GENETIC ANALYSIS OF 27 Y-STR LOCI IN DIFFERENT POPULATION GROUPS FROM SOUTH AFRICA FOR FORENSIC PURPOSES

By

Kyla Bianca Dooley

Submitted in fulfilment of the requirements in respect of the Master's Degree Magister Scientiae (Forensic Genetics) in the Department of Genetics in the Faculty of Natural and Agricultural Sciences at the University of the Free State.

Supervisor: Dr Karen Ehlers (Ph.D)

Co-Supervisor: M. Thabang Madisha (M.Tech)

October 2020

Declaration

I, *Kyla Bianca Dooley*, declare that the Master's Degree research dissertation or interrelated, publishable manuscripts/published articles, or coursework Master's Degree mini-dissertation that I herewith submit for the Master's Degree qualification, *MSc (Forensic Genetics),* at the University of the Free State is my independent work, and that I have not previously submitted it for a qualification at another institution of higher education.

Date: 12/10/2020

Table of Contents

Acknowledgements

I would like to acknowledge and sincerely thank the following people and organisations for their guidance and support during this degree.

My first thank you certainly has to go to you, Dr Ehlers. Thank you for being the most incredible supervisor. I have really enjoyed working with you over the past four years, and I have learnt so much from you, not only in the world of forensics, but also in life. Thank you for always listening to me and picking me back up when things did not go according to plan – and we both know how often that happened. I would not have been able to complete this degree without you, and I am so grateful to have had you as my supervisor.

To my co-supervisor and the best laboratory technician around, Thabang Madisha, thank you. I am so grateful to have had your guidance and assistance throughout this journey. From assisting me when things went wrong in the lab, teaching me everything there is to know about everything, letting me spend my days sharing your office, always telling me about all your research, and the endless conversations about anything and everything. The list goes on. You are so appreciated and I am really going to miss doing your admin – please keep your desk clean, for me.

To Sonja Strümpher, it has been an absolute blessing having your guidance during this process. I cannot thank you enough for everything that you have done to assist me throughout my research. Thank you for giving up your time to help me in the lab – you are the reason for getting things back on track – and for analysing results with me. I have thoroughly enjoyed learning from you, you have taught me so much. You, and everything you have done for me, are greatly appreciated.

A huge thank you to ThermoFisher Scientific for funding this research. I am honoured to have had the opportunity to conduct this kind of research using your products, and it would not have been possible without your financial support. A special thank you goes to Jenna Van Den Munckhof for facilitating the ordering of the kits and any consumables required, for making sure that we had what we needed, and for always dealing with my frequent phone calls and enquires with a smile. Another special mention goes to Polo Mokomo, who was so helpful throughout this whole process, fetching samples from Bloemfontein, and spending time with me in the laboratory to assist when things were not going according to plan. Your assistance and guidance is greatly appreciated Mama Polo.

The University of the Free State, thank you for the opportunity to study further at an institution of such high calibre and the financial support in the form of tuition fee bursaries. The Department of Genetics, and Prof Grobler in particular, thank you for allowing me, as an 'outsider' from a different university, to begin my career in Forensic Genetics at your institution. A very special thank you to Mrs Wessels, for your constant support and pick me ups. Thank you for letting me cry in your office when things went wrong, for always trying to help where you could, and for all the chats in between.

Thank you to all the men, whether they are students or staff members, who so graciously donated their DNA samples. This research would literally not have been possible without your input, so it is greatly appreciated.

To all my friends that I've made along the way, and to those that I've always had. Thanks for making my time in Bloemfontein a memorable one, and for all the support throughout this process. Thanks for always taking an interest in all my stories about my research and happenings in the laboratory – especially when you had no idea what I was talking about.

To my parents, Lyndsay and Paul Coelho, words cannot express my gratitude for your unwavering support throughout this entire journey. Whether I was on a high from fantastic results, or at my lowest when nothing worked, you were both by my side the entire time. You had to endure this degree as much as I did, and I thank you both so much for everything. To my brother, Justin Coelho, thank you. You have gone through this whole experience with me and have been a constant pillar of support. I could not have done it without you.

To my guardian angel – this one is for you, Grumps.

Abstract

South Africa has one of the highest rape statistics in the world, with an average of 117 rapes reported daily. Y-STR genotyping is becoming a popular tool in the analysis of DNA evidence collected after a crime of a sexual nature has been committed. Although there are some exceptions, most rape cases involve a male perpetrator and a female victim. By targeting male-specific (perpetrator) DNA, any female (victim) DNA is excluded from analysis, resulting in clear Y-STR profiles that could be used to obtain a match with a male suspect. A reduced genetic diversity at some core Y-STR loci, a limited number of markers investigated, and lack of haplotype frequency data present a challenge in the implementation of Y-STRs in South Africa's forensic laboratories. This dissertation represents a study aimed at investigating the forensic value of commercial Y-STR PCR Amplification kits in the South African population, as well as provide haplotype frequency data.

A total of 308 samples were collected from the African, Asian/Indian, Coloured, and Caucasian populations at the University of the Free State. These samples were amplified using ThermoFisher Scientific's Y-Filer® Plus PCR Amplification kit, and analysed using the GeneMapper™ ID-X Software. Statistical analysis was performed to estimate several forensic parameters to evaluate the performance of this kit. This set of markers was able to identify 261 unique haplotypes, with an overall discrimination capacity of 98.15%. Discrimination capacities ranged from 91.67% for the Asian/Indian population to 100% for the Coloured population. The haplotype diversity across the four populations is 0.9999, with an average gene diversity across all loci of 0.684. The Coloured population exhibited the highest gene and haplotype diversities, as well as the highest discrimination capacity, most likely due to the high levels of admixture in this population. These values are comparable to those of other populations around the world and are increased from those reported in previous South African studies.

This study also used the Y-Filer® Plus kit to genetically characterise the South African population through the use of allelic patterns, genetic distances, AMOVAs, and P-CoAs using the GenAlEx software. The Coloured population exhibited the highest number of different alleles, contributing to the high gene diversity in this population, while, interestingly, the African population was shown to have the highest number of private alleles. It was shown that the most genetic differences occurred between the African and Caucasian populations, while the Coloured population showed a closer affinity to the Caucasian population. The African and Caucasian populations formed two distinct clusters during P-CoA, with the Coloured population distributed across both clusters, albeit closer to the Caucasian cluster. These

results are consistent with the history of the South African population. Several profile variations were detected and analysed, including null alleles at *DYS390* and *DYS448,* duplications at *DYS458* and *DYS449,* triplications at *DYF387S1,* and several intermediate alleles at *DYS389II, DYS458,* and *DYF387S1.* Although these variations could introduce some challenges when used in forensic DNA analysis, additional knowledge gained through sequencing regarding the nature of these variations may circumvent these challenges.

The forensic parameters estimated in this study provide evidence for the potential use of commercial Y-STR PCR kits in a forensic application in South Africa. Even though this study does provide some haplotype frequency data, it is highly recommended that more haplotypic data is obtained for a more accurate estimation of match probabilities. Future studies should also focus on the performance of the Y-Filer® Plus kit when analysing DNA from closely related males in the South African population.

Keywords

Forensic Genetics

Y-STR markers

Y-Filer® Plus PCR Amplification Kit

South African Population

Population Data

Discrimination Capacity

Haplotype Diversity

Gene Diversity

Sexual Assault Investigation

List of Abbreviations

AMOVA	-	Analysis of Molecular Variance
AZF	-	Azoospermia Factor
CE	-	Capillary Electrophoresis
CODIS	-	Combined DNA Index System
D	-	Genetic distance
DC	-	Discrimination Capacity
DF	-	Degrees of Freedom
DNA	-	Deoxyribonucleic Acid
F_{ST}	-	Genetic distance unit
GD	-	Gene Diversity
Hd	-	Haplotype Diversity
Indel	-	Insertion/Deletion
MB	-	Mega (million) Base pairs
MH	-	Minimal Haplotype
MP	-	Match Probability
MPS	-	Massively Parallel Sequencing
MS	-	Mean Square
MSY	-	Male-Specific region of the Y-chromosome
NIST	-	National Institute of Standards and Technology
NGS	-	Next Generation Sequencing
NRY	-	Non-recombining region of the Y-chromosome
OL	-	Off Ladder
OMR	-	Off Marker Range
P-CoA	-	Principal Coordinates Analysis
PA	-	Private Alleles
PARs	-	Pseudo-Autosomal Regions

PCR	-	Polymerase Chain Reaction
RFLP	-	Restriction Fragment Length Polymorphism
RFU	-	Relative Fluorescence Units
RM	-	Rapidly Mutating
R_{ST}	-	Genetic distance unit
SAPS	-	South African Police Service
SM	-	Slowly Mutating
SNP	-	Single Nucleotide Polymorphism
STR	-	Short Tandem Repeat
STRBase	-	Short Tandem Repeat DNA Internet DataBase
SWGDAM	-	Scientific Working Group on DNA Analysis Methods
UH	-	Unique Haplotypes
UK	-	United Kingdom
USA	-	United States of America
YAP	-	Y-chromosome Alu Polymorphism
Yp	-	Short arm of the Y-chromosome
Yq	-	Long arm of the Y-chromosome
Y-STR	-	Y-chromosome STR
Y-SNP	-	Y-chromosome SNP
YHRD	-	Y-chromosome Haplotype Reference Database

List of Figures

List of Tables

CHAPTER 1: INTRODUCTION

1.1 Introduction

With a total number of 1 919 495 violent crimes reported to the South African Police Service (SAPS) in 2019/2020, it could indeed be said that crime is a huge challenge in South Africa (South African Police Service, 2020).[1] During the period from April 2019 to March 2020, the total number of sexual offences reported was 53 293, a slight increase from the 2018/2019 statistic of 52 420. The category of sexual offences includes rape, compelled rape, sexual assault, incest, bestiality, statutory rape, and the sexual grooming of children. In South Africa, rape is described as an act in which there is oral, anal, or vaginal penetration of an individual with a genital organ, or any object, without consent or agreement between the people involved (Africa Check, 2018). Of the 53 293 sexual offences reported in 2019/2020, there were 42 289 (79.3%) reports of rape (South African Police Service, 2020). This number is also slightly higher than the 41 583 cases reported in 2018/2019, with an average of 117 rapes reported each day. Despite these statistics already being alarmingly high, it is thought that they are higher in reality, as many sexual assaults go unreported (Rape Crisis, 2018).

When crimes of a sexual nature are committed, DNA evidence is collected and analysed in an attempt to identify and convict the perpetrator of the crime. The specific DNA regions that are analysed are known as short tandem repeats (STRs). STRs are short DNA segments that are repeated a certain number of times, with the number of repeats determined by the alleles at each locus (Moxon and Wills, 1999). Many different allelic combinations are possible, with the number of repeat units varying significantly between individuals. This variability makes it possible to identify and differentiate between individuals, which is the main objective of DNA evidence in forensic investigations (Jobling, 2001).

Autosomal STRs are found on chromosomes 1 to 22 while Y-STRs are located on the Y-chromosome, meaning that they are only found in individuals who are genetically male (XY genotype). In most sexual assault cases, the DNA evidence collected consists of both female (usually the victim) and male (usually the perpetrator) biological material. Consequently, the DNA profile generated using autosomal STRs would most likely be a mixture of both individuals' DNA. Often, the female DNA is higher in quantity than the male DNA, which could lead to the male DNA being undetectable in the profile. In cases such as this, the autosomal

[1] The researcher is aware that data collected is timeous and, therefore, may become outdated subsequent to the time of research. This is particularly true of data retrieved online, which is subject to real-time updates. Kindly refer to the reference list for specific dates of access relating to data retrieved via online sources, to better contextualise the time of validity for such sources.

STR profiles may fail to provide the necessary information to match a suspect's DNA to the evidentiary DNA. The alternative use of Y-STRs is especially beneficial when this happens.

Targeting the male-specific markers is advantageous as it excludes any female DNA from analysis, thereby eliminating the risk of it concealing the male DNA (Roewer, 2009). Generating male-specific DNA profiles also allows for easier comparison between the male suspect and the DNA evidence as the female DNA is no longer a factor that needs to be considered. Y-STR analysis does not require the presence of sperm cells, so it is possible to obtain DNA profiles for males who are azoospermic or oligospermic (Shewale *et al.,* 2004). In addition to this, Y-STR analysis produces DNA profiles that are less complex than mixture profiles as there would only be one allele per locus, with the exception of a few loci that can potentially be heterozygous. Despite its apparent benefits, Y-STRs are still under-utilised in forensic investigations.

One of the biggest concerns regarding the use of Y-STRs in forensic analysis is the fact that the Y-chromosome is passed down intact from father to son. Any differences between generations would only come about from mutational events; however, Y-STR mutation rates are low compared to autosomal markers, so significant differences between relatives are rare (Roewer, 2009). The consequence of this is that, when a match between a suspect and the evidence is obtained, any male relatives of the suspect cannot automatically be excluded.

An additional challenge encountered with the use of Y-STRs is the lack of available comprehensive population reference data. Once a match between a suspect and the evidence is obtained, statistical confidence should be included in the result to reinforce the significance of the match (Chakraborty and Kidd, 1991). This statistical confidence specifies the probability that the match occurred by chance, as well as the probability of any other individuals in the population having the same DNA profile. The calculation of these probabilities depends on having information regarding the haplotype frequencies of the reference population available.

STRs have been used in forensic investigations for longer than Y-STRs, with Y-STRs only being evaluated for potential use in forensic investigations by Kayser *et al.* in 1997. Autosomal STR databases are, therefore, currently more comprehensive than Y-STR databases. This lack of data is not an issue that is unique to South Africa. However, there is an initiative in place to rectify this and provide ample information regarding Y-STRs for use in forensic investigations across the world.

The Y-chromosome Haplotype Reference Database (YHRD) is a global Y-STR database that was established in 2001, and it has been growing ever since. There are currently 1 080 656 haplotypes on the database that contain various loci from several population groups (Roewer and Willuweit, 2020). However, the majority of these haplotypes have been contributed by individuals in more developed countries, with European and Asian samples dominating the database. To date, Sub-Saharan African haplotypes make up only 1% of the whole YHRD. As of October 2020, there have been 89 574 haplotypes consisting of the 27 Y-STR loci of ThermoFisher Scientific's Y-Filer® Plus PCR Amplification Kit uploaded to this database (www.yhrd.org). Despite this large number, there are no Y-Filer® Plus haplotypes on this database for the South African population. This indicates a gap in the knowledge of the 27 Y-Filer® Plus Y-STR loci in South Africa, a gap that this study aims to fill.

1.2 Aims and Objectives

The aim of this study is to investigate the forensic value of, as well as the characterisation of, 27 Y-STR loci in four different population groups from South Africa. This aim will be achieved by completing the following objectives:

Objective 1: Collect DNA buccal swabs from male individuals from the Asian/Indian, African, Coloured, and Caucasian population groups.

Objective 2: Generate DNA profiles of the 27 Y-STR loci included in ThermoFisher Scientific's Y-Filer® Plus PCR Amplification Kit.

Objective 3: Use forensic parameters to assess the viability of these 27 Y-STR markers for use in forensic investigations in South Africa.

Objective 4: Use the GenAlEx software to calculate the genetic diversity within and among the four population groups, as well as the allelic patterns and Y-STR profile variations observed within each group.

1.3 Dissertation Layout

Chapter 2 introduces a literature review on Y-STRs, their discovery and development, as well as their use in forensic investigations. The expanding value of the Y-chromosome in forensic genetics, the promising use of single nucleotide polymorphisms (SNPs) found on the Y-chromosome, the combination of Y-STRs and massively parallel sequencing, as well as, the current use of Y-STRs in South Africa will also be addressed in this chapter.

Chapter 3 focuses on evaluating the forensic value of the 27 Y-Filer® Plus loci in the South African population. Using forensic parameters such as the number of unique haplotypes and discrimination capacity, the potential to use these markers in forensic investigations in South Africa will be discussed. This chapter reports on answering the primary aim of this study, which is to investigate the forensic value of 27 Y-STR loci in four different population groups from South Africa

Chapter 4 assesses and characterises the 27 Y-Filer® Plus loci in the different population groups in South Africa. The genetic diversity within and among the population groups in South Africa, as well as the allelic patterns across the populations will be discussed. This chapter will also focus on various Y-STR profile variations observed in each population group.

Chapter 5 concludes the dissertation, providing a summary of all the results obtained during this study, the implications thereof, and any future recommendations for related research.

References

Africa Check. (2018) '*FACTSHEET: South Africa's crime statistics for 2017/18'*. [Online] Available at: www.africacheck.org/factsheets/factsheet-south-africas-crime-statistics-for-2017-18. [Accessed 11 Feb. 2019].

Chakraborty, R. and Kidd, K. K. (1991) 'The utility of DNA typing in forensic work'. *Science*, 254(5039), pp.1735-1739.

Jobling, M. (2001) 'Y-chromosomal SNP haplotype diversity in forensic analysis'. *Forensic Science International*, 118. pp.158-162

Kayser, M., Caglia, A., Corach, D., Fretwell, N., Gehrig, C., Graziosi, G., Heidorn, F., Herrmann, S., Herzog, B., Hidding, M., Honda, K., Jobling, M., Krawczak, M., Leim, K., Meuser, S., Meyer, E., Oesterreich, W., Pandya, A., Parson, W., Penacino, G., Perez-Lezaun, A., Piccinini, A., Prinz, M., Schmitt, C., Schneider, P. M., Szibor, R., Teifel-Greding, J., Weichhold, G., de Knijff, P., and Roewer, L. (1997) 'Evaluation of Y-chromosomal STRs: A multicenter study'. *International Journal of Legal Medicine*, 110(February 2014), pp.125-133.

Moxon, E. R., and Wills, C. (1999) 'DNA microsatellites: agents of evolution?'. *Scientific American*, 280(1), pp.94-99.

Rape Crisis. (2018) '*The real numbers on sexual offences*'. [Online]. Available at: https://rapecrisis.org.za/the-real-numbers-on-sexual-offence/. [Accessed 11 Feb. 2019]

Roewer, L. (2009) 'Y chromosome STR typing in crime casework'. *Forensic Science, Medicine, and Pathology*, 5, pp.77-84.

Roewer, L. and Willuweit, S. (2020) '*YHRD: Database Statistics'*. YHRD - Y chromosome STR haplotype reference database. [Online]. Available at: https://yhrd.org/pages/resources/stats#haplotype_counts. [Accessed 09 Oct. 2020].

Shewale, J. G., Nasir, H., Schneida, E., Gross, A. M., Budowle, B., and Sinha, S. K. (2004) 'Y-Chromosome STR System, Y-PLEX™ 12, for Forensic Casework: Development and Validation'. *Journal of Forensic Sciences*, 49(6), pp.1-13.

South African Police Service. (2012) '*Crime Situation in Republic of South Africa Twelve (12) Months (April to March 2019_20')*. [Online]. Available at: www.saps.gov.za/services/crimestats.php. [Accessed 06 Jun. 2020].

www.yhrd.org.

CHAPTER 2: LITERATURE REVIEW

2.1 DNA Evidence in Forensic Investigations

When a crime is committed, two types of investigations take place in an attempt to determine what happened during the crime. The first type is the criminal investigation which involves all the police activity focused on identifying, locating, and proving the guilt of a suspect (O'Hara and O'Hara, 1994). The second type of investigation is the forensic investigation. The forensic investigation aims to objectively analyse any physical evidence collected after a crime has been committed in order to determine the events that took place and to connect suspects to the crime using scientific methods (Eckert, 1997).

While the physical evidence is often not the only aspect taken into account during a criminal trial, eye-witness accounts are subjective and could sometimes be unreliable, and so the scientific facts are used to support or refute these testimonies. Forensic investigations are comprised of many facets used in combination to solve a crime. Such facets include, but are not limited to, bloodstain analysis, ballistics, chemistry, fingerprints, and deoxyribonucleic acid (DNA) analysis. During the investigation of a sexual crime, biological evidence would constitute the focus of analysis. Biological evidence includes, but is not limited to, blood and bloodstains, semen and semen stains, saliva, urine and other bodily fluids (Lee and Ladd, 2001). One of the most valuable aspects in the investigation of crimes of a sexual nature is DNA analysis. DNA material can be collected from many types of biological matter. Once the biological evidence is collected, the DNA is extracted from the sample and analysed to create a DNA profile, which can then be used to identify the individual from which the biological material originated. This identification could then be used to connect an accused suspect to the scene of the crime. This process is known as DNA profiling, or DNA fingerprinting, as developed in 1985 by Professor Alec Jeffreys (McKie, 2009).

The first time that DNA was used successfully in a court of law was during the trial of the rapes and murders of Lynda Mann and Dawn Ashworth in 1986 (Evans, 1996). Both girls had been found raped and strangled to death in England in 1983 and 1986 respectively. Semen samples from both bodies were collected and analysed for blood type, which was found to be blood group A. The police's main suspect was a 17-year-old male, Richard Buckland, who matched the blood type and ended up confessing to the murder of Dawn Ashworth during questioning. However, after the development of DNA profiling, this technique was used to examine the semen samples taken from each girl, and it was shown that both samples belonged to the same male, but not Richard Buckland. He became the first person to be exonerated with the use of DNA evidence. Following a mass screening in which over 5 000 males were tested using either blood or saliva samples, Colin Pitchfork was identified as a possible DNA match

and brought in for questioning. He eventually confessed to the rapes and murders of both girls and became the first individual to be identified and convicted using DNA evidence (Wambaugh, 1990). Since this first case, DNA evidence has become invaluable during the investigation of sexual crimes as a means of exonerating or convicting suspects.

2.2 Short Tandem Repeats (STRs)

DNA profiles for forensic analyses are created by targeted amplification of specific DNA regions known as short tandem repeats (STRs). STRs are regions of non-coding DNA in which di-, tri-, tetra-, and pentanucleotide motifs are repeated a certain number of times. The number of times that the motif is repeated depends on the alleles at each locus (Moxon and Wills, 1999). STR loci are hypervariable, meaning there is a high potential to detect a variety of alleles at a single locus, thereby allowing for different allelic combinations among individuals. This polymorphic state, along with the high mutation rate of STR loci, makes STRs an ideal method to differentiate between and identify individuals during forensic investigations (Jobling, 2001).

STRs are currently the preferred genetic markers used in identity testing. The reason for this is that STR markers are relatively short in length, which makes them ideal for the analysis of forensic DNA samples that are often degraded and yield low quantity and quality extracted DNA. Another advantage of STR markers, with their small size, is that they can be amplified efficiently using the standard PCR process (Tamaki and Jeffreys, 2005). STR profiles are presented in the form of electropherograms, as shown in Figure 2.1 (ThermoFisher Scientific, 2019a). Each peak represents an allele that has been detected and typed at each specific marker, which are shown across five different coloured panels.

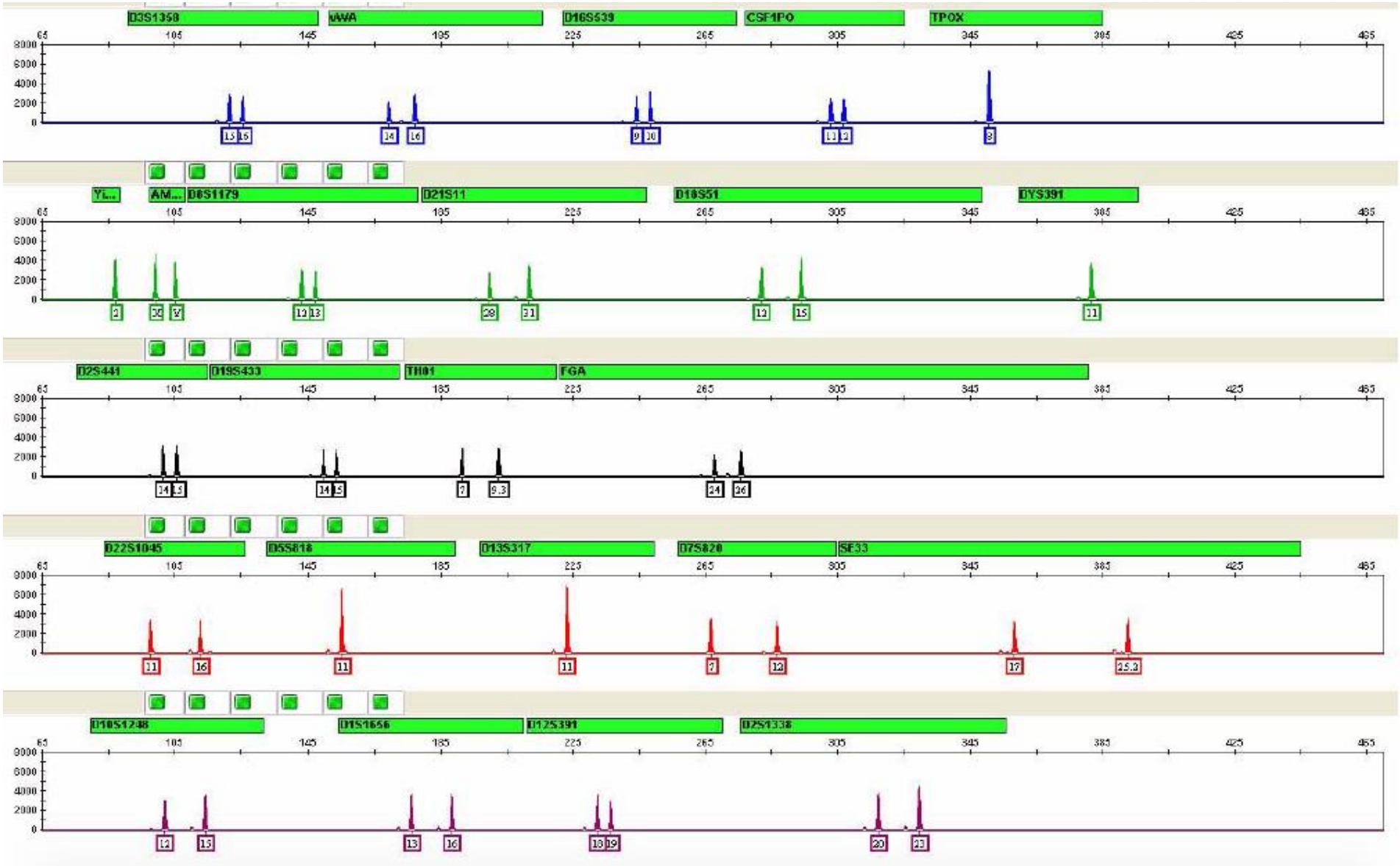


Figure 2.1: *An example of an autosomal STR profile, presented as an electropherogram, taken from ThermoFisher Scientific's GlobalFiler™ and GlobalFiler™ IQC PCR Amplification Kits: User Guide (2019a) – DNA Control 007.*

Once an STR profile is generated from the suspect's DNA sample, it is compared to that of the evidentiary sample. This comparison can result in one of three different outcomes, namely a non-match, match, or an inconclusive result, as shown in Table 2.1 (Chakraborty and Kidd, 1991).

Table 2.1: *The description and interpretation of the possible outcomes of STR profile comparison, adapted from Chakraborty and Kidd (1991).*

Outcome	Description	Interpretation
Non-match or exclusion	Profiles are different — no DNA match	This is evidence that because the profiles are different, they may have originated from different sources.
Null or inconclusive result	Profile comparison not possible	This outcome will be stated if the laboratory was unable to state a precise match or non-match based on the results due to insufficient DNA in the sample or technical issues during the test.
Match or inclusive	No differences were observed between the samples	The samples present a genetic similarity of several DNA loci and can be evidence that the two profiles share a common source.

A match result means that the suspect's DNA profile is identical to that of the evidentiary sample, and so it can be concluded that it was the suspect who left the biological material at the scene of the crime. Once a match result is obtained, statistical support is needed to prove the significance of this DNA match (Chakraborty and Kidd, 1991). The statistical support aims answer two questions: (1) what is the probability that the match occurred by chance, and the suspect is not connected to the evidence?; and (2) is it possible that there are other individuals in the same population group that have this exact DNA profile? Without these probabilities included in the match report, the evidence will hold no statistical value and would, therefore, not be used in a court of law. Match probabilities are calculated as the sum of the squared haplotype frequencies. To calculate these values, a reference database consisting of haplotype frequencies for the loci used in DNA profiling is required. These databases usually contain this information for populations that represent several racial and geographic groups.

Despite all the successes achieved with autosomal STRs, there are some occasions where they either fail to or do not provide sufficient or useful information, particularly during the analysis of DNA evidence from sexual assault cases. In these cases, the answer would be to rather focus on Y-STRs: STR loci that are found on the Y-chromosome.

2.3 The Y-Chromosome

Any karyotype is proof that the Y-chromosome is smaller than the X-chromosome. This was further reported by Buhler (1980), who showed that the human Y-chromosome is, in fact, one of the smallest chromosomes, with an average size of ~60 million/mega base pairs (Mb) compared to the 80 – 248 Mb range of the 22 other chromosomes. Figure 2.2 below provides the structure of the Y-chromosome (Gusmão *et al.,* 2008). The Y-chromosome is divided into two separate arms, with Yp being the short arm and Yq being the long arm. The pseudo-autosomal regions (PARs) are located at the tip of each arm, which contain sequences that are homologous to those on the X-chromosome, and, therefore, undergo genetic material exchange with the X-chromosome during meiosis (Quintana-Murci and Fellous 2001).

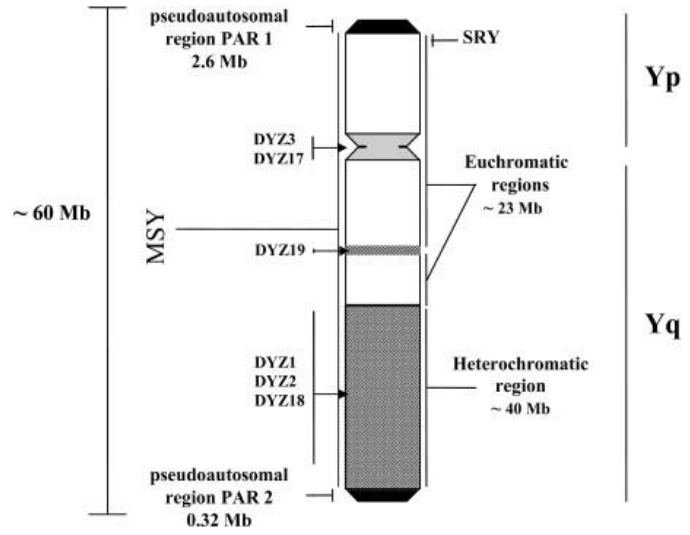


Figure 2.2: *The structure of the human Y-chromosome, showing the sizes of the different components, taken from Gusmão et al. (2008).*

The remaining portion of the Y-chromosome is known as the non-recombining region of the Y-chromosome (NRY), as it does not pair and exchange genetic material with the X-chromosome during meiosis. Owing to the lack of recombination at this region, the NRY is inherited in a haploid state—intact through paternal lineages—and has, therefore, been given the name male-specific region (MSY) as noted by Gusmão *et al.* (2008). The MSY would be passed down from father to son unchanged, unless a mutation occurs. The NRY/MSY consists of two portions, the euchromatin and heterochromatin (Gusmão *et al.,* 2008). The euchromatin is the region in which protein-coding genes and Y-specific repetitive sequences are found and is constant among males. On the other hand, the heterochromatin is known to be non-functional and hypervariable between individuals: to a point being undetectable in some males

(Gusmão *et al.*, 1999). Polymorphic regions have also been discovered within the heterochromatin, with many polymorphisms such as STRs and SNPs located in this area (Quintana-Murci and Fellous, 2001).

The location of polymorphisms in the heterochromatin within the NRY/MSY—and the male-specific inheritance pattern thereof—allows for the successful use of the Y-chromosome in male identity testing. Male identity testing has many applications including forensic casework on sexual assault evidence, paternity testing, missing person investigations, human migration and evolutionary studies, and historical and genealogical research (Butler, 2003). For the purposes of this study, focus falls on the use of Y-STRs from a forensic perspective during the investigation of sexual assaults.

2.4 Y-STRs

The use of Y-STRs during the investigation of sexual offences has proven to be significant since the discovery of the first Y-STR marker in 1992. A timeline is presented of the discovery of Y-STRs, development of Y-STR commercial kits and databases (Figure 2.3), along with the introduction of core Y-STR markers, as found in Table 2.2 (Butler, 2003; Kayser *et al.,* 2004; Ballantyne, 2012).

The concept of Y-STR DNA typing came about in 1992 when *Y-27H39*—the first Y-STR marker—was discovered (Butler, 2003). This marker has since been renamed to *DYS19.* In contrast to autosomal STRs, the identification and introduction of Y-STRs progressed at a much slower rate. By the beginning of 2002, only 30 Y-STR markers had been introduced (Table 2.2). Despite the slow start, Y-STRs suddenly expanded rapidly, with 149 and 52 markers being introduced in 2002 and 2003 respectively.

In 1997, Kayser *et al.* considered some Y-STR loci for potential use in a forensic application. It was in this year that the ‘minimal haplotype’ (MH) was established: a core set of eight loci deemed to be a sufficient representation for providing haplotypic information for forensic purposes. The loci included are *DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, and DYS385*. The addition of the *YCA II* locus resulted in the nine loci which make up the ‘extended haplotype’ (Butler, 2003). In 2003, the Scientific Working Group on DNA Analysis Methods (SWGDAM) recommended that two additional loci, *DYS438 and DYS439*, should be included in the MH, to replace the *YCA II* locus, as there were often technical difficulties experienced when attempting to type dinucleotide Y-STRs (Butler *et al.,* 2007).

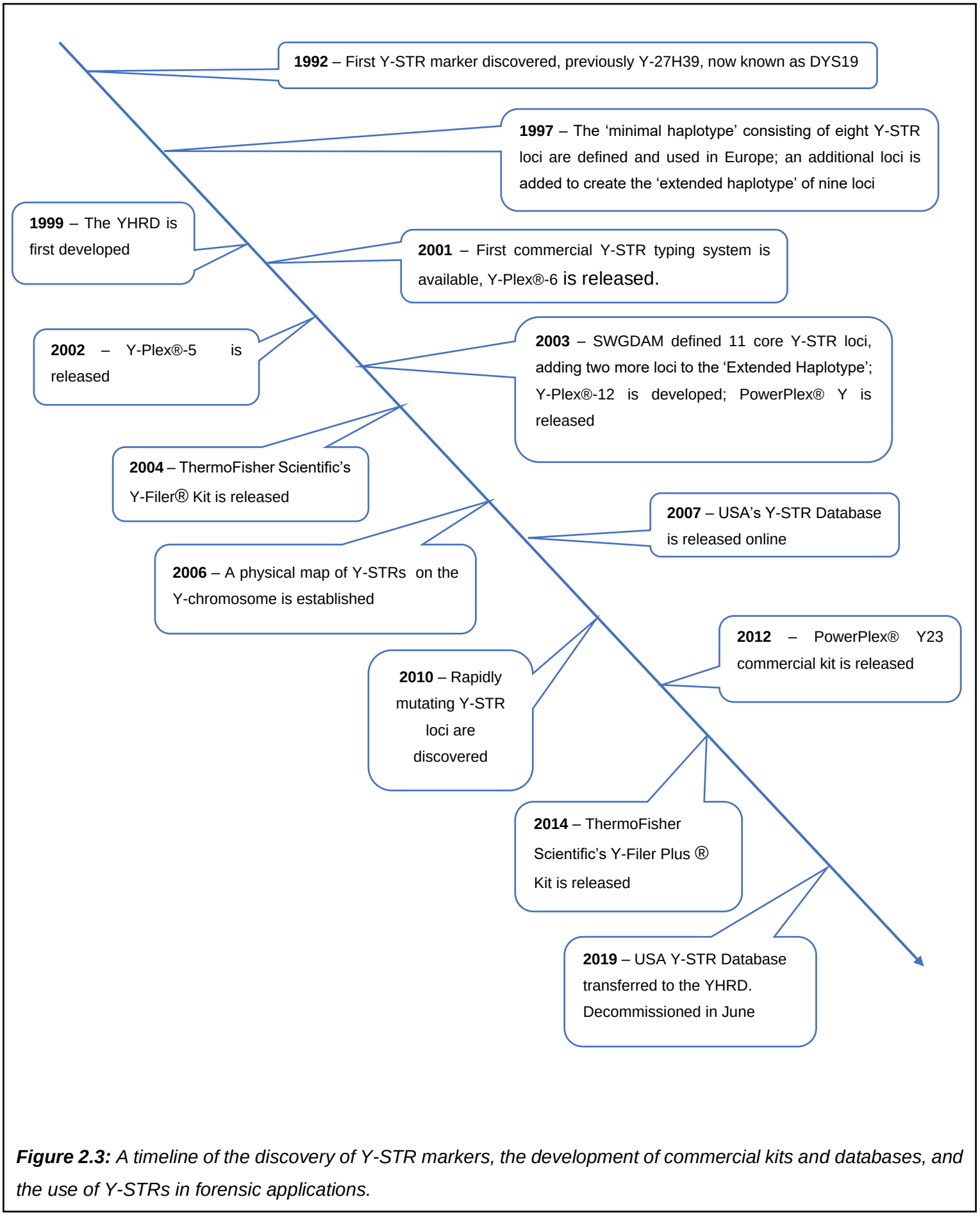


Figure 2.3: *A timeline of the discovery of Y-STR markers, the development of commercial kits and databases, and the use of Y-STRs in forensic applications.*

Table 2.2 below lists the many Y-STR loci that were discovered, and the year in which they were discovered, when Y-STRs were introduced to the forensic science community (Butler, 2003; Kayser *et al.,* 2004; Ballantyne, 2012).

Table 2.2: *The history of Y-STR marker discoveries between 1992 and 2003 (Butler, 2003).*

Year	Markers
1992	*DYS19 (*Previously known as *Y-27H39)*
1994	*YCAI; YCAII; YCAIII (DYS413); DXYS156*
1996	*DYS389I/II; DYS390; DYS391; DYS392; DYS393; DYF371; DYS425; DYS426*
1997	*DYS288; DYS388*
1998	*DYS385*
1999	*A7.1 (DYS460); A7.2 (DYS461); A10; C4; H4*
2000	*DYS434; DYS435; DYS436; DYS437; DYS438; DYS439*
2001	*DYS441; DYS442*
2002	*DYS443; DYS444; DYS445; DYS462; DYS446; DYS447; DYS448; DYS449; DYS450; DYS452; DYS453; DYS454; DYS455; DYS456; DYS458; DYS459; DYS463; DYS464; DYS468-DYS596 + 129 others*
2003	*DYS597-DYS645 + 50 others*

Figure 2.4 provides the position of several Y-STR loci on the Y-chromosome, with the MH loci indicated in blue (Hammer and Redd, 2016). As indicated in the figure, 11 loci are located on the short arm (p) and 15 loci on the long arm (q). The loci located on arm p are more closely positioned as opposed to the loci on arm q, as arm q is longer. It should be reiterated that Y-STR loci are found in the non-recombining portion of the Y-chromosome, giving them their uniparental inheritance pattern.

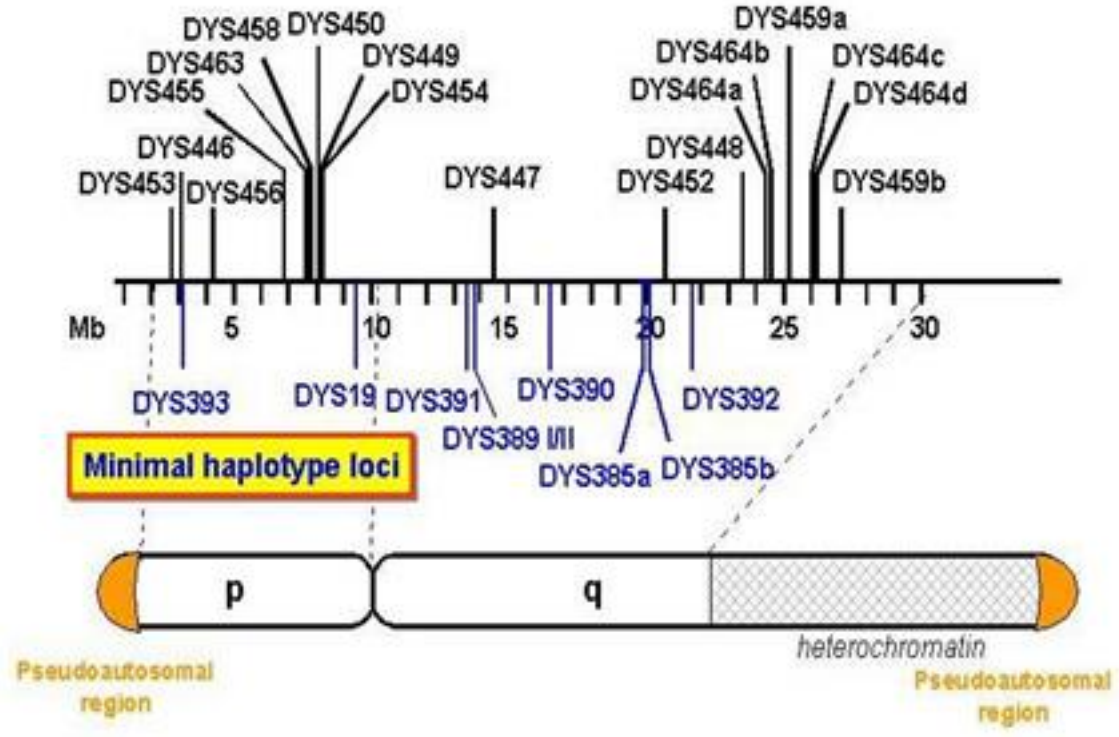


Figure 2.4: *The positions of some, but not all, Y-STRs along the Y-chromosome (Hammer and Redd, 2016).*

2.5 Commercial Y-STR Typing Kits

In 2001, the first commercial Y-STR typing kit was released. The Y-Plex®-6 kit simultaneously amplified six loci, *DYS393, DYS19, DYS389II, DYS390, DYS391, and DYS385* (Shewale, 2003). The development of the Y-Plex®-5 and Y-Plex®-12 kits followed in 2002 and 2003 respectively. The Y-Plex®-5 system used multiplex PCR to amplify five markers simultaneously: namely, *DYS389I, DYS389II, DYS439, DYS438,* and *DYS392*. When combined with the Y-Plex®-6 system, all nine loci included in the MH were analysed. The two additional loci, *DYS438* and *DYS439,* were later included so that the 11 core Y-STR loci recommended by SWGDAM were incorporated into the DNA analysis. Both Y-Plex®-6 and Y-Plex®-5 were deemed validated, sensitive, reliable, robust, and sufficient for use in analysing forensic evidence (Shewale, 2003). The Y-Plex®-12 kit was developed a year later for amplification of all 11 core Y-STR loci with the intention of combining the Y-Plex®-6 and Y-Plex®-5 kits (Shewale *et al.,* 2004). The Y-Plex®-12 system also included the sex-determining gene of amelogenin to serve as an internal control for PCR. This kit was also deemed validated, sensitive, reliable, robust, and useful in human forensic and male lineage

identification cases. None of these Y-Plex® kits are commercially available anymore. In 2003, Promega also developed and released their Power Plex® Y commercial kit (Krenke *et al.,* 2003). The PowerPlex® kit was expanded from 12 loci to 23 loci in 2012 when the PowerPlex® Y23 kit was released (Thompson *et al.,* 2012).

ThermoFisher Scientific released their commercial kit, Y-Filer®, in 2004, which types 17 different Y-STR loci (ThermoFisher Scientific, 2006). This kit was further developed in 2014, when 10 additional loci were included in the typing system, which resulted in the Y-Filer® Plus PCR Amplification Kit (ThermoFisher Scientific, 2019b). The Y-Filer® Plus kit has been approved for use in generating profiles for inclusion in the Combined DNA Index System (CODIS) database. Of the 27 Y-STR loci included in the Y-Filer® Plus Kit, seven loci are known to be rapidly mutating (RM) loci, making this kit especially useful in the analysis of male-specific forensic DNA evidence.

2.6 Use of Y-STRs in Forensic Investigations

During the forensic investigation of a violent crime, various types of biological evidence can be collected at the scene of a crime (Lee and Ladd, 2001). Typical forensic evidence collected at the scenes of crimes of a sexual nature includes vaginal swabs from the victim, semen, and saliva (Hall and Ballantyne, 2003).

Despite the successes experienced with autosomal STR markers over the years, there are occasions in which they could fail to provide sufficient DNA profiles for analysis. With the kinds of samples collected after a sexual assault, the biological material from a male perpetrator is frequently mixed with biological material from a female victim: often with the female DNA being higher in quantity than that of the male's. In such cases, it can become difficult to separate the male autosomal profile from the female's, as a large amount of female DNA could end up completely concealing the male DNA. Differential extraction techniques could be used to separate the sperm cells from the vaginal epithelial cells before DNA analysis; however, fresh samples are required for this technique and it is not always possible to collect such samples (Hall and Ballantyne, 2003). There is also the possibility that the semen samples collected could be from males who are azoospermic or oligospermic, or have had vasectomies or orchidectomies in the past, where autosomal cells would not be available (Shewale *et al.,* 2004).

Commercially available PCR kits do generally include the sex-determination marker amelogenin, although there have been reports that this marker is prone to typing errors and

is not always reliable (Kayser and Schneider, 2009). A deletion in the azoospermia factor (AZF) gene results in the non-amplification of the amelogenin gene, only detecting the X-chromosome, thereby resulting in a 'false female'. To overcome this problem, newly developed commercial kits, have begun to include other sex-determining markers such as a Y-STR marker (*DYS391*) and an insertion/deletion (indel) polymorphic marker on the Y-chromosome (Y indel) as described by ThermoFisher Scientific, (2019a). It is worth noting that including these markers is not always adequate in the analysis of complex mixture samples.

Given the abovementioned disadvantages of autosomal STRs, Y-STRs are becoming a popular alternative in investigations of rape and other crimes of a sexual nature. The ultimate goal for Y-STRs is not to replace autosomal STRs, but rather to be used in combination with autosomal STRs as a means of obtaining the best possible DNA profile for use as forensic evidence.

