Identification of novel SNPSTRs by 454 sequencing in Nguni and Sotho-Tswana Populations

---

A thesis submitted in the fulfilment of the requirements for the degree of

MASTER OF SCIENCE IN BIOCHEMISTRY

Rhodes University

by

Jo-Anne Elizabeth Laurence

February 2015

ABSTRACT

DNA profiling is currently performed by analysis of the electropherogram that results following the amplification of a panel of Short Tandem Repeat (STR) loci. A need has arisen, however, for the development of a typing method that generates results which are compatible and comparable with existing databases, but that have a higher discrimination power by supplying sequence data as well as repeat-number data. Recent studies that explore these alternative typing methodologies have revealed the existence of a number of STR variants. There is, however, little information about the exact nature and prevalence of these sub-alleles. There have also been limited population studies of the genetic profiles of sub-Saharan African populations, despite the fact that evidence suggests that there is greater genetic structure and genetic diversity in these populations. In this study, a processing protocol for the generation of 454 sequencing-ready amplicons of vWA, D2S441, D3S1358, D13S317, D21S11 and D7S820 loci was developed. This protocol was applied to buccal swabs collected from 144 individuals of the Nguni and Sotho-Tswana population groups. A total of 145 485 reads were obtained from the sequencing of these amplicons, of which 97 400 and 48 085 reads were obtained for the Nguni and Sotho-Tswana populations respectively. The proportional representation for each locus ranged from 8-20%, and the allele calls and observed frequencies of these alleles suggested a high degree of relatedness between population groups. The sequencing results, furthermore, enabled the identification of a number of previously undescribed STR variants and SNPSTRs; with allele 13´ for D13S317 representing a SNP that may be predictive of Nguni-ancestry. The results also demonstrated the usefulness of next generation sequencing for increasing the number of discernible alleles for STR profiling.

DECLARATIONS

I declare that this thesis is my own, unaided work. It is being submitted for the degree of Master of Science of Rhodes University. It has not been submitted previously for any degree or examination at any other university.

The financial assistance from Rhodes University Prestigious Scholarship towards the research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to Rhodes University.

The financial assistance of the Sandisa Imbewu grant from Rhodes University towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to Rhodes University.

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

Jo-Anne Elizabeth Laurence
10 February 2015

ACKNOWLEDGEMENTS

A thesis is the culmination of the work shared by a community, presented by an individual.

In light of this, I would like to say a massive THANK YOU to the following:

Firstly to my Dad and my Uncle John who made sure I would see it through the initial phases of higher education

The National Research Foundation for awarding me an Innovation Master's Scholarship; and Rhodes University for awarding me the Henderson Prestigious Scholarship, sending me on various conferences in 2015, and for supporting the research itself through the Sandisa Imbewu grant

Thank you too to my supervisors, Doctors Adrienne Edkins and Brendan Wilhelmi; who were always ready to help me

And a special thank you to Dr Edkins for the amazing speed at which she turned around thesis drafts

I am immensely grateful to those who ensured that I was able to obtain and process my sequencing results:

Dr Gwynneth Matcher, especially for being so helpful and patient with my barrage of email enquiries

Mr David Penkler for providing intellectual support with regards to understanding the potential and utility of bioinformatics

Mr Jeremy Baxter for the script which he kindly wrote for me to enable data sorting

Garry Jevons who assisted with creating the Microsoft Access data capturing interface

And, of course, the 144 individuals who kindly donated samples of their DNA, without whom there would little else left to say.

As crucial to the financial and intellectual support mentioned above, was the support given to me by my wonderful family, loving and patient boyfriend, and supportive friends.

A special thank you to Dave for his unconditional love, support, and for keeping me sane and smiling.

To my favourite sister, marvellous mom and ever-missed dad, who have been the most encouraging, motivating and loving family anyone could ask for, thank you!

To my friends and extended family, as well as Carol and the girls, thank you for always making me smile

Finally thank you God for the talents, opportunities and support that you have given me; and for helping me persevere when I thought I could not keep on

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1: *Introduction*

CHAPTER 2: *Protocol Development*

CHAPTER 3: *SNPSTR Discovery and Analysis*

LIST OF TABLES

CHAPTER 1: *Literature Review and Introduction*

CHAPTER 2: *Protocol Development*

CHAPTER 3: *SNPSTR Discovery and Analysis*

LIST OF ABBREVIATIONS

bp	base pairs
CE-FD	Capillary Electrophoresis-Fluorescent Detection/Detector
CODIS	Combined DNA Indexing System
DNA	Deoxyribonucleic Acid
ESS	European Standard Set
EVC	Externally Visible Characteristic
FD	Fluorescent Detector/Detection
FDP	Forensic DNA Phenotyping
GWA	Genome Wide Association
HWE	Hardy-Weinberg Equilibrium
LCN	Low Copy Number
LD	Linkage Disequilibrium
mtDNA	Mitochondrial DNA
NDNAD	National DNA Database
NFDD	National Forensic DNA Database of South Africa
NGS	Next Generation Sequencing
PAGE	Polyacrylamide Gel Electrophoresis
PBS	Primer Binding Site
PCR	Polymerase Chain Reaction
PD	Power of Discrimination
PE	Power of Exclusion
PI	Paternity Index
PM/P_I	Match Probability/Probability of identity
RFLP	Restriction Fragment Length Polymorphism
SGS/TGS	Second/Third Generation Sequencing
SLP	Single-Locus Probes
SNP	Single Nucleotide Polymorphism
ST	Sotho-Tswana
STR	Short Tandem Repeat
TR	Tandem Repeat

CHAPTER 1
Literature Review and Introduction

Sherlock Holmes said, "It has long been an axiom of mine that the little things are infinitely the most important" and this could not be truer than for DNA, the material that holds the code to life. Genomic material is comprised of conserved regions of DNA which are retained within species, as well as variable regions which can be used to differentiate individuals within a species (Jobling & Gill, 2004). Both of these regions have been exploited by forensic scientists for the purpose of identifying the source of biological material (Jobling & Gill, 2004). The variable regions in particular have been extensively used for the creation of DNA profiles to enable individuation from samples (Morling, 2004). These regions have also been used extensively for familial testing and, in more recent times, to investigate the evolution of humans, map human migration and to determine the biogeographic ancestry of individuals (Kayser and De Knijff, 2011).

ROUTINE DNA PROFILING

1.1 A brief history to autosomal DNA profiling

DNA profiling of autosomal DNA relies on the presence of polymorphic regions of DNA which occupy specific chromosomal locations, are flanked by conserved loci-specific (CLS) sequences, and whose inheritance follow a Mendelian pattern of inheritance (Society for Forensic Haemogenetics, 1989; International Society for Forensic Haemogenetics,1992). There are two main types of genomic polymorphisms that are exploited for DNA profiling; namely Tandem Repeats (TRs) and Single Nucleotide Polymorphisms (SNPs) (Kayser & de Knijff, 2011), schematics of which are displayed in **Figure 1**. TRs are characterised by variability in length, while SNPs characterised by a substitution mutation at a single conserved locus (Butler *et al.*, 2007; Jobling and Gill, 2004)

The first DNA profiles made use of polymorphic TRs called *minisatellites* (Jeffreys *et al.*, 1985). These hypervariable regions are comprised of 10-1000 tandemly repeated sequences of 10-100 bp, and were detected by restriction fragment length polymorphism (RFLP) analysis by Southern Blot, using a single probe (Jeffreys *et al.*, 1985). Later, single-locus

probes (SLPs) were developed for, typically, four loci to enable simplification of profile interpretation (Jobling & Gill, 2004).

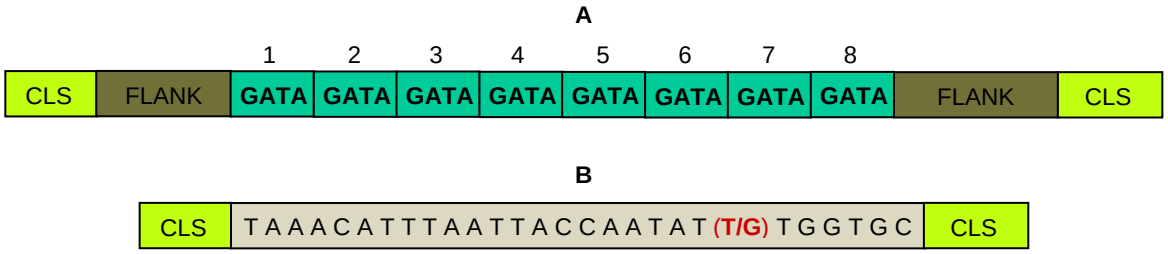


Figure 1: Schematic of polymorphisms used for DNA profiling

A depicts a Tandem Repeat, specifically a simple microsatellite (Short Tandem Repeat) with the 4 base-pair repeat unit GATA and a repeat number of 8. Microsatellites are usually comprised of 2-6 bp repeat units and repeat numbers of <50. Minisatellites are comprised of 10-100 bp repeat units with repeat numbers of 100-1000. Forensically relevant tandem repeats are flanked by conserved loci-specific (CLS) regions of DNA which are used for DNA profiling. A flanking region (FLANK) between the CLS region and the repeat units may be present. An example of a typical Single Nucleotide Polymorphism (T/G) is presented in **B**. SNPs used for DNA profiling are also flanked by CLS regions. CLS regions are used for primer design in PCR-based DNA profiling techniques.

Shortly after the development of RFLP-based analysis, Polymerase Chain Reaction (PCR) based techniques were developed; the first of which relied on the analysis of SNPs within a single polymorphic gene, HLA-DQA1 (Helmuth *et al.*, 1990). PCR allowed for profiles to be generated from less DNA template, however, because SNPs are usually biallelic, this technique posed difficulties for mixture interpretation and also had a far lower discrimination power to that of SLP analysis (Jobling & Gill, 2004). Consequently this PCR-based technique was usually used in combination with RFLP-based techniques (Jobling & Gill, 2004).

In the early 1990s the potential of microsatellites or Short Tandem Repeats (STRs) for human identity testing was identified (Edwards *et al.*, 1991). These markers are similar to minisatellites in that they are comprised of hypervariable tandemly repeated sequences; however the sizes and repeat numbers are far lower, being 2-6 bp and <50 respectively (Ellegren, 2004). The discovery of STRs for DNA profiling enabled the benefits of PCR-based techniques for typing lower quantities of DNA, and the high discrimination power of RFLP analysis to be combined. Consequently, STR analysis has been used as the standard method for DNA profiling for over 20 years (Butler & Hill, 2012).

1.2 Short Tandem Repeats

The human genome contains over a million STRs of which there are three different types; namely simple STRs which are comprised of single repeating units (illustrated in **Figure 1**), compound STRs which are comprised of 2 or more adjacent simple repeats, and complex STRs which contain one or more simple repeat unit interspersed with variable sequences (Butler, 2005; Collins *et al.*, 2003). These STRs may also contain partial repeat units, termed microvariants (Butler, 2005). The high degree of variability observed within STRs is a consequence of their high mutation rates, which result in frequent insertions, deletions and substitutions (Butler & Hill, 2012). The variability is retained within the genome because STRs are usually located within non-coding regions of the genome and, therefore, have no direct phenotypic consequence (Butler, 2006).

1.2.1 Selection of core STRs

The first STR multiplex (single reaction tube PCR that targets a number of loci) that was used for DNA profiling was developed by the Forensic Science Service (FSS) of the U.K. and contained primers specific to the four loci TH01, vWA, FES/FPS and F13A1 (Kimpton *et al.*, 1994). Shortly after its development, a second generation multiplex (SGM) was adopted which targeted the loci TH01, vWA, FGA, D8S1179, D18S51 and D21S11, outlined in **Table 1** (Kimpton *et al.*, 1996). The success of this multiplex for individuation spurred forth the development of a number of commercially produced STR multiplexes (Butler, 2006), all of which included an additional marker for sex-typing, Amelogenin. These multiplex kits were tested and validated in a collaborative effort, undertaken by laboratories across the USA, which culminated in the announcement of the USA core loci, the COmbined DNA Indexing System (CODIS) (Budowle *et al.*, 1998).

The STRs identified in these initial studies have been largely retained in the core loci of various countries today. In particular, the loci of the SGM, indicated in **Table 1** by the red boxes, comprise the bulk of the recommended Interpol Standard Set of Loci (ISSA) (Morling, 2004). The SGM has since been expanded to the SGM Plus which is the set of loci used in the UK. Germany and the Republic of South Africa (RSA) have adopted core loci from the kits which they routinely employ for DNA profiling; consequently, Germany is the only country to include SE33 as a core locus and the core loci of RSA are precisely those targeted in the kit AmpFℓSTR® Profiler Plus™, developed in 1996 (de Wet *et al.*, 2011). It should be noted however, that as of this year RSA will be expanding their core loci by the adoption of a

kit developed in 2001 the, AmpFℓSTR® Identifiler Plus™ (Heathfield, 2014). AmpFℓSTR® Identifiler Plus™ is comprised of all the CODIS loci plus D2S1338 and D19S433 (core loci of the UK).

Table 1: Genetic markers used for DNA profiling by Interpol and various countries

Loci	Interpol	USA	EU	U.K.	Germany	RSA
D1S1356		E	E			
D2S441		E	E			
D2S1338		E	K	C		E
TPOX		C (R)				E
D3S1358	C	C	C	C	C	C
FGA (4)	C	C	C	C	C	C
D5S818		C				C
CF1PO (5)		C				E
SE33		R	K		C	
D7S820		C				C
D8S1179	C	C	C	C	C	C
D10S1248		E	E			
TH01 (11)	C	C	C	C	C	E
D12S391		E	E			
vWA (12)	C	C	C	C	C	C
D13S317		C				C
Penta E (15)		K				
D16S539		C	K	C		E
D18S51	C	C	C	C	C	C
D19S433		E	K	C		E
D21S11	C	C	C	C	C	C
PentaD (21)		R				
D22S1045		R	E			
Amelogenin	C	E	K	C	C	C
DYS391		E				

Core (C) loci are the STRs required for DNA profiling and are country specific or, in the case of the Interpol Standard Set of Loci, are the STRs recommended to countries to enable international collaboration; The USA and EU have recently expanded their panels, CODIS and ESS respectively, to include recently identified STRs and the Republic of South Africa (RSA) is also in the process of expanding its core loci. These additional loci (E) will become part of the core (C) loci in due course. Some STRs (K) are included in national DNA databases due to their presence in STR multiplexing kits. The USA also recommends some STRs (R). TPOX in particular is notable because its status has been changed from core locus to recommended locus in the extended CODIS. Red boxes indicate STRs of the original Second Generation Multiplex (SGM). Data obtained from the STRbase (www.cstl.nist.gov/strbase).

1.2.2 Extension of core STRs

The completion of the Human Genome Project (HGP) by the International Human Genome Sequencing Consortium (2004) resulted in the emergence of a number of investigations

aimed at the identification of 'ideal' STRs for DNA profiling. One such study was performed in 2009 by the European Network of Forensic Science Institute (ENFSI). The ENFSI aimed to identify miniSTRs, STRs of 50-150 bp which are useful in the typing of degraded and low copy number (LCN) DNA (Coble & Butler, 2005); STRs with a low frequency of PCR stutter, a PCR artefact which results when the polymerase 'slips' during replication, resulting in a PCR product one repeat-unit shorter than the true allele (Bacher & Schumm, 1998); and STRs with lower Probability of Identity (P_I) values, resulting in profiles that enable more accurate individuation (Hill and Butler, 2012). This study lead to the expansion of the European Union's core loci, the European Standard Set (ESS), to include the loci D12S391, D1S1656, D2S441, D10S1248, and D22S1045, of which D2S441, D10S1248 and D22S1045 are miniSTRs (Hill & Butler, 2012). In 2011 the FBI Laboratory followed suit and recommended an expansion of CODIS to include loci which enabled greater international collaboration and increased the discrimination power of the profiles (Hares, 2012). Included in this expanded set of loci was the Y-chromosome STR DYS391 which serves to confirm the findings indicated by Amelogenin typing (Hill and Butler, 2012).

While profiles are being expanded to include novel STRs, it is unlikely that existing panels will be completely replaced by 'ideal' STRs since extensive population data, important to the determination of allele frequencies necessary for the calculation of statistical parameters required in court proceedings (discussed in *1.3.2* and *1.3.3*), do not exist for these STRs (Gill *et al.*, 2006). Furthermore, millions of profiles are stored in intelligence DNA databases across the globe (discussed in *1.3.4*) and to replace loci would be to lose that wealth of data (Gill *et al.*, 2006). The locus TPOX is a notable exception which is being altered from a core locus to a recommended locus in the USA due to its low P_I value, presented in **Table 2.** Conversely, it is being added to the core loci of RSA due to its inclusion in the Identifiler kit (Meintjies-van der Walt, 2011).

1.2.3 Characteristics of STRs used in DNA profiling

Table 2 serves to characterise the STRs frequently used for DNA profiling. It is important to note that while a DNA profile with triallelic patterns for D18S51 or D21S11 might suggest a chromosomal abnormality like trisomy-18 (Edward syndrome) or trisomy-21 (Down syndrome), all the STRs used routinely in DNA profiling are found in the introns of chromosomes and have no known direct linkage to disease causing genes (Butler, 2006). The STRs too are found on separate chromosomes and in the rare case that two STRs are located on the same chromosome, are spaced sufficiently far apart so as to ensure independent

segregation during meiosis, as is the case in, for example, CSF1PO and D5S818 or PentaD and D21S11 (Ardlie *et al.*, 2002). This characteristic ensures that STRs are inherited independently allowing for the application of the Product Rule in calculating the overall P_I of a panel of markers (Hill & Butler, 2012).

Table 2: Characteristics of core STRs used for DNA profiling

Loci	Chromosomal Location	Repeat Unit	P_I	Mutation Rate	Triallelic Patterns	Variants
D1S1656	1q42	TAGA	0.0224	-	None	9
D2S441	2p14	TCTA	0.0845	-	None	6
D2S1338	2q35	TGCC/TTCC	0.0220	0.12%	7	28
TPOX	2q25.3	AATG	0.1358	0.01%	18	23
D3S1358	3q21.31	TAGA/CAGA	0.0915	0.12%	11	30
FGA	4q28	CTTT	0.0308	0.28%	37	114
D5S818	5q23.2	AGAT	0.1104	0.11%	8	20
CSF1PO	5q33.1	AGAT	0.1054	0.16%	8	22
SE33	6q14	AAAG complex	0.0066	0.64%	3	25
D7S820	7q21.11	GATA	0.0726	0.10%	19	26
D8S1179	8q24.13	TCTA	0.0558	0.14%	19	24
D10S1248	10q26.3	GGAA	0.0845	-	1	0
TH01	11p15.5	TCAT	0.0766	0.01%	4	22
D12S391	12p13.2	AGAT/AGAC	0.0271	-	2	14
vWA	12p13.31	TCTA/TCTG/TCCA	0.0611	0.16%	24	20
D13S317	13q31.1	TATC	0.0765	0.14%	15	18
Penta E	15q26.2	AAAGA	0.0147	0.16%	15	36
D16S539	16q24.1	GATA	0.0749	0.11%	12	22
D18S51	18q21.33	AGAA	0.0258	0.22%	37	51
D19S433	19q12	AAGG/TAGG	0.0559	0.11%	10	33
D21S11	21q21.1	TCTA complex	0.0403	0.19%	24	42
PentaD	21q22.3	AAAGA	0.0382	0.14%	12	38
D22S1045	22q12.3	ATT	0.0921	-	None	0
DYS391	Y	TCTA	-	0.28%	3	0

Probability of Identity (P_I) is equal to the sum of the genotype frequencies squared and were calculated from results obtained from the analysis of 1036 unrelated Americans: 361 Caucasians, 342 African Americans, 236 Hispanics and 97 Asians (Hill and Butler, 2012). Other data was extracted from the STRbase (www.cstl.nist.gov/strbase).

Also indicated in **Table 2** is the fact that most loci employed in DNA profiling are simple tetranucleotide STRs with the only complex STRs used being D21S11 and SE33. Tetranucleotide loci are favourable because they are highly polymorphic and display heterozygousity values of >0.9 (Walsh *et al.*, 1996). PENTA D and PENTA E are notable as the only pentanucleotide loci, identified by Promega for their high variability and low

propensity for PCR stutter (Bacher & Schumm, 2001; Butler, 2006). The STRs exhibit a range in P_I values (obtained from a study performed on 1036 unrelated Americans belonging to a range of population groups by Hill and Butler, 2012), mutation rates and number of observed variants and triallelic patterns (obtained from STRbase, an online database of STR-related resources which can be accessed on www.ncbi.nlm.nih.gov). The reported mutation rates are based on the 2003 Annual Report of the American Association of Blood Banks and no data are, therefore, available for the STRs defined in 2009. Loci with lower mutation rates and higher P_I values are suited to kinship analysis while those with high mutation rates and low P_I values are ideal for individuation (Hill & Butler, 2012).

1.2.4 DYS391 and other Y-STRs

No P_I value was reported for the extended CODIS locus DYS391. This locus is the only non-autosomal, Y-chromosomal locus to be recommended as a core STR (Hill and Butler, 2012). There are, however, 219 known Y-chromosomal STRs and commercial kits that target panels of 9-11 non-recombining Y-STRs have been produced (Kayser *et* al, 2002; Butler, 2003; Jobling & Tyler-Smith, 2003). These kits are useful in cases where differential lysis (a technique that entails treating sexual assault samples with detergent and protease to lyse female epithelial cells and subsequent centrifugation to recover sperm nuclei) is unable to separate the male and female components in a sample, for example where the rapist is either vasectomised or azospermic. Applying autosomal STR profiling in these cases often results in profiles where the female alleles mask the alleles of the male contributor (Shewale *et al.*, 2003). Conversely, because Y-STRs are confined to men, the use of Y-STR profiling kits results in profiles specific for the male contributor (Shewale *et al.*, 2003). Y-STR kits are also useful for determining the number of male contributors in gang rape situations; simplified by the fact that most Y-STRs are single-copy loci (Prinz, 2003 & Butler, 2003). A limitation to these profiles is, however, that Y-chromosomes are paternally inherited and, consequently, all paternally related men, barring the effects of mutations, will have shared haplotypes (Jobling & Gill, 2004). Studies aimed at the identification of rapidly mutating (RM) Y-STRs to increase the P_I of STRs have been performed (Ballantyne *et al.*, 2010). It is predicted that with the production of commercially-available multiplexes which target RM Y-STRs, RM Y-STRs may replace existing Y-STRs for crime scene investigation (Kayser & de Knijff, 2011).

Although Y-STRs have been widely used in sexual assault cases internationally, despite the high incident of sexual assault cases in South Africa, Y-chromosomal DNA profiling is

currently not performed on forensic case work samples (Davison *et al.*, 2008). The University of the Western Cape is however currently investigating the use of Y-STRs in South African casework samples and is in the process of creating a database of Y-STR allele frequencies for local populations (Leat *et al.*, 2004; Leat *et al.*, 2006; Ehrenreich *et al.*, 2008).

1.3 Generation, application and significance of STR profiles

1.3.1 Generation of STR profiles

The initial step of STR-based DNA profiling is a multiplex PCR that amplifies the target STRs and Amelogenin. More specifically, loci are amplified using fluorescently-labelled primers (Butler, 2010; Sullivan *et al.*, 1993). Thereafter, amplicons are separated by size using capillary electrophoresis (CE) and STRs with alleles of overlapping sizes are differentiated by the discriminating fluorescent tags using a fluorescent detector (FD). A schematic of an electropherogram from the CE-FD of Amelogenin and the loci vWA, D21S11, D2S441, D3S1358 and D7S820 is presented in **Figure 2**. The sizes of alleles and their corresponding repeat number would have been determined by comparison of the allele peaks to sequenced allelic ladders (Butler, 2010).

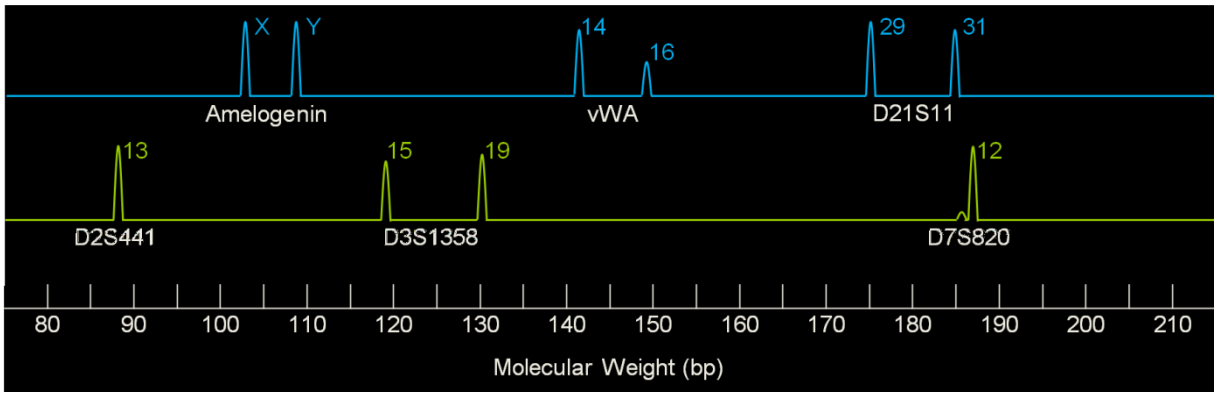


Figure 2: Schematic electropherogram DNA profile for 5 STRs and Amelogenin illustrating various stochastic effects of typing low copy number and/or degraded DNA

The STRs and sex-informative Amelogenin, amplified using fluorescently-labelled primers, are separated by capillary electrophoresis coupled to a fluorescent detector to create an electropherogram. In this case a two-colour (blue and green channels) fluorescent system is displayed. To create the profile, a third colour channel would be used for the size marker, containing a sequenced allelic ladder for each locus, to enable locus and allele calling. Locus names are given bellow the representative peaks and allele calls, representative of the number of repeat units, are given by the numbers adjacent to each peak. The peaks at 106 and 112 bp for amelogenin indicate that the donor was a male. vWA exhibits an imbalanced heterozygous peak, possibly due to a mutation in the primer binding site of the allele, while D21S11 and D3S1358 exhibit balanced heterozygous peaks. D2S441 and D7S820 exhibit single peaks, indicative of a homozygous allele, allele dropout, or the presence of a null allele. The peak adjacent to that of D7S820 is indicative of PCR stutter.

1.3.2 Utilising STR Profiles

In the cases of missing person identification, identification of mass-disaster victims and familial testing, or in cases were a suspect has been identified and a profile from a crime-scene has been obtained; DNA profiles can be matched directly (Kayser & de Knijff, 2011). Conversely, where DNA is obtained from an unknown donor, DNA profiles may be searched in intelligence databases (Jobling & Gill, 2004).

The composition and organisation of intelligence databases varies from country to country; however there are two sections that are usually present, the first comprising convicted offender and/or suspect profiles and the second containing profiles obtained from crime scene samples (Jobling and Gill, 2004). The CODIS database includes a third section of profiles from unidentified persons (National Research Council, 2009). These databases are useful for the identification of perpetrators and missing persons, the exoneration of suspects, and for the linking of crimes (Morling, 2004). In recent years, they have also been used for familial searching which aims to create investigative leads and limit suspect pools by searching an intelligence database for a profile of a relative to the donor of a sample from an unknown source (Gershaw *et al.*, 2011).

Table 3: The indices of the National Forensic DNA Database of South Africa

Index	Contents (Profiles from…)	Storage in NFDD
1. Crime Scene	Bodily samples collected at crime scenes	Indefinitely
2. Arrestee	Individuals accused of committing Schedule 1 offenses	3 years*
3. Convicted Offenders	All convicted offenders (past and current)	Indefinitely
4. Volunteer	Volunteers who have given written consent	As permitted
5. Elimination	Relevant officials e.g. forensic laboratory technicians	Indefinitely

* Profiles expunged if donor is acquitted; a decision not to prosecute was taken; the conviction was put aside on appeal/review; or no criminal proceedings were instituted

All data were obtained from the Criminal Law (Forensic Procedures) Amendment Act (2013)

An example of such a database is the National Forensic DNA Database of South Africa (NFDD); the content, organisation and management of which is outlined in the recently passed Criminal Law (Forensic Procedures) Amendment Act (2013) or "DNA Act." The database will comprise indices, outlined in **Table 3**. The DNA Act also prescribes that the profiles "shall not contain the following information derived from a bodily sample which was taken from a person: (a) The appearance of the person, other than indicating the sex; (b) medical information of the person; (c) historical information relating to the person; and (d) behavioural information of the person." With regards to the description of medical

conditions, concerns regarding the link between certain triallelic patterns and chromosomal abnormalities like trisomy-18 have been raised (Heathfield, 2014).

1.3.3 Describing the significance of STR profiles

In order for profile matches to be used as forensic evidence, statistical parameters that describe the significance of the matches are required (Huston, 1998). These statistical parameters, outlined in **Table 4**, are obtained by assessing allele frequencies ascertained from National DNA Databases (NDNAD) which contain profiles from representative population groups (defined in terms of racial group as well as geographic region) (Huston, 1998). It is important that loci used for DNA profiling exhibit low inter-population variances (F_{ST}) to ensure that all other parameters have similar values for the various population groups. These parameters include the Power of Exclusion (PE), which determines the usefulness of a particular panel of markers at excluding particular genotypes; and the Match probability (PM) or Probability of Identity (P_I), Power of Discrimination (PD) and Paternity Index (PI) which describe the specificity of a match to a particular individual (Tillmar. 2010). It is interesting to note that the reciprocal PMs for most STR panels used today exceed that of the human population, indicating that it is theoretically impossible for 2 unrelated individuals to have identical DNA profiles (Jobling & Gill, 2004).

The calculation of these parameters relies on two main factors; firstly, that there is no linkage or Linkage Disequilibrium (LD) between loci and alleles; and, secondly, that the alleles exhibit Hardy-Weinberg Equilibrium (HWE) (Hardy, 1908). Linkage occurs when alleles on a shared chromosome undergo co-segregation during meiosis and LD occurs when alleles of different loci are associated in a non-random manner (Ott, 1999). LD can arise when loci are insufficiently spaced from one another and are, therefore, inherited together; alternatively population genetic effects like selection, founder effects and admixture can result in LD (Ott, 1999). The effects of linkage and LD were alluded to in *1.2.3* with reference to the importance of core loci being on independent chromosomes or at a sufficient distance from one another. The HW theory states that alleles inherited in a Mendelian manner have predictable, constant frequencies in large, randomly breeding populations based on laws of probability (Planz, 2004). The result of HWE on allele frequencies is that the frequency of observed homozygotes is p^2 and the frequency of heterozygote profiles are 2pq; where p and q are the frequencies of respective alleles (Planz, 2004). Factors that disrupt this equilibrium include mutation, gene migration, selection and genetic drift (Planz, 2004). In cases where deviations from HWE occur due to inadequate sampling, the Bonferroni correction can result

in compliance of the data to HWE (Abrahams, *et al.*, 2011; Hill *et al.* , 2008; Schlebusch *et al.*, 2012)

Table 4: Statistical Parameters considered in the analysis of DNA profiles

Parameter	Symbol	Description
Inter-population Variance	F_{ST}	The proportion of inter-population genetic diversity to intra-population diversity
Power of Exclusion	PE	The proportion of individuals with a different profile to that of a randomly selected individual
Match Probability/ Probability of Identity	PM/P_I	The probability that 2 unrelated randomly-selected individuals in a population will have identical profiles
Power of Discrimination	PD	The probability that 2 randomly selected individuals in a population will have different profiles (1-PM)
Paternity Index	PI	The likelihood that the tested individual is the biological father of a subject, rather than a random member of the population

An example of an investigation aimed at determining whether forensic loci exhibited HWE and linkage equilibrium in a population was that performed by Lucassen *et al.* (2014). This study investigated the AmpFℓSTR® Identifiler Plus™ kit to assess its suitability for the individuation of members of the South African population so as to validate its implementation in forensic casework. Profiles from 1543 unrelated individuals of various population groups, defined as Black, White, Coloured and Indian, were analysed. The Bonferroni correction was required for the loci D3S1358 for the Indian population, D198S433 and TH01 for the Black population, and D19S433 and D16S539 for the Coloured population to bring them within normal expectations for HWE. LD and linkage had been excluded in previous studies of the loci (Butler, 2006) and results of this study confirmed these findings. The PM values ranged from $1/3.3\times10^{17}$ for the White population to $1/1.88\times10^{18}$ for the Coloured population and the results obtained from this study, therefore, served to validate AmpFℓSTR® Identifiler Plus™ for use in RSA and the profiles produced now form the population genetic database (reference database) of the Division of Forensic Services of the South African Police Service (SAPS).

1.3.4 Specialised applications of STR profiles

Studies, such as that performed by Luccassen *et* al (2014), aimed at investigating the allele frequencies of populations to enable the determination of statistical parameters have revealed that there is a difference in the allele proportions of different population groups, and that there is significantly (~20%) more STR diversity in African populations than either

Europeans or Asians, both in terms of largest total number of alleles and in terms of largest number of unique alleles (Jorde *et al.*, 1997; 2000). These results have led to a number of studies which use STRs to map human migration, evaluate population substructure, and which aim to determine whether biogeographic ancestry can be predicted from STR profiles.

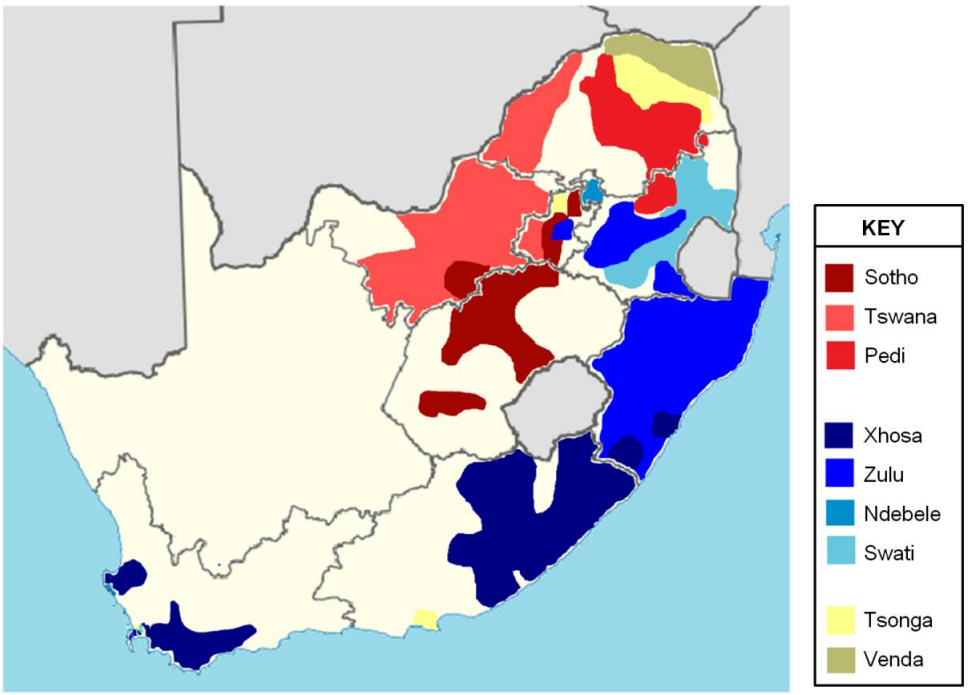


Figure 3: Geographic distribution of Niger-Kordofanian-speaking population groups within South Africa (Figure constructed from data taken from Statistics South Africa, 2011)

An example of a study which used STRs (among other genetic markers) to assess population substructure was that performed by Lane *et al.* (2002). This research group assessed the genetic differentiation between 7 South African Niger-Kordofanian-speaking groups; namely the Zulu, Xhosa, Tsonga/Shangaan, Southern Sotho, Pedi, Tswana and Venda populations. These groups, together with the Ndebele and Swati groups, comprise over 75% of the South African population, and the current geographic distribution of these population groups is displayed in **Figure 3** (Statistics South Africa, 2011). Lane *et al.* (2002) found that genetic distances correlated with geographic distances, as did genetic to linguistic distances; and that there was a genetic relatedness between the Xhosa and Zulu populations, and the Sotho, Tswana and Pedi populations. These results correlate to ceramic, linguistic, anthropological and archaeological evidence which suggest that 2 separate migration events from present-day Nigeria and Cameroon (in about 1100 AD) gave rise to the Nguni (Zulu, Xhosa, Ndebele and Swati) and Sotho-Tswana (Southern Sotho, Pedi and Tswana) populations (Huffman, 2006).

Studies aimed at assessing whether STR profiles can be used to predict biogeographic ancestries have yielded some promising results. Fosella *et al.* (2004), for example, were able to successfully discriminate between an Italian and Sub-Saharan population using 13 STRs. Lowe *et al.* (2001) illustrated that the accuracy of these STR-based predictions is population-dependent. Specifically, Lowe *et al.* (2001) examined 6 STRs and obtained correct predictions in 30-67% of cases, depending on whether the donor was Caucasian, Afro-Caribbean, Indian, Southeast Asian or Middle-Eastern.

1.4 Challenges to STR-based DNA profiling by PCR/CE-FD

While complete STR profiles generated by PCR/CE-FD can be used for familial testing, the identification of missing persons and mass disaster victims, and in the international fight against crime; STR-based profiling is not infallible and there are a number of challenges associated with it (Jobling & Gill, 2004). These challenges usually arise when sample DNA is of poor quality and/or low quantity; is comprised of DNA from numerous donors; or when there are sequence variations within template DNA (Kayser & de Knijff, 2011).

1.4.1 Degraded and/or low copy number template DNA

The effects of degradation and low copy number (LCN) template DNA on profile interpretation and analysis, are often relevant when generating profiles from samples obtained at crime scenes and mass disaster sites (Hughes-Stamm *et al.*, 2011). These samples often contain degraded DNA below the stochastic threshold of 100-200 pg required for reliable interpretation (Moretti *et al.*, 2001), which can result in profiles that exhibit allele and/or locus drop-out, imbalanced heterozygous alleles and enhanced PCR stutter (Pitterl *et al.*, 2010; Budowle *et al.*, 2009). Allele drop-out and imbalanced heterozygous alleles, the latter of which is illustrated by the peaks for vWA in **Figure 2**, present due to preferential amplification of certain loci (Schneider *et al.*, 2004; Butler *et al.*, 2003). These partial profiles are cumbersome to analyse and, when searched on established intelligence databases like that of CODIS and the UK National Database which contain millions of DNA profiles the number of hits returned can be unmanageable (Butler, 2006; Gill, 2002; Pitterl *et al.*, 2010).

The negative stochastic effects of typing degraded and LCN DNA are most frequently observed for the larger STRs (Butler *et al.*, 2003). Recent studies have therefore aimed to minimise amplicon size by designing primers which bind closer to the repeat unit of core loci (illustrated in **Figure 4**) (Grubwieser *et al.*, 2006; Butler *et al.*, 2003) and, as discussed in *1.2*,

have aimed to identify novel miniSTRs (Hill *et al.*, 2008). The applicability of these reduced-length and miniSTR loci is, however, limited by the fact that CE-FD distinguishes alleles based on size and fluorescence. The number of loci that can be analysed, therefore, is limited by the ability to resolve the STRs and by the number of fluorescent channels.

Figure 4: Schematic of an STR with original (v1) and reduced-length (v2) Primer Binding Sites (PBSs)
Primers are redesigned to bind closer to the STR repeat units (PBS v2) so as to minimise amplicon size and reduce the stochastic effects observed when degraded or LCN template DNA is used.

1.4.2 Mixed profiles

DNA profiling by PCR/CE-FD can also be complicated when sample DNA is comprised of DNA from multiple donors, as is often the case with samples collected from sexual assault victims and mass disasters sites (Brenner & Weir, 2003). While simple mixtures can often be resolved on the basis that peaks from a single individual will be of similar heights (comparing major and minor contributors); more complex mixtures require the use of likelihood ratios (LR) to give predictions on the allele calls for each profile (Jobling & Gill, 2004). Concerns relating to subjectivity and bias in forensic mixture interpretation have also been raised (Dror & Hampikian, 2011). It has been suggested that by increasing the number of loci profiled, and in particular including loci with low stutter frequencies, mixture interpretation may be simplified (Bacher & Schumm, 1998; Phillips *et al.*, 2014); however, as alluded to earlier, the number of loci that can be analysed in any one profile is limited by the nature of loci differentiation in CE-FD.

1.4.3 Sequence variations within STRs

Challenging and anomalous profiles can also occur despite sample DNA being of high quality and sufficient quantity. These anomalies tend to arise when template DNA contains mutations, either in the form of insertions, deletions, substitutions (including SNPs) or duplications and, depending on the nature and location of these mutations, different anomalies, summarised in **Figure 5**, can be observed (Huel *et al.*, 2007). Where DNA profiles are generated for the purpose of comparative searching in a NDNAD, or where a profile is being compared directly against another to identify whether or not there is a match, these anomalies have little effect except for increasing the PM. In cases where profiles are being used for familial testing, however, the presence of novel mutations that result in a

deviation from Mendelian inheritance can have drastic effects on the outcome of the analysis (Lane, 2013).

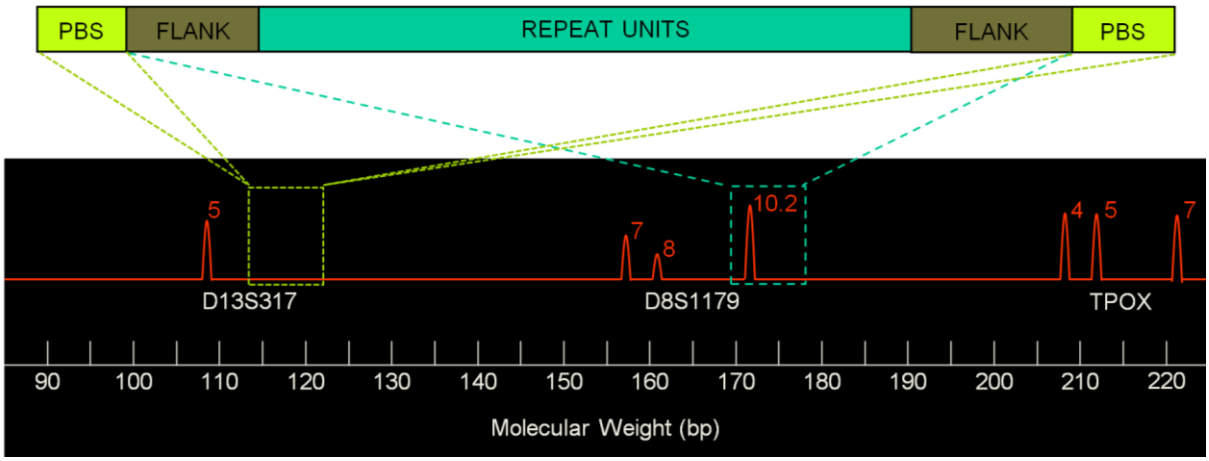


Figure 5: STR sequence variations and their effect on observed electropherogram peaks

Loci names are given below the peaks and allele calls, representative of the number of repeat units, are given by the numbers adjacent to each peak. Sequence variations in the primer-binding site (PBS) of the STR result in a null allele for D13S317, indicated by the **green** box. Sequence variations in the flanking region (FLANK) and/or repeat units result in an off-ladder allele, indicated by the **turquoise** box. D8S1179 and TPOX exhibit Type I and Type II triallelic patterns respectively.

