Validation and Application of a Highly Discriminating and Rapid 10-Locus Y-STR DNA Profiling System

Mohaimin Kasu

A thesis submitted for the partial fulfilment of a degree of Doctor of Philosophiae (Biotechnology)

Forensic DNA Laboratory

Department of Biotechnology

13 March 2019

Supervisor: Professor Maria Eugenia D’Amato

Keywords

DNA profiling

Forensic Genetics

Polymerase Chain Reaction

Rapidly mutating

Sexual assault

South Africa

Y-chromosome Short Tandem Repeats (Y-STRs)

Abstract

Mohaimin Kasu

PhD thesis title:

Validation and Application of a Highly Discriminating and Rapid 10-Locus Y-STR DNA Profiling System

DNA profiling the male specific region on the Y-chromosome is fundamental to forensic practise. Its recognised as a powerful analytical tool for investigation of sexual assault when the DNA evidence is highly admixed. Standard practises for processing sexual assault evidence include physically separate the sperm cell from the female fraction using differential extraction followed by autosomal DNA profiling. However, under specific scenarios of assault physical separation may not be possible due to the nature of the evidence.

The research presented in this thesis was focused on the development and validation of the UniQTyper™ Y-10 prototype for male specific DNA profiling. The prototype which contains 10 Y-STR markers was developed and validated to deliver a rapid and cost-effective system while maintaining a forensic applicable level of performance. An allelic ladder is produced with an allele cloning approach for which an overview of the workflow and technicalities presented herein is aimed to assists an efficient bulk production process.

In a second component novel sequence variation was reported across 153 sequenced alleles and submitted to Genbank. In this output the Y-STR panel was perused beyond the scope of length polymorphisms. In a proof of concept, its potential to discriminate between shared allele sizes by characterizing sequence structure variations is discussed.

In a final component we generate the largest Y-STR survey across South Africa to establish reference data and to comprehensively assess the forensic genetics parameters for the UniQTyper™ Y-10.

Declaration

I declare that “Validation and Application of a Highly Discriminating and Rapid 10-Locus Y-STR DNA Profiling System” is my own work, that it has not been submitted for any degree or examination in any other university, and that all the resources I have used or quoted, and all work which was the result of joint effort, have been acknowledged by complete references.

Mohaimin Kasu

Signature: ………………. Date: 14/12/2018

Acknowledgements

A sincere thanks you to my supervisor professor Maria Eugenia D'Amato for this PhD appointment. From my unscheduled interview appearance to your office with all my family present, you gave me this opportunity for which I'm forever grateful. I thank you for your belief in me, the courage you have installed in me and for your dedicated support through this research. Your wisdoms have been invaluable in overcoming many challenges and your perseverance admirable, I thank you from the bottom of my heart.

I would like to thank all the members of the Forensic DNA Laboratory for the upliftment during the days or even weeks when experimental results were not in favour. Thank you for all the knowledge you shared contributing to this research and assistance with experimental work. A special thank you to Dr Peter Ristow for taking the time to critically evaluate my results and for the training in R-Studio and STR-validator. A sincere thank you to Mr Kevin Cloete for the technical support and especially the administrative support.

I extend my appreciation to the staff of Inqaba biotechnical industries, particularly Mrs Mischa Fraser and Dr Christiaan Labuschagne for the cloning and sequencing aspect of the research. Thank you for responding to my thousands of emails and constant update requests.

I would like to express my deep gratitude to my family and friends, for without your loving support this would not have been possible. For keeping me calm, positive and helping me fit the pieces of my life together. I'm forever grateful for your love.

I would like to acknowledge the National Research Foundation of South Africa (NRF) for funding my PhD bursary. I also like to thank the Technology Innovation Agency (TIA) for seed funding and the South Africa Network for Bioscience (SANBio) BioFISAII Flagship Projects for funding experimental costs. This work was also funded by Inqaba biotechnical industries the commercial partner and manufacturer of the UniQTyper™ Y-10.

Abbreviations

°C	degrees celsius
ABI	Applied Biosystems (Thermo Fisher)
AJ	Ashkanazi Jew
AMOVA	Analysis of Molecular Variance
bp	base pair
C	Coloured
CE	Capillary Electrophoresis
cm	Centimetre
DC	Discrimination Capacity
DNA	Deoxyribonucleic acid
dNTP	Deoxyribonucleotide triphosphate
DYS	Deoxyribonucleic acid Y-chromosome segment
Eng	English
FTA	Flinders Technology Associates
GD	Gene Diversity
h	Hour
Hap	Haplotype
HD	Haplotype diversity
HiDi	Highly Deionized
hr	Hours
HRM	High Resolution Melt
I	Indian
KA	Known Alleles
M	Molar
MDS	Multi Dimensional Scaling
mg	Milligram
min	Minutes
ml	Millilitre
mM	Millimolar
MP	Mach Probability
N	Number of individuals/samples
NA	Not Available
NaCl	Sodium Chloride
NAS	Novel Allele Sequence
Ng	Nanogram

NGS	Next Generation Sequencing
NIST	National Institute of Standards and Technology
P	Pedi
PCR	Polymerase Chain Reaction
pg	Picogram
PM	Pattern Match
RFLP	Restriction Fragment Length Polymorphism
RFU	Relative Fluorescent Unit
RM Y-STR	Rapidly mutating Y-Chromosome Short Tandem Repeats
rpm	ramp per minute
RPV	Repeat Pattern Variants
Sd	Standard deviation
SDS	Sodium Dodecyl Sulphate
SGM	Second Generation Multiplex
SH	Size Homoplasy
SNP	Single Nucleotide Polymorphism
STR	Short Tandem Repeat
STRbase	Short Tandem Repeat DNA internet database
SWGDAM	Scientific working Group on DNA Analysis Methods
TE	Tris – EDTA Tris 2-amino-2 (hydroxymethyl)-1,3-propandiol
TS	Tsumkwe San
VNTR	Variable Number Tandem Repeat
X	Xhosa
YHRD	Y-chromosome haplotype reference database
Y-STR	Y Chromosome Short Tandem Repeat
µg	Microgram
µl	Microlitre
µM	Micromolar

List of Figures:

List of Tables

Background

The incidence of sexual offences in South Africa (SA) and neighbouring countries is amongst the highest in the world. The statistics in South Africa indicates a major crisis of gang rape against adolescents which includes victims younger than one years old (Hirschowitz et al. 2000 and Jewkes et al. 2009). The heinous nature of these crimes has been directly associated with the misconception that sexual intercourse with a virgin cures HIV/AIDS, commonly referred as the "virgin cure" (Groce and Trasi 2004 and Meel 2003). Over the past 20 years, as a result of the high incidence of sexual offences the casework demand has increased dramatically and evidently met with poor conviction rates (Jewkes et al. 2009). https://www.gov.za/documents/crime-statistics-20172018-11-sep-2018-0000).

DNA profiling systems that specifically identify male DNA using Y-chromosome markers may improve the conviction rates of sexual assault. Although several commercial Y-chromosome profiling kits are available, most of its DNA core markers are limited to maximize discrimination in South African Bantu men. Due to extensive practices of patrilocality and the historically isolated nature of certain metapopulations, the discriminatory capacity of commercial Y-chromosome markers maybe less informative to assist a conviction. Currently the Forensic community of South Africa are in dire need of a highly discriminatory commercial Y-chromosome DNA profiling kit. Our research outputs are focused on providing such a development in partnership with the industrial partner Indaba Biotechnical Industries. This development aims to offer cost effective and high throughput DNA profiling to assist the increase casework demand and poor conviction rates. Ultimately the intent of this research capacity is to protect society and provide justice to the innocent, vulnerable and wrongfully accused.

Contents

Chapter 1:

1.0 Introduction

1.1 Forensic DNA typing

It is well known that humans share a 99.7% genetic code identity and only genetically dissimilar in 0.3% of the genome. In forensic genetics we exploit DNA sequence variations within this minor variable regions to identify individual specific differences. DNA typing techniques for human identification has become a fundamental tool in civil and criminal legal proceeding (Jobling and Gill 2004). DNA typing is central to the forensic dogma which states that every contact leaves a trace. In essence, the victim, perpetrator and crime scene can all exchange microscopic and macroscopic biological evidence between each other which may be profiled by extracting the DNA content (Jobling and Gill 2004). The first successful conviction with DNA typing was achieved using traces of semen found in dual murder case in which two female victims were sexually assaulted (Wambaugh 1989). This application of DNA profiling took place in 1986, one year after Sir Alec J Jeffrey discovered the value of repetitive DNA elements called Variable Number of Tandem Repeats (VNTRs) for human identification (Jeffreys et al.1985). The conventional term "DNA fingerprinting" as referred to by Alec Jeffreys defines an abstraction of genetic elements presenting patterns that are unique an individual. These VNTRs were highly polymorphic (variable) in number and could be separated by size using the technique called restriction fragment length polymorphism (RFLP) to generate a unique DNA barcode (Jeffreys et al.1985). It was directly after VNTR application in the 1986 dual murder case that DNA profiling became widely recognised by the forensic community as a powerful analytical tool which can assist exoneration or conviction of the excused.

Although VNTRs were highly discriminatory, it’s typing methodologies were quickly recognised as the major limiting factor due to its lack of sensitivity and power to resolve mixtures of DNA (Panneerchelvam and Norazmi 2003). Its procedure which involved digesting the DNA with restriction enzymes and fragment separation with gel electrophoresis required a large amount of DNA (0.5-10 µg), was low throughput and labour intensive (Gill et al. 1985 and Tamaki and Jeffreys 2005). It was by the late 1980s that two more ground-breaking discoveries together lead the revolution giving rise to the mainstream DNA typing procedures to date. The invention of the polymerase chain reaction (PCR) by Mullis et al. 1986 allowed for an exponential magnification of minute amounts of DNA evidence to produce billions of copies of a selectively targeted DNA fragment. This PCR approach which could overcame the sensitivity limitation of RFLP was only applied to VNTR analysis in 1992 as during this period the beginning of a new exiting DNA marker emerged in forensics. Short Tandem Repeat (STR) polymorphisms as described by Edward et al.1991 arrived quickly onto the forensic scene. These elements which are 2-7 bp repeat motifs were more resistant to degradation due to its smaller length while at the same time remained highly variable in location and number of repeats (Gill et al. 1985 and 1994; Frégea and Fourney 1993)

1.2 Short tandem repeats (STR) for forensics

A larger number of STRs are identified scattered throughout the human genome (Weber and Broman 2001). Its positions can be referred to as a locus and its variable number of tandem repeats occurring between individuals as the alleles. The presentation of these different allelic combinations for a locus or panel of loci is referred to as an STR genotype which in essence defines the DNA profile. The value of STR in forensics is recognised due to their high variability and abundance.

The human genome is known to have 150, 000 highly polymorphic STRs and due to their high degree of mutability remains the mainstay for DNA typing (Weber and Wong 1993 and Weber and Broman 2001). Its preference over VNTRs is also due to its relatively short DNA amplicon length (<500 bp) which is suitable for degraded samples and limits the issue of differential amplifications between smaller and larger amplicons. Compared to VNTRs, STR alleles are also more distinct and amenable to multi-locus PCR based methodologies (Butler 2009).

The various types of STRs identified on both the autosomes and Y-chromosome have been categorised according to their sequence structure, the repeat length and repeat number. Simple repeats define STRs with repeat units identical in length and sequence, while compound repeats may contain more than two simple repeats neighbouring each other. The complex repeat category may contain many repeat blocks for which each can vary in number of repeat units and intermediate sequence (Urquhart et al. 1994). If we consider the nucleotide composition, the type of repeat units can be expressed as di, tri, tetra, penta, and hexanucleotide repeats which corresponds respectively to 2,3,4,5 and 6 nucleoids positioned in tandem (Urquhart et al. 1994 and Jin et al. 1994).

The general selection criteria for STR for forensics application include the following: obtaining discriminating power > 80%, a heterozygosity >70%, low percentage of stutter products and be amenable to multiplexing (Urquhart et al. 1994 and 1996). The selection should however also be made relative to the intended forensic application. For DNA profiling of skeletal remains miniSTRs (Opel et al.2006) would be ideal and for analysis of sexual assault evidence with DNA mixtures STRs with lowest stutter would be important. Furthermore, to analyse male-female DNA mixtures from sexual assault evidence using STRs specific to the Y-chromosome would be preferable (Roewer et al. 2009 and Kayser et al. 2017).

The selection of autosomal STRs for forensic applications between 1994-1996 were initially contributed from various research divisions, forensic service units and by major commercial companies. It was during this period that several small STR panels were being developed into commercial kits. The first to be released was a multi-locus kits with three autosomal markers (CSF1PO, TPOX, and TH01) which was made available from Promega® in 1994 and based on the selection made by the work of Edwards et al. 1991 and Hammond et al. 1994. Second generation multiplex (SGM) kits which soon followed by Applied Biosystems (now Thermo Fisher) was based on the work by Kimpton et al. 1996 and Sparkes et al. 1996. With the expectation of an efflux in new STRs and commercial prototypes the forensic community in 1997 selected a core set of 13 STRs to standardize and establish concordance between the different developments and divisions. The core 13 STRs was in 1998 integrated into a national DNA database in the United Stated called the Combined DNA Index System (CODIS) which grew after twelve years to more than nine million DNA profiles (Budowle et al. 1998 and Butler 2006). The 13 CODIS loci also overlapped at several loci within the European standard set (ESS) as represented in the SGM kit (Thermo Fisher) and for the same loci corresponding to the Interpol Standard Set of Loci (ISSOL). This overlap is a crucial aspect of DNA typing and databasing. This is because different primer sequences can be utilized to amplify the same STR locus across different systems. Thus, potential for discordance can existed across kits due to mutations in the primer binding site or STR-flanking regions (Butler et al. 2011 and Davis et al. 2012).

This rapid evolution of STR discovery on the Y-chromosome occurred in parallel to autosomal markers. First described and valued for sexual assault application by Rower et al.1992, the forensic community recognised its benefits and also standardized a panel of nine loci in 1997 referred to as the European Minimal Haplotype (EMP) (Kayser et al. 1997 and Pascali et al. 1998).

At the turn of the new millennium ReliaGene Technologies released the first commercial kit referred as the Y-PLEX™ 6 (Shina et al. 2003). Between 2000-2003, several in-house developments were being developed (Dekairelle and Hoste et al. 2001; Prinz and Sansone et sl. 2001; Corach et al. 2001 and Sibille et al. 2002). While several new polymorphic Y-STRs were also being discovered (White et al. 1999; Ayub et al. 2000; Redd et al. 2002 and Butler et al. 2003). The Scientific Working Group on DNA Analysis Methods (SWGDAM) made recommendation for a standardize set of eleven core markers which included the nine EMP markers with the addition of DYS438 and DYS439 (Ayub et al. 2000). This marker panel formed the core of commercial kits such as the Powerplex Y (2003) and Y-filer (2004) which became widespread for forensic application and Y-STR databasing.

1.3 The Y-chromosome in forensic

1.3.1 Y-STR applications

The Y-chromosome STRs are referred to as lineage markers as they are largely male specific and passed down the paternal lineage without genetic recombination of genes between generations (Roewer 1992 and 2009). We usually expect a single allele per individual and therefore instead of making referral to a genotype, Y-chromosome DNA profiles is called a haplotype. Y-STR analysis in general proposes more of a challenge to distinguishing paternally related males and therefore present much lower probabilities of inclusion compared with autosomal STRs (Caliebe et al. 2015 and Andersen and Balding et al. 2017). However, Y-STRs are still valued under specific forensic scenarios and with more highly polymorphic markers being characterized, the forensic community remains driven to individualize the male haplotype (Ballantyne et al. 2012 and 2014).

The standard approach in processing sexual assault evidence is to separate the male sperm from female components using differential DNA extraction procedures followed by autosomal STR genotyping (Gill et al.1985). However, Y-STR profiling is of particular value when utilized in adjunct to autosomal systems under specific case scenarios (Prinz and Sansone, 2001). This includes multiple perpetrator rape (gang rape) were a mixture of male DNA contributors must be distinguished and when the female DNA component is in large excess of the male. This is often the case when the test for semen is negative or too degraded for differential extraction. Furthermore, male cells collected from bite marks and fingernail scrapings of the victim would present an admixture of DNA which may only be resolved by specifically profiling the Y-chromosome as physical separation is not possible.

Obtaining a successful autosomal profile from a vaginal swab collected more than 24-36 hours after the intercourse is also less likely (Hall and Ballantyne 2003). The longer the post-coitus time period the more likely the sperm would be lost, which makes differential extraction less plausible. In these circumstances Y-STR analysis would be preferable as it is known to provide routine full profiles 3-4 days after intercourse (Mayntz-Press et al. 2008). Furthermore, using a simplified post PCR purification method this period can be extended to up to 7 days which is the duration approaching the limit of sperm detection in the cervix (Smith and Ballantyne 2007).

The value of Y-STRs is also recognised in scenarios were information on the paternal lineage and geographical origin can be useful. The application of Y-STRs to trace bio-geographical ancestry may lead the investigation when a suspect is unknown, this was applied for the first time in the 1999 Vaatstra murder-rape case reviewed in detail by Kayser et al. 2009. Although autosomal markers are routinely applied to paternity cases, Y-STRs can be valuable in non-putative paternity whereby the father of a male child is deceased.

In these situations, Y-STRs can be utilized to profile members of the deceased father's paternal lineage (uncles or nephews). This value was observed in 1998 when it implicated that the former Unites States president Thomas Jefferson had fathered a son from his domestic worker (Foster et al. 1998). Y-STR analysis would also come to assist the mass disaster victim identification process, identification of family groups in mass graves and for identifying historical human skeletal remains (Gill et al. 1994; Davison et al. 2008 and Corach et al. 2010).

1.3.2 Y-STR multiplex developments

Provided with its widespread application, the Y-STR selected for forensic applications can be more complicated compared to autosomal systems. In addition to the general STR selection criteria, for Y-STRs the primer combinations should be carefully selected considering no cross reactivity to the female X-chromosome. Furthermore, due to its patrilineal mode of inheritance, the impact of cultural practices, geographical factors and historic events can influence how informative a given Y-STR or panel may be (Oota et al. 2001; Kayser et al. 2003; Roewer et al. 2005 and Nuñez et al. 2015). Its selection should therefore consider large regional population studies to validate its informativeness before being implemented as a forensic tool. Due to the limited population data available when the eleven SWGDAM recommended set was selected, multiplex developments have since been continuously upgraded with more informative markers to make it more informative across globally dispersed populations

Currently the most informative panels and robust commercial Y-STR systems include the Powerplex® Y-23 system (Promega) and the AmpFlSTR® Yfiler-plus™ (Thermo Fisher) which provide 23 and 27 loci respectively. The Powerplex® Y-23 (PPY23) and AmpFlSTR® Yfiler-plus™ (Yfiler-plus) are 5-dye and 6-dye fluorescent base multiplex systems that both contain the 11 core SWGDAM recommended loci (DYS19, DYS385a/b, DYS389I/II, DYS390, DYS391, DYS392, and DYS393, DYS438 and DYS439). Since the comprehensive study on Y-STRs mutation rates in 2012 (Ballantyne et al. 2012), it was identified that the majority of Y-STRs used in forensics until then had mutation rates in the order less than 1 x 10^{-3}. This study, identified thirteen Y-STRs (DYF399S1, DYF387S1, DYS570, DYS576, DYS518, DYS526a/ b, DYS626, DYS627, DYF403-S1a + b, DYF404S1, DYS449, DYS547 and DYS612) which were classified as rapidly mutating marker > 1.0 x 10^{-2} (Ballantyne et al. 2012).

The PPY23 and Yfiler-plus which were developed subsequently therefore incorporated respectively two (DYS570 and DYS576) and seven (DYF387S1a/b, DYS449, DYS518, DYS570, DYS576, DYS627) rapidly mutating Y-STRs (RM-Y-STRs) to maximize discrimination between related individuals. These upgrades made to the marker panels significantly improved the discriminatory power compared to its respective predecessor versions (Purps et al. 2012 and Pickrahn et al. 2016). Concurrently, the PCR chemistries of these multiplex systems were also continuously upgraded to vastly improve its technical performance considering the plethora of challenging types of DNA samples encountered in forensics (Thompson et al. 2013 and Gopinath et al. 2016). Validation guidelines for DNA testing laboratories therefore recommends both population and technical validation studies for a given development before widespread implementation within a jurisdiction.

1.3.3 Y-STR developmental validations

Developmental validation studies considered for commercial, non-commercial and prototype kits are usually conducted following standardized guidelines. In 2004, the SWGDAM released revised validation guidelines adapted from those previously published between 1988-1995 by the Technical Working Group on DNA Analysis Methods (TWGDAM). The proposed procedure is aimed to validate that the adopted DNA typing technique is robust, reproducible and reliable for a set of criteria defined by a broad DNA typing community.

The validation studies published by research institutes, forensic services companies or commercial companies therefore provides performance traceability which may serve as guidelines for the forensic users. Two type of validations studies are required before a novel system is implemented for forensic purposes. 1) The developmental validation studies which is conducted by the organisation or body before a novel system is presented for forensic applications and 2) Internal validation which is the validation by a given Forensic institute or service company accrediting the reliability, reproducibility and limitations of the procedure before its applied for casework proceeding.

1.4 UniQTyper™ Y-10 multiplex

The UniQTyper™ Y-10 multiplex system represents an upgrade to the previously referred to UWC-10 plex development. The system is previously validated by D’Amato et al. 2011 for forensic utility and since then being perused commercially. The system represents an accumulative effort to ascertain a highly informative marker panel considering the rich ethnic and genetic diversity in South Africa. The Primer sequences of UniQ-Typer™ Y-10 is complementary to the previous developments while a new fluorescent dye panel was been selected. The primer dyes adopted for DYS612 DYS504, DYS626, DYS481, DYS385 and DYS644 are provided in Table 1.1. The dye 6-FAM for DYS710 and DYS518 primers were kept consistent with the previous development (Table 1.1 and Figure 1). For the customized dye panel, we previously designed a matrix standard for calibration of fluorescence detection on the capillary electrophoresis Genetic Analyser ABI3500 (Cloete et al. 2016). According to classification made by Ballantyne et al. 2010 the panel is host to 4 rapidly mutating markers (DYS449, DYS 518, DYS612, DYS626) and 5 markers with mutation rate in the order of 1.2 X 10^{-3} to 5 X10^{-3} (Table 1.2). A total of 5 Loci (DYS644, DYS504, DYS447, DYS626, DYS612) of the panel is not present in commercial Y-STR kits, while the overlapped with

Yfiler-plus and with the PPY-23 systems is indicated in Table 1.2. The UniQTyper™ Y-10 multiplex system presented in herein was optimized using with a high fidelity Taq polymerase in order to reduce the amplification time. Compared to the previous development the chemistry has also been reformulated as to provide resistance to various potential PCR inhibitors commonly associated with forensic DNA evidence.

Table 1.1: Locus dye label configuration, allele size range and repeat structure.

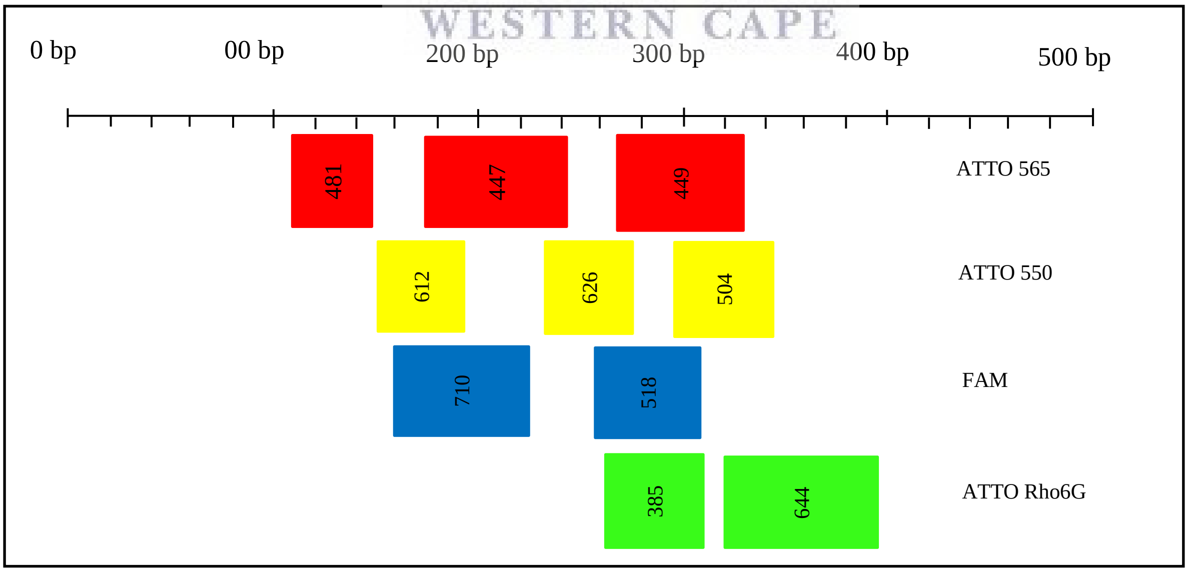
Locus	Dye	Allele size range (bp)	Repeat structure
DYS710	FAM	160-224	[**AAAG**]$_n$ [**AG**]$_n$ [**AAAG**]$_n$
DYS518	FAM	254-301	[AAAG]$_3$[GAAG]$_1$[**AAAG**]$_n$[GGAG]$_1$[AAAG]$_4$N$_6$[**AAAG]**$_n$N$_{27}$[AAGG]$_4$
DYS612	ATTO 550	154-194	[CCT]$_5$[CTT]$_1$[TCT]$_4$[CCT]$_1$[**TCT**]$_n$
DYS385	ATTO Rho6G	243-300	[**GAAA**]$_n$
DYS644	ATTO Rho6G	308-394	[**TTTTA]**$_n$ TTTA [**TTTTA**]$_n$
DYS504	ATTO 550	294-340	[**TCCT**]$_n$N$_7$[CCCT]$_3$
DYS626	ATTO 550	232-276	[**GAAA**]$_n$N$_{24}$[GAAA]$_3$N$_6$[GAAA]$_5$ [AAA]$_1$[**GAAA**]$_{2–3}$ [GAAG]$_1$(GAAA)$_3$
DYS481	ATTO 565	109-141	[**CTT**]$_n$
DYS447	ATTO 565	174-243	[**TAATA**]$_n$[TAAAA]$_1$[**TAATA**]$_n$[TAAAA]$_1$[**TAATA**]$_n$
DYS449	ATTO 565	271-328	[**TTTC**]$_n$N$_{50}$[**TTTC**]$_n$

Table 1.2: UniQTyper™ Y-10 loci mutation rates and overlap to similar commercial systems. Mutation rate are adapted from Ballantyne et al.2010. Powerplex® Y-23 (PPY23) and AmpFlSTR® Yfiler-plus™ (Yfiler-Plus). NA (Not Available)

UniQTyper™ Y-10 Loci	Mutation Rates	PPY23	Yfiler-Plus
DYS710	NA		
DYS518	1.84 x10^{-2}		x
DYS385	2.08 x10^{-3}	x	x
DYS644	3.22 x10^{-3}		
DYS504	3.24 x10^{-3}		
DYS612	1.45 x10^{-2}		
DYS626	1.22 x10^{-2}		
DYS481	4.97 x10^{-3}	x	x
DYS447	2.12 x10^{-3}		
DYS449	1.22 x10^{-2}		x

The selection of Y-STRs for the kit is represented by several years of research aimed to ascertain the most informative marker panel. This initial attempt by Leat et al.2007 described for the first time the value of locus DYS710 from a comprehensive survey of 27 Y-STRs in South African populations. A continued effort to screen various Y-STRs in D'Amato et al. 2009 identifies markers which provided the highest gene diversities and discriminatory capabilities. This study also identified markers which may be informative to make genealogical inference. In subsequent research attempts a set of Y-STRs were selected base on their predetermined forensic value and amenability to a multiplex (D'Amato et al. 2010 and 2011). This research which validated a carefully ascertained set of 10 Y-STR showed a significantly improved in discrimination of individuals in native populations groups. while other commercial kits show 7-10% of Bantu men sharing the same profile, with discrimination capacity (DC) ~ 0.8, the UWC 10-plex kit showed DC higher than 92% (D'Amato et al 2011).

Figure 1: UniQTyper™ Y-10 primer dye set configuration and size range spectrum



1.5 Specific research aims and objectives:

A primary goal of this PhD research is to produce a Y-STR commercial forensic kit intended for processing sexual assault DNA evidence. This prototype being developed is designed on male-specific STRs of the Y-Chromosome to provides two main functions: (1) The resolution of sexual offenses utilizing DNA evidence originating under various scenarios of assault and in (2) kinship analysis (paternal lineage).

The original development of the UWC-10plex system is herein optimized with a novel chemistry for delivery of a competitive commercial prototype presenting a desirable level of forensic performance. This PhD contributed to the design and construction of integral calibration components constituting the DNA profiling prototype. Furthermore, the aim was to study systems capacity to discriminate between males across diverse populations in South Africa, for which the performances are characterized on both a DNA profiling and sequencing level.

The aims are divided into the following objectives:

1) Optimize and validate a Y-STR multiplex chemistry with intent to provide rapid turnover time to result and cost-effective DNA profiling using a direct amplification approach.
2) Optimize a workflow to produce a bulk balanced allelic ladder as the calibration standard for the prototype.
3) Conduct a large Y-STR survey across the nine South Africa provinces to validate the informativeness of the UniQTyper™ Y-10 panel from a DNA profiling and allele sequencing perspective.

Altogether, the application for such a powerful analytical tool is intended to improve current conviction rates on sexual assault, exonerate innocent men accused of rape and alleviate the backlog of casework in South Africa and neighbouring countries.

1.6 References

Andersen, M. M., & Balding, D. J. (2017). How convincing is a matching Y-chromosome profile? *PLoS Genetics*, *13*(11), e1007028–e1007028. https://doi.org/10.1371/journal.pgen.1007028

Ayadi, I., et al. (2007). Combining autosomal and Y-chromosomal short tandem repeat data in paternity testing with male child: Methods and application. Journal of Forensic Sciences, 52, 1068–1072.

Ayub, Q., Mohyuddin, A., Qamar, R., Mazhar, K., Zerjal, T., Mehdi, S. Q., & Tyler-Smith, C. (2000). Identification and characterisation of novel human Y-chromosomal microsatellites from sequence database information. *Nucleic Acids Research*, *28*(2), e8–e8. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/10606676

Ballantyne, K. N., Keerl, V., Wollstein, A., Choi, Y., Zuniga, S. B., Ralf, A., 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), 208–218. https://doi.org/10.1016/j.fsigen.2011.04.017

Ballantyne, K. N., Ralf, A., Aboukhalid, R., Achakzai, N. M., Anjos, M. J., Ayub, Q., Kayser, M. (2014). Toward Male Individualization with Rapidly Mutating Y-Chromosomal Short Tandem Repeats. *Human Mutation*, *35*(8), 1021–1032. https://doi.org/10.1002/humu.22599

Budowle, B., Moretti, T. R., Niezgoda, S. J., & Brown, B. L. (1998). CODIS and PCR-based short tandem repeat loci: law enforcement tools. Proceedings of the Second European Symposium on Human Identification, 73–88.

Butler, J. M. (2003). Recent developments in Y-short tandem repeat and Y-single nucleotide polymorphism analysis. Forensic Science Review, 15, 91–111.

Butler, J. M. (2006). Genetics and Genomics of Core Short Tandem Repeat Loci Used in Human Identity Testing. *Journal of Forensic Sciences*, *51*(2), 253–265. https://doi.org/10.1111/j.1556-4029.2006.00046.x

Butler, J. M. (2009). Fundamentals of Forensic DNA Typing. Elsevier Science.

Butler, J. M., Becky Hill, C. R., Kline, M. C., Bastisch, I., Weirich, V., McLaren, R. S., & Storts, D. R. (2011). SE33 variant alleles: Sequences and implications. *Forensic Science International: Genetics Supplement Series*, *3*(1), e502–e503. https://doi.org/10.1016/j.fsigss.2011.10.002

Caliebe, A., Jochens, A., Willuweit, S., Roewer, L., & Krawczak, M. (2015). No shortcut solution to the problem of Y-STR match probability calculation. *Forensic Science International: Genetics*, *15*, 69–75. https://doi.org/10.1016/j.fsigen.2014.10.016

Cloete, K. W., Ristow, P. G., Kasu, M., & D'Amato, M. E. (2017). Design, installation, and performance evaluation of a custom dye matrix standard for automated capillary electrophoresis. *Electrophoresis*, *38*(5). https://doi.org/10.1002/elps.201600257

Corach, D. (2010). Mass Disaster Victim Identification Assisted by DNA Typing. In *Molecular Diagnosis* (pp. 407–415). https://doi.org/10.1016/B978-0-12-374537-8.00027-4

Corach, D., Filgueira Risso, L., Marino, M., Penacino, G., & Sala, A. (2001). Routine Y-STR typing in forensic casework. *Forensic Science International*, *118*(2), 131–135. https://doi.org/https://doi.org/10.1016/S0379-0738(00)00483-7

D'Amato ME, Benjeddou M, Davison S (2009) Evaluation of 21 Y-STRs for population and forensic studies. Forensic Sci Int Genet Suppl Ser 2:446–447. https://doi.org/10.1016/j.fsigss.2009.08.091

D'Amato ME, Bajic VB, Davison S (2011) Design and validation of a highly discriminatory 10-locus Y-chromosome STR multiplex system. Forensic Sci Int Genet 5:122–125. https://doi.org/10.1016/j.fsigen.2010.08.015

Davis, C., Ge, J., King, J., Malik, N., Weirich, V., Eisenberg, A. J., & Budowle, B. (2012). Variants observed for STR locus SE33: A concordance study. *Forensic Science International: Genetics*, *6*(4), 494–497. https://doi.org/10.1016/j.fsigen.2011.12.002

Davison, S., Benjeddou, M., & Amato, M. E. D. (2008). Molecular genetic identification of skeletal remains of apartheid activists in South Africa. *African Journal of Biotechnology*, *7*(25), 4750–4757.

Dekairelle, A. F., & Hoste, B. (2001). Application of a Y-STR-pentaplex PCR (DYS19, DYS389I and II, DYS390 and DYS393) to sexual assault cases. Forensic Science International, 118, 122–12

Edwards, A., Civitello, A., Hammond, H. A., & Caskey, C. T. (1991). DNA typing and genetic mapping with trimeric and tetrameric tandem repeats. *American Journal of Human Genetics*, *49*(4), 746–756. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/1897522

Foster, E. A., Jobling, M. A., Taylor, P. G., Donnelly, P., de Knijff, P., Mieremet, R., Tyler-Smith, C. (1998). Jefferson fathered slave's last child. *Nature*, *396*, 27. Retrieved from https://doi.org/10.1038/23835

Frégeau, C. J., & Fourney, R. M. (1993). DNA typing with fluorescently tagged short tandem repeats: a sensitive and accurate approach to human identification. *BioTechniques*, *15*(1), 100—119. Retrieved from http://europepmc.org/abstract/MED/8103347

Gill, P., Ivanov, P. L., Kimpton, C., Piercy, R., Benson, N., Tully, G., Sullivan, K. (1994). Identification of the remains of the Romanov family by DNA analysis. *Nature Genetics*, *6*, 130. Retrieved from https://doi.org/10.1038/ng0294-130

Gill, P., Jeffreys, A. J., & Werrett, D. J. (1985). Forensic application of DNA "fingerprints". *Nature*, *318*(6046), 577–579.

Gopinath, S., Zhong, C., Nguyen, V., Ge, J., Lagacé, R. E., Short, M. L., & Mulero, J. J. (2016). Forensic Science International : Genetics Developmental validation of the Yfiler®Plus PCR Amplification Kit : An enhanced Y-STR multiplex for casework and database applications, *24*, 164–175. https://doi.org/10.1016/j.fsigen.2016.07.006

Groce, N. E., & Trasi, R. (2004). Rape of individuals with disability: AIDS and the folk belief of virgin cleansing. *The Lancet*, *363*(9422), 1663–1664. https://doi.org/10.1016/S0140-6736(04)16288-0

Hall, A., & 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*(1), 58–72. https://doi.org/https://doi.org/10.1016/S0379-0738(03)00258-5

Hammond, H. A., Jin, L., Zhong, Y., Caskey, C. T., & Chakraborty, R. (1994). Evaluation of 13 short tandem repeat loci for use in personal identification applications. *Am J Hum Genet*, *55*(1), 175–189. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1918216/

Jeffreys, A. J., Wilson, V., & Thein, S. L. (1985). Hypervariable 'minisatellite' regions in human DNA. *Nature*, *314*, 67. Retrieved from https://doi.org/10.1038/314067a0

Jewkes, R. & Sikweyiya, Yandisa, Robert Morrell, K.D., 2009. Understanding men's health and use of violence: Interface of rape and HIV in South Africa, Available at: http://www.mrc.ac.za/gender/interfaceofrape&hivsarpt.pdf.

Jin, L., Zhong, Y., & Chakraborty, R. (1994). The exact numbers of possible microsatellite motifs. *American Journal of Human Genetics*, *55*(3), 582–583. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/8079998

Jobling, M. A., & Gill, P. (2004). Encoded evidence: DNA in forensic analysis. *Nature Reviews Genetics*, *5*(10), 739–751. https://doi.org/10.1038/nrg1455

Kayser, M., Brauer, S., Weiss, G., Schiefenhövel, W., Underhill, P., Shen, P., Stoneking, M. (2003). Reduced Y-chromosome, but not mitochondrial DNA, diversity in human populations from West New Guinea. *American Journal of Human Genetics*, *72*(2), 281–302. https://doi.org/10.1086/346065

Kayser, M., Caglia, A., Corach, D., Fretwell, N., Gehrig, C., Graziosi, G., Roewer, L. (1997). Evaluation of Y-chromosomal STRs: a multicenter study. *International Journal of Legal Medicine*, *110*(3), 125-133,141-149.

Kimpton, C. P., Oldroyd, N. J., Watson, S. K., Frazier, R. R. E., Johnson, P. E., Millican, E. S., Gill, P. (1996). Validation of highly discriminating multiplex short tandem repeat amplification systems for individual identification. *ELECTROPHORESIS*, *17*(8), 1283–1293. https://doi.org/10.1002/elps.1150170802

Leat, N., Ehrenreich, L., Benjeddou, M., Cloete, K., & Davison, S. (2007). Properties of novel and widely studied Y-STR loci in three South African populations. *Forensic Science International*, *168*(2–3), 154–161. https://doi.org/10.1016/j.forsciint.2006.07.009

Mayntz-Press, KA., Sims, L., Hall, A., & Ballantyne, J. (2008). Y-STR Profiling in Extended Interval (≥3 days) Postcoital Cervicovaginal Samples. *Journal of Forensic Sciences*, *53*, 342–348. https://doi.org/10.1111/j.1556-4029.2008.00672.x

Meel, B. L. (2003). The Myth of Child Rape as a Cure for HIV/AIDS in Transkei. *Medicine, Science and the Law*, *43*(1), 85–88. https://doi.org/10.1258/rsmmsl.43.1.85

Mullis, K. B., & Faloona, F. A. (1987). Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. *Recombinant DNA Part F*, *155*, 335–350. https://doi.org/10.1016/0076-6879(87)55023-6

Nuñez, C., Baeta, M., Fernández, M., Zarrabeitia, M., Martinez-Jarreta, B., & De Pancorbo, M. M. (2015). Highly discriminatory capacity of the PowerPlex® Y23 System for the study of isolated populations. *Forensic Science International: Genetics*, *17*, 104–107. https://doi.org/10.1016/j.fsigen.2015.04.005

Oota, H., Settheetham-Ishida, W., Tiwawech, D., Ishida, T., & Stoneking, M. (2001). Human mtDNA and Y-chromosome variation is correlated with matrilocal versus patrilocal residence. *Nature Genetics*, *29*, 20. Retrieved from https://doi.org/10.1038/ng711

Opel, K. L., Chung, D. T., Drábek, J., Tatarek, N. E., Jantz, L. M., & McCord, B. R. (2006). The Application of Miniplex Primer Sets in the Analysis of Degraded DNA from Human Skeletal Remains. *Journal of Forensic Sciences*, *51*(2), 351–356. https://doi.org/10.1111/j.1556-4029.2006.00077.x

P.S. White, O.L. Tatum, L.L. Deaven, J.L. Longmire, New, male-specific microsatellite markers from the human Y chromosome, Genomics 57 (1999) 433–437.

Panneerchelvam, S., & Norazmi, M. N. (2003). Forensic DNA profiling and database. *Malaysian Journal of Medical Sciences*, *10*(2), 20–26.

Pascali VL, Dobosz M, Brinkmann B. Coordinating Y-chromosomal STR research for the Courts. *International Journal of Legal Medicine*. 1998;112(1)

Pickrahn, I., Müller, E., Zahrer, W., Dunkelmann, B., Cemper-Kiesslich, J., Kreindl, G., & Neuhuber, F. (2016). Yfiler®Plus amplification kit validation and calculation of forensic parameters for two Austrian populations. *Forensic Science International: Genetics*, *21*, 90–94. https://doi.org/10.1016/j.fsigen.2015.12.014

Prinz, M., & Sansone, M. (2001). Y chromosome-specific short tandem repeats in forensic casework. Croatian Medical Journal, 42, 288–29

Redd, A. J., Agellon, A. B., Kearney, V. A., Contreras, V. A., Karafet, T., Park, H., … Hammer, M. F. (2002). Forensic value of 14 novel STRs on the human Y chromosome, *130*, 97–111.

Roewer, L. (2009). Y chromosome STR typing in crime casework. *Forensic Science, Medicine, and Pathology*, *5*(2), 77–84. https://doi.org/10.1007/s12024-009-9089-5

Roewer, L., Amemann, J., Spurr, N. K., Grzeschik, K.-H., & Epplen, J. T. (1992). Simple repeat sequences on the human Y chromosome are equally polymorphic as their autosomal counterparts. *Human Genetics*, *89*(4), 389–394. https://doi.org/10.1007/BF00194309

Roewer, L., Croucher, P. J. P., Willuweit, S., Lu, T. T., Kayser, M., Lessig, R., Krawczak, M. (2005). Signature of recent historical events in the European Y-chromosomal STR haplotype distribution. *Human Genetics*, *116*(4), 279–291. https://doi.org/10.1007/s00439-004-1201-z

Sibille, I., Duverneuil, C., de la Grandmaison, G. L., Guerrouache, K., Teissière, F., Durigon, M., & de Mazancourt, P. (2002). Y-STR DNA amplification as biological evidence in sexually assaulted female victims with no cytological detection of spermatozoa. *Forensic Science International*, *125*(2), 212–216. https://doi.org/https://doi.org/10.1016/S0379-0738(01)00650-8

Sinha, S. K., Budowle, B., Arcot, S. S., Richey, S. L., Chakraborty, R., Jones, M. D., … Shewale, J. G. (2003). Development and validation of a multiplexed Y-chromosome STR genotyping system, Y-PLEX$^{\text{TM}}$6, for forensic casework. *Journal of Forensic Sciences*, *48*(1), 93–103.

Smith, P. J., & Ballantyne, J. (2007). Simplified Low-Copy-Number DNA Analysis by Post-PCR Purification. *Journal of Forensic Sciences*, *52*(4), 820–829. https://doi.org/10.1111/j.1556-4029.2007.00470.x

Sparkes, R., Kimpton, C., Gilbard, S., Carne, P., Andersen, J., Oldroyd, N., Gill, P. (1996). The validation of a 7-locus multiplex STR test for use in forensic casework. (II), Artefacts, casework studies and success rates. *International Journal of Legal Medicine*, *109*(4), 195–204.

Tamaki, K., & Jeffreys, A. J. (2005). Human tandem repeat sequences in forensic DNA typing. *Legal Medicine*, *7*(4), 244–250. https://doi.org/10.1016/j.legalmed.2005.02.002

Urquhart, A. J., Oldroyd, N. J., Downes, T., Barber, M., Alliston-Greiner, R., Kimpton, C. P., & Gill, P. D. (1996). Selection of STR loci for forensic identification systems BT - 16th Congress of the International Society for Forensic Haemogenetics (Internationale Gesellschaft für forensische Hämogenetik e.V.), Santiago de Compostela, 12–16 September 1995. In A. Carracedo, B. Brinkmann, & W. Bär (Eds.) (pp. 115–117). Berlin, Heidelberg: Springer Berlin Heidelberg.

Urquhart, A., Kimpton, C. P., Downes, T. J., & Gill, P. (1994). Variation in Short Tandem Repeat sequences —a survey of twelve microsatellite loci for use as forensic identification markers. International Journal of Legal Medicine, 107(1), 13–20. https://doi.org/10.1007/BF01247268

Urquhart, A., Kimpton, C. P., Downes, T. J., & Gill, P. (1994). Variation in Short Tandem Repeat sequences —a survey of twelve microsatellite loci for use as forensic identification markers. *International Journal of Legal Medicine*, *107*(1), 13–20. https://doi.org/10.1007/BF01247268

Walsh, B., et al. (2008). Joint match probabilities for Y chromosomal and autosomal markers. Forensic Science International, 174, 234–238.

Wambaugh, J. (1989). *The Blooding*. Bantam Books. Retrieved from https://books.google.co.za/books?id=b8FrBNeslmsC

Weber, J. L., & Broman, K. W. (2001). Genotyping for human whole-genome scans: past, present, and future. *Advances in Genetics*, *42*, 77–96.

Weber, J. L., & Wong, C. (1993). Mutation of human short tandem repeats. *Human Molecular Genetics*, *2*(8), 1123–1128.

White, P. S., Tatum, O. L., Deaven, L. L., & Longmire, J. L. (1999). New, Male-Specific Microsatellite Markers from the Human Y Chromosome. *Genomics*, *57*(3), 433–437. https://doi.org/https://doi.org/10.1006/geno.1999.5782

Chapter 2

2.0 Validation of a rapid Y-chromosome multiplex for forensic application.

2.1 Introduction:

The male specific nature of Y-STRs makes it beneficial for sexual assault analysis when the female and male mixed evidence cannot be distinguished using autosomal STR genotyping (Roewer et al. 2009). The probative value of autosomal genotyping in sexual assault analysis is subject to the physical abilities to separate the female and male admixed evidence. This is achieved routinely using differential DNA extraction procedures to enrich the sperm cells from the excess female biological components (Gill et al. 1985). However, differential extraction is less plausible when virginal, anal or oral swabs of a female victim present a negative test for semen or the sperm is too degraded or minute. Likewise, differential extraction cannot physically separate admixed biological material in samples from multiple sperm contributors, bite marks, kissing and fingernail scrapings often collected from a victim. Autosomal genotyping systems are therefore routinely complemented by Y-STR analysis as it offers a higher degree of male sensitivity in excess of female DNA and also simplifies the resolution of mixed male haplotypes (Jobling and Gill 2004).

