Evaluation of insertion-deletion polymorphisms with the kit Qiagen Investigator® DIPplex for forensic application in South Africa

Gwynneth Jacobs

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

Supervisor: Assoc. Prof. Maria Eugenia D'Amato

Co-Supervisor: Prof Sean Davison

Keywords

Forensics

Genotyping

Polymerase Chain Reaction (PCR)

PCR Multiplex Reaction

Insertion-deletion polymorphism

Indels

Allele frequencies

Population genetics

Mutation

Null Allele

Abstract

Evaluation of insertion-deletion polymorphisms with the kit Qiagen Investigator® DIPplex for forensic application in South Africa

G. Jacobs

MSc Thesis, Department of Biotechnology, University of the Western Cape

Insertion-deletion polymorphisms (indels) have been underutilized in forensic identification of individuals in comparison with single nucleotide polymorphisms (SNPs) and short tandem repeat (STRs) systems. The use of indels for the purpose of human identification is more advantageous than previously used methods as it combines desirable characteristics of both the SNPs and STRs i.e. low costs and simplistic typing methods as well as indels having small amplicons size, making them suitable for genotyping highly degraded DNA. Currently there is only one commercial kit available for the forensic community, the Investigator® DIPlex kit (Qiagen), which cover a total of 30 indel loci distributed over 19 autosomal chromosomes.

The objective of this study was to evaluate the Qiagen Investigator® DIPplex kit for forensic application in South Africa. The kit's performance was evaluated by comparing different extraction methods; sensitivity, robustness and reproducibility were evaluated and forensic parameters (match probability, power of discrimination, polymorphism information content, power of exclusion and typical paternity index) were estimated based on population data generated from five South African populations (Afrikaner, Mixed Ancestry, Indian-Asian, Xhosa and Zulu). Population comparisons were performed using F_{ST}-analysis, factorial component analysis as well as phylogenetic tree construction.

DNA was extracted from buccal swabs and whole blood collected from a total of 512 individuals from the five South African population groups and genotyped using the Qiagen Investigator® DIPplex kit. Sanger DNA sequencing and sequence alignments confirmed the presence of a null allele at locus HLD97

which was present in high frequency in the Xhosa and Zulu populations. This observation was made in 14 individuals belonging to the Xhosa and Zulu populations. Null allele frequencies in all five South African populations were also estimated. Null alleles were estimated for all loci using analytical methods i.e. Charkraborty null allele estimator, Brookfield null allele estimators 1 and 2 and ML-NullFreq software program.

The kit performed well in the laboratory, not requiring any additional reagents or instrumentation and successfully generating profiles with input DNA amounts as low as 0.2 ng/µL.

Although well suited for forensic application, the Qiagen Investigator® DIPplex kit showed some drawbacks with regards to application on South African populations. The presence of a null allele at the HLD97 locus as well as indication of population substructure affects allele frequency estimates for the South African populations. Correction for population substructure as present within the South African populations should be considered using F_{ST} analysis and it is recommended that the HLD97 locus should be excluded from any kinship analysis performed on South African populations.

Date 1/12/2014

Declaration

I declare that 'Evaluation of insertion-deletion polymorphisms with the kit Qiagen Investigator® DIPplex for forensic application in South Africa' is my own work that has not been submitted for any degree or examination in any other university and that all the sources I have used have been indicated and acknowledged by complete references.

X

_____

Gwynneth Jacobs

Mrs. 1 May 2015

Acknowledgements

First and foremost I would like to thank my supervisor Assistant Professor Maria Eugenia D'Amato for granting me the opportunity to complete this research project. Your patience, guidance and endless support are what I could always rely on throughout this journey.

I would like to show my appreciation to my family and friends who have always encouraged me throughout my studies and believed in my abilities to follow this through.

A special thank you goes to my husband, Stanley Jacobs, for all his patience and support in the final stages of this project.

I would also like to acknowledge the staff and management of the SAPS Forensic Science Laboratory's Biology Unit for supporting me in my part time studies.

I would like to thank the students and staff of the UWC Forensics DNA Laboratory for their help and support and for welcoming me into their fold. I need to acknowledge Professor Sean Davison for his support and encouragement and for always being willing to lend a helping hand.

I would also like to thank Qiagen for their assistance in making this project possible and for donating the Qiagen Investigator® DIPplex kits used in this research project.

List of Abbreviations

^{o}C	-	Degrees celcius
AIM	-	Ancestry informative marker
Avg	-	Average
bp	-	Base pair
BSA	-	Bovine Serum Albumin
CDP	-	Combined power of discrimination
CMP	-	Combined match probability
CODIS	-	Combined DNA index system
CPE	-	Combined power of exclusion
ddNTP	-	dideoxynucleoside triphosphate
DNA	-	Deoxyribonucleic acid
dNTP	-	Deoxyribonucleic triphosphate
DP	-	Power of discrimination
EDNAP	-	European DNA Profiling Group
ENFSI	-	European Network of Forensic Science Institutes
FBI	-	Federal Bureau of Investigation
FCA	-	Factorial correspondence analysis
FGA	-	Alpha fibrinogen
GC	-	Guanine and cytocine
He	-	Expected heterozygosity
HLA	-	Human leukocyte antigen
HLD	-	Human Locus DIP
Ho	-	Observed heterozyogisity
HWE	-	Hardy Weinberg Equilibrium
IDT	-	Integrated DNA Technologies
ISFG	-	International Society for Forensic Genetics
ISO	-	International Organisation for Standardization
M	-	Molar
mM	-	Millimolar
MP	-	Match probability
NDIS	-	National DNA Index System

NFDD	-	National Forensic DNA Database
ng	-	Nanogram
NJ	-	Neighbour joining
NDNAD	-	National DNA Database
nm	-	nanometer
NRY	-	Non-recombining Y-chromosome region
PCR	-	Polymerase chain reaction
PE	-	Power of exclusion
PIC	-	Polymorphism information content
P-Value	-	Probability value
RFLP	-	Restriction fragment length polymorphism
RMSE	-	Root Mean Squared Error
RM Y-STR	-	Rapidly mutating Y-hromosome short tandem repeat
SANAS	-	South African National Accreditation System
SAPS	-	South African Police Service
SNP	-	Single nucleotide polymorphism
STR	-	Short tandem repeat
SWGDAM	-	Scientific Working Group on DNA Analysis
Ta	-	Annealing temperature
THO	-	Tyrosine hydroylase
Tm	-	Melting temperature
TPI	-	Typical paternity index
TPOX	-	Thyroid peroxidase gene
U	-	Unit
uM	-	Micromolar
uL	-	Microliter
UV	-	Ultra violet
UWC	-	University of the Western Cape
VNTR	-	Variable number tandem repeats
vWA	-	Von Willebrand Factor
Y-STR	-	Y Chromosome Short Tandem Repeat

List of Figures

List of Tables

Chapter One

Literature review

General Introduction

Forensic genetics has evolved into a ground breaking field with developments and advances that has placed it on the forefront of innovative and cutting edge research. With an ever increasing dependency on forensics in crime solving, there is a need for faster sample processing and higher output of results that meet high quality standards. In its quest to fulfil these requirements, the field of forensics has made ground breaking progress and breakthroughs.

1.1. A brief history of forensic genetics

Figure 1.1 summarises some of the developments and advances in forensics. The first steps in human identification were taken with the application of human blood groups and protein electrophoresis as methods of distinguishing between individuals (Reviewed in Jobling and Gill, 2004). The ABO, MN and Rh systems were the most common systems used in the 1900s. All three systems were based on distinguishing between individuals on the bases of the different phenotypes that exists for the blood groups. The protein electrophoresis method was based on the existence of protein polymorphisms and isoenzymes for red blood cells and blood serum. These applications had their own shortcomings. Human blood groups and protein techniques could only be applied to these types of samples. Although these early methods hinted at the forensic possibilities, they were limited by rapid DNA degradation due to bacterial and environmental exposure and the inability to distinguish DNA mixtures.

The first DNA typing method was described by Dr Alec Jeffreys in 1984 (Jeffreys *et al*, 1985; Gill *et al*, 1985). DNA Fingerprinting was the first genetic method applied in crime solving.

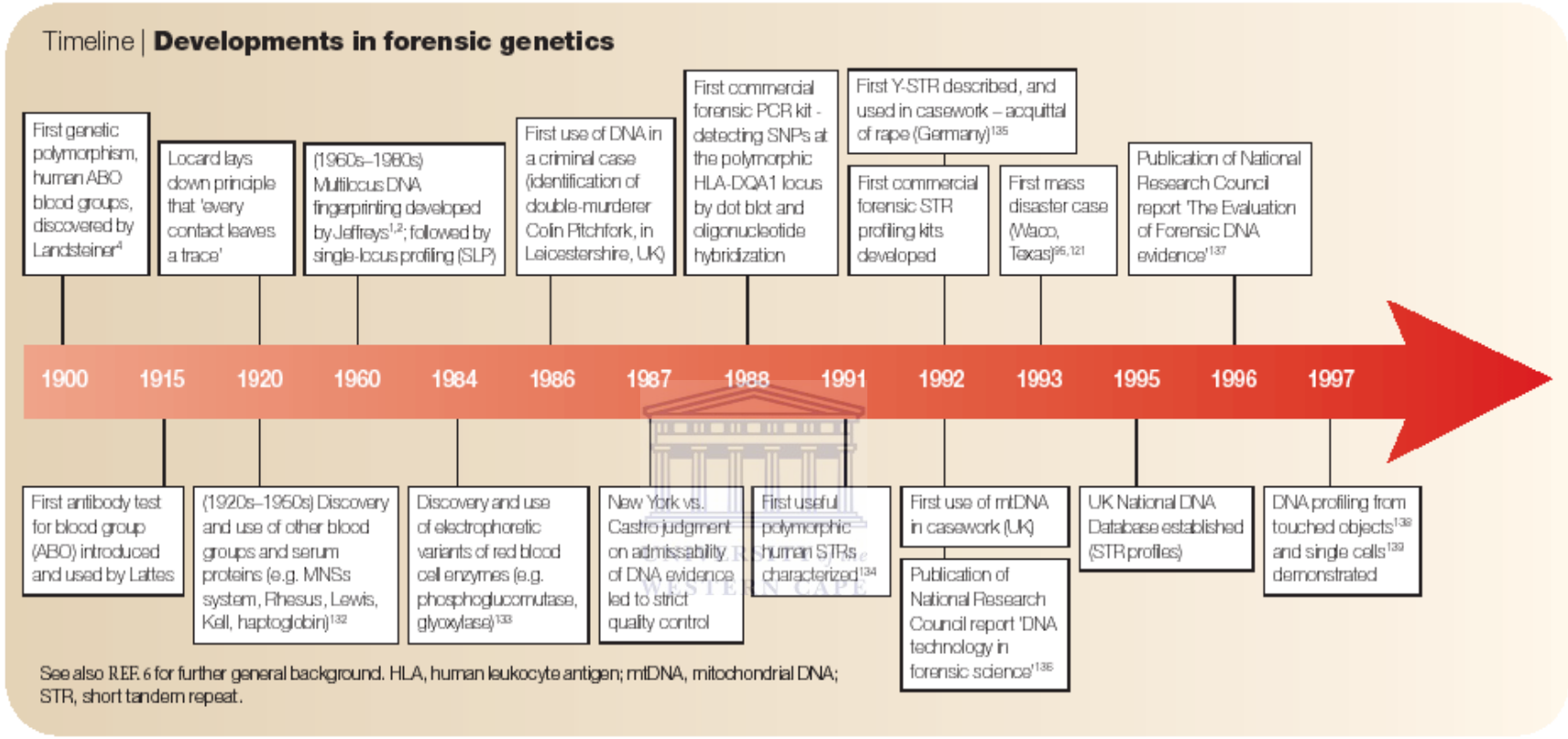


Figure 1.1 A timeline for the developments that have shaped forensics (Jobling and Gill, 2004)

The technique is based on the differences in length of repeat DNA sequences known as variable number tandem repeats (VNTRs). The VNTRs allow for human identification as the number of repeat sequences varies between individuals, variation that is due to genetic inheritance and recombination which results in a unique DNA fingerprint profile. Figure 1.2 illustrates the steps used to create a DNA fingerprint profile method known as restriction fragment length polymorphism (RFLP). This technique allowed for comparison between crime exhibits and reference DNA samples. Although the technique had its drawbacks like being labour intensive, it laid the foundation for forensic science.

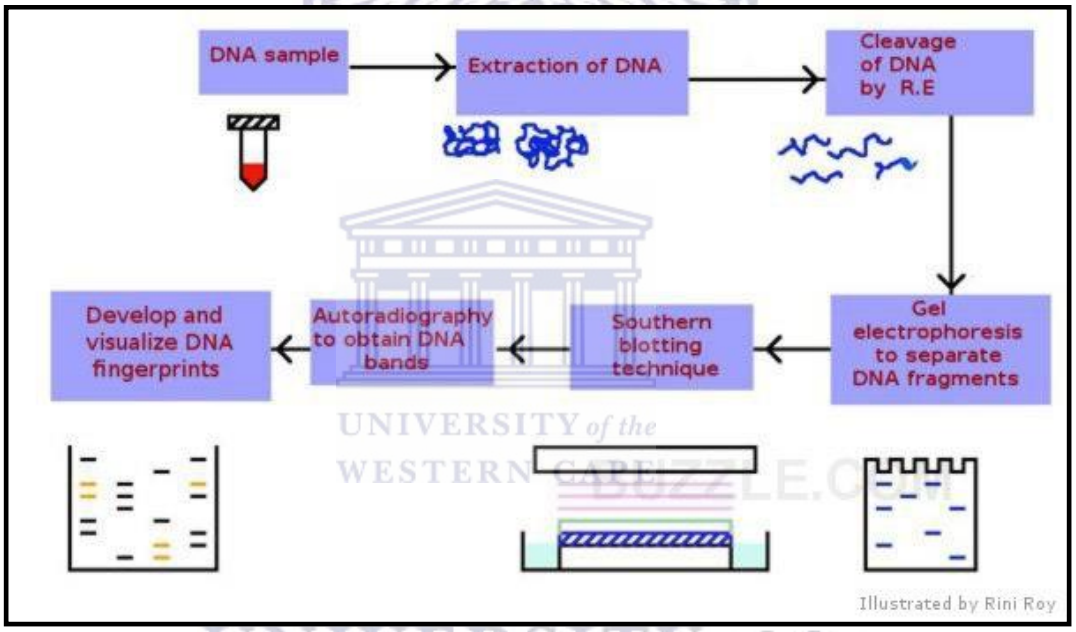


Figure 1.2 The DNA fingerprinting process (Image from www.buzzle.com).

Application of the revolutionary technique that is today used on all biological platforms, the polymerase chain reaction (PCR) has allowed a way to overcome the requirement of high DNA concentration by earlier methods (Saiki *et al*, 1985; Mullis and Faloona, 1987). Simply put, PCR allows amplification of DNA segments by denaturing a DNA strand and using the denatured single DNA strand as a template for making more DNA copies.

The idea of PCR developed with individual scientific discoveries that eventually lead to the PCR process that is known today. First and foremost the idea of PCR started with the discovery of the DNA double helix structure by James D. Watson

and Francis Crick and their postulation that DNA can be replicated (Watson and Crick, 1953). This was followed by the discovery of, among others, the first DNA polymerase by Arthur Kornberg by 1957 (Lehman *et al*, 1958) and the first taq enzyme *Thermus aquaticus* by Thomas D. Brock in 1969 (Brock, 1969). Kary Mullins is an American biochemist who worked on the process of PCR in the 1980s. He successfully demonstrated the PCR process in 1985 (Saiki *et al*, 1985). His improvements on PCR which were based on the use of taq polymerase which is heat resistant unlike DNA polymerase, revolutionised PCR and its application in science. For his work on PCR, Kary Mullins was awarded the Nobel Prize for Chemistry in 1993.

PCR is more sensitive, allowing amplification of minute amounts of DNA. It has also allowed the use of microsatellites and other genetic markers than being limited to using VNTRs. It has also allowed for simultaneous amplification

of more than one target sequence in one PCR reaction in a process called a multiplex PCR reaction.

1.2.Multiplex PCR reactions

Multiplexing has allowed amplification of multiple loci at the same time, an advantage in forensic application. (Chamberlain *et al*, 1988). It is cost efficient and less labour intensive than performing individual PCR reactions as there is a decrease in the PCR reagents used and preparation time is cut short.

Since a multiplex PCR reaction is essentially multiple PCR reactions taking place at once, the success of the reaction lies in getting all the different components from each reaction to work together to produce successful amplification of all loci that make up the multiplex. Some of key factors to consider include:

1.2.1. PCR primers- (length, GC content):

The length (18-24 bp) and Guanine and Cytocine (GC) content (35 – 60 %) are the two principle factors to keep in mind when designing primers (Reviewed in Edwards and Gibbs, 1994; Innis *et al*, 2012). The length and GC content has an influence on the formation of primer artefacts (primer dimers and hairpin formations) which has a direct impact on the PCR yield. Primer concentration also contributes to the formation of primer artefacts and should be considered as well.

1.2.2. Annealing temperature and cycling conditions:

The annealing temperature for a multiplex reaction is determined by the primer length and GC content. Therefore, the primers should be designed to have annealing temperatures in the same range (56 – 60 ^{o}C) although lowering the annealing temperature by 4 – 6 ^{o}C for co-amplification of multiple target areas has proven successful (Markoulatos *et al*, 2002). Extension times also need to be considered to allow complete amplification of all loci.

1.2.3. Competing components

Interaction of the other components in PCR amplification is also very important. With multiple reactions taking place simultaneously, the polymerase enzyme, buffers and dNTPs amounts should be balanced to allow sufficient reagent amounts for each individual loci being amplified (Reviewed in Edwards and Gibbs, 1994).

The discrimination power of a forensic marker is vital. As important as it is to include an individual as a donor of DNA it is just as important to be able to exclude an individual as a DNA donor. The application of PCR multiplexing forensic analysis has allowed an increase of loci to be amplified simultaneously as discrimination power increases with an increase of the number of loci included in a panel of forensic markers.

Multiplex PCR reactions are a key element to any forensic lab and have eliminated tedious laboratory work by allowing streamlining of amplifications and allowing for high throughput of sample processing.

1.3. DNA sources

With the advances in forensic technology, it broadened the type of samples to be analysed. PCR allowed analysis of samples on nuclear level and therefore any DNA sample could be analysed including blood, hair and tissue, allowing for potentially any type of body fluid evidence collected at crime scenes to be analysed for forensic purposes (Table 1.1).

Table 1.1 Common types of DNA samples collected for forensics analyses.

Type of crime	**Exhibit**	**DNA Source**
Sexual assault	Intimate swabs from genital areas, Oral swabs, Underwear, Condoms	Semen Saliva Epithelial cells, hair
Murder	Bloody clothing Murder weapons Human remains	Blood Epithelial cells, Saliva Bones, Human tissue, hair
Burglary	Food items, cigarettes, clothing items, weapons	Blood, Saliva, Epithelial cells

1.4. Forensic markers

The choice of forensic markers plays a key role, and depends on the type of forensic samples being processed and the different forensic analysis being performed. To be a useful forensic marker, the following qualities serve as a few key requirements:

i) high discrimination power

ii) amplification from low template concentration

iii) good multiplexing ability

iv) no complicated and expensive instrumentation or chemistries required

v) distribution throughout the genome

The different types of forensic markers have evolved from the protein genetic markers and the first single nucleotide polymorphism (SNP) markers in the human leukocyte antigen (HLA) locus (reviewed in Budowle and Van Daal, 2008). The protein markers, although polymorphic, lacked discrimination power and did not allow analysis of all body fluids or tissue. The human leukocyte antigen locus provided sensitivity but lacked in discrimination power and did not perform well with DNA mixtures.

The forensic markers currently used are at the genetic level, allowing any body fluid type that contains nucleated cells to be forensically analysed, overcoming the limitations of the initial human identification systems used, which only allowed specific types of body fluids to be analysed (reviewed in Budowle and Van Daal, 2008). The high concentration requirement was overcome with the use of PCR which allowed amplification of small amounts of DNA.

1.4.1. Current forensic DNA markers

1.4.1.1. Short Tandem Repeats (STRs)

STRs are the most popular choice of forensic marker (Edwards *et al*, 1991; reviewed in Guichoux *et al*, 2011). They are widespread throughout the human genome and contain repeat units of two to seven base pairs in length. STRs differ in the length and the number of repeat units. Autosomal STRs are the most commonly used type of STR used in forensics and are present on autosomal chromosomes. A huge factor for autosomal STRs as choice forensic marker is due to them meeting the requirements of a useful forensic marker including being highly polymorphic. Autosomal STRs are also used in kinship analysis. Their application in kinship analysis is common but is not without shortcomings. These

include the presence of mutations and null alleles which can lead to false exclusions (Chakraborty and Zhong, 1994; Brinkmann *et al,* 1998). Despite their usefulness, STRs have limitations when dealing with forensic samples of a degraded nature. To increase the success rate of DNA typing of degraded DNA, a different approach has been applied. MiniSTRs can successfully be used to amplify highly degraded DNA. The principle is based on decreasing the PCR product size being amplified by moving the primers closer to the STR region (Wiegland and Kleiber, 2001; Butler *et al*, 2003). The advantages of the technique rests on the fact that these minSTRs are still compatible with loci being used for database purposes while producing better results than the conventional STRs.

Y-STRs are another type of forensic marker that has become useful in forensics (Butler, 2003). Y-STRs are present on the male Y chromosome, specifically on the non-recombining portion of the Y chromosome (NRY). The NRY is useful in tracing paternal lineage as it is passed on from father to son, unchanged, with variation only arising from a mutational event (Reviewed in Jobling *et al*, 1997). This important feature is useful in paternity testing, paternal lineage studies as well as sexual assault cases where DNA mixtures are common and it is difficult to differentiate between multiple male donors. To date, Y-STRs have been extensively researched and resulted in a high amount of Y-STR loci, with over 400 Y-STRs having been characterized and 186 Y-STRs having been investigated (Jobling *et al*, 1997; Ballantyne *et al*, 2010).

Mutations play an important role in the use of forensic markers in forensics. Mutational events are measured by comparison of the genotype between parent and offspring. STRs also have low mutation rates with an average mutation rate of 10^{-3} (Reviewed in Goldstein and Pollock, 1997). Mutation rates vary between loci with certain loci having a higher mutation rate than others. The mutation rates of the thirteen core STR loci as used by the Federal Bureau of Investigation (FBI) for their Combined DNA Index System (CODIS) system have been researched and are available at (http://www.cstl.nist.gov/biotech/strbase/mutation.htm).

In kinship analysis, mutations have a direct impact on result interpretation as mutations could be the difference between exclusion and an inclusion. Paternity determination is based on the principle that a paternal parent will pass on the same Y-chromosome to his son based on the lack of recombination of the NRY portion. A mutation at the Y-STR loci will result in a different Y-STR haplotype being observed and consequently, exclusion. But the lack of recombination places a limitation on the use of Y-STRs when it comes to individual identification which is ideally required for forensic purposes. As the NRY portion is passed on through paternal lineage, the implication is that no one male individual can be identified as the donor of DNA as any of the male family members (brother, uncle etc.) will be in possession of that Y-STR haplotype.

Recent studies have uncovered the presence of rapidly mutating Y-STRs (RM Y-STRs) (Ballantyne *et al*, 2010; Ballantyne *et al*, 2012). These Y-STR loci have a higher mutation rate than other Y-STRs that have an average mutation rate of 2.8 X 10^{-3} (Kayser *et al*, 2000; Kayser and Sajantila, 2001). During a study by Ballantyne *et al* (2010) a mutation rate above 1 x 10^{-2} was noted for the RM-STRs investigated. The discovery of RM Y-STRs presents the possibility of their use in differentiating between individual male family members. Due to the high mutation rates of these Y-STRs, unique Y-STR haplotypes are assigned to male individuals that are consequently not identical to the haplotypes of their male family members. Future applications in population studies are also possible, especially in populations that have undergone certain changes like population substructure or cultural effects, resulting in limited Y-chromosome diversity.

1.4.1.2. Single Nucleotide Polymorphisms (SNPs)

Among human polymorphisms, single nucleotide polymorphisms are the most abundant class. SNPs have characteristics that make them favourable for forensic application: they have a low mutation rate (10^{-8}) and they can be amplified in short amplicons, they are ideal for high throughput processing and can be analysed using the current automated instrumentation (reviewed in Amorim and

Pereira, 2005). This is ideal when considering the needs in forensic analysis i.e. high throughput and low costs.

There are different DNA typing methodologies that can be applied, with the SNPs typing method chosen being determined by the outcome required. Applications of SNPs include ancestry informative markers (AIMs), identity testing for individualization and lineage SNPs located on the mitochondrial DNA genome and Y chromosome (Reviewed in Budowle and Van Daal, 2008). Each typing method also has its own requirements for the type of SNPs chosen. For example, the SNPs used for identity testing has to have high heterozygosity and low population heterogeneity whereas SNPs used as AIMs would require low heterozygosity and high population heterogeneity.

There has been an attempt to create a SNP multiplex assay for the purpose of human identification. The assay was created by the consortium group SNPforID group and consists of 52 unlinked autosomal SNPs which are highly polymorphic within the European, Asian and African populations (Sanchez *et al*, 2006). The aim of the project was to create a multiplex SNP assay which consisted of all the qualities needed to be used for forensic typing, which includes multiplexing and the use of high throughput DNA typing platforms. The project successfully identified 52 polymorphic SNPs and used standard instrumentations and methods available in forensic laboratories, further allowing the possibility of applying the 52 SNPs for other genotyping methods.

Due to the biallelic nature of SNPs, SNP application requires a higher number of SNPs loci to equal the discrimination power of the current STR kits in use (Sobrino *et al*, 2005; Reviewed in Budowle and Van Daal, 2008). Furthermore, the characteristics of the population of interest should be considered including allele frequencies and genomic location as these have an influence the discrimination power (Zhong *et al*, 1999). A minimum of 50 SNPs are needed to equal the discrimination power of 10 to 15 STRs (Gill, 2001; reviewed in Amorim and Pereira, 2005). The role of mutations and null alleles in SNPs, as with all forensic markers, and their influence in population genetics needs to be considered as well. Furthermore, the biallelic nature of SNPs makes interpretation

of mixture profiles difficult. The advantages of SNPs might also be considered as disadvantages, depending on the needs and outcomes required by the user. For forensic requirements, SNPs are difficult to streamline and might be a better alternative as a supplementary test in conjunction with other established DNA typing methods. Another drawback of SNP application is the time taken to analyse SNP results, as analysis of data for a SNP panel in excess of fifty SNPs is time consuming. However, where STR markers fail in analysis of degraded DNA, SNPs can be a welcome alternative in generating successful results.

1.4.1.3. Insertion-deletion polymorphisms (Indels)

Indels are known as length polymorphisms with the insertion or deletion of nucleotides in the genome (Weber *et al*, 2002). Insertion-deletion polymorphisms are a result of a mutational event that occur possibly only once in the human evolutionary past (Pereira and Gusmao, 2012). As a result of the mutational event, indels can have an ancestral and mutational state (two allelic states) and are therefore known as biallelic polymorphisms (Pereira and Gusmao, 2012). What have sparked interest in indels are the qualities they possess; combining the properties of SNPs and STRs that have made these polymorphisms favourable choices as forensic markers for forensics:

i) Indels are distributed throughout the genome, just like SNPs and STRs. Indels make up about 20 % of human polymorphisms (Weber *et al*, 2002).

ii) They possess low mutation rates. This is especially applicable to relationship testing and identification purposes.

iii) Due to the nature of forensic samples, the system chosen to analyse forensic samples need to be able to handle samples of a degraded nature. Indels can be amplified in short amplicons. This quality allows for PCR amplification of degraded DNA as well as being an advantage for multiplexing purposes.

iv) Genotyping can be performed using current genotyping methods (PCR and capillary electrophoresis) and does not require expensive chemistries or instrumentation. Therefore they are cost efficient and provide uncomplicated and speedy analysis.

1.5. Commercial kits

In forensics, multiplex PCR reactions have successfully been used for amplification of STR loci. Due to the time and costs involved in designing and optimising PCR multiplex reactions, most forensic laboratories make use of commercially manufactured multiplex kits. These kits contain all the components for successful PCR amplification with initial optimisation of reagent concentrations performed. The first multiplex system consisted of four STR markers (THO1, FES/FPS, vWA and F13A1) and was developed by the Forensics Science Services in the United Kingdom (Clayton *et al*, 1998). Multiplex system development has advanced to an increase in the number of loci as well as the inclusion of a gender marker (Amelogenin). As previously mentioned, an important advantage to the increase of the number of loci used is the increase in discrimination power. The commercial kits make use of established and known loci that have been extensively researched. Thirteen core STR loci (CSF1PO, FGA, THO1, TPOX, vWA, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51 and D21S11) were established for inclusion within CODIS (Budowle *et al*, 1998). These core thirteen loci as well as other autosomal STR loci have been incorporated by companies within and outside the United States for use in their STR multiplex kits. The amount of loci have also increased with as many as 24 STR loci now being amplified within PCR multiplex kits, as in the case of the GlobalFiler® Express PCR multiplex kit (Life Technologies) and PowerPlex® Fusion System (Promega). Y-STR multiplex systems also consist of increased Y-STR loci amounts such as 23 loci in PowerPlex® Y23 System (Promega) and 27 loci in Yfiler® Plus PCR Amplification Kit (Life Technologies).

The kits undergo stringent optimisation and validation by the manufacturer, essentially removing the time, cost and labour involved in developing and optimising an in-house PCR multiplex system. These developments have provided standardization of protocols between laboratories, with comparative DNA results being produced as laboratories are using similar protocols and forensic markers globally. Companies that commercially provide kits for autosomal STRs, Y-STRs and indels include Life Technologies, Promega Corporation, Qiagen and Biotype, to name a few. Over fifty kits are commercially available for autosomal STRs alone. These commercial kits need to meet strict quality requirements including the validation studies mentioned previously.

1.5.1. Qiagen® Investigator DIPplex kit

The Qiagen® Investigator DIPplex kit is the first commercial indel kit. It allows for multiplex PCR amplification of 30 indel loci including Amelogenin for gender determination. The 30 indel loci are located across 19 autosomal chromosomes. Twenty eight loci are located in intronic coding regions and two loci (HLD133 and HLD128) are located in intergenic regions (Rajeevan *et al*, 2003). The 30 loci are detected using fluorescence-labelled primers using 6-FAM™, BTG, BTY and BTR. The advantages of the kit include the absence of stutter peaks which further assists with DNA mixture interpretation and an increase of discrimination power when applied in addition to standard markers (Qiagen, 2011).

The Qiagen® Investigator DIPplex kit has been evaluated in other countries. Population data for European countries including Finland, Germany, Denmark and North Portugal have been generated using the kit (Friis *et al*, 2012; Qiagen; Neuvonen *et al*, 2012; Carvalho and Pinheiro, 2013), as well as Asian countries namely Chinese and Korean populations (Li *et al*, 2011; Liang *et al*, 2013; Kim *et al*, 2014). African American populations were also evaluated (Fondevila *et al*, 2012; LaRue *et al*, 2012).

In certain populations namely the Spanish (Martin et al¸ 2012), African American (Fondevila *et al*, 2012, LaRue *et al*, 2012), Danes (Friss *et al* , 2012) and Iranian

populations (Poulsen *et al*, 2015), the presence of a null allele was detected at the HLD97 locus. A lack of population data for African populations is evident with only two studies (Fondevila *et al*, 2012, LaRue *et al*, 2012) reporting on African populations which makes the evaluation of the kit on South African populations valid.

