Characterisation and differentiation of five UK populations using massively parallel sequencing of forensic STRs

Devesse, Laurence

Awarding institution:
King's College London

Characterisation and differentiation of five UK populations using massively parallel sequencing of forensic STRs

Laurence Alicia Elisabeth Devesse

This thesis is submitted in partial fulfilment for the degree of

Doctor of Philosophy (PhD)

at King's College London

Primary Supervisor: Dr. David Ballard

Secondary Supervisor: Professor Denise Syndercombe Court

February 2022

Cette thèse est dediée à mes parents, qui ont toujours eu confience que je réaliserais de grandes choses. Papa, j'aurais tellement aimé que tu puisse la voir terminée.

This thesis is dedicated to my mum and dad, who always believed I would achieve great things. Dad, I wish you could have seen it finished.

Abstract

The transition from capillary electrophoresis (CE) to massively parallel sequencing (MPS) in forensics presents an opportunity to review the choice of genetic markers used for identification and assess the ways in which we utilise them. In relation to short tandem repeat (STR) analysis, the move to assign alleles using sequence rather than length-based methodologies has highlighted the extent to which previous allelic variation was masked. In this work, 1000 samples from five UK-representative populations (White British, West African, North East African, East Asian and South Asian) were typed using the ForenSeq™ DNA Signature Prep kit and MiSeq FGx™ Forensic Genomics System. This thesis addresses some of the key questions associated with the characterisation of novel sequence variants, such as back-compatibility with CE results, power of discrimination and nomenclature. A concordance rate of over 99% was obtained when comparing results of the ForenSeq DNA Signature Prep kit with CE, making it highly compatible with current DNA databases. The increased power of discrimination when taking sequence-level variation into account was substantial, with an overall random match probability for the loci studied that was over 750 times lower than with length-based data alone. The added value of analysing flanking regions of STRs was found to be limited, although their inclusion in analysis is vital for accurate allele calling.

The data from this PhD contributed 214 novel sequences to a larger project cataloguing autosomal STR variation. The large number of variants characterised at select markers brings into question the strategies for producing representative population data, yet also provides an opportunity to use this diversity in unique ways. The presence of population-specific sequence variation in particular raises the prospect of using STR profiles for population identification, both on their own and in combination with ancestry-informative single nucleotide polymorphisms (SNPs). STRs have largely been discounted for geographic ancestry determination due to their high mutation rate, which in turn makes them well suited for individual identification. Being able to obtain a DNA profile that can simultaneously be used for geographical ancestry estimation and searching against offender databases would be a huge benefit to the field of forensic identification in terms of time, cost, and sample availability. Across the five populations studied, good differentiation was achieved using sequenced STR profiles – results which also showed a clear improvement over length-based data.

Table of Contents

List of Figures

List of Tables

Acknowledgments

It’s apparently a PhD student’s prerogative as to how long to make their acknowledgements section, and if not for careful re-writing, it could have been very long indeed. It’s also the first section I started writing, on a bus, on a casual Tuesday night, in July 2019. I wrote that it takes a village to raise a PhD student. My family may never fully understand what is in this thesis, but it’s thanks to their endless love, support and encouragement that I was able to follow this project through to the end. Thank you for always believing in me.

I am eternally thankful to my supervisors David and Denise, for their unfailing guidance and mentorship. And for agreeing to take me on in the first place. David, our pub meetings may not have contributed quite as much to the field of science as that of Watson and Crick, but I hope you know how much they meant to me. Many were the times where you patiently waited for me to reach the right conclusion, or talked me out of throwing my hands up and giving up, and into another round. All of that, all around the world. Thanks to the King’s Forensics PhD students over the years who often made the moments of madness seem more normal. To the DNA team in particular, for all the adventures – in the words of Harry Potter, “mischief managed!”. Special thanks to Lucinda and Ella for always being there, through the laughter, the tears, and copious volumes of prosecco. Much of this thesis was written in that little house by the seaside, and my confidence came from the people writing next to me.

Thanks to everyone at Verogen who tried to accomodate my dual personality/ part time role. To Nicola, Cydne and Kathy who pushed to make the collaboration a possibility when everyone said it couldn’t be done. To Melissa, thank you for being a good friend as well as an excellent manager. I would like to thank the Royal Commission for the Exhibition of 1851, without whom I would not have been able to complete my PhD. Through my Industrial Fellowship I was also able to travel the world and attend conferences that inspired me. I wish I could thank every forensic science rockstar I met and befriended along the way, including the team at NIST who hosted me, but at least they are all referenced at one point or another in the coming pages.

Nigel deserves his own acknowledgement because let’s be honest, no one else would have been willing to sit on my desk for all those hours, full of moral support, fluff and a bit of judgement.

Finally, there aren’t enough words to thank Dave, who asked me if I was sure when I first wondered out loud about doing a PhD. Thank you for never asking again. Thank you for believing I could do this, even when I was pretty sure I couldn’t. Thank you for being my rock, as well as my bio-statistician extraordinaire.

It takes a village to raise a PhD student, but it would take more than a thesis to thank them.

1. INTRODUCTION

1.1. DNA typing in forensic science

In a letter to her father, Rosalind Franklin wrote: “Science, for me, gives a partial explanation for life. In so far as it goes, it is based on fact, experience and experiment.” A decade after writing this, she was to play an instrumental role in elucidating the structure of DNA, the molecule of life. Understanding the structure of DNA led to advances in numerous fields of science, from evolution to disease. In forensic science, DNA can be used as a calling card of sorts for the individual it belongs to, and can be used to identify a victim, perpetrator or missing person, as well as resolve relationships between individuals and provide investigative insights. The technology for performing DNA typing has come a long way since its inception in the 1980’s [1], with technological advances pushing the boundaries of what was believed possible, whilst always relying on fact, experience and experiment.

1.1.1. DNA and individual identification

Deoxyribonucleic acid (DNA) is a molecule which is often referred to as the blueprint of all cells, which contains the genetic instructions for the development, functioning, growth and reproduction of all known living organisms. This molecule is present in virtually all cells in the body and is composed of two chains of nucleotide bases that coil around each other, forming what is known as the “double helix”. The nucleotide bases are Adenine (A), Thymine (T), Guanine (G) and Cytosine (C). In humans, the DNA molecule is composed of approximately 6 billion pairs of these bases, split into 23 pairs of chromosomes within the nucleus of every cell in our body, with one from each pair in a person being inherited from the mother and the other from the father. One of the pairs of chromosomes code for the genetic sex of an individual: the X and Y chromosomes, while the other 22 pairs form the autosomal genetic make-up of this person. Any two unrelated humans of the same sex will share approximately 99.9% of their DNA composition [2].

Identifying one individual from another is a core concept during forensic investigations – investigators must be able to accurately recognise the persons involved in a crime for example. Methods for distinguishing individuals range in means, accuracy, cost, and level of sophistication. Eye-witness testimony, when available, can be an extremely fast way of getting an initial description of a suspect fleeing a scene, but it relies on human eyesight and recall, and is often influenced by bias, making it frequently unreliable [3, 4]. More dependably, the evidence deposited at a crime scene can be used to make an identification. This evidence normally falls under two broad categories: marks and traces (e.g., fingermarks, footprints etc) and biological material (saliva, blood, semen etc).

In forensic science, DNA profiling, or typing, is most commonly used to match biological materials such as blood, semen or saliva to the individual they came from. The principle relies on the isolation of DNA from such a matrix, amplification of specific target areas which are known to vary between individuals, and visualisation of these areas. As much of the genome is identical between individuals, highly variable DNA targets must be used for DNA profiling. Once a DNA "profile" is obtained, it can be compared to another profile, or searched against a database of profiles in order to obtain a match.

1.1.2. DNA fingerprinting

On the 31st of July 1986, 15-year-old Dawn Ashworth went missing from her home in Enderby, England. Two days later, her body was found on a footpath, she had been beaten, raped and murdered. The similarity to another case in the area led the local police to believe they had a serial killer on their hands, and local outcry led to resources being poured into the investigation to catch the killer. A new technique was used in this case for the first time, "DNA fingerprinting". This method relied on comparing genetic signatures between two individuals, to identify similarities or differences, but it had yet to be used for criminal investigations. The police wanted confirmation that Richard Buckland, a local youth, had committed both crimes, given he had the same blood group as the sample found at the crime scenes, but he was exonerated when his "DNA fingerprint" did not match that of the semen found at the crime scenes. Later, the detectives working on the case came to decide that the technology that had exonerated Buckland should be used to catch the killer. They began a mass screening effort, in an attempt to collect blood samples from all males in the area of a certain age bracket. Suddenly, DNA testing was at the forefront of the national and international news. In August 1987, a man confessed to providing a sample on behalf of his colleague, Colin Pitchfork. Pitchfork was subsequently arrested and confessed to both crimes and was sentenced to life in prison, serving 33 years before his release in 2021 [5]. This case marked a key milestone in the use of DNA evidence for forensic investigations, and, to date, it is estimated that more than 125 million samples have been uploaded to DNA databases across 60 countries in the context of criminal investigations [6, 7].

The method of DNA fingerprinting was first developed by Dr. Alec Jeffries in 1985, when he found that certain areas of DNA were repeated next to each other, over and over again [8]. He advanced the idea that the number of repeated sequences might differ between individuals, and that they could therefore be used to distinguish people. These regions Jeffries described became known as variable number of tandem repeats (VNTRs), and the technique used to visualise them called restriction fragment length polymorphism (RFLP), because it utilised a restriction enzyme to cut the areas directly surrounding the VNTRs. This technique for DNA Fingerprinting revolutionised DNA typing, as it was significantly more discriminatory than ABO blood group typing (where there are only four possible phenotypes, and 40% of the population has blood group "O"). It did, however, require lengthy processing time and expertise to interpret results, making it poorly suited for large scale and rapid DNA identification. It also required large amounts of intact DNA, making it inappropriate for a wide variety of forensic-type samples which have often been exposed to the elements, causing the DNA to degrade and fragment.

1.1.3. STR profiling

The compromise for speed of analysis and power of discrimination quickly came in the form of short tandem repeat (STR) markers, shorter versions of the VNTRs first described in 1985. Typical VNTR probes and RFLP markers used in the late 1980's had repeat unit lengths of 9 (D1S7) to 38 (D17S79) base pairs, whereas STRs repeat unit length vary from 2 to 7 base pairs [9]. These markers were easier to amplify using a technique called polymerase chain reaction (PCR), which provides the capability to copy and label a specific DNA sequence in order to make it easier to detect.

For the last 20 years, forensic DNA typing has relied almost exclusively on the targeting of autosomal STRs (aSTRs) [10, 11]. STRs are ideal for human identification due to their highly variable nature and therefore high power of discrimination [12]. At each STR location, or locus, the number of repeats of a core DNA motif varies between individuals, and is known as an allele. In brief, STR-DNA profiling relies on successful extraction of DNA from a biological matrix, followed by amplification of multiple STR sequences (often referred to as a “multiplex”) and their subsequent separation and visualisation. The number of repeats, or alleles, for each STR locus detected can be determined, and a profile obtained (Figure 1.1). An STR-DNA profile can then be compared to other profiles. If the alleles in two profiles are different, the samples from which the profiles have been obtained can be excluded as coming from the same individual.

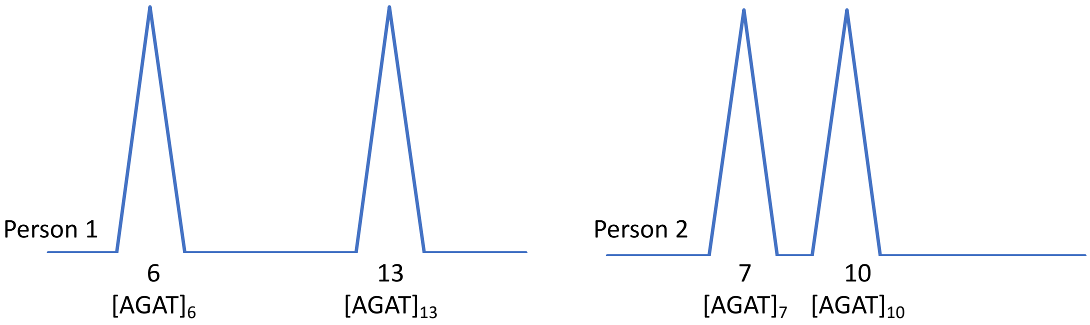


Figure 1.1: Theoretical example of STR profiles

The repeat structure in this example is $[AGAT]_n$*, and the figure shows that person 1 has a 6, 13 profile, whereas person 2 has a 7, 10 profile at this locus.*

Separation of PCR products can be performed using slab gel or capillary electrophoresis. The concept of electrophoresis relies on the negative charge carried by molecules of DNA, which under the influence of an electric field, will migrate through a matrix at different speeds depending on molecular weight, thus separating amplified products based on their size. Gels are rarely used in forensic laboratories these days due to the tedious and sometimes hazardous process for preparing and loading gels. The advent of capillary electrophoresis (CE) brought with it a number of advantages, including requirement for less input material, faster run time, easier laboratory manipulation and the fact that less sample needs to be injected, meaning the PCR product could be re-tested if needed [13].

Silver staining of polyacrylamide gels was sometimes used to visualise STR amplicons in laboratories looking to reduce costs, as it only requires a gel box for electrophoresis, silver nitrate and other developing chemicals [14]. Despite the low cost and ease of use, silver nitrate quickly went out of favour due to the complexities in interpretation and single colour (meaning PCR product size was the only method for differentiating alleles). The latter issue was somewhat resolved by switching to fluorescent detection, which enabled the labelling of potentially overlapping PCR products with different coloured fluorescent dyes [15, 16]. The dye is normally attached to a PCR primer that is incorporated into the amplified target region, which can then be measured by exciting the dye molecule and detecting the emitted light – visualised as bands on a gel, or peaks on an electropherogram. Allelic ladders are then used to assign the correct allelic designation to these peaks. An allelic ladder is an artificial mixture of all common alleles present in the human population for a particular set of STR markers, amplified using the same primers as the tested samples in order to provide a reference DNA amplicon size.

By the mid 1990's, DNA testing was entering its second decade and STR typing through PCR amplification, CE separation and fluorescent detection was fast becoming the technique of choice. In April 1995, the UK launched its first National DNA Database (see 1.1.7), leading the way for standardisation of STR marker typing.

1.1.4. Types of STR markers

Traditionally, STRs are described according to variation in repeat unit length: di-, tri-, tetra-, penta-, hexa-nucleotide markers. These markers also vary in the number of repeat units that can be observed (for example, a range of 3-14 repeats for the marker TH01 according to STRBase [17]), and in the rigour with which alleles conform to a specific repeat pattern. Simple repeat loci, such as TH01 are typically composed of repeat motifs of identical length and sequence (e.g. AATG). Compound repeat loci are composed of two or more adjacent simple repeats, such as D12S391 which has a combination of AGAT and AGAC. Finally, complex repeat loci may contain several repeat motifs of variable length, as well as variable intervening sequences. D21S11 is one example of a complex repeat locus and is best visualised in Table 1.1.

Designation of alleles is typically done according to nomenclature rules set out and updated by the DNA commission of the International Society of Forensic Genetics (ISFG) over the years [18]. For complex repeat loci especially, it is expected that there are single base changes at the sequence level which will not affect the allele designation. In the case of insertions or deletions, these will affect the size of the amplicon as separated with CE, but so long as the allele peak corresponds to an allele bin, it will be assigned that allele designation.

Table 1.1: STR repeat categories and example loci, repeat region sequence and designation

STR Repeat category	STR locus	Repeat region sequence	Allele designation
Simple repeat	TH01	$[AATG]_7$	Allele 7
Compound repeat	D12S391	$[AGAT]_{11}$ $[AGAC]_8$	Allele 19
Complex repeat	D21S11	$[TCTA]_6$ $[TCTG]_5$ $[TCTA]_3$ TA $[TCTA]_3$ TCA $[TCTA]_2$ TCCATA $[TCTA]_9$	Allele 28

1.1.5. Commercial STR kits

Most forensic laboratories do not have the time to design PCR primers or optimise multiplexes for targeted amplification of STRs of interest. Ready-made STR kits are popular within the forensic genetics community as they remove this time and resource constraint, whilst also facilitating the sharing of data between laboratories and standardising validation criteria and benchmarks for database uploads. Typical commercial STR kits consist of the following components: 1. A buffer containing deoxynucleotide triphosphates, $MgCl_2$ and other necessary reagents; 2. A PCR primer mix, containing primers designed to amplify and fluorescently tag a specific set of STR loci; 3. A DNA polymerase (sometimes combined with the buffer); 4. An allelic ladder with a mixture of common alleles for the STR loci being amplified and finally; 5. A positive control DNA sample. Discordance between results obtained with different kits can occur due to differences in primer design for the same loci. Rare changes in template sequences chosen as target binding sites for commercial primers may cause the primer to not bind at all, or with reduced affinity. This in turn can lead to what is called a null allele (i.e., an allele that did not amplify and therefore isn’t visible). If different commercial kits amplify the same loci using primers placed in different places, then it is possible that this change in sequence in the template may affect the set of primers in one kit but not in another.

1.1.6. Evaluating a DNA match

If two profiles cannot be differentiated, suggesting that the samples could have originated from the same donor, the strength of such a match must be evaluated. There are various methods for doing this, although most rely on statistical calculations such a Likelihood Ratio (LR) or Random Match Probability (RMP), i.e. the chance that a randomly selected, unrelated individual will have the same combination of alleles at the STRs tested [19]. To calculate this, the frequency of each allele for any given marker within a population must be known. Historically, data from a minimum of 100 individuals per population group was used to generate these frequencies, otherwise known as population databases [20, 21]. If a profile contains a "rare" allele, this will push the strength of a match further, making the results more discriminatory. As a general rule, increasing the number of markers targeted will also increase the power of discrimination of a DNA test. This is one of the reasons that commercially available STR kits have evolved from 10 to 24 markers, with kits targeting 17 markers recommended in England and Wales since 2014 [22].

1.1.7. DNA Databases and Core STR Loci

Forensic DNA databases are an essential part of forensic DNA profiling. Generally, after a sample is collected from a crime scene and a DNA profile obtained, the profile will be searched against a database of known offenders. The first national database to comprise STR DNA data was established in 1995 in the UK, and by the end of 1999 it contained over 700,000 profiles [23]. Six initial STR markers were used to form this database: TH01, vWA, FGA, D8S1179, D18S51 and D21S11. By 2005, there were seven core loci that overlapped between all of the European databases and laboratories, but this still wasn't sufficient to avoid adventitious matches occurring when comparing profiles across borders. As a result of this, the European Network of Forensic Science Institutes (ENFSI) and the European DNA Profiling Group (EDNAP) have worked collaboratively for a number of years to achieve standardisation of DNA profiling throughout Europe [24]. Whilst individual countries and laboratories have sought to increase the discrimination power and improve robustness and sensitivity of their assay of choice, the collaboration of these major networks has additionally caused a shift in emphasis to expand core marker sets in order to increase the efficiency of DNA databases. This in turn led to the incorporation of these markers into commercially available assays, to meet the demand of the field. If a larger number of core markers overlap between different countries, there can be a more efficient cross-border exchange of profile data [25]. Using larger and more discriminatory STR sets can also help for investigative tools such as familial searching [26], when a direct match is not obtained when searching for a profile on a DNA database. In these cases, relatives of the true source of the profile can be searched for in the database. In more recent years, recommendations from multiple working groups have led to the expansion of the European Standard Set [27], and of the CODIS STR markers in the USA [28, 29].

1.1.8. Limitations

Although advances in recent years have led to very sensitive assays for DNA typing, there are a number of limitations associated with current techniques. Primarily, the fact that most STR assays today rely on size-based amplicon separation and fluorescent dye detection has several implications. With all STR-CE kits, the number of markers that can be analysed simultaneously is limited by the fact that within a single dye channel, all amplified products must be distinguishable by size. Each dye channel is limited to a maximum of approximately 500 nucleotides, and the range of alleles for each STR locus must fit alongside the other loci in the same channel [30], with most commercially available assays targeting amplicons that range from 80 to 500 base pairs (bp) [31-33]. One of the most established CE platforms is capable of collecting data from 6 dye channels, and by extension, analysing up to 24 loci [34-36]. The latest innovation in CE technology suggests the possibility of collecting data from 8 dye channels and targeting up to 35 loci [37], but at the time of writing there were no scientific publications demonstrating performance or results for this.

This restriction to the number of loci that can be targeted and accompanying necessity for specific target sizes for accurate size-based separation have a number of implications for forensic type samples. Degraded DNA is commonly encountered in criminal casework, mass disaster victim identification and missing persons cases. Here, the exposure of biological materials to environmental factors, inhibitors, time since deposition etc. may lead any DNA present to become fractured [38]. The more degraded the sample, the less likely it is that recovering large portions of intact DNA will be possible. For STR analysis, higher molecular weight amplicons have an increased chance of not being amplified (also known as “drop out”). If too few loci are recovered, there may be insufficient data to obtain a useable result. Targeting smaller amplicons may improve the chances of obtaining results, but the restriction of amplicon size for CE analysis means that, despite best efforts to move primers closer to the target region, there are still design implications which limit the extent of what is possible [39, 40].

Where in the past, a large amount of material was needed to produce a DNA profile, today a profile can be obtained from minute amounts of sample collected from almost any surface. The fact that DNA typing techniques have evolved to become so sensitive means we can deliberate more complex scenarios, including the analysis of “touch DNA”, where surfaces that have been touched by multipe individuals are considered [41]. This in turn means that the odds of collecting biological samples containing DNA from more than one individual from a crime scene are constantly increasing. These “mixtures” have always been present at crime scenes, but due to lower sensitivity would not necessarily have been detected. As stated in the 2021 National Institute of Standards and Technology (NIST) Scientific Foundation Review on DNA Mixture Interpretation: “From a historical perspective, […] increase in DNA test method sensitivity and willingness to attempt examination of smaller quantities of DNA have resulted in an increase in samples and sample types submitted to forensic laboratories.”[42]. As the number of contributors to a DNA profile increases, the probability of resolving it into individual components fundamentally decreases [43]. Using the largest and most polymorphic set of STRs possible will increase the chance of distinguishing components in a mixture [13, 44]. Here again, the limitation to the number of markers that can be targeted using STR typing and CE means there is sometimes insufficient power of discrimination for certain lines of enquiry. Additionally, if an investigation requires testing of Y or X chromosome STRs, as is often the case for sexual assault cases or for establishing complex relationships, these markers must be analysed using additional testing, often subsequent to autosomal STR analysis. This type of iterative testing requires additional DNA, which may be limited in cases where the amount of sample is limited.

1.1.9. Single nucleotide polymorphisms

A single nucleotide polymorphism (SNP) is another type of marker sometimes used for DNA typing, and can be described as a single nucleotide position where at least two alleles can be found within a population. Here, alleles are defined by the different nucleotide that can be seen at this SNP position. Traditionally, the frequency of the minor allele has to be present at a significant global frequency of over 1%. In 2003, the HapMap Project identified approximately 10 million SNPs that met these criteria in the worldwide population [2]. The Human Genome Project went on to characterize 84.7 million SNPs, and estimated that on average, each person differs from the human reference genome at 4.1-5 million sites, with SNPs and Indels (insertions/ deletions) forming over 99.9% of this variation [45]. These markers have been used for a long time in medical genetics, but their use in forensic DNA typing is becoming more prevalent due to the number of advantages they have over STR-DNA analysis. Typically, given the target area is just one base, SNP amplicons are drastically shorter than STR amplicons (usually <150bp), making them well suited for the analysis of degraded DNA, where genomic fragmentation is likely to have occurred. SNPs are also not prone to stutter, the result of polymerase “slippage” observed when amplifying STR loci. They have a lower mutation rate compared to STRs, which is useful for certain applications such as kinship analysis (where decreased number of mutations over the course of generations will lead to a reduced chance of false exclusion). As a result of this, and their usually bi-allelic nature, SNPs also have significantly less discriminatory power than STRs for individual identification. Gill et al. predicted that a set of 50-100 SNPs would need to be amplified to match the power of discrimination and mixture resolution capability achieved with a kit targeting 16 STRs [46]. More recent research has shown that a panel of 50-60 well chosen identity informative SNPs can reach the same power of discrimination as commonly used STR kits [47, 48]

A major challenge for SNP analysis has been the simultaneous amplification of enough loci to obtain the required power of discrimination for forensic DNA typing. One technique for multiplex SNP analysis is called “SNaPshot” which relies on single base extension with fluorescent dye-labeled dideoxynucleotide triphosphates (ddNTPs) at the 3’ end of a primer directly upstream of the targeted SNPs, following amplification of the SNP region by PCR [49]. This has been the technique of choice for SNP analysis since the early 2000’s in forensic practise, as it allows alleles to be detected by CE, which was the instrumentation already in place for this type of laboratory [50]. Multiplexing capability with this technique is of about 30-50 SNPs, which can still be insufficient in a forensic context [51]. A huge benefit of using SNPs is that they can also be used for the prediction of certain traits such as phenotype and bio-geographical ancestry, which will be discussed later in this chapter.

1.2. Massively parallel sequencing

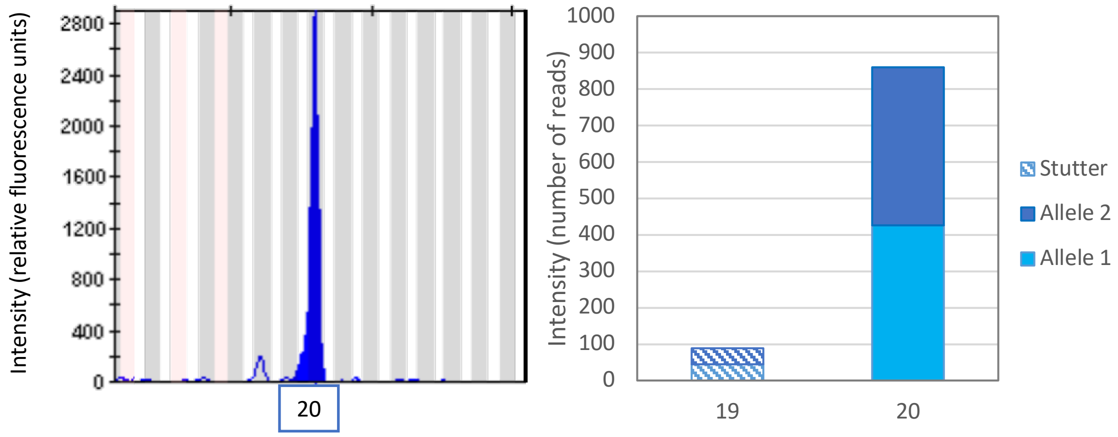
Fred Sanger believed that knowing the specific chemical structure of a biological molecule was necessary for a deeper understanding of its function and mechanism of action [52]. Around 1977, the method he developed for protein sequencing was first applied to DNA. Sanger sequencing, as it has come to be known, starts with the use of short primers binding near the region of interest. Each DNA strand is sequenced in a single reaction with the appropriate primer. In the presence of the four DNA nucleotides, a DNA polymerase will extend the primer by adding the complementary nucleotide from the DNA template strand. By using modified nucleotides with a removed hydroxyl group at the 3' end of the molecule called dideoxyribonucleotide triphosphates (ddNTPs), this ensures that the reaction stops here, i.e. the ddNTPs act as chain terminators. Extendable dNTPs are present alongside ddNTPs in the reaction mix so that some portions of the DNA are extended. Sanger sequencing results in the formation of different length fragments, which can be separated by CE to a resolution of one base. Each ddNTP is labelled with a different fluorescent dye, which can be identified and visualised as individual bases of a DNA molecule

Despite the success and continued use of Sanger sequencing, many applications required faster, higher throughput technology. In the field of forensic science, laboratories focussed on PCR amplification and CE for the analysis of STRs, although Sanger sequencing is still used in certain specific cases, and for individual SNP analysis.

Massively parallel sequencing (MPS), also referred to as second generation sequencing or as next generation sequencing (NGS) is a high throughput approach to genomic sequencing. The technology was first introduced in 2005, and is now commonly used in oncology, microbial genetics and disease genomics worldwide [53, 54]. MPS allows the entire human genome to be sequenced at once, which can provide access to all genetic variation between individuals including those occurring in coding, regulatory, and intronic regions. Whilst this can be valuable in the study of diseases and biological systems, it requires a significant sequencing and interpretation effort. More recently, MPS has been applied to the field of forensics, with researchers discussing a move from forensic genetics to forensic genomics [55], and from next generation sequencing to "now generation sequencing"[56]. In forensic DNA typing, targeted sequencing is more appropriate than whole genome sequencing; by sequencing a dense set of loci, casework and database efforts are directed towards genomic regions that best answer forensic questions. This relieves privacy concerns and produces less data than whole genome sequencing, thereby simplifying analysis.

1.2.1. Advantages over capillary electrophoresis

Unlike CE, which translates DNA molecule migration time into DNA fragment length, sequencing allows the visualisation of underlying bases of DNA molecules. This has a major impact on the analysis of STRs – given each base of the target DNA is sequenced, variation at the molecular level can now be investigated. As a result, alleles of the same size but differing in sequence, which would previously have been masked with CE, can now be differentiated using MPS [57-65]. These are sometimes referred to as “iso-alleles”, an example of which is shown in Figure 1.2. This alone provides one of the major strengths of MPS over CE, given a larger number of detectable alleles will result in a more discriminatory test.



AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAC AGAC AGAC AGAC AGAC AGAC AGAC AGAC AGAC
AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAC AGAC AGAC AGAC AGAC AGAC AGAC AGAC AGAT

Figure 1.2: Example of iso-alleles at D12S391

Both alleles are the same length (20 repeats) but differ at the sequence level (bottom, sequence). In a CE profile (top, left), these alleles would be indistinguishable. Using MPS (top, right) however, the difference can be visualised thanks to access to sequence-level information. The X axis for both graphs represents the alleles, and the Y axis represents the intensity (in relative fluorescence units for CE, and in number of reads for MPS).

Mixtures, or DNA profiles from several individuals, are a commonly encountered issue in forensic science, and sequence variation is likely to be a useful tool for deconvoluting contributions within such a sample [66]. It can also be used to separate a true allele from a stutter/artefact product, as demonstrated in Figure 1.3. Stutter is commonly encountered when analysing results from STRs amplified using PCR, and is the result of DNA polymerase slippage. Because of the repetitive nature of STR sequences, the DNA polymerase enzyme can "slip", usually backwards, resulting in a DNA fragment that is one repeat motif shorter than the true sequence [67]. This is often referred to as "N-1" stutter, and "N+1" can also occur, when the DNA polymerase slips forward by one repeat motif. With CE, stutter products appear as separately sized peaks on an electropherogram, and can be indistinguishable from a true allele of the same length. In a single source profile (from one individual), stutter peaks usually have an appreciably lower intensity than the true allele, and for data interpretation purposes, acceptable stutter percentages are usually established during method validation and applied to each STR locus targeted. When dealing with mixed source profiles (from more than one individual), stutter often complicates analysis as it may mask or confuse interepretation of a minor contributor allele [68]. Although stutter is still encountered with MPS due to the underlying reliance on PCR, there is a increased chance of differentiating it from a true allele if the sequences differ [69].

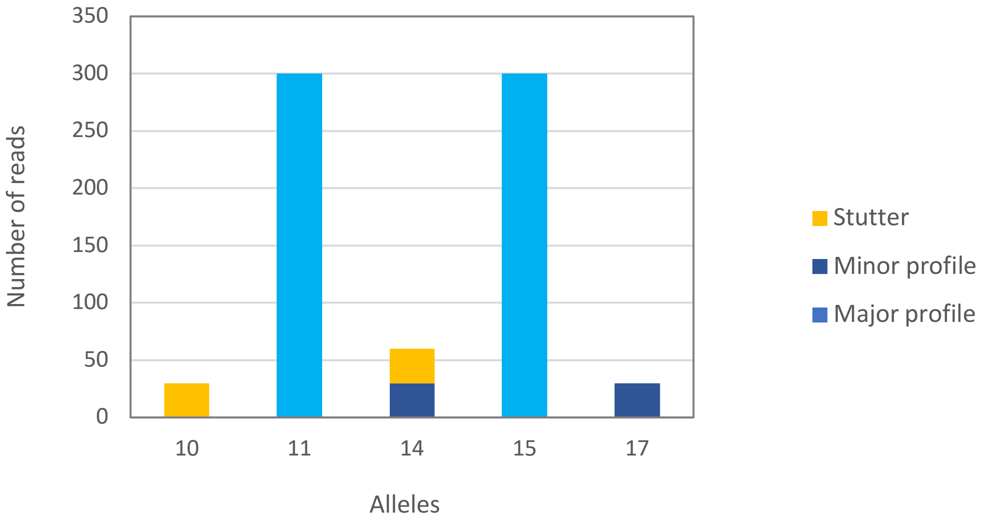


Figure 1.3: Theoretical example of a 2-person mixture

In this example, there is a mixture of two DNA profiles, where one is present at a higher concentration (the major contributor), and one at a lower concentration (the minor contributor). Sequence level variation is used to distinguish the minor profile allele 14 from the stutter product of the major allele 15.

Other advantages over CE include MPS' increased locus multiplexing ability, which makes it an ideal tool for DNA identification [62, 63, 69-74]. The lack of size restriction means a much larger number of autosomal STRs can be investigated, expanding capability past the core STR loci sets, as well as markers of different types. Targeting SNPs can strongly increase the chance of obtaining information from degraded samples, often encountered in a forensic context [69, 75]. Sex chromosome markers can help provide information in cases of complex relationship identification [76], disaster victim identification [77] or sexual assault [78], and are often run as an additional test to autosomal STRs with CE for such cases. With MPS, these markers can be run simultaneously, thus saving time and resources [79]. The sequence variants, as well as the additional markers, can increase the power of discrimination of a DNA test.

As MPS can be used to analyse a much wider range of targets than was previously available, there is now the potential to multiplex more than just identity informative markers. The ability to infer what someone looks like in terms of their externally visible characteristics, and where they are from in terms of global populations (where their ancestors are from, otherwise known as "bio-geographical ancestry") are useful tools to supplement traditional DNA profiling [80]. "No hit, no suspect" cases arise frequently, where a DNA profile returns no match from a database, and no suspect has been identified by either a victim or during police investigation. Here, the ability to determine externally visible traits such as hair colour, eye colour and bio-geographic ancestry from DNA evidence can provide investigative leads to help solve a case [81, 82]. Phenotypic and bio-geographical ancestry informative SNPs can also be useful in the case of historical remains, where a reference sample or next of kin may be lacking [83-86]. The use of MPS to investigate factors such as ancestry and physical characteristics has already been well documented [87-90], including research performed using custom SNP panels [90] and commercially available solutions such as the ForenSeq DNA Signature Prep Kit primer mix B panel discussed further[89].

De Knijf [56] suggests that one of the factors accounting for the late adoption of MPS in forensics is the often limited and degraded nature of the DNA encountered regularly from crime scenes. Forensic cases often involve samples where the DNA is found to be in poor condition. Exposure to the elements and other external factors can all contribute to DNA degradation, where the genetic material is fragmented. This means that, for these types of samples, smaller target regions are likely to yield better results [91]. The improved locus multiplexing ability and small amplicon size for both STRs and SNPs make MPS a helpful tool for analysing degraded samples. The applicability of MPS on degraded DNA has already been investigated in a number of publications, with markers such as Y-STRs and SNPs proving useful when autosomal STRs alone have not yielded sufficient information [90, 92-94].

1.2.2. Flanking regions of amplicons

Flanking regions refer to the sequences between where the PCR primers bind and the target region, which in the case of STRs is usually the repeated region. With CE analysis, these regions were only considered in so much as they contribute to the overall PCR amplicon size, and that deletions or insertions in these regions could cause issues with allele designation. Because MPS results provide the full sequence of all amplicons, it is now possible to view flanking regions. Although much of the initial published research on STR sequence variation [62, 70, 73, 95] has focussed on the repeat regions of STRs given these are historically of more importance, preliminary investigations into the flanking regions of these markers has shown that further variation can be observed here, thus further increasing the power of discrimination of the MPS test [63, 96-98].

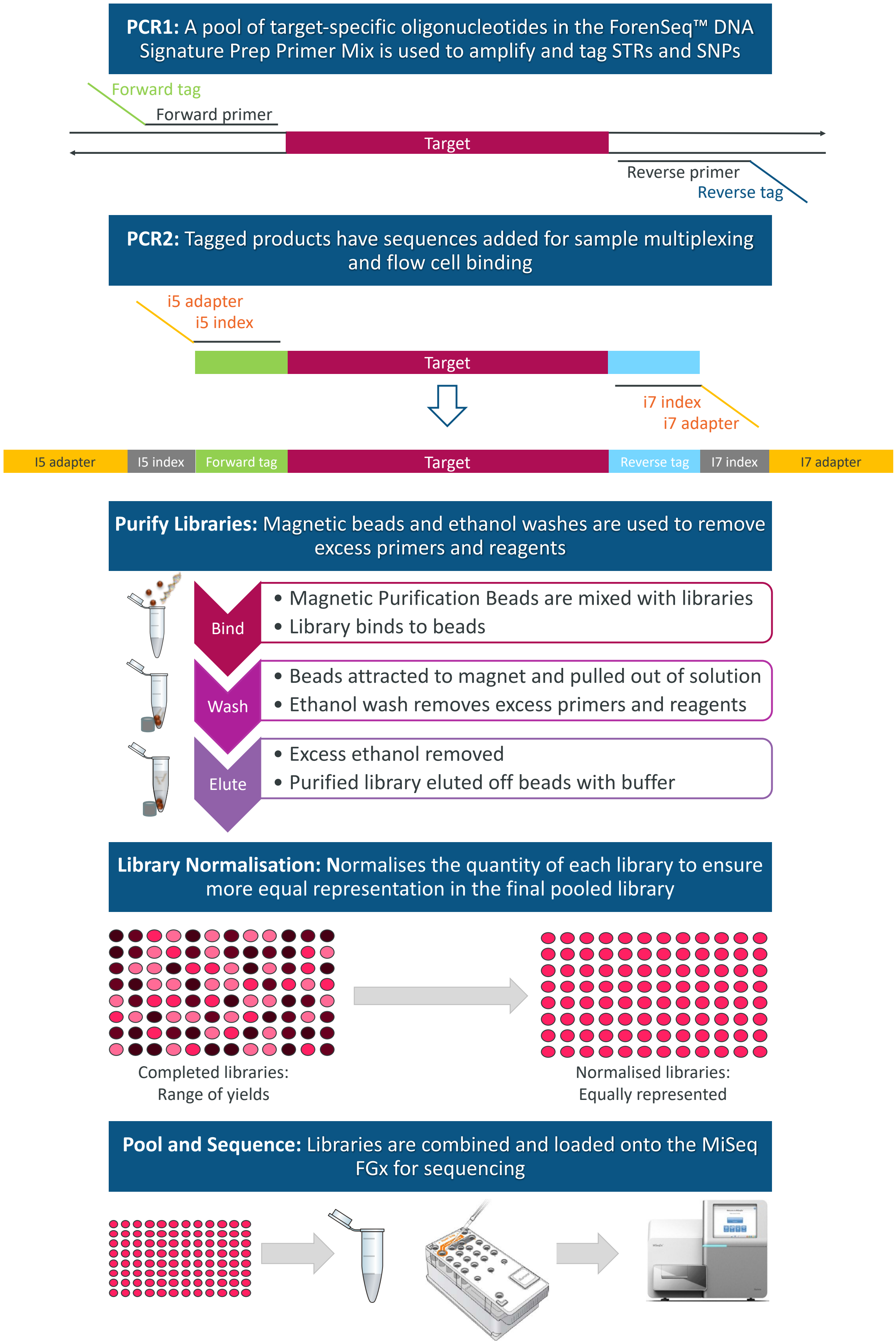
1.2.3. MiSeq FGx Forensic Genomics Solution

The MiSeq sequencer is a benchtop sequencer manufactured by Illumina, which allows for small genome and targeted sequencing [99]. The MiSeq FGx™ Forensic Genomics Solution (Verogen, San Diego) was first released by Illumina in 2015, comprising of an application specific MiSeq platform and dedicated library preparation kit.

1.2.3.1. ForenSeq DNA Signature Prep Kit

In order to prepare DNA for sequencing on an MPS platform, a process called library preparation must take place, with a "library" effectively being a DNA sample that is ready for sequencing. As with traditional DNA typing, this process starts with targeted amplification of areas of interest. The ForenSeq™ DNA Signature Prep Kit (Verogen), targets 27 autosomal STRs, 24 Y-STRs, 7 X-STRs and 94 identity-informative SNPs when using DNA primer mix A (DPMA) [100]. These markers cover the majority of autosomal STRs commonly targeted by commercially available and recommended STR kits such as the PowerPlex® ESI 17 and GlobalFiler® kits [34, 101], as well as Y STRs and X STRs routinely used in forensics – the advantage being that all markers can be amplified simultaneously in one reaction. The identity informative SNPs were selected from published research and chosen based on the fact that they are polymorphic across global populations (i.e. where both alleles are found at high frequency worldwide) [50, 102, 103]. There is also the option of using DNA primer mix B (DPMB), which contains primers for the same markers as DPMA, as well as those for phenotype and bio-geographical ancestry-informative SNPs. 54 ancestry informative SNPS (aSNPs) were selected from the Kidd lab aSNP panel [104, 105] and 22 phenotype informative SNPs (pSNPs) were selected from the HIrisPlex system, a set of markers specifically developed to determine the hair and eye colour of an individual based on their DNA [106, 107]. The amplicon range for this kit is 60-460 bp, with SNPs forming the majority of the markers under 200 bp [108]. In DPMB, 190 markers are below 200 bp in size [109]. Depending on the level of degradation of a DNA sample, this can offer an important improvement over current CE-based STR kits [110, 111]. For example, only 6 out of the 9 STR markers targeted in the AmpFℓSTR MiniFiler® PCR Amplification Kit, which was specifically designed for degraded samples, fall below 200 bp [112, 113].

The steps for library preparation are described in Figure 1.4. Following initial amplification of targeted STRs and SNPs, the amplicons are exposed to a second round of PCR to incorporate indices, which are short (8 base pairs) known sequences that act as barcodes, which will allow amplicons belonging to different samples to be teased apart during the data analysis process, and adapter sequences, which will be necessary for the sequencing process to begin. Once these have been added, the samples are effectually "DNA libraries" and can be sequenced. A purification step forms part of the protocol to remove unincorporated dNTPs, primers and other reagents, before a normalisation step to ensure equal representation of DNA from each sample. Finally, the libraries are pooled together, ready for sequencing. When using DPMA, up to 94 reference samples and 2 controls can be sequenced simultaneously. With DPMB, due to the higher number of markers targeted, the manufacturer recommendation is to sequence no more than 30 samples and 2 controls at the same time. The kit is used to prepare libraries for sequencing on the MiSeq® FGx system [114], and its performance has been evaluated in several publications, where it has been found to be robust and sensitive for forensic applications [73, 74, 89, 115-117]. More recently, a number of internal and inter-laboratory studies have also been undertaken to evaluate the kit for different forensic applications [118-120].



PCR1: A pool of target-specific oligonucleotides in the ForenSeq™ DNA Signature Prep Primer Mix is used to amplify and tag STRs and SNPs

PCR2: Tagged products have sequences added for sample multiplexing and flow cell binding

Purify Libraries: Magnetic beads and ethanol washes are used to remove excess primers and reagents

- Magnetic Purification Beads are mixed with libraries
- Library binds to beads

- Beads attracted to magnet and pulled out of solution
- Ethanol wash removes excess primers and reagents

- Excess ethanol removed
- Purified library eluted off beads with buffer

Library Normalisation: Normalises the quantity of each library to ensure more equal representation in the final pooled library

Pool and Sequence: Libraries are combined and loaded onto the MiSeq FGx for sequencing

Figure 1.4: ForenSeq DNA Signature Prep Kit protocol

1.2.3.2. Cluster generation

Prepared libraries are loaded into a reagent cartridge which is then placed in the MiSeq FGx (Figure 1.4, final stage). Cluster generation and sequencing take place on a glass slide called a flowcell. The flowcell has an etched lane, and at the start of any sequencing run the single stranded DNA library fragments and reagents are pumped along this lane, which is coated with two types of oligonucleotides – complementary to the i5 and i7 adapters bound to each fragment. Hybridisation occurs between the DNA library fragments and this "oligonucleotide lawn", and a complementary strand of each hybridised fragment is then synthesised. The double stranded molecule is subsequently denatured and the original template strands are washed away. The newly synthesised DNA fragment then bends over and forms a "bridge" to another oligonucleotide on the flowcell, and a complementary strand is synthesised, leading to two complementary strands being attached to the flowcell. These are denatured, bent over to form another bridge, hybridise to another oligonucleotide and new strands are synthesised. This process is repeated for a number of cycles and is referred to as "bridge amplification", leading to clusters which contain thousands of copies of the same initial DNA library fragment (Figure 1.5). After bridge amplification, all reverse strands (fragments with P5 complement sequence) are removed, leaving just the forward strands for sequencing by synthesis [121].

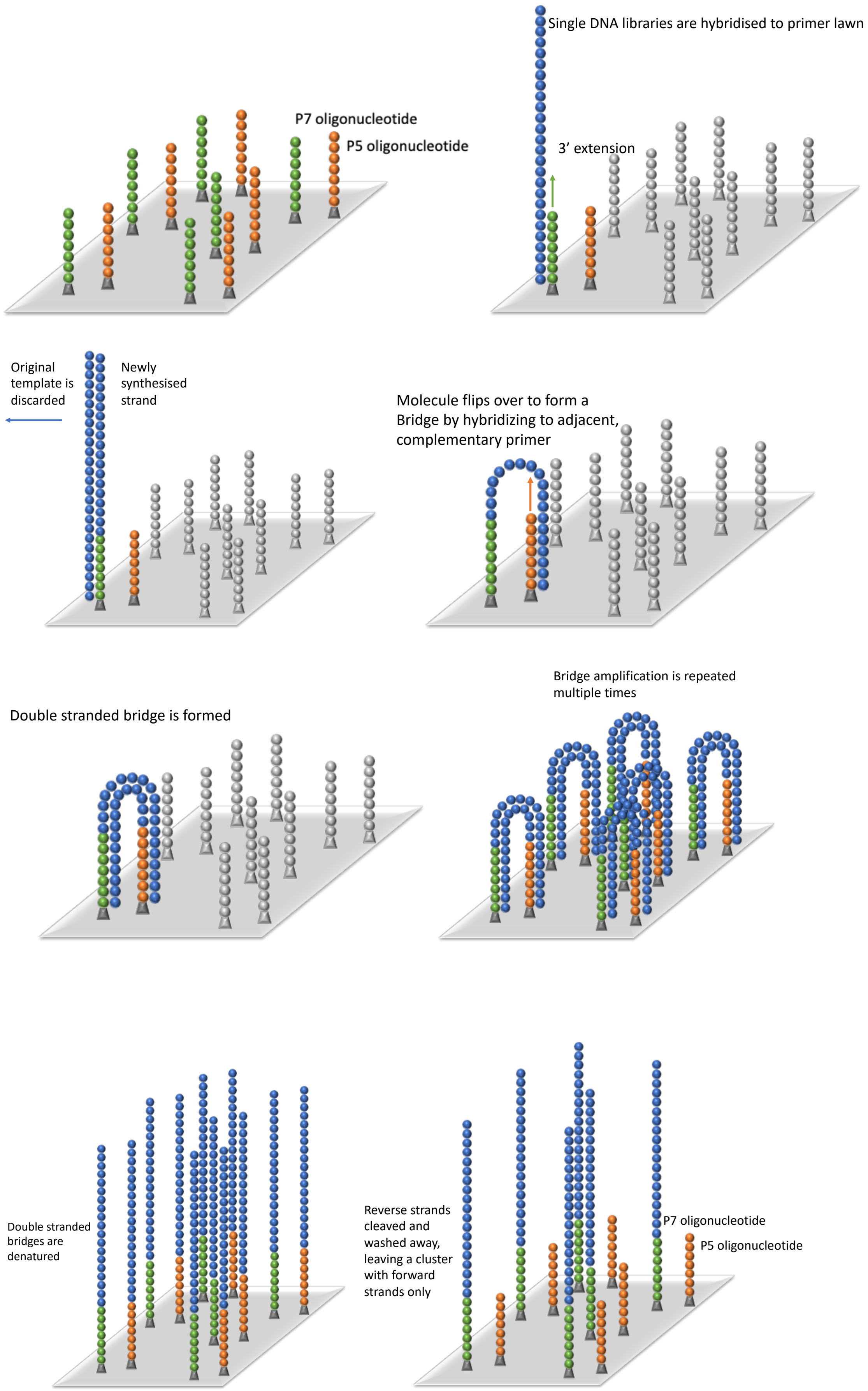

P7 oligonucleotide
P5 oligonucleotide
Single DNA libraries are hybridised to primer lawn
3’ extension
Original template is discarded
Newly synthesised strand
Molecule flips over to form a Bridge by hybridizing to adjacent, complementary primer
Double stranded bridge is formed
Bridge amplification is repeated multiple times
Double stranded bridges are denatured
Reverse strands cleaved and washed away, leaving a cluster with forward strands only
P7 oligonucleotide
P5 oligonucleotide

Figure 1.5: Cluster generation on a MiSeq flowcell

1.2.3.3. Sequencing by synthesis

Illumina’s sequencing by synthesis (SBS) technology is a widely accepted approach for sequencing PCR amplicons and is the basis of more than 3500 publications. The chemical concepts are similar to Sanger sequencing, with each nucleotide of a small fragment of DNA template sequentially identified from signals emitted during a series of synthesis cycles. The term “massively parallel sequencing” comes from the fact that each of the millions of clusters on a flowcell can be sequenced simultaneously.

At the start of a run, a sequencing primer hybridises to the 3’ end of the I5 adapter sequence of the forward strand of each molecule attached to the flowcell. A combination of polymerase and four fluorescently labelled deoxyribonucleotides (dNTPs) with 3’ reversible terminators are pumped over the flowcell surface, allowing the detection of each single complementary base that is incorporated into a growing DNA strand. The fact that the Illumina chemistry contains a mix of all four reversible terminator bound dNTPs during each sequencing cycle minimises competition for base pairing (“incorporation bias”). As each dNTP is added, a fluorescently labelled terminator is imaged and then cleaved to allow incorporation of the next base [121]. Fluorescence is captured by a camera which records four images during each incorporation, one for each of the dNTPs, which emit light at different wavelengths (Figure 1.6). As all the strands in a cluster have the same sequence, the light they emit is recorded as one single signal. The incorporation, detection and cleavage steps constitute a single cycle which is repeated a number of times to achieve the desired read length (i.e. 300 cycles would equate to 300 base pairs). MiSeq reagents typically enable up to 15 Gb of output per run, with 25 million sequencing reads and up to 600 cycles (often split into 2 X 300 bp read lengths) [99].

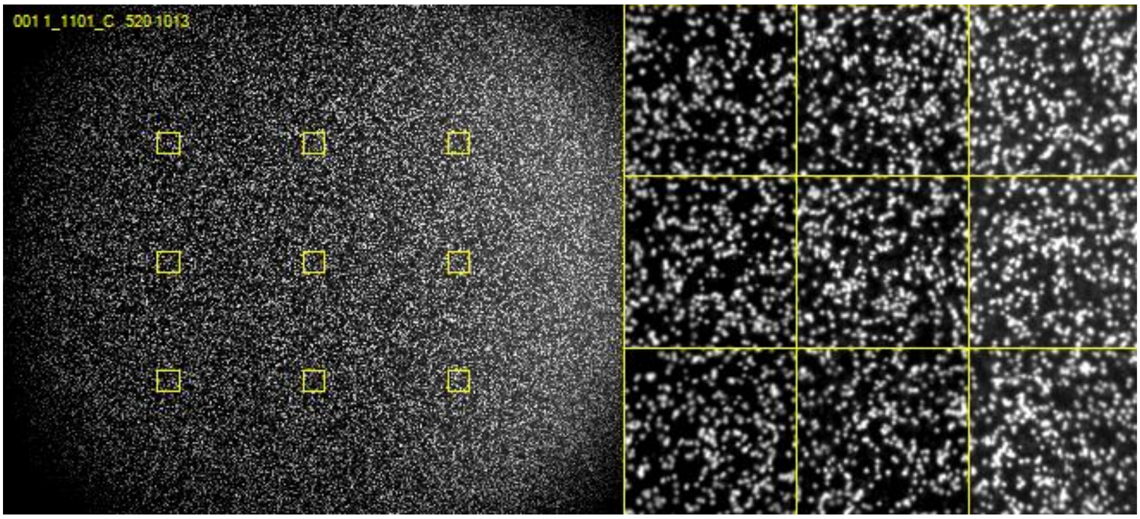


Figure 1.6: Image of a flowcell during sequencing on the MiSeq FGx

Light emitted for cycle 1, base C.

A ForenSeq DNA Signature Prep Kit run consists of 398 sequencing cycles. “Read 1”, described above, is sequenced for 351 cycles [122], after which the read 1 product is removed from the flowcell bound strand, and the i7 sequencing primer is hybridised to the 5’ end of the i7 adapter (Figure 1.7). “Index Read 1” consists of eight cycles, used to identify the i7 index associated with the sample. Once this is finished, the i7 read product is removed and the template bends over to hybridise to an adjacent P5 oligonucleotide on the flowcell, which then serves as the i5 sequencing primer for “Index Read 2” (also eight cycles). The i5 read product is then removed, and the original template strand is used to synthesise a complementary strand attached to the P5 oligonucleotide. The two strands are denatured, and the original forward strand is cleaved off, leaving only the single-stranded reverse strand bound to the P5 oligonucleotide. “Read 2” starts with the read 2 sequencing primer hybridising to the 3’ end of the i7 adapter, and in a ForenSeq DNA Signature Prep Kit run consists of 31 cycles.

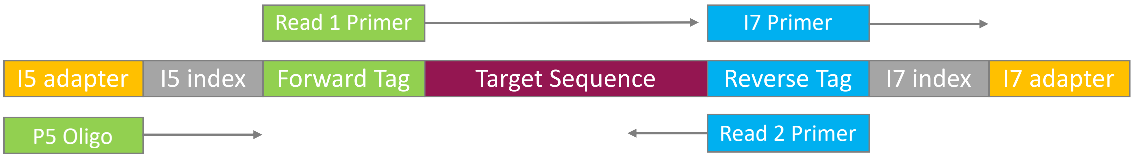


Figure 1.7: Sequencing by Synthesis of a ForenSeq DNA Signature Prep kit library

Read 1 sequencing begins with the hybridisation of the Read 1 sequencing primer to the 3' end of the i5 adapter sequence ("Forward tag"). Index read 1 starts with the hybridisation of the i7 sequencing primer to the 5' end of the i7 adapter sequence ("Reverse Tag"). The molecule then bends over to hybridise with an adjacent P5 oligonucleotide, which serves as the i5 sequencing primer for Index read 2. Following reverse strand synthesis and denaturing, Read 2 sequencing begins with the hybridisation of the Read 2 sequencing primer to the 3' end of the i7 adapter sequence.

1.2.3.4. MiSeq FGx sequencing metrics

When assessing the quality of a MiSeq FGx sequencing run, several metrics are taken into consideration:

a. Cluster density: The number of individual libraries that formed clusters on the flowcell. This metric is related to the loading concentration of the DNA library pool. Too high, and the clusters could overlap, leading to overlapping fluorescent signals and issues for base calling [123]. Too low, and the clusters could be spaced too far apart and result in low sequencing read coverage of some libraries.

b. Cluster Passing Filter: The percentage of clusters that passed an internal quality filter. This metric is related to cluster density: Ideally, each cluster is distinct, and this filter removes the least reliable data, often derived from overlapping clusters.

c. Phasing and prephasing: During sequencing, reversible terminators should ensure a single dNTP is incorporated during each cycle, however in some of the strands of a cluster more than one base will occasionally be added [124]. This can cause the strand to be ahead of the others and is known as "pre-phasing". Conversely, in some strands no dNTPs may be added, leading to "phasing", where these strands are lagging.

1.2.3.5. Data processing

Sequencing data is processed first on the MiSeq FGx, before being transferred to a server containing the Universal Analysis Software [122]. During the first few Read 1 sequencing cycles, the location coordinates of each cluster on the flowcell are determined. Fluorescent images for every cycle are aligned to these coordinates at the end of a sequencing run, and cluster intensities are extracted. A filtering step removes poor quality clusters caused by over-clustering, poor amplification or sequencing using a chastity filter. Clusters that pass this filter are converted to base calls based on the signal intensities for each cycle, with quality scores recorded in a .bcl file.

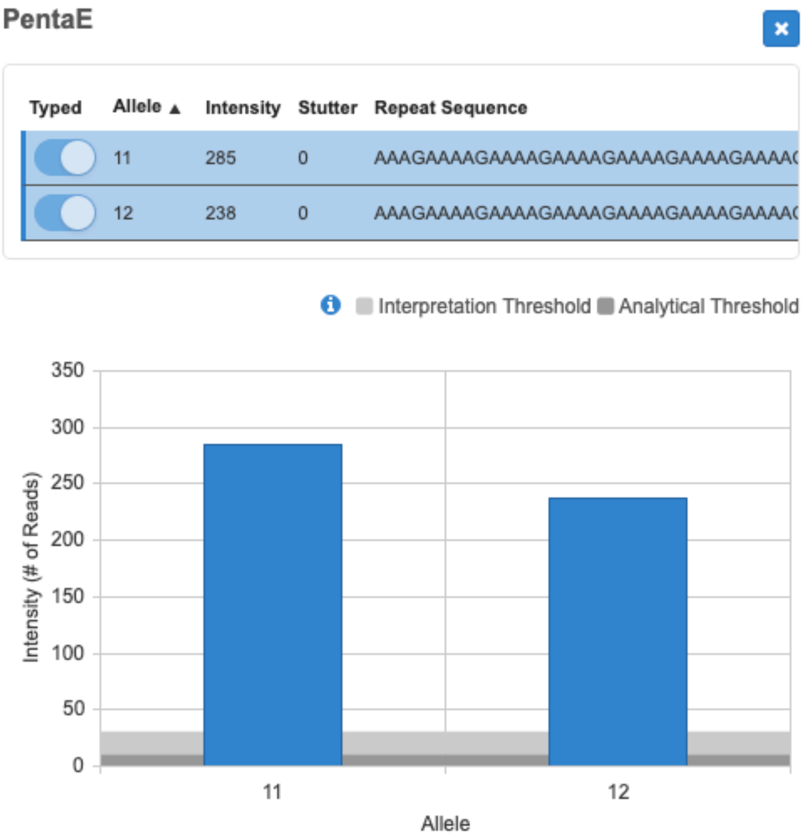
Prior to sequencing, the dual index combination for each sample is recorded in a sample sheet, which is then used for demultiplexing clusters belonging to individual libraries in a pool during data processing. Base calls for each identified cluster are combined and recorded in a .fastq file for each library.

1.2.3.6. Universal Analysis Software

The ForenSeq Universal Analysis Software (UAS) utilises two different bio-informatic methods for the extraction of target sequences for STRs and SNPs from the library fastq files.

1) SNP sequencing reads are aligned to the human reference genome (hg19) using a Smith-Waterman-Gotoh algorithm [125], whereby the SNP locus and strand is identified using the primer sequences, the reads are aligned, and the SNP locus is confirmed using the read 2 sequence.
2) STR sequencing reads are aligned using a Needleman-Wunsch algorithm [126], which identifies primer sequences and seed sequences in the flanking regions of both sides of the STR to identify the allele. For simple STR loci, the number of expected repeats for the locus are counted in order to name the allele (according to length-based nomenclature). For complex or compound STR loci, a more customised interrogation of the locus is conducted to correctly call the allele.

After allele calling, each read is assigned to a locus and the abundance of each sequence is recorded as “read coverage”, and genotyping thresholds are applied by the software. UAS uses a number of thresholds and parameters during genotyping. Two adjustable thresholds exist within the software: an analytical threshold, below which no alleles can be called and is set by default to 1.5% of the total reads aligned to that locus with a 11 read cut off; and an interpretation threshold, below which alleles are not automatically called but can be changed manually, set by default to 4.5% of the total reads and with a 30 read cut off. Figure 1.8 shows an example profile for the locus Penta E in UAS, including how the two thresholds are visualised. STR loci have an additional, loci specific, adjustable stutter threshold. A number of “flags” for identifying loci-level considerations such as heterozygous imbalance, elevated stutter, an incongruous number of alleles and other factors are also applied by the software [122].



PentaE

Typed	Allele ▲	Intensity	Stutter	Repeat Sequence
●	11	285	0	AAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAA
●	12	238	0	AAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAA

Figure 1.8: Example of UAS profile for STR locus Penta E

The two alleles in this heterozygous profile are shown in blue, with the allelic sequences given at the top. The analytical and interpretation thresholds are visualised on the chart in dark grey and light grey, respectively. Screenshot taken of actual software.

Once the results for a sample have been manually checked in UAS, reports can be downloaded at the sample or run level. Sample detail reports contain all of the sequences for the target region of the STR and SNP amplicons with 11 or more reads, including any possible stutter or sequencing artefacts. Sample summary reports are similar, but only contain sequences that have been designated as allele (either initially by the software algorithm or following manual change by the analyst). Earlier versions of the software did not include any form of analysis of the flanking regions of the STRs or SNPs, but version 1.3 included an additional feature to export a report containing full amplicon sequences with 11 or more reads, and has been available since April 2018.

UAS has been compared to other available software tools. In 2017, Wendt et al. found variations in the depth of coverage of alleles identified using UAS and a software called STRaitRazor [127], which they suggest is likely due to the two bio-informatic tools targeting different flanking region sequences for alignment – often called either "Seed" or "Anchor" regions depending on the software [96]. It is also probable that the UAS algorithm uses an internal chastity filter which removes read sequences that do not meet a certain quality criteria, leading to the slightly lower depth of coverage described in this article. The size of the amplicon may additionally affect this, as sequencing errors are more likely to occur towards the end of a long read sequence. Ultimately, the difference in depth of coverage did not lead to any discordance between the two software tools, which returned the same results for the samples analysed, making the difference between the two inconsequential. In 2021, Hoogenboom et al. analysed a series of samples that had already been processed through UAS using their "STRNaming" algorithm [128]. Here, they explain the differences in reporting strategy and reporting range for the loci in the ForenSeq DNA Signature Prep kit, with a major objective of the authors' research being in ensuring compatibility between results obtained using the different tools. To this end, STRNaming was specifically developed to be compatible with UAS.

1.2.4. Implementation of massively parallel sequencing

MPS is an appealing technology to overcome the limitations of CE and advance the field of forensic genomics, but there are challenges that need to be addressed before it can become part of routine analyses. In a survey conducted in 2017 by the DNASEQEX consortium, the major challenges to MPS implementation were outlined as follows [129]:

- Lack of compatibility with existing national DNA database infrastructure
- Lack of population data to support statistical calculations
- Lack of consistent nomenclature and reporting standards

In addition to these scientific challenges, the cost of running MPS-based assays remains higher than that of CE, and therefore a barrier to implementation for certain laboratories. A recent review by Alonso et al. suggests that the way forward will be in large-scale collaboration projects, which will contribute towards solving the three scientific challenges listed above, and will in turn help to reduce the cost of certain aspects of the validation and implementation of these new technologies [130].

1.2.4.1. Concordance with CE and compatibility with DNA Databases

Before any new STR kit can be implemented into a forensic casework laboratory, its concordance with previous kits must be demonstrated [131, 132]. In order to ensure compatibility with existing databases, results obtained from any given sample typed using different technologies should be the same. Many kits target the same markers, but differing kit configurations can mean different primer sequences are used to amplify the same STR [112]. Where discordant results are found, due for example to primer binding site mutations, these can be listed on a database, or information provided to the manufacturer to enable a reconfiguration of the primer design. It is especially important to check the concordance of MPS with CE-based technologies, as the manner in which alleles are designated is so inherently different. Concordance studies for the ForenSeq DNA Signature Prep Kit have already been undertaken as part of a developmental validation study performed by Illumina [108] and as part of several external studies [73, 74, 89, 115, 117]. Results have shown a level of concordance exceeding 99% between this kit and CE-based STR multiplexes, although the diversity of samples used in these studies have been limited both in terms of numbers and population groups investigated. In particular, the initial research outputs from the development and validation of the kit suggest that all testing was performed mainly using samples from three global population groups: Caucasian, African American (which is mostly equivalent to the West African ancestral population) and East Asian. Online data from 2019 suggests that approximately 23% of the world population are of South Asian ancestry (India, Bangladesh and Pakistan) [133], yet this population was not accounted for during the initial testing of the ForenSeq DNA Signature Prep kit.

1.2.4.2. Population databases

Population databases are used for the statistical evaluation of DNA profiles and contain allelic frequencies for markers targeted. In order to make use of the variation observed through the sequencing of STRs, new databases must be generated, containing allelic frequencies for sequence-based alleles rather than length-based ones alone. Although work has begun to characterise sequence variants for common STR loci [60, 62, 63, 74], there is a lack of population data necessary to the implementation of the technique. Novroski et al. [63] sequenced 777 unrelated individuals from four U.S. population groups (Caucasian, Hispanic, African American, and Chinese) and documented variation seen within both the repeat region and flanking region of the 58 ForenSeq STR loci. Fifty of the markers demonstrated an increase in allelic diversity when sequence-based alleles were compared to length-based alleles. Hussing et al. [134] sequenced 363 Danish population samples and found allelic diversity increased in 34 of the STR markers when comparing sequence-based data to that obtained with CE. At 10 markers, more than double the number of sequence-based alleles were observed compared to length-based alleles alone. Wendt et al. [96] sequenced 62 Native American samples and found significant variation in the flanking regions of 11 of the STRs targeted.

1.2.4.3. Nomenclature challenges

Nomenclature was first discussed in section 1.1.4, referring to the rules for designating STR alleles according to length and repeat type. The move to define alleles by their sequence rather than length brings with it a number of challenges which must be addressed before MPS can be implemented for routine forensic DNA testing. The annotation of each allelic sequence determined using MPS must now meet two criteria: First, it must be back-compatible with the (generally CE-generated) nomenclature used in national DNA databases (section 1.2.4.1) and secondly, it must encompass and identify all sequence level variation within the agreed range for any given marker. The latter is important given the need to account for relevant genetic variation, whilst also ensuring that searching can be accurately performed across different population databases (section 1.2.4.2).

In terms of compatibility with current DNA databases, as stated by Gettings et al. in 2015, "this concordance challenge is neither new nor insurmountable"[65]. Even between CE-based assays, discordances can arise between results obtained for the same sample run with two different kits, due to difference in primer design. As discussed early on in this chapter, primer binding site mutations can lead to inefficient or complete lack of amplification and may lead to different results for the same locus in the same sample, targeted with different primer sets. Additionally, insertions or deletions in the flanking regions of STRs can lead to different length amplicons if they are captured by the primer set in one kit but not another. With MPS, bioinformatic pipelines for data analysis need to be configured to ensure back-compatibility with results generated with CE, i.e., the length of amplicon sequenced must be rendered for allele designation.

A nomenclature that encompasses all possible sequence variants requires more thought. The DNA commission of the international society of forensic genetics (ISFG) set out an initial set of minimum criteria for standardisation of sequencing nomenclature at three hierarchical levels: The full sequence, the alignment of sequences relative to a reference sequence, and the annotation of alleles [135]. In 2016, the STR Sequence Working Group was formed, later formalised as the STRAND Working Group in 2018, with endorsement from the ISFG [136], with the aim of harmonising STR sequence nomenclature. In alignment with the considerations from the DNA commission, there are different types of naming formats, which fall into three broad categories:

1. **Short designator,** where a minimal code is used. The main advantage of using short designators is ease of databasing and in casework when needing to refer to alleles in a brief format. Several methods for implementing short designators have been published, such as the sequence identifier (SID) method by Young et al. [137], or the longest uninterrupted stretch (LUS) representation [138]. Short designators could be used to search against a catalogue of sequences such as the STRSeq BioProject [139]. The major drawback of reducing a sequence to a short designator is the loss of nucleotide level information.
2. **Bracketed repeat,** consists of condensing the repeat region of STRs into brackets, as is traditionally done for STR typing and examples of which can be seen in Table 1.2. This is a familiar framework for DNA analysts and provides a full understanding of the repeat sequence structure of alleles without needing the full string. This format requires the designation of an agreed start and end point of the repeat region. Here, using the historical start and end points may ignore neighbouring repetitive elements which could count for an important level of variation at certain loci, but defining new ones would involve a certain level of discussion and consensus amongst users.
3. **Full string,** as the name suggests, consists of providing the entire reported

sequence. This is by far the most comprehensive format as it “hides” nothing as such, but is ill-suited for reporting and may be incompatible for searching against certain types of databases such as CODIS which allows for a maximum number of characters to be searched.

