Investigation into implementing a massively parallel sequencing workflow for forensic human identification in South Africa

---

Donna-Lee Pamela Martin

Thesis Presented for the Degree of

DOCTOR OF PHILOSOPHY

in Forensic Genetics

In the Division of Forensic Medicine and Toxicology, Department of Pathology, Faculty of Health Sciences

UNIVERSITY OF CAPE TOWN

30 August 2024

Supervisor: Associate Professor Laura Heathfield

Declaration

I, *Donna-Lee Pamela Martin*, hereby declare that the work on which this thesis is based is my original work (except where acknowledgements indicate otherwise) and that neither the whole work nor any part of it has been, is being, or is to be submitted for another degree in this or any other university. I authorise the University to reproduce for the purpose of research either the whole or any portion of the contents in any manner whatsoever.

This thesis has been submitted to the Turnitin module (or equivalent similarity or originality checking software, and I confirm that my supervisor has seen my report, and any concerns revealed by such have been resolved with my supervisors in any manner whatsoever.

Signature: Signed by candidate

Date: 30 August 2024

Table of Contents

Abstract

South Africa faces grave challenges with high crime rates and associated unidentified bodies each year. DNA profiling using capillary electrophoresis (CE) is typically utilised for human identification purposes but is limiting when applied to degraded post-mortem samples. The ForenSeq™ DNA Signature Prep kit was the first massively parallel sequencing (MPS) workflow validated on the MiSeq FGx™ system, addressing several challenges identified in CE-based methods. With forensic laboratories in developing regions showing proclivity towards a seemingly impossible adoption of MPS, sequence-based studies in Africa are sorely needed to leverage emerging advancements for forensic human identification.

This study proposed a four-phased approach for laboratories to *facilitate* the implementation of MPS for forensic human identification, and included: optimisation, population data generation, internal validation and demonstration. An optimisation study was carried out to ensure high first-time success rates of analysing reference samples (crude buccal swab lysates) with the ForenSeq™ DNA Signature Prep kit. This entailed systematic adjustments to a direct PCR approach and the development of a lysate purification method. This optimised approach was subsequently used to conduct a population study comprising 463 consenting South African volunteers, wherein the first sequence-based allele frequency data pertaining to autosomal short tandem repeat (A-STR) markers were generated for South African populations. Rich variation was observed, where 80 novel allele sequences were recorded. An increase of 86% was observed in length- to sequence-based allele counts across several A-STR markers, with additional variation recorded in flanking regions. Furthermore, a concordance rate exceeding 99% was achieved. The novel findings and abundance of variation observed in the South African population surpasses that which has been previously characterised on a global scale,

warranting further research into characterising sequence data for other forensically relevant markers.

The final facet of this study involved the internal validation of the optimised MPS workflow, from sample preparation to sequencing. The workflow was deemed fit for purpose and reported the first performance parameters for post-mortem crude buccal swab lysates. The validated workflow was then applied to a forensic cold case to generate investigative leads from a severely decomposed body, demonstrating the comprehensive capability of the workflow. The synthesis of results obtained in this study have led to key recommendations for under-resourced laboratories to maximise resources for large-scale studies.

This thesis is dedicated to my nephews, Micah, Cyrus, Malakhai and Jacob.

You have been a constant source of joy and motivation.

I hope this inspires you to always follow your dreams, no matter how big they are.

Acknowledgements

In reaching the culmination of my PhD journey, I am reminded that no great achievement is accomplished alone. This work has come to completion through the collective support, guidance and encouragement from many amazing people. I want to praise and thank God for instilling this passion in me and making a way for me to bring it to fruition.

I want to express my deepest gratitude to my supervisor, Associate Professor Laura Heathfield, for the unconditional support you have given me during this journey. You have taught me some of the most valuable lessons I know today, but perhaps the one I'll always remember is that asking for help is a sign of bravery and not weakness. Arriving at the University of Cape Town as a master's student feeling uncertain and timid, your unwavering confidence in my abilities has been fundamental in shaping me into the bold and confident researcher I am today. I am profoundly grateful for your mentorship and your endless insights.

I would like to extend my thanks to the following institutions for their contributions toward funding this project, the University of Cape Town, the Forensic Pathology Services and the National Health Laboratory Services (Grant No. 95842 95530). Furthermore, I want to thank the University of Cape Town and the National Research Foundation (Reference: PMDS230720137366) for providing financial support toward my tuition and living expenses.

To the Molecular Forensics Research Group and Biomedical Forensic Science staff and students, your support and motivation has been paramount to this achievement. Amy Whittaker, thank you so much for your support and assistance in the lab whenever I was in a crunch. Lisa Malan, you have been a friend, a creative and intellectual muse, but most of all a light in this sometimes-dark journey. I will be forever grateful for you. Kate, your friendship, mentorship and guidance have been invaluable to the completion of this PhD, thank you for your unconditional support.

To the team at Kings College London, Dr David Ballard, Dr Laurence Devesse and Lucinda Davenport, you have made my visit to Kings College London worthwhile, and I want to thank you for your guidance and teaching regarding the analysis of population sequence data. I would

herewith also like to thank the University of Cape Town for the provision of travel funding that made my research visit to Kings College London possible.

To my mom and dad, Crystal and Deon Martin, you both have supported me every step of the way and sacrificed your own dreams so that I could achieve mine. Know that I could never have achieved any of this without you. To my siblings, Bruce, Stash, Jake, Vinny, Tammy and Jonny, your motivation, banter, advice and more-than-occasional provision of food and yummy snacks has undoubtedly made this journey lighter to bear. Stash, your daily motivation, phone calls and encouragement from day one has meant the world to me. Thank you for teaching me how to take criticism gracefully.

To my girls, Lauren, Andrea, Jaimeé and Michaela, your camaraderie, support, advice and shared struggles have made this experience both more bearable and enjoyable. To my partner Xylon, you have endured every moment of this journey with me and supported me in every way possible. No matter how tough things were, you helped pull me back on track, helped me overcome my doubts and kept me balanced. Thank you showing me the importance of discipline and consistency. This achievement is as much yours as it is mine.

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

List of figures

List of tables

List of abbreviations

Abbreviation, unit or symbol	Full name
~	approximately
%	percentage
<	less than
>	more than
±	plus, or minus
3'	three prime
5'	five prime
°C	degrees Celsius
µL	micro litre
µM	micro molar
A	adenine
AEC	animal ethics committee
ai	ancestry informative
A-STR	autosomal STR
AT	analytical threshold
BLAST	basic local alignment search tool
C	cytosine
CA	California
CE	capillary electrophoresis
CI	confidence interval
C_T	cycle threshold
CEPH	centre for the study of human polymorphism
CODIS	combined deoxyribonucleic acid index system
DI	degradation index
DNA	deoxyribonucleic acid
dNTP	deoxyribonucleotide triphosphate
ddNTP	dideoxy ribonucleotide triphosphate
DPM	deoxyribonucleic acid primer mix
dsDNA	double-stranded deoxyribonucleic acid
DTT	dithiothreitol
DV	developmental validation
e. g	for example
etc	etcetera
EDTA	ethylenediaminetetraacetic acid
ENFSI	European network of forensic science institutes
FGx	forensic genomics
FHS	faculty of health sciences
FSSG	forensic short tandem repeat sequence structure guide
F_{ST}	fixation index
FTA	fast technology for analysis of nucleic acids

g	gram
G	guanine
GD	genetic diversity
GRCh38	genome reference consortium human build 38
HIC	high income country
HID	human identification
HREC	human research ethics committee
HS	high sensitivity
HSC	human sequencing control
HWE	Hardy-Weinberg equilibrium
i.e.	Latin: id est, English: that is
ii	identity informative
IPC	internal positive control
ISFG	international society of forensic genetics
ISO	international organisation for standardization
IT	interpretation threshold
IV	internal validation
LB	length-based
LTDNA	low template DNA
LMIC	low to middle income country
LNB	library normalisation beads
LR	likelihood ratio
M	molar
MA	Maryland
MBG	molecular biology grade water
MCMC	Markov chain Monte Carlo
MO	Missouri
MP	match probability
MPS	massively parallel sequencing
n	sample size
NA	not applicable
NaOH	sodium hydroxide
NDIS	national deoxyribonucleic acid index system
NDNAD	national deoxyribonucleic acid database
NFDD	national forensic deoxyribonucleic acid database
ng	nanogram
NGS	next generation sequencing
NIST	national institute of standards and technology
NTC	no-template control
OL	off-ladder
PBS	phosphate-buffered saline
PCR	polymerase chain reaction
PIC	polymorphic information content

pi	phenotypic informative
PM	post-mortem
PRISMA	preferred reporting items for systematic reviews and meta-analyses
qPCR	real-time PCR
RMP	random match probability
rpm	rotations per minute
SA	South Africa
SADC	Southern African Development Community
SAPS	South African Police Services
SB	sequence-based
SNP	single nucleotide polymorphism
SNV	single nucleotide variation
SRM	Salt River Mortuary
ST	stochastic threshold
STR	short tandem repeat
STRAF	short tandem repeat analysis for forensics
STRAND	short tandem repeat: align, name define
STRidER	short tandem repeats for identity European network of forensic science institutes reference database
ssDNA	single-stranded deoxyribonucleic acid
SWGDAM	scientific working group on deoxyribonucleic acid analysis methods
T	thymine
TE	tris-ethylenediaminetetraacetic acid
UAS	universal analysis software
UCT	University of Cape Town
UHR	unidentified human remains
UK	United Kingdom
UMIC	upper to middle income country
USA	United States of America
VNTR	variable number tandem repeat
WMA	world medical association
X-STR	X-chromosomal short tandem repeat
Y-STR	Y-chromosomal short tandem repeat

Chapter 1: Introduction

1.1. Study background

The tombstone of the famous English poet, John Keats, bears the epitaph: “Here lies one whose name was writ in water”. This poignant epitaph emphasises the fear that human beings have of being forgotten by the world. For many, being buried nameless is a brutal reality. Establishing the identity of a deceased individual is therefore a crucial aspect of medico-legal death investigations. Globally, the number of individuals that remain unidentified each year is substantially greater in developing countries [1]. From a more local perspective, South Africa has been facing a public health crisis, exemplified by the alarming number of deceased individuals whose identities have yet to be determined [1-3]. More specifically, a review conducted by Reid *et al*., 2020 has highlighted that, over an eight-year period (2010 – 2017), 2476 bodies (9.2% of annual caseload) admitted to Salt River Mortuary in the Western Cape, South Africa were unidentified each year. These bodies must eventually be buried as paupers and represent bodies from just one out of more than a hundred mortuaries in the country [2].

The storage, maintenance and investigation into the high number of unidentified bodies places a massive strain on forensic mortuary resources and is a public health concern [1]. A more tragic consequence of the high number of bodies left unidentified is that the families of missing loved ones are left in distress and agony due to an absence of closure. This underscores the critical necessity to enhance the efficiency of existing human identification procedures, and more specifically deoxyribonucleic acid (DNA)-based methods [3]. This study was therefore developed within the context of a rising interest and need for facilitating the implementation of massively parallel sequencing (MPS) and to advance current forensic human identification methods in South Africa.

In South Africa, the Forensic Pathology Services investigates all unnatural deaths, and as far as possible assists the South African Police Services (SAPS) with the identification of the deceased (Inquests Act 58 of 1959) [4,5]. Current identification procedures at South African forensic mortuaries typically involve visual recognition by next-of-kin, except where there is an absence of visual identifiers due to trauma, injury, decomposition, burning, or where a body is incomplete [4]. While visual recognition serves as the quickest method for identifying deceased individuals, it can prove to be unreliable and complex in its execution, especially when not accompanied by corroborating evidence, as observed in a locally notorious case in South Africa [6]. In this case, a prisoner, Thabo Bester had faked his own death to escape from prison. Authorities had found the burnt remains of an individual in his cell after a fire, and they initially believed it belonged to Thabo Bester. Due to the body being visually unrecognisable, it was deliberately misidentified as Bester by his next-of-kin and collected from the mortuary under false pretences. Only a year later, DNA evidence was used to identify the body, which was confirmed to belong to Katlego Bereng, who was killed and his body used as a decoy [6].

The complexities related to the unreliability of visual recognition is exacerbated by fraudulent kinship claims with motives ranging from illicit insurance claims to potentially darker criminal intentions, as was demonstrated in the above-mentioned Thabo Bester case [7,8]. Therefore, in the absence of reliable and honest visual recognition by next-of-kin, the reliance on scientifically rigorous methods becomes critical to the identification of the deceased. These methods include fingerprint identification, dental and radiological analysis, anthropological assessment, and DNA-based identification [2,4].

Where bodies have reached advanced stages of decomposition, have been severely burnt or mutilated, the success of fingerprint identification is considerably impaired [1-3]. Similarly, dental and radiological identification are highly reliant on the availability of prior dental records or X-rays [9]. This is a requirement that is often unmet for migrants and low-income populations in South Africa [10]. Furthermore, cases involving remains that have been burnt or mutilated often preclude the possibility of comprehensive skeletal and/or anthropological analysis that could otherwise assist in generating leads about the identity of an individual. DNA-based identification serves as an alternate and supporting route for positive identification, presuming the availability of good quality and intact DNA samples [11].

The current methodologies employed in DNA-based identification exploit genetic markers known as short tandem repeats (STRs). These markers are ideal for individual identification, as they are highly variable between people [12]. By targeting these STRs, a DNA profile can be generated based on the separation of DNA fragments by size. This separation technique is known as capillary electrophoresis (CE), and the method has been prominent in the forensic genetics industry for over three decades [13]. Although CE-based methods have withstood the test of time, the method is not without shortcomings.

Particularly, challenges arise when the DNA within biological samples have become fragmented due to degradation, or when their quality has been compromised by substances that inhibit polymerase chain reaction (PCR) [14]. The method lacks sufficient sensitivity for compromised DNA samples and low template DNA (LTDNA) and is only able to analyse a limited number of STRs in a single reaction [15]. Global efforts to increase the multiplexing capacity of CE-based STR profiling kits have been made, although this number is approaching

a plateau due to the limitations of the dye labelling systems employed by CE technology [16,17].

Massively parallel sequencing (MPS) has emerged as a technique that can be leveraged to overcome the limitations posed by CE-based methods [18,19]. In 2015, Illumina Inc. and Verogen Inc. (now QIAGEN) released the MiSeq FGx™ platform, which was the first MPS system that was validated for forensic use [20,21]. The kit indeed addressed many of the shortcomings that came with CE-based methods but was mainly superior in its capacity to 1) multiplex more (*i.e.*, >200) genetic markers in a single reaction; 2) generate sequence information for alleles of the same length and 3) offer a higher sensitivity for low input DNA samples [20,21]. Forensic DNA profiling was thus no longer restricted to obtaining size-based information for up to ~30 STR markers, but rather sequence *and* size-based information for over 200 markers.

For DNA samples that have been degraded due to post-mortem environmental conditions, using a primer set that targeted more markers with shorter amplicons would enable a deeper and more insightful comparison to reference or familial reference samples for identification [15]. Furthermore, the primer set included in the ForenSeq™ DNA Signature Prep kit also enabled the prediction of phenotypic characteristics such as hair and eye colour, as well as biogeographical ancestry to generate investigative leads, especially where samples are obtained from bodies that are visually unrecognisable [22].

1.2. Rationale

Naturally, prior to the use of any new technology, kit or workflow in the field of forensic genetics, there are certain requirements that need to be met. Firstly, as with CE, to facilitate statistical interpretation of a DNA profile, allele frequency data from the population is required [23]. In this case*, sequence-based* allele frequencies are required for statistical interpretation of MPS-based DNA profiles. However, prior to undertaking a large-scale population study, it is essential that the workflow is optimised to a laboratory's resources and needs, but also that it is compatible with the sample types predominantly encountered and used in the laboratory [24]. Finally, the workflow, from sample preparation through to sequencing should be internally validated according to established guidelines to ensure that it is fit for purpose [25,26].

To facilitate the implementation of a forensically relevant MPS workflow for human identification, there is an urgent need for sequence-based population data in African countries. More specifically, no sequence-based allele frequency data exists for the South African population. It is widely known that a high degree of genetic variation exists in African individuals; yet no studies have been undertaken to map this variation from a forensic perspective [27]. To elucidate this gap, an in-depth review of the literature is required to understand the magnitude and direction of sequence-based population data that have been generated globally using the ForenSeq™ DNA Signature Prep kit. It was hypothesised that African populations would exhibit higher levels of sequence variation and analysis of markers would result in higher discriminatory statistics.

Furthermore, an optimised workflow for generating large-scale sequence-based population data using a streamlined and direct PCR approach is needed to reduce the time and costs associated with sample preparation, but herewith, also costs associated with potential failures

that could result from incompatibility of direct PCR sample types and kits with more sensitive MPS chemistries [28]. Crude buccal swab lysates, which are typically processed using a direct PCR approach, are routinely used in our Biomedical Forensic Science Laboratory as reference DNA samples for the purposes of generating population data, as well as for conducting DNA analyses on post-mortem samples collected during autopsy at the Observatory Forensic Pathology Institute in Cape Town, South Africa. Thus far, no studies have optimised the use of crude buccal swab lysates from both living and deceased individuals for MPS applications with the ForenSeq™ DNA Signature Prep kit.

After optimisation, a considerable component in the facilitation of implementation of MPS in a forensic laboratory is the internal validation of the MPS workflow, ensuring that it is fit for purpose in terms of sensitivity, accuracy, precision, repeatability, reproducibility, stability and species specificity, but more importantly that the data generated can be used in court and legal proceedings. The International Standards Organisation (ISO) 17025 standard, adapted for forensic laboratories by the Scientific Working Group on DNA Analysis Methods (SWGDAM) and the European Network of Forensic Science Institutes (ENFSI) requires that any method, technique or workflow used in a testing laboratory must be internally validated [25,26,29]. Very few studies using the ForenSeq™ DNA Signature Prep kit have focused on the internal validation of post-mortem samples, and no studies have internally validated the workflow using post-mortem crude buccal swab lysates using a direct PCR approach [30].

1.3. Aim and objectives

To address the gaps and shortcomings revealed in the rationale, the aim of this study was to facilitate the implementation of the ForenSeq™ DNA Signature Prep kit workflow for post-mortem human identification purposes.

The objectives were to:

1) Conduct a comprehensive review and meta-analysis of ForenSeq™ DNA Signature Prep kit population studies to establish evidence-based insight for optimising and informing the workflow in a South African MPS population study. To do so, the systematic review objectives were to:
    a) assess the forensic MPS community's inclination to using direct PCR approaches in MPS population studies,
    b) evaluate trends in genetic diversity in African and non-African populations,
    c) assess the level of standardisation for STR sequence nomenclature, and
    d) investigate concordance patterns to inform analysis of sequence data for the South African population.

2) Develop an optimised protocol for generating high first-time success rates of crude buccal swab lysates processed with the ForenSeq™ DNA Signature Prep kit, in preparation for use in an MPS population study. Achieving this objective required:
    a) initial testing of crude buccal swab lysates and investigation into their sub-optimal success rates with the ForenSeq™ DNA Signature Prep kit and
    b) assessing the impact of different optimisation methods to determine a method that results in adequate library quality and quantity that would improve MPS profile success rates.

3) Establish sequence-based data for the two major South African population groups. To achieve this objective, the following sub-objectives were identified:

a) carry out sequencing using the ForenSeq™ DNA Signature Prep kit on DNA samples from South African volunteers representing the two major population groups,
b) characterise and name genetic variants identified in the South African population in collaboration with Kings College London,
c) generate sequence-based allele frequency data, population, and forensic parameters for the two major South African population groups using A-STRs and
d) assess concordance between length-based genotypes generated using the ForenSeq™ DNA Signature Prep kit and CE-based methods.

4) Perform an internal validation of the ForenSeq™ DNA Signature Prep workflow with DNA primer mix B (DPMB) on the MiSeq FGx™ platform to assess if the workflow is fit for purpose in a mortuary-linked laboratory. The internal validation study thus aimed to:
   a) establish performance parameters for the workflow in terms of sensitivity, accuracy, precision, call rate, stability, species specificity, repeatability and reproducibility, as applied to control DNA and authentic forensic samples including bone, nail, teeth, blood, formalin-fixed paraffin embedded (FFPE) tissue, buccal cells on Flindlers® Technology Associates (FTA) cards, and for the first time, post-mortem crude buccal swab lysates and
   b) apply the internally validated workflow to a forensic cold case to demonstrate the fitness of purpose and applicability of the workflow in an authentic post-mortem context.

1.4. Study design

To carry out the objectives, a quantitative, cross-sectional design was used. This study employed a combination of descriptive, correlational and experimental research methods, as relevant to each objective. All laboratory work was undertaken in an academic laboratory that complies with ISO 17025 guidelines with an established quality management system, and is linked to a forensic mortuary in Cape Town, South Africa. The population assessed in this study consisted of 463 Black African (n = 216) and Admixed South African individuals (n = 247). These two populations were chosen for the study as they make up most of the South African population. They also represent the demographics of the unidentified human bodies in the Western Cape and are underrepresented by the current literature [1,2]. Another design consideration in the population study was the use of A-STRs alone. This was done firstly because A-STRs are primarily and routinely used in forensic casework in South Africa [31]. Focusing A-STRs ensures that the study addresses the most critical gaps first. The focus on A-STRs therefore serve as a strong foundation for analysis and research on sex-linked STRs and single nucleotide polymorphism (SNV) markers.

Secondly, as this is the first sequence-based population study in South Africa, and a high level of variation is expected within STR markers, the Y- and X-STR markers would also require in-depth analysis, characterisation and concordance assessment. This additional analysis does not fit into the time-scope of this project and is currently being undertaken in a separate study in our research group to increase capacity building and maximise technical resources. Furthermore, A-STRs are the only marker type that can currently undergo standardised sequence quality control [32]. Selecting A-STRs to focus on in the early stages of implementation has ensured that data generated meets required international standards. Since

Y- and X-STR sequence data would require manual curation and quality control, this is explored in a separate study.

1.5. Theoretical and conceptual background

1.5.1. The role of DNA in identification

In comprehending the role of DNA in human identification, it is crucial to understand its ability to distinguish between different people. This concept was first brought to light by Sir Alec Jeffreys, through his discovery of repetitive sequences of DNA present in the human genome. These sequences are known as variable number tandem repeats (VNTRs), in which the counts of repeats vary from person to person [12]. As it later emerged, this inherent variation set the stage for what would evolve into forensic DNA profiling through the discovery of STRs. This pivotal advancement has since proved instrumental in the field of forensic genetics [12]. This section will provide a theoretical and conceptual background into important principles regarding both conventional DNA profiling and MPS in a forensic context and highlight the limitations of CE technology that have been addressed with the onset of MPS. The section concludes with the elicitation of requirements for implementation of MPS in forensics.

1.5.2. STRs in forensic DNA profiling

An STR is a short repetitive sequence of DNA that is 2-6 bp in length. As with VNTRs, they are highly polymorphic in nature due the variability in the number of repeats between different individuals [12]. They are however shorter than VNTRs in length, making them suitable for analysing degraded DNA, which is common in forensic settings [14]. The technique typically used to assess the length of STR alleles in DNA profiling is called CE, which relies on the negatively charged DNA migrating through a capillary filled with a polymer matrix under an electric current [13]. An electric charge is applied, causing smaller fragments to migrate faster

than larger fragments. Fluorescently labelled fragments are detected using CE, and their overall size is determined relative to an internal size standard. Software is then used to determine the size of the STR (or the number of repeats), which is then designated as a length-based allele. In simple terms, a forensic DNA profile is made up of several allele pairs or genotypes present at various STR locations on the autosomal, X or Y chromosomes [13].

1.5.3. DNA databases

When a DNA profile is generated and the donor is unknown, the profile may be searched against another DNA profile, and it can be searched against a database. This database would contain DNA profiles of individuals falling into different categories [33]. Most national DNA databases have some form of index system, where DNA profiles from various sources are categorised into indices, and these often differ by country. For example, in the United States of America (USA), the Federal Bureau of Investigations National DNA Index System (NDIS) features multiple indices designed to assist with criminal investigations and identification processes [34]. The NDIS includes indices for convicted offenders, arrestees, crime scene samples, missing persons, relatives of missing persons and unidentified human remains.

The United Kingdom (UK) has one of the world's most advanced and longest-standing DNA databases [33]. Broadly, the profiles stored in the National DNA Database (NDNAD) can either fall into the category of an individual and is known as a "subject DNA profile" or the category of an "unidentified individual", which includes DNA profiles generated from DNA samples collected from crime scenes [35]. The NDNAD Delivery Unit also operates a Missing Persons DNA Database which includes DNA profiles from missing persons items and those from unidentified human bodies or body parts. This database is held separate from the NDNAD, and

searches are only conducted between and against missing persons and unidentified deceased persons [35].

In South Africa, the National Forensic DNA Database (NFDD) was created in 2015 and has demonstrated its efficiency and success by improving conviction rates in part through the generation of forensic investigative leads [36]. The NFDD comprises different indices for categorising forensic DNA profiles, and these include the crime scene index, arrestee index, convicted offender index, investigative index, elimination index and missing persons and unidentified human remains index [36]. Although, the lack of awareness and lack of adequate knowledge to provide these DNA samples and mistrust in police authorities in Africa, among other reasons, may prevent families from providing reference DNA samples in cases of missing persons [37].

The second type of database that is pivotal to forensic DNA statistics is a population database. This is different to a national DNA database and consists of allele frequencies for different forensically relevant markers [23]. Essentially a core set of loci are used. The standardisation of a marker set is yet to be achieved on a global scale, but efforts have been made to use the same core set of loci in different countries to enable international searches and collation of allele frequency data [38].

1.5.4. Match statistics

When a match is found between a DNA profile in question (unknown donor) and/or a DNA profile held in a database, then the confidence in this match is determined. This is called random match probability (RMP) and is the probability that the DNA profile in question matches the DNA profile from a random, unrelated individual in the population [39]. What is required for

this calculation is the frequencies of alleles found in the background population for a particular set of genetic markers (mostly STRs) [39]. If the number of markers tested increases, then the RMP will decrease as testing more markers will result in a higher power of discrimination [15,40]. This is where CE methods have fallen short, specifically with regards to more compromised samples that are often encountered in forensic mortuaries. This is because only a limited number of markers can be multiplexed using CE methods.

For several years, the forensic community has relied on the calculation of match statistics to determine the rarity of a DNA profile in a population group [39]. This calculation is based on the frequencies of alleles in the background/general population. The smaller the RMP value, the higher the probability is that two DNA profiles originate from the same source or donor. For A-STR markers, it is possible to use the product rule to determine the RMP of a DNA profile match. This is owed to the nature of autosomal STR markers when they are inherited independently [41,42]. The use of a likelihood ratio (LR) is also frequently used and is a ratio of the evidence given one hypothesis over the evidence given an alternative hypothesis. A higher LR indicates stronger evidence for the two profiles belonging to the same person [39].

Interpretation of DNA profiles by humans can become complex when generated from compromised DNA samples. Traditionally, the calculation of match statistics for single-source DNA profiles does not consider peak heights (or intensities) of stutter, artifacts, or unknown genotypes, as stutter is filtered out beforehand. However, these variables become especially important when interpreting mixed DNA profiles, where they contribute to uncertainty. These variables play a role in the size and accuracy of the match statistic, an important aspect in reporting. Accounting for these variables, specifically when dealing with mixtures, may change a match statistic from 13000 to 189 billion, as demonstrated in a murder case in Pennsylvania,

USA [43]. Shortcomings of traditional DNA profile interpretation software and approaches were met with the development or rather the use of probabilistic genotyping software. It emerged as a tool that is inherently unbiased in its approach at resolving complex DNA profiles and is widely used in the forensic community for resolving complex mixed DNA profiles (or mixtures).

With respect to mixtures, the probabilistic approach involves computational estimations of allele dropout and drop-in probabilities [44]. For example, TrueAllele® (Cybergenetics, Pittsburgh, USA) uses statistical modelling to infer a probability for each possible genotype in a DNA profile [45]. The model considers all data (*e.g.*, unknown genotypes, stutter, artifacts) and uses it to plot uncertainties for each possible explanation of the data. Through Markov chain Monte Carlo (MCMC) statistical sampling, thousands of genotype combinations are generated and proposed, which account for all data and their associated uncertainties. The software then generates a distribution of probabilities that could explain the DNA results. These results are used in LR calculations, where solutions or explanations that have higher probabilities explain the data better than solutions with lower probabilities. This method has been validated in the analysis of complex, low-level, and higher-order mixtures that were once considered too challenging to interpret [46]. It marks a major transformation in how DNA data is analysed but has not been fully explored in complex kinship cases [44].

In the assessment of kinship or parentage, a match would not occur between two DNA profiles, but rather a measure of relatedness would be determined. Kinship analysis tools estimate the LRs for determining relationships such as paternity, maternity, and sibling relationships (more commonly referred to as sibships) by making use of population allele frequency databases [47]. However, the variation in numerical precision across different software can lead to different

conclusions, which is particularly important when distinguishing between full and half siblings. This is because both full and half siblings could share no alleles at several markers due to inheriting the alternate allele from a parent.

A study by Mályusz et al., 2006, suggested that relying solely on autosomal STR typing is inadequate for accurately resolving cases involving potential half-sibship, although using a combination of both STRs and SNVs has shown to result in more reliable LR calculations [48,49]. This is a major limitation of current CE systems, but more so for the SAPS Forensic Science Laboratory, as only 15 STR markers are used in kinship tests, often resulting in inconclusive results (in-house data). The use of probabilistic genotyping software in kinship assessment, combined with the increase in genetic information obtained through MPS, could serve as a method that results in more informative sibship estimations.

1.5.5. Limitations of STR typing using CE technology

In forensic DNA profiling, a limited number of STRs can be multiplexed due to varying fluorescence wavelengths of dyes [13]. Fluorescent detection allows for the tagging or labelling of PCR products with overlapping sizes, such that they can be distinguished by both fluorescence and size. In a dye system, one of the dyes are used for labelling the size standard that would be added to each sample. Each of the fluorescent dyes added to the PCR reaction will emit a different maximum fluorescence wavelength. This difference in wavelength means that when the emissions are captured by a camera, the fluorescence will be separated and enables the visualisation of markers with the same or similar sizes [13].

With an increasing number of STRs that can be multiplexed, the more information there is to distinguish between different individuals, even if only a few markers are genotyped [41].

However, the number of markers that can be multiplexed in CE platforms have been reaching stagnation because there are a limited number of fluorescent dyes that can be used simultaneously without significant spectral overlap, as each dye must have a distinct emission wavelength to be able to separate markers with fragments of similar sizes. The maximum number of dyes has increased to eight, enabling the multiplexing of a maximum of 35 STR markers, including an increased number of mini-STRs [17]. However, it is not known at which rate the number of dyes will increase over the next few years. The current rate at which improvements are being made in this regard (for CE) poses a limitation for samples that are compromised, such as those that are degraded, or contain PCR inhibitors. Although, MPS chemistries are also not always robust, and may be similarly impacted by PCR inhibitors.

When a DNA sample is degraded, there is an abundance of shorter fragments available for amplification, and this results in what is known as preferential amplification during PCR. This means that markers targeting larger amplicons will ultimately be amplified at a lower rate or not be amplified at all, resulting in a failed or partial DNA profile [14]. When there are less markers available for comparison between an unknown donor sample and a reference profile, then the confidence in a match is reduced and may not be usable for court [48]. Due to this limitation, a secondary consequence is that when the number of A-STR markers that successfully amplified are too few, testing with different markers, such as X-STR and Y-STR testing may be undertaken. These analyses would typically need to be performed as additional tests, except for the latest commercial STR kit which couples both A-STR and Y-STR markers [17].

The principles of CE-based technology are widely used and understood by the forensic community, and because of this, many studies have been undertaken to improve its ability in

gaining maximum insights from compromised samples (*i.e.*, low input, degraded or inhibited). [14,50-52]. However, as previously highlighted, its capacity to increase the number of markers that can be multiplexed is limited by the number of dyes that can be included in a system. Consequentially, this limits insights from degraded samples, as more markers of different non-overlapping sizes need to be multiplexed to obtain more genetic information [15,53].

When bodies are visually unrecognisable and CE methods fall short, generating additional information from forensic DNA samples may be carried out. These include data to estimate externally visible characteristics to generate new investigative leads [54,55]. Therefore, the forensic community have tapped into using MPS to address the limitations of CE that have not been capable of being addressed through optimisation or troubleshooting.

1.5.6. Massively parallel sequencing: Entering the genomic era

The field of DNA sequencing has evolved tremendously since its inception, starting with the work of Frederick Sanger, who pioneered the first DNA sequencing method commonly known as Sanger sequencing, and known as the "chain termination method" [56]. In this method, DNA base pairs (bp) within a sequence are determined in a fragment of DNA, typically less than 1000 bp. What made Sanger sequencing different from conventional PCR was the incorporation of dideoxynucleotides triphosphates (ddNTPs), which lack a 3′ hydroxyl group into a reaction and allows for the termination of the growing strand of DNA when a ddNTP is incorporated. This method allowed for the accurate reading of DNA base pairs, albeit a manual and labour-intensive process. The detection of fragments has become largely automated, but this has not been suitable for generating large-scale sequence data [57]. More advanced and efficient methods of sequencing were needed that led to the onset of MPS technologies, introducing a paradigm shift in DNA analysis [58]. MPS technology enables the simultaneous

sequencing of a large number (millions) of DNA fragments in parallel, increasing the sequencing throughout and reducing the time and costs associated with sequencing large numbers of genetic markers or even entire genomes [19].

Although MPS has been widely used in many molecular biology fields since its inception in 2005, it has only recently been used in the field of forensic genetics [19]. In 2015, the MiSeq FGx™ platform, which was the first forensically validated MPS system that addressed many of the above-mentioned CE limitations [21,59]. What made this advancement significant was that the system with its bioinformatic pipeline, referred to as the ForenSeq™ Universal Analysis Software (UAS), was validated for forensic use, allowing it to gain high traction in the forensic community, as it was a huge step towards being able to use the technology for forensic casework [60].

The onset of MPS has highlighted significant advantages over CE in forensic DNA analysis, including enhanced resolution of complex mixtures, the ability to analyse a broader range of genetic markers simultaneously, and improved sensitivity and accuracy in detecting low-frequency alleles. The MiSeq FGx™ system was first coupled with the ForenSeq™ DNA Signature Prep kit. The ForenSeq™ DNA Signature Prep kit was developed to amplify forensically relevant markers including Amelogenin, 27 A-STRs, 24 Y-STRs, 7 X-STRs and 94 identity-informative SNVs (iiSNVs) with the DPMA. Additionally, the DPMB enabled further multiplexing of an additional 56 biogeographical ancestry-informative SNVs (aiSNVs) and 22 phenotypic-informative SNVs (piSNVs) [21]. Having the option to select a primer mix that either targets only STRs, only STRs and iiSNVs or include aiSNVs and piSNVs with DPMB are therefore ideal for countries where it is deemed legally and ethically inappropriate to reveal phenotypic or ancestry information.

1.5.6.1. The ForenSeq™ DNA Signature Prep Kit workflow with the MiSeq FGx™ system

Sample and library preparation

As with traditional CE workflows, the ForenSeq™ DNA Signature Prep kit workflow begins with a sample preparation step, whereby either extracted/purified DNA is prepared and diluted to a concentration of 0.2 ng/µL (5 µL), or a direct PCR approach is employed. The manufacturer's protocol requires that all extracted DNA samples are quantified using a fluorescence-based quantification method [21]. The kit has been evaluated for a range of different extracted DNA sample types including bone, teeth, blood, saliva and buccal cells, but research on its performance with direct PCR crude lysates is limited [20,30,61]. Nonetheless, the manufacturers recommend that crude lysates can be added directly to the first PCR reaction with the addition of nuclease-free water, while FTA cards require a washing step prior to addition to the PCR reaction. No quantification or quality control steps are required by the manufacturer when using direct PCR samples (such as FTA cards or crude lysates). The implication for this lack of quality control is further explored in Chapter 3.

Following sample preparation, a two-step PCR is required to first amplify targets and then to add adapters onto the DNA fragments. The adapters contain three essential sets of sequence tags including: a) the 3′ or 5′ sequence of each adapter that complements the initial primers used the in the first PCR step, b) the i5 and i7 indices that are both added to each target amplicon to create a unique index for each sample and c) a sequence tag commonly called "P5" or "P7", complementary to the oligonucleotide sequences present on the flow cell (**Figure 1.1**) [19].

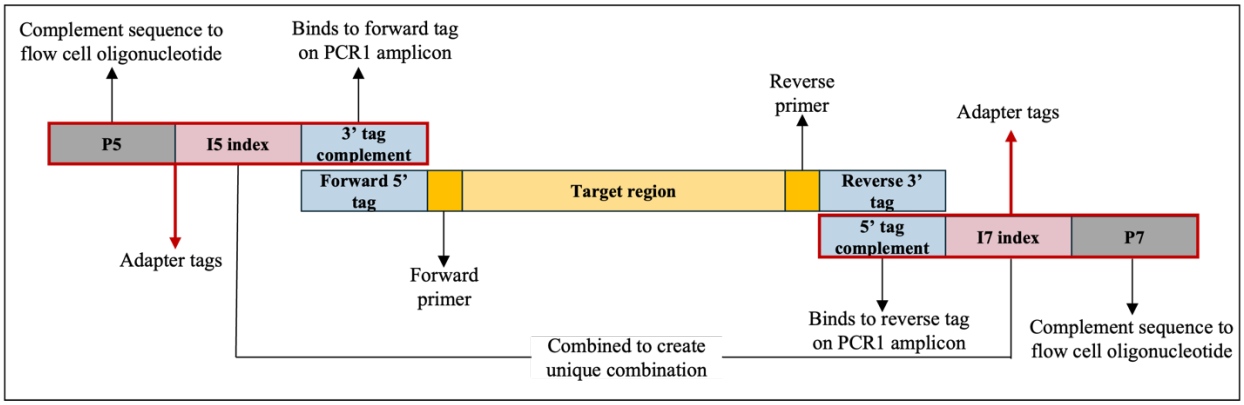

Figure 1.1: *Diagram illustrating adapter ligation of the target region. The three components of the adapter are highlighted in the red box, with the target region shown in yellow.*

Libraries are then purified through magnetic bead-based purification to remove free adapters and unbound primers. Following purification, the libraries do not undergo quantification as with most other library preparation protocols in molecular biology, instead they undergo magnetic bead-based normalisation, where the bead to sample ratio ensures that the binding capacity results in equal concentrations of each library being added to the flow cell for sequencing [62]. An equal volume of each library is then pooled into a single tube and diluted with a hybridisation buffer. Hereafter, a human sequencing control (HSC) is denatured and added to the pooled library. The HSC is used as a positive control consisting of 23 pooled STR loci and enables the MiSeq FGx™ sequencing run to reach completion. Denaturation of the pooled library is required prior to loading of the library onto the flow cell, and this consists of first heating and then immediately snap freezing the tube containing the pooled library, after which the normalised, pooled denatured library is added to a MiSeq FGx™ reagent cartridge [21,59].

Sequencing

A flow cell contains two types of oligonucleotides that are complimentary to each of the P5 and P7 adapter sequences added during adapter ligation. This makes it possible for fragments

to hybridise to the flow cell. Fragments added to the flow cell are single stranded DNA (ssDNA) molecules and the P5 and P7 adapter sequences present on these fragments bind to the P5 and P7 oligonucleotides on the flow cell. A primer mix containing a mixture of polymerase and nucleotides are added to the flow cell which enables synthesis of a new complementary strand, resulting in a now double-stranded DNA (dsDNA) molecule. The dsDNA molecule is then denatured, and the template strand is removed, leaving only newly synthesised ssDNA. The "free" P5 and P7 oligonucleotides present on the flow cell are then hybridised by this ssDNA by a bending or arching over action, resulting in a bridge-like formation. From this bridge, a new complementary strand is made, which undergoes another denaturation step, and the ssDNA arches over again to bind to the P5 and/or P7 adapters on the flow cell, forming another bridge and synthesising a new strand [63]. The process is repeated until there are sufficient clusters of each fragment. The final step in cluster generation is the washing away of all reverse strands, with only the forward strands remaining for sequencing by synthesis [21,59,64,65].

The MiSeq FGx™ reagent cartridge contain the primer mixes for the sequencing of "read 1" and "read 2" primers. The "read 1" primer binds to the P5 sequence tag, while the polymerase and all four fluorescently labelled deoxyribonucleotide triphosphates (dNTPs) are added to the flow cell at the same time (**Figure 1.2**). The dNTPs are then incorporated sequentially based on the nucleotides present on the forward strand. The reading of the sequences in the strand are made possible through fluorescence capturing by a camera. As each dNTP is tagged with a different fluorescent dye, the camera records different emission wavelengths for each nucleotide incorporated, bearing in mind that the signal intensifies as the number of clusters increase. For "read 1", this process involves 351 cycles of sequencing, after which all the

synthesised strands resulting from “read 1” are washed away, as the signals from each cluster would be recorded and stored for further processing [59].

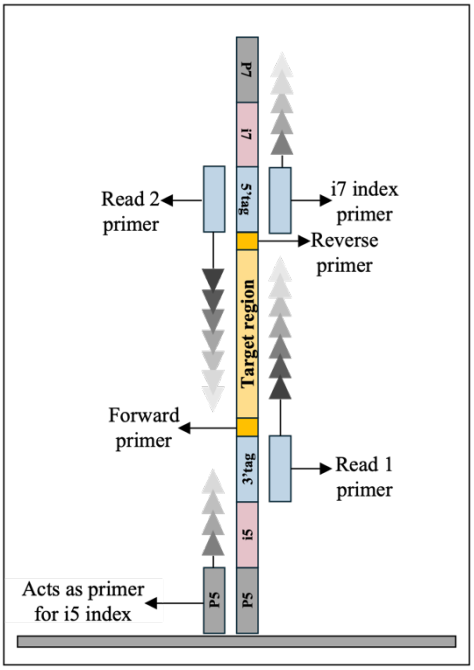

Figure 1.2: *Read 1 and Read 2 sequencing process.*

The i7 index sequencing is enabled by the addition of the i7 sequencing primer and is carried out for eight cycles, followed by eight cycles of sequencing for the i5 index sequencing using the P5 oligonucleotide attached to the flow cell as the sequencing primer. Once all i5 index read products (newly synthesised strands) are washed away, leaving only the template strand, a new complement strand is made, resulting in a dsDNA molecule that is denatured. The template strand is washed away, with the reverse strand remaining. Then only is the “read 2” primer introduced and attaches to the reverse tag on the 3´ end of the strand for 31 sequencing cycles.

Fluorescence imaging captures each nucleotide that was incorporated through its fluorescent label. The signal emitted from each cluster is recorded along with the wavelength of the different dyes for each nucleotide. After each cycle, a single image illustrates various clusters

of signals recorded by the camera, and the intensities are recorded based on their originally recorded co-ordinates on the flow cell. Base calls are recorded in a "bcl" file, dependent on whether clusters met Illumina's built in chastity filter, along with the Phred scores. All sequences in clusters with the same i5-i7 index will be combined to generate a "read 1 fastq" file with all forward strand sequences called per unique i5-i7 index, and a "read 2 fastq" file for the same sample but with the reverse strand sequencing reads [66]. Both these files can then be used with the integrated ForenSeq™ UAS for automated sequence alignment, extraction and repeat counting (for STRs) based on default or internally established thresholds (*i.e.*, analytical, interpretation, stutter and heterozygous balance thresholds) [60].

1.5.7. Requirements for implementation of MPS in post-mortem human identification strategies

A survey conducted in Europe by Alonso *et al*., 2017 to establish opinions, use and interest from forensic laboratories in MPS (hereon referred to as the European MPS survey) revealed reasons why laboratories did not purchase an MPS instrument. These reasons included the high costs, lack of sufficient funding for implementation, and reluctance to purchase when methods were still underdeveloped. Subsequently, a surge of studies was undertaken in the years that followed, which evaluated the performance of MPS in a forensic context. Further to this, sequence-based population studies were undertaken [67].

As with CE, background allele frequency data for the relevant population are needed to facilitate the calculation of match statistics [68]. These sequence-based allele frequency data need to be compatible *and* concordant with those obtained from CE [69]. Compatibility in this context means that an MPS profile should be stored in such a manner that it can easily be searched in a CE DNA profile database and/or population allele frequency database. Concordance refers to whether the length-based genotype calls made by each technology (CE

or MPS) are the same. These are non-negotiable requirements for implementing MPS in a forensic laboratory [69].

Several population studies have now been conducted using MPS, with most studies to date having used the ForenSeq™ DNA Signature Prep kit [70-75]. These studies have aimed to demonstrate the compatibility and concordance of the method; however, African countries have been slow to align with the MPS trajectory in the forensic space. This slow adoption is also evident in the implementation of CE technology, which only gained traction in African countries some 30 years after its inception, including in Burkina Faso, Ghana, Kenya and Sierra Leone [76-79]. It is particularly important for African countries to remain aligned with methods that would improve aspects of forensic human identification, as they endure a high number of unidentified human bodies each year [1]. The need for sequence-based population data in African countries is fully explored in the systematic review in Chapter 2, striving to facilitate the positioning and standardisation of MPS in both developed and developing regions.

Implementing a new and costly technology is a daunting endeavour for many African countries, especially when no specific sequence analysis and validation guidelines exist. This is compounded by the lack of consensus regarding nomenclature standards and the almost complete absence of sequence-based allele frequency data for African populations [32].

Additionally, MPS is not only more sensitive in its ability to generate more insight into samples than CE, but it is superior in its ability to resolve complex profiles that are challenging to solve through only using CE methods. Indeed, several studies have demonstrated the compatibility, concordance and allelic gain provided using MPS, and have demonstrated examples where MPS can provide an answer when CE-based methods cannot [22,80,81]. This is important for post-mortem samples that are degraded or compromised, as the limited number of markers used

in CE have not always been sufficient when performing kinship testing [48,49]. To this end, DNA profile success rates using the ForenSeq™ DNA Signature Prep kit have outperformed and supported CE analysis for post-mortem sample types, although showing lower performance on ancient human remains. This was revealed after a mini systematic search of the literature was performed, pertaining to post-mortem sample types processed with the ForenSeq™ DNA Signature Prep kit [81-85] (Appendix 1.1).

What is further needed for implementation of MPS in a forensic context is the establishment of performance parameters of a wide range of sample types [29]. Studies have validated the ForenSeq™ DNA Signature Prep kit workflow on several sample types and established performance parameters with respect to sensitivity, stability, accuracy, precision, call rate, mixture resolution and species specificity. Sensitivity studies have showed that a minimum input range of 0.25 ng – 1 ng is recommended [20,62]. For control DNA samples, studies have deemed the workflow to be accurate and precise under unstable conditions (*i.e.*, in the presence of PCR inhibitors and subjected to DNA degradation) [20,61]. However, few studies have validated the workflow in a post-mortem context and focused on authentic forensic samples, while to this author's knowledge, none have validated the workflow using a direct PCR approach for post-mortem samples [30]. The gaps highlighted above, in combination with the requirements for implementation of the workflow speak directly to the objectives of this study, and these requirements have been used as the basis for carrying out this study.

1.6. Thesis roadmap

The thesis consists of six chapters, and a concise illustration of the thesis roadmap can be found in **Figure 1.3**. Chapter 1 introduces important definitions, concepts and methods currently used in DNA-based human identification. This chapter identifies the position of an MPS workflow in forensic human identification, highlighting previous successes with the workflow and requirements needed to facilitate implementation.

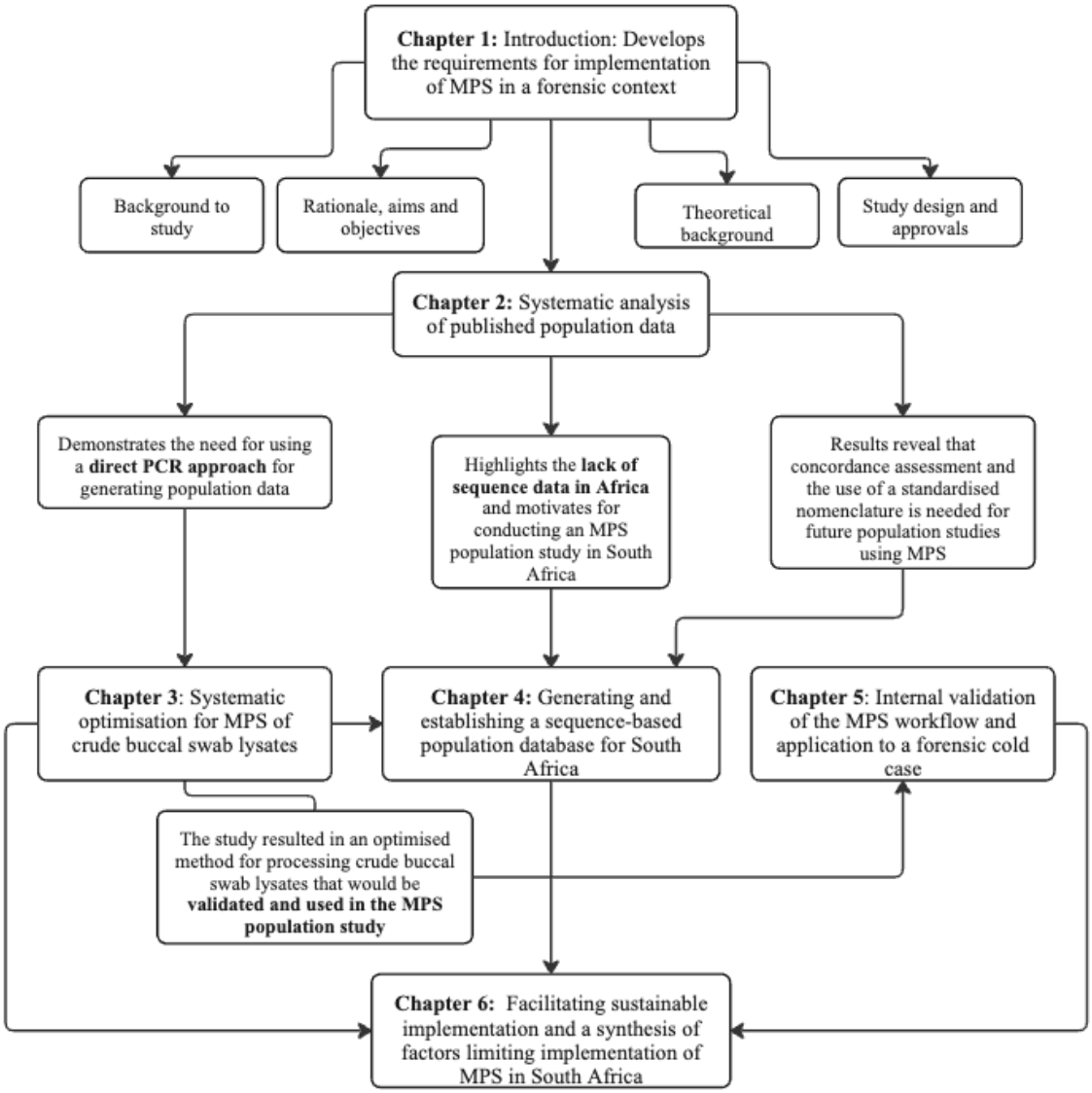

Figure 1.3: *Thesis roadmap*

Chapter 2 is a systematic review of the literature pertaining to published ForenSeq™ DNA Signature Prep kit population studies. A meta-analysis was conducted for a subset of studies to understand the increase in variation provided through sequence data compared to length-based data. Additionally, the review assesses the current consensus on concordance and discordance between data generated from the ForenSeq™ DNA Signature Prep kit and commercial CE kits. Finally, the chapter identifies the lack of sequence-based population data for African countries, whilst also demonstrating the high levels of variation present in African populations, based on currently published population data. The results of this chapter strongly argue the undertaking of a sequence-based population study for the South African population.

Prior to undertaking a sequence-based population study using the ForenSeq™ DNA Signature Prep kit, it was noted in the first few attempts at generating MPS data, that direct PCR crude buccal swab lysates that had previously given full profiles with CE-based methods performed poorly using the ForenSeq™ DNA Signature Prep kit. Therefore, a systematic optimisation study was performed to investigate and optimise the first-time success rates of crude buccal swab lysates generated using the SwabSolution™ kit (Promega Corporation, Madison, WI, USA) and the STR GO! Lysis buffer (QIAGEN, Hilden, Germany), and processed with the ForenSeq™ DNA Signature Prep kit. The results of Chapter 3 provide essential recommendations for processing crude buccal swab lysates using a direct PCR approach, as well as a modified lysate purification method that could be used in generating sequence-based population data in our laboratory, with a higher degree of confidence.

Chapter 4 describes the characterising of MPS population data for 463 South Africans with 27 autosomal STR markers utilising an optimised workflow developed in Chapter 3. It establishes the first STR sequence-based allele frequency data for two major South African population

groups. The chapter describes novel findings for both highly polymorphic and more conserved STR markers. Lastly, it provides an evaluation of the concordance levels between length-based genotypes generated using the ForenSeq™ DNA Signature Prep kit and CE-based methods, while providing an in-depth analysis of discordances. Chapters 2, 3 and 4 are written in manuscript format.

As an important step towards fulfilling ISO 17025 quality requirements, and thus facilitating implementation of MPS in our laboratory, the workflow was internally validated using control and forensic casework samples commonly encountered in forensic mortuaries such as bone, nail, teeth, blood, FFPE tissue, buccal cells on FTA cards and crude buccal swab lysates. Chapter 5 details experiments undertaken to establish whether the workflow meets the implicit criteria set through developmental validation studies, including sensitivity, accuracy, precision, repeatability, reproducibility, species specificity and stability. The chapter closes off with a concise discussion, and the demonstration of the workflow applicability to a forensic cold case from a severely decomposed, unidentified human body, where MPS data was used to contribute to generating a forensic investigative lead in a missing persons case.

The final chapter of this thesis (Chapter 6) synthesises the results from a broader perspective, offering recommendations for future studies and addressing the study's limitations and their impact on identification. The findings of this study emphasise that implementing MPS workflows in under-resourced regions requires sustainable practices and acknowledge that successful implementation is largely administrative, demanding collaborative efforts beyond the laboratory. The study identified that improving identification strategies hinges on addressing two key limitations: establishing a legal framework in South Africa for generating phenotype data to aid in forensic investigations and raising awareness and trust within

communities for submitting familial reference samples, which are crucial for assessment of their relationship to the deceased. Holistically, the successful generation of a forensic investigative lead, using an internally validated and optimised workflow, along with the use of the newly generated sequence-based population data, lead to the realisation of this study's aim, nudging South Africa one giant step closer to achieving implementation.

1.7. Ethics approval

Ethical approval to undertake this study was obtained by the University of Cape Town's (UCT) Faculty of Health Science (FHS) Human Research Ethics Committee (HREC:400/2021) (Appendix 1.2), as well as from the FHS Animal Ethics Committee (AEC) to perform species specificity experiments (Ref No: 021_010) (Appendix 1.3). The study was carried out in compliance with the Declaration of Helsinki (1964), as adapted at the 4th World Medical Association (WMA) General Assembly (Fortaleza, Brazil, October 2013).

Chapter 2: Systematic Literature Review

Systematic analysis of population studies performed with the ForenSeq™ DNA Signature Prep kit: Examining the scope and standardisation of global research efforts

2.1. Introduction

In Chapter 1, the need for sequence-based population data was mentioned as an essential requirement for the implementation of MPS methods to support CE-based protocols. Globally, several forensic laboratories have pursued sequence-based population studies to fortify this implementation [70,75,86-88]. Despite this, the exact number and identity of populations represented in the literature remain uncertain, which is what prompted the systematic review undertaken in this study, to better understand the current landscape of population representation. Furthermore, MPS has been shown to enhance allelic representation compared to CE methods, as reported in several studies [70,75,86-88]. It is hypothesised that for African population groups, the enhancement of allelic representation provided by MPS will be further exemplified [89]. Defining the breadth of the increase in allelic diversity for currently published sequence-based population studies establishes a rationale for undertaking sequence-based population studies in African and underrepresented population groups. Further to this, sizing the increase in genetic diversity may uncover insights into sample size and STR sequence nomenclature recommendations for more genetically diverse population groups.

A consequence for the lack of sequence-based population data is the delay in achieving a standardised nomenclature for characterising STR sequence variants [19]. Consequently, there is a need for assessing which naming conventions are reported in different population studies. These insights will shed light on the current stance of the forensic genetics community in working towards a consensus on the standardisation of STR sequence nomenclature. This consensus will serve as a concrete guide as to what should be implemented within a South

African context. Moreover, while studies have carried out concordance testing between CE and MPS methods, the level of agreement between studies regarding autosomal STR marker concordance is not firmly established [72,86,87,90,91]. There is a subsequent need to address the commonalities and disparities between population studies regarding concordance with CE methods.

Another challenge in fortifying the implementation of MPS is the high upfront cost. While the use of direct PCR approaches has been pursued for CE-based population data to mitigate time and cost constraints, this pursual is less evident with MPS [92-95]. Indeed, the ForenSeq™ DNA Signature Prep kit has been tested using direct PCR with FTA cards as well as crude lysates from buccal swabs [21]. Although, the number of population studies employing direct PCR approaches with the ForenSeq™ DNA Signature Prep kit is unknown.

Since the inception of MPS in forensic genetics almost a decade ago, reviews have assessed and evaluated the technology ahead of its implementation [19,55,62]. However, there has been no systematic review nor meta-analysis of collated data to date to address the gaps highlighted above. This literature review has therefore employed a systematic approach to summarise and collate data pertaining to all population genetic studies conducted using the ForenSeq™ DNA Signature Prep kit with the Miseq FGx™ system. The chapter thus aims to present a holistic and up to date account of the increase in genetic diversity obtained using MPS, to determine the proclivity of the forensic MPS community towards the use of direct PCR approaches, to pinpoint commonalities in concordance levels, and assess trends in STR sequence nomenclature. This was done to gain new insights and appreciate the potential, as well as the limitations of the technology ahead of its implementation within a South African context.

2.2. Methods

2.2.1. Search strategy

The PRISMA approach was used to guide the search strategy for this review [97,98] (Appendix 2.1). A search strategy was designed to answer the research objectives using MeSH terms. The following search terms were then entered into four databases, PubMed, Scopus, ScienceDirect and Web of Science: "ForenSeq", "MiSeq FGx", "population", "validation", "performance", "developmental validation". Search terms related to performance and validation studies were included, as it was established through an initial search that data pertaining to population groups were included in certain studies as an aspect of validation and evaluation studies. Boolean operators were used in each search query to result in a more focused search. Citations were exported into the EndNote version 21.3 reference manager software for further evaluation of records. This review covers articles published from January 2015 up until 30 June 2024. These dates were used as the ForenSeq™ DNA Signature Prep kit and MiSeq FGx™ system was released in 2015 [99].

2.2.2. Screening

Prior to evaluation of full text records for inclusion and exclusion, duplicate records were removed using EndNote's "remove duplicates" feature. The number of records retrieved from each database was recorded. Records were included if they comprised a population study using the ForenSeq™ DNA Signature Prep kit with the MiSeq FGx™ instrument, and if the study assessed A-STR, Y-STR, X-STR or iiSNV markers, or a combination of these markers. Articles were excluded if they were not relevant to the field of forensic human identification. Review, opinion and non-English articles were excluded. Once all the articles were assessed for eligibility, reference lists within the included articles were evaluated by applying the same inclusion and exclusion criteria as above until no new articles were retrieved.

2.2.3. Data collection

The full text of each included article was assessed, and data were extracted onto a standardised data collection Microsoft Excel® spreadsheet. Variables pertaining to publication details included the year of publication and journal name, while variables pertaining to the population group studied included sample size, country in which population study was performed and specific population groups. The sample type and sample preparation method used were also recorded. Data relating to concordance with CE methods for each study assessing STR data were collected. The STR sequence nomenclature formats used were also recorded. Variables pertaining to allelic gain were recorded from reported length- and sequence-based allele frequency data (**Table 2.1**)

Table 2.1: *Variables collected from population studies grouped into categories; publication details, sample preparation, population group, concordance, sequence variation and characterisation of sequence data.*

Category	**Variable**
Publication details	• Year of publication • Journal of publication • Markers assessed
Sample preparation	• Sample type • Sample preparation method used (Direct PCR or DNA extraction)
Population group	• Population group studied • Country in which study was conducted • Sample size
Concordance	• Level of concordance (%) with CE-based methods
Sequence variation and characterisation for STR data	• Number of alleles reported by length and by sequence • Length-based and sequence-based random match probability • Nomenclature format used

2.2.4. Meta-analysis

To infer statistical associations between length- and sequence-based variation reported for A-STRs, a meta-analysis was performed on a subset of studies that met additional inclusion criteria. Studies were included in the meta-analysis if allele frequency data were reported for both length- and sequence-based studies, and if they reported RMP values for both length- and sequence-based data. If a study reported allele frequency data for both length- and sequence-based alleles for A-STR markers but did not explicitly report RMP values, these were still included in the meta-analysis to perform statistical calculations on allelic gain.

To record the number of alleles reported by sequence and by length for 27 A-STR markers across different population groups, allele frequency data were used. More specifically, Microsoft® Excel "=COUNT()" and "=UNIQUE()" functions were used to count the number of unique length and sequence-based alleles for each marker and for each population group. Two studies did not report allele frequency data for the D22S1045 marker, and these were not included in statistical analyses. Additionally, the RMP was recorded for each population group. Where match statistics were reported by marker only, the product rule was applied to determine the RMP.

2.2.5. Data analysis

Collected data were collated into a Microsoft Excel® spreadsheet and statistical analysis was performed using R software (R Core Team, 2022). The RStudio package "ggplot2" and Microsoft Excel® were used for creating graphs and tables for the systematic review [100]. Heterogeneity was assessed using the "meta" package in RStudio, which determined three values pertaining to heterogeneity including Tau^2, I^2 and Cochranes Q-statistic (Appendix 2.2). To conduct the heterogeneity assessment, the sample size, length-based and sequence-based

allele counts for each population were used. These values were required to determine the effect size (*i.e.*, the ratio between length- and sequence-based allele proportions). The log of the effect size (hereafter named log Effect size) for each population group was calculated in RStudio. Thereafter, the standard error of the log Effect size was determined. A forest plot was generated in RStudio to visualise the effect size and weight of each population group, using a random effects model.

Descriptive statistics including mean, median, standard deviation and confidence intervals (CI) were determined for the number of length- and sequence-based alleles across population groups using the data collected from allele frequency tables. Histograms were created for length-based and sequence-based allele counts across population groups to assess normality of the data. A t-test was then performed to determine if the allelic gain was significant between length and sequence-based alleles. The average percentage increase in allele count was also determined for each of the 27 A-STR markers across the population groups. The decrease in RMP values from length- to sequence-based data for each population group was determined by calculating the fold-change.

Furthermore, a correlation test was performed using Pearson's correlation coefficient combined with a t-test using the "cor.test" function in RStudio (R Core Team, 2022). This was done to determine if there was a relationship between sample size and percentage increase in allele count in two scenarios; 1) across all population groups, and 2) across a merged dataset where populations were grouped into their major ancestral populations. This was done to understand the effect of sample size on allelic gain across populations and ancestral groups, and to test the sensitivity of the correlation. A significance cut-off value of 0.05 was used for all tests.

2.3. Results

2.3.1. General search results

The systematic search resulted in a total of 582 articles from the four databases searched, while 167 duplicates were removed, resulting in a total of 415 articles (**Figure 2.1**). After applying the inclusion and exclusion criteria to the abstracts, titles and methods sections, 44 articles remained. While every attempt was made to access all articles, one publication could not be accessed and was thus excluded from further analysis. Reasons for exclusion of articles are shown in Appendix 2.3. Additionally, two articles did not meet the inclusion criteria after full-text screening. Furthermore, one publication had been retracted and was therefore excluded from any further analysis. This resulted in a total of 40 included articles.

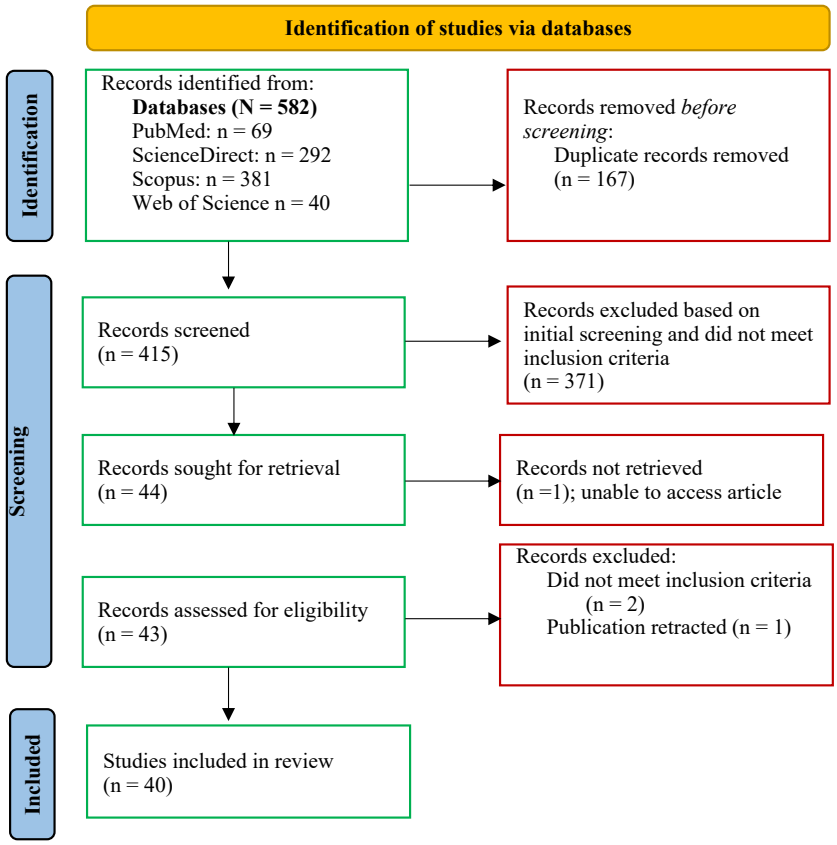

Figure 2.1: *Adapted PRISMA flow diagram depicting the number of records retrieved from four databases using three phases; 1) identification, 2) screening and 3) inclusion.*

2.3.2. General summary of literature

2.3.3.1. Country of publication

Population studies have been conducted in 16 different countries, evaluating 43 different population groups (**Table 2.2**). Many population studies utilising the ForenSeq™ DNA Signature Prep kit have been carried out in the USA (30%, n = 12/40) and China (20%, n = 8/40). No studies using the ForenSeq™ DNA Signature Prep kit have been conducted in Africa, although studies have assessed populations with individuals of African ancestry (**Figure 2.2**).