As already discussed, Y-STR loci are only found in individuals who are genotypically male (XY). The female DNA is, therefore excluded from any DNA analysis, eliminating the risk of it concealing the male DNA (Roewer, 2009). As a result of the female DNA being excluded, complete male profiles can be obtained even when mixed with large amounts of female biological material. An example of a male DNA profile is given in Figure 2.5. In comparison to the autosomal profile provided in Figure 2.1, there are fewer peaks displayed on the electropherogram, given the haploid state of the Y-chromosome.

Profiles consisting of male-only DNA are more straightforward to interpret than mixture profiles consisting of both male and female profiles. Mixture profiles of autosomal DNA can have up to four alleles per locus if two people contribute to the sample (Hu *et al*., 2014), which is often the case with rape samples. Autosomal STRs can also only be used to resolve mixture samples with one male donor, as more than one male profile becomes very complicated (Redd *et al.,* 2002). If the female DNA is excluded and only Y-DNA remains, there can only be one allele per locus (as seen in Figure 2.5) because Y-STR loci are haploid, with the exception of a few multicopy loci (indicated with red circles in Figure 2.5), resulting in fewer alleles to take into consideration. Fewer alleles to interpret gives Y-STR typing the advantage of being able to determine the number of male donors in a mixture sample and to resolve these profiles more efficiently, which is especially beneficial for samples from gang rape (Hall and Ballantyne, 2003).

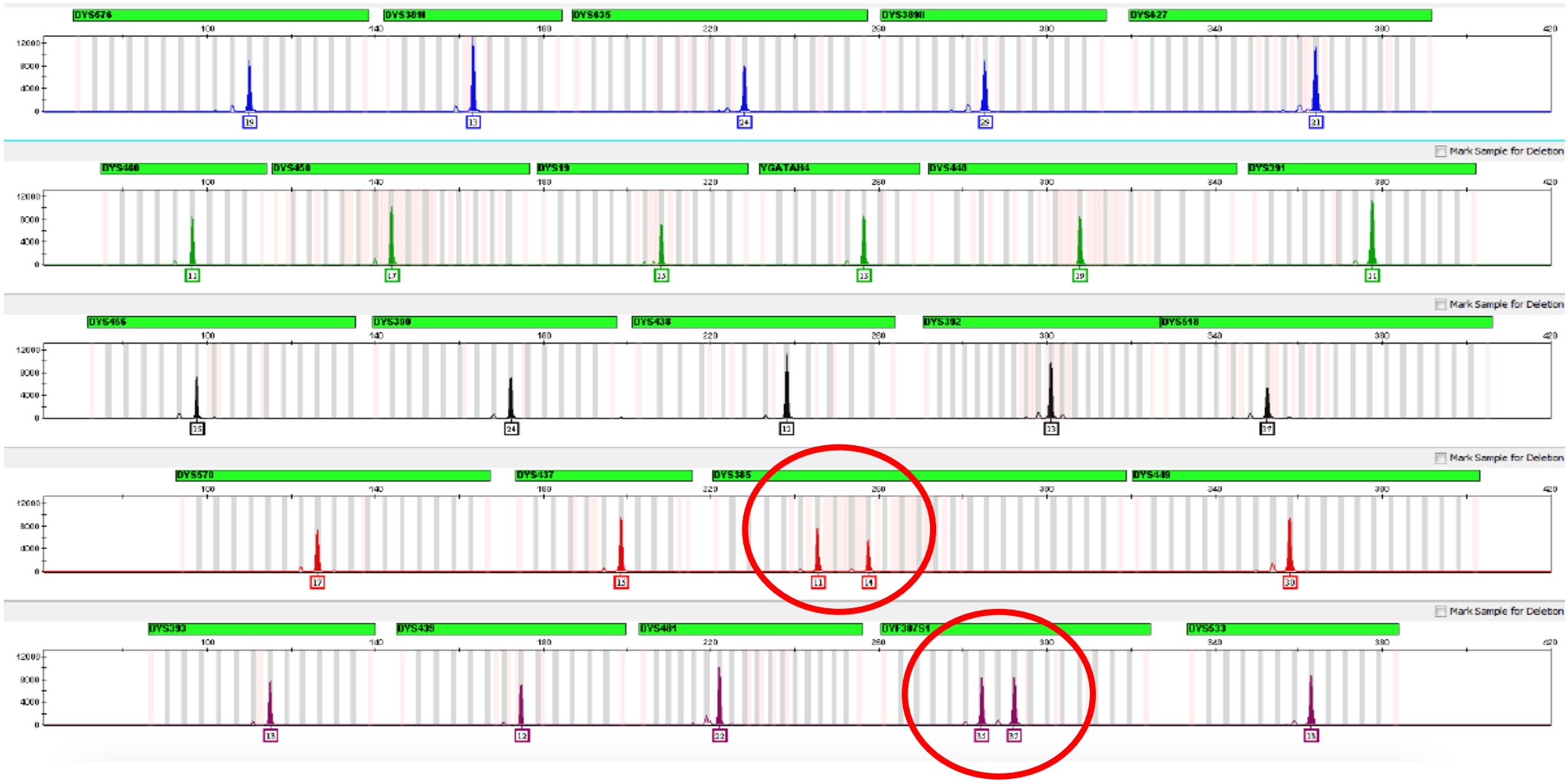


Figure 2.5: *An example of a Y-STR profile, presented as an electropherogram, taken from ThermoFisher Scientific's Yfiler™ Plus PCR Amplification Kit: User Guide (2019b) – DNA Control 007. Two heterozygous loci, DYS385 and DYF387S1, are encircled in red.*

Given the uniparental inheritance pattern of the Y-chromosome, it is possible to trace the parental lineage back to the origin thereof. Y-STRs can, therefore, provide investigators with the ethnic origin of a male individual from a DNA sample (Roewer, 2009). This information can prove to be useful during crime investigations when attempting to identify suspects.

Despite the progress made with the use of Y-STRs, there is still an uncertainty regarding the inclusion of these loci in forensic analysis. A significant challenge experienced with Y-STRs is the inability to differentiate between closely related males as they would all share the same Y-chromosome unless a mutation occurred during meiosis (Gill *et al*., 2001). When a match between a suspect and an evidence sample is found, any male relatives of the suspect cannot be excluded unless there is other evidence proving innocence (Roewer, 2009). The ability to differentiate between individuals, related or unrelated, using Y-STR loci is measured using the discrimination capacity. The discrimination capacity is a value, expressed as a percentage, that is calculated by dividing the number of different haplotypes observed in a population by the total number of samples in that population (Redd *et al.,* 2002). A discrimination capacity of 100% means that all the haplotypes in the population are different, and the combination of markers used can differentiate between 100% of the males in that population. The goal for selecting Y-STR loci for use in forensic analyses is, therefore, to achieve a discrimination capacity as close to 100% as possible. The challenge with the lower mutation rates associated with Y-STR markers means fewer differences between generations, which leads to reduced discrimination capacity between related individuals. Fortunately, the inclusion of additional markers, as well as rapidly mutating (RM) markers in commercial PCR amplification kits, has allowed for better resolution between related males.

2.7 Y-STR Mutation Rates

Considering the nature of the Y-chromosome once again, the NRY region does not undergo any form of recombination, and so there will be no genetic variation on the Y-chromosome between related males unless a mutation occurs during meiosis (Gusmão *et al.,* 2008). Fortunately, several molecular factors have an influence on mutation rates of STR markers on the Y-chromosome (Claerhout *et al.,* 2018). Such factors include: the length of the repeat motif; the average number of repeats per locus; the complexity of the repeat motif; and occasionally the age of the father at the time of Y-chromosome inheritance.

An accurate method for calculating the mutation rates of Y-STRs is the direct counting method, in which the total number of observed mutations is divided by the total number of meiosis events/generations, provided that a large number of father-son pairs is considered (Ballantyne

et al., 2010). A study performed by Goedbloed *et al.* (2009) revealed that mutation rates of the 17 loci included in ThermoFisher Scientific's Y-Filer® kit varied between 2×10^{-4} and 6.5×10^{-3} per locus per generation. Although there is no significant difference between these mutation rates and that of autosomal STRs (Gusmão**,** and Carracedo, 2003), such mutation rates can be problematic when these Y-STRs are used in a forensic setting. A lack of mutations during inheritance means that paternally related males would all share the same Y-chromosome, making it virtually impossible to differentiate between them. As a result, it would be difficult to exclude related male suspects in the investigation of a crime based on DNA evidence (Roewer, 2009).

The introduction of rapidly mutating (RM) Y-STR markers in 2012 by Ballantyne *et al.* has proved revolutionary in forensic DNA evidence from rape and sexual assault cases. Ballantyne *et al.* investigated 189 Y-STR markers and their mutation rates. Standard mutation rates varying from 1×10^{-4} to 1×10^{-3} were estimated for 176 of those markers. During this study, 13 Y-STR markers were proven to have notably higher mutation rates than the others, with rates ranging between 1.19×10^{-2} and 7.73×10^{-2}. It is the enhanced discrimination capacity between related males that is of particular interest within the forensic scope. Ballantyne *et al.* (2012) showed that these 13 new RM loci performed better when analysing related males than the 17 loci of the Y-Filer® commercial kit. The new RM loci were able to differentiate between the two males in 48.7% of father-son pairs, 60% of brother pairs, and 75% of cousin pairs, while Y-Filer® could only do this in 7.7%, 8.0%, and 25% of the respective pairs.

Several RM loci have already been included in commercially available kits as to enhance to use of Y-STRs in forensic identity testing. One such kit is ThermoFisher Scientific's Y-Filer® Plus Kit, which includes seven RM loci (Table 2.3; Goodur, 2018). Although the inclusion of RM in commercial Y-STR kits is highly advantageous in analysing forensic DNA evidence, there are a few cases in which it could prove to be unreliable (Baeta *et al.,* 2018). In cases where Y-STR profiles are compared to potential relatives—such as in paternity testing or missing person identification—high mutation rates could result in false exclusions. This predicament has led to the introduction of another type of Y-STR: the slowly mutating (SM) Y-STR markers. Six SM loci have been presented, *DYS388, DYS426, DYS461, DYS485, DYS525, and DY561*, with mutation rates between 3.98×10^{-4} and 9.89×10^{-4}. These loci are more stable and may be useful in cases where even the smallest of differences are crucial and are reported as exclusions, such as for paternity cases. In addition, SM loci could be beneficial in proving legitimate exclusions. A combination of SM and RM loci could prove to

be very valuable to the forensic community. Another solution to this problem could be the implementation of next-generation sequencing (NGS), or massively parallel sequencing (MPS), in a forensic setting (Qian *et al.,* 2017).

Table 2.3: *The mutation rates of the 27 Y-STR markers included in ThermoFisher Scientific's Y-Filer® Plus PCR Amplification Kit (Goodur, 2018). Rapidly mutating loci are shown in red.*

Marker	Mutation Rate (x 10^{-3})
DYS518	18.4
DYF387S1 a/b	15.9
DYS576	14.3
DYS570	12.4
DYS627	12.3
DYS449	12.2
DYS458	8.36
DYS460	6.22
DYS389I	5.51
DYS553	5.01
DYS481	4.97
DYS456	4.94
DYS19	4.37
DYS385 b	4.14
DYS635	3.85
DYS439	3.84
DYS389II	3.83
DYS391	3.23
Y GATA H4	3.22
DYS393	2.11
DYS385 a/b	2.08
DYS437	1.53
DYS390	1.52
DYS438	0.96
DYS392	0.97
DYS448	0.39

2.8 Y-SNPs

Budowle and Van Daal (2008) defined SNPs as 'base substitutions, insertions, or deletions that occur at single positions in the genome of any organism,' and Y-SNPs are simply those variations found on the Y-chromosome. The majority of known SNPs are biallelic, so they are not considered to be highly polymorphic or hypervariable. Consequently, SNPs are not as revealing on the locus level as most STR loci, and a higher number of markers would be needed to provide sufficient information. Nonetheless, SNPs account for ~85% of the total variation in the human genome (Budowle and Van Daal, 2008) and can be successfully amplified using very short fragments. Thus, SNPs would be especially useful when analysing heavily degraded DNA samples, which is often the case with forensic crime scene samples (Lessig *et al.,* 2005).

The first Y-SNP was discovered in 2003 through the identification of the Y-chromosome Alu Polymorphism (YAP) marker, an insertion variation (Butler, 2003). This Y-SNP was revealed to occur more frequently in the African population than in Europeans. In terms of Y-STRs, a haplotype is the set of Y-STR alleles that are directly inherited by the son from his father. Conversely, in terms of Y-SNPs, a haplogroup is the set of the Y-SNPs that are inherited down a paternal lineage (Qian *et al.,* 2017). As with haplotypes, paternal male relatives are thought to share the same haplogroup.

Regarding the needs of DNA analysis in forensic investigations, SNPs are certainly at a disadvantage compared to STRs, given their low discrimination capacities. Qian *et al.* (2017) reported that the mutation rate of Y-SNP markers, ranging between 1×10^{-8} and 1×10^{-9}—almost negligible values—is ~100 000 times lower than that of Y-STRs. However, this low mutation rate of Y-SNPs allows for the ancestral and population-specific origin of a male individual to be determined with relative ease (Lessig *et al.,* 2005). Although Y-SNPs do not have the same discrimination capacity as Y-STRs, they are advantageous in other circumstances. Lessig *et al.* (2005) showed that Y-SNP assays are exceptionally sensitive and could successfully genotype samples with less than 125 pg of DNA. Additionally, there were no problems experienced with detecting Y-SNPs in male-female mixture samples, with the Y-SNPs avoiding concealment by the female, which is a common occurrence when dealing with autosomal STRs.

The identification of a Swiss war hero from the 17th century is a perfect example of how Y-SNPs and Y-STRs can be used in combination during forensic DNA identity testing (Haas *et al.,* 2013). Jörg Jenatsch fought for independence and liberty in Switzerland in the 17th century,

and so participated in several acts of violence such as assassinations and murders. Some of his many victims were members of a noble family, a family who wanted revenge and exiled him. After seeking refuge in Venice and becoming a professional soldier and military entrepreneur, he was eventually assassinated. His body was buried in a cathedral with the exact location being unknown. When this burial place was discovered and his body exhumed, fabric from a piece of cloth was analysed, and three male members of the Jenatsch family were traced and the family tree reconstructed.

His body was exhumed again in 2012 and bone and teeth used to collect DNA samples. Based on the SNaPshot technique, 21 Y-SNPs deemed adequate for defining the most common European haplogroups were analysed (Haas *et al*., 2013). The skeleton and three supposed male relatives all belonged to the same Y-SNP haplogroup. When this haplogroup was compared against those in the YHRD, it was revealed that that specific haplogroup was quite common in the region of the supposed family and was not concrete evidence of a familial relationship. In addition to the Y-SNP data, a complete PowerPlex® Y23 profile was generated from the bone and teeth matter. A comparison of this profile to that of the three male relatives showed that the profiles were a match at 20 loci, but that there were mismatches at 3; however, mutations could undoubtedly have occurred over the generations resulting in those mismatches. Nonetheless, when combining the Y-SNP and the Y-STR results, the statistics revealed that it was at least 20 times more likely that the skeleton was, in fact, Jörg Jenatsch.

Although there have been several successes through using Y-SNPs as an additional forensic tool, it is not likely that Y-SNPs will replace Y-STRs as the primary method of Y-DNA-specific forensic identity testing any time soon (Budowle and Van Daal, 2008). Aside from the low mutation rates that could prove to be a disadvantage when trying to differentiate between related males, the large number of Y-SNPs that need to be analysed to provide the same information as Y-STRs prevent Y-SNPs from becoming an alternative genotyping system in the forensic community. However, Y-SNPs can be a valuable addition to forensic genetics when used in combination with Y-STR markers.

2.9 Y-STRs and Massively Parallel Sequencing

Once understanding how Y-STRs and Y-SNPs can be used in combination to identify and located male offenders, one can consider the use of sequencing techniques in a forensic setting. Next-generation sequencing (NGS), or hereafter massively parallel sequencing (MPS), is a sequencing technique that analyses millions of small, targeted fragments of multiple DNA samples at the same time, or in parallel (Sousa, 2017). MPS has recently

become of interest in the forensic community as a technology that has the potential to significantly improve the discrimination power of Y-STRs in forensic casework applications. The fact that Y-STRs are inherited as a whole haplotype intact from father to son remains a shortcoming in the use of Y-STR genotyping systems as it is difficult to distinguish between closely related males unless mutations occur (Gill *et al.,* 2001). This problem could persist even with the presence of RM loci in commercially available Y-STR kits.

Conventional STR profiles are generated using capillary electrophoresis (CE), with the alleles at each locus being detected according to their sizes; these differ depending on the number of repeat units (de Knijff, 2019). Y-STR profiles are generated in the same way. While Y-STRs have indeed proven to be useful in distinguishing between two male individuals, MPS may provide the technology to even further that level of distinction. MPS analyses Y-STR markers deeper than just the allelic level as it analyses it at the nucleotide level (Warshauer *et al.,* 2015). For example, two males may have allele 9 at the *DYS19* locus, and it would be detected as such via CE as both alleles would have nine repeat units. However, there may be a SNP or microvariant within the allele at a nucleotide level that results in a sequence difference between the two males without affecting the length of the allele. Conventional Y-STR typing would not detect this difference, but MPS would, allowing to differentiate between two seemingly identical profiles. This level of distinction is so powerful that it has been reported that MPS was able to distinguish between identical twins, something that is impossible with conventional STR typing (Bruijns *et al.,* 2018). This concept would imply that the detection technique of CE results in allelic frequencies being overestimated while sequence variation is underestimated (Sousa, 2017).

In addition to improving the discriminating power of Y-STR markers, MPS technologies are also beneficial in that they could easily detect unidentified SNPs and novel motif variants, and this would simply add to the knowledge that is needed in this industry (Warshauer *et al.,* 2015). As MPS is able to differentiate between individuals to such an extent, and identify novel Y-SNPS, further haplogrouping in Y-lineage studies would be possible. This could, in turn, prove to be advantageous in forensic investigations making use of Y-SNP analysis (Qian *et al.,* 2017). For instance, when comparing the evidence DNA sample to that of the suspect, a different haplogroup present in the suspect's sample would indicate that the two samples originated along different lineages and that the suspect should be excluded from the investigation.

MPS is not used in day-to-day forensic DNA analyses as of yet for two simple reasons: time and cost (Bruijns *et al.,* 2018). The entire process of DNA extraction, library generation, and

data analysis can take several days to complete, which is a disadvantage in the forensic field where timely delivery of results is crucial. In addition to time, the cost of installing the MPS machines and purchasing the necessary kits cannot be justified as yet, especially as MPS has not entirely replaced conventional STR typing. Nonetheless, MPS is most certainly an emerging technology for forensic purposes which has grown so much already in the past few years, and it will no doubt continue to provide valuable information for the forensic community.

In an attempt to promote the use of MPS in the analysis of forensic DNA evidence, de Knijff (2019) has specified several recommendations that should be implemented. The recommendations include: (1) a consistent nomenclature for MPS-based STR alleles; (2) suggestions regarding the number of reads needed to call an STR allele in various types of samples accurately; (3) provision of information concerning non-target, or error, reads; (4) recommendations on the MPS strategy to be used – for example, if an amplicon should be sequenced in full or partially and then assembled and aligned; (5) proposals on the storage of MPS results; (6) specific requirements of the analysis software; and (7) updated statistical software to interpret the evidential value of a match result based on the new allele designations provided by MPS.

2.10 Y-STR Forensic Databases

The need for DNA databases containing population reference data for forensic investigations has been evident since the very first database was established in the United Kingdom (UK) (Wallace, 2006). Reference databases are a necessity regardless of the type of DNA profiles they contain. Since the discovery of Y-STRs for use in forensic investigations, a reference database has been established in an attempt to: (1) identify polymorphisms capable of discriminating between unrelated lineages in a given population; (2) establish a database representative of geographical and ethnical groups of the populations of interest; and (3) create a database that would allow accurate frequency estimation for rare haplotypes.

Known as the Y-chromosome Haplotype Reference Database (YHRD), this database contains population-specific haplotype frequency data of multiple Y-STR loci (Roewer *et al*., 2001). Developed and released in 1999, the YHRD contains reference data for several sub-datasets of haplotypes including the minimal and extended haplotypes, PowerPlex® Y, Y-Filer®, PowerPlex® Y23, Y-Filer® Plus, maximal haplotypes, and even Y-SNPs. To date, the YHRD is the most comprehensive Y-STR reference database that is commercially and freely available online (www.yhrd.org). The United States of America (USA) also created its Y-STR reference database, US Y-STR, in 2007 with the intention of allowing estimation of haplotype

frequencies for five different American population groups, based on the 11 extended haplotype loci recommended by SWGDAM (Sousa, 2017). The US Y-STR Database was decommissioned in June 2019 and all the haplotypes moved to the YHRD for the long-term operation and stability of the haplotypes. At the time of decommission, the database contained a total of 32 972 haplotypes of the 11 SWGDAM loci (www.usystrdatabase.org).

The rape and murder of Marianne Vaatstra from the Netherlands is an excellent example of how the YHRD could be used for forensic purposes (Kayser, 2017). In May 1999, Vaatstra was raped and murdered, with semen found in and on her body. There were no human eyewitnesses available to provide information on what had happened. Semen samples were taken, and the DNA analysed, but there were no matches for the autosomal profile on the national criminal offender DNA database. Several suspects were investigated over the following months, but there were no profile matches for any of them. A Y-STR profile was generated from the semen sample and compared against the profiles in the YHRD. This comparison revealed that the semen donor's ancestry could be traced back to North-western Europe. The police could then refocus their search to look for a male from the Dutch European population in the area in which Vaatstra was killed. In September 2012, 13 years after the crime, 6 600 men from that region provided their DNA sample for Y-STR profiling in a DNA dragnet as a last resort to solve the case. The intention was to use the Y-DNA profile in a familial search to locate a male relative of the offender. Surprisingly, though, one of the profiles from the dragnet matched that of the semen sample. The suspect's sample was used to generate an autosomal profile which then also matched the semen sample. The donor of this sample was eventually found and arrested. Jasper S of Dutch European ancestry, who lived 2.5 km away from the murder site, confessed to the rape and murder. He was consequently convicted and sentenced in April 2013, 14 years after he had committed the crime.

Y-STR reference databases are not intended to operate like the standard CODIS software. In essence, it is not for identifying individuals based on their Y-STR profile, but rather to infer ancestral origin and calculate accurate haplotype frequencies and match probabilities (Kayser, 2017). Foreman and Evett (2001) emphasise the need to provide statistical value with any evidence presented in a court of law. In the case of DNA evidence, this value is the probability that the DNA match could occur with any other random individual in the population, aiding in the exclusion of suspects. In the case of Y-DNA profiles, it is the haplotype frequency—obtained from the reference database—that allows this probability to be calculated (Kayser, 2017). A disadvantage of using Y-STR profiles is that they are significantly more variable than single autosomal STR loci, meaning that Y-STR haplotype reference databases need to be considerably larger than standard autosomal STR databases in order to provide accurate and

reliable match probability values. Another disadvantage with Y-STR databases is that the haplotype frequencies may prove to be unreliable as the YHRD contains only unrelated males (Kayser and de Knijff, 2011). It has been suggested that databases should contain Y-STR profiles from unrelated as well as related individuals in order to provide more accurate, and consistent, haplotype frequency estimates and match probabilities.

As of October 2020, there are 1 080 656 haplotypes in total on the database from several population groups (www.yhrd.org; accessed 09 October 2020). However, the majority of these haplotypes have been contributed by individuals in more developed countries, with European and Asian samples dominating the database. Only 6% of the total YHRD samples were contributed by African countries, with Sub-Saharan African contributing a mere 1% to the whole database. The YHRD separates all the uploaded haplotypes into the 'National Databases.' The National Databases provide the observed and expected frequencies for the haplotypes detected in each of the 139 countries that have contributed to the YHRD, South Africa included. These haplotypes include loci in the minimal, PowerPlex® Y, Y-Filer®, PowerPlex® Y23, Y-Filer® Plus and maximal haplotypes.

To date, there have been 89 574 haplotypes that include the 27 Y-STR loci of ThermoFisher Scientific's Y-Filer® Plus PCR Amplification Kit uploaded to this database (Roewer and Willuweit, 2020). Despite this large number, there are no Y-Filer® Plus haplotypes on this database for the South African population. The South African population is not only lacking Y-Filer® Plus data, but haplotype data across the board. Currently, South Africa only has 1 572 minimal, 758 PowerPlex® Y, 758 Y-Filer®, and 740 PowerPlex® Y23 haplotypes uploaded to the database (www.yhrd.org). In addition, there is no available data for the maximal haplotype: a set of 27 loci defined by the YHRD. It is clear that this kind of reference data is crucial for the successful use of Y-STR analysis in forensic investigations. Thus, the aim of this research is also to provide knowledge regarding the haplotype frequencies of the South African population using the Y-Filer® Plus 27 loci.

2.11 Population Structure in South Africa

When forensic evidence is presented in a court of law—particularly DNA evidence—there must be statistical support for this evidence to be considered valuable. This statistical value of DNA evidence is determined by providing the probability that other individuals in the population have the exact DNA profile in question (Chakraborty and Kidd, 1991). Population

reference databases are required to calculate these statistics. However, in order to create these databases, background information on the local population structure is needed.

The 'Mid-Year Population Estimates' released by Statistics South Africa for 2020 reported that approximately 59.6 million people were living in South Africa by the end of July 2020. Table 2.4 shows the racial distribution of the population, as well as the number of males and females per population group at the time of the report's release (Statistics South Africa, 2020). Table 2.4 shows that the number of females in the South African population is slightly higher than that of the males. Despite this, females fall victim to sexual assaults more frequently than males. While men are also certainly abused, statistics show that just over twice as many females reported sexual offences by June 2018 (Statistics South Africa, 2018). Men are also more likely to be the perpetrators than the victims of sexual assault (Gordon, 2002).

Table 2.4: *The distribution of males and females in each population group in South Africa by July 2020 (Statistics South Africa, 2020).*

Population Group	**Males**		**Females**		**Total**
	Number	**%**	**Number**	**%**	
Asian/Indian	787 662	2.70	753 451	2.50	1 541 113
African	23 519 474	80.70	24 634 253	80.80	48 153 727
Coloured	2 555 204	8.80	2 692 536	8.80	5 247 740
Caucasian	2 266 535	7.80	2 413 235	7.90	4 679 770
Total	**29 128 875**	100	**30 493 475**	100	**59 622 350**

The majority of the South African population are placed in the African group. The category of African is too broad, however, and can be further divided into multiple ethnic subgroups, based on their home language and cultural practices. These subgroups include Zulu, Xhosa, Sotho, Pedi, Venda, Tswana, Tsonga, Swati, and Ndebele. Figure 2.6 shows the distribution of the South African population based on home language, reported during the census conducted in 2011 (Statistics South Africa, 2012). IsiZulu is the most predominant language in South Africa with 22.7% of the population speaking it as their home language. IsiZulu is then followed by IsiXhosa at 16%, Afrikaans at 13.7%, and English at 9.6%.

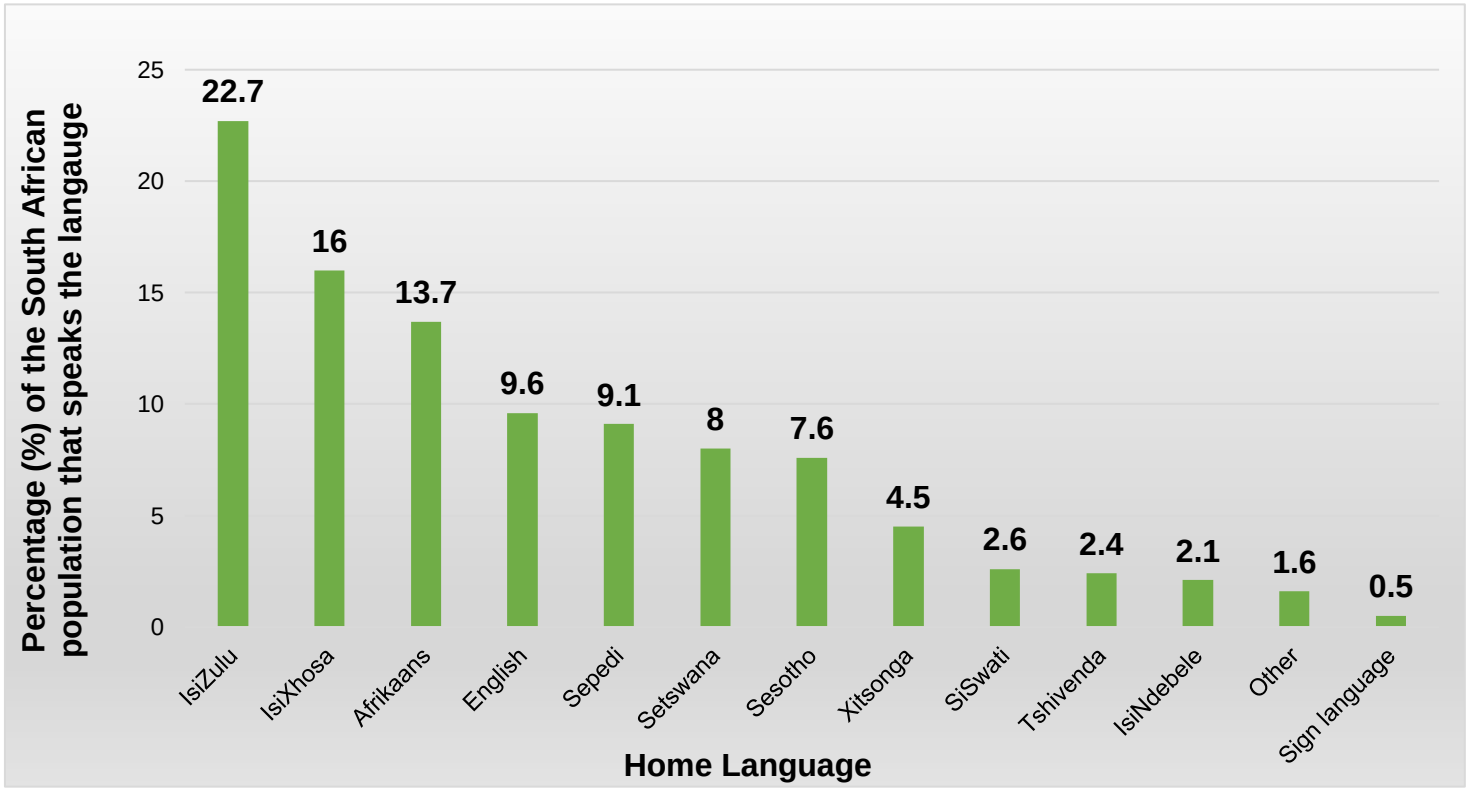


Figure 2.6: *The distribution of the South African population based on home language (Statistics, South Africa, 2012).*

2.12 Crime in South Africa

It could undoubtedly be said that South Africa is one of most violent countries in the world, with a total of 1 919 495 serious crimes being reported during 2019/2020 (South African Police Service, 2020). In the SAPS '*Crime Situation in Republic of South Africa'* report, it is stated that the crime category of sexual offences includes rape, sexual assault, attempted sexual offences, and contact sexual offences (South African Police Service, 2020). During the period from April 2019 to March 2020, the total number of sexual offences reported was 53 293, excluding the sexual assaults detected as a result of police action: a slight increase from the 2018/2019 statistics.

Figure 2.7 shows the number of incidents reported within the category of sexual offences in 2019/2020 (South African Police Service, 2020). Of the 53 293 sexual offences in 2019/2020, 42 289 (79.3%) reports were of rape. This value is also slightly increased from the 2018/2019 statistic, with an average of 117 rapes being reported each day.

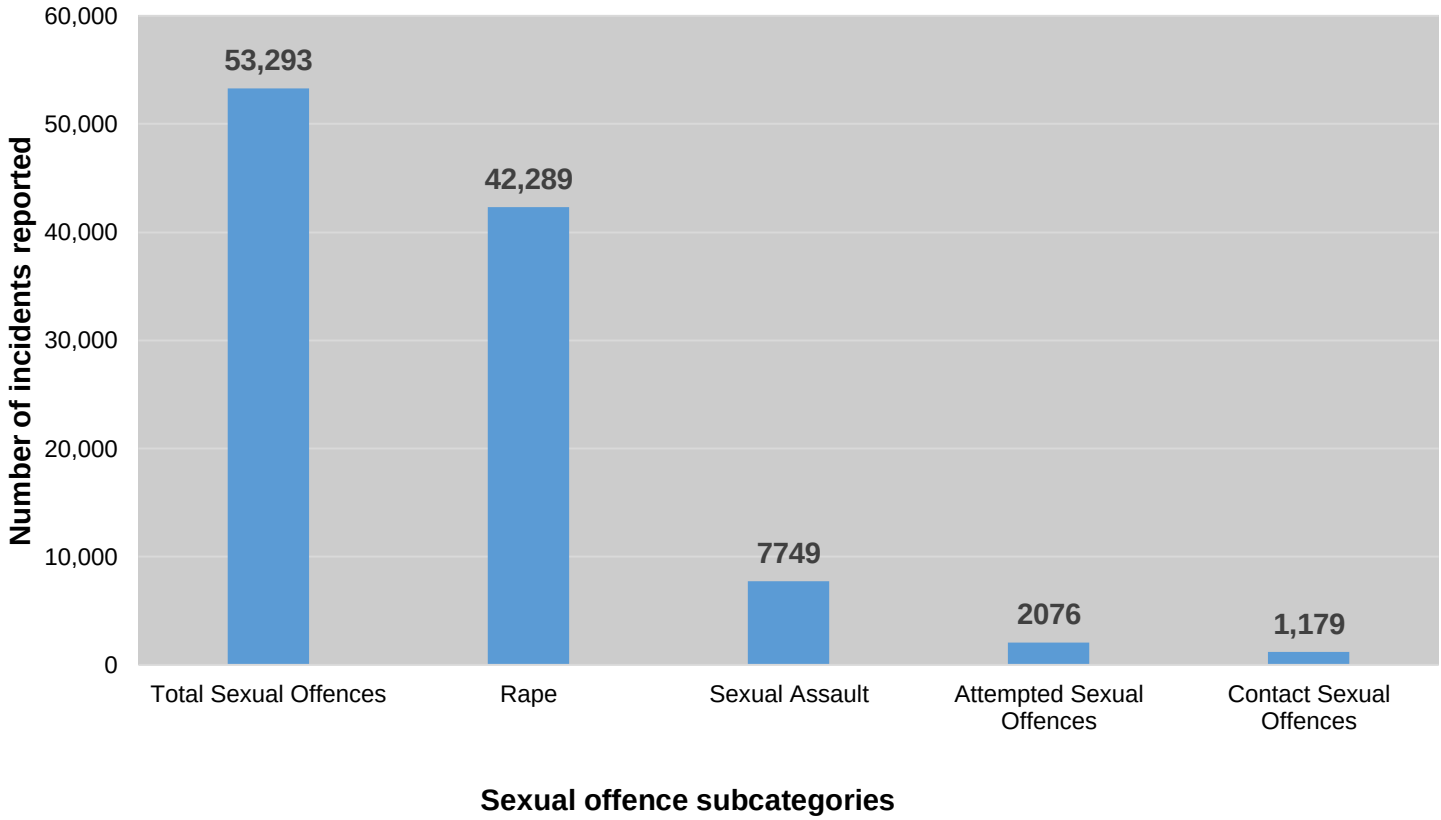


Figure 2.7: *The total number of sexual offences reported to the SAPS during the period from April 2019 to March 2020, as well as the number of incidents within each subcategory (Statistics South Africa, 2020).*

Figure 2.8 depicts the trend of the number of sexual assault cases from 2010 to 2020. It is clear from this figure that there has been a decrease in reported sexual assaults since 2010, with the lowest number of reports in 2016/2017. A reason for this decline could be the advances made in forensic technologies in recent years. Forensic evidence now has more evidentiary value in a court of law, allowing for an increased probability of conviction. Although the number of cases reported in 2019/2020 is considerably lower than that of 2010/2011, this value is has been gradually increasing again since 2016/2017. This is particularly concerning in that these statistics are probably a lot higher in reality, with a large number of sexual assaults going unreported (Rape Crisis, 2018).

There is no doubt, given these statistics, that something needs to be done in an attempt to reduce the number of sexual offences occurring daily in South Africa. From a scientific perspective, adequate forensic facilities, and comprehensive information regarding DNA profiles in the population to provide satisfactory evidence with statistical value, can aid in successfully identifying and convicting sexual assault offenders.

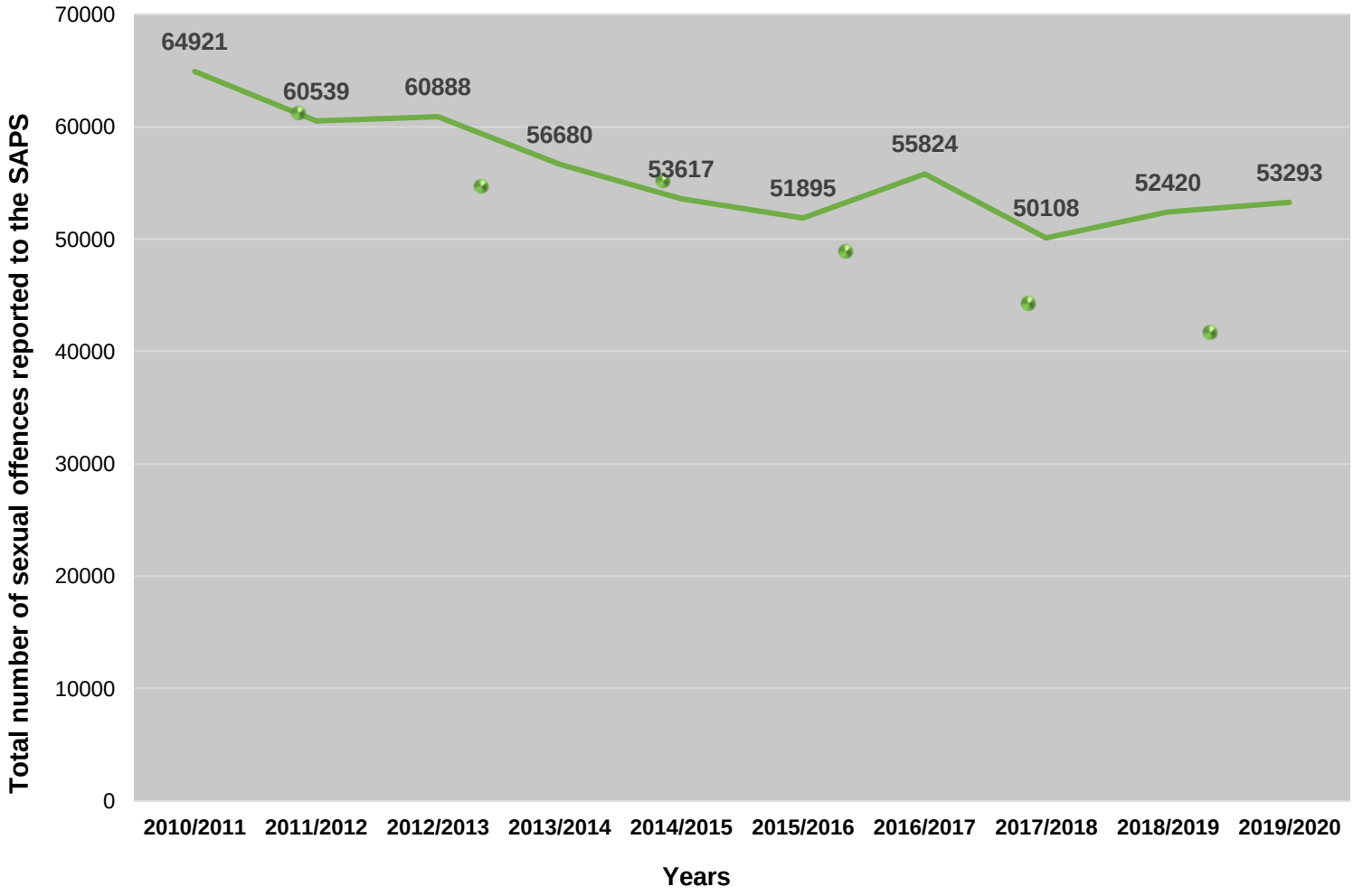


Figure 2.8: *The general trend of the total number of sexual offences reported to the SAPS between 2010 and 2020 (Statistics South Africa, 2020).*

2.13 Use of Y-STRs in South Africa

In contrast to the USA and Europe, Y-STRs are not commonly used in South African forensic casework for a number of reasons (D'Amato *et* al., 2011; D'Amato and Kasu, 2017). Firstly, there is a lack of a comprehensive population reference database; and secondly, some loci included in commercial testing kits are not suitable for use in South Africa's population. Table 2.5 shows the population groups that have been studied to date, as well as the markers that were tested in each population group.

Table 2.5: *The South African population groups that have been investigated using Y-STR markers.*

Population Group	Y-STR markers/kits tested	Reference/s
English Caucasian	MH; UniQ Typer™ Y-10; PowerPlex® Y; Y-Filer®	1; 2; 3; 5; 7; 8
Afrikaans Caucasian	MH; UniQ Typer™ Y-10	4; 8
Coloured	MH; UniQ Typer™ Y-10	4; 8
Asian/Indian	MH; UniQ Typer™ Y-10; PowerPlex® Y; Y-Filer®	2; 3; 5; 7; 8
Cape Muslim	MH and eight additional Y-STR markers*	6
Pedi	UniQ Typer™ Y-10	8
Xhosa	MH; UniQ Typer™ Y-10; PowerPlex® Y; Y-Filer®	1; 2; 3; 5; 7; 8
Venda	UniQ Typer™ Y-10	8
Zulu	UniQ Typer™ Y-10	8

[1](Leat *et al.,* 2004) [2](Ehrenreich, 2005) [3](Leat *et al.*, 2007) [4](Ehrenreich *et al.*, 2008) [5](D'Amato *et al.*, 2009) [6](Cloete *et al.*, 2010) [7](D'Amato *et al.*, 2011) [8](D'Amato and Kasu, 2017)

*The eight additional markers tested in the Cape Muslim population: *DYS449; DYS481; DYS518; DYS557; DYS570; DYS607; DYS612; DYS614*

The study conducted by Leat *et al.,* in 2004 was the beginning of the search for forensically valuable Y-STR markers in South Africa. Leat *et al.* discovered that the MH loci would be useful in the English-speaking Caucasian population for forensic purposes, as high genetic diversity levels were exhibited. Conversely, within the Xhosa population, two loci (*DYS391* and *DYS392*) were found to show relatively low levels of polymorphism. For loci *DYS391* and *DYS392*, 82% and 96% of the Xhosa group shared the same allele respectively. This similarity would suggest that these two loci may not show sufficient genetic variation within some Sub-Saharan population groups to reliably differentiate between male individuals during the course of forensic investigations. In addition to these two loci, a third MH locus—*DYS393*—was found to be lacking in variability within the English, Asian/Indian, and Xhosa population groups, adding evidence to the fact that the MH loci are not suitable for use in Sub-Saharan Africa (Ehrenreich, 2005; Leat *et al.*, 2007). On the other hand, it was found that some of the novel Y-STR markers—*DYS710, DYS711, DYS712, DYS713, and DYS714*—were highly variable. Therefore, South Africa did not necessarily need to rely on the MH loci anymore. In 2007, Leat *et al.* (2007) determined that a core set of eight loci—*DYS518, DYS626, DYS714, DYS449, DYS713, DYS712, DYS570 and DYS458*—were able to successfully resolve haplotypes in the Caucasian population, while the addition of *Y-GATA-A10* and one other marker pair increased resolution in the Xhosa population to an extent. However, the addition of more loci did not improve the resolution fully.

In 2008, a study conducted by Ehrenreich *et al*. focused on the Afrikaans Caucasian and Coloured population groups was conducted using the nine MH loci. Once again, the gene diversity ranged from relatively low to relatively high in both the populations. Locus *DYS391* was again proven to have exceptionally low levels of polymorphism, with 98% of the Afrikaans population carrying either allele 10 or 11: essentially making the locus biallelic. It is worth noting that the Coloured population had a higher discrimination capacity (79%) than the Afrikaans population at 59%. This result would have been expected given the diverse history of the Coloured population. Research conducted by D'Amato *et al.* (2009) was the first to be done that did not include any MH loci. The study focused again on the English Caucasian, Asian/Indian, and Xhosa groups, and introduced a set of 21 Y-STR markers. Some of these markers were proven to be more successful than the MH loci. Once again, the Xhosa group exhibited lower levels of genetic diversity than the other two populations. Despite the Xhosa group having the lowest gene diversity and discrimination capacity out of the three groups, this new set of 21 Y-STR loci yielded significantly better results than the MH. The overall discrimination capacity for the three populations was 95.8%, in contrast to the 77.3% with the MH. These novel Y-STR markers showed significant potential for use in forensic casework in South Africa.

Cloete *et al.* (2010) investigated a group that was known to have migrated to the Cape in South Africa between 1652 and 1834 during the slave trade: the Cape Muslim community. The migratory nature of this group's origins—as well as intermarriage, religious conversions, and mixture with other groups—has led to a complex population structure with a suspected high level of admixture. For this study, the MH loci were used, as well as the core set of eight additional markers described by Leat *et al.* (2007). The minimum gene diversity in this group was not as low as in the previously studied groups. When using only the nine MH loci, the population showed a relatively higher discrimination capacity of 86.6%, although this value increased to 92.4% when considering only the set of eight additional loci. An overall discrimination capacity of 99% validated that these eight markers would indeed be of forensic value as they could considerably increase the power of resolution of the MH loci.