1.6. Quality control and standardization

A very important and vital part of any laboratory is ensuring that quality standards and procedures are met and adhered to. In the field of forensics, this is even more so, as results obtained in forensic laboratories are used in criminal cases as crucial evidence. Furthermore, standardization of techniques are important for results to be replicated and compared for inter laboratory comparisons. An international forensic community exist consisting of organizations that determine standards and guidelines that aim to provide additional guidance to the bigger DNA community worldwide including Scientific Working Group on DNA Analysis Methods (SWGDAM) (www.swgdam.org), International Society for Forensic Genetics (ISFG) (www.isfg.org) and European DNA Profiling Group (EDNAP) (www.isfg.org/EDNAP), among others. Guidelines as recommended by SWGDAM can be accessed on the SWGDAM website (http://swgdam.org/docs.html).

Some of the main issues that are looked at include the following:

1.6.1. Laboratory staff competency

Personnel performing analysis need to be suitably qualified, trained and competent. This is a process that needs to be monitored regularly. One such measure is performing regular proficiency tests. Proficiency tests evaluate not only the analyst's ability but also the laboratory's ability to obtain consistent results (Butler, 2005).

1.6.2. Instrumentation and chemistries

Instruments have to serviced and maintained regularly. Instruments need to perform at their best with no possibility of incorrect results being reported due to technical errors. Servicing and maintenance is normally performed by the manufacturers of the instruments. These companies issue servicing and calibration certificates as proof of the maintenance performed which is often requested by the courts. The chemistries used in forensic science need to be evaluated and validated. Validation studies are carried out to ensure that the procedures and technologies laboratories implement are done so accurately and results obtained reflect these procedures. Developmental and internal validation studies should be carried out:

1.6.2.1. Developmental validation

These tests are performed, usually by the manufacturers of the kits, to test the chemistries and technologies used in the kits including the primer sets and the chosen marker sets. It includes a very extensive list of tests, testing for different variables including consistency, to test if a technique produces the same result each time it is performed, whether by the same or different laboratories. Environmental stress is another variable that is tested, again to determine if the same result is obtained repeatedly. Environmental factors play an important role in forensics as exhibits are often exposed to extreme weather conditions which influence the DNA quality. Validation tests include population studies, environmental studies and reproducibility studies.

1.6.2.2. Internal validation

These validation tests are performed by testing laboratories to test the experimental procedures and protocols established during validation, to see if the same processes can be successfully applied in one's own laboratory.

1.6.2.3. Laboratory competency

Besides the individual sections or components that the laboratory consists of, laboratories as a whole are also evaluated to assess how these individual components perform together. This is monitored through laboratory audits that evaluate the laboratory operation in its entirety (Butler, 2005). Laboratories can also go through the process of being accredited. The assessment is performed by an accreditation body which stringently inspects every aspect of the laboratory and its performance relating to good lab practices (Butler, 2005).

The process of accreditation is a method of determining the competency of laboratories to perform specific types of testing, measurement and calibration. Through accreditation, customers are provided assurance of reliability and competency in the services provided by laboratories. For laboratories, accreditation provides a means to determine workmanship quality and accuracy. Assessments are regularly performed to determine whether standards are maintained and to pinpoint possible areas for improvement. The process of accreditation is recognised nationally and internationally as a measure of competency.

Different accreditation bodies exist worldwide. The accreditation body recognised in South Africa is the South African National Accreditation System (SANAS) (http://www.home.sanas.co.za/).

Depending on the types of testing, measurement and calibrations performed by a laboratory, different standards can be implemented. For many of the accreditation bodies, ISO 17025 is now being used as the basis for accreditation for calibration and testing bodies and it is also the standard used by forensic science laboratories. This further supports inter-laboratory collaboration and result comparison.

1.7. Technological advances

1.7.1. Instrumentation

The forensic community has made huge strides in cutting down on labour and increasing sample processing in terms of speed and efficiency by improving the instrumentation used for forensic applications. It has moved from the forensic analyst doing most of the labour involved in analysis to semi automation, rendering the forensic analyst a mere operator of the instrumentation. Companies have seen the potential of instrumentation and have started focusing on manufacturing instruments to be used for routine analysis. Among the companies on the forefront of instrumentation development are Merck, Promega, Life Technologies and Bio-Rad. These companies have released instruments for performing PCR amplification, quantification, gel electrophoresis and capillary electrophoresis. Forensics has also seen the use of robotics increasing for routine analysis in forensic laboratories. These include the use of liquid handlers manufactured by companies like Hamilton and Tecan to perform routine laboratory techniques. Liquid handling work stations can be applied successfully at all levels within a forensic laboratory: DNA extraction, quantification, PCR setup and post PCR levels. The use of liquid handlers has helped to decrease the rate of errors due to human handling in forensic processing. One major advantage of robotic systems is the decrease in error rates as opposed to the error rate due to human handling. Although the time taken by robotic systems to complete the work is not necessarily less, the quality of the work is of a higher standard. The human factor can also never fully be removed or replaced, but the use of robotics does aid in decreasing the dependency on analysts to perform technical tasks.

1.7.2. Electrophoresis

The initial detection platforms for separation of genetic markers were labour intensive with a lot of time taken up by preparation and sample loading. These platforms were based on polyacrylamide slab-gels and silver staining (Deforce *et al*, 1998). A limitation of the earlier slab-gel method was that it did not allow for

overlapping of loci and therefore limited the amount of loci to be incorporated in multiplex systems, a limitation overcome by the latest capillary electrophoresis platforms incorporating the use of fluorescent labelling.

With the advancement of capillary electrophoresis platforms, development has contributed to increasing output and turnaround times. It has been successful thus far and has allowed for more samples to be amplified simultaneously (increased output) and turnaround time being decreased from hours to minutes (increased turnaround time), compared to the gel based detection systems. Capillary electrophoresis instruments have also become more sensitive with lower input DNA concentration being required, further contributing to successful detection of DNA profiles. Among the companies previously mentioned is Life Technologies, a leading manufacturer of genetic analysis systems including capillary detection instrumentation. Their instruments are being used in forensic laboratories across the world, including South African government forensic laboratories.

1.8. Statistical analysis of forensic profiles

1.8.1. The importance of population genetics in forensics

In forensics, the value of a match between DNA profiles is estimated using statistics. The strength of the match is demonstrated by calculating, through population genetic principles and statistical analysis, the rarity of a DNA profile within a population. Simply put, the rarer the occurrence of a DNA profile within a population, the less likely the possibility of someone else having left the DNA evidence at the crime scene. Reviewing the methods of evaluating the validity of a match is beyond the scope of this thesis. However, evaluating these indel loci for forensic applications are valid and it requires evaluation of forensic and population genetics parameters.

1.8.2. The basic principles of population genetics

Population genetics is the study of allele and genotype frequencies at different loci, within a population and the factors that influence them (Goodwin *et al*, 2011). The Hardy-Weinberg principle is based on the works of Godfrey Hardy and Wilhelm Weinberg and is a method which describes the relationship between allele and genotype frequencies within a population (Butler, 2005). Using the Hardy-Weinberg principle, allele and genotype frequencies can be estimated in the population. The Hardy-Weinberg principle requires that the population meets certain assumptions:

i) Infinite population size
ii) No natural selection has taken place
iii) Random mating takes within the population
iv) No mutations take place
v) The effects of migration does not influence the population

A naturally occurring population cannot meet all these requirements as it represents the qualities of a perfect population, which realistically does not exist. The Hardy-Weinberg equation can still be applied for estimation of allele frequencies and the amount of deviation from Hardy-Weinberg equilibrium within a population can be statistically calculated.

To be able to apply statistical analysis to a data set, there are requirements for the data set and the samples used to generate it. Firstly, a population is of infinite size and therefore immeasurable. Because we cannot sample all individuals within a population, samples from individuals belonging to a specific population represents a sample size of that population. The minimum amount of samples required to be representative of a population is a minimum of 100 samples (Chakraborty, 1992). Secondly, individuals that form part of population sampling must be non-related. The diversity in the population is measured by the allele frequencies within the population. If the samples representing a population consists mostly of related individuals they will share the same alleles, and not give a true reflection of the population's diversity.

1.8.3. The importance of population data

Genetic variation of populations is strongly influenced by evolution and geographical characteristics. These would include how isolated populations are, the size of populations, if there is migration between different populations and the ancestry and origin of individuals belonging to populations.

There are different methods that can introduce genetic variation. Mutation allows for new alleles to be created and can be caused by environmental or chemical changes. Migration occurs when individuals from different populations move between populations and introduce genetic variation. Another method is sampling errors due to genetic drift which results in allele frequency changes in populations of small sample size. Genetic drift is the random change in allele frequency between generations in populations due to finite samples of individuals, gametes and alleles that contribute to the next generation (Hamilton, 2009). Genetic drift is caused by lack in population size increases over generations, with random change increasing as the population size decreases. It can also be caused by the founder effect, when a population is established by a small amount of individuals resulting in limited genetic variability despite growth in population size over generations. All these factors have an influence on genetic variation whether it is a positive or negative effect. Evolutionary and geographical changes can influence the survival of populations by eliminating rare genetic variants or it can influence the establishment of new species.

To better understand the evolutionary and genetic changes that populations go through populations are studied through population genetics. Sample of populations are taken and population statistics are performed on the population data and the results analysed. Useful insights on populations can be obtained by studying populations in this way.

1.9. DNA databases

In forensics, DNA databases have emerged as a new and powerful tool to crime fighting and prevention using DNA. Different DNA databases exist to fulfil numerous functions. Criminal DNA databases are used to detect repeat offenders. A criminal database would contain DNA profiles of samples collected from crime scenes and DNA profiles of all offenders and individuals arrested on suspicion of criminal involvement (arrestees). With their first offence, a criminal's DNA profile is collected and added to the criminal DNA database. Upon a second offence, a search on the DNA database will pick up the criminal's DNA profile, ultimately linking the offender to the crime and revealing an existing criminal offence. One major advantage to a criminal database is the elimination of time and man power in searching for a possible suspect. Criminal databases also play a key role in excluding possible suspects and identifying the true perpetrator of a crime.

The first national population DNA database was a criminal DNA database established in the United Kingdom in April 1995 and is known as National DNA Database (NDNAD) (Martin, 2004; Butler, 2005). The NDNAD originally made use of six STR loci (vWA, D8S1179, D21S11, D18S51, THO and FGA) including a gender marker but there has been an addition of loci (D3S1358, D16S539, D2S1338 and D19S433) that has expanded the original six loci to ten loci (Butler, 2005). After the NDNAD many population DNA databases have been established around the world to serve the forensic community. Other European countries have now established their own NDNAD or are in the process of implementing a NDNAD like Scotland, Denmark and Hungary, among others (Martin, 2004). South Africa has also realised the power of a criminal database in the fight against crime and legislation has recently (January 2014) been passed for the establishment of a criminal DNA database.

The FBI manages CODIS which forms part of the National DNA Index System (NDIS) in the United States that was launched in 1998 (Budowle *et al*, 1998). The success of a DNA database is based on the amount of DNA profiles contained within. All fifty states within the United States are currently participating members that are making use of CODIS and resulting in its success. Further

aiding in its success is the use of thirteen core CODIS STR loci (CSF1PO, FGA, THO1, TPOX, vWA, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51 and D21S11) that are being used by participating laboratories in the United States. This allows for easier exchange of information and comparison of data between participating laboratories. Following successful application of CODIS, the FBI offers assistance to other countries making use of DNA databases. A stand-alone version of the CODIS software has been developed for use by other countries outside of the United States and the FBI also hosts a national CODIS conference, annually. Information on the CODIS system is available on the FBI website (http://www.fbi.gov/).

The value of databases have been realised in assisting with across boarder crimes, as criminals do not necessarily operate in one area or country With this in mind, many of the countries with operational NDNADs have adopted the use of the same loci within their databases. This allows collaboration and exchange of information between countries. Interpol is the largest international police organization with its main aim to assist police around the world to work together to prevent crime. Interpol has also realised the value of DNA databases and have created a platform for countries allowing data exchange called the Interpol Gateway which allows countries worldwide access to their database (http://www.interpol.int/en). Differences in legislation and database laws exist for different countries. The European Network of Forensic Science Institutes (ENFSI) is a DNA working group that aims to bring together organisations that pursue forensic DNA analysis methods for the purpose of exchanging and disseminating information on forensic applications (www.enfsi.eu). The ENFSI makes available information on DNA databases relating to the database laws that exist in European countries.

DNA databases can also serve a purpose in identifying missing persons or individuals killed in mass disasters. DNA profiles of recovered biological remains or unknown individuals are added to a DNA database and a subsequent search of possible relatives' DNA profiles against the DNA database can aid in locating a missing loved one. The application of DNA databases have been successfully

implemented in Spain. The database contains DNA profiles of samples donated voluntarily for identification of cadavers and human remains (Lorente *et al*, 2002). The same purpose can be fulfilled in cases of mass disasters where large scale deaths make identification of the missing and deceased difficult (Alonso *et al*, 2005), as in the case of the World Trade Centre disaster and the Thai Tsunami of December 2004.

DNA allele frequency databases are used in forensics to determine the occurrence of a DNA profile in a population. These DNA databases consist of DNA profiles sampled from individuals in a population with the aim of determining the occurrence of the alleles within the population (Butler, 2005). The identities of the individuals are unknown and the DNA profiles are used purely for the purpose of determining the allele frequencies, based on the individual's ethnicity. One such database is the United States Y-STR Database (www.usystrdatabase.org/). This database is used to estimate Y-STR haplotype population frequencies using 11 to 29 Y-STR haplotypes for four populations occurring in the United States i.e. African American, Caucasian, Asian and Native American. Another allele frequency database is the Y chromosome Haplotype Reference Database (YHRD) (Willuweit *et al*, 2007). This database was established in 1999 and provides Y-STR haplotype frequencies for populations worldwide as well as assessment of male population stratification based on Y-STR haplotype frequencies.

1.10. DNA forensics in South Africa

Crime statistics was released by the South African Police Service (SAPS) in September 2014 (www.saps.gov.za). The crime figures in South Africa do not create a positive picture with specific types of crime continuing to be on the increase, placing pressure on SAPS and the Criminal Justice System.

Currently all criminal forensic case work is being processed by the South African Police Service's Division Forensics at their forensic science laboratories across the country. The SAPS allele frequency database is managed by SAPS Forensics Science Laboratories. This database was established by using convenience

samples that represent the four South African populations (Caucasian, Black, Coloured and Asian). It is used to perform statistical calculations, to determine how rare a DNA profile is in the South African populations. The allele frequency database also contains DNA profiles obtained from volunteers and is not a criminal DNA database.

In September 2009 the Office for Criminal Justice System Reform made recommendations relating to a forensic DNA database allowing for the establishment of a criminal DNA database containing DNA profiles collected from suspects arrested for criminal activity. The criminal DNA database would be known as the National Forensic DNA Database (NFDD) and the main aims of the database are:

i) to serve as a criminal investigative tool;
ii) identify repeat offenders;
iii) assist in proving the innocence or guilt of accused individuals and
iv) assist in identification of human remains and missing persons.

The Portfolio Committee on Police consists of members of Parliament's National Assembly. The purpose of a portfolio committee is to consider bills, deal with departmental budget votes and oversee and make recommendations on the specific department they are responsible for. The Portfolio Committee on Police consulted with various entities including the National Prosecuting Authority and Department of Justice and Constitutional Development during drafting of the DNA Bill as well as embarking on a visit to Canada and the United Kingdom in an effort to familiarise themselves with issues relating to the DNA legislation. In April 2013 the Criminal Law (Forensic Procedures) Amendment Act [B9-2013] was approved by Cabinet and on 12 November 2013 the DNA Act was approved by the National Assembly. On 27 January 2014 the Criminal Law (Forensics Procedures) Amendment Act 37 of 2013 (DNA Act) was passed into law (Government Gazette, 2014).

The financial cost of implementing the DNA Act is high with training of police officials and forensic awareness programmes only, calculated to be approximately R22 million.

Despite the high financial cost to the South African government, establishment of the NFDD will prove to be a worthwhile investment in combatting crime in South Africa.

1.11. South African population

South Africa has a rich and colourful history which is deeply rooted in South Africa's people and their heritage. South Africa's population consists of different and a wide variety of languages. The different population groups within the South African populations originate from all over the world with the Afrikaner group being from French, Dutch and German descend and the Coloured population being an admixed group of Khoisan, Bantu, Asian and European ancestry.

The first people responsible for South Africa's diversity were the Khoisan people, also known as the Bushmen and Hottentots. They were hunter-gatherers who hunted using bow and arrow and lived off the land. The Khoisan people are sadly on the brink of extinction. They are being driven from their ancestral land and prevented from hunting.

The first Europeans to arrive in South Africa were European settlers from Britain and Netherlands. The Dutchman Jan Van Riebeeck arrived in the Cape of Good Hope in 1652 bringing with him his crew of ninety men. His purpose was to build a fort to supply Dutch ships. The White South African population of today are descendants from these European settlers.

The Coloured people are also referred to as people of mixed race and are found to reside predominantly in the Western Cape. The Coloureds originated from admixture between the Khoisan, Bantus, Europeans and slaves that the European settlers brought to South Africa. These slaves were brought from Madagascar and

South East Asia. Immigration was also encouraged which resulted in Indian and Chinese labourers establishing themselves in South Africa during the 1600s.

The largest population group that make up the South African population is the Black Africans known collectively as the Bantu who are descendants from Black Africans who took part in the Bantu expansion from South East Nigeria and Western Cameroon by A.D. 300 (Montano *et al*, 2011). The Bantu people consist of different ethnic groups which include Zulu, Xhosa and Tswana and speak different Bantu languages. They are followed by the Coloured, White and Indian population groups, respectively. This colourful mix of different ethnicity, cultures and religions makes up South Africa's 'rainbow nation' as originally referred to by Archbishop Desmond Tutu, when he referred to the post-Apartheid South African population.

From the last census conducted in 2011, South Africa's population was estimated at 51.8 million (Statistics South Africa, 2011). Black Africans make up 79.2 % of South Africa's population, Coloureds make up 8.9 % and Whites also make up 8.9 % with Indian Asians consisting of 2.5 %, with individuals not falling within the four groups being classified as "Other", making up 0.5 %.

The distribution of the populations in South Africa is illustrated in figure 1.3. The largest population group, the Black Africans can be seen to populate the major part of South Africa. The Coloured group is settled mainly in the Western and Northern Cape with the other racial groups i.e. Indian-Asian and White, found throughout South Africa.

1.12. Objectives of the project

The main objective of the present study was to evaluate the performance of the Qiagen® Investigator DIPplex kit for forensic application in South Africa. Firstly, the properties of the the Qiagen® Investigator DIPplex kit was evaluated by investigating the sensitivity, reproducibility and robustness of the kit. Secondly, the 30 loci in the Qiagen® Investigator DIPplex kit were evaluated in five South

African populations i.e. the Afrikaner, Mixed Ancestry, Indian-Asian and Bantu (Xhosa and Zulu) populations by estimating allele frequencies, forensic and population parameters to give valuable insights into the application of this kit in South Africa.

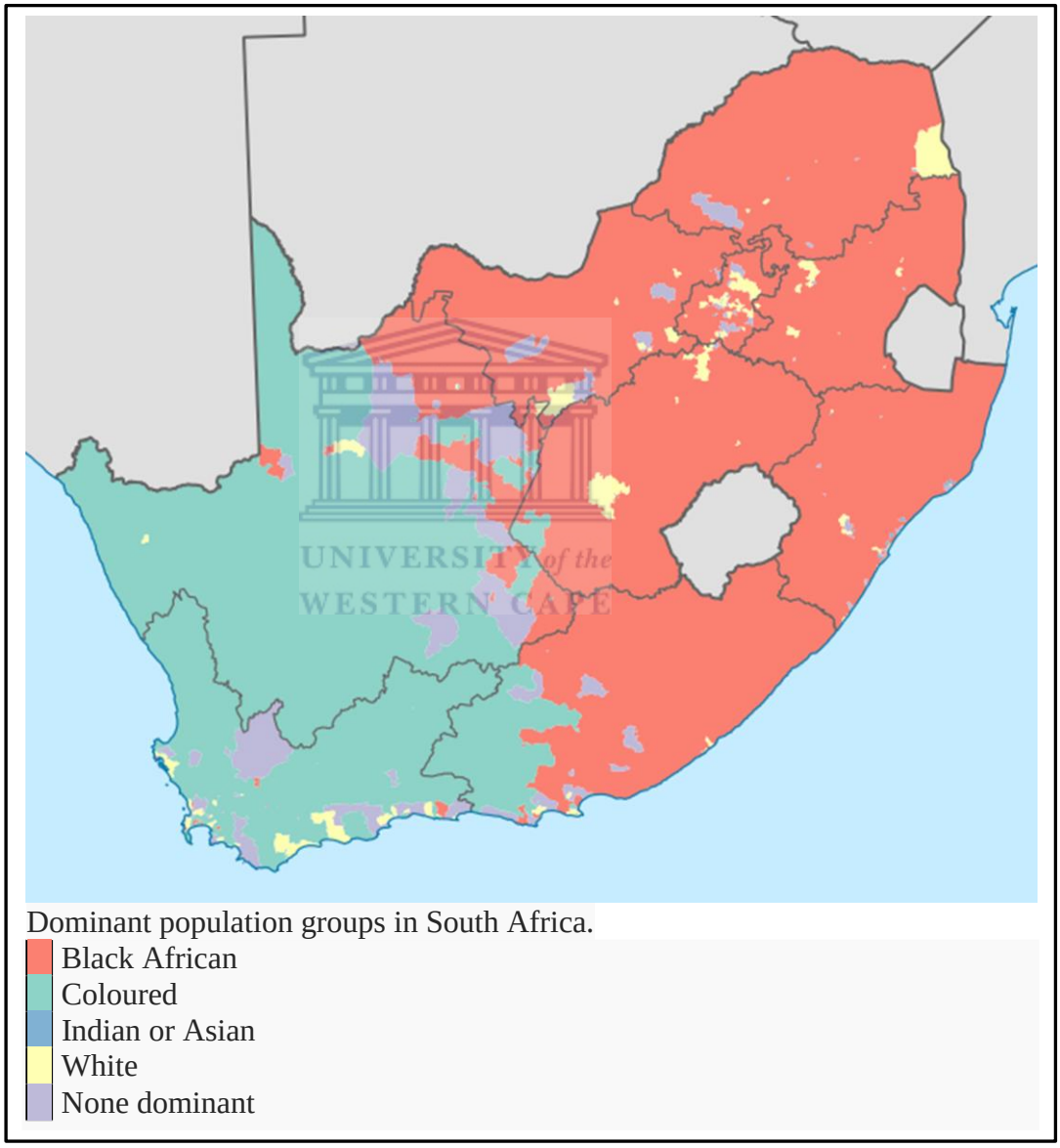


Figure 1.3. Distribution of the different population groups within South Africa based on their location within South Africa.

(Image from http://en.wikipedia.org/wiki/Ethnic_groups_in_South_Africa data provided by Stats SA)

Chapter 2

Evaluation of the sensitivity, reproducibility and robustness of the Qiagen Investigator DIPplex kit and generating population data for five South African populations

2.1. Introduction

For a PCR amplification system it is of vital importance to perform optimisation. Through optimisation it can be determined how the different elements that make up a system performs together to deliver maximum results at minimum costs. This includes optimisation of reagents and instrumentation. For a PCR system used for forensic purposes it is of even more importance to perform optimisation. The types of samples usually encountered in forensics are degraded DNA samples and DNA samples of small quantities. Therefore the PCR system is normally highly sensitive to ensure maximum detection and amplification of the DNA. Although increased sensitivity is an advantage for analysis of forensic samples, it is also a disadvantage as this also allows detection and PCR amplification of non-specific DNA, increasing the chances of DNA contamination (Sundquist and Bessetti, 2005; Champlot *et al*, 2010). Optimisation also means optimising the laboratory environment in which the work will be done. Again, due to the sensitivity of the PCR system used in forensics, the laboratory setup has a big impact on the quality of the work. Sharing of instrumentation and reagents like pipettes may be a potential contamination source. Inconsistency in decontamination and practical application of laboratory techniques among laboratory staff will allow potential contamination and make it difficult to detect and eliminate the source of contamination.

For the current project, a population dataset was generated to determine the occurrence of 30 indel alleles of the Qiagen® Investigator DIPplex kit representative of five South African populations (Afrikaner, Mixed Ancestry, Indian-Asian, Xhosa and Zulu subpopulations).

This chapter will investigate the multiplex PCR amplification system used in this study and the generation of the population dataset for the 30 indel alleles of the Qiagen® Investigator DIPplex kit.

2.2. Materials and methods

2.2.1. Contamination preventative measures and good lab practices

To limit possible contamination, a separate laboratory area was identified for the purpose of PCR amplification setup. UV lights were installed above the work benches and connected to timers to switch on daily at a specified time, for overnight exposure. A laminar flow was installed for the setup of PCR amplification. Extra attention was given to wearing full personal protective equipment namely gloves, lab coat, hair net and face mask. All work surfaces were decontaminated with a 5 % bleach solution before and after PCR setup. Gloves were changed regularly. Only filtered tips were used and all consumables were in a sealed condition before being opened and utilized. Samples were handled one at a time to prevent cross contamination. All gloves and tips were disposed of appropriately. Pipettes, tubes and filter tips were regularly exposed to UV light to further aid in contamination prevention.

The presence of contamination is indicated by the presence of amplification in the negative control and possibly by any additional peaks in the positive control. The data from any PCR batch where there were indications of contamination is excluded and PCR experiments repeated.

2.2.2. Sample collection and DNA extraction methods

To investigate the effect of different DNA extraction methods on the quality of the DNA results, DNA was extracted using two different DNA extraction methods: DNA extraction from buccal swabs and DNA extraction from whole blood (See below).

The South African populations investigated in this project consist of five population groups: Afrikaner, Mixed Ancestry (Coloured), Indian-Asian, Xhosa and Zulu (collectively referred to as Bantu). Samples were collected from 101

Afrikaner individuals, 104 Mixed Ancestry individuals, 102 Indian-Asian individuals, 102 Xhosa individuals and 103 Zulu individuals, with ethical approval granted from the Ethics committee of the University of the Western Cape (UWC). Whole blood samples were obtained from Western Province Blood Transfusion Services for individuals belonging to the Afrikaner populations, Mixed Ancestry populations and Asian Indian populations. Buccal (oral) swabs were collected from the Xhosa and Zulu populations, from the campus of the University of the Western Cape.

The ethnicity of study participants was determined using a process of self-classification, going as far back as three generations. To determine the correctness of the self-classification process participants were interviewed by laboratory staff members, with possible admixed individuals being excluded.

2.2.2.1. DNA extraction

2.2.2.1.1. DNA extraction from buccal swabs

DNA from buccal swabs was extracted using the Proteinase K and salting lysis extraction method (Medrano, 1990). The comprehensive protocol of the extraction method is presented in the appendix. The extracted DNA samples were stored at -20 ^{o}C.

2.2.2.1.2. DNA extraction from whole blood

DNA extraction from whole blood was performed using the Lahiri and Nurnberger (1991) extraction method. The comprehensive protocols of both methods are presented in the appendix. The extracted DNA samples were stored at -20 ^{o}C.

2.2.3. DNA quantification

DNA samples were quantified using the Nanodrop ND-1000 (Coleman Technologies Inc.) and the Nanodrop 2000 (Thermo Scientific). Absorbance readings at 230 nm, 260 nm and 280 nm were determined. Three samples were diluted to 0.1 ng/µL, 0.2 ng/µL and 0.5 ng/µL.

2.2.4. PCR amplification

Using buccal samples and samples from whole blood, PCR amplification was first performed as per manufacturer's recommendations. Amplification was performed in a final volume of 25 µL containing 2 µL 0.1-1.0 ng DNA, 5 µL Reaction Mix A, 5 µL Primer Mix DIPplex, 0.6 µL Multi Taq2 DNA Polymerase and 12.4 µL SABAX water (Qiagen, 2011).

PCR amplification was then repeated in a final volume that was half of the manufacturer's recommendations. Amplification took place in a final volume of 12.5 µL containing 2.5 µL 0.1-1.0 ng DNA, 2.5 µL Reaction Mix A, 2.5 µL Primer Mix DIPplex, 0.3 µL Multi Taq2 DNA Polymerase and 5.2 µL SABAX water.

Lastly, amplification was performed in a final volume of 10 µL with 40 % of the manufacturer's conditions. Amplification took place in a final volume of 10 µL containing 2 µL 0.1-1.0 ng DNA, 2 µL Reaction Mix A, 2 µL Primer Mix DIPplex, 0.24 µL Multi Taq2 DNA Polymerase and 3.76 µL SABAX water. Thermo cycling was conducted in a Veriti 96-well Thermal Cycler (Life Technologies) and Arktik™ Gradient Enabled 96-well Thermal Cycler (Thermo Scientific). Thermo cycling conditions as recommended by the manufacturers were followed: 94 ^{o}C for 4 minutes, 94 ^{o}C for 30 seconds, 30 cycles at 61 ^{o}C for 120 seconds, 68 ^{o}C for 60 minutes and a holding step at 10 ^{o}C. Samples were stored at -4 ^{o}C.

A negative control and a positive control were included with each round of PCR amplification. A volume of 2 µL SABAX water was added as a negative control.

A volume of 1 µL Control DNA XY5 serving as a positive control, as provided in the Qiagen® Investigator DIPplex kit, of the working stock solution in addition to 1 µL SABAX water was added.

Amplification of samples for the purpose of constructing a reference database was performed using the amplification protocol consisting of a final volume of 10 µL and thermo cycling conditions as recommended by the manufacturer. Optimal DNA concentration for input DNA was determined to be between 0.2 to 0.3 ng. PCR amplification was performed using input DNA of 0.2 ng for samples extracted from buccal samples and whole blood.

2.2.5. Evaluation of DNA concentration sensitivity

The DNA concentration range as recommended by the manufacturers is 0.2 - 0.5 ng with reliable results obtained for concentrations less than 0.1 ng during internal validations (Qiagen, 2011). To evaluate the system's sensitivity to different input DNA concentrations and to determine the optimal input DNA concentration for performing PCR amplification, three different input DNA concentrations were used during PCR amplification i.e. 0.1 ng/µL, 0.2 ng/µL and 0.5 ng/µL. PCR amplification was performed using three samples extracted from buccal swabs, the positive and a negative control, in a final volume of 25 µL as per manufacturer's recommendations and a final volume of 12.5 µL that was half of the manufacturer's recommendations. The impact of the different input DNA concentrations on heterozygote peak imbalance was also evaluated.

2.2.6. Evaluation of the robustness and reproducibility

To investigate the robustness and reproducibility of the system, PCR amplification was tested in two samples, the positive and a negative control, in 2X PCR duplicates, using the three final volume experimental conditions indicated in section 2.2.4.

2.2.7. Capillary Electrophoresis

PCR amplified product was prepared for capillary electrophoresis by loading 1 µL unpurified PCR amplicon in a 96 well PCR plate. A cocktail consisting of size standard and HiDi formamide was prepared by adding 0.3 µL DNA size standard 550 (BTO) (Qiagen) and 8.7 µL HiDi formamide together, per sample. The cocktail of HiDi formamide and size standard was mixed using the pipette and spun down. A final volume of 9 µL cocktail was loaded to the 1 µL unpurified PCR product. The samples were denatured at 95 ^{o}C for 5 minutes in the Veriti 96-well Thermal Cycler (Life Technologies) and snap cooled on ice. Capillary electrophoresis was performed on a 3500 Genetic Analyzer (Life Technologies).