Anomalous alleles that arise due to insertions or deletions within the sequence of STR flanking regions or the repeat units themselves, result in the occurrence of "off-ladder" or microvariant alleles (Allor *et al.*, 2005; Mizuno *et al.*, 2003; Heinrich, 2005; Grubwieser, 2005), illustrated by allele 10.2 of D8S1179 in **Figure 5**. The locus D7S820 is particularly prone to off ladder alleles because it contains a homopolymeric string of 8, 9 or 10 T residues in its 5´ flanking region (Egyed *et al.*, 2000). While off-ladder alleles of the type illustrated in **Figure 5** and described for D7S820 do not complicate allele-calling significantly, in other cases, off-ladder alleles fall within range of adjacent loci, which can complicate loci and allele calling considerably (Heinrich, 2005; Grubwieser, 2005).

Mutations within the primer binding sites (PBSs) of STRs can have two possible effects: The efficiency of primer binding may either be reduced, resulting in imbalanced heterozygous peaks; or, primer binding may be prevented completely, resulting in the presence (or lack thereof) of a null allele (Clayton *et al.*, 2004; Budowle *et al.*, 2001). Null alleles cause relevant loci to be erroneously called as homozygous, illustrated in **Figure 5** by D13S317. Concordance studies on a variety of commercial kits have been performed with the aim of identifying null alleles (Hill & Butler, 2012; Phillips *et al.*, 2014). Interestingly, many of these null-alleles are specific to certain population groups, for example the NGM Select kit

primers for SE33 and D2S441 result in null alleles for 5% and 6% of African Americans and Koreans respectively (Phillips *et al.*, 2014). The results of these studies have enabled the inclusion of degenerate primers, for example the degenerate primer for D16S539 which has been added to the Powerplex 1.1 system (Nelson *et al.*, 2002), so as to reduce the incidence of null alleles. To date (according to the author's knowledge) only one paper has been published which estimated, rather crudely, the null allele frequencies in South African populations (Lane, 2013). This is of particular concern to identity and familial testing in South Africa since African populations are known to have highest degree of genetic diversity (Jorde *et al.*, 1995) and are likely therefore to exhibit novel mutations that affect primer binding.

Triallelic patterns can also complicate profiles, especially when they arise during the profiling of mixed DNA samples. There are two main types of triallelic patterns observed during STR analysis: Type I alleles gives rise to two peaks, the sum of the area of which is equal to the area of the third peak. These alleles occur as a result of somatic mutations prior to segregation during development. Type II alleles, on the other hand, give rise to three peaks, each with the same area. These alleles occur either as a result of a localized duplication event, or as a result of chromosomal trisomy (Clayton *et al.*, 2004; Rolf *et al.*, 2002). A schematic of the typical triallelic patterns is depicted in **Figure 5** by the loci D8S1179 and TPOX respectively. TPOX is notable because it frequently displays Type II triallelic patterns, which has been hypothesised to be due to its location (see **Table 2**) near the tip of Chromosome 2 (Chakhparonian & Wellinger, 2003; Louis & Vershinin, 2005).

DEVELOPMENTS IN DNA PROFILING

1.5 Single Nucleotide Polymorphisms revisited

Not only did the completion of the Human Genome Project (International Human Genome Sequencing Consortium, 2004) result in investigations aimed at identifying 'ideal' STRs (discussed in *1.2.2*), but it also led to questions being raised as to whether STRs are, in fact, the most suitable genomic marker for DNA profiling. Consequently, and further stimulated by the development of HapMap (International HapMap Consortium, 2005), the applicability of a variety of SNPs (illustrated in **Figure 1 B**) to forensics has been reconsidered (Butler *et al.*, 2007). These applications include DNA profiling for individuation and familial testing, and phenotype and biogeographic ancestry prediction; all made possible by the abundance

and versatility of SNPs, which occur every several hundred bases in both coding and non-coding regions of nuclear and mitochondrial DNA (mtDNA).

1.5.1 SNPs for individuation and familial testing

SNPs have been reconsidered for their use in DNA profiling primarily because of their potential to type highly degraded DNA samples without falling victim to the stochastic effects observed for STRs discussed in *1.4.1*; and for their use in complicated familial testing scenarios, for example where non-Mendelian inheritance for a locus is observed. These applications are made possible by the small amplicon sizes required to type SNPs (50 bp as opposed to 100-400 bp for STRs) (Budowle, 2004) and by their low mutation rates (10^{-8} as opposed to 10^{-3} for STRs) (Butler *et al.*, 2007). During the identification of victims of the 9/11 attack, these two characteristics were exploited and DNA from family members was used as the reference DNA for kinship analysis and ultimately individual identification (Biesecker *et al.*, 2005).

The identification of 9/11 victims was a relatively specialised application of SNPs, and in order for SNPs to be applied routinely in human identification, loci must be selected which, like STRs, are unlinked and which exhibit high levels of population-independent allelic diversity, indicated by low F_{ST} values (discussed in *1.3.2*) and high heterozygousity values (Kidd *et al.*, 2006). This is vital to ensuring that the P_I value for the panel of markers is similar across different population groups (Kidd *et al.*, 2006). A study by Pakstis *et al.* (2010) identified 45 such loci which have subsequently been used to develop a multiplex for SNP-based DNA profiling (Rixun *et al.*, 2009); and a study by Sanchez *et al.* (2008) demonstrated that single base extension multiplexes followed by CE-FD with multicolour detection can be used to generate SNP-based DNA profiles for use in forensics.

One notable drawback to SNP-based DNA profiling, however, is that SNPs are usually biallelic in nature and, as evidenced above, require between 40 and 60 loci to obtain the same level of discrimination as that achieved using 13-15 STRs (Gill, 2001). This is likely not only to increase the cost of DNA profiling, but also to complicate and prolong the analysis process (Butler *et al.*, 2007). The biallelic nature of SNPs also makes mixture resolution virtually impossible; and while some triallelic SNPs have been characterised in an attempt to enable mixture resolution (Westen et al., 2009), the resolution obtained from these SNPs is still considerably less than that obtained using STR-based typing.

Another major disadvantage to DNA profiling by SNP analysis is that the infrastructure and expertise for STR-based profiling is established; as are many reference and intelligence DNA databases which contain millions of profiles (discussed in *1.3*). In a review by Buter *et al.* (2011) the observation was made that in countries that do not have established DNA databases, SNP analysis may be considered as an alternative method for DNA profiling. South Africa, which was an example of one such country, opted to retain STR-based typing. The DNA Act (2013) states explicitly that "Profiles may be shared between foreign states or recognised international organisations, tribunals or entities" and the decision to retain STRs as the loci of choice was likely to enable this international collaboration. It seems therefore that the conclusion reached by Butler *et al.* (2007) in a review considering the application of STRs versus SNPs that, 'STRs rather than SNPs will fulfill the dominant role in human identity testing for the foreseeable future' was correct.

1.5.2 Specialized SNP applications

Specialized forensic applications of SNPs, identified during genome-wide association (GWA) studies, have been developed and include the prediction of externally visible characteristics (EVCs) and biogeographic ancestry prediction (Kayser & de Knijff, 2011).

EVC prediction aims to enable forensic DNA phenotyping (FDP) to limit suspect pools, and makes use of multiplexes that target SNPs in genes associated with particular phenotypes (Kayser & de Knijff, 2011). An example of such a multiplex is the HIrisPlex system which targets 24 SNPs and enables the prediction of both eye, using the 6 SNPs of the forensically validated Irisplex system (Walsh et al, 2011), and hair colour (Walsh *et al.*, 2013). Recent studies have also identified 5 SNPs which account for 82% of skin colour variation (Lao *et al.*, 2007); over 180 SNPs which influence adult height (Lango *et al.*, 2010). The application of FDP is limited however, because phenotypic characteristics can be drastically altered by environmental factors and both pigmentation and height exhibit age-dependent variations (Kayser & de Knijff, 2011).

Biogeographic ancestry prediction may also be useful in limiting the suspect pool, and is performed using commercial microarrays which target genome-wide SNPs (Li *et al.*, 2008; Jakobsson *et al.*, 2008) or systems which target small sets of ancestry informative markers (AIMs) (Bamshad *et al.*, 2004). These markers often reside within the maternally inherited mtDNA or the paternally inherited Y-chromosome which do not undergo recombination and, therefore, exhibit reduced population size which makes them susceptible to revealing genetic

clusters and clines (Underhill & Kivisild, 2007). Clusters and clines result in different individuals exhibiting a decreased genetic overlap with increasing geographic distance (Kayser & de Knijff, 2011).

While the commercial microarrays are able to predict individual ancestry proportions, they are not suited to forensic applications because they consider tens-of-thousands of markers, and are therefore resource intensive to prepare and time consuming to analyse (Kayser & de Knijff, 2011). Conversely, kits that target small sets of AIMs are suited to forensic applications and have been shown capable of distinguishing between individuals from various continents and sub-continents; namely Africa, Asia, Europe and America (Hispanic Americans in particular) (Frudakis *et al.*, 2003). Again the application of these specialised SNPs is limited, this time, however, by the effects of admixture in many population groups and by the fact that biogeographic ancestry is not necessarily indicative of phenotype and therefore, may have limited use in suspect pool minimisation (Kayser & de Knijff, 2011).

It should be noted that in most countries the legal framework has not yet been developed to permit and regulate these specialised applications of SNPs (Koops *et al.*, 2008) and in South Africa, the DNA Act specifies that a DNA profile may not "contain any…physical information of that person other than the sex of that person."

1.5.3 SNPSTRs: SNPs within and associated with STRs

Studies investigating the evolution and migration of humans have not only led to the identification of SNP AIMs, but have also led to the discovery of a set compound AIMs called SNPSTRs, which were catalogued in the SNPSTR database (http://imperial.ac.za.uk /theoreticalgenomics/data-software) by Agrafioti & Stumpf (2007). Each SNPSTR is comprised of at least 1 SNP within 500 bp of a single STR (Mountain *et al.*, 2002). The usefulness of these markers for biogeographic ancestry prediction stems from the fact that, firstly, the markers are tightly linked and likely therefore to be inherited together without undergoing recombination and, secondly, that the markers have vastly different mutation rates (see *1.5.1*). The slower mutation rate of the SNPs enables the discrimination between homoplasmic STRs (STRs that have undergone convergent evolution and are therefore identical by state but not descent) and STRs that have shared identities due to evolution from a common ancestor (identical by state and descent) (Ramakrishnan & Mountain, 2004).

A recent study by Wang *et al.* (2013) demonstrated the applicability of SNPSTRs to forensics DNA profiling by analysing the forensically relevant STR D5S818 linked to the SNP

rs25768. The amplification refractory mutation system (ARMS), using an STR-specific forward primer and 2 reverse primers, each labelled with a different fluorescent tag and specific for a different haplogroup, coupled to CE-FD, was employed for the typing of the SNPSTR (Wang *et al.*, 2013). Given that the P_I of a DNA profile is decreased by increasing the number of loci examined, and given the already limited number of STRs that can be analysed using CE-FD, it seems unlikely that SNPSTRs assessed by CE-FD would replace currently used STRs, since a single SNPSTR is unlikely to give a lower P_I than two STRs. This technique, while effective, is therefore unsuitable for routine DNA profiling.

Conversely, the use of compound markers in complicated familial testing scenarios could be highly beneficial. This was demonstrated by Ye *et al.* (2014) where a son exhibited an allele for CSF1PO, 1 repeat unit smaller than the alleged father. After analysis of 5 CSF1PO-linked SNPs, the probability of paternity increased from 99.92% to 99.998% and the anomalous STR was, therefore, explained by a mutation. Ye *et al.* (2014) concluded that a panel of SNPSTRs enabled a higher discriminatory power than either STR or SNP analysis alone and was able to differentiate uncles from fathers in paternity testing.

The usefulness of SNPSTRs in forensics has also been demonstrated in studies aimed at assessing alternative detection platforms for DNA profiling. These alternative platforms include ICEMS, that is electrospray ionization quadrupole time-of-flight mass spectrometric detection of amplicons following ion-pair reversed-phase high-performance liquid chromatography (Oberacher *et al.*, 2008) and next generation sequencing (NGS) which will be discussed in detail in *1.6.* ICEMS reveals the existence of SNPs within STR amplicons, however, is unable to give the position of the SNPs within the amplicons; NGS, on the other hand, is able to give the exact position and identity of the SNP.

A series of studies, summarized in **Table 5**, was performed by Oberacher and Pitterl *et al.* (2008, 2009) using ICEMS as the detection platform for DNA profiling. The initial study investigated 21 forensically relevant STRs in an Austrian population and revealed that 11 of the investigated loci; namely SE33, D2S1338, vWA, D21S11, D3S1358, D16S539, D8S1179, D7S820, D13S317, D5S818 and D2S441; yielded additional allele variants due to the presence of SNPSTRs. This resulted in a 99.5% decrease in the combined PM for the loci typed by ICEMS versus CE-FD and “increased the overall information content of this set of loci by 20 to 30% (equaling two to three STRs)” (Oberacher *et al.*, 2008).

Table 5: Summary of STRs analysed and variant STRs observed during studies investigating the use of ICEMS as an alternative detection platform for DNA profiling

	Oberacher *et al.*, 2008		Pitterl *et al.*, 2010		Planz *et al.*, 2009	
Loci	Tested	SNPSTRs	Tested	SNPSTRs	Tested	SNPSTRs
SE33	Y (S)	✓	Y (S)	✓		
D2S1338	Y (S)	✓	Y (S)	✓		
VWA	Y (S)	✓	Y (M)	✓	Y	✓
D21S11	Y (S)	✓	Y (M)	✓	Y	✓
D3S1358	Y (S)	✓	Y (M)	✓	Y	✓
D16S539	Y (S)	✓	Y (M)	✓	Y	×
D8S1179	Y (S)	✓	Y (M)	✓	Y	✓
D7S820	Y (S)	✓			Y	✓
D13S317	Y (S)	✓			Y	✓
D5S818	Y (S)	✓			Y	✓
D2S441	Y (S)	✓				
TPOX	Y (S)	×			Y	×
CSF1PO	Y (S)	×			Y	~
D10S1248	Y (S)	×				
D22S1045	Y (S)	×				
TH01	Y (S)	×	Y (S)	✓	Y	×
FGA	Y (S)	×	Y (S)	✓	Y	~
D18S51	Y (S)	×	Y (S)	✓	Y	~
D19S433	Y (S)	×	Y (S)	✓		
Penta D	Y (S)	×				
Penta E	Y (S)	×				

Ion-pair reversed-phase high-performance liquid chromatography coupled to electrospray ionization quadrupole time-of-flight mass spectrometric (ICEMS) was used for allele typing of STRs. STRs included in the study, Y, were either amplified in singleplex (S) or multiplex (M) with the orange box indicating a multiplex developed by Pitterl *et al.*, 2008. Planz *et al.* (2009) specified that 8 PCRs were performed but did not specify which loci were amplified in multi- or singleplex. The ✓ indicates observed variant alleles (SNPSTRs) and the × indicates that no variant alleles were observed. The ~ indicates that very few variants were observed, and that no significant increase in DP was observed when using ICEMS in place of CE/FD (specific values were unreported).

Pitterl *et al.* (2008) targeted these same 21 forensically relevant STRs for multiplex development for use in ICEMS and succeeded in the generation of a multiplex (indicated by the orange box in **Table 5**) which amplified Amelogenin plus 13 STRs. This multiplex was optimised by increasing and decreasing the primer concentrations of the relatively under- and over-amplified loci respectively until even amplification of STRs resulted. The optimised multiplex was then used to amplify the loci D3S1358, D21S11, D8S1179, vWA, and D16S539, along with TH01, FGA, D18S51, D2S1338, D19S433 and SE33 amplified in singleplex, of the autochthonous Central Asian Yakut and South African Khoe-San populations (Pitterl *et al.*, 2010). These population groups were chosen because they are

likely to exhibit greater genetic diversity due to their autochthonous nature (Pitterl *et al.*, 2010). Results obtained demonstrated an increased allelic resolution for all loci, indicating the presence of SNPs within the flanking regions or STRs themselves for all loci.

Planz *et al.* (2009) used the protocol developed by Oberacher *et al.* (2008) to type the CODIS loci of American Caucasian, Hispanic and African American population groups. The results obtained by Planz *et al.* (2009) are summarised in **Table 5** and **Figure 6**, the latter of which depicts the increase in PD of each locus observed for each population group. Planz *et al.* (2009) reported that these 7 loci had a combined PD of over 99.999999% when typed using ICEMS, which is equivalent to the PD obtained for the analysis of 10 loci by CE-FD

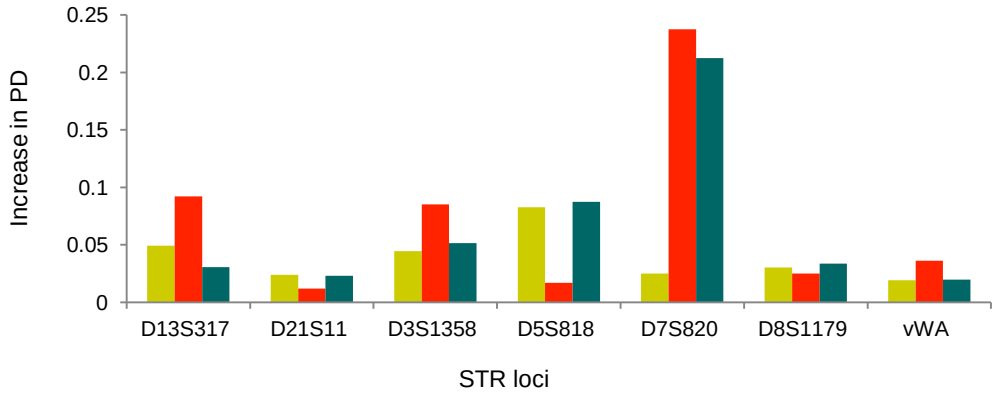


Figure 6: Increase in the discrimination power (PD) of STRs typed by ICEMS as opposed to CE-FD.
STRs were typed from American Caucasian (**lime**), African American (**orange**) and Hispanic (**blue**) population groups by Planz *et al.* (2009).

Analysis by ICEMS is promising for forensic applications since, not only does it reveal the presence of SNPSTRs thereby increasing the PD of a profile, but it also creates profiles that are compatible and comparable with existing databases (Oberacher *et al.*, 2008). The increased PD is particularly important where the quality of the sample only permits the generation of a partial profile (Oberacher *et al.*, 2008) and the typing of SNPs within STRs is also useful in complicated familial testing (Ye *et al.*, 2014). ICEMS, however, is unable to reveal the position of SNPs within STRs; and sequencing, therefore, retains the status of the method capable of yielding the highest PD from a given DNA profile (Pitterl et al., 2010). The number of loci that can be typed is also limited in ICEMS by the number of suitable mass differentiating markers available, a similar limitation to that of CE-FD discussed in *1.4.1* (Oberacher *et al.*, 2008).

1.6 Next Generation Sequencing

While sequencing is known to yield the highest possible PD for DNA profiling, the low throughput and sensitivity of Sanger sequencing, together with its high cost, have prevented it from being used for routine DNA profiling (Fullwood *et al.*, 2009). In the last 5 years, however, the applicability of sequencing to DNA profiling has been reconsidered due to the development of high throughput, low cost, Second and Third ('Next') Generation Sequencing (NGS); a comparison of some of the most frequently used NGS technologies to Sanger sequencing is presented in **Table 6**.

Table 6: A comparison of commonly used 1st, 2nd and 3rd generation (Gen.) sequencing technologies

Gen.	Method	Year	Throughput	Read length	Accuracy	Sequencing	Detection
1st	Sanger	1977	-	<1200 bp	~100%	Chain term.	Fluorescent
2nd	Roche 454	2005	500 Mbp/run	100-600 bp*	>99%*	Synthesis	Light
2nd	ABI SOLiD	2007	100 Gbp/run	50-100 bp	~100%	Ligation	Fluorescent
2nd	Illumina	2007	200 Gbp/run	2 x 250 bp**	99.9%	Synthesis	Fluorescent
2nd	Ion Torrent	2010	100 Mbp/run	<400 bp	99.8%	Synthesis	Semicond.
3rd	PacBio SMRT	2011	500 Mbp/run	>5 500bp	99.3%	Synthesis	Fluorescent

Details obtained from Phillips *et al.* (2014), Yang *et al.* (2014) and Manufacturer's websites; and accuracy is based on the substitution error-rate reported by Zascavage *et al.* (2013)

* Read lengths for GS Junior are 200-400 bp and read lengths for GS FLX are 400-600 bp; Accuracy for GS FLX Titanium is 99.995% while for GS Junior it is 99%.

** Read lengths were limited to 100 bp in the Illumina GAIIx but with the Illumina MiSeq read lengths of 250 bp are possible.

1.6.1 Forensic application of NGS

NGS enables the simultaneous sequencing of thousands to billions of sequences from hundreds of samples (differentiated by barcoding sequences), providing the potential for the simultaneous typing of mtDNA, sex chromosomes, STRs, miniSTRs and SNPs from hundreds of individuals (Yang *et al.*, 2014). Currently, however, only proof-of-concept studies for DNA-profiling by NGS have been performed, some of which are summarised in **Table 7**.

The Roche 454 sequencer was the first Second Generation Sequencing (SGS) technology developed, and given that forensically relevant STRs are between 100 and 400 bp in length (Sanchez *et al.*, 2006), it follows that 454 sequencing was the first SGS technology to be investigated for forensic application. Studies investigating the use of 454 have revealed that this technology has difficulty resolving homopolymers (sequences of >2 adjacent identical

nucleotides). This difficulty arises because the method relies on sequencing by synthesis (pyrosequencing) and, therefore, while the nucleotides in a homopolymeric stretch are added sequentially, the detection is simultaneous which can result in a non-linear increase in the signal emitted (Zascavage *et al.*, 2013). Nevertheless, Mikkelsen *et al.* (2014) reported an accuracy of 95% for homopolymers of up to 4 bases (or 6 bases if data were visually inspected); and a number of studies (Fordyce *et al.*, 2012; Scheible *et al.*, 2014 etc) have demonstrated that 454 can be used for the typing of STRs in both single contributor and mixed samples with a higher DP to CE-FD. For example, Scheible *et al.* (2014) observed 1 polymorphism for every 6 alleles in 50 analysed samples which were unobserved in CE-FD.

Table 7: A summary of papers investigating the used SGS for STR-based DNA profiling

Authors	Year	Platform	Loci	PCR	Samples
Fordyce *et al.*	2011	GS FLX	5 STRs	Singleplex	SC
Van Neste *et al.*	2012	GS FLX	Profiler Plus	Kit Multiplex	SC & MC
Bornman *et al.*	2012	Illumina GAIIx	13 CODIS	Singleplex	SC & MC
Van Neste *et al.*	2013	Illumina MiSeq	PPP STRs	Unoptimised Multiplex	MC
Scheible *et al.*	2011 2014	GS Junior	13 loci[1] 12 4-plexes	Multiplexes	Degraded SC
Gelardi *et al.*	2014	GS Junior	D3, D12, D21	Multiplex (D3 & D21) Singleplex (D12)	SC
Xiangpei *et al.*	2014	Illumina MiSeq	PPP Fusion loci	Prototype multiplex	SC
Fordyce *et al.*	2015	Ion torrent	10-plex panel	Optimised Multiplex	SC; MC & CS

[1]Multiplex designed by Pitterll *et al.* (2008) partially optimised for this study

The sequences of the STR-specific Promega PowerPlex (PPP) primers have been publicised and were used by Van Neste *et al.* (2014) and Zeng *et al.* (2014). The samples studied were either from single contributors (SC), multiple contributors (MC) to create mixed profiles, or case samples (CS).

Studies investigating the use of Illumina technology have revealed fewer sequencing errors than the 454 systems; however studies have been limited by the short read lengths of the system and allele dropout for the larger alleles was reported using the Illumina GAIIx. For example, Bornman *et al.* (2012) reported dropout of alleles 34.2 and 32.2 for D21S11. With the launch of the Illumina MiSeq system, however, read lengths have been increased from 2 x 100 bp to 2 x 250 bp and Promega has consequently redesigned its PowerPlex kit to enable STR amplification specifically for sequencing using the Illumina MiSeq system (Xiangpei *et al.*, 2014). Xiangpei *et al.* (2014) reported that the Profiler Plus Fusion kit is able to generate amplicons from as little as 62 ng template DNA, and that the Illumina MiSeq system is capable of generating reliable profiles from 6-200 ng amplicon libraries. The

Illumina MiSeq has also been used for the development of optimized approaches to mtDNA analysis (McElhoe *et al.*, 2014).

Sequencing using the Ion Torrent (IT) system is gaining increasing popularity, especially since the announcement by Roche of the discontinuation of the 454 systems, and a number of papers have been published in the last year or two investigating the application of IT to forensics (Parson *et al.*, 2013; Daniel *et al.*, 2014; Fordyce *et al.*, 2015). Parson *et* al (2013) investigated the suitability of IT to mtDNA analysis and Daniel *et al.* (2014) used 5 SNaPshot multiplexes (including the previously discussed IrisPlex system) to type 3 samples at various concentrations using IT. These papers are particularly interesting because they demonstrate the versatility in marker type that can be assessed using SGS. Fordyce *et al.* (2015) successfully typed single contributor, mixed and case work samples (some of which generated only partial profiles when typed with CE-FD) using a commercial STR multiplex developed by Thermo Fisher specifically for use in SGS using IT. Full profiles were obtained for most samples with input concentrations as low as 50 ng and mixture resolution was possible down to 20:1.

Third generation sequencing (TGS) is capable of the real-time sequencing of single, unamplified molecules and is able to detect epigenetic modifications which may be useful for the individuation of monozygotic twins. To the author's knowledge however, no studies investigating the applicability of TGS to forensics in particular have been performed; although Pacbio Single Molecule Real-Time (SMRT) technology has been used for targeted DNA analysis aimed at the identification of medically relevant SNPs and genetic variation (Carneiro *et al.*, 2012). Given that TGS is capable of the simultaneous sequencing of single molecules without the lengthy process of library preparation required in SGS, it is a promising tool for forensics.

1.6.2 Analysis of STR-data from NGS

It was noted by Zhang *et al.* (2011) that "the full benefit of NGS will not be achieved until the bioinformatics are able to maximally interpret and utilize these short-read sequences." In the case of STR analysis, this interpretation is further complicated by the tandem repeat (TR) nature of the loci since most data processing for allele calling is performed by aligning test sequences to reference sequences, and alignment programs have difficulty producing accurate alignments for TRs (Phillips *et al.*, 2014). Consequently, a number of papers have been published which present algorithms for the analysis of STR-data generated by NGS (Gymrek

et al., 2012; Van Neste *et al.*, 2014, 2013; Warshauer *et al.*, 2015). LobSTR, the program developed by Gymrek *et al.* (2012) was designed for the profiling of STRs within personal genomes and is therefore unsuitable for use in the forensic typing of targeted STRs. Conversely, STRait Razor and MyFLq, the programs created by Warshauer *et al.* (2015) and Van Neste *et al.* (2012;2014) respectively, were designed specifically for the analysis of STRs amplified using SGS. These programs have similar processing protocols in which reads are (1) separated into separate loci; (2) grouped according to sequence similarities; (3) filtered to remove erroneous sequences like PCR stutters; and (4) aligned against a library of known alleles to enable allele-calling and identification of variants. In the case of the MyFLq, an additional step of homopolymer compression (the conversion of strings of identical nucleotides to a single nucleotide) was included to avoid erroneous allele-calling for STRs typed with 454 and IT systems.

Until this year, all programs designed for STR analysis were operated by command-line and, therefore, required a level of bioinformatic experience to operate. This year, however, Van Neste *et al.* (2014) launched a web-based version of MyFLq which has a graphical user-interface. A number of parameters can be specified by the user, for example, whether or not homocompression is desired or not, and a level of understanding is required in order to apply the software to NGS data.

SUMMARY AND PROJECT AIMS

Analysis of STRs following PCR/CE-FD has been the method employed for routine DNA profiling for over 2 decades. There is, therefore, not only a thorough understanding of the characteristics of the core STRs, but there is also a wealth of expertise related to the generation and interpretation of profiles. Databases of STR profiles are also well established, which enable the significance of matches to be described and which are used for individuation and criminal investigations. Consequently, DNA profiling by PCR-CE-FD has been successfully employed for the identification of missing persons, the naming of suspects and linking of crimes, and the resolution of kinship.

There are, however, a number of challenges related to DNA profiling of STRs by PCR/CE-FD. When DNA is of low quality or quantity, for example, profiles can exhibit heterozygote imbalance, increased incidence of stutter, and allele drop-out. This results in the generation of partial profiles which have a lower PD than complete profiles and, when

searched in existing databases, return an unmanageable number of hits. Likewise, the presence of these partial profiles within databases gives rise to an unmanageable number of hits when searching complete profiles. Because smaller loci are less susceptible to stochastic effects observed when typing LCN/degraded DNA, mini-STRs and reduced-size amplicons have been introduced to core loci. Mixed DNA profiles are also challenging to resolve; as are profiles generated from DNA samples which contain sequence variations in the PBS, flanking regions or repeat regions. An increased number of loci would enable increased resolution when analysing mixtures, while providing increased PD in samples that exhibit allele drop-out due to sequence variations in the PBS. The number of small loci and the number of additional loci that can be analysed using CE-FD is, however, limited to the number of fluorescent labels and channels provided by a given system.

Alternative methods of DNA profiling have been considered to overcome these challenges. For example, SNPs have been proposed as an alternative marker for DNA profiling due to their small amplicon sizes and relative immunity to the stochastic effects of LCN/degraded DNA typing. Because of the lower PD of SNPs, increased difficulty to resolve mixtures, and the wealth of expertise and data available for STRs; SNPs are unlikely to replace STRs as the workhorse for routine DNA profiling. Rather, SNPs will be used in specialised applications like FDP, biogeographic ancestry prediction and resolution of complicated kinship analysis.

SNPSTRs, on the other hand, are promising supplementary markers as they provide increased PD for existing loci whilst producing data that are both compatible and comparable with existing databases. They also enable increased resolution useful for mixture interpretation and the calling of off-ladder alleles. The slower mutation rates of SNPs when compared to STRs provide insight for complicated familial testing and are also useful in biogeographic ancestry prediction. However, allele frequencies for SNPSTRs within different population groups are as yet not well understood.

Furthermore, in order to type SNPSTRs an alternative detection platform is required. While MS has been proposed as one such platform, NGS provides the highest possible PD for each locus as it provides complete sequence data. NGS also enables the simultaneous analysis of loci with overlapping sizes (enabling the simultaneous typing of a large number of mini- or reduced-size STRs). Furthermore, different kinds of markers, for example SNPs within mtDNA, Y-STRs and SNPs which determine EVCs, can be analysed together with autosomal STRs. While a number of proof-of-concept studies have been performed on various SGS

platforms, at the time of commencement of this study the Roche GS Junior system was the only platform capable of sequencing the full complement of core STRs and no STR-specific multiplexes had been developed specifically for analysis by NGS. The aim of this study was, therefore, to develop a two-step multiplex PCR for the amplification of reduced size STRs, and to use this protocol to identify novel SNPSTRs and other STR variants in South African populations.

RESEARCH QUESTIONS

1. Protocol Development

   Using the primer sequences published by Oberacher *et al.* (2008), can a multiplex which amplifies reduced size STRs be developed for analysis using the Roche GS Junior System?

2. SNPSTR Discovery and Analysis

   Can this protocol be used to identify novel SNPSTRs and STR variants in Nguni and Sotho-Tswana populations, and are the alleles of these SNPSTRs and STR variants associated with either population group?

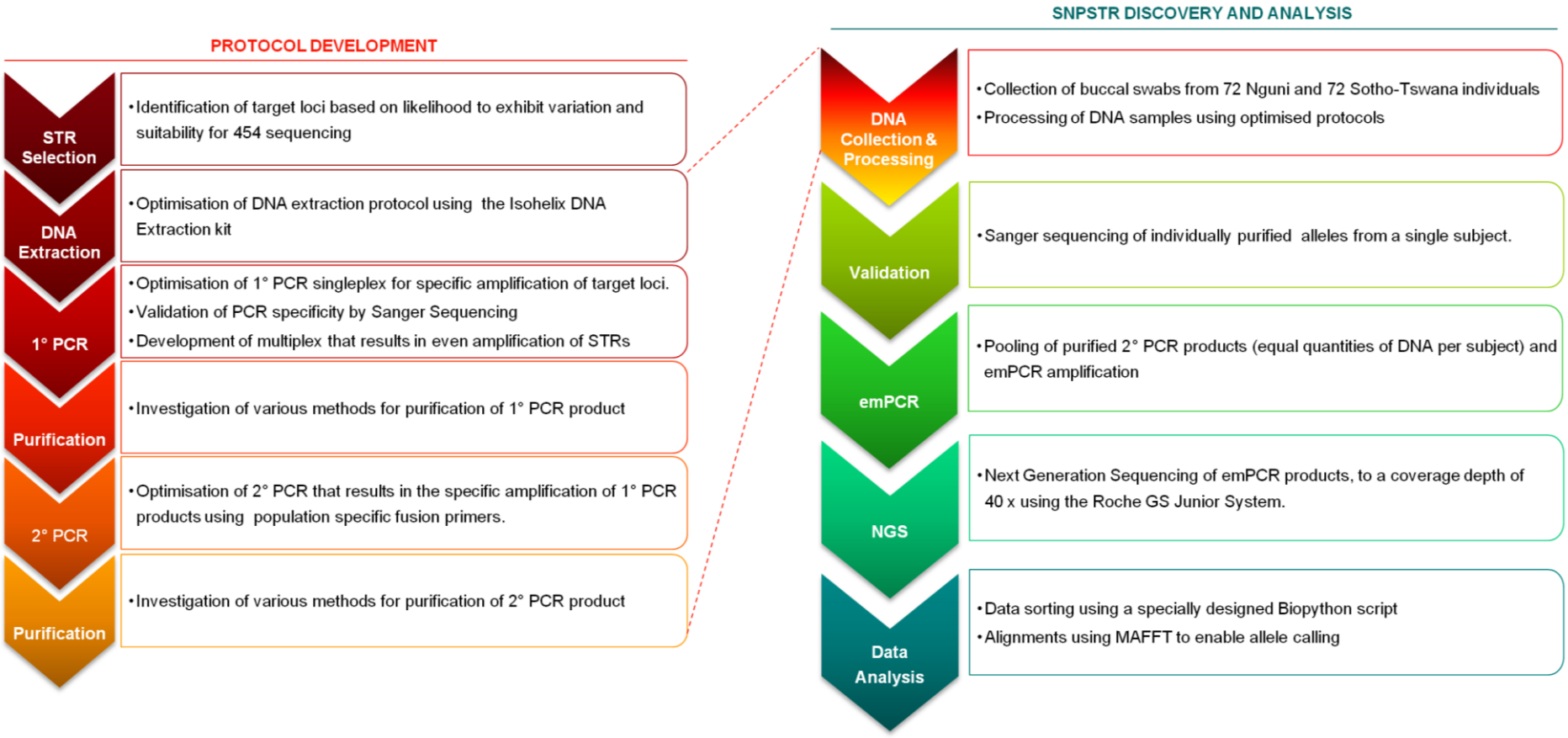


Figure 7: Outline of the objectives of and methodology employed for this study.

The protocol development was performed prior to sample collection. Thereafter, buccal swabs from 144 individuals were collected and the optimised protocol was used to prepare the amplcon library for emPCR and subsequent 454 sequencing.

CHAPTER 2
Protocol Development

This project was granted ethical clearance by the Ethical Standards Committee of Rhodes University, Grahamstown, South Africa (reference number 2013Q1-5). DNA processing protocols were optimised and validated using buccal swabs collected from five unrelated subjects, namely Subject 0 (of a Western-European population), and Subjects i-iv (of the Nguni population).

2.1 METHODOLOGY

2.1.1 Selection of target STRs

The autosomal STRs described in **Table 1** were considered as potential targets for this study. From this set, the loci vWA, D2S441, D3S1358, D21S11, D7S820 and D13S317 were selected for further analysis based on (1) the suitability of the loci to amplification by the 2-step PCR outlined in **Figure 8** and subsequent analysis using the Roche GS Junior System; and (2) the predicted likelihood of the STR-analysis to reveal STR variants and SNPSTRs.

2.1.2 Buffers and reagents

All buffers were prepared using Milli-Q water (Millipore Academic System fitted with a Quantem EX Cartridge) and autoclaved prior to use. All Tris-EDTA-based buffers were prepared using Sigma-Aldrich® Tris(hydroxymethyl)amino-methane (Tris) and Merck Ethylenediaminetetraacetic acid (EDTA). A solution of Tris-EDTA (TE) buffer (10 mM Tris, 1 mM EDTA) was prepared and acidified with Sigma Aldrich® hydrochloric acid (HCl) to a pH of 8.0. Tris-acetate-EDTA (TAE) buffer (40 mM Tris and 1.0 mM EDTA) was prepared and acidified with Sigma-Aldrich® Glacial acetic acid to pH 8.2. Tris-Borate-EDTA (TBE) buffer (89 mM Tris, 89 mM Boric acid, 2 mM EDTA) of pH 8.3 was also prepared. Ethidium Bromide from Sigma-Ardich® was dissolved in Milli-Q water at a concentration of 10 mg/mL and used for DNA visualisation. Lonza SeaKem® LE Agarose was used for all horizontal electrophoresis. Sigma Aldrich® molecular grade 40% $^{w}/_{v}$ Acrylamide/bis-Acrylamide (37:1 or 19:1), N,N,N',N'-tetramethylethylenediamine (TEMED) of purity

~99%, and molecular grade Ammonium Persulphate (APS) were used for all vertical electrophoresis.

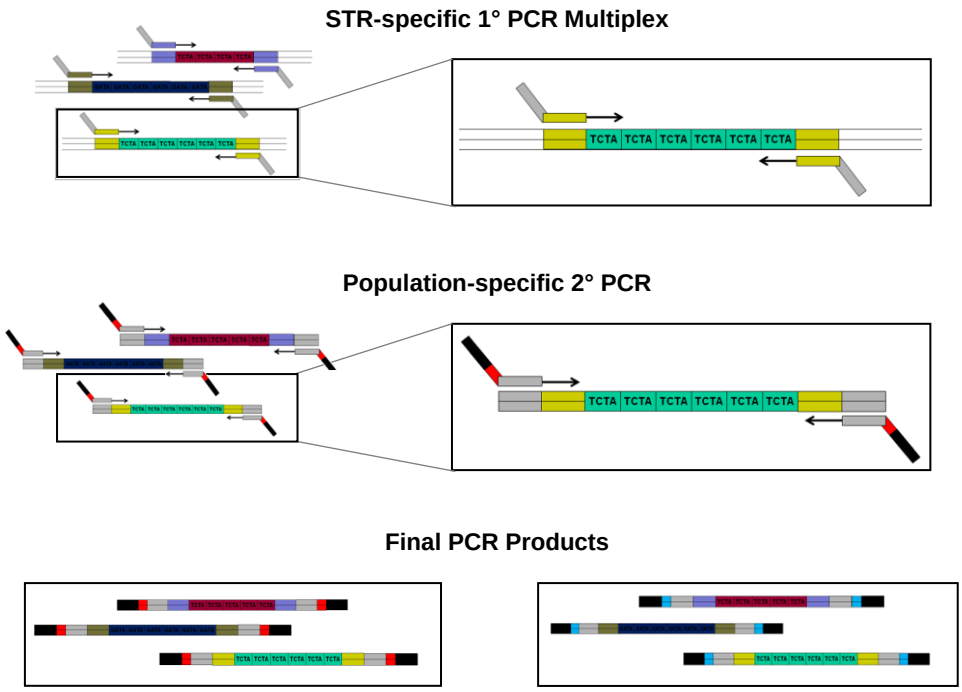


Figure 8: Schematic of the 2-step PCR protocol and the final amplicons used for 454 library preparation

The **STR-Specific 1° PCR Multiplex** makes use of STR-specific (**lime** in the enlarged schematic) M13-tailed (**grey**) primers. The **2° PCR** makes use of M13-specific primers labelled with population-specific MIDs (**red** in the enlarged schematic; and **red** or **blue** for the **Final PCR Products**) and fusion primers (**black**) required for 454 sequencing.

2.1.3 DNA quantitation by spectrophotometry and GelQuant.NET

DNA quantification and assessment of the success of DNA purification was achieved either by analysis of the absorbance spectrum from 220 to 330 nm using the ThermoScientific Nanodrop 2000 Spectrophotometer or by electrophoresis (see *2.1.4* and *2.1.5*) and subsequent analysis with GelQuant.NET software provided by biochemlabsolutions.com.

2.1.4 Genomic DNA extraction and purification

The Cell Projects Isohelix DNA Isolation Kit was used to extract genomic DNA, as per the manufacturer's instructions and including all optional steps. The quality and quantity of extracted DNA was assessed by agarose gel (0.7% $^{w}/_{v}$) electrophoresis (4 V/cm) in TAE, and spectrophotometry.

2.1.5 Polyacrylamide gel electrophoresis (PAGE) of amplicons

Singleplex amplicons were analysed using discontinuous, 4% $^{w}/_{v}$ stacking and 10% $^{w}/_{v}$ resolving, polyacrylamide (37:1 acrylamide:bis-acrylamide) gel electrophoresis in TAE at 4-5 V/cm, for 3-4 hours. Multiplex PCR products were also analysed using discontinuous PAGE, however, 19:1 acrylamide:bis-acrylamide and TBE buffer were used. Voltages of between 1.6 and 8 V/cm were applied for a minimum of 6 hours. All Acrylamide Gels were soaked in 0.15 µg/mL ethidium bromide for 15 minutes and visualised using long wave ultraviolet light with a 2-4 second exposure time.

2.1.6 Primer Synthesis

2.1.6.1 Selection, analysis and synthesis of 1° PCR Primers

The primer sequences (shown in **Table 8**), excluding the universal tails (M13), were obtained from literature (Oberacher *et al.*, 2008). Primers were assessed using the Basic Local Alignment Search Tool (blast.ncbi.nlm.nih.gov/Blast.cgi), FastPCR 6.3. and IDT Oligo Analyser 3.1 (http://eu.idtdna.com). Primers were synthesised by Integrated DNA Technologies (IDT) and purchased from WhiteSci Scientific (Pty) Ltd. Purification of primers was achieved by standard desalting.