The University of the Western Cape went as far as to develop a new multiplex which showed desirable performance (a high discriminating capacity) in native African populations (D'Amato *et al.,* 2011). This multiplex, named the UniQ Typer™ Y-10, consists of 10 Y-STR loci—*DYS710, DYS518, DYS385, DYS644, DYS612, DYS626, DYS504, DYS481, DYS447* and *DYS449*—two of which come from the set of eight loci by Leat *et al.* (2007) and also includes the controls and standards explicitly made for the kit. Of the 10 loci included in the kit, four are RM markers: *DYS518, DYS612, DYS626, and DYS449*. The study conducted by D'Amato and Kasu in 2017 formed part of an extensive genotyping plan to generate an extensive reference database. Consequently, nearly 1 000 Y-STR profiles were obtained from males from several South African population groups: English, Afrikaans, Asian/Indian, Coloured, Venda, Pedi, Zulu, and Xhosa (D'Amato and Kasu, 2017). The UniQ Typer™ Y-10 kit resulted in an overall discrimination capacity of 90.9%. The two most common Y-haplotypes were each shared between eight males who belonged to either the Xhosa or Zulu groups, emphasising that there is a lack of genetic variation with the Xhosa population group. This lack of genetic variation within certain groups prevents the currently tested commercial kits from effectively distinguishing between male individuals, unrelated and related.

It is evident from the previous research conducted on the South African population that the current knowledge available is simply not sufficient for the reliable use of Y-STR markers in forensic investigations. However, the combination of all 27 Y-STR loci included in ThermoFisher Scientific's Y-Filer® Plus PCR Amplification kit has not yet been investigated within the South African population. Of the 10 Y-STR loci investigated by D'Amato and Kasu (2017) using the UniQ Typer™ Y-10 kit, only four are included in the Y-Filer® Plus kit: *DYS518, DYS385, DYS481, and DYS449*. Therefore, this study investigates 23 different Y-loci in the South African population. Additionally, the Y-Filer® Plus kit includes seven RM loci,

which may prove to have sufficient levels of polymorphism in the population for forensic application. In addition to this, there is no national Y-STR database available for use in forensic casework in South Africa.

The research conducted during this study provides information on additional Y-STR loci and their viability for use in forensic investigations in South Africa, with the use of the Y-Filer® Plus testing kit. Given the crime statistics and high prevalence of rape in this country, this kind of research could be highly beneficial and conducive to reducing these statistics. Should these Y-STR loci prove to have a higher discriminating power in all the South African population groups than the previously tested commercial kits, the SAPS laboratories may consider making use of this kit during rape investigations. The statistical analyses performed in this study would also provide haplotype frequency data that could be used in a population reference database. A combination of novel Y-STR loci and a comprehensive reference database could aid in identifying, and consequently convicting, rapists in South Africa and reducing the sexual assault statistics. Rape and sexual assault are crimes that affect every individual in South Africa, in some way or another, so the research conducted during this study has the potential to benefit the whole South African population.

References

Baeta, M., Núñez, C., Villaescusa, P., Ortueta, U., Ibarbia, N., Herrera, R.J., Blazquez-Caeiro, J.L., Builes, J.J., Jiménez-Moreno, S., Martínez-Jarreta, B., and de Pancorbo, M.M. (2018) 'Assessment of a subset of Slowly Mutating Y-STRs for forensic and evolutionary studies'. *Forensic Science International: Genetics*, 34, pp.e7-e12.

Ballantyne, J. (2012) '*Current Status and Future of Y-STR Analysis*'. Presentation, Henry B. González Convention Center, San Antonio, Texas, USA.

Ballantyne, K.N., Goedbloed, M., Fang, R., Schaap, O., Lao, O., Wollstein, A., Choi, Y., van Duijn, K., Vermeulen, M., Brauer, S., and Decorte, R. (2010) 'Mutability of Y-chromosomal microsatellites: rates, characteristics, molecular bases, and forensic implications'. *The American Journal of Human Genetics*, 87(3), pp.341-353.

Ballantyne, K.N., Keerl, V., Wollstein, A., Choi, Y., Zuniga, S.B., Ralf, A., Vermeulen, M., de Knijff, P., and Kayser, M. (2012) 'A new future of forensic Y-chromosome analysis: rapidly mutating Y-STRs for differentiating male relatives and paternal lineages'. *Forensic Science International: Genetics*, 6(2), pp.208-218.

Bruijns, B., Tiggelaar, R., and Gardeniers, H. (2018) 'Massively parallel sequencing techniques for forensics: A review'. *Electrophoresis*, 39(21), pp.2642-2654.

Budowle, B. and Van Daal, A. (2008) 'Forensically relevant SNP classes'. *Biotechniques*, 44(5), pp.603-610.

Buhler, E. (1980) 'A synopsis of the human Y Chromosome'. *Human Genetics*, 55(2), pp.145-175.

Butler, J. M. (2003) 'Recent developments in Y-single tandem repeat and Y-single nucleotide polymorphism analysis'. *Forensic Science Review*, 15(91), pp.91-114.

Butler, J.M., Hill, C.R., Decker, A.E., Kline, M.C., Reid, T.M., and Vallone, P.M., (2007) 'New autosomal and Y-chromosome STR loci: characterization and potential uses'. *Proceedings of the Eighteenth International Symposium on Human Identification*. Promega Corporation Madison, WI.

Chakraborty, R. and Kidd, K. K. (1991) 'The utility of DNA typing in forensic work'. *Science*, 254(5039), pp.1735-1739.

Claerhout, S., Vandenbosch, M., Nivelle, K., Gruyters, L., Peeters, A., Larmuseau, M.H., and Decorte, R. (2018) 'Determining Y-STR mutation rates in deep-routing genealogies: Identification of haplogroup differences'. *Forensic Science International: Genetics*, 34, pp.1-10.

Cloete, K., Ehrenreich, L., D'Amato, M. E., Leat, N., Davison, S., and Benjeddou, M. (2010) 'Analysis of seventeen Y-chromosome STR loci in the Cape Muslim population of South Africa'. *Legal Medicine*, 12(1), pp.42-45.

D'Amato, M.E., Bajic, V.B. and Davison, S. (2011) 'Design and validation of a highly discriminatory 10-locus Y-chromosome STR multiplex system'. *Forensic Science International: Genetics*, 5(2), pp.122-125.

D'Amato, M. E., Benjeddou, M., and Davison, S. (2009) 'Evaluation of 21 Y-STRs for population and forensic studies'. *Forensic Science International: Genetics Supplement Series*, 2, pp.446-447.

D'Amato, M.E., Ehrenreich, L., Benjeddou, M., Davison, S. and Leat, N. (2008) 'Ancestry and genetic relationships between groups within the Cape Town metropolitan population inferred using Y-STRs genotyping'. *Forensic Science International: Genetics Supplement Series*, 1(1), pp.318-319.

D'Amato, M. E., and Kasu, M. (2017) 'Population analysis of African Y-STR profiles with UniQ Typer™ Y-10 genotyping system'. *Forensic Science International: Genetics Supplement Series*, 6(October), pp.e84-e85.

De Knijff, P. (2019) 'From next generation sequencing to now generation sequencing in forensics'. *Forensic Science International: Genetics*, 38, pp.175-180.

Ehrenreich, L. S. (2005) '*The evaluation of Y-STR loci for use in Forensics'*. Masters. University of the Western Cape.

Ehrenreich, L., Benjeddou, M., Davison, S., D'Amato, M., and Leat, N. (2008) 'Nine-locus Y-STR profiles of Afrikaner Caucasian and mixed ancestry populations from Cape Town, South Africa'. *Legal Medicine*, 10(4), pp.225-227.

Evans, C. (1996) '*The Casebook of Forensic Detection: How Science Solved 100 of the World's Most Baffling Crimes*'. New York: John Wiley & Sons Inc.

Eckert, D. A. (1997) '*Introduction the forensic science*'. Florida: CRC Press.

Foreman, L. and Evett, I. (2001) 'Statistical analyses to support forensic interpretation for a new ten-locus STR profiling system'. *International Journal of Legal Medicine*, 114(3), pp.147-155.

Gill, P., Brenner, C., Brinkmann, B., Budowle, B., Carracedo, A., Jobling, M. A., de Knijff, P., Kayser, M., Krawczak, M., Mayr, W. R., Morling, N., Olaisen, B., Pascali, V., Prinz, M., Roewer, L., Schneider, P. M., Sajantila, A., and Tyler-smith, C. (2001) 'DNA Commission of the International Society of Forensic Genetics: recommendations on forensic analysis using Y-chromosome STRs'. *Forensic Science International*, 124, pp.5-10.

Goedbloed, M., Vermeulen, M., Fang, R.N., Lembring, M., Wollstein, A., Ballantyne, K., Lao, O., Brauer, S., Krüger, C., Roewer, L., and Lessig, R. (2009) 'Comprehensive mutation analysis of 17 Y-chromosomal short tandem repeat polymorphisms included in the AmpFlSTR® Yfiler® PCR amplification kit'. *International Journal of Legal Medicine*, 123(6), p.471.

Goodur, P. (2018) '*The Y-filer® Plus PCR Amplification Kit'*. Presentation, ThermoFisher Scientific, Johannesburg, South Africa.

Gordon, L. P (2002) '*Violence Against Women'*. Nova Publishers. pp.4-6.

Gusmão, L. and Carracedo, A. (2003) 'Y chromosome-specific STRs'. *Profiles in DNA*, 6(1), pp.3-6.

Gusmão, L., Brion, M., Gonzjalez-Neira A., Lareu, M., and Carracedo, A. (1999) 'Y Chromosome specific polymorphisms in forensic analysis'. *Legal Medicine*, 1, pp.55-60.

Gusmão, L., Butler, J. M., Carracedo, A., Gill, P., Kayser, M., Mayr, W. R., Morling, N., Prinz, M., Roewer, L., Tyler-smith, C., and Schneider, P. M. (2006) 'DNA Commission of the International Society of Forensic Genetics (ISFG): An update of the recommendations on the use of Y-STRs in forensic analysis'. *Forensic Science International*, 157, pp.187-197.

Gusmão, L., Brión, M. and Gomes, I. (2008) 'The human Y chromosome male-specific polymorphisms and forensic genetics'. *Handbook of Analytical Separations*, 6, pp.969-1000.

Haas, C., Shved, N., Rühli, F.J., Papageorgopoulou, C., Purps, J., Geppert, M., Willuweit, S., Roewer, L., and Krawczak, M. (2013) 'Y-chromosomal analysis identifies the skeletal remains of Swiss national hero Jörg Jenatsch (1596–1639)'. *Forensic Science International: Genetics*, 7(6), pp.610-617.

Hall, A., and Ballantyne, J. (2003) 'Novel Y-STR typing strategies reveal the genetic profile of the semen donor in extended interval post-coital cervicovaginal samples'. *Forensic Science International*, 136, pp.58-72.

Hammer, M., and Redd, A.J. (2016) 'Forensic Applications of Y chromosome STRs and SNPs'. *Forensic Science International*. pp. 97-111.

Hu, N., Cong, B., Li, S., Ma, C., Fu, L., And Zhang, X. (2014) 'Current developments in forensic interpretation of mixed DNA samples (Review)'. *Biomedical Reports*, 2(3), pp.309-316.

Jobling, M. (2001) 'Y-chromosomal SNP haplotype diversity in forensic analysis'. *Forensic Science International*, 118. pp.158-162

Kayser, M. (2017) 'Forensic use of Y-chromosome DNA: a general overview'. *Human Genetics*, 136(5), pp.621-635.

Kayser, M., Caglia, A., Corach, D., Fretwell, N., Gehrig, C., Graziosi, G., Heidorn, F., Herrmann, S., Herzog, B., Hidding, M., Honda, K., Jobling, M., Krawczak, M., Leim, K., Meuser, S., Meyer, E., Oesterreich, W., Pandya, A., Parson, W., Penacino, G., Perez-Lezaun, A., Piccinini, A., Prinz, M., Schmitt, C., Schneider, P. M., Szibor, R., Teifel-Greding, J., Weichhold, G., de Knijff, P., and Roewer, L. (1997) 'Evaluation of Y-

chromosomal STRs: A multicenter study'. *International Journal of Legal Medicine*, 110(February 2014), pp.125-133.

Kayser, M. and de Knijff, P (2011) 'Improving human forensics through advances in genetics, genomics and molecular biology'. *Nature reviews, Genetics*, 12(3), pp.179-92.

Kayser, M., Kittler, R., Erler, A., Hedman, M., Lee, A.C., Mohyuddin, A., Mehdi, S.Q., Rosser, Z., Stoneking, M., Jobling, M.A., and Sajantila, A. (2004). 'A comprehensive survey of human Y-chromosomal microsatellites'. *The American Journal of Human Genetics*, 74(6), pp.1183-1197.

Kayser, M. and Schneider, P.M. (2009) 'DNA-based prediction of human externally visible characteristics in forensics: motivations, scientific challenges, and ethical considerations'. *Forensic Science International: Genetics,* 3(3), pp.154-161.

Krenke, B.E., Fulmer, P.M., Miller, K.D., and Sprecher, C.J. (2003) 'The PowerPlex® Y System'. *Profiles DNA*, 6(2), pp.6-9.

Leat, N., Benjeddou, M., and Davison, S. (2004) 'Nine-locus Y-chromosome STR profiling of Caucasian and Xhosa populations from Cape Town, South Africa'. *Forensic Science International*, 144(1), pp.73-75.

Leat, N., Ehrenreich, L., Benjeddou, M., Cloete, K., and Davison, S. (2007) 'Properties of novel and widely studied Y-STR loci in three South African populations'. *Forensic Science International*, 168(2–3), pp.154-161.

Lee, H. C., and Ladd, C. (2001) 'Preservation and Collection of Biological Evidence'. *Croatian Medical Journal*, 42(3), pp.225-228.

Lessig, R., Zoledziewska, M., Fahr, K., Edelmann, J., Kostrzewa, M., Dobosz, T., and Kleemann, W.J. (2005) 'Y-SNP-genotyping–a new approach in forensic analysis'. *Forensic Science International*, 154(2-3), pp.128-136.

Martin, P., Albarran, C., Garcia, O., Garcia, P., Sancho M., and Alonso, A. (1999) 'Application of Y-STR analysis to rape cases that cannot be solved by autosomal STR analysis'. *Progress in Forensic Genetics 8*. eds. Sensabaugh, G. F., Lincoln, P. J., and Olaisen, B. New York, USA: Excerpta Medica. pp.526-528.

McKie, R. (2009) '*Eureka moment that led to the discovery of DNA fingerprinting'*. The Guardian. [Online] Available at https://www.theguardian.com/science/2009/may/24/dna-fingerprinting-alec-jeffreys. [Accessed 15 Feb. 2019]

Moxon, E. R., and Wills, C. (1999) 'DNA microsatellites: agents of evolution?'. *Scientific American*, 280(1), pp.94-99.

O'Hara, C. E., and O'Hara, G. L. (1994) '*Fundamentals of Criminal Investigation*'. 6th edn.

Qian, X., Hou, J., Wang, Z., Ye, Y., Lang, M., Gao, T., Liu, J., and Hou, Y. (2017) 'Next generation sequencing plus (NGS+) with Y-chromosomal markers for forensic pedigree searches'. *Scientific Reports*, 7(1), pp.11324.

Quintana-Murci, L. and Fellous, M. (2001) 'The human Y chromosome: the biological role of a "functional wasteland"'. *BioMed Research International*, 1(1), pp.18-24.

Rape Crisis. (2018) '*The real numbers on sexual offences*'. [Online]. Available at: https://rapecrisis.org.za/the-real-numbers-on-sexual-offence/. [Accessed 11 Feb. 2019]

Redd, A.J., Agellona, A.B., Kearneya, V.A., Contrerasa, V.A., Karafeta, T., Park, H., de Knijff, P., John. J.M., and Hammer, M.F. (2002) 'Forensic value of 14 novel STRs on the Human Y chromosome'. *Forensic Science International*, 130, pp.97-111.

Roewer, L. (2009) 'Y chromosome STR typing in crime casework'. *Forensic Science, Medicine, and Pathology*, 5, pp.77-84.

Roewer, L. (2013) 'DNA fingerprinting in forensics: past, present, future'. *Investigative genetics*, 4 (1), p.22.

Roewer, L., Krawczak, M., Willuweit, S., Nagy, M., Alves, C., Amorim, A., Anslinger, K., Augustin, C., Betz, A., Bosch, E., Caglia, A., Carracedo, A., Corach, D., Dekairelle, A. F., Dobosz, T., Duput, B. M., Füredi, S., Gehrig, C., Gusmão, L., Henke, J., Henke, L., Hidding, M., Hohoff, C., Hoste, B., Jobling, M. A., Kargel, H. J., de Knijff, P., Lessig, R., Liebeherr, E., Lorente, M., Martinez-Jarreta, B., Nievas, P., Nowak, M., Parson, W., Pascali, V. L., Penacino, G., Ploski, R., Rolf, B., Sala, A., Schmidt, U., Schmitt, C., Schneider, P. M., Szibor, R., Teifel-Greding, J., and Kayser, M. (2001) 'Online reference

database of European Y-chromosomal short tandem repeat (STR) haplotypes'. *Forensic Science International*, 118, pp.106-113.

Roewer, L. and Willuweit, S. (2020) '*YHRD: Database Statistics'*. YHRD - Y chromosome STR haplotype reference database. {Online]. Available at: https://yhrd.org/pages/resources/stats#haplotype_counts. [Accessed 09 Oct. 2020].

Shewale, J.G. (2003) 'Y-Short Tandem Repeat Multiplex Systems-Y-PLEX™ 6 and Y-PLEX™ 5'. *Forensic science review*, 15(2), pp.115-135.

Shewale, J. G., Nasir, H., Schneida, E., Gross, A. M., Budowle, B., and Sinha, S. K. (2004) 'Y-Chromosome STR System , Y-PLEX™ 12 , for Forensic Casework: Development and Validation'. *Journal of Forensic Sciences*, 49(6), pp.1-13.

Sousa, A.S.A. (2017) '*Estimation of mutation rates at Y-STRs'*. Masters. University of Porto.

South African Police Service. (2020) '*Crime Situation in Republic of South Africa Twelve (12) Months (April To March 2019_20)'*. [Online]. Available at: www.saps.gov.za/services/crimestats.php. [Accessed 06 Aug. 2020].

Statistics South Africa. (2012) '*Census 2011 – Census in Brief*'. [Online] Available at: http://www.statssa.gov.za/publications/P03014/P030142011.pdf. [Accessed 05 Feb. 2019].

Statistics South Africa. (2018) '*Crime against Women in South Africa - An in-depth analysis of the Victims of Crime Survey data*'. [Online]. Available at: http://www.statssa.gov.za/publications/Report-03-40-05/Report-03-40-05June2018.pdf. [Accessed 05 Feb. 2019].

Statistics South Africa. (2020) '*Mid-year population estimates 2020*'. [Online]. Available at: http://www.statssa.gov.za/publications/P0302/P03022020.pdf. [Accessed 06 Aug. 2020].

Tamaki, K. and Jeffreys, A.J. (2005) 'Human tandem repeat sequences in forensic DNA typing'. *Legal Medicine*, 7(4), pp.244-250.

ThermoFisher Scientific. (2006) *AmpFlSTR® Yfiler® PCR Amplification Kit: User's Manual*. Foster City: Applied Biosystems. [Online]. Available at: https://www.thermofisher.com/content/dam/LifeTech/Documents/PDFs/Y-STR/yfiler-users-manual.pdf. [Accessed 25 Jan. 2019]

ThermoFisher Scientific. (2019a) *GlobalFiler™ and GlobalFiler™ IQC PCR Amplification Kits: User Guide*. Woolston: ThermoFisher Scientific. [Online]. Available at: https://assets.thermofisher.com/TFS-Assets/LSG/manuals/4477604.pdf. [Accessed 25 Jan. 2019]

ThermoFisher Scientific. (2019b) *Yfiler™ Plus PCR Amplification Kit: User Guide*. Woolston: ThermoFisher Scientific. [Online]. Available at: https://assets.thermofisher.com/TFS-Assets/LSG/manuals/4485610_YfilerPlus_UG.pdf. [Accessed 30 Jan. 2019]

Thompson, J.M., Ewing, M.M., Rabbach, D.R., Fulmer, P.M., Sprecher, C.J., and Storts, D.R. (2012) 'The PowerPlex® Y23 System: A new Y-STR multiplex for casework and database applications'. *Profiles in DNA 2012*.

Wallace, H. (2006) 'The UK National DNA Database: Balancing crime detection, human rights and privacy'. *EMBO reports*, 7(1S), pp.S26-S30.

Wambaugh, J. (1990) '*The Blooding: True Story of the Narborough Village Murders'*. Bantam Books.

Warshauer, D.H., Churchill, J.D., Novroski, N., King, J.L. and Budowle, B. (2015) 'Novel Y-chromosome short tandem repeat variants detected through the use of massively parallel sequencing'. *Genomics, Proteomics & Bioinformatics*, 13(4), pp.250-257.

Willuweit, S. and Roewer, L. (2015) 'The new Y chromosome haplotype reference database'. *Forensic Science International: Genetics*, 15, pp.43-48.

www.usystrdatabase.org

www.yhrd.org

CHAPTER 3: FORENSIC GENETIC VALUE OF 27 Y-STR LOCI (Y-FILER® PLUS) IN THE SOUTH AFRICAN POPULATION

3.1 Introduction

Y-STRs have proven to be valuable in the DNA analysis of biological evidence to overcome challenges encountered when using autosomal STRs. Y-STRs overcome the issue of overrepresented female DNA potentially preventing the male DNA from being detected during the analysis of crime scene evidence (Roewer, 2009). The ability to differentiate between male individuals, related or unrelated, using Y-STR loci is measured using the discrimination capacity (Redd *et al.,* 2002). The goal for selecting Y-STR loci for use in forensic analyses is, therefore, to achieve a discrimination capacity as close to 100% as possible.

Y-STRs are currently not utilised in forensic laboratories in South Africa due to the low discrimination capacity of commercially available PCR kits and the lack of population reference data (D'Amato *et al.,* 2011; D'Amato and Kasu, 2017). South Africa's population consists of many ethnic groups, with a significant amount of admixture between these groups, and a limited number of Y-STR markers have been evaluated for use in these different groups. The full combination of 27 Y-STR markers included in ThermoFisher Scientific's Y-Filer® Plus kit has not yet been comprehensively tested in the South African population and there is no published population data available.

Research into forensically valuable Y-STR markers in South Africa began in 2004 with Leat *et al.* conducting a study using the minimum haplotype (MH) loci. Leat *et al.* discovered that while the English-speaking Caucasian population exhibited high diversity, some loci (*DYS391* and *DYS392*) showed low diversity in the Xhosa population. It was also discovered that *DYS391* has low levels of polymorphism in the Afrikaans Caucasian population (Ehrenreich *et al.*, 2008). Additional studies showed that a third MH locus (*DYS393*) also lacked diversity in the English, Asian/Indian, and Xhosa populations (Ehrenreich, 2005; Leat *et al.*, 2007). D'Amato *et al.* (2009) introduced a set of 21 Y-STR markers which did not include any MH loci. The study undertaken by D'Amato *et al.* (2009) showed that, by increasing the number of markers used, the discrimination capacity can also be significantly increased. Cloete's 2010 study in showed that high levels of admixture in a population has the potential to increase the genetic diversity with a population and thus increase the discriminating power of a set of markers in that population. Although Cloete's study focused on the Cape Muslim population, the same hypothesis could be applied to the Coloured population.

D'Amato and Kasu (2017) developed a new genotyping system in an attempt to provide a combination of Y-STRs that could reliably discriminate between individuals in Sub-Saharan African populations. The UniQ Typer™ Y-10 multiplex included 10 Y-STR, of which four are

rapidly mutating. Despite only using 10 markers, this multiplex had desirable results—with the overall discrimination capacity increasing significantly—and indicated that the inclusion of rapidly (RM) loci may be the key to finding forensically valuable loci for use in South Africa. D'Amato and Kasu's study did, however, report that the Xhosa and Zulu groups still showed lower levels of genetic variation, even with the inclusion of the RM loci. It is this lower genetic diversity of some markers included in commercial kits that present a challenge in the forensic application of Y-STRs in South Africa.

While the current research conducted demonstrates the potential that Y-STRs have in South Africa, it is still evident that more loci need to be considered. Of the 27 Y-STR loci included in the Y-Filer® Plus kit, only five of those markers—*DYS438, DYS460, DYS533, DYS627,* and *DYF387S1a/b*—are yet to be investigated within the South African population. However, this specific combination of markers has not yet been comprehensively studied, nor has a study been conducted to date using this number of markers in a single multiplex. With the increased number of markers included and the addition of seven RM loci in the Y-Filer® Plus kit, it is expected that this multiplex could significantly increase the discrimination capacity in the South African population.

There is undoubtably a need for the implementation of Y-STRs in the SAPS forensic laboratories, given the alarmingly high rape statistics in this country. Bernitz *et al.* (2015) reports that approximately 70% of cases received by the SAPS each year are sexual assault cases. A total of 53 293 sexual offences were reported in 2019/2020 and 79% of those offences committed were rape (South African Police Service, 2020). While autosomal STRs certainly have a significant success rate in the analysis of DNA evidence from such cases, the addition of Y-STRs in DNA analysis has the potential to further increase the success rate of DNA evidence.

This chapter aims to evaluate the forensic value of the specific combination of the 27 Y-STR loci in ThermoFisher Scientific's Y-Filer® Plus PCR amplification and consider its potential in the SAPS forensic laboratories. Furthermore, the chapter will provide haplotype frequency data that could contribute towards building a comprehensive population reference database. A combination of Y-STR loci that results in increased resolution between haplotypes and additional population data could certainly be beneficial for South Africa's forensic community.

3.2 Materials and Methods

3.2.1 Sampling

Ethical clearance for this research was obtained through the General Human Research Ethics Committee at the University of the Free State (UFS-HSD-2018/0779/2908; see Appendix A on Page 109). Sampling was conducted on the Bloemfontein campus at the University of the Free State. Participants were approached and the study explained to them in detail using an information sheet, allowing them to give informed consent to participate. The information sheet and informed consent form are attached in Appendix B on Page 111. Male individuals who consented to participating were provided with two cotton buccal swabs to take themselves, as to be as non-invasive as possible and to ensure that the individuals were comfortable during the sampling process. They were instructed and shown to take the swab by rubbing it up and down on the inside of the cheek for five to ten seconds. The participants were instructed to use one swab on the inside of one cheek and the duplicate swab on the other. The swabs were then placed back into the tubes and stored at 4°C until processed. A unique code, specified according to the population group of the participant, was designated to each tube to ensure anonymity with downstream processing. Efforts were made to ensure that the participants were not eating or drinking at the time of sampling in order to reduce the potential of collecting food remnants that may inhibit PCR downstream.

In addition to providing a buccal swab, the participants also filled in a mini questionnaire (Appendix B, Page 111) to provide information regarding their population group and home language for both themselves and their parents. Each of the four main population groups of Asian/Indian, African, Coloured, and Caucasian were subdivided further using the home language information provided by the participants. Coloured and Caucasian individuals were asked to indicate if they are English or Afrikaans. African individuals were asked to specify their ethnic group. Some Asian/Indian individuals indicated that they were not born in South Africa, so this group was divided into South African Indians and Asian Indians (i.e. those who were not born in South Africa). Efforts were made to ensure, as far as possible, that all the individuals that participated in the research were not related. This was done to prevent potential bias in the statistical analysis as including related males is likely to under- or overestimate some results leading to false conclusions.

Chakraborty (1992) suggested that 100 to 150 individuals per population would be sufficient to conservatively estimate allele and genotype or haplotype frequencies at tandem repeat loci. Therefore, the intended number of samples per population group was 100 males, or 400

samples in total. Overall, 308 buccal swab samples were collected. A total of 13 Asian/Indian, 132 African, 44 Coloured, and 119 Caucasian samples were collected.

3.2.2 Direct Amplification

The buccal swabs collected were processed using direct amplification, eliminating the DNA extraction stage. The swabs were first lysed using 200 µl Prep 'n Go buffer (*Applied Biosystems*) for 20 minutes at room temperature (modified from ThermoFisher Scientific, 2019). The lysates were transferred to new Eppendorf tubes, and the swab heads discarded. The PCR reactions were set up according to the Y-Filer® Plus Amplification Kit user manual (ThermoFisher Scientific, 2019), although optimised for a half-reaction with a final volume of 13.5 µl containing 5 µl MasterMix, 5 µl low TE buffer, 2.5 µl Primer Set, and 1 µl sample lysate.

Amplification was performed using a Veriti™ 96-Well Thermal Cycler (*Applied Biosystems*), making use of the 9600 Emulation Mode. The amplification conditions were as follows: an initial incubation of 95°C for 1 minute; 29 cycles of denaturation at 94°C for 4 seconds and annealing/extension at 61.5°C; and a final extension at 60°C for 22 minutes.

3.2.3 Detection and Genotyping

After amplification, capillary electrophoresis (CE) was carried out on the PCR products using the *Applied Biosystem's* 3500 Genetic Analyser. The PCR products were prepared for CE by adding 1 µl PCR product to a mixture of 9.6 µl Hi-Di™ Formamide and 0.4 µl GeneScan™ 600 LIZ® Size Standard (*Applied Biosystems*). Before being loaded on the genetic analyser, the reaction plate was heat denatured at 95°C for 10 minutes and then immediately placed on ice for 3 minutes. The reaction plate was then loaded onto the 3500 Genetic Analyser and CE was performed using a 36 cm capillary and POP4-polymer (*Applied Biosystems*). The run conditions for CE were set up according to the Y-Filer® Plus Amplification Kit user manual (ThermoFisher Scientific, 2019). For quality control purposes, positive (DNA Control 007) and negative controls were run with each batch of samples, and the Yfiler™ Plus Allelic Ladder was included for every 23 samples run. The result profiles of the controls can be found in Appendix C on Page 116.

Upon completion of CE, the DNA profiles were generated and analysed using the GeneMapper™ ID-X Software v1.5 (*Applied Biosystems*). A peak threshold of 100 relative fluorescence units (RFU) was used to call the alleles in the profiles. A few samples that had

detected off-ladder (OL) peaks were reamplified using the duplicate swab to confirm the allele, and then those OL alleles were calculated as described by Butler (2005; see Appendix F, Page 140). Samples that had little or no amplification were excluded from further analysis.

3.2.4 Statistical Analysis

The samples that generated full profiles were used to calculate summary statistics. Allele and haplotype frequencies were calculated using the direct counting method (ThermoFisher Scientific, 2019). The number of unique haplotypes was determined: a unique haplotype being one that was observed only once in the population (ThermoFisher Scientific, 2019). The discrimination capacity (DC) was calculated as the number of different haplotypes divided by the total number of samples (Redd *et al.,* 2002). The Match probability (MP) was calculated as the sum of the squared haplotype frequencies. Haplotype diversity (Hd) was calculated using $Hd = \frac{N\,(1-\Sigma pi^2)}{N-1}$, where N and pi are the sample size and haplotype frequency respectively (Nei and Tajima, 1981). The gene diversity (GD) at each locus was calculated using $GD = \frac{N\,(1-\Sigma pi^2)}{N-1}$, where N and pi are the sample size and allele frequency, respectively (Johnson *et al.,* 2003). These statistics were calculated for the four population groups of Asian/Indian, African, Coloured, and Caucasian, as well as for the overall South African population. All statistics, excluding MP, were also calculated for the Afrikaans and English subgroups for the Caucasian population and Tswana, Xhosa, Zulu, Sotho, and Pedi subgroups of the African population group. This subdivision was done in order to evaluate the discriminating capabilities of the testing kit within these two population groups and to make accurate comparisons to previous research conducted on these specific subgroups.

3.3 Results and Discussion

3.3.1 Sampling

Overall, 308 buccal swab samples were collected. A total of 13 Asian/Indian, 132 African, 44 Coloured, and 119 Caucasian samples were collected.

Table 3.1 presents the racial distribution of males in South Africa and at the University of the Free State in 2020, as well as the number of samples collected for this study. Figure 3.1 further shows the distribution of the samples collected across the four population groups for this study.

Table 3.1: *The distribution of males in the four population groups registered at the University of the Free State in 2020 (University of the Free State, 2020), in comparison to the male population of South Africa based on the 2020 estimates (Statistics South Africa, 2020), as well as the number of samples collected.*

Population Group	University of the Free State		South Africa		Samples Collected	
	Number of males	% of male population	Number of males	% of male population	Number of samples	% of sample size
Asian/Indian	205	1.32	787 662	2.70	13	4.00
African	12 418	79.87	23 519 474	80.70	132	43.00
Coloured	689	4.43	2 555 204	8.80	44	14.00
Caucasian	2 236	14.48	2 266 535	7.80	119	39.00
Total	15 548	100	29 128 875	100	308	100

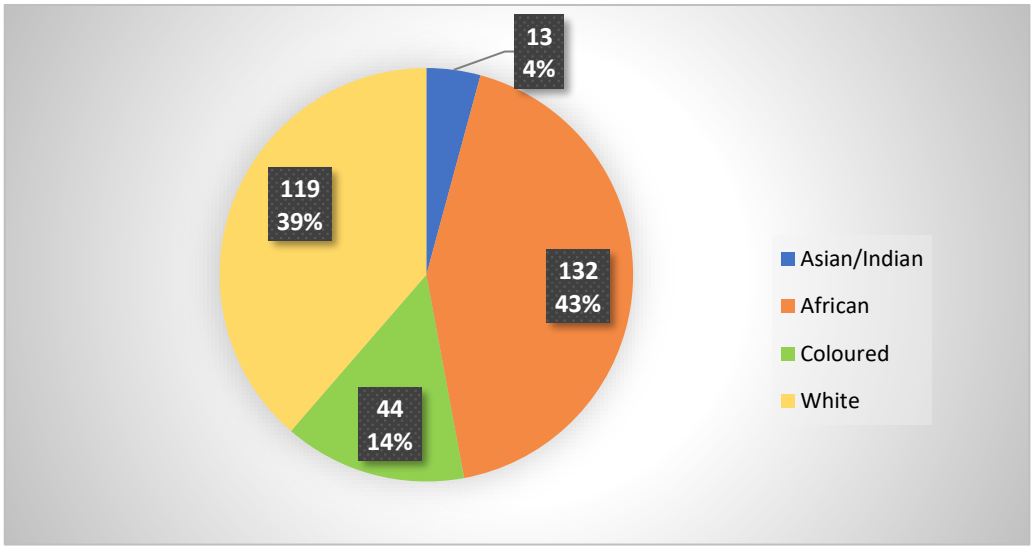


Figure 3.1: *The total number of samples collected per population group and the percentage of each group in the total number of samples.*

3.3.2 Direct Amplification, Detection, Genotyping, and Sample Exclusion

Figure 3.2 is an example of a standard full profile that was generated after sample processing. A standard Y-STR profile generally has only one peak per locus (with the exception of the markers *DYS385* and *DYF387S1*), peak heights above the threshold of 100 RFU, and exhibits both intra- and interlocus balance. Of the 308 samples processed in total, 271 samples produced full profiles that met the abovementioned criteria of a standard profile, and were thus included for further statistical analysis.

Figure 3.3 is indicative of a sample that was excluded from analysis as it only generated a partial profile, most likely due to a poor quality sample. In this specific sample, three markers—*DYS448, DYS393*, and *DYS533*—exhibited peaks that were too low to be detected, and other markers such as *DYS481* and *DYS19* showed increased stochastic amplification with additional peaks being detected. Several other markers were also flagged for low peak heights, indicating a lack of interlocus balance across the profile. After profile analysis, 37 samples were excluded from further analysis based on the abovementioned criteria.

After sample exclusion, a total of 271 samples (or DNA profiles) were used in statistical analysis. Overall, 12 DNA profiles were obtained for the Asian/Indian population, 113 for the African population, 43 for the Coloured population, and 103 for the Caucasian population. Excluding the 37 samples does not have a significant effect on the percentage contribution of each group to the overall sample set, and the numbers for each population still reflect the demographics of South Africa and the University of the Free State (Table 3.1). Despite having to exclude some profiles, the number of samples for the African and Caucasian populations were still above the required 100 samples for accurate allele frequency estimations (Chakraborty, 1992). The lower number of samples for the Asian/Indian and Coloured populations was taken into consideration when interpreting the statistical analyses.

The African and Caucasian population groups were further divided into several subgroups to evaluate the discrimination capacity within these two groups, and to make accurate comparisons to previous research conducted using these subgroups. The African population is comprised of many different ethnic groups, classified according to language (Lane *et al*., 2002). The Nguni group includes the Zulu, Xhosa, Tsonga, Mpondo, Ndebele, and Swati subgroups. The Sotho/Tswana group consists of Sotho, Pedi, and Tswana individuals. The third subgroup consists only of Venda individuals.

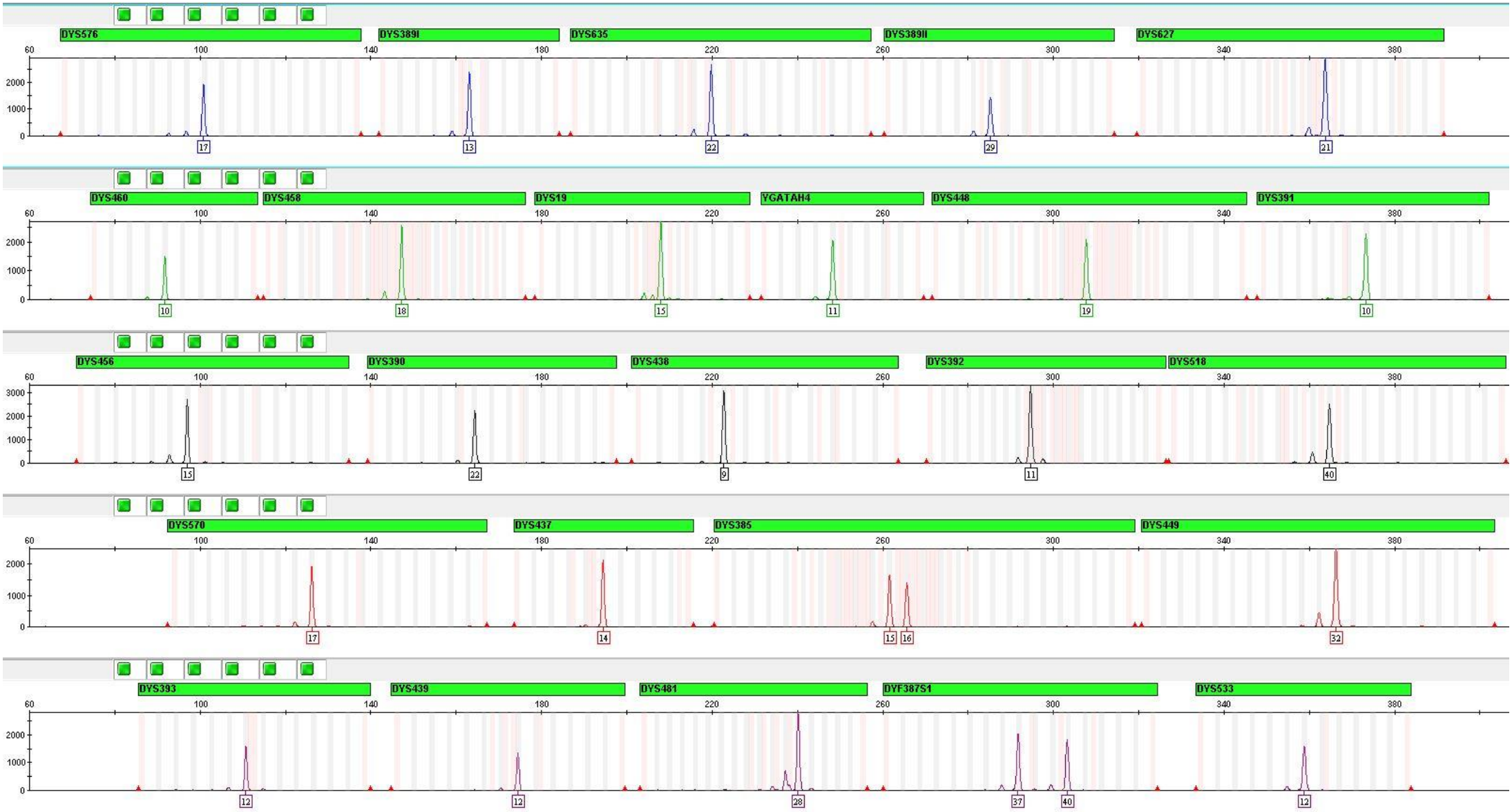


Figure 3.2: *A representative electropherogram of a standard full profile that was generated for a sample included in statistical analysis.*

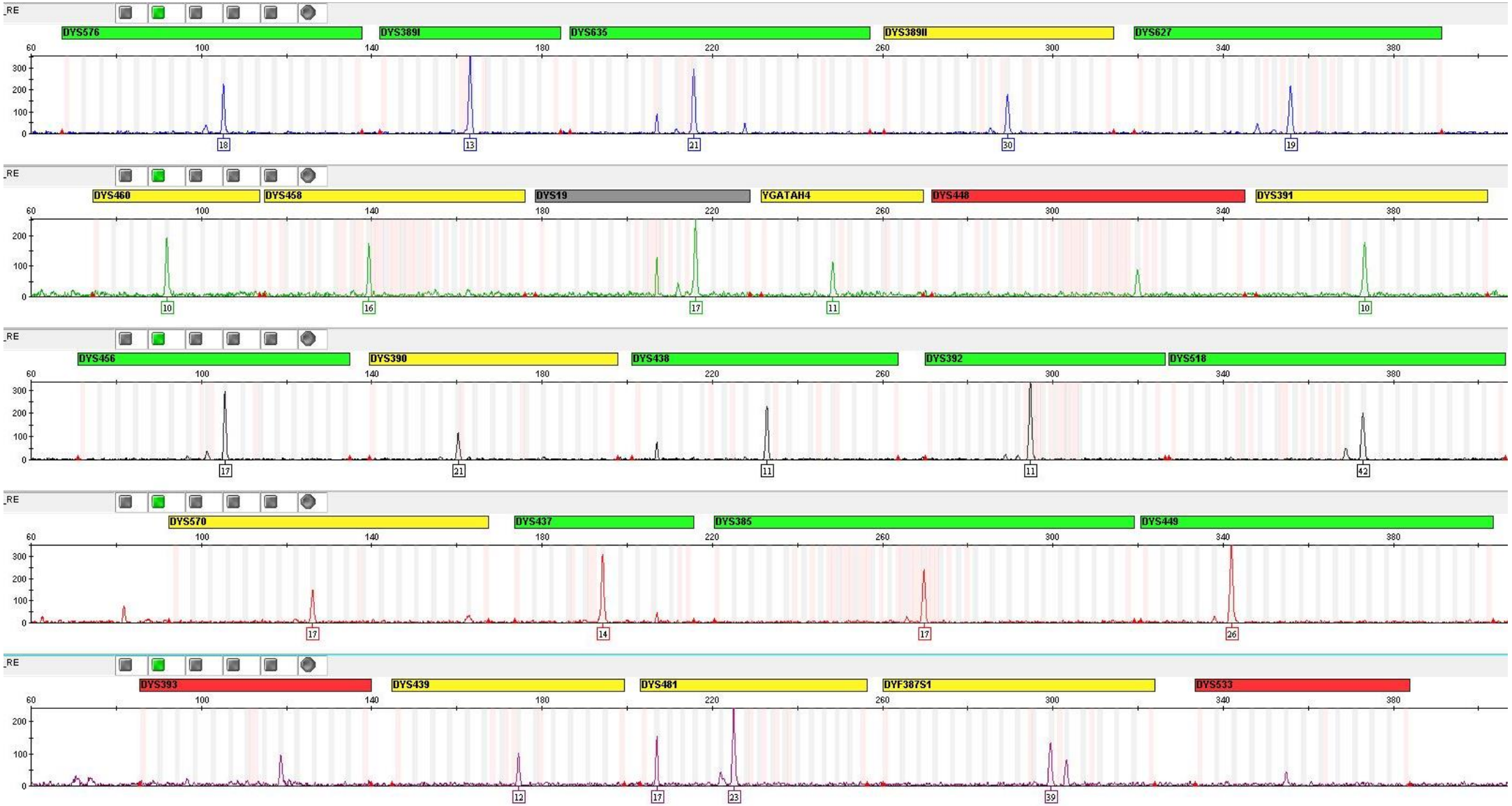


Figure 3.3: *A representative electropherogram of a partial profile generated for a sample that was excluded from further analysis.*

The Shona population is native to Zimbabwe, while Idoma and Igbo people are native to Nigeria. The subgroups of Venda, Swati, Tsonga, Shona, Ndebele, Mpondo, Idoma, and Igbo were not used in the statistical analysis as separate subgroups due to their small sample sizes The South African Caucasian population was divided into two subgroups: the Afrikaans and English subgroups. This subdivision was done in order to evaluate the capabilities of the testing kit within this population group, and to make accurate comparisons to previous research conducted on these two specific subgroups.

The pie charts in Figure 3.4 below show the number of profiles generated for each subgroup in the African and Caucasian population groups. The number of samples for the Asian/Indian and Coloured populations were too low to further divide, so the statistical analyses for these groups were calculated as an overall value.