Table 2.2: *Summary of meta-data for 40 population studies included in the systematic review. The table summarises the general information such as country of study, population group, sample size, sample type and sample preparation method used.*

Country	Sample size	Population Group	Sample type	Sample preparation method	Overall concordance (%)	Reference
Centre for the study of human polymorphism (CEPH) Dataset	102	Sub-Saharan African	Blood	Not reported	99.99	[72]
	158	European				
	163	Middle East				
	200	Central South Asian				
	229	East Asian				
	28	Oceanian				
	64	Native American				
China	85	Nigerian	Blood on FTA card	Direct PCR	Not reported	[101]
	90	Hui	Blood on FTA card	Direct PCR	Not reported	[102]
	70	Torghut Mongols	Blood on FTA card	Direct PCR	100	[70]
	88	Jalaid Mongols				
	635	Northern Han Chinese	Bloodstains	Trace Evidence DNA Isolation Kit (AusBio, Yantai, China)	99.94	[87,103]
	107	Tibetan	Blood on FTA card	Direct PCR	100	[104]
	220	Eastern Han Chinese	Blood	QiaAmp DNA Blood Mini Kit	Not reported	[105]
	136	Hainan Li	Blood and Buccal swabs on FTA cards	Direct PCR	100	[106]
Denmark	363	Danish	Blood and Buccal swabs on FTA cards	EZ1 DNA Investigator Kit (QIAGEN) and direct PCR	99.74	[74]
El Salvador	391	El Salvador	Buccal cells and saliva	Prepfiler BTA kit	Not reported	[107]

France	169	French	Buccal swabs	QIAamp DNA Mini Kit (QIAGEN) or NucleoSpin® Plasma XS kit (Macherey-Nagel)	Not reported	[108]
Korea	209	Korean	Blood	Prepfiler Forensic DNA Extraction Kit	99.90	[90]
Mexico	105	Mestizos	Blood	Prepfiler BTA kit	Not reported	[109]
Norway	371	Norwegian	Blood	Salting out	99.80	[91]
Peru	172	Chachapoya	Saliva	Modified high-salt DNA extraction protocol	Not reported	[75]
	25	Awajún				
	13	Wampís				
	9	Huancas				
	14	Cajamarca				
Philippines	143	Filipino	Blood on FTA card	Direct PCR	100%	[110]
Qatar	150	Qatari	Buccal swabs	QIAGEN Mini Kit	Not reported	[111-113]
Saudi Arabia	89	Arab	Buccal swabs and saliva	QiaAmp DNA Mini kit	Not reported	[114]
Spain	88	Spanish Roma	Saliva	Phenol-chloroform extraction	Not reported	[115]
	143	Catalans				
Thailand	182	Thai	Blood on FTA card	Direct PCR	76.47	[88]
United Kingdom	200	White-British	Buccal swabs	Chelex and DNA Investigator	99.98	[86,116-118]
	200	British - Chinese				
	200	South East Asian				
	201	West African				
	206	Northeast African				
United States of America	342	African American	Blood	QiaAmp DNA Mini Kit	99.97	[73,119,120], [121-127]
	361	Caucasian				
	97	East Asian				
	236	Hispanic				
	62	Yavapai Native American	Blood	QiaAmp DNA Blood Mini Kit	Not reported	[71,128]
Vietnam	148	Kinh	Blood on FTA card	Chelex	Not reported	[129]

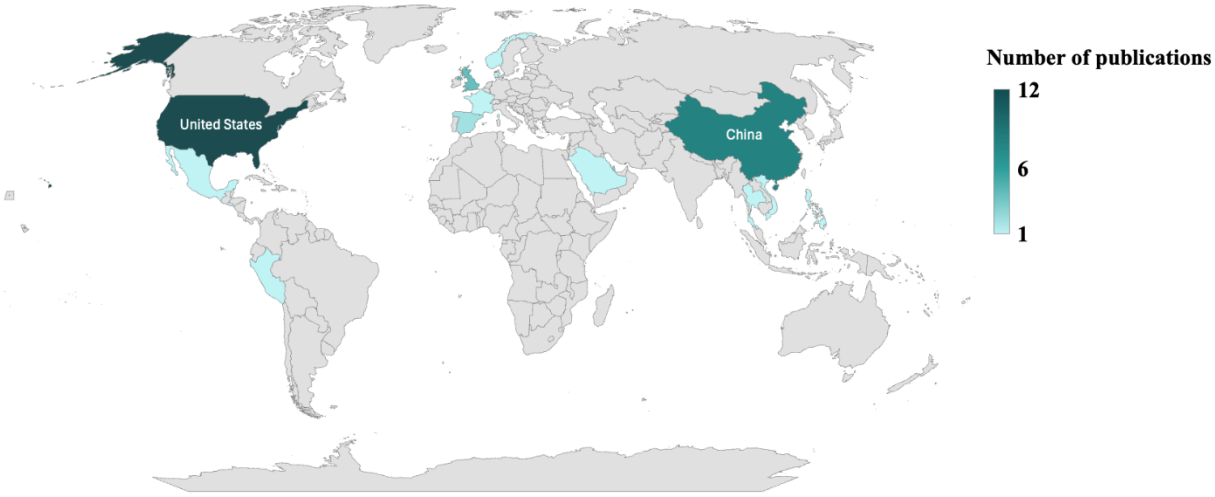

Figure 2.2: *The map illustrates the counts of population studies published globally. As indicated by the gradient key, dark green sections represent a higher number of population studies conducted in that region, whereas light green sections represent fewer population studies conducted in that region. Grey areas represent countries where no population studies have been conducted using the ForenSeq™ DNA Signature Prep kit.*

2.3.3.2.Classification of population studies according to income grouping

Among the countries that have published population data, 65% of studies (n = 26/40) have been conducted in high-income countries (HIC), 27.50% in upper-to-middle income countries (UMIC) and the least (7.50%, 3/40) in low-to-middle income countries (LMIC), and these were classified according to the World Bank Group for the 2024-2025 period (**Figure 2.3**).

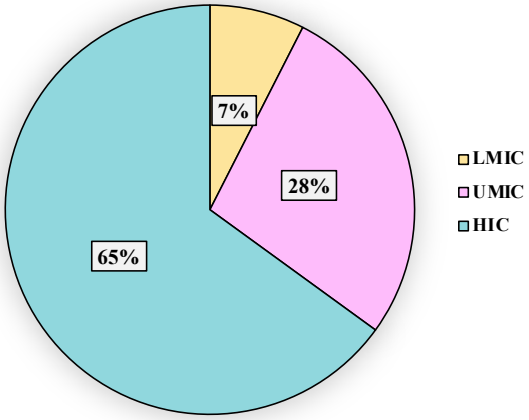

Figure 2.3: *Pie chart showing income category of countries that have published population data with the ForenSeq™ DNA Signature Prep kit. LMIC = Low to middle income, UMIC = Upper to middle income and HIC = High income.*

2.3.3.3. Year of publication and journal

The articles included were published over an eight-year period, with the first articles being published in 2016, after the introduction of the MiSeq FGx™ system (**Figure 2.4**). In 2018, most population studies were published (n = 7), with less articles published each subsequent year. The Forensic Science International: Genetics journal contained the highest number of population studies conducted with the ForenSeq™ DNA Signature Prep kit (Appendix 2.4).

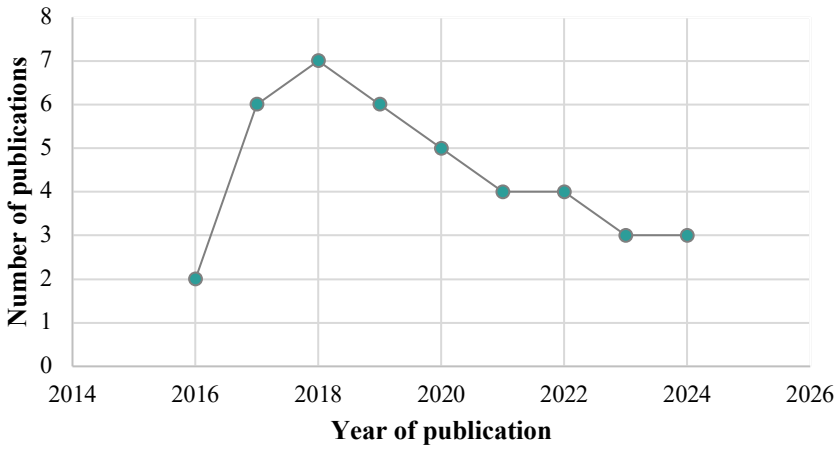

Figure 2.4: *Line graph showing the number of studies included in the systematic review that have been published between 2016 and 2024.*

2.3.4. Use of direct PCR in MPS population studies

Out of 40 studies, sample preparation methods were recorded for 23 population datasets. One study did not report whether direct PCR or conventional DNA extraction was used in sample preparation methods [72]. One study indicated that both direct PCR and DNA extraction was used during sample preparation [74]. DNA extraction (60.87%; 14/23) was more commonly employed in sample preparation methods compared to direct PCR methods (34.78%; 8/23) (**Figure 2.5**). Within DNA extraction techniques, spin-column purification was the most frequently used, while phenol-chloroform methods were the least utilised. When studies employed direct PCR, FTA cards were the only choice of substrate used.

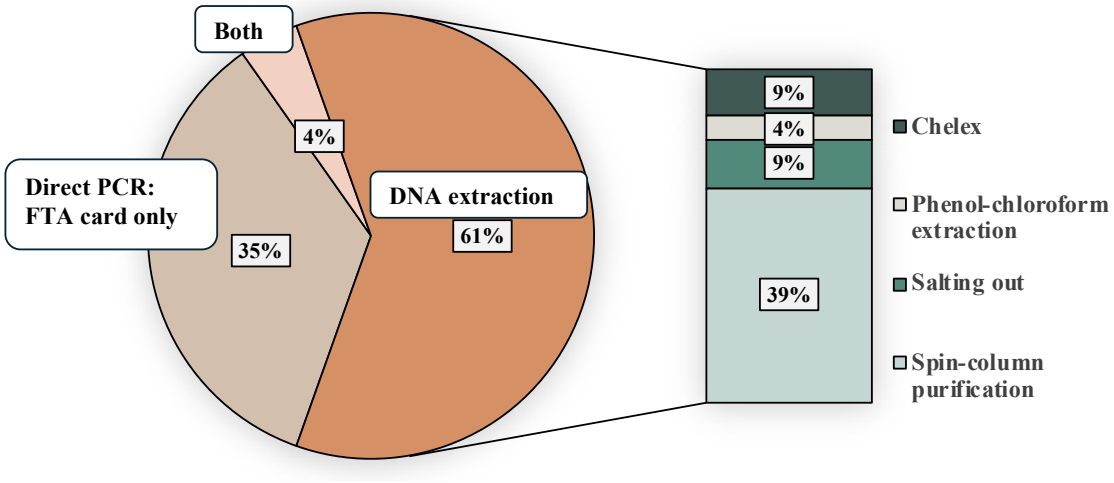

Figure 2.5: *Pie and bar chart illustrating percentage of studies employing either DNA extraction methods or direct PCR methods. Within DNA extraction methods used, a connecting bar chart illustrates the percentages of different extraction techniques used.*

2.3.5. Concordance with CE data

Concordance between length- and sequence-based alleles was only reported for 14 population studies that assessed A-, Y- and or X-STR markers (**Table 2.2**). Thirteen population studies reported a concordance level of above 99%, with one study reporting a 76.47% concordance level. The D7S820 marker was the most frequently observed marker causing discordance, followed by DXS10074 and D5S818 (**Table 2.3**).

The most reported reason for discordance was an insertion or deletion present in the upstream or downstream flanking region of STR markers. This type of discordance was investigated by designing primers extending the primer binding sequences of the ForenSeq™ DNA Signature Prep kit, followed by re-sequencing the sample. Where re-sequencing was not performed, raw data were re-analysed to include flanking region sequences through alignment with a reference genome.

A second common reason for discordance was allele dropout of the second, larger allele of a heterozygote. However, there were discrepancies in reporting allele dropout as discordance, as some studies did not classify allele dropout as a discordance, but rather as “allele ambiguity”. Other reasons reported for discordances included the presence of null alleles, bioinformatic configurations, microsatellite instability and miscounting of repeats by different analysis software tools (**Table 2.3**).

Table 2.3: *The table depicts a summary of observed discordances at STR markers reported by population studies using the ForenSeq™ DNA Signature Prep kit, along with reported reasons for the discordance.*

Discordant marker(s)	CE allele or genotype	MPS allele	Reported explanation for discordance	Article
D5S818	9,11	11,11	The discordance resulted from null allele caused by a SNV in the primer binding site	[88]
	10,12	10,10	Due to larger allele dropout in heterozygote (larger second allele fell below UAS threshold)	[72]
	9,13	9,9		
D7S820	10.3	11	Caused by an upstream single nucleotide deletion	[73]
	6.3	7	Discordance caused by a rare 1 bp deletion in the flanking region of the 7 allele, causing amplicon to be 1 bp shorter (*i.e.*, 6.3)	[86,116]
	7.3	8	Discordance was due to deletion on 5' side of repeat region (23 nucleotides upstream)	[72]
	9	9.1		
	10.3	11		
	11.3	12		
	7.3	8	Discordance was 1 bp deletion in flanking region	[87]
D12S391	Not reported	Not reported	Not reported	[90]
D13S317	GlobalFiler: 8, (OL - 28.2) PowerPlex: 8, 10, (OL = 28.2)	8, 28.2	MPS resolved discordance between two CE kits. MPS was concordant with GlobalFiler.	[86]
D21S11	29,30	29,29	Due to larger allele dropout in heterozygote (larger second allele fell below UAS threshold)	[72]
D22S1045	15.1	15	Caused by indel in flanking region	[73]
	Not reported	Not reported	Due to locus drop-outs	[91]
	14, 19	14, 14	Due to larger allele dropout in heterozygote (larger second allele fell below UAS threshold)	[87]
	17, 20	17, 17		
	16, 17	16, 16		
Penta D	Not reported	Not reported	Not reported	[90]
	9,11	9,9	A G > A change occurred before the repeat unit, resulting in a homopolymer of 12 A bases, contributing to sequencing or alignment issues	[86]
	13.4	14	Discordance caused by presence of a 1 bp deletion in the 3′ flanking region	[91]
Penta E	21,27	21,21	Not reported	[90]
	Not reported	Not reported	Due to locus drop-outs	[91]
SE33	14	13	Caused by a 4 bp deletion (TTTT) in the flanking region	[126]
	15	15		
	16	15		
	17	16		
	18	17		
	25.2	24.2	Caused by a 4 bp deletion (TCTT) in the flanking region	
	26.2	25.2		
	27.2	26.2		
	29.2	28.2		
	29.2	28.3	Caused by a 3 bp deletion (ACA) in the flanking region.	
DYS570	17,18	19	Not reported	[88]
DYS389 II	29	30	Not reported	
DYS3895a-b	Not reported	19	Not investigated, but reported that discordance could be due to poor intra-locus balance	[90]
DYF3875I	35,36,39	36,39	Reasons reported are due to flanking region SNV associated with the 35-allele interfering with primer binding or microsatellite instability	[72]
	35,37,38	37,38		
DXS10135	Not reported	Not reported	Allele dropout of larger allele with MPS sample level report and flanking region report. Sanger sequencing confirmed concordance.	[125]
DXS10103	19	16,19	Not reported	

	19	19.2	The discordance is caused by a dinucleotide repeat in the 5' flank of DXS10103	
DXS10074	No signal	7	It was concluded that the discordance was caused by a CE null allele	
	Not reported	Not reported	Discordance was caused by a 2 bp insertion in a dinucleotide repeat present in the 5' flanking region.	
	16	15,16	Not reported	
	14	15	A 4 bp deletion in flanking region	[87]
	16.3	17	A 1 bp deletion in flanking region	
HPRTB	Not reported	Not reported	Discordance was caused by a 4 bp deletion present in the 5' flanking region.	[125]
DXS7132	Three alleles present (not named)	No alleles present	The three alleles have a SNV in common that results in a bioinformatic null allele in the UAS sample level report.	
DXS58378	12	10,12	Not reported	

2.3.6. STR Sequence nomenclature formats

The sequence-based STR nomenclatures used had varied between four different formats. There were five (n = 5/40) studies that were linked with regard to the population groups studied, although different marker sets were analysed (*i.e.*, either A-, Y- or X-STRs) and these were grouped under one population dataset (**Table 2.4**).

Table 2.4*: STR sequence nomenclature formats used by population studies conducted with the ForenSeq™ DNA Signature Prep kit. The CSF1PO allele 9 is used as an example to visualise the nomenclature format. The population dataset contains the number of linked studies where A-, Y- and X-STRs have been published in separate publications but use the same population dataset.*

Nomenclature used	Example using CSF1PO, allele 9	Number of sample or population datasets
Bracketed repeat with defined coordinates	CSF1PO [CE 9]-GRCh38-chr5:150076318-150076389 [ATCT]9	10
Bracketed repeat only	[ATCT]9	8
Short designator	9a (first observation of allele 9 assigned an “a”, second observation assigned “b”)	4
Full UAS sequence string	CTTCCTATCTATCTATCTATCTATCTAT CTATCTATCTATCTAATCTATCTATCTT	4
Total		**26**

Additionally, three (n = 3/40) studies did not include sequence-based allele frequency data in their results or supplementary data, although they were included as they had conducted population studies using the ForenSeq™ DNA Signature Prep kit and had therefore met inclusion criteria. Furthermore, six (n = 6/40) studies did not assess STR data but rather iiSNV data and did not use a specific nomenclature format. Sequence-based STR nomenclature formats were thus reported by 26 studies and this number was used to determine the percentage of articles using a specific format.

The 26 studies included the following nomenclature formats; 1) the bracketed repeat region only, 2) the bracketed repeat region with defined co-ordinates according to Parson *et al*., 2016, 3) a short designator system using a combination of numbers (for allele length) and letters (for allele sequence) and lastly 4) the full sequence string as reported by the ForenSeq^{TM} UAS software (**Table 2.4**). The two most frequently used formats included the bracketed repeat region with defined coordinates, as recommended by Parson *et al*., 2016 (n = 10/26) and the bracketed repeat region only (n = 8/26). The least used sequence nomenclatures were the short designator (n = 4/26) method and the full UAS sequence string (n = 4/26).

2.3.7. Meta analysis

2.3.7.1. Assessment of heterogeneity

A total of 40 articles were included in this systematic review, and a subset of 20 met additional inclusion and exclusion criteria for conducting a meta-analysis on allelic diversity statistics. Allele counts by length and sequence for each population group are shown in Appendix 2.5. The 20 articles that met the additional inclusion criteria for meta-analysis encompassed 35 different population groups.

Heterogeneity analysis revealed a Tau^2 estimate of 0.0015 (95% CI [0.0000; 0.0085]), and an I^2 value of 23.5% (p = 0.1089) (**Figure 2.6**). Both values indicate low heterogeneity across population groups. The I^2 value suggests that only 23.5% of the variability in effect sizes are due to heterogeneity. While the I^2 value describes the proportion of total variation due to heterogeneity, the Tau^2 estimate quantifies the magnitude of this variation (Tau^2 = 0.0015), which is ultimately very small (Appendix 2.6).

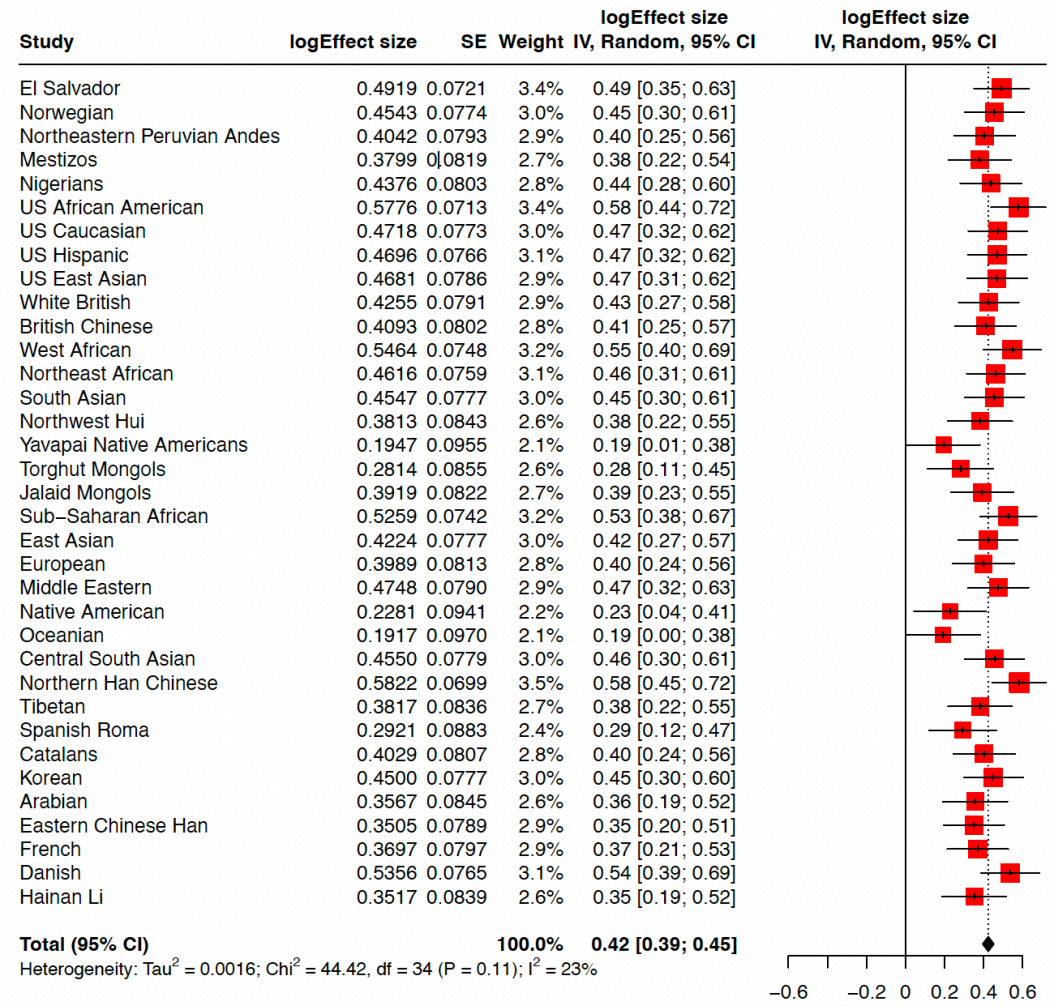

***Figure 2.6**: Forest plot showing log Effect sizes (i.e., log transformed ratio between proportions of length- and sequence-based allele counts), weight of each population group studied (%), 95% CI, and visualisation hereof. The dotted vertical line represents the value of the overall effect size (log of effect size), while the solid vertical line represents the line of "no effect". Pooled estimate values are shown in bold. Heterogeneity statistics are shown in the bottom left of the plot (Tau², and I²).*

From the heterogeneity assessment, it was noted that the population groups with the lowest weighting were the Oceanian and Yavapai Native American groups. These population group effect sizes had less influence on the pooled estimate due to their smaller population sizes (Oceanian: n = 28, Yavapai Native American: n = 62). Population groups with larger weights such as the El Salvador, US African American and Northen Han Chinese population groups

had the largest sample sizes (> 300) and thus contributed more to the overall pooled estimate and may be more representative of the true effect size across population groups.

2.3.7.2. Sensitivity analysis: Relationship between allelic gain and sample size

Assessment of whether there was a relationship between sample size and allelic gain was performed to understand the contribution sample size had in the increased variation in some populations. It was found that there was a significant positive correlation between sample size and percentage increase in allele counts (r = 0.66; p = 1.39E-05) (**Figure 2.7**). However, in assessing the robustness of this analysis, when merging data according to ancestral population group (African, East Asian, European, Latin American, Middle Eastern, Native American, Oceanian, South Asian), no correlation between sample size and increase in percentage allele counts were observed (**Figure 2.7**). That is, sample size may not have played a significant role in determining which *ancestral* population groups showed the highest percentage gains in alleles through inclusion of sequence data (r = 0.365, p = 0.373).

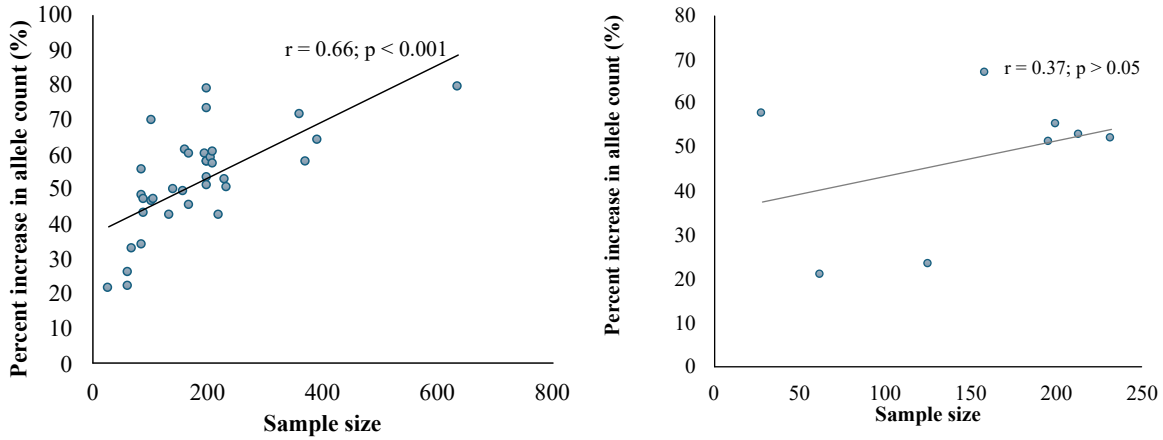

Figure 2.7: Left: Scatter plot with line of best fit to determine the relationship between sample size and percentage allelic gain across all population groups. Right: Scatter plot with line of best fit to determine the relationship between sample size and percentage allelic gain for a dataset merged by ancestral population group. Both plots indicate Pearson's correlation coefficient represented by 'r' and the p-value obtained from the two-tailed t-test.

2.3.7.3. Increase in variation by ancestral population group

To determine which ancestral groups contributed the most towards an increase in allelic variation, population groups were collated by their ancestral population. The average percentage increase in variation was then calculated across population groups within an ancestral group. The average sample size across ancestral population groups was 152 ± 74 (**Figure 2.8**). More so, the highest increase in allele counts were observed for the African ancestral population group with an increase of 66.46%, despite not having the highest sample size (average of 159 ± 60). The ancestral group with the highest sample size was Latin America (232 ± 119), with a 55.39% increase in allele count (Appendix 2.7).

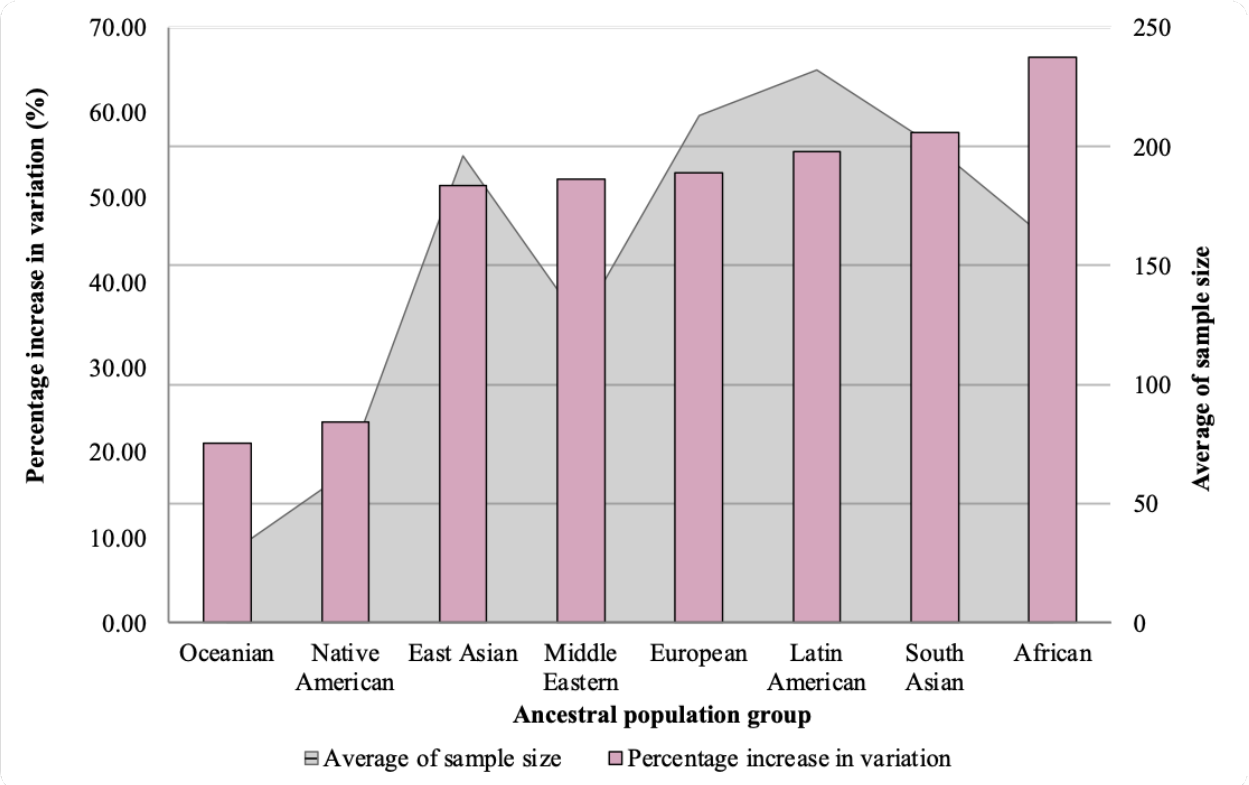

Figure 2.8: *Percentage increase (%) in allele counts from length to sequence-based alleles across eight ancestral population groups. Populations were grouped into their major ancestral population groups.*

2.3.7.4. Overall increase in allele counts

When the data for all markers and population groups were merged, the data for length- and sequence-based allele counts were found to be normally distributed (Appendix 2.8). A t-test revealed that the mean number of alleles reported for sequence-based alleles across 27 A-STR markers was higher than length-based alleles across 35 population groups (p = 5.987e-13, 95% CI: [111 – 166]) (**Figure 2.9**). More specifically, the mean number of alleles increased from 260 ± 28 length-based alleles to 398 ± 75 sequence-based alleles (53.08% increase).

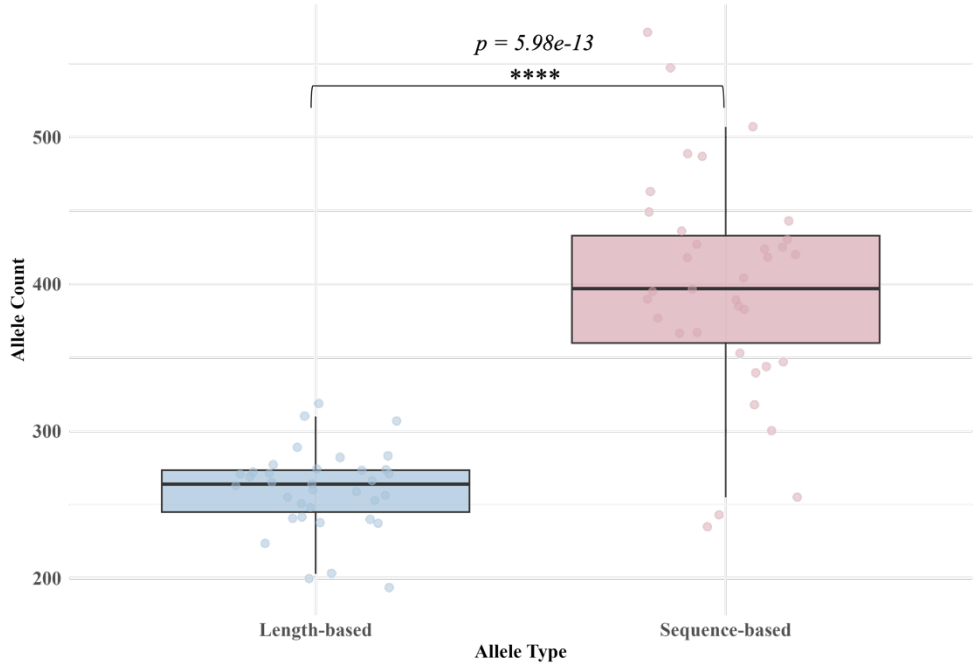

Figure 2.9: *The box-and-whisker plot shows the difference in the average number of length- (blue) and sequence- (pink) based alleles reported across the 20 population studies.*

2.3.7.5. Increase in allelic gain by marker

As data were collected per marker across all studies, it was possible to gain insight into the percentage increase in allelic gain by marker type. The most conserved marker across all 20 studies (and 35 population groups) was the TPOX marker, with no increase in allele counts across all population groups (**Figure 2.10**). The markers showing the largest percentage increase in allele counts were D21S11 (140% increase, 95% CI [123 – 157]), D2S1338 (172% increase, 95% CI [145 – 199]) and D12S391 (203% increase, 95% CI [183 – 223]).

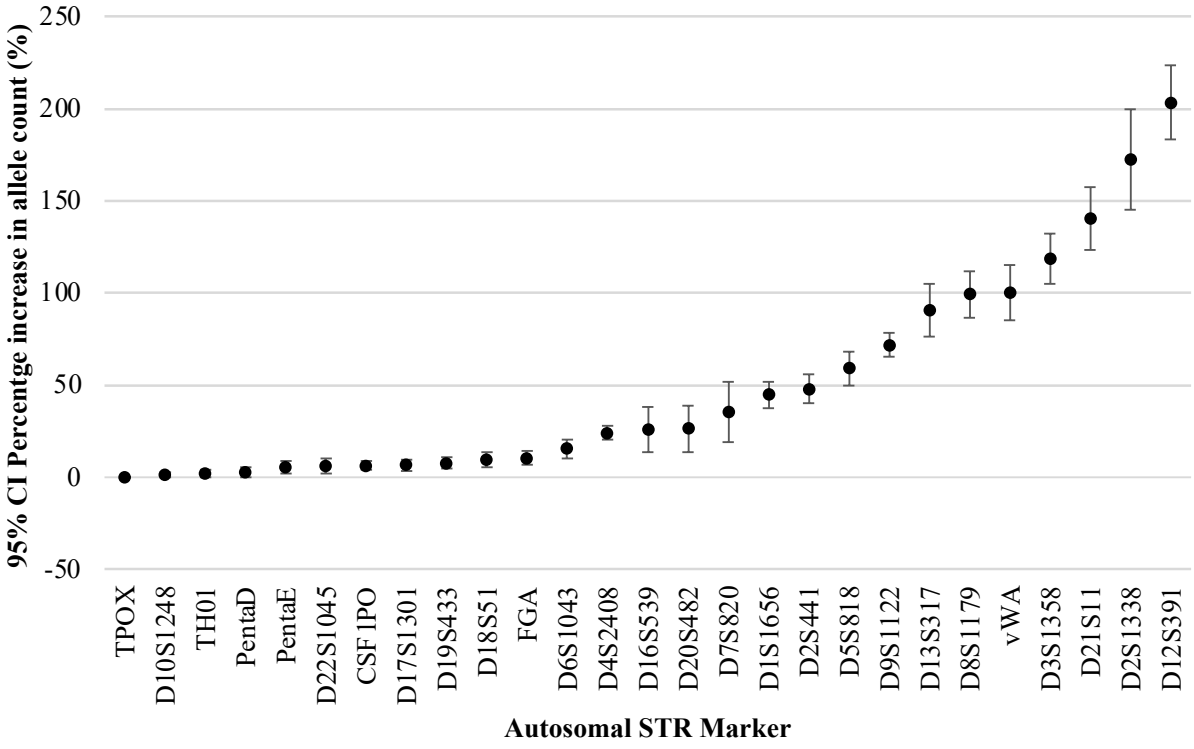

Figure 2.10: *Percentage increase in allele count across 35 population groups (20 studies) for 27 A-STR markers. The 95% CI is represented by the vertical line running through each data point. Markers are ordered by percentage increase in ascending order.*

2.3.7.6. Assessment of RMP

A subset of 17 studies included in the meta-analysis met additional inclusion criteria in that the RMP was reported for both length- and sequence-based data. These 17 studies reported RMP statistics for 26 different population groups. As expected, the RMP for sequenced-based data was lower than that for length-based data across all 26 populations and 17 studies. Of particular importance are the studies showing the largest decrease in RMP, which were those evaluating data from African population groups including the Nigerian population (2.41E+03 times lower) and US African American (6.52E+05 times lower) (**Table 2.5**).

Table 2.5*: Meta-data summary of length-based versus sequence-based variation, as determined through analysis of allele counts and random match probability. The table is sorted according to fold change in RMP in descending order.*

Population group	Sample size	Number of alleles by length	Number of alleles by sequence	Effective increase in number of alleles (%)	Length-based random match probability (A-STRs)	Sequence-based random match probability (A-STRs)	Fold change in RMP	Reference
US African American	200	307	547	78.18	8.54E-34	1.31E-39	651908	[73]
Nigerians	85	255	395	54.9	2.17E-32	9.02E-38	240576	[101]
Jalaid Mongols	88	248	367	47.98	1.10E-31	6.30E-36	17460	[70]
US Caucasian	210	272	436	60.29	6.28E-32	3.63E-36	17300	[58]
West African	201	282	487	72.7	4.39E-31	2.78E-35	15784	[86,116]
Northeast African	206	283	449	58.66	4.51E-31	3.04E-35	14836	[86,116]
Danish	363	271	463	70.85	4.8E−31	3.3E−35	14545	[74]
US Hispanic	198	277	443	59.93	1.51E-31	1.23E-35	12276	[73]
Norwegian	371	273	430	57.51	3.02E-35	3.09E-39	9773	[91]
Arabian	89	238	340	42.86	2.62E-30	3.49E-34	7507	[114]
US East Asian	169	263	420	59.7	6.37E-32	8.66E-36	7356	[73]
Tibetan	107	241	353	46.47	1.94E−30	3.21E−34	6038	[104]
Catalans	143	256	383	49.61	1.11E-31	1.92E-35	5762	[115]
Torghut Mongols	70	240	318	32.5	1.60E-31	3.41E-35	4691	[70]

El Salvador	391	310	507	63.55	6.79E-31	1.47E- 34	4619	[107]
Northern Han	635	319	571	79	1.52E-31	3.29E-35	4616	[87,103]
Eastern Han	220	274	389	41.97	6.05E-31	1.53E-34	3949	[105]
South Asian	200	271	427	57.56	5.07E-31	2.89E-34	1754	[86,116]
Spanish Roma	88	224	300	33.93	2.93E-30	1.87E-33	1570	[115]
French	169	266	385	44.74	8.99E-31	7.12E-34	1262	[108]
Northeastern Peruvian Andes	233	265	397	49.81	9.85E-28	9.64E-31	1022	[75]
Yavapai	62	200	243	21.5	2.37E-26	2.81E-29	843	[71,128]
White British	200	264	404	53.03	2.08E-30	2.82E-33	738	[86,116]
Hainan Li	136	242	344	42.15	9.07E-32	1.39E-34	653	[106]
Northwest Hui	90	237	347	46.41	3.06E-30	8.01E-33	382	[102]
British Chinese	200	259	390	50.58	7.31E-30	2.40E-31	30	[86,116]
Oceanian	28	194	235	21.13	Not reported	Not reported	Not reported	[72]
Native American	64	203	255	25.62	Not reported	Not reported	Not reported	[72]
Mestizos	105	251	367	46.22	Not reported	Not reported	Not reported	[109]
European	158	253	377	49.01	Not reported	Not reported	Not reported	[72]
East Asian	229	274	418	52.55	Not reported	Not reported	Not reported	[72]
Korean	209	271	425	56.83	Not reported	Not reported	Not reported	[90]
Central South Asian	200	269	424	57.62	Not reported	Not reported	Not reported	[72]
Middle Eastern	163	260	418	60.77	Not reported	Not reported	Not reported	[72]
Sub-Saharan African	102	295	489	65.76	Not reported	Not reported	Not reported	[72]

2.4. Discussion

In this systematic review, we reviewed population studies conducted globally to gain insight into sequence-based population data, to identify populations that were not yet adequately represented and to understand the current consensus, standardisations and gaps within a global dataset to inform population studies in South Africa. The review therefore focused on: 1) underrepresentation of African population groups in published sequence data, 2) the breadth of sequence variation across population groups, and more specifically, the extent of genetic variation in African populations compared to other regions, 3) the STR nomenclature formats used, 4) concordance and discordances observed, and 5) the use of direct PCR as an approach to mitigate time and cost constraints in LMICs. The findings of this review showed the absence of population data from African regions but has also revealed that much of the variation carried by sequence-based data is held in population groups with African ancestry. Unpacking the variation that exists in currently published population data has also revealed commonalities and disparities in concordance data as well as the levels of informativeness of certain STR markers which can be leveraged by forensic laboratories performing population studies with the ForenSeq™ DNA Signature Prep kit.

2.4.1. Exploration of reasons for underrepresentation of sequence data in low to middle income countries

Before diving into the increase in allelic variation that comes with MPS, it is important to first capture the magnitude of sequence-based population data that currently exists, specifically those generated using the ForenSeq™ DNA Signature Prep kit. Since the release of the first forensically validated MPS system in 2015, most population studies that made use of the ForenSeq™ DNA Signature Prep kit were published in 2018 (**Figure 2.4**). This highlights the inclination of the forensic community to consider more advanced methods of DNA analysis

when these methods are tailored for forensic use. This is also supported by the fact that sequence-based population studies have been performed in both developed and developing countries, despite the upfront costs for MPS methods being more expensive than conventional DNA profiling methods. The European MPS survey by Alonso *et al*., 2017 alluded to in Chapter 1 also supports the idea that forensic laboratories are certainly leaning towards implementation of MPS workflows, as the study highlighted that 52% of participating laboratories had already purchased an MPS instrument for forensic applications only two years after the release of the MiSeq FGx™ system [67].

Global population data from a wide variety of population groups are required to fully characterise STR sequence variation [73]. Based on the findings of this systematic review, this goal appears to be increasingly elusive, as the number of population studies conducted using the ForenSeq™ DNA Signature Prep kit has declined in the last six years (**Figure 2.4**). However, this finding may be biased, due to the limited inclusion criteria which only accounted for population studies carried out using the ForenSeq™ DNA Signature Prep kit. Since the release of the first forensically validated MiSeq FGx™ system used with the ForenSeq™ DNA Signature Prep kit, many companies have followed suit in the development of kits and MPS systems that target many of the same markers and address the same CE limitations [19]. Population studies conducted using these kits were not accounted for in this review, and inclusion of these studies may put forward a different conclusion, *i.e*., the forensic community are indeed making efforts to explore more sophisticated methods to address CE limitations.

Contrastingly, this review revealed that many of the sequence-based population studies conducted using the ForenSeq™ DNA Signature Prep kit have been performed in HICs and

UMICs, with less studies carried out in LMICs, and none from the Southern African Development Community (SADC) region (**Figure 2.2** and **Figure 2.3**). This finding is not unexpected given the cost demands of the entire workflow. Reviews on genomic studies outside of forensics have published similar conclusions, with genomic data being biased toward the Global North [130-132]. Global genetic diversity measures are thus represented by these skewed data. Herewith, a more representative dataset is required to account for the true global variation.

The European MPS survey mentioned above also highlighted that cost and lack of funding was the primary reason for laboratories not purchasing an MPS instrument, and it is presumable that this is even more true for LMICs [67]. Using MPS methods for generating DNA profiles are at first glance more costly than conventional DNA profiling methods, which poses an obstacle for resource-limited laboratories. Although from a zoomed-out perspective, a single MPS run with a batch of 32 samples (multiplexing between 173-231 markers using the ForenSeq™ DNA Signature Prep kit) is effectively comparable with a single CE run with the same number of samples (multiplexing 16-35 markers) [133].

2.4.2. Mitigation of time and cost: a warrant for the use of direct PCR in generating sequence-based population data

Sequencing methods have also become more affordable since the sequencing of the first human genome, with this reduction in cost being expected to decrease in the years to come [134,135]. Recent efforts by the Human Hereditary Health in Africa (H3Africa) Consortium have driven the development of genomics in Africa, yet none of these efforts have been leveraged or geared toward forensic genomic applications [136]. Therefore, given the challenges faced in implementing an MPS system even in HICs and UMICs, it is incumbent upon forensic

laboratories to make efforts to mitigate costs and resources, especially when generating large-scale MPS data. One way to achieve this is using a direct PCR approach for generating sequence-based population data. In this review, most population studies used DNA extraction, followed by quantification and dilution prior to library preparation. The use of direct PCR was less common and was only applied with blood spots on FTA cards (**Figure 2.5**). The ForenSeq™ DNA Signature Prep kit has been validated for both FTA cards and crude buccal swab lysates; however, no study utilised crude buccal swab lysates for generating population data.

The success rates of crude buccal swab lysates with MPS kits are poorly understood and warrant further research to garner trust and confidence before its wide-scale usage. This is especially relevant for laboratories who have made use of crude buccal swab lysates for generating CE population data, as these samples would likely be available for sequence-based population studies. However, the crude buccal swab lysates need to be chemically compatible with MPS kits to generate high first-time success rates and mitigate the potential for costly repeat sequencing (in-house data). It is therefore essential to consider optimisation of direct PCR methods to mitigate time and cost barriers, especially for underrepresented SADC regions, where resources are seldom allocated toward improving forensic human identification strategies [137].

2.4.3. Consequences for the absence of population genetic data in African countries (underrepresented regions)

One of the most important findings of the review, in the context of this study, was the underrepresentation of African countries in the literature on forensically relevant sequence data. No MPS-based population study has been carried out in any African country. Without

considering the sequence variation observed in population groups, as depicted in the meta-analysis component of this review, it is already widely accepted that much of the genetic variation in humans is held within African individuals and the African diaspora [138]. More so, Africa is regarded as a region of vast genetic diversity, as well as rich cultural, linguistic, and phenotypic diversity [139]. However, this variation is yet to be fully characterised and harnessed from a forensic genetic perspective.

The high genetic diversity in African individuals is attributed to the continent's status as the origin of anatomically modern humans around 200,000 years ago, allowing ample time to accumulate diversity. Additionally, gene flow through migration and inter-marriage has further enriched this genetic variation [27,140]. In contrast, when humans migrated out of Africa (assumed under the Out of Africa hypothesis), the founding populations were relatively small, carrying only a subset of the genetic diversity found in Africa [141,142].

Considering this substantial genetic diversity, African populations still represent only the minority in this review (**Table 2.2**), a finding that is sadly not unique to this study [131,140,143]. Studies have certainly aimed to be more inclusive in the populations studied, and although this change has been slowly taken up in studies pertaining to clinical interventions, drug discovery and historical adaptations, this is once again yet to spill over into the field of forensic genetics [134].

The populations with African ancestry included in this review consisted of individuals residing in the USA, the UK and China [73,86,101,116]. The CEPH panel included individuals from Sub-Saharan Africa, these were from multiple different population groups with sample sizes

below 25 [72]. Moreso, population genetic data reporting guidelines recommend that there should be a minimum of 50 samples per population group for MPS data [69]. Furthermore, although there are sequence data available for African Americans and Africans residing in the UK and China, these data do not fully represent African populations. This is because these are assumed to be individuals who descended from Africans that have migrated out of Africa, and only carry a subset of the genetic diversity found in African populations that have not migrated [131].

Implications for the absence of, or weak representation of certain population groups include inaccurate representation, specifically with regards to ancestry, hair and eye colour predictions made using the ForenSeq^{TM} DNA Signature Prep kit. Accurate representation may require reference data from African and other underrepresented population groups to ensure the reliability of these predictive models. Although, it may be that the models used for prediction are sufficient since they use SNV's in biological processes that determine phenotype that are independent of ancestry. Considering this, conducting sequence-based population studies in regions with rich genetic diversity could play an important role in generating more informative investigative leads for human identification purposes. From an ethical perspective, the global disparity in forensic data representation might exacerbate existing inequalities, where African populations are underrepresented in scientific research. This may potentially lead to mistrust or resistance toward participating in new forensic DNA projects.

Inclusion of African and highly diverse population groups in sequencing studies may enable the capturing of a wider range of sequence variation, lead to the development of novel markers, the improvement of STR nomenclature reporting guidelines, and a more fitting sample size

requirement for the establishment of population databases [19]. Furthermore, sequencing forensically relevant markers in African populations will provide a more accurate and comprehensive overview of the diversity that currently exists but is yet to be explored.

2.4.4. Gain in allelic diversity in African population groups

The population studies included in this review reported an effective increase in the number of alleles detected through sequence-based methods, compared with length-based alleles alone (**Figure 2.9**). This increase was due to the sequence variation present within the repeat motif of each STR, which is not captured through allele sizing (CE methods). Across all studies, markers showing an increase in the number of alleles through inclusion of sequence data also showed a notable decrease in RMP values (**Table 2.5**). This confirms the swell of information carried by inclusion of sequence data.

Indeed, the populations with African ancestry included in this review showcased the highest level of variation with respect to both increase in allele counts but also reduction in RMP values, when compared to non-African population groups (**Figure 2.8** and **Table 2.5**). These findings motivate for carrying out sequence-based population studies on African individuals *living in Africa*, especially given that these data are mainly extrapolated from population studies with individuals of African ancestry residing in non-African countries, and do not fully represent the vast level of variation present in African populations [131]. More importantly, this finding held true even without the effect of sample size (**Figure 2.8**). Non-African populations exhibited less variation with higher sample sizes than African population groups (**Table 2.5**). With the added variation observed in individuals of African ancestry, it is expected that additional novel sequences will be observed, especially in polymorphic markers, and sample size guidelines may be adapted to this variation. [86].

2.4.5. Commonalities in marker informativeness

This review found that markers such as D12S391, D2S1338 and D21S11, among others, provided a breadth of variation that would likely not be captured in the minimum required sample size of 50 samples per population group, as mentioned in Devesse *et al*., 2020 (**Figure 2.10**), and this is even more pertinent when considering the expected diversity in African populations [86]. What has thus not yet been established is how often new alleles are observed in markers in more genetically diverse populations. With this in mind, conserved markers *do* have lower novel allele observation rates than highly polymorphic markers when considering a sample size of ±200, as seen in the White-British and British-Chinese populations, but it cannot be assumed that additional novel alleles will not be observed at higher rates in more polymorphic markers in more diverse and underrepresented populations [86,117].

The highly conserved (TPOX, CSF1PO, TH01) and highly polymorphic (D2S1338, D21S11, D12S391) nature of markers observed in this review are based on what has been observed in non-African populations and this further motivates for the conducting of sequence-based population studies in Africa (**Figure 2.10**). Similarly, Novroski *et al*., 2016 brought attention to the fact that polymorphic loci showing a large increase in the effective number of alleles are also influenced by limited sequence-based allele data available for those markers [73]. The same could also be true for more conserved markers when there is a lack of sequence data available from both minority and diversity-rich population groups. Conducting sequence-based studies in highly diverse populations, such as in Africa, can thus provide comprehensive coverage of diversity to inform the inclusion of new markers in MPS kits.

2.4.6. Concordance between CE and MPS

The onset of MPS has introduced a torrent of variation to the forensic genetics field, as highlighted in the previous section. However, it has also triggered the onset of discordance issues due to the presence of insertions, deletions and substitutions in flanking regions (Table 2.3). Concordance with CE is a crucial element to consider in the implementation of MPS methods into forensic laboratories as CE profiling methods are still the gold standard for forensic DNA laboratories [144]. Discordances, even between different CE kits is not a newfound concern, and as with the development of new STR kits, there is always the possibility of finding discordances [145]. Resolving and reporting these discordances are thus essential for future studies carrying out population studies.

A limitation in the reporting of sequence-based population data is that many studies do not report concordance data, as seen in this review (**Table 2.2**), however there was a consensus regarding high concordance rates achieved between CE and MPS alleles. All population studies included in this review, except one, reported concordance levels exceeding 99%, which aligns with concordance levels reported from the developmental validation of the ForenSeq™ DNA Signature Prep kit, as well as concordance levels between different CE-based kits [20,146].

There were certainly commonalities amongst the discordances that have been evaluated as part of this review. The main reasons for discordances revealed in this review have been attributed to the presence of flanking region indels and differences in primer binding sites (**Table 2.3**). Fortunately, it is possible to resolve the discordance if both the length-based allele and the full sequence string or the flanking sequence is reported with variant information, as done by many studies included in the review.

Discordances that were due to flanking region indels were often resolved via analysis of raw data that included the flanking region sequences to enable variant identification and confirmed by re-sequencing using primers that extended the primer binding sequence [73,116,124]. Inclusion of flanking region data in DNA databases is thus an important consideration for forensic laboratories, as recommended by Parson *et al*., 2016 [69]. Including the full sequence string, or at least using a nomenclature reporting format where a flanking region insertion or deletion is easily identified as the cause for the discordance, will enable relatively seamless resolving of these types of discordances.

The commonalities in discordances reported as well as the paucity in concordance reporting calls for the *publishing* of population data *with* concordance data early-on in adoption to increase the familiarity with potentially discordant genotypes called in later studies. For example, if a discordance has been confirmed in many population groups at an early stage, then there can be a consensus and generality regarding the resolution of that discordance when discrepancies are observed in newer studies. However, at this point, consensus regarding concordance would be based on limited data, and with an absence of population *and* concordance data from more genetically diverse population groups, the consensus is yet to be reached.

2.4.7. STR sequence nomenclature formats

In this review, the most frequently used nomenclature format was the high-level bracketed repeat, consisting of the name of the STR marker, the length-based allele, the defined genome co-ordinates, the repeat motif in a bracketed format, with the number of repeats of each motif outside the bracket and the SNVs present within and around the repeat motif (**Table 2.4**). The

frequent use of this format may be attributed to it being a recommended format reported in the guidelines published by the STRAND working group [147]. This highlights the inclination of forensic laboratories to move towards a standardised STR nomenclature reporting approach, and the importance of inter-laboratory collaboration towards a specific goal, yet the guidelines proposed for sequence nomenclature have been developed with the absence of sequence data from more diverse population groups such as those in Africa.

The use of the bracketed repeat with the length-based allele, without genome co-ordinates or flanking region variant information, was also commonly considered as a reporting format among population studies, and this is likely due to it being the most easily human-readable format. As with the high-level nomenclature, this format also makes comparison between length-based allele frequency databases convenient, as no additional steps are needed to extract the length-based allele from the name of the STR sequence for compatibility purposes. Unsurprisingly, there were four different nomenclature formats used to report allele sequences in the population studies, which suggests a lack of standardisation of STR sequence nomenclature. However, most population studies in this review followed guidelines proposed for STR sequence nomenclature reporting, which showed some movement towards STR Sequence nomenclature standardisation [69,147]. The ability of several nomenclature formats to capture maximum diversity, while still being compatible with length- based population databases, is still being investigated, but the lack of sequence-based population data makes this process challenging [148].

A short designator system, which, in this review, consisted of a study reporting only the CE length with a lower-case letter for the level of sequence variation, was one of the less commonly

used formats. Alternative versions of short designator nomenclatures have been proposed in the literature, but preference to the type of short designator system used seemed to be laboratory specific [149,150]. All short designator nomenclature systems reveal no information about the actual sequence but is still able to differentiate between length and sequence-based variation level. The short designator nomenclature system also has the advantage of being easy to store and read and requires similar storage capabilities required for length-based alleles. However, additional steps are required to link the designator back to the full sequence details to dissect the variation present in each STR sequence, especially when full sequence information is not reported alongside reported alleles, which is often required for resolving discordances, as seen in this review (**Table 2.3**).

Considerations on minimal nomenclature requirements state that flanking region sequences should be included when reporting sequence data in order to account for variants occurring within this region [69,150]. To account for these nomenclature considerations, later versions of the ForenSeq™ UAS allowed for the generation of a flanking region report, which includes information about the length-based allele, and the full sequence string, including the flanking region variants. The use of the full ForenSeq™ UAS sequence string including was the least commonly used nomenclature. This is likely due to it being the least human-readable format. This nomenclature captures the full variation within and adjacent to the repeat region (if flanking regions are reported) but does not allow for the repeat motif to be visualised (as with the bracketed repeat).

In reporting the full sequence string only, the variation between sequences is not explicitly visible by a human analyst through *just* looking at the sequence string, therefore additional

characterisation would be required to visualise sequence and/or flanking region variation. The nature of the sequence string means that multiple steps are required to extract the repeat motif and the flanking region variation in order to characterise each STR sequence, adding further resistance to the standardisation of STR sequence nomenclature. This was also highlighted by Parson *et al*., 2016, where it was noted that specialised software was required to convert sequence strings into more human-readable formats that can be easily communicated in expert reports [69].

2.5. Conclusion

This review highlighted several critical issues and gaps in the current landscape of sequence-based population studies, particularly concerning African populations. Furthermore, although sequence data for African individuals residing in non-African countries were reported, these data do not fully capture the genetic diversity of African populations that have not migrated out of Africa. Indeed, our findings indicate that populations with African ancestry exhibit the highest level of variation in terms of increased allele counts and reduced RMP values compared to non-African population groups. This underscores the necessity for conducting sequence-based population studies on African populations to obtain a comprehensive understanding of the genetic diversity present. Furthermore, consensus on concordance is based on limited data, and the absence of both population data and concordance data from more genetically diverse groups means that a definitive consensus has yet to be achieved.

Addressing the gaps highlighted in this review is crucial for the advancement of population genetic studies and the implementation of MPS in forensic science, particularly in African contexts. Collaborative efforts must therefore be made to generate comprehensive population

data from Africa, improve concordance reporting, develop nomenclature systems that effectively capture genetic diversity, and explore avenues of cost mitigation through the use of validated direct PCR methods.

The review emphasised that the main contributing factor to the absence of population data in African countries is the high upfront costs of MPS implementation. It is therefore the responsibility of forensic laboratories to make every effort to explore avenues for mitigating costs, such as using optimised direct PCR methods with MPS.

Chapter 3: Optimisation study

Systematic optimisation of direct PCR protocols for crude buccal swab lysates using the ForenSeq™ DNA Signature Prep kit

3.1. Introduction

Chapter 2 revealed that although the ForenSeq™ DNA Signature Prep kit (Verogen, San Diego, CA, USA) has been developmentally validated for direct PCR methods, it is not used as often as conventional DNA extraction in MPS population studies [20]. The requirement for cost efficient, quick and reliable methods for generating DNA profiles, particularly for reference samples and allele frequency databases, resulted in the emergence of direct PCR methods for CE-based DNA profiling [28]. Direct PCR eliminates the need for DNA extraction and DNA quantification, thereby reducing the hands-on time and costs associated with sample preparation prior to PCR and DNA profiling [151]. With regards to CE-based DNA profiling, previous studies have addressed, reviewed and evaluated the use and performance of direct PCR methods as applied to both reference and casework samples in forensic laboratories [93,152]. Sample types which have been subjected to direct PCR methods for CE-based DNA profiling have ranged from blood, buccal swabs, soft tissue, hair, nails and trace DNA from various substrates, [93,152-156]. For the purposes of generating DNA profiles for establishing population databases, good quality samples such as blood and buccal swabs are routinely processed with direct PCR methods to generate DNA profiles [153,156].

The general method for processing buccal swabs with a direct PCR approach starts with a short lysis step, wherein the cells contained on a swab are separated from its substrate and lysed in a buffer at a specific temperature [157,158]. The resulting sample is known as a crude buccal swab lysate that can be added directly to a PCR reaction for amplification and subsequent profiling (CE or MPS). However, the risk associated with direct PCR methods is that, in

absence of purification, quantification and quality control steps, PCR inhibitors may cause interference with efficient PCR amplification [28].

Success rates of direct PCR methods have shown to be comparable with conventional methods that include DNA extraction and quantification prior to DNA profiling, however, some studies have shown sub-optimal first-pass rates for samples processed with direct PCR workflows [153,159,160]. Polymerases and PCR buffers in some commercial STR profiling kits have been designed to tolerate a certain level of inhibition and have a built-in buffering capacity to maintain optimal polymerase activity [161]. However, forensic kits designed for MPS are far more sensitive and could be overwhelmed by PCR inhibitors and sub-optimal polymerase conditions [162]. With the absence of purification, quantification and library quality assessment, the risk of sub-optimal success rates with sensitive MPS kit chemistries increases [28].

Chapter 2 revealed that no population studies using the ForenSeq™ DNA Signature Prep kit made use of crude buccal swab lysates, and it is therefore currently unclear how crude buccal swab lysates perform with MPS workflows. Crude buccal swab lysates generated using the SwabSolution™ kit (Promega Corporation, Madison, WI, USA) and the STR GO! Lysis buffer (QIAGEN, Hilden, Germany) are routinely used in our laboratory and have been used to successfully carry out population studies with CE systems using commercial STR profiling kits in our laboratory [163,164]. It was assumed that these same samples would perform similarly using the ForenSeq™ DNA Signature Prep kit (Verogen, San Diego, CA, USA), as the manufacturer's had tested crude buccal swab lysates and provided input recommendations in the manufacturer's protocol [21]. A population study was then initially attempted using the crude buccal swab lysates generated with the SwabSolution™ kit (Promega Corporation,

Madison, WI, USA) (hereon referred to as "SwabSolution™ lysates") and with the STR GO! Lysis Buffer (QIAGEN, Hilden, Germany) (hereon referred to as "STR GO! lysates") with the ForenSeq™ DNA Signature Prep kit (Verogen, San Diego, CA, USA) which resulted in failed MPS profiles.

This prompted a series of unforeseen troubleshooting experiments that evolved into a full optimisation study. The aim of this chapter was thus to investigate and optimise first-time success rates of crude buccal swab lysates generated using the SwabSolution™ kit (Promega Corporation, Madison, WI, USA) and the STR GO! Lysis buffer (QIAGEN, Hilden, Germany) that have been processed with the ForenSeq™ DNA Signature Prep kit (Verogen, San Diego, CA, USA), in preparation for use in a population study (Chapter 4).

3.2. Methods

3.2.1. Study design and overview

This study was carried out in three phases. A graphic is presented in **Figure 3.1**, where summary results are presented under each phase, to enhance clarity early-on of the systematic nature of this chapter. The phases included an initial testing phase to assess the first-time success rate of crude buccal swab lysates prepared using SwabSolution™ (Promega Corporation, Madison, WI, USA) or STR GO! Lysis buffer (QIAGEN, Hilden, Germany) and subsequently processed with the ForenSeq™ DNA Signature prep kit (Verogen, San Diego, CA, USA). Upon evaluation of call rates, poor performance in these lysates were noted. Phase 2 of this study therefore involved an investigation into the reasons for sub-optimal performance of crude buccal swab lysates processed with the ForenSeq™ DNA Signature Prep kit (Verogen, San Diego, CA, USA) using quantitative real-time PCR (qPCR). The results from phase 2 were used to identify and inform different optimisation methods which would be used in phase 3.

Phase 3 involved applying different methods to optimise the success rates of SwabSolution™ (Promega Corporation, Madison, WI, USA) and STR GO! Lysis buffer (QIAGEN, Hilden, Germany) crude buccal swab lysates with the ForenSeq™ DNA Signature prep kit (Verogen, San Diego, CA, USA). The quality of libraries for the different optimisation methods were assessed using two variables as a measure for potential profile success, namely, average library size and library concentration as assessed using TapeStation and Qubit™ fluorometry (Invitrogen, CA, USA), respectively.

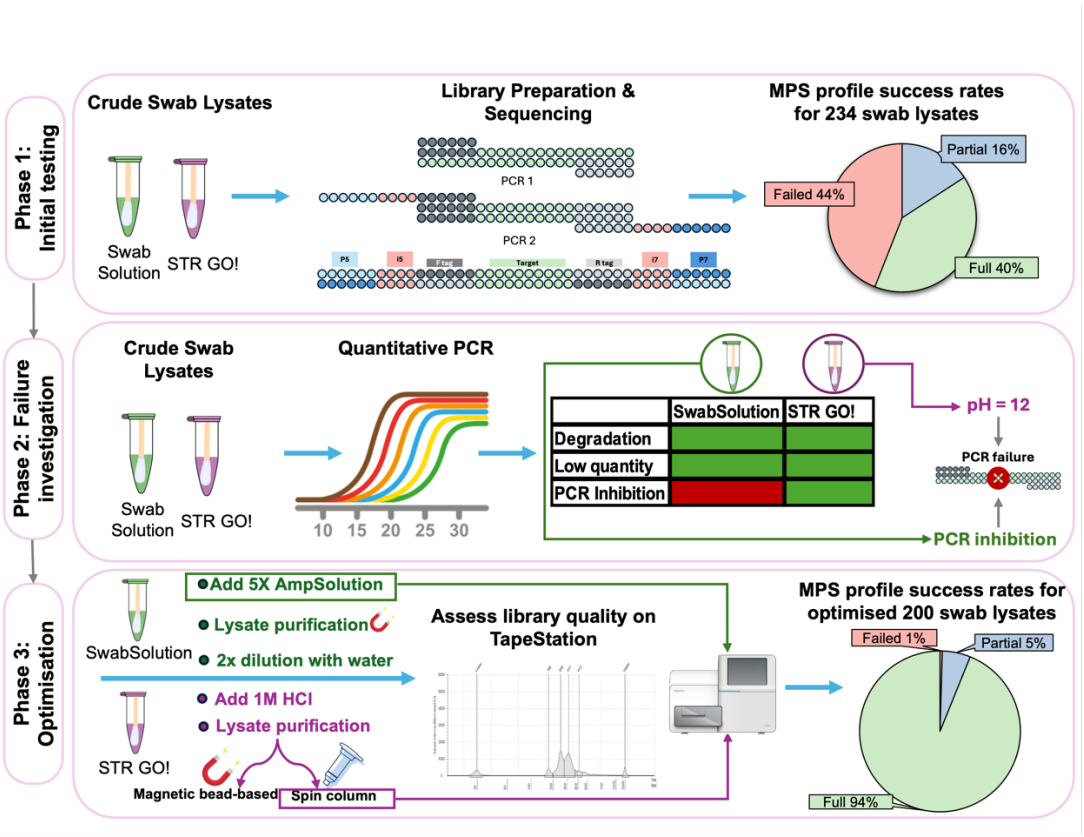

Figure 3.1: *Overview of the adaptations made to ensure a high-first time success rate with crude buccal swab lysates using the ForenSeq™ DNA Signature Prep kit workflow.*

3.2.2. Phase 1: Initial processing of crude buccal swab lysates with the ForenSeq™ DNA Signature prep kit

In an ongoing ethically approved study at the University of Cape Town (HREC: 342/2016), cotton buccal swabs were collected from living and deceased individuals, as described in Heathfield *et al*. 2024. Buccal swabs were processed using either the SwabSolution™ Kit (Promega Corporation, Madison, WI, USA) or the STR GO! Lysis Buffer (QIAGEN, Hilden, Germany) using the manufacturers' protocols [165,166]. Having obtained full conventional DNA profiles previously, as described in Heathfield *et al*., 2024, 234 of these samples underwent library preparation using a direct PCR approach with the ForenSeq™ DNA Signature Prep kit (Primer mix A and/or B) [21]. Libraries were purified, normalised, denatured and loaded as per the manufacturer's protocol, except that a loading volume of 12 µL of pooled, normalised library was used, as determined through prior internal laboratory optimisation experiments [21]. Approximately 600 µL of denatured Human Sequencing Control (HSC) and library were loaded onto a MiSeq FGx™ reagent cartridge and sequenced on the MiSeq FGx™ instrument (Verogen, San Diego, CA, USA) for 398 cycles in Forensic Genomics mode according to the manufacturer's protocol [59].

3.2.2.1. Data analysis

Primary data analysis was performed on the ForenSeq™ UAS, an analysis software integrated into the MiSeq FGx™ system [60]. Alleles were automatically called based on their read counts and whether the read count met the default analytical and interpretation thresholds of 1.5% and 4.5% respectively. The sample genotype reports generated on the ForenSeq™ UAS software were then exported into Microsoft Excel® to determine call rates. Call rate was calculated as the number of successfully called genotypes/haplotypes at each marker, divided by the total number of markers, and converted to a percentage.

3.2.3. Phase 2: Investigation into crude buccal swab lysate failure

3.2.3.1. Real-time PCR quantification

Crude buccal swab lysates processed in SwabSolution™ (Promega Corporation, Madison, WI, USA) and STR GO! Lysis buffer (QIAGEN, Hilden, Germany) with failed or partial MPS profiles with the ForenSeq™ DNA Signature prep kit (Verogen, San Diego, CA, USA) were identified and subjected to qPCR with the Quantifiler® Trio kit (Applied Biosystems, Foster City, USA), according to the manufacturers protocol, with the exception that half-volumes were used (as per optimised and internally validated procedures) [167]. Data were viewed using the HID real-time PCR software (Applied Biosystems, Foster City, USA) and further assessed using Microsoft Excel®. It was initially hypothesised that PCR inhibition played a role in sub-optimal lysate performance, therefore initial assessment of lysates included an assessment of PCR inhibition using qPCR. Lysates that were flagged for high IPC C_T (*i.e.,* IPC $C_T > 31$) values by the HID real-time PCR software were diluted two-fold with molecular biology grade water prior to re-quantification with the Quantifiler® Trio kit (Applied Biosystems, Foster City, USA) to confirm inhibition. The concentration and degradation index of all lysates were also assessed. These results informed the optimisation methods used in phase 2, which were used to overcome PCR inhibition prior to library preparation with the ForenSeq™ DNA Signature Prep kit (Verogen, San Diego, USA) [21].

3.2.3.2. pH assessment of STR GO! buccal swab lysates

Whilst helpful for the buccal swabs prepared in SwabSolution™ (Promega Corporation, Madison, WI, USA), the qPCR results provided little to no insight into MPS profile failure with STR GO! lysates. It was therefore hypothesised that sub-optimal performance of these lysates could be due to the incompatibility of lysis buffers with the ForenSeq™ DNA Signature Prep kits' PCR 1 components. Further investigation into this incompatibility involved personal

communication with the manufacturers to understand reasons behind potential incompatibility of buffers, and it was noted that the high pH of the STR GO! Lysis buffer (QIAGEN, Hilden, Germany) could not be negated due to the limited buffering capacity of ForenSeq™ DNA Signature Prep kit pre-PCR reagents (Personal Communication, QIAGEN, Hilden, Germany). This informed the optimisation methods used in Phase 3.