Table 8: Short Tandem Repeat (STR)-specific primers used for 1° PCR amplification

Primer	Sequence (5´-3´)
vWA F	CGCCAGGGTTTTCCCAGTCACGACCCCTAGTGGATGATAAGAATAATCAGTATG
vWA R	TCACACAGGAAACAGCTATGACGGACAGATGATAAATACATAGGATGGATGG
D2S441 F	CGCCAGGGTTTTCCCAGTCACGACCTGTGGCTCATCTATGAAAACTT
D2S441 R	TCACACAGGAAACAGCTATGACGAAGTGGCTGTGGTGTTATGAT
D3S1358 F	CGCCAGGGTTTTCCCAGTCACGACACTGCAGTCCAATCTGGGT
D3S1358 R	TCACACAGGAAACAGCTATGACATGAAATCAACAGAGGCTTG
D21S11 F	CGCCAGGGTTTTCCCAGTCACGACATATGTGAGTCAATTCCCCAAG
D21S11 R	TCACACAGGAAACAGCTATGACGGTAGATAGACTGGATAGATAGACGA
D7S820 F	CGCCAGGGTTTTCCCAGTCACGACAACACTTGTCATAGTTTAGAACGAAC
D7S820 R	TCACACAGGAAACAGCTATGACTCATTGACAGAATTGCACCA
D13S317 F	CGCCAGGGTTTTCCCAGTCACGACTCTGACCCATCTAACGCCTA
D13S317 R	TCACACAGGAAACAGCTATGACCAGACAGAAAGATAGATAGATGATTGA

The coloured region of the primer sequence is specific for the target STR. The remaining region is the universal (M13) primer sequence. F and R indicate the forward and reverse primers respectively.

2.1.6.2 Design and synthesis of 2° PCR primers

Secondary PCR primer sequences were defined by the specifications of the Roche 454 Sequencing System Guidelines (*http://454.com/downloads/my454/applications-info/GuidelinesForAmpliconExperimentalDesign_Nov2012.pdf*). The sequences (displayed in **Table 9**) are made up of three distinct regions: A 5´ 25-mer required for bead hybridization, emPCR amplification, and amplicon sequencing (underlined in **Table 9**); a population-specific Multiplex Identifier (MID) (bold in **Table 9)**; and the 3´ template-specific M13 sequence. Primers were analysed using IDT Oligo Analyser 3.1 and purchased from WhiteSci Scientific (Pty) Ltd. Purification of primers was achieved by High Performance Liquid Chromatography (HPLC).

Table 9: Population-specific fusion primers used for 2° PCR amplification

Primer	Sequence (5´-3´)
Nguni F	CGTATCGCCTCCCTCGCGCCATCAG**ACGAGTGCGT**CGCCAGGGTTTTCCCAGTCACGAC
Nguni R	CTATGCGCCTTGCCAGCCCGCTCAG**ACGAGTGCGT**TCACACAGGAAACAGCTATGAC
ST F	CGTATCGCCTCCCTCGCGCCATCAG**AGACGCACTC**CGCCAGGGTTTTCCCAGTCACGAC
ST R	CTATGCGCCTTGCCAGCCCGCTCAG**AGACGCACTC**TCACACAGGAAACAGCTATGAC

The underlined 5´ 25-mer is necessary for bead hybridization, emPCR amplification, and amplicon sequencing. The **bold** nucleotides indicate the population-specific MID. MIDs for Nguni-specific and Sotho-Tswana (ST)-specific primers are Roche-defined MID 1 and 3 respectively. The 3´ region is template-specific to M13. F and R indicate the forward and reverse primers respectively.

2.1.7 Development of singleplex 1° PCR for amplification of target STRs

Target loci were amplified in singleplex using KAPA HIFI^{TM} HotStart ReadyMix as per manufacturer's instructions and using a reduced primer concentration of 0.2 µM each. The primers used were those depicted in **Table 8**. Cycling conditions entailed initial denaturation at 95 °C for 5 minutes; 25 cycles of denaturation at 98 °C for 20 seconds, annealing for 15 seconds, and extension at 72 °C for 15 seconds; followed by a final extension at 72 °C for 30 seconds. The success of each reaction was assessed based on the yield of the target amplicon, the presence/absence of primer dimers, and by the specificity of the reaction. Specificity was evaluated based on the absence of non-specific bands (i.e. bands greater or smaller than the expected size range of the STR of interest). The optimum annealing temperature for the primers was determined by amplification of the loci using gradient PCR. Annealing temperatures between 50 and 71 °C were investigated. The amount of template

DNA was optimised by preparing reactions with a range in quantity of DNA template, specifically 50, 80 and 100 ng.

2.1.8 Validation of the specificity of PCR amplification by Sanger sequencing

Following PAGE of the singleplex PCR products of Subject 0, individual alleles were excised and purified using the "crush and soak" method outlined in *2.1.10.4*. The extracted DNA was quantified as described in *2.1.3* and 10 ng of the PCR product was used as a template for a second round of amplification. PCR conditions and cycling parameters were as described in *2.1.7*, with the exception of the annealing temperature and the cycle number which were set to 64 °C and 35 cycles respectively. Gel purification and Sanger sequencing was performed by Inqaba Biotech™. The resultant forward and reverse sequences were aligned using MAFFT version 7 (online) using default settings and visualised using BioEdit Sequence Alignment Editor version 7.2.5. to create a consensus sequence. M13 tail regions were removed and sequencing errors in the STR-specific primer regions of the consensus sequences were corrected. The modified consensus sequences were then aligned against a database of known alleles using MAFFT and a Sequence Identity Matrix of the alignment was generated using BioEdit to enable allele calling.

2.1.9 Development and optimisation of 1° PCR multiplex

The optimised conditions, established for the singleplex PCR of the STRs, were used as a basis for the development of the multiplex PCR. These conditions included the optimal DNA concentration of 80 ng per 25 µL PCR reaction mixture, and optimised cycling conditions of 95 °C for 5 minutes; 25 cycles of denaturation at 98 °C for 20 seconds, annealing at 64 °C for 15 seconds, and extension at 72 °C for 15 seconds; followed by a final extension at 72 °C for 30 seconds. Primer combinations and relative primer concentrations were varied, as described by Schoske *et al.* (2003), some combinations of which are outlined in **Figures 13** and **14**. Initially, equimolar amounts of each primer were added to the reactions. In each case, the primer concentrations for every subsequent reaction were altered slightly. Primer concentrations for loci that were over-amplified, in comparison to other loci, were decreased; while primer concentrations for loci that were relatively under-amplified were increased. This continued until a multiplex that yielded bands of equivalent intensity, when electrophoresed on 19:1 acrylamide:bis-acrylamide gel as described in *2.1.5*.

During optimisation of the multiplex, the optimum template concentration and cycle number for the reaction was also investigated. Some of these variations are outlined in G4-G7 in **Figure 14**. Template quantities of 25 to 200 ng were used in samples with 25 cycles. The reactions with a decreased cycle number (15, 18 or 21 cycles) had an increased template quantity of 300 ng and an increased total volume of 50 µL (as opposed to 25 µL).

2.1.10 Purification of 1° PCR products

A number of purification methods were investigated for the purification of the 1° PCR products, with the aim of finding a method of purification that would remove all primer artefacts. In order to assess the success of any purification method, products were visualised using polyacrylamide gel electrophoresis as described in *2.1.5*.

2.1.10.1 PCR purification

PCR purification using the Thermo Scientific GeneJET PCR Purification kit, as per manufacturer's instructions but excluding the optional step for DNA smaller than 500 bp, was used to purify the 1° PCR from Subject 0. Purified amplicons were eluted in 25 µL of TE buffer.

2.1.10.2 Agencourt AMPure XP Beads

Agencourt AMPure XP beads at a variety of ratios (0.6, 0.8, 1.0, 1.2, 1.3, 1.4, 1.5, and 1.6 to 1 $^{v}/_{v}$ AMPure XP beads to PCR product), was used to purify the 1° PCR product from Subject i. The beads were used according to the manufacturer's instructions.

2.1.10.3 Agarose gel purification

The 1° PCR product from Subject 0, alongside 1 µL per mm (where the mm refer to the width of the gel well) Thermo Scientific GeneRuler 50 bp, was electrophoresed in TAE buffer at 5 V/cm in a 2% $^{w}/_{v}$ agarose gel containing 0.4 µg/mL ethidium bromide. Following electrophoresis, the gel was visualised using long wave UV light and the target bands (identified by comparison to the marker) were excised. The DNA was extracted and purified using the Thermo Scientific GeneJET Gel Extraction kit, as per manufacturer's instructions and including the optional step for DNA <500 bp. DNA was eluted in 20 µL of TE buffer.

2.1.10.4 Polyacrylamide gel purification: The "crush and soak" method

The 1° PCR product from Subject 0, alongside 0.07 µL per mm (where the mm refer to the width of the gel well) Thermo Scientific GeneRuler 50 bp, was electrophoresed as described in *2.1.5*. Following electrophoresis, the gel was visualised using long wave UV light and the target bands (identified by comparison to the marker) were excised. The excised gel was submerged in TE buffer for 20 minutes, whereafter the buffer was removed and replaced with 1:1 ($^{w}/_{v}$ gel:buffer) fresh TE buffer. Using the end of a melted-closed P1 tip, the polyacryalamide gel was crushed and the suspension was incubated overnight on a shaker at 37 °C. The suspension was centrifuged at 14 000 *g* for 1 minute and the supernatant retrieved. The DNA in the supernatant was concentrated using the Thermo Scientific GeneJET Gel Extraction kit, as per manufacturer's instructions and including the optional step for DNA <500 bp.

2.1.11 Optimisation of 2° PCR for addition of fusion primers

A stock of 1° PCR products for Subjects 0 and i-iv was prepared using the "crush and soak" method outlined in *2.1.10.4*. These purified products were used as the template DNA for 2° PCR optimisation using the KAPA HIFI^{TM} HotStart ReadyMix as per manufacturer's instructions and with a reduced primer concentration, the sequences of which can be viewed in **Table 9**. The success of each reaction was investigated by electrophoresis of the PCR products as described in *2.1.5* and was assessed based on the yield and specificity of the reaction (as in *2.1.7*). A number of conditions were analysed during optimisation of the 2° PCR, some of which are depicted in **Figure 16**. The optimum annealing temperature for the primers was determined by amplification of the loci using gradient PCR in which annealing temperatures of between 61 and 72 °C were investigated. Following annealing temperature optimisation, cycling parameters were adjusted slightly to include a 5 minute denaturation step at 98 °C; 20 cycles of 98 °C for 30 seconds, 64 °C for 15 seconds and 72 °C for 10 seconds; and a final elongation at 72 °C for 30 seconds. The optimum primer and template DNA concentration was investigated by preparing reactions with a range in quantity of primer and DNA template, specifically 0.06-0.2 µM and 1-2.6 ng respectively. The effect of adding DMSO (3 and 5% $^{V}/_{V}$) and increasing the cycle number to 25 was also investigated. Following optimisation of the 2° PCR amplification protocol, the use of template DNA that had either not been purified or that had been purified using PCR purification (as described in *2.1.10.1*) was also investigated.

2.1.12 Purification of the 2° PCR product

Secondary PCR products from Subjects iii and 0 were purified using agarose gel purification and the "crush and soak" method respectively (as outlined in *2.1.10.3* and *2.1.10.4*). The success of purification was assessed by electrophoresis of the purified products as outlined in *2.1.5.* Products were deemed successfully purified if no primer artefacts or non-target bands were present.

2.2 RESULTS

2.2.1 Selection of target loci

Before STR selection could commence, the total number of loci permitted by the experimental design, outlined in **Figure 7**, was defined: Since the desired depth of coverage for the sequencing run was 40x, the number of populations analysed was 2, and the number of individuals per population tested was 72 [40 x 72 x 2 = 5760], and given that the average Roche GS Junior sequencing run produces ~70 000 reads (http://my454.com/downloads/my454/applications-info/454SequencingSystem_GuidelinesforAmpliconExperimentalDesign_July2011.pdf), a maximum of 12 unique sequences per individual could be sequenced [70 000/5760 = 12]. Therefore, given that an individual may be heterozygous for all loci, the number of target STRs was limited to 6 [12/2 = 6].

The experimental design not only limited the number of loci that could be analysed; but also required that target loci be compatible, both with regards to primer sequence for 1° PCR multiplexing and with regards to allele size for 454 sequencing. Given that a range of primer sequences, which have been used for the multiplex PCR of STRs, are available on the STRbase (www.cstl.nist.gov/strbase), pools of compatible primers for the amplification of 6 STRs were readily available. The effect of the M13 tail on primer compatibility was, however, uncertain and is expanded on in *2.2.3*. STR allele-size compatibility was somewhat more difficult to ascertain since the amplicon size is linked directly to the primer sequence. For example, D8S1179 has reported amplicon size ranges of 123-175, 157-209 and 203-255 bp depending on which primer set is used for amplification. Allele-size is important for 454 sequencing because the technology is suited only to the sequencing of amplicons of 200-600 bp with maximum size ranges of 150 bp (http://my454.com/downloads/my454/applications-info/454SequencingSystem_GuidelinesforAmpliconExperimentalDesign_July20

11.pdf). Since the 2-step PCR, outlined in **Figure 8,** resulted in the addition of tails of 46 and 70 bp respectively (116 bp in total), and since all STRs recorded in the STRbase (www.cstl.nist.gov/strbase) are 80-400 bp, all core STRs were considered suitable for analysis by 454 sequencing. The STRs that could be analysed together, however, were limited by the maximum size range of 150 bp required to avoid PCR and sequencing bias.

Bearing in mind the limitations with regards to the total number of target STRs, and both primer and amplicon-size compatibility, the STRs multiplexed by Pitterl *et al.* in 2008 (indicated by the red box in **Table 5**) using the primer sequences published by Oberacher *et al* (2008) were identified as potential target STRs. These loci were selected because their primer sequences were not only compatible, but also gave rise to reduced-length PCR products which exhibited a maximum allele size range of 157 bp, only slightly higher than the recommended threshold. **Figure 9** shows the possible allele size ranges for the STRs included in the Pitterl multiplex. TPOX however, although part of the Pitterl multiplex, was excluded from the figure and from the loci selection process due to its low P_I value (discussed in *1.2.2*).

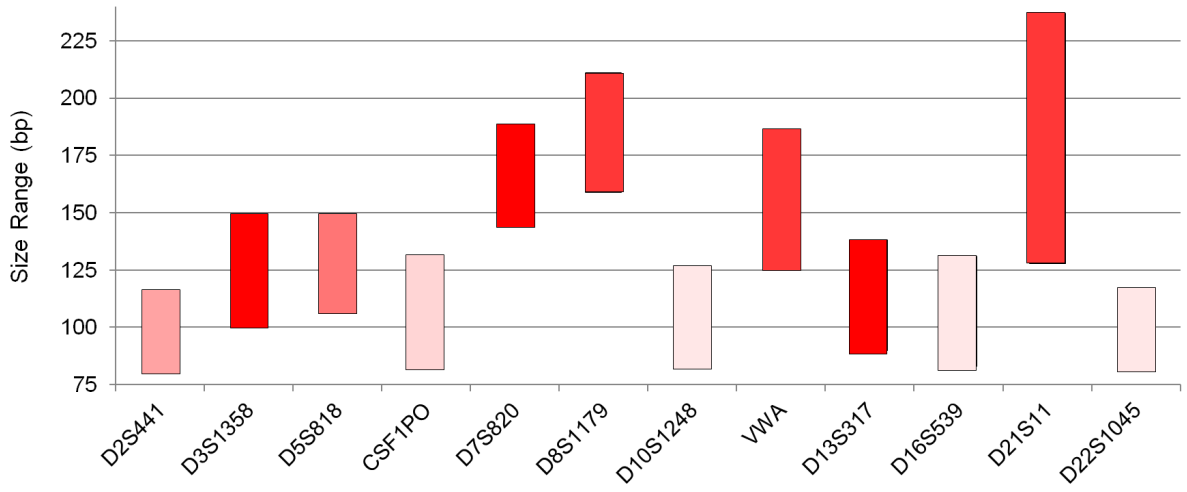


Figure 9: Possible allele-size ranges of STRs multiplexed by Pitterl *et al.* (2008), excluding the locus TPOX

Darker shades of red are indicative of higher predicted likelihood to present novel STR variants, as determined by the data in **Table 10.** The known alleles were determined from the STRbase (www.cstl.nist.gov/strbase) and the possible allele sizes were determined by linking the known alleles to the example alleles presented by Oberacher *et al.* (2008).

The STRs amplified by the Pitterl multiplex were also of interest because they are included in most commercial DNA profiling kits and National DNA databases (evidenced in **Table 1**). D8S1179, vWA, D3S1358 and D21S11, for example, form part of the core loci of all

countries considered; and, with the exception of the miniSTRs D2S441, D10S1248 and D22S1045, all loci included form part of the South African extended core loci.

From these STRs, the six loci D2S441, D3S1358, D7S820, vWA, D13S317 and D21S11 were selected for analysis. These loci were selected based on the predicted likelihood of sequence analysis revealing novel STR variants and SNPSTR data. These predictions were made based on analysis of the mutation rates and the numbers of known and variant alleles for each STR with higher values considered more likely to present variant alleles (values obtained from the STRbase, www.cstl.nist.gov/strbase); the presence of SNPs within or associated with the STRs (obtained from HapMap, http://hapmap.ncbi.nlm.nih.gov/); and the results of relevant literature (described in **Table 5**), all of which are summarised in **Table 10**. Loci with higher predicted likelihoods of revealing novel STRs are coloured in incrementally darker shades of red in the **Figure 9**.

Table 10: Summary of variance-predictive data for STRs multiplexed by Pitterl *et al.* (2008)

Loci	STRbase			SNPs		Literature
	MR $x10^{-2}$%	Alleles	Variants	Associated SNPs[1]	African* Data	PD increase
D2S441	-	14	6	-	-	Yes[1]
D3S1358	12	27	30	-	-	**Yes**[1,2,3,4,6]
D5S818	11	16	20	rs25768:T/C	Present	Yes[1,3,5,6]
CSF1PO	16	23	22	-	-	Yes[3]
D7S820	10	32	26	rs7789995: T/A rs7786079: T/G	Present Present	**Yes**[1,3,4]
D8S1179	14	19	24	-	-	Yes[1,2,3,4,6]
D10S1248	-	12	0	-	-	None
vWA	16	28	20	-	-	Yes[1,2,3,6]
D12S391	-	23	14	-	-	n/a
D13S317	14	20	18	rs9546005:A/T	None**	**Yes**[1,3,4,5,6]
D21S11	19	92	42	-	-	Yes[1,2,3,5,6]
D22S1045	-	13	0	-	None	None

Mutation Rates (MR), allele and variant numbers were obtained from the STRbase (www.cstl.nist.gov/strbase).

SNPs within the amplicons were taken from [1] Oberacher *et al.* (2008) and population data for these SNPs were obtained from HapMap (http://hapmap.ncbi.nlm.nih.gov/). *African population data were obtained from African Americans living in Southwest U.S.A, Massai individuals from Kenya, and Yorubans from Nigeria for rs7789995 and from Yorubans alone for rs25768 and rs7786079. **This SNP has no population data whatsoever.

Literature references were as follows [1] Oberacher *et al.*, 2008; [2] Pitterl *et al.*, 2010; [3] Planz *et al.*, 2009, with **bold** references indicating a dramatic increase in power of discrimination for African populations; [4] Divne *et al.*, 2010; [5] Fordyce *et al.*, 2011; [6] Van Neste, 2012.

D3S1358, D7S820 and D13S317 were identified as the loci most likely to reveal allele variants. This prediction was made based on the high mutation rates and number of alleles of

these loci, as well as the reported STR variants reported for the loci when analysed using ICEMS, pyrosequencing and NGS. Of these loci, those that exhibited enhanced PDs for individuals of African descent were of particular interest (**Figure 6**). D7S820 and D13S317, furthermore, were of particular interest due to their associated SNPs (for which data for African populations were absent in some cases). D2S441 was the fourth locus selected for analysis despite its exclusion from the South African core loci. This locus was of interest because it is a novel miniSTR with an, as yet, unknown mutation rate. Despite its comparatively young age in the field of forensics, 6 variants and an increase in PD by Oberacher *et al.* (2008) have been reported, suggesting that 454 sequencing may reveal additional novel variants. The remaining two target STRs were chosen from between D5S818, D8S1179, vWA and D21S11; the loci second most likely to exhibit STR variants as per **Table 10**. vWA and D21S11 were selected due to their higher mutation rates and number of reported alleles.

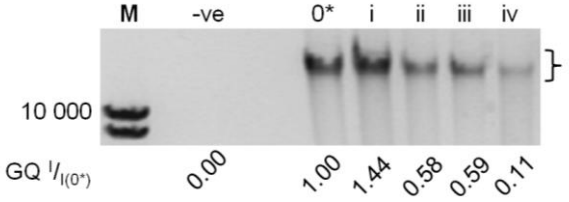
2.2.2 Extraction and purification of template DNA for PCR optimisation studies

In order to obtain material for the optimisation of the PCR reactions, DNA from five Subjects, namely Subjects 0, i-iv, was extracted using the Isohelix DNA extraction kit. Subject 0 DNA was DNA extracted from myself, a female of Western European descent. This DNA was used in the optimisation studies for a variety of reasons: Firstly, because I do not fall into either target population group, my profile serves to show that the amplification reactions are not specific to any given population group; secondly, my DNA was readily available and could be extracted as needed and; thirdly, during collection and amplification of DNA samples for the Next Generation Sequencing study, if contamination were to occur, the DNA most likely to be the contaminant would be DNA from the sampler and tester, i.e. myself. It was therefore important to know my profile as a control for contamination. Subjects i-iv were unrelated isiXhosa individuals (3 males and 1 female). These subjects, therefore, fell into one of the population groups (Nguni) of interest. Consequently, differences between their profiles indicated that variances between profiles are individual-specific and not population-specific, although this is to be expected since STRs are well established individuation markers (Jobling and Gill, 2004). These subjects were also selected for the optimisation studies because their DNA was readily available for re-sampling if necessary.

A

Subject	[DNA] (ng/µL)	$^{260}/_{280}$	$^{260}/_{230}$	$^{[DNA]}/_{[DNA(0^*)]}$
-ve	-16.1	1.3	-3.55	-
0	485.7	1.59	0.7	2.52
0*	192.1	1.75	0.91	1.00
i	681.8	1.53	0.68	3.55
ii	394.5	1.62	0.7	2.05
iii	298.9	1.7	0.83	1.56
iv	452.8	1.6	0.67	2.36

B



C

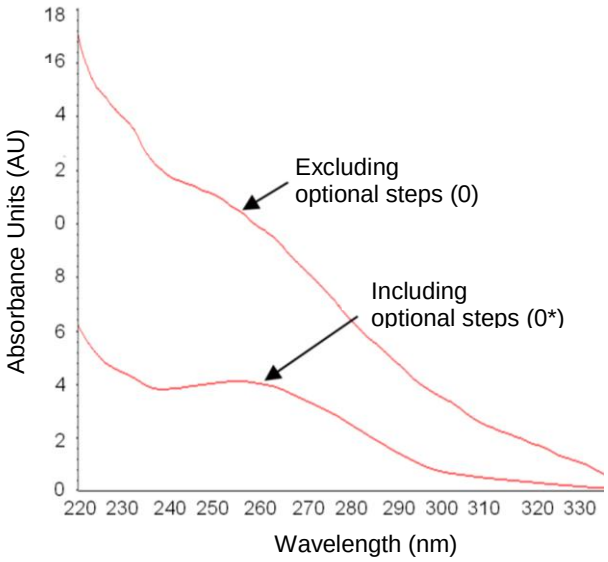


Figure 10: Assessment of the Isohelix DNA Extraction kit using buccal swabs of Subjects 0, and i-iv.

Subject DNA was extracted either excluding (Subjects 0 and i-iv) or including (Subject 0*) the optional steps outlined in manufacturer's manual. The negative control (-ve) was prepared by following the Isohelix protocol on the DNA Stabilisation Solution (the solution used to store buccal swabs). **A** is a table constructed after measuring the absorbance spectrum from 220-330 nm using the ThermoScientific Nanodrop 2000. The 260/280 and 260/230 ratios of between 1.7 and 1.9 and ~1 respectively are indicative of pure DNA. The relative concentration of DNA extracted from Subjects 0 and i-iv to that of Subject 0* is given. **B** is an agarose gel (0.7% $^{w}/_{v}$ in TAE buffer) electrophoresis (4V/cm) of DNA extracted from Subject 0 (*including the optional extraction steps) and Subjects i-iv (extracted excluding the optional steps). The bracket indicates the high molecular weight genomic DNA extracted from the subjects. Fermentas MassRuler HighRange DNA Ladder was loaded into lane **M.** Bands were analysed using GelQuant.NET software provided by biochemlabsolutions.com. GQ $^{I}/_{I(0^*)}$ is the ratio of the signal intensity of each band to that for Subject 0* obtained using GelQuant is displayed below the figure. **C** shows the spectrograms for DNA of Subject 0 extracted both excluding and including (*) the optional extraction steps.

Two DNA extraction methods, using the Isohelix DNA Extraction Kit, were explored on samples obtained from Subjects 0 and i-iv. The first method, performed on the DNA extracted from all the subjects, aimed to minimise the extraction time by omission the optional steps in the manufacturer's protocol; while the second method, performed initially only on the sample extracted from Subject 0 (Sample 0*), included an extra centrifugation to remove undissolved particulates and a final incubation step at 80 °C. **Figure 10** demonstrates that the additional steps were necessary in order to obtain an accurate measurement of the concentration of DNA in the extraction. **Figure 10 A** shows that when the additional steps were included for the sample collected from Subject 0, the apparent nucleic acid

concentration decreased from 485.7 ng/µL (Sample 0) to 192.1 ng/µL (Sample 0*). **Figure 10 A and C** demonstrated that the optional steps were important for the removal of contaminants. The spectrograms obtained for Subject 0 show that exclusion of the optional steps results in an atypical spectrogram. Ideal absorbance ratios for wavelengths of 260/280 and 260/230 for pure DNA are between 1.7 and 1.9, and ~1 respectively (Siwach and Singh, 2007). Where optional steps were included, the 260/280 and 260/230 absorbance ratios were ideal at 1.75 and 0.91; where optional steps were excluded, the mean 260/280 ratio was 1.61 and the mean 260/230 ratio was 0.72. These low ratios are indicative of contamination with proteins and/or phenol (maximum absorbance at 280 and 270 nm respectively) and ethanol (maximum absorbance at 230 nm) (Siwach and Singh, 2007).

As indicated by comparison of the DNA concentration for Sample 0 and 0* (2.52 times decrease in concentration), the nucleic acid concentrations reported in **Figure 10 A** for Subjects 0 and i-iv, extracted excluding the optional steps, were likely exaggerated. This was further illustrated in **Figure 10 B** which shows the electrophoresis of equivalent volumes of DNA from Subject 0 (Sample 0*), extracted including the additional centrifugation and incubation, and Subjects i-iv, extracted excluding these steps. The resultant bands were compared using GelQuant. The intensity of the electrophoresed DNA, which is an indication of DNA concentration, was found to be 1.44, 0.58, 0.59 and 0.11 times that of Subject 0* for Subjects i, ii, iii and iv respectively. These values are on average 2.88 times lower for Subjects i-iii than those observed in **Figure 10 A**, and 21 times lower than that observed for Subject iv. The discrepancy in estimated DNA concentration when compared to Subject 0* was consistent when considering Sample 0 and Samples i-iii, however the discrepancy for Subject iv was to a far greater degree. The reason for this is unclear, although it may be a result of ineffective sample loading prior to electrophoresis caused by ethanol contamination.

In order to ensure consistent results during the optimisation of the primary PCR reaction, accurate quantification of template DNA was necessary. Consequently, all DNA extractions made use of the optional steps outlined in the manufacturer's instructions. It is important to note, however, that the genomic DNA extracted both with and without the optional steps was of good quality. All the bands in **Figure 10 B** were of a high molecular weight, and were sharp and compact. Furthermore, PCR amplification was not compromised when template DNA extracted excluding the optional steps was used (data not shown). In cases where accurate quantification of DNA was less important, therefore, additional extraction steps may be excluded in favour of reducing the time per extraction.

2.2.3 Analysis of loci-specific primers to determine suitability for multiplexing

The primers used in the 1° PCR reaction were comprised of two parts, namely the universal tail region (M13) and the loci-specific region obtained from literature (Oberacher *et al*., 2008). In order to ensure successful development of the multiplex STR PCR reaction, two parameters needed to be considered: Firstly, whether the specificity of the primers for the target STR was retained once the universal tail had been added and; secondly, whether the primers were compatible i.e. whether they were suitable for multiplexing.

Table 11: Predicted specificity of primers for the target loci, and general primer characteristics and compatibility

Primer	Min E-value		T_m (°C)	CG (%)	Strongest Folding T_m (°C)	Strongest T_m (°C) of cross dimers
	Incl. M13	Excl. M13				
vWA F	3e-07	9e-08	68.2	48.1	38.9	4.0 with D13 F
vWA R	3e-07	9e-08	66.1	42.3	34.6	4.0 with D13 F
D2S441 F	0.003	6e-04	69.5	51.1	45.4	n/a
D2S441 R	0.047	0.008	67.1	45.5	54.0	4.0 with D7 F
D3S1358 F	0.046*	0.059*	71.7	58.1	40.6	4.0 with D13 F
D3S1358 R	0.044	0.023	65.0	42.9	33.0	5.3 with D7 F
D21S11 F	0.013	0.009	69.0	52.2	51.5	4.0 with D13 R
D21S11 R	6e-05	7e-05	65.6	43.8	29.3	4.0 with D7 F
D7S820 F	6e-05	2e-05	68.0	48	53.0	5.3 with D3 R
D7S820 R	0.044	0.023	65.8	42.9	33.0	4.0 with D7 F and D2 R
D13S317 F	0.003	0.023	70.6	56.8	44.7	4.0 with D13 R, D7 F, D3, vWA R
D13S317 R	1e-05	4e-06	64.6	38.8	32.0	4.0 with D3

F indicates the forward primer while R indicates the reverse primer specific to a given locus. M13 is a universal primer sequence that was added as a 5´-tail to primers. Primers were analysed using BLAST, IDT Oligo Analyser 3.1. and FastPCR 6.3.

E-values were obtained using BLAST. The closer the E-value is to 0, the less likely it is that the match arose by chance i.e. the more significant the match

* The minimum E-value did not represent the target locus

To assess the specificity of the primers, the Basic Local Alignment Search Tool (BLAST) was used to screen the sequences. The Expect (E)-values, obtained following BLAST analysis of the primers both with (incl. M13) and without (excl. M13), are presented in **Table 11**. E-values closer to 0 indicate matches that are less likely to have arisen by chance and that are, therefore, more significant. E-values for the primers linked to M13 were, on average, higher than those obtained for the primers without the universal tail. E-values of M13 primers were between 1.44 and 5.88 times greater (D21S11 F and D2S441 R respectively) than those of the purely locus-specific primers. This was expected since the

M13 sequences yield high minimum E-values of 2.8 (forward sequence) and 0.47 (reverse sequence) when searched in isolation. Exceptions included D13S317 F, D3S1358 F and D21S11 R whose un-tailed primer E-values were 7.67, 1.17 and 1.28 times lower than the tailed-primers. D13S317 and D3S1358 are among the shortest of the primers analysed. This may account for the higher E-value observed for the primers lacking the M13 tail since shorter sequences are likely to have higher E-values; the length of the sequence is a parameter taken into account during calculation of the E-value (*Frequently Asked Questions*, http://www.ncbi.nlm.nih.gov/blast)

Primers, both with and without universal M13 tails, yielded minimum E-values and corresponding highest 'hits' for the target chromosomal regions. One exception was that for the forward primer of D3S1358. In the case of the D3S1358 primer linked to M13, the target locus was the 5th hit with an E-value of 0.18. It was, however, the only hit that yielded a Query Coverage and Sequence Identity of 100% (the top hit was obtained for a match on chromosome 2 which yielded 97% query coverage with 96% sequence identity). Without the universal M13 tail, the target locus for the D3S1358 primer was the second hit, while the top thirteen hits all yielded identical E-values (0.059), Max Scores (38.2) and Query Coverages (100%). Hits were found on regions of chromosomes 1, 3, 6, 14, and 19. Nevertheless, this primer has been successfully used for the selective amplification of D3S1358 (Oberacher *et al*., 2008).

Primer compatibility is vital in order to ensure successful multiplex PCR. Compatible primers have similar annealing temperatures and GC contents, do not form hairpin structures, and do not interact with themselves or other primers in the multiplex to form dimers (Butler, 2011). To assess the suitability of the primers used in this study for multiplex PCR, IDT Oligo Analyser 3.1. and FastPCR 6.3 were used. The results of these analyses are presented in **Table 11**. Theoretical melting temperatures were found to range from 64.6 to 71.7 °C (a difference of 7.1 °C). GC content ranged from 38.8% to 58.1%. Ideally melting temperatures should be within 5 °C of each other for multiplex PCR and GC content should range from 40% to 60% (Butler, 2011). The melting temperature range and GC content values obtained for the primer set were deemed sufficiently similar to the ideal values to continue with *in situ* annealing temperature studies (See section 4.2.3). The strongest folding melting temperature of any intra-primer secondary structure was 54.0 °C (observed for the reverse primer of D2S441). Likewise, the maximum melting temperature of any cross-dimer was found to be 5.3 °C (observed for the cross-dimer between the forward primer of D7S820 and reverse

primer of D3S1358). Since the melting temperature of the primers were all at least 10 °C higher (minimum of 64 °C) than the melting temperatures for these intra-primer secondary structures, self- and cross-dimers, these interactions were likely to be destabilised during PCR and, therefore, were predicted to have no effect on the multiplex optimisation process.

Following bioinformatic analysis of the primers, the complete primer set was determined to be suitable for use in the multiplex optimisation studies. Primer specificity for the target locus was not compromised by the universal M13 tail region; primers had sufficiently similar melting temperature and GC contents; and primers were unlikely to form hairpins, self- or cross-primers during PCR.

2.2.4 Singleplex PCR development to predict multiplex-compatibility of primers

Analysis of the loci-specific primers suggested that they were both specific to their target loci, and compatible for use in multiplex PCR. To confirm these theoretical findings, an investigation was performed to determine whether each locus, in singleplex and under identical conditions, could be specifically amplified using the M13 tailed-primers. Preliminary studies were performed using the inexpensive KAPA HotStart (to prevent secondary structure formation during PCR set-up and subsequent non-target amplification) and the standard KAPA ReadyMix (to ensure inter-reaction consistency). Since the overall aim of this study was to identify SNPs and other STR variants, sequence accuracy during amplification was vital and subsequent reactions were, therefore, performed using the high-fidelity kit, KAPA HiFi HotStart ReadyMix. This kit was used because its polymerase has a 3´-5´ exonuclease activity and had the lowest error rate of all B-family DNA polymerases (1 per 3.6 x 10^6 nucleotides) (Technical Data Sheet, KAPA Biosystems). Different loci that could be amplified under identical conditions using the KAPA HiFi HotStart ReadyMix were taken to be promising candidates for multiplex PCR development.

The preliminary studies for the singleplex amplification of target STRs were performed using the KAPA HotStart self-made mastermix (1x buffer, 1.5 mM $MgCl_2$, 0.1 mM each dNTP, 0.1 µM each forward and reverse primer, 1 U polymerase and 384 ng DNA from Subject 0 and ii each) and the KAPA HiFi ReadyMix (150 ng template DNA from Subject 0, i and ii). The self-made mastermix was used to determine the approximate annealing temperature of the primers using a wide range of annealing temperatures (50, 55.9, 64, 67.2 and 72 °C) and with cycling parameters otherwise consistent with those described in the Technical Data

Sheet. The PCR products were visualised using both agarose (1.6-2.5% $^{w}/_{v}$) and polyacrylamide (10% $^{w}/_{v}$) gels with TAE buffer. **Figure 11** enables the comparison of agarose to PAGE gels for a representative set of amplified STRs, namely vWA, D2S441 and D3S1358 from Subjects 0 and ii. The profiles that resulted on the agarose gel (**Figure 11 A**) were less informative than those of the polyacrylamide gel (**Figure 11 B**) as it was impossible to differentiate inter- and intra-individual alleles on the agarose. Both gels, however, revealed that annealing temperatures above 64 °C prohibited annealing (no bands of the expected size range were present, data not shown) whereas temperatures below 64 °C resulted in an increased number and intensity of bands outside of the expected size range for the loci (all bands not indicated by the brackets in **Figure 11**). This is particularly notable for D3S1358 amplified using an annealing temperature of 50 °C.

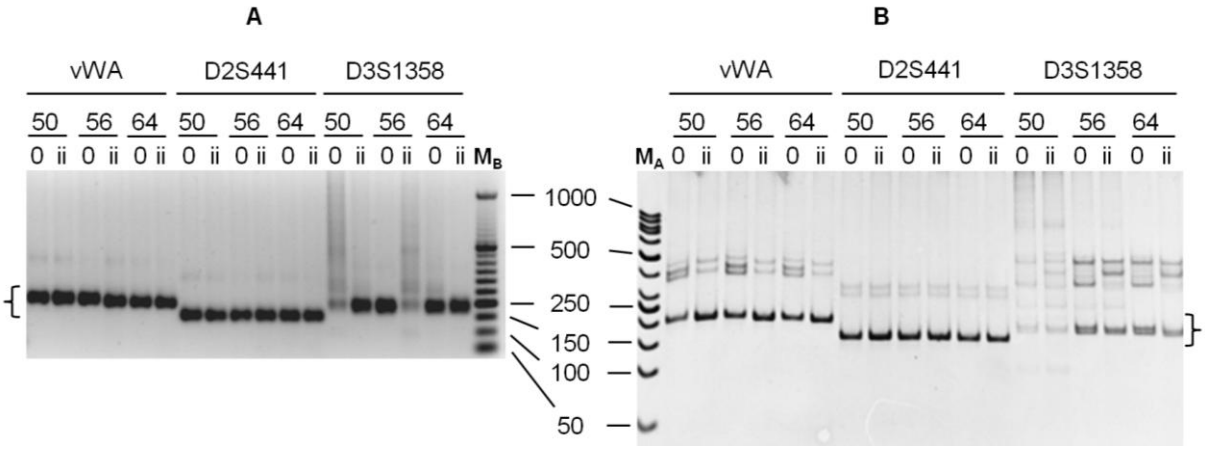


Figure 11: Preliminary annealing temperature optimisation for vWA, D2S441 and D3S1358 amplified using the KAPA self-made mastermix

STRs from Subject 0 and ii were amplified using annealing temperatures of 50, 55.9 and 64 °C and electrophoresed on **A** agarose (1.6% $^{w}/_{v}$, 4V/cm) and **B** polyacrylamide (10% $^{w}/_{v}$ 37.5:1 acrylamide:bisacrylamide, 5 V/cm) in TAE. The markers used in each gel were $\mathbf{M_A}$ (O'Range Ruler 50 bp DNA ladder at 1 µL/mm) and $\mathbf{M_B}$ (Thermoscientific GeneRuler 50 bp at 1 µL/mm). Bands marked by the brackets fall within the expected size range of each STR.

To ensure inter-reaction consistency, the second KAPA kit, namely the KAPA Readymix, was used. The reaction conditions were altered from those of the KAPA HotStart in an attempt to prevent non-specific amplification: The extension time and cycle number were decreased (from 60 to 30 seconds and 35 to 26 cycles respectively). This, however, was unsuccessful in eliminating the non-specific amplification and primer artefacts were present in all samples. Thereafter, the template and primer concentrations were decreased (from 384 to 150 ng, and 0.2 to 0.12 µM). This resulted in the specific amplification of the target

amplicons, however the PCR product was low in yield and primer dimers were still present (data not shown).

Having obtained conditions that enabled the specific amplification of the loci, albeit with primer artefacts, a second round of singleplex PCR optimisation was commenced using the KAPA HiFi HotStart ReadyMix. All reactions made use of a reduced primer concentration (0.2 µM instead of the 0.3 µM suggested in the Technical Data Sheet) in an attempt to avoid amplification of the primers. Two optimisation studies were performed; namely optimisation of the annealing temperature (**Figure 12 A**) and optimisation of the template concentration (**Figure 12 B**). Since the preliminary studies suggested that the minimum optimum annealing temperature that resulted in specific amplification of all the loci was 64 °C, the temperatures used in the second round of optimisation were 62.2, 64 and 65.3 °C.

During this optimisation study, 80 ng of template DNA was used because it fell into the upper range of the recommended amount of genomic DNA to add to PCR reactions (the Technical Data Sheet for the ReadyMix suggested 10-100 ng genomic DNA). Again, the optimum annealing temperature was found to be 64 °C. At temperatures above 64 °C amplification was negligible, and for loci D2S441 and D21S11 it was undetectable. While amplification with an annealing temperature of 62.2 °C produced higher amplicon yields than at 64 °C for all loci, a non-target band was observed at ~250 bp for D3S1358 (outlined in **Figure 12 A**). The presence of this non-target band prevented the use of the lower annealing temperature in further PCRs since the ideal multiplex would enable the amplification of all loci.

The amount of template DNA for the singleplex PCR was thereafter optimised. This was achieved by using a range in template DNA quantity (namely 50, 80 and 100 ng of genomic DNA) and comparing the amplification results (**Figure 12 B**). For all loci, 50 ng of DNA resulted in the lowest level of amplification for the STRs. Varying results were observed for 80 and 100 ng samples. vWA, D21S11 and D7S820 exhibited the highest yield when 80 ng of template DNA was used. D21S11 and D7S820 exhibit very low levels of amplification when 100 ng of DNA were used. The loci D2S441, D3S1358 and D13S317, conversely, exhibited the highest yield when 100 ng of template DNA was used. D13S317, in particular, is notable because as the template DNA concentration increased, so the presence of a primer dimer at ~80 bp decreased. While the use of 100 ng of template DNA resulted in the complete disappearance of the dimer for D13S317, because the levels of amplification were

negligible for D21S11 and D7S820 using this DNA concentration, 80 ng genomic DNA was selected for use in the initial multiplex development studies.

