00 carries the ancestral state for all SNPs that define the Y-chromosome tree (Mendez *et al.*, 2013). The ancestral state for all SNPs was found in an African American Y chromosome sample. According to Mendez *et al.* (2013) haplogroup A00 was found in the Mbo individuals from Western Cameroon (Figure 1.6).

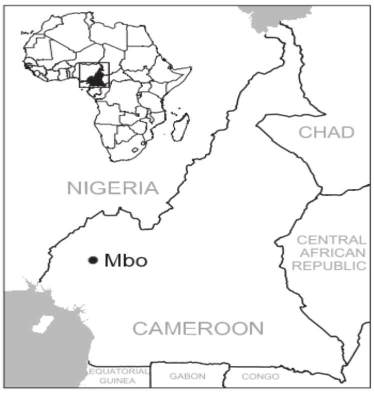


Figure 1.6 Indicating where derived samples for haplogroup A00 are found.

SNP V148 defines haplogroup A0 and is found mainly in central and West Africa. SNP M31 defines haplogroup A1 and is found in Western and North Africa (Batini *et al.*, 2011). SNP V50 defines haplogroup A2 and can be found mainly in Southern and Central Africa (Batini *et al.*, 2011). The last sub-clade in haplogroup A is haplogroup A3 and is defined by SNP M32. Haplogroup A3 is mainly found in Southern and Eastern Africa (Batini *et al.*, 2011).

1.2.5.2 Haplogroup B

Haplogroup B is mainly found in Africa, mostly in Central, Eastern and Southern Africa. Haplogroup B has a high frequency among Pygmies (Underhill *et al.,* 2001; Cruciani *et al.*, 2002; Karafet *et al.*, 2008). According to Van Oven *et al.* (2014) haplogroup B is defined by the following SNPs: M60 (rs2032623), M182 (rs2032601), M150 (rs371646183) and M112 (rs111725135).

Comparing the SNPs seen on Phylotree (http://www.phylotree.org/Y/tree/index.htm) (accessed 1/09/2017) and the SNPs seen in Figure 1.5 one can see that SNPs have been added to further define the B haplogroup. The following SNPs have been added: SNP M236 has been added to define haplogroup B1 and SNP L1388 has been added to define haplogroup B3.

SNP M60 defines the basal branch of haplogroup B. This haplogroup consists of 3 sub-haplogroups but it can be seen in Figure 1.5 that only haplogroup B2 is focused on and the two sub-haplogroups within haplogroup B2.

SNP M182 defines haplogroup B2, with SNP M150 defining haplogroup B2a and SNP M112 defining haplogroup B2b. Haplogroup B2a is mainly found in Eastern, Central and Southern Africa. Haplogroup B2b is mainly distributed in Central and Southern Africa.

1.2.5.3 Haplogroup E

According to Van Oven *et al.* (2014) haplogroup E basal branch is defined by the following SNPs: M96 (rs9306841), M40 (rs9786608) and P29 (rs60115999). The sub-haplogroups are defined by the following SNPs: M75 (rs2032639), P147 (rs16980577), M33 (rs368762706), P177 (rs16980473), P2 (rs9785756), V38 (rs768983) and M215 (rs2032654). Haplogroup E is mainly found in Africa.

The SNPs seen on Phylotree (http://www.phylotree.org/Y/tree/index.htm) (accessed 1/09/2017) and the SNPs seen in Figure 1.5 correspond to one another. This shows that the SNP names that define haplogroup E have not changed.

Haplogroup E is by far the most diverse of all major Y-chromosome clades. Haplogroup E1b1 which is defined by SNP P2 has the greatest geographical distribution as it has two major sub-lineages. Haplogroup E1b1a which is defined by SNP V38 can be seen from West to Central to Eastern and Southern Africa. Haplogroup E1b1b which is defined by SNP M215 can be seen in Northeast Africa and then expanded West. SNP M75 defines haplogroup E2 and can be found mainly from central Africa (Cruciani *et al.*, 2002, 2004, 2006, 2007; Karafet *et al.*, 2008).

1.2.5.4 Haplogroup R

According to Van Oven *et al.* (2014) haplogroup R is defined by the following SNPs: M207 (rs2032658), M479 (rs372157627), M173 (rs2032624), M420 (rs17250535), M343 (rs9786184) and V88 (rs180946844). Haplogroup R is found in Europe, West Asia, Central Asia, South Asia, North and Central Africa.

The SNPs seen on Phylotree (http://www.phylotree.org/Y/tree/index.htm) (accessed 1/09/2017) and the SNPs seen in Figure 1.5 correspond to one another. This shows that the SNP names that define haplogroup R have not changed.

Haplogroup R is split into two sub-clades: haplogroup R1 which is defined by SNP M173 and haplogroup R2 which is defined by SNP M479. Haplogroup R2 is mainly found in south Asia and parts of central and west Asia.

Haplogroup R1 is split into two sub-haplogroups:

Haplogroup R1a is defined by SNP M420 and is found mainly in eastern and northern parts of Europe, Southern Russia and Central Asia.

Haplogroup R1b, which is defined by SNP M343, is found mainly in Western Europe. It is also found in South Asia and among some Sub-Saharan African populations. It is rare to find haplogroup R1b in Africa but it was found in a central-western African population. SNP V88 is used to define R1b found in Africa. It is thought to be the back migration that occurred in prehistoric times (Cruciani *et al.*, 2010).

1.3 Genotyping methods

There are many methods of SNP typing. An important aspect of a SNP assay is the ability to examine multiple markers simultaneously. Table 1.2 is a summary of all the different SNP analysis techniques that are implemented in the field.

Table 1.1 A summary of all the different SNP analysis techniques. Taken from Butler (2012a)

Method	**Description**	**References**
Reverse dot blot or linear arrays	A series of allele-specific probes are attached to a nylon test strip at separate sites; biotinylated PCR products hybridize to their complementary probes and are then detected with a colorimetric reaction and evaluated visually.	(Saiki *et al.*, 1989; Reynolds *et al.*, 2000; Gabriel *et al.*, 2001)
Genetic bit analysis	Primer extension with ddNTPs is detected with a colorimetric assay in a 96-well format.	Nikiforov *et al.* (1994)
Direct sequencing	PCR products are sequenced and compared to reveal SNP sites.	Kwok *et al.*, (1994)
Denaturing HPLC	Two PCR products are mixed and injected on an ion-paired reversed-phase HPLC; single base differences in the two amplicons will be revealed by extra heteroduplex peaks.	(Hecker *et al.*, 1999)
TaqMan 5’ nuclease assay	A fluorescent probe consisting of reporter and quencher dyes is added to a PCR reaction; amplification of a probe-specific product causes cleavage of the probe and generates an increase in fluorescence.	(Livak, 1999)
Fluorescence polarization	Primer extension across the SNP site with dye-labelled ddNTPs; monitoring changes in	(Chen *et al.*, 1999)

	fluorescence polarization reveals which dye is bound to the primer.	
Mass spectrometry	Primer extension across the SNP site with ddNTPs; mass difference between the primer and extension product is measured to reveal nucleotide(s) present.	(Haff and Smirnov, 1997; Li *et al.*, 1999)
High-density arrays (Affymetrix chip)	Thousands of oligonucleotide probes are represented at specific locations on a microchip array; fluorescently labelled PCR products hybridize to complementary probes to reveal SNPs.	Wang et al. (1998), (Sapolsky *et al.*, 1999)
Electronic dot blot (Nanogen chip)	Potential SNP alleles are placed at discrete locations on a microchip array; an electric field at each point in the array is used to control hybridization stringency.	(Sosnowski *et al.*, 1997), Gilles *et al*., (1999)
Molecular beacons	Hairpin stem on oligonucleotide probe keeps fluorophore and its quencher in contact until hybridization to DNA target, which results in fluorescence.	(Giesendorf *et al.*, 1998)
Oligonucleotide ligation assay (OLA)	Colorimetric assay in microtiter 96-well format involving ligation of two probes if the complementary base is present.	(Delahunty *et al.*, 1996)
T_m-shift genotyping	Allelic-specific PCR is performed with a GC-tail attached to one of the forward allele-specific primers; amplified allele with GC-tailed primer will exhibit a melting curve at a higher temperature.	(Germer *et al.*, 1999)
Pyrosequencing	Sequencing by synthesis of 20–30 nucleotides beyond primer site; dNTPs are added in a specific	(Ahmadian *et al.*, 2000), Andreasson

	order and those incorporated result in release of pyrophosphate and light through an enzyme cascade.	*et al*., (2002)
Allele-specific hybridization (Luminex 100)	Dye-labelled PCR products hybridize to oligonucleotide probes (representing the various SNP types) attached to as many as 100 different coloured beads; each bead is interrogated to determine its colour and whether or not a PCR product is attached as the beads pass two lasers in a flow cytometer.	(Armstrong *et al.*, 2000), Budowle *et al.*, (2004)
Minisequencing (SNaPshot assay)	Allele-specific primer extension across the SNP site with fluorescently labelled ddNTPs; mobility modifying tails can be added to the 5'-end of each primer in order to spatially separate them during electrophoresis.	(Tully *et al.*, 1996)
SNPstream UHT	High-tech version of genetic bit analysis with a 384-well tag array and 12plex PCR.	(Bell *et al.*, 2002)
Sanger Sequencing	Sequencing based on selective incorporation of chain-terminating dideoxynucleotides by DNA polymerase during DNA replication	Sanger *et al*., (1975)
High Resolution Melting (HRM)	Detects small differences in PCR melting curves. It uses dsDNA binding dyes used in conjunction with real-time PCR.	(Wittwer *et al.*, 2003)

1.3.1 High resolution melting (HRM)

High resolution melting analysis is a fast, simple and inexpensive method compared to alternative approaches requiring separations or labelled probes (Wittwer, 2009). The advantages of HRM are that it is a low consumption with little waste protocol. There is a simple workflow to follow.

Real-time PCR and melting curve analysis was first introduced in 1997 (Ririe *et al.*,1997; Wittwer *et al.*, 2003). HRM is an entirely closed tube procedure that requires a DNA intercalating dye (Reja *et al.*, 2010; Li *et al.*, 2014). Double stranded DNA (dsDNA) dye SYBR® Green 1 was used for amplicon melting analysis which helped to provide a rough characterisation of what was amplified (Wittwer *et al.*, 2003).

The increasing temperature allows for the intercalating dyes to dissociate from the double stranded DNA, allowing the dye to be released from the DNA and fluorescence decreases. Fluorescence measurements are collected throughout a range of temperatures. The melting temperature (Tm) corresponds to the amplicon. Melting temperature is the temperature at which the double DNA strand dissociates from one another to become single DNA strands (Lando *et al.*, 2015). The melting temperature varies depending on the dNTPs composition or GC content of the sample. Fluorescence intensity values are first normalised between 0-100 by defining the linear baseline before and after melting transition of each sample (Wittwer *et al.*, 2003). The normalised data are then transformed into melting curves by superimposing the curves over a certain fluorescence interval. These form a melting curve, the curve shape allows for samples to be compared to each other. HRM helps to distinguish genotypes by comparing the relative positions and shapes of the melting curves (Ririe, Kirk M., Rasmussen, Randy P., Witter, 1997; Wittwer *et al.*, 2003). The melting curves of each sample are either in the derived or ancestral state for that particular SNP. Each SNP is defined at a mutation point; the mutation location will have a specific nucleotide if it is in the derived state and another nucleotide if it is the ancestral state. Figure 1.7 gives a clear representation of a HRM melting curve that shows both derived and ancestral states for each haplogroup presented.

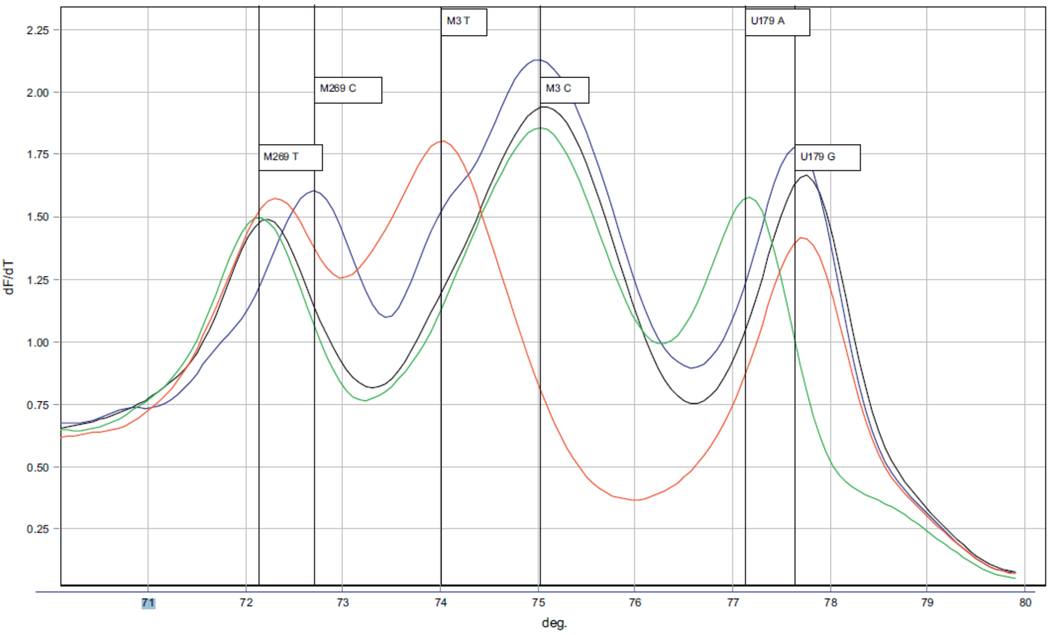


Figure 1.7 Illustrates the first derivative melting curves of a Y-SNP triplex. Haplogroup R1b1b2 (blue), haplogroup Q1a3a (red) and haplogroup I (green) are represented in the above image. The black sample is ancestral for all three haplogroups. Image taken from Zuccarelli *et al.* (2011)

1.4 Aims and Objectives of project

Population genetics is the study of factors affecting the allele and genotype frequencies of different loci in a population group. Some DNA markers contain information about the genetic ancestry of a person. These markers, called AIM (ancestry-informative markers) can be autosomal or uniparental (located in mtDNA or Y-chromosome).

The inference of ancestry is done by genotyping with ancestry-informative markers. Among AIMs, SNPs are possibly the most utilized. Single nucleotide polymorphisms (SNPs) typing has become increasingly popular in the forensic field. SNPs have a low mutation rate and can be analysed using short amplicons which help with analysing DNA from degraded samples. SNPs are found mainly in the non-coding regions of genomes. SNPs have been used to define Y-chromosome and mitochondrial DNA haplogroups. Y-chromosomes contain a recombining and non-recombining region. The non-recombining region makes up the majority of the Y-chromosome. It is within the non-recombining region that the SNPs are found that are used to determine paternal ancestral lines.

The paternal lineages will be determined by using High Resolution Melting (HRM) analysis. HRM is a highly powerful method for SNP genotyping. HRM technology characterizes nucleic acid samples based on their disassociation and detects small sequence differences in the PCR amplified sequences by direct melting. This helps one to determine which haplogroup the sample belongs to.

The aim of this project is to produce a cost effective multiplex system that is able to deduce paternal ancestral origin using SNPs and utilising HRM analysis mainly focusing on African ancestry. Y-Chromosome primers have been designed in-house and will be used to create the needed multiplexes. The multiplexes will be used to screen the Johannesburg Coloured population.

Chapter 2: Materials and Methods

2.1 DNA extraction methods

There are multiple methods of extracting DNA, whether by using a DNA isolation kit or manually performing the DNA extraction.

2.1.1 Organic Extraction

Phenol-chloroform is a commonly used method for purifying DNA from a sample. Phenol-Chloroform-Isoamyl (25:24:1) is added to the sample in the lysis buffer. Phenol is denser than the aqueous solution containing the DNA; it sits on the bottom of the tube. It is also non-polar so it is immiscible in aqueous solution. The proteins in the aqueous phase are denatured and move into the phenol phase while the DNA stays in aqueous phase (Oswald, 2008; KIRBY, 1956).

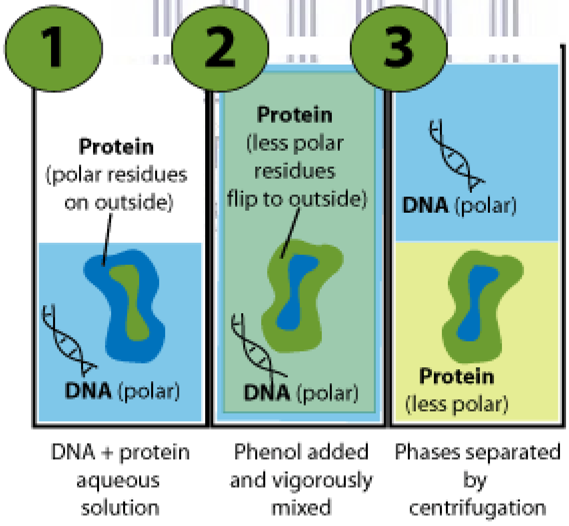


Figure 2.1 Image showing how the DNA is separated from the protein. Taken from Oswald (2008)

The aqueous solution is polar solvent because of the oxygen atoms that are electronegative, meaning the charge is polarised within the molecule. Phenol is less polar than the aqueous solution. Although phenol contains electronegative oxygen its phenyl ring counteracts this charge as it is also electronegative, meaning the charge is not as polarised in the phenol molecule. As DNA is negatively charged due to its phosphate backbone it is soluble in the aqueous phase and less in the phenol, so when mixed with phenol it remains in the aqueous phase. (Oswald, 2008)

Proteins have both polar and non-polar charged amino acids. They fold so that the non-polar charges are inside and the polar charges are facing outwards. When phenol is added the way the protein folds changes. The non-polar residues want to react with the less polar phenol so that the protein flips inside out, denaturing the proteins. Therefore the proteins become more soluble in the phenol separating the proteins from the DNA (Oswald, 2008). In Figure 2.1 it shows how the separation of DNA and proteins occur in phenol-chloroform extraction. Though the proteins are denatured by phenol RNase activity is not completely inhibited. This is why a small amount of isoamyl alcohol is added. Isoamyl alcohol ensures that RNase activity is further deactivated (Sigma Aldrich, 1989).

2.1.2 Salting-Out Extraction

Salting out is a non- toxic extraction method (Javadi and Shamaei, 2014). Salting out is an extraction based on electrolyte - nonelectrolyte interactions. There are hydrophilic and hydrophobic amino acids in proteins. The hydrophobic amino acids are protected from the hydrophilic amino acids and if there are enough hydrophilic amino acids on the protein's surface the protein is soluble in water. When a higher concentration of salt is added to the DNA + protein aqueous solution some of the water molecules are attracted by the salt ions. This decreases the number of water molecules that can interact with the charged part of the protein. This leads to an increased demand for solvent molecules where the protein molecules form hydrophobic interactions with each other (Miller *et al.*, 1988).

Looking at Figure 2.2 one can see the process of a salting out extraction. Cell lysis first occurs to remove the proteins from the DNA, after cell lysis samples are centrifuged to separate the protein precipitation from the DNA sample. The DNA sample is then removed and isopropanol is added to start the DNA precipitation process. After DNA precipitation the DNA samples are centrifuged and washed then dried and finally resuspended.

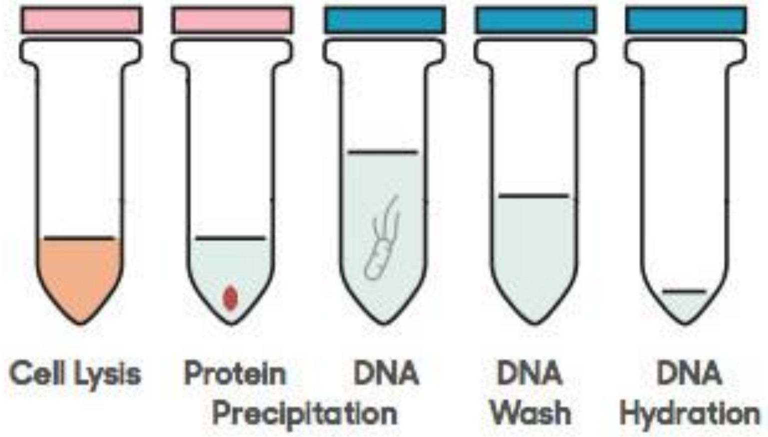


Figure 2.2 Representing the workflow of salting out extraction. Taken from Protocol (no date)

Comparing the two extractions they both have their own advantages and disadvantages. Phenol-chloroform has a high DNA yield with a high purity rate while salting out has a variable DNA yield and purity rate. However, phenol-chloroform is very time consuming and uses toxic compounds whereas salting out is a non-toxic method (Miller *et al.*, 1988).

2.2 Analytic tools

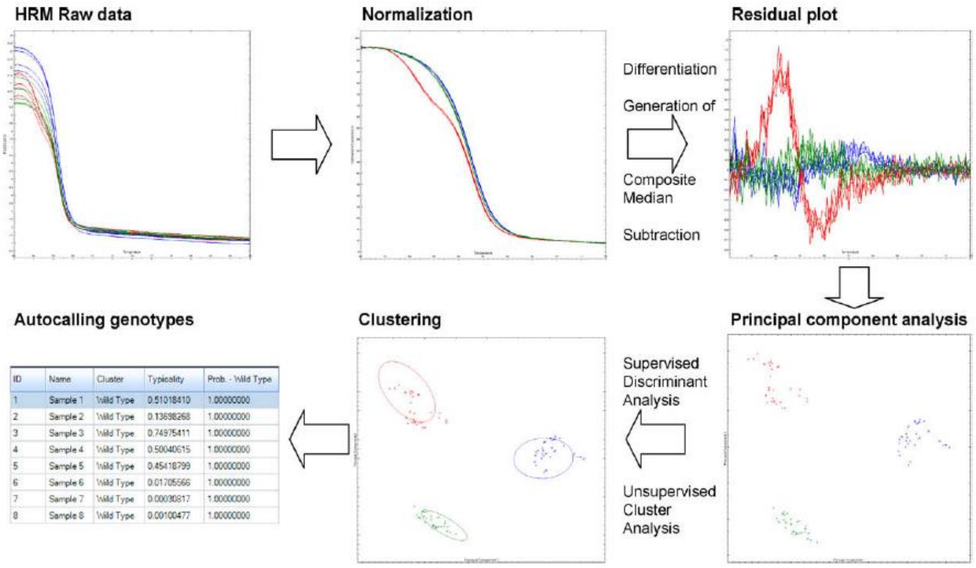
To determine that samples have been correctly placed into haplogroups an analytic tool is used in conjunction with visual analysis.

Amplicon DNA melting analysis for HRM has normalised both the temperature and fluorescence data (Herrmann *et al.*, 2006). This type of analysis removes the majority of the HRM curve information which leaves one with the general shape of the curve intact (Reja *et al.*, 2010).

Rotor-Gene® ScreenClust HRM® software was specifically designed for Rotor-Gene HRM and is the tool that will be implemented in this study.

2.2.1 Rotor-Gene® ScreenClust HRM® software

Rotor-Gene® ScreenClust HRM® is a tool for analysing high-resolution melting data from the Rotor-Gene 6000 cyclers. Samples are grouped into clusters based on their melting curve characteristics. This software performs applications such as genotyping and mutation scanning (QIAGEN, 2009). Figure 2.3 represents a summarised workflow of ScreenClust software (Herrmann *et al.*, 2006).