Developing rapid PCR thermal cycling protocols my significantly reduce the overall turnover times to results and therefore improve the throughput of a forensic workflow (Vallone et al. 2008; Butts and Vallone 2014 and Romos and Vallone 2015). A significant reduction in PCR time have been achieved since the development of the early forensic STR commercial kits which had a duration of ~ 2-3 h. This reduction in time is mainly attributed by the development of Taq polymerases with much faster processivity and the subsequent replacement of 3-step thermal cycling with rapid 2-step parameters (Butts and Vallone 2014 and Romos and Vallone 2015). The autosomal STR typing systems such as the Powerplex 18D (Promega) and Global

Filer Express (Thermo Fisher) which adopts 2-step thermal cycling currently provides a PCR duration of 90 min and 40 min respectively (Oostdik et al. 2013 and Hennessy et al. 2014). For Y-chromosome STR typing the fastest durations are currently offered by the 3-step protocol of Powerplex Y23 (~ 90min) and the 2-step protocol of Yfiler® Plus which deliver robust typing in less than 95 min (Thompson et al. 2013 and Gopinath et al. 2016).

Validation studies exploring higher fidelity Taq polymerases have identified that 2-step cycling in additions to its rapid duration may also offer improved performance compared to 3-step cycling with regard to the sensitivity, allele peak heights and the degree of artefacts (Butts and Vallone 2014 and Romos and Vallone 2015). Characterizing the forensic performance of PCR protocols shorter than 1 hour in this regard are relatively well documented for autosomal systems (Vallone et al. 2008; Butts and Vallone 2014 and Ramos and Vallone 2015), while similar validations are certainly limited in Y-STR developments and validations.

This chapter introduces the validation of a rapid Y-STR typing system defied throughout as the beta version of the UniQTyper™ Y-10 commercial prototype. Its multiplex chemistry which was optimized for a 45 min 2-step thermal cycling protocol is herein being validate for its forensic applicability using recommendations made by the SWGDAM. It presents the potential of utilizing one of the fastest Y-STR typing systems to specifically target the male DNA components of sexual assault evidence presenting: low copy number template, a high degree of admixture and traces of common inhibitors. The developmental validation herein presents a significant improvement to the overall performance compared to the previous prototype (UWC 10-plex). The aspects of PCR inhibitor tolerance and the direct application capabilities are herein validated for the first time. For the sensitivity and mixture studies the validation within the thesis also explores larger range for DNA concentration and mixture proportion.

Ultimately this chapter aims to present the technical performances and limitation of the UniQTyper™ Y-10 (beta version) regarding its genotyping capabilities and its potential for reference genotyping using a direct amplification approach.

2.2 Material and methods

2.2.1 Sample collection and preparation

The male DNA samples utilized for this study were obtained from four males by informed consent and ethical approval form the University of the Western Cape (10/3/39) and (15/4/97). A 5 ml saliva sample was collected from each donor using an in-house preservation buffer formula (Burrows et al. 2017). Collection of crude saliva (2 ml) without preservation was collected for direct amplification testing and for transfer onto FTA cards. A buccal swab taken from participants were used to transfer buccal cells immediately onto FTA cards. A 5ml blood sample was collected in blue cap VD Vacutainer® tubes and spotted directly into FTA cards. Saliva and blood samples were all stored at -20°C and FTA cards stored at room temperature in a moisture free environment. Commercial control DNA utilized include DNA 007 (Thermo Fisher), 2800M (Promega) and the National Institute of Standards and Technology (NIST) components A-C.

Blood stained blue jeans, green leather and a humus rich soil was prepared using a 50 µl blood sample taken form the VD Vacutainer tubes. The jeans and leather were left to dry for 24 h at room temperature and the blood stained (Shahzad et al. 2009 and Kasu and Shires 2015) soil were dried for 12 h, 24 h and 36 h in an outdoors protected area with an average day time high temperature of 29-32°C.

2.2.1 DNA extraction

2.2.1.1 Saliva DNA extraction

Saliva samples collected in the preservation buffer were extracted using the salting out procedure adapted form Miller et al. 1988 and Medrano et al. 1990. To achieve cell lysis a 500 µl aliquot of the saliva-storage buffer mixture was incubated at 56℃ overnight with 6 µl of 20 mg/ml Proteinase K. Lysates were centrifuged briefly at 15 000 rpm for 1 min and the salting out of proteins achieved by adding 150 µl of 4 M NaCl, incubation on ice for 1 h and centrifugation at 15 000 rpm for 30 min. The supernatant was carefully transferred to a new 2 ml eppendorf and the protein pellet discarded. DNA precipitation was achieved in two volumes of 100% Ethanol with incubation at -20°C for 2 h. The samples were centrifuged at 15 000 rpm for 30 min and the DNA pellets washed twice in 70% Ethanol with centrifugation at 15 000 rpm for 10 min between washes. The pellets were dried at room temperature and DNA resuspended in a low 1X TE buffer (Tris-EDTA; 10 mM Tris base, 0.1 mM EDTA) (Sigma Aldrich).

2.2.1.1 Blood DNA extraction

DNA extractions from blood stained jean and leather were achieved using a Phenol Chloroform Isoamyl protocol adapted from Comey et al. 1994. A sample area (2 X 2 cm) of the blood-stained jeans and leather was excised using a sterile surgical blade and transferred to a 2 ml eppendorf. Samples were transferred to 500µl lysis buffer containing (10 mM Tris, 10 mM EDTA, pH 8.0, 100 mM NaCL, 2% SDS, and 0.5 mg/ mL Proteinase K) and incubated at 56℃ overnight. The denim and leather material were removed, and the lysate mixed with saturated phenol: chloroform: isoamyl (25:24:1). Samples were vortexed briefly for 10s and the phases separated by centrifugation at 15 000 rpm for 10 min. The supernatant was carefully transferred to a clean 1.5 ml eppendorf without disturbing the interphase layer containing proteins. Residual traces of phenol were removed by mixing an equal volume of chlorophorm with the supernatant and centrifugation at 15 000 rpm for 10 min. The supernatant was transferred to a clean 2 ml eppendorf and DNA precipitated as described above.

2.2.3 DNA amplification

DNA samples were amplified using 10.5 µl UniQTyper™ Y-10 master mix containing (5 X reaction buffer, Taq polymerase, 200 uM dNTPs mix and additives), 2.5 µl UniQTyper™ Y-10 primer mix (10X) and adjusted accordingly with nuclease free water (Thermo Fisher) to give a 25 µl PCR reaction volume. The 2-step thermal cycling protocol contained an initial denaturation of 98°C for 30s, 30 cycles of denaturation at 98°C for 5s, 65°C annealing for 50s and a final extension at 72°C for 2 min and a final 4°C hold. For 3-step cycling all conditions were kept the same with inclusion of an extension at 72°C for 25s repeated for 30 cycles. Thermal conditions were programmed using the Arktik (Thermo Fisher), ABI 2720 (Thermo Fisher) and Veriti (Thermo Fisher) for performance studies across thermal cyclers and for the core of the validation using the Arktik (Thermo Fisher).

2.2.4 PCR based procedures

Validation testing of a genotyping kit also considers PCR based procedures that may result in variations known to influence optimal performances. These procedures that are most influential may include pipetting errors of the primer mix or Taq master mix and the temperature variation that may exist between different thermal cyclers. These tests were all conducted for 2-step thermal cycling for 30 cycles at 0.5 ng DNA of 2 male positive controls amplified in duplicate.

2.2.5 Annealing temperature gradient

The optimal annealing temperature was determined in silicon using Primer3 online-software (Untergasse et al. 2012) and a gradient PCR in the range from 59-68°C. Validations were performed at the optimal temperature of 65°C for all tests. Suboptimal annealing temperatures were evaluated at 61°C and 63°C annealing by gradient PCR on the Veriti thermal cycler.

2.2.6 Primer mix titration

Primer concentrations were optimized using recommendations made by D’Amato et al. 2011. Adjustment to the concentration of each primer set was made to determine the optimal working proportions. The optimal conditions were determined by evaluating peak height intensity and balance at 0.5 ng DNA and low copy number template. Optimal specifications were used formulate a 10 X bulk primer mix which was predetermined optimal at final concentration of 1X in a 25 µl PCR reaction. The bulk primer mix was titrated to provide final concentration of 0.7, 0.8, 0.9, 1.0 and 1.5 X in 25 µl PCR.

2.2.7 Taq polymerase titration

The concentration of Taq polymerase for the UniQTyper™ PCR master mix was tested at various concentrations by titration at 1.2, 1.0, 0.8, 0.7 and 0.6 Units (U) Taq in a 25 µl PCR reaction.

2.2.8 Rapid thermal cycling: 2-step vs 3-step

The validation of a rapid 45 min 2-step thermal cycling protocol was established in direct comparison to a 60 min 3-step cycling protocol using all three thermal cyclers. Tests were conducted using 1 ng DNA from three donors amplified in duplicate for each protocol across all thermal cyclers and repeated at low copy number template amplification using 125 pg and 62.5 pg DNA. The Y-STR profiles were compared between thermal protocols in respect to allele peak heights (measured in relative fluorescent units or RFUs), peak height balance, stutter production and the occurrence of artefacts.

2.2.9 Multiplex validation and performance

2.2.9.1 Sensitivity study

All multiplex performance studies referred to hereafter were conducted using the 2-step thermal cycling protocol programmed on the Arkik thermal cycler. Sensitivity was conducted with titration of 3 male DNA samples between (500 pg, 250 pg, 125 pg, 62.5 pg and 31.2 pg) for amplification in duplicate at 28, 30 and 32 cycles.

2.2.9.2 Mixture study

Male/male DNA admixtures were prepared for two males with unshared alleles and with alleles that were not overlapped with respective stutter products. The admixed DNA was amplified at ratios (1:4; 1:9 and 1:19) in four replicates using a total of 2 ng DNA. Female/male mixtures were conducted for a single mixture amplified in duplicate at 1:100, 1:200, 1:400, 1:800, 1:2000 with the female DNA kept constant at 200 ng and decreasing concentration of the male component.

2.2.9.3 Stability study

The two common PCR inhibitors tested at increasing concentrations were humic acid (Sigma Aldrich) and hematin (Sigma Aldrich). Male DNA from tree donors kept constant at 1ng were amplified in duplicate at humic acid concentrations (0, 5, 10, 25, 50, 100 ng /µl) and for hematin at (0, 5, 10, 50, 100 µM). Blood stained denim jeans and green leather as a source of Indigo and tannic acid PCR inhibitors respectively were prepared using 50 µl blood. Samples were allowed to dry for 24 hrs and extracted for two donors' samples and the DNA amplified in duplicate using 1 ng/µl DNA from each donor. Humus rich blood-stained soil samples were prepared for two donors and exposed for a maximum of 6 h to high humidity and direct sunlight for every 12 h outdoor exposure period. Samples were collected at a 12 h, 24 h and 36 h outdoor exposure period, the DNA extracted as indicated above and genotyped in duplicate using 1 ng/µl DNA.

2.2.9.4 Direct amplification studies

Direct amplifications were validated using blood and saliva sample for two donors processed in three different formats. A 50 µl volume of blood was transferred onto FTA® cards (Whatman) and allowed to dry at toom temperature. Buccal cells were transferred onto the FTA® cards using a cotton swab. For blood and saliva FTA tests a single 1.2 mm disc was sampled with the Harris Uni-Core punch (Whatman), washed briefly in nuclease free water and transferred directly into a 25 µl PCR reaction. Direct amplification was also tested with addition of 2 µl crude saliva or blood added directly in to the PCR reaction. All samples types blood FTA, buccal FTA, crude saliva and crude blood were amplified in triplicates for each donor using the 30 PCR cycles.

2.2.9.5 Species specificity

Primer interactions with non-human species were characterized for 10 ng DNA of male and female Chimpanzee *(Pan troglodytes)*, Gorilla *(gorilla gorilla)*, Vervet monkey *(Chlorocebus pygerythrus)*, cat *(Felis felis)* and dog *(Canis familiaris) species*. DNA from primate were previously donated by Jacqui Bishop (University of Cape Town) and Desiree Dalton (National Zoological Gardens of South Africa).

2.2.10 DNA quantification

Extracted DNA stock and dilutions were quantified using the Qubit dsDNA high sensitivity assay and measurements obtained on a Qubit® 2.0 Fluorometer. Genomic DNA extracted from blood, saliva and blood stained materials were measured by Nanodrop™ Thermo (Fisher) for an assessment of DNA purity.

2.2.11 Capillary electrophoresis and data analysis.

Amplified PCR products were electrophoresed on an ABI3500 (Thermo Fisher) Genetic Analyzer with spectral calibration achieved using an in house developed 5-dye custom matrix (Cloete et al.2016). Samples were prepared using 9.7 µl HiDi™ formamide (Thermo Fisher) and 0.3 µl GeneScan™ LIZ500® (Thermo Fisher), 1 µl PCR product and allelic ladder. Samples were denatured at 95°C for 5 min and snap cooled on ice for 3 min. Samples were injected in an 8-capillary 36 cm array for 15s at 1.2kV and fragment separation achieved using a POP4 polymer for 20 min at 15kV at a run temperature of 60°C. Data was analysed using GeneMapper® ID-X software v1.4 using a 50 RFU cut-off threshold. Exported data was analysed for graphical presentations using statistical packages R-studio (RStudio Team 2015) and STR validator (Hansson et al. 2014).

2.3 Results and discussion

2.3.1 Rapid thermal cycling validation.

A multiplex optimized to adopt an annealing temperature of 65°C allowed the annealing and extension phase to be merged to yield 2-step thermal cycling. The performance of 2-step cycling was compared to 3-step cycling which included an extension at 75°C for 25s during each cycle. The 2-step protocol provided higher average allele peak heights at all loci with the most notable increase at locus DYS385ab as displayed in (Table 1.3 and Figure 2.1). In the box plot of Figure 2.1 with the median given as the 50 percentiles, the allele peak balance was generally more consistent for 2-step cycling. The Y-STR haplotypes were devoid of spilt peaks and non-specific artefacts in the detection range for both 2- step (45 min) and 3-step (60 min) PCR protocols (Figure 2.2). The occurrence of stutter was similar between the protocols (Table 1.3), with the majority of loci only presenting stutter one repeat unit shorter (n-1) than the true allele. The exception was DYS710 and DYS612 which also presented n-2 and n+1 stutter product respectively (Table 1.3). Stutter was absent at DYS644 and below 5% for DYS626, DYS504 and DYS447 for both protocols. Although loci DYS449, DYS481, DYS518 and DYS385ab presented up to 2-3 fold higher stutter percentages (Table 1.3), these values were lower to those reported for these respective loci of the Powerplex® Y-23 and Yfiler® plus systems (Thompson et al. 2013 and Gopinath et al. 2016).

To validate the robustness of the 45 min 2-step PCR protocol the DNA from 3 donors were amplified in duplicate across three thermal cyclers (Figure 2.3 and Supplementary 1 Figure 1.1). Full Y-STR haplotypes were obtained with an allele peak height ranging on average from 2322.4 to 11258.5 RFUs between all thermal cyclers. The average allele peak height across all loci for the Arktik (Thermo Fisher), Veriti (Thermo Fisher) and 2720 (Thermo Fisher) was (7810.2 RFUs; *sd*± 3208.6); (4876 RFUs; *sd*± 2470); (5233 RFUs; *sd* ±2193) respectively.

In the box plot of Figure 2.3 the average RFUs annotated respectively per thermal cycler indicates superior performance by the Arktik cycler, while the Veriti and ABI2720 gave similar performances.

To evaluate if the variable performance between thermal cyclers could have a negative effect on abilities to produce full profiles at low copy number template the following tests were conducted. The protocol for 2-step and 3-step using 32 cycles was validated for the Arktik and Veriti thermal cyclers using sub optimal DNA concentrations. In Figure 2.4, full Y-STR haplotypes were obtained at 125 pg and 62.5 pg for all conditions despite the 2 fold variation in the average allele peak height between the Arktik and the Veriti thermal cyclers. The 2 fold reduction in RFU’s evident for the Veriti at 125 pg and 62.5 pg as previously observed at optimal DNA concentrations had not comprised the production of full haplotypes. The allele peak heights averaged between the 2-step and 3-step protocols for the Arktik and Veriti respectively was (2338.5 *sd*±1473) and (1116. 6 *sd*±593.8) for 125pg DNA and (1139.9 *sd*± 684.3) and (557.7 *sd*±290.3 respectively for 62.5 pg DNA.

Figure 2.1: Box and whisker plot indicating Interlocus peak heights in RFUs between two step (red) and three step (turquoise) thermal cycling. The 50 percentile indicates the median and the whiskers indicates the variance. (Rstudio plot).

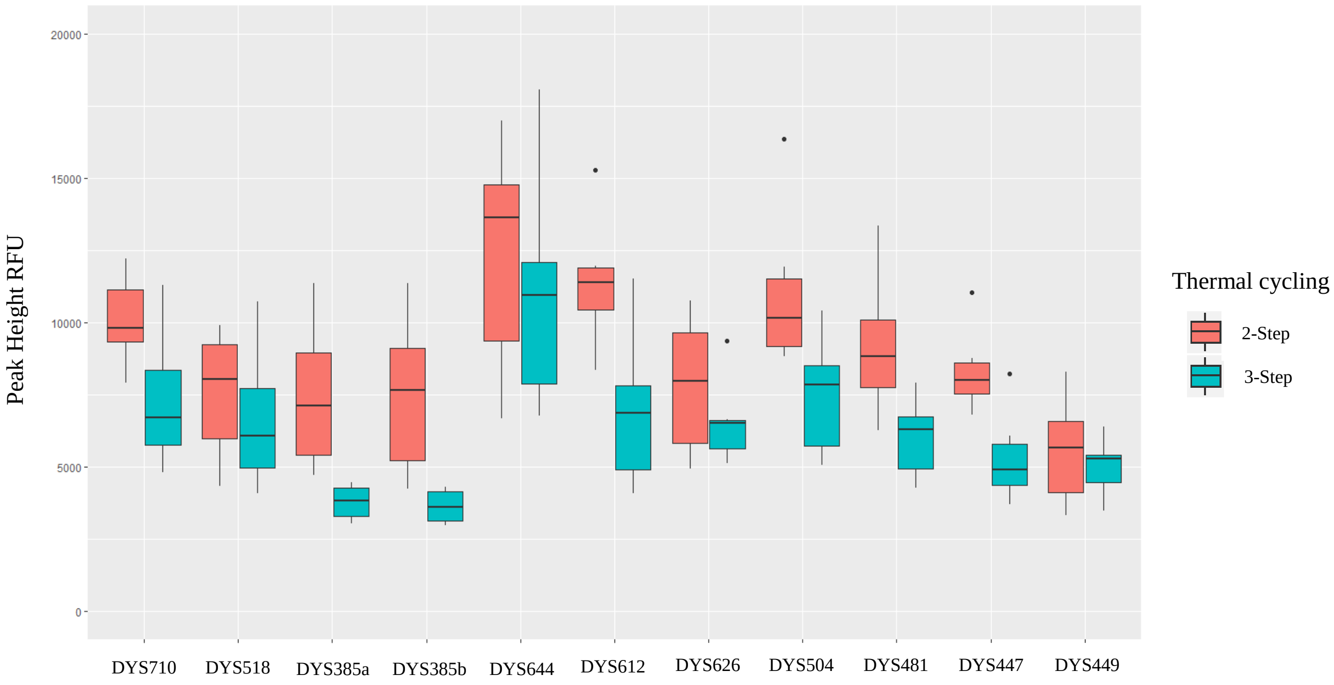


Figure 2.2: Y-STR profile comparative between 2-step and 3-step thermal cycling. Y-axis scale 18000 RFU. A) 2-step and B) 3-step thermal cycling amplified for 30 cycles.

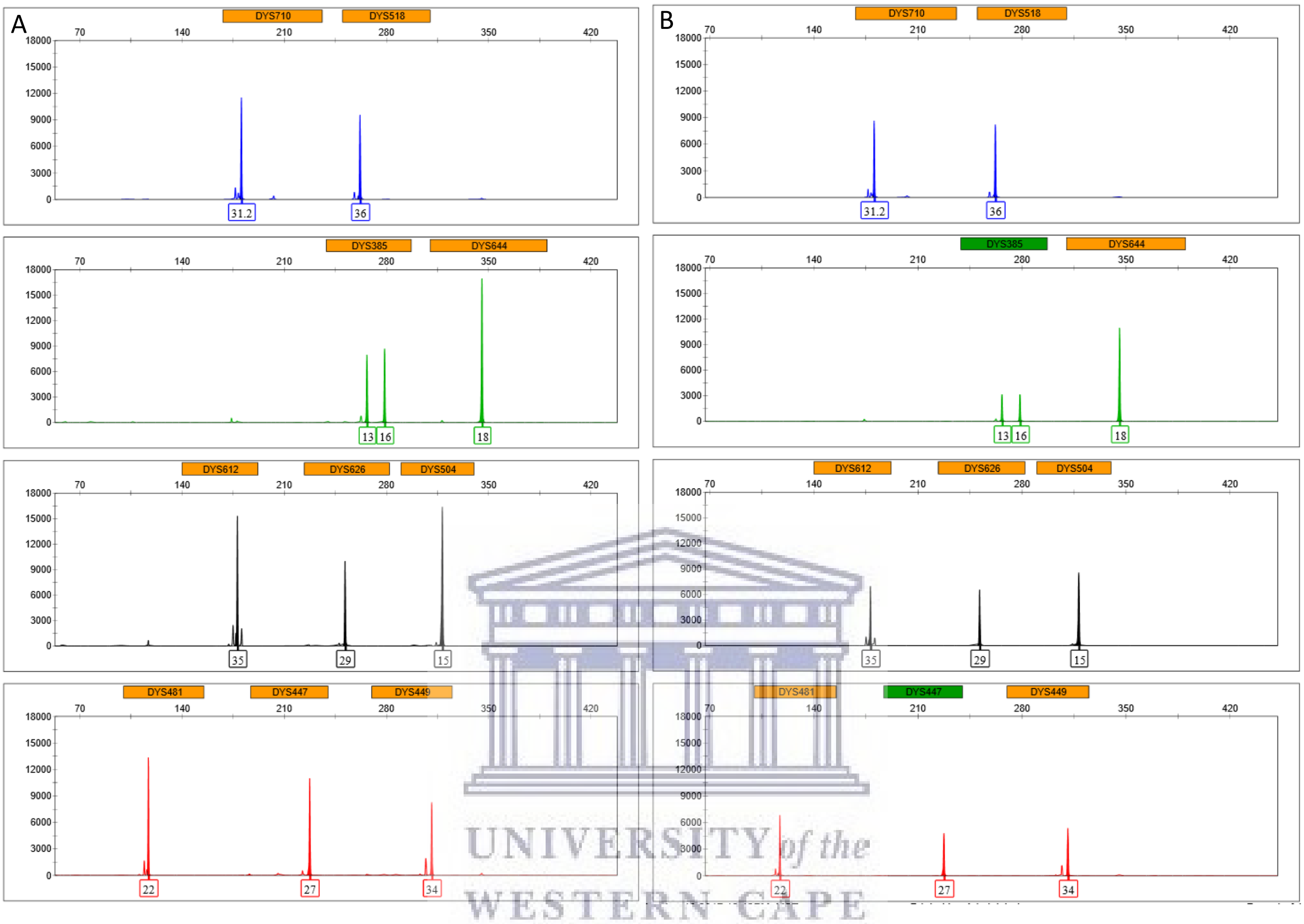


Figure 2.3: Box and whisker plot indicating allele peak height variation for the 2-step cycling for amplification across three thermal cyclers. ABI2720 (red), Arktik (green) and Veriti (blue), dots represent data points for DNA of 3 donors amplified in duplicate at 1ng for each thermal cycler. (Rstudio plot).

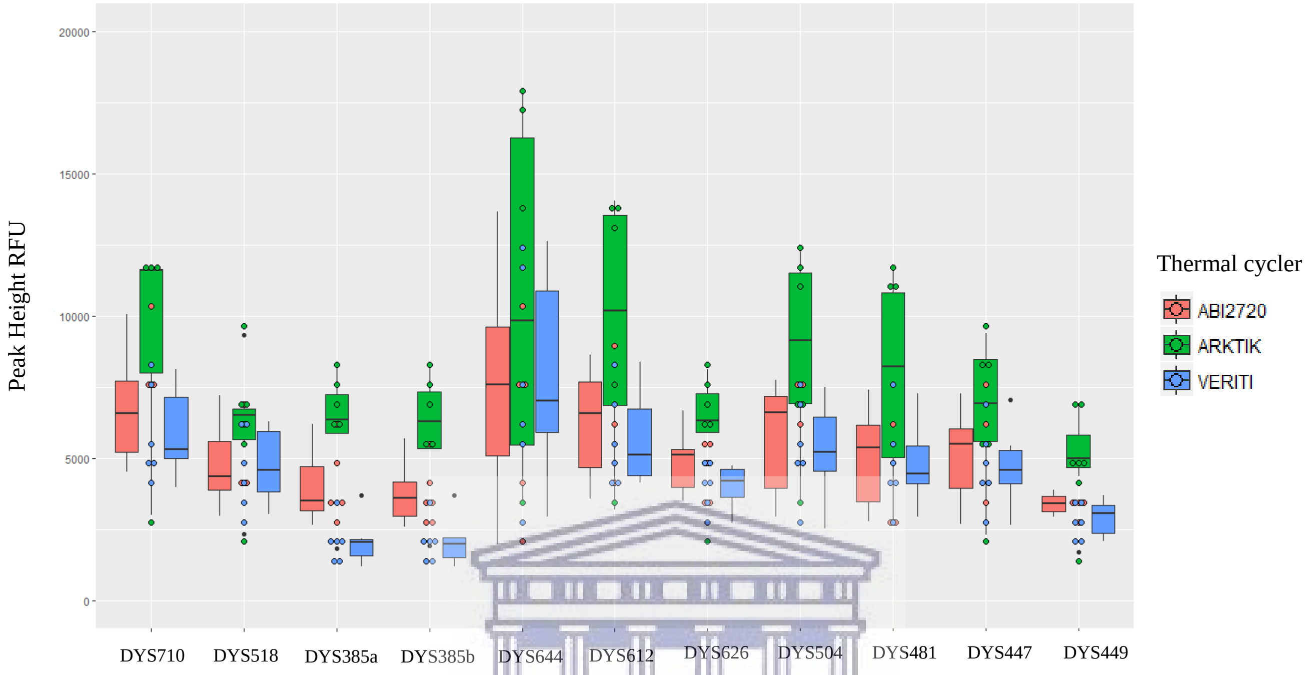


Figure 2.4: Box and whisker plot indicating allele peak height distribution between the Arktik and Veriti using low copy number templates. The 50 percentile indicates the median between 2-step and 3-step data and the whiskers indicates the variance around the upper and lower quartiles using 125 pg and 62.5 pg DNA from three donors amplified in duplicate for 32 cycles. (Rstudio plot)

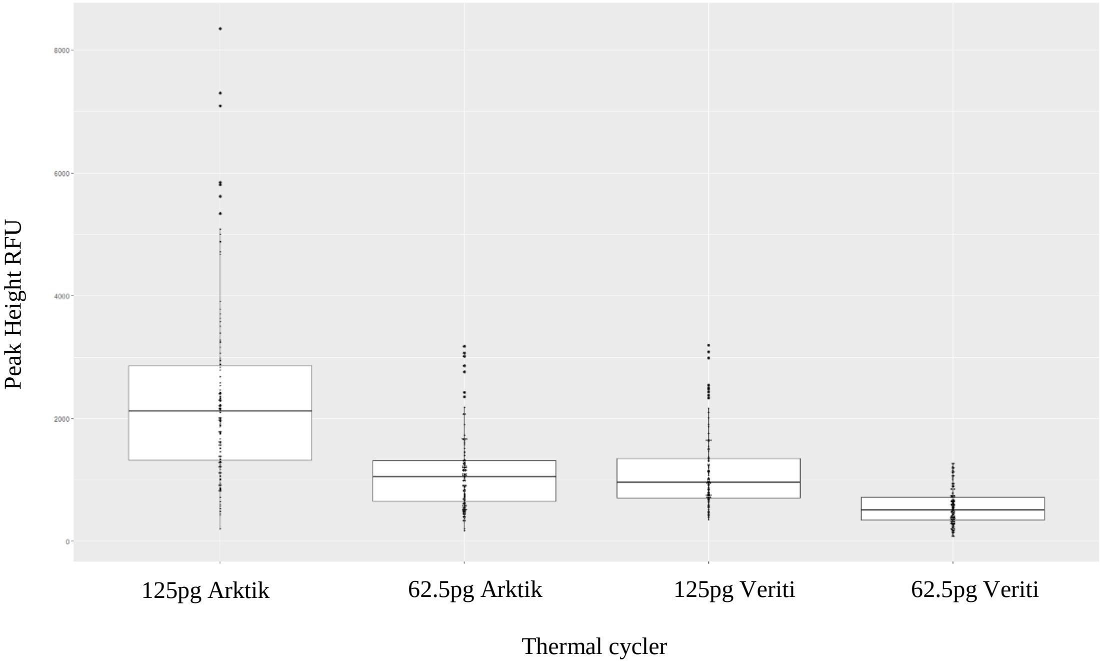


Table 1.3: Tabulated difference in allele peak heights and stutter percentages between 3-step and 2-step thermal cycling conditions.

	DYS710		DYS518	DYS385a	DYS385b	DYS644	DYS612		DYS626	DYS504	DYS481	DYS447	DYS449
	3-step cycling												
Max	11316		10718	4448	4298	18069	11525		9375	10404	7897	8214	6376
Min	4820		4064	3032	2978	6779	4091		5126	5058	4258	3695	3325
Ave RFU	7315,3		6636,5	3773,8	3629,5	11010,3	6954,8		6586,2	7485,5	6020,8	5330,7	4721,3
sd	2393,8		2463,9	615,0	596,8	4171,1	2737,7		1512,6	2083,1	1403,0	1626,9	1189,8
	2-step cycling												
Max	12218		9905	11377	11377	16998	15304		10753	16384	13359	11051	8279
Min	7922		4323	4701	4246	6690	8360		4944	8822	6266	6807	3933
Ave RFU	10087,8		7534,8	7452,2	7489,2	12354,5	11439,3		7823,7	11055,8	9199,0	8340,2	5888,3
sd	1570,9		2257,0	2585,8	2775,4	4044,0	2294,1		2404,4	2846,5	2485,6	1482,8	1654,7
	Stutter % for 3-step and 2-step cycling												
Stutter type	n-2	n-1	n-1	n-1	n-1		n-1	n+1	n-1	n-1	n-1	n-1	n-1
3-Step: **Ave %** ***sd***	11,00 0,35	6,90 1,13	10,44 0,52	10,39 0,54	11,17 2,40	No stutter	16,49 1,07	14,10 0,19	3,53 0,85	3,65 0,29	16,55 1,91	2,20 0,15	21,81 1,85
2-Step: **Ave %** ***sd***	11,50 0,42	7,15 0,42	10,98 1,75	11,07 2,43	13,03 0,96	No stutter	17,31 1,77	13,72 0,02	3,88 1,06	3,63 0,57	14,71 1,39	3,79 1,67	22,21 2,27

2.3.2 Sensitivity

To determine the fastest PCR duration and cycle number required to achieve desirable sensitivity, the DNA from 3 titrated between (500 pg, 250 pg, 125 pg, 62.5 pg and 31.2 pg) where amplified at 28, 30 and 32 cycles for the 2-step protocol on the Arktik thermal cycler. In Figure 2.5, full haplotypes were achieved at 28, 30 and 32 cycles up to 125 pg DNA, however the reproducibility to achieve full profiles at less than 125 pg was improved using 30-32 cycles. The fastest reliable PCR duration was achieved at 30 cycles to provide a 45 min assay which was selected for further validation hereafter.

2.3.3 Mixture study

Forensic casework and sexual assault evidence often present a mixture of biological evidence which cannot be physically separated before genotyping. To determine the extent of the UniQTyper™ Y-10 to resolve the male-male mixture components, a single male mixture with unshared alleles was amplified at ratios (1:4; 1:9 and 1:19) in four replicates. All the minor alleles were identified with peak heights in the range of 244 to 4892 RFUs across all the ratios tested (Figure 2.6). The minor allele average peak height across all loci for each respective mixture ratio was (2753.9 RFUs; *sd*± 885); (1400 RFUs; *sd*± 426.8) and (633.8 RFUs; *sd*± 264.3). See Figure 2.7 for an example the minor allele DNA profile resolution relative to the major contributing allele.

Female-male mixtures were tested to validate the sensitivity of the multiplex in excess of the female component. The female DNA was kept constant at 200 ng while decreasing the amount of male DNA between (2 ng, 1 ng, 400 pg, 200 pg and 100 pg) to yield ratios (1:100, 1:200, 1:500, 1:1000 and 1:2000) respectively (Figure 2.8). The average allele peak heights for female-male mixtures presented in Figure 2.8 indicates that the excess had not compromised the sensitivity to detect the minor male component. The average allele peak height for the minor male component across all loci for each respective mixture ratio was (10167.6 RFUs; *sd*± 3537.7); (8360.9 RFUs; *sd*±3429.3); (3267.4 RFUs; *sd*± 1370.1); (1628.3 RFUs; *sd*± 658.0) and (701.3 RFUs; *sd*± 357.7) (see profiles in Figure 2.9)

Figure 2.5: Titration of male DNA for amplification reactions containing 500 pg, 250 pg, 125 pg 62.5 pg and 31.2 pg at increasing cycle number. The DNA of 3 male donors were amplified in duplicate for each concentration for the 2-step protocol on the Arktik for 28, 30 and 32 cycles.

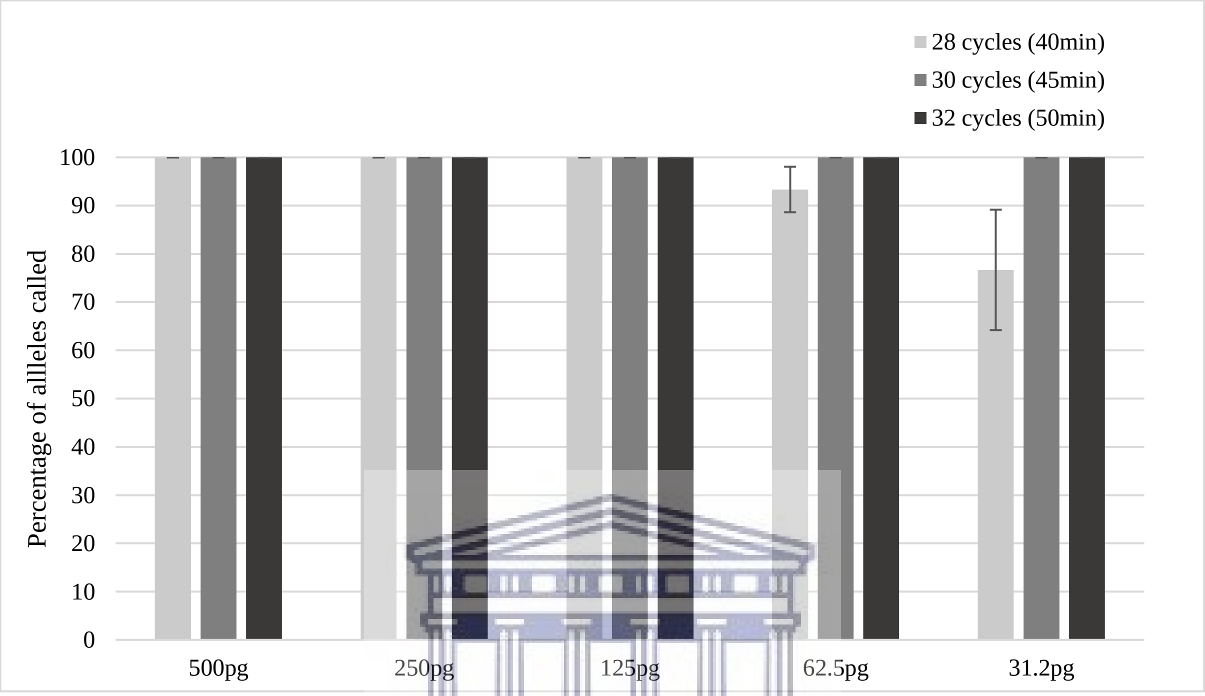


Figure 2.6: Allele peak heights of the minor alleles for male-male reaction mixtures at ratios of 1:4, 1:9 and 1:19 for a total of 2ng DNA.

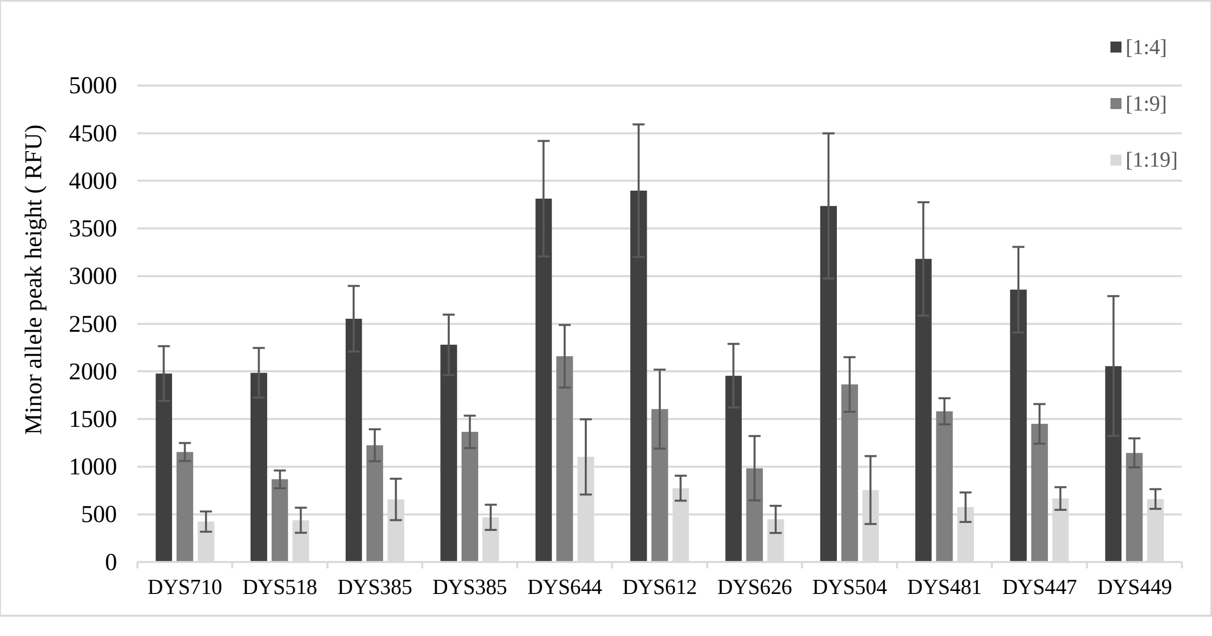


Figure 2.7: Male: male mixture haplotype resolution for A) 1:9 and B) 1:19 for a total of 2ng DNA. Y-axis scale 12000 RFU.

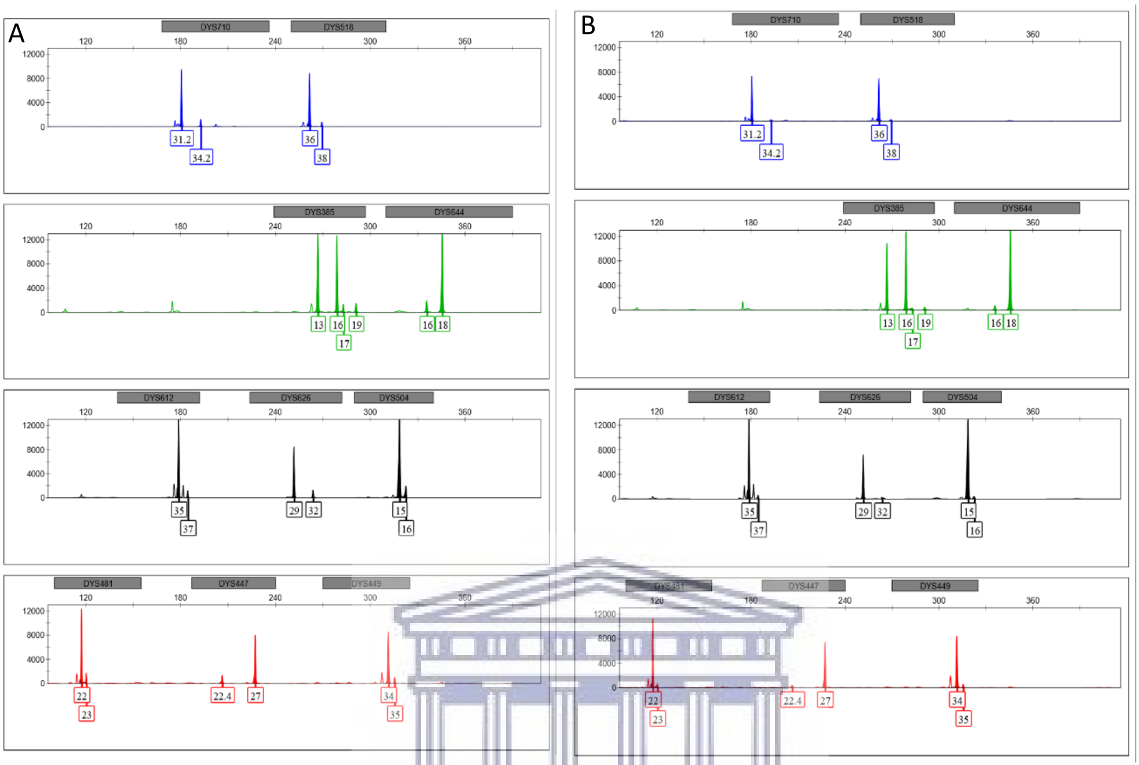


Figure 2.8: Allele peak heights of the minor male alleles for female-male reaction mixtures at ratios of (1:100, 1:200, 1:500, 1:1000 and 1:2000) with the female constant at 200 ng and male at decreasing concentrations.

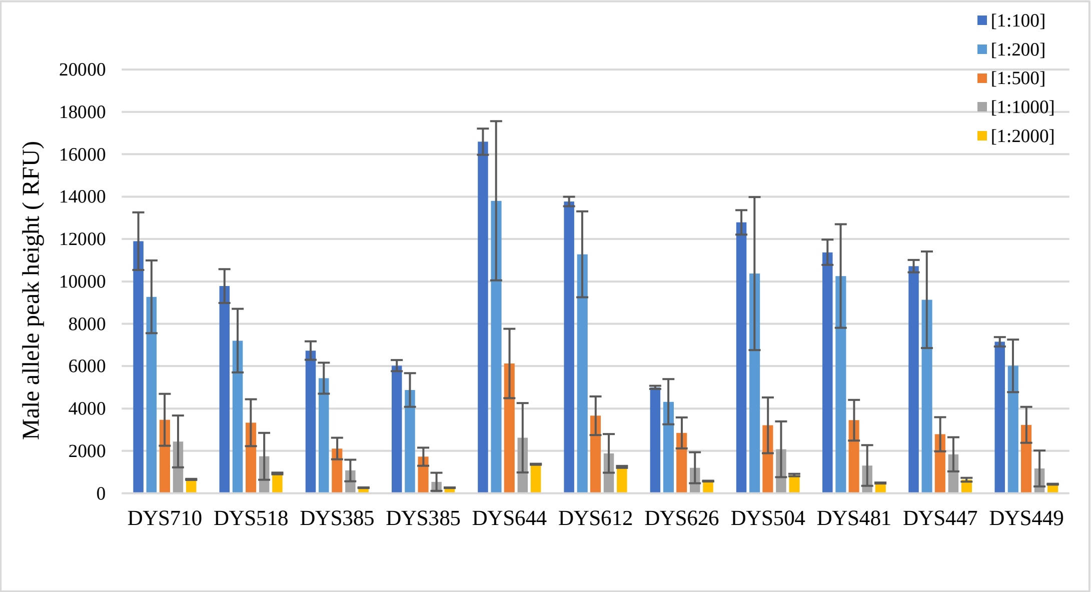
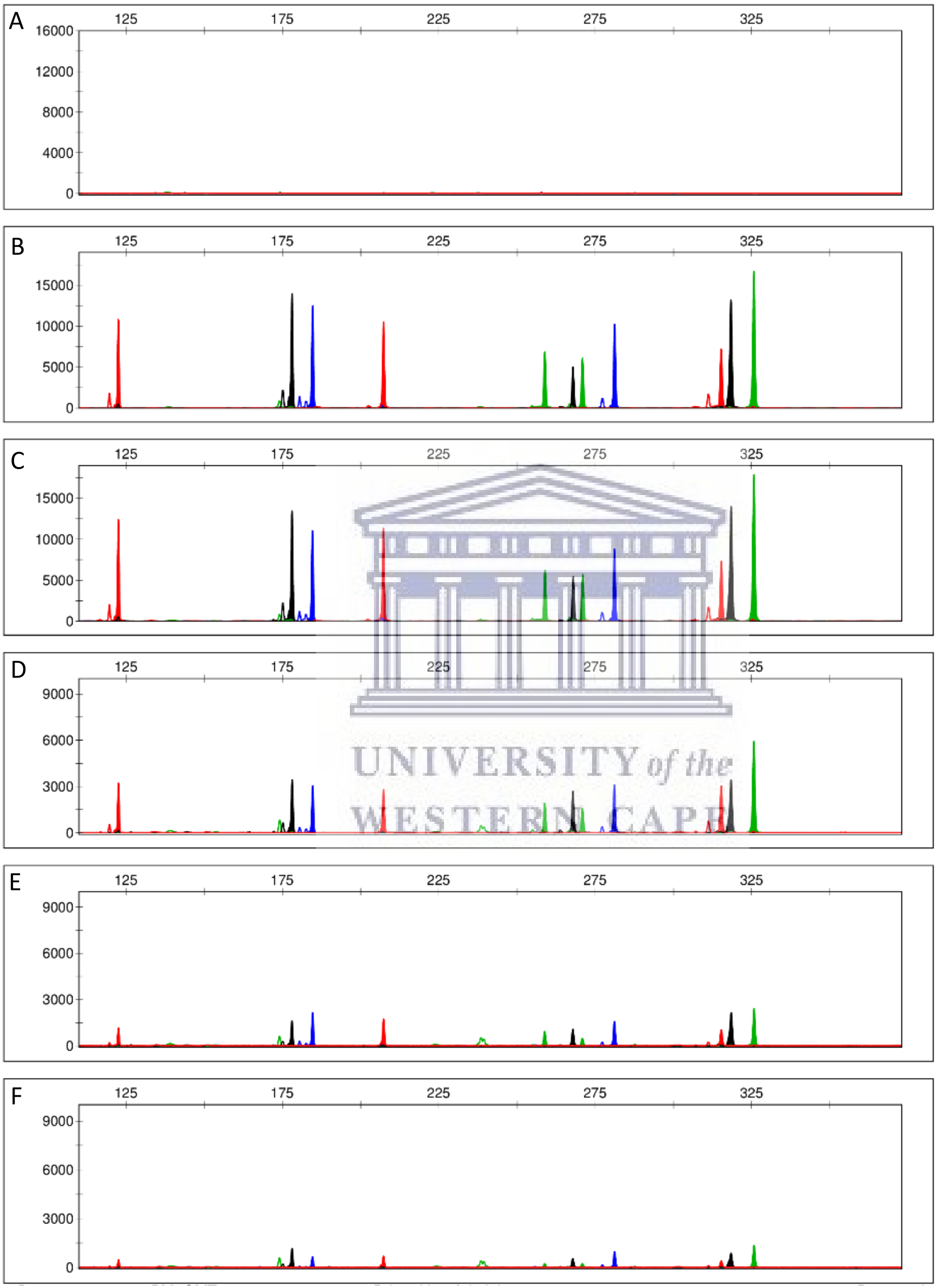


Figure 2.9: Male haplotype resolution in excess of the female DNA component. Female DNA kept constant at 200 ng and male decreased to provide ratios A) 0:1, B) 1:100, C) 1:200, D) 1:500, E) 1:1000, F) 1:2000. For visual purpose Y-axis scaled to 16000 RFU (A-C) and 9000 RFU (D-F).