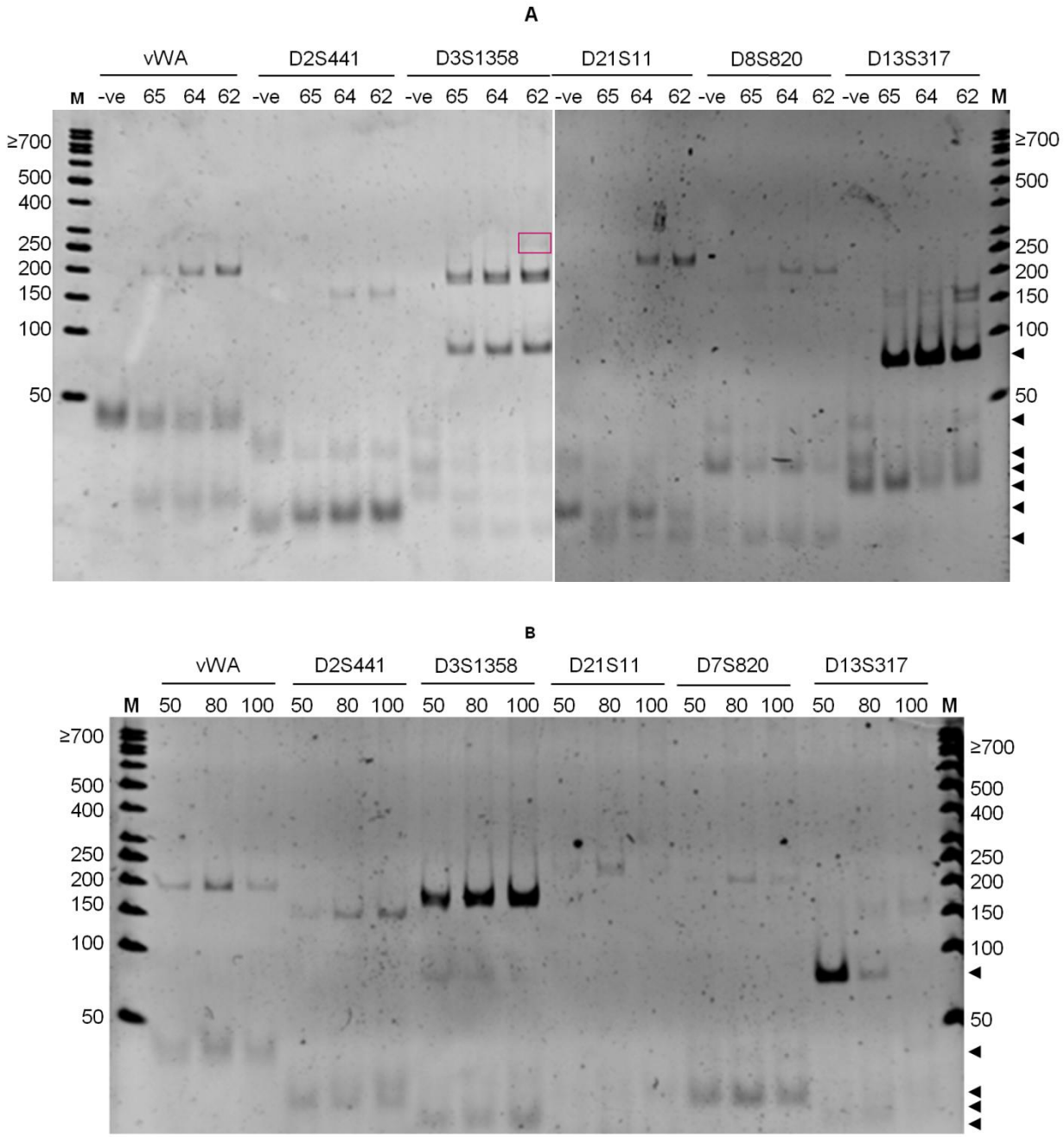


Figure 12: Identification of the annealing temperature (**A**) and template DNA concentration (**B**) that enable specific amplification of target STRs

STRs from Subject 0 amplified using the KAPA HiFi HotStart ReadyMix and visualised by polyacrylamide (4% $^{w}/_{v}$ stacking, 10% $^{w}/_{v}$ resolving 37.5:1 acrylamide:bisacrylamide) gel electrophoresis (5 V/cm) in TAE. The marker used for size determination (**M**) was the Thermoscientific GeneRuler 50 bp (1 µL/mm). ◄ indicates the primer artefacts. The annealing temperatures explored were 62.2 °C, 64.0 °C, and 65.4 °C; indicated in **A** by 62, 64 and 65 respectively. The pink box in **A** indicates the non-specific band observed for D2S1358.The DNA quantities explored in **B** were 50, 80 and 100 ng of template DNA genomic DNA.

2.2.5 Sanger sequencing to validate specificity of the amplification

Before commencing with multiplex development, the specificity of the PCR reactions was verified by Sanger Sequencing of the amplicons. Separate alleles were purified from 37:1 acrylamide-bisacrylamide gels following electrophoresis and re-amplified prior to sequencing. The sequences obtained are displayed in **Table 12** and the general characteristics of the alleles are summarised in **Table 13**. Amplicons were sequenced in both the forward and reverse directions and were aligned using MAFFT (accessed on http://mafft.cbrc.jp/alignment/server/ and run using default settings) to create a consensus sequence using BioEdit Sequence Alignment Editor version 7.2.5. A notable exception was vWA which, although apparently homozygous following electrophoresis on the 37:1 acrylamide:polyacrylamide, resulted in two distinct forward and reverse sequences. While these sequences included nucleotide errors (indicated in **Table 12** by the italic letters) in the primer sequences, two alleles could be resolved since the region flanking the STR prior to the primer binding site, demarcated the start and end of each allele (Oberacher *et al.*, 2008).

The remaining sequences were of generally high quality with little ambiguity in the sequencing chromatogram. Some sequencing errors (indicated in **Table 12** by the italic letters for substitutions and underscores for deletions) were, however, noted in the primer binding sites of the STRs. More specifically, a single substitution and a single deletion were observed in the primer binding sites for the larger alleles of D21S11 and D13S317 respectively. The larger allele of D3S1358 also exhibited 2 substitutions in the forward and 7 substitutions in the reverse primer binding sites. These errors, however, were not necessarily unexpected because errors near the ends of sequences are not uncommon when using Sanger Sequencing, especially when large primers are used (recommended primer sizes are 18-23 bp, http://www.lifesciences.sourcebioscience.com/). Furthermore, the errors were not present in the different alleles of the same locus and it is unlikely that they were a consequence of inaccurate primer synthesis.

Table 12: Sanger sequencing results to verify the specificity of PCRs used to amplify the STRs of Subject 0

Locus	Sequence
vWA[1]	TCGCCCGCCAGGGTT*CAGTCA*GTCCCGACCCCTGGTGGA*AGGA*AAGAACAATCAGTA*AG*TGA*G*TTGGATTGA**TCTATCTGTCTGTCTGTCTRTCTATCTATCTATCTATCTATCTAT CTATCTATCTA**TCC**T**ATCTATCCAT
	*AA*TTGGATTGA**TCTATCTGTCTGTCTGTCTGTCTATCTATCTATCTATCTATCTATCTATCTATCTATCATCTATCTA**TCYATCTATCCATCCATCCTATGTATTTATCAACTGTCCGTCA TACCTATTTCCTGCGTGA*ATRAA*
D2S441	CGCCAGGGTTTTCCCGTCACGACCTGTGGCTCATCTATGAAAAMA**TCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATT TATCTATCTA**TATCATAACACCACAGCCACTTCGTC ATAGCTGTTTCCTGTGTGA
D3S1358	*T*GCCAGGGTTTT*A*CCAGTCACGACACTGCAGTCCAATCTGGGTGACAG**C**AGC_AGACCCTGTCTCA**TAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGAC AGACAGATAGA**TACATGCAAGCCTCTGTTGATTTCTTGTC*TTATTT*GTTGCCT*TATT*GA
	CGCCAGGGTTTTCCCAGTCACGACACTGCAGTCCAATCTGGGTGACAGAGCAAGACCCTGTCTCAT**AGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATA GATAGACAGACAGATAGATAGATAGAT**ACATGCAAGCCTCTGTTGATTTCATGTCATAGCTGTTTCCTGTGTGA
D21S11	CGCCAGGGTTTTCCCAGTCACGACATATGTGAGTCAATTCCC*AAAG*TGAATTGCCT**TCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTAT CTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTA**TCGTCTATCTATCCAGTCTATCTACCGTCATAGCTGTTTCCTGTGTGA
	CGCCAGGGTTTTCCCAGTCACGACATATGTGAGTCAATTCCCCAAGTGAATTGCCT**TCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTAT CTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATATCTA**TCGTCTATCTATCCAGTCTATCTACCGTCATAGCTGTTTCCTGTGT GA
D7S820	CGCCAGGGTTTTCCCAGTCACGA*C*AACACTTGTCATAGTTTAGAACGAACTAAC**GATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATA**GACAGATTGATAGT TTTTTTTAATCTCACTAAATAGTCTATAGTAAACATTTAATTACCAATATTTTGGTGCAATTCTGTCAATGAGTCATAGCTGTTTCCTGTGTGA
D13S317	CGCCAGGGTTTTCCCAGTCACGACTCTGACCCATCTAACGCC_ATCTGTATTTACAAATACAT**TATCTATCTATCTATCTATCTATCTATCTATCTATCA**ATCAATCATCTATCTATCTT TCTGTCTGGTCATAGCTGTTTCCTGTGTGA
	CGCCAGGGTTTTCCCAGTCACGACTCTGACCCATCTAACGCCTATCTGTATTTACAAATACAT**TATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCA**ATCA ATCATCTATCTATCTTTCTGTCTGGTCATAGCTGTTTCCTGTGTGA

Sequencing was performed by Inqaba Biotech in the forward and reverse directions, with the exception of vWA which was sequences only in one direction. Forward and reverse sequences were aligned using MAFFT v 7 and visualised using BioEdit v 7.5.2 to obtain a consensus sequence. Underlined regions correspond to the primer binding sites; **bold** regions are the STR units. Known SNPs (A/T and T/G) are highlighted. _ and *italic* indicate deletions and insertions respectively when compared to the reference sequences.

Some variations, when compared to the sequences recorded by Oberacher *et al.* (2008), were present in the flanking regions of the STRs. The cause of these variations, whether due to sequencing errors or variations in the allele itself, was uncertain. It is likely, however, that they were a result of sequencing errors since there was ambiguity in the sequencing chromatograms for both vWA and the smaller allele of D3S1358, the alleles which exhibited single nucleotide substitutions and/or deletions. All other flanking regions and STR repeat sequences were consistent with those recorded by Oberachter *et al.* (2008). D7S820 and D13S317, in particular, are notable because of the known SNPs that are indicated by the highlighting in the **Table 12** (blue indicating an A/T SNP and yellow indicating a T/G SNP).

Table 13 compares the expected amplicon size, assessed using GelQuant to analyse the band migration compared to that of the marker in **Figure 12 B**, to the actual amplicon size by number of nucleotides sequenced in **Table 12**. The expected and actual sizes of the loci correlate to within 6 nucleotides of each other, with the exception of vWA. The similarity between these sizes, together with the complementarity of the alleles to their reverse sequences, indicates that the full length of each amplicon was successfully sequenced and that the sequencing results are representative of the full target amplicons. As previously discussed, vWA was a notable exception. The alleles for this locus were not sequenced in the forward and reverse directions and, consequently, only part of the sequence length is displayed in **Table 12**.

Table 13: Allele calls and comparison of actual to expected sizes for sequenced STRs

Locus	Allele Call	Identity Matrix Score [allele]	Actual Size (bp)	Expected Size (bp)
vWA[1]	14.1(16".1)	0.956 [14(16")]	132	197
	16.3 (18.3)	0.972 [16(18)]	142	197
D2S441	13´	0.989 [13´]	142	142
D3S1358	15´'	0.96 [15]	173	169
	19'	0.965 [19]	185	181
D21S11	29	1.000 [29]	231	226
	31.2	1.000 [31.2]	241	236
D7S820	12_AT	1.000 [12]	209	203
D13S317	9_A	1.000 [9]	149	145
	13_A	1.000 [13]	166	167

Allele calls were made by aligning consensus sequences (excluding M13 and with corrected STR-specific primer regions) to a database of known alleles. The SNP calls for loci containing known SNPs are indicated adjacent to the underscore, in order of 3´ to 5´ in the case of D7S820. The maximum Identity Matrix Score was generated using BioEdit v 7.5.2 with 1.000 representing 100% identity. The expected size of each amplicon was estimated from **Figure 12 B** using GelQuant. The actual size of the amplicon was given by the number of nucleotides sequenced.

[1]Electrophoresis of vWA on 37:1 acrylamide:bis-acrylamide did not enable the resolution of separate alleles.

Results in **Table 13** also served to call the allele number for each of the amplified STRs and compare them to those recorded in the STRBase. The maximum Sequence Identity Matrix score (generated by analysis of alignments using BioEdit) for each allele is given to illustrate the degree of identity between the aligned sequences. Alleles for loci D21S11, D7S820 and D13S317 yielded identities scores of 1.000 indicating perfect matches to the database. The alleles for loci vWA, D2S441 D3S1358 may represent variant alleles and the names indicated in the 'Allele call' column of **Table 13** are not present in the STRbase. To definitively identify these alleles, amplification, purification and sequencing needs to be repeated because some ambiguity in the sequences was observed. Allele 19 of D3S1358, for example, had the sequence $[TAGA]_{13}\,[CAGA]_{2}\,[TAGA]_{4}$ with peaks for C in [CAGA] being ambiguous (C or T with C slightly higher than T). The sequence may, therefore, have been $[TAGA]_{19}$ or $[TAGA]_{13/12}\,[CAGA]_{2/1}\,[TAGA]_{4/6}$.

Nevertheless, the allele calling was sufficient for the purpose of determining whether the PCR reactions specifically amplified the target loci. The conserved primer binding site and flanking region, together with the expected repeat sequences observed in **Table 12** and confirmed in **Table 13**, as well as the fact that the full length of the amplicons (with the exception of vWA) was successfully sequenced, indicated that specific amplification of the target STRs had been achieved.

2.2.6 Development and optimisation of multiplex polymerase chain reaction

The 1° PCR multiplex development entailed the altering of primer concentrations, as described by Schoske *et al.* (2003), until the even amplification of the 6 target STRs, in the minimum number of reactions, was achieved. Optimal singleplex PCR reaction conditions, including the annealing temperature of 64 °C, were used as a base for multiplex development and, while not all multiplex combinations are presented, **Figures 13** and **14** provide insight into the multiplex development.

Figure 13 presents an investigation of primer compatibility using different combinations of primers. **Figure 13 A** contains the relative primer concentrations used in each reaction, while **Figures 13 B** and **C** show the polyacrylamide gels resulting from the electrophoresis of each reaction product. **Figure 13 B** made use of 37.5:1 acrylamide:bis-acrylamide during electrophoresis and, consequently, the primer artefacts were visualised (indicated by the arrow head). Conversely, **Figure 13 C** made use of 19:1 acrylamide:bis-acrylamide and the

resolution of the loci, therefore, was higher than in **Figure 13 B**. Some primer combinations were only attempted once, for example reactions A, C, E and H. Other combinations were further developed and optimised, for example reactions B, D, F and G. Reaction combinations 1 and 2 represent two pairs of PCRs, the former of which is optimised, that resulted in the amplification of all target loci.

While reaction combinations 1 and 2 resulted in the amplification of the target loci; $G_{2,}$ visualised in **Figure 13 B**, represented a single multiplex that resulted in the, albeit uneven, amplification of all target loci. G_2, therefore, was selected for further optimisation, the process of which is outlined in **Figure 14.**

A in **Figure 14** outlines the various primer combinations, template quantities and cycle numbers considered for G reactions, and the electrophoresed PCR products generated using these parameters are depicted in **Figure 14 B**. Primer concentrations in each subsequent reaction were altered, as previously described, with the aim of creating an un-biased multiplex reaction. G_3 showed more even loci amplification than G_1 and G_2, however, a relative under-amplification of D13S317 and D2S441 and an over-amplification of D3S1358 were still observed. Reactions G_4-G_6 still aimed to optimise primer combinations, however other parameters were also considered.

G_4 served to investigate whether a reduced cycle number could be used. The reaction made use of an increased PCR volume of 50 µL and a subsequent concentration step (elution into 20 µL TE buffer using the Thermo Scientific GeneJET PCR Purification kit) prior to electrophoresis. Nevertheless, the bands resulting from the electrophoresis of the PCR product resulting from 21 cycles were barely visible, and those from ≤18 cycles were not visible at all.

Template concentrations of 25-200 ng in a 25 µL reaction were also explored, of which the products resulting from 100-200 ng are depicted in G_5 and $G_{6.}$ G_5 showed similar band intensity for 125-200 ng template DNA, whereas G_6, which used a template range of 100-150 ng, showed a distinct increase in band intensity with increasing template quantity. The intensity of bands in G_5 was far greater for the reaction containing 125 ng of template than those observed for $G_{6.}$ The reason for this is uncertain, however, variation in the intensity of control bands (marker bands) in different electrophoresis runs was not uncommon.

A

Reaction	Primer Concentration (µM) for each locus (each forward and reverse)						
	vWA	D2S441	D3S1358	D21S11	D7S820	D13S317	
A		0.2	0.2			0.2	
B_1			0.2			0.2	
C				0.2	0.2		
D_1	0.2	0.2		0.2	0.2		
E				0.2	0.2	0.2	
F_1	0.2	0.2	0.2	0.2	0.2		
G_1	0.2	0.2	0.2	0.2	0.2	0.2	
G_2	0.2	0.2	0.2	0.4	0.4	0.34	
F_2	0.2	0.3	0.16	0.4	0.4		
H	0.24	0.4	0.16	0.8	0.48		
I			0.2	0.48	0.44	0.56	1
J	0.2	0.32					
D_2	0.2	0.24		0.32	0.32		2
B_2			0.16			0.3	

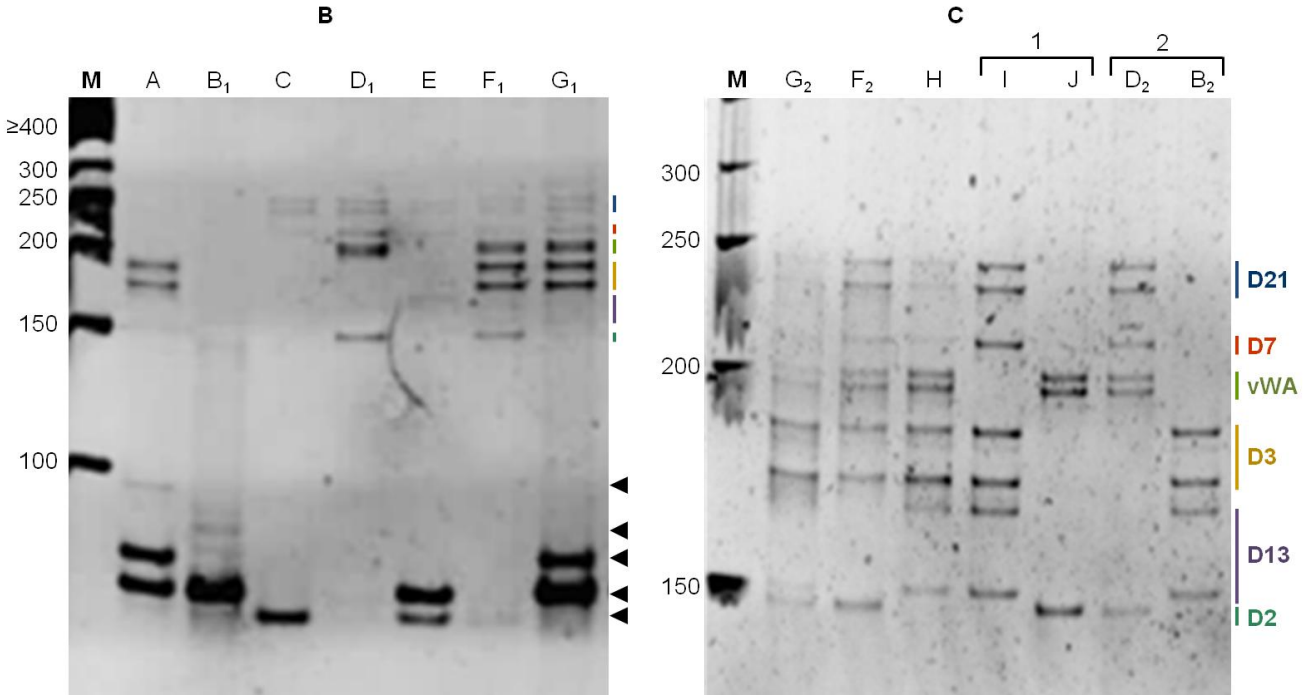


Figure 13: Investigation of various primer combinations for the multiplexing of target STRs using DNA (80-100 ng) extracted from Subject, amplified using KAPA HiFi HotStart ReadyMix.

A gives the primer concentrations of each multiplex visualised in **B** and **C**. Amplicons in **B** and **C** were separated using polyacrylamide gel electrophoresis [4% $^{w}/_{v}$ stacking, 10% $^{w}/_{v}$ resolving 37.5:1 (**B**) and 19:1 (**C**) acrylamide:bisacrylamide respectively at 5 V/cm in TAE]. The marker used for size determination (**M**) was the Thermoscientific GeneRuler 50 bp (1 µL/mm). ◄ marks the primer artefacts present in the multiplexes. 1 (I and J) and 2 (D_2 and B_2) are two multiplex pairs that resulted in the even amplification of all the loci.

A

Reaction	Primer Concentrations (µM) for each Locus (each F and R)						Template (ng)	Cycles
	vWA	D2S441	D3S1358	D21S11	D7S820	D13S317		
G_1	0.2	0.2	0.2	0.2	0.2	0.2	80	25
G_2	0.2	0.2	0.2	0.4	0.4	0.34	100	25
G_3	0.24	0.28	0.2	0.56	0.56	0.52	100	25
G_4	0.26	0.36	0.14	0.3	0.3	0.52	300	15,18,21,25
G_5	0.24	0.36	0.14	0.56	0.56	0.52	125,150,175, 200	25
G_6	0.2	0.36	0.14	0.4	0.4	0.52	2 x 100,125,150	25
G_7	0.2	0.4	0.14	0.4	0.4	0.56	150 (Subj. 0, i & ii)	25

B

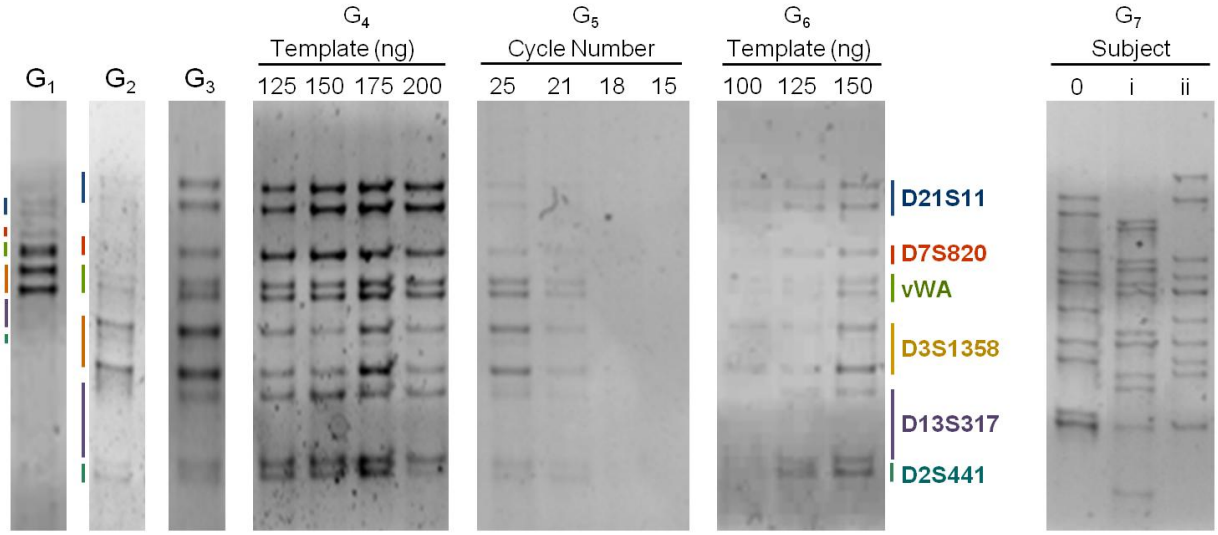


Figure 14: Optimisation of 1° multiplex PCR for the even amplification of target STRs, achieved by variation of relative primer concentrations; template quantity and cycle number.

All PCRs were performed using KAPA HiFi HotStart ReadyMix. **A** gives the primer concentrations, DNA template quantity and cycle number for the multiplexes visualised in **B**. All reactions except for G_4 had a total volume of 25 µL (G_4 had a total volume of 50 µL). Amplicons in **B** were separated using polyacrylamide gel electrophoresis (4% $^{w}/_{v}$ stacking, 10% $^{w}/_{v}$ resolving 19:1 acrylamide:bisacrylamide at 5-8 V/cm in TBE). DNA extracted from Subject 0 was used in multiplexes G_1 to G_7. G_8 illustrates the reproducibility of the optimised multiplex and used DNA extracted from Subjects 0, i and ii.

G_7 is the optimised 1° PCR multiplex and was performed on DNA extracted from the buccal swabs of Subject 0 and Subjects i and ii. The cycle number used was 25 and the template quantity was 150 ng to ensure that bands would be visible following electrophoresis. The results in **Figure 14 B** show that approximately even amplification of all target loci was achieved with minimal inter-subject variation.

2.2.7 Purification of 1° PCR Products

As illustrated in **Figure 13 B**, a high abundance of primer artefacts were present in the 1° PCR products. Consequently, prior to 2° PCR development, a method for 1° PCR purification was necessary to avoid the amplification of primer artefacts. **Figure 15** shows the products of the 1° PCR, purified using various techniques, and visualised using PAGE (electrophoresed for 5-18 hours). In **Figures 15 A-C** the unpurifed (U) PCR product was electrophoresed alongside the purified (P) PCR product. **Figure 15 D** contains only the electrophoresed purified product. DNA from Subject 0 was used for all analyses except for **Figure 15 B** where DNA from Subject i was used because the allele exhibited by this subject for D2S441 corresponded to the minimum size limit for this (the smallest) locus.

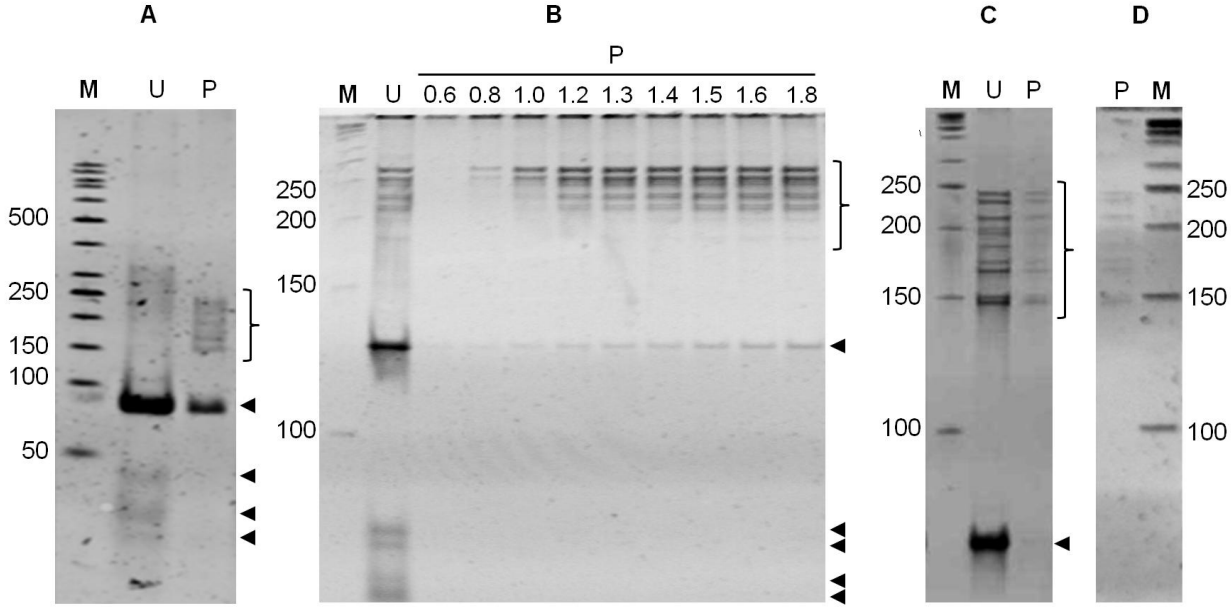


Figure 15: Investigation of various purification strategies for the removal of primer artefacts from the 1° PCR product.

Unpurified (U) and Purified (P) PCR products, generated using the KAPA HiFi HotStart ReadyMix, were visualised using polyacrylamide gel electrophoresis (4% $^{w}/_{v}$ stacking, 10% $^{w}/_{v}$ resolving 37.5:1 or 19:1 acrylamide:bisacrylamide at 5-8 V/cm in TBE). Thermo Scientific GeneRuler 50 bp (**M**) was used to determine the sizes of amplicons. Template DNA used in the PCR reactions was that extracted from from Subject 0 (**A**, **C** and **D**) or Subject i (**B**)**. A** investigated the use of the Thermo Scientific GeneJET PCR Purification kit. **B** investigated the applicability of Agencourt AMPure XP beads. Various ratios of beads to PCR product ($^{v}/_{v}$) were used for purification. **C** and **D** utilised the Thermo Scientific GeneJET Gel Extraction kit. **A** depicts the use of the kit following extraction of the target amplicons from an agarose gel; **B** depicts the use of the kit following extraction of the target amplicons from a polyacrylamide gel using the "crush-and-soak" method.

In **Figures 15 A** and **B** the two types of primer artefacts resulting from the 1° PCR product were visible in the unpurified products, namely the faint bands <50 bp in size which are

likely amplified primer, and the bright band at 73 bp which is similar to the non-target bands observed in the singleplex amplification reactions of D3S1358 and D13S317 (**Figure 12**). In a reaction that excluded the primers for D2S441, this latter primer artefact was not present in the PCR product (results not shown).

The use of the Thermo Scientific GeneJET PCR Purification kit in **Figure 15 A** resulted in the complete removal of the <50 bp artefacts and in a reduction in concentration of the larger artefact. The larger artefact was therefore too large to be removed using the PCR Purification kit.

The Agencourt AMPure XP beads in **Figure 15 B** also resulted in the complete removal of the smaller bands and, as the ratio of beads:PCR product decreased, so too did the concentration of the 73 bp artefact. An undesired decrease in the concentration of the smaller loci (D2S441 and D13S317) however also occurred, and when the ratio of beads to PCR product was sufficiently low (0.6:1) that the entire artefact was removed, the complete complement of loci were also removed. The size difference between the target loci and the artefact was, therefore, deemed too small to purify the PCR product using the AMPure beads.

The methods employed in both **Figures 15 C** and **D** involved the use of electrophoresis to separate the target amplicons from the primer artefacts, and subsequent excision of the target bands from the gel. Agarose gel electrophoresis did not adequately resolve the bands and, consequently, the 73 bp primer artefact was faintly visible in the purified product in **Figure 15 C**. While an extended electrophoresis time would have enabled adequate separation of the bands, the quantity of gel to be excised would have become unmanageable (>1 g) and purification tedious (the purification column can hold only 800 µL DNA binding buffer, which must be mixed at a 1:1 ratio with the melted DNA). Conversely, PAGE did enable adequate separation of the target amplicons from the artefacts. The "crush and soak" method, demonstrated in **Figure 15 D**, successfully removed all primer artefacts, and was therefore employed to prepare the 1° PCR products required for 2° PCR development.

2.2.8 Optimisation of 2° PCR for addition of fusion primers

Using purified 1° PCR product from Subject 0 and Subjects ii and iv, the 2° PCR was developed and optimised. The characteristics of the primers used for these reactions are depicted in **Table 14**. Forward and reverse primers for each population group were differentiated only by the MID sequence and consequently have near identical characteristics.

All the 2° PCR primers are relatively long and, consequently, have high melting temperatures of between 72.6 and 74.9 °C. These temperatures would suggest an optimal annealing temperature of ~70 °C. The difference in melting temperature between the forward and reverse primers is sufficiently low as to ensure primer compatibility. The forward primers are notable for their high GC content (64.4%) and the Nguni-specific forward primer has the highest strongest folding melting temperature of all the primers (50.8 °C). It was concluded that the high GC content of the forward primers may cause the 2° PCR to require an extended denaturation time or an elevated denaturation temperature, while the strongest folding melting temperature was unlikely to affect the success of amplification since the PCR cycling parameters would not drop below 60 °C (given the predicted annealing temperature of 70 °C).

Table 14: General characteristics of population-specific primers used for 2° PCR amplification

Primer	Size (bp)	T_m (°C)	CG (%)	Strongest Folding T_m (°C)	E-value
Nguni F	59	74.9	64.4	50.8	7e-09
ST F	59	74.9	64.4	46.2	7e-09
Nguni R	57	72.8	57.8	40.5	n/a
ST R	57	72.6	57.8	41.6	n/a

Primers were analysed using IDT Oligo Analyser 3.1 and BLAST. The maximum E-values obtained following BLAST analysis corresponded to a 16s RNA sequence obtained from a sample taken from Madagascan soil. The reverse primers had no significant sequence similarity in the BLAST database. ST represents primers specific for the Sotho-Tswana population group. F and R indicate the forward and reverse primers respectively.

To commence 2° PCR optimisation, an initial gradient PCR using annealing temperatures of 61-72 °C was prepared according to manufacturer's instructions and with 1° PCR product from Subject 0 and Nguni-specific primers (data not shown). The resultant products showed that many non-specific amplicons of >350 bp were present. These non-target amplicons were present across the temperature range at approximately constant intensities, while the target loci intensities diminished as the annealing temperature increased above 65 °C. The reduction in target amplicon amplification at higher temperatures was unexpected since the predicted annealing temperature was 70 °C. To ascertain the cause of these non-specific bands, the primer sequences were assessed using BLAST, the results of which are displayed in **Table 14**. The minimum E-value for the forward primers was 7e-09, corresponding to the 16S rRNA of a Madagascan soil sample, the source of which is unlikely to be contaminating the genomic DNA samples, and no significantly similar sequences were obtained for the BLAST analysis of the reverse sequences. Both sequences yielded no significant similarity when searched against the human database. Consequently, it was deduced that the non-target amplicons were

a consequence of ineffective cycling parameters and PCR conditions, as opposed to the amplification of contaminating DNA.

A

Reaction	Template DNA		Primers		DMSO	Annealing Temperature	Cycle Number
	Quantity	Subject	Type	µM			
1	2.6 ng	0 and ii	N and S	0.2	-	60 °C	20
2	2.6 ng	0 and ii	N and S	0.2	-	64 °C	20
3	2.6 ng	0	S	0.12	-	64 °C	20
4	2.6 ng	0	S	0.06	-	64 °C	20
5	0.05 ng	0	S	0.2	-	64 °C	20
6	0.5 ng	0	S	0.2	-	64 °C	20
7	1.3 ng	0	S	0.2	-	64 °C	20
8	2.6 ng	0	S	0.2	3 %	64 °C	20
9	2.6 ng	0	S	0.2	5 %	64 °C	20
10	2.6 ng	0	N	0.2	-	64 °C	20
11	50 pg	0	N	0.2	-	64 °C	20
12	4 pg	0	N	0.2	-	64 °C	20
13	3 pg	0	N	0.2	-	64 °C	20
14	1 pg	0	N	0.2	-	64 °C	20
15	1 pg	iv	N	0.2	-	64 °C	25
16	1 pg	iv	N	0.2	-	64 °C	20

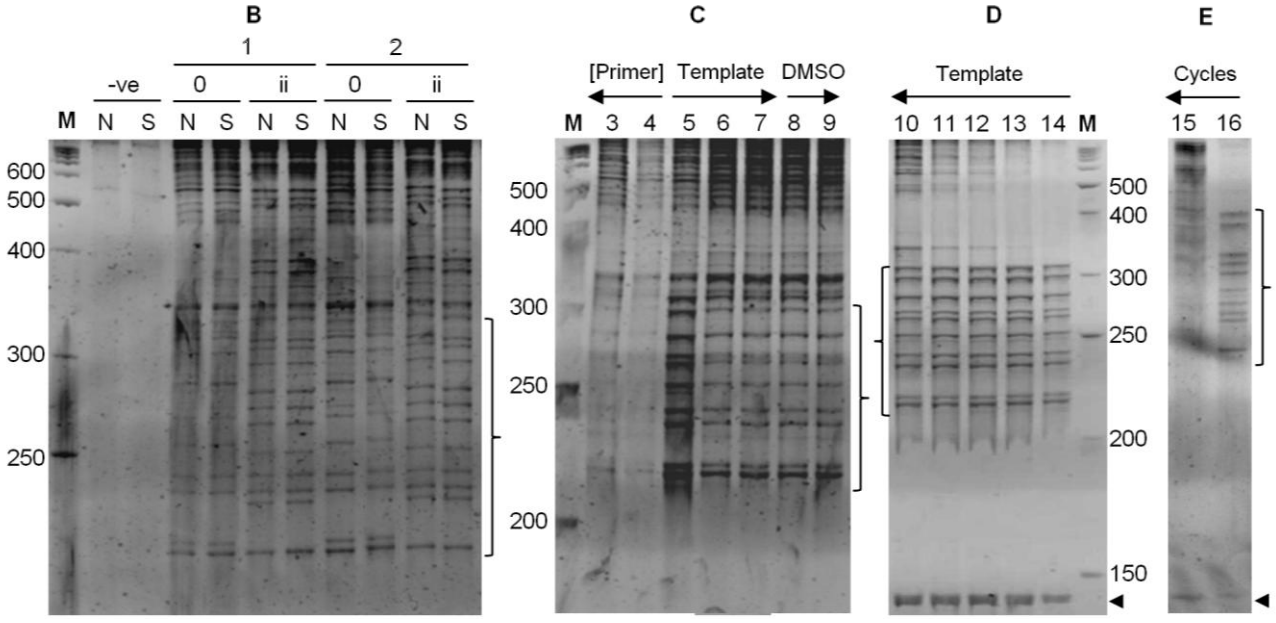


Figure 16: Development and optimisation of 2° PCR by variation of PCR conditions.

All amplifications were performed using KAPA Hifi Hotstart Readymix and template DNA extracted using the "crush-and-soak" method and Thermo Scientific GeneJET Gel Extraction kit. Primers specific to either the Nguni (N) or Sotho-Tswana (S) population groups were used. **A** outlines the parameters used for each PCR reaction visualised by polyacrylamide gel (4% $^{w}/_{v}$ stacking, 10% $^{w}/_{v}$ resolving 19:1 acrylamide:bisacrylamide) electrophoresis (5-8 V/cm in TBE) in **B-E**. The red boxes in **A** indicate the parameter of interest for the particular PCR reaction. The marker (**M**) used to identify target amplicons in **B-E** was the Thermo Scientific GeneRuler 50 bp (0.07 µL/mm). Brackets indicate the target amplicons while ◄ indicates the primer artefacts.

A number of reaction conditions were investigated in an attempt to optimise the 2° PCR reaction. The PCR conditions and cycling parameters of which are outlined in **Figure 16 A** and the results of which are displayed in **Figure 16 B-E**. Reactions 1 and 2 considered in **Figure 16 B** aimed to verify that the previously described non-target amplicons were neither specific to the amplification of the 1° PCR product of Subject 0, nor to the Nguni-specific primers. The PCR was therefore repeated using the 1° PCR product from both Subject 0 and Subject ii, and using both Nguni- and ST-specific primers. In these reactions however, only two annealing temperatures were investigated, namely 60 and 64 °C.

The banding pattern observed in **Figure 16 B** when comparing the use of the Nguni- and ST-specific primers, following electrophoresis of the PCR products, was indistinguishable. Moreover, the previously described large non-target amplicons were observed for both Subjects 0 and ii. Interestingly, however, the banding pattern of the non-target amplicons was distinct for each subject. The intensity of all bands (target and non-target) decreased when an annealing temperature of 60 °C was used as opposed to 64 °C.

The results observed for reactions 1 and 2 in **Figure 16 B** indicated that the different population-specific primers produced indistinguishable results, and that the presence of non-target bands was not subject-specific but that their banding pattern was. The results also suggested that an annealing temperature of 64 °C should be retained to ensure that the yield of target amplicons remained sufficient for visualisation using PAGE. Consequently, subsequent reaction conditions investigated made use of DNA from a single subject and a single population-specific primer set. In an attempt to reduce non-target amplicons, the cycling parameters were also adjusted from those outlined in the manufacturer's instructions. The initial denaturation was increased to 98 °C and the cycling denaturation time was extended by 10 seconds to ensure the complete denaturation of the GC-rich forward primers. The extension time was also reduced from 15 to 10 seconds to prevent the amplification of long non-target amplicons.

Reactions 3 to 9 in **Figure 16 C** investigated a number of parameters in an attempt to increase the specificity of the PCR. Reduced primer concentrations were used in reactions 3 and 4; varying template concentrations were used in reactions 5-7 and the effect of adding DMSO was investigated in reactions 8 and 9. Reducing the primer concentration resulted in a marked reduction in the yield of the PCR, both with regards to the target and non-target amplicons (reactions 3 and 4). DMSO had no apparent effect (reactions 8 and 9). Conversely,

the addition of less template resulted in an increased yield of target amplicons and a corresponding decrease of non-target amplicons (reactions 5-7).

Reactions 10-14 in **Figure 16 D** further investigated the effect of using reduced template concentration. Reaction 14, which contained only 1 pg of template DNA, resulted in the specific amplification of the target amplicons. Yields were however low and, since the purification strategy for 2° PCR would be a gel-based technique, an attempt to increase the yield slightly by increasing the cycle number was explored in reaction 15 presented in **Figure 16 E**. Reaction 15 made used of template DNA from Subject iv, and was accompanied by a control reaction, reaction 16, which used identical parameters to those of reaction 14. Reaction 15 resulted in the renewed presence of non-target amplicons; Reaction 16 served to illustrate that the optimised protocol was not template-specific.

2.2.9 Analysis of whether complete purification of 1° PCR product is necessary for successful 2° PCR amplification

The protocol optimised in this chapter was to be used to prepare 144 samples for 454 sequencing. Consequently, minimisation of the processing time and simplification of the processing protocol was crucial.

Minimisation of the processing time and simplification of the processing protocol was imperative too if the protocol was to be eligible for use in the forensic setting in place of standard 1-step PCR followed by capillary electrophoresis. Consequently, the optimised 2° PCR protocol was used to amplify 1° PCR products that were unpurified, purified using only the Thermo Scientific PCR Purification kit (column purification), or purified using the 'crush and soak" method. These results are displayed in **Figure 17**.

Figure 17 A shows the unpurified and PCR purification kit purified 1° PCR products, and the 2° products that resulted from the amplification of these products. In both 2° PCR products, no bands were observed in the expected size range and identical non-target bands of >450 bp were observed. Conversely, 1° PCR products purified using the "crush and soak" method, gave rise only to bands within the expected size range (with the exception of a ~140 bp primer artefact). These results are displayed in **Figure 17 B** which contains both the 1° PCR product purified using the "crush and soak" method, and the corresponding 2° PCR product. The banding pattern observed for the 2° PCR reaction is identical to that for the 1° PCR product with the exception of the expected size increase of 70 bp per band, indicative of

successful fusion primer addition. Complete purification of the 1° PCR product was therefore necessary for successful, selective 2° PCR amplification.