ID	Name	Cluster	Typicality	Prob. - Wild Type
1	Sample 1	Wild Type	0.51018410	1.00000000
2	Sample 2	Wild Type	0.13698268	1.00000000
3	Sample 3	Wild Type	0.74975411	1.00000000
4	Sample 4	Wild Type	0.50040615	1.00000000
5	Sample 5	Wild Type	0.45418799	1.00000000
6	Sample 6	Wild Type	0.01705566	1.00000000
7	Sample 7	Wild Type	0.00030817	1.00000000
8	Sample 8	Wild Type	0.00100477	1.00000000

Figure 2.3 Illustration representing the workflow of ScreenClust software. Taken from Herrmann *et al.* (2006)

2.2.1.1 Normalisation

The raw HRM data is uploaded onto the Rotor-Gene® ScreenClust HRM® software and normalisation of the HRM curves takes place. Normalisation occurs when a curve scale is applied to the line of best fit of the data set being analysed. This is done so that the highest fluorescence value is equivalent to 100 and the lowest fluorescence is equivalent to 0. A region is selected before and after the melt curve transition. The average fluorescence and best fit line are calculated and applied in the normalisation (QIAGEN, 2009; Reja *et al.*, 2010). Figure 2.4 indicates how normalisation is done using ScreenClust.

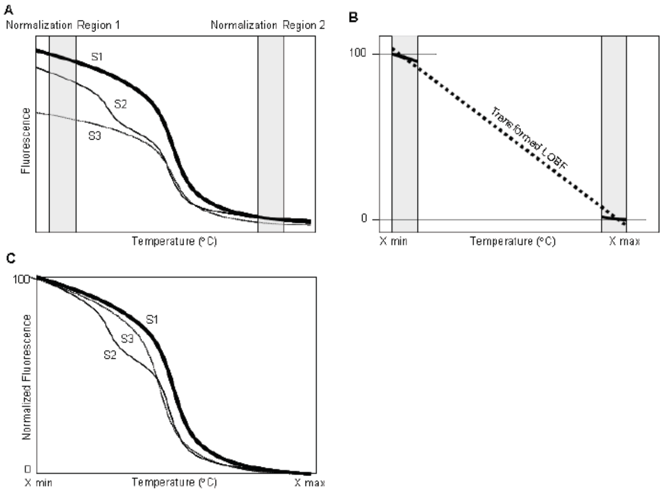


Figure 2.4 (A) in the first step in HRM curve normalization, 2 normalization regions are selected. The software focuses on points within the normalization regions. (B) A line of best fit (LOBF) is determined for each HRM curve between points X min and X max. Each individual HRM curve is forced to have the same LOBF from a fluorescence range of 0 to 100. (C) The software transforms each curve to have this LOBF using the following formula: y = (m2/m1) x F(x) + (m1 x b2 – m2 x b1) / m1; where m1 is the gradient of the individual HRM curve, m2 is the gradient of the required transformed LOBF, b1 is the intercept for the individual HRM curve, b2 is the intercept for the required transformed LOBF, and F(x) is fluorescence at temperature X. Taken from QIAGEN (2009) .

2.2.1.2 Generation of residual plot

After the melt curves are normalised from each sample they are differentiated to emphasise the rate of change in melting temperature. A composite median curve is created by taking the median fluorescence of all curves for each coordinate along the temperature axis. Residual plots are then drawn from the composite median curve for each sample being analysed. The principal components are determined from the residual plots (QIAGEN, 2009; Reja *et al.*, 2010). This can be seen in Figure 2.5.

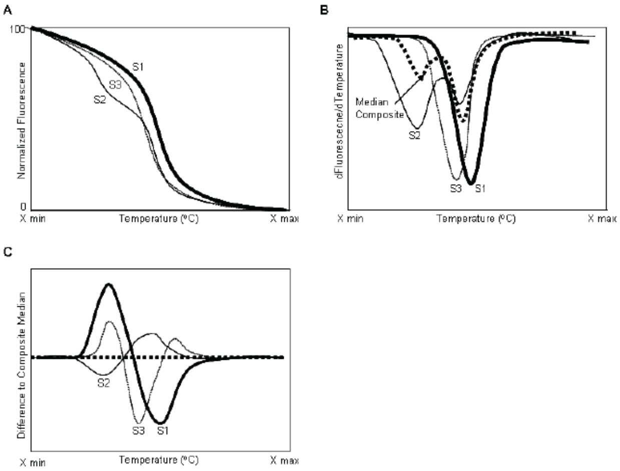


Figure 2.5 Illustration of residual plots. Normalised samples are seen in (A) the samples are then differentiated to emphasise the rate of change in the HRM curves (B). Composite median curve is then subtracted from each individual differentiated sample to give a residual plot (C). Taken from QIAGEN (2009)

2.2.1.3 Principal component analysis

Although principal component analysis (PCA) is a well-known form of data analysis Rotor-Gene$^{®}$ ScreenClust HRM$^{®}$ software is the first software application that applies it to HRM data. PCA determines the cluster patterns from multi-dimensional data. This steps allows similarities and differences to be highlighted (QIAGEN, 2009; Reja *et al.*, 2010).

Three principal components (PC) are used in the process of HRM data analysis. All three principal components allow for variation among the samples being analysed. PC1 variation among the samples is done by using a linear combination of data vectors. If there are more variations that have not been characterised in PC1 this is done with the second principal component (PC2). The third component characterises the remaining variations that have not been accounted in PC1 and PC2. This type of analysis allows for the maximum amount of separation of the samples (QIAGEN, 2009; Reja *et al.*, 2010).

Principal component analysis reduces the dimensionality of a particular data set that consists of a large number of interrelated variables. The PCs are uncorrelated components that are ordered so that the first few PCs retain most of the variation from the original variables of the data set being analysed (Jolliffe, 2002).

The samples are scored against each principal component. Depending on how many principal components are chosen the analysis will depict if it will be plotted in 2 or 3 dimensions (QIAGEN, 2009; Reja *et al.*, 2010).

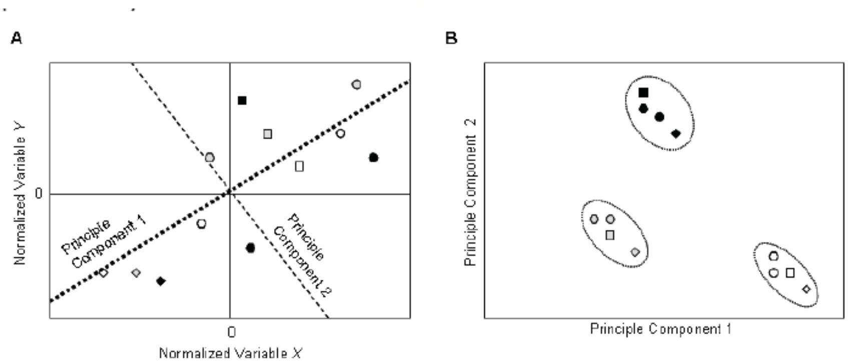


Figure 2.6 Example of principal component analysis. PC1 and PC2 are shown in (A). Individual samples are scored against each PC and drawn (B). Taken from QIAGEN (2009)

2.2.1.4 Clustering

After principal analysis has been done samples are clustered into groups. This can be done in either supervised or unsupervised mode.

2.2.1.4.1 Supervised mode

Unknown samples are classified into known clusters using known control samples. This is achieved by using linear discrimination analysis (LDA). The cluster distribution is calculated by using the known control samples. The cluster centres are set as the mean of the controls and the unknown samples are allocated to a cluster based on their closeness to the mean points of the controls (QIAGEN, 2009; Reja *et al.*, 2010).

Once all unknown samples have been classified the covariance of each cluster is determined and helps calculate posterior probabilities. A posterior probability is the probability of each sample belonging to a particular cluster appropriately. Samples are placed into clusters if they have the highest probability for the cluster.

2.2.1.4.2 Unsupervised mode

Rotor-Gene® ScreenClust HRM® software determines the ideal number of clusters automatically in this mode. This is calculated by using k-means for cluster analysis. The cluster centres are assigned randomly via k-means. Cluster distribution is first calculated from the many random assignments with the least sum of squares. The optimal clusters are then calculated from the sum of the distance between all the samples. The samples are then assigned to the clusters that have the lowest cluster centre for each sample respectively (QIAGEN, 2009).

The optimal amount of clusters is determined by combining gap statistics and k-means (Hartigan and Wong, 1979). This is done by allowing gap statistics to run a within-sum of squares analysis as a cluster quality measure for each cluster number determined by k-means algorithm and plots the gap score for each cluster number as well as the standard deviation (QIAGEN, 2009).

Once the analysis is done the gap score and its error are plotted against the cluster count. The optimal cluster number is determined via the first cluster count where the gap score is greater or equal to the gap score of the following cluster count minus its error. The highest gap score determines how many principal components are used (QIAGEN, 2009).

2.3 Multiplex design according to the Y- PhyloTree

2.3.1. Primer design

The primers that were used for the project were designed in-house by supervisors Dr PG Ristow and Prof. ME D'Amato. In order for primers to be designed relevant literature was read. The Y-SNP tree (http://www.phylotree.org/Y/tree/) and ISOGG 2015 Y-DNA Haplogroup Tree (https://isogg.org/tree/ISOGG_YDNA_SNP_Index.html) were used to select the important SNPs to design primers from. Once the correct SNPs were chosen to be used, primers were designed using Oligo v7 (Wojciech, 2010). BLAT search genome (https://genome.ucsc.edu/cgi-bin/hgBlat) was used to determine the specificity of the designed primers. Oligo v7 (Wojciech, 2010) was used to estimate the melting temperatures of each primer set. This was done to make sure that each primer set used in a multiplex would not overlap the other primer set.

2.3.2. SNPs chosen for project

A total of 21 SNPs were chosen and forward and reverse primers were designed. The SNPs focused on for this study were mainly haplogroups found in Africa and Southern Africa. These SNPs were grouped together according to the hierarchy of the Y-SNP tree. This helped with the design of the primers as the melting temperatures all had to be significantly different. The theoretical ancestral and derived melting temperatures were determined using Oligo v7 (Wojciech, 2010).

Each SNP has a specific nucleotide present when the sample would be ancestral for that specific haplogroup and a specific nucleotide present when the sample would be derived for that specific haplogroup. We evaluated the Tm of the alternative amplicons carrying either the ancestral or derived form of the SNP. Determining the ancestral and derived Tm was completed using Oligo v7 (Wojciech, 2010). Primer sets chosen were designed specifically that the derived and ancestral states differed by > 0.3ºC. Table 2.1 indicates the SNPs that were chosen and the theoretical ancestral and derived melting temperatures for each amplicon. SNPs in Table 2.1 are grouped according to the multiplexes they were placed in.

According to KapaBiosystems (2016) for optimal SNP detection PCR products should range between 40-200 bp in length. This is why all product sizes for each SNP are between this product size range (Table 2.1).

Table 2.1 Indicating the SNPs chosen and their theoretical ancestral and derived melting temperatures

Haplogroup (SNP marker: YCC nomenclature)	rs SNP ID	Forward and reverse primer (‘5-3’)	Mutation	Product size (bp)	Ancestral melting temperature (°C)	Derived melting temperature (°C)	
DE (PF1439)	rs16980598	GGTTTTCAGTTCTGTTCTTGG ACAGTCATCTCTAATGGATCT	G→A	72	74,79	74,23	Multiplex 1
CF (P143)	rs4141886	AAACTGTGAAAATGTGTGGG GCAAATTTTCGTGTAGTCCATG	G→A	118	77,75	76,8	
K(xLT) (M526)	rs2033003	TTAGAGGCAGGGTGTTGCT CTCTGAGCCCAAAAGTCTGG	A→C	73	81,1	81,7	
AOT (AF3)		CGGTAGCCTGTCTGATC CTTAGGAACCAGGAGATACC	CCT→del	96		83,4	
A1a (M31)	rs369315948	GAACCAGACAATACGAAATAGAAG ATGATAATTCACAGATGTCAGGAC	G→C	141	80.8	80.4	Multiplex 2
A0 (L1055)	rs766814820	ATAGGCAAAATTGGATACATGTG TGTACATGGCAGTTACATTCCTC	G→A	85	69,68	69,03	
B (M181)	rs2032599	GATTTTTCTCCTGGACAACTTG AACTGGCAATATTTACTATTTGG	T→C	94	75.1	75.5	
B2 (M182)	rs2032601	TATTCAAAGACTTAAAGCAGTGG ACTTGCATTTGTCCAACTTG	C→T	104	64,74	62,22	Multiplex 3
B2b (M112)	rs111725135	GTTGCAGAATTATCTACCTCTTT AAGAGGTGAGATAAAAACAAAGC	C→T	102	75,7	75,3	
B2a (M152)	rs371646183	CTGCCCACACACACAGATAG AGGAGAAGGAGGGTATCCTG	C→T	110	82,91	82,49	
E2 (M75)	rs2032639	AGACAATTATCAAACCACATCC AAGACATTTATTGAACAGAGGC	G→A	90	73,6	73,1	Multiplex 4
E1 (P147)	rs16980577	ACAAGGACTGGGCAAGTAAT AGTTCCCCAAAAGTTCTCTT	T→A	72	76,2	76,7	
E (P152)	rs9786634	AGTCTCAGAAATCAAGTTAGCT TGTGACTCAGAACTATTTCCTT	G→C	81	77,1	77,7	
D (M174)	rs2032602	GTACGTTTTTGGTTTACTCATAA AAAAGGAGAAGGACAAGAC	T→C	162	80,8	81,1	

E1a (M132)	rs2032617	CTACACTTAATAGAACACAAGCG CTTCGGTTATGTTTTCTCCAC	G→T	93	66,2	65,76	Multiplex 5A
E1b1b (M215)	rs2032654	CTTGCTGCATTAAGACAAACT TCCAGCACAGAAGCATCAG	A→G	59	77,3	78	
E1b1a7 (U186)	rs16980370	CCTTCTCGTAAGGGGCTG CTGGATAAGGAGTCCTTGGAG	A→G	73	78,8	79, 4	Multiplex 5B
E1b1a8 (U175)	rs16980588	GCTTATACTGGTCACACTAAGGC TCTAATGACCAGGAGAAGTCAAG	G→A	79	81,1	80,6	
E1b1a1 (M2)	rs9785941	CCCTGTTTAAAAATGTAGGTTT CCCTTTATCCTCCACAGATC	A→G	77	74	74,5	
R1b1a2 (V88)	rs2032658	ATTTCTCAGAGCAGGGAAC GCCTTACAACACCATCAAAATA	C→T	105	76,6	76,2	Multiplex 6
R (M207)	rs180946844	TATGGGGCAAATGTAAGTCAAG GAAGGAAAAGTGGAGTCTGA	A→G	124	77,9	78,2	

2.3.3 Controls chosen for project

Controls were chosen to show ancestral or derived states for the multiplex system from Ristow *et al.* snpchip analysed unpublished data that was generated at the University of Copenhagen in collaboration with UWC. The chip was used to either confirm the SNPs or the haplotype. This was done by extracting the SNPs with the Ytool (http://www.mitotool.org/ytool/index.html). The system used was the Infinium OmniExpress-24 kit.

2.4 High resolution melting (HRM)

KAPA™ HRM FAST PCR kit was used to perform this study. The kit came with a ready to use master mix that was designed for high performance detection of DNA sequences using HRM analysis. It contained a novel DNA polymerase for fast and effective DNA amplification. KAPA™ HRM FAST PCR kit contains EvaGreen® ; this is a next generation saturating fluorescent dye that binds to dsDNA. EvaGreen® can be used at higher concentrations without PCR inhibitions compared to SYBR® Green dye 1. EvaGreen® shows equal binding affinity to GC and AT rich regions. Because of the engineered polymerase and EvaGreen® combined in this product it allows for amplification and discrimination of the most challenging sequence differences (KapaBiosystems, 2016).

2.5 Optimisation of multiplexes

2.5.1 Single Plex

The manufacturer recommendations were followed for HRM reaction setup and cycling conditions. The manufacturer's conditions did not work with the primer sets that were used so optimisation took place. All amplification and HRM melting of samples was done as shown in Table 2.2 and cycling conditions as shown in Table 2.3. The manufacturer's recommendation of 45 cycles for the cycling conditions was kept as the primer sets amplified best with 45 cycles. Samples were amplified and melted on the Rotor-Gene® Q (36 sample carousels, Qiagen).

Table 2.2 Reagents and concentrations used for Y-SNP amplification

Reagents	Stock concentration	Volume (µl)	Final concentration
HRM Kit	2x	10	1x
Primers	10x	2	1x
$MgCl_2$	25 mM	2	2.5 mM
Sub-total		14	
DNA	20 ng/µl		2 ng/µl
Sub-total		16	
H_2O			
Final Volume		20	

Table 2.3 Cycling conditions and HRM melt conditions used for Y-SNP amplification

Stage	Temperature	Time		Cycles
Enzyme activation	95°C	5min		
Denaturation	95°C	5s	45 cycles	Acquire on HRM/green channel
Annealing and extension	60°C	25s		
Denaturation	95°C	1min		
HRM	65°C – 95°C	2s per temperature	Ramp at 0.1°C	Acquire on HRM channel

All primers were first amplified in single plex to check that they amplified correctly. This was done by taking 0.2 µM of each primer set. Set up conditions and cycling conditions were followed as directed in the tables above (Table 2.2-2.3). Primers that did not amplify at 0.2 µM were increased to a concentration of 0.3 µM or until the desired concentration was reached. The theoretical and experimental melting temperatures were compared after each primer set was run as a single plex. The primers were then grouped into multiplexes.

The multiplexes were arranged hierarchically with Multiplex 1 being the starting point. Figure 2.7 below is an illustration explaining the screening process that each sample should need to go through in order to infer its haplogroup.

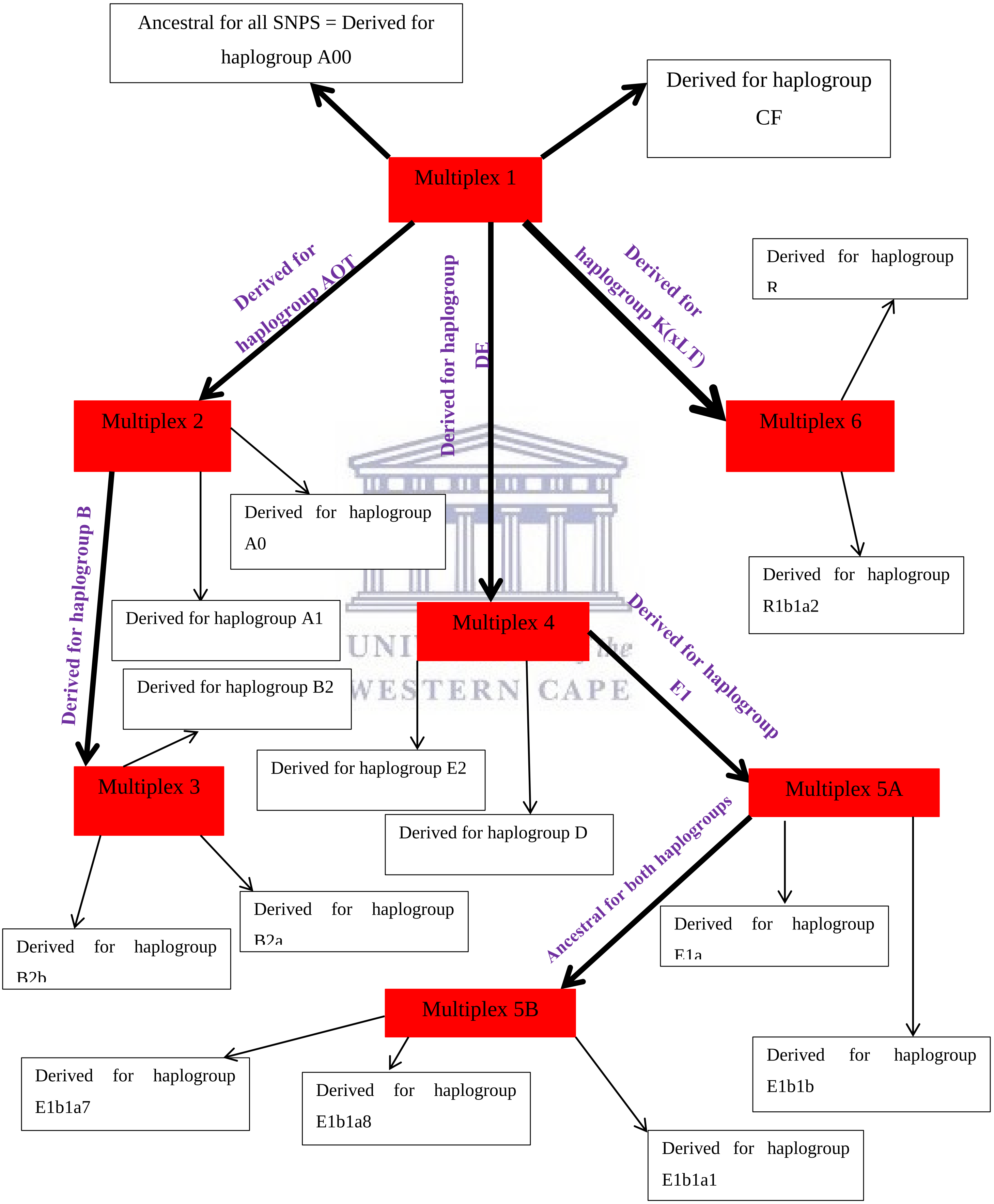

Ancestral for all SNPS = Derived for haplogroup A00
Derived for haplogroup CF
Multiplex 1
Derived for haplogroup AOT
Derived for haplogroup DE
Derived for haplogroup K(xLT)
Derived for haplogroup R
Multiplex 2
Multiplex 6
Derived for haplogroup A0
Derived for haplogroup B
Derived for haplogroup A1
Multiplex 4
Derived for haplogroup R1b1a2
Derived for haplogroup E1
Derived for haplogroup B2
Multiplex 3
Derived for haplogroup E2
Multiplex 5A
Derived for haplogroup D
Ancestral for both haplogroups
Derived for haplogroup B2a
Derived for haplogroup B2b
Derived for haplogroup E1a
Multiplex 5B
Derived for haplogroup E1b1b
Derived for haplogroup E1b1a7
Derived for haplogroup E1b1a8
Derived for haplogroup E1b1a1

Figure 2.7 Hierarchical screening processes of multiplexes. Red boxes indicate the 6 different multiplex systems. The bold black arrows indicate the direction that is followed in the screening process. Purple text indicates what haplogroup samples have to be derived for in order to continue to the next multiplex. The normal arrows pointing to the text boxes indicate what haplogroups samples could be derived for. The screening process always starts with Multiplex 1.

2.5.2 Multiplexes

Multiplexes were optimised by using the concentrations of the primer set single plexes. The same set up and cycling conditions were used for the multiplexes that were used for the single plexes (Table 2.2-2.3). Concentrations of the primer sets were altered as multiplexes were optimised. This can be seen below. Known controls were used to optimise each multiplex; these controls were sourced in-house.

2.5.2.1 Multiplex 1

Multiplex 1 was designed to be a starting point for each sample screened with this system. Table 2.4 below indicates which SNPs were targeted in this multiplex and how much of each primer set was used. Table 2.4 also indicates the theoretical ancestral and derived melting temperatures for each SNP in Multiplex 1. It indicates that there are significant melting temperature differences between each SNP. The significant melting temperature differences make sure that no SNP in the multiplex will overlap any other SNP. Fig. 2.8 illustrates the focus of this multiplex; the red circles show which haplogroups are being focused on.