2.2.7.1. Allelic ladder

An allelic ladder is a DNA sample consisting of all possible alleles. It is used for correct sizing of DNA fragments. A volume of 1 µL allelic ladder DIPplex (Qiagen) was loaded with each batch of samples prepared for capillary electrophoresis. A cocktail consisting of size standard and HiDi formamide was prepared by adding 0.3 µL DNA size standard 550 (BTO) (Qiagen) and 8.7 µL HiDi formamide together, per sample. The cocktail of HiDi formamide and size standard was mixed using the pipette and spun down. A volume of 9 µL of size standard/ formamide cocktail was added to the allelic ladder.

2.2.7.2. Data Collection and Data Analysis

Data was collected using the 3500 data collection software v1.0 (Life Technologies). Data analysis and allele designation was performed using the Qiagen® Investigator DIPplex template files (Qiagen) and the Genemapper® ID-X software v1.2 (Life Technologies).

2.2.8. DNA profile quality

The quality of the DNA profiles was assessed by looking at peak height, peak height ratio, allele designation and overall DNA profile quality. For analysis performed using Genemapper® ID-X software v1.2 (Life Technologies), quality settings according to the manufacturer was used which includes the following criteria: a minimum peak height for homozygous samples of 200 relative fluorescent units (rfu) and 100 rfu for heterozygous samples; maximum peak height of 5000 rfu; minimum peak height ratio of 0.7 for heterozygote samples.

2.3. Results

2.3.1. Evaluation of DNA extraction methods

DNA extracted from buccal swabs and whole blood, using two different extraction methods, was used to perform PCR amplification in a final volume of 10 µL with 40 % of the manufacturer's conditions. The DNA quality and quantity was evaluated by assessing the absorbance readings ratios 260/280 determined in Section 2.2.3. A 260/280 ratio in the range of 1.8 is indicative of good DNA purity. Readings lower than this expected value may be indicative of the presence of contaminants.

On average the 260/280 ratio for DNA extracted from whole blood was higher than 1.8 compared to the 260/280 ratio for DNA extracted from buccal samples. The DNA concentration readings for DNA extracted from buccal samples was lower overall when compared to DNA extracted from whole blood. However, DNA profiles obtained from both DNA extraction methods still resulted in good, usable DNA profiles despite the differences in the DNA quality. No PCR amplification was detected in the negative controls, indicating absence of contamination. Figure 2.1 shows an example of an electropherogram depicting the DNA profile of the DNA XY5 positive control provided in the Qiagen® Investigator DIPplex kit generated using 10 µL with 40 % of the manufacturer's conditions.

2.3.2. Evaluation of DNA sensitivity

Sensitivity was evaluated by looking at the DNA profile quality characteristics as listed section 2.2.5. PCR amplification using input DNA of 0.1 ng proved to be insufficient, resulting in DNA profiles of poor quality and incomplete profiles with allelic dropout (Figure 2.2). Input DNA of 0.2 ng to 0.3 ng resulted in DNA profiles of better quality consisting of amplification of all 30 loci (full DNA profiles) and good peak morphology (Figure 2.3). DNA quality and concentration of input DNA had a big effect on the DNA profiles obtained. A high concentration of input DNA (0.5 ng and higher) resulted in off scale peaks and in some cases, no profiles being visible due to the peaks being above the detection level of the 3500 Genetic Analyzer (Life Technologies) (see section 2.2.8) as well as massive peak imbalance. The fine balance of the system was demonstrated as too much input DNA affected capillary electrophoresis resulting in off scale peaks and pull up peaks (Figure 2.4). Samples extracted from whole blood delivered overall better quality DNA profiles than samples extracted from buccal samples. Heterozygote peak imbalance was observed across all three concentrations tested, with allelic dropout observed in samples using input DNA of 0.1 ng and allelic drop in observed in samples using input DNA of 0.5 ng and higher. In samples using input DNA in the range of 0.2 - 0.3 ng heterozygote peak imbalances were also observed however it was not high enough to impact allele calling and interpretation.

2.3.3. Evaluation of the robustness and reproducibility

The two samples and the positive control were successfully amplified using the three PCR amplification protocols. All 30 loci were amplified using the three PCR protocols (not shown). No differences in the results obtained from the two different thermal cyclers were observed. The DNA profile generated using the three protocols were identical showing that the system is robust and results were reproducible.

2.3.4. Sample data

No PCR amplification was detected in the negative controls, indicating absence of contamination.

2.4. Summary

The PCR system was successfully evaluated for a template DNA within the range of 0.1 ng to 0.5 ng, depending on the DNA quality of the input DNA. The profiles produced with DNA within the range of 0.2 ng to 0.3 ng were determined to be reproducible and robust. Full DNA profiles were obtained with amplification at all 30 indel loci, for both buccal and blood samples.

The effect of DNA extracted using different DNA extraction methods demonstrated the effect DNA quality has on the DNA profiles but that the system was sensitive enough to generate DNA profiles of usable quality by adjusting the concentration of input DNA. Samples extracted from whole blood delivered samples of better DNA quality than samples extracted from buccal samples with higher purity and higher yield of DNA concentration. Samples extracted from whole blood resulted in sample profiles of higher overall quality than samples extracted from buccal samples, using the PCR protocol of 10 µL final volume with 40 % manufacturer's recommendations with input DNA in the range of 0.2 ng - 0.3 ng.

For samples extracted using both extraction methods, the DNA input range of 0.2 ng - 0.3 ng proved to be sufficient for generating profiles of good quality. DNA input concentrations outside the range of 0.2 ng- 0.3 ng resulted in loss of quality.

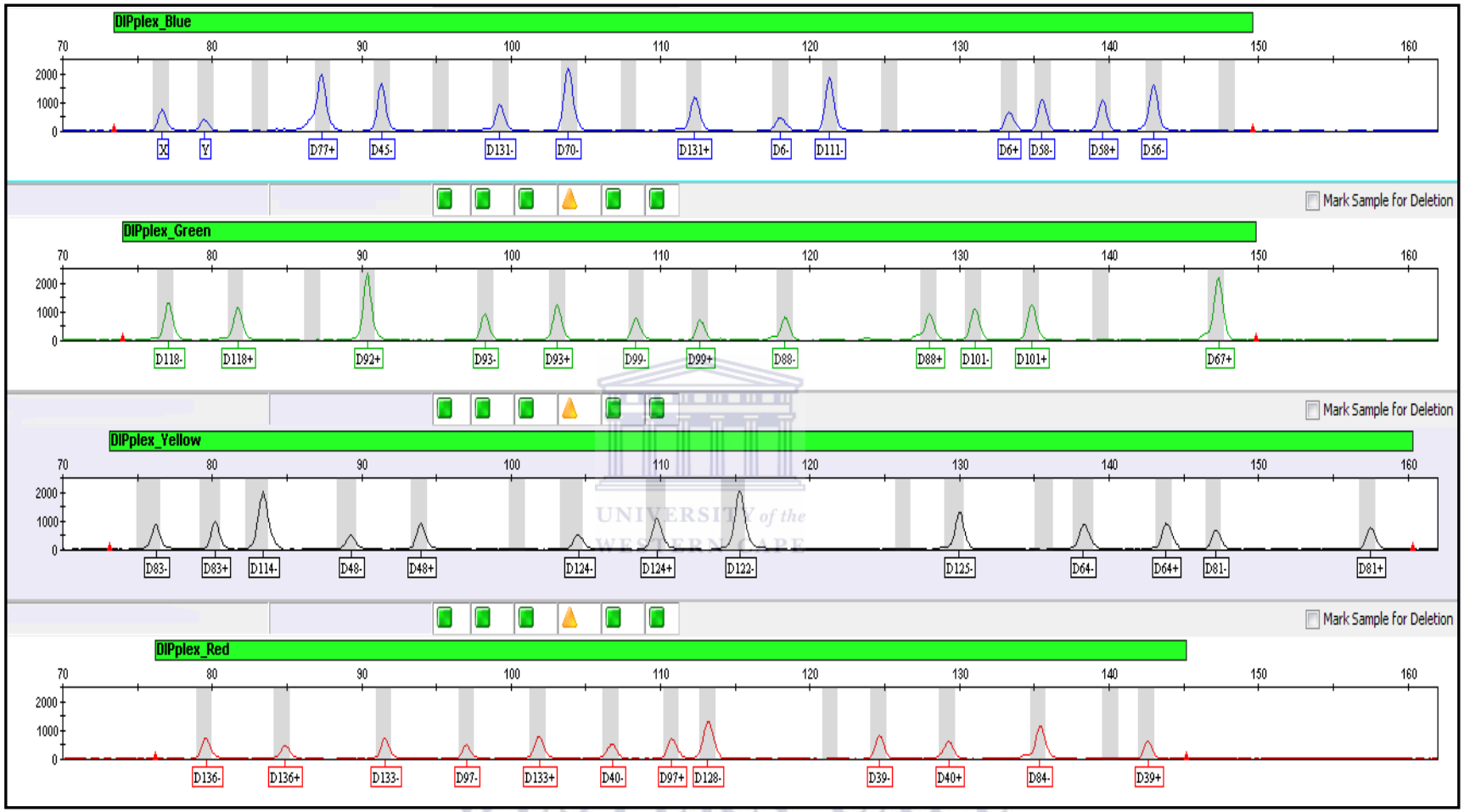


Figure 2.1 An electropherogram of a DNA profile obtained from the positive control DNA XY5 using 10 µL with 40 % of the manufacturer's conditions.

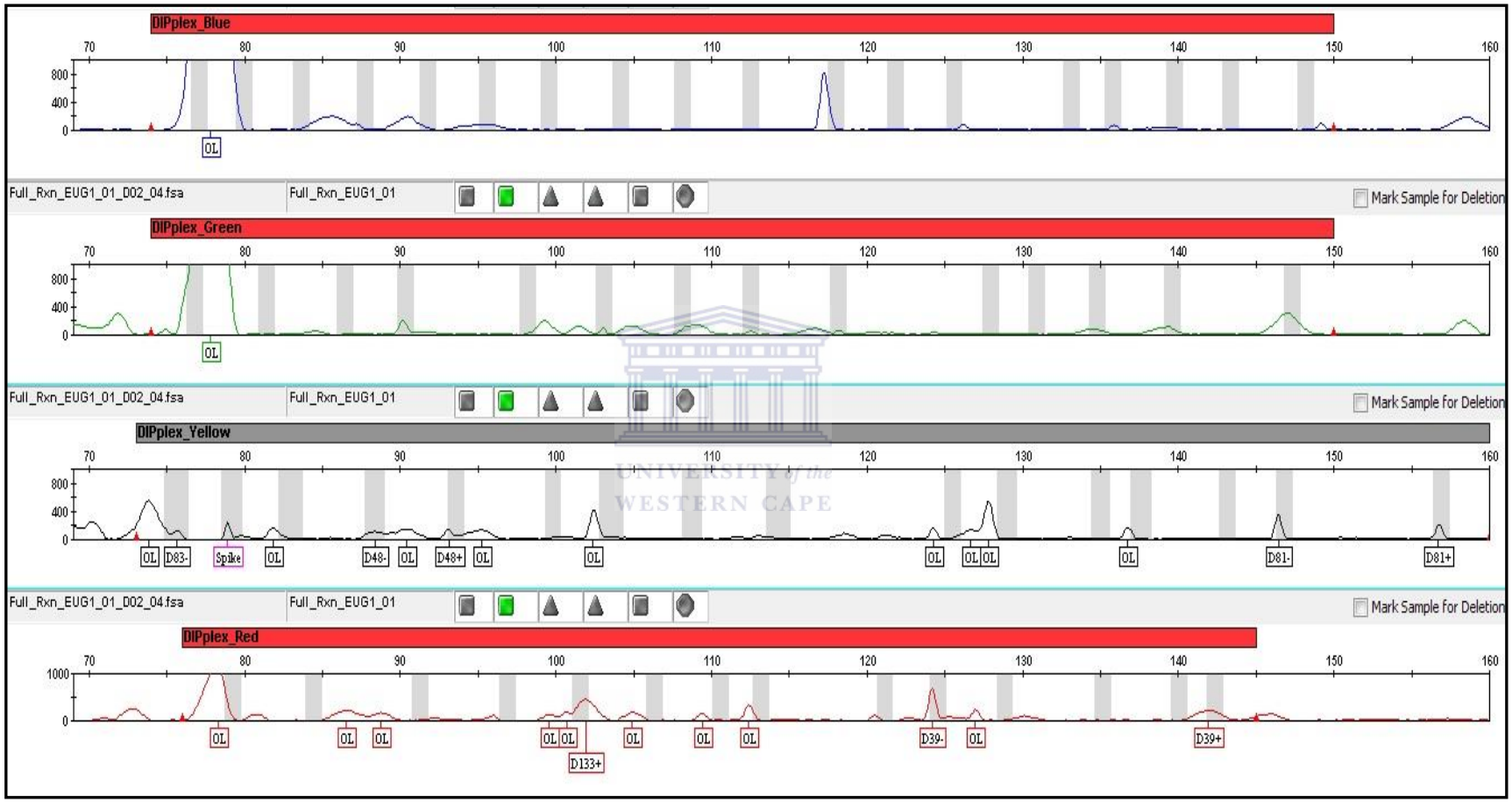


Figure 2.2 An electropherogram of a DNA profile obtained with input DNA of 0.1 ng. OL = Off Ladder

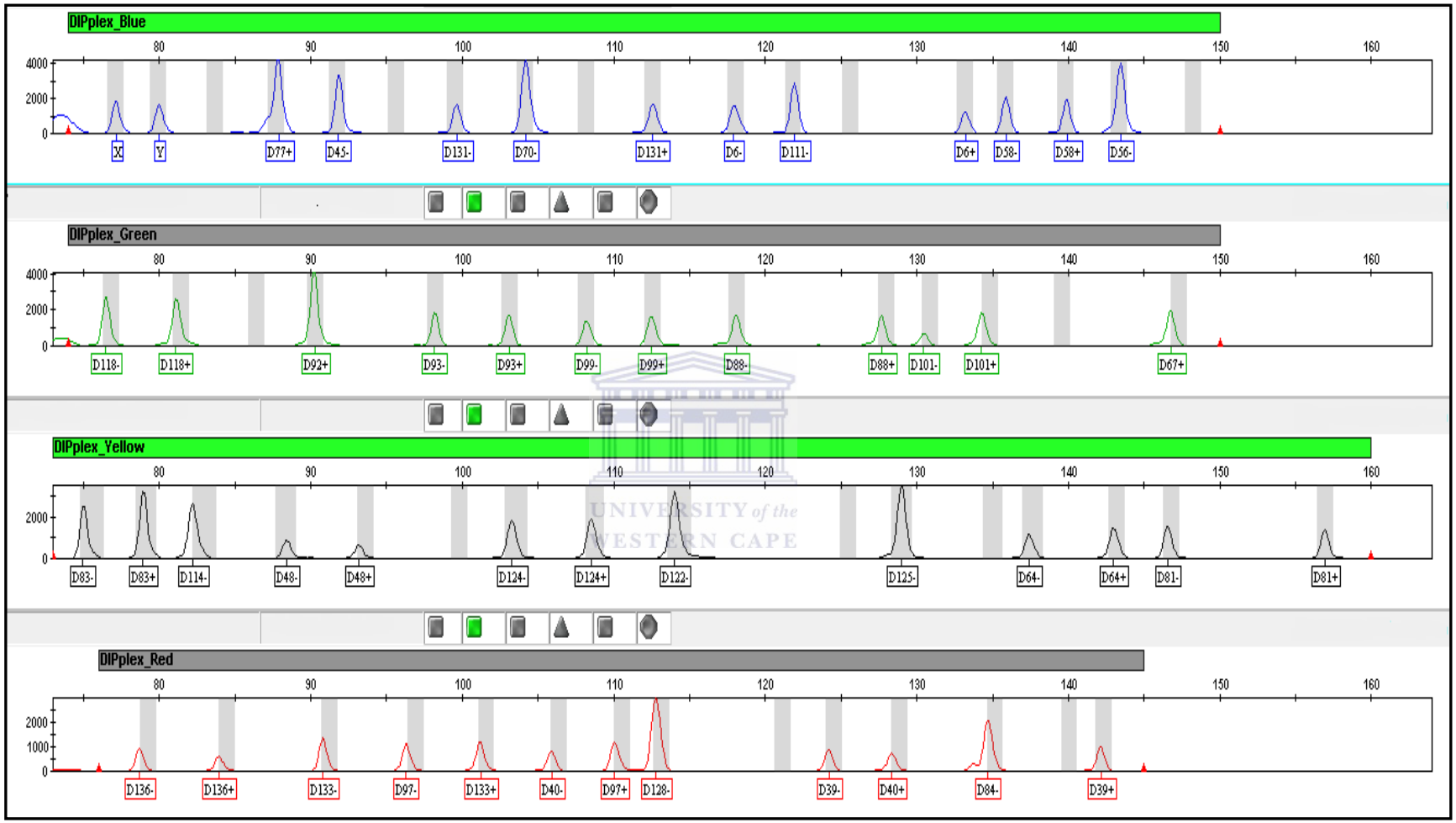


Figure 2.3 An electropherogram of a DNA profile obtained from with input DNA of 0.2 ng in a final volume of 10 µL.

.

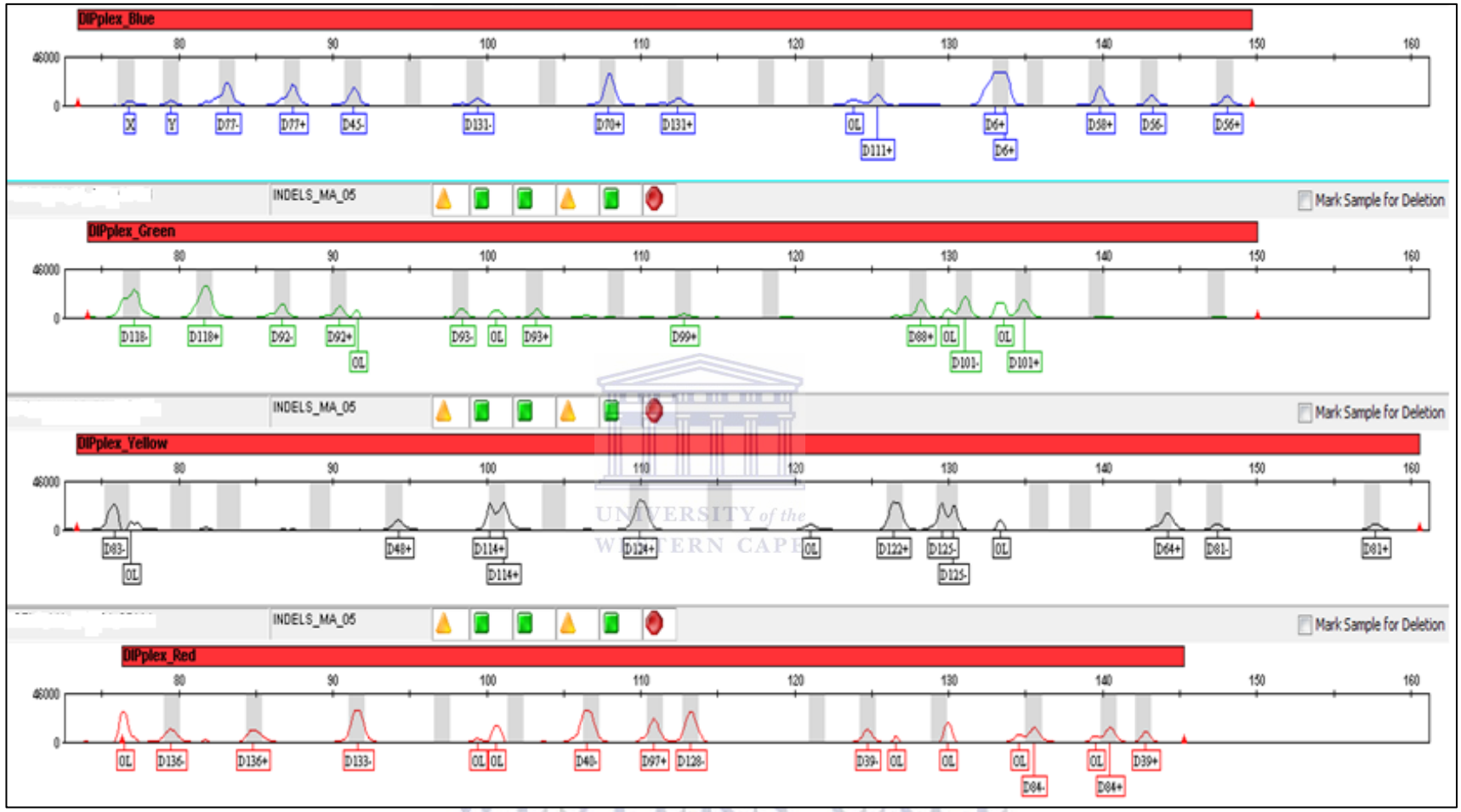


Figure 2.4 An electropherogram of a DNA profile obtained from with input DNA of > 0.5 ng. OL = Off Ladder

All three PCR protocols evaluated proved to be successful in delivering usable results for DNA input concentration of 0.2 ng. This concentration was within the manufacturer recommended concentration range of 0.1 - 0.5 ng.

Despite the the Qiagen Investigator DIPplex kit already being optimised, I would recommend further evaluation and adaptation of the system to suit specific research requirements which would include consideration of DNA quality as a result of different DNA extraction methods. Due to the high sensitivity of the system, careful attention needs to be given to contamination preventative measures. These would include the basic decontamination method, good laboratory practices and restricting communal use of laboratory instruments like pipettes (see section 2.2.1).

Based on the above evaluations, the conditions for generating population data using samples from both extraction methods were as follows: DNA input concentration range of 0.2 ng to 0.3 ng using the PCR protocol consisting of a final volume of 10 µL (using 40 % manufacturer's recommendations), shown in the next chapter.

Chapter 3

Investigation of the presence of a null allele at HLD97 locus

3.1. Introduction

The presence of null alleles has been reported in humans and in other species (Paetkau and Strobeck, 1995; Pemberton *et al*, 1995; Neumann and Wetton, 1996). A null allele refers to an allele that repeatedly fails to PCR amplify at a locus due to unsuccessful primer binding (Dakin and Avise, 2004). During this study lack of PCR amplification was observed at HLD97 locus in six samples from the Xhosa population and eight samples from the Zulu population. For six samples from the Xhosa population and eight samples from the Zulu population successful amplification was achieved at all loci except the HLD97 locus; neither the HLD97 insertion nor deletion allele was amplified. The DNA result obtained for these 14 samples from the Bantu populations at all the other loci were of good quality with peak heights well over 100 rfu and good peak morphology observed. Therefore the possibility of this result being obtained by chance, allelic dropout or poor DNA quality was discarded. It is hypothesized that the possible presence of a null allele is responsible for the lack of PCR amplification at the HLD97 locus.

To investigate the possible presence of a null allele at the HLD97 locus, the following strategies were applied. One cause of null alleles is unsuccessful primer binding which is due to variations in the nucleotide sequence within the primer binding site. Therefore two tests were performed to investigate this. Firstly, the binding properties of the manufacturer's primers for the HLD97 locus were investigated by performing PCR amplification using a lower annealing temperature, which decreases the primer binding stringency. The second test was to perform DNA sequencing and sequence alignments on selected samples and comparison to the Genbank sequence (Benson *et al*, 2013) to establish the presence of any nucleotide differences that could lead to the primer binding failure and consequently, unsuccessful PCR amplification. The selected samples were (1) samples for which unsuccessful PCR amplification was observed at

HLD97, and (2) samples representing the other genotypes observed, i.e. homozygote for the deletion allele, homozygote for the insertion allele, heterozygote for the deletion and insertion alleles and homozygote for the null allele, for comparison purposes. Lastly, statistical analysis was performed to determine null allele frequencies and to establish possible consequences of the presence of the HLD97 null allele on the populations.

3.2. Materials and methods

3.2.1. Samples selection

Generation of population profiles for the five populations was performed as previously described (section 2.2.4.1.). Successful amplification was observed at all loci except the HLD97 locus with neither the HLD97 insertion nor deletion allele being amplified in 14 Bantu (N=6 Xhosa) and N=8 Zulu) samples.

One sample from each Bantu population (N= 1 Xhosa and N= 1 Zulu) was chosen to represent each of the four genotypes namely homozygote for the deletion allele, homozygote for the insertion allele, heterozygote for the deletion and insertion alleles and 4X homozygotes for the null allele (N=2 Xhosa and N=2 Zulu), based on their observed phenotypes through fragment analysis. An insertion homozygote individual carries the insertion allele, a 14 bp sequence (5'- AGAGAAAGCTGAAG-3') in both chromosomes. A heterozygote individual carries both the insertion and deletion allele. The positive control as provided in the kit is a heterozygote for the deletion and insertion alleles and was also included in sequencing analysis performed. For statistical analysis of the HLD97 null allele, samples from all five populations (Afrikaner, Mixed Ancestry, Indian Asian, Xhosa and Zulu) were used. No samples containing the HLD97 null allele were excluded.

3.2.2. Investigating the primer binding properties of the Investigator DIPplex HLD97 primers (Qiagen)

To investigate the primer binding stringency of the manufacturer's Investigator DIPplex primers (Qiagen), N= 4 (see above) null allele homozygote samples were PCR amplified at a lower annealing temperature. PCR amplification was performed as previously discussed (section 2.2.8). Thermocycling conditions as recommended by the manufacturers were followed, with the annealing temperature of the primers adjusted to 5 ^{o}C lower than the manufacturers' recommended annealing temperature of 61 ^{o}C.

3.2.3. Primer design and PCR amplification

The accession number rs17238892 was used to obtain the reference sequence from Genbank (Figure 3.1). A primer pair consisting of a forward and reverse primer was designed for the purpose of PCR amplification of a DNA sequence containing the HLD97 indel site (Table 3.1) using the primer design software Oligo 1.4 Primer Analysis Software (Rychlik, 2007) and the Genbank reference sequence. Primers were designed taking into consideration the percentage guanine and cytosine, primer length, melting temperature (Tm) and annealing temperature (Ta).The primers were designed for PCR amplification of a 527 bp amplicon with the forward primer binding site being 349 bp from the indel site and the reverse primer being 123 bp from the indel site. Primers were supplied by Integrated DNA Technologies (IDT), Whitehead Scientific. The optimal primer annealing temperature was determined empirically by gradient PCR. Two reference samples, a negative control and the positive control were PCR amplified for this purpose.

Table 3.1 PCR primer set and sequences used for amplification of a selected region of the HLD97 indel.

Primer	Sequence	Product Size (bp)	%GC	Tm (^{o}C)	Ta (^{o}C)
HLD97FWD	5'- AAGGCAAATCGTGATTGTGAC-3'	527	43	60	62
HLD97REV	5'- ACACACCAGCAATGAGTGTCC-3'		52	64	62

bp, base pairs; %GC, percentage Guanine and Cytosine; Tm, melting temperature; Ta, annealing temperature; ^{o}C, degrees Celsius.

3.2.4. PCR amplification

3.2.4.1. PCR amplification for primer optimisation

PCR amplification was performed in a final volume of 20 µL consisting of 2 ng

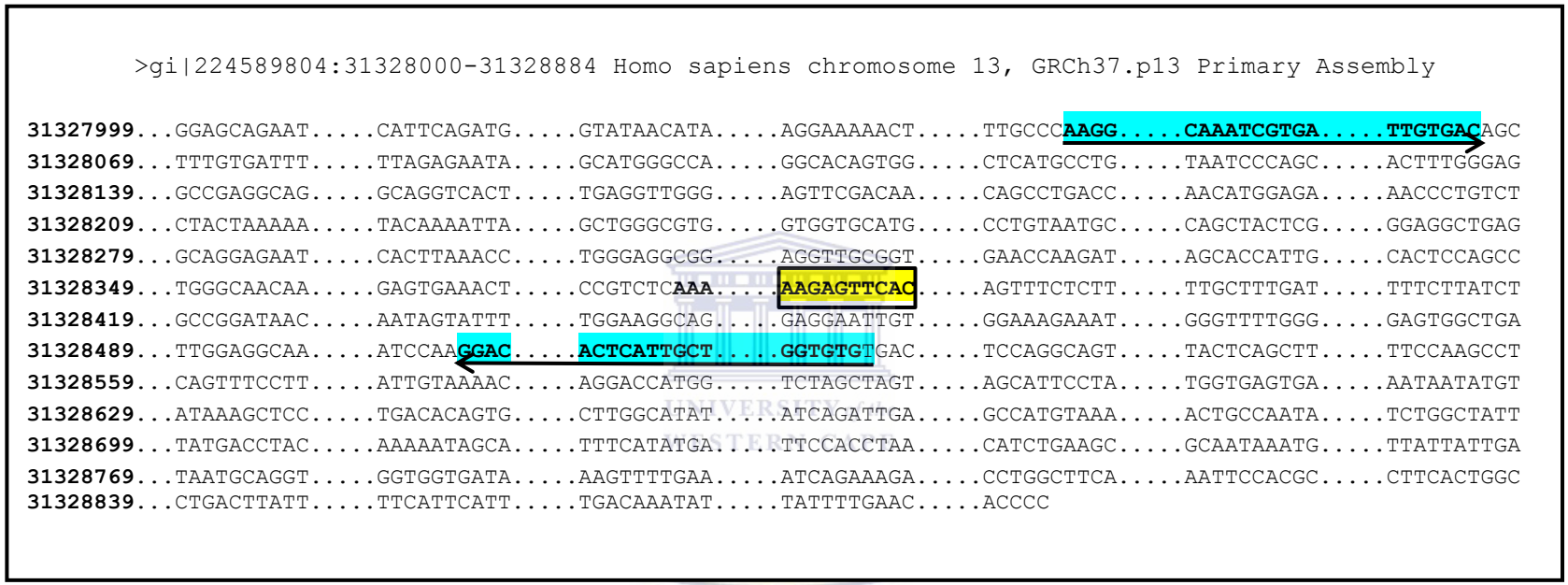


Figure 3.1.The reference sequence obtained from Genbank (rs17238892) with the nucleotide position (in bp) indicated in bold. The HLD97 indel site is located between 31328384 and 31328385. The region containing the HLD97 indel site is highlighted in yellow. The forward and reverse primer position is highlighted in blue.

DNA, 2 µL S-T GOLD buffer (1X), 2 µL dNTPs [0.2 mM], 2 µL Ultrapure BSA [0.4 mg/µL] (Ambion®), 0.2 µL Super Therm GOLD Taq polymerase [0.05 U], 2 µL forward primer [0.3 µM], 2 µL reverse primer [0.3 µM] and 7.8 µL SABAX water. Thermo cycling was conducted in an Arktik™ Gradient Enabled 96-well Thermal

Cycler (Thermo Scientific). The thermo cycling conditions were 94 °C for 10 minutes, 35 cycles at 94 °C for 40 seconds, Tm range from 58 °C to 66 °C for 45seconds and 72 °C for 45 seconds, 60 °C for 45 minutes and a holding step at 10 °C.

3.2.4.2. PCR amplification of a DNA fragment containing the HLD97 indel site

PCR amplification was performed in a final volume of 30 µL consisting of 2 ng DNA, 3 µL S-T GOLD buffer (1X), 3 µL Ultrapure BSA [0.4 mg/µL] (Ambion®), 3 µL forward primer [0.3 µM], 3 µL reverse primer [0.3 µM], 0.3 Super Therm GOLD Taq polymerase [0.05 U], 3 µL of dNTP mix [0.2 mM] and 12.7 µL SABAX water. Thermo cycling was conducted in a Veriti 96-well Thermal Cycler (Life Technologies) and Arktik™ Gradient Enabled 96-well Thermal Cycler (Thermo Scientific). Thermo cycling conditions were as follows: 94 °C for 10 minutes, 35 cycles at 94 °C for 40 seconds, 60.1 °C for 5 seconds and 72 °C for 45 seconds, 60 °C for 45 minutes and a holding step at 10 °C. Samples were stored at -4 °C.