Although the “final” agreement on the best system, if there is to be one, had yet to be reached at the time of writing, there are a number of factors generally agreed upon as necessary for reaching a nomenclature consensus. The primary necessity is that of large scale, multi-population databases which can be used for the generation of frequencies (and subsequent implementation of MPS in casework as discussed earlier), but also to gain an understanding of the breadth of sequence variation expected across all commonly used STRs. It is likely that a minimal reported range will be defined for all loci, encapsulating both the repeat region and a certain amount of flanking region which can be obtained using commercially available solutions. For this, it is useful to understand which parts of the sequence are more likely to vary, and so a wealth of population data will help to ensure the establishment of a nomenclature system capable of coping with the observed variation.

Table 1.2: List of autosomal STR markers in the ForenSeq DNA Signature Prep Kit

aSTR Locus	Chromosome	Strand	Repeat type	Locus type	Most common repeat motif [a]	Amplicon Length Range (bp)
D1S1656	1	R	Tetra	Compound	[TAGA]n [TGA]0-1 [TAGA]n [TAGG]0-1 **[TG]5**	133–192
TPOX	2	F	Tetra	Simple	[AATG]n	61–109
D2S441	2	F	Tetra	Compound [b]	[TCTA]n; [TCTA]n [TNNN] [TCTA]n	137–177
D2S1338	2	R	Tetra	Compound	[TGCC]n [TTCC]n	110–203
D3S1358	3	F	Tetra	Compound	[TCTA] [TCTG]n [TCTA]n	138–194
D4S2408	4	R	Tetra	Simple	[ATCT]n	98–118
FGA	4	R	Tetra	Compound	[TTTC]3 [TTTT] [TTCT] [CTTT]n [CTCC] [TTCC]2	150–312
D5S818	5	F	Tetra	Simple	[AGAT]n **AGAG**	98–162
CSF1PO	5	F	Tetra	Simple	[AGAT]n	72–120
D6S1043	6	F	Tetra	Compound [b]	[AGAT]n; [AGAT]n [ACAT]n [AGAT]n	154–226
D7S820	7	R	Tetra	Simple	[GATA]n	118–183
D8S1179	8	F	Tetra	Compound [b]	[TCTA]n; [TCTA]n [TCTG]n [TCTA]n	82–138
D9S1122	9	R	Tetra	Compound [c]	[TAGA]n; [TAGA] [TCGA] [TAGA]n	104–132
D10S1248	10	F	Tetra	Simple	[GGAA]n	124–176
TH01	11	R	Tetra	Simple	[AATG]n	96–140
vWA	12	F	Tetra	Compound	[TCTA] [TCTG]n [TCTA]n **TCCA TCTA**	135–195
D12S391	12	F	Tetra	Compound	[AGAT]n [AGAC]n [AGAT]n	229–289
D13S317	13	R	Tetra	Simple	[TATC]n **[AATC]2**	138–186
Penta E	15	R	Penta	Simple	[AAAGA]n	362–481
D16S539	16	F	Tetra	Simple	[GATA]n	132–184
D17S1301	17	F	Tetra	Simple	[AGAT]n	130–154
D18S51	18	F	Tetra	Simple	[AGAA]n	136–272
D19S433	19	R	Tetra	Compound	[AAGG] **AAAG** [AAGG] **TAGG** [AAGG]n	148–240
D20S482	20	F	Tetra	Simple	[AGAT]n	125–157
D21S11	21	F	Tetra	Complex	[TCTA]n [TCTG]n [TCTA]n **TA** [TCTA]n **TCA** [TCTA]n **TCCATA** [TCTA]n	147–265
Penta D	21	F	Penta	Simple	[AAAGA]n	209–298
D22S1045	22	R	Tri	Compound	[ATT]n [ACT] [ATT]2	201-245

[a] Nucleotides in bold are not counted towards the allele designation.

[b] Smaller alleles at D2S441, D6S1043, D8S1179 have simple repeat motifs, whereas larger alleles for these markers can be composed of a compound motif.

[c] D9S1122 has two common repeat motifs, one simple, one compound.

1.3. Ancestry estimation

The ability to accurately determine the genetic ancestry of an individual is of high interest in multiple areas. In the field of genetic epidemiology, genetic variants associated with disease risk can be geographically restricted due to evolutionary forces such as mutations, genetic drift and natural selection [140, 141], with incidence of breast cancer and diabetes for example differing in prevalence and severity across different ancestral populations [142, 143]. There is also a risk of false positive association in case-control association studies, where ancestry differences may accidentally be highlighted as disease vectors, for example where an allele is present at a higher rate in the case group than the control group because of unidentified population structure (or substructure) rather than being related to the disease [141, 144]. For these reasons, assessing the genetic background of potential research study individuals is crucial.

Ancestry estimation has also been a hot topic in the field of personal genomics, with individuals submitting samples to companies such as 23andme and Ancestry.com to gain insight into their biogeographic ancestry. In the field of forensics, as stated early on, a DNA profile obtained from a sample is typically compared to a reference profile (belonging to a victim, suspect, or from an elimination sample), or searched against a database. If no matches are obtained, a traditional profile is considered to provide no information more than the sex of the person who contributed the DNA. Additional tools, in the context of investigative intelligence, have been developed for the inference of bio-geographical ancestry [145, 146] and externally visible characteristics [107, 147]. There are a number of reasons why ancestry inference can be important in the context of forensic genetics. Given that eyewitness testimony is notoriously unreliable even when available [3], techniques that would allow the estimation of what a person looks like can be hugely beneficial for criminal investigations. It can also be used in the context of missing persons or mass disaster victim identification, in order to achieve more complete identifications and also confirm donors' self-declared ancestries in order to maintain the accuracy of population databases [148].

Markers used for the inference of bio-geographical ancestry are called ancestry informative markers (AIMs) and are generally SNPs [149], but other markers such as microhaplotypes [150, 151], nucleotide insertions or deletions [152] or STRs have also been used [152-155].

1.3.1. Ancestry informative SNPs

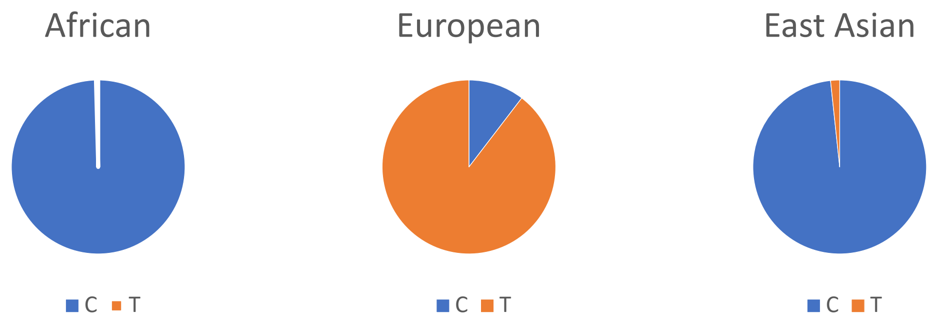
Autosomal SNPs have traditionally been the AIM of choice due to their stability, density of distribution and range of allelic frequencies across global populations. Identifying candidate SNPs involves looking for the most pronounced allele frequency differences between populations. Phillips et al. identified a 34-ancestry informative SNP panel [146] useful at distinguishing three global populations (sub-Sahara African, European and East Asian) by selecting markers that met one of the following criteria:

1) Loci with an allele detected in one or two populations but absent in the other(s).
2) Loci with a common allele in one population that is rare in others.
3) Tri-allelic SNPs.
4) Loci where one allele is seen exclusively in one population group and the alternative allele is seen exclusively in the others.

An example of a SNP included in the panel that corresponds to category 4 listed above is shown in Figure 1.9. “Fixed” difference markers have one allele seen exclusively in one population, with the alternative allele seen exclusively in the others. The SNP rs16891982 is found within the SLC45A2 gene and has been shown to be associated with skin de-pigmentation in Europe [156]. This is an example of gene variation that has been subjected to strong regional positive selection, leading to a locus that can be immensely useful for ancestry determination.

A large number of SNP panels have been developed over the years, often designed to discern ancestry between specific populations [157, 158], or at the global level [146, 159, 160]. In 2013, the genetics department at Yale university published a highly discriminative 41-SNP panel that aimed to meet two forensically relevant criteria: firstly, the ability to distinguish ancestral origin at the continental level and secondly, an assay that reduced both the cost and quantity of DNA required to obtain useable results.

These SNPs have been leveraged by a large number of laboratories in the forensic community, as well as by commercial companies. The 56 SNPs targeted by primer mix B in the ForenSeq DNA Signature Prep kit were selected from the since extended Yale university SNP panel, sometimes referred to simply as the “Kidd SNPs” [105], and have been demonstrated to accurately estimate ancestry of individuals from European and East Asian origin [116]. In the past, SNP-AIMs for these purposes would have been detected using Sanger sequencing or the SNaPshot primer extension assay [107, 161], but the advent of MPS has made the process considerably easier by allowing a greater number and type of markers to be sequenced simultaneously.



		C	T
	AFR	**0.996**	**0.004**
C	**0.996**	0.99168	0.00415
T	**0.004**		0.00002

	C	T
EUR	**0.104**	**0.896**
0.104	0.01085	0.09332
0.896		0.80252

	C	T
E. ASN	**0.983**	**0.017**
0.983	0.96653	0.01659
0.017		0.00028

Figure 1.9: Allelic frequency distribution for rs16891982

Top: Allelic frequency distribution in the three populations studied by Phillips et al. Bottom: Allelic frequencies (in bold) and likelihood of genotypes. Created using supplementary data from Phillips et al.[146] .

1.3.2. Consumer DNA testing and array genotoyping

Outside of the field of forensics, companies such as Ancestry.com, 23andme and even National Geographic offer commercial ancestry testing, promising consumers to connect to long-lost relatives and/or predict what parts of the world their ancestors came from. Users send in saliva samples, and can expect results within 6-8 weeks. Most of the technical testing details are proprietary to these companies, with 23andme for example giving no details about how many markers they target. Ancestry.com state that they target 700,000 locations using micro-array based genotyping of autosomal DNA [162]. SNP micro-arrays rely on the hybridisation of fragmented, single stranded DNA to slides containing hundreds of thousands of unique DNA probes. Each probe is designed to bind to a target sequence, which for this application are usually ancestry-informative SNPs. Priced at approximately £99 depending on the tests requested, this method offers a relatively cheap option for ancestry analysis. The lack of transparency in terms of testing methodology, prediction algorithms and accuracy especially make this type of testing entirely inadequate for forensic applications. With regards to accuracy, Ancestry.com state in their frequenctly asked questions that: “During the testing process, each DNA sample is held to a quality standard of at least a 98% call rate” which in itself gives no information as to the accuracy of genotyping or ancestry prediction. In fact, there are several instances where individuals or groups have submitted the same sample multiple times and under different names – and obtained different results from the same company [163-165].

Array SNP genotyping offers a relatively low cost option for wide-scale SNP typing and ancestry estimation which could theoretically be adopted by forensic laboratories, outside of consumer testing companies. One reason why it isn’t appropriate for forensic casework situations is the sample quantity requirements, with the aforementioned companies requiring “approximately a teaspoon” amount of saliva, stored in specific conditions and collected in a sterile tube. Prior to the implementation of massively parallel sequencing, research had been carried out to see whether array-based methods could be applied to forensics, but it was found to require too much input DNA. Krjutško et al. showed that 50 ng of input DNA was required for accurate analysis of 46 autosomal SNPs [166], compared to the 0.5 ng used to analyse 52 SNPs for human identity using the SNaPshot method and CE around the same time [50].

1.3.3. STRs for ancestry estimation

Traditionally, microsatellite markers such as autosomal and Y chromosome STRs have not been considered to any serious extent in the field of ancestry estimation due to the limited contrast in allelic frequencies between populations. Core STR loci were primarily selected for their highly polymorphic qualities, enabling the discrimination of unrelated individuals. Prior to the application of MPS technologies to forensics, there was already a significant push to improve the discrimination power of these STRs. The analysis of STR genotypes to infer genetic ancestry has been studied in the context of length-based allelic variation by a number of groups, but results have invariably highlighted limitations and inferior capabilities compared to SNP multiplexes designed for this reason. There are generally two broad approaches for consideration; either the adoption of specific STRs with strong population differentiation [152, 153] or looking at traditional markers for the ability to distinguish populations [154, 155].

Rosenberg et al. [167] used 377 STRs to successfully differentiate the major global population groups, but an assay of this size would be inappropriate in the context of a forensic scenario, where there is often insufficient DNA to target this many markers. Phillips et al. [168] used frequency data from this 377 marker dataset to identify tetranucleotide makers with the highest level of population informativeness. They generated a 12-plex of AIM-STRs that could be used as a stand-alone test or combined with identity informative STRs or even an AIM-SNP panel. Moriot et al. [152] genotyped the CEPH Human Genome Diversity panel (CEPH-HGDP) for 23 deletion-insertion polymorphism (DIP)-STR markers, which combine an insertion/deletion and a closely linked STR, selected from a set developed for improved mixture deconvolution. The rationale behind this marker choice lay in the fact that they are not only forensically relevant, but the combination of fast and slow- mutating markers would be beneficial for both individual and global population differentiation. Preliminary results from this study were promising, and clustering was considerably better when combining the data than looking at each type of variation individually (STR or DIP), however this set requires more markers in order to efficiently distinguish Eurasian populations.

Whilst these panels can be powerful tools when combined with standard STR typing, the fact remains that core forensic STRs are always typed first and foremost in routine profiling. If data from these STRs could be used effectively to distinguish global populations, this would be a huge advantage to the field of investigative intelligence. In 2001, Lowe et al. [169] suggested an approach for inferring ethnic origin using the 6 STR loci utilised by the UK Forensic Science Service at the time. They gathered allelic frequencies from the National DNA Database for five UK populations (Caucasian, Afro-Carribean, Indian sub-continent, Southeast Asian and Middle Eastern) to estimate the population proportion of a given profile to any of these populations. This method was applied to a case involving the rape of a woman where DNA recovered could either have been from a consensual encounter with a Caucasian partner, or assault by an Afro-Carribbean man as described by the victim. The method developed by Lowe et al. returned a result suggesting that the profile obtained was 28 times more likely to have originated from a Caucasian individual than an Afro-Carribean individual. Whilst promising from a research perspective, the authors emphasise the limitation of this work and the need for a larger number of more informative loci. Londin et al. [140] highlighted the need for small panels for accurate ancestral determination when studying diseases in populations, and initially looked at genotyping samples using the Identifiler and Coriell Identity mapping kits, which combined target 19 STRs. They established that these markers were not sufficient in number or ancestry informativeness for accurate ancestry determination and went on to develop a panel of specific AIM STRs as the groups above have.

During their investigation into global variability of the 15 established and 5 new European Standard Set (ESS) STRs, Phillips et al. concluded that the CEPH populations of Europe, Middle East and South Asia did not show sufficiently differentiated allele variation using these core STRs [170]. Using the program STRUCTURE [171], they did however show clear differentiation between the African, European and American populations in the CEPH group. When combining data from the 20 STR set with 34 AIM-SNPs, they were even able to differentiate the Oceanian population, which they were not able to previously do using AIM-SNPs alone.

Algee-Hewitt et al. conducted ancestry estimation using core STR sets, and found that a reduced number of markers limits the resolution of ancestry inference [155]. The same group later assessed the utility of these markers for ancestry estimation from post-mortem blood cards [172]. Using a single cluster approach to ancestry inference, 17 of the 20 samples tested classified into an ancestry group corresponding to their self-reported ancestry. These findings are limited and show the lack of accuracy obtained when looking at length-based allelic data from just 13 core STR loci. In 2010, Pereira et al. [173] published a new online calculator, designed to assign samples to one of three main population groups; Eurasian, East Asian or sub-Saharan African based on length-based data from 17 autosomal STR loci. They tested this tool on 48 samples from the three ancestral groups, and obtained 86% accuracy for individual population affiliation.

The results from the publications discussed hereabove seem to highlight two key points when it comes to using STR data for population affiliation:

1) Length-based data from core STR loci data can be used to roughly distinguish between ancestrally different populations such as Europe, East Asia and West Africa.
2) A higher number of STRs, or specifically chosen AIMs, are necessary for more accurate and reproducible population differentiation from autosomal STRs.

As discussed earlier, MPS offers increased discrimination of STRs through access to sequence level information, as well as a higher multiplexing capability. The large number of alleles for forensic STR markers is an indication of their instability over population divergence time, whereas SNPs in the flanking regions are likely to be more stable and therefore offer better ancestry resolution [152].

1.4. Thesis aims

Massively Parallel Sequencing offers a number of benefits to the forensic community including increased sensitivity, power of discrimination and multiplexing. It is not however, simply a matter of replacing the currently established technology. World-wide, DNA databases have been established using data generated by CE, and so any new technology must provide comparable data. This research first aims to assess the suitability of the MiSeq FGx Forensic Genomics System by verifying back-compatibility of results for searching against currently available databases. This is of particular importance for groups such as the South Asian population, which is a significantly represented group both in the UK and world-wide, yet no samples from this part of the world were tested as part of the original development or validation of the ForenSeq DNA Signature Prep kit or MiSeq FGx sequencing platform. Results discussed in the following chapters were generated by preparing and sequencing over one thousand samples using the ForenSeq DNA Signature Prep Kit and MiSeq FGx. The concordance of this technology with well characterised commercial STR-CE kits will provide insights into its applicability and ease of implementation in routine forensic DNA testing.

Characterising sequence variation and the nomenclature of STR alleles will form an important part of the dialogue throughout this thesis, starting with the traditional repeat region of these markers, which is no longer quite so traditional once we delve into the data at the nucleotide level. The increased granularity of analysis and considerations from other laboratories means it is important to also consider flanking region variation, and so part way through the project, this work shifted from characterising sequence variation simply within the traditionally defined repeat region of STR markers to looking outside of them as well. In order to make use of all this newly characterised sequence variation, a database of sequence-based allelic frequencies will be generated for the 5 UK relevant population groups. These population databases will enable laboratories to implement MPS for forensic DNA typing, as the frequencies will allow statistical evaluations to be made from sequence-based data.

The added value of providing investigative leads such as bio-geographic ancestry estimation is already a major, well document advantage of using MPS, although it mostly relies on the targeting of specific ancestry informative SNP marker typed as a standalone panel or in tandem with STRs. The final part of this thesis will look to address the possibility of using traditional autosomal STR markers for ancestry inference, based on sequencing data. Results will be compared to those obtained using ancestry informative SNPs.

2. MATERIALS AND METHODS

2.1. Samples

Samples from five well defined, UK relevant population groups were selected to evaluate the concordance of massively parallel sequencing (MPS) with capillary electrophoresis (CE) for commonly amplified autosomal STRs. The most recent guidelines for the publication of genetic data and for submission to online population STR databases require data from a minimum of 500 samples per population group for CE-generated data, but from just 50 samples per population group for MPS-generated data [174, 175]. Internally, the decision was made to analyse approximately 200 samples per population in order to generate an allelic frequency database, and to try to capture the breadth of sequence-level variation for the 27 autosomal STRs in the ForenSeq DNA Signature Prep kit.

Buccal swab samples from at least 200 unrelated individuals from multiple population groups had already been extracted and analysed using multiple commercial CE-STR kits as part of a large-scale concordance study performed by Gabriella Mason-Buck (King's College London), and these formed the pool from which samples were selected for MPS analysis. Sections 2.1.2 and 2.1.3 provide brief details of the work that was done prior to the commencement of this PhD project in order to obtain extracted DNA ahead of library preparation as well as CE genotypes.

2.1.1. Sample selection

A minimum of 200 samples from each of the following population groups were selected: White British, British Chinese, North East African, South Asian and West African, from individuals who are resident in the United Kingdom. For the latter three populations, the predominant ancestries were India, Bangladesh and Pakistan (South Asian population group); Nigeria, Ghana and Caribbean-Jamaica (West African population group); Somalia and Ethiopia (North East African population group). Later, a smaller sample set of Middle Eastern ancestry (n=110) were also selected. Ancestry information for each individual was self-declared at the time of sample collection. Individuals gave informed consent for their DNA to be used for research purposes and ethical approval for this work was granted by the King's College London research ethics subcommittee (HR-16/17-2594).

2.1.2. DNA extraction

Before DNA can be analysed, it must first be extracted from a biological matrix (for example blood, saliva, or semen). DNA was extracted from the buccal swabs using one of two methods. Most samples were extracted using the Chelex method, which consists of cutting approximately three-millimetre lengths from each swab (Isohelix, Cell Projects Ltd, Kent, UK), and incubating them in 1 mL of de-ionised water at room temperature for 30 minutes. Samples were then centrifuged for 5 minutes at 15,600 g to pellet cellular material. 180 µL of 5 % Chelex® (Sigma-Aldrich, St Louis, MO, USA) solution was added to each sample, according to standard protocols for buccal swab reference profiles [176]. The samples were then subjected to a 20-minute step at 56 °C on a shaking incubator followed by an 8-minute step at 100 °C. A minority of samples were extracted using an alternative method, with the EZ1® DNA Investigator® Kit (Qiagen, Hilden, Germany) on the BioRobot EZ1 (Qiagen), using the pre-programmed DNA Investigator protocol card.

2.1.3. Capillary electrophoresis testing

CE data for the samples were obtained for the STR markers contained within the GlobalFiler® Express (Applied Biosystems, Foster City, USA) and PowerPlex® 16 HS (Promega, Madison, USA) kits as per manufacturer's guidelines [31, 177]. Input DNA of approximately 1 ng was used for amplification, and injection was performed at 1.2 kV for 23 seconds on the AB Prism 3130xl Genetic Analyzer (Applied Biosystems) for separation and detection of autosomal STR loci. A detection threshold of 50 relative fluorescence units (RFU) was imposed during data analysis using GeneMapper®IDX v1.4 software (Applied Biosystems).

2.2. Library preparation and massively parallel sequencing

Prior to sequencing with MPS, DNA samples must be prepared in a way that ensures that the target sequences have the necessary adapters and indices. This process is called library preparation, and the protocol used throughout this work was that of the ForenSeq™ DNA Signature Prep Kit, although some samples were analysed in a different manner when discrepancies were observed.

2.2.1. ForenSeq DNA Signature Prep Kit

The ForenSeq DNA Signature Prep Kit (Verogen, San Diego, USA) was used to prepare samples for sequencing [100]. An initial PCR was set up to amplify and tag regions of interest (STRs and SNPs) with an input volume of 5 µL for all samples apart from those extracted with the Qiagen DNA Investigator Kit, where 2 µL of sample and 3 µL of nuclease free water was added to the initial amplification step. Sample extracts were not quantified, but Chelex extracted samples are known within the lab to have an average quantification of 1 ng/ µL, and preliminary experiments showed that using 5 µL of extract gave good results. Similarly, the samples extracted using the Qiagen method generally have considerably higher quantification values than those extracted using Chelex, and initial results led to the decision to use 2 µL. Verogen recommends an input amount of 1 ng, with the risk of dropout if using less DNA [100]. Although adding more than 1 ng is not recommended, the normalisation step in the protocol ensures that excess DNA is removed prior to sequencing, leading to a negligible negative overall impact on results. In addition to the 5 µL of sample, each reaction contained 4.7 µL of a PCR reaction mix containing dNTPs, buffer, and other reagents necessary for the PCR process, 0.3 µL of enzyme mix, and 5 µL of primer mix. The final volume per reaction was of 15 µL. All samples were initially analysed using DNA Primer Mix A, which contains primers for the amplifications of 27 autosomal STRs, 7 X-STRs, 24 Y-STRs and 94 identity informative SNPs (listed in Table 2.2). A positive amplification control (2800M) and negative amplification control were run alongside the samples for each run. Samples were prepared in batches of 96 (including the two controls) using 96-well semi skirted PCR plates, as per manufacturer's guidelines when using Primer Mix A.

The next step of library preparation consists of a second PCR to ligate adapters and indices to the amplicons from PCR1. For each sample, 4 µL of each type of index adapter (i5 and i7) were added, in addition to 27 µL of PCR2 reaction mix, leading to a total volume of 50 µL per reaction. The ForenSeq DNA Signature Prep Kit contains eight different i5 index adapter tubes and twelve i7 index adapter tubes. Given that each i5 can be combined with one i7, there is a total of 96 possible unique combinations of indices. This means that a maximum of 96 reactions can be pooled at the end of the library preparation, and data isolated and allocated to their respective samples through bio-informatic sorting in downstream data analysis. For both PCR steps, a Veriti 96-well thermal cycler (Thermofisher, Waltham, USA) was used, and the cycling conditions are given in Table 2.1.

Table 2.1: Cycling conditions for PCR1 and PCR2

PCR1		8 cycles of:				10 cycles of:			
	98°C	96°C	80°C	54°C	68°C	96°C	68°C	68°C	10°C
	3 min	45 sec	30 sec	2 min*	2 min*	30 sec	3 min*	10 min	Hold
	**Specified ramp rate of 4%*								
PCR2		8 cycles of:							
	98°C	98°C	66°C	68°C	68°C	10°C			
	30 sec	20 sec	30 sec	10 min	10 min	Hold			

Following successful target amplification, tagging and addition of indices and adapters, the samples are referred to as libraries. The libraries were purified by incubating 45 µL of PCR product with 45 µL of magnetic purification beads. DNA binds to these magnetic beads, and by using a magnetic plate, unwanted reagents (unincorporated dNTPs, excess primers etc) were washed off in a two-step process using 80% ethanol. Purified libraries were subsequently eluted using 50 µL of elution buffer. Following purification, libraries were normalised to ensure equal representation of samples in the final pool, using a bead-based process which relies on saturation of beads and washing off of excess DNA. 20 µL of purified product was added to 45 µL of a master mix containing normalisation beads and library normalisation additives. Once the DNA bound to the beads, it was washed twice using a wash buffer and finally eluted off the beads using 0.1M NaOH, which also ensures that the libraries are single stranded. As with the previous steps, all reagents are provided with the ForenSeq DNA Signature Prep kit, and the steps followed are that of the Verogen protocol [100]. Finally, 5 µL of each library are added to a single tube.

The manufacturer protocol recommends combining 7 µL of these pooled libraries with 591 µL of hybridisation buffer [100], but following low cluster density being obtained for the first few runs, these volumes were adjusted to 12 µL of pooled libraries and 586 µL of hybridisation buffer. In addition, 2 µL of denatured Human Sequencing Control (Verogen) was added to the same tube to provide a control for the sequencing, independent of library preparation. This 600 µL reaction was subjected to a 2 minute incubation at 96°C to ensure all DNA strands are single stranded, and a 5 minute snap cool immediately prior to being loaded onto a sequencing cartridge for sequencing as described in section 2.2.2 [100, 114].

The entire process for library preparation was performed 15 times, including some re-runs to account for poor performing samples, to ensure results for at least 200 samples from each of the main five populations studied were obtained.

2.2.1.1. Primer mix B SNPs

In order to acquire ancestry-informative SNP genotypes for a subset of the samples, 47 samples from each group taken forward for concordance and frequency generation were selected from the samples described in 2.1.1. These were analysed using a custom primer mix provided by Verogen, containing only the primers for the 22 phenotype and 56 ancestry- informative SNPs usually found in primer mix B (Table 2.2). Library preparation was performed in batches of 96 reactions (including a positive and negative amplification control) and sequenced in three runs.

Table 2.2: List of all markers amplified using the ForenSeq DNA Signature Prep Kit

DNA Primer Mix B				
DNA Primer Mix A			Custom Mix	
aSTRs	X STRs	iSNPs	pSNPs	aSNPs
D1S1656	DXS10074	rs826472	rs28777	rs2238151
TPOX	DXS10103	rs964681	rs12203592	rs671
D2S441	DXS10135	rs10488710	rs4959270	rs1572018
D2S1338	DXS7132	rs1498553	rs683	rs2166624
D3S1358	DXS7423	rs2076848	rs1042602	rs7326934
D4S2408	DXS8378	rs901398	rs1393350	rs7997709
FGA	HPRTB	rs10773760	rs12821256	rs9522149
D5S818	rs10495407	rs2107612	rs12896399	rs200354
CSF1PO	rs1294331	rs2111980	rs2402130	rs12439433
D6S1043	rs1413212	rs2269355	rs1800407	rs1426654
D7S820**	rs1490413	rs2920816	N29insA	rs1800414
D8S1179	rs560681	rs1058083	rs1110400	rs735480
D9S1122	rs891700	rs1335873	rs11547464	rs12913832*
D10S1248	rs1109037	rs1886510	rs1805005	rs459920
TH01	rs12997453	rs354439	rs1805006	rs11652805
vWA	rs876724	rs1454361	rs1805007	rs17642714
D12S391	rs907100	rs4530059	rs1805008	rs2593595
D13S317	rs993934	rs722290	rs1805009	rs4411548
PentaE	rs1355366	rs873196	rs201326893	rs4471745
D16S539	rs1357617	rs1528460	rs2228479	rs2042762
D17S1301	rs2399332	rs1821380	rs885479	rs3916235
D18S51	rs4364205	rs8037429	rs2378249	rs4891825
D19S433	rs6444724	rs1382387	rs2814778	rs7226659
D20S482	rs1979255	rs2342747	rs3737576	rs7251928
D21S11	rs2046361	rs430046	rs7554936	rs310644
PentaD	rs279844	rs729172	rs10497191	rs2024566
D22S1045	rs6811238	rs740910	rs1834619	
Y STRs	rs13182883	rs8078417	rs1876482	
DYF387S1	rs159606	rs938283	rs260690	
DYS19	rs251934	rs9905977	rs3827760	
DYS385a-b	rs338882	rs1024116	rs6754311	
DYS389I	rs717302	rs1493232	rs798443	
DYS389II	rs13218440	rs1736442	rs12498138	
DYS390	rs1336071	rs9951171	rs1919550	
DYS391	rs214955	rs576261	rs1229984	
DYS392*	rs727811	rs719366	rs3811801	
DYS437	rs321198	rs1005533	rs4833103	
DYS438	rs6955448	rs1031825	rs7657799	
DYS439	rs737681	rs1523537	rs7722456	
DYS448	rs917118	rs445251	rs870347	
DYS460	rs10092491	rs221956	rs16891982*	
DYS481	rs2056277	rs2830795	rs192655	
DYS505	rs4606077	rs2831700	rs3823159	
DYS522	rs763869	rs722098	rs917115	
DYS533	rs1015250	rs914165	rs1462906	
DYS549	rs10776839	rs1028528	rs1871534	
DYS570	rs1360288	rs2040411	rs2196051	
DYS576	rs1463729	rs733164	rs6990312	
DYS612	rs7041158	rs987640	rs3814134	
DYS635	rs3780962		rs4918664	
DYS643	rs735155		rs1079597	
Y-GATA-H4	rs740598		rs174570	

*SNPs used for both phenotype and ancestry prediction

2.2.2. MiSeq FGx sequencing

The entire volume (600 µL) of pooled, diluted and denatured libraries for each run were loaded into the appropriate well of a single use, defrosted, MiSeq FGx cartridge, which was then inserted into the chiller compartment of the MiSeq FGx instrument. Each sequencing cartridge comes with a single use flowcell and sequencing buffer bottle, both of which were also loaded on the MiSeq FGx prior to each run (Figure 2.1). Prior to loading, the flowcell was cleaned using de-ionised water and lint-free paper wipes.

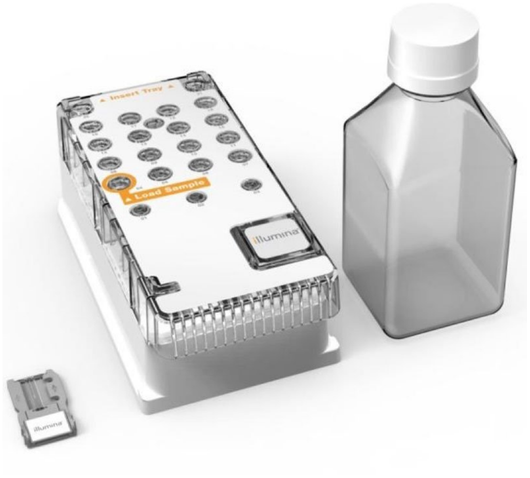


Figure 2.1: MiSeq FGx Reagent kit components – needed for MiSeq FGx sequencing
From left to right: flowcell, cartridge and buffer bottle.

A sample sheet containing the sample names and associated index combinations was created for each run using Microsoft Excel, saved as comma delimited format (.csv) and uploaded to the Universal Analysis Software (UAS) according to the layout described in the software protocol [122]. During run set up on the MiSeq FGx, the relevant sample sheet was selected after choosing to use the instrument in “Forensic Genomics” mode. The machine performs a series of automated checks such as ensuring that there is sufficient disk space and that all compartment doors are closed, before allowing the user to press start on the sequencing run. All ForenSeq DNA Signature Prep runs for the initial part of this project were performed using MiSeq FGx Reagent sequencing kits (Verogen), with a run time of approximately 28 hours. The runs for SNP analysis were performed using MiSeq FGx Reagent Micro sequencing kits (Verogen), with a run time of approximately 22 hours [178]. The chemistry of the two kits is identical, with the micro kit sequencing up to 5 million paired reads compared to the 12.5 million paired reads of the standard sequencing kits [178]. Given the reduced number of loci targeted by the custom panel, the micro kits were deemed to have sufficient output for the SNP runs, which was confirmed following analysis of the first of these runs. Once a run has completed, data is extracted from image files stored on the MiSeq FGx and transferred automatically to the UAS server for demultiplexing and secondary analysis.

2.2.3. Additional sequencing

To investigate the cause of null or imbalanced alleles at D5S818, D10S1248, D21S11 and Penta D, new primers were designed outside the range of the ForenSeq DNA Signature Prep kit amplicons. The full amplicon sequences for the ForenSeq alleles were obtained by manually aligning raw data as described further (2.3.1.3), and although the primer sequences are proprietary, the 5' end of both primers can be identified given they mark the end of the amplicons. The 3' ends of the primers fall within the amplicon and are unknown but estimated to fall within 20-25 bp inside the sequence. Custom primers were designed by copying the ForenSeq total amplicon sequence and specifying it as the minimum area to be amplified using the Primer3 software [179]. Default conditions were applied: optimum primer size of 20 bp; optimum primer melting temperature (Tm) of 60°C; and 40-60% target GC content. Primer sequences are shown in Table 2.3. Each PCR reaction was set up by combining 0.6 µL of diluted forward and reverse primers (5 mM) with 5 µL of QIAGEN Multiplex PCR Mix (QIAGEN, Hilden, Germany) 3.4 µL of nuclease free H_20 and 1 µL of template DNA (at 1 ng/ µL), leading to a final primer concentration of 0.3 µM. Cycling conditions used consisted of 95°C for 15 minutes followed by 30 cycles of 95°C for 15 seconds/ 60°C for 15 seconds/ 70°C for 30 seconds. PCR products were prepared for sequencing using the KAPA Hyper Prep Kit for Illumina Platforms (Roche, Basel, Switzerland) according to manufacturer's guidelines, using TruSeq indexes (Illumina, San Diego, USA), and sequenced on the MiSeq FGx in RUO mode, using a MiSeq v2 300 cycle (Illumina) cartridge.

Table 2.3: Primers used for additional sequencing

	Forward primer	Reverse primer
D21S11	ACTGCCAGCTTCCCTGATTC	AGCCATAAACACTGAGAAGGGA
D5S818	TCCCATCTGGATAGTGGACCT	GCTTCTAATTAAAGTGGTGTCCCA
D10S1248	GTAAAAAGCAAACCTGAGCATTAG	GGTGGGATACAGAGGTTTTAGCA
Penta D	GAAGGTCGAAGCTGAAGTG	TTGGGTTGTCTTATTGATGTG

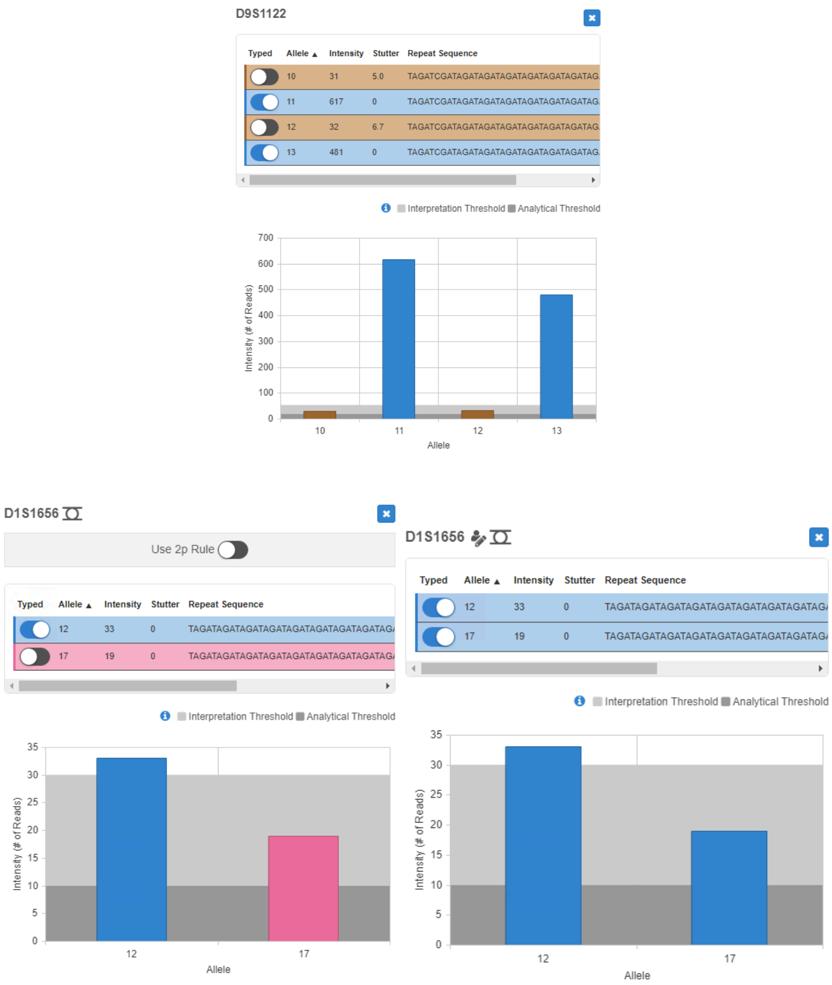
2.3. Data Analysis

Data analysis was performed sequentially and falls under four broad methods: concordance of sequenced STR genotypes with CE, sequence variant characterisation, SNP genotyping, and ancestry estimation. The methods used for each step of the data analysis process are listed and described in the following sections.

2.3.1. Concordance testing of autosomal STRs

2.3.1.1. Universal Analysis Software

Preliminary data analysis was performed using the ForenSeq™ Universal Analysis Software (UAS, Verogen) [122]. This software comes pre-installed on the server which accompanies the MiSeq FGx instrument, and contains a module tailored for the visualisation of STR and SNPS data generated using the ForenSeq DNA Signature Prep Kit and MiSeq FGx. The software calls alleles based on read counts, and makes use of an analytical and interpretation threshold, which are determined as a percentage of the total number of reads per locus. The thresholds used were 1.5% and 4.5% for the analytical and interpretation thresholds, respectively, taken from the developmental validation of the software [108]. The software also contains lower limits for both thresholds of 11 and 30 reads. For the purpose of this study, and as samples were all known to be single source with confirmed CE genotypes, autosomal STR alleles below the interpretation threshold and above the analytical threshold were manually called in the software and used for downstream concordance assessment and allele frequency calculations. Figure 2.2 shows an example of where an allele in a heterozygous profile at D1S1656 was manually called in between the two thresholds. Once all genotypes were manually verified in UAS, sample summary reports were exported in Microsoft Office Excel format.



D9S1122

Typed	Allele ▲	Intensity	Stutter	Repeat Sequence
○	10	31	5.0	TAGATCGATAGATAGATAGATAGATAGATAGATAG.
●	11	617	0	TAGATCGATAGATAGATAGATAGATAGATAGATAG.
○	12	32	6.7	TAGATCGATAGATAGATAGATAGATAGATAGATAG.
●	13	481	0	TAGATCGATAGATAGATAGATAGATAGATAGATAG.

D1S1656

Use 2p Rule

Typed	Allele ▲	Intensity	Stutter	Repeat Sequence
●	12	33	0	TAGATAGATAGATAGATAGATAGATAGATAGATAGA
○	17	19	0	TAGATAGATAGATAGATAGATAGATAGATAGATAGA

D1S1656

Typed	Allele ▲	Intensity	Stutter	Repeat Sequence
●	12	33	0	TAGATAGATAGATAGATAGATAGATAGATAGATAGA
●	17	19	0	TAGATAGATAGATAGATAGATAGATAGATAGATAGA

Figure 2.2: Example of locus genotypes in UAS

At D9S1122 (top), two alleles in a heterozygous profile were automatically called by UAS as they are both above the interpretation threshold (blue bars), with stutter sequences falling below the interpretation threshold and stutter filter (brown). At D1S1656, one allele in a heterozygous profile was above the analytical threshold but below the interpretation threshold (bottom left, in pink) and so was manually called (bottom right, in blue).

Following initial analysis in UAS, some sample results were deemed to be too poor to take forward. If a sample showed drop out at two or more loci used for concordance testing, it was re-extracted using the EZ1 DNA Investigator Kit and re-run. If the sample still showed multiple dropouts, it was removed from any further analysis. Results from 1018 samples were taken forward for concordance testing, with at least 200 from each of the following groups: White British (n=207), British Chinese (n=200), North East African (n=209), South Asian (n=200) and West African (n=202).

2.3.1.2. Allele comparison and concordance testing

Length-based allele calls from UAS for 22 autosomal loci and amelogenin were compared to CE results using in-house Microsoft Office Excel workbooks, and any discordant results recorded and further investigated. Discordance was defined as any instance where an allele observed using one technique was not observed with the other. Rarely, one allele in a known heterozygote profile was seen below the interpretation threshold in UAS while the other was not seen; this was considered allelic drop out rather than discordancy. In a casework scenario, the absence of one allele would not indicate discordancy in these instances given the fact that the other allele was present below the interpretation threshold. Given that the primary aim of this part of the project was to check concordance rather than performance of the kit, the genotypes with drop out were not considered moving forward for this purpose as they were not useful for assessing concordance. Instances of allelic drop out were primarily associated with poorer quality samples, however they were investigated further to determine whether there was a systematic issue that would affect the generation of accurate allelic frequencies – for example if longer alleles specifically were dropping out (see section 2.3.2.7).

Concordance was not assessed for the following loci: D4S2408, D6S1043, D9S1122, D17S1301, and D20S482, given they are not present in either of the CE STR kits used. Although D4S2408 is not one of the markers traditionally found in CE STR kits, a triple genotype in one North East African sample was verified using published primers [180] and amplified using the method for PCR described in section 2.2.3. Following amplification with FAM labelled primers, 1 µL of amplification product was combined with 10 µL of Hi-Di Formamide (Thermofisher) and 0.4 µL of ROX500 internal size standard (Thermofisher) and then analysed and visualised using the method for CE described in section 2.1.3.

D22S1045 is known to have poor heterozygous balance in the ForenSeq DNA Signature Prep Kit [100, 181], and drop out was frequently observed at this locus. Because of this, concordance was not assessed for this locus, and sequence-based allelic frequencies were not calculated, although observed sequence variants were still characterised as described in section 2.3.2. At Penta E, three instances of drop out occurred where the called allele was seen between 31 and 41 reads – manual realignment of the data as described in section 2.3.1.3 below revealed the secondary allele to be present below the 11 read analytical threshold and therefore these instances were not considered discordant from the CE genotypes. In all three cases, the alleles differed in size by at least 6 repeats (30 base pairs).

2.3.1.3. Additional sequencing analysis

For samples sequenced using the method described in section 2.2.3, data was analysed by exporting FASTQ files from the MiSeq FGx. These were aligned to a bespoke reference genome containing the flanking region of the STR sequences of interest using the mem algorithm within Burrow Wheeler Aligner (BWA) [182], and visualised using the Integrative Genomics Viewer (IGV) [183]. The same method was also used for some samples sequenced using the ForenSeq DNA Signature Prep Kit, when trying to visualise alleles below the 11 read cut off, by exporting FASTQ files from the UAS server.

2.3.2. Sequence characterisation and verification

2.3.2.1. UAS sample summary reports

Once genotypes were verified through comparison to CE results, allelic sequences for all samples were extracted from the same UAS reports in order to beginning characterising sequence variation within the repeat region of STRs. When exporting data from UAS, one option is to export sample summary reports, which contain the size-based allelic designation for the genotypes, as well as the repeat region sequence, for all alleles considered “typed” by the software (either automatically according to internal analysis thresholds, or manually after review). Sample summary reports were already downloaded and collated for concordance analysis, after which the sequences were extracted for alleles at all loci in all samples and collated using Excel workbooks.

2.3.2.2. Repeat region sequence characterisation and allele naming

In order to facilitate data visualisation, sequences were broken down according to repeat type and traditional allele designation in STRBase [17]. This was done using a VLOOKUP function in Excel and a compiled list of alleles observed in the growing database. When the sequence was unseen, it was added to the database. Given the lack of sequence-based naming nomenclature at the time this work was performed, each variant/allele was also assigned a short designator. This internal system is based on an initial number to designate the alleles by CE fragment length, followed by additional numbers to denote intra-sequence variation. For example, the following sequence at D12S391:

AGATAGATAGATAGATAGATAGATAGATAGATAGACAGACAGACAGACAGACAGACAGACAGAT

This sequence would be broken down into a bracket format: **[AGAT]$_8$ [AGAC]$_7$ AGAT** and given the internal allele name of "1601" because it is the first observed sequence version of an allele 16. The next version has the following bracket format: **[AGAT]$_9$ [AGAC]$_6$ AGAT** and is called "1602". This system isn't appropriate for large scale databasing, and it is expected that once a naming system has been agreed, it will be a simple exercise of replacing the internal names by those implemented by the forensic community [17].

2.3.2.3. STRaitRazor 2.0

In order to investigate sequence variation in the flanking regions of autosomal STRs, FASTQ files were extracted for all samples, and an analysis pipeline using a modified version of STRait Razor 2.0 (STR Allele Identification Tool - Razor) [127] was used for evaluation. This Perl-based bioinformatic software package can be used for variant analysis of raw STR-MPS data. Alleles are detected by matching areas in the forward and reverse flanking regions of STRs, using specific, known anchor sequences. The software allows for user-defined stringency parameters, and calls alleles by comparing the length for the repeat region with known allele lengths to ensure back-compatibility with CE results. Modifications for this work involved repositioning the anchor sequences to allow for additional analysis of the flanking regions [98]. An allelic balance threshold of 20% was used for analysis, in order to reduce the number of artefacts requiring manual inspection. Results from this method were verified against alleles characterised using the repeat region sequences extracted from the UAS sample summary reports.

2.3.2.4. UAS flanking region reports

Part-way through this project, Verogen released a new version of the UAS which enabled the generation of flanking region reports. These reports contain the full sequence string of each amplicon sequenced, meaning that data was now available for the flanking region as well as the repeat region of autosomal STRs. Flanking region reports were downloaded for all samples and sorted to remove stutter and sequencing artefacts according to the alleles already verified using repeat region variant characterisation above.

2.3.2.5. Flanking region sequence characterisation and allele naming

The results from the modified version of STRait Razor provided full sequence strings for all alleles, as well as a bracket annotation and the RS numbers for any SNPs in the flanking regions when known. Nomenclature described in the literature and the ISFG considerations [63, 74, 135] suggest that all markers should be reported on the forward strand, which is the nomenclature used by STRait Razor. UAS reports the following markers on the reverse strand: D1S1656, D2S1338, D5S818, D6S1043, D7S820, D19S433, CSF1PO, FGA, Penta E and vWA, so both the full string and the bracket allelic annotations were kept in both directions for these markers in order to ensure compatibility and enable comparisons with other publications. An additional concordance check was performed between the sequences validated using STRait Razor and the UAS flanking region reports.

Flanking region variation is defined as any change occurring outside of the repeat region of STRs, with analysis ranges for the ForenSeq DNA Signature Prep Kit provided in the supplementary file of Gettings et al. [136]. Flanking region SNPs were identified manually by comparing sequences in Excel, and characterised using the "Forensic STR Sequence Structure Guide" (available from https://strider.online/nomenclature) [184]. Any SNPs found that were not present in this guide were submitted to dbSNP in order to obtain an rs number [185]. For the following markers, a short portion of sequence directly adjacent to the repeat region is reported by UAS in the sample summary reports: D13S317, D18S51, D19S433, D1S1656, D5S818, D7S820, vWA. Because these count as part of the flanking region in the "Forensic STR Sequence Structure Guide", they have been referred to as "short flank" in this work.

Variants observed on the basis of changes in flanking region sequence were added to the internal database generated using repeat-region variation described in 2.3.2.2. The addition of a letter to the internal short designator naming system was used to denote variation in the flanking region of STRs.

2.3.2.6. Familias

Although all samples were selected on the basis that they were from unrelated individuals, an additional verification check for any genetic relatedness was performed. This assessment was carried out using the Blind Search functionality of the Familias software [186, 187], which was downloaded freely from https://familias.no/. This software uses DNA data (e.g. STR and/or SNP genotypes) to calculate which of two or more proposed hypotheses of genetic relatedness is most likely in the form of a likelihood ratio (LR). The 'Blind Search' tool can be used to search for specified relationships between any and all individuals within a DNA dataset. Pair-wise comparisons are carried out between each individual against all other individuals within the dataset to calculate an LR for a selected relationship (e.g. parent-child, full-siblings, half-siblings or first cousins) against an unrelated hypothesis. This function also allows the user to search for any direct matches amongst the DNA data, which would indicate duplicated samples. An initial search was carried out using STR genotypes and an LR threshold of 100 was used as a cut-off for all tested relationships. All pairs obtaining LR values under this threshold were considered to be unrelated, and all sample pairs with a value above this threshold were subjected to further investigation using available identity informative SNP genotypes. Any pairs still above an LR of 100 when including SNP data were considered to be related and one sample from the pair was removed from the dataset.

2.3.2.7. Allele frequencies

Once all sequence variants were characterised, sequence-based allelic frequencies were calculated and verified, in order to be able to use results from autosomal STRs sequenced with the ForenSeq DNA Signature Prep Kit and MiSeq FGx in statistical calculations for DNA typing. Frequencies can essentially be worked out by counting the number of observed alleles and dividing them by the total allele number. Arlequin software (v 3.5.2) was used to generate these frequencies [188]. Genotype data was formatted in Microsoft excel, so that sample names were in column A; frequency in column B (set to an absolute value of 1 for all genotypes as during the computations, Arlequin will compare all genotypes and recompute the frequencies [189]); the two alleles for the first STR locus in column C (across two rows); the two alleles for the second STR locus in column D and so on. Missing genotypes were replaced with “?”. In Arlequin, a new project was created following these steps: First, the “Microsat” option selected. Under “Data Type”, “Genotypic Data” was ticked, and under “Number of Samples”, the number of different populations being tested was inputted. As each population group’s frequencies were calculated separately, this was always set to 1. “TAB” was selected under “Locus Separator” and “?” for “Missing Data”. Once the project was created, genotype data formatted as described above was pasted in the open project under "Samples/Sample Data". Sample name and size were changed to the required values, before saving the project. Settings were selected to generate frequencies based on the genotypes inputted, and to check for non-random association of alleles by testing whether genotypes for each STR marker were in Hardy-Weinberg equilibrium (HWE) [23], applying a Bonferroni correction for multiple comparisons [190]. Doing this ensures that the expected number of homozygote and heterozygote genotypes are present within a dataset. Whilst loci not in HWE can be due to specific genetic factors such as selection or genetic drift, the primary benefit of verifying whether results were in agreement with HWE in this context is to check for serious genotyping errors, for example an excess of homozygotes which could indicate allelic drop out.

Extra quality control steps were taken when generating allele frequencies from population data containing allelic dropout to avoid bias. Allelic dropout was not random (the larger allele was more likely to drop out) so it would be unacceptable to include only the single allele for those samples that did amplify, however at the same time, removing the entire sample can also lead to bias if, for example, an 11,18 genotype is more likely to suffer allelic dropout than an 11,13 genotype. To ensure allele frequencies were as accurate as possible, frequencies for each size-based allele were calculated from the known CE results, and if allelic dropout was observed in the sequence data then the sequence frequencies were scaled accordingly. This adjustment made a negligible difference to allele frequencies apart from in D1S1656 where a disproportionately high percentage of dropout had affected the 17.3 allele and this correction allowed the true 17.3 frequency to be more accurately reported.

An example of this approach is described here for the D1S1656 marker in the White British population, where only 385 out of the expected 414 alleles could be detected in this sample set. From the 385 detected sequenced alleles, a total of 56 alleles were observed that corresponded to a '12' allele by CE. This was broken down into 20 alleles observed with the motif [TAGA]12 and 36 with the motif [TAGA]11TAGG, equating to frequencies of 0.052 and 0.094 respectively. The known '12' allele frequency from the CE data on this sample set was 0.139, however the sequence allele frequency is calculated as an inflated 0.146 (0.052+0.094) due to the preferential drop-out of higher molecular weight alleles. The sequence frequencies are therefore adjusted down to the known value of 0.139 keeping the ratio between the 2 sequence motifs that was observed for this ’12’ length allele, i.e. the adjusted frequency for [TAGA]12 is 0.050 (20/56*0.139) and for [TAGA]11TAGG is 0.090 (36/56*0.139). In this way, all allele frequencies can be adjusted for this marker within this population, the result generally being to give a minor boost to allele frequencies for higher molecular weight alleles and a slight reduction to allele frequencies for lower molecular weight alleles.

2.3.2.8. STRIDER

STRidER (STRs for Identity ENFSI Reference Database) is a curated online STR allele frequency population database, used to provide STR genotype probability estimates and quality control of autosomal STR data. All sequence-based alleles were submitted to STRidER (accession number: STR000292) for quality control verification prior to publication and in order to contribute to this STR population database [175]. Data was first formatted according to the submission requirements – one excel spreadsheet was compiled for each of the 26 STR loci, with the two length-based alleles and allelic sequences for each individual given on two lines. The submission also contained a file containing background information on the dataset, a spreadsheet of length-based (CE) genotypes, and finally a description of the QC procedures undertaken (Familias and Arlequin steps described previously).

2.3.2.9. FORSTAT

FORSTAT (forensic statistics analysis toolbox) [191] is a webtool available from https://fdl-uwc.shinyapps.io/forstat/, used for the evaluation of genetic markers. It can be used for calculations such as homozygosity, heterozygosity, match probability, power of discrimination, paternity index, power of exclusion etc. Prior to publication of the frequencies, all genotypes were formatted to GenePop input [192] and FORSTAT was used to investigate locus diversity and overall match probability of the marker set.

2.3.3. SNP analysis

2.3.3.1. Data extraction

As described in section 2.2.1.1, a subset of the samples from each population group studied were sequenced a second time using a custom panel containing primers for the phenotype and ancestry informative SNPs found in the ForenSeq DNA Signature Prep Kit DNA primer mix B. Analysis was not performed using UAS for this part of the study in order to be able to visualise and interpret data below the 11 read cut off applied by the software. Instead, FASTQ files were initially exported from the UAS server computer, before being renamed for ease of analysis. This involved creating a reference file containing the sample numbers and corresponding FASTQ name, which normally contains the index combination as an identifier (e.g. “R701-A501_S1_L001_R1_001”). RStudio [193] was then used to rename all FASTQ files and generate a new folder containing the renamed files.

2.3.3.2. SNP genotyping

The end file for SNP analysis is a variant call format (.VCF) file which is a standard output used in MPS genotyping, and highlights the target SNPs and variants compared to a reference genome. In order to generate these files, a script was used to undertake the following steps:

1. BWA: The sequences in the FASTQ files were aligned to a reference file containing the target sequences of interest (DNA primer mix B SNPs) using the mem algorithm within BWA [182]. This reference file was created by searching for the target SNPs on dbSNP [185] and taking approximately 100 bases on either side of the SNP. For the HIrisPlex SNPs, the sequence between the published primers was used [147]. BWA creates sequence alignment map (SAM) files that are aligned to the reference sequences provided.
2. SAMTools: SAM files are converted to BAM files, sorted indexed and intermediary files removed using SAMTools [194].
3. GATK: The Genome Analysis Toolkit [195] is then used to highlight all variants to the reference sequences in the initial reference file.

A final RStudio script was used to modify the files into a more user-friendly format, such as providing heterozygous allele balance and adding conditional formatting to the values in the spreadsheet for easier visualisation. The files generated from this script were Excel files and were then collated to provide all genotypes for all samples in a run in a single workbook.

Results were manually verified using a set of genotyping “rules”, namely:

- Minimum number of reads to consider an allele genuine: 3 (i.e., an allele with 2 or less reads was considered drop out).

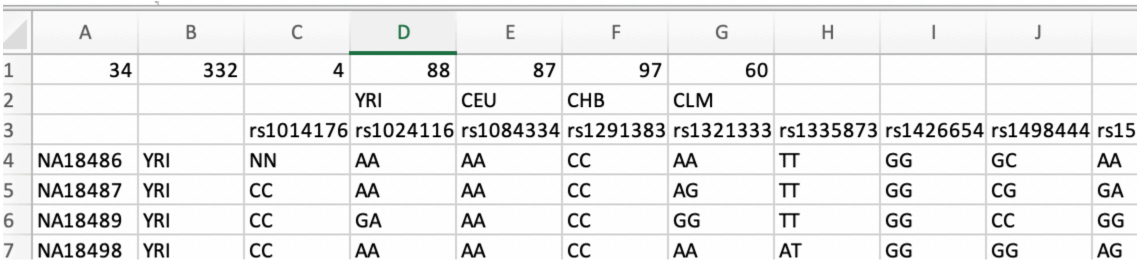
- Minimum total number of reads to consider a locus typed: 6 (i.e., a locus with less than 6 reads overall was considered drop out).
- Minimum number of read to consider a homozygous genotype genuine: 20 (i.e., a locus with one allele present at 19 reads or less was considered potentially heterozygous with possible drop out).
- Heterozygous balance: anything below 0.85 was considered imbalanced.
- Samples with 15 or more poor genotypes (imbalanced, or where drop out has occurred) were removed from further analysis.

Based on these criteria, a final number of 266 samples were taken forward for SNP ancestry analysis (White British, n=42; South Asian, n=39; North East African, n=45; British Chinese, n=47 samples; West African, n=47 and Middle Eastern, n=47).

2.3.4. Ancestry analysis

2.3.4.1. Data formatting

Verified genotypes were collated according to what was being analysed (loci type, populations etc) in Microsoft Excel spreadsheets. Additional information was added to the first line of the data sheet as follows and as shown in Figure 2.3: A1= number of loci; B1= number of samples; C1= number of populations; D1- G1= samples per population.



	A	B	C	D	E	F	G	H	I	J	
1	34	332	4	88	87	97	60				
2				YRI	CEU	CHB	CLM				
3			rs1014176	rs1024116	rs1084334	rs1291383	rs1321333	rs1335873	rs1426654	rs1498444	rs15
4	NA18486	YRI	NN	AA	AA	CC	AA	TT	GG	GC	AA
5	NA18487	YRI	CC	AA	AA	CC	AG	TT	GG	CG	GA
6	NA18489	YRI	CC	GA	AA	CC	GG	TT	GG	CC	GG
7	NA18498	YRI	CC	AA	AA	CC	AA	AT	GG	GG	AG

Figure 2.3: Example of data formatted for downstream ancestry analysis

GenAlEx (Genetic Analysis in Excel) was downloaded freely as an add-on to Microsoft Excel [196]. This was used to format the data for both STR and SNP results into a format that would work for downstream ancestry analysis. Two main functions were used:

- Split data (genotypes in two columns): Manage data > edit raw data > split codom
- Change letters for numbers: manage data > edit raw data > alpha to numeric

The resulting dataset contained the following for each analysis: column A contained the sample name; column B the population identifier (1 for White British, 2 for British Chinese, 3 for North East African, 4 for South Asian, 5 for West African – these identifiers were mainly used for subsequent graphical display in CLUMPAK); column C contained the first allele for the first genetic marker (STR or SNP); column D contained the second allele for the same marker; columns E and F contained the first and second allele for the second marker and so on. The “split codom” function of GenAlEx was used to split the data for the individual markers into two columns, and the “alpha to numeric” function represent alleles as numbers rather than letters. Any missing data was represented by -9. The spreadsheet was exported to “STRUCTURE format”, using GenAlEx, as a “tab delimited text file”.

2.3.4.2. STRUCTURE

Ancestry estimation was performed using the program STRUCTURE [171], which uses a model-based clustering algorithm to infer population structure from multi-locus genotype data. The underlying algorithm looks for K different genetic signatures within a dataset of individuals. Individuals are then assigned to one of the K clusters using either a non-admixture model which assigns individuals in a yes/no approach to each cluster, or by breaking down each individual’s results by the proportion of ancestry which can be assigned to each of the K populations. If an individual is of mixed ancestry, using the admixture model of clustering would results in components being attributed to both/ multiple ancestral populations. A graphical display of the admixture model results shows each individual as a single vertical line, and the membership of proportion to each inferred K group is represented by splitting this line into different colours. The model uses a Bayesian approach to discerning K genetic clusters within the datal, through the use of allelic frequencies. These frequencies are assumed to be in Hardy-Weinberg equilibrium and genetic markers are assumed to be in linkage equilibrium.

Version 2.3.4 of the STRUCTURE software is freely available and was downloaded from https://web.stanford.edu/group/pritchardlab/structure.html and used throughout this work. For each STRUCTURE run, a new project was created, and the appropriately formatted dataset file uploaded. When prompted, information was entered regarding number of individuals and markers present in the data, and options were ticked specifying that data included marker names in row 1 and that data was stored on a single row for each individual (as opposed to on two rows). Options were also ticked to denote that samples names were entered in column A, and putative populations in column B. Parameters were set at 100,000 for burn-in and 100,000 Markov Chain Monte Carlo repetitions, using the admixture model of analysis. K values were set depending on the data being analysed.

2.3.4.3. CLUMPAK

Results from STUCTURE were displayed graphically using the program CLUMPAK (Clustering Markov Packager Across K) [197]. According to instructions, the results folder generated by STRUCTURE was zipped, and then uploaded to http://clumpak.tau.ac.il/ alongside a label file denoting which population identifier corresponds to which ancestral population, and a colours file to assign each cluster a specific colour. Once a CLUMPAK run was completed, a results folder and PDF file containing graphical display of data were downloaded. An example output graph from CLUMPAK can be seen in Figure 2.4.

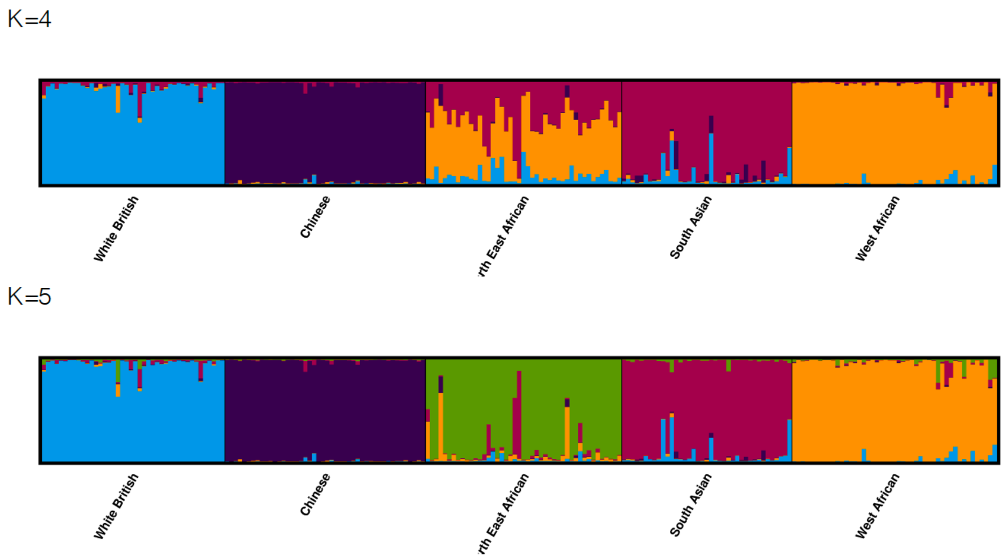


Figure 2.4: Example of CLUMPAK graphical display

These plots were generated using results from STRUCTURE for 989 individuals using aSNP data for K=4 and K=5 population groups.

2.3.4.4. Rosenberg informativeness for assignment measure

The informativeness for assignment (I_n) measure is used to determine the amount of information that multi-allelic markers provide about individual ancestry [198]. Rosenberg et al. proposed this value with the purpose of reducing the genotyping required for ancestry inference, i.e., using a smaller subset of markers of highest informativeness will reduce the number of markers needing to be targeted whilst achieving the desired result in terms of ancestry estimation.

Infocalc is a script used to calculate statistics that measure the ancestry information content of genetic markers [198], including the I_n measure. This perl script can be downloaded from https://rosenberglab.stanford.edu/infocalc.html. Genotypes formatted for STRUCTURE were used as they were already purely numerical, and after removing all spaces in the spreadsheet, the sheet was saved as a .STRU file extension. This was done for all genotypes for the five population groups, and then individual input files containing just combinations of populations (e.g. White British and British Chinese, British Chinese and West African and so on.) were also saved. In the end, one input file containing genotypes for all five populations and 10 input files containing pairs of populations were used for length-based alleles, repeat-region variant alleles and flanking-region variant alleles (resulting in 33 input files).

In a command line, the directory was changed to the folder containing the input file, and the following was inputted:

```
./infocalc -input infile.stru -numpops 5
```

For each input file, the “infile.stru” was changed to the appropriate name (e.g. 5_pop_infile.stru) and the number of populations investigated was modified accordingly (either 5 or 2 populations at a time). The analyses were performed without a weightfile, meaning that each population is equally likely to be the source population. Given numbers of samples per population were practically identical, this was considered acceptable. The output file created for each run of Infocalc was copied into an excel spreadsheet.

2.3.4.5. FROG-kb

The genotypes for the 55 SNPs described in Kidd et al. [105] were extracted from the 56 SNP results provided by UAS for a few specific samples from each population for review. The FROG-kb website (https://frog.med.yale.edu/FrogKB/index.jsp) provides the ability to calculate relative likelihoods of ancestry from different reference populations for uploaded aiSNP genotypes, derived underlying allelic frequencies from the ALlele FREquency Database (ALFRED, http://alfred.med.yale.edu). The “AISNP” radio button allows the user to enter genotypes of an individual at multiple SNPs, and the probabilities of that multisite genotype in each of several populations is calculated and provided. The “KiddLab - Set of 55 AISNPs” option was selected, followed by “Data Entry”, “File Upload” and finally “Input Genotype for a Panel”. Here, the genotypes for the 55 SNPs for a sample was pasted into the box, formatted so that the first line starts with 'ai55' (to denote an ancestry inference 55 SNP Set). The second line on the file contains the column heads (ALFRED_UID, dbSNP_rsnumber, chrom, chrom_pos, alleles, genotype), and the rest of the line contain the appropriate data. The results obtained were copied and pasted into Excel for graphical rendering.

2.3.4.6. Snipper

The length-based genotypes for several samples were uploaded to the program *Snipper*, accessed from http://mathgene.usc.es/snipper/frequencies_new.html, in order to classify them using a frequency-based training set of 32 markers. Frequencies for 22 out of the 27 autosomal STR markers targeted by the ForenSeq DNA Signature Prep kit are available as part of this set, so profiles were modified by removing data for the additional 5 loci (D4S2408, D6S1043, D9S1122, D17S1301 and D20S482) and formatted according to the requirements. For each input profile to be classified, markers are separated by slashes, both alleles of each marker by commas – as in the example below.

9,10/15,15/13,15/19,20/15,17/11,13/10,11/14,17/13,14/17,18/8,12/12,12/14,16/14,16/30,31.2/15,15/24,29/6,9.3/8,8/15,16/9,11/9,14

2.3.4.7. PopAfiliator

Length-based genotypes for several samples were manually input into the PopAfiliator 2 website (http://cracs.fc.up.pt/~nf/popaffiliator2/) for population estimation, for the following loci: CSF1PO, D2S1338, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, D19S433, D21S11, FGA, Penta D, Penta E, TH01, TPOX, vWA.