According to Van Oven *et al.* (2014) haplogroup DE is defined by SNP M145 (Figure 1.5). In this multiplex system haplogroup DE is defined by SNP PF1439 according to the paper written by Francalacci *et al.* (2013). In its ancestral state PF1439 (rs16980598) has a G nucleotide and in its derived state PF1439 has an A nucleotide.

Haplogroup CF is defined by SNP P143 (Karafet *et al.*, 2008; Van Oven *et al.*, 2014). This can be seen in Figure 1.5. SNP P143 (rs4141886) is ancestral when a G nucleotide is present and derived when an A nucleotide is present.

Haplogroup K(xLT) is defined by SNP M526 (Wang *et al.*, 2010; Van Oven *et al.*, 2014). This can be seen in Figure 1.5. SNP M526 (rs2033003) in its ancestral state SNP M526 has an A nucleotide and in its derived state it has a G nucleotide.

According to Van Oven *et al.* (2014) haplogroup AOT is defined by SNP L1085. This can be seen in Figure 1.5. In this multiplex system haplogroup AOT is defined by the SNP AF3 according to Mendez *et al.* (2013). SNP AF3 is derived when the CCT nucleotides are present; in its ancestral state the CCT nucleotides have been deleted.

Table 2.4 Indicates the SNP markers, concentration used of each primer set in the multiplex and the theoretical ancestral and derived melting temperatures for Multiplex 1

Haplogroup (SNP Marker: YCC nomenclature)	**[primer set] used in multiplex**	**Ancestral melting temperature (°C)**	**Derived melting temperature (°C)**	**Difference between ancestral and derived melting temperatures.**
DE (PF1439)	0.3 µM	74,79	74,23	0.56
CF (P143)	0.12 µM	77,75	76,8	0.95
K(xLT) (M526)	0.13 µM	81,1	81,7	0.6
AOT (AF3)	0.13 µM		83,4	

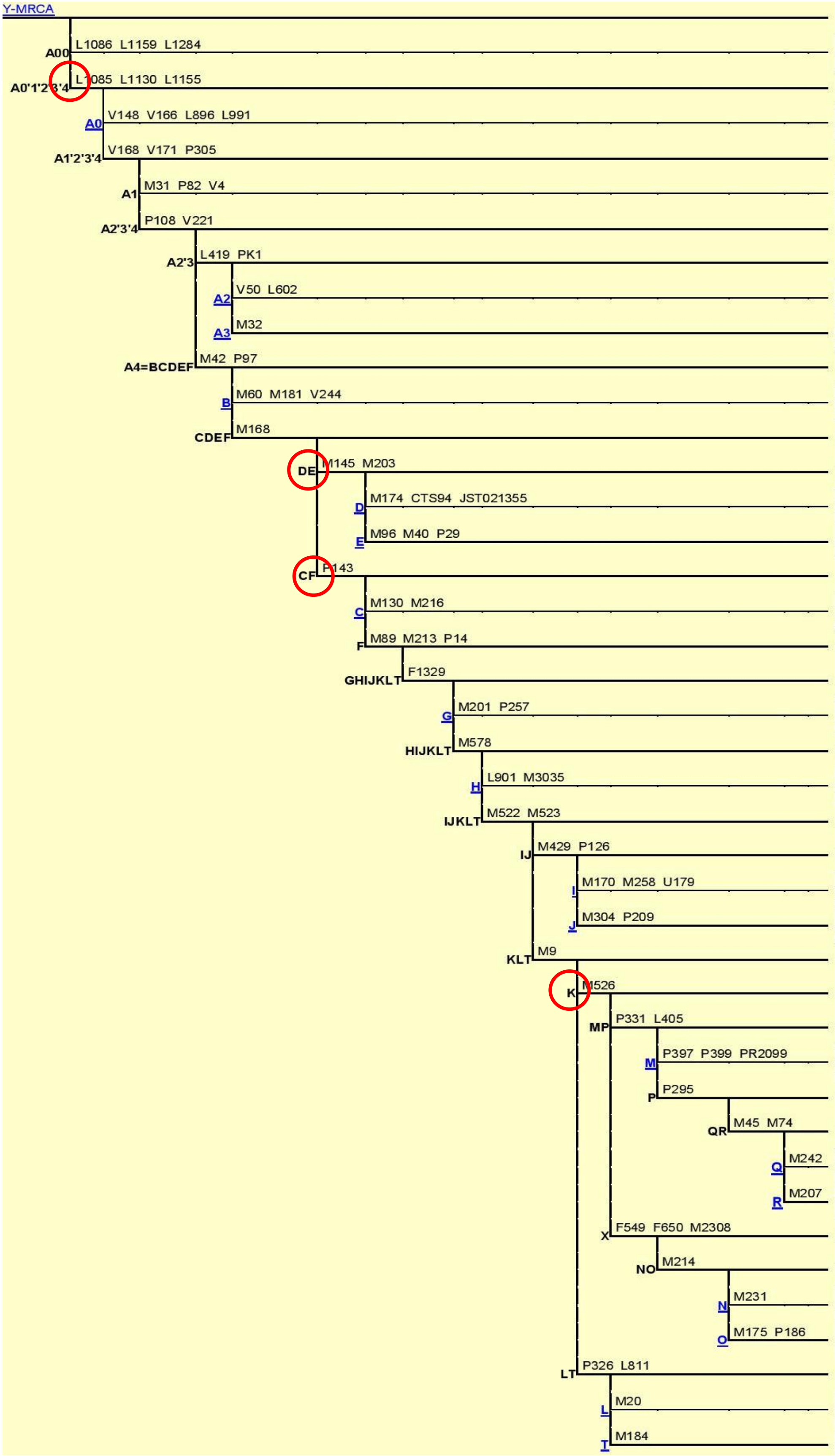


Figure 2.8 Skeleton of human Y-chromosome phylogeny. Red circles illustrate which haplogroups are being targeted in Multiplex 1. (Phylotree Y, last updated 9/03/2016, accessed 2017)

2.5.2.2 Multiplex 2

Multiplex 2 was designed to cover the top part of the Y-SNP tree. This multiplex contains primers to identify haplogroups A1, A0 and B. Table 2.5 below indicates which SNPs were targeted in this multiplex and how much of each primer set was used. Table 2.5 also indicates the theoretical ancestral and derived melting temperatures for each SNP in Multiplex 2. It indicates that there are significant melting temperature differences between each SNP. The significant melting temperature differences make sure that no SNP in the multiplex will overlap any other SNP. Fig. 2.9 illustrates the focus of this multiplex. The red circles show which haplogroups are being focused on.

Haplogroup A0 is defined by SNP V148 according to Van Oven *et al.* (2014) as can be seen in Figure 1.5. In this multiplex system haplogroup A0 is defined by SNP L1055 (https://isogg.org/tree/ISOGG_HapgrpA.html). In its ancestral state L1055 has a G nucleotide and in its derived state L1055 has an A nucleotide.

According to Van Oven *et al.* (2014) and Phylotree Y (http://www.phylotree.org/Y/tree/index.htm) SNP M31 defines haplogroup A1. However, according to ISOGG SNP M31 defines haplogroup A1a. This multiplex system originally defined SNP M31 as haplogroup A1 but has subsequently changed it to haplogroup A1a in order to be in compliance with ISOGG. However, Figure 2.9 still indicates that SNP M31 defines A1 and not haplogroup A1a as indicated in the previous sentence, as this image was taken from Phylotree Y. For the purpose of this thesis Figure 2.9 is an indication of where Multiplex 2 is focusing but the ISOGG change has been taken into consideration. SNP M31 (rs369315948) is ancestral when a G nucleotide is present and its derived state is when a C nucleotide is present.

Haplogroup B is defined by SNP M60 according to Van Oven *et al.* (2014). In this multiplex system haplogroup B is defined by SNP M181 (Underhill *et al.*, 2001). In its ancestral state SNP M181 (rs2032599) has a T nucleotide present and in its derived state it has a C nucleotide present.

Table 2.5 Indicates the SNP markers, concentration used of each primer set in the multiplex and the theoretical ancestral and derived melting temperatures for Multiplex 2

Haplogroup (SNP Marker: YCC nomenclature)	[primer set] used in multiplex	Ancestral melting temperature (°C)	Derived melting temperature (°C)	Difference between ancestral and derived melting temperatures.
A1a (M31)	0.2 µM	80.8	80.4	0.4
A0 (L1055)	0.2 µM	69,68	69,03	0.65
B (M181)	0.3 µM	75.1	75.5	0.4

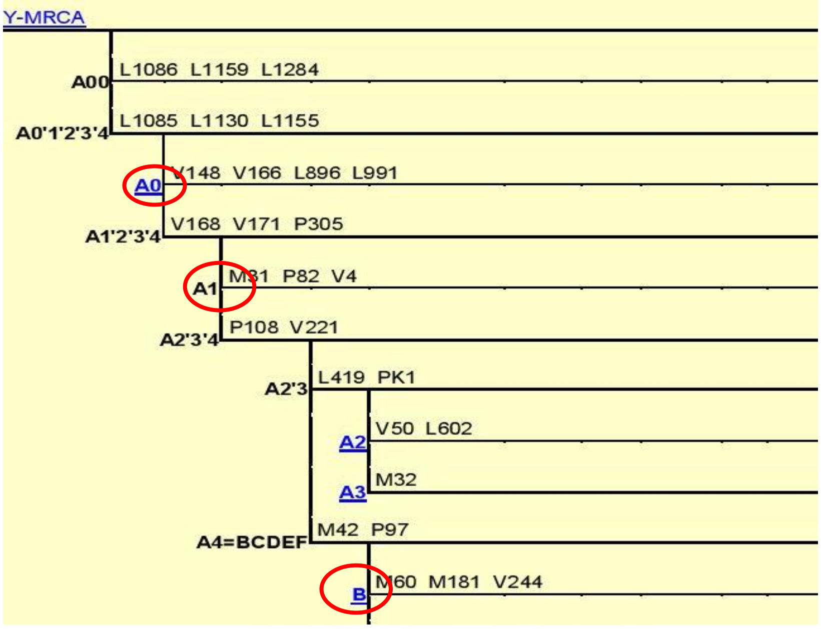


Figure 2.9 Skeleton of human Y-chromosome phylogeny. Red circles illustrate which haplogroups are being targeted in Multiplex 2. (Phylotree Y, last updated 9/03/2016, accessed 2017)

2.5.2.3 Multiplex 3

Multiplex 3 was designed to cover haplogroup B. This multiplex contains primers that identify haplogroups B2, B2b and B2a. Table 2.6 below indicates which SNPs were targeted in this multiplex and how much of each primer set was used. Table 2.6 also indicates the theoretical ancestral and derived melting temperatures for each SNP in Multiplex 3. It indicates that there are significant melting temperature differences between each SNP. The significant melting temperature differences make sure that no SNP in the multiplex will overlap any other SNP. Fig. 2.10 illustrates the focus of this multiplex. The red circles show which haplogroups are being focused on.

B2 is defined by SNP M182 (Underhill *et al.*, 2001; Van Oven *et al.*, 2014). SNP M182 (rs2032601) in its ancestral state has a C nucleotide and in its derived state a T nucleotide.

Haplogroup B2b is defined by SNP M112 (Underhill *et al.*, 2001; Van Oven *et al.*, 2014). SNP M112 (rs111725135) is ancestral if a G nucleotide is present and derived if an A nucleotide is present.

Haplogroup B2a is defined by SNP M150 according to Van Oven *et al.* (2014). This haplogroup was originally defined by SNP M152 (Underhill *et al.*, 2001) and this is the SNP that was used for this multiplex system. SNP M152 is now defined as haplogroup B2a1a according to Phylotree Y and ISOGG (http://www.phylotree.org/Y/tree/B.htm; https://isogg.org/tree/ISOGG_HapgrpB.html). M152 (rs371646183) is ancestral if a C nucleotide is present and derived if a T nucleotide is present.

Table 2.6 Indicates the SNP markers, concentration used of each primer set in the multiplex and the theoretical ancestral and derived melting temperatures for Multiplex 3

Haplogroup (SNP Marker: YCC nomenclature)	**[primer set] used in multiplex**	**Ancestral melting temperature (°C)**	**Derived melting temperature (°C)**	**Difference between ancestral and derived melting temperatures.**
B2 (M182)	0.2 µM	64,74	62,22	2.52
B2b (M112)	0.2 µM	75,7	75,3	0.4
B2a (M152)	0.2 µM	82,91	82,49	0.42

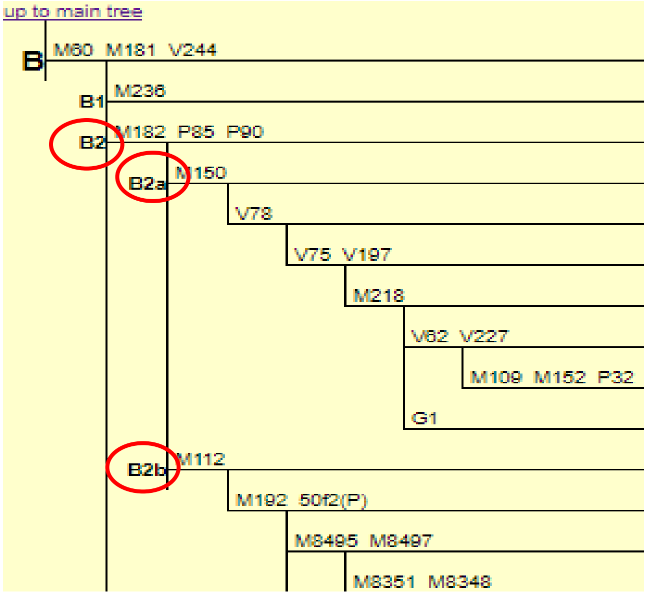


Figure 2.10 Skeleton of human Y-chromosome phylogeny. Red circles illustrate which haplogroups are being targeted in Multiplex 3. (Phylotree Y, last updated 9/03/2016, accessed 2017)

2.5.2.4 Multiplex 4

Multiplex 4 was designed to further investigate haplogroup DE. This multiplex contains primers to identify haplogroups E2, E1, E and D. Table 2.7 below indicates which SNPs were targeted in this multiplex and how much of each primer set was used. Table 2.7 also indicates the theoretical ancestral and derived melting temperatures for each SNP in Multiplex 4. It indicates that there are significant melting temperature differences between each SNP. The significant melting temperature differences make sure that no SNP in the multiplex will overlap any other SNP. Fig. 2.11 illustrates the focus of this multiplex. The red circles show which haplogroups are being focused on.

Haplogroup E2 is defined by SNP M75 (Underhill *et al.*, 2001; Van Oven *et al.*, 2014). SNP M75 (rs2032639) is ancestral if a G is present and is derived if an A is present.

Haplogroup E1 is defined with SNP P147 (Karafet *et al.*, 2008; Van Oven *et al.*, 2014). This can be seen in Figure 1.5. SNP P147 (rs16980577) is ancestral when a T nucleotide is present and derived if an A nucleotide is present.

Haplogroup E is defined by SNP M96 according to Van Oven *et al.* (2014). In this multiplex system haplogroup E is defined by SNP P152 according to Karafet *et al.* (2008). P152 (rs9786634) is derived when a C nucleotide is present and ancestral if a G nucleotide is present.

Haplogroup D is defined by SNP M174 (Underhill *et al.*, 2001; Van Oven *et al.*, 2014). M174 (rs2032602) is derived when a C nucleotide is present and ancestral when a T nucleotide is present.

Table 2.7 Indicates the SNP markers, concentration used of each primer set in the multiplex and the theoretical ancestral and derived melting temperatures for Multiplex 4

Haplogroup (SNP Marker: YCC nomenclature)	[primer set] used in multiplex	Ancestral melting temperature (°C)	Derived melting temperature (°C)	Difference between ancestral and derived melting temperatures.
E2 (M75)	0.1 µM	73,6	73,1	0.5
E1 (P147)	0.1 µM	76,2	76,7	0.5
E (P152)	0.08 µM	77,1	77,7	0.6
D (M174)	0.22 µM	80,8	81,1	0.3

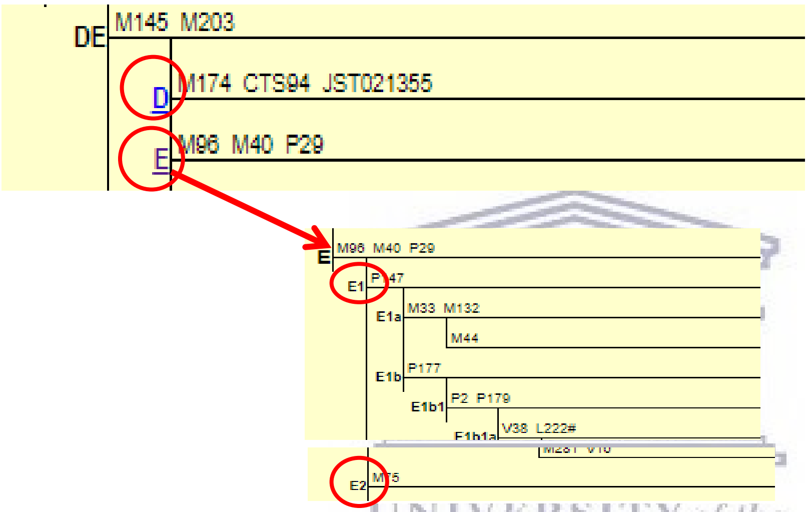


Figure 2.11 Skeleton of human Y-chromosome phylogeny. Red circles illustrate which haplogroups are being targeted in Multiplex 4. (Phylotree Y, last updated 9/03/2016, accessed 2017)

2.5.2.5 Multiplex 5A

Multiplex 5A was designed to further investigate haplogroup E1. This multiplex contains primers to identify haplogroups E1a and E1b1b. Table 2.8 below indicates which SNPs were targeted in this multiplex and how much of each primer set was used. Table 2.8 also indicates the theoretical ancestral and derived melting temperatures for each SNP in Multiplex 5A. It indicates that there are significant melting temperature differences between each SNP. The significant melting temperature differences make sure that no SNP in the multiplex will overlap any other SNP. Fig. 2.12 illustrates the focus of this multiplex. The red circles show which haplogroups are being focused on.

Haplogroup E1a is defined by SNP M33 according to Van Oven *et al.* (2014). This can also be seen in Figure 1.5. In this multiplex system haplogroup E1a is defined by SNP M132 ((Underhill *et al.*, 2001). SNP M132 (rs2032617) is ancestral when a G nucleotide is present and derived when a T nucleotide is present.

Haplogroup E1b1b is defined by the SNP M215 (Underhill *et al.*, 2001; Van Oven *et al.*, 2014). M215 (rs2032654) in its ancestral state has an A nucleotide and in its derived state has a G nucleotide.

Table 2.8 Indicates the SNP markers, concentration used of each primer set in the multiplex and the theoretical ancestral and derived melting temperatures for Multiplex 5A

Haplogroup (SNP Marker: YCC nomenclature)	**[primer set] used in multiplex**	**Ancestral melting temperature (°C)**	**Derived melting temperature (°C)**	**Difference between ancestral and derived melting temperatures.**
E1a (M132)	0.4 µM	66,2	65,76	0.44
E1b1b (M215)	0.3 µM	77,3	78	0.7

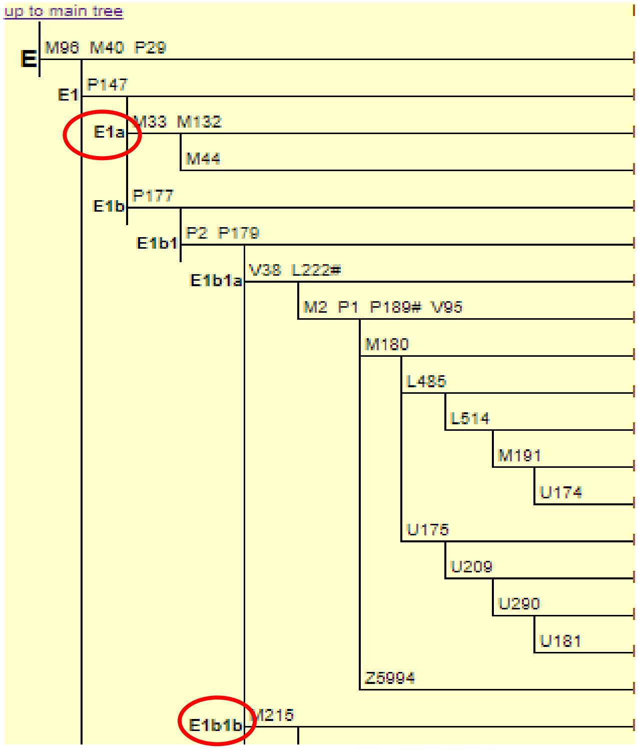


Figure 2.12 Skeleton of human Y-chromosome phylogeny. Red circles illustrate which haplogroups are being targeted in Multiplex 5A. (Phylotree Y, last updated 9/03/2016, accessed 2017)

Multiplex 5A had difficulty amplifying on the Rotor-Gene® Q (36 sample carousels, Qiagen). A temperature gradient was performed on the multiplex ranging from 54-64°C; this can be seen in red in Table 2.9. The amplification was performed on Veriti 96 well thermal cycler (*Applied Biosystems*). Samples were then melted on the Rotor-Gene® Q (36 sample carousels, Qiagen) in the same tubes following the same melting conditions of every other multiplex system (Table 2.10).

Table 2.9 Cycling conditions for temperature gradient amplification

Stage	Temperature	Time (seconds)	Cycles
Enzyme activation	95°C	300	
Denaturation	95°C	5	
Annealing and extension	54-64°C	25	45 cycles
Denaturation	95°C	60	

Table 2.10 Melting conditions for Rotor-Gene® Q (36 sample carousels, Qiagen)

HRM	65°C – 95°C	2s per temperature	Ramp at 0.1°C	Acquire on HRM channel

It was noted that Multiplex 5A amplified the best at 62°C. Amplification was then compared between two thermocyclers, the Veriti 96 well thermal cycler (*Applied Biosystems*) and the Thermo Scientific Arktik Thermal Cycler. The Thermo Scientific Arktik Thermal Cycler amplified Multiplex 5A the best as two peaks were visible after melting on the Rotor-Gene® Q (36 sample carousels, Qiagen). The amplification of Multiplex 5A on the Veriti 96 well thermal cycler (*Applied Biosystems*) only produced a peak for SNP M215 after melting on the Rotor-Gene® Q (36 sample carousels, Qiagen).

Cycling conditions for Multiplex 5A were set up on the Thermo Scientific Arktik Thermal Cycler as stated in the lilac part of Table 2.11. The Rotor-Gene® Q (36 sample carousels, Qiagen) was used for identifying the melting temperatures of the samples. Rotor-Gene® Q (36 sample carousels, Qiagen) melting conditions were set up as stated in the light blue of Table 2.11.