3.2.5. Agarose gel electrophoresis

Success of PCR amplification was checked by loading the PCR products on a 1 % agarose gel (see Appendix). A volume of 1 µL PCR product was mixed using a pipette with 4 µL loading dye (Bioline) that has been pre-mixed with GelRed® Nucleic Acid Stain (Biotium) before loading onto the agarose gel. A volume of 1 µL of the negative control, positive control and Hyper Ladder V (Bioline) each

was loaded and run in addition to the PCR amplicon samples. Gel electrophoresis was performed on a GEL XL Ultra V-2 (Labnet Intl, Inc.) at 25 V for 25 minutes. Following gel electrophoresis an image was captured using an Alphamager® HP Gel Imager.

3.2.6. Capillary Electrophoresis

Capillary electrophoresis was performed as previously described (section 2.2.7).

3.2.7. Data Collection

Data collection was performed as previously described (section 2.2.7.2).

3.2.8. DNA Sequence Analysis

3.2.8.1. DNA sequencing

DNA sequencing is a method of determining the nucleotide order of a DNA molecule. Using dideoxynucleoside triphosphates (ddNTPs) and a denatured DNA strand as a template, new DNA strands are synthesised through strand elongation. Sanger DNA sequencing is a method of DNA sequencing that was developed by Frederick Sanger in 1975. It incorporates chain-terminating ddNTPs by DNA polymerase.

The DNA nucleotide sequence information of the selected samples was determined by Sanger DNA sequencing to confirm the genotypes of the heterozygote-, deletion homozygote- and insertion homozygote samples and to investigate the possible presence of nucleotide variations occurring in the primer binding sites of the null allele homozygote samples.

The forward and reverse primers used for PCR amplification were also used for Sanger DNA sequencing. The PCR amplified DNA samples and primers were

submitted to Macrogen Inc. (Netherland) for PCR purification and Sanger DNA sequencing.

3.2.8.2. Sequencing data analysis

DNA sequencing results were obtained from Macrogen Inc. (Netherland). The chromatograms were visualised using BioEdit software (Hall, 1999), Chromas Lite software 2.1 (http://www.technelysium.com.au) and FinchTV software (www.geospiza.com/finchtv).

3.2.8.3. Sequence alignments

Consensus sequences were constructed using the forward and reverse sequence reads using BioEdit software (Hall, 1999). Multiple consensus sequence alignments were constructed using BioEdit software (Hall, 1999) along with the reference sequence obtained from Genbank using the accession number/SNP ID (rs17238892) (Qiagen, 2011). Sequence alignments were also performed with the HLD97 null allele sequences and the Genbank reference sequence. Sequence alignments were not performed with heterozygous individuals due to the presence of both alleles which results in mixed traces, making base pair assignments difficult.

3.2.9. Statistical data analysis of the HLD97 null allele

3.2.9.1. Estimation of HLD97 null allele frequencies

Null allele frequencies were calculated using five methods:

i) the Hardy-Weinberg equation

$$\mathbf{1 = p^2 + 2pq + q^2}$$

where p and q represent the allele frequencies of the two alleles at the a locus with q^2 representing the D97 null allele. By taking the frequency of the null allele

homozygote individuals in a population and calculating the square root of that frequency, an estimate of the null allele frequency is calculated.

ii) a null allele frequency estimator Chakraborty *et al* (1992):

$$\mathbf{r = (He - Ho) / He + Ho)}$$

Chakraborty *et al*'s null allele estimator is based on the assumption that any missing data is due to failed amplification as a result of degraded DNA, etc. or due to the presence of a homozygous null allele.

iii) null allele estimator 1 by Brookfield (1996):

$$\mathbf{r = (He - Ho) / (1 + He)}$$

where r is null allele frequency, He is expected heterozygosity calculated as the sum of the product of all observed allele frequencies ($\sum pi\, pj$, with i≠j) and Ho is observed heterozygosity calculated as n2 / (n1+n2) with n1 being the number of one-banded individuals and n2 the number of two-banded individuals. This null allele estimator is applied under the assumption that no null allele homozygotes are present and ignores all non-amplified samples as degraded DNA or human error.

iv) null allele estimator 2 by Brookfield (1996):

$$\mathbf{r = A + \sqrt{(A2 + B)/ 2(1 + He)}}$$

in which A = He(1 + N) - Ho, and B = 4N (1 - He^2). Here, the observed heterozygosity (Ho) is measured from the data as n2 / (n1 + n2), where n1 is the number of one-banded individuals and n2 is the number of two-banded individuals in the sample. N represents the proportion of samples not amplified due to a null allele presence. The expected heterozygosity (He) is calculated as the

sum of the product of all observed allele frequencies ($\sum pi\,pj$, with i≠j). This null allele estimator does acknowledge the presence of null allele homozygotes.

v) ML-NullFreq software program (Kalinowski and Taper, 2006). This program uses a maximum likelihood estimator that is based on actual genotype counts and not allele frequencies. It is not based on the presence or absence of data:

$$\text{RMSE}= Avg\sqrt{\left(\frac{\sum_{i=1}^{k}(p_i-\hat{p}_i)^2}{k}\right)}$$

where the root mean squared error (RMSE) is calculated from the difference between the actual null allele frequency (p_i), the estimated frequency ($\widehat{Pi}$) and the visible alleles where *k* is the number of visible alleles at a locus.

3.3. Results

3.3.1. Basis of investigation for the presence of a null allele at the HLD97 locus

As previously mentioned (section 3.1), observations of successful PCR amplification at all loci except the HLD97 locus was noted in 14 samples with good quality profiles obtained at the remaining loci, as well as good peak morphology and high relative fluorescent units. Based on these observations it was hypothesized that the possible presence of a null allele, was responsible for the failure of PCR amplification at the HLD97 locus.

3.3.2. Investigating the primer binding properties of the Investigator DIPplex HLD97 primers (Qiagen)

Figure 3.2 shows an electropherogram of a sample with PCR amplification at all loci including the HLD97 locus using the manufacturer's recommended annealing temperature of 61 ^{o}C, in a final volume of 10 uL using 40 % of manufacturer's

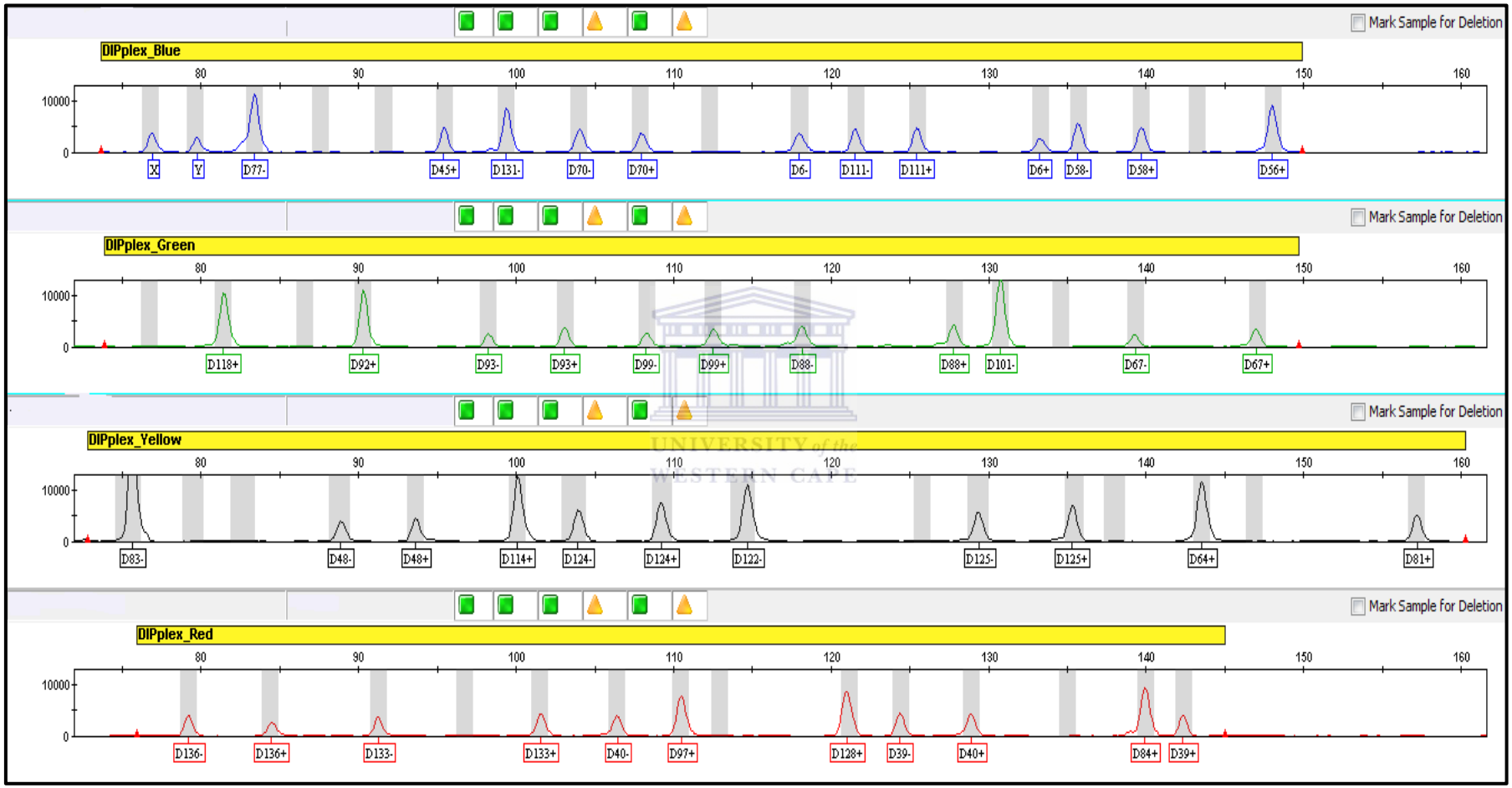


Figure 3.2 An electropherogram of a DNA profile with PCR amplification at all of the 30 loci as well as the HLD97 locus using manufacturer's recommended annealing temperature of 61 ^{o}C using 10 uL.(Qiagen, 2011).

conditions (Qiagen, 2011). Figure 3.3 shows an electropherogram of the red fluorescent dye labelled alleles of a sample containing a HLD97 null allele amplified using the manufacturer's recommended annealing temperature of 61 ^{o}C. The absence of both HLD97 alleles is clearly visible. The annealing temperature as per manufacturer's recommendation was decreased by 5 ^{o}C to lower the stringency of the primer binding to the target sequence. Successful PCR multiplex amplification of both alleles at the HLD97 locus was obtained at a lower annealing temperature of 56 ^{o}C in four samples in which no amplification was obtained for HLD97 using manufacturer's PCR annealing temperature. Figure 3.4 shows an electropherogram of the same sample amplified using an annealing temperature of 56 ^{o}C. The presence of the HLD97 insertion allele is visible. This confirms that the HLD97 primer binding ability is affected. The most likely explanation is the presence of a mutation in the primer binding site.

3.3.3. Primer optimisation and PCR amplification

The optimal primer annealing temperature for the designed primer pair as determined using the Oligo 1.4 primer analysis software (Rychlik, 2007) was 62 ^{o}C. Successful PCR amplification was observed for the temperature range 58 ^{o}C to 65.8 ^{o}C (Figure 3.5). The best PCR amplification was observed for annealing temperatures ranging between 60.1 ^{o}C and 64.6 ^{o}C as indicated by the strong intensity of the PCR bands. The amplified PCR product size was within the expected size range (450-500 bp) as indicated by the ladder used. No contamination was present as indicated by the absence of amplification in the negative control (not shown).

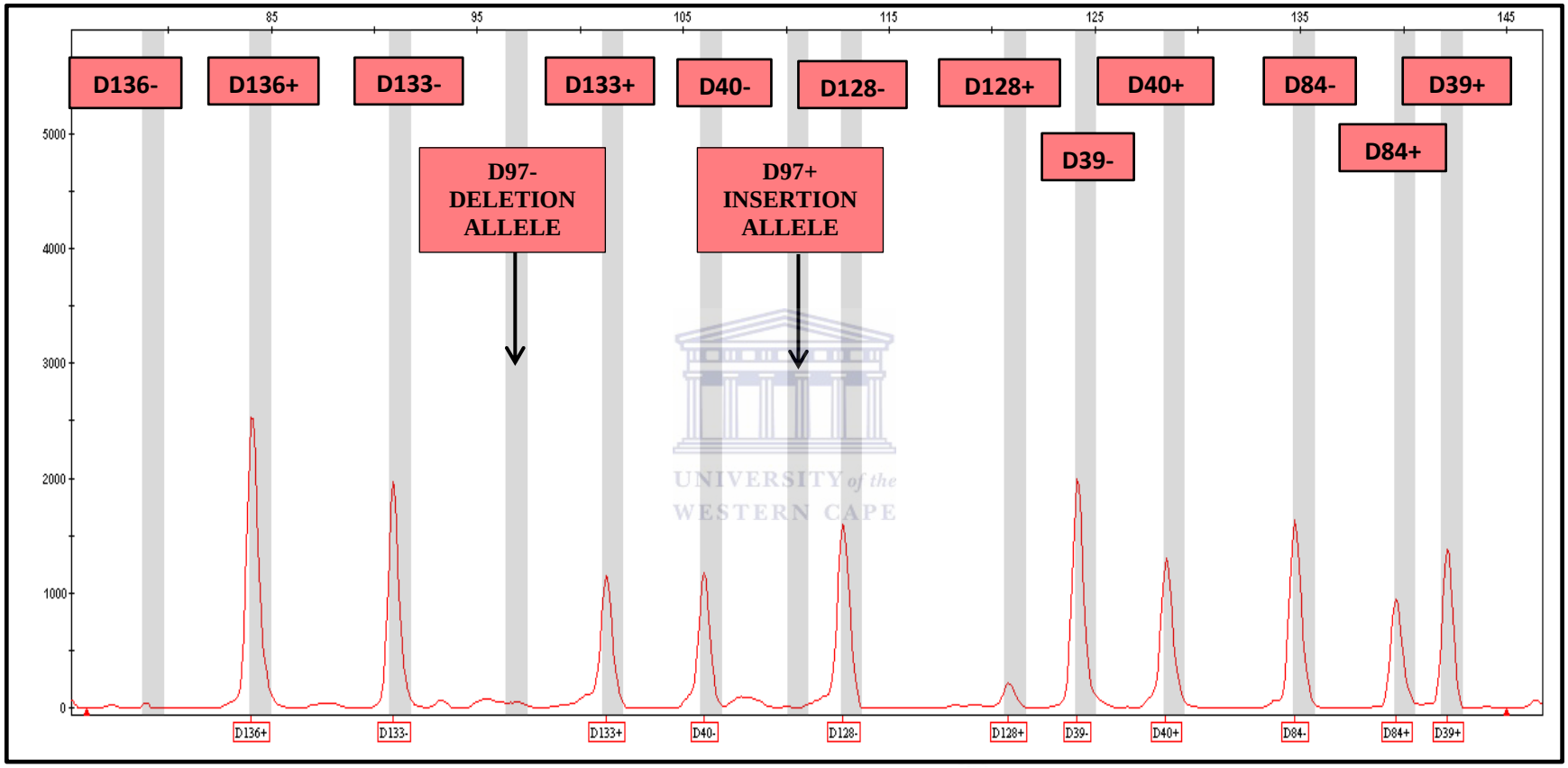


Figure 3.3 An electropherogram showing the red fluorescent dye labels of a HLD97 null allele amplified with Tm of 61 ^{o}C. Black arrows indicate the absence of the HLD97 insertion and deletion alleles. The loci names are indicated. (+) indicates insertion allele; (-) indicates deletion allele.

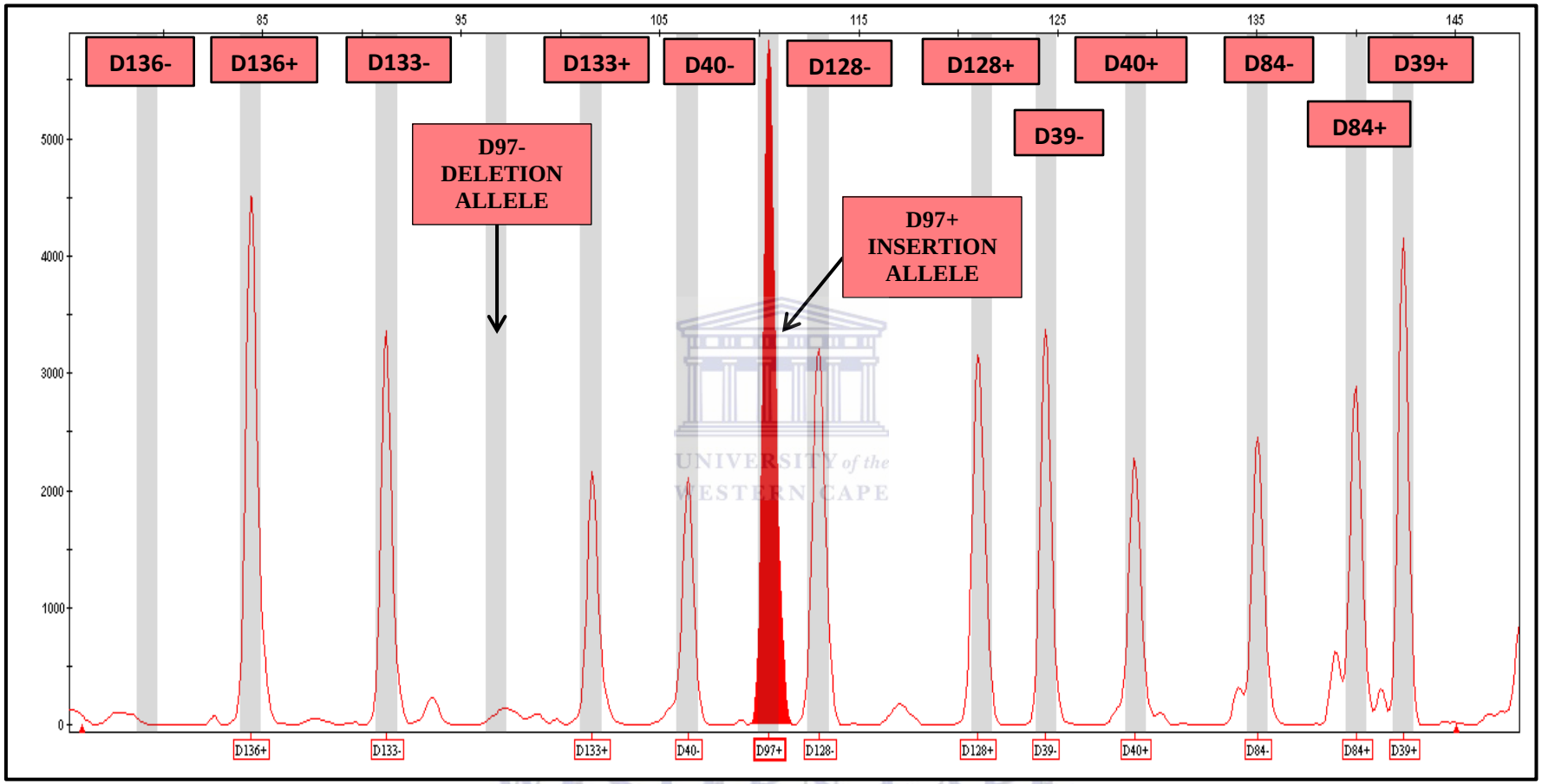


Figure 3.4 An electropherogram showing the red fluorescent dye labels of a HLD97 null allele amplified with Tm of 56 ^{o}C. The presence of the HLD97 insertion allele is highlighted in red. The loci names are indicated. (+) indicates insertion allele; (-) indicates deletion allele

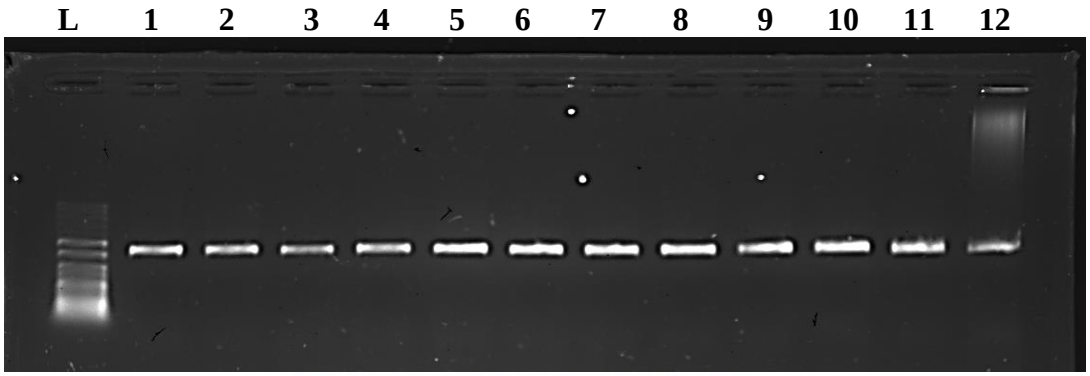


Figure 3.5 Gradiënt PCR of sample 2

L = Hyper Ladder V; Lane 1, 58 ^{o}C; Lane 2, 58.2 ^{o}C; Lane 3, 58.6 ^{o}C; Lane 4, 59.2 ^{o}C; Lane 5, 60.1 ^{o}C; Lane 6, 61.2 ^{o}C; Lane 7, 62.6 ^{o}C; Lane 8, 63.8 ^{o}C; Lane 9, 64.6 ^{o}C; Lane 10, 65.4 ^{o}C; Lane 11, 65.8 ^{o}C; Lane 12, 66 ^{o}C

3.3.4. DNA sequencing analysis

3.3.4.1. DNA sequencing

As mentioned in section 3.2.8.1, Sanger DNA sequencing was performed to determine the nucleotide sequences of the selected samples' genotypes and to investigate the possible presence of nucleotide variations. DNA sequencing confirmed the genotypes of the insertion homozygote- (Figure 3.6), deletion homozygote-(Figure 3.7) and heterozygote- individuals (Figure 3.8). The positive control as provided by the kit is a heterozygote for the HLD97 locus and was also sequenced. The DNA sequence of the heterozygote positive control reflects this by the presence of a mixed DNA sequence trace starting at the indel insertion site which is caused by the presence of both alleles (Figure 3.8).

3.3.4.1.1. HLD97 null allele sequencing results

Figure 3.9 shows the chromatogram of a HLD97 homozygote null allele individual sequenced using the forward primer. The 14 bp insertion sequence is present. Sequencing of the homozygote HLD97 null allele samples therefore confirmed that the true genotype of the HLD97 null allele homozygote samples is that of an insertion homozygote.

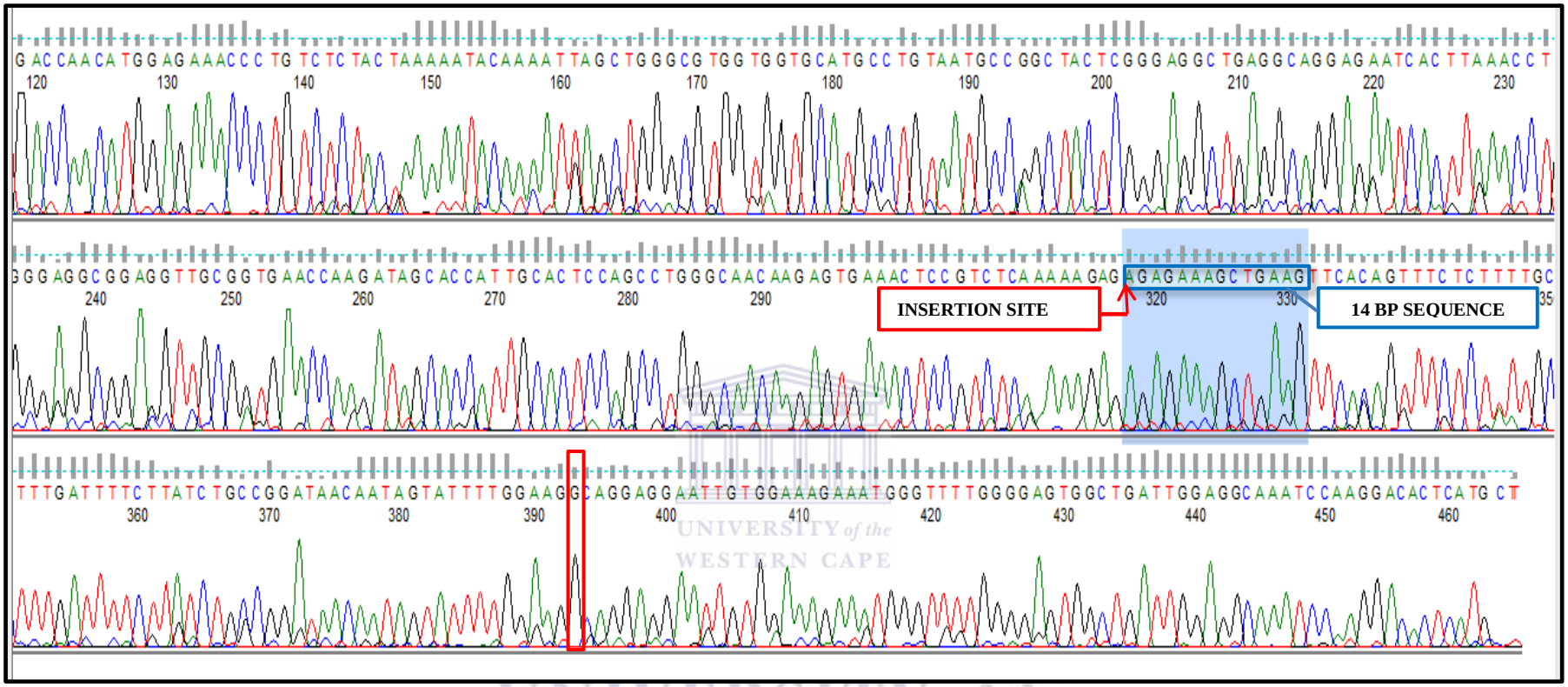


Figure 3.6. A chromatogram of a HLD97 insertion allele homozygote individual. The insertion site and insertion sequence (shaded) are indicated. The wild type G-nucleotide is indicated in red.

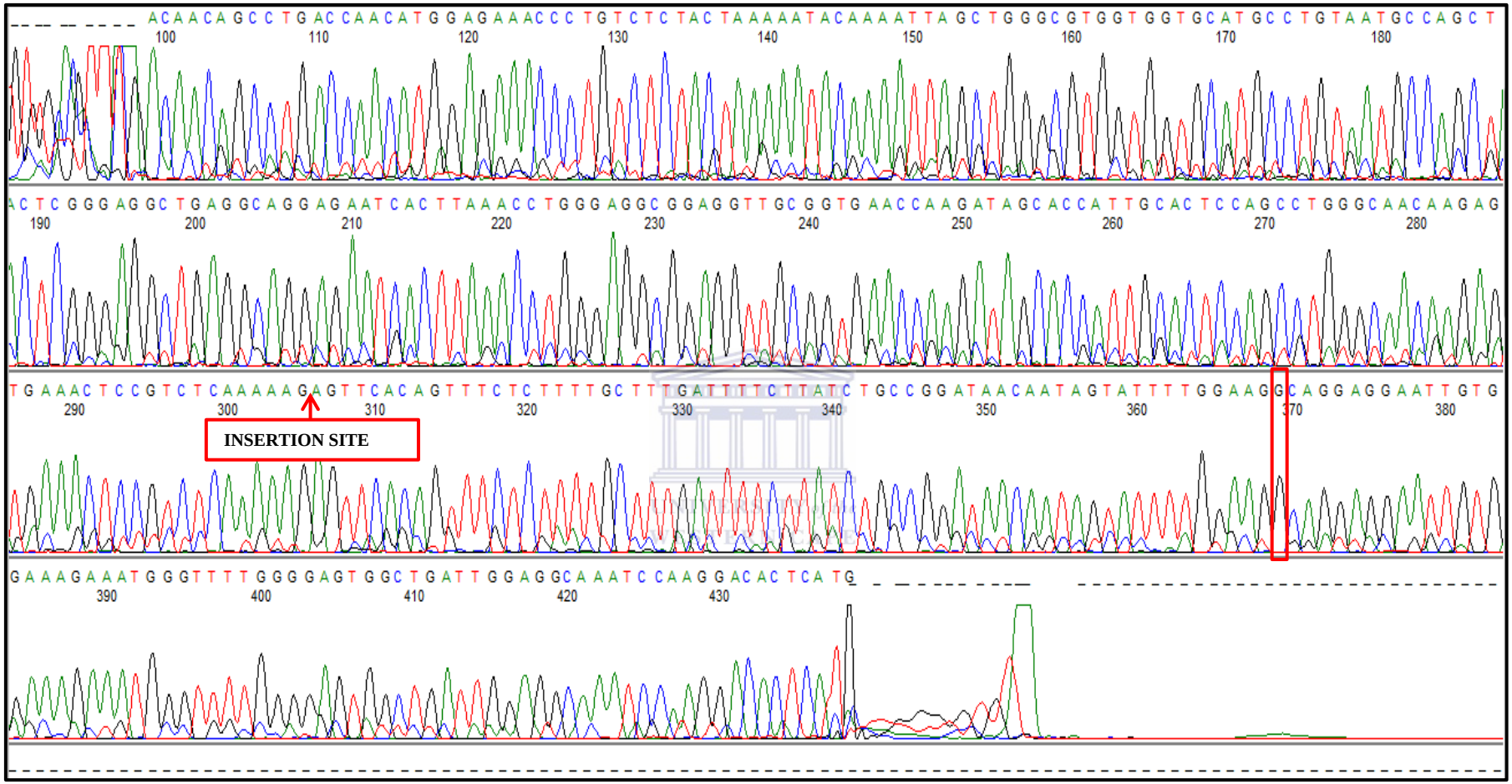


Figure 3.7 A chromatogram of an HLD97 deletion homozygote individual. The insertion site is indicated. The wild type G-nucleotide is indicated in red.

