EVALUATION OF COMMERCIALLY AVAILABLE DNA TESTING KITS TO RESOLVE COMPLEX KINSHIP DISPUTES IN A SOUTH AFRICAN SETTING

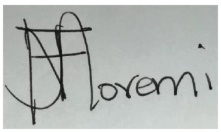
Mashela Mahlatse Moremi

A dissertation submitted to the faculty of Human Sciences, University of the Witwatersrand, Johannesburg, in fulfilment of the requirements for the degree of Master of Science in Medicine

Johannesburg, 2020

DECLARATION

I, Mashela Mahlatse Moremi, declare that this dissertation is my own work. It is being submitted for the degree of Master of Science in Medicine in the University of Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at this or any other University.



12th Day of October 2020 in Johannesburg

DEDICATION

As I prepare to populate this page, my eyes well up with tears, as I look back at the road travelled thus far, and I am so thankful to the Almighty God, for making everything possible, for opening doors that I never imagined would open for me. All I can say is that this is a testimony of the wondrous works that only You Lord can do! I am forever grateful for Your love, mercy, grace, wisdom and ever-present support and help throughout this journey and beyond. To You alone be the honour, glory, exaltation and praise, now and always. You are truly faithful in all your ways. You are a covenant-keeping, mountain-moving, destiny-changing God, Hallelujah!

To my husband and my children, this was a tough journey that I could not have overcome without your support, love and understanding. For all the times when I was not there for you, all the lonely times that you had to endure, thank you. You are all very dear to my heart, and all this would be meaningless if I could not share the fruits of my labour with you. I love you all so much. To my mum, thank you for your prayers and encouragement throughout this journey. You are really a precious gift from God, a jewel and treasure to always cherish. I love you so much.

PRESENTATIONS ARISING FROM THIS RESEARCH PROJECT

Pathology Research and Development Congress (PathRed), 18-21 July 2019, Destiny Hotel and Convention Centre, Ekurhuleni, Johannesburg, South Africa.

ABSTRACT

Background: Human identification techniques using biological material have been under development for over a century. By the early 2000s, routine DNA profiling, using short tandem repeats (STRs), was in place as a standard of practice in both forensic science and in parentage testing laboratories. Use of the commercial kits, containing less than 20 STR loci, led to inconclusive results in certain cases, particularly when kinship testing had to be performed in the absence of one or both parents. It is proposed that better resolution can be achieved when more STR loci are used to calculate Likelihood Ratios (LRs) for establishing biological relationships. Recently, newer kits containing more than 20 STR loci, have become available. The present study aimed to look at the utility of two new commercial kits with expanded numbers of STR loci and to assess the validity of outcomes when fewer loci were utilized for assessment of complex kinship queries, in a large parentage testing laboratory within the National Health Service in South Africa. Further, the study aimed to evaluate different software packages available for performing the statistical calculations associated with this genetic testing approach.

Subjects and methods: A total of 729 samples, collected from unrelated individuals from seven ethnic groups, namely: South African Bantu-speakers (n=366), South African Whites (n=84), South African mixed ancestry (Coloureds) (n=40), South African Indians (n=26), Zimbabwean Bantu-speakers (n=113), Nigerian Bantu-speakers (n=73) and Namibian Bantu-speakers (n=27), were used in the present study. STR-based polymerase chain reaction (PCR) amplification was performed using either the PowerPlex® Fusion system (Promega Corporation) or the VeriFiler™ Express PCR amplification kit (Life Technologies), both of which contain more than 20 STR loci, compared to the DNA Profiling Laboratory (DPL)'s currently used commercial kits which contain 16 or 18 STR loci. Allele frequencies were calculated for each ethnic group where n>40, using Microsoft Excel® 2016 (counting method). Hardy-Weinberg

Equilibrium (HWE), linkage disequilibrium (LD) and population genetic variation tests were performed using *GenePop*. Kinship queries, especially those with inconclusive outcomes when 15 autosomal STR loci were used in the initial, routine testing, were retested using the new expanded kits. Outcomes obtained from using four different analysis software packages, namely, the In-house Microsoft Excel macro-enabled software, *Converge*™ Software, *GeneMarker*® *HID* and *Familias* were compared to assess if there was a statistical difference in the outcomes obtained when calculating LRs using these different analysis tools.

Results: The genotype results showed that the probability of identity (PI) increased as the number of STR loci contained in a commercial kit increased. Thus, a significantly higher discrimination power was achieved when the PowerPlex® Fusion system and the VeriFiler™ Express PCR kit were used, as compared to the routinely used kits which contain 16 or 18 STR loci. More than 50% of cases with initially inconclusive outcomes, when retested with expanded kits, could now be resolved based on additional STR loci. Populations investigated in the present study were found to be in HWE when assessed for heterozygosity excess. Departures from HWE were detected in the SA and Nigerian Bantu-speakers, when assessed for heterozygosity deficiency. Linkage disequilibrium was observed (erroneously) between loci located on different chromosomes. Genetic diversity tests (F_{st}) showed that the SA and Zimbabwean Bantu-speakers can be clustered together, and the same allele frequencies can be used for calculating LRs involving these two groups. The analysis software packages evaluated all gave *p*-values > 0.05 when compared to each other, indicating that there was no significant difference in the Likelihood Ratios obtained when different analysis software packages were used.

Conclusion: The inclusion of additional STR loci in the newer commercial kits for human identity testing was shown to be beneficial in kinship analysis. It is recommended that all kinship queries should be tested using the newer kits as a first line testing approach. Additional testing, using lineage markers such as X and Y chromosomes and mitochondrial DNA, when applicable, is recommended as part of first line testing, to reduce long turn-around times and ensure conclusive outcomes. The In-house data analysis software package currently used in DPL was validated for routine use.

ACKNOWLEDGEMENTS

To my supervisors, Professor Himla Soodyall, Dr Venesa Sahibdeen and Dr Robyn Kerr, thank you so much for the guidance, feedback and support throughout this journey. I remember how every time I walked into Prof Soodyall' s office, she would give me such good advice that I always took a pen and a notebook with me so that I would not miss any of your lessons. I had planned to buy a voice recorder, so that I could record all of our conversations, but I guess I was late to act on this idea, as you resigned before I could. You are the one person, in my entire life, who was able to tell me the truth to my face, no matter how hurtful. I appreciate that because you said it with so much love and those hurtful truths continue to change my life for the better, thank you so much. To Dr Sahibdeen, thank you so much "V" for all your help with my writing, for all your timeous feedbacks and for not giving up on me, even though I would have understood if you did. To Dr Kerr, thank you for helping me with the pitch for my writing. I hope I managed to improve on my writing. I do understand that the journey continues and that I will continue to improve as I go. For how far we have come, thank you all!

To Dr Anthony Lane, thank you for your help with putting the updated database together, and thank you for always being there to help me whenever I reached out to you. For your help with constructing pedigrees on *Familias* software, you have been truly amazing.

To Ms Angela Turner, thank you for being more than my laboratory manager. Thank you for being my friend, mentor and sounding board. Thank you for always being willing to listen and for always giving me sound advice. I will forever treasure the relationship we have.

To Mr Skhumbuzo Mkhwanazi (Skeem saam), thank you for all your help with printing out the reports and for always being there to help whenever I reached out to you. You have shared this load with me without complaining. Thank you for making the load lighter for me!

Mr Andrew May, thank you for your help with finding and sharing the software to analyse my data, for the editing tips, and resources to use during writing up. Thank you for always keeping your door and heart open, your ears attentive, for all your advice and for always welcoming me with a warm smile.

I would like to thank Professor Zané Lombard for advice on how to present some of the findings in my write up. Thank you for always being willing to listen and help where you could. Thank you for the time you spent helping me to put a plan in place to get all the work done on time.

Professor Amanda Krause, the Head of Department for Human Genetics, thank you for giving me the opportunity to register for the course, access to data, and setting up progress meetings to see to it that the project progresses towards completion. Thank you for suggesting a good plan to get everything wrapped up.

My heartfelt gratitude goes to Dr Teresa Snyder-Leiby for providing training on *GeneMarker HID* (online meetings), for helping with trouble shooting and continued support throughout the project. For always attending speedily to my queries and helping me beyond the scope of *GeneMarker HID*.

Ms Parmowtee Goodur, thank you for helping me with troubleshooting for VeriFiler™Express PCR amplification kit, for help with the *Converge*™ Software, and for always being willing to help and providing support whenever I reached out to you. Thank you for your friendship, we have become like sisters through this experience!

To the Anatech Team; Ms Barbara Moore and Mr Gerald Kimbi Chiafiinii and the Promega Genetic Identity Team; Mr Stephan Kutranov and Mr Andy Hopwood. Thank you for your help with trouble shooting for the PowerPlex® Fusion system and for giving us a free kit to generate data for the present study. Thank you for always responding to our queries speedily and for keeping us up to date with new developments at Promega Corporation.

Lastly, thank you to the NHLS for funding my tuition fees and NHLS Research Trust (NHLSRT) for funding the project.

TABLE OF CONTENTS

DECLARATION II
DEDICATION III
PRESENTATIONS ARISING FROM THIS RESEARCH PROJECT IV
ABSTRACT V
ACKNOWLEDGEMENTS VIII
TABLE OF CONTENTS XI
LIST OF FIGURES XIV
LIST OF TABLES XV
LIST OF ABBREVIATIONS XVII
CHAPTER 1 INTRODUCTION 1
1.1 An overview of the South African population groups 1
1.2 Human identification testing using DNA 3
1.3 DNA profiling using short tandem repeats 5
1.4 Commercially available STR based kits used in human identification 7
1.5 Mutation rate for STRs 9
1.6 Parentage testing 11
1.7 Statistical calculations 13
1.8 Population specific allele frequencies 18
1.9 Kinship testing 21
1.10 DNA profiling at the National Health Laboratory Service 22
1.11 The current testing strategy at the DNA Profiling Laboratory 23
1.12 Additional testing available to resolve inconclusive outcomes 24
1.13 Challenges faced with the current testing strategy 28
1.14 What causes kinship analysis to be complex? 28
1.15 Potential solutions regarding inconclusive kinship outcomes 29
1.16 Comparison of software packages used to calculate Likelihood Ratios 32
1.17 Study aim and objectives 33

CHAPTER 2 MATERIALS AND METHODS....34
2.1 Subjects....34
2.2 Methods....37
2.2.1 DNA extraction....37
2.2.2 DNA amplification....37
2.2.3 Capillary electrophoresis....42
2.3 Data analysis....43
2.3.1 Updating of the database to include the newly generated allele frequencies....44
2.3.2 Retesting and evaluation of kinship cases previously tested using PowerPlex® 16 HS and PowerPlex® 18D....45
2.3.3 Comparison of Likelihood Ratios using different data analysis software packages....46
CHAPTER 3 RESULTS....48
3.1 Outcomes of allele frequency calculations....48
3.2 Number of alleles observed per locus....50
3.3 Number of observed genotypes....52
3.4 Observed heterozygosity....55
3.5 Probability of Identity....58
3.6 The Polymorphic Information Content....62
3.7 The ranking of the STR loci contained in the commercial kits assessed in the present study, per population group, based on the PI value....65
3.8 Hardy-Weinberg equilibrium....67
3.9 Assessment of linkage disequilibrium....68
3.10 Genetic variations between the population groups....69
3.11 Retesting and evaluation of kinship cases previously tested using PowerPlex® 16 HS or PowerPlex® 18D....69
3.12 Comparison of Likelihood Ratios when different data analysis software packages were used....71

CHAPTER 4 DISCUSSION 73
4.1 Outcomes of allele frequency calculations 73
4.2 Number of alleles observed 74
4.3 Number of observed genotypes and heterozygosity 75
4.4 Probability of Identity 77
4.5 Polymorphic Information Content and ranking of STR loci 79
4.6 Assessment of Hardy-Weinberg equilibrium 81
4.7 Assessment of linkage disequilibrium 82
4.8 Genetic variation between population groups 83
4.9 Self-declared ethnic groups and relatedness between individuals included in the study. 84
4.10 Reassessment of kinship cases previously tested with PowerPlex® 16HS and PowerPlex® 18D systems 85
4.11 Comparison of Likelihood Ratios when different data analysis software packages were used 87
CHAPTER 5 CONCLUSION 90
CHAPTER 6 REFERENCES 93
6.1 ELECTRONIC RESOURCES 99
CHAPTER 7 APPENDICES 102
7.1 APPENDIX A1: Ethics Clearance 102
7.2 APPENDIX A2: Change of dissertation title 103
7.3 APPENDIX B1: Consent form used in the DNA Profiling Laboratory (2018 to date) 104
7.4 APPENDIX B2: Consent form used in the DNA Profiling lab prior to 2018 105
7.5 APPENDIX C: Sample Map template 106
7.6 APPENDIX D: Population allele frequencies for the different ethnic groups. 107
7.7 APPENDIX E: Linkage disequilibrium results 116
7.8 APPENDIX F: Validation report for comparing different data analysis software packages used in kinship analysis 135

LIST OF FIGURES

Figure 1.1 Bantu migrations 2
Figure 1.2 An overview of advances experienced in human identification testing. 4
Figure 1.3 Showing the different number of repeats at an STR locus, and the flanking regions where the primers bind. 6
Figure 1.4 Mutation observed as a result of strand slippage. 10
Figure 1.5 Paternity testing outcome illustration. 12
Figure 1.6 Scenarios where X chromosome testing will be applicable. 26
Figure 1.7 Maternal (mtDNA) and paternal (Y chromosome) patterns of inheritance. 27
Figure 2.1 A flow chart showing student involvement in different parts of the workflow. 36
Figure 3.1 An electropherogram showing STR loci (green blocks)and alleles (represented as peaks under the STR loci), observed when VeriFiler Express PCR amplification kit was used. 49
Figure 3.2 A chart showing the number of alleles observed per STR locus across the different ethnic groups included in the present study. 51
Figure 3.3 Chart showing the number of observed genotypes per locus, across the five population groups included in the present study. 54
Figure 3.4 Chart showing observed heterozygosity per STR locus, across the five population groups included in the present study. 57
Figure 3.5 Chart showing the Probability of Identity for all the tested STR loci, per population group included in the present study. 61
Figure 3.6 A chart showing Polymorphic information content values for all the tested STR loci, per population group included in the present study 64

LIST OF TABLES

Table 1.1 Commercially available human identification kits 8
Table 1.2 Mutation rate frequencies 11
Table 1.3 Inheritance coefficients for common biological relationships 14
Table 1.4 Formulae for calculating Likelihood Ratios for a paternity trio. 16
Table 1.5 Formulae for calculating Likelihood Ratios for kinships (sibships).....21
Table 1.6 Number of kinship requests processed by the DPL from 2014 to 2019.....23
Table 1.7 The probability of identity (PI) of 6 different DNA testing commercial kits and the kit prices 31
Table 2.1 Summary of the different population groups included in the study according to geographic location and self-reported ethnicity 34
Table 2.2 Summary of loci included in the commercial STR kits evaluated in the current study.....38
Table 2.3 Scenarios of family relationships used in the evaluation of the utility of additional STR loci included in the expanded commercial kits and in the validation of the In-house data analysis software 45
Table 2.4 Capabilities of the different data analysis software for calculating Likelihood Ratios for kinship testing 47
Table 3.1 Calculated combined Probability of Identity calculated for the five ethnic groups when different STR kits were used. 58
Table 3.2 Ranking of STR loci contained in the commercial kits included in the present study.....66
Table 3.3 Assessment of Hardy-Weinberg equilibrium using the exact *p*-values by the Markov chain method 67
Table 3.4 Global Hardy-Weinberg tests- results by population, when H1= heterozygote deficit and when H2= heterozygote excess (multi-locus test).....68
Table 3.5 Pairwise F_{st} values between population groups for all loci.....69
Table 3.6 Comparison of outcomes when 14 and 23 polymorphic STR loci were used .70

Table 3.7 *P*-values for Student's T-test conducted to compare Likelihood Ratio outcomes obtained using four different data analysis software packages 71
Table 7.1 An example of a sample plate map for loading on the genetic anlyser 106
Table 7.2 Population allele frequencies for the SA Bantu-speakers (SAFBS), SA Coloureds (SAFMX), SA Whites (SAFWH), Zimbabwean Bantu-speakers (ZIMBS) and Nigerian Bantu-speakers (NGRBS), grouped per STR locus. 107
Table 7.3 *P-values* for pairwise analysis of all the STR loci for the SA Bantu-speakers 116
Table 7.4 *P-values* for pairwise analysis of loci all the STR for the SA Whites.....120
Table 7.5 *P*-values for pairwise analysis of all the STR loci for the SA Coloureds 124
Table 7.6 *P*-values for pairwise analysis of all the STR loci for the Zimbabwean Bantu-speakers.....128
Table 7.7 *P*-values for pairwise analysis of all the STR loci for the Nigerian Bantu-speakers.....131
Table 7.8 Formulae for calculating Likelihood Ratios for a paternity trio. 136
Table 7.9 Formulae for calculating Likelihood Ratios for kinships (sibships).....137
Table 7.10 Scenarios of relationships used in the validation of the In-house data analysis software.....139
Table 7.11 Capabilities of the different data analysis software for calculating Likelihood Ratios for kinship testing 140
Table 7.12 Paired two sample T-test conducted to compare outcomes obtained using four different data analysis software packages.....142

LIST OF ABBREVIATIONS

˚C	Degrees Celsius
µl	Microliter
AABB	American Association of Blood Bank
ANOVA	Analysis of variance
CE	Capillary electrophoresis
CLR	Combined Likelihood Ratio
CODIS	Combined DNA Index System
DBA	DNA Advisory Board of Quality Assurance Standards
DBS	Dry blood spots
DHA	Department of Home Affairs
DNA	Deoxyribonucleic acid
DPL	DNA Profiling Laboratory
EDTA	Ethylene diamine tetra-acetic acid
ESS	European Standard Set
FBI	Federal Bureau of Investigation
F_{st}	F-statistic used to measure correlation of alleles within the same subpopulation relative to the entire population
G_{obs}	Number of genotypes observed
H_0	Null hypothesis
H_1	Alternative hypothesis
H_2O	Water
HCl	Hydrochloric acid
HID	Human Identification
H_{obs}	Observed heterozygosity
HREC	Human Research Ethics Committee
HWE	Hardy-Weinberg equilibrium
IBD	Identity by descent
IBS	Identity by state

ILS	Internal lane standard
ISFG	International Society of Forensic Genetics
k	Coefficient of inheritance
LD	Linkage disequilibrium
LE	Linkage equilibrium
LR	Likelihood Ratio
MAF	Minimum allele frequency
Mb	Mega bases
ml	millilitres
mm	millimetres
MPS	Massive parallel sequencing
mtDNA	Mitochondrial DNA
NAMBS	Namibian Bantu speakers
ng/ µl	Nano gram per microliter
NGRBS	Nigerian Bantu speakers
NGS	Next Generation sequencing
NIST	National institute for Standards and Technology
P	Prior odds
P (%)	Probability of relatedness
p-arm	Short arm of a chromosome
PCR	Polymerase chain reaction
PE	Power of exclusion
PI	Paternity index
PI	Probability of identity
PIC	Polymorphic Information Content
PO	Posterior odds
Pr	Prior probability
q-arm	Long arm of a chromosome
RFLP	Restriction fragment length polymorphism
RI	Residual index

SA	South Africa(n)
SAFBS	South African Bantu-speakers
SAFIN	South African Indians
SAFMX	South African mixed ancestry (Coloureds)
SAFWH	South African Whites
SEB	South Eastern Bantu
SEV	Standard Elution volume
Stats	Statistics
STR	Short tandem repeat
SWB	South West Bantu
SWGDAM	Scientific Working Group on DNA Analysis Methods
TE buffer	Tris EDTA buffer
US	United States
VNTR	Variable number tandem repeat
ZIMBS	Zimbabwean Bantu-speakers

CHAPTER 1 INTRODUCTION

1.1 An overview of the South African population groups

According to Statistics South Africa (Stats SA), the population of South Africa (SA) was estimated at 57.7 million as at the 1st July 2018 (Statistics South Africa, Republic of South Africa, 2018). According to Stats SA, the main causes of changes in the population composition are births, deaths and migration. A report from Stats SA showed that South Africa is estimated to receive a net immigration of 1,02 million people between 2016 and 2021. The number of people in SA, per population group, as per the 2011 census was as follows: 79.2% Black Africans, 8.9% Coloureds, 2.5% Indians/Asians, 8.9% Whites and 0.5% Other (Statistics South Africa, not dated). The present study seeks to establish different population group specific statistical data to be utilized for human identification applications.

The following population groups were thus included in the present study: South African Bantu-speakers (made up of the Zulu, Xhosa, Ndebele, Northern Sotho, Tswana, Venda and Tsonga ethnic groups), SA Whites, SA Coloureds, SA Indians, Zimbabwean Bantu-speakers, Namibian Bantu-speakers and Nigerian Bantu-speakers. Bantu is the parent language in sub-Saharan Africa (Grollemund et al., 2015). A study previously done in the same department as the present study illustrated that the SA Bantu-speakers can be clustered together, and the same population specific allele frequencies utilised, when performing human identification testing, particularly with regards to family relationship testing (Lane et al., 2002).

The Bantu-speaking groups included in the present study migrated from the West Central Africa approximately 5000 years ago, to eventually settle in the southern part of Africa, forming two distinct clusters: The South Eastern Bantu (SEB) and the South Western (SWB) groups (Grollemund et al., 2015). The migration of the Nigerian Bantu-speakers, took place much later and continues to increase as documented in the Stats SA census findings of 2018. Figure 1.1 shows the migration routes of the Bantu-speakers.

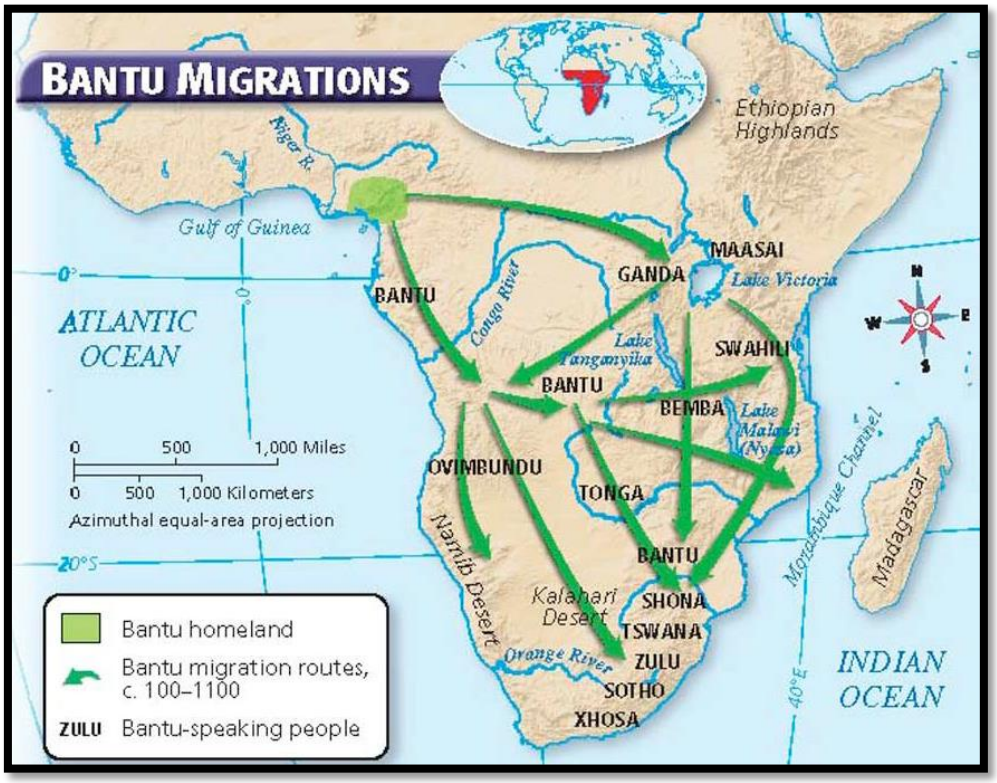


Figure 1.1 Bantu migrations

The Bantu-speaking population groups originally lived in West Central Africa, on the border of modern Nigeria and Cameroon. Evidence suggests that the migration of the Bantu people spread out toward the east (forming the South Eastern Group) and the west (forming the South Western Group in the Namib Desert) through a series of migrations over time.
(https://www.longbranch.k12.nj.us/cms/lib/NJ01001766/Centricity/Domain/635/Bantu.pdf)

1.2 Human identification testing using DNA

Deoxyribonucleic acid (DNA) based human identification (HID) is a way of uniquely identifying a person using DNA extracted from white blood cells obtained from traces of tissues including blood, buccal cells, tissue biopsies, hair, saliva, teeth, bones, semen, amniotic fluid or chorionic villus. Biological relatedness between individuals can be established by using molecular methods such as DNA profiling (Chakraborty and Jin, 1983). For more than three decades, DNA profiling has been used worldwide for HID purposes. This technique was developed by Sir Alec Jeffreys in 1985, when he showed that variable number tandem repeats (VNTRs) could be interrogated by restriction fragment length polymorphism (RFLPs) analysis of DNA to demonstrate genetic linkages between individuals (McKiernan and Danielson, 2017). Based on genetic profiling, a DNA fingerprint, a similar identifier to an actual fingerprint, could be produced, which is unique for each individual. DNA profiling is used worldwide in parentage, kinship and forensic investigations. Prior to the invention of DNA profiling, forensic casework was performed using protein biomarkers (McKeirnan and Danielson, 2017). However, these had a limited number of possible genotypes at each locus, resulting in low marker variability (Chakraborty and Jin, 1983). The field of forensic science has evolved in the past decades, as evidenced by advancements in the techniques currently used that are illustrated in Figure 1.2.

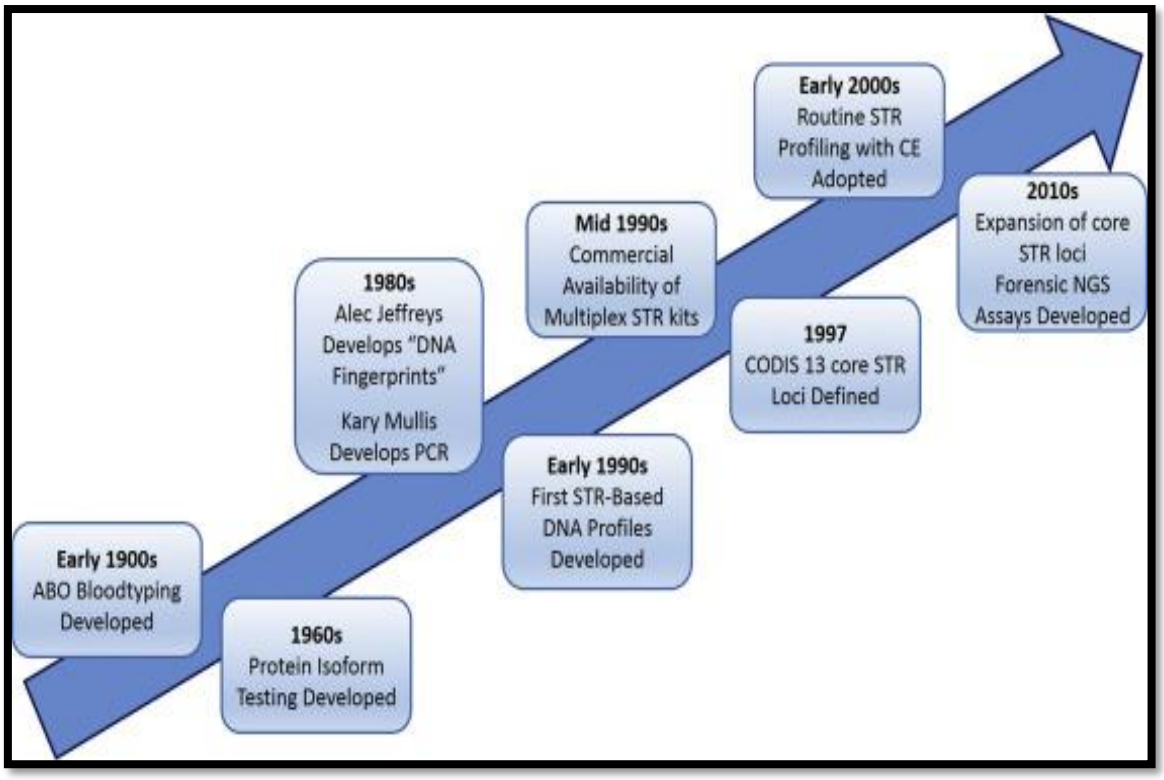


Figure 1.2 An overview of advances experienced in human identification testing.

Prior to the advent of DNA profiling, human identification using biological markers for forensic casework was performed using protein biomarkers (McKeirnan and Danielson, 2017). This was followed by the introduction of DNA fingerprinting using RFLP. In the early 1990s, the first short tandem repeat (STR)-based DNA profiles were developed. Since then, efforts have been focused on improving the multiplex STR kits available, by increasing the number of STR loci and also improving sample throughput. [CODIS-combined DNA Index System; NGS-next generation sequencing; PCR-polymerase chain reaction, CE-capillary electrophoresis.] Image reproduced, with permission (license number 4795851263624) from McKiernan and Danielson, 2017.

DNA profiling, involves several processes in which an individual's DNA is extracted from a tissue source, followed by polymerase chain reaction (PCR) using fluorescently labelled primers to amplify selected target regions of the genome. Then capillary electrophoresis (CE) is performed to characterize alleles, based on their sizes, inferred from the known sizes on the internal lane standard (ILS). Allele sizing is based on the migration time a DNA fragment takes to pass through the laser. The use of different

fluorescent dyes helps to distinguish alleles with overlapping size ranges. The composite information product is a DNA profile depicting all the alleles and their sizes, which is unique per individual. The DNA profile obtained can be used for a variety of downstream applications including direct identification, such as in cases of missing person investigations, forensic casework and for establishing biological relationships such as parentage and kinship testing (Carracedo et al, 2008).

1.3 DNA profiling using short tandem repeats

Short tandem repeats (STRs) are tandem (head to tail) arrays of short repeated sequences involving repetitive units of one to six base pairs, forming a series with resulting lengths of up to 100 nucleotides (Thompson, 1999). They are widely dispersed throughout the human genome, and account for approximately 3% of the entire genome (Fan and Chu, 2007). Most STRs are found in the non-coding regions, with only 8% located in the coding regions. The use STRs in DNA profiling is based on detection of copy number variations of repeat sequences at a locus, and has been a common practice in forensic science since 1993 (Thompson, 1999). Alleles are therefore distinguished by the length of a DNA fragment that contains different numbers of copies of the repeat unit at that locus, following PCR (Chakraborty and Jin, 1983).

STRs can be classified into different types based on the length of the repeat unit or the repeat structure. With regards to length of the repeat unit, STRs are classified into mono-, di-, tri, tetra, penta- and hexa-nucleotide repeats (Fan and Chu, 2007). The most common are dinucleotides. However, for human identification applications, tetra-nucleotide repeats are preferred as they can be amplified using PCR with greater efficiency than dinucleotides (National Institute of standards and Technology, 2007). PCR fragment sizing can more reliably be interpreted in terms of number of repeats present at the locus. The repeat structures can be simple, containing repeat units of

identical length and sequence, for example, $[AATC]_5$ is a simple tetra-nucleotide repeat structure. Compound repeats comprise two or more different repeat units, and complex repeats consist of complex repeat structures, with hyper-variability in the repeat units, length and sequence (Fan and Chu, 2007). Figure 1.3 shows an example of different number of repeats at a locus, based on the length and number of the repeat units present. The repeat region is variable between different people, but the flanking regions are the same, hence primers are designed to bind within the flanking regions.

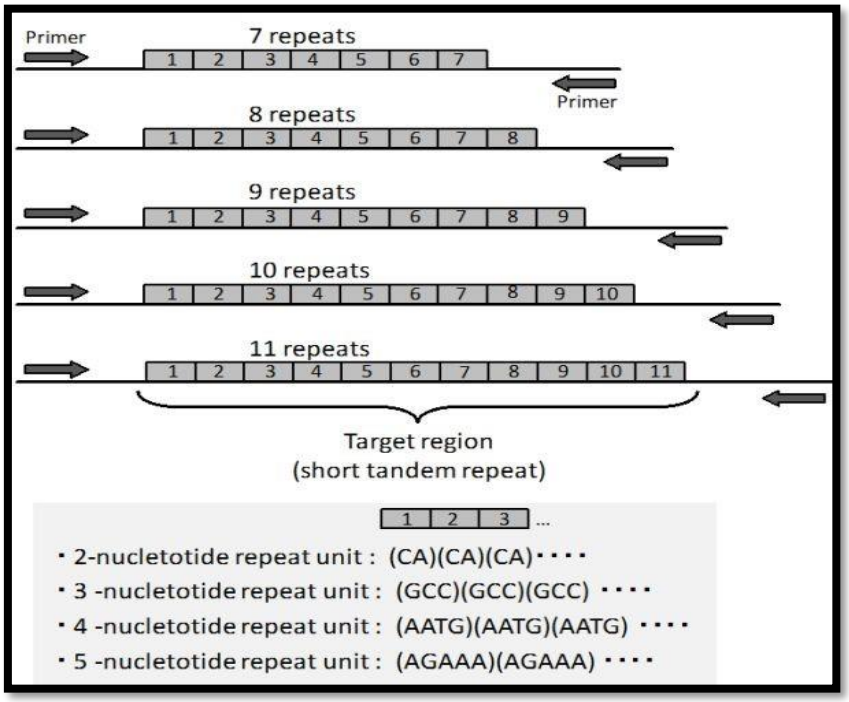


Figure 1.3 Showing the different number of repeats at an STR locus, and the flanking regions where the primers bind.

The repeat structures can be di-, tri, tetra or hexa-nucleotides, with simple repeat sequences. The flanking region where the primer binds is the same between different individuals. However, the target regions (allele sizes) are different.
https://www.researchgate.net/publication/221912832_DNA_biometrics/figures?lo=1 Image available under a Creative Commons licence.

STRs exhibit variability between individuals and thus can be utilised as a tool when undertaking human identification. The following features are desirable for STRs: high heterozygosity, high number of distinguishable alleles, high number of observed genotypes, regular repeat units and high fidelity amplification (National Institute of standards and Technology, 2007) . All these features help to increase the genetic variability of STR loci between individuals and result in a high degree of discrimination when multiple STR loci are examined. When a multi-locus commercial kit is used, a DNA profile is generated – this is a presentation of copy number variations at a number of loci tested, which correspond to genotypes observed per locus.

1.4 Commercially available STR based kits used in human identification

In 1997, the Federal Bureau of investigations (FBI) established the Combined DNA Index System (CODIS), which was composed of 13 autosomal STRs (namely TPOX, D3S1358, FGA, D5S818, CSF1PO, D7S820, D8S1179, TH01, vWA, D13S317, D16S539, D18S51 and D21S11), in addition to *Amelogenin*, a sex determination locus (Butler and Hill, 2013). Subsequently, in 2001, the European Standard Set (ESS) was established, which comprised seven autosomal STR loci, namely, D3S1358, FGA, TH01, vWA, D8S1179, D18S51 and D21S11 (Corte-Real, 2004). These CODIS core and ESS loci were primarily used in forensic science to facilitate searches in state and national databases (as discussed in O' Connor et al., 2010), and became the standard sets for performing human identification testing worldwide. The most commonly used STR based commercial kits include these loci, even though there are other commercial kits available containing different STR loci.

Presently, the three main manufacturers for human identification commercial kits are Promega Corporation, Life Technologies (formerly known as Applied Biosystems) and Qiagen. These commercial kits are manufactured following the Scientific Working Group on DNA Analysis Methods (SWGDAM), American Association of Blood Bank (AABB) guidelines and International Society of Forensic Genetics (ISFG) recommendations (Gjertson et al, 2007). Some of the commonly used commercial kits are shown in Table 1.1. The SWGDAM guidelines stipulate the criteria for accepted quality of the commercial kits produced, the validation process, as well as the robustness of the analysis software used to analyse the results obtained after electrophoresis of amplified products (SWGDAM, 2010). Commercial testing laboratories are expected to establish inclusion and exclusion criteria based on the relationships tested and the outcomes obtained.

Table 1.1 Commercially available human identification kits

Commercial kit name	No. of Loci	Manufacturer
CODIS 13		
PowerPlex® 16/ PowerPlex® 16HS	16	Promega
PowerPlex® 18D	18	Promega
PowerPlex® ESI 16 / ESX 16 / NGM™	16	Promega
PowerPlex® ESI 17 / ESX 17 / NGM SElect™	17	Promega
IdentiFiler®/ IdentiFiler® Plus	16	Life Technologies
CODIS 20		
GlobalFiler™	25	Life Technologies
PowerPlex® Fusion	24	Promega
PowerPlex® 21	21	Promega
VeriFiler™ Express	25	Life Technologies
Investigator 24plex GO! Kit	24	Qiagen
Investigator 24plex QS Kit	24	Qiagen
Investigator 26plex QS Kit	26	Qiagen
Others*		
ESS 12	12	Promega
SGM Plus®	11	Promega
SEfiler Plus™	12	Life Technologies
COfiler®	7	Life Technologies
Profiler Plus™ (Discontinued since 2018)	10	Life Technologies
PowerPlex® CS7	7	Promega

*Commercial kits with fewer STR loci are usually used as supplementary kits or used for research purposes.

A panel of 13 markers has a 99% inclusion power and 100% exclusion power for parentage testing (Pu and Linacre, 2007). Too few loci tested may lead to results whereby multiple people share the same DNA profile, by chance. The commercial kits with fewer STR loci are usually used to supplement the ones with 13 and more loci. Over the past few years, there was a call to expand the number of loci contained in the CODIS 13 and ESS. Since then, the FBI has expanded its core set of 13 autosomal STRs to 18, and the ESS core set has increased from seven to 12 loci (Butler and Hill, 2013). Expanded commercial kits have been manufactured by Promega Corporation, Life Technologies and Qiagen to meet these requirements (see kits listed under CODIS 20 in Table 1.1).

1.5 Mutation rate for STRs

A mutation at an STR locus is defined as the observance of one or more genetic inconsistencies (mismatches) between two individuals who are biologically related, such as parent and offspring, who, under normal circumstances, should show a match at all loci tested. The average rate of STR mutation frequency is 0.001 (O' Connor et al., 2010), and has been reported to be locus and gender specific. Mutations result from DNA changes at particular loci resulting in the observance of allele inconsistencies from one generation to the next. There are three possible causes of mutations, namely 1) unequal cross over during meiosis, 2) retro transposition and 3) strand slippage during replication (Fan and Chu, 2007). Of the three, strand slippage has been associated with common mutations in STRs. An illustration of strand slippage is shown in Figure 1.4.