Table 2.11 Cycling (lilac) and melting (light blue) conditions used for Multiplex 5A

Stage	**Temperature**	**Time (seconds)**		**Cycles**
Enzyme activation	95°C	300		
Denaturation	95°C	5		
Annealing and extension	62°C	25	45 cycles	Acquire on HRM/green channel
Denaturation	95°C	60		
HRM	65°C – 95°C	2s per temperature	Ramp at 0.1°C	Acquire on HRM channel

2.5.2.6 Multiplex 5B

Multiplex 5B was designed to further investigate haplogroup E1. This multiplex contains primers to identify haplogroups E1b1a1, E1b1a7 and E1b1a8. Table 2.12 below indicates which SNPs were targeted in this multiplex and how much of each primer set was used. Table 2.12 also indicates the theoretical ancestral and derived melting temperatures for each SNP in Multiplex 5B. It indicates that there are significant melting temperature differences between each SNP. The significant melting temperature differences make sure that no SNP in the multiplex will overlap any other SNP. Fig. 2.13 illustrates the focus of this multiplex. The red circles show which haplogroups are being focused on.

Haplogroup E1b1a1 is defined by SNP M2 (Selelstad et al., 1994). M2 (rs9785941) is ancestral if an A nucleotide is present and derived if a G nucleotide is present.

Haplogroup E1b1a7 is defined by SNP U186 (Sims, Garvey and Ballantyne, 2007). U186 (rs16980370) is in an ancestral state if an A nucleotide is present. It has a derived state if a G nucleotide present; this makes the melting temperature higher.

Haplogroup E1b1a8 is defined by SNP U175 (Sims, Garvey and Ballantyne, 2007). U175 (rs16980588) is ancestral when a G nucleotide is present and derived when an A nucleotide is present.

Table 2.12 Indicates the SNP markers, concentration used of each primer set in the multiplex and the theoretical ancestral and derived melting temperatures for Multiplex 5B

Haplogroup (SNP Marker: YCC nomenclature)	**[primer set] used in multiplex**	**Ancestral melting temperature (°C)**	**Derived melting temperature (°C)**	**Difference between ancestral and derived melting temperatures.**
E1b1a7 (U186)	0.2 µM	78,8	79,4	0.6
E1b1a8 (U175)	0.2 µM	81,1	80,6	0.5
E1b1a1 (M2)	0.3 µM	74	74,5	0.5

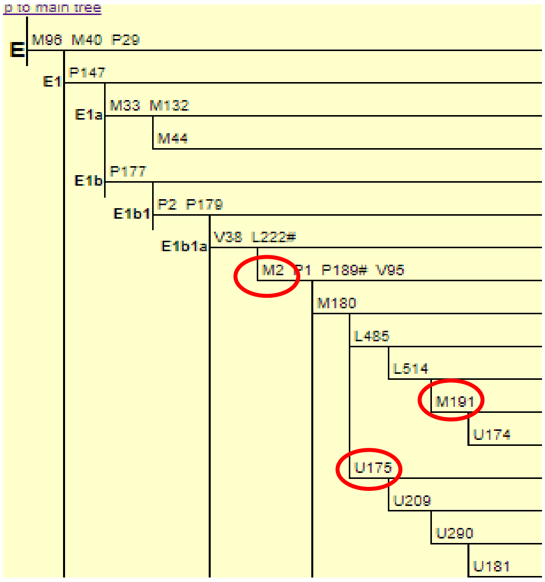


Figure 2.13 Skeleton of human Y-chromosome phylogeny. Red circles illustrate which haplogroups are being targeted in Multiplex 5B. (Phylotree Y, last updated 9/03/2016, accessed 2017)

Multiplex 5B had difficulty amplifying on the Rotor-Gene® Q (36 sample carousels, Qiagen). A temperature gradient was performed on the multiplex ranging from 54-64°C; this can be seen in red in Table 2.13. The amplification was performed on Veriti 96 well thermal cycler (*Applied Biosystems*). Samples were then melted on the Rotor-Gene® Q (36 sample carousels, Qiagen) in the same tubes following the same melting conditions of every other multiplex system (Table 2.14).

Table 2.13 Cycling conditions for temperature gradient amplification

Stage	Temperature	Time (seconds)	Cycles
Enzyme activation	95°C	300	
Denaturation	95°C	5	
Annealing and extension	54-64°C	25	45 cycles
Denaturation	95°C	60	

Table 2.14 Melting conditions for Rotor-Gene® Q (36 and 72 sample carousels, Qiagen)

HRM	65°C – 95°C	2s per temperature	Ramp at 0.1°C	Acquire on HRM channel

It was noted that Multiplex 5B amplified the best at 56°C.

Cycling conditions for Multiplex 5B were set up on the Veriti 96 well thermal cycler (*Applied Biosystems*) as stated in the light green part of Table 2.15. The Rotor-Gene® Q (36 sample carousels, Qiagen) was used for identifying the melting temperatures of the samples. Rotor-Gene® Q (36 and 72 sample carousels, Qiagen) melting conditions were set up as stated in the light pink of Table 2.15.

Table 2.15 Cycling (light green) and melting (light pink) conditions used for Multiplex 5B

Stage	**Temperature**	**Time (seconds)**		**Cycles**
Enzyme activation	95°C	300		
Denaturation	95°C	5	45 cycles	Acquire on HRM/green channel
Annealing and extension	62°C	25		
Denaturation	95°C	60		
HRM	65°C – 95°C	2s per temperature	Ramp at 0.1°C	Acquire on HRM channel

2.5.2.7 Multiplex 6

Multiplex 6 was designed to further investigate samples that are derived for haplogroup K(xLT). This multiplex contains primers to identify haplogroups R and R1b1a2. Table 2.16 below indicates which SNPs were targeted in this multiplex and how much of each primer set was used. Table 2.16 also indicates the theoretical ancestral and derived melting temperatures for each SNP in Multiplex 6. It indicates that there are significant melting temperature differences between each SNP. The significant melting temperature differences make sure that no SNP in the multiplex will overlap any other SNP. Fig. 2.14 illustrates the focus of this multiplex. The red circles show which haplogroups are being focused on.

Haplogroup R is defined by SNP M207 (Underhill *et al.*, 2001; Van Oven *et al.*, 2014). M207 (rs2032658) is ancestral if an A nucleotide is present and derived if a G nucleotide is present. This would mean that the melting temperature of samples that are derived for this haplogroup will be higher than those that are ancestral for the haplogroup.

Haplogroup R1b1a2 is defined by SNP V88 (Cruciani *et al.*, 2002; Van Oven *et al.*, 2014). SNP V88 (rs180946844) is ancestral if a C nucleotide is present and derived if a T nucleotide is present.

Table 2.16 SNP markers and concentrations of each primer in Multiplex 6

Haplogroup (SNP Marker: YCC nomenclature)	[primer set] used in multiplex	Ancestral melting temperature (°C)	Derived melting temperature (°C)	Difference between ancestral and derived melting temperatures.
R1b1a2 (V88)	0.4 µM	76,6	76,2	0.4
R (M207)	0.2 µM	77,9	78,2	0.3

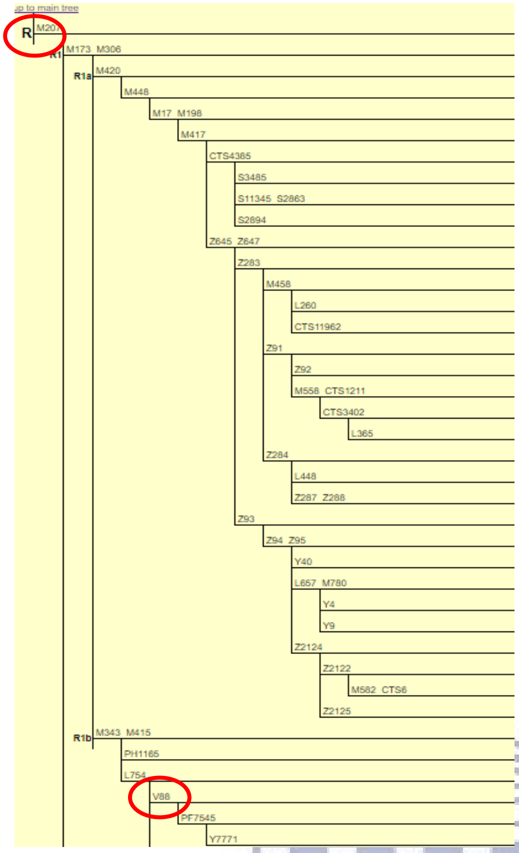


Figure 2.14 Skeleton of human Y-chromosome phylogeny. Red circles illustrate which haplogroups are being targeted in Multiplex 6. (Phylotree Y, last updated 9/03/2016, accessed 2017)

2.6 Johannesburg Coloured (Admixed) population samples

The newly developed multiplexes were validated by screening previously collected samples from the Johannesburg Coloured population.

2.6.1 Collection of samples

Ethical clearance for this project was obtained from the University of the Western Cape (15-4-97). Each donor had to sign a consent form permitting the lab to use their DNA post sampling. Johannesburg Coloured samples were collected in 2014. Saliva and buccal swab samples were collected from 143 individuals that classified themselves as Coloured. To help the donors increase salivation they were given sweets. Once they had produced enough saliva they spat into a 15ml tube; equal amounts of storage buffer was added to the saliva sample. Samples were kept at room temperature until -20°C storage could be accessed. The samples were stored at -20ºC to help prolong the preservation of the DNA in the saliva.

2.6.2 DNA Extraction

2.6.2.1 Organic DNA extraction for saliva samples

The saliva samples were first extracted. An optimised organic DNA extraction protocol was used. In a clean Eppendorf tube 5 µl of 10mg/ml proteinase K was added to 500 µl of saliva sample and incubated at 60°C for 4 hours with orbital shaking at 40rpm on a Labnet Vortemp 1550. After incubation 600 µl Phenol: Chloroform: Isoamyl (PCI, 25:24:1) was added to the Eppendorf tubes and each sample was vortexed for 15 seconds. The samples were then centrifuged for 10 minutes at 10 000rpm at room temperature. The supernatant was then removed and placed into a new Eppendorf tube and an equal amount of chloroform was added. The tubes were then vortexed for 30 seconds and centrifuged for 10 minutes at 10 000rpm. The supernatant was removed and placed into a new Eppendorf tube and an equal amount of chloroform was added. The tubes were then vortexed for 30 seconds and centrifuged for 10 minutes at 10 000rpm. The supernatant was then removed and placed into a new Eppendorf tube.

Double the volume of ice cold isopropanol was added to the samples. The tubes were then placed into -20°C freezer overnight to precipitate the DNA. The tubes were then centrifuged for 15 minutes at 14 000rpm. The supernatant was removed and the pellet was washed with 200 µl ice cold 70% ethanol. The tubes were then centrifuged for 15 minutes at 14 000rpm. The supernatant was removed again and washed with 70% ethanol again. The supernatant was then removed and inverted on a paper towel until the tube was dry. The pellet was then resuspended in 50 µl 1X TE buffer and incubated at 55°C for 15 minutes with orbital shaking at 30rpm on a Labnet Vortemp 1550. Tubes then stood at room temperature for 45 minutes before quality and quantity of DNA was checked. The quantity and quality of the DNA was then checked using a Nanodrop ND-2000 UV spectrophotometer.

2.6.2.2 Salting out DNA extraction for buccal swabs

Some Johannesburg saliva samples did not have a good DNA concentration yield; therefore DNA was extracted from the swab samples according to Medrano (1990). The swab tips were cut off using a sterilised pair of scissors and placed into an Eppendorf tube that contained 600 µl lysis buffer and 3 µl Proteinase K (20mg/ml). Samples were vortexed for 30 seconds then incubated at 56°C overnight with orbital shaking at 40rpm on a Labnet Vortemp 1550. Lysis buffer that contained biological material was removed and placed into a new clean Eppendorf tube. The biological material still trapped in the swab tip was recovered by placing the swab tip in a perforated 0.5ml tube which was placed inside a 1.5ml Eppendorf tube. The tubes were then centrifuged for 5 minutes at 14 000rpm. The collected volume was added to the previously collected volume. Precipitation took place by adding 1/3 volume of 5M NaCl and tubes were shaken vigorously for 15 seconds. Samples were then centrifuged for 15 minutes at 10 000rpm.

The supernatant was transferred into a clean Eppendorf tube and equal volume of ice cold isopropanol was added to each tube. Tubes were placed into the -20°C freezer overnight. The tubes were then centrifuged for 15 minutes at 14 000rpm. Supernatant was removed and pellet was washed with 100 µl of ice cold 70% ethanol. Tubes were centrifuged for 15 minutes at 14 000rpm. The ethanol was discarded and pellet was dried. DNA pellet was dissolved in 50 µl 1X TE buffer. The quantity and quality of the DNA was then checked using a Nanodrop ND-2000 UV spectrophotometer.

2.6.3 Screening samples using Y-SNP multiplex system

A total of 69 male Johannesburg samples were genotyped using the Y-SNP multiplex system. All samples were first screen with Multiplex 1; then screening process illustrated in Figure 2.7 was followed.

2.7 Analysis of data

2.7.1 Rotor-Gene® ScreenClust HRM® software

Rotor-Gene ScreenClust HRM software was used to analysis the Johannesburg samples. Rotor-Gene ScreenClust HRM software is a powerful analysis tool for high resolution melting. The software groups samples into clusters based on their melting temperatures. This tool was used to confirm the haplogroup assignment from the visual analysis.

2.7.2 Chi-Square test analysis

A comparison of the relative contributions of African, European and Asian ancestry between the Johannesburg Coloured population and the Western Cape Coloured population was performed through a Chi-square test in excel.

Chapter 3: Results and Discussion

3.1 Multiplex Optimisation

3.1.1 Multiplex 1

As stated in Chapter 2 the multiplex systems were designed according to the hierarchy of the Y-SNP tree. Multiplex 1 consists of four SNPs and it infers samples to basal branch haplogroups. Figure 3.1 represents a melt curve analysis of Multiplex 1; it demonstrates how all four SNPs melt at different temperatures.

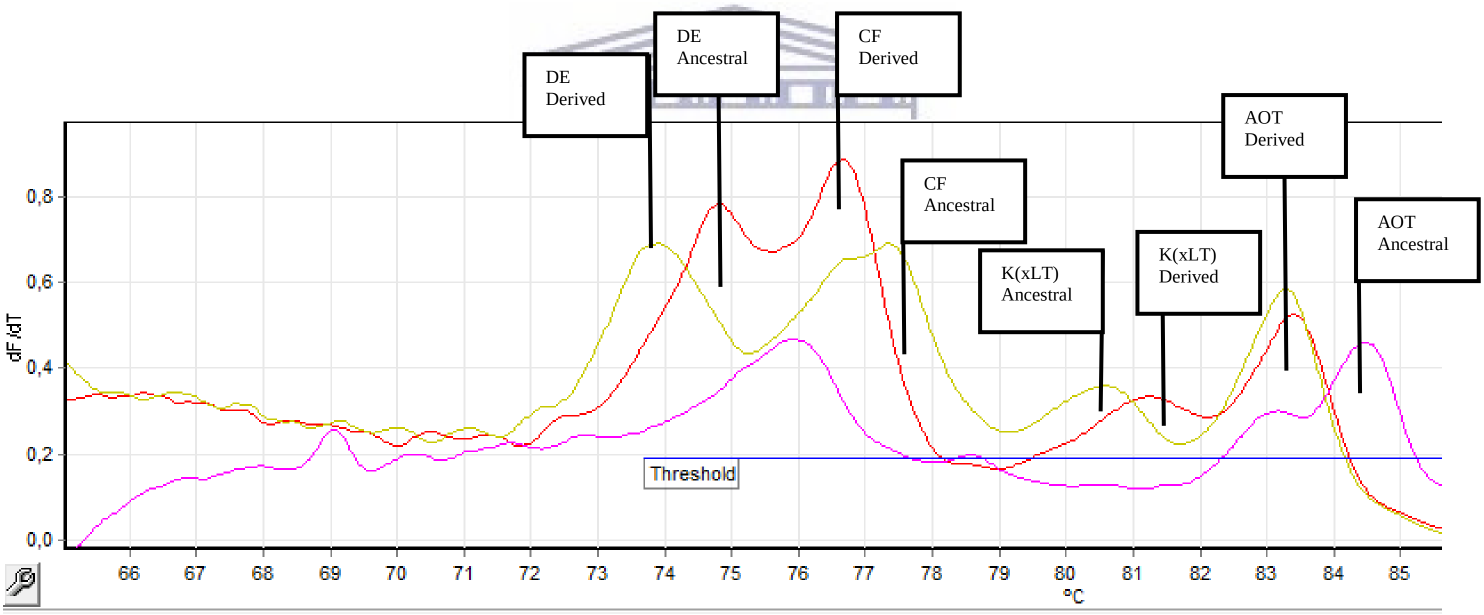


Figure 3.1 Melting curve analysis of Multiplex 1. Sample that belongs to haplogroup DE (yellow), sample belonging to haplogroup CF and K(xLT) (red) and sample ancestral for haplogroup AOT (pink) are shown.

The yellow curve is a sample that is derived for haplogroup DE. It can be seen in Figure 3.1 that the yellow curve has a lower temperature than the red curve indicating that it is indeed derived for haplogroup DE.

It can be seen in Figure 3.1 that the red curve is derived for haplogroup CF as the red curve has a lower temperature than the yellow curve.

The red curve in Figure 3.1 is derived for haplogroup K(xLT) as its melting temperature is higher than the yellow curve temperature.

Both the yellow and red curve are derived for haplogroup AOT. The pink curve is ancestral for haplogroup AOT and has a higher melting temperature than the two samples that are derived for the haplogroup (Figure 3.1).

The multiplex system was simple and clear to understand once the ancestral and derived melting temperatures were clarified. It was noted that the ancestral and derived melting temperatures did shift slightly from the theoretical to the experimental ones. This can be seen in Table 3.1. The theoretical ancestral and derived melting temperatures were determined using Oligo v7 (Wojciech, 2010). The shift in temperature could be due to the purity of the DNA sample that was used. Samples that had low concentrations or were not clean tended to shift to the right in temperature.

Table 3.1 Theoretical (Blue) and experimental (Purple) Tm of ancestral and derived states of amplicons for Multiplex 1. Average temperature of each SNP was calculated over 4 different runs and standard deviation was calculated from the average temperature of each run.

Haplogroup (SNP marker: YCC nomenclature)	Ancestral melting temperature (°C)	Derived melting temperature (°C)	Ancestral melting temperature (°C) (average temperature in degree Celsius, **σ**)	Derived melting temperature (°C) (average temperature in degree Celsius, **σ**)
DE (PF1439)	74,79	74,23	74,82 (74.36, **0.17**)	73,9 (73.77, **0.15**)
CF (P143)	77,75	76,8	77,35 (77.00, **0.13**)	76,65 (76.30, **0.07**)
K(xLT) (M526)	81,1	81,7	80,6 (80.35, **1.17**)	81,23 (80.99, **0.17**)
AOT (AF3)		83,4	84,48	83,4 (82.86, **0.11**)

3.1.2 Multiplex 2

Multiplex 2 consists of three SNPs that are used if samples are only derived for haplogroup AOT. In Figure 3.2 the ancestral state of haplogroup A0 and A1a can be seen. The pink curve is derived for haplogroup B. This can be seen in Figure 3.2 where the pink curve’s temperature is higher than that of the red curve.

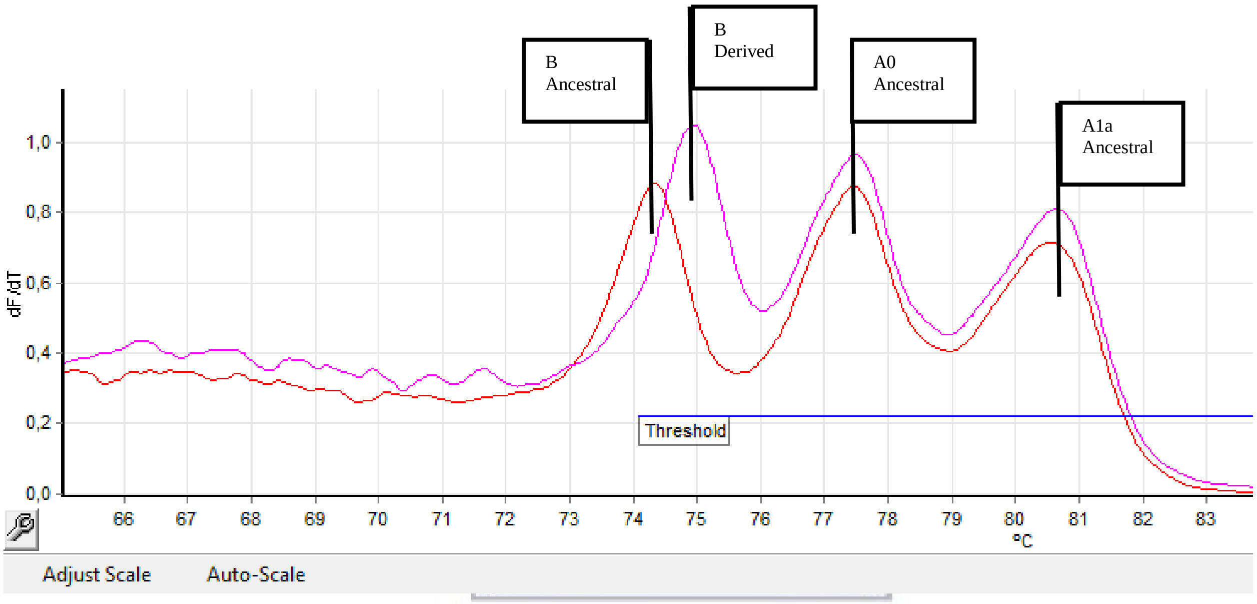


Figure 3.2 Melting curve analysis of Multiplex 2. Sample derived for haplogroup B (red) and sample ancestral for haplogroup B (pink) are shown.

Multiplex 2 was easy to interpret; it was unfortunate that positive controls for haplogroup A1a and haplogroup A0 were not available in-house. It can be seen in Table 3.2 that the theoretical temperatures are different to the experimental ancestral temperatures. The experimental Tm of the amplicon for the ancestral state of SNP M31 showed a slight decrease in comparison with the theoretical Tm. The experimental Tm of the amplicon for the ancestral state of SNP L1055 showed an increase of 6ºC in comparison with the theoretical Tm. The experimental Tm of the amplicon for the ancestral and derived state of SNP M181 showed a decrease by 2ºC in comparison with the theoretical Tm. The shift in temperatures could be due to the reaction mixture of the concentration of each primer set used in the reaction mix. Although there was a shift in temperatures for each primer set the temperatures did not overlap one another so it was easy to identify each haplogroup.

Table 3.2 Theoretical (Blue) and experimental (Purple) Tm of ancestral and derived states of amplicons for Multiplex 2. Average temperature of each SNP was calculated over 4 different runs and standard deviation was calculated from the average temperature of each run.

Haplogroup (SNP marker: YCC nomenclature)	Ancestral melting temperature (°C)	Derived melting temperature (°C)	Ancestral melting temperature (°C) (average temperature in degree Celsius, **σ**)	Derived melting temperature (°C) (average temperature in degree Celsius, **σ**)
A1a (M31)	80.8	80.4	79.92 (80.07, **0.13**)	-
A0 (L1055)	69,68	69,03	76.55 (76.72, **0.37**)	-
B (M181)	75.1	75.5	73.37 (73.58, **0.27**)	73.91 (74.11, **0.40**)

3.1.3 Multiplex 3

Multiplex 3 consists of three SNPs that are used if samples are derived for haplogroup B in Multiplex 2. In Figure 3.3 both curves are derived for B2 and in Figure 3.4 the yellow and pink curve are derived for haplogroup B2. The blue curve in Figure 3.3 and the pink curve in Figure 3.4 are derived for haplogroup B2b. In both Figures 3.3 and 3.4 the derived sample for B2b has a lower melting temperature than the ancestral samples.