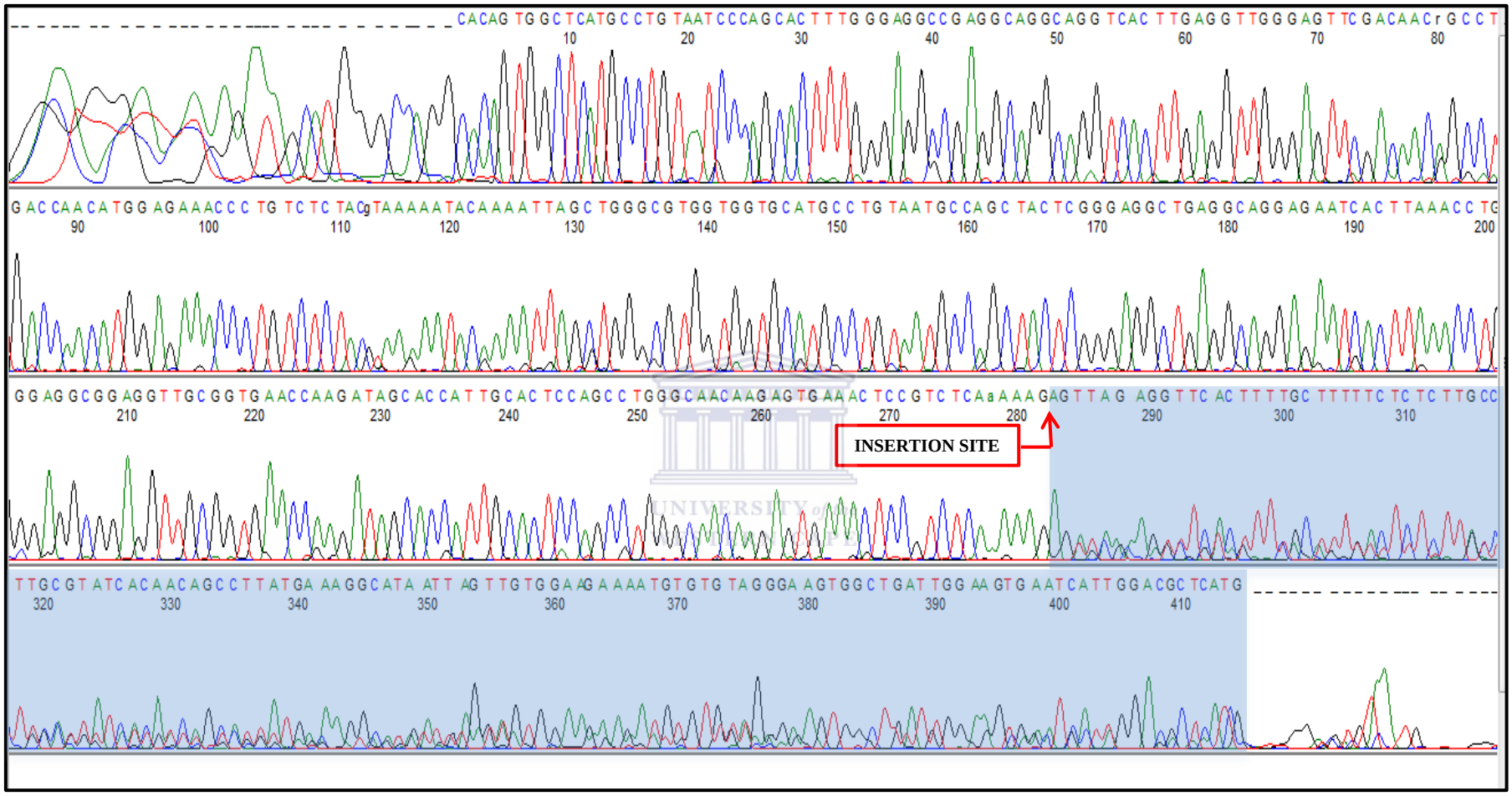


Figure 3.8 A chromatogram of an HLD97 heterozygote individual. The visible mixed trace and decrease in trace quality from the insertion site (indicated) onwards is visible and shaded. It is caused by the presence of both the insertion and deletion alleles.

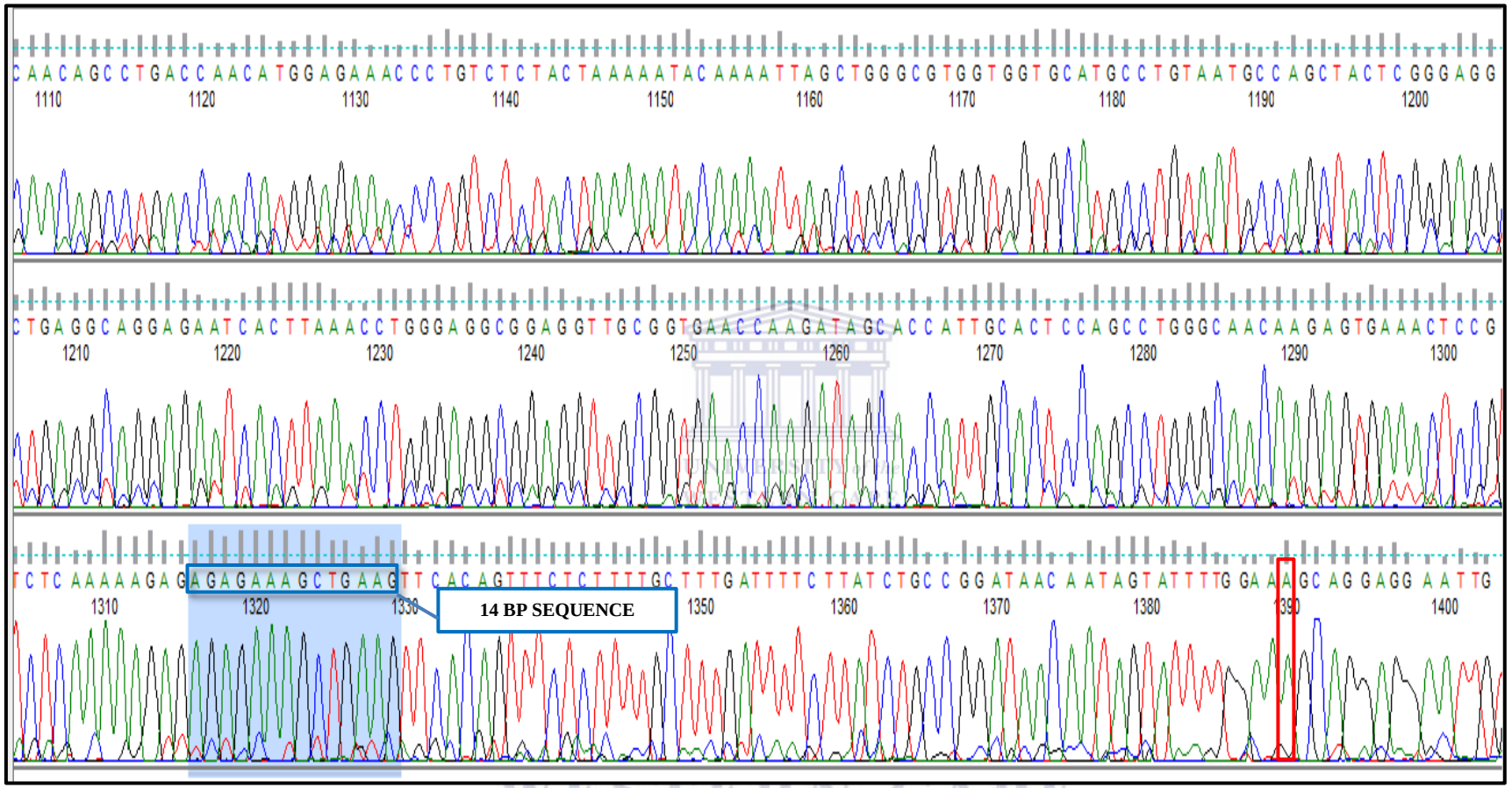


Figure 3.9 A sequence chromatogram using the reverse primer of an HLD97 null allele individual. The presence of the 14 bp insertion sequence is indicated and shaded, confirming the true genotype to be insertion homozygote. The mutant A nucleotide is indicated in red.

3.3.4.2. Sequence alignments

Consensus sequences were constructed as mentioned in section 3.2.8.3. Comparison between the consensus sequences and the Genbank reference sequence is shown in Figure 3.10.

3.3.4.2.1. HLD97 null allele sequence alignments

The sequence alignments constructed using the HLD97 null allele sequences is shown in Figure 3.11. It confirms the actual genotype of HLD97 null allele homozygote samples as being insertion homozygote as stated in section 3.3.4.1.1.

In all of the sequenced HLD97 null allele homozygote samples, the sequencing results revealed the presence of a G to A substitution located 75 bp downstream from the HLD97 indel site (Figure 3.11 and Figure 3.12). This mutation would explain the lack of PCR amplification at the HLD97 locus, as it is likely located within the primer binding site. This is also supported by successful PCR amplification obtained at a lower annealing temperature that caused a decrease One individual genotyped as a HLD97 deletion homozygote revealed a heterozygote genotype after sequencing. This result is explained in the next section 3.3.4.2.1 (Figure 3.13).

3.3.5. Statistical and analytical estimation of null allele frequencies

The null allele estimators by Chakraborty *et al* and Brookfield are based on the assumption of homozygote excess due to null allele presence in the population but are not reliable due to these estimators not accounting for the presence of missing data. These estimators are calculated based on allele frequencies (Brookfield estimators) and observed and expected heterozygosity values (Chakraborty *et al*'s estimator).

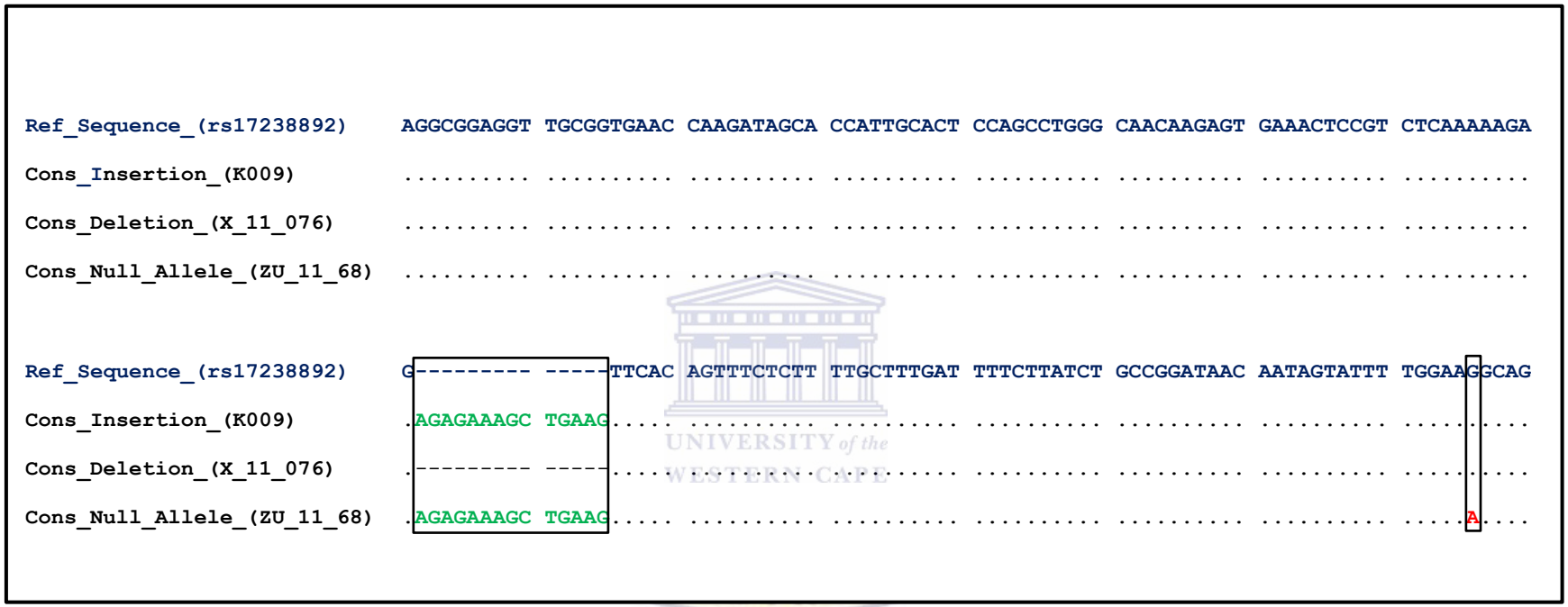


Figure 3.10 The sequence alignments performed with consensus sequences representing homozygote individuals for the insertion (K009), deletion (X_11_76) and null allele (ZU_11_68). The 14 bp insertion sequence is indicated in green. The wild type G nucleotide present in the GenBank reference sequence is substituted with an A nucleotide in the null allele consensus sequence (ZU_11_68) as indicated in red. Cons=Consensus

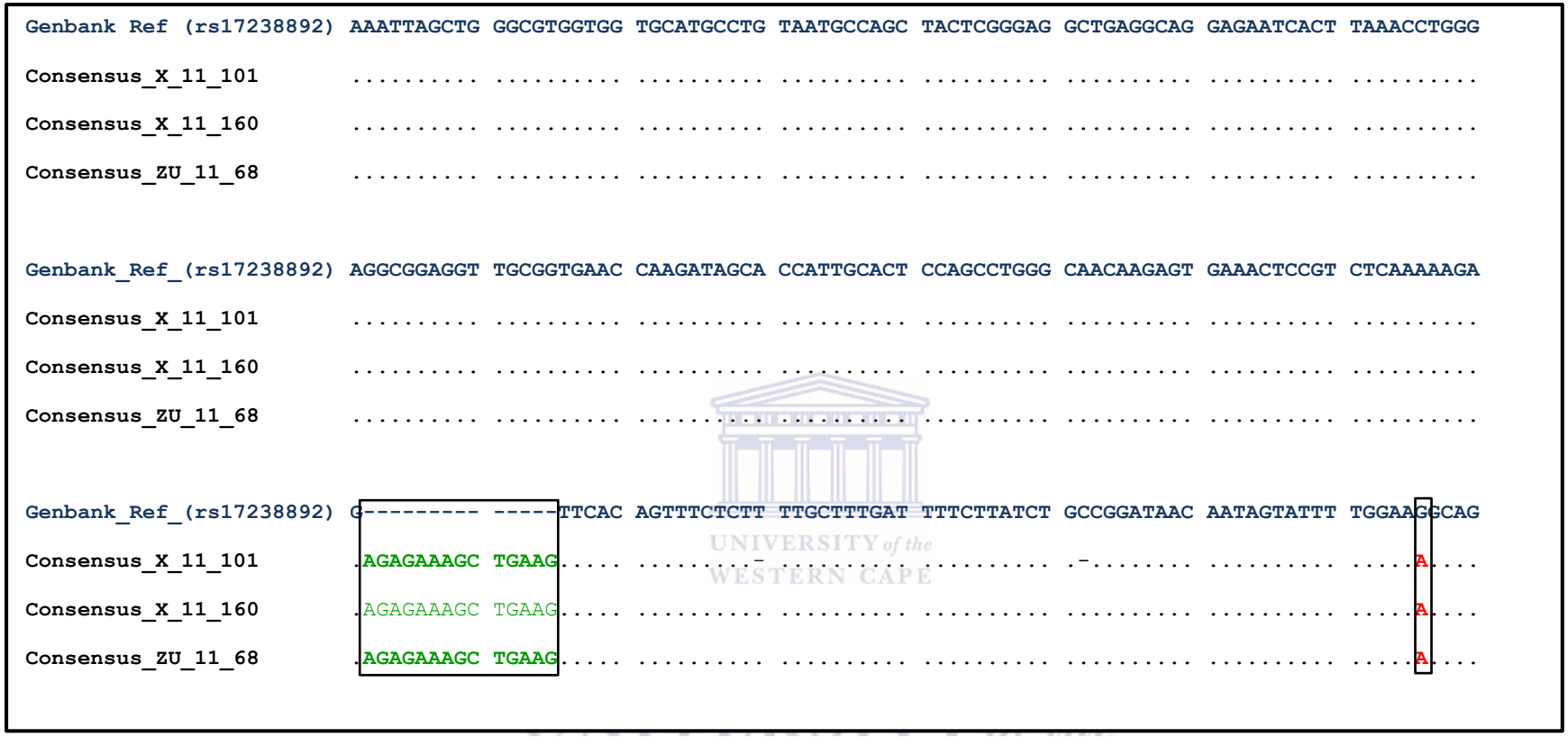
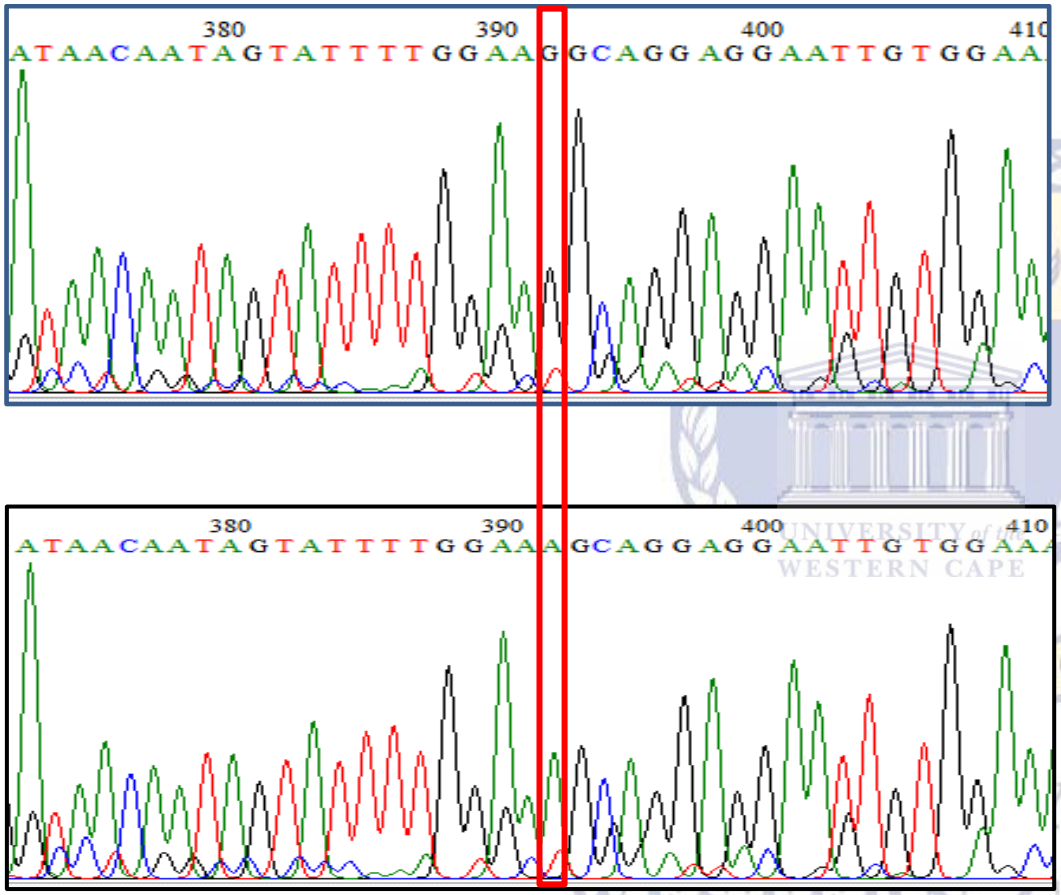


Figure 3.11 Sequence alignments using three of the HLD97 null allele samples with the 14 bp insertion sequence indicated in green. The wild type G nucleotide in the Genbank reference sequence is substituted with an A nucleotide in the null allele samples as indicated in red.



HLD97 INSERTION HOMOZYGOTE

HLD97 NULL ALLELE HOMOZYGOTE

Figure 3.12. Chromatograms indicating the presence of the wild type G nucleotide (**A**) and the mutant A nucleotide resulting in the HLD97 null allele (**B**).

The fourth null allele frequency estimator used is the ML-NullFreq software program (Kalinowski and Taper, 2006) and is likely the most reliable estimator for this particular case. This estimator is based on actual genotype counts and calculates null allele frequencies whether missing data is present or not.

The estimated null allele frequencies are listed in Table 3.2 (A), (B), (C), (D) and (E). Null allele frequencies were calculated using the four different methods for comparison purposes. The presence of a null allele at the HLD97 locus has been demonstrated in the Xhosa and Zulu populations. The allele frequencies for HLD97 calculated using all four estimators listed in Table 3.3. in the primer binding stringency as described in section 3.3.2.Furthermore, the contradicting results obtained for genotyping (insertion homozygote) and direct sequencing (heterozygote) of one individual as described in section 3.3.4.1 is also explained by the presence of a mutation in the primer binding site (Figure 3.13). As a heterozygote individual possesses the insertion as well as the deletion allele, a mutation in one of the two alleles can result in PCR amplification of one allele only. In the case of this heterozygote sample, the insertion allele possesses the mutation, similar to the samples containing the HLD97 null allele. This results in only the deletion allele being detected.

Using the Brookfield estimator 2, allele frequencies at the HLD97 locus in these two populations were higher than 60 % ($0.449 \geq 0.620$). Using the ML-NullFreq program, allele frequencies were between 26 % and 33 % ($0.261 \geq 0.327$). This is expected considering that individuals with no observed amplification at the HLD97 locus were observed in the Xhosa and Zulu populations. In both the Indian and Coloured populations the calculated null allele frequencies were high ($0.118 \geq 0.334$) with the Indian population HLD97 null allele frequency well above 30 %, providing strong evidence for the possible presence of a null allele at this locus as opposed to the Afrikaner population which had null allele frequencies at the HLD97 locus just below 15 % ($0.061 \geq 0.147$).

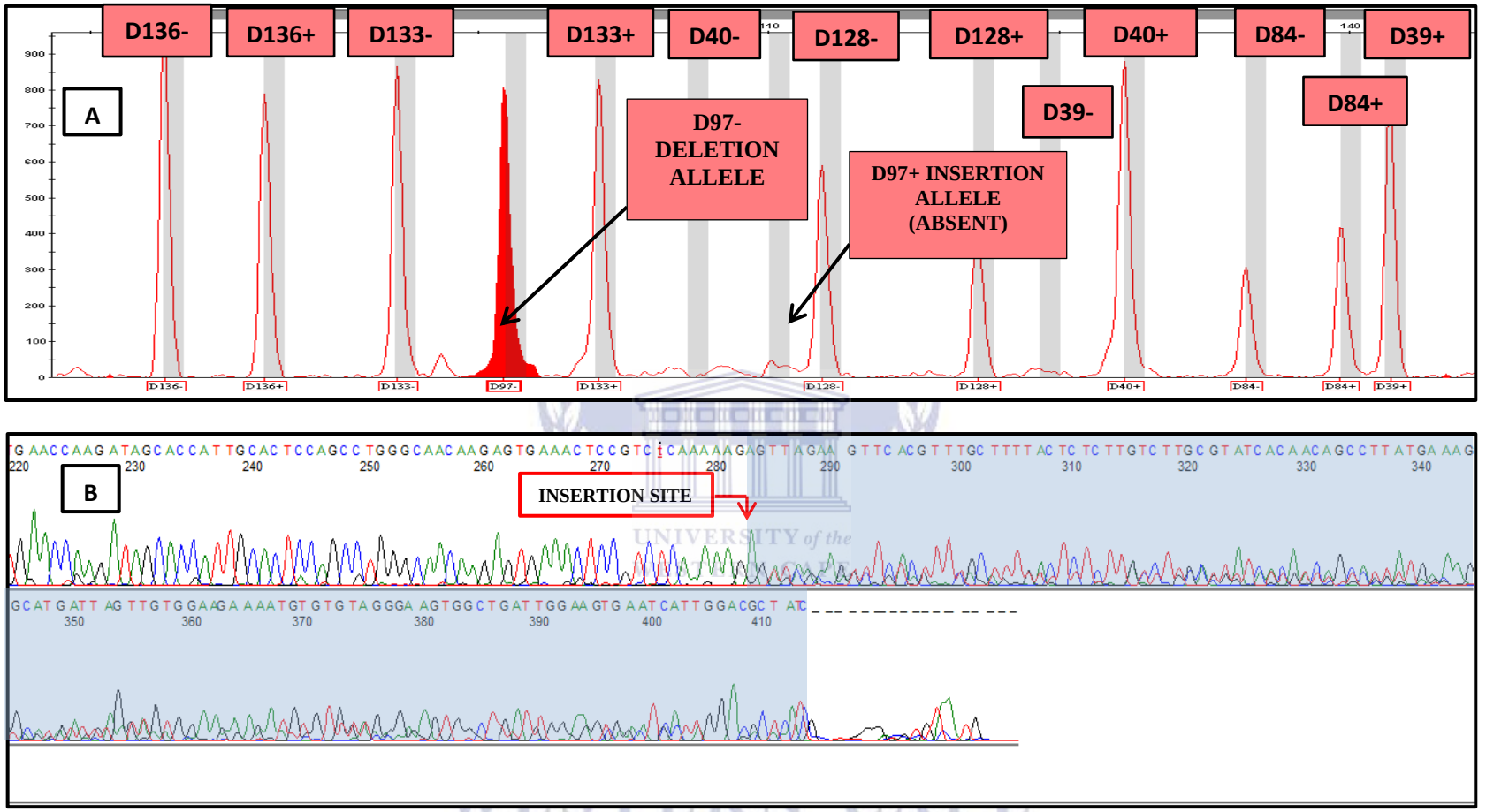


Figure 3.13 (A) An electropherogram of an HLD97 deletion homozygous individual as genotyped. **(B)** The true genotype is confirmed to be an HLD97 heterozygote by direct sequencing**.** The visible mixed trace and decrease in trace quality starting from the insertion site (indicated) is visible.

Table 3.2 (A) Null allele frequency estimates of the 30 indel loci for the Afrikaner population calculated using four methods: Charkraborty null allele estimator [a], Brookfield null allele estimators[b] 1 and 2 and ML-NullFreq software program [c].

	Afrikaner			
	Null Allele Estimator			
Locus	**Charkraborty**	**Brookfield 1**	**Brookfield 2**	**ML-NullFreq**
HLD77	0.007	0.005	0.012	0.003
HLD45	-0.006	-0.004	-0.002	0.000
HLD131	-0.042	-0.029	-0.018	0.000
HLD70	0.048	0.031	0.072	0.029
HLD6	0.024	0.016	0.037	0.014
HLD111	-0.052	-0.036	-0.023	0.000
HLD58	-0.050	-0.035	-0.022	0.000
HLD56	0.045	0.027	0.062	0.027
HLD118	0.017	0.011	0.026	0.009
HLD92	-0.003	-0.002	-0.001	0.000
HLD93	-0.031	-0.022	-0.014	0.000
HLD99	-0.031	-0.021	-0.014	0.000
HLD88	0.016	0.010	0.025	0.009
HLD101	-0.042	-0.029	-0.019	0.000
HLD67	-0.022	-0.015	-0.009	0.000
HLD83	-0.067	-0.048	-0.030	0.000
HLD114	-0.022	-0.015	-0.009	0.000
HLD48	0.060	0.037	0.088	0.036
HLD124	-0.031	-0.021	-0.013	0.000
HLD122	-0.045	-0.031	-0.020	0.000
HLD125	0.017	0.011	0.026	0.009
HLD64	0.009	0.006	0.014	0.005
HLD81	0.016	0.010	0.025	0.009
HLD136	0.017	0.011	0.026	0.009
HLD133	-0.095	-0.069	-0.043	0.000
HLD97	**0.103**	**0.062**	**0.147**	**0.061**
HLD40	-0.059	-0.042	-0.027	0.000
HLD128	0.017	0.011	0.026	0.009
HLD39	0.001	0.001	0.002	0.000
HLD84	0.031	0.020	0.047	0.019

HLD97 locus frequencies with known presence of null allele highlighted in bold

[a] Charkraborty (Charkraborty, 1992)

[b] Brookfield (Brookfield, 1996)

[c] Kalinowski and Taper (2006)

Table 3.2 (B) Null allele frequency estimates of the 30 indel loci for the Mixed Ancestry population calculated using three methods: Charkraborty null allele estimator [a], Brookfield null allele estimators[b] 1 and 2 and ML-NullFreq software program [c].

	Mixed Ancestry			
	Null Allele Estimator			
Locus	**Charkraborty**	**Brookfield 1**	**Brookfield 2**	**ML-NullFreq**
HLD77	-0.044	-0.030	0.000	0.000
HLD45	-0.018	-0.012	0.000	0.000
HLD131	-0.029	-0.020	0.000	0.000
HLD70	-0.016	-0.009	0.000	0.000
HLD6	0.048	0.029	0.039	0.029
HLD111	-0.025	-0.016	0.000	0.000
HLD58	-0.028	-0.018	0.000	0.000
HLD56	0.086	0.051	0.052	0.051
HLD118	-0.041	-0.029	0.000	0.000
HLD92	0.027	0.017	0.023	0.016
HLD93	0.058	0.036	0.034	0.035
HLD99	0.053	0.030	0.007	0.032
HLD88	0.002	0.002	0.000	0.000
HLD101	0.014	0.009	0.000	0.008
HLD67	0.059	0.035	0.025	0.036
HLD83	0.002	0.002	0.018	0.000
HLD114	-0.009	-0.006	0.006	0.000
HLD48	-0.005	-0.003	0.024	0.000
HLD124	0.028	0.018	0.013	0.017
HLD122	-0.007	-0.004	0.000	0.000
HLD125	-0.007	-0.004	0.013	0.000
HLD64	-0.016	-0.008	0.023	0.000
HLD81	0.101	0.059	0.082	0.059
HLD136	-0.001	0.000	0.029	0.000
HLD133	0.053	0.031	0.025	0.032
HLD97	**0.256**	**0.134**	**0.118**	**0.134**
HLD40	-0.069	-0.045	0.000	0.000
HLD128	0.010	0.007	0.015	0.005
HLD39	-0.015	-0.010	0.000	0.000
HLD84	-0.021	-0.013	0.003	0.000

HLD97 locus frequencies with known presence of null allele highlighted in bold

[a] Charkraborty (Charkraborty, 1992)

[b] Brookfield (Brookfield, 1996)

[c] Kalinowski and Taper (2006)

Table 3.2 (C) Null allele frequency estimates of the 30 indel loci for the Indian-Asian population calculated using three methods: Charkraborty null allele estimator [a], Brookfield null allele estimators[b] 1 and 2 and ML-NullFreq software program [c].

	Indian-Asian			
	Null Allele Estimator			
Locus	**Charkraborty**	**Brookfield 1**	**Brookfield 2**	**ML-NullFreq**
HLD77	-0.024	-0.013	-0.007	0.000
LD45	-0.019	-0.012	-0.007	0.000
HLD131	0.000	0.000	0.000	0.000
HLD70	0.010	0.006	0.013	0.005
HLD6	-0.038	-0.026	-0.017	0.000
HLD111	-0.029	-0.018	-0.011	0.000
HLD58	-0.068	-0.044	-0.026	0.000
HLD56	-0.007	-0.004	-0.002	0.000
HLD118	-0.054	-0.038	-0.024	0.000
HLD92	0.035	0.022	0.053	0.021
HLD93	0.030	0.019	0.045	0.018
HLD99	-0.024	-0.016	-0.010	0.000
HLD88	0.066	0.040	0.092	0.039
HLD101	0.043	0.027	0.065	0.026
HLD67	-0.008	-0.005	-0.003	0.000
HLD83	0.041	0.024	0.055	0.025
HLD114	-0.003	-0.002	-0.001	0.000
HLD48	-0.026	-0.018	-0.011	0.000
HLD124	-0.034	-0.023	-0.015	0.000
HLD122	-0.012	-0.008	-0.005	0.000
HLD125	-0.049	-0.034	-0.022	0.000
HLD64	-0.040	-0.019	-0.010	0.000
HLD81	0.048	0.029	0.068	0.029
HLD136	0.002	0.001	0.003	0.000
HLD133	0.051	0.031	0.072	0.031
HLD97	**0.327**	**0.147**	**0.334**	**0.155**
HLD40	0.044	0.024	0.053	0.026
HLD128	-0.046	-0.032	-0.020	0.000
HLD39	-0.007	-0.005	-0.003	0.000
HLD84	0.023	0.014	0.031	0.014

HLD97 locus frequencies with known presence of null allele highlighted in bold

[a] Charkraborty (Charkraborty, 1992)

[b] Brookfield (Brookfield, 1996)

[c] Kalinowski and Taper (2006)

Table 3.2 (D) Null allele frequency estimates of the 30 indel loci for the Xhosa population calculated using three methods: Charkraborty null allele estimator [a], Brookfield null allele estimators[b] 1 and 2 and ML-NullFreq software program [c].