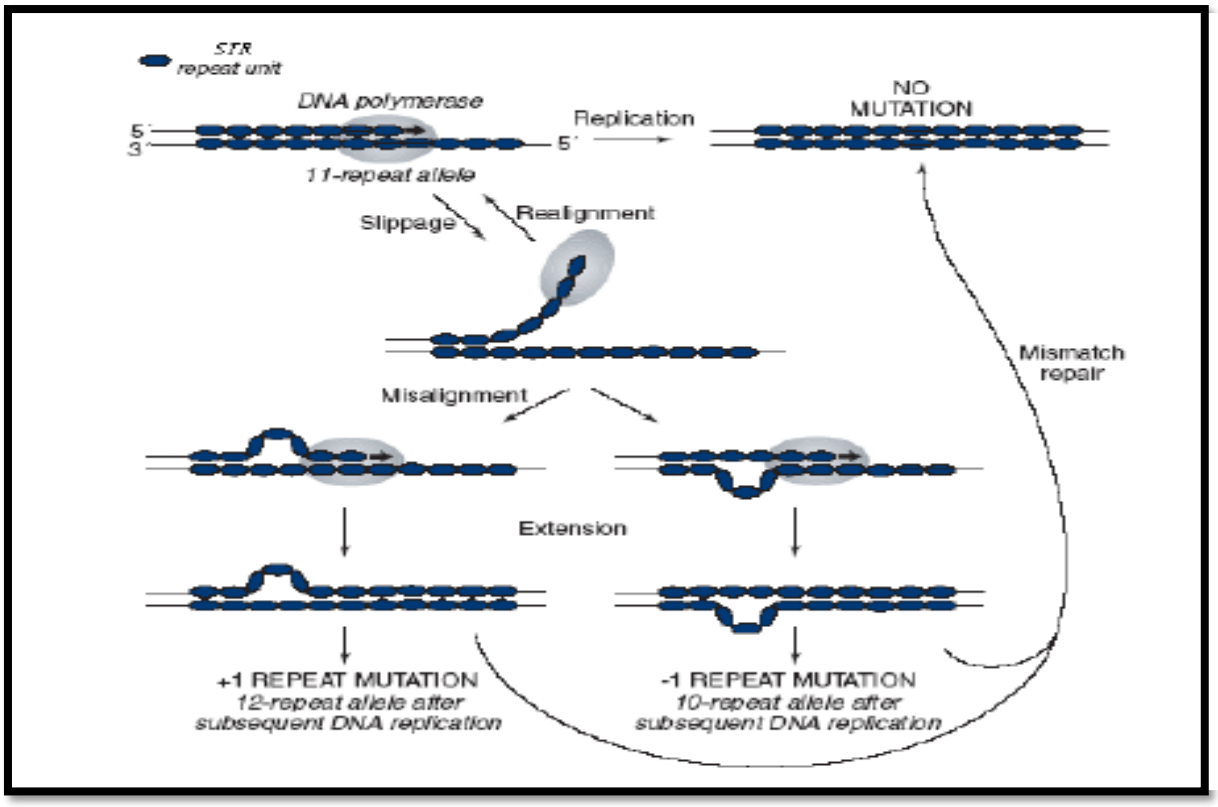


Figure 1.4 Mutation observed as a result of strand slippage.

The 11 repeat allele mutates to a 12 repeat allele following a +1 repeat mutation or a 10 repeat allele following a -1 repeat mutation (Fan and Chu, 2007). Image available under a Creative Commons licence.

Great variation in mutation rates has been observed among different loci (discussed in Chakraborty et al., 1993). More gains, of one or two complete repeat units, have been observed than losses (Dauber et al., 2003). Also, Dauber et al., 2003 have reported that more mutations have been observed in male than female germlines and that higher mutation rates have been observed in loci with longer uninterrupted repeats. Table 1.2 presents a list of STR loci contained in the VeriFiler™ Express kit and their mutation rates. As can be seen from Table 1.2, mutation rates in males are generally higher (up to tenfold) as compared to females, although there are exceptions to this rule.

Table 1.2 Mutation rate frequencies

STR locus	Paternal mutation rate	Maternal mutation rate
D3S1358	0.001691	0.000211
vWA	0.003258	0.000494
D16S539	0.001127	0.000481
CSF1PO	0.002021	0.000283
TPOX	0.000130	0.000081
D8S1179	0.002031	0.000333
D21S11	0.001709	0.001295
D18S51	0.002530	0.000748
Penta_E	0.000260	<0.000253
D2S441	0.000898	0.000069
D19S433	0.000745	0.000596
TH01	0.000070	0.000043
FGA	0.003713	0.000522
D22S1045	0.000216	0.000084
D5S818	0.001742	0.000300
D13S317	0.001743	0.000436
D7S820	0.001348	0.000073
D6S1043	0.001200	0.000400
D10S1248	0.000637	0.000250
D1S1656	0.001144	0.000330
D12S391	0.002900	0.000837
D2S1338	0.001526	0.000245
Penta_D	0.000259	<0.000253

Paternal and maternal mutation rates of STR loci contained in the VeriFiler™ Express PCR amplification kit (The Relationship Testing Program Unit, 2008). The mutation rates for the STR loci indicated in bold are from (Li et al., 2011), with the exception of the D6S1043 locus, sourced from (Gaviria et al., 2017).

1.6 Parentage testing

In parentage testing, DNA profiling is performed on a child and its parent(s). The aim of this test is to establish if the alleged parent, usually the alleged father, is indeed the biological parent of the child in question or if a random man is the biological father. The resulting genotypes may be assessed by noting if the individuals tested share at least one allele by descent at all loci tested. This assessment is based on the fact that an individual receives half of his/her nuclear DNA from the mother and the other half from

the father. This is determined at conception when haploid gametes fuse - sperm from the father fertilizes the mother's egg. The resulting diploid zygote contains a unique set of DNA harbouring information from the two parental lines and their ancestors going back many generations. Consequently, with the exception of identical twins, no two individuals share the identical combination of DNA sequences. A simplified pedigree illustrating how parents pass on STRs to their off-springs is illustrated in Figure 1.5.

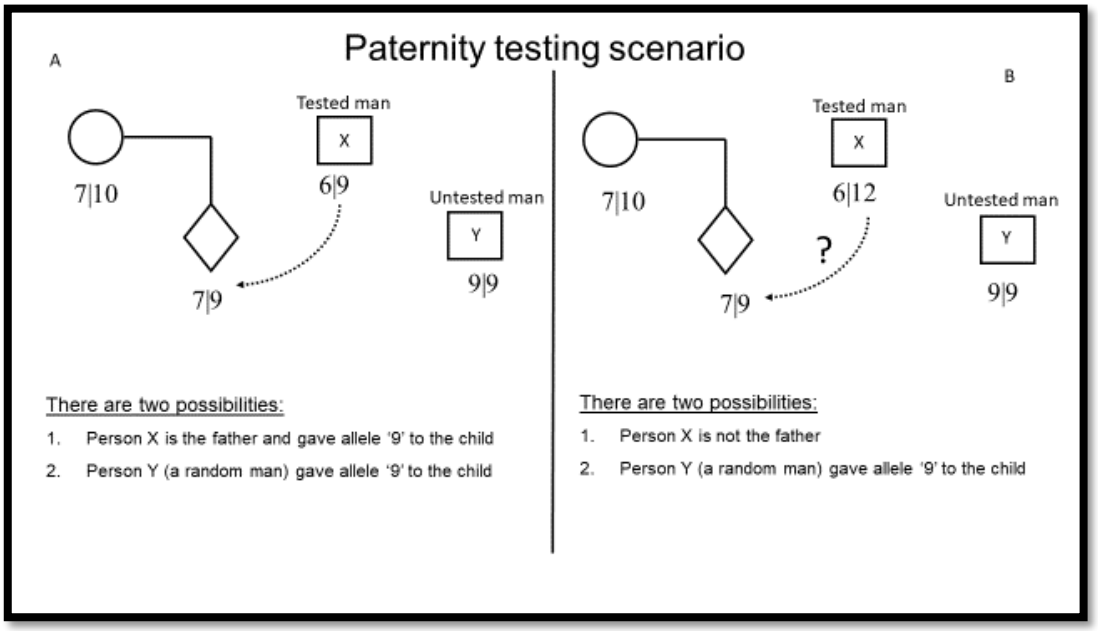


Figure 1.5 Paternity testing outcome illustration.

The tested man (X) may be excluded or not excluded as being the biological father of the child. Circles, females; squares, males; diamond, unknown sex. The numbers under the shapes represent alleles. Parents pass on 50% of their alleles to their offspring. In A, the child has inherited allele 7 from the mother and so the biological father must possess the obligatory allele 9. In B, the tested man does not have the obligatory allele that the child must have received from its biological father.

In a paternity query, the alleged father may be “not excluded”, by showing that he passed on the obligatory allele that the child received from its biological father, or he may be “excluded” when there are three and more mismatches between himself and the child. The tested man in Figure 1.5 has allele 9, but it could be by chance. The sharing of one allele must be demonstrated across a battery of STR loci tested to confirm a biological relationship. If three or more mismatches are observed in parentage testing, then the alleged mother/father is not the biological parent, a random untested individual is the biological parent.

1.7 Statistical calculations

All statistical calculations used to establish relatedness between individuals are based on Likelihood Ratio principles recommended by the International Society for Forensic Genetics (ISFG) (Gjertson et al. 2007). The DNA profiles obtained from any set of polymorphic STR loci (based on the commercial kit used) from different individuals may be analysed by assessing if the individuals share zero, one or two alleles at a locus, taking into account the number of independently segregating loci tested. Table 1.3 shows inheritance coefficients (k_0, k_1 and k_2) for common biological relationships between individuals.

Any biological relationship between two individuals can thus be defined by the following relationship between the coefficients: $k_0 + k_1 + k_2 = 1$ (Chakraborty and Jin, 1983). Table 1.3 shows the expected coefficients between relationships for any pair of genotypes, when individuals share zero, one or two alleles. Half siblings, uncle/aunt-niece and grandparent-grandchild relationships all share on average 25 % of their DNA and belong to the same equivalent class of kinships, exhibiting $k_0 = 1/2$, $k_1 = 1/2$ and $k_2 = 0$. These relationships will always be difficult to distinguish from genotype data alone. Using this

approach and based on genotypic data, some relationships such as parent-offspring pair, unrelated and whether the genotypes represent data from the same individual (or from monozygotic twins) yield a high discrimination power and therefore easy to establish, even in single parent (motherless/fatherless) parentage analysis (Chakraborty and Jin, 1993). This approach implies that a Likelihood Ratio (LR) is calculated based on allele sharing between individuals. For relationships sharing less DNA (i.e. distant biological relationships) the proportion of allele sharing is reduced, making it difficult to establish true relationships based on genotypic data only.

Table 1.3 Inheritance coefficients for common biological relationships

		Expected		
Relationship	**Symbol**	K_0	K_1	K_2
Identical twins (monozygotic)	MZ	0	0	1
Parent-offspring	P-O	0	1	0
Full siblings	FS	¼	½	¼
Half siblings/ uncle or aunt-niece/ Grandparent/grandchild	HS/Un-N/Au-N/GP-GC	½	½	0
First cousins	1C	¾	¼	0
Second cousins	2C	$^{15}/_{16}$	$1/16$	0
Unrelated	U	1	0	0

Likelihood Ratios indicate the likelihood of obtaining the observed genotypes under two mutually exclusive hypotheses (Sozer et al., 2010). The first hypothesis is that the tested individuals are related as indicated in the pedigree showing the proposed relationship. The second hypothesis is that the tested individuals are unrelated, and assumes that the

pedigree provided is not true. A relationship index (RI) is calculated for each of the tested STR loci. A combined relationship index (CRI), also called a combined Likelihood Ratio (CLR) is calculated by multiplying relationship indices from all the independently segregating STR loci, as follows:

$$CLR = LR (STR_1) \times LR(STR_2) \times \ldots LR (STR_n).$$

If genetically linked loci are used, the haplotype frequencies must be used instead of STR individual LRs for the affected markers (O' Connor et al., 2011; Sozer et al., 2010).

Establishing the LR for parentage testing is based on the calculation of the paternity Indices (PIs). A Paternity Index calculates the odds in favour of the proposed relationship, given the genotypes of the people tested (Gjertson et al., 2007). The following equation is used to calculate the PI:

PI = X/Y, where

- PI is the Paternity Index per locus tested,
- X represents the probability of paternity (that the tested man is the father) - based on the alleles observed and their frequency in the specific population group and
- Y represents the probability of non-paternity (that the tested man is unrelated to the child/ a random man is the father) (Gjertson et al., 2007).

The Combined Paternity Index (CPI) is calculated by multiplying paternity indices (PIs) obtained for each locus tested, if the population is in Hardy Weinberg equilibrium. Large CPI values (>100), support the proposed relationship (Gjertson et al., 2007).

Statistical calculations take the following considerations into account:

i) The alleged person's ethnic group- how common is this DNA profile in that ethnic group.
ii) Allele frequencies for the STRs tested, specific to the tested individual's ethnic group,
iii) Commercial kit used (number of STR loci tested and their mutation rates).

The formulae for calculating Likelihood Ratios for parentage testing are indicated in Table 1.4. The LR can be calculated manually, however, it is advisable that kinship analysis software be utilized to prevent human error and to streamline workflow in testing laboratories (Drábek, 2009).

Table 1.4 Formulae for calculating Likelihood Ratios for a paternity trio.

Mother	Child	Alleged father	Formula	Mother	Child	Alleged father	Formula
A,A	A,A	A,B	$1/2a^2$	A,A	A,A	A,A	$1/a^2$
A,A	A,B	A,B	1/4ab	A,B	A,A	A,A	$1/2a^2$
A,A	A,B	B,C	1/4ab	B,B	A,B	A,A	1/2ab
A,B	A,A	A,B	$1/4a^2$	B,C	A,B	A,A	1/4ab
A,B	A,A	A,C	$1/4a^2$	A,B	A,B	A,C	1/8ab
B,C	A,B	A,B	1/8ab	A,B	A,B	A,A	1/4ab
B,C	A,B	A,C	1/8ab	A,B	A,B	A,B	1/4ab
B,D	A,B	A,C	1/8ab				

Mother, child and alleged father are tested. The formulae are based on the inheritance pattern. A, B, C and D depict alleles obtained at a particular locus. a and b represent the allele frequencies of alleles A and B. Profiles are denoted by genotypes. Adapted from (Sozer et al., 2010)

Bayes Theorem

The Bayes Theorem specifies the need to combine information from genotypes and non-DNA data. The non-DNA data is taken into account in the form of prior odds (P), prior probability (Pr), posterior odds (PO) and probability of relatedness [P (%)]. (Sozer et al., 2010). Prior probability is the proposed strength of the evidence that individuals are related as indicated on the pedigree based on non-DNA evidence. For parentage testing, the Pr is mostly 0.5, meaning that there is a 50% probability that the tested people are related and 50% probability that they are not related. Prior odds are calculated as follows:

Prior odds= Prior probability (Pr)/(1-Pr)

The posterior odds give a numerical weight to the opinion of relatedness, and is calculated as follows:

PO = (combined Likelihood Ratio) x (prior odds)

The probability of a relationship, also known as the posterior probability, expresses the Combined Likelihood Ratio (CLR) as a percentage and is calculated as follows:

P (%)= CLR/1+CLR.

The inclusion criteria for a paternity test that is admissible in court, is defined as a Combined Paternity Index (CPI) above 100 and a P (%) above 99%.

1.8 Population specific allele frequencies

Local population databases with allele frequency data indicating the common and rare alleles present within the different ethnic groups should be generated. This information helps to determine alleles and allele frequency averages in the population groups, since the entire population group can obviously not be sampled to obtain a more accurate representation of alleles present. This information is used to make statistical inferences on biological relationships and prior calculation and data-basing of population specific allele frequencies helps to avoid under- or over-estimating the calculated statistical evidence in DNA testing. Ideally, population specific allele frequencies must be generated per commercial kit, as the primers used in each kit are different. This may result in certain alleles not been amplified at all in a given population group, if there is a mutation in the primer's binding site, resulting in partial or non-amplification (null alleles). The use of inappropriate population data may lead to misrepresentation of statistical evidence. The same profile assessed using different population groups' allele frequencies, will show different matching probabilities as the common and rare alleles are different across population groups. If the ethnic group of an individual is not indicated, one may use a range of population data to calculate a probability match and report on the lowest outcome, as this conservative approach is preferred over inflating statistical evidence.

When calculating allele frequencies, the following assumptions are taken into account: 1) the organisms tested are diploid, 2) reproduction is sexual, 3) generations are non-overlapping, 4) the genes under consideration have two alleles, 5) alleles are identical in males and females, 6) individuals of all genotypes have equal rates of survival and equal reproductive success (i.e. no selection) (Butler, 2005). Hardy Weinberg equilibrium (HWE) test is a test which is used to assess the evolutionary changes in a population. This test determines if the allele frequencies, and consequently, genotype frequencies in a population remain constant over time (Waples, 2015). HWE assumes

that the following processes are not taking place in the population: 1) No mutations, 2) no selection, 3) no migration, 4) random mating takes place and 5) that the population is infinitely large (discussed in Waples, 2015). Although natural populations violate HWE, this does not affect the statistics calculated based on population allele frequencies since an allowance of a certain level of uncertainty is taken into account. However, three factors that result in major departures from HWE are 1) inbreeding within a population (i.e. the presence of a high number of related individuals in a population that are mating), 2) selection and 3) genotyping errors which result in loss of certain genotypes (Butler, 2005). Alleles are assumed to be in linkage equilibrium when recombination takes place between the STR loci. This concept gives an indication of calculating statistical evidence without bias.

Evolutionary forces that manipulate allele frequencies in a population are: 1) mating patterns, 2) migration, 3) mutations, 4) recombination between genes, 5) selection and 6) genetic drift. These forces can introduce new alleles previously not seen in a population group and may cause the common frequency to increase or decrease. Heterozygosity deficiency and heterozygosity excess may be observed where population bottlenecks occur. Heterozygosity excess occurs when the observed heterozygosity in a population sample is larger than the expected heterozygosity when a population is assumed to be in HWE (Conuet and Luikart, 1996). Heterozygosity deficiency then occurs when the observed heterozygosity in a population sample is smaller than the expected heterozygosity.

Statistical calculations performed in human identification testing assumes that populations are in HWE and that linkage disequilibrium (LD) (non-random association of STR loci) exists between the STR loci used. The calculations for establishing biological relatedness are based on 1) the algorithm of Identity by descent (IBD), 2) inheritance pattern and 3) the allele frequencies observed in specific population groups (Cho et al.,

2017). IBD is defined as two or more individuals sharing the same alleles at given loci, inherited from the common ancestor, versus sharing alleles by chance because the alleles are common in the population, which is identity by state (IBS) (Thompson, 2013).

Genetic variation between population groups

Genetic variation between populations can be studied by comparing genotypic similarities within and among population groups based on the average number of shared alleles and their frequencies (Chakraborty and Jin, 1993). Wright's F-statistics, particularly F_{st} estimates, are among the widely used descriptive statistics that provide important insights into the processes that influence genetic variation within and among population groups (Holsinger and Weir, 2009). F_{st} measures the similarities of allele frequencies within the same subpopulation relative to the entire population. It measures the similarities and correlation of alleles within population groups and the degree of variance observed between different population groups (Holsinger and Weir, 2009). Differentiations between populations are said to be caused by genetic drift and selection over time. Establishing genetic variation between population groups provides statistical information with regards to how similar or how different the population groups are from each other. When populations are shown to be similar, clustering can be inferred, and the same allele frequency data can be used in Likelihood Ratio calculations. F_{st} values range from 0 to 1. A value of 0 means that the two population groups are similar (same alleles are present in both populations at the same frequency). An F_{st} value of 1 means that the two population group are different (there is no overlap between the alleles and their frequencies). According to Rosenberg et al., 2002, differences among major population groups constitute 3 to 5%, whereas within population differences account for 95 to 97%.

1.9 Kinship testing

Another type of testing that is performed routinely in DPL is kinship testing. This type of testing is performed mainly in applications of genetics to social issues such as change of surname, immigration, estate settlements and applications for identity documents or unabridged birth certificates. For many, individuals require this testing in the absence of their parents and need to be tested with relatives, such as siblings, half siblings, aunts/uncles and grand-parents (Gjertson et al., 2007). In immigration applications, for instance, kinship testing is required to verify the biological relationship between the prospective immigrant and his/her sponsor, based on National immigration laws (Karlsson et al., 2007). Statistical calculations for kinship analysis are more complex than those used for paternity testing and these calculations are usually done using a software package which takes into account different permutations to calculate CLRs (Gjertson et al., 2007). Table 1.5 shows formulae used when performing kinship analysis between alleged full and half siblings.

Table 1.5 Formulae for calculating Likelihood Ratios for kinships (sibships)

Sibling	Possible full sibling	Formula	Half sibling	Possible half sibling	Formula
A,B	A,B	(1+a+b+2ab)/8ab	A,B	A,B	(a+b+4a)/8ab
A,A	A,A	$(1+a)^2/(2a^2)$	A,A	A,A	(1+a)/2a
A,A	A,B	(1+a)/4a	A,A	A,B	(1+2a)/4a
A,B	A,C	(1+2a)/8a	A,B	A,C	(1+4a)/8a
A,B	C,D	¼	A,B	C,D	½

Formulae for calculating Likelihood Ratios when siblings, full and half, are tested. The formulae are based on the inheritance pattern. A, B, C and D depict alleles obtained at a particular locus. a and b represent the allele frequencies of alleles A and B. Profiles are denoted by genotypes. Adapted from (Sozer et al., 2010)

The formulae for half siblings will be the same as those used for testing relationships between uncle/aunt- niece/nephew and grandparent-grandchild as they all share the same coefficient scores (k_0 = 1/2, k_1= 1/2 and k_2= 0). The question in this type of testing is not, are these people related but rather, does the genetic evidence support the proposed relationship? It is therefore important to have a family pedigree indicating the proposed relationship before testing. Calculating LRs, indicates how much the genotypes obtained from the people tested support the proposed relationship versus the alternative hypothesis, and takes into account the probability of allele sharing, based on the proposed relationship, the frequency of alleles in their population group(s) and where applicable, takes into account the possibility of a mutation. A LR will be calculated for each genetically independent STR locus, and all the LRs for the different loci can be multiplied to calculate a CLR.

1.10 DNA profiling at the National Health Laboratory Service

The Division of Human Genetics, National Health Laboratory Service (NHLS) and School of Pathology, Faculty of Health Sciences, The University of Witwatersrand, DNA Profiling Laboratory (DPL), Braamfontein, Johannesburg, South Africa has been involved in offering internationally recognized DNA profiling services to the public for parentage and kinship testing for nearly three decades.

Since 2014, the Department of Home Affairs (DHA) has introduced amendments to national regulations, some resulting in the need for DNA testing to resolve parentage and kinship claims, particularly when applications are made for unabridged birth certificates, where one of the parents is not a South African citizen (Department of Home Affairs, 2020a) or in instances where the biological parents are not available to apply for birth certificates for minor children/late birth registrations, and for immigration claims, to

mention a few (Department of Home Affairs, 2020a, 2020b). As a result of these changes, the number of requests for kinship DNA testing has increased dramatically as illustrated in Table 1.6.

Table 1.6 Number of kinship requests processed by the DPL from 2014 to 2019

Year	Number of kinship requests	Number of conclusive outcomes	Number of inconclusive outcomes
2014	18	12	**6 (66%)**
2015	31	12	**19 (61%)**
2016	93	66	**27 (29%)**
2017	192	112	**80 (42%)**
2018	500	254	246 (49%)
2019	783	***Not available**	***Not available**

*Not available: Please note that these numbers are currently not available due to a backlog in the laboratory

1.11 The current testing strategy at the DNA Profiling Laboratory

The PowerPlex® 18D system (Promega Corporation, Madison, WI) is currently (as of 2018) used by the DNA profiling laboratory, to resolve parentage and kinships queries. This kit utilizes a multiplex system, where 18 STR loci are co-amplified simultaneously (Promega Corporation, 2017a). The PowerPlex® 16HS (Promega Corporation, 2016) was replaced by the PowerPlex® 18D when direct amplification was introduced in the laboratory. However, the In-house software and database were never updated to include the two additional STR loci (D2S1338 and D19S433) contained the PowerPlex® 18D system. Because of this, only the overlapping 15 autosomal loci, present in the

PowerPlex® 16HS and the PowerPlex® 18D system are reported on. It is well established within the DPL that the 15 common STR loci contained in both kits are effective in resolving parentage queries, and most kinship disputes that have been processed. However, ss an increase in the number of requests for kinship testing from 2014 has been observed, so has the number of inconclusive outcomes also increased (Table 1.6).

The DNA Profiling Laboratory reports its outcomes based on the following three-zone partitioning for kinship testing based on findings by (Coste and Pouchot, 2003; Giroti et al., 2007; Pu and Linacre, 2008):

a) The **inclusion criteria** - CLR of 100 and more, and a probability of relatedness above 99%.
b) A CLR between 0.1 - 99 is considered to be **inconclusive**, meaning, there is insufficient statistical evidence to either support or not support the proposed relationship. The findings in this inconclusive range showed high levels of false positives and false negatives, when supplementary testing was performed.
c) A CLR below 0.1 is considered as **no support** for the proposed relationship.

1.12 Additional testing available to resolve inconclusive outcomes

When an outcome is inconclusive, lineage genetic markers, namely X chromosome , Y chromosome and mitochondrial DNA (mtDNA) testing can be used to supplement autosomal STR loci outcomes under certain circumstances (Maguire and Woodward, 2008). The nature of the biological relationships in question must lend themselves to either Y chromosome, X chromosome or mtDNA testing. A further consideration of downstream testing is at the additional cost to the client (both financially and with regards to time).

The X, Y chromosome and mtDNA are inherited without recombination, they are inherited essentially as a haplotype. To assess the outcome derived from linkage markers, the product rule cannot be used. Instead, the counting method is used to establish how common the profile is in the population. The calculation, using the counting method, should be conducted using the correct population database. The larger the database, the more accurate the probability will be. This is achieved using the following formula:

P (probability) =X (number of profiles in the database with the exact profile) /

N (total number of profiles in the database) (Butler, 2005)

Figure 1.6 shows several scenarios where X chromosome testing can be informative as a supplementary test to an inconclusive outcome based on autosomal DNA testing. Females have two X chromosomes, and can pass on either one to their offspring. This transmission can be traced when testing a female child and the alleged paternal grandmother (Figure 1.6C). It is also passed on from the father, unchanged to all his daughters. Female children claiming to share a father, even though they have different mothers, will share one X chromosome (Figure 1.6D). Also, in motherless paternity testing, where there is one mismatch between the alleged father and a female child suspected to be due to a mutation, X chromosome testing can supplement autosomal DNA testing outcomes (Figure 1.6A). Commercial kits such as Investigator Argus X-12 (Qiagen) have been validated for use in kinship testing. The STR loci located on the X chromosome can sometimes be more informative than autosomal STRs, depending on the biological relationship being tested (Tillmar et al., 2017)

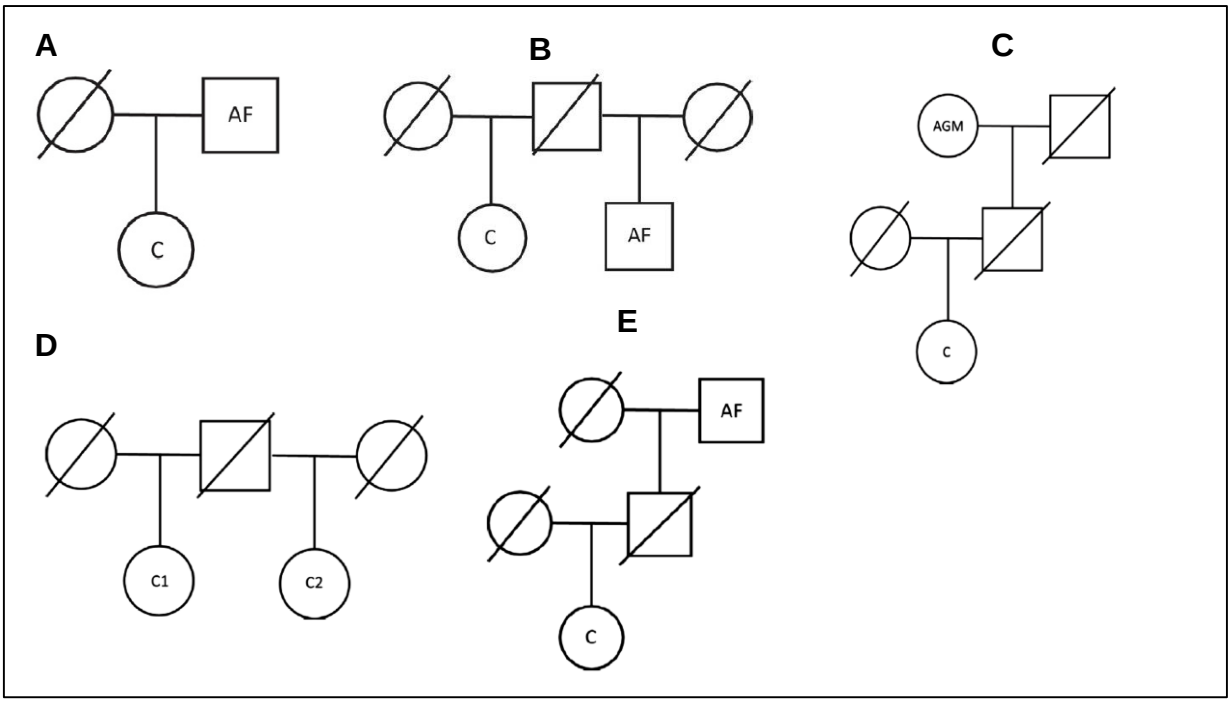


Figure 1.6 Scenarios where X chromosome testing will be applicable.

Different instances where X chromosome testing can be used in kinship testing. A: In motherless paternity testing where the presence of a mutation leads to a Combined Paternity Index less than 100. B: to resolve paternity when the biological father is the father of the alleged father. C: When autosomal DNA testing between child and alleged paternal grandmother yielded an inconclusive outcome. D: When two half-sisters claim to share a father. E: When the biological father is the son of the alleged father. C, child; AF, alleged father; AGM, alleged grandmother.

The Y chromosome test is recommended when the alleged father is not available for testing and inferences are to be made by testing individuals from his male lineage, such as his father, brothers and male cousins (Gill et al., 2001). This test is commonly used in forensic investigation to resolve rape cases. The Y-chromosome is passed on from fathers to sons, as a result, males in the same paternal lineage will possess an identical Y chromosome fingerprint, (see Figure 1.7A). (Father to son Y transmission will be identical except in cases where a mutation took place during gametogenesis; the

mutation rate for the Y chromosomes is approximately 0.2% per generation (Gill et al., 2001).

Mitochondrial DNA haplotypes are inherited uniparentally (Parson et al., 2014). Females pass on their mtDNA to all their offsprings, males and females, however, only females can pass it on to the next generation. This testing can only be performed on individuals from the same maternal lineage, such as aunt-niece/nephew, grandmother-grandchild relationships (see Figure 1.7B). Mitochondrial DNA typing is achieved by sequencing the non-recombining, non-coding hypervariable control region, and comparing the similarities of the sequences obtained.

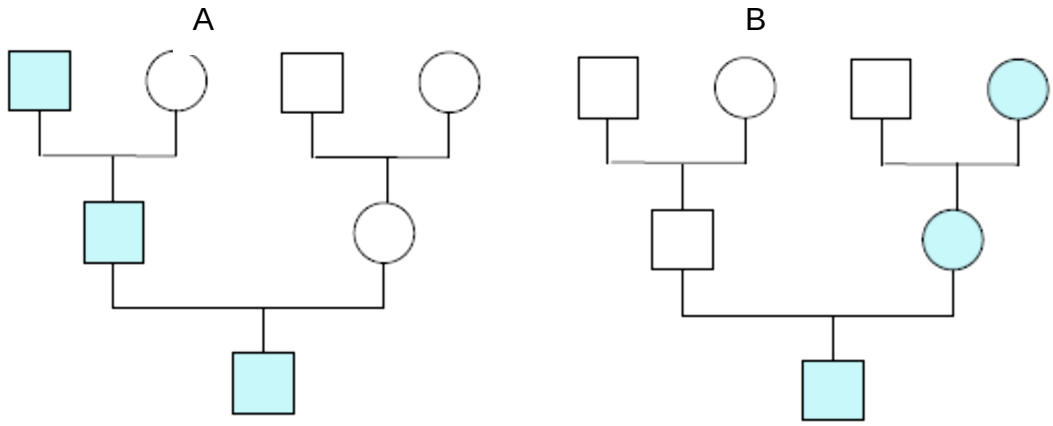


Figure 1.7 Maternal (mtDNA) and paternal (Y chromosome) patterns of inheritance.

Pedigree A shows the inheritance pattern of the Y chromosome. The shaded blue squares demonstrate that a father passes on his Y chromosome to his son(s) who in turn pass(es) it on, unchanged to their son(s). Pedigree B shows the inheritance pattern of the mtDNA. It is passed on by the females to all their offsprings, male and female. However, only females can pass it on to the next generation.

1.13 Challenges faced with the current testing strategy

The main challenge faced with the current testing strategy is the high number of inconclusive kinship outcomes for complex kinship queries, when a commercial kit with 15 polymorphic STRs is used for generating genotypic data (see Table 1.6). The implication of inconclusive outcomes is devastating for the families tested, as additional testing has to be performed (where it is applicable) at an additional cost. This prolongs the time it takes to receive the final DNA test outcome and subsequent outcome of the application for legal documentation, for example. Sometimes, the families cannot afford to pay for additional testing.

1.14 What causes kinship analysis to be complex?

As indicated in Table 1.3, allele sharing between individuals decreases as the genetic distance increases and this makes it difficult to observe sufficient genotypic similarities between individuals tested in a kinship case, resulting in equivocal assignment of relationships based on genotype data only. There can be several contributing factors to an inconclusive result including 1) the number of STR loci used, 2) the number of segregating alleles at each locus, 3) the extent of polymorphism at the loci tested, 4) incorrect pedigree provided and 5) the presence of a mutational event (Chakraborty and Jin, 1993). The presence and source of a mutational event may or may not be detected in complex kinship analysis. This is because, when testing full siblings, a mismatch is expected in 25% of cases per locus (k_0=1/4). The value increases to 50% (k_0=1/2) for half siblings and 75% (k_0=3/4) for first cousins. Also, the individuals tested provide a pedigree indicating their hypothesized relationship (social pedigree). However, the biological correctness of the stated relationship between individuals presented for testing is not always valid (may not match the genetic pedigree). This may result from

unstated adoptions, undetected baby swops or undisclosed extramarital relationships in the previous generation(s).

1.15 Potential solutions regarding inconclusive kinship outcomes

In order to diminish the assignment of equivocal relationship outcomes between individuals, where a true biological relationship exists, previous studies have demonstrated that increasing the number of STR loci employed during genetic testing can significantly decrease the number of inconclusive outcomes (Musanovic et al., 2012). This is because analysis of STRs is based on allele sharing between tested individuals. The more STR loci are tested, the better the resolution power of the test.

In 2011, in order to support law enforcement DNA databases, the FBI made known its intention to expand the number of the CODIS core loci set from 13 to 20, which would result in increasing the discrimination power of the kits used to aid in resolution of criminal and missing person cases (Hares, 2012). The expansion of the CODIS core loci set and subsequent expansion of the ESS resulted in the manufacturing of commercial kits with an extended number of STR loci (Table 1.1- kits listed under CODIS 20). The inclusion of additional STR loci would greatly increase the discrimination power of the test and promote compatibility for international data sharing, and assist with resolving complex kinship analysis (Butler and Hill, 2013). The PowerPlex® Fusion system (Promega Corporation, 2017b) and the VeriFiler™ Express PCR amplification kits (Life Technologies Holdings Pte Ltd, 2017) possess the capability of simultaneously amplifying 24 and 25 different STR loci respectively, resulting in a higher discrimination power (Oostdik et al., 2014). The PowerPlex® Fusion and the VeriFiler™Express kits have seven and eight additional polymorphic STR loci respectively. There is an overlap of 16 STR loci contained in these commercial kits and those contained in the PowerPlex® 16 HS system.

The probability of identity (PI) value, which is defined as the probability that randomly selected individuals will possess an identical genotype at all the tested locus (Butler, 2014; Musanovic et al., 2012), was shown (see Table 1.7) to be significantly higher for the PowerPlex® Fusion, GlobalFiler™ and VeriFiler™ Express kits than that of PowerPlex® 16 HS and PowerPlex® 18D, offering better discrimination power because of the increased number of STR loci (Butler, 2012, Coble and Butler, 2013; Hill et al., 2013). These commercial kits have now gained prominence when conducting parentage or kinship testing worldwide (Oostdik et al., 2014; Turrina et al., 2014).

With regards to cases where a mutational event took place from one generation to the next, this can be mitigated by testing as many people in the family as possible and building up pedigrees for the parents in question based on the people tested. However, this approach is costly and not always applicable. The use of lineage markers, where applicable, would be ideal.

Table 1.7 The probability of identity (PI) of 6 different DNA testing commercial kits and the kit prices

Amplification kit	# of loci	PI Caucasians)	PI (African-Americans)	PI (Hispanics)	PI (Asians)	Kit price	Cost per reaction
PowerPlex® 16 HS	16	4.24×10^{-18}	6.09×10^{-19}	1.26×10^{-18}	2.55×10^{-17}	R88 831 (400 rxns)	R222
IdentiFiler™ Direct	16	6.87×10^{-18}	1.04×10^{-18}	2.73×10^{18}	5.31×10^{-18}	R70 400.00 (200 rxns)	R352
PowerPlex® 18D	18	9.82×10^{-21}	5.60×10^{-22}	2.54×10^{-21}	7.92×10^{-20}	R160 000 (800 rxns)	R200
PowerPlex® Fusion	24	2.35×10^{-27}	1.59×10^{-28}	2.12×10^{-27}	1.42×10^{-25}	R94 374 (200 rxns)	R472
GlobalFiler™	24	1.30×10^{-36}	3.20×10^{-27}	2.27×10^{-26}	1.81×10^{-24}	R 63 600 (200 rxns)	R318
VeriFiler™ Express	25	4.69×10^{-28}	2.55×10^{-30}	2.95×10^{-29}	4.69×10^{-28}	R37 572 (200 rxns)	R188

*Prices as of March 2020. Adapted from (Butler, 2012). The VeriFiler™ Express has been shown to have the lowest cost per reaction when compared to other kits, both expanded and non-expanded kits

1.16 Comparison of software packages used to calculate Likelihood Ratios

In paternity and kinship testing, the impact of using the correct software for calculating Likelihood Ratios has huge legal implications, as incorrect calculations can result in individuals being erroneously classified as related or unrelated. False positive outcomes may result in undue access to estate benefits and false negative may deprive an individual estate benefits or approval of an application for maintenance or citizenship. A negative outcome may even lead to divorce in false non paternity scenarios. For these reasons, all laboratories offering this service should validate or use validated software that meets good quality and assurance standards. A high level of concordance between outcomes obtained when using different software packages is expected for calculation of Likelihood Ratios (Azevedo et al., 2011). It is worrying to know that some commercial and freely available software packages for calculating Likelihood Ratios are not sufficiently validated to meet the required standards for competent testing (Drábek, 2009). Algebraic formulae that are based on the Mendelian inheritance and probability of transmission have been published by the American Association of Blood Banks (AABB) (see Table 1.4 and 1.5) and a concordance study should be conducted to validate the results obtained when using any software package, based on these formulae (Azevedo et al., 2011).

Calculation of Likelihood Ratios or paternity indices can be done manually, however, this approach can be very complex and prone to human error, especially in complicated kinship cases. In order to aid the process of resolving complex kinship cases, the In-house data analysis software used in DPL for calculation of Likelihood Ratios was validated, as no validation records could be found. This was because of change of management in the laboratory. Three analysis software packages, namely, *Familias, GeneMarker® HID* and *Converge™* Software were used in the validation process (see Appendix F). The SWAGDAM and the DNA Advisory Board of Quality Assurance Standards (DBA) provide guidelines for validation of software used in HID Likelihood Ratios calculations (National Institute of Standards, 2009; SWGDAM,

2012). Validation is done to ensure that the same accurate result is obtained each time the calculation is performed (Drábek, 2009).

1.17 Study aim and objectives

The aim of the present study was to evaluate the utility of the additional polymorphic STR loci present in the evaluated commercial kits (eight in the VeriFiler™ Express PCR amplification kit and seven in the PowerPlex® Fusion™ system) and the validity of kinship testing results by comparing the calculated Likelihood Ratio outcomes with those obtained when using PowerPlex® 16 HS and PowerPlex® 18D.