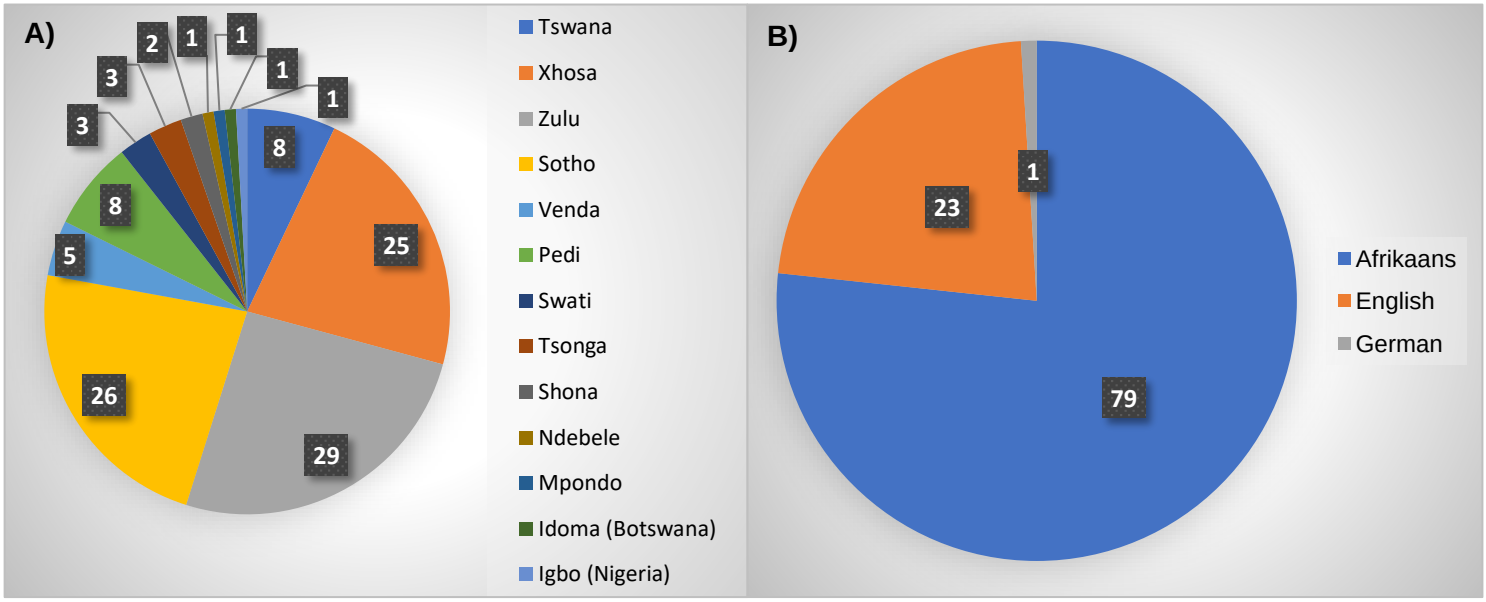


Figure 3.4: *The number of samples in each subgroup for the A) African and B) Caucasian populations.*

3.3.3 Statistical Analysis

The allelic and haplotype frequencies were calculated from the 271 Y STR profiles obtained using the direct counting method and are presented in Appendix D (Page 120) and Appendix E (Page 127) respectively. Additional forensic parameters are presented in Table 3.2, including: the number of unique haplotypes (UHs); the number of different haplotypes; discrimination capacity (DC); match probability (MP); haplotype diversity (Hd); and the number of private alleles (PAs).

Table 3.2: *Summary statistics for the Asian/Indian, African, Coloured, and Caucasian populations.*

	Asian/Indian	African	Coloured	Caucasian	Overall
n	12	113	43	103	271
# of UHs	10	109	43	99	261
# of different haplotypes	11	111	43	101	266
DC (%)	91.67	98.23	100	98.06	98.15
MP	0.0972	0.0092	0.0233	0.0101	0.0038
MP (1 in …)	10	109	43	99	261
Hd	0.9848	0.9997	1.0000	0.9996	0.9999
# of PA	1	17	13	9	-

n = sample size UHs = unique haplotypes DC = discrimination capacity

MP = match probability PA = private alleles Hd = haplotype diversity

3.3.3.1 Unique Haplotypes

A unique haplotype is one that is observed only once in a population (ThermoFisher Scientific, 2019). There were no shared haplotypes between the four main population groups. In the Coloured population, all the haplotypes detected were found to be unique (100%), as shown in Table 3.2. The Coloured population having the highest percentage of unique haplotypes between the population groups is in line with previous studies (Ehrenreich *et al.,* 2008). This is most likely due to the high levels of admixture within this population group. The Asian/Indian population has one shared haplotype, resulting in 10 unique haplotypes (83%) being detected. There are two shared haplotypes in both the African and Caucasian populations. Consequently, the African and Caucasian populations have 109 (96%) and 99 (96%) unique haplotypes respectively. The percentages of unique haplotypes in each population group in this study is increased from those reported in previous studies (Leat *et al.,* 2004; Ehrenreich, 2005; Leat *et al.*, 2007).

Overall, out of the 271 haplotypes analysed in total, 261 haplotypes (96%) were detected as being unique. This high percentage of unique haplotypes is already a promising indicator of the discriminating power of these 27 Y-STR markers in South Africa. This value is slightly higher than the 91% unique haplotypes that were reported using the UniQ TYPER™ Y-10 genotyping system (D'Amato and Kasu, 2017).

The number of different haplotypes observed in the Asian/Indian, African, Coloured, and Caucasian groups are 11, 111, 43, and 101 respectively. The number of different haplotypes detected in a population directly influences the discrimination capacity of the Y-STR testing kit, as it is the value used to calculate the discrimination capacity. The higher the number of different haplotypes detected, the more efficient the testing kit becomes in differentiating between individuals.

3.3.3.2 Discrimination Capacity

The number of different haplotypes in a population was divided by the total number of samples in that population to calculate the discrimination capacity. As indicated in Table 3.2, an overall discrimination capacity of 98.15% was achieved for the South African population. The Coloured population showed the highest DC at 100%, as all 43 haplotypes in the group were different. A DC of 100% is the ideal target as it means that this particular set of markers is able to differentiate between 100% of all the males in that population. The DC obtained for the Coloured population in this study supports the idea by Quintana-Murci *et al.* (2010) when they noted that high levels of admixture results in increased genetic diversity within a population group, which would make it easier to differentiate between individuals in that population. The DCs for the African and Caucasian groups were slightly lower at 98.23% and 98.06% respectively. While these DCs are not at the desired target of 100%, they are still comparable with that of several other countries (Olofsson *et al.,* 2015; Rapone *et al.,* 2016; Khubrani *et al.,* 2018). Of particular note is that the DC values obtained in this study are higher than that calculated for Eastern and Northern African populations using the same set of loci (Iacovacci *et al.,* 2017; D'Atanasio *et al.,* 2019).

With the release of the Y-Filer® Plus kit, ThermoFisher Scientific (2019) reported on DCs for the African-American, Caucasian, Hispanic, and Asian groups in the USA. The USA Asian DC was significantly lower than the other groups at 94.5%, and this is also reflected in the South African Asian/Indian group, with a DC of 91.67%. Even though this value is low in comparison to the rest of the population groups, the sample size of the Asian/Indian group is smaller in comparison and, therefore, even just one shared haplotype would have a larger effect on the DC calculation. The USA Caucasian DC was 98.5%, which is only marginally higher than for the South African Caucasian group. The African-American DC is particularly high at 99.6%, while South Africa's African DC is 98.23%. This comparison does provide additional evidence that some markers used in commercial testing kits have lower resolution capabilities in South Africa's African population, which is in line with other studies conducted on the South African population (Leat *et al.,* 2004; Ehrenreich, 2005; Leat *et al.*, 2007). However, it is important to

consider that the population size per group in the USA study was significantly larger than in the current study, with 557 African-American, 533 Caucasian, 391 Hispanic, and 340 Asian samples. As with the Asian/Indian group, the smaller population size of the African population in this study in comparison with the USA African-American group could be the reason for the African DC seeming lower than in the USA study.

Obtaining DCs for the four South African population groups that are comparable with the published data for other populations already suggests that the Y-Filer® Plus kit has performed desirably in the South African population. However, comparing the overall DC of 98.15% to the results of previous research conducted on the South African population emphasises the success of these 27 markers. Using the MH loci only resulted in an overall DC of 77.3%, while the UniQ Typer™ Y-10 loci resulted in an overall DC of 90.9% (D'Amato *et al.,* 2008, D'Amato and Kasu, 2017). It is evident from the results presented in Table 3.2 that using this specific set of 27 Y-STR loci significantly increased the discrimination capacity in the South African population. It is possible that the increase in the number of loci amplified could be the reason for the increased discrimination capacity. By increasing the number of markers included, the chances for detectable variation is increased. It is, however, more likely that it is more about which markers are used as opposed to how many (Ballantyne *et al.,* 2012). The inclusion of seven RM loci in the Y-Filer® Plus kit most likely played a role in increasing the discrimination capacities detected using this kit.

3.3.3.3 Match Probability

The match probability is the probability of any two individuals in the same population having the same DNA profile (Chakraborty and Kidd, 1991). This statistic can be expressed as a decimal or as a value of 'one in … chance of any two haplotypes matching in a population'. The product rule is generally used to calculate the MP when dealing with autosomal STRs; however, this rule cannot be applied to Y-STRs given the non-recombination and linkage disequilibrium of the Y-chromosome (Walsh *et al.,* 2008). For this reason, the MPs when using Y-STRs are often much higher than that of autosomal STRs. The higher the MP, the more chance of two haplotypes in a population matching, and the less statistical confidence the DNA evidence would hold. It is evident in Table 3.2 that these MPs are especially high and the reason for this could be that the sample sizes of the populations are too small. For example, the MP for the African population is one in 109 individuals. This means that even in the context of the University of the Free State, with an African male population of 12 418, approximately 114 individuals might share a haplotype. While 100 samples per population group may be sufficient for accurate allele frequency estimations (Chakraborty, 1992), the

estimation of accurate MPs requires large, comprehensive reference databases as the estimation is based on haplotype frequencies in the population in question (Kayser, 2017). As the number of individuals that are genotyped increases—and the more comprehensive the reference database becomes—the match probabilities calculated are reduced and the confidence in the DNA evidence increases. Should the use of Y-STRs be considered in the SAPS forensic laboratories, a large reference Y-STR database of the South African male population would have to be established too.

3.3.3.4 Haplotype Diversity and Private Alleles

There are multiple factors that could have an influence on the discrimination capacity. The DC is a ratio of the different haplotypes in a population and the total number of haplotypes in that population. The haplotype diversity across populations influences the DC, as it indicates how many differences there are between haplotypes, thereby indicating how many different haplotypes are present (Kayser *et al.,* 2004). Again, the number of loci analysed affects the diversity detected, as more loci included means more chances for variation being detected. The overall Hd for the South African population in this study is 0.9999 (Table 3.2) which is significantly higher than 0.67 detected using the MH loci (D'Amato *et al.,* 2008). The Hds calculated in this study ranged from 0.9848 in the Asian/Indian population to 1.000 in the Coloured population. The African and Caucasian populations had Hds of 0.9997 and 0.9996, respectively. The Hd for the Coloured population is 1.000, which indicates that all 43 haplotypes are 100% different, resulting in the DC of 100%. These increased Hds relate to the increased discrimination capacities calculated in this study. The correlation between DC and Hd is shown in the Asian/Indian population whereby the DC is decreased to 91.67% and there is a corresponding decrease in Hd with 0.9848. This decrease is exaggerated due to the small population size, but the sentiment remains.

The number of private alleles (PAs), or alleles that are unique to a specific population, is a simple measure of the genetic diversity between populations (Kalinowski, 2004). The presence of private alleles in each population could prevent shared haplotypes between populations, and thus increase the diversity between populations. Increased variation between populations results in an increase in the overall discrimination capacity of the testing kit (Kayser *et al.,* 2004). Private alleles were detected in each population group. The Asian/Indian population only had one PA, while the African population had the most PAs at 17. The Coloured and Caucasian populations had 13 and 9 PAs respectively. A list of the private alleles and their frequencies for each population group can be found in Table 3.3 below. This

high number of PAs detected, 30 in total, correlates to the high overall haplotype diversity of 0.9999. In turn, this haplotype diversity correlates to the increased overall discrimination capacity of 98.15%.

Table 3.3: *List of private alleles for the Asian/Indian, African, Coloured, and Caucasian populations*

Population	Locus	Allele	Frequency
Asian/Indian	DYS635	18	0.167
African	DYS389I	10	0.018
	DYS389I	11	0.009
	DYS635	19	0.009
	DYS389II	32.3	0.009
	DYS390	20	0.010
	DYS518	43	0.089
	DYS518	46	0.009
	DYS570	14	0.009
	DYS570	22	0.009
	DYS449	25	0.009
	DYS449	36	0.044
	DYS449	37	0.018
	DYS393	16	0.018
	DYS439	14	0.035
	DYS439	15	0.009
	DYF387S1	41	0.010
	DYF387S1	42.2	0.005
Coloured	DYS635	29	0.023
	DYS458	13	0.047
	DYS448	16	0.024
	DYS391	8	0.023
	DYS390	18	0.023
	DYS390	19	0.023
	DYS392	15	0.023
	DYS518	35	0.093
	DYF387S1	32	0.013
	DYF387S1	40	0.038
	DYF387S1	40.2	0.013
	DYF387S1	45.2	0.013
	DYS533	8	0.047
Caucasian	DYS576	12	0.010
	DYS576	22	0.010

Population	Locus	Allele	Frequency
	DYS389I	15	0.010
	DYS627	15	0.010
	DYS460	9	0.039
	DYS458	17.2	0.010
	DYS392	9	0.010
	DYS518	36	0.107
	DYS385	9	0.005

3.3.3.5 Summary statistics for the subgroups

These forensic statistics were also calculated for the Afrikaans, English, Xhosa, Zulu, and Sotho subgroups in order to allow for comparisons with previous research, as well as to investigate the capability of the testing kit within each population group (Table 3.4). Previous research has indicated that the Xhosa group exhibits particularly low levels of genetic diversity and, therefore, lower discrimination capacities (Leat *et al.,* 2004; D'Amato *et al.,* 2008; D'Amato *et al.,* 2017).

The Xhosa group in this study has proven that, while the diversity within this group is lower than the other subgroups, the detectable diversity is significantly increased using this set of 27 markers. With 23 out of 25 haplotypes being unique, a haplotype diversity of 0.9967, and a discrimination capacity of 96%, the supposed lack of genetic diversity within the Xhosa population is not of concern when using the Y-Filer® Plus kit. While the Xhosa subgroup only makes up ~22% of the total African sample population in this study, it could be this group's lower diversity that reduces the overall African discrimination capacity.

Table 3.4: *Number of unique haplotypes (UH), discrimination capacities (DC), and haplotype diversities (Hd) for the Afrikaans, English, Xhosa, Zulu, and Sotho population subgroups.*

	Afrikaans	English	Xhosa	Zulu	Sotho
n	79	23	25	29	26
# of UH	75	23	23	29	26
DC (%)	97.46	100	96	100	100
Hd	0.9994	1.000	0.9967	1.000	1.0000

n = sample size UH = unique haplotypes DC = discrimination capacity

Hd = haplotype diversity

The Zulu and Sotho populations both exhibit 100% diversity with all the haplotypes being unique, haplotype diversities of 1.000, and discrimination capacities of 100%. The English subgroup exhibits the same complete diversity. The Afrikaans group, however, consists of two shared haplotypes, reducing the number of unique haplotypes to 75. This resulted in a haplotype diversity of 0.9994 and a discrimination capacity of 97.46%. Even when divided into the subgroups, the South African population still exhibits high levels of diversity and discriminating power. This is testament to the success of the Y-Filer® Plus kit in the South African population.

Figure 3.5 reiterates the relationship between the haplotype diversity and discrimination capacity. As seen with the Afrikaans and Xhosa groups, a reduced haploid diversity is detected in conjunction with a reduced discrimination capacity. It can, therefore, be concluded that the diversity of individual loci directly influences the haplotype diversity and therefore the discrimination capacity too. The alleles that are detected at each loci affect the diversity thereof. When considering the overall population groups only, all loci are 100% polymorphic in this study.

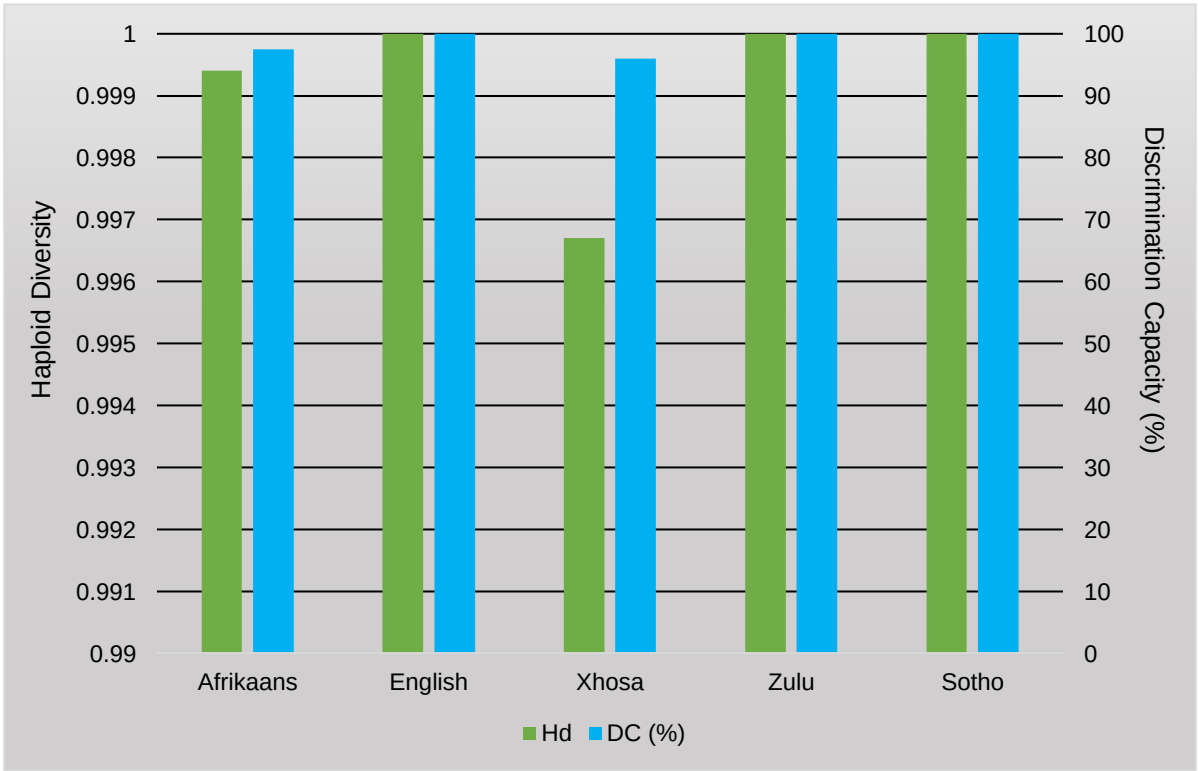


Figure 3.5: *The Discrimination Capacity (DC) and Haplotype diversity (Hd) for the Afrikaans, English, Xhosa, Zulu, and Sotho population subgroups.*

3.3.3.6 Gene diversity

The biggest contributor to the increased discrimination capacities, and the success rate of the Y-Filer® Plus kit in the South African population, is the gene diversity at each Y-STR marker. Table 3.5 presents the gene diversity (GD) calculated for each Y-STR marker in the four main population groups. There are four markers—*DYS391, DYS392, DYS393, and DYS437*—that exhibit extremely low levels of gene diversity, and these markers in the relevant populations are highlighted in red. The seven RM loci (*DYS449, DYS518, DYS570, DYS576, DYS627,* and *DYF387S1a/b*) are highlighted in yellow. The marker with the highest gene diversity in each population is highlighted in green.

The GD values for the Asian/Indian population range from 0.167 to 0.894, with a mean of 0.705. It is interesting to note that the locus with the highest GD value in this population, and in all the populations (*DYS635)*, is not one of the RM loci. With the lowest mean GD of 0.625, the GD values for the African population range from 0.052 to 0.872. The Coloured population presented the highest mean GD at 0.729, with values ranging from 0.464 to 0.889. The Caucasian population had a mean of 0.676 and GDs ranging from 0.361 to 0.849.

The gene diversities obtained in this study are comparable to the data relating to the USA published by ThermoFisher Scientific (2019). It has, however, been considered that the sample sizes of the population groups in this study are not as large as those in this published data, so it is possible that the GD values could be slightly overestimated. Despite this, the GD values for the South African populations are on par with the USA population using the 27 Y-Filer® Plus markers. The mean GDs for the USA African-American, Caucasian, Hispanic, and Asian populations are 0.698, 0.658, 0.701, and 0.726 respectively. These values are very similar to those calculated in this study. While larger sample populations should be further investigated in the South African population to avoid any overestimations, this study certainly provides evidence for the great potential that this set of 27 Y-Filer® Plus markers has in this country.

The two markers that have the highest average GD across all four populations are *DYS518* and *DYS385*. *DYS518* is a RM locus, so the diversity is expected to be increased compared to other non-RM loci. *DYS385* is not a RM locus so it would not necessarily be expected to be as diverse as it is presented to be. However, *DYS385* is a multicopy marker, and includes *DYS385a* and *DYS385b*.

Table 3.5: *The Gene Diversity (GD) at each locus for each population. RM loci are highlighted in yellow. Loci with particularly low levels of gene diversity are highlighted in red. The locus with the highest gene diversity in each population is highlighted in green.*

Locus	Asian/Indian (n*=12)	African (n=113)	Coloured (n=43)	Caucasian (n=103)	Average
DYS576	0.803	0.798	0.782	0.757	0.785
DYS389I	0.621	0.654	0.504	0.572	0.588
DYS635	0.894	0.752	0.827	0.643	0.779
DYS389II	0.712	0.802	0.756	0.709	0.745
DYS627	0.864	0.729	0.829	0.849	0.818
DYS460	0.485	0.535	0.594	0.579	0.548
DYS458	0.742	0.689	0.812	0.771	0.754
DYS19	0.652	0.651	0.684	0.570	0.639
YGATAH4	0.803	0.627	0.630	0.609	0.667
DYS448	0.652	0.658	0.805	0.624	0.684
DYS391	0.167	0.221	0.464	0.530	0.345
DYS456	0.591	0.554	0.619	0.726	0.622
DYS390	0.745	0.659	0.822	0.745	0.743
DYS438	0.712	0.449	0.730	0.621	0.628
DYS392	0.561	0.119	0.699	0.607	0.497
DYS518	0.848	0.872	0.849	0.830	0.850
DYS570	0.864	0.789	0.794	0.775	0.805
DYS437	0.409	0.052	0.547	0.646	0.413
DYS385a/b	0.874	0.851	0.877	0.797	0.850
DYS449	0.758	0.864	0.889	0.819	0.832
DYS393	0.848	0.601	0.598	0.361	0.602
DYS439	0.712	0.626	0.633	0.648	0.655
DYS481	0.758	0.821	0.888	0.790	0.814
DYF387S1a/b	0.830	0.805	0.853	0.777	0.816
DYS533	0.712	0.442	0.740	0.542	0.609
Mean GD	**0.705**	**0.625**	**0.729**	**0.676**	**0.684**

*n = sample size

The presence of two alleles are common at this marker and, therefore, increases the diversity at this marker despite it not being a RM locus. Despite *DYF387S1a/b* being both a multicopy and RM locus, the GD is not as high as one would expect, particularly in the Caucasian population. Unfortunately, there is no published data available on the *DYF387S1a/b* locus in South Africa, and so a comparison within the South African context cannot be made. Despite this, the seven RM loci do generally exhibit higher GD values than the other loci across all the populations, with values ranging from 0.729 at *DYS627* to 0.872 at *DYS518.*

Highlighted in red in Table 3.5, there are four loci (*DYS391, DYS392, DYS393, and* DYS437) that have particularly low gene diversity. The low diversity values at these markers are consistent with findings from previous research conducted in South Africa (Leat *et al.,* 2004; Leat *et al.,* 2007). *DYS391* exhibited low diversity in the Asian/Indian population (0.167 – the lowest GD in this population), with all individuals but one sharing allele 10. *DYS391* also lacked diversity in the African population (0.221) as ~87% of the group shared allele 10. *DYS392* showed even less diversity in the African population at 0.119 with ~93% of the population sharing allele 11. *DYS437* was especially low in diversity (0.052) with ~97% of the African population sharing allele 14 and only three individuals sharing allele 15, making *DYS437* biallelic in the African population. The last marker with reduced gene diversity, *DYS393*, occurred in the Caucasian population with a GD of 0.361. While this value is certainly not as low as the other three markers, it is low in comparison with the rest of the markers for the Caucasian population. At this locus, ~78% of the population shared allele 13, ~13% shared allele 14, and ~7% shared allele 12. A decreased GD (0.363) at *DYS393* was also reported in the USA Caucasian population, so this is not unique to the South African Caucasian population (ThermoFisher Scientific, 2019).

These four markers that showed low gene diversity were evaluated in the Afrikaans, English, Tswana, Xhosa, Zulu, Sotho, and Pedi subgroups to determine which subgroups are experiencing low diversity and contributing to the overall low diversity. Figure 3.6 illustrates the gene diversity, or lack thereof, at these four markers in these specific population subgroups. It is also evident from the graph that the gene diversity at *DYS391, DYS392*, and *DYS437* is much higher in the Afrikaans and English (Caucasian) subgroups. Conversely, the gene diversity at *DYS393* is higher in the African subgroups (Tswana, Xhosa, Zulu, Sotho, Pedi) than in the Afrikaans and English subgroups.

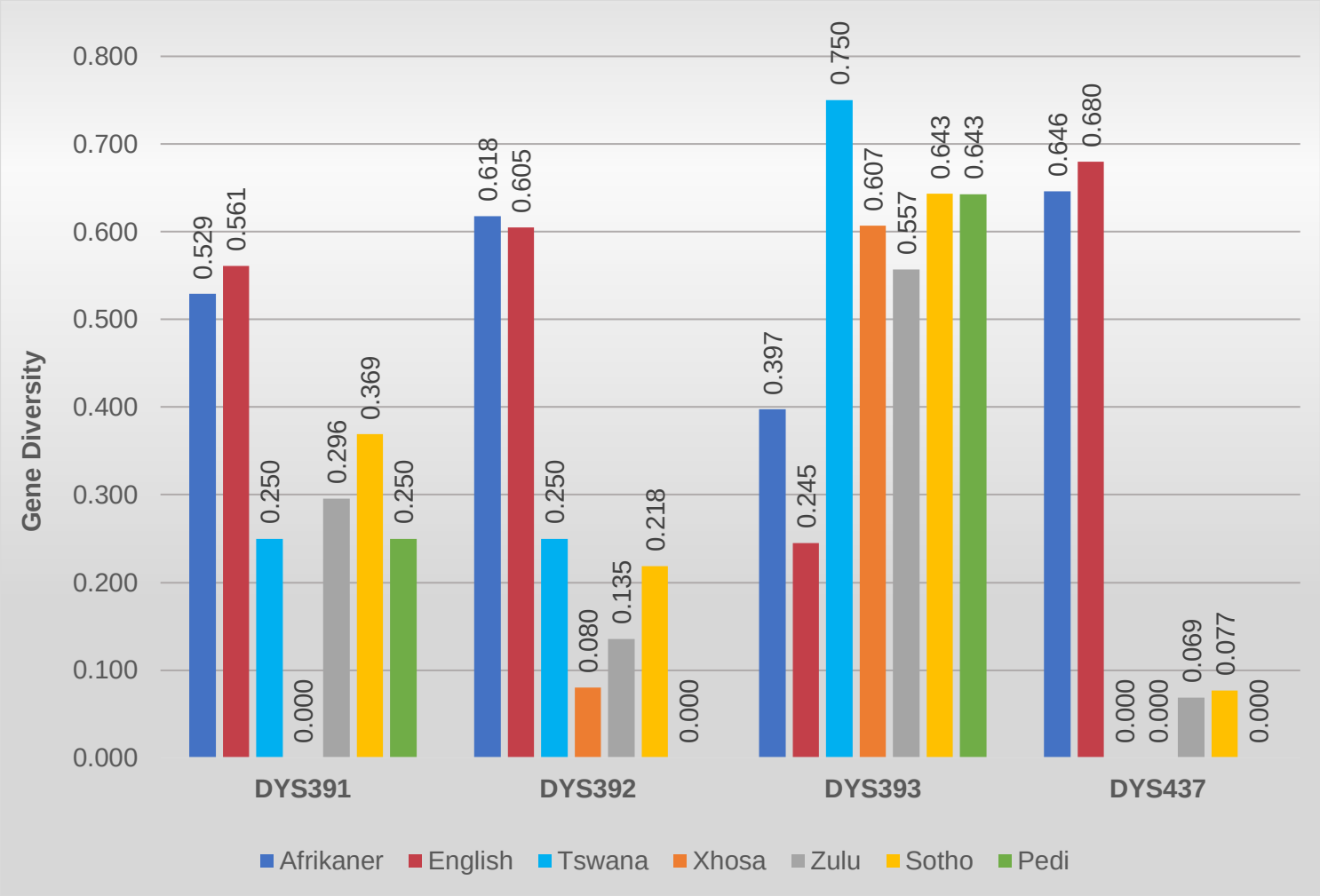


Figure 3.6: *The gene diversity in the different Caucasian and African population subgroups at the four markers DYS391, DYS392, DYS437, and DYS393.*

The GD at *DYS393* is low in the Caucasian population overall. However, Figure 3.6 indicates that it is the English subgroup, in particular, that is lacking gene diversity. As already mentioned, when considering the four main population groups of Asian/Indian, African, Coloured, and Caucasian overall, all loci were polymorphic. However, the graph in Figure 3.6 indicates that, when considering separate population subgroups, not all the loci are polymorphic. The gene diversity at both *DYS391* and *DYS437* for the Xhosa subgroup were calculated to be 0.000, meaning that these two markers are completely monomorphic. All the Xhosa individuals in this study share allele 10 at *DYS391*, and allele 14 at *DYS437*. This low gene diversity detected in the Xhosa subgroup is consistent with previous research (Leat *et al.,* 2004; D'Amato *et al.,* 2009). The Pedi subgroup is also monomorphic at two markers, with all individuals sharing allele 11 at *DYS391* and allele 14 at *DYS437.* Additionally, the Tswana subgroup is monomorphic at *DYS437*, with all individuals also sharing allele 14. While the

Zulu and Sotho subgroups are not monomorphic at any of these markers, the diversity is significantly lower at *DYS437*.

All the statistics calculated and represented throughout this chapter have indicated that the Xhosa subgroup has the lowest diversity. It is difficult to conclusively state the reason for the lack of diversity in this subgroup. Traditionally individuals in the Nguni group (Xhosa and Zulu) marry outside of their clans, and a man can have multiple wives (Carver, 2007). For this reason, the expectation would be an increase in genetic diversity within these subgroups, but the results of this study appear to contradict this. On the other hand, the Sotho-Tswana subgroups (including Pedi) favour marriage within their cultures (Kuper, 1975). This is expected to show a reduction in the overall genetic diversity in these three groups, but again, this does not seem to be the case.

3.3.3.7 Father-son pairs in South Africa: A recommendation for future studies

The uniparental inheritance pattern of the Y-chromosome and the low mutation rates of Y-STRs may be of concern in studies aimed at differentiating between closely related males, such as father and son or brothers (Roewer, 2009). When analysing DNA evidence using Y-STRs, and a match occurs between a suspect and the evidence, closely related males of the suspect could not be immediately excluded: unless a mutation occurs and the relative does not match the evidence, or if there is other evidence proving innocence. Although the inclusion of RM Y-STR markers has been successful in increasing resolution between related males, it may not always prove beneficial in analyses where false exclusions could hinder investigations relating to missing persons cases, victim identifications, and paternity testing (Baeta *et al.,* 2018). Regardless of this concern, being able to distinguish between closely related males in a population would greatly increase the discrimination capacity of a set of Y-STR markers.

While potential bias was avoided during the sampling process for the study undertaken here, through sampling only unrelated males, it would be interesting to investigate the resolution of closely related males using the 27 Y-Filer® Plus loci in South Africa. In this study, these 27 markers have proven to successfully resolve 96% of all the haplotypes and significantly increased the discrimination capacity in the Asian/Indian, African, Coloured, and Caucasian populations. Even with the decreased level of gene diversity in some of the African subgroups, this set of markers has great potential to differentiate between closely related males based on the high haplotype diversities and discrimination capacities detected in this study. Future studies using the Y-Filer® Plus PCR Amplification kit in South Africa should focus on

investigating father-son pairs and the ability of these 27 Y-STR markers to successfully distinguish between closely related males.

3.4 Conclusion

An overall discrimination capacity of 98.15% indicates that the 27 Y-Filer® Plus markers perform remarkably well in the South African population. This is further supported by the high number of unique haplotypes and high haplotype diversities observed for the four population groups and the overall population. All except four markers exhibited moderate to high gene diversity in all four population groups. The four markers with low gene diversity, however, did not greatly impact on the overall gene diversity, nor did it significantly reduce the discriminatory power of the multiplex. The results obtained are comparable with those for populations in the USA, Europe, and the Middle East, and suggest that this commercial Y-STR testing kit is a feasible option for use in forensic investigations in South Africa.

Despite the abovementioned success, there is still a factor that hinders the plausible implementation of this DNA technology in forensic laboratories in South Africa: population reference data. It is evident with the South African population that, even though more than 100 samples were used and the allele frequencies are considered as conservatively estimated, the match probabilities are still very high for a forensic application. It could, therefore, be said that, when working with Y-STRs, 100 samples per population group may not be sufficient for the accurate estimation of match probabilities. Future studies on the South African population should increase genotyping efforts using the Y-Filer® Plus PCR Amplification kit, in order to establish a comprehensive reference database for accurate and reliable match probability calculations.

This study has proven that this marker set is a viable option for use of a commercial kit in the SAPS forensic laboratories. South Africa has one of the highest rape statistics in the world, with an average of 117 rapes being reported per day (South African Police Service, 2020). A combination of comprehensive population reference data and forensically valuable Y-STRs to supplement autosomal STRs during DNA analysis could drastically increase the efforts to reduce this statistic.

References

Baeta, M., Núñez, C., Villaescusa, P., Ortueta, U., Ibarbia, N., Herrera, R.J., Blazquez-Caeiro, J.L., Builes, J.J., Jiménez-Moreno, S., Martínez-Jarreta, B., and de Pancorbo, M.M. (2018) 'Assessment of a subset of Slowly Mutating Y-STRs for forensic and evolutionary studies'. *Forensic Science International: Genetics*, 34, pp.e7-e12.

Ballantyne, K.N., Keerl, V., Wollstein, A., Choi, Y., Zuniga, S.B., Ralf, A., Vermeulen, M., de Knijff, P., and Kayser, M. (2012) 'A new future of forensic Y-chromosome analysis: rapidly mutating Y-STRs for differentiating male relatives and paternal lineages'. *Forensic Science International: Genetics*, 6(2), pp.208-218

Bernitz, H., Kenyhercz, M., Kloppers, B., L'Abbé, E.N., Labuschagnes, G.N., Olckers, A., Myburgh, J., Saayman, G., Steyn, M., and Stull, K. (2015) 'The history and current status of forensic science in South Africa'. In: Ubelake, D.H. ed., *The Global Practice of Forensic Science*. Chichester: John Wiley & Sons, pp.241-260.

Butler, J. M. (2005) 'Microvariants and 'Off-Ladder' Alleles'. In: Butler, J.M ed., *Forensic DNA Typing: Biology, Technology, And Genetics of STR Markers*. 2nd ed. Elsevier, pp.130-131.

Carver, D.H. (2007) 'The Xhosa and the truth and reconciliation commission: African ways.' *Tribal Law Journal*, 8, pp.34-52.

Chakraborty, R. and Kidd, K. K. (1991) 'The utility of DNA typing in forensic work'. *Science*, 254(5039), pp.1735-1739.

Chakraborty, R. (1992) 'Sample size requirements for addressing the population genetic issues of forensic use of DNA typing'. *Human Biology*, pp.141-159.

Cloete, K., Ehrenreich, L., D'Amato, M. E., Leat, N., Davison, S., and Benjeddou, M. (2010) 'Analysis of seventeen Y-chromosome STR loci in the Cape Muslim population of South Africa'. *Legal Medicine*, 12(1), pp.42-45.

D'Amato, M.E., Bajic, V.B. and Davison, S. (2011) 'Design and validation of a highly discriminatory 10-locus Y-chromosome STR multiplex system'. *Forensic Science International: Genetics*, 5(2), pp.122-125.

D'Amato, M. E., Benjeddou, M., and Davison, S. (2009) 'Evaluation of 21 Y-STRs for population and forensic studies'. *Forensic Science International: Genetics Supplement Series*, 2, pp.446-447.

D'Amato, M.E., Ehrenreich, L., Benjeddou, M., Davison, S. and Leat, N. (2008) 'Ancestry and genetic relationships between groups within the Cape Town metropolitan population inferred using Y-STRs genotyping'. *Forensic Science International: Genetics Supplement Series*, 1(1), pp.318-319.

D'Amato, M. E., and Kasu, M. (2017) 'Population analysis of African Y-STR profiles with UniQ Typer™ Y-10 genotyping system'. *Forensic Science International: Genetics Supplement Series*, 6(October), pp.e84-e85.

D'Atanasio, E., Iacovacci, G., Pistillo, R., Bonito, M., Dugoujon, J.M., Moral, P., El-Chennawi, F., Melhaoui, M., Baali, A., Cherkaoui, M., and Sellitto, D. (2019) 'Rapidly mutating Y-STRs in rapidly expanding populations: Discrimination power of the Yfiler Plus multiplex in northern Africa'. *Forensic Science International: Genetics*, 38, pp.185-194.

Ehrenreich, L., Benjeddou, M., Davison, S., D'Amato, M., and Leat, N. (2008) 'Nine-locus Y-STR profiles of Afrikaner Caucasian and mixed ancestry populations from Cape Town, South Africa'. *Legal Medicine*, 10(4), pp.225-227.

Iacovacci, G., D'Atanasio, E., Marini, O., Coppa, A., Sellitto, D., Trombetta, B., Berti, A., and Cruciani, F. (2017) 'Forensic data and microvariant sequence characterization of 27 Y-STR loci analysed in four Eastern African countries'. *Forensic Science International: Genetics*, 27, pp.123-131.

Johnson, C.L., Warren, J.H., Giles, R.C., and Staub, R.W. (2003) 'Validation and uses of a Y-chromosome STR 10-plex for forensic and paternity laboratories'. *Journal of Forensic Sciences*, 48(6), pp.1260-1268.

Kalinowski, S.T. (2004) 'Counting alleles with rarefaction: private alleles and hierarchical sampling designs'. *Conservation genetics*, 5(4), pp.539-543.

Kayser, M., Kittler, R., Erler, A., Hedman, M., Lee, A.C., Mohyuddin, A., Mehdi, S.Q., Rosser, Z., Stoneking, M., Jobling, M.A., and Sajantila, A. (2004). 'A comprehensive survey of

human Y-chromosomal microsatellites'. *The American Journal of Human Genetics*, 74(6), pp.1183-1197.

Khubrani, Y.M., Wetton, J.H., and Jobling, M.A. (2018) 'Extensive geographical and social structure in the paternal lineages of Saudi Arabia revealed by analysis of 27 Y-STRs'. *Forensic Science International: Genetics*, 33, pp.98-105.

Kuper, A. (1975) 'The social structure of the Sotho-speaking peoples of Southern Africa'. *Africa*, 45(1), pp.67-81.

Lane, A.B., Soodyall, H., Arndt, S., Ratshikhopha, M.E., Jonker, E., Freeman, C., Young, L., Morar, B., and Toffie, L. (2002) 'Genetic substructure in South African Bantu-speakers: Evidence from autosomal DNA and Y-chromosome studies'. *American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists*, 119(2), pp.175-185.

Leat, N., Benjeddou, M., and Davison, S. (2004) 'Nine-locus Y-chromosome STR profiling of Caucasian and Xhosa populations from Cape Town, South Africa'. *Forensic Science International*, 144(1), pp.73-75.

Leat, N., Ehrenreich, L., Benjeddou, M., Cloete, K., and Davison, S. (2007) 'Properties of novel and widely studied Y-STR loci in three South African populations'. *Forensic Science International*, 168(2–3), pp.154-161.

Nei, M. and Tajima, F. (1981) 'DNA polymorphism detectable by restriction endonucleases'. *Genetics*, 97(1), pp.145-163.

Olofsson, J.K., Mogensen, H.S., Buchard, A., Børsting, C., and Morling. N. (2015) 'Forensic and population genetic analyses of Danes, Greenlanders and Somalis typed with the Yfiler® Plus PCR amplification kit'. *Forensic Science International: Genetics*,16, pp.232-236.

Quintana-Murci, L., Harmant, C., Quach, H., Balanovsky, O., Zaporozhchenko, V., Bormans, C., van Helden, P.D., Hoal, E.G. and Behar, D.M. (2010) 'Strong maternal Khoisan contribution to the South African coloured population: a case of gender-biased admixture'. *The American Journal of Human Genetics*, 86(4), pp.611-620.

Rapone, C., D'Atanasio, E., Agostino, A., Mariano, M., Papaluca, M.T., Cruciani, F., and Berti, A. (2016) 'Forensic genetic value of a 27 Y-STR loci multiplex (Yfiler® Plus kit) in an Italian population sample'. *Forensic Science International: Genetics*, 21, pp.e1-e5.

Redd, A.J., Agellon, A.B., Kearney, V.A., Contreras, V.A., Karafet, T., Park, H., De Knijff, P., Butler, J.M., and Hammer, M.F. (2002). 'Forensic value of 14 novel STRs on the human Y chromosome'. *Forensic Science International*, 130(2-3), pp.97-111.

Roewer, L. (2009) 'Y chromosome STR typing in crime casework'. *Forensic Science, Medicine, and Pathology*, 5, pp.77–84.

South African Police Service. (2020) '*Crime Situation in Republic of South Africa Twelve (12) Months (April to March 2019_20')*. [Online]. Available at: www.saps.gov.za/services/crimestats.php. [Accessed 06 Aug. 2020]

Statistics South Africa. (2020) '*Mid-year population estimates 2020*'. [Online] Available at: http://www.statssa.gov.za/publications/P0302/P03022020.pdf. [Accessed 06 Aug. 2020].

ThermoFisher Scientific. (2019) *Yfiler™ Plus PCR Amplification Kit: User Guide*. Woolston: ThermoFisher Scientific. [Online]. Available at: https://assets.thermofisher.com/TFS-Assets/LSG/manuals/4485610_YfilerPlus_UG.pdf. [Accessed 30 Jan. 2019]

University of the Free State. (2020) '*Management Information – Registrations Profile*'. [Online]. Available at: http://iinfo.ufs.ac.za/powerheda/Dashboard.aspx. [Accessed 21 Apr. 2020].

Walsh, B., Redd, A.J., and Hammer, M.F. (2008) 'Joint match probabilities for Y chromosomal and autosomal markers'. *Forensic Science international*, 174(2-3), pp.234-238.

CHAPTER 4: GENETIC CHARACTERISATION OF THE SOUTH AFRICAN POPULATION USING THE 27 Y-FILER® PLUS Y-STR LOCI

4.1 Introduction

The population in South Africa is made up of 59.6 million people, of which 29.1 million are males (Statistics South Africa, 2020). The group of African descent is the largest, making up 80.70% of the male population, while the Asian/Indian group is the smallest at 2.70% of the total male population. The Coloured and Caucasian populations comprise 8.80% and 7.80% of the male population respectively. The African group can be further divided into the Nguni, Sotho-Tswana (includes Pedi), and Venda groups (Lane *et al.,* 2002). The Nguni group includes the Zulu, Xhosa, Tsonga, Mpondo, Ndebele, and Swati groups. The Caucasian population group consists of the Afrikaans and English subgroups.

The Coloured population in South Africa has a particularly interesting history, resulting in one of the highest levels of admixture globally (Tishkoff *et al.,* 2009). The Coloured population in South Africa originated during the colonisation by the Dutch Settlers in the Western Cape, and was derived from the integration of the indigenous Khoisan women and European men by marriage (de Wit *et al.,* 2010). The Coloured population also includes natives who were brought to South Africa during the slave trade from Indonesia, Malaysia, the Indian subcontinent, the east coast of Africa, and Madagascar.

The ability to detect high levels of intra- and interpopulation variance increases the efficiency of genetic markers used in forensic DNA analysis (Chakraborty and Deka, 2009). The increase in efficiency of genetic markers leads to an increased ability to distinguish between male individuals. However, this is not the only use of population genetics in a forensic application. Being able to identify from which population group DNA evidence originated could be beneficial if a suspect is not known, such as in the case of Marianne Vaatstra, as discussed in Section 2.9 of this dissertation. Biogeographic ancestry also has its uses in victim identification whereby identification of biological remains is not possible. Additionally, data for distinct populations is needed to provide accurate haplotype frequencies in the calculation of match probabilities for forensic purposes. For these reasons, it is important to be able to differentiate between population groups, and higher levels of interpopulation variance allows for this to happen.

Previous studies using fewer Y-STR loci suggested that there is little genetic differentiation between the four main population groups subgroups, in South Africa (Lane *et al.,* 2002; D'Amato *et al.,* 2008). However, the addition of more loci—specifically markers with higher mutation rates—in a testing kit can increase the amount of variation detected in a population (D'Amato and Kasu, 2017). It could be expected, then, that the set of 27 Y-Filer® Plus markers

(with seven RM loci) used in this study would increase the variation detected between the four populations groups in South Africa.

Factors that could increase variation between individuals and populations are allelic variations such as null alleles, duplications, triplications, and microvariant alleles. As suggested by Budowle *et al.* (2008), each of these variations could increase the intra- (within) and inter- (among) specific diversity of populations, as it results in distinguishable profiles which contribute to haplotype diversity. Not only does this increase the ability to differentiate between population groups, it could also increase the statistical confidence in a match between the two profiles. Therefore, having these types of variations present in the South African population would be beneficial to a forensic application.

Occasionally the primer fails to bind during PCR and this results in what is called a null allele (Kline *et al.,* 2011). The allele fails to amplify due to sequence variation in the template DNA at the primer binding site: such as an indel, frameshift, or simply a mutation of a nucleotide. Allelic dropout is generally associated with partial amplification of degraded forensic casework samples, so null alleles should be verified before being used in profile analysis (Lee and Ladd, 2001; Bender *et al.,* 2004). However, there are means of overcoming this potential issue with null alleles in profile interpretation, and, ultimately, null alleles are not a cause for concern when encountered in forensic DNA analysis (Budowle *et al.,* 2008).

In the Y-Filer® Plus kit, markers *DYS385* and *DYF387S1* are multicopy loci so two alleles are frequently observed at these markers. Duplications (and even triplications) at other loci can, occasionally, be detected. A second allele (or third, in the case of triplications) is detected if the duplicated region independently mutated over time by gaining or losing repeat units, resulting in two different alleles (Butler *et al.,* 2005). Triplications occur through either the second duplication of one of the already duplicated alleles, or the duplication of the entire region containing the two different alleles. The latter would be the most likely scenario for multicopy markers such as *DYS385* and *DYF387S1.* The result of such an event is the presence of three alleles at a single locus. As with duplications, the presence of triplications can severely complicate the analysis of mixture samples, and could potentially result in false inclusions or exclusions of major or minor contributors (Butler, 2005). The presence of multiple alleles at one locus also introduces the possibility of contamination, another major concern in forensic DNA testing.