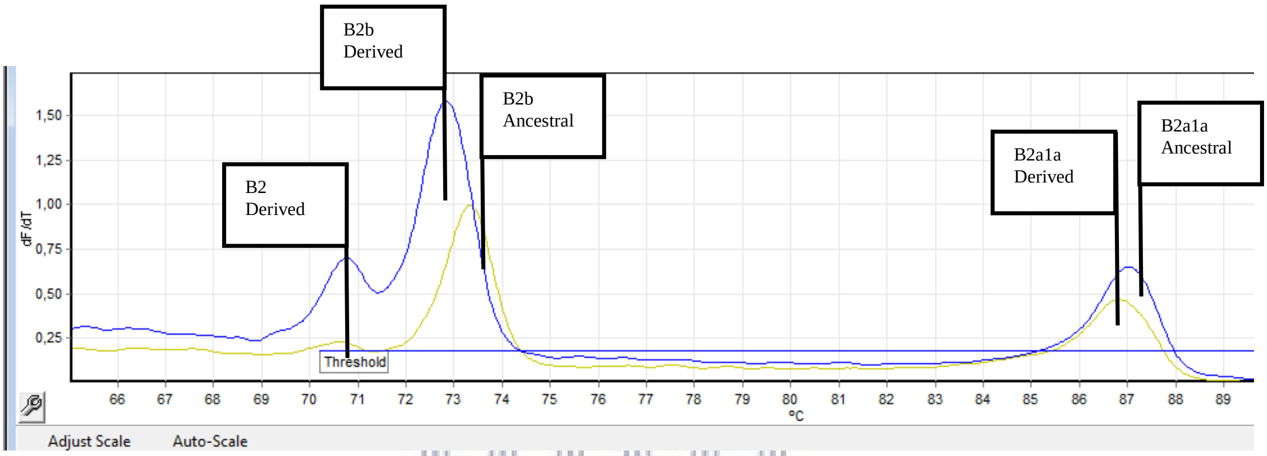


Figure 3.3 Melting curve analysis of Multiplex 3. Samples derived for haplogroup B2 (yellow and blue), sample derived for haplogroup B2b (blue) and sample derived for haplogroup B2a1a (yellow) are shown.

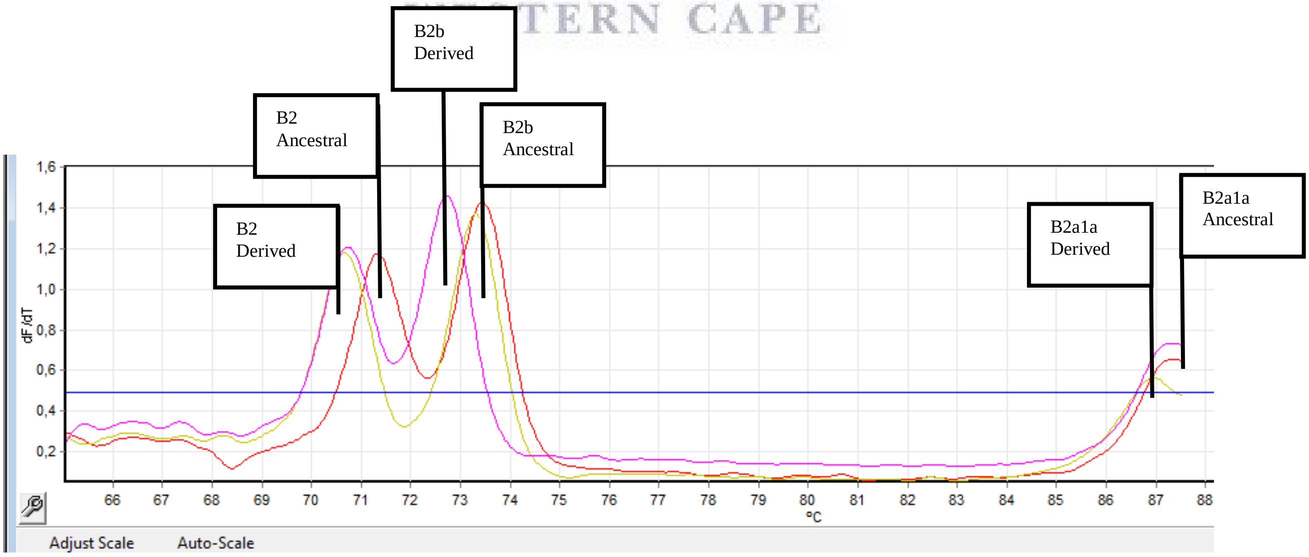


Figure 3.4 Melting curve analysis of Multiplex 3. Ancestral states are shown for all haplogroups. However, B2a melting curve was cut off due to a file corruption. Sample ancestral for all three haplogroups (red), samples derived for haplogroup B (pink and yellow), sample derived for haplogroup B2b (pink) and sample derived for haplogroup B2a1a (yellow) are shown.

In both Figure 3.3 and 3.4 the yellow curve is derived for haplogroup B2a1a. In both figures one can see that the derived samples have a lower melting temperature than the samples that are ancestral for haplogroup B2a1a.

The experimental Tm of the amplicon for the ancestral and derived state of SNP M182 showed a vast increase in temperature in comparison with the theoretical Tm. The experimental Tm of the amplicon for the ancestral and derived state of SNP M152 showed an increase in temperature in comparison with the theoretical Tm as well. The experimental Tm of the amplicon for the ancestral and derived state of SNP M112 showed a decrease in temperature in comparison with the theoretical Tm. All this can be seen in Table 3.3. Even though the temperatures had shifted this did not hinder the identification of the three different SNPs present in Multiplex 3. Temperatures could have shifted due to the concentration of each primer set used in Multiplex 3.

Table 3.3 Theoretical (Blue) and experimental (Purple) Tm of ancestral and derived states of amplicons for Multiplex 3. Average temperature of each SNP was calculated over 4 different runs and standard deviation was calculated from the average temperature of each run.

Haplogroup (SNP marker: YCC nomenclature)	Ancestral melting temperature (°C)	Derived melting temperature (°C)	Ancestral melting temperature (°C) (average temperature in degree Celsius, **σ**)	Derived melting temperature (°C) (average temperature in degree Celsius, **σ**)
B2 (M182)	64,74	62,22	71,77 (71.63, **0.27**)	70,73 (71.16, **0.35**)
B2b (M112)	75,7	75,3	73,33 (73.75, **0.34**)	72,87 (72.74, **0.69**)
B2a (M152)	82,91	82,49	87,37 (87.28, **0.17**)	86,95 (86.83, **0.26**)

3.1.4 Multiplex 4

Multiplex 4 consists of four SNPs that are used if samples are derived for haplogroup DE in Multiplex 1.

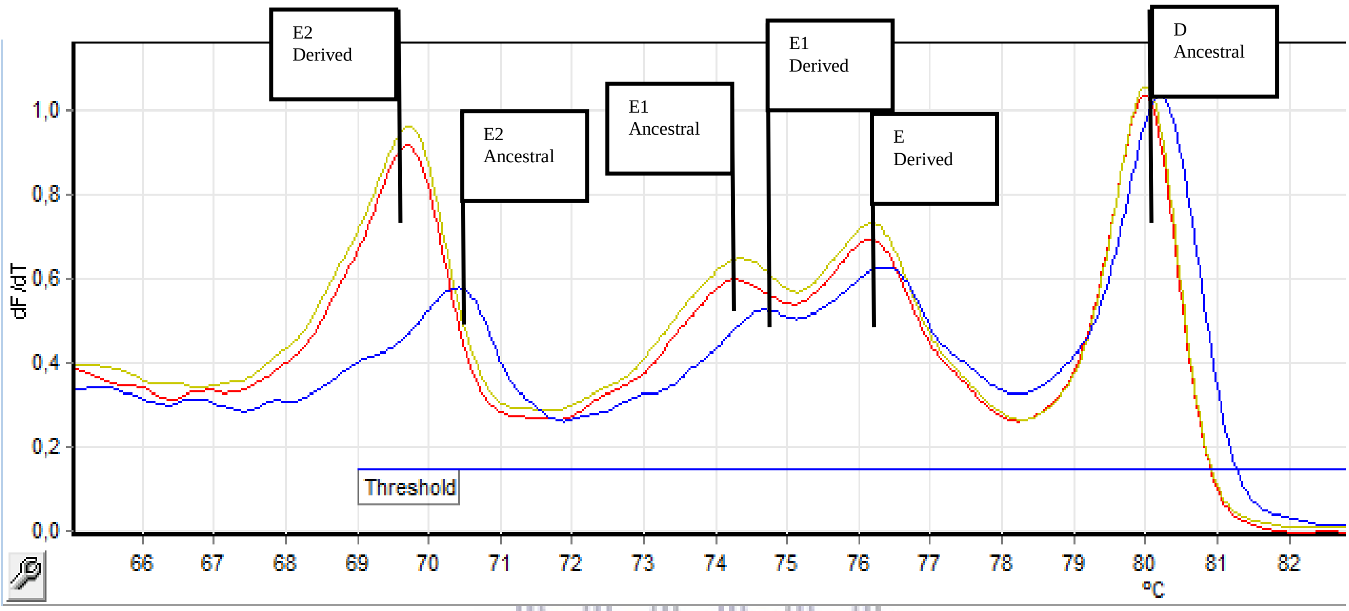


Figure 3.5 Melting curve analysis of Multiplex 4. Samples derived for haplogroup E2 (yellow and red), sample derived for haplogroup E1 (blue), samples derived for haplogroup E1 (red, yellow and blue) and samples ancestral for haplogroup D (red, yellow and blue) are shown.

The yellow and red curves are derived for haplogroup E2 in Figure 3.5. This can be seen in Figure 3.5 as both the yellow and red curve have a lower melting temperature compared to the ancestral sample. The blue curve in Figure 3.5 shows the derived state for haplogroup E1 as its melting temperature is higher than the melting temperature of the ancestral samples. All samples are derived for haplogroup E in Figure 3.5. Figure 3.5 shows the ancestral state of haplogroup D.

Table 3.4 represents the theoretical and experimental ancestral and derived melting temperatures for Multiplex 4.

Table 3.4 Theoretical (Blue) and experimental (Purple) Tm of ancestral and derived states of amplicons for Multiplex 4. Average temperature of each SNP was calculated over 4 different runs and standard deviation was calculated from the average temperature of each run.

Haplogroup (SNP marker: YCC nomenclature)	Ancestral melting temperature (°C)	Derived melting temperature (°C)	Ancestral melting temperature (°C) (average temperature in degree Celsius, **σ**)	Derived melting temperature (°C) (average temperature in degree Celsius, **σ**)
E2 (M75)	73,6	73,1	70,40 (70.39, **0.15**)	69.70 (69.92, **0.38**)
E1 (P147)	76,2	76,7	74,25 (74.08, **0.46**)	74,68 (74.48, **0.11**)
E (P152)	77,1	77,7	76,95 -	76,47 (76.19, **0.22**)
D (M174)	80,8	81,1	80,37 (80.06, **0.16**)	- -

3.1.5 Multiplex 5A

Multiplex 5A consists of two SNPs that are used when a sample is derived for haplogroup E1 in Multiplex 4.

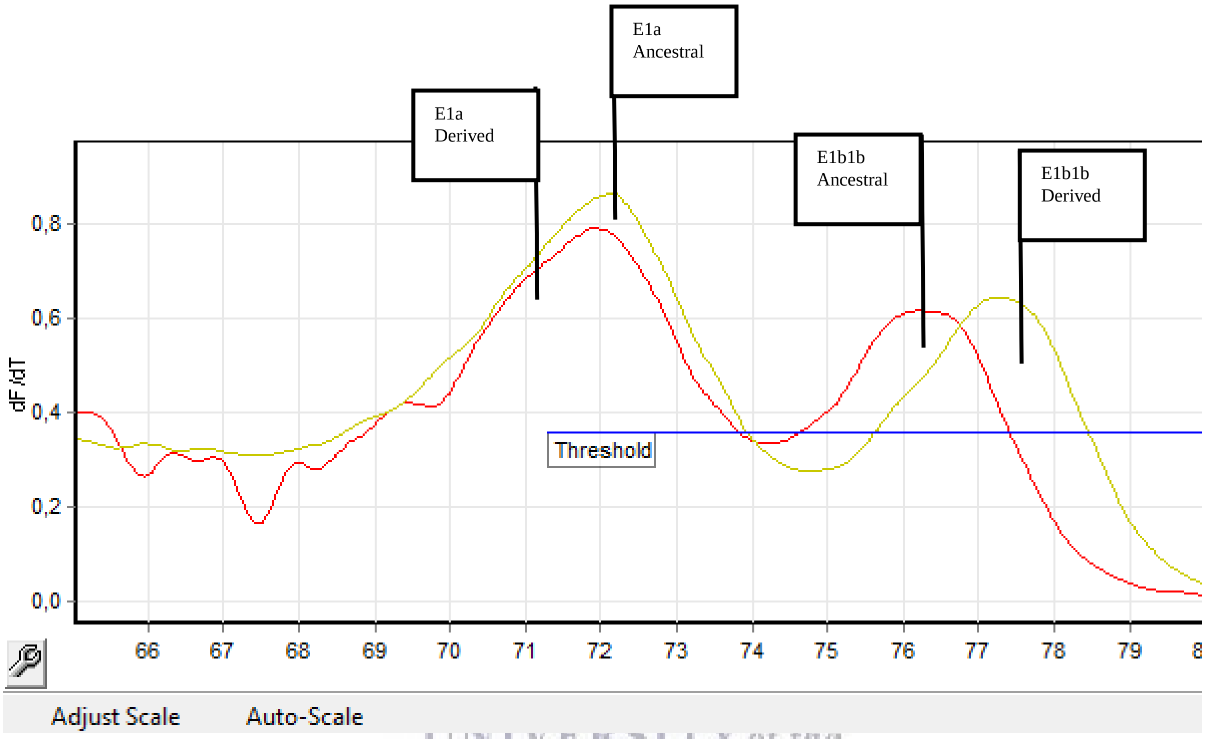


Figure 3.6 Melting curve analysis of Multiplex 5A. Sample belonging to haplogroup E1a (red) and sample belonging to haplogroup E1b1b (yellow) are shown.

The red curve in Figure 3.6 represents haplogroup E1a. In Figure 3.6 the red curve’s melting temperature is lower than that of the yellow curve. The yellow curve in Figure 3.6 is derived for haplogroup E1b1b as its melting temperature is higher than that of the ancestral sample.

It can be seen in Table 3.5 that the theoretical temperatures are different to the experimental ancestral temperatures. The experimental Tm of the amplicon for both ancestral and derived state of SNP M132 showed an increase of temperature in comparison to the theoretical Tm. The experimental Tm of the amplicon for both ancestral and derived state of SNP M215 showed a slight decrease in comparison with the theoretical Tm. This made Multiplex 5A difficult to interpret as SNP M132 had such a small shift in temperature. Multiplex 5A was also a difficult multiplex to work with due to the fact that three different thermocyclers had to be used to get the best results. It was also a difficult multiplex as both primers were sensitive. Out of all the multiplexes, Multiplex 5A was the most difficult to optimise and interpret.

Table 3.5 Theoretical (Blue) and experimental (Purple) Tm of ancestral and derived states of amplicons for Multiplex 5A. Average temperature of each SNP was calculated over 3 different runs and standard deviation was calculated from the average temperature of each run.

Haplogroup (SNP marker: YCC nomenclature)	Ancestral melting temperature (°C)	Derived melting temperature (°C)	Ancestral melting temperature (°C) (average temperature in degree Celsius, **σ**)	Derived melting temperature (°C) (average temperature in degree Celsius, **σ**)
E1a (M132)	66,2	65,76	71,9 (71.5, **0.47**)	71,95 (71.79, **0.22**)
E1b1b (M215)	77,3	78	76,53 (76.46, **0.31**)	77,43 (77, **0.51**)

3.1.6 Multiplex 5B

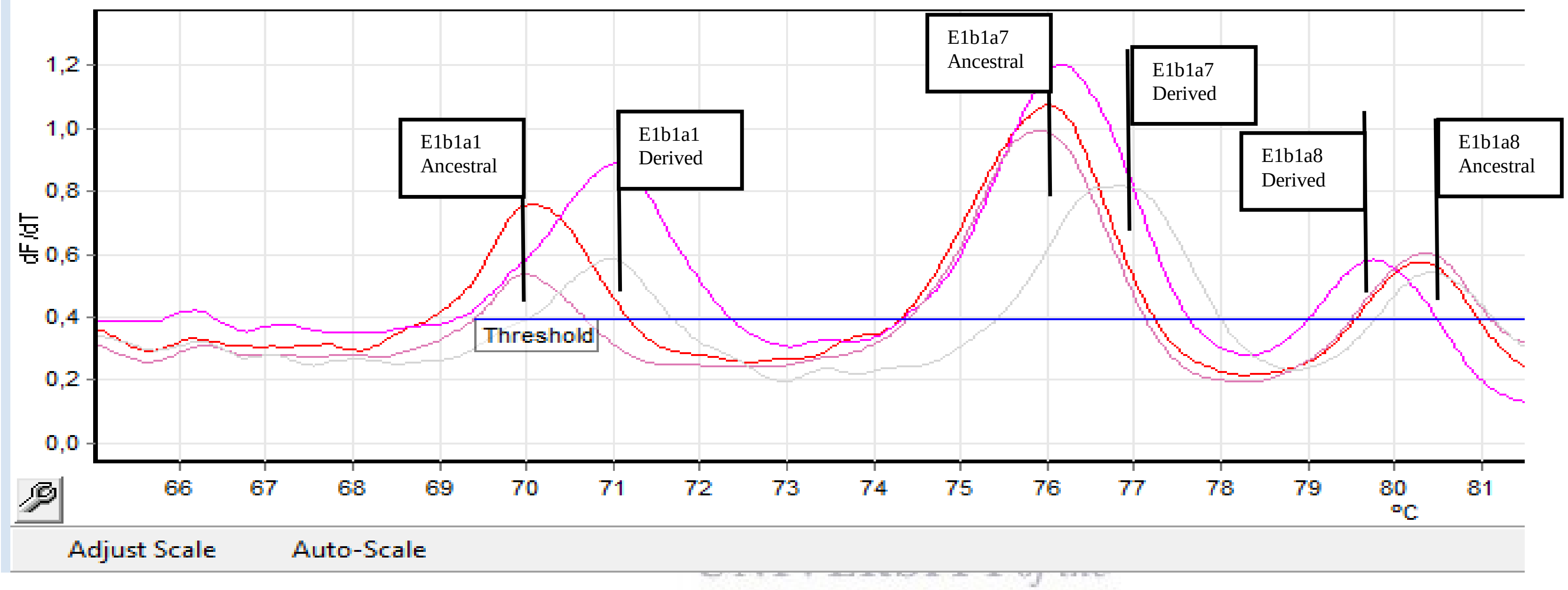


Figure 3.7 Melting curve analysis of Multiplex 5B. Samples derived for haplogroup E1b1a1 (purple and grey), samples ancestral for haplogroup E1b1a1 (red and pink), sample derived for haplogroup E1b1a7 (grey) and sample derived for haplogroup E1b1a8 (purple) are shown.

Multiplex 5B consists of three SNPs that are used if samples are ancestral for all in Multiplex 5A. In Figure 3.7 one can see that the derived samples for haplogroup E1b1a1 have a higher melting temperature than the ancestral samples. This can be seen in Figure 3.7 where the sample that is derived for haplogroup E1b1a7 has a higher melting temperature. In Figure 3.7 the sample that is derived for haplogroup E1b1a8 has a lower melting temperature than the samples that are ancestral for this haplogroup.

It can be seen in Table 3.5 that the theoretical temperatures are different to the experimental ancestral temperatures. The experimental Tm of the amplicon for both ancestral and derived state of SNP U186 showed a decrease of temperature in comparison to the theoretical Tm. The experimental Tm of the amplicon for both ancestral and derived state of SNP M2 showed a decrease of temperature in comparison to the theoretical Tm. The experimental Tm of the amplicon for both ancestral and derived state for SNP U175 had a slight but insignificant decrease in temperature in comparison to the theoretical Tm. The shift in temperature could be due to the DNA samples that were used. If an increased amount of DNA sample was used because it had a low concentration this increased the melting temperatures of the multiplex. This is because the DNA samples could contain excess salt so therefore increasing the amount of DNA added into the reaction would increase the amount of excess salt in the reaction leading to higher temperatures for the multiplex. Although the experimental temperatures of Multiplex 5B were lower than the experimental it did not compromise the ease of determining each haplogroup. None of the temperatures overlapped each other so the different haplogroups were easy to detect. This made Multiplex 5B easy to interpret.

Table 3.6 Theoretical (Blue) and experimental (Purple) Tm of ancestral and derived states of amplicons for Multiplex 5B. Average temperature of each SNP was calculated over 4 different runs and standard deviation was calculated from the average temperature of each run.

Haplogroup (SNP marker: YCC nomenclature)	Ancestral melting temperature (°C)	Derived melting temperature (°C)	Ancestral melting temperature (°C) (average temperature in degree Celsius, **σ**)	Derived melting temperature (°C) (average temperature in degree Celsius, **σ**)
E1b1a7 (U186)	78,8	79,4	76,02 (75.99, **0.19**)	76,85 (76.80, **0.21**)
E1b1a8 (U175)	81,1	80,6	80,36 (80.38, **0.27**)	79,75 (79.72, **0.18**)
E1b1a1 (M2)	74	74,5	70,02 (70.20, **0.20**)	71,01 (70.95, **0.22**)

3.1.7 Multiplex 6

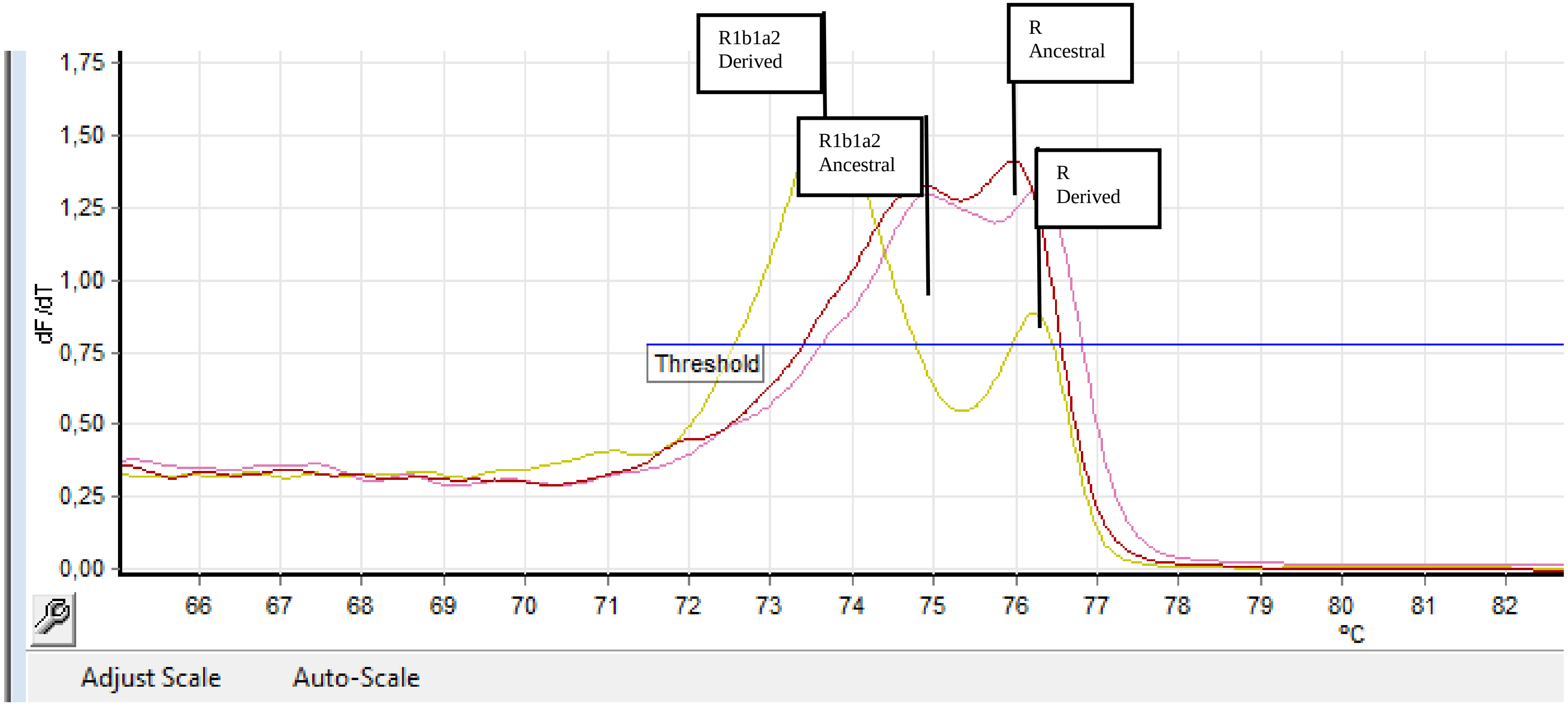


Figure 3.8 Melting curve analysis of Multiplex 6. Sample derived for haplogroup R1b1a2 (yellow), samples derived for haplogroup R (yellow and pink) and sample ancestral for both haplogroups (red) are shown.