3.2.4. Phase 3: Method optimisation

3.2.4.1. SwabSolution™ crude buccal swab lysates

As PCR inhibition was confirmed in SwabSolution™ lysates in phase 2, a subset of 10 SwabSolution™ lysates that had failed MPS profiles were subjected to three different optimisation methods prior to library preparation with the ForenSeq™ DNA Signature Prep kit (Verogen, San Diego CA, USA):

- A two-fold dilution of crude buccal swab lysates with nuclease-free water prior to library preparation according to manufacturer's recommendations was performed [21].
- The addition of 3 µL 5X AmpSolution® reagent (Promega Corporation, Madison, WI, USA) to 2 µL of the crude buccal swab lysate (instead of 3 µL nuclease-free water).
- Purification of 100 µL of lysate using the Mag-Bind® Blood DNA HV kit (Omega Bio-tek, Norcross, GA, USA) eluted into 30 µL of TE buffer [168].

3.2.4.2. Lysate preparation methods for STR GO! Buffer lysates

To overcome the high pH of the STR GO! lysates, 10 STR GO! crude buccal swab lysates that had failed or sub-optimal MPS call rates were subjected to optimisation.

- Purification of 100 µL of lysate using the QiaAmp® DNA Investigator kit, eluted into 30 µL of ATE buffer (QIAGEN, Hilden, Germany) [169].

- Purification of 100 µL of lysate using the Mag-Bind® Blood DNA HV kit (Omega Bio-tek, Norcross, GA, USA) eluted into 30 µL of TE buffer [168].
- Addition of 3 µL 1M hydrochloric acid (HCl) (Sigma-Aldrich) to 2 µL of the crude buccal swab lysate (instead of 3 µL nuclease-free water) to reduce the pH without further dilution of the lysate.

3.2.4.3. Optimisation of the spin-column purification method to improve starting concentration

Sub-optimal DNA concentrations were obtained for lysates purified with using spin column-based purification methods and were thus further optimised. Fresh buccal swabs (n = 20) were collected from consenting individuals and processed using the STR GO! Lysis buffer (QIAGEN, Hilden, Germany). These lysates were used to determine if the spin-column-based purification method could be optimised to improve extracted DNA concentrations prior to sequencing. The first optimisation method evaluated two modifications to spin-column purification with the QiaAmp® DNA Investigator kit (QIAGEN, Hilden, Germany), where the lysate either underwent total purification according to the manufacturer's protocol or partial purification (*i.e.*, only the lysate was added directly to the spin-column). The second method evaluated the use of a protocol designed to purify DNA from crude buccal swab lysates using the DNeasy Blood and Tissue kit (QIAGEN, Hilden, Germany) and the adapted protocol for crude buccal swab lysates [170]. An additional modification was made to the "DNeasy" protocol to include the addition of 10% acetic acid to bring the pH of the lysate down to at least 7 (**Table 3.1**).

Table 3.1: *Modifications applied to two spin-column purification kits and protocols; namely, the QiaAmp® DNA Investigator kit and protocol, and the DNeasy kit used with the protocol designed for the purification of DNA from crude buccal swab lysates.*

Spin-column purification method	Modification 1	Modification 2
QiaAmp® DNA Investigator Kit (QIAGEN, Hilden, Germany)	▪ 100 µL of the STR GO! Lysate was added directly to a spin column. ▪ purification according to the manufacturer's protocol. ▪ elution volume was reduced to 30 µL.	▪ 100 µL of STR GO! Lysate was lysed with Buffer AL and proteinase K. ▪ purification according to the manufacturer's protocol. ▪ elution volume was reduced to 30 µL.
DNeasy Protocol (QIAGEN, Hilden Germany	▪ 200 µL of the STR GO! Lysate was purified using the DNeasy. ▪ elution volume was reduced to 50 µL.	▪ 20 µL of 10% Acetic acid was added to 200 µL of the lysate to reduce the pH to 7. ▪ 220 µL of the STR GO! Lysate was purified using the DNeasy kit. ▪ elution volume was reduced to 50 µL.

3.2.4.4. Quantification of extracted DNA

Where lysates were purified, DNA was quantified with the Quantifiler® Trio Kit (Applied Biosystems, Foster City, USA) as described in section 3.2.3.1, and diluted to 0.2ng/µL with nuclease-free water.

3.2.5. Library preparation

Extracted DNA and crude buccal swab lysates prepared in phase 3 were processed using the ForenSeq™ DNA Signature Prep kit (Verogen, San Diego, CA, USA), as stipulated in the manufacturer's protocol, with modifications as presented in phase 3 above [21]. All libraries were prepared alongside a positive control using 5 µL of the 2800M control DNA (Promega Corporation, Madison, WI, USA), which was prepared by dilution to 0.2 ng/µL with nuclease-free water, along with a negative control consisting of 5 µL nuclease-free water. The amplification and tagging of targets, enrichment of targets and library purification was performed according to the manufacturer's guidelines. Purified libraries were stored in a -20

°C freezer until further use. The 'safe stopping point' (as per the manufacturer's protocol) was used to assess library quality and quantity [21].

3.2.6. Assessment of library quality and quantity of crude buccal swab lysates

The purified libraries of the positive control, the lysates subjected to the different library preparation methods as well as previously processed lysates that had initially resulted in failed or partial profiles were processed using the High Sensitivity D1000 Screen Tape Assay with the 2200 TapeStation (Agilent Technologies) following the manufacturer's protocol [171]. The average library sizes were recorded from the TapeStation Controller Software. Library concentration was assessed using the Qubit 2.0 Fluorometer with the Qubit™ 1x dsDNA High Sensitivity Assay (Thermo Fisher Scientific, Waltham, MA, USA) [172].

3.2.7. Statistical analysis for assessment of optimisation methods

To evaluate whether the lysate preparation method used influenced library quantity and average library size, two-sided Friedman's tests were performed after the assumption of normality was not met. This was used to test for significant differences in library concentration and library sizes between the different lysate preparation methods. To determine the most optimised protocol for STR GO! purified lysates, their library concentration and average library sizes of the purified libraries were compared to the 2800M control purified library metrics (library concentration and size) using a one-sample Wilcoxon test, as both used the same input amount of 1 ng. The Friedman test was also used to compare the difference in concentrations obtained from the four different silica-based methods tested to determine which approach resulted in the highest extracted DNA concentration. All statistical analyses were conducted using the SciPy. Stats package in Python version 3.7 [173].

3.2.8. Library normalisation and sequencing

Following statistical assessment of the lysate preparation method best suited to each type of lysate (SwabSolution™ and STR GO! lysates), the selected purified libraries were normalised and sequenced on the MiSeq FGx™ sequencer in batches of 32 according to the manufacturer's protocol, and as described in section 3.2.2. Additional samples (n = 212) were sequenced using the optimised protocols as part of a population study (Chapter 4) and call rates compared to the previously failed crude buccal swab lysates.

3.3. Results

3.3.1. Phase 1: Initial assessment into failure rates

3.3.1.1. Quality metrics

The quality metrics for the seven experimental runs that included lysates processed in phase 1 are shown in **Table 3.2**. All quality metrics were within the recommended ranges reported by the manufacturer, except for three runs, where cluster density was below the recommended range. This was likely due to the sample variability within the pooled library, where some samples have expected lower concentrations resulting in an overall lower cluster density. Although, it has been established and reported that quality metrics falling outside the recommended ranges are still likely provide an adequate output of usable data [59].

***Table 3.2**: Quality metrics obtained for sequencing runs performed with the ForenSeq™ DNA Signature Prep kit on the MiSeq FGx™ sequencer. Bold text indicates values that were below the manufacturer's recommended ranges.*

Experiment Number	Cluster Density (k/mm²)	Clusters Passing Filter (%)	Phasing	Pre-phasing
Run 1	777	93.25	0.132	0.093
Run 2	1069	91.00	0.112	0.095
Run 3	**681**	92.63	0.158	0.113
Run 4	**456**	96.35	0.219	0.098
Run 5	1000	88.95	0.156	0.115
Run 6	1187	87.87	0.143	0.114
Run 7	**520**	91.07	0.205	0.094

3.3.1.2. Call rates

Preliminary assessment of call rates of crude buccal swab lysates processed with the ForenSeq™ DNA Signature Prep kit and MiSeq FGx™ workflow revealed that a substantial portion of crude buccal swab lysates resulted in failed or partial profiles. The failed profiles accounted for 44% (n = 103) of the crude buccal swab lysates processed, 16% (n = 37) were partial profiles, and 40% (n = 94) were full profiles. SwabSolution™ lysates accounted for 61% (n = 63) of failed samples, while STR GO! lysates made up 39% (n=40) (**Table 3.3**).

***Table 3.3:** Number of full, partial and failed profiles obtained for SwabSolution™ lysates and STR GO! lysates when processed with the ForenSeq™ DNA Signature Prep kit. Italic text represents the breakdown of the number of samples with failed profiles.*

Success category	Number of samples
Partial	**37**
Full	**94**
Failed	**103**
-SwabSolution™	*63*
-STR GO!	*40*
Total processed	**234**

3.3.2. Phase 2: Investigation into lysate failure

3.3.2.1. SwabSolution™ crude buccal swab lysates

The subset of lysates (n = 103) that resulted in failed MPS profiles were categorised into lysate type; SwabSolution™ lysates and STR GO! lysates. Real-time qPCR results showed that 93.65% (n = 59/63) of SwabSolution™ lysates had IPC C_T values above 31 and were thus likely to be affected by PCR inhibition. The DI values for SwabSolution™ lysates indicated that the lysates were unlikely degraded, and this was ruled out as a possible reason for failure. Additionally, where samples were not affected by inhibition, concentrations were within adequate ranges for sequencing. It should be noted that many lysates that were flagged for inhibition had unknown DI and concentration values, therefore concentration and DI values could not be properly assessed for those samples. In these cases, the reason for failure with MPS methods was taken as inhibition. Dilution of SwabSolution™ crude buccal swab lysates that were identified as potentially inhibited showed an improvement in the IPC C_T value, where 87.30 % (n = 55/63) of lysates had new IPC C_T values < 30, supporting the hypothesis that these samples were initially inhibited.

3.3.2.2. STR GO! crude buccal swab lysates

The real-time PCR results for STR GO! (QIAGEN, Hilden, Germany) crude buccal swab lysates showed that 100% of lysates had an IPC C_T value below 31 and were thus *unlikely* to be inhibited. The DI values indicated little to no degradation. Concentrations were above 0.2 ng/µL in all samples. The failure of lysates with MPS could thus not be attributed to low input concentration, degradation or inhibition. Further investigation into reasons for failure involved personal communication with the manufacturers, during which it was established that the pH of the STR GO! Lysis buffer (QIAGEN, Hilden, Germany) was too high to be compatible with the ForenSeq™ DNA Signature Prep kit PCR 1 buffers (Personal communication, QIAGEN,

Hilden, Germany). This information was pivotal to establish methods to improve compatibility between the STR GO! Lysis buffer (QIAGEN, Hilden, Germany) and the ForenSeq™ DNA Signature Prep kit pre-PCR reagent chemistry. Assessment of the pH levels of the STR GO! Lysis buffer (QIAGEN, Hilden, Germany) indicated a buffer of a highly alkaline nature (pH = 12).

3.3.3. Phase 3: Method optimisation

3.3.3.1. SwabSolution™ lysate library quality

Libraries prepared with 5X AmpSolution® reagent resulted in the highest average library concentrations, followed by lysates prepared with magnetic bead purification and lysates prepared according to the manufacturers protocol (*i.e.,* "original") (**Figure 3.2**). Lysates diluted 2-fold prior to library preparation resulted in the lowest library concentrations. Lysates spiked with 5X AmpSolution® resulted in significantly higher library concentrations than all other optimisation methods tested ($p < 0.05$). Furthermore, lysates that underwent magnetic bead purification also resulted in significantly higher library concentrations than lysates prepared with the manufacturers protocol and lysates diluted two-fold ($p < 0.05$). No significant differences were observed between library concentrations of lysates diluted two-fold and lysates prepared according to the manufacturer's protocol.

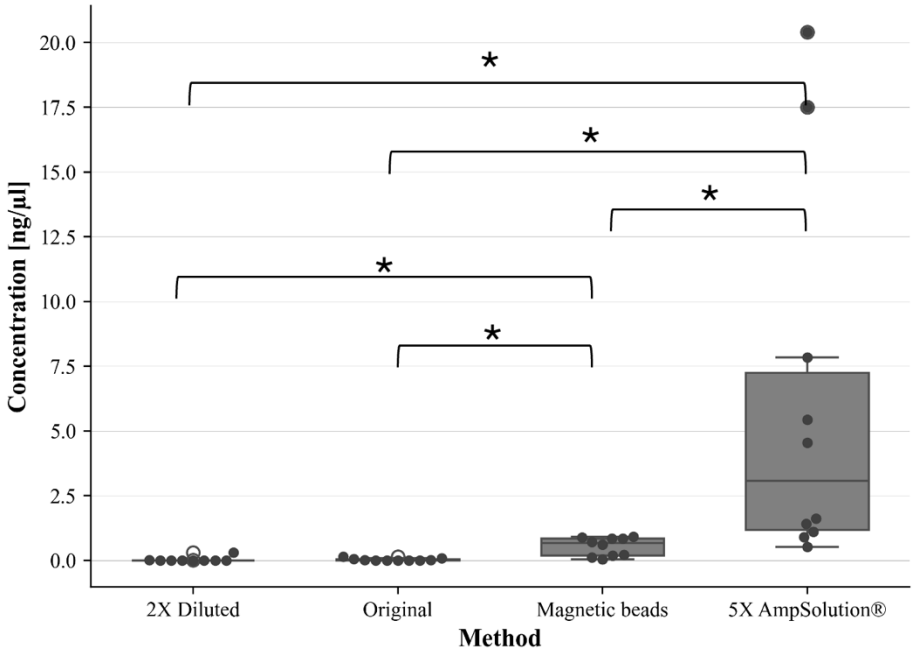

Figure 3.2: *Box-and-whisker plot representing average library concentration for SwabSolution™ lysates prepared with modifications to the manufacturer's protocol.*

Lysates prepared with 5X AmpSolution® resulted in the highest overall library size compared to lysates prepared with lysates purified with magnetic beads, lysates diluted two-fold, and lysates prepared with the manufacturer's protocol (**Figure 3.3** and **Figure 3.4**). Friedman's test revealed significant between-group differences, while post-hoc comparisons revealed that significant differences in library size existed between lysates prepared with 5X AmpSolution® reagent and all other methods tested ($p < 0.05$).

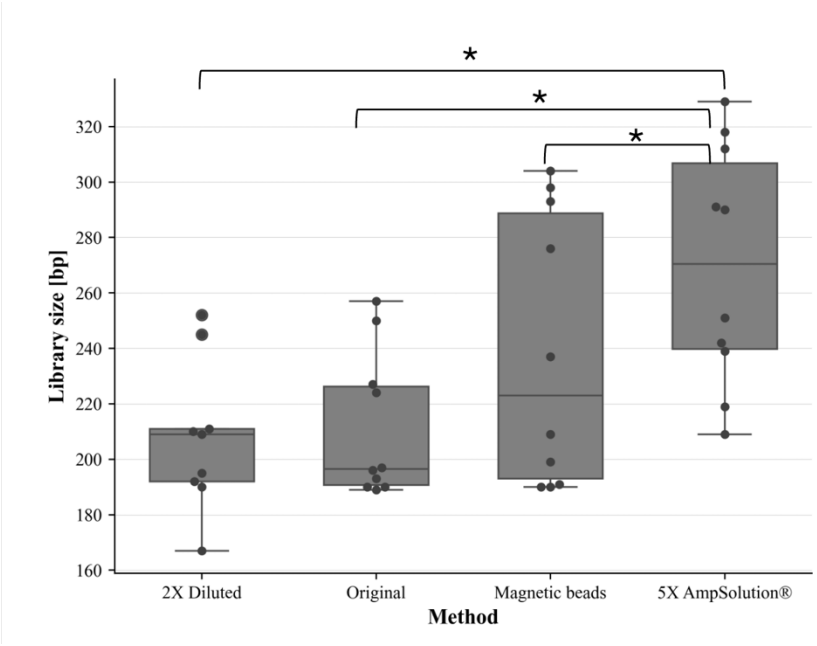

***Figure 3.3**: Box and whisker plot representing average library size (bp) for SwabSolution™ lysates prepared with modifications to the manufacturer's protocol.*

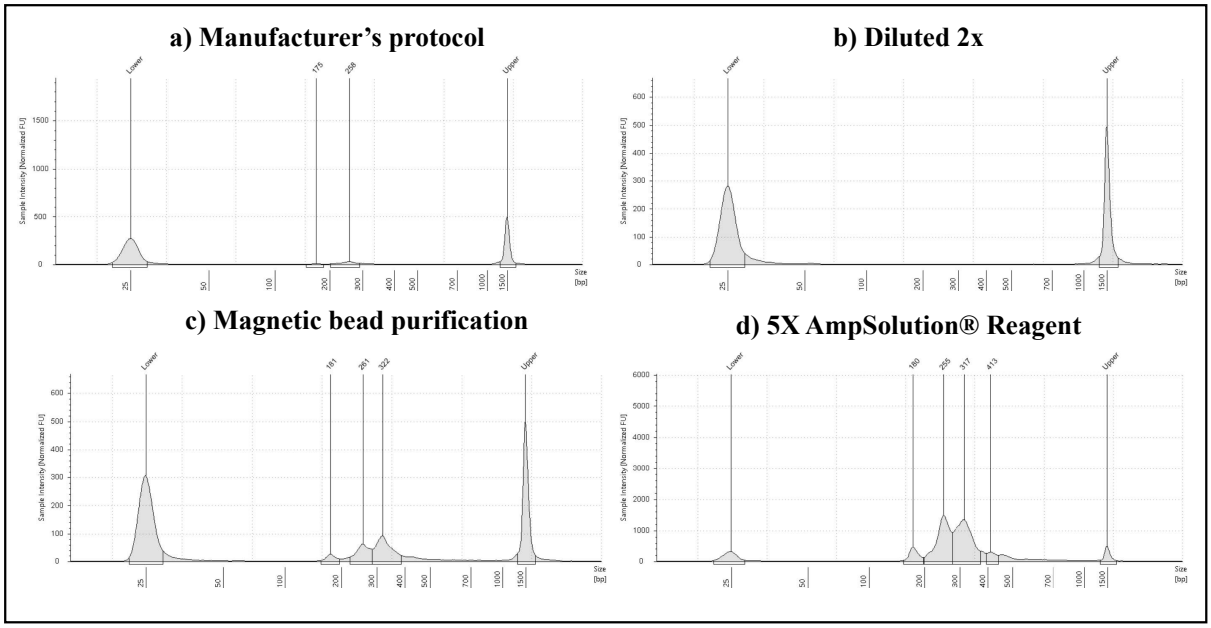

Figure 3.4: *TapeStation traces with sample intensity shown on each Y-axis in fluorescence units and library size shown in base pairs (bp) on the X-axis. The traces are shown for a SwabSolution™ crude buccal swab lysate prepared with a) the manufacturer's protocol, b) a modified protocol where the lysate was diluted 2x with nuclease-free water, c) modified protocol whereby the lysate was purified using the Mag-Bind® Blood DNA HV kit and d) a modified protocol where 3uL of 5X AmpSolution® reagent was added to the PCR 1 reaction.*

3.3.3.2. Sequencing success of optimised crude buccal swab lysate protocol: SwabSolution™ lysates

From these results, the addition of 5X AmpSolution® reagent improved both the quantity and quality of libraries generated for SwabSolution™ (Promega Corporation, Madison, WI, USA) crude buccal swab lysates. While the aim of lysate purification was to reduce the number of inhibitors present, the addition of 5X AmpSolution® reagent enabled a higher PCR efficiency, overcoming PCR inhibition even with no purification steps. A direct PCR approach could thus be maintained. This method was selected to be used in the population study (Chapter 4). SwabSolution™ lysates that previously failed, as well as previously unprocessed SwabSolution™ lysates resulted in improved call rates (**Figure 3.5**). The mean call rate for DPMA markers across 93 SwabSolution™ lysates processed before optimisation was 30.03%

± 37.72 This increased to 96.39% ± 10.71 across 136 samples when adding 5X AmpSolution®. There were nine samples (n = 9) that resulted in partial profiles (call rates between 30.13% and 64.13%) even after 5X AmpSolution® was added, but upon evaluation of qPCR data, these were found to be moderately degraded.

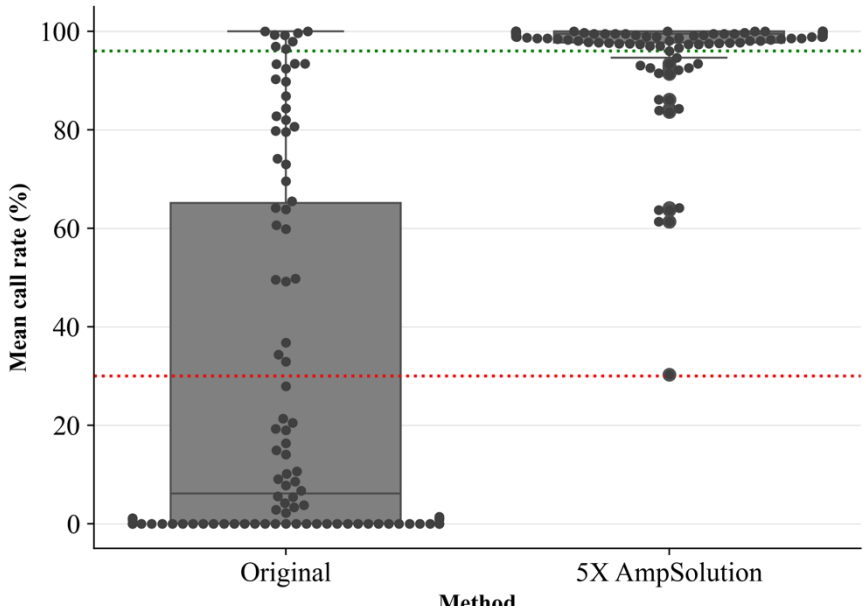

Figure 3.5*: Box-and-whisker plot comparing call rates in percentage (%) across DPMA markers for SwabSolution™ lysates processed before (Original) and after optimisation (5X AmpSolution® added). The red dotted line represents the mean call rate (30.03%) for lysates processed using the original protocol, while the green dotted line represents the mean call rate (96.39%) for lysates processed with 5X AmpSolution®.*

3.3.3.3. STR GO! Lysate library quality assessment

Lysates purified with the QiaAmp® DNA Investigator kit (QIAGEN, Hilden, Germany) resulted in higher library concentrations than all other methods. Significant differences in library concentration occurred between lysates purified with the QiaAmp® DNA Investigator kit and the lysates which were modified to include a 1 M HCl addition ($p < 0.05$) (**Figure 3.6**). Although a marked improvement in library concentration was observed in when using magnetic bead purification, the median concentration was lower than that observed for the 2800M control DNA (6.26 ng/µL).

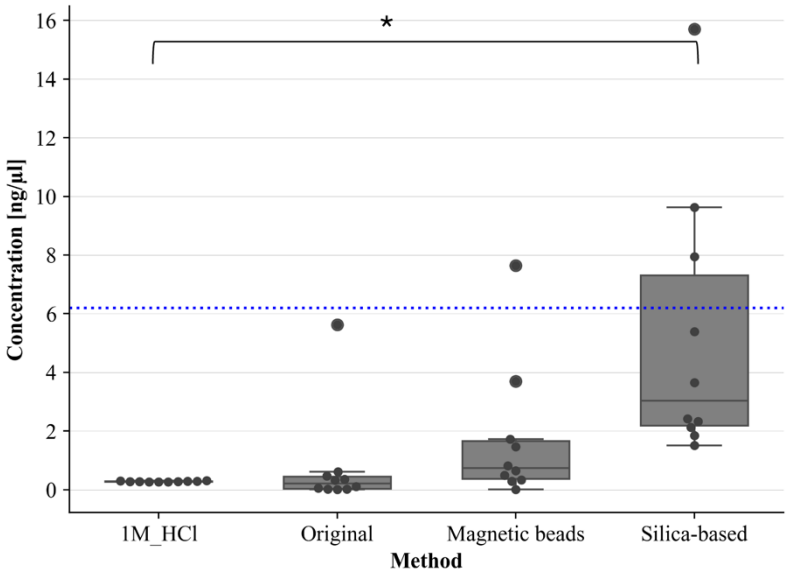

Figure 3.6: *Box-and-whisker plot representing library concentration for STR GO! lysates prepared with modifications to the manufacturer's protocol. The blue-dotted line represents the library concentration of the 2800M control DNA library of 6.26 ng/µL.*

For assessment of library size, lysates purified with the QiaAmp® DNA Investigator kit (QIAGEN, Hilden, Germany) resulted in a significantly higher average library size (Median: 284.5bp ± 17.49) than lysates prepared with prepared according to the manufacturer's protocol, which resulted in the lowest library sizes when all methods were compared [(Median: 239bp ± 45), ($p < 0.05$)] (**Figure 3.7**).

As both purified lysates and the 2800M control DNA had the same input amount (1 ng), they could be directly compared. A one-sample Wilcoxon test to compare library concentrations of each method of lysate preparation against the average library concentration of the control DNA indicated that the average library concentration of the 2800M control DNA library resulted in significantly higher library concentrations than lysates prepared with; a) 1 M HCl, b) the manufacturer's protocol c) magnetic bead purification ($p < 0.05$). Only the *average library concentration* of lysates prepared with the QiaAmp® DNA Investigator kit did not differ

significantly from the library concentration of the control DNA library ($p > 0.05$). There were no significant differences in *average library size* for all methods compared to the 2800M control DNA library size of 276 bp (**Figure 3.7**).

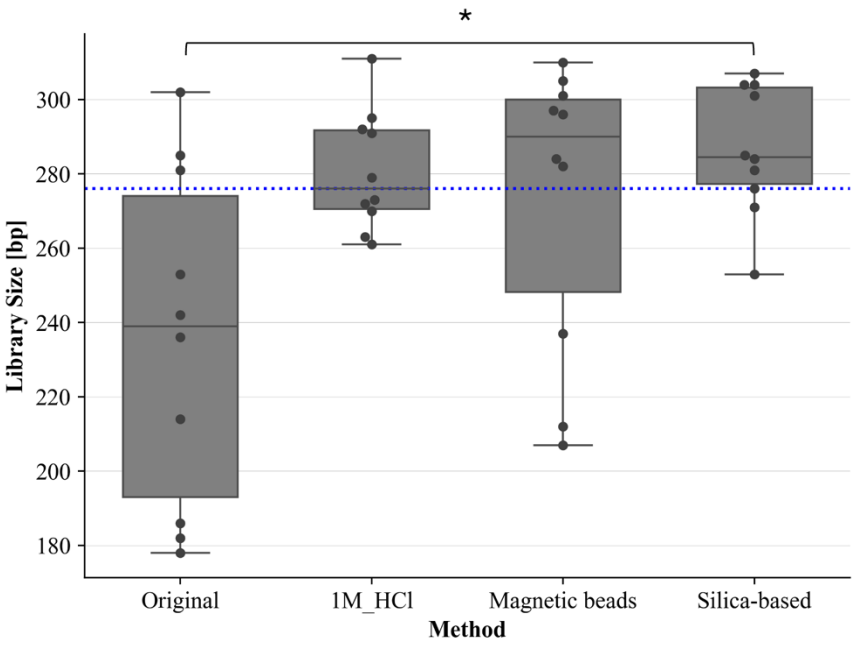

***Figure 3.7**: Box-and-whisker plots representing average library size for STR GO! lysates prepared with modifications to the manufacturer's protocol. The blue-dotted line represents the average library size of the 2800M control DNA library of 276 bp*

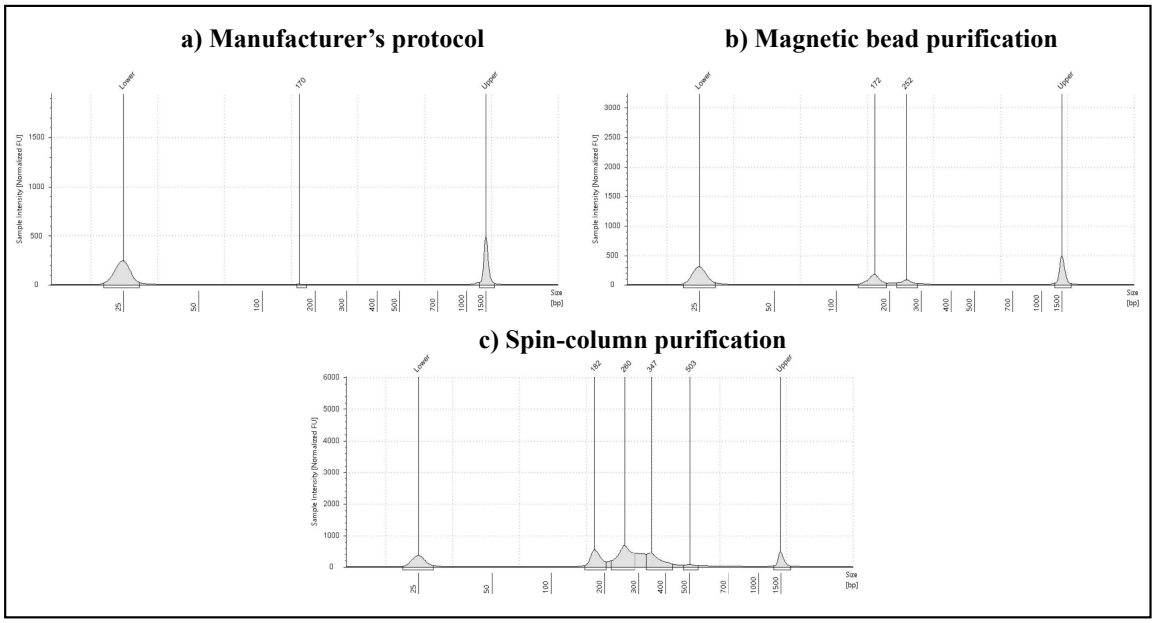

Figure 3.8: *TapeStation traces with sample intensity shown on each Y-axis in fluorescence units and library size shown in base pairs (bp) on the X-axis. The traces are shown for a STR GO! crude buccal swab lysate prepared with a) the manufacturer's protocol, b) a modified protocol whereby the lysate was purified using the Mag-Bind Blood DNA HV kit and c) a modified protocol whereby the lysate was partially purified using spin-columns with the QiaAmp® DNA Investigator kit.*

3.3.3.4. Sequencing success of optimised crude buccal swab lysate protocol: STR GO! lysates

It can be observed from the results that purification with the QiaAmp® DNA Investigator kit resulted in the most improved library concentration. Although average library sizes were similar across all methods, the addition of 1 M HCl and silica-based purification resulted in a more consistent average library concentration across the 10 samples, showing the least variation. However, the addition of 1 M HCl resulted in extremely low concentrations when compared to the silica-based method. For this reason, STR GO! lysates (QIAGEN, Hilden, Germany) purified using spin-column purification was used and selected for full sequencing with the ForenSeq^{TM} DNA Signature prep kit for the population study (Chapter 4).

The STR GO! lysates that that had failed in the initial testing phase, as well as previously unprocessed STR GO! lysates were processed using the optimised protocol (Table 3.1, Modification 1). A marked improvement in call rate was noted, where the original protocol resulted in a mean call rate of 58.64% ± 36.85 across 145 samples, increasing to 94.73% ± 12.27 when using the optimised protocol across 76 samples. Thus, for both SwabSolution™ and STR GO! lysates, the call rate improved from 44.35% to 93.38%. It should be noted that four samples did not result in full profiles, and this was hypothesised to be due to low input concentrations, which was further investigated and reported in the section below.

3.3.3.5. Improving input concentration of spin column-based purified lysates

Upon evaluation of library quality and concentration, the spin-column based purification method was selected as the most appropriate method for purification of STR GO! lysates (QIAGEN, Hilden, Germany). However, the average concentration of *extracted DNA* samples was 1.19 ng/µL. Although this concentration meets the input recommendations for the ForenSeq™ DNA Signature Prep kit for extracted DNA, an improved concentration was required for consistent performance of purified lysates. The standard lysate purification method which included adding the lysate directly to the spin column resulted in the lowest average concentration across 20 samples (1.19 ng/µL) (**Figure 3.9**).

Total purification of the lysate (*i.e.*, including lysis and binding steps stipulated in the manufacturer's protocol) resulted in slightly higher concentrations (2.79 ng/µL) (**Figure 3.9**). It was therefore hypothesized that the pH of the STR GO! lysate resulted in sub-optimal binding to the spin-column, causing lower than expected recovery. This warranted using a method suited for lysate purification specifically. Using the DNeasy protocol for lysate purification resulted in significantly higher extracted DNA concentrations than both QiaAmp® DNA

Investigator kit (QIAGEN, Hilden, Germany) protocols mentioned above. Furthermore, it was found that adjusting the pH to 7 prior to adding the lysate to the spin column result resulted in slightly higher concentrations (7.83 ng/µl) than when the pH was not adjusted (7.31 ng/µl), although this increase was not statistically significant.

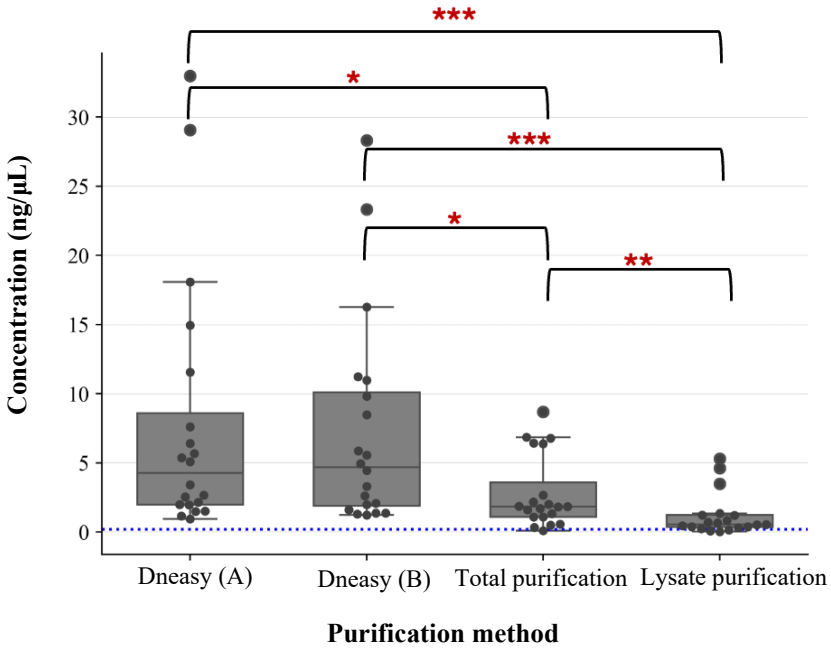

Figure 3.9: *Box and whisker plot representing four modifications spin-column purification aimed at improving extracted DNA concentration. The DNeasy method refers to the use of the protocol titled 'Purification of total DNA from crude lysates with the DNeasy Blood and Tissue kit' [170], while* ***(A)*** *denotes the modification of the protocol through the addition of 10% acetic acid and* ***(B)*** *without addition of any acid.*

3.4. Discussion

The aim of this study was to develop an optimised approach toward processing crude buccal swab SwabSolution™ (Promega Corporation, Madison, WI, USA) and STR GO! (QIAGEN, Hilden, Germany) lysates with the ForenSeq™ DNA Signature prep kit, given sub-optimal first-time success rates with the manufacturer's recommended protocol. Real-time PCR of all lysates provided a confirmation of the hypothesis that PCR inhibitors were present in SwabSolution™ crude buccal swab lysates. The addition of 5X AmpSolution® to ForenSeq™ DNA Signature Prep kit libraries significantly improved library quantity and quality compared to other methods (dilution and purification). Real-time PCR results provided little to no information to inform MPS profiling success for STR GO! lysates, while the pH of the STR GO! lysis buffer provided a better indication of PCR failure. Purification of STR GO! lysates using spin column-based methods resulted in overall improved library quantity and quality. The mechanisms by which library quality and quantity have been improved through these modifications is highlighted in this discussion.

3.4.1. PCR inhibition in SwabSolution™ Lysates

It was hypothesised that lysates contained PCR inhibitors interfering with ForenSeq™ DNA Signature Prep kit PCR buffers. For lysates prepared with SwabSolution™ (Promega Corporation, Madison, WI, USA), the PCR inhibition hypothesis was confirmed with real-time PCR, as SwabSolution™ lysates had IPC C_T values above 31, and when diluted two-fold, resulted in IPC C_T values below 31. This suggests that if a PCR inhibitor was present in the undiluted lysate, the effect of diluting the sample would essentially dilute the inhibitor, thereby enabling successful real-time PCR amplification.

Previous studies have shown high success rates with the SwabSolution™ kit (Promega Corporation, Madison, WI, USA) when applied to conventional DNA profiling methods [174-176]. For the samples used in this study, all SwabSolution™ lysates resulted in high success with CE kits [164,177]. The ForenSeq™ DNA Signature Prep kit manufacturers protocol states that the SwabSolution™ kit (Promega Corporation, Madison, WI, USA) has been validated with the ForenSeq™ DNA Signature Prep kit and recommends adding 3 µL water and 2 µL of the crude buccal swab lysate directly to the PCR 1 reaction, without prior assessment or quantification of the lysate, and without library quality assessment [21]. However, studies that have used the SwabSolution™ kit (Promega Corporation, Madison, WI, USA) with the ForenSeq™ DNA Signature Prep kit are limited. In one such study, authors stipulated that a direct PCR approach was not used, but rather that lysates were purified, quantified and diluted prior to processing with MPS kits [30].

In another study where the SwabSolution™ kit (Promega Corporation, Madison, WI, USA) was used for MPS with the ForenSeq™ DNA Signature Prep kit, authors quantified and diluted lysates to 0.2 ng/µL prior to processing with the ForenSeq™ DNA Signature Prep kit [178]. Therefore, neither of these studies employed a direct PCR approach when using the SwabSolution™ kit (Promega Corporation, Madison, WI, USA). Although it cannot be assumed that the authors also experienced first-time failure with the ForenSeq™ DNA Signature prep kit when processing SwabSolution™ lysates modifications were made to the manufacturer's protocol, suggesting that authors suspected that using a full direct PCR approach would have a higher risk of sub-optimal first-time success rates with MPS [30,178].

In this study, three methods were evaluated and compared to overcome PCR inhibition in SwabSolution™ lysates, including dilution with water, addition of 5X AmpSolution®, and

purification with magnetic beads. It was concluded that dilution of lysates was not a suitable method for overcoming inhibition in SwabSolution™ lysates. The poor library quality of lysates diluted with water (**Figure 3.4**) was surprising as quantification results showed an acceptable concentration, DI value and IPC C_T value. Real-time PCR has been used as a method for assessing downstream profiling success in many studies, however, a few sources have highlighted its shortcomings when using real-time PCR results to inform to profiling success [179-181]. In this study, real-time PCR was not a suitable and consistent proxy for determination of profiling success of diluted SwabSolution™ lysates with the ForenSeq™ DNA Signature Prep kit. However, it was successfully used to detect PCR inhibition in undiluted lysates. To this end, the Quantifiler® Trio kit was more tolerant to PCR inhibitors in the SwabSolution™ lysates than the ForenSeq™ DNA Signature Prep kit but was thus minimally informative for predicted MPS profile success.

The second method tested for overcoming inhibition was purification of lysates using the Mag-Bind Blood DNA HV kit (Omega Bio-tek, Norcross, GA, USA). Purification of the lysates with this method resulted in lower IPC C_T values, suggesting that PCR inhibition was successfully overcome. Additionally, improved library quality was achieved when compared to libraries where lysates were diluted with water, but this improvement was not significantly better than libraries of lysates that had resulted in failed MPS profiles (**Figure 3.3**)

Lastly, the addition of 5X AmpSolution® significantly improved library concentration and library size, while showing similar concentration and library size to that of the 2800M positive control. The use 5X AmpSolution® is not stipulated in the manufacturer's protocol [21] but is recommended when performing direct amplification of SwabSolution™ lysates on specific PowerPlex® systems [165]. Although the composition of the 5X AmpSolution® reagent is

proprietary, it is known that the reagent works to reduce the effect of inhibitors in the lysate. This approach was selected as the best method to overcome inhibition, while maintaining a direct-PCR approach. It is therefore recommended that when processing SwabSolution™ lysates using a direct PCR approach with the ForenSeq™ DNA Signature Prep kit, 5X AmpSolution® must be added to the reaction to maximise the first-time success rates.

3.4.2. Limited pH buffering capacity in STR GO! lysates

Lysates that failed due to high pH causing issues with the ForenSeq™ PCR 1 buffering capacity were subjected to silica-based and magnetic bead-based purification methods to lower the pH to an acceptable range while maintaining an adequate input concentration for MPS. A direct approach was also tested which included the addition of 1 M HCl to the PCR 1 reaction containing the crude buccal swab lysate. While both purification methods resulted in improved library concentrations compared to the original method, the silica-based purification resulted in higher library concentrations than magnetic-bead based purification and significantly higher concentrations than the direct approach in which 1 M HCl was added to the reaction. Although a slight improvement was noted with the Mag-Bind Blood DNA HV kit, recovery was lower than with the QiaAmp® DNA Investigator kit (QIAGEN, Hilden, Germany) (**Figure 3.6**). This is likely due to improved compatibility between QiaAmp® DNA Investigator Kit (QIAGEN, Hilden, Germany) and the STR GO! Lysis buffer (QIAGEN, Hilden, Germany) components.

Furthermore, assessment of extracted DNA concentrations of STR GO! lysates where 100 µL of lysate was added directly to the spin column showed that concentrations were minimally adequate for input amounts required for MPS and required further optimisation to improve recovery. This was warranted by internally established methods, as well as the manufacturer's internal validation, indicating that the average concentration of a buccal swab extracted with

the QiaAmp® DNA Investigator kit (QIAGEN, Hilden, Germany) results in concentrations ranging between 5-10 ng/µL [182]. A protocol was developed by QIAGEN to purify DNA from crude lysates using the DNeasy Blood and Tissue kit (QIAGEN, Hilden, Germany) [170], which, in this study was successfully used to improve the DNA recovery from STR GO! lysates compared to the QiaAmp® DNA Investigator kit (QIAGEN, Hilden, Germany) protocol [183]. The pH adjustment prior to adding the lysate to a spin-column resulted in the highest average concentrations, which underscores the importance of creating optimal binding conditions prior to purification.

Although spin-column purification resulted in adequate library quality and concentration, the method steers away from using a direct PCR approach. A direct-PCR approach was tested on STR GO! lysates by the addition of an acid to the PCR reaction to reduce the pH of the lysate to an acceptable range; however, this was at the cost of diluting the lysate to a level that was found to be too low for efficient PCR amplification.

A highly alkaline environment is usually overcome through built-in buffering capacity of PCR reagents in CE kits [184]; however, it is anticipated that the sensitivity of MPS kit chemistries may not include sufficient buffering capacity to reduce pH to an acceptable range for optimal polymerase activity. For this reason, it is recommended that the manufacturer's produce a reagent that would enable the direct amplification of samples lysed in STR GO! Lysis buffer (QIAGEN, Hilden, Germany) that addresses the high pH while simultaneously maintaining a suitable input concentration needed for successful amplification of fragments. This is especially important for laboratories already making use of STR GO! lysates (QIAGEN, Hilden, Germany) for reference buccal swabs used for large-scale sequence-based population studies, to make use of samples that have already been collected.

3.4.3. Importance of a quality control step prior to and post library preparation when conducting databasing studies

A drawback of the low first-time success rates achieved with STR GO! lysates in the context of large-scale population studies is that where length-based allele frequency data has been generated using STR GO! lysates, such as in Heathfield *et al*., 2024 and Whittaker and Heathfield 2024, concordance studies will be challenging when conducting *sequence-based* population studies using the same samples unless modifications are made to the protocol [164,177]. Considering this, an update to the manufacturer's protocol is proposed to purify STR GO! lysates prior to library preparation or to include a pre-sequencing quality control step on a small subset of purified libraries of STR GO! lysates. This study has demonstrated that TapeStation is a suitable quality control method to review the quality of library traces prior to normalisation and sequencing. Although not required in the manufacturers protocol, the use of crude buccal swab lysates with the ForenSeq™ DNA Signature Prep kit does in fact require quality control steps prior to library preparation and prior to sequencing.

3.5. Conclusion

Forensic laboratories moving towards the adoption of MPS for forensic casework or reference samples should consider the compatibility of previously collected samples used for direct PCR with CE technology with sensitive MPS kit chemistries. This paper has provided valuable insights into direct PCR approaches that can be adapted for MPS with the ForenSeq™ DNA Signature Prep kit. The systematic approach used to elucidate and identify optimal methods for sequencing of crude buccal swab lysates have informed recommendations for current and future researchers conducting large-scale sequence-based population studies. For SwabSolution™ lysates, the addition of 5X AmpSolution® to the lysate prior to PCR steps is recommended to overcome potential PCR inhibition, while spin-column purification of the STR GO! lysate with the DNeasy Blood and Tissue kit protocol and adapted for crude lysates

is recommended. Furthermore, it is highly recommended that the SwabSolution™ lysis buffer be used over the STR GO! lysis buffer, particularly when using ForenSeq™ kit chemistries, given that a direct PCR approach can be maintained by simply adding 5X AmpSolution®. The insights in this chapter have achieved the aim of improving low first-pass success rates obtained with the SwabSolution™ and STR GO! lysates when processed with the ForenSeq™ DNA Signature Prep kit. The results of this study have streamlined the manufacturer's protocol to avoid repeat sampling and re-sequencing. This is crucial for low-resource laboratories that may rely on readily available population or reference samples for establishing allele frequency databases, avoiding significant financial constraints. This research has thus crucially established a trusted workflow using crude buccal swab lysates for the purposes of carrying out a sequence-based population study for the South African population.

Chapter 4: Population study

Characterisation of autosomal STR sequence variation for the South African population using the ForenSeq™ DNA Signature Prep kit

4.1. Introduction

The breadth of sequence-based population data that exist in the global forensic community was extensively explored in Chapter 2 [71,72,90,108,109,116,121]. The chapter importantly highlighted that there are currently no sequence-based allele frequency data for forensically relevant STRs for South African population groups. To leverage MPS within a forensic casework context in South Africa, sequence-based population data from the reference population is required.

South Africa's current population stands at 62 million individuals belonging to a diverse range of ethnic groups. The two most prevalent population groups in South Africa are the Black African group, making up 81% of the total population, followed by the South African Admixed population group, making up 8.2% [185]. The Black African population group is made up of individuals with primarily African descent but may be differentiated with respect to their specific ancestral and ethno-linguistic groups [177]. To enable maximum local relevance and translation to the police and forensic casework, individuals in the Black African population group were grouped into one major group, as per previous South African population studies [164,186,187].

Regarding the Admixed population group, the commonly held belief is that the heritage of the "Coloured" population, herein referred to as "Admixed" originates from the historical encounters between European settlers and the indigenous Khoisan groups of the Cape region. From these early encounters, subsequent generations of mixed heritage were formed through

relationships between the settlers, the Khoisan, and enslaved individuals brought from the East, culminating in the establishment of a community with varied racial and ethnic roots [188].

Chapter 2 also revealed that much of the variation held in MPS data lie within populations with African ancestry, although Africa is the continent in which no MPS population study has been conducted. It is therefore hypothesised that the South African population will exhibit variation that exceeds what has been previously revealed in non-African population groups. The aim of this chapter was thus to; 1) explore sequence variation of A-STR loci (as justified under the study design in Chapter 1), 2) generate a sequence-based allele frequency database using the ForenSeq™ DNA Signature Prep kit using an optimised approach for crude buccal swab lysates, as established in Chapter 3, and 3) evaluate concordance with CE methods for South African population Admixed and Black African population groups.

4.2. Methods

4.2.1. Samples

Buccal swab and whole blood samples were previously collected with informed consent from a total of 463 individuals: Black African (n = 216) and Admixed (n = 247). Samples were collected as part of an umbrella study [164]. Institutional ethics approval was obtained for the use of these samples in the current study (HREC: 400/2021). The methods used were carried out in accordance with the latest guidelines published by the International Society of Forensic Genetics (ISFG) [189].

4.2.2. Sample preparation and processing

Manufacturer's protocols were used throughout this chapter, except where modifications are specified. Crude buccal swab lysates had been prepared from 224 buccal swabs using the

SwabSolution™ Kit (Promega, Madison, WI, USA) and from 134 buccal swabs using the STR GO! Lysis buffer (QIAGEN, Hilden, Germany) [165,166]. Crude buccal swab lysates processed with the STR GO! Lysis buffer (QIAGEN, Hilden, Germany) with the manufacturer's protocol, although those that had shown poor amplification success (n = 95) with the ForenSeq™ DNA Signature Prep Kit (Verogen, San Diego, CA, USA) in an initial testing phase were purified using the QiaAmp® DNA Investigator Kit (QIAGEN, Hilden, Germany), as described under Modification 1, **Table 3.1** (Chapter 3) [190]. Whole blood samples (n = 105) were collected and extracted using a standard salting-out procedure as outlined in Miller *et al.*, 1988 in a previous study (HREC: 445/2015), and ethical approval was obtained for the use of these samples in the current study (HREC: 400/2021) [191]. All extracted DNA samples were quantified using the Quantifiler® Trio DNA Quantification Kit on the Applied Biosystems 7500 Fast Real-Time PCR System (Applied Biosystems, Foster City, USA) and diluted to 0.2 ng/µL using nuclease-free water. The concentration of diluted samples was confirmed using Qubit™ fluorometry with the Qubit™ dsDNA High Sensitivity Assay (Invitrogen, CA, USA) [192].

4.2.3. Library Preparation with the ForenSeq™ DNA Signature Prep kit

Libraries were prepared from samples using the ForenSeq™ DNA Signature Prep Kit (Verogen, San Diego, CA), with Primer Mix B targeting both identity loci as well as biogeographic ancestry and phenotype markers [21]. Although all data were generated for all the loci targeted in Primer Mix B, this study focused on A-STRs, as justified under the study design in Chapter 1, section 1.4. The manufacturer's protocol was adhered to, with modifications to crude buccal swab lysate processing steps as follows: 2 µL SwabSolution™ (Promega Corporation, Madison, WI, USA) crude buccal swab lysate was used as template with 3 µL 5X AmpSolution® instead of water, as informed by internal optimisation

experiments (Chapter 3). For extracted DNA samples and purified lysates, 1 ng of template was used [21]. Thermal cycling was performed on a Bio-Rad T100 Thermal Cycler (Bio-Rad, Hercules, CA). Libraries were purified using Agencourt AMPure XP beads (Beckman Coulter, Brea, CA), normalised, and pooled in batches ranging from 32-96 samples before sequencing. Volumes of pooled libraries ranging between 8 µL and 12 µL were denatured with NaOH, diluted in Hybridization Buffer, and spiked with a 4 µL Human Sequencing Control (HSC) mixture. Pooled, denatured and diluted libraries were loaded onto MiSeq FGx™ Reagent cartridges and sequenced on the MiSeq FGx™ System (Illumina, San Diego, USA) in Forensic Genomics mode [59].

4.2.4. Data analysis

An initial analysis of the sequenced libraries was performed using ForenSeq™ Universal Analysis Software (UAS) version 1.3 (Verogen, San Diego, USA) with default settings of a 1.5% analytical threshold (AT) and 4.5% interpretation threshold (IT) for allele calling [60]. Flanking region reports were exported from UAS as Microsoft® Excel files for each sequencing run. Flanking region reports were filtered to remove stutter, remove erroneous sequences and formatted to group alleles according to marker type to facilitate variant characterisation.

4.2.5. Concordance with length-based genotypes

After removal of stutter and erroneous sequences, sequence-based genotypes were compared to CE genotypes generated with the QIAGEN Investigator 24plex GO! Kit (QIAGEN, Hilden, Germany) for a subset of 229 samples that were previously published [164]. Concordance was assessed for 20 overlapping autosomal STR markers. Alleles were grouped into three

categories: a) discordant, b) concordant and c) ambiguous (*i.e.*, genotype undetermined due to imbalance or allele dropout).

Where an allele was noted to be discordant, samples were re-processed using the GlobalFiler® PCR Amplification Kit (Applied Biosystems, Foster City, USA), due to the extended allelic ranges that the kit provides for some markers [16]. Amplification of target region was performed using the Bio-Rad T100 Thermal Cycler (Bio-Rad, Hercules, CA), while post-PCR fragment separation was performed on the ABI 3130xL Genetic Analyser (Applied Biosystems, Foster City, CA, USA) using POP-4™ polymer (Applied Biosystems, Foster City, USA). Allele sizing was facilitated by the GeneScan™ 600 LIZ size standard (Applied Biosystems, Foster City, CA, USA). The electropherograms were analysed using the GeneMapper® ID-X software version 1.5 (Applied Biosystems, Foster City, CA, USA).

4.2.6. Short-hand naming

Sequences were characterised using a naming system developed by Kings College London [116]. The naming system consists of sequences found in five different British population groups [116]. Each unique sequence has been named according to a) the length-based allele, which was represented as the length-based allele itself, b) a sequence annotation, which used a new number each time new variation is observed within the repeat region of the sequence (e.g., S1 for the first sequence variation observed for a length-based allele, S2 for the second sequence variation observed, etc) and c) a version number, which denoted version of variation in the flanking region of the sequence-based allele (e.g., V1 for the first flanking region variation observed for a specific length-based allele, V2 for the second flanking region variation observed).

To facilitate inter-laboratory data comparison, a lookup of the FASTA sequences seen at each marker in each South African population group was performed in Microsoft® Excel, and where an allele name was not retrieved in any of the sequences in the British population groups, (*i.e.*, the VLOOKUP resulted in a "#N/A"), these alleles were investigated for novelty. When a sequence *was* previously seen in the British population groups, the sequence was assigned a shorthand name as described above, and the sequence divided into its repeat motifs (Figure 4.1). The use of the Microsoft® Excel® naming system enabled the visualisation of sequence variants by repeat regions (**Figure 4.1**). All alleles were initially visualised using this system to characterise repeat region variants and to name alleles according to the shorthand sequence and bracketed repeat notation. This system has been developed for A-STRs, but is currently under development regarding X-STR, Y-STR and iiSNV markers (Personal Communication, David Ballard).

Flanking sequence	Length-based allele	Allele name	Sequence Breakdown													
CAGAGAGAA	13	v1_13_S1	AGAT	AGAT	AGAT	AGAT	AGAT	AGAT	AGAT	AGAC	AGAC	AGAC	AGAC	AGAC	AGAT	
CAGAGAGAA	14	v1_14_S1	AGAT	AGAT	AGAT	AGAT	AGAT	AGAT	AGAC	AGAC	AGAC	AGAC	AGAC	AGAC	AGAC	AGAT
CAGAGAGAA	14	v1_14_S2	AGAT	AGAT	AGAT	AGAT	AGAT	AGAT	AGAT	AGAC	AGAC	AGAC	AGAC	AGAC	AGAC	AGAT

Figure 4.1: *Snippet of Excel-based naming system developed by Kings College London (Kings Forensics) used to assign the shorthand sequence annotation and visualise repeat region variation.*

4.2.7. Bracketed repeat nomenclature

To systematically count, group and annotate consecutive tetranucleotide A-STRs, a custom function was implemented in RStudio v4.2.2 (RStudio Team, 2020) [100]. The repeat motifs that were broken down in Microsoft® Excel were used as an input file for bracketed repeat generation and converted into a vector for further processing. The first step in the function was to identify unique repeat motifs and count their occurrences. When a repeat was encountered more than once, it was annotated by enclosing the motif in square brackets, followed by the number of repeats. For example, seven consecutive occurrences of "AGAT" were annotated as

[AGAT]7. Additionally, one occurrence of a motif was retained without annotation. The same logic was applied to compound and complex repeats, with an additional step to concatenate all annotated repeats for a sequence into a single string. For example, seven instances of "AGAT" would be combined with five instances of "AGAC", followed by one instance of "AGAT", resulting in [AGAT]7 [AGAC]5 AGAT. A full description of the function is provided in Appendix 4.1. This function was applied to all the repeat motifs for each A-STR marker. The output was stored in spreadsheet format and used in subsequent allele reporting. Sequence-based alleles were recorded and stored, alongside the length-based allele, bracketed repeat nomenclature, the shorthand sequence annotation described in section 4.2.6, flanking region variants, 5′ and 3′ flanking sequences and FASTA sequences.

4.2.8. Variant characterisation

All sequences were aligned with the Forensic STR Sequence Structure Guide version 5 (FSSG v5, available at https://STRidER.online) to name and identify sequence variants [32,69]. All sequence strings were split using a Microsoft® Excel function to separate nucleotides such that each nucleotide was contained within a separate cell. Conditional formatting was then used to align data to the FSSG v5 to visualise variants [32]. Variants that were initially identified for novelty were subjected to a BLAST search for comparison to sequences housed in the STRSeq BioProject, according to the process outlined in Gettings *et al.* 2017 [193,194]. Alleles that did not result in 100% identity with another sequence in the STRSeq BioProject were reported as novel alleles, and chromosomal position and SNV information were noted for these alleles.

4.2.9. Statistical analysis

Allele frequencies were determined using STRAF v2.6, while F_{ST} values and assessment of departure from Hardy Weinberg Equilibrium (HWE) were determined using Arlequin v3.5.2.2

[195,196]. Forensic and population statistical parameters were also calculated for sequence-based alleles using STRAF v2.6 [196]. These included statistics for Genetic Diversity (GD), Polymorphic Information Content (PIC), Observed Heterozygosity (Hobs), Match Probability (MP). The RMP values were manually calculated from MP statistics using the product rule in Microsoft® Excel for each population group. Testing for departure from Hardy Weinberg Equilibrium (HWE) included an adjusted significance level of 0.002 after applying a Bonferroni correction for multiple comparisons for each population group.

4.2.10. Quality control

Length and sequence-based genotype data were submitted to STRs for Identity ENFSI Reference Database (STRidER) to verify the quality of the data [32]. Prior to submission, sequence-based data were checked to ensure that no duplicates of samples existed. As samples were collected from random, reportedly unrelated, individuals in the population, tests for genetic relatedness were not performed. Samples with tri-allelic genotypes were removed from the dataset prior to submission. Additionally, as no missing data was allowed in STRidER submissions, the D22S1045, Penta D and Penta E markers were removed from the dataset due to the high levels of allele dropout exhibited at these markers. Both length and sequence-based genotype data were formatted according to STRidER guidelines and submitted with the following accession numbers for the Black African and Admixed population groups, respectively: STR000421 and STR000422 [32].

4.3. Results

4.3.1. Quality metrics

Samples used for the purposes of establishing the allele frequency database were sequenced in 24 batches, which each included a wide range of sample types. The quality metrics were within adequate ranges for cluster density, phasing and prephasing for all sequencing runs (Appendix 4.2) [59]. The clusters passing filter for eleven runs did not meet the recommended chastity filter and ranged between 56% and 79%. Investigation into the low clusters passing filter was undertaken with the manufacturer and their recommendations indicated that usable data can still be generated for runs that did not fully meet the recommended ranges for quality metrics. Indeed, the data generated from these eleven runs resulted in sufficient coverage and sequence data, as evidenced by the performance of the 2800M control DNA and the HSC, as well as the high call rates obtained for population study samples.

4.3.2. Allele frequencies and sequence variation

The allele frequencies for 24 A-STR markers are reported in Appendix 4.3. A shorthand annotation which includes the length-based allele, the level of repeat region variation and the level of flanking region variation is shown within the dataset, alongside the sequence-based allele frequencies. Additional information on all sequences, including the bracketed repeat nomenclature, flanking region variants and shorthand sequence annotations are provided for compatibility with other published sequenced data. The FASTA sequences have not been reported as part of this thesis due to physical restrictions of the document processor (*i.e.*, Microsoft Word®) but will be included as part of the submission of the population study for publication, as supplementary data.

In the Black African population group, a total of 295 length-based alleles and 500 sequence-based alleles were observed. An increase of 70% was observed in the number of alleles when accounting for sequence variation. Inclusion of flanking region variants increased the number of alleles to 552, an 87.11% increase from length-based alleles alone and a 10.4% increase from sequence-based alleles. In the Admixed population group, 327 different length-based alleles were observed, whereas 591 sequence-based alleles were observed. Including flanking region variation increased the number of unique alleles to 650, indicating a 98.78% increase from length-based alleles and a 9.98% increase from sequence-based alleles alone (**Figure 4.2**).

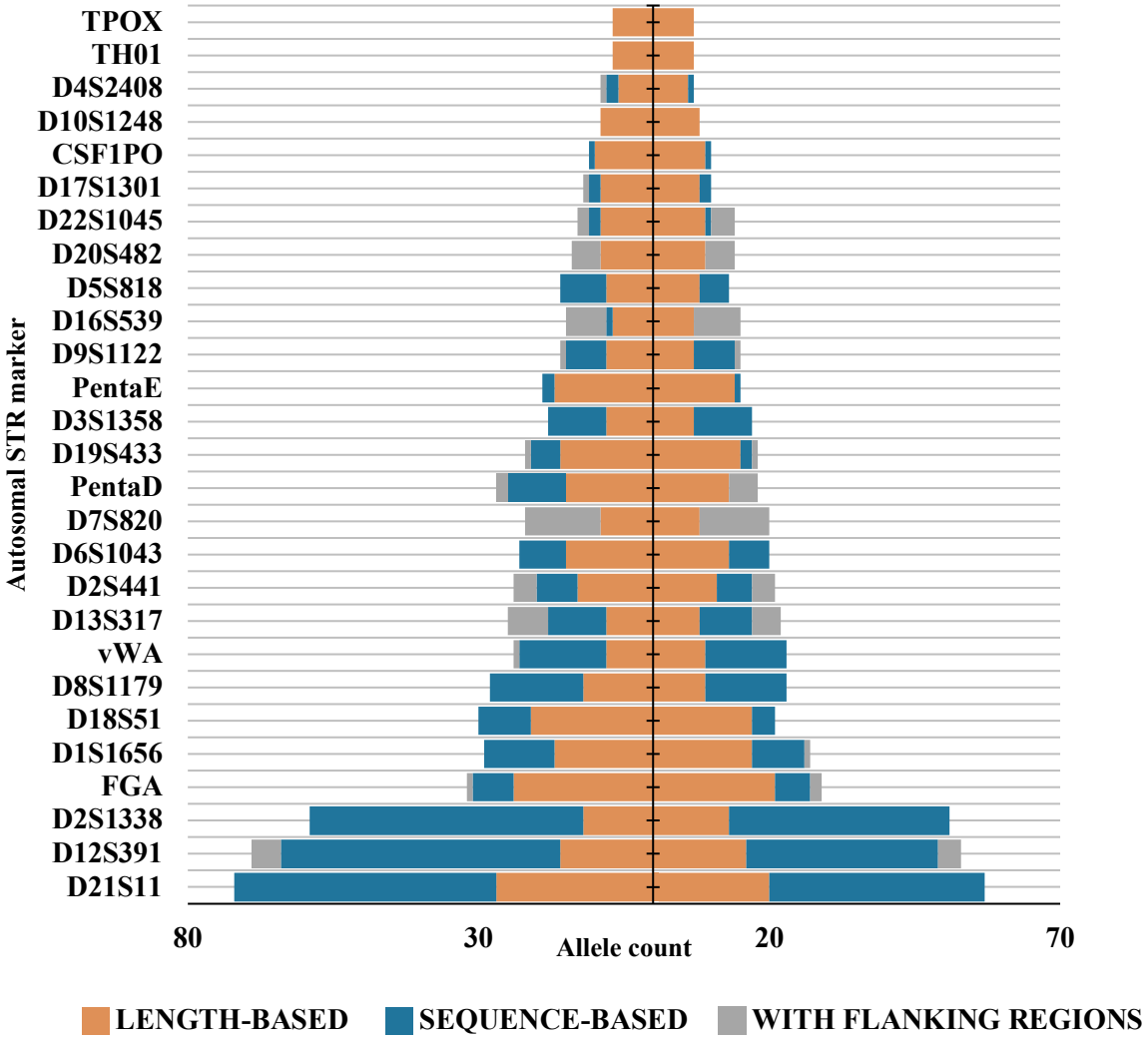

Figure 4.2: *Stacked column chart representing allele counts for length-based, sequence-based and sequence-based alleles with flanking regions included for 27 autosomal STR markers and for the Admixed (left) and Black African (right) population groups.*

For both population groups the two markers showing the highest degree of variation by sequence were the D21S11 and D12S391 markers. The length-based allele counts obtained for the D21S11 marker was 19 in the Black African and 27 in Admixed population groups whereas sequence-based allele counts increased to 56 and 72 in the Black African and Admixed populations respectively. No additional alleles were observed when including flanking region sequences. Singleton sequences (*i.e.*, only seen once in the combined population dataset accounted for 9/27 of the unique length-based alleles observed at the D21S11 marker. In comparing the length- and sequence-based allele frequencies obtained for this marker, allele 28 can be used as an example. The allele 28 occurred at a frequency of 0.206 in the Admixed population group, but four different variations of allele 28 was observed in sequence-based data at lower frequencies. These variations occurred within in the repeat region of this complex marker. In the Black African population group, allele 28 occurred at a frequency of 0.3, while repeat-region variation resulted in three additional alleles observed at lower frequencies. The D12S391 marker, also found to be highly polymorphic, resulted in 16 length-based alleles in both the Black African and Admixed population groups, 49 and 64 sequence-based alleles, and 53 and 69 sequence-based alleles with flanking region variants accounted for, for the Black African and Admixed population groups, respectively.

Markers showing the least amount of variation included TPOX, TH01 and D10S1248, with no sequence variation observed in the repeat motif or the flanking regions of these markers. At the TPOX locus, seven tri-allelic genotype combinations in 12 individuals were observed, all concordant with CE genotypes (**Table 4.1**). Each of the tri-allelic genotypes observed included the allele 10, which was discussed in Heathfield *et al*., 2024 [164]. There were rare variants in traditionally conserved markers: a rare microvariant with length 12.1 was observed at the

CSF1PO marker, and flanking region variant was observed in the highly conserved D4S2408 marker.

Table 4.1: *Seven tri-allelic genotype combinations observed in the South African Admixed and Black African population groups in 12 individuals.*

Length-based genotype	Sequence-based genotype	Observations
6, 8, 10	[AGAT]6, [AGAT]8, [AGAT]10	1
6, 9, 10	[AGAT]6, [AGAT]9, [AGAT]10	2
6, 10, 11	[AGAT]6, [AGAT]10, [AGAT]11	2
7, 10, 11	[AGAT]7, [AGAT]10, [AGAT]11	1
8, 9, 10	[AGAT]8, [AGAT]9, [AGAT]10	3
8, 10, 11	[AGAT]8, [AGAT]10, [AGAT]11	1
9, 10, 11	[AGAT]9, [AGAT]10, [AGAT]11	2

Variations in the flanking regions also increased the number of alleles. For example, variation in the FGA marker has primarily fallen within the repeat region, while in this study, two instances of flanking region variations occurred and were found to be novel, showcasing the first record of flanking region variation for this marker. Two markers, namely D7S820 and D20S842, showed no variation in the repeat regions, but the inclusion of flanking region variants increased the number of alleles observed at these markers. In the D7S820 marker, a SNV present in the 5′ flanking region (rs7789995 T>A) was observed in several length-based alleles for this marker, along with a 3′ flanking region SNV (rs16887642 G>A) resulting in an increase in allele counts with inclusion of flanking region sequences. The increase in the number of alleles in the D20S842 marker was also largely due to the presence of a 5′ flanking region SNV (rs77560248 C>T). **Table 4.2** lists all known flanking region variants observed in the dataset for the South African populations.