2.3.4 Species specificity

Since the current Y-STR panel is intended for human identification purposes, cross reactivity was characterized for male and female non-human primates and domestic animals. The Y-STR primers had no cross reactivity to female Chimpanzee, Gorilla, Vervet monkey and to female/ male cat and dog species. Primer interaction with male chimpanzee DNA presented products at loci DYS385, DYS612 and DYS481 of respective size (298.3 bp, 192.2 bp and 113.3 bp). In Gorilla loci DYS385 presented products of size (249.0 bp, 255.13 bp and 276.19 bp) and single products at loci DYS612, DYS504 and DYS481 of size (190.8 bp, 295.0 bp and 123.0 bp) respectively. For the male Vervet monkey peaks were present at DYS385 of size (289.0 bp and 293.0 bp). See Figure 2.10 for examples of the amplification products detected in higher primates and monkey.

2.3.5 Stability study

The stability of the multiplex was tested under the adverse conditions for PCR inhibition and DNA degradation which is often encountered for casework samples. The two common PCR inhibitors humic acid from soil and hematin from blood was tested to evaluate the level of inhibitor tolerance. Full profiles were achieved using 1ng DNA at 25 ng/µl humic acid and 10 µM hematin for 3 donors DNA amplified in duplicate. On average 97% of alleles were called when humic acid was increased to 100 ng/µl and 53% at 200 ng/µl humic acid for which the larger alleles were the first to drop out. When increasing hematin to 50 and 100 µM the average alleles called were 83% and 17% respectively, while complete drop out occurred at 200 µM hematin. Full profiles were obtained from dirty DNA extracts of blood-stained denim jeans and leather which presented a source for PCR inhibitors indigo and tannic acid respectively. The blood-stained soil samples exposed for 6h to high humidity and direct sunlight every 12 h outdoor period provided full profiles outdoor exposure of 12hs, while the 24and 36h outdoor samples failed to yield haplotypes (Supplementary 1 Figure 1.2).

Figure 2.10: Species specificity test for non-human primates and domestic animal. A) Chimpanzee *(Pan troglodytes)*, B) Gorilla *(gorilla gorilla)*, C) Vervet monkey *(Chlorocebus pygerythrus)*, D) cat *(Felis felis)*, E) dog *(Canis familiaris)*, F). For visual purpose Y-axis scaled to 30 000 RFU (A-B) and 1500 RFU (C-E).

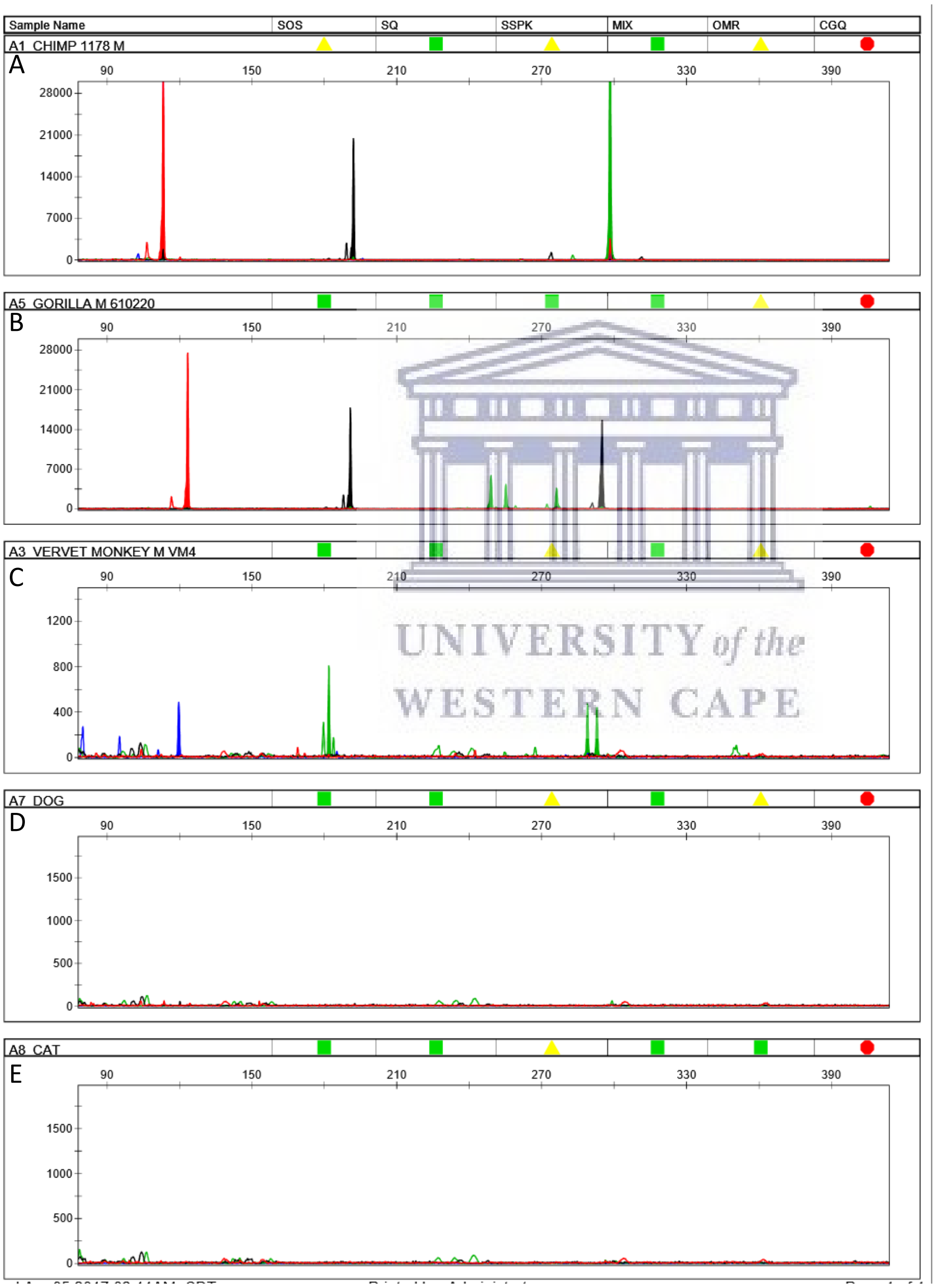


Figure 2.11: Average percentage of alleles detected at increasing concentration of the respective blood and soil PCR inhibitors humic acid and hematin.

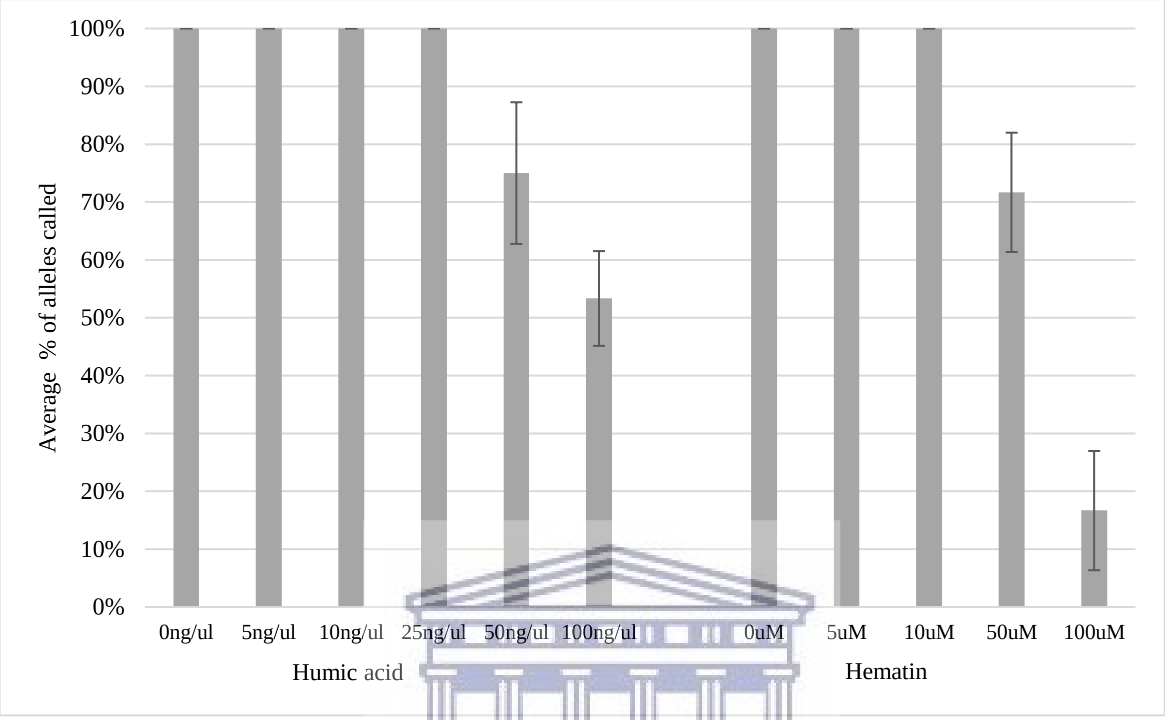
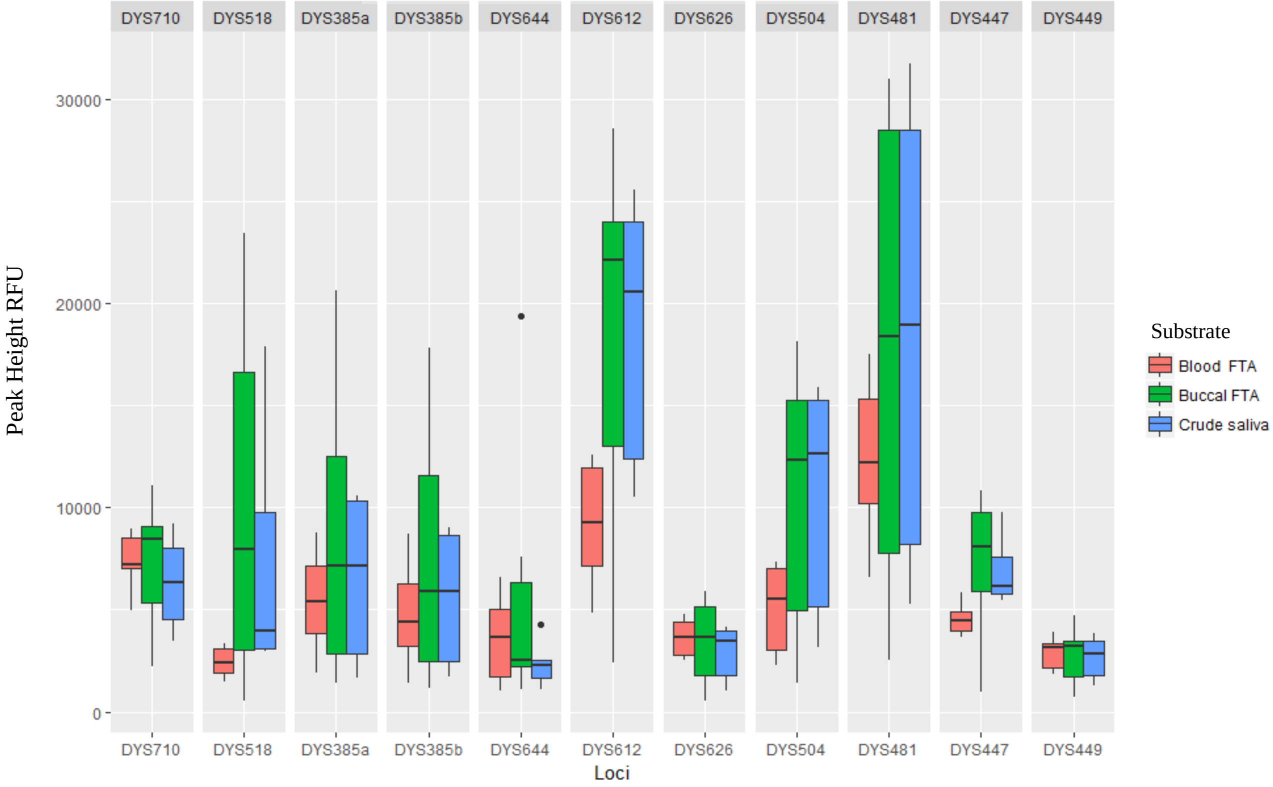


Figure 2.12: Variability in allele peak heights across the profiles generated from direct amplification substrates. Blood stored on FTA (Red), Buccal swab transfer to FTA (green) and crude saliva (blue).



2.3.6 Direct amplification

Direct amplification from biological material without DNA extraction can provide rapid turnover times for generating reference genotypes. For these tests blood and saliva of two donors were amplified from FTA cards and crude biological material in full reaction volumes. For blood and buccal cells collected on FTA a single 1.2 mm disc provided full haplotypes for all replicates. In tests using untreated crude saliva and blood for direct amplification, full profiles were achieved routinely using crude saliva, while adding blood directly to the PCR failed to yield amplification. The profiles for buccal cells stored on FTA, crude saliva and blood FTA provided allele peak heights at > 5000 RFUs for 54%; 60% and 58% of the data points for each respective sample type. In Figure 2.12 the box plot with the median given as the 50 percentile indicates a larger peak height variation using buccal FTA and crude saliva samples compared to the peak heights for FTA blood, the later also presented more consistent allele peak balance across all loci (Figure 2.12). The minimum RFUs were (519; 1027; and 1043) and maximum RFUs (23442; 31699 and 17531) for buccal FTA, crude saliva and blood FTA respectively (Supplementary 1 Figure 1.3).

2.3.7 PCR based procedures

Various technical factors which may influence the genotyping performances were considered to evaluate the impact of potential sub optimal conditions. These technical factors previously referred to as PCR based procedure by Thompson et al. 2012 is of great importance during technical validation as subtle changes in accuracy and precision of the instruments may have a significant effect on genotyping performances. We observed a prominent decline in the allele peak heights at sub optimal annealing temperatures tested at 2°C and 4°C below the optimal temperature of 65°C (Figure 2.13). Loci most affected were DYS644, DYS626 DYS447 and DYS449 which presented a 2-3 fold under suboptimal temperatures. Complete allele dropout and peak height lower that 50 RFU were observed at loci DYS644, DYS626 and DYS447 for annealing at 61°C. Temperature variation which may exist between various commercial PCR machines can be a major source of allele peak height variability and may lead to discrepancies in genotyping performances across platforms (Ho et al. 2008).

The potential of pipetting errors of the master mix or primer mix may result in subtle changes in the Taq polymerase or the primer concentration in the PCR reaction. The effect of subtle reduction in Taq polymerase from 1 U to 0.8-0.7 U per reaction gave a pronounced decrease in allele peak height across the profiles (Figure 2.14).We also identified during the optimization (data not shown) that the thermal cycling is also sensitive to Taq polymerase concentrations above 2 U per reaction for which we of observed a 5 fold reduction in peak heights. We observed no major negative or positive correlation in allele peak height with subtle fluctuations less than and more than the optimal final concentration of 1 X in a 25 µl PCR reaction. It was evident that subtle variation in primer input had less of an influence on allele peak height compared to Taq polymerase.

Figure 2.13: Average allele peak height variations at optimal and sub optimal annealing temperatures.

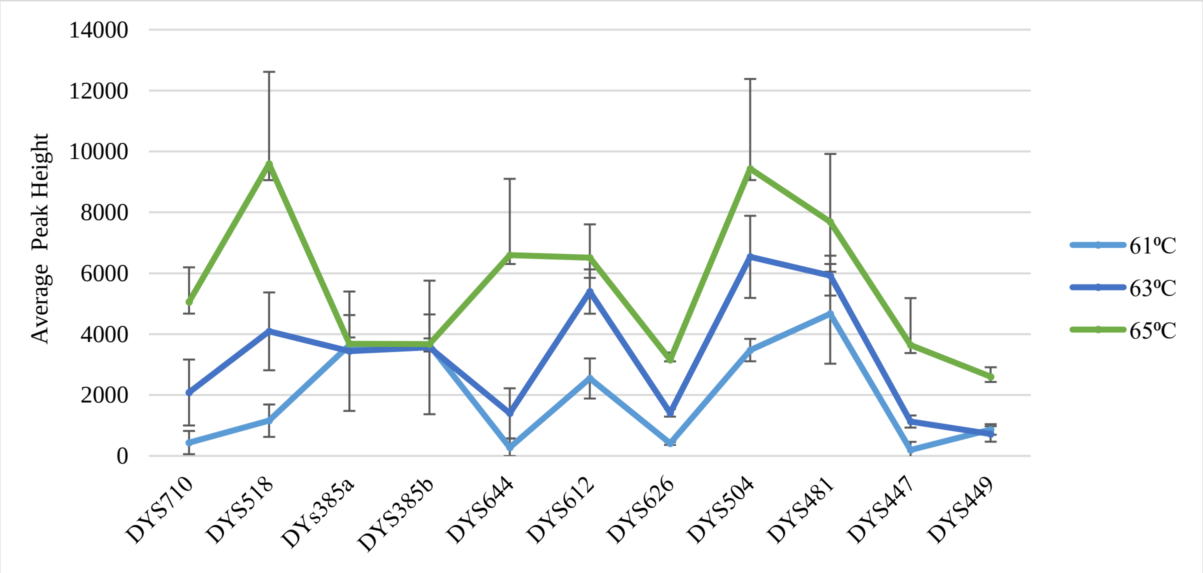


Figure 2.14: Allele peak heights obtained using subtle variation Taq polymerase and input primer concentrations. Top panel: Titration of Taq polymerase (1.2, 1.0, 0.8, 0.7 and 0.6 U) in the mater mix and bottom panel titration of primer input (0.7, 0.8, 0.9, 1.0 and 1.5 X).

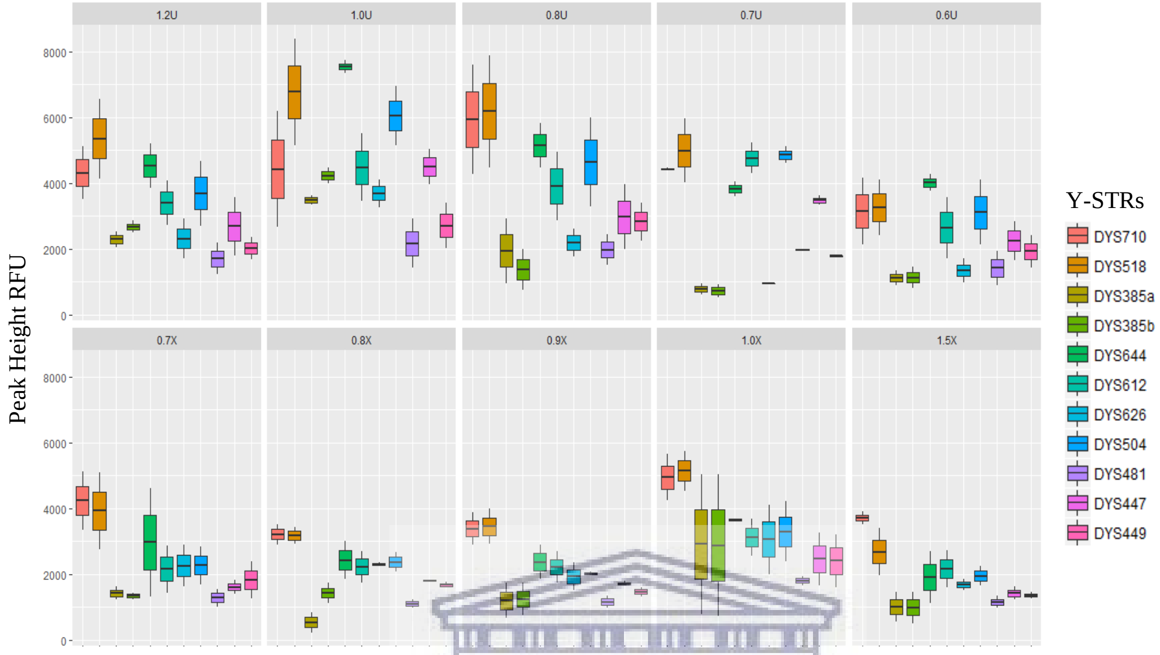
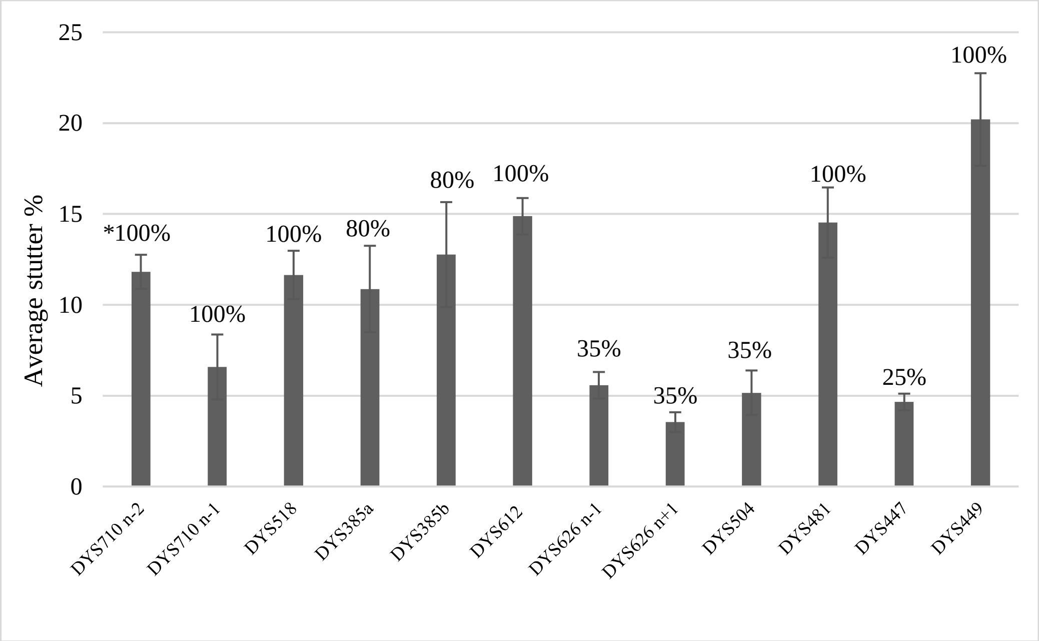


Figure 2.15: Average stutter percentages obtained across the Y-STR loci.

The astrix (*) indicates the percentage of stutter occurrence per locus (i.e. proportion of tested individuals showing stutter). Stutter occurrences and percentage against the true allele was estimated from N=100 haplotypes. DYS644 was not presented due to the absence of stutter products.



2.3.8 Concordance, quality control and stutter calculations

The DNA labelled FDLP001 and FDLK003 in Table 1.4 below represents the in-house control DNA samples confirmed for the correct allele repeat number trough the sequencing of each of the 10-Y STR alleles. For the in-house controls and commercial control DNA 007, 2800M and NIST components A-C, alleles at loci DYS481 and DYS385 were in concordance with the Powerplex® Y23 genotyping system (data not shown) and the sequence data available form STRbase for NIST components A-C for loci DYS710, DYS449 and DYS481 (https://strbase.nist.gov/srm2395.htm). Population data was also generated internally for n=95 out of the 114 Xhosa individuals previously profiled by Purps et al.2014. For the shared loci DYS481 and DYS385 alleles were in concordance. In Figure 2.15, as previously observed stutter was absent at DYS644. This observation which is beneficial for mixture analysis is a valued maker of the panel. Locus DYS644 is a pentanucleotide repeat which is a type of structure well known to present reduced levels of stutter (de la Puente et al. 2017). Stutter was less frequently observed for DYS626, DYS504 and DYS447 (< 30%, N=100 haplotypes). These loci also had the lowest percentage of stutter, observed on average 4 orders lower than the highest recorded stutter at DYS449 (> 20%). Stutter in the range of 10-15% was observed across loci (DYS710, DYS518, DYS385, DYS612 and DYS481).

Table 1.4: Haplotypes for control DNA used to establish concordance. Internal sequenced controls (FDLP001 and FDLK003), 007 DNA (Thermo Fisher), 2800M (Promega) and NIST A-C (National Institute of Standards and Technology).

	DYS710	DYS518	DYS385a/b	DYS644	DYS612	DYS626	DYS504	DYS449	DYS447	DYS481
FDLP001	32.2	41	11_14	14	35	33	15	35	23	24
FDLK003	33	39	16_16	24.4	37	25	16	30	26	24
007 DNA	35	37	11_14	17	37	30	17	30	25	22
2800M DNA	31.2	36	13_16	18	35	29	15	34	27	22
NIST comp A	36	39	12_15	16	38	31	16	28	24	22
NIST comp B	34.2	40	14_17	16	38	34	17	32	25	23
NIST comp C	35.2	38	17_20	23.4	34	28	13	30	25	28

2.4 Conclusion

This invention referred to as the UniQTyper Y-10™ kit is previously known as to maximize male discrimination in South Africa populations. In this chapter a novel multiplex PCR chemistry was optimized and validated to provide one of the fastest known Y-STR profiling assays to date. The rapid nature of the system was shown to have no adverse effect on the ability to deliver a sensitive, reproducible and robust profiling system which may be applicable for forensic applications. The sensitivity of the assay was improved with a longer PCR duration, however full haplotypes were obtainable at 62.5 and 31.2 pg in less than one hour. The fastest PCR duration to deliver reliable performance was obtained for a PCR time of 45min for both forensic type samples and database types using a direct application approach. Compared to the previously developed UWC-10plex systems the developmental validation performed herein demonstrated a distinguished performance. Most noteworthy was the reduction in the PCR duration from 2h:30min to a 45min while maintaining a superior level of sensitivity compared to validation in D'Amato et al. 2011. The validation herein also supports the improvements to the male: female admixture proportion and the delivery of a genotyping system with capabilities of direct application.

2.5 References

Burrows AM, Ristow PG, Amato MED (2017) Preservation of DNA from saliva samples in suboptimal conditions. Forensic Sci Int Genet Suppl Ser 6:e80–e81. https://doi.org/10.1016/j.fsigss.2017.09.050

Butts, E. L. R., & Vallone, P. M. (2014). Rapid PCR protocols for forensic DNA typing on six thermal cycling platforms. *ELECTROPHORESIS*, *35*(21–22), 3053–3061. https://doi.org/10.1002/elps.201400179

Butts, E. L. R., & Vallone, P. M. (2014). Rapid PCR protocols for forensic DNA typing on six thermal cycling platforms. *ELECTROPHORESIS*, *35*(21–22), 3053–3061. https://doi.org/10.1002/elps.201400179

Comey, C.T et al (1994). DNA extraction strategies for amplified fragment length polymorphism analysis. *J Forensic Sci*, *39*, 1254–1269. Retrieved from http://ci.nii.ac.jp/naid/10012199854/en/

D'Amato ME, Bajic VB, Davison S (2011) Design and validation of a highly discriminatory 10-locus Y-chromosome STR multiplex system. Forensic Sci Int Genet 5:122–125. https://doi.org/10.1016/j.fsigen.2010.08.015

de la Puente, M., Phillips, C., Fondevila, M., Gelabert-Besada, M., Carracedo, Á., & Lareu, M. V. (2017). A forensic multiplex of nine novel pentameric-repeat STRs. *Forensic Science International: Genetics*, *29*, 154–164. https://doi.org/https://doi.org/10.1016/j.fsigen.2017.04.007

Gill, P., Jeffreys, A. J., & Werrett, D. J. (1985). Forensic application of DNA "fingerprints". Nature, 318(6046), 577–579.

Gopinath, S., Zhong, C., Nguyen, V., Ge, J., Lagacé, R. E., Short, M. L., & Mulero, J. J. (2016). Forensic Science International : Genetics Developmental validation of the Y-filer™ Plus PCR Amplification Kit : An enhanced Y-STR multiplex for casework and database applications, *24*, 164–175. https://doi.org/10.1016/j.fsigen.2016.07.006

Hansson, O., Gill, P., & Egeland, T. (2014). STR-validator: An open source platform for validation and process control. *Forensic Science International: Genetics*, *13*, 154–166. https://doi.org/https://doi.org/10.1016/j.fsigen.2014.07.009

Ho Kim, Y., Yang, I., Bae, Y.-S., & Park, S.-R. (2008). Performance evaluation of thermal cyclers for PCR in a rapid cycling condition. *BioTechniques*, *44*(4), 495–505. https://doi.org/10.2144/000112705

Jovanovich, S., Williams, S., Park, C., & Gangano, S. (2014). Forensic Science International : Genetics Developmental validation of the GlobalFiler® express kit , a 24-marker STR assay , on the RapidHIT® System. *Forensic Science International: Genetics*, *13*, 247–258. https://doi.org/10.1016/j.fsigen.2014.08.011

Kasu, M., & Shires, K. (2015). The validation of forensic DNA extraction systems to utilize soil contaminated biological evidence. *Legal Medicine*, *17*(4). https://doi.org/10.1016/j.legalmed.2015.01.004

Medrano, J. F., Aasen, E., & Sharrow, L. (1990). DNA extraction from nucleated red blood cells. *BioTechniques*, *8*(1), 43.

Miller, S. A., Dykes, D. D., & Polesky, H. F. (1988). A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Research*, *16*(3), 1215. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/3344216

Oostdik, K., French, J., Yet, D., Smalling, B., Nolde, C., Vallone, P. M., Sprecher, C. (2013). Developmental validation of the PowerPlex 18D System, a rapid STR multiplex for analysis of reference samples. *Forensic Science International: Genetics*, *7*(1), 129–135. https://doi.org/10.1016/j.fsigen.2012.07.008

Roewer, L. (2009). Y chromosome STR typing in crime casework. *Forensic Science, Medicine, and Pathology*, *5*(2), 77–84. https://doi.org/10.1007/s12024-009-9089-5

Romsos, E. L., & Vallone, P. M. (2015). Forensic Science International : Genetics Rapid PCR of STR markers : Applications to human identi fi cation, *18*, 90–99. https://doi.org/10.1016/j.fsigen.2015.04.008

RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/

Science International : Genetics Rapid PCR of STR markers : Applications to human identi fi cation, *18*, 90–99. https://doi.org/10.1016/j.fsigen.2015.04.008

Shahzad, M. S., Bulbul, O., Filoglu, G., Cengiz, M., & Cengiz, S. (2009). Effect of blood stained soils and time period on DNA and allele drop out using Promega 16 Powerplex® kit. *Forensic Science International: Genetics Supplement Series*, *2*(1), 161–162. http://doi.org/10.1016/J.FSIGSS.2009.08.192

Thompson, J. M., Ewing, M. M., Frank, W. E., Pogemiller, J. J., Nolde, C. A., Koehler, D. J., Storts, D. R. (2013). Developmental validation of the PowerPlex® Y23 System: A single multiplex Y-STR analysis system for casework and database samples. *Forensic Science International: Genetics*, *7*(2), 240–250. https://doi.org/https://doi.org/10.1016/j.fsigen.2012.10.013

Vallone, P. M., Hill, C. R., & Butler, J. M. (2008). Forensic Science International : Genetics Demonstration of rapid multiplex PCR amplification involving 16 genetic loci §, *3*, 42–45. https://doi.org/10.1016/j.fsigen.2008.09.005

Chapter 3:

3.0 Novel Y-chromosome short tandem repeat sequence variation for loci DYS710, DYS518, DYS385, DYS644, DYS612, DYS626, DYS504, DYS481, DYS447 and DYS449

3.1 Introduction:

The forensic application of Y-chromosome short tandem repeat (Y-STR) genotyping has become a fundamental tool for male identification in sexual offense, paternity, familial search and missing victim investigations (Roewer et al. 1992; Roewer 2009 and Kayser et al. 2017). The male specific and hemizygous state of Y-STRs make it particularly favourable for sexual assault analysis when the female and/or male mixed genotype cannot be distinguished using autosomal STR systems (Roewer 2009).

The general limitation of Y- STR genotyping is its reduced power to exclude paternal relatives as potential contributors to the DNA evidence. Y-STRs are located on the non-recombinant region of the Y-chromosome, therefore in absence of mutational events, paternal relatives will share identical haplotypes. The discrimination between males also presents a challenge when the events of inbreeding, population isolation and patrilocal practices reduce the Y-chromosome diversity within a population (Oota et al. 2001; Ballantyne et al. 2012 and Nuñez et al. 2015). Improving Y-STR discriminatory capabilities is currently achieved by three main approaches: 1) by utilizing markers of higher mutation rates (Ballantyne et al. 2010 and 2012), 2) developing larger Y-STR panels for capillary electrophoresis (CE) (Thompson et al. 2013 and Purps et al. 2014) and 3) Sequence discrimination by defining Y-STR pattern variants and single nucleotide polymorphisms (SNPs) between isometric alleles (Garza and Freimer 1996 and Warshauer et al. 2015).

The latter approach has rapidly gained attention for the potential of making distinctions between shared male haplotypes in forensic applications (Warshauer et al. 2015 and Huszar et al. 2018). The increase in autosomal and Y- chromosome STR mutability is well known to be influenced by the repeat motif length, nucleotide composition and complexity of the repeat structure (Ballantyne et al. 2010, Garza and Friemer 1996). For Y-STR analysis these structure characteristics may potentially offer improved discriminatory capabilities from a sequence perspective as opposed to length polymorphisms (Ballantyne et al. 2010 and Warshauer et al. 2015). The characterization of Y-STR allele sequence variation through both Sanger and Next Generation Sequencing (NGS) has indeed presented great promise for forensic applications through identification of novel Y-STR repeat pattern arrangements and SNPs between which are identical in size (i.e. isometric) (Garza and Friemer 1996, Warshauer et al. 2015, Huszar et al. 2018).

In this chapter we present 153 Y-STR sequences across 10 Y-STR markers from the forensic applicable multiplex panel developed by D'Amato et al. 2011. We present a series of novel allele sequences (n=23), repeat pattern variants (n=37) and SNP variants (n=2) observed across the South African population groups: English, Asian Indian, Coloured (admixed) and seven native Bantu populations. Furthermore, a comprehensive summary of all known variation reported to date for these 10 Y-STRs is provided and the most informative loci in the panel recommended for the discrimination of isometric alleles.

3.2 Material and methods

3.2.1 Sample collection and DNA extraction

Saliva samples were collected with informed consent from unrelated male voluntary donors. Ethical Approval was granted by the University of the Western Cape Senate Research Committee (15/4/97 and 10/3/39). Saliva samples were preserved in the buffer formulation of Burrows et al. 2017 and the DNA extracted by salting out adapted from Miller et al. 1988. The extracted DNA samples were quantified using the Nanodrop 2000 (Thermo Fisher) and the DNA concentration normalized to 2 ng/µl for singleplex PCR reactions.

3.2.2 Criteria for allele sequencing

Population genotyping of 1447 male haplotypes from (D'Amato and Kasu et al. 2017 and D'Amato, Kasu, Lesaoana, unpublished) presented a set of alleles that we considered for sequencing. The criteria for selecting alleles for sequencing were based on: 1) Novel off ladder alleles; 2) Alleles not previously observed (e.g. out of range) and 3) alleles previously observed with no reported DNA sequence. As a quality control measure between sequencing and CE at least one isomeric allele was selected for each Locus for either a comparison within this report or to our previously published sequences (D'Amato et al. 2017). The aim of allele screening and sequencing was also intended to ultimately construct an allelic ladder comprising > 150 cloned and sequenced Y-STR alleles (covered in Chapter 3).

3.2.3 Singleplex PCR for cloning and Sanger sequencing.

DNA samples for cloning were amplified in singleplex using amplification conditions and primer sequences previously published (D'Amato et al. 2010 and 2011).

3.2.4 Allele Cloning

PCR products were cloned into linearized pMiniT 2.0 in a 10µl ligation reaction following the manufacturer's instructions for the NEB cloning kit (NEB #E1202). Plasmids were transformed into NEB 10-beta Competent *E. coli* (NEB #C3019) by incubating the heat shocked component *E. coli* cells with the NEB 10-beta Stable Outgrowth Medium for 1 h at 37°C. Outgrowth cultures were plated onto agar palates containing 100 µg/ml ampicillin and inverted plates were incubated overnight at 37°C. Screening of inserts with colony PCR was conducted using the OneTaq 2X Master Mix and 0,3 µM F and R cloning analysis primer in a 50 µl reaction volume. The PCR cycle conditions were programmed as recommended (NEB #E1202) on the Arktik thermal cycler (Thermo Fisher). The amplicons of positively transformed colonies were purified with Exonuclease I (NEB#M0293) for removal of excess primers and the dephosphorylation of dNTP's using Shrimp Alkaline Phosphatase NEB #M0371). All cloning was conducted at the facilities of the industrial partner Inqaba Biotec.

3.2.5 Sequencing of cloned DNA

Cycle sequencing reactions were prepared using the Big Dye Terminator v. 3.1 kit (Thermo Fisher) following the manufacturer recommended conditions for a 10 µl reaction volume. Sequencing reactions were purified using ZR-96 DNA Sequencing Clean-up Kit™ (Zymo Research) and sequencing performed on the Genetic Analyser ABI3500xl (Thermo Fisher). A total of 153 allele sequences across loci DYS710, DYS510, DYS385, DYS644, DYS504, DYS612, DYS626, DYS481, DYS449, DY477 and DYS449 were aligned to GenBank reference AC007972.4, AC010972.3, AC022486.4, AC006462.3, AC006383.2, AC007320.3, AC010972.3, AC016991.5, NT_011875.11 and AC051663.9 respectively and also the reference sequences from D'Amato et al. 2010. Alignments were assembled using the software package MEGA 7 (Kumar et al. 2008).

3.2.6 Capillary electrophoresis

Fluorescently labelled PCR products were electrophoresed on an ABI3500 (Thermo Fisher) and spectral calibration achieved using an in house developed 5-dye matrix standard (Cloete et al. 2016). Samples were prepared using 9.7 µl HiDi™ formamide (Thermo Fisher), 0.3 µl GSLIZ® 500 (Thermo Fisher) and 1 µl PCR product and allelic ladder. Sample plates were denatured at 95ºC for 5min and snap cooled on ice for 3min. PCR products were injected in an 8-capillary 36 cm array for 15s at 1.2kV and fragment separation achieved using a POP4 polymer for 20 min at 15kV at a run temperature of 60ºC. Data was analysed using GeneMapper® ID-X software v1.4.

3.3 Results and discussion:

3.3.1 Classification of allele sequences

The total sequences generated per loci were: DYS710 (n=15), DYS518 (n=9), DYS385 (n=17), DYS644 (n= 31), DYS612 (n=12), DYS626 (n=11), DY504 (n= 11), DYS481 (n= 13), DYS447 (n= 17) and DYS449 (n= 17). A total of 94 of the 153 Y -STR sequences (GenBank accession numbers =MK005372 - MK005525, Table 2.2) represented sequences that were not previously reported. These were subdivided into 3 categories: 1) Size Homoplasy (SH), representing alternative repeat pattern arrangements for allele sequences, 2) Novel Allele Sequence (NAS), representing sequences for alleles not previously reported and 3) Sequences of Known Alleles (KA), representing alleles previously observed by CE for which no sequence data has been published (Tables 2.1 and 2.2). A further classification was made for a repeat pattern match (PM), which represented a total of 59 sequences that matched the repeat pattern structures previously published for the respective allele. In Table 2.1, SH allele sequences were reported for loci DYS449 (n = 11); DYS447 (n=11); DYS710 (n= 6); DYS518 (n= 4) and DYS644 (n= 5). Newly observed alleles sequences were reported for loci DYS644 (n=18), DYS710 (n=1), DYS447 (n=1) and DYS504 (n=3). A total of 34 sequences were reported for alleles previously observed with CE for which no sequence data was available (Table 2.1).

3.3.2 Size homoplasy

Size homoplasy (SH) at Y-STRs may occur when alleles which are identical by state (i.e. same length) present sequence pattern variation due to replication slippage or point mutations in the STR repeat region or flanking region (Garza and Friemer 1996 and Warshauer et al. 2015). This format of variation also referred to by others as Repeat Pattern Variants (RPV) is considered for distinguishing alleles shared between male haplotypes which is often encountered during Y-STR analysis by CE detection (Ballantyne et al. 2010 and Warshauer et al. 2015). In Table 2.2, SH was detected between populations within this report and compared to others for Locus DYS449 alleles 25, 28, 29, 30 and 35 (Ruitberg et al. 2001; Redd et al. 2002; D'Amato et al.2010 and Warshauer et al. 2015) ; DYS447 alleles 19, 21, 22, 24, 25, 26 and 29 (Ruitberg et al. 2001; Redd et al. 2002; Rodig et al. 2007; D'Amato et al. 2009 and Park et al. 2012,). SH was observed at the DYS710 alleles 28.2, 30.2, 32, 33 and 35 (D'Amato et al. 2010); DYS644 alleles 24.4 and 26.4 (D'Amato et al. 2010) and at DYS518 for allele 39 and 44 (D'Amato et al. 2010 and Warshauer et al. 2015).

SH was also observed in comparison to D'Amato et al. 2010 and in this report for the same population group at locus DYS449 alleles 27, 28 and 29; at DYS644 allele 20.4, 21.4 and 22.4 and at DYS518 allele 34. Comparing the SH events for all available sequence data published for each locus, we identify up to 3-4 variable structure patterns for loci DYS449, DYS447 and DYS710 (Supplementary 2 Table 1). Alleles presenting 4 variable structure patterns were observed at locus DYS449 for allele 28 and DYS447 for allele 24 and 26 (Ruitberg et al. 2001; Redd et al. 2002; Rodig et al. 20027; D'Amato et al. 2009). A total of 3 alternative repeat arrangements were found for alleles 29 and 35 at DYS449 (Ruitberg et al. 2001; Redd et al. 2002 and D'Amato et al. 2010), alleles 32 and 35 at DYS710 D'Amato et al. 2010 and allele 25 at DYS447 (Ruitberg et al. 2001; Redd et al. 2002). This variability in the pattern arrangement was detected respectively within and between population group (Supplementary 2 Table 1). We observed that locus DYS710 with a di-tetra complex structure and loci DYS449 and DYS447 which have interrupted tetra and penta nucleotide repeats respectively were more informative for sequence discrimination by identifying SH. This observation may be attributed to the interrupted repeats of loci DYS447 and DYS449 which is known to increase the probability for replication slippage occurring between its variable regions (Ballantyne et al. 2010). For DYS710 mutability is likely attributed by having a total of 3 variable regions and two variable blocks of [AAAG] asymmetric purine and pyrimidine repeats which is known to have a positive correlation with locus mutability (Ballantyne et al. 2010). The loci DYS481, DYS626 and DYS612 with its perfect repeat structure did not allow for the means to detect SH but may instead aid sequence discrimination by the presence of point mutations within the variable block region or rare events of SH as recently shown for DYS481 due to a SNPs in the flanking region (Warshauer et al. 2015, Kwon et al. 2016, Hutsar et al. 2018). In this report we encounter 2 intra repeat SNPs, one for allele 32 at DYS481 and 35 at DYS449, for which both were defined by a T/C transition represented as and $[\mathrm{CTT}]_2[\mathrm{C\underline{C}T}]_1[\mathrm{CTT}]_{29}$ and $[\mathrm{TTTC}]_5[\mathrm{T\underline{C}TC}]_1$ $[\mathrm{TTTC}]_{12}\mathrm{N50}[\mathrm{TTTC}]_{18}$ respectively .

We observed a rare sequence variant for allele 21 at DYS385 containing 8 tandem [AAGG] repeats in the invariant 5’ flanking region located upstream to the first DYS385 [GAAA] site. Its structure consequently presents size homoplasy with the consensus $[\mathrm{GAAA}]_{21}$ allele sequences which only contain 6 [AAGG] repeats in the flanking region (Kayser et al. 1997 and Gusmão et al. 2006). Similar variant structures were reported in (Gusmão et al. 2006 Novroski et al. 2016) and for allele 21 more recently by Huszar et al.2018, who proposed that these observed variants be represented as $[\mathrm{AAGG}]_{5\text{-}9}[\mathbf{GAAA}]_n$.

3.3.3 Novel Allele Sequences

The occurrence of a single base deletion at DYS644 presented sequence structures which deviated from the traditional motif arrangement. A series of 14 novel off ladder DYS644 alleles were observed adjacent to the bins for micro variant alleles 22.4, 23.4, 24.4, and 25.4 when using CE detection (Supplementary 2 Figure 1). Subsequent to their sequencing, we identified a 1bp deletion in position “A” of the first repeat unit [TTTT$\underline{\mathrm{A}}$] for all 14 variants. These variant alleles represented by the repeat motif $[\mathrm{TTTT}\mathit{del}]_{0\text{-}1}$**[TTTTA)** $_n$TTTA[**TTTTA**]$_m$ (Table 2.1 and 2.2) were subsequently scored as alleles 22.3, 23.3, 24.3, and 25.3 respectively. The deletion event was identified in the Bantu groups Venda and Pedi, in the Khoe group Nama and the admixed Coloured populations, all from the NE and N South Africa. The occurrence of these microvariants provided added value to DYS644, which in addition to its value for inferring ancestry (D'Amato et al. 2010) may also potentially allow for bio-geographical inferences. Sequences for alleles detected outside the known allele range are reported for DYS710 allele 26 and DYS504 alleles 9 and 20. Sequencing of alleles presented as off ladder (OL) also confirmed novel microvariants at DYS644 alleles 19.4 and 27.4 and DYS447 allele 24.4 (Table 2.2).

3.3.4 Updates to the repeat structure variable range

Alleles from this report and previously published were compiled to provide a summarized update to the loci variable region repeat range (Table 2.1). For DYS612 the previously reported repeat unit variable range 19-31 (bold) has been previously expressed as [CCT]$_5$[CTT]$_1$[TCT]$_4$[CCT]$_1$[**TCT**]$_{19-31}$ (Ballantyne et al. 2010 and 2012). This range was updated to 14-31 as indicated by an asterisk (*) in the motif [CCT]$_5$[CTT]$_1$[TCT]$_4$[CCT]$_1$ [**TCT**]$_{14*-31}$ in Table 2.1. Similarly, the repeat block range for DYS626 was updated from [**GAAA**]$_{14-23}$ (Kayser et al. 2004 and Ballantyne et al. 2012) to [**GAAA**]$_{11*-23}$ considering sequences for alleles 24, 25 and 26 in Table 2.2. The update for DYS518 from the range previously reported by (Ballantyne et al. 2010) was assembled as [AAAG]$_3$[GAAG]$_1$ [**AAAG**]$_{10*,13*-22}$[GGAG]$_1$[AAAG]$_4$N$_6$[**AAAG**]$_{11-19}$N$_{27}$**[**AAGG]$_4$ and identified that repeats 11 and 12 in the first [AAAG] variable block was not observed to date (Kayser et al. 2004, Ballantyne et al. 2010 and D'Amato et al. 2010). For DYS447 the first and second variable blocks were expanded by 4-5 repeats and 6-7 repeats respectively. At DYS447 truncated repeat patterns were observed for alleles 17, 19 and 21, which were also previously reported for allele 19 in a Korean population group (Park et al. 2012) and for allele 24 in an English (UK) population (Redd et al. 2002).