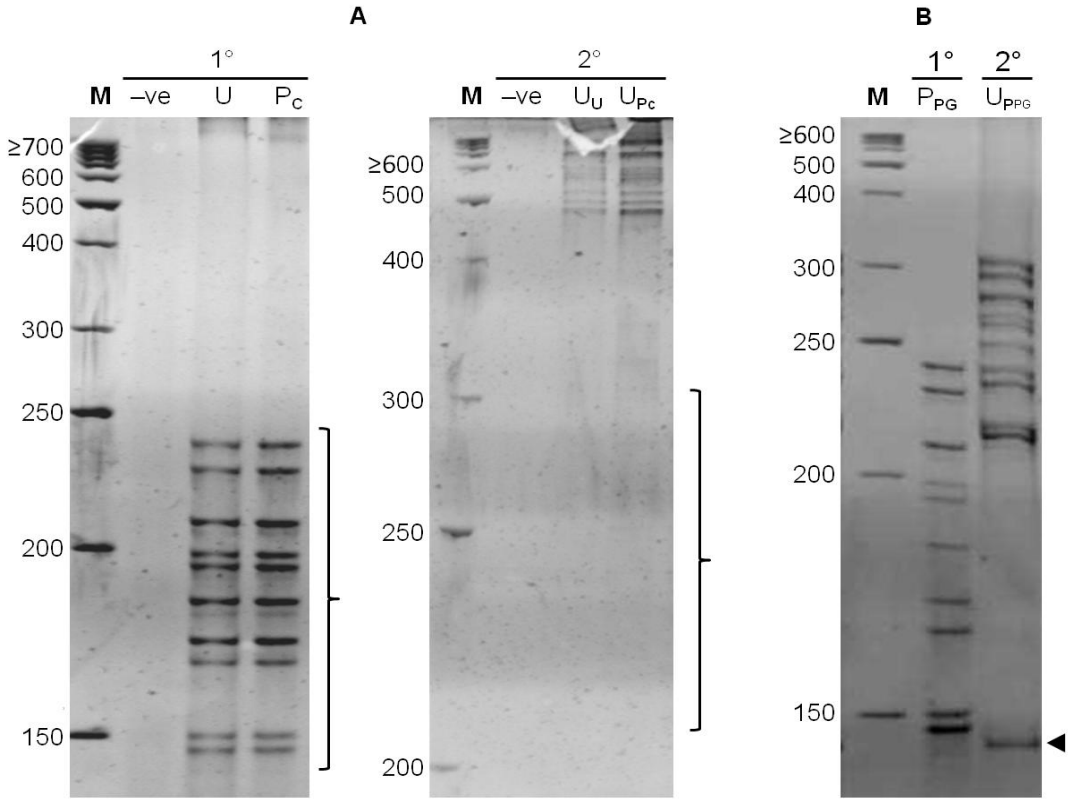


Figure 17: Assessment of whether complete purification of the 1° PCR product is necessary for successful 2° PCR amplification.

Unpurified (U) and Purified (P) PCR products, generated using the KAPA HiFi HotStart ReadyMix and Subject 0 template DNA, were visualised using polyacrylamide gel electrophoresis (4% $^{w}/_{v}$ stacking, 10% $^{w}/_{v}$ resolving 19:1 acrylamide:bisacrylamide at 5-8 V/cm in TBE). 0.07 µL/mm Thermo Scientific GeneRuler 50 bp (**M**) was used determine the sizes of amplicons and negative controls (-ve) were prepared by adding dH_2O in place of template DNA. In **A,** 1° PCR products were electrophoresed prior to purification (P_C) and subsequent to purification with Thermo Scientific GeneJET PCR Purification kit (P_C). 1 pg of these 1° PCR products was used as the template for the 2° PCR amplifications visualised in **A**. U_U was generated using the unpurified 1° PCR product as a template; U_{Pc} was generated using PCR purified 1° PCR product as a template. In **B**, the 1° PCR product (P_{PG}) was purified using the "crush and soak" method and concentrated using the Thermo Scientific GeneJET Gel Extraction kit. 1 pg of this purified product was used as the template for the 2° PCR product (U_{PPG}) visualised in **B**. ◄ marks the primer artefacts present in U_{PPG}.

2.2.10 Purification of the 2° PCR product

Purification of the 2° PCR product was necessary prior to 454 sequencing to remove the primer artefact observed in **Figure 16 B**. The use of the Thermo Scientific GeneJET Purification kit and Agencourt AMPure XP beads were not considered during purification development for the 2° PCR products since the results obtained in *2.2.7* indicated that these kits would be unable to selectively remove primer artefacts of >75 bp. Consequently, only

agarose gel purification and purification using the "crush and soak" method were investigated as methods for 2° PCR purification, the results of which are depicted in **Figure 18**.

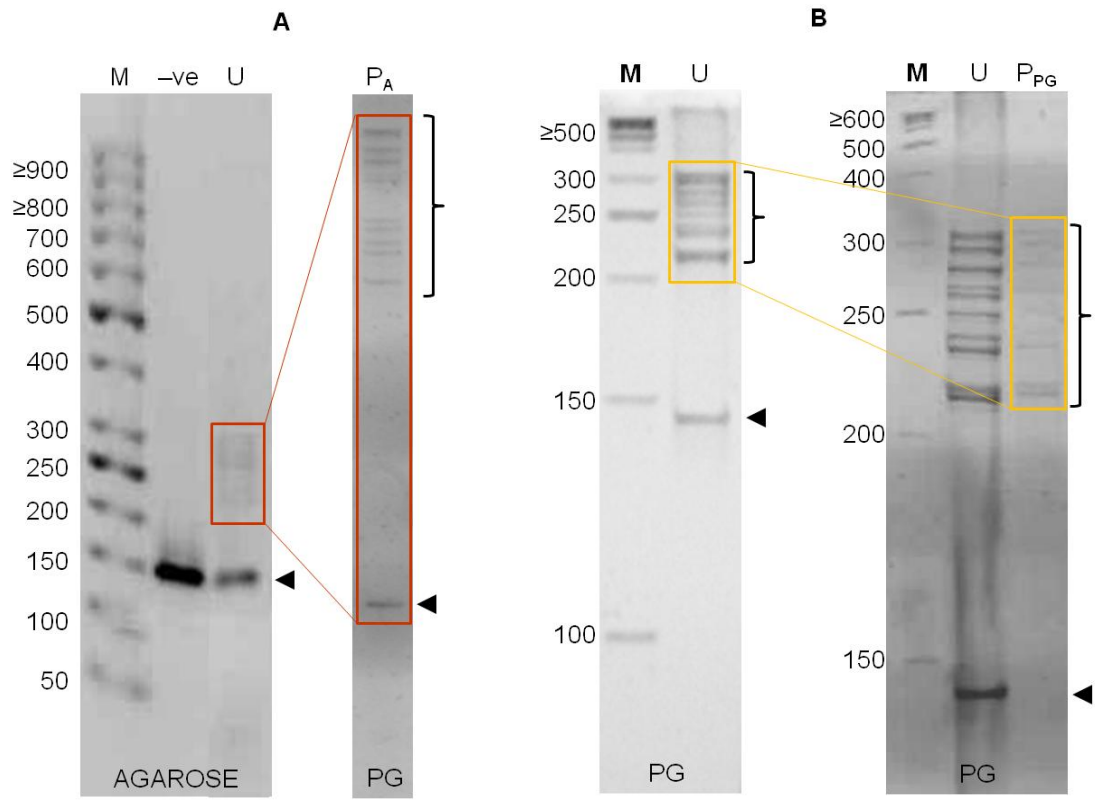


Figure 18: Comparison of gel purification strategies for purification of 2° PCR product

All PCR products were prepared using the KAPA HiFi HotStart ReadyMix and the marker (**M**) used for size determination of DNA products was the Thermo Scientific GeneRuler 50 bp. ◄ marks all primer artefacts. **A** depicts the agarose gel (2% $^{w}/_{v}$) electrophoresis (5V/cm) of 1 µL/mm **M**, a negative control (-ve) prepared by adding dH_2O in place of template during 2° PCR amplification, and unpurified 2° PCR product for Subject iii. The region outlined in red is the region of the agarose gel that was excised and purified using the Thermo Scientific Gel Extraction kit. The purified DNA (P_A) was visualised by polyacrylamide gel (PG) electrophoresis (4% $^{w}/_{v}$ stacking, 10% $^{w}/_{v}$ resolving 19:1 acrylamide:bisacrylamide at 5-8 V/cm in TBE). **B** depicts the PG electrophoresis (as described for A) of unpurified 2° PCR product (U) and the yellow box is the region that was excised for purification using the "crush and soak" method and the Thermo Scientific GeneJET Gel Extraction kit.

Figure 18 A shows the agarose gel from which the PCR product was excised, as well as the purified PCR product of Subject iii. Agarose gel purification using the Thermo Scientific GeneJET Gel Extraction kit was unsuccessful at removing the ~140 bp primer artefact. Conversely, the purification of the 2° PCR product from a polyacrylamide gel (U in **Figure 18 B**) using the "crush and soak" method resulted in the complete removal of the primer artefact from the PCR product

2.3 DISCUSSION

The aim of Chapter 2 was to identify STRs whose sequence analysis was likely to reveal novel STR variants and SNPSTRs; and to develop a protocol that would create emPCR-ready amplicons from these loci. The STRs D2S441, D3S1358, D7S820, vWA, D13S317 and D21S11 were selected based on their primer and size compatibility, as well as their predicted likelihood to reveal novel STR variants. The emPCR-ready amplicons were prepared using a 2-step PCR protocol. During the 1° multiplex amplification, target STRs were amplified, using M13-tailed primers, from 150 ng input DNA purified from buccal swabs using the complete Isohelix DNA Extraction kit. During the 2° amplification 1 pg of 1° PCR product was amplified using M13-directed population-specific fusion primers. Both PCRs made use of the minimum possible cycle number that permitted visualisation of the PCR product by PAGE, namely 25 and 20 cycles, and required clean-up using the "crush-and-soak" method.

At the time of commencement of this study, to the author's knowledge, only 4 papers (outlined in **Table 7**) had been published which considered the application of SGS to STR analysis. Namely, that by Bornman (2012) which made use of the Illumina GAIIx system; and those by Scheible *et al.* (2011), Fordyce *et al.* (2011) and Van Neste *et al.* (2012) which made use of the Roche GS systems. While Bornman *et al.* (2012) successfully sequenced the 13 CODIS loci (amplified in singleplex) using Illumina GAIIx system, because the system was limited to sequencing amplicons of 150 bp, larger alleles (for example alleles >32.2 for D21S11) were unable to be sequenced. Due to this size-limitation of Illumina sequencing systems, the remaining 3 publications made use of Roche GS systems.

The first, by Scheible *et al.* (2011), made use of the GS Junior system to sequence amplicons generated using the 40 cycle multiplex published by Pitterl *et al.* (2008). Specific results regarding STR variants were not given, except to say that the number of reads per locus was imbalanced and that the multiplex was undergoing further optimisation to prevent this bias. The second and third publications, by Fordyce *et al.* (2011) and Van Neste *et al.* (2012) respectively, made use of the GS FLX Titanium system. Fordyce *et al.*, used a 30 cycle singleplex PCR to amplify CSF1PO, D13S317, D21S11, D5S818 and TH01 from 1 µL DNA. Van Neste *et al.* (2012) used the non-high fidelity AmpFlSTR® Profiler Plus® kit (Applied Biosystems) (including fluorescently labelled primers) to amplify D3S1358, D5S818, D7S820, D8S1179, D13S317, D18S51, D21S11, FGA and vWA in a 34 cycle multiplex PCR using 1 ng input DNA. In each of these cases, the amplicons underwent adaptor ligation with

individual-specific MIDs as per the Rapid Library Preparation Manual (http://454.com/downloads/my454/documentation/gs-flx-plus/Rapid-Library-Preparation-Method-ManualXLPlusMay2011.pdf) and clean-up using Agencourt AMPure beads.

In contrast to the singleplex and unoptimised multiplexes used in the aforementioned studies, the protocol outlined in Chapter 2 entailed the use of a multiplex reaction (the 1° PCR) designed and optimised specifically for this study. This reaction was also unique because it resulted in the addition of a universal primer sequence, required to enable for 2° PCR amplification using M13-directed fusion primers. Since adaptor ligation requires a DNA input of >10 ng/uL, fusion primer addition by 2° PCR was necessitated by the low yield obtained following purification of the 1° PCR product (3.32-11.69 ng/uL, discussed in *3.2.1*). However, an investigation into the yields obtained for the multiplex STR excluding the universal tail have not been performed and may well provide sufficient yields to enable direct adaptor ligation. This too would be useful for reducing the total cycle number (currently at 20 + 25) important for ensuring that sequence variations observed in 454 reads are *bona fide* genomic mutations. Alternatively, an investigation into the use of population- or individual-specific fusion-tailed STR-specific primers could be performed. A limitation with this latter experimental design is however the high cost associated with fusion primer synthesis.

A second notable difference between the study outlined in Chapter 2 and those described in literature was the clean-up method employed, namely the "crush-and-soak" method as opposed to the use of Agencourt AMPure beads. While the "crush-and-soak" method is laborious and vulnerable to contamination, it was the only method which enabled the complete removal of primer-dimers from PCR products. It also enabled a quality-check following each PCR, useful since this part of the study was aimed at protocol development. Interestingly, none of the aforementioned papers reported analysis of purified products to ensure dimer removal and, based on the results in **Figure 15 B** it is possible that dimer removal may have been unsuccessful given that the minimum amplicon size for the smallest locus with ligated adaptor is 148 and primer dimers are 70 bp. Analysis of the sequencing results for these studies, however, would enable identification of contaminating primer dimers.

In the last 30 months, an additional 5 papers (outlined in **Table 7**) have been published which consider the application of NGS to STR analysis. 2014 saw the publication of papers that

utilised the GS Junior (Van Neste *et al.*, 2014; Scheible *et al.*, 2014; Kim *et al.*, 2014 and Gelardi *et al.*, 2014) and Illumina MiSeq (Zeng *et al.*, 2014) systems; Ion Torrent, the platform that is replacing 454 sequencing, has been investigated in a paper published in 2015 (Fordyce *et al.*, 2015). With the exception of that published by Gelardi *et al.* (2014), the protocols for library preparation used in these publications entailed multiplex PCR for STR amplification followed by adaptor ligation. Gelardi *et al.* (2014) performed single- or duplex PCR using fusion-tailed STR-specific primers. All protocols included clean-up using Agencourt AMPure beads.

In these more recent papers, focus has shifted from mere proof-of-concept studies, to testing the protocols using template DNA which mimic forensic samples (mixed and degraded DNA) or that are genuine casework samples (Fordyce *et al.* (2015). Zeng *et al.* (2014) and Fordyce *et al.* (2015), for example, reported minimum sensitivities of 62 and 50 pg input DNA respectively; and Fordyce *et al.* (2015) were able to resolve mixtures down to a 20:1 ratio.

These recent papers highlight the downfalls of the experimental design developed in Chapter 2, with regards to suitability to the forensic field. Firstly, input DNA required for the 1° PCR reaction was large, at 3 000 times that of the minimum sensitivity reported by Fordyce *et al.* (2015). The sample processing protocol, too, was laborious with its 2-step PCR and "crush-and-soak" method of purification; which also makes it vulnerable to contamination. Furthermore, the total PCR cycle of 45 is concerning since the likelihood of introducing sequencing errors increases with each PCR cycle. Nevertheless, the protocol enabled the specific amplification of STRs with universal-tailed primers, validated by Sanger sequencing. It also enabled the addition of population-specific fusion primers to 1° PCR products. The yield of the PCRs was sufficient to enable visualisation of PCR products following PAGE, useful for quality assessment of amplicons. Furthermore, the "crush-and-soak" method was successfully employed to remove primer dimers from PCR products. The protocol, therefore, was deemed suited to the creation of population-specific emPCR-ready STR amplicons.

CHAPTER 3

SNPSTR Discovery and Analysis

The optimised protocol described in Chapter 2 and outlined in Figure 7 was used to process buccal swabs collected from study subjects.

3.1 METHODOLOGY

3.1.1 DNA collection, purification and quantification

DNA collection was performed at Rhodes University, South Africa from non-related subjects of self-defined Nguni or ST population groups. Subjects were required to give informed consent prior to donation of the buccal swab, and to fill out a questionnaire (**Figure i** in the **Appendix**) stating their gender, age, family region of origin, and population group. All DNA was purified as described in *2.1.4* and quantified spectrophotometrically as described in *2.1.3*.

3.1.2 DNA Processing

3.1.2.1 1° PCR amplification and purification

Purified genomic DNA (150 ng) was amplified using KAPA HIFI^{TM} HotStart ReadyMix as per manufacturer's instructions, using cycling parameters outlined in *2.1.9* with 25 cycles and 10.5 µL primer mastermix. The final primer concentrations (forward and reverse) in the PCR reaction mixtures were 0.2 µM vWA; 0.4 µM D2S441, D21S11 and D7S820; 0.14 µM D3S1358; and 0.56 µM D13S31. Negative and positive controls were prepared for each PCR run by replacing template DNA with dH_2O *2.1.5* and the target bands were excised and purified using the "crush and soak" method described in *2.1.10.4*. Thereafter, amplicons were quantified spectrophotometrically as described in *2.1.3*.

3.1.2.2 2° PCR amplification and purification

Purified 1° PCR product (1 pg) was amplified using either the Nguni-specific or ST-specific primers at 0.2 µM (forward and reverse primers each) and the KAPA HIFI^{TM} HotStart ReadyMix with the adjusted cycling parameters outlined in *2.1.9*. Amplicons were purified and spectrophotometrically quantified as described in *2.10.4* and *2.1.3*. respectively. Negative

and positive controls were prepared for each 2° PCR run by replacing template DNA with the purified 1° PCR products from the negative and positive controls prepared in *3.1.2.1*.

3.1.3 Pre-NGS quality check

3.1.3.1 PAGE of randomly selected purified 2° PCR products

The controls and six arbitrarily selected purified 2° PCR products (10 µL) were electrophoresed and visualised as described in *2.1.5* to assess the success of the sample processing protocol. The negative and positive controls prepared for each 2° PCR run were concentrated using the Thermo Scientific GeneJET PCR Purification Kit as per manufacturer's instructions and including the optional step for DNA <500 bp. The arbitrarily selected samples were the Nguni Subjects 14, 15, 68 and 84 and ST Subjects 101 and 144.

3.1.3.2 Sanger sequencing of purified 2° PCR product

The 1° PCR product for Subject 15 was used as a template to prepare 19 parallel 2° PCRs. Following amplification, the PCR products were electrophoresed as described in *2.1.5* for an extended time period (28 hours). Individual alleles for each locus were excised, pooled, purified using the "crush and soak" method described in *2.1.10.4*, and concentrated using the Thermo Scientific GeneJET PCR Purification Kit as per manufacturer's instructions and including the optional step for DNA <500 bp. Sanger sequencing of the purified alleles was performed by Inqaba Biotech™. Consensus sequences were created as described in *2.1.8.* Fusion primer and M13 tail regions were removed from sequences and sequencing errors in the STR-specific primer regions were corrected. Allele calling was performed as described in *2.1.8*.

3.1.4 Library preparation, emPCR amplification and 454 Sequencing

Purified 2° PCR products of similar concentrations were concentrated using the Thermo Scientific Scientific GeneJET PCR Purification kit, as per manufacturer's instructions and including the optional steps for DNA <500 bp in size. The resulting 15 samples were diluted to give a single sample library containing 2.4 ng 2° PCR product per subject. A total of 0.2 molecules per capture bead were targeted for emPCR and sequencing was performed as per Roche GS Junior Titanium series instructions by Dr Gwynneth Matcher of the Rhodes University Sequencing Facility.

3.1.5 Data Sorting and Analysis

A Biopython script, the functions of which are outlined in **Figure 19**, for the processing and sorting of raw sequencing data were prepared by Mr Jeremy Baxter of the Rhodes University Statistics Department.

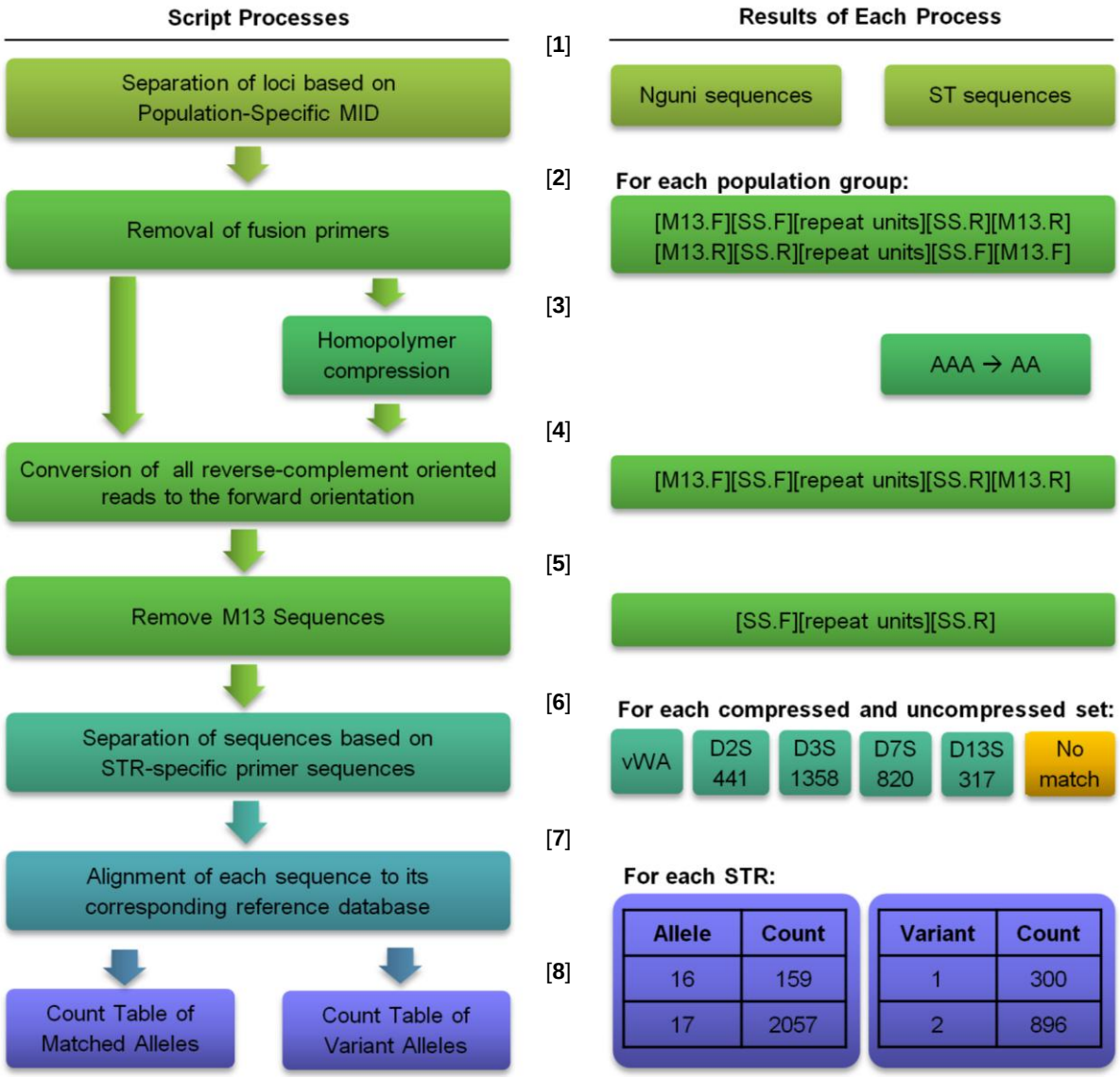


Figure 19: Outline of the Biopython script functions used for data processing and sorting, together with a schematic of the results generated during each process.

Universal forward and reverse primer sequences are indicated by M13.F and M13.R respectively. STR-specific forward and reverse primer sequences are indicated by SS.F and SS.R respectively.

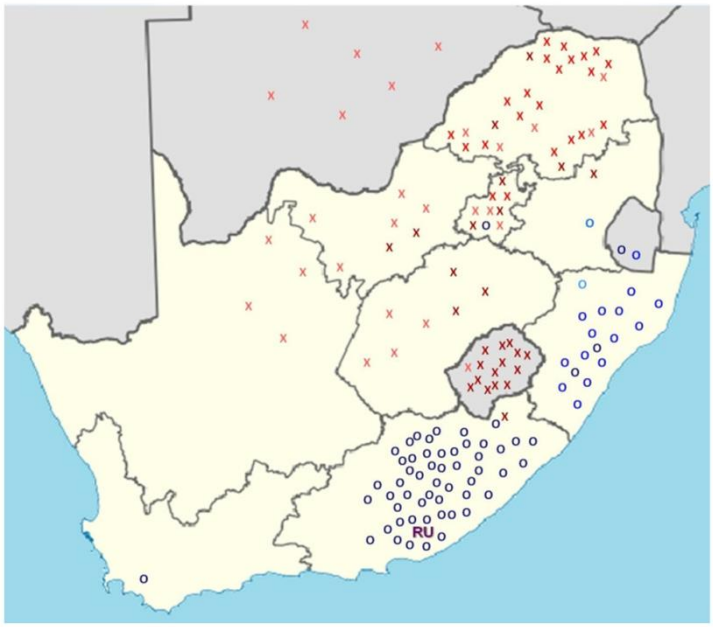
Data were initially separated based on population-specific MIDs (**Figure 19** [**1**]) whereafter fusion primer sequences were removed from the sequenced reads (**Figure 19** [**2**]). Sequences were then processed in 2 parallel streams. In the first stream, homopolymers of greater than 3 nucleotides, in both the sequenced data and the reference sequences (presented in **Table i of**

Appendix), were compressed to 2 nucleotides (**Figure 19** [**3**]), for example 'GGGG' was converted to 'GG'. In the second stream, data did not undergo homopolymer compression. All sequences were then converted to the forward orientation using the M13 forward and reverse primers (either compressed or uncompressed depending on the data set) as a reference (**Figure 19** [**4**]). M13 primer sequences were subsequently removed from the sequenced reads (**Figure 19** [**5**]) and data were separated into their respective STR types using both the forward and reverse STR-specific primer sequences (again either compressed or uncompressed depending on the data set) (**Figure 19** [**6**]). Sequences that did not match 100% to both the forward and reverse STR-specific primers were placed in a separate folder to signify that "no match" was possible. Sequences that did match a particular primer pair were aligned to the reference database presented in **Table i** of the **Appendix** (**Figure 19** [**7**]). This database was compiled using the flanking regions reported by Oberacher *et al.* (2008) and the repeat units per allele listed in the STRbase (www.cstl.nist.gov/strbase) for each locus. Sequences which matched a particular reference allele were named according to their allele call; while sequences that did not match a particular allele were placed in a separate folder containing putative variants. Count tables of these matched alleles and putative variants were prepared so that each unique sequence was represented once, together with its count value (the number of times it was observed) (**Figure 19** [**8**]).

The sum of the total number of reads obtained for each population group, subsequent to M13 removal, was taken as the total number of successfully sequenced target amplicons. This amount was divided by the maximum possible number of unique amplicons (144 individuals times 12 alleles) to give the minimum count number for a sequenced read to be considered a 'true' amplicon, as opposed to a sequencing error. These retained sequences were used to create lists of alleles with their corresponding sequence read number, for each population group, for both compressed and uncompressed sequences, and for every STR; resulting in a total of 24 lists plus 2 lists, for compressed and uncompressed data, per population group for sequences that did not match any STR. The frequencies of each sequence were then normalised for each population group by dividing the number of reads per allele by the total number of retained reads of the relevant population group (both for compressed and uncompressed data). Each variant allele was then aligned and visualised as described in *2.1.8* to enable variant allele calling and identification of the nature of the variation (SNP, STR variant, flanking region indel etc).

3.2.1 DNA collection and Processing

Buccal swabs from 144 subjects, 72 Nguni and 72 ST individuals were collected at Rhodes University, South Africa. All subjects were required to give informed consent and to fill out a Microsoft Office Access Questionnaire. The questionnaire and information sheets are presented in **Figures i-ii** of the **Appendix**. Using the information collected from the donors, **Figure 20** was constructed. This figure illustrates where in Southern Africa the participants' family originate, together with their self-defined population-group. It should be noted however that 12.5% of ST donors were unable to give complete paternal data and that the self-defined ancestry and region of origin for these subjects were, therefore, limited to maternal data.



KEY

Marker	Population Group	
X	Sotho	Sotho-Tswana
X	Tswana	
X	Pedi	
O	Xhosa	Nguni
O	Zulu	
O	Ndebele	
O	Swati	

Figure 20: Self-defined family region of origin of Subjects 1-144, mapped to illustrate the distribution of Nguni and Sotho-Tswana populations in South Africa.

DNA samples were collected at Rhodes University (RU), Eastern Cape, South Africa.

The distribution of the population groups in **Figure 20** correlates with that presented in **Figure 3** – with the Sotho, Tswana and Pedi populations clustering in central and North-Western Southern Africa; and the Xhosa, Zulu, Ndebele and Swati populations clustering to the South-East of the country. Two notable exceptions for the Nguni population originated from Cape Town and Johannesburg, South Africa's two major cities. All other

willing participants that originated from either of these cities were unsuitable donors for this study since they were of combined ST and Nguni ancestry, highlighting the effect of admixture observed in cities.

The protocol developed in Chapter 2 was used to process the buccal swabs collected from each donor. Briefly, genomic DNA was extracted using the Isohelix DNA extraction kit and amplified in multiplex using STR-specific M13-tailed primers. Thereafter, 1° PCR products were purified and used as the template DNA for 2° amplification with M13-directed population-specific fusion primers. An example of the polyacrylamide gels from which PCR products (both 1° and 2°) were excised and purified are displayed in **Figure 21**. The results obtained by analysis of DNA using the Nanodrop 2000, following each purification step, are displayed in **Table 15**.

The profiles in **Figure 21** showed evidence of both inter- and intra-individual variation in profile and amplicon intensity. While these variations may be a consequence of the electrophoresis protocol used for DNA visualisation, evidenced by the variability observed in the profiles for the replicates of Subject 15 in **Figure 23**; they too may be a consequence of inconsistent template quantity, allele homozygousity, or mutations within the primer binding sites of STRs. Interestingly, the effect of PCR bias toward smaller amplicons was not observed, likely because the maximum range between allele sizes was small at 157 bp.

The inconsistency in template concentrations can be explained by assessing the results displayed in **Table 15**. In the case of the 1° PCR, the cause of this inconsistency is the high standard deviation of the $^{260}/_{230}$ absorbance ratio. Samples with lower $^{260}/_{230}$ absorbance ratios due to contaminating ethanol (Siwach and Singh, 2007) will have displayed an exaggerated DNA concentration, resulting in insufficient DNA being added to the PCR and, consequently, a lower yield of PCR product. The converse can be said for samples with higher $^{260}/_{230}$ absorbance ratios. In the case of the inter-individual variation observed in the 2° PCR product, the mean concentration obtained following purification of the 1° PCR product should be considered. The mean concentration was 7.51 ng/µL with a large standard deviation of 4.18 ng/µL. Given that the lower limit of detection of the NanoDrop 2000 is 2 ng/µL (*Product Specifications*, http://www.thermoscientific.com/) and that the $^{260}/_{280}$ and $^{260}/_{230}$ absorbance ratios were erratic with a median of 2.16 and 0.02 respectively, it is likely that the quantification was inconsistent and that there was variation in the template concentration used in 2° PCRs.

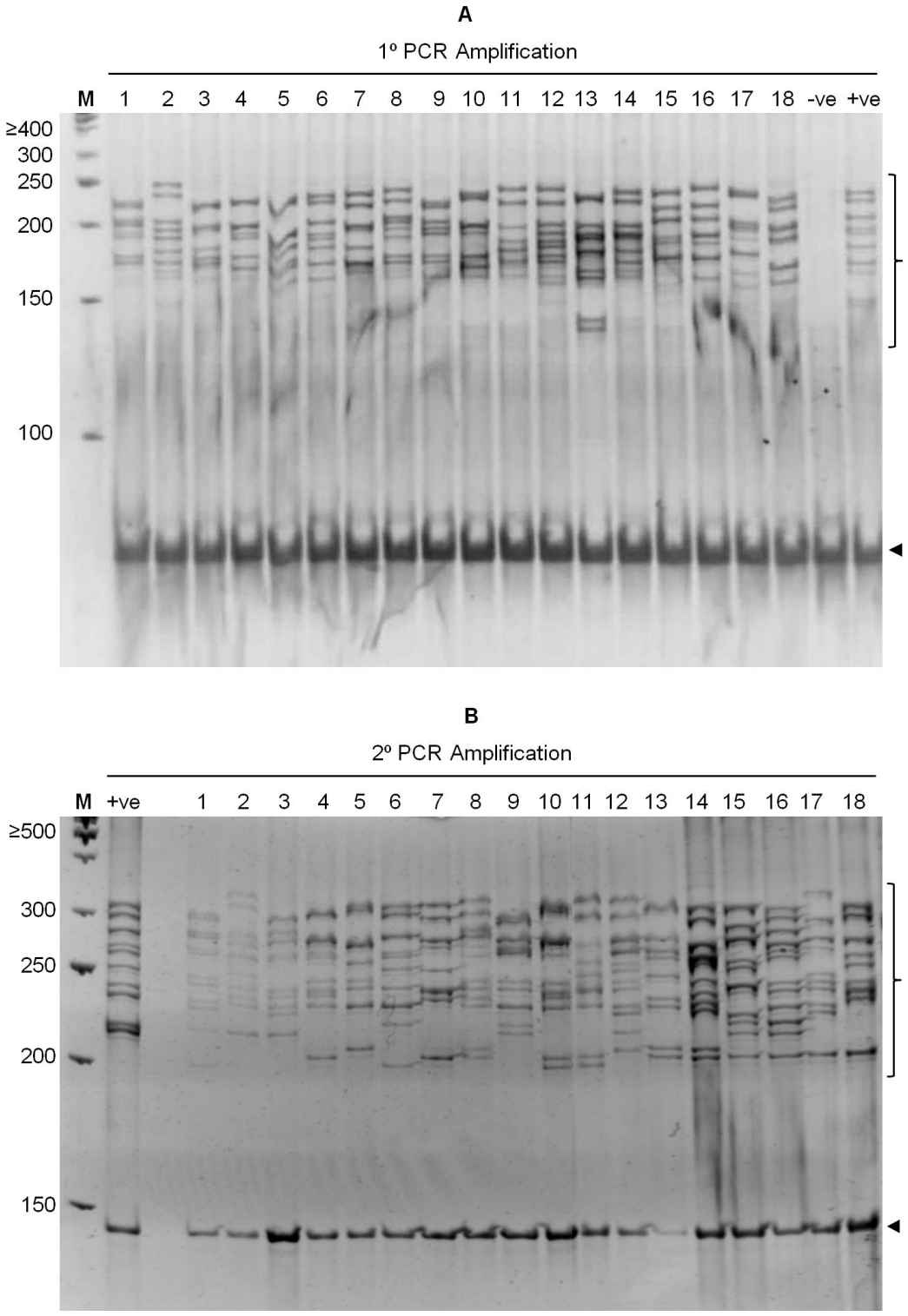


Figure 21: Example of sample processing protocol illustrated using DNA extracted from Nguni Subjects 1-18.

Polyacrylamide gel electrophoresis (4% $^{w}/_{v}$ stacking, 10% $^{w}/_{v}$ resolving 19:1 acrylamide:bisacrylamide at 5-8 V/cm in TBE) of 1° (**A**) and 2° (**B**) PCR products for Subjects 0 (+ve) and Nguni Subjects 1-18 amplified using Kapa HiFi HotStart ReadyMix. Thermo Scientific GeneRuler 50 bp at 0.07 µL/mm (**M**) was used to identify target amplicons (indicated by brackets). These bands were excised and purified using the "crush and soak" method and the Thermo Scientific GeneJET Gel Extraction kit. The negative control (-ve) was prepared by using dH_2O in place of template DNA. Purified 1° PCR products from **A** were used as the template for the 2° PCR, the amplicons of which are displayed in **B**. The ◄ marks all primer artefacts.

Table 15: Average concentrations and absorbance ratios for purified DNA measured using the Nanodrop 2000

Statistical Parameter	Genomic DNA			1° Product	2° Product
	ng/µL	$^{260}/_{280}$	$^{260}/_{230}$	ng/µL	ng/µL
Average	180.31	1.63	0.94	7.51	7.3
SD	92.91	0.06	0.16	4.18	3.1

Samples were measured at 220-330 nm using the ThermoScientific Nanodrop 2000. Standard Deviations (SD) and mean (average) readings for 144 subjects are presented.

Intra-individual variation in band intensity of the profiles was also observed, for which there are two likely causes: Firstly, homozygous alleles which result in bands roughly 2 times brighter than those arising from heterozygous alleles. This was evident in the band arising for D2S441 versus those for D13S317 for Subject 0. Secondly, mutations within primer binding sites which result in alterations in the efficiency of primer annealing. Whether this latter cause was in fact the case in any of the subjects tested was, however, beyond the scope of this study.

3.2.2 Pre-NGS quality check

3.2.2.1 PAGE of randomly selected purified 2° PCR products

Following PAGE purification of the 2° PCR products for Subjects 0-144, six arbitrarily selected PCR products, together with the pooled positive and negative controls used for DNA processing, were electrophoresed to ascertain whether the protocol was successful and free from contamination. The resulting gel is displayed in **Figure 22**.

All samples were free from primer artefacts and contamination. The negative control was free from bands and the profile observed for Subject 0 was consistent with previously observed results. Intra-individual variation within profiles and inter-individual variation between profiles for the Nguni and ST subjects was observed, the reasons for which are likely as discussed in *4.2.2*. The results observed for Subject 101 are notable because no lower size range bands were observed. Since the theoretical upper range for the smaller loci D2S441, D13S317 and D3S1358 is 230, 252 and 263 bp respectively, it is possible that the lack of ~200 bp products is not indicative of an incomplete profile, but is due rather to large alleles for these loci for this particularsubject.

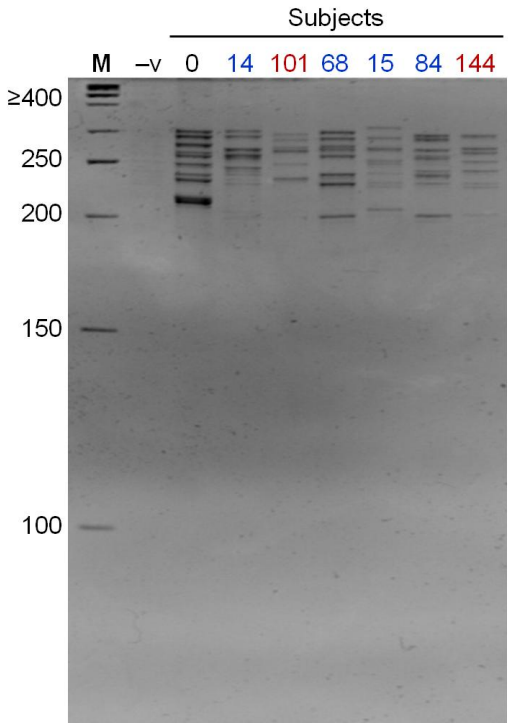


Figure 22: Assessment of the quality of DNA used for preparation of emPCR library by polyacrylamide gel electrophoresis (4% $^{w}/_{v}$ stacking, 10% $^{w}/_{v}$ resolving 19:1 acrylamide:bisacrylamide at 8 V/cm in TBE).

All PCR products were prepared using the Kapa HiFi HotStart ReadyMix and 2° PCR primers labelled with either Nguni or Sotho-Tswana MIDs. The samples were purified using the "crush and soak" method and concentrated using the Thermo Scientific GeneJET Gel Extraction kit. The marker (**M**) used for size determination of DNA products was the Thermo Scientific GeneRuler 50 bp (0.07 µL/mm).The negative (-ve) and positive (Subject 0) controls were prepared by pooling all the 2° PCR amplicons for each control and concentrating them using the Thermo Scientific GeneJET columns.

3.2.2.2 Sanger sequencing of purified 2° PCR product

DNA from Subject 15 of the Nguni-population (**Figure 21**) was selected for pre-NGS quality checking by Sanger sequencing. This sample was selected because its profile exhibited high resolution between amplicons, suitable for individual band excision. The yield of the 2° PCR product was increased to a concentration sufficient for Sanger sequencing by preparing 18 replicates of the reaction, the electrophoresis products of which can be observed in **Figure 23**. Amplicons were excised from this gel, combined with like amplicons, and purified using the "crush and soak" method. Thereafter the amplicons were concentrated and sent for Sanger sequencing. During the gel excision of Amplicon 4, 9 of the replicate samples were contaminated and were consequently discarded. Only 9 replicates of this amplicon were therefore sent for sequencing.

Both intra- and inter-reaction variation in band intensity are evident in **Figure 23**. The causes of intra-reaction variation are likely as previously described. Conversely, the inter-reaction

variation is likely a result of ineffective mixing of the running buffer over the extended electrophoresis time, which resulted in centre samples running hotter and faster than the samples near the edges of gels (*Performing Electrophoresis*, www.bio-rad.com).

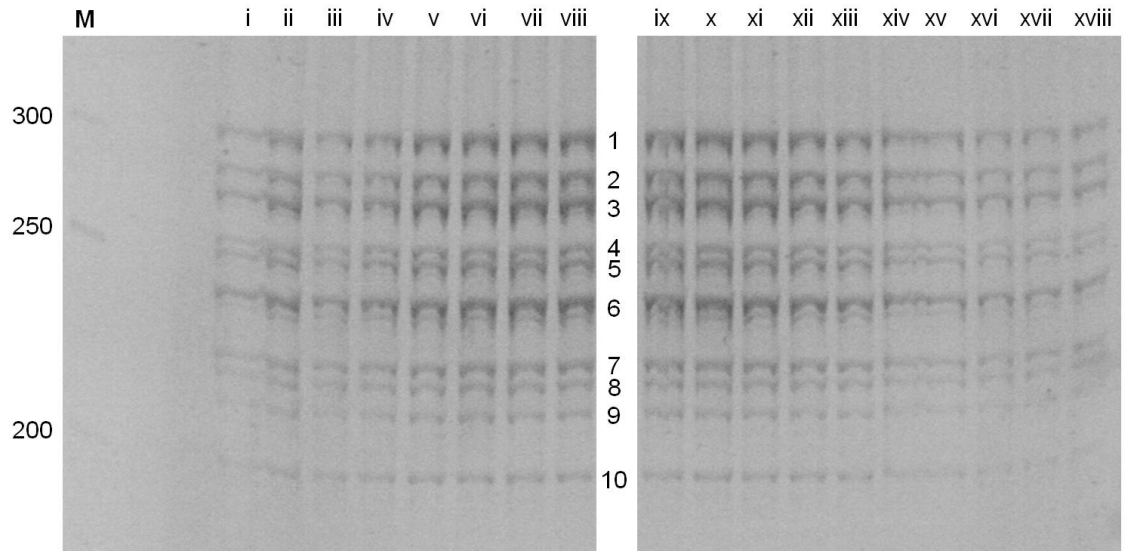


Figure 23: Polyacrylamide gel (4% $^{w}/_{v}$ stacking, 10% $^{w}/_{v}$ resolving 19:1 acrylamide:bisacrylamide electrophoresed at 8 V/cm in TBE)l used for purification of alleles amplified using 2° PCR for validation of DNA processing protocol

Lanes i-xviii are replicate PCRs performed using the Kapa HiFi HotStart ReadyMix, 1° PCR product for Subject 15, and 2° PCR primers labelled with Nguni MIDs. Each amplicon (1-10) was purified using the "crush and soak" method and concentrated using the Thermo Scientific GeneJET Gel Extraction kit. The marker (**M**) used for size determination of DNA products was the Thermo Scientific GeneRuler 50 bp (0.07 µL/mm).