Table 4.2: *Known flanking region variants for 11 A-STR markers are shown with their corresponding start-stop coordinates, chromosome number and frequency in the South African Admixed and Black African population groups.*

Marker	Chromosome	rs number	Coordinate start	Coordinate stop	Variant	Type	Frequency in Admixed (2N = 494)	Frequency in Black African (2N = 432)
D2S441	2	rs74640515	68011922	68011922	G>A	Substitution	0.04453	0.00202
D5S818	5	rs73801920	123775552	123775552	C>A	Substitution	0.25911	0.26721
D7S820	7	rs7789995	84160204	84160204	T>A	Substitution	0.93927	0.05668
		rs897512434	84160203	84160203	[T/-]	Deletion	0.00202	-
		rs16887642	84160286	84160286	G>A	Substitution	0.14575	0.10324
D9S1122	9	rs149309595	77073812	77073812	T>C	Substitution	0.00202	0.00202
D12S391	12	rs138635218	12297179	12297179	C>G	Substitution	0.02834	0.02632
VWA	12	rs75219269	5983970	5983970	A>G	Substitution	0.09312	0.04453
D13S317	13	rs73250432	82148000	82148000	C>T	Substitution	0.00405	-
		rs146621667	82148001	82148001	G>A	Substitution	0.00810	0.02024
		rs561167308	82148090	82148093	[TCTG/-]	Deletion	0.00810	0.00607
		rs1442523705	82148077	82148080	[ATCT/-]	Deletion	0.00202	-
		rs9546005	82148069	82148069	A>T	Substitution	0.42713	0.31984
		rs202043589	82148073	82148073	A>T	Substitution	0.00607	-
D16S539	16	rs563997442	86352692	86352692	C>G	Substitution	0.00810	-
		rs11642858	86352761	86352761	A>G	Substitution	0.23887	0.17206
		rs114697632	86352779	86352779	T>C	Substitution	0.00607	0.01215
D18S51	18	rs535823682	63281740	63281740	A>G	Substitution	0.03441	0.04858
D19S433	19	rs745607776	29926229	29926230	[CT/-]	Deletion	0.00202	-
D20S482	20	rs77560248	4525680	4525680	C>T	Substitution	0.08097	0.10324

4.3.4. Novel alleles

A total of 80 novel alleles were observed among the two South African population groups studied, with 16 shared between the two population groups and across 15 different A-STR markers Appendix 4.4 and 4.5). More specifically, the Admixed population presented with a total of 43 novel alleles that were population specific, while the Black African population consisted of 21 population specific novel alleles. Markers consisting of novel flanking region variants are summarised with start-stop positions and variant frequencies in (Appendix 4.4).

The D21S11 marker resulted in the most novel alleles (23), followed by the D12S391 marker which resulted 10 novel alleles. The FGA marker consisted of three novel microvariants, four novel repeat region alleles and two novel flanking region variants.

4.3.5. Forensic and population statistics

The allele frequencies did not depart from HWE at any loci, and upon assessment of population differentiation between the Black African and Admixed population groups, an F_{ST} value of 0.0079 ($p < 0.0001$) was obtained. This indicated little population differentiation between the two population groups. Assessment of population differentiation using length-based data for the same population groups was assessed in Heathfield et al., 2024, and also showed little genetic differentiation between the Admixed and Black African population groups [164]. In both population groups, the markers showing the highest levels heterozygosity and genetic diversity were D12S391 and D2S1339. In the Admixed population group, the D13S317 showed the highest *increase* in heterozygosity from length to sequence-based alleles (18.47%), while in the Black African population group, it was the D9S122 marker (15.38%) (Appendix 4.6). Similar to gains in heterozygosity, the D13S317 and the D9S122 markers showed the highest gains in genetic diversity from the Admixed and Black African population groups, respectively. Length-based data revealed an RMP value of 3.88E-28 for the Black African population group across STR markers, and a drastic reduction to 4.17E-33 was observed when assessing sequence-based data including flanking regions (Appendix 4.7). An even larger decrease in the RMP value from length- to sequence-based data with flanking regions was observed for the Admixed population group from 3.04E-29 to 1.93E-34.

4.3.6. Concordance assessment

An overall concordance rate of 99.10% was achieved across all markers, with markers D7S820 and D1S1656 showing potentially discordant or ambiguous genotypes (Appendix 4.8). However, both could be resolved through additional review and analysis, which will be described below (section 4.3.6.1). Individual instances of allele dropout were also observed in D5S818, D7S820, D16S539 and D19S433 while several instances of dropout were observed in D22S1045, Penta D and Penta E, which were excluded from allele frequency and forensic parameter calculations. Instances of incorrect repeat size estimation was observed for the D7S820 and D13S317 markers, flanking region polymorphisms, also which will be described below.

4.3.6.1. Flanking region variant impact on repeat size estimation

A discordance was observed in one sample in the South African Admixed population group at the D7S820 marker upon comparison to CE genotypes. The ForenSeqTM UAS genotype was "10, 11", whereas a "10, 10.3" was consistently obtained with CE using the Investigator 24Plex GO! Kit and the GlobalFiler® PCR assay. Investigation into the full sequence, including the flanking region of UAS allele "11" in this sample revealed a deletion in the 5' flanking region of the amplicon. The deletion (rs897512434) resulted in a CE length of 10.3. The deletion also resulted in the extension of the minimum sequence range for this marker (**Figure 4.3**). The software subsequently included 1 bp of the primer in the sequence output, violating the minimum sequence range, as detected during STRidER quality control checks (shown in grey in **Figure 4.3**). The forward sequence is shown to enable comparison to the reference sequence.

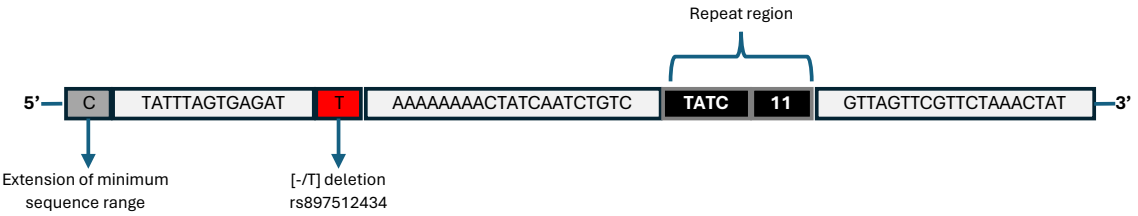

Figure 4.3: *The 5'-3' sequence breakdown for an allele 11 present in the D7S820 marker that resulted in a CE-discordance is shown. The "C" is coloured in grey in the 5' flank. The deletion of a "T" nucleotide is shown in the red box (rs897512434). Flanking regions are shown in in white text boxes, while the repeat region is shown in a black text box.*

Upon alignment to reference sequences reported the FSSG v5, several repeat region and flanking region variants were observed in the D13S317. The D13S317 marker consists of a simple tetranucleotide repeat motif [TATC]. Submission of sequences to STRidER for quality control revealed several instances where the number of repeats derived from the sequence did not match the length of the allele called by the UAS. Five samples in the South African Admixed population were noted to have a different number of repeats compared to the UAS-derived length. For example, where a UAS-called allele was originally "9", when counting the "TATC" repeats for this allele, 10 "TATC" repeats were present. This discrepancy was also seen in three instances in the South African Black population group. Investigation into the flanking region sequences of these alleles revealed flanking region variation that were flagged during STRidER quality control checks. The first was in a sample from the South African Admixed population, where an allele was called as 10 by the UAS, but upon characterisation of the sequence, an A>T substitution (rs9546005) in the flanking region was noted and resulted in the extension of the "TATC" repeat (**Figure 4.4**). In the same sample, an insertion/deletion (InDel) was found 20bp upstream of the first SNV, which was a "TCTG" deletion (rs561167308) that has been previously characterised. The A>T substitution, in combination

with the InDel results in the extension of the minimum sequence range and the misrepresentation of the number of TATC repeats in the sequence. The same combination of flanking region variation was present in three other samples in the South African Admixed population, and three samples in the South African Black population, which caused a discrepancy to occur between the number of repeats in the sequence and the length derived from the sequence.

In one other sample from the South African admixed population, a sample with the same A>T substitution in the flanking region as mentioned above (rs9546005), presented with an "ATCT" deletion in the upstream flanking region (rs14425237035). The influence of this 4 bp-deletion along with the A>T substitution resulted in a sequence, which when the repeat units were manually derived from the sequence, a length of 10 was obtained, however, the UAS length-based allele (9) correctly matched the CE length of 9.

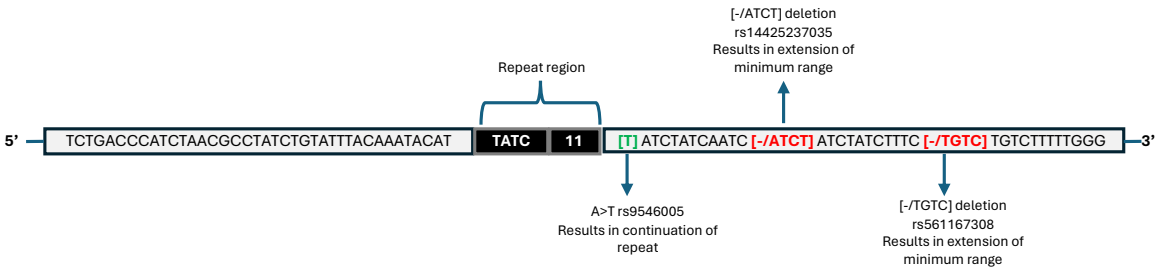

Figure 4.4: *The 5'-3' sequence breakdown for an allele 11 in the D13S317 marker is shown. The flanking regions are represented by white text boxes to the left and right of the repeat region, represented by a black text box. The substitution (rs9546005 A>T) is shown in green in square brackets. The 4 bp deletions (rs14425237035 and rs561167308) are shown in red in square brackets.*

4.3.6.2. Alleged discordance at D1S1656

One sample in the South African Admixed population at the D1S1656 marker sequenced with the ForenSeq™ DNA Signature Prep kit revealed a heterozygous "8, 16" whereas CE methods resulted in a homozygous genotype "16, 16" with the Investigator® 24Plex GO! Kit (QIAGEN,

Hilden Germany). Upon evaluation of the Investigator® 24Plex GO! electropherogram, an additional corresponding to the 8-allele superimposed on the DYS391 marker, as described in Heathfield *et al*., 2024 [164]. This superimposition was due to the limited allelic range for the Investigator® 24Plex GO! kit (10 – 20.3). When re-typed with the GlobalFiler® PCR Amplification kit (Applied Biosystems, Foster City, USA), a heterozygous "8, 16" was observed, concordant with MPS. The genotype was therefore concordant with MPS, as both systems detected the same length-based allele. In the same sample, flanking region variation consisting of a "CCTA" deletion at the 5′ start of the repeat region had resulted in 8 repeats, and this allele was therefore outside the allelic ladder ranges for both the Investigator® 24Plex GO! kit and the GlobalFiler® PCR Amplification kit.

4.4. Discussion

The aim of this chapter was to establish allele frequency data, characterise STR sequences generated with the ForenSeq™ DNA Signature Prep kit for the South African population and establish concordance with CE genotypes. An important overall finding of this study was the rich diversity introduced through the inclusion of sequence-based data, as seen through assessment of the increase in the number of alleles obtained by sequence, the unexpected variation in conserved markers and the presence of several novel alleles. To leverage the increase in sensitivity provided through the detection of additional variation when incorporating MPS methods, the data needs to comply with internationally suggested recommendations with respect to standardisation and back-compatibility. An important outcome of this study was thus the reporting of sequence data characterised according to established nomenclature guidelines to contribute to global nomenclature standardisation efforts. Finally, concordance rates exceeding 99% were achieved in this population study, whereas several flanking region indels were characterised that caused incorrect length-based

allele counting by the ForenSeq™ UAS, bringing into question the true concordance status of our population dataset.

4.4.1. Richness in variation

A notable increase in the number of alleles obtained by sequence for both South African populations was observed in this dataset. More specifically, a 70% and 80% increase was observed for the Black African and Admixed population, respectively. A slightly higher increase in the number of alleles were observed when including flanking regions (**Figure 4.2**). Nonetheless, the higher number of alleles observed by sequence for South African populations, and more specifically the Admixed population, is largely because populations in Africa have higher levels of genetic diversity [89].

Chapter 2 revealed that African populations exhibited great variation within population groups, even across different regions, a finding strongly supported by non-forensic genetic studies [89,131]. More importantly, African population groups were found to be highly under-represented in genetic population studies [89]. This has implications for forensic casework laboratories, and more specifically those that generate reference population data for determining match probabilities.

Highly polymorphic markers observed in this study included the D12S391, D2S13338 and D21S11 markers, in both population groups. Several studies have reported high genetic diversity and increase in number of alleles by sequence for these markers [71-73,115,117,197]. These markers consist of compounding repeat units, with much of the variation arising from the varying repeat counts within the sequence. The STRSeq project contains sequence data arising from several different population groups with sequences for over 4500 individuals, where 172 different alleles have been characterised for the D12S391 marker, 111 for the

D2S1338 marker and 192 for the D21S11 marker [194]. In this study, 79, 66 and 86 different alleles were characterised for the D12S391, D2S1338 and D21S11 markers, respectively, from only 463 individuals.

For highly polymorphic markers such as D12S291, a sample size of 200 may not be sufficient to capture and identify the variation present at this marker [117,124]. This is especially true for genetically diverse African population groups. The same can be assumed for the D2S1338 and D21S11 markers [124]. Implications for not capturing the maximum breadth of variation in more diverse population groups, especially at highly polymorphic markers is match statistics and estimates may be inaccurate. However, this implication may not be entirely relevant in forensic casework where MPS is used, as the increase in the number of markers notably improves the discriminatory power, without yet exploring sequence-level discrimination.

The CSF1PO marker is known to be a highly conserved marker, as reported in several population studies [73,75,108,116,198]. There is little sequence variation within the repeat and flanking regions of this marker. Additionally, alleles typed for this marker have ranged between 5 and 17. In this study, a rare microvariant allele, 12.1, was observed in the South African Admixed population (Appendix 4.4), whose sequence has not previously been characterised in MPS studies. The length-based allele has been observed in three individuals in the Navajo population in a study conducted by Budowle *et al.*, 2001 [199]. The CSF1PO marker may provide little informativeness in complex DNA profile interpretation such as mixtures, however, previous studies have not characterised sequences for this marker from more diverse populations, such as those in Africa. Further unexpected and novel repeat and flanking region variants were observed in the FGA marker. To the best of our knowledge, no flanking region

variation has been previously characterised for the FGA marker. This finding in this study demonstrates the richness of genetic variation observed in the South African population.

A total of 80 novel alleles were identified in this study, and 15 markers presented with novel alleles in both the repeat and flanking regions (Appendix 4.4). It is expected that as more population studies are conducted, less novel alleles are described in later population studies, however, as there is limited population data for African populations, there is much still to characterised. In this study, most novel alleles were due to variations in the repeat region. For example, the D21S11 consists of a complex repeat structure, and several variations were observed in which the count of different repeats varied greatly within the sequence. Larger datasets of polymorphic loci from under-characterised population groups are needed to fully capture the number of isometric alleles at this marker [115].

Novel alleles by flanking region variation were not as common as those attributed by repeat sequence variation but contributed to the characterisation of 14 previously undescribed alleles. This finding is expected, given that flanking region variation only contributed to ~10% increase in the number of alleles obtained (in comparison to the allelic gain from length- to sequence-based alleles without flanking regions). Similar findings have been reported in the literature, where flanking region variation was assessed, and found to have low contributions to allelic gain, yet were useful in the resolution of discordances [73,86,108,124].

4.5. Evaluation of concordances and discordances

A concordance rate of 99.10% was achieved in this study, in accord with previous concordance assessments using the ForenSeq™ DNA Signature prep kit, including those determined by the developmental validation [20,61,84,90,198]. Although a high concordance rate was achieved for a subset of MPS markers that overlapped with the CE kit markers, it is not known whether

non-overlapping markers would result in similar concordance levels. Markers showing initial discordance could be resolved upon further analysis of flanking regions. The D7S820 marker discordance resulted from a deletion in the flanking region, causing both incorrect counting of repeats and the extension of the minimum sequence range (rs897512434) (**Figure 4.3**). This variant has been previously characterized in the Lebanese population at a frequency of 0.0026 (1/195), and at a frequency of 0.0024155 in the White British population group [117,200]. However, no discordance resulted from this variant in both the White British and Lebanese population datasets, as was observed in the South African Admixed population group. Although discordance between methodologies could result in interpretation challenges, these can be overcome through detailed reporting practices, where both the full sequence string is reported alongside the length-based allele.

The naming discrepancy observed at the D13S317 was due to miscounting of repeats by the ForenSeq™ UAS (**Figure 4.4**) and was resolvable through flanking region assessment. This miscounting was detected through the STRidER quality control process, which is not routinely performed by laboratories when generating sequence data for forensic casework. Submission of data to STRidER is only a requirement for publishing of allele frequency data. Absence of the same quality control process in a laboratory that uses MPS data for forensic casework means that instances of repeat miscounting by the ForenSeq™ UAS may go undetected, and ultimately result in a non-match at one or more markers. This calls for an update in the ForenSeq™ UAS algorithm with a new requirement to include an integrated quality control process for allele size estimation with consideration of flanking region indels, as their presence may influence repeat counting.

Without a quality control check to ensure that STR sequences are concordant with the ForenSeq™ UAS allele size estimate, reporting of flanking region data in both DNA profile and allele frequency databases becomes critical [69,117,124]. To contribute to the transparency and standardisation of STR sequence reporting practices, variants in the flanking region should easily identifiable, and even more so when it results in an instance of discordance [32]. However, consequences for including these data in DNA profile databases for determining matches would give rise to certain complications if DNA matching systems do not cater for a level of leniency. For example, an exact match would not be found due to the discrepancy in allele calling but if the system allowed for near to full matches, then this consequence could be overcome, as also suggested by Hölzl-Müller *et al.*, 2021 [197].

A second alleged discordance was observed in the D1S1656 marker, where the sequenced genotype (8, 16) was discordant with the CE genotype (16, 16) due to the limited locus range of the Investigator® 24Plex GO kit. Although the peak was present in the CE electropherogram, it superimposed on an adjacent marker (DYS391), causing incorrect haplotype calls. The discordance for this sample was previously documented by Heathfield *et al*, 2024 [164]. The implication of this limited allelic ladder and locus range is that an incorrect CE genotype may be inferred if sequencing is not used for confirmation. This underscores the importance of incorporating MPS in forensic casework, at least for confirmatory purposes, where a full transition is not feasible.

A final aspect of concordance was the ambiguous genotypes observed at problematic markers such as D22S1045, Penta D and Penta E. Several studies have reported allele dropout at these markers, as also observed in this study [20,74,116,198,201,202]. Reasons for consistently poor heterozygote balance and ultimately performance at the D22S1045 high salt/ion concentrations

in PCR 1 buffers. High heterozygous imbalance can also be caused by mutations occurring in the primer binding regions. This ultimately affects primer hybridisation and less copies of the targeted region, leading to failed amplification of an allele [203].

The high imbalance and allele dropout observed at these markers have implications for complex DNA profile interpretation such as in mixtures and degraded samples, but also in reference profile comparison, where it may be challenging to discern stutter artefacts from true alleles [121]. Even in good quality reference samples, allele dropout and high imbalance is expected in these markers; therefore, even if quantitative cut-offs are implemented (for non-direct PCR samples) as suggested by Foley *et al.*, 2024, there is a high possibility that dropout will occur in either good quality reference samples or casework samples, where sample quality is often compromised [204]. Similar to other studies, it was decided that these markers should be excluded from allele frequency calculations, as their inclusion would undoubtedly result in an increase in false homozygotes [20,198]. However, an implication of their exclusion is that the discriminatory power that could be leveraged through MPS was reduced. These markers require further finetuning in future MPS kits to harness the full potential of MPS for forensic human identification.

5. Conclusions

The findings from this chapter have established the first sequence-based allele frequency data for the South African population and demonstrated that a high level of concordance exists between MPS and CE methods. The generation of population data and the compatibility between MPS and CE are requirements for the implementation of MPS in our forensic laboratory, and the findings of this chapter thus contributes to fulfilling the aim of this thesis. This chapter also highlighted the disparities between manually derived lengths based on the

FASTA sequence and the automatic length calls made by the ForenSeq™ UAS in the presence of flanking region variants. The discussion called for integrated analysis software to consider flanking region variants in their allele naming scripts. The findings of this chapter serve as a launch pad for other African laboratories to adopt an MPS workflow using a collaborative approach to improve standardisation and increase the quality of data published. Finally, the generation of allele frequency data also fulfils a critical requirement of internal validation studies that need to be conducted to forward implementation.

Chapter 5: Internal validation study

Internal validation of the ForenSeq™ DNA Signature Prep workflow on the MiSeq FGx™ platform for human identification applications

5.1. Introduction

Implementation of a new forensic DNA analysis workflow or method requires adherence to quality standards. The ISO17025 standards, which established guidelines for competence in laboratory practices have been adapted to for forensic laboratories [25]. These adapted guidelines highlight the requirement for method validation for forensic testing laboratories [26,29,205]. This validation is essential for ensuring the that the workflow is fit for its purpose and legally defensible. The ForenSeq™ DNA Signature Prep kit has been developmentally validated on the MiSeq FGx™ benchtop sequencer and has been shown to overcome several limitations of CE methods, as explored in Chapter 1 [20]. The aim of this thesis was to facilitate implementation of this workflow in our forensic mortuary-linked laboratory, of which a major requirement was its internal validation. as it verifies that the MPS workflows meet the necessary quality benchmarks for implementation.

This component of the overall study thus encompassed the internal validation of the ForenSeq™ DNA Signature prep workflow on the MiSeq FGx™ platform (DPMB) implemented in the Biomedical Forensic Science laboratory at the University of Cape Town. To this end, the developmental validation study was used as a blueprint to design and carry out equivalent experiments to assess the performance parameters of the sequencing workflow in our laboratory. These include the assessment of the workflow in terms of accuracy, call rate, precision, sensitivity, repeatability and reproducibility and species specificity on a range of sample types. The samples ranged from good quality control samples to compromised samples, such as degraded DNA and those containing PCR inhibitors. Studies have evaluated the use of

the workflow using a range of sample types, but published information on the performance of post-mortem samples, and more specifically crude buccal swab lysates is lacking and needed. This internal validation therefore aimed to establish performance parameters for both control DNA samples, authentic forensic post-mortem samples, and crude buccal swab lysates for the first time. While the developmental validation includes mixture studies, this internal validation focused on single-source post-mortem samples. This is because single-source samples are most commonly encountered in the mortuary setting. The findings are scrutinised against internally established acceptance criteria, and based on these results, the workflow is assessed to determine if is fit for forensic human identification purposes. The study was designed and carried out according to SWGDAM validation guidelines where possible and adhered to internally established quality standards [205]. To demonstrate the value, purpose and impact of the validates workflow, the last section of this chapter describes the application of the workflow to a forensic cold case, involving a severely decomposed human body.

5.2. Methods

5.2.1. Samples

Control DNA samples, including the 2800M positive amplification control (Promega Corporation, Madison, WI, USA) and the Standard Reference Material (SRM) 2372a (NIST, MA, USA) were prepared through dilution with nuclease-free water to concentrations required for each internal validation study. The SRM 2372a (NIST, MA, USA) control DNA is often used as a quantification standard and was therefore diluted to concentration ranges used in the Quantifiler® Trio Kit (Applied Biosystems, Foster City, USA) for assessment of sensitivity and PCR inhibition. These were assessed for call rate, accuracy and precision with the ForenSeq^{TM} DNA Signature Prep kit (Verogen, San Diego, CA, USA) to enable qPCR-to-sequencing cross validation.

Forensic post-mortem samples previously collected from deceased individuals: a meta-carpal and molar sample, were previously collected as part of routine service delivery by the Forensic Pathology Services, and samples left over from primary analyses were approved to be used in this internal validation study, while nail, buccal swabs and FFPE tissue were obtained for ancillary investigation under three respective studies, HREC: 136/2021, HREC:342/2016 and HREC: 445/2015). Ethical approval was obtained for the collection and use of all samples used in this internal validation study from the University of Cape Town (HREC: R036/2014; HREC:400/2021, AEC: 021_010). All samples were prepared, and DNA was extracted using the methods described below, alongside reagent blanks to assess background contamination.

5.2.2. DNA extraction and preparation

5.2.2.1. Bone and teeth

The bone and tooth sample surfaces were decontaminated with bleach, ultra-pure water and 70% ethanol, followed by ultra-pure water again. The surface of the bone was sanded using a Dremel drill with a sanding band attachment. A fragment of the bone (~1 cm^3) was cut using a cutting band on the same drill. The tooth sample was prepared by removing of the enamel using the Dremel tool with the sanding band attachment. The hard tissue samples were placed in pre-cooled grinding jars with liquid nitrogen. The samples were ground into a fine powder using the Tissue Lyser II (QIAGEN, Hilden, Germany). Bone and tooth powder were transferred into 1.5 mL micro-centrifuge tubes and stored at -20°C until it was required for DNA extraction. Approximately 0.05 g input mass of the bone and tooth powdered samples were demineralised in a buffers solution with 1210 µL UltraPure™ 0.5 M, pH 8.0 ethylenediaminetetraacetic acid (EDTA) (Thermo Fisher Scientific, Waltham, MA, USA), 396 µL Buffer ATL (QIAGEN, Hilden, Germany), 22 µL proteinase K (QIAGEN, Hilden, Germany), 0.0121 g sodium dodecyl

sulphate (Sigma-Aldrich, MO, USA) and 22 µL of 1 M dithiothreitol (DTT) (Sigma-Aldrich, MO, USA) for 20 hours at 56 ℃ and shaken at 450 rotations per minute (rpm) in an Eppendorf ThermoMixer® C (Eppendorf, Hamburg, Germany). Thereafter, demineralised samples were purified using the QiaAmp DNA Investigator kit (QIAGEN, Hilden, Germany) according to the manufacturer's protocol, with the exception that an elution volume of 30 µL was used [190].

5.2.2.2. Nail

For removal of exogenous DNA, approximately 3 mg of the fingernail clipping was added to 1X phosphate-buffered saline (PBS), pH 7.4 (Invitrogen, MA, USA) and Tween®20 (Sigma-Aldrich, MO, USA) and incubated in an Eppendorf ThermoMixer® C (Eppendorf, Hamburg, Germany) at room temperature for 10 minutes, shaken at 750 rpm. The cleaned fingernail clipping was dried with paper towel and 2 mg of the clipping was placed in a 2.0 mL cap-lock micro-centrifuge tube. The nail clipping was incubated in a buffer solution containing 20 µL of 1 M DTT (Sigma-Aldrich, MO, USA), 300 µL Buffer ATL (QIAGEN, Hilden, Germany) and 20 µL proteinase K (QIAGEN, Hilden, Germany) at 56 ℃ for 16 hours, shake at 900 rpm, in an Eppendorf ThermoMixer® C (Eppendorf, Hamburg, Germany). Thereafter, DNA was extracted with the QiaAmp DNA Investigator kit (QIAGEN, Hilden, Germany), according to the manufacturer's protocol, with the exception that the elution volume was reduced to 30 µL.

5.2.2.3. Blood

Blood samples underwent DNA extraction using a standard salting out procedure as previously described in Chapter 4, section 4.2.2 [191]. Exceptions were made to the protocol in that the proteinase K digestion step was extended to 24 hours, and an elution was made into 300 µL Tris-EDTA.

5.2.2.4. Crude buccal swab lysates

Crude lysates were prepared from cotton buccal swabs in 1 mL SwabSolution™ (Promega Corporation, Madison, WI, USA), according to the manufacturer's protocol. The additional buccal sample that was collected using an EasiCollect™ Buccal Collection Kit (QIAGEN, Hilden, Germany) was punched in triplicate using a 1.2 mm Harris Uni-Core™ Punch kit (Sigma-Aldrich). These punches were treated with 100 µL 1X Tris-EDTA buffer, as recommended by the ForenSeq™ DNA Signature prep kit manual (Verogen, San Diego, CA, USA) prior to library preparation.

5.2.2.5. FFPE lung tissue sample

The deparaffinised lung tissue section underwent DNA extraction using the QiaAmp® DNA FFPE Tissue Kit (QIAGEN, Hilden, Germany) and was eluted into 50 µL [206].

5.2.2.6. Non-human samples

Assessment of the workflow for species specificity consisted of collection of buccal swabs from a domestic cat (*Felis catus)* and dog (*Canis lupus familiaris).* Meat from domesticated chicken (*Gallus domesticus)* and domesticated cow (*Bos taurus)* were purchased from a butchery and then swabbed. Blood was collected from a Rhesus macaque (*Macaca mulatta*) by a trained veterinarian. DNA was extracted from swabs and blood with the QiaAmp® DNA Investigator kit according to the manufacturer's protocol (QIAGEN, Hilden, Germany) with no deviations.

5.2.3. DNA Quantification

All extracted DNA samples, and control DNA samples were quantified with the Quantifiler® Trio kit (Applied Biosystems, Foster City, USA) on the 7500 real time thermal cycler, as described in Chapter 3, section 3.2.3.1. Extracted DNA samples and positive amplification control samples were diluted to amounts listed in **Table 5.1**, and dependent on the type of internal validation study. The Table 5.1 also lists the preparation and processing details, as well as how many replicates of each sample was used in each study. The concentrations were verified using the Qubit™ dsDNA High-Sensitivity Assay on the Qubit™ 4 Fluorometer (Thermo Fisher Scientific, Waltham, MA, USA) prior to library preparation and sequencing.

Table 5.1: *Study-specific preparation and processing details*

Study type	Sample	Preparation and processing details
Sensitivity	2800M control DNA (Promega Corporation, Madison, WI, USA)	Diluted to 1 ng, 0.25 ng, 0.0625 ng and 0.01525 ng and processed in triplicate.
	SRM 2372a (NIST, USA) quantification standard	Diluted to 2.5 ng, 0.25 ng, 0.025 ng and 0.0025 ng and processed in triplicate.
	Blood (extracted DNA)	Diluted to 1 ng, 0.25 ng, 0.0625 ng (processed in triplicate) and 0.01525 ng (processed in quadruplicate).
	SwabSolution™ (Promega Corporation, Madison, WI, USA) crude buccal swab lysate	3 µL 5X AmpSolution® added to undiluted lysate, then further diluted in nuclease-free water: 1 in 10, 1 in 100 and 1 in 1000 and processed in quadruplicate.
Stability: Degradation	2800M control DNA (Promega Corporation, Madison, WI, USA)	Diluted to 1 ng, subjected to incubation at 95 °C for 20, 40 and 60 minutes on the Eppendorf® Thermo-Mixer C (Eppendorf, Hamburg, Germany) and processed in triplicate
	SRM 2372a (NIST, USA) quantification standard	
Stability: Inhibition	2800M control DNA (Promega Corporation, Madison, WI, USA)	Diluted to 1 ng, spiked with 400 µM and 600 µM humic acid (Sigma-Aldrich, MO, USA), and processed in triplicate.
	SRM 2372a (NIST, USA) quantification standard	
	2800M control DNA (Promega Corporation, Madison, WI, USA)	Diluted to 1 ng, spiked with 1% ethanol (Sigma-Aldrich, MO, USA) and processed in triplicate.
	SRM 2372a (NIST, USA) quantification standard	
Accuracy and precision	2800M control DNA (Promega Corporation, Madison, WI, USA)	Diluted to 1 ng and processed in triplicate.
	Authentic forensic samples: Blood, bone, crude buccal swab lysates, FFPE tissue, FTA card (buccal), nail and tooth	Diluted to 1 ng and processed in triplicate.
Repeatability	2800M control DNA (Promega Corporation, Madison, WI, USA)	Diluted to 1 ng and processed in triplicate by the same user on different runs.
	Authentic forensic samples: Blood, bone, crude buccal swab lysates, FFPE tissue, FTA card (buccal), nail and tooth	Diluted to 1 ng and processed in triplicate by the same user on different runs.

Reproducibility	2800M control DNA (Promega Corporation, Madison, WI, USA)	Diluted to 1 ng and processed in triplicate by a different user.
	Authentic forensic samples: Blood, bone, crude buccal swab lysates, FFPE tissue, FTA card (buccal), nail and tooth	Diluted to 1 ng and processed in triplicate by a different user.
Species specificity	Extracted DNA from Rhesus blood, cat buccal swab, dog buccal swab, domesticated chicken swab, domesticated cow swab	Diluted to 1 ng, where necessary and processed once.

5.2.4. ForenSeq™ DNA Signature Prep kit library preparation and sequencing

Libraries were prepared using DNA primer mix B in batches of 32 samples, following the manufacturers' protocol (Verogen, San Diego, CA, USA) [59]. Each of the six plates included 1 ng of the 2800M positive control (Promega Corporation, Madison, WI, USA) and nuclease-free water as a no template control. For buccal swabs processed in SwabSolution™ (Promega Corporation, Madison, WI, USA), 2 µL of the buccal swab lysate was added to the PCR 1 master mix, and 3 µL of 5X AmpSolution® (Promega Corporation, Madison, WI, USA), instead of nuclease-free water, as determined through optimisation experiments (Chapter 3). Amplification steps were performed using a T100 thermal cycler (BioRad, Hercules, CA, USA).

Normalised libraries from each plate of 32 samples were pooled. Sequencing was prepared according to the manufacturer's protocol except that the volume of pooled library added to the MiSeq FGx™ reagent cartridge was adjusted to 10 µL. Sequencing was carried out on the MiSeq FGx™ benchtop sequencer as described in Chapter 3, section 3.2.2.

5.2.5. Accuracy assessment with conventional STR profiling

To assess concordance (and therefore accuracy) within the internal validation study, the forensic samples underwent conventional DNA profiling. This was achieved using the GlobalFiler® PCR Amplification Kit (Applied Biosystems, Foster City, USA) for autosomal STRs, the Investigator Argus X-12 QS Kit (QIAGEN, Hilden, Germany) for X-STRs, and the

PowerPlex® Y23 System (Promega Corporation, Madison, WI, USA) for Y-STRs [16,207,208]. All PCR products were separated on the ABI 3130xL Genetic Analyser (Applied Biosystems, Foster City, CA, USA) using a POP-4™ polymer (Applied Biosystems, Foster City, USA). The GlobalFiler® PCR Amplification kit made use of the GeneScan™ 600 LIZ size standard (Applied Biosystems, Foster City, CA, USA). The Investigator Argus X-12 QS Kit used the DNA size standard 550 (BTO) and the PowerPlex® Y23 System used the internal lane DNA standard ILS 500. The resulting electropherograms were viewed and analysed using GeneMapper® ID-X software version 1.5 (Applied Biosystems, Foster City, CA, USA).

5.2.6. Data analysis

Genotype and flanking region reports were exported from the ForenSeq™ UAS into Microsoft® Excel® for further analysis. Erroneous sequences and stutter in flanking region reports were identified and separated from the primary dataset. Data for each study were copied into separate sheets for individual analysis. The call rate for each sample was calculated using the number of genotypes that met the AT. Precision was determined by using the count of the most frequent genotype across replicates. As alluded to above, accuracy was determined if the allele called in MPS was correct as per the CE genotypes. A criterion of acceptance for call rate, precision and accuracy of control DNA and authentic forensic sample types was established as 10% of the call rate, precision and accuracy obtained in the developmental validation for control DNA samples. The accuracy assessment pertains to the concordance of the genotypes, while precision refers to the most frequently observed genotype across all loci, as outlined in the developmental validation. In this internal validation, instances where no results were obtained (*i.e.,* failed amplification of an allele or genotype at a specific locus) contributed to a decrease in precision. Additionally, cases of discordant results, where two

distinct genotypes were observed across replicates, provided both were successfully amplified, impacted the percentage accuracy.

5.2.7. Analytical and stochastic threshold setting

All negative amplification control and extraction blank samples were used to establish the analytical threshold, while the results from the sensitivity studies were used to determine the stochastic threshold. While some genotyping software tools such as TrueAllele®, do not require the establishment of thresholds and considers all data available for probabilistic genotyping, thresholds were established here for the purposes of comparison to the developmental validation data. Additionally, our laboratory currently utilises the ForenSeq™ UAS for genotyping of MPS data, which requires the input of a threshold.

All samples that undergo DNA extraction in our laboratory are processed alongside a DNA extraction blank consisting of nuclease-free water or unused swabs instead of a biological sample. The DNA extraction blanks generated in this internal validation study were used to establish the analytical threshold. Additionally, all libraries were prepared alongside negative amplification control samples containing nuclease-free water. In addition to being a quality control measure for assessing possible contamination, they were used in the internal validation study to establish read counts for the noise threshold determination.

For establishing the AT, both the AT and IT on the ForenSeq™ UAS software for all markers were initially set to 0, after which genotype reports for all NTC, and extraction blank samples were exported into Microsoft® Excel® documents. The AT was determined by using the average reads detected in NTC and blank samples (**Equation 5.1**). After assessment of these reads, the analytical threshold was set to 1.9% (12 reads) across all markers (Appendix 5.1).

$$Threshold = Average\ read\ count + (2\ x\ Standard\ Deviation)$$

Equation 5.1: *Equation used to determine analytical threshold*

The IT was determined by setting the newly determined AT and then exporting genotype and flanking region reports of samples used in sensitivity studies. Due to the high variability expected for direct PCR samples, equation 1 was not used for the IT calculation. It was hypothesised that the of this equation for setting the IT would result in an unnecessary loss of allelic data. The average read count across all markers where the second allele dropped out was thus used and found to be 35 ± 19 reads. The new IT was set to 5% (35 reads) for all further analyses.

5.3. Results

5.3.1. Quality Metrics

The quality metrics obtained for all sequencing runs were found to be within the recommended ranges. Cluster density for all runs was in the range of 754 – 1246 k/mm^2. The recommended range for cluster density is between 700 – 1400 k/mm^2. The recommended percentage of clusters passing the Illumina chastity filter is 80%, and this requirement was met for all runs. The percentage of molecules that fell behind (phasing) or jumped ahead (pre-phasing) of a sequencing cycle were also within recommended ranges for each run (**Table 5.2**). Additionally, the HSC met the minimum criteria for intensity and concordance.

Table 5.2: *Run quality metrics obtained for six validation experiments performed with the ForenSeq™ DNA Signature Prep kit on the MiSeq FGx™ instrument.*

Run number	Cluster Density (K/mm2)	Clusters PF (%)	Phasing	Pre-phasing
Run 1	1020	91.3	0.205	0.08
Run 2	1223	88.79	0.197	0.09
Run 3	754	93.79	0.216	0.027
Run 4	921	92.05	0.208	0.08
Run 5	1085	89.84	0.222	0.046
Run 6	1246	86.59	0.179	0.076

5.3.2. Performance parameters

5.3.2.1. Sensitivity study

2800M Control DNA

The aim of the sensitivity study was to determine the limit of detection, which is defined as the lowest input of DNA at which a full profile (or 100% call rate) can be achieved. The internally established criterion of acceptance for sensitivity studies was 0.25 ng. Thus, below 0.25 ng, it is expected to see some alleles drop out, and above it, a 100% call rate is expected for control

DNA samples, as was observed in for the 2800M (Promega Corporation, Madison, WI, USA) control DNA and crude buccal swab lysates this study (**Table 5.3**, Appendix 5.2).

Table 5.3: *Average and standard deviation of call rates (%) obtained for the 2800M control DNA sample sensitivity assessment.*

Sample	Marker	Average call rate (%) ± standard deviation			
		1 ng	**0.25 ng**	**0.0625 ng**	**0.01525 ng**
2800M	**A-STR**	100 ± 0	98.81 ± 2.06	97.62 ± 4.12	73.81 ± 10.91
	Y-STR	100 ± 0	100 ± 0	100 ± 0	59.72 ± 18.79
	X-STR	100 ± 0	100 ± 0	85.71 ± 0	71.43 ± 24.74
	iiSNV	100 ± 0	95.39 ± 6.23	76.6 ± 2.81	33.33 ± 13.22
	aiSNV	98.81 ± 2.06	95.24 ± 2.06	82.74 ± 2.73	36.31 ± 9.83
	piSNV	100 ± 0	100 ± 0	90.91 ± 4.55	62.12 ± 20.99

The call rates obtained across all markers for the 2800M (Promega Corporation, Madison, WI, USA) control sample were 100% at 1 ng input amounts, while an average call rate of 98.55% was obtained for an input amount of 0.25 ng (**Table 5.3**). Full profiles (100% call rates) were obtained for A- and Y-STRs for input amounts as low as 0.0625 ng, which is in alignment with the lowest input amounts at which 100% call rates were obtained in the developmental validation [20]. Call rates of 86% were obtained for A-STRs, X-STRs and piSNVs at an input amount of 0.01525 ng. At lower input amounts, the iiSNV markers resulted in lower call rates than other markers.

SRM 2372a quantification standard

The call rates for the SRM 2372a (NIST, MA, USA) quantification standard showed lower call rates in comparison to the 2800M (Promega Corporation, Madison, WI, USA) (**Table 5.4**, Appendix 5.3). More specifically, a 94% call rate for A-, Y- and X-STRs, and 98% for iiSNVs was observed for samples diluted to 2.5 ng, and the minor allele loss was attributed to the amount of DNA being too high (above the manufacturer's recommended 1 ng input), thereby

reducing PCR amplification efficiency for larger markers. This was confirmed by processing the same SRM 2372a (NIST, MA, USA) sample diluted to the recommended 1 ng, where a 100% call rate was observed. Therefore, the lowest input amount for the SRM 2372a standard at which a 100% call rate could be obtained was the recommended 1 ng.

Table 5.4: *Average and standard deviation of call rates (%) obtained for the SRM2372a quantification standard assessed as part of the sensitivity study.*

Sample	**Marker**	**Average call rate (%) ± standard deviation**				
		1 ng (processed once)	**2.5 ng**	**0.25 ng**	**0.025 ng**	**0.0025 ng**
SRM 2372a	**A-STR**	100	92.86 ± 0	83.33 ± 2.06	53.57 ± 6.19	3.57 ± 3.57
	Y-STR	100	88.89 ± 2.41	75 ± 4.17	55.56 ± 6.36	9.72 ± 6.36
	X-STR	100	100 ± 0	85.71 ± 14.29	71.43 ± 14.29	0 ± 0
	iiSNV	100	97.87 ± 1.84	84.75 ± 2.21	49.29 ± 6.41	7.45 ± 1.06
	aiSNV	100	98.21 ± 0	90.48 ± 2.06	55.95 ± 7.43	4.76 ± 4.49
	piSNV	100	100 ± 0	93.94 ± 2.62	66.67 ± 9.46	4.55 ± 4.55

Blood

Genomic input amounts of 1 ng replicates of blood samples resulted in an average call rate of 99.89% across all markers, while an average call rate of 94.09% was obtained for input amounts of 0.25 ng for all markers (**Table 5.5**, Appendix 5.4). The iiSNV markers performed poorly at 0.0625 ng input amounts (48.0%), in contrast to the performance of other markers. At 0.1525 ng input amounts, all markers resulted in an average call rate of 32.94%. When using blood as a sample type, an input amount of 1ng is recommended, although at 0.25 ng input amounts, a full profile is expected across all markers.

Table 5.5: *Average and standard deviation of call rates (%) obtained for the blood sample assessed as part of the sensitivity study.*

Sample	Marker	Average call rate (%) ± standard deviation			
		1 ng	**0.25 ng**	**0.0625 ng**	**0.01525 ng**
Blood	**A-STR**	100 ± 0	94.64 ± 4.72	72.02 ± 2.73	25.6 ± 5.74
	Y-STR	100 ± 0	94.44 ± 9.62	72.22 ± 4.81	41.67 ± 4.17
	X-STR	100 ± 0	90.48 ± 8.25	80.95 ± 8.25	33.33 ± 21.82
	iiSNV	99.29 ± 0.61	84.04 ± 5.63	47.52 ± 4.43	14.18 ± 3.42
	aiSNV	100 ± 0	98.21 ± 1.79	82.14 ± 4.72	42.86 ± 11.71
	piSNV	100 ± 0	100 ± 0	90.91 ± 4.55	48.48 ± 6.94

Crude buccal swab lysates

One of the primary objectives of this internal validation study was to validate the direct PCR workflow for buccal swabs collected from deceased individuals. When undiluted lysates were processed, 100% call rates were obtained across all markers (**Table 5.6**).

Table 5.6: *Average and standard deviation of call rates (%) obtained for the post-mortem crude buccal swab lysate assessed as part of the sensitivity study.*

Sample	Marker	Average call rate (%) ± standard deviation			
		Undiluted	**1 in 10**	**1 in 100**	**1 in 1000**
Crude buccal swab lysate	**A-STR**	100 ± 0	98.81 ± 2.06	78.57 ± 6.19	16.67 ± 5.46
	Y-STR	100 ± 0	98.72 ± 2.22	76.92 ± 0	19.23 ± 10.18
	X-STR	100 ± 0	95.24 ± 8.25	80.95 ± 8.25	28.57 ± 0
	iiSNV	100 ± 0	99.65 ± 0.61	78.01 ± 2.21	10.64 ± 2.81
	aiSNV	100 ± 0	97.62 ± 1.03	80.36 ± 4.72	12.5 ± 3.57
	piSNV	100 ± 0	100 ± 0	90.91 ± 0	19.7 ± 18.37

Call rates above 99% were obtained across all markers for 1 in 10 dilutions, except for X-STRs, where a call rate of 95.24% was obtained across replicates. Call rates of above 75% were obtained for 1 in 100 dilutions, however, call rates were less consistent across replicates (**Figure 5.1**). Lower success rates were observed for lysates diluted 1 in 1000, with an average call rate of 21.49% for all STRs, 14.28% for all SNVs. The lowest dilution factor at which a 100% call rate could be obtained for crude buccal swab lysates was thus a 1 in 10 dilution. The

optimal dilution for crude buccal lysates therefore remains 2 µL of undiluted lysate with 3 µL of 5X AmpSolution®.

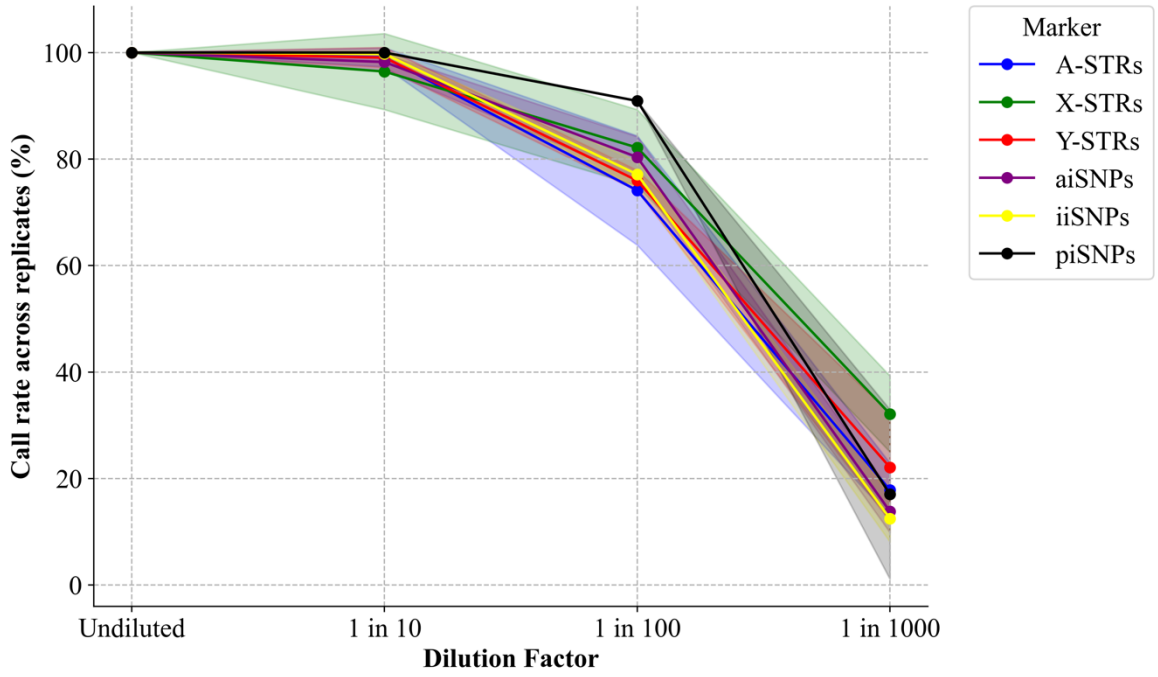

***Figure 5.1**: Average call rates (%) of autosomal STRs (blue), X-STRs (green), Y-STRs (red), iiSNVs (yellow), aiSNVs (purple) and piSNVs (black) using a crude buccal swab lysate for the following dilutions; undiluted, 1 in 10, 1 in 100 and 1 in 1000. The shaded regions around each line depict the standard deviation observed across replicate samples.*

To gain further insight into the performance of crude buccal swab lysates, read counts were assessed for all crude buccal swab lysate dilutions. Call rates above 80% were obtained when read counts were below the 85 000 read count guideline [20] (***Figure 5.2***). The lowest read count at which a full profile could be obtained for an undiluted lysate was 20 000. All lysates that were diluted with water did not meet the 85 000 read count guideline, however, 1 in 10 dilutions still resulted in high call rates (above 80%) with read counts as low as 18 000.

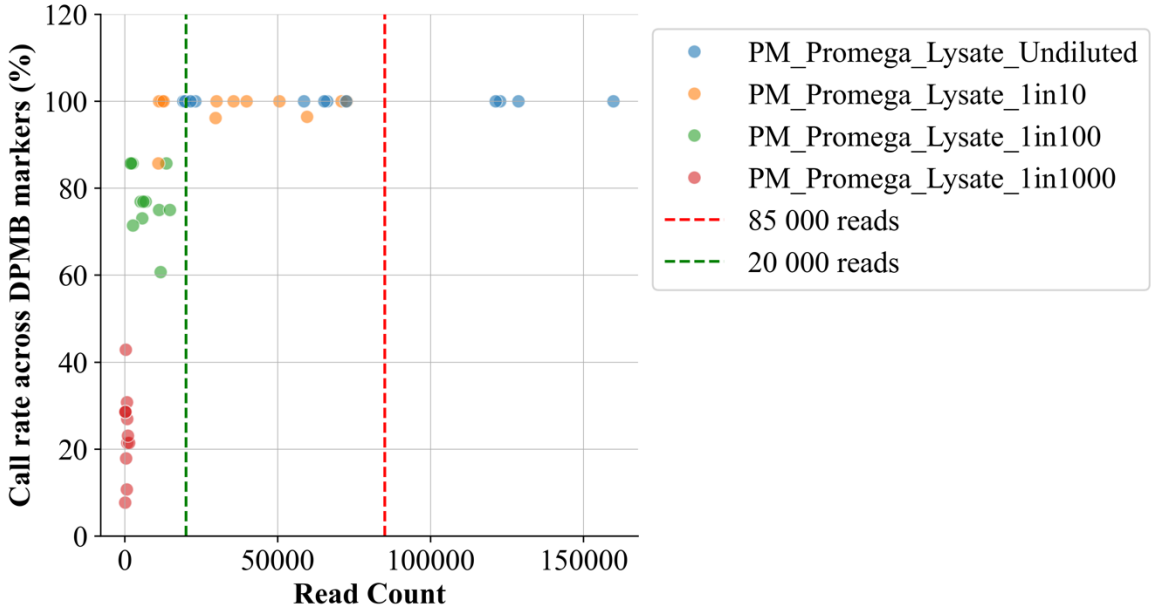

Figure 5.2: *Scatter plot illustrating read counts for all undiluted and diluted lysates processed as part of sensitivity studies and their corresponding call rates (%). PM = post-mortem.*

5.3.2.2. Stability studies

Degradation study

Heat-degraded control DNA samples were degraded at 95°C for 20, 40 and 60 minutes, and were found to range between mildly degraded, and severely degraded, as determined by the DI obtained through qPCR experiments. There was an observed and expected increase in the DI for samples heated longer time periods (Appendix 5.5). The control DNA that was not subjected to heat-degradation experiments resulted in a DI value below 1, indicating that this DNA sample was intact, as categorised according to Vernarecci *et al*., 2015 [209] (Appendix 5.6). The mean call rate across all markers was 100% for the untreated control DNA sample. For the 20- and 40-minute heating period, mean call rates of 81.45% and 76.5% was observed, respectively, for A-, Y- and X- STR markers (**Figure 5.3**). Interestingly, the control sample, degraded for 60 minutes, with the highest mean DI value of 30, resulted in the highest call rates for both STRs (82.64%) and iiSNVs (93.35%). The A-STR markers which frequently failed to

amplify across all time periods were the larger markers such as Penta D, Penta E, D22S1045 and D5S818. The average call rate for iiSNVs were generally higher for all degradation time periods. Furthermore, all sequence-based genotypes of degraded control DNA samples were found to be 100% concordant with length-based genotypes.

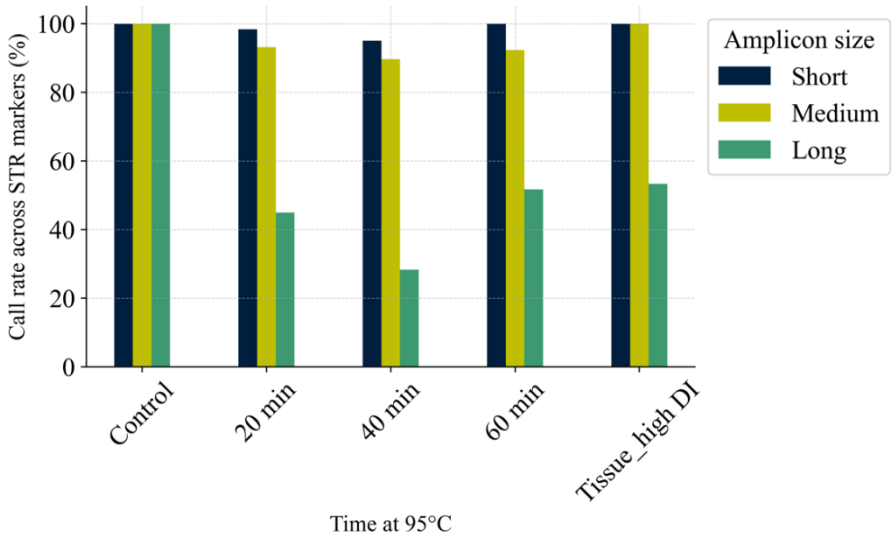

Figure 5.3: *Bar chart illustrating the average call rate (%) for the 2800M control DNA sample subjected to heat-degradation for different time intervals at 95°C, separated according to amplicon size (short amplicons shown in blue, medium amplicons shown in yellow and long amplicons shown in green), as categorised according to the developmental validation[20].*

Inhibition study

The control samples spiked with 1% ethanol, were not flagged as inhibited in qPCR experiments, but despite this, they were automatically marked as "low quantity." In every instance, these ethanol-spiked controls achieved a 100% call rate and precision for all markers, and, except for aiSNVs. A 100% accuracy for CE-shared STR markers was achieved, maintaining this rate even when flagged for low quantity in qPCR (**Figure 5.4** and Appendix 5.6). In contrast to samples spiked with 1% ethanol, control DNA samples 2800M (Promega Corporation, Madison, WI, USA) and SRM 2372a (NIST, MA, USA) spiked with humic acid at 400 µM and 600 µM resulted in completely failed ForenSeq™ libraries. Quantification

results with the Quantifiler® Trio kit (Applied Biosystems, Foster City, USA) showed no indication of inhibition.

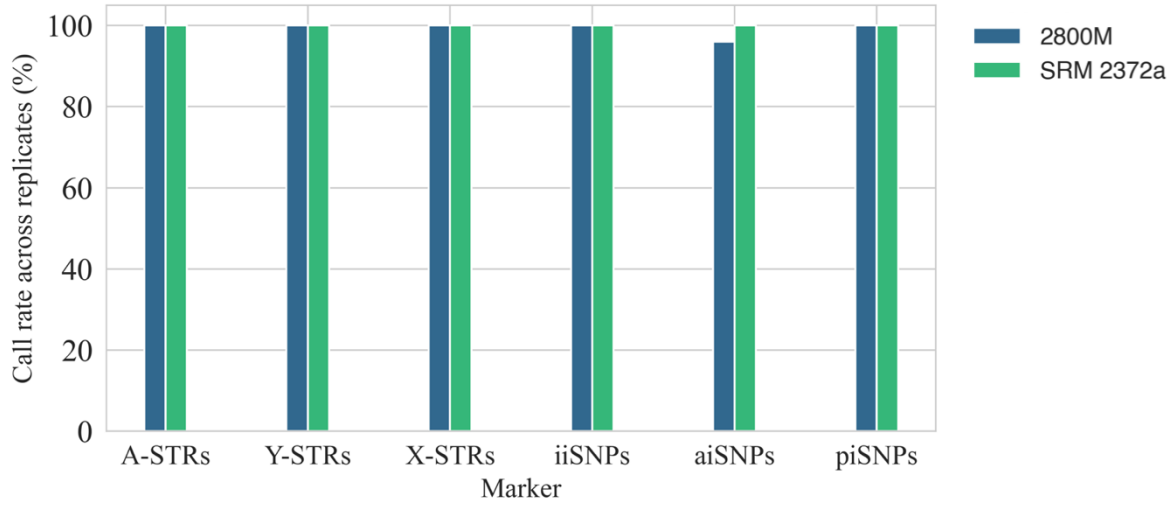

Figure 5.4: *Bar chart showing call Rates for A- Y-, X-STRs, iiSNVs, aiSNVs and piSNVs for 2800M control DNA samples (shown in blue) and SRM 2372a (shown in green) spiked with 1% ethanol.*

5.3.2.3. Accuracy, precision and call rate

The data obtained through repeatability and reproducibility studies for the 2800M control DNA were used to assess accuracy, call rate and precision, which has also been done in the developmental validation. A 100% accuracy of genotype calls was obtained across the CE-shared STR markers. Accuracy of SNV markers were assessed through comparison with the SNV profile published in the manufacturer's protocol and was found to be 100% accurate across all markers. Call rates across replicates was determined to be 100% for all markers except for aiSNVs, where the precision was 97.53% due to one of the 2800M control DNA replicates consisting of incomplete genotypes at three aiSNV markers (**Table 5.7**). However, all accuracy, precision and call rate statistics were in alignment with those obtained in the developmental validation.

Table 5.7: *Accuracy, precision and call rate obtained for the 2800M control DNA replicates across 231 markers, including A-STRs, Y-STRs, X-STRs, iiSNVs, aiSNVs and piSNVs. IV^a = performance metric obtained for internal validation study. DV^b = performance metric obtained for developmental validation study.*

Parameter	A-STRs		Y-STRs		X-STRs		iiSNVs		aiSNVs		piSNVs		Within 10% of DV^b result
	IV^a	DV^b	IV^a	DV^b	IV^a	DV^b	IV^a	DV^b	IV^a	DV^b	IV^a	DV^b	
Accuracy (%)	100	100	100	100	100	100	100	99.73	100	100	100	100	✓
Precision (%)	100	100	100	99.74	100	100	100	99.45	99.1	100	100	100	✓
Call Rate (%)	100	100	100	99.74	100	100	100	100	97.5	100	100	100	✓

5.3.2.4. Sample type studies

The sample types tested were those commonly encountered in forensic mortuaries, and included bone, teeth, nail, blood, buccal swabs, FFPE tissue and FTA cards (buccal). Call rates of 100% were obtained for nail, bone, tooth, blood, buccal swab lysates (from both living and deceased individuals) across STR markers (**Figure 5.5**). Across all markers, buccal swab lysates spiked with 5X AmpSolution® resulted in the overall highest call rate. Sample types with call rates below 100% were frequently due to the markers prone to allele dropout such as Penta D, Penta E and D22S1045. A small allelic loss was observed in iiSNV markers in the FTA card and Y-STR markers for the FFPE sample. It was also possible to generate full phenotype estimations for all authentic forensic sample types tested.

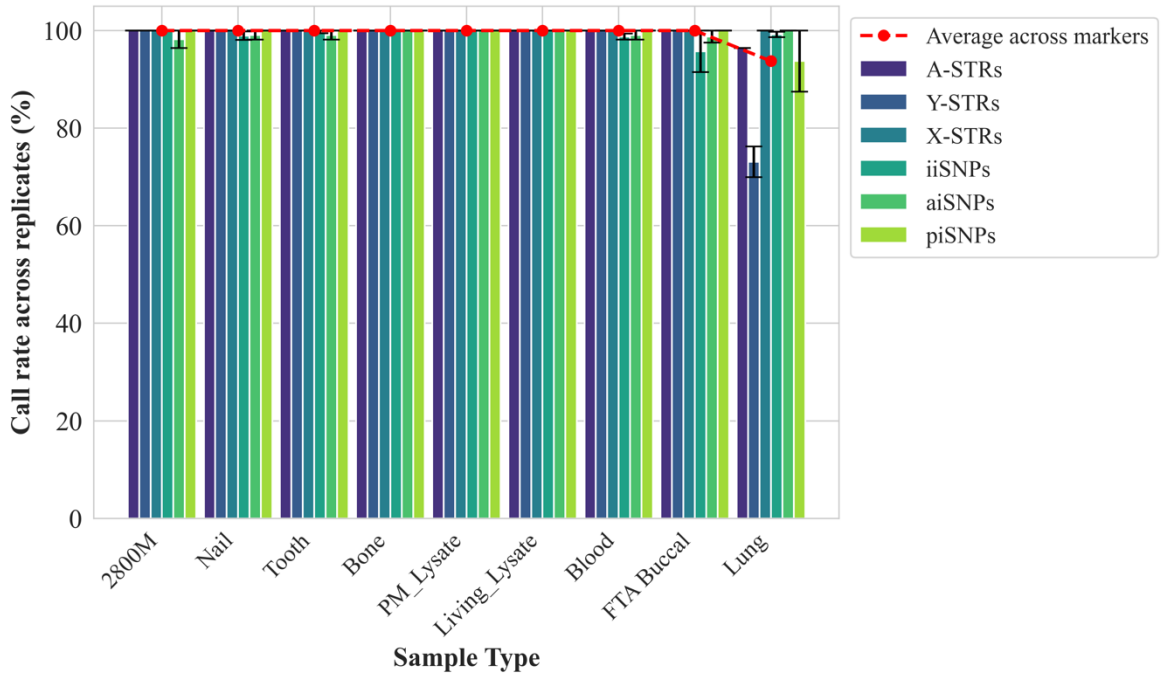

Figure 5.5: *Call rate (in percentage) for the 2800M control DNA sample and authentic forensic sample types across all markers. PM = post-mortem, Lung = FFPE tissue.*

The precision and accuracy statistics across all sample types and across all markers exceeded 99% (**Table 5.8**), indicating that the workflow is deemed fit for purpose for the forensic sample types tested in this validation study.

Table 5.8*: Accuracy (%) and intra-assay precision (%) across forensic sample replicates. The inter-assay precision for the FFPE tissue sample was the lowest and is emboldened.*

	Intra-assay precision (%)						**Accuracy (%)**
Sample Type	**A-STRs**	**Y-STRs**	**X-STRs**	**iiSNVs**	**aiSNVs**	**piSNVs**	**STRs only**
Nail	100	100	100	97.16	100	100	100
Tooth	100	100	100	99.19	100	100	100
Bone	100	100	100	99.47	100	100	100
Crude buccal swab lysate (post-mortem)	100	100	100	99.73	100	100	100
Crude buccal swab lysate (living)	100	100	100	100	100	100	100
FTA buccal	100	100	100	95.74	100	100	100
Blood	100	100	100	98.76	100	100	100
FFPE tissue	96.43	**73.08**	100	99.2	99.07	100	100
Average (%)	99.6	97.01	100	98.81	99.9	100	100

Determining repeatability involved calculating inter-assay precision between two identical runs performed by the same analyst, as described in **Table 5.1**. The inter-assay precision for the same analyst was found to be above 99% across all markers for bone, tooth, and crude buccal swab lysate samples, indicating that the workflow was highly repeatable for these sample types (**Table 5.9**). The inter-assay precision across the three runs was also above 99% across all markers for bone, tooth, and crude buccal swab lysates, indicating that the workflow was highly reproducible. However, iiSNV marker precision statistics were below 98% for the nail sample.

Table 5.9: *Inter-assay precision calculated across runs for repeatability and reproducibility studies. The numbers, "1-2" shows the inter-assay precision between two runs performed by the same analyst. The numbers, "1-2-3" shows the inter-assay precision between three runs, where run by the same analyst, and the third by a different analyst. The developmental validation (DV) precision statistic is shown for a control sample, and whether each marker met this result within 10% is shown with a tick mark (✓).*

Inter-assay precision (%)										
Marker	**Nail**		**Tooth**		**Bone**		**Lysate**		**DV precision (%)**	**Within 10% of DV result**
	1-2	**1-2-3**	**1-2**	**1-2-3**	**1-2**	**1-2-3**	**1-2**	**1-2-3**		
A-STRs	99.11	98.31	100	99.7	100	99.7	100	100	100	✓
Y-STRs	99.5	98.91	100	100	100	100	100	100	99.74	✓
X-STRs	100	100	100	100	100	100	100	100	100	✓
iiSNVs	97.82	97.79	99.34	98.18	99.6	98.4	99.87	99.9	99.45	✓
aiSNVs	100	100	100	100	100	100	100	100	100	✓
piSNVs	100	100	100	100	100	100	100	100	100	✓

5.3.2.5. Species specificity study

The dog, chicken and cow samples resulted in no sequencing reads, while a total read count of 83 was obtained for the cat sample across all STR and iiSNV markers. The rhesus macaque sample resulted in the highest total of 14 484 reads. For the rhesus macaque, five A-STR (5/28)

genotypes were called. The Amelogenin genotype obtained for the rhesus macaque was “X/X”. No Y or X STRs were called, and eight iiSNVs were called (8/94). The developmental validation reported a total read count range of ~159000 to ~178000 for a rhesus monkey [20]. Cross reactivity was expected for non-human primates, and this was evident in the higher read counts obtained for the rhesus macaque blood sample. Given the low call rate of this species’ sample across all markers, usable information from the primate for matching purposes was deemed unlikely. With this in mind, the read counts for the rhesus primate obtained in this study were below those obtained by the developmental validation [20].

5.4. Application to a forensic cold case

In addition to the assessment of the performance parameters, this chapter comprised a component to demonstrate the suitability of the workflow in the authentic forensic context. To this end, the internally validated workflow was applied to a sample from an unidentified deceased individual whose investigation had become cold with state authorities. This component was performed under a different umbrella project with its own ethical approval (HREC: 692/2021) (referred to as the Western Cape Cold Case Consortium (W4C) project) and ethical approval was obtained to include this component in this doctoral thesis (HREC: 400/2021).

A decomposed body was found in an aquatic environment in the southern region of Cape Town in South Africa several years ago. The details surrounding the case are not reported in this thesis due to privacy and the (now) ongoing investigation. Thus, no genotype will be reported; however, information related to DNA sample preparation, library preparation, sequencing, overall performance and DNA profile frequency are given in this study to demonstrate the value of the workflow for its intended purpose.

5.4.1. Sample processing

A toenail clipping was collected from the body by a state forensic pathologist and sent to our forensic laboratory, following chain of custody protocols. Although hard tissue samples were also requested, these had already undergone maceration in formalin and usable DNA from these samples was impossible to obtain. The toenail clipping underwent washing and incubation as described in section 5.2.2.2. However, purification was done using magnetic beads, with the Mag-Bind® Blood DNA HV Kit (Omega Bio-tek), adjusted for using an input volume of 500 µL, and eluted into 30 µL [168]. The magnetic-bead method was done with future automation intentions in mind.

DNA was quantified using the Quantifiler Trio® kit (Applied Biosystems, Foster City, USA) as described earlier in section 5.2.3. Furthermore, FTA cards (buccal) from two alleged siblings were submitted to our laboratory for conventional DNA profiling with the GlobalFiler® Amplification kit (Applied Biosystems, Foster City, USA), and processed, as described in this chapter (section 5.2.5).

5.4.2. Library preparation, quality control and sequencing

The DNA extracted from the toenail sample underwent library preparation and sequencing according to section 5.2.4 of this chapter. However, prior to library normalisation, the purified library of this sample, as well as that of the 2800M (Promega Corporation, Madison, WI, USA) control DNA, were assessed using the TapeStation 2200 with the High Sensitivity D1000 Screen Tape Assay (Agilent Technologies) to confirm library quality [171].

5.4.3. Data analysis

Data analysis was performed using the ForenSeq™ UAS using thresholds established during internal validation (AT = 12 reads; IT = 35 reads). Data for the 2800M and extracted DNA sample were exported into Microsoft® Excel for further analysis. A bio-geographical ancestry and phenotype report was generated on the ForenSeq™ UAS for the unidentified individual. Hair and eye colour were estimated using all the available aiSNV and piSNV data by entering the SNV genotypes into the HIrisPlex-S System [210-212]. The hair and eye colour results were also checked using the Snipper App Suite version 2.5 (http://mathgene.usc.es/snipper/), while bio-geographical ancestry was estimated using the ForenSeq™ UAS. The DNA profile frequency of the deceased individual was calculated using the sequence-based allele frequency data generated as part of this study (Chapter 4). All data generated for the unidentified deceased individual and alleged siblings were submitted TrueAllele® (Cybergenetics, Pittsburgh, USA) for assessment of kinship.