	Xhosa			
	Null Allele Estimator			
Locus	**Charkraborty**	**Brookfield 1**	**Brookfield 2**	**ML-NullFreq**
HLD77	0.017	0.010	0.021	0.010
HLD45	-0.034	-0.022	-0.014	0.000
HLD131	0.057	0.032	0.071	0.034
HLD70	-0.036	-0.009	-0.003	0.000
HLD6	0.017	0.010	0.024	0.010
HLD111	0.013	0.008	0.020	0.007
HLD58	-0.030	-0.012	-0.005	0.000
HLD56	0.074	0.045	0.106	0.044
HLD118	0.009	0.006	0.013	0.004
HLD92	0.022	0.014	0.032	0.013
HLD93	-0.070	-0.051	-0.032	0.000
HLD99	0.021	0.012	0.027	0.012
HLD88	-0.042	-0.026	-0.015	0.000
HLD101	0.057	0.020	0.041	0.032
HLD67	0.009	0.005	0.013	0.004
HLD83	-0.012	-0.008	-0.005	0.000
HLD114	0.038	0.016	0.034	0.022
HLD48	-0.054	-0.033	-0.018	0.000
HLD124	-0.036	-0.018	-0.009	0.000
HLD122	0.007	0.004	0.010	0.003
HLD125	0.107	0.036	0.073	0.055
HLD64	0.018	0.008	0.017	0.010
HLD81	0.020	0.013	0.031	0.012
HLD136	-0.037	-0.022	-0.013	0.000
HLD133	0.012	0.008	0.018	0.007
HLD97	**0.398**	**0.190**	**0.449**	**0.261**
HLD40	0.026	0.014	0.030	0.015
HLD128	-0.108	-0.079	-0.049	0.000
HLD39	0.020	0.013	0.030	0.012
HLD84	-0.008	-0.004	-0.002	0.000

HLD97 locus frequencies with known presence of null allele highlighted in bold

[a] Charkraborty (Charkraborty, 1992)

[b] Brookfield (Brookfield, 1996)

[c] Kalinowski and Taper (2006)

Table 3.2 (E) Null allele frequency estimates of the 30 indel loci for the Zulu populationcalculated using three methods: Charkraborty null allele estimator [a], Brookfield null allele estimators[b] 1 and 2 and ML-NullFreq software program [c].

	Zulu			
	Null Allele Estimator			
Locus	**Charkraborty**	**Brookfield**	**Brookfield 2**	**ML-NullFreq**
HLD77	-0.017	-0.011	-0.006	0.000
HLD45	-0.015	-0.009	-0.005	0.000
HLD131	-0.073	-0.042	-0.023	0.000
HLD70	-0.041	-0.011	-0.004	0.000
HLD6	-0.079	-0.049	-0.027	0.000
HLD111	-0.015	-0.009	-0.005	0.000
HLD58	-0.025	-0.009	-0.004	0.000
HLD56	-0.039	-0.027	-0.017	0.000
HLD118	0.007	0.005	0.011	0.003
HLD92	0.008	0.005	0.012	0.004
HLD93	-0.043	-0.030	-0.019	0.000
HLD99	0.104	0.057	0.129	0.060
HLD88	-0.031	-0.017	-0.009	0.000
HLD101	-0.051	-0.017	-0.007	0.000
HLD67	0.068	0.038	0.086	0.040
HLD83	0.060	0.036	0.084	0.036
HLD114	-0.054	-0.018	-0.008	0.000
HLD48	0.035	0.020	0.046	0.021
HLD124	0.042	0.016	0.032	0.024
HLD122	-0.025	-0.016	-0.009	0.000
HLD125	0.070	0.027	0.055	0.039
HLD64	-0.030	-0.014	-0.007	0.000
HLD81	0.073	0.045	0.106	0.044
HLD136	-0.049	-0.030	-0.018	0.000
HLD133	0.026	0.016	0.036	0.015
HLD97	**0.560**	**0.255**	**0.620**	**0.327**
HLD40	0.024	0.012	0.025	0.014
HLD128	0.043	0.025	0.058	0.026
HLD39	-0.063	-0.042	-0.025	0.000
HLD84	0.023	0.013	0.029	0.013

HLD97 locus frequencies with known presence of null allele highlighted in bold

[a] Charkraborty (Charkraborty, 1992)

[b] Brookfield (Brookfield, 1996)

[c] Kalinowski and Taper (2006)

Table 3.3. Null allele frequency estimates of the HLD97 locus for all populations calculated using four methods: Charkraborty null allele estimator [a], Brookfield null allele estimators[b] 1 and 2 and ML-NullFreq software program [c].

Population	Null Allele Estimator			
	Charkraborty	**Brookfield 1**	**Brookfield 2**	**ML- NullFreq**
Afrikaner	0.103	0.062	0.147	0.061
Mixed Ancestry	0.256	0.134	0.118	0.134
Indian-Asian	0.327	0.147	0.334	0.155
Xhosa	0.398	0.190	0.449	0.261
Zulu	0.560	0.255	0.620	0.327
Average	**0.329**	**0.158**	**0.333**	**0.188**

[a] Charkraborty (Charkraborty, 1992)

[b] Brookfield (Brookfield, 1996)

[c] Kalinowski and Taper (2006)

The observed null allele frequencies across all other loci were much lower than the frequencies observed at the HLD97 locus suggesting the absence of null alleles at these loci.

The fifth method to estimate null allele frequency is using the Hardy-Weinberg equation ($1 = p^2 + 2pq + q^2$). By applying this method of null allele estimation the null allele frequency at locus HLD97 was calculated as 0.343 and 0.279 in the Xhosa and Zulu populations, respectively.

3.4. Summary

No alleles were observed for HLD97 in six of the 102 individuals from the Xhosa population and in eight of the 103 individuals from the Zulu populations, with successful PCR amplification detected at all other loci. This observation led to the hypothesis that the possible presence of a null allele is responsible for the lack of PCR amplification at the HLD97 locus.

PCR amplification using less stringent conditions of four the null allele homozygotes confirmed the true genotype to be that of an insertion homozygote. Confirmation of these 4 individual genotypes via sequencing determined the true genotype to be homozygote for the HLD97 insertion allele. DNA sequencing identified the presence of a G to A substitution in the samples homozygous for the HLD97 null allele, located 75 bp downstream from the HLD97 indel site, on chromosome 13. The presence of the G to A substitution in the primer binding site would affect the primer binding properties, resulting in primer binding failure.

Similar observations have previously been reported at the HLD97 locus with an allele imbalance observed in heterozygotes (Alvarez *et al*, 2011; Fondevila *et al*, 2012; Martin *et al*, 2012) and a partially silent allele at HLD97 reported by Friis *et al*, 2012. In these cases the presence of a G/A transition 75 bp from the insertion site was confirmed through DNA sequencing in all four studies. LaRue *et al* (2012) made the same observation of allele imbalance at HLD97 but did not confirm the presence of the G/A transition through sequencing. The peak

imbalance was hypothesized to be caused by insufficient primer annealing due to the presence of the G/A substitution within the primer binding site.

The null allele frequency estimators indicated that a high number of individuals within the Mixed Ancestry, Indian Asian, Xhosa and Zulu populations may contain the HLD97 insertion null allele. Therefore, considering the three possible alleles i.e. wild type insertion-, mutant insertion allele (null allele), and wild type deletion; six possible genotypes are produced of which only four genotypes are visibly observable:

i) deletion homozygote

ii) insertion homozygote

iii) heterozygote

iv) homozygote null allele

The mutation is carried on the insertion allele. In a heterozygote individual who may possess the mutant insertion allele and the wild type deletion allele, only the deletion allele will be amplified. This will result in the heterozygote individual falsely being observed as a deletion homozygote and consequently an excess of homozygotes.

Null alleles have direct influences on population and forensic genetics. Their presence increases the levels of homozygosity by being falsely genotyped as homozygotes and consequently decreases heterozygosity levels. High enough null allele frequencies can cause a shift in HWE. In doing so null alleles affect population structure estimates by interfering with measurements of genetic diversity and decreasing estimates of relatedness (Kalinowski and Taper, 2006). A more serious implication of the presence of a null allele is the affect it has on possible false exclusions in parentage determination.

With the possible presence of a null allele in a sample population, solutions should be considered to deal with its presence and the effect it might have on population data and kinship analysis:

1. Redesigning the primers so that the primer binding sites are not located in or near the mutation site is a possible option.

2. Another alternative is to design degenerate primers, taking into consideration the mutation site and using these primers in addition to the original primer set. This will accommodate individuals that possess the polymorphic site and result in successful amplification of all carriers.

3. A third option is to exclude the locus containing the null allele altogether.

All three options have considerations to bear in mind. If considering redesigning the primers, the effect of the newly designed primers on the multiplex PCR should be investigated and optimization of multiplex PCR reaction may need to be repeated which might have further consequences on time and labour costs. This approach was chosen when a primer binding polymorphism at D7S820 led to allele dropout using the Promega Powerplex® 1.1 kit (Promega) (Schumm *et al*, 1996).

This also counts for the option of degenerate primers and the time and costs involved needs to be considered in cases where only a few samples containing null alleles are present in proportion to the samples not containing the null allele. Discrepancies in results for the D16S539 locus were detected in data generated by two independent laboratories, using the Promega Powerplex® 1.1 (Promega) (Nelson *et al*, 2002). Upon further investigation a T/A mutation located within the primer binding site was found to be the cause of the null allele presence. As a solution to the problem of the D16S539 null allele, a degenerate primer was designed and included in the Promega Powerplex® 1.1 kit (Promega).

The option to exclude the locus is a viable option when primer binding problems are detected early in development. This option was chosen in the early stages of development of the Forensic Science Service's second generation multiplex when the D19S253 locus was excluded (Urquhart *et al*, 1995).

For application in South Africa, I would suggest excluding the locus containing the null allele from kinship analysis. This will be the most affordable option as

well as being less time consuming and labour intensive, considering the other options relating to primer design.

Chapter 4

Statistical Analysis on population data from five South African populations

4.1. Introduction

Statistics is the science of uncertainty and its measurement and it is used to provide a sense of how reliable a measurement is if made multiple times (Butler, 2005). In forensics, statistical analyses of DNA profiles are mainly performed to establish the occurrence of a DNA profile within a population. This is performed by estimating certain forensic parameters. These include match probability, power of exclusion and discrimination power. All of these are estimates that give an indication of how rare a profile is within a certain population.

Statistical analyses can also reveal information about the population's genetic diversity and history. In addition to the information obtainable on the population using statistics, information can also be obtained on the actual forensic kit used to generate the genetic information eg. the quality of the markers chosen. These forensic parameters can be used to determine a forensic kit's ability to be used as a forensic tool.

This chapter looks at statistical analyses performed on the population data and its application in forensics in the South African population.

4.2 Materials and methods

4.2.1 Samples

DNA was extracted from blood and buccal swabs from Afrikaner (N = 101), Mixed Ancestry (N = 104), Indian-Asian (N = 102), Xhosa (N = 102) and Zulu (N = 103) populations are included in the following analysis. For more details see chapter 2, section 2.2.2).

4.2.2 Estimation of population and forensic parameters

To assess the Investigator kit's efficiency for application in individual identification, certain forensic parameters are estimated. The forensic parameters match probability (MP), power of discrimination (DP), polymorphism information content (PIC), power of exclusion (PE) and typical paternity index (TPI) were estimated using PowerStats (Tereba, 1999). Allele frequencies, expected and observed heterozygosity, as well as probability values (p-values) for Hardy-Weinberg equilibrium were estimated using Arlequin (Excoffier *et al*, 2005).

4.2.3. Population Comparisons

4.2.3.1. Factorial Correspondence Analysis (FCA)

The genetic relationship between the five populations was investigated through Factorial Correspondence Analysis (FCA). FCA is a multivariate statistical method that is used to extract information from contingency tables i.e. allele frequency tables. The aim of FCA is to build modalities of variables to determine if any correlations exist and allows graphical summaries of large datasets. Scatter plots of individuals and populations were constructed through FCA using the computer program Genetix v.4.05.2. (Belkhir *et al*, 2002).

4.2.3.2. F_{ST}-analysis

F_{ST}- analysis is a parameter for estimating genetic differentiation amongst populations. A large θ-value (closer to 1) is indicative of more differentiation whereas a small θ-value (closer to 0) indicates less differentiation amongst populations. The amount of genetic differentiation is related to evolutionary processes that populations experience i.e. genetic drift, migration and mutation. F_{ST}- analysis was estimated using Arlequin (Excoffier *et al*, 2005) with genetic distance being estimated using Slatkin's method (Slatkin, 1995) with 1000 permutations applied for estimation of significance.

4.2.3.3. Phylogenetic tree construction

Phylogenetic tree construction is a way to investigate genetic differences amongst populations. A phylogenetic tree is constructed using genetic distance where the genetic distance calculated between populations is equivalent to the tree branch distance between populations on the phylogenetic tree (Kalinowski, (2009). Genetic similaries between populations are indicated by short branch lengths; the shorter the branches the higher the genetic similaries.

Phylogenetic tree construction was conducted using the program Treefit (Kalinowski, 2009) and visualised using TreeView (Page, 1996). The progam TreeFit allows for calculation of the degree of fit between observed data and phylogenetic tree construction using R^2 statistics; the closer the R^2-value is to 1.0, the better the tree summarises the genetic relationships between the populations (Kalinowski, (2009).

Using the program Treefit (Kalinowski, 2009), the Neighbour-Joining (NJ) method was used to estimate genetic distance using Weir and Cockerham's method (1984) for the phylogenetic tree construction. This method of F_{ST}-analysis is used to estimate population differentiation where the θ-value ranges from 0 to 1 with a value of 0 indicating random mating or panmixis and a value of 1 indicating total population isolation with no genetic diversity between

populations. The NJ method does not assume equal evolutionary rates between populations.

4.3. Results

4.3.1 Estimation of population and forensic parameters

4.3.1.1 Heterozygosity and Hardy-Weinberg Equilibrium

Table 4.1 (A), (B), (C), (D) and (E) summarizes the calculated values for allele frequency and observed and expected heterozygosity for all five populations. The mean observed heterozygosity values for the Afrikaner, Mixed Ancestry, Indian-Asian, Xhosa and Zulu populations were 0.501, 0.454, 0.452, 0.396 and 0.376, respectively. In all but Afrikaner population, observed heterozygosity values at the HLD97 locus were considerably lower compared to the expected heterozygosity values.

The probability values (p-values) for Hardy-Weinberg equilibrium are listed in Table 4.1 (A), (B), (C), (D) and (E) for each locus in the five populations. For the Afrikaner population, all loci were in Hardy-Weinberg equilibrium ($p > 0.05$) except HLD133 ($p < 0.041$). For the other four populations deviation from Hardy-Weinberg equilibrium was detected in the Xhosa population at locus HLD128 ($p < 0.023$) as well as at locus HLD97 in all populations except the Afrikaner population. After Bonferroni correction the significance level changes from 0.05 to 0.00167 to account for multiple testing (Bland and Altman, 1995). After Bonferroni correction, all p-values were greater than 0.00167 and all loci including HLD28 were therefore in Hardy-Weinberg equilibrium (HWE) except the HLD97 locus in the Mixed Ancestry, Indian Asian, Xhosa and Zulu population

4.3.1.2 Forensic parameters

Table 4.2 (A), (B), (C), (D) and (E) lists the estimated values for power of discrimination, match probability, power of exclusion and typical paternity index per locus, for each of the five populations. Polymorphic information content (PIC) is a measure of the polymorphic level of a locus (Goodwin *et al*, 2011). It was determined that a diallelic locus can have a maximum PIC value of 0.375 (Hildebrand *et al*, 1992). The mean PIC values of each of the five populations are listed in Table 4.2 (A), (B), (C), (D) and (E). Again, the highest PIC value (mean = 0.37) was observed in the Afrikaner population for loci HLD77, HLD92, HLD93, HLD99, HLD83, HLD125, HLD136, HLD40 and HLD128. This value is quite high and therefore the loci are considered as informative. For the remaining populations the mean PIC value was in the range of 0.308 - 0.355.

The combined indices (combined power of discrimination, combined match probability and combined power of exclusion) for all five populations are summarized in Table 4.3. The highest values for all three indices were calculated for the Afrikaner population. The highest combined power of discrimination (CDP) as calculated for the Afrikaner population was 0.9999999999996. This value can be interpreted as 99.99999999996% of the Afrikaner population can be excluded as donors of a chosen DNA profile. In the Afrikaner population the Combined Power of Exclusion (CPE) was calculated as 0.9983 with the lowest CPE observed for the Zulu population at 0.9783. For the remaining three populations the CPE was 0.9938, 0.9944 and 0.9838 for the Mixed Ancestry, Indian-Asian and Xhosa populations, respectively. The combined match probability values as listed in Table 4.3 was calculated as 4.25×10^{-13} for the Afrikaner and 3.55×10^{-11} for the Xhosa population.

Table 4.1 (A) Allele frequencies and heterozygosity estimates for the the 30 Indels for the Qiagen Investigator Kit in the Afrikaner population.

	Afrikaner (N=101)			
Locus	**Deletion Allele Frequency**	**Observed Heterozygosity (He)**	**Expected Heterozygosity (Ho)**	**P-Value (HWE)**
HLD77	0.505	0.495	0.502	1.000
HLD45	0.460	0.505	0.499	1.000
HLD131	0.426	0.535	0.491	0.417
HLD70	0.475	0.455	0.501	0.426
HLD6	0.545	0.475	0.499	0.691
HLD111	0.465	0.554	0.500	0.320
HLD58	0.475	0.554	0.501	0.321
HLD56	0.347	0.416	0.455	0.389
HLD118	0.574	0.475	0.491	0.839
HLD92	0.520	0.505	0.502	1.000
HLD93	0.485	0.535	0.502	0.555
HLD99	0.505	0.535	0.502	0.556
HLD88	0.530	0.485	0.501	0.843
HLD101	0.470	0.545	0.501	0.427
HLD67	0.361	0.485	0.464	0.672
HLD83	0.495	0.574	0.502	0.167
HLD114	0.639	0.485	0.464	0.672
HLD48	0.426	0.436	0.491	0.308
HLD124	0.381	0.505	0.474	0.533
HLD122	0.550	0.545	0.498	0.421
HLD125	0.520	0.485	0.502	0.842
HLD64	0.381	0.465	0.474	1.000
HLD81	0.530	0.485	0.501	0.842
HLD136	0.520	0.485	0.502	0.842
HLD133	0.426	0.594	0.491	0.041*
HLD97	0.460	0.406	0.499	0.072
HLD40	0.520	0.564	0.502	0.233
HLD128	0.480	0.485	0.502	0.842
HLD39	0.589	0.485	0.487	1.000
HLD84	0.441	0.465	0.495	0.549
Mean		0.501	0.493	
s.d.		0.045	0.013	

*** p < 0.00167; HWE = Hardy Weinberg Equilibrium**

Table 4.1 (B) Allele frequencies and heterozygosity estimates for the the 30 Indels for the Qiagen Investigator Kit in the Mixed Ancestry population.

	Mixed Ancestry (N= 104)			
Locus	**Deletion Allele Frequency**	**Observed Heterozygosity (He)**	**Expected Heterozygosity (Ho)**	**P-Value (HWE)**
HLD77	0.615	0.519	0.476	0.406
HLD45	0.572	0.510	0.492	0.840
HLD131	0.457	0.529	0.499	0.558
HLD70	0.255	0.394	0.382	0.802
HLD6	0.635	0.423	0.466	0.402
HLD111	0.615	0.500	0.476	0.680
HLD58	0.654	0.481	0.455	0.666
HLD56	0.370	0.394	0.469	0.139
HLD118	0.558	0.538	0.496	0.427
HLD92	0.587	0.462	0.487	0.687
HLD93	0.409	0.433	0.486	0.311
HLD99	0.308	0.385	0.428	0.357
HLD88	0.380	0.471	0.473	1.000
HLD101	0.317	0.423	0.435	0.822
HLD67	0.346	0.404	0.455	0.284
HLD83	0.620	0.471	0.473	1.000
HLD114	0.399	0.490	0.482	1.000
HLD48	0.409	0.490	0.486	1.000
HLD124	0.543	0.471	0.499	0.693
HLD122	0.668	0.452	0.446	1.000
HLD125	0.567	0.500	0.493	1.000
HLD64	0.212	0.346	0.335	1.000
HLD81	0.375	0.385	0.471	0.093
HLD136	0.462	0.500	0.499	1.000
HLD133	0.663	0.404	0.449	0.378
HLD97	0.587	0.288	0.487	0.000*
HLD40	0.683	0.500	0.435	0.172
HLD128	0.466	0.490	0.500	0.847
HLD39	0.563	0.510	0.495	0.842
HLD84	0.327	0.462	0.442	0.823
Mean		0.454	0.466	
s.d.		0.059	0.037	

*** p < 0.00167; HWE = Hardy Weinberg Equilibrium**

Table 4.1 (C) Allele frequencies and heterozygosity estimates for the the 30 Indels for the Qiagen Investigator Kit in the Indian-Asian population.

	Indian-Asian (N=102)			
Locus	**Deletion Allele Frequency**	**Observed Heterozygosity (He)**	**Expected Heterozygosity (Ho)**	**P-Value (HWE)**
HLD77	0.779	0.363	0.346	0.775
HLD45	0.328	0.461	0.443	0.823
HLD131	0.578	0.490	0.490	1.000
HLD70	0.275	0.392	0.400	0.809
HLD6	0.534	0.539	0.500	0.549
HLD111	0.696	0.451	0.425	0.641
HLD58	0.681	0.500	0.436	0.174
HLD56	0.206	0.333	0.329	1.000
HLD118	0.485	0.559	0.502	0.320
HLD92	0.436	0.461	0.494	0.549
HLD93	0.539	0.471	0.499	0.690
HLD99	0.387	0.500	0.477	0.679
HLD88	0.373	0.412	0.470	0.288
HLD101	0.515	0.461	0.502	0.433
HLD67	0.520	0.510	0.502	1.000
HLD83	0.681	0.402	0.436	0.493
HLD114	0.588	0.490	0.487	1.000
HLD48	0.500	0.529	0.502	0.693
HLD124	0.407	0.520	0.485	0.543
HLD122	0.662	0.461	0.450	0.829
HLD125	0.451	0.549	0.498	0.323
HLD64	0.181	0.324	0.298	0.513
HLD81	0.382	0.431	0.475	0.402
HLD136	0.485	0.500	0.502	1.000
HLD133	0.632	0.422	0.467	0.394
HLD97	0.696	0.216	0.425	0.000*
HLD40	0.730	0.363	0.396	0.451
HLD128	0.529	0.549	0.501	0.428
HLD39	0.510	0.510	0.502	1.000
HLD84	0.299	0.402	0.421	0.643
Mean		0.452	0.455	
s.d.		0.079	0.056	

*** p < 0.00167; HWE = Hardy Weinberg Equilibrium**

Table 4.1 (D) Allele frequencies and heterozygosity estimates for the the 30 Indels for the Qiagen Investigator Kit in the Xhosa population.

	Xhosa (N= 102)			
Locus	**Deletion Allele Frequency**	**Observed Heterozygosity (He)**	**Expected Heterozygosity (Ho)**	**P-Value (HWE)**
HLD77	0.730	0.382	0.396	0.802
HLD45	0.632	0.500	0.467	0.528
HLD131	0.294	0.373	0.417	0.340
HLD70	0.074	0.147	0.137	1.000
HLD6	0.681	0.422	0.436	0.820
HLD111	0.598	0.471	0.483	0.839
HLD58	0.868	0.245	0.231	1.000
HLD56	0.417	0.422	0.489	0.220
HLD118	0.608	0.471	0.479	1.000
HLD92	0.603	0.461	0.481	0.683
HLD93	0.495	0.578	0.502	0.164
HLD99	0.284	0.392	0.409	0.807
HLD88	0.314	0.471	0.433	0.490
HLD101	0.132	0.206	0.231	0.374
HLD67	0.353	0.451	0.459	1.000
HLD83	0.417	0.500	0.489	0.840
HLD114	0.172	0.265	0.286	0.486
HLD48	0.270	0.441	0.396	0.316
HLD124	0.794	0.353	0.329	0.553
HLD122	0.632	0.461	0.467	1.000
HLD125	0.868	0.186	0.231	0.069
HLD64	0.186	0.294	0.305	0.745
HLD81	0.466	0.480	0.500	0.695
HLD136	0.275	0.431	0.400	0.467
HLD133	0.623	0.461	0.472	0.835
HLD97	0.637	0.216	0.501	0.000*
HLD40	0.755	0.353	0.372	0.599
HLD128	0.397	0.598	0.481	0.023*
HLD39	0.373	0.451	0.470	0.832
HLD84	0.270	0.402	0.396	1.000
Mean		0.396	0.405	
s.d.		0.113	0.098	

*** p < 0.00167; HWE = Hardy Weinberg Equilibrium**

Table 4.1 (E) Allele frequencies and heterozygosity estimates for the the 30 Indels for the Qiagen Investigator Kit in the Zulu population.

	Zulu (N=103)			
Locus	**Deletion Allele Frequency**	**Observed Heterozygosity (He)**	**Expected Heterozygosity (Ho)**	**P-Value (HWE)**
HLD77	0.675	0.456	0.441	0.824
HLD45	0.709	0.427	0.415	0.816
HLD131	0.243	0.427	0.369	0.175
HLD70	0.083	0.165	0.152	1.000
HLD6	0.728	0.466	0.398	0.088
HLD111	0.709	0.427	0.415	0.815
HLD58	0.874	0.233	0.222	1.000
HLD56	0.505	0.544	0.502	0.436
HLD118	0.621	0.466	0.473	1.000
HLD92	0.597	0.476	0.484	1.000
HLD93	0.456	0.544	0.499	0.427
HLD99	0.311	0.350	0.430	0.067
HLD88	0.248	0.398	0.374	0.602
HLD101	0.102	0.204	0.184	0.593
HLD67	0.301	0.369	0.423	0.241
HLD83	0.374	0.417	0.470	0.293
HLD114	0.107	0.214	0.192	0.599
HLD48	0.306	0.398	0.427	0.494
HLD124	0.859	0.223	0.243	0.414
HLD122	0.684	0.456	0.434	0.652
HLD125	0.850	0.223	0.257	0.235
HLD64	0.170	0.301	0.283	0.731
HLD81	0.437	0.427	0.494	0.226
HLD136	0.301	0.466	0.423	0.353
HLD133	0.646	0.437	0.460	0.669
HLD97	0.563	0.155	0.551	0.000*
HLD40	0.796	0.311	0.326	0.760
HLD128	0.330	0.408	0.444	0.501
HLD39	0.345	0.515	0.454	0.196
HLD84	0.282	0.388	0.407	0.635
Mean		0.376	0.388	
s.d.		0.113	0.106	

*** p < 0.00167; HWE = Hardy Weinberg Equilibrium**

Table 4.2 (A) Forensic parameters for the 30 Indels from the Qiagen Investigator Kit in the Afrikaner population

HLD	Chromosomal Location	GenBank SNP ID	Afrikaner				
			PD	MP	PE	TPI	PIC
HLD77	**7q31.1**	**rs1611048**	0.627	0.373	0.183	0.990	0.375
HLD45	**2q31.1**	**rs2307959**	0.619	0.381	0.192	1.010	0.373
HLD131	**7q36.2**	**rs1611001**	0.595	0.405	0.220	1.074	0.369
HLD70	**6q16.1**	**rs2307652**	0.643	0.357	0.151	0.918	0.374
HLD6	**16q13**	**rs1610905**	0.632	0.368	0.167	0.953	0.373
HLD111	**17p11.2**	**rs1305047**	0.591	0.409	0.240	1.122	0.374
HLD58	**5q14.1**	**rs1610937**	0.592	0.408	0.240	1.122	0.374
HLD56	**4q25**	**rs2308292**	0.609	0.391	0.124	0.856	0.350
HLD118	**20p11.1**	**rs16438**	0.625	0.375	0.167	0.953	0.369
HLD92	**11q22.2**	**rs17174476**	0.622	0.378	0.192	1.010	0.375
HLD93	**12q22**	**rs2307570**	0.605	0.395	0.220	1.074	0.375
HLD99	**14q23.1**	**rs2308163**	0.606	0.394	0.220	1.074	0.375
HLD88	**9q22.32**	**rs8190570**	0.630	0.370	0.175	0.971	0.374
HLD101	**15q26.1**	**rs2307433**	0.598	0.402	0.230	1.098	0.374
HLD67	**5q33.2**	**rs1305056**	0.594	0.406	0.175	0.971	0.355
HLD83	**8p22**	**rs2308072**	0.580	0.420	0.261	1.174	0.375
HLD114	**17p13.3**	**rs2307581**	0.594	0.406	0.175	0.971	0.355
HLD48	**2q11.2**	**rs28369942**	0.640	0.360	0.137	0.886	0.369
HLD124	**22q12.3**	**rs6481**	0.594	0.406	0.192	1.010	0.360
HLD122	**21q22.11**	**rs8178524**	0.595	0.405	0.230	1.098	0.373
HLD125	**22q11.23**	**rs16388**	0.631	0.369	0.175	0.971	0.375
HLD64	**5q12.3**	**rs1610935**	0.612	0.388	0.159	0.935	0.360

Table continued

HLD	Chromosomal Location	GenBank SNP ID	Afrikaner				
			PD	MP	PE	TPI	PIC
HLD81	**7q21.3**	**rs17879936**	0.630	0.370	0.175	0.971	0.374
HLD136	**22q13.1**	**rs16363**	0.631	0.369	0.175	0.971	0.375
HLD133	**3p22.1**	**rs2067235**	0.554	0.446	0.284	1.232	0.369
HLD97	**13q12.3**	**rs17238892**	0.656	0.344	0.118	0.842	0.373
HLD40	**1p32.3**	**rs2307956**	0.586	0.414	0.250	1.148	0.375
HLD128	**1q31.3**	**rs2307924**	0.631	0.369	0.175	0.971	0.375
HLD39	**1p22.1**	**rs17878444**	0.616	0.384	0.175	0.971	0.367
HLD84	**8q24.12**	**rs3081400**	0.633	0.367	0.159	0.935	0.371
Mean			0.612	0.388	0.191	1.009	0.370

PD, Power of Discrimination; MP, Match Probability; PE, Power of Exclusion; TPI, Typical Paternity Index; PIC, Polymorphic Information Content

Table 4.2 (B) Forensic parameters for the 30 Indels from the Qiagen Investigator Kit in the Mixed Ancestry population

HLD	Chromosomal Location	GenBank SNP ID	Mixed Ancestry				
			PD	MP	PE	TPI	PIC
HLD77	**7q31.1**	**rs1611048**	0.588	0.412	0.205	1.040	0.361
HLD45	**2q31.1**	**rs2307959**	0.610	0.390	0.196	1.020	0.370
HLD131	**7q36.2**	**rs1611001**	0.606	0.394	0.214	1.061	0.373
HLD70	**6q16.1**	**rs2307652**	0.541	0.459	0.110	0.825	0.308
HLD6	**16q13**	**rs1610905**	0.618	0.382	0.129	0.867	0.356
HLD111	**17p11.2**	**rs1305047**	0.598	0.402	0.188	1.000	0.361
HLD58	**5q14.1**	**rs1610937**	0.587	0.413	0.171	0.963	0.350
HLD56	**4q25**	**rs2308292**	0.627	0.373	0.110	0.825	0.358
HLD118	**20p11.1**	**rs16438**	0.597	0.403	0.223	1.083	0.372
HLD92	**11q22.2**	**rs17174476**	0.627	0.373	0.156	0.929	0.367
HLD93	**12q22**	**rs2307570**	0.635	0.365	0.135	0.881	0.367
HLD99	**14q23.1**	**rs2308163**	0.589	0.411	0.105	0.813	0.335
HLD88	**9q22.32**	**rs8190570**	0.609	0.391	0.163	0.945	0.360
HLD101	**15q26.1**	**rs2307433**	0.588	0.412	0.129	0.867	0.339
HLD67	**5q33.2**	**rs1305056**	0.612	0.388	0.116	0.839	0.350
HLD83	**8p22**	**rs2308072**	0.609	0.391	0.163	0.945	0.360
HLD114	**17p13.3**	**rs2307581**	0.609	0.391	0.179	0.981	0.365
HLD48	**2q11.2**	**rs28369942**	0.613	0.387	0.179	0.981	0.367
HLD124	**22q12.3**	**rs6481**	0.634	0.366	0.163	0.945	0.373
HLD122	**21q22.11**	**rs8178524**	0.589	0.411	0.149	0.912	0.345
HLD125	**22q11.23**	**rs16388**	0.616	0.384	0.188	1.000	0.370
HLD64	**5q12.3**	**rs1610935**	0.500	0.500	0.084	0.765	0.278