The following study objectives were established:

1) To generate allele frequencies for the additional loci included in the commercial kits to be evaluated, for the population groups routinely tested in DPL. These seven population groups are: South African Bantu-speakers, South African Whites, South African Coloureds, South African Indians, Namibian Bantu-speakers, Nigerian Bantu-speakers and Zimbabwean Bantu-speakers.

2) To use the allele frequency data to:

i) Update the database to include the newly generated allele frequencies, and

ii) Retest and compare the outcomes for kinship cases previously tested using only 15 polymorphic autosomal STR loci.

3) To compare different kinship data analysis software programs. This involved comparing the In-house database with other reputable data analysis software programs, to assess if there was a significant difference in the combined Likelihood Ratios obtained, especially with regards to complex kinship disputes. The three data analysis software packages evaluated are *Familias, GeneMarker® HID* and *Converge™* Software were used in the comparison.

CHAPTER 2 MATERIALS AND METHODS

2.1 Subjects

A total of 729 samples from seven different ethnic groups (Table 2.1), were selected from samples received in the DNA profiling laboratory between 2016 and 2019. Samples were included in the study based on tested individuals' self-reported ethnic affinities.

Table 2.1 Summary of the different population groups included in the study according to geographic location and self-reported ethnicity

Country	Variables (Ethnic groups)	No. of samples (n)	Commercial kit used
South Africa (SA)	1.SA Bantu speakers (SAFBS) (South Eastern Bantu (SEB)	366	VeriFiler™ Express (n=319) PowerPlex® Fusion system (n=47)
	• Ndebele	13	
	• Northern Sotho	66	
	• Tsonga/Shangaan	37	
	• Tswana	36	
	• Venda	23	
	• Xhosa	39	
	• Zulu	128	
	• Alleged father excluded, no mother	12	
	2. SA Indians	26	Only VeriFiler™ Express used (n=363)
	3. SA Whites (SAFWH) (European ancestry)	84	
	4. SA mixed ancestry (SAFMX) (Coloureds*)	40	
Namibia	5. Bantu speaking individuals (NAMBS) (SWB)	27	
Nigeria	6. Bantu speaking individuals (NGRBS)	73	
Zimbabwe	7. Bantu speaking individuals(ZIMBS)	113	
Total (N)		729	

*The term Coloureds is used as a descriptive term to distinguish the group of admixture included in the study, it is not meant to be offensive.

The subjects selected to determine population-based allele frequencies included: child-alleged father pairs (from cases of non-paternity); mothers and alleged fathers from same families (assuming no consanguinity), and unrelated individuals from kinship testing queries. In cases where the ethnicity was not recorded at the time of sampling or where the alleged father was excluded (shown to not be the biological father), the ethnicity of children was inferred based on the mother's self-reported ethnicity, if applicable, or indicated as unknown within the applicable ethnic group.

Additionally, a retrospective assessment of kinship reports from 63 families, where the initial outcome was reported based on a commercial kit with 16 STR loci, was conducted using genotypes obtained from commercial kits with 24-25 STR loci. The individuals included in the assessment were all from the SA Bantu-speaking population group.

All samples used were anonymized by allocation of sample codes. Ethics approval was obtained from the University of Witwatersrand Medical Human Research Ethics Committee (HREC), protocol number M170474 (Appendix A). At the time of sample collection for parentage testing, the individuals signed a consent form (see Appendix B). This form includes a clause which reads "I allow the laboratory to store my sample/DNA and use it for research purposes". A flow chart showing my involvement on different stages of the workflow is presented in Figure 2.1.

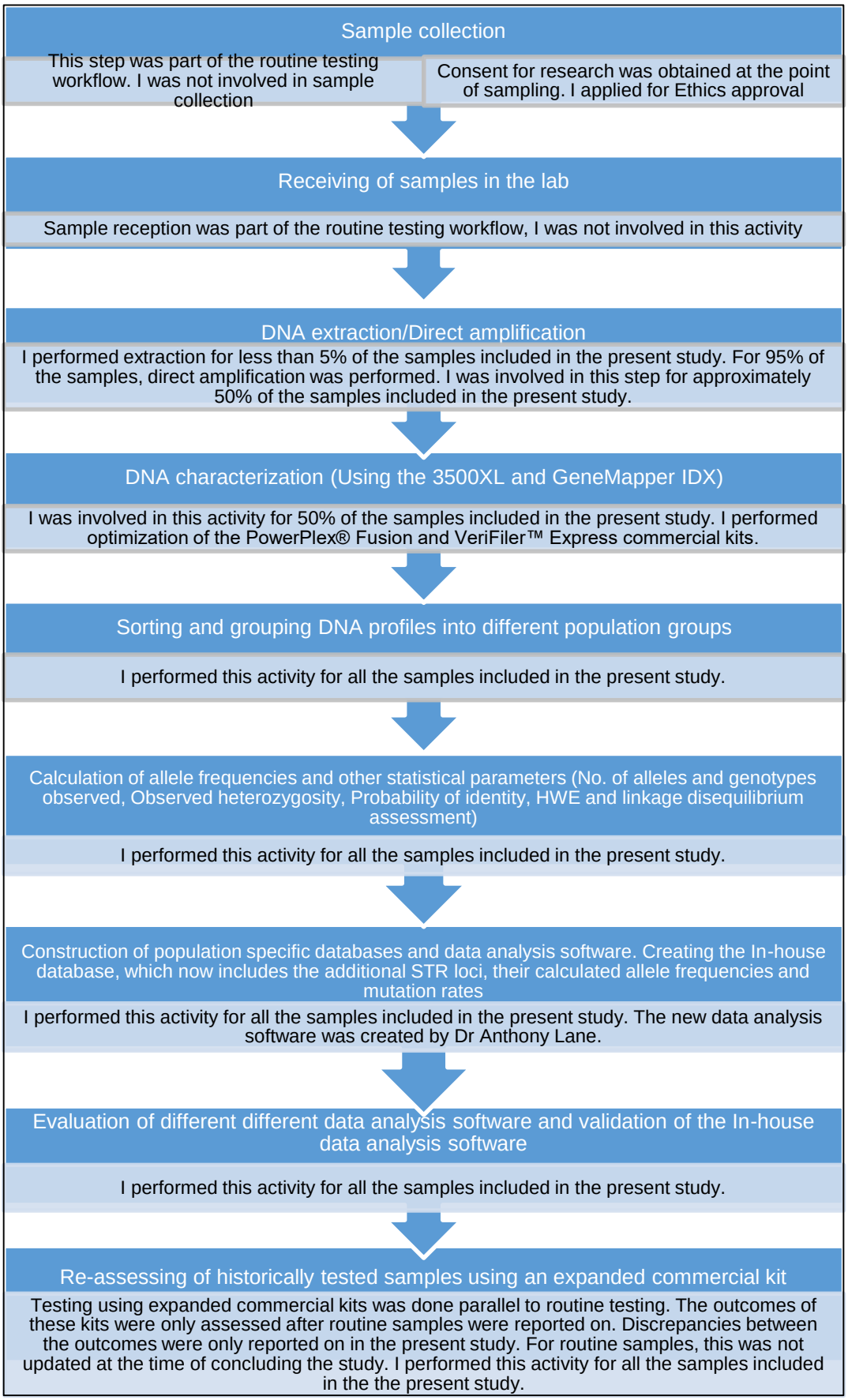


Figure 2.1 A flow chart showing student involvement in different parts of the workflow.

2.2 Methods

2.2.1 DNA extraction

DNA from liquid blood was extracted using the semi-automated Maxwell® 16 system (Promega Corp., Madison, WI), according to the manufacturers' instructions, and following the Standard Elution Volume (SEV) protocol. DNA from untreated blood storage cards was extracted using the PrepGEM™ Universal kit (MicroGem, New Zealand), according to the manufacturer's instructions and following the "blood on storage cards" protocol. Blood samples stored on pre-treated blood storage cards required no DNA extraction. Direct amplification of DNA was performed following punching of the blood cards using the 1-2mm (millimetres in diameter) Harris micro-punch and the Harris punching mat (Merck). Some samples were processed using the semi-automated Delfia® Dried blood spot puncher (Perkin Elmer), with a punching head size of 1.3 mm. The process of punching using specific head-size punches was to standardize the input sample size.

2.2.2 DNA amplification

DNA amplification was performed according to the manufacturer's instructions using PowerPlex® Fusion system (Promega Corp., Madison, WI) and VeriFiler™ Express PCR amplification kit (Life Technologies, Foster City, CA) technical manuals. A summary of the STR loci included in each kit respectively is presented in Table 2.2.

Table 2.2 Summary of loci included in the commercial STR kits evaluated in the current study.

Locus	PowerPlex® 16 HS system	PowerPlex® Fusion system	VeriFiler™ Express PCR amplification kit
D3S1358	√	√	√
vWA	√	√	√
D16S539	√	√	√
CSF1PO	√	√	√
TPOX	√	√	√
D8S1179	√	√	√
D21S11	√	√	√
D18S51	√	√	√
Penta_E	√	√	√
D2S441	-	√	√
D19S433	-	√	√
TH01	√	√	√
FGA	√	√	√
D22S1045	-	√	√
D5S818	√	√	√
D13S317	√	√	√
D7S820	√	√	√
D6S1043	-	-	√
D10S1248	-	√	√
D1S1656	-	√	√
D12S391	-	√	√
D2S1338	-	√	√
Penta_D	√	√	√
Amelogenin	√	√	√
Y-indel	-	-	√
DYS391	-	√	-

*The eight autosomal and two Y markers not present in the 16 HS kit are highlighted in bold.

Direct amplification of DNA from storage card punches using PowerPlex® Fusion system

The PowerPlex® Fusion system is a five-colour multiplex DNA amplification kit which co-amplifies 24 STR loci (Promega Corporation, 2017b). The STR loci amplified using this commercial kit are listed in Table 2.2. The majority of the STR loci included in this commercial kit are polymorphic, except *Amelogenin* and DYS391, which are non-polymorphic markers, and used for sex determination.

Following punching of the dry blood spots (DBS) and placing one 1.2/1.3 mm storage card punch per sample (depending on the punching method used) into a clean 96 well PCR plate or 0.2 millilitres (ml) reaction tubes, the following reagents were thawed and centrifuged: PowerPlex® Fusion 5x Master mix, PowerPlex® Fusion Primer mix and the Amplification grade water (all reagents and water were supplied in the PowerPlex® Fusion system kit). A sample map (Appendix C), indicating the name and location of each sample on the 96 well plate, was constructed during punching, for samples prepared in the 96 well PCR plate. Where 0.2ml PCR tubes used, the tubes were clearly labelled with the sample names. For each set-up, the number of reactions to be set up was predetermined, including the positive and negative control reactions required. The PCR amplification mix was prepared by adding 5 microliters (µl) of the PowerPlex® Fusion 5x Master mix, 5 µl of the PowerPlex® Fusion Primer mix and 15 µl of water (H_2O) for one full reaction. The total reaction volume was 25 µl per reaction. This was multiplied by the number of reactions to be set up. The PCR amplification mix was vortexed and aliquots of 25 µl were added to each sample. For the positive control, 1 µl of the 10 nanogram(ng)/µl control DNA provided in the PowerPlex® Fusion System was added to 24 µl of the PCR amplification mix. For the negative control, no DNA was added to the 25 µl PCR amplification mix. The 96 well plates were sealed using the MicroAmp® Clear adhesive film (Life Technologies). The sealed plates or closed 0.2 ml PCR tubes were centrifuged and then placed in the GeneAmp® PCR system 9700 (Life Technologies) for PCR amplification.

Thermal cycling

The cycling conditions for the PowerPlex® Fusion system were as follows: 96 degrees Celsius (°C) for 1 minute (min) at maximum ramp, followed by 27 cycles of denaturation at 94°C for 10 seconds (sec), annealing at 59°C for 1 minute, and extension at 72°C for 30 sec. Final extension was done at 60°C for 20 minutes and a final hold step at 4°C indefinitely.

Direct amplification of DNA from storage card punches using the VeriFiler™ Express PCR amplification kit

The VeriFiler™ Express PCR amplification kit is a six-colour system multiplex DNA amplification kit which co-amplifies 25 STR loci (Thermo Fischer Scientific, 2017). The STR loci amplified in this commercial kit are listed in Table 2.2. The majority of the STR loci included in this commercial kit are polymorphic, with the exception of *Amelogenin*, which is used for sex determination (Thermo Fischer Scientific, 2017).

Following punching of the dry blood spots (DBS) and placing one 1.2/1.3 mm storage card punch per sample (depending on the punching method used) into a clean 96 well PCR plate or 0.2 ml reaction tubes, the following reagents were thawed and centrifuged: Master Mix and Primer Set (both provided by the manufacturer). A sample map (Appendix C), indicating the name and location of each sample on the 96 well plate, was constructed during punching, for samples prepared in the 96 well PCR plate. Where 0.2ml PCR tubes used, the tubes were clearly labelled with the sample names. The protocol required the use of Low-Tris-Ethylene diamine tetra-acetic acid (EDTA) (TE) Buffer, which was prepared as follows: 10 ml of 1 molar (M) Tris-Hydrochloric acid (HCl), pH 8.0, 0.2 ml of 0.5 M EDTA, pH 8.0 and 990 ml of deionized water were mixed together and the solution was autoclaved. The solution was divided into four 250 ml aliquots and all the aliquots were stored at room temperature and used when required. The PCR amplification mix was prepared by adding 10 µl of the Master Mix, 10 µl of the Primer set and 5 µl of the Low-TE Buffer. The final reaction volume was 25 µl per reaction. This was multiplied by the number of reactions set up. The PCR amplification mix was vortexed and 25 µl of PCR mix was added to each sample. For the positive control, 1 ng/µl of the control DNA provided in the VeriFiler™ Express PCR amplification kit was added to 24 µl of the PCR amplification mix. For the negative control, no DNA was added to the 25 µl PCR amplification mix. The 96 well plate(s) were sealed using the MicroAmp® Clear adhesive film (Life Technologies). The sealed plates or 0.2 ml PCR tubes were

centrifuged and then placed in the GeneAmp® PCR system 9700 (Life Technologies) for PCR amplification.

Thermal cycling

The cycling conditions for the VeriFiler™ Express PCR amplification kit were as follows: 95°C for 1 minute, followed by denaturation at 94°C for 3 seconds, annealing at 59°C for 16 seconds, and extension at 65°C for 29 seconds; for 28 cycles Final extension was at 60°C for 5 minutes and a final hold at 4°C.

Optimization of the VeriFiler™ Express PCR reaction volumes and thermal cycling conditions

There was a need to optimize reaction volumes, thermal cycling conditions and injection volume for the VeriFiler™ Express PCR amplification kit in order to get rid of the stutter and shouldered peaks which resulted in inaccurate sizing of some of the alleles. The relative fluorescent units (RFUs) observed were also very high, indicating that the amount of input DNA was initially too much. Unfortunately, the input DNA could not be varied as dry blood spots were used instead of extracted DNA.

The PCR amplification reaction volumes and the thermal cycling conditions were adjusted as follows, after conducting several trouble shooting exercises: The volumes of the PCR amplification reagents were reduced to half of the original volumes-5 µl of the Master Mix, 5 µl of the Primer set and 2.5 µl of the Low- TE Buffer. The final reaction volume was 12.5 µl per reaction. This reaction volume was used for bloods stored on storage cards. For the positive control and for samples where DNA was extracted, 1 ng/µl of the (control) DNA was added to a quarter of the

initial reaction volume, resulting in combined PCR amplification mix volume/DNA volume of 7.25 µl per reaction.

Thermal cycling conditions

The cycling conditions were adjusted as follows: 95°C for 1 minute, followed by denaturation at 94°C for 3 seconds, annealing at 59°C for 16 seconds, and extension at 65°C for 29 seconds for **25** cycles, final extension at 60°C for 5 minutes and a final hold at 4°C.

2.2.3 Capillary electrophoresis

Following PCR amplification, the loading mix was prepared by adding 9.6 µl of Hi-Di™ Formamide (Life Technologies), 0.4 µl of the GeneScan 600 LIZ (Life Technologies) per sample, for samples amplified using the VeriFiler™ Express PCR reaction kit**.** For samples amplified using the PowerPlex® Fusion system and 0.4 µl of WEN ILS and 9.6 µl of Hi-Di™ Formamide (Life Technologies) were used. The volume of the reaction mix was 10 µl for both commercial kits. This volume was aliquoted into each sample well on the 96 well PCR plate. 1 µl of the PCR product was added to the mix for each sample, including the positive and negative controls. An allelic ladder, which facilitates accurate sizing of alleles, was included in each set up.

The amplified PCR products were loaded on the ABI PRISM® 3500 xl Genetic analyser (Life Technologies, Foster City, CA, USA). Following, electrophoresis, the results were analysed using the GeneMapper ID-X v1.5 (Life Technologies, Foster City, CA, USA). The panels and bins for PowerPlex® Fusion system were downloaded from the Promega website and imported onto the GeneMapper ID-X. For the VeriFiler™ Express PCR amplification kits, the panels and bins came already loaded on the GeneMapper ID-X v1.5 software. The panel file contains information

about the list of loci contained in a kit, while the bin file provides an exact location of each allele within the locus. Analysis of the samples' DNA profiles was preceded by assessment of the size standard, allelic ladder and correct expected full control DNA profile. The negative control was expected to show no peaks, as peaks would indicate DNA contamination. The DNA profiles were accepted as good quality when the allele peaks were more than 100 RFUs and no significant background peaks observed, according to the manufacturer's recommendations (Promega Corporation, 2017b; Thermo Fischer Scientific, 2017) .

2.3 Data analysis

The flow chart for data analysis is summarised in Figure 2.1. Allele frequencies for all the polymorphic STR loci present in the PowerPlex® Fusion system (22) and VeriFiler™ Express PCR amplification kit (23) were calculated using Microsoft Excel 2016 (counting method). The allele frequencies were then confirmed using an online Microsoft Excel macro enabled software (Duewer, 2016). The following statistical parameters were also evaluated using calculated outcomes from the online Microsoft Excel macro enabled software (Duewer, 2016): Number of alleles observed per STR locus, number of genotypes observed per STR locus, observed heterozygosity per STR locus, Probability of identity, Polymorphic information content and STR loci were ranked based on their calculated probability of identity

Following calculation of allele frequencies, the following statistical tests were performed, using *GenePop* (Raymond and Rousset, 2019, 1995; Rousset, 2008) : 1) Hardy Weinberg equilibrium (HWE) test which tests for evolutionary changes in a population. This test determines if the allele frequencies, and consequently, genotype frequencies remain the same across generations (Waples, 2015), and 2) Linkage disequilibrium (LD) test for locus independence. To assess departures from HWE, the following parameters were selected on *GenePop*: Estimation of exact *P*- values by the Markov chain method. The number of dememorization steps, which is the number of steps performed to reach a random starting point before comparing the

alternative contingency table probabilities to those of the observed table (O' Connor et al., 2011), 1000, number of batches, 100 and number of iterations per batch, 1000. These are default values on the software. The HW probability test was employed. Linkage disequilibrium assessment was performed using the same number of dememorization steps, batches and iterations per batch used, similar to those used for HWE assessment. The genotypic disequilibrium test for each pair of loci in each population, using the probability tests sub-option was used.

Genetic variations between the population groups evaluated in the present study were calculated using *GenePop*. This was accomplished by calculating pairwise F_{sts}, using one locus estimates and following standard analysis of variation (ANOVA) (Weir and Cockerham, 1984). ANOVA is a statistical method which is based on testing whether the means between two or more groups are equal, and therefore can be used to assess the degree of differentiation between population groups (Holsinger and Weir, 2009).

2.3.1 Updating of the database to include the newly generated allele frequencies

The In-house data analysis software was updated with the help of Dr Anthony Lane (former Head of the DPL section). This involved creating a new Microsoft Excel macro enabled software to include all the 23 STR loci, their allele frequencies and mutation rates for the different population groups. Formulae for calculating Likelihood Ratios were incorporated and the new calculations were based on the 23 polymorphic STR loci instead of the former 15 STR loci.

2.3.2 Retesting and evaluation of kinship cases previously tested using PowerPlex® 16 HS and PowerPlex® 18D

Following routine procedures, kinship cases were initially analysed using 15 STR loci present in the PowerPlex® 16 HS system and PowerPlex® 18D. In the present study, a total of 63 scenarios (family relationships), were retested using either PowerPlex® Fusion or VeriFiler Express commercial kit, containing 22 or 23 polymorphic STR loci, respectively. Of these 63 scenarios, 17 relationship scenarios were initially reported as not excluded, three as excluded and 43 were initially reported as inconclusive. The breakdown of the different scenarios tested is shown in Table 2.3. The selection of these relationship scenarios was based on the common test requests received in the DPL. The different relationship scenarios tested represented instances of who was available in a family at the time of testing and not necessarily a choice on the part of the individuals requiring DNA testing. It has been our experience that sometimes suitable close relatives were available for testing, but refused to be tested. This assessment was done to evaluate the utility of kinship analysis using the additional STR markers.

Table 2.3 Scenarios of family relationships used in the evaluation of the utility of additional STR loci included in the expanded commercial kits and in the validation of the In-house data analysis software

Relationship (Scenario)	Number of cases (n)
Siblings	29
Child and alleged grand(mother/father)	8
Child and alleged full aunt/uncle	8
Child and alleged half aunt/uncle	5
Mother, child and alleged sibling	1
Mother, child and alleged grandparent	3
Child 1, Child 2 and alleged grandparent	2
Child 1, Child 2 and alleged aunt/uncle	3
Mother, child 1 and child 2, alleged grandmother	1
Mother, child and alleged half aunt	1
Mother, child 1 and child 2, alleged grandmother	1
First cousins	1

2.3.3 Comparison of Likelihood Ratios using different data analysis software packages

The following data analysis software packages were used to compare calculated Combined Likelihood Ratio outcomes: The In-house Microsoft Excel Macro enabled software (Dr Anthony Lane, unpublished), *Converge™ Software* v.2.1, *Familias* v3.2.2 and *GeneMarker® HID* v3.0.0.0. The Student T-tests were conducted to calculate the significance of the differences obtained when the different software packages were used. The analysis was done using Microsoft Excel 2016, Data analysis tool, t-Test: Paired Two Sample for Means. The null hypothesis (H_o) stated that there was no significant difference in the outcomes obtained when the different analysis software packages were used to calculate CLRs. The alternative hypothesis (H_1) stated that there was a significant difference in outcomes obtained when the different data analysis software packages were used. A threshold of *p*-value less than 0.05 ($p < 0.05$) was used to indicate outcomes that are statistically significant. It was accepted that all these software packages make use of the algebraic formulae published by the AABB, according to the ISFG, SWGDAM and AABB recommendations and guidelines. The different capabilities of each software package are indicated in Table 2.4. The relationship scenarios listed in Table 2.3 were used in the evaluation.

Table 2.4 Capabilities of the different data analysis software for calculating Likelihood Ratios for kinship testing

Software capability and maintenance	In-house software	*Converge™ Software*	*Familias*	*GeneMarker® HID*
Trio calculations	√	√	√	√
Duo calculations	√	√	√	√
Motherless	√	√	√	√
Grandparents/Half siblings/uncle-nephew (Duos)	√	√	√	√
Parent, child and relative	√	√	√	Pair wise only
Kinship- More than 3 family members together	Pair wise only	√	√	Pair wise only
Mutation calculation Incorporated	√	√	√	√
Null alleles calculation Incorporated	Done separately	√	√	√
Simulations	Not capable	Not capable	√	√
Updates	Not documented	√	√	√
User manual	Not available	√	√	√

CHAPTER 3 RESULTS

3.1 Outcomes of allele frequency calculations

Allele frequencies for the STR loci investigated per ethnic group studied are presented in Appendix D, and were calculated using Microsoft Excel® (2016) and confirmed using an online Microsoft Excel macro enabled software tool (Duewer, 2016). Databases for the 23 polymorphic STR loci were constructed for five of the seven ethnic groups studied, namely SA Bantu-speakers, SA Coloureds, SA Whites, Zimbabwean and Nigerian Bantu-speakers (Appendix D). A sufficient sample size could not be obtained to perform a meaningful data analysis for two ethnic groups, namely SA Indians and Namibian Bantu-speakers. According to the recommendations of the United States (US) National Institute of Standards and Technology (NIST), a minimum of 40 samples is required for meaningful data analysis (Duewer, 2016). Only 26 samples could be obtained for the SA Indians and 27 for the Namibian Bantu-speakers. These two ethnic groups were subsequently excluded from further data analysis. The laboratory will collect more samples on these population groups and publish on these findings in future studies.

All samples analysed yielded complete DNA profiles of good quality (Figure 3.1). The positive and negative controls showed expected outcomes when analysed with the PowerPlex® Fusion system and the VeriFiler™ Express PCR amplification kit. When a comparison was made between these commercial kits and the currently used PowerPlex® 16HS or PowerPlex® 18D systems, the 15 overlapping STR loci were in concordance. The overlapping STR loci between all the commercial kits and the additional STR loci assessed in the present study are listed in Table 2.2.

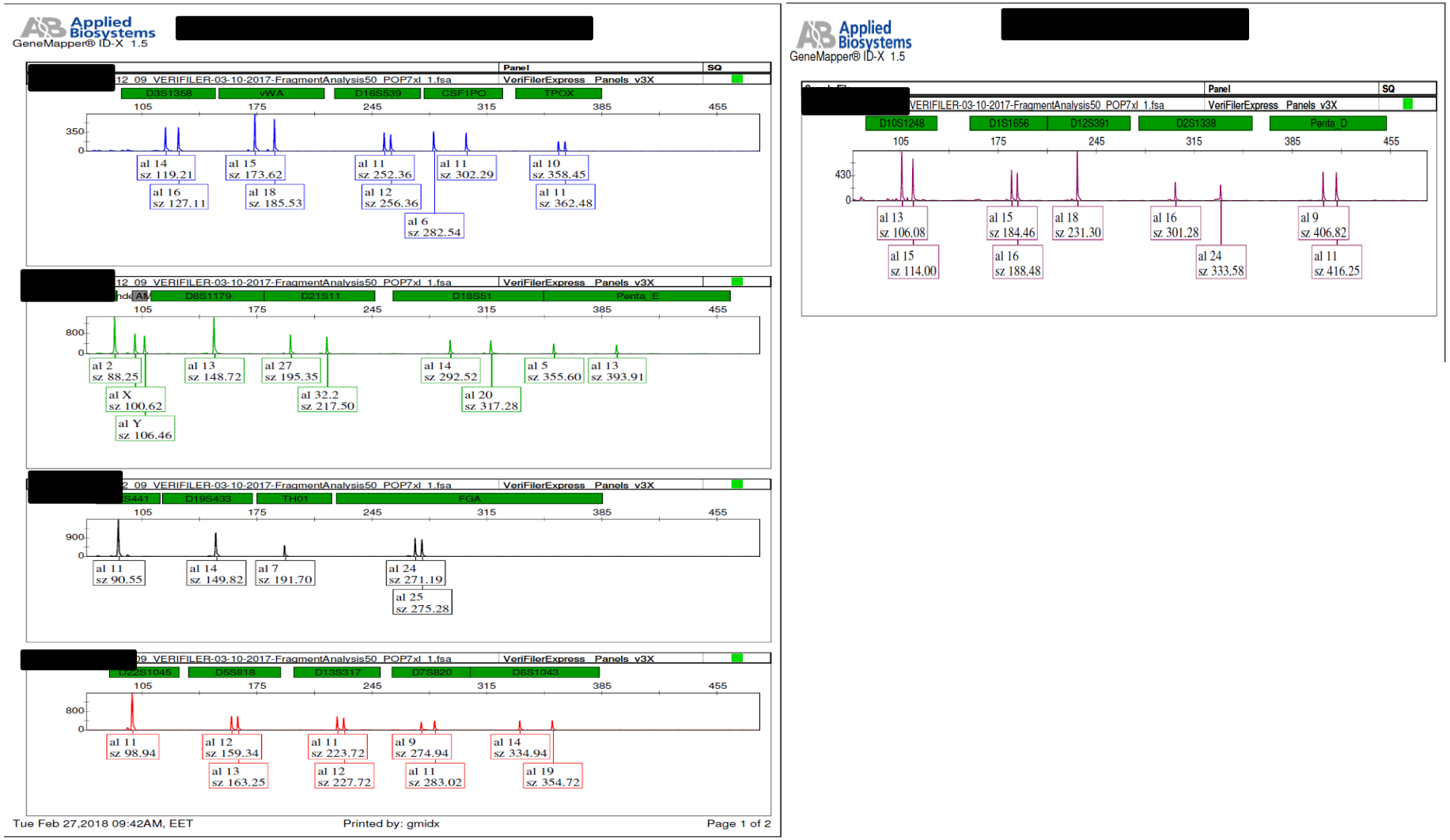


Figure 3.1 An electropherogram showing STR loci (green blocks)and alleles (represented as peaks under the STR loci), observed when VeriFiler Express PCR amplification kit was used.

The minimum allele frequency (MAF), also referred to as the minor allele frequency, is an estimate given when an allele is observed less than five times at a locus. MAF was calculated using the formula 5/2N for each locus, where N represents the number of individuals tested in a population group, and 2N represents the number of chromosomes tested (Crow, 1996). Where the MAF was calculated, these are indicated with an asterisk in Appendix D. The MAFs for the SA Bantu-speakers, SA Whites, SA Coloureds, Zimbabwean and Nigerian Bantu-speakers are 0.0068, 0.0298, 0.0625. 0.0221 and 0.0338, respectively (Appendix D).

3.2 Number of alleles observed per locus

The number of alleles observed gives an indication of the possible number of different alleles that can be observed at an STR locus in a given population group. Figure 3.2 presents a summary of the number of alleles observed per STR locus across the different ethnic groups. The high number of alleles observed at a locus is associated with a high locus variability. The increased locus variability reduces the likelihood of unrelated individuals having the same alleles at a locus (Butler, 2012).

As shown in Figure 3.2, for the SA Bantu-speakers, the FGA locus exhibited the highest number of alleles (29) followed by the D21S11 (23) and then the D18S51 (20) loci. The FGA locus also has the highest number of alleles in the Zimbabwean Bantu-speakers (22), but not for the SA Whites (13), Coloureds (14) and Nigerian Bantu-speakers (14). The latter exhibit Penta_E (16), D21S11 (16) and D21S11/Penta_E (16) loci as having the highest number of alleles respectively. The D3S1358 locus has the least number of alleles in the SA and the Nigerian Bantu speaking groups (6). TH01 locus has the least number of alleles in the SA Whitess (5), Coloureds (5) and Zimbabwean Bantu speakers (5).

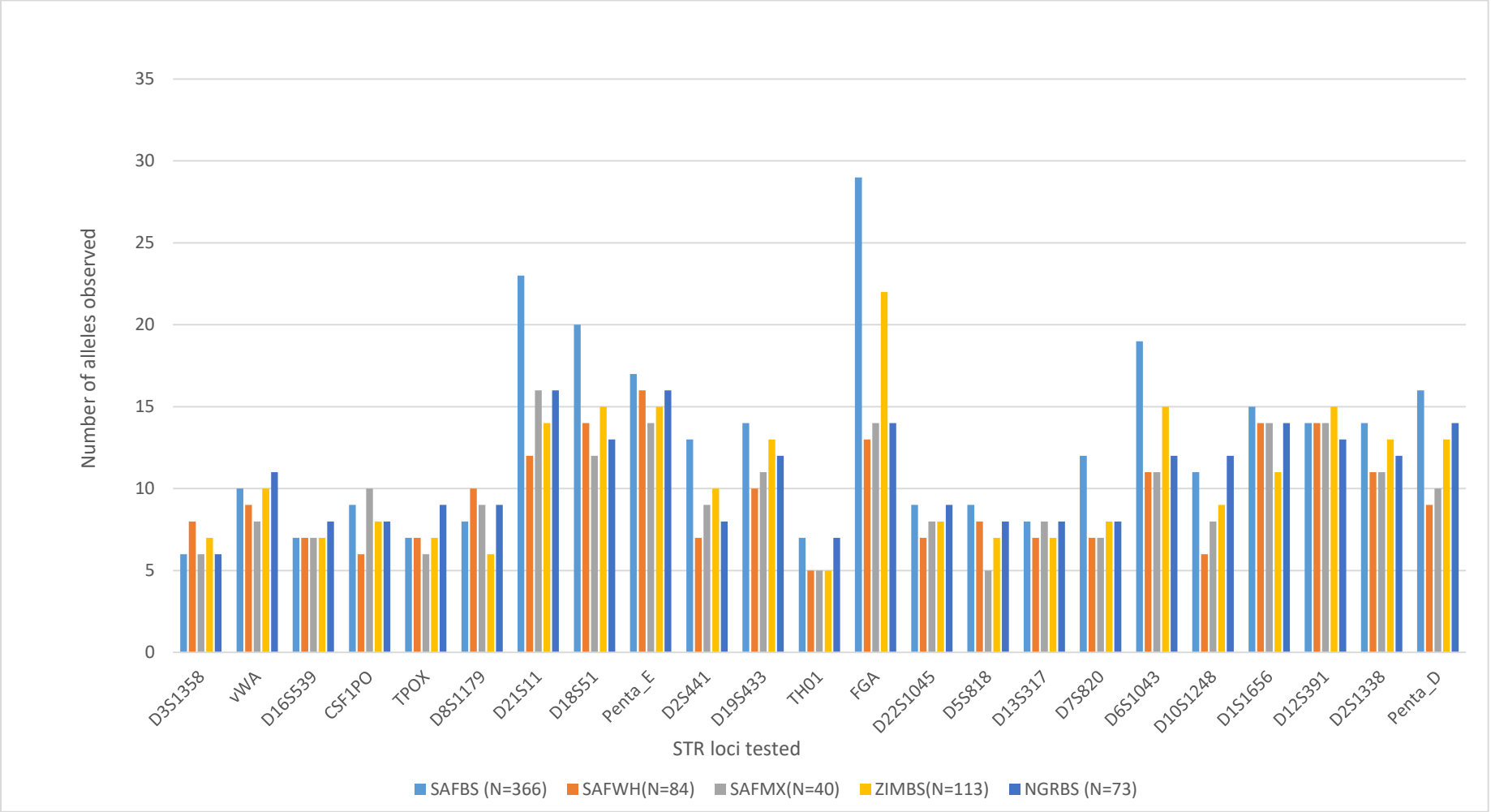


Figure 3.2 A chart showing the number of alleles observed per STR locus across the five different population groups included in the present study.

3.3 Number of observed genotypes

Ideally, a genetic marker used in human identity testing should be highly polymorphic, i.e. displaying many alleles and a high number of observed genotypes ($\#G_{obs}$). Our data, in Figure 3.3 showed that with respect to the SA Bantu-speakers, D21S11 locus displayed the highest number of genotypes (91), followed by the Penta_E locus (85) and then D18S51 (77). The D6S1043 locus displayed 72 observed genotypes. The TPOX and D13S317 loci showed the same number (21) of observed genotypes. The CSF1PO locus exhibited the least number of observed genotypes (9).

With regards to the SA Whites, the D1S1656 locus showed the highest number of observed genotypes (45), followed by the Penta_E and D12S391 loci, each having 42 observed genotypes. The TPOX locus displayed the least number of observed genotypes (13). The CSF1PO (14), THO1 (14), D22S1045 (15), D16S539 (16), D5S818 (16) loci showed low numbers of observed genotypes (Figure 3.3).

The SA Coloured population group showed the Penta_E locus as having the highest number of observed genotypes (32), followed by the D21S11 locus (31), D1S1656 locus (29), D12S391 and D2S1338 loci, both showing 28 observed genotypes. The D5S818 locus showed the least number of observed genotypes (11), following the TPOX and D3S1358 loci, both showing 13 observed genotypes (Figure 3.3).

With regards to the Zimbabwean Bantu-speaking population group, the Penta_E locus (57) showed the highest number of observed genotypes, followed by the FGA (53), D2S1338 (52) and D6S1043 (50) loci. The THO1 locus showed the least number of observed genotypes (11), following the D3S1358 locus (15) (Figure 3.3).

The Nigerian Bantu-speaking population group showed that the Penta_E and Penta_D loci both have the highest number of observed genotypes (37), followed by the FGA (36), D1S1656 (35), D6S1043 (35), D2S1338 (34) D19S433 (34) and D21S11 (34) loci. The following loci showed low numbers of observed genotypes: D3S1358 (12), D13S317 (12) and THO1 (13) (Figure 3.3).

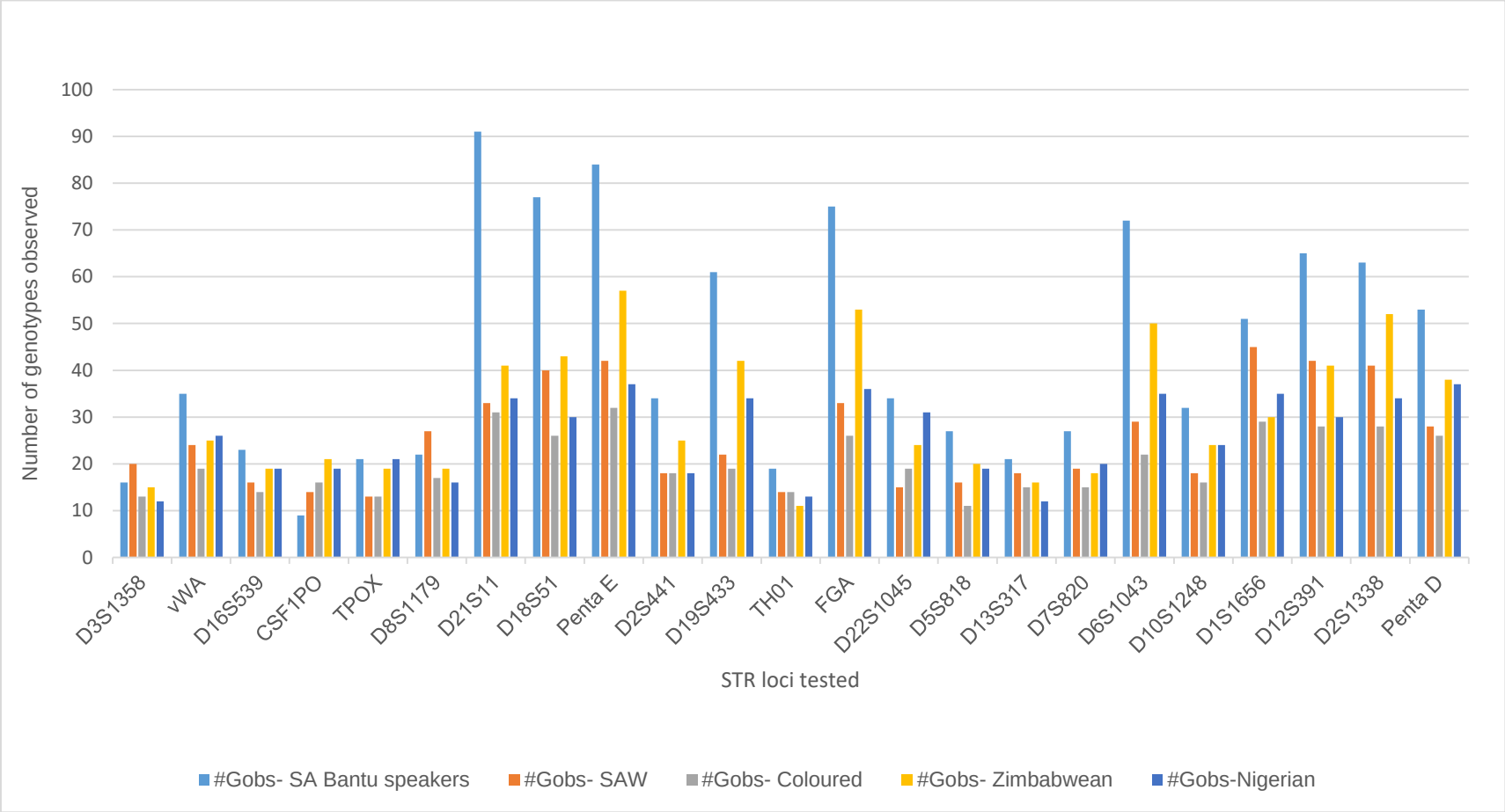


Figure 3.3 Chart showing the number of observed genotypes per STR locus, across the five population groups included in the present study.