5.4.4. Profile success

The MPS profile resulted in successfully typed Amelogenin marker, 22/27 A-STRs, 17/24 Y-STRs, 5/7 X-STRs, 73/94 iiSNVs, 19/22 piSNVs, and 49/56 aiSNVs. The profile resulted in successfully typed genotypes and/or haplotypes at 186/231 markers, resulting in an overall profile success of 80.52% (**Figure 5.6**).

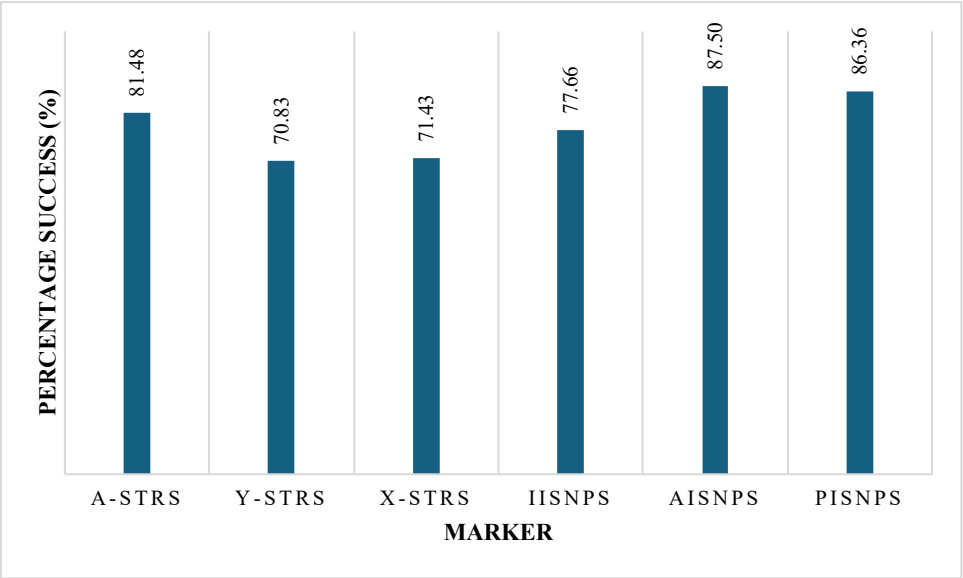

***Figure 5.6**: Profile success across all markers. Success rate is shown as a percentage of successfully typed genotypes or haplotypes for each marker.*

The DNA profile frequency was 3.38E-34 and was calculated using the sequence-based allele frequencies. Rare sequence alleles were observed at D2S1338, D9S1122 and D12S391, where the minimum allele frequency approach (5/2N) was used. Biogeographical ancestry, hair and eye colour were successfully estimated and used by a forensic artist in the generation of a digital facial image. Facial depictions are typically done in greyscale, however, these results provided insight into colour, for the first time. All results obtained from this cold case were handed over to the main study, which was then handed over to the SAPS for dissemination into the public.

5.5. Discussion

This internal validation was designed, carried out and reported for the ForenSeq™ DNA Signature Prep kit with the MiSeq FGx™ sequencing platform for a forensic mortuary-linked DNA laboratory. All parameters assessed in the internal validation study were commensurate with the equivalent performance parameters reported in the developmental validation study (where comparison was possible), with the exception that the workflow was not tolerant to high concentrations of humic acid, and it was not able to distinguish between non-human primate and human DNA. For the first time, performance parameters for post-mortem crude buccal swab lysates were established, thus fulfilling a primary objective of this study.

This internal validation established performance parameters for post-mortem crude buccal swab lysates processed in SwabSolution™ using a direct PCR approach (Promega Corporation, Madison, WI, USA) for the first time, as previously studies have focused on using extracted DNA, FTA cards, buccal swabs processed in QuickExtract™ DNA Extraction Solution (Epicentre®, Madison, WI, USA) and samples collected from living individuals [20]. This is important because it expands the samples that can be tested. It also validated the optimised methods for crude buccal lysates developed in chapter 3.

MPS profiles generated using a direct PCR approach showed a high performance with regards to accuracy, precision and call rate (**Table 5.8**). More so, complete biogeographical ancestry and phenotypes were generated for all casework samples, demonstrating a clear advantage over traditional CE-based STR profiling methods. FTA cards were validated using DPMB for the first time, as the developmental validation tested FTA cards with DPMA [20]. Although lower call rates were observed for FTA cards, this could be attributed to factors such as inadequate buccal cells present in the inner cheek of the deceased, or inadequate sample transfer from the swab onto the FTA card. However, call rates of above 95% were still obtained across all STR

markers, which is sufficient for comparison to reference samples, and similar allelic loss was observed in the developmental validation for FTA cards [20].

The sensitivity results shows that the optimal input range across sample types tested is between 0.0625 ng and 1 ng (**Table 5.3**). This finding is concordant with both CE and MPS studies, where they have reported accurate and precise genotype calling above between 0.05 ng and 0.2 ng [84], [202]. Based on sensitivity results obtained, the following recommendation is made; input amounts between 0.25 ng and 1 ng are sufficient for generating accurate and precise genotype calls.

It should be noted that the authentic forensic sample types commonly encountered in our mortuary, such as bone, nail and teeth samples generally have concentration ranges above 0.2 ng/µL. Furthermore, although lower call rates are observed below 0.0625 ng, the amount of genetic information that can be obtained from a partial MPS profile is substantially greater than a partial or even full CE-based STR profile. Extracted DNA samples such as bone, nail, teeth and soft tissue, that often pose challenges in traditional DNA profiling methods, performed exceptionally well with the tested sequencing workflow, exceeding call rates previously obtained for similar post-mortem sample types by other studies (Appendix 1.1).

The workflow was found to be both stable and reliable in highly degraded samples, with DI values ranging between mildly to severely degraded, as determined through qPCR experiments (Appendix 5.5). Longer amplicons in degraded samples had expectantly lower call rates than medium and shorter amplicons, due to the preferential amplification of shorter amplicons during the first PCR amplification step [14,53]. This is a known phenomenon that occurs during PCR amplification in degraded DNA samples. Similar results have been obtained in the

developmental validation and studies testing the stability of the workflow using artificially degraded samples [20,61]. However, the high call rates (mean of above 84% across all markers) observed in this study (Appendix 5.5) conflict with results obtained by Fattorini *et al.*, 2017, who obtained call rates of 5% for STRs and 11% for SNVs [213]. The inconsistency in call rates may be due to different levels of degradation in samples, as the method used by Fattorini *et al*., 2017 to artificially induce degradation in control DNA samples was hydrolysis [213].

The accurate and precise genotype calls generated from this workflow in severely degraded control samples, as well as in a severely degraded FFPE tissue sample was an important finding of this internal validation study. This is because it aligns directly with the rationale for this validation, which was to test the performance of the workflow on challenging mortuary sample types. Unidentified remains are often in a state of advanced decomposition due to exposure to conditions that degrade DNA within the cells. However, with the number of markers (231) amplified in the DPMB of the ForenSeq™ DNA Signature Prep kit, even with preferential amplification of shorter amplicons, a significant amount of information was still obtained from a severely degraded sample.

Humic acid and ethanol have been known to inhibit PCR amplification, or at least reduce the efficiency of PCR amplification enzymes such as DNA polymerases [162,214,215]. In the developmental validation study, the ForenSeq™ DNA Signature Prep kit demonstrated tolerance to certain concentrations of humic acid, as high as 66.67 µM, with a high genotyping accuracy and call rate, but complete inhibition at 133 µM [20]. In this study, higher concentrations of humic acid were used, and the workflow showed complete inhibition at 400 µM and 600 µM. Further experimentation would require testing of humic acid at lower concentrations prior to spiking of control DNA samples to test the true tolerance of the workflow to this inhibitor in our laboratory.

Control DNA samples spiked with 1% ethanol showed no inhibition and resulted in 100% accuracy, precision and call rates (**Figure 5.4**). Ethanol plays a role in the DNA extraction and purification process, where it is used to separate the DNA from other components that are present, such as salts that are added to enable precipitation. When ethanol is not adequately removed, it can be co-extracted with DNA, causing issues with PCR amplification. Ethanol can inhibit DNA polymerase by disrupting the enzyme's structure and function, it can thus also impact primer binding, thereby leading reduced efficiency or complete failure of the PCR process [215,216]. Ethanol, commonly employed as a sterilising agent in mortuaries and as a washing reagent during DNA extraction, can therefore inhibit DNA polymerase [217]. However, the ForenSeq™ DNA Signature Prep kit displayed complete tolerance to ethanol's inhibitory effects, with all control samples providing full STR profiles that were entirely concordant.

The call rates obtained for samples spiked with 1% ethanol was unexpected, as the Quantifiler® Trio kit did not identify the samples as inhibited by ethanol but indicated them as "low quantity," suggesting concentrations were below the limit of detection of the assay (Appendix 5.6). The sensitivity study showed that such low concentrations typically yield partial profiles (**Table 5.4**). The low quantities obtained in the absence of inhibition indicators (*i.e.*, IPC $_{CT}$ values > 31) may thus incorrectly inform decision-making processes for downstream analyses.

Furthermore, the IPC C_T values obtained for both humic and ethanol-spiked samples were below 31, suggesting that the IPC target seemed to be unaffected by the inhibitor. However, the high C_T values (>31) of the autosomal targets *were* affected by the presence of the inhibitors. This suggests that the IPC target of the Quantifiler® Trio kit is ineffective at

detecting inhibitors. The Quantifiler® Trio kit was therefore deemed unsuitable for the reliable detection of PCR inhibitors of both humic acid and ethanol, a finding that also held true for crude buccal swab lysates assessed using qPCR (Chapter 3).

The internal validation also assessed species specificity and found that the workflow resulted in high overall specificity, except for its capability to distinguish between non-human primate and human DNA. This finding is consistent with the developmental validation and is expected due to the genetic similarities between primates [20]. Although, encountering this species is uncommon in our setting. Despite this, results would need to be interpreted with caution if a case involved an attack or presence of a non-human primate. Furthermore, no other study, apart from the developmental validation, has assessed species specificity with the ForenSeq DNA Signature Prep kit, and this study therefore contributes valuable insights to this growing, yet limited body of knowledge related to species specificity testing with forensic MPS kits.

5.6. Conclusion

This chapter demonstrated that the ForenSeq™ DNA Signature Prep kit workflow is fit for purpose in a mortuary-linked laboratory. First-time insights into the performance of post-mortem crude buccal swab lysates processed using a direct PCR approach were gained. Furthermore, acceptable performance parameters were established for both control and authentic forensic sample types in terms of accuracy, precision and call rate. Sensitivity studies provided insight into the recommended minimum DNA input range, which was commensurate with the developmental validation findings. The workflow demonstrated stability under conditions of degradation and inhibition, except under the influence of humic acid. Furthermore, the workflow was repeatable and reproducible with high precision rates for both control and authentic forensic sample types.

Finally, this chapter demonstrated the successful application of the workflow to a cold case dealing with a severely decomposed body. The generation of a forensic investigative lead such as hair colour, eye colour and bio-geographical ancestry is a first for Africa and has contributed to the generation of a digital facial image that was handed over to the South African Police Services. Although performed under a research banner, this presents a new opportunity to revisit the legal framework surrounding the generation of externally visible characteristics for humanitarian purposes, which will be discussed in the final chapter of this thesis.

Chapter 6: Discussion

A sustainable approach to facilitating implementation: optimise, generate, validate and demonstrate

The overall aim of this thesis was to facilitate the implementation of the ForenSeq™ DNA Signature Prep kit workflow for post-mortem human identification. To do so, the workflow required internal validation. However, an aspect of internal validation is the generation of population allele frequency data, and the systematic review revealed that these MPS data do not exist for the South African population. A population study was then attempted using crude buccal swab lysates, but it was discovered that the workflow required significant optimisation prior to use on a large scale. Thus, to facilitate implementation, the study required the *optimisation* of the workflow for crude buccal swab lysates in preparation for use in the population study, the *generation* of MPS population data using the optimised protocol, internal *validation* of the workflow and *demonstration* of its applicability to a forensic cold case. This chapter addresses whether these objectives have been achieved and their findings are discussed holistically. The findings obtained from this study will be drawn upon for translation to our own forensic laboratory, and other under-resourced laboratories in Africa.

6.1. Facilitating sustainable implementation

Facilitating implementation of MPS technology in a forensic setting requires the generation of allele frequency data, just as for implementing CE technology. However, undertaking a large-scale population study using MPS is a massive undertaking for laboratories with limited resources. It is therefore imperative that sustainable practices are formed to leverage available data *and* laboratory resources to maximise MPS reagents. As improvements are made to current profiling methods, aspects related to sample preparation should be transferrable between

methods (*i.e.*, between CE and MPS). This is an especially important consideration for laboratories with limited resources.

One way in which available resources could be leveraged is through using readily available DNA samples in the laboratory. Samples used for establishing CE-based allele frequency databases could potentially also be used to conduct MPS-based population studies. In this study, this approach was used for a set of crude buccal swab lysates, but was met with high failure rates, as crude buccal swab lysates had not been thoroughly tested in the developmental validation of the ForenSeq™ DNA Signature Prep kit. This section of the discussion therefore elaborates on the findings pertaining to the first two objectives of this study, which were carried out in Chapter 2 (systematic review) and Chapter 3 (optimisation study)

In our laboratory, crude buccal swab lysate samples were initially used for the establishment of CE allele frequency databases. These samples performed well with many commercial STR typing kits and they provided a streamlined approach to generating a large number of DNA profiles [117,163,164]. When this MPS-based population study was first attempted, 234 crude buccal swab lysate samples were processed, and 44% resulted in failed MPS profiles (Chapter 3). This was attributed to the crude nature of the lysates combined with the chemistry of the ForenSeq™ DNA Signature Prep kit. The lack of published research of these samples with the chemistry was also demonstrated in Chapter 2. After initial sequencing results were obtained, a research objective was formulated to investigate reasons for failure, and to determine ways to overcome what was later concluded to be PCR inhibition in the SwabSolution™ lysates and high pH in STR GO! lysates (Chapter 3).

Whilst the manufacturer's protocol does not make comment with regards to assessing crude buccal swab lysates or libraries in terms of quality within the workflow, we recommended that, as least in the initial testing phases of a new kit, workflow or technology, that quality control on a subset of samples prior to and after library preparation is carried out [21]. This was adopted after the poor sequencing results that were initially obtained, providing insight into sample selection for successful sequencing and reassurance that expensive sequencing consumables would be well-spent. This intervention would allow other laboratories, who are also wishing to adopt this workflow for the first time, to gain insight into samples and libraries prior to sequencing. This stresses the importance of thorough testing and optimisation prior to the use of new technology for large-scale population and validation studies.

In working towards leveraging research efforts, the point at which it is globally agreed upon that MPS is generally concordant with CE, without assessing concordance needs to be established. The systematic literature review in Chapter 2 revealed high concordance levels (exceeding 99%) across studies using the ForenSeq^{TM} DNA Signature Prep kit, consistent with the concordance findings of this thesis (Chapter 4: population study and Chapter 5: internal validation study). Conventional DNA profiling using CE, over many years became the gold standard. After its accuracy was proven, there was no need to confirm every single genotype with a different method. Based on these combined findings, one might consider that a consensus is being reached regarding concordance rates as well as commonly observed discordances. However, drawing on the results from Chapter 4 (population study), this may not hold true for more genetically diverse populations, where an increase in flanking region variants may elevate discordance rates.

On the one hand, African populations exhibit high levels of genetic diversity, which can be leveraged for improving discriminatory power, but also comes with higher levels of discordances, as observed in Chapter 4. On the other hand, laboratories in Africa are often under-resourced and forensic genomic studies are seldom funded. Moreso, concordance assessment requires CE-based DNA profiles, and if a laboratory do not have these data, samples will need to be collected, profiles generated using CE, and then a concordance assessment can be made. Concordance assessment is thus a large undertaking for laboratories already limited in funding, resources and equipment [67]. These two compounding aspects related to African countries underscore the need for more extensive and thorough concordance assessments in more genetically diverse populations. More importantly, they highlight the urgent need for strategies to overcome resource constraints associated with more exhaustive concordance testing in these already under-resourced laboratories.

Strategies aimed at making concordance evaluation more sustainable include the use of a subset of samples or a subset of overlapping markers to check concordance between MPS and CE kits, and then use it to inform the overall level of concordance and was also done in this study [116]. In contrast, the population data generated in Chapter 5 was submitted to STRidER, in which instances of discordances caused by incorrect allele naming by the ForenSeq™ UAS were detected in non-overlapping STR markers. It is therefore suggested that when assessing concordance using a subset of markers, that sequence data also be submitted to STRidER (or equivalent) for quality checking [32]. In light of this recommendation, the population study (Chapter 4) results also show that the ForenSeq™ UAS could be improved in its ability to include, and perhaps automate the same process undertaken by STRidER for quality control, thereby considering the flanking region variants when naming sequences, and accounting for possible discordances that would result from them.

With respect to concordance, this study also demonstrated the trade-off between an increase in variation, and the potential loss of harmony between CE and MPS methods. The streaming influx of genomic variation has had a slightly polarising effect on two important aspects in the analysis of forensic DNA profiles. A primary advantage of MPS has been the notable effect on improving discriminatory power in the statistical interpretation of DNA profile data. Contrastingly, it has simultaneously disrupted harmonisation between alleles reported by conventional CE-based DNA profiling and MPS. Moreso, inclusion of such a plethora of variation can be a double-edged sword, in that there may be an increase in the number of discordances observed between methods, as well as the allele reporting formats, as observed in Chapter 2 and Chapter 4. Whether these discordances are resolvable is linked to whether sufficient information on variation has been generated and/or reported. This provides a strong motivation for laboratories to adopt comprehensive allele reporting practices during initial implementation stages, and to collaborate with laboratories that have already conducted sequence-based population studies.

One of the biggest challenges in implementing MPS-methods into forensic laboratories is the lack of standardised STR sequence nomenclature requirements, although guidelines have been published [69,147,150]. To foster transnational collaboration between laboratories performing sequence-based population studies, a standardised STR nomenclature system would enable a more seamless use and implementation of MPS data across laboratories. To this end, a critical requirement of STR sequence nomenclature is that it must be compatible with conventional STR population databases, thus the nomenclature must also report the length-based allele [150]. Secondly, the STR sequence nomenclature should be able to capture the sequence variation within the repeat motif, especially between two alleles of the same size with different repeat sequences [147,150].

What facilitated efforts towards a standardised nomenclature approach in this study was the collaboration formed with Kings College London. Leveraging existing resources and specialist expertise have contributed towards using a standardised nomenclature between the two institutions. In this study, using a sequence-naming database already containing sequences from five British populations enabled both quick identification of both known and unknown sequences through lookup functionality. Forensic laboratories in Africa wishing to adopt this workflow are therefore encouraged to collaborate with more established laboratories that have already implemented MPS in their forensic workflows.

The sluggish embrace of new technology in African countries is arguably due to the absence of funding and resources allocated to forensic laboratories, and the lack of government participation in improving humanitarian issues. Transnational collaborative research efforts will have unyielding benefits for new MPS users with respect to increasing the magnitude of population data published by underrepresented and under-resourced laboratories by reducing the cost associated with large scale MPS population studies. These collaborative efforts hold potential to maximise variant characterisation, especially in diversity rich regions, and reinforce data- and expertise-sharing between laboratories, and could further contribute towards global MPS nomenclature standardisation.

6.2. A sustainable approach to validation

Performing an internal validation is thought to be much simpler than performing a developmental validation, as the developmental validation can be used as a blueprint to conduct the internal validation in ones' own laboratory. However, having unlimited resources to conduct exactly the same studies as manufacturer's is a luxury not afforded to laboratories in LMICs. The internal validation experiments are thus almost always adapted to match the

resources available to the laboratory. For example, the developmental validation of the ForenSeq™ DNA Signature Prep kit with the MiSeq FGx™ workflow tested reproducibility by testing the control DNA 2800M (Promega Corporation, Madison, WI, USA) with 95 replicates per plate and tested across five different laboratories on a MiSeq FGx™ sequencer. The precision statistics obtained with such a high volume of replicates will undoubtedly be more reliable than testing in triplicate, as was performed in this study. Although, as stated by John Butler in an interview aimed at debunking validation myths, "…validation should not require large numbers of samples to confirm that an instrument or method is working properly in your laboratory." [219]

The insights gained from the internal validation study (Chapter 5) have allowed for recommendations to be made regarding the number of replicates used for precision testing. This prompted the question; "how many replicates are enough?". Firstly, there are of course acceptance criteria that need to be met regarding the level of variation obtained for each test. The establishment of these criteria are based on ideal conditions, which have been formulated based on testing multiple (much more than three) replicates. The ENFSI DNA Working Group recommends a minimum of 5 replicates for validating an electrophoresis instrument [29]. Although there are general guidelines for validation of MPS instruments or workflows, no guidelines exist regarding the number of replicates that need to be tested for MPS kits or workflows (SWGDAM). It is indeed more expensive to increase the number of replicates for MPS kits and workflows than it is for electrophoresis workflows.

Secondly, the criteria for establishing precision for forensic DNA profiles is surprisingly more forgiving, as precision is based on whether the same genotype and profile is observed in each replicate [20]. It is unlikely to obtain contradicting observations of alleles when testing

conditions remain constant (*i.e.*, same run conditions, same plates set-up, same laboratory). For this reason, it is suggested that fewer replicates are required when testing for precision of forensic DNA profiling workflows (*i.e.*, when assessing whether the same DNA profile will be obtained across replicates). In this study, a high precision was obtained across all control DNA replicates, but also for bone, teeth, nail and crude buccal swab lysates (Chapter 5).

Lastly, the number of replicates will also depend on what test is being conducted. For example, it is acceptable to test a control DNA sample, or a known, good quality sample in triplicate at minimum. However, when testing for sensitivity, it is expected that low input DNA samples will result in higher stochastic variation, impacting precision negatively [220]. In this study, for input amounts known to produce good quality profiles in the developmental validation (*i.e.*, 1ng - 0.25 ng), three replicates were used, however, when testing lower input amounts, the replicates were increased to gain more insight into the low input DNA sample performance.

This variability in the number of replicates tested for different input amounts may be a limitation of this study, as it introduces bias in the lower input amounts tested. Although this is justified given that every effort was to be made to maximise the reagents available for the internal validation study. What *was* prioritised in the internal validation study was thus obtaining more insight into more challenging samples in order to understand the increase in stochastic artefacts and variation in genotypes obtained.

Furthermore, it would be imprudent to test fewer replicates of low input DNA samples when there would be a possibility of the results being too variable to obtain any meaningful insights from it. This threads into the theme of this study, which was the leveraging of information that already exists in a meaningful way and adapting it for a low-resourced laboratory. More

specifically, as developmental validation *experiments* can be used as blueprints for designing internal validation experiments, the *results* obtained from both developmental validation studies, as well as those published by other users of the workflow can be used to inform internal validation experiments.

For example, when it is known that a control DNA sample with a 1 ng input would result in a full DNA profile over 95 replicates (such as in the developmental validation), then internal validation experiments need not test as many replicates for those samples, but rather spare resources for testing more replicates of challenging samples and low input DNA samples, where increased variability is expected [20]. Additionally, this also emphasises the need for all laboratories to publish their internal validation results, so that later users of the kit or workflow can leverage those results to inform their own experiments in a sustainable way.

6.3. Technical limitations and areas for improvement

A technical limitation of the systematic review (Chapter 2) is that the search was only limited to population studies conducted with the ForenSeq™ DNA Signature Prep kit. While there are other MPS kits developed for forensic use, e.g., IDSeek® (Nimagen, Nijmegen, Netherlands), ForenSeq™ MainstAY kit (Verogen, San Diego, CA, USA), ForenSeq™ Kintelligence kit (Verogen, San Diego, CA, USA), Precision ID Identity Panel (Thermo Fisher Scientific, Waltham, MA, USA), and the PowerSeq® 46GY Panel (Promega Corporation, Madison, WI, USA), they are relatively new to the market and thus the number of publications resulting from them were low. Moreover, the ForenSeq™ DNA Signature Prep kit was the first forensically validated kit, and the aim of the systematic review was to assess the consensus on an increase in variation and concordance between MPS and CE over a set of markers. Inclusion of additional sequencing kits in inclusion criteria may not allow for a direct comparison between

markers and across populations. However, once more MPS kits are as firmly established as the ForenSeq™ DNA Signature Prep kit, future research can assess current advances with respect to forensically validated MPS kits that encompass a range of different sequencing technologies, kits and markers. Additionally, the systematic search was not undertaken by a second reviewer, although the results were reviewed. The systematic review will benefit from a second reviewer, and this will be done prior to submission of the work for publication.

Chapter 3 focused on novel adaptations to the protocols used to generate MPS profiles from crude buccal swab lysates. The first limitation of this study was the small number of samples used (n = 10). Due to the initial unforeseen failures resulting from the initial testing of crude buccal swab lysates, resources were severely constrained. There were 50 leftover library preparation reactions available, and so to make the best use of available resources, 10 samples were subjected to 5 different optimisation methods, followed by library preparation with the ForenSeq™ DNA Signature Prep kit. Although the sample size was small, the paired data design reduced variability and increased statistical power of the experiment.

A second limitation of this optimisation study was that only methods that resulted in optimal library quality, as determined by TapeStation, were taken through the full sequencing workflow, while other optimisation methods tested in Chapter 3, that did not result in optimal library quality and quantity were not tested. This is because the sequencing cartridges had already been allocated for the population and validation studies, with some of it having been used in the earlier failed experiments. There were insufficient funds to purchase an additional cartridge, and thus the libraries generated from all optimisation experiments were not sequenced. However, the quality and quantity of these libraries were assessed on TapeStation to inform sequencing success. While the use of the library quality control step allowed for

meaningful insights to be obtained within the imposed resource limitations, a smaller study can and should be undertaken to fully understand the effects of all optimisation methods used in Chapter 3, on the sequencing success of crude and purified buccal swab lysates.

In the population study conducted on the South African population groups (Chapter 4), a limitation in the reporting of allele frequency data was the absence of three markers that exhibited allele dropout and are known to be problematic with the ForenSeq™ DNA Signature Prep kit. The absence of data from these markers meant that only allele frequencies from 24 A-STR markers were reported. Inclusion of their data however would have resulted in false levels of homozygosity due to high counts of allele dropout at each of these markers. Their inclusion would certainly improve the discriminatory power of the database but would require further optimisation in primer design to obtain more reliable genotype calls, prior to being included in this sequence-based allele frequency database.

In highlighting future research endeavours, the additional data generated for X-STRs, Y-STRs and iiSNVs are being actively established in our research group at the University of Cape Town. The A-STR allele frequency database was generated first as a pilot for all other markers to follow. Procedures for data cleaning, variant characterisation and naming were established for A-STRs and thereby, in-house capacity was developed in preparation for generating sequence-based allele frequency data for the additional markers. Furthermore, population data for these additional markers may improve kinship assessments, as LRs would be based on an increased number of markers for comparison, with available population allele and/or haplotype frequencies. Additionally, this study only focused on the two major South African population groups, due to them making up the majority of unidentified human remains in the Western Cape and making up the largest proportions of the South African population [185]. A

continuation of this study should include the sequencing of samples from the South African White and Indian/Asian population groups, to ensure that our allele frequency data base is representative of the population.

In the internal validation component of this study (Chapter 5), three replicates were used for input amounts known to produce good quality profiles (1 ng and 0.25 ng) and for all control DNA samples. When testing the sensitivity of the workflow applied to forensic sample types, the number of replicates was increased for low input (0.01525 ng) samples to achieve a more reliable precision value, which may limitation of this study as the number of replicates across samples were not consistent. However, this approach aimed to maximise reagent use and gain maximum insights into challenging samples, focusing on understanding the rise in stochastic effects and variation, as discussed in section 6.2. Furthermore, the developmental validation could not be replicated in many aspects due to the number of samples used, the availability of alternative control DNA samples the testing of a range of different PCR inhibitors. However, this internal validation was designed to leverage the developmental validation experiments and results in way that would enable the maximising of resources, a more importantly it was tailored to the samples that were relevant to our laboratory (post-mortem sample types). The validation could be extended to include additional sample types such as mixtures, as well as test a broader range of PCR inhibitors.

6.4. Future directions: Applications of the workflow to forensic cold cases

The developmental validation and several evaluations of performance studies show that the ForenSeq™ DNA Signature Prep kit workflow outperforms CE in its ability to produce DNA profiles with a higher resolution power, produces significantly more information that can be used as investigative leads and enables a one-test strategy for obtaining maximum genetic

insights for DNA samples that are compromised, as opposed to using several different STR multiplexes to obtain data for all STR marker types (A-, Y- and X-STRs) [20,81-85] .

This thesis has contributed to the data that supports these conclusions. Importantly, the thesis has provided novelty in terms of 1) insights into failed crude buccal swab lysates, and ways to ensure their high first-time success rates, 2) understanding the global disparities that exist regarding published sequence-based population data, and generating this data for Africa and South Africa for the first time, 3) validation of post-mortem crude buccal swab lysates, which has not been validated before and 4) application of the internally validated workflow to a cold case, where for the first time in Africa, additional phenotype information could be layered into digitally recreated facial images of an unidentified deceased individual. With these objectives having been met, it is important to position this workflow in the context of solving cases involving compromised human bodies that are visually unrecognisable.

When a body is found that is visually unrecognisable, available and suitable samples are collected for DNA analysis [3]. If a sample of adequate quality is collected for DNA analysis, generating the DNA profile would be a manageable process using traditional CE systems. Although, in cases where no match is found in a countries national forensic DNA database, the DNA profiling results would be uninformative evidence. The crucial dilemma at hand in the absence of a DNA match, is thus the absence of familial reference data from potential family members that can be linked to the deceased. One way to link visually unrecognisable bodies of deceased individuals to family relatives is through generating possible investigative leads with respect to the phenotype of an individual [221].

In this study, the generation of hair colour, eye colour and biogeographical ancestry, in combination with forensic art and forensic anthropology, resulted in a facial image with colour, that was distributed by the South African Police Services to aid in identifying a severely decomposed human body (Chapter 5). Importantly, this digitally recreated image reached potential family members, motivating them to provide DNA samples for kinship assessment. However, a significant limitation of the phenotype and ancestry results, is that they were based on reference population data from non-African countries. Although the reference databases contain a large contribution from European countries, the major ancestries are represented. Inclusion of reference data from South African individuals may improve the accuracy of the predictions made.

Further information that could be ascertained from this study regarding the identity of the individual was the DNA profile frequency, which was determined to be very low, despite having used a minimum allele frequency approach at three markers, D12S391, D2S1338 and D9S1122. This links to a point made in Chapter 4's discussion, where it was mentioned that "for highly polymorphic markers such as D12S291, a sample size of ±200 may not be sufficient to capture and identify the variation present at this marker", as also pointed out by Devesse *et al*., 2020 [117]. The discussion further emphasised that this implication might not be entirely problematic in forensic casework using MPS, as the increase in the number of markers significantly enhances discriminatory power, even without delving into sequence-level discrimination. Indeed, a rare allele was observed a the D12S391 marker in the DNA sample from the cold case.

This demonstrates that irrespective of whether the allele frequency data generated in Chapter 4 was not entirely representative of the South African population, the increased number of

markers still resulted in a high level of discrimination and affirms that the sample sizes used in our population study was sufficient in this instance. Having fulfilled the objectives relating to generating sequence-based allele frequency data and internally validating the workflow, the findings from these chapters were fully harnessed to make informative contributions to an ongoing cold case investigation and demonstrates value of the workflow in its entirety.

6.5.1. Adapting the legal and qualitative framework for forensic humanitarian purposes

This thesis presents the first time in South Africa that MPS was used to generate phenotypic characteristics for a cold case, under a research banner. The facial image is currently the only way for a potential family member to be motivated to come forward to provide a reference sample, provided the image is recognisable and accessible by a potential family member, as was demonstrated in this study. It is therefore strongly warranted that for humanitarian purposes, there is some legal revision with regards to generating phenotypic investigative leads for linking of missing persons and unidentified human bodies to their families.

One limiting factor in the use of MPS to generate investigative and phenotypic leads in a South African context is that the Criminal Law (Forensic Procedures) Amendment Act 37 of 2013 (the DNA Act) prohibits the generation of DNA profile data that could infer physical or mental characteristics of an individual, and this therefore makes inferring phenotype challenging from a forensic perspective [222]. The ethical implication of estimation of phenotypic traits is that by virtue of having the same hair and eye colour as the suspect, one could be indirectly thought of a suspect, and this is particularly an issue with minority groups. However, it has also been argued that these traits are already externally visible and by nature are not private [223].

This implication may hold true for criminal suspects, but a different consideration should be made for unidentified human remains. From a humanitarian perspective, generating externally visible characteristics are one of the very few ways in which an image can be created from a face that has been severely decomposed and is visually unrecognisable, or where only parts of a body are used for DNA profiling. When combined with forensic art, the generation of externally visible characteristics have shown to generate accurate investigative leads that have contributed to the closing of cold cases [224].

The issue regarding the absence of familial reference data may be exacerbated by family members not always be willing to provide a DNA sample to the police, as there is often a lack of awareness in what happens to a DNA profile once it is added to a national forensic DNA database [37]. A survey conducted in Nigeria also revealed that among 458 participants, approximately 50% were unaware of what a forensic DNA database was [225]. This finding may hold true in other African countries. Furthermore, there is fear and reprise of what can be learned from this genetic information, and with whom it is shared [37].

This mistrust is not necessarily misplaced in a South African context, given the politicised discourse around the corruptive stigma embodying law enforcement agencies in South Africa [226]. Subsequently, family members of missing persons may end up not submitting a sample to the South African Police Services, and no link would be established between the deceased and families of missing persons. The mistrust in state authorities is a crucial dilemma, and one that must be addressed to increase trust and awareness in police bodies.

Matching STR profiles against a database of missing persons or their families has low success rates in South Africa, primarily because families are often reluctant to approach the police to

provide DNA samples. This issue called for the need to generate phenotype data, and thus DPMB was validated and used throughout all experiments in this study. A further justification for the use of this panel is needed, since it is up to three times the cost of the DPMA [73]. This added cost is due to the multiplexing capacity of each primer set: only 32 samples can be batched when using DPMB, whereas 96 reference samples can be batched using DPMA. Despite the higher cost, the DPMB was chosen in this study. The first reason for this was the throughput consideration in the local setting, where it was anticipated that batch sizes of 32 samples would be processed regardless, thus there would be no financial saving if DPMA was used. Secondly, the nature of cases encountered at the forensic mortuary was considered: sample types that are encountered are often far too compromised to generate STR profiles that could be compared to alleged family members, as reference and familial profiles are absent in most cases. In these cases, the DPMB would offer insight into phenotype and ancestry, as demonstrated in the application of this workflow to a cold case. The third reason for choosing DPMB was that initial testing showed that DPMB outperformed DPMA regarding call rates and overall performance. The decision was thus fortified as using DPMB maximised success rates and would have reduced the need for re-testing. While these findings require further exploration, DPMB was adopted during this study to optimise outcomes in a costly process.

With respect to cold cases, DPMB could be advantageous where CE is unsuccessful, and no match to a family member is found. One notable benefit is that generating phenotype characteristics and facial images could reduce the time spent tracing family members by allowing the information to reach the broader community at a faster rate. Furthermore, the mortuary is currently used as a place of storage for bodies that are both unclaimed and/or unidentified. This cost has been previously alluded to but given the reprise of family members to provide a DNA sample to the SAPS, there are currently limited methods that can be used to

link these bodies to their families. Using the DPMB to generate new investigative leads may lead to earlier identification, which in turn, would alleviate these storage costs.

On the other hand, it may also be argued that once the DPMB is used to generate an investigative lead, more resources and time would be allocated to such a case, since more families would enquire about these cases, and thus more investigative work and DNA analyses would need to be conducted. If DPMA is used, the cost of analysis can potentially be reduced to one third of the cost of the DPMB panel. However, this means that DNA analysis using MPS would need to wait until a batch of 96 samples is received. This, together with the shelf-life of the kit, was not deemed a pragmatic approach.

When assessing the cost requirements of a cold case, it is also imperative to determine whether the increase in output provided with MPS is necessary for *all* casework. In cases where unidentified bodies are fresh, and facial photography has been undertaken, CE analysis may be enough for comparison to a family member. Where samples are degraded or compromised, newly developed CE kits can be used that multiplex and incorporate more mini-STR markers. This comparison would still require that family members are confident in law enforcement and that they regularly come forward to provide DNA samples, which, as discussed earlier, is seldom the case in a South African context. The increase in the number of alleles provided with MPS is likely not necessary all cases. In fact, the increase in allele counts may not have a high impact on identification at all, but the ability to generate phenotype and ancestry does. Thus, the context-specific benefit of MPS lies in its ability to address a critical operational niche; where bodies visually unrecognisable bodies, and no STR match is made to a database.

Having addressed the costs of MPS, and the adaptation of a legal framework for consideration of generating phenotypic data, and increasing the trust and awareness in state authorities, another factor that may enable the gaining of maximum insights from a forensic MPS workflow is the adaptation of stringent SOPs informing downstream analyses. Expanding on this, a forensic MPS workflow will likely be used for samples that are compromised, given that it is positioned to be applied to cases where bodies are severely decomposed, mutilated or charred [20,81-85]. These samples may have to undergo optimisation to maximise every opportunity to obtain a full DNA profile on CE and MPS systems. Furthermore, even samples with seemingly low concentrations can still generate informative data, and it is suggested that SOPs should remain open to processing samples despite this.

It is not uncommon for State Forensic Science Laboratories to implement quantitative and qualitative cut-off values for DNA samples to inform their potential value in downstream processing. Thus, when a sample does not meet the quantitative and qualitative criterion, no DNA profile information will be generated from the DNA sample. When drawing on the results from this study, it is observed that the use of the Quantifiler® Trio kit was unreliable in informing downstream processing. Laboratories that have these qualitative and quantitative cut-off values in place should therefore consider the reliability of their quantitative and qualitative assessments, and whether having these cut-off values in place may come at a greater cost due to reduced case resolution. It is therefore suggested that these laboratories increase the adaptability of their quality management systems to accommodate cases where samples that might require further optimisation, or perhaps more daringly, the abolition of those cut-off values for humanitarian purposes altogether.

With that being said, this recommendation may come at a cost if a sample is taken through the workflow and does *not* yield adequate results. However, there is also a cost associated with the management and storage of unidentified bodies at the Observatory Forensic Pathology Institute in the Western Cape, and with resources required to investigate cases with no leads. A cost-benefit analysis, such as one carried out by Budowle *et al.*, 2021 regarding the cost benefits of investing in large SNV panels to generate forensic investigate leads may be used in a future study to ascertain how these costs would compare against one another [227].

It is acknowledged that Chapter 3 of this thesis recommended that a subset of samples should undergo qualitative and quantitative screening prior to sequencing, but this recommendation is made with respect to studies where a sample type will be used on a large-scale (population or validation study) and therefore requires rigorous optimisation *prior* to implementation. The recommendation made here is with respect to cold cases, where the use of the above-mentioned cut-off values may limit important case insights provided by MPS, especially in its infant stages.

Nonetheless, with the increase in sensitivity and insights provided by MPS as opposed to CE, as demonstrated in Chapter 5 (application to a forensic cold case), even degraded and low concentration samples can still provide plenty of insight to generate forensic investigative leads. Having quantitative or qualitative cut-off values in place could prevent a laboratory from leveraging the full benefit of using MPS for forensic humanitarian purposes and result in a missed opportunity to generate an investigative lead that could reach a family member.

6.6. Conclusion

This thesis aimed to facilitate the implementation of the ForenSeq™ DNA Signature Prep kit workflow for forensic human identification in South Africa. Through a series of inter-related studies, this work has advanced the understanding of how to effectively and sustainably implement a forensic MPS workflow in a South African context. The systematic review highlighted critical gaps in MPS population studies, especially the lack of sequence data from African countries. These gaps emphasised the need for more population data from developing regions, improved concordance reporting and establishment of direct PCR protocols to mitigate time and cost constraints of MPS population studies. Subsequent studies addressed these gaps.

For example, the optimisation study successfully addressed the low first-pass success rates of crude lysates when processed with the ForenSeq™ DNA Signature Prep kit, with call rates improving from 44% to 95%. This served as a foundational step for mitigating costs when conducting sequence-based population studies in developing regions. Subsequently, the first sequence-based allele frequency data for South African population groups were generated, with findings indicating high concordance between MPS and CE methods. This marks a significant step toward the implementation of MPS in South Africa. The findings contribute novel insights, including previously undescribed length-and sequence-based alleles and markers with unexpected flanking region variation.

The generation of these novel and rich data presented only for A-STRs in this study, represents a cog in the machine that is driving implementation of MPS. It is however acknowledged that achieving implementation will require characterisation of additional markers for the South African population, such as sex-linked markers and SNVs. At the same time, the increased genetic variation offered by MPS, and especially for the South African population, brought

new challenges, particularly in maintaining compatibility with CE methods, as it led to more discordances. Therefore, laboratories in the infancy stages of MPS implementation should prioritise comprehensive allele reporting and concordance practices and collaborate with institutions that have already conducted sequence-based population studies. Submitting sequence data to quality control databases like STRidER is recommended to identify and correct anomalies caused by the ForenSeq™ UAS's allele naming discrepancies. By integrating these strategies into the implementation process, laboratories can improve the quality of sequence data and streamline the implementation of MPS workflows, particularly for highly diverse populations and in resource-limited settings.

Taking a further step towards implementation, the validation study confirmed that the ForenSeq™ DNA Signature Prep kit workflow is fit for purpose in our mortuary-linked forensic laboratory. This study provided first-time insights into the performance parameters of post-mortem crude buccal swab lysates processed using a direct PCR approach. Acceptable performance parameters were established across various forensic sample types. The workflow demonstrated stability under degradation and inhibition conditions. However, further work is required to test the workflows' tolerance to humic acid as an inhibitor. Despite this limitation, the validated workflow was successfully applied to a cold case, generating a forensic investigative lead that contributed to the creation of a digital facial image, marking a first for Africa.

The main obstacle faced in the investigation of the forensic cold case was the absence of familial reference data, and it remains a critical challenge to link unidentified bodies to family members. The forensic investigative lead generated with the validated workflow, and its use in the creation of a digital facial image demonstrate that the workflow can be used as an

investigative approach to link families to their loved ones. The success of this approach first requires state authorities to prove that communities can trust them with sensitive data. It secondly necessitates revisions to the legal framework to allow for the generation of phenotype data, at least for humanitarian purposes. Lastly, the success of the MPS workflow requires the flexibility of SOPs, especially when dealing with challenging cases, such as cold cases involving degraded samples, as important insights can still be gained from compromised samples, as demonstrated in this thesis.

This doctoral thesis led to the generation of a forensic investigative lead that made a valuable contribution to a cold case investigation. This achievement demonstrates the value of the validated MPS workflow to support humanitarian efforts, particularly in identifying severely decomposed bodies. Achieving this final objective was a testament to the achievement of all other objectives in this study and is what drove the main aim towards completion. By addressing the specific challenges and opportunities represented in an African context, this thesis contributes to uplifting forensic genomic research in Africa and adds to the growing body of global forensic genomic research. Considering these findings, recommendations, and limitations, this thesis has made substantial leaps toward implementing the ForenSeq™ DNA Signature Prep kit workflow in South Africa, while also identifying critical areas that must be addressed to see implementation to fruition.

Reference list

[1] K.M. Reid, L.J. Martin, L.J. Heathfield, Understanding the Burden of Unidentified Bodies: A Systematic Review, International Journal of Legal Medicine 137(4) (2023) 1193-1202 https://doi.org/10.1007/s00414-023-02968-5.

[2] K.M. Reid, L.J. Martin, L.J. Heathfield, Bodies without Names: A Retrospective Review of Unidentified Decedents at Salt River Mortuary, Cape Town, South Africa, 2010 - 2017, South African Medical Journal 110(3) (2020) 223-228 https://doi.org/10.7196/SAMJ.2020.v110i3.14192.

[3] C.A. Keyes, T.J. Mahon, A. Gilbert, Human Decedent Identification Unit: Identifying the Deceased at a South African Medico-Legal Mortuary, International Journal of Legal Medicine 136(6) (2022) 1889-1896 https://10.1007/s00414-022-02893-z.

[4] South Africa, National Health Act: Regulations: Rendering of Forensic Pathology Service, 2018.

[5] South African Government, Inquests Act 58 of 1959, 58, 1959.

[6] L. Solomons, 'My Heart Is Shattered': Body Found in Thabo Bester's Cell Finally Identified. <https://www.news24.com/news24/southafrica/news/my-heart-is-shattered-body-found-in-thabo-besters-cell-finally-identified-20230422>, 2023 (accessed 12 August.2024).

[7] G. Majola, Financial Desperation Driving Fraudulent Claims High. <https://www.iol.co.za/personal-finance/financial-planning/financial-desperation-driving-fraudulent-claims-high-5dbc9fb8-4a0c-4e5d-82b9-aefcd74c13a8>, 2024 (accessed 26 August.2024).

[8] R.L. Sykes LM, Bernitz H, Fraudulent Records - Grave Forensic Consequences, South African Dental Journal 76(5) (2020) 3 http://dx.doi.org/10.17159/2519-0105/2020/v75no5a8

[9] I.A. Bianchi, M.B. Focardi, R. Grifoni, S. Raddi, A. Rizzo, B. Defraia, V. Pinchi, Dental Identification of Unknown Bodies through Antemortem Data Taken by Non-Dental X-Rays. Case Reports, Journal of Forensic Odonto-Stomatology 39(3) (2021) 49-57.

[10] T.J. Chirau, J. Shirinde, C. McCrindle, Access to Healthcare by Undocumented Zimbabwean Migrants in Post-Apartheid South Africa, African Journal of Primary Health Care & Family Medicine 16(1) (2024) 1-8 https://doi.org/10.4102/phcfm.v16i1.4126.

[11] E. Ziętkiewicz, M. Witt, P. Daca, J. Żebracka-Gala, M. Goniewicz, B. Jarząb, M. Witt, Current Genetic Methodologies in the Identification of Disaster Victims and in Forensic Analysis, Journal of Applied Genetics 53(1) (2012) 41-60 https://doi.org/10.1007/s13353-011-0068-7.

[12] J.M. Butler, Overview and History of DNA Typing, Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers, Elsevier Academic Press, Burlington, USA, 2005.

[13] J.M. Butler, DNA Separation Methods: Slab Gel and Capillary Electrophoresis, Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers, Elsevier Academic Press, Burlington, USA, 2005.

[14] J.M. Butler, Forensic Issues: Degraded DNA, PCR Inhibtion, Contamination, Mixed Samples and Low Copy Number, Forensic DNA Typing: Biology, Technology, and Genetics of Str Markers, Elsevier Academic Press, Burlington, USA, 2005.

[15] J.M. Butler, The Future of Forensic DNA Analysis, Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 370(1674) (2015) https://10.1098/rstb.2014.0252.

[16] Applied. Biosystems, Globalfiler™ and Globalfiler™ IQC PCR Amplification Kits User Guide, 2019.

[17] Promega, Powerplex® 35GY System for Use on the Spectrum Compact Ce System Technical Manual, 2023.

[18] M.J. Alvarez-Cubero, M. Saiz, B. Martínez-García, S.M. Sayalero, C. Entrala, J.A. Lorente, L.J. Martinez-Gonzalez, Next Generation Sequencing: An Application in Forensic Sciences?, Annals of Human Biology 44(7) (2017) 581-592 https://10.1080/03014460.2017.1375155.

[19] D. Ballard, J. Winkler-Galicki, J. Wesoły, Massive Parallel Sequencing in Forensics: Advantages, Issues, Technicalities, and Prospects, International Journal of Legal Medicine 134(4) (2020) 1291-1303 https://10.1007/s00414-020-02294-0.

[20] A.C. Jäger, M.L. Alvarez, C.P. Davis, E. Guzmán, Y. Han, L. Way, P. Walichiewicz, D. Silva, N. Pham, G. Caves, J. Bruand, F. Schlesinger, S.J.K. Pond, J. Varlaro, K.M. Stephens, C.L. Holt, Developmental Validation of the MiSeq Fgx™ Forensic Genomics System for Targeted Next Generation Sequencing in Forensic DNA Casework and Database Laboratories, Forensic Sciece International: Genetics 28 (2017) 52-70 https://10.1016/j.fsigen.2017.01.011.

[21] Verogen, Forenseq™ DNA Signature Prep Kit Reference Guide, 2022.

[22] A.D. Ambers, J.D. Churchill, J.L. King, M. Stoljarova, H. Gill-King, M. Assidi, M. Abu-Elmagd, A. Buhmeida, M. Al-Qahtani, B. Budowle, More Comprehensive Forensic Genetic Marker Analyses for Accurate Human Remains Identification Using Massively Parallel DNA Sequencing, BMC Genomics 17(Suppl 9) (2016) 750 https://10.1186/s12864-016-3087-2.

[23] J.M. Butler, STR Population Database Analyses, Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers, Elsevier Academic Press, Burlington, USA, 2005.

[24] W.J. Amankwaa Aaron, McNevin Allan et al.,, Forensic Evidence Processing in Gender-Based Violence Cases: Handbook for Criminal Justice Practitioners., United Nations Office on Drugs and Crime,, Vienna, Austria, 2024.

[25] I.S. Organisation, ISO/IEC 17025:2017: General Requirements for the Competence of Testing and Calibration Laboratories., ISO, Geneva, 2017.

[26] ILAC. International Laboratory Accreditation Cooperation, Modules in a Forensic Science Process, ILAC-G19:06/2022., ILAC, Silverwater, Australia, 2022.

[27] M.C. Campbell, S.A. Tishkoff, African Genetic Diversity: Implications for Human Demographic History, Modern Human Origins, and Complex Disease Mapping, The Annual Review of Genomics and Human Genetics 9 (2008) 403-433 https://doi.org/10.1146/annurev.genom.9.081307.164258.

[28] S.E. Cavanaugh, A.S. Bathrick, Direct PCR Amplification of Forensic Touch and Other Challenging DNA Samples: A Review, Forensic Science International: Genetics 32 (2018) 40-49 https://doi.org/10.1016/j.fsigen.2017.10.005.

[29] ENFSI Working Group, Recommended Minimum Criteria for the Validation of Various Aspects of the DNA Profiling Process 2010.

[30] A. Senst, A. Caliebe, E. Scheurer, I. Schulz, Validation and Beyond: Next Generation Sequencing of Forensic Casework Samples Including Challenging Tissue Samples from Altered Human Corpses Using the Miseq Fgx™ System, Journal of Forensic Science 67(4) (2022) 1382-1398 https://10.1111/1556-4029.15028.

[31] M. Watney, DNA in the Courtroom: Principles and Practice, L. Meintjes-Van Der Walt : Book Review, Journal of South African Law / Tydskrif vir die Suid-Afrikaanse Reg 2013(1) (2013) 195-196 https://doi.org/:10.10520/EJC130599.

[32] C. Phillips, K.B. Gettings, J.L. King, D. Ballard, M. Bodner, L. Borsuk, W. Parson, "The Devil's in the Detail": Release of an Expanded, Enhanced and Dynamically Revised Forensic Str Sequence Guide, Forensic Science International: Genetics 34 (2018) 162-169 https://doi.org/10.1016/j.fsigen.2018.02.017.

[33] P. Johnson, R. Williams, P. Martin, Genetics and Forensics: Making the National DNA Database, Science and Technology Studies 16(2) (2003) 22-37.

[34] FBI. Federal Bureau of Investigation, CODIS-NDIS Statistics. <https://le.fbi.gov/science-and-lab/biometrics-and-fingerprints/codis/codis-ndis-statistics>, 2024 (accessed 27 August.2024).

[35] National DNA Database, National DNA Database Annual Report 2011-2012, 2013.

[36] J. Smith, J. Horne, The Value of Forensic DNA Investigative Leads in South Africa, Journal of Forensic Science and Criminal Investigation, 17 (2024) 1177-1184 https://juniperpublishers.com/jfsci/pdf/JFSCI.MS.ID.555969.pdf

[37] A. Ahmed, Ethical Concerns of DNA Databases Used for Crime Control. <https://blog.petrieflom.law.harvard.edu/2019/01/14/ethical-concerns-of-dna-databases-used-for-crime-control/>, 2019).

[38] J.M. Butler, Combined DNA Index System: CODIS and the Use of DNA Databases, Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers, Elsevier Academic Press, Burlington, USA, 2005.

[39] J.M. Butler, Profile Frequency Estimates, Likelihood Ratios, and Source Attribution, Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers, Elsevier Academic Press, Burlington, USA, 2005.

[40] J. Ge, H. Sun, H. Li, C. Liu, J. Yan, B. Budowle, Future Directions of Forensic DNA Databases, Croat Med J 55(2) (2014) 163-166 https://10.3325/cmj.2014.55.163.

[41] J.M. Butler, Statistical Interpretation: Evaluating the Strength of Forensic DNA Evidence, in: J.M. Butler (Ed.), Fundamentals of Forensic DNA Typing, Academic Press, San Diego, 2010, https://doi.org/10.1016/B978-0-12-374999-4.00011-4, pp. 229-258.

[42] J.M. Butler, Basic Genetic Principles, Statistics, and Probability, Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers, Elsevier Academic Press, Burlington, USA, 2005.

[43] Cybergenetics, NBC Dateline Presents the First Trueallele Case, Pittsburgh, USA, 2019.

[44] M.D. Coble, J.-A. Bright, Probabilistic Genotyping Software: An Overview, Forensic Science International: Genetics 38 (2019) 219-224 https://doi.org/10.1016/j.fsigen.2018.11.009.

[45] M.W. Perlin, M.M. Legler, C.E. Spencer, J.L. Smith, W.P. Allan, J.L. Belrose, B.W. Duceman, Validating Trueallele® DNA Mixture Interpretation, Journal of Forensic Sciences 56(6) (2011) 1430-1447 https://doi.org/10.1111/j.1556-4029.2011.01859.x.

[46] D.W. Bauer, N. Butt, J.M. Hornyak, M.W. Perlin, Validating Trueallele(®) Interpretation of DNA Mixtures Containing up to Ten Unknown Contributors, Journal of Forensic Sciences 65(2) (2020) 380-398 https://10.1111/1556-4029.14204.

[47] J.M. Butler, Kinship and Parentage Testing, Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers, Elsevier Academic Press2005.

[48] V. Mályusz, M. Poetsch, E. Simeoni, T. Schwark, O. Manfred, N. von Wurmb-Schwark, Problems of Assessing Sibship Probabilities by Means of Genetic Analysis, Arch Kriminol 218(1-2) (2006) 44-54.

[49] R. Li, H. Li, D. Peng, B. Hao, Z. Wang, E. Huang, R. Wu, H. Sun, Improved Pairwise Kinship Analysis Using Massively Parallel Sequencing, Forensic Science International: Genetics 38 (2019) 77-85 https://10.1016/j.fsigen.2018.10.006.

[50] D.T. Chung, J. Drábek, K.L. Opel, J.M. Butler, B.R. McCord, A Study on the Effects of Degradation and Template Concentration on the Amplification Efficiency of the STR Miniplex Primer Sets, J Forensic Sci 49(4) (2004) 733-740.

[51] N. Czado, B. LaRue, A. Wheeler, R. Houston, A. Holmes, K. Grisedale, S. Hughes, The Effectiveness of Various Strategies to Improve DNA Analysis of Formaldehyde-Damaged Tissues from Embalmed Cadavers for Human Identification Purposes, Journal of Forensic Sciences 68(2) (2023) 596-607 https://10.1111/1556-4029.15200.

[52] B. Jiang, W. He, C. Jin, Y. Liu, D. Wen, C. Wang, M.M.J. Zeye, J. Li, L. Zha, Developmental Validation of the STRscan-17LC Kit: A 6 Dye STR Kit Enhanced Stability and Ability to Detect Degraded Samples, International Journal of Legal Medicine 135(2) (2021) 431-440 https://10.1007/s00414-020-02490-y.

[53] J.M. Butler, Y. Shen, B.R. McCord, The Development of Reduced Size STR Amplicons as Tools for Analysis of Degraded DNA, Journal of Forensic Sciences 48(5) (2003) 1054-1064.

[54] K. Breslin, B. Wills, A. Ralf, M. Ventayol Garcia, M. Kukla-Bartoszek, E. Pospiech, A. Freire-Aradas, C. Xavier, S. Ingold, M. de La Puente, K.J. van der Gaag, N. Herrick, C. Haas, W. Parson, C. Phillips, T. Sijen, W. Branicki, S. Walsh, M. Kayser, Hirisplex-S System for Eye, Hair, and Skin Color Prediction from DNA: Massively Parallel Sequencing Solutions for Two Common Forensically Used Platforms, Forensic Science International: Genetics 43 (2019) https://10.1016/j.fsigen.2019.102152.

[55] T.M.T. Carratto, V.M.S. Moraes, T.S.F. Recalde, M.L.G. de Oliveira, C.T. Mendes-Junior, Applications of Massively Parallel Sequencing in Forensic Genetics, Genetics and Molecular Biology 45(3) (2022) 10.1590/1678-4685-GMB-2022-0077.

[56] F. Sanger, Sequences, Sequences, and Sequences, Annu Rev Biochem 57 (1988) 1-28 https://doi.org/10.1146/annurev.bi.57.070188.000245.

[57] Applied Biosystems 3500/3500xl Genetic Analyzer User Guide, 2010.

[58] S. Goodwin, J.D. McPherson, W.R. McCombie, Coming of Age: Ten Years of Next-Generation Sequencing Technologies, Nature Reviews Genetics 17(6) (2016) 333-351 https://doi.org/10.1038/nrg.2016.49.

[59] VEROGEN, Miseq Fgx Sequencing System Reference Guide, 2021.

[60] VEROGEN, Forenseq™ Universal Analysis Software Guide, 2018.

[61] P. Mueller, C. Sell, T. Hadrys, J. Hedman, S. Bredemeyer, F.X. Laurent, L. Roewer, S. Achtruth, M. Sidstedt, T. Sijen, M. Trimborn, N. Weiler, S. Willuweit, I. Bastisch, W. Parson, S.T.R.C. SeqFor, Inter-Laboratory Study on Standardized Mps Libraries: Evaluation of Performance, Concordance, and Sensitivity Using Mixtures and Degraded DNA, International Journal of Legal Medicine 134(1) (2020) 185-198 https://doi.org/10.1007/s00414-019-02201-2.

[62] R.H. England, S., A Review of the Method and Validation of the MiSeq Fgx™ Forensic Genomics Solution, WIREs Forensic Science 2(1) (2020) e1351 https://doi.org/10.1002/wfs2.1351.

[63] K. Shao, W. Ding, F. Wang, H. Li, D. Ma, H. Wang, Emulsion PCR: A High Efficient Way of PCR Amplification of Random DNA Libraries in Aptamer Selection, PLoS ONE 6(9) (2011) e24910 https://10.1371/journal.pone.0024910.

[64] S. Ambardar, R. Gupta, D. Trakroo, R. Lal, J. Vakhlu, High Throughput Sequencing: An Overview of Sequencing Chemistry, Indian Journal of Microbiology 56(4) (2016) 394-404 https://10.1007/s12088-016-0606-4.

[65] Illumina, Illumina Adapter Portfolio. <https://knowledge.illumina.com/library-preparation/general/library-preparation-general-reference_material-list/000003275>, 2023 (accessed 27 August.2024).

[66] Illumina, An Introduction to Next-Generation Sequencing Technology, 2017.

[67] A. Alonso, P. Müller, L. Roewer, S. Willuweit, B. Budowle, W. Parson, European Survey on Forensic Applications of Massively Parallel Sequencing, Forensic Science International: Genetics 29 (2017) 23-25 https://doi.org/10.1016/j.fsigen.2017.04.017.

[68] National Research Council (US) Committee on DNA ForensicScience: Population Genetics, An Update: The Evaluation of Forensic DNA Evidence, National Academic Press (US), Washington (DC), 1996, https://10.17226/5141, pp. 89-124.

[69] W. Parson, D. Ballard, B. Budowle, J.M. Butler, K.B. Gettings, P. Gill, L. Gusmão, D.R. Hares, J.A. Irwin, J.L. King, P.D. Knijff, N. Morling, M. Prinz, P.M. Schneider, C.V. Neste, S. Willuweit, C. Phillips, Massively Parallel Sequencing of Forensic STRs: Considerations of the DNA Commission of the International Society for Forensic Genetics (ISFG) on Minimal Nomenclature Requirements, Forensic Science International: Genetics 22 (2016) 54-63 https://10.1016/j.fsigen.2016.01.009.

[70] R. Wu, R. Li, N. Wang, D. Peng, H. Li, Y. Zhang, C. Zheng, H. Sun, Genetic Polymorphism and Population Structure of Torghut Mongols and Comparison with a Mongolian Population 3000 Kilometers Away, Forensic Science International: Genetics 42 (2019) 235-243 https://10.1016/j.fsigen.2019.07.017.

[71] F.R. Wendt, J.D. Churchill, N.M.M. Novroski, J.L. King, J. Ng, R.F. Oldt, K.L. McCulloh, J.A. Weise, D.G. Smith, S. Kanthaswamy, B. Budowle, Genetic Analysis of the Yavapai Native Americans from West-Central Arizona Using the Illumina MiSeq Fgx™ Forensic Genomics System, Forensic Science International: Genetics 24 (2016) 18-23 https://10.1016/j.fsigen.2016.05.008.

[72] C. Phillips, L. Devesse, D. Ballard, L. van Weert, M. de la Puente, S. Melis, V. Álvarez Iglesias, A. Freire-Aradas, N. Oldroyd, C. Holt, D. Syndercombe Court, Á. Carracedo, M.V. Lareu, Global Patterns of Str Sequence Variation: Sequencing the CEPH Human Genome Diversity Panel for 58 Forensic STRs Using the Illumina Forenseq™ DNA Signature Prep Kit, Electrophoresis 39(21) (2018) 2708-2724 https://doi.org/10.1002/elps.201800117

[73] N.M.M. Novroski, J.L. King, J.D. Churchill, L.H. Seah, B. Budowle, Characterization of Genetic Sequence Variation of 58 STR Loci in Four Major Population Groups, Forensic Science International: Genetics 25 (2016) 214-226 https://doi.org/10.1016/j.fsigen.2016.09.007.

[74] C. Hussing, R. Bytyci, C. Huber, N. Morling, C. Børsting, The Danish STR Sequence Database: Duplicate Typing of 363 Danes with the Forenseq™ DNA Signature Prep Kit, International Journal of Legal Medicine 133(2) (2019) 325-334 https://doi.org/10.1007/s00414-018-1854-0.

[75] E.K. Guevara, J.U. Palo, J.L. King, M.M. Buś, S. Guillén, B. Budowle, A. Sajantila, Autosomal STR and SNV Characterization of Populations from the Northeastern Peruvian Andes with the Forenseq™ DNA Signature Prep Kit, Forensic Science International Genetics 52 (2021) 102487 https://doi.org/10.1016/j.fsigen.2021.102487.

[76] M.M.J. Zeye, S.Y. Ouedraogo, P. Bado, A.A. Zoure, F.W. Djigma, X. Wu, J. Simpore, Forensic Autosomal and Gonosomal Short Tandem Repeat Marker Reference Database for Populations in Burkina Faso, Scientific Reports 14(1) (2024) 7369 https://doi.org/10.1038/s41598-024-58179-4.

[77] A.E. Kofi, H.M. Hakim, H.O. Khan, S.A. Ismail, A. Ghansah, A.R.N. Haslindawaty, S. Shamsuddin, M.Y. Aziz, G.K. Chambers, H.A. Edinur, Population Dataset for 21 Simple Tandem Repeat Loci in the Akan Population of Ghana, Data in Brief 31 (2020) 105746 https://doi.org/10.1016/j.dib.2020.105746.

[78] J.M. Muinde, D.R. Chandra Bhanu, R. Neumann, R.O. Oduor, W. Kanja, J.K. Kimani, M.W. Mutugi, L. Smith, M.A. Jobling, J.H. Wetton, Geographical and Linguistic Structure in the People of Kenya Demonstrated Using 21 Autosomal STRs, Forensic Science International: Genetics 53 (2021) 102535 https://doi.org/10.1016/j.fsigen.2021.102535.

[79] J. Li, L. Zha, Forensic Characteristics and Genetic Structure of 18 Autosomal STR Loci in the Sierra Leone Population, International Journal of Legal Medicine 135(2) (2021) 455-456 https://doi.org/10.1007/s00414-020-02487-7.

[80] L. Barbaric, I. Horjan-Zanki, Challenges in the Recovery of the Genetic Data from Human Remains Found on the Western Balkan Migration Route, International Journal of Legal Medicine 137(1) (2023) 181-193 https://doi.org/10.1007/s00414-022-02829-7.

[81] C. Finaughty, K.M. Reid, I.H. Alli, L.J. Heathfield, A First for Forensic Genetics in Africa: Successful Identification of Skeletal Remains from the Marine Environment Using Massively

Parallel Sequencing, Forensic Science International: Genetics 49 (2020) https://doi.org/10.1016/j.fsigen.2020.102370.

[82] F. Calafell, R. Anglada, N. Bonet, M. González-Ruiz, G. Prats-Muñoz, R. Rasal, C. Lalueza-Fox, J. Bertranpetit, A. Malgosa, F. Casals, An Assessment of a Massively Parallel Sequencing Approach for the Identification of Individuals from Mass Graves of the Spanish Civil War (1936-1939), Electrophoresis 37(21) (2016) 2841-2847 https://doi.org/10.1002/elps.201600180.

[83] P. Sukawutthiya, T. Sathirapatya, K. Vongpaisarnsin, Considering a Performance Test of Forensic Genomics System on Massively Parallel Sequencing Technology, Forensic Science International: Genetics Supplement Series 6 (2017) e599-e600 https://doi.org/10.1016/j.fsigss.2017.10.007.

[84] C. Xavier, W. Parson, Evaluation of the Illumina Forenseq™ DNA Signature Prep Kit - Mps Forensic Application for the Miseq Fgx™ Benchtop Sequencer, Forensic Science International-Genetics 28 (2017) 188-194 https://doi.org/10.1016/j.fsigen.2017.02.018.

[85] J.D. Churchill, S.E. Schmedes, J.L. King, B. Budowle, Evaluation of the Illumina® Beta Version Forenseq™ DNA Signature Prep Kit for Use in Genetic Profiling, Forensic Science International-Genetics 20 (2016) 20-29 https://doi.org/10.1016/j.fsigen.2015.09.009.

[86] L. Devesse, L. Davenport, L. Borsuk, K. Gettings, G. Mason-Buck, P.M. Vallone, D. Syndercombe Court, D. Ballard, Classification of STR Allelic Variation Using Massively Parallel Sequencing and Assessment of Flanking Region Power, Forensic Science International: Genetics 48 (2020) https://doi.org/10.1016/j.fsigen.2020.102356.

[87] F. Guo, Z. Liu, G. Long, B. Zhang, X. Dong, D. Liu, S. Yu, High-Resolution Genotyping of 58 STRs in 635 Northern Han Chinese with Miseq Fgx ® Forensic Genomics System, Forensic Sci Int Genet 65 (2023) 102879 https://doi.org/10.1016/j.fsigen.2023.102879.

[88] T. Sathirapatya, P. Sukawutthiya, K. Vongpaisarnsin, Massively Parallel Sequencing of 24 Y-STR Loci in Thai Population, Forensic Science International: Genetics Supplement Series 6 (2017) e310-e313 10. https://doi.org/1016/j.fsigss.2017.09.129.

[89] A. Pfennig, L.N. Petersen, P. Kachambwa, J. Lachance, Evolutionary Genetics and Admixture in African Populations, Genome Biol Evol 15(4) (2023) https://doi.org/10.1093/gbe/evad054.

[90] S.Y. Kim, H.C. Lee, U. Chung, S.K. Ham, H.Y. Lee, S.J. Park, Y.J. Roh, S.H. Lee, Massive Parallel Sequencing of Short Tandem Repeats in the Korean Population, Electrophoresis 39(21) (2018) 2702-2707 https://doi.org/10.1002/elps.201800090.

[91] M.M.A. Agudo, H.; Albert, M.; Janssen, K.; Gill, P.; Bleka, Ø., An Overview of Autosomal STRs and Identity SNVs in a Norwegian Population Using Massively Parallel Sequencing, Forensic Science International Genetics 71 (2024) 103057 https://doi.org/10.1016/j.fsigen.2024.103057.