Multiplex 6 consists of two SNPs that are used when samples are derived for K(xLT) in Multiplex 1. In Figure 3.8 the yellow and light pink curve are both derived for haplogroup R. The maroon curve is ancestral as its temperature is lower than the two samples that are derived for haplogroup R. In Figure 3.8 the yellow curve is derived for R1b1a2 as its melting temperature is lower than the ancestral samples.

Multiplex 6 had its challenges when it came to the ancestral temperatures for both SNPs as they are close to one another. $MgCl_2$ concentrations were increased and decreased to see if that would help with the separation of the samples. 2.5mM as stated in Chapter 2 was still the best $MgCl_2$ concentration for Multiplex 6. It can be seen in Table 3.7 that the theoretical temperatures are different to the experimental ancestral temperatures. The experimental Tm of the amplicon for both ancestral and derived state of SNP V88 showed a decrease of temperature in comparison to the theoretical Tm. The experimental Tm of the amplicon for both ancestral and derived state of SNP M207 showed a decrease of temperature in comparison to the theoretical Tm.

Table 3.7 Theoretical (Blue) and experimental (Purple) Tm of ancestral and derived states of amplicons for Multiplex 6. Average temperature of each SNP was calculated over 3 different runs and standard deviation was calculated from the average temperature of each run.

Haplogroup (SNP marker: YCC nomenclature)	Ancestral melting temperature (°C)	Derived melting temperature (°C)	Ancestral melting temperature (°C) (average temperature in degree Celsius, **σ**)	Derived melting temperature (°C) (average temperature in degree Celsius, **σ**)
R1b1a2 (V88)	76,6	76,2	74,05 (74.22, **0.21**)	73,25 (73.5, **0.28**)
R (M207)	77,9	78,2	75,35 (75.56, **0.19**)	75,68 (75.77, **0.09**)

3.2 Johannesburg Samples

3.2.1 Quality and Quantity of extracted DNA

A total of 77 male DNA Johannesburg samples were successfully extracted from the 143 sample group. The quantity and quality of the DNA was then checked using a Nanodrop ND-2000 UV spectrophotometer after DNA extractions were completed. Table 3.1 indicates the DNA concentrations in ng/µl and the 360/380 purity ratio. Table 3.8 also indicates which samples did not yield any DNA concentration. A total of 56 samples were successfully genotyped using Multiplex 1 from the 77 male samples that were successfully extracted.

Table 3.8 Indicates the successful Nanodrop results from the Johannesburg samples.

Sample Code	Concentration (ng/µl)	260/380 ratio
S14-001	103,8	1,91
S14-003	142,5	1,81
S14-011	33,3	1,97
S14-013	32,2	1,98
S14-014	36,5	1,98
S14-017	25,5	1,78
S14-018	22,6	1,8
S14-020	143,7	1,93
S14-021	60,8	2,04
S14-022	31,1	1,94
S14-023	33,2	2,07
S14-028	14	1,85
S14-030	55,7	1,93
S14-031	47,4	1,98
S14-032	5,9	1,51
S14-034	43,4	1,98
S14-035	226	1,92

S14-036	13,5	2,12
S14-037	11	2,02
S14-042	18,2	1,89
S14-045	5,1	1,75
S14-047	8,5	1,15
S14-049	54,9	1,92
S14-050	17,1	2,09
S14-051	18,3	1,97
S14-055	16,1	1,93
S14-057	9,1	1,98
S14-058	140,7	1,92
S14-061	53,2	1,8
S14-068	51,4	2
S14-069	72,4	1,93
S14-071	16,7	2,04
S14-072	46,5	1,79
S14-073	198,6	1,92
S14-074	22,6	1,86
S14-075	51	1,82
S14-076	47,7	1,98
S14-077	95,2	1,83
S14-079	62,5	1,72
S14-080	16,3	1,86
S14-081	7,6	1,87
S14-084	29,4	1,92
S14-085	206,7	1,93
S14-086	42,2	1,92
S14-091	78,8	1,96
S14-092	11	1,27
S14-095	27,1	1,73
S14-097	7,4	1,96
S14-099	4,6	1,6
S14-100	3,6	1,86

S14-101	63,4	1,93
S14-103	86,8	1,98
S14-104	31,7	1,88
S14-105	24,2	1,7
S14-106	42,7	1,92
S14-107	19,9	2,1
S14-110	20,2	1,84
S14-111	14	2,07
S14-114	48	1,9
S14-116	10,4	2,06
S14-120	6,3	1,97
S14-124	14,5	2,11
S14-129	13	2,02
S14-135	13,2	1,96
S14-137	22,8	1,76
S14-138	40,9	2,1
S14-140	23,4	2,02
S14-142	19,8	1,89
S14-143	50,6	1,87

.

3.2.2 ScreenClust Analysis

ScreenClust analysis was performed on certain PCR melt analysis. ScreenClust was performed to compare visual analysis of the Johannesburg samples. It was seen that the samples were clustered into the correct haplogroups and coincided with the visual analysis of each sample. There are samples that fell out of the clusters due to their temperatures but were actually visually placed into a haplogroup. Further explanation of each ScreenClust analysis for each multiplex system will be done in this part of Chapter 3.

3.2.2.1 Multiplex 1

The ScreenClust analysis confirmed the haplogroups that the Johannesburg samples were allocated to after being genotyped with Multiplex 1. The samples that are circled in pink fell outside of the DE haplogroup cluster but were identified as being derived for haplogroup DE. This can be seen in Figure 3.9 below.

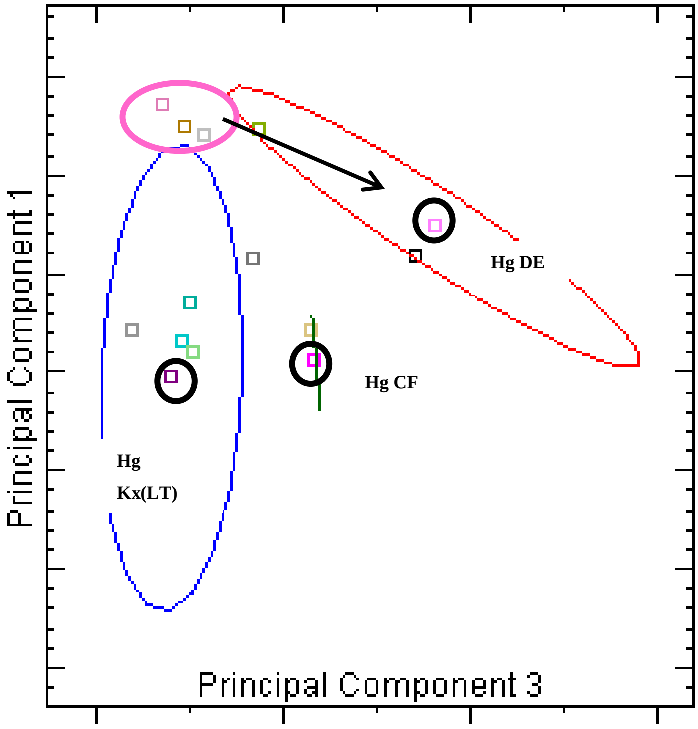


Figure 3.9 Supervised ScreenClust analyses of Johannesburg samples genotyped with Multiplex 1. Red cluster represents haplogroup DE, blue cluster represents haplogroup K(xLT) and the green line represents haplogroup CF. All samples circled in black are the positive controls used to identify each

haplogroup in Multiplex 1. The pink cluster represents the samples that were not clustered correctly. The black arrow indicates which haplogroup the samples belong to.

3.2.2.2 Multiplex 2

As only one control was used in Multiplex 2 it was difficult for the ScreenClust to identify which haplogroups each sample belonged to. It was also difficult as not many samples were derived for haplogroup AOT only. It was easier to do a visual analysis of the samples then move onto Multiplex 3 if samples were derived for haplogroup B. The two samples that were derived for haplogroup AOT only in Multiplex 1 were both derived for haplogroup B in Multiplex 2. This can be seen in the melting curve analysis below (Figure 3.10). The melting difference can also be seen in the normalised HRM curve graph (Figure 3.11). The purple circle indicates that there is a temperature change between the derived samples for B and the ancestral sample (Figure 3.11). This temperature change can be seen in Figure 3.10 as well.

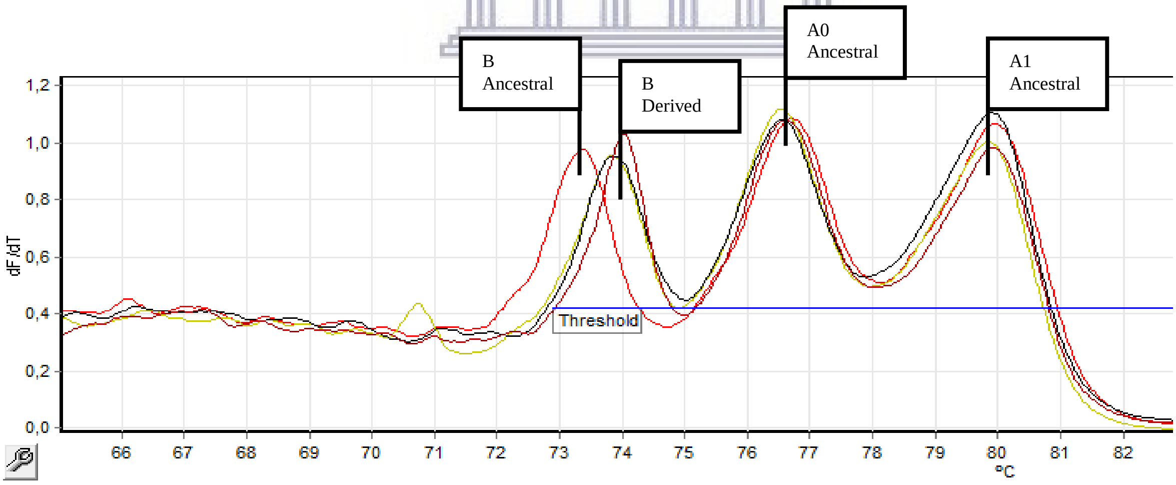


Figure 3.10 Indications of samples being derived for haplogroup B in Multiplex 2.

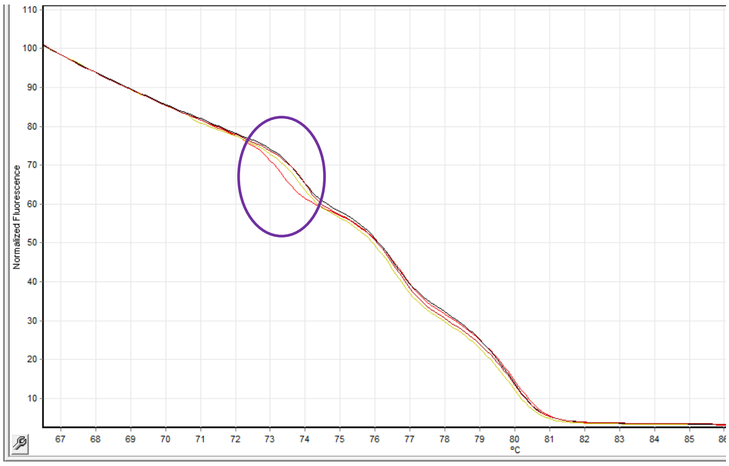


Figure 3.11 Normalised HRM curve of samples derived for haplogroup B in Multiplex 2.

3.2.2.3 Multiplex 3

It can be seen in Figure 3.12 that the Johannesburg samples were derived for haplogroup B2a. The ScreenClust analysis coincides with the visual analysis of the samples. As there were no other samples for haplogroup B2b the sample circled in the red is also the positive control used to represent the haplogroup. The ancestral for all SNPs in Multiplex 3 sample is the positive control used for haplogroup R. It was used to make sure that ancestral and derived temperatures could be seen in Multiplex 3.

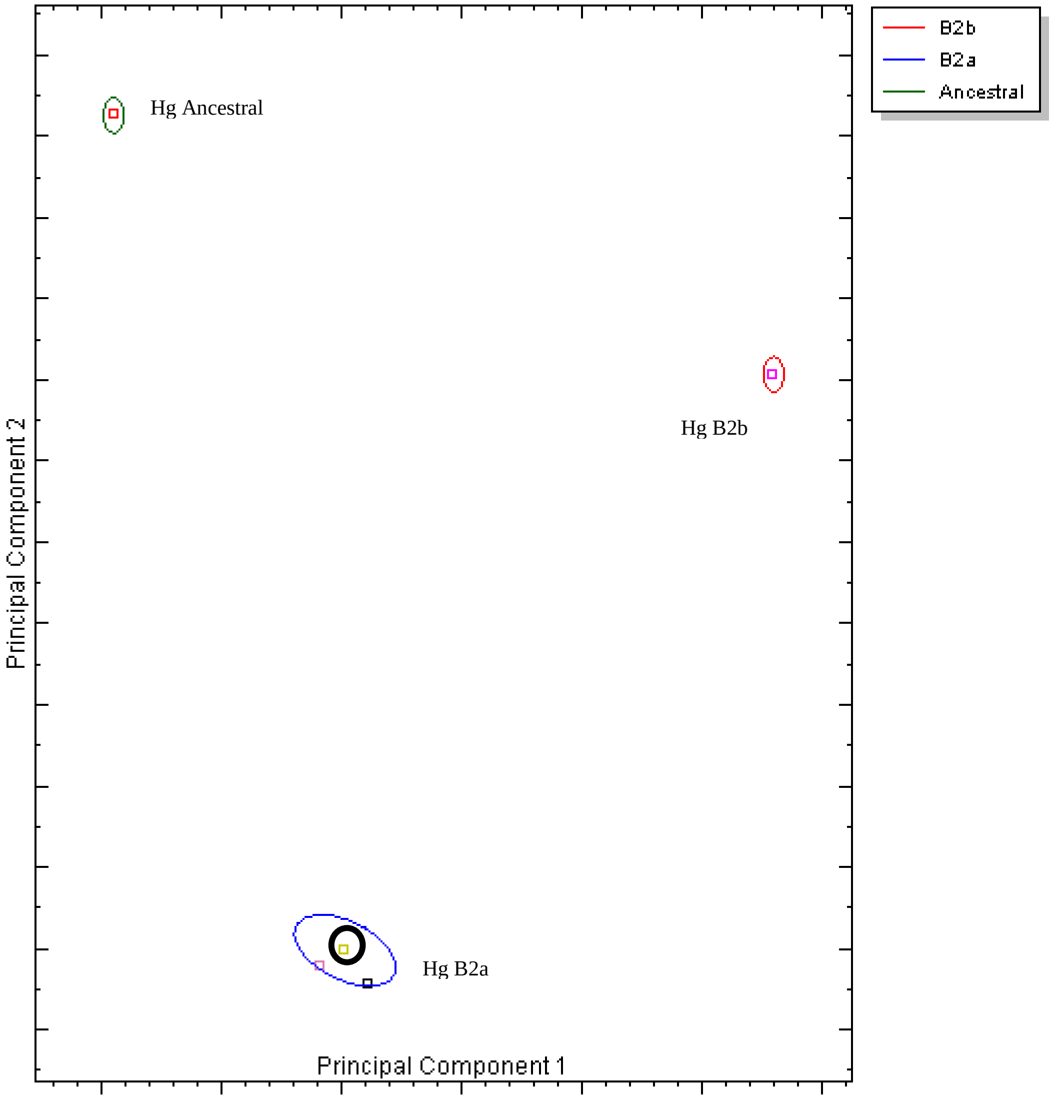


Figure 3.12 Supervised ScreenClust analyses of Johannesburg samples genotyped with Multiplex 3. Red cluster represents haplogroup B2b. Blue cluster represents haplogroup B2a and green cluster represents ancestral for all sample. Sample circled in black represents the positive control used for the haplogroup.

3.2.2.4 Multiplex 4

The ScreenClust analysis of the Johannesburg genotyped with Multiplex 4 (Figure 3.13) coincides with the visual analysis of the samples. There are, however, two outliers in this screen analysis. When this sample set was analysed visually the two outlier samples were determined to be part of haplogroup E1. They could be out of the cluster due to their temperatures.

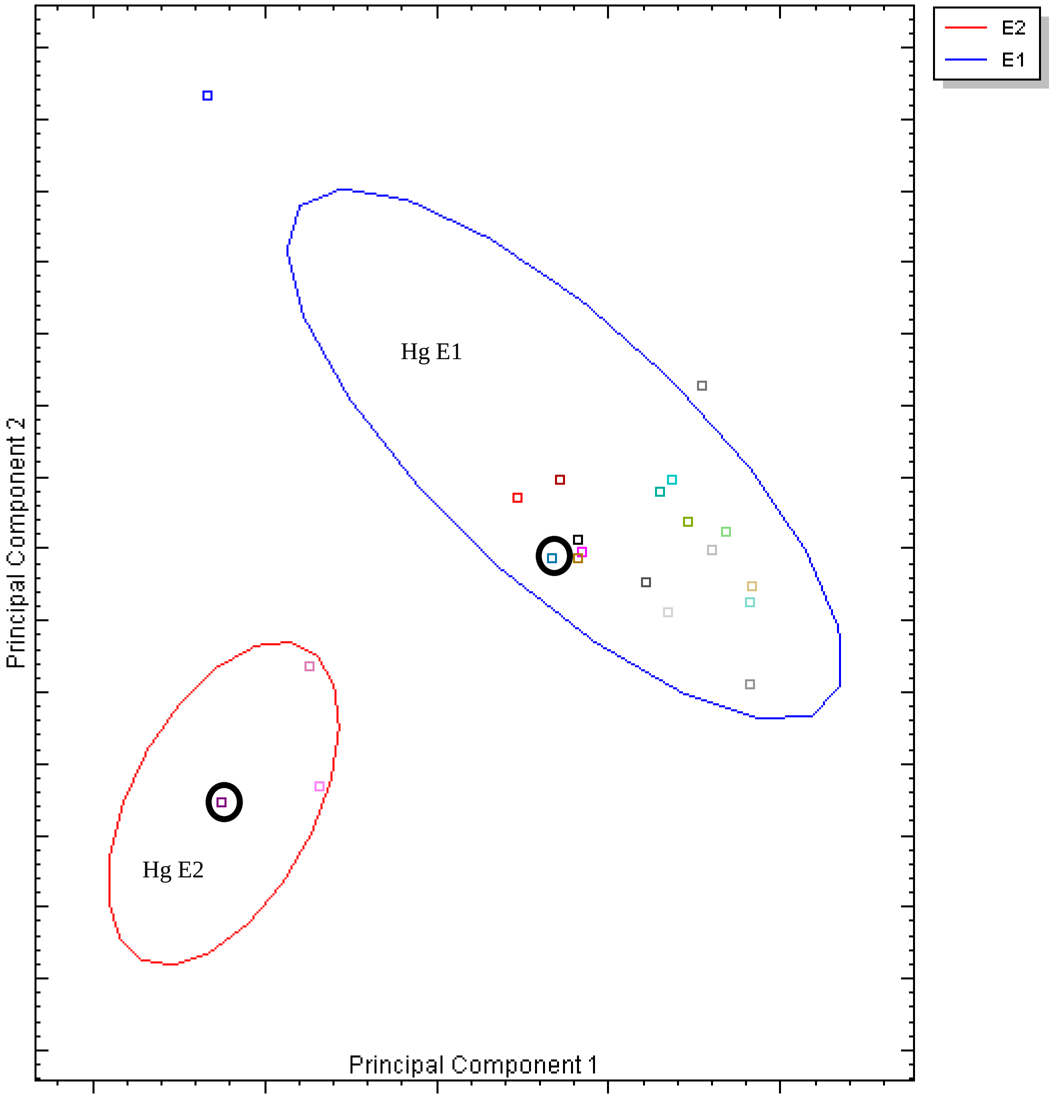


Figure 3.13 Supervised ScreenClust analyses of Johannesburg samples genotyped with Multiplex 4. Red cluster represents haplogroup E2 and blue cluster represents haplogroup E1. All samples circled in black are the positive controls used to identify each haplogroup in Multiplex 4.

3.2.2.5 Multiplex 5A

Multiplex 5A is the first multiplex to decide which E1 haplogroup samples belong to if they were derived for haplogroup E1 in Multiplex 4. The samples that are not clustered represent the samples that were ancestral for both haplogroups. These samples were then run with Multiplex 5B. Figure 3.15 corresponds with the visual analysis of the Johannesburg samples for this multiplex.