3.4 Observed heterozygosity

A high heterozygosity value is associated with high allelic diversity at that locus and a high number of observed genotypes. This then implies that a locus with a high observed heterozygosity (H_{obs}) value is more informative, and results in a decrease in the chance of seeing a random match between unrelated individuals at that locus, within a population group. Our data show that with respect to the SA Bantu-speakers, the D6S1043 locus has the highest observed heterozygosity value (90.6%), followed by the Penta_D locus (89.89%). Generally, for the SA Bantu-speakers, the observed heterozygosity value for the majority of the STR loci tested is above 70%, with the exception of the D13S317 locus (68.58%) (Figure 3.4).

The SA Whites population group showed that the Penta_E locus has the highest observed heterozygosity value (90.48%), followed by the D3S1358 and D1S1656 loci (86.9%). The heterozygosity value for the majority of the tested STR loci is above 70%, with the exception of the CSF1PO (67.86%), D5S818 (60.71%) and TPOX (54.76%) (Figure 3.4).

The SA Coloureds group showed that the Penta_E and D12S391 loci have the highest observed heterozygosity value (92.5%), followed by the FGA, D1S1656 and Penta_D loci (90%). The heterozygosity values for all the tested loci were above 70% (Figure 3.4).

With regards to the Zimbabwean Bantu-speakers, the Penta_E locus showed the highest observed heterozygosity value (92.92%), followed by the D2S1338 locus (91.15%). The heterozygosity value for the majority of the tested loci was above 70%, with the exception of the THO1 (69.91%) and D13S317 (65.49%) loci (Figure 3.4).

The Nigerian Bantu-speaking group showed that the Penta_D locus displayed the highest observed heterozygosity value (89.19%), followed by the D19S433 and D2S1338 loci (87.84%). The heterozygosity value for the majority of the tested loci was above 70%, with the exception of the D13S317 locus (68.92%) (Figure 3.4).

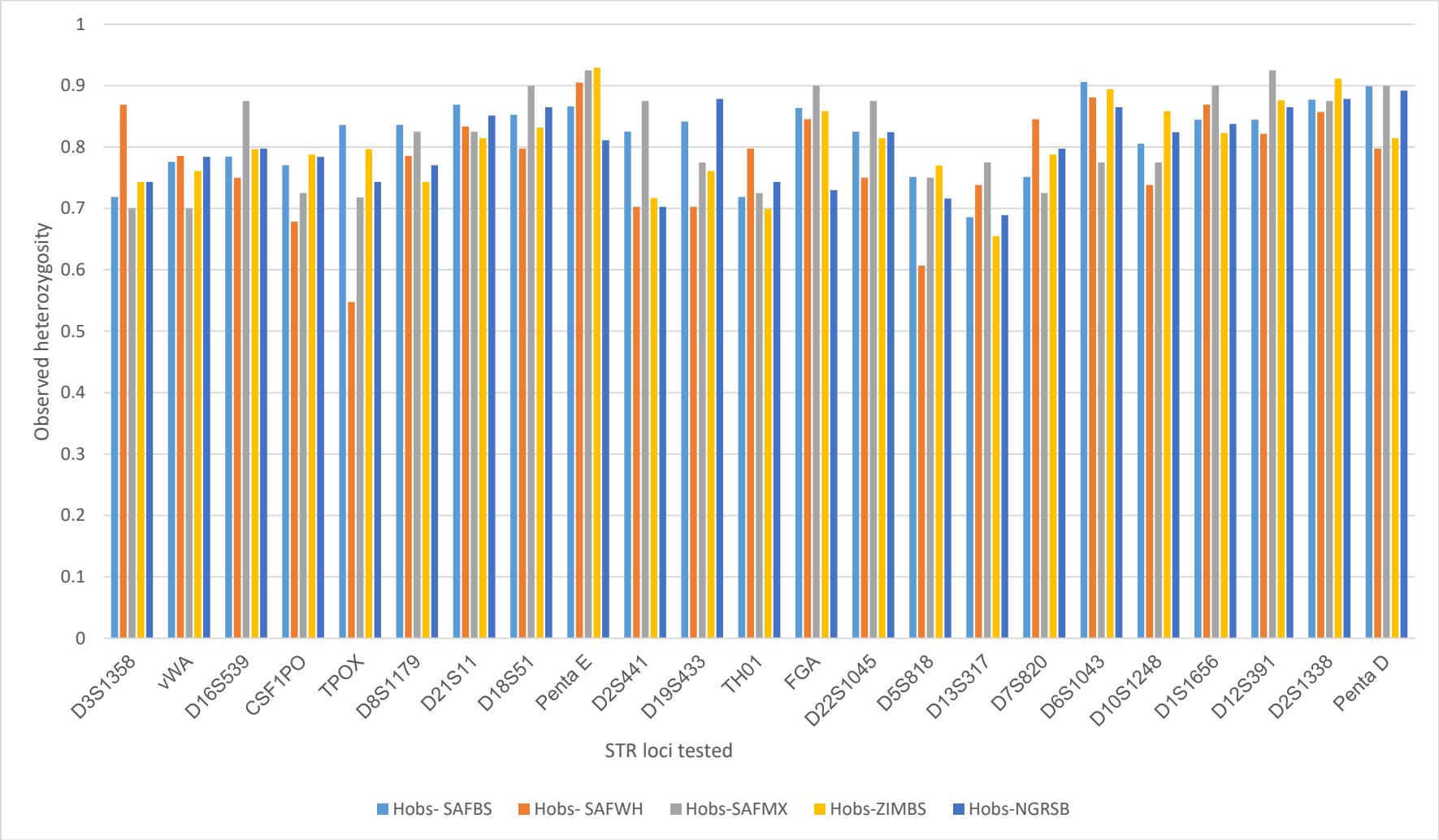


Figure 3.4 Chart showing observed heterozygosity per STR locus, across the five population groups included in the present study.

3.5 Probability of Identity

The probability of identity (PI) is the probability that any two unrelated individuals, selected at random, will possess the same genotype at any given locus (Butler, 2014). This is also referred to as the matching probability. The PI of a single locus is calculated as follows: the sum of the square of the observed genotype frequencies (Butler, 2012). The lower the PI value, the more variable the STR locus is in a population because this indicates that there are more genotypes occurring at a lower frequency (Butler, 2012). Individual PI values can be multiplied together for independently inherited loci (i.e. the product rule) to create a combined profile PI. The same rule can be applied to determine the discrimination power of a commercial kit. The combined PIs observed for the different commercial kits evaluated in the present study are listed in Table 3.1.

Table 3.1 Calculated combined Probability of Identity calculated for the five ethnic groups when different STR kits were used.

STR kit	# of STR loci	SAFBS (N=366)	SAFWH (N=84)	SAFMX (N=40)	ZIMBS (N=113)	NGRBS (N=73)
PowerPlex® 16 HS	16	1.74×10^{-18}	1.65×10^{-17}	3.27×10^{-17}	6.70×10^{-16}	4.80×10^{-18}
PowerPlex® 18D	18	2.44×10^{-21}	6.71×10^{-20}	1.08×10^{-19}	2.69×10^{-17}	6.45×10^{-21}
PowerPlex® Fusion	24	1.55×10^{-27}	7.56×10^{-26}	1.33×10^{-25}	3.09×10^{-23}	1.35×10^{-26}
VeriFiler™ Express	25	4.06×10^{-29}	4.57×10^{-27}	7.83×10^{-27}	1.16×10^{-24}	3.94×10^{-28}

The VeriFiler™ Express PCR amplification kit, showed the highest combined PI than all the commercial kits evaluated in the present study. The combined PI value for the VeriFiler™ Express PCR amplification kit was one to two-fold higher than that of the PowerPlex® Fusion system and more than eight-fold higher than that of the PowerPlex® 16HS (Table 3.1).

Figure 3.5 displays the PI per STR locus tested using the VeriFiler™ Express PCR amplification kit, for the different population groups included in the present study. The STR loci with high variability, based on the PI values, for the SA Bantu-speakers are: The Penta_E (0.0219), D6S1043 (0.0262), FGA (0.0288), D2S1338 (0.029) and D18S51 (0.0309), with the Penta_E locus showing the highest variability. The STR loci that showed low variability are: D2S441 (0.1039), D5S818 (0.1042), D13S317 (0.1265), TH01 (0.1306) and D3S1358 (0.1316), with the D3S1358 locus showing the least variability in the SA Bantu-speaking population group (Figure 3.5).

With regards to the SA Whites, the STR loci that showed high variability are: D1S1656 (0.0309), D12S391 (0.0317), D18S51 (0.0366), D2S1338 (0.0368) and Penta_E (0.0371). The D1S1656 locus exhibited the highest variability. The STR loci that showed low variability are: D19S433 (0.11), CSF1PO (0.1213), D22S1045 (0.1375), D5S818 (0.1593) and TPOX (0.1905). The TPOX locus showed the least variability (Figure 3.5).

The STR loci which showed high variability for the SA Coloured population group are: D21S11 (0.0375), Penta_E (0.0388), D1S1656 (0.0425), D2S1338 (0.0425) and D12S391 (0.0433). The D21S11 locus exhibited the highest variability. The STR loci that showed low variability are: D5S818 (0.11), D13S317 (0.1118), TPOX (0.1124), D8S1179 (0.1125) and D3S1358 (0.1363). The D3S1358 locus showed the least variability (Figure 3.5).

The STR loci which showed high variability for the Zimbabwean Bantu-speaking group are: FGA (0.0275), D6S1043 (0.0291), Penta_E (0.0294), D2S1338 (0.0311) and Penta_D (0.0411). The FGA locus exhibited the highest variability. The STR loci that showed low variability are: D5S818 (0.1042), D2S441(0.1063), D3S1358 (0.11), TH01 (0.1320) and D13S317 (0.1342). The D13S317 locus showed the least variability (Figure 3.5).

With regards to the Nigerian Bantu-speaking population group, the STR loci that showed high variability are: D6S1043 (0.0376), Penta_D (0.0391), D1S1656 (0.0427), FGA (0.0446) and D2S1338 (0.0446). The D6S1043 locus exhibited the highest variability. The STR loci that showed low variability are: D13S317 (0.1527), TH01 (0.1366), D3S1358 (0.1297), D2S441 (0.115) and CSF1PO (0.1132). The D13S317 locus showed the least variability (Figure 3.5).

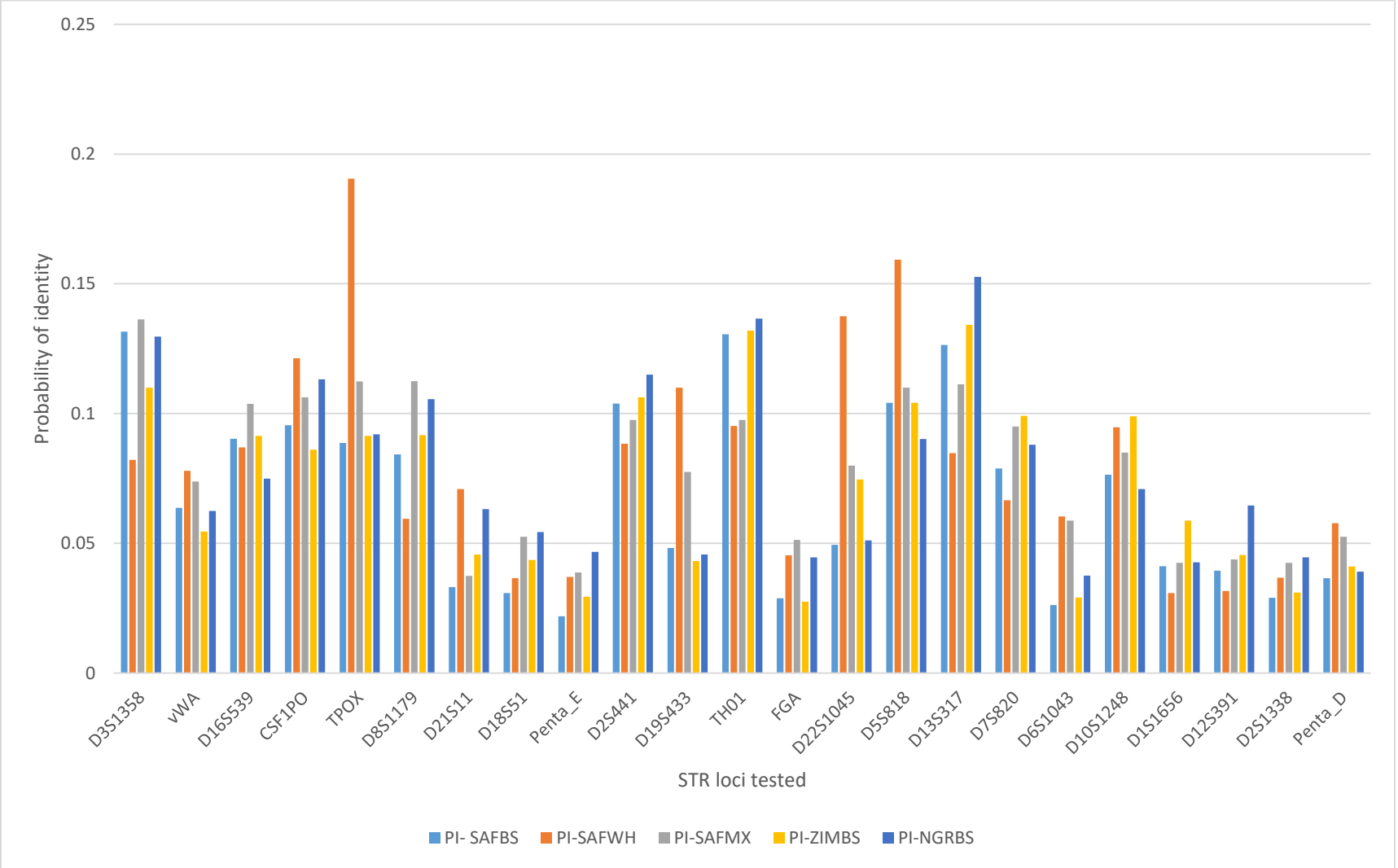


Figure 3.5 Chart showing the Probability of Identity for all the tested STR loci, for population groups included in the present study.

3.6 The Polymorphic Information Content

The Polymorphic Information Content (PIC) is a function of the number of loci examined and their degree of heterozygosity. It gives information about how variable and polymorphic an STR locus is in a population. Figure 3.6 shows the PIC values of all the STR loci tested in the present study, across the different population groups.

The STR loci with high PIC values, for the SA Bantu-speakers are: The Penta_E (88.66%), D6S1043 (87.06%), FGA (86.51%), D2S1338 (86.47%) and D18S51 (85.91%), with the Penta_E locus showing the highest PIC value. The STR loci that showed low PIC values are: D2S441 (70.9%), D5S818 (71.03%), TH01 (66.69%), D13S317 (66.52%), and D3S1358 (66.41%), with the D3S1358 locus showing the lowest PIC value in the SA Bantu-speaking population group (Figure 3.6).

The STR loci which showed high PIC values, in the SA White population group are: The Penta_E (88.68%), D1S1656 (87.67%), D12S391 (87.47), D6S1043 (87.06%) and D2S1338 (86.77%); with the Penta_E locus showing the highest PIC value. The STR loci which showed low PIC values are: D19S433 (70.57%), CSF1PO (69.09%), D22S1045 (66.36%), D5S818 (65.04%) and TPOX (58.28%); with the TPOX locus showing the lowest PIC value in the SA White population group (Figure 3.6).

The STR loci which showed high PIC values in the SA Coloureds are: Penta_E (88.88%), D21S11 (88.22%), D2S1338 (88.73%), D1S1656 (86.34%), and D12S391 (86.72%). The Penta_E locus exhibited the highest PIC value. The STR loci which showed low PIC values are: D13S317 (71.59%), CSF1PO (71.25%), D5S818 (69.93%), TPOX (68.9%), and D3S1358 (67.18%). The D3S1358 locus showed the least PIC value (Figure 3.6).

The STR loci which showed high PIC values for the Zimbabwean Bantu-speaking group are: Penta_E (89.96%), D2S1338 (87.85%), FGA (87.31%), D6S1043 (87%), and D18S51 (85.22%). The Penta_E locus exhibited the highest PIC value. The STR loci which showed low PIC values are: D5S818 (71.24%), D2S441(70.71%), D3S1358 (69.73%), TH01 (66.35%) and D13S317 (64.79%). The D13S317 locus showed the least PIC value (Figure 3.6).

With regards to the Nigerian Bantu-speaking population group, the STR loci which showed high PIC values are: D6S1043 (86.57%), Penta_D (86.35%), FGA (86.02%), D2S1338 (85.87%) and D1S1656 (84.83%). The D6S1043 locus exhibited the highest PIC value. The STR loci which showed low PIC values are: D8S1179 (71.9%), D2S441 (69.94%), TH01 (67.39%), D3S1358 (67.15%) and D13S317 (63.4%). The D13S317 locus showed the least PIC value (Figure 3.6).

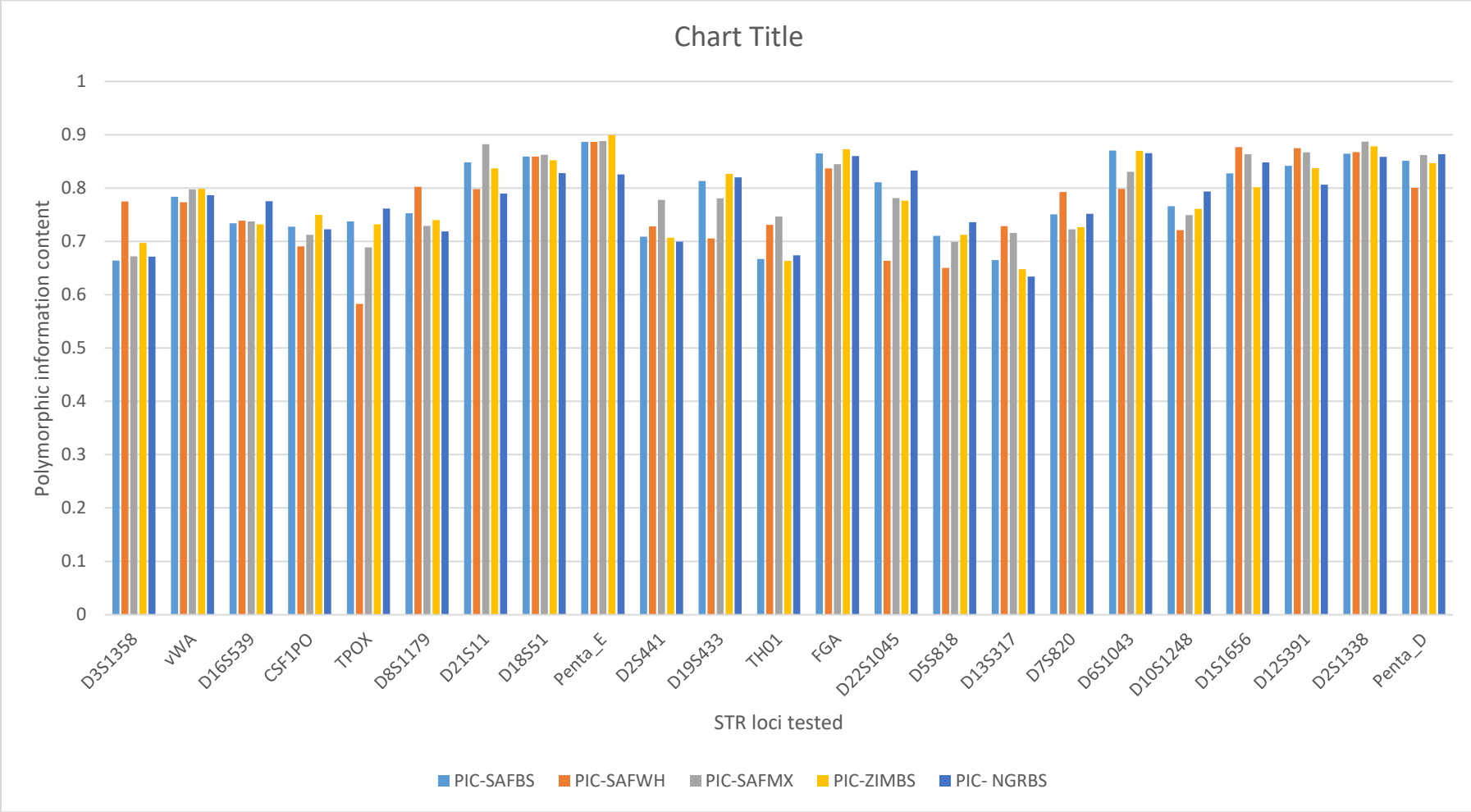


Figure 3.6 A chart showing Polymorphic information content values for all the tested STR loci, for population groups included in the present study

3.7 The ranking of the STR loci contained in the commercial kits assessed in the present study, per population group, based on the PI value

The PI values were used to rank STR loci, in an attempt to establish which loci contained in the commercial kits provide the most discrimination between unrelated individuals. Table 3.2 shows that for the SA Bantu-speakers, the Penta_E locus is most discriminative, followed by the D6S1043 locus, then the FGA locus. The least informative locus is the D3S1358 locus.

The SA Whites exhibited D1S1656 locus as most discriminative, followed by the D12S391, then the D18S51 loci. The TPOX locus is least discriminative for this population group (Table 3.2)

The D21S11 locus was shown to be the most discriminative locus in the SA Coloureds population group, followed by the Penta_E and D1S1656 loci. The D3S1358 was shown to be least discriminative locus in this population group (Table 3.2).

In the Zimbabwean Bantu-speaking population group, the FGA locus was shown to be the most discriminative locus, followed by the D6S1043 and Penta_E loci. The D13S317 locus showed the least discriminative power (Table 3.2).

With regards to the Nigerian Bantu-speakers, the D6S1043 locus was shown to be the most discriminative, followed by the Penta_D and D1S1656 loci. The CSF1PO locus showed least discrimination (Table 3.2).

Table 3.2 Ranking of STR loci contained in the commercial kits included in the present study

Locus	Ranking-SAFBS	Locus	Ranking-SAFWH	Locus	Ranking-SFAMX	Locus	Ranking-ZIMBS	Locus	Ranking-NGRBS
Penta_E	1	**D1S1656**	1	D21S11	1	FGA	1	**D6S1043**	1
D6S1043	2	**D12S391**	2	Penta_E	2	**D6S1043**	2	Penta_D	2
FGA	3	D18S51	3	**D1S1656**	3	Penta_E	3	**D1S1656**	3
D2S1338	4	**D2S1338**	4	**D2S1338**	4	**D2S1338**	4	FGA	4
D18S51	5	Penta_E	5	**D12S391**	5	Penta_D	5	**D2S1338**	5
D21S11	6	FGA	6	FGA	6	**D19S433**	6	**D19S433**	6
Penta_D	7	Penta_D	7	D18S51	7	D18S51	7	Penta_E	7
D12S391	8	D8S1179	8	Penta_D	8	**D12S391**	8	**D22S1045**	8
D1S1656	9	**D6S1043**	9	**D6S1043**	9	D21S11	9	D18S51	9
D19S433	10	D7S820	10	vWA	10	vWA	10	vWA	10
D22S1045	11	D21S11	11	**D19S433**	11	**D1S1656**	11	D21S11	11
vWA	12	vWA	12	**D22S1045**	12	**D22S1045**	12	**D12S391**	12
D10S1248	13	D3S1358	13	**D10S1248**	13	CSF1PO	13	**D10S1248**	13
D7S820	14	D13S317	14	D7S820	14	D16S539	14	D16S539	14
D8S1179	15	D16S539	15	**D2S441**	15	TPOX	15	D7S820	15
TPOX	16	**D2S441**	16	TH01	16	D8S1179	16	D5S818	16
D16S539	17	**D10S1248**	17	D16S539	17	**D10S1248**	17	TPOX	17
CSF1PO	18	TH01	18	CSF1PO	18	D7S820	18	D8S1179	18
D2S441	19	**D19S433**	19	D5S818	19	D5S818	19	CSF1PO	19
D5S818	20	CSF1PO	20	D13S317	20	**D2S441**	20	**D2S441**	20
D13S317	21	**D22S1045**	21	TPOX	21	D3S1358	21	D3S1358	21
TH01	22	D5S818	22	D8S1179	22	TH01	22	TH01	22
D3S1358	23	TPOX	23	D3S1358	23	D13S317	23	D13S317	23

3.8 Hardy-Weinberg equilibrium

The HW probability test was performed using *GenePop*, estimation of the exact *p*-values by the Markov chain method. The results are listed in Table 3.3 for each locus per population group. No departures were detected for the majority of the STR loci across all the population groups studied using the exact test (statistical significance level, $p<0.05$), with the exception of the D7S820 locus in the SA Bantu-speaking population group ($p=0.0233$), CSF1PO locus ($p=0.0298$), Penta_E locus ($p=0.0031$) and D5S818 locus ($p=0.0001$) in the SA Whites, vWA locus ($p=0.0401$) in the SA Coloureds, D18S51 locus ($p=0.0035$), Penta_E locus ($p=0.0072$) and D10S1248 locus ($p=0.0304$) loci in the Zimbabwean Bantu-speaking population group and TPOX ($p=0.0187$) and FGA ($p=0021$) loci in the Nigerian Bantu-speakers (Table 3.3).

Table 3.3 Assessment of Hardy-Weinberg equilibrium using the exact *p*-values by the Markov chain method

Locus	Population (*p*-values) per locus, significance level $p<0.05$				
	SAFBS (N=366)	SAFWH (N=84)	SAFMX (N=40)	ZIMBS (N=113)	NGRBS (N=73)
D3S1358	0.8674	0.3863	0.7038	0.6461	0.5113
vWA	0.2398	0.2905	**0.0401**	0.3158	0.9776
D16S539	0.2146	0.1506	0.7094	0.6828	0.2623
CSF1PO	0.6588	**0.0298**	0.6973	0.0884	0.1679
TPOX	0.2472	0.2334	0.9653	0.9627	**0.0187**
D8S1179	0.8630	0.1142	0.3873	0.2715	0.2744
D21S11	0.6550	0.5373	0.4001	0.1478	0.9916
D18S51	0.3331	0.2280	0.3051	**0.0035**	0.1823
Penta_E	0.2513	**0.0031**	0.7798	**0.0072**	0.4402
D2S441	0.4099	0.5285	0.2014	0.4997	0.4139
D19S433	0.3176	0.2579	0.1863	0.6134	0.9980
TH01	0.7003	0.9295	0.2914	0.4892	0.5768
FGA	0.3327	0.7000	0.7281	0.8518	**0.0021**
D22S1045	0.9913	0.6391	0.8454	0.1564	0.2139
D5S818	0.7441	**0.0001**	0.8304	0.9762	0.6443
D13S317	0.4434	0.8761	0.4327	0.9057	0.6955
D7S820	**0.0233**	0.3043	0.8579	0.9146	0.8712
D6S1043	0.9631	0.6069	0.0521	0.9402	0.1819
D10S1248	0.3369	0.9588	0.7396	**0.0304**	0.3752
D1S1656	0.1014	0.6273	0.8709	0.2934	0.4096
D12S391	0.2606	0.1450	0.7496	0.1759	0.3150
D2S1338	0.4140	0.2354	0.1685	0.5933	0.1294
Penta_D	0.1860	0.8261	0.3247	0.0611	0.6031

-The *p*-values where deviation from HWE was observed at a locus are indicated in bold.

Global HW tests were also performed to assess the presence or absence of heterozygosity deficiency and heterozygosity excess for each population group. The default Markov chain parameters were used for all the tests: Dememorization-1000, Batches-100 and Iterations per batch- 1000.

When analysing HWE per population under two independent hypotheses, H1: assuming heterozygosity deficiency and H2: assuming heterozygosity excess, no departures were observed for H2 (assuming heterozygosity excess). Departures from HWE were observed for heterozygosity deficiency in the SA and Nigerian Bantu-speakers, (p=0.0174) and (p=0.0049), respectively (Table 3.4).

Table 3.4 Global Hardy-Weinberg tests- results by population, when H1= heterozygote deficit and when H2= heterozygote excess (multi-locus test)

Population	*p*-value (HWE)- H1= Heterozygosity deficit	*p*-value (HWE)- H2= Heterozygosity excess
SAFBS (N=366)	**0.0174**	0.9826
SAFWH(N=84)	0.0506	0.9494
SAFMX(N=40)	0.4192	0.5808
ZIMBS(N=113)	0.1076	0.8924
NGRBS (N=73)	**0.0049**	0.9951

The *p*-values where deviations from HWE were observed are indicated in bold, significance level $p<0.05$

3.9 Assessment of linkage disequilibrium

Linkage disequilibrium analysis was conducted on the genotype results of all the samples belonging to the five ethnic groups evaluated in the present study. The *GenePop*, module 2 for analysing genotypic linkage disequilibrium was used. The results, consisting of *p*-value pairwise analysis of all loci, can be found in Appendix E. The following LD associations, were observed between pairs of loci across the different ethnic groups: 16 inter-chromosomal pairs showed LD ($p<0.05$) in the SA Bantu-speaking population group, 21 in the SA Whites, seven in the SA Coloureds, eight in the Zimbabwean Bantu-speakers and 35 in the Nigerian Bantu-speakers.

3.10 Genetic variations between the population groups

Genetic variations between the five population groups in the present study were evaluated using GenePop v4.7.5. This was accomplished by calculating pairwise F_{sts}, using one locus estimates and following standard ANOVA. Table 3.5 shows the calculated outcomes. The F_{st} values indicate that the genetic variation between SA Bantu-speakers, Zimbabwean Bantu-speakers and Nigerian Bantu-speakers versus the SA Whites population group is about 3%. The SA Whites and Coloureds groups show a genetic variation of ~1%. SA Bantu-speakers and Coloureds show variation of 0.5%. The genetic variation between SA Bantu-speakers and Zimbabwean Bantu-speakers was 0.05 % (Table 3.5).

Table 3.5 Pairwise F_{st} values between population groups for all loci

	SAFBS (n=366)	**SAFWH (n=84)**	**SAFMX (n=40)**	**ZIMBS (n=113)**
SAFWH (n=84)	0.0328 (3.28%)	-	-	-
SAFMX (n=40)	0.0058 (0.58%)	0.0131 (1.31%)		-
ZIMBS (n=113)	0.0005 (0.05%)	0.0316 (3.16%)	0.0041(0.41%)	-
NGRBS (n=73)	0.0031 (0.31%)	0.0345 (3.45%)	0.0074 (0.74%	0.0026 (0.26%)

* Estimates following standard ANOVA discussed by Weir and Cockerham (1984)

3.11 Retesting and evaluation of kinship cases previously tested using PowerPlex® 16 HS or PowerPlex® 18D

To assess the utility of the additional markers contained in the newer commercial kits and the validity of outcomes obtained when fewer loci were used, a total of 63 scenarios (family relationships), whose genotypes were originally generated using a commercial kit containing 15 autosomal STR loci, had their genotypes regenerated using commercial kits with 23 polymorphic STR loci and the results are shown in Table 3.6.

Table 3.6 Comparison of outcomes when 14 and 23 polymorphic STR loci were used

Number of reassessed cases (n=63)	Outcomes obtained when 5 STR loci were used	Outcomes obtained when 23 STR loci were used	
17/63 (27%)	Not excluded	16/17 (94%) remained not excluded, and had an improved Likelihood Ratio	1/17 (5.9%) changed to inconclusive
3/63 (5%)	Excluded	1/3 (33.3%) remained excluded	- 2/3 (67%) changed to inconclusive
43/63 (68%)	Inconclusive	-26/43 (60%) remained inconclusive. -Of this 26/43, 16/23 had an increased Likelihood Ratio and -10/23 had a decreased LR, all within the inconclusive range.	- 7/43 (16%) changed to excluded - 10/43 (23%) changed to NOT excluded

Of the 63 scenarios that were retested, 17/63 cases were initially reported as not excluded. Following retesting, the outcome for 94% (16/17), remained the same, and the Likelihood Ratio value increased in contrast to when 15 polymorphic markers were used. Of these 17 cases, 5.9% (1/17) changed from not excluded, to inconclusive. The initial test showed a Likelihood Ratio of 209.22. When retested with VeriFiler™Express, the Likelihood Ratio obtained was 30.43.

A total of 3/63 cases were initially reported as excluded (not related) when 15 polymorphic STR loci were used. The LRs observed were 0.0006 (alleged full siblings), 0.0004 (alleged full siblings) and 0,026 (alleged full siblings). Subsequent to retesting, 33.3% (1/3) remained excluded (0.0006 changed to 3.04 $x10^{-5}$). However, 67% (2/3) changed to inconclusive (0.0004 changed to 0.80 and 0.03 changed to 0.2817).

In the final category, 68% (43/63) of cases that were initially reported as inconclusive, showed the following results: Sixty percent (26/43) of the cases remained inconclusive. Sixteen percent (7/43) cases changed to excluded and 23% (10/43) of the cases changed to not excluded.

3.12 Comparison of Likelihood Ratios when different data analysis software packages were used

A comparison of Likelihood Ratios calculated using 4 different software packages was conducted using the In-house software, *Converge™ software*, *GeneMarker® HID* and *Familias* software packages. The *p*-values obtained when using the Student's T-test, paired two samples for means method are listed in Table 3.7. Student's T-test calculations were performed using Microsoft Excel 2016.

Table 3.7 *P*-values for Student's T-test conducted to compare Likelihood Ratio outcomes obtained using four different data analysis software packages

N=63	***P*-values (at a significance level of *p*<0.05)**		
	In-house software	***Familias***	***GeneMarker® HID***
Familias	0.331	-	-
GeneMarker® HID	0.372	0.386	-
Converge™ Software	0.366	0.347	0.389

The null hypothesis (H_o) was that there was no significant difference in the CLR outcomes obtained when the different data analysis software packages were used particularly for kinship testing. The alternative hypothesis (H_1) was that there was a significant difference in the CLR outcomes obtained when the different data analysis software packages were used. A significance level of $p<0.05$ was chosen. The following *p*-values were obtained: In-house vs *Familias* (p=0.331), In-house vs *GeneMarker™ HID* (p=0.372), In-house vs *Converge™* Software (p=0.366), *Converge™* Software vs *Familias* (p=0.389) and lastly, *GeneMarker™ HID* and *Familias* (p=0.386).

The calculated *p*-values when comparing all the data analysis software packages was >0.05. Therefore, the In-house data analysis software was deemed fit for use in routine diagnostic testing.

CHAPTER 4 DISCUSSION

4.1 Outcomes of allele frequency calculations

Local population databases are generally compiled to keep a record of allele frequencies that are specific to different population groups of interest. Allele frequencies for polymorphic markers are determined in order to identify all common and rare alleles in a given population group. The allele frequencies and statistical data were calculated using the counting method and confirmed using an online Microsoft Excel macro enabled software (Duewer, 2016). Random sampling helps to determine average alleles observed in a population group, as the entire population cannot be sampled (Butler, 2005). We successfully calculated allele frequencies for five population groups; the SA Bantu-speakers, SA Whites, SA Coloureds, Zimbabwean Bantu-speakers and Nigerian Bantu-speakers. This information will be used to update the current databases in DPL and aid in analysis of complex kinship testing requests. The allele frequencies calculated in the present study are comparable with similar studies previously conducted in the SA population groups for overlapping STR loci (Fujitani et al., 2003; Lucassen et al., 2013). The allele frequencies calculated in a study previously done on Africans and Europeans from South Africa by Kido et al., 2007, also showed similar allele frequencies for the overlapping STR loci tested.

For kinship testing, we evaluated whether the additional STR loci, found in two commercially available kits, would help to resolve complex queries, some of which were previously unresolved when 15 STR loci were utilised. This was based on the premise that the additional markers would increase the discrimination power of the newer commercial kits.

4.2 Number of alleles observed

Our data showed the FGA locus as having the highest number of alleles in the SA Bantu-speaking population group followed by the D21S11, D6S1043, Penta_E and Penta_D loci. When compared with the other Bantu-speaking groups included in the present study, the Zimbabwean Bantu-speaking population group also showed that the FGA locus as having the highest number of alleles followed by the D18S51, Penta_E, D6S1043 and D12S391 loci. An overlap of three loci between the SA and Zimbabwean Bantu-speaking groups was observed, showing the FGA, Penta_E and D6S1043 loci as having highest number of alleles in common between the two groups. The Nigerian Bantu-speaking population group exhibited the D21S11 locus as having the highest number of alleles, followed by the Penta_E, FGA, D1S1656 and Penta_D loci. An overlap of four loci, showing a common high number of alleles between the Nigerian and the SA Bantu-speaking groups was observed, at the FGA, D21S11, Penta_E and Penta_D loci. An overlap of two loci showing a high number of alleles was observed at the FGA and Penta_E loci when the Zimbabwean and Nigerian Bantu-speaking population groups were compared. Four loci were noted showing the least number of alleles in common between the three Bantu-speaking groups included in the present study: D3S1358, D16S539, THO1 and D13S317 loci. The SA and Zimbabwean groups exhibited two more loci in common as having the least number of alleles: The TPOX and D8S1179 loci. The Zimbabwean and Nigerian Bantu-speaking groups showed one more locus in common with the least number of alleles: The D5S818 locus.

The Penta_E and FGA loci showed high number of alleles in the SA White and SA Coloured population groups. The D1S1656 locus showed a high number of alleles in the SA Whites, SA Coloureds and Nigerian Bantu-speakers. The D12S391 locus showed a high number of alleles in the SA White, SA Coloured and the Zimbabwean Bantu-speaking groups.

The THO1 locus showed the least number of alleles in all the population groups assessed in the present study. The D3S1358 locus showed the least number of alleles in four of the five population groups assessed, with the exception of the SA Whites population group. The TPOX locus showed the least number of alleles in four of the five population groups, with the exception of the Nigerian Bantu-speaking group.

A high number of alleles observed at a locus is associated with a high locus variability. D18S51 locus was previously reported to exhibit the highest locus variability in the CODIS 13 pool of STR loci, and the TPOX locus was reported to have the least number of alleles in the following United States (US) populations: Caucasians, African American, Asian and Hispanic (Butler, 2012). The finding on the D18S51 and TPOX loci was different from the observations noted in the present study, particularly for four out of five population groups, with the exception of the SA Whites, which showed concordance with the US findings. With regards to the other four population groups, the D18S51 locus showed high, but not the highest variability among the CODIS 13 pool, and TPOX showed low, but not the lowest variability.

4.3 Number of observed genotypes and heterozygosity

The number of genotypes at an STR locus determines its heterozygosity value. The fewer the number of genotypes observed at a locus, the more likely that there will be a low heterozygosity value and the higher the probability that unrelated individuals will have the same genotype at that locus. The higher the heterozygosity value, the higher the allelic diversity at that locus. This then implies that a locus is more informative as there is a lower chance of seeing a random match between unrelated individuals within a population.

Our data showed that with respect to the SA Bantu-speakers, the D21S11 locus exhibited the highest number of genotypes (91) and a heterozygosity value of 86.89%. Interestingly, the D6S1043 locus displayed 72 genotypes and a heterozygosity value above that of the D21S11 locus, 90.60%, which is the highest in this group. This observation suggests that even though the number of genotypes was higher for the D21S11 locus, the allele frequencies across the observed genotypes are not as evenly distributed as those for the D6S1043 locus (Butler, 2012). The statement above was confirmed when the TPOX and D13S317 loci both showed 21 observed genotypes, but different heterozygosity values, 83.67% and 68.58%, respectively. The CSF1PO locus showed the least number of observed genotypes (9), but showed a heterozygosity value of 77.05%, which is higher than that of the D13S317 locus (68.58%), which exhibited 21 observed genotypes.