[92] A. Barbaro, P. Cormaci, A. Teatino, A. Barbaro, Use of "Anydirect PCR Buffer" for PCR Amplification of Washed Bloodstains: A Case Report, Forensic Science International: Genetics Supplement Series 1(1) (2008) 11-12 https://doi.org/10.1016/j.fsigss.2007.10.201.

[93] A. Barbaro, P. Cormaci, S. Votano, Direct PCR by the AmpFLSTR® Ngm™ Kit for Database Purpose, Forensic Science International: Genetics Supplement Series 3(1) (2011) 103-104 https://doi.org/10.1016/j.fsigss.2011.08.051.

[94] B.A. Myers, J.L. King, B. Budowle, Evaluation and Comparative Analysis of Direct Amplification of STRs Using Powerplex® 18D and Identifiler® Direct Systems, Forensic Science International: Genetics 6(5) (2012) 640-645 https://doi.org/10.1016/j.fsigen.2012.02.005.

[95] B. Martin, A. Linacre, Direct PCR: A Review of Use and Limitations, Science & Justice 60(4) (2020) 303-310 https://doi.org/10.1016/j.scijus.2020.04.003.

[96] T.M.T. Carratto, V.M.S. Moraes, T.S.F. Recalde, M.L.G. Oliveira, C. Teixeira Mendes-Junior, Applications of Massively Parallel Sequencing in Forensic Genetics, Genetics and Molecular Biology 45(3 Suppl 1) (2022) e20220077 https://doi.org/10.1590/1678-4685-gmb-2022-0077.

[97] D. Moher, L. Shamseer, M. Clarke, D. Ghersi, A. Liberati, M. Petticrew, P. Shekelle, L.A. Stewart, P.-P. Group, Preferred Reporting Items for Systematic Review and Meta-Analysis

Protocols (Prisma-P) 2015 Statement, Systematic Reviews 4(1) (2015) 1 https://doi.org/10.1186/2046-4053-4-1.

[98] M.J. Page, J.E. McKenzie, P.M. Bossuyt, I. Boutron, T.C. Hoffmann, C.D. Mulrow, L. Shamseer, J.M. Tetzlaff, E.A. Akl, S.E. Brennan, R. Chou, J. Glanville, J.M. Grimshaw, A. Hróbjartsson, M.M. Lalu, T. Li, E.W. Loder, E. Mayo-Wilson, S. McDonald, L.A. McGuinness, L.A. Stewart, J. Thomas, A.C. Tricco, V.A. Welch, P. Whiting, D. Moher, The Prisma 2020 Statement: An Updated Guideline for Reporting Systematic Reviews, BMJ 372 (2021) n71 https://doi.org/10.1136/bmj.n71.

[99] Illumina, Targeted Next-Generation Sequencing for Forensic Genomics, 2015.

[100] R. Team, Rstudio: Integrated Development for R., PBC, Boston, MA, 2020.

[101] R. Wu, D. Peng, H. Ren, R. Li, H. Li, N. Wang, X. Shen, E. Huang, Y. Zhang, H. Sun, Characterization of Genetic Polymorphisms in Nigerians Residing in Guangzhou Using Massively Parallel Sequencing, Forensic Science International: Genetics 48 (2020) https://doi.org/10.1016/j.fsigen.2020.102323.

[102] C. Chen, X. Jin, X. Zhang, W. Zhang, Y. Guo, R. Tao, A. Chen, Q. Xu, M. Li, Y. Yang, B. Zhu, Comprehensive Insights into Forensic Features and Genetic Background of Chinese Northwest Hui Group Using Six Distinct Categories of 231 Molecular Markers, Frontiers in Genetics 12 (2021) https://doi.org/10.3389/fgene.2021.705753.

[103] F. Guo, Z. Liu, G.N. Long, B. Zhang, D.H. Liu, S.B. Yu, Performance and Characterization of 94 Identity-Informative SNVs in Northern Han Chinese Using Forenseq™ DNA Signature Prep Kit, Journal of Forensic and Legal Medicine 103 (2024) 13 https://doi.org/10.1016/j.jflm.2024.102678.

[104] D. Peng, Y. Zhang, H. Ren, H. Li, R. Li, X. Shen, N. Wang, E. Huang, R. Wu, H. Sun, Identification of Sequence Polymorphisms at 58 STRs and 94 iiSNVs in a Tibetan Population Using Massively Parallel Sequencing, Scientific Reports 10(1) (2020) 12225 https://doi.org/10.1038/s41598-020-69137-1.

[105] R. Tao, X. Dong, X. Zhen, R. Xia, Y. Qu, S. Liu, S. Zhang, C. Li, Population Genetic Analyses of Eastern Chinese Han Nationality Using Forenseq™ DNA Signature Prep Kit, Mol Genet Genomics 299(1) (2024) 9 https://doi.org/10.1007/s00438-024-02121-w.

[106] H. Fan, Z. Du, F. Wang, X. Wang, S.Q. Wen, L. Wang, P. Du, H. Liu, S. Cao, Z. Luo, B. Han, P. Huang, B. Zhu, P. Qiu, The Forensic Landscape and the Population Genetic Analyses of Hainan Li Based on Massively Parallel Sequencing DNA Profiling, International Journal of Legal Medicine 135(4) (2021) 1295-1317 https://doi.org/10.1007/s00414-021-02590-3.

[107] F. Casals, R. Rasal, R. Anglada, M. Tormo, N. Bonet, N. Rivas, P. Vásquez, F. Calafell, A Forensic Population Database in El Salvador: 58 STRs and 94 SNVs, Forensic Science International: Genetics 57 (2022) https://doi.org/10.1016/j.fsigen.2021.102646.

[108] A. Delest, D. Godfrin, Y. Chantrel, A. Ulus, J. Vannier, M. Faivre, C. Hollard, F.X. Laurent, Sequenced-Based French Population Data from 169 Unrelated Individuals with Verogen's Forenseq DNA Signature Prep Kit, Forensic Science International: Genetics 47 (2020) 102304 https://doi.org/10.1016/j.fsigen.2020.102304.

[109] J.A. Aguilar-Velázquez, M.Á. Duran-Salazar, M.F. Córdoba-Mercado, C.E. Coronado-Avila, O. Salas-Salas, G. Martinez-Cortés, F. Casals, F. Calafell, B. Ramos-González, H. Rangel-Villalobos, Characterization of 58 STRs and 94 SNVs with the Forenseq™ DNA Signature Prep Kit in Mexican-Mestizos from the Monterrey City (Northeast, Mexico), Mol Biol Rep (2022) https://doi.org/10.1007/s11033-022-07575-y.

[110] J.M. Salvador, D.L.T. Apaga, F.C. Delfin, G.C. Calacal, S.E. Dennis, M.C.A. De Ungria, Filipino DNA Variation at 12 X-Chromosome Short Tandem Repeat Markers, Forensic Science International Genetics 36 (2018) e8-e12 https://doi.org/10.1016/j.fsigen.2018.06.008.

[111] E. Almohammed, A. Iyengar, D. Ballard, L. Devesse, S. Hadi, Evaluation of Forenseq DNA Signature Kit for Qatari Population, Forensic Science International: Genetics Supplement Series 6 (2017) 596-598 https://doi.org/10.1016/j.fsigss.2017.10.003.

[112] E. Almohammed, S. Hadi, The Study of 95 Identity SNVs for Qatari Population Using Massively Parallel Sequencing (MPS), Forensic Science International: Genetics Supplement Series 7(1) (2019) 869-871 https://doi.org/10.1016/j.fsigss.2019.11.006.

[113] E. Almohammed, S. Hadi, The Study of Novel Sequence Alleles for Qatari Population Using Forenseqtm DNA Kit, Forensic Science International: Genetics Supplement Series 7(1) (2019) 872-874 https://doi.org/10.1016/j.fsigss.2019.11.007.

[114] Y.M. Khubrani, P. Hallast, M.A. Jobling, J.H. Wetton, Massively Parallel Sequencing of Autosomal STRs and Identity-Informative SNVs Highlights Consanguinity in Saudi Arabia, Forensic Sci Int Genet 43 (2019) 102164 https://doi.org/10.1016/j.fsigen.2019.102164.

[115] F. Casals, R. Anglada, N. Bonet, R. Rasal, K.J. van der Gaag, J. Hoogenboom, N. Solé-Morata, D. Comas, F. Calafell, Length and Repeat-Sequence Variation in 58 STRs and 94 SNVs in Two Spanish Populations, Forensic Science International: Genetics 30 (2017) 66-70 https://doi.org/10.1016/j.fsigen.2017.06.006.

[116] L. Devesse, D. Ballard, L. Davenport, I. Riethorst, G. Mason-Buck, D. Syndercombe Court, Concordance of the Forenseq™ System and Characterisation of Sequence-Specific Autosomal STRs Alleles across Two Major Population Groups, Forensic Science International: Genetics 34 (2018) 57-61 https://doi.org/10.1016/j.fsigen.2017.10.012.

[117] L.A. Devesse, D.J. Ballard, L.B. Davenport, K.B. Gettings, L.A. Borsuk, P.M. Vallone, D. Syndercombe Court, The Tao of MPS: Common Novel Variants, Forensic Science International: Genetics Supplement Series 6 (2017) 579-581 https://doi.org/10.1016/j.fsigss.2017.09.222.

[118] L. Davenport, L. Devesse, D. Syndercombe Court, D. Ballard, Forensic Identity STRs: Characterisation of Flanking Region Variation Using Massively Parallel Sequencing, Forensic Science International: Genetics 64 (2023) 102847 https://doi.org/10.1016/j.fsigen.2023.102847.

[119] C.R. Steffen, T.I. Huszar, L.A. Borsuk, P.M. Vallone, K.B. Gettings, A Multi-Dimensional Evaluation of the 'NIST 1032' Sample Set across Four Forensic Y-STR Multiplexes, Forensic Science International: Genetics 57 (2022) 102655 https://doi.org/10.1016/j.fsigen.2021.102655.

[120] J.L. King, J.D. Churchill, N.M.M. Novroski, X. Zeng, D.H. Warshauer, L.-H. Seah, B. Budowle, Increasing the Discrimination Power of Ancestry- and Identity-Informative SNV Loci within the Forenseq™ DNA Signature Prep Kit, Forensic Science International: Genetics 36 (2018) 60-76 https://doi.org/10.1016/j.fsigen.2018.06.005.

[121] J.D. Churchill, N.M.M. Novroski, J.L. King, L.H. Seah, B. Budowle, Population and Performance Analyses of Four Major Populations with Illumina's FGx Forensic Genomics

System, Forensic Science International: Genetics 30 (2017) 81-92 https://doi.org/10.1016/j.fsigen.2017.06.004.

[122] K.M. Kiesler, L.A. Borsuk, C.R. Steffen, K.B. Gettings, P.M. Vallone, Population Data for 94 Identity SNVs in Four U.S. Population Groups, Forensic Science International: Genetics Supplement Series 8 (2022) 7-8 https://doi.org/10.1016/j.fsigss.2022.09.003.

[123] C.R. Steffen, K.B. Gettings, K.M. Kiesler, L.A. Borsuk, P.M. Vallone, Sequence Variation Observed in 27 Y-STR Markers with U.S. Population Samples, Forensic Science International: Genetics Supplement Series 7(1) (2019) 520-521 https://doi.org/10.1016/j.fsigss.2019.10.074 .

[124] K.B. Gettings, L.A. Borsuk, C.R. Steffen, K.M. Kiesler, P.M. Vallone, Sequence-Based U.S. Population Data for 27 Autosomal STR Loci, Forensic Science International: Genetics 37 (2018) 106-115 https://doi.org/10.1016/j.fsigen.2018.07.013.

[125] L.A. Borsuk, C.R. Steffen, K.M. Kiesler, P.M. Vallone, K.B. Gettings, Sequence-Based U.S. Population Data for 7 X-STR Loci, Forensic Science International: Reports 2 (2020) https://doi.org/10.1016/j.fsir.2020.100160.

[126] L.A. Borsuk, K.B. Gettings, C.R. Steffen, K.M. Kiesler, P.M. Vallone, Sequence-Based Us Population Data for the SE33 Locus, Electrophoresis 39(21) (2018) 2694-2701 https://doi.org/10.1002/elps.201800091.

[127] K.M. Kiesler, L.A. Borsuk, C.R. Steffen, P.M. Vallone, K.B. Gettings, Us Population Data for 94 Identity-Informative SNV Loci, Genes 14(5) (2023) https://doi.org/10.3390/genes14051071.

[128] F.R. Wendt, J.L. King, N.M.M. Novroski, J.D. Churchill, J. Ng, R.F. Oldt, K.L. McCulloh, J.A. Weise, D.G. Smith, S. Kanthaswamy, B. Budowle, Flanking Region Variation of Forenseq™ DNA Signature Prep Kit STR and SNV Loci in Yavapai Native Americans, Forensic Science International: Genetics 28 (2017) 146-154 https://doi.org/10.1016/j.fsigen.2017.02.014.

[129] N.S. Pham, H.L. Tran, T.H.T. Nguyen, V.H. Nguyen, H. Hoang, Q.N. Tung, Q.T. Phi, The First Autosomal STR Population Data of Kinh Ethinic Group in Vietnam by Using

Massively Parallel Sequencing, Russian Journal of Genetics 57(8) (2021) 985-988 https://doi.org/10.1134/S102279542108010X.

[130] J. Atutornu, R. Milne, A. Costa, C. Patch, A. Middleton, Towards Equitable and Trustworthy Genomics Research, EBioMedicine 76 (2022) 103879 https://doi.org/10.1016/j.ebiom.2022.103879.

[131] S. Fatumo, T. Chikowore, A. Choudhury, M. Ayub, A.R. Martin, K. Kuchenbaecker, A Roadmap to Increase Diversity in Genomic Studies, Nature Medicine 28(2) (2022) 243-250 https://doi.org/10.1038/s41591-021-01672-4.

[132] O.E. Omotoso, J.O. Teibo, F.A. Atiba, T. Oladimeji, A.O. Adebesin, A.O. Babalghith, Bridging the Genomic Data Gap in Africa: Implications for Global Disease Burdens, Globalization and Health 18(1) (2022) 103 https://doi.org/10.1186/s12992-022-00898-2.

[133] B. Budowle, A. Arnette, A. Sajantila, A Cost–Benefit Analysis for Use of Large SNV Panels and High Throughput Typing for Forensic Investigative Genetic Genealogy, International Journal of Legal Medicine 137(5) (2023) 1595-1614 https://doi.org/10.1007/s00414-023-03029-7.

[134] A.B. Popejoy, S.M. Fullerton, Genomics Is Failing on Diversity, Nature 538(7624) (2016) 161-164 https://doi.org/10.1038/538161a.

[135] P.M. Sajeer, Disruptive Technology: Exploring the Ethical, Legal, Political, and Societal Implications of Nanopore Sequencing Technology: Exploring the Ethical, Legal, Political, and Societal Implications of Nanopore Sequencing Technology, EMBO Reports 24(5) (2023) e56619 https://doi.org/10.15252/embr.202256619.

[136] C. Rotimi, A. Abayomi, A. Abimiku, V.M. Adabayeri, C. Adebamowo, E. Adebiyi, A.D. Ademola, A. Adeyemo, Research Capacity. Enabling the Genomic Revolution in Africa, Science 344(6190) (2014) 1346-1348 10.1126/science.1251546.

[137] L. Daniel, SA's DNA Backlog Won't Be Cleared before 2023, at the Current Processing Rate. <https://www.news24.com/news24/bi-archive/south-africa-dna-backlog-wont-be-cleared-by-2023-2021-8>, 2021 (accessed 27 August.2024).

[138] S. Fan, M.E.B. Hansen, Y. Lo, S.A. Tishkoff, Going Global by Adapting Local: A Review of Recent Human Adaptation, Science 354(6308) (2016) 54-59 https://doi.org/doi:10.1126/science.aaf5098.

[139] S.A. Tishkoff, F.A. Reed, F.R. Friedlaender, C. Ehret, A. Ranciaro, A. Froment, J.B. Hirbo, A.A. Awomoyi, J.M. Bodo, O. Doumbo, M. Ibrahim, A.T. Juma, M.J. Kotze, G. Lema, J.H. Moore, H. Mortensen, T.B. Nyambo, S.A. Omar, K. Powell, G.S. Pretorius, M.W. Smith, M.A. Thera, C. Wambebe, J.L. Weber, S.M. Williams, The Genetic Structure and History of Africans and African Americans, Science 324(5930) (2009) 1035-1044 https://doi.org/10.1126/science.1172257.

[140] F.A. Reed, S.A. Tishkoff, African Human Diversity, Origins and Migrations, Current Opinion in Genetics & Development 16(6) (2006) 597-605 https://doi.org/10.1016/j.gde.2006.10.008.

[141] S. Ramachandran, O. Deshpande, C.C. Roseman, N.A. Rosenberg, M.W. Feldman, L.L. Cavalli-Sforza, Support from the Relationship of Genetic and Geographic Distance in Human Populations for a Serial Founder Effect Originating in Africa, Proceedings of the National Academy of Sciences 102(44) (2005) 15942-15947 https://doi.org/10.1073/pnas.0507611102.

[142] J.Z. Li, D.M. Absher, H. Tang, A.M. Southwick, A.M. Casto, S. Ramachandran, H.M. Cann, G.S. Barsh, M. Feldman, L.L. Cavalli-Sforza, R.M. Myers, Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation, Science 319(5866) (2008) 1100-1104 doi: https://doi.org/10.1126/science.1153717.

[143] A.R. Bentley, S. Callier, C.N. Rotimi, Diversity and Inclusion in Genomic Research: Why the Uneven Progress?, Journal of Community Genetics 8(4) (2017) 255-266 https://doi.org/10.1007/s12687-017-0316-6.

[144] D. Jordan, D. Mills, Past, Present, and Future of DNA Typing for Analyzing Human and Non-Human Forensic Samples, Frontiers in Ecology and Evolution 9 (2021) https://doi.org/10.3389/fevo.2021.646130.

[145] K.B. Gettings, R.A. Aponte, P.M. Vallone, J.M. Butler, STR Allele Sequence Variation: Current Knowledge and Future Issues, Forensic Science International: Genetics 18 (2015) 118-130 https://doi.org/10.1016/j.fsigen.2015.06.005.

[146] C.R. Hill, M.C. Kline, D.L. Duewer, J.M. Butler, Concordance Testing Comparing STR Multiplex Kits with a Standard Data Set, Forensic Science International: Genetics Supplement Series 3(1) (2011) e188-e189 https://doi.org/10.1016/j.fsigss.2011.08.094.

[147] K.B. Gettings, D. Ballard, M. Bodner, L.A. Borsuk, J.L. King, W. Parson, C. Phillips, Report from the Strand Working Group on the 2019 STR Sequence Nomenclature Meeting, Forensic Science International: Genetics, 2019, https://doi.org/10.1016/j.fsigen.2019.102165

[148] J. Hoogenboom, T. Sijen, K.J. van der Gaag, STRNaming: Generating Simple, Informative Names for Sequenced STR Alleles in a Standardised and Automated Manner, Forensic Science International: Genetics 52 (2021) https://doi.org/10.1016/j.fsigen.2021.102473.

[149] B. Young, T. Faris, L. Armogida, A Nomenclature for Sequence-Based Forensic DNA Analysis, Forensic Science International: Genetics 42 (2019) 14-20 https://doi.org/10.1016/j.fsigen.2019.06.001.

[150] K.B. Gettings, M. Bodner, L.A. Borsuk, J.L. King, D. Ballard, W. Parson, C.C.G. Benschop, C. Børsting, B. Budowle, J.M. Butler, K.J. van der Gaag, P. Gill, L. Gusmão, D.R. Hares, J. Hoogenboom, J. Irwin, L. Prieto, P.M. Schneider, M. Vennemann, C. Phillips, Recommendations of the DNA Commission of the International Society for Forensic Genetics (ISFG) on Short Tandem Repeat Sequence Nomenclature, Forensic Science International: Genetics 68 (2024) https://doi.org/10.1016/j.fsigen.2023.102946.

[151] S.J. Park, J.Y. Kim, Y.G. Yang, S.H. Lee, Direct STR Amplification from Whole Blood and Blood- or Saliva-Spotted Fta® without DNA Purification, Journal of Forensic Sciences 53(2) (2008) 335-341 https://doi.org/10.1111/j.1556-4029.2008.00666.x.

[152] D.Y. Wang, S. Gopinath, R.E. Lagacé, W. Norona, L.K. Hennessy, M.L. Short, J.J. Mulero, Developmental Validation of the Globalfiler® Express PCR Amplification Kit: A 6-Dye Multiplex Assay for the Direct Amplification of Reference Samples, Forensic Science International: Genetics 19 (2015) 148-155 https://doi.org/10.1016/j.fsigen.2015.07.013.

[153] Y.C. Swaran, L. Welch, A Comparison between Direct PCR and Extraction to Generate DNA Profiles from Samples Retrieved from Various Substrates, Forensic Science International: Genetics 6(3) (2012) 407-412 https://doi.org/10.1016/j.fsigen.2011.08.007.

[154] P. Thanakiatkrai, B. Rerkamnuaychoke, Direct STR Typing from Fired and Unfired Bullet Casings, Forensic Sci Int 301 (2019) 182-189 https://doi.org/10.1016/j.forsciint.2019.05.037.

[155] P. Shrivastava, T. Jain, R.K. Kumawat, Direct PCR Amplification from Saliva Sample Using Non-Direct Multiplex STR Kits for Forensic DNA Typing, Scientific Reports 11(1) (2021) 7112 https://doi.org/10.1038/s41598-021-86633-0.

[156] D.Y. Wang, C.-W. Chang, N.J. Oldroyd, L.K. Hennessy, Direct Amplification of STRs from Blood or Buccal Cell Samples, Forensic Science International: Genetics Supplement Series 2(1) (2009) 113-114 https://doi.org/10.1016/j.fsigss.2009.08.069.

[157] M. Scherer, B. Alsdorf, L. Bochmann, A. Prochnow, H. Engel, Development of the Investigator STR Go! Kits for the Direct Amplification of Reference Samples, Forensic Science International: Genetics Supplement Series 4(1) (2013) e15-e16 https://doi.org/10.1016/j.fsigss.2013.10.007.

[158] M. Date Chong, J. Wallin, A Single Direct Amplification Method for Forensic Casework References on a Variety of Substrates, Forensic Science International: Reports 5 (2022) 100260 https://doi.org/10.1016/j.fsir.2022.100260.

[159] P. Brito, V. Lopes, V. Bogas, F. Balsa, L. Andrade, A. Serra, M.S. Bento, A.M. Bento, P. Cunha, M. Carvalho, F. Corte-Real, M.J. Anjos, Amplification of Non-Fta Samples with AmpFLSTR® Identifiler® Direct PCR Amplification Kit, Forensic Science International: Genetics Supplement Series 3(1) (2011) e371-e372 https://doi.org/10.1016/j.fsigss.2011.09.047.

[160] K.M. Reid, L.J. Heathfield, Evaluation of Direct PCR for Routine DNA Profiling of Non-Decomposed Deceased Individuals, Science and Justice 60(6) (2020) 567-572 10.1016/j.scijus.2020.08.004.

[161] J. Hedman, A. Nordgaard, C. Dufva, B. Rasmusson, R. Ansell, P. Rådström, Synergy between DNA Polymerases Increases Polymerase Chain Reaction Inhibitor Tolerance in Forensic DNA Analysis, Analytical Biochemistry 405(2) (2010) 192-200 https://doi.org/10.1016/j.ab.2010.06.028.

[162] M. Sidstedt, P. Rådström, J. Hedman, PCR Inhibition in qPCR, dPCR and MPS-Mechanisms and Solutions, Analytical and Bioanalytical Chemistry 412(9) (2020) 2009-2023 https://doi.org/10.1007/s00216-020-02490-2.

[163] K.M. Reid, L.J. Heathfield, Allele Frequency Data for 23 Y-Chromosome Short Tandem Repeats (STRs) for the South African Population, Forensic Sci Int Genet 46 (2020) 102270 https://doi.org/10.1016/j.fsigen.2020.102270.

[164] L.J. Heathfield, L. Nel, K.M. Reid, Evaluation of the Investigator® 24plex GO! Kit and Associated Allele Frequency Data for Four South African Population Groups, Forensic Science International: Reports 9 (2024) 100357 https://doi.org/10.1016/j.fsir.2024.100357.

[165] Promega, Swabsolutiontm Kit Technical Manual, 2016.

[166] QIAGEN, Investigator® 24plex Go! Handbook, 2021.

[167] Applied Biosystems, Quantifiler™ HP and Trio DNA Quantification Kits User Guide, 2018.

[168] Omega Bio-tek, Mag-Bind® Blood & Tissue DNA HDQ 96 DNA Extraction Kit, 2024.

[169] QIAGEN, Qiaamp® DNA Investigator Handbook, 2020.

[170] QIAGEN, Purification of Total DNA from Crude Lysates Using the Dneasy® Blood & Tissue Kit, 2023.

[171] A. Technologies, High Sensitivity D1000 Screentape Assay Quick Guide for Tapestation Systems, 2021.

[172] Invitrogen, Qubit™ 1x Dsdna Hs Assay Kit User Guide, 2020.

[173] P. Virtanen, R. Gommers, T.E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S.J. van der Walt, M. Brett, J. Wilson, K.J. Millman, N. Mayorov, A.R.J. Nelson, E. Jones, R. Kern, E. Larson, C.J. Carey, İ. Polat, Y. Feng, E.W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E.A. Quintero, C.R. Harris, A.M. Archibald, A.H. Ribeiro, F. Pedregosa, P. van Mulbregt, Scipy 1.0: Fundamental Algorithms for Scientific Computing in Python, Nature Methods 17(3) (2020) 261-272 https://doi.org/10.1038/s41592-019-0686-2.

[174] R.S. McLaren, J. Bourdeau-Heller, J. Patel, J.M. Thompson, J. Pagram, T. Loake, D. Beesley, M. Pirttimaa, C.R. Hill, D.L. Duewer, M.C. Kline, J.M. Butler, D.R. Storts, Developmental Validation of the PowerPlex® ESI 16/17 Fast and Powerplex® ESX 16/17 Fast Systems, Forensic Science International: Genetics 13 (2014) 195-205 https://doi.org/10.1016/j.fsigen.2014.08.004.

[175] K. Oostdik, J. French, D. Yet, B. Smalling, C. Nolde, P.M. Vallone, E.L.R. Butts, C.R. Hill, M.C. Kline, T. Rinta, A.M. Gerow, S.R. Allen, C.K. Huber, J. Teske, B. Krenke, M. Ensenberger, P. Fulmer, C. Sprecher, Developmental Validation of the Powerplex® 18d System, a Rapid STR Multiplex for Analysis of Reference Samples, Forensic Science International: Genetics 7(1) (2013) 129-135 https://doi.org/10.1016/j.fsigen.2012.07.008.

[176] J. Lal, L. McIntosh, D. Qin, Validation of Buccal Swab DNA Isolation Using Promega SwabSolution and Powerplex 16HS System for Full-Reaction and Half-Reaction Volumes for PCR Amplification, American Journal of Clinical Pathology 142(suppl_1) (2014) A177-A177 https://doi.org/10.1093/ajcp/142.suppl1.177.

[177] A.L. Whittaker, L.J. Heathfield, The First X-STR Population Study for the South African Population, Forensic Science International: Reports 9 (2024) https://doi.org/10.1016/j.fsir.2024.100359.

[178] A.P. Revoir, H. Lancaster, C. Ames, Initial Assessment of NGS as a Tool to Augment Routine CE Analysis of STRs, Forensic Science International: Genetics Supplement Series 7(1) (2019) 747-749 https://doi.org/10.1016/j.fsigss.2019.10.162.

[179] L.J. Heathfield, A.N. Hitewa, A. Gibbon, C.G. Mole, The Effect of Nucleospin® Forensic Filters on DNA Recovery from Trace DNA Swabs, Science & Justice 62(3) (2022) 284-287 https://doi.org/10.1016/j.scijus.2022.03.001.

[180] C.M. Cupples, J.R. Champagne, K.E. Lewis, T.D. Cruz, STR Profiles from DNA Samples with “Undetected” or Low Quantifiler™ Results, Journal of Forensic Sciences 54(1) (2009) 103-107 https://doi.org/10.1111/j.1556-4029.2008.00914.x.

[181] D.N.O. Bonsu, D. Higgins, C. Simon, C.S. Goodwin, J.M. Henry, J.J. Austin, Quantitative PCR Overestimation of DNA in Samples Contaminated with Tin, Journal of Forensic Sciences 68(4) (2023) 1302-1309 https://doi.org/10.1111/1556-4029.15312.

[182] QIAGEN, Developmental Validation of the Qiaamp® DNA Investigator® Kit, 2019.

[183] QIAGEN, QiaAmp® DNA FFPE Advanced Handbook, 2020.

[184] C. McDonald, D. Taylor, A. Linacre, PCR in Forensic Science: A Critical Review, Genes (Basel) 15(4) (2024) https://doi.org/10.3390/genes15040438.

[185] Statistics South Africa, Statistical Release P0302: Mid-Year Population Estimates, 2022. <https://census.statssa.gov.za/#/>, 2022 (accessed 30 June.2024).

[186] A. Lucassen, K. Ehlers, P.J. Grobler, A.L. Shezi, Allele Frequency Data of 15 Autosomal STR Loci in Four Major Population Groups of South Africa, International Journal of Legal Medicine 128(2) (2014) 275-276 https://doi.org/10.1007/s00414-013-0898-4.

[187] P.G. Ristow, K.W. Cloete, M.E. D'Amato, Globalfiler® Express DNA Amplification Kit in South Africa: Extracting the Past from the Present, Forensic Science International: Genetics 24 (2016) 194-201 https://doi.org/10.1016/j.fsigen.2016.07.007.

[188] P. Theodore, I.-M. Wendy, The Multiple Meanings of Coloured Identity in South Africa, Africa Insight 42(1) (2012) 87-102 https://doi.org/10.10520/EJC125075.

[189] Á. Carracedo, J.M. Butler, L. Gusmão, A. Linacre, W. Parson, L. Roewer, P.M. Schneider, New Guidelines for the Publication of Genetic Population Data, Forensic Science International: Genetics 7(2) (2013) 217-220 https://doi.org/10.1016/j.fsigen.2013.01.001.

[190] QIAGEN, Qiaamp DNA Investigator Handbook, 2012.

[191] S.A. Miller, D.D. Dykes, H.F. Polesky, A Simple Salting out Procedure for Extracting DNA from Human Nucleated Cells, Nucleic Acids Research 16(3) (1988) 1215-1215 https://doi.org/10.1093/nar/16.3.1215.

[192] Thermo Fisher Scientific. Qubit™ dsDNA HS Assay Kits, (2015).

[193] K.B. Gettings, L.A. Borsuk, D. Ballard, M. Bodner, B. Budowle, L. Devesse, J. King, W. Parson, C. Phillips, P.M. Vallone, STRseq: A Catalog of Sequence Diversity at Human Identification Short Tandem Repeat Loci, Forensic Science International: Genetics 31 (2017) 111-117 https://doi.org/10.1016/j.fsigen.2017.08.017.

[194] K.B. Gettings, L.A. Borsuk, P.M. Vallone, Performing a Blast Search of the STRseq Bioproject, Forensic Science International: Genetics Supplement Series 6 (2017) e372-e374 https://doi.org/10.1016/j.fsigss.2017.09.173.

[195] L. Excoffier, H.E. Lischer, Arlequin Suite Ver 3.5: A New Series of Programs to Perform Population Genetics Analyses under Linux and Windows, Molcular Ecology Resources 10(3) (2010) 564-567 https://doi.org/10.1111/j.1755-0998.2010.02847.x.

[196] A. Gouy, M. Zieger, STRAF—a Convenient Online Tool for STR Data Evaluation in Forensic Genetics, Forensic Science International: Genetics 30 (2017) 148-151 https://doi.org/10.1016/j.fsigen.2017.07.007.

[197] P. Hölzl-Müller, M. Bodner, B. Berger, W. Parson, Exploring STR Sequencing for Forensic DNA Intelligence Databasing Using the Austrian National DNA Database as an Example, International Journal of Legal Medicine 135(6) (2021) 2235-2246 https://doi.org/10.1007/s00414-021-02685-x.

[198] R.S. Just, L.I. Moreno, J.B. Smerick, J.A. Irwin, Performance and Concordance of the Forenseq™ System for Autosomal and Y Chromosome Short Tandem Repeat Sequencing of Reference-Type Specimens, Forensic Science International Genetics 28 (2017) 1-9 https://doi.org/10.1016/j.fsigen.2017.01.001.

[199] B.S. Bruce Budowle, Stephen Niezgoda, Ranajit Chakraborty, Codis STR Loci Data from 41 Sample Populations, Journal of Forensic Sciences 3(43) (2001) 37.

[200] S. Riman, M. Ghemrawi, L.A. Borsuk, R. Mahfouz, S. Walsh, P.M. Vallone, Sequence-Based Allelic Variations and Frequencies for 22 Autosomal STR Loci in the Lebanese Population, Forensic Science International: Genetics 65 (2023) https://doi.org/10.1016/j.fsigen.2023.102872.

[201] M. Magdalena, M. Wróbel, A. Parys-Proszek, T. Kupiec, Evaluation of the Performance of the Beta Version of the Forenseq™ DNA Signature Prep Kit on the Miseq Fgx™ Forensic Genomics System, Forensic Science International: Genetics Supplement Series 7(1) (2019) 585-586 https://doi.org/10.1016/j.fsigss.2019.10.099.

[202] F. Guo, J. Yu, L. Zhang, J. Li, Massively Parallel Sequencing of Forensic STRs and SNVs Using the Illumina(®) Forenseq™ DNA Signature Prep Kit on the Miseq Fgx™

Forensic Genomics System, Forensic Sci Int Genet 31 (2017) 135-148 10.1016/j.fsigen.2017.09.003.

[203] J.M. Butler, Data Interpretation Overview, in: J.M. Butler (Ed.), Advanced Topics in Forensic DNA Typing: Interpretation, Academic Press, San Diego, 2015, https://doi.org/10.1016/B978-0-12-405213-0.00001-4, pp. 3-24.

[204] M.M. Foley, G. Koehler, J. Fu, R. Allen, J.R. Wagner, An Exploratory View into Allelic Drop-out of Sequenced Autosomal STRs, Journal of Forensic Sciences 69(3) (2024) 825-835 https://doi.org/10.1111/1556-4029.15504.

[205] Scientific Working Group on DNA Analysis Methods, Addendum to "SWGDAM Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories" to Address Next Generation Sequencing, SWGDAM, 2019.

[206] R. Viljoen, K.M. Reid, C.G. Mole, M. Rangwaga, L.J. Heathfield, Towards Molecular Autopsies: Development of a FFPE Tissue DNA Extraction Workflow, Science & Justice 62(2) (2022) 137-144 https://doi.org/10.1016/j.scijus.2021.12.005.

[207] QIAGEN, Investigator® Argus X-12 Qs Handbook, 2022.

[208] Promega, Powerplex® Y23 System for Use on the Applied Biosystems® Genetic Analyzers Technical Manual, 2023.

[209] S. Vernarecci, E. Ottaviani, A. Agostino, E. Mei, L. Calandro, P. Montagna, Quantifiler® Trio Kit and Forensic Samples Management: A Matter of Degradation, Forensic Science International: Genetics 16 (2015) 77-85 https://doi.org/10.1016/j.fsigen.2014.12.005.

[210] S. Walsh, L. Chaitanya, L. Clarisse, L. Wirken, J. Draus-Barini, L. Kovatsi, H. Maeda, T. Ishikawa, T. Sijen, P. de Knijff, W. Branicki, F. Liu, M. Kayser, Developmental Validation of the Hirisplex System: DNA-Based Eye and Hair Colour Prediction for Forensic and Anthropological Usage, Forensic Science International: Genetics 9 (2014) 150-161 https://doi.org/10.1016/j.fsigen.2013.12.006.

[211] S. Walsh, L. Chaitanya, K. Breslin, C. Muralidharan, A. Bronikowska, E. Pospiech, J. Koller, L. Kovatsi, A. Wollstein, W. Branicki, F. Liu, M. Kayser, Global Skin Colour

Prediction from DNA, Human Genetics 136(7) (2017) 847-863 https://doi.org/10.1007/s00439-017-1808-5.

[212] L. Chaitanya, K. Breslin, S. Zuñiga, L. Wirken, E. Pośpiech, M. Kukla-Bartoszek, T. Sijen, P. Knijff, F. Liu, W. Branicki, M. Kayser, S. Walsh, The Hirisplex-S System for Eye, Hair and Skin Colour Prediction from DNA: Introduction and Forensic Developmental Validation, Forensic Science International: Genetics 35 (2018) 123-135 https://doi.org/10.1016/j.fsigen.2018.04.004.

[213] P. Fattorini, C. Previderé, I. Carboni, G. Marrubini, S. Sorçaburu-Cigliero, P. Grignani, B. Bertoglio, P. Vatta, U. Ricci, Performance of the Forenseq™ DNA Signature Prep Kit on Highly Degraded Samples, Electrophoresis 38(8) (2017) 1163-1174 https://doi.org/10.1002/elps.201600290.

[214] K. Elwick, X. Zeng, J. King, B. Budowle, S. Hughes-Stamm, Comparative Tolerance of Two Massively Parallel Sequencing Systems to Common PCR Inhibitors, International Journal of Legal Medicine 132(4) (2018) 983-995 https://doi.org/10.1007/s00414-017-1693-4.

[215] M. Sidstedt, C.R. Steffen, K.M. Kiesler, P.M. Vallone, P. Rådström, J. Hedman, The Impact of Common PCR Inhibitors on Forensic MPS Analysis, Forensic Science International-Genetics 40 (2019) 182-191 https://doi.org/10.1016/j.fsigen.2019.03.001.

[216] C. Schrader, A. Schielke, L. Ellerbroek, R. Johne, PCR Inhibitors – Occurrence, Properties and Removal, Journal of Applied Microbiology 113(5) (2012) 1014-1026 https://doi.org/10.1111/j.1365-2672.2012.05384.x.

[217] E. Jue, D. Witters, R.F. Ismagilov, Two-Phase Wash to Solve the Ubiquitous Contaminant-Carryover Problem in Commercial Nucleic-Acid Extraction Kits, Scientific Reports 10(1) (2020) 1940 https://doi.org/10.1038/s41598-020-58586-3.

[219] J.M. Butler, Debunking Some Urban Legends Surrounding Validation within the Forensic DNA Community, (2006).

[220] J.C. Cihlar, C. Amory, R. Lagacé, C. Roth, W. Parson, B. Budowle, Developmental Validation of a MPS Workflow with a PCR-Based Short Amplicon Whole Mitochondrial Genome Panel, Genes (Basel) 11(11) (2020) https://doi.org/10.3390/genes11111345.

[221] M. Kayser, W. Branicki, W. Parson, C. Phillips, Recent Advances in Forensic DNA Phenotyping of Appearance, Ancestry and Age, Forensic Science International: Genetics 65 (2023) https://doi.org/10.1016/j.fsigen.2023.102870.

[222] South Africa, Act No 37 of 2013: Criminal Law (Forensic Procedure) Amendment Act of 2013, in: South African Government Gazette (Ed.) 2013.

[223] N. Slabbert, L.J. Heathfield, Ethical, Legal and Social Implications of Forensic Molecular Phenotyping in South Africa, Developing World Bioethics 18(2) (2018) 171-181 https://doi.org/10.1111/dewb.12194.

[224] P. Hunter, Cold Cases and Ancient Trade Routes: DNA Phenotyping and Isotope Analysis Extend Forensic Science into New Domains, EMBO Reports 22(12) (2021) e54188 https://doi.org/10.15252/embr.202154188.

[225] A.C. Nwawuba SU, Awareness Level on the Role of Forensic DNA Database in Criminal Investigation in Nigeria: A Case Study of Benin City., Journal of Forensic Science and Research 4 (2020) 007-014 https://doi.org/10.29328/journal.jfsr.1001019.

[226] D.T. Masiloane, An Enemy from Within : A Critical Analysis of Corruption in the South African Police Service, South African Journal of Criminal Justice 20(1) (2007) 46-59 https://doi.org/doi:10.10520/EJC52902.

[227] B. Budowle, A. Arnette, A. Sajantila, A Cost-Benefit Analysis for Use of Large SNV Panels and High Throughput Typing for Forensic Investigative Genetic Genealogy, International Journal of Legal Medicine 137(5) (2023) 1595-1614 https://doi.org/10.1007/s00414-023-03029-7.

[228] A. Ambers, M.M. Bus, J.L. King, B. Jones, J. Durst, J.E. Bruseth, H. Gill-King, B. Budowle, Forensic Genetic Investigation of Human Skeletal Remains Recovered from the La Belle Shipwreck, Forensic Science International 306 (2020) https://doi.org/10.1016/j.forsciint.2019.110050.

[230] G. Kulstein, T. Hadrys, P. Wiegand, As Solid as a Rock—Comparison of Ce- and Mps-Based Analyses of the Petrosal Bone as a Source of DNA for Forensic Identification of Challenging Cranial Bones, International Journal of Legal Medicine 132(1) (2018) 13-24 https://doi.org/10.1007/s00414-017-1653-z.

Appendices

Appendix 1.1: MPS profile success rates (in percentage) for post-mortem sample types processed with the ForenSeq™ DNA Signature Prep kit

Table A1.1: Meta data from a systematic search of the literature pertaining to post-mortem sample types and their call rates (%) with the ForenSeq™ DNA Signature Prep kit (FSP)

Source	Sample type	DNA profile success rate with the FSP (%)			Reference
		STRs (n=59)	SNVs (173)		
Decomposed skeleton in marine environment	Bone	93.22	Not reported		[81]
	Bone	100.00			
	Nail	89.83			
	Soft tissue	100.00			
	Soft tissue	98.31			
Mass grave	Tooth	88.14	93.64		[82]
	Tooth	84.75	63.01		
	Tooth	5.08	4.62		
	Tooth	59.32	63.01		
	Tooth	50.85	16.18		
	Tooth	0.00	0.00		
	Tooth	91.53	75.14		
	Tooth	25.42	18.50		
	Tooth	64.41	35.84		
	Tooth	0.00	1.16		
Blood FTA	Living	90-100	Not reported		[83]
5th-13th century skeletal remains	Molar	77.97	87.23 (94 iiSNVs)	Not reported	[84]
	Molar	38.98	70.21 (94 iiSNVs)		
	Molar	74.58	90.42 (94 iiSNVs)		
	Femur	76.27	86.17 (94 iiSNVs)		
	Molar	52.54	65.96 (94 iiSNVs)		
	Molar	67.80	85.11 (94 iiSNVs)		
Not reported	Bone	100	98.95 (94 iiSNVs)	Not reported	[85]
	Bone	88.9	96.84 (94 iiSNVs)		
	Bone	98.4	100 (94 iiSNVs)		
	Bone	17.5	26.32 (94 iiSNVs)		
	Bone	77.8	93.68 (94 iiSNVs)		
Full skeletons recovered from shipwreck	Bone	71.21	100		[228]
	Bone	30.22	79.84		
5th-8th century hard tissues	Bone	0	1.66 (94 iiSNVs)	Not reported	[61]
	Bone	79.71	93.46 (94 iiSNVs)		
	Tooth	57.84	91.99 (94 iiSNVs)		
	Tooth	94.24	97.27 (94 iiSNVs)		
Exhumed remains	Femur	69.50	95.69		[22]
Cranial bones	Bone	91.05	98.57		[230]
	Bone	76.27	95.83		
	Bone	79.66	81.50		
	Bone	89.83	99.63		
	Bone	98.31	99.63		
	Bone	72.88	65.57		
	Bone	100	98.67		
	Bone	83.9	79.77		

Appendix 1.2: Human research ethics approval from the HREC UCT

UNIVERSITY OF CAPE TOWN
Faculty of Health Sciences
Human Research Ethics Committee

Room G50- Old Main Building
Groote Schuur Hospital
Observatory 7925
Telephone [021] 406 6492
Email: hrec-submissions uct.ac.za
Website: www.health.uct.ac.za fhs research humanethics forms

08 July 2021

HREC REF: 400/2021

Dr L Heathfield
Division of Forensic Medicine& Toxicology
Falmouth Building-FHS
Email: Laura.heathfield uct ac za
Student: MRTDON003 m uct.ac.za

Dear Dr Heathfield

PROJECT TITLE: INVESTIGATION TO IMPLEMENT A MASSIVELY PARALLEL SEQUENCING WORKFLOW IN FORENSIC HUMAN IDENTIFICATION IN SOUTH AFRICA. (SUB-STUDY – 342/2016) (PHD DEGREE – MISS DONNA-LEE MARTIN)

Thank you for submitting your study to the Faculty of Health Sciences Human Research Ethics Committee for review.

It is a pleasure to inform you that the HREC has **formally approved** the above-mentioned study.

This approval is subject to strict adherence to the HREC recommendations regarding research involving human participants during COVID -19, dated 17 March 2020 & 06 July 2020.

Approval is granted for one year until the 30 July 2022.

Please submit a progress form, using the standardised Annual Report Form if the study continues beyond the approval period. Please submit a Standard Closure form if the study is completed within the approval period.
(Forms can be found on our website: www.health.uct.ac.za fhs research humanethics forms)

The HREC acknowledge that the student: Miss Donna-Lee Martin will also be involved in this study.

Please quote the HREC REF 400/2021 in all your correspondence.

Please note that the ongoing ethical conduct of the study remains the responsibility of the principal investigator.

Please note that for all studies approved by the HREC, the principal investigator **must** obtain appropriate institutional approval, where necessary, before the research may occur

Yours sincerely

PROFESSOR M BLOCKMAN
CHAIRPERSON, FACULTY OF HEALTH SCIENCES HUMAN RESEARCH ETHICS COMMITTEE
Federal Wide Assurance Number: FWA00001637.
Institutional Review Board (IRB) number: IRB00001938
NHREC-registration number: REC-210208-007

This serves to confirm that the University of Cape Town Human Research Ethics Committee complies to the Ethics Standards for Clinical Research with a new drug in patients, based on the Medical Research Council (MRC-SA), Food and Drug Administration (FDA-USA), International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use: Good Clinical Practice (ICH GCP), South African Good Clinical Practice Guidelines (DoH 2020), based on the Association of the British Pharmaceutical Industry Guidelines (ABPI), and Declaration of Helsinki (2013) guidelines. The Human Research Ethics Committee granting this approval is in compliance with the ICH Harmonised Tripartite Guidelines E6: Note for Guidance on Good Clinical Practice (CPMP/ICH/135/95) and FDA Code Federal Regulation Part 50, 56 and 312.

HREC/REF400/2021sa

UNIVERSITY OF CAPE TOWN

FACULTY OF HEALTH SCIENCES
Human Research Ethics Committee

FHS016: Annual Progress Report / Renewal

HREC office use only (FWA00001637; IRB00001938) This serves as notification of annual approval, including any documentation described below.			
☑ Approved	Annual progress report	Approved until/next renewal date	30/06/25
☐ Not approved	See attached comments		
Signature Chairperson of the HREC/ Designee		Date Signed	28/5/24

Note: Please email this form and supporting documents (if applicable) in a combined pdf-file to hrec-enquiries@uct.ac.za.

Please use the latest form found on our website:
http://www.health.uct.ac.za/fhs/research/humanethics/forms

Comments to PI from the HREC

Principal Investigator to complete the following:

1. Protocol information

Date (when submitting this form)	May 2023		
HREC REF Number	HREC:400/2021	Current Ethics Approval was granted until	30 June 2024
Protocol title	Investigation to implement a massively parallel sequencing workflow for forensic human identification in South Africa		
Protocol number (if applicable)			
Are there any sub-studies linked to this study?		☐ Yes	☑ No
If yes, could you please provide the HREC Reference number for all sub-studies? **Note:** A separate FHS016 must be submitted for each sub-study.			

HUMAN RESEARCH ETHICS COMMITTEE
FHS016
28 MAY 2024
HEALTH SCIENCES FACULTY
UNIVERSITY OF CAPE TOWN

Appendix 1.3: Animal ethics approval from AEC UCT

UNIVERSITY OF CAPE TOWN
Faculty of Health Sciences
Animal Ethics Committee

Room G50 Old Main Building
Groote Schuur Hospital
Observatory 7925
Website: www.health.uct.ac.za/fhs/research/animalethics/forms

26 July 2021

Dr Laura Heathfield
Division of Forensic Medicine and
Toxicology
Department of Pathology
Faculty of Health Sciences
University of Cape Town

Dear Dr Heathfield

PROTOCOL TITLE: Investigation to implement a massively parallel sequencing workflow in forensic human identification in South Africa.
FHS AEC REF NO: 021_010

Thank you for submitting your request for authorisation of use of animal material for scientific purposes to the Faculty of Health Sciences (FHS) Animal Ethics Committee (AEC).

I am pleased to inform you that the FHS AEC has **authorised** your protocol, which will terminate on **31 July 2024.**

Number of animals & species: 10 µl of DNA from 8 different species (2 animals per species) namely Cat (*Felis catus*), dog (*Canis lupus familiaris*), vervet monkey (*Chlorocebus pygerythrus*), macaque (*Macaca*), cow (*Bos taurus*), sheep (*Ovis*), domesticated chicken (*Gallus gallus domesticus*) and mouse (*Mus musculus*).

Please quote the FHS AEC REF NO (above) in all future correspondence.

Please note that the authorisation of this protocol imposes the following obligations on the principal investigator (PI):

1. To submit an annual mandatory progress report. The first annual report for this protocol is due on **28 February 2022**. The forms can be accessed from http://www.health.uct.ac.za/fhs/research/animalethics/forms
2. To submit a final mandatory report on the **31 July 2024**, please access the final report form from: http://www.health.uct.ac.za/fhs/research/animalethics/forms
3. Ensuring that all study participants perform within the confines of the procedures and experimental design of the protocol as authorised, or as amended.

4. Ensuring that all study participants comply with all applicable national legislation, UCT policies, FHS AEC policies and standard operating procedures (SOPs) and national standards (SANS 10386: 2008).
5. Ensuring compliance with DAFF Section 20 requirements.
6. Ensuring that you as the PI immediately alert the FHS AEC to any event involving the welfare of the animals which has occurred during the course of the study, as well as the actions that were taken to respond to these events.
7. Ensuring that you as the PI alert the FHS AEC to any new or unexpected ethical issues that arose during the course of the study, and how these issues were addressed.
8. Ensuring that all study participants are registered with or have been authorised by the South African Veterinary Council (SAVC) to perform the procedures on animals or will be performing the procedures under the direct and continuous supervision of SAVC-registered veterinary professionals or SAVC-registered para-veterinary professionals.
9. If the PI or any study participant is in any way uncertain how to respond to any of these obligations or deal with any of the issues referred to above, they must consult with FHS AEC.
10. All animals found dead must be reported to the RAF on the appropriate form: http://www.health.uct.ac.za/fhs/research/animalethics/forms
11. All animals found in distress must be reported to the RAF on the appropriate form.

My best wishes for a successful research and /or teaching endeavour.

Yours sincerely

PROF G. LOUW
CHAIR, FHS AEC

Appendix 1.4: Outputs related to this thesis

Chapter in book:

This book chapter stems from the research done by the PhD candidate and supervisor on quality management and internal validation

L.J. Heathfield and **D.P Martin**, Quality Assurance Processes in Forensic Science, Forensic Evidence Processing in Gender-Based Violence Cases: Handbook for Criminal Justice Practitioners, United Nations Office on Drugs and Crime, Austria, 2024

http://www.unodc.org/rosaf/uploads/documents/Publication/13.5.2024_Forensic_Evidence_Processing_in_Gender-Based_Violence_Cases.pdf

Appendix 1.5: Data Management Plan

Investigation to implement a massively parallel sequencing workflow for forensic human identification in South Africa - Student Outline DMP

1. General guidelines

I understand the Outline DMP template is a projection of my anticipated data management planning requirements and should be updated as my project develops.

2. Authors and supervisors

Investigation to implement a massively parallel sequencing workflow for forensic human identification in South Africa.

Donna-Lee Martin
MRTDON003

Dr Laura Heathfield

3. Data Collection/Generation

I intend to produce original quantitative data for frequencies of genetic variations within the South African population. The data is primarily quantitative. The sample size is approximately 500, and the size of the output files is unknown at this time. The format of this data will be tabular and will be saved as xlsx. or CSV files.

I do intend on using existing data that formed part of my Masters project. The research is ongoing and will be continued in this project. The Masters project and this PhD project are both linked to an umbrella study for which the research is ongoing.

Anonymous population data consisting of allele frequencies will be published in manuscripts in scientific journals and shared with a quality control platform (STRidER: https://STRidER.online/). Statistics from internal validation and optimization experiments will also be published in the form of averaged statistical values in a manuscript or report for a scientific journal.

4. Data Storage

20GB or less
The anticipated dataset size is less than 20GB. Additional storage facilities will not be required.

The data will be saved to an external hard drive securely housed in an access-controlled laboratory and will be backed up monthly. Every three months, the data will be backed up on the students UCT OneDrive.

5. Data Centre(s)/Repositories

All specimens and data will be archived according to standard operating procedures within the Division of Forensic Medicine & Toxicology.

Metadata will be included with certain input files on MS Excel and additional text files explaining how input files were formatted. All statistical analysis will be accompanied by raw data files with steps explaining how statistical analysis was performed on the dataset.

6. Budget

I do not anticipate any large data costs as my data is less than 20GB, and I will be using a 1TB hard drive and the UCT OneDrive to store and manage my data.

Appendix 2.1: PRISMA Checklist

Section and Topic	Item #	Checklist item	Location where item is reported
TITLE			
Title	1	Identify the report as a systematic review.	Chapter 2
ABSTRACT			
Abstract	2	See the PRISMA 2020 for Abstracts checklist.	NA
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of existing knowledge.	2.1
Objectives	4	Provide an explicit statement of the objective(s) or question(s) the review addresses.	2.1
METHODS			
Eligibility criteria	5	Specify the inclusion and exclusion criteria for the review and how studies were grouped for the syntheses.	2.2.2
Information sources	6	Specify all databases, registers, websites, organisations, reference lists and other sources searched or consulted to identify studies. Specify the date when each source was last searched or consulted.	2.2.1
Search strategy	7	Present the full search strategies for all databases, registers and websites, including any filters and limits used.	2.2.1
Selection process	8	Specify the methods used to decide whether a study met the inclusion criteria of the review, including how many reviewers screened each record and each report retrieved, whether they worked independently, and if applicable, details of automation tools used in the process.	NA
Data collection process	9	Specify the methods used to collect data from reports, including how many reviewers collected data from each report, whether they worked independently, any processes for obtaining or confirming data from study investigators, and if applicable, details of automation tools used in the process.	2.2.3
Data items	10a	List and define all outcomes for which data were sought. Specify whether all results that were compatible with each outcome domain in each study were sought (e.g. for all measures, time points, analyses), and if not, the methods used to decide which results to collect.	Table 2.1
	10b	List and define all other variables for which data were sought (e.g. participant and intervention characteristics, funding sources). Describe any assumptions made about any missing or unclear information.	Table 2.1
Study risk of bias assessment	11	Specify the methods used to assess risk of bias in the included studies, including details of the tool(s) used, how many reviewers assessed each study and whether they worked independently, and if applicable, details of automation tools used in the process.	NA
Effect measures	12	Specify for each outcome the effect measure(s) (e.g. risk ratio, mean difference) used in the synthesis or presentation of results.	Figure 2.6
Synthesis methods	13a	Describe the processes used to decide which studies were eligible for each synthesis (e.g. tabulating the study intervention characteristics and comparing against the planned groups for each synthesis (item #5)).	2.2.4
	13b	Describe any methods required to prepare the data for presentation or synthesis, such as handling of missing summary statistics, or data conversions.	2.2.4
	13c	Describe any methods used to tabulate or visually display results of individual studies and syntheses.	2.2.5
	13d	Describe any methods used to synthesize results and provide a rationale for the choice(s). If meta-analysis was performed, describe the model(s), method(s) to identify the presence and extent of statistical heterogeneity, and software package(s) used.	
	13e	Describe any methods used to explore possible causes of heterogeneity among study results (e.g. subgroup analysis, meta-regression).	2.2.4
	13f	Describe any sensitivity analyses conducted to assess robustness of the synthesized results.	2.2.5
Reporting bias assessment	14	Describe any methods used to assess risk of bias due to missing results in a synthesis (arising from reporting biases).	NA
Certainty assessment	15	Describe any methods used to assess certainty (or confidence) in the body of evidence for an outcome.	NA
RESULTS			

Section and Topic	Item #	Checklist item	Location where item is reported
Study selection	16a	Describe the results of the search and selection process, from the number of records identified in the search to the number of studies included in the review, ideally using a flow diagram.	2.3.1
	16b	Cite studies that might appear to meet the inclusion criteria, but which were excluded, and explain why they were excluded.	Appendix 2.5
Study characteristics	17	Cite each included study and present its characteristics.	Table 2.2
Risk of bias in studies	18	Present assessments of risk of bias for each included study.	NA
Results of individual studies	19	For all outcomes, present, for each study: (a) summary statistics for each group (where appropriate) and (b) an effect estimate and its precision (e.g. confidence/credible interval), ideally using structured tables or plots.	Figure 2.6
Results of syntheses	20a	For each synthesis, briefly summarise the characteristics and risk of bias among contributing studies.	NA
	20b	Present results of all statistical syntheses conducted. If meta-analysis was done, present for each the summary estimate and its precision (e.g. confidence/credible interval) and measures of statistical heterogeneity. If comparing groups, describe the direction of the effect.	Figure 2.6
	20c	Present results of all investigations of possible causes of heterogeneity among study results.	2.3.7.1
	20d	Present results of all sensitivity analyses conducted to assess the robustness of the synthesized results.	2.3.7.2
Reporting biases	21	Present assessments of risk of bias due to missing results (arising from reporting biases) for each synthesis assessed.	NA
Certainty of evidence	22	Present assessments of certainty (or confidence) in the body of evidence for each outcome assessed.	NA
DISCUSSION			
Discussion	23a	Provide a general interpretation of the results in the context of other evidence.	2.4
	23b	Discuss any limitations of the evidence included in the review.	2.4 and 6.3
	23c	Discuss any limitations of the review processes used.	2.4 and 6.3
	23d	Discuss implications of the results for practice, policy, and future research.	2.4 and 6.3
OTHER INFORMATION			
Registration and protocol	24a	Provide registration information for the review, including register name and registration number, or state that the review was not registered.	NA
	24b	Indicate where the review protocol can be accessed, or state that a protocol was not prepared.	NA
	24c	Describe and explain any amendments to information provided at registration or in the protocol.	NA
Support	25	Describe sources of financial or non-financial support for the review, and the role of the funders or sponsors in the review.	Pg 11
Competing interests	26	Declare any competing interests of review authors.	NA
Availability of data, code and other materials	27	Report which of the following are publicly available and where they can be found: template data collection forms; data extracted from included studies; data used for all analyses; analytic code; any other materials used in the review.	Appendices for Chapter 2

Appendix 2.2: RStudio script used to conduct meta-analysis

Instructions are preceded by "##"

```
##Data for length-based and sequence-based allele counts as well as sample size were imported
## 'data' is used here as a generic name for the dataset used
## proportions for length-based alleles were calculated and a variable created – 'prop_before'
## proportions for sequence-based alleles were calculated and a variable created – 'prop_after'

data$prop_before <- data$LB / data$`Sample size`
data$prop_after <- data$SB / data$`Sample size`


##Effect size was calculated as the ratio between 'length-based' and 'sequence-based'
proportions calculated above
data$EffectSize <- data$prop_after / data$prop_before

##The log transformed values of the effect sizes were calculated
data$logEffectSize <- log(data$EffectSize)

##The standard error for the log effect sizes were calculated
data$se_logEffectSize <- sqrt((1/data$LB) + (1/data$SB))

##The 'meta' package was installed and loaded to perform the meta-analysis
install.packages("meta")
library(meta)

##The 'metagen' function was used to perform a meta-analysis of the log effect sizes

meta_analysis <- metagen(TE = logEffectSize, seTE = se_logEffectSize, data = data, sm =
"logEffect size",
                         studlab= data$Population, comb.fixed = FALSE)
```

TE = logEffectSize: This specifies the treatment effect, which in this case is the log-transformed effect size. It's the measure of effect that you are interested in combining across studies.

##seTE = se_logEffectSize: This specifies the standard error of the log-transformed effect size. It's the measure of the variability or precision of the effect size estimate.

##data = data: This indicates the dataset being used for the analysis. data is the name of the dataframe containing the relevant information for the meta-analysis.

##sm = "logEffect size": This specifies the summary measure, which in this case is the log effect size. This tells the metagen() function what type of effect measure is being used.

##studlab = data$Population: This argument is used to label the studies (which in this study was the population groups) in the meta-analysis.

##data$Population refers to the column in the dataset that contains the study labels or names.