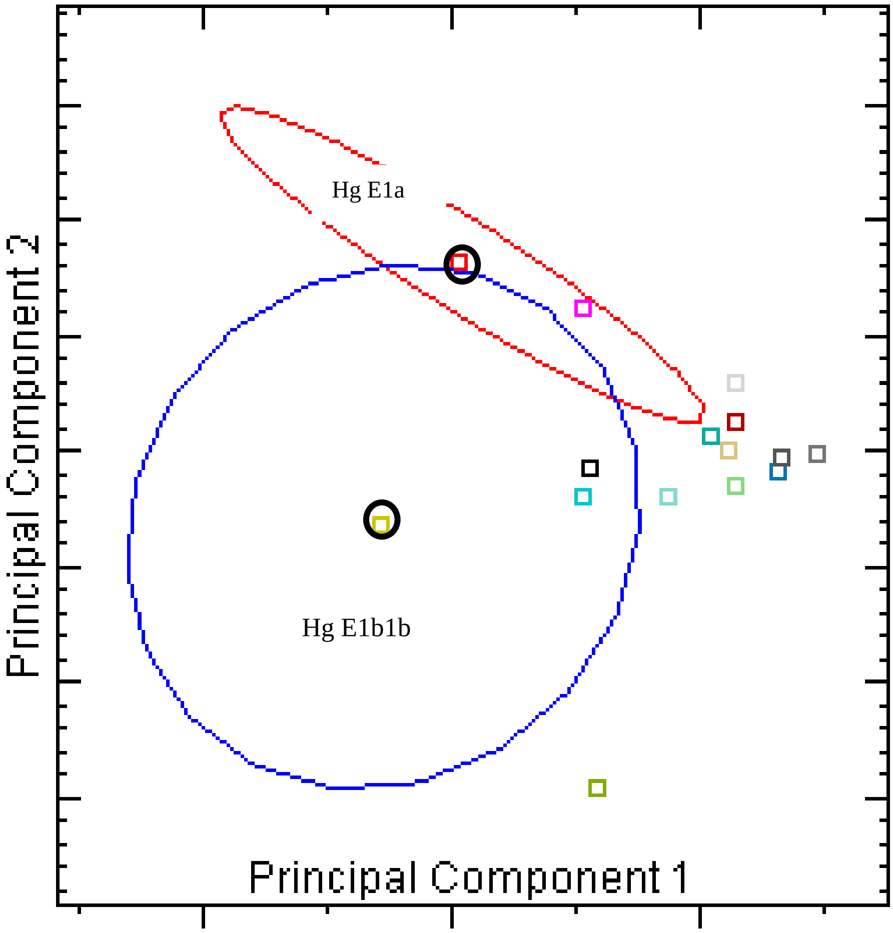


Figure 3.14 Supervised ScreenClust analyses of Johannesburg samples genotyped with Multiplex 5A. Red cluster represents haplogroup E1a and blue cluster represents haplogroup E1b1b. The samples outside of the two clusters are ancestral for both haplogroups. All samples circled in black are the positive controls used to identify each haplogroup in Multiplex 5A.

3.2.2.6 Multiplex 5B

The ScreenClust analysis of the Johannesburg genotyped with Multiplex 5B (Figure 3.15) coincides with the visual analysis of the samples. Ancestral samples were used to see that all derived and ancestral melting temperatures could be seen for Multiplex 5B.

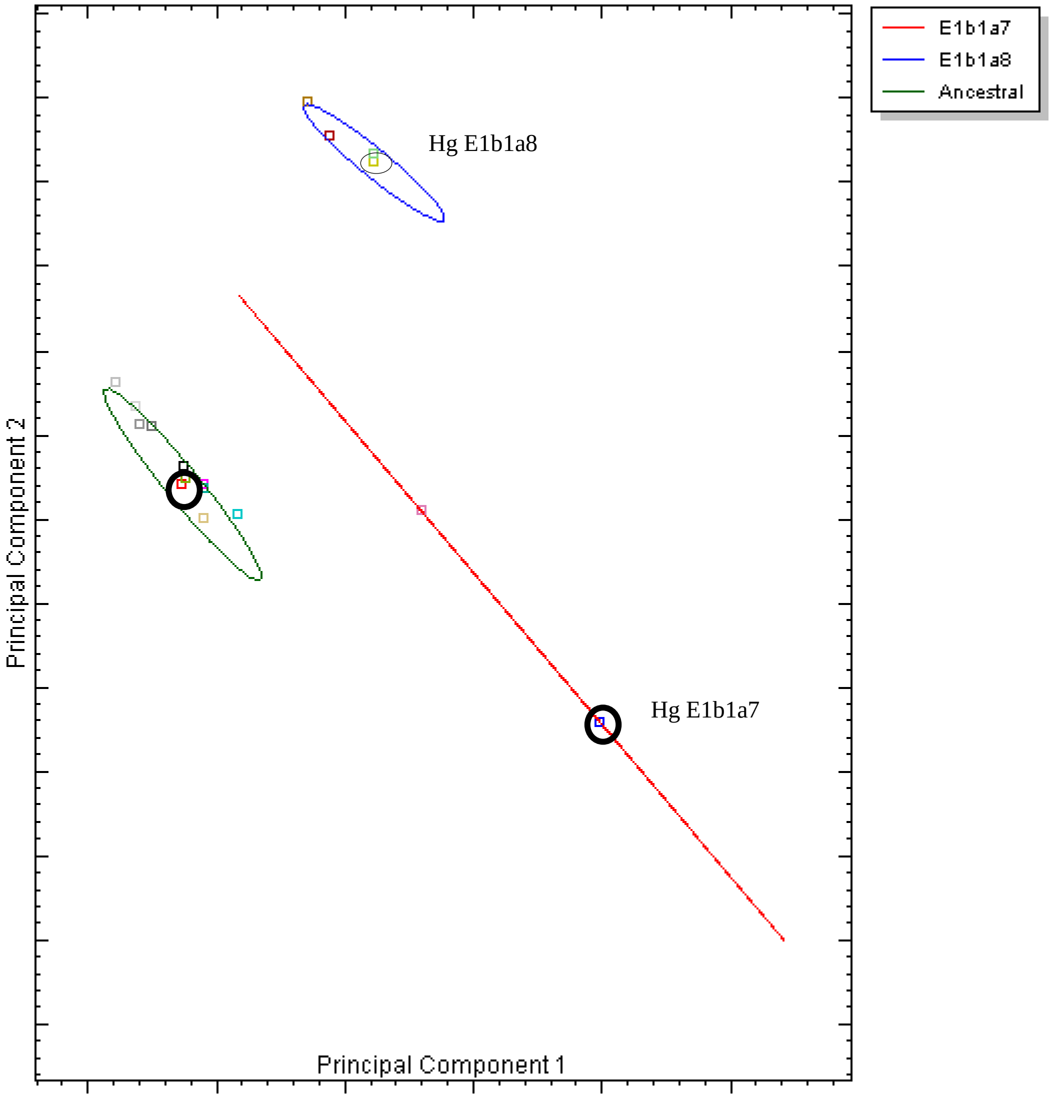


Figure 3.15 Supervised ScreenClust analyses of Johannesburg samples genotyped with Multiplex 5B. Red cluster represents haplogroup E1b1a7 and the blue cluster represents haplogroup E1b1a8. The green cluster represents all the ancestral samples. All samples circled in black are the positive controls used to identify each haplogroup in Multiplex 5B.

3.2.2.7 Multiplex 6

The ScreenClust analysis of the Johannesburg genotyped with Multiplex 6 (Figure 3.16) coincides with the visual analysis of the samples. As there were no other samples for haplogroup R1b1a2 the sample circled in the blue is also the positive control used to represent the haplogroup. The ancestral for all SNPs in Multiplex 6 sample is a sample derived for haplogroup E1. It was expected for the samples to be derived for haplogroup R as haplogroup R is from European descent. European lineages were mixed with African ones when South Africa was first colonised by the Dutch.

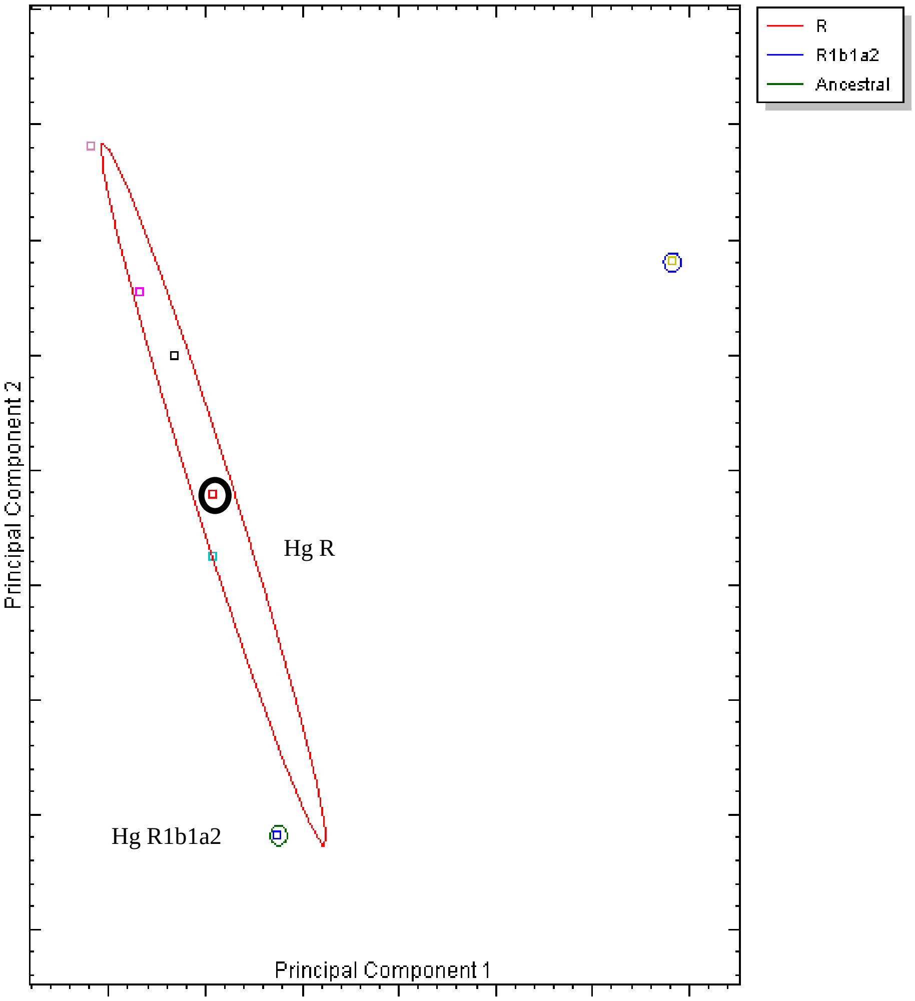


Figure 3.16 Supervised ScreenClust analyses of Johannesburg samples genotyped with Multiplex 6. Red cluster represents haplogroup R. Blue cluster represents haplogroup R1b1a2 and green cluster represents ancestral sample. All samples circled in black are the positive controls used to identify each haplogroup in Multiplex 6.

3.2.3 Genotyping Results

A total of 56 samples were successfully genotyped using Multiplex 1. The rest of the samples did not genotype correctly due to their DNA concentration. Some samples failed completely where others had an inconclusive profile using Multiplex 1.

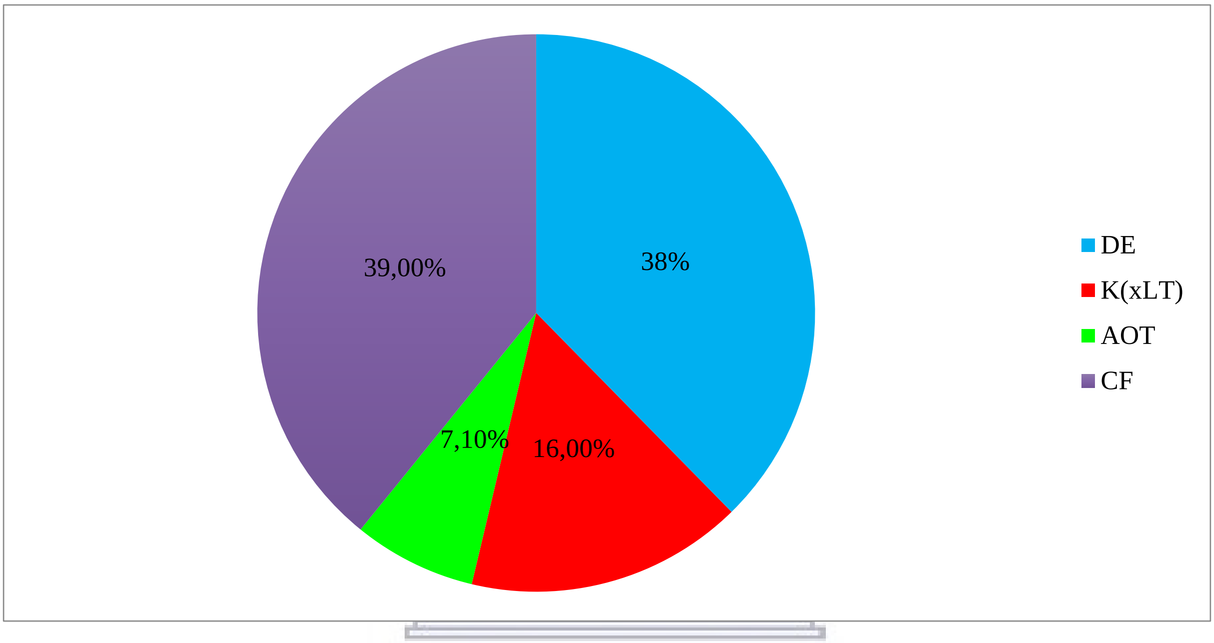


Figure 3.17 Pie chart indicates the haplogroup distribution of the Johannesburg samples when genotyping with Multiplex 1. Light blue represents DE haplogroup, purple represents samples only derived for haplogroup CF, green represents samples only derived for haplogroup AOT and red represents haplogroup K (xLT).

It was expected that haplogroup DE would be the dominant haplogroup from Multiplex 1 however one can see in Figure 3.17 that haplogroup CF was the dominate haplogroup from Multiplex 1. It can be seen that 38% of the samples were derived for haplogroup DE (Figure 3.17).

Haplogroup CF had 39% of the samples belonging to haplogroup CF (Figure 3.17). Haplogroup CF is a macro haplogroup and includes lineages mainly outside of Africa. This is the furthest this haplogroup is defined in this multiplex system. The Johannesburg samples derived for haplogroup CF suggest that they could be from Asian descent.

Haplogroup K(xLT) made up 16% of the samples that were genotyped with Multiplex 1 (Figure 3.17). Samples derived for haplogroup K(xLT) were then run with Multiplex 6 and were found to be derived for haplogroup R. Haplogroup R is derived from European, West, Central and South Asian descent.

Haplogroup AOT only made up 7.10% of the samples that were genotyped with Multiplex 1 (Figure 3.17). Samples that were only derived for haplogroup AOT were run with Multiplex 2; it was seen that the samples were derived for haplogroup B. This led to the samples being run with Multiplex 3 and it was found that the samples were derived for haplogroup B2a1a. This haplogroup is found mainly in African population groups.

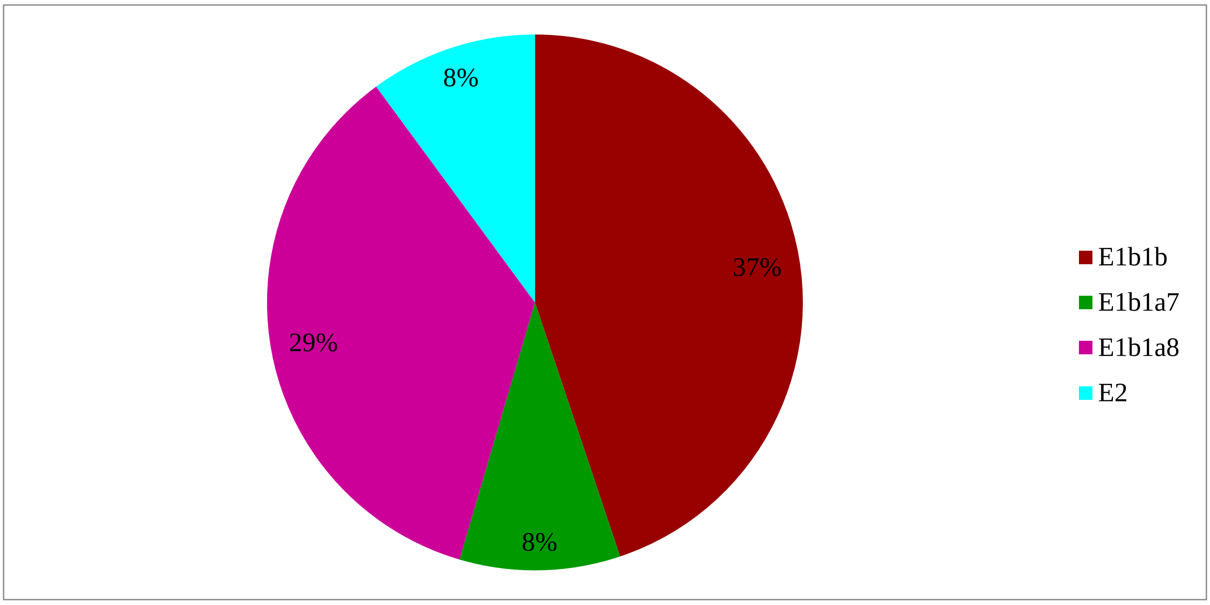


Figure 3.18 Pie chart indicating the haplogroup distribution of the Johannesburg samples when genotyping with Multiplex 4, 5A and 5B. Maroon represents E1b1b haplogroup, green represents E1b1a7 haplogroup, light purple represents E1b1a8 haplogroup and light blue represents E2 haplogroup.

A total of 20 samples were successfully genotyped using Multiplex 4, Multiplex 5A and Multiplex 5B.

A total of 37% of the samples that were derived for haplogroup DE from Multiplex 1 were derived for haplogroup E1b1b from Multiplex 5A (Figure 3.18). Haplogroup E1b1b has a wide geographical distribution which ranges from Northern and Eastern Africa to Europe and Western Asia (Cruciani *et al.*, 2004).

Figure 3.19 represents the distribution of haplogroup E1b1b. Although it is mainly found in the horn of Africa the distribution of this haplogroup has moved to the Southern parts of Africa. This could be the reason that 37% of the Johannesburg samples are derived for this haplogroup.

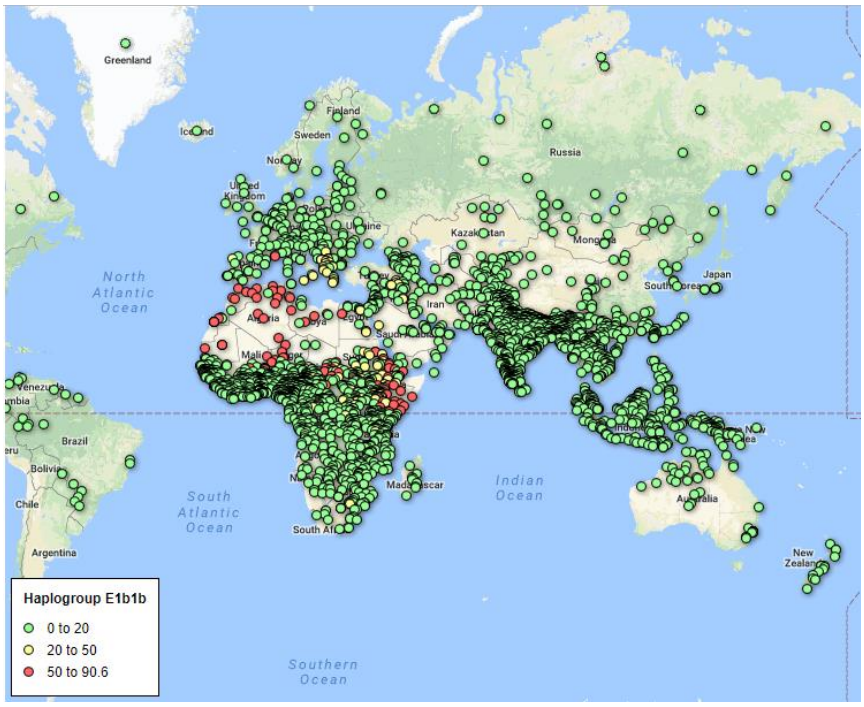


Figure 3.19 Haplomap indicating the distribution of haplogroup E1b1b. Taken from https://haplomaps.com/haplogroup-e1b1b/

A total of 29% of the samples derived for haplogroup DE from Multiplex 1 were derived for haplogroup E1b1a8 from Multiplex 5B (Figure 3.18). A total of 8% of the samples that were derived for haplogroup DE from Multiplex 1 were derived for haplogroup E1b1a7 from Multiplex 5B (figure 3.18). Haplogroups E1b1a8 and E1b1a7 are sub-haplogroups of haplogroup E1b1a1. Haplogroup E1b1a1 expanded from central Africa to west and southern Africa. It is thought that the distribution of E1b1a1 was due to the Bantu-expansion that took place (De Filippo *et al.*, 2013). In Figure 3.20 it illustrates that haplogroups E1b1a8 and E1b1a7 can be found in South Africa.

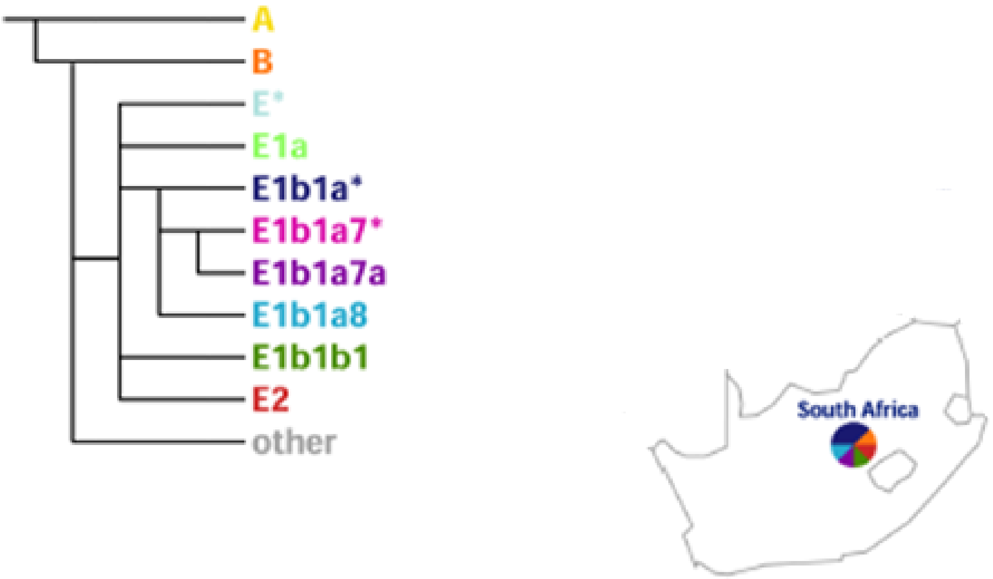


Figure 3.20 Haplogroup composition of the combined data set used in (De Filippo *et al.*, 2013). Image has been adapted to indicate only the haplogroups found in South Africa and not the whole of Africa as the original image did in the paper. The illustration indicates that haplogroups E1b1a7 (SNP U174) and E1b1a8 (SNP U175) are found in South Africa.

A total of 8% of samples genotyped with Multiplex 4 were derived for haplogroup E2 (Figure 3.18). Haplogroup E2 is found mainly in central Africa. Figure 3.21 indicates the distribution of haplogroup E2; one can see the red dot on the map is near central Africa.