Microvariant alleles occur when the last repeat unit of the allele is incomplete (Butler, 2005). Any form of sequence variation—such as the deletion of a base pair or a base pair mutation—could cause a partial repeat motif, resulting in the microvariant allele. While calculations can

be done using the alleles detected during STR genotyping to estimate the number of bases—and, thus, the length—of the partial repeat unit, it remains only an estimation as the variation occurs at the sequence level. These intermediate alleles would also contribute to haplotype rarity and increased haplotype diversity.

The aim of this chapter is to represent and discuss the investigation into the ability of the Y-Filer® Plus Amplification Kit to detect significant genetic differentiation between the four main population groups in South Africa. The allelic patterns and genetic distances between population subgroups were used to characterise the variation within the South African population. This chapter also aims to evaluate the DNA profile variations observed during analysis and their occurrences in the South African population. Knowledge of such variations and their prevalence in the population could increase the confidence in using variant alleles/markers in forensic DNA testing in South African laboratories.

4.2 Materials and Methods

4.2.1 Sampling to Genotyping

A total of 308 samples from the Asian/Indian (13), African (132), Coloured (44), and Caucasian (119) population groups were collected at the University of the Free State. These samples were processed using 200 µl of Prep 'n Go Buffer (*Applied Biosystems*), and PCR amplification was performed using the Y-Filer® Plus Amplification Kit (ThermoFisher Scientific, 2019) on a Veriti™ 96-Well Thermal Cycler (*Applied Biosystems*). Following amplification, capillary electrophoresis (CE) was performed on a 3500 Genetic Analyser (*Applied Biosystems*). The DNA profiles were generated and analysed using the GeneMapper™ ID-X Software v1.5 (*Applied Biosystems*). Confirmed off-ladder (OL) peaks were calculated using the method described by Butler (2005, refer to Appendix F, Page 140) and labelled in the profiles.

4.2.2 Statistical Analysis

The GenAlEx Software v6.5 was used to perform statistical analyses. The allelic patterns were analysed across the four main population groups using the 271 samples that generated full profiles. One sample from the African population (Sotho subgroup) was found to have suspected duplications at two markers, *DYS485,* and *DYS449,* which complicated statistical

analysis and for this reason, the sample was excluded from further analysis as represented in this chapter.

Genetic distance (D) between the four population groups was calculated using Nei's unbiased formula (Nei, 1978). An Analysis of Molecular Variance (AMOVA) was also performed using the four main population groups (Asian/Indian, African, Coloured, and Caucasian) to evaluate the genetic differences between these population groups (interpopulation variance), as it considers the differences in the number of repeat units, as well as the molecular relationship of alleles, rather than just allelic frequencies (www.yhrd.org). A second AMOVA was performed using the population subgroups (Tswana, Xhosa, Zulu, Sotho, Venda, Pedi, Coloured, Afrikaans, and English) to evaluate the intrapopulation variance. To further detail the degree of genetic relationship between the four groups, Principal Coordinates Analysis (P-CoA) was performed. P-CoA was performed twice: once including all 270 samples and once excluding any profiles that showed any form of profile variations to evaluate the grouping of the samples without the influence of shared profile variations. Several profile variations were analysed and characterised.

4.3 Results and discussion

4.3.1 Allelic patterns

The number of different alleles, shared alleles, private alleles, and the mean gene diversity across all loci in the four main population groups are listed in Table 4.1. The Coloured population had the highest number of different alleles with 163 detected in total. The African and Caucasian populations each had 157 and 152 different alleles respectively. The number of different alleles detected in the Asian/Indian population was significantly lower than the other three populations at 111. The number of different alleles detected at a locus influences the level of gene diversity at that locus. The more alleles there are, the higher the diversity at the locus, and the more efficient that marker becomes in differentiating between individuals. Conversely, with fewer alleles detected at a locus, the gene diversity at the locus is reduced. The results in Table 4.1 are in line with this, as the Coloured population exhibited the highest number of different alleles, as well as the highest mean gene diversity. An anomaly to this, however, is the Asian/Indian population: this group has the fewest number of different alleles, but a relatively high gene diversity. It is possible that the gene diversity for this population was overestimated due to the smaller sample size. Figure 4.1 depicts the relationship between the number of different alleles, private alleles, and overall gene diversity for the Asian/Indian,

African, Coloured, and Caucasian populations. It is clear from this graph, and Table 4.1, that there is a weak correlation between the number of alleles and the overall gene diversity in this study.

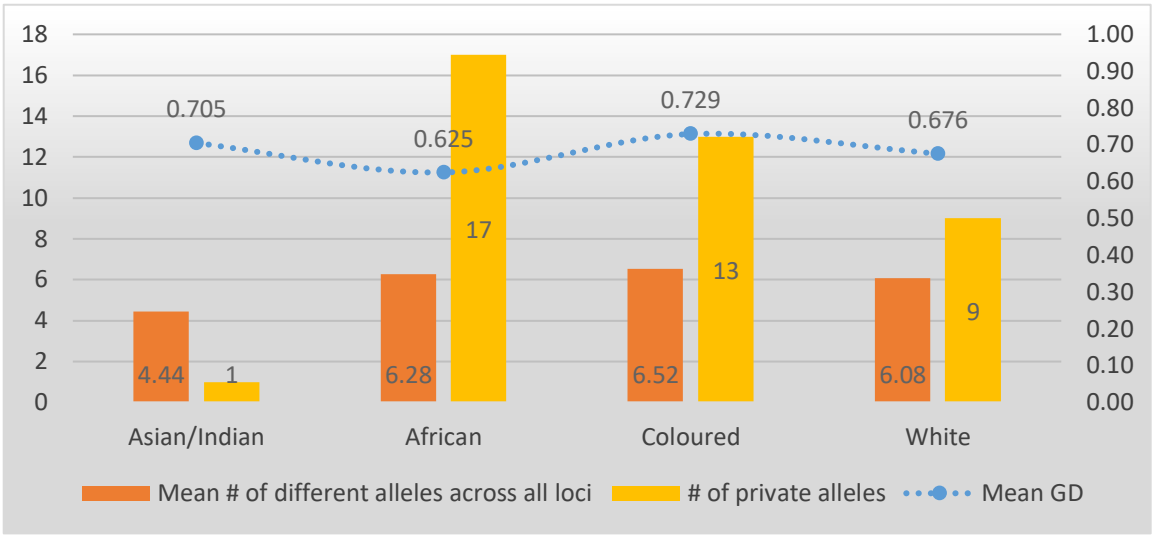


Figure 4.1: *The mean number of different alleles per locus, number of private alleles, and mean gene diversity for the four population groups.*

Table 4.1: *Allelic patterns across all loci for the Asian/Indian, African, Coloured, and Caucasian populations.*

	Asian/Indian	African	Coloured	Caucasian
# of different alleles across all loci	111	157	163	152
# shared alleles	110	140	150	143
# of private alleles	1	17	13	9
Mean gene diversity	0.705	0.625	0.729	0.676

However, it is also clear that there is no correlation between the number of private alleles in a population and the gene diversity within that population. Private alleles are those that are detected only in one population, and Szpiech and Rosenberg (2011) suggest that increased genetic differentiation between populations increases the frequency of these private alleles. A list of the private alleles, and their frequencies, is included in Table 3.4 in Chapter 3. The Caucasian and Asian/Indian populations have nine and one private alleles respectively. Despite being the population with the lowest mean gene diversity, the African population has the highest number of private alleles, with 17 in total. Conversely, the Coloured population had fewer private alleles (13) and the highest gene diversity (0.729). The higher number of private alleles in the African population is consistent with previous findings by Szpiech and

Rosenberg (2011) and is thought to be due to the 'Out of Africa' human migration. The fact that the Coloured population has a smaller sample size than the African and Caucasian populations, but the highest number of different alleles, reiterates the idea that this population has higher levels of gene diversity resulting from the high levels of admixture, as suggested by Quintana-Murci *et al.* (2010). These results support the notion that it is not necessarily which alleles (private or shared) that are detected that contribute to gene diversity, but rather the number of different alleles. A higher number of private alleles does not necessarily result in a higher diversity within populations, although it could increase the variance between populations.

4.3.2 Genetic Distance and AMOVA

The ability to differentiate between population groups is advantageous in forensic investigations, and crucial when calculating match probabilities. The differentiation between the populations in this study was evaluated using several statistical analyses. The genetic distance (D) between the four population groups is presented in Table 4.2 below. It is generally accepted that genetic distance values below 0.050 represent little to no genetic differentiation; values between 0.050 and 0.150 indicates moderate genetic differentiation; values between 0.150 and 0.250 represent great differentiation; and values above 0.250 indicate very great differentiation (Hartl, and Clark, 1997; Frankham *et al.,* 2002).

Table 4.2: *The genetic distances between the four populations, calculated as D values based on Nei's unbiased formula.*

Asian/Indian	**African**	**Coloured**	**Caucasian**	
0.000	-	-	-	**Asian/Indian**
0.152	0.000	-	-	**African**
0.059	0.129	0.000	-	**Coloured**
0.251	0.416	0.080	0.000	**Caucasian**

It can be seen in Table 4.2 that the largest genetic distance is between the African and Caucasian population groups (D = 0.416). Given the history of the South African population, the indigenous population and several colonisation events, this differentiation would be expected. The Asian/Indian population shows only slight differentiation from the Coloured population, which could also be due to the history of the South African Asian/Indian population. Many individuals from Asian countries were brought to the Western Cape during the time of

Dutch colonisation and essentially incorporated into the 'Cape Coloured' population (South African History Online, 2010). Admixture between these groups during this time would result in little genetic variation between them. Conversely, the Asian/Indian population exhibited great differentiation from the African and Caucasian populations. The Coloured population exhibited greater differentiation from the African population (D = 0.129) than the Caucasian population (D = 0.080). This result is consistent with the origins of the Coloured population: from the union of indigenous African women and European men (de Wit *et al.,* 2010). Thus, while the Coloured population might show more genetic similarities with African population when using autosomal STRs, the use of Y-STRs considers the paternal lineage only, resulting in an affinity to the Caucasian population because of the European descent.

An AMOVA was performed to quantify the variance detected within and between the population groups. Figure 4.2 shows that only 19% of all variance occurred among the populations while the majority of variance (81%) occurred within the populations. An associated p value of 0.001 indicates that variance between the populations is, however, significant ($p < 0.05$). Having most of the variance occur within the subpopulations is consistent with previous studies, however, the amount of variance between the populations is increased from those previously reported using fewer Y-STR loci for the South African population (D'Amato *et al.,* 2008; D'Amato and Kasu, 2017). Similar findings have been reported for the USA population (Hammer *et al.,* 2006). These results align with the general trend of increased *within* variance being detected while the *between* variance is decreased with the inclusion of additional Y-STR loci—and RM loci in particular (Purps *et al.,* 2014).

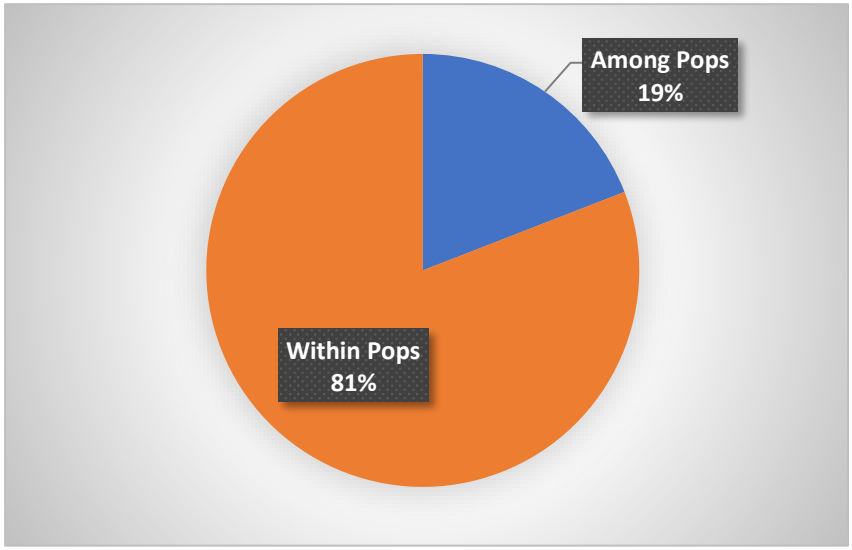


Figure 4.2: *Percentages of Molecular Variance – the amount of variance within and among populations (pops).*

The AMOVA was also used to confirm the differentiation between populations detected using Nei's formula. Nei's formula assumes that genetic differences are brought about by genetic drift and mutations, and is based on allelic frequencies. The AMOVA outputs PhiPT values, which are analogous to R_{ST}, as well as associated p values to indicate the significance of the results (Peakall and Smouse, 2006 and 2012). The R_{ST} estimation is based on genetic distance estimates that assume a stepwise mutation model (Slatkin, 1995). The calculated PhiPT values and associated p values are presented in Table 4.3 below.

Table 4.3: *The PhiPT* values and associated p values calculated during AMOVA.*

Asian/Indian	**African**	**Coloured**	**Caucasian**	
-	0.228	0.409	0.001	**Asian/Indian**
0.016	-	0.002	0.001	**African**
0.000	**0.097**	-	0.001	**Coloured**
0.254	**0.281**	**0.089**	-	**Caucasian**

*PhiPT values are below the diagonal and **bolded** and the associated p values are above the diagonal.

The PhiPT values are reduced in comparison to those in Table 4.2, which demonstrates the inflation of differentiation when using Nei's formula. However, the same general trends of differentiation are shown. The largest genetic differentiation occurred between the African and Caucasian populations, while there were no genetic differences between the Asian/Indian and Coloured populations. The associated p values for the differentiation between the Asian/Indian and Caucasian, Coloured and African, Coloured and Caucasian, and African and Caucasian populations are all smaller than 0.05, indicating that the differences between these groups are significant.

A second AMOVA was performed with some of the population subgroups to further investigate the variation within the main population groups and between the specific subgroups. The subgroups included in this analysis are Tswana, Xhosa, Zulu, Sotho, Pedi, Coloured, Afrikaans, and English. This AMOVA revealed that, when considering the subgroups, 80% of the variance detected occurred within the groups and 20% occurred between the subgroups. However, the associated p value of 0.001 indicated that the differences between the subgroups are significant. In combination with the initial AMOVA, these results indicate that majority of the detectable genetic variance occurs *within* populations, and population subgroups, as opposed to *between* them.

Table 4.4 presents the PhiPT values and associated p values calculated during the AMOVA performed with the population subgroups (Tswana, Xhosa, Zulu, Sotho, Venda, Pedi, Coloured, Afrikaans, and English). There was no differentiation detected between the Tswana, Xhosa, Zulu, Sotho, and Venda subgroups in this study. This result differs from the significant differences reported between the Xhosa, Zulu, Venda, and Pedi groups in previous research (D'Amato and Kasu, 2017). The lack of detectable variance between these groups in this study is not necessarily a true representation of the population, given the small sample sizes of the subgroups. It would be premature and incorrect to conclude that the set of 27 Y-Filer® Plus loci does not have the ability to detect variance between the population subgroups based on this analysis using smaller sample sizes.

Table 4.4: *The PhiPT* values and associated p values calculated during AMOVA performed with the population subgroups Tswana, Xhosa, Zulu, Sotho, Venda, Pedi, Coloured, Afrikaans, and English*

Tswana	Xhosa	Zulu	Sotho	Venda	Pedi	Coloured	Afrikaans	English	
-	0.373	0.348	0.352	0.237	0.027	0.031	0.001	0.001	Tswana
0.000	-	0.379	0.378	0.309	0.001	0.001	0.001	0.001	Xhosa
0.000	**0.000**	-	0.383	0.435	0.001	0.001	0.001	0.001	Zulu
0.000	**0.000**	**0.000**	-	0.396	0.001	0.001	0.001	0.001	Sotho
0.056	**0.000**	**0.000**	**0.000**	-	0.022	0.124	0.001	0.001	Venda
0.368	**0.246**	**0.316**	**0.287**	**0.348**	-	0.001	0.001	0.001	Pedi
0.109	**0.127**	**0.069**	**0.072**	**0.059**	**0.460**	-	0.001	0.009	Coloured
0.460	**0.382**	**0.294**	**0.297**	**0.382**	**0.698**	**0.087**	-	0.377	Afrikaans
0.462	**0.311**	**0.218**	**0.223**	**0.321**	**0.628**	**0.046**	**0.001**	-	English

*PhiPT values are below the diagonal and **bolded** and the associated p values are above

The Coloured population, again, exhibited very little genetic differentiation from the Afrikaans and English subgroups, whereas it showed the largest genetic differentiation from the Pedi subgroup. While the Coloured population has its origins in the Western Cape, the Pedi

population settled in Limpopo, geographically located on the other side of the country (Lane *et al.,* 2002). The geographical distance between these two subgroups could be the reason for the increased genetic differences between them. The largest and most significant differences were observed between the English/Afrikaans groups and the Pedi, Tswana, Venda, Xhosa, Zulu, and Sotho groups. This reveals the distinction between African and non-African population groups.

Principal Coordinates Analysis (P-CoA) was performed to further detail the genetic relationship between these four population groups. The first P-CoA included all 270 samples and this result is shown in Figure 4.3. It is clear, in Figure 4.3, that regardless of the population group, variant samples cluster together and distinctly away from the rest of the samples. The group of samples that is the most separated consists of 13 samples—12 African and one Asian/Indian—that all exhibit a null allele at *DYS390.* A second distinct group was formed, consisting of 18 samples—16 African and two Coloured—with each containing a microvariant allele, although this group is clustered in closer proximity to the rest of the samples. There is no clear correlation between the population group of the samples and the profile variation, so the grouping is random in terms of population.

In order to evaluate the distinction of samples without the influence of profile variations, a second P-CoA was done, excluding all samples that had any form of profile variation: null alleles; duplications at loci except *DYS385* or *DYF387S1*; triallelic patterns; and microvariants. The second P-CoA was, therefore, performed using 230 profiles: 11 Asian/Indian, 81 African, 36 Coloured, and 102 Caucasian samples. The result from this second P-CoA is shown in Figure 4.4.

Despite the two distinct clusters of variant profiles in Figure 4.3, there is still a separation of the African and Caucasian populations, with the Asian/Indian samples clustered in the middle and the Coloured samples scattered across both groups. This distinction of the populations becomes more clear when excluding the variant samples, as evidenced in Figure 4.4. This figure clearly shows a polarisation of the African and Caucasian populations. There is an overlap between the two groups, with two African samples clustering in with some Caucasian sample. However, it is minimal. Although there is a higher number of Coloured samples clustering with the Caucasian population, the Coloured population is clearly distributed across both the African and Caucasian populations, reiterating the levels of admixture in this population..

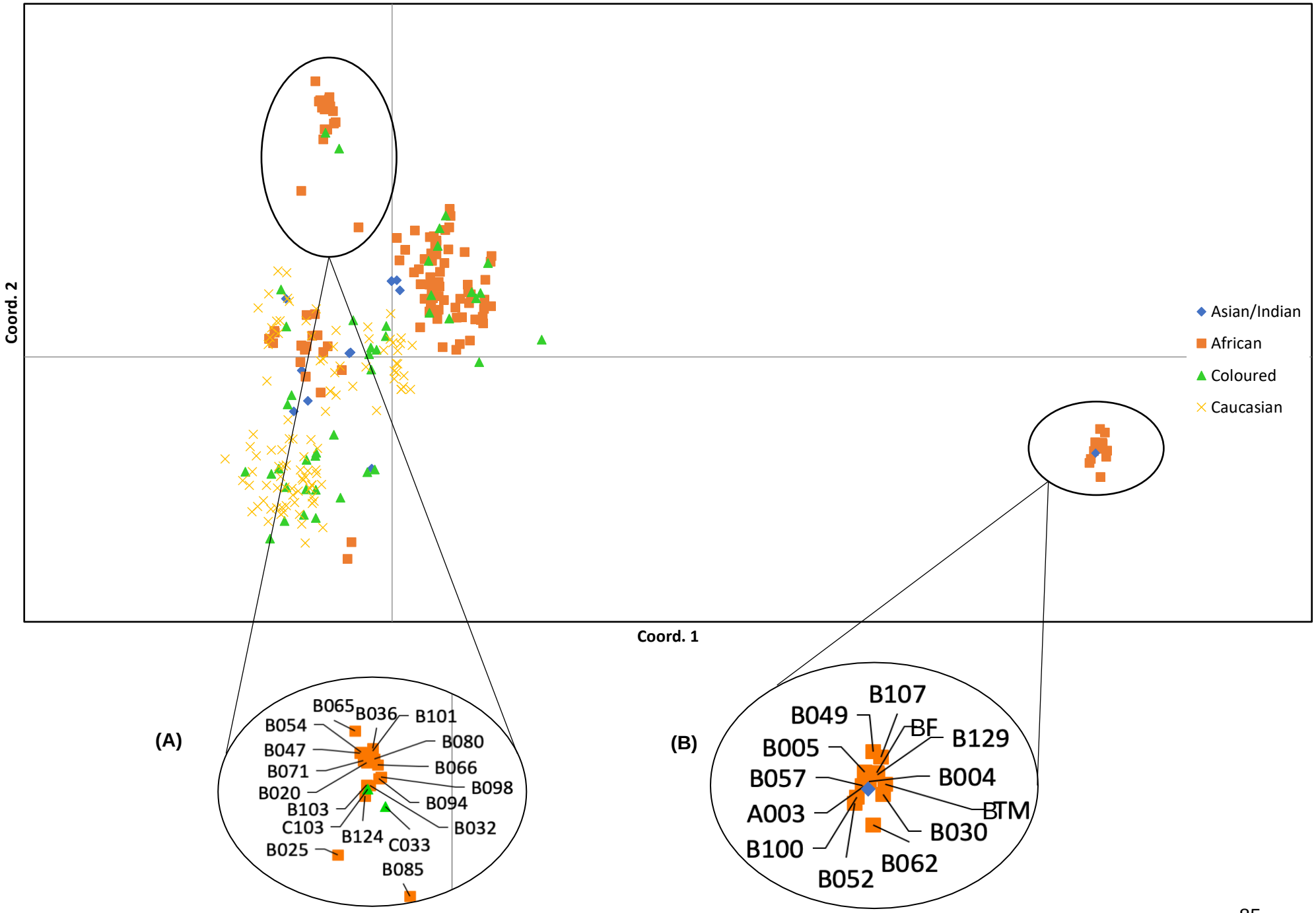


Figure 4.3: *The P-CoA graph showing the grouping of samples, when all the samples are included. Samples that distinctly grouped together are enlarged. (A) All these samples had microvariant alleles (B) All these samples had null alleles at DYS390.*

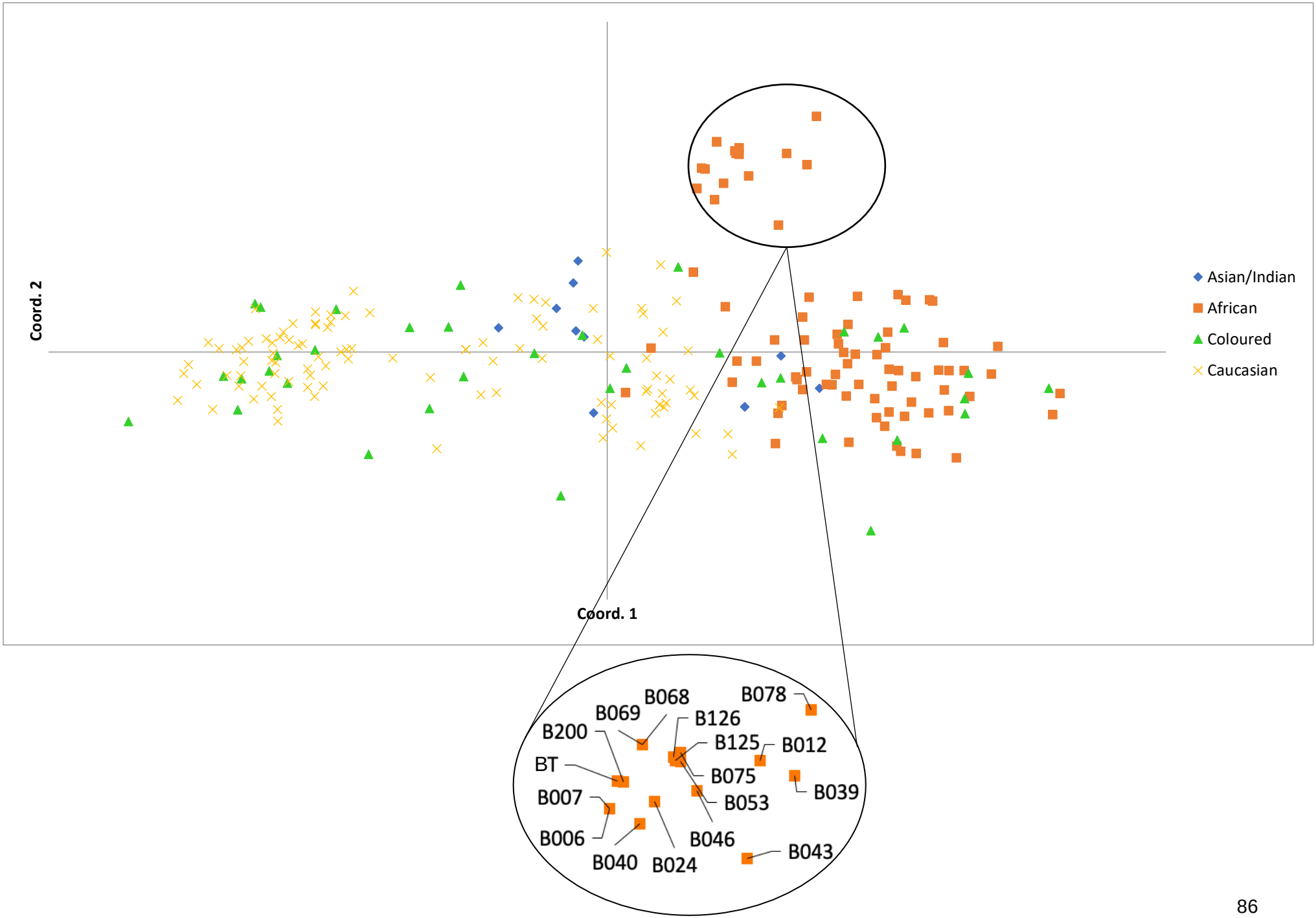


Figure 4.4: *The P-CoA graph showing the grouping of samples, where all samples with any form of profile variation are excluded from analysis. A seemingly random grouping of samples is enlarged.*

This clear distinction between population groups demonstrates the efficiency of the 27 Y-Filer® Plus markers in differentiating between the four main population groups in South Africa, and, as suggested by Kayser (2017), this is highly advantageous from a forensic perspective There is, however, a group of 18 African samples that weakly cluster together away from the main African group. This appears to be a random group of samples clustering together, as none of these samples exhibited any form of profile variation. All these individuals self-identified as being either Xhosa, Sotho, Zulu, Venda, or Shona and indicated that their fathers were of the same ethnicity. These ethnic groups are also scattered across South Africa geographically (Lane *et al.,* 2002), which would restrict gene flow between them. It is a possibility that there has been some gene flow between these groups, because people migrate and interact more frequently than they used to. It is, therefore, certainly not uncommon to marry outside one's ethnic group. Since this cluster consists of only a few samples (7%) out of the whole sample population, it is more likely that this random group clustered together simply by chance. However, it is also possible that the participants identified themselves as belonging to a specific ethnic group, but that their true genetic background is not reflecting this.

4.3.3 Profile Variations

A standard Y-STR profile will have only one allele detected per locus, given the haploid nature of the Y-chromosome. An exception to this is the two markers *DYS385* and *DYF387S1*, which are multicopy loci, and so two alleles are frequently detected at those markers. Other variations to the standard profile include null alleles, duplications, triplications, and microvariant alleles. In this study, a total of 44 samples deviated from the standard profile and exhibited some form of profile variation.

4.3.3.1 Null Alleles

Several null alleles were detected at two different loci. Four null alleles were seen at *DYS448* (Figure 4.5) and 13 at *DYS390* (Figure 4.6). A total of 729 null alleles at *DYS448* have been recorded on the Y-chromosome Haplotype Reference Database (YHRD), whereas there have been only six reported for *DYS390* across all populations included in the YHRD (www.yhrd.org; accessed 09 October 2020). This difference is interesting to note, as it would then be thought that null alleles would be more common at the *DYS448* marker. However, in this study, there were more null alleles observed at *DYS390* than *DYS448*.

Of the four null alleles at *DYS448*, two were in African population samples (Sotho and Zulu individuals) and two were in Coloured population samples. One of the null alleles detected at *DYS390* was an Asian/Indian population sample and the other 12 belonged to the African subpopulations Zulu, Sotho, Tsonga, Xhosa, and Pedi. No null alleles were detected in the Caucasian population.

There have been reports of null alleles at both loci in several populations around the world, such as South Korea, the USA, and Canada (https://strbase.nist.gov; accessed 09 October 2020). There does not appear to be any correlation or pattern regarding the population groups within South Africa in which the null alleles are found. Although it could be said that the null alleles are more common in the African population subgroups in this study, the small sample sizes prevent drawing definitive conclusions regarding the population-specific prevalence of null alleles within the South African population.

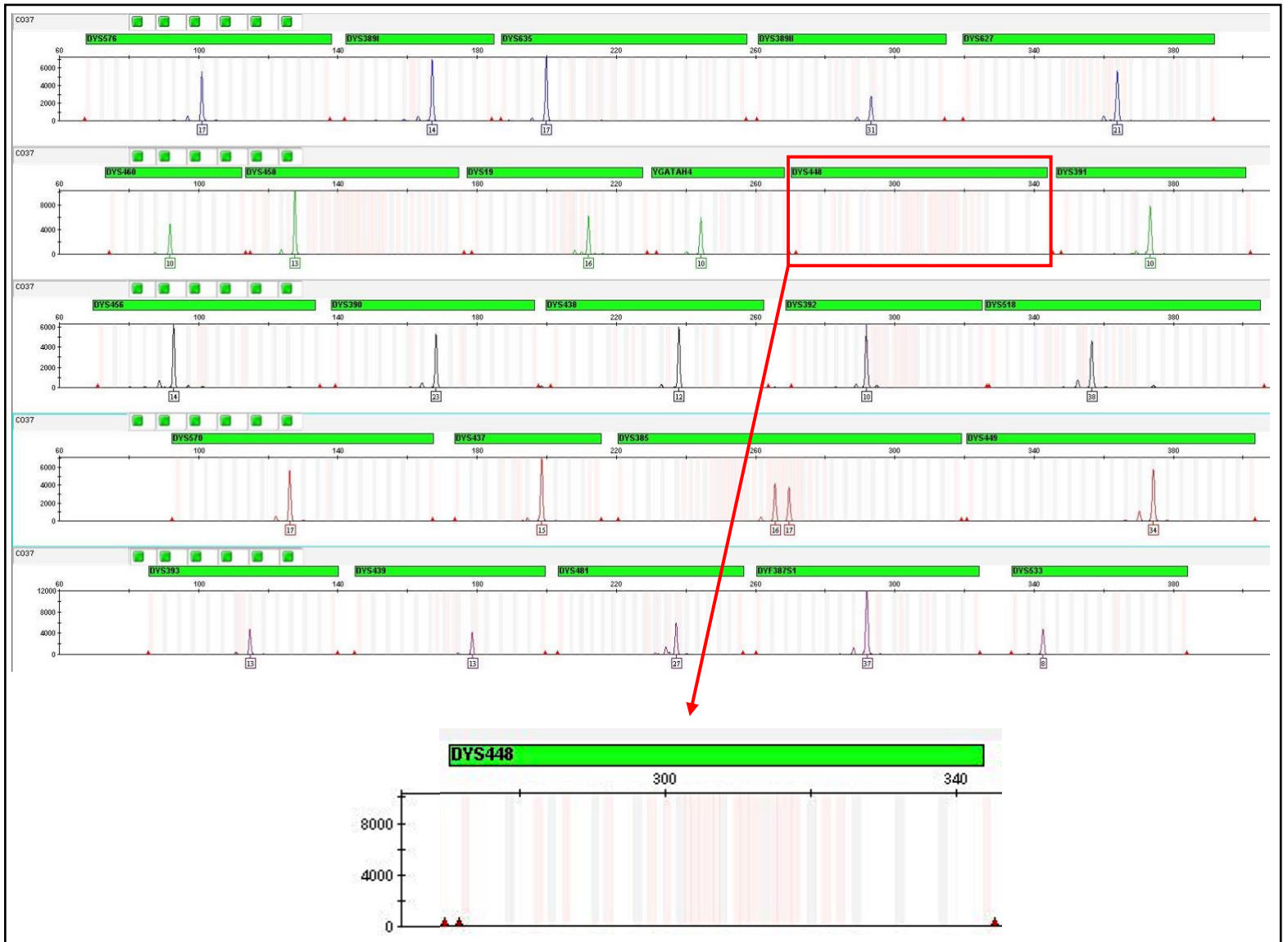


Figure 4.5: *The null allele detected at DYS448.*

Without performing DNA sequencing at this locus to assess the null allele on a sequence level, one could only speculate as to why the locus failed to amplify. Some kind of sequence polymorphism would have taken place either within the marker region on the template DNA, in the primer binding region, or in one of the flanking regions on either side of the primer binding site (Butler, 2010). This polymorphism could be an indel, resulting in a frameshift, or a point mutation which alters the sequence. It has been reported that the null allele at *DYS448* has occurred as a result of large locus deletions, as well as several variations at the primer binding sites (Balaresque *et al*., 2008; Budowle *et al.,* 2008). Sequence data for the variation resulting in the null allele *DYS390* is not readily available. DNA sequencing was beyond the scope of this study, and thus not performed on any samples that exhibited any form of variation. It is, therefore, recommended that the *DYS448* and *DYS390* loci in South African samples exhibiting null alleles be sequenced to determine if the mechanism for allelic dropout in this population is similar to those already reported on.

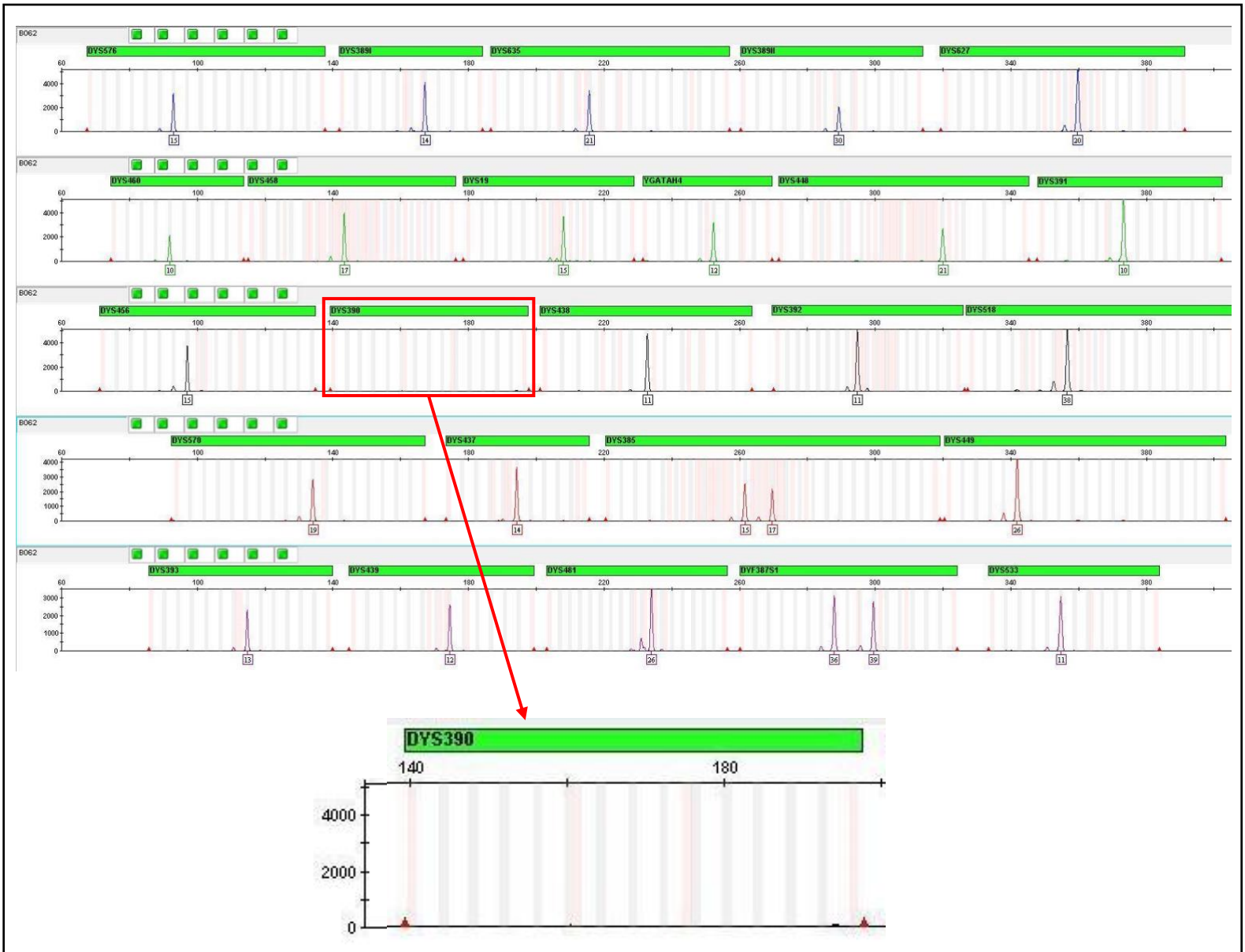


Figure 4.6: *The null allele detected at DYS390.*

4.3.3.2 Duplications

Markers *DYS385* and *DYF387S1* are multicopy loci and many samples in this study were found to have two alleles at one or both of these markers. However, one individual belonging to the Sotho ethnic group, exhibited two alleles at both *DYS385* and *DYF387S1*, as well as duplications at two additional markers, *DYS458* and *DYS449*, indicated in Figure 4.7. It has been noted that most duplicated loci would exhibit two alleles that are a single repeat unit apart, following the single step mutation model (Butler *et al.,* 2005). The duplications detected in this study are consistent with this, with alleles 16 and 17 observed at *DYS458* and alleles 28 and 29 at *DYS449.*

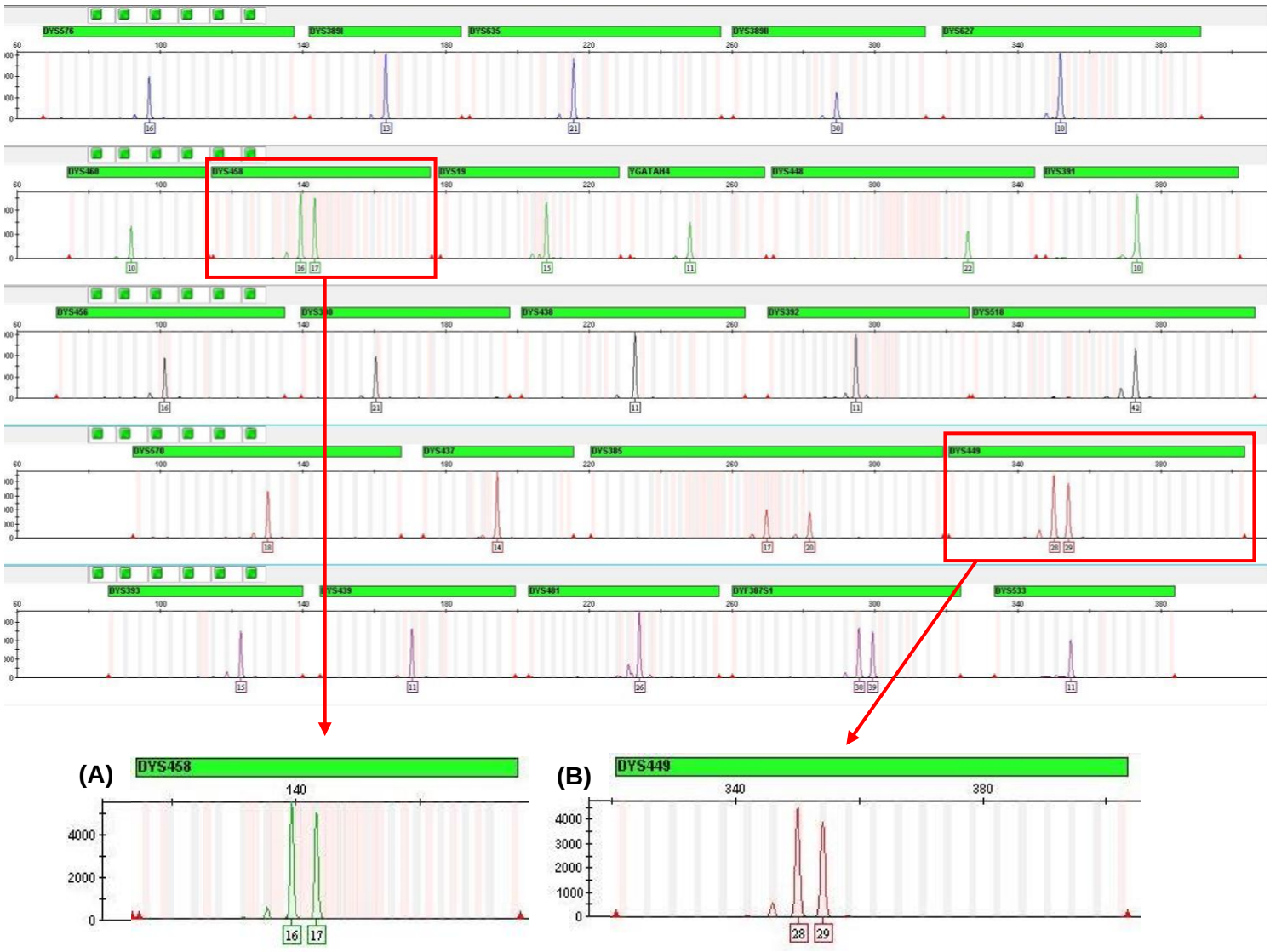


Figure 4.7: *The duplications detected at (A) DYS458 and (B) DYS449.*

As with null alleles, only the end result of the variation is detected during genotyping, Therefore, without performing sequencing, it is impossible to know exactly where and how the duplication event took place to result in two different alleles. It has been suggested by Butler (2005) that there are many regions of the Y-chromosome with duplication potential, and that markers that are located in close proximity to one another would be duplicated together, resulting in multiple duplicated loci in the same DNA sample. The *DYS458* and *DYS449* loci are located close together on the p arm of the Y-chromosome, at positions 7.833 Mb and 8.183 Mb respectively. Based on this, it is likely that a duplication of that whole region on the p arm occurred.

Based on the current knowledge, there has been no indication of duplications at these two markers in the South African population. The most recent update on the Short Tandem Repeat DNA Internet DataBase (STRBase; April 2016) shows that duplications at these two markers are not common (National Institute of Standards and Technology (NIST); https://strbase.nist.gov). STRBase does not have any reports of a duplication at *DYS449*. However, a study conducted in the United Kingdom also reported on two samples that exhibited simultaneous duplication of *DYS458* and *DYS449* (Aliferi *et al.,* 2018).

4.3.3.3 Triplications

In addition to duplications that can occur at specific loci, triplications are also a possibility. These triplications can be presented in two different types of patterns (NIST; https://strbase.nist.gov). The first pattern has three allelic peaks where the sum of the height of two peaks equals the height of the third, while the second pattern results in three allelic peaks of approximately the same height. Figure 4.8 depicts these two triallelic patterns that were detected at *DYF387S1* in this study.

In this study, five triplications were detected at *DYF387S1.* Three of these triplications occurred in African population samples (Sotho and Pedi subgroups) and two in Coloured population samples. Of these five triplications, two exhibited the first pattern with the height of two peaks equalling the height of the third, and three samples exhibited three allelic peaks of approximately the same height. The five triplications included the following alleles in combination: (A) 35, 36, 40; (B) 36, 37, 39; (C) 36, 39, 40; (D) 36, 39, 41; and (E) 36, 37, 40. Four out of the five triplications resulted in two alleles that are a single repeat unit apart, keeping in line with the single step mutation model suggested by Butler *et al. (*2005). It is one of these alleles that most likely underwent a second duplication and mutational event to result in the two alleles, while the third peak (more than one repeat unit apart) exists simply because

DYF387S1 is multicopy. In combination D, with alleles 36, 39, and 41, it is possible that either 39 or 41 duplicated and underwent a two-step mutation to result in alleles that are two repeat units apart.