These structures which were missing the DYS447 2nd non-variable block [TAAAA] and 3rd variable block [TAATA] were annotated by 0 in the updated structure. The DYS447 was updated to [**TAATA]** $_{4*-21}$[TAAAA] $_{1}$**[TAATA]** $_{6*-13}$[TAAAA] $_{0-1}$**[TAATA]** $_{0, 5-9}$ from that previously observed (Redd et al. 2002, Parkin et al. 2006 and Park et al. 2012). For DYS644, the updated structure from Kayser et.al. 2004 can be represented as [TTTT*del*]$_{0-1}$[**TTTTA**] $_{10*-19*}$ [TTTA]$_{0-1}$[**TTTTA**]$_{0, 9*-15*}$ considering the novel alleles, the deletion event and micro variants we report (Table 2.1). The sequence structure of DYS710 is described as a di-tetra complex with two variable tetra nucleotide blocks separated by a variable block of dinucleotide repeats. Provided with the DYS710 sequences in Table 2.2 and those previously published by Leat et al. 2007 and D’Amato et al. 2010) the observed repeat counts in each variable block was assembled as [**AAAG**]$_{12*-20*}$[**AG**]$_{9*-20*}$[**AAAG**]$_{8*-14*}$.

3.4 Conclusion

In this study, we described the sequence structure of novel alleles and sequence variation encountered for 153 sequences across the 10-Y STR markers DYS710, DYS518, DYS385, DYS644, DYS612, DYS626, DYS504, DYS481, DYS447 and DYS449. For the total of 94 novel sequences, 36% were for previously observed alleles for which sequence data was not available, 39% for sequences structures which presented size homoplasy (SH) and 24% for newly observe alleles or variants detected by CE. The variation attributed by SH was evidently more informative than the occurrences of SNPs for distinguishing isometric alleles. From the 10-Y-STR panel, DYS710, DYS449 and DYS447 contributed 76% of the total SH events and given its sequence structure complexity and nucleotide compositions may provide a more informative means to resolve shared male haplotypes. Africa is known to harbour a high degree of human diversity (Tishkoff et al 2002) and evidently host to unique genetic variation for both autosomal (Ristow et al. 2016) and Y-chromosome STRs (Warshauer et al. 2015). The majority of novel alleles and variation we encountered were all associated with individuals of African paternal ancestry which in many cases originated exclusively from the rarely studied population in South Africa. The degree of sequence variation we identified from our relatively small sample size encourages the need for large scale variant characterizations which may assist the individualization of the male haplotype.

Table 2.1: Summary of categorised sequences representing updates to the loci repeat structure and variable block repeat range.

Size Homoplasy (SH), Known Alleles (KA) and Novel Allele Sequence (NAS) represents the sequence categories. (*) Updates to the allele range or repeat range provided in this study. The variable region for each variable block(s) is shown in bold. [a] Updates proposed to the repeat structures considering variation reported for DYS481 (Warshauer et al. 2015 and Hutszar et al. 2018), DYS385 (Hutszar et al. 2018) and DYS644 (this study). For visual purpose references to the repeat structure are represented as follows [1] Leat eat al. 2007; [2] Ballantyne et al. 2012; [3] Kayser et al. 1997; [4] Kayser et al. 2004; [5] Lim et al. 2007; [6] Redd et al. 2002.

Loci	**Repeat type**	**Reference repeat structure**	**Ref**	**Updates to repeat structure from (this study) and from [9,10]**	**Known allele range**	**(SH)**	**(KA)**	**(NAS)**
DYS710	Di-Tetra, complex	[**AAAG**]$_n$ [**AG**]$_n$ [**AAAG**]$_n$	[1]	[**AAAG**]$_{12*-20*}$ [**AG**]$_{9*-20*}$ [**AAAG**]$_{8*-14*}$	***26-42**	6	7	1
DYS518	Tetra complex	[AAAG]$_3$[GAAG]$_1$[**AAAG**]$_n$[GGAG]$_1$[AAAG]$_4$N$_6$[**AAAG]**$_n$N$_{27}$[AAGG]$_4$	[2]	[AAAG]$_3$ [GAAG]$_1$ [**AAAG**]$_{10*,13*–22}$ [GGAG]$_1$[AAAG]$_4$N$_6$[**AAAG**]$_{11–19}$N$_{27}$[AAGG]$_4$	**23-35**	4	**3**	0
DYS385	Multicopy complex	[**GAAA**]$_n$	[3]	[a][AAAG]$_{6-8}$[**GAAA**] $_{6-28}$	**6-28**	0	0	0
DYS644	Penta complex	[**TTTTA**]$_\mathbf{n}$ TTTA [**TTTTA**]$_\mathbf{n}$	[4]	[a][TTTT*del*]$_{0-1}$[**TTTTA**]$_{12*-19*}$ [TTTA]$_{0-1}$ [**TTTTA**]$_{0, 9*-15*}$	**12-27.4***	5	3	18
DYS612	Tri complex	[CCT]$_5$[CTT]$_1$[TCT]$_4$[CCT]$_1$[**TCT**]$_n$	[2]	[CCT]$_5$[CTT]$_1$[TCT]4[CCT]$_1$[**TCT**]$_{14*–31}$	**14-31**	0	4	0
DYS626	Tetra complex	[**GAAA**]$_n$N$_{24}$[GAAA]$_3$N$_6$[GAAA]$_5$ [AAA]$_1$[**GAAA**]$_{2–3}$ [GAAG]$_1$(GAAA)$_3$	[2]	[**GAAA**]$_{11*–23}$N$_{24}$ [GAAA]$_3$N$_6$ [GAAA]$_5$ [AAA]1 [**GAAA**]$_{2–3}$ [GAAG]$_1$ [GAAA]$_3$	**11-23**	0	3	0
DYS504	Tetra simple	[**TCCT**]$_n$N$_7$[CCCT]$_3$	[4]	[**TCCT**]$_{11-19*}$N$_7$ [CCCT]$_3$	***9-20***	0	4	3
DYS481	Tri simple	[**CTT**]$_n$	[5]	[a][CTG]$_{0-2}$[**CTT**]$_{16-32*}$	**16-34**	0	5	0
DYS447	Penta complex	[**TAATA**]$_n$[TAAAA]$_1$[**TAATA**]$_n$[TAAAA]$_1$ [**TAATA**]$_n$	[6]	[**TAATA**]$_{4*,8-21}$[TAAAA]$_1$[**TAATA**]$_{6*-13}$[TAAAA]$_{0*-1}$[**TAATA**]$_{0*,5-9}$	**15-36**	11	3	1
DYS449	Tetra complex	[**TTTC**]$_n$N$_{50}$[**TTTC**]$_n$	[6]	[**TTTC**]$_{11–19}$ N$_{50}$ [**TTTC**]$_{12–19}$	**22-42**	11	2	0
Total						37	34	23

Table 2.2: Summary of repeat structures for all 153 categorised sequences.

Known alleles (KA); Size Hompoplasy (SH); Pattern Match (PM) and Novel Allele Sequence (NAS) represents allele reporting categories for sequences ^{a}Population group: E (European English); I (Asian Indian); C (Coloured); AF (Afrikaner); X (Xhosa); Z (Zulu); P (Pedi); V (Venda); N (Ndebele) and S (Sotho). For visual purpose the references to sequence comparison studies are as follows: [6] Redd et al. 2002; [7] D'Amato et al. 2010; [8] Warshauer et al. 2015; [9] Ruitberg et al. 2001; [10] Huszar et al. 2018; [11] Novroski et al. 2016;[12] D'Amato et al. 2009; [13] Kwon et al. 2016; [14] Park et al. 2012; [15] Rodig et al. 2007.

Locus	^{b}Population group	Genbank Accession Number	Allele	Repeat structure	KA	SH	PM	NAS	Sequence comparison
DYS710									
	C	MK005372	26	[AAAG]$_{12}$[AG]$_{12}$[AAAG]$_{8}$				1	
	C	MK005373	27	[AAAG]$_{12}$[AG]$_{12}$[AAAG]$_{9}$		1			
	C	MK005374	28,2	[AAAG]$_{14}$[AG]$_{9}$[AAAG]$_{10}$		1			[7]
	I	MK005375	30,2	[AAAG]$_{15}$[AG]$_{11}$[AAAG]$_{10}$		1			[7]
	AF	MK005376	32	[AAAG]$_{13}$[AG]$_{12}$[AAAG]$_{13}$		1			[7]
	C	MK005377	33	[AAAG]$_{13}$[AG]$_{12}$[AAAG]$_{14}$		1			[7]
	I	MK005378	35	[AAAG]$_{17}$[AG]$_{12}$[AAAG]$_{11}$		1			[7]
	I	MK005379	37,2	[AAAG]$_{16}$[AG]$_{15}$[AAAG]$_{14}$			1		[7]
	C	MK005381	38	[AAAG]$_{19}$[AG]$_{14}$[AAAG]$_{12}$	1				
	I	MK005380	38,2	[AAAG]$_{20}$[AG]$_{9}$[AAAG]$_{14}$	1				
	C	MK005382	39	[AAAG]$_{18}$[AG]$_{16}$[AAAG]$_{13}$	1				
	X	MK005383	40	[AAAG]$_{20}$[AG]$_{18}$[AAAG]$_{11}$	1				
	C	MK005384	41	[AAAG]$_{19}$[AG]$_{20}$[AAAG]$_{12}$	1				
	I	MK005385	41,2	[AAAG]$_{18}$[AG]$_{19}$[AAAG]$_{14}$	1				
	E	MK005386	42	[AAAG]$_{20}$[AG]$_{18}$ [AAAG]$_{13}$	1				
DYS518									
	S	MK005441	33	[AAAG]$_{3}$[GAAG]$_{1}$[AAAG]$_{11}$[GGAG]$_{1}$[AAAG]$_{4}$N6[AAAG]$_{13}$	1				
	C	MK005442	34	[AAAG]$_{3}$[GAAG]$_{1}$[AAAG]$_{13}$[GGAG]$_{1}$[AAAG]$_{4}$N6[AAAG]$_{12}$		1			
	C	MK005443	34	[AAAG]$_{3}$[GAAG]$_{1}$[AAAG]$_{10}$[GGAG]$_{1}$[AAAG]$_{4}$N6[AAAG]$_{15}$		1			
	I	MK005444	38	[AAAG]$_{3}$[GAAG]$_{1}$[AAAG]$_{15}$[GGAG]$_{1}$[AAAG]$_{4}$N6[AAAG]$_{14}$			1		[8]
	C	MK005448	39	[AAAG]$_{3}$[GAAG]$_{1}$[AAAG]$_{17}$[GGAG]$_{1}$[AAAG]$_{4}$N6[AAAG]$_{13}$		1			[8]
	AF	MK005449	41	[AAAG]$_{3}$[GAAG]$_{1}$[AAAG]$_{16}$[GGAG]$_{1}$[AAAG]$_{4}$N6[AAAG]$_{16}$			1		[8]
	C	MK005445	44	[AAAG]$_{3}$[GAAG]$_{1}$[AAAG]$_{20}$[GGAG]$_{1}$[AAAG]$_{4}$N6[AAAG]$_{15}$		1			[7]
	C	MK005446	45	[AAAG]$_{3}$[GAAG]$_{1}$[AAAG]$_{20}$[GGAG]$_{1}$[AAAG]$_{4}$N6[AAAG]$_{16}$	1				[7]

	Z	MK005447	46	$[AAAG]_{3}[GAAG]_{1}[AAAG]_{20}[GGAG]_{1}[AAAG]_{4}N6[AAAG]_{17}$	1				
DYS385									
	I	MK005508	7	$[GAAA]_{7}$			1		[9]
	C	MK005509	8	$[GAAA]_{8}$			1		[9]
	I	MK005510	9	$[GAAA]_{9}$			1		[9]
	E	MK005511	10	$[GAAA]_{10}$			1		[9]
	I	MK005512	11	$[GAAA]_{11}$			1		[9]
	C	MK005513	12	$[GAAA]_{12}$			1		[9]
	I	MK005514	13	$[GAAA]_{13}$			1		[9]
	I	MK005515	14	$[GAAA]_{14}$			1		[9]
	I	MK005516	15	$[GAAA]_{15}$			1		[9]
	I	MK005517	16	$[GAAA]_{16}$			1		[9]
	I	MK005518	17	$[GAAA]_{17}$			1		[9]
	Z	MK005519	17	$[GAAA]_{17}$			1		[9]
	X	MK005520	17	$[GAAA]_{17}$			1		[9]
	C	MK005521	18	$[GAAA]_{18}$			1		[9]
	C	MK005522	19	$[GAAA]_{19}$			1		[9]
	I	MK005523	20	$[GAAA]_{20}$			1		[9]
	C	MK005524	21	$[AAGG]_{8}[GAAA]_{13}$			1		[10]
DYS644									
	C	MK005387	10	$[TTTTA]_{10}$				1	
	C	MK005388	10	$[TTTTA]_{10}$				1	
	E	MK005389	12	$[TTTTA]_{12}$	1				
	I	MK005390	13	$[TTTTA]_{13}$			1		[7]
	E	MK005391	14	$[TTTTA]_{14}$	1				
	I	MK005392	15	$[TTTTA]_{15}$	1				
	I	MK005393	16	$[TTTTA]_{16}$			1		[7]
	I	MK005394	17	$[TTTTA]_{17}$			1		[7]
	X	MK005395	19.4	$[TTTTA]_{10}TTTA[TTTTA]_{9}$				1	
	X	MK005396	20.4	$[TTTTA]_{10}TTTA\ [TTTTA]_{10}$		1			[7]
	X	MK005397	21.4	$[TTTTA]_{11}TTTA\ [TTTTA]_{10}$		1			[7]
	X	MK005398	22.4	$[TTTTA]_{11}TTTA\ [TTTTA]_{11}$		1			[7]
	C	MK005399	24.4	$[TTTTA]_{12}TTTA\ [TTTTA]_{12}$		1			[7]
	Z	MK005400	25.4	$[TTTTA]_{11}TTTA\ [TTTTA]_{14}$			1		
	X	MK005401	25.4	$[TTTTA]_{11}TTTA[TTTTA]_{14}$			1		

	Z	MK005402	26.4	[TTTTA]$_{12}$TTTA[TTTTA]$_{14}$		1			[7]
	P	MK005403	27.4	[TTTTA]$_{12}$TTTA[TTTTA]$_{15}$				1	
	C	MK005405	22.3	[TTTT*del*]$_{1}$[TTTTA]$_{10}$TTTA[TTTTA]$_{11}$				1	
	C	MK005404	23.3	[TTTT*del*]$_{1}$[TTTTA]$_{11}$TTTA[TTTTA]$_{11}$				1	
	P	MK005416	23.3	[TTTT*del*]$_{1}$[TTTTA]$_{11}$TTTA[TTTTA]$_{11}$				1	
	C	MK005408	24.3	[TTTT*del*]$_{1}$[TTTTA]$_{11}$TTTA[TTTTA]$_{12}$				1	
	C	MK005406	24.3	[TTTT*del*]$_{1}$[TTTTA]$_{11}$TTTA[TTTTA]$_{12}$				1	
	C	MK005410	24.3	[TTTT*del*]$_{1}$[TTTTA]$_{11}$TTTA[TTTTA]$_{12}$				1	
	C	MK005407	24.3	[TTTT*del*]$_{1}$[TTTTA]$_{11}$TTTA[TTTTA]$_{12}$				1	
	C	MK005409	24.3	[TTTT*del*]$_{1}$[TTTTA]$_{11}$TTTA[TTTTA]$_{12}$				1	
	V	MK005411	24.3	[TTTT*del*]$_{1}$[TTTTA]$_{11}$TTTA[TTTTA]$_{12}$				1	
	P	MK005412	24.3	[TTTT*del*]$_{1}$[TTTTA]$_{11}$TTTA[TTTTA]$_{12}$				1	
	P	MK005414	24.3	[TTTT*del*]$_{1}$[TTTTA]$_{11}$TTTA[TTTTA]$_{12}$				1	
	P	MK005415	24.3	[TTTT*del*]$_{1}$[TTTTA]$_{11}$TTTA[TTTTA]$_{12}$				1	
	P	MK005413	24.3	[TTTT*del*]$_{1}$[TTTTA]$_{11}$TTTA[TTTTA]$_{12}$				1	
	C	MK005417	25.3	[TTTT*del*]$_{1}$[TTTTA]$_{12}$TTTA[TTTTA]$_{12}$				1	
DYS612									
	C	MK005429	25	[CCT]$_{5}$[CTT]$_{1}$[TCT]$_{4}$[CCT]$_{1}$[TCT]$_{14}$	1				
	C	MK005430	27	[CCT]$_{5}$[CTT]$_{1}$[TCT]$_{4}$[CCT]$_{1}$[TCT]$_{16}$			1		[7]
	C	MK005431	29	[CCT]$_{5}$[CTT]$_{1}$[TCT]$_{4}$[CCT]$_{1}$[TCT]$_{18}$	1				
	C	MK005432	30	[CCT]$_{5}$[CTT]$_{1}$[TCT]$_{4}$[CCT]$_{1}$[TCT]$_{19}$	1				
	C	MK005433	31	[CCT]$_{5}$[CTT]$_{1}$[TCT]$_{4}$[CCT]$_{1}$[TCT]$_{20}$			1		[7]
	I	MK005434	32	[CCT]$_{5}$[CTT]$_{1}$[TCT]$_{4}$[CCT]$_{1}$[TCT]$_{21}$			1		[11]
	E	MK005435	35	[CCT]$_{5}$[CTT]$_{1}$[TCT]$_{4}$[CCT]$_{1}$[TCT]$_{24}$			1		[7],[11]
	I	MK005436	36	[CCT]$_{5}$[CTT]$_{1}$[TCT]$_{4}$[CCT]$_{1}$[TCT]$_{25}$			1		[7],[11]
	I	MK005437	37	[CCT]$_{5}$[CTT]$_{1}$[TCT]$_{4}$[CCT]$_{1}$[TCT]$_{26}$			1		[7],[11]
	C	MK005438	37	[CCT]$_{5}$[CTT]$_{1}$[TCT]$_{4}$[CCT]$_{1}$[TCT]$_{26}$			1		[7],[11]
	C	MK005439	38	[CCT]$_{5}$[CTT]$_{1}$[TCT]$_{4}$[CCT]$_{1}$[TCT]$_{27}$			1		[7],[11]
	I	MK005440	39	[CCT]$_{5}$[CTT]$_{1}$[TCT]$_{4}$[CCT]$_{1}$[TCT]$_{28}$			1		[7],[11]
DYS626									
	AF	MK005418	24	[GAAA]$_{11}$N24[GAAA]$_{3}$N$_{6}$[GAAA]$_{5}$[AAA]$_{1}$[GAAA]$_{2}$ [GAAG]$_{1}$[GAAA]3	1				
	AF	MK005419	25	[GAAA]$_{12}$N24[GAAA]$_{3}$N$_{6}$[GAAA]$_{5}$ [AAA]$_{1}$[GAAA]$_{2}$ [GAAG]$_{1}$[GAAA]$_{3}$	1				

	C	MK005420	25	[GAAA]$_{12}$N24[GAAA]$_{3}$N$_{6}$[GAAA]$_{5}$[AAA]$_{1}$[GAAA]$_{2}$ [GAAG]$_{1}$[GAAA]$_{3}$	1				
	AF	MK005421	26	[GAAA]$_{13}$N24[GAAA]$_{3}$N$_{6}$[GAAA]$_{5}$[AAA]$_{1}$[GAAA]$_{2}$ [GAAG]$_{1}$[GAAA]$_{3}$	1				
	C	MK005422	27	[GAAA]$_{14}$N24[GAAA]$_{3}$N$_{6}$[GAAA]$_{5}$[AAA]$_{1}$[GAAA]$_{2}$ [GAAG]$_{1}$[GAAA]$_{3}$			1		[7]
	C	MK005423	28	[GAAA]$_{15}$N24[GAAA]$_{3}$N$_{6}$[GAAA]$_{5}$[AAA]$_{1}$[GAAA]$_{2}$ [GAAG]$_{1}$[GAAA]$_{3}$			1		[7]
	I	MK005424	29	[GAAA]$_{16}$N24[GAAA]$_{3}$N$_{6}$[GAAA]$_{5}$[AAA]$_{1}$[GAAA]$_{2}$ [GAAG]$_{1}$[GAAA]3	1				
	I	MK005425	30	[GAAA]$_{17}$N24[GAAA]$_{3}$N$_{6}$[GAAA]$_{5}$[AAA]$_{1}$[GAAA]$_{2}$ [GAAG]$_{1}$[GAAA]$_{3}$	1				
	I	MK005426	31	[GAAA]$_{18}$N24[GAAA]$_{3}$N$_{6}$ [GAAA]$_{5}$[AAA]$_{1}$[GAAA]$_{2}$ [GAAG]$_{1}$[GAAA]$_{3}$	1				
	AF	MK005427	33	[GAAA]$_{20}$N24[GAAA]$_{3}$N$_{6}$[GAAA]$_{5}$ [AAA]$_{1}$[GAAA]$_{2}$ [GAAG]$_{1}$[GAAA]$_{3}$	1				
	I	MK005428	35	[GAAA]$_{22}$N24[GAAA]$_{3}$N$_{6}$[GAAA]$_{5}$[AAA]$_{1}$ [GAAA]$_{2}$ [GAAG]$_{1}$[GAAA]$_{3}$	1				
DYS504									
	S	MK005450	9	[TCCT]$_{9}$				1	
	C	MK005451	12	[TCCT]$_{12}$	1				
	AF	MK005452	13	[TCCT]$_{13}$			1		[12]
	AF	MK005453	14	[TCCT]$_{14}$			1		[12]
	C	MK005454	15	[TCCT]$_{15}$			1		[12]
	AF	MK005457	16	[TCCT]$_{16}$	1				
	AF	MK005456	17	[TCCT]$_{17}$	1				
	E	MK005455	18	[TCCT]$_{18}$			1		[9]
	N	MK005458	19	[TCCT]$_{19}$	1				
	C	MK005459	20	[TCCT]$_{20}$				1	
	C	MK005460	20	[TCCT]$_{20}$				1	
DYS481									
	C	MK005461	22	[CTT]$_{22}$			1		[13],[11]
	I	MK005462	23	[CTT]$_{23}$			1		[13],[11]
	C	MK005463	24	[CTT]$_{24}$			1		[13],[11]
	I	MK005464	25	[CTT]$_{25}$			1		[13],[11]

	I	MK005465	26	[CTT]$_{26}$			1		[13],[11]
	C	MK005466	27	[CTT]$_{27}$			1		[7],[13],[11]
	X	MK005467	28	[CTT]$_{28}$			1		[13],[11]
	C	MK005468	29	[CTT]$_{29}$			1		[13],[11]
	C	MK005469	30	[CTT]$_{30}$			1		[11]
	C	MK005470	30	[CTT]$_{30}$			1		[11]
	V	MK005471	31	[CTT]$_{31}$			1		[11]
	C	MK005472	32	[CTT]$_{2}$[CTT]$_{30}$			1		[11]
	P	MK005473	32	[CTT]$_{32}$			1		[11]
DYS447									
	X	MK005491	17	[TAATA]$_{8}$[TAAAA]$_{1}$[TAATA]$_{8}$	1				
	I	MK005492	19	[TAATA]$_{12}$[TAAAA]$_{1}$[TAATA]$_{6}$		1			[14]
	I	MK005493	20	[TAATA]$_{4}$[TAAAA]$_{1}$[TAATA]$_{8}$[TAAAA]$_{1}$[TAATA]$_{6}$			1		[15],[12]
	I	MK005494	21	[TAATA]$_{7}$[TAAAA]$_{1}$[TAATA]$_{6}$[TAAAA]$_{1}$[TAATA]$_{6}$		1			[15]
	C	MK005495	21	[TAATA]$_{13}$[TAAAA]$_{1}$[TAATA]$_{7}$		1			[15]
	I	MK005496	22	[TAATA]$_{6}$[TAAAA]$_{1}$[TAATA]$_{8}$[TAAAA]$_{1}$[TAATA]$_{6}$		1			[15]
	C	MK005497	22	[TAATA]$_{5}$[TAAAA]$_{1}$[TAATA]$_{9}$[TAAAA]$_{1}$[TAATA]$_{6}$		1			[15]
	I	MK005498	22.4	[TAATA]$_{7}$[TAAAA]$_{1}$[TAATA]$_{7.4}$[TAAAA]$_{1}$[TAATA]$_{6}$			1		[7]
	I	MK005499	24	[TAATA]$_{7}$[TAAAA]$_{1}$[TAATA]$_{8}$[TAAAA]$_{1}$[TAATA]$_{7}$		1			[6], [15], [12],[9]
	C	MK005500	24.4	[TAATA]$_{7}$[TAAAA]1[TAATA]$_{7.4}$[TAAAA]$_{1}$[TAATA]$_{8}$				1	
	I	MK005501	25	[TAATA]$_{7}$[TAAAA]$_{1}$[TAATA]$_{9}$[TAAAA]$_{1}$[TAATA]$_{7}$		1			[6], [9]
	I	MK005502	26	[TAATA]$_{9}$[TAAAA]$_{1}$[TAATA]$_{8}$[TAAAA]$_{1}$[TAATA]$_{7}$		1			[6], [9]
	C	MK005503	26	[TAATA]$_{7}$[TAAAA]$_{1}$[TAATA]$_{9}$[TAAAA]$_{1}$[TAATA]$_{8}$		1			[6],[15],[9]
	I	MK005504	28	[TAATA]$_{7}$[TAAAA]$_{1}$[TAATA]$_{12}$[TAAAA]$_{1}$[TAATA]$_{7}$	1				
	I	MK005505	29	[TAATA]$_{7}$[TAAAA]$_{1}$[TAATA]$_{13}$[TAAAA]$_{1}$[TAATA]$_{7}$		1			
	C	MK005506	29	[TAATA]$_{7}$[TAAAA]$_{1}$[TAATA]$_{12}$[TAAAA]$_{1}$[TAATA]$_{8}$		1			
	X	MK005507	30	[TAATA]$_{9}$[TAAAA]$_{1}$[TAATA]$_{11}$[TAAAA]$_{1}$[TAATA]$_{8}$	1				
DYS449									
	I	MK005474	24	[TTTC]$_{11}$N$_{50}$[TTTC]$_{13}$	1				
	I	MK005476	25	[TTTC]$_{13}$N$_{50}$[TTTC]$_{12}$		1			[8]
	I	MK005477	26	[TTTC]$_{11}$ N$_{50}$[TTTC]$_{15}$			1		[7]
	I	MK005475	27	[TTTC]$_{11}$ N$_{50}$[TTTC]$_{15}$		1			[7], [9]
	I	MK005478	28	[TTTC]$_{15}$ N$_{50}$[TTTC]$_{13}$		1			[6],[9]
	I	MK005479	28	[TTTC]$_{11}$ N$_{50}$[TTTC]$_{17}$		1			[6],[9]
	I	MK005480	29	[TTTC]$_{12}$ N$_{50}$[TTTC]$_{17}$		1			[6], [7], [9]

	C	MK005481	30	[TTTC]$_{14}$ N$_{50}$[TTTC]$_{16}$		1			[9]
	I	MK005482	30	[TTTC]$_{14}$ N$_{50}$ [TTTC]$_{16}$		1			[9]
	I	MK005483	32	[TTTC]$_{17}$ N$_{50}$ [TTTC]$_{15}$			1		[7], [9]
	I	MK005484	32	[TTTC]$_{17}$ N$_{50}$ [TTTC]$_{15}$			1		[7],[9]
	I	MK005485	33	[TTTC]$_{16}$ N$_{50}$ [TTTC]$_{17}$			1		[7]
	I	MK005486	35	[TTTC]$_{18}$ N$_{50}$ [TTTC]$_{17}$		1			[7]
	AF	MK005487	35	[TTTC]$_{5}$[T$\underline{\text{C}}$TC]$_{1}$[TTTC]$_{12}$N$_{50}$[TTTC]$_{18}$		1			[7]
	N	MK005488	37	[TTTC]$_{17}$ N$_{50}$ [TTTC]$_{20}$		1			
	S	MK005489	37	[TTTC]$_{16}$ N$_{50}$ [TTTC]$_{21}$		1			
	S	MK005490	38	[TTTC]$_{16}$ N$_{50}$ [TTTC]$_{22}$	1				

3.5 References

Ballantyne KN, Goedbloed M, Fang R, et al (2010) Mutability of Y-Chromosomal Microsatellites: Rates , Characteristics , Molecular Bases , and Forensic Implications. Am J Hum Genet 87:341–353. https://doi.org/10.1016/j.ajhg.2010.08.00

Ballantyne KN, Keerl V, Wollstein A, et al (2012) A new future of forensic Y-chromosome analysis: Rapidly mutating Y-STRs for differentiating male relatives and paternal lineages. Forensic Sci Int Genet 6:208–218. https://doi.org/10.1016/j.fsigen.2011.04.017

Burrows AM, Ristow PG, Amato MED (2017) Preservation of DNA from saliva samples in suboptimal conditions. Forensic Sci Int Genet Suppl Ser 6: e80–e81. https://doi.org/10.1016/j.fsigss.2017.09.050

Cloete KW, Ristow PG, Kasu M, D’Amato ME (2016) Design, installation, and performance evaluation of a custom dye matrix standard for automated capillary electrophoresis. Electrophoresis 38:617–623. https://doi.org/10.1002/elps.201600257

D’Amato ME, Bajic VB, Davison S (2011) Design and validation of a highly discriminatory 10-locus Y-chromosome STR multiplex system. Forensic Sci Int Genet 5:122–125. https://doi.org/10.1016/j.fsigen.2010.08.015

D’Amato ME, Benjeddou M, Davison S (2009) Evaluation of 21 Y-STRs for population and forensic studies. Forensic Sci Int Genet Suppl Ser 2:446–447. https://doi.org/10.1016/j.fsigss.2009.08.091

D’Amato ME, Ehrenreich L, Cloete K, et al (2010) Characterization of the highly discriminatory loci DYS449, DYS481, DYS518, DYS612, DYS626, DYS644 and DYS710. Forensic Sci Int Genet 4:104-10. https://doi.org/10.1016/j.fsigen.2009.06.011

D’Amato ME, Kasu M (2017) Population analysis of African Y-STR profiles with UniQ TYPER™ Y-10 genotyping system. Forensic Sci Int Genet Suppl Ser 6: e84 – e85. https://doi.org/10.1016/j.fsigss.2017.09.048

Garza JC, Freimer NB (1996) Homoplasy for Size at Microsatellite Loci in Humans and Chimpanzees. 211–217. https://doi.org/10.1101/gr.6.3.211

Gusmão L, Butler JM, Carracedo A, et al (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 Sci Int 157:187–197. https://doi.org/10.1016/j.forsciint.2005.04.002

Huszar TI, Jobling MA, Wetton JH (2018) A phylogenetic framework facilitates Y-STR variant discovery and classification via massively parallel sequencing. Forensic Sci Int Genet 35:97–106. https://doi.org/10.1016/j.fsigen.2018.03.012

Kayser M (2017) Forensic use of Y-chromosome DNA : a general overview. Hum Genet 136:621–635. https://doi.org/10.1007/s00439-017-1776-91

Kayser M, Caglià A, Corach D, et al (1997) Evaluation of Y-chromosomal STRs: A multicenter study. Int J Legal Med 110:125–133. https://doi.org/10.1007/s004140050051

Kayser M, Kittler R, Erler A, et al (2004) A Comprehensive Survey of Human Y-Chromosomal Microsatellites. Am J Hum Genet 74:1183–1197. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1182082/

Kumar S, Nei M, Dudley J, Tamura K (2008) MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences. Brief Bioinform 9:299–306. https://doi.org/10.1093/bib/bbn017

Kwon SY, Lee HY, Kim EH, et al (2016) Investigation into the sequence structure of 23 Y chromosomal STR loci using massively parallel sequencing. Forensic Sci Int Genet 25:132–141. https://doi.org/10.1016/j.fsigen.2016.08.010

Leat N, Ehrenreich L, Benjeddou M, et al (2007) Properties of novel and widely studied Y-STR loci in three South African populations. Forensic Sci Int 168:154–61. https://doi.org/10.1016/j.forsciint.2006.07.009

Lim S-K, Xue Y, J Parkin E, Tyler-Smith C (2007) Variation of 52 new Y-STR loci in the Y Chromosome Consortium worldwide panel of 76 diverse individuals. Int J Legal Med 121:124–127. https://doi.org/10.1007/s00414-006-0124-8

Miller S.A, Dykes D.D, Polesky H.F (1988) A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Research 16:1215. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC334765/

Novroski NMM, King JL, Churchill JD, et al (2016) Characterization of genetic sequence variation of 58 STR loci in four major population groups. Forensic Sci Int Genet 25:214–226. https://doi.org/10.1016/j.fsigen.2016.09.007

Nuñez C, Baeta M, Fernández M, et al (2015) Highly discriminatory capacity of the PowerPlex® Y23 System for the study of isolated populations. Forensic Sci Int Genet 17:104–107. https://doi.org/10.1016/j.fsigen.2015.04.005

Oota H, Settheetham-Ishida W, Tiwawech D, et al (2001) Human mtDNA and Y-chromosome variation is correlated with matrilocal versus patrilocal residence. Nat Genet 29:201. http://doi.org/10.1038/ng711

Park MJ, Lee HY, Yang WI (2012) Understanding the Y chromosome variation in Korea — relevance of combined haplogroup and haplotype analyses. Int J Legal Med 126:589–599. https://doi.org/10.1007/s00414-012-0703-9

Parkin EJ, Kraayenbrink T, van Driem GL, et al (2006) 26-Locus Y-STR typing in a Bhutanese population sample. Forensic Sci Int 161:1–7. https://doi.org/10.1016/j.forsciint.2005.10.008

Purps, J., Siegert, S., Willuweit, S., Nagy, M., Alves, C., Salazar, R., Roewer, L. (2014). A global analysis of Y-chromosomal haplotype diversity for 23 STR loci. *Forensic Science International: Genetics*, *12*, 12–23. https://doi.org/10.1016/j.fsigen.2014.04.008

Redd AJ, Agellon AB, Kearney VA, et al (2002) Forensic value of 14 novel STRs on the human Y chromosome. Forensic Sci Int 130:97–11. https://doi.org/10.1016/S0379-0738(02)00347-X

Ristow PG, Cloete KW, D'Amato ME (2016) GlobalFiler® Express DNA amplification kit in South Africa: Extracting the past from the present. Forensic Sci Int Genet 24:194–201. https://doi.org/10.1016/j.fsigen.2016.07.007

Rodig H, Grum M, Grimmecke H-D (2007) Population study and evaluation of 20 Y-chromosome STR loci in Germans. Int J Legal Med 121:24–27. https://doi.org/10.1007/s00414-005-0075-5

Roewer L (2009) Y chromosome STR typing in crime casework. Forensic Sci Med Pathol 5:77–84. https://doi.org/10.1007/s12024-009-9089-51

Roewer L, Arnemann J, Spurr NK, et al (1992) Simple repeat sequences on the human Y chromosome are equally polymorphic as their autosomal counterparts. Hum Genet 89:389–3941. https://doi.org/10.1007/bf00194309

Ruitberg CM, Reeder DJ, Butler JM (2001) STRBase: a short tandem repeat DNA database for the human identity testing community. Nucleic Acids Res 29:320–322. https://doi.org/10.1093/nar/29.1.320

Thompson JM, Ewing MM, Frank WE, et al (2013) Developmental validation of the PowerPlex® Y23 System : A single multiplex Y-STR analysis system for casework and database samples. Forensic Sci Int Genet 7:240–25010. https://doi.org/1016/j.fsigen.2012.10.013

Tishkoff SA, Williams SM (2002) Genetic analysis of African populations: human evolution and complex disease. Nat Rev Genet 3:611. http://dx.doi.org/10.1038/nrg865

Warshauer DH, Churchill JD, Novroski N, et al (2015) Novel Y-chromosome Short Tandem Repeat Variants Detected Through the Use of Massively Parallel Sequencing. Genomics Proteomics Bioinformatics 13:250–257. https://doi.org/10.1016/j.gpb.2015.08.001

Chapter 4:

4.0 Technical overview for construction of a balanced allelic ladder.

4.1 Introduction:

The allelic ladder supplied with a commercial Forensic STR typing kit is by far its most critical component. The allelic ladder which is hosts to common and rare alleles also captures information on unique variants observed in vivo for the STR marker or panel (Sajantila et al.1992, Gill et al.1996 and Barber and Parkin 1996). Its composition is largely a synthetic mixture of PCR products which have been accurately defined in length by sequencing across the repetitive DNA motif (Puers et al. 1993). Its primary role is to ensure accurate allele designation, size precision and concordance between different DNA typing workflows. The characterization of the allele by sequencing is therefore a crucial aspect of the ladder construction as the nucleotide composition is well known to impact the STR migration through polyacrylamide gels (Frankand and Köster 1979) and a capillary (Mansfield et al. 1996). The ladder therefore allows for inter-laboratory data comparison by adjusting for different size measurements which can occur under different experimental conditions or between different Genetic Analyser instruments (Butler 2014). This size precision assurance is also of great importance for characterizing microvariants as even a 1bp deletion as seen in Chapter 3 can result in an allele shift during capillary electrophoresis (CE). A parallel analysis with the allelic ladder may also compensate for various technical factors which are known to influence the CE sizing precision. This include the type of polymer used, the size and age of the capillary and even room temperature fluctuations (Butler 2014).

The importance of allelic ladder construction in forensics is recognised since the beginning of the DNA typing era (Wayman and White 1980). The conventional approach included PCR enrichment of the observed alleles followed by the polyacrylamide gel purification (PAGE) of each individual STR allele. The extracted alleles are amplified for a second time to enrich the product for sequencing and to achieve a balanced pooling of individual alleles for the locus (locus specific ladders) (Sajantila et al. 1992, Gill et al. 1996 and Barber and Parkin 1996). It was also common practice to bulk the ladder form the locus specific mix with a dilution and re-amplification strategy using reduced PCR cycles (Griffiths et al. 1998 and Butler et al. 2003). This methodology has been the most widespread approach used in forensic to date, however the PAGE purification step may be recognised as low throughput, rate limiting for DNA recovery and not ideal for long term storage of the individual allele isolates. Evidently in some methodologies, the PAGE purification step has been omitted and therefore the locus specific ladder mix is produced directly after the first PCR amplification from genomic DNA with subsequent dilution and re-amplified to generate a bulk (Hill et al. 2008, Alghafri et al. 2015 and Shao et al. 2015). Although this approach is less labour intensive and provides faster turnover, it demands frequent utilization of genomic DNA especially for the rare alleles which may become depleted during bulk productions and for the balancing of the locus allele mix. Allelic ladder construction by cloning of the alleles is therefore a promising alternative for massive allele enrichment which may be more suitable for bulk productions and long-term storage (Bai et al. 2010, Wang et al. 2014 and Zhang et al. 2015). The major advantage of the cloning method is that recombinant bacterial colonies can be stored in glycerol for long term (Feltham et al. 2018) and plated for outgrowth when required. The PCR allele amplification which is performed on the plasmid extracts usually requires massive dilution factors which may generate a lifetime supply of the rare STR alleles. This approach also provides for a workflow more conducive for automating the process for bulk production.

In this chapter, technical recommendations are made for producing a bulk supply of a balanced allelic ladder using the cloning and sequencing methodologies described in Chapter 3. It presents the analytical and technical challenges that may be encountered for approving an allelic ladder for forensic application and describes a workflow which may maximize throughput and productivity. We herein show how massive allele enrichment achieved by amplification from plasmid extracts can generate a lifetime supply of a bulk balanced allelic ladder suitable for forensic applications.

This work was achieved by an on-going collaboration with our industrial partner Inqaba Biotec. All work related to cloning and sequencing of alleles were conducted at the facilities of the industrial partner. Storage of clones and the extracted plasmids were also located at the facilities of the industrial partner. The optimization of a workflow to construct the allelic ladder was conducted by the PhD candidate at the UWC facilities.

4.2 Results and discussion

4.2.1 Locus specific ladders

In order to produce a Locus specific ladder from plasmid extracts certain quality assurance measures should be considered. Plasmid extractions are only perused for positive recombinant bacterial colonies for which the inserts have been approved by sequencing. However, the quality of the plasmid extract needs to be defined using several criteria to achieve desirable downstream results. 1) Plasmid extract concentrations should be measured by Nanodrop or Qubit and a minimal cut-off set with regard to amplification performance. 2) Singleplex amplification using labelled primers should confirm the absence of non-specific products from a normalized dilution of the plasmid extracts 3) The size of the allele from CE should correspond with the sequencing results.

In Figure 3.1 and Table 3.1 below shows an example for a batch of DYS481 plasmid extract validation to achieve a balanced locus specific ladder. In our experimental design the minimum plasmid extract concentration accepted was 5ng/µl, this was determined enough to provide an off-scale CE signal using the largest dilution factor to achieve normalization across the locus. The DYS481 batch of 11 alleles had an averaged plasmid extract concertation of 24 ng/µl ±7.0. The maximum dilution factor of 1:500 was determined optimal to provide an off-scale CE signal for each allele. In Figure 3.1 we observe the plasmid extract amplifications were devoid of non-specific products and the respective allele calling corresponded with sequence information (see Chapter 3). Furthermore, the predetermined optimal dilution factor must be established also considering that mixing each PCR product together as in Figure 3.2 would impact its signal in the locus specific ladder. The reduction in the CE signal once all PCR products were mix is indicated in Table 3.1 and the reduction factor presented accordingly. The total volume of this locus specific ladder was 220 µl (11 alleles at 20 µl each).

This mixture was also diluted as 1:50, 1:100 and 1:200 to measure the dilution impact on the mix considering that another 9 loci are to be combined. These dilutions are also reserved for the re-amplification approach to generate more ladder and for ladder updates with novel alleles. Interestingly, diluting the locus specific mix up to 200 fold (Figure 3.2D) only gave an average reduction factor of 5.0 ± 0.58. Using this approach, a balanced locus specific ladder is generated for each of the 10 Y-STRs, in addition each mix is diluted accordingly to establish its dilution factor to present each allele between 1000-1200 RFU, which is the peak height range intended for the full composite ladder. As indicated in Figure 3.2B the dilution of 1:50 for the DYS481 locus specific ladder was optimal to reach the desired peak height (threshold at 1200 RFU is indicated accordingly).

Table 3.1: DYS481 plasmid extracts validated by capillary electrophoresis (CE) detection.

Plasmid Extract	Allele	CE Size(bp)	[DNA] Plasmid extract ng/µl	Maximum dilution factor	RFUs	RFUs for locus mix	Reduction Factor
481-19-A	19	108.65	26.0	1:500	30399	2206	13,8
481-20-A	20	111.54	19.3	1:500	29724	2405	12,4
481-21-A	21	114.42	36.6	1:500	30699	2608	11,8
481-22-A	22	117.52	33.4	1:500	28676	2630	10,9
481-23-A	23	120.07	20.2	1:500	30874	2699	11,4
481-24-A	24	122.94	19.6	1:500	30647	2619	11,7
481-25-A	25	126.01	15.6	1:500	30585	2350	13,0
481-26-A	26	129.0	19.9	1:500	28961	2282	12,7
481-27-A	27	131.86	26.3	1:500	29743	2622	11,3
481-28-A	28	135.26	32.4	1:500	28763	2560	11,2
481-29-A	29	139.0	15.6	1:500	30598	1807	16,9

Figure 3.1: PCR amplifications for DYS481 plasmid extractions showing an absence of non-specific products. Y-axis scale 30 000 RFU.

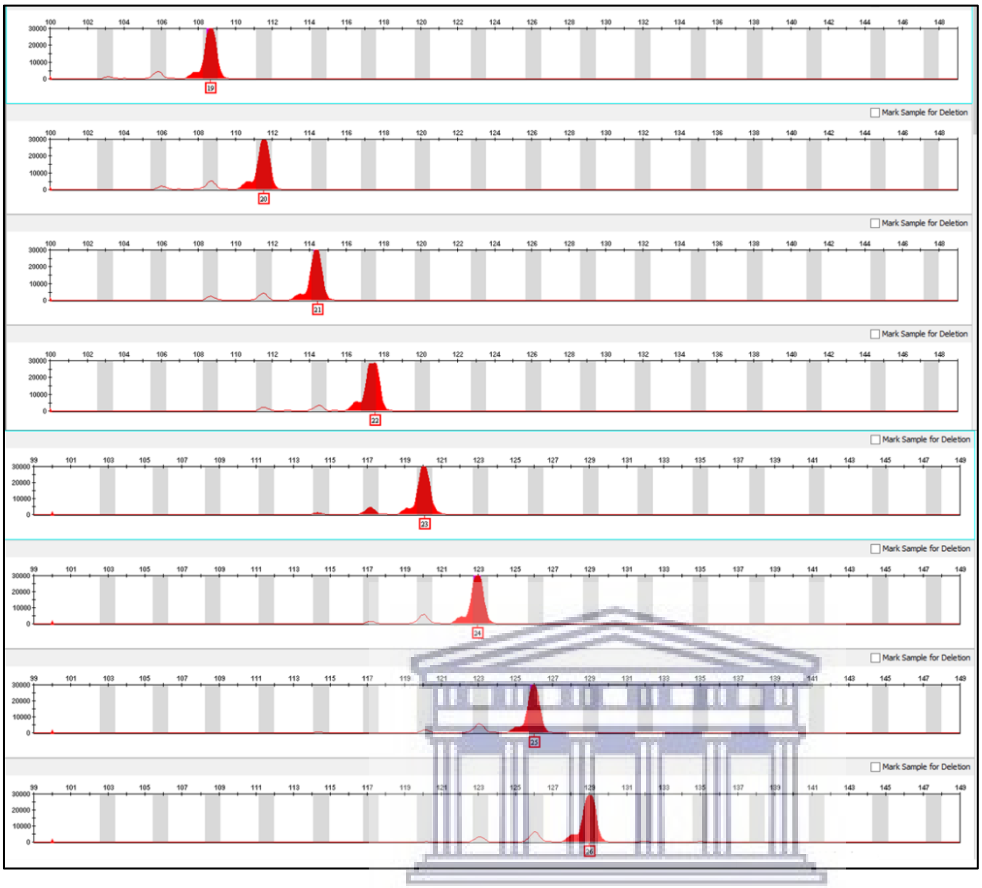
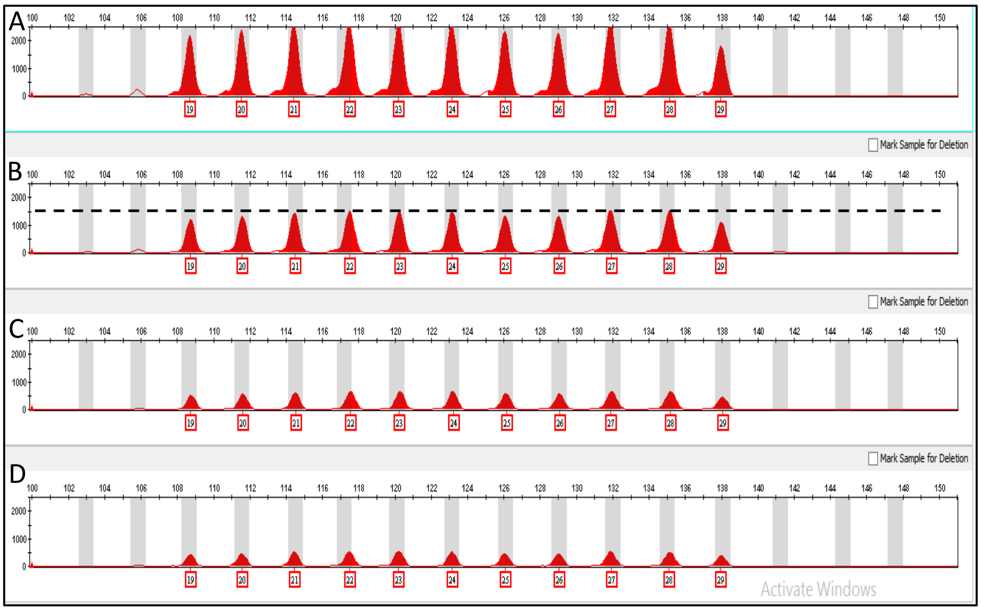


Figure 3.2 : DYS481 locus specific ladder mix for a batch amplification of 11 cloned alleles. A) Undiluted DYS481 locus specific mix (1:1) and its dilutions B) 1:50, C) 1:100, D)1:200. A Dashed line (---) indicates the 1200 RFU threshold at a 1:50 dilution. Y-axis scale 2000 RFU



Criteria were set to evaluate the acceptable level of non-specific amplification from the plasmid extracts. Some of the extracts gave variable degrees of non-specific amplifications after plasmid dilution, these products were either observed at the same locus or adjacent loci. The criteria were established to decide if re-cloning is required or if a dilution and reamplification approach is required. Prominent non-specific products as in Figure 3.3 for DYS710 would have resulted in re-cloning, however non-specific products in the range of 1000-1200 RFU as presented in Figure 3.4 for the DYS518 plasmid extract applications was also interpreted as not optimal. Serial dilutions for these DYS518 PCR products up to 1:200 did not remove its presence but did significantly reduced the peak heights of the specific allele. The problem continuing with non-specific products even less than 50 RFU is that using the re-amplification approach to bulk the locus specific mix also enhanced the non-specific products (Figure 3.4B). To overcome this, PCR products were diluted until disappearance on CE without losing the signal from the specific allele, this is followed by re-amplification of the dilution. For our workflow we determined that the maximum peak height acceptable for a non-specific signal was 500 RFU provided that the specific allele is amplified off-scale. This cut-off should be established considering that re-amplification enrichment of the single specific allele after dilution may still enhance the presence of non-specific products not visible with CE.