With respect to the SA Whites, the Penta_E locus showed 42 observed genotypes, and the highest heterozygosity value of 90.48%, which is comparable to the American Caucasian heterozygosity value of 89.96% (Butler, 2012). The D1S1656 locus showed the highest number of observed genotypes (45), and a heterozygosity value similar to that of the Penta_E locus: 90%. The TPOX locus showed the lowest heterozygosity value of 54.76% and was reportedly the lowest in the Caucasian group with a heterozygosity value of 69.02%, according to Butler et al., 2012.

The following observations were noted in the SA Coloureds, the D12S391 (28) and Penta_E (32) loci showed the same (and highest) heterozygosity value of 92.50%, despite the different number of observed genotypes, followed by the D1S1656 (29), D18S51 (26), FGA (26) and Penta_D (26) loci, all with a heterozygosity value of 90%. The D3S1358 and vWA loci both exhibited the lowest heterozygosity value of 70%, despite different number of observed genotype, (13) and (19), respectively.

With regards to the Zimbabwean Bantu-speakers, the Penta_E locus showed the highest number of genotypes (57) and the highest heterozygosity value of 92.92%. The D13S317 locus exhibited 15 observed genotypes, and the lowest heterozygosity value of 65.49%.

With respect to the Nigerian Bantu-speaking population group, the Penta_E and Penta_D loci both exhibited the highest number of genotypes of (37), however, the Penta_D locus showed the highest heterozygosity value of 89.19% followed by D2S1338 locus with (34) observed genotypes and a heterozygosity value of 87.84%. The Penta_E locus showed a heterozygosity value of 81.08%, despite having the highest number of genotypes. The D13S317 locus showed the lowest heterozygosity value of 68.92%, and the same number of genotypes (12) as the D3S1358 locus, but the D3S1358 locus showed a heterozygosity value of 74.32%.

These observations point to a non-linear relationship between the number of genotypes and the heterozygosity values. The even distribution of alleles is preferred over a high number of genotypes, as observing a high frequency of common alleles and few rare alleles lowers the genetic variability of a locus. It is better to have alleles frequencies evenly distributed across a locus in order to attain a high heterozygosity value (Butler, 2015).

4.4 Probability of Identity

The lower the PI value, the more variable the STR locus is in a population because the PI value gives an indication that there are more genotypes occurring at a lower frequency, as discussed in Butler et al., 2012. Previous publications have suggested that the PI value is a better measure of a locus' variability than the total number of observed alleles and genotypes because specific common alleles may occur in a relatively high frequency and reduce the overall variability, especially if the number of rare alleles occur at that locus (discussed in Butler, 2012). Ideally, observing a fairly

even level of variation across many genotypes at each locus is desirable, as there will be a greater chance of finding a difference between two unrelated individuals selected at random.

The probability of identity values generated using the different commercial kits: PowerPlex® 16 HS or PowerPlex® 18D, PowerPlex® Fusion and VeriFiler™ Express, across all population groups included in the present study were found to be comparable with published data (Butler, 2012), and are presented in Table 3.1 and Table 1.7. The outcomes of the PI values showed that the VeriFiler™ Express PCR amplification kit exhibited the most discrimination of all the commercial kits assessed in the present study, as it contains the most number of polymorphic autosomal STR loci (23), compared to the PowerPlex® Fusion system, which contains 22 polymorphic autosomal STR loci. The PowerPlex® 16 HS and the PowerPlex® 18D contain 15 and 17 polymorphic autosomal STR loci respectively.

Figure 3.5 shows the PI values of all the STR loci contained in the VeriFiler™ Express PCR amplification kit across the five population groups included in the present study. The level of variability of these loci in these population group is comparable with their observed variability in the US African American group (Butler, 2012).

Our data showed the following PI values for the SA Bantu-speaking group: Penta_E displayed the highest variability, followed by the D6S1043, FGA, D2S1338 and D18S51. The D2S1338 locus showed high variability across all the population groups tested. The Penta_E locus showed high variability across four out of the five population groups, with the exception of the Nigerian Bantu-speaking group. The Penta_E, D6S1043 and FGA showed high variability among the Bantu-speaking population groups. The D1S1656 and D12S391 showed high variability among the SA Whites, Coloureds and Nigerian Bantu-speaking groups. The D21S11 locus showed the highest variability in the SA Coloureds group, but not in the other four

population groups evaluated in the present study, despite its high number of observed genotypes in the latter.

The D2S441 and D3S1358 showed low variability across all the Bantu-speaking groups included in the present study. The D5S818 locus showed low variability among the three out of five population groups, with the exception of the SA Whites and Nigerian Bantu-speaking groups. The TPOX locus showed low variability in the SA Whites and SA Coloureds.

4.5 Polymorphic Information Content and ranking of STR loci

The PIC is closely related to the PI. A high PIC value indicates a more variable and polymorphic locus, and is associated with a low PI. The relationship is however not always consistent as seen between the Penta_D and D21S11 loci in the SA Bantu-speaking population group. The Penta_D locus showed a PI value of 0.0366 and a PIC value of 85.15%, and was ranked number 7 (with regards to level of variability) based on its PI value. The D21S11 locus showed a PI value of 0.0332 and a PIC value of 84.82% and was ranked number 6. The PI value for the D21S11 locus was less than that of the Penta_D locus, indicating that the D21S11 locus displayed more variability than the Penta_D locus. When the PIC values of these two STR loci were compared, the Penta_D locus showed a higher PIC value than the D21S11 locus.

The STR loci which showed the highest PIC values in the SA Bantu-speakers were Penta_E, D6S1043, FGA, D2S1338 and D21S11. The STR loci which showed the least PIC values in this population group were D3S1358, TH01, D13S317, D5S818 and D2S244.

With regards to the SA Whites, the STR loci which showed the highest PIC values were Penta_E, D1S1656, D12S391, D2S1338 and D18S51. The least PIC values were observed in the TPOX, D5S818, D22S1045, CSF1PO and D19S433 loci. The degree of variability of the STR loci in the SA Whites population group was comparable with those observed in the US Caucasian and European groups (Butler, 2012; Kido et al., 2007).

The STR loci with highest PIC values in the SA Coloured group were D21S11, Penta_E, D1S1656, D2S1338 and D12S391. The STR loci showing the least PIC values were D5S818, D13S317, TPOX, D8S1179 and D3S1358. The Penta_E, D2S1338 and D12S391 loci were among the STR loci with high PIC values in the SA Bantu-speakers and the SA Whites. The D5S818, D13S317 and D3S1358 STR loci also showed the least PIC values in the SA Bantu-speakers. The TPOX and D5S818 loci were also among STR loci with the least PIC values in the SA Whites ethnic group. The trend seen in the SA Coloured population group with regards to the above mentioned STR loci is consistent with this group being admixed.

The STR loci with highest PIC values in the Zimbabwean Bantu-speakers were FGA, D6S1043, Penta_E, D2S1338 and Penta_D. The STR loci with the least PIC values were D5S818, D2S441, D3S1358, TH01 and D13S317. The FGA, D6S1043, Penta_E and D2S1338 loci were also among the STR loci with highest variability in the SA Bantu-speakers and the STR loci with least PIC values in this group also matched those observed in the SA Bantu-speakers.

The STR loci with highest PIC values in the Nigerian Bantu-speakers were D6S1043, D1S1656, FGA, D19S433 and Penta_E. The STR loci with least PIC values were D3S1358, TH01, D13S317, D5S818 and D2S441. The D6S1043, FGA and Penta_E loci were also among the STR loci with highest PIC values in the SA Bantu-speakers. The D3S1358, TH01, D13S317 loci were also among the STR loci with the least PIC values in the SA and Zimbabwean Bantu-speaking groups. The D1S1656 and Penta_E loci also showed the highest PIC values in the SA Whites and Coloureds.

The D19S433 locus showed high variability in the Bantu-speaking groups and a lower PIC value in the SA Whites.

Using the PI to determine an STR locus' level of variability showed that number of alleles, number of observed genotypes and observed heterozygosity values were not good indicators for evaluating STR locus variability. PI and PIC are better indicators and their outcomes are comparable, with minor discrepancies. The differences in STR locus variability observed when comparing the locus variability across different population groups are comparable with those observed by Fujitani et al., 2003; Kido, et al., 2007, Butler 2012 and Lucassen et al., 2013.

The Penta_E and D2S1338 loci showed high PIC values, while the D5S818 showed low PIC values across all five population groups included in the present study. The D6S1043 and FGA loci showed high PIC values, while the D3S1358, TH01, D13S317, D5S818 and D2S441 showed the least PIC values among the Bantu-speaking groups.

4.6 Assessment of Hardy-Weinberg equilibrium

The HW probability test showed that departures were observed in the D7S820 locus in the SA Bantu-speaking population group (p=0.0233). The following departures from HWE were observed in the SA Whites population group: CSF1PO (p=0.0298), Penta_E (p=0.0031) and D5S818 (p=0.0001). In the SA Coloureds group, departures form HWE were observed in the vWA locus (p=0.0401). The Zimbabwean Bantu-speaking population group showed departures from HWE in the D18S51 (p=0.0035), Penta_E (p=0.0072) and D10S1248 (p=0.0304) loci. The TPOX (p=0.0187) and FGA (p=0021) loci in the Nigerian Bantu-speakers also showed departures from the HWE (Table 3.3). The significance of these findings is not known as further tests, which make use of a correction factors, were not performed to confirm the validity of these observations.

When the population groups in the present study were assessed for Heterozygosity excess, no departures from HWE were observed. Departures from HWE were observed in the SA- and Nigerian Bantu–speaking population groups, when assessed for heterozygosity deficiency (Table 3.4). This type of departure from HWE was reported previously in other studies to be associated with mutations, genetic drift, inbreeding, selection and population substructure (Chen et al., 2017).

A further analysis should be performed to establish if the observed departures represented true departures from HWE. Some studies report that the use of a correction factor, such as the *Bonferonni* approach for data involving STR loci, provide clarification of departures from HWE (Moretti et al., 2016; Ramos-González et al., 2016; Wang et al., 2013) . A correction factor was not used in the present study, as a result, the observed departures from HWE could not be validated, or classified into major or minor, and the source of the departures could not be established. Major departures from HWE are due to selection, inbreeding, population substructure and genotypic errors (Chen et al., 2017).

4.7 Assessment of linkage disequilibrium

Linkage disequilibrium, which is non-random association of STR loci was assessed. The STR loci included in the PowerPlex® Fusion and the VeriFiler™Express commercial kits are located on different chromosomes or on different arms of the same chromosome, and some, on the same chromosome arm. The expectation from this analysis is that LD should not be detected between STR loci located on different chromosomes. The implication, with regards to DNA profiling, is that STR loci that are in LD should not both be included in the product rule when calculating Likelihood Ratios for proposed biological relationships (O' Connor et al., 2010, Butler and Hill, 2013). In the present study, LD was observed between the D21S11 and Penta_D loci, both of which are on chromosome 21, and are 24.5 mega bases (Mb) apart (Butler and Hill, 2013). However, previous studies have established that there is no LD between these loci (Butler and Hill, 2013). In the present study, LD was not

observed between the vWA and D12S391 loci, both of which are on chromosome 12, and are 6.3 Mb apart. A study by O' Connor et al., 2010 indicated that these two loci are actually in LD. The following syntenic STR loci are included in the commercial kits evaluated: vWA-D12S391 (chromosome 12p, 6.3 Mb apart), D21S11-Penta_D, (chromosome 21q, 24.5 Mb apart), D5S818-CSF1PO (chromosome 5q, 33.4 Mb apart) and TPOX-D2S441 (chromosome 2p, 66.6 Mb apart) (Butler and Hill, 2013).

Surprisingly, LD was erroneously detected between STR loci that are located on different chromosomes (Appendix E). Wang et al illustrated that loci on different chromosomes can show LD because of population substructure or natural selection. Several studies observed similar outcomes when assessing LD on different population groups (Moretti et al., 2016; Ramos-González et al., 2016; Wang et al., 2013).

4.8 Genetic variation between population groups

There is supporting evidence stated by Rosenberg et al., 2002 that genetic differences between major population groups account for 3-5%. SA Whites and Coloureds groups showed a genetic variation of ~1%. SA Bantu-speakers and Coloureds showed a genetic variation of 0.5%, suggesting that the admixture in the samples tested was predominantly comprised of Bantu-speakers than the SA Whites populations. The genetic variation between SA Bantu-speakers and Zimbabwean Bantu-speakers was 0.05 %. This suggests that there is little evidence of genetic differentiation between the two population groups, and further supports that the same allele frequency data can be used across the two population groups when calculating CLRs for family relationship testing. The genetic variation between SA and Zimbabwean Bantu-speaking population groups versus that of the Nigerian Bantu-speakers was shown to be approximately 0.3 %. This suggests that there is evidence of genetic differentiation, and calculation of LRs should take into account allele frequencies of both population groups. The differences observed between the two groups could be attributed to the different times that the groups migrated towards the

southern part of Africa. The SA and Zimbabwean Bantu-speakers migrated from the Western Central Africa more than 3 centuries ago and have been living in close proximity to each other. The Nigerian Bantu-speakers only migrated towards the southern part of Africa in recent years. The differences observed can thus be attributed to population substructure and genetic drift. Thus, the SA/Zimbabwean Bantu-speaking population groups cannot be clustered together with the Nigerian Bantu-speaking population group for the purpose of using the same allele frequencies for family relationship testing. The same conclusion applies to SA Bantu-speakers versus SA Coloured and White population groups.

4.9 Self-declared ethnic groups and relatedness between individuals included in the study.

Unfortunately, ethnic groups could not be confirmed in the present study using Ancestry –informative markers to aid in confident clustering of the population groups as this was beyond the scope of the study. As a result of this, there could be skewing of frequencies obtained. It is also important to note that relatedness between individuals studied could not be completely eliminated using lineage markers. This may cause skewness in the allele frequencies observed, because of people identifying their ethnic groups differently, often socially and not necessarily biologically. This could have contributed to the observed departures from HWE. Ideally, only unrelated individuals should be tested. Despite these limitations, we are happy with the allele frequencies obtained, as they are compare well with published data.

4.10 Reassessment of kinship cases previously tested with PowerPlex® 16HS and PowerPlex® 18D systems

A threshold Likelihood Ratio of 100 is used in DPL as an accepted level where the statistical evidence is believed to support a true biological relationship, based on genotypic profiles of the tested individuals. A Likelihood Ratio of between 0.1 and 99 is classified as inconclusive. This means that further testing is required to unequivocally confirm the relationship, as 25% of cases in this category have been shown to be unrelated, following supplementary testing or testing of additional family members.

An outcome observed for one of the families retested showed the following: the initial testing showed a Likelihood Ratio of 209.22. When retested with VeriFiler™Express kit, the Likelihood Ratio obtained was 30.43. The outcome was similar when the Likelihood Ratio was calculated using different software packages. The social pedigree in this particular case, indicated the relationship between the two individuals as child and paternal aunt. The In-house data analysis software only does calculations for full aunts, and the initial outcome was based on this calculation. When reviewing the two profiles, the observation noted was that the outcome from 15 autosomal STR loci showed 5 mismatches (5/15), which is a 33% mismatch. When using 23 STR loci, 10/23 mismatches were detected, resulting in a 43% mismatch, confirming the reduced Likelihood Ratio. Analysis using different possible scenarios showed a Likelihood Ratio of 66.28 when the individuals' DNA profiles were analysed as child and alleged half aunt (*GeneMarker™ HID* and *Familias*). This means that the tested pair were more likely child and paternal half-aunt as opposed to child and full aunt. The probable paternal half aunt happened to have more in common with the half-niece by chance, resulting in a high probability when fewer STR loci were used, skewing the outcome towards false positive. It must be noted that initially, an incorrect calculation was done, based on the wrong proposed relationship. The reassessment process indicated that the use of more loci (and correct scenario calculations) decreased the number of false positives as compared to using fewer loci.

A total of 3/63 cases were initially reported as excluded (not related) when 15 polymorphic STR loci were used. The LRs observed were 0.0006 (alleged full siblings), 0.0004 (alleged full siblings) and 0,026 (alleged full siblings). Subsequent to retesting, 33.3% (1/3) remained excluded (0.0006 changed to 3.04 $x10^{-5}$). However, 67% (2/3) changed to inconclusive (0.0004 changed to 0.80 and 0.03 changed to 0.2817). This finding showed that using more STR loci decreased the number of false negatives. Also, this finding highlighted that the number of false negatives in DPL was; high (67%) observed in the present study. This lead to the suggestion that all negative kinship testing outcomes (where there is insufficient DNA evidence to support the biological relationship) should be accompanied by supplementary testing to exclude false negatives.

When inconclusive outcomes, obtained when the samples were tested with the commercial kits with 15 autosomal STR loci, 16% of the initial outcomes changed from inconclusive to excluded, and 23% of the initial outcomes changed from inconclusive to not excluded. This finding showed that the use of more loci significantly improved the rate of unequivocal assignment of relationships. Of the 68% inconclusive outcomes obtained in the present study, 41% remained inconclusive, 11% changed to excluded and 16% changed to not excluded. This means that the number of inconclusive outcomes improved from 68% to 41%, showing a 27% improvement, based on the use of additional STR loci only.

It is important to note that there was a huge overlap in the Likelihood Ratio values of cases that changed from inconclusive to excluded and the ones that changed from inconclusive to not excluded, showing that a clear cut-off value cannot be established based on the Likelihood Ratio values only. Some studies have divided the inconclusive range into the following subdivisions: 1) A LR of 1 to 10 was classified as limited support of the relationship being true, 2) 10 to 100 moderate support 3) LR values above 100 were classified as having strong support and 4) LR values more than 1000 classified as very strong support (Butler, 2005). We recommend the use of additional testing, using either lineage markers or testing additional family members to unequivocally assign relatedness to proposed relationships.

Based on the current testing strategy in DPL, for all the cases with an initial outcome in the inconclusive range, additional testing would be recommended, either using lineage markers or testing additional family members. This is currently done in a tiered approach, and at an additional cost. The use of extended STR kits showed the ability to resolve the outcomes from inconclusive to excluded or not excluded, without the need for further testing in 27% of the time, and eliminated the need for further testing in those scenarios. It is therefore our recommendation that all kinship cases would benefit from using 23 polymorphic STR loci as the first line of testing approach. Approximately 27% of the previously inconclusive outcomes would be resolved with the use of the PowerPlex® Fusion system or the VeriFiler™ Express PCR amplification kit instead of using 15 STR loci. The cost of the commercial kits gives bias to DPL to choose the VeriFiler™ Express PCR amplification kit as the primary testing kit, as it costs less per reaction (R188) compared to the PowerPlex® Fusion system (R472), (see kit prices in Table 1.7). Moreover, the VeriFiler™ Express PCR amplification kit contains an additional STR locus; D6S1043, which was shown in the present study to have high variability in different population groups (Table 3.2), resulting in the kit's higher discrimination power compared to the PowerPlex® Fusion system. It must however be noted that in some instances, lineage markers or testing of additional family members may not be applicable, in which case, no further recommendations would be given.

4.11 Comparison of Likelihood Ratios when different data analysis software packages were used

The null hypothesis, stating that there is no difference in the Likelihood Ratios obtained when different data analysis software packages was proved to be true. The calculated *p*-values across all the different analysis software packages were more than 0.05, thus supporting the null hypothesis. *GeneMarker® HID* and *Familias* software packages have been cited in many studies, and most researchers have confidence in these software packages because they have been adequately validated for calculation of Likelihood Ratios in complex kinship queries (Drábek, 2009).

Familias was found to have extensive capability to simultaneously analyze data from multiple people in a family. Testing of a mother, multiple children and a relative of the alleged father can now be done, assuming different permutations, using *Familias*. This could not be carried out previously using the In-house software which mainly calculates pairwise LRs for kinship testing. The use of the *Familias* for calculating LRs when testing families with multiple individuals has the potential to further decrease the inconclusive rate, as testing multiple people has been shown to improve unequivocal assignment of relationships between individuals in our laboratory.

GeneMarker® HID showed a limited capability in that only pairwise comparisons can be performed. Similarly, the In-house software calculates pairwise comparisons for full and half siblings, uncle-nephew, grandparent grandchild relationships (these relationships share 25% of DNA). Also, when using the In-house data analysis software, the only calculation that takes into account more than two people in kinship testing is that which incorporates testing between the mother, child and a relative of the alleged father (such as sibling, parent or child). *Familias, Converge™ Software* and *GeneMarker® HID* software packages can do calculations for more distant relationships compared to the In-house software. *Converge™ Software* and *GeneMarker® HID* software packages can test up to third degree relatives (i.e. relationships sharing about 12.5% of DNA). *Familias*' capability to test relationships beyond third degree relationships has not been explored. When using *Familias*, it has been demonstrated that testing more people in a family improves the chance of proving or disproving a relationship, thus improving the rate of assigning unequivocal relationships to claimed relationships (Drábek, 2009).

During the assessment process, it was very important that the same population data (allele frequencies and mutation rates) be used across all the different data analysis software packages, otherwise, the outcomes were incomparable. This supports the importance of compiling adequate population data for different population groups in order to draw appropriate conclusions during analysis.

Following the generation of population data, a new analysis software was constructed, which included the additional STR loci, their allele frequencies and mutation rates for the different population groups included in the present study. Following this evaluation, we validated the DPL In-house analysis software. This software package will be used in the analysis of LRs testing whenever the VeriFiler™ Express kit was used to generate DNA profiles.

CHAPTER 5 CONCLUSION

The aim of the present study was to evaluate the utility of the additional polymorphic STR loci present in the evaluated commercial kits (eight in the VeriFiler™ Express PCR amplification kit and seven in the PowerPlex® Fusion system). We successfully generated population data (allele frequency data) for five population groups and used it to evaluate the genetic variability of each STR locus and how they performed collectively. The probability of identity obtained for VeriFiler™ Express and PowerPlex® Fusion provided evidence that the inclusion of additional markers improves the discrimination power of these commercial kits, compared to kits containing fewer STR loci.

The genetic variations between the population groups studied show that the SA Bantu-speakers and the Zimbabwean Bantu-speakers may be clustered together as there is no evidence of genetic differentiation between them (F_{st} ~=0). The SA Coloureds group tested shows evidence of being comprised predominantly of SA Bantu-speakers than SA Whites, with an estimated ratio for the group admixture of 2 to 1. The Nigerian Bantu-speakers showed evidence of genetic differentiation when compared to the SA and Zimbabwean Bantu-speakers. The genetic differentiation between SA Whites and the Bantu-speaking groups was about 3%, supporting the published data that genetic variation between major population groups ranges between 3-5%.

One of the objectives was to retest and reassess kinship cases with inconclusive outcomes using commercial kits with expanded number of STR loci. The high inconclusive rate (approximately 68%), previously observed in DPL was shown to be reduced by 49% based on the use of additional STR loci only. This information will be used to amend the current testing strategy for kinship analysis. Going forward, all kinship requests will be processed using the VeriFiler™ Express PCR amplification kit as the primary testing kit. Retested historical requests where the outcome changed when the relationships were re-analysed using the VeriFiler™ Express PCR

amplification kit, amended reports will be compiled, and the results will be communicated to the relevant families and their referring authorities, where applicable. We also recommend that the current testing strategy for kinships be refined as follows, to further reduce the number of inconclusive outcomes: Assess suitable family members available for testing, establish the applicable supplementary test, test the samples using the VeriFiler™ Express kit, and perform the confirmatory supplementary test as a standard approach.

While the present study has provided evidence of the benefit of the additional STR loci, it has also highlighted that the validity of outcomes obtained when fewer loci are used, particularly with regards to kinship testing, is questionable. This finding necessitates the need to reassess all the kinship cases previously tested and reported on using 15 STR loci only, as some false positive and false negative outcomes were detected in the present study. The presence of false positives may be attributed largely to sharing of alleles by chance, and the presence of false negatives may be attributed to the presence of mutations, null alleles and allelic dropout during PCR amplification (Wenk et al., 2003). Both these processes pose a serious problem when kits with fewer loci are used for complex kinship DNA testing.

Using commercial kits with additional loci together with lineage markers as a first line of testing will resolve a majority of previously inconclusive complex kinship testing outcomes. However, for queries where the use of lineage markers is not applicable requires a different approach. Recommendations include using commercial kits with even more markers than the ones included in the study, such as the PowerPlex® Fusion 6C system which contains 23 autosomal STRs, three Y-STRs and *Amelogenin* to further increase the power of discrimination. PowerPlex® Fusion 6C system is based on two of the three Y-STRs being rapidly mutating, giving the user the ability to discriminate between closely related males (Promega Corporation, 2018). Other studies recommend the use of massively parallel sequencing (MPS), which enables simultaneous analysis of hundreds of STR loci and provides the nucleotide sequence variations for the alleles, routinely undetected by capillary electrophoresis, thus differentiating identity by descent from identity by state

(Novroski et al., 2016). The DPL intends to explore available options in order to provide sufficient service to its clientele, as the outcomes of the test results usually have huge legal implications, such as approval of an application for an identity document/birth certificate, immigration or settlement of an estate query, among many reasons for DNA testing.

CHAPTER 6 REFERENCES

Azevedo, D.A., Souza, G.R.B., Silva, I.H.E.F., et al., 2011. Genetic kinship analysis: A concordance study between calculations performed with the software Familias and algebraic formulas of the American Association of Blood Banks. *Forensic Science International: Genetics Supplement Series,* 3(1), e186–e187. doi:10.1016/j.fsigss.2011.08.093

Butler, J.M., 2005. Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers. 2nd ed*.* Elsevier.

Butler, J.M., Hill, C.R., 2013. Biology and Genetics of New Autosomal STR Loci Useful for Forensic DNA Analysis. *Forensic DNA Analysis,* 181-198. doi:10.1201/b15361-13

Butler, J.M., 2015. *Advanced Topics in Forensic DNA Typing: Interpretation*. 3rd ed. Academic Press.

Chakraborty, R., Jin, L., 1993. Determination of Relatedness between Individuals Using DNA Fingerprinting. *Human Biology* 65(6), 875–895. doi:10.2307/41464929

Chakraborty R., Jin L. (1993) A unified approach to study hypervariable polymorphisms: Statistical considerations of determining relatedness and population distances. In: Pena S.D.J., Chakraborty R., Epplen J.T., Jeffreys A.J. (eds). DNA Fingerprinting: State of the Science. Progress in Systems and Control Theory. Birkhäuser, Basel. doi:10.1007/978-3-0348-8583-6_14

Chen, B., Cole, J.W., Grond-Ginsbach, C., 2017. Departure from Hardy Weinberg Equilibrium and Genotyping Error. *Front Genet*, 8:167. doi:10.3389/fgene.2017.00167

Cho, S., Shin, E.S., Yu, H.J., et al., 2017. Set up of cut-off thresholds for kinship determination using SNP loci. *Forensic Sci Int Genet*, 29, 1–8. doi:10.1016/j.fsigen.2017.03.009

Coble, M.D., Butler, J. M., 2013. Haplotype data for 23 Y-chromosome markers in four U.S. population groups. *Forensic Science International. Genetics* ,7(3), e66–e68. doi:10.1016/j.fsigen.2013.03.006.

Converge™ version 2.1, 2018, computer software, Thermo Fisher Scientific Inc.

Cornuet, J. M., Luikart, G. 1996. Description and Power Analysis of Two tests for detecting recent population bottlenecks from allele frequency data. *Genetics,* 144(4), 2001-2014. https://www.genetics.org/content/144/4/2001

Corte-Real, F., 2004. Forensic DNA databases. *Forensic Science International* ,146, S143–S144. doi:10.1016/j.forsciint.2004.09.043

Coste, J., Pouchot, J., 2003. A grey zone for quantitative diagnostic and screening tests. *International Journal of Epidemiology, 32(2)* 304–313. doi:10.1093/ije/dyg054

Dauber, E.M., Bar, W, Klintschar, M., Neuhuber, et al., 2003. Mutation rates at 23 different short tandem repeat loci. *International Congress Series,1239, 565–567.* doi:10.1016/s0531-5131(02)00217-0

Drábek, J., 2009. Validation of software for calculating the Likelihood Ratio for parentage and kinship. *Forensic Science International Genetics* 3(2), 112–118. doi:10.1016/j.fsigen.2008.11.005

Fan, H., Chu, J.-Y., 2007. A brief review of short tandem repeat mutation. *Genomics Proteomics Bioinformatics* 5(1), 7–14. doi:10.1016/S1672-0229(07)60009-6

Fujitani, N., Hara, M., Kido, A., 2003. Population genetic data on five STR loci, vWA, TH01, TPOX, F13A01 and CSF1PO, in South African blacks and South African whites. *Legal Medicine, 5(2), 112–113.* doi:10.1016/s1344-6223(03)00008-7

GeneMarker® HID version 3.0.0 (35-day trial), 2019, computer software, SoftGenetics LLC.

Gill, P., Brenner, C., Brinkmann, B., et al., 2001. DNA Commission of the International Society of Forensic Genetics: recommendations of forensic analysis using Y-chromosome STRs. *Int J Leg Med*, *114, 5*–10. doi:10.1007/s004140100232

Giroti, R.I., Verma, S., Singh, K., Malik, R., Talwar, I., 2007. A grey zone approach for evaluation of 15 short tandem repeat loci in sibship analysis: a pilot study in Indian subjects. *J Forensic Leg Med*, 14(5), 261–265. doi:10.1016/j.jcfm.2006.08.002

Gjertson, D.W., Brenner, C.H., Baur, M.P., et al., 2007. ISFG: recommendations on biostatistics in paternity testing. *Forensic Science International: Genetics,* 1(3-4), 223–231. doi:10.1016/j.fsigen.2007.06.006

Grollemund, R., Branford, S., Bostoen, K. et al., 2015. Bantu expansion shows that habitat alters the route and pace of human dispersals. *PNAS,* 112(43), 13296-13301. doi:10.1073/pnas.1503793112

Hares, D.R., 2012. Expanding the CODIS core loci in the United States. *Forensic Sci Int Genet,* 6(1), e52-54. doi:10.1016/j.fsigen.2011.04.012

Hill, C.R., Kline, M.C., Coble, M.D., 2013. U.S. population data for 29 autosomal STR loci. *Forensic Science International. Genetics,* 7 (3), e82–e83. doi:10.1016/j.fsigen.2012.12.004

Holsinger, K.E., Weir, B.S., 2009. Genetics in geographically structured populations: defining, estimating and interpreting FST. *Nature Reviews Genetics* ;10: 639–650. doi:10.1038/nrg2611

Karlsson, A.O., Holmlund, G., Egeland, T., et al., 2007. DNA-testing for immigration cases: the risk of erroneous conclusions. *Forensic Science International*, 172(2-3), 144–149. doi:10.1016/j.forsciint.2006.12.015

Kido, A., Dobashi, Y., Fujitani, N., et al., 2007. Population data on the AmpFlSTR IdentiFiler loci in Africans and Europeans from South Africa. *Forensic Science International,* 168(2-3), 232– 235*.* doi:10.1016/j.forsciint.2006.02.040

Kling, D, Tillmar, A.O., Egeland, T., 2014. Familias 3- Extensions and new functionality. *Forensic Science International: Genetics, (13)* 121–127. doi:10.1016/j.fsigen.2014.07.004

Lane, A. B., Soodyall, H., Arndt, S., et al 2002. Genetic substructure in South African Bantu-speakers: Evidence from autosomal DNA and Y-chromosome studies. *American Journal of Physical Anthropology, 119(2), 175–185.* doi:10.1002/ajpa.10097

Li, H.X., Tong, D.Y., Lu, H.L., et al., 2011. Mutation analysis of 24 autosomal STR loci using in paternity testing. *Forensic Science International: Genetics Supplement Series, Progress in Forensic Genetics,* 3(1), e159–e160. doi:10.1016/j.fsigss.2011.08.080

Lucassen, A., Ehlers, K., Grobler, P.J. & Shezi, A.L., 2013. 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. doi:10.1007/s00414-013-0898-4

McKiernan, H.E, Danielson, P.B., 2017. Molecular Diagnostic Applications in Forensic Science. In Patrinos, G. P. (Eds)*. Molecular Diagnostics,* (pp 371-394), 3rd ed. Academic Press. doi:10.1016/B978-0-12-802971-8.00021-3

Moretti, T.R., Moreno, L.I., Smerick, J.B., et al., 2016. Population data on the expanded CODIS core STR loci for eleven populations of significance for forensic DNA analyses in the United States. *Forensic Science International: Genetics,* 25, 175–181. doi:10.1016/j.fsigen.2016.07.022

Musanovic, J., Filipovska-Musanovic, M., Kovacevic, L., et al., 2012. Determination of combined sibship indices “grey zone” using 15 STR loci for central Bosnian human population. *Mol Biol Rep,* 39(5), 5195–5200. doi:10.1007/s11033-011-1316-y

Novroski, N.M.M., King, J.L., Churchill, J.D., et al., 2016. Characterization of genetic sequence variation of 58 STR loci in four major population groups. *Forensic Science International: Genetics,* 25, 214–226. doi:10.1016/j.fsigen.2016.09.007

O'Connor, K. L., Hill, C. R., Vallone, P. M., & Butler, J. M. 2011. Linkage disequilibrium analysis of D12S391 and vWA in U.S. population and paternity samples. *Forensic Science International: Genetics, 5(5), 538–540.* doi:10.1016/j.fsigen.2010.09.003

Oostdik, K., Lenz, K., Nye, J., et al., 2014. Developmental validation of the PowerPlex® Fusion System for analysis of casework and reference samples: A 24-locus multiplex for new database standards. *Forensic Science International: Genetics*,12, 69–76. doi:10.1016/j.fsigen.2014.04.013

Parson, W., Gusmao, L., Hares, D.R., et al., 2014. DNA Commission of the International Society for Forensic Genetics: revised and extended guidelines for mitochondrial DNA typing. *Forensic Science International: Genetics,* (13), 134–142. doi: 10.1016/j.fsigen.2014.07.010

Pu, C.E., Linacre, A., 2008. Systematic evaluation of sensitivity and specificity of sibship determination by using 15 STR loci. *J Forensic Leg Med,* 15, 329–334. doi:10.1016/j.jflm.2007.12.018

Ramos-González, B., Aguilar-Velázquez, J.A., Chávez-Briones, M. de L., et al., 2016. Population data of 24 STRs in Mexican-Mestizo population from Monterrey, Nuevo Leon (Northeast, Mexico) based on PowerPlex® Fusion and GlobalFiler® kits. *Forensic Sci Int Genet,* 21, e15-e17. doi:10.1016/j.fsigen.2015.12.004

Raymond, M., Rousset, F., 1995. GENEPOP (v1.2): Population genetics software for exact tests and ecumenisms. *Journal of Heredity,* 86(3), 248–249. doi:10.1093/oxfordjournals.jhered.a111573

Rosenberg, N.A., 2002. Genetic Structure of Human Populations. *Science,* 298(5602), 2381–2385. doi: 10.1126/science.1078311

Rousset, F., 2008. Genepop'007: A complete reimplementation of GenePop software for Windows and Linux. *Molecular Ecology Resources,* 8(1), 103–106. doi:10.1111/j.1471-8286.2007.01931

Thompson, E.A., 2013. Identity by Descent: Variation in Meiosis, Across Genomes, and in Populations. *Genetics,* 194(2), 301–326. doi:10.1534/genetics.112.148825

Tillmar, A.O., Mostad, P., Egeland, T., et al, 2008. Analysis of linkage and linkage disequilibrium for eight X-STR markers. *Forensic Science International: Genetics,* 3(1), 37–41. doi:10.1016/j.fsigen.2008.09.006

Turrina, S., Ferrian, M., Caratti, S., et al., 2014. Evaluation of genetic parameters of 22 autosomal STR loci (PowerPlex® Fusion System) in a population sample from Northern Italy. *Int. J. Legal Med.,* 128(2), 281–283. doi:10.1007/s00414-013-0934-4

Wang, H., Shen, C., Liu, W., et al., 2013. Allelic frequency distributions of 21 non-combined DNA index system STR loci in a Russian ethnic minority group from Inner Mongolia, China. *J. Zhejiang Univ. Sci. B,* 14(6), 533–540. doi:10.1631/jzus.B1200262

Waples, R.S., 2015. Testing for Hardy–Weinberg Proportions: Have We Lost the Plot? *Journal of Heredity,* 106(1), 1–19. doi:10.1093/jhered/esu062

Weir, B.S., Cockerham, C.C., 1984. Estimating F-Statistics for the Analysis of Population Structure. *Evolution,* 38(6), 1358–1370. doi: 10.2307/2408641

Wenk, R.E., Chiafari, F.A., Gorlin, J., et al., 2003. Better tools are needed for parentage and kinship studies. *Transfusion,* 43(7), 979–981. doi:10.1046/j.1537-2995.2003.00439.x

6.1 ELECTRONIC RESOURCES

Botstein, D. White, R.L., 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1686077/ (accessed 25/03/2020).

Butler, J.M., 2012. Variability of New STR Loci and Kits in US Population Groups. https://worldwide.promega.com/resources/profiles-in-dna/2012/variability-of-new-str-loci-and-kits-in-us-population-groups/ (accessed 21/02/2020).

Crow, J.F., 1996. The 1996 NRC Report: Another Look - Promega Corporation.https://www.promega.com/-/media/files/resources/conference-proceedings/ishi-09/oral-presentations/24.pdf?la=en (accessed 12/02/2020)

Department of Home Affairs, 2020a. Civic services http://www.dha.gov.za/index.php/civic-services/citizenship (accessed 12/02/2020).

Department of Home Affairs, 2020b. Immigration http://www.dha.gov.za/index.php/civic-services/amendments (accessed 12/02/2020).

Duewer, D., 2016. *STR_Genotype*. computer software. https://strbase.nist.gov/software.htm (accessed 18/07/2019)

Life Technologies Holdings Pte Ltd, 2017. VeriFiler™ Express PCR Amplification Kit User guide.https://assets.thermofisher.com/TFS-Assets/LSG/manuals/100043588_VFE_UG.pdf (accessed 06/04/2020).

Maguire, C., Woodward, M., 2008. DNA-Based Kinship Analysis. *Profiles in DNA* 11, 3–6.https://www.promega.co.uk/resources/profiles-in-dna/2008/dnabased-kinship-analysis/ (accessed 20/08/2018)

National Institute of standards and Technology, 2007. Brief introduction to STRs https://strbase.nist.gov/intro.htm#STR%20Loci (accessed 25/03/2020).

National Institute of Standards, 2009. DNA Advisory Board Quality Assurance Standards for forensic DNA testing laboratories *(No. Forensic Science Communication* 2 (2000) https://strbase.nist.gov/dabqas.htm. (accessed 12/02/2020)

O'Connor, K. L., 2010. The Role of DNA in Kinship Analysis. https://strbase.nist.gov//pub_pres/Lewis-LM-kinship-Apr2010.pdf (accessed 12/02/2020)

Promega Corporation, 2017. powerplex-18d-system-protocol.pdf https://www.promega.com/-/media/files/resources/protocols/technical-manuals/101/powerplex-18d-system-protocol.pdf (accessed 03/03/2018).

Promega Corporation, 2017b. powerplex-fusion-system-protocol.pdf https://www.promega.com/~/media/files/resources/protocols/technical-manuals/101/powerplex-fusion-system-protocol.pdf (accessed 18/02/2020)

Promega Corporation, 2018. powerplex-fusion-6c-system-protocol.pdf https://www.promega.com/-/media/files/resources/protocols/technical-manuals/101/powerplex-fusion-6c-system-protocol.pdf (accessed 03/03/2020)

Raymond, M., Rousset, F., 2019. GenePop on the Web, V 4.7.5 https://genepop.curtin.edu.au/ (accessed 07/04/2020).