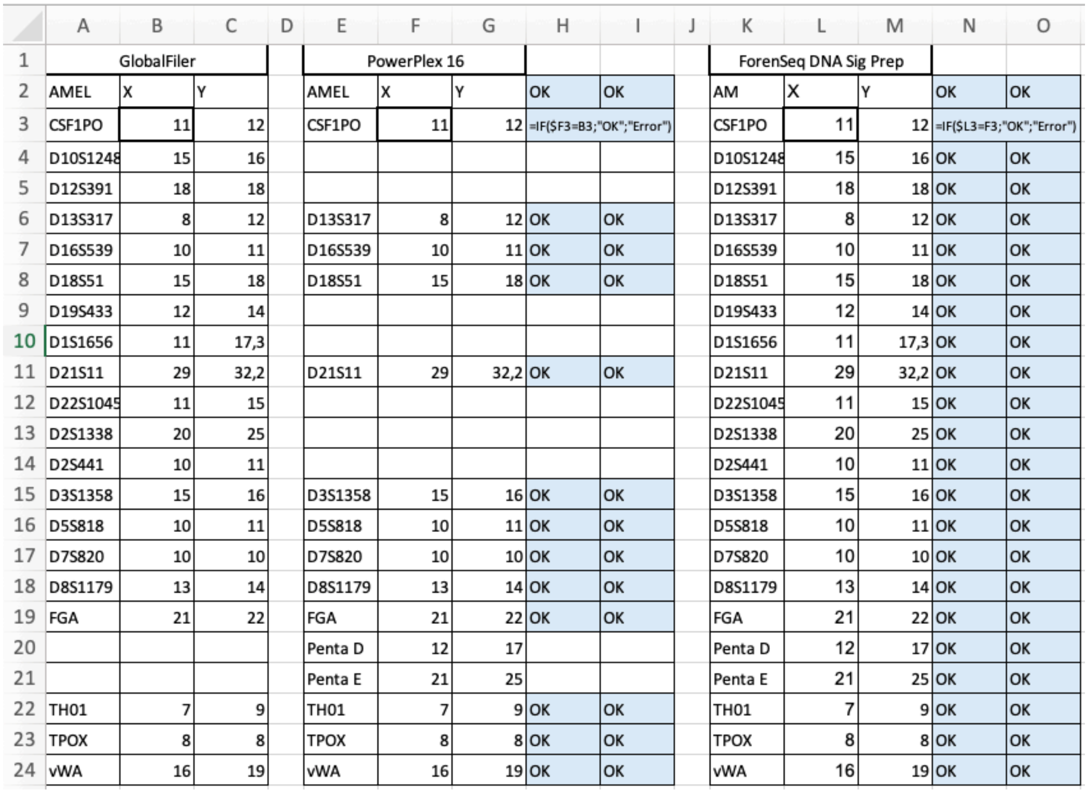
3. CONCORDANCE WITH CAPILLARY ELECTROPHORESIS AND SEQUENCE-BASED ALLELE CHARACTERISATION

3.1. Concordance with capillary electrophoresis

Before a new STR kit can be incorporated into the routine work of a forensic casework laboratory, its concordance with the previously implemented method must be demonstrated [131, 132]. Many commercial STR kits target the same markers, but differing kit configurations can mean different primer sequences are used to amplify the same locus, and therefore different primer binding sequences around the STR repeat region are targeted [112]. Concordance testing involves comparing results for identical markers obtained by different kits on the same set of samples. This kind of study can highlight a number of possible causes for discordances between kits. A mutation in the primer binding site for one kit can lead to inefficient amplification that may skew heterozygous genotypes, or in more drastic cases could cause a null allele (i.e., allele drop out). If a different kit targets a different primer binding region, this mutation would not affect amplification for the same sample, leading to a discordant result between kits. Similarly, the difference in primer design amongst kits may lead to discordant results if there is an insertion or a deletion in the flanking region of an STR locus (the area outside of the repeat region which is still captured within the amplicon) which is only included within the amplicon length amplified by one set of primers. Discordances can impact DNA databases and, in extreme cases, lead to incorrect comparisons. For example, if a profile recovered from a crime scene was compared to a suspect reference profile, it is important that the methods used to generate both profiles are considered concordant. The lack of a match between two genuinely identical DNA samples is a routine inconvenience to forensic laboratories. Databases deal with this by having algorithms for "near-miss" reports, which show profiles that nearly match the one being searched – which are useful for taking into account possible clerical errors or discordancy between kits. Despite the use of near-miss reports, it is vital to understand the scope and scale of discordancy between commercial kits. Where discordant results are found, these can be listed on a database, or information provided to the manufacturer to enable a reconfiguration of primer design. It is especially important to check the concordance of MPS with CE-based technologies, as the manner in which alleles are designated is so inherently different. Concordance studies for the ForenSeq DNA Signature Prep kit have already been undertaken as part of a developmental validation of the kit [108] and as part of several external studies [73, 74]. Results have shown a high level of concordance between this kit and CE-based STR multiplexes, although the diversity of samples used in these studies have been limited both in terms of numbers and population groups investigated.

Although the process of library preparation and amplicon detection differs from traditional STR-CE kits in a multitude of ways, the underlying principle of the ForenSeq DNA Signature Prep kit still relies on PCR amplification of target STR sites, just like any CE-based kit. This means that prior to adoption and implementation of this kit, it was important to show that it is concordant with profiles, and by extension DNA databases of any kind, generated using CE kits. Most population studies aiming to investigate concordance and/or generate allelic STR frequencies are conducted using at least 200 samples per population group, with the most recent guidelines for population data submission requiring data from a minimum of 500 samples per population group for CE-generated data, and 50 samples per population group for MPS-generated data [174, 175]. In this work, following initial analysis in the universal analysis software (UAS), results from 1018 samples were taken forward for concordance testing, with at least 200 from each of the following groups: White British (n=207), British Chinese (n=200), North East African (n=209), South Asian (n=200) and West African (n=202).

Length-based allelic designations for all alleles genotyped using the ForenSeq DNA Signature Prep kit and MiSeq FGx were compared to genotypes obtained for the STR markers contained within the GlobalFiler Express and PowerPlex 16 HS kits using a simple comparison tool designed using Microsoft Excel, as shown in Figure 3.1.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	GlobalFiler				PowerPlex 16						ForenSeq DNA Sig Prep				
2	AMEL	X	Y		AMEL	X	Y	OK	OK		AM	X	Y	OK	OK
3	CSF1PO	11	12		CSF1PO	11	12	=IF($F3=B3;"OK";"Error")			CSF1PO	11	12	=IF($L3=F3;"OK";"Error")	
4	D10S1248	15	16								D10S1248	15	16	OK	OK
5	D12S391	18	18								D12S391	18	18	OK	OK
6	D13S317	8	12		D13S317	8	12	OK	OK		D13S317	8	12	OK	OK
7	D16S539	10	11		D16S539	10	11	OK	OK		D16S539	10	11	OK	OK
8	D18S51	15	18		D18S51	15	18	OK	OK		D18S51	15	18	OK	OK
9	D19S433	12	14								D19S433	12	14	OK	OK
10	D1S1656	11	17,3								D1S1656	11	17,3	OK	OK
11	D21S11	29	32,2		D21S11	29	32,2	OK	OK		D21S11	29	32,2	OK	OK
12	D22S1045	11	15								D22S1045	11	15	OK	OK
13	D2S1338	20	25								D2S1338	20	25	OK	OK
14	D2S441	10	11								D2S441	10	11	OK	OK
15	D3S1358	15	16		D3S1358	15	16	OK	OK		D3S1358	15	16	OK	OK
16	D5S818	10	11		D5S818	10	11	OK	OK		D5S818	10	11	OK	OK
17	D7S820	10	10		D7S820	10	10	OK	OK		D7S820	10	10	OK	OK
18	D8S1179	13	14		D8S1179	13	14	OK	OK		D8S1179	13	14	OK	OK
19	FGA	21	22		FGA	21	22	OK	OK		FGA	21	22	OK	OK
20					Penta D	12	17				Penta D	12	17	OK	OK
21					Penta E	21	25				Penta E	21	25	OK	OK
22	TH01	7	9		TH01	7	9	OK	OK		TH01	7	9	OK	OK
23	TPOX	8	8		TPOX	8	8	OK	OK		TPOX	8	8	OK	OK
24	vWA	16	19		vWA	16	19	OK	OK		vWA	16	19	OK	OK

Figure 3.1: Example of the allele comparison spreadsheet used for concordance testing *Example given for one sample. Where data is available for the two CE kits and the ForenSeq DNA Signature Prep kit, all three were compared (as seen on line 3 for CSF1PO). If the locus is only amplified in one CE kit, the comparison was made with that kit.*

A concordance rate exceeding 99% was observed for the STR markers compared with CE data in all five populations. This value is comparable to that observed when comparing CE-based kits [132], and suggests that the ForenSeq DNA Signature Prep kit is compatible with current technologies. Discordances were detected at D5S818 in the White British (n=1/207) and South Asian population (n=1/200), at D7S820 in the White British population (n=1/207) at Penta D in the South Asian population (n=1/200) and at D21S11 in the North East African (n=3/209) and West African (n=1/202) populations. Further details are given for each discordance event in the following subsections.

Data for the STR locus D22S1045 was discounted early on due to severe heterozygote imbalance. This marker is known to perform poorly in this kit [199], also seen by the manufacturer, and a note about interpretation of homozygous genotypes is included in the protocol [122]. Because homozygous genotypes were not reliable, D22S1045 was not analysed in the context of concordance or for frequency generation, although sequences observed for this locus were characterised.

3.1.1. Discordances at D5S818

D5S818 is a simple tetranucleotide repeat marker composed of a core [AGAT] repeat unit, with alleles ranging in size from 7-16 seen in this work. The D5S818 genotype for one White British sample was reported by UAS as a homozygous 11 genotype at D5S818, whereas a heterozygous 9, 11 genotype was observed using the GlobalFiler and PowerPlex 16 CE kits. Custom primers were used to sequence a larger amplicon containing the primer binding regions, and the null allele was found to be caused by a SNP 22 bp away of the traditionally defined repeat region, in the reverse primer binding site of the ForenSeq DNA Signature Prep kit primer. The sequence of this null 9 allele is shown in Figure 3.2, with the base change generating the null allele highlighted. This mutation does not have an assigned RS number and is therefore assumed to be rare.

tgattttcctctttggtatccttacgtaatattttgaAGATAGATAGATAGATAGATAGATAGATAGATAGATagag
gtataaataaggatacacataaagatacaaatgttgt

Figure 3.2: G>C mutation in a null 9 allele at D5S818

The traditionally defined repeat region of D5S818 is underlined and sequence given in capitals. Flanking region sequence is provided in lower case, with the G>C mutation highlighted in red.

The discordance at D5S818 in a South Asian sample was caused by a null allele in the ForenSeq DNA Signature Prep kit result, where one allele in a heterozygous profile had 37 reads, whilst the other allele (allele 13) was not detected. This sample was sequenced using primers designed to fall outside of the ForenSeq DNA Signature Prep kit primer binding sites and a SNP was observed in the reverse primer binding region: rs25768 [184]. This primer binding site SNP is well characterized and is an example of where the alternate G allele is considerably more common than the A allele observed in the reference genome, with global frequencies of 0.157 for the A allele and 0.843 for the G allele according to Phase 3 of the 1000 Genomes Project data [45]. Figure 3.3 highlights the position of the rs25768 SNP, 13 bp from the repeat region. Because D5S818 is one of the markers reported on the reverse strand by UAS, the SNP is visualised as a T base, although it is still referred to as a G/A SNP here in order to stay consistent with Phillips et al. [184].

agggtgattttcctctttggtatcctta**t**gtaatattttgaAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATagaggtataaataaggatacagataaagatacaaatgttgtaaactgtggctatgattggaatca

Figure 3.3: rs25768 SNP minor allele in a null 13 allele at D5D818

The traditionally defined repeat region of D5S818 is underlined and sequence given in capitals. Flanking region sequence is provided in lower case, with the mutation shown in red.

The A base was associated with the null allele in the single South Asian sample but upon further investigation was also found to be associated in all populations with numerous instances of heterozygous imbalance. Heterozygous balance was measured at this locus by dividing the intensity of the minimum intensity allele by the intensity of the maximum intensity allele in all samples. A heterozygous balance ratio of under 0.6 is flagged by UAS as an imbalanced genotype [122] and was observed 223 times in this dataset (n= 1018) at D5S818. The average heterozygous balance ratio for these genotypes was 0.43. Additional sequencing of a subset of the samples with this imbalanced genotype revealed an "A" allele at the rs25768 SNP within the D5S818 reverse primer binding site of the underrepresented allele in each case. Interestingly, the affected alleles in the imbalanced genotypes were all 10-14 alleles based on length-based designation, but all had the same repeat structure of $[AGAT]_n$ AGAG (including the "short flank") rather than the alternative $[AGAT]_n$ AGAT motif.

When visualising the sequencing reads using IGV, it became apparent that there is a redundant primer included within the ForenSeq DNA Signature Prep Kit to counteract the presence of this primer binding-site SNP. In these cases, this was evidently not sufficient to allow for efficient amplification, leading to lower coverage for certain alleles, at the most extreme level leading to the allelic dropout.

3.1.2. Discordance at D7S820

D7S820 is a simple tetranucleotide repeat marker composed of repeats of a core [GATA] unit, with alleles ranging in size from 6-14 observed in this work. A White British sample presented with a 7 allele at D7S820, despite it being genotyped as a 6.3 with several CE-based methods [31-33]. Here, raw data was re-analysed in order to look at the sequence outside of the repeat region for this allele. The discrepancy was determined to be due to a rare deletion (rs540346880) found in the flanking region of the 7 allele [65, 200] as shown in Figure 3.4. This variant has been described in the literature, with a frequency of less than 0.01% in a Caucasian population [201]. In the case of the CE-based methods, the deletion would have caused the amplicon to be 1 base pair shorter than expected, hence the resulting 6.3 genotype. If flanking regions were considered during allele designation, it would have been immediately apparent that this would be a 6.3 allele according to length-based allele calling. This issue raises the question regarding whether flanking region information should be reported for all markers when using MPS, to facilitate a nomenclature that offers full back-compatibility with CE-based methods.

ataaagggtatgatagaacacttgtcatagtttagaacgaactaacGATAGATAGATAGATAGATAGATAGATaga
cagattgatagttttttttaatctcactaaatagtctatagtaaacatttaattaccaatatttggtg

Figure 3.4: rs540346880 deletion in the flanking region of an allele 7 at D7S820

The repeat region of D7S820 is underlined and sequence given in capitals. Flanking region sequence is provided in lower case, and the deletion is shown in red.

3.1.3. Imbalance at D10S1248

D10S1248 is a simple tetranucleotide repeat marker composed of a core [GGAA] unit, with alleles ranging in size from 9-18 seen in this work. Whilst no discordant results were obtained at D10S1248, heterozygous imbalance was noticed in some South Asian samples. Five instances where the heterozygous genotype balance was recorded as being less than 0.3 were detected and in every case the allele with a reduced read number was observed to be a 13. Sequencing with primers set outside of the ForenSeq DNA Signature Prep kit amplicon revealed a C to T mutation in these five 13 alleles that interferes with the binding of the reverse primer, as shown in Figure 3.5. This SNP has been previously characterised: rs531980552 [184], with a frequency of 0.012 for the T allele in the South Asian population according to Phase 3 of the 1000 Genomes Project data [45]. Manual alignment of the sequencing reads showed that there is no redundant primer present to account for this SNP.

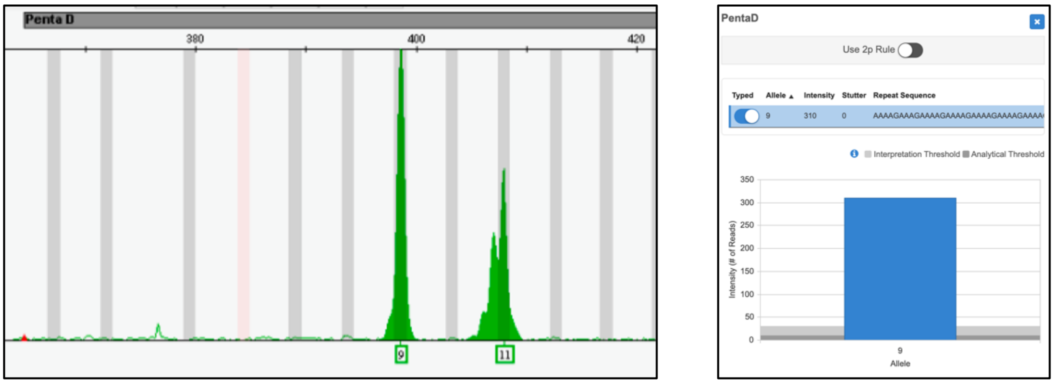
gaccaatctggtcacaaacatattaatgaattgaacaaatgagtgagtGGAAGGAAGGAAGGAAGGAAGGAA
GGAAGGAAGGAAGGAAGGAAGGAAGGAAatgaagacaatacaaccagagttgtttctttaataa

Figure 3.5: rs531980552 C to T mutation at D10S1248

The repeat region of D10S1248 is underlined and sequence given in capitals. Flanking region sequence is provided in lower case, and the C>T mutation is shown in red.

3.1.4. Discordance at Penta D

Penta D is a simple pentanucleotide repeat marker composed of a core [AAAGA] repeat unit, with alleles ranging in size from 2.2-17 observed in this work. For one South Asian sample, a heterozygous 9, 11 profile was obtained with CE, whereas a 9 homozygous profile with 310 reads was seen in UAS. The 11 CE peak was abnormal, as shown in Figure 3.6, with a very large shoulder on the left-hand side of the peak. Sequencing with custom primers showed that there was no mutation associated with the 11 allele in the area of sequence assumed to correspond with the ForenSeq DNA Signature Prep kit primers, however there was a G>A change just before the repeat unit that created a homopolymer run of 12 A bases. This seems to have caused an unusual stutter, explaining the shoulder on the CE 11 allele, which corresponds to an amplicon product with only 11 A bases in this poly-A area. It would appear that this long homopolymer stretch, combined with the subsequent AAAAG repeat, is either causing problems in the sequencing or alignment of this allele.



Typed	Allele ▲	Intensity	Stutter	Repeat Sequence
●	9	310	0	AAAAGAAAGAAAAGAAAAGAAAAGAAAAGAAAA

Figure 3.6: Penta D discordance

Instance of discordance observed in one South Asian sample, with a 9/11 genotype with CE (left) but a 9 homozygote in UAS, with 310 reads (right). The 11 CE peak is abnormal, with a large shoulder on the left-hand side of the peak.

3.1.5. Discordances at D21S11

D21S11 is a complex tetranucleotide repeat marker that follows this motif: [TCTA]$_n$ [TCTG]$_n$ [TCTA]$_n$ **TA** [TCTA]$_n$ **TCA** [TCTA]$_n$ **TCCATA** [TCTA]$_n$, where the bases in bold are not counted towards the length-based designation of alleles. In this work, alleles ranging in size from 24.3-37 were observed. The four recorded occurrences of discordance at this locus all involved drop out of a 24.3 allele observed with CE. Sequencing with custom primers revealed that the 24.3 allele is in fact a 28 allele with a thirteen base pair deletion in the flank which would appear to be causing the drop out. The flanking region sequence for the 28 allele was obtained from Phillips et al. [184] and is shown in Table 3.1. There is very limited flanking region sequence provided even by the UAS flanking report for this marker (just “CTATCTAT” on the 3’ end of the amplicon), further indicating that the reverse primer is likely to bind close to the repeat region of D21S11. As no other 24.3 alleles were seen in any populations with the ForenSeq DNA Signature Prep kit, it is possible that this deletion is predominantly associated with 24.3 alleles. The alleles completely dropping out is likely due to the fact that the deletion is found 26 bases inside the 3’ end of the amplicon, suggesting it is towards the 5’ end of the primer binding site.

Table 3.1: Sequence of null 24.3 alleles at D21S11

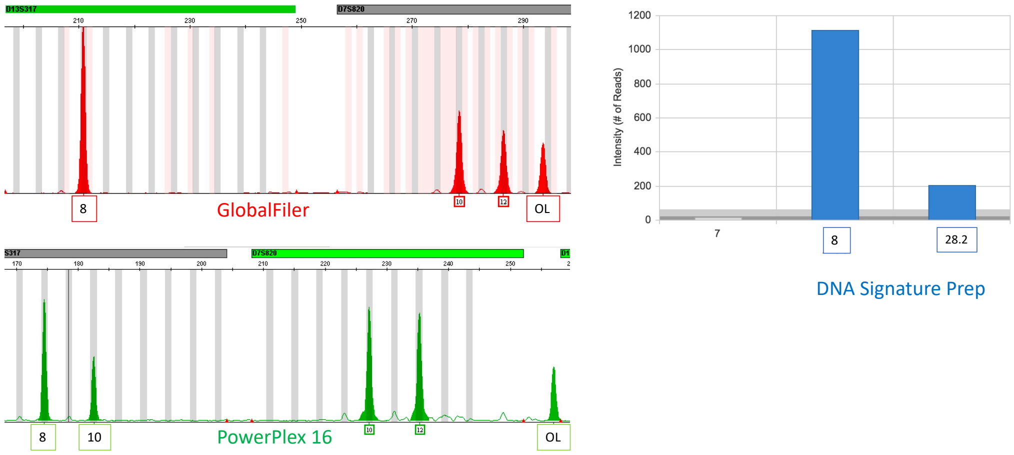
Sequencing the null 24.3 allele obtained with the ForenSeq DNA Signature Prep kit reveals a 28 allele with a 13 bp deletion, shown striked through, in the flanking region, which likely falls within the primer binding site.

Allele	Repeat Region	Flanking Region
2803	[TCTA]5 [TCTG]6 [TCTA]3 **TA** [TCTA]3 **TCA** [TCTA]2 **TCCATA** [TCTA]9	tc gtctatctatcca gtctatctacc
24.3	[TCTA]5 [TCTG]6 [TCTA]3 **TA** [TCTA]3 **TCA** [TCTA]2 **TCCATA** [TCTA]9	tc ~~gtctatctatcca~~ gtctatctacc

3.1.6. CE kit discordance at D13S317

In addition to the discordances described above in the ForenSeq DNA Signature Prep kit, a discordance was observed between two CE kits, GlobalFiler and PowerPlex 16, in a West African sample at D13S317. This marker is a simple tetranucleotide repeat marker composed of a [TATC] repeat unit. As shown in Figure 3.7A, an 8 allele was obtained with GlobalFiler, whilst alleles 8 and 10 were obtained with PowerPlex 16. An out-of-bin allele was also observed in both profiles, corresponding to a 28.2 allele when extrapolated from the highest ladder allele. Sequencing this sample revealed the ForenSeq DNA Signature Prep kit profile to be a heterozygous 8/ 28.2, resulting in a concordant genotype with GlobalFiler, when taking into account the out-of-range 28.2 allele. Sequence-level data suggests the underlying reason for the discordance between the two CE kits. Figure 3.7B shows the allelic sequences obtained from the ForenSeq DNA Signature Prep kit data, where it can be observed that the 28.2 allele is formed from a partial fusion of the repeat region and flanking region of a 10 and 7 allele. The small portion of flanking region duplicated contains the primer binding site sequence for PowerPlex 16 (TCTGTCTTTTTGGGCTGCC) [202], which explains the triple genotype observed with this kit (8, 10, OL/28.2). This is a clear example of the benefit of massively parallel sequencing in resolving discordances.

A



B

Allele 8

[TATC]8 aatcaatcatctatctatctttctgtctgtctttttggg

Allele 28.2

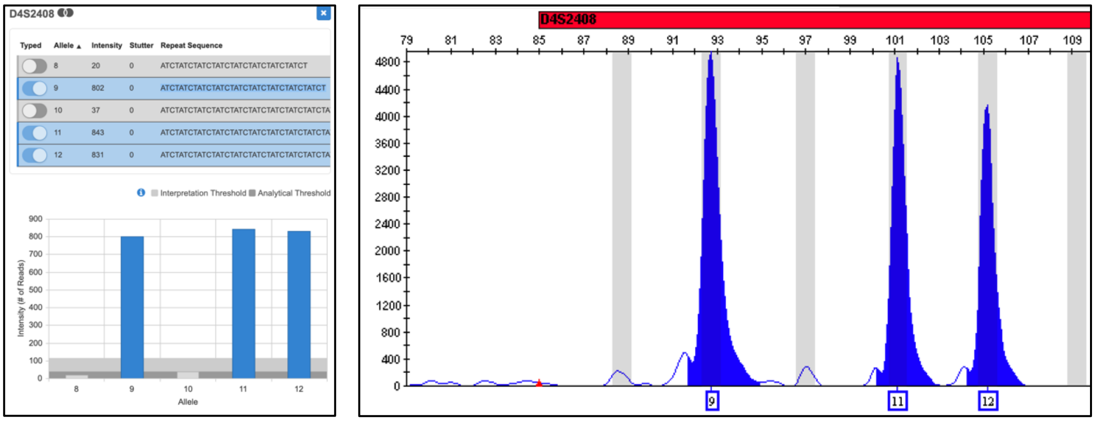
[TATC]10 AATCAATCATCTATCTATCTTTCTGTCTGTCTTTTTGGGCTGCCTA [TATC]7 aatcaatcatctatctatctttctgtctgtctttttggg

Figure 3.7: Discordance observed between two CE kits resolved using MPS

The profiles obtained using GlobalFiler and the ForenSeq DNA Signature Prep kit are concordant, but different to the result obtained from PowerPlex 16, where an additional 10 allele is observed (A). The sequence for the 28.2 allele (B) shows a partial fusion of the repeat region and flanking region of a 10 and 7 allele, leading to the 28.2 genotype. The flanking regions are shown in lower case, and the primer binding sequence for PowerPlex 16 is underlined in the 28.2 allele, which led to the triple genotype with that kit.

3.1.7. Triple genotype at D4S2408

D4S2408 is a simple tetranucleotide repeat with a [ATCT] repeat unit that is not traditionally targeted by CE STR kits, with allele ranging in size from 7-13 seen in this work. One instance of a triple genotype was observed in a North East African sample, and given the maker isn't part of any standard CE kit, the genotype was verified using published primers [180] and CE. The genotype was confirmed to be genuine and would appear to be the first instance of this occurrence. Given the balanced nature of this triplicate genotype, combined with the impossibility of having 3 complete copies of chromosome four, the cause of this is likely to be chromosomal rearrangement; similar to that observed at the TPOX locus where the third allele is in fact located on the X chromosome [203, 204]. Both the ForenSeq DNA Signature Prep kit and custom CE profiles are shown in Figure 3.8.



D4S2408

Typed	Allele ▲	Intensity	Stutter	Repeat Sequence
○	8	20	0	ATCTATCTATCTATCTATCTATCTATCTATCT
●	9	802	0	ATCTATCTATCTATCTATCTATCTATCTATCTATCT
○	10	37	0	ATCTATCTATCTATCTATCTATCTATCTATCTATCTA
●	11	843	0	ATCTATCTATCTATCTATCTATCTATCTATCTATCTA
●	12	831	0	ATCTATCTATCTATCTATCTATCTATCTATCTATCTA

Figure 3.8: D4S2408 Triple genotype

Instance of a triple genotype observed at D4S2408 in a North East African sample in UAS (left) and confirmed using published primers and CE (right).

3.1.8. Discussion on concordance

As mentioned at the start of this chapter, a concordance rate exceeding 99% across all STR markers investigated is comparable to the value observed when comparing any two CE kits. The fact that different commercial kits target different primer binding sites to amplify the same markers can lead to poor amplification of an allele or even drop out when a mutation exists in the region where the primer for one kit binds, leading to a discordant genotype. The discordances observed at D5S818 (section 3.1.1) and D21S11 (section 3.1.5) are examples of this and simply highlight the fact that the ForenSeq DNA Signature Prep kit relies on primer-based target amplification, where these primers can differ to those used in CE-based kits. Although certain rare mutations can always occur, such as the first discussed at D5S818, Verogen may want to consider increasing the ratio of the redundant primer for the rs25768 SNP at D5S818 and adding one for the rs531980552 SNP at D10S1248 in future iterations of this kit to avoid any issues with concordance. The deletion in the primer binding site of D21S11 in the four African samples is more complicated and could require the design of a new primer to resolve.

The discordance at D7S820 (section 3.1.2) introduces the need to account for flanking region sequences for nomenclature purposes, which will form a major part of the discussion later in this thesis. By looking at the repeat region alone, it is normal that the UAS would identify the $[GATA]_7$ sequence as an allele 7, but back compatibility to CE is crucial, and therefore the full length of the amplicon must be considered in order to avoid discordant results like the one observed in this case. Other software tools such as STRait Razor [127, 205] and FDSTools [206] have taken this onboard already and look at both sequence and amplicon size to name autosomal STR alleles.

The discordance at Penta D (section 3.1.4) highlights the added complexity associated with comparing methods that rely on different molecular biology principles, while the use of the ForenSeq DNA Signature Prep kit to resolve a discordance between two CE kits at D13S317 (section 3.1.6) features a major advantage of using sequencing rather than size-based separation for the interpretation of results. Overall, these results indicate that the ForenSeq DNA Signature Prep kit is highly concordant with the PowerPlex and GlobalFiler kits, and results obtained using the different kits can be compared in the context of DNA typing. Large scale concordance studies such as this one are important to find out how likely it is to observe a discordant genotype when using multiple kits for analysis.

Since the beginning of this work, numerous other groups have also published concordance studies between MPS assays and CE and have also found very few discordances [62, 63, 207-210]. Just et al. obtained an autosomal STR concordance rate of 99.96% when comparing the ForenSeq DNA Signature Prep kit and PowerPlex Fusion CE kit [74] across 103 individuals. Novroski et al. compared the ForenSeq DNA Signature Prep kit with GlobalFiler across 170 samples and observed just two discordances, resulting once more in a concordance rate exceeding 99% [63]. As in this work, one discordance was observed due to a point deletion in the flanking region of the D7S829 marker, causing a 10.3 allele identified with CE to be typed as an 11 with the UAS. The authors came to the same conclusion as that discussed above regarding the need to incorporate flanking regions and overall amplicon size when naming alleles. Gettings et al. make a similar point when discussing what they refer to as “Flanking Region InDel type null alleles” and suggest moving the bio-informatic recognition sites in order to improve concordance with length-based data [62].

3.2. Characterisation of sequence variants

The results discussed above highlight the fact that STRs can be successfully typed using massively parallel sequencing, and the genotypes produced are highly concordant with those generated using CE. This back compatibility is of course vital for the implementation of MPS in forensic DNA typing, due to the millions of CE-generated STR profiles in national DNA and population databases. So, length-based genotypes obtained using MPS can be used to search a profile against a DNA database in the same way as any adopted CE kit results. The rationale for implementing a new technology is not simply for it to match what can currently be done however, and therefore the added benefits of sequencing STRs need to be investigated. MPS derived genotypes provide additional information compared to those obtained through size separation, by capturing the full underlying nucleotide sequence of the repeat units of STRs as well as neighbouring flanking regions. This added level of granularity for analysing STRs in turn leads to an increase in the power of discrimination of the test. In this section, the increase in alleles observed across commonly used STRs when applying MPS will be presented and discussed. The addition of sequence level information brings into question how we can easily refer to the "new" alleles observed, and so the strategy for allele characterisation will also be presented. Interesting sequences were investigated further, and the impact of omitting flanking region sequences for allele characterisation will be highlighted prior to further discussion in the next chapter.

3.2.1. Repeat region sequence variation

The extent of sequence variation in the 26 autosomal STRs studied was first investigated in the context of variation within the repeat region of these markers. This region has traditionally been the focus of STR analysis, and initial versions of the data analysis pipelines (namely UAS and STRait Razor) used in this work did not yet allow for visualisation of flanking regions of STRs.

3.2.1.1. Characterisation of repeat region variant alleles

In order to make use of the additional data provided by MPS results, alleles were characterised according to the method described in the previous chapter. In brief, sequences were extracted from UAS sample reports and used to compile an internal database. As more samples were sequenced and analysed, a VLOOKUP function was used in Excel to identify allelic sequences, which would return the bracket annotation and internal short designator based on whether the FASTA sequence inputted had been seen before. If this lookup returned no result, the sequence was added to the database. Table 3.2 shows an example of the compiled sequences for one marker, D17S1301.

The short designators used to assign a name to the sequence-based alleles were solely used for internal purposes, in order to avoid issues during downstream data analysis and frequency generation. The bracket annotation is consistent with how autosomal STRs are usually referred to in the literature, and described in databases such as STRBase [17]. This nomenclature system is purely used to describe the repeat region of STRs however and did not take into account flanking regions at this stage. Although the latest ISFG considerations suggest reporting all alleles on the forward strand, UAS still reports the following markers on the reverse strand, as historically reported [65]: D1S1656, D2S1338, D5S818, D6S1043, D7S820, D19S433, CSF1PO, FGA, Penta E and vWA. In this work, the bracket allelic annotations were kept in both directions for these markers to enable comparisons with other publications.

Table 3.2: Example of the table used to compile and then identify repeat-region sequence variants

This table shows the alleles observed for D17S1301. Once a reasonable number of sequences had been seen, the FASTA sequences for newly sequenced samples were searched against this database, and if the same FASTA sequence was present in the table, the VLOOKUP function would return the corresponding short identifier (Allele (SB)) and bracket annotation. Length-based = LB; Sequence-based = SB.

Allele (LB)	Allele (SB)	FASTA	Bracket
7	7	AGATAGATAGATAGATAGATAGATAGAT	[AGAT]7
8	8	AGATAGATAGATAGATAGATAGATAGATAGAT	[AGAT]8
9	9	AGATAGATAGATAGATAGATAGATAGATAGATAGAT	[AGAT]9
10	10	AGATAGATAGATAGATAGATAGATAGATAGATAGATAGAT	[AGAT]10
11	1101	AGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGAT	[AGAT]11
11	1102	AGATAGATAGATAGATAGATAGATAGATAGATAGATAGATCGAT	[AGAT]10 CGAT
11.3	11.3	AGATAGATAGATAGATAGATAGATAGATAGATGATAGATAGATAGAT	[AGAT]8 GAT [AGAT]3
12	1201	AGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGAT	[AGAT]12
12	1202	AGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATCGAT	[AGAT]11 CGAT
13	1301	AGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGAT	[AGAT]13
13	1302	AGATAGATACATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGAT	[AGAT]2 ACAT [AGAT]10
14	14	AGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGAT	[AGAT]14
15	15	AGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGAT	[AGAT]15

As mentioned in the previous chapter, UAS reports a small portion of sequence directly adjacent to the repeat region for several markers. Two STR loci showed considerable variation in this “short flank” portion of sequence. Lack of comprehensive allele sequence information at the time of definition of the D5S818 and D13S317 loci (~25 years ago) means that the traditionally defined repeat region of these markers does not take into account all variation which affects their repeat structure. In both cases, there are additional repeat units directly next to the repeat regions. In the case of D5S818, rs73801920 is a A/C SNP, where the C allele causes the four base pairs directly neighbouring the initially defined repeat region to act as an additional [ATCT] repeat. Figure 3.9 shows the structure of an 11 allele at D5S818, which when present with the

C allele at rs73801920 (and therefore a structure of "[ATCT]$_{12}$") has a global frequency of 0.054 in this dataset. Figure 3.9 also shows the structure of an 11 allele at D13S317, which has two SNPs which affect the repeat region. Having a T allele at rs9546005 is common (global frequency of 0.42 according to dbSNP and the 1000 genomes project [45, 185]), with the "[TATC]$_{12}$" version of an allele 11 having an average frequency of 0.15 across the five populations studied. Having a T allele at both rs9546005 and rs202043589 is more uncommon, with the "[TATC]$_{13}$" version of an allele 11 having only been observed at a frequency of 0.018 in the British Chinese population in this work.

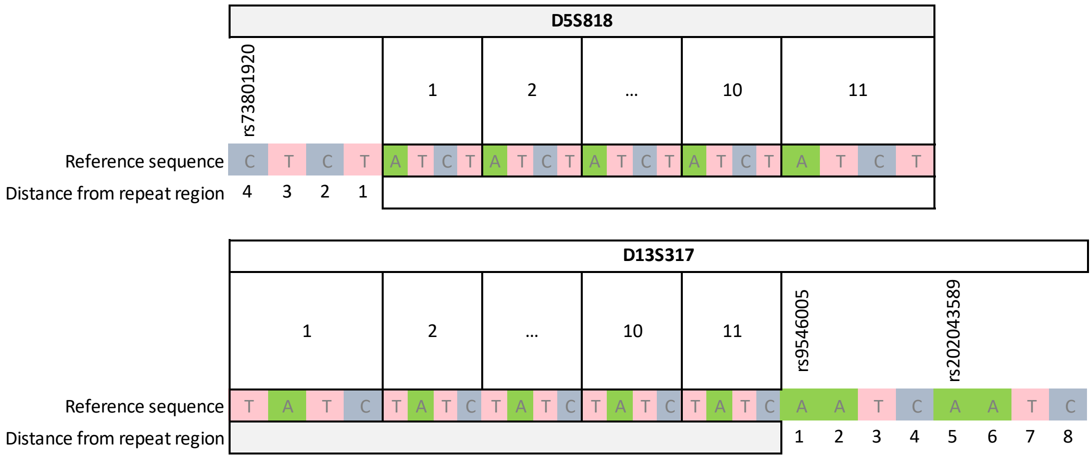

D5S818
rs73801920
1
2
...
10
11
Reference sequence
C T C T A T C T A T C T A T C T A T C T A T C T
Distance from repeat region
4 3 2 1
D13S317
1
2
...
10
11
rs9546005
rs202043589
Reference sequence
T A T C T A T C T A T C T A T C T A T C A A T C A A T C
Distance from repeat region
1 2 3 4 5 6 7 8

Figure 3.9: Sequence structure of D5S818 an D13S317

Sequences were taken and adapted from the updated 'Forensic STR Sequence Structure Guide' associated with the publication by Phillips et al. [184]

Although other publications have referred to the variation in these regions as affecting the flanking regions of the STRs [98, 211], it has been included as part of the repeat region structure in this work, for consistency with the UAS output and for scientific accuracy. Theoretically, the knowledge we now have about the structure of these markers could be used to change the way in which we report them, but given the extensive number of databases that already have D5S818 and D13S317 results in the current format, careful consideration is required. One option would be to adopt a nomenclature design that includes these variants as part of the flanking region of these loci but with a commentary on location and effect on repeat region structure. Another option would be to keep the length-based designator the same as for CE (i.e. an 11 allele for the examples discussed above), but with the entire repeated region in bracker format (e.g. 11 [TATC]$_{12}$).

3.2.1.2. Additional variation observed compared to length-based alleles

The number of individual alleles characterised by length was compared to those obtained by looking at variation within the repeat region of the 26 autosomal STRs studied, as shown in Figure 3.10 and then again in Figure 3.11. Samples were split according to population group for this second figure, as variation in marker discrimination is expected between populations. D12S391 is the most highly polymorphic autosomal STR within the group of markers studied in this work, with 88 distinguishable alleles in the repeat region sequence-based results compared to 25 alleles based on length. TPOX and TH01 are the two least polymorphic markers, with 9 alleles distinguishable based on size for both and just one additional allele characterised using repeat region variation for TH01. Only two of the twenty-six autosomal loci showed no gain in the number of alleles seen using sequence information compared to length-based analysis: TPOX and D10S1248. These results differ slightly from the results published by Gettings et al. [62], where increased variation was seen at D10S1248, although no sequence variation was observed in their study at D7S820 and D13S317. Novroski et. al. [63] reported sequence variation at all autosomal loci except TPOX, whereas Delest et al. [212] did not observe any sequence variation at TPOX, D17S1301, TH01, CSF1PO, D10S1248, or Penta E in their database of 169 French individuals. This is likely due to differences in the number of samples and populations investigated between the different studies.

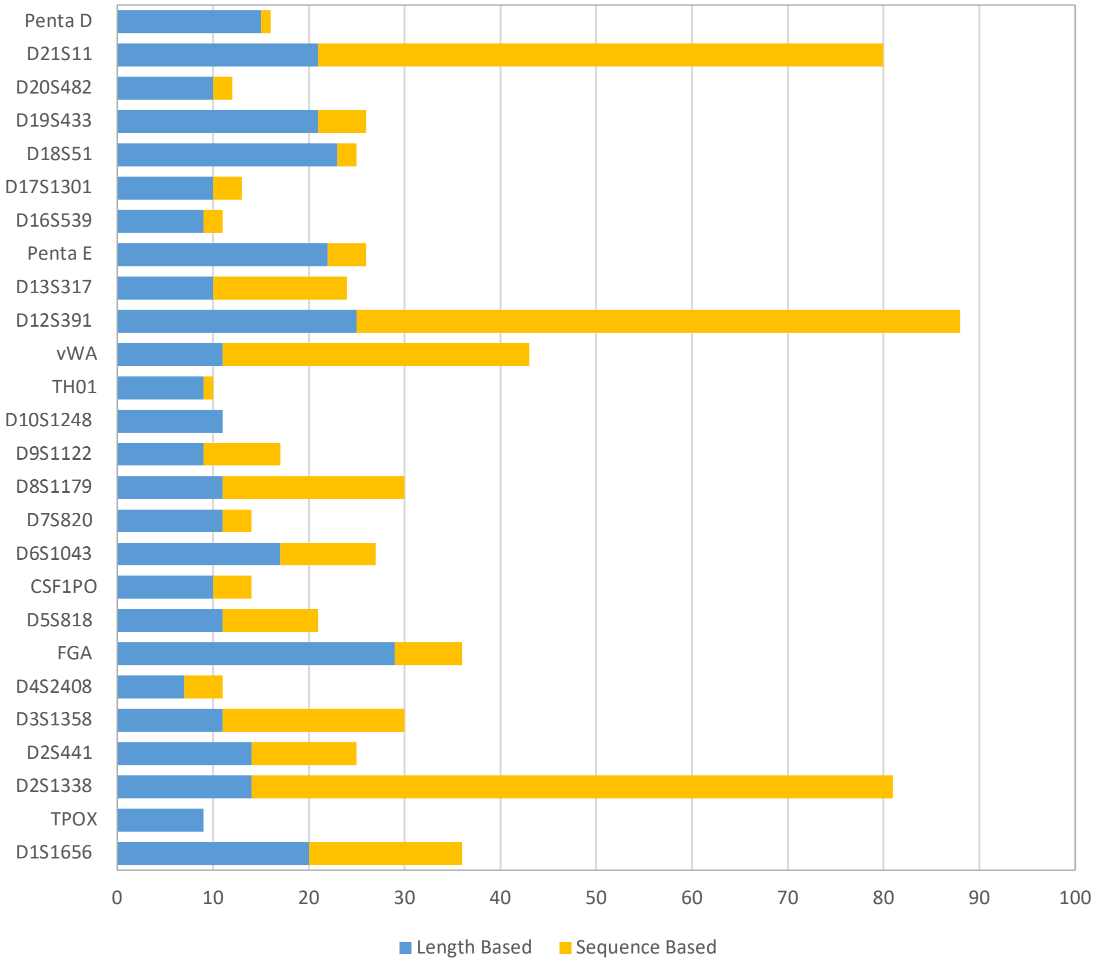


Figure 3.10: Increase in the number of STR alleles observed with repeat region variation, across all five population groups studies

This graph shows the increase in the number of alleles seen across 26 autosomal STR markers targeted by the ForenSeq DNA Signature Prep kit when taking sequence-based variation into account compared to length-based data.

This substantial increase in the number of alleles observed will directly impact forensic investigations, as it will increase the power of discrimination of STR testing. The larger the number of alleles that can be distinguished from commonly targeted loci, the lower the individual allelic frequencies and therefore the higher the power of discrimination. Allelic frequencies and their impact will be discussed in more detail in the next chapter.

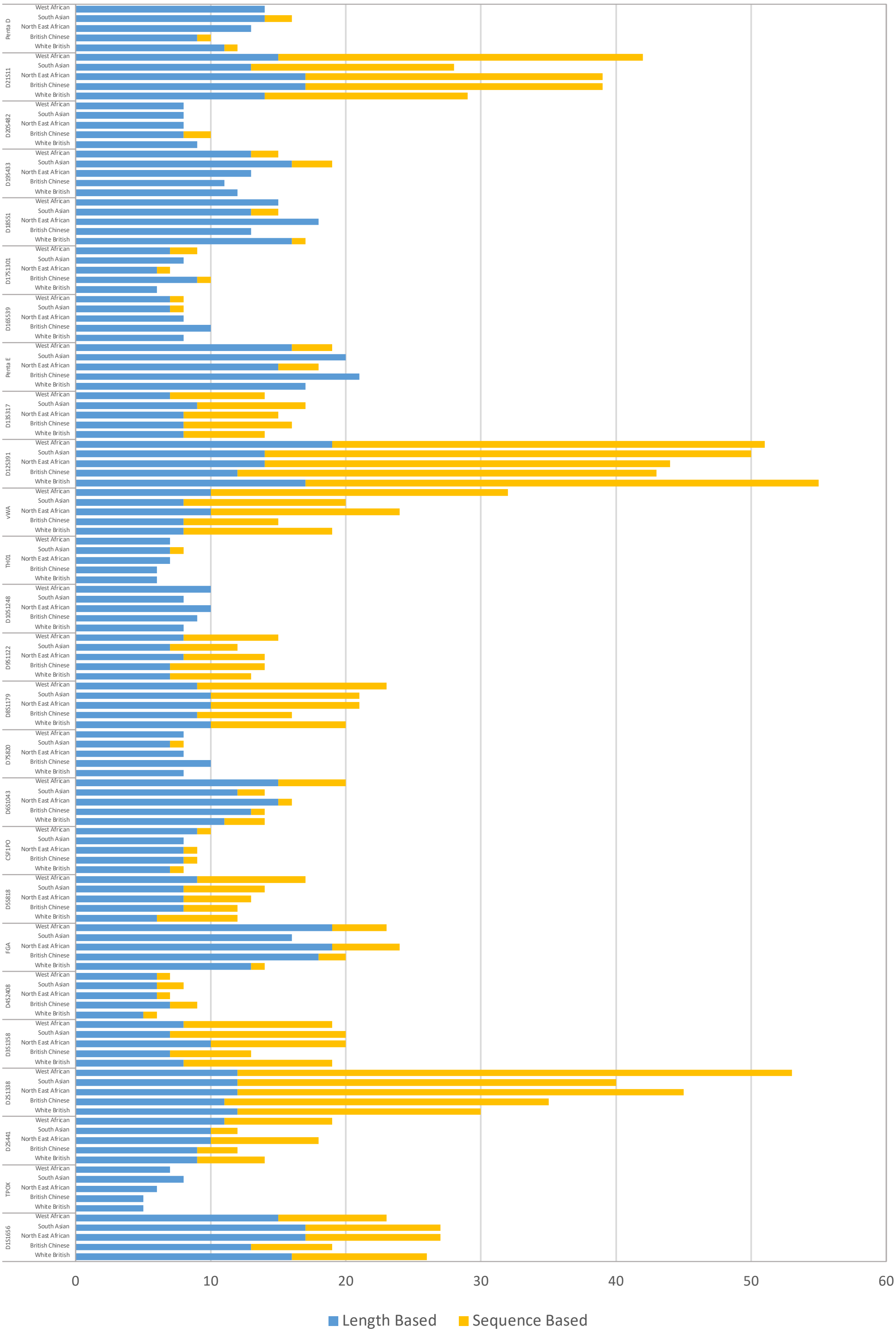


Figure 3.11: Increase in the number of STR alleles observed, split by population

This graph shows the same data on the increase in the number of alleles seen using sequence-based allelic data as Figure 3.10, but split according to the different population groups studied.

3.2.1.3. Interesting sequences in the "short flank"

During the characterisation of sequence-based alleles, several interesting variants were identified from the UAS reports, found in a portion of sequence not traditionally associated with the repeat region of STRs. For the following markers, a short section of sequence directly adjacent to the repeat region is reported by UAS in the sample summary reports: D13S317, D18S51, D19S433, D1S1656, D5S818, D7S820, vWA. This region of sequence was already discussed specifically for D5S818 and D13S317 in section 3.2.1.1, and for all of these markers count as part of the flanking region in the "Forensic STR Sequence Structure Guide" [184]. They have been referred to as "short flank" in this work. One novel variant was observed at D18S51, a locus which exhibits limited repeat region sequence variation. This variant was an 18.1 length-based allele, with the following sequence: $[AGAA]_{14}$ AAAG AGAG AG GAA [AGAA] AAAG AGAG AG. Although a size-based 18.1 allele has been seen before, no sequence information is available in the literature [17]. D18S51 is a simple tetranucleotide repeat marker where the common repeat motif is $[AGAA]_n$. The "short flank" of this locus consists of a 10bp sequence found after the repeat region (AAAG AGAG AG), and in the case of the allele, this sequence can be seen to occur twice, with one occurrence of the [AGAA] repeat motif in the middle. The divergence at this allele would suggest that either $[AGAA]_{14}$ AAAG AGAG AG is the original allele and a GAA [AGAA] AAAG AGAG AG has then been appended on, or the AAAG AGAG AG GAA sequence has been inserted inside a normal 15 allele. Because UAS uses a short amount of flanking region for allele designation at this marker, a discordance was not observed with the CE data for this sample. The utility of this short sequence of flanking region for allele designation is also demonstrated at D7S820, where a 9.2 allele was observed which is composed of 10 tetra-nucleotide repeats within its repeat region: $[GATA]_{10}$ GACA GATT GA-- GTTT. In the "short flank" (in grey), two bases are missing due to a deletion. Looking at the repeat region sequence alone in this case would not match to the size-based allele and would have therefore been discordant with the CE result.

3.2.2. Flanking region variation

In 2016, the DNA Commission of the International Society for Forensic Genetics (ISFG) published a series of considerations on minimal nomenclature requirements for the massively parallel sequencing of forensic STRs [135]. The article includes a discussion on the importance of including flanking region sequences when characterising STR alleles, both because of the potential of added variation provided in these regions, and because the mapping of insertions or deletion can inform the assignment of size-based alleles. This second point was already considered during this project due to the discordance observed between MPS and CE results for a sample with a one base deletion in the flanking region. One major consideration listed by the DNA Commission of the ISFG was as follows: "To account for relevant genetic variation outside common repeat regions, STR sequences stored as sequence strings should include flanking sequences as well as the genome coordinates of the sequence read start and end." The provision on genome coordinates is important in the context of ensuring compatibility between different primer sets. If a flanking region variant was observed 100 bp upstream from the repeat region for example, it may be considered a "novel" allele, but the same variant would not be seen when using a kit that only includes 50 bp upstream of the repeat region. To investigate flanking regions, data for all samples initially used for concordance and the repeat region variation characterisation were re-analysed using two data analysis tools.

3.2.2.1. Characterisation of flanking region variant alleles

The output from the modified STRait Razor pipeline introduced in the methods chapter included full FASTA sequences for all alleles (including flanking regions), as well as the size-based allelic designation and bracket nomenclature for the repeat region sequence. A comparison to the alleles characterised based on repeat-region variation revealed full concordance, apart for the allele at D7S820 discussed in section 3.1.2, where STRait Razor provided the same result as that obtained for CE given it takes amplicon length into account. Once the option was available in the software, UAS flanking region reports were downloaded and used to check against the sequence output from STRait Razor. At this point, it was confirmed that there were no discrepancies observed between the sequence output from the two methods for data analysis. The full FASTA sequences, including flanking regions, for all the alleles already validated and checked for concordance and repeat region variation were added to the internal database described earlier. For the following markers, the bracket annotation was added in both reporting directions to ensure back compatibility with both UAS flanking region reports and STRait Razor (and STRSeq, discussed below): D1S1656, D2S1338, D5S818, D6S1043, D7S820, D19S433, CSF1PO, FGA, Penta E and vWA. Table 3.3 shows an example of what the list of alleles looks like for the marker D1S1656, which can then be used both as a measure of the extent of sequence variation observed at this marker, and to identify any future sequence from a sample genotyped using the ForenSeq DNA Signature Prep kit and analysed using the UAS. As part of the publication process, all genotypes were also submitted to STRidER, a publicly available, centrally curated online allele frequency database and quality control platform for autosomal STRs [175]. The full list of sequences characterised in the course of this PhD is given at the end of this chapter, showing the bracket annotation for the repeat region of all alleles, as well as flanking region sequences, short flank and STRSeq record description where appropriate (Table 3.5).

Table 3.3: Example of the table used to identify flanking-region variants

This table shows a subsection of the alleles observed for D1S1656. Using this table, FASTA sequences for newly sequenced samples (either repeat region - RR or full amplicon - FR) can be searched against the internal database, and if the same FASTA sequence is present in the table, the VLOOKUP function will return the short identifier (Allele (SB)) and the bracket annotation. The "Short Flank", [TG]$_5$ sequence, reported by UAS for this marker is seen in grey in the bracket annotation. Length-based = LB; Sequence-based = SB.

Allele (LB)	Allele (SB)	UAS FASTA (RR)	UAS FASTA (FR)	Bracket (UAS)	STRSeq Record
17	1701a	TAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGGTGTGTGTGTG	TAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGGTGTGTGTGTGTTTAATTGTATGTATATATATTTGGTTCCCTAGTGATTCTATTTCTCTGAA	[TAGA]16 TAGG [TG]5	CCTA [TCTA]16
17	1701b	TAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGGTGTGTGTGTG	TAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGGTGTGTGTGTGTTTAATTGTATGTATATATATTTGGTTCCCTAGTGATTCCATTTCTCTGAA	[TAGA]16 TAGG [TG]5	CCTA [TCTA]16 rs541123499
17	1702	TAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATGTGTGTGTG	TAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATGTGTGTGTGTGTTTAATTGTATGTATATATATTTGGTTCCCTAGTGATTCTATTTCTCTGAA	[TAGA]17 [TG]5	[TCTA]17
17.3	17.3	TAGATAGATAGATAGATGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGGTGTGTGTGTG	TAGATAGATAGATAGATGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGGTGTGTGTGTGTTTAATTGTATGTATATATATTTGGTTCCCTAGTGATTCTATTTCTCTGAA	[TAGA]4 TGA [TAGA]12 TAGG [TG]5	CCTA [TCTA]12 TCA [TCTA]4 rs4847015
18	1801	TAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATGTGTGTGTG	TAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATGTGTGTGTGTGTTTAATTGTATGTATATATATTTGGTTCCCTAGTGATTCTATTTCTCTGAA	[TAGA]18 [TG]5	[TCTA]18
18	1802	TAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGGTGTGTGTGTG	TAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGGTGTGTGTGTGTTTAATTGTATGTATATATATTTGGTTCCCTAGTGATTCTATTTCTCTGAA	[TAGA]17 TAGG [TG]5	CCTA [TCTA]17

3.2.2.2. Additional variation observed

No variation was observed in the flanking regions of STRs that did not show variation in the repeat region, namely TPOX and D10S1248. This is consistent with other studies such as that by Phillips et al. [213] who recorded no sequence variation at these loci, and the later study by Gettings et al. [98] who observed one instance of repeat variation for both TPOX and D10S1248 but no flanking region variation. The loci showing the most allelic gains by sequence when including both repeat and flanking region variation are vWA, D21S11, D2S1338, and D12S391, which are all compound or complex STRs [65]. This is once again consistent with other studies looking at sequence variation both within and outside of the repeat regions of these STRs [62, 63, 95, 96, 213, 214]. Figure 3.12 shows that for virtually all markers, variants not associated with the repeat region account for a small proportion of the increase in alleles observed due to sequence variation. At D7S820, D16S539, D20S482 and Penta D, this increase is more pronounced. Table 3.4 shows the length of flanking region sequence provided by UAS for each locus, and number of alleles identified based on variation in the flank. There is a strong correlation between the two which was found to be significant (Pearson's R= 0.694, p=0,0000834), suggesting, perhaps unsurprisingly, that there is a higher chance of observing flanking region polymorphisms in longer flanking sequence stretches. The usefulness of flanking region variation is better demonstrated by considering frequencies, which will be discussed in the following chapter.

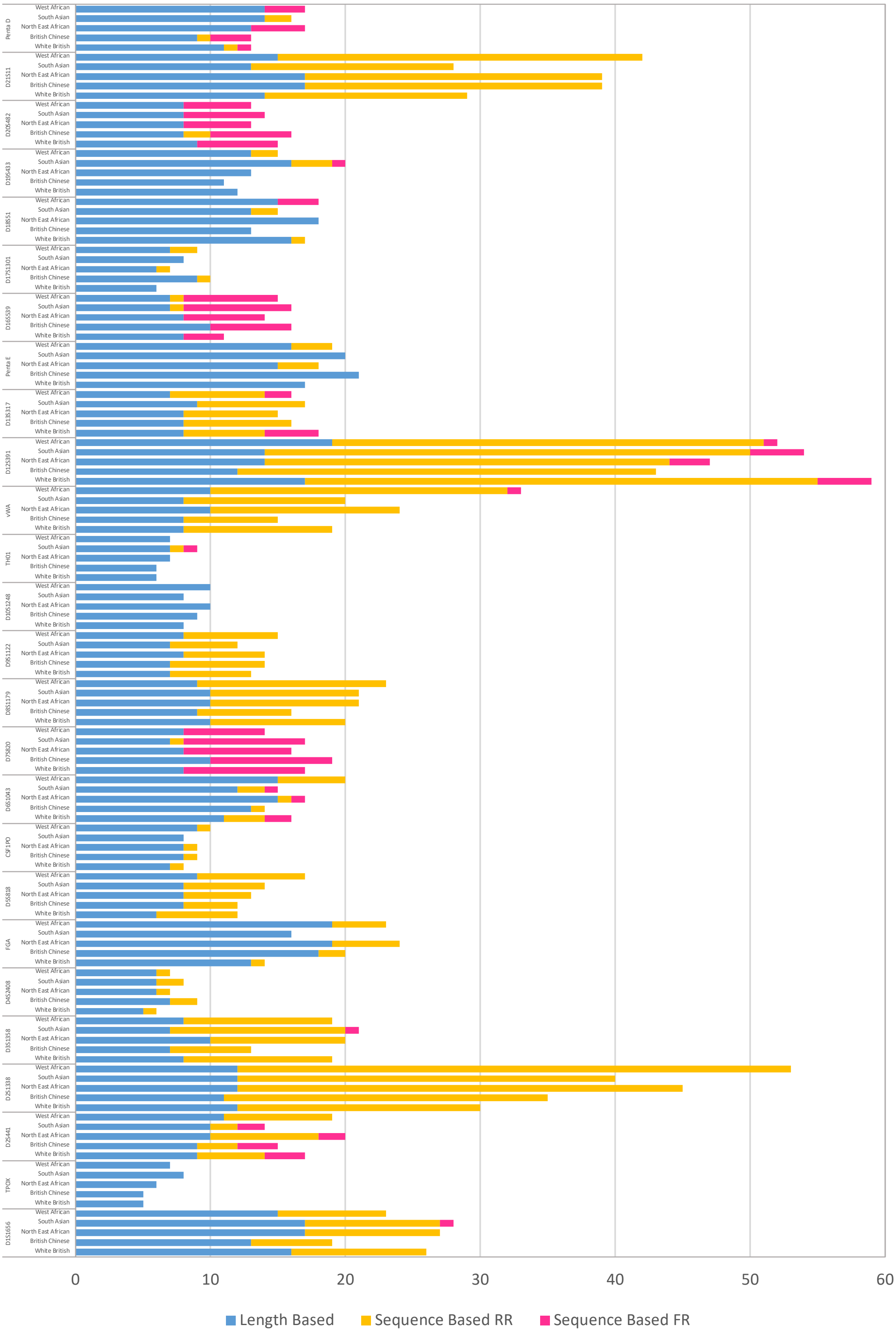


Figure 3.12: Increase in the number of STR alleles observed with flanking region variation, split by population

As in Figure 3.11, this graph shows the increase in the number of alleles observed when considering sequence variation (RR: Repeat Region; FR: Flanking Region).

Table 3.4: Amplicon length vs variation in flanking regions

Length of flanking region sequence provided by the UAS output (flanking region report) per locus in base pairs, and number of alleles characterised by variation in flanking region (FR alleles).

STR Locus	FR length (bp)	Number of FR alleles
D1S1656	51	1
TPOX	5	0
D2S1338	17	0
D2S441	52	3
D3S1358	66	1
D4S2408	14	0
FGA	40	0
D5S818	16	0
CSF1PO	20	0
D6S1043	85	3
D7S820	39	12
D8S1179	5	0
D9S1122	24	0
D10S1248	41	0
TH01	44	1
vWA	24	1
D12S391	133	9
D13S317	78	6
Penta E	19	0
D16S539	74	12
D17S1301	25	0
D18S51	53	3
D19S433	66	1
D20S482	42	6
D21S11	46	0
Penta D	153	12

3.2.3. STRSeq

The full sequences for 1100 samples (including a small number of Middle Eastern samples) were submitted as part of the creation of the STR Sequencing (STRSeq) BioProject [139], alongside data from the National Institute of Standards and Technology (NIST) [98, 180, 215], the University of North Texas (UNT) [63, 64] and the University of Santiago de Compostela (USC) [213]. This curated catalogue of sequence diversity was set up to facilitate the description of sequence-based alleles at forensically relevant STR loci, in a format consistent with the DNA Commission for the ISFG's considerations [135]. The STRSeq catalogue will evolve alongside tools for bio-informatic data analysis, and requirements for sequence-based allelic nomenclature, facilitating inter-laboratory communications. The future goal for this database will be for it to be used to check all submitted MPS-obtained STR results, by integrating it with STRidER which could then query the STRSeq database, enabling researchers to identify whether the sequences they have observed are novel or not, and what the correct nomenclature is. If STRidER were to find no matching sequence in STRSeq, a process to evaluate the sequence would be initiated (checking of sequence range, flanking region polymorphisms etc). To this end, it was important that a range of large-scale population databases be used to compile the initial list of "known sequences". Figure 3.13 shows the overlap of sequences submitted by all four contributing laboratories for the locus D12S391. Of the 97 sequences observed for this locus as part of this work, 58 were seen by all laboratories and 17 were seen solely in the KCL database. This is likely due to the fact that different populations were sampled, with the work in this project focussing on UK-relevant population groups.

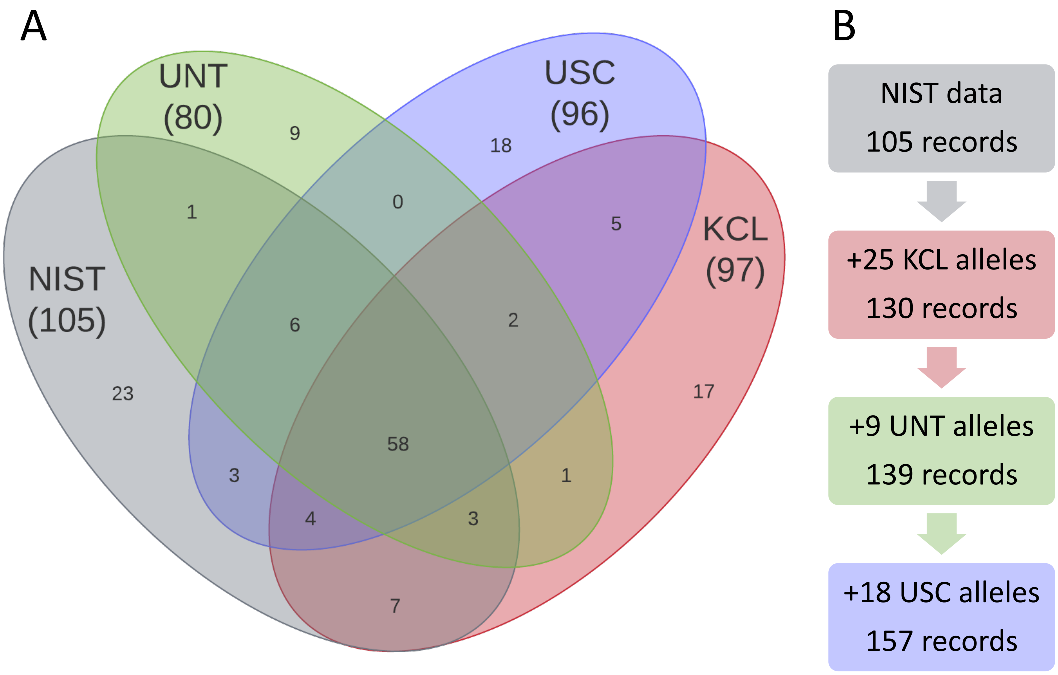


Figure 3.13: D12S391 sequence-based alleles submitted to STRSeq

A: Venn diagram showing the overlap of sequence-based alleles initially submitted to the STRSeq BioProject for the D12S391 locus by the four participating laboratories, with the total number of unique sequences observed for each given in brackets. B: Sequential submission of unique sequences from each laboratory led to a final 157 records for D12S391, with the data from this project (KCL) providing 25 unique sequences which had not been seen in the NIST data set. Figure recreated and adapted from Gettings et al. [139].

3.3. Discussion on implementation

A crucial factor for the implementation of MPS in forensic DNA testing was its compatibility with existing, CE-generated databases. This work has shown that genotypes obtained using the ForenSeq DNA Signature Prep kit and MiSeq FGx can be converted to length-based results which are concordant with CE genotypes and can be compared to a CE profile or searched against a database. The small number of discordant results observed have been fully investigated, and suggestions as to how discrepancies can be avoided discussed earlier on in this chapter. Characterising sequence-based alleles is important in the context of comparison with CE results, but also to understand the added value of using sequencing over size-based allelic separation. At the start of this work, it was impossible to know what the breadth of variation would be for some of the autosomal STRs that have been in common use for the past 20 years. Whilst there is a clear correlation between the complexity of the repeat structure of these loci and the amount of sequence variation observed, it took looking at over 1000 samples to gain a true understanding of how the markers behave at the sequence level. D12S391 showed an increase of almost 300% in the number of alleles observed by sequence compared to the number observed by length, and it's likely that considerably more variation will be seen when looking at more populations. These results have also been published (Appendix I and II), and form part of a global online database that is available to the wider forensic community. As of April 2021, the article introducing STRSeq had been cited 41 times, with a NIST grant summary report breaking down usage of the STRSEq BioProject showing that since 2019, researchers have been mostly using the database to assess if a sequence is known (or to check its correct formatting), and in the development of data analysis frameworks for MPS data [216]. The data from King's College London contributed 214 unique sequences to this dataset, demonstrating the value of such a large-scale population database.

As well as gaining an understanding of the breadth of variation seen at traditional loci when accounting for sequence information, characterising the alleles provided the foundation for the population databases which will be presented in the next chapter. Alleles were given a short, internal designator which allowed for data to be input into certain bioinformatic software tools, which links back to the full sequence string and a bracket annotation. This designator involves different characters to denote repeat and flanking region variants, to enable their comparison in later work. Nomenclature remains an important barrier to the implementation of MPS, but a growing understanding of how STRs vary will help inform decisions by bodies such as the STRAND working group [136]. The bracket annotation method referred to in this project for example works well when referring to the repeat region of autosomal STRs, and provides an ease of translation to CE, length-based designation, but crucially misses out information about the flanking region sequences. The incorporation of flanking region sequences complicates the matter of STR nomenclature on several levels. One major consideration relates to profile comparison or database searching. Because different commercial MPS-based kits target different length amplicons through differing primers, the same sample could be sequenced with two different kits and provided discordant results if one result reports a differently named allele due to flanking region variances. In order to avoid this, it is vital that a nomenclature system be adopted which takes this into consideration. Currently, multiple research groups and consortiums are working towards a standardised, easy to use method, but the preferred tactic for flanking region sequence inclusion has yet to be agreed. Using the full sequence string for referring to MPS-generated data, whilst the most technically straight forward, has a number of important limitations. For one, many forensic DNA databases are not currently equipped to deal with such long strings of characters, and so these cannot be adequately stored or searched for [136]. For database searching and mixture analysis software tools, the use of a short designator would be preferred due to the reduced number of characters [137, 138]. These short designators would refer to an online (or offline, with periodic updates) database which refers back to the full string sequence. The problem with this solution is that two laboratories could theoretically discover a new allele and assign them difference short designator if the database isn't updated continuously. Using a bracketed repeat format for naming alleles enables a certain level of understanding of the allele structure, and enables easy comparison between results while maintaining a relatively low number of characters. The main drawbacks of this solution are potential lack of compatibility with the CE allele, and complications when trying to include flanking region variants. Hoogenboom et al. [128] suggest a new algorithm called "STRNaming", which would return a bracketed repeat type name for any sequenced allele which include the length-based designation, bracket repeat and flanking region variants by means of variants calls upstream or downstream of the repeat region. This method was developed following extensive collaboration with laboratories who generated large scale population databases, to ensure results were accurate, comparable, and met the needs of the forensic community. The results from this PhD have contributed in a very real way towards the solutions for how laboratories will interpret and apply results, leading to global adoption of the technology. Studies such as this one, and the others that have contributed to the STRSeq BioProject and STRNaming, are vital to help move towards a solution that starts with the most comprehensive amount of information possible. Although the utility of analysing flanking regions remains to be seen in terms of power of discrimination, and will be discussed in the next chapter, they absolutely must be considered when characterising alleles, to avoid the kind of discrepancies discussed at the start of this chapter.