Table continued

HLD	Chromosomal Location	GenBank SNP ID	Mixed Ancestry				
			PD	MP	PE	TPI	PIC
HLD81	**7q21.3**	**rs17879936**	0.631	0.369	0.105	0.813	0.359
HLD136	**22q13.1**	**rs16363**	0.622	0.378	0.188	1.000	0.374
HLD133	**3p22.1**	**rs2067235**	0.606	0.394	0.116	0.839	0.347
HLD97	**13q12.3**	**rs17238892**	0.649	0.351	0.059	0.703	0.367
HLD40	**1p32.3**	**rs2307956**	0.558	0.442	0.188	1.000	0.339
HLD128	**1q31.3**	**rs2307924**	0.627	0.373	0.179	0.981	0.374
HLD39	**1p22.1**	**rs17878444**	0.612	0.388	0.196	1.020	0.371
HLD84	**8q24.12**	**rs3081400**	0.582	0.418	0.156	0.929	0.343
Mean			0.603	0.397	0.155	0.926	0.355

PD, Power of Discrimination; MP, Match Probability; PE, Power of Exclusion; TPI, Typical Paternity Index; PIC, Polymorphic Information Content

Table 4.2 (C) Forensic parameters for the 30 Indels from the Qiagen Investigator Kit in the Asian-Indian population

HLD	Chromosomal Location	GenBank SNP ID	Asian Indian				
			PD	MP	PE	TPI	PIC
HLD77	**7q31.1**	**rs1611048**	0.509	0.491	0.093	0.785	0.285
HLD45	**2q31.1**	**rs2307959**	0.583	0.417	0.155	0.927	0.344
HLD131	**7q36.2**	**rs1611001**	0.617	0.383	0.179	0.981	0.369
HLD70	**6q16.1**	**rs2307652**	0.560	0.440	0.109	0.823	0.319
HLD6	**16q13**	**rs1610905**	0.601	0.399	0.224	1.085	0.374
HLD111	**17p11.2**	**rs1305047**	0.569	0.431	0.148	0.911	0.334
HLD58	**5q14.1**	**rs1610937**	0.559	0.441	0.188	1.000	0.340
HLD56	**4q25**	**rs2308292**	0.494	0.506	0.078	0.750	0.274
HLD118	**20p11.1**	**rs16438**	0.590	0.410	0.244	1.133	0.375
HLD92	**11q22.2**	**rs17174476**	0.634	0.366	0.155	0.927	0.371
HLD93	**12q22**	**rs2307570**	0.635	0.365	0.163	0.944	0.373
HLD99	**14q23.1**	**rs2308163**	0.600	0.400	0.188	1.000	0.362
HLD88	**9q22.32**	**rs8190570**	0.625	0.375	0.121	0.850	0.358
HLD101	**15q26.1**	**rs2307433**	**0.642**	0.358	0.155	0.927	0.375
HLD67	**5q33.2**	**rs1305056**	0.619	0.381	0.196	1.020	0.375
HLD83	**8p22**	**rs2308072**	0.594	0.406	0.115	0.836	0.340
HLD114	**17p13.3**	**rs2307581**	0.614	0.386	0.179	0.981	0.367
HLD48	**2q11.2**	**rs28369942**	0.609	0.391	0.215	1.063	0.375
HLD124	**22q12.3**	**rs6481**	0.597	0.403	0.205	1.041	0.366
HLD122	**21q22.11**	**rs8178524**	0.590	0.410	0.155	0.927	0.347
HLD125	**22q11.23**	**rs16388**	0.592	0.408	0.234	1.109	0.373
HLD64	**5q12.3**	**rs1610935**	0.463	0.537	0.074	0.739	0.253

Table continued

HLD	Chromosomal Location	GenBank SNP ID	Asian Indian				
			PD	MP	PE	TPI	PIC
HLD81	**7q21.3**	**rs17879936**	0.613	0.387	0.148	0.911	0.358
HLD136	**22q13.1**	**rs16363**	0.625	0.375	0.188	1.000	0.375
HLD133	**3p22.1**	**rs2067235**	0.620	0.380	0.128	0.864	0.357
HLD97	**13q12.3**	**rs17238892**	0.569	0.431	0.034	0.638	0.334
HLD40	**1p32.3**	**rs2307956**	0.559	0.441	0.093	0.785	0.316
HLD128	**1q31.3**	**rs2307924**	0.595	0.405	0.234	1.109	0.374
HLD39	**1p22.1**	**rs17878444**	0.620	0.380	0.196	1.020	0.375
HLD84	**8q24.12**	**rs3081400**	0.579	0.421	0.115	0.836	0.331
Mean			0.589	0.411	0.157	0.931	0.349

PD, Power of Discrimination; MP, Match Probability; PE, Power of Exclusion; TPI, Typical Paternity Index; PIC, Polymorphic Information Content

Table 4.2 (D) Forensic parameters for the 30 Indels from the Qiagen Investigator Kit in the Xhosa population

HLD	Chromosomal Location	GenBank SNP ID	Xhosa				
			PD	MP	PE	TPI	PIC
HLD77	**7q31.1**	**rs1611048**	0.557	0.443	0.104	0.810	0.316
HLD45	**2q31.1**	**rs2307959**	0.590	0.410	0.188	1.000	0.357
HLD131	**7q36.2**	**rs1611001**	0.580	0.420	0.098	0.797	0.329
HLD70	**6q16.1**	**rs2307652**	0.251	0.749	0.017	0.586	0.127
HLD6	**16q13**	**rs1610905**	0.589	0.411	0.128	0.864	0.340
HLD111	**17p11.2**	**rs1305047**	0.619	0.381	0.163	0.944	0.365
HLD58	**5q14.1**	**rs1610937**	0.385	0.615	0.043	0.662	0.203
HLD56	**4q25**	**rs2308292**	0.641	0.359	0.128	0.864	0.368
HLD118	**20p11.1**	**rs16438**	0.615	0.385	0.163	0.944	0.363
HLD92	**11q22.2**	**rs17174476**	0.617	0.383	0.167	0.954	0.365
HLD93	**12q22**	**rs2307570**	0.577	0.423	0.266	1.186	0.375
HLD99	**14q23.1**	**rs2308163**	0.568	0.432	0.109	0.823	0.324
HLD88	**9q22.32**	**rs8190570**	0.569	0.431	0.163	0.944	0.338
HLD101	**15q26.1**	**rs2307433**	0.372	0.628	0.031	0.630	0.203
HLD67	**5q33.2**	**rs1305056**	0.603	0.397	0.148	0.911	0.352
HLD83	**8p22**	**rs2308072**	0.611	0.389	0.188	1.000	0.368
HLD114	**17p13.3**	**rs2307581**	0.444	0.556	0.050	0.680	0.244
HLD48	**2q11.2**	**rs28369942**	0.543	0.457	0.141	0.895	0.316
HLD124	**22q12.3**	**rs6481**	0.493	0.507	0.088	0.773	0.274
HLD122	**21q22.11**	**rs8178524**	0.607	0.393	0.155	0.927	0.357
HLD125	**22q11.23**	**rs16388**	0.364	0.636	0.026	0.614	0.203
HLD64	**5q12.3**	**rs1610935**	0.468	0.532	0.061	0.708	0.257

Table continued

HLD	Chromosomal Location	GenBank SNP ID	Xhosa				
			PD	MP	PE	TPI	PIC
HLD81	**7q21.3**	**rs17879936**	0.632	0.368	0.171	0.962	0.374
HLD136	**22q13.1**	**rs16363**	0.551	0.449	0.134	0.879	0.319
HLD133	**3p22.1**	**rs2067235**	0.612	0.388	0.155	0.927	0.360
HLD97	**13q12.3**	**rs17238892**	0.606	0.394	0.054	0.689	0.346
HLD40	**1p32.3**	**rs2307956**	0.536	0.464	0.088	0.773	0.302
HLD128	**1q31.3**	**rs2307924**	0.540	0.460	0.289	1.244	0.364
HLD39	**1p22.1**	**rs17878444**	0.613	0.387	0.148	0.911	0.358
HLD84	**8q24.12**	**rs3081400**	0.553	0.447	0.115	0.836	0.316
Mean			0.544	0.456	0.126	0.858	0.316

PD, Power of Discrimination; MP, Match Probability; PE, Power of Exclusion; TPI, Typical Paternity Index; PIC, Polymorphic Information Content

Table 4.2 (E) Forensic parameters for the 30 Indels from the Qiagen Investigator Kit in the Zulu population

HLD	Chromosomal Location	GenBank SNP ID	Zulu				
			PD	MP	PE	TPI	PIC
HLD77	**7q31.1**	**rs1611048**	0.583	0.417	0.152	0.920	0.343
HLD45	**2q31.1**	**rs2307959**	0.566	0.434	0.131	0.873	0.328
HLD131	**7q36.2**	**rs1611001**	0.521	0.479	0.131	0.873	0.300
HLD70	**6q16.1**	**rs2307652**	0.276	0.724	0.021	0.599	0.140
HLD6	**16q13**	**rs1610905**	0.536	0.464	0.159	0.936	0.318
HLD111	**17p11.2**	**rs1305047**	0.566	0.434	0.131	0.873	0.328
HLD58	**5q14.1**	**rs1610937**	0.372	0.628	0.039	0.652	0.196
HLD56	**4q25**	**rs2308292**	0.600	0.400	0.229	1.096	0.375
HLD118	**20p11.1**	**rs16438**	0.611	0.389	0.159	0.936	0.360
HLD92	**11q22.2**	**rs17174476**	0.617	0.383	0.167	0.954	0.365
HLD93	**12q22**	**rs2307570**	0.596	0.404	0.229	1.096	0.373
HLD99	**14q23.1**	**rs2308163**	0.595	0.405	0.086	0.769	0.337
HLD88	**9q22.32**	**rs8190570**	0.533	0.467	0.113	0.831	0.303
HLD101	**15q26.1**	**rs2307433**	0.325	0.675	0.031	0.628	0.166
HLD67	**5q33.2**	**rs1305056**	0.586	0.414	0.096	0.792	0.332
HLD83	**8p22**	**rs2308072**	0.624	0.376	0.125	0.858	0.359
HLD114	**17p13.3**	**rs2307581**	0.336	0.664	0.034	0.636	0.173
HLD48	**2q11.2**	**rs28369942**	0.585	0.415	0.113	0.831	0.334
HLD124	**22q12.3**	**rs6481**	0.390	0.610	0.036	0.644	0.213
HLD122	**21q22.11**	**rs8178524**	0.576	0.424	0.152	0.920	0.339
HLD125	**22q11.23**	**rs16388**	0.404	0.596	0.036	0.644	0.223
HLD64	**5q12.3**	**rs1610935**	0.447	0.553	0.064	0.715	0.242

Table continued

HLD	Chromosomal Location	GenBank SNP ID	Zulu				
			PD	MP	PE	TPI	PIC
HLD81	**7q21.3**	**rs17879936**	0.645	0.355	0.131	0.873	0.371
HLD136	**22q13.1**	**rs16363**	0.561	0.439	0.159	0.936	0.332
HLD133	**3p22.1**	**rs2067235**	0.608	0.392	0.138	0.888	0.353
HLD97	**13q12.3**	**rs17238892**	0.632	0.368	0.046	0.669	0.362
HLD40	**1p32.3**	**rs2307956**	0.569	0.431	0.201	1.030	0.350
HLD128	**1q31.3**	**rs2307924**	0.601	0.399	0.119	0.844	0.344
HLD39	**1p22.1**	**rs17878444**	0.569	0.431	0.201	1.030	0.350
HLD84	**8q24.12**	**rs3081400**	0.567	0.433	0.107	0.817	0.323
Mean			0.533	0.467	0.118	0.839	0.308

PD, Power of Discrimination; MP, Match Probability; PE, Power of Exclusion; TPI, Typical Paternity Index; PIC, Polymorphic Information Content

Table 4.3 Combined indices for the 30 Indels from the Qiagen Investigator Kit in five South African populations

Population Group	CDP	CMP	CPE
Afrikaner	0.9999999999996	4.25×10^{-13}	0.9983
Mixed Ancestry	0.9999999999992	8.50×10^{-13}	0.9938
Indian Asian	0.999999999998	2.23×10^{-12}	0.9944
Xhosa	0.99999999996	3.55×10^{-11}	0.9838
Zulu	0.99999999994	6.42×10^{-11}	0.9783

CDP, Combined Power of Discrimination; CMP, Combined Match Probability; CPE, Combined Power of Exclusion

4.3.2 Population Comparisons

4.3.2.1 Factorial Correspondence Analysis (FCA)

Figure 4.1 represents a graph of the individuals belonging to the five South African populations. The graph illustrates the relationship between the individuals from the five populations. Individuals from each population can be seen clustered together. Individuals from the Xhosa and Zulu populations are closely clustered together, an indication of genetic similarities between the two populations. The Mixed Ancestry individuals are dispersed throughout the graph, indicating genetic similaries shared with the other populations whereas the Afrikaner and Indian-Asian individuals are more distinctive indicating decreased genetic relationship with these populations. Figure 4.2 shows the FCA relationship between the five populations and their genetic relatedness. The genetic distance between the populations is shown, with populations sharing genetic similarities being located close to each other and populations located far apart not sharing a high degree of genetic similarity. The centralised positioning of the Mixed Ancestry population indicates this population's shared genetic relationship with all four populations.

4.3.2.2. F_{ST}-Analysis

Table 4.4 shows the pairwise genetic distance as calculated between the five populations by the program Arlequin (Excoffier *et al*, 2005) with the probability values of the pairwise genetic distances shown. The genetic distance estimates results show the Zulu population being closer to the Xhosa population ($\theta = 0.00125$, $-0.01235 < p < 0.16504$) and being the furthest from the Afrikaner ($\theta = 0.11886$, $p < 0.05$) and Indian-Asian populations ($\theta = 0.10909$, $p < 0.05$). There is indication of genetic similarity between the Mixed Ancestry population and each of the four populations i.e. Afrikaner population ($\theta = 0.03020$, $p < 0.00000$), Indian-Asian population ($\theta = 0.02149$, $p < 0.00000$), Xhosa population ($\theta = 0.03568$, $p < 0.00000$) and Zulu population ($\theta = 0.04800$, $p < 0.00000$).

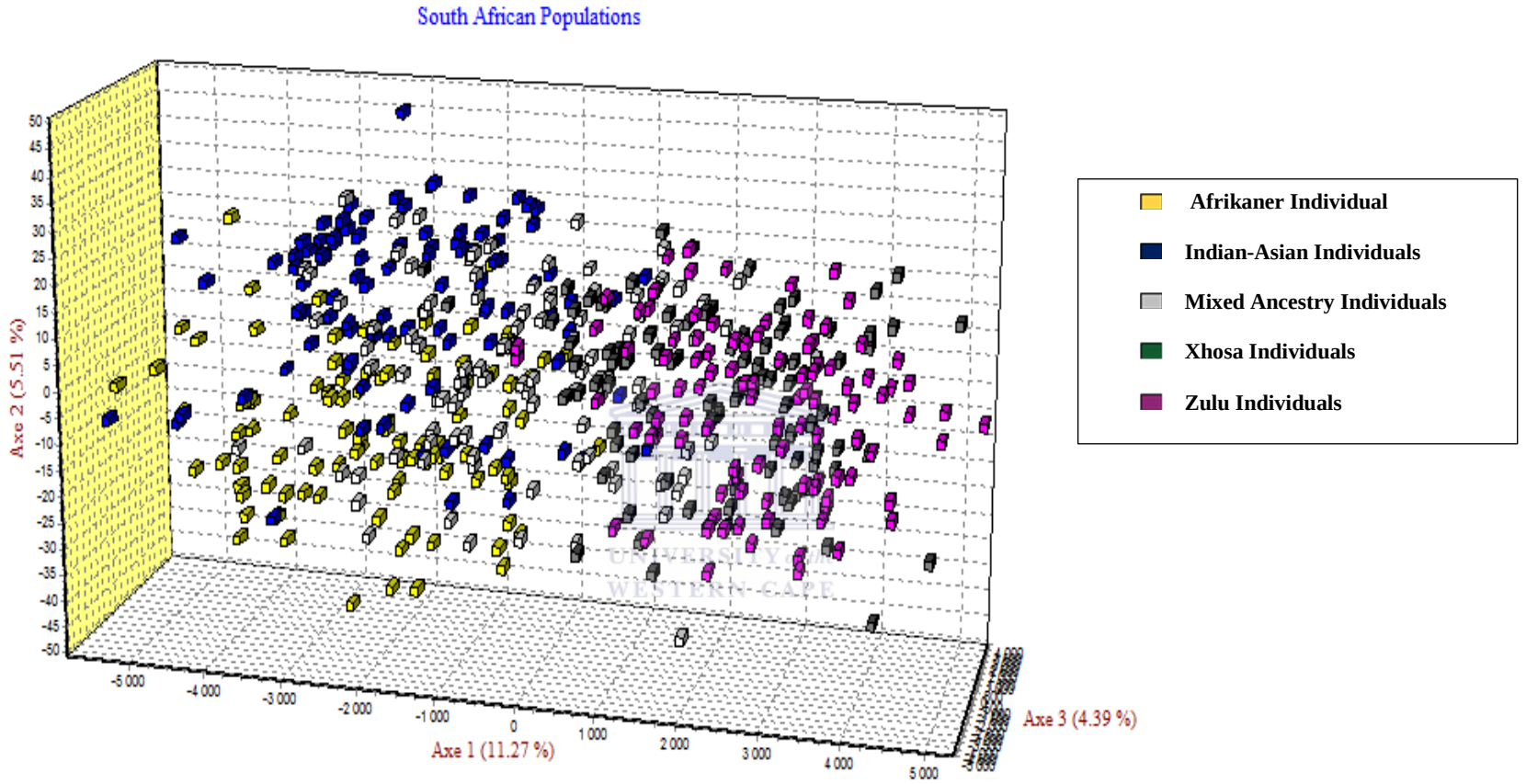


Figure 4.1 Factorial Correspondence Analysis (FCA) of five South African populations' individuals. A graphical illustration of the individuals belonging to the five South African populations as constructed using Genetix v.4.05.2. (Belkhir *et al*, 2002).

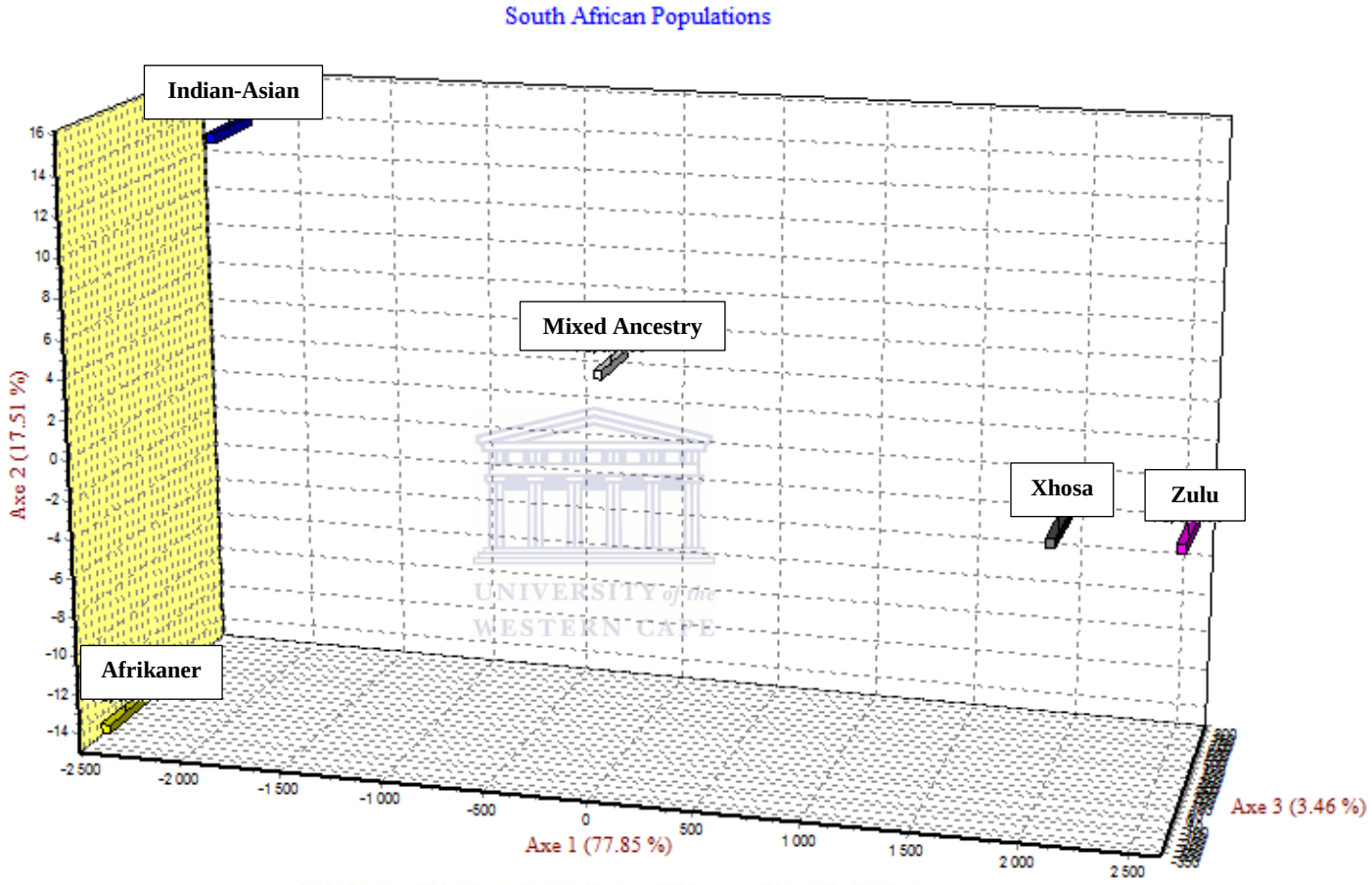


Figure 4.2 Factorial Correspondence Analysis (FCA) of five South African populations. A graphical illustration of the five South African populations as constructed using Genetix v.4.05.2. (Belkhir *et al*, 2002).

Table 4.4 Population pairwise genetic distances among the five populations (F_{ST} values below diagonal) using Arlequin (Excoffier *et al*, 2005) with corresponding *p*-values (above diagonal) (p > 0.05)

Populations	Afrikaner	Indian-Asian	Coloured	Xhosa	Zulu
Afrikaner	-	**0.00000+-0.0000***	**0.00000+-0.0000***	**0.00000+-0.0000***	**0.00000+-0.0000***
Indian-Asian	0.03919	-	**0.00000+-0.0000***	**0.00000+-0.0000***	**0.00000+-0.0000***
Coloured	0.03020	0.02149	-	**0.00000+-0.0000***	**0.00000+-0.0000***
Xhosa	0.09887	0.08573	0.03568	-	0.16504+-0.0123
Zulu	0.11886	0.10909	0.04800	0.00125	-

***p < 0.05**

4.3.2.3 Phylogenetic tree construction

Table 4.5 shows the genetic distance as calculated between the five populations by the program TreeFit. The genetic distance estimates results are similar to the results shown in section 4.3.2.2 with the Zulu population being closer to the Xhosa population (θ = 0.0012) and being the furthest from the Afrikaner (θ = 0.119) and Indian-Asian populations (θ = 0.0992).

The degree of fit degree of fit between observed data and phylogenetic tree construction is summarized in Table 4.6 with an R^2-value of 0.998. The phylogenetic tree constructed using the NJ method is shown in Figure 4.3. The tree illustrates the same interpopulation relationships between the five populations as observed with the FCA method. The Xhosa and Zulu populations are clustered closely together on the tree, an indication of close genetic similarities shared.

Table 4.5 Population comparison of the five populations using genetic distance (F_{ST}) (Weir and Cockerham, 1984) by Treefit.

Populations	Afrikaner	Indian-Asian	Coloured	Xhosa	Zulu
Afrikaner	0.0000				
Indian-Asian	0.0392	0.0000			
Coloured	0.0302	0.0214	0.0000		
Xhosa	0.0992	0.0857	0.0356	0.0000	
Zulu	0.119	0.1089	0.0477	0.0012	0.0000

4.4 Summary

A higher number of heterozygotes are indicative of high genetic diversity within a population whereas a low number of heterozygotes would indicate low genetic diversity. Genetic diversity was high especially in the Afrikaner population. The kit is shown to be highly polymorphic. For all forensic indices the Afrikaner population had the highest values. Surprisingly, compared to the Afrikaner population, lower estimates in genetic diversity and forensic indices were

Table 4.6 Summary of observed genetic distances (Weir and Cockerham's 1984) versus data fitted in the phylogenetic tree (NJ method) as calculated using Treefit.

		Observed D	Fitted D
Afrikaner	**Indian-Asian**	0.0392	0.0392
Afrikaner	**Coloured**	0.0302	0.0309
Afrikaner	**Xhosa**	0.0992	0.0991
Afrikaner	**Zulu**	0.119	0.1176
Indian-Asian	**Coloured**	0.0214	0.0207
Indian-Asian	**Xhosa**	0.0857	0.0889
Indian-Asian	**Zulu**	0.1089	0.1073
Coloured	**Xhosa**	0.0356	0.0325
Coloured	**Zulu**	0.0477	0.0509
Xhosa	**Zulu**	0.0012	0.0012

R^2 = 0.998

D = Degree of Fit

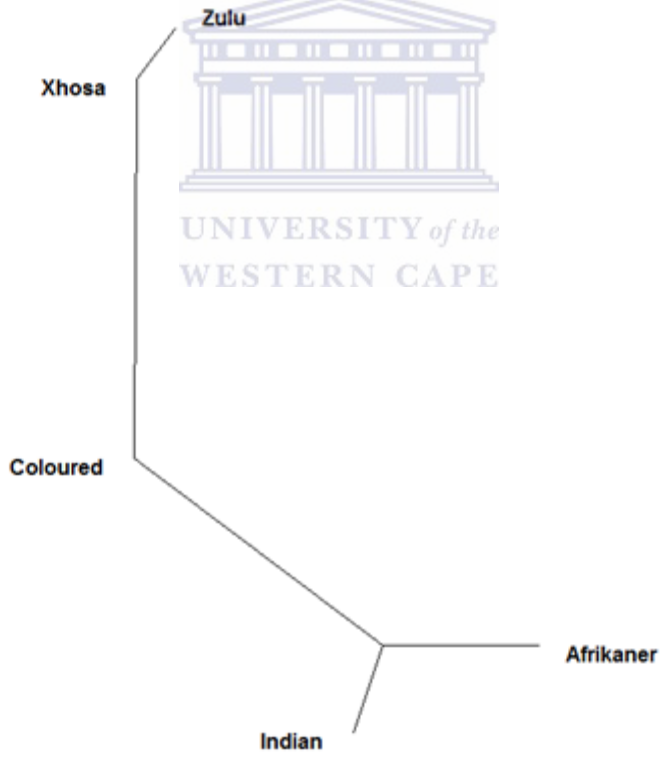


Figure 4.3 Phylogenetic tree construction of the five South African populations. A phylogenetic tree constructed using Treefit (Kalinowski, 2009) and visualised using TreeView (Paper, 2006). The Neighbour-Joining method was used and genetic distances estimated using Weir and Cockerham (1984).

observed in the other four populations. For all forensic parameters, the highest values were estimated in the Afrikaner populations and the lowest values were estimated in the Black populations. This was also observed in other populations using the Investigator kit. In Finnish and Danish populations combined match probability values of 3.54 X 10^{-13} (Neuvonen *et al,* 2012) and 3.3 X 10^{-13} (Friis *et al*, 2012) were estimated, respectively. While in the African American population a combined match probability value of 1.43×10^{-11} (LaRue *et al,* 2012) was estimated. The Mixed Ancestry population is historically and genetically linked to all the other populations. The statistical analysis performed shows evidence of this. It is reflected in the genetic distance measures which show genetic differentiation between the Mixed Ancestry and each of the remaining four populations. FCA and phylogenetic tree construction illustrates the genetic relationship between the Mixed Ancestry population with the remaining four populations, with the Mixed Ancestry population being located centrally on both the phylogenetic tree as well as the scatter plots constructed using FCA. Population comparison also confirmed close genetic links between the Xhosa and Zulu populations. The Mixed Ancestry and Indian-Asian populations also share genetic similarities with the Xhosa and Zulu populations and might also genetically be predisposed to the mutation causing the HLD97 null allele.

The data strongly indicates the presence of a null allele at locus HLD97 in the Mixed Ancestry and Indian-Asian populations i.e. high null allele frequencies and Hardy-Weinberg disequilibrium being detected by statistical analysis. Null alleles negatively affect heterozygosity levels and therefore genetic diversity. The issue of null alleles should be addressed if the presence of null alleles is suspected.

The Afrikaner population shared the least genetic similarity with the four populations. The results for population comparison indicate the presence of population substructure resulting in individuals from the same population sharing more genetic similarities. The implication of subpopulations in forensics is that this can lead to errors when it comes to estimation of population frequencies and therefore correction for population substructure using a tetha value should be considered and implemented (Balding and Nichols, 1994).

Chapter 5

Conclusions and recommendations

The Investigator DIPplex kit was assessed based on its application in forensics, specifically on South African populations as well as looking at qualities which will aid in improving turnaround time, streamlining efficiency and delivering of quality results.

On its performance in the laboratory the Investigator DIPplex kit performed well. As the kit is already optimised, it makes it easy to use and less labour intensive. The low amount of input DNA required makes it ideal for application in forensic work. In terms of laboratory application, the Investigator DIPplex kit basically has no drawbacks. The kit has been shown to be robust and the results reproducible, all qualities that are ideal for streamlining laboratory work and increasing turnaround time. The marker panel consists of thirty indels. PCR amplification and data analysis with the Investigator kit is fast with high output of good quality data in a short amount of time. The kit can be used with older or newer genotyping platforms i.e. the 3100 to 3500xl sequencing instruments. This allows for delivering of results using simple PCR amplification to capillary electrophoresis processing.

Despite these advantages, indels are di-allelic markers. This one characteristic makes the kit more ideal as a supplementary forensic kit than a stand-alone kit as the amount of markers are not sufficient compared to STRs, with the statistical values of the combined match probability being lower than that of the current forensic kit applied by SAPS. SAPS forensics laboratories are in the process of migrating to chemistries which incorporates more loci, as much as fifteen STR loci. However, currently, SAPS forensics laboratories are using the AmpFLSTR Profiler Plus™ PCR Amplification kit (Life Technologies). The kit amplifies nine STR loci namely D3S1358, D5S818, D7S820, D8S1179, D13S317, D18S51, D21S11, FGA, and vWA as well as a tenth gender determining locus, Amelogenin. SAPS forensic laboratories deal with cases of missing persons and

victim identification as well as mass disasters. SAPS forensic laboratories can benefit from making use of this kit to aid in DNA profiling of degraded DNA as STR analysis is not as successful in DNA typing badly degraded DNA, as in the case of mass disasters.

This is the first time Investigator DIPplex kit has been applied on South African populations. The data presented has demonstrated the presence of a null allele in elevated frequency in the Xhosa and Zulu populations at locus HLD97. Null alleles have a negative effect on population genetic diversity estimation by decreasing heterozygosity, as well as in kinship analysis. I recommend that the manufacturers of the kit look at the HLD97 locus and possibly redesigning the primers so that the primer binding sites are located away from the problem area. Another option the manufacturers should consider is designing degenerate primers, taking into consideration the mutation site and these primers can be used from the onset, in addition to the original primer set to prevent the lack of PCR amplification at the HLD97 locus. This will accommodate individuals that possess the polymorphism site and result in successful amplification of all individuals in the South African population.