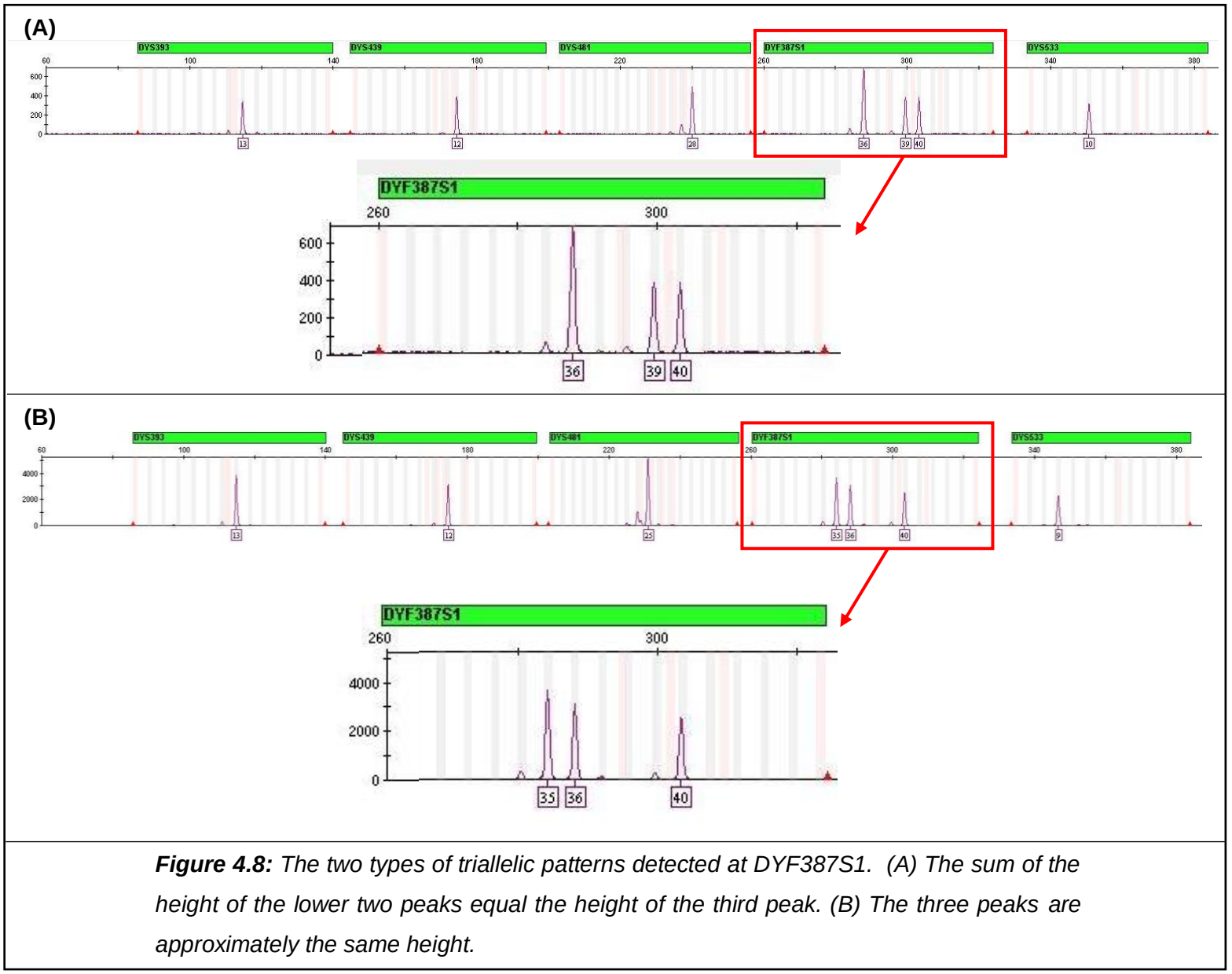


Figure 4.8: *The two types of triallelic patterns detected at DYF387S1. (A) The sum of the height of the lower two peaks equal the height of the third peak. (B) The three peaks are approximately the same height.*

Reports of triallelic patterns occurring at *DYF387S1* are becoming more frequent (NIST; https://strbase.nist.gov). As of November 2017, there have been nine reports of triplications at *DYF387S1*, with the combination of alleles 37, 38, and 39 being the most common. However, this update is not that recent. Not all variants are uploaded to the database, and so it is likely that more triplications have been observed and reported. Despite the lack of records on the STRBase website, triplications at *DYF387S1* are being reported in populations around the world (Gettings, 2014; Olofsson *et al.,* 2015; Iacovacci *et al.,* 2017; Fan *et al,* 2018; Khubrani *et al.,* 2018). Interestingly, none of the combinations of alleles detected in the

triplications in this study have been reported. Based on the current knowledge, there has been no indication of triplications at this marker in the South African population, but the *DYF387S1* locus has yet to be fully investigated in this population.

4.3.3.4 Intermediate alleles

Five different microvariant alleles were detected in 21 samples in this study. Sequencing was beyond the scope of this study, so the microvariants presented here are based on the calculations detailed in Appendix F (Page 140) and are only an estimate of the allele. It is thus recommended that samples containing such intermediate alleles be sequenced in order to conclusively determine the length of the incomplete repeat unit. As with all the other profile variations discussed in the chapter, the inclusion of intermediate alleles in forensic DNA testing has the potential to increase haplotype diversity: especially if certain microvariants are population specific. The confidence in a match between a suspect reference sample and a casework sample could be increased with the presence of an intermediate allele, and this could be advantageous in forensic DNA testing in South Africa.

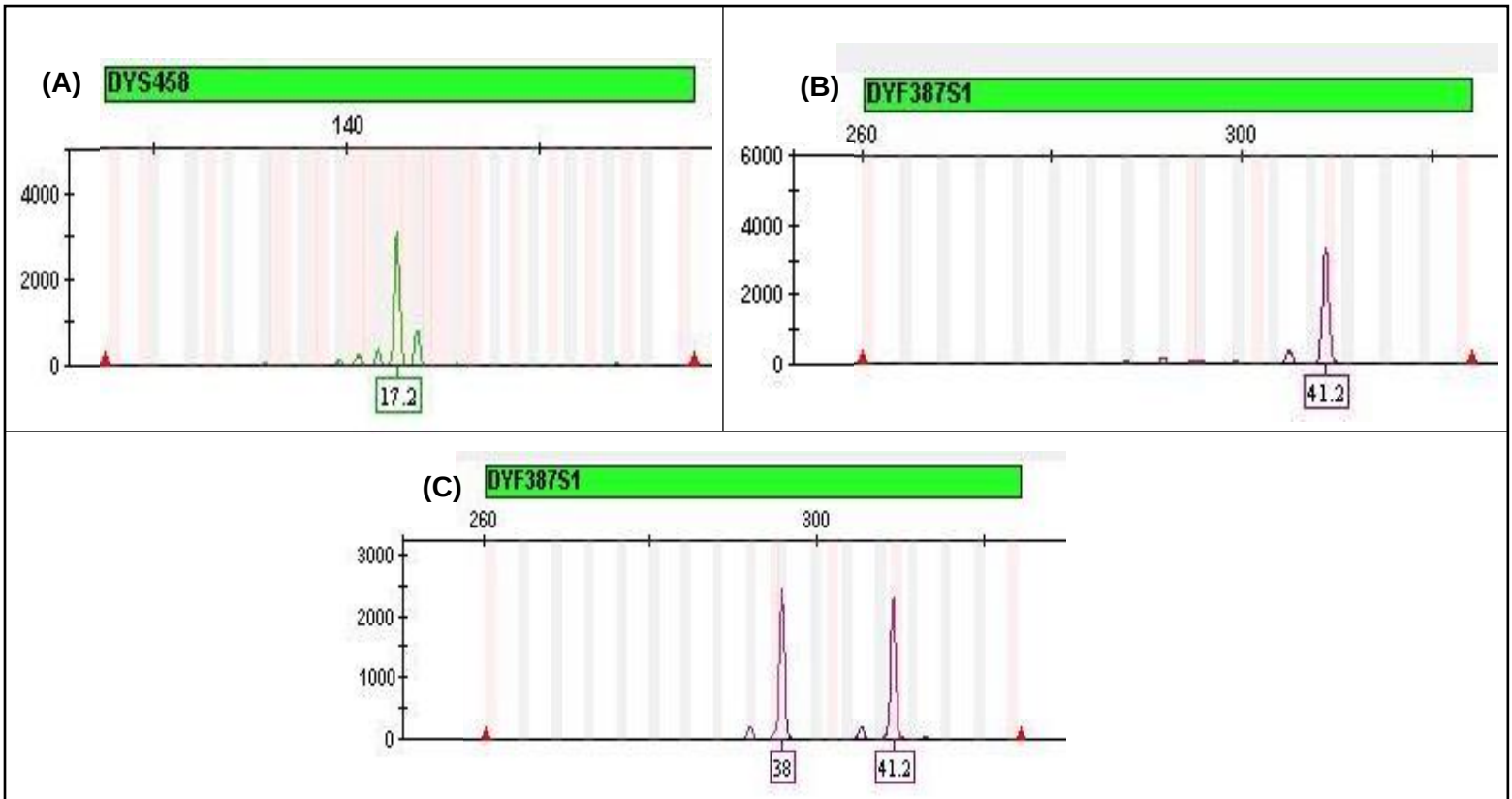


Figure 4.9: *The microvariants detected. (A) 17.2 at DYS458, (B) 41.2 at DYF387S1 occurring as a single allele, and (C) 41.2 at DYF387S1 occurring in a duplication.*

GeneMapper™ ID-X Software (*Applied Biosystems*) is able to automatically detect and label certain microvariant alleles, as it would any other allele. Two microvariant alleles in this study were detected this way: 17.2 at *DYS458* and 41.2 at *DYF387S1.* The 17.2 allele occurred once in a Caucasian sample, and the 41.2 allele was detected in 16 individuals of which 15 were African (Zulu, Sotho, Xhosa, and Swati) and one was a Coloured individual. There did not appear to be any pattern as to which population groups the variant alleles occurred in. The 41.2 allele at *DYF387S1* occurred as a single allele once and as part of a duplication 15 times. Figure 4.9 illustrates these variant alleles detected by GeneMapper™ ID-X during profile analysis.

Some of the other variant alleles were initially detected as off-ladder (OL) alleles by the GeneMapper™ ID-X Software, but they did appear to be true alleles and not artefacts. Samples with these OL alleles were reamplified and genotyped using the duplicate swabs and all OLs were confirmed to be true alleles. These alleles were calculated using the size of the alleles in conjunction with the size of the nearest allele in the allelic ladder (Appendix F, Page 140). Figure 4.10 shows these alleles that were detected as OL alleles at the two markers.

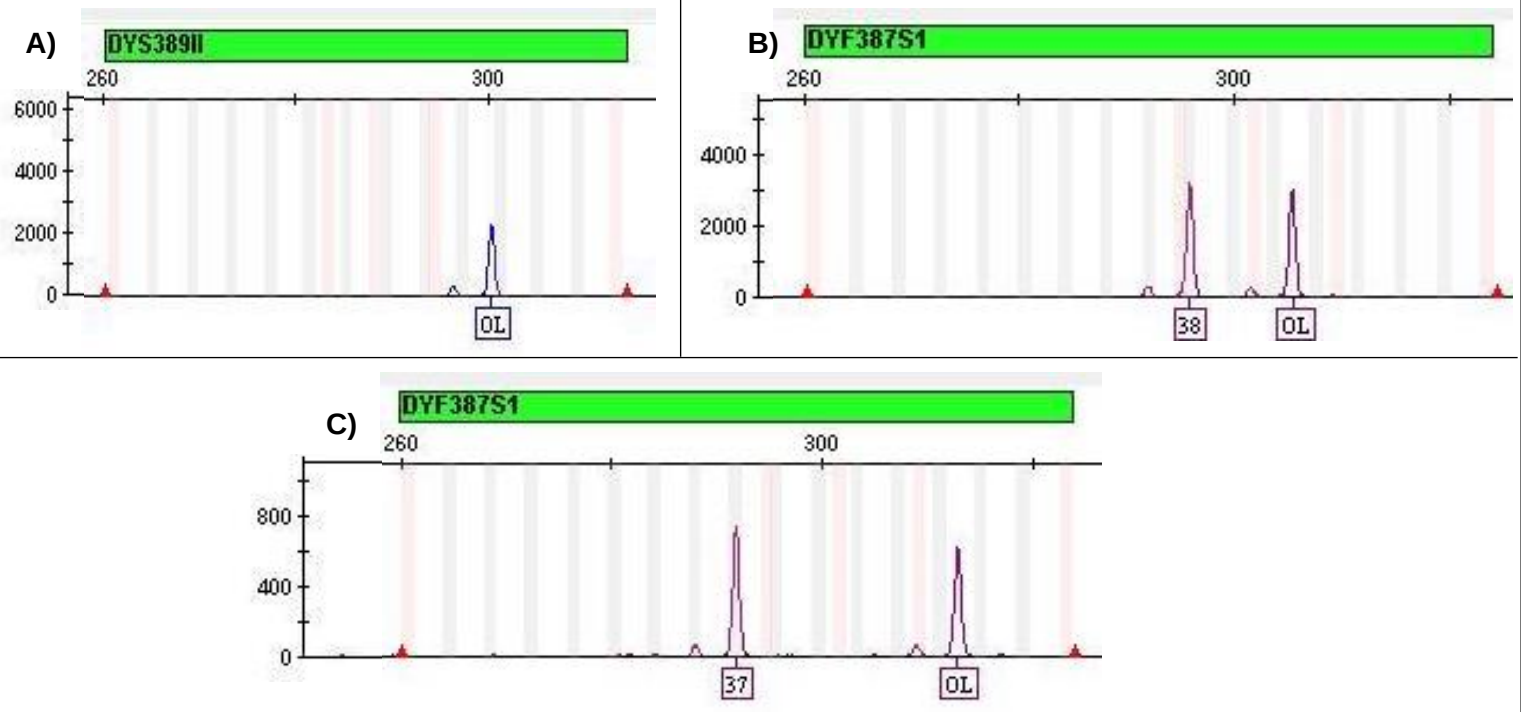


Figure 4.10: *These microvariants were not included in the virtual bin set and were therefore detected as off-ladder (OL) alleles. The value of the OL was calculated and the microvariant used in statistical analyses. A) 32.3 at DYS389II, B) 40.2 at DYF387S1 and C) 42.2 at DYF387S1.*

Allele 32.3 at *DYS389II* was detected in one African (Zulu) population sample, allele 40.2 at *DYF387S1* belonged to a Coloured population sample, and allele 42.2 at *DYF387S1* occurred in an African (Venda) population sample.

Occasionally a microvariant allele can actually fall outside of the marker size range, resulting in an off marker range (OMR) allele. Even though it may not be allocated to a specific locus, the OMR allele could still be a true allele instead of being an artefact. An OMR allele was detected between *DYF387S1 and DYS533* in one of the Coloured population samples (Figure 4.11). This sample was reamplified and genotyped using the duplicate swab and the OMR allele was confirmed. It is more likely that this allele had occurred as a part of the *DYF387S1* marker than *DYS533* and was, therefore, calculated to be allele 45.2 at DYF387S1 (detailed in Appendix F, Page 140). It is clear in Figure 4.11 that visually the OMR allele was closer to *DYF387S1.* If it was allocated to *DYS533*, it would be calculated to be allele 4.3, which is well below the smallest observed allele, allele 6, at this locus (www.yhrd.com). There are also reports of a 45.2 variant allele being observed at *DYF387S1* (www.yhrd.com). Despite this speculation and justification, only sequencing of this sample would be able to definitively conclude which marker this allele would fall under and how many repeats (full and partial) it would have.

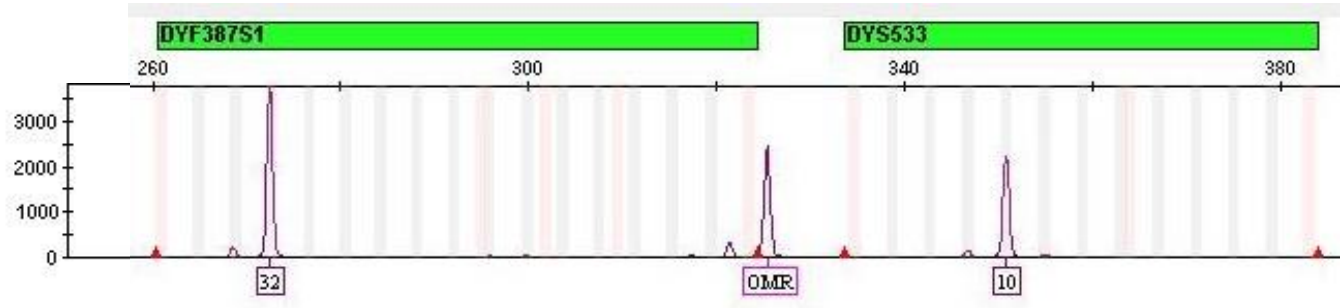


Figure 4.11: *The off marker range (OMR) allele detected between DYF387S1 and DYS533. This allele was calculated to be the microvariant 45.2 at DYF387S1.*

All these intermediate alleles, except for 32.3 at *DYS389II,* have been observed before and reported on various platforms (NIST; https://strbase.nist.gov; www.yhrd.org). *DYF387S1* has not yet been fully investigated in South Africa, so no comparisons can be made for the microvariants detected at this marker. While some intermediate alleles have been detected at various loci in the South African population (Ehrenreich, 2005), the specific microvariants detected in this study have not been reported on in South Africa.

Several microvariant alleles have been characterised at *DYS458* and *DYF387S1* in the Eastern and Northern African populations (Iacovacci *et al.,* 2017; D'Atanasio *et al.,* 2019). A range of alleles from 16.2 to 20.2 were detected and sequenced at *DYS458. DYS458* is a tetranucleotide marker and the repeat motif is $[GAAA]_n$ (NIST; https://strbase.nist.gov; www.yhrd.org). Iacovacci *et al.* (2017) reported that the incomplete repeat unit occurred due to a dinucleotide indel within this repeat motif. Because this sequence polymorphism occurs within the [GAAA] motif, it is not clear if it is a GA deletion or AA insertion (Myres *et al.,* 2007). The tetranucleotide motif of *DYF387S1* is $[AAAG]_n$, and it has been reported that the .2 incomplete repeat arises due to a two base indel on the 5' end of the last full repeat unit (Iacovacci *et al.,* 2017; D'Atanasio *et al.,* 2019). Future studies should use sequencing to confirm these polymorphisms in the South African population and to investigate the cause of the other microvariant alleles observed in this study.

4.3.3.5 Recommendation for future studies: Sequencing of profile variations in South Africa

Y-STR genotyping only reveals the end result of DNA polymorphisms on the allelic level. Null alleles are presented as the absence of an allele, duplications as two alleles, triplications as three alleles, and intermediate alleles as alleles falling between the allelic bins in the GeneMapper™ ID-X Software. All of these variations arise from some kind of change at the molecular level: a change in the DNA sequence. Without the use of sequencing to analyse these changes, one can only speculate as to what caused the variant allele. DNA sequencing would reveal if the polymorphism causing the variation was an indel, a frameshift from an indel elsewhere, a base pair mutation, duplication of a single locus, or duplication of a whole region.

Some of the variations detected in this study have been sequenced and characterised in other populations around the world. The null allele at *DYS448* has been said to occur due to large locus deletions, as well as several variations at the primer binding sites (Balaresque *et al*., 2008; Budowle *et al.,* 2008). There is limited information available regarding the null allele at *DYS390.* The simultaneous duplication at two closely located markers—*DYS458* and *DYS449*—suggests that the whole region was duplicated. However, it is possible that there were two independent duplication events. Triplications could result from a second duplication of an already duplicated allele, or the duplication of an entire region containing the two different alleles (Butler *et al.,* 2005). The latter is more likely the cause of the triplications observed at *DYF387S1* in this study, given the multicopy property of this marker. Only DNA sequencing could confirm any of these speculations. Iacovacci *et al.* (2017) and D'Atanasio *et al.* (2019) detail the polymorphisms that resulted in the intermediate alleles at *DYS458* and *DYF387S1* in the Eastern and Northern populations, respectively.

It is certainly possible that these would be the same polymorphisms causing the variant alleles observed in the South African population in this study. However, genetic drift and chance mutations could mean that the sequence variations in the South African population might differ from other previous reported cases. Future studies using the 27 Y-Filer® Plus loci should consider sequencing any detected profile variations to gain a better understanding of the molecular mechanisms behind these variations in the South African population.

4.4 Conclusion

This study showed that while there is a weak correlation between the number of different alleles detected and the mean gene diversity in a population. Furthermore, there is no clear correlation between the number of private alleles in a population and the gene diversity. This is evident with the African population having the highest number of private alleles while having the lowest mean gene diversity. It could then be said that it is not necessarily which alleles (private or shared) that are detected that contributes to gene diversity, but rather the number of different alleles.

The findings of this study further showed that the Y-Filer® Plus DNA testing kit is able to successfully detect significant genetic differentiation between the Asian/Indian, African, Coloured, and Caucasian groups in South Africa. As suggested by Kayser (2017), the ability of a commercial testing kit to distinguish between population groups is highly advantageous in a forensic setting. The Y-Filer® Plus kit was, however, not able to detect significant variance between some of the African subgroups, which is in contrast to previous findings (D'Amato and Kasu, 2017). Despite this, it would be premature, and incorrect, to make conclusions regarding this based on this analysis using the smaller sample sizes for each African subgroup ($n < 100$). However, it can then be recommended that future studies conducted using the Y-Filer® Plus testing kit make use of larger sample sizes for the specific population subgroups in order to draw accurate conclusions.

In this study, 44 samples in total were identified as deviating from the standard profile and exhibiting some form of profile variation. Of these 44 variations, there were 17 null alleles, two duplications at unexpected loci, five triplications, and 21 intermediate alleles identified. There was no correlation between the variations detected and the population groups in which they were observed. Sequencing of these variations was beyond the scope of this study. Therefore, it is recommended that future studies based on the South African population using the 27 Y-Filer® Plus loci should focus on generating sequence data for such variations to gain better insights to the sequence variance within this population.

References

Aliferi, A., Thomson, J., McDonald, A., Paynter, V.M., Ferguson, S., Vanhinsbergh, D., Court, D.S., and Ballard, D. (2018) 'UK and Irish Y-STR population data – A catalogue of variant alleles'. *Forensic Science International: Genetics*, 34, pp.e1-e6.

Balaresque, P., Bowden, G.R., Parkin, E.J., Omran, G.A., Heyer, E., Quintana-Murci, L., Roewer, L., Stoneking, M., Nasidze, I., Carvalho-Silva, D.R., and Tyler-Smith, C. (2008) 'Dynamic nature of the proximal AZFc region of the human Y chromosome: multiple independent deletion and duplication events revealed by microsatellite analysis'. *Human mutation*, 29(10), pp.1171-1180.

Bender, K., Farfán, M.J., and Schneider, P.M. (2004) 'Preparation of degraded human DNA under controlled conditions'. *Forensic Science International*, 139, pp.135-140.

Budowle, B., Aranda, X. G., Lagace, R. E., Hennessy, L. K., Planz, J. V., Rodriguez, M., and Eisenberg, A. J. (2008) 'Null allele sequence structure at the DYS448 locus and implications for profile interpretation'. *International Journal of Legal Medicine*, 122(5), 421-427.

Butler, J. M. (2005) 'Microvariants and 'Off-Ladder' Alleles'. In: *Forensic DNA Typing: Biology, Technology, And Genetics of STR Markers*. 2nd ed. Elsevier, pp.130-131.

Butler, J.M. (2010) 'STR Genotyping and Data Interpretation'. In: J. Butler, ed., *Fundamentals of Forensic DNA Typing*, 1st ed. Academic Press: Elsevier, pp.205-228.

Butler, J.M. (2012) 'Y-Chromosome DNA Testing'. In: J. Butler, ed., *Advanced Topics in Forensic DNA Typing: Methodology*, 1st ed. Academic Press: Elsevier, pp.371-403.

Butler, J.M., Decker, A.E., Kline, M.C., and Vallone, P.M. (2005) 'Chromosomal duplications along the Y-chromosome and their potential impact on Y-STR interpretation'. *Journal of Forensic Science*, 50(4), pp.853-859.

Chakraborty, R. and Deka, R. (2009) 'DNA Forensics: A Population Genetic and Biological Anthropological Perspective'. In: P. Rudan, ed., *Physical (Biological) Anthropology*. EOLSS Publications, pp.216-245.

D'Amato, M.E., Ehrenreich, L., Benjeddou, M., Davison, S. and Leat, N. (2008) 'Ancestry and genetic relationships between groups within the Cape Town metropolitan population inferred using Y-STRs genotyping'. *Forensic Science International: Genetics Supplement Series*, 1(1), pp.318-319.

D'Amato, M. E., and Kasu, M. (2017) 'Population analysis of African Y-STR profiles with UniQ Typer™ Y-10 genotyping system'. *Forensic Science International: Genetics Supplement Series*, 6(October), pp.e84-e85.

D'Atanasio, E., Iacovacci, G., Pistillo, R., Bonito, M., Dugoujon, J.M., Moral, P., El-Chennawi, F., Melhaoui, M., Baali, A., Cherkaoui, M., and Sellitto, D. (2019) 'Rapidly mutating Y-STRs in rapidly expanding populations: Discrimination power of the Yfiler Plus multiplex in northern Africa'. *Forensic Science International: Genetics*, 38, pp.185-194.

de Wit, E., Delport, W., Rugamika, C.E., Meintjes, A., Möller, M., van Helden, P.D., Seoighe, C., and Hoal, E.G. (2010). Genome-wide analysis of the structure of the South African Coloured Population in the Western Cape'. *Human genetics*, 128(2), pp.145-153.

Ehrenreich, L. S. (2005) '*The evaluation of Y-STR loci for use in Forensics'*. Masters. University of the Western Cape.

Fan, H., Wang, X., Chen, H., Zhang, X., Huang, P., Long, R., Liang, A., Song, T., and Deng, J. (2018) 'Population analysis of 27 Y-chromosomal STRs in the Li ethnic minority from Hainan province, southernmost China'. *Forensic Science International: Genetics*, 34, pp.e20-e22.

Frankham, R., Ballou, J., and Briscoe, D. eds. (2002) 'Population Fragmentation'. In: *Introduction to Conservation Genetics*, 1st ed. Cambridge, United Kingdom: Cambridge University Press, pp.309-335.

Gettings, K.B. (2014) '*Y-Filer® Plus Kit – Improved Haplotype Discrimination using "Rapidly Mutating" Y-STR Markers in a Large Multiplex Kit*' Presentation. Future Trends in Forensic DNA Technology, NIST, USA.

Hammer, M.F., Chamberlain, V.F., Kearney, V.F., Stover, D., Zhang, G., Karafet, T., Walsh, B. and Redd, A.J. (2006) 'Population structure of Y chromosome SNP haplogroups in

the United States and forensic implications for constructing Y chromosome STR databases'. *Forensic Science International*, 164(1), pp.45-55.

Hartl, D.L. and Clark, A.G. eds. (1997) 'Inbreeding, Population Subdivision, And Migration'. In: *Principles of population genetics*. Sunderland, Massachusetts: Sinauer Associates, Inc. Publishers, pp. 275-287.

Iacovacci, G., D'Atanasio, E., Marini, O., Coppa, A., Sellitto, D., Trombetta, B., Berti, A., and Cruciani, F. (2017) 'Forensic data and microvariant sequence characterization of 27 Y-STR loci analysed in four Eastern African countries'. *Forensic Science International: Genetics*, 27, pp.123-131.

Kayser, M. (2017) 'Forensic use of Y-chromosome DNA: a general overview'. *Human Genetics*, 136(5), pp.621-635.

Kline, M.C., Hill, C.R., Decker, A.E., and Butler, J.M. (2011) 'STR sequence analysis for characterizing normal, variant, and null alleles'. *Forensic Science International: Genetics*, 5(4), pp.329-332.

Khubrani, Y.M., Wetton, J.H., and Jobling, M.A. (2018) 'Extensive geographical and social structure in the paternal lineages of Saudi Arabia revealed by analysis of 27 Y-STRs'. *Forensic Science International: Genetics*, 33, pp.98-105.

Lane, A.B., Soodyall, H., Arndt, S., Ratshikhopha, M.E., Jonker, E., Freeman, C., Young, L., Morar, B., and Toffie, L. (2002) 'Genetic substructure in South African Bantu-speakers: Evidence from autosomal DNA and Y-chromosome studies'. *American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists*, 119(2), pp.175-185.

Lee, H. C., and Ladd, C. (2001) 'Preservation and Collection of Biological Evidence'. *Croatian Medical Journal*, 42(3), pp.225-228.

Myres, N.M., Ekins, J.E., Lin, A.A., Cavalli-Sforza, L.L., Woodward, S.R., and Underhill, P.A. (2007) 'Y-chromosome short tandem repeat DYS458. 2 non-consensus alleles occur independently in both binary haplogroups J1-M267 and R1b3-M405'. *Croatian Medical Journal*, 48(4.), pp.450-459.

Nei, M. (1978). 'Estimation of average heterozygosity and genetic distance from a small number of individuals'. *Genetics*, 89(3), pp.583-590.

Olofsson, J.K., Mogensen, H.S., Buchard, A., Børsting, C., and Morling. N. (2015) 'Forensic and population genetic analyses of Danes, Greenlanders and Somalis typed with the Yfiler® Plus PCR amplification kit'. *Forensic Science International: Genetics*,16, pp.232–236.

Peakall, R. and Smouse, P.E. (2006) 'GenAlEx 6: genetic analysis in Excel. Population genetic software for teaching and research'. *Molecular Ecology Notes*, 6, pp.288-295.

Peakall, R. and Smouse, P.E. (2012) 'GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research – an update'. *Bioinformatics*, 28, 2537-2539.

Purps, J., Siegert, S., Willuweit, S., Nagy, M., Alves, C., Salazar, R., Angustia, S.M., Santos, L.H., Anslinger, K., Bayer, B. and Ayub, Q. (2014) 'A global analysis of Y-chromosomal haplotype diversity for 23 STR loci'. *Forensic Science International: Genetics*, 12, pp.12-23.

Quintana-Murci, L., Harmant, C., Quach, H., Balanovsky, O., Zaporozhchenko, V., Bormans, C., van Helden, P.D., Hoal, E.G. and Behar, D.M. (2010) 'Strong maternal Khoisan contribution to the South African coloured population: a case of gender-biased admixture'. *The American Journal of Human Genetics*, 86(4), pp.611-620.

Slatkin, M. (1995) 'A measure of population subdivision based on microsatellite allele frequencies'. *Genetics*, 139(1), pp.457-462.

South African History Online (2010) '*From Bondage To Freedom - The 150Th Anniversary Of The Arrival Of Indian Workers In South Africa*'. [Online] Available at: https://www.sahistory.org.za/article/indian-south-africans. [Accessed 15 May 2020].

Statistics South Africa. (2020) '*Mid-year population estimates 2020*'. [Online] Available at: http://www.statssa.gov.za/publications/P0302/P03022020.pdf. [Accessed 06 Aug. 2020].

Szpiech, Z.A. and Rosenberg, N.A. (2011) 'On the size distribution of private microsatellite alleles'. *Theoretical population biology*, 80(2), pp.100-113.

Tishkoff, S.A., Reed, F.A., Friedlaender, F.R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J.B., Awomoyi, A.A., Bodo, J.M., Doumbo, O., and Ibrahim, M. (2009) 'The genetic structure and history of Africans and African Americans'. *Science*, 324(5930), pp.1035-1044.

ThermoFisher Scientific. (2019) *Yfiler™ Plus PCR Amplification Kit: User Guide*. Woolston: ThermoFisher Scientific. [Online]. Available at: https://assets.thermofisher.com/TFS-Assets/LSG/manuals/4485610_YfilerPlus_UG.pdf. [Accessed 30 Jan. 2020]

www.yhrd.org

https://strbase.nist.gov

CHAPTER 5: CONCLUDING REMARKS AND RECOMMENDATIONS FOR FUTURE RESEARCH

5.1 Summary

The use of Y-STRs in forensic DNA testing excludes any female (victim) DNA, simplifying the analysis of mixture samples that are often a consequence of sexually based crimes (Roewer, 2009). However, Y-STR technology is currently not employed in South Africa's forensic laboratories due to reduced discrimination capacities of some commercially available testing kits, a lack of reference population data, and more recent commercial testing kits not yet being investigated in the South African population. The purpose of this study was to investigate the forensic potential of a commercial Y-STR kit developed using the American population, and directing this towards a South African population application. Previous research revealed that the African population exhibited low levels of gene diversity at some core Y-STR loci used in commercial testing kits, resulting in the reduced discrimination capabilities of these kits in this population (D'Amato *et al.,* 2008; D'Amato *et al.,* 2009).

The first aim of this study was to investigate the forensic value of the Y-Filer® Plus kit's 27 Y-STR markers in the South African population. This was accomplished by collecting 308 buccal swab samples from male individuals at the University of the Free State and amplifying them using the 27 Y-STR loci included in the Y-Filer® Plus testing kit. A total of 271 full DNA profiles were generated and used in statistical analysis. Overall, there were 12 Asian/Indian, 113 African, 43 Coloured, and 103 Caucasian samples that were included in the final statistical analyses. The profiles were used to calculate several forensic parameters to assess the forensic value of the Y-Filer® Plus kit for the South African male population.

Of the 271 profiles analysed, this testing kit was able to detect 261 (96%) unique haplotypes, of which 10 belonged to the Asian/Indian population, 109 to the African, 43 to the Coloured, and 99 to the Caucasian populations. The discrimination capacity of a testing kit is the ability of the set of markers to distinguish between individuals (Redd *et al.,* 2002), and is one of the most important parameters to consider when contemplating the use of a specific testing kit in forensic laboratories. In this study, the Y-Filer® Plus markers showed an overall discrimination capacity of 98.15%, with values ranging from 91.67% for the Asian/Indian population to 100% for the Coloured population. The African and Caucasian population exhibited discrimination capacities of 98.23% and 98.06% respectively. These values are significantly higher than those reported in previous studies conducted in South Africa. This parameter alone demonstrates the forensic potential that the Y-Filer® Plus testing kit has in the South African population.

The low mutation rate of markers on the Y-chromosome presents a challenge in the exclusion of related males when a match between DNA evidence and a male suspect occurs (Roewer,

2009). Given the uniparental inheritance and low mutation rates of Y-STRs, biologically related males would have the same haplotype, unless a mutation has taken place. The sampling for this study was random, but efforts were made to sample unrelated males, so any shared haplotypes detected in this study occurred by chance: purposefully including related males increases the probability of detecting shared haplotypes, thus reducing the discrimination capacity. However, it has been suggested that Y-STR reference databases should contain Y-STR profiles from related individuals, in addition to unrelated males, in order to provide more accurate and consistent haplotype frequency estimates (Kayser and de Knijff, 2011). The performance of the Y-Filer® Plus testing kit when working with father/son or brother pairs is yet to be investigated in the South African population.

In addition to the discrimination capacity, the haplotype diversity and gene diversity at each locus were investigated. The haplotype diversities ranged from 0.9848 in the Asian/Indian population to 1.0000 in the Coloured population, with an overall diversity of 0.9999. The overall gene diversity across all loci and all four population groups was 0.684. The gene diversity across all loci ranged from 0.625 in the African population to 0.729 in the Coloured population. It is not surprising that the Coloured population exhibited the highest level of gene and haplotype diversities, as well as the highest discrimination capacity, given the origins of this population group and the high levels of admixture detected (Tishkoff *et al.,* 2009). These diversities are comparable to those reported for other populations around the world, demonstrating the performance ability of this testing kit for the South African population. Consistent with previous research, however, low gene diversity was detected at several core loci in the African population. The *DYS391*, *DYS392*, and *DYS437* markers showed particularly low gene diversities, with several African population subgroups being monomorphic at these markers. Nonetheless, the low diversities at these markers did not have a significant impact on the overall efficiency of the Y-Filer® Plus kit.

Although the Y-Filer® Plus kit showed potential as a viable commercial kit option in South Africa, the match probabilities calculated in this study are still very high for a forensic application. While 100 samples may be sufficient for accurate allele frequency estimations, this study showed that this was not enough samples to calculate the reliable match probabilities that are needed to add confidence to the DNA evidence. In contrast to autosomal STRs, much larger reference databases are needed for match probabilities calculations when using Y-STRs and, currently, South African does not have extensive population reference data using the Y-Filer® Plus set of markers readily available.

The second aim of this study was to determine the ability of the Y-Filer® Plus kit to differentiate between the four main population groups in the South African population. The results of these

statistical analyses revealed that the largest amount of genetic differences occurred between the African and Caucasian population groups, while the Asian/Indian and Coloured populations were the most closely related population groups. In addition, the Coloured population showed a closer genetic relationship with the Caucasian population. This result is in line with the history of the Coloured population in South Africa, when exclusively considering the Y-lineage, as this group originated through the union of European men and African women during the Dutch settlement of the Western Cape (de Wit *et al.,* 2010). The AMOVAs performed showed that the majority of the variance detected occurred within populations as opposed to between populations, although the variation between populations was indicated to be significant. The Y-Filer® Plus kit is, therefore, able to successfully distinguish between the four main population groups in South Africa. This differentiation between populations is important in forensic applications, as it allows for the identification of the bioancestry of the individual from which the DNA originated, not only for accuracy in haplotype frequency estimation for match probability calculations, but may also prove to be a powerful investigative tool when suspect leads are lacking (Kayser, 2017).

Several different profile variations were also detected and analysed in this study. These variations include null alleles at markers *DYS448* and *DYS390*, duplications at *DYS458* and *DYS449*, and triplications at *DYF387S1.* Multiple samples also exhibited microvariant alleles across several loci. These intermediate alleles include 17.2 at *DYS458*, 32.3 at *DYS389II*, 40.2, 41.2, 42.2, and 45.2 at *DYF387S1*. All of these variations could complicate analysis if encountered during forensic DNA testing. However, if they are fully understood, and verified as true alleles during sample processing, they could contribute to haplotype rarity, thereby increasing the confidence in matches between the suspect and the DNA evidence (Budowle *et al.,* 2008). These variations have all been detailed through sequencing in other populations, so the molecular mechanisms behind them are relatively well understood (Butler *et al., 2005;* Balaresque *et al*., 2008; Budowle *et al.,* 2008; Iacovacci *et al.,* 2017; D'Atanasio *et al.,* 2019). However, there is not much information available regarding the polymorphisms that could result in these specific variations in the South African population.

Sexual assault—and rape in particular—is an enormous challenge faced by South Africans daily. Every person in this country has been affected by this type of crime in some way or another, so the impact of this research could be far reaching. Often DNA evidence in forensic investigations of sexually based crimes cannot be used to obtain a match between the evidence and a suspect, because autosomal STRs may not always provide sufficient information to make a conclusive inclusion (Hall and Ballantyne, 2003). While DNA evidence is not the only evidence used in a court of law, physical evidence plays a significant role in the outcome of a trial (Eckhert, 1997). It is, therefore, crucial to obtain the most reliable results.

The use of Y-STRs in forensic DNA testing could improve the chances of obtaining conclusive results, whether it be inclusions or exclusions, in cases of a sexual nature. Consequently, this could contribute to an increased conviction rate for perpetrators of such crimes, and could ultimately play a role in reducing the sexual offence statistic in South Africa. For these reasons, the following recommendations can be made:

1. Future research conducted in South Africa using this testing kit should consider using father/son and brother pairs as the sample population, to investigate the ability of the kit to distinguish between closely related South African males. Many populations around the world have reported on the effectiveness of the Y-Filer® Plus kit in differentiating between closely related males. It could, therefore, be expected that it would perform equally well in the South African population.

2. Increased genotyping efforts using the Y-Filer® Plus testing kit should be implemented in the South African population in order to establish a comprehensive reference database for accurate and reliable match probability calculations. The establishment of such a database could facilitate the use of Y-STRs in forensic investigations in South Africa's forensic laboratories.

3. Future studies should also focus on further detailing the profiles variations discussed in this study—including their underlying molecular causes—in the South African population. Knowledge regarding these variations, their causes, and their implications in forensic DNA testing could increase the confidence in using such variations in forensic investigations.

References

Balaresque, P., Bowden, G.R., Parkin, E.J., Omran, G.A., Heyer, E., Quintana-Murci, L., Roewer, L., Stoneking, M., Nasidze, I., Carvalho-Silva, D.R., and Tyler-Smith, C. (2008) 'Dynamic nature of the proximal AZFc region of the human Y chromosome: multiple independent deletion and duplication events revealed by microsatellite analysis'. *Human mutation*, 29(10), pp.1171-1180.

Budowle, B., Aranda, X. G., Lagace, R. E., Hennessy, L. K., Planz, J. V., Rodriguez, M., and Eisenberg, A. J. (2008) 'Null allele sequence structure at the DYS448 locus and implications for profile interpretation'. *International Journal of Legal Medicine*, 122(5), 421-427.

Butler, J.M., Decker, A.E., Kline, M.C., and Vallone, P.M. (2005) 'Chromosomal duplications along the Y-chromosome and their potential impact on Y-STR interpretation'. *Journal of Forensic Science*, 50(4), pp.853-859.

D'Amato, M. E., Benjeddou, M., and Davison, S. (2009) 'Evaluation of 21 Y-STRs for population and forensic studies'. *Forensic Science International: Genetics Supplement Series*, 2, pp.446-447.

D'Amato, M.E., Ehrenreich, L., Benjeddou, M., Davison, S. and Leat, N. (2008) 'Ancestry and genetic relationships between groups within the Cape Town metropolitan population inferred using Y-STRs genotyping'. *Forensic Science International: Genetics Supplement Series*, 1(1), pp.318-319.

D'Atanasio, E., Iacovacci, G., Pistillo, R., Bonito, M., Dugoujon, J.M., Moral, P., El-Chennawi, F., Melhaoui, M., Baali, A., Cherkaoui, M., and Sellitto, D. (2019) 'Rapidly mutating Y-STRs in rapidly expanding populations: Discrimination power of the Yfiler Plus multiplex in northern Africa'. *Forensic Science International: Genetics*, 38, pp.185-194.

de Wit, E., Delport, W., Rugamika, C.E., Meintjes, A., Möller, M., van Helden, P.D., Seoighe, C., and Hoal, E.G. (2010). Genome-wide analysis of the structure of the South African Coloured Population in the Western Cape'. *Human genetics*, 128(2), pp.145-153.

Eckert, D. A. (1997) '*Introduction the forensic science'*. Florida: CRC Press.

Hall, A., and Ballantyne, J. (2003) 'Novel Y-STR typing strategies reveal the genetic profile of the semen donor in extended interval post-coital cervicovaginal samples'. *Forensic Science International*, 136, pp.58-72.

Iacovacci, G., D'Atanasio, E., Marini, O., Coppa, A., Sellitto, D., Trombetta, B., Berti, A., and Cruciani, F. (2017) 'Forensic data and microvariant sequence characterization of 27 Y-STR loci analysed in four Eastern African countries'. *Forensic Science International: Genetics*, 27, pp.123-131.

Kayser, M. (2017) 'Forensic use of Y-chromosome DNA: a general overview'. *Human Genetics*, 136(5), pp.621-635.

Kayser, M. and de Knijff, P (2011) 'Improving human forensics through advances in genetics, genomics and molecular biology'. *Nature reviews, Genetics*, 12(3), pp.179-92.

Redd, A.J., Agellon, A.B., Kearney, V.A., Contreras, V.A., Karafet, T., Park, H., De Knijff, P., Butler, J.M., and Hammer, M.F. (2002). 'Forensic value of 14 novel STRs on the human Y chromosome'. *Forensic Science International*, 130(2-3), pp.97-111.

Roewer, L. (2009) 'Y chromosome STR typing in crime casework'. *Forensic Science, Medicine, and Pathology*, 5, pp.77-84.

Tishkoff, S.A., Reed, F.A., Friedlaender, F.R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J.B., Awomoyi, A.A., Bodo, J.M., Doumbo, O., and Ibrahim, M. (2009) 'The genetic structure and history of Africans and African Americans'. *Science,* 324(5930), pp.1035-1044.

APPENDIX A

Ethical Clearance Certificate

UNIVERSITY OF THE FREE STATE
UNIVERSITEIT VAN DIE VRYSTAAT
YUNIVESITHI YA FREISTATA

UFS UV

GENERAL/HUMAN RESEARCH ETHICS COMMITTEE (GHREC)

30-Aug-2019

Dear Miss Dooley, Kyla KB

Application Approved

Research Project Title:
Investigating the genetic diversity at 27 Y-STR loci in different population groups from South Africa.

Ethical Clearance number:
UFS-HSD2018/0779/2908

We are pleased to inform you that your application for ethical clearance has been approved. Your ethical clearance is valid for twelve (12) months from the date of issue. We request that any changes that may take place during the course of your study/research project be submitted to the ethics office to ensure ethical transparency. furthermore, you are requested to submit the final report of your study/research project to the ethics office. Should you require more time to complete this research, please apply for an extension. Thank you for submitting your proposal for ethical clearance; we wish you the best of luck and success with your research.

Yours sincerely

Prof Derek Litthauer
Chairperson: General/Human Research Ethics Committee

Digitally signed by Derek Litthauer
Date: 2019.08.30 19:25:01 +02'00'

205 Nelson Mandela Drive/Rylaan
Park West/Parkwes
Bloemfontein 9301
South Africa/Suid-Afrika

P.O. Box / Posbus 339
Bloemfontein 9300
South Africa / Suid-Afrika
T: +27(0)51 401 2116
F: +27(0)51 401 3752
WitthuhnRC@ufs.ac.za
www.ufs.ac.za

APPENDIX B

The information sheet that participants were given prior to providing a sample, which was used to fully explain the project to them

University of the Free State
PO Box 339
Bloemfontein 9300
South Africa

Telephone Numbers: +27 76 516 1653/+27 51 401 3978 Fax Number: +27 86 518 7317
Email Addresses: 2016108389@ufs4life.ac.za; EhlersK@ufs.ac.za

Information Sheet

Project Title: *Genetic diversity at 27 Y-STR loci in different population groups from South Africa for use in a forensic reference database.*

SAMPLE BARCODE

What is this study about?

DNA is the material, found in all of the cells in our body, which contains all the information to determine the characteristics of an individual. Short tandem repeat (STR) markers are DNA regions used to create DNA profiles which can be used to identify individuals during a forensic investigation. Y-STR markers are those gene regions found on the Y chromosome, meaning that they are only found in individuals that are genetically male. Y-STRs are commonly used in the investigation of crimes of a sexual nature. By targeting the male DNA only, analysis of the evidence is simpler and can provide a more accurate DNA profile for use in suspect identification. However, when using DNA to identify a suspect, it is necessary to compare the results to a database of the population, to determine how common or unique the DNA profile is – the more unique it is, the more reliable the DNA results are. There is currently no database available for the different population groups in South Africa for the specific set of Y-STR markers used in the Y-Filer Plus analysis kit. The aim of this study is to establish a database of these markers for the different ethnic groups in the South African population.

What are the benefits of this research in South Africa?

South Africa is one of the most violent countries in the world, with alarmingly high crime rates. The number of reported sexual assaults in South Africa is exceptionally high, with 49 660 reports in total in 2016/2017, of which 39 828 of those reports were of actual rape. Although this value is already high enough, it is possible that the actual incidence rate of sexual assaults could be much higher because a lot of assaults actually go unreported. There has been research conducted into the use of Y-STRs in investigations in South Africa, however, some markers prove to be unreliable for certain population groups. The Y-Filer Plus analysis kit could be very beneficial in analysing evidence more accurately and as a result, more criminals could be identified and stopped. However, a database needs to be established before this kit can be used in forensic investigations.

What will you be asked to do if you agree to participate?