The results obtained following Sanger Sequencing of the amplicons are displayed in **Tables 16** and **17**. As in *2.2.5* **Table 16** shows the actual sequences obtained following Sanger sequencing and alignment, while **Table 17** shows the allele calls and sizes, both expected and actual, of the amplicons. **Table 17** also includes a locus call determined by the STR-specific primer sequence observed in **Table 16.** No sequence was obtained for Amplicon 4 because the concentration of the sample was too low to permit accurate base-calling. Nevertheless, the results demonstrated that the protocol resulted in the amplification of all target loci using M13- and then fusion-tailed primers. The results in **Table 17** also indicate that (with the exception of amplicon 4) sequencing results represent each amplicon in its entirety, since amplicon sizes estimated from **Figure 21** correspond to within 2 bp of those obtained by sequencing.

Table 16: Sanger sequencing results for 2° PCR products of Subject 15 to validate the DNA processing protocol

Amplicon	Sequence
1	*T*CGTATSGCCTC**ACGAGTGCGT**SGSCAGGGTKKTCCCAGTCASGACAKATGTGAGTCAATTCCCCAAGTGAATTGCCT**TCTATCTATCTATCTATCTATCTATCTGTCTGTCTGT CTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTA**TCGTCTATCTATCCA GTCTATCTACCGTCATAGCTGTTKCCTGTGTGA**ACGCACTCGT**CTGAGCGGGCTGGCAAGGCGCATAG*ACAC*
2	*T*CGTATCGCCTCCCTCGCGCCATCAG**ACGAGTGCGT**CGCCAGGGTTTTCCCAGTCACGACATATGTGAGTCAATTCCCCAAGTGAATTGCCT**TCTATCTATCTATCTATCTGTC TGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCTATCTATCTA**TCGTCTATCTATCCAGTCTA TCTACCGTCATAGCTGTTTCCTGTGTGA**ACGCACTCGT**CTGAGCGGGCTGGCAAGGCGCATAG*A*
3	*GGT*CGTATCGCCTCCCTCGCGCCATCAGACGAGTGCGTCGCCAGGGTTTTCCCAGTCACGACAACACTTGTCATAGTTTAGAACGAACTAAC**GATAGATAGATAGATAGATA GATAGATAGATAGATAGATAGATAGATA**GACAGATTGATAGTTTTTTTTTATCTCACTAAATAGTCTATAGTAAACATTTAATTACCAATATGTGGTGCAATTCTGTCAATGAGTC ATAGCTGTTTCCTGTGTGA**ACGCACTCGT**CTGAGCGGGCTGGCAAGGCGCATAG*A*
5	*T*CGTATCGCCTCCCTCGCGCCATCAG**ACGAGTGCGT**CGCCAGGGTTTTCCCAGTCACGACCYMTAGTGGATGATAAGAATAATCAGTATGTGACTTGGATTGA**TCTATCTGTC TGTCTGTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCCATCTA**TCCATCCATCCTATGTATTTATCATCTGTCCGTCATAGCTGTTTCCTGTGTGA**ACGCA CTCG**TCTGAGCGGGCTGGCAAGGCGCATAGAAG
6	CGTATCGCCTCCCTCGCGCCATCAG**ACGAGTGCGT**SGCCAGGGTTTTCCCAGTCACGACACTGCAGTCCAATCTGGGTGACAGAGCAAGACCCTGTCTCA**TAGATAGATAGA TAGATAGATAGATAGATAGATAGATAGATAGATAGACAGACAGACAGATAGA**TACATGCAAGCCTCTGTTGATTTCATGTCATAGCTGTTTCCTGTGTGA**ACGCACTCG**TCTG AGCGGGCTGGCAAGGCGCATAG*A*
7	TCGTATCGCCTCCCTCGCGCCATCAG**ACGAGTGCGT**CGCCAGGGTTTTCCCAGTCACGACTCTGACCCATCTAACGCCTATCTGTATTTACAAATACATT**ATCTATCTATCTAT CTATCTATCTATCTATCTATCTATCTATCT**ATCAATCATCTATCTATCTTTCTGTCTGGTCATAGCTGTTTCCTGTGTGA**ACGCACTCG**TCTGAGCGGGCTGGCAAGGCGCATAG
8	*CGT*CGTATCGCCTCCCTCGCGCCATCAGACGAGTGCGTCGCCAGGGTTTTCCCAGTCACGACTCTGACCCATCTAACGCCTATCTGTATTTACAAATACAT**TATCTATCTATCT ATCTATCTATCTATCTATCTATCTATCT**ATCAATCATCTATCTATCTTTCTGTCTGGTCATAGCTGTTTCCTGTGTGA**ACGCACTCG**TCTGAGCGGGCTGGCARGGCGCATAG*AA G*
9	*T*CGTATCGCCTCCCTCGCGCCATCAG**ACGAGTGCGT**CGCCAGGGTTTTCCCAGTCACGACCTGTGGCTCATCTATGAAAACT**TCTATCTATCTATCTATCTATCTATCTATCTA TCTATCTATCTATCTATCTATCTA**TATCATAACACCACAGCCACTTCGTCATAGCTGTTTCCTGTGTGA**ACGCACTCG**TCTGAGCGGGCTGGCAAGGCGCATAG*AAG*
10	*T*CGTATCGCCTCCCTCGCGCCATCAG**ACGAGTGCGT**CGCCAGGGTTTTCCCAGTCACGACCTGTGGCTCATCTATGAAAACT**TCTATCTATCTATCTATCTATCTATCTATCTA TCTATCTATCTA**TATCATAACACCACAGCCACTTCGTCATAGCTGTTTCCTGTGTGA**ACGCACTCG**TCTGAGCGGGCTGGCAAGGCGCATAG*A*

Sequencing was performed by Inqaba Biotech in the forward and reverse directions. Forward and reverse sequences were aligned using MAFFT v 7 and visualised using BioEdit v 7.5.2 to obtain a consensus sequence. Underlined regions correspond to the primer binding sites; blue regions are fusion primer sequences with **bold** regions indicating the Nguni-Specific MID. Coloured regions are STR-specific primer binding sites. **Bold** segments are STR units. Known SNPs (A/T and T/G) are highlighted.

Table 17: Locus and Allele calls and comparison of actual to expected sizes for sequenced STRs

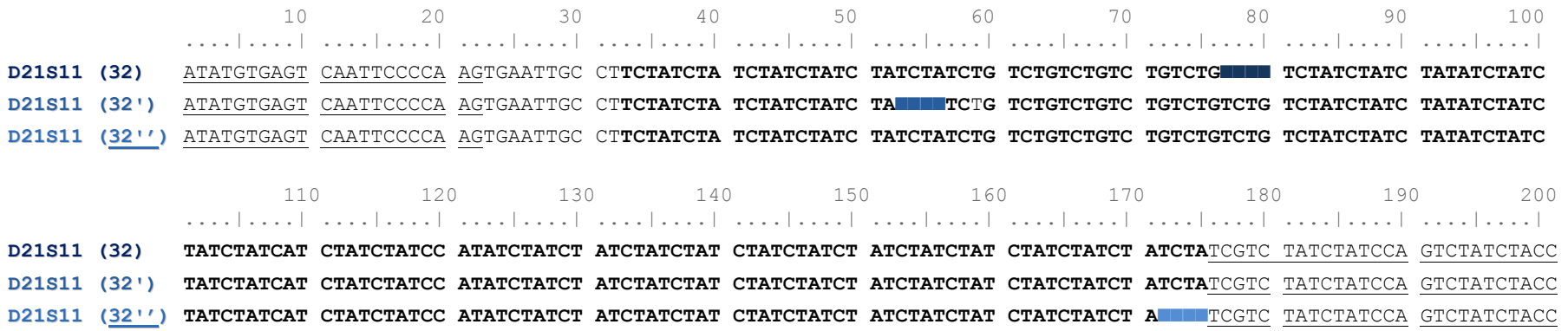
Amplicon	Locus	Allele Call	Identity Matrix Score	Actual Size (bp)	Expected Size (bp)
1	D21S11	32''	0.944 [32]	313	313
2	D21S11	27	1.000	293	293
3	D7S820	12_TG	1.000	279	278
4	n/a	n/a	n/a	n/a	262
5	vWA	15´(17')	1.000	258	256
6	D3S1358	16''	0.991 [16]	247	245
7	D13S317	11_T	1.000	228	227
8	D13S317	10_T	1.000	224	223
9	D2S441	14	1.000	217	218
10	D2S441	11	1.000	205	206

Allele calls were made by aligning consensus sequences (excluding fusion primers and M13 and with corrected STR-specific primer regions) to a database of known alleles. The SNP calls for loci containing known SNPs are indicated adjacent to the underscore, in order of 3´ to 5´ in the case of D7S820. Allele calls in red indicate putative novel variant alleles. The maximum Identity Matrix Score was generated using BioEdit v 7.5.2 with 1.000 representing 100% identity. [Square brackets] enclose the allele for which the maximum Identity Matrix Score was obtained for putative variant alleles. The expected size of each amplicon was estimated from **Figure 21** using GelQuant. The actual size of the amplicon was given by the number of nucleotides sequenced.

The loci and allele calls displayed in **Table 17** illustrate that 7 of the 9 sequences corresponded to known alleles. The remaining two sequences represent putative variants for the loci D21S11 and D2S1358 since the quality of the sequencing results was high and there were no ambiguous base-calls within the STR-specific primer binding region, flanking regions or repeat units. Alignments of these putative variants to the alleles which they exhibited the highest Identity Matrix Score are displayed in **Figure 24**. The STR sequence for the consensus sequence of D21S11 in **Figure 24** is $[TCTA]_6$ $[TCTG]_{6...}[TCTA]_{12}$ whereas the known sequences for alleles of the same size are $[TCTA]_6$ $[TCTG]_5$…$[TCTA]_{13}$ and $[TCTA]_5$ $[TCTG]_6$…$[TCTA]_{13}$. The STR sequence for the consensus sequence of D3S1358 in **Figure 24** is $[TAGA]_{11}$ $[CAGA]_3$ $[TAGA]_2$ whereas the known sequences for alleles of the same size are $[TAGA]_{12}$ $[CAGA]_3$ [TAGA] and $[TAGA]_{13}$ $[CAGA]_2$ [TAGA]. In both cases, the putative variants are consistent with the type of variants currently recorded for these STRs (STRbase, www.cstl.nist.gov/strbase).

.

Amplicon 1



Amplicon 6

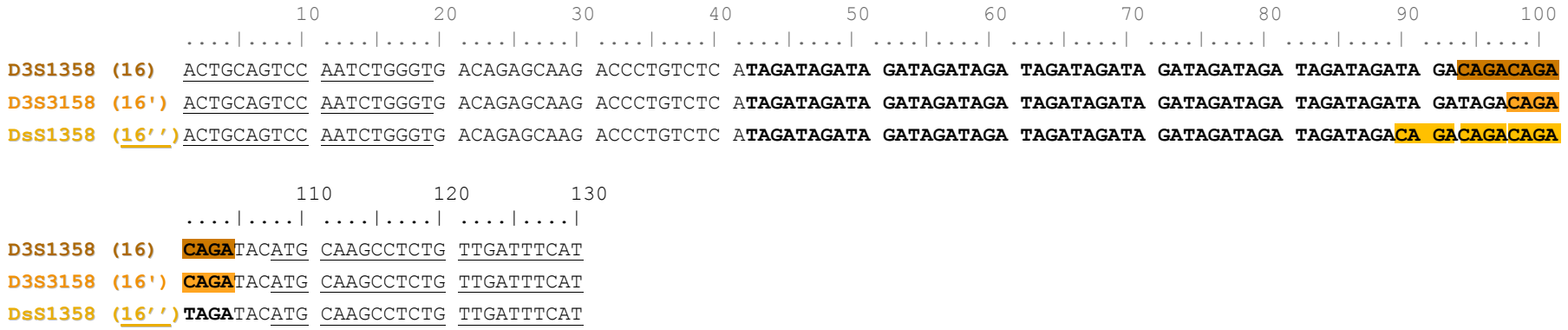


Figure 24: Putative novel variants for loci D21D11 and D3S1358 obtained by Sanger Sequencing and aligned against alleles of corresponding lengths

UNDERLINED allele calls indicate the putative novel variants. These sequences were obtained by creating a consensus sequence in BioEdit v 7.5.2 from the aligned (in MAFFT v 7) forward and reverse sequences obtained using Sanger Sequencing by Inqaba Biotech. Underlined regions correspond to the STR-specific primer binding sites and **bold** segments to STR units. Blue boxes (■■■) indicate insertions of 4-bp for each allele required to enable alignment of **Amplicon 1** to like-sized alleles. **CAGA** repeat units in **Amplicon 6** are highlighted in various shades of yellow.

3.2.3 454 Sequencing Results

The results obtained in *3.2.2* indicated that the DNA processing protocol produced samples containing target STRs with M13- and fusion-tails, uncontaminated by primer dimers or non-target amplicons. The samples were, therefore deemed ready for library preparation and subsequent 454 sequencing, using the protocol outlined in *3.1.4*. The pooled library, containing 2.4 ng of DNA per individual as measured by the Nanodrop 2000, was used for emPCR and sequencing using the Roche GS Junior System.

3.2.3.1 Overview of processed sequencing results

The sequencing reaction produced 145 485 reads, of which 97 400 were labelled with the MID specific to the Nguni population and 48 085 with the MID specific to the ST population. These reads were processed as outlined in *3.1.5*, the results of which (following step **[6]** of **Figure 19**) are summarised in **Figure 25**.

The number of reads obtained for the Nguni population was roughly double that obtained for the ST population (evidenced in **Figure 25 A**); with the proportion of reads per locus per population group being similar (evidenced in **Figure 25 B**). The proportion of reads per locus within each population group, however, exhibited some imbalance; with vWA, in particular, being under-represented. This observed imbalance did not, however, seem to favour a particular size range or GC-content of amplicons. For example, D2S441 and D21S11 are the loci with the smallest and largest amplicons respectively, with D2S441 containing TCTA repeat units and D21S11 containing TCTA and TCTG repeat units. Despite these differences, the proportion of sequences obtained for D21S11 was higher than that obtained for D2S441 for the ST population and proportions for each of these loci were similar for the Nguni population, evidenced in **Figure 25 B**.

The aforementioned loci proportions were determined using uncompressed sequence data. **Figure 25 A** enables the comparison of data obtained with and without homopolymer compression. These data reveal that loci exhibited small increases of 0.9-2.6% in the number of reads obtained when homopolymer compression was applied. D2S441 was an exception, which exhibited a large increase of 11.8 and 9.8% in the number of reads for Nguni and ST populations respectively; likely due to the adenine 4-mer (depicted in **Table 8**) found in the forward primer sequence of D2S441 which is vulnerable to the inaccuracies associated with 454 sequencing of homopolymeric regions (Zascavage *et al.*, 2013). Conversely, the number

of reads which did not match any known sequence, decreased by 16.5 and 12.2% when homopolymer compression was applied.

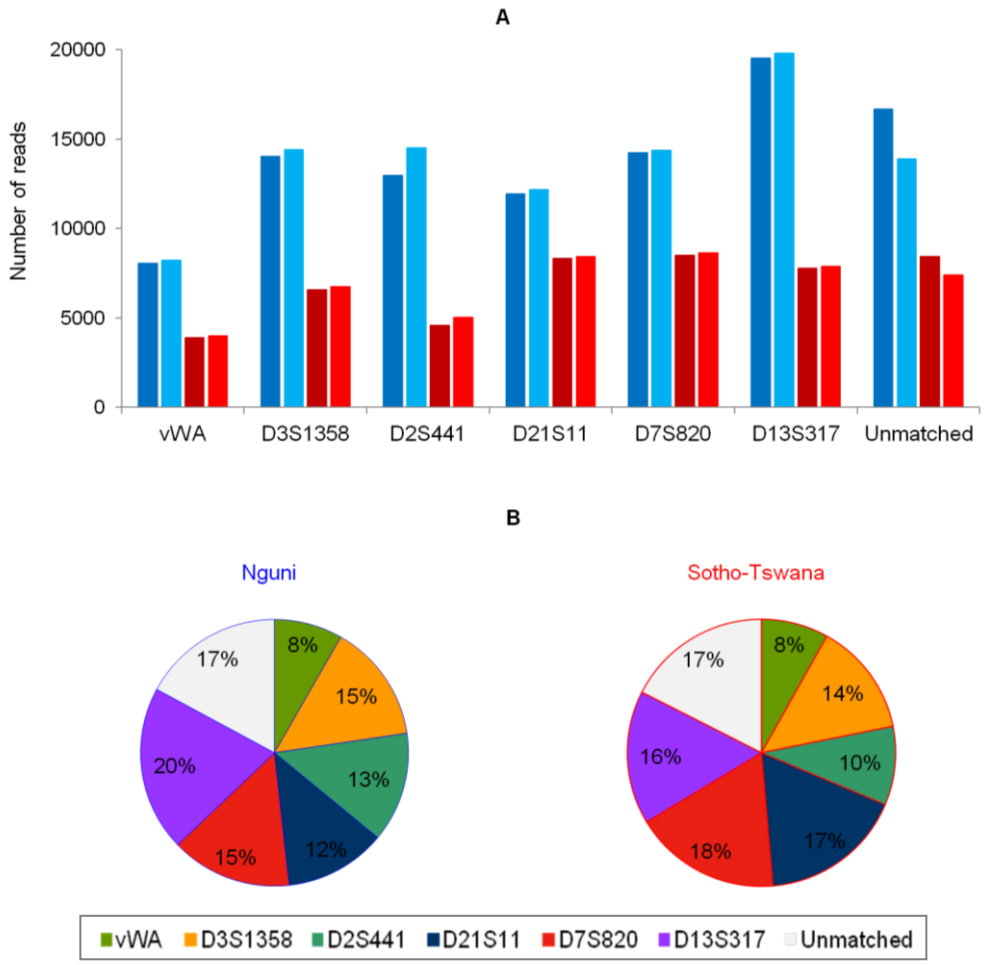


Figure 25: The number of reads obtained per locus per population group using raw and/or homopolymer compressed data
The **blue** and **red** bars in **A** represent data for the Nguni and Sotho-Tswana (ST) populations respectively. Darker shaded bars (on the left) were constructed using uncompressed data while the lighter shaded bars (bars on the right) were constructed from homopolymer compressed data. **B** was constructed using the uncompressed sequencing data and illustrates the proportion of each locus within a population-specific set of sequences.

The results presented in **Figure 26** further serve to illustrate the similarity in sequencing data with and without application of homopolymer compression. These graphs were constructed by plotting the frequencies for each called allele obtained from uncompressed data against the frequencies for each corresponding allele obtained from homopolymer compressed data. Allele calls were made following completion of step **[8]** (**Figure 19**) and subsequent to the identification of 'true' alleles. Alleles were deemed 'true' if they were represented by at least 84 reads, determined by dividing the total number of reads (145 485) by the theoretical minimum number of reads per allele (144*12=1728).

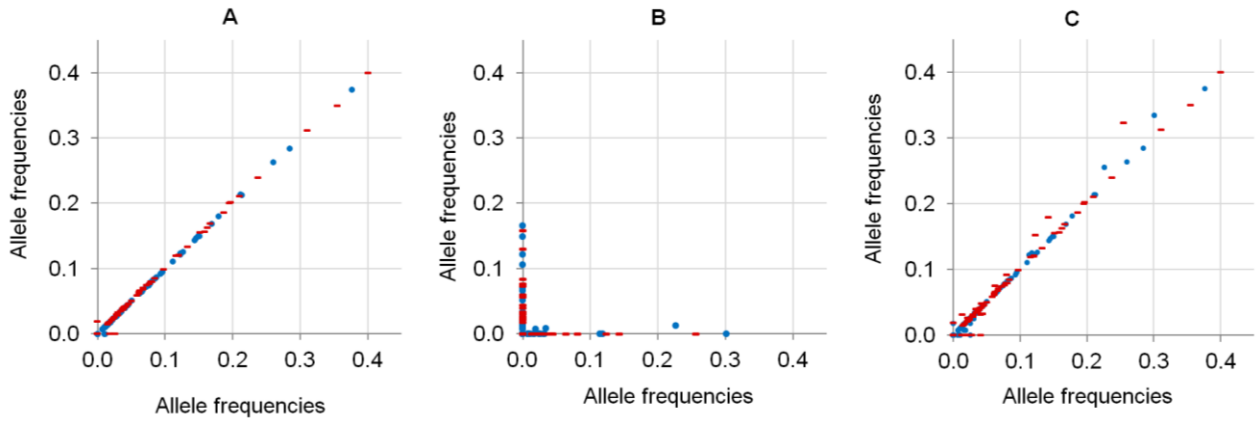


Figure 26: Scatter plot of the allele frequencies obtained with (x-axis) and without (y-axis) homopolymer compression for Nguni (**blue**) and Sotho-Tswana (**red**) populations

A was constructed using the allele frequencies for D2S441, D3S1358, vWA, D13S317 and D21S11. **B** was constructed using the allele frequencies for D7S820. **C** was constructed using the allele frequencies for all loci. The allele calls for the raw data of D7S820 in **C** were modified so that all alleles identical but for the number of T's in the 5´ 8-10mer, were regarded as a single allele.

Figure 26 A demonstrates that the allele frequencies and allele calls for all loci, with the exception of D7S820, for each population group were almost identical, exhibiting a linear relationship with gradients of 1.00 and a R^2 values of >0.99. This indicates that homopolymer compression of these loci had little to no effect on the allele calls or observed allele frequencies. Conversely, the results depicted in **Figure 26 B** illustrate clearly that homopolymer compression had a marked effect on both allele calling and allele frequencies obtained for D7S820. This discrepancy was resolved by sequence analysis. The homopolymeric string of 8-10 thymine residues in the 5´ flanking region of the locus, described in *1.4.3*, were truncated in the sequencing data to either 5, 6 or 7 residues. This truncation resulted in an exaggeration in the total number, and number of putative variant alleles, reported for D7S820.

Because analysis of uncompressed data enables the highest resolution of sequences, and given the high similarity between compressed and uncompressed data for all loci except for D7S820, it was decided that uncompressed data should be used for allele-calling of vWA, D2S441, D3S1358, D13S317 and D21S11. To enable D7S820 to be analysed alongside these loci, the sequences for the uncompressed D7S820 were manually edited to remove the effects of the erroneously called poly-thymine region. More specifically, any alleles that were identical but for the number of thymine residues were called as a single allele. Those which were identical to known alleles but for the number of thymine residues were, similarly, called as the known allele. The resultant alleles and corresponding allele frequencies, together with those for all other loci, were plotted as before to create the scatter plot in **Figure 26 C**. The

correlation between data was slightly less than that observed when D7S820 was excluded, with the gradients and R^2 values of the trendlines being 1.05 and 1.07, and 0.99 and 0.97 for Nguni and ST populations respectively. In most cases, points that deviated from the linear relationship were a result of higher allele frequencies for homopolymer compressed, as opposed to uncompressed, data for common alleles of D7S820. Some of these deviations, however, were a result of allele calls specific to either compressed or uncompressed data (points on the axes of the graph). These unique alleles were observed at low frequencies of below 0.04 for D7S820 and 0.026 for other loci.

Using the uncompressed data and the modified data for D7S820, a complete list of observed alleles and corresponding allele frequencies was constructed, which can be viewed in **Table ii** of the **Appendix**. **Table 18** was also constructed from this data, which summarises the processed and sorted sequencing data and describes the numbers of unique and retained unique sequences (sequences with at least 84 reads), as well as the proportion of the sum of retained sequences to that of the total number of reads per locus.

Table 18: Summary of unique sequence data obtained following processing and sorting of 454 sequencing data for Nguni and Sotho-Tswana (ST) populations.

Loci	Unique Sequences		Retained unique sequences		Proportion of the sum of reads retained	
	Nguni	ST	Nguni	ST	Nguni	ST
vWA	340	179	16	11	0.880	0.807
D3S1358	572	254	17	14	0.919	0.906
D2S441	183	80	14	10	0.953	0.909
D21S11	938	605	21	14	0.720	0.722
D7S820	785	645	12	11	0.783	0.647
D13S317	391	182	10	9	0.938	0.934
Unmatched	7741	4116	9	2	0.273	0.270

The number of unique sequences observed for each locus was counted (**unique sequences**). Sequences observed at least 84 (the theoretical minimum read number per unique allele) times were retained (**retained unique sequences**). The **proportion of the sum of reads retained** was calculated by dividing the absolute number of sequences contained within the **retained unique sequences** set by the total number of reads obtained for the corresponding locus.

The imbalance between Nguni and ST data was evidenced in **Table 18** as in **Figure 25**, with the number of unique reads per locus for Nguni data often being double that obtained for ST data. D21S11 and D7S820 were notable exceptions, corresponding to the results depicted in **Figure 25 B** for these loci. Interestingly however, considerably more alleles were retained for D21S11 of the Nguni as opposed to the ST population, despite the high total number of unique sequences for this allele. There does not seem to be a relationship, therefore, between total number of sequences, total number of unique sequences, and total number of retained

unique sequences. This is also evidenced when considering the proportion of the sum of retained reads to the total number of sequenced reads per locus, with proportions ranging from 0.647 for ST D7S820 sequences to 0.953 for Nguni D2S441 sequences.

The unmatched sequences described in **Table 18** were comprised largely of unique sequences, for which only 9 and 2, for Nguni and ST population groups respectively, were represented by at least 84 reads. These retained unmatched reads were primer dimers, the sequences for which are listed in **Table iii** of the **Appendix**. The low proportion of retained unmatched reads (0.27) is an indication that the unmatched reads were likely a result of low-frequency sequencing errors.

3.2.3.2 Population specificity of allele-calls

Using the allele calls and corresponding frequencies listed in **Table ii**, **Figures 27** and **28** were constructed. **Figure 27** provides an overview of the similarity in allele calls and frequencies between the populations; while **Figure 28** depicts the precise allele calls, the nature of the particular allele (variant or reported), as well as the frequencies of each allele per population group.

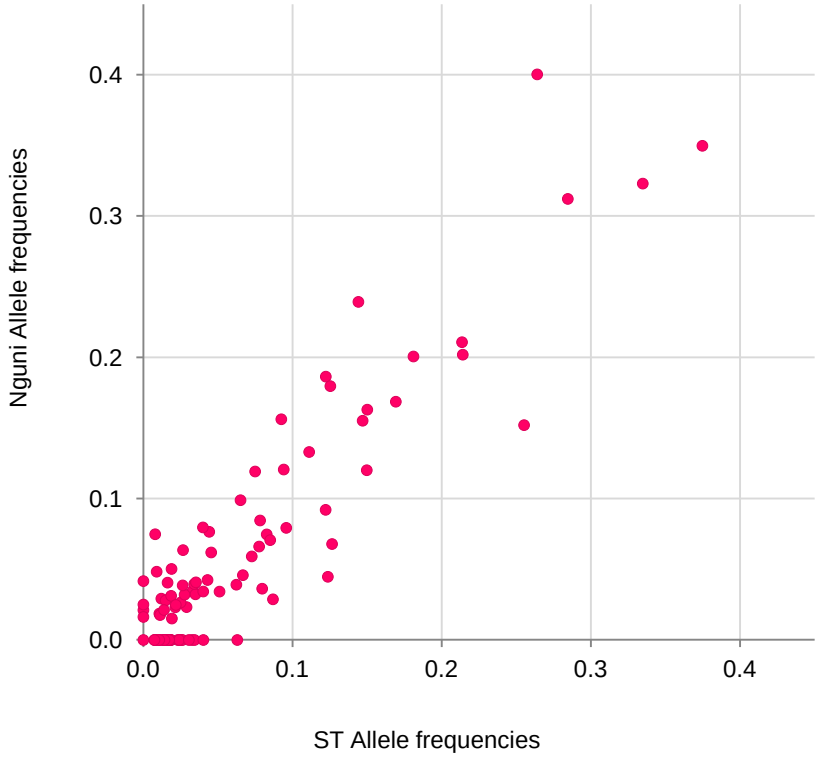


Figure 27: Correlation between the allele frequencies obtained for Nguni (y-axis) and Sotho-Tswana (ST) (x-axis) populations.

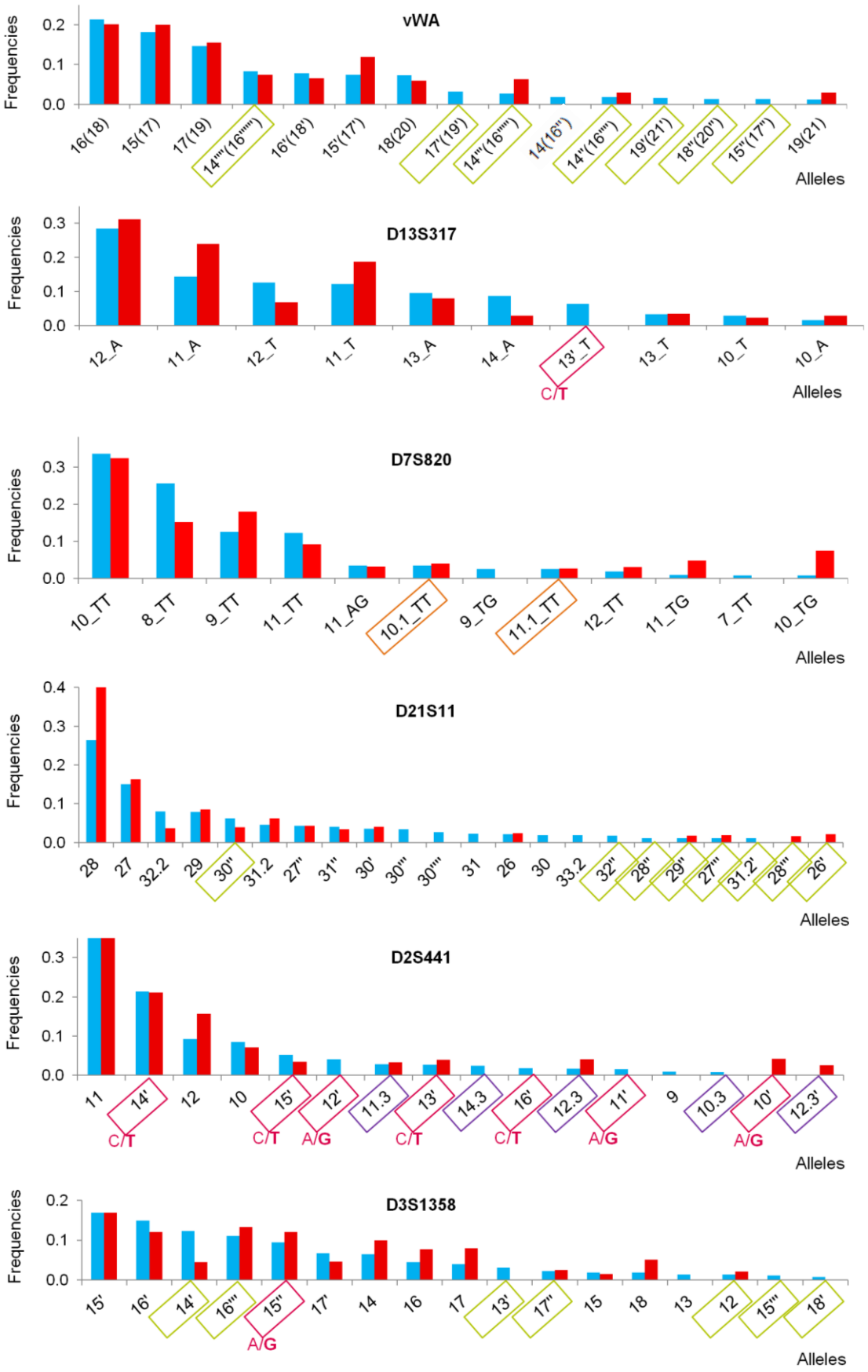


Figure 28: Allele frequencies for target STRs for Nguni (**blue**) and Sotho-Tswana (**red**) population groups.

Coloured outlines indicate putative novel STR variants. **Green** outlines indicate repeat unit variations; **pink** outlines indicate SNPs, with the bold letter indicating the base to which the unbolded letter has been replaced; **orange** outlines indicate flanking region insertions; and **purple** outlines indicate single base pair deletions.

The results in **Figure 27** illustrate the similarity in allele calls and frequencies obtained for Nguni and ST populations. Some higher frequency alleles were however observed slightly more often in one or other population group. For example, the allele 8_TT for D7S820 was observed 1.6 times more frequently in the Nguni population group than the ST population group; whereas the allele 9_TT for D7S820 was observed 1.4 times more frequently in the ST population group than in the Nguni population group. Furthermore, some alleles, all with allele frequencies of below 0.1, were observed in only the Nguni or ST population groups (indicated by points on the axes).

The data used to construct **Figures 27** and **28** were also assessed by ANOVA (p=0.05), both for single STR loci and altogether. The results obtained revealed consistently lower F-statistic values to the F-critical values, indicating that no significant difference between the population groups was observed. These data, however, should be approached with caution due to the imbalance in sequencing data. More specifically, although equimolar amounts of DNA per individual were added to the sequencing reaction, the sequencing results favoured the Nguni population 2-fold and exhibited imbalance in loci representation. Furthermore, if we consider the 9 allele-calls for Subject 15, made in *3.2.2* (**Table 17**), 4 of these alleles (allele 11 and 14 of D2S441; allele 12_TG of D7S820; and variant allele 16'' of D3S1358) were not represented in the sequencing results. This suggests that there was bias not only in population and locus representation; but also in individual representation.

Also notable in the sequencing results were the SNP allele frequencies observed for rs7789995 and rs7786079 (within D7S820) and rs9546005 (within D13S317), summarised in **Table 19**. The frequency of the T allele of rs7789995 was similar to that reported in HapMap for African populations (0.007-0.058) (http://hapmap.ncbi.nlm.nih.gov/). Furthermore, the frequency at which this rare allele was observed in the ST population group was nearly double that for the Nguni population group. Similarly, the rare allele of rs7786079 exhibited a frequency in the ST population group three times that of the frequency observed in the Nguni population. Only a single comparison to African population data (people of the Yoruban tribe in Nigeria) could be made for this latter SNP (http://hapmap.ncbi.nlm.nih.gov/). The frequency of the G allele in Yorubans was 0.730, considerably lower than that reported for the Nguni and ST population groups in this study. In contrast to the results observed for the D7S820 SNPs, the comparatively rare T allele for rs9546005 was observed at a higher frequency in the Nguni population group compared to the ST population group, and the locus exhibited a more even allelic distribution. No comparisons to existing population data could

be made for rs9546005 however, since there is no population data available for the SNP. It should be noted that, as with the STR allele frequencies, the value of the SNP frequency data outlined in **Table 19** is limited by the previously discussed sequencing biases

Table 19: The SNP ratios observed for known SNPs in loci D7S820 and D13S317.

Alleles	rs7789995		rs7786079		rs9546005	
	Nguni	ST	Nguni	ST	Nguni	ST
A	0.958	0.926	-	-	0.624	0.688
T	0.042	0.074	-	-	0.373	0.312
G	-	-	0.945	0.833	-	-
T	-	-	0.055	0.167	-	-

.

3.2.3.3 Identification of putative novel STR variants

While comparison of allele frequencies within STRs and between population groups was limited by the experimental design and sequencing biases, qualitative analysis with regards to identification of novel STRs variants (alleles unreported in the STRBase, www.cstl.nist.gov/strbase) was possible. **Figure 28** provides a complete list of the allele calls, both variant and reported, for each locus. **Tables 20** provides a summary of the characteristics of the STR variants and **Table 21** provides the sequences of these alleles.

Table 20: Description of the novel STR variants observed during the analysis of 454 sequencing data for Nguni and Sotho-Tswana population groups

Locus	Type	Description	Frequency	
			Nguni	ST
vWA	STR Variations	Various novel variations to the reported repeat unit sequences	0.206	0.168
D13S317	SNP (C/**T**)	Single SNP within the eighth repeat unit of allele 13	0.063	0.000
D7S820	Flanking insertion	Observed in alleles 10_TT and 11_TT	0.059	0.066
D21S11	STR Variations	Various novel variations to the reported repeat unit sequences	0.087	0.074
D2S441	SNP (A/**G**)	Second last repeat unit of alleles 10, 11 and 12	0.056	0.042
	SNP (C/**T**)	Third last repeat unit of alleles 13, 14, 15 and 16	0.308	0.284
	Deletion [TC_A]	Third or fourth repeat units of alleles 11, 12, 13 and 15	0.075	0.098
D3S1358	STR Variations	Various novel variations to the reported repeat unit sequences	0.512	0.345
	SNP (A/**G**)	First and eleventh repeat unit of allele 15	0.011	0.000

Sequences unreported in the STRBase (www.cstl.nist.gov/strbase) were classified as novel STR variants.

The data in **Figure 28** demonstrated that different loci exhibited different propensities for containing novel variants and for the type of variants which they contained. The complex and compound STRs, namely D21S11, vWA and D3S1358 discussed in *1.2.3* were more susceptible to containing variations in the repeat units of identically-sized STRs. For example, the compound marker vWA exhibited 4 variations of allele 14(16'') (**Table ii**). This allele has an additional sequence variation which is presented in **Table i**, namely 14'(16'''), which was not observed in this study. The proportion of novel variant reads to the total number of reads obtained for each these loci was, however, distinct. Novel variant reads accounted for 0.168-0.206 and 0.345-0.512 of the reads for vWA and D3S1358 respectively; whereas the novel variant alleles observed for the complex STR D21S11, accounted for only 0.168-0.206 of the total locus reads.

Table 21: Putative novel variant alleles obtained by 454 sequencing

Locus	Putative Allele Call	Sequence	Description
vWA	14"(16"")	TCTA [TCTG]$_3$ [TCTA]$_{10}$ **TCCA** TCTA	STR variation
	14""(16""")	TCTA [TCTG]$_4$ [TCTA]$_9$ **TCCA** TCTA	STR variation
	14"""(16""""")	TCTA **TCTG TCTA** [TCTG]$_4$ [TCTA]$_3$ **TCCA** [TCTA]$_3$ **[TCCA]$_2$**	STR variation
	15´'(17")	TCTA[TCTG]$_4$[TCTA]$_{12}$	STR variation
	17'(19')	TCTA [TCTG]$_3$ [TCTA]$_{13}$ TCCA TCTA	STR variation
	18"(20")	TCTA [TCTG]$_6$ [TCTA]$_{11}$ TCCA TCTA	STR variation
	19'(21')	TCTA [TCTG]$_6$ [TCTA]$_{12}$ TCCA TCTA	STR variation
D13S317	13´_T	[TATC]$_7$ TAT**T** [TATC]$_5$	SNP
D7S820	10.1_TT	..[GATA]$_{10}$ CAGATTG**T**ATAGTTTT..	Flanking insertion
	11.1_TT	..[GATA]$_{11}$ CAGATTG**T**ATAGTTTT..	Flanking insertion
D21S11	26'	[TCTA]$_4$ [TCTG]$_5$ [TCTA]$_3$ TA [TCTA]$_3$TCA [TCTA]$_2$ TCCA TA [TCTA]$_9$	STR variation
	27"'	[TCTA]$_4$ [TCTG]$_5$ [TCTA]$_3$ TA [TCTA]$_3$TCA [TCTA]$_2$ TCCA TA [TCTA]$_{10}$	STR variation
	28"	[TCTA]$_5$ [TCTG]$_5$ [TCTA]$_3$ TA [TCTA]$_3$ TCA [TCTA]$_2$ TCCA TA [TCTA]$_{10}$	STR variation
	28"'	[TCTA]$_4$ [TCTG]$_7$ [TCTA]$_3$ TA [TCTA]$_3$ TCA [TCTA]$_2$ TCCA TA [TCTA]$_9$	STR variation
	29"	[TCTA]$_5$ [TCTG]$_6$ [TCTA]$_3$ TA [TCTA]$_3$TCA [TCTA]$_2$ TCCA TA [TCTA]$_{10}$	STR variation
	30""	[TCTA]$_5$ [TCTG]$_5$ [TCTA]$_3$ TA [TCTA]$_3$ TCA [TCTA]$_2$ TCCA TA [TCTA]$_{12}$	STR variation
	31.2'	[TCTA]$_5$ [TCTG]$_6$ [TCTA]$_3$ TA [TCTA]$_2$ TCA [TCTA]$_2$ TCCA TA [TCTA]$_{12}$ TA TCTA	STR variation
	32"	[TCTA]$_6$ [TCTG]$_6$ [TCTA]$_3$ TA [TCTA]$_3$ TCA [TCTA]$_2$ TCCA TA [TCTA]$_{12}$	STR variation
D2S441	10'	[TCTA]$_8$ [TCT**G**] [TCTA]	SNP
	10.3	[TCTA]$_3$ **[TCA]** [TCTA]$_7$	Deletion
	11'	[TCTA]$_9$ [TCT**G**] [TCTA]	SNP
	11.3	[TCTA]$_4$ **[TCA]** [TCTA]$_7$	Deletion

D2S441 continued	12'	$[TCTA]_{10}$ [TCT**G**] [TCTA]	SNP
	12.3	$[TCTA]_{3}$ **[TCA]** $[TCTA]_{9}$	Deletion
	12.3´	$[TCTA]_{4}$ **[TCA]** $[TCTA]_{8}$	Deletion
	13´	$[TCTA]_{10}$ T**T**TA $[TCTA]_{2}$	SNP
	14'	$[TCTA]_{11}$ T**T**TA $[TCTA]_{2}$	SNP
	14.3	$[TCTA]_{3}$ **[TCA]** $[TCTA]_{11}$	Deletion
	15´	$[TCTA]_{12}$ T**T**TA $[TCTA]_{2}$	SNP
	16'	$[TCTA]_{13}$ T**T**TA $[TCTA]_{2}$	SNP
D3S1358*	12	**$[TAGA]_{10}$ [CAGA] [TAGA]**	Unreported allele
	13´	$[TAGA]_{11}$ $[CAGA]_{1}$ [TAGA]	STR variation
	14'	$[TAGA]_{12}$ $[CAGA]_{1}$ [TAGA]	STR variation
	15´'	$[TAGA]_{13}$ $[CAGA]_{1}$ [TAGA]	STR variation
	15´''	[T**G**GA] $[TAGA]_{10}$ [TAG**G**] [TAGA] $[CAGA]_{1}$ [TAGA]	SNP
	16'''	$[TAGA]_{14}$ $[CAGA]_{1}$ [TAGA]	STR variation
	17''	$[TAGA]_{15}$ $[CAGA]_{1}$ [TAGA]	STR variation
	18'	$[TAGA]_{15}$ $[CAGA]_{2}$ [TAGA]	STR variation

*Allele sequences are the reverse complement to those reported in the STRbase (www.cstl.nist.gov/strbase)
Sequence variations from those reported in the STRBase (www.cstl.nist.gov/strbase) are indicated in **red**.