Sozer, A., Baird, M., Beckwith, M., et al., 2010. Guidelines for Mass Fatality DNA Identification Operations, 58. http://www.aabb.org/programs/disasterresponse/Pages/massfatality.aspx (accessed 18/02/2020)

Statistics South Africa, n.d., 2011 Census products | Statistics South Africa. http://www.statssa.gov.za/?page_id=3955 (accessed 14/04/2020).

Statistics South Africa, Republic of South Africa, 2018. Migrants flock to Gauteng statssa.gov.za. http://www.statssa.gov.za/?p=11331 (accessed 14/04/2020).

SWGDAM, 2012. Validation guidelines for DNA Analysis Methods. https://1ecb9588-ea6f-4feb-971a-73265dbf079c.filesusr.com/ugd/4344b0_813b241e8944497e99b9c45b163b76bd.pdf (accessed 20/02/2020)

SWGDAM, 2010. SWGDAM Autosomal STR Interpretation Guidelines 40. https://1ecb9588-ea6f-4feb-971a-73265dbf079c.filesusr.com/ugd/4344b0_50e2749756a242528e6285a5bb478f4c.pdf (accessed 20/02/2020)

The Relationship Testing Program Unit, 2008. ANNUAL REPORT SUMMARY FOR TESTING IN 2008 http://www.aabb.org/sa/facilities/Pages/relationshipreports.aspx (accessed 31/03/2020).

Thermo Fischer Scientific, 2017. VeriFiler™ Express PCR Amplification Kit.https://assets.thermofisher.com/TFS-Assets/LSG/manuals/100043588_VFE_UG.pdf (accessed 06/04/2020).

CHAPTER 7 APPENDICES

7.1 APPENDIX A1: Ethics Clearance

R14/49 Ms Mahlatse Moremi

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

CLEARANCE CERTIFICATE NO. M170474

NAME: (Principal Investigator)	Ms Mahlatse Moremi
DEPARTMENT:	Human Genetics National Health Laboratory Services
PROJECT TITLE:	Evaluation of the PowerPlex Fusion Kit to Resolve Complex Kinship Disputes in a South African Setting
DATE CONSIDERED:	05/05/2017
DECISION:	Approved uncoonditionally
CONDITIONS:	
SUPERVISOR:	Prof Himla Soodyall and Dr Venesa Sahibdeen
APPROVED BY:	Professor CB. Penny Co-Chairperson, HREC (Medical)
DATE OF APPROVAL:	04/09/2017

This clearance certificate is valid for 5 years from date of approval. Extension may be applied for.

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Research Office Secretary in Room 10004,10th floor, Senate House/3rd floor, Phillip Tobias Building, Parktown, University of the Witwatersrand. I/We fully understand the conditions under which I am/we are authorised to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions. Should any departure be contemplated, from the research protocol as approved, I/we undertake to resubmit to the Committee. **I agree to submit a yearly progress report**. The date for annual re-certification will be one year after the date of convened meeting where the study was initially reviewed. in this case, the study was initially review April and will therefore be due in the month of April each year. Unreported changes to the application may invalidate the clearance given by the HREC (Medical).

Moremi
Principal Investigator Signature

08 - SEPTEMBER-2017
Date

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

7.2 APPENDIX A2: Change of dissertation title

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

Private Bag 3 Wits, 2050

Fax: 027117172119
Tel: 02711 7172076
Reference: Mrs Sandra Benn

E-mail: sandra.benn@wits.ac.za

04 January 2019

Miss MM Moremi
P. O. Box
31119
Braamfontein
2017
South Africa

Person No: 9603072E
TAA

Dear Miss Mashela Moremi

Master of Science in Medicine: Change of title of research

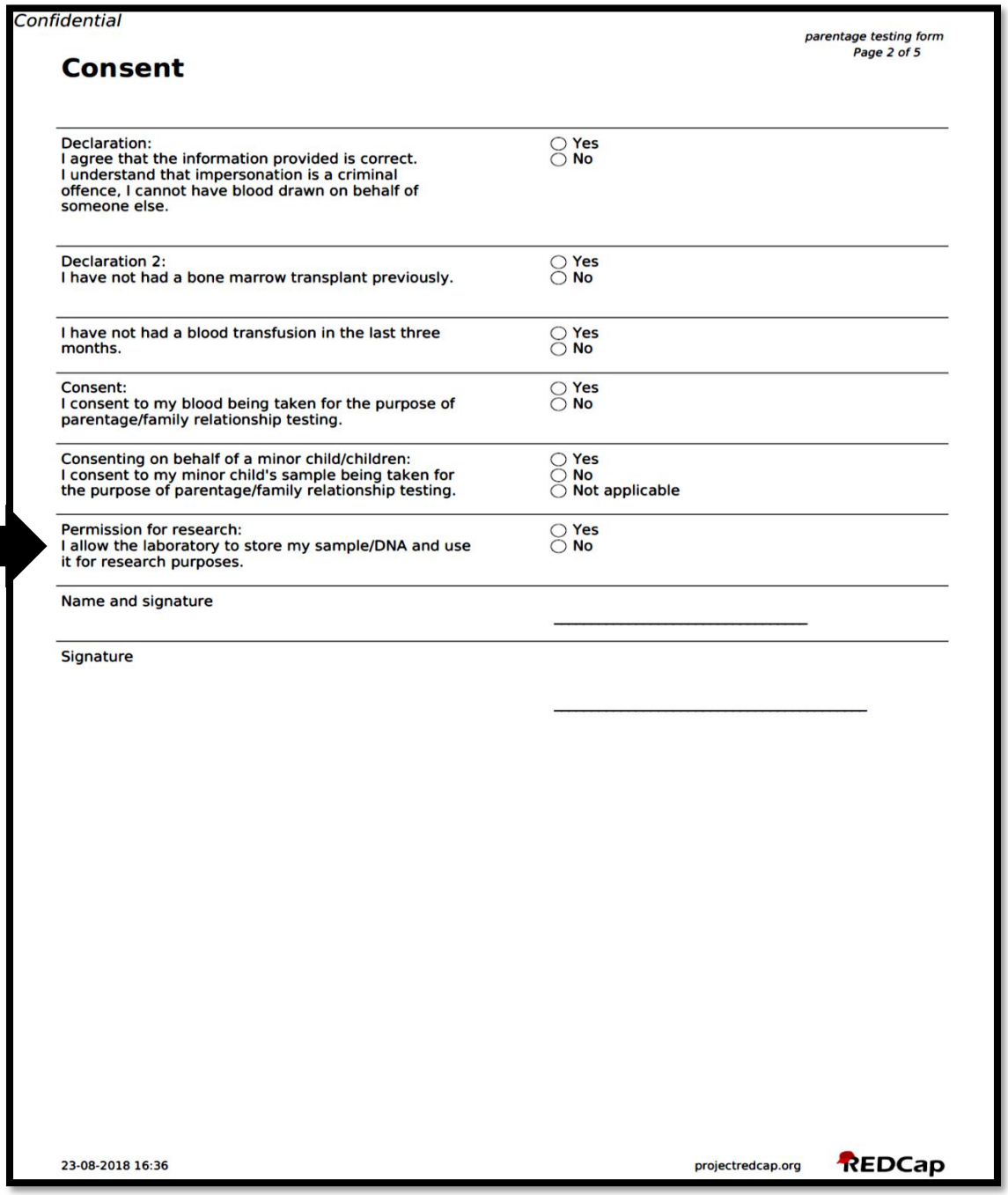
I am pleased to inform you that the following change in the title of your Dissertation for the degree of **Master of Science in Medicine** has been approved:

From:	**Evaluation of the PowerPlex Fusion kit to resolve complex kinship disputes in a South African setting**
To:	**Evaluation of commercially available parentage testing kits to resolve complex kinship disputes in a South African setting**

Yours sincerely

Mrs Sandra Benn
Faculty Registrar
Faculty of Health Sciences

7.3 APPENDIX B1: Consent form used in the DNA Profiling Laboratory (2018 to date)



Confidential

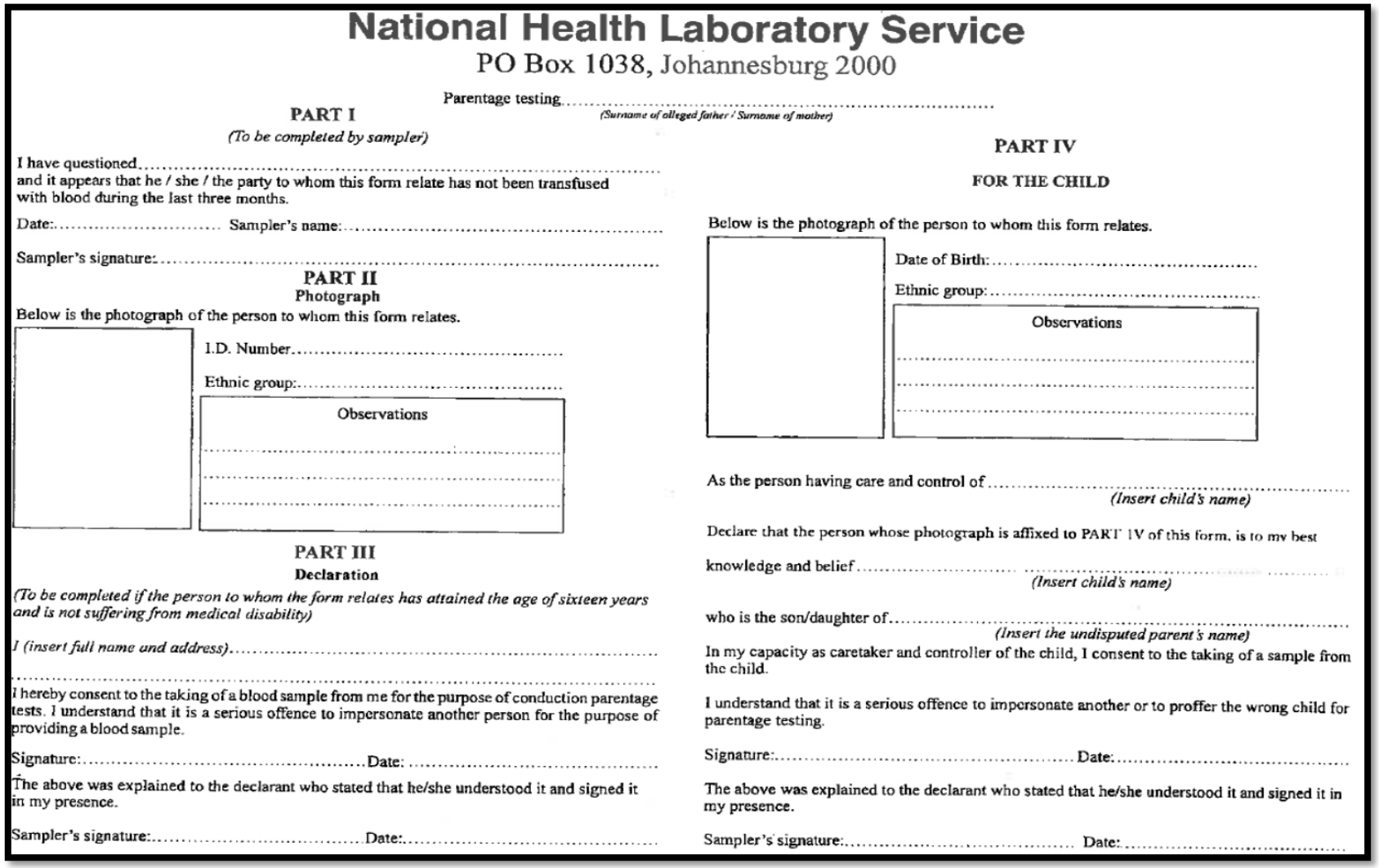
parentage testing form
Page 2 of 5

Consent

Declaration: I agree that the information provided is correct. I understand that impersonation is a criminal offence, I cannot have blood drawn on behalf of someone else.	◯ Yes ◯ No
Declaration 2: I have not had a bone marrow transplant previously.	◯ Yes ◯ No
I have not had a blood transfusion in the last three months.	◯ Yes ◯ No
Consent: I consent to my blood being taken for the purpose of parentage/family relationship testing.	◯ Yes ◯ No
Consenting on behalf of a minor child/children: I consent to my minor child's sample being taken for the purpose of parentage/family relationship testing.	◯ Yes ◯ No ◯ Not applicable
Permission for research: I allow the laboratory to store my sample/DNA and use it for research purposes.	◯ Yes ◯ No
Name and signature	_____
Signature	_____

23-08-2018 16:36 projectredcap.org REDCap

7.4 APPENDIX B2: Consent form used in the DNA Profiling lab prior to 2018



National Health Laboratory Service

PO Box 1038, Johannesburg 2000

Parentage testing.....
(Surname of alleged father / Surname of mother)

PART I

(To be completed by sampler)

I have questioned.....
and it appears that he / she / the party to whom this form relate has not been transfused with blood during the last three months.

Date:..... Sampler's name:.....

Sampler's signature:.....

PART II

Photograph

Below is the photograph of the person to whom this form relates.

I.D. Number.....

Ethnic group:.....

Observations

.....

.....

.....

PART III

Declaration

(To be completed if the person to whom the form relates has attained the age of sixteen years and is not suffering from medical disability)

I (insert full name and address).....

.....

I hereby consent to the taking of a blood sample from me for the purpose of conduction parentage tests. I understand that it is a serious offence to impersonate another person for the purpose of providing a blood sample.

Signature:.....Date: .....

The above was explained to the declarant who stated that he/she understood it and signed it in my presence.

Sampler's signature:.....Date:.....

PART IV

FOR THE CHILD

Below is the photograph of the person to whom this form relates.

Date of Birth:.....

Ethnic group:.....

Observations

.....

.....

.....

As the person having care and control of .....
(Insert child's name)

Declare that the person whose photograph is affixed to PART IV of this form, is to my best knowledge and belief ....
(Insert child's name)

who is the son/daughter of.....
(Insert the undisputed parent's name)

In my capacity as caretaker and controller of the child, I consent to the taking of a sample from the child.

I understand that it is a serious offence to impersonate another or to proffer the wrong child for parentage testing.

Signature:.....Date:.....

The above was explained to the declarant who stated that he/she understood it and signed it in my presence.

Sampler's signature:..... Date:.....

7.5 APPENDIX C: Sample Map template

Table 7.1 An example of a sample plate map for loading on the genetic anlayser

	1	2	3	4	5	6	7	8	9	10	11	12
A	A ladder	SAFBS6	SAFBS14	SAFBS22	SAFWH1	SAFWH9	SAFMX1	SAFMX9	ZIMBS1	ZIMBS9	NGRBS1	NGRBS9
B	Pos con	SAFBS7	SAFBS15	SAFBS23	SAFWH2	SAFWH10	SAFMX2	SAFMX10	ZIMBS2	ZIMBS10	NGRBS2	NGRBS10
C	Neg con	SAFBS8	SAFBS16	SAFBS24	SAFWH3	SAFWH11	SAFMX3	SAFMX11	ZIMBS3	ZIMBS11	NGRBS3	NGRBS11
D	SAFBS1	SAFBS9	SAFBS17	SAFBS25	SAFWH4	SAFWH12	SAFMX4	SAFMX12	ZIMBS4	ZIMBS12	NGRBS4	NGRBS12
E	SAFBS2	SAFBS10	SAFBS18	SAFBS26	SAFWH5	SAFWH13	SAFMX5	SAFMX13	ZIMBS5	ZIMBS13	NGRBS5	NGRBS13
F	SAFBS3	SAFBS11	SAFBS19	SAFBS27	SAFWH6	SAFWH14	SAFMX6	SAFMX14	ZIMBS6	ZIMBS14	NGRBS6	A. ladder
G	SAFB4	SAFBS12	SAFBS20	SAFBS28	SAFWH7	SAFWH15	SAFMX7	SAFMX15	ZIMBS7	ZIMBS15	NGRBS7	Pos cont
H	SAFBS5	SAFBS13	SAFBS21	SAFBS29	SAFWH8	SAFWH16	SAFMX8	SAFMX16	ZIMBS8	ZIMBS16	NGRBS8	Neg cont

7.6 APPENDIX D: Population allele frequencies for the different ethnic groups.

Table 7.2 Population allele frequencies for the SA Bantu-speakers (SAFBS), SA Coloureds (SAFMX), SA Whites (SAFWH), Zimbabwean Bantu-speakers (ZIMBS) and Nigerian Bantu-speakers (NGRBS), grouped per STR locus.

STR locus: D3S1358	**SAFBS (N=366)**	**SAFMX (N=40)**	**SAFWH (N=84)**	**ZIMBS(N=113)**	**NGRBS (N=73)**
11	**-**	**0.0625**	**-**	**-**	**-**
12	0.0068*	-	-	0.0221*	0.0338*
13	-	-	0.0298*	-	
14	0.0997	0.0625*	0.1488	0.0885	0.0878
15	0.2910	0.3250	0.2500	0.2699	0.3649
16	0.4030	0.3750	0.2321	0.3584	0.3108
17	0.1721	0.1625	0.2083	0.2168	0.2027
18	0.0314	0.0750	0.1131	0.0575	0.0338*
19	-	-	0.0357	0.0221*	-
20	-	-	0.0298*	-	-
STR locus: vWA	**SAFBS (N=366)**	**SAFMX (N=40)**	**SAFWH (N=84)**	**ZIMBS(N=113)**	**NGRBS (N=73)**
11	0.0068*	**-**	**-**	0.0221*	**-**
13	0.0096	0.0625*	0.0298*	0.0221*	0.0338*
14	0.0820	0.0750	0.1488	0.0708	0.0338*
15	0.1899	0.2250	0.0774	0.2035	0.2635
16	0.2842	0.2125	0.2202	0.2522	0.2500
17	0.1899	0.2125	0.3036	0.1947	0.1689
18	0.1653	0.1625	0.1607	0.1504	0.1419
19	0.0519	0.0625*	0.0655	0.0796	0.0676
20	0.0178	0.0625*	0.0298*	0.0221*	0.0338*
21	0.0068*	-	0.0298*	0.0221*	-
22	-	-	-	-	0.0338*
28	-	-	-	-	0.0338*
32.2	-	-	-	-	0.0338*
STR locus: D16S539	**SAFBS (N=366)**	**SAFMX (N=40)**	**SAFWH (N=84)**	**ZIMBS(N=113)**	**NGRBS (N=73)**
8	0.0287	0.0625*	0.0298*	0.0221*	0.0541
9	0.1899	0.1750	0.1488	0.2345	0.2230
10	0.1598	0.1375	0.0833	0.1283	0.1149
11	0.3730	0.3750	0.3155	0.3673	0.3108
12	0.1557	0.1875	0.2798	0.1283	0.1419
13	0.0861	0.0625*	0.1369	0.1062	0.1216
14	0.0068*	0.0625*	0.0298*	0.0265	0.0338*
16	-	-	-	-	0.0338*

Table 7.2, continued

STR locus: CSF1PO	SAFBS (N=366)	SAFMX (N=40)	SAFWH (N=84)	ZIMBS(N=113)	NGRBS (N=73)
6	0.0137	0.0625*	-	-	-
7	0.0642	0.0625*	-	0.0442	0.0338*
7.3	-	0.0625*	-	-	-
8	0.0355	0.0625*	-	0.0442	0.0473
9	0.0314	0.0625*	0.0298*	0.0796	0.0338*
10	0.2650	0.2375	0.2560	0.2743	0.2905
11	0.2158	0.2125	0.3036	0.1903	0.2500
12	0.3333	0.3750	0.3095	0.3053	0.2905
13	0.0396	0.0750	0.0893	0.0575	0.0541
14	0.0068*	0.0625*	0.0298*	0.0221*	-
16	-	-	-	-	0.0338*
STR locus: TPOX	**SAFBS (N=366)**	**SAFMX (N=40)**	**SAFWH (N=84)**	**ZIMBS(N=113)**	**NGRBS (N=73)**
6	0.1134	0.0625*	0.0298*	0.1018	0.0946
7	0.0205	-	0.0298*	0.0221*	0.0338*
8	0.2404	0.3718	0.5595	0.2876	0.3176
9	0.2240	0.2051	0.1429	0.1504	0.2162
10	0.0861	0.0641	0.0893	0.0973	0.0946
11	0.3019	0.2821	0.1845	0.3319	0.2027
12	0.0068*	0.0625*	0.0298*	0.0221*	0.0405
19	-	-	-	-	0.0338*
23	-	-	-	-	0.0338*
10;11	0.0068*	-	-	-	-
8;10	0.0068*	-	-	-	-
9;10	0.0068*	-	-	-	-
STR locus: D8S1179	**SAFBS (N=366)**	**SAFMX (N=40)**	**SAFWH (N=84)**	**ZIMBS(N=113)**	**NGRBS (N=73)**
7	-	-	-	-	0.0338*
8	-	0.0625*	0.0357	-	-
9	-	0.0625*	0.0298*	-	-
10	0.0068*	0.0625*	0.1190	-	0.0338*
11	0.0328	0.0625*	0.0655	0.0221*	0.0338*
12	0.1366	0.0875	0.1607	0.1283	0.0811
13	0.2281	0.2250	0.3095	0.2522	0.1622
14	0.2650	0.4000	0.1488	0.3186	0.3919
15	0.2596	0.1250	0.1131	0.1858	0.2230
16	0.0669	0.0625*	0.0298*	0.0929	0.1014
17	0.0096	-	0.0298*	-	0.0338*

Table 7.2, continued

STR locus: D21S11	SAFBS (N=366)	SAFMX (N=40)	SAFWH (N=84)	ZIMBS(N=113)	NGRBS (N=73)
17	-	-	-	-	0.0338*
18	-	-	-	-	0.0338*
24.3	0.0068*	-	-	0.0221*	-
26	0.0068*	-	-	-	-
27	0.0820	0.0750	0.0298*	0.0796	0.0743
28	0.2828	0.1625	0.1369	0.2522	0.3446
29	0.1448	0.1875	0.2619	0.1549	0.1757
29.2	-	-	0.0298*	-	-
29.3	-	0.0625*	-	-	-
30	0.1161	0.1250	0.2679	0.1726	0.1689
30.2	0.0096	0.0625*	0.0298*	-	0.0338*
31	0.0888	0.1000	0.0655	0.1106	0.0405
31.2	0.0587	0.0750	0.0595	0.0664	0.0338*
32	0.0205	-	0.0298*	0.0221*	0.0338*
32.2	0.0601	0.0625*	0.1012	0.0531	0.0405
33	0.0137	-	-	0.0221*	0.0338*
33.1	0.0082	-	-	0.0221*	-
33.2	0.0301	0.0625*	0.0357	0.0221*	0.0338*
34	0.0109	-	-	-	0.0338*
34.1	0.0068*	0.0625*	-	-	-
34.2	0.0068*	0.0625*	-	-	-
34.3	0.0068*	-	-	-	-
35	0.0369	0.0625*	0.0298*	0.0575	0.0338*
35.1	0.0068*	0.0625*	-	-	-
35.2	-	0.0625*	-	-	-
36	0.0123	-	-	0.0221*	0.0338*
36.1	-	0.0625*	-	-	-
37	0.0068*	-	-	-	-
38	0.0068*	-	-	-	-
STR locus: D2S441	**SAFBS (N=366)**	**SAFMX (N=40)**	**SAFWH (N=84)**	**ZIMBS(N=113)**	**NGRBS (N=73)**
7	0.0068*	-	-	-	-
7.3	0.0068*	-	-	-	-
9	-	-	-	0.0221*	-
10	0.0451	0.1250	0.2321	0.0619	0.0405
10.1	0.0068*	-	-	-	-
11	0.3675	0.3125	0.3155	0.3673	0.3784
11.3	0.0246	0.0625*	0.0298*	0.0221*	0.0608
12	0.1667	0.1375	0.0298*	0.1991	0.1622
12.3	0.0301	0.0625*	-	0.0221*	-
13	0.0232	0.0625*	0.0655	0.0354	0.0338*
13.3	-	-	-	0.0221*	-
14	0.2910	0.2375	0.2679	0.2699	0.2905
14.3	0.0068*	-	-	-	-
15	0.0383	0.0625*	0.0595	0.0221*	0.0338*
16	0.0068*	0.0625*	-	-	0.0338*

Table 7.2, continued

STR locus: D18S51	SAFBS (N=366)	SAFMX (N=40)	SAFWH (N=84)	ZIMBS(N=113)	NGRBS (N=73)
9	0.0068*	-	-	-	-
10	0.0068*	-	-	-	-
10.2	0.0178	0.0625*	0.0298*	-	0.0338*
11	0.0068*	0.0625*	0.0298*	0.0221*	-
12	0.0219	0.0625*	0.0476	0.0354	0.0811
13	0.0396	0.1250	0.1131	0.0310	0.0338*
13.2	-	-	-	-	0.0338*
14	0.0560	0.1000	0.1786	0.0575	0.0405
14.2	0.0068*	-	-	0.0221	-
15	0.1393	0.1250	0.1548	0.1726	0.1554
15.2	0.0123	0.0625*	-	0.0221*	0.0338*
16	0.1995	0.1875	0.1310	0.1991	0.2635
17	0.1858	0.1625	0.1726	0.1549	0.1959
18	0.1148	0.0750	0.0774	0.1504	0.0743
19	0.1025	0.1125	0.0476	0.0929	0.0811
20	0.0669	0.0625*	0.0298*	0.0442	0.0338*
20.2	0.0068*	-	-	-	-
21	0.0205	-	0.0298*	0.0354	0.0338*
21.2	0.0068*	-	-	-	-
22	0.0068*	-	0.0298*	0.0221*	-
22.2	0.0068*	-	-	-	-
23	-	-	0.0298*	0.0221*	-
STR locus: Penta_E	**SAFBS (N=366)**	**SAFMX (N=40)**	**SAFWH (N=84)**	**ZIMBS(N=113)**	**NGRBS (N=73)**
3	0.0068*	-	-	-	-
5	0.0915	0.1250	0.0714	0.0796	0.1014
6	-	-	0.0298*	-	-
7	0.1284	0.1000	0.1667	0.0531	0.1149
8	0.1831	0.0625*	0.0298*	0.1814	0.3311
9	0.0615	0.0625*	0.0298*	0.0619	0.0541
10	0.0902	0.0875	0.0833	0.0752	0.0338*
10.4	0.0068*	-	-	-	-
11	0.0601	0.0750	0.1071	0.0664	0.0338*
12	0.0984	0.1375	0.1667	0.0885	0.0676
13	0.1257	0.1750	0.1012	0.1106	0.1216
14	0.0396	0.0750	0.0476	0.0708	0.0338*
15	0.0451	0.0625*	0.0595	0.0619	0.0405
16	0.0260	0.0625*	0.0655	0.0796	0.0338*
16.3	0.0068*	-	-	-	-
17	0.0232	-	0.0595	0.0398	0.0338*
18	0.0109	0.0625*	0.0298*	0.0221*	0.0338*
19	0.0068*	-	0.0298*	0.0221*	0.0338*
20	-	0.0625*	-	0.0221*	-
24	-	0.0625*	0.0298*	-	-
28	-	-	-	-	0.0338*
32.2	-	-	-	-	0.0338*

Table 7.2, continued

STR locus: D19S433	SAFBS (N=366)	SAFMX (N=40)	SAFWH (N=84)	ZIMBS(N=113)	NGRBS (N=73)
7	0.0314	-	-	0.0221*	-
8	0.0068*	-	-	-	-
10	0.0178	0.0625*	-	0.0221*	-
11	0.0464	0.0625*	0.0298*	0.0708	0.0743
12	0.1530	0.0875	0.0774	0.1903	0.1014
12.2	0.0355	0.0625*	0.0298*	0.0708	0.0338*
13	0.3005	0.2750	0.2083	0.2699	0.2635
13.2	0.0724	0.0625*	0.0298*	0.0664	0.0811
14	0.1954	0.3000	0.4226	0.1637	0.2432
14.2	0.0519	0.0625*	0.0298*	0.0575	0.0676
15	0.0587	0.1125	0.1726	0.0354	0.0338*
15.2	0.0150	0.0625*	0.0357	0.0310	0.0608
16	0.0068*	0.0625*	0.0298*	-	0.0338*
16.2	0.0137	-	-	0.0221*	0.0338*
17.2	-	-	-	0.0221*	
18.2	-	-	-	-	0.0338*
STR locus: TH01	**SAFBS (N=366)**	**SAFMX (N=40)**	**SAFWH (N=84)**	**ZIMBS(N=113)**	**NGRBS (N=73)**
5	0.0068*	-	-	-	-
6	0.0970	0.2000	0.2381	0.0929	0.1284
7	0.3852	0.2000	0.1845	0.3319	0.4459
8	0.3128	0.2125	0.1071	0.3496	0.2162
9	0.1680	0.2875	0.1369	0.2080	0.1486
9.3	0.0205	0.1000	0.3333	0.0221*	0.0473
10	0.0150	-	-	-	-
19	-	-	-	-	0.0338*
23	-	-	-	-	0.0338*
STR locus: D22S1045	**SAFBS (N=366)**	**SAFMX (N=40)**	**SAFWH (N=84)**	**ZIMBS(N=113)**	**NGRBS (N=73)**
8	-	-	-	-	0.0338*
10	0.1079	0.0625*	-	0.0796	0.0743
11	0.1653	0.2625	0.1369	0.1903	0.1892
12	0.0383	0.0625*	0.0298*	0.0310	0.0541
13	0.0068*	0.0625*	0.0298*	-	-
14	0.0874	0.0625*	0.0417	0.0442	0.0676
15	0.2459	0.2625	0.3869	0.2965	0.1486
16	0.1817	0.1500	0.3333	0.1770	0.1959
17	0.1598	0.1500	0.0833	0.1770	0.1892
18	0.0123	-	-	0.0221*	0.0608

Table 7.2, continued

STR locus: FGA	SAFBS (N=366)	SAFMX (N=40)	SAFWH (N=84)	ZIMBS(N=113)	NGRBS (N=73)
7	-	-	-	-	0.0338*
15	0.0068*	-	-	-	-
16	0.0068*	-	-	-	-
16.1	0.0068*	0.0625*	-	-	-
17	0.0068*	-	-	0.0221*	-
17.2	-	-	-	0.0221*	0.0338*
18	0.0068*	-	0.0298*	0.0221*	-
18.2	0.0068*	-	-	0.0221*	0.0338*
19	0.0656	0.0625*	0.0357	0.0664	0.0946
19.2	0.0068*	-	-	0.0221*	0.0338*
20	0.0464	0.0875	0.1845	0.0531	0.0338*
20.2	-	-	0.0298*	0.0221*	
21	0.0806	0.0875	0.1845	0.0841	0.1554
21.2	0.0068	-	0.0298*	-	-
22	0.1503	0.2625	0.1726	0.1681	0.1622
22.2	0.0068*	-	0.0298*	-	-
22.3	0.0068	-	-	0.0221*	-
23	0.2063	0.1625	0.1607	0.1460	0.1486
23.2	0.0068*	0.0625*	-	-	-
23.3	0.0068	-	-	-	-
24	0.1544	0.1250	0.0952	0.1637	0.1622
25	0.1052	0.0875	0.1012	0.1416	0.1149
25.2	-	0.0625*	-	-	-
26	0.0888	0.0625*	0.0357	0.0575	0.0608
27	0.0423	0.0625*	0.0298*	0.0354	0.0338*
28	0.0068	-	-	0.0221*	0.0338*
29	0.0068*	-	-	0.0221*	-
29.2	0.0068*	0.0625*	-	-	-
30	0.0068*	-	-	-	-
30.2	0.0068*	-		0.0221*	-
31.2	-	-		0.0221*	
41.2	0.0068*	-		-	-
42.2	0.0068*	-		-	-
43.2	-	-		0.0221*	
44.2	0.0068*	0.0625*		-	-
45.2	-	-		0.0221*	
STR locus: D5S818	**SAFBS (N=366)**	**SAFMX (N=40)**	**SAFWH (N=84)**	**ZIMBS(N=113)**	**NGRBS (N=73)**
7	0.0068*	-	-	-	-
8	0.0587	0.0750	0.0298*	0.0708	0.0676
9	0.0219	-	0.0536	0.0354	0.0338*
10	0.0683	0.1000	0.0417	0.0531	0.0811
11	0.2500	0.3375	0.3690	0.2124	0.1689
12	0.3538	0.3125	0.3750	0.3584	0.3446
13	0.2309	0.1750	0.1310	0.2566	0.2635
14	0.0123	-	0.0298*	0.0221*	0.0473
15	0.0068*	-	0.0298*	-	0.0338*

Table 7.2, continued

STR locus: D13S317	SAFBS (N=366)	SAFMX (N=40)	SAFWH (N=84)	ZIMBS(N=113)	NGRBS (N=73)
7	-	0.0625*	-	-	-
8	0.0068*	0.0625*	0.1429	0.0133	-
9	0.0082	0.0625*	0.0774	0.0221*	0.0338*
10	0.0423	0.0625*	0.0476	0.0221*	-
11	0.3183	0.2500	0.3036	0.3097	0.2635
12	0.3962	0.4000	0.3274	0.4204	0.4595
13	0.1339	0.1125	0.0893	0.1593	0.1757
14	0.0970	0.1000	0.0298*	0.0531	0.0676
15	0.0068*	-	-	0.0133	0.0338*
28	-	-	-	-	0.0338*
32.2	-	-	-	-	0.0338*
STR locus: D7S820	**SAFBS (N=366)**	**SAFMX (N=40)**	**SAFWH (N=84)**	**ZIMBS(N=113)**	**NGRBS (N=73)**
7	-	-	0.0298*	0.0221*	
7.3	0.0068*	-	-	0.0221*	-
8	0.1503	0.3000	0.1667	0.1858	0.1554
8.2	0.0068*	-	-	-	-
8.3	0.0068*	-	-	-	-
9	0.1407	0.1250	0.2024	0.1327	0.1554
10	0.3101	0.3250	0.2440	0.3628	0.3243
10.1	0.0068*	-	-	-	-
10.3	0.0068*	-	-	-	-
11	0.2555	0.1500	0.1369	0.1327	0.2365
12	0.1107	0.0750	0.1786	0.3628	0.0811
13	0.0232	0.0625*	0.0417	0.1327	0.0338*
14	0.0068*	0.0625*	-	-	0.0338*
16	-	-	-	-	0.0338*
STR locus: D10S1248	**SAFBS (N=366)**	**SAFMX (N=40)**	**SAFWH (N=84)**	**ZIMBS(N=113)**	**NGRBS (N=73)**
7	0.0068	0.0625*	-	0.0221*	-
8	-	-	-	-	0.0338*
9	-	-	-	-	0.0338*
10	0.0068*	-	-	0.0221*	0.0338*
11	0.0615	0.0625*	-	0.0752	0.0338*
12	0.1079	0.0625*	0.0357	0.1106	0.1351
13	0.2404	0.2625	0.3274	0.2389	0.2432
14	0.3128	0.3125	0.2798	0.3274	0.2432
15	0.1530	0.1875	0.1488	0.1460	0.1689
16	0.0997	0.1000	0.1726	0.0796	0.1216
17	0.0109	0.0625*	0.0357	0.0221*	0.0338*
18	0.0068*	-	-	-	-
19	0.0068*	-	-	-	0.0338*
22	0.0068*	-	-	-	-
23	-	-	-	-	0.0338*
16;18	0.0068*	-	-	-	-

Table 7.2, continued

STR locus: D6S1043	SAFBS (N=319)	SAFMX (N=40)	SAFWH (N=84)	ZIMBS(N=113)	NGRBS (N=73)
9	0.0110	-	-	0.0221*	-
10	0.0078*	0.0625*	0.0298*	0.0221*	-
10.3	0.0078*	-	-	-	-
11	0.0690	0.1750	0.2976	0.0973	0.0676
12	0.2335	0.2750	0.2500	0.2257	0.2095
12.3	0.0078*	-	-	-	-
13	0.0940	0.0875	0.0774	0.0973	0.1014
14	0.0784	0.0625*	0.0655	0.0442	0.0541
15	0.0737	0.0625*	0.0357	0.0664	0.0473
16	0.0392	0.0625*	-	0.0398	0.0541
17	0.0846	0.0875	0.0476	0.0929	0.0811
18	0.0611	0.0625*	0.1071	0.0841	0.1216
19	0.1379	0.1500	0.0476	0.1460	0.1689
20	0.0862	0.0625*	0.0417	0.0619	0.0743
21	0.0078*	-	0.0298*	0.0221*	0.0338*
21.3	0.0078*	-	-	-	-
22	0.0078*	-	-	-	-
23	0.0078*	-	-	0.0221*	-
24	0.0078*	-	-	0.0221*	0.0338*
STR locus: D1S1656	**SAFBS (N=366)**	**SAFMX (N=40)**	**SAFWH (N=84)**	**ZIMBS(N=113)**	**NGRBS (N=73)**
7	-	-	-	-	0.0338*
8	-	0.0625*	-	-	-
10	0.0068*	-	-	-	0.0338*
11	0.0820	0.0875	0.0655	0.0708	0.0473
12	0.0314	0.0625*	0.1012	0.0442	0.0405
13	0.1052	0.0625*	0.0357	0.0885	0.1149
13.3	0.0068*	-	0.0298*	-	-
14	0.2678	0.1750	0.0952	0.3186	0.2297
14.3	0.0273	0.0625*	0.0298*	0.0398	0.0338*
15	0.1694	0.2125	0.1905	0.1726	0.1824
15.3	0.0068*	0.0625*	0.0536	0.0221*	0.0405
16	0.1612	0.1500	0.1131	0.1283	0.1419
16.3	0.0861	0.0750	0.0476	0.1195	0.1014
17	0.0423	0.0625*	0.0298	0.0221*	0.0338*
17.3	0.0068*	0.0625*	0.1548	-	0.0338*
18	0.0068*	0.0625*	-	-	-
18.3	0.0096	0.0625*	0.0952	0.0221*	0.0338*
19.3	-	-	0.0298*	-	-

Table 7.2, continued

STR locus: D2S1338	SAFBS (N=366)	SAFMX (N=40)	SAFWH (N=84)	ZIMBS(N=113)	NGRBS (N=73)
14	-	-	-	0.0221*	-
15	0.0068*	-	-	-	0.0338*
16	0.0888	0.0875	0.0298*	0.0575	0.0338*
17	0.0533	0.1000	0.2024	0.0619	0.0676
18	0.0492	0.1000	0.0774	0.0575	0.0405
19	0.1626	0.1625	0.0952	0.1460	0.1892
19.2	0.0068*	-	-	-	-
20	0.0929	0.0625*	0.1548	0.0796	0.0878
21	0.1708	0.1250	0.0655	0.1637	0.1622
22	0.1872	0.1000	0.0417	0.1726	0.1824
23	0.0792	0.0875	0.1250	0.0796	0.0878
24	0.0369	0.0625*	0.0774	0.0708	0.0946
25	0.0505	0.0875	0.1071	0.0575	0.0338*
26	0.0219	0.0625*	0.0298*	0.0442	0.0338*
27	0.0068*	-	-	0.0221*	-
STR locus: Penta_D	**SAFBS (N=366)**	**SAFMX (N=40)**	**SAFWH (N=84)**	**ZIMBS(N=113)**	**NGRBS (N=73)**
2.2	0.1489	0.0625*	-	0.1770	0.1486
3.2	0.0068*	-	-	0.0221*	-
5	0.0478	0.0625*	-	0.0398	0.0608
6	0.0246	-	-	0.0221*	0.0338*
6.4	0.0068*	-	-	-	-
7	0.0273	0.0625*	0.0298*	0.0398	0.0338*
8	0.0915	0.0750	0.0417	0.1195	0.1014
9	0.1639	0.1375	0.2857	0.1593	0.1622
10	0.1475	0.1375	0.1071	0.1504	0.1014
11	0.1762	0.1500	0.1369	0.1858	0.1757
11.4	0.0068*	-	-	-	-
12	0.1393	0.1750	0.1429	0.0752	0.1284
12.1	0.0068*	-	-	-	-
12.2	0.0068*	-	-	0.0221*	-
13	0.0205	0.1375	0.1964	0.0221*	0.0541
14	0.0068*	0.0625*	0.0655	0.0221*	0.0338*
15	-	-	0.0298*	-	0.0338*
28	-	-	-	-	0.0338*
32.2	-	-	-	-	0.0338*