If you agree to participate, you will be asked to provide a DNA sample by means of a buccal swab, taken from the inside of your cheek. Two swabs will be taken using buccal swabs (safe for human use). The sample will be taken by gently rubbing the swab on the inside of your cheek for approximately 10 seconds – one swab will be used on the left side and the other on the right. You will also be asked to provide your race (White, Black, Coloured, or Asian/Indian),

as well as the population group (as specific as possible) and home language of yourself and your parents. You will not be asked to provide your name.

How will your personal details be kept confidential in this study?

You will not be providing your name for this study, and so your sample will remain anonymous and cannot be traced back to you. Your sample is coded using a unique identification code with reference to your race, and will be kept in a secure location until used. Only the researcher and study coordinator will have access to your samples and the information that you provide. The samples you donate will have a one-time use and therefore will be destroyed once analysed, however, the duplicate sample will be stored until the study is complete.

What are the benefits and risks of participating in this study?

There are no personal economic benefits for participating in this study. There are no risks of any side effects, health-related or otherwise, from participating.

Will you be able to withdraw from the study?

You would be able to withdraw from the study at any given time, without giving any reason. Should you want to withdraw, please contact the researcher using the contact details provided and provide your sample barcode number (found at the top of Page 1), and your sample and information will be destroyed. If your sample has already been analysed, the resulting data will also be removed and deleted.

How will the results of the study be published?

This study is being conducted in partial fulfilment of the degree Master of Science (Forensic Genetics), and as such, all analysed data will be available to the final examiners. However, only the researcher and study coordinator will have access to the raw data, and any donor information. This study may result in peer-reviewed articles or reports which would be published and made publicly available.

Who can you contact if you have any further questions about this research?

For any further queries, you may contact the researcher; +27 76 516 1653 or 2016108389@ufs4life.ac.za, or the study coordinator, Dr Karen Ehlers; +27 51 401 3978 or EhlersK@ufs.ac.za.

Who can you contact, besides the researcher, for assistance if you have been negatively affected by participating in this study or any other problems you might have experienced?

You can contact the Head of Department, Professor JP Grobler, on +27 51 401 3844 or via email, GroblerJP@ufs.ac.za. This study has received ethical approval with the number *xxxxxx*.

Study Coordinator

Name: Dr Karen Ehlers
Contact number: +27 51 401 3978
Email Address: EhlersK@ufs.ac.za
Address: Office 169, Department of Genetics,
Biology Building, Bloemfontein Campus,
University of the Free State, Bloemfontein,
South Africa, 9300

_____
Signature

_____
Date

The informed consent form that participants were asked to complete before providing their DNA samples

University of the Free State
PO Box 339
Bloemfontein 9300
South Africa

Telephone Numbers: +27 76 516 1653/+27 51 401 3978 Fax Number: +27 86 518 7317
Email Addresses: 2016108389@ufs4life.ac.za; EhlersK@ufs.ac.za

Consent Form

Project Title: *Genetic diversity at 27 Y-STR loci in different population groups from South Africa for use in a forensic reference database.*

SAMPLE BARCODE

By signing this consent form, I am acknowledging and agreeing to the following:

- ☐ I am over the age of 18 years old.
- ☐ The project has been fully explained to me, in a language that I understand, and any questions that I had have been answered.
- ☐ I will not provide my name and thus will be kept anonymous when giving a sample, and I will be able to withdraw from the study at any point.
- ☐ The genetic material for analysis will be obtained from the buccal swab sample I am donating.
- ☐ My sample will be given a unique identification code in reference to my race, and that it will be stored in a secure location until the study is completed and then destroyed.
- ☐ I, freely and voluntarily, agree to provide a DNA sample, as well information regarding the ethnicity of myself and my parents.

YES	NO

_____ _____

Participant's Signature Date

Should the participant have any further questions regarding the study, or wish to report any problems they have experienced during this process, they should contact the study coordinator.

Study Coordinator

Name: Dr Karen Ehlers
Contact number: +27 51 401 3978
Email Address: EhlersK@ufs.ac.za
Address: Office 169, Department of Genetics, Biology Building, Bloemfontein Campus, University of the Free State, Bloemfontein, South Africa, 9300

_____
Signature

_____
Date

The questionnaire that participants were asked to complete when providing a DNA sample

Information provided for a voluntary DNA donation – University of the Free State

Race: WHITE | BLACK | COLOURED | ASIAN/INDIAN

Date: _____

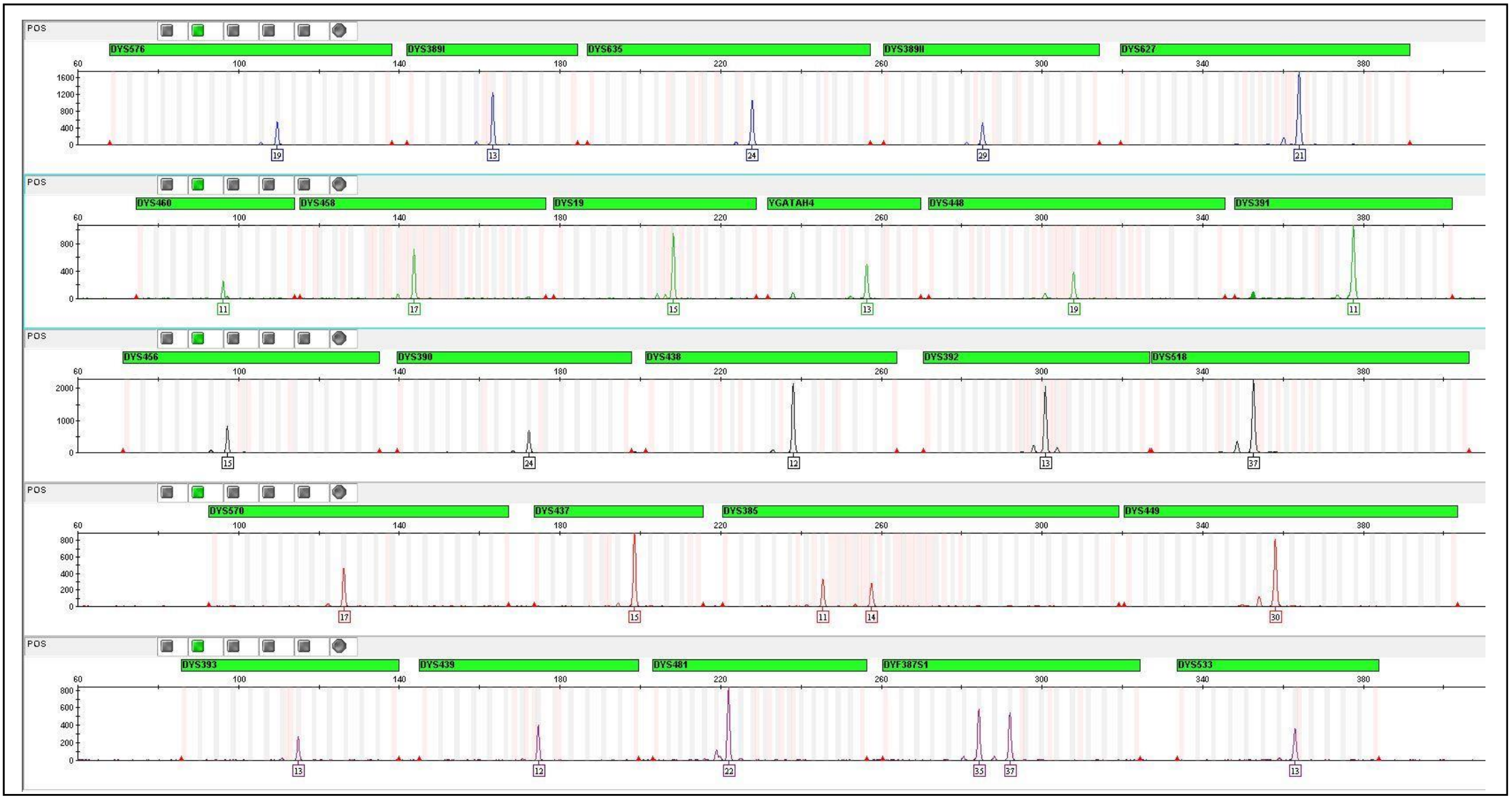
SAMPLE BARCODE

	MYSELF	MOTHER	FATHER
POPULATION GROUP			
HOME LANGUAGE			

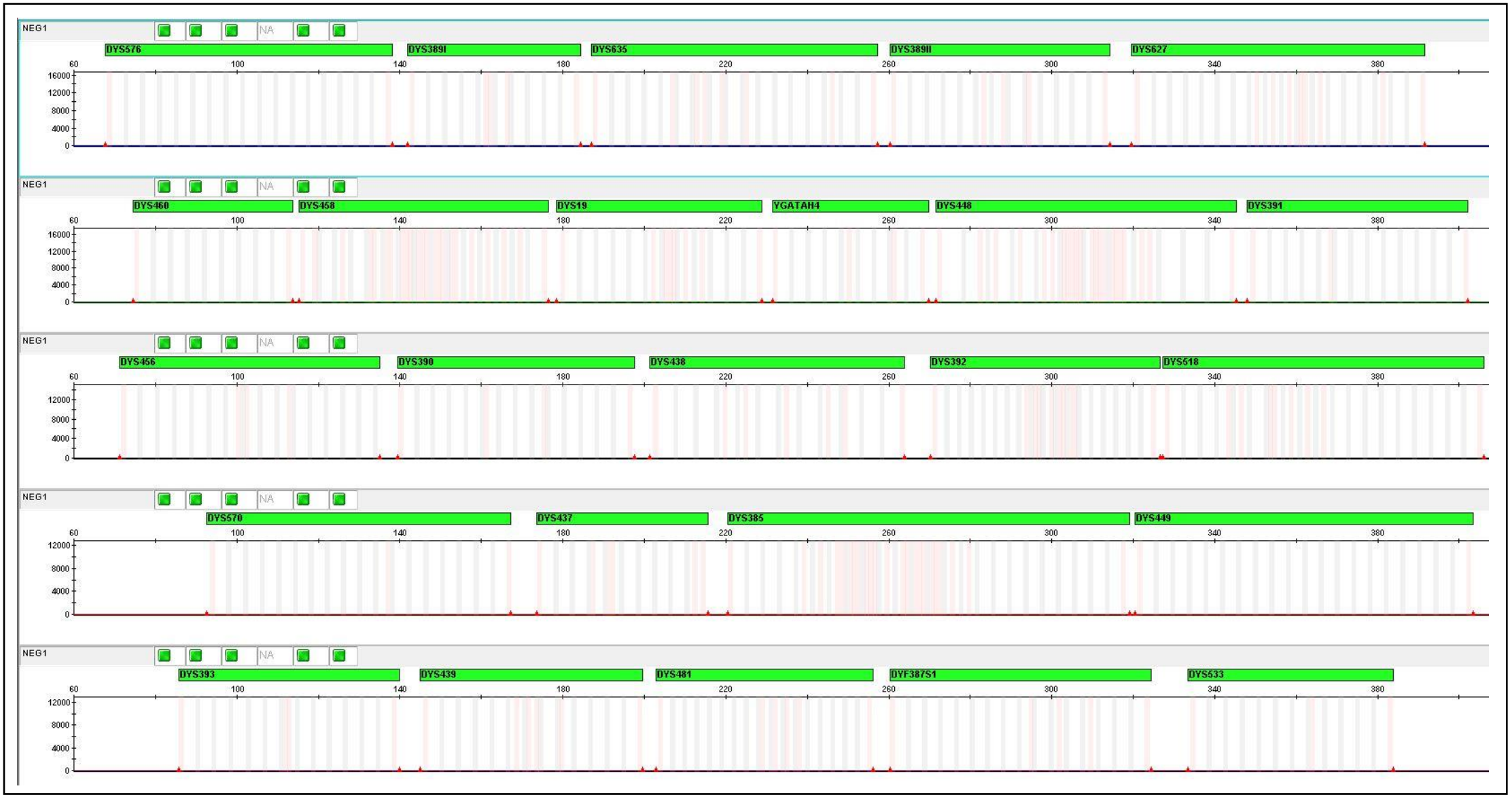
Page 2 of 2

APPENDIX C

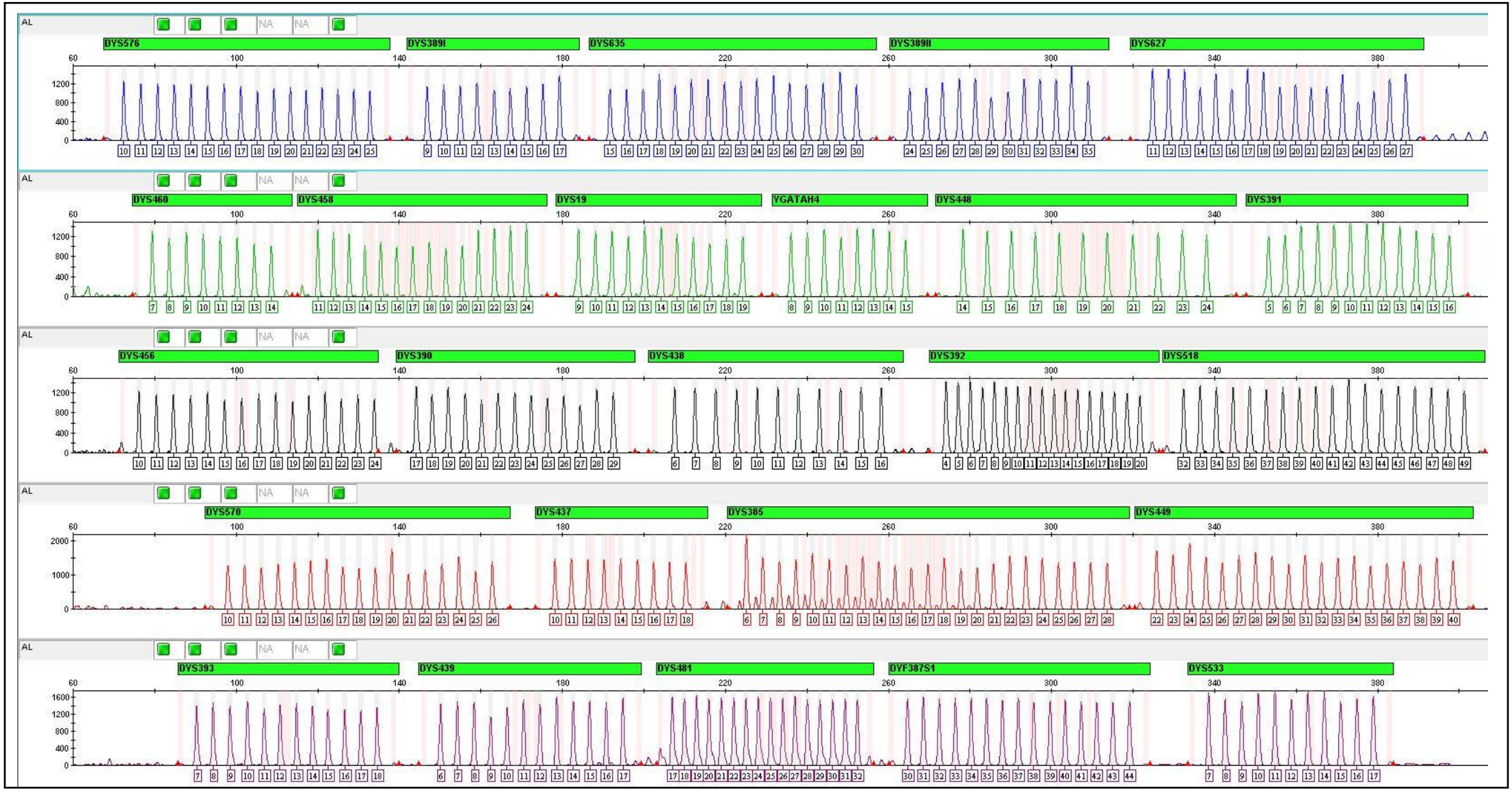
The resulting DNA profile of the positive control (DNA Control 007)



The resulting DNA profile of a negative control



The resulting DNA profile of the Yfiler™ Plus Allelic Ladder



APPENDIX D

Allele frequency table for the Asian/Indian, African, Coloured, and Caucasian population groups in South Africa

Locus	Allele	Asian/Indian	African	Coloured	Caucasian
DYS576	**n***	12	113	43	103
	12	0.000	0.000	0.000	0.010
	14	0.000	0.097	0.070	0.019
	15	0.083	0.319	0.116	0.029
	16	0.250	0.150	0.093	0.223
	17	0.417	0.186	0.395	0.379
	18	0.083	0.195	0.209	0.223
	19	0.083	0.053	0.093	0.068
	20	0.083	0.000	0.000	0.029
	21	0.000	0.000	0.023	0.010
	22	0.000	0.000	0.000	0.010
DYS389I	**n**	12	113	43	103
	10	0.000	0.018	0.000	0.000
	11	0.000	0.009	0.000	0.000
	12	0.167	0.195	0.163	0.243
	13	0.583	0.460	0.674	0.592
	14	0.250	0.319	0.163	0.155
	15	0.000	0.000	0.000	0.010
DYS635	**n**	12	113	43	103
	17	0.000	0.168	0.140	0.000
	18	0.167	0.000	0.000	0.000
	19	0.000	0.009	0.000	0.000
	20	0.083	0.062	0.023	0.049
	21	0.250	0.425	0.233	0.146
	22	0.167	0.159	0.140	0.126
	23	0.167	0.071	0.279	0.563
	24	0.000	0.106	0.140	0.058
	25	0.167	0.000	0.023	0.058
	29	0.000	0.000	0.023	0.000
DYS389II	**n**	12	113	43	103
	27	0.000	0.009	0.023	0.010
	28	0.167	0.159	0.163	0.175
	29	0.500	0.106	0.419	0.456
	30	0.250	0.274	0.186	0.223
	31	0.000	0.248	0.140	0.087
	32	0.083	0.177	0.070	0.039
	32.3	0.000	0.009	0.000	0.000
	33	0.000	0.018	0.000	0.010

*n – the total number of alleles observed at each locus in each population group

Allele frequency calculated using the count method, therefore:

Allele frequency = $\frac{Number\ of\ times\ the\ allele\ was\ observed\ at\ the\ locus}{Total\ number\ of\ alleles\ observed\ at\ the\ locus}$

Locus	Allele	Asian/Indian	African	Coloured	Caucasian
DYS627	**n**	12	113	43	103
	15	0.000	0.000	0.000	0.010
	16	0.000	0.000	0.047	0.039
	17	0.000	0.009	0.023	0.039
	18	0.167	0.071	0.023	0.049
	19	0.083	0.301	0.326	0.107
	20	0.333	0.398	0.140	0.194
	21	0.167	0.142	0.140	0.252
	22	0.167	0.044	0.186	0.175
	23	0.083	0.027	0.070	0.087
	24	0.000	0.009	0.047	0.049
DYS460	**n**	12	113	43	103
	9	0.000	0.000	0.000	0.039
	10	0.667	0.319	0.442	0.398
	11	0.333	0.602	0.465	0.515
	12	0.000	0.080	0.093	0.049
DYS458	**n**	12	114	43	103
	13	0.000	0.000	0.047	0.000
	14	0.083	0.018	0.000	0.010
	15	0.167	0.035	0.163	0.243
	16	0.083	0.281	0.233	0.320
	17	0.500	0.430	0.279	0.223
	17,2	0.000	0.000	0.000	0.010
	18	0.167	0.228	0.209	0.155
	19	0.000	0.009	0.047	0.019
	20	0.000	0.000	0.023	0.019
DYS19	**n**	12	113	43	103
	13	0.083	0.018	0.023	0.078
	14	0.167	0.168	0.279	0.621
	15	0.583	0.522	0.465	0.194
	16	0.167	0.221	0.186	0.068
	17	0.000	0.071	0.047	0.039
YGATAH4	**n**	12	113	43	103
	9	0.000	0.018	0.023	0.000
	10	0.167	0.035	0.070	0.068
	11	0.333	0.522	0.233	0.427
	12	0.250	0.310	0.558	0.456
	13	0.250	0.089	0.116	0.039
	14	0.000	0.027	0.000	0.010

*n – the total number of alleles observed at each locus in each population group

Allele frequency calculated using the count method, therefore:

Allele frequency = $\frac{Number\ of\ times\ the\ allele\ was\ observed\ at\ the\ locus}{Total\ number\ of\ alleles\ observed\ at\ the\ locus}$

Locus	Allele	Asian/Indian	African	Coloured	Caucasian
DYS448	**n**	12	111	41	103
	16	0.000	0.000	0.024	0.000
	18	0.167	0.000	0.098	0.078
	19	0.583	0.171	0.317	0.505
	20	0.167	0.153	0.220	0.340
	21	0.083	0.532	0.220	0.078
	22	0.000	0.036	0.073	0.000
	23	0.000	0.108	0.049	0.000
DYS391	**n**	12	113	43	103
	8	0.000	0.000	0.023	0.000
	9	0.000	0.009	0.000	0.029
	10	0.917	0.876	0.674	0.447
	11	0.083	0.115	0.302	0.524
DYS456	**n**	12	113	43	103
	13	0.083	0.124	0.047	0.019
	14	0.000	0.027	0.047	0.175
	15	0.583	0.646	0.581	0.320
	16	0.333	0.097	0.209	0.369
	17	0.000	0.089	0.093	0.107
	18	0.000	0.018	0.023	0.010
DYS390	**n**	11	101	43	103
	18	0.000	0.000	0.023	0.000
	19	0.000	0.000	0.023	0.000
	20	0.000	0.010	0.000	0.000
	21	0.000	0.535	0.209	0.000
	22	0.455	0.099	0.209	0.136
	23	0.182	0.000	0.256	0.311
	24	0.091	0.188	0.186	0.311
	25	0.273	0.119	0.093	0.223
	26	0.000	0.050	0.000	0.019
DYS438	**n**	12	113	43	103
	9	0.250	0.000	0.070	0.029
	10	0.417	0.221	0.256	0.340
	11	0.333	0.708	0.372	0.117
	12	0.000	0.071	0.279	0.505
	13	0.000	0.000	0.023	0.010

*n – the total number of alleles observed at each locus in each population group

Allele frequency calculated using the count method, therefore:

Allele frequency = $\frac{Number\ of\ times\ the\ allele\ was\ observed\ at\ the\ locus}{Total\ number\ of\ alleles\ observed\ at\ the\ locus}$

Locus	Allele	Asian/Indian	African	Coloured	Caucasian
DYS392	**n**	12	113	43	103
	9	0.000	0.000	0.000	0.010
	10	0.083	0.035	0.116	0.000
	11	0.667	0.938	0.488	0.408
	12	0.000	0.027	0.047	0.049
	13	0.167	0.000	0.233	0.476
	14	0.083	0.000	0.093	0.058
	15	0.000	0.000	0.023	0.000
DYS518	**n**	12	113	43	103
	35	0.000	0.000	0.093	0.000
	36	0.000	0.000	0.000	0.107
	37	0.083	0.062	0.163	0.155
	38	0.250	0.142	0.186	0.194
	39	0.250	0.168	0.256	0.272
	40	0.167	0.133	0.116	0.146
	41	0.250	0.177	0.140	0.097
	42	0.000	0.142	0.047	0.010
	43	0.000	0.089	0.000	0.000
	44	0.000	0.080	0.000	0.019
	46	0.000	0.009	0.000	0.000
DYS570	**n**	12	113	43	103
	14	0.000	0.009	0.000	0.000
	15	0.167	0.000	0.047	0.000
	16	0.167	0.035	0.116	0.068
	17	0.333	0.283	0.326	0.340
	18	0.083	0.265	0.233	0.214
	19	0.167	0.221	0.209	0.223
	20	0.083	0.115	0.070	0.126
	21	0.000	0.062	0.000	0.029
	22	0.000	0.009	0.000	0.000
DYS437	**n**	12	113	43	103
	14	0.750	0.973	0.605	0.320
	15	0.250	0.027	0.302	0.456
	16	0.000	0.000	0.093	0.223

*n – the total number of alleles observed at each locus in each population group

Allele frequency calculated using the count method, therefore:

Allele frequency = $\frac{Number\ of\ times\ the\ allele\ was\ observed\ at\ the\ locus}{Total\ number\ of\ alleles\ observed\ at\ the\ locus}$

Locus	Allele	Asian/Indian	African	Coloured	Caucasian
DYS385	**n**	23	197	79	196
	9	0.000	0.000	0.000	0.005
	10	0.000	0.000	0.025	0.026
	11	0.043	0.076	0.190	0.265
	12	0.000	0.000	0.013	0.041
	13	0.261	0.005	0.063	0.107
	14	0.000	0.041	0.177	0.316
	15	0.174	0.137	0.139	0.138
	16	0.174	0.218	0.127	0.051
	17	0.087	0.223	0.139	0.015
	18	0.043	0.117	0.076	0.036
	19	0.130	0.066	0.025	0.000
	20	0.087	0.107	0.013	0.000
	21	0.000	0.010	0.013	0.000
DYS449	**n**	12	114	43	103
	25	0.000	0.009	0.000	0.000
	26	0.083	0.018	0.023	0.010
	27	0.167	0.123	0.047	0.058
	28	0.083	0.281	0.140	0.184
	29	0.083	0.079	0.186	0.272
	30	0.000	0.070	0.186	0.252
	31	0.083	0.114	0.140	0.097
	32	0.500	0.140	0.093	0.049
	33	0.000	0.035	0.093	0.029
	34	0.000	0.018	0.047	0.029
	35	0.000	0.044	0.023	0.019
	36	0.000	0.044	0.000	0.000
	37	0.000	0.018	0.000	0.000
	38	0.000	0.009	0.023	0.000
DYS393	**n**	12	113	43	103
	11	0.083	0.000	0.047	0.000
	12	0.250	0.009	0.140	0.078
	13	0.250	0.566	0.605	0.786
	14	0.250	0.221	0.163	0.136
	15	0.167	0.186	0.047	0.000
	16	0.000	0.018	0.000	0.000

*n – the total number of alleles observed at each locus in each population group

Allele frequency calculated using the count method, therefore:

Allele frequency = $\frac{Number\ of\ times\ the\ allele\ was\ observed\ at\ the\ locus}{Total\ number\ of\ alleles\ observed\ at\ the\ locus}$

Locus	Allele	Asian/Indian	African	Coloured	Caucasian
DYS439	**n**	12	113	43	103
	10	0.250	0.018	0.000	0.058
	11	0.167	0.283	0.465	0.398
	12	0.500	0.531	0.372	0.427
	13	0.083	0.124	0.163	0.117
	14	0.000	0.035	0.000	0.000
	15	0.000	0.009	0.000	0.000
DYS481	**n**	12	113	43	103
	19	0.000	0.009	0.000	0.010
	20	0.000	0.000	0.023	0.019
	21	0.000	0.000	0.047	0.058
	22	0.000	0.018	0.186	0.379
	23	0.083	0.044	0.140	0.165
	24	0.500	0.133	0.093	0.078
	25	0.167	0.283	0.209	0.175
	26	0.083	0.257	0.047	0.058
	27	0.083	0.097	0.070	0.058
	28	0.083	0.089	0.116	0.000
	29	0.000	0.044	0.047	0.000
	30	0.000	0.027	0.023	0.000
DYF387S1	**n**	24	195	79	178
	32	0.000	0.000	0.013	0.000
	34	0.000	0.000	0.051	0.017
	35	0.042	0.041	0.139	0.264
	36	0.250	0.138	0.228	0.303
	37	0.250	0.113	0.215	0.169
	38	0.083	0.282	0.139	0.185
	39	0.167	0.277	0.127	0.045
	40	0.208	0.056	0.038	0.017
	40,2	0.000	0.000	0.013	0.000
	41	0.000	0.010	0.013	0.000
	41,2	0.000	0.077	0.013	0.000
	42,2	0.000	0.005	0.000	0.000
	45,2	0.000	0.000	0.013	0.000
DYS533	**n**	12	113	43	103
	8	0.000	0.000	0.047	0.000
	9	0.000	0.018	0.023	0.019
	10	0.250	0.053	0.070	0.019
	11	0.417	0.717	0.349	0.291
	12	0.333	0.212	0.349	0.612
	13	0.000	0.000	0.163	0.058

*n – the total number of alleles observed at each locus in each population group

Allele frequency calculated using the count method, therefore:

Allele frequency = $\frac{Number\ of\ times\ the\ allele\ was\ observed\ at\ the\ locus}{Total\ number\ of\ alleles\ observed\ at\ the\ locus}$

APPENDIX E

Haplotype frequency table for the Asian/Indian, African, Coloured, and Caucasian population groups in South Africa

ID	Pop	DYS 576	DYS 389I	DYS 635	DYS 389II	DYS 627	DYS 460	DYS 458	DYS 19	YGA TAH 4	DYS 448	DYS 391	DYS 456	DYS 390	DYS 438	DYS 392	DYS 518	DYS 570	DYS 437	DYS 385	DYS 449	DYS 393	DYS 439	DYS 481	DYF 387S1	DYS 533	Freq
A001	A/I	16	13	18	29	20	11	17	15	10	19	10	15	22	10	11	38	16	14	13;19	32	15	10	24	39;40	11	0.0074
A002	A/I	16	13	18	29	20	11	17	15	10	19	10	15	22	10	11	38	16	14	13;19	32	15	10	24	39;40	11	0.0074
A003	A/I	15	12	21	30	20	10	17	15	12	21	10	15	0	11	11	38	20	14	16	28	13	12	26	36;39	11	0.0037
A004	A/I	18	14	22	30	23	10	17	15	11	18	10	15	25	11	13	41	17	14	15;20	27	14	12	24	36;37	10	0.0037
A006	A/I	16	13	21	29	18	11	16	16	11	19	10	13	23	9	11	37	15	15	13;18	31	12	12	23	36;38	10	0.0037
A007	A/I	17	12	25	28	21	10	14	14	12	19	10	16	22	10	13	39	15	15	13;17	27	11	13	25	35;39	12	0.0037
A008	A/I	17	14	25	30	20	10	18	14	12	19	10	16	23	11	10	39	19	15	13;17	32	14	11	24	37;40	11	0.0037
A010	A/I	17	14	23	32	18	11	15	15	13	20	11	16	25	11	11	41	19	14	11;15	32	13	10	24	37;38	12	0.0037
A011	A/I	17	13	20	29	19	10	17	15	13	19	10	15	22	9	11	39	17	14	16;19	32	12	11	27	36;37	11	0.0037
A016	A/I	19	13	21	28	22	10	17	13	11	20	10	15	25	10	11	40	17	14	13;16	26	13	12	25	36;40	12	0.0037
A017	A/I	17	13	22	29	21	10	18	15	11	19	10	15	22	9	11	40	17	14	15;16	32	12	12	28	37;40	12	0.0037
A018	A/I	20	13	23	29	22	10	15	16	13	18	10	16	24	10	14	41	18	14	15;20	29	14	12	24	36;37	10	0.0037
B001	A	17	13	22	30	18	11	16	15	11	20	10	15	21	11	11	39	18	14	17;18	32	15	12	26	38;39	11	0.0037
B002	A	16	14	21	30	21	11	17	16	11	21	10	15	21	12	11	43	17	14	16;20	29	15	13	25	37;38	12	0.0037
B003	A	15	14	23	32	19	11	16	15	12	21	10	15	22	11	11	40	19	14	16;17	28	13	13	29	35;38	11	0.0037
B004	A	15	14	21	32	20	10	17	15	12	21	10	15	0	11	11	38	21	14	15;16	28	13	12	25	37;39	11	0.0037
B005	A	15	14	21	31	19	10	17	15	12	21	10	15	0	11	12	39	19	14	15;20	27	13	12	26	39	12	0.0037
B006	A	15	12	24	28	23	11	18	14	11	19	10	15	25	11	11	42	19	14	15	28	13	11	26	38	12	0.0074
B007	A	15	12	24	28	23	11	18	14	11	19	10	15	25	11	11	42	19	14	14;21	28	13	11	26	38	12	0.0074
B009	A	15	12	21	29	18	11	16	15	11	21	10	15	21	11	11	39	20	14	16	29	13	11	28	37;39	11	0.0037
B010	A	14	13	20	31	19	11	18	15	12	21	11	16	21	11	11	39	19	14	13;19	27	14	14	29	36;39	11	0.0037
B012	A	14	12	24	28	20	11	18	14	11	19	10	15	25	10	11	44	19	14	16;20	28	13	11	25	38	12	0.0037

Table Legend:

ID – Sample Code **Pop** – Population Group **Freq** – Haplotype Frequency **A/I** – Asian/Indian **A** – African **Co** – Coloured **Ca** – Caucasian

– Marker with at least one duplication/triplication seen – Null allele – Microvariant – Duplication – Duplication and Microvariant – Triplication

ID	Pop	DYS 576	DYS 389I	DYS 635	DYS 389II	DYS 627	DYS 460	DYS 458	DYS 19	YGA TAH 4	DYS 448	DYS 391	DYS 456	DYS 390	DYS 438	DYS 392	DYS 518	DYS 570	DYS 437	DYS 385	DYS 449	DYS 393	DYS 439	DYS 481	DYF 387S1	DYS 533	Freq
B013	A	14	12	22	29	18	10	17	16	12	21	11	15	21	11	11	40	18	14	16;17	29	13	12	27	36;39	11	0.0037
B014	A	15	12	23	28	22	11	18	14	11	19	10	15	25	12	11	39	18	14	15;20	27	13	11	27	38	12	0.0037
B015	A	17	13	21	32	19	10	17	16	12	20	10	15	21	12	11	40	19	14	16;17	28	13	12	28	36;39	11	0.0037
B016	A	14	14	21	31	19	10	16	15	12	21	10	15	21	11	11	39	20	14	16;19	29	14	11	30	37;39	11	0.0037
B017	A	17	14	23	31	21	11	17	17	11	20	10	17	21	11	11	41	19	14	16;17	32	14	13	24	39	11	0.0037
B018	A	18	13	21	31	19	10	17	16	12	20	10	15	21	12	11	38	18	14	16;17	28	13	12	28	36;39	11	0.0037
B019	A	19	10	21	28	19	11	17	13	11	20	11	18	24	10	11	39	18	14	16;17	31	14	10	25	36;37	11	0.0037
B020	A	18	14	17	32	20	11	17	15	14	23	10	13	24	10	11	41	17	14	11	36	13	12	25	38;41.2	11	0.0037
B023	A	15	13	22	30	20	11	16	15	12	21	10	15	20	11	11	39	18	14	15;20	31	13	11	28	39	11	0.0037
B024	A	15	12	23	28	21	11	18	14	11	19	10	15	25	11	11	43	18	14	15;20	28	13	11	25	38	12	0.0037
B025	A	17	13	17	31	20	11	17	15	13	23	10	14	24	10	11	42	17	14	11	34	13	12	19	38;41.2	11	0.0037
B026	A	17	13	22	30	20	10	17	15	11	21	11	15	21	11	11	41	21	14	16	30	15	15	23	35;40	10	0.0037
B027	A	15	13	21	31	20	10	15	15	12	21	10	15	21	11	11	37	20	14	17;18	28	13	12	28	36;39	11	0.0037
B029	A	16	13	21	29	19	11	18	15	11	20	10	17	21	11	11	43	18	14	17;18	32	15	12	26	37;39	11	0.0037
B030	A	15	14	21	32	21	10	18	15	12	21	10	15	0	11	11	38	21	14	15;17	27	13	12	26	37;39	11	0.0037
B031	A	15	13	20	31	22	10	16	15	12	21	10	15	21	11	11	37	19	14	17	27	13	12	30	38;39	11	0.0037
B032	A	18	13	17	31	20	11	16	17	12	22	10	13	24	10	11	41	18	15	11	35	13	12	25	37;42.2	12	0.0037
B033	A	16	13	22	33	20	12	16	15	11	21	10	16	21	11	11	44	17	14	16;17	31	16	13	26	37;38	11	0.0037
B034	A	15	14	17	30	22	11	14	15	10	0	10	15	22	12	10	37	16	15	14;19	30	13	12	25	35;36;40	9	0.0037
B035	A	18	13	21	30	20	11	16	17	12	21	10	17	21	11	11	41	17	14	17;18	31	14	12	24	39;40	11	0.0037
B036	A	18	13	17	31	18	11	17	15	13	23	10	13	24	10	11	41	17	14	11	36	13	12	26	37;41.2	11	0.0037
B037	A	19	13	21	30	19	10	16	17	11	21	10	17	21	11	11	41	17	14	17;18	26	14	12	25	38;40	11	0.0037
B038	A	17	13	21	30	20	11	17	16	11	21	10	16	21	11	11	42	18	14	17;18	32	15	13	25	38	11	0.0037
B039	A	14	12	23	28	20	11	18	14	11	20	10	15	25	10	11	44	18	14	16;20	28	13	11	25	38	12	0.0037

Table Legend:

ID – Sample Code **Pop** – Population Group **Freq** – Haplotype Frequency **A/I** – Asian/Indian **A** – African **Co** – Coloured **Ca** – Caucasian

– Marker with at least one duplication/triplication seen – Null allele – Microvariant – Duplication – Duplication and Microvariant – Triplication

ID	Pop	DYS 576	DYS 389I	DYS 635	DYS 389II	DYS 627	DYS 460	DYS 458	DYS 19	YGA TAH 4	DYS 448	DYS 391	DYS 456	DYS 390	DYS 438	DYS 392	DYS 518	DYS 570	DYS 437	DYS 385	DYS 449	DYS 393	DYS 439	DYS 481	DYF 387S1	DYS 533	Freq
B040	A	14	12	22	28	24	11	17	14	11	19	11	15	24	11	11	43	20	14	14;19	27	13	11	25	38;39	12	0.0037
B041	A	16	14	21	31	18	11	16	16	11	21	11	15	22	11	11	39	17	14	17	33	15	13	26	37;39	11	0.0037
B042	A	16	13	21	30	18	10	16;17	15	11	22	10	16	21	11	11	42	18	14	17;20	28;29	15	11	26	38;39	11	0.0037
B043	A	14	13	22	29	22	10	18	14	11	19	10	15	25	11	11	42	19	14	14;20	27	13	11	27	38	12	0.0037
B044	A	17	14	21	32	19	12	16	16	11	21	10	16	21	11	11	41	18	14	17;20	33	15	12	27	36;39	11	0.0037
B045	A	17	14	21	31	19	11	16	16	11	21	10	15	21	11	11	42	18	14	17;20	32	14	12	27	36;39	11	0.0037
B046	A	15	12	24	28	21	12	18	14	11	19	10	15	25	10	11	43	18	14	15;20	28	13	11	26	38	12	0.0037
B047	A	18	13	17	32.3	20	11	17	15	12	23	10	13	24	10	11	42	16	14	11	36	13	11	25	38;41.2	11	0.0037
B048	A	15	13	21	31	20	10	18	15	13	21	10	15	21	11	11	39	19	14	16;17	29	13	14	28	35;39	11	0.0037
B049	A	15	14	21	31	18	10	18	15	12	21	10	16	0	11	11	39	20	14	15;16	28	13	13	26	36;39	11	0.0037
B050	A	17	13	20	31	20	11	16	15	11	21	10	15	21	11	11	40	19	14	17;18	32	15	12	23	38;40	11	0.0037
B051	A	17	12	17	29	19	11	17	16	9	19	11	15	22	11	10	38	18	14	17	29	13	12	27	36;39	10	0.0037
B052	A	15	14	21	32	20	10	17	15	12	21	10	15	0	11	11	38	20	14	15;16	28	13	12	24	36;39	11	0.0037
B053	A	15	12	24	28	20	11	17	14	11	19	10	15	25	10	11	44	18	14	16;20	28	13	11	24	38	12	0.0037
B054	A	19	14	17	32	20	11	18	15	13	23	10	13	24	10	11	42	17	14	11	36	13	12	25	38;41.2	11	0.0037
B055	A	17	13	21	31	19	10	17	16	12	20	10	15	21	12	11	39	19	14	17	28	13	11	28	36;39	11	0.0037
B056	A	18	13	21	30	19	11	17	15	11	20	10	15	21	11	11	38	17	14	17;18	31	15	12	24	38;40	11	0.0037
B057	A	15	14	21	32	21	10	17	15	12	21	10	15	0	11	11	38	19	14	15;16	28	13	12	26	36;39	11	0.0037
B058	A	17	13	22	30	20	10	17	15	11	21	10	15	21	11	11	40	21	14	15;16	30	16	13	23	35;40	10	0.0037
B059	A	17	13	22	30	19	10	16	15	12	21	11	14	21	11	11	41	21	14	16	31	14	12	22	35;41	11	0.0037
B060	A	18	14	21	31	19	11	16	16	11	21	10	15	21	11	11	42	18	14	16;20	30	15	12	26	36;39	11	0.0037
B061	A	18	13	20	31	21	11	17	16	11	21	10	17	22	11	11	40	18	14	17;18	32	14	12	25	39	11	0.0037
B062	A	15	14	21	30	20	10	17	15	12	21	10	15	0	11	11	38	19	14	15;17	26	13	12	26	36;39	11	0.0037
B063	A	14	13	21	31	19	11	17	15	12	20	10	15	21	11	11	39	19	14	15;19	27	14	12	29	36;39	11	0.0037

Table Legend:

ID – Sample Code **Pop** – Population Group **Freq** – Haplotype Frequency **A/I** – Asian/Indian **A** – African **Co** – Coloured **Ca** – Caucasian

– Marker with at least one duplication/triplication seen – Null allele – Microvariant – Duplication – Duplication and Microvariant – Triplication

ID	Pop	DYS 576	DYS 389I	DYS 635	DYS 389II	DYS 627	DYS 460	DYS 458	DYS 19	YGA TAH 4	DYS 448	DYS 391	DYS 456	DYS 390	DYS 438	DYS 392	DYS 518	DYS 570	DYS 437	DYS 385	DYS 449	DYS 393	DYS 439	DYS 481	DYF 387S1	DYS 533	Freq
B064	A	16	13	21	30	19	11	16	16	11	21	10	15	21	11	12	41	17	14	16;17	32	15	12	26	38;39	11	0.0037
B065	A	18	14	17	32	20	11	19	15	13	23	10	13	24	10	11	42	17	14	11	38	13	12	25	37;41.2	11	0.0037
B066	A	19	14	17	32	20	11	18	15	13	23	10	13	24	10	11	41	18	14	11	35	13	12	26	38;41.2	11	0.0037
B067	A	16	13	21	29	19	11	16	16	11	21	10	17	21	11	11	39	17	14	16;19	31	15	11	25	39	11	0.0037
B068	A	15	12	24	28	20	12	18	14	11	19	10	15	26	11	11	44	17	14	15;20	28	13	11	25	38	12	0.0074
B069	A	15	12	24	28	20	12	18	14	11	19	10	15	26	11	11	44	17	14	15;20	28	13	11	25	38	12	0.0074
B071	A	18	14	17	32	20	11	17	15	14	23	10	13	24	10	11	40	17	14	11	37	13	12	25	38;41.2	11	0.0037
B074	A	18	13	20	31	20	11	16	15	11	21	10	15	21	11	11	40	19	14	17;18	33	15	12	24	37;40	11	0.0037
B075	A	15	12	24	28	19	12	18	14	11	19	10	15	26	11	11	44	18	14	15;20	28	13	11	25	38	12	0.0037
B078	A	15	12	24	28	20	11	17	14	11	19	10	15	25	10	11	46	19	14	16;20	28	13	11	25	38	12	0.0037
B080	A	18	14	17	33	20	11	18	15	12	23	9	13	24	10	11	42	17	14	11	35	13	12	25	38;41.2	11	0.0037
B081	A	17	13	21	30	20	11	17	16	11	20	10	16	21	11	11	39	17	14	17;19	31	14	11	24	38;39	11	0.0037
B082	A	16	13	23	32	20	11	16	15	12	21	10	15	22	11	11	41	19	14	16;17	29	13	11	28	35;39	11	0.0037
B083	A	14	14	21	31	19	10	15	15	12	21	10	15	21	11	11	40	19	14	16;19	28	14	11	30	36;39	11	0.0037
B085	A	18	13	19	29	21	10	18	15	11	21	10	15	24	10	11	37	17	14	14;15	31	14	13	28	41.2	11	0.0037
B088	A	15	14	17	30	23	11	14	15	10	0	11	15	22	12	10	37	16	15	14;19	30	13	12	25	36;40	9	0.0037
B090	A	18	12	17	29	20	11	17	16	9	19	11	15	22	11	10	39	18	14	17	29	13	12	27	36;37;39	10	0.0037
B091	A	16	13	22	31	21	10	17	16	12	22	10	15	21	11	11	38	19	14	14;19	28	12	12	26	37;38	11	0.0037
B093	A	16	13	21	30	20	11	17	16	11	21	10	15	21	11	11	43	16	14	16;17	31	15	12	23	38;39	12	0.0037
B094	A	19	14	17	32	20	11	18	15	13	23	10	13	24	10	11	41	17	14	11	34	13	12	26	38;41.2	11	0.0037
B098	A	18	14	17	32	20	11	17	15	13	23	10	13	24	10	11	40	17	14	11	35	13	14	26	38;41.2	11	0.0037
B099	A	16	13	22	30	21	10	16	16	11	21	10	14	21	11	11	43	17	14	16;18	31	14	12	26	38;39	11	0.0037
B100	A	15	14	22	31	19	10	17	15	12	21	10	15	0	11	11	38	19	14	15;17	28	13	12	25	37;39	11	0.0037
B101	A	19	14	17	32	20	11	17	15	13	23	10	13	24	10	11	41	18	14	11	36	13	12	26	38;41.2	11	0.0037

Table Legend:

ID – Sample Code **Pop** – Population Group **Freq** – Haplotype Frequency **A/I** – Asian/Indian **A** – African **Co** – Coloured **Ca** – Caucasian

– Marker with at least one duplication/triplication seen – Null allele – Microvariant – Duplication – Duplication and Microvariant – Triplication