The remaining loci were all of the simple STR class. The sequencing data for loci D13S317 and D7S820 contained only 1 and 2 novel STR variants respectively. The novel variant observed for D13S317 was a SNP in the seventh repeat unit of allele 13. It was not observed in the ST population group. However, its frequency was relatively low in the Nguni population and its lack of representation in the ST population group may, therefore, be due to sequencing bias. The novel variants observed in D7S820 were T insertions 8 bp downstream of the repeat unit, observed in alleles 10 and 11. These variants were observed at similar frequencies in both population groups (0.059 and 0.066 respectively).

D2S441 was the locus whose allele calls contained of the greatest number, both absolutely and relatively, of novel STR variants; which proportionately, comprised nearly half of the reads obtained for D2S441. The variants fell into classes, namely C/T and A/G SNPs and single base pair deletions. The C to T SNP in the third last repeat unit of alleles 13-16 was observed with high frequencies of 0.308 and 0.284 for Nguni and ST populations respectively; with allele 14' and 15´ being the alleles with the second and fifth highest frequencies for D2S441 respectively. The A to G SNP observed in the second last repeat unit of alleles 10-12, and the deletion observed in the third or fourth repeat units of alleles 11-13 and 15, were observed with lower frequencies of 0.056 and 0.042, and 0.075 and 0.095 for Nguni and ST populations respectively. This locus is also notable because it contained a number of lower frequency alleles that were unique to either the Nguni or ST population

group. Again however, these data need to be interpreted in appreciation of the observed sequencing bias.

3.2.3.4 Evaluating the usefulness of 454 sequencing for STR Typing

Analysis of 454 sequencing data enabled SNP-typing of known SNPs and the identification of putative novel STR variants. **Figure 29** summarises the fold increase in the number of alleles for the loci typed using 454 sequencing, compared to that which would have been permitted using conventional CE-FD. The fold increase ranged from between 0.333 (for D2S441) to 1.667 (for vWA). Because there were a greater proportion of novel variant alleles observed for D2S441, this value was underestimated and a second analysis was therefore preformed which included STRbase reference sequences. The fold increase was constant for all alleles but D2S441 and vWA, which increased to 0.917 and 1.667 respectively.

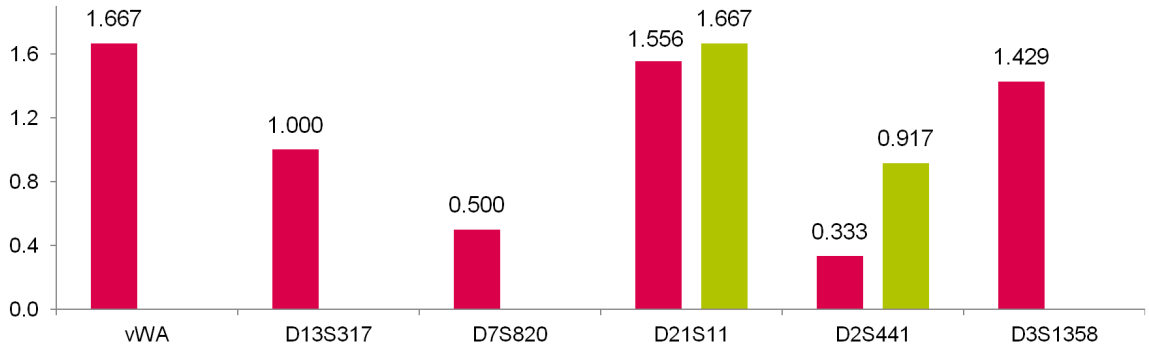


Figure 29: Fold increase in the number of observed alleles when differentiating alleles by sequence as well as by size

The **pink** bars illustrate the fold increase in the number of observed alleles for the sequences obtained from 454 sequencing. The **green** bars illustrate the fold increase in the number of observed alleles compared to those (of corresponding sizes) reported in the STRBase. The fold increases were determined by calculating the difference between the 'number of observed alleles by sequence' to that of 'alleles by size'; and dividing that difference by the 'number of observed alleles by sequence'. The number of STRBase alleles, of corresponding sizes to the alleles called in this study, was added to the 'number of observed alleles by sequence' for the construction of the green bars.

3.3 DISCUSSION

This chapter aimed to identify novel SNPSTRs and STR variants, and to investigate whether these variants exhibit specificity to the Nguni or ST populations. To achieve this aim, buccal samples from 144 individuals of self-defined Nguni or ST population groups were collected and processed using the protocol developed in Chapter 2. Notably, 12.5% of ST donors were unable to give paternal ancestry information. Following a pre-NGS quality check, these samples were pooled and sequenced using the Roche GS Junior system, whereafter the data were process and sorted as described in *3.1.5*.

The number of sequencing reads obtained in this study was more than double that reported for the average sequencing run using the Roche GS Junior system (http://my454.com/downloads/my454/applications-info/454SequencingSystem_GuidelinesforAmpliconExperimentalDesign_July2011.pdf). Of these sequences, more than 80% represented full length STR reads; and of these sequences, between 70 and 90% of reads were retained for use in allele calling. These results contrast those of Van Neste *et al.* (2012) who reported that only 48% of wells were successfully sequenced using 454 sequencing, of which only 25% represented full length STR reads. The results also contrasted those of Van Neste *et al.* (2012) with regards to the necessity of homopolymer compression. With the exception of the thymine 8-10mer in the flanking region of D7S820 which was sequenced as a 5-7mer, it was found that homopolymer compression of sequences had little effect on the allele calls and frequencies obtained. These results were in concordance with those reported by Mikkelson *et al.* (2014) who reported 95% accuracy for homopolymers of up to 6 bases. In subsequent analyses therefore, a script which compresses homopolymers of >5 base pairs in sequence-reads and reference databases should be sufficient for removing the effects of erroneously called nucleotides.

While the sequencing reaction was successful, biases were observed with regards to locus representation, with vWA in particular being represented less frequently than the other loci. As discussed in *3.2.3.1*, however, this imbalance did not seem to be related to allele length or GC content. These results contrast those of Gelardi *et al.* (2014) who found that sequencing using the Roche GS Junior system tended to favour the shorter alleles within a system; and those of Fordyce *et al.* (2015) who found that D7S820 was consistently underrepresented. Fordyce *et al.* (2015) made use of the Ion Personal Genome Machine (PGM), which is susceptible to the same errors as 454 sequencing technologies associated with homopolymer

sequencing, and it was suggested, therefore, that the poor performance of D7S820 was due to the thymine 8-10mer (Fordyce *et al.*, 2015). It is hypothesised that the M13 sequence in this study had a 'buffering' effect on the homopolymeric region, enabling D7S820 to be sequenced at a similar rate to other loci (if with an incorrectly called polymeric stretch).

Despite some imbalance in the proportion of reads per STR, there was no evidence of allele dropout for the larger alleles, as was observed on other SGS platforms, for example the Illumina MiSeq which resulted in the allele dropout of D21S11 alleles with >32 repeat units (Bornman *et al.*, 2012). Fordyce *et al.* (2015) made use of a 10-plex designed by Thermo Fisher for analysis using the Ion PGM, which exhibited an amplicon size range of 103-205 bp. The highly polymorphic complex locus D21S11 and the SNP-informative locus D13S317 were excluded from this panel; presumably due to concerns with regards to sequencing bias due to their longer and shorter lengths respectively, as well as the complexity exhibited by D21S11 (Fordyce *et al.*, 2015). The results obtained in this study, however, suggest that the sequence biases are less predictable than previously thought; with the smallest (D2S441) and largest (D21S11) loci being represented nearly equally in the Nguni population and with the proportion of the larger locus exceeding that of the smaller in the ST population.

Imbalance in sequencing data was also observed with regards to population representation. Because samples were labelled per population group as opposed to per individual (as was the case in the NGS papers discussed in *2.3*), these imbalances had the effect of casting doubt on any conclusions that may have otherwise been drawn with regards to the predictive power of allele frequencies, both with regards to comparing the Nguni and ST population groups and with regards to comparing these data with frequencies from other population groups. This lack of certainty with regards to allele frequencies was further demonstrated by the allele calls for Subject 15, made in *3.1.2*, which were not fully represented in the 454 sequencing data. F_{ST} values could, therefore, not be calculated and specific results could not be compared to those of, for example, Lane *et al.* (2002). Such results would, furthermore, be questionable due to the incomplete population data obtained for 12.5% of the ST population who were unable to give paternal information.

Despite these limitations however, a comparison of allele calls and frequencies for the two population groups was performed, which revealed a high similarity between population groups, with common alleles being shared and population-specific alleles exhibiting low

frequencies. The frequencies obtained for the SNPs rs7789995 and rs7786079 (within D7S820) and rs9546005 were also calculated and compared to those available for African populations. While allele frequencies for rs7789995 and rs7786079 were similar to those of the reference frequencies, the allele frequency for rs9546005 differed from that reported for the reference frequency. This discrepancy may be accounted for by the fact that the only African reference frequency available was for Yorubans in Nigeria, and may suggest that there are SNPs which are predictive of population group within continental groups.

The labelling by population group as opposed to by individual furthermore made it impossible to differentiate stutter sequences from genuine alleles, or to detect allele dropout and triallelic patterns. While this did not affect the calling of SNPs or indels within or associated with STRs, it did create uncertainty with regards to the validity of the putative novel variants observed for the compound and complex STRs; D21S11, vWA and D3S1358. The variants observed for these loci were however consistent with the type of variants reported in the STRbase (http://www.cstl.nist.gov/strbase) and literature. Gelardi *et al.* (2014), for example reported 10 novel variants for D21S11 which had the sequences $[TCTA]_x [TCTG]_y \ldots [TCTA]_z$ with x,y and z having varying copy numbers. This was the case for all variants reported for D21S11 in this study. Further considering the novel variant alleles reported for D21S11 in this study, it is noteworthy that two of the alleles were reported as novel variants in previous papers, namely allele 29'' (Gelardi *et al.*, 2014) and allele 31.2' (Rockenbauer *et al.*, 2014). This was also the case for T to C SNPs (described as 'STR variations' as opposed to SNPs due to their prevalence) within repeat units of D3S1358, reported by Divne *et al.* (2010), and specifically for allele 14' of this locus which was reported by Gelardi *et al.* (2014). One completely novel allele for D3S1358 was observed which contained two A to G SNPs. It, however, had a very low frequency and should be further investigated to confirm its validity. The vWA variant alleles, while unreported in literature, correspond precisely to the kind of variants observed in the STRBase and listed in the reference database.

The variants observed for the simple STRs were less susceptible to erroneous allele calls caused by stutter, since the variants were characterised by single SNPs and single base-pair insertions or deletions. Divne *et al.* (2015) reported a T/A SNP in the last repeat of D13S317. This variant was unobserved in the study considered here; however a previously undescribed variant characterised by a C to T SNP in the eighth repeat unit of allele 13 was observed. This novel allele was observed only in the Nguni population group and may, therefore, be a

population informative marker. Further studies, labelling samples by individual to ensure accurate frequency calculations, will enable this hypothesis to be tested.

The variants observed for the remaining two loci, namely D7S820 and D2S441 were characterised by both indels and SNPs. D7S820 contained a T insertion in the flanking regions of alleles 10 and 11. To the author's knowledge, this variant is previously unreported and may account for some of the off-ladder alleles reported in the STRBase (http://www.cstl.nist.gov/strbase) for D7S820. Similarly, the single base pair deletions observed for D2S441 may account for the off-ladder (x.3) alleles reported in the STRBase (http://www.cstl.nist.gov/strbase). The observed SNPs within D2S441 were previously undescribed, and their characterisation in this study supports the ICEMS data presented by Oberacher *et al.* (2008) and SGS data presented by Scheible *et* al (2014) for this locus.

Altogether, the results presented in this chapter illustrate the usefulness of SGS for the typing of STRs. The sequencing data were of high quality and quantity; and the effects of homopolymers on the accuracy of base calling of STRs, was shown to be less than previously reported. Sequencing biases, however, were observed for both population- and loci-specific data. The source of these biases was uncertain and seemed not to favour smaller or less complex alleles as predicted in literature. The effect of these biases, together with the experimental design of labelling samples by population group as opposed to by individual, was that limited conclusions could be drawn from allele frequencies. The data obtained did, however, suggest that the Nguni and ST populations are highly similar. Almost all alleles were observed in both population groups at similar frequencies. One exception was the variant allele observed for D13S317 which was observed only in the Nguni population and, with further investigation, may prove to be predictive for Nguni ancestry. The sequencing results furthermore revealed a number of putative novel variants for the other loci investigated, some of which overlapped with recent publications; and illustrated the increase in number of observable alleles when using sequence- as well as size-based data.

CONCLUSION

SGS offers a promising alternative to conventional CE-FD-based STR profiling. Not only does the method provide sequence data which enables profiles to possess higher PD values, but it also creates the potential for many loci of overlapping sizes to be typed simultaneously. While the former characteristic is particularly important for mixture resolution and familial testing; the latter is useful for enabling the multiplexing of mini- and reduced-sized STRs useful for the generation of profiles from LCN and degraded DNA. The data generated by SGS are, furthermore, both compatible and comparable with existing databases, vital for the application of the technology to forensics. SGS, furthermore, provides the potential for various genetic markers, including autosomal- and Y-STRs, as well as mitochondrial DNA, and SNPs to be typed together.

While a number of SGS technologies are available, at the time that this project was commenced, 454 sequencing was the only platform capable of sequencing full length STRs for the core loci. Since this time, however, the Roche GS Junior has been discontinued and replaced with the Ion PGM, and the Illumina Miseq platform has increased its read length to 2 x 300 bp. The number of papers investigating SGS for use as an alternative platform for STR profiling has more than doubled; and, from there being only a single paper that utilised a multiplex (unoptimised) PCR for STR amplification prior to sequencing, two commercial prototype kits (by Thermo Fisher and Promega repectively) for STR profiling by SGS have now been released.

A notable difference between the experimental design developed in this study and that employed in the aforementioned publications, is that this study explored the use of a two-step PCR protocol with population-specific MIDs; as opposed to single-step PCR followed by adaptor ligation with individual-specific MIDs. The purification strategy of the "crush and soak" method employed in this study, furthermore, contrasted that of the Agencourt AMPure bead purification, and the input DNA utilised in this study was considerably higher than that used in literature. These differences had the effect of making the protocol developed in the study unsuitable for use in routine forensic analysis. The 2-step PCR and "crush-and-soak" purification were laborious and vulnerable to contamination. Furthermore, the input DNA quantity was far in excess of that usually obtained from forensic samples. The labelling of samples by population group as opposed to by individual, furthermore, had the effect of preventing the discrimination of stutter from true alleles, identifying allele drop-out and

trialleleic patterns; thereby reducing the information that could be gleaned from allele frequencies.

Despite the drawbacks of the protocol, it was successfully employed to amplify the STRs, vWA, D2S441, D3S1358, D7S820, D13S317 and D21S11, using M13- and fusion-tailed primers from buccal swabs collected from 144 individuals of Nguni and ST ancestry. The pooled libraries generated high quality sequencing data, the processing of which demonstrated discrepancies in the sequencing of >8 bp homopolymers (which were reported as 5-8 bp homopolymers). These results suggested that future studies should employ homopolymer compression for homopolymers >4 bp. Furthermore, while the sequencing data was of high quality, biases were observed between population groups and loci. These biases were however, not related to sequence length or complexity; and did not prevent the identification of a number of previously unreported STR variants.

The results obtained in this study demonstrate the power of SGS for increasing the number of discernible alleles for STR profiling. While the protocol is, in its current form, unsuitable for use in routine forensic analysis, revisions with regards to the PCR set-up and purification strategy may change this. For example, the success of the 1° PCR multiplex suggests that the protocol may be modified to a single-step PCR which utilises individual-specific STR-directed fusion primers. Such an experimental design would remove the need for the laborious 2° PCR used in this experimental design, and for the costly adaptor ligation used in the various SGS publications. The problem of the high input DNA and prevalence of primer dimers would, however, need to be addressed. Furthermore, because of the massive impact that an incorrectly called allele could have on an individual's life; thorough validation of any such SGS protocol for use in DNA profiling would need to be performed before it could be considered for routine forensic application. Data-processing would also need to be streamlined and nomenclature agreed upon by the forensic community. Furthermore, research to define sub-allele frequencies for various population groups would need to be undertaken to enable accurate PD calculations.

REFERENCE LIST

Journal Articals

Abrahams, Z., D'Amato, M. E., Davison, S. & Benjeddou, M. (2011) Allele frequencies of six non-CODIS miniSTR loci (D1S1627, D3S4529, D5S2500, D6S1017, D8S1115 and D9S2157) in three South African populations, *Forensic Science International: Genetics*, 5(4):354–355.

Agrafioti, I. & Stumpf, M. P. H. (2007) SNPSTR: a database of compound microsatellite-SNP markers, *Nucleic Acids Research*, 35(Database issue):71–75.

Allor, C., Einum, D.D. & Scarpetta, M. (2005) Identification and characterization of variant alleles at CODIS STR loci, *Journal of Forensic Science,* 50:1128-1133.

Ardlie, K.G., Kruglyak, L. & Seielstad, M. (2002) Patterns of linkage disequilibrium in the human genome, *Nature Reviews Genetics,* 3:299–309.

Bacher, J., Schumm, J.W. (1998) Development of highly polymorphic pentanucleotide tandem repeat loci with low stutter, *Profiles in DNA*, 2(2):3–6.

Ballantyne, K. N. *et al*. (2010) Mutability of Y-chromosomal microsatellites: rates, characteristics, molecular bases, and forensic implications, *American Journal of Human Genetics,* 87:341–353.

Bamshad, M., Wooding, S., Salisbury, B. A. & Stephens, J. C. (2004) Deconstructing the relationship between genetics and race, *Nature Reviews:Genetics*, 5(8):598–609.

Biesecker, L.G., Bailey-Wilson, J.E., Ballantyne, J., Baum, H., Bieber, F.R., Brenner, C., Budowle, B., Butler, J.M., Carmody, G., Conneally, P.M., Duceman, B., Eisenberg, A., Forman, L., Kidd, K.K., LeClair, B., Niezgoda, S., Parsons, T., Pugh, E., Shaler, R., Sherry, S.T., Sozer, A. & Walsh, A. (2005) DNA identifications after the 9/11 World Trade Center attack, *Science*, 310:1122–1123.

Bornman, D. M., Hester, M. E., Schuetter, J. M., Kasoji, M. D., Minard-smith, A., Barden, C.A. Nelson S.C., Goldbold, G.D., Baker, C.H., Yang, B., Walther, J.E., Tornes, I.E., Yan, P.S. Rodriguez, B., Bundschuh, R., Dickens, M.L., Young, B.A. & Faith, A. (2012) Short-read, high-throughput sequencing technology for STR genotyping, *Biotech Rapid Dispatches,* 1–6.

Brenner, C.H. & Weir, B.S. (2003) Issues and strategies in the DNA identification of World Trade Center victims, *Theoretical Population Biology,* 63 (3):173–178.

Brookes, A.J. (1999) The essence of SNPs, *Gene*, 234:177–186.

Budowle B, Masibay A, Anderson SJ, Barna C, Biega L, Brenneke S, et al (2001) STR primer concordance study, *Forensic Science International,*124:47-54.

Budowle, B., Eisenberg, A.J. & van Daal, A. (2009) Validity of Low Copy Number Typing and Applications to Forensic Science, *Croatia Medical Journal*, 50:207-217.

Butler, J.M. (2003) Recent developments in Y-short tandem repeat and Y-single nucleotide polymorphism analysis, *Forensic Science Review*, 15(2):91–111.

Butler, J. M. & Hill, C. R. (2012) Biology and Genetics of New Autosomal STR Loci Useful for Forensic DNA Analysis Biology and Genetics of New Autosomal STR Loci Useful for Forensic DNA Analysis, *Forensic Science Review*, 24(1):15–26.

Butler, J. M., Coble, M. D. & Vallone, P. M. (2007) STRs vs. SNPs: thoughts on the future of forensic DNA testing, *Forensic Science, Medicine, and Pathology*, 3(3):200–205.

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

Carneiro, M.O., Russ, C., Ross, M.G., Gabriel, S.b., Nusbaum, C. & DePristo, M.A. (2012) Pacific Biosciences Sequencing Technology for Genotyping and Variation Discovery in Human Data, *BioMed Central Genomics*, 13:375.

Chakhparonian, M. & Wellinger, R.J. (2003) Telomere maintenance and DNA replication: how closely are these two connected? *Trends in Genetics,*19(8): 439–446.

Clayton, T.M., Guest, J.L., Urquhart, A.J. & Gill, P.D. (2004) A genetic basis for anomalous band patterns encountered during DNA STR profiling, *Journal of Forensic Science*,49:1207-1214.

Clayton, T.M., Hill, S.M., Denton, L.A., Watson, S.K. & Urquhart, A.J. (2004) Primer binding site mutations affecting the typing of STR loci contained within the AMPFlSTR SGM Plus kit, *Forensic Science International,*139:255-259.

Clayton, T. M., Whitaker, J. P., Sparkes, R. L. & Gill, P. (1998) Analysis and intepretation of mixed forensic stains using DNA STR profiling, *Forensic Science International,* 91:55–70.

Coble, M.D. & Butler, J.M. (2005) Characterization of new miniSTR loci to aid analysis of degraded DNA, *Journal of Forensic Science,* 50:43-53.

Collins JR, Stephens RM, Gold B, Long B, Dean M, Burt SK. (2003) An ex- haustive DNA micro-satellite map of the human genome using high performance computing, *Genomics*,82:10–19.

Committee on Identifying the Needs of the Forensic Sciences Community (2009) Strengthening Forensic Science in the United States: A path forward, *The National Academies Press,* 228091.

Daniel, R., Santos, C., Phillips, C., Fondevila, M., van Oorchot, R.A.H., Carracedo, A., Lareu, M.V. & McNevin, D. (2014) A SNaPshot of next generation sequencing for forensic SNP analysis, *Forensic Science International: Genetics,* 14:50–60.

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

Dror, I.E. & Hampikian, G. (2011) Subjectivity and bias in forensic DNA mixture interpretation, *Science and Justice,* 51:204-208.

Editorial (1992) Report concerning recommendations of the DNA Commission of the International Society for Forensic Haemogenetics relating to the use of DNA polymorphisms, *Forensic Science International,* 52:125–130.

Editorial (1989) Recommendations of the Society for Forensic Haemogenetics concerning DNA polymorphisms, *Forensic Science International,*43:109–111.

Edwards, A,, Civitello, A., Hammond, H.A. & Caskey, C.T. (1991) DNA typing and genetic mapping with trimeric and tetrameric tandem repeats, *American Journal of Human Genetics,*49:746–756.

Ellegren, H. (2004) Microsatellites: simple sequences with complex evolution, *Nature Reviews Genetics,* 5:435–445.

Fang, R., Pakstis, A.J., Hyland, F., Wang, D., Shewale, J., Kidd, J.R., Kidd, K.K. & Furtado, M.R. (2009) Multiplexed SNP detection panels for human identification. *Forensic Science International: Genetics*. 2:538–539.

Fordyce, S. L., Ávila-arcos, M. C., Rockenbauer, E., Børsting, C. & Frank-hansen, R. (2011) High-throughput sequencing of core STR loci for forensic genetic investigations using the Roche Genome Sequencer FLX platform, *Biotechniques*, 51:123-133.

Fordyce, S. L., Smidt, H., Børsting, C., Lagace, R. E., Chang, C., Rajagopalan, N. & Morling, N. (2015) Second-generation sequencing of forensic STRs using the Ion Torrent TM HID STR 10-plex and the Ion PGM TM, *Forensic Science International* :Genetics, 14:132–140.

Frudakis, T. Venkateswarlu, K., Thomas, M.J., Gaskin, Z., Ginjupalli, S., Gunturi, S., Ponnuswamy, V., Natarajan, S. & Nachimuthu, P.K. (2003) A classifier for the SNP-based inference of ancestry, *Journal of Forensic Science,* 48:1–12.

Fullwood, M.J., Wei, C.L., Liu, E.T. & Ruan, Y. (2009) Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses, *Genome Research*, 19:521–532.

Gelardi, C., Rockenbauer, E. Dalsgaard, S. Borsting, C. & Morling, N. (2014) Second generation sequencing of three STRs D3S1358, 2 D12S391 and D21S11 in Danes and a new nomenclature for 3 sequenced STR alleles, *Forensic Science International: Genetics,* 12:38-41.

Gershaw, C. J., Schweighardt, A. J., Rourke, L. C. & Wallace, M. M. (2011) Forensic utilization of familial searches in DNA databases, *Forensic Science International: Genetics*, 5:16–20.

Gill, P., Fereday, L., Morling, N. &Schneider, P.M. (2006) The evolution of DNA databases-Recommendations for new European STR loci, *Forensic Science International,* 156:242-244.

Gill, P. (2002) Role of short tandem repeat DNA in forensic casework in the UK – past, present, and future perspectives, *BioTechniques*, 32:366–372.

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., Whitaker, J., Flaxman, C., Brown, N. & Buckleton, J. (2000) An investigation of the rigor of interpretation rules for STRs derived from less than 100 pg of DNA. *Forensic Science International,*112:17–40.

Grubwieser, P., Mühlmann, R., Niederstätter, H., Pavlic, M. & Parson, W. (2005) Unusual variant alleles in a commonly used short tandem repeat loci, *International Journal of Legal Medicine,*119:164-166.

Grubwieser, P., Mühlmann, R., Berger, B., Niederstätter, H., Pavlic, M. & Parson, W. (2006). A new “miniSTR-multiplex” displaying reduced amplicon lengths for the analysis of degraded DNA, *International Journal of Legal Medicine*, 120(2):115–120.

Gymrek, M., Golan, D., Rosset, S. & Erlich, Y. (2012) lobSTR: A short tandem repeat profiler for personal genomes, *Genome Research*, 22(6):1154–1162.

Hardy, G.H. (1908) Mendelian proportion in a mixed population, *Science,* 28:49-50.

Hares, D.R. (2012) Expanding the CODIS core loci in the United States, *Forensic Science International: Genetics,* 6:e52.

Heathfield, L. (2014) Policy required for entry of DNA profiles onto the National Forensic DNA Database of South Africa, *South African Journal of Sciences,* 110(7):8–10.

Heinrich, M., Felske-Zech, H., Brinkmann, B. & Hohoff, C. (2005) Characterization of variant alleles in the STR systems D2S1338, D3S1358 and D19S433, *International Journal of Legal Medicine,* 119:310-313.

Helmuth, R. Fildes, N., Blake, E., Luce, M.C., Chimera, J., Madej, R., Gorodezky, C., Stoneking, M., Schmill, N. & Klitz, W. (1990) HLA-DQ alpha allele and genotype frequencies in various human populations, determined by using enzymatic amplification and oligonucleotide probes, *American Journal of Human Genetics*, 47:515-523.

Hill, C. R., Kline, M. C., Coble, M. D., & Butler, J. M. (2008) Characterization of 26 miniSTR loci for improved analysis of degraded DNA samples. *Journal of Forensic Sciences*, 53(1):73–80.

Hughes-Stamm, S. R., Ashton, K. J., & van Daal, A. (2011) Assessment of DNA degradation and the genotyping success of highly degraded samples, *International Journal of Legal Medicine*, 125(3), 341–348.

Huston, K.A. (1998) Statistical Analysis of STR Data, *Profiles in DNA, Technical Tips*, *GenePrint*™, 14-15.

International HapMap Consortium (2005) A haplotype map of the human genome, *Nature*, 437:1229–1320.

International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome, *Nature*, 431:931–945.

Jeffreys, A.J., Wilson, V. & Thein, S.L. (1985) Hypervariable 'minisatellite' regions in human DNA, *Nature,* 314:67-79.

Jeffreys, A.J., Wilson, V. & Thein, S.L.(1985) Individual-Specific 'fingerprints' of human DNA, *Nature,* 316:76-79.

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

Jobling, M.A. & Tyler-Smith, C. (2003) The human Y chromosome: an evolutionary marker comes of age, *Nature Reviews Genetics,*4:598-612.

Jorde, L. B., Bamshad, M. J., Watkins, W. S., Zenger, R., Fraley, A.E., Krakowiak, P.A., Carpenter, K.D., Soodyall, H., Jenkins, T. &Rogers, A.R. (1995) Origins and affinities of modern humans: a comparison of mitochondrial and nuclear genetic data, *American Journal of Human Genetics*, 57(3):523–538.

Jorde, L. B., Watkins, W. S., Bamshad, M. J., Dixon, M. E., Ricker, C. E., Seielstad, M. T. & Batzer, M. A. (2000) The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data, *American Journal of Human Genetics*, 66(3):979–988.

Kayser, M., Kittler, R., Erler, A., Hedman, M., Lee, A.C., Mohyuddin, A., Mehdi, S.Q., Rosser, Z., Stoneking, M., Jobling, M.A., Sajantila, A. & Tyler-Smith, C. (2004) A comprehensive survey of human Y- chromosomal microsatellites, *American Journal of Human Genetics,* 74:1183–1197.

Kayser, M., Caglia, A., Corach, D., Fretwell, N., Gehrig, C., Graziosi, G., Heidorn, F., Herrmann, S., Herzog, B., Hidding, M., Honda, K., Jobling, M., Krawczak, M., Leim, K., Meuser, S., Meyser, E., Oesterreich, W., Pandya, A., Parson, W., Penacino, G., Perez-Lezaun, A., Piccinini, A., Prinz, M., Schmitt, C. & Roewer, L. (1997) Evaluation of Y-chromosomal STRs: a multicenter study, *International Journal of Legal Medicine,* 110:125–133, 141–149.

Kayser, M. & de Knijff, P. (2011) Improving human forensics through advances in genetics, genomics and molecular biology, *Nature Reviews: Genetics*, 12(3):179–192.

Kidd, K.K., Pakstis, A.J., Speed, W.C., Grigorenko, E.L., Kajuna, S.L., Karoma, N.J., Kungulilo, S., Kim, J.J., Lu, R.B., Odunsi, A., Okonofua, F., Parnas, J., Schulz, L.O., Zhukova, O.V. & Kidd, J.R. (2006) Developing a SNP panel for forensic identification of individuals, *Forensic Science International,*164:20–32.

Kimpton, C.P., Fisher, D., Watson, S., Adams, M., Urquhart, A., Lygo, J. & Gill, P. (1994) Evaluation of an automated DNA profiling system employing multiplex amplification of four tetrameric STR loci, *International Journal of Legal Medicine*,106:302– 311.

Kimpton, C.P., Oldroyd, N.J., Watson, S.K., Frazier, R.R.E., Johnson, P.E., Millican, E.S., Urquhart, A., Sparkes, B.L. & Gill, P.(1996) Validation of highly discriminating multiplex short tandem repeat amplification systems for individual identification, *Electrophoresis*, 17(8):1283–1293.

Lane, A.B. (2013). STR null alleles complicate parentage testing in South Africa, *Molecular Genetics,* 103(12):1004–1008.

Lango, A. H. *et al.* (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height, *Nature,* 467:832–838.

Lao, O., de Gruijter, J. M., van Duijn, K., Navarro, A. & Kayser, M. (2007) Signatures of positive selection in genes associated with human skin pigmentation as revealed from analyses of single nucleotide polymorphisms, *Annals of Human Genetics*, 71:354–369.

Leat, N., Ehrenreich, L., Benjeddou, M. & Davison, S. (2006) Properties of novel and widely studied Y-STR loci in three South African populations, *Forensic Science International,* 157:210-217.

Leat, N., Ehrenreich, L., Benjeddou, M. & Davison, S. (2004) Developments in the use of Y-chromosome markers in forensic genetics, *African Journal of Biotechnology*, 3: 637-642.

Liu, F., van der Lijn, F., Schurmann, C., Zhu, G., Chakravarty, M. M., Hysi, P. G., … Kayser, M. (2012) A genome-wide association study identifies five loci influencing facial morphology in Europeans, *PLoS Genetics*, 8(9):e1002932.

Louis, E.J. & Vershinin, A.V. (2005) Chromosome ends: different sequences may provide conserved functions, *BioEssays*, 27:685–697.

Lucassen, A., Ehlers, K., Grobler, P. J., & Shezi, A. L. (2014) Allele frequency data of 15 autosomal STR loci in four major population groups of South Africa, *International Journal of Legal Medicine,* 128(2):275–276.

McElhoe, J.A., Holland, M. M., Makova, K. D., Su, M. S.-W., Paul, I. M., Baker, C. H., … Young, B. (2014) Development and assessment of an optimized next-generation DNA sequencing approach for the mtgenome using the Illumina MiSeq, *Forensic Science International: Genetics*, 13, 20–29.

Meintjies-van der Walt, L. (2010) DNA in the courtroom: Principles and practice, South African Journal of Science, *Juta*.

Mizuno, N., Sekiguchi, K., Sato, H. & Kasai, K. (2003) Variant alleles on the penta E locus in the PowerPlex 16 kit, *Journal of Forensic Science,* 48:358-361.

Moretti, T.R., Baumstark, A.L., Defenbaugh, D.A., Keys, K.M., Brown, A.L. & Budowle, B. (2001) Validation of STR typing by Capillary Electrophoresis, *Journal of Forensic Science,* 46:647-676.

Morling, N. (2004). Forensic genetics, *The Lancet,* 364:10–11.

Nelson, M.S., Levedakou, E.N., Matthews, J.R., Early, B.E., Freeman. D.A., Kuhn, C.A., Sprecher, C. J., Amin, A.S., McElfresh, K.C. & 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 Science*, 47:345–349.

Oberacher, H., Pitterl, F., Huber, G., Niederstätter, H. & Parson, W. (2008) Increased forensic efficiency of DNA fingerprints through simultaneous resolution of length and nucleotide variability by high-performance mass spectrometry, *Human Mutation,* 29:427–432.

Pakstis, A. J. Speed, W.C., Fang, R., Furtado, M.R. & Kidd, K.K. (2010) SNPs for a universal individual identification panel, *Human Genetics,* 127:315–324.

Parra, E. J., Marcini, A., Akey, J., Martinson, J., Batzer, M.A., Cooper, R., Forrester, T., Allison, D.B., Deka,R., Ferrell, R.E. & Shriver, M.D. (1998) Estimating African–American admixture proportions by use of population-specific alleles, *American Journal of Human Genetics,* 63:1839–1851.

Parson, W. C., Strobl, G., Huber, B., Zimmermann, S.M., Gomes, L., Souto, Fendt, L., Delport, R., Langit, R., Wootton, S., Lagace, R. & Irwin, J. (2013) Evaluation of next generation mtGenome sequencing using the Ion Torrent Personal Genome Machine (PGM), *Forensic Science International: Genetics*, 7(5):543–549.

Phillips, C., Gelabert-Besada, M., Fernandez-Formoso, L., García-Magariños, M., Santos, C., Fondevila, M., Ballard, D., Syndercombe, C.D., Carracedo, A. & Lareu, M.V. (2014) “New turns from old STaRs”: Enhancing the capabilities of forensic short tandem repeat analysis, *Electrophoresis*, 35(21-22):3173–3187.

Pitterl, F., Niederstätter, H., Huber, G., Zimmermann, B., Oberacher, H. & Parson, W. (2008) The next generation of DNA profiling--STR typing by multiplexed PCR--ion-pair RP LC-ESI time-of-flight MS, *Electrophoresis*, 29(23):4739–4750.

Pitterl, F., Schmidt, K., Huber, G., Zimmermann, B., Delport, R., Amory, S., Ludes, B., Oberacher, H. & Parson, W. (2010) Increasing the discrimination power of forensic STR testing by employing high-performance mass spectrometry, as illustrated in indigenous South African and Central Asian populations, *International Journal of Legal Medicine*, 124(6):551–558.

Planz, J. V., Budowle, B., Hall, T., Eisenberg, A. J., Sannes-Lowery, K. A. & Hofstadler, S. A. (2009) Enhancing resolution and statistical power by utilizing mass spectrometry for detection of SNPs within the short tandem repeats, *Forensic Science International: Genetics,* Supplement Series, 2(1):529–531.

Prinz, M. (2003) Advantages and disadvantages of Y-short tandem repeat testing in forensic casework, *Forensic Science Review*,15:189–196.

Ramakrishnan, U. & Mountain, J. L. (2004) Precision and accuracy of divergence time estimates from STR and SNPSTR variation, *Molecular Biology and Evolution*, 21(10):1960–1971.

Roewer, L. Y. (2009) Chromosome STR typing in crime casework. *Forensic Science, Medicine and Pathology,* 5(2):77–84.

Rolf, B., Wiegand, P. & Brinkmann, B. (2002) Somatic mutations at STR loci – a reason for three-allele pattern and mosaicism. *Forensic Science International,* 26:200-202.

Sanchez, J. J., Børsting, C., Balogh, K., Berger, B., Bogus, M., Butler, J. M., … Morling, N. (2008) Forensic typing of autosomal SNPs with a 29 SNP-multiplex--results of a collaborative EDNAP exercise, *Forensic Science International: Genetics*, 2(3):176–183.

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.

Sanger, F., Nicklen, S. & Coulson A.R. (1997) DNA sequencing with chain- terminating inhibitors, *Proceedings of the National Academy of Sciences,* 74:5463–5467.

Scheible, M., Loreille, O., Just, R. & Irwin, J. (2011) Short tandem repeat sequencing on the 454 platform, *Forensic Science International: Genetics* Supplement Series, 3(1):e357–e358.

Scheible, M., Loreille, O., Just, R. & Irwin, J. (2014) Short tandem repeat typing on the 454 platform: strategies and considerations for targeted sequencing of common forensic markers, *Forensic Science International: Genetics* , 12:107–119.

Schlebusch, C. M., Soodyall, H. & Jakobsson, M. (2012) Genetic variation of 15 autosomal STR loci in various populations from southern Africa, *Forensic Science International: Genetics*, 6(1):e20–e21.

Schneider, P.M., Bender, K., Mayr, W.R., Parson, W., Hoste, B., Decorte, R., ... Greenhalgh, M. (2014) STR analysis of artificially degraded DNA-results of a collaborative European exercise, *Forensic Science International,*139:123–134.

Schoske, R., Vallone, P. M., Ruitberg, C. M. & Butler, J. M. (2003) Multiplex PCR design strategy used for the simultaneous amplification of 10 Y chromosome short tandem repeat (STR) loci, *Analytical and Bioanalytical Chemistry*, 375(3):333–343.

Shewale, J. G., Sikka, S. C., Schneida, E. & Sinha, S. K. (2003) DNA profiling of azoospermic semen samples from vasectomized males by using Y-PLEX 6 amplification kit, *Journal of Forensic Science*, 48:127–129.

Sullivan, K. M., Mannucci, A., Kimpton, C. P. & Gill, P. (1993) A rapid and quantitative DNA sex test: fluorescence-based PCR analysis of X-Y homologous gene amelogenin, *Biotechniques,* 15:636–641.

Underhill, P. A. & Kivisild, T. (2007) Use of Y chromosome and mitochondrial DNA population structure in tracing human migrations, *Annual Review of Genomics and Human Genetics,* 41:539–564.

Van Neste, C., Van Nieuwerburgh, F., Van Hoofstat, D. & Deforce, D. (2012) Genetics Forensic STR analysis using massive parallel sequencing, *Forensic Science International*: Genetics, 6(6):810–818.

Van Neste, C., Gansemans, Y., De Coninck, D., Van Hoofstat, D., Van Criekinge, W., Deforce, D. & Van Nieuwerburgh, F. (2014) Forensic massively parallel sequencing data analysis tool: Implementation of MyFLq as a standalone web- and Illumina BaseSpace(®)-application, *Forensic Science International: Genetics*, 15:2-7.

Van Neste, C., Vandewoestyne, M., Van Criekinge, W., Deforce, D., & Van Nieuwerburgh, F. (2013) My-Forensic-Loci-queries (MyFLq) framework for analysis of forensic STR data generated by massive parallel sequencing, *Forensic Science International: Genetics*, 9:1-8.

Walsh, P.S., Fildes, N.J. & Reynolds, R. (1996) Sequence analysis and characterization of stutter products at the tetranucleotide repeat locus vWA, *Nucleic Acid Resolution,* 24:2807–2812.

Walsh, S., Liu, F., Wollstein, A., Kovatsi, L., Ralf, A., Kosiniak-Kamysz, A., … Kayser, M. (2013) The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA. *Forensic Science International: Genetics*, 7(1):98–115.

Wang, L., Schneider, P.M., Rothschikld, M.A., Bai, P., Liang, W. & Zhang, L. (2013) SNP-STR polymorphism: A sensitive marker for forensic genetic applications, *Forensic Science International: Genetics* Supplementary Series:e206-e207.

Warshauer, D. H., King, J. L. & Budowle, B. (2015) STRait Razor v2.0: The improved STR Allele Identification Tool – Razor, *Forensic Science International: Genetics*, 14:182–186.

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

Westen, A.A., Matai, A.S., Laros, J.F.J., Meiland, H. C., Jasper, M., de Leeuw, W. J. F., de Knijff, P. & Sijen, T. (2009) Tri-allelic SNP markers enable analysis of mixed and degraded DNA samples, *Forensic Science International: Genetics,* 3(4):233–241.

Yang, Y., Xie, B. & Yan, J. (2014) Application of Next-generation Sequencing Technology in Forensic Science, *Genomics*, *Proteomics & Bioinformatics*, 12(5):190–197.

Ye, Y., Luo, H., Liao, L., Zhang, J., Wei, W., Wang, Z. & Hou, Y. (2014) A case of SNPSTR efficiency in paternity testing with locus incompatibility, *Forensic Science International: Genetics,* 9:72-75.

Yoshida, K., Yayama, K., Hatanaka, A. & Tamaki, K. (2011) Efficacy of extended kinship analyses utilizing commercial STR kit in establishing personal identification, *Journal of Legal Medicine,* 13:12–15 .