Figure 3.3: DYS710 plasmid DNA amplification showing large non-specific products (arrows). Y-axis scaled 20 000 RFU.

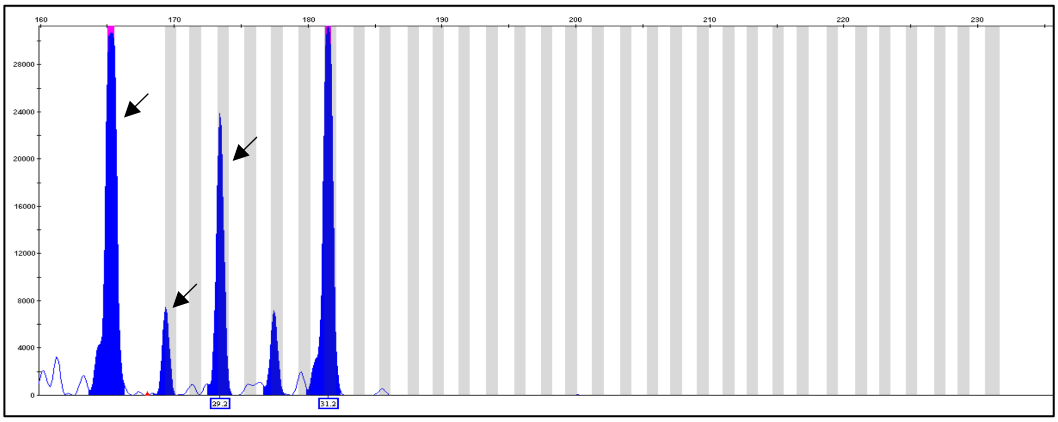
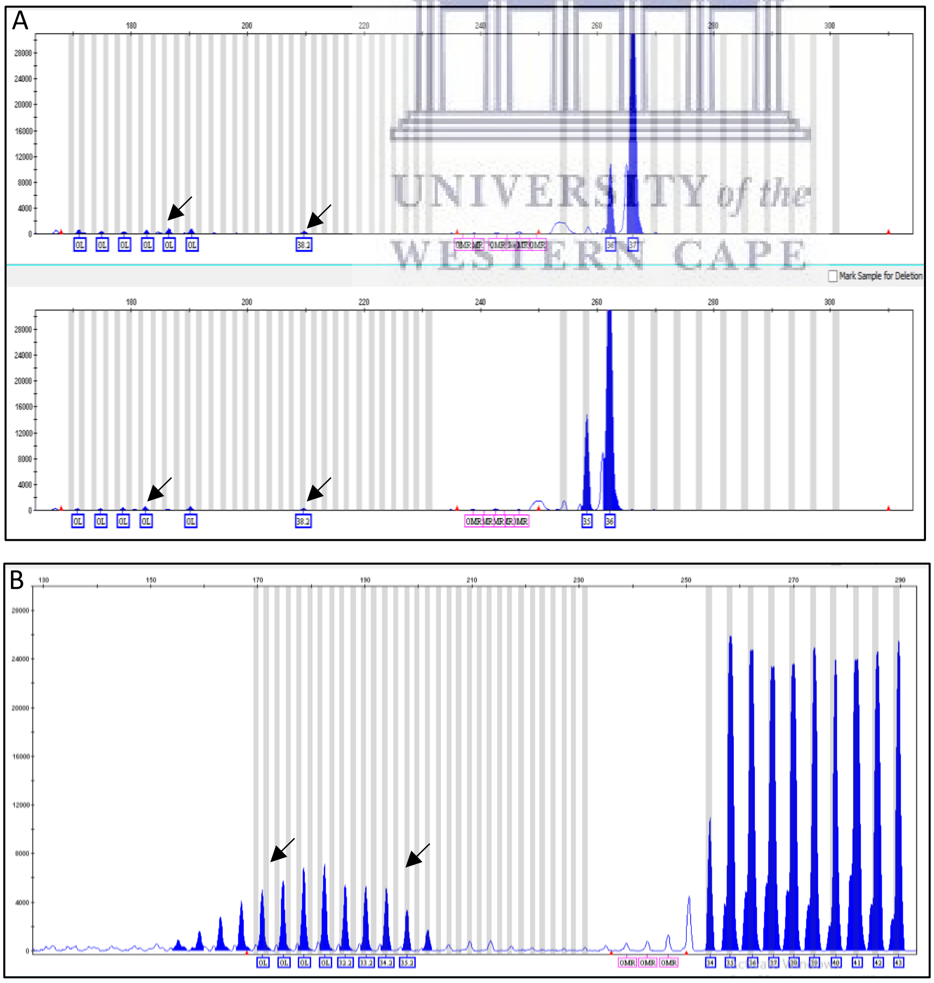


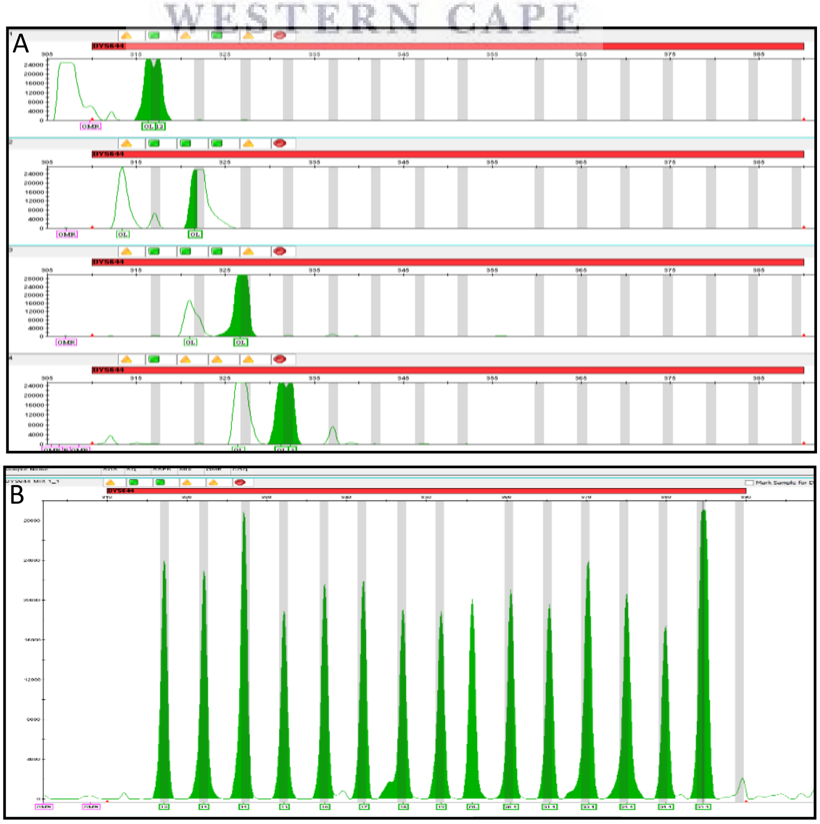
Figure 3.4: DYS518 plasmid extract amplifications showing nonspecific at DYS710.A) Singleplex of plasmid extracts with subtle non-specific. B) DYS518 locus specific mix re-amplification showing enhancement of non-specific products (arrows). Y-axis scale 20 000 RFU.



4.2.2 Post injection hybridisation

Amplification of the diluted plasmid extracts often presented a large degree of PCR products which can result in CE artefacts. Post injection hybridisation on the ABI3500 due to massive allele enrichment generates artefacts which can impact the quality evaluation process. The common feature of this overloaded is wide peak morphology, split peaks and large secondary peaks adjacent to the specific allele. As indicated in Figure 3.5A or DYS644 this could easily be interpreted as a poor-quality plasmid extract having amplified a 1:500 dilution for all DYS644 alleles. Upon pooling of the 15 DYS644 alleles in a 1:1 ratio we generate a total of 300 µl of a locus specific allele mix free of any non-specific products or CE artefacts (Figure 3.5B). Achieving a balanced DYS644 locus mix borderline 20 000 RFU is testimony to the massive allele enrichment achieved from just one round of PCR amplification for a 1:500 dilution of the plasmid extracts.

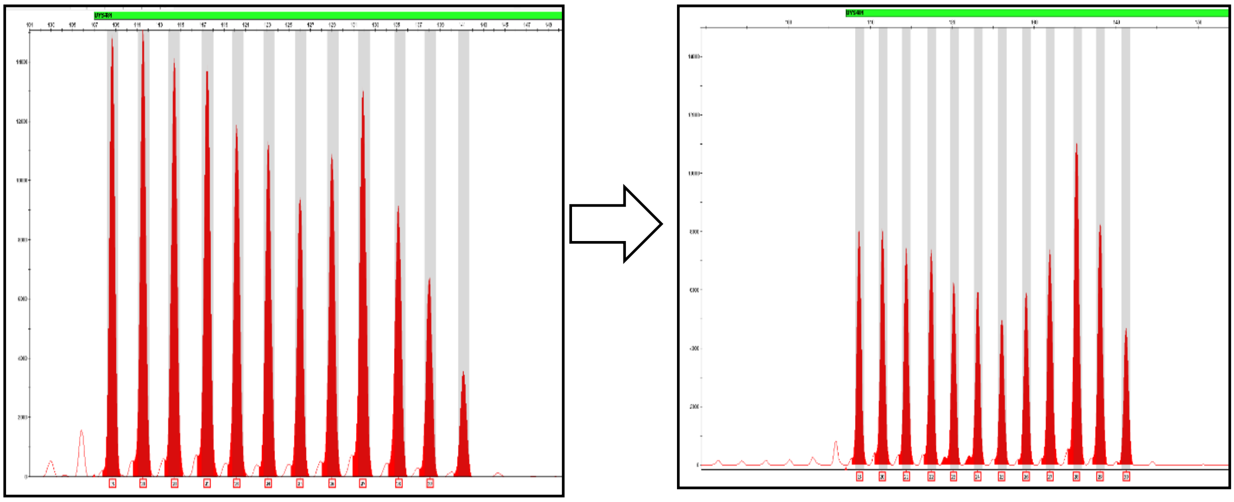
Figure 3.5: DYS644 plasmid extract amplification and validation. A) Singleplex alleles showing post hybridisation artefacts. B) Locus specific ladder mix with disappearance of the artefacts. Y-axis scale 30 000 RFU.



4.2.3 Ladder updates and enrichment

Producing a locus specific ladder allows for fine adjustments to allele balance and faster updates with new alleles or variants which can then be bulked using a simple re-amplification approach. For the singleplex amplifications of diluted plasmid extracts the degree of off- scale amplification can be variable and thus difficult to quantify across the loci, this can result in an imbalance for a 1:1 allele mix as demonstrated for DYS481 (Figure 3.6). To improve the balance a re-adjustment to the prepared mixture is made by individual allele additions (Figure 3.6) which can be re-amplified for enrichment. Updating the ladder with novel alleles can be archived with the same approach, by either making the addition to the stock locus specific ladder or its 1:50, 1:100 or 1:200 dilution kept as a reserve. The re-amplification methodology needs to be optimized for each given marker as to maximize enrichment. In general, these amplifications performed on the locus mix usually needs 12-15 PCR cycles and longer elongation times as described in Butler et al.2007.

Figure 3.6: DYS481 locus specific ladder re-adjusted to improved balance. Y axis scale 16 000 RFU



4.2.4 Ladder compilation

The most delicate step in the production is pooling of each locus specific mix to provide a complete balanced ladder compilation. Considering the following technical and analytical guidelines may assist passing the ladder using the genotyping software such as GeneMapper ® (ThermoFisher) which impart specific criteria for its forensic application (Supplementary 3 Figure 1).

The most important factor would be to determine the optimal peak height range for the composite ladder. The major peak height restriction is usually due to pull-up fluorescence that can impact correct alleles designation between the fluorescent spectrum (Cloete et al.2016). For a system intended for forensic applications the pull-up threshold should be established depending on the fluorescent dyes being utilized and the respective calibration standard (Cloete et al.2016). This would also apply to the allelic ladder component for it pass the quality assessments performed by the GeneMapper® ID-X software (GeneMapper user manual, Rev. A). In the Figure 3.7A below, DYS385 had a locus allele mix peak height in the range of 5000-10000 RFU which and presented pull-up in the range of 50-900 RFU in the blue, yellow and red channel. In Figure 3.7B we again observe pull-up in the blue and red spectrum for the DYS626 mix for an allele peak height range of (4000-8000 RFU). The optimal peak height for the balanced composite ladder was therefore determined to be in the range of 1200-2000 RFU for which pull-up was not detected between fluorescent spectrum. Once the optimal range is determined all locus specific ladders are mixed in a proportion as to normalize each of its alleles within the desired range. This can easily be achieved by first performing a composite mix of all loci using a 1:1 ratio to compare the peak height across the entire spectrum. Using a graphical comparison (Figure 3.8) based on peak heights we can identify which locus specific mixes can be pooled in equal proportions, which require dilution beforehand and which locus mix need to be bulked by re-amplification. Evidently DYS449 was the limiting factor giving a locus specific allele peak height borderline 1000-2000 RFU, it was therefore considered for re-amplification before generating a complete ladder mix. From the comparison DYS644, DYS385 and DYS481 would either require further dilution or its proportion adjusted accordingly for the mix. For assistance the 1:1 complete ladder mix can be visually overlaid as shown in Figure 3.9 to estimate the best proportional mix. For this batch production subsequent to improving the DY449 locus specific peak heights, the loci DYS710, DYS518, DYS612, DYS626, DYS504 DYS447, DYS447 could be mixed in equal proportions, while DYS385 and DYS644 required a 1:30 dilution for the composite ladder mix. For this batch DY449 as the limiting factor was added last to provide a respective degree of balance. For this batch a total of 1,5ml total ladder was produced from a single production to provide a composite ladder which passed the Genemapper® quality assessment (Figure 3.10).

4.3 Conclusion

In this chapter we successfully optimized a workflow for construction an allelic ladder for more than 150 cloned alleles. Using a cloning approach, we demonstrate that massive allele enrichment can be achieved to generate a copious supply of a plasmid extract which may be suitable for bulk production and of long-term storage. With a single batch amplification, a total of 1500 µl of a ladder could be produced considering the DYS449 as the limiting factor. In perspective this would be enough to process 500 X 96 well genotyping plates if a ladder is injected thrice per plate. Generating a locus specific mix with allele peak height in the order of 20 000 RFU demonstrates the potential to efficiently generate bulk supplies using the dilution and reamplification approach. We herein provide guidelines to synthesis a balanced allelic ladder which may be utilize for forensic applications.

Figure 3.7: DYS385 and DYS504 locus specific ladder showing pull-up across the fluorescence spectrum. A) DYS385 alleles in green (scale 20 000 RFU) with pull-up into blue (scale 300 RFU), red (scale 600 RFU) and black (scale 1000 RFU). B) DYS504 alleles in black (scale 10 000 RFU) with pull-up into blue (scale 120 RFU), red (scale 320 RFU). For visualization the yellow fluorescent channel is referred to here as black and Y-axis scales are varied.

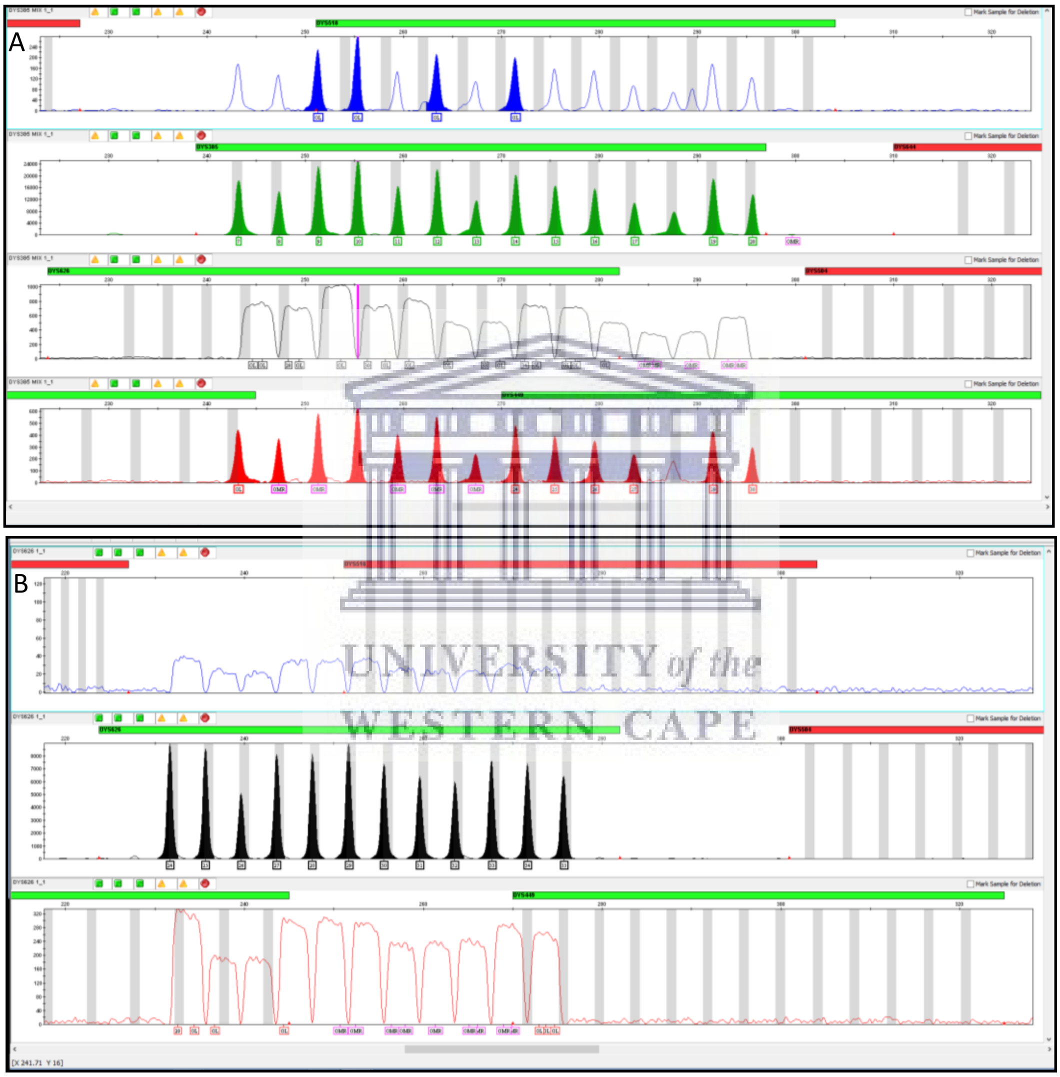


Figure 3.8: Peak height in Relative Fluorescent Units (RFUs) for each locus specific ladder mixed in a 1:1 proportion.

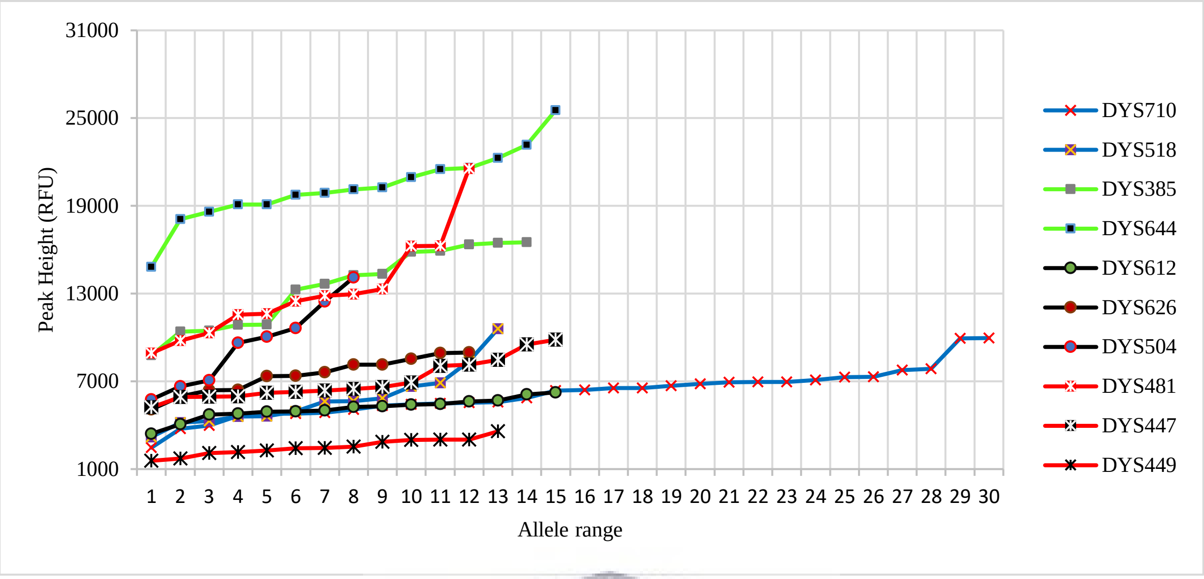


Figure 3.9: Visual overlay of the composite ladder representing each locus specific ladder mixed in a 1:1 proportion. Dashed line represents 2000 RFU threshold. Y-axis scaled to 30 000 RFU

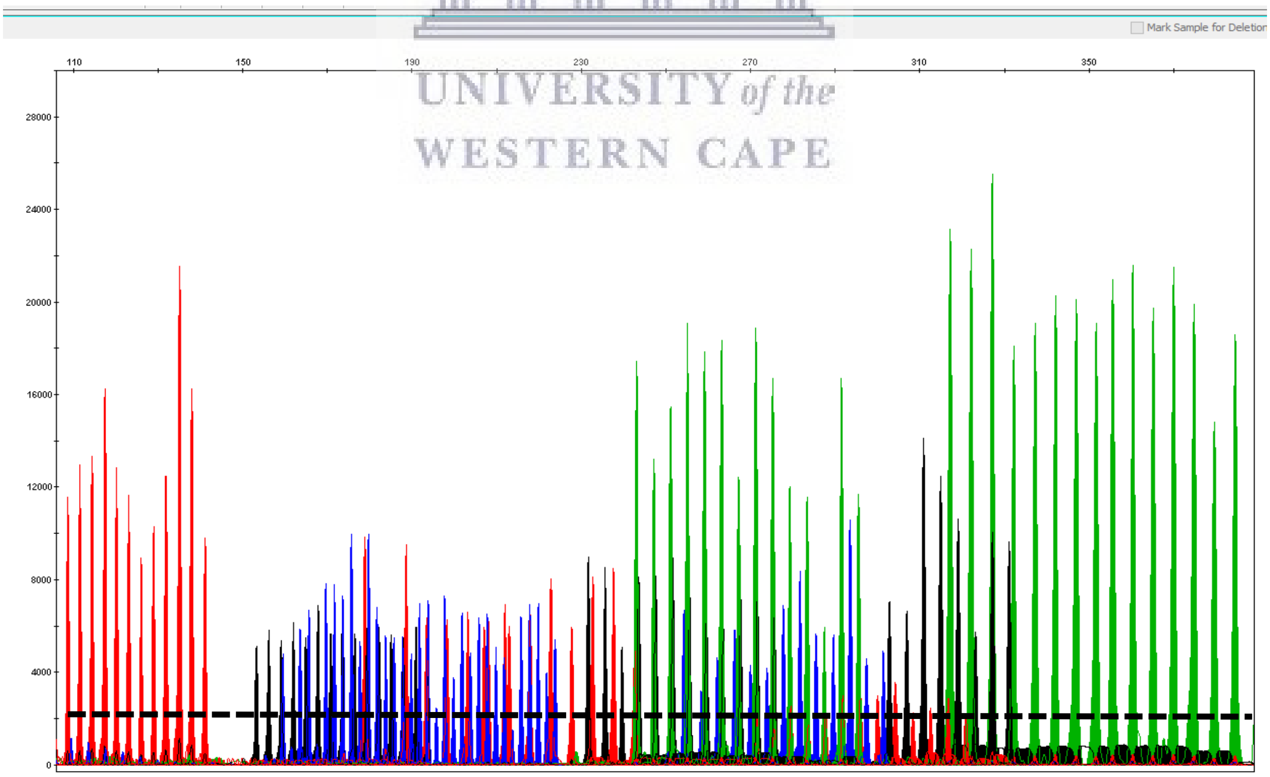
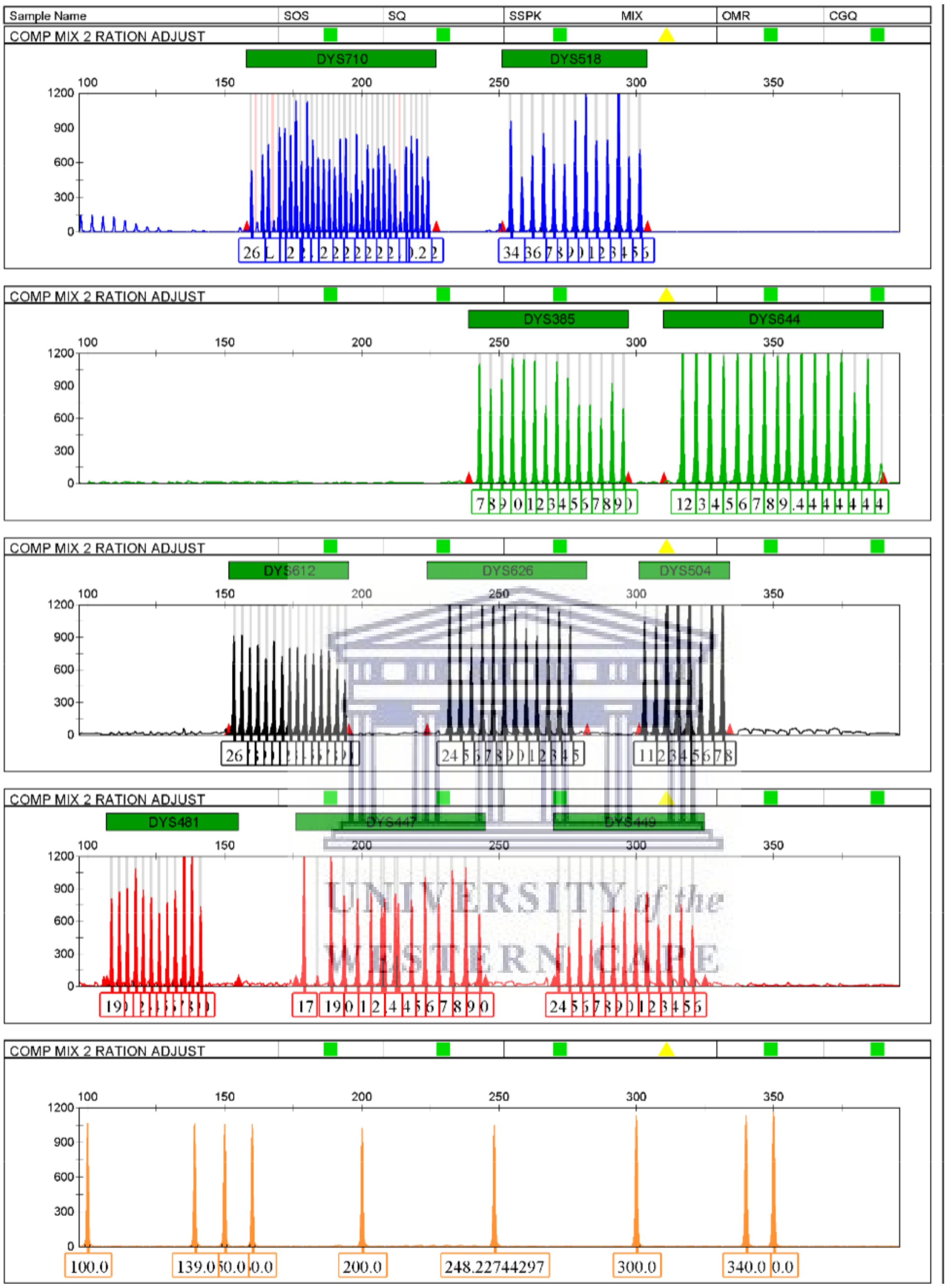


Figure 3.10: The composite balanced allelic ladder. Y-axis scaled to 1200 RFU



4.4 References

Alghafri, R., Goodwin, W., Ralf, A., Kayser, M., & Hadi, S. (2015). A novel multiplex assay for simultaneously analysing 13 rapidly mutating Y-STRs. *Forensic Science International: Genetics*, *17*, 91–98. https://doi.org/10.1016/j.fsigen.2015.04.004

Bai, X., Li, S., Cong, B., Li, X., Guo, X., He, L., Pei, L. (2010). Construction of two fluorescence-labelled non-combined DNA index system miniSTR multiplex systems to analyze degraded DNA samples in the Chinese Han Population. *ELECTROPHORESIS*, *31*(17), 2944–2948. https://doi.org/10.1002/elps.201000163

Barber, M. D., & Parkin, B. H. (1996). Sequence analysis and allelic designation of the two short tandem repeat loci D18S51 and D8S1179. *International Journal of Legal Medicine*, *109*(2), 62–65. https://doi.org/10.1007/BF01355518

Butler, J. M. (2014). *Advanced Topics in Forensic DNA Typing: Interpretation*. Elsevier Science. Retrieved from https://books.google.co.za/books?id=reBQBAAAQBAJ

Butler, J. M., Shen, Y., & Mccord, B. R. (2003). The Development of Reduced Size STR Amplicons as Tools for Analysis of Degraded DNA *. *Journal of Forensic Science*, *48*(5), 1054–1064.

Cloete, K. W., Ristow, P. G., Kasu, M., & D'Amato, M. E. (2017). Design, installation, and performance evaluation of a custom dye matrix standard for automated capillary electrophoresis. *Electrophoresis*, *38*(5). https://doi.org/10.1002/elps.201600257

Feltham, R. K. A., Power, A. K., Pell, P. A., & Sneath, P. H. A. (2018). A Simple Method for Storage of Bacteria at - 76°C. *Journal of Applied Bacteriology*, *44*(2), 313–316. https://doi.org/10.1111/j.1365-2672.1978.tb00804.x

Frank, R., & Köster, H. (1979). DNA chain length markers and the influence of base composition on electrophoretic mobility of oligodeoxyribonucleotides in plyacrylamide-gels. *Nucleic Acids Research*, *6*(6), 2069–2087. https://doi.org/10.1093/nar/6.6.2069

GeneMapper® ID-X Software Version 1.5 Reference Guide (Pub. no. 100031707 Rev. A) https://assets.thermofisher.com/TFSAssets/LSG/manuals/100031707_GeneMapIDX_ver1_5_ReferenceGuide.pdf

Gill, P., Urquhart, A., Millican, E., Oldroyd, N., Watson, S., Sparkes, R., & Kimpton, C. P. (1996). A new method of STR interpretation using inferential logic -development of a criminal intelligence database. *International Journal of Legal Medicine*, *109*(1), 14–22. https://doi.org/10.1007/BF01369596

Griffiths, R. A. L., Barber, M. D., Johnson, P. E., Gillbard, S. M., Haywood, M. D., Smith, C. D., Gill, P. (1998). New reference allelic ladders to improve allelic designation in a multiplex STR system. *International Journal of Legal Medicine*, *111*(5), 267–272. https://doi.org/10.1007/s004140050167

Hill, C. R., Kline, M. C., Coble, M. D., & Butler, J. M. (2008). Characterization of 26 miniSTR loci for improved analysis of degraded DNA samples. *Journal of Forensic Sciences*, *53*(1), 73–80. https://doi.org/10.1111/j.1556-4029.2008.00595.x

Mansfield, E. S., Vainer, M., Enad, S., Barker, D. L., Harris, D., Rappaport, E., & Fortina, P. (1996). Sensitivity, reproducibility, and accuracy in short tandem repeat genotyping using capillary array electrophoresis. *Genome Research*, *6*(9), 893–903. https://doi.org/10.1101/gr.6.9.893

Puers, C., Hammond, H. A., Jin, L., Thomas Caskey, C., & Schummt, J. W. (1993). Identification of Repeat Sequence Heterogeneity at the Polymorphic Short Tandem Repeat Locus HUMTHO I [AATG]n and Reassignment of Alleles in Population Analysis by Using a Locus-specific Allelic Ladder. *Am. J. Hum. Genet*, *53*, 953–958. https://doi.org/10.1016/j.jcp.2011.08.012

Sajantila, A., Puomilahti, S., Johnsson, V., & Ehnholm, C. (1992). Amplification of reproducible allele markers for amplified fragment length polymorphism analysis. *BioTechniques*, *12*(1), 16,18,20-22.

Shao, C., Zhang, Y., Zhou, Y., Zhu, W., Xu, H., Liu, Z., Xie, J. (2015). Identification and characterization of the highly polymorphic locus D14S739 in the Han Chinese population. *Croatian Medical Journal*, *56*(5), 482–489. https://doi.org/10.3325/cmj.2015.56.482

Wang, L., Zhao, X. C., Ye, J., Liu, J. J., Chen, T., Bai, X., Wang, F. (2014). Construction of a library of cloned short tandem repeat (STR) alleles as universal templates for allelic ladder preparation. *Forensic Science International: Genetics*, *12*, 136–143. https://doi.org/10.1016/j.fsigen.2014.06.005

Wyman, A. R., & White, R. (1980). A highly polymorphic locus in human DNA. *Proceedings of the National Academy of Sciences*, *77*(11), 6754 LP-6758. Retrieved from https://doi.org/10.1073/pnas.77.11.6754

Zhang, S., Bian, Y., Tian, H., Wang, Z., Hu, Z., & Li, C. (2015). Development and validation of a new STR 25-plex typing system. *Forensic Science International: Genetics*, *17*, 61–69. https://doi.org/10.1016/j.fsigen.2015.03.008

Chapter 5:

5.0 The UniQTyper™ Y-10: Forensic genetics and haplotype diversities across South African populations.

5.1 Introduction

5.1.1 Y-STR background

Although more than a decade has passed since the development of the first commercial forensic Y-STR systems, many countries do not exercise the benefits of Y-STR analysis due to various limiting factors. The general limitation of Y- STRs is that it cannot independently exclude paternal relatives as potential contributors to the DNA evidence (Roewer et al. 2009). The Y-chromosome haplotypes also have the tendency to cluster geographically and therefore demands thorough population and sub-population representation to provide meaningful match probability estimates (Butler et al. 2005). Furthermore, the exclusion of non-related males may be a challenge when inbreeding, population isolation and patrilocal practises has influenced low Y-chromosome diversities (Oota et al.2001, Kyser et al. 2003, Nuñez et al. 2015). In such homogenous populations the core Y-STR panels of commercial kits such as the PowerPlex®Y (Promega) and AmpFLSTR®Y-filer™ (Thermo Fisher) where shown limited for maximizing male discrimination (Nuñez et al. 2015 and D'Amato 2010). These panels have therefore been supplemented using Y-STR of higher mutation rates in-order to improve male discrimination between homogenous paternal lineages and also between paternal relatives (Redd et al. 2002, Kayser et al. 2004; Decker et al. 2007; Hanson et al. 2007; Ballantyne. 2010 and D'Amato et al. 2010-2011).

The UniQ Y-Typer™ Y-10 presented herein is a 10 Y-STR multiplex that contains 4 rapidly mutating markers. The multiplex initially validated for forensic application by D'Amato et al (2011) is was characterised as a highly discriminatory platform. Utilizing a carefully ascertained set of highly polymorphic Y-STRs was shown to significantly improve discrimination amongst South African Bantu men compared to the 17 markers of the AmpFLSTR®Y-filer™ kit (Thermo Fisher) (D'Amato et al. 2011). The AmpFLSTR® YFiler-plus (Thermo Fisher) and PowerPlex®Y23 (Promega) which have 7 and 2 rapidly mutating markers respectively provides currently the highest discriminatory commercial systems. However, population data for these systems is still limited for South African and the performance of its minimal haplotype composition is known to provided limited discriminatory capabilities in bantu men (DC = 0.63) (D'Amato 2010). Furthermore, the loci DYS391, DYS392 and DYS437 are also known to be largely monomorphic amongst the Xhosa (Leat et al. 2007).

Previous studies show that with minimum of 13 rapidly mutating Y-STRs a higher degree of haplotype resolution can be obtained for various worldwide population in comparison the standard markers of commercial kits (Ballantyne 2012 and 2014). In this global collaboration we contributed N=116 Xhosa haplotypes and obtained 100% discrimination. The potential to obtain a high level of discrimination amongst diverse population groups using a minimal maker set was a key aspect in the design of the UniQ Y-Typer™ Y-10.

5.1.2 South African ethno-linguistic diversity

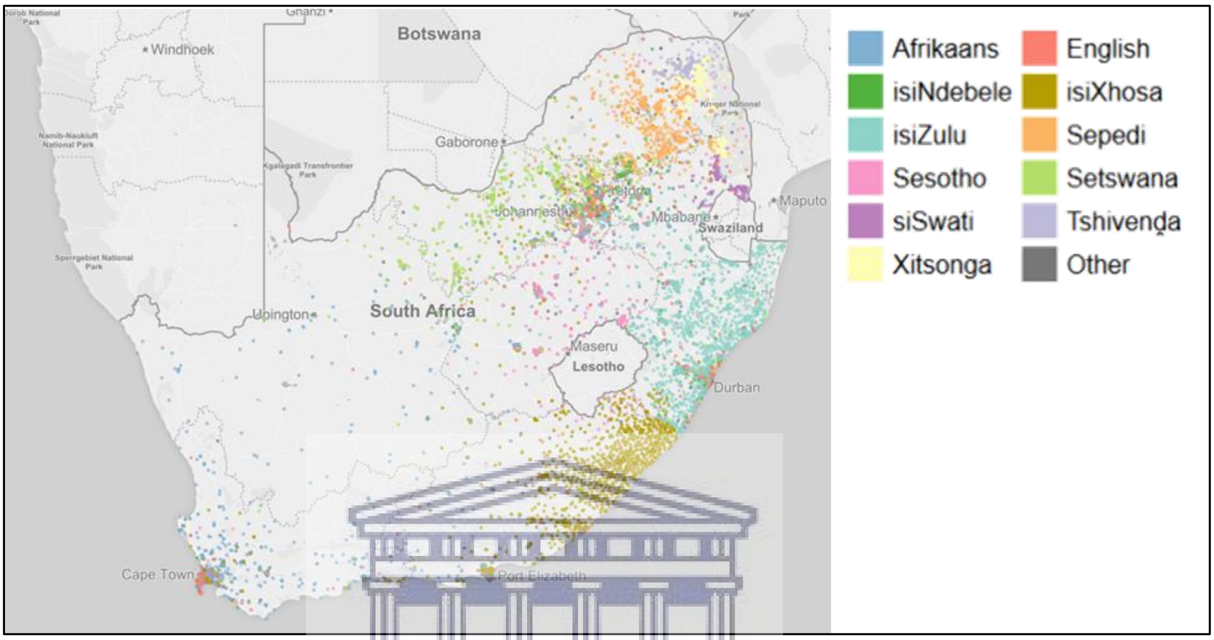
South Africa with its more than 57 million inhabitants is known for being one of the most ethno-linguistic diverse countries in the world (Statistics South Africa 2018 and Fearon 2003). According the latest national census the country is predominated by Black Africans (79.2%) which is a makeup of nine Southern Bantu-speaking groups that originated from West-Central Africa through the Bantu expansion (Nurse and Philippson 2006; de Filippo et al. 2012 and Marks et al. 2015). In comparison the ethnic minority White (9.1%) and Asian Indian (2.5%) populations are predominately immigrants which established settlements in South Africa between the 16th to 19th century.

South Africa with 11 official languages is host to a rich ethno-linguistic diversity which can be found distributed across its nine provinces as shown in Figure 4.1. Majority of South Africans speak one or more Bantu languages which have two main classifications: The Nguni branch which include Zulu, Xhosa, Swazi, Ndebele and the Sotho–Tswana branch which include Sotho, Pedi and Tswana. Other South African Bantu languages include Tsonga which originating from Southern Mozambique groups of the Bantu Expansion and Venda which is mainly spoken in the Northern tip of South Africa bordering Zimbabwe. The White population can be linguistically separated into Afrikaans and English-speaking groups. The Afrikaner (Afrikaans speaking) are largely descendants from Dutch settlers arriving in 1652, while the English speaking are predominantly from of European British colonist whom displaced the Dutch rule in 1806 (Lloyd 1844). In modern day South Africa both these groups are most concentrated around the Western Cape and Gauteng regions and dispersed in smaller pockets for in rest of the country.

The Indian population of South Africa is also largely an English speaking group who are largely the descendants of Indian-Asian immigrants first arriving in the late 18th century to work in the sugarcane plantations in the East coast of South Africa.

The largest admixed group is by far the native Coloured, which constituents ~8.8% of the South African population. The Coloured which are located predominantly in the Western Cape and Northern Cape provinces represent a complex mixed ancestry between indigenous Khoi-San (Khoi and San), Bantu speaking and immigrant groups. The coloured population can be represented by two main groups, namely the Cape coloured and Griqua. The Cape Coloured which is a highly admixed population of Khoi-San, Europeans, Bantu speaking and Indian (de Wit et al.2010) constitutes the largest population in the Western Cape. The Coloured group referred to as Griqua on the other hand which also originated from the Cape is mainly found to date in the Northern Cape in a much smaller community. The Griqua ancestry is known to be predominated by the descendants of relations between Dutch men and Khoi-San women, however evidence of non Khoi-San influence in the Northern Cape Griqua is also apparent (Nurse and Jenkins 1975 and Morris 1997). The non-Bantu indigenous people of South Africa (Khoi-San) whom are largely pastoralists and hunter-gatherers speak a complex click language very different to the Bantu languages. The main Khoe-San language families include Ju (Northern Khoisan), Khoe-Kwadi (Central Khoisan) and Tuu (Southern Khoisan), for which many of its different dialects are extinct. Under the apartheid regime in South Africa all Khoi-San were forced to register as coloured and to date completely lost their identity. Currently the only remaining distinct Khoe group from the Khoe-Kwadi family in South Africa the Nama people, which can be found still living in Namibia and in the Northern Cape region bordering Namibia.

Figure 4.1: The distribution pattern for the 11 official South African languages according to population density data provided in the national census of 2011.



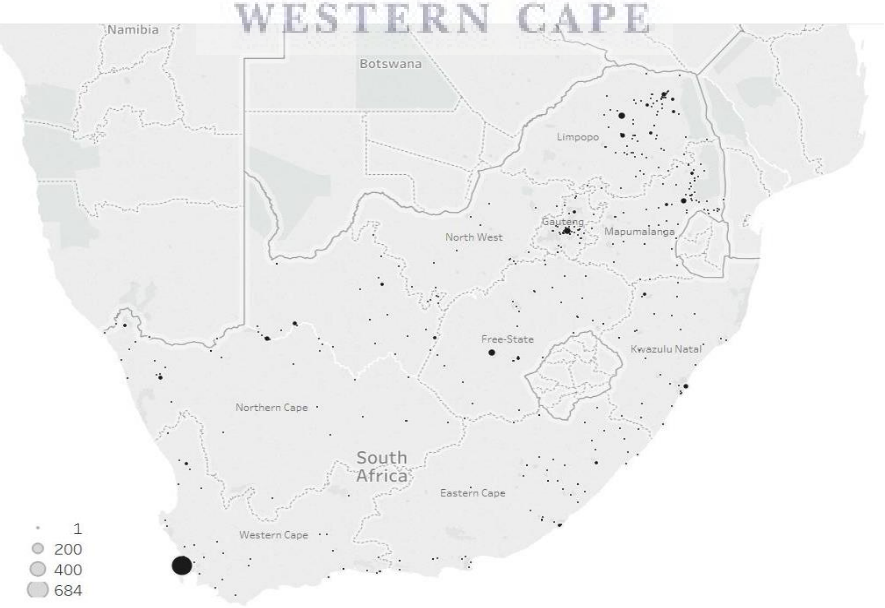
In this study, a comprehensive Y-STR analysis within the 9 provinces of South Africa is performed for a total of 2201 male haplotypes sampled from 15 ethno-linguistic populations. The chapter aims to validate UniQTyper™ Y-10 panel intended for forensic investigations, genealogy studies and kinship analysis. In this work we report forensic summary statistics, allele frequencies, novel allele variants and population differences based on genetic distances between groups.

5.2 Material and methods:

5.2.1 Sample collection

The DNA samples utilized for this population study were obtained by informed consent and ethical approval form the University of the Western Cape (10/3/39) and (15/4/97). For the 2201 unrelated male DNA samples collected, participants were obtained across the 9 South African provinces as depicted on the map below (Figure 4.2). Participants were requested to share information of their own ethnic group, the paternal relative's ethnicity and the spoken home language. All samples were assigned a unique barcode and grouped according to the ethnic origin of the paternal lineage. For this study a total of 15 ethnic groups were represented as follows : Afrikaner (161); English (111); Indian(104); Coloured (500); Griequa (68); Nama (47); Pedi (198); Venda (122); Southern Sotho (70); Twana (99); Tsonga (118); Swazi (104); Ndebele (16); Zulu (180) and Xhosa (303).