7.7 APPENDIX E: Linkage disequilibrium results

Table 7.3 *P-values* for pairwise analysis of all the STR loci for the SA Bantu-speakers

Population	Locus#1	Locus#2	*P*-Value	Population	Locus#1	Locus#2	*P*-Value
SAFBS366	D3S13581	vWA	0.3995	SAFBS366	CSF1PO	D2S441	0.2862
SAFBS366	D3S13581	D16S539	0.8329	SAFBS366	TPOX	D2S441	0.3177
SAFBS366	vWA	D16S539	0.6588	SAFBS366	D8S1179	D2S441	0.4952
SAFBS366	D3S13581	CSF1PO	0.0976	SAFBS366	D21S11	D2S441	**0.0326**
SAFBS366	vWA	CSF1PO	0.3201	SAFBS366	D18S51	D2S441	0.3462
SAFBS366	D16S539	CSF1PO	0.9806	SAFBS366	Penta_E	D2S441	0.8428
SAFBS366	D3S13581	TPOX	0.9443	SAFBS366	D3S13581	D19S433	0.2591
SAFBS366	vWA	TPOX	0.8430	SAFBS366	vWA	D19S433	0.7587
SAFBS366	D16S539	TPOX	0.2805	SAFBS366	D16S539	D19S433	0.7914
SAFBS366	CSF1PO	TPOX	0.2301	SAFBS366	CSF1PO	D19S433	0.0536
SAFBS366	D3S13581	D8S1179	0.9943	SAFBS366	TPOX	D19S433	0.8823
SAFBS366	vWA	D8S1179	0.2744	SAFBS366	D8S1179	D19S433	0.8484
SAFBS366	D16S539	D8S1179	0.3069	SAFBS366	D21S11	D19S433	0.5659
SAFBS366	CSF1PO	D8S1179	0.9313	SAFBS366	D18S51	D19S433	0.1572
SAFBS366	TPOX	D8S1179	0.2630	SAFBS366	Penta_E	D19S433	0.1891
SAFBS366	D3S13581	D21S11	0.1425	SAFBS366	D2S441	D19S433	0.2107
SAFBS366	vWA	D21S11	0.6514	SAFBS366	D3S13581	TH01	0.8598
SAFBS366	D16S539	D21S11	0.3088	SAFBS366	vWA	TH01	0.8045
SAFBS366	CSF1PO	D21S11	**0.0000**	SAFBS366	D16S539	TH01	0.3842
SAFBS366	TPOX	D21S11	0.0690	SAFBS366	CSF1PO	TH01	0.8092
SAFBS366	D8S1179	D21S11	0.5180	SAFBS366	TPOX	TH01	0.1248
SAFBS366	D3S13581	D18S51	0.7733	SAFBS366	D8S1179	TH01	0.7292
SAFBS366	vWA	D18S51	0.2757	SAFBS366	D21S11	TH01	0.3225
SAFBS366	D16S539	D18S51	**0.0107**	SAFBS366	D18S51	TH01	0.1015
SAFBS366	CSF1PO	D18S51	0.7044	SAFBS366	Penta_E	TH01	0.7463
SAFBS366	TPOX	D18S51	0.3998	SAFBS366	D2S441	TH01	0.8885
SAFBS366	D8S1179	D18S51	0.5402	SAFBS366	D19S433	TH01	0.9029
SAFBS366	D21S11	D18S51	0.9432	SAFBS366	D3S13581	FGA	0.2295
SAFBS366	D3S13581	Penta_E	0.7640	SAFBS366	vWA	FGA	0.1558
SAFBS366	TPOX	D13S317	0.3012	SAFBS366	D22S1045	D6S1043	0.3294

*Loci in LD are indicated in bold

Table 7.3, continued

Population	Locus#1	Locus#2	*P*-Value	Population	Locus#1	Locus#2	*P*-Value
SAFBS366	vWA	Penta_E	0.9685	SAFBS366	D16S539	FGA	0.6537
SAFBS366	D16S539	Penta_E	0.4051	SAFBS366	CSF1PO	FGA	0.2492
SAFBS366	CSF1PO	Penta_E	**0.0000**	SAFBS366	TPOX	FGA	0.3089
SAFBS366	TPOX	Penta_E	0.9326	SAFBS366	D8S1179	FGA	0.4585
SAFBS366	D8S1179	Penta_E	0.4198	SAFBS366	D21S11	FGA	0.9578
SAFBS366	D21S11	Penta_E	0.1118	SAFBS366	D18S51	FGA	**0.0347**
SAFBS366	D18S51	Penta_E	0.1442	SAFBS366	Penta_E	FGA	0.5115
SAFBS366	D3S13581	D2S441	0.9638	SAFBS366	D2S441	FGA	0.7186
SAFBS366	vWA	D2S441	0.2755	SAFBS366	D19S433	FGA	0.1699
SAFBS366	D16S539	D2S441	0.5891	SAFBS366	TH01	FGA	0.1667
SAFBS366	vWA	D22S1045	0.2195	SAFBS366	D5S818	D13S317	0.2118
SAFBS366	D16S539	D22S1045	**0.0101**	SAFBS366	D3S13581	D7S820	0.5289
SAFBS366	CSF1PO	D22S1045	0.9516	SAFBS366	vWA	D7S820	0.3037
SAFBS366	TPOX	D22S1045	0.3111	SAFBS366	D16S539	D7S820	0.6300
SAFBS366	D8S1179	D22S1045	0.6800	SAFBS366	CSF1PO	D7S820	0.2488
SAFBS366	D21S11	D22S1045	0.5339	SAFBS366	TPOX	D7S820	0.4047
SAFBS366	D18S51	D22S1045	0.2303	SAFBS366	D8S1179	D7S820	0.7464
SAFBS366	Penta_E	D22S1045	0.9529	SAFBS366	D21S11	D7S820	**0.0088**
SAFBS366	D2S441	D22S1045	0.5838	SAFBS366	D18S51	D7S820	0.6005
SAFBS366	D19S433	D22S1045	0.5118	SAFBS366	Penta_E	D7S820	0.9746
SAFBS366	TH01	D22S1045	0.9024	SAFBS366	D2S441	D7S820	0.0536
SAFBS366	FGA	D22S1045	0.4901	SAFBS366	D19S433	D7S820	0.1778
SAFBS366	D3S13581	D5S818	0.6894	SAFBS366	TH01	D7S820	**0.0188**
SAFBS366	vWA	D5S818	0.7860	SAFBS366	FGA	D7S820	0.1914
SAFBS366	D16S539	D5S818	0.3313	SAFBS366	D22S1045	D7S820	0.0708
SAFBS366	CSF1PO	D5S818	0.2838	SAFBS366	D5S818	D7S820	**0.0119**
SAFBS366	TPOX	D5S818	0.7092	SAFBS366	D13S317	D7S820	0.6868
SAFBS366	D8S1179	D5S818	0.8444	SAFBS366	D3S13581	D6S1043	0.7364
SAFBS366	D21S11	D5S818	0.7281	SAFBS366	vWA	D6S1043	0.4876
SAFBS366	D18S51	D5S818	0.9737	SAFBS366	D16S539	D6S1043	0.7501
SAFBS366	Penta_E	D5S818	0.8796	SAFBS366	CSF1PO	D6S1043	**0.0277**
SAFBS366	D2S441	D5S818	1.0000	SAFBS366	TPOX	D6S1043	0.6845

*Loci in LD are indicated in bold

Table 7.3, continued

Population	Locus#1	Locus#2	*P*-Value	Population	Locus#1	Locus#2	*P*-Value
SAFBS366	D19S433	D5S818	0.0881	SAFBS366	D8S1179	D6S1043	0.8733
SAFBS366	TH01	D5S818	0.0555	SAFBS366	D21S11	D6S1043	**0.0333**
SAFBS366	D8S1179	D13S317	0.6250	SAFBS366	D5S818	D6S1043	0.2222
SAFBS366	D21S11	D13S317	0.5123	SAFBS366	D13S317	D6S1043	0.3681
SAFBS366	D18S51	D13S317	0.3761	SAFBS366	D7S820	D6S1043	0.1669
SAFBS366	Penta_E	D13S317	0.5545	SAFBS366	D3S13581	D10S1248	0.2669
SAFBS366	D2S441	D13S317	0.3566	SAFBS366	vWA	D10S1248	0.3360
SAFBS366	D19S433	D13S317	0.5045	SAFBS366	D16S539	D10S1248	0.4005
SAFBS366	TH01	D13S317	0.7423	SAFBS366	CSF1PO	D10S1248	0.3613
SAFBS366	FGA	D13S317	0.1555	SAFBS366	TPOX	D10S1248	0.1202
SAFBS366	D22S1045	D13S317	0.1075	SAFBS366	D8S1179	D10S1248	0.3132
SAFBS366	D21S11	D10S1248	0.5030	SAFBS366	D2S441	D12S391	0.4869
SAFBS366	D18S51	D10S1248	0.1976	SAFBS366	D19S433	D12S391	0.7593
SAFBS366	Penta_E	D10S1248	1.0000	SAFBS366	TH01	D12S391	0.2033
SAFBS366	D2S441	D10S1248	0.8196	SAFBS366	FGA	D12S391	0.0988
SAFBS366	D19S433	D10S1248	0.3372	SAFBS366	D22S1045	D12S391	0.8640
SAFBS366	TH01	D10S1248	0.2611	SAFBS366	D5S818	D12S391	0.5836
SAFBS366	FGA	D10S1248	0.3651	SAFBS366	D13S317	D12S391	**0.0011**
SAFBS366	D22S1045	D10S1248	0.5220	SAFBS366	D7S820	D12S391	0.2708
SAFBS366	D5S818	D10S1248	0.0938	SAFBS366	D6S1043	D12S391	0.6888
SAFBS366	D13S317	D10S1248	0.5591	SAFBS366	D10S1248	D12S391	0.9288
SAFBS366	D7S820	D10S1248	0.3576	SAFBS366	D1S1656	D12S391	0.2179
SAFBS366	D6S1043	D10S1248	0.2702	SAFBS366	D3S13581	D2S1338	0.9366
SAFBS366	D3S13581	D1S1656	0.2803	SAFBS366	vWA	D2S1338	0.3890
SAFBS366	vWA	D1S1656	0.9466	SAFBS366	D16S539	D2S1338	0.0807
SAFBS366	D16S539	D1S1656	0.9899	SAFBS366	CSF1PO	D2S1338	0.8002
SAFBS366	CSF1PO	D1S1656	0.2688	SAFBS366	TPOX	D2S1338	0.7922
SAFBS366	TPOX	D1S1656	0.7758	SAFBS366	D8S1179	D2S1338	0.3938
SAFBS366	D8S1179	D1S1656	0.1309	SAFBS366	D21S11	D2S1338	0.7856
SAFBS366	D21S11	D1S1656	0.9203	SAFBS366	D18S51	D2S1338	0.1363
SAFBS366	D18S51	D1S1656	0.2030	SAFBS366	Penta_E	D2S1338	0.2057
SAFBS366	Penta_E	D1S1656	0.6057	SAFBS366	D2S441	D2S1338	0.3082
SAFBS366	D2S441	D1S1656	0.1582	SAFBS366	D19S433	D2S1338	**0.0273**

*Loci in LD are indicated in bold

Table 7.3, continued

Population	Locus#1	Locus#2	*P*-Value	Population	Locus#1	Locus#2	*P*-Value
SAFBS366	D19S433	D1S1656	0.7954	SAFBS366	TH01	D2S1338	0.0873
SAFBS366	TH01	D1S1656	0.2632	SAFBS366	FGA	D2S1338	0.4127
SAFBS366	D10S1248	D1S1656	0.7872	SAFBS366	FGA	D5S818	0.5345
SAFBS366	D1S1656	D2S1338	0.8956	SAFBS366	D22S1045	D5S818	0.7597
SAFBS366	Penta_E	Penta_D	0.4817	SAFBS366	D3S13581	D13S317	0.3829
SAFBS366	D3S13581	D12S391	0.6844	SAFBS366	vWA	D13S317	0.0601
SAFBS366	vWA	D12S391	0.2930	SAFBS366	D16S539	D13S317	0.6222
SAFBS366	D12S391	D2S1338	0.2003	SAFBS366	CSF1PO	D13S317	0.7965
SAFBS366	D3S13581	Penta_D	0.8419	SAFBS366	D18S51	D6S1043	0.9866
SAFBS366	D2S441	Penta_D	0.2751	SAFBS366	Penta_E	D6S1043	0.1424
SAFBS366	D19S433	Penta_D	0.3837	SAFBS366	D2S441	D6S1043	0.9518
SAFBS366	TH01	D6S1043	0.6652	SAFBS366	D19S433	D6S1043	**0.0147**
SAFBS366	TH01	Penta_D	0.1006	SAFBS366	FGA	D6S1043	0.4138
SAFBS366	FGA	Penta_D	0.3535	SAFBS366	D2S1338	Penta_D	0.7406
SAFBS366	D22S1045	Penta_D	0.0739	SAFBS366	D3S13581	D22S1045	0.9845
SAFBS366	D5S818	Penta_D	0.8060	SAFBS366	vWA	Penta_D	0.4875
SAFBS366	D13S317	Penta_D	0.9295	SAFBS366	D16S539	Penta_D	0.4390
SAFBS366	D7S820	Penta_D	0.3190	SAFBS366	CSF1PO	Penta_D	0.9532
SAFBS366	D6S1043	Penta_D	0.3673	SAFBS366	TPOX	Penta_D	0.6386
SAFBS366	D10S1248	Penta_D	0.8419	SAFBS366	D8S1179	Penta_D	0.8674
SAFBS366	D1S1656	Penta_D	0.2925	SAFBS366	D21S11	Penta_D	0.4848
SAFBS366	D12S391	Penta_D	0.2438	SAFBS366	D18S51	Penta_D	0.0830
SAFBS366	FGA	D1S1656	0.8994	SAFBS366	D22S1045	D2S1338	**0.0252**
SAFBS366	D22S1045	D1S1656	0.8180	SAFBS366	D5S818	D2S1338	0.0960
SAFBS366	D5S818	D1S1656	0.8909	SAFBS366	D13S317	D2S1338	**0.0169**
SAFBS366	D13S317	D1S1656	0.1287	SAFBS366	D7S820	D2S1338	0.4855
SAFBS366	D7S820	D1S1656	0.0775	SAFBS366	D6S1043	D2S1338	0.6698
SAFBS366	FGA	D1S1656	0.8994	SAFBS366	D22S1045	D2S1338	**0.0252**
SAFBS366	D6S1043	D1S1656	0.6238	SAFBS366	D10S1248	D2S1338	0.0699

* Loci in LD are indicated in bold

Table 7.4 *P-values* for pairwise analysis of loci all the STR for the SA Whites

Population	Locus#1	Locus#2	*P*-Value	Population	Locus#1	Locus#2	*P*-Value
SAFWH84	D3S1358	vWA	0.5873	SAFWH84	D16S539	D2S441	0.0673
SAFWH84	D3S1358	D16S539	0.3681	SAFWH84	CSF1PO	D2S441	0.4818
SAFWH84	vWA	D16S539	0.4066	SAFWH84	TPOX	D2S441	0.8437
SAFWH84	D3S1358	CSF1PO	0.1355	SAFWH84	D8S1179	D2S441	0.1337
SAFWH84	vWA	CSF1PO	0.6831	SAFWH84	D21S11	D2S441	0.3233
SAFWH84	D16S539	CSF1PO	0.1500	SAFWH84	D18S51	D2S441	0.1611
SAFWH84	D3S1358	TPOX	0.8767	SAFWH84	Penta_E	D2S441	**0.0000**
SAFWH84	vWA	TPOX	0.0664	SAFWH84	D3S1358	D19S433	0.5558
SAFWH84	D16S539	TPOX	0.2215	SAFWH84	vWA	D19S433	0.1135
SAFWH84	CSF1PO	TPOX	0.1128	SAFWH84	D16S539	D19S433	0.8706
SAFWH84	D3S1358	D8S1179	0.1677	SAFWH84	CSF1PO	D19S433	0.8564
SAFWH84	vWA	D8S1179	0.8987	SAFWH84	TPOX	D19S433	**0.0127**
SAFWH84	D16S539	D8S1179	0.3457	SAFWH84	D8S1179	D19S433	0.6980
SAFWH84	CSF1PO	D8S1179	0.6726	SAFWH84	D21S11	D19S433	0.3768
SAFWH84	TPOX	D8S1179	0.6557	SAFWH84	D18S51	D19S433	0.7517
SAFWH84	D3S1358	D21S11	0.1881	SAFWH84	Penta_E	D19S433	1.0000
SAFWH84	vWA	D21S11	0.8618	SAFWH84	D2S441	D19S433	0.8958
SAFWH84	D16S539	D21S11	**0.0000**	SAFWH84	D3S1358	TH01	0.7429
SAFWH84	CSF1PO	D21S11	**0.0005**	SAFWH84	vWA	TH01	0.1535
SAFWH84	TPOX	D21S11	0.2927	SAFWH84	D16S539	TH01	0.0486
SAFWH84	D8S1179	D21S11	0.9055	SAFWH84	CSF1PO	TH01	0.9855
SAFWH84	D3S1358	D18S51	0.3854	SAFWH84	TPOX	TH01	**0.0446**
SAFWH84	vWA	D18S51	0.4222	SAFWH84	D8S1179	TH01	0.4160
SAFWH84	D16S539	D18S51	0.8917	SAFWH84	D21S11	TH01	0.7306
SAFWH84	CSF1PO	D18S51	0.5168	SAFWH84	D18S51	TH01	0.2422
SAFWH84	TPOX	D18S51	0.6542	SAFWH84	Penta_E	TH01	0.7440
SAFWH84	D8S1179	D18S51	0.5798	SAFWH84	D2S441	TH01	0.4441
SAFWH84	D21S11	D18S51	0.9766	SAFWH84	D19S433	TH01	0.8504
SAFWH84	D3S1358	Penta_E	0.1271	SAFWH84	D3S1358	FGA	0.7345
SAFWH84	vWA	Penta_E	0.8387	SAFWH84	vWA	FGA	**0.0311**
SAFWH84	D16S539	Penta_E	0.2745	SAFWH84	D16S539	FGA	0.1139
SAFWH84	CSF1PO	Penta_E	0.5651	SAFWH84	CSF1PO	FGA	**0.0199**
SAFWH84	TPOX	Penta_E	0.9228	SAFWH84	TPOX	FGA	0.2330

* Loci in LD are indicated in bold

Table 7.4, continued

Population	Locus #1	Locus #2	*P*-Value	Population	Locus #1	Locus #2	*P*-Value
SAFWH84	D8S1179	Penta_E	0.4202	SAFWH84	D8S1179	FGA	0.8225
SAFWH84	D21S11	Penta_E	0.6711	SAFWH84	D21S11	FGA	0.5158
SAFWH84	D18S51	Penta_E	0.3624	SAFWH84	D18S51	FGA	0.0533
SAFWH84	D3S1358	D2S441	0.3247	SAFWH84	Penta_E	FGA	0.2705
SAFWH84	vWA	D2S441	0.2241	SAFWH84	D2S441	FGA	0.3529
SAFWH84	D3S1358	D22S1045	0.6190	SAFWH84	D22S104	D13S317	0.9369
SAFWH84	vWA	D22S1045	0.9559	SAFWH84	D5S818	D13S317	0.8866
SAFWH84	D16S539	D22S1045	0.5374	SAFWH84	D3S1358	D7S820	0.1962
SAFWH84	CSF1PO	D22S1045	0.1762	SAFWH84	vWA	D7S820	0.6146
SAFWH84	TPOX	D22S1045	0.8195	SAFWH84	D16S539	D7S820	0.5637
SAFWH84	D8S1179	D22S1045	0.8366	SAFWH84	CSF1PO	D7S820	0.5954
SAFWH84	D21S11	D22S1045	0.8022	SAFWH84	TPOX	D7S820	0.5529
SAFWH84	D18S51	D22S1045	0.6751	SAFWH84	D8S1179	D7S820	0.1711
SAFWH84	Penta_E	D22S1045	0.7931	SAFWH84	D21S11	D7S820	0.6018
SAFWH84	D2S441	D22S1045	0.2466	SAFWH84	D18S51	D7S820	0.5453
SAFWH84	D19S433	D22S1045	0.1271	SAFWH84	Penta_E	D7S820	0.2106
SAFWH84	TH01	D22S1045	0.8476	SAFWH84	D2S441	D7S820	0.0633
SAFWH84	FGA	D22S1045	0.4445	SAFWH84	D19S433	D7S820	0.5272
SAFWH84	D3S1358	D5S818	0.1144	SAFWH84	TH01	D7S820	0.1841
SAFWH84	vWA	D5S818	0.1142	SAFWH84	FGA	D7S820	0.7406
SAFWH84	D16S539	D5S818	0.8063	SAFWH84	D22S104	D7S820	0.1453
SAFWH84	CSF1PO	D5S818	0.2029	SAFWH84	D5S818	D7S820	0.3994
SAFWH84	TPOX	D5S818	0.3575	SAFWH84	D13S317	D7S820	0.6583
SAFWH84	D8S1179	D5S818	0.3659	SAFWH84	D3S1358	D6S1043	**0.0188**
SAFWH84	D21S11	D5S818	0.8175	SAFWH84	vWA	D6S1043	0.4115
SAFWH84	D18S51	D5S818	0.4107	SAFWH84	D16S539	D6S1043	0.7697
SAFWH84	Penta_E	D5S818	0.4668	SAFWH84	CSF1PO	D6S1043	0.3313
SAFWH84	D2S441	D5S818	0.7175	SAFWH84	TPOX	D6S1043	0.8872
SAFWH84	D19S433	D5S818	0.1401	SAFWH84	D8S1179	D6S1043	**0.0356**
SAFWH84	TH01	D5S818	0.9341	SAFWH84	D21S11	D6S1043	0.4189
SAFWH84	FGA	D5S818	0.1818	SAFWH84	D18S51	D6S1043	0.9182
SAFWH84	D22S104	D5S818	0.8670	SAFWH84	Penta_E	D6S1043	0.2251
SAFWH84	D3S1358	D13S317	0.7852	SAFWH84	D2S441	D6S1043	0.0853

* Loci in LD are indicated in bold

Table 7.4, continued

Population	Locus#1	Locus#2	*P*-Value	Population	Locus#1	Locus#2	*P*-Value
SAFWH84	D16S539	D13S317	0.7759	SAFWH84	TH01	D6S1043	0.9116
SAFWH84	D16S539	D13S317	0.7759	SAFWH84	TH01	D6S1043	0.9116
SAFWH84	CSF1PO	D13S317	0.0707	SAFWH84	FGA	D6S1043	0.1551
SAFWH84	TPOX	D13S317	0.4849	SAFWH84	D22S104	D6S1043	0.6288
SAFWH84	D8S1179	D13S317	0.8742	SAFWH84	D5S818	D6S1043	0.5101
SAFWH84	D21S11	D13S317	0.2935	SAFWH84	D13S317	D6S1043	0.3253
SAFWH84	D18S51	D13S317	0.3465	SAFWH84	D7S820	D6S1043	0.5951
SAFWH84	Penta_E	D13S317	0.2841	SAFWH84	D3S1358	D10S1248	0.2376
SAFWH84	D2S441	D13S317	0.0892	SAFWH84	vWA	D10S1248	0.1717
SAFWH84	D19S433	D13S317	0.0934	SAFWH84	D16S539	D10S1248	0.7087
SAFWH84	TH01	D13S317	0.6641	SAFWH84	CSF1PO	D10S1248	0.0672
SAFWH84	FGA	D13S317	0.2335	SAFWH84	TPOX	D10S1248	0.1958
SAFWH84	D8S1179	D10S1248	0.0828	SAFWH84	Penta_E	D12S391	0.6692
SAFWH84	D21S11	D10S1248	0.3466	SAFWH84	D2S441	D12S391	0.6779
SAFWH84	D18S51	D10S1248	0.9445	SAFWH84	D19S433	D12S391	0.8255
SAFWH84	Penta_E	D10S1248	0.1762	SAFWH84	TH01	D12S391	0.4052
SAFWH84	D2S441	D10S1248	0.6608	SAFWH84	FGA	D12S391	0.5631
SAFWH84	D19S433	D10S1248	0.6687	SAFWH84	D22S104	D12S391	0.6781
SAFWH84	TH01	D10S1248	0.5195	SAFWH84	D5S818	D12S391	**0.0449**
SAFWH84	FGA	D10S1248	0.2378	SAFWH84	D13S317	D12S391	0.1525
SAFWH84	D22S104	D10S1248	0.2528	SAFWH84	D7S820	D12S391	**0.0035**
SAFWH84	D5S818	D10S1248	0.2924	SAFWH84	D6S1043	D12S391	0.6860
SAFWH84	D13S317	D10S1248	0.7537	SAFWH84	D10S124	D12S391	**0.0000**
SAFWH84	D7S820	D10S1248	0.5208	SAFWH84	D1S1656	D12S391	0.4778
SAFWH84	D6S1043	D10S1248	0.5050	SAFWH84	D3S1358	D2S1338	0.1855
SAFWH84	D3S1358	D1S1656	**0.0000**	SAFWH84	vWA	D2S1338	0.8703
SAFWH84	vWA	D1S1656	0.4745	SAFWH84	D16S539	D2S1338	0.7820
SAFWH84	D16S539	D1S1656	0.0648	SAFWH84	CSF1PO	D2S1338	0.9752
SAFWH84	CSF1PO	D1S1656	0.0873	SAFWH84	TPOX	D2S1338	0.2272
SAFWH84	TPOX	D1S1656	0.3421	SAFWH84	D8S1179	D2S1338	0.7869
SAFWH84	D8S1179	D1S1656	0.3320	SAFWH84	D21S11	D2S1338	0.3110
SAFWH84	D21S11	D1S1656	0.1035	SAFWH84	D18S51	D2S1338	0.0538
SAFWH84	D18S51	D1S1656	0.4619	SAFWH84	Penta_E	D2S1338	**0.0001**

* Loci in LD are indicated in bold

Table 7.4, continued

Population	Locus#1	Locus#2	*P*-Value	Population	Locus#1	Locus#2	*P*-Value
SAFWH84	Penta_E	D1S1656	0.3421	SAFWH84	D2S441	D2S1338	0.9868
SAFWH84	D2S441	D1S1656	0.9791	SAFWH84	D19S433	D2S1338	0.1495
SAFWH84	D19S433	D1S1656	0.7052	SAFWH84	TH01	D2S1338	**0.0158**
SAFWH84	TH01	D1S1656	0.1682	SAFWH84	FGA	D2S1338	**0.0306**
SAFWH84	FGA	D1S1656	0.3021	SAFWH84	D22S104	D2S1338	0.7492
SAFWH84	D22S104	D1S1656	0.3008	SAFWH84	D5S818	D2S1338	0.1093
SAFWH84	D5S818	D1S1656	0.8619	SAFWH84	D13S317	D2S1338	0.0780
SAFWH84	D13S317	D1S1656	0.9741	SAFWH84	D7S820	D2S1338	0.2810
SAFWH84	D7S820	D1S1656	0.3247	SAFWH84	D6S1043	D2S1338	0.6161
SAFWH84	D6S1043	D1S1656	0.5728	SAFWH84	D10S124	D2S1338	0.6335
SAFWH84	D10S124	D1S1656	0.0868	SAFWH84	D1S1656	D2S1338	0.5855
SAFWH84	D3S1358	D12S391	0.4314	SAFWH84	D12S391	D2S1338	0.8964
SAFWH84	vWA	D12S391	**0.0197**	SAFWH84	D3S1358	Penta_D	0.0721
SAFWH84	D16S539	D12S391	0.7070	SAFWH84	vWA	Penta_D	0.2159
SAFWH84	CSF1PO	D12S391	0.0272	SAFWH84	D16S539	Penta_D	0.7457
SAFWH84	TPOX	D12S391	0.5787	SAFWH84	CSF1PO	Penta_D	0.9961
SAFWH84	D8S1179	D12S391	0.4300	SAFWH84	TPOX	Penta_D	0.5898
SAFWH84	D21S11	D12S391	**0.0042**	SAFWH84	Penta_E	Penta_D	0.6377
SAFWH84	D8S1179	Penta_D	0.6645	SAFWH84	D2S441	Penta_D	0.5079
SAFWH84	D18S51	D12S391	0.3187	SAFWH84	D19S433	Penta_D	**0.0413**
SAFWH84	D21S11	Penta_D	0.4099	SAFWH84	TH01	Penta_D	0.6069
SAFWH84	D18S51	Penta_D	0.6430	SAFWH84	FGA	Penta_D	0.8499
SAFWH84	D22S104	Penta_D	0.7264	SAFWH84	D10S124	Penta_D	0.3440
SAFWH84	D5S818	Penta_D	0.8842	SAFWH84	D1S1656	Penta_D	0.9069
SAFWH84	D13S317	Penta_D	0.3731	SAFWH84	D12S391	Penta_D	0.1143
SAFWH84	D7S820	Penta_D	0.5527	SAFWH84	D2S1338	Penta_D	0.6137
SAFWH84	D6S1043	Penta_D	0.0708	SAFWH84	D10S124	Penta_D	0.3440

* Loci in LD are indicated in bold

Table 7.5 *P*-values for pairwise analysis of all the STR loci for the SA Coloureds

Population	Locus#1	Locus#2	*P*-Value	Population	Locus#1	Locus#2	*P*-Value
SAFMX40	D3S1358	vWA	0.3709	SAFMX40	CSF1PO	D2S441	0.8982
SAFMX40	D3S1358	D16S539	0.9532	SAFMX40	TPOX	D2S441	0.9762
SAFMX40	vWA	D16S539	0.1197	SAFMX40	D8S1179	D2S441	0.2717
SAFMX40	D3S1358	CSF1PO	0.2249	SAFMX40	D21S11	D2S441	1.0000
SAFMX40	vWA	CSF1PO	0.9735	SAFMX40	D18S51	D2S441	0.7113
SAFMX40	D16S539	CSF1PO	0.2484	SAFMX40	Penta_E	D2S441	1.0000
SAFMX40	D3S1358	TPOX	0.8544	SAFMX40	D3S13581	D19S433	0.5625
SAFMX40	vWA	TPOX	0.6218	SAFMX40	vWA	D19S433	0.3804
SAFMX40	D16S539	TPOX	0.5694	SAFMX40	D16S539	D19S433	0.8792
SAFMX40	CSF1PO	TPOX	0.5053	SAFMX40	CSF1PO	D19S433	0.7136
SAFMX40	D3S1358	D8S1179	0.0831	SAFMX40	TPOX	D19S433	0.7000
SAFMX40	vWA	D8S1179	0.7852	SAFMX40	D8S1179	D19S433	0.0532
SAFMX40	D16S539	D8S1179	0.3696	SAFMX40	D21S11	D19S433	1.0000
SAFMX40	CSF1PO	D8S1179	0.9428	SAFMX40	D18S51	D19S433	0.7937
SAFMX40	TPOX	D8S1179	0.4513	SAFMX40	Penta_E	D19S433	**0.0358**
SAFMX40	D3S1358	D21S11	0.7807	SAFMX40	D2S441	D19S433	0.2092
SAFMX40	vWA	D21S11	1.0000	SAFMX40	D3S13581	TH01	0.6323
SAFMX40	D16S539	D21S11	1.0000	SAFMX40	vWA	TH01	0.5023
SAFMX40	CSF1PO	D21S11	1.0000	SAFMX40	D16S539	TH01	0.8607
SAFMX40	TPOX	D21S11	1.0000	SAFMX40	CSF1PO	TH01	0.8167
SAFMX40	D8S1179	D21S11	1.0000	SAFMX40	TPOX	TH01	0.2198
SAFMX40	D3S1358	D18S51	1.0000	SAFMX40	D8S1179	TH01	0.1870
SAFMX40	vWA	D18S51	0.2951	SAFMX40	D21S11	TH01	1.0000
SAFMX40	D16S539	D18S51	0.6151	SAFMX40	D18S51	TH01	1.0000
SAFMX40	CSF1PO	D18S51	0.8733	SAFMX40	Penta_E	TH01	1.0000
SAFMX40	TPOX	D18S51	0.8617	SAFMX40	D2S441	TH01	0.5568
SAFMX40	D8S1179	D18S51	0.8670	SAFMX40	D19S433	TH01	0.6205
SAFMX40	D21S11	D18S51	1.0000	SAFMX40	D3S13581	FGA	0.8940
SAFMX40	D3S13581	Penta_E	1.0000	SAFMX40	vWA	FGA	0.0270
SAFMX40	vWA	Penta_E	**0.0183**	SAFMX40	D16S539	FGA	1.0000

* Loci in LD are indicated in bold

Table 7.5, continued

Population	Locus#1	Locus#2	*P*-Value	Population	Locus#1	Locus#2	*P*-Value
SAFMX40	D16S539	Penta_E	1.0000	SAFMX40	CSF1PO	FGA	1.0000
SAFMX40	CSF1PO	Penta_E	0.2812	SAFMX40	TPOX	FGA	0.4892
SAFMX40	TPOX	Penta_E	1.0000	SAFMX40	D8S1179	FGA	0.8568
SAFMX40	D8S1179	Penta_E	0.4178	SAFMX40	D21S11	FGA	1.0000
SAFMX40	D21S11	Penta_E	1.0000	SAFMX40	D18S51	FGA	1.0000
SAFMX40	D18S51	Penta_E	1.0000	SAFMX40	Penta_E	FGA	1.0000
SAFMX40	D3S13581	D2S441	0.6210	SAFMX40	D2S441	FGA	0.6133
SAFMX40	vWA	D2S441	0.5060	SAFMX40	D19S433	FGA	0.3570
SAFMX40	D16S539	D2S441	0.1508	SAFMX40	TH01	FGA	0.8148
SAFMX40	D2S1338	Penta_D	0.3991	SAFMX40	D3S13581	D22S1045	0.9909
SAFMX40	vWA	D22S1045	0.4515	SAFMX40	D5S818	D13S317	0.6914
SAFMX40	D16S539	D22S1045	0.2947	SAFMX40	D3S13581	D7S820	0.2129
SAFMX40	CSF1PO	D22S1045	0.1948	SAFMX40	vWA	D7S820	0.9769
SAFMX40	TPOX	D22S1045	0.4997	SAFMX40	D16S539	D7S820	0.8777
SAFMX40	D8S1179	D22S1045	0.9152	SAFMX40	CSF1PO	D7S820	0.4562
SAFMX40	D21S11	D22S1045	1.0000	SAFMX40	TPOX	D7S820	0.6427
SAFMX40	D18S51	D22S1045	0.7895	SAFMX40	D8S1179	D7S820	0.9222
SAFMX40	Penta_E	D22S1045	0.1078	SAFMX40	D21S11	D7S820	0.5308
SAFMX40	D2S441	D22S1045	0.3249	SAFMX40	D18S51	D7S820	1.0000
SAFMX40	D19S433	D22S1045	0.4758	SAFMX40	Penta_E	D7S820	1.0000
SAFMX40	TH01	D22S1045	0.5127	SAFMX40	D2S441	D7S820	0.5475
SAFMX40	FGA	D22S1045	0.7961	SAFMX40	D19S433	D7S820	0.9683
SAFMX40	D3S13581	D5S818	0.7982	SAFMX40	TH01	D7S820	0.6773
SAFMX40	vWA	D5S818	0.9862	SAFMX40	FGA	D7S820	0.1993
SAFMX40	D16S539	D5S818	0.2992	SAFMX40	D22S1045	D7S820	0.3097
SAFMX40	CSF1PO	D5S818	0.8289	SAFMX40	D5S818	D7S820	0.0872
SAFMX40	TPOX	D5S818	0.2387	SAFMX40	D13S317	D7S820	0.7342
SAFMX40	D13S317	D10S1248	0.9899	SAFMX40	D10S1248	D12S391	0.5854
SAFMX40	D7S820	D10S1248	0.3096	SAFMX40	D1S1656	D12S391	1.0000
SAFMX40	D6S1043	D10S1248	0.5543	SAFMX40	D3S13581	D2S1338	0.2158
SAFMX40	D3S13581	D1S1656	0.2208	SAFMX40	vWA	D2S1338	1.0000
SAFMX40	vWA	D1S1656	**0.0000**	SAFMX40	D16S539	D2S1338	0.4117

* Loci in LD are indicated in bold

Table 7.5, continued

Population	Locus#1	Locus#2	*P*-Value	Population	Locus#1	Locus#2	*P*-Value
SAFMX40	D8S1179	D5S818	0.9234	SAFMX40	D3S13581	D6S1043	0.8347
SAFMX40	D21S11	D5S818	1.0000	SAFMX40	vWA	D6S1043	1.0000
SAFMX40	D18S51	D5S818	0.5676	SAFMX40	D16S539	D6S1043	0.6156
SAFMX40	Penta_E	D5S818	0.5325	SAFMX40	CSF1PO	D6S1043	0.7763
SAFMX40	D2S441	D5S818	0.9849	SAFMX40	TPOX	D6S1043	0.6036
SAFMX40	D19S433	D5S818	0.6760	SAFMX40	D8S1179	D6S1043	0.9442
SAFMX40	TH01	D5S818	0.3805	SAFMX40	D21S11	D6S1043	1.0000
SAFMX40	FGA	D5S818	0.6991	SAFMX40	D18S51	D6S1043	0.3116
SAFMX40	D22S1045	D5S818	0.5007	SAFMX40	Penta_E	D6S1043	1.0000
SAFMX40	D3S13581	D13S317	**0.0111**	SAFMX40	D2S441	D6S1043	0.5506
SAFMX40	vWA	D13S317	0.2596	SAFMX40	D19S433	D6S1043	0.3894
SAFMX40	D16S539	D13S317	0.3743	SAFMX40	TH01	D6S1043	1.0000
SAFMX40	CSF1PO	D13S317	0.4034	SAFMX40	FGA	D6S1043	0.1227
SAFMX40	TPOX	D13S317	0.6411	SAFMX40	D22S1045	D6S1043	0.8199
SAFMX40	D8S1179	D13S317	0.4504	SAFMX40	D5S818	D6S1043	0.7671
SAFMX40	D21S11	D13S317	1.0000	SAFMX40	D13S317	D6S1043	0.7890
SAFMX40	D18S51	D13S317	0.7153	SAFMX40	D7S820	D6S1043	0.6656
SAFMX40	Penta_E	D13S317	0.4060	SAFMX40	D3S13581	D10S1248	0.7720
SAFMX40	D2S441	D13S317	0.4192	SAFMX40	vWA	D10S1248	0.9675
SAFMX40	D19S433	D13S317	0.9096	SAFMX40	D16S539	D10S1248	0.9267
SAFMX40	TH01	D13S317	0.9052	SAFMX40	CSF1PO	D10S1248	0.1862
SAFMX40	FGA	D13S317	0.8827	SAFMX40	TPOX	D10S1248	0.7588
SAFMX40	D22S1045	D13S317	0.3617	SAFMX40	D8S1179	D10S1248	0.5940
SAFMX40	D21S11	D10S1248	1.0000	SAFMX40	D2S441	D12S391	1.0000
SAFMX40	D18S51	D10S1248	0.2150	SAFMX40	D19S433	D12S391	1.0000
SAFMX40	Penta_E	D10S1248	1.0000	SAFMX40	TH01	D12S391	0.7428
SAFMX40	D2S441	D10S1248	0.2511	SAFMX40	FGA	D12S391	0.3900
SAFMX40	D19S433	D10S1248	0.7552	SAFMX40	D22S1045	D12S391	0.0124
SAFMX40	TH01	D10S1248	0.2449	SAFMX40	D5S818	D12S391	0.7883
SAFMX40	FGA	D10S1248	0.3369	SAFMX40	D13S317	D12S391	0.7558
SAFMX40	D22S1045	D10S1248	0.9825	SAFMX40	D7S820	D12S391	0.1811
SAFMX40	D5S818	D10S1248	0.5423	SAFMX40	D6S1043	D12S391	1.0000