Zeng, X., King, J., Stoljarova, M., Warshauer, D., LaRue, B., Sajantila, A., Patel, J., Storts, D. & Budowle, B. (2014) High Sensitivity Multiplex Short Tandem Repeat Loci Analyses with Massively Parallel Sequencing, *Forensic Science International: Genetics,* 16:38-47.

Zhang, J., Chiodini, R., Badr, A. & Zhang, G. (2011) The impact of next-generation sequencing on genomics. *Journal of Genetics and Genomics*, 38(3):95–109.

Books

Butler, J.M. (2010) *Fundamentals of Forensic DNA Typing*, Elsevier Academic Press, San Diego.

Butler, J.M. (2005) *Forensic DNA Typing: biology, technology, and genetics of STR markers,* Elsevier Academic Press, London.

Zascavage, R.R., Shewale, S.J. & Planz, J.V. “Deep-Sequencing Technologies and Potential Applications in Forensic DNA Testing” 312-338 in Shewale, J.G. & Liu, R.H. (2014) *Forensic DNA Analysis: Current Practices and Emerging Technologies*, CRC Press Taylor and Francis Group, LLC, U.S.A.

Ott, J. (1999) *Analysis of human genetic linkage*. 3rd edition, Johns Hopkins University Press, Baltimore.

Rudmin, N. & Indman, K. (2002). *An Introduction to Forensic DNA Analysis* 2nd Ed. CRC Florida, USA.

Tillmar, A. (2010) *Populations and Statistics in Forensic Genetic*s. Linköping University Medical Dissertations No. 1175, Faculty of Health Sciences, Linköping University, Linköping, Sweden.

Websites

Blast frequently asked questions
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?CMD=Web&PAGE_TYPE=B lastDocs&DOC_TYPE=FAQ
Accessed: March 2013

GelQuant.NET
biochemsolutions.com
Accessed: July 2013

Performing Electrophoresis
www.bio-rad.com 17/11/2014 Bi-Rad Laboratories, Inc. 2014
Accessed: Octeober 2014

STR Fact Sheets
www.cstl.nist.gov/strbase/coreSTRs.htm
Accessed: 2013-2015

HapMap
http://hapmap.ncbi.nlm.nih.giv
Accessed: 2013-2015

Primer Scepcifications
www.lifesciences.sourcebiosciences.com
Accessed: November 2014

Library Preparation Manual
http://454.com/ downloads/my454/documentation/gs-flx-plus/Rapid-Library-Preparation-Method-ManualXL PlusMay2011.pdf
Accessed: 2013-2015

Product Specifications
http://www.thermoscientific.com/\
Accessed: July 2013

FastPCR 6.3. and IDT Oligo Analyser 3.1
http://eu.idtdna.com
Accessed: 2013-2014

Conference Presentations

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

Hill, B. & Butler, J. (2012) “Characterization of Additional STR Loci Beyond the 13 CODIS Loci.” *National Institute of Standards and Technology. Bode Technology 1st Annual Advanced DNA Technology Mid-Atlantic Workshop* Charlottesville, VA..

Planz, J.V. (2004) “Forensic Statistics: From the Ground up” *15th International Symposium on Human Identification*, UNT Health Science.

A. DNA Collection

Subjects were required to fill in the questionnaire below (**Figure i**) after giving informed consent to participating in the study. The information sheet, which accompanied a verbal explanation of the study, is presented in **Figure ii.** A copy of the signed consent forms can be obtained from the author upon request.

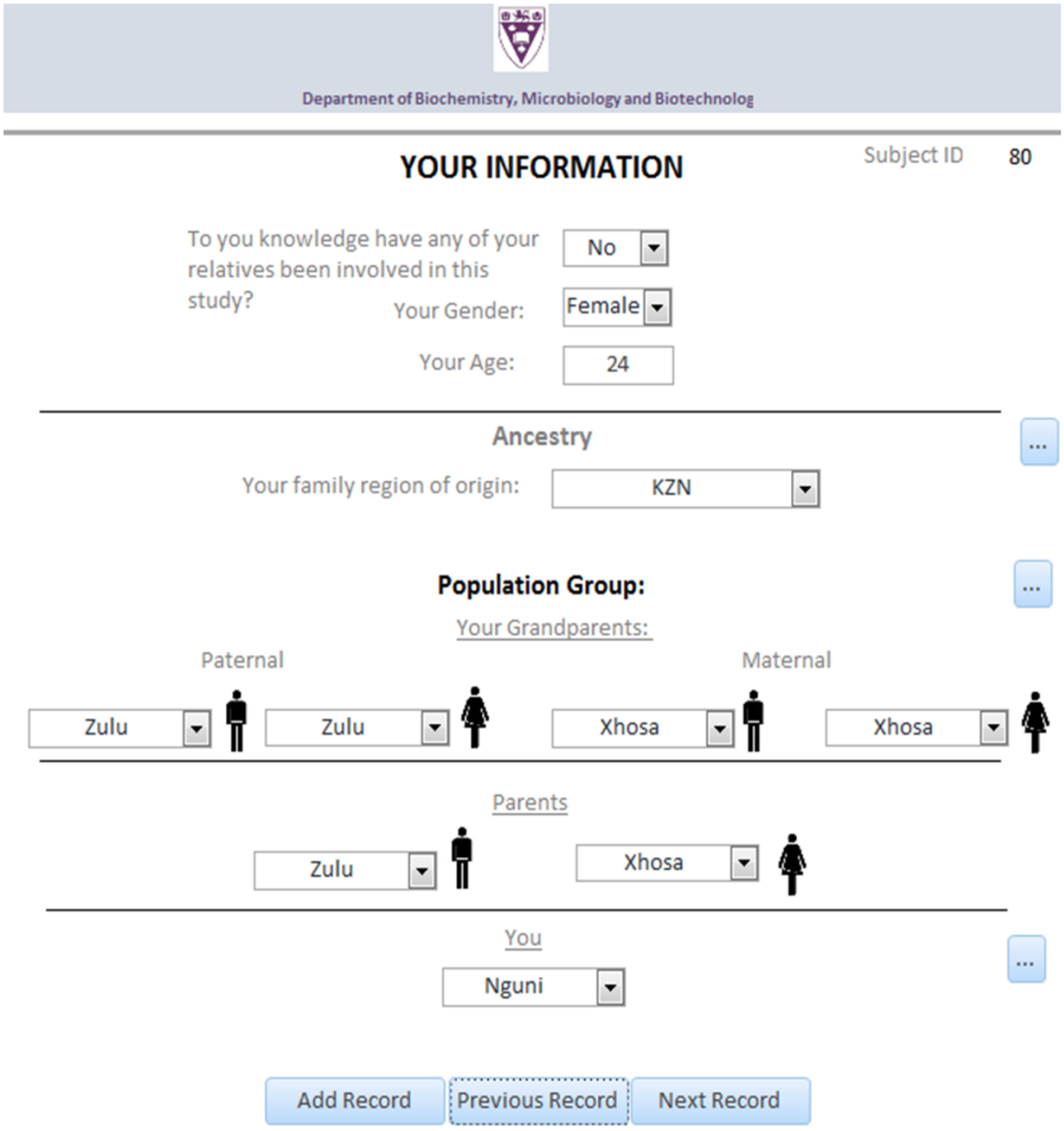


Figure i: Screenshot of the questionnaire (Microsoft Access) used for data collection

Subject ID No.

Department of Biochemistry, Microbiology and Biotechnology
Rhodes University

Sequence Variation of Autosomal Microsatellites in South African sub-Populations, Characterised using Next Generation Sequencing

Researcher: Jo-Anne Laurence (g09l1091@campus.ru.ac.za, Lab 235 of the Biological Sciences Building)
Supervisors: Dr Adrienne Edkins and Dr Brendan Wilhelmi ; Head of Department: Dr Jo Dames

Some background information:

Current methodology for DNA Fingerprinting

Step 1: DNA is extracted from the biological sample

Step 2: Regions of the DNA called *Short Tandem Repeats* (STRs) are analysed.

SHORT TANDEM REPEATS: Your DNA can be imagined as a book, the book is made up of words, some of which carry meaning (genes) and others which do not add much to the story-line (non-coding regions). In some places, these 'unnecessary' sections contain nonsensical letters (nucleotides) which are repeated (STR). The number of these repeats varies between individuals.
A person can be identified by their unique number and combination of repeats.

	 BRIAN		 LETHI	
	STR sequence	No. of Repeats	STR Sequence	No. of Repeats
STR 1	PIG.PIG.PIG.PIG.PIG	5	PIG.PIG.PIG	3
STR 2	CAR.CAR.CAR	3	CAR.CAR.CAR.CAR.CAR.CAR	6
STR 3	ROOM.ROOM	2	ROOM.ROOM.ROOM	3
FINAL BARCODE		5,3,2		3,6,3

This Study

1. Development of a novel method for DNA profiling

In order to identify a suspect with a higher degree of certainty, this study will investigate the spelling of the "nonsensical' regions so as to determine if inter-individual variation exists. If this variation does exist, the reading (sequencing) of the STRs will increase the discrimination power of the genetic profile.

	BRIAN	LETHI	JULIA
STR 1	PIG.PIG.PIG.PIG.PIG	PIG.PIG.PIG	PIG.PIG.PUG.PIG.PIG.PIG

2. Analysis of whether variations in DNA sequence are related to ancestry

This may enable the prediction of the biogeographic origin (ancestry) of a given suspect.

Note:

- *All samples will be labelled numerically and the subject's name will not be recorded.*
- *All DNA samples will be destroyed upon completion of this study*
- *If a subject desires to withdraw from the study, he/she may do so within 3 days of sample donation.*

Your anonymity is guaranteed!

Figure ii: Copy of the information sheet given to DNA donors

B. 454 Sequencing

Table i: Reference sequences used for identifying alleles that matched those reported in the STRbase (www.cstl.nist.gov/strbase)

Loci	Allele	Sequence (Forward primer | Sequence | Reverse Primer)
vWA	10	CCCTAGTGGATGATAAGAATAATCAGTATGTGACTTGGATTGATCTATCTGTCTATCTGTCTGTCTGTCTGTCTATCTATCTATCCATCCATCCTATGTATTTATCATCTGTC
	11(13')	CCCTAGTGGATGATAAGAATAATCAGTATGTGACTTGGATTGATCTATCTGTCTGTCTATCTATCTATCTATCTATCTATCTATCCATCCATCCTATGTATTTATCATCTGTC
	12	CCCTAGTGGATGATAAGAATAATCAGTATGTGACTTGGATTGATCTATCTGTCTGTCTGTCTGTCTATCTATCTATCTATCTATCTATCTATCTATCCATCCATCCTATGTATTTATCATCTGTC
	13	CCCTAGTGGATGATAAGAATAATCAGTATGTGACTTGGATTGATCTATCTATCTGTCTGTCTGTCTGTCTATCTATCTATCTATCCATCTATCTATCTATCCATCCATCCTATGTATTTATCATCTGTC
	13(15)	CCCTAGTGGATGATAAGAATAATCAGTATGTGACTTGGATTGATCTATCTGTCTGTCTGTCTGTCTATCTATCTATCTATCTATCTATCTATCTATCTATCCATCTATCCATCCATCCTATGTATTTATCATCTGTC
	13(15'')	CCCTAGTGGATGATAAGAATAATCAGTATGTGACTTGGATTGATCTATCTGTCTGTCTGTCTGTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCCATCCATCCTATGTATTTATCATCTGTC
	14(16'')	CCCTAGTGGATGATAAGAATAATCAGTATGTGACTTGGATTGATCTATCTGTCTGTCTGTCTGTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCCATCCATCCTATGTATTTATCATCTGTC
	14'(16''')	CCCTAGTGGATGATAAGAATAATCAGTATGTGACTTGGATTGATCTATCTGTCTATCTGTCTGTCTGTCTGTCTATCTATCTATCCATCTATCTATCTATCCATCCATCCTATGTATTTATCATCTGTC
	15(17)	CCCTAGTGGATGATAAGAATAATCAGTATGTGACTTGGATTGATCTATCTGTCTGTCTGTCTGTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCCATCTATCCATCCATCCTATGTATTTATCATCTGTC
	15(17')	CCCTAGTGGATGATAAGAATAATCAGTATGTGACTTGGATTGATCTATCTGTCTGTCTGTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCCATCTATCCATCCATCCTATGTATTTATCATCTGTC
	16(18)	CCCTAGTGGATGATAAGAATAATCAGTATGTGACTTGGATTGATCTATCTGTCTGTCTGTCTGTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCCATCTATCCATCCATCCTATGTATTTATCATCTGTC
	16(18')	CCCTAGTGGATGATAAGAATAATCAGTATGTGACTTGGATTGATCTATCTGTCTGTCTGTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCCATCTATCCATCCATCCTATGTATTTATCATCTGTC
	17(19)	CCCTAGTGGATGATAAGAATAATCAGTATGTGACTTGGATTGATCTATCTGTCTGTCTGTCTGTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCCATCTATCCATCCATCCTATGTATTTATCATCTGTC
	18(20)	CCCTAGTGGATGATAAGAATAATCAGTATGTGACTTGGATTGATCTATCTGTCTGTCTGTCTGTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCCATCTATCCATCCATCCTATGTATTTATCATCTGTC
	18'(20')	CCCTAGTGGATGATAAGAATAATCAGTATGTGACTTGGATTGATCTATCTGTCTGTCTGTCTGTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCCATCTATCCATCCATCCTATGTATTTATCATCTGTC
	18.1	CCCTAGTGGATGATAAGAATAATCAGTATGTGACTTGGATTGATCTATCTGTCTGTCTGTCTGTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTAATCTATCCATCTATCCATCCATCCTATGTATTTATCATCTGTC
	19(21)	CCCTAGTGGATGATAAGAATAATCAGTATGTGACTTGGATTGATCTATCTGTCTGTCTGTCTGTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCCATCTATCCATCCATCCTATGTATTTATCATCTGTC
	20 (22)	CCCTAGTGGATGATAAGAATAATCAGTATGTGACTTGGATTGATCTATCTGTCTGTCTGTCTGTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCCATCTATCCATCCATCCTATGTATTTATCATCTGTC
	21 (23)	CCCTAGTGGATGATAAGAATAATCAGTATGTGACTTGGATTGATCTATCTGTCTGTCTGTCTGTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCCATCTATCCATCCATCCTATGTATTTATCATCTGTC
	22 (24)	CCCTAGTGGATGATAAGAATAATCAGTATGTGACTTGGATTGATCTATCTGTCTGTCTGTCTGTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCCATCTATCCATCCATCCTATGTATTTATCATCTGTC
D3S1358	13	ACTGCAGTCCAATCTGGGTGACAGAGCAAGACCCTGTCTCATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGACAGACAGATAGATACATGCAAGCCTCTGTTGATTTCAT
	14	ACTGCAGTCCAATCTGGGTGACAGAGCAAGACCCTGTCTCATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGACAGACAGATAGATACATGCAAGCCTCTGTTGATTTCAT
	15	ACTGCAGTCCAATCTGGGTGACAGAGCAAGACCCTGTCTCATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGACAGACAGACAGATAGATACATGCAAGCCTCTGTTGATTTCAT
	15'	ACTGCAGTCCAATCTGGGTGACAGAGCAAGACCCTGTCTCATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGACAGACAGACAGATAGATACATGCAAGCCTCTGTTGATTTCAT
	16	ACTGCAGTCCAATCTGGGTGACAGAGCAAGACCCTGTCTCATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGACAGACAGACAGATAGATACATGCAAGCCTCTGTTGATTTCAT
	16'	ACTGCAGTCCAATCTGGGTGACAGAGCAAGACCCTGTCTCATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGACAGACAGACAGATAGATACATGCAAGCCTCTGTTGATTTCAT
	17	ACTGCAGTCCAATCTGGGTGACAGAGCAAGACCCTGTCTCATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGACAGACAGACAGATAGATACATGCAAGCCTCTGTTGATTTCAT
	17'	ACTGCAGTCCAATCTGGGTGACAGAGCAAGACCCTGTCTCATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGACAGACAGACAGATAGATACATGCAAGCCTCTGTTGATTTCAT
	18	ACTGCAGTCCAATCTGGGTGACAGAGCAAGACCCTGTCTCATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGACAGACAGACAGATAGATACATGCAAGCCTCTGTTGATTTCAT
	19	ACTGCAGTCCAATCTGGGTGACAGAGCAAGACCCTGTCTCATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGACAGACAGACAGATAGATACATGCAAGCCTCTGTTGATTTCAT

Underlined regions indicate the STR-specific primer binding sequences

Table i continued: Reference sequences used for identifying alleles that matched those reported in the STRbase (www.cstl.nist.gov/strbase)

Loci	Allele	Sequence (Forward primer | Sequence | Reverse Primer)
D7S820AT	6	AACACTTGTCATAGTTTAGAACGAACTAACGATAGATAGATAGATAGATAGATAGATAGACAGATTGATAGTTTTTTTTAATCTCACTAAATAGTCTATAGTAAACATTTAATTACCAATATTTGGTGCAATTCTGTCAATGA
	7	AACACTTGTCATAGTTTAGAACGAACTAACGATAGATAGATAGATAGATAGATAGATAGACAGATTGATAGTTTTTTTTAATCTCACTAAATAGTCTATAGTAAACATTTAATTACCAATATTTGGTGCAATTCTGTCAATGA
	8	AACACTTGTCATAGTTTAGAACGAACTAACGATAGATAGATAGATAGATAGATAGATAGATAGACAGATTGATAGTTTTTTTTAATCTCACTAAATAGTCTATAGTAAACATTTAATTACCAATATTTGGTGCAATTCTGTCAATGA
	9	AACACTTGTCATAGTTTAGAACGAACTAACGATAGATAGATAGATAGATAGATAGATAGATAGACAGATTGATAGTTTTTTTTAATCTCACTAAATAGTCTATAGTAAACATTTAATTACCAATATTTGGTGCAATTCTGTCAATGA
	10	AACACTTGTCATAGTTTAGAACGAACTAACGATAGATAGATAGATAGATAGATAGATAGATAGACAGATTGATAGTTTTTTTTAATCTCACTAAATAGTCTATAGTAAACATTTAATTACCAATATTTGGTGCAATTCTGTCAATGA
	11	AACACTTGTCATAGTTTAGAACGAACTAACGATAGATAGATAGATAGATAGATAGATAGATAGACAGATTGATAGTTTTTTTTAATCTCACTAAATAGTCTATAGTAAACATTTAATTACCAATATTTGGTGCAATTCTGTCAATGA
	12	AACACTTGTCATAGTTTAGAACGAACTAACGATAGATAGATAGATAGATAGATAGATAGATAGACAGATTGATAGTTTTTTTTAATCTCACTAAATAGTCTATAGTAAACATTTAATTACCAATATTTGGTGCAATTCTGTCAATGA
	13	AACACTTGTCATAGTTTAGAACGAACTAACGATAGATAGATAGATAGATAGATAGATAGATAGACAGATTGATAGTTTTTTTTAATCTCACTAAATAGTCTATAGTAAACATTTAATTACCAATATTTGGTGCAATTCTGTCAATGA
	14	AACACTTGTCATAGTTTAGAACGAACTAACGATAGATAGATAGATAGATAGATAGATAGATAGACAGATTGATAGTTTTTTTTAATCTCACTAAATAGTCTATAGTAAACATTTAATTACCAATATTTGGTGCAATTCTGTCAATGA
D7S820TT	6	AACACTTGTCATAGTTTAGAACGAACTAACGATAGATAGATAGATAGATAGATAGACAGATTGATAGTTTTTTTTTATCTCACTAAATAGTCTATAGTAAACATTTAATTACCAATATTTGGTGCAATTCTGTCAATGA
	7	AACACTTGTCATAGTTTAGAACGAACTAACGATAGATAGATAGATAGATAGATAGACAGATTGATAGTTTTTTTTTATCTCACTAAATAGTCTATAGTAAACATTTAATTACCAATATTTGGTGCAATTCTGTCAATGA
	8	AACACTTGTCATAGTTTAGAACGAACTAACGATAGATAGATAGATAGATAGATAGATAGACAGATTGATAGTTTTTTTTTATCTCACTAAATAGTCTATAGTAAACATTTAATTACCAATATTTGGTGCAATTCTGTCAATGA
	9	AACACTTGTCATAGTTTAGAACGAACTAACGATAGATAGATAGATAGATAGATAGATAGATAGACAGATTGATAGTTTTTTTTTATCTCACTAAATAGTCTATAGTAAACATTTAATTACCAATATTTGGTGCAATTCTGTCAATGA
	10	AACACTTGTCATAGTTTAGAACGAACTAACGATAGATAGATAGATAGATAGATAGATAGATAGACAGATTGATAGTTTTTTTTTATCTCACTAAATAGTCTATAGTAAACATTTAATTACCAATATTTGGTGCAATTCTGTCAATGA
	11	AACACTTGTCATAGTTTAGAACGAACTAACGATAGATAGATAGATAGATAGATAGATAGATAGACAGATTGATAGTTTTTTTTTATCTCACTAAATAGTCTATAGTAAACATTTAATTACCAATATTTGGTGCAATTCTGTCAATGA
	12	AACACTTGTCATAGTTTAGAACGAACTAACGATAGATAGATAGATAGATAGATAGATAGATAGACAGATTGATAGTTTTTTTTTATCTCACTAAATAGTCTATAGTAAACATTTAATTACCAATATTTGGTGCAATTCTGTCAATGA
	13	AACACTTGTCATAGTTTAGAACGAACTAACGATAGATAGATAGATAGATAGATAGATAGATAGACAGATTGATAGTTTTTTTTTATCTCACTAAATAGTCTATAGTAAACATTTAATTACCAATATTTGGTGCAATTCTGTCAATGA
	14	AACACTTGTCATAGTTTAGAACGAACTAACGATAGATAGATAGATAGATAGATAGATAGATAGACAGATTGATAGTTTTTTTTTATCTCACTAAATAGTCTATAGTAAACATTTAATTACCAATATTTGGTGCAATTCTGTCAATGA

Loci	Allele	Sequence (Forward primer | Sequence | Reverse Primer)
D7S820AG	6	AACACTTGTCATAGTTTAGAACGAACTAACGATAGATAGATAGATAGATAGATAGACAGATTGATAGTTTTTTTTAATCTCACTAAATAGTCTATAGTAAACATTTAATTACCAATATGTGGTGCAATTCTGTCAATGA
	7	AACACTTGTCATAGTTTAGAACGAACTAACGATAGATAGATAGATAGATAGATAGATAGACAGATTGATAGTTTTTTTTAATCTCACTAAATAGTCTATAGTAAACATTTAATTACCAATATGTGGTGCAATTCTGTCAATGA
	8	AACACTTGTCATAGTTTAGAACGAACTAACGATAGATAGATAGATAGATAGATAGATAGACAGATTGATAGTTTTTTTTAATCTCACTAAATAGTCTATAGTAAACATTTAATTACCAATATGTGGTGCAATTCTGTCAATGA
	9	AACACTTGTCATAGTTTAGAACGAACTAACGATAGATAGATAGATAGATAGATAGATAGATAGACAGATTGATAGTTTTTTTTAATCTCACTAAATAGTCTATAGTAAACATTTAATTACCAATATGTGGTGCAATTCTGTCAATGA
	10	AACACTTGTCATAGTTTAGAACGAACTAACGATAGATAGATAGATAGATAGATAGATAGATAGACAGATTGATAGTTTTTTTTAATCTCACTAAATAGTCTATAGTAAACATTTAATTACCAATATGTGGTGCAATTCTGTCAATGA
	11	AACACTTGTCATAGTTTAGAACGAACTAACGATAGATAGATAGATAGATAGATAGATAGATAGACAGATTGATAGTTTTTTTTAATCTCACTAAATAGTCTATAGTAAACATTTAATTACCAATATGTGGTGCAATTCTGTCAATGA
	12	AACACTTGTCATAGTTTAGAACGAACTAACGATAGATAGATAGATAGATAGATAGATAGATAGACAGATTGATAGTTTTTTTTAATCTCACTAAATAGTCTATAGTAAACATTTAATTACCAATATGTGGTGCAATTCTGTCAATGA
	13	AACACTTGTCATAGTTTAGAACGAACTAACGATAGATAGATAGATAGATAGATAGATAGATAGACAGATTGATAGTTTTTTTTAATCTCACTAAATAGTCTATAGTAAACATTTAATTACCAATATTTGGTGCAATTCTGTCAATGA
	14	AACACTTGTCATAGTTTAGAACGAACTAACGATAGATAGATAGATAGATAGATAGATAGATAGACAGATTGATAGTTTTTTTTAATCTCACTAAATAGTCTATAGTAAACATTTAATTACCAATATGTGGTGCAATTCTGTCAATGA
D7S820TG	6	AACACTTGTCATAGTTTAGAACGAACTAACGATAGATAGATAGATAGATAGATAGACAGATTGATAGTTTTTTTTTATCTCACTAAATAGTCTATAGTAAACATTTAATTACCAATATGTGGTGCAATTCTGTCAATGA
	7	AACACTTGTCATAGTTTAGAACGAACTAACGATAGATAGATAGATAGATAGATAGACAGATTGATAGTTTTTTTTTATCTCACTAAATAGTCTATAGTAAACATTTAATTACCAATATGTGGTGCAATTCTGTCAATGA
	8	AACACTTGTCATAGTTTAGAACGAACTAACGATAGATAGATAGATAGATAGATAGACAGATTGATAGTTTTTTTTTATCTCACTAAATAGTCTATAGTAAACATTTAATTACCAATATGTGGTGCAATTCTGTCAATGA
	9	AACACTTGTCATAGTTTAGAACGAACTAACGATAGATAGATAGATAGATAGATAGATAGACAGATTGATAGTTTTTTTTTATCTCACTAAATAGTCTATAGTAAACATTTAATTACCAATATGTGGTGCAATTCTGTCAATGA
	10	AACACTTGTCATAGTTTAGAACGAACTAACGATAGATAGATAGATAGATAGATAGATAGATAGACAGATTGATAGTTTTTTTTTATCTCACTAAATAGTCTATAGTAAACATTTAATTACCAATATGTGGTGCAATTCTGTCAATGA
	11	AACACTTGTCATAGTTTAGAACGAACTAACGATAGATAGATAGATAGATAGATAGATAGATAGACAGATTGATAGTTTTTTTTTATCTCACTAAATAGTCTATAGTAAACATTTAATTACCAATATGTGGTGCAATTCTGTCAATGA
	12	AACACTTGTCATAGTTTAGAACGAACTAACGATAGATAGATAGATAGATAGATAGATAGATAGACAGATTGATAGTTTTTTTTTATCTCACTAAATAGTCTATAGTAAACATTTAATTACCAATATGTGGTGCAATTCTGTCAATGA
	13	AACACTTGTCATAGTTTAGAACGAACTAACGATAGATAGATAGATAGATAGATAGATAGATAGACAGATTGATAGTTTTTTTTTATCTCACTAAATAGTCTATAGTAAACATTTAATTACCAATATGTGGTGCAATTCTGTCAATGA
	14	AACACTTGTCATAGTTTAGAACGAACTAACGATAGATAGATAGATAGATAGATAGATAGATAGACAGATTGATAGTTTTTTTTTATCTCACTAAATAGTCTATAGTAAACATTTAATTACCAATATGTGGTGCAATTCTGTCAATGA

Underlined regions indicate the STR-specific primer binding sequences. Sequences deduced from the repeat unit sequences recorded on STRbase and the flanking regions reported by Oberacher *et al.* (2008).

Table i continued: Reference sequences used for identifying alleles that matched those reported in the STRbase (www.cstl.nist.gov/strbase)

Loci	Allele	Sequence (Forward primer | Sequence | Reverse Primer)
D13S317A	7	TCTGACCCATCTAACGCCTATCTGTATTTACAAATACATTATCTATCTATCTATCTATCTATCTATCAATCAATCATCTATCTATCTTTCTGTCTG
	8	TCTGACCCATCTAACGCCTATCTGTATTTACAAATACATTATCTATCTATCTATCTATCTATCTATCTATCAATCAATCATCTATCTATCTTTCTGTCTG
	9	TCTGACCCATCTAACGCCTATCTGTATTTACAAATACATTATCTATCTATCTATCTATCTATCTATCTATCTATCAATCAATCATCTATCTATCTTTCTGTCTG
	10	TCTGACCCATCTAACGCCTATCTGTATTTACAAATACATTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCAATCAATCATCTATCTATCTTTCTGTCTG
	11	TCTGACCCATCTAACGCCTATCTGTATTTACAAATACATTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCAATCAATCATCTATCTATCTTTCTGTCTG
	12	TCTGACCCATCTAACGCCTATCTGTATTTACAAATACATTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCAATCAATCATCTATCTATCTTTCTGTCTG
	13	TCTGACCCATCTAACGCCTATCTGTATTTACAAATACATTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCAATCAATCATCTATCTATCTTTCTGTCTG
	14	TCTGACCCATCTAACGCCTATCTGTATTTACAAATACATTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCAATCAATCATCTATCTATCTTTCTGTCTG
	15	TCTGACCCATCTAACGCCTATCTGTATTTACAAATACATTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCAATCAATCATCTATCTATCTTTCTGTCTG
D13S317T	7	TCTGACCCATCTAACGCCTATCTGTATTTACAAATACATTATCTATCTATCTATCTATCTATCTATCTATCTATCAATCATCTATCTATCTTTCTGTCTG
	8	TCTGACCCATCTAACGCCTATCTGTATTTACAAATACATTATCTATCTATCTATCTATCTATCTATCTATCTATCAATCATCTATCTATCTTTCTGTCTG
	9	TCTGACCCATCTAACGCCTATCTGTATTTACAAATACATTATCTATCTATCTATCTATCTATCTATCTATCTATCAATCATCTATCTATCTTTCTGTCTG
	10	TCTGACCCATCTAACGCCTATCTGTATTTACAAATACATTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCAATCATCTATCTATCTTTCTGTCTG
	11	TCTGACCCATCTAACGCCTATCTGTATTTACAAATACATTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCAATCATCTATCTATCTTTCTGTCTG
	12	TCTGACCCATCTAACGCCTATCTGTATTTACAAATACATTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCAATCATCTATCTATCTTTCTGTCTG
	13	TCTGACCCATCTAACGCCTATCTGTATTTACAAATACATTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCAATCATCTATCTATCTTTCTGTCTG
	14	TCTGACCCATCTAACGCCTATCTGTATTTACAAATACATTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCAATCATCTATCTATCTTTCTGTCTG
	15	TCTGACCCATCTAACGCCTATCTGTATTTACAAATACATTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCAATCATCTATCTATCTTTCTGTCTG
D2S441	8	CTGTGGCTCATCTATGAAAACTTCTATCTATCTATCTATCTATCTATCTATCTATCTATATCATAACACCACAGCCACTTC
	9	CTGTGGCTCATCTATGAAAACTTCTATCTATCTATCTATCTATCTATCTATCTATCTATATCATAACACCACAGCCACTTC
	10	CTGTGGCTCATCTATGAAAACTTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATATCATAACACCACAGCCACTTC
	11	CTGTGGCTCATCTATGAAAACTTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATATCATAACACCACAGCCACTTC
	12	CTGTGGCTCATCTATGAAAACTTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATATCATAACACCACAGCCACTTC
	13	CTGTGGCTCATCTATGAAAACTTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATATCATAACACCACAGCCACTTC
	14	CTGTGGCTCATCTATGAAAACTTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATATCATAACACCACAGCCACTTC
	15	CTGTGGCTCATCTATGAAAACTTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATATCATAACACCACAGCCACTTC
	16	CTGTGGCTCATCTATGAAAACTTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATATCATAACACCACAGCCACTTC
	17	CTGTGGCTCATCTATGAAAACTTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATATCATAACACCACAGCCACTTC

Underlined regions indicate the STR-specific primer binding sequences. Sequences deduced from the repeat unit sequences recorded on STRbase and the flanking regions reported by Oberacher *et al.* (2008).

Table i continued: Reference sequences used for identifying alleles that matched those reported in the STRbase (www.cstl.nist.gov/strbase)

Loci	Allele	Sequence (Forward primer | Sequence | Reverse Primer)
D21S11	24	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCGTCTATCTATCCAGTCTATCTACC
	24.2	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCTATCTATCTATCTATCTATCTATCGTCTATCTATCCAGTCTATCTACC
	25	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCTATCGTCTATCTATCCAGTCTATCTACC
	25.2	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCTATCGTCTATCTATCCAGTCTATCTACC
	26	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCGTCTATCTATCCAGTCTATCTACC
	27	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCGTCTATCTATCCAGTCTATCTACC
	27'	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCGTCTATCTATCCAGTCTATCTACC
	27''	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCGTCTATCTATCCAGTCTATCTACC
	28.2	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCGTCTATCTATCCAGTCTATCTACC
	28'	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCGTCTATCTATCCAGTCTATCTACC
	28.2'	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCGTCTATCTATCCAGTCTATCTACC
	29	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCGTCTATCTATCCAGTCTATCTACC
	29'	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCGTCTATCTATCCAGTCTATCTACC
	29.1	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTGTCTATCTATCTATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCGTCTATCTATCCAGTCTATCTACC
	29.2	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCGTCTATCTATCCAGTCTATCTACC
	30	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCGTCTATCTATCCAGTCTATCTACC
	30'	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCGTCTATCTATCCAGTCTATCTACC
	30''	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCGTCTATCTATCCAGTCTATCTACC
	30'''	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCGTCTATCTATCCAGTCTATCTACC
	30.2	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCGTCTATCTATCCAGTCTATCTACC
	30.2'	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCGTCTATCTATCCAGTCTATCTACC
	31	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCGTCTATCTATCCAGTCTATCTACC
	31'	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCGTCTATCTATCCAGTCTATCTACC
	31''	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCGTCTATCTATCCAGTCTATCTACC
	31'''	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCGTCTATCTATCCAGTCTATCTACC
	31.2	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCGTCTATCTATCCAGTCTATCTACC
	32	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCGTCTATCTATCCAGTCTATCTACC
	32'	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCGTCTATCTATCCAGTCTATCTACC
	32.2	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCGTCTATCTATCCAGTCTATCTACC
	32.2'	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCGTCTATCTATCCAGTCTATCTACC

Underlined regions indicate the STR-specific primer binding sequences. Sequences deduced from the repeat unit sequences recorded on STRbase and the flanking regions reported by Oberacher *et al.* (2008).

Table i continued: Reference sequences used for identifying alleles that matched those reported in the STRbase (www.cstl.nist.gov/strbase)

Loci	Allele	Sequence (Forward primer | Sequence | Reverse Primer)
D21S11	32.2''	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATATCTATCGTCTATCTATCCAGTCTATCTACC
	33	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCGTCTATCTATCCAGTCTATCTACC
	33.2	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATATCTATCGTCTATCTATCCAGTCTATCTACC
	33.2′	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATATCTATCGTCTATCTATCCAGTCTATCTACC
	33.2''	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATATCTATCGTCTATCTATCCAGTCTATCTACC
	33.3	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATATCTATCGTCTATCTATCCAGTCTATCTACC
	34	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATATCTATCGTCTATCTATCCAGTCTATCTACC
	34.2	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATATCTATCGTCTATCTATCCAGTCTATCTACC
	34	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATATCTATCGTCTATCTATCCAGTCTATCTACC
	34.3	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATATCTATCGTCTATCTATCCAGTCTATCTACC
	35	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATATCTATCGTCTATCTATCCAGTCTATCTACC
	35.1	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATATCTATCGTCTATCTATCCAGTCTATCTACC
	35.2	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATATCTATCGTCTATCTATCCAGTCTATCTACC
	36	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATATCTATCGTCTATCTATCCAGTCTATCTACC
	36′	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATATCTATCGTCTATCTATCCAGTCTATCTACC
	36''	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATATCTATCGTCTATCTATCCAGTCTATCTACC
	36.2	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATATCTATCGTCTATCTATCCAGTCTATCTACC
	37	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATATCTATCGTCTATCTATCCAGTCTATCTACC
	37′	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATATCTATCGTCTATCTATCCAGTCTATCTACC
	37.2	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATATCTATCGTCTATCTATCCAGTCTATCTACC
	37.2′	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATATCTATCGTCTATCTATCCAGTCTATCTACC
	37.2''	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATATCTATCGTCTATCTATCCAGTCTATCTACC
	37.2'''	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCGTCTATCTATCCAGTCTATCTACC
	37.2''',	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCGTCTATCTATCCAGTCTATCTACC
	38	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCGTCTATCTATCCAGTCTATCTACC
	38′	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATATCTATCGTCTATCTATCCAGTCTATCTACC
	38''	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCGTCTATCTATCCAGTCTATCTACC
	38'',	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCGTCTATCTATCCAGTCTATCTACC
	38.2	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCGTCTATCTATCCAGTCTATCTACC
	38.2′	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCGTCTATCTATCCAGTCTATCTACC
	38.2''	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCGTCTATCTATCCAGTCTATCTACC
	39.2	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCGTCTATCTATCCAGTCTATCTACC
	39.2′	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCGTCTATCTATCCAGTCTATCTACC
	41.2	ATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCGTCTATCTATCCAGTCTATCTACC

Underlined regions indicate the STR-specific primer binding sequences. Sequences deduced from the repeat unit sequences recorded on STRbase and the flanking regions reported by Oberacher *et al.* (2008).

Table ii: Allele frequencies for STRs sequenced using the Roche GS Junior system from Nguni and Sotho-Tswana (ST) populations.

Locus	Allele	Frequency	
		Nguni	ST
vWA	14(16'')	0.019	0.000
	14''(16'''')	0.018	0.030
	14''''(16'''''')	0.027	0.064
	14''''(16'''''')	0.083	0.075
	15(17)	0.181	0.201
	15´(17')	0.075	0.119
	15´'(17'')	0.013	0.000
	16(18)	0.214	0.202
	16'(18')	0.078	0.066
	17(19)	0.147	0.155
	17'(19')	0.033	0.000
	18(20)	0.073	0.059
	18''(20'')	0.014	0.000
	19(21)	0.012	0.029
	19'(21')	0.015	0.000
D13S317	10_A	0.015	0.028
	10_T	0.029	0.023
	11_A	0.144	0.239
	11_T	0.122	0.186
	12_A	0.285	0.312
	12_T	0.127	0.068
	13´_T	0.063	0.000
	13_A	0.096	0.079
	13_T	0.033	0.035
	14_A	0.087	0.029
D7S820	7_TT	0.008	0.000
	8_TT	0.255	0.152
	9_TT	0.125	0.180
	9_TG	0.025	0.000
	10_TT	0.335	0.323
	10.1_TT	0.034	0.040
	10_TG	0.008	0.075
	11_TT	0.122	0.092
	11.1_TT	0.025	0.027
	11_AG	0.035	0.032
	11_TG	0.009	0.048
	12_TT	0.019	0.031
D21S11	26	0.021	0.023
	26'	0.000	0.021
	27	0.150	0.163
	27''	0.043	0.042
	27'''	0.011	0.019
	28(28.2)*	0.264	0.400
	28'	0.000	0.000
	28''	0.011	0.000
	28'''	0.000	0.016

Continued..			
	29	0.078	0.085
	29''	0.011	0.018
	30	0.018	0.000
	30'	0.035	0.041
	30''	0.062	0.039
	30'''	0.034	0.000
	30''''	0.026	0.000
D21S11	31	0.023	0.000
	31''	0.040	0.034
	31.2	0.045	0.062
	31.2'	0.010	0.000
	32''	0.017	0.000
	32.2	0.080	0.036
	33.2	0.018	0.000
D2S441	9	0.008	0.000
	10	0.085	0.071
	10'	0.000	0.042
	10.3	0.007	0.000
	11	0.375	0.350
	11'	0.016	0.000
	11.3	0.028	0.032
	12	0.092	0.156
	12'	0.040	0.000
	12.3	0.016	0.041
	12.3´	0.000	0.025
	13´	0.026	0.039
	14'	0.214	0.211
	14.3	0.024	0.000
	15´	0.051	0.034
	16'	0.017	0.000
D3S1358	12	0.014	0.021
	13	0.014	0.000
	13´	0.030	0.000
	14	0.065	0.099
	14'	0.124	0.045
	15	0.019	0.015
	15´	0.169	0.169
	15´'	0.094	0.121
	15´''	0.011	0.000
	16	0.044	0.077
	16'	0.150	0.120
	16'''	0.111	0.133
	17	0.040	0.080
	17'	0.067	0.046
	17''	0.022	0.025
	18	0.019	0.050
	18'	0.007	0.000

Putative novel variants, which exhibited sequence variations from those reported in the STRBase (www.cstl.nist.gov/strbase), are indicated in **red**.

Table iii: Sequences obtained by 454 sequencing which did not align to the forward and reverse primer sequences of any target STR

Variant No.	MID	Sequence
1	Nguni and ST	CGCCAGGGTTTTCCCAGTCACGACTCTGACCCATCTAACACCACAGCCACTTCGTCATAGCTGTTTCCTGTGTGA
2	Nguni and ST	CGCCAGGGTTTTCCCAGTCACGACTCTGACCCATCTAACGCCACAGCCACTTCGTCATAGCTGTTTCCTGTGTGA
3	Nguni	CGCCAGGGTTTTCCCAGTCACGACTCTGACCCATCTATCTATCTTTCTGTCTGGTCATAGCTGTTTCCTGTGTGA
4	Nguni	CGCCAGGGTTTTCCCAGTCACGACCTGTGGCTCATCTATGAAACTTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATATCATAACACC ACAGCCACTTCGTCATAGCTGTTTCCTGTGTGA
5	Nguni	CGCCAGGGTTTCCCAGTCACGACCTGTGGCTCATCTATGAAACTTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATATCATAACACCA CAGCCACTTCGTCATAGCTGTTTCCTGTGTGA
6	Nguni	CGCCAGGGTTTTCCCAGTCACGACCTGTGGCTCATCTATGAAACTTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATTTATCTATCTA TATCATAACACCACAGCCACTTCGTCATAGCTGTTTCCTGTGTGA
7	Nguni	CGCCAGGGTTTTCCCAGTCACGACCCCTAGTGGATGATAAGAATAATCAGTATGTGACTTGGATTGATCTATCTGTCTGTCTGTCTGTCTATCTATCCA GTCTATCTACCGTCATAGCTGTTTCCTGTGTGA
8	Nguni	CGCCAGGGTTTCCCAGTCACGACCTGTGGCTCATCTATGAAACTTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATTTATCTATCTAT ATCATAACACCACAGCCACTTCGTCATAGCTGTTTCCTGTGTGA
9	Nguni	CGCCAGGGTTTTCCCAGTCACGACTCTGACCCATCTAACGCCTCGCCAGGGTTTTCCCAGTCACGACTCTGACCCATCTAACACCACAGCCACTTCG TCATAGCTGTTTCCTGTGTGA