Figure 4.2: sampling distributions across 9 South African provinces. Size of the dots indicate number of individuals. The co-ordinates represent the birth place of individual participants or collection point when birth location is not specified.



5.2.2 DNA extraction and genotyping

DNA samples were extracted from saliva using the adapted salting out procedure described in Chapter 2. Population genotyping was performed using the UniQTyper™ Y-10 master mix and primer mix scaled proportionally for a 15 µl PCR reaction. An average sample DNA input of 4ng was utilized in order to streamline a high success rate in genotyping from 96 well optical plates (Thermo Fisher). As a quality control measure the positive control 2800M (Promega) and 007 (Thermo Fisher) was genotyped in parallel per PCR plate. The PCR thermal cycling parameters were programmed as described in Chapter 2 on an Arktik thermal cycler and amplified for 30 cycles. Y-STR haplotypes were resolved by capillary electrophoresis using the parameters and conditions described in chapter 2.

5.2.3 Quality control

The genotyping workflow was subjected to routine quality assurance standards. This included the typing of the NIST-SRM 2395 components A-D and the internal control male DNA samples FDLP001- FDLK003 which have been sequenced for each Y-STR. Novel alleles and variants (duplication/ triplications/nulls/ intermediates) obtained during multiplex genotyping were confirmed in singleplex. When possible, the novel, rare and intermediate alleles with an exception to duplications and triplications were sequenced by the Sanger method as described in Chapter 2 to confirm concordance between capillary electrophoresis and sequencing (Chapter 2).

5.2.4 Forensic summary statistics

In total, N=2201 haplotypes were considered for the forensic summary statistics. Allele frequencies and haplotype frequencies were computed in Arlequin v3.5.1.2 (Excoffer et al. 2007) using the counting method. For this analysis unique fantasy numbers were assigned for each different duplication, triplication and null. The fantasy numbers were selected higher than the largest alleles as not to overlap with designated alleles in the dataset. The discrimination capacity (DC) was calculated as the proportion of distinct haplotypes and the total number of haplotypes and the random Match probability (MP) calculated from the sum of squared haplotype frequencies.

5.2.5 Population diversity and structure.

Haplotype diversity (HD) and Gene diversity (GD) were calculated accordingly as n(1−∑pi2)/(n−1) (Nei and Tajima, 1981), where n represent the total number of samples and Pi the relative frequency of the i-th haplotype for (HD) or allele for (GD) calculations. Population genetic structure for the dataset was assessed with the analysis of molecular variance (AMOVA) for which the genetic distances between groups were determined using the R_{st} like distance matrix in Arlequin v3.5.1.2 (Excoffer et al. 2007) using a 5% significance and 10. 000 permutations per comparison. For a non- hierarchical AMOVA analysis the DYS385ab locus was excluded, likewise haplotypes presenting deletions, null alleles, duplications and triplications were all removed, leaving a total of 2153 haplotypes. Multidimensional scaling (MDS) on the reduced dataset was based on Slatkin's linearized pairwise R_{st} values (Slatkin 1995) with graphical outputs obtained with the software package SPSS (IBM corp. in Armonk, NY).

5.2.6 Network analysis

Median joining networks were produced using the software package Network v5.0.0.3 (Bandelt et al.1999). The analysis was performed with exclusion of locus DYS385ab for high frequency haplotypes and other haplotypes differing by one mutation event.

5.2.7 Haplotype and haplogroup external data

Haplotype data previously published for the Powerplex® Y-23 (Purps et al. 2014) for the same N=95 Xhosa individuals genotyped herein were utilized for a comparative of concordance and discriminatory performance. The lower number of samples we have profiled compared to Purps et al.2014 was due to the unavailability of the DNA source. Haplogroup information for small subset of samples were obtained from an in-house Y-SNP assay described in (Burrows 2018) or provided by C. Capelli (unpublished data).

5.3 Results:

5.3.1 Alleles frequencies and novel variants

In this study, a total of 206 unique alleles were identified across the 10 Y-STR loci for the 2201 male haplotypes. The loci DYS710 and DYS644 which displayed 39 and 24 different alleles respectively showed higher than the average number of alleles [16; *sd ± 3*] observed at DYS518, DYS612, DYS626, DYS504, DYS481, DYS447 and DYS449 (Table 4.1 and Supplementary 4 Figures 1-11). Intermediate alleles were observed across the DYS447 (n=4), DYS481 (n=2) and DYS385(n=1). In total, 21 duplications occurred across DYS710 (n=17), DYS518 (n=1), DY644 (n=1), DYS612 (n=1) and DYS481 (n=1). A novel triplication at DYS385 and a total of 25 haplotypes presenting a null allele at DYS626 is reported. For all 17 of the DYS710 duplications identified, 9 were unique (Table 4.1, supplementary 4 Figure 1) and for the majority associated with DYS644 alleles variants (23.3 to 24.3) which we previously characterized as 1bp deletion variants determining x.3 alleles (see Chapter 3). Notably, these DYS644 x.3 alleles were only observed amongst the haplotypes sampled from the northern South African regions. An association between the x.3.DYS644 alleles and 21 out of the 25 DYS626 null variants were also majorly refined to haplotypes form the northern parts of South Africa. The duplications observed at DYS481, and DYS612 in the Venda and Afrikaner groups respectively (Table 4.1) have been previously reported (YHRD, STRbase and Ballantyne et al. 2012), while the duplications at DYS710, DYS518 and DY644 and the triplication at DYS385 (10/11/15) we observed herein for the first time.

Bimodal non-overlapping allele frequency distributions within African and non-African groups were most prominent for DYS644 as previously observed by D'Amato et al. 2010. In total, 92% of DYS644 alleles ≤ 19 repeats were distributed across English, Afrikaner, Indian and the admixed groups (Coloured and Griequa), while alleles larger than 19 repeats (either 1bp del variants or microvariants) were represented proportionally as 82.4% in African groups, 15.4% in admixed and 2.2% across English, Afrikaner and Indian groups. Locus DYS504 presented a clear indication of model alleles between the different population groups (Supplementary 4 Figure 8). The DYS504 alleles 16-17 were largely observed amongst the English European, allele 15 for the Indian and alleles 12-13 in African Bantu groups. It was therefore noteworthy to observe the larger rare alleles 19 and 20 of DYS504 was only observed in admixed coloured and Bantu groups. Rare alleles that were only observed in non-African populations were identified at DYS 710 (alleles 41 and 41.2), for DYS518 (alleles 30-33) and at for locus DYS626 (alleles 24 and 36). The rapidly mutating loci DYS518, DYS612, DYS626 and DYS449 had generally a more homogenous distribution of allele size across population groups. (Table 4.1 and Supplementary 4 Figures 2; 6;7 and 11).

5.3.2 Forensic parameters and gene diversity

The 10 markers of the UniQTyper™ Y-10 were assessed with regard to their discrimination capacity (DC), haplotype diversity (HD) and random match probability (MP) for the 15 populations (Table 4.2). Of the 2201 haplotypes, 81.8% were identified as unique. In all populations, with exception to Xhosa and Nama groups a DC (> 0.90) was obtained (Table 5.2). The English, Griequa and Ndebele had the highest DC (~1.0), followed by the Indian, Tswana and Pedi (DC ~0.96), then S.Sotho, Afrikaner, Coloured, Venda and Tsonga (DC > 0.94) and finally Swazi and Zulu (DC~ 0.92). The Xhosa population which showed the largest amount of non-unique haplotypes had the lowest DC (~0.79).

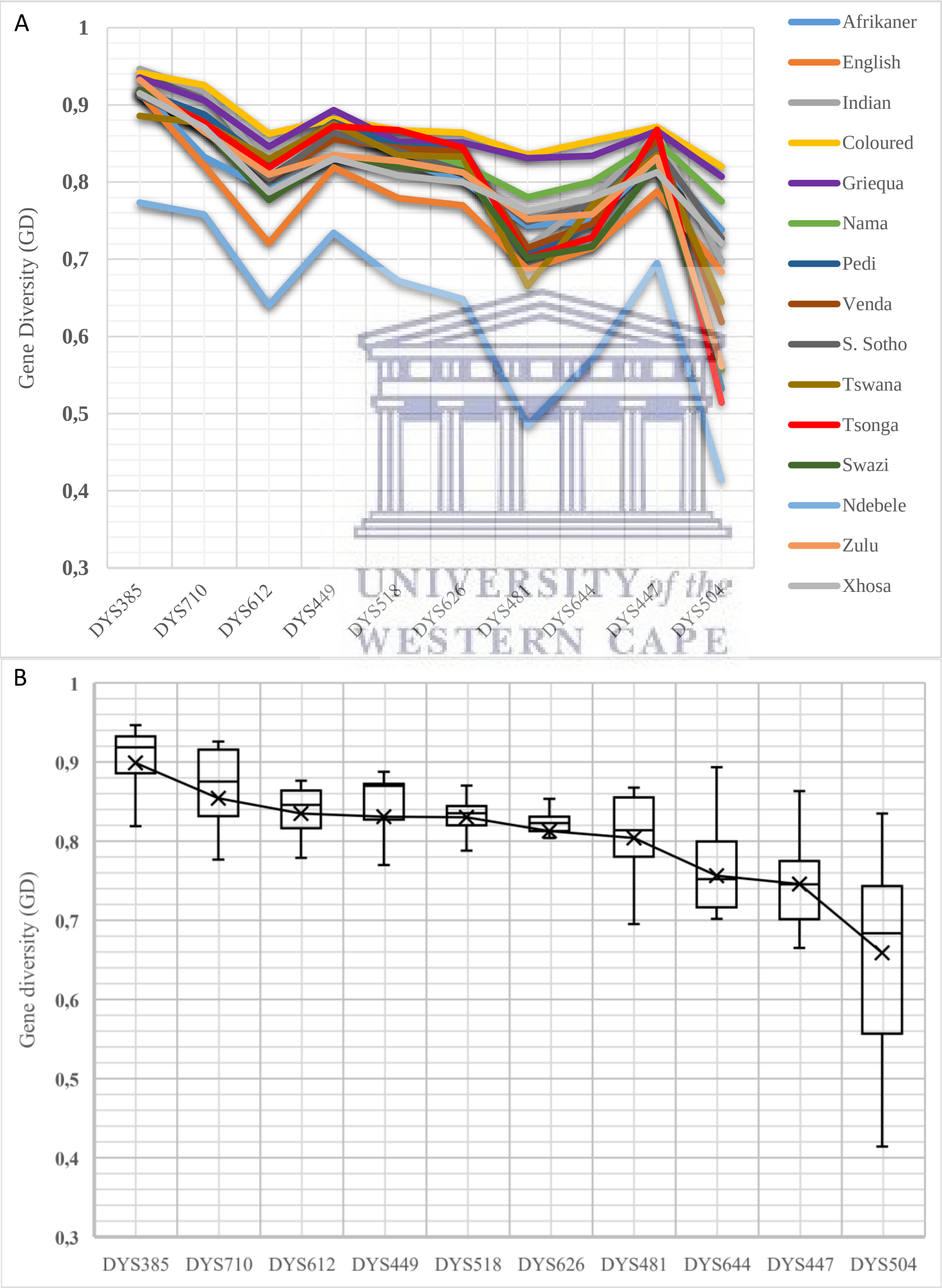
Table 4.2: Forensic summary statistic across the 15 population groups.

	Overall	Afrikaner	English	Indian	Coloured	Griequa	Nama	Pedi	Venda	S. Sotho	Tswana	Tsonga	Swazi	Ndebele	Zulu	Xhosa
Sample size (N)	2201	161	111	104	500	68	47	198	122	70	99	118	104	16	180	303
N.Hap	1950	152	111	100	469	68	42	191	115	67	96	111	96	16	166	240
n=1 (unique)	1801	137	111	93	423	58	33	173	107	59	83	102	78	12	138	194
HD	0.9995	0.9992	1.0000	0.9993	0.9996	1.0000	0.9944	0.9996	0.9988	0.9975	0.9994	0.9988	0.9981	1.0000	0.9982	0.9967
DC	0.8860	0.9441	1.0000	0.9615	0.9380	1.0000	0.8936	0.9646	0.9426	0.9571	0.9697	0.9407	0.9231	1.0000	0.9222	0.7921
MP	0.0070	0.0070	0.0090	0.0104	0.0024	0.0147	0.0267	0.0055	0.0094	0.0167	0.0107	0.0096	0.0115	0.0625	0.0073	0.0066

The Gene diversity (GD) was analysed as an average across populations, the highest GD was obtained for DYS385ab (0.8991 ± 0.00487) of which values (>0.9) were obtained in 11 populations (Figure 4.3A-B). Locus DYS385ab sowed a total of 71 different haplotypes for which the frequencies are presented in Table 4.1 and Supplementary 4 Figure 12). The most frequent haplotype (11,14) was repeated 226 times across population Afrikaner (27%), English (19%), Indian (6%), Admixed (44%) and 8% across African populations. This was on contrary the second most frequent haplotype (16,17) which was repeated 199 times and shared 2,5% across Afrikaner and Indian haplotypes and absent within the English group. This unique distribution was also observed for the Pseudohomozygous haplotype (11,11) which was largely observed to African and Admixed populations (Table 5.1).

The GD values at DYS710 were (> 0.9) in Afrikaner, English, Indian, Coloured and Griequa populations, (> 0.8) in 8 of the studied populations and (>0.7) in the Swazi and Ndebele groups. According to the ranked GD values (Figure 4.3B), the rapidly mutation loci DYS612, DYS449, DYS518, DYS626 followed in descending order with GD values (>0.8). According to the overall performances across populations, lower GD values were obtained for DYS481 (0.804 ± 0.06); DYS644 (0.756 ± 0.09); DYS447 (0.745 ± 0.05) followed by DYS504 (0.659 ± 0.12) in descending order. Although, the GD values for the lower ranked loci (DYS447, DYS644 and DYS504) were somewhat variable between the different populations (Figure 4.3A). In general, all GD values were above 0.5 with an exception to the Ndebele group which was most likely due to the poor population coverage.

Figure 4.3: Ranked gene diversity (GD) for the 10 Y-STRs across 15 South African populations. For the computation of GD, DYS385ab was analysed as a single marker. A) gene diversities ranked across loci per population, B) Box and Whisker blot for ranked gene diversities across each locus (X marks the mean value across population groups, whiskers indicate the standard error).



5.3.3 Shared haplotypes

A total of 1801 unique haplotypes were detected amongst the 2201 individuals. The proportion of non-unique haplotypes within a single population group was largest for the Xhosa (~21%) and Nama (~11%), followed by the Zulu and Swazi with the average of (~7.7%) non-unique cases. The most frequent haplotype (H1) was repeated n=17 times distributed between the population groups Zulu (n=7), S.Sotho (n=4), Xhosa (n=3), Swati (n=2) and Ndebele (n=1) (Table 4.3). For the highly shared haplotypes n=13 (H2) and n=9 (H3), majority was from by the Xhosa population which contributed n=10 and n=7 haplotypes respectively (Table 4.3). The Xhosa group also show repeated haplotypes n=7 (H7) and n=5 (H11) that were not observed in other populations and were host to the highest number shared between two individuals within any given group (supplementary 4 Table 1). Intrapopulation n=2 shared haplotypes were not observed in the Griequa, S.Sotho and Ndebele, rather all its n=2 shared events occurred between groups (supplementary 4 Table 1). The English were the only population with an absence of shared haplotypes from the comparisons within and between groups.

For the Network analysis, we used haplotypes differing by 1 step mutation from the most highly shared haplotypes to investigate 1) marker contribution to the mutations, 2) geneflow between populations and 3) common ancestry. In Figure 4.3, we produce a network analysis for the n=17 shared haplotypes (H1) and a total of 17 haplotypes found one-step mutation away (Table 4.4). From this network, all 9 loci contributed a total 11 mutations defining the haplotypes annotated as H1- (1-11), for which 50% were contributed by the rapidly mutating markers. This cluster of related haplotypes which was shared between the 6 Bantu speaking groups and in the Admixed (coloured) provides evidence of geneflow and/or shared paternal ancestry. Similarly, in Figure 4.4 for the network between the shared haplotypes n=13 (H2) and a total of 16 haplotypes one mutation away, we identify that 8 out of 9 loci contributed a total of 10 mutations across the haplotypes annotated as H2-(1-10). For this network the rapidly mutating markers again contributed 50% of the mutation events. Interestingly, the H3 cluster (n=7) of Figure 4.4 shared the same one step allele mutations at respective loci in comparison to the mutations for the H2 cluster. These two clusters were separated from each other by two admixed (coloured) haplotypes (H2-1 and H2-4) defined by mutations at DYS710 and DYS644 (Table 4.4, Figure 4.4). From these networks it was particularly noteworthy to observe loci DYS710, DYS644, DYS447 and DYS504 provided one mutation step away from the highly shared clusters proportionally to the rapidly mutating markers.

Table 4.3: Summary of shared haplotype distributions within and between population groups.

			Population group														
Shared Haplotypes	**Number of Observations**	**Haplotype Name**	Afrikaner	English	Indian	Coloured	Griequa	Nama	Pedi	Venda	S. Sotho	Tswana	Tsonga	Swazi	Ndebele	Zulu	Xhosa
n=17	1	H1									4			2	1	7	3
n=13	1	H2							1			1				1	10
n=9	1	H3				1										1	7
n=8	2	H4				7		1									
		H5				2											6
n=7	2	H6				2										1	4
		H7															7
n=5	4	H8				1								3		1	
		H9												2		1	2
		H10													1		4
		H11															5
n=4	4	H12			1	1											2
		H13					1		1		1						1
		H14									1			1		1	1
		H15									1	1				2	

Table 4.4: Highly shared haplotypes n=17 (H1), n=13 (H2) and n=7 (H3) with respective one-step mutations used for Network analysis. The Y-SNP information was either provided for a single representative haplotype within a given cluster or haplotypes 1 step away from the shared cluster due to data availability. Locus DYS385 was excluded for the network analysis.

	Haplotype	**N.Hap**	**DYS710**	**DYS518**	**DYS644**	**DYS612**	**DYS626**	**DYS504**	**DYS481**	**DYS447**	**DYS449**	**Y-SNP (mutation)**
Network (Figure 4.4)	**H1**	n=17	33,2	38	22,4	33	29	13	26	25	28	E1b1a8a (U209)
	H1-1	n=2	32,2	38	22,4	33	29	13	26	25	28	E1b1a8a (U209)
	H1-2	n=2	33,2	37	22,4	33	29	13	26	25	28	
	H1-3	n=3	33,2	39	22,4	33	29	13	26	25	28	
	H1-4	n=1	33,2	38	23,4	33	29	13	26	25	28	
	H1-5	n=1	33,2	38	22,4	34	29	13	26	25	28	
	H1-6	n=1	33,2	38	22,4	33	28	13	26	25	28	
	H1-7	n=1	33,2	38	22,4	33	29	12	26	25	28	E1b1a8a (U209)
	H1-8	n=1	33,2	38	22,4	33	29	13	25	25	28	E1b1a8a (U209)
	H1-9	n=1	33,2	38	22,4	33	29	13	26	26	28	
	H1-10	n=3	33,2	38	22,4	33	29	13	26	25	29	
	H1-11	n=1	33,2	38	22,4	33	29	13	26	25	27	E1b1a8a (U209)
Network (Figure 4.5)	**H2**	n=13	36	44	24,4	27	28	14	25	23	28	E1b1a8a (U209)
	H2-1	n=1	35	44	24,4	27	28	14	25	23	28	
	H2-2	n=2	37	44	24,4	27	28	14	25	23	28	
	H2-3	n=4	36	43	24,4	27	28	14	25	23	28	
	H2-4	n=1	36	44	25,5	27	28	14	25	23	28	
	H2-5	n=2	36	44	23,4	27	28	14	25	23	28	
	H2-6	n=2	36	44	24,4	28	28	14	25	23	28	
	H2-7	n=1	36	44	24,4	27	27	14	25	23	28	
	H2-8	n=1	36	44	24,4	27	28	15	25	23	28	
	H2-9	n=1	36	44	24,4	27	28	14	25	23	27	
	H2-10	n=1	36	44	24,4	27	28	14	25	23	29	
	H3	n=7	35	44	25,4	27	28	14	25	23	28	
	H3-1	n=3	35	43	25,4	27	28	14	25	23	28	E1b1a8a (U209)
	H3-2	n=2	35	45	25,4	27	28	14	25	23	28	
	H3-3	n=1	35	44	25,4	27	28	15	25	23	28	
	H3-4	n=1	35	44	25,4	27	28	14	25	23	29	

5.3.4 Population structure

5.3.4.1 Pairwise population difference

On a population level, according to Figure 4.5 the strongest genetic structure was observed with average R_{st} (0.48 *sd ± 0.03*) (Table 4.5) between the African populations (Xhosa, Zulu, Ndebele, Swazi, Tsonga, Twana, S.Sotho, Venda, Pedi) and Non-African population groups (Afrikaner, English and Indian) . An (R_{st} ~0.5) obtained by comparing the English to the African population suggest the capacity of the panel to distinctively separate these groups genetically. The genetic distances between the non-African populations were generally much smaller, the largest was obtained between the English and Indian (R_{st} ~0.1236) and in comparison ~10 fold smaller between the English and Afrikaner (R_{st}~0.012). The average (R_{st} = 0.21 *sd±0.02*) obtained by comparing the Admixed (coloured) to African populations was also significantly larger in comparison to its distance to the Griequa and Nama groups. Upon the AMOVA analysis its determined that 94.9% of the overall variation was obtained within populations and 5.1% amongst the various groups.

Figure 4.4: Network analysis between the n=17 (H1) shared haplotypes and a total of 17 haplotypes one-step mutation away.

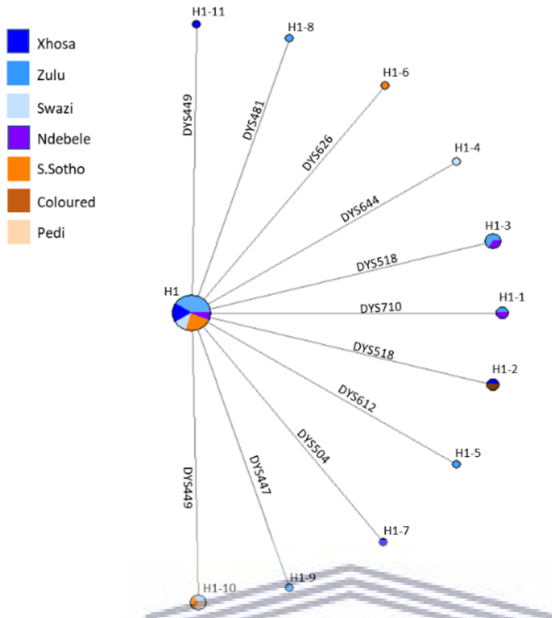


Figure 4.5: Network analysis between the n=13 (H2) and n=7 (H3) shared haplotypes and a total of 16 and 7 haplotypes one-step mutation away from each cluster respectively.

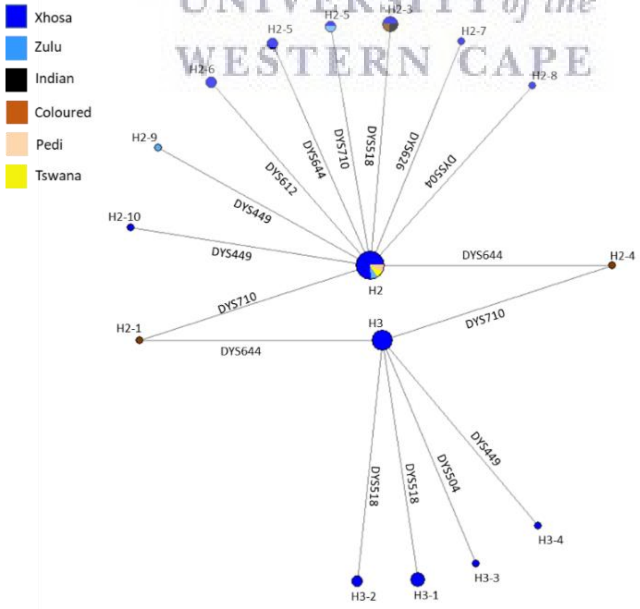


Figure 4.6: Genetic pairwise R_{st} distances between all 15 populations.

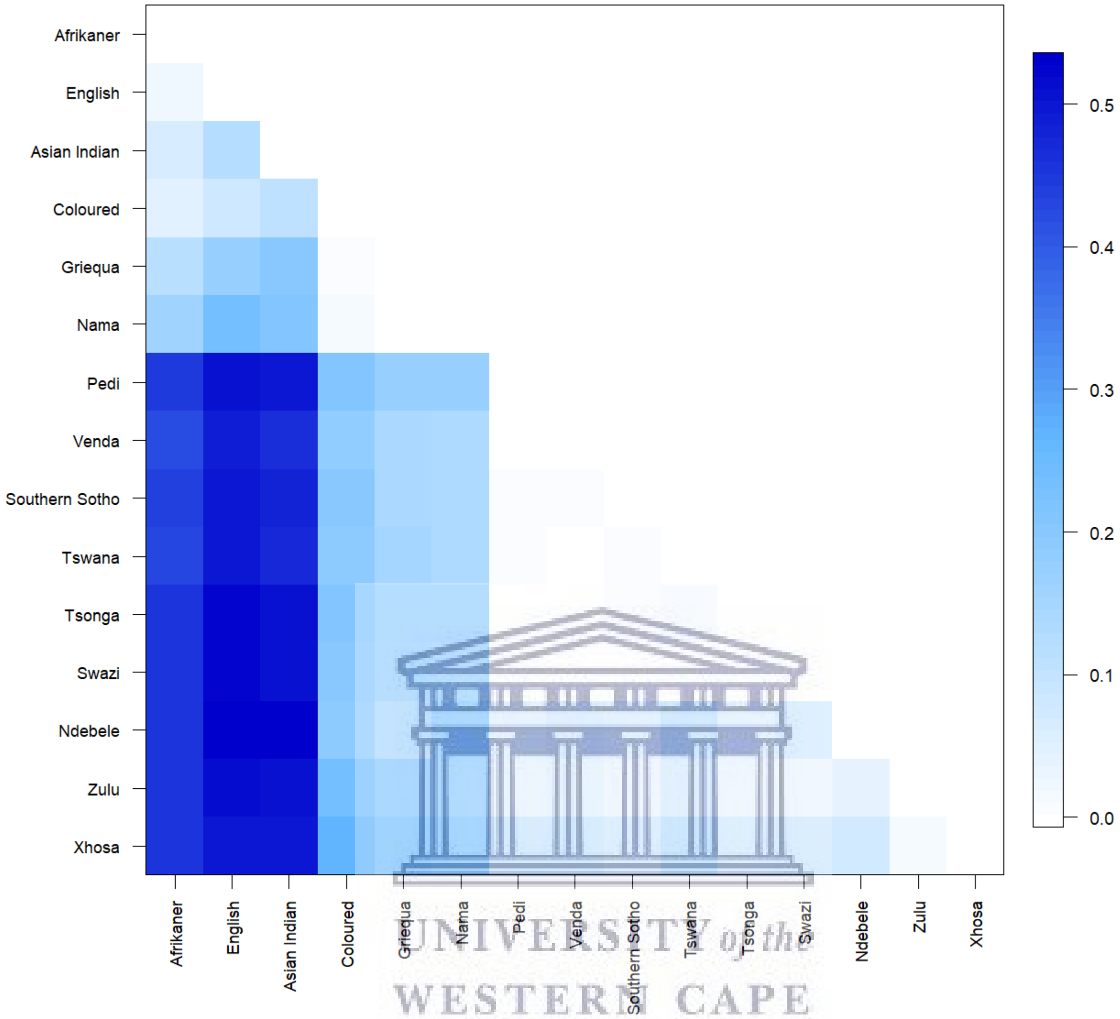


Table 4.5: Pairwise genetic distance (R_{st}) below diagnol and respective (P-values) above diagnol.

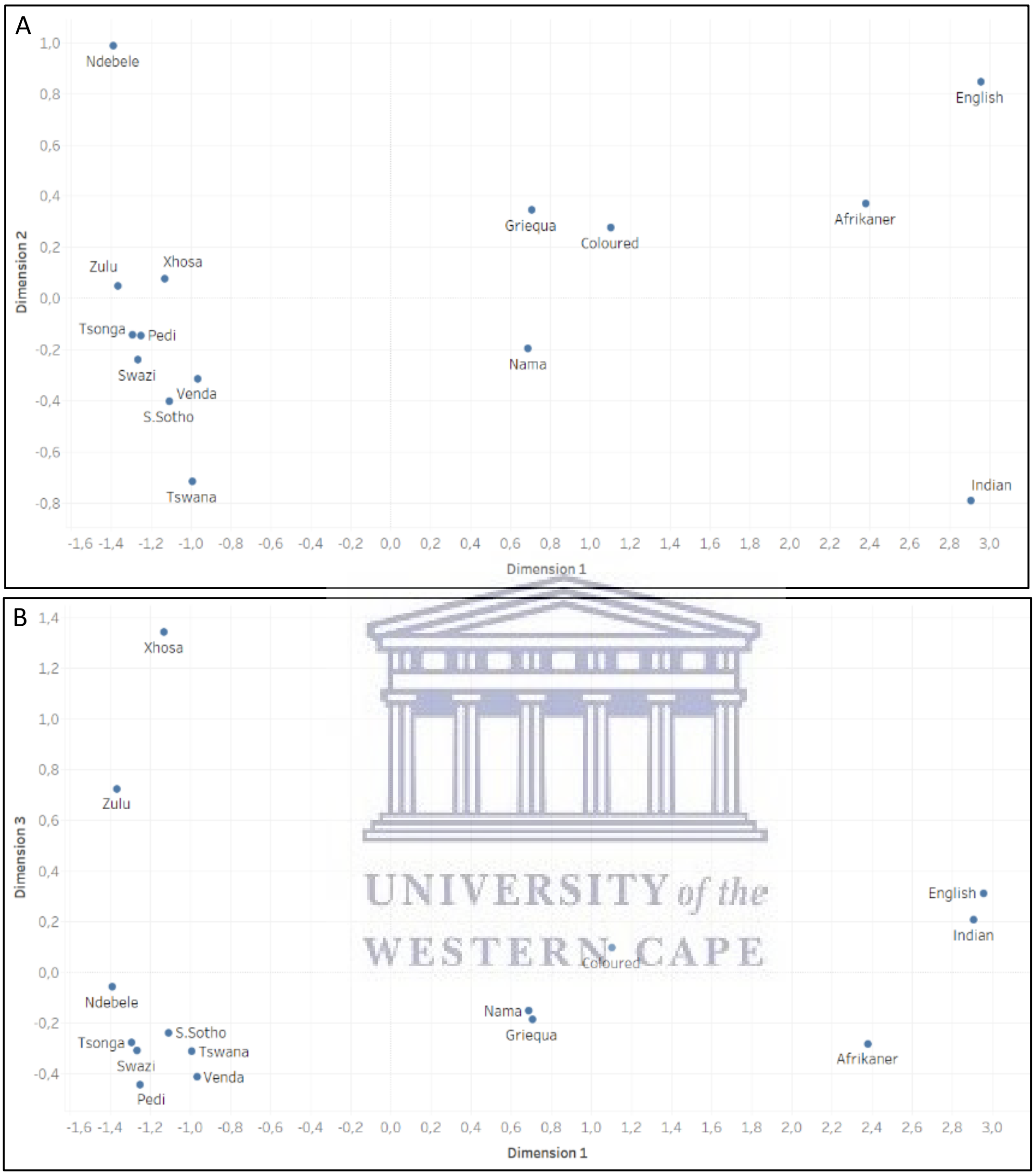
Sinificant R_{st} values ($P = 0.05$) are indicated as (+) and non-sinificant values as (-).

	Afrikaner	English	Indian	Coloured	Griequa	Nama	Pedi	Venda	S.Sotho	Tswana	Tsonga	Swazi	Ndebele	Zulu	Xhosa
Afrikaner		+	+	+	+	+	+	+	+	+	+	+	+	+	+
English	0.01856		+	+	+	+	+	+	+	+	+	+	+	+	+
Indian	0.06449	0.12358		+	+	+	+	+	+	+	+	+	+	+	+
Coloured	0.04621	0.08012	0.10683		-	-	+	+	+	+	+	+	+	+	+
Griequa	0.11982	0.17727	0.19645	0.00782		-	+	+	+	+	+	+	+	+	+
Nama	0.15449	0.23312	0.21278	0.01669	-0.00093		+	+	+	+	+	+	+	+	+
Pedi	0.44897	0.51018	0.50074	0.21310	0.17386	0.17226		-	-	-	-	-	+	+	+
Venda	0.41912	0.48730	0.46567	0.18370	0.13942	0.13588	-0.00316		-	-	-	-	+	+	+
S.Sotho	0.43683	0.50115	0.48163	0.19928	0.14449	0.13454	0.00656	0.00669		-	-	-	+	+	+
Tswana	0.42857	0.49916	0.47387	0.19092	0.14731	0.13133	0.00627	0.00095	0.00333		+	-	+	+	+
Tsonga	0.45565	0.52360	0.50818	0.20544	0.17224	0.17412	-0.00260	-0.00050	0.00369	0.01339		-	-	+	+
Swazi	0.45217	0.52164	0.50227	0.20292	0.16565	0.15857	-0.00143	-0.00109	-0.00717	-0.00040	-0.00325		+	+	+
Ndebele	0.45574	0.53496	0.53578	0.19060	0.14023	0.19495	0.04820	0.06266	0.05959	0.10248	0.04584	0.07141		-	+
Zulu	0.45920	0.51332	0.50880	0.23284	0.19769	0.18525	0.03930	0.04605	0.03047	0.06382	0.02901	0.03267	0.04272		+
Xhosa	0.45776	0.50118	0.49634	0.26738	0.22941	0.20861	0.08820	0.09247	0.07644	0.11309	0.07285	0.07774	0.07497	0.01197	

5.3.4.2 Multidimensional scaling

In comparison, Multidimensional Scaling (MDS) analysis was computed using the Slatkin's linearized R_{st} distances. According to the visual scree plot a clear "elbow" indicated that 3 dimensions were optimal, upon computation a Kruskal's stress value (0,03289) was obtained with an RSQ value (0,99674) representing the proportion of variance of the scaled data. In Figure 4.6A the first MDS component clearly separates the African and non-African population groups on the far left and right, while the populations (Coloured, Griequa and Nama) for which we expect a higher degree of admixture grouped centrally. The MDS plots confirmed a closer relationship between the English, Afrikaner and Indian groups originating mainly from the Western-Eastern Cape. The MDS analysis in the third dimension also illustrated distinct differences between the Nguni groups (Xhosa and Zulu) sampled from the Western and Eastern Cape compared to the bantu groups (Pedi, Venda, Tsonga and Tswana) which originated mainly from further North of South Africa (Figure 4.6B).

Figure 4.7: Multidimensional scaling based on slatkin's linearized R_{st} values obtained across the 15 populations for A) two dimensional and B) tree dimensional comparison



5.4 Discussion

In this chapter, the largest known compilation of Y-chromosomal haplotypes for South Africa (SA) was analysed to evaluating the forensic performance of the UniQTyper™ Y-10. The panel is known to be highly informative within South Africa particularly for maximizing discrimination in Bantu men compared to the Y-filer and Powerplex Y commercial panels (Leat et al. 2007, Cloete et al. 2010 and D’Amato et al. 2009, 2010 and 2011). Although successor Y-STR commercial kits proposed a better value proposition due to an increase in panel size and incorporation of rapidly mutating markers, their performances on a large scale in South Africa remains undetermined. Evidently, from the data available, overcoming the degree of non-unique haplotypes within the bantu Xhosa population still propose a challenge for even the highly discriminating commercial panels (Purps et al. 2014). In our experience the Xhosa group which generally presents the lowest Y-chromosome diversity demands the use of more rapidly mutating markers to maximize haplotype resolution as achieved in Ballantyne et al.2014. Providing this, we conduct for the first time a large Y-STR survey across diverse South African populations to comprehensively asses the informativeness of 10 Y-STRs of which 4 are rapidly mutating.

5.4.1 Genetic diversity and allele variation

In this study, the respective gene diversities obtained for the rapidly mutating markers illustrated its value within all populations studied. However, locus DYS710 due to its larger number of alleles clearly outranked their performances as previously observed (D’Amato et al. 2009 and 2010). Although initially characterized by Leat et al. 2006, DYS710 has only been valued elsewhere in populations from Tunisia (Makki-Rmida et al. 2015) and Chinese Han (Zhang et al. 2012). In comparison to the data available, we describe a significant degree of novel allele variation at DYS710 from the northern South African regions particularly in the form of duplications. Considering these DYS710 duplications and the allele 45 only reported previously in Tunisia (Makki-Rmida et al. 2015) a total of 41 alleles is observed to date. For the more frequently studied Y-STRs of the panel, some of the rare atypical allelic patterns have been reported elsewhere. The DYS626 (31,33) duplication observed in Afrikaner was previously reported in Ballantyne al. 2012 amongst African American populations. Similarly, the rare DYS481 duplication (25,26) observed in the Pedi has been reported by Purps et al. 2012 amongst the Kenya Maasai and Japanese.

Similar rare cases observed in the Pedi were however not previously reported to STRbase and YHRD (accessed 31 November 2018), this included the triplications at DYS385 (10,11,13), the DYS385 haplotype (13.2, 16) and the duplication at DYS518 (41,42). Evidently, the northern South African regions particularly in the Mpumalanga and Limpopo provinces were host to most of the novelties described, the most significant being the unique association between the DYS710 duplications and the DYS644 X.3 microvariants sequenced in Chapter 3. The DYS644 non-overlapping allele frequency distribution between African and non-African groups also complimented its value for ancestry inference and as a diagnostic marker for haplogroup E (D’Amato et al. 2009-2011). There is also ancestry informative potential for DYS504 due to its evident model alleles in different continental groups. However, their value for ancestral inference should be explored further considering Y-SNP haplogroup typing. Although, DYS644 and DYS504 were ranked lower in gene diversity, its informativeness for ancestral and potential geographical inferences could indeed be of great value in forensics (D’Amato et al. 2009).

5.4.2 Population specific diversities

In this study, the Xhosa population which had the highest degree of shared haplotypes also presented the lowest DC. The value obtained was significantly lower compared to previous observations having herein analysed a larger sample size with a wider regional coverage. In a comparative analysis between the Powerplex Y-23 (Promega) and the UniQTyper™ Y-10 panel the DC obtained was (86%) and (81%) respectively using a subset of N=95 Xhosa individuals genotyped in Purps et al. 2012. This recurrent lower level of Y-chromosome diversity in the Xhosa may be attributed to various cultural and historical factors. The South-East Bantu groups such as the Xhosa and Zulu are known for strict patrilocal marriage practices (Mesthrie et al. 2002), whereby the married couple settle in the same geographical territory as the male's paternal ancestors. Furthermore, the Xhosa traditional practice of manhood initiation (Ulwaluko) which involves male circumcision is associated with high morbidity and mortality rates (Froneman and Kapp 2017). The negative outcome of these practices also has a direct impact on the male status in society, which may have altogether contributed to reduced reproductive success (Rueden and Jaggi 2016). A reduction in the male population size due to the mass famine amongst the Xhosa settlers in the Eastern Cape could also be an underlying contributor. A historic event referred to as the "Cattle Killings" (Peires 1989) presents a potential bottleneck event which resulted in genetic drift. From April 1856 to February 1857, the Xhosa killed ~400, 000 cattle and destroyed all crops which lead to mass human starvation and death (Peires 1989). It's believed this was an instruction from the spiritual ancestors in order bring an end to the colonial rule. The Xhosa population in the Eastern Cape is recorded to have declined in one year from 104,721 to 37,679 individuals for which more than 40 000 where known to have starved to death (Mclean 1866).

The highest Y-chromosome diversity was seen within the immigrant European English population which is largely an outbred population occupying the Western Cape. The complete haplotype resolution obtained in the English was also observed previously, which confirms the value of the panel also within immigrant Europeans compared to the Y-filer and Powerplex Y markers (D'Amato et al. 2011). In the immigrant Indian population, the discriminatory capacity improved slightly with the increase in sample size in comparison to (D'Amato et al. 2011). Although full haplotype resolution was obtained in the Ndebele, a re-evaluation with a larger sample size would be required as haplotypes were shared with the Zulu and Xhosa. The Ndebele people are split into northern (Zimbabwe) and southern groups (Limpopo, Gauteng and Mapumalanga) and speak a district variation of isiNdebele which similar to Xhosa and Zulu. According to historical records the Ndebele groups are genealogically related as they are descendant for the same ancestral king (Wilkes et al. 2001). For the Griequa and Nama groups particularly we would also need to improve the population size as we expect an increase in shared haplotypes across these groups due to the admixture influence.

The population pairwise comparisons showed that the migrant populations (English, Afrikaner and Indian) were genetically distinct from populations of African ancestry. This was evident by a clear separation in the MDS analysis which was similar to observation previously made using the minimal haplotype (D'Amato et al. 2008). In our MDS analysis the admixed Coloured and Griequa groups were also separated from the African groups in the first MDS component. The Coloured and Griequa which were relatively similar have been known to receive significant European influence in Southern Africa (Petersen et al. 2013). Furthermore, although the sample size was relatively small, the Nama due to its known Eurasian influence (Pickrell et al. 2014) also clustered with the admixed groups as observed previously with the Powerplex Y panel (Schlebusch 2010). These findings are indeed testimony to the balanced marker composition of the UniQTyper™ Y-10 panel, while it contains a majority of highly polymorphic markers the system can still be of value for providing information about ancestry and history.

5.5 Conclusion

In summary, the UniQTyper™ Y-10 kit provided discriminatory capabilities substantial for forensic practices across diverse South African populations. The exception was seen for the Xhosa, whereby the influence of traditional and historical factors on overall Y-chromosome diversity may suggest the need to employ additional rapidly mutating markers. Providing coverage of scarcely studies population in the northern South African regions we identified a significant degree of novel variations. This suggests that a large amount genetic variation may still be undetermined and that scope to improve match probability estimates may still exist. The Y-STRs studied herein also revealed a remarkable degree of population structure for which its potential for ancestral and geographical inferences would be highly valued when a suspect in unknown. Altogether, we can recommend the UniQTyper™ Y-10 panel for forensic investigations of sexual assault and for genealogy studies in South Africa.

Table 4.1: Allele range and frequencies across the 15 population groups for 10 Y-STR markers.