Table 3.5: Characterised sequences for 1018 samples sequenced using the ForenSeq DNA Signature Prep Kit

Length-based (LB) and sequence-based (SB) allelic designations are given for each sequence characterised, as well as the repeat region bracket annotation according to UAS, and the STRSeq record description for the markers where this is different to UAS reporting. The "short flank" included as part of the UAS repeat region output is shown in grey for the relevant markers, and left and right flank sequences are provided, with flanking region variants highlighted in pink (RS number given when known).

D1S1656							
Allele (LB)	Allele (SB)	Bracket (UAS)	STRSeq Record Description		Left Flank	Right Flank	
7	701	[TAGA]7 [TG]5	[TCTA]7		NA	TGTGTGTGTGTTTAATTGTATGTATATATATTTGGTTCCCTAGTGATTCTATTTCTCTGAA	
8	801	[TAGA]8 [TG]5	[TCTA]8		NA	TGTGTGTGTGTTTAATTGTATGTATATATATTTGGTTCCCTAGTGATTCTATTTCTCTGAA	
9	901	[TAGA]9 [TG]5	[TCTA]9		NA	TGTGTGTGTGTTTAATTGTATGTATATATATTTGGTTCCCTAGTGATTCTATTTCTCTGAA	
10	1001	[TAGA]10 [TG]5	[TCTA]10		NA	TGTGTGTGTGTTTAATTGTATGTATATATATTTGGTTCCCTAGTGATTCTATTTCTCTGAA	
10	1002	[TAGA]9 TAGG [TG]5	CCTA [TCTA]9		NA	TGTGTGTGTGTTTAATTGTATGTATATATATTTGGTTCCCTAGTGATTCTATTTCTCTGAA	
11	1101	[TAGA]11 [TG]5	[TCTA]11		NA	TGTGTGTGTGTTTAATTGTATGTATATATATTTGGTTCCCTAGTGATTCTATTTCTCTGAA	
11	1102	[TAGA]10 TAGG [TG]5	CCTA [TCTA]10		NA	TGTGTGTGTGTTTAATTGTATGTATATATATTTGGTTCCCTAGTGATTCTATTTCTCTGAA	
12	1201	[TAGA]12[TG]5	[TCTA]12		NA	TGTGTGTGTGTTTAATTGTATGTATATATATTTGGTTCCCTAGTGATTCTATTTCTCTGAA	
12	1202	[TAGA]11 TAGG [TG]5	CCTA [TCTA]11		NA	TGTGTGTGTGTTTAATTGTATGTATATATATTTGGTTCCCTAGTGATTCTATTTCTCTGAA	
13	1301	[TAGA]13[TG]5	[TCTA]13		NA	TGTGTGTGTGTTTAATTGTATGTATATATATTTGGTTCCCTAGTGATTCTATTTCTCTGAA	
13	1302	[TAGA]12 TAGG [TG]5	CCTA [TCTA]12		NA	TGTGTGTGTGTTTAATTGTATGTATATATATTTGGTTCCCTAGTGATTCTATTTCTCTGAA	
13	1303	[TAGA]11 TAGC TAGA [TG]5	TCTA GCTA [TCTA]11		NA	TGTGTGTGTGTTTAATTGTATGTATATATATTTGGTTCCCTAGTGATTCTATTTCTCTGAA	
14	1401	[TAGA]14[TG]5	[TCTA]14		NA	TGTGTGTGTGTTTAATTGTATGTATATATATTTGGTTCCCTAGTGATTCTATTTCTCTGAA	
14	1402	[TAGA]13 TAGG [TG]5	CCTA [TCTA]13		NA	TGTGTGTGTGTTTAATTGTATGTATATATATTTGGTTCCCTAGTGATTCTATTTCTCTGAA	
14	1403	[TAGA]13 TAAG [TG]5	CTTA [TCTA]13		NA	TGTGTGTGTGTTTAATTGTATGTATATATATTTGGTTCCCTAGTGATTCTATTTCTCTGAA	
14.3	14.301	[TAGA]4 TGA [TAGA]9 TAGG [TG]5	CCTA [TCTA]9 TCA [TCTA]4		NA	TGTGTGTGTGTTTAATTGTATGTATATATATTTGGTTCCCTAGTGATTCTATTTCTCTGAA	
14.3	14.302	[TAGA]2 TGA [TAGA]11 TAGG [TG]5	CCTA [TCTA]11 TCA [TCTA]2		NA	TGTGTGTGTGTTTAATTGTATGTATATATATTTGGTTCCCTAGTGATTCTATTTCTCTGAA	
15	1501	[TAGA]14 TAAG [TG]5	CTTA [TCTA]14		NA	TGTGTGTGTGTTTAATTGTATGTATATATATTTGGTTCCCTAGTGATTCTATTTCTCTGAA	
15	1502	[TAGA]15[TG]5	[TCTA]15		NA	TGTGTGTGTGTTTAATTGTATGTATATATATTTGGTTCCCTAGTGATTCTATTTCTCTGAA	
15	1503	[TAGA]14 TAGG [TG]5	CCTA [TCTA]14		NA	TGTGTGTGTGTTTAATTGTATGTATATATATTTGGTTCCCTAGTGATTCTATTTCTCTGAA	
15.3	15.301	[TAGA]4 TGA [TAGA]10 TAGG [TG]5	CCTA [TCTA]10 TCA [TCTA]4		NA	TGTGTGTGTGTTTAATTGTATGTATATATATTTGGTTCCCTAGTGATTCTATTTCTCTGAA	
15.3	15.302	[TAGA]3 TGA [TAGA]11 TAGG [TG]5	CCTA [TCTA]11 TCA [TCTA]3		NA	TGTGTGTGTGTTTAATTGTATGTATATATATTTGGTTCCCTAGTGATTCTATTTCTCTGAA	
15.3	15.303	[TAGA]2 TGA [TAGA]12 TAGG [TG]5	CCTA [TCTA]12 TCA [TCTA]2		NA	TGTGTGTGTGTTTAATTGTATGTATATATATTTGGTTCCCTAGTGATTCTATTTCTCTGAA	
16	1601	[TAGA]15 TAAG [TG]5	CTTA [TCTA]15		NA	TGTGTGTGTGTTTAATTGTATGTATATATATTTGGTTCCCTAGTGATTCTATTTCTCTGAA	
16	1602	[TAGA]16 [TG]5	[TCTA]16		NA	TGTGTGTGTGTTTAATTGTATGTATATATATTTGGTTCCCTAGTGATTCTATTTCTCTGAA	
16	1603	[TAGA]15 TAGG [TG]5	CCTA [TCTA]15		NA	TGTGTGTGTGTTTAATTGTATGTATATATATTTGGTTCCCTAGTGATTCTATTTCTCTGAA	
16.1	16.1	[TAGA]15 [TAAG] G [TG]5	C CTTA [TCTA]15		NA	TGTGTGTGTGTTTAATTGTATGTATATATATTTGGTTCCCTAGTGATTCTATTTCTCTGAA	

16.3	16.3	[TAGA]4 TGA [TAGA]11 TAGG [TG]5	CCTA [TCTA]11 TCA [TCTA]4		NA	TGTGTGTGTGTTTAATTGTATGTATATATATTTGGTTCCCTAGTGATTCTATTTCTCTGAA	
17	1701a	[TAGA]16 TAGG [TG]5	CCTA [TCTA]16		NA	TGTGTGTGTGTTTAATTGTATGTATATATATTTGGTTCCCTAGTGATTCTATTTCTCTGAA	
17	1701b	[TAGA]16 TAGG [TG]5	CCTA [TCTA]16		NA	TGTGTGTGTGTTTAATTGTATGTATATATATTTGGTTCCCTAGTGATTCCATTTCTCTGAA	rs541123499
17	1702	[TAGA]17 [TG]5	[TCTA]17		NA	TGTGTGTGTGTTTAATTGTATGTATATATATTTGGTTCCCTAGTGATTCTATTTCTCTGAA	
17.3	17.3	[TAGA]4 TGA [TAGA]12 TAGG [TG]5	CCTA [TCTA]12 TCA [TCTA]4		NA	TGTGTGTGTGTTTAATTGTATGTATATATATTTGGTTCCCTAGTGATTCTATTTCTCTGAA	
18	1801	[TAGA]18 [TG]5	[TCTA]18		NA	TGTGTGTGTGTTTAATTGTATGTATATATATTTGGTTCCCTAGTGATTCTATTTCTCTGAA	
18	1802	[TAGA]17 TAGG [TG]5	CCTA [TCTA]17		NA	TGTGTGTGTGTTTAATTGTATGTATATATATTTGGTTCCCTAGTGATTCTATTTCTCTGAA	
18.3	18.3	[TAGA]4 TGA [TAGA]13 TAGG [TG]5	CCTA [TCTA]13 TCA [TCTA]4		NA	TGTGTGTGTGTTTAATTGTATGTATATATATTTGGTTCCCTAGTGATTCTATTTCTCTGAA	
19.3	19.3	[TAGA]4 TGA [TAGA]14 TAGG [TG]5	CCTA [TCTA]14 TCA [TCTA]4		NA	TGTGTGTGTGTTTAATTGTATGTATATATATTTGGTTCCCTAGTGATTCTATTTCTCTGAA	
20.3	20.3	[TAGA]4 TGA [TAGA]15 TAGG [TG]5	CCTA [TCTA]15 TCA [TCTA]4		NA	TGTGTGTGTGTTTAATTGTATGTATATATATTTGGTTCCCTAGTGATTCTATTTCTCTGAA	

TPOX							
Allele (LB)	Allele (SB)	Bracket			Left Flank	Right Flank	
6	6	[AATG]6			NA	TTTGG	
7	7	[AATG]7			NA	TTTGG	
8	8	[AATG]8			NA	TTTGG	
9	9	[AATG]9			NA	TTTGG	
10	10	[AATG]10			NA	TTTGG	
11	11	[AATG]11			NA	TTTGG	
12	12	[AATG]12			NA	TTTGG	
13	13	[AATG]13			NA	TTTGG	
14	14	[AATG]14			NA	TTTGG	

D2S441							
Allele (LB)	Allele (SB)	Bracket			Left Flank	Right Flank	
7	7	[TCTA]7			CCAGGAACTGTGGCTCATCTATGAAAACT	TATCATAACACCACAGCCACTTA	
8	8	[TCTA]8			CCAGGAACTGTGGCTCATCTATGAAAACT	TATCATAACACCACAGCCACTTA	
9	9a	[TCTA]9			CCAGGAACTGTGGCTCATCTATGAAAACT	TATCATAACACCACAGCCACTTA	
9	9b	[TCTA]9		rs74640515	CCAGAAACTGTGGCTCATCTATGAAAACT	TATCATAACACCACAGCCACTTA	
9.1	9.1	A [TCTA]9		rs74640515	CCAGAAACTGTGGCTCATCTATGAAAACT	TATCATAACACCACAGCCACTTA	
10	1001a	[TCTA]10			CCAGGAACTGTGGCTCATCTATGAAAACT	TATCATAACACCACAGCCACTTA	
10	1001b	[TCTA]10		rs74640515	CCAGAAACTGTGGCTCATCTATGAAAACT	TATCATAACACCACAGCCACTTA	
10	1002	[TCTA]8 TCTG TCTA			CCAGGAACTGTGGCTCATCTATGAAAACT	TATCATAACACCACAGCCACTTA	
11	1101a	[TCTA]11			CCAGGAACTGTGGCTCATCTATGAAAACT	TATCATAACACCACAGCCACTTA	
11	1101b	[TCTA]11		rs74640515	CCAGAAACTGTGGCTCATCTATGAAAACT	TATCATAACACCACAGCCACTTA	
11	1102	[TCTA]9 TCTG TCTA			CCAGGAACTGTGGCTCATCTATGAAAACT	TATCATAACACCACAGCCACTTA	
11.3	11.301	[TCTA]4 TCA [TCTA]7			CCAGGAACTGTGGCTCATCTATGAAAACT	TATCATAACACCACAGCCACTTA	
11.3	11.302	[TCTA]3 TCA [TCTA]8			CCAGGAACTGTGGCTCATCTATGAAAACT	TATCATAACACCACAGCCACTTA	
12	1201	[TCTA]12			CCAGGAACTGTGGCTCATCTATGAAAACT	TATCATAACACCACAGCCACTTA	
12	1202	[TCTA]10 TCTG TCTA			CCAGGAACTGTGGCTCATCTATGAAAACT	TATCATAACACCACAGCCACTTA	

12	1203	[TCTA]9 TTTA [TCTA]2			CCAGGAACTGTGGCTCATCTATGAAAACT	TATCATAACACCACAGCCACTTA	
12.3	12.301	[TCTA]4 TCA [TCTA]8			CCAGGAACTGTGGCTCATCTATGAAAACT	TATCATAACACCACAGCCACTTA	
12.3	12.302	[TCTA]4 TCATCCA [TCTA]7			CCAGGAACTGTGGCTCATCTATGAAAACT	TATCATAACACCACAGCCACTTA	
13	1301	[TCTA]13			CCAGGAACTGTGGCTCATCTATGAAAACT	TATCATAACACCACAGCCACTTA	
13	1302	[TCTA]10 TTTA [TCTA]2			CCAGGAACTGTGGCTCATCTATGAAAACT	TATCATAACACCACAGCCACTTA	
13	1303	[TCTA]11 TCTG TCTA			CCAGGAACTGTGGCTCATCTATGAAAACT	TATCATAACACCACAGCCACTTA	
13.3	13.3	[TCTA]4 TCA [TCTA]9			CCAGGAACTGTGGCTCATCTATGAAAACT	TATCATAACACCACAGCCACTTA	
14	1401	[TCTA]9 CCTA TCTA TTTA [TCTA]2			CCAGGAACTGTGGCTCATCTATGAAAACT	TATCATAACACCACAGCCACTTA	
14	1402	[TCTA]11 TTTA [TCTA]2			CCAGGAACTGTGGCTCATCTATGAAAACT	TATCATAACACCACAGCCACTTA	
14	1403	[TCTA]11 TTTA TCTA TGTA			CCAGGAACTGTGGCTCATCTATGAAAACT	TATCATAACACCACAGCCACTTA	
15	15	[TCTA]12 TTTA [TCTA]2			CCAGGAACTGTGGCTCATCTATGAAAACT	TATCATAACACCACAGCCACTTA	
16	1601	[TCTA]13 TTTA [TCTA]2			CCAGGAACTGTGGCTCATCTATGAAAACT	TATCATAACACCACAGCCACTTA	
16	1602	[TCTA]13 TTTA TCTA TGTA			CCAGGAACTGTGGCTCATCTATGAAAACT	TATCATAACACCACAGCCACTTA	

D2S1338							
Allele (LB)	Allele (SB)	Bracket (UAS)	STRSeq Record Description		Left Flank	Right Flank	
14	14	[TGCC]6[TTCC]8	[GGAA]8 [GGCA]6		AAATGGCTTGGCCTTGCC	CTC	
16	1601	[TGCC]7[TTCC]9	[GGAA]9 [GGCA]7		AAATGGCTTGGCCTTGCC	CTC	
16	1602	[TGCC]6[TTCC]10	[GGAA]10 [GGCA]6		AAATGGCTTGGCCTTGCC	CTC	
16	1603	[TGCC]5[TTCC]11	[GGAA]11 [GGCA]5		AAATGGCTTGGCCTTGCC	CTC	
16	1604	[TGCC]4[TTCC]12	[GGAA]12 [GGCA]4		AAATGGCTTGGCCTTGCC	CTC	
17	1701	[TGCC]6[TTCC]11	[GGAA]11 [GGCA]6		AAATGGCTTGGCCTTGCC	CTC	
17	1702	[TGCC]5[TTCC]12	[GGAA]12 [GGCA]5		AAATGGCTTGGCCTTGCC	CTC	
17	1703	[TGCC]4[TTCC]13	[GGAA]13 [GGCA]4		AAATGGCTTGGCCTTGCC	CTC	
17	1704	[TGCC]7[TTCC]10	[GGAA]10 [GGCA]7		AAATGGCTTGGCCTTGCC	CTC	
18	1801	[TGCC]7[TTCC]11	[GGAA]11 [GGCA]7		AAATGGCTTGGCCTTGCC	CTC	
18	1802	[TGCC]6[TTCC]12	[GGAA]12 [GGCA]6		AAATGGCTTGGCCTTGCC	CTC	
18	1803	[TGCC]4[TTCC]14	[GGAA]14 [GGCA]4		AAATGGCTTGGCCTTGCC	CTC	
18	1804	[TGCC]5[TTCC]13	[GGAA]13 [GGCA]5		AAATGGCTTGGCCTTGCC	CTC	
18	1805	[TGCC]3[TTCC]15	[GGAA]15 [GGCA]3		AAATGGCTTGGCCTTGCC	CTC	
18	1806	[TGCC]2[TTCC]16	[GGAA]16 [GGCA]2		AAATGGCTTGGCCTTGCC	CTC	
18	1807	[TGCC]7[TTCC]8[GTCC][TTCC]2	[GGAA]2 GGAC [GGAA]8 [GGCA]7		AAATGGCTTGGCCTTGCC	CTC	
19	1901	[TGCC]8[TTCC]11	[GGAA]11 [GGCA]8		AAATGGCTTGGCCTTGCC	CTC	
19	1902	[TGCC]7[TTCC]12	[GGAA]12 [GGCA]7		AAATGGCTTGGCCTTGCC	CTC	
19	1903	[TGCC]6[TTCC]10[GTCC][TTCC]2	[GGAA]2 GGAC [GGAA]10 [GGCA]6		AAATGGCTTGGCCTTGCC	CTC	
19	1904	[TGCC]6[TTCC]13	[GGAA]13 [GGCA]6		AAATGGCTTGGCCTTGCC	CTC	
19	1905	[TGCC]5[TTCC]14	[GGAA]14 [GGCA]5		AAATGGCTTGGCCTTGCC	CTC	
19	1906	[TGCC]7[TCCC][TTCC]11	[GGAA]11 GGGA [GGCA]7		AAATGGCTTGGCCTTGCC	CTC	
19	1907	[TGCC]7[TTCC]9[GTCC][TTCC]2	[GGAA]2 GGAC [GGAA]9 [GGCA]7		AAATGGCTTGGCCTTGCC	CTC	
19	1908	[TGCC]3[TTCC]16	[GGAA]16 [GGCA]3		AAATGGCTTGGCCTTGCC	CTC	

19	1909	[TGCC]4[TTCC]15	[GGAA]15 [GGCA]4		AAATGGCTTGGCCTTGCC	CTC	
20	2001	[TGCC]7[TCCC][TTCC]12	[GGAA]12 GGGA [GGCA]7		AAATGGCTTGGCCTTGCC	CTC	
20	2002	[TGCC]8[TTCC]12	[GGAA]12 [GGCA]8		AAATGGCTTGGCCTTGCC	CTC	
20	2003	[TGCC]7[TTCC]10[GTCC][TTCC]2	[GGAA]2 GGAC [GGAA]10 [GGCA]7		AAATGGCTTGGCCTTGCC	CTC	
20	2004	[TGCC]7[TTCC]13	[GGAA]13 [GGCA]7		AAATGGCTTGGCCTTGCC	CTC	
20	2005	[TGCC]7[TTCC]2[TTTC][TTCC]10	[GGAA]10 GAAA [GGAA]2 [GGCA]7		AAATGGCTTGGCCTTGCC	CTC	
20	2006	[TGCC]6[TTCC]11[GTCC][TTCC]2	[GGAA]2 GGAC [GGAA]11 [GGCA]6		AAATGGCTTGGCCTTGCC	CTC	
20	2007	[TGCC]6[TTCC]14	[GGAA]14 [GGCA]6		AAATGGCTTGGCCTTGCC	CTC	
20	2008	[TGCC]4[TTCC]16	[GGAA]16 [GGCA]4		AAATGGCTTGGCCTTGCC	CTC	
20	2009	[TGCC]5[TTCC]15	[GGAA]15 [GGCA]5		AAATGGCTTGGCCTTGCC	CTC	
21	2100	[TGCC]7[TCCC][TTCC]13	[GGAA]13 GGGA [GGCA]7		AAATGGCTTGGCCTTGCC	CTC	
21	2101	[TGCC]8[TTCC]13	[GGAA]13 [GGCA]8		AAATGGCTTGGCCTTGCC	CTC	
21	2102	[TGCC]7[TTCC]11[GTCC][TTCC]2	[GGAA]2 GGAC [GGAA]11 [GGCA]7		AAATGGCTTGGCCTTGCC	CTC	
21	2103	[TGCC]7[TTCC]14	[GGAA]14 [GGCA]7		AAATGGCTTGGCCTTGCC	CTC	
21	2104	[TGCC]7[TTCC]2[TTTC][TTCC]11	[GGAA]11 GAAA [GGAA]2 [GGCA]7		AAATGGCTTGGCCTTGCC	CTC	
21	2105	[TGCC]6[TTCC]12[GTCC][TTCC]2	[GGAA]2 GGAC [GGAA]12 [GGCA]6		AAATGGCTTGGCCTTGCC	CTC	
21	2106	[TGCC]6[TTCC]15	[GGAA]15 [GGCA]6		AAATGGCTTGGCCTTGCC	CTC	
21	2107	[TGCC]8[TTCC]10[GTCC][TTCC]2	[GGAA]2 GGAC [GGAA]10 [GGCA]8		AAATGGCTTGGCCTTGCC	CTC	
21	2108	[TGCC]5[TTCC]16	[GGAA]16 [GGCA]5		AAATGGCTTGGCCTTGCC	CTC	
21	2109	[TGCC]9[TTCC]12	[GGAA]12 [GGCA]9		AAATGGCTTGGCCTTGCC	CTC	
21	2110	[TGCC]4[TTCC]17	[GGAA]17 [GGCA]4		AAATGGCTTGGCCTTGCC	CTC	
22	2201	[TGCC]9[TTCC]13	[GGAA]13 [GGCA]9		AAATGGCTTGGCCTTGCC	CTC	
22	2202	[TGCC]9[TTCC]6[TTTC][TTCC]6	[GGAA]6 GAAA [GGAA]6 [GGCA]9		AAATGGCTTGGCCTTGCC	CTC	
22	2203	[TGCC]8[TTCC]14	[GGAA]14 [GGCA]8		AAATGGCTTGGCCTTGCC	CTC	
22	2204	[TGCC]7[TTCC]12[GTCC][TTCC]2	[GGAA]2 GGAC [GGAA]12 [GGCA]7		AAATGGCTTGGCCTTGCC	CTC	
22	2205	[TGCC]7[TTCC]15	[GGAA]15 [GGCA]7		AAATGGCTTGGCCTTGCC	CTC	
22	2206	[TGCC]6[TTCC]13[GTCC][TTCC]2	[GGAA]2 GGAC [GGAA]13 [GGCA]6		AAATGGCTTGGCCTTGCC	CTC	
22	2207	[TGCC]5[TTCC]14[GTCC][TTCC]2	[GGAA]2 GGAC [GGAA]14 [GGCA]5		AAATGGCTTGGCCTTGCC	CTC	
22	2208	[TGCC]4[TTCC]15[GTCC][TTCC]2	[GGAA]2 GGAC [GGAA]15 [GGCA]4		AAATGGCTTGGCCTTGCC	CTC	
22	2209	[TGCC]6[TTCC]16	[GGAA]16 [GGCA]6		AAATGGCTTGGCCTTGCC	CTC	
23	2301	[TGCC]9[TTCC]14	[GGAA]14 [GGCA]9		AAATGGCTTGGCCTTGCC	CTC	
23	2302	[TGCC]8[TTCC]12[GTCC][TTCC]2	[GGAA]2 GGAC [GGAA]12 [GGCA]8		AAATGGCTTGGCCTTGCC	CTC	
23	2303	[TGCC]7[TTCC]13[GTCC][TTCC]2	[GGAA]2 GGAC [GGAA]13 [GGCA]7		AAATGGCTTGGCCTTGCC	CTC	
23	2304	[TGCC]7[TTCC]16	[GGAA]16 [GGCA]7		AAATGGCTTGGCCTTGCC	CTC	
23	2305	[TGCC]6[TTCC]14[GTCC][TTCC]2	[GGAA]2 GGAC [GGAA]14 [GGCA]6		AAATGGCTTGGCCTTGCC	CTC	
23	2306	[TGCC]5[TTCC]15[GTCC][TTCC]2	[GGAA]2 GGAC [GGAA]15 [GGCA]5		AAATGGCTTGGCCTTGCC	CTC	
23	2307	[TGCC]4[TTCC]16[GTCC][TTCC]2	[GGAA]2 GGAC [GGAA]16 [GGCA]4		AAATGGCTTGGCCTTGCC	CTC	
23	2308	[TGCC]7[TTCC][TGCC]3[TTCC]12	[GGAA]12 [GGCA]3 GGAA [GGCA]7		AAATGGCTTGGCCTTGCC	CTC	
23	2309	[TGCC]8[TTCC]15	[GGAA]15 [GGCA]8		AAATGGCTTGGCCTTGCC	CTC	
24	2401	[TGCC]9[TTCC]15	[GGAA]15 [GGCA]9		AAATGGCTTGGCCTTGCC	CTC	
24	2402	[TGCC]8[TTCC]13[GTCC][TTCC]2	[GGAA]2 GGAC [GGAA]13 [GGCA]8		AAATGGCTTGGCCTTGCC	CTC	

24	2403	[TGCC]7[TTCC]14[GTCC][TTCC]2	[GGAA]2 GGAC [GGAA]14 [GGCA]7		AAATGGCTTGGCCTTGCC	CTC	
24	2404	[TGCC]6[TTCC]15[GTCC][TTCC]2	[GGAA]2 GGAC [GGAA]15 [GGCA]6		AAATGGCTTGGCCTTGCC	CTC	
24	2405	[TGCC]5[TTCC]16[GTCC][TTCC]2	[GGAA]2 GGAC [GGAA]16 [GGCA]5		AAATGGCTTGGCCTTGCC	CTC	
24	2406	[TGCC]8[TTCC]16	[GGAA]16 [GGCA]8		AAATGGCTTGGCCTTGCC	CTC	
24	2407	[TGCC]7[TTCC]13[TTTC][GTCC][TTCC]2	[GGAA]2 GGAC AGAA [GGAA]13 [GGCA]7		AAATGGCTTGGCCTTGCC	CTC	
25	2501	[TGCC]8[TTCC]14[GTCC][TTCC]2	[GGAA]2 GGAC [GGAA]14 [GGCA]8		AAATGGCTTGGCCTTGCC	CTC	
25	2502	[TGCC]7[TTCC]15[GTCC][TTCC]2	[GGAA]2 GGAC [GGAA]15 [GGCA]7		AAATGGCTTGGCCTTGCC	CTC	
25	2503	[TGCC]6[TTCC]16[GTCC][TTCC]2	[GGAA]2 GGAC [GGAA]16 [GGCA]6		AAATGGCTTGGCCTTGCC	CTC	
25	2504	[TGCC]9[TTCC]13[GTCC][TTCC]2	[GGAA]2 GGAC [GGAA]13 [GGCA]9		AAATGGCTTGGCCTTGCC	CTC	
25	2505	[TGCC]5[TTCC]17[GTCC][TTCC]2	[GGAA]2 GGAC [GGAA]17 [GGCA]5		AAATGGCTTGGCCTTGCC	CTC	
26	2601	[TGCC]8[TTCC]15[GTCC][TTCC]2	[GGAA]2 GGAC [GGAA]15 [GGCA]8		AAATGGCTTGGCCTTGCC	CTC	
26	2602	[TGCC]7[TTCC]16[GTCC][TTCC]2	[GGAA]2 GGAC [GGAA]16 [GGCA]7		AAATGGCTTGGCCTTGCC	CTC	
26	2603	[TGCC]6[TTCC]17[GTCC][TTCC]2	[GGAA]2 GGAC [GGAA]17 [GGCA]6		AAATGGCTTGGCCTTGCC	CTC	
27	2701	[TGCC]7[TTCC]17[GTCC][TTCC]2	[GGAA]2 GGAC [GGAA]17 [GGCA]7		AAATGGCTTGGCCTTGCC	CTC	
27	2702	[TGCC]8[TTCC]16[GTCC][TTCC]2	[GGAA]2 GGAC [GGAA]16 [GGCA]8		AAATGGCTTGGCCTTGCC	CTC	
28	2801	[TGCC]7[TTCC]18[GTCC][TTCC]2	[GGAA]2 GGAC [GGAA]18 [GGCA]7		AAATGGCTTGGCCTTGCC	CTC	

D3S1358							
Allele (LB)	Allele (SB)	Bracket			Left Flank	Right Flank	
11	11	TCTA [TCTG]2 [TCTA]8			TTTGGGGGCATCTCTTATACTCATGAAATCAACAGAGGCTTGCATGTA	TGAGACAGGGTCTTGCTC	
12	1201	TCTA [TCTG]1 [TCTA]10			TTTGGGGGCATCTCTTATACTCATGAAATCAACAGAGGCTTGCATGTA	TGAGACAGGGTCTTGCTC	
12	1202	TCTA [TCTG]2 [TCTA]9			TTTGGGGGCATCTCTTATACTCATGAAATCAACAGAGGCTTGCATGTA	TGAGACAGGGTCTTGCTC	
13	13	TCTA [TCTG]1 [TCTA]11			TTTGGGGGCATCTCTTATACTCATGAAATCAACAGAGGCTTGCATGTA	TGAGACAGGGTCTTGCTC	
14	1401	TCTA TCTG [TCTA]12			TTTGGGGGCATCTCTTATACTCATGAAATCAACAGAGGCTTGCATGTA	TGAGACAGGGTCTTGCTC	
14	1402	TCTA [TCTG]2 [TCTA]11			TTTGGGGGCATCTCTTATACTCATGAAATCAACAGAGGCTTGCATGTA	TGAGACAGGGTCTTGCTC	
14	1403	TCTA [TCTG]3 [TCTA]10			TTTGGGGGCATCTCTTATACTCATGAAATCAACAGAGGCTTGCATGTA	TGAGACAGGGTCTTGCTC	
15	1501	TCTA TCTG [TCTA]13			TTTGGGGGCATCTCTTATACTCATGAAATCAACAGAGGCTTGCATGTA	TGAGACAGGGTCTTGCTC	
15	1502	TCTA [TCTG]2 [TCTA]12			TTTGGGGGCATCTCTTATACTCATGAAATCAACAGAGGCTTGCATGTA	TGAGACAGGGTCTTGCTC	
15	1503	TCTA [TCTG]3 [TCTA]11			TTTGGGGGCATCTCTTATACTCATGAAATCAACAGAGGCTTGCATGTA	TGAGACAGGGTCTTGCTC	
16	1601	TCTA TCTG [TCTA]14			TTTGGGGGCATCTCTTATACTCATGAAATCAACAGAGGCTTGCATGTA	TGAGACAGGGTCTTGCTC	
16	1602	TCTA [TCTG]2 [TCTA]13			TTTGGGGGCATCTCTTATACTCATGAAATCAACAGAGGCTTGCATGTA	TGAGACAGGGTCTTGCTC	
16	1603	TCTA [TCTG]3 [TCTA]12			TTTGGGGGCATCTCTTATACTCATGAAATCAACAGAGGCTTGCATGTA	TGAGACAGGGTCTTGCTC	
16	1604	TCTA [TCTG]3 TCTA TCTG [TCTA]10			TTTGGGGGCATCTCTTATACTCATGAAATCAACAGAGGCTTGCATGTA	TGAGACAGGGTCTTGCTC	
16	1605	TCTA [TCTG]4 [TCTA]11			TTTGGGGGCATCTCTTATACTCATGAAATCAACAGAGGCTTGCATGTA	TGAGACAGGGTCTTGCTC	
16.2	16.2	TCTA [TCTG]3 TC [TCTA]12			TTTGGGGGCATCTCTTATACTCATGAAATCAACAGAGGCTTGCATGTA	TGAGACAGGGTCTTGCTC	
17	1701	[TCTA]2 [TCTG]3 [TCTA]12			TTTGGGGGCATCTCTTATACTCATGAAATCAACAGAGGCTTGCATGTA	TGAGACAGGGTCTTGCTC	
17	1702	TCTA TCTG [TCTA]15			TTTGGGGGCATCTCTTATACTCATGAAATCAACAGAGGCTTGCATGTA	TGAGACAGGGTCTTGCTC	
17	1703	TCTA [TCTG]2 [TCTA]14			TTTGGGGGCATCTCTTATACTCATGAAATCAACAGAGGCTTGCATGTA	TGAGACAGGGTCTTGCTC	
17	1704	TCTA [TCTG]2 TCTC [TCTA]13			TTTGGGGGCATCTCTTATACTCATGAAATCAACAGAGGCTTGCATGTA	TGAGACAGGGTCTTGCTC	
17	1705a	TCTA [TCTG]3 [TCTA]13			TTTGGGGGCATCTCTTATACTCATGAAATCAACAGAGGCTTGCATGTA	TGAGACAGGGTCTTGCTC	

17	1705b	TCTA [TCTG]3 [TCTA]13		rs1559501767	TTTGGGGGCATCTCTTATACTCATGAAATCAACAGATGCTTGCATGTA	TGAGACAGGGTCTTGCTC	
17	1706	TCTA [TCTG]4 [TCTA]12			TTTGGGGGCATCTCTTATACTCATGAAATCAACAGAGGCTTGCATGTA	TGAGACAGGGTCTTGCTC	
18	1801	TCTA [TCTG]1 [TCTA]16			TTTGGGGGCATCTCTTATACTCATGAAATCAACAGAGGCTTGCATGTA	TGAGACAGGGTCTTGCTC	
18	1802	TCTA [TCTG]2 [TCTA]15			TTTGGGGGCATCTCTTATACTCATGAAATCAACAGAGGCTTGCATGTA	TGAGACAGGGTCTTGCTC	
18	1803	TCTA [TCTG]3 [TCTA]14			TTTGGGGGCATCTCTTATACTCATGAAATCAACAGAGGCTTGCATGTA	TGAGACAGGGTCTTGCTC	
18	1804	TCTA [TCTG]4 [TCTA]13			TTTGGGGGCATCTCTTATACTCATGAAATCAACAGAGGCTTGCATGTA	TGAGACAGGGTCTTGCTC	
18	1805	TCTA [TCTG]2 TCTC [TCTA]14			TTTGGGGGCATCTCTTATACTCATGAAATCAACAGAGGCTTGCATGTA	TGAGACAGGGTCTTGCTC	
18.2	18.2	TCTA [TCTG]3 TC [TCTA]14			TTTGGGGGCATCTCTTATACTCATGAAATCAACAGAGGCTTGCATGTA	TGAGACAGGGTCTTGCTC	
19	1901	TCTA [TCTG]2 [TCTA]16			TTTGGGGGCATCTCTTATACTCATGAAATCAACAGAGGCTTGCATGTA	TGAGACAGGGTCTTGCTC	
19	1902	TCTA [TCTG]3 [TCTA]15			TTTGGGGGCATCTCTTATACTCATGAAATCAACAGAGGCTTGCATGTA	TGAGACAGGGTCTTGCTC	

D4S2408							
Allele (LB)	Allele (SB)	Bracket			Left Flank	Right Flank	
7	7	[ATCT]7			CTATGC	AATGGTTA	
8	8	[ATCT]8			CTATGC	AATGGTTA	
9	901	[ATCT]9			CTATGC	AATGGTTA	
9	902	ATCT GTCT [ATCT]7			CTATGC	AATGGTTA	
10	1001	[ATCT]10			CTATGC	AATGGTTA	
10	1002	ATCT GTCT [ATCT]8			CTATGC	AATGGTTA	
11	11	[ATCT]11			CTATGC	AATGGTTA	
12	1201	[ATCT]12			CTATGC	AATGGTTA	
12	1202	[ATCT]8 CTCT [ATCT]3			CTATGC	AATGGTTA	
12	1203	[ATCT]8 ATTT [ATCT]3			CTATGC	AATGGTTA	
13	13	[ATCT]13			CTATGC	AATGGTTA	

FGA							
Allele (LB)	Allele (SB)	Bracket (UAS)	STRSeq Record Description		Left Flank	Right Flank	
17	17	[TTTC]3[TTTT][TTCT][CTTT]9[CTCC][TTCC]2	[GGAA]2 GGAG [AAAG]9 AGAA AAAA [GAAA]3		GCATATTTACAAGCTAG	TTTCTTCCTTTCTTTTTTGCTGG	
18	18	[TTTC]3[TTTT][TTCT][CTTT]10[CTCC][TTCC]2	[GGAA]2 GGAG [AAAG]10 AGAA AAAA [GAAA]3		GCATATTTACAAGCTAG	TTTCTTCCTTTCTTTTTTGCTGG	
18.2	18.2	[TTTC]3[TTTT][TT][CTTT]11[CTCC][TTCC]2	[GGAA]2 GGAG [AAAG]11 AA AAAA [GAAA]3		GCATATTTACAAGCTAG	TTTCTTCCTTTCTTTTTTGCTGG	
19	19	[TTTC]3[TTTT][TTCT][CTTT]11[CTCC][TTCC]2	[GGAA]2 GGAG [AAAG]11 AGAA AAAA [GAAA]3		GCATATTTACAAGCTAG	TTTCTTCCTTTCTTTTTTGCTGG	
19.2	19.2	[TTTC]3[TTTT][TT][CTTT]12[CTCC][TTCC]2	[GGAA]2 GGAG [AAAG]12 AA AAAA [GAAA]3		GCATATTTACAAGCTAG	TTTCTTCCTTTCTTTTTTGCTGG	
20	20	[TTTC]3[TTTT][TTCT][CTTT]12[CTCC][TTCC]2	[GGAA]2 GGAG [AAAG]12 AGAA AAAA [GAAA]3		GCATATTTACAAGCTAG	TTTCTTCCTTTCTTTTTTGCTGG	

20.2	20.2	[TTTC]3[TTTT][TT][CTTT]13[CTCC][TTCC]2	[GGAA]2 GGAG [AAAG]13 AA AAAA [GAAA]3		GCATATTTACAAGCTAG	TTTCTTCCTTTCTTTTTTGCTGG	
21	21	[TTTC]3[TTTT][TTCT][CTTT]13[CTCC][TTCC]2	[GGAA]2 GGAG [AAAG]13 AGAA AAAA [GAAA]3		GCATATTTACAAGCTAG	TTTCTTCCTTTCTTTTTTGCTGG	
21.2	21.2	[TTTC]3[TTTT][TT][CTTT]14[CTCC][TTCC]2	[GGAA]2 GGAG [AAAG]14 AA AAAA [GAAA]3		GCATATTTACAAGCTAG	TTTCTTCCTTTCTTTTTTGCTGG	
22	22	[TTTC]3[TTTT][TTCT][CTTT]14[CTCC][TTCC]2	[GGAA]2 GGAG [AAAG]14 AGAA AAAA [GAAA]3		GCATATTTACAAGCTAG	TTTCTTCCTTTCTTTTTTGCTGG	
22.2	22.2	[TTTC]3[TTTT][TT][CTTT]15[CTCC][TTCC]2	[GGAA]2 GGAG [AAAG]15 AA AAAA [GAAA]3		GCATATTTACAAGCTAG	TTTCTTCCTTTCTTTTTTGCTGG	
23	23	[TTTC]3[TTTT][TTCT][CTTT]15[CTCC][TTCC]2	[GGAA]2 GGAG [AAAG]15 AGAA AAAA [GAAA]3		GCATATTTACAAGCTAG	TTTCTTCCTTTCTTTTTTGCTGG	
23.2	23.2	[TTTC]3[TTTT][TT][CTTT]16[CTCC][TTCC]2	[GGAA]2 GGAG [AAAG]16 AA AAAA [GAAA]3		GCATATTTACAAGCTAG	TTTCTTCCTTTCTTTTTTGCTGG	
23.3	23.3	[TTTC]3[TTTT][TTCT][CTTT]15[CTT][CTTT][CTCC][TTCC]2	[GGAA]2 GGAG AAAG AAG [AAAG]14 AGAA AAAA [GAAA]3		GCATATTTACAAGCTAG	TTTCTTCCTTTCTTTTTTGCTGG	
24	24	[TTTC]3[TTTT][TTCT][CTTT]16[CTCC][TTCC]2	[GGAA]2 GGAG [AAAG]16 AGAA AAAA [GAAA]3		GCATATTTACAAGCTAG	TTTCTTCCTTTCTTTTTTGCTGG	
24.2	24.2	[TTTC]3[TTTT][TT][CTTT]17[CTCC][TTCC]2	[GGAA]2 GGAG [AAAG]17 AA AAAA [GAAA]3		GCATATTTACAAGCTAG	TTTCTTCCTTTCTTTTTTGCTGG	
24.3	24.3	[TTTC]3[TTTT][TTCT][CTTT]15[CTT][CTTT][CTCC][TTCC]2	[GGAA]2 GGAG AAAG AAG [AAAG]15 AGAA AAAA [GAAA]3		GCATATTTACAAGCTAG	TTTCTTCCTTTCTTTTTTGCTGG	
25	25	[TTTC]3[TTTT][TTCT][CTTT]17[CTCC][TTCC]2	[GGAA]2 GGAG [AAAG]17 AGAA AAAA [GAAA]3		GCATATTTACAAGCTAG	TTTCTTCCTTTCTTTTTTGCTGG	
25.2	25.2	[TTTC]3[TTTT][TT][CTTT]18[CTCC][TTCC]2	[GGAA]2 GGAG [AAAG]18 AA AAAA [GAAA]3		GCATATTTACAAGCTAG	TTTCTTCCTTTCTTTTTTGCTGG	
26	2601	[TTTC]3[TTTT][TTCT][CTTT]12[CCTT][CTTT]5[CTCC][TTCC]2	[GGAA]2 GGAG [AAAG]5 AAGG [AAAG]12 AGAA AAAA [GAAA]3		GCATATTTACAAGCTAG	TTTCTTCCTTTCTTTTTTGCTGG	
26	2602	[TTTC]3[TTTT][TTCT][CTTT]18[CTCC][TTCC]2	[GGAA]2 GGAG [AAAG]18 AGAA AAAA [GAAA]3		GCATATTTACAAGCTAG	TTTCTTCCTTTCTTTTTTGCTGG	
26	2603	[TTTC]3[TTTT][TTCT][CTTT]16[GTTT][CTTT][CTCC][TTCC]2	[GGAA]2 GGAG AAAG AAAC [AAAG]16 AGAA AAAA [GAAA]3		GCATATTTACAAGCTAG	TTTCTTCCTTTCTTTTTTGCTGG	
26.2	26.2	[TTTC]3[TTTT][TT][CTTT]19[CTCC][TTCC]2	[GGAA]2 GGAG [AAAG]19 AA AAAA [GAAA]3		GCATATTTACAAGCTAG	TTTCTTCCTTTCTTTTTTGCTGG	
27	2701	[TTTC]3[TTTT][TTCT][CTTT]13[CCTT][CTTT]5[CTCC][TTCC]2	[GGAA]2 GGAG [AAAG]5 AAGG [AAAG]13 AGAA AAAA [GAAA]3		GCATATTTACAAGCTAG	TTTCTTCCTTTCTTTTTTGCTGG	
27	2702	[TTTC]3[TTTT][TTCT][CTTT]19[CTCC][TTCC]2	[GGAA]2 GGAG [AAAG]19 AGAA AAAA [GAAA]3		GCATATTTACAAGCTAG	TTTCTTCCTTTCTTTTTTGCTGG	
28	2801	[TTTC]3[TTTT][TTCT][CTTT][CTCT][CTTT]18[CTCC][TTCC]2	[GGAA]2 GGAG [AAAG]18 AGAG AAAG AGAA AAAA [GAAA]4		GCATATTTACAAGCTAG	TTTCTTCCTTTCTTTTTTGCTGG	
28	2802	[TTTC]3[TTTT][TTCT][CTTT]20[CTCC][TTCC]2	[GGAA]2 GGAG [AAAG]20 AGAA AAAA [GAAA]3		GCATATTTACAAGCTAG	TTTCTTCCTTTCTTTTTTGCTGG	

28	2803	[TTTC]3[TTTT][TTCT][CTTT]14[CCTT][CTTT]5[CTCC][TTCC]2	[GGAA]2 GGAG [AAAG]5 AAGG [AAAG]14 AGAA AAAA [GAAA]3		GCATATTTACAAGCTAG	TTTCTTCCTTTCTTTTTTGCTGG	
28	2804	[TTTC]3[TTTT][TTCT][CTTT]8[CTTC][CTTT]5[CCTT][CTTT]5[CTCC][TTCC]2	[GGAA]2 GGAG [AAAG]5 AAGG [AAAG]5 GAAG [AAAG]8 AGAA AAAA [GAAA]3		GCATATTTACAAGCTAG	TTTCTTCCTTTCTTTTTTGCTGG	
28	2805	[TTTC]3[TTTT][TTCT][CTTT]8[CTTC]2[CTTT]4[CCTT][CTTT]5[CTCC][TTCC]2	[GGAA]2 GGAG [AAAG]5 AAGG [AAAG]4 [GAAG]2 [AAAG]8 AGAA AAAA [GAAA]3		GCATATTTACAAGCTAG	TTTCTTCCTTTCTTTTTTGCTGG	
29	29	[TTTC]3[TTTT][TTCT][CTTT]15[CCTT][CTTT]5[CTCC][TTCC]2	[GGAA]2 GGAG [AAAG]5 AAGG [AAAG]15 AGAA AAAA [GAAA]3		GCATATTTACAAGCTAG	TTTCTTCCTTTCTTTTTTGCTGG	
30	30	[TTTC]3[TTTT][TTCT][CTTT]16[CCTT][CTTT]5[CTCC][TTCC]2	[GGAA]2 GGAG [AAAG]5 AAGG [AAAG]16 AGAA AAAA [GAAA]3		GCATATTTACAAGCTAG	TTTCTTCCTTTCTTTTTTGCTGG	
31.2	31.2	[TTTC]4[TTTT][TT][CTTT]15[CTTC]3[CTTT]3[CTCC][TTCC]4	[GGAA]4 GGAG [AAAG]3 [GAAG]3 [AAAG]15 AA AAAA [GAAA]4		GCATATTTACAAGCTAG	TTTCTTCCTTTCTTTTTTGCTGG	
32.2	32.2	[TTTC]4[TTTT][TT][CTTT]16[CTTC]3[CTTT]3[CTCC][TTCC]4	[GGAA]4 GGAG [AAAG]3 [GAAG]3 [AAAG]16 AA AAAA [GAAA]4		GCATATTTACAAGCTAG	TTTCTTCCTTTCTTTTTTGCTGG	
42.2	42.2	[TTTC]4[TTTT][TT][CTTT]11[CTGT]2[CTTT]13[CTTC]3[CTTT]3[CTCC][TTCC]4	[GGAA]4 GGAG [AAAG]3 [GAAG]3 [AAAG]13 [ACAG]2 [AAAG]11 AA AAAA [GAAA]4		GCATATTTACAAGCTAG	TTTCTTCCTTTCTTTTTTGCTGG	
43.2	43.2	[TTTC]4[TTTT][TT][CTTT]9[CTGT]3[CTTT]15[CTTC]3[CTTT]3[CTCC][TTCC]4	[GGAA]4 GGAG [AAAG]3 [GAAG]3 [AAAG]15 [ACAG]3 [AAAG]9 AA AAAA [GAAA]4		GCATATTTACAAGCTAG	TTTCTTCCTTTCTTTTTTGCTGG	

D5S818

Allele (LB)	Allele (SB)	Bracket (UAS)	STRSeq Record Description		Left Flank	Right Flank	
7	701	[AGAT]7[AGAG]	[ATCT]7 rs25768		ATTTTGAAGAT	GTATAAATA	
7	702	[AGAT]7[AGAT]	[ATCT]7 rs73801920		ATTTTGAAGAT	GTATAAATA	
8	8	[ATCT]8[AGAT]	[ATCT]8 rs73801920 rs25768		ATTTTGAAGAT	GTATAAATA	
9	901	[AGAT]9[AGAG]	[ATCT]9 rs25768		ATTTTGAAGAT	GTATAAATA	
9	902	[AGAT]9[AGAT]	[ATCT]9 rs73801920 rs25768		ATTTTGAAGAT	GTATAAATA	
10	1001	[AGAT]10[AGAG]	[ATCT]10		ATTTTGAAGAT	GTATAAATA	
10	1002	[AGAT]10[AGAT]	[ATCT]10 rs73801920 rs25768		ATTTTGAAGAT	GTATAAATA	
11	1101	[AGAT]11[AGAG]	[ATCT]11		ATTTTGAAGAT	GTATAAATA	
11	1102	[AGAT]11[AGAT]	[ATCT]11 rs73801920 rs25768		ATTTTGAAGAT	GTATAAATA	
11.1	11.1	AGAT T [AGAT]10[AGAT]	[ATCT]10 A ATCT rs73801920		ATTTTGAAGAT	GTATAAATA	
12	1201	[AGAT]12[AGAG]	[ATCT]12		ATTTTGAAGAT	GTATAAATA	
12	1202	[AGAT]12[AGAT]	[ATCT]12 rs73801920 rs25768		ATTTTGAAGAT	GTATAAATA	
12	1203	[AGAT]2 AAAT [AGAT]9[AGAT]	[ATCT]9 ATTT [ATCT]2 rs73801920		ATTTTGAAGAT	GTATAAATA	
13	1301	[AGAT]9 ACAT [AGAT]3[AGAG]	[ATCT]3 ATGT [ATCT]9 rs25768		ATTTTGAAGAT	GTATAAATA	
13	1302	[AGAT]13[AGAG]	[ATCT]13		ATTTTGAAGAT	GTATAAATA	
13	1303	[AGAT]13[AGAT]	[ATCT]13 rs73801920 rs25768		ATTTTGAAGAT	GTATAAATA	

14	1401	[AGAT]14[AGAG]	[ATCT]14		ATTTTGAAGAT	GTATAAATA	
14	1402	[AGAT]14[AGAT]	[ATCT]14 rs73801920 rs25768		ATTTTGAAGAT	GTATAAATA	
15	1501	[AGAT]11 ACAT [AGAT]3[AGAG]	[ATCT]3 ATGT [ATCT]11 rs25768		ATTTTGAAGAT	GTATAAATA	
15	1502	[AGAT]15[AGAG]	[ATCT]15		ATTTTGAAGAT	GTATAAATA	
16	1601	[AGAT]12 ACAT [AGAT]3[AGAG]	[ATCT]3 ATGT [ATCT]12		ATTTTGAAGAT	GTATAAATA	

CSF1PO

Allele (LB)	Allele (SB)	Bracket (UAS)	STRSeq Record Description		Left Flank	Right Flank	
6	6	[AGAT]6	[ATCT]6		AAGATAGATAGATT	AGGAAG	
7	7	[AGAT]7	[ATCT]7		AAGATAGATAGATT	AGGAAG	
8	8	[AGAT]8	[ATCT]8		AAGATAGATAGATT	AGGAAG	
9	9	[AGAT]9	[ATCT]9		AAGATAGATAGATT	AGGAAG	
10	10	[AGAT]10	[ATCT]10		AAGATAGATAGATT	AGGAAG	
11	1101	[AGAT]11	[ATCT]11		AAGATAGATAGATT	AGGAAG	
11	1102	[AGAT]3[ATAT][AGAT]7	[ATCT]7 ATAT [ATCT]3		AAGATAGATAGATT	AGGAAG	
12	1201	[AGAT]12	[ATCT]12		AAGATAGATAGATT	AGGAAG	
12	1202	[AGAT]6[AGAC][AGAT]5	[ATCT]5 GTCT [ATCT]6		AAGATAGATAGATT	AGGAAG	
12	1203	[AGAT]7[AGAC][AGAT]4	[ATCT]4 GTCT [ATCT]7		AAGATAGATAGATT	AGGAAG	
12	1204	[AGAT]8[ACCT][AGAT]3	[ATCT]8 ACCT [ATCT]3		AAGATAGATAGATT	AGGAAG	
13	13	[AGAT]13	[ATCT]13		AAGATAGATAGATT	AGGAAG	
14	14	[AGAT]14	[ATCT]14		AAGATAGATAGATT	AGGAAG	
15	15	[AGAT]15	[ATCT]15		AAGATAGATAGATT	AGGAAG	

D6S1043

Allele (LB)	Allele (SB)	Bracket (UAS)	STRSeq Record Description		Left Flank	Right Flank	
9	9	[AGAT]9	[ATCT]9		GATCAATAGATTGATAGATC	AGGATTTATTATGGGAAGTGGCTCATACAATTATGGAAGCTGAGAAATTTCACAATATGCCATCT	
10	10	[AGAT]10	[ATCT]10		GATCAATAGATTGATAGATC	AGGATTTATTATGGGAAGTGGCTCATACAATTATGGAAGCTGAGAAATTTCACAATATGCCATCT	
11	11	[AGAT]11	[ATCT]11		GATCAATAGATTGATAGATC	AGGATTTATTATGGGAAGTGGCTCATACAATTATGGAAGCTGAGAAATTTCACAATATGCCATCT	
12	1201a	[AGAT]12	[ATCT]12		GATCAATAGATTGATAGATC	AGGATTTATTATGGGAAGTGGCTCATACAATTATGGAAGCTGAGAAATTTCACAATATGCCATCT	
12	1201b	[AGAT]12	[ATCT]12	rs529713981	GATCAATAGATTGATAGATT	AGGATTTATTATGGGAAGTGGCTCATACAATTATGGAAGCTGAGAAATTTCACAATATGCCATCT	
12	1202	[AGAT]8 ACAT [AGAT]3	[ATCT]3 ATGT [ATCT]8		GATCAATAGATTGATAGATC	AGGATTTATTATGGGAAGTGGCTCATACAATTATGGAAGCTGAGAAATTTCACAATATGCCATCT	
13	13	[AGAT]13	[ATCT]13		GATCAATAGATTGATAGATC	AGGATTTATTATGGGAAGTGGCTCATACAATTATGGAAGCTGAGAAATTTCACAATATGCCATCT	

14	1401	[AGAT]12 ACAT AGAT	ATCT ATGT [ATCT]12		GATCAATAGATTGATAGATC	AGGATTTATTATGGGAAGTGGCTCATACAATTATGGAAGCTGAGAAATTTCACAATATGCCATCT	
14	1402a	[AGAT]14	[ATCT]14		GATCAATAGATTGATAGATC	AGGATTTATTATGGGAAGTGGCTCATACAATTATGGAAGCTGAGAAATTTCACAATATGCCATCT	
14	1402b	[AGAT]14	[ATCT]14		GATCAATAGATTGATAGATC	AGGATTTATTATGGGAAGTGGCTCTTACAATTATGGAAGCTGAGAAATTTCACAATATGCCATCT	rs577490589
15	1501	[AGAT]9 ACAT [AGAT]5	[ATCT]5 ATGT [ATCT]9		GATCAATAGATTGATAGATC	AGGATTTATTATGGGAAGTGGCTCATACAATTATGGAAGCTGAGAAATTTCACAATATGCCATCT	
15	1502	[AGAT]15	[ATCT]15		GATCAATAGATTGATAGATC	AGGATTTATTATGGGAAGTGGCTCATACAATTATGGAAGCTGAGAAATTTCACAATATGCCATCT	
15	1503	[AGAT]13 ACAT AGAT	ATCT ATGT [ATCT]13		GATCAATAGATTGATAGATC	AGGATTTATTATGGGAAGTGGCTCATACAATTATGGAAGCTGAGAAATTTCACAATATGCCATCT	
16	1601	[AGAT]10 ACAT [AGAT]5	[ATCT]5 ATGT [ATCT]10		GATCAATAGATTGATAGATC	AGGATTTATTATGGGAAGTGGCTCATACAATTATGGAAGCTGAGAAATTTCACAATATGCCATCT	
16	1602	[AGAT]11 ACAT [AGAT]4	[ATCT]4 ATGT [ATCT]11		GATCAATAGATTGATAGATC	AGGATTTATTATGGGAAGTGGCTCATACAATTATGGAAGCTGAGAAATTTCACAATATGCCATCT	
16	1603	[AGAT]14 ACAT AGAT	ATCT ATGT [ATCT]14		GATCAATAGATTGATAGATC	AGGATTTATTATGGGAAGTGGCTCATACAATTATGGAAGCTGAGAAATTTCACAATATGCCATCT	
17	1701	[AGAT]4 AGAC [AGAT]6 ACAT [AGAT]5	[ATCT]5 ATGT [ATCT]6 GTCT [ATCT]4		GATCAATAGATTGATAGATC	AGGATTTATTATGGGAAGTGGCTCATACAATTATGGAAGCTGAGAAATTTCACAATATGCCATCT	
17	1702	[AGAT]11 ACAT [AGAT]5	[ATCT]5 ATGT [ATCT]11		GATCAATAGATTGATAGATC	AGGATTTATTATGGGAAGTGGCTCATACAATTATGGAAGCTGAGAAATTTCACAATATGCCATCT	
18	18a	[AGAT]12 ACAT [AGAT]5	[ATCT]5 ATGT [ATCT]12		GATCAATAGATTGATAGATC	AGGATTTATTATGGGAAGTGGCTCATACAATTATGGAAGCTGAGAAATTTCACAATATGCCATCT	
18	18b	[AGAT]12 ACAT [AGAT]5	[ATCT]5 ATGT [ATCT]12	rs529713981	GATCAATAGATTGATAGATT	AGGATTTATTATGGGAAGTGGCTCATACAATTATGGAAGCTGAGAAATTTCACAATATGCCATCT	
19	1901	[AGAT]13 ACAT [AGAT]5	[ATCT]5 ATGT [ATCT]13		GATCAATAGATTGATAGATC	AGGATTTATTATGGGAAGTGGCTCATACAATTATGGAAGCTGAGAAATTTCACAATATGCCATCT	
19	1902	[AGAT]12 ACAT [AGAT]6	[ATCT]6 ATGT [ATCT]12		GATCAATAGATTGATAGATC	AGGATTTATTATGGGAAGTGGCTCATACAATTATGGAAGCTGAGAAATTTCACAATATGCCATCT	
20	2001	[AGAT]14 ACAT [AGAT]5	[ATCT]5 ATGT [ATCT]14		GATCAATAGATTGATAGATC	AGGATTTATTATGGGAAGTGGCTCATACAATTATGGAAGCTGAGAAATTTCACAATATGCCATCT	
20	2002	[AGAT]13 ACAT [AGAT]6	[ATCT]6 ATGT [ATCT]13		GATCAATAGATTGATAGATC	AGGATTTATTATGGGAAGTGGCTCATACAATTATGGAAGCTGAGAAATTTCACAATATGCCATCT	
21	2101	[AGAT]14 ACAT [AGAT]6	[ATCT]6 ATGT [ATCT]14		GATCAATAGATTGATAGATC	AGGATTTATTATGGGAAGTGGCTCATACAATTATGGAAGCTGAGAAATTTCACAATATGCCATCT	
21	2102	[AGAT]15 ACAT [AGAT]5	[ATCT]5 ATGT [ATCT]15		GATCAATAGATTGATAGATC	AGGATTTATTATGGGAAGTGGCTCATACAATTATGGAAGCTGAGAAATTTCACAATATGCCATCT	
21.3	21.3	[AGAT]13 GAT [AGAT]2 ACAT [AGAT]5	[ATCT]5 ATGT [ATCT]2 ATC [ATCT]13		GATCAATAGATTGATAGATC	AGGATTTATTATGGGAAGTGGCTCATACAATTATGGAAGCTGAGAAATTTCACAATATGCCATCT	
22	22	[AGAT]16 ACAT [AGAT]5	[ATCT]5 ATGT [ATCT]16		GATCAATAGATTGATAGATC	AGGATTTATTATGGGAAGTGGCTCATACAATTATGGAAGCTGAGAAATTTCACAATATGCCATCT	

23	23	[AGAT]11 ACAT [AGAT]4 ACAT [AGAT]6	[ATCT]6 ATGT [ATCT]4 ATGT [ATCT]11		GATCAATAGATTGATAGATC	AGGATTTATTATGGGAAGTGGCTCATACAATTATGGAAGCTGAGAAATTTCACAATATGCCATCT	
24	24	[AGAT]11 ACAT [AGAT]4 ACAT [AGAT]7	[ATCT]7 ATGT [ATCT]4 ATGT [ATCT]11		GATCAATAGATTGATAGATC	AGGATTTATTATGGGAAGTGGCTCATACAATTATGGAAGCTGAGAAATTTCACAATATGCCATCT	

D7S820							
Allele (LB)	Allele (SB)	Bracket (UAS)	STRSeq Record Description		Left Flank	Right Flank	
6	6a	[GATA]6 GACA GATT GATA GTTT	[TATC]6		ATAGTTTAGAACGAACTAAC	GACAGATTGATAGTTTTTTTTTTATCTCACTAAATA	rs7789995
6.3	6.3	[GATA]7 GACA GATT GATA GTTT	[TATC]7		ATAGTTTAGAACGAACTAAC	GACAGATTGATAGTTTTTTTTT_ATCTCACTAAATA	rs897512434
7	7a	[GATA]7 GACA GATT GATA GTTT	[TATC]7		ATAGTTTAGAACGAACTAAC	GACAGATTGATAGTTTTTTTTTTATCTCACTAAATA	rs7789995
7	7b	[GATA]7 GACA GATT GATA GTTT	[TATC]7	rs16887642	ATAGTTTAGAATGAACTAAC	GACAGATTGATAGTTTTTTTTTTATCTCACTAAATA	rs7789995
8	8a	[GATA]8 GACA GATT GATA GTTT	[TATC]8		ATAGTTTAGAACGAACTAAC	GACAGATTGATAGTTTTTTTTTTATCTCACTAAATA	rs7789995
8	8b	[GATA]8 GACA GATT GATA GTTT	[TATC]8	rs16887642	ATAGTTTAGAATGAACTAAC	GACAGATTGATAGTTTTTTTTTTATCTCACTAAATA	rs7789995
8	8c	[GATA]8 GACA GATT GATA GTTT	[TATC]8		ATAGTTTAGAACGAACTAAC	GACAGATTGATAGTTTTTTTTTAATCTCACTAAATA	
9	9a	[GATA]9 GACA GATT GATA GTTT	[TATC]9		ATAGTTTAGAACGAACTAAC	GACAGATTGATAGTTTTTTTTTTATCTCACTAAATA	rs7789995
9	9b	[GATA]9 GACA GATT GATA GTTT	[TATC]9	rs16887642	ATAGTTTAGAATGAACTAAC	GACAGATTGATAGTTTTTTTTTTATCTCACTAAATA	rs7789995
9	9c	[GATA]9 GACA GATT GATA GTTT	[TATC]9		ATAGTTTAGAACGAACTAAC	GACAGATTGATAGTTTTTTTTTAATCTCACTAAATA	
9.2	9.2	[GATA]9 GACA GATT GA-- GTTT	[TATC]10		ATAGTTTAGAACGAACTAAC	GACAGATTGA--GTTTTTTTTTTATCTCACTAAATA	rs7789995, rs1259806300
10	10a	[GATA]10 GACA GATT GATA GTTT	[TATC]10		ATAGTTTAGAACGAACTAAC	GACAGATTGATAGTTTTTTTTTTATCTCACTAAATA	rs7789995
10	10b	[GATA]10 GACA GATT GATA GTTT	[TATC]10	rs16887642	ATAGTTTAGAATGAACTAAC	GACAGATTGATAGTTTTTTTTTTATCTCACTAAATA	rs7789995
10	10c	[GATA]10 GACA GATT GATA GTTT	[TATC]10		ATAGTTTAGAACGAACTAAC	GACAGATTGATAGTTTTTTTTTAATCTCACTAAATA	
10.1	10.101	A [GATA]10 GACA GATT GATA GTTT	[TATC]10 T		ATAGTTTAGAACGAACTAAC	GACAGATTGATAGTTTTTTTTTTATCTCACTAAATA	rs7789995
10.1	10.102	[GATA]10 GACA GATT GATA GTTT	[TATC]10		ATAGTTTAGAACGAACTAAC	GACAGATTGATAGTTTTTTTTTTTATCTCACTAAATA	rs7789995, rs1463708262
11	1101a	[GATA]11 GACA GATT GATA GTTT	[TATC]11		ATAGTTTAGAACGAACTAAC	GACAGATTGATAGTTTTTTTTTTATCTCACTAAATA	rs7789995
11	1101b	[GATA]11 GACA GATT GATA GTTT	[TATC]11	rs16887642	ATAGTTTAGAATGAACTAAC	GACAGATTGATAGTTTTTTTTTTATCTCACTAAATA	rs7789995
11	1101c	[GATA]11 GACA GATT GATA GTTT	[TATC]11		ATAGTTTAGAACGAACTAAC	GACAGATTGATAGTTTTTTTTTAATCTCACTAAATA	
11	1102a	[GATA]3 GGTA [GATA]7 GACA GATT GATA GTTT	TATC]7 TACC [TATC]3		ATAGTTTAGAACGAACTAAC	GACAGATTGATAGTTTTTTTTTTATCTCACTAAATA	rs7789995
12	1201a	[GATA]12 GACA GATT GATA GTTT	[TATC]12		ATAGTTTAGAACGAACTAAC	GACAGATTGATAGTTTTTTTTTTATCTCACTAAATA	rs7789995

12	1201b	[GATA]12 GACA GATT GATA GTTT	[TATC]12	rs16887642	ATAGTTTAGAATGAACTAAC	GACAGATTGATAGTTTTTTTTTATCTCACTAAATA	rs7789995
12	1201c	[GATA]12 GACA GATT GATA GTTT	[TATC]12		ATAGTTTAGAACGAACTAAC	GACAGATTGATAGTTTTTTTTAATCTCACTAAATA	
12	1202a	[GATA]5 GACA [GATA]6 GACA GATT GATA GTTT	[TATC]6 TGTC [TATC]5		ATAGTTTAGAACGAACTAAC	GACAGATTGATAGTTTTTTTTTATCTCACTAAATA	rs7789995
13	13a	[GATA]13 GACA GATT GATA GTTT	[TATC]13		ATAGTTTAGAACGAACTAAC	GACAGATTGATAGTTTTTTTTTATCTCACTAAATA	rs7789995
13	13c	[GATA]13 GACA GATT GATA GTTT	[TATC]13		ATAGTTTAGAACGAACTAAC	GACAGATTGATAGTTTTTTTTAATCTCACTAAATA	

D8S1179							
Allele (LB)	Allele (SB)	Bracket			Left Flank	Right Flank	
8	8	[TCTA]8			NA	TTCCC	
9	9	[TCTA]9			NA	TTCCC	
10	10	[TCTA]10			NA	TTCCC	
11	1101	[TCTA]11			NA	TTCCC	
11	1102	[TCTA]2 TCTG [TCTA]8			NA	TTCCC	
11	1103	TCTA TCTG [TCTA]9			NA	TTCCC	
12	1201	[TCTA]12			NA	TTCCC	
12	1202	[TCTA]2 TCTG [TCTA]9			NA	TTCCC	
12	1203	TCTA TCTG [TCTA]10			NA	TTCCC	
13	1301	[TCTA]13			NA	TTCCC	
13	1302	[TCTA]2 TCTG [TCTA]10			NA	TTCCC	
13	1303	TCTA TCTG [TCTA]11			NA	TTCCC	
14	1401	[TCTA]14			NA	TTCCC	
14	1402	[TCTA]2 TCTG [TCTA]11			NA	TTCCC	
14	1403	[TCTA]2 [TCTG]2 [TCTA]10			NA	TTCCC	
14	1404	TCTA TCTG [TCTA]12			NA	TTCCC	
14	1405	TCTA TCTG TGTA [TCTA]11			NA	TTCCC	
15	1501	[TCTA]15			NA	TTCCC	
15	1502	[TCTA]2 TCTG [TCTA]12			NA	TTCCC	
15	1503	[TCTA]2 [TCTG]2 [TCTA]11			NA	TTCCC	
15	1504	TCTA TCTG [TCTA]2 CCTA [TCTA]10			NA	TTCCC	
15	1505	TCTA TCTG [TCTA]13			NA	TTCCC	
15	1506	TCTA [TCTG]2 [TCTA]12			NA	TTCCC	
15	1507	TCTA [TCTG]3 [TCTA]11			NA	TTCCC	
16	1601	[TCTA]2 TCTG [TCTA]13			NA	TTCCC	
16	1602	[TCTA]2 [TCTG]2 [TCTA]12			NA	TTCCC	
16	1603	TCTA TCTG [TCTA]14			NA	TTCCC	
17	1701	[TCTA]2 TCTG [TCTA]14			NA	TTCCC	