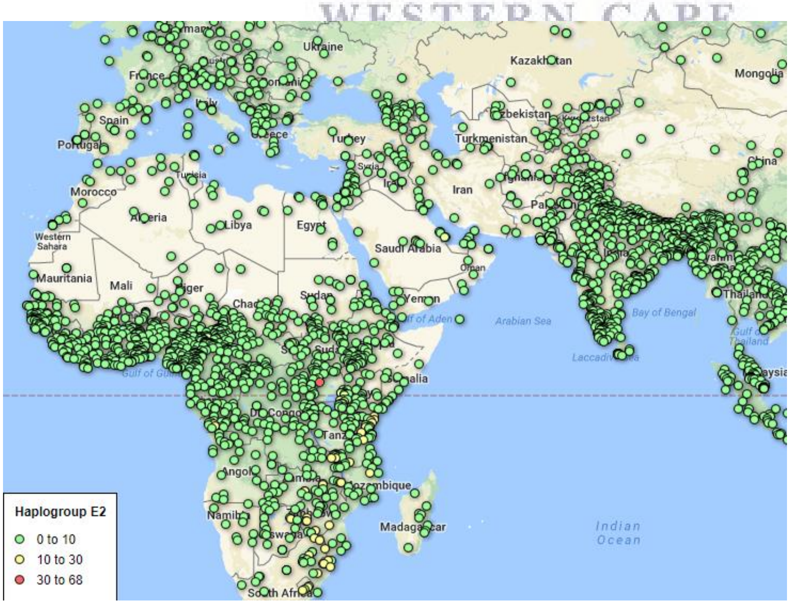


Figure 3.21 Haplogroup indicating the distribution of haplogroup E2. Taken from https://haplomaps.com/haplogroup-e2/

3.2.4 Chi-Square Test Analysis

The results of the Chi-square test analysis can be observed in Table 3.9 below. The Western Cape Coloured population ancestry groups were compared to the Johannesburg Coloured population ancestry groups.

Table 3.9 Comparison of the relative contributions of African, European and Asian ancestry between the Western Cape Coloured population and the Johannesburg Coloured population. Western Cape Coloured Population data taken from Quintana-Murci *et al.* (2010).

	African ancestry	European ancestry	Asian ancestry	Total
Actual Values				
Western Cape coloured	103	86	39	228
Johannesburg coloured	25	9	22	56
Total	128	95	61	284
	African ancestry	European ancestry	Asian ancestry	Total
Expected Values				
Western Cape coloured	102.7605634	76.26760563	48.97183099	228
Johannesburg coloured	25.23943662	18.73239437	12.02816901	56
Total	128	95	61	284
P value				**0.0002486700602**

Chapter 4: Conclusion and future work

4.1 Conclusion

It can be seen that the multiplex system that was designed and implemented in this thesis does work. The multiplex system can infer an individual's paternal ancestral lineage.

At first glance at the results one would think that the Johannesburg Coloured population and the Western Cape Coloured population that was tested by Quintana-Murci *et al.* (2010) had similar paternal results. But with further analysis of the Johannesburg Coloured population data one can deduce that the Johannesburg Coloured population ancestry is different to the Western Cape Coloured population ancestry.

In Quintana-Murci *et al.* (2010) the most frequent paternal lineage originated from Sub-Saharan African haplogroups at 45.2%. The Johannesburg Coloured population's second most frequent paternal lineage were from Sub-Saharan African haplogroups at 38%.

The next most frequent paternal lineages from the Western Cape Coloured population were West Eurasian haplogroups (37.7%) and South/ Southeast Asian haplogroups (17.1%). With regards to the Johannesburg samples, the most frequent paternal lineage was Asian haplogroups at 39%. It should be noted that the samples derived for haplogroup CF (39%) have to be further analysed as haplogroup CF is a base haplogroup and many other haplogroups fall under it.

The European haplogroups were the smallest paternal lineage group for the Johannesburg Coloured population only 16% of the Johannesburg samples were derived for this haplogroup. These results indicate that the Johannesburg Coloured population has a greater Asian ancestral history than a European one compared to the Western Cape Coloured population group if you compare the results found in Quintana-Murci *et al.* (2010) and this study.

A reason the Johannesburg Coloured population group could have an increased Asian ancestral history could be due to the slave trade that occurred during 1658-1806 (De Wit *et al.*, 2010). This population group could have moved from the coast inland with their masters. This could also be an indication of why the CF haplogroup of the Johannesburg Coloured population being 39%. One must take into consideration that this study only had a small sample size compared to Quintana-Murci *et al.* (2010), this haplogroup's percentages could potentially change if more male Johannesburg Coloureds are genotyped.

If we investigate further into the Sub-Saharan haplogroups of the Johannesburg Coloured population 37% of the samples were derived for haplogroup E1b1b (M215), 8% for haplogroup E2 (M75), 8% for haplogroup E1b1a7 (U186) and 29% for haplogroup E1b1a8 (U175).

In the Quintana-Murci *et al.* (2010) study 24.9% of the Western Cape Coloured samples were derived for haplogroup E-M2. These days E-M2 is classified as haplogroup E1b1a1. Haplogroups E1b1a7 (8%) and E1b1a8 (29%) are both sub groups of haplogroup E1b1a1. This indicates that these Johannesburg samples have Sub-Saharan African lineages; they could possibly be descendants from the Bantu population.

In contrast to the Western Cape study conducted by Quintana-Murci *et al.* (2010) 37% of the Johannesburg Coloured population was derived for haplogroup E1b1b, indicating that the Johannesburg sample's paternal ancestry is more Northern and Eastern Africa as that is where haplogroup E1b1b mainly originates. Table 4.1 below compares the paternal ancestral lineages of the two sample sets in descending order.

Table 4.1 Comparison of the sub E haplogroups of each sample set. This is indicated in descending order of the most frequent haplogroup of each sample set. Western Cape Coloured Population data taken from Quintana-Murci *et al.* (2010).

Western Cape Coloured Population **(Haplogroup)** **(103 males genotyped)**	**Johannesburg Coloured Population** **(Haplogroup)** **(20 males genotyped)**
24.12% E-M2 (E1b1a1)	37% M215 (E1b1b)
5.70% E-P68 (E2)	29% U175 (E1b1a8)
3.5% E-M5 (E1b1b1)	8% M75 (E2)
0.44% E-M123 (E1b1b1c)	8% U186 (E1b1a7)

A Chi-square test was performed to compare the relative contributions of African, European and Asian ancestry of both population groups result was significant (p= 0.00025).

The result of this study gives one more insight into the Coloured population group. This study shows that Coloureds from a different region within South Africa have their own unique paternal ancestry.

4.2 Future Work

Positive controls have to be sequenced with SNPs that were used in this project to validate this system. The lab was not able to send samples for sequencing due to lack of funding.

The multiplex system can be tested on a wider range of samples. This can be both in sample size and in different population groups. This multiplex system can also be compared to Y-STRs to see that the haplotype and the haplogroup for a given sample are similar.

A downfall of this system would be that one is required to have DNA samples that are high in concentration but should especially have a clean 260/280 purity ratio. Samples that did not work with the system or that were difficult to genotype were the samples that had bad DNA concentrations and bad 260/280 purity ratios.

It would also be an interesting study to compare the Johannesburg Coloured population's paternal lineages to the maternal lineages to see what genetic difference there is. It would also be interesting to see if the maternal lineages are similar to the ones found in the Western Cape Coloured population researched by Quintana-Murci *et al.* (2010). It would be an excellent way to investigate the population migration within South Africa.

References

Andréasson, H., *et al,*. (2002). ‘Mitochondrial sequence analysis for forensic identification using pyrosequencing technology.’ *Biotechniques, 32*, 124–133.

Ahmadian, A. *et al.* (2000) ‘Single-nucleotide polymorphism analysis by pyrosequencing’, *Analytical Biochemistry*, 280(1), pp. 103–110. doi: 10.1006/abio.2000.4493.

Armstrong, B., Stewart, M. and Mazumder, A. (2000) ‘Suspension arrays for high throughput, multiplexed single nucleotide polymorphism genotyping’, *Cytometry*, 40(2), pp. 102–108. doi: 10.1002/(SICI)1097-0320(20000601)40:2<102::AID-CYTO3>3.0.CO;2-4.

Batini, C. *et al.* (2011) ‘Early Y chromosome lineages and the peopling of Africa’, *Molecular Biology*.

Bell, P. A. *et al.* (2002) ‘SNPstream?? UHT: Ultra-high throughput SNP genotyping for pharmacogenomics and drug discovery’, *BioTechniques*, 32(6 SUPPL.).

Butler, J. M. (2012a) ‘Single Nucleotide Polymorphisms and Applications’, in *Advanced Topics in Forensic DNA Typing*, pp. 347–369. doi: 10.1016/B978-0-12-374513-2.00012-9.

Butler, J. M. (2012b) ‘Y-Chromosome DNA Testing’, in *Advanced Topics in Forensic DNA Typing*, pp. 371–403. doi: 10.1016/B978-0-12-374513-2.00013-0.

Chen, X. N., Levine, L. and Kwok, P. Y. (1999) ‘Fluorescence polarization in homogeneous nucleic acid analysis’, *Genome Research*, 9(5), pp. 492–498. doi: 10.1101/gr.156601.

Consortium, T. Y. C. (2002) ‘A nomenclature system for the tree of human Y-Chromosomal binary haplogroups’, *Genome Research*, 12(2), pp. 339–348. doi: 10.1101/gr.217602.

Cruciani, F. *et al.* (2002) ‘A Back Migration from Asia to Sub-Saharan Africa Is Supported by High-Resolution Analysis of Human Y-Chromosome Haplotypes’, *The American Journal of Human Genetics*, 70(5), pp. 1197–1214. doi: 10.1086/340257.

Cruciani, F. *et al.* (2004) ‘Phylogeographic Analysis of Haplogroup E3b (E-M215) Y Chromosomes Reveals Multiple Migratory Events Within and Out Of Africa’, *The American Journal of Human Genetics*, 74(5), pp. 1014–1022. doi: 10.1086/386294.

Cruciani, F. *et al.* (2006) ‘Molecular dissection of the Y chromosome haplogroup E-M78

(E3b1a): a posteriori evaluation of a microsatellite-network-based approach through six new biallelic markers.', *Human mutation*, 27(8), pp. 831–832. doi: 10.1002/humu.9445.

Cruciani, F. *et al.* (2007) 'Tracing past human male movements in northern/eastern Africa and western Eurasia: New clues from Y-chromosomal haplogroups E-M78 and J-M12', *Molecular Biology and Evolution*, 24(6), pp. 1300–1311. doi: 10.1093/molbev/msm049.

Cruciani, F. *et al.* (2010) 'Human Y chromosome haplogroup R-V88: a paternal genetic record of early mid Holocene trans-Saharan connections and the spread of Chadic languages.', *European journal of human genetics : EJHG*. Nature Publishing Group, 18(7), pp. 800–807. doi: 10.1038/ejhg.2009.231.

Delahunty, C. *et al.* (1996) 'Testing the feasibility of DNA typing for human identification by PCR and an oligonucleotide ligation assay.', *American journal of human genetics*, 58(6), pp. 1239–46. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1915044&tool=pmcentrez&rendertype=abstract.

De Filippo, C. *et al.* (2013) 'Europe PMC Funders Group Y-chromosomal variation in Sub-Saharan Africa : insights into the history of Niger-Congo groups', 2011(3), pp. 1255–1269. doi: 10.1093/molbev/msq312.Y-chromosomal.

Francalacci, P. *et al.* (2013) 'Low-pass DNA sequencing of 1200 Sardinians reconstructs European Y-chromosome phylogeny', *Science*, 341(6145), pp. 565–569. doi: 10.1126/science.1237947.

Gabriel, M. N. *et al.* (2001) 'Population variation of human mitochondrial DNA hypervariable regions I and II in 105 Croatian individuals demonstrated by immobilized sequence-specific oligonucleotide probe analysis.', *Croatian medical journal*, 42(3), pp. 328–335.

Germer, S. *et al.* (1999) 'Single-Tube Genotyping without Oligonucleotide Probes Single-Tube Genotyping without Oligonucleotide Probes', pp. 72–78. doi: 10.1101/gr.9.1.72.

Giesendorf, B. A. J. *et al.* (1998) 'Molecular beacons: A new approach for semiautomated mutation analysis', *Clinical Chemistry*, 44(3), pp. 482–486.

Gilles, P. N., *et al,*. (1999). ‘Single nucleotide polymorphic discrimination by an electronic dot bot assay on semiconductor microchips.’ *Nature Biotechnology*, *17*, 365–370.

Haff, L. A. and Smirnov, I. P. (1997) ‘Single-nucleotide polymorphism identification assays using a thermostable DNA polymerase and delayed extraction MALDI-TOF mass spectrometry’, *Genome Research*, 7(4), pp. 378–388. doi: 10.1101/gr.7.4.378.

Hartigan, J. A. and Wong, M. A. (1979) ‘Algorithm AS 136: A K-Means Clustering Algorithm’, *Applied Statistics*, 28(1), p. 100. doi: 10.2307/2346830.

Hecker, K. H., Taylor, P. D. and Gjerde, D. T. (1999) ‘Mutation detection by denaturing DNA chromatography using fluorescently labeled polymerase chain reaction products.’, *Analytical biochemistry*, 272(2), pp. 156–64. doi: 10.1006/abio.1999.4171.

Herrmann, M. G. *et al.* (2006) ‘Amplicon DNA melting analysis for mutation scanning and genotyping: Cross-platform comparison of instruments and dyes’, *Clinical Chemistry*, 52(3), pp. 494–503. doi: 10.1373/clinchem.2005.063438.

Javadi, A. and Shamaei, M. (2014) ‘Qualification Study of Two Genomic DNA Extraction Methods in Different Clinical Samples’, 13(4), pp. 41–47.

Jobling, M. A. and Tyler-Smith, C. (2003) ‘The human Y chromosome: An evolutionary marker comes of age’, *Nature Reviews Genetics*, 4(8), pp. 598–612. doi: 10.1038/nrg1124.

Jolliffe, I. T. (2002) *Principal Component Analysis, Second Edition*, *Encyclopedia of Statistics in Behavioral Science*. doi: 10.2307/1270093.

KapaBiosystems (2016) ‘Technical Data Sheet KAPA Hyper Prep Kit’, (January), pp. 1–16. Available at: https://www.sigmaaldrich.com/content/dam/sigma-aldrich/docs/Roche/Datasheet/1/hrmfastkbdat.pdf.

Karafet, T. M. *et al.* (2008) ‘New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree’, *Genome Research*, 18(5), pp. 830–838. doi: 10.1101/gr.7172008.

Kayser, M. *et al.* (2004) ‘A comprehensive survey of human Y-chromosomal microsatellites’, *American Journal of Human Genetics*, 74(6), pp. 1183–1197. doi: 10.1086/421531.

Kayser, M. (2017) ‘Forensic use of Y-chromosome DNA: a general overview’, *Human Genetics*. Springer Berlin Heidelberg, 136(5), pp. 621–635. doi: 10.1007/s00439-017-1776-9.

Kirby, K. S. (1956) ‘A new method for the isolation of ribonucleic acids from mammalian tissues.’, *The Biochemical journal*, 64(3), pp. 405–8. doi: 10.1042/bj0640405.

de Knijff, P. (2000) ‘Messages through bottlenecks: on the combined use of slow and fast evolving polymorphic markers on the human Y chromosome.’, *American journal of human genetics*, 67(5), pp. 1055–1061. doi: 10.1086/321215.

Kong, Q. P. *et al.* (2006) ‘Updating the East Asian mtDNA phylogeny: A prerequisite for the identification of pathogenic mutations’, *Human Molecular Genetics*, 15(13), pp. 2076–2086. doi: 10.1093/hmg/ddl130.

Kwok, P.-Y. (2001). ‘Methods for genotyping single nucleotide polymorphisms.’ *Annual Reviews of Genomics and Human Genetics, 2*, 235–258.

Lando, D. Y. *et al.* (2015) ‘Determination of melting temperature and temperature melting range for DNA with multi-peak differential melting curves’, *Analytical Biochemistry*, 479, pp. 28–36. doi: 10.1016/j.ab.2015.01.018.

Li, J. *et al.* (1999) ‘Single nucleotide polymorphism determination using primer extension and time-of-flight mass spectrometry’, *Electrophoresis*, 20(6), pp. 1258–1265. doi: 10.1002/(SICI)1522-2683(19990101)20:6<1258::AID-ELPS1258>3.0.CO;2-V.

Li, M. *et al.* (2014) ‘Genotyping accuracy of high-resolution DNA melting instruments’, *Clinical Chemistry*, 60(6), pp. 864–872. doi: 10.1373/clinchem.2013.220160.

Livak, K. J. (1999) ‘Allelic discrimination using fluorogenic probes and the 5’ nuclease assay’, *Genetic Analysis - Biomolecular Engineering*, 14(5–6), pp. 143–149. doi: 10.1016/S1050-3862(98)00019-9.

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

Mendez, F. L. *et al.* (2013) ‘An African American paternal lineage adds an extremely ancient root to the human y chromosome phylogenetic tree’, *American Journal of Human Genetics*. The American Society of Human Genetics, 92(3), pp. 454–459. doi: 10.1016/j.ajhg.2013.02.002.

Miller, S. a., Dykes, D. D. and Polesky, H. F. (1988) 'A simple salting out procedure for extracting DNA from human nucleated cells', *Nucleic Acids Research*, 16(3), p. 1215. doi: 10.1093/nar/16.3.1215.

Nikiforov, T. T., *et al*. (1994). 'Genetic Bit Analysis: A solid phase method for typing single nucleotide polymorphisms.' *Nucleic Acids Research*, *22*, 4167–4175.

Oswald, N. (2008). 'The Basics: How Phenol Extraction Works: Bitesize Bio.' http://bitesizebio.com/384/the-basics-how-phenol-extraction-works/

Van Oven, M. *et al.* (2014) 'Seeing the wood for the trees: A minimal reference phylogeny for the human Y chromosome', *Human Mutation*, 35(2), pp. 187–191. doi: 10.1002/humu.22468.

Petersen, D. C. *et al.* (2013) 'Complex Patterns of Genomic Admixture within Southern Africa', *PLoS Genetics*, 9(3), pp. 10–13. doi: 10.1371/journal.pgen.1003309.

PhyloTree Y- Minimal Y tree.2017. *PhyloTree Y- Minimal Y tree.* [ONLINE] Available at: http://www.phylotree.org.Y/tree/index.htm. [Acessed 07 December 2017]

Protocol, Q. (no date) *Danagene saliva kit*. Available at: http://bioted.es/protocolos/DANAGENE-SALIVA-KIT(EDUC)-ENG.pdf.

QIAGEN (2009) 'For use with Rotor-Gene cyclers Sample & Assay Technologies QIAGEN Sample and Assay Technologies', (September). Available at: https://www.qiagen.com/us/resources/download.aspx?id=af33be05-14c6-4ac3...%0A.

Quintana-Murci, L. *et al.* (2010) 'Strong Maternal Khoisan Contribution to the South African Coloured Population: A Case of Gender-Biased Admixture', *American Journal of Human Genetics*, 86(4), pp. 611–620. doi: 10.1016/j.ajhg.2010.02.014.

Reja, V. *et al.* (2010) 'ScreenClust: Advanced statistical software for supervised and unsupervised high resolution melting (HRM) analysis', *Methods*. Elsevier Inc., 50(4), pp. S10–S14. doi: 10.1016/j.ymeth.2010.02.006.

Reynolds, R., *et al*. (2000). 'Detection of sequence variation in the HVII region of the human mitochondrial genome in 689 individuals using immobilized sequence-specific oligonucleotide probes', *Journal of Forensic Sciences*, *45*, 1210–1231.

Ririe, Kirk M., Rasmussen, Randy P., Witter, C. T. (1997) 'Product Differentiation by

Analysis of DNA Melting Curves during the Polymerase Chain Reaction', *Analytical Biochemistry*, 245(245), pp. 154–160. doi: 10.1006/ABIO.1996.9916.

Saiki, R. K. *et al.* (1989) 'Genetic analysis of amplified DNA with immobilized sequence-specific oligonucleotide probes.', *Proceedings of the National Academy of Sciences of the United States of America*, 86(August), pp. 6230–6234. doi: 10.1073/pnas.86.16.6230.

Salas, A. *et al.* (2002) 'The making of the African mtDNA landscape.', *American Journal of Human Genetics, The*, 71(5), pp. 1082–111. doi: 10.1086/344348.

Sanger, F., Coulson, A.R., 1975. 'A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase.' J. Mol. Biol. 94, 441–448

Sapolsky, R. J. *et al.* (1999) 'High-throughput polymorphism screening and genotyping with high-density oligonucleotide arrays', *Genetic Analysis - Biomolecular Engineering*, 14(5–6), pp. 187–192. doi: 10.1016/S1050-3862(98)00026-6.

Selelstad, M. T. *et al.* (1994) 'Construction of human Y-chromosomal haplotypes using a new polymorphic A to G transition', *Human Molecular Genetics*, 3(12), pp. 2159–2161. doi: 10.1093/hmg/3.12.2159.

Sigma Aldrich (1989) 'Phenol:chloroform: isoamyl alcohol', 3, p. 3803.

Sims, L. M., Garvey, D. and Ballantyne, J. (2007) 'Sub-populations within the major European and African derived haplogroups R1b3 and E3a are differentiated by previously phylogenetically undefined Y-SNPs.', *Human mutation*, 28(1), p. 97. doi: 10.1002/humu.9469.

Skaletsky, H. *et al.* (2003) 'The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes', *Nature*, 423(6942), pp. 825–837. doi: 10.1038/nature01722.

Sosnowski, R. G. *et al.* (1997) 'Rapid determination of single base mismatch mutations in DNA hybrids by direct electric field control.', *Proceedings of the National Academy of Sciences of the United States of America*, 94(4), pp. 1119–23. doi: 10.1073/pnas.94.4.1119.

Statistics South Africa (2012) *Census 2011 - Census in brief*, *World Wide Web*. doi: ISBN 978-0-621-41388-5.

Tishkoff, S. A. *et al.* (2007) ‘History of click-speaking populations of Africa inferred from mtDNA and Y chromosome genetic variation’, *Molecular Biology and Evolution*, 24(10), pp. 2180–2195. doi: 10.1093/molbev/msm155.

Tully, G. *et al.* (1996) ‘Rapid detection of mitochondrial sequence polymorphisms using multiplex solid-phase fluorescent minisequencing’, *Genomics*, 34(1), pp. 107–113. doi: 10.1006/geno.1996.0247.

Underhill, P. A. *et al.* (2001) ‘The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations.’, *Annals of human genetics*, 65, pp. 43–62. doi: 10.1046/j.1469-1809.2001.6510043.x.

Wang, D. G., *et al.* (1998). ‘Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome.’ *Science*, *280,* 1077–1082.

Wang, W. *et al.* (2010) ‘Correction for Chiaroni et al., Y chromosome diversity, human expansion, drift, and cultural evolution’, *Proceedings of the National Academy of Sciences*, 107(30), pp. 13556–13556. doi: 10.1073/pnas.1008738107.

Wiesner, G. L. and Slavin, T. P. (2009) ‘Colorectal Cancer’, *Colorectal Cancer*, pp. 879–897.

De Wit, E. *et al.* (2010) ‘Genome-wide analysis of the structure of the South African Coloured Population in the Western Cape’, *Human Genetics*, 128(2), pp. 145–153. doi: 10.1007/s00439-010-0836-1.

Wittwer, C. T. *et al.* (2003) ‘High-resolution genotyping by amplicon melting analysis using LCGreen’, *Clinical Chemistry*, 49(6), pp. 853–860. doi: 10.1373/49.6.853.

Wittwer, C. T. (2009) ‘High-resolution DNA melting analysis: Advancements and limitations’, *Human Mutation*, 30(6), pp. 857–859. doi: 10.1002/humu.20951.

Wojciech, R. (2010) ‘Oligo v7: Primer analysis software’, p. 209. Available at: http://www.oligo.net/.

Zuccarelli, G. *et al.* (2011) ‘Rapid screening for Native American mitochondrial and Y-chromosome haplogroups detection in routine DNA analysis’, *Forensic Science International: Genetics*. Elsevier Ireland Ltd, 5(2), pp. 105–108. doi: 10.1016/j.fsigen.2010.08.018.