##comb.fixed = FALSE: This indicates that a fixed-effect model should not be used. Instead, the analysis will use a random-effects model, which accounts for variability both within and between studies.

```
## To view the results of the meta-analysis
print(meta_analysis)

## To create a forest plot
forest(meta_analysis)

## A new forest plot was created to align with the Cochrane Handbook guidelines for layout and formatting of a forest plot
meta::forest(meta_analysis,
        layout = "RevMan5")

## To save the forest plot, a pdf file was created as the ‘meta’ package does not have functionality to save the forest plot directly

pdf(file = "forestplot8.pdf", width = 280, height = 240)
```

Appendix 2.3: Reasons for exclusion of studies generated from search

Table A2.3: Reasons for exclusion of studies from further inclusion in systematic review

Reason for exclusion	**Number of publications**
Does not report on allele frequency data	4
Does not use the ForenSeq™ DNA Signature Prep kit	32
Does not use the ForenSeq™ DNA Signature Prep kit and focused on ancestry prediction	1
Excluded due to expression of concern regarding informed consent	1
Is a performance evaluation study	1
Is a population study but does not make use of MPS	3
Is a population study but does not use the ForenSeq™ DNA Signature Prep kit	16
Is a population study; uses ForenSeq™ DNA Signature Prep kit for confirmatory purposes; does not include sequence-based allele frequency data	2
Is a protocol in a book	1
Is a validation study; does not include population study experiments	15
Is a validation study; ForenSeq™ DNA Signature Prep kit; does not include population study experiments or data	1
Is a validation study; focused on ancestry prediction	1
Not a population study	***205***
Not a population study, based on CE data only	2
Not a population study, focused on mitochondrial DNA	2
Not a population study; focused on ancestry prediction	7
Not a population study; focused on casework samples	27
Not a population study; focused on non-human species	7
Not a population study; focused on phenotype prediction	4
Not a population study; mixture analysis	11
Not a population study; not relevant to forensic genetics	3
Not a population study; software focused	19
Review article	6
Would be included but publication was retracted	2
Total	**372**

Appendix 2.4: Journals in which included studies were published

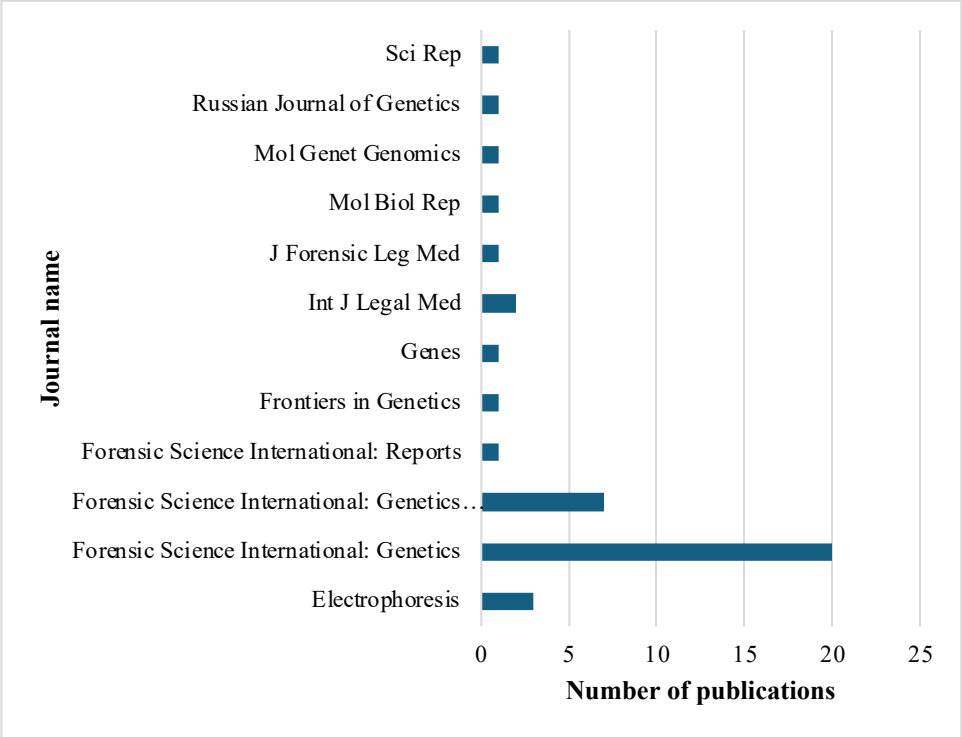

Figure A2.4: Number of publications published from each journal

Appendix 2.5: Meta data for length- and sequence-based allele counts

Table A2.5: Length- and sequence-based allele counts derived from allele frequency data reported in studies included in the meta-analysis

Population	CSF1PO		D10S1248		D12S391		D13S317		D16S539		D17S1301		D18S51	
	LB	SB	LB	SB	LB	SB	LB	SB	LB	SB	LB	SB	LB	SB
El Salvador	10	11	10	10	17	53	9	19	9	10	6	6	19	25
Norwegian	9	9	8	8	16	64	7	7	7	7	8	8	15	17
Northeastern Peruvian Andes	8	8	7	7	14	38	8	17	7	11	8	8	15	16
Mestizos	8	8	9	9	13	34	8	14	8	8	6	7	14	17
Nigerians	7	8	11	11	14	37	6	12	8	14	5	5	14	14
US African American	8	9	7	7	17	52	8	18	7	13	8	10	17	26
US Caucasian	9	9	7	7	16	56	8	16	7	14	8	9	13	16
US Hispanic	8	8	9	9	14	44	9	15	8	14	7	8	14	18
US East Asian	8	9	7	8	13	41	7	16	7	13	7	8	14	14
White British	7	7	8	8	17	55	8	14	8	8	6	6	16	17
British Chinese	8	9	9	9	11	42	8	16	9	9	9	10	13	13
West African	9	10	10	10	19	51	7	15	7	8	7	9	15	16
Northeast African	8	9	10	10	14	44	8	14	8	8	6	7	18	18
South Asian	8	8	8	8	14	50	9	16	7	8	8	8	13	15
Northwest Hui	7	7	7	7	10	33	8	18	7	7	7	7	15	15
Yavapai Native Americans	6	6	7	7	8	16	7	9	6	6	5	5	11	12
Torghut Mongols	7	8	7	7	13	26	7	16	6	10	7	7	14	14
Jalaid Mongols	6	7	7	7	12	29	8	19	7	12	8	8	14	14
Sub-Saharan African	9	9	12	12	19	48	8	17	7	8	6	6	14	16
East Asian	7	7	7	7	12	46	7	18	7	7	9	10	16	16
European	9	10	7	7	16	47	7	13	9	9	6	6	12	13
Middle Eastern	8	9	8	8	15	54	7	15	8	9	6	7	14	14
Native American	5	5	8	8	9	16	7	13	5	5	6	6	11	11
Oceanian	6	7	6	6	9	23	6	6	7	7	5	5	11	11
Central South Asian	6	6	8	8	17	51	8	17	7	7	8	8	14	17
Northern Han Chinese	9	10	10	10	15	61	9	23	8	15	10	11	17	19
Tibetan	7	7	6	6	11	29	7	16	7	12	6	6	15	15
Spanish Roma	6	6	7	7	13	27	7	12	7	7	6	6	10	10
Catalans	8	8	8	8	15	45	7	13	8	8	6	6	13	14
Korean	7	8	8	8	13	48	7	17	6	7	9	10	16	17
Arabian	6	7	8	8	13	38	7	12	7	7	5	5	11	11
Eastern Chinese Han	9	10	8	8	12	40	7	8	7	7	8	9	16	17
French	8	8	8	8	15	48	8	8	8	8	6	6	15	17
Danish	9	9	7	8	17	65	8	14	8	12	7	7	13	15
Hainan Li	7	8	8	9	11	34	7	13	6	6	6	8	15	16

Appendix 2.5 continued

Population	D19S433		D1S1656		D20S482		D21S11		D22S1045		D2S1338		D2S441		D3S1358	
	LB	SB	LB	SB	LB	SB	LB	SB	LB	SB	LB	SB	LB	SB	LB	SB
El Salvador	16	19	15	23	10	10	17	46	-	-	14	42	10	13	8	22
Norwegian	12	16	14	25	9	9	16	34	8	8	12	33	11	15	9	19
Northeastern Peruvian Andes	12	14	14	19	8	13	13	31	9	9	11	25	7	11	8	16
Mestizos	14	15	8	8	12	28	6	6	13	17	10	26	8	12	7	16
Nigerians	12	13	13	18	7	11	16	36	9	11	11	31	9	12	7	14
US African American	16	20	15	24	8	15	18	52	9	13	12	49	11	19	8	18
US Caucasian	12	12	15	24	8	13	12	29	8	8	13	32	8	11	8	14
US Hispanic	12	13	14	21	8	12	16	35	10	11	12	31	9	14	7	18
US East Asian	13	14	13	17	8	13	12	41	7	7	12	30	9	17	6	13
White British	12	12	16	24	9	9	14	29	7	7	13	31	9	14	8	19
British Chinese	11	11	13	17	8	10	15	37	7	7	11	35	9	12	7	13
West African	13	15	15	22	8	8	15	42	-	-	12	53	11	19	8	19
Northeast African	13	13	17	27	8	8	17	39	-	-	12	45	10	18	10	20
South Asian	16	19	17	26	8	8	13	28	-	-	12	40	10	12	7	20
Northwest Hui	9	10	11	16	7	8	12	31	6	6	10	27	9	12	5	11
Yavapai Native Americans	9	9	10	10	5	5	8	16	5	5	9	14	5	7	5	8
Torghut Mongols	10	10	15	17	7	10	11	20	6	6	11	20	8	11	6	6
Jalaid Mongols	10	10	12	15	8	12	12	33	8	8	12	24	10	17	6	11
Sub-Saharan African	13	13	14	23	8	8	18	45	9	11	11	48	13	21	6	15
East Asian	10	10	14	18	9	10	15	39	10	10	11	34	9	14	7	16
European	11	11	13	23	8	8	11	26	7	7	12	27	10	13	6	12
Middle Eastern	10	11	14	24	8	8	9	23	9	10	10	38	10	13	8	21
Native American	10	10	11	14	5	5	10	21	5	5	8	11	7	10	6	10
Oceanian	6	6	11	14	6	6	9	17	5	5	8	12	3	3	4	8
Central South Asian	14	15	16	26	7	7	14	38	8	8	11	36	7	10	7	17
Northern Han Chinese	16	19	13	18	8	16	18	60	10	11	12	50	13	23	8	20
Tibetan	12	13	11	14	6	11	12	28	5	5	11	26	5	11	5	8
Spanish Roma	8	8	13	19	6	6	11	21	6	7	11	22	7	8	6	14
Catalans	11	13	14	26	5	5	15	28	7	7	12	27	8	11	7	17
Korean	12	13	14	18	7	8	14	44	7	7	12	34	10	14	5	13
Arabian	11	12	13	18	7	7	13	21	6	6	12	29	9	13	7	19
Eastern Chinese Han	12	13	11	16	8	8	15	40	8	8	12	31	10	14	6	13
French	15	14	13	23	10	10	13	31	9	9	11	25	9	12	8	18
Danish	14	17	17	28	7	12	13	37	1	1	13	31	10	17	8	22
Hainan Li	13	13	10	13	7	7	10	30	7	9	11	24	9	12	7	12

Appendix 2.5 continued

Population	D4S2408		D5S818		D6S1043		D7S820		D8S1179		D9S1122		FGA		PentaD		PentaE		TH01		TPOX		vWA	
	LB	SB	LB	SB	LB	SB	LB	SB	LB	SB	LB	SB	LB	SB	LB	SB	LB	SB	LB	SB	LB	SB	LB	SB
El Salvador	6	7	9	16	20	25	9	10	10	24	11	18	17	22	13	13	18	21	7	7	9	9	11	26
Norwegian	6	7	6	6	13	19	8	8	9	20	7	15	14	16	12	12	17	17	8	8	5	5	7	18
Northeastern Peruvian Andes	5	7	8	13	16	17	8	14	7	13	7	12	12	14	12	13	20	21	6	6	8	8	7	16
Mestizos	5	6	8	13	14	17	8	9	7	17	7	11	13	15	10	10	17	20	5	5	6	6	7	13
Nigerians	5	5	6	11	13	14	6	8	8	19	8	14	11	15	11	12	13	13	6	6	8	8	11	23
US African American	6	7	8	15	15	19	9	18	11	24	7	14	20	23	14	16	15	19	7	8	7	7	19	32
US Caucasian	6	7	8	13	13	19	8	17	11	19	7	13	15	16	11	15	17	17	9	9	7	7	8	14
US Hispanic	5	6	8	13	16	19	7	12	10	24	6	12	14	20	11	11	19	20	6	6	7	7	11	22
US East Asian	5	7	7	12	13	15	8	18	9	19	8	15	16	16	11	11	19	19	6	6	9	9	9	12
White British	5	6	6	12	11	14	7	7	10	20	7	13	14	15	11	11	16	16	6	6	5	5	8	19
British Chinese	7	9	7	11	13	14	8	9	9	16	7	14	18	19	9	9	14	14	6	6	5	5	8	14
West African	6	7	9	17	15	20	8	8	9	23	8	15	19	23	13	13	15	18	7	7	7	7	10	32
Northeast African	6	7	8	13	15	16	8	8	10	21	8	14	19	24	12	12	15	17	7	7	6	6	10	24
South Asian	6	8	8	14	12	14	7	8	10	21	7	12	16	16	14	14	18	18	7	8	8	8	8	20
Northwest Hui	5	7	8	11	13	13	7	9	10	18	6	11	15	15	10	10	15	15	6	6	6	6	6	11
Yavapai Native Americans	5	6	7	10	12	12	8	8	7	12	6	8	9	10	9	9	15	15	5	5	4	4	7	9
Torghut Mongols	5	7	8	11	11	11	7	15	9	15	7	10	11	11	10	10	18	18	6	6	6	6	7	10
Jalaid Mongols	5	7	7	10	11	11	6	16	11	20	7	11	14	14	10	10	17	17	6	6	6	6	8	13
Sub-Saharan African	7	9	9	15	14	18	10	11	8	22	8	14	18	21	13	13	15	21	8	9	7	7	11	29
East Asian	5	6	7	13	19	21	8	11	10	18	7	11	18	18	10	10	21	22	6	6	6	6	7	14
European	7	8	8	15	13	13	9	9	6	17	7	10	14	15	12	12	12	16	7	7	6	6	8	17
Middle Eastern	5	7	7	14	12	17	8	8	10	20	6	10	17	19	13	13	18	18	6	6	7	7	7	15
Native American	5	6	6	8	13	13	7	7	7	11	6	9	11	11	5	5	15	16	4	4	4	4	7	11
Oceanian	4	5	6	7	11	11	7	7	8	10	6	9	11	11	9	9	13	13	5	5	5	5	7	7
Central South Asian	6	7	6	11	15	17	7	9	10	19	8	14	18	20	10	11	18	18	6	7	6	6	7	14
Northern Han Chinese	6	8	10	18	17	24	13	26	10	21	9	14	19	21	11	12	24	26	6	6	8	8	10	21
Tibetan	5	6	7	11	16	16	7	15	10	18	8	12	15	17	10	10	18	18	6	6	6	6	7	11
Spanish Roma	5	6	7	11	11	11	9	9	8	13	7	11	14	14	7	7	15	15	6	6	5	5	6	12
Catalans	6	7	7	11	11	14	8	9	10	20	6	11	14	16	12	12	16	16	8	8	7	7	7	16
Korean	6	9	8	13	15	17	11	14	12	20	8	13	17	17	9	9	19	19	6	6	6	6	7	16
Arabian	5	5	7	11	10	10	7	7	10	15	6	12	13	15	11	11	15	15	6	7	6	6	7	13
Eastern Chinese Han	5	7	10	10	17	17	10	10	9	17	8	13	14	14	11	11	20	20	7	7	5	5	9	16
French	5	6	8	8	12	15	8	8	10	19	7	12	15	16	9	9	15	15	7	7	5	5	9	20
Danish	5	6	9	16	12	14	9	11	11	26	8	15	13	14	11	12	17	17	8	8	7	7	9	22
Hainan Li	5	6	8	11	12	14	8	9	9	14	7	12	14	14	9	9	18	18	6	6	5	5	7	12

Appendix 2.6: Output for meta-analysis computed using the "meta" package in RStudio

Model	RR	95% CI	z	p-value
Random effects model	1.5272	[1.4825; 1.5733]	27.94	< 0.0001

Quantifying heterogeneity:

Parameter	Estimate	95% CI
Tau^2	0.0016	[0.0000; 0.0085]
tau	0.0406	[0.0000; 0.0922]
I^2	23.50%	[0.0%; 49.8%]
H	1.14	[1.00; 1.41]

Test of heterogeneity: Cochrane's Q-statistic

Q	d.f.	p-value
44.42	34	0.1089

Details on methods:

Inverse variance method used to assign weights

Restricted maximum-likelihood estimator for tau^2

Q-Profile method for confidence interval of tau^2 and tau

Appendix 2.7: Percentage increase in allele count by major ancestral population group

Table A2.7: Table showing percentage increase in allele count per major ancestral population group.

Major ancestral population group	Percentage increase in variation (%)
African	67.16101695
East Asian	51.39470014
European	52.86195286
Middle Eastern	52.20883534
Native American	23.57320099
Oceanian	21.13402062
South Asian	57.59259259
Latin American	55.39437897

Appendix 2.8: Histograms displaying normality of allele count distribution

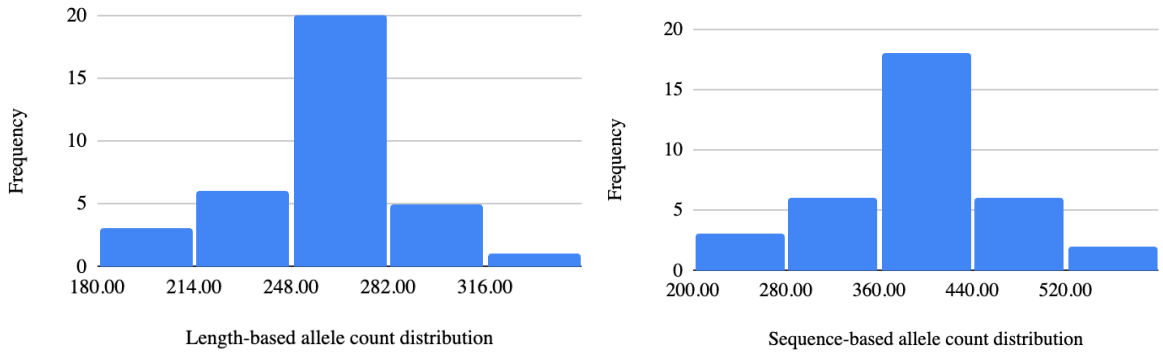

Figure A2.8: (Left) Length-based allele count distribution, (right) sequence-based allele count distributions

Appendix 4.1: RStudio script used for creating bracketed repeat naming format from a tab-separated sequence motifs as input file

```
##Grouping function
group_sequence <- function(seq){
  run_lengths <- rle(seq)
  values <- run_lengths$values
  lengths <- run_lengths$lengths
  res <- character()
  for (i in seq_along(values)) {
    res <- if (lengths[i] > 1) {
      c(res, paste0("[", values[i], "]", lengths[i]))
    } else {
      c(res, values[i])
    }
  }
  return(paste(res, collapse = " "))
}


##Apply the grouping function to the input file containing broken down repeats for the D9S1122 marker (see Appendix 4.2 for example of input file)
D9S1122 <- readxl::read_excel("File path/D9S1122_R_NAMING.xlsx") #import the excel file

##Create data frame called 'DFMarkerName' from the xlsx. file containing broken down repeats
DFD9S1122 <- as.data.frame(D9S1122)

 ##Apply the grouping function to the input file for this marker

DFD9S1122$output <- apply(DFD9S1122[,1:17], 1, function(x) group_sequence(x[!is.na(x)]))

##Check the output (this is optional)
print(DFD9S1122)

 ##write the output to an xlsx (or CSV or tab-delimited txt file), to any file name of your choosing
write_xlsx(DFD9S1122, "outputD9S1122.xlsx")
```

Appendix 4.1 continued: Input file for creating bracketed repeat format

#INSTRUCTIONS:

1. Ensure FASTA sequences are sorted by length in ascending order
2. Use FASTA sequences to generate repeat motifs separated by a tab (this was generated using an Excel-based naming system developed by Kings Forensics)
3. Insert a header row
4. Number each column containing a motif sequence starting from 1
5. An example of the output is shown in column named 'output' - this is not part of the original input file

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	output	
TAGA	TCGA	TAGA	TAGA	TAGA	TAGA	TAGA											TAGA TCGA [TAGA]5	
TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA									[TAGA]9	
TAGA	TCGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA									TAGA TCGA [TAGA]7	
TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA								[TAGA]10	
TAGA	TCGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA								TAGA TCGA [TAGA]8	
TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA							[TAGA]11	
TAGA	TCGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA							TAGA TCGA [TAGA]9	
TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA						[TAGA]12	
TAGA	TCGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA						TAGA TCGA [TAGA]10	
TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA					[TAGA]13	
TAGA	TCGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA					TAGA TCGA [TAGA]11	
TAGA	TCGA	TCGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA					TAGA [TCGA]2 [TAGA]10	
TAGA	TCGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA					TAGA TCGA [TAGA]11	
TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA				[TAGA]14	
TAGA	TCGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA				TAGA TCGA [TAGA]12	
TAGA	TCGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA			TAGA TCGA [TAGA]13	
TAGA	TCGA	TCGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA			TAGA [TCGA]2 [TAGA]12	
TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA			[TAGA]15	
TAGA	TCGA	TA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA		TAGA TCGA TA [TAGA]13	
TAGA	TCGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA		TAGA TCGA [TAGA]14	
TAGA	TCGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA	TAGA TCGA [TAGA]15	

Appendix 4.2: Quality metrics for population study

Table A4.2: Quality metrics for the 24 MiSeq FGx™ runs containing samples used for population databasing. Blocks highlighted in grey fall outside of the manufacturer's recommended ranges.

Experiment Number	Cluster Density (k/mm²)	Clusters Passing Filter (%)	Phasing	Pre-phasing
Run 1	777	93.25	0.132	0.093
Run 2	1069	91.00	0.112	0.095
Run 3	***681***	92.63	0.158	0.113
Run 4	***456***	96.35	0.219	0.098
Run 5	1000	88.95	0.156	0.115
Run 6	1187	87.87	0.143	0.114
Run 7	***520***	91.07	0.205	0.094
Run 8	819	82.73	0.241	0.078
Run 9	864	***78.90***	0.259	0.06
Run 10	1526	***62.26***	0.204	0.066
Run 11	1226	***72.90***	0.198	0.069
Run 12	1101	***77.09***	0.188	0.082
Run 13	1086	***66.56***	0.219	0.112
Run 14	1087	***67.25***	0.213	0.116
Run 15	1175	***78.73***	0.175	0.079
Run 16	993	***56.92***	0.304	0.128
Run 17	1008	***75.21***	0.221	0.087
Run 18	1157	***64.23***	0.307	0.148
Run 19	1133	***66.28***	0.288	0.166
Run 20	1457	84.32	0.168	0.119
Run 21	1498	83.70	0.15	0.124
Run 22	1430	83.95	0.178	0.125
Run 23	1491	84.21	0.173	0.129
Run 24	1246	86.59	0.179	0.076

Appendix 4.3: Sequence-based allele frequency data

Table A4.3: Sequence based allele frequency data for 24 autosomal STRs for 247 Admixed individuals and 216 Black African individuals from the South African population, generated with the ForenSeqTM DNA Signature Prep kit.

Green shaded cells: Potentially novel allele (not seen in STRSeq BioProject)

Marker	Length-based allele	Shorthand name (KCL)	Bracketed Repeat (UAS)	Variant	Allele frequency: Admixed (n = 247)	Allele frequency: Black African (n = 216)
CSF1PO	6	v1_6_S1	[ATCT]6		0.028340081	0.018518519
CSF1PO	7	v1_7_S1	[ATCT]7		0.010121457	0.048611111
CSF1PO	8	v1_8_S1	[ATCT]8		0.010121457	0.037037037
CSF1PO	9	v1_9_S1	[ATCT]9		0.030364372	0.034722222
CSF1PO	10	v1_10_S1	[ATCT]10		0.271255061	0.293981481
CSF1PO	11	v1_11_S1	[ATCT]11		0.259109312	0.196759259
CSF1PO	11	v1_11_S3	[ATCT]7 ACCT [ATCT]3		0.002024291	0.00462963
CSF1PO	12	v1_12_S1	[ATCT]12		0.317813765	0.305555556
CSF1PO	**12.1**	**v1_12.1_s1**	**[ATCT]7 A [ATCT]5**		**0.002024291**	**0**
CSF1PO	13	v1_13_S1	[ATCT]13		0.060728745	0.05787037
CSF1PO	14	v1_14_S1	[ATCT]14		0.008097166	0.002314815
D10S1248	7	v1_7	[GGAA]7		0.002024291	0
D10S1248	10	v1_10	[GGAA]10		0	0.006944444
D10S1248	11	v1_11	[GGAA]11		0.030364372	0.06712963
D10S1248	12	v1_12	[GGAA]12		0.050607287	0.125
D10S1248	13	v1_13	[GGAA]13		0.244939271	0.229166667
D10S1248	14	v1_14	[GGAA]14		0.309716599	0.328703704
D10S1248	15	v1_15	[GGAA]15		0.192307692	0.115740741
D10S1248	16	v1_16	[GGAA]16		0.139676113	0.113425926
D10S1248	17	v1_17	[GGAA]17		0.030364372	0.013888889
D12S391	15	v1_15_S1	[AGAT]8 [AGAC] 6AGAT		0.052631579	0.087962963

D12S391	15	v1_15_S2	[AGAT]9 [AGAC] 5AGAT		0.004048583	0
D12S391	16	v1_16_S1	[AGAT]8 [AGAC]7 AGAT		0.008097166	0.006944444
D12S391	16	v1_16_S2	[AGAT]9 [AGAC]6 AGAT		0.026315789	0.030092593
D12S391	16	v1_16_S3	[AGAT]10 [AGAC]5 AGAT		0.010121457	0.016203704
D12S391	17	v1_17_S1	[AGAT]9 [AGAC]7 AGAT		0.012145749	0.034722222
D12S391	17	v1_17_S2	[AGAT]10 [AGAC]6 AGAT		0.089068826	0.076388889
D12S391	17	v1_17_S3	[AGAT]11 [AGAC]5 AGAT		0.022267206	0.064814815
D12S391	17	v1_17_S4	[AGAT]12 [AGAC]4 AGAT		0	0.002314815
D12S391	**17**	**v1_17_S5**	**[AGAT]8 [AGAC]8 AGAT**		**0.002024291**	**0**
D12S391	17.3	v1_17.3_S1	AGAT GAT [AGAT]8 [AGAC]7 AGAT		0.010121457	0
D12S391	18	v1_18_S3	[AGAT]10 [AGAC]7 AGAT		0.04048583	0.041666667
D12S391	18	v1_18_S4	[AGAT]11 [AGAC]7		0	0.002314815
D12S391	18	v1_18_S5	[AGAT]11 [AGAC]6 AGAT		0.174089069	0.145833333
D12S391	18	v1_18_S6	[AGAT]12 [AGAC]5 AGAT		0.012145749	0.013888889
D12S391	18	v1_18_S7	[AGAT]13 [AGAC]4 AGAT		0	0.011574074
D12S391	**18**	**v1_18_S8**	**[AGAT]12 [AGAC]6**		**0.002024291**	**0**
D12S391	18	v2_18_S3	[AGAT]10 [AGAC]7 AGAT	rs138635218	0.004048583	0.011574074
D12S391	18	v2_18_S5	[AGAT]11[AGAC]6AGAT	rs138635218	0	0.002314815
D12S391	18.3	v1_18.3_S1	AGAT GAT [AGAT]9 [AGAC]7 AGAT		0.006072874	0.002314815
D12S391	19	v1_19_S4	[AGAT]11 [AGAC]8		0.006072874	0.011574074
D12S391	19	v1_19_S5	[AGAT]11 [AGAC]7 AGAT		0.020242915	0.030092593
D12S391	19	v1_19_S6	[AGAT]12 [AGAC]7		0.006072874	0.002314815
D12S391	19	v1_19_S7	[AGAT]12 [AGAC]6 AGAT		0.076923077	0.092592593
D12S391	19	v1_19_S8	[AGAT]13 [AGAC]5 AGAT		0.004048583	0.011574074
D12S391	19	v2_19_S5	[AGAT]11 [AGAC]7 AGAT	rs138635218	0.020242915	0.011574074
D12S391	19	v2_19_S7	[AGAT]12 [AGAC]6 AGAT	rs138635218	0.002024291	0
D12S391	19.1	v1_19.1_S1	AGAT T [AGAT]11 [AGAC]6 AGAT		0	0.00462963
D12S391	19.3	v1_19.3_S1	[AGAT]5 GAT [AGAT]7 [AGAC]6 AGAT		0.002024291	0
D12S391	19.3	v1_19.3_S3	AGAT GAT [AGAT]10 [AGAC]7 AGAT		0.002024291	0
D12S391	20	v1_20_S1	[AGAT]10 [AGAC]9 AGAT		0.004048583	0.002314815

D12S391	20	v1_20_S2	[AGAT]11 [AGAC]9		0.008097166	0
D12S391	20	v1_20_S3	[AGAT]11 [AGAC]8 AGAT		0.006072874	0.016203704
D12S391	20	v1_20_S4	[AGAT]12 [AGAC]8		0.030364372	0.027777778
D12S391	20	v1_20_S5	[AGAT]12 [AGAC]7 AGAT		0.024291498	0.032407407
D12S391	20	v1_20_S6	[AGAT]13 [AGAC]7		0	0.009259259
D12S391	20	v1_20_S7	[AGAT]13 [AGAC]6 AGAT		0.048582996	0.069444444
D12S391	20	v1_20_S8	[AGAT]14 [AGAC]5AGAT		0.004048583	0.00462963
D12S391	20	v2_20_S5	[AGAT]12 [AGAC]7 AGAT	rs138635218	0	0.00462963
D12S391	20.1	v1_20.1_S1	AGAT T [AGAT]12 [AGAC]6 AGAT		0	0.002314815
D12S391	21	v1_21_S1	[AGAT]11 [AGAC]10		0	0.002314815
D12S391	**21**	**V1_21_S12**	**[AGAT]4 AGGT [AGAT]9 [AGAC]6 AGAT**		**0.002024291**	**0**
D12S391	21	v1_21_S3	[AGAT]12 [AGAC]9		0.028340081	0.002314815
D12S391	21	v1_21_S4	[AGAT]12 [AGAC]8 AGAT		0.012145749	0.020833333
D12S391	21	v1_21_S5	[AGAT]13 [AGAC]8		0.002024291	0.006944444
D12S391	21	v1_21_S6	[AGAT]13 [AGAC]7AGAT		0.008097166	0.006944444
D12S391	21	v1_21_S8	[AGAT]14 [AGAC]6 AGAT		0.010121457	0
D12S391	22	v1_22_S2	[AGAT]12 [AGAC]10		0.004048583	0
D12S391	22	v1_22_S4	[AGAT]13 [AGAC]9		0.044534413	0.011574074
D12S391	22	v1_22_S5	[AGAT]13 [AGAC]8 AGAT		0.020242915	0.006944444
D12S391	22	v1_22_S6	[AGAT]14 [AGAC]8		0.010121457	0.002314815
D12S391	22	v1_22_S8	[AGAT]15 [AGAC]6 AGAT		0.002024291	0
D12S391	**22**	**v3_22_S4**	**[AGAT]13 [AGAC]9**	**rs unknown**	**0.002024291**	**0**
D12S391	22.3	v1_22.3_S1	[AGAT]7 GAT [AGAT]6 [AGAC]8 AGAT		0	0.002314815
D12S391	23	v1_23_S1	[AGAT]12 [AGAC]11		0.006072874	0
D12S391	23	v1_23_S3	[AGAT]13 [AGAC]10		0.004048583	0.002314815
D12S391	23	v1_23_S5	[AGAT]14 [AGAC]9		0.01417004	0
D12S391	23	v1_23_S6	[AGAT]14 [AGAC]8 AGAT		0.028340081	0.020833333
D12S391	23	v1_23_S7	[AGAT]15 [AGAC]8		0.002024291	0.002314815
D12S391	23	v1_23_S8	[AGAT]15 [AGAC]7 AGAT		0.002024291	0
D12S391	**23**	**v1_23_S9**	**[AGAT]11 [AGAC]11 AGAT**		**0.002024291**	**0**

D12S391	23	v2_23_S6	[AGAT]14 [AGAC]8 AGAT	rs138635218	0.002024291	0
D12S391	24	v1_24_S1	[AGAT]14 [AGAC]10		0.002024291	0
D12S391	24	v1_24_S2	[AGAT]14 [AGAC]9 AGAT		0	0.002314815
D12S391	24	v1_24_S3	[AGAT]15 [AGAC]9		0.006072874	0
D12S391	24	v1_24_S4	[AGAT]15 [AGAC]8 AGAT		0.008097166	0
D12S391	24	v1_24_S6	[AGAT]11 [AGAC]12 AGAT		0.006072874	0
D12S391	24	v1_24_S7	[AGAT]13 [AGAC]11		0.004048583	0
D12S391	**24**	**v1_24_S8**	**[AGAT]12 [AGAC]11 AGAT**		**0.004048583**	**0.009259259**
D12S391	25	v1_25_S2	[AGAT]15 [AGAC]10		0.002024291	0
D12S391	25	v1_25_S4	[AGAT]16 [AGAC]9		0.002024291	0
D12S391	25	v1_25_S5	[AGAT]16 [AGAC]8 AGAT		0.006072874	0.00462963
D12S391	**25**	**v1_25_S7**	**[AGAT]13 [AGAC]11 AGAT**		**0.002024291**	**0.00462963**
D12S391	25	v1_25_S8	[AGAT]12 [AGAC]12 AGAT		0	0.00462963
D12S391	26	v1_26_S4	[AGAT]17 [AGAC]8 AGAT		0.002024291	0
D12S391	**26**	**v1_26_S5**	**[AGAT]16 [AGAC]9 AGAT**		**0.004048583**	**0**
D12S391	**27**	**v1_27_S1**	**[AGAT]14 [AGAC]12 AGAT**		**0.002024291**	**0**
D12S391	**27**	**V2_27_S2**	**[AGAT]15 [AGAC]11 AGAT**		**0.002024291**	**0.002314815**
D13S317	8	v1_8_S2	[TATC]8 AATC		0.004048583	0
D13S317	8	v1_8_S1	[TATC]8 [AATC]2		0.099190283	0.00462963
D13S317	9	v1_9_S1	[TATC]9 [AATC]2		0.050607287	0.00462963
D13S317	10	v1_10_S2	[TATC]10 AATC		0.004048583	0.002314815
D13S317	10	v5_10_S2	[TATC]11 AATC	rs1442523705	0.002024291	0
D13S317	10	v4_10_S1	[TATC]11 AATC	rs561167308	0.002024291	0.002314815
D13S317	10	v1_10_S1	[TATC]10 [AATC]2		0.044534413	0.00462963
D13S317	11	v1_11_S2	[TATC]11 AATC		0.024291498	0.002314815
D13S317	11	v3_11_S2	[TATC]11 AATC	rs146621667	0.002024291	0.009259259
D13S317	**11**	**v4_11_S2**	**[TATC]12 AATC**	**rs561167308**	**0.004048583**	**0.00462963**
D13S317	11	v1_11_S1	[TATC]11 [AATC]2		0.119433198	0.152777778
D13S317	12	v1_12_S2	[TATC]12 AATC		0.141700405	0.141203704
D13S317	**12**	**v1_12_S5**	**[TATC]6 TATT [TATC]5 AATC**		**0.002024291**	**0**

D13S317	12	v2_12_S2	[TATC]12 AATC	rs73250432	0.004048583	0
D13S317	**12**	**v4_12_S2**	**[TATC]13 AATC**	**rs561167308**	**0.002024291**	**0**
D13S317	13	v6_13_S3	[TATC]13	rs202043589	0.006072874	0
D13S317	12	v1_12_S1	[TATC]12 [AATC]2		0.165991903	0.261574074
D13S317	13	v1_13_S2	[TATC]13 AATC		0.168016194	0.113425926
D13S317	13	v1_13_S4	[TATC]7 TATT [TATC]5 AATC		0.004048583	0.025462963
D13S317	13	v3_13_S2	[TATC]13 AATC	rs146621667	0.006072874	0.00462963
D13S317	13	v1_13_S1	[TATC]13 [AATC]2		0.062753036	0.127314815
D13S317	14	v1_14_S2	[TATC]14 AATC		0.036437247	0.041666667
D13S317	14	v3_14_S2	[TATC]14 AATC	rs146621667	0	0.009259259
D13S317	14	v1_14_S1	[TATC]14 [AATC]2		0.030364372	0.076388889
D13S317	15	v1_15_S2	[TATC]15 AATC		0.012145749	0.00462963
D13S317	**15**	**v1_15_S3**	**[TATC]5 TGTC [TATC]9 AATC**		**0.002024291**	**0.002314815**
D13S317	15	v1_15_S1	[TATC]15 [AATC]2		0	0.002314815
D13S317	16	v1_16_S2	[TATC]5 TGTC [TATC]10 AATC		0	0.002314815
D16S539	8	v1_8_S1	[GATA]8		0.022267206	0.039351852
D16S539	8	v2_8_S1	[GATA]8		0.008097166	0
D16S539	8	v3_8_S1	[GATA]8		0	0.002314815
D16S539	9	v1_9_S1	[GATA]9		0.066801619	0.076388889
D16S539	9	v3_9_S1	[GATA]9		0.125506073	0.108796296
D16S539	10	v1_10_S1	[GATA]10		0.044534413	0.092592593
D16S539	10	v3_10_S1	[GATA]10		0.078947368	0.071759259
D16S539	11	v1_11_S1	[GATA]11		0.285425101	0.349537037
D16S539	11	v3_11_S1	[GATA]11		0.028340081	0.011574074
D16S539	11	v5_11_S1	[GATA]11		0.006072874	0.011574074
D16S539	12	v1_12_S1	[GATA]12		0.178137652	0.138888889
D16S539	12	v3_12_S1	[GATA]12		0.006072874	0.002314815
D16S539	12	v5_12_S1	[GATA]12		0	0.002314815
D16S539	13	v1_13_S1	[GATA]13		0.125506073	0.074074074
D16S539	**13**	**v1_13_S2**	**GATA GGTA [GATA]11**		**0.002024291**	**0**

D16S539	**13**	**v6_13_S1**	**[GATA]13**	**rs unknown**	**0.002024291**	**0**
D16S539	14	v1_14_S1	[GATA]14		0.020242915	0.018518519
D17S1301	8	v1_8_S1	[AGAT]8		0.002024291	0
D17S1301	9	v1_9_S1	[AGAT]9		0.006072874	0.00462963
D17S1301	10	v1_10_S1	[AGAT]10		0.01417004	0.002314815
D17S1301	11	v1_11_S1	[AGAT]11		0.230769231	0.150462963
D17S1301	11.3	v1_11.3_S1	[AGAT]8 GAT [AGAT]3		0	0.002314815
D17S1301	12	v1_12_S1	[AGAT]12		0.439271255	0.502314815
D17S1301	**12**	**v2_12_S1**	**[AGAT]12**	**rs unknown**	**0.002024291**	**0**
D17S1301	13	v1_13_S1	[AGAT]13		0.234817814	0.289351852
D17S1301	13	v1_13_S3	[AGAT]12 CGAT		0.002024291	0.00462963
D17S1301	14	v1_14_S1	[AGAT]14		0.058704453	0.030092593
D17S1301	**14**	**v1_14_S2**	**[AGAT]13 CGAT**		**0.004048583**	**0.002314815**
D17S1301	15	v1_15_S1	[AGAT]15		0.004048583	0.011574074
D17S1301	16	v1_16_S1	[AGAT]16		0.002024291	0
D18S51	9	v2_9_S1	[AGAA]9 AG [AGAG]2	rs535823682	0.002024291	0
D18S51	10	v1_10_S1	[AGAA]10 AAAG AGAG AG		0.002024291	0
D18S51	10	v2_10_S1	[AGAA]10 AG [AGAG]2	rs535823682	0.008097166	0.006944444
D18S51	10.2	v1_10.2_S1	[AGAA]10 AG [AGAG]2 AG		0.004048583	0.020833333
D18S51	11	v1_11_S1	[AGAA]11 AAAG AGAG AG		0.016194332	0.002314815
D18S51	12	v1_12_S1	[AGAA]12 AAAG AGAG AG		0.091093117	0.027777778
D18S51	13	v1_13_S1	[AGAA]13 AAAG AGAG AG		0.091093117	0.025462963
D18S51	13	v2_13_S1	[AGAA]13 AG [AGAG]2	rs535823682	0.006072874	0.006944444
D18S51	14	v1_14_S1	[AGAA]14 AAAG AGAG AG		0.141700405	0.060185185
D18S51	14	v1_14_S2	AGAA AGCA [AGAA]12 AAAG AGAG AG		0.004048583	0
D18S51	**14**	**v1_14_S3**	**[AGAA]12 GGAA AGAA AAAG AGAG AG**		**0.002024291**	**0**
D18S51	14	v2_14_S1	[AGAA]14 AG [AGAG]2	rs535823682	0.002024291	0.002314815
D18S51	14.2	v2_14.2_S1	[AGAA]14 AG [AGAG]2 AG		0.002024291	0.002314815
D18S51	15	v1_15_S1	[AGAA]15 AAAG AGAG AG		0.137651822	0.12962963
D18S51	**15**	**v2_15_S3**	**[AGAA]13 GGAA AGAA AAAG AGAG AG**	**rs535823682**	**0.002024291**	**0**

D18S51	15.2	v2_15.2_S1	[AGAA]15 AG [AGAG]2 AG	rs535823682	0.006072874	0.016203704
D18S51	16	v1_16_S1	[AGAA]16 AAAG AGAG AG		0.143724696	0.162037037
D18S51	**16**	**V1_16_S2**	**[AGAA]14 GGAA AGAA AAAG AGAG AG**		**0.002024291**	**0**
D18S51	16	v2_16_S1	[AGAA]16 AG [AGAG]2	rs535823682	0.002024291	0
D18S51	16.2	v2_16.2_S1	[AGAA]16 AG [AGAG]2 AG	rs535823682	0.002024291	0
D18S51	17	v1_17_S1	[AGAA]17 AAAG AGAG AG		0.099190283	0.145833333
D18S51	**17**	**v1_17_S2**	**[AGAA]15 GGAA AGAA AAAG AGAG AG**		**0**	**0.009259259**
D18S51	18	v1_18_S1	[AGAA]18 AAAG AGAG AG		0.087044534	0.136574074
D18S51	19	v1_19_S1	[AGAA]19 AAAG AGAG AG		0.060728745	0.136574074
D18S51	19	v1_19_S2	[AGAA]15 GGAA [AGAA]3 AAAG AGAG AG		0	0.002314815
D18S51	20	v1_20_S1	[AGAA]20 AAAG AGAG AG		0.05465587	0.078703704
D18S51	20.2	v1_20.2_S1	[AGAA]19 AG AGAA AAAG AGAG AG		0.002024291	0
D18S51	21	v1_21_S1	[AGAA]21 AAAG AGAG AG		0.020242915	0.020833333
D18S51	21.2	v1_21.2_S1	[AGAA]20 AG AGAA AAAG AGAG AG		0	0.00462963
D18S51	22	v1_22_S1	[AGAA]22 AAAG AGAG AG		0.002024291	0.002314815
D18S51	**22.2**	**v1_22.2_S1**	**[AGAA]21 AG AGAA AAAG AGAG AG**		**0.002024291**	**0**
D18S51	23	v1_23_S1	[AGAA]23 AAAG AGAG AG		0.004048583	0
D19S433	**7**	**v1_7_S1**	**[CCTT]5 CCTA CCTT CTTT CCTT**		**0.026315789**	**0.013888889**
D19S433	**7**	**v3_7_S1**	**[CCTT]5 CCTA CCTT CTTT CCTT**	**rs unknown**	**0**	**0.002314815**
D19S433	8	v1_8_S1	[CCTT]6 CCTA CCTT CTTT CCTT		0.002024291	0
D19S433	**8**	**v3_8_S1**	**[CCTT]6 CCTA CCTT CTTT CCTT**	**rs unknown**	**0.002024291**	**0.006944444**
D19S433	10	v1_10_S1	[CCTT]8 CCTA CCTT CTTT CCTT		0.004048583	0.002314815
D19S433	11	v1_11_S1	[CCTT]9 CCTA CCTT CTTT CCTT		0.034412955	0.055555556
D19S433	12	v1_12_S1	[CCTT]10 CCTA CCTT CTTT CCTT		0.091093117	0.12962963
D19S433	12.2	v1_12.2_S1	[CCTT]11 CCTA CCTT TT CCTT		0.028340081	0.043981481
D19S433	12.2	v1_12.2_S2	[CCTT]4 CCCT [CCTT]6 CCTA CCTT TT CCTT		0	0.002314815
D19S433	13	v1_13_S1	[CCTT]11 CCTA CCTT CTTT CCTT		0.25708502	0.321759259
D19S433	13	v1_13_S2	[CCTT]5 CTTT [CCTT]5 CCTA CCTT CTTT CCTT		0.008097166	0
D19S433	**13**	**v1_13_S3**	**[CCTT]13 CTTT CCTT**		**0.002024291**	**0**
D19S433	13.2	v1_13.2_S1	[CCTT]12 CCTA CCTT TT CCTT		0.046558704	0.048611111

D19S433	13.2	v1_13.2_S2	[CCTT]12 CCTA CCTT CTTT CCTT	rs745607776	0.002024291	0
D19S433	14	v1_14_S1	[CCTT]12 CCTA CCTT CTTT CCTT		0.234817814	0.219907407
D19S433	14.2	v1_14.2_S1	[CCTT]13 CCTA CCTT TT CCTT		0.066801619	0.069444444
D19S433	15	v1_15_S1	[CCTT]13 CCTA CCTT CTTT CCTT		0.101214575	0.046296296
D19S433	15.2	v1_15.2_S1	[CCTT]14 CCTA CCTT TT CCTT		0.068825911	0.020833333
D19S433	16	v1_16_S1	[CCTT]14 CCTA CCTT CTTT CCTT		0.012145749	0.00462963
D19S433	16.2	v1_16.2_S1	[CCTT]15 CCTA CCTT TT CCTT		0.008097166	0.006944444
D19S433	16.2	v1_16.2_S2	[CCTT]14 CTTT CCTA CCTT TT CCTT		0	0.002314815
D19S433	17	v1_17_S1	[CCTT]15 CCTA CCTT CTTT CCTT		0.004048583	0.002314815
D1S1656	8	v1_8_S1	[CA]5 [TCTA]8		0.006072874	0
D1S1656	10	v1_10_S1	[CA]5 [TCTA]10		0.002024291	0.013888889
D1S1656	11	v1_11_S1	[CA]5 [TCTA]11		0.099190283	0.064814815
D1S1656	11	v1_11_S2	[CA]5 CCTA [TCTA]10		0.004048583	0
D1S1656	12	v1_12_S1	[CA]5 [TCTA]12		0.04048583	0.034722222
D1S1656	12	v1_12_S2	[CA]5 CCTA [TCTA]11		0.024291498	0.009259259
D1S1656	13	v1_13_S1	[CA]5 [TCTA]13		0.048582996	0.092592593
D1S1656	13	v1_13_S2	[CA]5 CCTA [TCTA]12		0.012145749	0.020833333
D1S1656	13	v1_13_S3	[CA]5 TCTA GCTA [TCTA]11		0.002024291	0
D1S1656	**13.3**	**v1_13.3_S1**	**[CA]5 CCTA [TCTA]11 TCA TCTA**		**0**	**0.00462963**
D1S1656	14	v1_14_S1	[CA]5 [TCTA]14		0.034412955	0.046296296
D1S1656	14	v1_14_S2	[CA]5 CCTA [TCTA]13		0.099190283	0.206018519
D1S1656	14.3	v1_14.3_S1	[CA]5 CCTA [TCTA]9 T CA [TCTA]4		0.004048583	0.025462963
D1S1656	14.3	v1_14.3_S3	[CA]5 CCTA [TCTA]12 T CA TCTA		0.006072874	0.006944444
D1S1656	15	v1_15_S1	[CA]5 CCTA [TCTA]14		0.182186235	0.113425926
D1S1656	15	v1_15_S3	[CA]5 [TCTA]15		0.046558704	0.05787037
D1S1656	**15**	**v3_15_S1**	**[CA]5 CCTA [TCTA]14**	**rs unknown**	**0**	**0.002314815**
D1S1656	15.2	v1_15.2_S1	[CA]6 [TCTA]15		0	0.002314815
D1S1656	15.3	v1_15.3_S1	[CA]5 CCTA [TCTA]10 T CA [TCTA]4		0.032388664	0.011574074
D1S1656	15.3	v1_15.3_S2	[CA]5 CCTA [TCTA]11 T CA [TCTA]3		0.004048583	0
D1S1656	15.3	v1_15.3_S3	[CA]5 CCTA [TCTA]12 T CA [TCTA]2		0.002024291	0

D1S1656	15.3	v1_15.3_S4	[CA]5 CCTA [TCTA]13 T CA TCTA		0.002024291	0
D1S1656	16	v1_16_S1	[CA]5 CCTA [TCTA]15		0.147773279	0.108796296
D1S1656	16	v1_16_S3	[CA]5 [TCTA]16		0.008097166	0.011574074
D1S1656	16.3	v1_16.3_S1	[CA]5 CCTA [TCTA]11 T CA [TCTA]4		0.060728745	0.087962963
D1S1656	**16.3**	**v1_16.3_S4**	**CA CG [CA]3 CCTA [TCTA]11 TCA [TCTA]4**	**rs unknown**	**0.002024291**	**0**
D1S1656	17	v1_17_S1	[CA]5 [TCTA]17		0	0.00462963
D1S1656	17	v1_17_S2	[CA]5 CCTA [TCTA]16		0.05465587	0.043981481
D1S1656	17.3	v1_17.3_S1	[CA]5 CCTA [TCTA]12 T CA [TCTA]4		0.028340081	0.00462963
D1S1656	18	v1_18_S1	[CA]5 [TCTA]18		0	0.002314815
D1S1656	18	v1_18_S2	[CA]5 CCTA [TCTA]17		0.010121457	0.006944444
D1S1656	18.3	v1_18.3_S1	[CA]5 CCTA [TCTA]13 T CA [TCTA]4		0.030364372	0.011574074
D1S1656	19.3	v1_19.3_S1	[CA]5 CCTA [TCTA]14 T CA [TCTA]4		0.004048583	0.00462963
D1S1656	20.3	v1_20.3_S1	[CA]5 CCTA [TCTA]15 T CA [TCTA]4		0.002024291	0
D20S482	9	v1_9_S1	[AGAT]9		0.008097166	0.002314815
D20S482	10	v1_10_S1	[AGAT]10		0.012145749	0.00462963
D20S482	11	v1_11_S1	[AGAT]11		0.012145749	0.00462963
D20S482	11	v2_11_S1	[AGAT]11	rs77560248	0	0.002314815
D20S482	12	v1_12_S1	[AGAT]12		0.058704453	0.027777778
D20S482	12	v2_12_S1	[AGAT]12	rs77560248	0.004048583	0
D20S482	13	v1_13_S1	[AGAT]13		0.232793522	0.173611111
D20S482	13	v2_13_S1	[AGAT]13	rs77560248	0.026315789	0.025462963
D20S482	14	v1_14_S1	[AGAT]14		0.406882591	0.474537037
D20S482	14	v2_14_S1	[AGAT]14	rs77560248	0.028340081	0.046296296
D20S482	15	v1_15_S1	[AGAT]15		0.147773279	0.141203704
D20S482	15	v2_15_S1	[AGAT]15	rs77560248	0.020242915	0.039351852
D20S482	16	v1_16_S1	[AGAT]16		0.038461538	0.050925926
D20S482	16	v2_16_S1	[AGAT]16	rs77560248	0.002024291	0.00462963
D20S482	17	v1_17_S1	[AGAT]17		0.002024291	0.002314815
D21S11	26	v1_26_S1	[TCTA]4 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]8		0.002024291	0.002314815

D21S11	26	v1_26_S2	[TCTA]4 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]9		0.002024291	0.011574074
D21S11	27	v1_27_S1	[TCTA]6 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]8		0.002024291	0
D21S11	27	v1_27_S2	[TCTA]5 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]9		0.010121457	0.030092593
D21S11	27	v1_27_S3	[TCTA]4 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]9		0.020242915	0.048611111
D21S11	28	v1_28_S1	[TCTA]6 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]9		0.008097166	0
D21S11	28	v1_28_S3	[TCTA]5 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]9		0.002024291	0
D21S11	28	v1_28_S4	[TCTA]4 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]10		0.194331984	0.284722222
D21S11	28	v1_28_S5	[TCTA]5 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]10		0	0.009259259
D21S11	28	v1_28_S6	[TCTA]4 [TCTG]7 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]9		0	0.006944444
D21S11	28	v1_28_S7	[TCTA]4 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]11		0.002024291	0
D21S11	29	v1_29_S1	[TCTA]6 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]10		0.030364372	0.002314815
D21S11	**29**	**v1_29_S10**	**[TCTA]4 [TCTG]7 [TCTA]3 TA [TCTA]2 TCA [TCTA]2 TCCA TA [TCTA]11**		**0.008097166**	**0**
D21S11	29	v1_29_S2	[TCTA]6 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]9		0	0.00462963
D21S11	29	v1_29_S4	[TCTA]5 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]11		0	0.002314815

D21S11	29	v1_29_S6	[TCTA]5 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]10		0.01417004	0.018518519
D21S11	29	v1_29_S7	[TCTA]4 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]11		0.111336032	0.081018519
D21S11	29	v1_29_S9	[TCTA]4 [TCTG]7 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]10		0.002024291	0.025462963
D21S11	29.2	v1_29.2_S2	[TCTA]5 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]9 TA TCTA		0.002024291	0
D21S11	30	v1_30_S2	[TCTA]6 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]11		0.026315789	0.002314815
D21S11	30	v1_30_S3	[TCTA]6 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]10		0.002024291	0.009259259
D21S11	30	v1_30_S5	[TCTA]5 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]11		0.038461538	0.043981481
D21S11	30	v1_30_S6	[TCTA]4 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]12		0.066801619	0.030092593
D21S11	30	v1_30_S7	[TCTA]4 [TCTG]7 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]11		0.01417004	0.009259259
D21S11	30	v1_30_S9	[TCTA]5 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]12		0.006072874	0
D21S11	30.2	v1_30.2_S1	[TCTA]5 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]11 TA TCTA		0.002024291	0
D21S11	30.2	v1_30.2_S2	[TCTA]5 [TCTG]6 [TCTA]3 TA [TCTA]2 TCA [TCTA]2 TCCA TA [TCTA]11 TA TCTA		0	0.011574074
D21S11	30.2	v1_30.2_S3	[TCTA]5 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]10 TA TCTA		0.01417004	0.006944444
D21S11	31	v1_31_S1	[TCTA]8 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]10		0.012145749	0

D21S11	**31**	**v1_31_S10**	**[TCTA]5 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA TCTA GCTA [TCTA]10**		**0.002024291**	**0**
D21S11	31	v1_31_S2	[TCTA]7 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]11		0.006072874	0
D21S11	31	v1_31_S3	[TCTA]7 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]10		0.006072874	0.00462963
D21S11	31	v1_31_S4	[TCTA]6 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]12		0.012145749	0
D21S11	31	v1_31_S5	[TCTA]6 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]11		0.016194332	0.034722222
D21S11	31	v1_31_S6	[TCTA]5 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]12		0.030364372	0.041666667
D21S11	31	v1_31_S7	[TCTA]4 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]13		0.01417004	0.00462963
D21S11	**31**	**v1_31_S9**	**[TCTA]4 [TCTG]7 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]12**		**0.008097166**	**0.00462963**
D21S11	31.2	v1_31.2_S1	[TCTA]5 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]12 TA TCTA		0.010121457	0
D21S11	31.2	v1_31.2_S2	[TCTA]5 [TCTG]6 [TCTA]3 TA [TCTA]2 TCA [TCTA]2 TCCA TA [TCTA]12 TA TCTA		0.006072874	0.039351852
D21S11	31.2	v1_31.2_S3	[TCTA]5 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]11 TA TCTA		0.072874494	0.039351852
D21S11	31.2	v1_31.2_S4	[TCTA]4 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]12 TA TCTA		0.002024291	0.00462963
D21S11	32	v1_32_S10	[TCTA]8 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]10		0	0.002314815
D21S11	32	v1_32_S2	[TCTA]8 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]11		0.002024291	0
D21S11	32	v1_32_S4	[TCTA]7 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]12		0.004048583	0

D21S11	32	v1_32_S5	[TCTA]6 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]12		0	0.00462963
D21S11	32	v1_32_S7	[TCTA]5 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]13		0.004048583	0.002314815
D21S11	**32**	**v1_32_S9**	**[TCTA]4 [TCTG]7 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]13**		**0.004048583**	**0.002314815**
D21S11	32.2	v1_32.2_S1	[TCTA]6 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]11 TA TCTA		0.002024291	0
D21S11	32.2	v1_32.2_S3	[TCTA]5 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]12 TA TCTA		0.111336032	0.06712963
D21S11	32.2	v1_32.2_S7	[TCTA]5 [TCTG]6 [TCTA]3 TA [TCTA]2 TCA [TCTA]2 TCCA TA [TCTA]13 TA TCTA		0.002024291	0.002314815
D21S11	**32.2**	**v1_32.2_S8**	**[TCTA]4 [TCTG]7 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]12 TA TCTA**		**0.004048583**	**0**
D21S11	**32.2**	**V1_32.2_S9**	**[TCTA]5 [TCTG]2 TCCG [TCTG]3 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]12 TA TCTA**		**0.002024291**	**0**
D21S11	**33**	**v1_33_S2**	**[TCTA]7 [TCTG]9 [TCTA]2 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]10**		**0.002024291**	**0.002314815**
D21S11	33.1	v1_33.1_S1	[TCTA]5 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]6 TCA [TCTA]6 TA TCTA		0.002024291	0.006944444
D21S11	33.2	v1_33.2_S1	[TCTA]5 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]13 TA TCTA		0.020242915	0.025462963
D21S11	33.2	v1_33.2_S2	[TCTA]6 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]12 TA TCTA		0.004048583	0
D21S11	33.2	v1_33.2_S3	[TCTA]5 [TCTG]7 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]12 TA TCTA		0	0.002314815

D21S11	34	v1_34_S2	[TCTA]10 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]11		0.002024291	0.002314815
D21S11	**34**	**v1_34_S7**	**[TCTA]7 [TCTG]9 [TCTA]2 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]11**		**0.002024291**	**0.002314815**
D21S11	**34.1**	**v1_34.1_S1**	**[TCTA]5 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]9 TCA [TCTA]4 TA TCTA**		**0.002024291**	**0.006944444**
D21S11	34.2	v1_34.2_S1	[TCTA]5 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]14 TA TCTA		0.002024291	0.006944444
D21S11	34.2	v1_34.2_S2	[TCTA]6 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]13 TA TCTA		0.002024291	0.002314815
D21S11	35	v1_35_S2	[TCTA]11 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]11		0	0.002314815
D21S11	35	v1_35_S3	[TCTA]10 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]12		0.002024291	0.011574074
D21S11	35	v1_35_S5	[TCTA]5 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]8 TCA [TCTA]3 TCA [TCTA]2 TA TCTA		0.010121457	0.009259259
D21S11	**35**	**v1_35_S6**	**[TCTA]10 [TCTG]4 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]13**		**0**	**0.002314815**
D21S11	35	v1_35_S7	[TCTA]7 [TCTG]9 [TCTA]2 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]12		0.002024291	0.002314815
D21S11	**35**	**v1_35_S8**	**[TCTA]11 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]10**		**0**	**0.002314815**
D21S11	**35.1**	**v1_35.1_S1**	**[TCTA]5 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]10 TCA [TCTA]4 TA TCTA**		**0.008097166**	**0**

D21S11	35.2	v1_35.2_S1	[TCTA]5 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]15 TA TCTA		0.002024291	0
D21S11	35.2	v1_35.2_S2	[TCTA]6 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]14 TA TCTA		0.002024291	0
D21S11	36	v1_36_S1	[TCTA]10 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]13		0.002024291	0
D21S11	**36**	**v1_36_S3**	**[TCTA]7 [TCTG]9 [TCTA]2 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]13**		**0.002024291**	**0**
D21S11	**36**	**v1_36_S4**	**[TCTA]12 [TCTG]6 [TCTA]2 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]11**		**0**	**0.002314815**
D21S11	**36**	**v1_36_S5**	**[TCTA]11 [TCTG]7 [TCTA]2 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]11**		**0**	**0.002314815**
D21S11	**36**	**v1_36_S6**	**[TCTA]12 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]11**		**0**	**0.002314815**
D21S11	**36**	**v1_36_S7**	**[TCTA]10 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]12**		**0**	**0.002314815**
D21S11	36.1	v1_36.1_S1	[TCTA]5 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]11 TCA [TCTA]4 TA TCTA		0.004048583	0
D21S11	36	v1_36_S2	[TCTA]11 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]13		0	0.00462963
D21S11	**37**	**v1_37_S3**	**[TCTA]11 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]12**		**0.002024291**	**0**
D21S11	**37**	**v1_37_S4**	**[TCTA]13 [TCTG]6 [TCTA]2 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]11**		**0.002024291**	**0**
D21S11	**37.1**	**v1_37.1_S1**	**[TCTA]5 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]12 TCA [TCTA]4 TA TCTA**		**0.002024291**	**0**
D21S11	**37.2**	**V1_37.2_S1**	**[TCTA]11 [TCTG]10 [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]12**		**0.002024291**	**0**
D21S11	38	v1_38_S1	[TCTA]13 [TCTG]7 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]10		0.002024291	0
D21S11	**39**	**v1_39_S1**	**[TCTA]15 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]9**		**0.002024291**	**0**
D21S11	**39**	**v1_39_S2**	**[TCTA]14 [TCTG]7 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]9**		**0**	**0.002314815**
D2S1338	**14**	**v1_14_S2**	**[GGAA]10 [GGCA]4**		**0**	**0.002314815**
D2S1338	**15**	**v1_15_S1**	**[GGAA]12 [GGCA]3**		**0**	**0.002314815**
D2S1338	16	v1_16_S1	[GGAA]9 [GGCA]7		0.002024291	0
D2S1338	16	v1_16_S2	[GGAA]10 [GGCA]6		0.016194332	0.009259259
D2S1338	16	v1_16_S3	[GGAA]11 [GGCA]5		0.006072874	0.032407407
D2S1338	16	v1_16_S4	[GGAA]12 [GGCA]4		0.012145749	0.06712963

D2S1338	**16**	**v1_16_S5**	**[GGAA]13 [GGCA]3**		**0.004048583**	**0**
D2S1338	17	v1_17_S1	[GGAA]10 [GGCA]7		0.008097166	0.013888889
D2S1338	17	v1_17_S2	[GGAA]11 [GGCA]6		0.0951417	0.025462963
D2S1338	17	v1_17_S3	[GGAA]12 [GGCA]5		0.012145749	0.016203704
D2S1338	17	v1_17_S4	[GGAA]13 [GGCA]4		0.004048583	0
D2S1338	17	v1_17_S5	[GGAA]14 [GGCA]3		0.006072874	0.00462963
D2S1338	18	v1_18_S1	[GGAA]11 [GGCA]7		0.04048583	0.002314815
D2S1338	18	v1_18_S2	[GGAA]12 [GGCA]6		0.046558704	0.006944444
D2S1338	18	v1_18_S3	[GGAA]13 [GGCA]5		0.01417004	0.016203704
D2S1338	18	v1_18_S4	[GGAA]14 [GGCA]4		0	0.00462963
D2S1338	18	v1_18_S5	[GGAA]15 [GGCA]3		0.004048583	0.00462963
D2S1338	19	v1_19_S1	[GGAA]11 [GGCA]8		0.004048583	0
D2S1338	19	v1_19_S2	[GGAA]12 [GGCA]7		0.074898785	0.083333333
D2S1338	19	v1_19_S3	[GGAA]13 [GGCA]6		0.028340081	0.06712963
D2S1338	19	v1_19_S4	[GGAA]14 [GGCA]5		0.022267206	0.013888889
D2S1338	19	v1_19_S5	[GGAA]15 [GGCA]4		0.008097166	0.00462963
D2S1338	19	v1_19_S6	[GGAA]16 [GGCA]3		0.006072874	0
D2S1338	20	v1_20_S1	[GGAA]12 [GGCA]8		0.006072874	0.011574074
D2S1338	20	v1_20_S2	[GGAA]13 [GGCA]7		0.087044534	0.069444444
D2S1338	20	v1_20_S3	[GGAA]14 [GGCA]6		0.012145749	0.006944444
D2S1338	20	v1_20_S4	[GGAA]15 [GGCA]5		0.006072874	0.002314815
D2S1338	20	v1_20_S5	[GGAA]16 [GGCA]4		0.002024291	0.002314815
D2S1338	20	v1_20_S6	[GGAA]2 GGAC [GGAA]10 [GGCA]7		0.012145749	0
D2S1338	20	v1_20_S7	[GGAA]2 GGAC [GGAA]11 [GGCA]6		0	0.002314815
D2S1338	21	v1_21_S1	[GGAA]12 [GGCA]9		0	0.002314815
D2S1338	21	v1_21_S2	[GGAA]13 [GGCA]8		0.008097166	0.023148148
D2S1338	21	v1_21_S3	[GGAA]14 [GGCA]7		0.034412955	0.025462963
D2S1338	21	v1_21_S4	[GGAA]15 [GGCA]6		0.006072874	0.00462963
D2S1338	21	v1_21_S5	[GGAA]16 [GGCA]5		0.004048583	0.00462963
D2S1338	21	v1_21_S6	[GGAA]17 [GGCA]4		0.004048583	0

D2S1338	21	v1_21_S8	[GGAA]2 GGAC [GGAA]11 [GGCA]7		0.016194332	0.046296296
D2S1338	21	v1_21_S9	[GGAA]2 GGAC [GGAA]12 [GGCA]6		0.008097166	0.053240741
D2S1338	22	v1_22_S1	[GGAA]13 [GGCA]9		0.008097166	0.018518519
D2S1338	22	v1_22_S10	[GGAA]17 [GGCA]5		0.006072874	0
D2S1338	22	v1_22_S2	[GGAA]14 [GGCA]8		0.008097166	0.00462963
D2S1338	22	v1_22_S3	[GGAA]15 [GGCA]7		0.032388664	0.064814815
D2S1338	22	v1_22_S4	[GGAA]16 [GGCA]6		0.004048583	0.011574074
D2S1338	22	v1_22_S5	[GGAA]2 GGAC [GGAA]12 [GGCA]7		0.032388664	0.037037037
D2S1338	22	v1_22_S6	[GGAA]2 GGAC [GGAA]13 [GGCA]6		0.018218623	0.027777778
D2S1338	23	v1_23_S1	[GGAA]14 [GGCA]9		0.002024291	0
D2S1338	**23**	**v1_23_S11**	**[GGAA]17 [GGCA]6**		**0.004048583**	**0.002314815**
D2S1338	23	v1_23_S2	[GGAA]15 [GGCA]8		0.002024291	0
D2S1338	23	v1_23_S3	[GGAA]16 [GGCA]7		0.01417004	0.013888889
D2S1338	23	v1_23_S4	[GGAA]2 GGAC [GGAA]12 [GGCA]8		0.002024291	0
D2S1338	23	v1_23_S5	[GGAA]2 GGAC [GGAA]13 [GGCA]7		0.085020243	0.069444444
D2S1338	23	v1_23_S6	[GGAA]2 GGAC [GGAA]14 [GGCA]6		0.012145749	0.009259259
D2S1338	24	v1_24_S2	[GGAA]16 [GGCA]8		0	0.002314815
D2S1338	24	v1_24_S4	[GGAA]2 GGAC [GGAA]14 [GGCA]7		0.050607287	0.053240741
D2S1338	24	v1_24_S5	[GGAA]2 GGAC [GGAA]15 [GGCA]6		0.006072874	0.00462963
D2S1338	24	v1_24_S6	[GGAA]2 GGAC [GGAA]16 [GGCA]5		0.004048583	0
D2S1338	24	v1_24_S7	[GGAA]2 GGAC AGAA [GGAA]13 [GGCA]7		0	0.002314815
D2S1338	**24**	**v1_24_S8**	**[GGAA]17 [GGCA]7**		**0.004048583**	**0**
D2S1338	25	v1_25_S2	[GGAA]2 GGAC [GGAA]14 [GGCA]8		0.004048583	0
D2S1338	25	v1_25_S3	[GGAA]2 GGAC [GGAA]15 [GGCA]7		0.056680162	0.030092593
D2S1338	25	v1_25_S4	[GGAA]2 GGAC [GGAA]16 [GGCA]6		0.008097166	0
D2S1338	26	v1_26_S1	[GGAA]2 GGAC [GGAA]15 [GGCA]8		0.006072874	0.006944444
D2S1338	26	v1_26_S2	[GGAA]2 GGAC [GGAA]16 [GGCA]7		0.004048583	0.009259259
D2S1338	26	v1_26_S3	[GGAA]2 GGAC [GGAA]17 [GGCA]6		0.002024291	0
D2S1338	27	v1_27_S3	[GGAA]2 GGAC [GGAA]18 [GGCA]6		0.002024291	0
D2S441	9	v1_9_S1	[TCTA]9		0.002024291	0

D2S441	9.1	v1_9.1_S1	A [TCTA]9		0.002024291	0
D2S441	10	v1_10_S1	[TCTA]10		0.022267206	0.018518519
D2S441	10	v1_10_S2	[TCTA]8 TCTG TCTA		0.113360324	0.011574074
D2S441	10	v2_10_S1	[TCTA]10	rs74640515	0.018218623	0
D2S441	**10.3**	**v1_10.3_S1**	**[TCTA]3 TCA [TCTA]7**		**0**	**0.00462963**
D2S441	11	v1_11_S1	[TCTA]11		0.327935223	0.331018519
D2S441	11	v1_11_S2	[TCTA]9 TCTG TCTA		0.018218623	0.018518519
D2S441	11	v2_11_S1	[TCTA]11	rs74640515	0.022267206	0.002314815
D2S441	11.3	v1_11.3_S1	[TCTA]4 TCA [TCTA]7		0.066801619	0.018518519
D2S441	11.3	v1_11.3_S2	[TCTA]3 TCA [TCTA]8		0.004048583	0.002314815
D2S441	12	v1_12_S1	[TCTA]12		0.0951417	0.168981481
D2S441	12	v1_12_S2	[TCTA]10 TCTG TCTA		0.008097166	0.009259259
D2S441	12	v2_12_S1	[TCTA]12	rs74640515	0.002024291	0
D2S441	12	v3_12_S2	[TCTA]10 TCTG TCTA	rs1033072608	0.002024291	0.023148148
D2S441	**12**	**v4_12_S1**	**[TCTA]12**	**rs unknown**	**0**	**0.002314815**
D2S441	12.3	v1_12.3_S1	[TCTA]4 TCA [TCTA]8		0.004048583	0.002314815
D2S441	12.3	v1_12.3_S3	[TCTA]3 TCA [TCTA]9		0.010121457	0.009259259
D2S441	13	v1_13_S1	[TCTA]13		0.004048583	0.011574074
D2S441	13	v1_13_S2	[TCTA]11 TCTG TCTA		0.004048583	0
D2S441	13	v1_13_S3	[TCTA]10 TTTA [TCTA]2		0.026315789	0.018518519
D2S441	13.3	v1_13.3_S2	[TCTA]3 TCA [TCTA]10		0.012145749	0
D2S441	14	v1_14_S1	[TCTA]11 TTTA [TCTA]2		0.208502024	0.291666667
D2S441	**14**	**v5_14_S1**	**[TCTA]11 TTTA [TCTA]2**	**rs unknown**	**0**	**0.002314815**
D2S441	14.3	v1_14.3_S1	[TCTA]3 TCA [TCTA]11		0.002024291	0.00462963
D2S441	15	v1_15_S1	[TCTA]12 TTTA [TCTA]2		0.020242915	0.046296296
D2S441	16	v1_16_S1	[TCTA]13 TTTA [TCTA]2		0.004048583	0.002314815
D3S1358	12	v1_12_S2	TCTA [TCTG]2 [TCTA]9		0	0.002314815
D3S1358	13	v1_13_S1	TCTA TCTG [TCTA]11		0	0.002314815
D3S1358	13	v1_13_S2	TCTA [TCTG]2 [TCTA]10		0.004048583	0
D3S1358	14	v1_14_S1	TCTA TCTG [TCTA]12		0.036437247	0.05787037

D3S1358	14	v1_14_S2	TCTA [TCTG]2 [TCTA]11		0.068825911	0.023148148
D3S1358	15	v1_15_S1	TCTA TCTG [TCTA]13		0.08097166	0.173611111
D3S1358	15	v1_15_S2	TCTA [TCTG]2 [TCTA]12		0.210526316	0.138888889
D3S1358	15	v1_15_S3	TCTA [TCTG]3 [TCTA]11		0.012145749	0.009259259
D3S1358	15.2	v1_15.2_S1	TCTA [TCTG]3 TC [TCTA]11		0.002024291	0
D3S1358	16	v1_16_S1	TCTA TCTG [TCTA]14		0.060728745	0.131944444
D3S1358	16	v1_16_S2	TCTA [TCTG]2 [TCTA]13		0.16194332	0.148148148
D3S1358	16	v1_16_S3	TCTA [TCTG]3 [TCTA]12		0.052631579	0.041666667
D3S1358	17	v1_17_S1	TCTA TCTG [TCTA]15		0.01417004	0.027777778
D3S1358	17	v1_17_S2	TCTA [TCTG]2 [TCTA]14		0.109311741	0.099537037
D3S1358	17	v1_17_S4	TCTA [TCTG]3 [TCTA]13		0.08097166	0.085648148
D3S1358	18	v1_18_S1	TCTA TCTG [TCTA]16		0.002024291	0.011574074
D3S1358	18	v1_18_S2	TCTA [TCTG]2 [TCTA]15		0.028340081	0.016203704
D3S1358	18	v1_18_S3	TCTA [TCTG]3 [TCTA]14		0.068825911	0.030092593
D3S1358	18	v1_18_S4	TCTA [TCTG]4 [TCTA]13		0.002024291	0
D3S1358	19	v1_19_S2	TCTA [TCTG]3 [TCTA]15		0.004048583	0
D4S2408	8	v1_8_S1	[ATCT]8		0.170040486	0.111111111
D4S2408	9	v1_9_S1	[ATCT]9		0.17611336	0.122685185
D4S2408	9	v1_9_S2	ATCT GTCT [ATCT]7		0.046558704	0.00462963
D4S2408	10	v1_10_S1	[ATCT]10		0.248987854	0.247685185
D4S2408	11	v1_11_S1	[ATCT]11		0.240890688	0.335648148
D4S2408	**11**	**v2_11_S1**	**[ATCT]11**	**rs unknown**	**0.002024291**	**0**
D4S2408	12	v1_12_S1	[ATCT]12		0.111336032	0.175925926
D4S2408	13	v1_13_S1	[ATCT]13		0.004048583	0.002314815
D5S818	8	v1_8_S1	[ATCT]8	rs73801920	0.026315789	0.085648148
D5S818	8	v1_8_S2	[ATCT]8		0.002024291	0
D5S818	9	v1_9_S1	[ATCT]9		0.004048583	0
D5S818	9	v1_9_S2	[ATCT]9	rs73801920	0.020242915	0.027777778
D5S818	10	v1_10_S1	[ATCT]10		0.087044534	0.050925926
D5S818	10	v1_10_S2	[ATCT]10	rs73801920	0.018218623	0.018518519