Locus	Allele	Afrikaner	English	Indian	Coloured	Griequa	Nama	Pedi	Venda	S. Sotho	Tswana	Tsonga	Swazi	Ndebele	Zulu	Xhosa
DYS710																
	26				0,0040		0,0213									
	28				0,0040	0,0147										
	28,2	0,0062	0,0180		0,0060	0,0147		0,0051								0,0033
	29	0,0124	0,0270		0,0060			0,0051			0,0101	0,0085				
	29,2	0,0124	0,0270	0,0096	0,0080						0,0101		0,0096			
	30	0,0062	0,0090	0,0289	0,0180	0,0147	0,0213	0,0101	0,0246						0,0056	0,0033
	30,2	0,0497	0,0270	0,0289	0,0220	0,0588		0,0455	0,0164	0,0143	0,0101	0,0339			0,0333	0,0198
	31	0,0497		0,0481	0,0580	0,0441	0,0426		0,0164		0,0101	0,0085	0,0096		0,0111	0,0066
	31,2	0,0870	0,0090	0,1058	0,0700	0,0441	0,0426	0,0404	0,0328	0,0143	0,0505	0,0085	0,0289			0,0099
	32	0,0621	0,0180	0,0962	0,0620	0,1177	0,0638	0,1616	0,1312	0,1143	0,2121	0,1271	0,1442	0,0625	0,1111	0,0429
	32,2	0,0435	0,0721	0,0769	0,0620	0,0735	0,0638	0,0960	0,1312	0,1143	0,1717	0,1017	0,0289	0,0625	0,0722	0,0528
	33	0,0745	0,0631	0,0481	0,0680	0,1618	0,0638	0,1263	0,1230	0,2286	0,1212	0,1525	0,1250		0,1000	0,0594
	33,2	0,1056	0,0991	0,1731	0,1580	0,1324	0,3192	0,2323	0,3115	0,2571	0,1414	0,2288	0,4135	0,5625	0,3500	0,3828
	34	0,0932	0,1261	0,0385	0,0880	0,0294	0,0213	0,0606	0,0902	0,0143	0,0707	0,0763	0,0192		0,0333	0,0198
	34,2	0,0808	0,0721	0,0673	0,0520	0,0735	0,0851	0,0707	0,0492	0,0857	0,0404	0,0678	0,0962	0,0625	0,0722	0,0726
	35	0,1367	0,1532	0,0673	0,0900	0,1177	0,0851	0,0253	0,0164	0,0571	0,0303	0,0170	0,0096	0,0625	0,0278	0,0891
	35,2	0,0621	0,0811	0,0385	0,0480	0,0294	0,1064	0,0556	0,0246	0,0143	0,0606	0,0509	0,0577	0,0625	0,0500	0,0627
	36	0,0745	0,0901	0,0192	0,0740	0,0441	0,0426	0,0202		0,0143	0,0303	0,0763	0,0192	0,1250	0,0833	0,1320
	36,2	0,0186	0,0451	0,0673	0,0240			0,0051	0,0246	0,0143	0,0101	0,0170	0,0096		0,0111	0,0099
	37	0,0124	0,0270	0,0192	0,0180			0,0101				0,0170	0,0192		0,0167	0,0297
	37,2	0,0062		0,0481	0,0100			0,0101	0,0082						0,0111	
	38	0,0062	0,0180		0,0160	0,0147						0,0085			0,0056	0,0033
	38,2		0,0090	0,0096	0,0060											
	39				0,0040											
	39,2				0,0020											
	40				0,0080											
	40,2				0,0020											
	41				0,0060											
	41,2			0,0096												
	42		0,0090													
	26,33.2						0,0213									
	30,30.2					0,0147										
	30.2,31.2							0,0051								
	30.2,32				0,0020					0,0143						
	30.2,33				0,0040			0,0152		0,0286	0,0101					
	30.2,34									0,0143						
	31,33										0,0101					
	32.2,33.2														0,0056	
	33,37												0,0096			
DYS518																
	30		0,0090													
	32		0,0090		0,0020											
	33	0,0062			0,0040											
	34	0,0062			0,0140		0,0213	0,0051	0,0082	0,0143	0,0101					
	35		0,0090	0,0096	0,0480	0,0294	0,0213					0,0085	0,0096		0,0056	0,0066
	36	0,0373	0,0901	0,0673	0,0340	0,0588	0,0213	0,0253	0,0082	0,0857	0,0303	0,0424	0,0096	0,1250	0,0111	
	37	0,1926	0,2072	0,1058	0,1520	0,2353	0,0426	0,1010	0,0984	0,0143	0,0707	0,1017	0,0673		0,0444	0,0594
	38	0,2298	0,3423	0,1731	0,1700	0,1765	0,2128	0,1364	0,1557	0,2429	0,0909	0,1186	0,1442	0,3125	0,1556	0,1254
	39	0,2112	0,1802	0,1058	0,1820	0,1618	0,1702	0,2222	0,2541	0,1571	0,2121	0,2627	0,2212	0,3125	0,1722	0,1716
	40	0,1367	0,0901	0,2308	0,1740	0,1177	0,2128	0,1970	0,1557	0,1714	0,2424	0,1695	0,2885	0,1250	0,2778	0,1155
	41	0,1242	0,0541	0,2019	0,1140	0,1177	0,2340	0,1667	0,1230	0,2714	0,2020	0,1525	0,1250	0,1250	0,1556	0,1947
	42	0,0373		0,0385	0,0540	0,0441		0,0960	0,1066	0,0286	0,0606	0,0763	0,0673		0,0500	0,0792
	43	0,0062	0,0090	0,0577	0,0200	0,0294	0,0213	0,0303	0,0820		0,0303	0,0424	0,0385		0,0444	0,0759
	44	0,0124		0,0096	0,0240	0,0147	0,0426	0,0101			0,0303	0,0254	0,0289		0,0556	0,1254
	45				0,0080	0,0147		0,0051		0,0143	0,0202				0,0167	0,0429
	46							0,0051							0,0111	0,0033
	41,42								0,0082							
DYS644																
	10				0,0040											
	12		0,0090		0,0020											
	13	0,0435	0,0270	0,0385	0,0180	0,0147			0,0082	0,0143						0,0033
	14	0,1056	0,0451	0,3462	0,0860	0,0882	0,0851		0,0164	0,0143	0,0202				0,0056	0,0066
	15	0,0497	0,0721	0,2019	0,1080	0,0735	0,1064		0,0246						0,0056	0,0132
	16	0,3789	0,3694	0,2212	0,2020	0,1177	0,0638	0,0202	0,0164		0,0101		0,0192		0,0111	0,0198
	17	0,2919	0,3604	0,0865	0,1760	0,1765	0,0851						0,0096		0,0056	0,0099

	18	0,0062	0,0451	0,0385	0,0200	0,0147					0,0202					0,0033
	19			0,0289	0,0060					0,0143						
	19,4								0,0164			0,0424				0,0033
	20,4				0,0060	0,0147	0,0213				0,0101	0,0170			0,0056	0,0165
	21,4	0,0186			0,0940	0,1029	0,2128	0,3081	0,3115	0,2429	0,2626	0,3136	0,3462	0,1875	0,3444	0,2937
	22,3				0,0020											0,0033
	22,4	0,0683	0,0451	0,0096	0,0940	0,1471	0,0851	0,4091	0,3607	0,3571	0,3030	0,4237	0,3462	0,6875	0,3167	0,2706
	23,3				0,0020		0,0426			0,0429	0,0404	0,0085	0,0096			0,0033
	23,4	0,0186			0,0620	0,0735		0,1768	0,1475	0,1857	0,2525	0,1271	0,2019		0,1278	0,0858
	24,3				0,0200	0,0441	0,1489	0,0253	0,0164	0,0429	0,0202		0,0096	0,0625	0,0056	
	24,4	0,0124	0,0090	0,0289	0,0640	0,0882	0,1277	0,0303	0,0492	0,0143	0,0505	0,0170		0,0625	0,0389	0,1353
	25,3						0,0213									
	25,4	0,0062	0,0180		0,0260	0,0147		0,0152		0,0143	0,0101	0,0424	0,0481		0,1000	0,1188
	26,3											0,0085				
	26,4				0,0080			0,0101	0,0328	0,0571			0,0096		0,0333	0,0099
	27,4							0,0051								
	24.4,25.4					0,0147										0,0033
DYS612																
	25				0,0020											
	26															0,0033
	27	0,0124		0,0096	0,0340	0,0147		0,0303	0,0246	0,0571	0,0202	0,0339	0,0577		0,1389	0,2244
	28				0,0060	0,0147	0,0213	0,0051							0,0222	0,0264
	30				0,0080	0,0294						0,0170				
	31							0,0354	0,0656	0,0143		0,0509	0,0096		0,0389	0,0165
	32	0,0062			0,0100	0,0294	0,0213	0,0202	0,0246	0,0286	0,0101	0,0170			0,0167	0,0066
	33	0,0248	0,0270	0,0192	0,0880	0,0735	0,1064	0,1111	0,0902	0,1429	0,1111	0,1864	0,1827	0,3750	0,1889	0,2145
	34	0,0435	0,0360	0,0962	0,0880	0,1471	0,1064	0,2071	0,1885	0,2286	0,2525	0,1441	0,2212	0,3125	0,1611	0,2343
	35	0,0932	0,0631	0,1539	0,1020	0,1177	0,1064	0,1869	0,2377	0,1429	0,1212	0,1780	0,2115		0,1611	0,0528
	36	0,3106	0,3243	0,2596	0,2020	0,1177	0,1915	0,1717	0,1475	0,1000	0,2020	0,1441	0,0962	0,1250	0,1389	0,0792
	37	0,2360	0,2523	0,2308	0,2000	0,2941	0,1489	0,1061	0,0984	0,1000	0,1212	0,0763	0,0385	0,0625	0,0389	0,0759
	38	0,1677	0,1982	0,1635	0,1360	0,0441	0,1064	0,0808	0,0574	0,1286	0,1414	0,1186	0,1539	0,0625	0,0667	0,0429
	39	0,0808	0,0811	0,0481	0,1020	0,0882	0,1277	0,0404	0,0656	0,0429	0,0202	0,0170	0,0289	0,0625	0,0167	0,0165
	40	0,0248	0,0180	0,0192	0,0180	0,0294	0,0426	0,0051				0,0170			0,0111	0,0033
	41				0,0020		0,0213									0,0033
DYS626																
	24	0,0062	0,0090													
	25	0,0435	0,0541		0,0640	0,0735	0,1277	0,0051	0,0164		0,0202			0,0625	0,0111	0,0198
	26	0,0124	0,0090		0,0060	0,0147		0,0051	0,0082	0,0286						
	27	0,0186	0,0180	0,0289	0,0140	0,0147		0,0152	0,0164	0,0143	0,0101		0,0096		0,0111	0,0264
	28	0,0745	0,1171	0,0481	0,0760	0,0588	0,0426	0,0505		0,1000	0,0404	0,0678	0,0962	0,0625	0,1556	0,2376
	29	0,2671	0,2252	0,0769	0,2280	0,2206	0,0638	0,2121	0,1967	0,2143	0,2020	0,2542	0,2308	0,5625	0,2778	0,2607
	30	0,2236	0,2523	0,2115	0,1980	0,2647	0,0426	0,2475	0,1803	0,1571	0,2222	0,2288	0,2019	0,1250	0,1833	0,0792
	31	0,1242	0,2072	0,1058	0,1480	0,0588	0,1702	0,2071	0,2951	0,2143	0,1616	0,1610	0,2308	0,1250	0,1833	0,0891
	32	0,1553	0,0451	0,2789	0,1280	0,1765	0,2766	0,1111	0,1475	0,1571	0,1414	0,1441	0,1635	0,0625	0,1333	0,2079
	33	0,0373	0,0451	0,1346	0,0580	0,0735	0,0426	0,1010	0,1230	0,0857	0,1616	0,1102	0,0289		0,0278	0,0627
	34	0,0124	0,0180	0,0962	0,0580	0,0294	0,0213	0,0354	0,0082	0,0143	0,0101	0,0170	0,0385		0,0111	0,0165
	35	0,0124		0,0192	0,0080			0,0051	0,0082		0,0101					
	36	0,0062														
	31,33	0,0062														
	NULL				0,0140	0,0147	0,2128	0,0051		0,0143	0,0202	0,0170			0,0056	
DYS504																
	11				0,0020	0,0147		0,0152		0,0714			0,0096		0,0056	0,0132
	12	0,0186			0,0640	0,1324	0,0426	0,1717	0,2295	0,2857	0,3030	0,2288	0,2404	0,1250	0,1556	0,2838
	13	0,0683	0,0360	0,0192	0,1840	0,2206	0,2340	0,6566	0,5574	0,4143	0,5051	0,6525	0,6154	0,7500	0,6167	0,3630
	14	0,1553	0,0541	0,0865	0,1400	0,0294	0,1064	0,0707	0,1230	0,0714	0,0606	0,0848	0,0769		0,1833	0,2541
	15	0,1056	0,0811	0,4231	0,1640	0,1618	0,2340	0,0253	0,0410	0,0429		0,0085	0,0192			0,0264
	16	0,1739	0,2432	0,3077	0,1720	0,2206	0,2553	0,0152	0,0246	0,0143	0,0505	0,0170	0,0096	0,0625	0,0222	0,0297
	17	0,4286	0,4865	0,1442	0,2200	0,1471	0,1064	0,0303	0,0246	0,0857	0,0303		0,0192		0,0111	0,0231
	18	0,0497	0,0991	0,0192	0,0500	0,0441	0,0213	0,0152		0,0143	0,0404	0,0085			0,0056	0,0066
	19				0,0020	0,0147					0,0101		0,0096	0,0625		
	20				0,0020	0,0147										
DYS481																
	19				0,0020					0,0143	0,0101					0,0066
	20	0,0062	0,0180		0,0080		0,0426									
	21	0,0808	0,0721	0,0385	0,0700	0,0441	0,0213						0,0096			0,0033
	22	0,3913	0,4865	0,1058	0,2040	0,1765	0,0638	0,0051			0,0707	0,0170	0,0096		0,0111	0,0132
	23	0,1242	0,1261	0,4712	0,1220	0,1029	0,1277	0,0556	0,0820	0,0286	0,0303	0,1102	0,0385		0,0611	0,0462
	24	0,0870	0,0721	0,1923	0,1120	0,1324	0,1915	0,1616	0,1393	0,1857	0,0909	0,1610	0,1250		0,1222	0,0462
	24,2				0,0040								0,0096		0,0056	
	25	0,1304	0,1351	0,0962	0,1760	0,1765	0,3617	0,2121	0,2131	0,2714	0,2525	0,1780	0,2404	0,3750	0,2500	0,3366
	26	0,0745	0,0270	0,0289	0,1200	0,1029	0,1702	0,1263	0,1885	0,2286	0,2424	0,1356	0,2692	0,2500	0,2778	0,1584
	27	0,0621	0,0541	0,0481	0,0920	0,0588		0,1111	0,1230	0,1000	0,1111	0,0763	0,1154		0,1556	0,2244
	28	0,0311		0,0192	0,0680	0,1618	0,0213	0,2121	0,1230	0,1143	0,1414	0,1695	0,1154	0,3125	0,0833	0,1386
	29	0,0062			0,0060	0,0147		0,0657	0,0410	0,0286	0,0202	0,0678	0,0385	0,0625	0,0111	0,0066
	29,2											0,0085				
	30	0,0062	0,0090		0,0120	0,0294		0,0455	0,0738	0,0286	0,0202	0,0678	0,0289		0,0222	0,0198

	31								0,0082		0,0101	0,0085				
	32				0,0040				0,0082							
	25,26							0,0051								
DYS447																
	17							0,0051								
	19			0,0192												
	20		0,0090	0,0096	0,0020											
	21		0,0180	0,0192	0,0080											
	21,4			0,0096												
	22	0,0124	0,0270	0,0289	0,0360	0,0294	0,0213	0,0101								0,0033
	22,4			0,1539	0,0220		0,1064								0,0056	
	23	0,2174	0,1351	0,1539	0,1380	0,1324	0,0851	0,0455	0,0656	0,0714	0,0404	0,0509	0,0481		0,1556	0,2607
	23,4			0,0385	0,0020											0,0066
	24	0,1677	0,1712	0,1539	0,1360	0,1324	0,1064	0,0657	0,0656	0,0857	0,0606	0,0339	0,0192	0,0625	0,0333	0,0330
	24,3				0,0040											
	24,4						0,0213									
	25	0,3913	0,5045	0,1731	0,2820	0,3235	0,0213	0,3131	0,2869	0,3429	0,2727	0,3898	0,3846	0,4375	0,2778	0,1419
	26	0,1801	0,0901	0,1731	0,2060	0,1765	0,2128	0,3131	0,4180	0,2857	0,4849	0,2203	0,3269	0,3125	0,3111	0,2739
	27	0,0248	0,0451	0,0481	0,1340	0,1324	0,3830	0,2273	0,1393	0,2000	0,1414	0,2542	0,2019	0,1875	0,2056	0,2706
	28	0,0062		0,0096	0,0240	0,0588	0,0426	0,0152	0,0164	0,0143		0,0509	0,0192		0,0111	0,0099
	29			0,0096	0,0060	0,0147										
	30								0,0082							
	32							0,0051								
DYS449																
	23				0,0020											
	24			0,0096	0,0040											
	25	0,0062		0,0096	0,0060	0,0147		0,0202			0,0101				0,0111	0,0033
	26	0,0186	0,0180	0,0577	0,0400	0,0441		0,0152			0,0101	0,0085	0,0096			0,0165
	27	0,0497	0,0541	0,0962	0,0820	0,0294	0,0213	0,1515	0,1639	0,0714	0,0505	0,1441	0,0865	0,0625	0,1278	0,0726
	28	0,1926	0,0901	0,0577	0,1300	0,1324	0,0638	0,1616	0,1475	0,2286	0,2121	0,1356	0,2019	0,6250	0,3222	0,3894
	29	0,2360	0,3423	0,0769	0,1980	0,2353	0,0638	0,1111	0,1312	0,1286	0,1212	0,1356	0,1250	0,0625	0,0889	0,1221
	30	0,2298	0,2883	0,1346	0,1600	0,1029	0,1489	0,0859	0,1230	0,0857	0,0707	0,0932	0,0962		0,0611	0,0264
	31	0,1180	0,1081	0,0962	0,1280	0,1912	0,1915	0,1313	0,1639	0,1000	0,1515	0,1780	0,1154	0,1250	0,0944	0,0990
	32	0,0808	0,0811	0,2404	0,1380	0,1471	0,2766	0,1263	0,0820	0,1571	0,0808	0,1525	0,1827	0,1250	0,1556	0,1881
	33	0,0311	0,0180	0,1346	0,0460	0,0147	0,0213	0,0303	0,0246	0,0286	0,0404	0,0509	0,0192		0,0611	0,0297
	34	0,0124		0,0385	0,0260	0,0147	0,0213	0,0354	0,0246	0,0143	0,0202	0,0254			0,0056	0,0132
	35	0,0248		0,0385	0,0200	0,0147		0,0404	0,0492	0,0429	0,1010	0,0254	0,0385		0,0111	0,0033
	36			0,0096	0,0040		0,1064	0,0758	0,0328	0,0857	0,1212	0,0509	0,0865		0,0611	0,0297
	37				0,0080	0,0588	0,0426	0,0101	0,0574	0,0429	0,0101		0,0289			0,0066
	38				0,0080		0,0426	0,0051		0,0143			0,0096			
DYS385ab																
	9,16	0,0123			0,0096											
	8,14			0,0020												
	8,11							0,0051				0,0254	0,0096			
	7,16				0,0096											
	7,15				0,0096											
	19,20								0,0082							
	19,19			0,0020											0,0056	
	18,20						0,0213	0,0051	0,0082						0,0056	
	18,19	0,0062		0,0020				0,0101			0,0101	0,0085			0,0056	0,0033
	18,18	0,0062		0,0020			0,0213	0,0101	0,0164	0,0143	0,0101					0,0033
	17,21			0,0020			0,0213				0,0101				0,0056	0,0231
	17,20		0,0090	0,0040			0,0426	0,0202				0,0085	0,0192		0,0278	0,0561
	17,19			0,0120	0,0096	0,0441		0,0303	0,0164	0,0571	0,0303	0,0085	0,0481		0,0389	0,0099
	17,18	0,0062	0,0360	0,0300	0,0096	0,0735	0,0213	0,0960	0,1148	0,1143	0,1111	0,1271	0,0673	0,1250	0,0667	0,0495
	17,17			0,0260	0,0192		0,0213	0,0960	0,1230	0,1143	0,0505	0,1017	0,0673		0,0833	0,0561
	16,21			0,0040									0,0096			0,0099
	16,20			0,0280		0,0147	0,0213				0,0101	0,0169	0,0096		0,0611	0,0924
	16,19	0,0123	0,0090	0,0120		0,0147	0,0213	0,0303	0,0328	0,0143	0,0505	0,0169	0,0096		0,0389	0,0165
	16,18	0,0370	0,0270	0,0300	0,0096	0,0147		0,0859	0,0902	0,0286	0,0303	0,1102	0,1538	0,0625	0,1000	0,0231
	16,17	0,0123		0,0640	0,0288	0,0147	0,1064	0,1818	0,1230	0,0714	0,1717	0,1186	0,1154	0,3125	0,1111	0,1056
	16,16	0,0062	0,0090	0,0320	0,0385	0,0588	0,0851	0,0202	0,0738	0,0429	0,0808	0,1017	0,0962		0,0667	0,0429
	15,21			0,0060			0,0213		0,0082							0,0033
	15,20	0,0062		0,0240	0,0096	0,0294	0,0213	0,0101	0,0164	0,0143	0,0101	0,0339	0,0192		0,0333	0,2046
	15,19			0,0180		0,0147		0,0101	0,0246	0,0143	0,0404	0,0424	0,0192	0,0625	0,0111	0,0429
	15,18			0,0040	0,0192	0,0588	0,1064	0,0455	0,0082		0,0101	0,0085	0,0096		0,0111	0,0231
	15,17	0,0123	0,0090	0,0360	0,0962	0,0735	0,1915	0,0606	0,0328	0,0714	0,0202	0,0254	0,0385	0,0625	0,0556	0,0561
	15,16	0,0123		0,0460	0,0288	0,0441	0,0213	0,0152	0,0082	0,1143	0,0303	0,0424	0,0577	0,3125	0,1000	0,0462
	15,15	0,0309		0,0240	0,0096	0,0441	0,0213	0,0202	0,0164	0,0143		0,0085	0,0096		0,0056	0,0099
	14,22			0,0040												
	14,21											0,0085			0,0056	
	14,20			0,0040	0,0096			0,0101		0,0143		0,0085	0,0096		0,0222	0,0033
	14,19	0,0062		0,0100	0,0096			0,0101	0,0164				0,0096		0,0222	0,0033
	14,18			0,0080	0,0385			0,0152	0,0328	0,0143			0,0096		0,0111	
	14,17			0,0120	0,0769		0,0426	0,0101	0,0246	0,0143	0,0101	0,0254				
	14,16	0,0123		0,0180	0,0096	0,0294		0,0051		0,0286	0,0303		0,0096		0,0056	0,0066
	14,15	0,0802	0,0721	0,0300	0,0192	0,0294	0,0213	0,0253	0,0410			0,0085		0,0625		0,0033

	14,14	0,0185	0,0360	0,0460	0,0192	0,0147	0,0213	0,0051		0,0143		0,0085	0,0096			0,0066
	13,20			0,0040	0,0288											
	13,19			0,0060	0,0385											0,0033
	13,18	0,0062		0,0060	0,0481	0,0294	0,0213		0,0082							0,0033
	13,17			0,0160	0,0962	0,0147				0,0143						
	13,16	0,0185	0,0090	0,0200	0,0288	0,0147			0,0164				0,0096			0,0033
	13,15	0,0309	0,0180	0,0200	0,0192	0,0147					0,0101				0,0056	
	13,14	0,0617	0,0541	0,0260	0,0096		0,0426	0,0051				0,0085				0,0033
	13,13	0,0185	0,0090	0,0060	0,0096											
	13,2,16							0,0051								
	12,22			0,0020												
	12,20			0,0020												
	12,19								0,0082							
	12,18			0,0040												
	12,17			0,0040	0,0096											
	12,16	0,0123	0,0180	0,0020		0,0147							0,0096			
	12,15	0,0062	0,0090	0,0280		0,0294										
	12,14	0,0123	0,0270	0,0120	0,0096	0,0147					0,0101					0,0066
	12,13			0,0080		0,0147										0,0033
	12,12	0,0062		0,0100				0,0051		0,0143			0,0096			0,0033
	11,18							0,0051								
	11,17	0,0062														
	11,16	0,0062	0,0090	0,0060												
	11,15	0,0926	0,1081	0,0280	0,0481	0,0294		0,0051					0,0096			
	11,14	0,3765	0,3874	0,1760	0,1346	0,1765	0,0638				0,0101				0,0056	0,0099
	11,13	0,0185	0,0631	0,0240	0,0192	0,0147		0,0101				0,0085	0,0096		0,0056	0,0033
	11,12	0,0062	0,0270	0,0040		0,0294		0,0202	0,0082	0,0143		0,0085	0,0192		0,0056	
	11,11	0,0123	0,0270	0,0240		0,0294		0,1061	0,1230	0,1714	0,2323	0,1017	0,1250		0,0778	0,0561
	10,11,13							0,0051								
	10,17			0,0020						0,0143						
	10,15			0,0020							0,0101					
	10,14	0,0185	0,0180	0,0100	0,0096		0,0213									
	10,13		0,0090													0,0033
	10,12			0,0020												
	10,11	0,0062		0,0040								0,0085				

5.6 References

Ballantyne, K. N., Keerl, V., Wollstein, A., Choi, Y., Zuniga, S. B., Ralf, A., 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), 208–218. https://doi.org/10.1016/j.fsigen.2011.04.017

Ballantyne, K. N., Ralf, A., Aboukhalid, R., Achakzai, N. M., Anjos, M. J., Ayub, Q., Kayser, M. (2014). Toward Male Individualization with Rapidly Mutating Y-Chromosomal Short Tandem Repeats. *Human Mutation*, *35*(8), 1021–1032. https://doi.org/10.1002/humu.22599

Bandelt, H. J., Forster, P., & Röhl, A. (1999). Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution*, *16*(1), 37–48. Retrieved from http://dx.doi.org/10.1093/oxfordjournals.molbev.a026036

Burrows 2018. A Comparative ancestry analysis of Y-chromosome DNA haplogoups using high resolution melting. Master thesis 2018, University of the Western Cape.

Butler, J. M. (2011). *Advanced Topics in Forensic DNA Typing: Methodology*. Elsevier Science. Retrieved from https://books.google.co.za/books?id=9MJpGiDwUbAC

Cloete, K., Ehrenreich, L., D'Amato, M. E., Leat, N., Davison, S., & Benjeddou, M. (2010). Analysis of seventeen Y-chromosome STR loci in the Cape Muslim population of South Africa. *Legal Medicine*, *12*(1), 42–45. https://doi.org/10.1016/j.legalmed.2009.10.001

D'Amato, M. E., Bajic, V. B., & Davison, S. (2011). Design and validation of a highly discriminatory 10-locus Y-chromosome STR multiplex system. *Forensic Science International: Genetics*, *5*(2), 122–125. https://doi.org/10.1016/j.fsigen.2010.08.015

D'Amato, M. E., Benjeddou, M., & Davison, S. (2009). Evaluation of 21 Y-STRs for population and forensic studies. *Forensic Science International: Genetics Supplement Series*, *2*(1), 446–447. https://doi.org/10.1016/j.fsigss.2009.08.091

D'Amato, M. E., Ehrenreich, L., Cloete, K., Benjeddou, M., & Davison, S. (2010). Characterization of the highly discriminatory loci DYS449, DYS481, DYS518, DYS612, DYS626, DYS644 and DYS710. *Forensic Science International: Genetics*, *4*(2), 104–110. https://doi.org/10.1016/j.fsigen.2009.06.011

de Filippo, C., Bostoen, K., Stoneking, M., & Pakendorf, B. (2012). Bringing together linguistic and genetic evidence to test the Bantu expansion. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1741), 3256 LP-3263. Retrieved from http://rspb.royalsocietypublishing.org/content/279/1741/3256

de Wit, E., Delport, W., Rugamika, C. E., Meintjes, A., Möller, M., van Helden, P. D., Hoal, E. G. (2010). Genome-wide analysis of the structure of the South African Coloured Population in the Western Cape. *Human Genetics*, *128*(2), 145–153. https://doi.org/10.1007/s00439-010-0836-1

Decker, A. E., Kline, M. C., Vallone, P. M., & Butler, J. M. (2007). The impact of additional Y-STR loci on resolving common haplotypes and closely related individuals. *Forensic Science International: Genetics*, *1*(2), 215–217. https://doi.org/https://doi.org/10.1016/j.fsigen.2007.01.012

Excoffier, L., Laval, G., & Schneider, S. (2007). Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online*, *1*, 47–50. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/19325852

Fearon, J. D. (2003). Ethnic and Cultural Diversity by Country*. *Journal of Economic Growth*, *8*(2), 195–222. https://doi.org/10.1023/A:1024419522867

Froneman, S., & Kapp, P. A. (2017). An exploration of the knowledge, attitudes and beliefs of Xhosa men concerning traditional circumcision. *African Journal of Primary Health Care & Family Medicine* , *9*, 1–8. Retrieved from http://www.scielo.org.za/scielo.php?script=sci_arttext&pid=S2071-29362017000100056&nrm=iso

IBM Corp. Released 2016. IBM SPSS Statistics for Windows, Version 24.0. Armonk, NY: IBM Corp.

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

Kayser, M., Brauer, S., Weiss, G., Schiefenhövel, W., Underhill, P., Shen, P., Stoneking, M. (2003). Reduced Y-chromosome, but not mitochondrial DNA, diversity in human populations from West New Guinea. *American Journal of Human Genetics*, *72*(2), 281–302. https://doi.org/10.1086/346065

Kayser, M., Kittler, R., Erler, A., Hedman, M., Lee, A. C., Mohyuddin, A., Tyler-Smith, C. (2004). A comprehensive survey of human Y-chromosomal microsatellites. *American Journal of Human Genetics*, *74*(6), 1183–1197. https://doi.org/10.1086/421531

Kirsty Hynes 2015. A dual analysis of the South African Griqua population using ancestry informative mitochondrial DNA and discriminatory short tandem repeats on the Y chromosome. Masters dissertation- Kirsty Hynes University of the Western Cape 2015.https://etd.uwc.ac.za/handle/11394/5056

Leat, N., Benjeddou, M., & Davison, S. (2004). Nine-locus Y-chromosome STR profiling of Caucasian and Xhosa populations from Cape Town, South Africa. *Forensic Science International*, *144*(1), 73–75. https://doi.org/10.1016/j.forsciint.2004.02.022

Lloyd, T. O. (1984). *The British Empire*. Oxford University Press. Retrieved from https://books.google.co.za/books?id=BRM0AAAAIAAJ

Maclean, J. (1866). A Compendium of Kafir Laws and Customs: Including Genealogical Tables of Kafir Chiefs and Various Tribal Census Returns. *S. Solomon & Company, printers*. Retrieved from https://books.google.co.za/books?id=JwQyAQAAMAAJ

Makki-Rmida, F., Kammoun, A., Mahfoudh, N., Ayadi, A., Gibriel, A. A., Mallek, B., Masmoudi, S. (2015). Genetic diversity and haplotype structure of 21 Y-STRs, including nine noncore loci, in South Tunisian Population: Forensic relevance. *ELECTROPHORESIS*, *36*(23), 2908–2913. https://doi.org/10.1002/elps.201500204

Marks, S. J., Montinaro, F., Levy, H., Brisighelli, F., Ferri, G., Bertoncini, S., Capelli, C. (2015). Static and Moving Frontiers: The Genetic Landscape of Southern African Bantu-Speaking Populations. Molecular Biology and Evolution, 32(1), 29–43. Retrieved from http://dx.doi.org/10.1093/molbev/msu263

Mesthrie, R. (2002). *Language in South Africa*. Cambridge University Press. Retrieved from https://books.google.co.za/books?id=Y25iAAAAMAAJ

Morris, A. G. (1997). The Griqua and the Khoikhoi: biology, ethnicity and the construction of identity. *Kronos*, (24), 106–118. Retrieved from http://www.jstor.org/stable/41056392

Nei, M., & Tajima, F. (1981). DNA polymorphism detectable by restriction endonucleases. *Genetics*, *97*(1), 145–163. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/6266912

Nuñez, C., Baeta, M., Fernández, M., Zarrabeitia, M., Martinez-Jarreta, B., & De Pancorbo, M. M. (2015). Highly discriminatory capacity of the PowerPlex® Y23 System for the study of isolated populations. *Forensic Science International: Genetics*, *17*, 104–107. https://doi.org/10.1016/j.fsigen.2015.04.005

Nurse, D., & Philippson, G. (2006). *The Bantu Languages*. Taylor & Francis. Retrieved from https://books.google.co.za/books?id=M8cHBAAAQBAJ

Nurse, G. T., & Jenkins, T. (1975). The Griqua of Campbell, Cape Province, South Africa. *American Journal of Physical Anthropology*, *43*(1), 71–78. https://doi.org/10.1002/ajpa.1330430111

Oota, H., Settheetham-Ishida, W., Tiwawech, D., Ishida, T., & Stoneking, M. (2001). Human mtDNA and Y-chromosome variation is correlated with matrilocal versus patrilocal residence. *Nature Genetics*, *29*, 20. Retrieved from https://doi.org/10.1038/ng711

Peires, J. B. (1989). The Dead Will Arise: Nongqawuse and the Great Xhosa Cattle-killing Movement of 1856-7. *Ravan Press*. Retrieved from https://books.google.co.za/books?id=Rcqy3c0go7QC

Pickrell, J. K., Patterson, N., Loh, P.-R., Lipson, M., Berger, B., Stoneking, M., Reich, D. (2014). Ancient west Eurasian ancestry in southern and eastern Africa. *Proceedings of the National Academy of Sciences*, *111*(7), 2632–2637. https://doi.org/10.1073/pnas.1313787111

Purps, J., Siegert, S., Willuweit, S., Nagy, M., Alves, C., Salazar, R., Roewer, L. (2014). A global analysis of Y-chromosomal haplotype diversity for 23 STR loci. *Forensic Science International: Genetics*, *12*, 12–23. https://doi.org/10.1016/j.fsigen.2014.04.008

Roewer, L. (2009). Y chromosome STR typing in crime casework. *Forensic Science, Medicine, and Pathology*, *5*(2), 77–84. https://doi.org/10.1007/s12024-009-9089-5

Schlebusch 2010. Genetic variation in Khoisan-speaking populations from southern Africa. PhD University of the Witwatersrand

Statistics South Africa 2018. Mid - year population estimates, Statistical release P0302.http://www.statssa.gov.za/publications/P0302/P03022018.pdf

Zhang, G. Q., Yang, S. Y., Niu, L. L., & Guo, D. W. (2012). Structure and polymorphism of 16 novel Y-STRs in Chinese Han population. *Genetics and Molecular Research : GMR*, *11*(4), 4487–4500. https://doi.org/10.4238/2012.October.11.1

Chapter 6

6.0 Conclusive statement

To facilitating the criminal justice systems approach to sexual offenses, the UniQTyper™ Y-10 was developed as a forensic tool that may efficiently expedites current standard procedures. Current practices in South Africa rely solely on autosomal typing which is known to have limitation for processing highly admixed biological evidence. Evidently, with alarming incidence of assault and slow processivity due to large backlog, sexual offences have been historically met with poor conviction rates.

The UniQTyper™ Y-10 was developed and validated to deliver one of the fastest Y-chromosome profiling systems reported to date while maintaining a desirable level of sensitivity and specificity suitable for forensic application. The system is shown to tolerate respective levels of DNA profiling inhibitors and environmental insult often associated with compromised evidence. With a direct DNA profiling approach, reference samples could be perused without the need for DNA extraction.

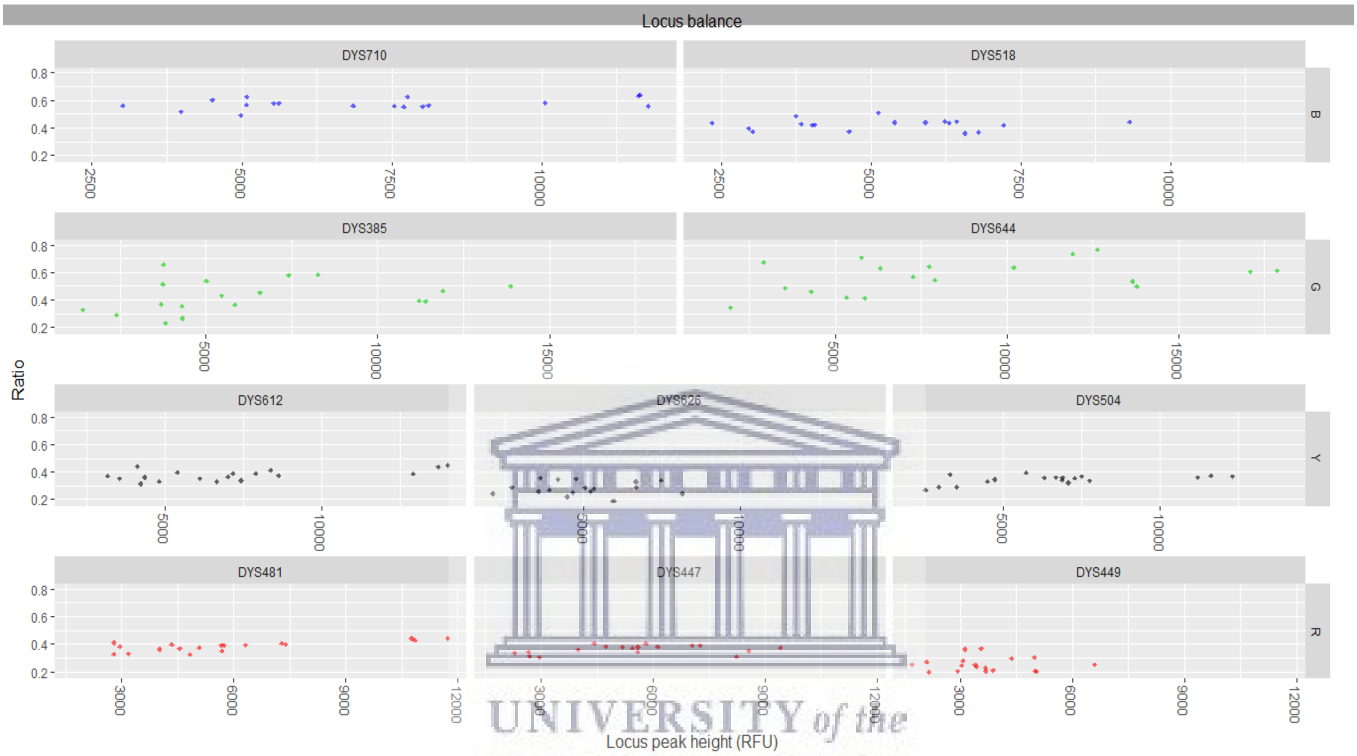
In this research output, the workflow for construction of a balanced allelic ladder has a major impact on the bulk production process and efficiency of upgrades to the observed allele range for the UniQTyper™ Y-10 kit.

The compilation of the largest Y-STR haplotype data and sequence data available for South Africa given with a highly discriminatory panel may provide for more comprehensive match probability estimates or sequence discrimination of shared alleles and haplotypes. The panel of markers also gave a strong signal that a possible bio-geographic ancestral approach may aid an investigative lead to a suspect.

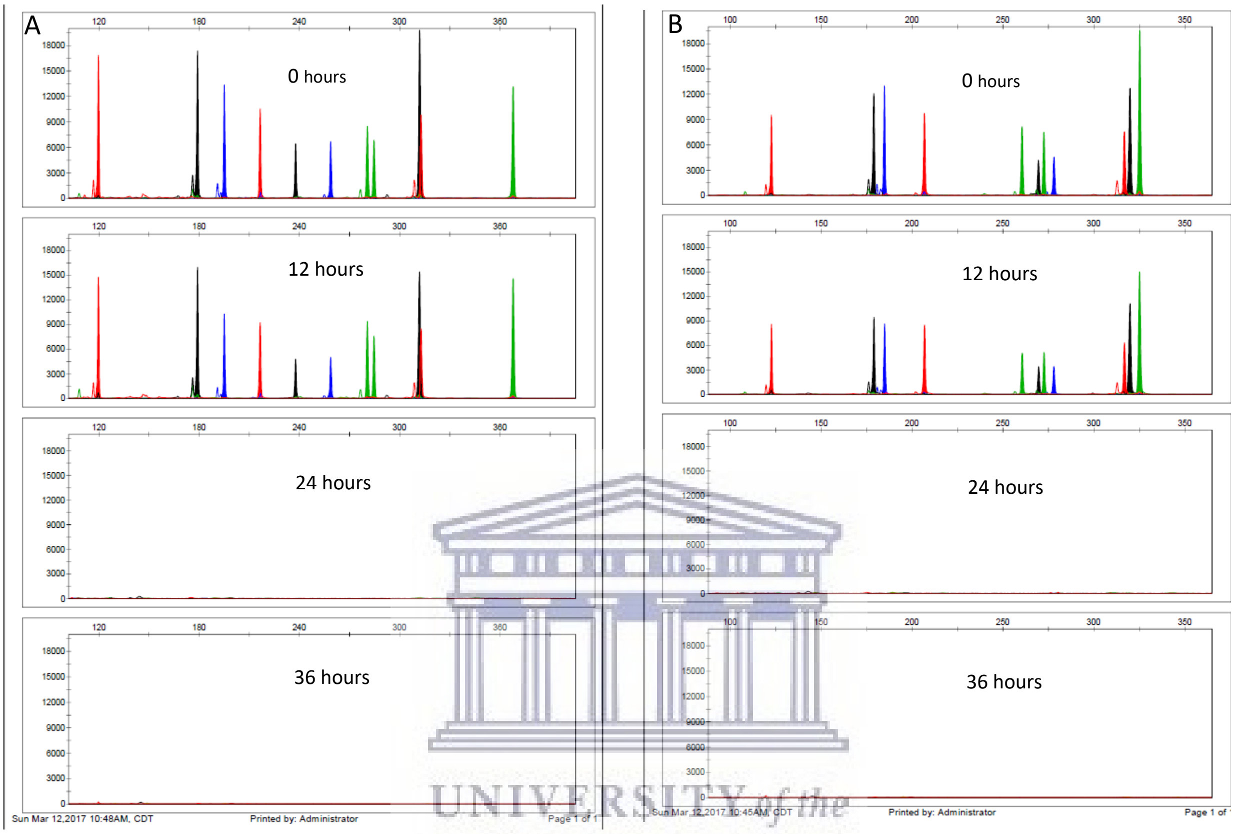
This research output which is at the forefront of forensic genetic technologies may significantly improve the power to discriminate between males in South African populations and potentially the conviction rate in sexual assault investigations.

7.0 Supporting Information

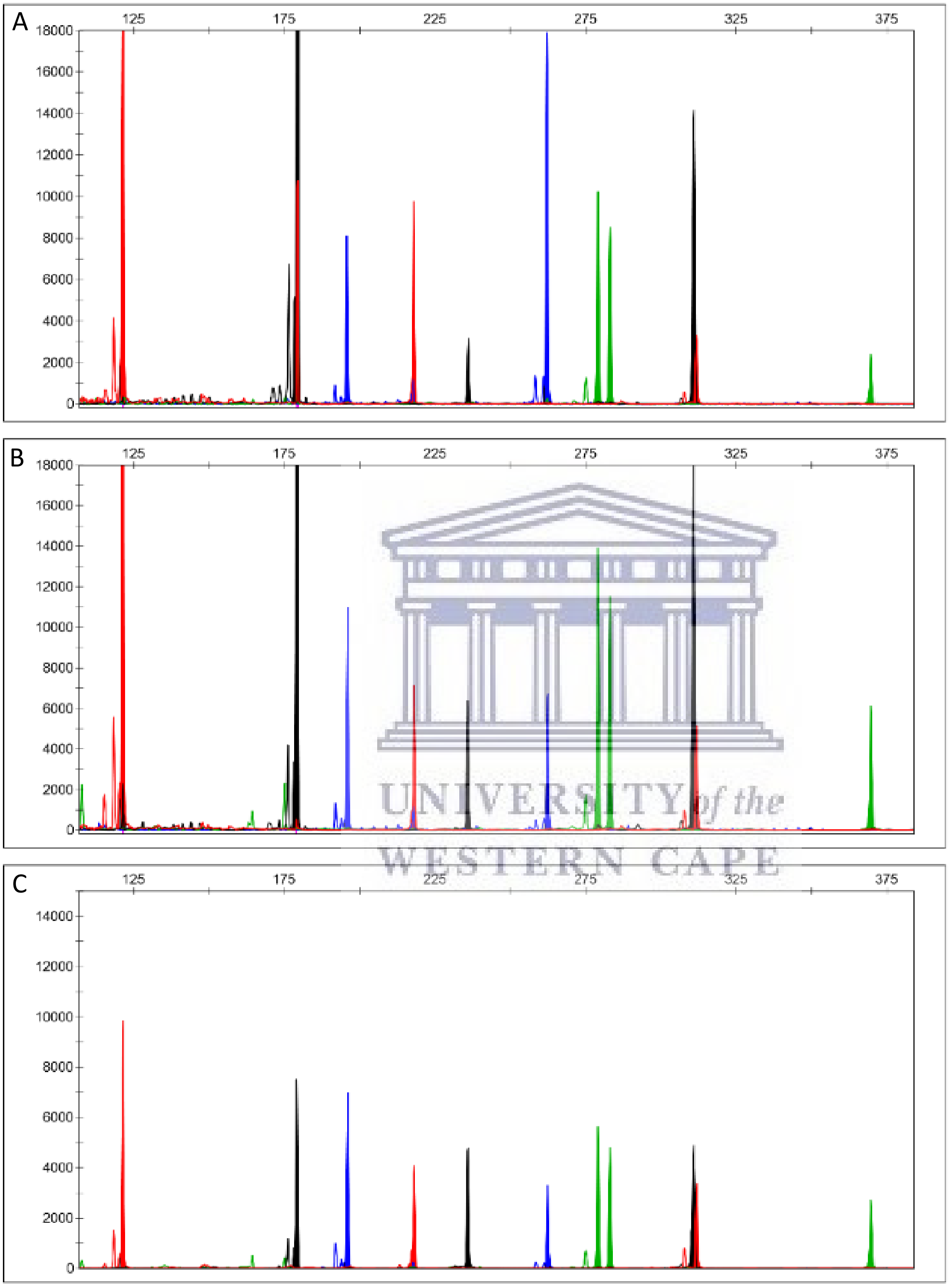
Supplementary 1 Figure 1.1: Inter-locus peak height balance for 2 step cycling across three thermal cyclers. The balance ratio was calculated in STR-validator using the total locus peak height as a proportion of the total profile peak height plotted by the mean peak height of the locus. (STR-validator plot).



Supplementary 1 Figure 1.2: A) Blood stained soil and B) denim jeans were exposed to outdoor conditions for 0, 12, 24 and 36 hours. The total direct sunlight exposure was 6 hours for every 12 hour outdoor exposure period. (i.e. samples exposed for 12 hours would have a total direct sunlight exposure of 6 hour, while samples exposed for 36 hours outdoors had a total of 18 hours of direct sunlight exposure. Y-axis scale 18 000 RFU.



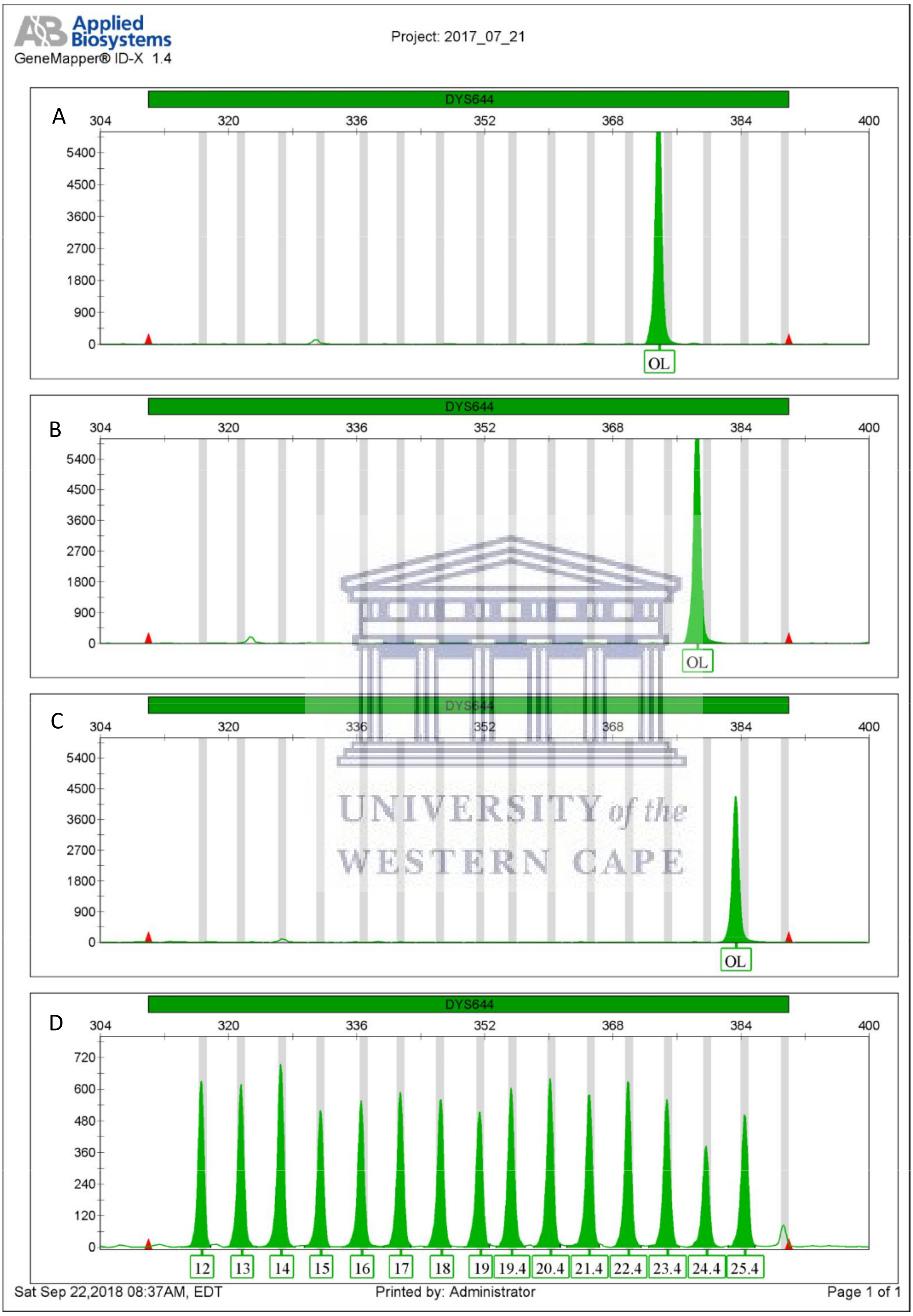
Supplementary 1 Figure 1.3: Direct amplification profiles for various substrates. A) Blood FTA, B) Buccal cells transfer to FTA and) crude saliva. Amplification was performed on 1 X 1.2 mm discs and 2µl crude saliva. Y-axis scaled 18 000 RFU (A-B) and 14000 RFU in (C).



Supplementary 2 Table 1: Alleles of the same size presenting 3-5 alternative repeat structure arrangements. ^{b}Population group: Eng (English); I (Asian Indian); C (Coloured); X (Xhosa); P (Pedi); TS (Tsumkwe San), AJ (Ashkanazi Jew), NA(information not available). For visual purpose the references to sequence comparison studies are as follows [6] Redd et al. 2002; [7] D’Amato et al. 2010; [9] Ruitberg et al. 2001; [15] Rodig et al. 2007.