My recommendation to the users is to exclude the locus containing the null allele from kinship analysis or repeating the PCR using a lower Tm which will result in lower primer binding stringency and amplification of the HLD97 null allele.

The data presented in this study has further indicated the presence of population substructure within the South African population. Considering the kit is designed for application in forensics, this will have an effect on allele frequencies used to calculate match probabilities.

Due to the fact that populations are not homogeneous and allele frequencies vary across populations, there is strong bias based on which allele frequencies are used. By using allele frequencies that are not population specific it can negatively affect the calculated match probabilities, and lead to stronger biased evidence against the suspect. Or in the case of missing persons, it could lead to weak statistical

evidence resulting in unsuccessful victim identification. Therefore allele frequencies need to be sampled from specific populations.

I suggest applying allele frequencies obtained from the specific subpopulation to which a person of interest belong to, to be representative and give accurate estimations of match probabilities which would work in the defence's favour, especially in kinship analysis. I recommend applying a theta value to correct for population substructure which is present within the South African populations. In this study theta values in the range of 0.00116 and 0.11903 were estimated using F_{ST} analysis.

Furthermore, in the event of person of interest's population of origin being unknown, a theta value should be used for calculating match probabilities, by applying allele frequencies from population groups which would best be representative of that individual's population. In cases of victim identification I recommend using the population group that possible relatives of the missing person belong to and making use of allele frequencies sampled from the geographical area that the missing person is from, for kinship analysis.

However, caution should be applied when choosing allele frequencies for F_{ST} analysis i.e. using overall allele frequencies. As allele frequencies between different geographical regions vary, the choice of allele frequencies should be as representative as possible as this will directly affect F_{ST} analysis and the calculated theta value (Toscanini *et al,* 2012). Currently, SAPS Forensic Science Laboratories apply the equations as recommended by the National Research Council (1996) for correction of population substructure by using a theta value of 0.01 for urban populations and 0.03 for rural populations and make use of allele frequencies from the four main population groups in South Africa i.e. Black, Coloured, Indian and White.

Overall, the kit will perform well and will serve a better purpose as a supplementary kit, specifically where other forensic markers are unsuccessful like genotyping of degraded DNA.

References

Act No. 37 of 2013 Criminal Law (Forensic Procedures) Amendment Act, 2013. GOVERNMENT GAZETTE, 27 January 2014.

Alvarez M.F.P.R., Gusmao L., Phillips C., Butler J., Lareu M.V.,Carracedo A., Vallone P.M. (2011) Forensic Performance of Short Amplicon Insertion-Deletion (INDEL) Markers. In: 22nd International Symposium on Human Identification, National Harbor, MD, 5 October 2011

Amorim, A. & Pereira, L. (2005). Pros and cons in the use of SNPs in forensic kinship investigation: a comparative analysis with STRs. Forensic science international, 150(1), 17-21.

Balding, D.J., & Nichols, R.A. (1994). DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands. Forensic Science International, 64(2), 125-140.

Ballantyne, K. N., Goedbloed, M., Fang, R., Schaap, O., Lao, O., Wollstein, A. & Kayser, M. (2010). Mutability of Y-chromosomal microsatellites: rates, characteristics, molecular bases, and forensic implications. The American Journal of Human Genetics, 87(3), 341-353.

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.

Belkhir, K., Borsa, P., Chikhi, L., Raufaste, N. & Catch, F. 1996-2002 GENETIX 4.05, software under Windows TM for the genetics of the populations. Laboratory Genome, Populations, Interactions, CNRS UMR 5000, University of Montpellier II, Montpellier (France).

Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. & Sayers, E.W. (2013). GenBank. Nucleic Acids Res. Jan;41 (Database issue):D36-42.

Bland J.M. & Altman D.G. (1995). Multiple significance tests: the Bonferroni method. BMJ: British Medical Journal, 310(6973), 170.

Brinkmann, B., Klintschar, M., Neuhuber, F., Hühne, J., & Rolf, B. (1998). Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. The American Journal of Human Genetics,62 (6), 1408-1415.

Brock T.D. Freeze H. Thermus aquaticus, a Non sporulating Extreme Thermophile. (1969). J. Bact. vol. 98 (1) pp. 289-297.

Brookfield J.F.Y. (1996). A simple new method for estimating null allele frequency from heterozygote deficiency. Molecular Ecology, 5(3), 453-455.

Budowle B., Moretti T.R., Niezgoda S.J. & Brown B.L. (1998). CODIS and PCR-based short tandem repeat loci: Law enforcement tools; Proceedings — The Second European Symposium on Human Identification; Innsbruck, Austria; pp 73–88.

Budowle, B. & Van Daal, A. (2008). Forensically relevant SNP classes. BioTechniques: The international journal of life science methods, 44(5), 603.

Butler J.M. (2003). Recent developments in Y-single tandem repeat and Y-single nucleotide polymorphism analysis. Forensic Sci Rev 15:91.

Butler J.M., (2005). Forensic DNA Typing: Biology and Technology behind STR Markers, Second Edition. London: Academic Press.

Butler, J.M., Shen, Y. & McCord, B.R. (2003). The Development of Reduced Size STR Amplicons as Tools for Analysis of Degraded DNA* J Forensic Sci, Vol. 48, No. 5.

Butler, J.M. and Reeder, D.J. (2010). STRBase - National Institute of Standards and Technology. [ONLINE] Available at: http://www.cstl.nist.gov/strbase/kits/Identifiler.htm. [Accessed 28 November 2014].

Butler, J.M. and Reeder, D.J. (2010). STRBase - National Institute of Standards and Technology. [ONLINE] Available at: http://www.cstl.nist.gov/strbase/multiplx.htm. [Accessed 24 November 2014].

Butler, J.M. and Reeder, D.J. 2010. STRBase - National Institute of Standards and Technology. [ONLINE] Available at: http://www.cstl.nist.gov/biotech/strbase/mutation.htm. [Accessed 19 November 2014].

Buzzle.com, 2013. buzzle. [ONLINE]. Available at: http://www.buzzle.com/articles/dna-fingerprinting-process.html. [Accessed 11 November 2013].

Carvalho A. & Pinheiro M.F. (2013) Population data of 30 insertion/deletion polymorphisms from a sample taken in the North of Portugal. Int J Legal Med 127:65–67.

Chamberlain, J.S., R.A. Gibbs, J.E. Ranier, P.N. Nguyen & C.T. Caskey. (1988). Deletion screening of the Duchenne muscular dystrophy locus via multiplex DNA amplification. Nucleic Acids Res. 16:11141-11156.

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

Chakraborty R., Andrade M.D., Daiger S.P., & Budowle B. (1992). Apparent heterozygote deficiencies observed in DNA typing data and their implications in forensic applications.Annals of Human Genetics, 56 (1), 45-57.

Chakraborty, R., Jin, L., & Zhong, Y. (1994). Paternity evaluation in cases lacking a mother and nondetectable alleles. International journal of legal medicine,107 (3), 127-131.

Champlot S., Berthelot C., Pruvost M., Bennett E.A., Grange T., *et al*. (2010). An Efficient Multistrategy DNA Decontamination Procedure of PCR Reagents for Hypersensitive PCR Applications. PLoS ONE 5(9): e13042. doi:10.1371/journal.pone.0013042.

Chromas Lite version 2.1 (2012). Technelysium Pty Ltd, South Brisbane, Queensland, Australia.

Clayton, T. M., Whitaker, J. P., Sparkes, R., & Gill, P. (1998). Analysis and interpretation of mixed forensic stains using DNA STR profiling. Forensic Science International, 91(1), 55-70.

Dakin E.E. & Avise, J.C. (2004). Microsatellite null alleles in parentage analysis. Heredity, 93(5), 504-509.

Deforce, D.L., Millecamps, R.E., Van Hoofstat, D. & Van den Eeckhout, E.G. (1998). Comparison of slab gel electrophoresis and capillary electrophoresis for the detection of the fluorescently labeled polymerase chain reaction products of short tandem repeat fragments. Journal of Chromatography A, 806(1), 149-155.

Edwards, A., Civitello, A., Hammond, H.A. & Caskey, C.T. DNA typing and genetic mapping with trimeric and tetrameric tandem repeats. (1991). Am. J. Hum. Genet. 49 746–756.

Edwards, M.C. & Gibbs, R.A. (1994). Multiplex PCR: advantages, development, and applications. Genome Research. 3(4), S65-S75.

Excoffier L., Laval G., & Schneider S. (2005). Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evolutionary bioinformatics online, 1, 47.

Fondevila, M., C. Phillips, C. Santos, R. Pereira, L. Gusmão, A. Carracedo, J. M. Butler, M. V. Lareu, & P. M. Vallone. (2012). Forensic performance of two insertion–deletion marker assays." International Int J Legal Med 126, no. 5: 725-737.

Friis, S. L., Børsting, C., Rockenbauer, E., Poulsen, L., Fredslund, S. F., Tomas, C., & Morling, N. (2012). Typing of 30 insertion/deletions in Danes using the first commercial indel kit—Mentype® DIPplex. Forensic Science International: Genetics, 6(2), e72-e74.

Gill, P. (2001). An assessment of the utility of single nucleotide polymorphisms (SNPs) for forensic purposes. International Journal of Legal Medicine, 114(4-5), 204-210.

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

Goldstein, D.B. & Pollock, D.D. (1997). Launching Microsatellites: A Review of Mutation Processes and Methods of Phylogenetic Inference. Journal of Heredity, 88:335-342.

Goodwin W., Linacre A., & Hadi S. (2011). An introduction to forensic genetics Second Edition. Wiley-Blackwell.

Green gazette. (2006). The government gazette of South Africa. [ONLINE] Available at: http://www.greengazette.co.za/notices/criminal-law-forensic-procedures-amendment-bill-2013-notice-of-intention-to-introduce-the-criminal-law-forensic-procedures-amendment-bill-2013-in-the-national-assembly-and-publication_20130426-GGN-36415-00435.pdf. [Accessed 28 November 2013].

Guichoux, E., Lagache, L., Wagner, S., Chaumeil, P., Le Ger, P., Lepais, O., Lepoittevin, C., Malausa, T., Revardel, E., Salin, F. & Ptit, R.J. (2011). Molecular Ecology Resources 11, 591–611.

Hall, T.A. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT Nucleic Acids Symposium Series, Vol. 41 (1999), pp. 95-98.

Hamilton M. (2009) Population Genetics. First Edition. Wiley-Blackwell.

Hildebrand C.E., Torney D.C., & Wagner R.P. (1992). Informativeness of polymorphic DNA markers. Los Alamos Sci, 20, 100-102.

Innis, M.A., Gelfand, D.H. Sninsky, J.J. & White, T.J. (Eds.). (2012). PCR protocols: a guide to methods and applications. Academic press.

Jeffreys, A .J., Wilson, V. & Thein, S.L. (1985). Hypervariable'minisatellite' regions in human DNA. Nature, 314(6006), 67-73.

Jobling, M.A. & Gill, P. (2004). Encoded evidence: DNA in forensic analysis. Nature Reviews Genetics, 5(10), 739-751.

Jobling, M.A., Pandya, A. & Tyler-Smith, C. (1997). The Y chromosome in forensic analysis and paternity testing. Int J Legal Med 110: 118–124.

Kalinowski S.T. (2009). How well do evolutionary trees describe genetic relationships between populations? Heredity 102:506-513.

Kalinowski, S.T. & Taper, M.L. (2006). Maximum likelihood estimation of the frequency of null alleles at microsatellite loci. Conservation Genetics, 7(6), 991-995.

Kayser, M., Roewer, L., Hedman, M., Henke, L., Henke, J., Brauer, S.,Krüger, C., Michael Krawczak, M., Marion Nagy, M., Dobosz, T., Reinhard Szibor, R., Peter de Knijff, P, Stoneking, M., & Sajantila, A. (2000). Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs. The American Journal of Human Genetics, 66(5), 1580-1588.

Kayser, M. & Sajantila, A. (2001). Mutations at Y-STR loci: implications for paternity testing and forensic analysis. Forensic Science International, 118(2), 116-121.

Kim, E. H., Lee, H. Y., Yang, I. S., Yang, W. I. & Shin, K. J. (2014). Population data for 30 insertion–deletion markers in a Korean population. Int J Legal Med, 128(1), 51-52.

Lahiri, D. and Nurnberger, J. (1991). A rapid non-enzymatic method for the preparation of HMW DNA from blood for RFLP studies. Nucleic Acids Research 19, 5444.

LaRue, B.L., Ge, J., King, J.L., & Budowle, B. (2012). A validation study of the Qiagen Investigator DIPplex® kit; an INDEL-based assay for human identification. International journal of legal medicine, 126(4), 533-540

Lehman, I.R., Bessman, M.J., Simms, E.S. & Kornberg, A. (1958). Enzymatic synthesis of deoxyribonucleic acid. J. biol. Chem, 233, 163-170.

Li, C. T., Zhang, S. H., & Zhao, S. M. (2011). Genetic analysis of 30 InDel markers for forensic use in five different Chinese populations. Genet Mol Res, 10(2), 964-979.

Liang, W., D. Zaumsegel, M. A. Rothschild, M. Lv, L. Zhang, J. Li, F. Liu, J. Xiang & P. M. Schneider. (2013). Genetic data for 30 insertion/deletion polymorphisms in six Chinese populations with Qiagen Investigator DIPplex Kit. Forensic Science International: Genetics Supplement Series, 4(1), e268-e269.

Lorente, J.A., Entrala, C., Alvarez, J.C., Lorente, M., Arce, B., Heinrich, B. & Villanueva, E. (2002). Social benefits of non-criminal genetic databases: Missing persons and human remains identification. International Journal of Legal Medicine, 116(3), 187-190.

Markoulatos, P., Siafakas, N. & Moncany, M. (2002). Multiplex polymerase chain reaction: a practical approach. Journal of clinical laboratory analysis, 16(1), 47-51.

Martın P., Garcıa O., Heinrichs B., Yurrebaso I., Andoni A. & Alonso A. (2012). Population genetic data of 30 autosomal indels in Central Spain and the Basque Country populations Forensic Science International: Genetics 7 e27–e30.

Martin, P.D. (2004, April). National DNA databases-practice and practicability. A forum for discussion. In International Congress Series (Vol. 1261, pp. 1-8). Elsevier.

Medrano J.F., Aasen E.& Sharrow L. (1990). DNA extraction from nucleated red blood cells. Biotechniques. Jan;8(1):43.

Montano, V., Ferri, G., Marcari, V., Batini, C., Anyaele, O., Destro-Bisol, G. & Comas, D. (2011). The Bantu expansion revisited: a new analysis of Y chromosome variation in Central Western Africa. Molecular ecology, 20(13), 2693-2708.

Mullis, K. B. & Faloona, F. A. (1987). Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. Methods Enzymol. 155, 335-350.

Neumann, K. & Wetton, J.H. (1996). Highly polymorphic microsatellites in the house sparrow Passer domesticus. Molecular Ecology, 5(2), 307-309.

Nelson, M. S., Levedakou, E. N., Matthews, J. R., Early, B. E., Freeman, D. A., Kuhn, C. A.,Sprecher, C.J., Ashima S.A., McElfresh, K.C. and Schumm, J.W. (2002). Detection of a Primer-Binding Site Polymorphism for the STR Locus D16S539 Using the PowerPlex® 1.1. System and Validation of a Degenerate Primer to Correct for the Polymorphism. Journal of forensic sciences, 47(2), 345-349.

Neuvonen, A.M., Palo, J.U., Hedman, M. & Sajantila, A. (2012). Discrimination power of Investigator DIPplex loci in Finnish and Somali populations. Forensic Science International: Genetics, 6(4), e99-e102.

Paetkau, D. & Strobeck, C. (1995). The molecular basis and evolutionary history of a microsatellite null allele in bears. Molecular Ecology 4(4), 519-520.

Page, R.D.M. 1996. TREEVIEW: An application to display phylogenetic trees on personal computers. Computer Applications in the Biosciences 12: 357-358.

Pemberton, J.M., Slate, J., Bancroft, D.R. & Barrett, J.A. (1995). Nonamplifying alleles at microsatellite loci: a caution for parentage and population studies. Molecular Ecology, 4(2), 249-252.

Pereira, R, & Gusmão, L. (2012). Capillary electrophoresis of 38 noncoding biallelic mini-Indels for degraded samples and as complementary tool in paternity testing. In DNA Electrophoresis Protocols for Forensic Genetics (pp. 141-157). Humana Press.

Poulsen, L., Farzad, M. S., Børsting, C., Tomas, C., Pereira, V. & Morling, N. (2015). Population and forensic data for three sets of forensic genetic markers in four ethnic groups from Iran: Persians, Lurs, Kurds and Azeris. Forensic Science International: Genetics, 17, 43-46.

Qiagen supplementary material: population data for analysis of results from the Investigator DIPplex kit, Qiagen, 2010. Available at

https://www.qiagen.com/za/resources/resourcedetail?id=4d594b28-026b-49bd-a251-7dafff2831c7&lang=en. Accessed August 2015.
Qiagen, (2011). Invesigator DIPplex handbook.

Rajeevan H, Osier MV, Cheung KH, Deng H, Druskin L, Heinzen R, Kidd JR, Stein S, Pakstis AJ, Tosches NP, Yeh CC, Miller PL, Kidd KK. (2003). ALFRED – the ALlele FREquency Database – update. Nucleic Acids Research. 31(1):270-271.

Rychlik, W. (2007) OLIGO 7 Primer Analysis Software, Methods in Molecular Biology Vol. 402: PCR Primer Design; Ed. A. Yuryev; Humana Press Inc., Totowa, NJ. pp. 35-59.

Saiki, R.K., Scharf, S., Faloona, F., Mullis, K.B., Horn, G.T., Erlich, H.A. & Arnheim, N. (1985). Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. Science, 230(4732), 1350-1354.

Sanchez, J.J., Phillips, C., Børsting, C., Balogh, K., Bogus, M., Fondevila, M. & Morling, N. (2006). A multiplex assay with 52 single nucleotide polymorphisms for human identification. Electrophoresis, 27(9), 1713-1724.

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

Sobrino, B., Brión, M., & Carracedo, A. (2005). SNPs in forensic genetics: a review on SNP typing methodologies. Forensic science international, 154(2), 181-194.

Statistics South Africa, Census 2011 Statistical Release P0301.4, Pretoria 2012

Sundquist T. and Bessetti J. (2005). Identifying and Preventing DNA Contamination in a DNA-Typing Laboratory. Promega Corporation www.promega.com

Tereba A. (1999). Tools for analysis of population statistics. Profiles in DNA, vol. 2, (www.Promega.com)

The evaluation of Forensic DNA Evidence, Committee on DNA Forensic Science, National Research Council 1996, National Aademy Press.

Toscanini, U., Garcia-Magariños, M., Berardi, G., Egeland, T., Raimondi, E. & Salas, A. (2012). Evaluating methods to correct for population stratification when estimating paternity indexes. PloS one, 7(11), e49832.

Urquhart, A., Chiu, C.T., Clayton, T.M., Downes, T., Frazier, R.R.E., Jones, S., Kimpton, C.P., Lareu, M.V., Millican, E.S., Oldroyd, N.J., Thompson, C., Watson, S., Whitaker, J.P. and Gill, P. (1995) Multiplex STR systems with fluorescent detection as human identification markers. Proceedings from the 5th International Symposium on Human Identification (1994). pp. 73-83.

Watson J.D., Crick F.H.C. A Structure for Deoxyribose Nucleic Acid. Nature. Vol. 171, pp. 737-738 (1953)

Weber, J.L., David, D., Heil, J., Fan, Y., Zhao, C., & Marth, G. (2002). Human diallelic insertion/deletion polymorphisms. The American Journal of Human Genetics, 71 (4), 854-862.

Weir, B.S. & Cockerham, C.C. (1984). Estimating F-statistics for the analysis of population structure. evolution, 1358-1370

Wiegand, P. & Kleiber, M. (2001). Less is more–length reduction of STR amplicons using redesigned primers. International journal of legal medicine, 114(4-5), 285-287.

Wikipedia.org. 2014. Wikipedia The Free Encyclopedia . [ONLINE] Available at: http://en.wikipedia.org/wiki/Ethnic_groups_in_South_Africa. [Accessed 19 November 14].

Willuweit, S., Roewer, L. & International Forensic Y Chromosome User Group. (2007). Y chromosome haplotype reference database (YHRD): update. Forensic Science International: Genetics, 1(2), 83-87.

Zhong, Y., Budowle, B. & Center, FBI (1999). The utility of short tandem repeat loci beyond human identification: implications for development of new DNA typing systems. Electrophoresis, 20, 1682-1696.

Internet Resources

ALFRED database	http://alfred.med.yale.edu/alfred/
BioEdit software	http://www.mbio.ncsu.edu/bioedit/bioedit.html
Brand South Africa	http://www.southafrica.info
Chromas Lite	http://www.technelysium.com.au
European DNA Profiling Group (EDNAP)	www.isfg.org/EDNAP
European Network Of Forensic Science Insitutes	www.enfsi.eu
Federal Bureau of Investigation	http://www.fbi.gov
FinchTV software	www.geospiza.com/finchtv
Genbank.	http://www.ncbi.nlm.nih.gov/genbank
Guidelines as recommended by SWGDAM	http://swgdam.org/docs.html
Interpol	http://www.interpol.int/en

International Society for Forensic Genetics (ISFG)	www.isfg.org
Macrogen Europe (Netherlands)	http://dna.macrogen.com/
NCBI website	http://www.ncbi.nlm.nih.gov/
Oligo Primer Analysis Software	www.oligo.net
Scientific Working Group on DNA Analysis Methods (SWGDAM)	www.swgdam.org
South African National Accreditation System	http://www.home.sanas.co.za/
South African Police Service	www.saps.gov.za
Statistics South Africa	www.statssa.gov.za
United States Y-STR database	http://www.usystrdatabase.org
Y-STR Haplotype Reference Database	www.yhrd.org
Parliament of the Republic of South Africa	www.parliament.gov.za

Appendix

1 DNA Extraction, Quantification and Working Stock Solutions

1.1 Lahiri and Nurnberger (1991) DNA Extraction Protocol

1.1.1 Reagents (To be autoclaved immediately after preparation)

TKM I (100 ml)

1M Tris, pH 8.0 (Merck Laboratory Supplies) 1 ml

100 mM KCl (Merck Laboratory Supplies) 10 ml

200 mM $MgCl_2$ (Merck Laboratory Supplies) 5 ml

100 mM EDTA (Merck Laboratory Supplies) 2 ml

Distilled H_2O 82 ml

TKM I + Nonidet P40 (100 ml)

1 M Tris, pH 8.0 (Merck Laboratory Supplies) 1 ml

100 mM KCl (Merck Laboratory Supplies) 10 ml

200 mM $MgCl_2$ (Merck Laboratory Supplies) 5 ml

100 mM EDTA (Merck Laboratory Supplies) 2 ml

Distilled H_2O 82 ml

Nonidet P40 (Sigma) 2.25 ml

TKM II (100 ml)

1 M Tris, pH 8.0 (Merck Laboratory Supplies) 1 ml

100 mM KCl (Merck Laboratory Supplies) 10 ml

200 mM $MgCl_2$ (Merck Laboratory Supplies) 5 ml

100 mM EDTA (Merck Laboratory Supplies) 2 ml

2 M NaCl (Merck Laboratory Supplies) 20 ml

Distilled H_2O 62 ml

10 % w/v SDS (100 ml)

SDS (*Merck Laboratory Supplies*) 10 g

Distilled H_2O 100 ml

2 M NaCl (100 ml)

NaCl (*Merck Laboratory Supplies*) 11.67 g

Distilled H_2O 100 ml

1.1.2 DNA Extraction steps

1. 0.5 ml of carefully mixed whole blood was transferred to a clean, dry, APPROPRIATELY LABELLED 1.5ml microfuge tube.

2. 0.5 ml of sterile [TKM I + Nonidet P-40] was added to the 0.5 ml blood and the contents of the tube mixed gently by inversion.

3. The tube was then placed in a bench-top centrifuge (*Eppendorf, 5415 D*) and the mixture centrifuged at 5000 rpm for 10 minutes to pellet the nuclei.

4. The supernatant was then carefully removed, paying attention to not disturb the pellet.

5. The pellet was then washed by adding 0.5 ml sterile TKM I to it.

6. The tube was centrifuged for 10 minutes at 5000 rpm.

7. The supernatant was removed again. This time, extra care was taken to remove as much of it as possible, without disturbing the pellet.

8. 70 µL of sterile TKM II was added to the pellets and tube vortexed (*Stuart Scientific, vortex mixer SA3*) until the pellet was completely re-suspended in the liquid.

9. 4.37 µL of sterile 10 % w/v SDS was added to this mix and the tube vortexed briefly.

10. The tube was then incubated in a water-bath (*Memmert*) at 55 ^{o}C for 10 minutes.

11. After incubation, 264 ul of sterile 2 M NaCl was added and the tube vortexed briefly.

12. The tube was then centrifuged at 13000 rpm for five minutes to pellet extra-cellular components.

13. After centrifugation, the supernatant was transferred to a clean, dry and CORRECTLY LABELLED, 1.5ml microfuge tube.

14. To this supernatant, 677 µL of absolute ethanol (room temperature) was added and the contents of the tube mixed gently by inverting it a few times. At this point the DNA became visible.

15. The tube was then centrifuged at 13000 rpm for five minutes to pellet the DNA. (The DNA did not always end up at the bottom of the tube, but sometimes got stuck on the side of the tube).

16. The supernatant was then removed and the tube centrifuged again at 13000 rpm.

17. The supernatant was removed after centrifugation and 250 µL of ice-cold 70 % ethanol added to wash the DNA.

18. The tube was then centrifuged at 13000 rpm for five minutes and the supernatant removed.

19. Step 17 and 18 were repeated.

21. The DNA was dried at room temperature, re-suspended in 100 µL of sterile distilled water and the DNA allowed to go into solution at room temperature overnight.

All the blood waste from the DNA extraction process were decanted into appropriate containers, sealed properly, clearly labelled as blood waste and discarded in an appropriate manner by Waste Tech, an accredited waste removal company.

2. DNA Extraction, Quantification and Working Stock Solutions from buccal swabs

2.1 Proteinase K and Salting Lysis method (Medrano, 1990).

2.1.1. Preparation of the lysis buffer:

1. Proteinase K stock is 20mg/ml and stored at -20 ^{o}C.

2. Add 400 mM NaCl [2 M], 10 mM Tris-ClH pH8 [1 M] and 2 mM EDTA [0.5 M] in 80 % of their final volume, in distilled water.

3. Add 1 % SDS [MW = 288.4]. Leave in the oven at 60 ^{o}C until fully dissolved. Transfer to a volumetric flask and add the required volume of SABAX water to get 100 % desired volume.

2.1.2. DNA sample collection using buccal swabs

1. Cut off the surface of the swab by using a clean scalpel or surgery blade, working on a clean surface.

2. Prepare 1.5 ml or 2 ml Eppendorf tubes and add a volume of lysis buffer and proteinase K. The final working solution of proteinase K should be 0.1mg/ml.

3. Add pieces of the cut swab to the Eppendorf tubes and vortex for 30 seconds. Incubate overnight at 56 ^{o}C.

4. Transfer the total volume to a clean Eppendorf tube.

5. To recover the lysis solution and biological material trapped in the swab pieces the following steps can be taken: perforate the end of a 0.5 ml tube with a needle. Place the perforated tube inside a 1.5 ml Eppendorf tube. Spin for 1 minute in a microcentrifuge and add the collected volume to the previously collected lysis material.

6. Add 1/3 volume of 6 M NaCl and precipitate by shaking the tube(s) vigorously for 15 seconds.

7. Centrifuge the tube(s) for 15 minutes at 5000rpm and transfer the supernatant containing the DNA to another tube.

8. Add an equal volume of cold isopropanol to the tube(s). Leave to stand overnight at -2 ^{o}C or at -80 ^{o}C for 30 minutes.

9. Centrifuge at 14000rpm for 30 minutes to pellet the DNA. Remove salts by washing the pellet with 100uL of 70 % ethanol.

10. Repeat the centrifugation at 14000rpm for 8 minutes to 30 minutes. Dry the pellet briefly in a SpeedyVac or at 65 ^{o}C. Take care to keep the drying process short as over drying will prevent subsequent dissolving in 50uL SABAX water. Store at -20 ^{o}C.

3. Positive control sample for PCR amplification

Control DNA XY5 (Qiagen) [2.0 ng/ul]

3.1 Dilution of positive control

A working stock solution was prepared by diluting the positive control from 2 ng to 0.35 ng by adding 3.5 µL positive control to 16.5 µL SABAX water. The stock solution mixed by pipette and stored at -20 ^{o}C.

Preparation of dNTPs mix

-Deoxynuceotide Triphosphate Set PCR grade (Roche Diagnostics)

Mix each tube of dNTP well by vortexing. Take 10 µL of each dNTP and add to 98 µL SABAX water. Mix well by vortexing and spin down.

4. Gel electrophoresis and working stock solutions

4.1 Reagents

10X TBE Buffer 1 L

Tris-HCl 108 g

Boric acid 55 g

0.5 M EDTA 40ml

dH_2O 1 L

4.2.1 Preparation of a 1 % agrose gel

Add 1.0 g SeaKem® LE agarose (WhiteSci) to 50 ml 1 X TBE buffer. Mix the solution and heat in a microwave, mixing the solution until the agarose gel is dissolved. Take care not to stir too vigorously as to have bubbles forming. Taking care for bubble formation, pour the agarose mixture in a 100 ml loading tray and leave to set.

3. DNA sequencing

3.1 Genbank sequence of HLD97

>gi|224589804:31328000-31328884 Homo sapiens chromosome 13, GRCh37.p10 Primary Assembly

GGAGCAGAATCATTCAGATGGTATAACATAAGGAAAAACTTTGCCCAAGGCAAATCGTGATTGT
GACAGCTTTGTGATTTTTAGAGAATAGCATGGGCCAGGCACAGTGGCTCATGCCTGTAATCCCA
GCACTTTGGGAGGCCGAGGCAGGCAGGTCACTTGAGGTTGGGAGTTCGACAACAGCCTGACCA
ACATGGAGAAACCCTGTCTCTACTAAAAATACAAAATTAGCTGGGCGTGGTGGTGCATGCCTGT
AATGCCAGCTACTCGGGAGGCTGAGGCAGGAGAATCACTTAAACCTGGGAGGCGGAGGTTGCG
GTGAACCAAGATAGCACCATTGCACTCCAGCCTGGGCAACAAGAGTGAAACTCCGTCTCAAAA
AGAGTTCACAGTTTCTCTTTTGCTTTGATTTTCTTATCTGCCGGATAACAATAGTATTTTGGAAGG
CAGGAGGAATTGTGGAAAGAAATGGGTTTTGGGGAGTGGCTGATTGGAGGCAAATCCAAGGAC
ACTCATTGCTGGTGTGTGACTCCAGGCAGTTACTCAGCTTTTCCAAGCCTCAGTTTCCTTATTGTA
AAACAGGACCATGGTCTAGCTAGTAGCATTCCTATGGTGAGTGAAATAATATGTATAAAGCTCC
TGACACAGTGCTTGGCATATATCAGATTGAGCCATGTAAAACTGCCAATATCTGGCTATTTATGA
CCTACAAAAATAGCATTTCATATGATTCCACCTAACATCTGAAGCGCAATAAATGTTATTATTGA
TAATGCAGGTGGTGGTGATAAAGTTTTGAAATCAGAAAGACCTGGCTTCAAATTCCACGCCTTC
ACTGGCCTGACTTATTTTCATTCATTTGACAAATATTATTTTGAACACCCC