17	1702	[TCTA]2 [TCTG]2 [TCTA]13			NA	TTCCC	
18	18	[TCTA]2 TCTG [TCTA]15			NA	TTCCC	

D9S1122							
Allele (LB)	Allele (SB)	Bracket			Left Flank	Right Flank	
7	7	TAGA TCGA [TAGA]5			AGATAACTGTAGATAGG	TATTAAT	
9	901	[TAGA]9			AGATAACTGTAGATAGG	TATTAAT	
9	902	TAGA TCGA [TAGA]7			AGATAACTGTAGATAGG	TATTAAT	
10	1001	[TAGA]10			AGATAACTGTAGATAGG	TATTAAT	
10	1002	TAGA TCGA [TAGA]8			AGATAACTGTAGATAGG	TATTAAT	
11	1101	[TAGA]11			AGATAACTGTAGATAGG	TATTAAT	
11	1102	TAGA TCGA [TAGA]9			AGATAACTGTAGATAGG	TATTAAT	
12	1201	[TAGA]12			AGATAACTGTAGATAGG	TATTAAT	
12	1202	TAGA TCGA [TAGA]10			AGATAACTGTAGATAGG	TATTAAT	
13	1301	[TAGA]13			AGATAACTGTAGATAGG	TATTAAT	
13	1302	TAGA TCGA [TAGA]11			AGATAACTGTAGATAGG	TATTAAT	
13	1303	TAGA TCGATCGA [TAGA]10			AGATAACTGTAGATAGG	TATTAAT	
14	1401	[TAGA]14			AGATAACTGTAGATAGG	TATTAAT	
14	1402	TAGA TCGA [TAGA]12			AGATAACTGTAGATAGG	TATTAAT	
15	1501	TAGA TCGA [TAGA]13			AGATAACTGTAGATAGG	TATTAAT	
15	1502	TAGA TCGATCGA [TAGA]12			AGATAACTGTAGATAGG	TATTAAT	
16	16	TAGA TCGA [TAGA]14			AGATAACTGTAGATAGG	TATTAAT	

D10S1248							
Allele (LB)	Allele (SB)	Bracket			Left Flank	Right Flank	
8	8	[GGAA]8			TTGAACAAATGAGTGAGT	ATGAAGACAATACAACCAGAGTT	
9	9	[GGAA]9			TTGAACAAATGAGTGAGT	ATGAAGACAATACAACCAGAGTT	
10	10	[GGAA]10			TTGAACAAATGAGTGAGT	ATGAAGACAATACAACCAGAGTT	
11	11	[GGAA]11			TTGAACAAATGAGTGAGT	ATGAAGACAATACAACCAGAGTT	
12	12	[GGAA]12			TTGAACAAATGAGTGAGT	ATGAAGACAATACAACCAGAGTT	
13	13	[GGAA]13			TTGAACAAATGAGTGAGT	ATGAAGACAATACAACCAGAGTT	
14	14	[GGAA]14			TTGAACAAATGAGTGAGT	ATGAAGACAATACAACCAGAGTT	
15	15	[GGAA]15			TTGAACAAATGAGTGAGT	ATGAAGACAATACAACCAGAGTT	
16	16	[GGAA]16			TTGAACAAATGAGTGAGT	ATGAAGACAATACAACCAGAGTT	
17	17	[GGAA]17			TTGAACAAATGAGTGAGT	ATGAAGACAATACAACCAGAGTT	
18	18	[GGAA]18			TTGAACAAATGAGTGAGT	ATGAAGACAATACAACCAGAGTT	

TH01							
Allele (LB)	Allele (SB)	Bracket			Left Flank	Right Flank	

5	5	[AATG]5			TGCAGGTCACAGGGAACACAGACTCCATGGTG	AGGGAAATAAGG	
6	6	[AATG]6			TGCAGGTCACAGGGAACACAGACTCCATGGTG	AGGGAAATAAGG	
7	7a	[AATG]7			TGCAGGTCACAGGGAACACAGACTCCATGGTG	AGGGAAATAAGG	
7	7b	[AATG]7			TGCAGGTCACAGGGAACACAGACTCCATGGTG	TGGGAAATAAGG	rs1564921257
8	8	[AATG]8			TGCAGGTCACAGGGAACACAGACTCCATGGTG	AGGGAAATAAGG	
9	901	[AATG]9			TGCAGGTCACAGGGAACACAGACTCCATGGTG	AGGGAAATAAGG	
9	902	AGTG [AATG]8			TGCAGGTCACAGGGAACACAGACTCCATGGTG	AGGGAAATAAGG	
9.3	9.3	[AATG]6 ATG [AATG]3			TGCAGGTCACAGGGAACACAGACTCCATGGTG	AGGGAAATAAGG	
10	10	[AATG]10			TGCAGGTCACAGGGAACACAGACTCCATGGTG	AGGGAAATAAGG	
10.3	10.3	[AATG]7 ATG [AATG]3			TGCAGGTCACAGGGAACACAGACTCCATGGTG	AGGGAAATAAGG	
11	11	[AATG]11			TGCAGGTCACAGGGAACACAGACTCCATGGTG	AGGGAAATAAGG	

vWA

Allele (LB)	Allele (SB)	Bracket (UAS)	STRSeq Record Description		Left Flank	Right Flank	
11	1101	[TCTA] [TCTG]3[TCTA]7TCCA TCTA	[TAGA]7 [CAGA]3 TAGA		ATTGA	TCCATCCATCCTATGTATT	
12	1201	[TCTA] [TCTG]4[TCTA]7TCCA TCTA	[TAGA]7 [CAGA]4 TAGA		ATTGA	TCCATCCATCCTATGTATT	
13	1301	[TCTA]2 [TCTG]4[TCTA]3[TCCA][TCTA]3 TCCA TCCA	[TAGA]3 TGGA [TAGA]3 [CAGA]4 [TAGA]2		ATTGA	TCCATCCATCCTATGTATT	
13	1302	[TCTA] [TCTG]4[TCTA]8TCCA TCTA	[TAGA]8 [CAGA]4 TAGA		ATTGA	TCCATCCATCCTATGTATT	
13	1303b	[TCTA] [TCTG]4[TCTA]8 TCTA TCTA	[TAGA]8 [CAGA]4 TAGA		ATTGA	TCCATCTATCCTATGTATT	rs771794429
14	1401	[TCTA] [TCTG] [TCTA] [TCTG]4[TCTA]3[TCCA][TCTA]3 TCCA TCCA	[TAGA]3 TGGA [TAGA]3 [CAGA]4 TAGA CAGA TAGA		ATTGA	TCCATCCATCCTATGTATT	
14	1402	[TCTA] [TCTG]3[TCTA]10 TCCA TCTA	[TAGA]10 [CAGA]3 TAGA		ATTGA	TCCATCCATCCTATGTATT	
14	1403	[TCTA] [TCTG]4[TCTA]9 TCCA TCTA	[TAGA]9 [CAGA]4 TAGA		ATTGA	TCCATCCATCCTATGTATT	
14	1404b	[TCTA] [TCTG]4[TCTA]9 TCTA TCTA	[TAGA]9 [CAGA]4 TAGA		ATTGA	TCCATCTATCCTATGTATT	rs771794429
15	1501	[TCTA] [TCTG] [TCTA] [TCTG]4[TCTA]3[TCCA] [TCTA]3[TCCA] TCCA TCCA	TGGA [TAGA]3 TGGA [TAGA]3 [CAGA]4 TAGA CAGA TAGA		ATTGA	TCCATCCATCCTATGTATT	
15	1502	[TCTA] [TCTG]3[TCTA]11 TCCA TCTA	[TAGA]11 [CAGA]3 TAGA		ATTGA	TCCATCCATCCTATGTATT	
15	1503	[TCTA] [TCTG]4[TCTA]10 TCCA TCTA	[TAGA]10 [CAGA]4 TAGA		ATTGA	TCCATCCATCCTATGTATT	
15	1504a	[TCTA] [TCTG]4[TCTA]10 TCTA TCTA	[TAGA]10 [CAGA]4 TAGA		ATTGA	TCCATCCATCCTATGTATT	
15	1504b	[TCTA] [TCTG]4[TCTA]10 TCTA TCTA	[TAGA]10 [CAGA]4 TAGA		ATTGA	TCCATCTATCCTATGTATT	rs771794429
15	1505	[TCTA] [TCTG]5[TCTA]9 TCCA TCTA	[TAGA]9 [CAGA]5 TAGA		ATTGA	TCCATCCATCCTATGTATT	
15	1506	[TCTA] [TCTG]4[TCTA]3[TCTG][TCTA]6 TCCA TCTA	[TAGA]6 CAGA [TAGA]3 [CAGA]4 TAGA		ATTGA	TCCATCCATCCTATGTATT	
16	1601	[TCTA] [TCTG]3[TCTA]11[TCCA] TCCA TCTA	TGGA [TAGA]11 [CAGA]3 TAGA		ATTGA	TCCATCCATCCTATGTATT	
16	1602	[TCTA] [TCTG]3[TCTA]12 TCCA TCTA	[TAGA]12 [CAGA]3 TAGA		ATTGA	TCCATCCATCCTATGTATT	
16	1603	[TCTA] [TCTG]4[TCTA]7[CCTA][TCTA]3 TCCA TCTA	[TAGA]3 TAGG [TAGA]7 [CAGA]4 TAGA		ATTGA	TCCATCCATCCTATGTATT	
16	1604	[TCTA] [TCTG]4[TCTA]11 TCCA TCTA	[TAGA]11 [CAGA]4 TAGA		ATTGA	TCCATCCATCCTATGTATT	
16	1605	[TCTA] [TCTG]4[TCTA]11 TCTA TCTA	[TAGA]11 [CAGA]4 TAGA		ATTGA	TCCATCCATCCTATGTATT	
17	1701	[TCTA] [TCTG]3[TCTA]13 TCCA TCTA	[TAGA]13 [CAGA]3 TAGA		ATTGA	TCCATCCATCCTATGTATT	
17	1702	[TCTA] [TCTG]4[TCTA]12 TCCA TCTA	[TAGA]12 [CAGA]4 TAGA		ATTGA	TCCATCCATCCTATGTATT	
17	1703	[TCTA] [TCTG]5[TCTA]11 TCCA TCTA	[TAGA]11 [CAGA]5 TAGA		ATTGA	TCCATCCATCCTATGTATT	

17	1704	[TCTG]4[TCTA]13 TCCA TCTA	[TAGA]13 [CAGA]4		ATTGA	TCCATCCATCCTATGTATT	
17	1705	[TCTC][TCTG]4[TCTA]13 TCCA TCTA	[TAGA]13 [CAGA]3 GAGA		ATTGA	TCCATCCATCCTATGTATT	
18	1801	[TCTA] [TCTG]3[TCTA]14 TCCA TCTA	[TAGA]14 [CAGA]3 TAGA		ATTGA	TCCATCCATCCTATGTATT	
18	1802	[TCTA] [TCTG]4[TCTA]13 GCCA TCTA	[TAGA]13 [CAGA]4 TAGA		ATTGA	TCCATCCATCCTATGTATT	
18	1803	[TCTA] [TCTG]4[TCTA]13 TCCA TCTA	[TAGA]13 [CAGA]4 TAGA		ATTGA	TCCATCCATCCTATGTATT	
18	1804	[TCTA] [TCTG]5[TCTA]12 TCCA TCTA	[TAGA]12 [CAGA]5 TAGA		ATTGA	TCCATCCATCCTATGTATT	
18	1805	[TCTA] [TCTG]6[TCTA]11 TCCA TCTA	[TAGA]11 [CAGA]6 TAGA		ATTGA	TCCATCCATCCTATGTATT	
18	1806	[TCTA] [TCTG]4[TTTA][TCTA]12 TCCA TCTA	[TAGA]12 TAAA [CAGA]4 TAGA		ATTGA	TCCATCCATCCTATGTATT	
18	1807	[TCTG]5[TCTA]13 TCCA TCTA	[TAGA]13 [CAGA]5		ATTGA	TCCATCCATCCTATGTATT	
18	1808	[TCTG]4[TCTA]14 TCCA TCTA	[TAGA]14 [CAGA]4		ATTGA	TCCATCCATCCTATGTATT	
19	1901	[TCTA] [TCTG]3[TCTA]15 TCCA TCTA	[TAGA]15 [CAGA]3 TAGA		ATTGA	TCCATCCATCCTATGTATT	
19	1902	[TCTA] [TCTG]4[TCTA]14 TCCA TCTA	[TAGA]14 [CAGA]4 TAGA		ATTGA	TCCATCCATCCTATGTATT	
19	1903	[TCTA] [TCTG]5[TCTA]13 TCCA TCTA	[TAGA]13 [CAGA]5 TAGA		ATTGA	TCCATCCATCCTATGTATT	
19	1905	[TCTA] [TCTG]6[TCTA]12 TCCA TCTA	[TAGA]12 [CAGA]6 TAGA		ATTGA	TCCATCCATCCTATGTATT	
20	2001	[TCTA] [TCTG]4[TCTA]15 TCCA TCTA	[TAGA]15 [CAGA]4 TAGA		ATTGA	TCCATCCATCCTATGTATT	
20	2002	[TCTA] [TCTG]5[TCTA]14 TCCA TCTA	[TAGA]14 [CAGA]5 TAGA		ATTGA	TCCATCCATCCTATGTATT	
20	2003	[TCTA] [TCTG]6[TCTA]13 TCCA TCTA	[TAGA]13 [CAGA]6 TAGA		ATTGA	TCCATCCATCCTATGTATT	
21	2101	[TCTA] [TCTG]4[TCTA]16 TCCA TCTA	[TAGA]16 [CAGA]4 TAGA		ATTGA	TCCATCCATCCTATGTATT	
21	2102	[TCTA] [TCTG]5[TCTA]15 TCCA TCTA	[TAGA]15 [CAGA]5 TAGA		ATTGA	TCCATCCATCCTATGTATT	
21	2103	[TCTA] [TCTG]6[TCTA]14 TCCA TCTA	[TAGA]14 [CAGA]6 TAGA		ATTGA	TCCATCCATCCTATGTATT	

					D12S391		
Allele (LB)	Allele (SB)	Bracket			Left Flank	Right Flank	
13	1301	[AGAT]7 [AGAC]5 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATG CATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGA AAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
14	1401	[AGAT]6 [AGAC]7 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATG CATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGA AAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
15	1501	[AGAT]8 [AGAC]6 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATG CATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGA AAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
15	1502	[AGAT]9 [AGAC]5 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATG CATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGA AAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
15.1	15.1	AGAT T [AGAT]7 [AGAC]6 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATG CATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGA AAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
16	1601	[AGAT]8 [AGAC]7 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATG CATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGA AAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
16	1602	[AGAT]9 [AGAC]6 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATG CATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGA AAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
16	1603	[AGAT]10 [AGAC]5 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATG CATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGA AAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	

17	1701	[AGAT]9 [AGAC]7 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
17	1702a	[AGAT]10 [AGAC]6 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
17	1702b	[AGAT]10 [AGAC]6 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCGAGGGACACTA	rs138635218
17	1703	[AGAT]11 [AGAC]5 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
17	1704	[AGAT]12 [AGAC]4 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
17.1	17.1	AGAT T [AGAT]9 [AGAC]6 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
17.3	17.3	AGAT GAT [AGAT]8 [AGAC]7 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
18	1801	[AGAT]10 [AGAC]8			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
18	1802a	[AGAT]10 [AGAC]7 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
18	1802b	[AGAT]10 [AGAC]7 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCGAGGGACACTA	rs138635218
18	1803	[AGAT]11 [AGAC]7			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
18	1804	[AGAT]11 [AGAC]6 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
18	1805	[AGAT]12 [AGAC]5 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
18	1806	[AGAT]13 [AGAC]4 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
18.3	18.3	AGAT GAT [AGAT]9 [AGAC]7 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
19	1900	AGAC [AGAT]11 [AGAC]6 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
19	1901	[AGAT]9 [AGAC]9 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
19	1902	[AGAT]10 [AGAC]8 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
19	1903	[AGAT]11 [AGAC]8			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
19	1904a	[AGAT]11 [AGAC]7 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
19	1904b	[AGAT]11 [AGAC]7 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCGAGGGACACTA	rs138635218

19	1905a	[AGAT]12 [AGAC]6 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
19	1905b	[AGAT]12 [AGAC]6 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCGAGGGACACTA	rs138635218
19	1906	[AGAT]13 [AGAC]5 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
19	1907	[AGAT]10 [AGAC]9			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
19	1908	AGGT [AGAT]11 [AGAC]6 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
19.1	19.1	AGAT T [AGAT]11 [AGAC]6 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
19.2	19.2	[AGAT]6 AT [AGAT]6 [AGAC]6 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCGAGGGACACTA	rs138635218
19.3	19.301	[AGAT]5 GAT [AGAT]7 [AGAC]6 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
19.3	19.303	AGAT GAT [AGAT]10 [AGAC]7 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
20	2000	[AGAT]10 [AGAC]9 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
20	2001	[AGAT]11 [AGAC]9			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
20	2002	[AGAT]11 [AGAC]8 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
20	2003	[AGAT]12 [AGAC]8			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
20	2004a	[AGAT]12 [AGAC]7 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
20	2004b	[AGAT]12 [AGAC]7 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCGAGGGACACTA	rs138635218
20	2005	[AGAT]13 [AGAC]7			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
20	2006a	[AGAT]13 [AGAC]6 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
20	2006b	[AGAT]13 [AGAC]6 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCGAGGGACACTA	rs138635218
20	2007	[AGAT]14 [AGAC]5 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
20.1	20.1	AGAT T [AGAT]12 [AGAC]6 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
20.3	20.3	[AGAT]3 GAT [AGAT]10 [AGAC]6 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	

21	2101	[AGAT]11 [AGAC]10			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
21	2102	[AGAT]11 [AGAC]9 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
21	2103	[AGAT]12 [AGAC]9			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
21	2104	[AGAT]12 [AGAC]8 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
21	2105	[AGAT]13 [AGAC]8			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
21	2106	[AGAT]13 [AGAC]7 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
21	2107	[AGAT]14 [AGAC]7			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
21	2108	[AGAT]14 [AGAC]6 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
21	2109	[AGAT]14 [AGAC]5 [AGAT]2			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
21	21010	AGGT [AGAT]11 [AGAC]9			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
21	21011	[AGAT]15 [AGAC]5 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
21.1	21.1	AGAT T [AGAT]13 [AGAC]6 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
22	2201	[AGAT]11 [AGAC]10 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
22	2202	[AGAT]12 [AGAC]10			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
22	2203	[AGAT]12 [AGAC]9 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
22	2204	[AGAT]13 [AGAC]9			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
22	2205a	[AGAT]13 [AGAC]8 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
22	2205b	[AGAT]13 [AGAC]8 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCGAGGGACACTA	rs138635218
22	2206	[AGAT]14 [AGAC]8			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
22	2207	[AGAT]14 [AGAC]7 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
22	2208	AGGT [AGAT]13 [AGAC]7 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	

22	2209	AGGT [AGAT]12 [AGAC]9			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
22	2210	[AGAT]15 [AGAC]6 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
23	2301	[AGAT]12 [AGAC]11			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
23	2302	[AGAT]12 [AGAC]10 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
23	2303	[AGAT]13 [AGAC]10			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
23	2304	[AGAT]13 [AGAC]9 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
23	2305	[AGAT]14 [AGAC]9			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
23	2306	[AGAT]14 [AGAC]8 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
23	2307	[AGAT]15 [AGAC]8			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
23	2308	[AGAT]15 [AGAC]7 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
24	2401	[AGAT]14 [AGAC]10			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
24	2402	[AGAT]14 [AGAC]9 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
24	2403	[AGAT]15 [AGAC]9			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
24	2404a	[AGAT]15 [AGAC]8 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
24	2404b	[AGAT]15 [AGAC]8 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCGAGGGACACTA	rs138635218
25	2500	[AGAT]14 [AGAC]10 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
25	2501	[AGAT]15 [AGAC]10			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
25	2502	[AGAT]15 [AGAC]9 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
25	2503	[AGAT]16 [AGAC]9			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
25	2504a	[AGAT]16 [AGAC]8 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
25	2504b	[AGAT]16 [AGAC]8 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCGAGGGACACTA	rs138635218

25	2505	[AGAT]14 [AGAC]11			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
26	2601	[AGAT]16 [AGAC]10			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
26	2602	[AGAT]17 [AGAC]9			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
26	2603	[AGAT]12 [AGAC]13 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	
28	28	AGGT [AGAT]18 [AGAC]8 AGAT			CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGT	GAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCTGGAGACCCAGGGACACTA	

D13S317							
Allele (LB)	Allele (SB)	Bracket			Left Flank	Right Flank	
7	701	[TATC]7 [AATC]2			TCTGACCCATCTAACGCCTATCTGTATTTACAAATACAT	ATCTATCTATCTTTCTGTCTGTCTTTTTGGG	
7	702	[TATC]7 [TATC][AATC]			TCTGACCCATCTAACGCCTATCTGTATTTACAAATACAT	ATCTATCTATCTTTCTGTCTGTCTTTTTGGG	
8	801	[TATC]8 [AATC]2			TCTGACCCATCTAACGCCTATCTGTATTTACAAATACAT	ATCTATCTATCTTTCTGTCTGTCTTTTTGGG	
9	901	[TATC]9 [AATC]2			TCTGACCCATCTAACGCCTATCTGTATTTACAAATACAT	ATCTATCTATCTTTCTGTCTGTCTTTTTGGG	
9	902	[TATC]9 [TATC][AATC]			TCTGACCCATCTAACGCCTATCTGTATTTACAAATACAT	ATCTATCTATCTTTCTGTCTGTCTTTTTGGG	
9	903a	[TATC]10 [TATC][AATC]			TCTGACCCATCTAACGCCTATCTGTATTTACAAATACAT	ATCTA----TCTTTCTGTCTGTCTTTTTGGG	rs1442523705
9	903b	[TATC]10 [TATC][AATC]			TCTGACCCATCTAACGCCTATCTGTATTTACAAATACAT	ATCTATCTATCTTT----CTGTCTTTTTGGG	rs561167308
10	1001	[TATC]10 [AATC]2			TCTGACCCATCTAACGCCTATCTGTATTTACAAATACAT	ATCTATCTATCTTTCTGTCTGTCTTTTTGGG	
10	1002a	[TATC]10 [TATC][AATC]			TCTGACCCATCTAACGCCTATCTGTATTTACAAATACAT	ATCTATCTATCTTTCTGTCTGTCTTTTTGGG	
10	1002b	[TATC]10 [TATC][AATC]		rs73250432	TCTGACCCATCTAATGCCTATCTGTATTTACAAATACAT	ATCTATCTATCTTTCTGTCTGTCTTTTTGGG	
10	1003	[TATC]10 [TATC]2			TCTGACCCATCTAACGCCTATCTGTATTTACAAATACAT	ATCTATCTATCTTTCTGTCTGTCTTTTTGGG	
11	1101	[TATC]11 [AATC]2			TCTGACCCATCTAACGCCTATCTGTATTTACAAATACAT	ATCTATCTATCTTTCTGTCTGTCTTTTTGGG	
11	1102a	[TATC]11 [TATC][AATC]			TCTGACCCATCTAACGCCTATCTGTATTTACAAATACAT	ATCTATCTATCTTTCTGTCTGTCTTTTTGGG	
11	1102b	[TATC]11 [TATC][AATC]		rs73250432	TCTGACCCATCTAATGCCTATCTGTATTTACAAATACAT	ATCTATCTATCTTTCTGTCTGTCTTTTTGGG	
11	1102c	[TATC]11 [TATC][AATC]		rs146621667	TCTGACCCATCTAACACCTATCTGTATTTACAAATACAT	ATCTATCTATCTTTCTGTCTGTCTTTTTGGG	
11	1103	[TATC]11 [TATC]2			TCTGACCCATCTAACGCCTATCTGTATTTACAAATACAT	ATCTATCTATCTTTCTGTCTGTCTTTTTGGG	
11	1104	[TATC]8[TGTC][TATC]2 [TATC][AATC]			TCTGACCCATCTAACGCCTATCTGTATTTACAAATACAT	ATCTATCTATCTTTCTGTCTGTCTTTTTGGG	
12	1200	[TATC]5[TAAC][TATC]6 [TATC][AATC]			TCTGACCCATCTAACGCCTATCTGTATTTACAAATACAT	ATCTATCTATCTTTCTGTCTGTCTTTTTGGG	
12	1201	[TATC]12 [AATC]2			TCTGACCCATCTAACGCCTATCTGTATTTACAAATACAT	ATCTATCTATCTTTCTGTCTGTCTTTTTGGG	
12	1202a	[TATC]12 [TATC][AATC]			TCTGACCCATCTAACGCCTATCTGTATTTACAAATACAT	ATCTATCTATCTTTCTGTCTGTCTTTTTGGG	
12	1202b	[TATC]12 [TATC][AATC]		rs73250432	TCTGACCCATCTAATGCCTATCTGTATTTACAAATACAT	ATCTATCTATCTTTCTGTCTGTCTTTTTGGG	
12	1202c	[TATC]12 [TATC][AATC]		rs146621667	TCTGACCCATCTAACACCTATCTGTATTTACAAATACAT	ATCTATCTATCTTTCTGTCTGTCTTTTTGGG	
12	1203	[TATC]12 [TATC]2			TCTGACCCATCTAACGCCTATCTGTATTTACAAATACAT	ATCTATCTATCTTTCTGTCTGTCTTTTTGGG	
12	1204	[TATC]7 TATT [TATC]4 [TATC][AATC]			TCTGACCCATCTAACGCCTATCTGTATTTACAAATACAT	ATCTATCTATCTTTCTGTCTGTCTTTTTGGG	
13	1301	[TATC]13 [AATC]2			TCTGACCCATCTAACGCCTATCTGTATTTACAAATACAT	ATCTATCTATCTTTCTGTCTGTCTTTTTGGG	
13	1302	[TATC]13 [TATC][AATC]			TCTGACCCATCTAACGCCTATCTGTATTTACAAATACAT	ATCTATCTATCTTTCTGTCTGTCTTTTTGGG	
14	1401	[TATC]14 [AATC]2			TCTGACCCATCTAACGCCTATCTGTATTTACAAATACAT	ATCTATCTATCTTTCTGTCTGTCTTTTTGGG	
14	1402	[TATC]14 [TATC][AATC]			TCTGACCCATCTAACGCCTATCTGTATTTACAAATACAT	ATCTATCTATCTTTCTGTCTGTCTTTTTGGG	

15	1501	[TATC]15 [AATC]2			TCTGACCCATCTAACGCCTATCTGTATTTACAAATACAT	ATCTATCTATCTTTCTGTCTGTCTTTTTGGG	
28.2	28.2	[TATC]10 [AATC]2 * [TATC]7[AATC]2			TCTGACCCATCTAACGCCTATCTGTATTTACAAATACAT	ATCTATCTATCTTTCTGTCTGTCTTTTTGGG	
		* ATCTATCTATCTTTCTGTCTGTCTTTTTGGGCTGCCTA					

Penta E

Allele (LB)	Allele (SB)	Bracket (UAS)	STRSeq Record Description		Left Flank	Right Flank	
5	5	[AAAGA]5	[TCTTT]5		NA	AAATTGTAAGGAGTTTTCT	
6	6	[AAAGA]6	[TCTTT]6		NA	AAATTGTAAGGAGTTTTCT	
7	7	[AAAGA]7	[TCTTT]7		NA	AAATTGTAAGGAGTTTTCT	
8	8	[AAAGA]8	[TCTTT]8		NA	AAATTGTAAGGAGTTTTCT	
9	9	[AAAGA]9	[TCTTT]9		NA	AAATTGTAAGGAGTTTTCT	
10	10	[AAAGA]10	[TCTTT]10		NA	AAATTGTAAGGAGTTTTCT	
11	11	[AAAGA]11	[TCTTT]11		NA	AAATTGTAAGGAGTTTTCT	
12	12	[AAAGA]12	[TCTTT]12		NA	AAATTGTAAGGAGTTTTCT	
13	1301	[AAAGA]13	[TCTTT]13		NA	AAATTGTAAGGAGTTTTCT	
13	1302	[AAAGA]12[AAATA]	TATTT [TCTTT]12		NA	AAATTGTAAGGAGTTTTCT	
14	14	[AAAGA]14	[TCTTT]14		NA	AAATTGTAAGGAGTTTTCT	
15	1501	[AAAGA]15	[TCTTT]15		NA	AAATTGTAAGGAGTTTTCT	
15	1502	[AAAGA]14[AAATA]	TATTT [TCTTT]14		NA	AAATTGTAAGGAGTTTTCT	
15.4	15.4	[AAAGA][AAGA][AAAGA]14	[TCTTT]14 CTTT TCTTT		NA	AAATTGTAAGGAGTTTTCT	
16	1601	[AAAGA]16	[TCTTT]16		NA	AAATTGTAAGGAGTTTTCT	
16	1602	[AAAGA]15[AAATA]	TATTT [TCTTT]15		NA	AAATTGTAAGGAGTTTTCT	
16.4	16.4	[AAGA][AAAGA]16	[TCTTT]16 TCTT		NA	AAATTGTAAGGAGTTTTCT	
17	1701	[AAAGA]17	[TCTTT]17		NA	AAATTGTAAGGAGTTTTCT	
17	1702	[AAAGA]16[AAATA]	TATTT [TCTTT]16		NA	AAATTGTAAGGAGTTTTCT	
18	18	[AAAGA]18	[TCTTT]18		NA	AAATTGTAAGGAGTTTTCT	
18.4	18.4	[AAAGA]6[AAAA][AAAGA]12	[TCTTT]12 TTTT [TCTTT]6		NA	AAATTGTAAGGAGTTTTCT	
19	19	[AAAGA]19	[TCTTT]19		NA	AAATTGTAAGGAGTTTTCT	
20	20	[AAAGA]20	[TCTTT]20		NA	AAATTGTAAGGAGTTTTCT	
21	21	[AAAGA]21	[TCTTT]21		NA	AAATTGTAAGGAGTTTTCT	
22	22	[AAAGA]22	[TCTTT]22		NA	AAATTGTAAGGAGTTTTCT	
23	23	[AAAGA]23	[TCTTT]23		NA	AAATTGTAAGGAGTTTTCT	

D16S539

Allele (LB)	Allele (SB)	Bracket			Left Flank	Right Flank	
8	8a	[GATA]8			TCCTCTTCCCTAGATCAATACAGACAGACAGACAGGTG	TCATTGAAAGACAAAACAGAGATGGATGATAGATAC	
8	8b	[GATA]8		rs563997442	TCCTCTTCCCTAGATCAATACAGACAGAGAGACAGGTG	TCATTGAAAGACAAAACAGAGATGGATGATAGATAC	
8	8c	[GATA]8			TCCTCTTCCCTAGATCAATACAGACAGACAGACAGGTG	TCATTGAAAGACAAACCAGAGATGGATGATAGATAC	rs11642858
9	9a	[GATA]9			TCCTCTTCCCTAGATCAATACAGACAGACAGACAGGTG	TCATTGAAAGACAAAACAGAGATGGATGATAGATAC	

9	9b	[GATA]9			TCCTCTTCCCTAGATCAATACAGACAGACAGACAGGTG	TCATTGAAAGACAAACCAGAGATGGATGATAGATAC	rs11642858
9	9c	[GATA]9		rs555136289	TCCTCTACCCTAGATCAATACAGACAGACAGACAGGTG	TCATTGAAAGACAAACCAGAGATGGATGATAGATAC	rs11642858
10	1001a	[GATA]10			TCCTCTTCCCTAGATCAATACAGACAGACAGACAGGTG	TCATTGAAAGACAAAACAGAGATGGATGATAGATAC	
10	1001b	[GATA]10			TCCTCTTCCCTAGATCAATACAGACAGACAGACAGGTG	TCATTGAAAGACAAACCAGAGATGGATGATAGATAC	rs11642858
10	1001c	[GATA]10		rs555136289	TCCTCTACCCTAGATCAATACAGACAGACAGACAGGTG	TCATTGAAAGACAAACCAGAGATGGATGATAGATAC	rs11642858
10	1002	[GATA]3 GATTGATT [GATA]5			TCCTCTTCCCTAGATCAATACAGACAGACAGACAGGTG	TCATTGAAAGACAAAACAGAGATGGATGATAGATAC	
11	1101a	[GATA]11			TCCTCTTCCCTAGATCAATACAGACAGACAGACAGGTG	TCATTGAAAGACAAAACAGAGATGGATGATAGATAC	
11	1101b	[GATA]11			TCCTCTTCCCTAGATCAATACAGACAGACAGACAGGTG	TCATTGAAAGACAAACCAGAGATGGATGATAGATAC	rs11642858
11	1101c	[GATA]11			TCCTCTTCCCTAGATCAATACAGACAGACAGACAGGTG	TCATTGAAAGACAAAACAGAGATGGATGATAGACAC	rs114697632
11	1103	[GATA]5 GACA [GATA]5			TCCTCTTCCCTAGATCAATACAGACAGACAGACAGGTG	TCATTGAAAGACAAAACAGAGATGGATGATAGATAC	
12	12a	[GATA]12			TCCTCTTCCCTAGATCAATACAGACAGACAGACAGGTG	TCATTGAAAGACAAAACAGAGATGGATGATAGATAC	
12	12b	[GATA]12			TCCTCTTCCCTAGATCAATACAGACAGACAGACAGGTG	TCATTGAAAGACAAACCAGAGATGGATGATAGATAC	rs11642858
12	12c	[GATA]12			TCCTCTTCCCTAGATCAATACAGACAGACAGACAGGTG	TCATTGAAAGACAAAACAGAGATGGATGATAGACAC	rs114697632
13	13a	[GATA]13			TCCTCTTCCCTAGATCAATACAGACAGACAGACAGGTG	TCATTGAAAGACAAAACAGAGATGGATGATAGATAC	
13	13b	[GATA]13			TCCTCTTCCCTAGATCAATACAGACAGACAGACAGGTG	TCATTGAAAGACAAACCAGAGATGGATGATAGATAC	rs11642858
14	14	[GATA]14			TCCTCTTCCCTAGATCAATACAGACAGACAGACAGGTG	TCATTGAAAGACAAAACAGAGATGGATGATAGATAC	
15	15	[GATA]15			TCCTCTTCCCTAGATCAATACAGACAGACAGACAGGTG	TCATTGAAAGACAAAACAGAGATGGATGATAGATAC	
16	16a	[GATA]16			TCCTCTTCCCTAGATCAATACAGACAGACAGACAGGTG	TCATTGAAAGACAAAACAGAGATGGATGATAGATAC	
16	16b	[GATA]16			TCCTCTTCCCTAGATCAATACAGACAGACAGACAGGTG	TCATTGAAAGACAAACCAGAGATGGATGATAGATAC	rs11642858

D17S1301							
Allele (LB)	Allele (SB)	Bracket			Left Flank	Right Flank	
7	7	[AGAT]7			ATATGTGTG	CCATCATAGGAATTTT	
8	8	[AGAT]8			ATATGTGTG	CCATCATAGGAATTTT	
9	9	[AGAT]9			ATATGTGTG	CCATCATAGGAATTTT	
10	10	[AGAT]10			ATATGTGTG	CCATCATAGGAATTTT	
11	1101	[AGAT]11			ATATGTGTG	CCATCATAGGAATTTT	
11	1102	[AGAT]10 CGAT			ATATGTGTG	CCATCATAGGAATTTT	
11.3	11.3	[AGAT]8 GAT [AGAT]3			ATATGTGTG	CCATCATAGGAATTTT	
12	1201	[AGAT]12			ATATGTGTG	CCATCATAGGAATTTT	
12	1202	[AGAT]11 CGAT			ATATGTGTG	CCATCATAGGAATTTT	
13	1301	[AGAT]13			ATATGTGTG	CCATCATAGGAATTTT	
13	1302	[AGAT]2 ACAT [AGAT]10			ATATGTGTG	CCATCATAGGAATTTT	
14	14	[AGAT]14			ATATGTGTG	CCATCATAGGAATTTT	
15	15	[AGAT]15			ATATGTGTG	CCATCATAGGAATTTT	

D18S51							
Allele (LB)	Allele (SB)	Bracket			Left Flank	Right Flank	

10	10	[AGAA]10 AAAG AGAG AG			GTCTC	AAAGAGAGAGGAAAGAAAGAGAAAAAGAAAAGAAATAGTAGCAACTGTTATTGTAAGA	
11	11	[AGAA]11 AAAG AGAG AG			GTCTC	AAAGAGAGAGGAAAGAAAGAGAAAAAGAAAAGAAATAGTAGCAACTGTTATTGTAAGA	
12	12	[AGAA]12 AAAG AGAG AG			GTCTC	AAAGAGAGAGGAAAGAAAGAGAAAAAGAAAAGAAATAGTAGCAACTGTTATTGTAAGA	
12.2	12.2	[AGAA]12 AG [AGAG]2 AG			GTCTC	AGAGAGAGAGGAAAGAAAGAGAAAAAGAAAAGAAATAGTAGCAACTGTTATTGTAAGA	rs535823682
13	1301	[AGAA]13 AAAG AGAG AG			GTCTC	AAAGAGAGAGGAAAGAAAGAGAAAAAGAAAAGAAATAGTAGCAACTGTTATTGTAAGA	
13	1302	AGAA AGCA [AGAA]11 AAAG AGAG AG			GTCTC	AAAGAGAGAGGAAAGAAAGAGAAAAAGAAAAGAAATAGTAGCAACTGTTATTGTAAGA	
13.2	13.2	[AGAA]13 AG [AGAG]2 AG			GTCTC	AGAGAGAGAGGAAAGAAAGAGAAAAAGAAAAGAAATAGTAGCAACTGTTATTGTAAGA	rs535823682
14	1401	[AGAA]14 AAAG AGAG AG			GTCTC	AAAGAGAGAGGAAAGAAAGAGAAAAAGAAAAGAAATAGTAGCAACTGTTATTGTAAGA	
14	1402	AGAA AGCA [AGAA]12 AAAG AGAG AG			GTCTC	AAAGAGAGAGGAAAGAAAGAGAAAAAGAAAAGAAATAGTAGCAACTGTTATTGTAAGA	
15	15	[AGAA]15 AAAG AGAG AG			GTCTC	AAAGAGAGAGGAAAGAAAGAGAAAAAGAAAAGAAATAGTAGCAACTGTTATTGTAAGA	
15.2	15.2	[AGAA]15 AG [AGAG]2 AG			GTCTC	AGAGAGAGAGGAAAGAAAGAGAAAAAGAAAAGAAATAGTAGCAACTGTTATTGTAAGA	rs535823682
16	16	[AGAA]16 AAAG AGAG AG			GTCTC	AAAGAGAGAGGAAAGAAAGAGAAAAAGAAAAGAAATAGTAGCAACTGTTATTGTAAGA	
16.2	16.2	[AGAA]16 AG [AGAG]2 AG			GTCTC	AGAGAGAGAGGAAAGAAAGAGAAAAAGAAAAGAAATAGTAGCAACTGTTATTGTAAGA	rs535823682
17	17a	[AGAA]17 AAAG AGAG AG			GTCTC	AAAGAGAGAGGAAAGAAAGAGAAAAAGAAAAGAAATAGTAGCAACTGTTATTGTAAGA	
17	17b	[AGAA]17 [AGAG]2 AG			GTCTC	AGAGAGAGAGGAAAGAAAGAGAAAAAGAAAAGAAATAGTAGCAACTGTTATTGTAAGA	rs535823682
17	17c	[AGAA]17 AAAG AGAG AG			GTCTC	AAAGAGAGAGGAAAGAAAGAGAAAAAGAAAAGAAATAGTAGCAGCTGTTATTGTAAGA	rs141950432
17.2	17.2	[AGAA]17 AG [AGAG]2 AG			GTCTC	AGAGAGAGAGGAAAGAAAGAGAAAAAGAAAAGAAATAGTAGCAACTGTTATTGTAAGA	rs535823682
18	18a	[AGAA]18 AAAG AGAG AG			GTCTC	AAAGAGAGAGGAAAGAAAGAGAAAAAGAAAAGAAATAGTAGCAACTGTTATTGTAAGA	
18	18b	[AGAA]18 AAAG AGAG AG			GTCTC	AAAGAGAGAGGAAAGAAAGAGAAAAAGAAAAGAAATAGTAGCAGCTGTTATTGTAAGA	rs141950432
18.1	18.1	[AGAA]14[AAAG][AGAG]AG[GAA][AGAA] AAAG AGAG AG			GTCTC	AAAGAGAGAGGAAAGAAAGAGAAAAAGAAAAGAAATAGTAGCAACTGTTATTGTAAGA	
18.2	18.2	[AGAA]18 AG [AGAG]2 AG			GTCTC	AGAGAGAGAGGAAAGAAAGAGAAAAAGAAAAGAAATAGTAGCAACTGTTATTGTAAGA	rs535823682
19	19	[AGAA]19 AAAG AGAG AG			GTCTC	AAAGAGAGAGGAAAGAAAGAGAAAAAGAAAAGAAATAGTAGCAACTGTTATTGTAAGA	
20	20	[AGAA]20 AAAG AGAG AG			GTCTC	AAAGAGAGAGGAAAGAAAGAGAAAAAGAAAAGAAATAGTAGCAACTGTTATTGTAAGA	
20.2	20.2	[AGAA]19 AG AGAA AAAG AGAG AG			GTCTC	AAAGAGAGAGGAAAGAAAGAGAAAAAGAAAAGAAATAGTAGCAACTGTTATTGTAAGA	
21	21	[AGAA]21 AAAG AGAG AG			GTCTC	AAAGAGAGAGGAAAGAAAGAGAAAAAGAAAAGAAATAGTAGCAACTGTTATTGTAAGA	
22	22	[AGAA]22 AAAG AGAG AG			GTCTC	AAAGAGAGAGGAAAGAAAGAGAAAAAGAAAAGAAATAGTAGCAACTGTTATTGTAAGA	
23	23	[AGAA]23 AAAG AGAG AG			GTCTC	AAAGAGAGAGGAAAGAAAGAGAAAAAGAAAAGAAATAGTAGCAACTGTTATTGTAAGA	
24	24	[AGAA]24 AAAG AGAG AG			GTCTC	AAAGAGAGAGGAAAGAAAGAGAAAAAGAAAAGAAATAGTAGCAACTGTTATTGTAAGA	

D19S433							
Allele (LB)	Allele (SB)	Bracket (UAS)	STRSeq Record Description		Left Flank	Right Flank	
4	4	[AAGG][AAAG][AAGG][TAGG][AAGG]2 AGAG AGGA AGAA AGAG AG	[CCTT]2 ccta CCTT cttt CCTT		AGCTATAATTGTACCACTGCACTCCAGCCTGGGCAACAGAATAAGATTCTGTTGA	AGAGAGGAAGAAAGAGAGAAGATTTTTATT	

8	8	[AAGG][AAAG][AAGG][TAGG][AAGG]6 AGAG AGGA AGAA AGAG AG	[CCTT]6 ccta CCTT cttt CCTT		AGCTATAATTGTACCACTGCACTCCAGCCTG GGCAACAGAATAAGATTCTGTTGA	AGAGAGGAAGAAAGAGAGAAGATTTTTATT	
9	9	[AAGG][AAAG][AAGG][TAGG][AAGG]7 AGAG AGGA AGAA AGAG AG	[CCTT]7 ccta CCTT cttt CCTT		AGCTATAATTGTACCACTGCACTCCAGCCTG GGCAACAGAATAAGATTCTGTTGA	AGAGAGGAAGAAAGAGAGAAGATTTTTATT	
10	10	[AAGG][AAAG][AAGG][TAGG][AAGG]8 AGAG AGGA AGAA AGAG AG	[CCTT]8 ccta CCTT cttt CCTT		AGCTATAATTGTACCACTGCACTCCAGCCTG GGCAACAGAATAAGATTCTGTTGA	AGAGAGGAAGAAAGAGAGAAGATTTTTATT	
11	11	[AAGG][AAAG][AAGG][TAGG][AAGG]9 AGAG AGGA AGAA AGAG AG	[CCTT]9 ccta CCTT cttt CCTT		AGCTATAATTGTACCACTGCACTCCAGCCTG GGCAACAGAATAAGATTCTGTTGA	AGAGAGGAAGAAAGAGAGAAGATTTTTATT	
11.2	11.2	[AAGG][AA][AAGG][TAGG][AAGG]10 AGAG AGGA AGAA AGAG AG	[CCTT]10 ccta CCTT tt CCTT		AGCTATAATTGTACCACTGCACTCCAGCCTG GGCAACAGAATAAGATTCTGTTGA	AGAGAGGAAGAAAGAGAGAAGATTTTTATT	
12	12	[AAGG][AAAG][AAGG][TAGG][AAGG]10 AGAG AGGA AGAA AGAG AG	[CCTT]10 ccta CCTT cttt CCTT		AGCTATAATTGTACCACTGCACTCCAGCCTG GGCAACAGAATAAGATTCTGTTGA	AGAGAGGAAGAAAGAGAGAAGATTTTTATT	
12.1	12.1	[AAGG][AAAG][AAGG][TAGG][AAGG]5A [AAGG]5 AGAG AGGA AGAA AGAG AG	[CCTT]5 T [CCTT]5 ccta CCTT cttt CCTT		AGCTATAATTGTACCACTGCACTCCAGCCTG GGCAACAGAATAAGATTCTGTTGA	AGAGAGGAAGAAAGAGAGAAGATTTTTATT	
12.2	12.201	[AAGG][AA][AAGG][TAGG][AAGG]11 AGAG AGGA AGAA AGAG AG	[CCTT]11 ccta CCTT tt CCTT		AGCTATAATTGTACCACTGCACTCCAGCCTG GGCAACAGAATAAGATTCTGTTGA	AGAGAGGAAGAAAGAGAGAAGATTTTTATT	
12.2	12.202	[AAGG][AA][AAGG][TAGG][AAGG]6[AGGG][AAGG]4 AGAG AGGA AGAA AGAG AG	[CCTT]4 CCCT [CCTT]6 ccta CCTT tt CCTT		AGCTATAATTGTACCACTGCACTCCAGCCTG GGCAACAGAATAAGATTCTGTTGA	AGAGAGGAAGAAAGAGAGAAGATTTTTATT	
13	1301	[AAGG][AAAG][AAGG][TAGG][AAGG]11 AGAG AGGA AGAA AGAG AG	[CCTT]11 ccta CCTT cttt CCTT		AGCTATAATTGTACCACTGCACTCCAGCCTG GGCAACAGAATAAGATTCTGTTGA	AGAGAGGAAGAAAGAGAGAAGATTTTTATT	
13	1302	[AAGG][AAAG][AAGG][TAGG][AAGG]5[AAAG][AAGG]5 AGAG AGGA AGAA AGAG AG	[CCTT]5 CTTT [CCTT]5 ccta CCTT cttt CCTT		AGCTATAATTGTACCACTGCACTCCAGCCTG GGCAACAGAATAAGATTCTGTTGA	AGAGAGGAAGAAAGAGAGAAGATTTTTATT	
13.2	13.201	[AAGG][AA][AAGG][TAGG][AAGG]12 AGAG AGGA AGAA AGAG AG	[CCTT]12 ccta CCTT tt CCTT		AGCTATAATTGTACCACTGCACTCCAGCCTG GGCAACAGAATAAGATTCTGTTGA	AGAGAGGAAGAAAGAGAGAAGATTTTTATT	
13.2	13.202	[AAGG][AAAG][AAGG][TAGG][AAGG]12 AGAG --GA AGAA AGAG AG	[CCTT]12 ccta CCTT cttt CCTT		AGCTATAATTGTACCACTGCACTCCAGCCTG GGCAACAGAATAAGATTCTGTTGA	AGAG--GAAGAAAGAGAGAAGATTTTTATT	rs745607776
14	1401a	[AAGG][AAAG][AAGG][TAGG][AAGG]12 AGAG AGGA AGAA AGAG AG	[CCTT]12 ccta CCTT cttt CCTT		AGCTATAATTGTACCACTGCACTCCAGCCTG GGCAACAGAATAAGATTCTGTTGA	AGAGAGGAAGAAAGAGAGAAGATTTTTATT	
14	1401b	[AAGG][AAAG][AAGG][TAGG][AAGG]12 AGAG AGGA AGAA AGAG AG	[CCTT]12 ccta CCTT cttt CCTT	rs533519464	AGCCATAATTGTACCACTGCACTCCAGCCTG GGCAACAGAATAAGATTCTGTTGA	AGAGAGGAAGAAAGAGAGAAGATTTTTATT	
14	1402	[AAGG][AAAG][AAGG]14 AGAG AGGA AGAA AGAG AG	[CCTT]12 cctt CCTT cttt CCTT		AGCTATAATTGTACCACTGCACTCCAGCCTG GGCAACAGAATAAGATTCTGTTGA	AGAGAGGAAGAAAGAGAGAAGATTTTTATT	
14.2	14.2	[AAGG][AA][AAGG][TAGG][AAGG]13 AGAG AGGA AGAA AGAG AG	[CCTT]13 ccta CCTT tt CCTT		AGCTATAATTGTACCACTGCACTCCAGCCTG GGCAACAGAATAAGATTCTGTTGA	AGAGAGGAAGAAAGAGAGAAGATTTTTATT	
15	15	[AAGG][AAAG][AAGG][TAGG][AAGG]13 AGAG AGGA AGAA AGAG AG	[CCTT]13 ccta CCTT cttt CCTT		AGCTATAATTGTACCACTGCACTCCAGCCTG GGCAACAGAATAAGATTCTGTTGA	AGAGAGGAAGAAAGAGAGAAGATTTTTATT	
15.2	15.2	[AAGG][AA][AAGG][TAGG][AAGG]14 AGAG AGGA AGAA AGAG AG	[CCTT]14 ccta CCTT tt CCTT		AGCTATAATTGTACCACTGCACTCCAGCCTG GGCAACAGAATAAGATTCTGTTGA	AGAGAGGAAGAAAGAGAGAAGATTTTTATT	

15.3	15.3	[AAGG][AAAG][AAGG][TAGG][AAGG]8[AAG][AAGG]5 AGAG AGGA AGAA AGAG AG	[CCTT]5 CTT [CCTT]8 ccta CCTT cttt CCTT		AGCTATAATTGTACCACTGCACTCCAGCCTGGGCAACAGAATAAGATTCTGTTGA	AGAGAGGAAGAAAGAGAGAAGATTTTTATT	
16	16	[AAGG][AAAG][AAGG][TAGG][AAGG]14 AGAG AGGA AGAA AGAG AG	[CCTT]14 ccta CCTT cttt CCTT		AGCTATAATTGTACCACTGCACTCCAGCCTGGGCAACAGAATAAGATTCTGTTGA	AGAGAGGAAGAAAGAGAGAAGATTTTTATT	
16.2	16.201	[AAGG][AA][AAGG][TAGG][AAGG]15 AGAG AGGA AGAA AGAG AG	[CCTT]15 ccta CCTT tt CCTT		AGCTATAATTGTACCACTGCACTCCAGCCTGGGCAACAGAATAAGATTCTGTTGA	AGAGAGGAAGAAAGAGAGAAGATTTTTATT	
16.2	16.202	[AAGG][AA][AAGG][TAGG]AAAG][AAGG]14 AGAG AGGA AGAA AGAG AG	[CCTT]14 CTTT ccta CCTT tt CCTT		AGCTATAATTGTACCACTGCACTCCAGCCTGGGCAACAGAATAAGATTCTGTTGA	AGAGAGGAAGAAAGAGAGAAGATTTTTATT	
17	17	[AAGG][AAAG][AAGG][TAGG][AAGG]15 AGAG AGGA AGAA AGAG AG	[CCTT]15 ccta CCTT cttt CCTT		AGCTATAATTGTACCACTGCACTCCAGCCTGGGCAACAGAATAAGATTCTGTTGA	AGAGAGGAAGAAAGAGAGAAGATTTTTATT	
17.2	17.2	[AAGG][AA][AAGG][TAGG][AAGG]16 AGAG AGGA AGAA AGAG AG	[CCTT]16 ccta CCTT tt CCTT		AGCTATAATTGTACCACTGCACTCCAGCCTGGGCAACAGAATAAGATTCTGTTGA	AGAGAGGAAGAAAGAGAGAAGATTTTTATT	
18	18	[AAGG][AAAG][AAGG][TAGG][AAGG]16 AGAG AGGA AGAA AGAG AG	[CCTT]16 ccta CCTT cttt CCTT		AGCTATAATTGTACCACTGCACTCCAGCCTGGGCAACAGAATAAGATTCTGTTGA	AGAGAGGAAGAAAGAGAGAAGATTTTTATT	

D20S482

Allele (LB)	Allele (SB)	Bracket			Left Flank	Right Flank	
9	9	[AGAT]9			AGACACCGAACCAATAAGAGAT	TTATTATAGGAATTGATT	
10	10a	[AGAT]10			AGACACCGAACCAATAAGAGAT	TTATTATAGGAATTGATT	
10	10b	[AGAT]10		rs77560248	AGACACTGAACCAATAAGAGAT	TTATTATAGGAATTGATT	
11	11	[AGAT]11			AGACACCGAACCAATAAGAGAT	TTATTATAGGAATTGATT	
12	12a	[AGAT]12			AGACACCGAACCAATAAGAGAT	TTATTATAGGAATTGATT	
12	12b	[AGAT]12		rs77560248	AGACACTGAACCAATAAGAGAT	TTATTATAGGAATTGATT	
13	1301a	[AGAT]13			AGACACCGAACCAATAAGAGAT	TTATTATAGGAATTGATT	
13	1301b	[AGAT]13		rs77560248	AGACACTGAACCAATAAGAGAT	TTATTATAGGAATTGATT	
13	1302	[AGAT]12 AGCT			AGACACCGAACCAATAAGAGAT	TTATTATAGGAATTGATT	
14	14a	[AGAT]14			AGACACCGAACCAATAAGAGAT	TTATTATAGGAATTGATT	
14	14b	[AGAT]14		rs77560248	AGACACTGAACCAATAAGAGAT	TTATTATAGGAATTGATT	
15	15a	[AGAT]15			AGACACCGAACCAATAAGAGAT	TTATTATAGGAATTGATT	
15	15b	[AGAT]15		rs77560248	AGACACTGAACCAATAAGAGAT	TTATTATAGGAATTGATT	
16	1601a	[AGAT]16			AGACACCGAACCAATAAGAGAT	TTATTATAGGAATTGATT	
16	1601b	[AGAT]16		rs77560248	AGACACTGAACCAATAAGAGAT	TTATTATAGGAATTGATT	
16	1602	ACAT [AGAT]15			AGACACCGAACCAATAAGACAT	TTATTATAGGAATTGATT	
17	17	[AGAT]17			AGACACCGAACCAATAAGAGAT	TTATTATAGGAATTGATT	
19	19	[AGAT]19			AGACACCGAACCAATAAGAGAT	TTATTATAGGAATTGATT	

D21S11

Allele (LB)	Allele (SB)	Bracket			Left Flank	Right Flank	
24.3	24.3	[TCTA]5 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]9			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CT------	
26	2601	[TCTA]4 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]8			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
26	2602	[TCTA]4 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]9			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
27	2701	[TCTA]6 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]8			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
27	2702	[TCTA]5 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]9			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
27	2703	[TCTA]4 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]9			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
27	2704	[TCTA]4 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]10			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
28	2801	[TCTA]6 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]9			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
28	2802	[TCTA]5 [TCTG]5 [TCTA]3 TA [TCTA]2 TCA [TCTA]2 TCCATA [TCTA]11			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
28	2803	[TCTA]5 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]9			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
28	2804	[TCTA]4 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]10			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
28.2	28.2	[TCTA]5 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]8 TA TCTA			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
29	2901	[TCTA]6 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]10			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
29	2902	[TCTA]6 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]9			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
29	2903	[TCTA]5 [TCTG]5 [TCTA]3 TA [TCTA]2 TCA [TCTA]2 TCCATA [TCTA]12			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
29	2904	[TCTA]5 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]11			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
29	2905	[TCTA]5 [TCTG]6 [TCTA]3 TA [TCTA]2 TCA [TCTA]2 TCCATA [TCTA]11			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
29	2906	[TCTA]5 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]10			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
29	2907	[TCTA]4 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]11			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
29	2908	[TCTA]4 [TCTG]6 [TCTA]3 TA [TCTA]4 TCA [TCTA]2 TCCATA [TCTA]10			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
29	2909	[TCTA]4 [TCTG]7 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]10			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
29.2	29.201	[TCTA]5 [TCTG]6 [TCTA]3 TA [TCTA]2 TCA [TCTA]2 TCCATA [TCTA]10 TA TCTA			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
29.2	29.202	[TCTA]5 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]9 TA TCTA			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
30	3001	[TCTA]7 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]10			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
30	3002	[TCTA]6 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]11			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
30	3003	[TCTA]6 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]10			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
30	3004	[TCTA]5 [TCTG]6 [TCTA]3 TA [TCTA]2 TCA [TCTA]2 TCCATA [TCTA]12			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
30	3005	[TCTA]5 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]11			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
30	3006	[TCTA]4 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]12			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
30	3007	[TCTA]4 [TCTG]7 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]11			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
30	3008	[TCTA]2 TATA [TCTA]3 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]11			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
30.2	30.201	[TCTA]5 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]11 TA TCTA			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
30.2	30.202	[TCTA]5 [TCTG]6 [TCTA]3 TA [TCTA]2 TCA [TCTA]2 TCCATA [TCTA]11 TA TCTA			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
30.2	30.203	[TCTA]5 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]10 TA TCTA			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
30.3	30.3	[TCTA]6 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]5 TCA [TCTA]6			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
31	3101	[TCTA]8 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]10			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
31	3102	[TCTA]7 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]11			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
31	3103	[TCTA]7 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]10			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
31	3104	[TCTA]6 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]12			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	

31	3105	[TCTA]6 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]11			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
31	3106	[TCTA]5 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]12			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
31	3107	[TCTA]4 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]13			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
31	3108	[TCTA]2 TATA [TCTA]3 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]12			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
31.2	31.201	[TCTA]5 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]12 TA TCTA			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
31.2	31.202	[TCTA]5 [TCTG]6 [TCTA]3 TA [TCTA]2 TCA [TCTA]2 TCCATA [TCTA]12 TA TCTA			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
31.2	31.203	[TCTA]5 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]11 TA TCTA			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
31.2	31.204	[TCTA]4 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]12 TA TCTA			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
32	3201	[TCTA]8 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]11			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
32	3202	[TCTA]7 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]11			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
32	3203	[TCTA]7 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]12			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
32	3204	[TCTA]6 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]13			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
32	3205	[TCTA]5 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]13			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
32	3206	[TCTA]4 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]14			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
32	3207	[TCTA]6 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]12			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
32	3208	[TCTA]10 [TCTG]4 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]10			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
32.2	32.201	[TCTA]6 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]11 TA TCTA			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
32.2	32.202	[TCTA]5 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]13 TA TCTA			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
32.2	32.203	[TCTA]5 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]12 TA TCTA			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
32.2	32.204	[TCTA]5 [TCTG]6 [TCTA]3 TA [TCTA]4 TCA [TCTA]2 TCCATA [TCTA]11 TA TCTA			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
32.2	32.205	[TCTA]5 [TCTG]7 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]11 TA TCTA			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
32.2	32.206	[TCTA]4 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]13 TA TCTA			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
33	3301	[TCTA]6 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]14			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
33.2	33.201	[TCTA]5 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]13 TA TCTA			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
33.2	33.202	[TCTA]6 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]12 TA TCTA			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
33.2	33.203	[TCTA]5 [TCTG]7 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]12 TA TCTA			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
34	3401	[TCTA]10 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]11			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
34	3402	[TCTA]9 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]12			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
34	3403	[TCTA]5 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]11 TA [TCTA]2 TA TCTA			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
34	3404	[TCTA]11 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]10			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
34	3405	[TCTA]8 [TCTG]7 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]11			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
34.2	34.201	[TCTA]5 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]14 TA TCTA			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
35	3501	[TCTA]10 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]12			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
35	3502	[TCTA]10 [TCTG]7 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]10			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
35	3503	[TCTA]5 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]8 TCA [TCTA]3 TCA [TCTA]2 TA TCTA			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
35	3504	[TCTA]12 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]10			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
35	3505	[TCTA]11 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]11			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
36	3601	[TCTA]10 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]13			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
36	3602	[TCTA]5 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]13 TA [TCTA]2 TA TCTA			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	
37	3701	[TCTA]12 [TCTG]7 [TCTA]2 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]11			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	

37	3702	[TCTA]11 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]13			AAATATGTGAGTCAATTCCCCAAGTGAATTGCCT	CTATCTAT	

Penta D							
Allele (LB)	Allele (SB)	Bracket			Left Flank	Right Flank	
2.2	2.2	[AAAGA]5		rs1190908807	AGTAGGATCACTTGAGCCTGGAAGGTCGAAGCTGAAGTGAGCCATGATCACAC CACTACACTCCAGCCTAGGTGACAGAGCAAGACACCATCTCAAG--------	AAAAACGAAGGGGAAAAAAAGAGAATCATAAACATAAATGTAA AATTT-----	
3.2	3.2	[AAAGA]6		rs1190908807	AGTAGGATCACTTGAGCCTGGAAGGTCGAAGCTGAAGTGAGCCATGATCACAC CACTACACTCCAGCCTAGGTGACAGAGCAAGACACCATCTCAAG--------	AAAAACGAAGGGGAAAAAAAGAGAATCATAAACATAAATGTAA AATTT-----	
5	5	[AAAGA]5			GATCACTTGAGCCTGGAAGGTCGAAGCTGAAGTGAGCCATGATCACACCACTAC ACTCCAGCCTAGGTGACAGAGCAAGACACCATCTCAAGAAAGAAAAAAAAG	AAAAACGAAGGGGAAAAAAAGAGAATCATAAACATAAATGTAA AATTTCTCAA	
6	6	[AAAGA]6			GATCACTTGAGCCTGGAAGGTCGAAGCTGAAGTGAGCCATGATCACACCACTAC ACTCCAGCCTAGGTGACAGAGCAAGACACCATCTCAAGAAAGAAAAAAAAG	AAAAACGAAGGGGAAAAAAAGAGAATCATAAACATAAATGTAA AATTTCTCAA	
7	7	[AAAGA]7			GATCACTTGAGCCTGGAAGGTCGAAGCTGAAGTGAGCCATGATCACACCACTAC ACTCCAGCCTAGGTGACAGAGCAAGACACCATCTCAAGAAAGAAAAAAAAG	AAAAACGAAGGGGAAAAAAAGAGAATCATAAACATAAATGTAA AATTTCTCAA	
8	8a	[AAAGA]8			GATCACTTGAGCCTGGAAGGTCGAAGCTGAAGTGAGCCATGATCACACCACTAC ACTCCAGCCTAGGTGACAGAGCAAGACACCATCTCAAGAAAGAAAAAAAAG	AAAAACGAAGGGGAAAAAAAGAGAATCATAAACATAAATGTAA AATTTCTCAA	
8	8c	[AAAGA]8			GATCACTTGAGCCTGGAAGGTCGAAGCTGAAGTGAGCCATGATCACACCACTAC ACTCCAGCCTAGGTGACAGAGCAAGACACCATCTCAAGAAAGAAAAAAAAG	AAAGACGAAGGGGAAAAAAAGAGAATCATAAACATAAATGTAA AATTTCTCAA	rs186259515
9	9a	[AAAGA]9			GATCACTTGAGCCTGGAAGGTCGAAGCTGAAGTGAGCCATGATCACACCACTAC ACTCCAGCCTAGGTGACAGAGCAAGACACCATCTCAAGAAAGAAAAAAAAG	AAAAACGAAGGGGAAAAAAAGAGAATCATAAACATAAATGTAA AATTTCTCAA	
9	9e	[AAAGA]9		ss3798736957	GACCACTTGAGCCTGGAAGGTCGAAGCTGAAGTGAGCCATGATCACACCACTAC ACTCCAGCCTAGGTGACAGAGCAAGACACCATCTCAAGAAAGAAAAAAAAG	AAAAACGAAGGGGAAAAAAAGAGAATCATAAACATAAATGTAA AATTTCTCAA	
9	9f	[AAAGA]9		rs927345580	GATCACTTGAGCCTGGAAGGTCGAAGCTGAAGTGAGCCATGATCACACCACTAC ACTCCAGCCTAGGTGACACAGCAAGACACCATCTCAAGAAAGAAAAAAAAG	AAAAACGAAGGGGAAAAAAAGAGAATCATAAACATAAATGTAA AATTTCTCAA	
10	10a	[AAAGA]10			GATCACTTGAGCCTGGAAGGTCGAAGCTGAAGTGAGCCATGATCACACCACTAC ACTCCAGCCTAGGTGACAGAGCAAGACACCATCTCAAGAAAGAAAAAAAAG	AAAAACGAAGGGGAAAAAAAGAGAATCATAAACATAAATGTAA AATTTCTCAA	
10	10c	[AAAGA]10			GATCACTTGAGCCTGGAAGGTCGAAGCTGAAGTGAGCCATGATCACACCACTAC ACTCCAGCCTAGGTGACAGAGCAAGACACCATCTCAAGAAAGAAAAAAAAG	AAAGACGAAGGGGAAAAAAAGAGAATCATAAACATAAATGTAA AATTTCTCAA	rs186259515
11	11a	[AAAGA]11			GATCACTTGAGCCTGGAAGGTCGAAGCTGAAGTGAGCCATGATCACACCACTAC ACTCCAGCCTAGGTGACAGAGCAAGACACCATCTCAAGAAAGAAAAAAAAG	AAAAACGAAGGGGAAAAAAAGAGAATCATAAACATAAATGTAA AATTTCTCAA	
11	11c	[AAAGA]11			GATCACTTGAGCCTGGAAGGTCGAAGCTGAAGTGAGCCATGATCACACCACTAC ACTCCAGCCTAGGTGACAGAGCAAGACACCATCTCAAGAAAGAAAAAAAAG	AAAGACGAAGGGGAAAAAAAGAGAATCATAAACATAAATGTAA AATTTCTCAA	rs186259515
12	1201a	[AAAGA]12			GATCACTTGAGCCTGGAAGGTCGAAGCTGAAGTGAGCCATGATCACACCACTAC ACTCCAGCCTAGGTGACAGAGCAAGACACCATCTCAAGAAAGAAAAAAAAG	AAAAACGAAGGGGAAAAAAAGAGAATCATAAACATAAATGTAA AATTTCTCAA	
12	1201b	[AAAGA]12		rs181880885	GATCACTTGAGCCTGGAAGGCCGAAGCTGAAGTGAGCCATGATCACACCACTAC ACTCCAGCCTAGGTGACAGAGCAAGACACCATCTCAAGAAAGAAAAAAAAG	AAAAACGAAGGGGAAAAAAAGAGAATCATAAACATAAATGTAA AATTTCTCAA	
12	1201c	[AAAGA]12			GATCACTTGAGCCTGGAAGGTCGAAGCTGAAGTGAGCCATGATCACACCACTAC ACTCCAGCCTAGGTGACAGAGCAAGACACCATCTCAAGAAAGAAAAAAAAG	AAAGACGAAGGGGAAAAAAAGAGAATCATAAACATAAATGTAA AATTTCTCAA	rs186259515
12	1202	[AAAGA]11 AAAGG			GATCACTTGAGCCTGGAAGGTCGAAGCTGAAGTGAGCCATGATCACACCACTAC ACTCCAGCCTAGGTGACAGAGCAAGACACCATCTCAAGAAAGAAAAAAAAG	AAAAACGAAGGGGAAAAAAAGAGAATCATAAACATAAATGTAA AATTTCTCAA	

13	13a	[AAAGA]13			GATCACTTGAGCCTGGAAGGTCGAAGCTGAAGTGAGCCATGATCACACCACTACACTCCAGCCTAGGTGACAGAGCAAGACACCATCTCAAGAAAGAAAAAAAAG	AAAAACGAAGGGGAAAAAAAGAGAATCATAAACATAAATGTAAAATTTCTCAA	
13	13b	[AAAGA]13		rs181880885	GATCACTTGAGCCTGGAAGGCCGAAGCTGAAGTGAGCCATGATCACACCACTACACTCCAGCCTAGGTGACAGAGCAAGACACCATCTCAAGAAAGAAAAAAAAG	AAAAACGAAGGGGAAAAAAAGAGAATCATAAACATAAATGTAAAATTTCTCAA	
13	13c	[AAAGA]13			GATCACTTGAGCCTGGAAGGTCGAAGCTGAAGTGAGCCATGATCACACCACTACACTCCAGCCTAGGTGACAGAGCAAGACACCATCTCAAGAAAGAAAAAAAAG	AAAGACGAAGGGGAAAAAAAGAGAATCATAAACATAAATGTAAAATTTCTCAA	rs186259515
14	14a	[AAAGA]14			GATCACTTGAGCCTGGAAGGTCGAAGCTGAAGTGAGCCATGATCACACCACTACACTCCAGCCTAGGTGACAGAGCAAGACACCATCTCAAGAAAGAAAAAAAAG	AAAAACGAAGGGGAAAAAAAGAGAATCATAAACATAAATGTAAAATTTCTCAA	
14	14b	[AAAGA]14		rs181880885	GATCACTTGAGCCTGGAAGGCCGAAGCTGAAGTGAGCCATGATCACACCACTACACTCCAGCCTAGGTGACAGAGCAAGACACCATCTCAAGAAAGAAAAAAAAG	AAAAACGAAGGGGAAAAAAAGAGAATCATAAACATAAATGTAAAATTTCTCAA	
15	15a	[AAAGA]15			GATCACTTGAGCCTGGAAGGTCGAAGCTGAAGTGAGCCATGATCACACCACTACACTCCAGCCTAGGTGACAGAGCAAGACACCATCTCAAGAAAGAAAAAAAAG	AAAAACGAAGGGGAAAAAAAGAGAATCATAAACATAAATGTAAAATTTCTCAA	
15	15c	[AAAGA]15			GATCACTTGAGCCTGGAAGGTCGAAGCTGAAGTGAGCCATGATCACACCACTACACTCCAGCCTAGGTGACAGAGCAAGACACCATCTCAAGAAAGAAAAAAAAG	AAAGACGAAGGGGAAAAAAAGAGAATCATAAACATAAATGTAAAATTTCTCAA	rs186259515
16	16	[AAAGA]16			GATCACTTGAGCCTGGAAGGTCGAAGCTGAAGTGAGCCATGATCACACCACTACACTCCAGCCTAGGTGACAGAGCAAGACACCATCTCAAGAAAGAAAAAAAAG	AAAAACGAAGGGGAAAAAAAGAGAATCATAAACATAAATGTAAAATTTCTCAA	
17	17a	[AAAGA]17			GATCACTTGAGCCTGGAAGGTCGAAGCTGAAGTGAGCCATGATCACACCACTACACTCCAGCCTAGGTGACAGAGCAAGACACCATCTCAAGAAAGAAAAAAAAG	AAAAACGAAGGGGAAAAAAAGAGAATCATAAACATAAATGTAAAATTTCTCAA	
17	17b	[AAAGA]17		rs181880885	GATCACTTGAGCCTGGAAGGCCGAAGCTGAAGTGAGCCATGATCACACCACTACACTCCAGCCTAGGTGACAGAGCAAGACACCATCTCAAGAAAGAAAAAAAAG	AAAAACGAAGGGGAAAAAAAGAGAATCATAAACATAAATGTAAAATTTCTCAA	
The 2.2 and 3.2 alleles have a bracket annotation that matches a 5 or 6 allele but have a large deletion in FR.							
The UAS left flanking sequence is shifted for the 2.2. and 3.2 alleles, giving an additional 5 bp on the left flank (assumed to be part of the primer sequence), and missing 5bp on the right flank.							