ID	Pop	DYS 576	DYS 389I	DYS 635	DYS 389II	DYS 627	DYS 460	DYS 458	DYS 19	YGA TAH 4	DYS 448	DYS 391	DYS 456	DYS 390	DYS 438	DYS 392	DYS 518	DYS 570	DYS 437	DYS 385	DYS 449	DYS 393	DYS 439	DYS 481	DYF 387S1	DYS 533	Freq
B102	A	17	13	21	30	17	11	17	17	11	21	10	15	21	11	11	41	14	14	16;18	25	15	12	23	37;38	11	0.0037
B103	A	18	14	17	32	20	11	17	15	13	21	10	13	24	10	11	40	20	14	11	37	13	12	25	38;41.2	11	0.0037
B104	A	16	13	21	30	19	11	16	16	11	21	10	15	21	11	12	40	17	14	16	31	15	12	26	38;39	10	0.0037
B105	A	17	12	23	29	20	10	16	14	11	21	10	15	21	11	11	42	20	14	16	30	14	13	22	39	10	0.0037
B106	A	17	13	22	30	19	11	17	17	11	20	10	16	21	11	11	39	17	14	17;18	31	14	11	24	39	11	0.0037
B107	A	15	14	21	32	19	10	17	15	12	21	10	15	0	11	11	38	21	14	15;16	28	13	12	26	37;39	11	0.0037
B110	A	17	13	21	30	19	11	16	17	11	21	10	17	21	11	11	41	17	14	17;18	30	14	14	24	38;39	11	0.0037
B111	A	15	13	21	31	20	10	16	15	12	21	10	15	21	11	11	38	17	14	15;18	27	13	12	29	36;37	11	0.0037
B112	A	14	13	21	30	19	11	16	16	11	20	10	15	21	11	11	40	20	14	15;19	27	14	13	27	36;39	11	0.0037
B113	A	18	14	21	31	19	11	16	15	11	21	10	15	21	11	11	41	18	14	16;20	32	15	12	27	36;39	11	0.0037
B114	A	16	13	22	29	19	11	18	15	10	20	10	15	21	11	11	40	20	14	16	28	14	13	26	38;40	11	0.0037
B115	A	16	13	21	30	21	11	17	16	11	21	10	15	21	11	11	41	18	14	17	30	14	13	25	38	12	0.0037
B117	A	16	13	22	30	19	12	18	15	10	20	10	15	21	11	11	39	20	14	16	27	14	12	26	38;39	11	0.0037
B118	A	15	13	21	31	21	10	15	15	12	21	11	15	21	11	11	37	22	14	16;18	28	13	11	28	37;39	11	0.0037
B119	A	17	13	21	30	20	10	16	17	12	21	10	17	21	11	11	42	17	14	16;18	32	14	12	24	38;39	11	0.0037
B120	A	18	13	20	30	20	11	16	16	11	21	10	17	22	11	11	40	18	14	17;18	32	13	12	24	39	11	0.0037
B122	A	18	13	20	30	20	11	17	16	11	21	10	17	22	11	11	42	18	14	17;18	32	14	11	24	39	11	0.0037
B123	A	16	13	22	29	19	11	17	15	11	21	10	15	21	11	11	43	17	14	16;18	32	15	12	25	38;39	11	0.0037
B124	A	18	14	17	32	21	11	17	15	14	22	10	13	24	10	11	41	17	14	11	35	13	12	25	38;41.2	11	0.0037
B125	A	15	12	23	28	21	12	17	14	11	19	10	15	26	11	11	44	19	14	15;20	28	13	11	25	38	12	0.0037
B126	A	15	12	24	28	20	11	18	14	11	19	10	15	25	10	11	44	18	14	15;20	28	13	11	24	38	12	0.0037
B127	A	17	14	21	31	19	11	16	15	11	21	10	15	21	11	11	41	18	14	17;19	32	15	11	27	36;39	11	0.0037
B128	A	18	10	21	27	19	10	17	13	11	20	11	18	24	10	11	37	20	14	16;17	32	14	10	25	35;37	11	0.0037
B129	A	15	14	21	30	20	10	17	15	12	21	10	15	0	11	11	38	19	14	15;17	27	13	12	26	36;41	11	0.0037

Table Legend:

ID – Sample Code **Pop** – Population Group **Freq** – Haplotype Frequency **A/I** – Asian/Indian **A** – African **Co** – Coloured **Ca** – Caucasian

– Marker with at least one duplication/triplication seen – Null allele – Microvariant – Duplication – Duplication and Microvariant – Triplication

ID	Pop	DYS 576	DYS 389I	DYS 635	DYS 389II	DYS 627	DYS 460	DYS 458	DYS 19	YGA TAH 4	DYS 448	DYS 391	DYS 456	DYS 390	DYS 438	DYS 392	DYS 518	DYS 570	DYS 437	DYS 385	DYS 449	DYS 393	DYS 439	DYS 481	DYF 387S1	DYS 533	Freq
B130	A	16	13	22	30	21	10	15	15	11	20	10	16	21	11	11	39	17	14	17	32	14	12	26	38	11	0.0037
B200	A	15	12	24	29	20	12	18	14	11	19	10	15	26	11	11	43	17	14	15;20	28	13	11	25	38	12	0.0037
B_C	A	17	13	21	30	22	11	17	16	11	21	10	16	21	12	11	42	17	14	17;18	33	15	13	24	38	12	0.0037
B_F	A	15	14	21	32	20	10	17	15	12	21	10	15	0	11	11	38	21	14	15;16	28	13	12	26	36;39	11	0.0037
B_T	A	15	12	24	28	21	11	18	14	11	19	10	15	25	10	11	43	18	14	16;20	28	13	11	24	38	12	0.0037
B_TM	A	15	14	22	31	19	10	17	15	12	21	10	15	0	11	11	38	20	14	15;17	27	13	12	27	37;39	11	0.0037
B_TK	A	14	11	21	28	18	11	16	15	12	21	11	16	21	11	11	38	19	14	15;19	27	13	12	29	36;39;40	11	0.0037
C001	Co	17	12	17	29	18	11	18	16	9	19	11	15	22	11	10	39	18	14	17	29	13	12	28	36;39;41	10	0.0037
C002	Co	17	13	22	30	19	11	17	16	11	20	10	16	21	11	11	39	17	14	17;18	31	14	11	24	38;39	11	0.0037
C003	Co	17	13	21	29	21	11	15	15	10	20	10	14	23	10	12	40	19	14	15	27	15	11	26	37;38	13	0.0037
C004	Co	17	13	23	30	16	12	15	16	13	20	11	16	25	11	11	40	20	14	11;14	31	13	11	23	36;38	12	0.0037
C005	Co	15	13	22	31	19	10	17	15	12	21	10	15	21	11	11	38	19	14	15;16	31	13	12	28	36	11	0.0037
C007	Co	18	13	23	29	22	11	17	14	12	18	10	16	24	12	13	35	18	16	11;13	30	13	11	21	35;36	13	0.0037
C008	Co	17	13	29	28	19	10	17	15	11	22	8	16	19	10	11	39	15	14	12;15	30	14	11	25	32; 45.2	10	0.0037
C009	Co	18	13	21	29	20	10	18	15	12	19	10	17	22	9	11	39	17	14	15;17	32	12	11	24	37	12	0.0037
C010	Co	17	12	25	28	21	10	15	14	12	18	10	15	23	10	14	38	15	15	13;17	26	11	13	24	35;40	12	0.0037
C011	Co	19	13	23	29	21	12	16	14	12	19	11	15	23	12	13	41	18	14	11;14	29	12	12	22	35;36	12	0.0037
C012	Co	19	14	21	30	19	11	18	15	11	21	11	15	23	9	11	38	19	14	15;19	32	12	13	22	36;38	12	0.0037
C013	Co	17	13	22	29	19	10	16	15	12	19	10	16	22	9	11	38	17	14	15;17	32	12	11	23	36;37;40	11	0.0037
C014	Co	17	13	23	30	22	10	18	15	12	20	10	15	25	10	12	35	18	15	11;18	31	13	11	25	36;38	13	0.0037
C015	Co	18	13	23	29	16	11	15	16	13	20	10	15	25	11	11	41	19	14	11;14	32	13	12	24	38	12	0.0037
C017	Co	19	13	17	29	22	11	17	17	10	20	11	15	18	11	10	37	18	15	14;16	28	13	12	29	36;37	10	0.0037
C018	Co	14	12	21	28	17	10	16	15	12	21	10	17	21	11	11	39	19	14	17	28	13	11	28	37;39	11	0.0037
C019	Co	17	13	24	28	21	11	17	14	12	19	11	16	23	12	13	37	18	14	10;14	30	13	12	23	35;37	12	0.0037

Table Legend:

ID – Sample Code **Pop** – Population Group **Freq** – Haplotype Frequency **A/I** – Asian/Indian **A** – African **Co** – Coloured **Ca** – Caucasian

– Marker with at least one duplication/triplication seen – Null allele – Microvariant – Duplication – Duplication and Microvariant – Triplication

ID	Pop	DYS 576	DYS 389I	DYS 635	DYS 389II	DYS 627	DYS 460	DYS 458	DYS 19	YGA TAH 4	DYS 448	DYS 391	DYS 456	DYS 390	DYS 438	DYS 392	DYS 518	DYS 570	DYS 437	DYS 385	DYS 449	DYS 393	DYS 439	DYS 481	DYF 387S1	DYS 533	Freq
C020	Co	16	13	21	31	19	11	17	15	12	21	10	15	21	11	11	38	20	14	16;17	29	13	12	28	36;39	11	0.0037
C021	Co	15	13	24	32	20	11	16	15	13	21	10	15	22	11	11	41	19	14	16;17	29	13	12	28	35;39	11	0.0037
C022	Co	16	12	24	28	23	10	19	15	12	20	10	15	22	11	10	41	18	16	13;21	33	13	11	26	36;37	13	0.0037
C023	Co	16	13	24	29	22	10	15	13	12	20	10	15	24	10	11	40	19	15	17	33	13	11	20	34;37	12	0.0037
C024	Co	19	13	23	29	24	11	17	14	12	19	11	15	24	12	13	40	17	15	11;14	29	13	12	22	36;37	13	0.0037
C025	Co	18	13	23	29	23	11	17	15	12	19	10	16	23	12	13	38	17	15	11;14	30	13	12	22	35;36	11	0.0037
C026	Co	18	13	23	29	22	12	16	14	12	18	11	17	24	12	14	39	17	15	11;13	30	13	12	25	35;36	13	0.0037
C027	Co	18	13	23	29	24	10	19	14	12	19	11	15	24	13	13	37	18	15	11;14	30	13	13	23	35;37	11	0.0037
C028	Co	14	14	21	31	19	10	16	15	12	21	10	15	21	11	11	40	18	14	16;19	28	14	11	30	36;39	11	0.0037
C029	Co	14	12	20	29	19	10	16	15	12	22	10	15	22	10	11	35	17	16	14	33	13	11	21	39	9	0.0037
C030	Co	17	12	24	28	20	11	18	14	12	18	11	15	25	12	13	39	17	14	11;14	30	13	11	22	34;37	12	0.0037
C031	Co	15	14	21	31	19	10	17	15	12	21	10	15	21	11	11	39	19	14	16;18	28	14	12	29	36;38	11	0.0037
C032	Co	17	12	22	28	21	10	20	16	12	16	10	15	23	10	14	37	16	14	14;15	30	13	11	25	38	12	0.0037
C033	Co	18	14	17	32	19	11	16	15	11	23	11	13	24	10	11	39	17	14	11;13	34	13	13	25	38;41.2	11	0.0037
C034	Co	18	13	22	30	20	11	18	15	11	19	10	15	22	10	14	35	16	16	10;16	27	12	12	23	37;40	12	0.0037
C035	Co	15	13	23	27	22	11	18	14	12	19	10	18	23	12	13	37	16	15	11;15	29	13	11	22	35;37	12	0.0037
C036	Co	17	13	21	30	19	10	18	16	11	21	10	15	21	11	11	42	16	14	15;18	31	14	11	25	37;39	11	0.0037
C037	Co	17	14	17	31	21	10	13	16	10	0	10	14	23	12	10	38	17	15	16;17	34	13	13	27	37	8	0.0037
C038	Co	17	14	21	31	19	11	16	16	11	21	10	15	21	11	11	41	18	14	16;20	33	15	12	27	36;39	11	0.0037
C039	Co	15	13	17	30	23	10	13	17	11	0	10	15	22	12	10	38	16	14	15;17	38	13	13	25	36;37	8	0.0037
C040	Co	17	13	24	29	22	10	15	14	12	19	10	15	22	11	15	39	17	15	14;18	28	11	11	23	35;39	12	0.0037
C041	Co	17	13	22	29	19	11	15	15	11	20	10	16	23	10	11	39	20	14	15	28	14	11	27	37;38	12	0.0037
C100	Co	21	13	23	29	19	12	17	14	12	19	11	16	23	12	13	37	19	15	11;14	29	13	12	22	34	13	0.0037
C101	Co	17	13	21	30	20	11	16	15	11	22	10	17	21	11	11	42	17	14	16;18	31	14	11	25	40;41	11	0.0037
C102	Co	16	13	23	29	22	10	17	14	13	19	11	15	24	12	13	37	17	15	11;14	29	12	13	22	34;35	12	0.0037

Table Legend:

ID – Sample Code **Pop** – Population Group **Freq** – Haplotype Frequency **A/I** – Asian/Indian **A** – African **Co** – Coloured **Ca** – Caucasian

– Marker with at least one duplication/triplication seen – Null allele – Microvariant – Duplication – Duplication and Microvariant – Triplication

ID	Pop	DYS 576	DYS 389I	DYS 635	DYS 389II	DYS 627	DYS 460	DYS 458	DYS 19	YGA TAH 4	DYS 448	DYS 391	DYS 456	DYS 390	DYS 438	DYS 392	DYS 518	DYS 570	DYS 437	DYS 385	DYS 449	DYS 393	DYS 439	DYS 481	DYF 387S1	DYS 533	Freq
C103	Co	18	14	17	32	20	11	18	15	13	23	10	13	24	10	11	41	17	14	11	35	13	12	25	38;40.2	11	0.0037
W_J	Ca	17	12	21	28	20	10	15	16	11	20	11	15	22	11	11	39	21	16	10;13	28	14	11	25	37;40	11	0.0037
W001	Ca	18	13	23	29	22	12	18	14	10	19	11	15	24	12	13	36	17	16	9;14	30	13	12	22	36	12	0.0037
W002	Ca	18	13	23	29	20	11	16	14	12	19	10	16	24	12	13	37	19	15	11;15	28	13	12	22	35;36	13	0.0037
W003	Ca	18	14	22	31	18	11	18	13	11	19	10	15	26	11	13	36	18	15	14;18	30	13	11	24	36;38	12	0.0037
W005	Ca	19	14	23	30	21	10	18	14	11	19	9	15	24	12	13	38	18	15	11;14	30	13	12	22	35;36	12	0.0037
W006	Ca	19	13	23	30	21	10	18	14	11	19	11	16	24	12	13	40	17	15	11;14	29	13	12	22	35;36	12	0.0037
W007	Ca	17	13	23	29	21	11	17	14	12	19	11	15	24	12	13	36	17	15	11;14	29	13	12	23	35;36	11	0.0037
W008	Ca	17	14	23	29	22	10	16	14	11	18	11	15	23	12	13	40	17	14	11;14	30	13	12	22	35;36	12	0.0037
W009	Ca	17	13	23	29	23	10	19	14	12	19	11	15	24	12	14	38	19	15	11;14	29	13	13	22	35;36	12	0.0037
W010	Ca	17	13	20	30	20	11	16	16	11	21	10	15	22	10	12	38	20	14	14;15	28	14	11	25	36;38	12	0.0037
W011	Ca	16	12	22	28	21	10	15	13	11	19	11	14	24	10	11	37	19	16	14;15	30	13	12	27	37;38	11	0.0037
W012	Ca	17	12	24	28	19	11	15	15	10	21	10	15	22	10	11	40	19	15	13;14	26	13	11	26	38	11	0.0037
W013	Ca	18	14	23	30	20	11	17	14	12	19	11	16	23	12	13	38	18	15	11;14	30	13	12	22	35;36	12	0.0037
W014	Ca	14	12	21	28	20	10	15	14	11	20	10	14	23	10	11	36	20	16	13;15	28	13	11	25	37	11	0.0037
W015	Ca	16	13	23	30	20	11	18	14	12	19	11	16	24	12	13	37	19	15	11;14	29	13	12	23	35;36	13	0.0037
W017	Ca	17	13	22	30	23	10	15	14	12	20	10	15	25	10	11	40	19	14	16;18	34	13	12	22	36;37	12	0.0037
W019	Ca	17	12	22	28	18	10	15	14	11	20	10	14	22	10	11	39	19	16	13	28	13	11	26	37;38	11	0.0037
W020	Ca	17	13	23	29	23	11	17	15	12	19	10	16	24	12	13	36	16	15	11;14	30	13	11	23	35;37	12	0.0037
W022	Ca	16	12	22	28	20	10	17	15	14	20	10	14	22	10	11	38	20	15	13;14	30	13	11	26	37;39	11	0.0037
W023	Ca	16	12	22	28	23	10	16	14	12	19	11	16	25	12	13	37	18	15	11;15	29	13	12	22	34;36	12	0.0037
W024	Ca	18	13	23	29	24	11	16	15	12	19	11	15	23	12	13	39	17	15	11;12	30	13	11	23	35;36	12	0.0037
W026	Ca	18	13	23	31	23	11	17	14	12	18	10	15	25	12	14	37	17	15	11;13	32	13	12	25	35;36	12	0.0037
W028	Ca	17	12	23	30	19	11	15	16	12	19	11	15	25	11	11	40	19	14	11;14	35	13	10	23	37;38	12	0.0037
W029	Ca	21	13	21	30	21	11	17.2	14	11	21	11	15	23	10	11	39	18	14	13;15	27	12	12	24	34;38	11	0.0037

Table Legend:

ID – Sample Code **Pop** – Population Group **Freq** – Haplotype Frequency **A/I** – Asian/Indian **A** – African **Co** – Coloured **Ca** – Caucasian

– Marker with at least one duplication/triplication seen – Null allele – Microvariant – Duplication – Duplication and Microvariant – Triplication

ID	Pop	DYS 576	DYS 389I	DYS 635	DYS 389II	DYS 627	DYS 460	DYS 458	DYS 19	YGA TAH 4	DYS 448	DYS 391	DYS 456	DYS 390	DYS 438	DYS 392	DYS 518	DYS 570	DYS 437	DYS 385	DYS 449	DYS 393	DYS 439	DYS 481	DYF 387S1	DYS 533	Freq
W030	Ca	17	13	23	29	22	10	16	15	11	18	11	16	23	12	13	38	17	14	11;15	29	13	13	22	35;38	11	0.0037
W031	Ca	18	13	23	29	24	11	17	14	12	18	11	17	25	12	14	39	17	14	14	30	13	12	24	36;37	12	0.0037
W032	Ca	18	13	23	30	20	11	18	14	11	19	11	17	23	12	13	39	18	15	11;14	30	14	12	22	36;37	12	0.0037
W033	Ca	18	13	23	29	24	11	16	14	12	19	11	16	24	12	13	37	17	15	11;14	30	12	12	22	35	12	0.0037
W034	Ca	17	14	20	32	21	11	15	15	10	20	10	14	23	10	12	37	20	14	14;16	28	14	11	27	37	11	0.0037
W038	Ca	16	12	21	28	20	11	15	14	10	20	10	16	22	10	11	39	19	16	13;14	28	13	11	25	38	11	0.0037
W039	Ca	16	13	23	29	21	11	16	14	12	19	10	15	25	12	13	38	18	15	11;15	28	13	12	22	35	12	0.0037
W040	Ca	17	13	23	30	22	11	17	14	13	21	10	14	25	9	14	44	18	14	12	29	13	12	27	38	12	0.0037
W041	Ca	19	13	23	29	21	12	16	15	12	19	11	15	23	12	13	42	18	14	11;14	29	12	13	22	35	12	0.0037
W043	Ca	18	15	23	32	16	11	15	17	11	19	10	16	25	11	11	41	20	14	11;14	33	13	11	24	38	12	0.0037
W044	Ca	15	14	23	32	19	10	18	15	12	21	10	17	24	10	11	37	16	16	13;16	28	13	11	19	39	10	0.0037
W046	Ca	19	13	23	30	21	10	20	14	11	19	11	16	24	12	13	39	17	15	11;14	29	13	12	22	35;36	12	0.0037
W047	Ca	16	12	22	29	19	11	15	14	11	20	10	13	22	10	11	38	20	16	13;16	28	13	11	25	37	11	0.0074
W048	Ca	16	12	23	29	19	10	16	14	11	20	10	15	24	10	11	41	20	16	13;14	27	13	11	25	38	11	0.0037
W049	Ca	18	14	23	31	22	11	17	14	12	20	11	15	24	12	13	39	17	15	11;15	30	13	12	23	35;36	12	0.0037
W050	Ca	18	13	23	29	16	11	16	16	12	19	10	18	25	11	11	44	18	14	12;14	32	13	11	25	37;38	11	0.0037
W051	Ca	17	13	25	29	21	11	16	14	11	19	11	16	24	12	13	38	18	15	11;14	29	12	11	22	35;36	12	0.0037
W053	Ca	22	13	20	29	20	11	15	15	11	19	10	16	23	10	12	39	19	14	15;17	28	14	11	26	36;40	12	0.0037
W054	Ca	17	14	23	29	21	10	16	14	11	18	11	16	23	12	13	40	17	14	11;14	29	13	13	22	35;36	12	0.0037
W055	Ca	16	13	25	29	20	10	18	14	12	19	11	16	23	13	13	38	17	15	11;14	30	13	11	20	35;36	12	0.0037
W056	Ca	17	14	23	30	20	10	16	14	10	20	11	16	23	9	11	39	19	15	12	33	12	11	23	37;40	11	0.0037
W057	Ca	16	12	23	28	23	11	17	14	11	19	11	17	25	11	14	36	17	15	11;14	31	13	12	23	36;37	13	0.0037
W058	Ca	17	12	22	28	19	10	15	14	11	20	10	14	23	10	11	39	18	16	14;15	28	13	11	26	36;38	12	0.0037
W059	Ca	19	13	23	29	18	11	16	14	12	19	11	16	24	12	13	39	16	15	11;15	30	13	11	24	35;37	13	0.0037

Table Legend:

ID – Sample Code **Pop** – Population Group **Freq** – Haplotype Frequency **A/I** – Asian/Indian **A** – African **Co** – Coloured **Ca** – Caucasian

– Marker with at least one duplication/triplication seen – Null allele – Microvariant – Duplication – Duplication and Microvariant – Triplication

ID	Pop	DYS 576	DYS 389I	DYS 635	DYS 389II	DYS 627	DYS 460	DYS 458	DYS 19	YGA TAH 4	DYS 448	DYS 391	DYS 456	DYS 390	DYS 438	DYS 392	DYS 518	DYS 570	DYS 437	DYS 385	DYS 449	DYS 393	DYS 439	DYS 481	DYF 387S1	DYS 533	Freq
W060	Ca	17	13	23	29	21	11	17	14	11	19	11	16	23	12	13	38	17	16	11;14	29	14	11	23	34;35	12	0.0037
W061	Ca	17	13	23	30	22	10	20	14	11	19	11	15	23	12	13	38	17	15	11;14	30	13	11	22	35;36	12	0.0037
W062	Ca	16	12	21	28	21	10	18	15	11	20	10	14	23	10	11	39	16	16	13;15	30	13	11	21	37;38	11	0.0074
W063	Ca	17	14	21	33	21	11	15	15	10	20	10	14	23	10	12	38	20	14	14;15	28	14	11	27	37	11	0.0037
W064	Ca	16	12	21	29	20	10	16	13	11	20	10	14	22	10	11	39	19	16	13;14	27	13	11	25	37;39	11	0.0037
W065	Ca	17	13	24	30	22	10	19	14	11	19	11	15	23	12	13	39	17	15	11;14	29	13	11	22	35;36	12	0.0037
W066	Ca	18	13	23	29	22	11	17	14	12	19	10	16	25	12	13	38	19	15	11;15	28	13	13	23	36	12	0.0037
W068	Ca	20	13	23	29	22	11	17	14	12	19	11	15	24	12	13	36	19	15	11;14	29	14	13	24	35;36	12	0.0037
W070	Ca	16	12	21	28	21	10	18	15	11	20	10	14	23	10	11	39	16	16	13;15	30	13	11	21	37;38	11	0.0074
W072	Ca	16	12	22	28	21	10	15	13	11	19	11	14	24	10	11	37	19	16	14;15	29	13	13	27	37	11	0.0037
W073	Ca	16	12	21	28	20	10	14	14	11	20	11	14	22	10	11	39	20	16	13;14	29	13	12	25	38;39	11	0.0037
W074	Ca	15	13	23	29	22	11	18	14	12	19	11	15	24	12	13	36	16	15	11	30	13	12	23	35	12	0.0037
W075	Ca	18	13	23	29	22	11	16	15	12	19	11	15	25	12	14	40	17	15	11;14	29	13	12	22	35	12	0.0037
W076	Ca	18	13	23	29	20	12	17	14	11	19	11	17	23	12	13	39	18	15	12;14	29	13	12	22	36	12	0.0037
W077	Ca	17	14	21	32	21	10	17	14	11	19	10	15	23	9	13	36	17	14	14;15	35	13	11	22	37	12	0.0037
W078	Ca	16	12	21	28	19	11	15	14	10	20	10	16	22	10	11	39	19	16	13;14	28	13	11	25	38	11	0.0037
W079	Ca	15	12	20	29	15	9	17	16	13	21	10	14	22	10	11	38	18	16	14	29	14	10	21	36;38	9	0.0037
W080	Ca	17	13	23	29	17	11	16	17	12	20	10	16	25	11	11	41	18	14	10;14	31	13	12	24	36;38	12	0.0037
W081	Ca	17	13	24	29	22	11	17	14	11	19	11	16	23	12	13	38	17	15	11;14	29	13	11	22	35;37	12	0.0037
W082	Ca	16	12	21	28	18	10	16	14	11	20	10	14	23	10	11	39	21	16	13;14	27	14	11	25	38;39	11	0.0037
W083	Ca	17	13	23	29	23	11	17	14	12	19	11	15	24	12	13	38	17	15	11;14	29	13	12	21	35;36	11	0.0037
W084	Ca	17	13	21	29	20	9	15	13	11	19	11	17	24	10	11	41	21	14	17;18	30	13	12	23	35;38	12	0.0037
W085	Ca	18	13	23	29	21	10	17	14	12	19	10	16	24	12	13	37	16	15	11;14	30	13	13	22	35;38	12	0.0037
W087	Ca	17	13	23	31	20	9	16	13	11	20	9	16	24	10	11	40	19	14	16;18	31	13	12	22	36;37	12	0.0037

Table Legend:

ID – Sample Code **Pop** – Population Group **Freq** – Haplotype Frequency **A/I** – Asian/Indian **A** – African **Co** – Coloured **Ca** – Caucasian

– Marker with at least one duplication/triplication seen – Null allele – Microvariant – Duplication – Duplication and Microvariant – Triplication

ID	Pop	DYS 576	DYS 389I	DYS 635	DYS 389II	DYS 627	DYS 460	DYS 458	DYS 19	YGA TAH 4	DYS 448	DYS 391	DYS 456	DYS 390	DYS 438	DYS 392	DYS 518	DYS 570	DYS 437	DYS 385	DYS 449	DYS 393	DYS 439	DYS 481	DYF 387S1	DYS 533	Freq
W088	Ca	17	14	23	31	22	11	17	14	12	20	11	15	24	12	13	39	17	15	11;15	30	13	12	23	35;36	12	0.0037
W089	Ca	18	13	24	30	17	11	17	16	12	20	10	16	25	11	11	41	17	14	10;14	31	13	10	25	36;38	12	0.0037
W090	Ca	16	14	25	30	21	10	18	14	12	19	10	16	23	12	13	36	17	14	11;15	29	13	11	22	35;36	12	0.0037
W091	Ca	14	13	23	30	21	10	15	13	12	21	11	15	25	10	11	39	19	14	15;17	32	14	12	25	37	11	0.0037
W092	Ca	16	12	22	28	19	11	16	14	11	20	10	14	23	10	11	39	19	16	13;14	29	13	11	26	38;39	11	0.0037
W093	Ca	18	13	23	30	21	10	18	14	11	20	11	16	24	12	13	39	17	15	11;14	29	13	12	22	35;36	12	0.0037
W094	Ca	17	13	23	31	17	11	15	17	13	20	11	16	25	11	11	41	17	14	11;14	31	13	10	25	36;38	12	0.0037
W096	Ca	17	13	23	29	21	10	16	14	12	20	11	16	23	12	13	41	17	15	11;14	30	13	12	22	36	12	0.0037
W097	Ca	20	13	23	29	20	11	17	15	11	19	10	16	24	12	12	37	18	15	11;15	31	13	12	23	35;36	13	0.0037
W098	Ca	18	13	23	29	23	11	17	14	12	19	11	16	25	12	13	38	18	15	13;14	30	13	11	22	36	11	0.0037
W099	Ca	12	12	20	28	18	10	16	14	12	21	9	15	22	10	11	37	17	16	14	27	14	13	21	38;39	9	0.0037
W100	Ca	17	14	23	29	22	10	16	14	11	18	11	15	23	12	13	40	17	14	11;14	29	13	13	22	35;36	12	0.0037
W101	Ca	18	14	21	31	19	11	15	15	11	20	10	16	23	10	11	39	20	14	15	28	14	11	27	37;38	12	0.0037
W102	Ca	17	13	24	29	20	11	18	14	12	18	11	16	26	12	13	37	18	15	11;15	31	13	12	22	35;38	12	0.0037
W103	Ca	18	13	23	29	21	10	16	14	12	19	11	15	25	12	13	38	19	15	12;14	32	12	12	23	35;36	12	0.0037
W104	Ca	16	12	22	29	19	11	15	14	11	20	10	13	22	10	11	38	20	16	13;16	28	13	11	25	37	11	0.0074
W105	Ca	17	13	23	29	22	10	16	14	12	19	11	16	23	12	13	40	17	15	11;15	30	13	12	22	35;36	12	0.0037
W106	Ca	17	13	22	30	24	10	15	15	12	20	10	17	25	10	11	40	19	14	16;18	34	13	12	22	36;37	12	0.0037
W108	Ca	17	13	22	30	24	10	15	14	12	20	10	17	25	10	11	40	19	14	16;18	34	13	12	22	35;37	12	0.0037
W109	Ca	17	13	23	29	17	11	16	17	12	20	10	16	24	11	11	40	18	14	10;14	31	13	12	24	36;38	12	0.0037
W110	Ca	17	14	24	31	23	11	16	15	12	19	10	16	25	12	13	37	17	15	11;14	29	13	12	21	35;36	12	0.0037
W111	Ca	18	13	23	29	22	12	18	14	12	19	11	14	24	12	13	39	17	15	11;14	31	13	12	22	36	11	0.0037
W112	Ca	16	12	23	27	19	11	16	14	11	20	11	14	22	10	11	40	20	16	13	28	13	11	25	38;39	10	0.0037
W113	Ca	16	13	23	29	21	11	17	15	11	18	11	16	24	12	13	37	18	14	12;16	28	13	13	22	35;36	13	0.0037

Table Legend:

ID – Sample Code **Pop** – Population Group **Freq** – Haplotype Frequency **A/I** – Asian/Indian **A** – African **Co** – Coloured **Ca** – Caucasian

– Marker with at least one duplication/triplication seen – Null allele – Microvariant – Duplication – Duplication and Microvariant – Triplication

ID	Pop	DYS 576	DYS 389I	DYS 635	DYS 389II	DYS 627	DYS 460	DYS 458	DYS 19	YGA TAH 4	DYS 448	DYS 391	DYS 456	DYS 390	DYS 438	DYS 392	DYS 518	DYS 570	DYS 437	DYS 385	DYS 449	DYS 393	DYS 439	DYS 481	DYF 387S1	DYS 533	Freq
W114	Ca	17	13	23	29	22	10	18	14	12	19	11	17	24	12	9	39	17	15	11;15	27	13	12	22	35;36	12	0.0037
W115	Ca	16	13	25	29	21	10	16	15	12	19	10	17	23	12	13	36	17	15	11;15	30	13	11	20	35;36	12	0.0037
W116	Ca	17	13	21	30	22	9	15	13	12	20	10	17	24	10	11	41	19	14	16;18	33	14	13	23	35;38	12	0.0037
W117	Ca	20	13	23	29	21	12	16	14	12	19	11	15	23	12	13	41	18	14	11;14	29	12	12	22	35	12	0.0037
W118	Ca	19	13	25	31	16	11	16	15	12	19	10	15	25	11	11	40	20	14	10;11	31	13	10	25	38	12	0.0037
W119	Ca	18	13	25	30	20	11	16	14	11	19	11	16	24	12	13	37	18	15	11;14	29	12	11	22	35;36	12	0.0037
W120	Ca	19	13	23	30	16	11	15	16	13	20	10	15	25	11	11	41	17	14	12	32	13	10	23	37;38	11	0.0037

Table Legend:

ID – Sample Code **Pop** – Population Group **Freq** – Haplotype Frequency **A/I** – Asian/Indian **A** – African **Co** – Coloured **Ca** – Caucasian

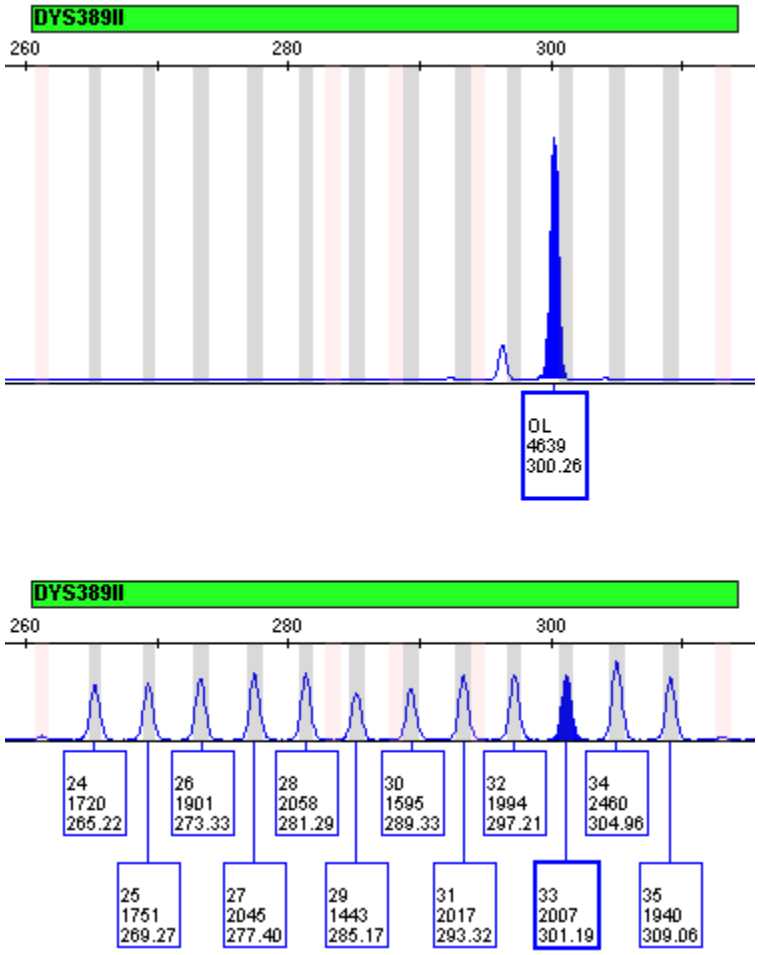
– Marker with at least one duplication/triplication seen – Null allele – Microvariant – Duplication – Duplication and Microvariant – Triplication

APPENDIX F

Calculation of the off-ladder (OL) and off marker range (OMR) alleles

Off-ladder (OL) alleles were calculated using the size of the allele (in bases) and the size of the closest allele from the allelic ladder. These OLs were confirmed by reamplifying the duplicate swab for samples B047, B032, C103, and C008.

Allele 32.3 at *DYS389II* for sample B047

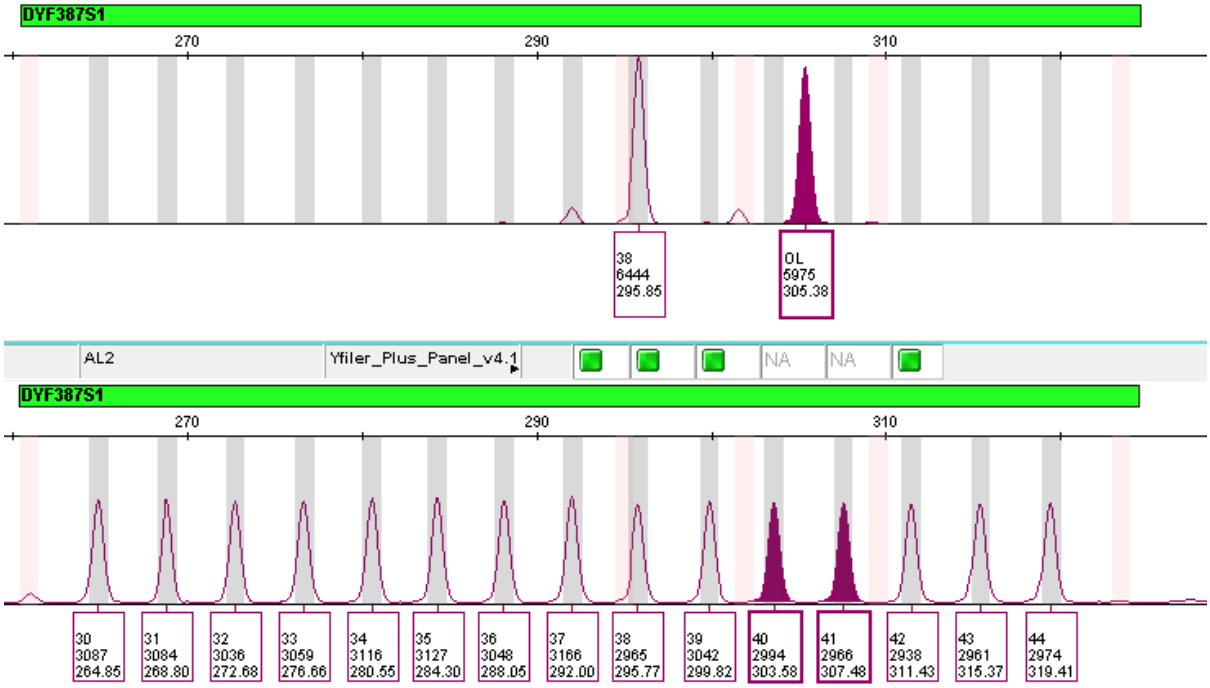


This allele occurred as a single allele at the *DYS389II* marker, between the 32 and 33 bins. It is closer to the 33 bin, so the 33 allele from the allelic ladder was used in this calculation.

$$Size\ of\ 33\ allele - Size\ of\ OL$$
$$= 301.9\ bases - 300.26\ bases$$
$$= 0.93\ base$$
$$\approx 1\ base$$

Therefore, the OL allele is 1 base shorter than the 33 allele. *DYS389II* has a tetranucleotide repeat motif, meaning that the last incomplete repeat only has 3 bases following the 32nd full repeat. Therefore, this allele can be labelled as the microvariant 32.3.

Allele 40.2 at *DYF387S1* for sample C103

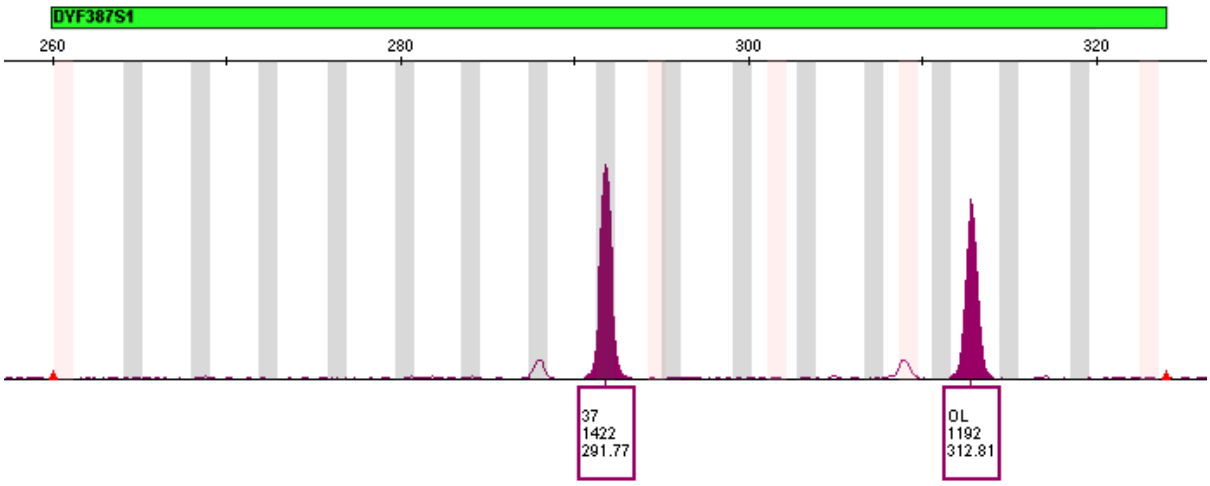


This allele occurred as part of a duplication at the *DYF387S1* marker, between the 40 and 41 bins. In this case, the 41 allele from the allelic ladder was used in the calculation.

$$\begin{aligned} &Size\ of\ 41\ allele - Size\ of\ OL \\ &= 307.48\ bases - 305.38\ bases \\ &= 2.1\ bases \\ &\approx 2\ bases \end{aligned}$$

Therefore, the OL allele is 2 bases shorter than the 41 allele. *DYS389II* has a tetranucleotide repeat motif, meaning that the last incomplete repeat only has 2 bases following the 40th full repeat. Therefore, this allele can be labelled as the microvariant 40.2.

Allele 42.2 at *DYF387S1* for sample B032



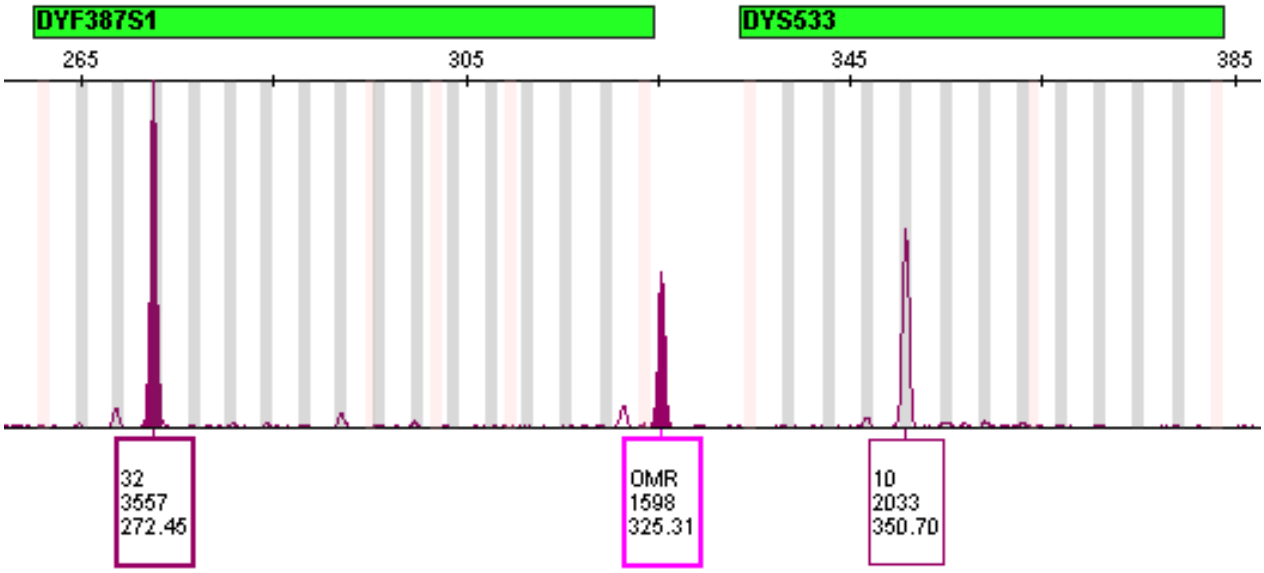
This allele occurred as part of a duplication at the DYF387S1 marker, between the 42 and 43 bins. In this case, the 43 allele from the allelic ladder was used in the calculation.

$$Size\ of\ 43\ allele - Size\ of\ OL$$
$$= 315.37\ bases - 312.81\ bases$$
$$= 2.56\ bases$$
$$\approx 2\ bases$$

Therefore, the OL allele is 2 bases shorter than the 43 allele. *DYS389II* has a tetranucleotide repeat motif, meaning that the last incomplete repeat only has 2 bases following the 42nd full repeat. Therefore, this allele can be labelled as the microvariant 42.2.

Allele 45.2 at *DYF387S1* for sample C008

This off marker range (OMR) peak was confirmed to be a true allele by amplifying the duplicate swab. It is much more likely that this allele has occurred as a part of the *DYF387S1* marker than *DYS533*. Visually and in terms of size at 325.31 bases, the allele is much closer to the range of *DYF387S1* (~260 to ~320 bases) than *DYS533* (~335 to ~385 bases). If it was allocated to *DYS533*, it would be calculated to be allele 4.3, which is well below the smallest observed allele, allele 6, at this locus (www.yhrd.com). There are also reports of a 45.2 microvariant being observed at *DYF387S1* (www.yhrd.com). Despite this speculation and justification, only sequencing of this sample would be able to truly conclude which marker this allele would fall under and how many repeats it would have.



Therefore, this allele occurred as part of a duplication at the *DYF387S1* marker, after the 44 bin, outside the marker range. The allelic ladder was not used for this calculation as the marker fell outside the marker range. There is a second allele detected in the same marker region, and that allele (32) is used in this calculation.

$$Size\ of\ OL - Size\ of\ 32\ allele$$
$$= 325.31\ bases - 272.45\ bases$$
$$= 52.86\ bases$$

DYF387S1 has a tetranucleotide repeat motif, so the number of bases is converted to the number of repeats by dividing by four.

$$Number\ of\ repeats = \frac{Number\ of\ bases}{4}$$
$$= \frac{52.86\ bases}{4}$$
$$= 13.215\ repeats$$

Therefore, the OL allele is 13.2 repeats more than the 32 allele, and this OL can be labelled as a 45.2 microvariant.