Loci	Allele	Population group	Repeat structure	Sequence comparison
DYS449				
	28	I	[TTTC]$_{15}$N$_{50}$[TTTC]$_{13}$	this study
		I	[TTTC]$_{11}$N$_{50}$[TTTC]$_{17}$	this study
		TS	[TTTC]$_{12}$N$_{50}$[TTTC]$_{16}$	[6]
		AJ	[TTTC]$_{14}$N$_{50}$[TTTC]$_{14}$	[6]
	29	I	[TTTC]$_{12}$N$_{50}$[TTTC]$_{17}$	this study
		I	[TTTC]$_{16}$N$_{50}$[TTTC]$_{13}$	[7]
		TS	[TTTC]$_{15}$N$_{50}$[TTTC]$_{14}$	[6]
	35	I	[TTTC]$_{18}$N$_{50}$[TTTC]$_{17}$	this study
		Eng (SA)	[TTTC]$_{17}$N$_{50}$[TTTC]$_{18}$	this study
		I	[TTTC]$_{16}$N$_{50}$[TTTC]$_{19}$	[7]
DYS447				
	24	I	[TAATA]$_{7}$[TAAAA]$_{1}$[TAATA]$_{8}$[TAAAA]$_{1}$[TAATA]$_{7}$	this study
		Eng (UK)	[TAATA]$_{7}$[TAAAA]$_{1}$[TAATA]$_{16}$	[6]
		Eu	[TAATA]$_{7}$[TAAAA]$_{1}$[TAATA]$_{6}$[TAAAA]$_{1}$[TAATA]$_{9}$	[15]
		NA	[TAATA]$_{6}$[TAAAA]$_{1}$[TAATA]$_{12}$[TAAAA]$_{1}$[TAATA]$_{4}$	[9] SRM2395
	25	I	[TAATA]$_{7}$[TAAAA]$_{1}$[TAATA]$_{9}$[TAAAA]$_{1}$[TAATA]$_{7}$	this study
		P	[TAATA]$_{7}$[TAAAA]$_{1}$[TAATA]$_{8}$[TAAAA]$_{1}$[TAATA]$_{8}$	[6]
		NA	[TAATA]$_{9}$[TAAAA]$_{1}$[TAATA]$_{8}$[TAAAA]$_{1}$[TAATA]$_{6}$	[9] SRM 2395
	26	I	[TAATA]$_{9}$[TAAAA]$_{1}$[TAATA]$_{8}$[TAAAA]$_{1}$[TAATA]$_{7}$	this study
		C	[TAATA]$_{7}$[TAAAA]$_{1}$[TAATA]$_{9}$[TAAAA]$_{1}$[TAATA]$_{8}$	this study
		Eng(US)	[TAATA]$_{6}$[TAAAA]$_{1}$[TAATA]$_{12}$[TAAAA]$_{1}$[TAATA]$_{6}$	[6]
		NA	[TAATA]$_{7}$[TAAAA]$_{1}$[TAATA]$_{11}$[TAAAA]$_{1}$[TAATA]$_{6}$	[9] SRM 2395
DYS710				
	32	C	[AAAG]$_{13}$[AG]$_{12}$ [AAAG]$_{13}$	this study
		I	[AAAG]$_{15}$[AG]$_{12}$[AAAG]$_{11}$	[7]
		I	[AAAG]$_{14}$[AG]$_{12}$[AAAG]$_{12}$	[7]
	35	I	[AAAG]$_{17}$[AG]$_{12}$[AAAG]$_{11}$	this study
		I	[AAAG]$_{16}$[AG]$_{10}$[AAAG]$_{14}$	[7]
		X	[AAAG]$_{16}$[AG]$_{14}$[AAAG]$_{12}$	[7]

Supplementary 2 Figure 1: DYS644 variants containing 1bp deletion for sequenced alleles, A) 22.3; B) 23.3; C) 24.3 and D) Allelic ladder



Applied Biosystems
GeneMapper® ID-X 1.4

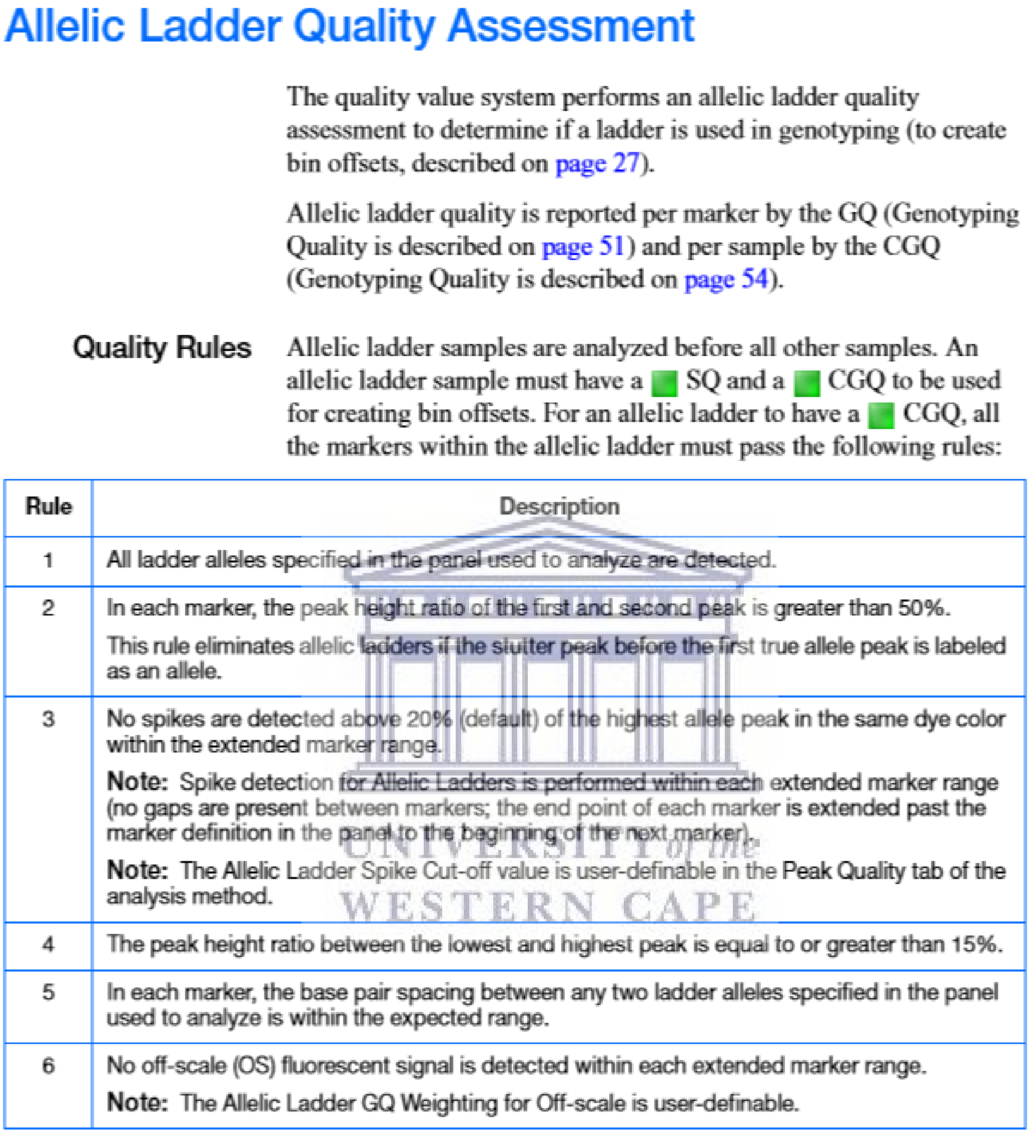
Project: 2017_07_21

Sat Sep 22,2018 08:37AM, EDT

Printed by: Administrator

Page 1 of 1

Supplementary 3 Figure 1 : GeneMapper® ID-X Quality assessment rules for passing an allelic ladder (GeneMapper® ID-X Software Version 1.5 Reference Guide (Pub. no. 100031707 Rev. A).



Allelic Ladder Quality Assessment

The quality value system performs an allelic ladder quality assessment to determine if a ladder is used in genotyping (to create bin offsets, described on page 27).

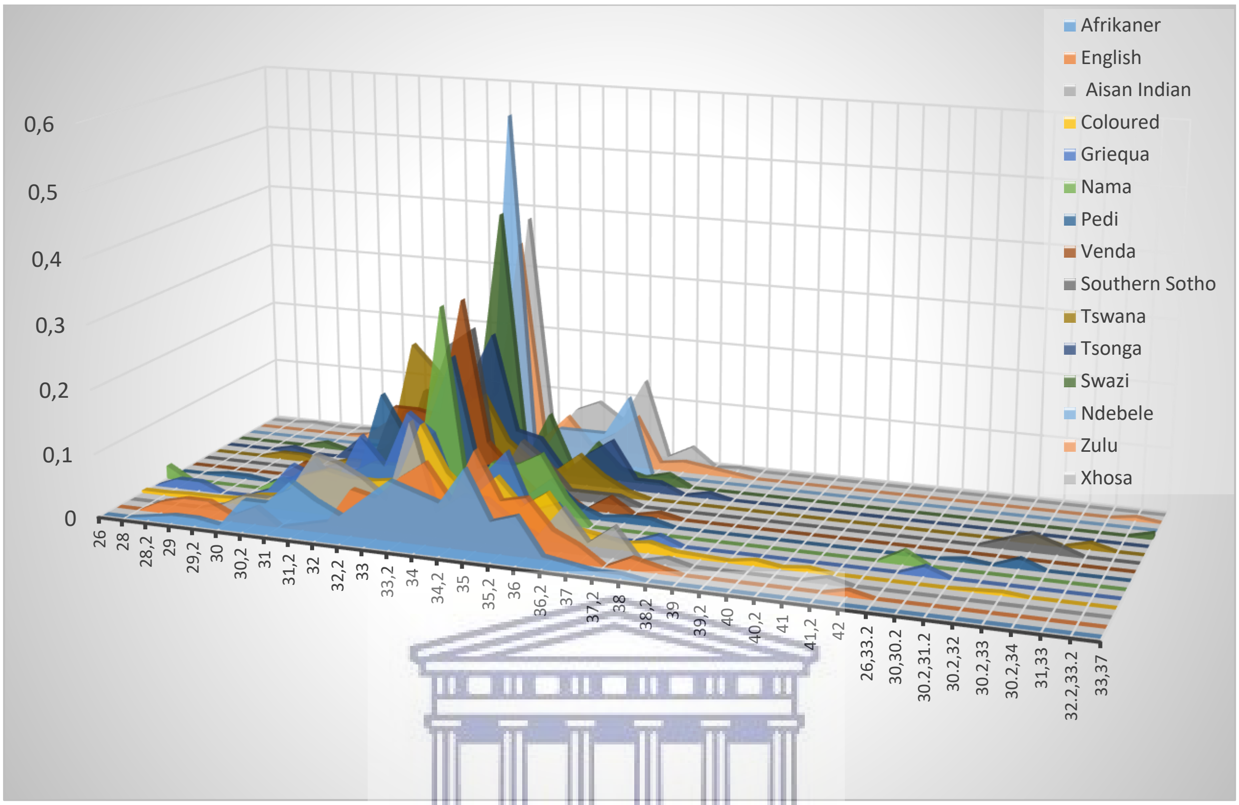
Allelic ladder quality is reported per marker by the GQ (Genotyping Quality is described on page 51) and per sample by the CGQ (Genotyping Quality is described on page 54).

Quality Rules

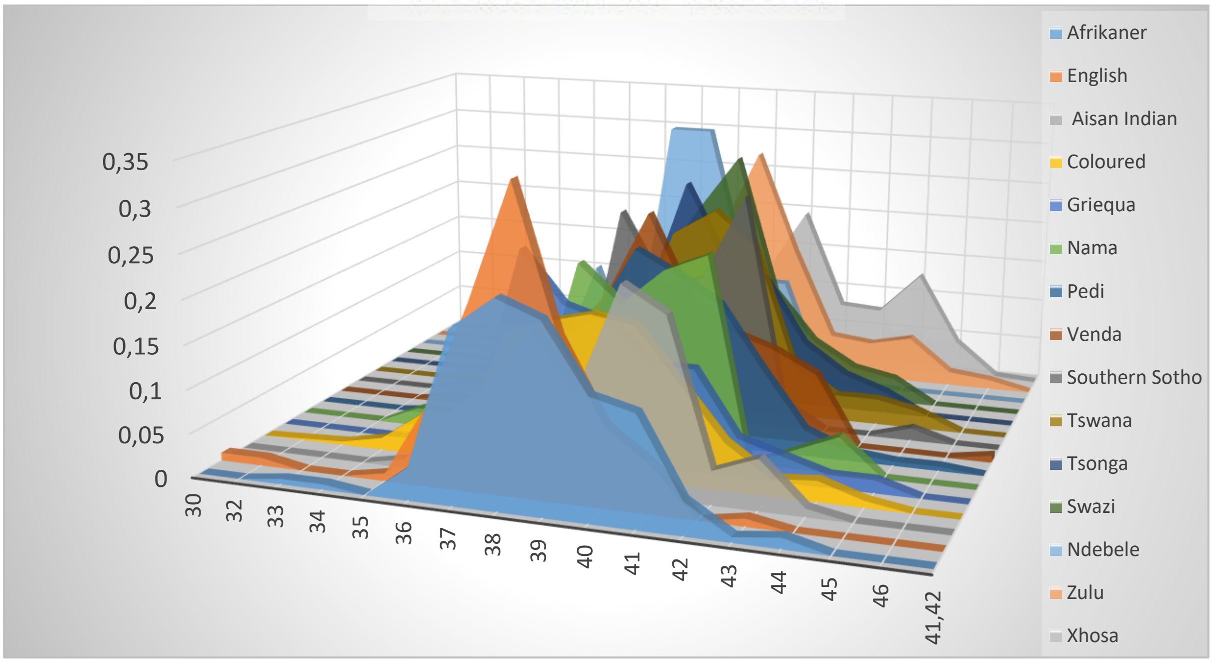
Allelic ladder samples are analyzed before all other samples. An allelic ladder sample must have a 🟩 SQ and a 🟩 CGQ to be used for creating bin offsets. For an allelic ladder to have a 🟩 CGQ, all the markers within the allelic ladder must pass the following rules:

Rule	Description
1	All ladder alleles specified in the panel used to analyze are detected.
2	In each marker, the peak height ratio of the first and second peak is greater than 50%. This rule eliminates allelic ladders if the stutter peak before the first true allele peak is labeled as an allele.
3	No spikes are detected above 20% (default) of the highest allele peak in the same dye color within the extended marker range. **Note:** Spike detection for Allelic Ladders is performed within each extended marker range (no gaps are present between markers; the end point of each marker is extended past the marker definition in the panel to the beginning of the next marker). **Note:** The Allelic Ladder Spike Cut-off value is user-definable in the Peak Quality tab of the analysis method.
4	The peak height ratio between the lowest and highest peak is equal to or greater than 15%.
5	In each marker, the base pair spacing between any two ladder alleles specified in the panel used to analyze is within the expected range.
6	No off-scale (OS) fluorescent signal is detected within each extended marker range. **Note:** The Allelic Ladder GQ Weighting for Off-scale is user-definable.

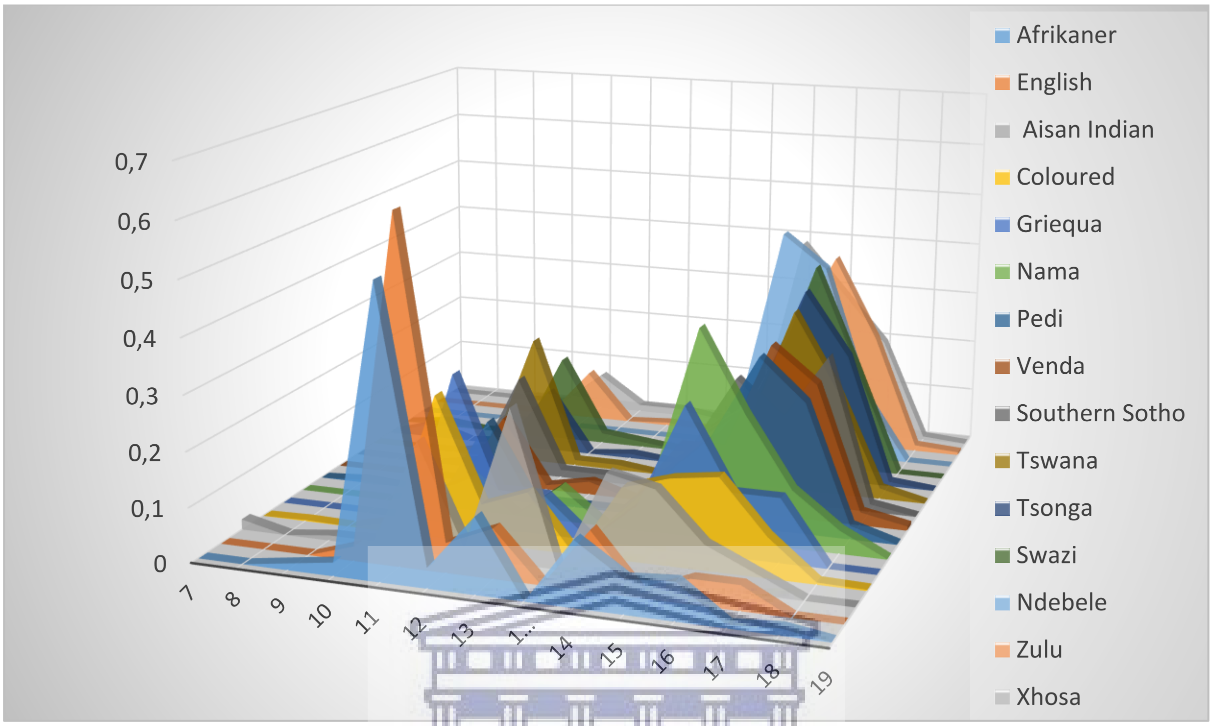
Supplementary 4 Figure 1: Allele frequencies for DYS710



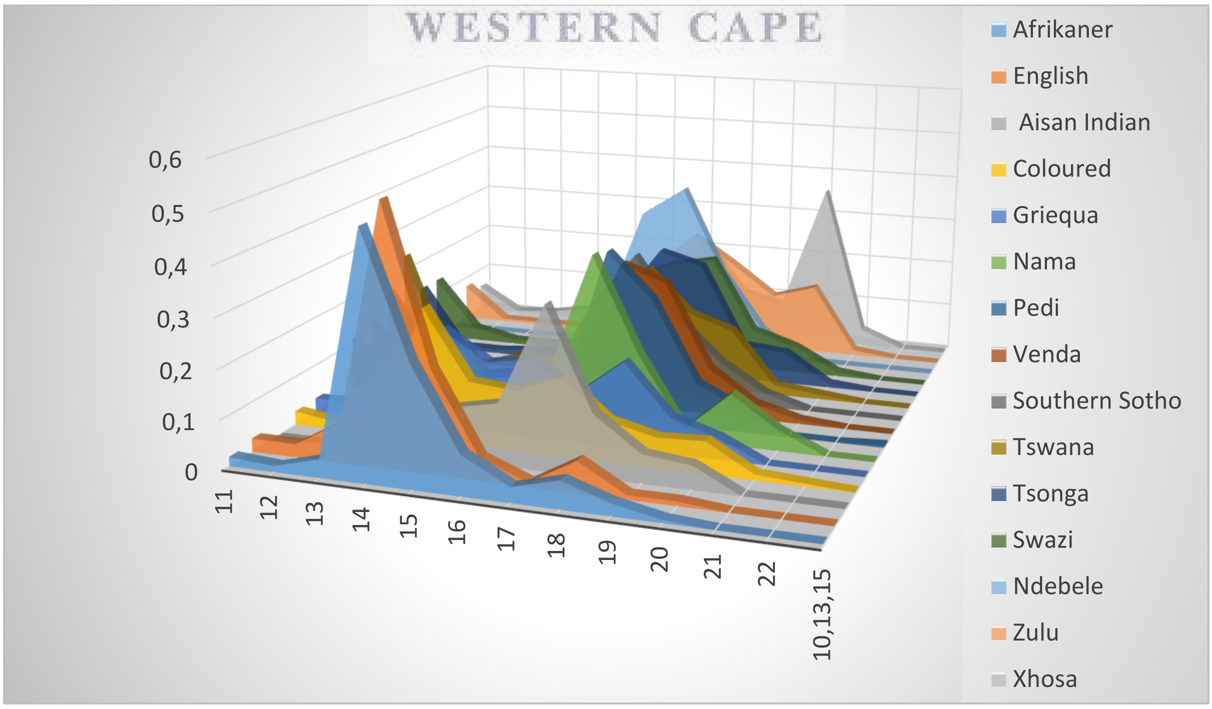
Supplementary 4 Figure 2: Allele frequencies for DYS518



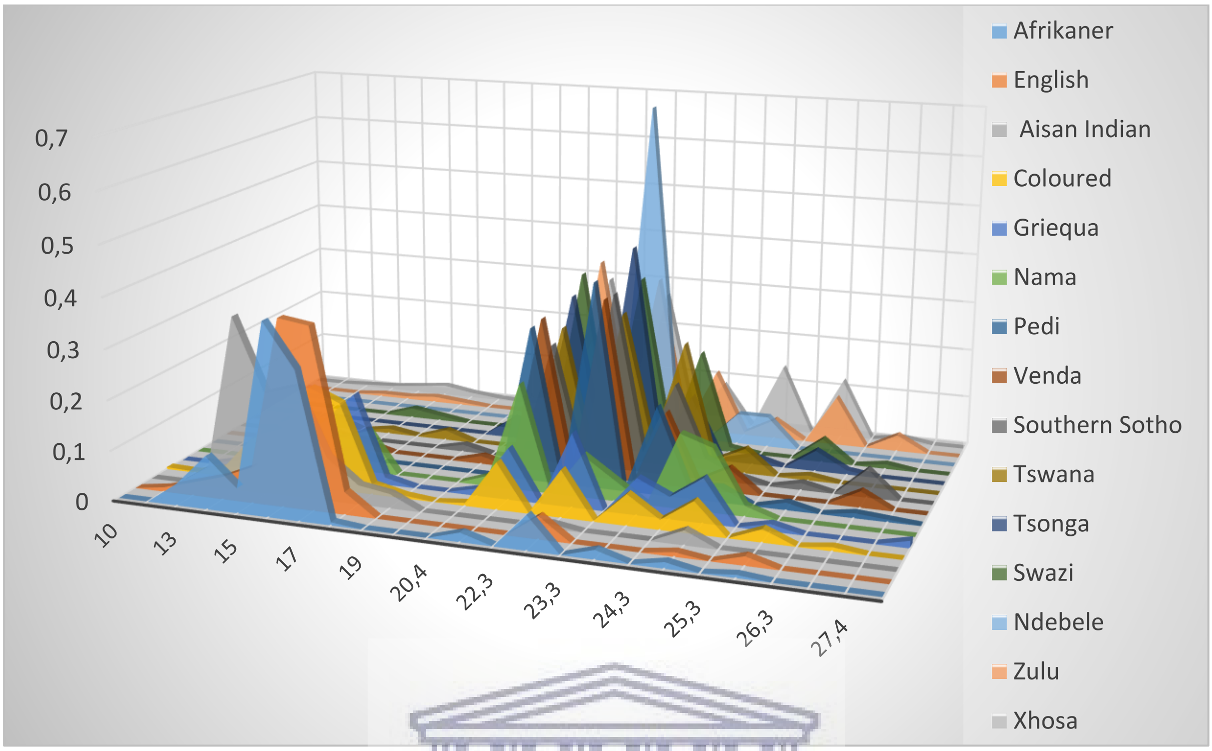
Supplementary 4 Figure 3: Allele frequencies for DYS385a



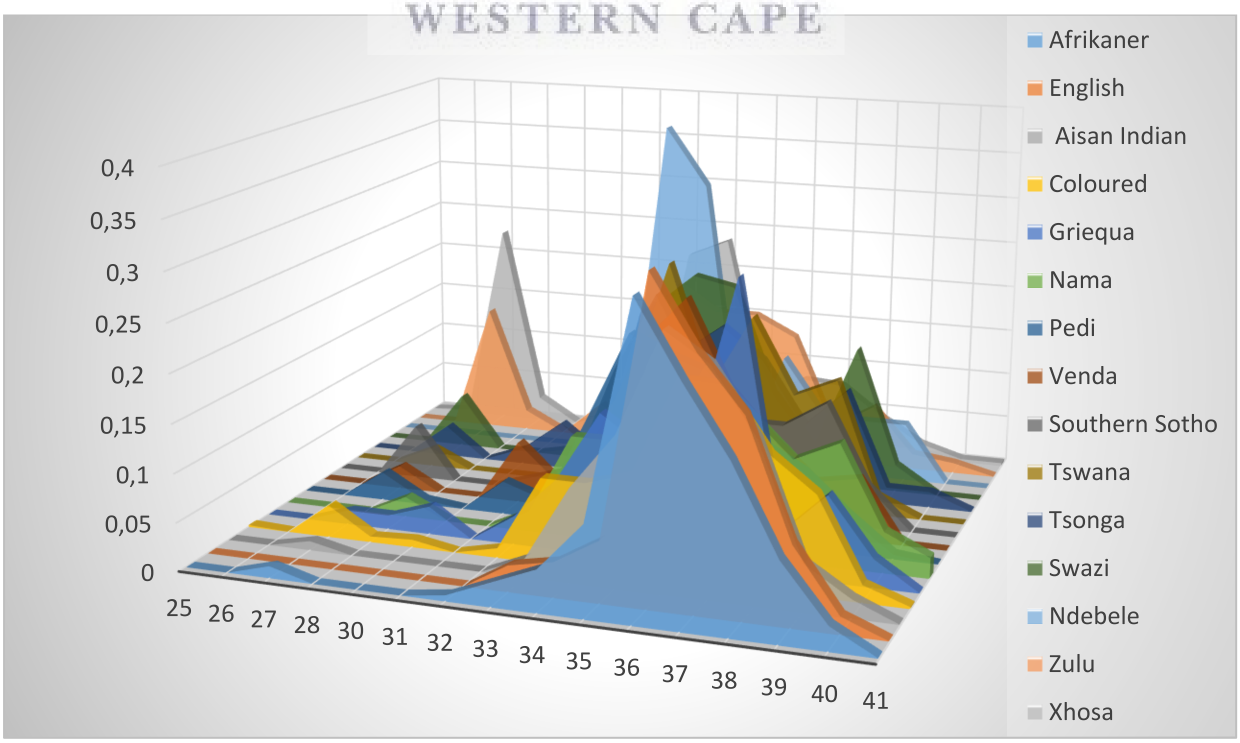
Supplementary 4 Figure 4: Allele frequencies for DYS385b



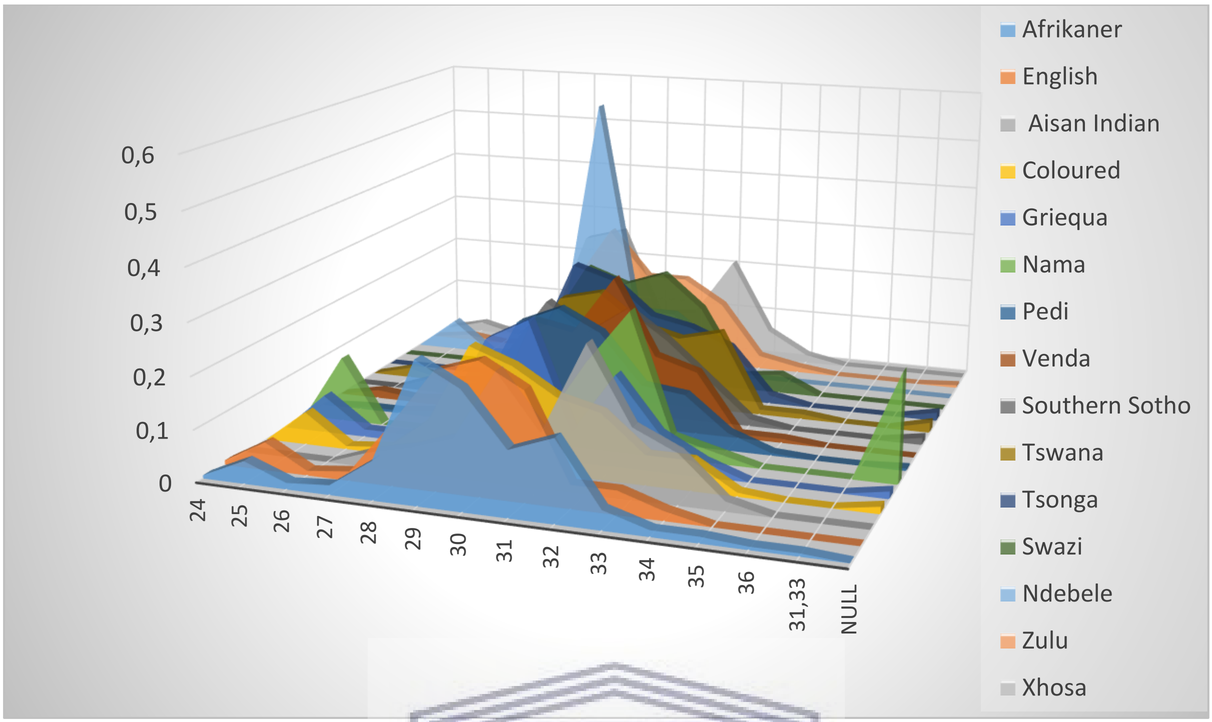
Supplementary 4 Figure 5: Allele frequencies for DYS644



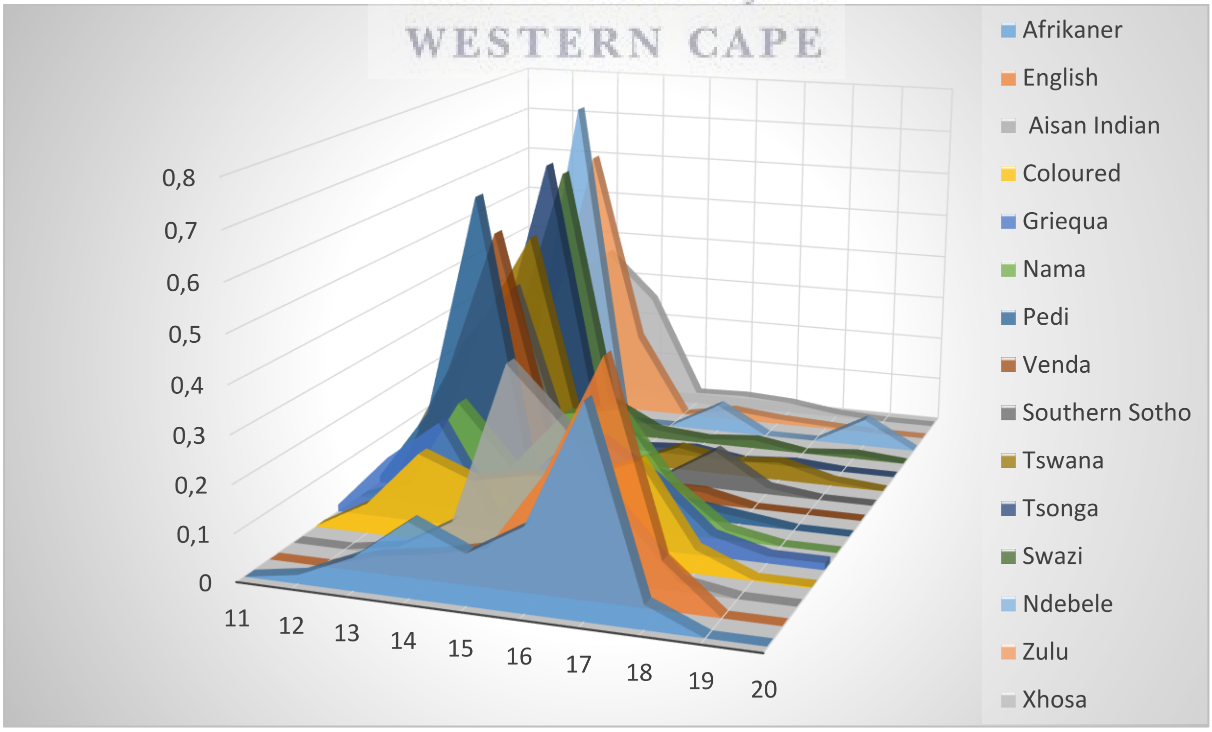
Supplementary 4 Figure 6: Allele frequencies for DYS612



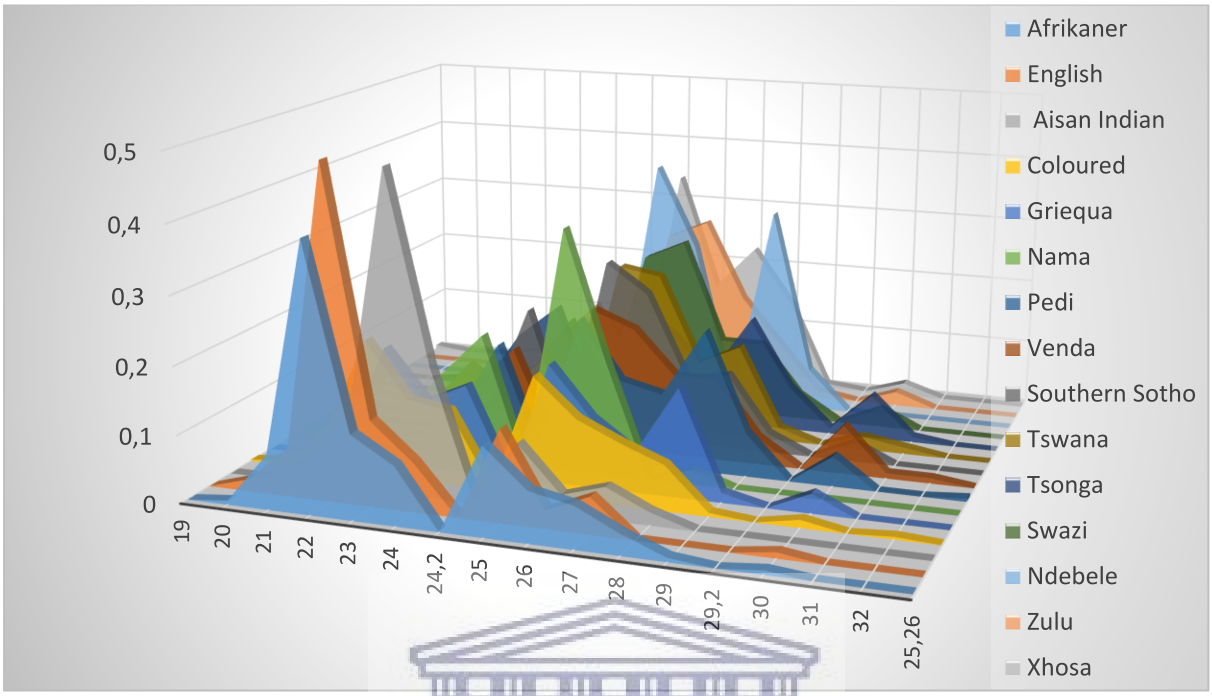
Supplementary 4 Figure 7: Allele frequencies for DYS626



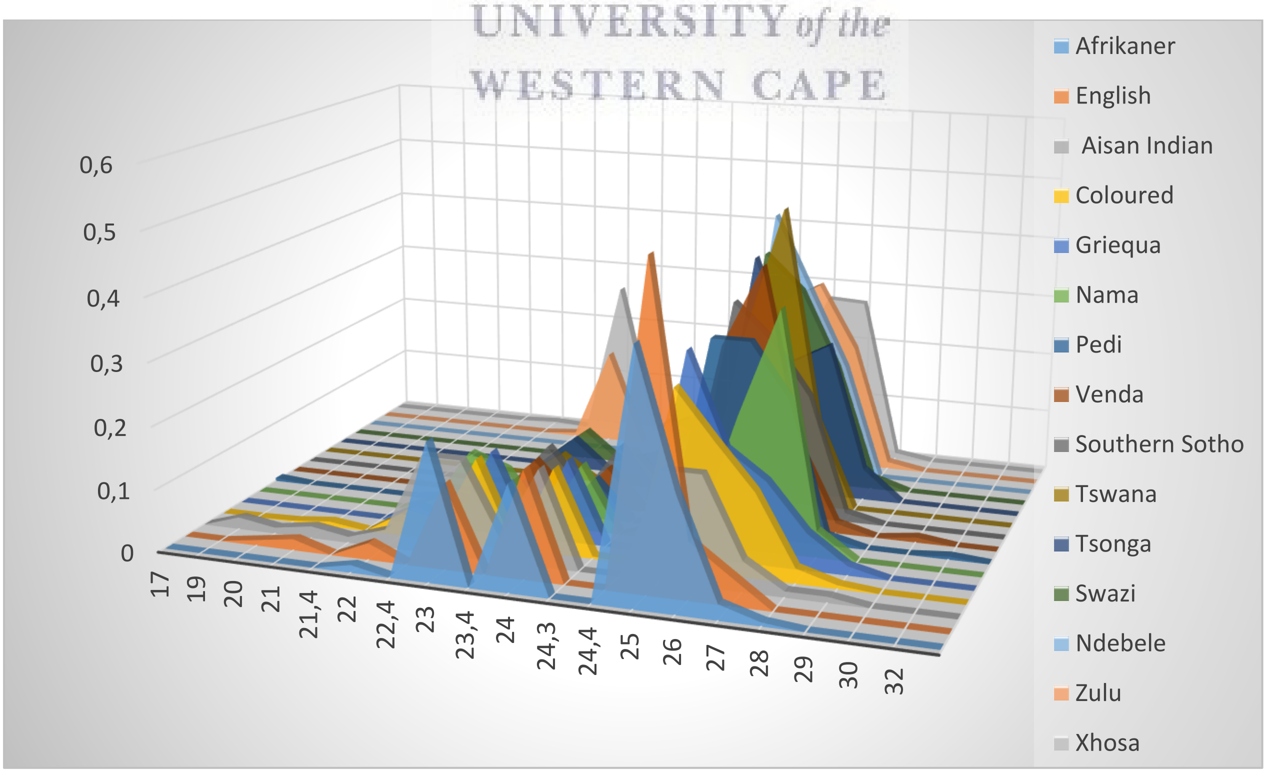
Supplementary 4 Figure 8: Allele frequencies for DYS504



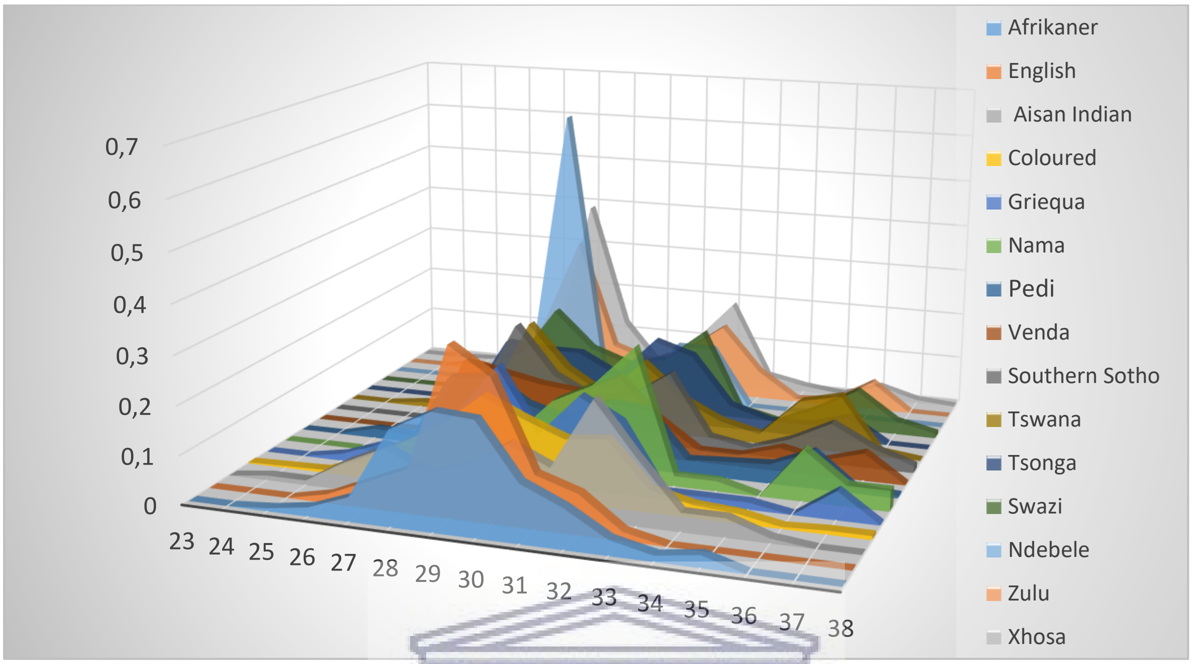
Supplementary 4 Figure 9: Allele frequencies for DYS481



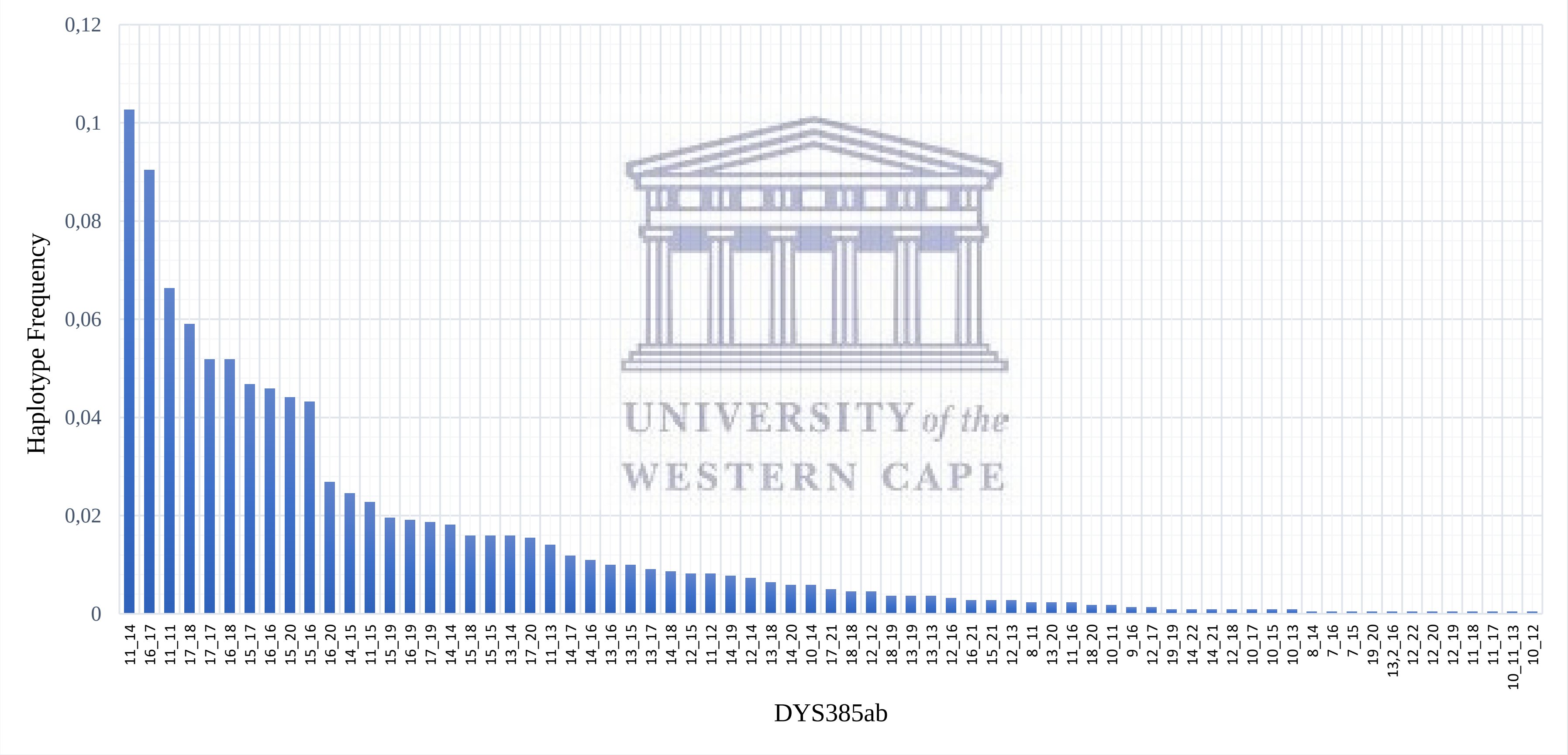
Supplementary 4 Figure 10: Allele frequencies for DYS447



Supplementary 4 Figure 11: Allele frequencies for DYS449



Supplementary 4 Figure 12 : Ranked haplotype frequency for DYS385ab across the 15 populations.



Supplementary 4 Table 1: Summary of n=2 and n=3 shared haplotype within groups and distributed between groups.

Haplotype	Afrikaner	English	Indian	Coloured	Griequa	Nama	Pedi	Venda	S. Sotho	Tswana	Tsonga	Swazi	Ndebele	Zulu	Xhosa
H16	3														
H17	1			2											
H18				2	1										
H19				3											
H20				1											2
H21				3											
H22						3									
H23							1				2				
H24							3								
H25							2					1			
H26							1	2							
H27							1			1		1			
H28								3							
H29								3							
H30									1			1			1
H31									1	1					1
H32										1				2	
H33											3				
H34											2	1			
H35												3			
H36													1	2	
H37														2	1
H38														1	2
H39															3
H40	1			1											
H41	2														
H42	2														
H43	2														
H44	1				1										
H45	2														
H46	2														
H47	1			1											
H48	2														
H49	1			1											
H50	2														
H51	1			1											
H52	1			1											
H53			2												
H54			2												
H55			1												1
H56			2												
H57			2												
H58			1	1											
H59				2											
H60				2											
H61				2											
H62				1											1
H63				2											
H64				2											
H65				2											
H66				2											
H67				2											
H68				1								1			
H69				2											
H70				2											

H71				2											
H72				2											
H73				2											
H74				2											
H75				1										1	
H76				1											1
H77				1										1	
H78				2											
H79				2											
H80				1	1										
H81				1	1										
H82				1	1										
H83				1		1									
H84				2											
H85				1						1					
H86				1		1									
H87				1			1								
H88					1									1	
H89					1	1									
H90					1										1
H91					1										1
H92						1									1
H93						2									
H94						2									
H95						2									
H96							1					1			
H97							2								
H98							1			1					
H99							2								
H100							2								
H101							1		1						
H102							1	1							
H103							1					1			
H104							2								
H105							1			1					
H106								2							
H107								2							
H108								1			1				
H109								1						1	
H110									1	1					
H111										2					
H112										2					
H113										2					
H114										1					1
H115											2				
H116											2				
H117											1	1			
H118											1				1
H119											1			1	
H120											2				
H121												2			
H122												1		1	
H123												1		1	
H124												2			
H125												1		1	
H126													1	1	
H127														1	1
H128														2	
H129														1	1
H130														2	
H131														2	
H132														2	
H133														1	1
H134															2

H135															2
H136															2
H137															2
H138															2
H139															2
H140															2
H141															2
H142															2
H143															2
H144															2
H145															2
H146															2
H147															2
H148															2
H149															2
H150															2
H151															2
H152															2