4. POPULATION DATABASES AND ASSESSMENT OF FLANKING REGION POWER

4.1. Sample selection

Following initial analysis in the Universal Analysis Software (UAS), 1018 samples from five different population groups were taken forward for concordance analysis and sequence characterisation in the previous chapter. Although all samples were initially selected on the basis that they were from unrelated individuals, an additional check for any genetic relatedness was performed prior to frequency generation as described in the materials and methods chapter of this thesis. In brief, the blind search function of the Familias software [186, 187] was used to search for specified relationships between all individuals within each population group. Pair-wise comparisons were carried out between each individual against all other individuals within the dataset to calculate a likelihood ratio (LR) for a selected relationship (e.g. parent-child, full-siblings, half-siblings) against an unrelated hypothesis. This search was carried out using STR genotypes and an LR threshold of 100 was used as cut-off for all tested relationships. Pairs obtaining LR values under this threshold were presumed to be unrelated, whereas pairs above this threshold were further investigated using available identity informative SNP genotypes. Any pairs still with an LR of 100 for a selected relationship when including SNP data were considered to be related and one sample from the pair was removed from the dataset.

A total of 29 samples were found to be related to another sample within the dataset and were therefore removed from any future analysis and frequency generation. Samples were initially obtained from individuals undergoing relationship testing who gave consent for their samples to be used for research, and although everything was done to avoid selecting related samples in the first instance, individuals tested across different cases were not immediately assumed to be related. For example, if two individuals came in requesting a sibling test, only one sample would have been taken forward for this research. If the other individual returned a year later requesting a cousin test with another person, they would not always flag as related to the initial sample. Genotypes for 989 samples were used to generate sequence-base allelic frequencies from the following five population groups: White British (n=207), British Chinese (n=193), North East African (n=198), South Asian (n=189) and West African (n=202).

4.2. Population databases

The transition from the current CE method to massively parallel sequencing of autosomal STRs leads to a more discriminating marker system, as demonstrated by the extensive increase in the number of alleles observed at certain loci discussed in the previous chapter. This is expected to result in more powerful match statistics and a greater ability to differentiate individuals, which first requires the generation of population databases. Population databases are used for the statistical evaluation of DNA profiles and contain allelic frequencies for a set of given markers. These databases are normally composed of a set number of samples extracted from unrelated individuals, from laboratory or country- relevant population groups. To make use of the variation observed through the sequencing of STRs, new databases must be generated, containing allelic frequencies for sequence-based alleles rather than length-based ones alone. Several laboratories have characterised sequence variation of autosomal STRs, and provided frequencies for the observed sequence-based variants, but there is still a lack of population diversity in the data available, with most publications focussing on European and American – relevant populations [60, 62-64, 74, 98, 212]. In terms of the UK, there is a lack of frequency data for the South Asian and North East African populations especially.

4.2.1. Sequence-based allelic frequencies

Sequence-based allelic frequencies were generated using Arlequin software [188] as described in the materials and methods chapter of this thesis, and the frequencies for the 26 loci targeted for the five population groups are provided in Table 4.1. This table refers to the length-based allelic designation, and short sequence-based designator for all alleles observed within this work and characterised in the previous chapter. Full fasta sequences, flanking region SNPs and a breakdown of the repeat region sequences are provided in Table 3.5, and in the supplementary materials of Devesse et al. [217] (Appendix II). For some markers, as previously discussed, UAS output includes a small portion of the directly adjacent flanking region (“short flank” at D1S1656, D5S818, D7S820, vWA, D13S317, D18S51, and D19S433). These regions were included in grey in the bracket annotation corresponding to the UAS output. UAS also reports the following markers on the reverse strand: D1S1656, D2S1338, FGA, D5S818, CSF1PO D6S1043, D7S820, vWA, Penta E, and D19S433. Because current guidelines recommend reporting all sequences on the forward strand, for these markers the bracket annotation is given to correlate with both UAS and STRSeq [135, 139, 184]. All loci were found not to deviate significantly from Hardy Weinberg equilibrium (HWE) following a Bonferroni correction for multiple testing, as shown in

Table ***4.2***.

Table 4.1: Allelic frequencies for all sequence-based alleles observed

For each STR locus, the length-based (LB) and sequence-based (SB) allelic designations are given, as well as the calculated frequencies of all alleles in each of the 5 populations studied.

D1S1656						
Allele (LB)	Allele (SB)	White British	British Chinese	North E. African	South Asian	West African
7	701				0.003	
8	801			0.005	0.034	
9	901				0.005	
10	1001			0.003	0.005	0.007
10	1002	0.002				
11	1101	0.087	0.067	0.013	0.169	0.031
11	1102			0.005		0.014
12	1201	0.049	0.028	0.008	0.061	0.051
12	1202	0.089	0.021	0.023	0.016	0.026
13	1301	0.024	0.049	0.061	0.081	0.131
13	1302	0.022	0.049	0.013	0.014	0.055
13	1303	0.002			0.014	
14	1401	0.002	0.016	0.051	0.016	0.037
14	1402	0.072	0.059	0.149	0.066	0.191
14	1403				0.003	
14.3	14.301			0.008		0.010
14.3	14.302	0.002				
15	1501	0.003		0.008		
15	1502	0.010	0.006	0.076	0.011	0.055
15	1503	0.110	0.331	0.134	0.167	0.123
15.3	15.301	0.029	0.003	0.035	0.013	0.012
15.3	15.302	0.043			0.003	
15.3	15.303			0.005		
16	1601	0.012		0.003	0.005	
16	1602	0.002		0.010		0.022
16	1603	0.084	0.194	0.104	0.159	0.077
16.1	16.1		0.003			
16.3	16.3	0.070	0.008	0.172	0.029	0.094
17	1701a	0.060	0.073	0.040	0.058	0.012
17	1701b				0.003	
17	1702			0.003	0.005	
17.3	17.3	0.147	0.065	0.030	0.037	0.032
18	1801					0.002
18	1802	0.005	0.005	0.005	0.008	0.002
18.3	18.3	0.058	0.023	0.030	0.013	0.012
19.3	19.3	0.010		0.008	0.003	
20.3	20.3	0.002		0.003		0.002

TPOX						
Allele (LB)	Allele (SB)	White British	British Chinese	North E. African	South Asian	West African
6	6			0.005		0.097
7	7				0.003	0.022
8	8	0.543	0.596	0.381	0.399	0.300
9	9	0.075	0.119	0.237	0.164	0.245
10	10	0.048	0.029	0.119	0.087	0.097

11	11	0.283	0.238	0.247	0.291	0.228
12	12	0.051	0.018	0.010	0.050	0.012
13	13				0.003	
14	14				0.003	

D2S441

Allele (LB)	Allele (SB)	White British	British Chinese	North E. African	South Asian	West African
7	7		0.003			
8	8					0.002
9	9a	0.002		0.005		0.002
9	9b				0.003	
9.1	9.1		0.036			
10	1001a	0.056	0.065	0.035	0.032	0.052
10	1001b	0.002	0.057	0.008	0.042	
10	1002	0.138	0.132	0.005	0.217	0.002
11	1101a	0.275	0.279	0.174	0.349	0.297
11	1101b	0.042	0.013	0.033	0.019	
11	1102	0.012	0.021	0.025	0.019	0.077
11.3	11.301	0.060	0.078	0.116	0.042	0.042
11.3	11.302					0.005
12	1201	0.014	0.168	0.088	0.048	0.161
12	1202	0.002	0.005	0.005		0.007
12	1203	0.002		0.005		0.002
12.3	12.301	0.005			0.005	0.010
12.3	12.302			0.003		
13	1301	0.002	0.021	0.005		0.005
13	1302	0.043		0.071	0.021	0.030
13	1303			0.005		
13.3	13.3					0.002
14	1401					0.002
14	1402	0.292	0.106	0.346	0.172	0.285
14	1403			0.005		0.002
15	15	0.051	0.016	0.056	0.026	0.012
16	1601			0.008	0.005	
16	1602			0.003		

D2S1338

Allele (LB)	Allele (SB)	White British	British Chinese	North E. African	South Asian	West African
14	14	0.002		0.030		
16	1601		0.008			
16	1602	0.041		0.043	0.003	0.005
16	1603	0.002				0.005
16	1604	0.002		0.003		0.042
17	1701	0.205	0.078	0.159	0.048	0.069
17	1702			0.008	0.008	0.007
17	1703					0.005
17	1704					0.002
18	1801	0.041	0.062	0.020	0.063	0.002
18	1802	0.060	0.034	0.028	0.042	0.020
18	1803					0.005
18	1804			0.003	0.045	0.002
18	1805			0.003		0.002
18	1806			0.003		
18	1807			0.003		
19	1901	0.002	0.008		0.008	
19	1902	0.092	0.194	0.086	0.095	0.072
19	1903			0.093		0.015
19	1904	0.019	0.010	0.066	0.019	0.077
19	1905				0.013	0.007
19	1906				0.003	
19	1907				0.003	
19	1908			0.003		
19	1909			0.010		
20	2001	0.007			0.003	0.002
20	2002		0.005	0.020	0.003	0.010
20	2003	0.022	0.003	0.013	0.029	0.002
20	2004	0.109	0.080	0.051	0.058	0.054
20	2005		0.010		0.003	0.002
20	2006					0.005
20	2007	0.002	0.003	0.033	0.019	0.012

20	2008			0.035		0.002
20	2009			0.003		
21	2100	0.002				
21	2101		0.003	0.008	0.011	0.012
21	2102	0.012	0.003	0.003	0.024	0.054
21	2103	0.022	0.026		0.011	0.030
21	2104		0.003			
21	2105			0.003		0.032
21	2106		0.005	0.005	0.011	
21	2107				0.003	
21	2108					0.002
21	2109			0.005		0.002
21	2110			0.003		
22	2201			0.015		0.005
22	2202					0.002
22	2203				0.003	0.010
22	2204	0.039	0.016	0.051	0.034	0.064
22	2205	0.002		0.003	0.005	0.020
22	2206	0.002	0.018	0.053	0.024	0.035
22	2207		0.003			
22	2208		0.003			
22	2209			0.003		
23	2301			0.005		0.012
23	2302	0.002	0.003			0.002
23	2303	0.087	0.153	0.051	0.183	0.087
23	2304					0.005
23	2305	0.002	0.036	0.015	0.008	0.015
23	2306		0.008			
23	2307		0.003			
23	2308			0.005		
23	2309			0.003		
24	2401					0.010
24	2402	0.010	0.003	0.003	0.008	0.005
24	2403	0.094	0.124	0.038	0.074	0.064
24	2404	0.005	0.023	0.005	0.024	0.007
24	2405		0.003			
24	2406				0.003	
24	2407					0.002
25	2501	0.005		0.003	0.019	0.002
25	2502	0.089	0.036	0.008	0.071	0.052
25	2503		0.023		0.008	0.002
25	2504				0.003	
25	2505					0.002
26	2601	0.002	0.003	0.003	0.005	0.005
26	2602	0.014	0.008		0.005	0.022
26	2603		0.003	0.003		
27	2701	0.002				
27	2702				0.003	
28	2801					0.002
D3S1358						
Allele (LB)	Allele (SB)	White British	British Chinese	North E. African	South Asian	West African
11	11	0.005		0.005		
12	1201					0.010
12	1202		0.005		0.003	
13	13	0.002		0.003		0.002
14	1401	0.002		0.035	0.003	0.069
14	1402	0.121	0.026	0.015	0.050	0.047
14	1403				0.003	
15	1501	0.027	0.005	0.124	0.063	0.186
15	1502	0.215	0.342	0.141	0.225	0.099
15	1503	0.007		0.005	0.011	0.022
16	1601	0.007	0.003	0.051	0.008	0.198
16	1602	0.186	0.215	0.109	0.185	0.106
16	1603	0.056	0.073	0.111	0.090	0.047
16	1604	0.002				0.002
16	1605				0.003	
16.2	16.2			0.003		
17	1701		0.003			
17	1702	0.002			0.003	0.015

17	1703	0.128	0.171	0.152	0.093	0.089
17	1704	0.002		0.018	0.003	
17	1705a	0.092	0.101	0.149	0.130	0.062
17	1705b				0.003	
17	1706				0.013	
18	1801					0.005
18	1802	0.002	0.021	0.013	0.005	0.010
18	1803	0.135	0.034	0.058	0.098	0.025
18	1804				0.003	
18	1805			0.003		
18.2	18.2			0.003		
19	1901	0.002		0.003		0.002
19	1902	0.005	0.003	0.003	0.008	0.002
D4S2408						
Allele (LB)	Allele (SB)	White British	British Chinese	North E. African	South Asian	West African
7	7		0.003			
8	8	0.229	0.218	0.222	0.283	0.082
9	901	0.283	0.088	0.124	0.220	0.166
9	902	0.046	0.246		0.011	
10	1001	0.271	0.319	0.164	0.209	0.235
10	1002		0.005		0.003	
11	11	0.152	0.093	0.396	0.228	0.396
12	1201	0.019	0.026	0.078	0.042	0.111
12	1202					0.002
12	1203			0.010		
13	13		0.003	0.005	0.005	0.007
FGA						
Allele (LB)	Allele (SB)	White British	British Chinese	North E. African	South Asian	West African
17	17		0.005			
18	18	0.017	0.026	0.013	0.005	0.007
18.2	18.2					0.017
19	19	0.077	0.047	0.018	0.045	0.062
19.2	19.2				0.003	0.005
20	20	0.155	0.041	0.040	0.095	0.059
20.2	20.2	0.002				0.005
21	21	0.169	0.106	0.129	0.159	0.074
21.2	21.2	0.007	0.003	0.003	0.003	
22	22	0.188	0.202	0.212	0.132	0.205
22.2	22.2	0.012	0.003		0.013	
23	23	0.167	0.218	0.199	0.204	0.144
23.2	23.2	0.005	0.005	0.003	0.005	
23.3	23.3			0.023		
24	24	0.106	0.171	0.152	0.164	0.178
24.2	24.2		0.008		0.003	
24.3	24.3			0.003		0.005
25	25	0.070	0.104	0.086	0.108	0.097
25.2	25.2		0.003		0.005	
26	2601			0.023		0.022
26	2602	0.017	0.044	0.018	0.053	0.037
26	2603	0.007				
26.2	26.2		0.005			
27	2701		0.004	0.010		0.035
27	2702		0.004	0.003	0.003	0.010
28	2801		0.003			
28	2802			0.003		0.005
28	2803			0.033		0.017
28	2804			0.003		0.002
28	2805			0.003		
29	29			0.020		0.005
30	30			0.003		0.002
31.2	31.2					0.002
32.2	32.2					0.002
42.2	42.2			0.003		
43.2	43.2			0.003		
D5S818						
Allele (LB)	Allele (SB)	White British	British Chinese	North E. African	South Asian	West African
7	701		0.029			

7	702				0.003	
8	8			0.076		0.059
9	901	0.002			0.005	0.002
9	902	0.046	0.073	0.013	0.040	0.005
10	1001	0.036	0.187	0.038	0.063	0.042
10	1002	0.017	0.005	0.005	0.034	0.015
11	1101	0.348	0.286	0.197	0.280	0.183
11	1102	0.036	0.042	0.141	0.024	0.030
11.1	11.1			0.003		
12	1201	0.249	0.184	0.220	0.275	0.280
12	1202	0.080	0.029	0.093	0.048	0.109
12	1203				0.003	
13	1301			0.040		0.017
13	1302	0.135	0.153	0.136	0.161	0.158
13	1303	0.036	0.010	0.035	0.053	0.069
14	1401	0.012	0.003	0.003	0.003	0.017
14	1402	0.002				0.002
15	1501				0.008	0.002
15	1502					0.002
16	1601					0.005

CSP1PO

Allele (LB)	Allele (SB)	White British	British Chinese	North E. African	South Asian	West African
6	6					0.002
7	7		0.003	0.013		0.067
8	8	0.002		0.053	0.003	0.067
9	9	0.019	0.036	0.081	0.011	0.040
10	10	0.263	0.246	0.288	0.183	0.280
11	1101	0.297	0.238	0.235	0.312	0.238
11	1102			0.003		
12	1201	0.326	0.389	0.298	0.386	0.245
12	1202		0.005			
12	1203	0.002				
12	1204					0.002
13	13	0.072	0.075	0.025	0.090	0.050
14	14	0.017	0.005	0.005	0.011	0.010
15	15		0.003		0.005	

D6S1043

Allele (LB)	Allele (SB)	White British	British Chinese	North E. African	South Asian	West African
9	9			0.013	0.003	0.005
10	10	0.010	0.023	0.033	0.021	0.010
11	11	0.295	0.127	0.210	0.323	0.079
12	1201a	0.312	0.130	0.247	0.217	0.215
12	1201b				0.003	
12	1202				0.003	
13	13	0.063	0.106	0.106	0.066	0.092
14	1401	0.002				
14	1402a	0.056	0.179	0.096	0.085	0.045
14	1402b	0.005				
15	1501			0.030		0.050
15	1502		0.029	0.028	0.008	0.005
15	1503				0.003	
16	1601	0.002	0.005	0.008		0.025
16	1602		0.003			0.002
16	1603	0.002				
17	1701					0.002
17	1702	0.063	0.034	0.056	0.050	0.129
18	18a	0.072	0.158	0.078	0.108	0.121
18	18b			0.003		
19	1901	0.080	0.145	0.056	0.074	0.126
19	1902					0.002
20	2001	0.034	0.047	0.030	0.032	0.072
20	2002					0.002
21	2101	0.002				
21	2102	0.002	0.010	0.003	0.005	0.005
21.3	21.3			0.003		
22	22		0.005			
23	23			0.003		0.005
24	24					0.007

D7S820						
Allele (LB)	Allele (SB)	White British	British Chinese	North E. African	South Asian	West African
6	6a					0.002
6.3	6.3	0.002				
7	7a	0.010		0.008	0.034	0.002
7	7b	0.010				
8	8a	0.086	0.019	0.083	0.114	0.111
8	8b	0.074	0.145	0.071	0.103	0.111
8	8c			0.003	0.003	
9	9a	0.140	0.010	0.066	0.101	0.067
9	9b	0.005	0.023	0.025	0.003	0.054
9	9c	0.005	0.003	0.003	0.003	
9.2	9.2		0.003			
10	10a	0.213	0.156	0.338	0.243	0.302
10	10b		0.002	0.003		0.007
10	10c	0.053	0.001	0.063	0.032	
10.1	10.101		0.000			
10.1	10.102			0.010		
11	1101a	0.184	0.301	0.164	0.175	0.201
11	1101b		0.005		0.013	
11	1101c	0.017	0.043	0.010	0.008	0.005
11	1102a		0.016			
12	1201a	0.150	0.234	0.124	0.114	0.111
12	1201b		0.003			0.002
12	1201c	0.034	0.013	0.018	0.029	0.002
12	1202a				0.008	
13	13a	0.017	0.023	0.013	0.016	0.020
13	13c	0.002			0.003	

D8S1179						
Allele (LB)	Allele (SB)	White British	British Chinese	North E. African	South Asian	West African
8	8	0.012		0.003	0.016	
9	9	0.014		0.005	0.003	0.002
10	10	0.111	0.127	0.030	0.164	0.015
11	1101	0.053	0.096	0.051	0.061	0.017
11	1102					0.012
11	1103		0.003		0.003	
12	1201	0.155	0.083	0.068	0.071	0.035
12	1202			0.033		0.062
12	1203	0.005	0.021	0.018	0.021	0.010
13	1301	0.068	0.062	0.051	0.034	0.035
13	1302	0.007		0.018	0.011	0.047
13	1303	0.271	0.122	0.078	0.119	0.129
14	1401	0.029	0.026		0.021	0.007
14	1402	0.029	0.080	0.088	0.069	0.092
14	1403					0.005
14	1404	0.109	0.093	0.164	0.143	0.252
14	1405	0.002				
15	1501	0.014			0.005	
15	1502	0.051	0.130	0.232	0.111	0.153
15	1503			0.030		0.010
15	1504	0.002				
15	1505	0.036	0.023	0.040	0.063	0.047
15	1506		0.003			0.002
15	1507			0.008		
16	1601	0.022	0.109	0.066	0.058	0.042
16	1602			0.005	0.008	0.010
16	1603	0.007		0.003	0.008	0.002
17	1701		0.021	0.008	0.005	0.007
17	1702	0.002		0.003	0.005	0.005
18	18		0.003			

D9S1122						
Allele (LB)	Allele (SB)	White British	British Chinese	North E. African	South Asian	West African
7	7			0.003		
9	901	0.002		0.005	0.008	0.007
9	902	0.002		0.005		0.030
10	1001	0.027	0.049	0.003	0.013	0.012
10	1002	0.005	0.005	0.003	0.005	0.007

11	1101	0.200	0.096	0.124	0.228	0.030
11	1102	0.034	0.070	0.141	0.085	0.149
12	1201	0.147	0.054	0.207	0.188	0.082
12	1202	0.208	0.241	0.144	0.183	0.300
13	1301	0.075	0.031	0.111	0.063	0.057
13	1302	0.249	0.365	0.215	0.188	0.272
13	1303		0.003			0.002
14	1401	0.007	0.008	0.013	0.003	0.002
14	1402	0.034	0.067	0.023	0.034	0.037
15	1501	0.010	0.005	0.005	0.003	0.010
15	1502		0.003			
16	16		0.003			0.002
D10S1248						
Allele (LB)	Allele (SB)	White British	British Chinese	North E. African	South Asian	West African
8	8		0.003			0.002
9	9	0.002		0.023		
10	10			0.005		0.002
11	11	0.005	0.003	0.018	0.011	0.057
12	12	0.019	0.078	0.066	0.024	0.151
13	13	0.280	0.350	0.253	0.135	0.248
14	14	0.314	0.215	0.285	0.296	0.240
15	15	0.188	0.223	0.247	0.296	0.198
16	16	0.145	0.111	0.086	0.190	0.082
17	17	0.046	0.016	0.015	0.045	0.015
18	18		0.003	0.003	0.003	0.005
TH01						
Allele (LB)	Allele (SB)	White British	British Chinese	North E. African	South Asian	West African
5	5					0.005
6	6	0.196	0.111	0.260	0.275	0.114
7	7a	0.172	0.285	0.343	0.159	0.451
7	7b				0.003	
8	8	0.114	0.057	0.091	0.119	0.183
9	901	0.147	0.448	0.192	0.265	0.153
9	902				0.003	
9.3	9.3	0.360	0.034	0.093	0.167	0.079
10	10	0.012	0.065	0.018	0.008	0.015
10.3	10.3				0.003	
11	11			0.003		
vWA						
Allele (LB)	Allele (SB)	White British	British Chinese	North E. African	South Asian	West African
11	1101					0.002
12	1201			0.003		
13	1301			0.003	0.003	
13	1302			0.005		0.015
13	1303b					0.012
14	1401	0.080	0.259	0.056	0.090	0.007
14	1402	0.027		0.005	0.005	0.007
14	1403	0.002	0.003	0.008	0.003	0.015
14	1404b					0.012
15	1501	0.005				
15	1502	0.068	0.005	0.033	0.040	0.057
15	1503	0.012	0.026	0.146	0.024	0.186
15	1504a	0.002				0.002
15	1504b					0.005
15	1505					0.002
15	1506					0.002
16	1601					0.002
16	1602	0.053	0.008	0.025	0.011	0.077
16	1603					0.002
16	1604	0.184	0.150	0.217	0.235	0.210
16	1605					0.002
17	1701	0.024		0.023	0.008	0.027
17	1702	0.228	0.262	0.210	0.272	0.141
17	1703			0.010	0.005	0.002
17	1704				0.005	
17	1705			0.003		
18	1801		0.008	0.005		0.010

18	1802		0.003			
18	1803	0.204	0.171	0.098	0.198	0.079
18	1804	0.005		0.005	0.008	0.010
18	1805			0.018		0.010
18	1806	0.002				
18	1807		0.005			
18	1808				0.003	
19	1901	0.002		0.003	0.003	0.002
19	1902	0.083	0.083	0.020	0.058	0.059
19	1903	0.005			0.013	0.005
19	1905			0.068		0.012
20	2001	0.010	0.013	0.005	0.013	
20	2002				0.003	0.007
20	2003			0.030		0.007
21	2101	0.002				0.002
21	2102		0.005			0.002
21	2103			0.003		

D12S391

Allele (LB)	Allele (SB)	White British	British Chinese	North E. African	South Asian	West African
13	1301					0.002
14	1401			0.005		0.002
15	1501	0.031	0.008	0.033	0.011	0.082
15	1502					0.007
15.1	15.1					0.002
16	1601		0.003			0.002
16	1602	0.034		0.013	0.005	0.035
16	1603			0.008		0.025
17	1701	0.002		0.048	0.003	0.020
17	1702a	0.114	0.106	0.152	0.130	0.082
17	1702b					0.002
17	1703	0.005	0.003	0.083	0.003	0.047
17	1704					0.002
17.1	17.1					0.002
17.3	17.3	0.017			0.011	
18	1801	0.002				
18	1802a	0.002	0.016	0.020	0.011	0.030
18	1802b			0.003		
18	1803		0.011			
18	1804	0.174	0.215	0.104	0.249	0.215
18	1805	0.007		0.063		0.035
18	1806		0.003			
18.3	18.3	0.014		0.003	0.013	
19	1900		0.003		0.003	
19	1901			0.003		0.002
19	1902		0.003	0.015		0.002
19	1903		0.010		0.011	0.002
19	1904a	0.017	0.023	0.023	0.013	0.035
19	1904b	0.007		0.003	0.008	
19	1905a	0.092	0.138	0.091	0.079	0.089
19	1905b	0.007			0.005	
19	1906		0.008	0.008		0.002
19	1907				0.003	
19	1908				0.003	
19.1	19.1					0.017
19.2	19.2				0.003	
19.3	19.301	0.005				0.002
19.3	19.303	0.007				
20	2000					0.002
20	2001	0.022	0.005	0.003	0.011	
20	2002	0.002	0.045	0.003	0.008	0.012
20	2003	0.007	0.005		0.013	0.020
20	2004a	0.019	0.040	0.023	0.021	0.027
20	2004b				0.003	
20	2005	0.002		0.003		0.012
20	2006a	0.041	0.082	0.040	0.061	0.040
20	2006b	0.002				
20	2007	0.002	0.003	0.008	0.003	0.010
20.1	20.1					0.002
20.3	20.3	0.002				

21	2101	0.012	0.003		0.003	0.002
21	2102	0.002	0.003	0.005	0.005	
21	2103	0.051	0.029	0.020	0.050	0.002
21	2104		0.049	0.013	0.019	0.010
21	2105	0.019	0.003		0.003	0.007
21	2106	0.012	0.016	0.013	0.016	0.012
21	2107					0.010
21	2108		0.021	0.003	0.013	0.007
21	2109	0.002				
21	21010	0.002				
21	21011			0.003		
21.1	21.1					0.002
22	2201	0.005				0.002
22	2202	0.014	0.010		0.008	0.002
22	2203	0.002	0.003	0.015	0.003	
22	2204	0.068	0.042	0.051	0.058	0.002
22	2205a	0.014	0.023	0.025	0.005	0.027
22	2205b			0.003		
22	2206	0.017			0.011	0.007
22	2207	0.005	0.013	0.013	0.003	
22	2208	0.002				
22	2209				0.003	
22	2210			0.003	0.003	
23	2301	0.002		0.003		
23	2302	0.002				
23	2303	0.014	0.005	0.008	0.008	
23	2304	0.005	0.005		0.005	
23	2305	0.034	0.016	0.010	0.048	0.002
23	2306	0.022	0.008	0.033	0.005	0.012
23	2307	0.005			0.003	0.005
23	2308				0.005	
24	2401	0.002	0.003		0.005	
24	2402	0.002	0.003	0.003		
24	2403	0.012	0.009	0.003	0.013	
24	2404a	0.002	0.006	0.008	0.003	0.002
24	2404b				0.003	
25	2500					0.002
25	2501		0.003		0.005	
25	2502	0.002			0.005	
25	2503	0.002		0.003	0.005	
25	2504a	0.007	0.003		0.005	0.005
25	2504b			0.003		
25	2505			0.010		
26	2601	0.005				
26	2602	0.007				
26	2603			0.003		
28	28	0.002				

D13S317						
Allele (LB)	Allele (SB)	White British	British Chinese	North E. African	South Asian	West African
7	701		0.003			
7	702	0.002			0.005	
8	801	0.119	0.339	0.169	0.185	0.010
9	901	0.065	0.124	0.025	0.095	0.002
9	902		0.021		0.003	
9	903a				0.003	
9	903b			0.010		0.002
10	1001	0.044	0.026	0.035	0.021	0.007
10	1002a	0.002	0.085	0.018	0.042	0.015
10	1002b	0.002				
10	1003		0.013	0.005		
11	1101	0.148	0.026	0.149	0.066	0.213
11	1102a	0.167	0.163	0.086	0.196	0.097
11	1102b	0.017				
11	1102c					0.005
11	1103		0.018			
11	1104		0.010		0.003	
12	1200	0.002				
12	1201	0.177	0.018	0.232	0.116	0.354
12	1202a	0.116	0.111	0.131	0.153	0.082

12	1202b	0.002				
12	1202c					0.010
12	1203		0.008			
12	1204					0.002
13	1301	0.065		0.056	0.053	0.116
13	1302	0.024	0.023	0.053	0.034	0.030
14	1401	0.034		0.023	0.013	0.050
14	1402	0.012	0.010	0.005	0.008	0.005
15	1501				0.003	
28.2	28.2			0.003		

Penta E

Allele (LB)	Allele (SB)	White British	British Chinese	North E. African	South Asian	West African
5	5	0.070	0.034	0.038	0.088	0.112
6	6					0.002
7	7	0.189	0.005	0.063	0.086	0.114
8	8	0.012		0.056	0.013	0.194
9	9	0.007	0.019	0.025	0.008	0.062
10	10	0.083	0.042	0.053	0.019	0.032
11	11	0.117	0.146	0.124	0.174	0.075
12	12	0.177	0.133	0.093	0.104	0.087
13	1301	0.107	0.053	0.047	0.053	0.124
13	1302			0.009		
14	14	0.046	0.088	0.124	0.070	0.055
15	1501	0.049	0.088	0.149	0.110	0.059
15	1502			0.010		0.004
15.4	15.4				0.003	
16	1601	0.053	0.077	0.078	0.120	0.021
16	1602					0.014
16.4	16.4				0.003	
17	1701	0.051	0.080	0.040	0.072	0.028
17	1702					0.004
18	18	0.015	0.074	0.053	0.032	0.007
18.4	18.4		0.005			
19	19	0.010	0.064	0.023	0.016	0.005
20	20	0.010	0.045	0.015	0.019	
21	21	0.005	0.029		0.011	
22	22		0.013			
23	23		0.005			

D16S539

Allele (LB)	Allele (SB)	White British	British Chinese	North E. African	South Asian	West African
8	8a	0.010	0.003	0.038	0.032	0.027
8	8b		0.007		0.040	0.002
8	8c		0.003			
9	9a	0.056	0.024	0.053	0.037	0.072
9	9b	0.056	0.216	0.091	0.146	0.158
9	9c				0.003	
10	1001a	0.012	0.003	0.030	0.032	0.079
10	1001b	0.053	0.130	0.035	0.061	0.047
10	1001c				0.003	
10	1002					0.002
11	1101a	0.285	0.242	0.283	0.249	0.240
11	1101b	0.007	0.029	0.005	0.053	0.012
11	1101c			0.003		0.010
11	1103				0.003	
12	12a	0.324	0.205	0.273	0.172	0.218
12	12b		0.011	0.003	0.021	0.002
12	12c					0.002
13	13a	0.172	0.107	0.149	0.132	0.106
13	13b			0.008	0.003	
14	14	0.022	0.013	0.025	0.016	0.020
15	15		0.005	0.005		
16	16a	0.005				
16	16b		0.003			

D17S1301

Allele (LB)	Allele (SB)	White British	British Chinese	North E. African	South Asian	West African
7	7		0.003			
8	8		0.008		0.003	

9	9	0.002	0.026		0.008	
10	10	0.031	0.057	0.028	0.008	0.007
11	1101	0.307	0.189	0.187	0.235	0.146
11	1102					0.005
11.3	11.3					0.002
12	1201	0.473	0.435	0.525	0.468	0.542
12	1202		0.005			0.002
13	1301	0.155	0.228	0.194	0.241	0.243
13	1302			0.003		
14	14	0.031	0.044	0.056	0.032	0.047
15	15		0.005	0.008	0.005	0.005

D18S51

Allele (LB)	Allele (SB)	White British	British Chinese	North E. African	South Asian	West African
10	10	0.002			0.008	
11	11	0.010		0.010	0.008	0.002
12	12	0.147	0.044	0.066	0.074	0.050
12.2	12.2			0.005		
13	1301	0.130	0.163	0.043	0.151	0.040
13	1302				0.003	
13.2	13.2					0.010
14	1401	0.155	0.194	0.093	0.296	0.047
14	1402	0.005			0.008	
15	15	0.150	0.199	0.131	0.175	0.176
15.2	15.2			0.005		
16	16	0.128	0.140	0.088	0.132	0.186
16.2	16.2			0.043		
17	17a	0.133	0.057	0.154	0.066	0.149
17	17b					0.002
17	17c					0.002
17.2	17.2			0.035		
18	18a	0.075	0.062	0.111	0.026	0.101
18	18b					0.002
18.1	18.1	0.002				
18.2	18.2			0.005		
19	19	0.027	0.039	0.104	0.026	0.136
20	20	0.012	0.021	0.068	0.019	0.052
20.2	20.2					0.002
21	21	0.014	0.013	0.030	0.003	0.015
22	22	0.005	0.039	0.005	0.005	0.020
23	23	0.002	0.021	0.003		0.007
24	24	0.002	0.008			

D19S433

Allele (LB)	Allele (SB)	White British	British Chinese	North E. African	South Asian	West African
4	4		0.003			
8	8				0.003	
9	9				0.003	
10	10					0.002
11	11			0.020		0.097
11.2	11.2				0.003	
12	12	0.065	0.036	0.134	0.056	0.136
12.1	12.1	0.002				
12.2	12.201		0.010	0.003	0.013	0.025
12.2	12.202					0.005
13	1301	0.251	0.298	0.268	0.267	0.297
13	1302				0.016	
13.2	13.201	0.012	0.047	0.048	0.013	0.042
13.2	13.202				0.005	
14	1401a	0.370	0.241	0.303	0.220	0.196
14	1401b				0.003	
14	1402				0.005	
14.2	14.2	0.043	0.111	0.045	0.074	0.062
15	15	0.169	0.101	0.088	0.116	0.030
15.2	15.2	0.031	0.122	0.033	0.108	0.050
15.3	15.3				0.003	
16	16	0.039	0.008	0.035	0.056	0.005
16.2	16.201	0.010	0.023	0.018	0.026	0.050
16.2	16.202					0.002
17	17	0.005		0.003	0.005	

17.2	17.2			0.003	0.005	0.002
18	18	0.002				
D20S482						
Allele (LB)	Allele (SB)	White British	British Chinese	North E. African	South Asian	West African
9	9	0.017		0.058	0.003	0.005
10	10a	0.002	0.022		0.003	0.002
10	10b				0.003	0.002
11	11	0.014	0.008	0.003	0.008	0.012
12	12a	0.010	0.030	0.018	0.050	0.037
12	12b	0.002	0.003	0.013	0.003	0.002
13	1301a	0.169	0.274	0.199	0.233	0.210
13	1301b	0.042	0.022	0.040	0.013	0.005
13	1302		0.003			
14	14a	0.425	0.386	0.366	0.405	0.448
14	14b	0.043	0.019	0.033	0.037	0.005
15	15a	0.196	0.177	0.146	0.159	0.201
15	15b	0.007	0.008	0.015	0.026	
16	1601a	0.058	0.041	0.091	0.048	0.067
16	1601b	0.010	0.003	0.013	0.011	0.002
16	1602		0.003			
17	17	0.005		0.005		
19	19		0.003			
D21S11						
Allele (LB)	Allele (SB)	White British	British Chinese	North E. African	South Asian	West African
24.3	24.3			0.005		0.002
26	2601	0.002			0.003	
26	2602			0.003		
27	2701	0.007	0.003		0.005	
27	2702			0.020		0.040
27	2703	0.024		0.028	0.016	0.032
27	2704			0.003		
28	2801	0.002	0.023		0.011	
28	2802		0.008			
28	2803		0.008	0.035		0.017
28	2804	0.159	0.018	0.058	0.106	0.203
28.2	28.2		0.003			
29	2901	0.072	0.177	0.013	0.056	
29	2902			0.008		0.002
29	2903		0.003			
29	2904			0.003		0.005
29	2905		0.003			0.002
29	2906	0.007	0.005	0.071	0.016	0.052
29	2907	0.140	0.091	0.149	0.138	0.111
29	2908					0.002
29	2909					0.037
29.2	29.201					0.005
29.2	29.202		0.003		0.003	
30	3001	0.005	0.023		0.003	
30	3002	0.135	0.130	0.013	0.048	0.002
30	3003			0.008		0.007
30	3004		0.003			
30	3005	0.031	0.065	0.162	0.069	0.121
30	3006	0.077	0.047	0.109	0.058	0.035
30	3007	0.002		0.003		0.002
30	3008				0.011	
30.2	30.201	0.022			0.003	0.005
30.2	30.202			0.003		0.010
30.2	30.203	0.005	0.008		0.034	0.010
30.3	30.3		0.003			
31	3101	0.002	0.005			
31	3102	0.005	0.021			
31	3103			0.005		0.020
31	3104	0.022	0.010		0.011	0.002
31	3105			0.005		0.017
31	3106	0.036	0.052	0.018	0.013	0.035
31	3107	0.007		0.018	0.013	0.002
31	3108			0.003		
31.2	31.201		0.005	0.013		

31.2	31.202					0.022
31.2	31.203	0.097	0.049	0.073	0.127	0.032
31.2	31.204					0.002
32	3201		0.010			
32	3202					0.002
32	3203		0.010			
32	3204	0.010	0.003			
32	3205		0.026	0.005		0.010
32	3206	0.002				
32	3207					0.005
32	3208					0.002
32.2	32.201	0.002			0.003	
32.2	32.202		0.005		0.003	
32.2	32.203	0.087	0.127	0.078	0.161	0.064
32.2	32.204		0.005			
32.2	32.205		0.003		0.003	
32.2	32.206					0.005
33	3301	0.002				
33.2	33.201	0.027	0.034	0.030	0.077	0.022
33.2	33.202				0.003	
33.2	33.203		0.003			0.002
34	3401			0.003		0.005
34	3402		0.003			
34	3403	0.002			0.003	
34	3404			0.005		
34	3405			0.003		
34.2	34.201	0.005	0.005	0.003	0.008	0.007
35	3501			0.013		0.012
35	3502					0.002
35	3503					0.020
35	3504			0.010		
35	3505			0.005		
36	3601			0.010		
36	3602			0.008		
37	3701			0.003		
37	3702			0.003		
			Penta D			
Allele (LB)	Allele (SB)	White British	British Chinese	North E. African	South Asian	West African
2.2	2.2			0.082	0.003	0.173
3.2	3.2			0.022	0.003	0.012
5	5					0.067
6	6			0.014	0.003	0.002
7	7	0.010	0.005	0.044	0.008	0.027
8	8a	0.012	0.057	0.093	0.016	0.176
8	8c			0.011		
9	9a	0.193	0.349	0.161	0.233	0.106
9	9e		0.006			
9	9f			0.003		
10	10a	0.145	0.136	0.199	0.161	0.094
10	10c		0.004			
11	11a	0.145	0.127	0.134	0.251	0.176
11	11c			0.003		0.003
12	1201a	0.210	0.148	0.095	0.116	0.086
12	1201b					0.005
12	1201c	0.010				
12	1202				0.003	
13	13a	0.176	0.109	0.057	0.116	0.062
13	13b					0.005
13	13c			0.005		
14	14a	0.065	0.052	0.065	0.050	
14	14b					0.002
15	15a	0.022	0.004	0.003	0.024	
15	15c		0.004			
16	16	0.007			0.011	
17	17a	0.005			0.003	
17	17b					0.002

Table 4.2: Hardy-Weinberg Equilibrium

130 tests, 0.05 significance level with Bonferroni correction means a p-value of 0.00038 needs to be achieved to reach a significant departure from HWE.

	White British						British Chinese					
Locus	#Genot	Obs.Het.	Exp.Het.	P-value	s.d.	Steps done	#Genot	Obs.Het.	Exp.Het.	P-value	s.d.	Steps done
D1S1656	185	0.89730	0.92173	0.00192	0.00003	1001000	150	0.76667	0.80892	0.03445	0.00014	1001000
D2S1338	207	0.90338	0.90493	0.32899	0.00026	1001000	193	0.93264	0.90271	0.46406	0.00016	1001000
D2S441	206	0.78641	0.80925	0.49837	0.00029	1001000	190	0.85789	0.85270	0.61927	0.00050	1001000
D3S1358	207	0.85990	0.85959	0.86756	0.00019	1001000	193	0.78756	0.79186	0.93524	0.00022	1001000
D4S2408	207	0.78261	0.77051	0.41461	0.00045	1001000	193	0.73057	0.77537	0.32702	0.00032	1001000
D5S818	207	0.74396	0.78788	0.50443	0.00032	1001000	192	0.85417	0.81978	0.50466	0.00040	1001000
D6S1043	206	0.77670	0.79512	0.09172	0.00018	1001000	193	0.84456	0.87528	0.13941	0.00028	1001000
D7S820	206	0.83495	0.86331	0.41476	0.00035	1001000	187	0.83422	0.81430	0.99146	0.00006	1001000
D8S1178	207	0.88889	0.86670	0.41579	0.00027	1001000	193	0.90155	0.90562	0.86377	0.00024	1001000
D9S1122	207	0.84058	0.82626	0.74614	0.00028	1001000	193	0.79275	0.78545	0.67611	0.00035	1001000
D10S1248	207	0.80193	0.76573	0.41163	0.00032	1001000	193	0.78238	0.76508	0.20261	0.00028	1001000
D12S391	206	0.93204	0.93409	0.59214	0.00007	1001000	190	0.90526	0.90739	0.34914	0.00018	1001000
D13S317	206	0.87864	0.88066	0.95905	0.00013	1001000	193	0.83938	0.82171	0.43283	0.00039	1001000
D16S539	207	0.76812	0.77668	0.99884	0.00004	1001000	189	0.82011	0.82404	0.26971	0.00023	1001000
D17S1301	207	0.70048	0.65747	0.66661	0.00059	1001000	193	0.72539	0.71871	0.92314	0.00022	1001000
D18S51	207	0.85990	0.87620	0.29916	0.00031	1001000	193	0.89119	0.86533	0.03731	0.00023	1001000
D19S433	207	0.74879	0.76468	0.47463	0.00031	1001000	193	0.81865	0.81365	0.75862	0.00040	1001000
D20S482	206	0.75728	0.74579	0.27765	0.00034	1001000	184	0.71739	0.74257	0.14774	0.00025	1001000
D21S11	207	0.92754	0.90580	0.87141	0.00014	1001000	192	0.86979	0.91391	0.48341	0.00015	1001000
CSF1PO	207	0.71981	0.73192	0.81370	0.00033	1001000	193	0.78756	0.72647	0.20254	0.00040	1001000
Penta D	195	0.80513	0.84108	0.74039	0.00029	1001000	146	0.85616	0.79450	0.02850	0.00013	1001000
Penta E	162	0.81481	0.88275	0.48212	0.00040	1001000	57	0.84211	0.92004	0.03911	0.00015	1001000
FGA	207	0.87440	0.86330	0.78552	0.00034	1001000	192	0.85938	0.85478	0.71521	0.00030	1001000
TH01	207	0.76812	0.76989	0.60342	0.00050	1001000	193	0.72539	0.69874	0.16050	0.00035	1001000
TPOX	207	0.55556	0.61574	0.30624	0.00039	1001000	193	0.57513	0.57430	0.86126	0.00030	1001000
vWA	207	0.87923	0.85229	0.60034	0.00026	1001000	191	0.77487	0.80624	0.28961	0.00026	1001000

	North East African						South Asian					
Locus	#Genot	Obs.Het.	Exp.Het.	P-value	s.d.	Steps done	#Genot	Obs.Het.	Exp.Het.	P-value	s.d.	Steps done
D1S1656	198	0.89394	0.90399	0.25507	0.00017	1001000	188	0.87766	0.89713	0.22359	0.00017	1001000
D2S1338	198	0.95455	0.93694	0.92207	0.00007	1001000	189	0.93122	0.93023	0.94110	0.00008	1001000
D2S441	198	0.81818	0.81936	0.52038	0.00027	1001000	189	0.77249	0.79477	0.62244	0.00036	1001000
D3S1358	198	0.90909	0.88968	0.11548	0.00014	1001000	189	0.84656	0.86733	0.73785	0.00020	1001000
D4S2408	198	0.70707	0.74681	0.00059	0.00002	1001000	189	0.74603	0.77633	0.93189	0.00021	1001000
D5S818	198	0.82323	0.85752	0.42571	0.00051	1001000	189	0.77778	0.80923	0.19967	0.00024	1001000
D6S1043	198	0.79798	0.86031	0.07860	0.00017	1001000	189	0.76720	0.81804	0.50047	0.00025	1001000
D7S820	198	0.83838	0.82369	0.81545	0.00025	1001000	189	0.81481	0.86231	0.48848	0.00029	1001000
D8S1178	198	0.88889	0.88797	0.88567	0.00021	1001000	189	0.87831	0.90494	0.10708	0.00019	1001000
D9S1122	198	0.82828	0.84404	0.60105	0.00023	1001000	189	0.80952	0.83390	0.56457	0.00039	1001000
D10S1248	198	0.75253	0.78276	0.64038	0.00039	1001000	189	0.75132	0.76926	0.37326	0.00043	1001000
D12S391	198	0.93434	0.93662	0.94307	0.00011	1001000	189	0.88360	0.90226	0.50388	0.00008	1001000
D13S317	198	0.85354	0.86399	0.91606	0.00022	1001000	189	0.80423	0.87264	0.51162	0.00036	1001000
D16S539	198	0.80303	0.81004	0.48581	0.00035	1001000	189	0.80423	0.86001	0.50774	0.00029	1001000
D17S1301	198	0.64141	0.64910	0.92769	0.00021	1001000	189	0.65079	0.66794	0.65880	0.00040	1001000
D18S51	198	0.90404	0.90669	0.12874	0.00022	1001000	189	0.85185	0.83185	0.36179	0.00030	1001000
D19S433	198	0.78788	0.80540	0.54021	0.00035	1001000	189	0.84656	0.84420	0.80077	0.00030	1001000
D20S482	198	0.77778	0.79143	0.22351	0.00031	1001000	189	0.75132	0.75152	0.36329	0.00025	1001000
D21S11	198	0.91919	0.91713	0.35626	0.00019	1001000	189	0.87831	0.90823	0.63620	0.00021	1001000
CSF1PO	198	0.76263	0.76494	0.98259	0.00012	1001000	189	0.74603	0.71358	0.19357	0.00031	1001000
Penta D	182	0.89011	0.88398	0.69396	0.00030	1001000	181	0.82873	0.83003	0.96187	0.00013	1001000
Penta E	122	0.91803	0.92009	0.29483	0.00024	1001000	160	0.88750	0.90374	0.50639	0.00031	1001000
FGA	198	0.86869	0.86538	0.27295	0.00023	1001000	189	0.86772	0.86525	0.52251	0.00031	1001000
TH01	198	0.75253	0.76218	0.78085	0.00033	1001000	189	0.70370	0.78917	0.17067	0.00026	1001000
TPOX	198	0.64141	0.72463	0.04853	0.00019	1001000	189	0.68783	0.72057	0.16707	0.00032	1001000
vWa	198	0.83333	0.86802	0.75840	0.00020	1001000	189	0.82540	0.81879	0.22230	0.00019	1001000

	West African					
Locus	#Genot	Obs.Het.	Exp.Het.	P-value	s.d.	Steps done
D1S1656	196	0.86224	0.90409	0.22623	0.00022	1001000
D2S1338	202	0.94059	0.95424	0.13012	0.00007	1001000
D2S441	202	0.77228	0.79519	0.29364	0.00033	1001000
D3S1358	202	0.84158	0.88481	0.12056	0.00026	1001000
D4S2408	202	0.76238	0.74305	0.91049	0.00023	1001000
D5S818	202	0.84158	0.84146	0.71533	0.00025	1001000
D6S1043	202	0.83663	0.88346	0.08239	0.00017	1001000
D7S820	202	0.82178	0.82550	0.91560	0.00025	1001000
D8S1178	202	0.89109	0.87635	0.57596	0.00024	1001000
D9S1122	202	0.80198	0.80266	0.65374	0.00031	1001000
D10S1248	202	0.75248	0.81091	0.08093	0.00021	1001000
D12S391	202	0.90594	0.92186	0.04674	0.00003	1001000
D13S317	202	0.81188	0.79798	0.92220	0.00017	1001000
D16S539	202	0.81188	0.84553	0.67680	0.00045	1001000
D17S1301	202	0.65347	0.62520	0.13993	0.00027	1001000
D18S51	202	0.91584	0.87620	0.33082	0.00025	1001000
D19S433	202	0.88119	0.83569	0.15076	0.00026	1001000
D20S482	202	0.72772	0.71048	0.99847	0.00003	1001000
D21S11	202	0.92079	0.91653	0.74381	0.00012	1001000
CSF1PO	202	0.75248	0.79416	0.30650	0.00030	1001000
Penta D	194	0.90206	0.87433	0.76994	0.00026	1001000
Penta E	152	0.88816	0.89406	0.42383	0.00036	1001000
FGA	202	0.87129	0.88143	0.45609	0.00034	1001000
TH01	202	0.74752	0.72226	0.87370	0.00032	1001000
TPOX	202	0.77723	0.78104	0.96279	0.00018	1001000
vWA	202	0.89604	0.88234	0.55597	0.00015	1001000

Three tri-allelic genotypes were omitted from the frequency data: two at TPOX in the

West African population and one at D4S2408 in the North East African population (discussed in the previous chapter). The genotype for the sample with a discordance at D7S820 caused by a flanking region SNP was given to be concordant with CE (6.3 rather than 7), whereas those where the discordances at D5S818 and Penta D led to drop out were omitted. For the 5 samples with missing 24.3 alleles at D21S11, given that the "true" genotypes were confirmed with CE and re-sequencing with custom primers, these were included in the frequencies so as not to skew the data by missing out this allele altogether.

4.2.1.1. Allele frequency distribution

Some of the frequency data from Table 4.1 has been condensed in the following figures to better discuss certain findings. Figure 4.1 shows a graphical representation of the allele frequency distribution across all samples. Frequencies for each allele were averaged across the five population groups studied, to get an initial idea of the increase in diversity gained when using sequence-based allelic frequencies compared to length-based ones. Globally, the six markers showing the most substantial gains in number of observable alleles compared to CE in the previous chapter were D13S317 (+20 alleles), D3S1358 (+20 alleles), vWA (+33 alleles), D21S11 (+59 alleles), D2S1338 (+67 alleles) and D12S391 (+72 alleles). This is reflected in Figure 4.1 where it is obvious for markers such as D3S1358 and D2S1338 especially that the more common length-based alleles are now "split" because of variation seen at the sequence level in these alleles. Similar results are seen as D13S317 and D21S11. At vWA, the correlation is less obvious, suggesting that many of the sequence-based variants at this locus are quite uncommon. When averaging the frequencies for loci across the five population groups, a frequency of between 0.002 and 0.003 would indicate that the allele has only been observed once or twice throughout the entire dataset and is therefore considered rare. At vWA, of the 33 additional alleles observed when differentiating based on sequence rather than length, 14 are rare by this definition. At D12S391, of the 72 additional sequence-based alleles observed, 27 are rare. This can be seen in Figure 4.2, where many of the alleles characterised by repeat region or flanking region sequence are depicted with very thin bands in the Sankey diagram, representing low frequencies. Most flanking region variants also appear limited to a few length-based alleles for this marker.

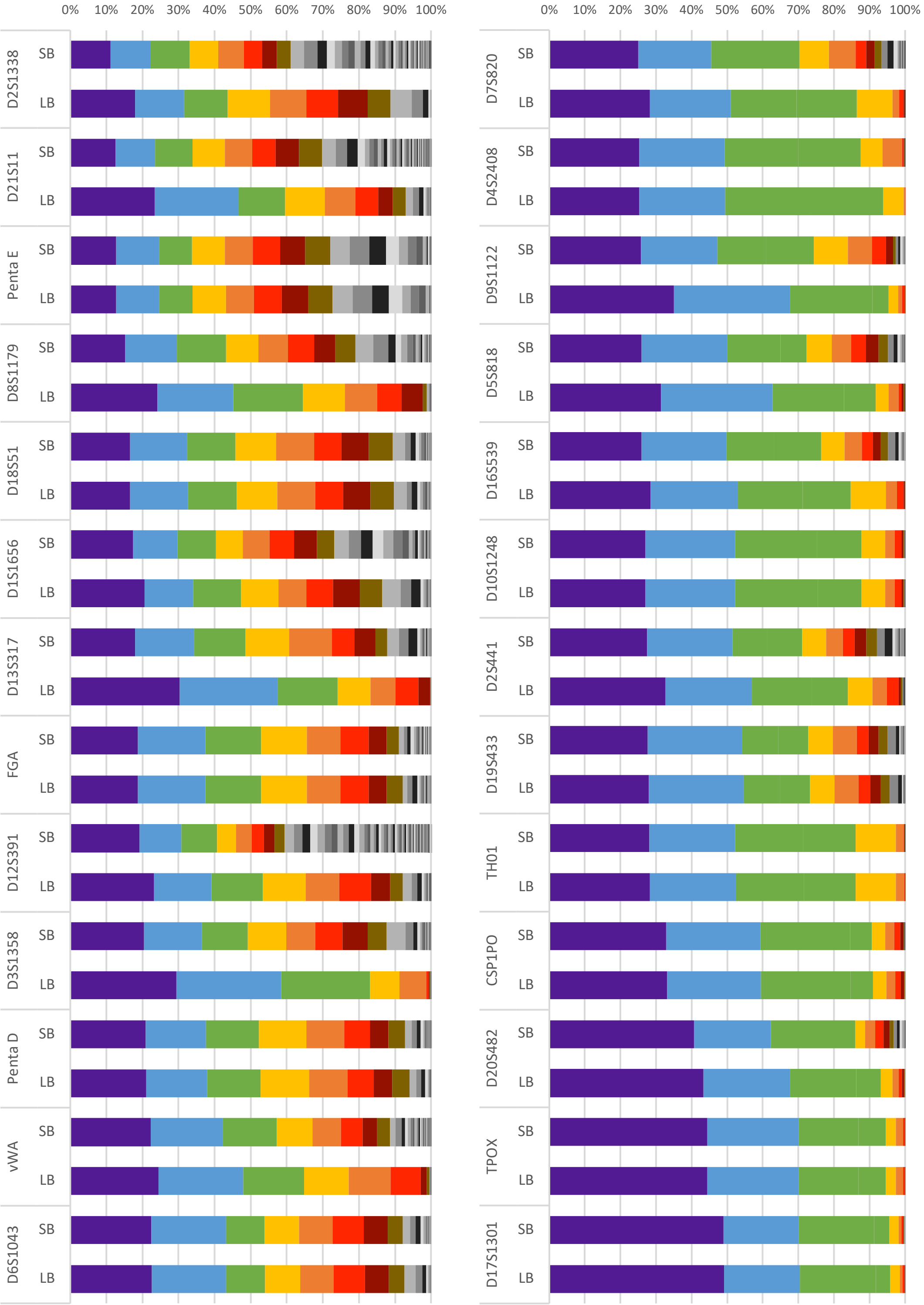


Figure 4.1: Allele frequency distribution per locus

Distribution of allelic frequencies across all populations (n=989) per locus, sequence-based (SB) and length-based (LB). The 8 most common alleles for each locus are coloured, and any remaining alleles are shown in greyscale. Loci are sorted in increasing order of sequence-based frequency of the most common allele at each locus (i.e. D2S1338 has the lowest frequency most common allele, and D17S1301 has the highest frequency most common allele).

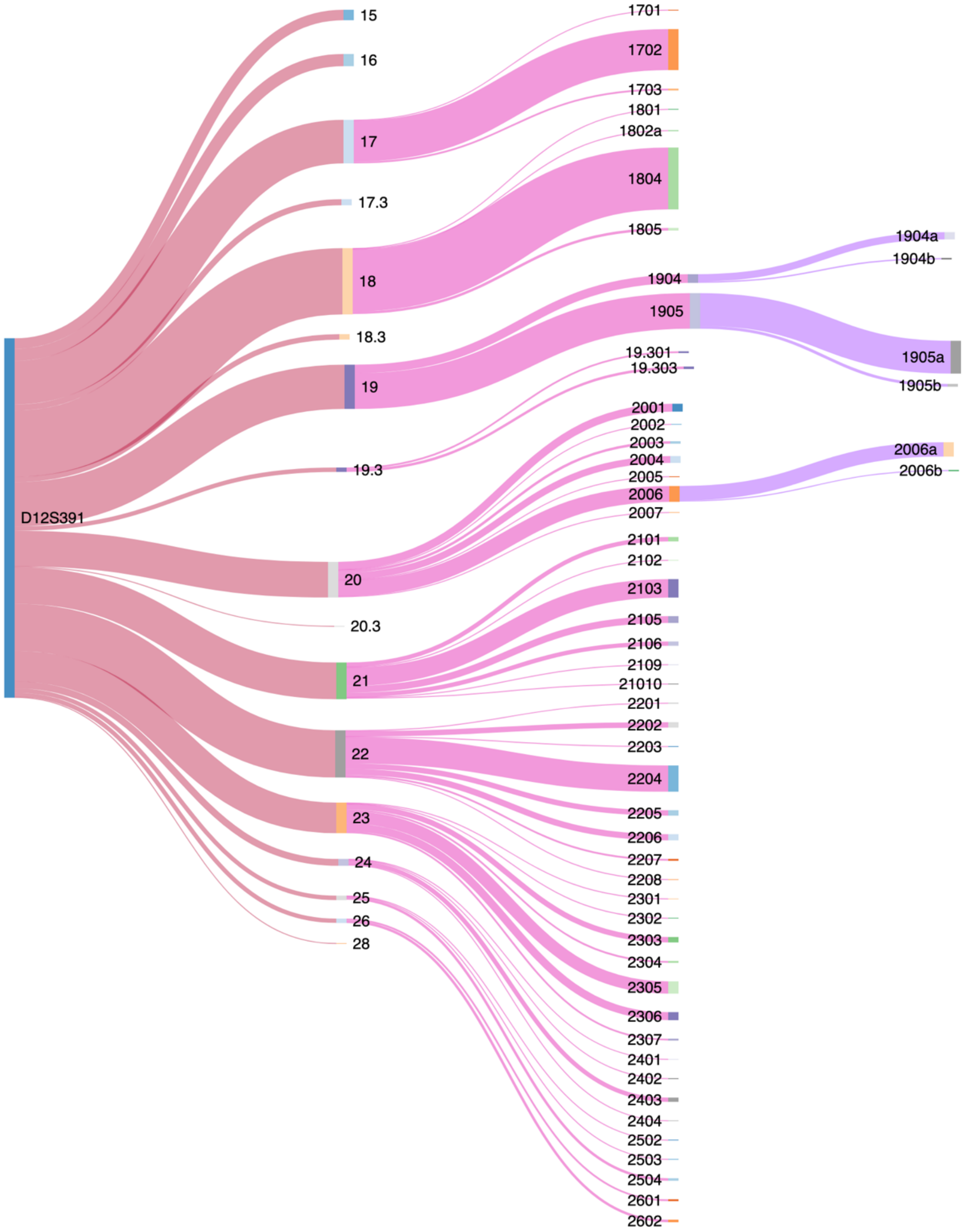


Figure 4.2: Sankey Diagram of D12S391 allelic frequencies in the White British population

The thickness of each band represents the frequency of the corresponding allele. Alleles are split from left to right from length-based (red), repeat region sequence-based (pink) and finally flanking region sequence-based (purple). This diagram was created using http://sankeymatic.com/build/.

Figure 4.3 provides a more detailed view of the frequency distribution for vWA in the West African population. This population was chosen as it is the one showing the greatest allelic gain for this marker. The graph clearly shows that the most common motif for this locus is [TCTA] [TCTG]$_4$ [TCTA]$_n$, which makes up the majority of the common alleles by length. The most common allele by length is a 16, with a frequency of 0.29 in the West African population. Breaking it down by sequence, the most common sequence-based allele, [TCTA] [TCTG]$_4$[TCTA]$_{11}$ still has a frequency of 0.21, as seen in Figure 4.3. As suggested above, most of the alleles gained by sequence for this marker are in fact quite rare. Variants found within the flanking regions are coloured in yellow and orange (with the rs numbers given in the figure legend), and account for a small fraction of the sequence variation observed at this locus in this population.

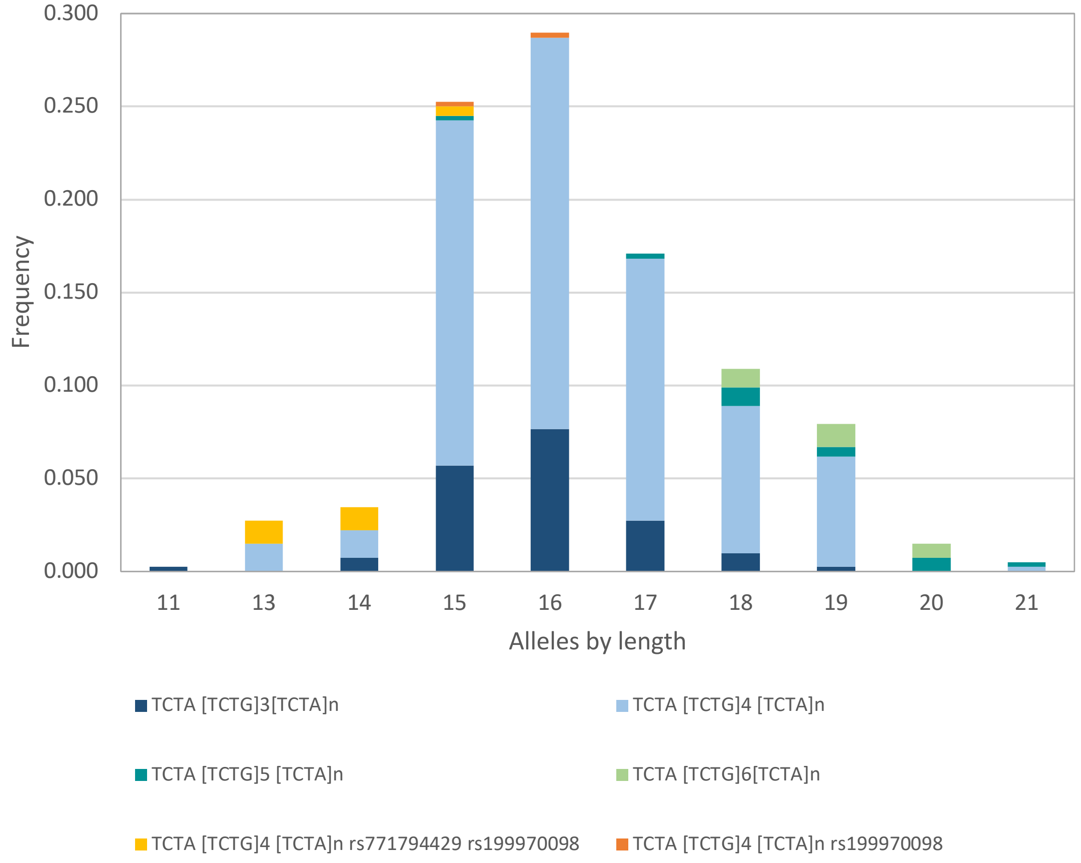


Figure 4.3: Allele frequency distribution for vWA in the West African population

The primary motif for vWA is as follows: [TCTA] [TCTG]$_{3\text{-}6}$ [TCTA]$_n$. The different length-based alleles observed at vWA are split into the main sequence motifs reported in this work. For simplicity, four additional rare motif alleles present in the West African dataset were not included.

A more detailed view of the frequency distribution for D2S1338 in the West African population is shown in Figure 4.4. As before, this population was chosen for demonstration as it is the one showing the greatest allelic gain for this marker. Here, the frequency for the most common allele by length, a 19, is effectively at least halved when considering sequence-level information. A length-based 19 allele at this locus has a frequency of 0.171 in the West African population. When accounting for sequence variation, there are three different "versions" of a 19 allele observed with the following repeat motifs: $[TGCC]_5[TTCC]_{14}$, $[TGCC]_6[TTCC]_{13}$, $[TGCC]_6[TTCC]_{10}[GTCC][TTCC]_2$, and $[TGCC]_7[TTCC]_{12}$ with frequencies of 0.00742, 0.0767, 0.0149 and 0.0718, respectively. With a frequency of 0.0767, the $[TGCC]_6[TTCC]_{13}$ allele is the most common by sequence for D2S1338, but is still 50% less common than the 19 allele by length in the population.