* Loci in LD are indicated in bold

Table 7.5, continued

Population	Locus#1	Locus#2	*P*-Value	Population	Locus#1	Locus#2	*P*-Value
SAFMX40	D16S539	D1S1656	0.3887	SAFMX40	CSF1PO	D2S1338	0.8091
SAFMX40	CSF1PO	D1S1656	1.0000	SAFMX40	TPOX	D2S1338	1.0000
SAFMX40	TPOX	D1S1656	0.5810	SAFMX40	D8S1179	D2S1338	1.0000
SAFMX40	D8S1179	D1S1656	0.0919	SAFMX40	D21S11	D2S1338	1.0000
SAFMX40	D21S11	D1S1656	0.4531	SAFMX40	D18S51	D2S1338	1.0000
SAFMX40	D18S51	D1S1656	0.1146	SAFMX40	Penta_E	D2S1338	1.0000
SAFMX40	Penta_E	D1S1656	1.0000	SAFMX40	D2S441	D2S1338	0.7343
SAFMX40	D2S441	D1S1656	1.0000	SAFMX40	D19S433	D2S1338	0.4831
SAFMX40	D19S433	D1S1656	1.0000	SAFMX40	TH01	D2S1338	0.7325
SAFMX40	TH01	D1S1656	0.1921	SAFMX40	FGA	D2S1338	1.0000
SAFMX40	FGA	D1S1656	1.0000	SAFMX40	D22S1045	D2S1338	0.5383
SAFMX40	D22S1045	D1S1656	0.6053	SAFMX40	D5S818	D2S1338	0.7374
SAFMX40	D5S818	D1S1656	0.8091	SAFMX40	D13S317	D2S1338	1.0000
SAFMX40	D13S317	D1S1656	**0.0000**	SAFMX40	D7S820	D2S1338	0.5725
SAFMX40	D7S820	D1S1656	0.1951	SAFMX40	D6S1043	D2S1338	0.4567
SAFMX40	D6S1043	D1S1656	1.0000	SAFMX40	D10S1248	D2S1338	0.5087
SAFMX40	D10S1248	D1S1656	0.6945	SAFMX40	D1S1656	D2S1338	1.0000
SAFMX40	D3S13581	D12S391	0.1143	SAFMX40	D12S391	D2S1338	1.0000
SAFMX40	vWA	D12S391	1.0000	SAFMX40	D3S13581	Penta_D	0.8506
SAFMX40	D16S539	D12S391	0.8449	SAFMX40	vWA	Penta_D	0.7977
SAFMX40	CSF1PO	D12S391	0.3289	SAFMX40	D16S539	Penta_D	0.4527
SAFMX40	TPOX	D12S391	1.0000	SAFMX40	CSF1PO	Penta_D	0.2303
SAFMX40	D8S1179	D12S391	0.2900	SAFMX40	TPOX	Penta_D	0.5954
SAFMX40	D21S11	D12S391	1.0000	SAFMX40	D8S1179	Penta_D	0.0593
SAFMX40	D18S51	D12S391	0.2566	SAFMX40	D19S433	Penta_D	0.6904
SAFMX40	D21S11	Penta_D	1.0000	SAFMX40	TH01	Penta_D	0.1670
SAFMX40	Penta_E	D12S391	1.0000	SAFMX40	FGA	Penta_D	1.0000
SAFMX40	D18S51	Penta_D	0.4608	SAFMX40	D22S1045	Penta_D	0.8417
SAFMX40	Penta_E	Penta_D	1.0000	SAFMX40	D5S818	Penta_D	0.8649
SAFMX40	D2S441	Penta_D	0.5029	SAFMX40	D13S317	Penta_D	0.0471
SAFMX40	D7S820	Penta_D	0.5235	SAFMX40	D1S1656	Penta_D	0.5616
SAFMX40	D6S1043	Penta_D	1.0000	SAFMX40	D12S391	Penta_D	0.2164
SAFMX40	D10S1248	Penta_D	0.7154				

* Loci in LD are indicated in bold

Table 7.6 *P*-values for pairwise analysis of all the STR loci for the Zimbabwean Bantu-speakers

Population	Locus#1	Locus#2	*P*-Value	Population	Locus#1	Locus#2	*P*-Value
ZIMBS113	D21S11	Penta_E	0.8791	ZIMBS113	D18S51	FGA	0.9230
ZIMBS113	D18S51	Penta_E	0.9767	ZIMBS113	Penta_E	FGA	0.0130
ZIMBS113	D3S13581	D2S441	0.2211	ZIMBS113	D2S441	FGA	0.9749
ZIMBS113	vWA	D2S441	0.9869	ZIMBS113	D19S433	FGA	0.5191
ZIMBS113	D16S539	D2S441	0.4168	ZIMBS113	TH01	FGA	0.3646
ZIMBS113	vWA	D22S1045	0.0836	ZIMBS113	D5S818	D13S317	0.8866
ZIMBS113	D16S539	D22S1045	0.8268	ZIMBS113	D3S13581	D7S820	0.7034
ZIMBS113	CSF1PO	D22S1045	0.7897	ZIMBS113	vWA	D7S820	0.2821
ZIMBS113	TPOX	D22S1045	0.8126	ZIMBS113	D16S539	D7S820	0.9886
ZIMBS113	D8S1179	D22S1045	0.6191	ZIMBS113	CSF1PO	D7S820	0.5225
ZIMBS113	D21S11	D22S1045	0.3382	ZIMBS113	TPOX	D7S820	0.6649
ZIMBS113	D18S51	D22S1045	0.8382	ZIMBS113	D8S1179	D7S820	0.8307
ZIMBS113	Penta_E	D22S1045	0.6011	ZIMBS113	D21S11	D7S820	0.5006
ZIMBS113	D2S441	D22S1045	0.6843	ZIMBS113	D18S51	D7S820	0.1039
ZIMBS113	D19S433	D22S1045	0.6465	ZIMBS113	Penta_E	D7S820	0.5342
ZIMBS113	TH01	D22S1045	0.9526	ZIMBS113	D2S441	D7S820	0.2814
ZIMBS113	FGA	D22S1045	0.0718	ZIMBS113	D19S433	D7S820	0.9548
ZIMBS113	D3S13581	D5S818	0.4750	ZIMBS113	TH01	D7S820	0.5733
ZIMBS113	vWA	D5S818	0.9044	ZIMBS113	FGA	D7S820	0.1789
ZIMBS113	D16S539	D5S818	0.3513	ZIMBS113	D22S1045	D7S820	0.8245
ZIMBS113	CSF1PO	D5S818	0.6273	ZIMBS113	D5S818	D7S820	0.7230
ZIMBS113	TPOX	D5S818	0.1738	ZIMBS113	D13S317	D7S820	0.3273
ZIMBS113	D8S1179	D5S818	0.9403	ZIMBS113	D3S13581	D6S1043	0.1327
ZIMBS113	D21S11	D5S818	0.5528	ZIMBS113	vWA	D6S1043	0.7854
ZIMBS113	D18S51	D5S818	0.7540	ZIMBS113	D16S539	D6S1043	0.9102
ZIMBS113	Penta_E	D5S818	0.0467	ZIMBS113	CSF1PO	D6S1043	1.0000
ZIMBS113	D2S441	D5S818	0.4755	ZIMBS113	TPOX	D6S1043	0.0611
ZIMBS113	D19S433	D5S818	0.8716	ZIMBS113	D8S1179	D6S1043	0.2317
ZIMBS113	TH01	D5S818	0.4028	ZIMBS113	D21S11	D6S1043	0.7512
ZIMBS113	FGA	D5S818	0.3717	ZIMBS113	D18S51	D6S1043	0.9662
ZIMBS113	D22S1045	D5S818	0.9842	ZIMBS113	Penta_E	D6S1043	1.0000
ZIMBS113	D3S13581	D13S317	0.7549	ZIMBS113	D2S441	D6S1043	0.9084

* Loci in LD are indicated in bold

Table 7.6, continued

Population	Locus#1	Locus#2	*P*-Value	Population	Locus#1	Locus#2	*P*-Value
ZIMBS113	vWA	D13S317	0.3116	ZIMBS113	D19S433	D6S1043	0.1833
ZIMBS113	D16S539	D13S317	0.0946	ZIMBS113	TH01	D6S1043	0.6081
ZIMBS113	CSF1PO	D13S317	0.9497	ZIMBS113	FGA	D6S1043	1.0000
ZIMBS113	TPOX	D13S317	0.9989	ZIMBS113	D22S1045	D6S1043	0.5845
ZIMBS113	D8S1179	D13S317	0.2884	ZIMBS113	D5S818	D6S1043	0.0999
ZIMBS113	D21S11	D13S317	0.4499	ZIMBS113	D13S317	D6S1043	0.1732
ZIMBS113	D18S51	D13S317	0.7679	ZIMBS113	D7S820	D6S1043	0.4670
ZIMBS113	Penta_E	D13S317	0.3110	ZIMBS113	D3S13581	D10S1248	0.7303
ZIMBS113	D2S441	D13S317	0.5437	ZIMBS113	vWA	D10S1248	0.2323
ZIMBS113	D19S433	D13S317	0.1841	ZIMBS113	D16S539	D10S1248	0.3192
ZIMBS113	TH01	D13S317	0.8597	ZIMBS113	CSF1PO	D10S1248	0.0603
ZIMBS113	FGA	D13S317	0.0968	ZIMBS113	TPOX	D10S1248	0.8336
ZIMBS113	D22S1045	D13S317	0.6936	ZIMBS113	D8S1179	D10S1248	0.3545
ZIMBS113	D21S11	D10S1248	0.7469	ZIMBS113	D2S441	D12S391	0.8489
ZIMBS113	D18S51	D10S1248	0.9976	ZIMBS113	D19S433	D12S391	0.9410
ZIMBS113	Penta_E	D10S1248	0.0935	ZIMBS113	TH01	D12S391	0.5675
ZIMBS113	D2S441	D10S1248	0.8371	ZIMBS113	FGA	D12S391	0.5348
ZIMBS113	D19S433	D10S1248	1.0000	ZIMBS113	D22S1045	D12S391	0.0663
ZIMBS113	TH01	D10S1248	0.3089	ZIMBS113	D5S818	D12S391	0.1404
ZIMBS113	FGA	D10S1248	0.3618	ZIMBS113	D13S317	D12S391	0.1609
ZIMBS113	D22S1045	D10S1248	0.8532	ZIMBS113	D7S820	D12S391	0.5785
ZIMBS113	D5S818	D10S1248	**0.0058**	ZIMBS113	D6S1043	D12S391	0.8374
ZIMBS113	D13S317	D10S1248	0.3485	ZIMBS113	D10S1248	D12S391	0.9227
ZIMBS113	D7S820	D10S1248	0.1394	ZIMBS113	D1S1656	D12S391	0.6171
ZIMBS113	D6S1043	D10S1248	0.8914	ZIMBS113	D3S13581	D2S1338	0.8906
ZIMBS113	D3S13581	D1S1656	0.2617	ZIMBS113	vWA	D2S1338	0.2374
ZIMBS113	vWA	D1S1656	0.2403	ZIMBS113	D16S539	D2S1338	0.4015
ZIMBS113	D16S539	D1S1656	0.9567	ZIMBS113	CSF1PO	D2S1338	0.8500
ZIMBS113	CSF1PO	D1S1656	0.8696	ZIMBS113	TPOX	D2S1338	0.5962
ZIMBS113	TPOX	D1S1656	0.7778	ZIMBS113	D8S1179	D2S1338	0.6763
ZIMBS113	D8S1179	D1S1656	0.6931	ZIMBS113	D21S11	D2S1338	0.6564
ZIMBS113	D21S11	D1S1656	0.7702	ZIMBS113	D18S51	D2S1338	0.7643

* Loci in LD are indicated in bold

Table 7.6, continued

Population	Locus#1	Locus#2	*P*-Value	Population	Locus#1	Locus#2	*P*-Value
ZIMBS113	D18S51	D1S1656	**0.0206**	ZIMBS113	Penta_E	D2S1338	0.7260
ZIMBS113	Penta_E	D1S1656	0.9613	ZIMBS113	D2S441	D2S1338	0.8866
ZIMBS113	D2S441	D1S1656	0.2839	ZIMBS113	D19S433	D2S1338	1.0000
ZIMBS113	D19S433	D1S1656	0.0966	ZIMBS113	TH01	D2S1338	0.9995
ZIMBS113	TH01	D1S1656	0.3970	ZIMBS113	FGA	D2S1338	0.4186
ZIMBS113	FGA	D1S1656	0.7040	ZIMBS113	D22S1045	D2S1338	0.6999
ZIMBS113	D22S1045	D1S1656	0.3210	ZIMBS113	D5S818	D2S1338	**0.0002**
ZIMBS113	D5S818	D1S1656	0.5785	ZIMBS113	D13S317	D2S1338	0.2178
ZIMBS113	D13S317	D1S1656	0.9276	ZIMBS113	D7S820	D2S1338	0.9762
ZIMBS113	D7S820	D1S1656	0.3013	ZIMBS113	D6S1043	D2S1338	0.3679
ZIMBS113	D6S1043	D1S1656	0.6976	ZIMBS113	D10S1248	D2S1338	0.9422
ZIMBS113	D10S1248	D1S1656	0.9924	ZIMBS113	D1S1656	D2S1338	0.8782
ZIMBS113	D3S13581	D12S391	0.6197	ZIMBS113	D12S391	D2S1338	0.6445
ZIMBS113	vWA	D12S391	0.2638	ZIMBS113	D3S13581	Penta_D	0.8664
ZIMBS113	D16S539	D12S391	0.7908	ZIMBS113	vWA	Penta_D	0.5588
ZIMBS113	CSF1PO	D12S391	0.4693	ZIMBS113	D16S539	Penta_D	0.7239
ZIMBS113	TPOX	D12S391	0.0899	ZIMBS113	CSF1PO	Penta_D	0.9663
ZIMBS113	D8S1179	D12S391	0.2856	ZIMBS113	TPOX	Penta_D	0.6332
ZIMBS113	D21S11	D12S391	0.1118	ZIMBS113	FGA	Penta_D	0.7400
ZIMBS113	D18S51	D12S391	0.5800	ZIMBS113	D22S1045	Penta_D	0.7423
ZIMBS113	Penta_E	D12S391	0.1358	ZIMBS113	D5S818	Penta_D	**0.0395**
ZIMBS113	D8S1179	Penta_D	0.6480	ZIMBS113	D13S317	Penta_D	0.7835
ZIMBS113	D21S11	Penta_D	**0.0000**	ZIMBS113	D7S820	Penta_D	0.3214
ZIMBS113	D18S51	Penta_D	0.4369	ZIMBS113	D6S1043	Penta_D	0.9025
ZIMBS113	Penta_E	Penta_D	0.0875	ZIMBS113	D10S1248	Penta_D	0.3717
ZIMBS113	D2S441	Penta_D	0.1816	ZIMBS113	D1S1656	Penta_D	0.4831
ZIMBS113	D19S433	Penta_D	0.2608	ZIMBS113	D12S391	Penta_D	0.1563
ZIMBS113	TH01	Penta_D	0.8875	ZIMBS113	D2S1338	Penta_D	0.3031

* Loci in LD are indicated in bold

Table 7.7 *P*-values for pairwise analysis of all the STR loci for the Nigerian Bantu-speakers

Population	Locus#1	Locus#2	*P*-Value	Population	Locus#1	Locus#2	*P*-Value
NGRBS73	D3S13581	vWA	0.3646	NGRBS73	TPOX	D2S441	0.0610
NGRBS73	D3S13581	D16S539	**0.0005**	NGRBS73	D8S1179	D2S441	0.4708
NGRBS73	vWA	D16S539	0.6290	NGRBS73	D21S11	D2S441	0.1689
NGRBS73	D3S13581	CSF1PO	0.4935	NGRBS73	D18S51	D2S441	0.3645
NGRBS73	vWA	CSF1PO	0.8058	NGRBS73	Penta_E	D2S441	**0.0080**
NGRBS73	D16S539	CSF1PO	0.0471	NGRBS73	D3S13581	D19S433	0.4797
NGRBS73	D3S13581	TPOX	0.6331	NGRBS73	vWA	D19S433	**0.0373**
NGRBS73	vWA	TPOX	**0.0157**	NGRBS73	D16S539	D19S433	0.3694
NGRBS73	D16S539	TPOX	**0.0183**	NGRBS73	CSF1PO	D19S433	0.4842
NGRBS73	CSF1PO	TPOX	**0.0219**	NGRBS73	TPOX	D19S433	0.3232
NGRBS73	D3S13581	D8S1179	0.9544	NGRBS73	D8S1179	D19S433	0.7210
NGRBS73	vWA	D8S1179	0.1099	NGRBS73	D21S11	D19S433	0.3720
NGRBS73	D16S539	D8S1179	0.2260	NGRBS73	D18S51	D19S433	0.4466
NGRBS73	CSF1PO	D8S1179	0.2076	NGRBS73	Penta_E	D19S433	0.1809
NGRBS73	TPOX	D8S1179	0.8925	NGRBS73	D2S441	D19S433	**0.0313**
NGRBS73	D3S13581	D21S11	0.1918	NGRBS73	D3S13581	TH01	0.1120
NGRBS73	vWA	D21S11	**0.0269**	NGRBS73	vWA	TH01	0.5168
NGRBS73	D16S539	D21S11	0.3241	NGRBS73	D16S539	TH01	0.5930
NGRBS73	CSF1PO	D21S11	0.0523	NGRBS73	CSF1PO	TH01	0.1674
NGRBS73	TPOX	D21S11	0.4042	NGRBS73	TPOX	TH01	0.7270
NGRBS73	D8S1179	D21S11	0.2206	NGRBS73	D8S1179	TH01	0.1202
NGRBS73	D3S13581	D18S51	0.3971	NGRBS73	D21S11	TH01	0.6417
NGRBS73	vWA	D18S51	0.6689	NGRBS73	D18S51	TH01	0.5117
NGRBS73	D16S539	D18S51	0.0551	NGRBS73	Penta_E	TH01	0.2326
NGRBS73	CSF1PO	D18S51	0.0602	NGRBS73	D2S441	TH01	0.1968
NGRBS73	TPOX	D18S51	0.1390	NGRBS73	D19S433	TH01	0.1113
NGRBS73	D8S1179	D18S51	0.1244	NGRBS73	D3S13581	FGA	0.7989
NGRBS73	D21S11	D18S51	0.8513	NGRBS73	vWA	FGA	0.4924
NGRBS73	D3S13581	Penta_E	**0.0151**	NGRBS73	D16S539	FGA	0.3153
NGRBS73	vWA	Penta_E	0.6351	NGRBS73	CSF1PO	FGA	0.3044
NGRBS73	D16S539	Penta_E	0.0897	NGRBS73	TPOX	FGA	0.7142
NGRBS73	CSF1PO	Penta_E	0.9265	NGRBS73	D8S1179	FGA	0.2530

* Loci in LD are indicated in bold

Table 7.7, continued

Population	Locus#1	Locus#2	*P*-Value	Population	Locus#1	Locus#2	*P*-Value
NGRBS73	TPOX	Penta_E	0.1487	NGRBS73	D21S11	FGA	0.0648
NGRBS73	D8S1179	Penta_E	0.1452	NGRBS73	D18S51	FGA	0.2541
NGRBS73	D21S11	Penta_E	0.3089	NGRBS73	Penta_E	FGA	0.1661
NGRBS73	D18S51	Penta_E	0.0260	NGRBS73	D2S441	FGA	0.4054
NGRBS73	D3S13581	D2S441	0.0547	NGRBS73	D19S433	FGA	0.0818
NGRBS73	vWA	D2S441	0.3926	NGRBS73	TH01	FGA	0.0481
NGRBS73	D16S539	D2S441	0.3208	NGRBS73	D3S13581	D22S1045	0.3000
NGRBS73	vWA	D22S1045	0.0734	NGRBS73	D5S818	D13S317	0.5405
NGRBS73	D16S539	D22S1045	**0.0115**	NGRBS73	D3S13581	D7S820	0.6794
NGRBS73	CSF1PO	D22S1045	0.2795	NGRBS73	vWA	D7S820	0.6441
NGRBS73	TPOX	D22S1045	0.7363	NGRBS73	D16S539	D7S820	0.5975
NGRBS73	D8S1179	D22S1045	0.1673	NGRBS73	CSF1PO	D7S820	0.4406
NGRBS73	D21S11	D22S1045	0.5328	NGRBS73	TPOX	D7S820	0.1337
NGRBS73	D18S51	D22S1045	0.0941	NGRBS73	D8S1179	D7S820	0.2249
NGRBS73	Penta_E	D22S1045	0.1088	NGRBS73	D21S11	D7S820	0.3021
NGRBS73	D2S441	D22S1045	0.5583	NGRBS73	D18S51	D7S820	0.1479
NGRBS73	D19S433	D22S1045	0.7434	NGRBS73	Penta_E	D7S820	0.5982
NGRBS73	TH01	D22S1045	0.5961	NGRBS73	D2S441	D7S820	0.8076
NGRBS73	FGA	D22S1045	**0.0096**	NGRBS73	D19S433	D7S820	0.0998
NGRBS73	D3S13581	D5S818	0.8690	NGRBS73	TH01	D7S820	0.0980
NGRBS73	vWA	D5S818	**0.0162**	NGRBS73	FGA	D7S820	0.1963
NGRBS73	D16S539	D5S818	**0.0133**	NGRBS73	D22S1045	D7S820	**0.0000**
NGRBS73	CSF1PO	D5S818	0.4215	NGRBS73	D5S818	D7S820	0.5817
NGRBS73	TPOX	D5S818	**0.0442**	NGRBS73	D13S317	D7S820	0.4359
NGRBS73	D8S1179	D5S818	0.1165	NGRBS73	D3S13581	D6S1043	0.4274
NGRBS73	D21S11	D5S818	**0.0381**	NGRBS73	vWA	D6S1043	**0.0373**
NGRBS73	D18S51	D5S818	**0.0402**	NGRBS73	D16S539	D6S1043	0.3180
NGRBS73	Penta_E	D5S818	0.8016	NGRBS73	CSF1PO	D6S1043	0.7191
NGRBS73	D2S441	D5S818	0.4948	NGRBS73	TPOX	D6S1043	0.5240
NGRBS73	D19S433	D5S818	0.5540	NGRBS73	D8S1179	D6S1043	0.1452
NGRBS73	TH01	D5S818	0.3423	NGRBS73	D21S11	D6S1043	0.9314
NGRBS73	FGA	D5S818	0.4301	NGRBS73	D18S51	D6S1043	0.7616

* Loci in LD are indicated in bold

Table 7.7, continued

Population	Locus#1	Locus#2	*P*-Value	Population	Locus#1	Locus#2	*P*-Value
NGRBS73	D22S1045	D5S818	0.2360	NGRBS73	Penta_E	D6S1043	0.5106
NGRBS73	D3S13581	D13S317	0.4885	NGRBS73	D2S441	D6S1043	0.1471
NGRBS73	vWA	D13S317	0.3379	NGRBS73	D19S433	D6S1043	0.4496
NGRBS73	D16S539	D13S317	0.2309	NGRBS73	TH01	D6S1043	0.9699
NGRBS73	CSF1PO	D13S317	0.0482	NGRBS73	FGA	D6S1043	**0.0360**
NGRBS73	TPOX	D13S317	**0.0338**	NGRBS73	D22S1045	D6S1043	0.8762
NGRBS73	D8S1179	D13S317	0.5387	NGRBS73	D5S818	D6S1043	0.3122
NGRBS73	D21S11	D13S317	0.4707	NGRBS73	D13S317	D6S1043	0.0513
NGRBS73	D18S51	D13S317	0.6277	NGRBS73	D7S820	D6S1043	0.3736
NGRBS73	Penta_E	D13S317	0.5743	NGRBS73	D3S13581	D10S1248	0.3654
NGRBS73	D2S441	D13S317	0.7378	NGRBS73	vWA	D10S1248	0.2737
NGRBS73	D19S433	D13S317	0.8589	NGRBS73	D16S539	D10S1248	0.8014
NGRBS73	TH01	D13S317	0.6550	NGRBS73	CSF1PO	D10S1248	**0.0056**
NGRBS73	FGA	D13S317	0.2506	NGRBS73	TPOX	D10S1248	0.3213
NGRBS73	D22S1045	D13S317	0.0600	NGRBS73	D8S1179	D10S1248	0.6927
NGRBS73	D21S11	D10S1248	**0.0120**	NGRBS73	D2S441	D12S391	0.5887
NGRBS73	D18S51	D10S1248	0.2462	NGRBS73	D19S433	D12S391	**0.0000**
NGRBS73	Penta_E	D10S1248	0.7871	NGRBS73	TH01	D12S391	0.0957
NGRBS73	D2S441	D10S1248	0.2694	NGRBS73	FGA	D12S391	0.2014
NGRBS73	D19S433	D10S1248	0.0825	NGRBS73	D22S1045	D12S391	**0.0033**
NGRBS73	TH01	D10S1248	0.2543	NGRBS73	D5S818	D12S391	0.1691
NGRBS73	FGA	D10S1248	**0.0204**	NGRBS73	D13S317	D12S391	0.0806
NGRBS73	D22S1045	D10S1248	0.8332	NGRBS73	D7S820	D12S391	0.6534
NGRBS73	D5S818	D10S1248	0.0926	NGRBS73	D6S1043	D12S391	0.4159
NGRBS73	D13S317	D10S1248	0.5288	NGRBS73	D10S1248	D12S391	0.5001
NGRBS73	D7S820	D10S1248	0.5433	NGRBS73	D1S1656	D12S391	0.0701
NGRBS73	D6S1043	D10S1248	**0.0201**	NGRBS73	D3S13581	D2S1338	0.7266
NGRBS73	D3S13581	D1S1656	0.1669	NGRBS73	vWA	D2S1338	0.5487
NGRBS73	vWA	D1S1656	0.8305	NGRBS73	D16S539	D2S1338	**0.0324**
NGRBS73	D16S539	D1S1656	0.0468	NGRBS73	CSF1PO	D2S1338	0.6817
NGRBS73	CSF1PO	D1S1656	0.0579	NGRBS73	TPOX	D2S1338	0.2363
NGRBS73	TPOX	D1S1656	**0.0335**	NGRBS73	D8S1179	D2S1338	0.5614

* Loci in LD are indicated in bold

Table 7.7, continued

Population	Locus#1	Locus#2	*P*-Value	Population	Locus#1	Locus#2	*P*-Value
NGRBS73	D8S1179	D1S1656	**0.0000**	NGRBS73	D21S11	D2S1338	**0.0096**
NGRBS73	D21S11	D1S1656	0.0798	NGRBS73	D18S51	D2S1338	0.3657
NGRBS73	D18S51	D1S1656	0.1467	NGRBS73	Penta_E	D2S1338	0.0459
NGRBS73	Penta_E	D1S1656	0.0658	NGRBS73	D2S441	D2S1338	0.1581
NGRBS73	D2S441	D1S1656	1.0000	NGRBS73	D19S433	D2S1338	**0.0270**
NGRBS73	D19S433	D1S1656	0.1615	NGRBS73	TH01	D2S1338	0.0673
NGRBS73	TH01	D1S1656	0.7562	NGRBS73	FGA	D2S1338	0.1620
NGRBS73	FGA	D1S1656	0.0141	NGRBS73	D22S1045	D2S1338	0.1277
NGRBS73	D22S1045	D1S1656	0.0691	NGRBS73	D5S818	D2S1338	0.0479
NGRBS73	D5S818	D1S1656	0.2031	NGRBS73	D13S317	D2S1338	0.8928
NGRBS73	D13S317	D1S1656	0.5631	NGRBS73	D7S820	D2S1338	0.6054
NGRBS73	D7S820	D1S1656	0.2063	NGRBS73	D6S1043	D2S1338	**0.0225**
NGRBS73	D6S1043	D1S1656	0.4277	NGRBS73	D10S1248	D2S1338	0.2943
NGRBS73	D10S1248	D1S1656	0.3450	NGRBS73	D1S1656	D2S1338	0.6776
NGRBS73	D3S13581	D12S391	0.0358	NGRBS73	D12S391	D2S1338	0.1151
NGRBS73	vWA	D12S391	0.5712	NGRBS73	D3S13581	Penta_D	0.4617
NGRBS73	D16S539	D12S391	0.2051	NGRBS73	vWA	Penta_D	0.5550
NGRBS73	CSF1PO	D12S391	0.6160	NGRBS73	D16S539	Penta_D	0.0597
NGRBS73	TPOX	D12S391	0.2849	NGRBS73	CSF1PO	Penta_D	0.0952
NGRBS73	D8S1179	D12S391	0.7898	NGRBS73	TPOX	Penta_D	0.7114
NGRBS73	D21S11	D12S391	0.8195	NGRBS73	D8S1179	Penta_D	0.1966
NGRBS73	D18S51	D12S391	0.6241	NGRBS73	D5S818	Penta_D	0.6969
NGRBS73	D21S11	Penta_D	**0.0148**	NGRBS73	D13S317	Penta_D	0.8452
NGRBS73	Penta_E	D12S391	0.7353	NGRBS73	D7S820	Penta_D	0.3568
NGRBS73	D18S51	Penta_D	**0.0146**	NGRBS73	D6S1043	Penta_D	0.3822
NGRBS73	Penta_E	Penta_D	0.0155	NGRBS73	D10S1248	Penta_D	0.2011
NGRBS73	D2S441	Penta_D	0.1063	NGRBS73	D1S1656	Penta_D	0.0671
NGRBS73	D19S433	Penta_D	0.5783	NGRBS73	D12S391	Penta_D	**0.0225**
NGRBS73	TH01	Penta_D	0.2667	NGRBS73	D2S1338	Penta_D	0.6628
NGRBS73	FGA	Penta_D	0.0929	NGRBS73	CSF1PO	D2S441	0.2541
NGRBS73	D22S1045	Penta_D	0.7317				

* Loci in LD are indicated in bold

7.8 APPENDIX F: Validation report for comparing different data analysis software packages used in kinship analysis

Title: Validation of the In-house data analysis software for calculation of Likelihood Ratios in kinship testing

Written by: Mahlatse Moremi

Introduction

In order to improve the process of resolving complex kinship outcomes observed in the Division of Human Genetics, National Health Laboratory Service (NHLS) and School of Pathology, Faculty of Health Sciences, The University of Witwatersrand, DNA Profiling Laboratory (DPL), Braamfontein, Johannesburg, South Africa, we decided to validate the In-house data analysis software used in the present laboratory for calculation of Likelihood Ratios. A total of 63 scenarios (family relationships) was assessed. Of the 63 scenarios, 17 relationships were reported as not excluded, 43 as inconclusive and three as excluded when the samples were processed and analysed using commercial kits with 15 STR loci.

Three data analysis software packages, namely, *Familias*, *GeneMarker HID* and *Converge*™ Software were used in the validation process. The Scientific Working Group on DNA Analysis Methods and the DNA Advisory Board of Quality Assurance Standards (DBA) provide guidelines for validation of software used in human identification Likelihood Ratios calculations (National Institute of Standards, US Department of commerce, 2009; SWGDAM, 2012). Validation is done to ensure that the same accurate

result is obtained each time the calculation is performed (Drábek, 2009). Calculation of Likelihood Ratios or paternity indices can be done by hand, however, this approach can be very complex and prone to human error, especially in complex kinship queries. In paternity and kinship testing, the impact of using the correct software for calculating Likelihood Ratios has huge legal implications, as incorrect calculations can result in people classified as unrelated, and not eligible for estate benefits or approval of applications for maintenance or citizenship. A negative outcome may lead to a divorce in false non paternity scenarios. For these reasons, all laboratories offering this service should validate or use validated software that meets good quality and assurance standards. Some level of concordance between outcomes obtained when using different software packages is expected for calculation of Likelihood Ratios (Azevedo et al., 2011). It is worrying to know that some commercial and freely available software packages for calculating Likelihood Ratios were not sufficiently validated to meet the required standards for competence testing (Drábek, 2009). Algebraic formulae that are based on the Mendelian inheritance and probability of transmission have been published by the American Association of Blood Banks (AABB) (see Table 7.8 and 7.9) (Azevedo et al., 2011).

Table 7.8 Formulae for calculating Likelihood Ratios for a paternity trio.

Mother	Child	Alleged father	Formula	Mother	Child	Alleged father	Formula
A,A	A,A	A,B	$1/2a^2$	A,A	A,A	A,A	$1/a^2$
A,A	A,B	A,B	1/4ab	A,B	A,A	A,A	$1/2a^2$
A,A	A,B	B,C	1/4ab	B,B	A,B	A,A	1/2ab
A,B	A,A	A,B	$1/4a^2$	B,C	A,B	A,A	1/4ab
A,B	A,A	A,C	$1/4a^2$	A,B	A,B	A,C	1/8ab
B,C	A,B	A,B	1/8ab	A,B	A,B	A,A	1/4ab
B,C	A,B	A,C	1/8ab	A,B	A,B	A,B	1/4ab
B,D	A,B	A,C	1/8ab				

Mother, child and alleged father are tested. The formulae are based on the inheritance pattern. A, B, C and D depict alleles obtained at a particular locus. a and b represent the allele frequencies of alleles A and B. Profiles are denoted by genotypes. Adapted from (Sozer et al., 2010)

Table 7.9 Formulae for calculating Likelihood Ratios for kinships (sibships)

Sibling	Possible full sibling	Formula	Half sibling	Possible half sibling	Formula
A,B	A,B	(1+a+b+2ab)/8ab	A,B	A,B	(a+b+4a)/8ab
A,A	A,A	$(1+a)^2/(2a^2)$	A,A	A,A	(1+a)/2a
A,A	A,B	(1+a)/4a	A,A	A,B	(1+2a)/4a
A,B	A,C	(1+2a)/8a	A,B	A,C	(1+4a)/8a
A,B	C,D	¼	A,B	C,D	½

Formulae for calculating Likelihood Ratios when siblings, full and half, are tested. The formulae are based on the inheritance pattern. A, B, C and D depict alleles obtained at a particular locus. a and b represent the allele frequencies of alleles A and B. Profiles are denoted by genotypes. Adapted from (Sozer et al., 2010)

Aim and objective

The aim of this assessment was to validate the In-house data analysis software package currently used in DPL.

The objective was to compare Combined Likelihood Ratios (CLRs) obtained when different data analysis software packages were used, under the premise that the outcomes obtained should not be significantly different. This assessment would then be used to validate the In-house data analysis software, particularly for kinship testing

Responsible staff members

Validation protocol: Mahlatse Moremi

Sample processing: Mahlatse Moremi

Results analysis: Mahlatse Moremi

Validation report: Mahlatse Moremi

Location

Division of Human Genetics, National Health Laboratory Service (NHLS) and School of Pathology, Faculty of Health Sciences, The University of Witwatersrand, DNA Profiling Laboratory (DPL), Braamfontein, Johannesburg, South Africa

Protocol

Genotypes already generated using the VeriFiler™Express PCR amplification kit were used. The relationship scenarios indicated in Table 7.8 were assessed in the validation process. The following data analysis software packages were included in the study, to compare Likelihood Ratio (LR) outcomes: In-house Microsoft Excel Macro enabled software (Lane, unpublished), *Converge*™ Software, *Familias* and *GeneMarker HID* The capabilities of each data analysis software package are indicated in Table 7.11.

Table 7.10 Scenarios of relationships used in the validation of the In-house data analysis software

Relationship (Scenario)	**Number of cases (n)**
Siblings	29
Child and alleged grand(mother/father)	8
Child and alleged full aunt/uncle	8
Child and alleged half aunt/uncle	5
Mother, child and alleged sibling	1
Mother, child and alleged grandparent	3
Child 1, Child 2 and alleged grandparent	2
Child 1, Child 2 and alleged aunt/uncle	3
Mother, child 1, child 2 and alleged grandmother	1
Mother, child and alleged half aunt	1
Mother, child 1, child 2 and alleged grandmother	1
First cousins	1

Table 7.11 Capabilities of the different data analysis software for calculating

Software capability	**In-house software**	**Converge ™ Software V2.1**	**Familias v3.2.2**	**GeneMarker® HID v3.0.0**
Trio calculations	√	√	√	√
Duo calculations	√	√	√	√
Motherless	√	√	√	√
Grandparents/Half siblings/uncle-nephew (Duos)	√	√	√	√
Parent, child and relative	√	√	√	Pair wise only
Kinship- More than 3 family members together	Pair wise only	√	√	Pair wise only
Mutation calculation Incorporated	√	√	√	√
Null alleles calculation Incorporated	Separately	√	√	√
Simulations	Not capable	Not capable	√	√
Updates	Not documented	√	√	√
User manual	Not available	√	√	√

Likelihood Ratios for kinship testing

Data analysis

The analysis of data obtained when different data analysis software packages were used was done using Microsoft Excel 2016, Data analysis tool, t-Test: Paired Two Sample for Means. The null hypothesis ($H_{o)}$ was that there was no significant difference in the CLR outcomes obtained when the different data analysis software packages were used. The alternative hypothesis (H_1) was that there was a significant difference in the CLR outcomes obtained when different analysis software packages were used. A significance level of 0.05 was used.

Acceptance criteria:

If *p*-value >0.05, accept the null hypothesis

If *p*-value<0.05, accept the alternative hypothesis (i.e. reject the null hypothesis)

The null hypothesis must be true to continue using the In-house data analysis software to calculate Likelihood Ratios for kinship testing.

Results

The *p*-values shown in Table 7.12 were obtained when the Student's T-test was used to compare the CLR outcomes from different data analysis software packages. All the *p*-values were above 0.05. This finding supports the null hypothesis.

Table 7.12 Paired two sample T-test conducted to compare outcomes obtained using four different data analysis software packages

	P-values (significance level of 0.05)		
	In-house software	*Familias*	*GeneMarker® HID*
Familias	0.331	-	-
GeneMarker® HID	0.372	0.386	-
Converge™ Software	0.366	0.347	0.389

Conclusion

The null hypothesis ($H_{0)}$ was that there was no significant difference in the CLR outcomes obtained when the different data analysis software packages were used. This was also the acceptance criteria for the test. The calculated *p*-values when comparing the In-house data analysis software to other published software packages were greater than 0.05. Therefore, the In-house data analysis software was deemed fit for use in routine diagnostic testing.

References

Azevedo, D.A., Souza, G.R.B., Silva, I.H.E.F., et al., 2011. Genetic kinship analysis: A concordance study between calculations performed with the software Familias and algebraic formulas of the American Association of Blood Banks. *Forensic Science International: Genetics Supplement Series* 3, e186–e187.

Converge™ Software version 2.1, 2018, computer software, Thermo Fisher Scientific Inc.

Drábek, J., 2009. Validation of software for calculating the Likelihood Ratio for parentage and kinship. *Forensic Sci Int Genet* 3, 112–118.

GeneMarker® *HID* version 3.0.0 (35-day trial), 2019, computer software, SoftGenetics LLC.

Kling, D, Tillmar, A.O., Egeland, T., 2014. Familias 3- Extensions and new functionality. *Forensic Science International: Genetics* 121–127.

National Institute of Standards, US Department of commerce, 2009. DNA Advisory Board Quality Assurance Standards for forensic DNA testing laboratories (*No. Forensic Science Communication* 2 (2000)).