D5S818	11	v1_11_S1	[ATCT]11		0.222672065	0.199074074
D5S818	11	v1_11_S2	[ATCT]11	rs73801920	0.050607287	0.048611111
D5S818	12	v1_12_S1	[ATCT]12		0.240890688	0.206018519
D5S818	12	v1_12_S2	[ATCT]12	rs73801920	0.103238866	0.104166667
D5S818	13	v1_13_S1	[ATCT]13		0.159919028	0.210648148
D5S818	13	v1_13_S2	[ATCT]13	rs73801920	0.04048583	0.020833333
D5S818	13	v1_13_S3	[ATCT]3 ATGT [ATCT]9		0.010121457	0.009259259
D5S818	14	v1_14_S1	[ATCT]14		0.012145749	0.013888889
D5S818	15	v1_15_S1	[ATCT]15		0.002024291	0.00462963
D6S1043	9	v1_9_S1	[ATCT]9		0	0.009259259
D6S1043	10	v1_10_S1	[ATCT]10		0.020242915	0.00462963
D6S1043	11	v1_11_S1	[ATCT]11		0.21659919	0.118055556
D6S1043	12	v1_12_S1	[ATCT]12		0.228744939	0.212962963
D6S1043	13	v1_13_S1	[ATCT]13		0.066801619	0.074074074
D6S1043	**13**	**v1_13_S2**	**ATCT ATGT [ATCT]11**		**0.002024291**	**0**
D6S1043	**13.3**	**v1_13.3_S1**	**[ATCT]9 ATC [ATCT]4**		**0.002024291**	**0**
D6S1043	14	v1_14_S1	[ATCT]14		0.085020243	0.048611111
D6S1043	14	v1_14_S3	[ATCT]5 ATGT [ATCT]8		0.020242915	0.011574074
D6S1043	15	v1_15_S1	[ATCT]15		0.010121457	0.00462963
D6S1043	15	v1_15_S2	[ATCT]5 ATGT [ATCT]9		0.036437247	0.097222222
D6S1043	15	v1_15_S3	ATCT ATGT [ATCT]13		0.002024291	0
D6S1043	16	v1_16_S1	[ATCT]5 ATGT [ATCT]10		0.008097166	0.039351852
D6S1043	16	v1_16_S2	[ATCT]4 ATGT [ATCT]11		0.004048583	0.006944444
D6S1043	16	v1_16_S3	ATCT ATGT [ATCT]14		0.002024291	0
D6S1043	**16**	**v1_16_S4**	**[ATCT]5 ATGT [ATCT]7 ATGT [ATCT]2**		**0**	**0.002314815**
D6S1043	17	v1_17_S2	[ATCT]5 ATGT [ATCT]11		0.044534413	0.0625
D6S1043	17	v1_17_S3	[ATCT]5 ATGT [ATCT]8 ATGT [ATCT]2		0.004048583	0.020833333
D6S1043	17	v1_17_S4	[ATCT]4 ATGT [ATCT]12		0	0.002314815
D6S1043	**17**	**v1_17_S5**	**ATCA [ATCT]4 ATGT [ATCT]11**		**0**	**0.002314815**
D6S1043	18	v1_18_S1	[ATCT]5 ATGT [ATCT]12		0.087044534	0.092592593

D6S1043	19	v1_19_S1	[ATCT]5 ATGT [ATCT]13		0.082995951	0.134259259
D6S1043	20	v1_20_S1	[ATCT]5 ATGT [ATCT]14		0.060728745	0.053240741
D6S1043	21	v1_21_S1	[ATCT]6 ATGT [ATCT]14		0.002024291	0
D6S1043	21	v1_21_S2	[ATCT]5 ATGT [ATCT]15		0.006072874	0
D6S1043	21.3	v1_21.3_S1	[ATCT]5 ATGT [ATCT]2 ATC [ATCT]13		0.008097166	0.002314815
D7S820	7	v1_7_S1	[TATC]7	rs7789995	0.018218623	0.002314815
D7S820	8	v1_8_S1	[TATC]8	rs7789995	0.115384615	0.122685185
D7S820	8	v2_8_S1	[TATC]8	rs7789995, rs16887642	0.072874494	0.05787037
D7S820	**8**	**v4_8_S1**	**[TATC]8**	**rs7789995**	**0.002024291**	**0**
D7S820	9	v1_9_S1	[TATC]9	rs7789995	0.087044534	0.081018519
D7S820	9	v2_9_S1	[TATC]9	rs7789995, rs16887642	0.028340081	0.030092593
D7S820	9	v3_9_S1	[TATC]9		0	0.002314815
D7S820	10	v1_10_S1	[TATC]10	rs7789995	0.224696356	0.314814815
D7S820	10	v2_10_S1	[TATC]10	rs7789995, rs16887642	0.020242915	0.00462963
D7S820	10	v3_10_S1	[TATC]10		0.010121457	0.006944444
D7S820	**10**	**v5_10_S1**	**[TATC]10**	**rs7789995, rs unknown**	**0.002024291**	**0**
D7S820	11	v1_11_S1	[TATC]11	rs7789995	0.165991903	0.185185185
D7S820	11	v2_11_S1	[TATC]11	rs7789995, rs16887642	0.012145749	0.016203704
D7S820	11	v3_11_S1	[TATC]11		0.024291498	0.050925926
D7S820	12	v1_12_S1	[TATC]12	rs7789995	0.13562753	0.094907407
D7S820	12	v2_12_S1	[TATC]12	rs7789995, rs16887642	0.010121457	0.006944444
D7S820	12	v3_12_S1	[TATC]12		0.022267206	0.00462963
D7S820	**12**	**v6_12_S1**	**[TATC]12**	**rs7789995, rs unknown**	**0.002024291**	**0**
D7S820	13	v1_13_S1	[TATC]13	rs7789995	0.038461538	0.013888889
D7S820	13	v2_13_S1	[TATC]13	rs7789995, rs16887642	0.002024291	0.002314815
D7S820	13	v3_13_S1	[TATC]13		0.002024291	0
D7S820	14	v1_14_S1	[TATC]14	rs7789995	0.002024291	0.002314815
D7S820	**15**	**v1_15_S1**	**[TATC]15**	**rs7789995**	**0.002024291**	**0**
D8S1179	8	v1_8_S1	[TCTA]8		0.010121457	0

D8S1179	9	v1_9_S1	[TCTA]9		0.012145749	0
D8S1179	10	v1_10_S1	[TCTA]10		0.066801619	0.002314815
D8S1179	11	v1_11_S1	[TCTA]11		0.060728745	0.013888889
D8S1179	11	v1_11_S2	[TCTA]2 TCTG [TCTA]8		0.008097166	0.025462963
D8S1179	11	v1_11_S3	TCTA TCTG [TCTA]9		0	0.002314815
D8S1179	12	v1_12_S1	[TCTA]12		0.072874494	0.046296296
D8S1179	12	v1_12_S2	[TCTA]2 TCTG [TCTA]9		0.022267206	0.037037037
D8S1179	12	v1_12_S3	TCTA TCTG [TCTA]10		0.016194332	0.006944444
D8S1179	13	v1_13_S1	[TCTA]13		0.044534413	0.006944444
D8S1179	13	v1_13_S2	[TCTA]2 TCTG [TCTA]10		0.036437247	0.097222222
D8S1179	13	v1_13_S3	TCTA TCTG [TCTA]11		0.137651822	0.157407407
D8S1179	14	v1_14_S1	[TCTA]14		0.010121457	0.006944444
D8S1179	14	v1_14_S2	[TCTA]2 TCTG [TCTA]11		0.111336032	0.104166667
D8S1179	14	v1_14_S4	TCTA TCTG [TCTA]12		0.129554656	0.1875
D8S1179	14	v1_14_S5	TCTA TCTG TGTA [TCTA]11		0.002024291	0
D8S1179	15	v1_15_S1	[TCTA]15		0.004048583	0
D8S1179	15	v1_15_S2	[TCTA]2 TCTG [TCTA]12		0.125506073	0.150462963
D8S1179	15	v1_15_S3	[TCTA]2 [TCTG]2 [TCTA]11		0.01417004	0.037037037
D8S1179	15	v1_15_S5	TCTA TCTG [TCTA]13		0.034412955	0.025462963
D8S1179	16	v1_16_S1	[TCTA]2 TCTG [TCTA]13		0.062753036	0.064814815
D8S1179	16	v1_16_S2	[TCTA]2 [TCTG]2 [TCTA]12		0	0.011574074
D8S1179	16	v1_16_S3	TCTA TCTG [TCTA]14		0.002024291	0.009259259
D8S1179	17	v1_17_S1	[TCTA]2 TCTG [TCTA]14		0.006072874	0.006944444
D8S1179	17	v1_17_S2	[TCTA]2 [TCTG]2 [TCTA]13		0.002024291	0
D8S1179	17	v1_17_S3	TCTA TCTG [TCTA]15		0.002024291	0
D8S1179	18	v1_18_S1	[TCTA]2 TCTG [TCTA]15		0.002024291	0
D8S1179	18	v1_18_S2	[TCTA]2 [TCTG]2 [TCTA]14		0.002024291	0
D8S1179	19	v1_19_S1	[TCTA]2 [TCTG]2 [TCTA]15		0.002024291	0
D9S1122	9	v1_9_S1	[TAGA]9		0.002024291	0.009259259
D9S1122	9	v1_9_S2	TAGA TCGA [TAGA]7		0.016194332	0.048611111

D9S1122	10	v1_10_S1	[TAGA]10		0.010121457	0.002314815
D9S1122	10	v1_10_S2	TAGA TCGA [TAGA]8		0.006072874	0.00462963
D9S1122	11	v1_11_S1	[TAGA]11		0.093117409	0.071759259
D9S1122	11	v1_11_S2	TAGA TCGA [TAGA]9		0.070850202	0.071759259
D9S1122	12	v1_12_S1	[TAGA]12		0.16194332	0.090277778
D9S1122	12	v1_12_S2	TAGA TCGA [TAGA]10		0.232793522	0.275462963
D9S1122	13	v1_13_S1	[TAGA]13		0.052631579	0.020833333
D9S1122	13	v1_13_S2	TAGA TCGA [TAGA]11		0.277327935	0.335648148
D9S1122	13	v2_13_S1	TAGA TCGA [TAGA]11	rs149309595	0.002024291	0.002314815
D9S1122	14	v1_14_S1	[TAGA]14		0.010121457	0.002314815
D9S1122	14	v1_14_S2	TAGA TCGA [TAGA]12		0.05465587	0.060185185
D9S1122	15	v1_15_S1	[TAGA]15		0.002024291	0.002314815
D9S1122	15	v1_15_S2	TAGA TCGA [TAGA]13		0.006072874	0.002314815
D9S1122	16	v1_16_S1	TAGA TCGA [TAGA]14		0.002024291	0
FGA	16	v1_16_S1	[GGAA]2 GGAG [AAAG]8 AG [AA]3 [GAAA]3		0.002024291	0
FGA	**16.1**	**v1_16.1_S1**	**[GGAA]2 GGAG [AAAG]3 A [AAAG]5 AG [AA]3 [GAAA]3**		**0**	**0.013888889**
FGA	17	v1_17_S1	[GGAA]2 GGAG [AAAG]9 AG [AA]3 [GAAA]3		0.002024291	0
FGA	18	v1_18_S1	[GGAA]2 GGAG [AAAG]10 AG [AA]3 [GAAA]3		0.010121457	0.00462963
FGA	18.2	v1_18.2_S1	[GGAA]2 GGAG [AAAG]11 [AA]3 [GAAA]3		0.004048583	0.016203704
FGA	19	v1_19_S1	[GGAA]2 GGAG [AAAG]11 AG [AA]3 [GAAA]3		0.074898785	0.050925926
FGA	19.2	v1_19.2_S1	[GGAA]2 GGAG [AAAG]12 [AA]3 [GAAA]3		0.008097166	0.025462963
FGA	20	v1_20_S1	[GGAA]2 GGAG [AAAG]12 AG [AA]3 [GAAA]3		0.109311741	0.06712963
FGA	20.2	v1_20.2_S1	[GGAA]2 GGAG [AAAG]13 [AA]3 [GAAA]3		0	0.002314815
FGA	21	v1_21_S1	[GGAA]2 GGAG [AAAG]13 AG [AA]3 [GAAA]3		0.129554656	0.087962963
FGA	21.2	v1_21.2_S1	[GGAA]2 GGAG [AAAG]14 [AA]3 [GAAA]3		0.006072874	0.002314815
FGA	22	v1_22_S1	[GGAA]2 GGAG [AAAG]14 AG [AA]3 [GAAA]3		0.159919028	0.166666667
FGA	**22**	**v1_22_S2**	**GGAG GGAA GGAG [AAAG]14 AG [AA]3 [GAAA]3**		**0.002024291**	**0.002314815**
FGA	22.2	v1_22.2_S1	[GGAA]2 GGAG [AAAG]15 [AA]3 [GAAA]3		0.004048583	0.002314815
FGA	23	v1_23_S1	[GGAA]2 GGAG [AAAG]15 AG [AA]3 [GAAA]3		0.141700405	0.164351852
FGA	23	v1_23_S2	GGAG GGAA GGAG [AAAG]15 AG [AA]3 [GAAA]3		0.002024291	0.002314815

FGA	23.2	v1_23.2_S1	[GGAA]2 GGAG [AAAG]16 [AA]3 [GAAA]3		0.006072874	0.002314815
FGA	24	v1_24_S1	[GGAA]2 GGAG [AAAG]16 AG [AA]3 [GAAA]3		0.13562753	0.143518519
FGA	24	v1_24_S2	[GGAA]2 GGAG [AAAG]16 AG [AA]3 GAAA GCAA GAAA		0	0.00462963
FGA	**24**	**v1_24_S3**	**GGAG GGAA GGAG [AAAG]16 AG [AA]3 [GAAA]3**		**0.002024291**	**0**
FGA	**24**	**v2_24_S1**	**[GGAA]2 GGAG [AAAG]16 AG [AA]3 [GAAA]3**	**rs unknown**	**0.002024291**	**0.002314815**
FGA	25	v1_25_S1	[GGAA]2 GGAG [AAAG]17 AG [AA]3 [GAAA]3		0.093117409	0.12037037
FGA	**25**	**v1_25_S2**	**GGAG GGAA GGAG [AAAG]17 AG [AA]3 [GAAA]3**		**0.004048583**	**0**
FGA	25.2	v1_25.2_S1	[GGAA]2 GGAG [AAAG]18 [AA]3 [GAAA]3		0.002024291	0
FGA	26	v1_26_S1	[GGAA]2 GGAG [AAAG]5 AA GGAA AG [AAAG]11 AG [AA]3 [GAAA]3		0.006072874	0.016203704
FGA	26	v1_26_S2	[GGAA]2 GGAG [AAAG]18 AG [AA]3 [GAAA]3		0.046558704	0.032407407
FGA	**26**	**v1_26_S4**	**GGAG GGAA GGAG [AAAG]18 AG [AA]3 [GAAA]3**		**0.002024291**	**0**
FGA	**26**	**v2_26_S2**	**[GGAA]2 GGAG [AAAG]18 AG [AA]3 [GAAA]3**	**rs unknown**	**0**	**0.009259259**
FGA	27	v1_27_S1	[GGAA]2 GGAG [AAAG]5 AA GGAA AG [AAAG]12 AG [AA]3 [GAAA]3		0.016194332	0.039351852
FGA	27	v1_27_S2	[GGAA]2 GGAG [AAAG]19 AG [AA]3 [GAAA]3		0.01417004	0.002314815
FGA	28	v1_28_S2	[GGAA]2 GGAG [AAAG]20 AG [AA]3 [GAAA]3		0	0.00462963
FGA	28	v1_28_S3	[GGAA]2 GGAG [AAAG]5 AA GGAA AG [AAAG]13 AG [AA]3 [GAAA]3		0.002024291	0.006944444
FGA	29	v1_29_S1	[GGAA]2 GGAG [AAAG]5 AA GGAA AG [AAAG]14 AG [AA]3 [GAAA]3		0.006072874	0
FGA	29	v1_29_S2	[GGAA]2 GGAG [AAAG]21 AG [AA]3 [GAAA]3		0	0.002314815
FGA	29.2	v1_29.2_S1	[GGAA]4 GGAG [AAAG]3 [GAAG]3 [AAAG]13 [AA]3 [GAAA]4		0	0.002314815
FGA	**30.2**	**v1_30.2_S1**	**[GGAA]4 GGAG [AAAG]3 [GAAG]3 [AAAG]14 [AA]3 [GAAA]4**		**0**	**0.002314815**
FGA	31.2	v1_31.2_S1	[GGAA]4 GGAG [AAAG]3 [GAAG]3 [AAAG]15 [AA]3 [GAAA]4		0.004048583	0
FGA	**43.2**	**v1_43.2_S2**	**[GGAA]4 GGAG [AAAG]3 [GAAG]4 [AAAG]13 [ACAG]5 [AAAG]8 [AA]3 [GAAA]4**		**0.002024291**	**0**
TH01	5	v1_5_S1	[AATG]5		0.002024291	0.00462963
TH01	6	v1_6_S1	[AATG]6		0.170040486	0.12037037
TH01	7	v1_7_S1	[AATG]7		0.246963563	0.319444444
TH01	8	v1_8_S1	[AATG]8		0.251012146	0.337962963
TH01	9	v1_9_S1	[AATG]9		0.184210526	0.171296296
TH01	9.3	v1_9.3_S1	[AATG]6 ATG [AATG]3		0.133603239	0.034722222

TH01	10	v1_10_S1	[AATG]10		0.012145749	0.011574074
TPOX	6	v1_6	[AATG]6		0.016194332	0.097222222
TPOX	7	v1_7	[AATG]7		0.002024291	0.018518519
TPOX	8	v1_8	[AATG]8		0.400809717	0.305555556
TPOX	9	v1_9	[AATG]9		0.236842105	0.208333333
TPOX	10	v1_10	[AATG]10		0.060728745	0.094907407
TPOX	11	v1_11	[AATG]11		0.269230769	0.263888889
TPOX	12	v1_12	[AATG]12		0.01417004	0.011574074
VWA	11	v1_11_S1	TAGA TGGA [TAGA]7 [CAGA]3 TAGA		0.010121457	0.002325581
VWA	13	v1_13_S1	TAGA TGGA [TAGA]8 [CAGA]4 TAGA		0	0.004651163
VWA	13	v2_13_S2	[TAGA]10 [CAGA]4 TAGA	rs771794429	0.002024291	0.004651163
VWA	14	v1_14_S1	TAGA TGGA [TAGA]9 [CAGA]4 TAGA		0.008097166	0.004651163
VWA	14	v1_14_S2	[TAGA]11 [CAGA]4 TAGA		0.002024291	0
VWA	14	v1_14_S3	TAGA TGGA [TAGA]10 [CAGA]3 TAGA		0.004048583	0.002325581
VWA	14	v1_14_S4	[TGGA]2 [TAGA]3 TGGA [TAGA]3 [CAGA]4 TAGA CAGA TAGA		0.072874494	0.074418605
VWA	14	v2_14_S4	[TAGA]11 [CAGA]4 TAGA	rs771794429	0.010121457	0.023255814
VWA	15	v1_15_S2	TAGA TGGA [TAGA]10 [CAGA]4 TAGA		0.099190283	0.095348837
VWA	15	v1_15_S3	[TAGA]12 [CAGA]4 TAGA		0.004048583	0.002325581
VWA	15	v1_15_S5	TAGA TGGA [TAGA]11 [CAGA]3 TAGA		0.034412955	0.065116279
VWA	16	v1_16_S1	TAGA TGGA [TAGA]11 [CAGA]4 TAGA		0.204453441	0.190697674
VWA	16	v1_16_S4	TAGA TGGA [TAGA]12 [CAGA]3 TAGA		0.030364372	0.039534884
VWA	17	v1_17_S1	TAGA TGGA [TAGA]11 [CAGA]5 TAGA		0.002024291	0
VWA	17	v1_17_S2	TAGA TGGA [TAGA]12 [CAGA]4 TAGA		0.218623482	0.195348837
VWA	17	v1_17_S3	TAGA TGGA [TAGA]13 [CAGA]3 TAGA		0.026315789	0.030232558
VWA	17	v1_17_S4	TAGA TGGA [TAGA]13 [CAGA]4		0.002024291	0.002325581
VWA	18	v1_18_S1	TAGA TGGA [TAGA]11 [CAGA]6 TAGA		0.006072874	0
VWA	18	v1_18_S2	TAGA TGGA [TAGA]12 [CAGA]5 TAGA		0.002024291	0
VWA	18	v1_18_S3	TAGA TGGA [TAGA]13 [CAGA]4 TAGA		0.168016194	0.174418605
VWA	18	v1_18_S6	TAGA TGGA [TAGA]14 [CAGA]3 TAGA		0.010121457	0.011627907
VWA	18	v1_18_S8	TAGA TGGA [TAGA]14 [CAGA]4		0.002024291	0

VWA	19	v1_19_S1	TAGA TGGA [TAGA]12 [CAGA]6 TAGA		0.016194332	0.009302326
VWA	19	v1_19_S2	TAGA TGGA [TAGA]13 [CAGA]5 TAGA		0.010121457	0.002325581
VWA	19	v1_19_S3	TAGA TGGA [TAGA]14 [CAGA]4 TAGA		0.038461538	0.048837209
VWA	19	v1_19_S5	TAGA TGGA [TAGA]14 [CAGA]4 TAGT		0.002024291	0
VWA	20	v1_20_S1	TAGA TGGA [TAGA]13 [CAGA]6 TAGA		0.002024291	0.002325581
VWA	20	v1_20_S2	TAGA TGGA [TAGA]14 [CAGA]5 TAGA		0.002024291	0.006976744
VWA	20	v1_20_S3	TAGA TGGA [TAGA]15 [CAGA]4 TAGA		0.010121457	0.004651163
VWA	21	v1_21_S1	TAGA TGGA [TAGA]14 [CAGA]6 TAGA		0	0.002325581

Appendix 4.4: Novel allele sequences

Table A4.4: Novel allele sequences and their frequencies in two South African population groups

Marker	Allele	Sequence	Admixed (n = 247)	Black African (n = 216)
CSF1PO	12.1	CTTCCTATCTATCTATCTATCTATCTATCTATCTAATCTATCTATCTATCTATCTAATCTATCTATCTT	0.002024291	0
D12S391	17	CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGTAGATAGATAGATAGATAGATAGATAGATAGATAGACAGACAGACAGACAGACAGACAGACAGACAGATGAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	0.002024291	0
D12S391	18	CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGTAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGACAGACAGACAGACAGACAGACGAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	0.002024291	0
D12S391	21	CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGTAGATAGATAGATAGGTAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGACAGACAGACAGACAGACAGACAGACAGATGAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	0.002024291	0
D12S391	23	CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGTAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGACAGACAGACAGACAGACAGACAGACAGACAGATGAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	0.002024291	0
D12S391	24	CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGTAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGATGAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	0.004048583	0.009259259
D12S391	25	CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGTAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGATGAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	0.002024291	0.004629629
D12S391	26	CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGTAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGATGAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	0.004048583	0
D12S391	27	CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGTAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGATGAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	0.002024291	0
D12S391	27	CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGTAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGATGAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	0.002024291	0.002314815
D12S391	22	CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGTAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACGAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACGA	0.002024291	0
D13S317	11	TCTGACCCATCTAACGCCTATCTGTATTTACAAATACATTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCAATCATCTATCTATCTTTCTGTCTTTTTGGG	0.004048583	0.004629629
D13S317	12	TCTGACCCATCTAACGCCTATCTGTATTTACAAATACATTATCTATCTATCTATCTATCTATTTATCTATCTATCTATCTATCAATCATCTATCTATCTTTCTGTCTGTCTTTTTGGG	0.002024291	0

D13S317	12	TCTGACCCATCTAACGCCTATCTGTATTTACAAATACATTATCTATCT ATCTATCTATCTATCTATCTATCTATCTATCTATCTATCAATCAT CTATCTATCTTTCTGTCTTTTTGGG	0.002024291	0
D13S317	15	TCTGACCCATCTAACGCCTATCTGTATTTACAAATACATTATCTATCT ATCTATCTATCTGTCTATCTATCTATCTATCTATCTATCTATCTA TCAATCATCTATCTATCTTTCTGTCTGTCTTTTTGGG	0.002024291	0.0023 14815
D16S539	13	TCCTCTTCCCTAGATCAATACAGACAGACAGACAGGTGGATAGGTAG ATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATATCA TTGAAAGACAAAACAGAGATGGATGATAGATAC	0.002024291	0
D16S539	13	TCCTCTTCCCTAGATCAATACAGACAGACAGACAGGTGGATAGATAG ATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATATCA TTGAAAGACAAAACAGAGATGGATGATAGTTAC	0.002024291	0
D17S1301	14	ATATGTGTGAGATAGATAGATAGATAGATAGATAGATAGATAGATA GATAGATAGATAGATCGATCCATCATAGGAATTTT	0.004048583	0.0023 14815
D17S1301	12	ATATGTGCGAGATAGATAGATAGATAGATAGATAGATAGATAGATA GATAGATAGATCCATCATAGGAATTTT	0.002024291	0
D18S51	15	GTCTCAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAA AGAAAGAAAGAAGGAAAGAAAAAGAGAGAGGAAAGAAAGAGAAA AAGAAAAGAAATAGTAGCAACTGTTATTGTAAGA	0.002024291	0
D18S51	16	GTCTCAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAA AGAAAGAAAGAAAGAAGGAAAGAAAAAGAGAGAGGAAAGAAAGA GAAAAAGAAAAGAAATAGTAGCAACTGTTATTGTAAGA	0.002024291	0
D18S51	17	GTCTCAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAA AGAAAGAAAGAAAGAAAGAAGGAAAGAAAAAGAGAGAGGAAAGA AAGAGAAAAAGAAAAGAAATAGTAGCAACTGTTATTGTAAGA	0	0.0092 59259
D18S51	22.2	GTCTCAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAA AGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAA AGAGAAAAAGAGAGAGGAAAGAAAGAGAAAAAGAAAAGAAATAGT AGCAACTGTTATTGTAAGA	0.002024291	0
D18S51	14	GTCTCAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAA AGAAAGAAGGAAAGAAAAAGAGAGAGGAAAGAAAGAGAAAAAGA AAAGAAATAGTAGCAACTGTTATTGTAAGA	0.002024291	0
D19S433	7	AATAAAAATCTTCTCTCTTTCTTCCTCTCTCCTTCCTTCCTTCCTTCCT TCCTACCTTCTTTCCTTCAACAGAATCTTATTCTGTTGCCCAGGCTGG AGTGCAGTGGTACAATTATAGCT	0.026315789	0.0138 88889
D19S433	7	AATAAAAATATTCTCTCTTTCTTCCTCTCTCCTTCCTTCCTTCCTTCCT TCCTACCTTCTTTCCTTCAACAGAATCTTATTCTGTTGCCCAGGCTGG AGTGCAGTGGTACAATTATAGCT	0	0.0023 14815
D19S433	13	AATAAAAATCTTCTCTCTTTCTTCCTCTCTCCTTCCTTCCTTCCTTCCT TCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCTTTCCTTCAACAGAA TCTTATTCTGTTGCCCAGGCTGGAGTGCAGTGGTACAATTATAGCT	0.002024291	0
D19S433	8	AATAAAAATATTCTCTCTTTCTTCCTCTCTCCTTCCTTCCTTCCTTCCT TCCTTCCTACCTTCTTTCCTTCAACAGAATCTTATTCTGTTGCCCAGG CTGGAGTGCAGTGGTACAATTATAGCT	0.002024291	0.0069 44444
D1S1656	13.3	TTCAGAGAAATAGAATCACTAGGGAACCAAATATATATACATACAA TTAAACACACACACACCTATCTATCTATCTATCTATCTATCTATCTAT CTATCTATCTATCTATCATCTA	0	0.0046 2963
D1S1656	15	TTCAGAGAAATAGAATCACTAGGGAACCAAATATATATACCTACAAT TAAACACACACACACCTATCTATCTATCTATCTATCTATCTATCTATC TATCTATCTATCTATCTATCTATCTA	0	0.0023 14815
D1S1656	16.3	TTCAGAGAAATAGAATCACTAGGGAACCAAATATATATACATACAA TTAAACACGCACACACCTATCTATCTATCTATCTATCTATCTATCTAT CTATCTATCTATCTATCATCTATCTATCTATCTA	0.002024291	0
D21S11	29	AAATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATC TATCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTAT CTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCTAT CTATCTATCTATCTATCGTCTATCTAT	0.008097166	0
D21S11	31	AAATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATC TATCTATCTGTCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTAT CTATCTATCATCTATCTATCCATATCTAGCTATCTATCTATCTATCTAT CTATCTATCTATCTATCTATCTATCGTCTATCTAT	0.002024291	0
D21S11	31	AAATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATC TATCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTAT CTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTAT CTATCTATCTATCTATCTATCTATCGTCTATCTAT	0.008097166	0.0046 2963
D21S11	32	AAATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATC TATCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTAT CTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTAT CTATCTATCTATCTATCTATCTATCGTCTATCTAT	0.004048583	0.0023 14815

D21S11	32.2	AAATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATC TATCTGTCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTAT CTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTAT CTATCTATCTATCTATCTATCTATATCTATCGTCTATCTAT	0.004048583	0
D21S11	32.2	AAATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATC TATCTATCTGTCTGTCCGTCTGTCTGTCTGTCTATCTATCTATATCTAT CTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTAT CTATCTATCTATCTATCTATCTATATCTATCGTCTATCTAT	0.002024291	0
D21S11	33	AAATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATC TATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCT GTCTATCTATATCTATCTATCATCTATCTATCCATATCTATCTAT CTATCTATCTATCTATCTATCTATCTATCTATCGTCTATCTAT	0.002024291	0.0023 14815
D21S11	34	AAATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATC TATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCT GTCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTAT CTATCTATCTATCTATCTATCTATCTATCTATCTATCGTCTATCTAT	0.002024291	0.0023 14815
D21S11	34.1	AAATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATC TATCTATCTGTCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTAT CTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTAT CTATCTATCTATCATCTATCTATCTATATCTATCGTCTATCTAT	0.002024291	0.0069 44444
D21S11	35	AAATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATC TATCTATCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTATCT ATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTAT CTATCTATCTATCTATCTATCTATCTATCTATCTATCGTCTATCT AT	0	0.0023 14815
D21S11	35	AAATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATC TATCTATCTATCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCT GTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATAT CTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCGTCTATCT AT	0	0.0023 14815
D21S11	35.1	AAATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATC TATCTATCTGTCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTAT CTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTAT CTATCTATCTATCTATCATCTATCTATCTATCTATATCTATCGTCTATC TAT	0.008097166	0
D21S11	36	AAATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATC TATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCT GTCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTAT CTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCGTCT ATCTAT	0.002024291	0
D21S11	36	AAATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATC TATCTATCTATCTATCTATCTATCTATCTATCTGTCTGTCTGTCT GTCTGTCTGTCTATCTATATCTATCTATCTATCATCTATCTATCCATAT CTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCGTCT ATCTAT	0	0.0023 14815
D21S11	36	AAATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATC TATCTATCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCT GTCTGTCTGTCTATCTATATCTATCTATCTATCATCTATCTATCCATAT CTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCGTCT ATCTAT	0	0.0023 14815
D21S11	36	AAATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATC TATCTATCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCT GTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATAT CTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCGTCT ATCTAT	0	0.0023 14815
D21S11	36	AAATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATC TATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTGTCT GTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTAT CTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCGTCT ATCTAT	0	0.0023 14815
D21S11	37	AAATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATC TATCTATCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCT GTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATAT CTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATC GTCTATCTAT	0.002024291	0
D21S11	37	AAATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATC TATCTATCTATCTATCTATCTATCTATCTATCTGTCTGTCTGTCC ATATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATC GTCTATCTAT	0.002024291	0

D21S11	37.1	AAATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATC TATCTATCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTAT CTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTAT CTATCTATCTATCTATCTATCTATCATCTATCTATCTATCTATATCTAT CGTCTATCTAT	0.002024291	0
D21S11	37.2	AAATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATC TATCTATCTATCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCT GTCTGTCTGTCTGTCTGTCTGTCTATCTATCTATCATCTATCTATCCAT ATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTA TCGTCTATCTAT	0.002024291	0
D21S11	39	AAATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATC TATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCT GTCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATC ATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCTATCTATC TATCTATCGTCTATCTAT	0.002024291	0
D21S11	39	AAATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATC TATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTGTCT GTCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATC ATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCTATCTATC TATCTATCGTCTATCTAT	0	0.0023 14815
D2S1338	14	GAGGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGG GCAGGCAGGCAGGCAAGGCCAAGCCATTT	0	0.0023 14815
D2S1338	15	GAGGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGG GAAGGAAGGCAGGCAGGCAAGGCCAAGCCATTT	0	0.0023 14815
D2S1338	16	GAGGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGG GAAGGAAGGAAGGCAGGCAGGCAAGGCCAAGCCATTT	0.004048583	0
D2S1338	23	GAGGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGG GAAGGAAGGAAGGAAGGAAGGAAGGCAGGCAGGCAGGCAGG CAGGCAAGGCCAAGCCATTT	0.004048583	0.0023 14815
D2S1338	24	GAGGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGG GAAGGAAGGAAGGAAGGAAGGAAGGCAGGCAGGCAGGCAGG CAGGCAGGCAAGGCCAAGCCATTT	0.004048583	0
D2S441	10.3	CCAGGAACTGTGGCTCATCTATGAAAACTTCTATCTATCTATCATCTA TCTATCTATCTATCTATCTATCTATATCATAACACCACAGCCACTTA	0	0.0046 2963
D2S441	12	CCAGGAACTGTGGCTCATCTATGAAAACTTCTATCTATCTATCTATCT ATCTATCTATCTATCTATCTATCTATATCATAACACCACAGTCA CTTA	0	0.0023 14815
D2S441	14	CCAGGAACTGTGGCTCATCTATGAAAACTTCTATCTATCTATCTATCT ATCTATCTATCTATCTATCTATTTATCTATCTATATCATAATACC ACAGCCACTTA	0	0.0023 14815
D4S2408	11	CTATGCATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATC TAATGTTTA	0.002024291	0
D6S1043	13	AGATGGCATATTGTGAAATTTCTCAGCTTCCATAATTGTATGAGCCA CTTCCCATAATAAATCCTATCTATGTATCTATCTATCTATCTATCTAT CTATCTATCTATCTATCTATCTGATCTATCAATCTATTGATC	0.002024291	0
D6S1043	13.3	AGATGGCATATTGTGAAATTTCTCAGCTTCCATAATTGTATGAGCCA CTTCCCATAATAAATCCTATCTATCTATCTATCTATCTATCTATC TATCTATCATCTATCTATCTATCTGATCTATCAATCTATTGATC	0.002024291	0
D6S1043	16	AGATGGCATATTGTGAAATTTCTCAGCTTCCATAATTGTATGAGCCA CTTCCCATAATAAATCCTATCTATCTATCTATCTATCTATGTATCTAT CTATCTATCTATCTATCTATCTATGTATCTATCTGATCTATCAATCTAT TGATC	0	0.0023 14815
D6S1043	17	AGATGGCATATTGTGAAATTTCTCAGCTTCCATAATTGTATGAGCCA CTTCCCATAATAAATCCTATCAATCTATCTATCTATCTATGTATCTAT CTATCTATCTATCTATCTATCTATCTATCTATCTATCTGATCTATCAAT CTATTGATC	0	0.0023 14815
D7S820	8	TATTTAGTGAGATAAAAAAAACTATCAATCTGTCTATCTATCTATCTA TCTATCTATCTATCTATCGTTAGTTCGTTCTAAACTAT	0.002024291	0
D7S820	10	TATTTAGTGAGATAAAAAAAAACTATCAATCTGTCTATCTATCTATCT ATCTATCTATCTATCTATCTATCTATCGTTGGTTCGTTCTAAACTAT	0.002024291	0
D7S820	12	TATTTAGTGAGATAAAAAAAAACTATCAATCTATCTATCTATCTATCT ATCTATCTATCTATCTATCTATCTATCTATCGTTAGTTCGTTCTA AACTAT	0.002024291	0
D7S820	15	TATTTAGTGAGATAAAAAAAAACTATCAATCTGTCTATCTATCTATCT ATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCGT TAGTTCGTTCTAAACTAT	0.002024291	0
FGA	16.1	CCAGCAAAAAAGAAAGGAAGAAAGGAAGGAAGGAGAAAGAAAGA AAGAAAAGAAAGAAAGAAAGAAAGAGAAAAAAGAAAGAAAGAAA CTAGCTTGTAAATATGC	0	0.0138 88889

FGA	22	CCAGCAAAAAAGAAAGGAAGAAAGGAGGGAAGGAGAAAGAAAGA AAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGA AAGAGAAAAAAGAAAGAAAGAAACTAGCTTGTAAATATGC	0.002024291	0.0023 14815
FGA	24	CCAGCAAAAAAGAAAGGAAGAAAGGAGGGAAGGAGAAAGAAAGA AAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGA AAGAAAGAAAGAGAAAAAAGAAAGAAAGAAACTAGCTTGTAAATA TGC	0.002024291	0
FGA	25	CCAGCAAAAAAGAAAGGAAGAAAGGAGGGAAGGAGAAAGAAAGA AAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGA AAGAAAGAAAGAAAGAGAAAAAAGAAAGAAAGAAACTAGCTTGTA AATATGC	0.004048583	0
FGA	26	CCAGCAAAAAAGAAAGGAAGAAAGGAGGGAAGGAGAAAGAAAGA AAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGA AAGAAAGAAAGAAAGAAAGAGAAAAAAGAAAGAAAGAAACTAGCT TGTAAATATGC	0.002024291	0
FGA	30.2	CCAGCAAAAAAGAAAGGAAGAAAGGAAGGAAGGAAGGAAGGAGA AAGAAAGAAAGGAAGGAAGGAAGAAAGAAAGAAAGAAAGAAAGAAAGA AAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAAAAGAA AGAAAGAAAGAAACTAGCTTGTAAATATGC	0	0.0023 14815
FGA	43.2	CCAGCAAAAAAGAAAGGAAGAAAGGAAGGAAGGAAGGAAGGAGA AAGAAAGAAAGGAAGGAAGGAAGAAAGAAAGAAAGAAAGAAAGAAAGA AAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGACAGACAGA CAGACAGACAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGA AAAAAGAAAGAAAGAAAGAAACTAGCTTGTAAATATGC	0.002024291	0
FGA	24	CCAGCAAAACAGAAAGGAAGAAAGGAAGGAAGGAGAAAGAAAGAA AGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAA AGAAAGAAAGAGAAAAAAGAAAGAAAGAAACTAGCTTGTAAATAT GC	0.002024291	0.0023 14815
FGA	26	CCAGCAAAACAGAAAGGAAGAAAGGAAGGAAGGAGAAAGAAAGAA AGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAA AGAAAGAAAGAAAGAAAGAGAAAAAAGAAAGAAAGAAACTAGCTT GTAAATATGC	0	0.0092 59259

Appendix 4.5: Sunburst chart showing novel alleles across STR markers

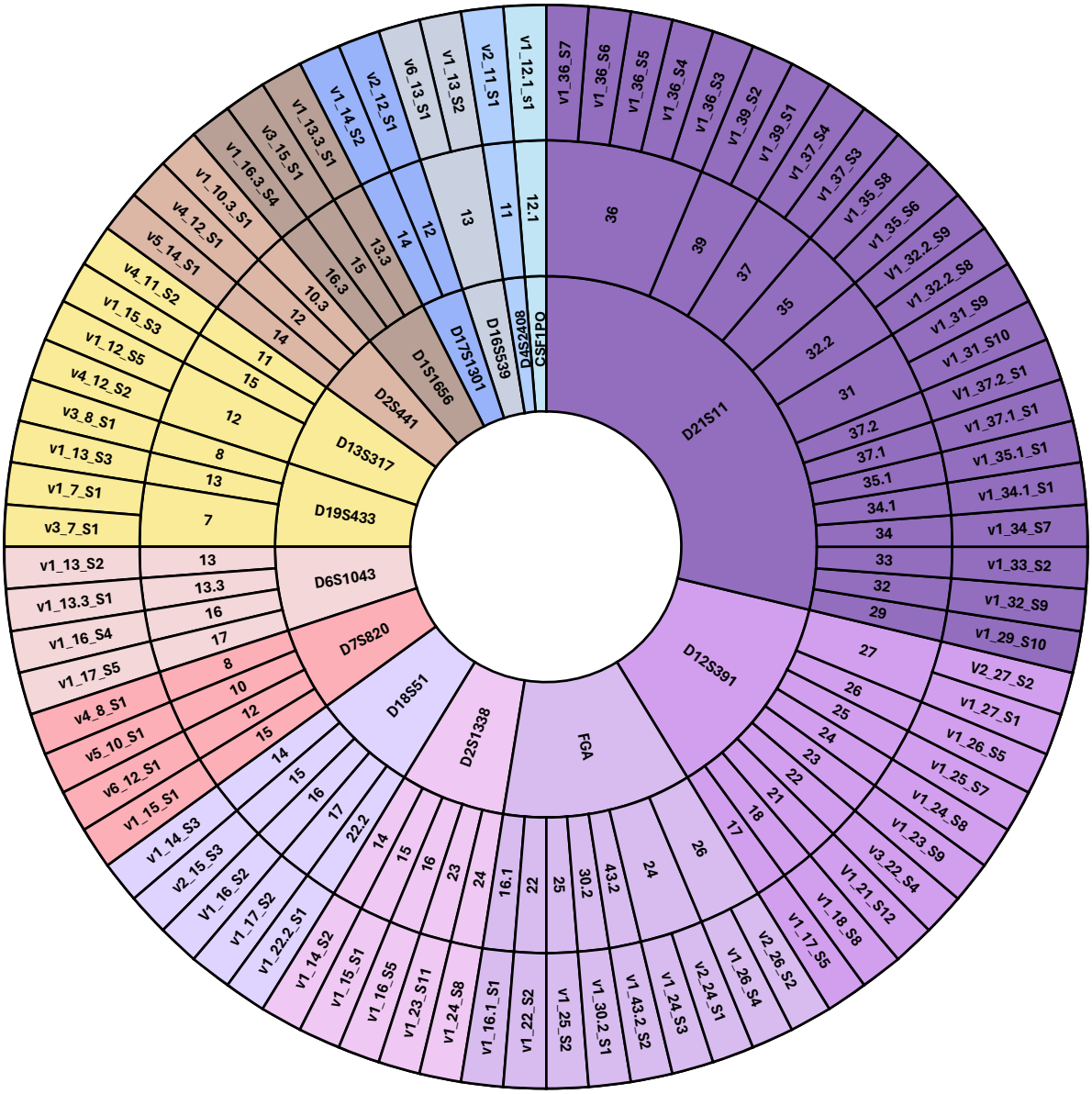

Figure A4.5: Sunburst chart showing 15 different A-STR markers where novel alleles were found for both the Admixed and Black African population groups. The inner circle represents the A-STR marker, the middle-most circle represents the length-based allele, whilst the outer circle represents the shorthand sequence annotation that differentiates between repeat region and flanking region variants.

Appendix 4.6: Population parameters

Table A4.6: Population statistics for the Admixed and Black African population groups. A p-value of 0.002 was used after applying Bonferroni correction

Marker	Total number of alleles		Effective number of alleles		Genetic Diversity (GD)		Observed Heterozygosity		Expected Heterozygosity		p-value (HW)	
	Black African	Admixed	Black African	Admixed	Black African	Admixed	Black African	Admixed	Black African	Admixed	Black African	Admixed
CSF1PO	432	494	10	11	0.774630489	0.754112227	0.78241	0.76518	0.77463	0.75411	0.93928	0.62585
D10S1248	432	494	8	8	0.79464209	0.78476813	0.75	0.77328	0.79464	0.78477	0.22461	0.41238
D12S391	432	494	50	66	0.940201512	0.941603502	0.92593	0.93927	0.9402	0.9416	0.81634	0.51332
D13S317	432	494	22	25	0.85264673	0.890187319	0.86111	0.89069	0.85265	0.89019	0.39498	0.5029
D16S539	432	494	14	15	0.821377932	0.842466597	0.83333	0.84615	0.82138	0.84247	0.22278	0.72877
D17S1301	432	494	10	12	0.641703188	0.696323427	0.75463	0.76113	0.6417	0.69632	0.0395	0.06778
D18S51	432	494	21	29	0.88782547	0.900542822	0.87037	0.88259	0.88783	0.90054	0.10524	0.11989
D19S433	432	494	18	19	0.818080261	0.847566334	0.79167	0.83401	0.81808	0.84757	0.11006	0.06855
D1S1656	432	494	26	29	0.903841196	0.909264111	0.89352	0.92713	0.90384	0.90926	0.31367	0.50631
D20S482	432	494	14	14	0.718613045	0.752724376	0.74074	0.79352	0.71861	0.75272	0.56827	0.47432
D21S11	432	494	55	69	0.894893443	0.922387104	0.85648	0.91498	0.89489	0.92239	0.22686	0.78701
D2S1338	432	494	48	58	0.956324654	0.955293132	0.95833	0.95951	0.95632	0.95529	0.34718	0.95833
D2S441	432	494	21	24	0.774028959	0.821106832	0.74537	0.82996	0.77403	0.82111	0.02976	0.6023
D3S1358	432	494	16	18	0.888233651	0.88773189	0.82407	0.88664	0.88823	0.88773	0.13138	0.10015
D4S2408	432	494	7	8	0.76939933	0.807096928	0.75463	0.78947	0.7694	0.8071	0.10273	0.37041
D5S818	432	494	13	15	0.850530635	0.844380025	0.86574	0.82591	0.85053	0.84438	0.13476	0.19076
D6S1043	432	494	20	22	0.889812666	0.868219855	0.89815	0.89879	0.88981	0.86822	0.49245	0.45941
D7S820	432	494	18	22	0.830433101	0.874641746	0.85648	0.86235	0.83043	0.87464	0.26513	0.92038
D8S1179	432	494	20	27	0.888179943	0.914618423	0.89352	0.91498	0.88818	0.91462	0.91615	0.05793
D9S1122	432	494	15	16	0.788283063	0.824334201	0.78704	0.8502	0.78828	0.82433	0.85732	0.83215
FGA	432	494	29	30	0.893131821	0.89169014	0.87963	0.8583	0.89313	0.89169	0.19086	0.00608
TH01	432	494	7	7	0.740257369	0.796766061	0.71759	0.82186	0.74026	0.79677	0.46596	0.1496
TPOX	432	494	7	7	0.776456561	0.708050357	0.82407	0.73684	0.77646	0.70805	0.10342	0.617
VWA	432	494	24	28	0.872412858	0.86354715	0.86111	0.87854	0.87992	0.85372	0.3359	0.50928

Appendix 4.7: Forensic parameters

Table A4.7: Forensic parameters for 24 A-STR markers

Marker	Polymorphic Information Content				Match Probability				Power of Discrimination			
	Black		Coloured		Black		Coloured		Black		Coloured	
	Length-based	Sequence-based	Length-based	Sequence-based	Length-based	Sequence-based	Length-based	Sequence-based	Length-based	Sequence-based	Length-based	Sequence-based
CSF1PO	0.73643296	0.73893972	0.71008377	0.71151047	0.08487654	0.0835048	0.10567293	0.10446	0.91512346	0.9164952	0.89432707	0.89554
D10S1248	0.76491339	0.76491339	0.75072589	0.75072589	0.07064472	0.07064472	0.08147978	0.08147978	0.92935528	0.92935528	0.91852022	0.91852022
D12S391	0.83126598	0.93485335	0.86313828	0.93710125	0.04003772	0.01015947	0.02984806	0.00965431	0.95996228	0.98984053	0.97015194	0.99034569
D13S317	0.66166872	0.834464	0.74445673	0.87821599	0.13451646	0.0420096	0.08328279	0.02561917	0.86548354	0.9579904	0.91671721	0.97438083
D16S539	0.74114225	0.80249385	0.76424919	0.82360814	0.08624829	0.05178326	0.0721369	0.04427216	0.91375171	0.94821674	0.9278631	0.95572784
D17S1301	0.57756839	0.58196314	0.64058883	0.64495789	0.22243656	0.21532064	0.16743431	0.16520513	0.77756344	0.78467936	0.83256569	0.83479487
D18S51	0.87016123	0.87489474	0.88456257	0.88996578	0.03039266	0.02773491	0.02453736	0.0222754	0.96960734	0.97226509	0.97546264	0.9777246
D19S433	0.79537577	0.79578535	0.82331275	0.82972059	0.0571845	0.0567987	0.04787818	0.04535396	0.9428155	0.9432013	0.95212182	0.95464604
D1S1656	0.84702676	0.89437553	0.86467342	0.9006651	0.03725137	0.020619	0.03112656	0.01981675	0.96274863	0.979381	0.96887344	0.98018325
D20S482	0.60808074	0.68886126	0.66595194	0.72013658	0.17146776	0.11213992	0.13547182	0.10042781	0.82853224	0.88786008	0.86452818	0.89957219
D21S11	0.8409913	0.88799129	0.85547264	0.91598949	0.03849451	0.020619	0.03322461	0.01345703	0.96150549	0.979381	0.96677539	0.98654297
D2S1338	0.86708424	0.95218476	0.88277344	0.9514363	0.02820645	0.00844479	0.02424232	0.00732679	0.97179355	0.99155521	0.97575768	0.99267321
D2S441	0.70254831	0.74042523	0.75064815	0.80056792	0.11265432	0.09203532	0.07774263	0.05518858	0.88734568	0.90796468	0.92225737	0.94481142
D3S1358	0.69239314	0.87545538	0.73163347	0.87587355	0.1144976	0.02640604	0.0989854	0.02686489	0.8855024	0.97359396	0.9010146	0.97313511
D4S2408	0.72978066	0.73140971	0.75395562	0.77678554	0.09066358	0.09057785	0.08092249	0.06925208	0.90933642	0.90942215	0.91907751	0.93074792
D5S818	0.73550774	0.83115406	0.70911012	0.82457908	0.09203532	0.04663923	0.10688587	0.04607517	0.90796468	0.95336077	0.89311413	0.95392483
D6S1043	0.87043072	0.87795403	0.84730653	0.85376555	0.02829218	0.02533436	0.03896146	0.03647003	0.97170782	0.97466564	0.96103854	0.96352997
D7S820	0.73783708	0.81047042	0.78550889	0.86062274	0.09194959	0.05122599	0.06358078	0.0295858	0.90805041	0.94877401	0.93641922	0.9704142
D8S1179	0.74609839	0.87575262	0.8090786	0.90628461	0.08166152	0.02614883	0.05384451	0.01893163	0.91833848	0.97385117	0.94615549	0.98106837
D9S1122	0.66014745	0.7594466	0.65072044	0.80087497	0.12847222	0.07253086	0.14986313	0.05525414	0.87152778	0.92746914	0.85013687	0.94474586
FGA	0.87457988	0.88142385	0.87452094	0.87969136	0.02700617	0.02503429	0.02745497	0.02607812	0.97299383	0.97496571	0.97254503	0.97392188
TH01	0.69471596	0.69471596	0.76318688	0.76318688	0.11659808	0.11659808	0.07915226	0.07915226	0.88340192	0.88340192	0.92084774	0.92084774
TPOX	0.73950138	0.73950138	0.6547658	0.6547658	0.09597908	0.09597908	0.1460932	0.1460932	0.90402092	0.90402092	0.8539068	0.8539068
vWA	0.79660158	0.8574285	0.7838852	0.84765836	0.05847051	0.03307734	0.06971103	0.03922372	0.94152949	0.96692266	0.93028897	0.96077628

Appendix 4.8: Concordance

Table A4.8: Concordance data for 24 autosomal STR markers based on 229 samples for the South African Admixed and Black African population groups. The total number of alleles, concordant alleles, discordant alleles, ambiguous alleles and percentage concordance (%) are shown. The D22S1045 marker shown in red font was excluded from allele frequency calculations but included for concordance assessment. Discordances are shown in grey-shaded blocks

Marker	Alleles (N)	Concordant (N)	Discordant	Discordant due to imbalance or allele dropout	Concordance (%)
CSF1PO	456	456			100.00
D10S1248	456	456			100.00
D12S391	440	440			100.00
D13S317	456	456			100.00
D16S539	*456*	*454*		*2*	99.56
D18S51	454	454			100.00
D19S433	*452*	*451*		*1*	*99.78*
D1S1656	*440*	*439*	*1*		*99.77*
D21S11	442	442			100.00
D22S1045	***408***	***345.00***		***63***	***84.56***
D2S1338	458	458			100.00
D2S441	456	456			100.00
D3S1358	456	456			100.00
D5S818	*442*	*436*		*6*	*98.64*
D7S820	**440**	**440**	**1**	**2**	**99.77**
D8S1179	452	452			100.00
FGA	456	456			100.00
TH01	458	458			100.00
TPOX	459	459			100.00
vWA	454	454			100.00
Total	**8991**	**8987**	**1**	**76**	**99.10**

Appendix 5.1: Average read counts for blank and NTC samples

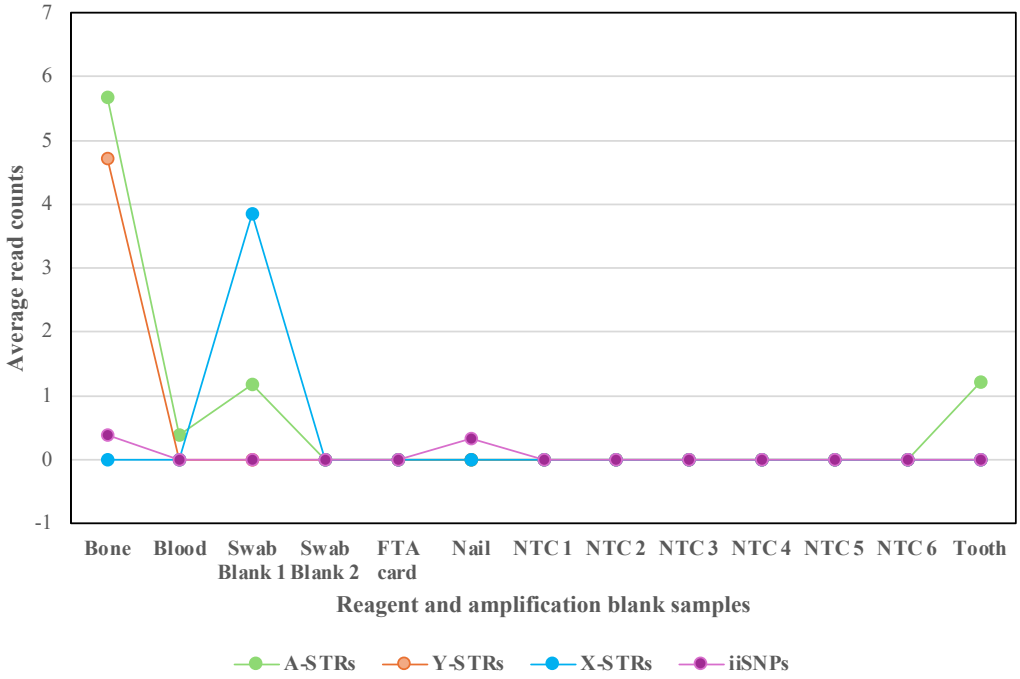

Figure A5.1: Line graph showing average read counts across DPMA markers for NTC and reagent or extraction blank samples.

Analytical threshold calculation

Marker	Average	SD	2 x SD
STRs	0.49	5.58	**11.64**
iiSNVs	0.06	1.22	2.50
aiSNVs	0.41	3.07	6.56
piSNVs	0.55	2.56	5.66

Appendix 5.2: 2800M sensitivity

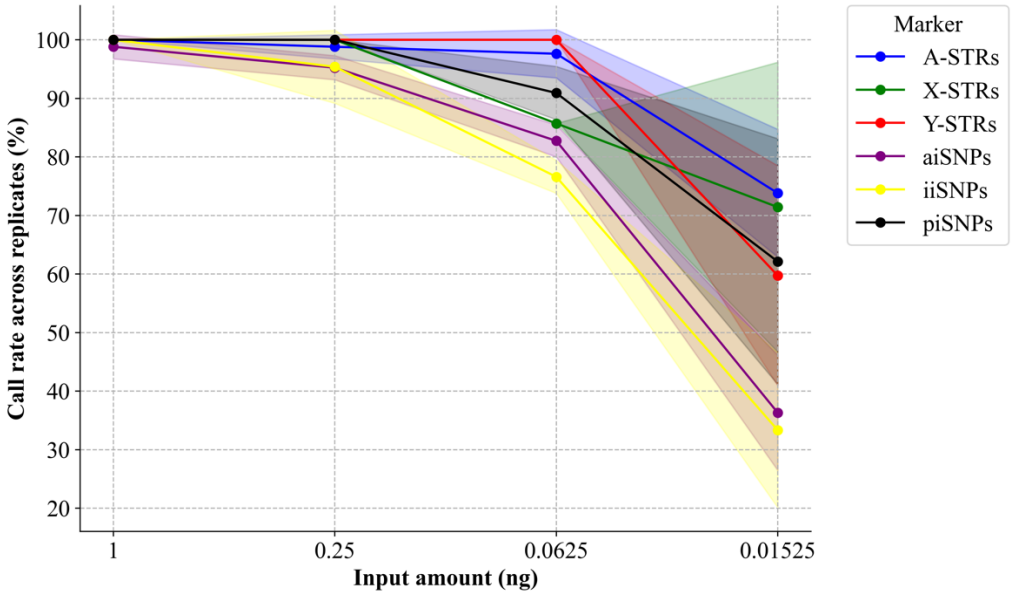

Figure A5.2: *Average call rates (%) of autosomal STRs (blue), X-STRs (green), Y-STRs (red), and iiSNVs (purple) for the 2800M control DNA sample diluted to 1ng, 0.25 ng, 0.0625 ng and 0.01525pg. The shaded regions around each line depict the variation observed across replicate samples.*

Appendix 5.3: SRM 2372a sensitivity

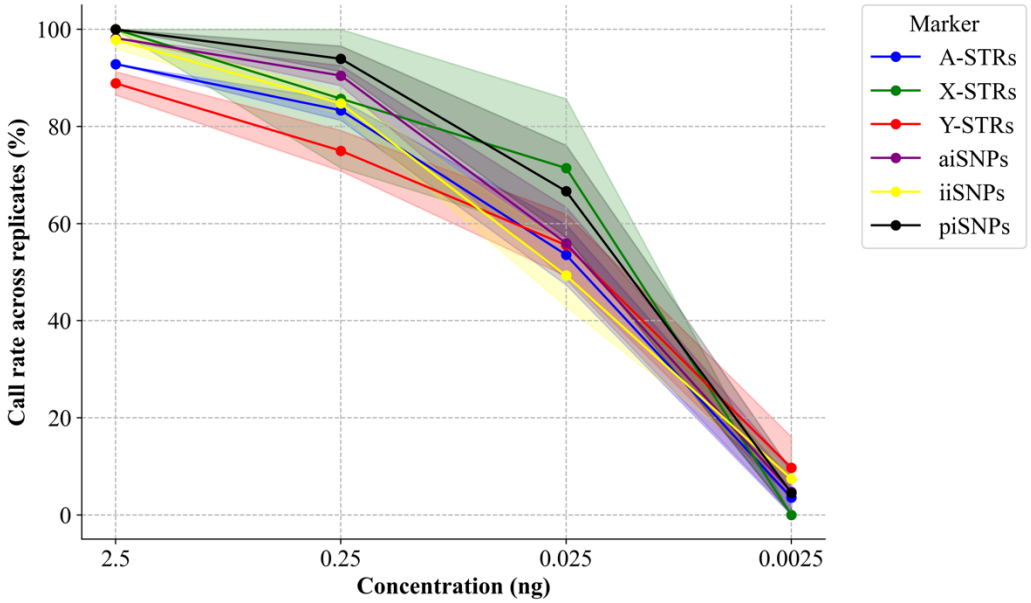

Figure A5.3: *Average call rates (%) of autosomal STRs (blue), X-STRs (green), Y-STRs (red), iiSNVs (yellow), aiSNVs (purple) and piSNVs (black) for the SRM 2372a control DNA*

sample diluted to 2.5 ng, 0.25 ng, 0.025 ng and 0.0025 ng. The shaded regions around each line depict the variation observed across replicate samples.

Appendix 5.4: Blood sensitivity

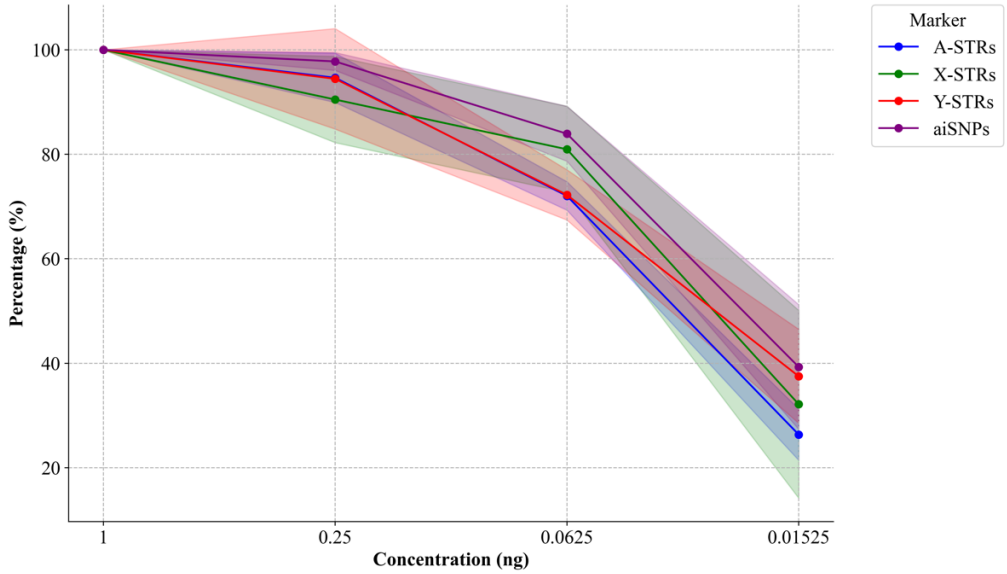

Figure A5.4: Average call rates (%) of autosomal STRs (blue), X-STRs (green), Y-STRs (red), iiSNVs (yellow), aiSNVs (purple) for the authentic forensic blood sample diluted to 1 ng, 0.25 ng, 0.0625 ng and 0.1525 ng. The shaded regions around each line depict the variation observed across replicate samples.

Appendix 5.5: Call rates – degradation study

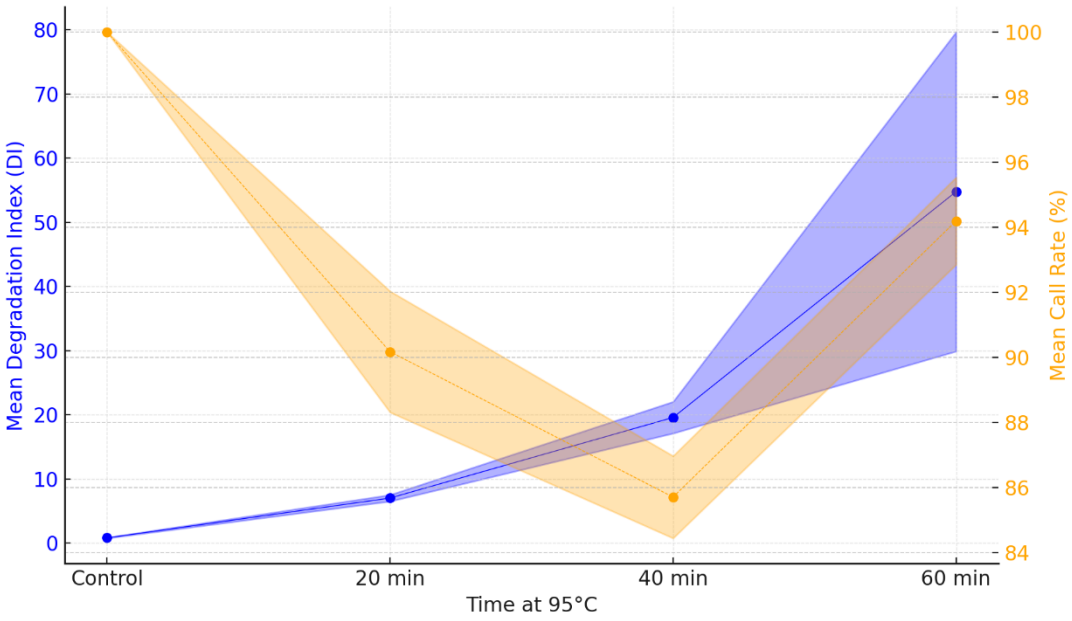

Figure A5.5: Mean call rates (%) across all DMPA markers plotted in comparison with DI values obtained through qPCR experiments. The x-axis shows the time periods at which the control DNA was incubated at 95°C.

Appendix 5.6: qPCR results for inhibition study

Table A5.6: qPCR results for the 2800M control sample spiked with 1% ethanol and 600uM humic acid. The IPC C_T, C_T for the large, small and Y targets are shown, as well as the DI.

	Sample ID	IPC CT	Large		Small		Y		DI
			CT	Concentration	CT	Concentration	CT	Concentration	
2800M 0.2ng/ul	2800M_0.2_1	27.7507114	26.08556938	0.299823135	28.3430271	0.244586796	27.6419353	0.253301531	0.815770268
	2800M_0.2_2	27.6974201	26.07898903	0.301141083	28.3543091	0.242697358	27.5054779	0.277751148	0.805925786
	2800M_0.2_3	27.8548355	25.94379616	0.329537839	27.9647732	0.317213297	27.4545059	0.287477642	0.962600529
	2800M_0.2_4	27.5452099	25.98905182	0.319745779	28.4149742	0.232784986	27.642807	0.25315249	0.728031456
	Average	**27.7120442**	**26.0243516**	**0.312561959**	**28.2692709**	**0.259320609**	**27.5611815**	**0.267920703**	**0.82808201**
	SD	**0.12901171**	**0.069445708**	**0.014520122**	**0.20544244**	**0.03894064**	**0.09603207**	**0.017425382**	**0.097890794**
	% CV	**0.46554383**	**0.266848948**	**4.645517867**	**0.72673413**	**15.01640767**	**0.34843235**	**6.503932692**	**11.82138877**
Ethanol 1%	ET_1.1	27.6054287	31.62330627	0.007478508	34.0635529	0.004794148	32.2623024	0.011185406	0.641056716
	ET_1.2	27.3159485	31.42961884	0.008509089	33.6849251	0.006219299	33.0437584	0.006599014	0.730900645
	ET_1.3	27.6825886	31.44646263	0.008414091	34.0201988	0.004939167	33.0875893	0.006406562	0.587011337
	ET_1.4	27.2552147	31.7028656	0.007092255	33.692131	0.006188569	33.1793098	0.006021806	0.872581363
	Average	**27.4647951**	**31.55056334**	**0.007873486**	**33.865202**	**0.005535296**	**32.89324**	**0.007553197**	**0.707887515**
	SD	**0.2107851**	**0.134104491**	**0.00069823**	**0.2047929**	**0.000774445**	**0.42439983**	**0.002433334**	**0.124807951**
	% CV	**0.76747376**	**0.425046266**	**8.868116974**	**0.60472961**	**13.99103545**	**1.29023418**	**32.21594299**	**17.63104277**
Humic acid 600ng/ul	Humic acid_600_1	29.2566509	Undetermined	Undetermined	33.7232056	0.006057783	34.9024925	0.001880961	Undetermined
	Humic acid_600_2	27.7846794	32.15677643	0.005240676	34.2239838	0.004293572	32.9744186	0.006915345	0.81927824
	Humic acid_600_3	27.5758495	32.2831955	0.004817169	34.0333252	0.004894802	33.5695076	0.004626961	1.016116023
	Humic acid_600_4	27.6145573	31.85386086	0.006413197	33.7693863	0.005868507	32.7440491	0.008079285	0.915067255
	Average	**28.0579343**			**33.9374752**	**0.005278666**	**33.547617**	**0.005375638**	
	SD	**0.8042758**			**0.23483309**	**0.000831203**	**0.96789198**	**0.002735793**	
	% CV	**2.86648259**			**0.69195804**	**15.74646388**	**2.88512886**	**50.89243307**	