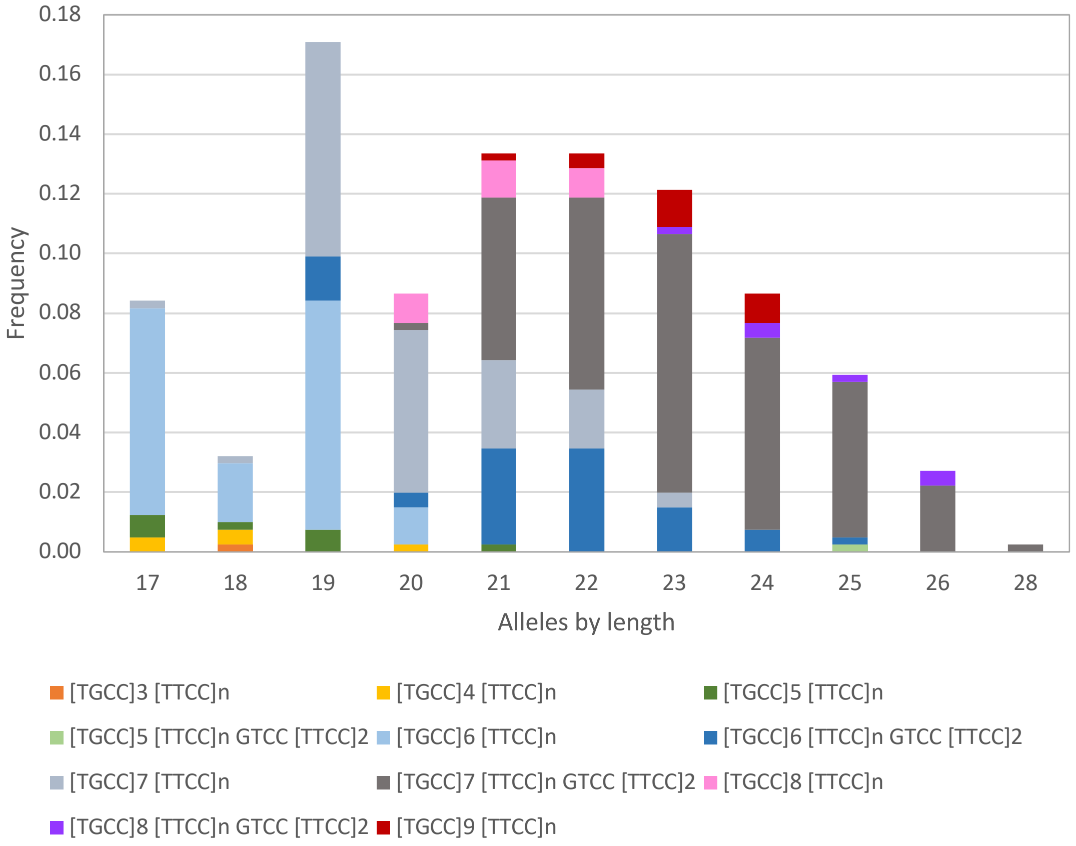


Figure 4.4: Allele frequency distribution for D2S1338 in the West African population

The two primary motifs for D2S1338 are as follows: $[TGCC]_{3-9}$ $[TTCC]_n$ and $[TGCC]_{3-9}$ $[TTCC]_n$ GTCC $[TTCC]_2$. The different length-based alleles observed at D2S1338 are split into the main sequence motifs reported in this work. For simplicity, four additional rare motif alleles were not included.

By increasing the number of alleles distinguishable at common autosomal STR loci, MPS offers an improved power of discrimination for these tests. The more common an allele is within a population, the more likely it is that two individuals at random would share that allele. The increased granularity achieved through sequence-level information means that for many markers, these "common alleles" are, in fact, less common than previously thought. Lower frequencies will lead to more pronounced likelihood ratios, in favour or against the hypothesis that two profiles come from unrelated people for example. The population genetics parameters discussed in the next section offer a different view of the increased power of STR testing using MPS.

4.2.2. Locus diversity

4.2.2.1. Expected heterozygosity

The expected heterozygosity of a locus within a population gives an indication of genetic variability. Expected heterozygosity (H_{exp}), also referred to as a genetic diversity (D), ranges from zero to almost 1 (for a system with a large number of equally frequent alleles) and represents the number of heterozygotes that would be expected under Hardy Weinberg Equilibrium, based on the observed allele frequencies in the sampled population group. H_{exp} values for all loci in each population group were calculated using Arlequin [188, 218] and Forstat [191].

Table 4.3 compares the heterozygosity by length and by sequence for the 26 autosomal STRs targeted in this study, sorted in descending order of average absolute increase of heterozygosity. D9S1122 showed the largest average increase in H_{exp} across all five populations, and is also the one showing the most pronounced increase in the White British, North East African and South Asian populations. Other loci showing an average absolute increase in heterozygosity of over 5% by sequence compared to length are D3S1358, D13S317, D5S818, D8S1179, D21S11, D12S391, D7S820, D2S1338 and D2S441. This general trend of locus diversity is consistent with previous studies for the markers targeted by the ForenSeq DNA Signature Prep kit in similar populations [63, 98]. Of these ten loci, eight correspond to markers which showed 100% or more gain in number of distinguishable alleles (presented in the previous chapter). Figure 4.5 shows a comparison between the number of alleles gained by sequence for each marker across all five populations, and the average gain in heterozygosity. This suggests a general correlation between an increase in alleles and an increase in heterozygosity. D9S1122 and D5S818 did not show such a high increase in allelic numbers, and yet based on H_{exp} show a high increase in diversity when taking sequence-based alleles into account. This is due to the fact that although there isn't a great number of new, rare sequence-based alleles, the sequence diversity at these loci has effectively split up some of the more frequent length-based alleles (as demonstrated for D2S1338 in Figure 4.4). Interestingly, vWA shows an increase in number of alleles of 300% across all population, and yet only shows an increase in H_{exp} of over 5% in the West African population. As suggested in the previous section, it is likely that the large number of rare alleles at this locus cause it to show a huge increase in allelic numbers, yet the frequencies are less useful than anticipated. Frequency distribution within a population and initial heterozygosity by length are likely to affect the gains in heterozygosity by sequence. These results highlight the importance of taking allele frequencies and locus diversity into account when assessing a marker set. Across all populations and all markers, average STR marker diversity was 0.785 when analysed by length and 0.82 when sequence information was considered. This equates to an increase in average diversity of 0.035 with a maximal increase of 0.145 at D3S1358 in the West African population.

Table 4.3: Expected heterozygosity for length and sequence-based allelic data

For each population, length-based (LB), sequence-based (SB), and the increase (Inc) in expected heterozygosity is provided. Loci are listed in decreasing order of average increase in expected heterozygosity across all five populations, and values over 0.05 are highlighted in bold.

Locus	White British			British Chinese			North East African			South Asian			West African		
	LB	SB	Inc	LB	SB	Inc	LB	SB	Inc	LB	SB	Inc	LB	SB	Inc
D9S1122	0.713	0.826	**0.113**	0.719	0.785	**0.066**	0.701	0.844	**0.143**	0.702	0.834	**0.132**	0.711	0.803	**0.091**
D3S1358	0.792	0.860	**0.067**	0.718	0.792	**0.074**	0.747	0.890	**0.143**	0.757	0.867	**0.110**	0.740	0.885	**0.145**
D13S317	0.772	0.881	**0.109**	0.783	0.822	0.038	0.769	0.864	**0.095**	0.803	0.873	**0.070**	0.677	0.798	**0.121**
D5S818	0.712	0.788	**0.076**	0.779	0.818	0.039	0.737	0.858	**0.121**	0.746	0.809	**0.063**	0.738	0.841	**0.103**
D8S1179	0.801	0.867	**0.065**	0.856	0.906	**0.050**	0.797	0.888	**0.091**	0.843	0.905	**0.062**	0.770	0.876	**0.107**
D21S11	0.839	0.906	**0.067**	0.816	0.914	**0.098**	0.828	0.917	**0.089**	0.855	0.908	**0.053**	0.858	0.917	**0.059**
D12S391	0.891	0.934	0.043	0.840	0.907	**0.068**	0.841	0.937	**0.096**	0.860	0.902	0.042	0.850	0.922	**0.072**
D7S820	0.809	0.864	**0.055**	0.755	0.807	**0.053**	0.755	0.824	**0.069**	0.804	0.862	**0.058**	0.786	0.826	0.040
D2S1338	0.882	0.905	0.022	0.861	0.903	0.042	0.854	0.937	**0.083**	0.878	0.930	**0.052**	0.890	0.954	**0.064**
D2S441	0.761	0.809	0.047	0.790	0.851	**0.061**	0.790	0.819	0.030	0.733	0.795	**0.062**	0.742	0.795	**0.053**
vWA	0.809	0.852	0.043	0.797	0.807	0.010	0.821	0.868	0.047	0.794	0.819	0.025	0.800	0.882	**0.082**
D20S482	0.692	0.747	**0.055**	0.713	0.743	0.030	0.744	0.791	0.047	0.706	0.752	0.046	0.703	0.710	0.007
D16S539	0.765	0.777	0.012	0.795	0.825	0.030	0.790	0.810	0.020	0.805	0.860	**0.055**	0.801	0.846	0.044
D1S1656	0.904	0.922	0.018	0.818	0.831	0.012	0.860	0.904	0.044	0.882	0.898	0.016	0.856	0.905	0.049
D4S2408	0.745	0.771	0.026	0.729	0.775	0.047	0.745	0.747	0.002	0.771	0.776	0.006	0.743	0.743	0.001
PentaD	0.838	0.843	0.004	0.801	0.807	0.006	0.879	0.883	0.003	0.828	0.828	0.001	0.871	0.874	0.003
D19S433	0.765	0.765	0.000	0.814	0.814	0.000	0.805	0.805	0.000	0.832	0.844	0.012	0.835	0.836	0.000
D17S1301	0.657	0.657	0.000	0.714	0.719	0.005	0.648	0.649	0.001	0.668	0.668	0.000	0.621	0.625	0.004
D18S51	0.875	0.876	0.001	0.865	0.865	0.000	0.907	0.907	0.000	0.826	0.832	0.006	0.874	0.876	0.002
CSF1PO	0.730	0.732	0.002	0.722	0.726	0.004	0.764	0.765	0.001	0.714	0.714	0.000	0.793	0.794	0.001
D6S1043	0.793	0.794	0.001	0.875	0.875	0.000	0.858	0.860	0.002	0.816	0.818	0.002	0.881	0.883	0.002
PentaE	0.888	0.888	0.001	0.917	0.917	0.000	0.911	0.916	0.005	0.905	0.905	0.000	0.896	0.898	0.002
FGA	0.863	0.863	0.000	0.856	0.856	0.000	0.864	0.865	0.001	0.865	0.865	0.000	0.879	0.881	0.003
TH01	0.770	0.770	0.000	0.699	0.699	0.000	0.762	0.762	0.000	0.787	0.789	0.002	0.722	0.722	0.000
D10S1248	0.766	0.766	0.000	0.765	0.766	0.000	0.783	0.766	0.000	0.769	0.766	0.000	0.811	0.766	0.000
TPOX	0.616	0.616	0.000	0.574	0.616	0.000	0.725	0.616	0.000	0.721	0.616	0.000	0.781	0.616	0.000

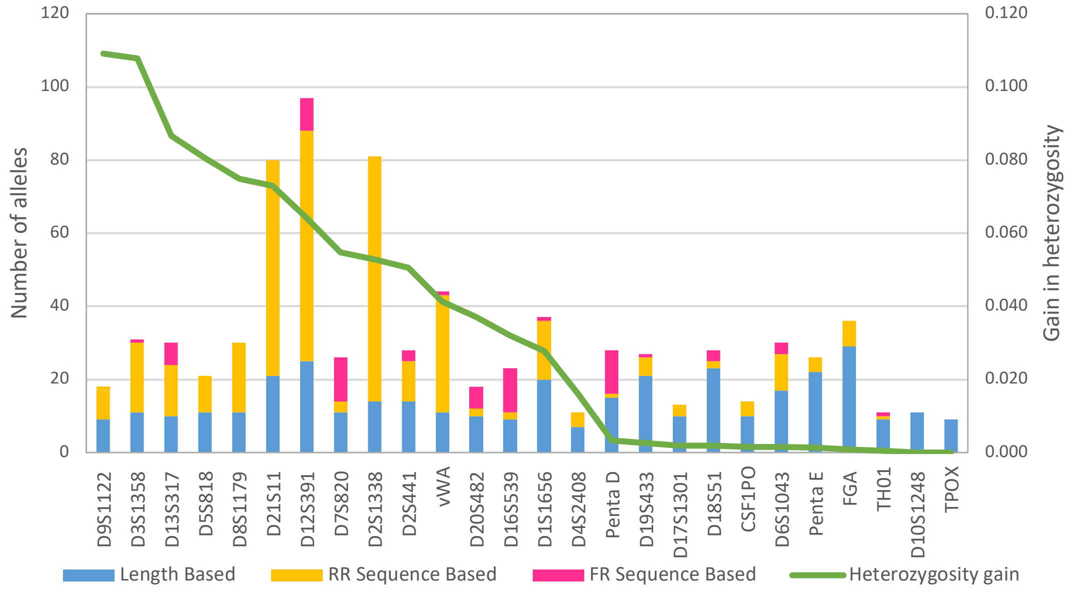


Figure 4.5: Allelic gain by sequence compared to average gains in heterozygosity

Different colours indicate the number of alleles observable by length (blue), repeat region sequence (RR, yellow) and flanking region sequence (FR, pink) across all five populations studied. The green line indicates average gain in heterozygosity, using a secondary Y axis, and STRs are listed in descending order of average increase in expected heterozygosity on the X axis as in Table 4.3.

4.2.2.2. Match probability

Another measure of locus diversity is match probability (MP), which is the probability that a person at random within a population would have a certain genotype [212, 219]. The lower the match the probability, the higher the power of discrimination of a locus. Table 4.4 compares the match probability values for the 26 loci in the five populations for length and sequence-based genotypes, and gives the combined MP for this marker set in each population. Including sequence variation for both the repeat and flanking regions made the average combined MP across all populations 2,756 times lower than looking at length-based alleles alone. The overall increase in heterozygosity and decrease in combined match probability when moving from length-based to sequence-based allele differentiation prove the added value of sequencing STRs using MPS.

Table 4.4: Loci match probabilities for length and sequence-based allelic data

For each population, length-based (LB), sequence-based (SB), and the decrease (Dec) in match probability is provided. Loci are listed in decreasing order of average decrease in match probability across all five populations.

Locus	White British			British Chinese			North East African			South Asian			West African		
	LB	SB	Dec	LB	SB	Dec	LB	SB	Dec	LB	SB	Dec	LB	SB	Dec
D9S1122	0.138	0.058	0.080	0.131	0.078	0.053	0.151	0.047	0.104	0.141	0.052	0.089	0.132	0.065	0.067
D3S1358	0.077	0.039	0.039	0.136	0.073	0.064	0.114	0.029	0.086	0.104	0.034	0.070	0.107	0.027	0.080
D5S818	0.127	0.068	0.059	0.087	0.062	0.025	0.119	0.039	0.080	0.108	0.065	0.043	0.111	0.044	0.067
D13S317	0.089	0.029	0.060	0.082	0.055	0.027	0.084	0.035	0.050	0.069	0.031	0.037	0.166	0.068	0.098
D8S1179	0.073	0.034	0.038	0.042	0.021	0.022	0.071	0.025	0.045	0.050	0.022	0.028	0.089	0.031	0.058
D2S441	0.094	0.059	0.035	0.084	0.044	0.040	0.078	0.051	0.026	0.111	0.070	0.041	0.107	0.072	0.035
D7S820	0.066	0.038	0.028	0.101	0.057	0.045	0.098	0.054	0.044	0.066	0.037	0.029	0.081	0.054	0.027
D20S482	0.148	0.102	0.046	0.131	0.103	0.028	0.109	0.073	0.036	0.133	0.096	0.037	0.130	0.123	0.007
D21S11	0.049	0.021	0.028	0.058	0.017	0.041	0.053	0.018	0.036	0.041	0.020	0.021	0.041	0.017	0.024
D12S391	0.026	0.012	0.014	0.049	0.018	0.031	0.048	0.011	0.037	0.038	0.020	0.017	0.043	0.016	0.027
vWA	0.073	0.044	0.029	0.075	0.066	0.009	0.059	0.032	0.027	0.080	0.065	0.015	0.077	0.031	0.046
D16S539	0.094	0.084	0.010	0.075	0.058	0.017	0.079	0.065	0.014	0.064	0.037	0.028	0.069	0.043	0.026
D2S1338	0.029	0.020	0.008	0.042	0.025	0.016	0.041	0.011	0.030	0.034	0.013	0.021	0.027	0.009	0.018
D4S2408	0.116	0.096	0.021	0.121	0.087	0.034	0.114	0.113	0.001	0.092	0.086	0.006	0.111	0.111	0.001
D1S1656	0.023	0.016	0.006	0.056	0.052	0.004	0.042	0.023	0.019	0.029	0.024	0.006	0.042	0.020	0.022
Penta.D	0.048	0.044	0.003	0.072	0.070	0.002	0.031	0.028	0.004	0.054	0.048	0.006	0.034	0.031	0.003
D17S1301	0.187	0.187	0.000	0.122	0.120	0.002	0.171	0.171	0.001	0.166	0.166	0.000	0.207	0.201	0.007
CSF1PO	0.121	0.119	0.002	0.139	0.135	0.004	0.092	0.091	0.001	0.142	0.142	0.000	0.077	0.077	0.001
D19S433	0.087	0.087	0.000	0.066	0.066	0.000	0.062	0.062	0.000	0.050	0.043	0.006	0.056	0.056	0.000
D6S1043	0.067	0.066	0.002	0.033	0.033	0.000	0.039	0.038	0.001	0.056	0.055	0.001	0.030	0.029	0.001
D18S51	0.034	0.033	0.000	0.044	0.044	0.000	0.022	0.022	0.000	0.053	0.051	0.002	0.035	0.034	0.001
FGA	0.039	0.039	0.000	0.042	0.042	0.000	0.039	0.039	0.001	0.038	0.038	0.000	0.031	0.029	0.002
THO1	0.092	0.092	0.000	0.144	0.144	0.000	0.096	0.096	0.000	0.077	0.076	0.002	0.115	0.115	0.000
D10S1248	0.098	0.098	0.000	0.096	0.096	0.000	0.081	0.081	0.000	0.092	0.092	0.000	0.063	0.063	0.000
Penta.E	0.025	0.025	0.000	0.017	0.017	0.000	0.020	0.020	0.000	0.021	0.021	0.000	0.024	0.024	0.000
TPOX	0.207	0.207	0.000	0.235	0.235	0.000	0.120	0.120	0.000	0.122	0.122	0.000	0.084	0.084	0.000
Combined	2.08E-30	3.51E-34		7.31E-30	3.51E-33		4.51E-31	1.79E-36		5.07E-31	4.76E-35		4.39E-31	6.20E-36	

4.2.3. Flanking region analysis of autosomal STRs

As discussed in the previous chapter, it became clear when characterising autosomal STR alleles using MPS that there is considerably more variation in the repeat region of the amplicons compared to the flanking regions. It is important to understand what this means in terms of allelic frequencies and locus diversity, as this is in fact more important when looking at implementing and applying MPS data to forensic casework. When characterising flanking region variation, D7S820, D16S539, D20S482 and Penta D showed the most pronounced increase in allelic number, with flanking region variants accounting for approximately half of the characterised alleles for the first three STRs.

4.2.3.1. Flanking region allele frequency distribution

Figure 4.6 shows the distribution of allelic frequencies at D7S820 and D16S359 in two different populations, and the distribution of frequencies at a marker showing only variation in the repeat region, D2S1338, for comparison. SNP rs11642858, located within the flanking region of D16S539, is seen from 1000 Genomes data to be observed in East Asian populations at a frequency of 0.56 for the major allele and 0.44 for the minor allele [185], suggesting that the inclusion of this SNP when analysing D16S539 STR alleles should significantly increase the allelic diversity. Haplotypic data for this locus in the British Chinese population, as displayed in Figure 4.6, shows a different picture however, with the rs11642858 variant being highly associated with only a few STR alleles. It is observed with almost all 10 alleles and the majority of allele 9s, but rarely associated with an 11 or 12 allele and never observed with an allele number higher than 12. The result of this is that locus diversity is seen to only increase from 0.795 to 0.825 in the British Chinese population when considering flanking region variation despite the added presence of a SNP showing high diversity at population level. This locus diversity increase is only slightly more pronounced in the West African population, going from 0.802 to 0.846. A similar pattern is seen at locus D7S820, where it is again evident that specific flanking region SNPs are strongly associated with particular alleles, reflecting the evolutionary and mutational history of these variants. This was also observed to a certain extent at D12S391, as seen in Figure 4.2, where despite substantial sequence variation, the flanking region SNPs are once again strongly associated with particular allele length classes. A founder effect would lead to certain SNPs being highly associated, presumably, with the STR allele upon which they first arose, which were then only distributed to close alleles that are one, or at most two, mutational steps away. This is in contrast to much repeat region variation, especially for compound STRs such as D2S1338, where the increase of diversity can be quite pronounced, here increasing from 0.890 to 0.954 when taking repeat region sequence variation into account in the West African population.

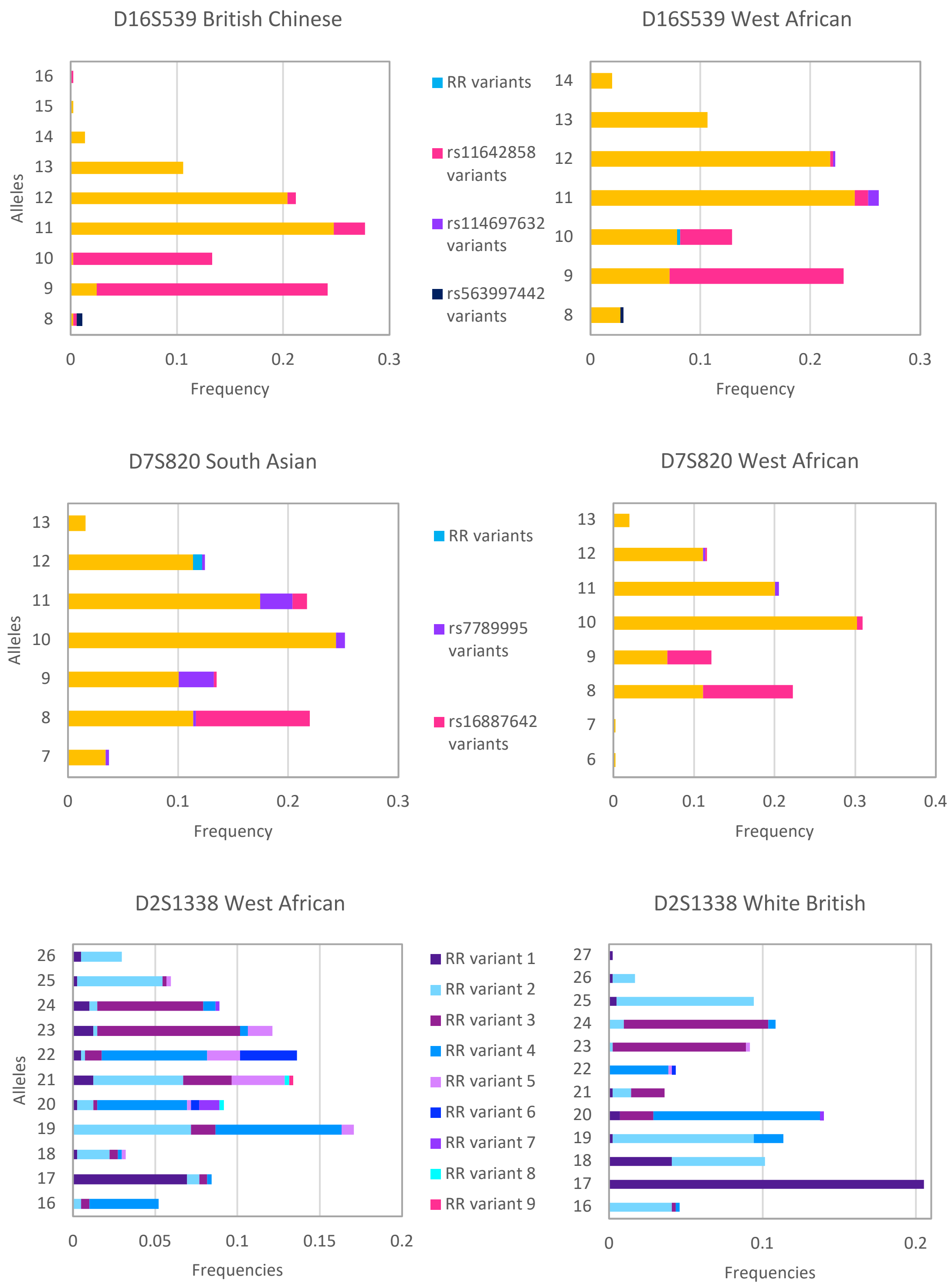


Figure 4.6: Distribution of allelic frequencies at D16S359, D7S820 and D2S1338

For D16S359 and D7S820, the frequency of the most common repeat motif allele is shown in yellow, with other colours representing either repeat region (RR) variants or flanking region variants, where the rs number of the flanking region SNP is given. For D2S1338, which shows no flanking region variation, the different levels of repeat region variants are highlighted.

4.2.3.2. Effect of flanking regions on locus diversity

Figure 4.7 shows the gains in heterozygosity split by repeat and flanking region sequences. Of the 15 loci which showed an average gain in heterozygosity of over 0.01 (from the data in Table 4.3), 11 show an increase in heterozygosity compared to length-based alleles almost exclusively due to repeat region variation. At D7S820, D16S539 and D20S482 the reverse is true, with gains in heterozygosity mostly linked to flanking region variation. This is concordant with the markers showing the highest gain in flanking region allele numbers in the previous chapter. Although the number of alleles observed at Penta D in the global sample set practically doubles due to flanking region variants, the effect of increased heterozygosity is negligible, highlighting once more that certain variants at specific loci are likely to be rare. D2S441 shows considerable population-specific differences, with flanking region variants accounting for 50% of the increase in heterozygosity in the White British population but having little to no effect in the two African populations for example. The potential increase in diversity when analysing flanking region variation is much less pronounced than might be expected given the mutation rate of STRs and the accompanying initial assumption that flanking region variations would spread throughout the allele range of an STR marker. Of the 0.035 increase in average STR marker diversity across all populations and all markers, when sequence information was considered, only 0.005 is attributed to flanking region sequences.

The match probabilities of the combined 26 autosomal STR marker set when considering flanking region variation tells a similar story, with the average combined MPs going from 2.16 X 10^{-30} to 2.73 X 10^{-33} and 7.82 X 10^{-34} for length-based, sequence-based without flanking regions and sequence-based with flanking regions respectively across all populations. This means that including the sequence variation within repeat regions made the average combined RMP 789 times lower than with length-based alleles alone. Including the sequence variation within the flanking regions only resulted in an average combined MP that a was further 3.5 times lower. This finding is similar to that of Delest et al., who observed combined MPs for the autosomal STRs in the ForenSeq DNA Signature Prep Kit in their French population data of 8.99 X 10^{-31}, 7.12 X 10^{-34} and 7.12 X 10^{-34} for length-based, sequence-based without flanking regions and sequence-based with flanking regions respectively [212].

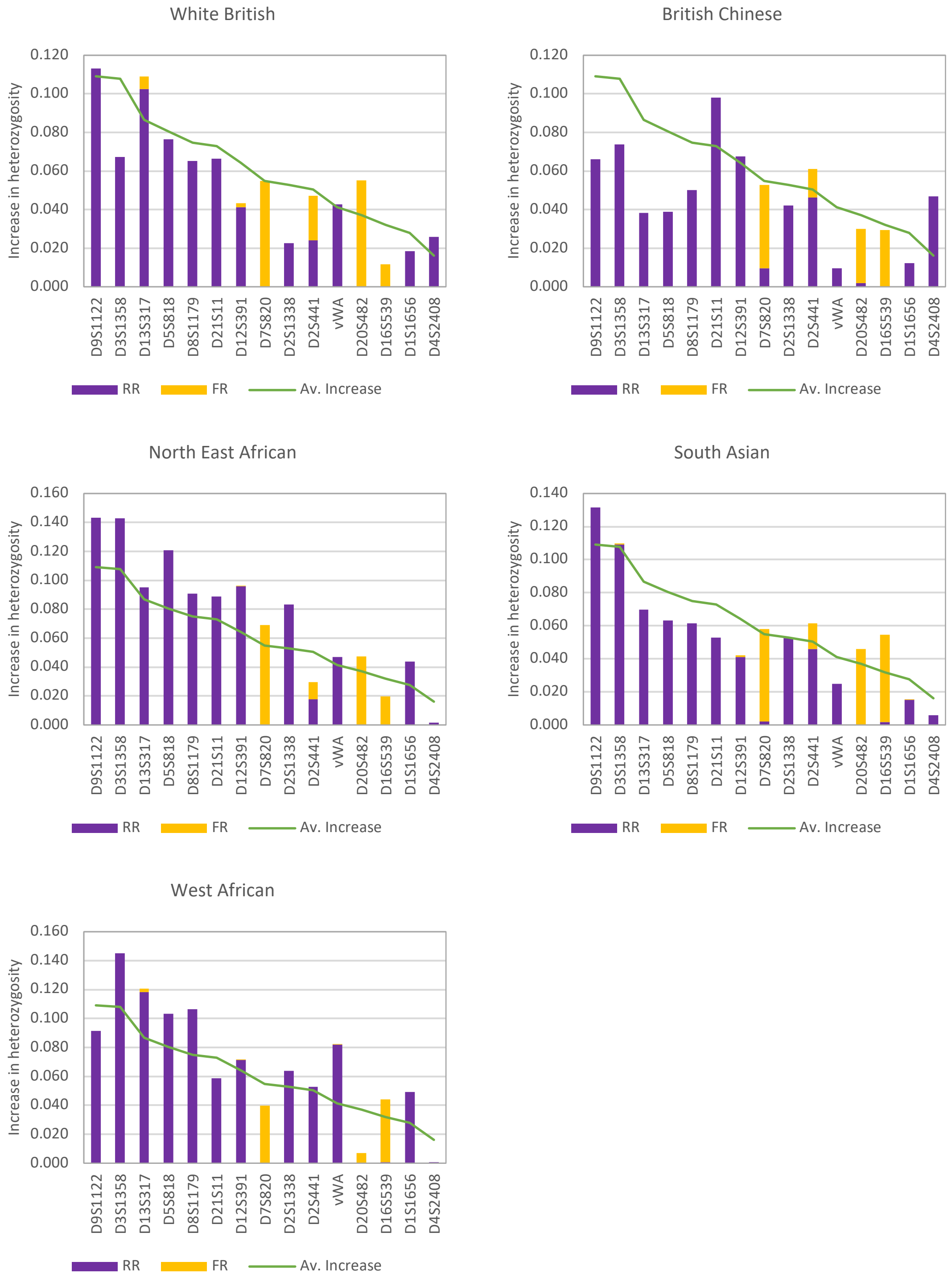


Figure 4.7: Gains in heterozygosity split by repeat and flanking region sequences

The five populations are split into different graphs, showing the gain in heterozygosity when taking repeat region (RR, purple) and flanking region (FR, yellow) sequence-based allelic frequencies into account. The green line indicates average gain in heterozygosity across all populations. STRs are listed in descending order of average increase in expected heterozygosity on the X axis, and 11 loci with an average heterozygosity gain of under 0.01 are not represented.

4.2.3.3. Power of flanking region variation

The results discussed so far indicate that flanking region variation is markedly less useful in terms of increasing the power of discrimination of the autosomal STR marker set studied than repeat region variation, making their power for forensic identification purposes limited. Flanking region analysis is more complex than repeat region analysis and complicates the naming of sequence-based alleles. This, in turn, makes it more difficult to compare results between different commercial kits, and may require analysts to have an in depth understand of allelic sequences rather than just being able to rely on a more standard format of allelic designation. The limited additional power of discrimination provided by STR flanking regions, as well as the complexity of analysis, has been highlighted in previously published research. Wendt et al. observed some flanking region variation at 6 autosomal STRs in the Yavapai population but note that with refinement of primer sequences for commercially available MPS kits, genomic coordinates for amplicons may change, influencing the exact flanking region sequence that may be seen at a target marker, making analysis more complicated [96].

There does, however, remain advantages to having characterised these flanking region sequences and included them in sequence-based frequencies. First, as discussed in the previous chapter, using the full sequence available is vital for accurate allele characterisation, especially in relation to back compatibility with length-based allele designation. Secondly, there is no additional cost to looking at flanking regions save perhaps the added time to investigate them. It is likely that, in the future, commercial data analysis pipelines will incorporate flanking region sequences into the standard output of sequencing data – certain bio-informatic pipelines such as STRNaming do so already [128]. This in turn should remove some of the hurdles of flanking region analysis, and lead to a more global adoption of these sequences as part of MPS analysis. Finally, despite the fact that the inclusion of flanking regions has a limited impact on overall marker diversity and match probability, they may be useful in specific cases, in the same way as any other rare allele. At D16S539 in the British Chinese population for example, although the rs11642858 is rarely associated with an allele 11 as discussed earlier, this allele does have a frequency of 0.0289 in that population. If this allele was shared between two profiles in a criminal, relationship, or victim identification scenario, it could make a huge difference to resulting likelihood ratio – appreciably more so than an 11 allele by length (given there is no repeat region variation), which has a frequency of 0.271 in the British Chinese population.

4.3. Additional considerations

4.3.1. Sample size

Population DNA database are important to understand how rare a genetic profile might be. In an ideal world, a database used for identification would include STR genotypes from every individual within a population – but of course this is unlikely to be practical for cost and time reasons and would have serious ethical and legal implications to consider. The statistics used in forensic DNA analysis rely on the premise that data extracted from a subset of the population can be used to generate representative allelic frequencies, which in turn can be used to assess the rarity of genotypes. In 1992, Chakraborty discussed the need for large population databases to observe all possible genotypes for polymorphic marker systems used for individualisation, and wrote "… adequate estimation of genotypic probabilities must be based on allele frequencies, and the sample size needed to represent all possible alleles is far more reasonable" [21]. The key is therefore to collect results from enough individuals to reliably estimate the frequency of major alleles [220]. A population sample size of approximately 200 has been used as the recent benchmark for generating length-based allelic frequencies for current commercial kits, is estimated to encompass all common alleles, and provide their representative frequencies within a population. This benchmark is used globally, and is included in the UK Forensic Regulator guidance of 2020 [22], although they acknowledge that the latest ISFG guidelines do suggest an increased minimum of 500 samples genotyped where possible [175]. The large number of sequence variants characterised at select markers with MPS brings into question the strategy for producing representative population data with this technology [98].

4.3.1.1. How many samples are required?

As shown earlier in this chapter, some of the common alleles detected at autosomal STRs are now drastically less common due to sequence-level variation, and so further research was needed to establish whether 200 samples were sufficient to capture all common sequence-based alleles. As described in the materials and methods chapter, sequence-based alleles were added to an internal database created to "name" sequences. To gain an idea of how many samples must be typed to identify the common variants at any given STR locus, sample order was randomised, and “novel” alleles were recorded as they were added to the database (i.e. the first allele added is the first allele seen, the second allele would either be the same as the first, or a new allele etc.). The number of new sequence-based alleles were plotted against the total number of alleles sequenced. For markers showing limited diversity, such as TPOX or TH01, a very limited number of samples needed to be sequenced to capture the full breadth of variation for each locus. For highly polymorphic markers such a vWA and D12S391, new variants were regularly observed as more samples were sequenced. Figure 4.8 shows the number of alleles observed against alleles typed for four markers: TPOX, CSF1PO, vWA and D12S391. The graph for TPOX, which showed very limited allele diversity by size and no sequence-variation, indicates that 200 samples (i.e., 400 alleles) is more than sufficient for the line to “plateau”, meaning all expected alleles at any appreciable frequency within the population have been observed at this point, and no new alleles are discovered as more sample data is added. This is in contrast to the graphs for vWA and D12S391, which show that new variants are still being observed when 400 alleles have been typed, suggesting more samples need to be analysed to see the full range of commonly expected alleles for these loci. At vWA, a plateau is seen for some populations (North East African, green line and British Chinese, purple line), whereas all five populations show no decrease in the rate of new allele discovery for D12S391.

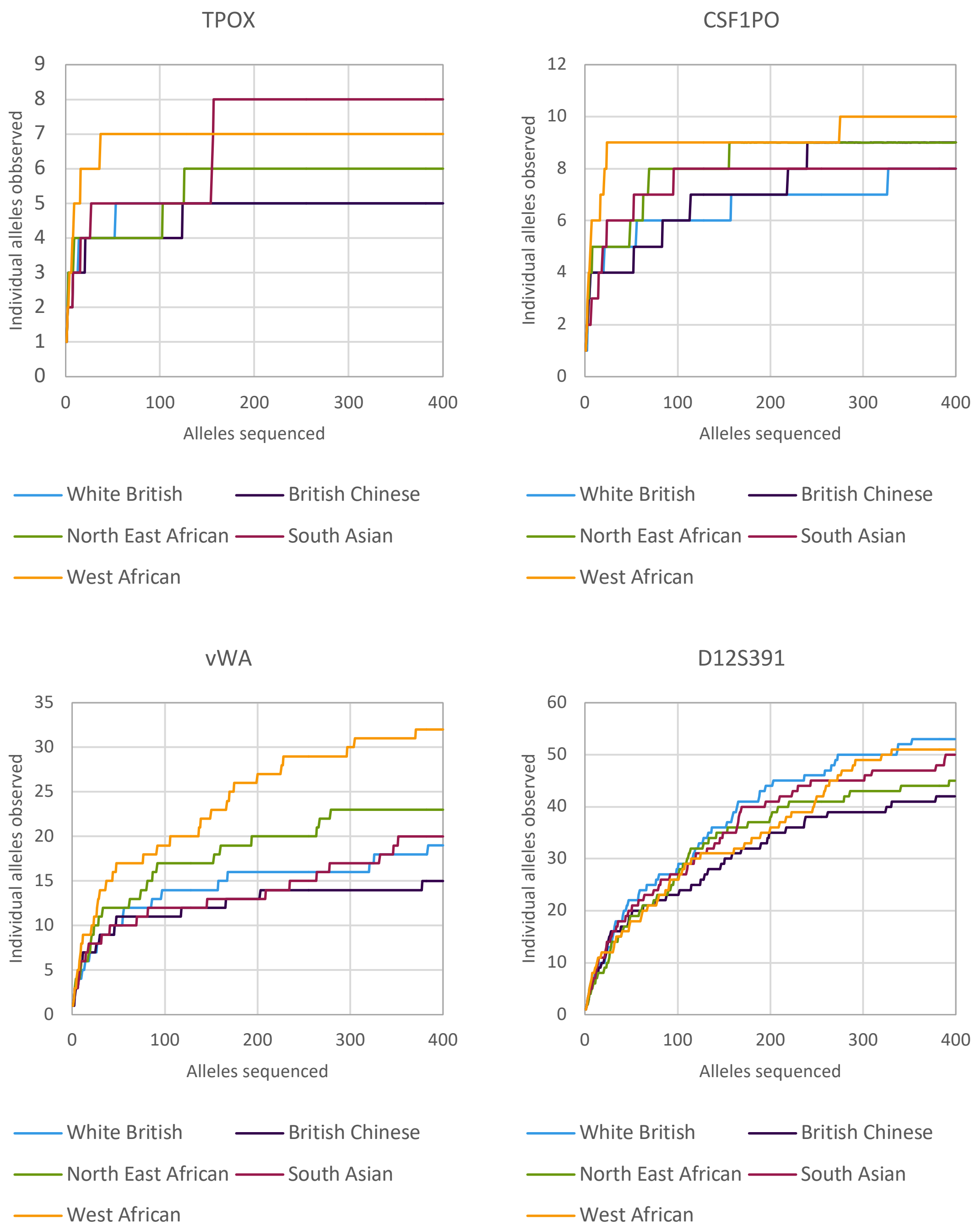
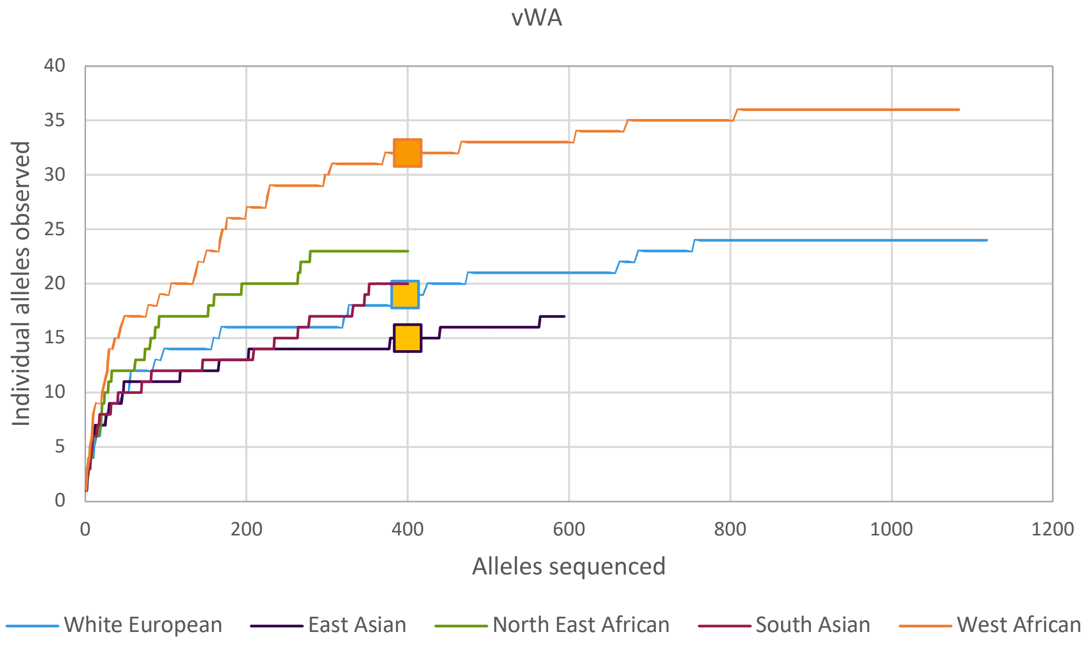


Figure 4.8: Number of individual alleles observed against number of alleles sequenced

Sample order was randomised, and each newly observed allele plotted against the total number of alleles sequenced. The five differently coloured lines represent the populations studied, and all show the increase in number of alleles observed as more samples are sequenced. TPOX and CSF1PO were chosen as examples of low allelic diversity markers, whereas vWA and D12S391 were chosen as examples of highly polymorphic markers.

Sequencing hundreds more samples from each population to identify the ideal sample size was not operationally feasible, but a collaboration with the National Institute of Standards and Technology (NIST), USA, meant data could be shared and compared for this purpose. Results for a subset of the loci studied, from samples analysed at NIST (and subsequently published in [98]) were added to those of the samples sequenced in this work, for three sets of comparable population groups: White British and Caucasian; West African and African American; British Chinese and Asian. For ease of analysis and comparison, only repeat-region sequence-based alleles were considered. Results from this collaborative study demonstrate that the number of samples needed to capture the breadth of allelic variation is highly dependent on the individual marker and the extent of its sequence variability. Figure 4.9 shows that at vWA, adding more samples appears to achieve the expected plateau in the graph, suggesting that closer to 400 samples (i.e., 800 alleles) need to be sequenced to see all common variation at this locus. Although the NIST dataset contained fewer East Asian samples than for the West African and White European groups, a levelling of the graph is still seen. The addition of alleles observed within the comparable population groups at NIST show that even above 1000 alleles, i.e., 500 samples, novel alleles are still being discovered for D12S391. This marker was the most polymorphic within the 27 STRs studied, but as discussed earlier in this chapter, many of the "new" sequence-based alleles observed are quite rare and it is therefore perhaps unsurprising that all possible sequence variants have not yet been characterised.



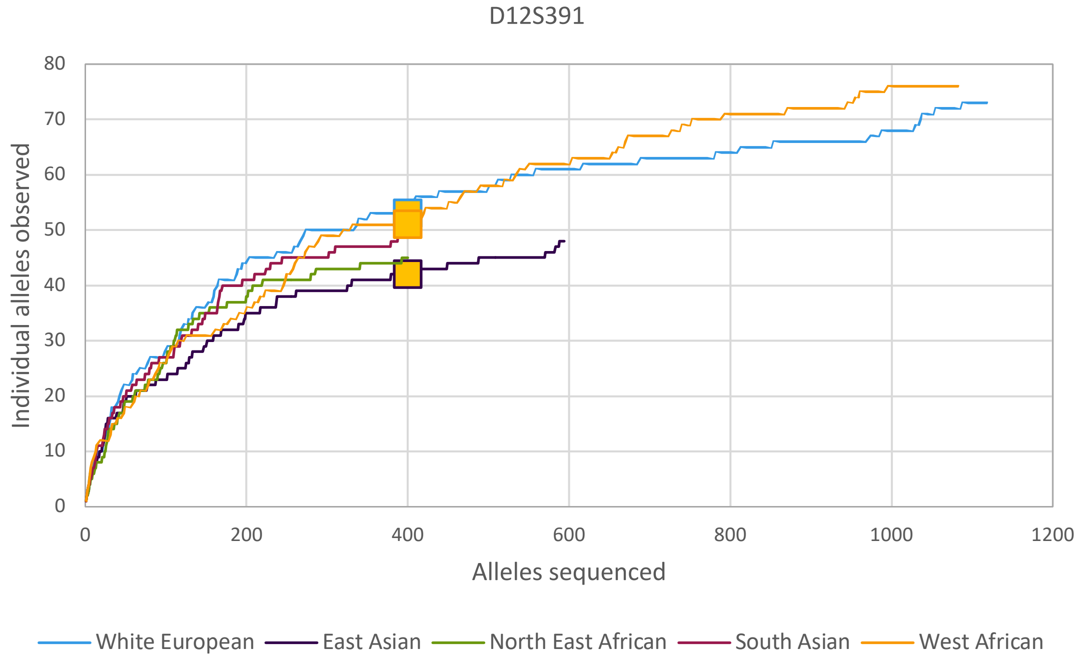


Figure 4.9: Number of individual alleles observed against number of alleles sequenced including NIST data

Combined results for the number of alleles observed plotted against number of alleles sequenced, including NIST data, for vWA (top) and D12S391 (bottom). The yellow box on the White European (White British + Caucasian), West African (West African + African American) and East Asian (British Chinese + Asian) graphs highlight allele 400 – after which all additional alleles sequenced are from the NIST data set.

Because of the nature of STRs, and their high genetic variability when accounting for sequence variation in particular, it is expected that certain loci will show a predominance of rare alleles. In their 1992 paper, Chakraborty wrote about VNTR allele distribution: "… even when the total number of alleles is large, the expected number of alleles having frequency *p* or above is generally below 10 for p=0.001, 0.01, or 0.05." [21]. Although this was written about VNTRs, this is likely to be applicable to the highly polymorphic sequenced-based STR loci, and relates to the sample size needed to adequately estimate the frequency of all common alleles. If there is a very large number of alleles identified, this is not a problem so long as a majority (in this description, 10 or more) are considered rare. In the combined dataset of 989 samples, all 26 loci meet Chakraborty's description for *p*=0.05, i.e., none have 10 or more sequence-based alleles with an average global frequency of 0.05 or higher, highlighting the very large proportion of rare alleles in this dataset. Eight loci meet this description for *p*=0.01, meaning 16 out of the 26 loci have 10 or more alleles with a frequency equal to or above 0.01. One locus meets the description for *p*=0.001, with all other loci (n=25) having 10 or more alleles with a frequency of 0.001 or higher. This locus is in fact TPOX, which is not very polymorphic and where only 9 alleles have been observed in total. The NIST dataset shows similar results, with no loci having 10 or more alleles with a frequency of 0.05 or higher, 10 loci matching the above description for *p*=0.01 and 5 loci for *p*=0.001. Gettings et al. remark on the fact that for several loci the finding is irrespective of sequence variation, stating that "There appear to be five loci for which sequencing will substantially increase the number of alleles beyond the expected range at the given *p*: D1S1656, D2S1338, D8S1179, D12S391, and D21S11" [98]. This is also the case in this work, for example with at locus D1S1656 which has 13 alleles with a frequency of above 0.001 when using length-based data, increasing to 37 alleles with a frequency of above 0.001 when using sequence-based data.

Given the number of samples tested, for large population groups with high random mating, this study should represent the most common alleles and give reliable frequencies. Other populations with more substructure due to geographical, cultural, religious, linguistic, or other reasons, may benefit from further sampling. While the samples used should be representative of the British Chinese or South Asian populations, for example, they may not represent the diversity within these large geographical areas. Outside of the scope of this PhD, the samples in the CEPH panel were analysed using the ForenSeq DNA Signature Prep kit [213]. The CEPH human diversity genome diversity is a sample set of 944 individuals from 51 globally distributed ancestral populations, including many populations not studied as part of this work such as Oceanian and Native American. At D12S391, as highlighted by the contributions to the STRSeq Bioproject, 69 of the same sequences were observed in the data from this project and that of the CEPH panel (97 and 96 submissions, respectively). The fact that such a high number of alleles were shared across both studies, despite substantial differences in population groups and sample size, gives high confidence that this study has captured the more common alleles for the population groups discussed herein.

The STRSeq BioProject [139] will continue to accept submissions for novel alleles, and its future incorporation into a database such as STRidER [175] will help provide allelic nomenclature for novel alleles and population frequencies on a global scale. This is especially important for under-represented population groups, given that the majority of the results published so far focus on White European/Caucasian, Hispanic, East Asian and West African/ African American populations. In the meantime, an approach for minimum allele frequency may need to be adopted such as that currently used by many laboratories for rare alleles. This is often described as 5/2N, where N is the number of individuals sampled from a population. Additionally, or alternatively, as suggested by Phillips et al., a generalised minimum allele frequency of 1% could be applied from previously unobserved sequences [213].

4.3.2. Population specific alleles

Due to genetic evolution, it is expected that certain alleles will be more widespread in some populations than others, such as the well documented prevalence of a 9.3 allele at TH01 in the Caucasian population [221], or of a 2.2 allele at Penta D which has a frequency of over 11% in the West African population [98]. It was assumed that this would also hold true for sequence-based alleles, although it is only now possible to easily investigate this through the use of MPS. Gettings et al. [98] reported multiple examples of apparent population-specific enriched frequencies – where certain motif alleles had frequencies that were over 20% higher in one population compared to the others studied: D3S1358 in the African American population (TCTA TCTG [TCTA]$_n$), a specific 9 allele at D4S2408 in the Asian population (ATCT GTCT [ATCT]$_7$), and D16S539 in the Asian population ([GATA]$_n$ rs11642858). This phenomenon was observed during this work, where it also became apparent that certain alleles were seen only in one, or two more closely related, populations. Figure 4.10 shows the distribution of alleles at each locus according to how many alleles are seen in one, two, three, four or even all five populations. As expected, certain very common alleles are seen in all populations, such as allele 11 at TPOX which has a frequency of over 0.2 in the five population groups studied. At certain markers, there is a surprisingly high proportion of "population-specific" alleles, such as at D19S433 where over half the alleles observed are only seen in one population.

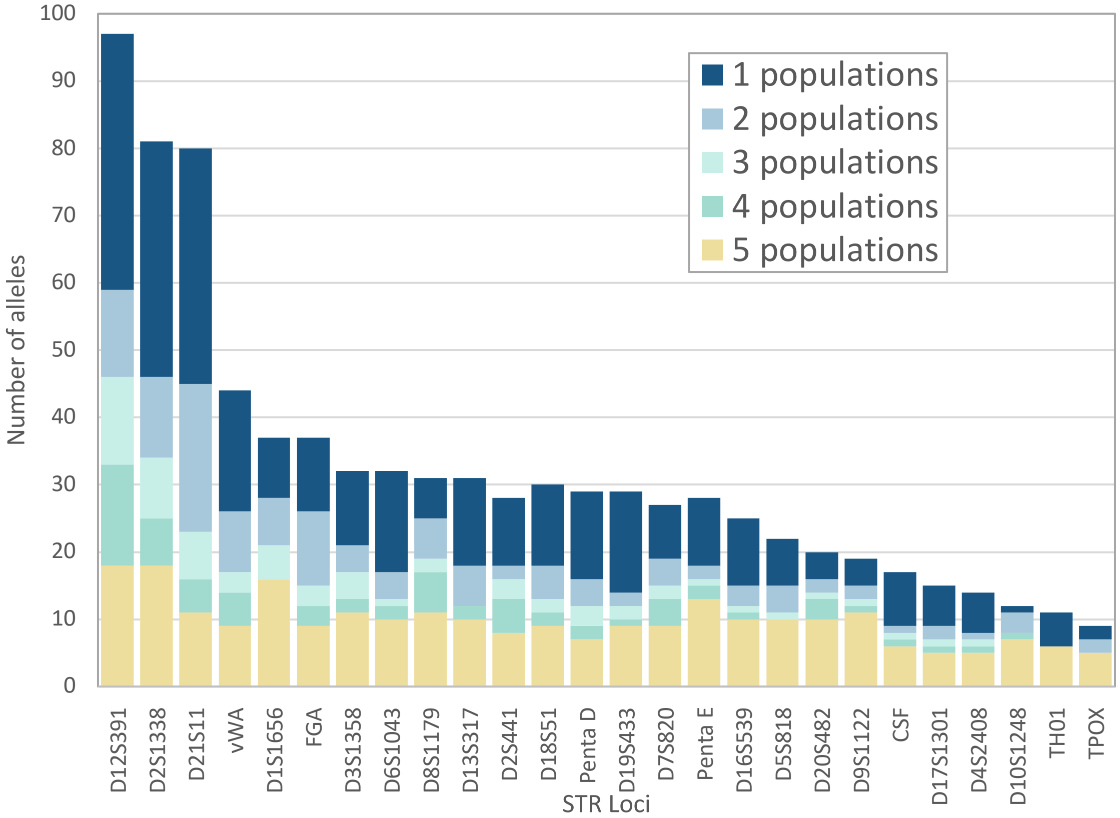


Figure 4.10: Number of discernible alleles at each locus, split by number of populations in which they have been recorded

The number of discernible alleles at each autosomal STR locus studied are split into 5 categories, from those observed in all five populations (yellow) to those seen in just one population (dark blue). This graph does not take into account the different characterisation of alleles (length-based or sequence-based), or their frequency within a population.

For certain rare alleles, the fact that they are only seen once mean it is hard to draw any meaningful conclusions regarding their population specificity. At CSF1PO for example, all variation observed where the sequence diverged from the traditional $[AGAT]_n$ motif was population specific, and only seen once or twice across the entire dataset, suggesting these could simply be one-off mutational events. The sequences were compared with alleles genotyped in the Caucasian, African-American and East Asian population samples published by NIST [98] and the University of North Texas (UNT) [63]. One allele, which was seen only once in a West African sample (allele 1202 $[AGAT]_8$ ACCT $[AGAT]_3$) was seen twice in the NIST African American population, and not in any other population. This suggests this sequence variant could be specific to the West African population, and once again demonstrates the utility of larger scale databases to properly capture all expected variation.

Figure 4.11 illustrates the frequencies for all population-specific alleles observed across the 5 populations. The thicker the band, the more common the allele in the population it was recorded in. Certain alleles immediately jump out as being pointedly more common than others, despite the fact that they are only seen in one population. The thick band going from D19S433 in the South Asian population corresponds to allele 1302, which has a frequency of 0.016 in this population (i.e. 6/378 alleles). Other notable common, population-specific alleles include one sequence-based allele at D5S818, observed at a frequency of 0.029 in the British Chinese population (11/396) and a sequence-based allele at D21S11, observed at a frequency of 0.037 in the West African population (n=16/404). The thick bands going from D18S51 to the two African populations correspond to the length-based .2 alleles at this marker. Neither Figure 4.10 nor Figure 4.11 take into account the fact that some of the population specific alleles are length-based ones, and so this breakdown is shown in Figure 4.12. This figure shows that the vast majority of population specific alleles are in fact repeat-region sequence variants, although some of the highest frequency alleles in this category are length-based alleles, such as an allele 5 by length at Penta D in the West African population (frequency of 0.067).

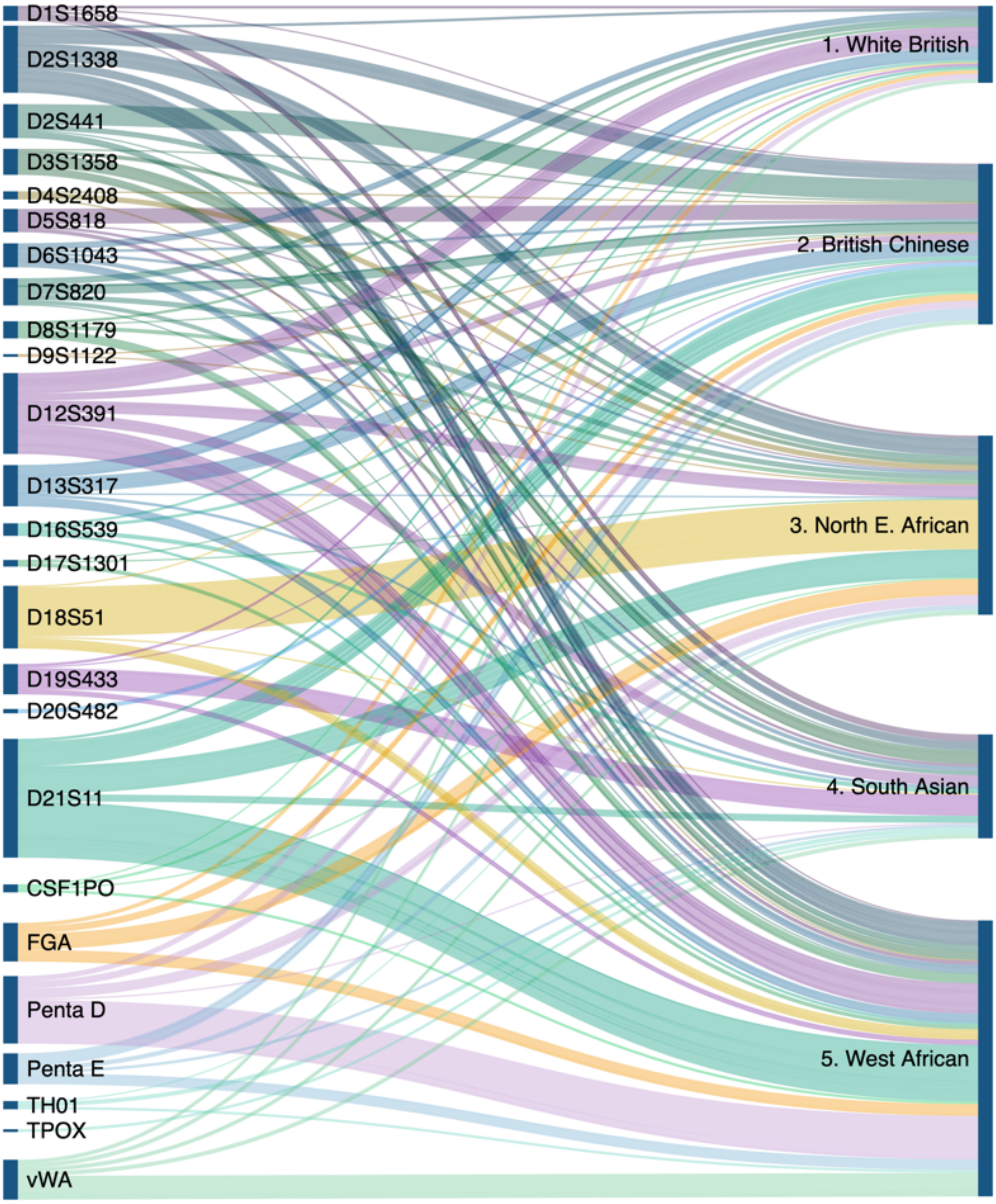


Figure 4.11: Sankey diagram for all population specific alleles

The frequency for each population specific allele (alleles seen in only one population) for the 26 autosomal STRs is represented by a single line. The thicker the line, the higher the frequency of the allele within the population. On the right-hand side of the graph, the corresponding population in which each allele is observed is provided.

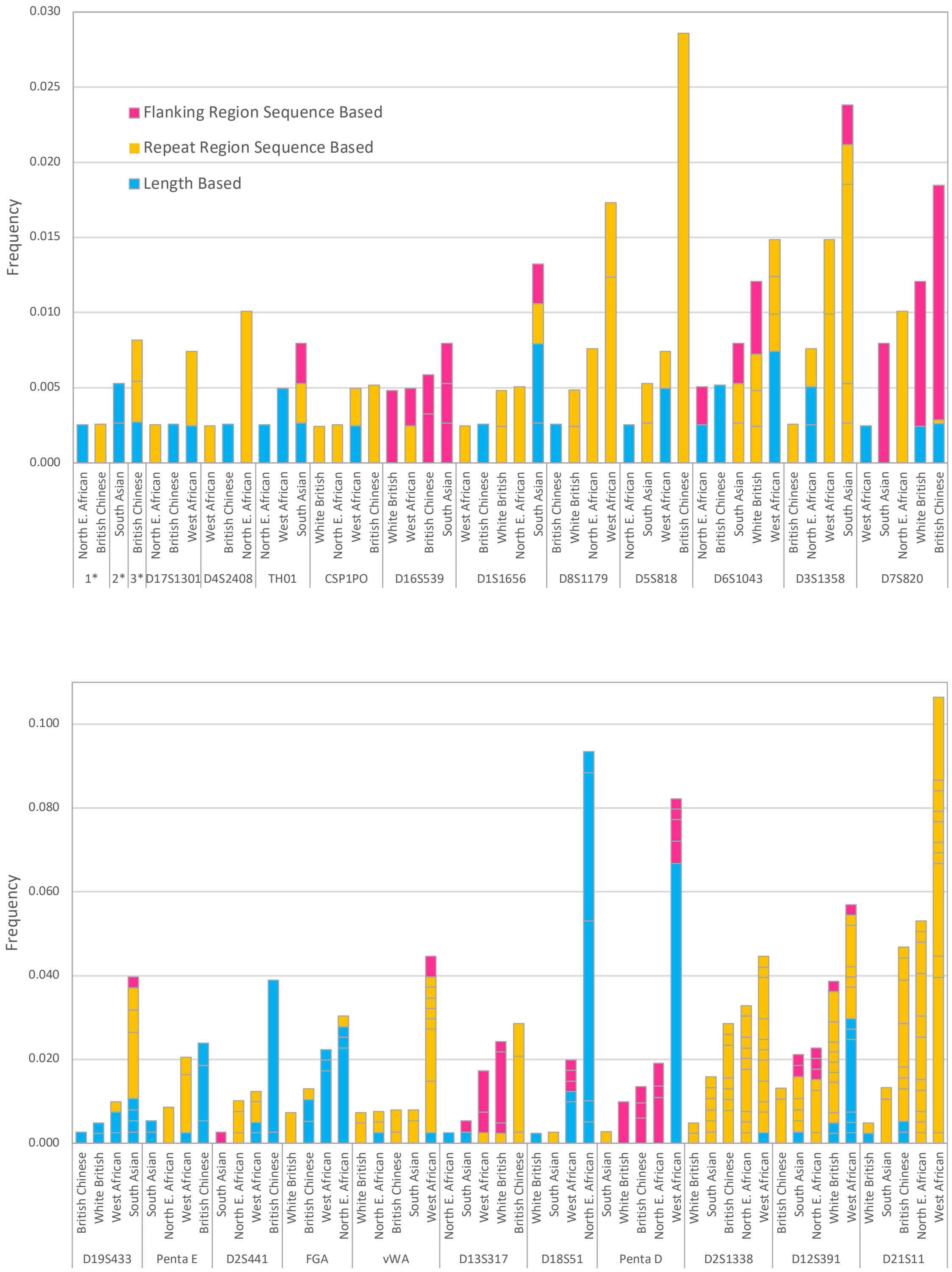


Figure 4.12: Frequency distribution for all population specific alleles

STRs are ordered from left to right, top to bottom, in order of increasing overall frequency of population specific alleles. Within each STR "bin", populations are provided from left to right in order of increasing overall frequency of population specific alleles. NB: For space, the following three markers had to be condensed: 1- D9S1122, *2- TPOX, *3- D20S482. D10S1248 was not included as it does not show any population-specific variation. Please note that the scale on the Y axis is different between the two panels.*

4.4. Discussion on population data

Population databases such as the one discussed in this chapter will form an intrinsic part of the implementation of massively parallel sequencing in forensics. The sequence-based allelic frequencies for all five population groups are now published, which enables equipped laboratories to use results obtained from MPS for DNA identification. Genotypes were verified and submitted to STRidER prior to publication in FSI: Genetics as a quality control method and have also contributed a large proportion of the initial STRSeq BioProject database for recording known sequence variants. This work complements other available sequence-based allelic frequency databases published by laboratories such as NIST, the University of North Texas, and the Institut National de Police Scientifique [98, 212, 222], but also fills a large data gap for the South Asian and North East African populations.

The highly polymorphic nature of certain STRs, such as D12S391, brings into question whether all "common" variants have been found, but this is a global picture which will grow over time and as more data is submitted. Aside from being one of the earliest and most comprehensive studies published on sequence-based frequencies, they are known to also have been used by the DNA Analysis at King's casework laboratory for live kinship relationship casework, as well as for further research outside of the scope of this PhD. Figure 4.13 shows an example of a real relationship testing case, where the two hypotheses were that the two individuals tested were either father-son or uncle-nephew. Using an in-house panel of 44 CE-STRs, the likelihood ratio indicated a value of 601 times in favour of the father-son hypothesis, which is an inconclusive result according to the internal guidelines set by DNA Analysis at King's. By analysing the samples from this case using the ForenSeq DNA Signature Prep kit, and including sequence-based allelic variation, the likelihood ratio rises to 175,644,600 more likely in favour of the father-son hypothesis, and a result could be reported.

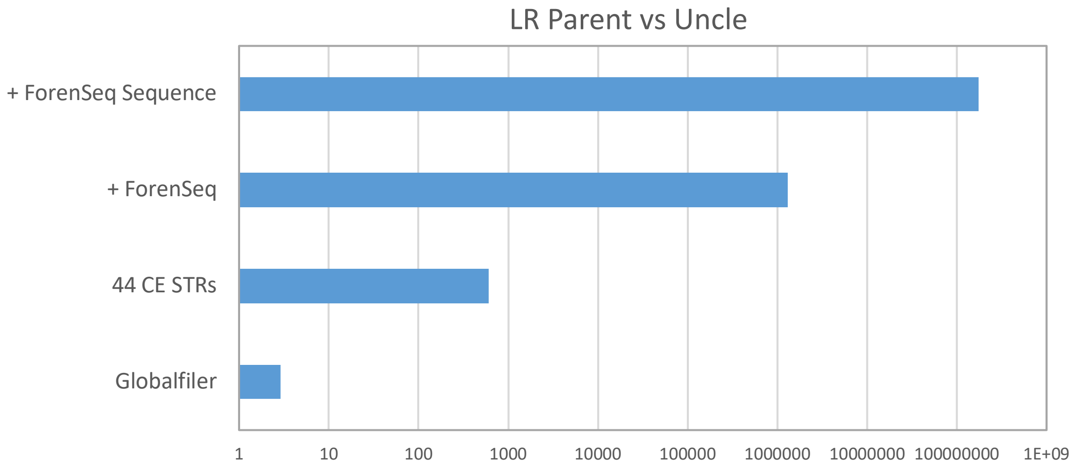


Figure 4.13: Parent vs avuncular relationship testing case

Graph showing likelihood ratio (LR) values obtained when testing for a parent vs uncle relationship in a real relationship testing case. The case was analysed using the markers in the GlobalFiler STR kit, as well as an in-house panel of 44 CE-STRs, and the ForenSeq DNA Signature Prep kit. For the latter kit, length-based allelic frequencies, and sequence-based allelic frequencies were applied. Data obtained from Dr. David Ballard, King's College London (personal communication).

The steps taken by the forensic community and guidelines to include flanking region sequences for all MPS-derived data came part-way through this project and involved re-analysing results as described in the Materials and Methods chapter, and discussed in the previous results chapter. Although there is certainly a necessity to include these regions for accurate allelic designation, results for marker diversity and match probability indicate quite clearly that flanking regions add limited value to the power of MPS for autosomal STR analysis. In order to make use of what diversity these regions do bring, and to minimise the complexity of analysis, significant strides will be needed in the context of sequence-based allelic nomenclature, and bio-informatic strategies. The most valuable approach would appear to be the clear delineation of "start" and "stop" points, otherwise described as a range of bases upstream and downstream of the repeat region of STRs. This would enable compatible nomenclature across different commercial kits, and a standard format for reporting alleles from a bio-informatic standpoint.

The presence of population specific alleles was an interesting observation to come out of this work and during the assembly of data. These will be explored further in the next chapter, with emphasis on whether this population segmented variation can be used in the context of ancestry determination.

5. ANCESTRY INFORMATIVENESS OF AUTOSOMAL STR AND ANCESTRY SNP MARKERS IN THE FORENSEQ DNA SIGNATURE PREP KIT

5.1. Sample selection

The same autosomal STR genotypes for 989 samples that were used to generate sequence-base allelic frequencies were used for this work, from the following five population groups: White British (n=207), British Chinese (n=193), North East African (n=198), South Asian (n=189) and West African (n=202). A subset of 47 samples from each population were selected at random for re-analysis for the ancestry informative SNPs in DNA Primer Mix B of the ForenSeq DNA Signature Prep kit. Following initial review using the Universal Analysis Software (UAS), samples with no drop out at any of the 56 ancestry-informative SNPs were taken forward for analysis for the White British (n=42), British Chinese (n=46), North East African (45), South Asian (n=39), and West African (n=47) populations.

5.2. Using autosomal STRs for population differentiation

The possibility of inferring bio-geographical ancestry using DNA markers with population differentiated variation provides an opportunity to verify eyewitness testimony, or in its absence, gain information about an unknown sample where no database match is obtained. Core autosomal STR loci, such as those included for analysis in kits for capillary electrophoresis or the ForenSeq DNA Signature Prep Kit, were chosen initially for a number of reasons, including their power of discrimination for individual identification. A number of studies have looked at the use of autosomal STRs for ancestry estimation, as presented in the introduction chapter of this thesis, but the overwhelming findings have been that a considerably larger number of non-core STR markers were needed for differentiation of global population groups. Londin et al. [140] assessed the ancestry estimation potential of the Identifiler kit but failed to differentiate a global sample set consisting of 7 populations. Phillips et al. were able to use the program STRUCTURE for ancestry assignment of the HGDP-CEPH panel into four global population groups (African, East Asian, American and European) using the 15 STR loci in Identifiler, with 5 additional extended-ESS STRs [170]. The addition of 34 ancestry-informative SNPs was required to adequately differentiate the 5th population group present in the subset of samples tested (Oceania). They also excluded the Middle East and Central-South Asian groups due to geographic proximity with the European group and consequential poor population differentiation.

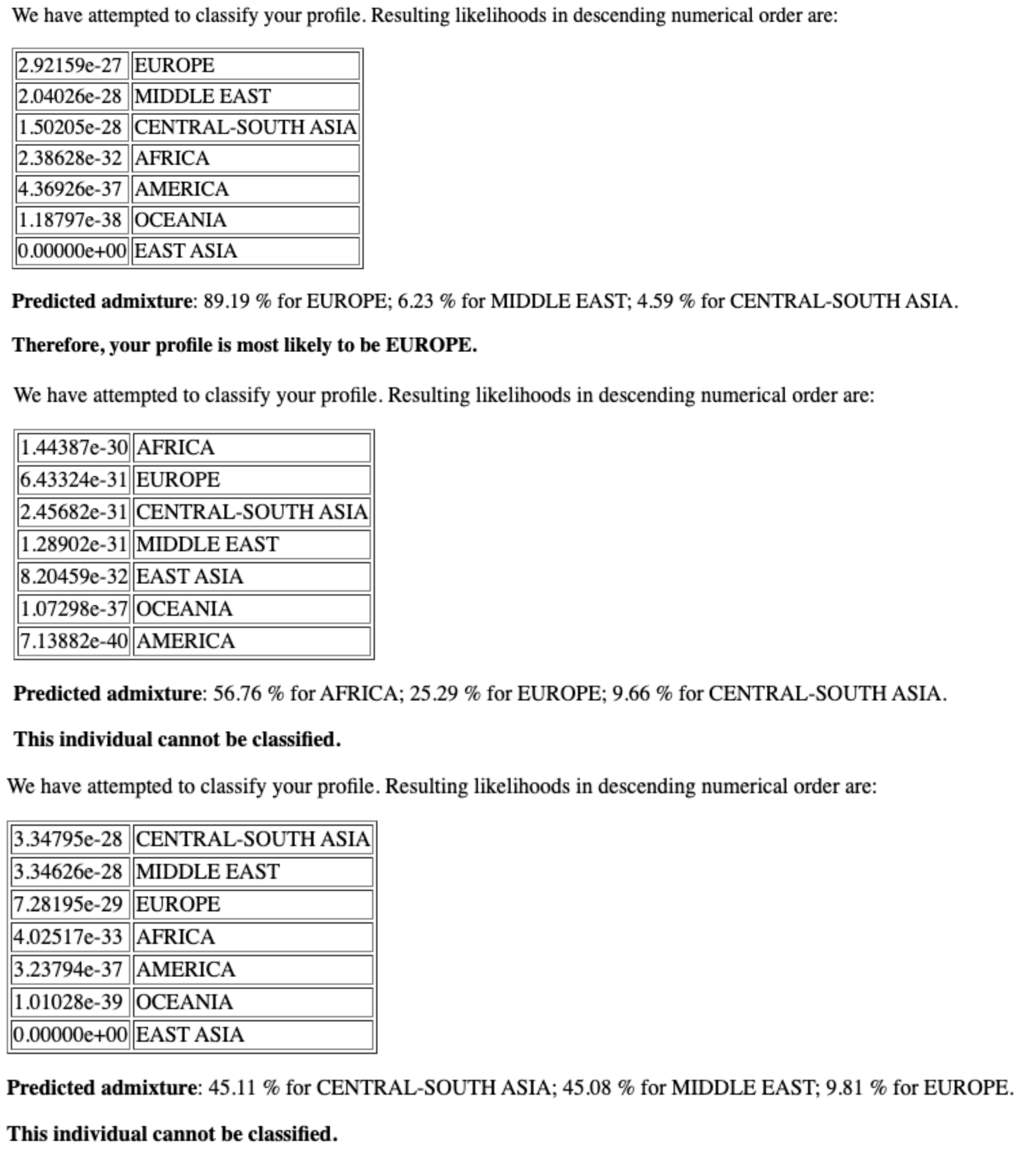
Although the authors of these papers, and others similar, present their results as not particularly promising for the future of autosomal STRs for ancestry estimation, Chris Phillips makes a key point in his 2015 review: “Despite these results, it is important to explore how effectively core STRs can infer ancestry as the data is generated in almost all forensic tests” [148]. The primary focus, in the majority of forensic identification cases, is to identify the individual to whom a DNA profile belongs. It is only once the possibility of a profile comparison or database search has yielded no results that investigations may turn to the use of bio-geographical estimation. The theoretical possibility of getting this information from a sample that has already been extracted, amplified, and analysed is an attractive one, due to the lack of additional sample, cost and time required.

As mentioned towards the end of the previous chapter, there are well known (and well documented) length-based alleles that are more prevalent in one population, such as the 9.3 allele at TH01 in the Caucasian population [221], or the 2.2 allele at Penta D which has a frequency of over 11% in the West African population [98]. There has been limited research on the presence of sequence-based population specific alleles due to the reduced number of population STR databases generated using massively parallel sequencing compared to capillary electrophoresis, but the results from this project suggest that there are in fact many sequence-based STR alleles whose frequencies are enriched in one population. Following on from this finding, this section will focus on the ancestry informativeness of the autosomal STRs in the ForenSeq DNA Signature Prep kit, and their power to distinguish ancestral populations.

5.2.1. Currently available tools for ancestry estimation from STRs

There are multiple tools available online which allow the upload of SNP genotypes for ancestry estimation, including the program FROG-kb which will be discussed further in section 5.3.2. Because STRs are not usually used for ancestry inference, there are limited options for genotype uploads from these markers. In recent years, *Snipper*, an online

Bayesian classifier established to analyse SNP data [223] was modified to accommodate STR profiles [148], meaning length-based autosomal STR genotypes can now be uploaded and used to classify the profile against a frequency-based training set. This training set contains the HGDP-CEPH panel frequencies for 32 autosomal STR markers for the following re-classified populations: Oceania, America, Europe, East Asia, Central-South Asia, Africa, Middle East, evaluated by Phillips et al. [224]. The results for three samples chosen at random from the White British, North East African and South Asian populations are shown in Figure 5.1. The software was able to predict the White British sample's European ancestry but struggled more with the other two population groups.



We have attempted to classify your profile. Resulting likelihoods in descending numerical order are:

2.92159e-27	EUROPE
2.04026e-28	MIDDLE EAST
1.50205e-28	CENTRAL-SOUTH ASIA
2.38628e-32	AFRICA
4.36926e-37	AMERICA
1.18797e-38	OCEANIA
0.00000e+00	EAST ASIA

Predicted admixture: 89.19 % for EUROPE; 6.23 % for MIDDLE EAST; 4.59 % for CENTRAL-SOUTH ASIA.

Therefore, your profile is most likely to be EUROPE.

We have attempted to classify your profile. Resulting likelihoods in descending numerical order are:

1.44387e-30	AFRICA
6.43324e-31	EUROPE
2.45682e-31	CENTRAL-SOUTH ASIA
1.28902e-31	MIDDLE EAST
8.20459e-32	EAST ASIA
1.07298e-37	OCEANIA
7.13882e-40	AMERICA

Predicted admixture: 56.76 % for AFRICA; 25.29 % for EUROPE; 9.66 % for CENTRAL-SOUTH ASIA.

This individual cannot be classified.

We have attempted to classify your profile. Resulting likelihoods in descending numerical order are:

3.34795e-28	CENTRAL-SOUTH ASIA
3.34626e-28	MIDDLE EAST
7.28195e-29	EUROPE
4.02517e-33	AFRICA
3.23794e-37	AMERICA
1.01028e-39	OCEANIA
0.00000e+00	EAST ASIA

Predicted admixture: 45.11 % for CENTRAL-SOUTH ASIA; 45.08 % for MIDDLE EAST; 9.81 % for EUROPE.

This individual cannot be classified.

Figure 5.1: Screenshot of results for a White British sample (top), a North East African sample (middle), and a South Asian sample (bottom) using *Snipper*

Another option available for ancestry estimation from STR data is PopAfiliator [173], which is a free online tool for evaluating population assignment of an individual’s 17 locus STR profile. The website states that accuracy of individual population affiliation assignment to three population groups (Asia, Eurasia, sub-Saharan Africa) is approximately 90%. This decreases to 65% when five population groups are considered, due to the addition of two groups that are more genetically similar to the Eurasian group than the initial set (North Africa and Near East). Length-based allelic genotypes for the same samples tested with *Snipper* were input into Popafiliator on a number of occasions but always returned an error message suggesting the website was not working, and so no results were obtained. Given the lack of South Asian samples in the database and under representation of African samples (1.42% sub-Saharan African and 1.43% North African), it is unlikely that this tool would have showed improved results compared to *Snipper.* The fact that both websites only allow the upload of length-based allelic data generally meant they would not be useful in the context of this research.

5.2.2. Population differentiation using STRUCTURE

To investigate whether autosomal STR data could be used to differentiate the five global populations studied in this work, genotypes for the 26 autosomal markers discussed in the previous two chapters were run through STRUCTURE (27 STRs in the ForenSeq DNA Signature Prep kit minus D22S1045). This program was used to discern genetic clusters based on individuals’ similarity or dissimilarity to others within the sample set, following the method described in Chapter 2. An initial aim was to check whether sequence-based alleles, particularly those including flanking region sequences, might be more useful for ancestry inference than length-based allelic data alone. Moriot et al. stipulate that haplotypes composed of slow and fast-evolving loci might combine the advantages of identity and ancestry-informative marker types [152]. The instability of STR markers has led to a large divergence in number of alleles in a population over time, whereas flanking region SNPs or insertions/ deletions should be more stable through the course of evolution, possibly allowing for greater conservation within populations.

Analyses were run for all samples (n=989) using data for length-based alleles (Figure 5.2), sequence-based without flanking regions alleles (Figure 5.3) and sequence-based with flanking regions included alleles (Figure 5.4). In these figures, each vertical line represents one sample, and the colour composition of that line reflects the proportion of membership for each calculated genetic cluster. Colour assignments correspond to the population group with the largest membership in that cluster. Samples are grouped together by self-declared ancestry in the diagrams for simplicity, with the five populations separated by black lines. The K value for each STRUCTURE analysis refers to the number and patterns of genetic clusters found, and is user defined. STRUCTURE plots were run using K=2 to K=5 to see how many groups (/populations) the program could separate by genetic stratification of the data. K=6 was also run, to ensure no additional substructure was being picked up, which would possibly indicate sub-populations. Although STRUCTURE is primarily a way of clustering individual samples rather than classifying them, looking at the number of samples which cluster incorrectly may provide a measure of the STR set for ancestry inference.

The differences between STRUCTURE results obtained using length-based and sequence-based (RR and FR) data are discussed in the following sections in the context of the number of clusters differentiated, proportion of membership and incorrect cluster assignment.

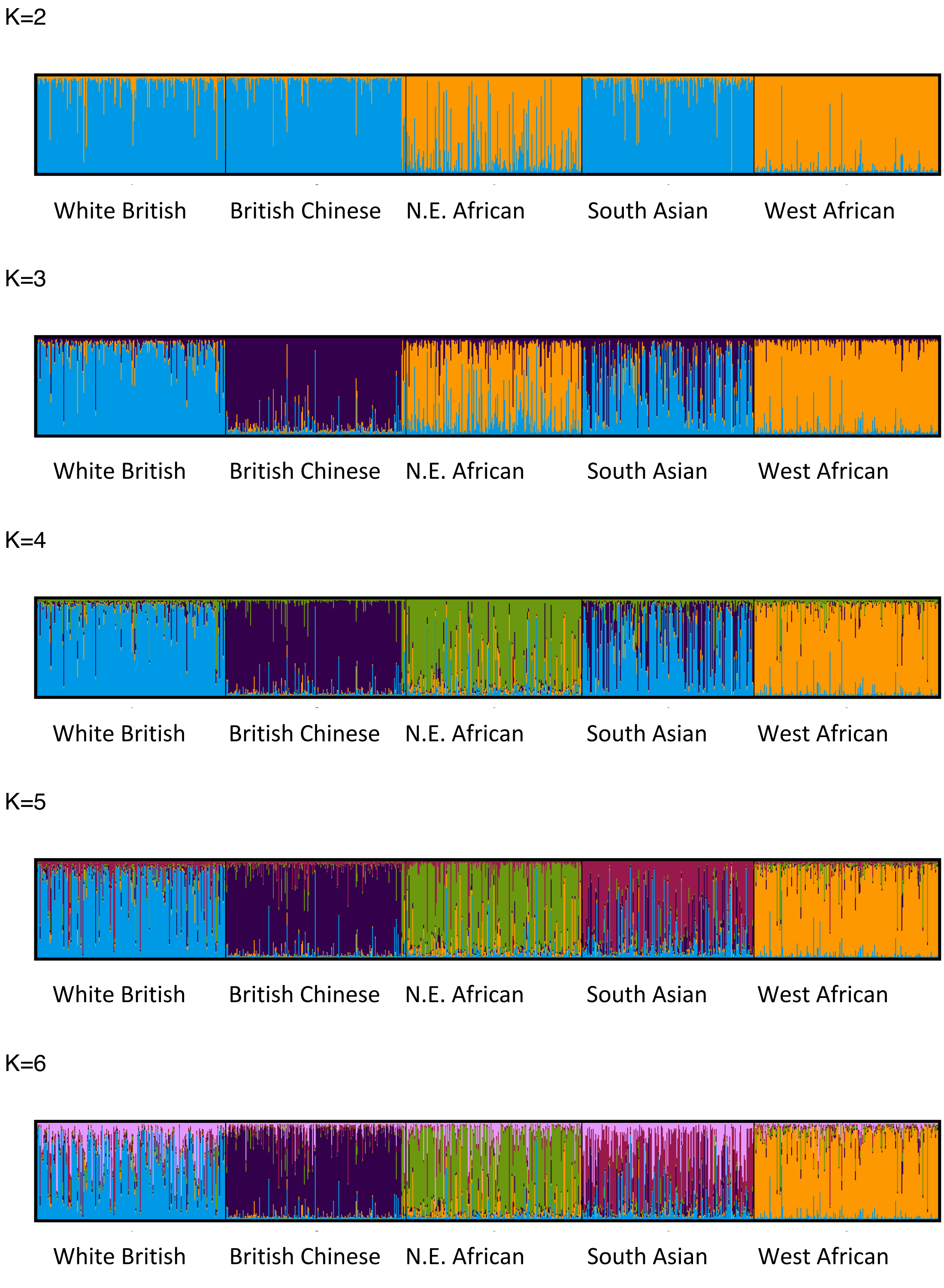


Figure 5.2: STRUCTURE plots for the five populations, generated using length-based allelic data for 26 autosomal STRs in the ForenSeq DNA Signature Prep Kit

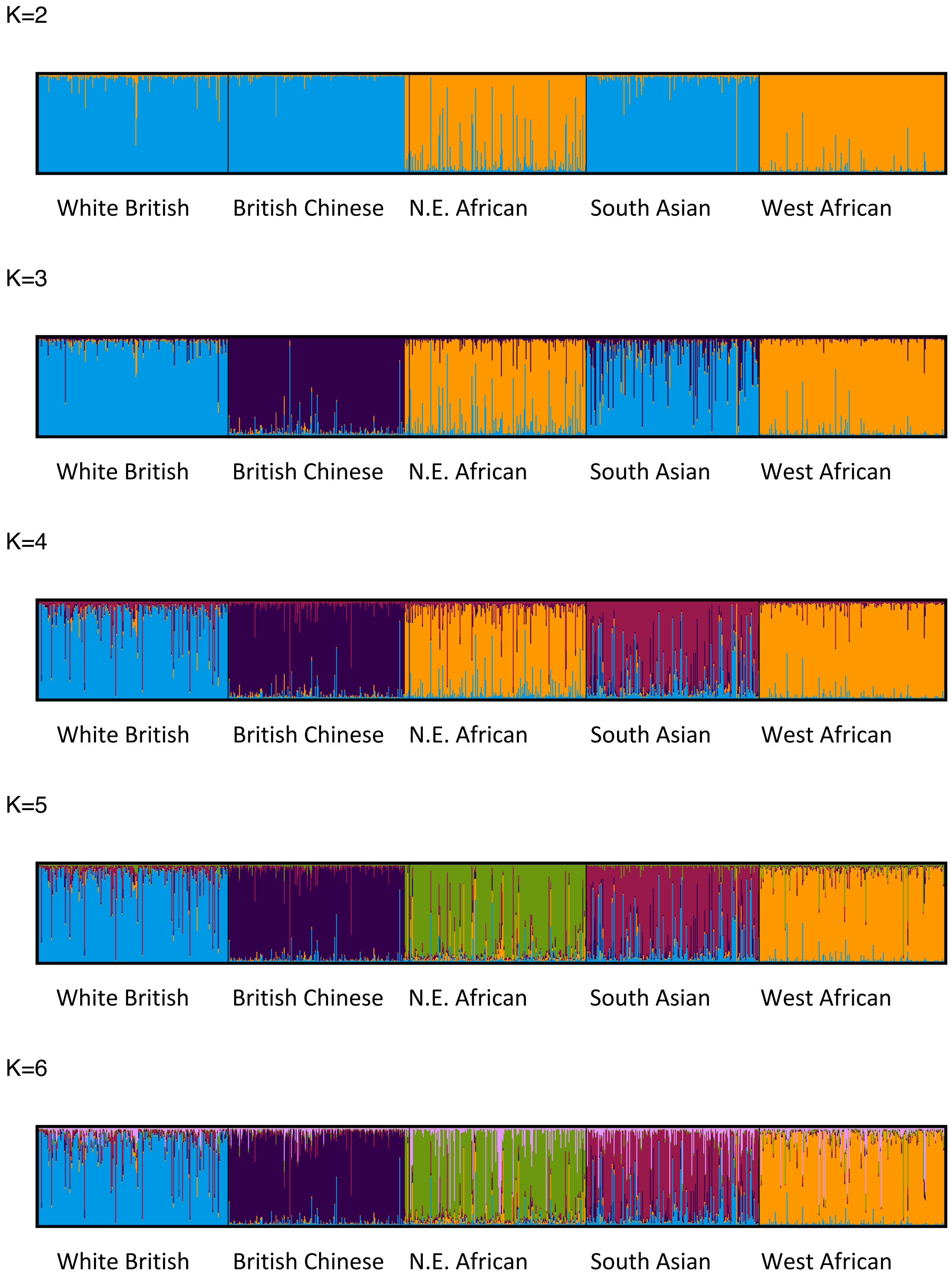


Figure 5.3: STRUCTURE plots for the five populations, generated using repeat region sequence-based allelic data for 26 autosomal STRs in the ForenSeq DNA Signature Prep Kit

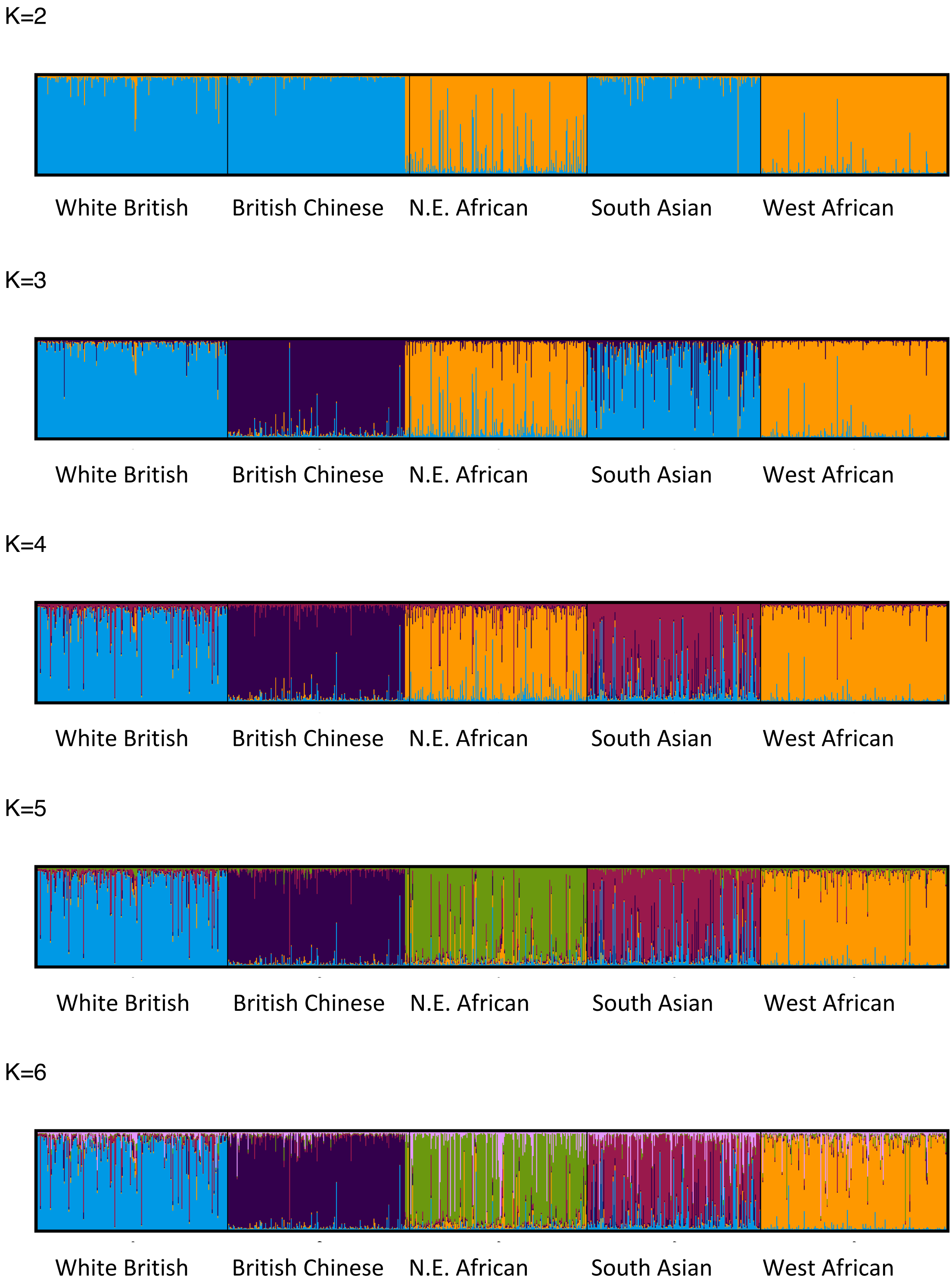


Figure 5.4: STRUCTURE plots for the five populations, generated using sequence-based allelic data which includes flanking region sequences for 26 autosomal STRs in the ForenSeq DNA Signature Prep Kit

5.2.2.1. K number of clusters

When applying K=2, it is clear that the software is able to distinctly separate the African populations from the others, highlighting the genetic dissimilarity between African and non-African ancestral populations. STRUCTURE defines genetic clusters without prior knowledge of population affiliation, hence the fact that the African populations has been successfully separated confirms that the STRs provide good African: Non-African differentiation. The next cluster to be distinguished with K=3 is the British Chinese. With K=4, the trend differs between the length-based and sequence-based plots, with the former separating a cluster for the North East African group whilst the two sequence-based plots differentiate the South Asian group first. Whether using length-based or full sequence-based allelic data (including flanking regions), STRUCTURE appears to be able to distinguish 5 (i.e., K=5) distinct genetic clusters, which correspond to the five ancestral population groups. In order to confirm which value of K best represented the data in a more objective way, results from STRUCTURE were uploaded to STRUCTURE Harvester [225]. One of the plots produced by this program provides an indication of how likely each K value tested is, as shown in Figure 5.5 and Figure 5.6 for the length-based and full sequence-based allelic data.

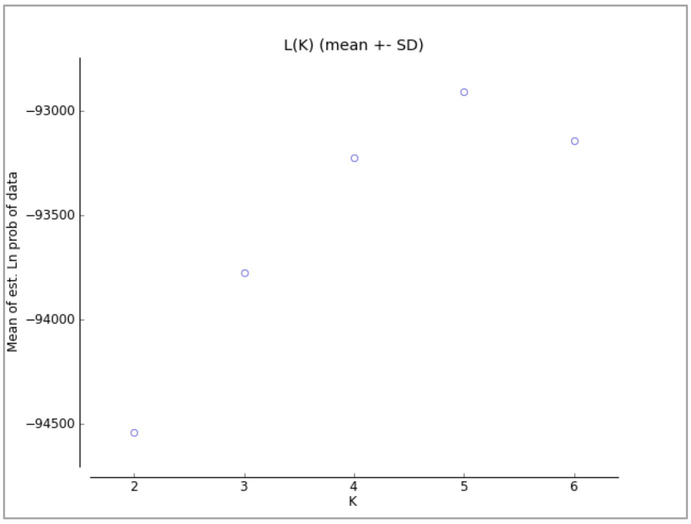


Figure 5.5: STRUCTURE Harvester plot showing the mean likelihood of K for length-based allelic data

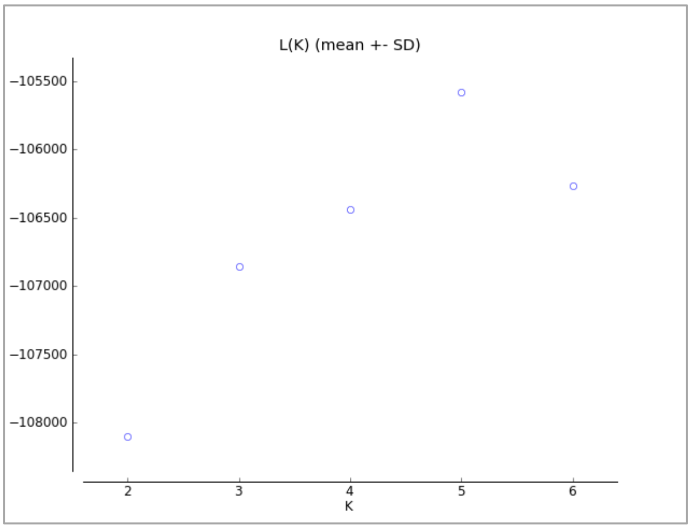


Figure 5.6: STRUCTURE Harvester plot showing the mean likelihood of K for sequence-based data (including flanking regions)

The results from STRUCTURE Harvester confirm what could be seen the STRUCTURE plots for K=2 to K=6, which is that a partition of the data into 5 genetic clusters is the most likely scenario. From here on, STRUCTURE data focusses on plots generated using a k=5 assumption.

5.2.2.2. Proportion of membership

For all STRUCTURE analyses run with K=5, the average proportion of membership of each pre-defined population in each of the 5 clusters was extracted from the results file and collated in Table 5.1. This provides the average proportion of membership coefficient for the samples in each population group (listed in the first column) assigned to each of the five clustered inferred with K=5. A proportion of membership of 1 would indicate 100% assignment to one cluster. For example, the samples in the White British group have an average proportion of membership of 0.73 to cluster 1 in the STRUCTURE plot run with length-based allelic data, whereas this goes up to 0.83 when taking full sequence-based allelic data into account. Overall, the strongest average proportion of membership for each group (highest value assigned to one cluster) for all groups occur when including flanking region variation, apart from the North East African group where membership of population drops by 0.002 when going from repeat region alleles to flanking region data. These results can be better visualised in Figure 5.7 .

Table 5.1: Proportion of membership for each of the 5 clusters inferred with K=5

STRUCTURE analyses were run for length-based (LB), sequence-based including just repeat region (SB-RR) and sequence-based including flanking regions (SB-FR). The proportion of membership for the correct cluster is highlighted in colours matching the colours used for earlier STRUCTURE plots.

		Inferred Clusters				
Given Population	**STRUCTURE Run**	**1**	**2**	**3**	**4**	**5**
White British	LB	0.733	0.05	0.047	0.146	0.024
	SB-RR	0.823	0.032	0.019	0.111	0.015
	SB-FR	0.831	0.028	0.019	0.106	0.016
British Chinese	LB	0.048	0.822	0.049	0.059	0.022
	SB-RR	0.034	0.876	0.033	0.042	0.015
	SB-FR	0.03	0.887	0.032	0.036	0.016
North East African	LB	0.064	0.038	0.708	0.087	0.103
	SB-RR	0.044	0.024	0.799	0.06	0.073
	SB-FR	0.041	0.024	0.797	0.063	0.075
South Asian	LB	0.173	0.153	0.037	0.608	0.029
	SB-RR	0.162	0.094	0.023	0.699	0.021
	SB-FR	0.154	0.09	0.022	0.713	0.021
West African	LB	0.037	0.027	0.067	0.035	0.834
	SB-RR	0.024	0.016	0.053	0.021	0.886
	SB-FR	0.024	0.015	0.046	0.021	0.895

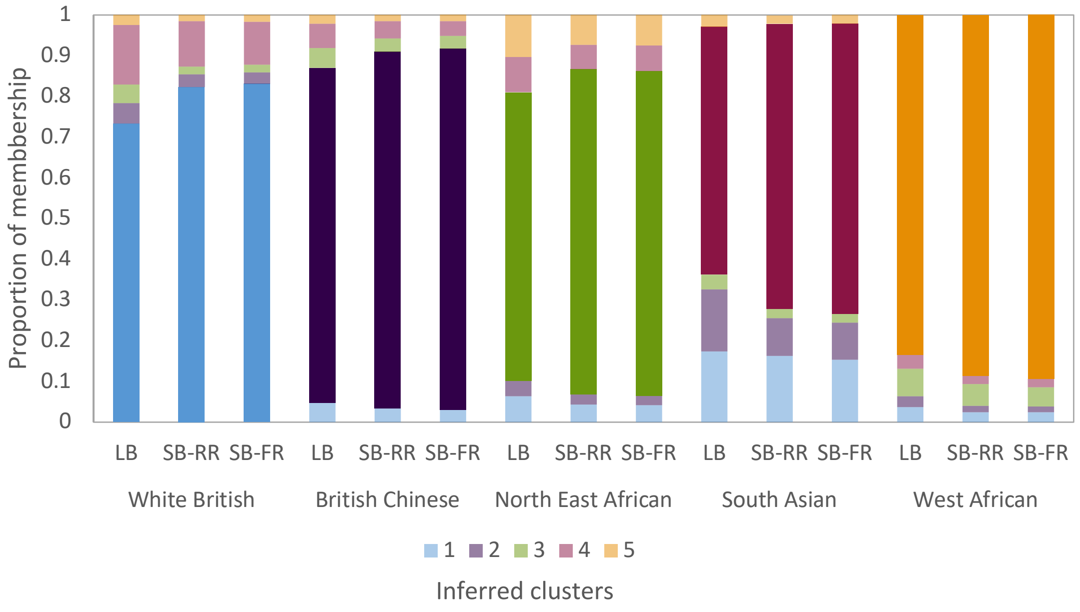


Figure 5.7: Proportion of membership of each cluster for each of the 5 pre-defined populations

The main inferred cluster for each pre-defined population group is colour coded, with all other cluster assignments shown in a more transparent colour.

5.2.2.3. Incorrect cluster assignment

Although promising in terms of population differentiation, a number of samples in each population group are still being assigned incorrect cluster membership. Phillips et al. measured the relative ability of a forensic STR set to differentiate ancestries by measuring group misclassification, defined as the number of samples with less than 0.5 group membership proportion for their true population of origin [170]. In the White British population, 18 samples out of 207 did not achieve a proportion of membership higher than 0.5 for the correct cluster (sequence data including flanking regions). A similar phenomenon was seen in the British Chinese (4/193), North East African (27/198), South Asian (37/189) and West African (8/202) population groups. These numbers, and the more general assignment results presented in Figure 5.7 and Table 5.1 above also highlight the fact that certain populations are considerably easier to genetically separate from the others, namely the West African and British Chinese. Although some of these samples don't show a proportion of membership coefficient higher than 0.5 for any of the groups, 67 samples in total appear to show assignment to the wrong cluster and are shown in Table 5.2. This will be investigated further in the context of ancestry estimation, but is already an important improvement over the fact that with the length-based data, 185 samples do not have a proportion of membership higher than 0.5 for the correct cluster, with 108 of these having a coefficient of over 0.5 for the wrong cluster.

These values effectively correspond to the samples represented by the "wrong" colour in the STRUCTURE plots seen earlier. One example of this is the White British sample WB85 which shows a proportion of membership of over 0.95 to the cluster associated with the South Asian samples. Figure 5.8 shows a zoomed in view of the STRUCTURE plot for the British Chinese (flanking region sequences, K=5) and displays quite clearly the two samples which have a high proportion of membership for the incorrect clusters that are in Table 5.2. The vast majority of samples are assigned to one cluster associated with the British Chinese population group, represented in dark purple. One sample is represented as a mostly burgundy line, sample BC275 from the table below, which has a coefficient of 0.859 for the cluster associated with South Asian samples. Another sample had a coefficient of 0.716 for the cluster associated with White British samples and is visualised as a blue line on the plot.

Table 5.2: Samples with a proportion of membership of > 0.5 for the incorrect cluster

For each sample, the value in black corresponds to the proportion of assignment for the correct cluster, and the value in red corresponds to the highest proportion of assignment. All other values are shown in grey. WB= White British, BC= British Chinese, NEA= North East African, SA= South Asian, WA= West African.

Sample	**Population**	**1. WA**	**2. WB**	**3. NEA**	**4. SA**	**5. BC**
WB4	WB	0.004	**0.271**	0.007	**0.680**	0.038
WB15	WB	0.021	**0.235**	0.016	**0.719**	0.009
WB35	WB	0.021	**0.118**	0.007	**0.830**	0.024
WB51	WB	0.006	**0.069**	0.006	**0.874**	0.045
WB85	WB	0.008	**0.019**	0.009	**0.957**	0.007
WB114	WB	0.011	**0.045**	0.020	**0.916**	0.008
WB148	WB	0.063	**0.244**	0.054	**0.601**	0.038
WB154	WB	0.012	**0.291**	0.022	**0.601**	0.074
WB158	WB	0.007	**0.339**	0.087	**0.560**	0.006
WB166	WB	0.014	**0.063**	0.018	**0.886**	0.018
WB191	WB	0.005	**0.170**	0.005	**0.790**	0.030
WB193	WB	0.008	**0.376**	0.007	**0.585**	0.025
WB197	WB	0.026	**0.143**	0.132	0.183	**0.516**
BC275	BC	0.021	0.065	0.035	**0.859**	**0.020**
BC395	BC	0.006	**0.716**	0.047	0.019	**0.213**
NEA404	NEA	**0.910**	0.026	0.011	0.041	**0.012**
NEA409	NEA	**0.635**	0.037	**0.025**	0.133	0.170
NEA429	NEA	0.007	0.284	**0.061**	**0.628**	0.020
NEA438	NEA	0.167	0.131	**0.134**	**0.557**	0.012
NEA439	NEA	0.024	0.019	**0.372**	**0.512**	0.073
NEA447	NEA	0.038	0.037	**0.057**	**0.852**	0.016
NEA465	NEA	**0.800**	0.018	**0.173**	0.004	0.005
NEA474	NEA	0.289	0.032	**0.006**	**0.663**	0.010
NEA475	NEA	0.232	0.171	**0.165**	0.360	0.072
NEA477	NEA	**0.851**	0.016	**0.039**	0.019	0.076
NEA478	NEA	0.197	**0.698**	**0.020**	0.043	0.042
NEA496	NEA	0.123	**0.599**	**0.022**	0.060	0.196
NEA508	NEA	**0.873**	0.011	**0.085**	0.018	0.013
NEA519	NEA	0.074	0.025	**0.015**	**0.826**	0.060
NEA525	NEA	**0.537**	0.160	**0.101**	0.049	0.153
NEA558	NEA	0.023	0.047	**0.117**	**0.743**	0.069
NEA576	NEA	0.077	0.019	**0.061**	**0.838**	0.005
NEA580	NEA	**0.628**	0.162	**0.075**	0.105	0.029
SA606	SA	0.013	**0.768**	0.018	**0.132**	0.069
SA609	SA	0.006	0.019	0.010	**0.370**	**0.594**
SA613	SA	0.012	**0.770**	0.009	**0.179**	0.030
SA615	SA	0.009	**0.875**	0.006	**0.088**	0.022
SA617	SA	0.102	0.164	0.042	**0.091**	**0.601**
SA627	SA	0.052	**0.583**	0.034	**0.318**	0.013
SA629	SA	0.078	**0.747**	0.031	**0.073**	0.071
SA639	SA	0.015	**0.655**	0.008	**0.313**	0.009
SA647	SA	0.005	**0.519**	0.011	**0.396**	0.068
SA651	SA	0.005	0.041	0.005	**0.275**	**0.675**
SA672	SA	0.018	0.086	0.009	**0.343**	**0.544**
SA678	SA	0.008	**0.878**	0.010	**0.093**	0.011
SA693	SA	0.007	**0.820**	0.009	**0.150**	0.014
SA702	SA	0.005	**0.529**	0.007	**0.142**	0.317
SA703	SA	0.009	**0.855**	0.012	**0.110**	0.015
SA716	SA	0.013	0.010	0.036	**0.387**	**0.554**
SA720	SA	0.019	0.143	0.022	**0.285**	**0.531**
SA733	SA	0.024	0.081	0.024	**0.094**	**0.777**
SA736	SA	0.014	0.013	0.031	**0.013**	**0.928**
SA744	SA	0.020	**0.704**	0.177	**0.030**	0.068
SA746	SA	0.014	**0.887**	0.009	**0.069**	0.022
SA758	SA	0.016	**0.888**	0.007	**0.077**	0.012
SA760	SA	0.013	**0.801**	0.012	**0.143**	0.031
SA762	SA	0.008	**0.519**	0.125	**0.339**	0.010
SA763	SA	**0.967**	0.004	0.016	**0.008**	0.005
SA768	SA	0.184	0.397	0.037	**0.092**	0.290
SA769	SA	0.014	0.123	0.036	**0.261**	**0.565**
SA774	SA	0.004	**0.834**	0.004	**0.147**	0.012
SA778	SA	0.018	**0.733**	0.107	**0.114**	0.028
SA779	SA	0.012	**0.840**	0.030	**0.081**	0.037
WA816	WA	**0.296**	0.014	**0.674**	0.011	0.004
WA871	WA	**0.374**	0.071	0.009	**0.537**	0.010
WA945	WA	**0.037**	0.010	**0.922**	0.024	0.008
WA950	WA	**0.066**	0.138	**0.628**	0.058	0.110

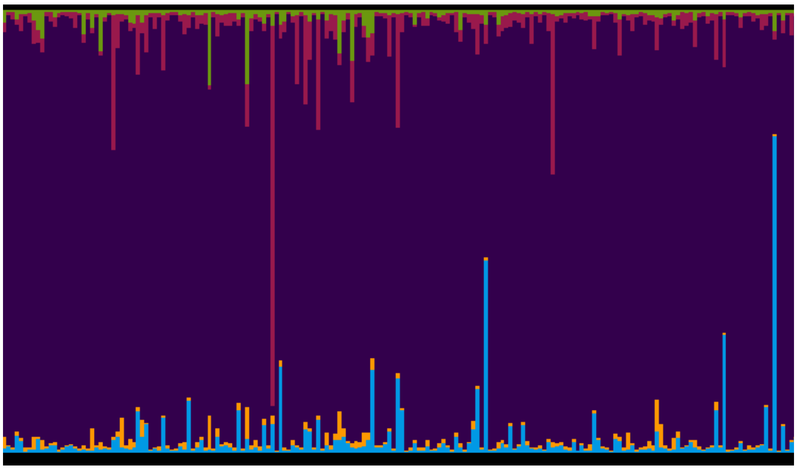


Figure 5.8: Enlarged view of the British Chinese population group in STRUCTURE

Zoomed in view of the STRUCTURE plot for the British Chinese (taken from Figure 5.4 flanking region sequences, K=5).

A more realistic indication of the rate of incorrect assignment would be to use a more conservative proportion of membership coefficient as the cut-off. If a proportion of membership to the correct cluster of over 0.7 is taken as a "correct" population assignment, with anything below classifying as "inconclusive", 84% of samples (832 out of 989) are assigned to the correct group. This is a noteworthy improvement to the length-based data and a very slight improvement over the repeat-region sequence-based data, where 72% (719 out of 989) and 83% (818 out of 989) of samples are assigned correctly, respectively. The number of samples which are assigned to the incorrect group with over 0.7 proportion of membership is inversely related, with 56 samples being assigned the wrong group for length-based data, 38 samples with the repeat-region sequence-based data, and 35 samples being assigned incorrectly for the full flanking region sequence-based data. This is equivalent to a general error rate of 3.5% for this STR set's ability to assign correct group membership.

Of the three samples tested with *Snipper*, the two which the software was unable to classify were also specifically checked. The North East African sample which had a predicted admixture of 57% African and 25% European has a proportion of membership for the North East African cluster of 0.94 with the sequence-based (FR) STRUCTURE run. This sample already has a proportion of membership of 0.91 with length-based data alone, suggesting perhaps that the loci targeted by the ForenSeq DNA Signature Prep kit for which frequencies are not available for the *Snipper* estimation are useful for this population. More likely, the poor classification by *Snipper* is due to the lack of North East African samples included in the African population training set (n=103), where most samples are in fact West African. The second sample was a South Asian sample with a predicted admixture of 45% for both the Central-South Asian and European populations. This sample has a proportion of membership of over 0.5 for the incorrect cluster (White British) in the length-based data STRUCTURE run, but this changes to 0.59 for the correct cluster (South Asian) in the sequence-based (FR) run, suggested the improved ancestry inference when taking sequence variation into account.

With 84% of samples assigned to the correct cluster according to the five pre-determined population groups, and over 90% of samples assigned for the correct cluster with a proportion of 0.5, it is fair to state that there is ancestry-informative data within the autosomal STR results for this dataset. The following section will look at improving the percentage of samples that are correctly assigned, as well as reducing the number of samples that are assigned to an incorrect population.

5.2.3. Ancestry informativeness, I_n

It has been stipulated that the use of highly informative markers can reduce the amount of genotyping required for ancestry inference, as using markers with the highest level of informativeness can reduce the number of overall markers needed [198]. The number of markers analysed in this dataset form a commercial panel, and are therefore always amplified together, so it is not a case of needing to reduce the number of markers. However, the use of highly uninformative markers (expected given the loci type, autosomal STRs), could be adding noise to the STRUCTURE plots presented above. In order to assess the ancestry informativeness of the autosomal STR loci studied, and therefore identify which contribute more meaningfully to the population clustering in the STRUCTURE plots, the Informativeness for Assignment value, I_n, was calculated for each locus. The I_n metric is highly correlated to the fixation index (F_{ST}), which is often used to measure population differentiation of SNPs [148], but is better suited for multi-locus data according to Rosenberg et al. [198]. Table 5.3 shows the I_n values for all loci when comparing the five population groups at once. The table highlights the top 6, 12 and 24 most informative markers ahead of further analysis. This table shows that the ancestry informativeness of all markers increase when taking sequence variation into account. Xu et al. state that I_n values of over 0.2 can be considered as “the signal of very great genetic difference between populations."[226], and the results presented here show that no length-based locus data gave an I_n of above 0.2.

Table 5.3: I_n values for all markers

I_n values are given for 26 loci, across the 5 population groups studied, using length-based (LB), sequence-based repeat-region (SB-RR) and sequence-based flanking-region (SB-FR) data.

	5 pops LB		5 pops RR		5 pops FR	
	Locus	I_n	Locus	I_n	Locus	I_n
6 most informative	Penta D	0.168	D21S11	0.320	D21S11	0.320
	Penta E	0.163	D2S1338	0.300	D2S1338	0.300
	D1S1656	0.147	D12S391	0.248	D12S391	0.261
	D18S51	0.145	D1S1656	0.220	D1S1656	0.220
	TH01	0.117	D13S317	0.203	D13S317	0.216
	D2S441	0.113	vWA	0.199	vWA	0.199
12 most informative	D13S317	0.110	D3S1358	0.184	Penta D	0.195
	D2S1338	0.103	Penta D	0.173	D3S1358	0.185
	D19S433	0.100	Penta E	0.173	D2S441	0.177
	D6S1043	0.092	D8S1179	0.161	Penta E	0.173
	FGA	0.091	D2S441	0.156	D8S1179	0.161
	D12S391	0.090	D18S51	0.147	D18S51	0.149
24 most informative	D21S11	0.087	D4S2408	0.145	D4S2408	0.145
	D4S2408	0.080	D5S818	0.125	D5S818	0.125
	TPOX	0.079	TH01	0.117	TH01	0.118
	D5S818	0.078	D19S433	0.111	D7S820	0.116
	vWA	0.074	D6S1043	0.109	D6S1043	0.112
	D8S1179	0.074	FGA	0.107	D19S433	0.112
	D10S1248	0.061	D9S1122	0.088	FGA	0.107
	CSF1PO	0.061	TPOX	0.079	D9S1122	0.088
	D7S820	0.051	CSF1PO	0.065	D16S539	0.084
	D3S1358	0.034	D10S1248	0.061	TPOX	0.079
	D16S539	0.030	D7S820	0.058	CSF1PO	0.065
	D20S482	0.030	D17S1301	0.034	D10S1248	0.061
All 26 markers	D17S1301	0.029	D20S482	0.032	D20S482	0.051
	D9S1122	0.028	D16S539	0.031	D17S1301	0.034

Penta D is the marker with most ancestry informativeness when only accounting for length-based data, which is unsurprising given the known prevalence of specific length variants such as the 2.2 and 3.2 alleles in the African populations. Interestingly, these are not true length variants and are in fact caused by a 13 base pair deletion in the flanking region. This makes these alleles particular in the sense that they combine a slowly mutating marker (the deletion) and a faster mutating one (the STR), which had been suggested to be beneficial in the context of population-specific allelic enrichment

[152]. When adding sequence-based data, Penta D is far outstripped in terms of ancestry informativeness by other markers. The marker showing the most pronounced increase in I_n is D21S11, which is also one of the marker showing the highest level of sequence variation as discussed in Chapter 3, and highest expected heterozygosity as discussed in Chapter 4. There is little difference when adding the flanking regions in, although the markers showing the most increase in I_n do correlate with those showing the most variation in the flanking regions: D7S820, D16S539, D20S482 and Penta D. The improvement in I_n at D7S820 due to the addition of flanking region data even pushes it out of the 6 least informative markers in the set.

Despite the limited improvement in ancestry informativeness of the markers when including flanking region information compared to repeat region alone, the decision was made to use the full sequence, including these regions, going forward for analysis. Given the fact that all sequences were characterised in the previous chapters, and there is no additional time or cost consideration, there would be no reason to not use the full spectrum of data available.

The 3 most informative markers for differentiating the five population groups using full sequence-based allelic data are D21S11, D2S338 and D12S391, which corresponds to the 3 markers showing the highest frequency of population specific alleles discussed in the previous chapter. Figure 5.9 shows the allelic frequency distribution across the different populations for two loci on opposite ends of the I_n spectrum. The frequencies for D17S1301 clearly show that there is very little divergence in distribution across the different populations, supporting the result that it has the lowest ancestry informative coefficient. D1S1656 was chosen as an example of very informative marker for ease of visualisation as it has a relatively high I_n value for length-based as well as sequence-based allelic data. These results clearly indicate that divergence in loci informativeness is dictated by contrasting allelic frequencies between populations

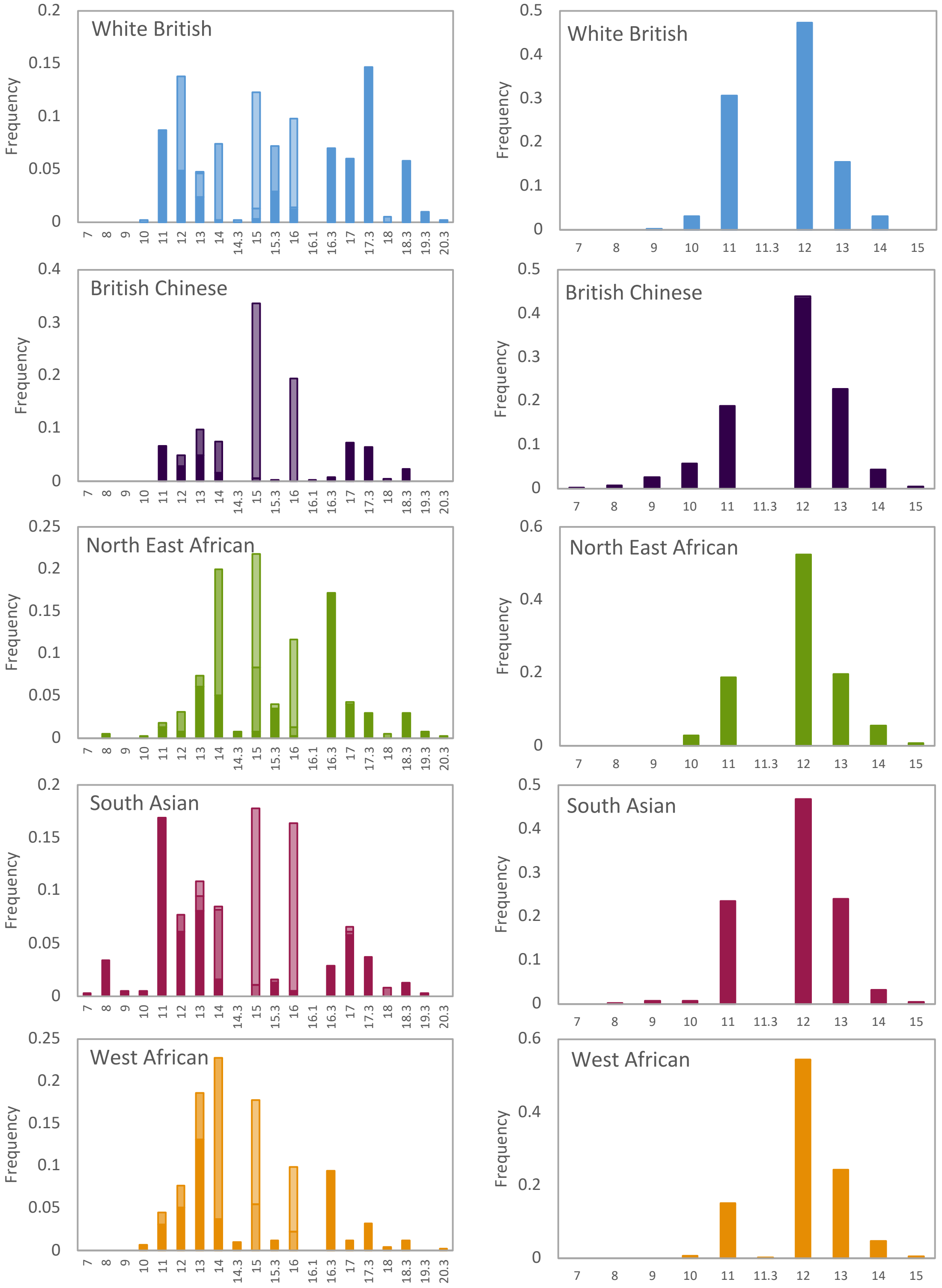
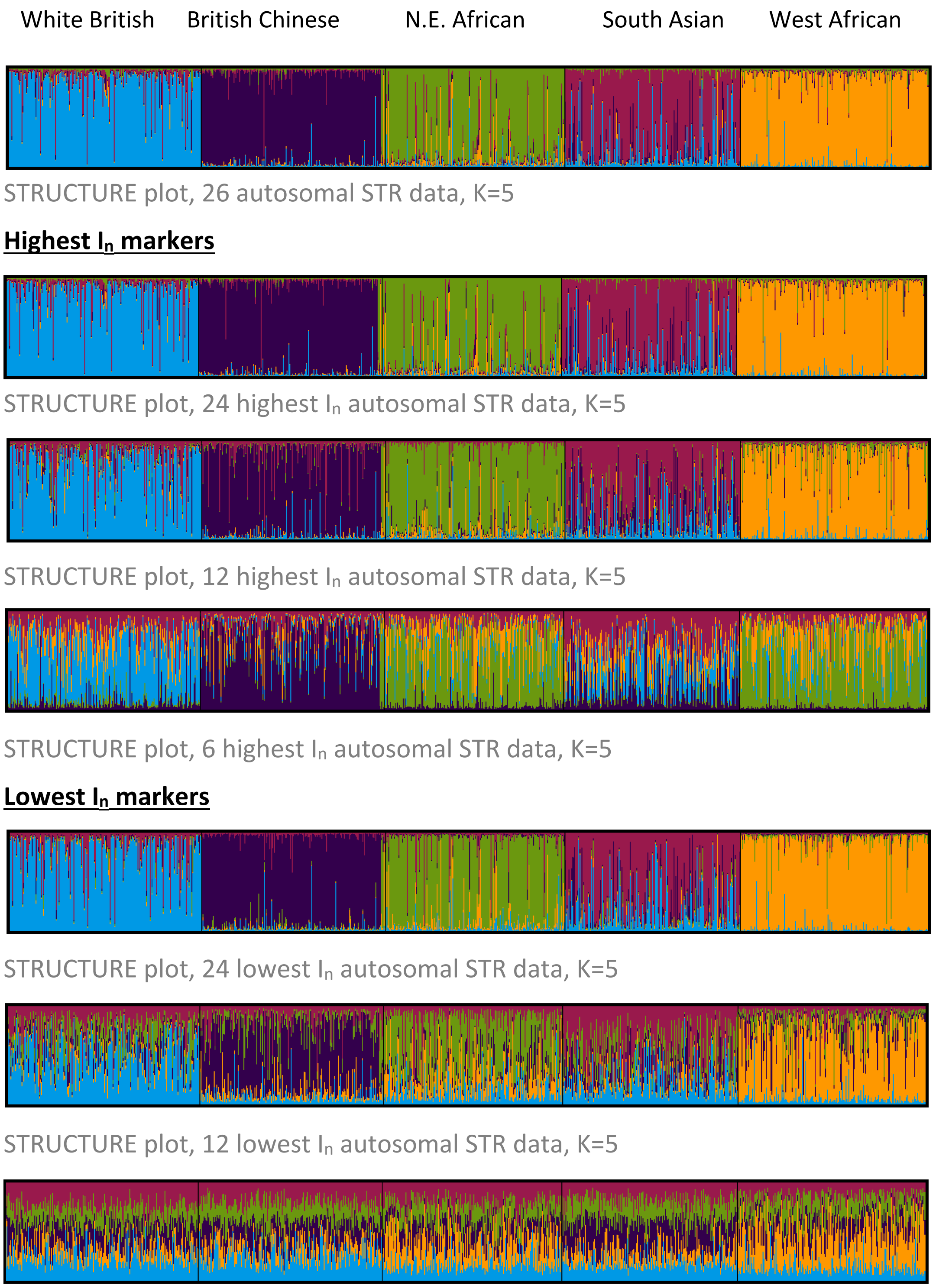


Figure 5.9: Autosomal STR frequency distribution for D1S1656 and D17S1301

Allelic frequency distribution for D1S1S656 (left panels) and D17S1301 (right panels) in all five populations. The X axis for each graph provides the length-based alleles, and frequencies are divided in sequence-based alleles where appropriate using shading.

5.2.3.1. Loci of most and least informativeness

Figure 5.10 shows the STRUCTURE plots run using data for the 6, 12 and 24 most informative STR markers (highest I_n), as well as for the 6, 12 and 24 least informative STR markers (lowest I_n). From looking at these plots, there does not appear to be much difference between the 24 most and 24 least informative markers, or with the plot generated using data for all 26 markers. The difference is markedly more apparent between the top and bottom 6 informative markers, which is to be expected.



STRUCTURE plot, 26 autosomal STR data, K=5

Highest I_n markers

STRUCTURE plot, 24 highest I_n autosomal STR data, K=5

STRUCTURE plot, 12 highest I_n autosomal STR data, K=5

STRUCTURE plot, 6 highest I_n autosomal STR data, K=5

Lowest I_n markers

STRUCTURE plot, 24 lowest I_n autosomal STR data, K=5

STRUCTURE plot, 12 lowest I_n autosomal STR data, K=5

STRUCTURE plot, 6 lowest I_n autosomal STR data, K=5

Figure 5.10: STRUCTURE plots for highest and lowest I_n markers

These STRUCTURE plots were constructed using data for all five populations and autosomal STR genotypes for the highest 6, 12 and 24 most ancestry informative markers based on I_n, as well as the lowest 2, 12 and 24 least ancestry informative markers.

The average proportion of membership, generated by STRUCTURE for each population group with varying number of markers, is shown in Figure 5.11. Although only the STRUCTURE plots for the top 6, 12 and 24 most informative loci are shown above, they were run for more combinations of markers when it became apparent that the “ideal” number of markers could be somewhere between 12 and 24. With the data for the 6 most informative markers, the average proportion of membership for the White British samples is of just over 0.5 for the cluster associated with the White British population. This jumps to over 0.7 when using the 12 most informative loci, and over 0.8 when using data from the top 22-26 loci. Interestingly, Figure 5.11 also shows that it is virtually impossible to distinguish the North East African from the West African clusters with the top 6 most informative markers, as 50% of the West African samples effectively have a proportion of membership to the same cluster as the majority of the North East African samples (in green). Overall, the figure suggests that there is little difference in terms of proportion of membership between using data for the 22 most informative markers and all 26 markers.

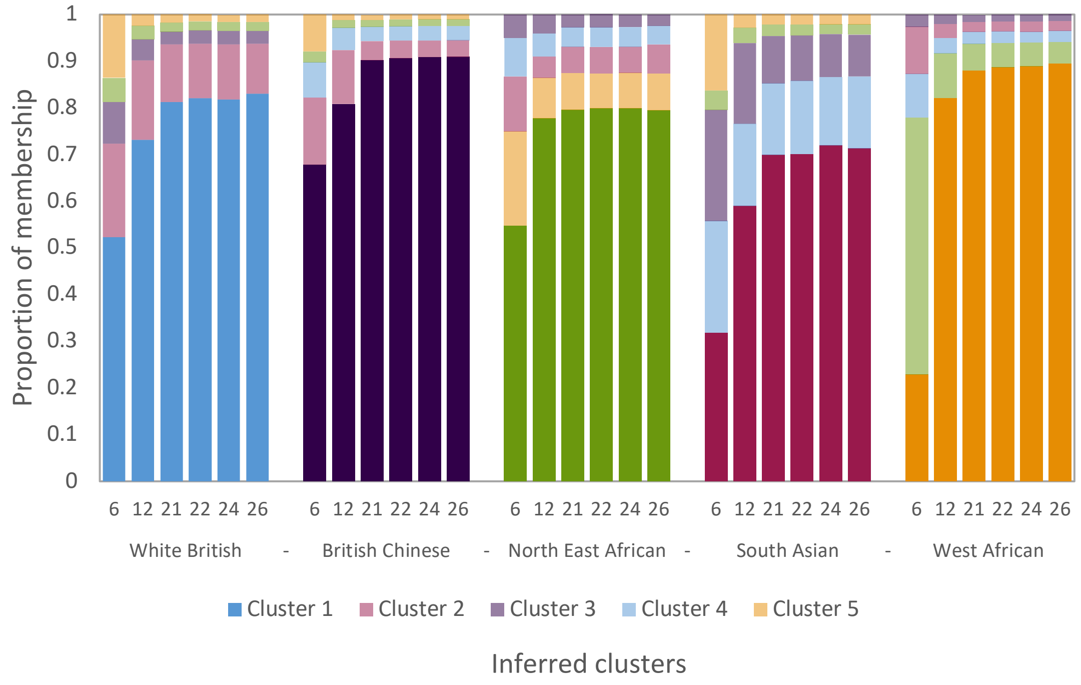
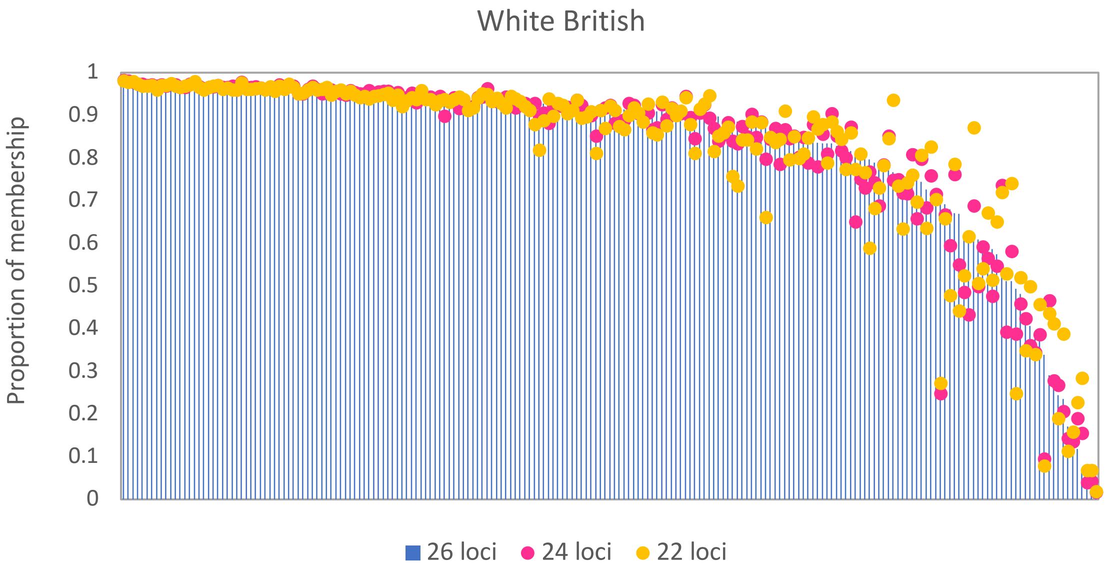


Figure 5.11: Proportion of membership obtained using top 6, 12, 21, 22, 24 most informative loci, as well as all 26 loci

The main inferred cluster for each pre-defined population group is colour coded, with all other cluster assignments shown in a more transparent colour.

The White British samples have the highest proportion of membership to the correct cluster when using data for all 26 loci (0.83). This is also the case for the West African samples and British Chinese samples, which have an average proportion of membership of 0.90 and 0.91 when using all 26 loci, respectively. The North East African and South Asian samples appear to have the highest proportion of membership when using 24 loci (0.80 and 0.72, respectively) – suggesting that removing data for the two markers with the lowest I_n, D20S482 and D17S1301, helps push the assignment in the right direction for these two groups. Using 22 loci led to the same results as 24 loci for the North East African cluster, but when decreasing the number of loci used beyond that, the proportion of membership decreased for all population groups.

Figure 5.12 shows the proportion of membership to the correct cluster for all White British and North East African samples when using data for all 26 loci compared to that of the top 22 and 24 most informative loci. These graphs show that removing 2 or 4 loci does appears to improve ancestry inference for some of the samples in the North East African population, but overall does not show a substantial improvement over using the data from all 26 loci targeted, as expected from the average proportion of membership plots. One sample in the White British population group stands out, which has a proportion of membership of 0.7 when using results from 26 loci, which goes down to below 0.3 when using results for the top 22 or 24 loci. For the latter two STRUCTURE runs, this sample appears to have a proportion of membership exceeding 0.7 for the South Asian population. The reason for this occurrence is unknown, although all alleles seen at the two markers removed from the 24 loci run for that sample did have alleles that are more common in the White British population than the South Asian one.



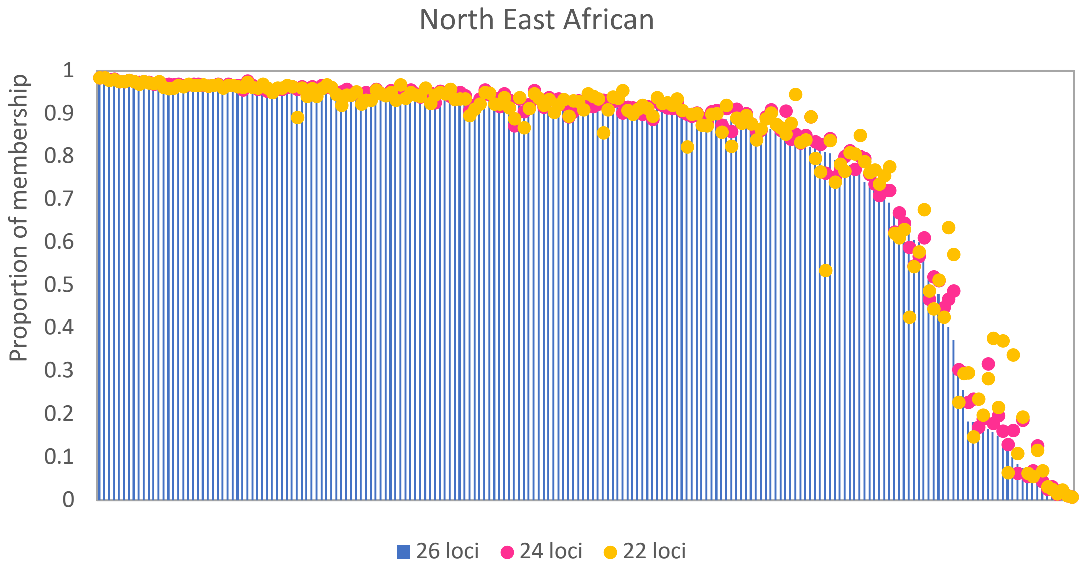


Figure 5.12: Proportion of membership of White British and North East African samples to the correct cluster using differing number of loci of most informativeness

Each line represents a White British (top) or North East African (bottom) sample, which are ordered in decreasing proportion of membership for the correct cluster using data from all 26 loci (blue bars). Pink and yellow scatter dots are used to represent proportion of membership for the correct cluster when using the 24 and 22 loci of highest informativeness, respectively.

When looking at the initial STRUCTURE run generated using data for the 26 autosomal STR loci, there were 67 samples that had a proportion of membership of above 0.5 for the incorrect cluster (Table 5.2). Of these, 35 were assigned to the incorrect cluster with over 0.7 proportion of membership. The assignment for these 35 samples in the STRUCTURE runs generated using the top 22 and 24 most informative loci was plotted against the results for the 26 loci run and is shown in Figure 5.13.

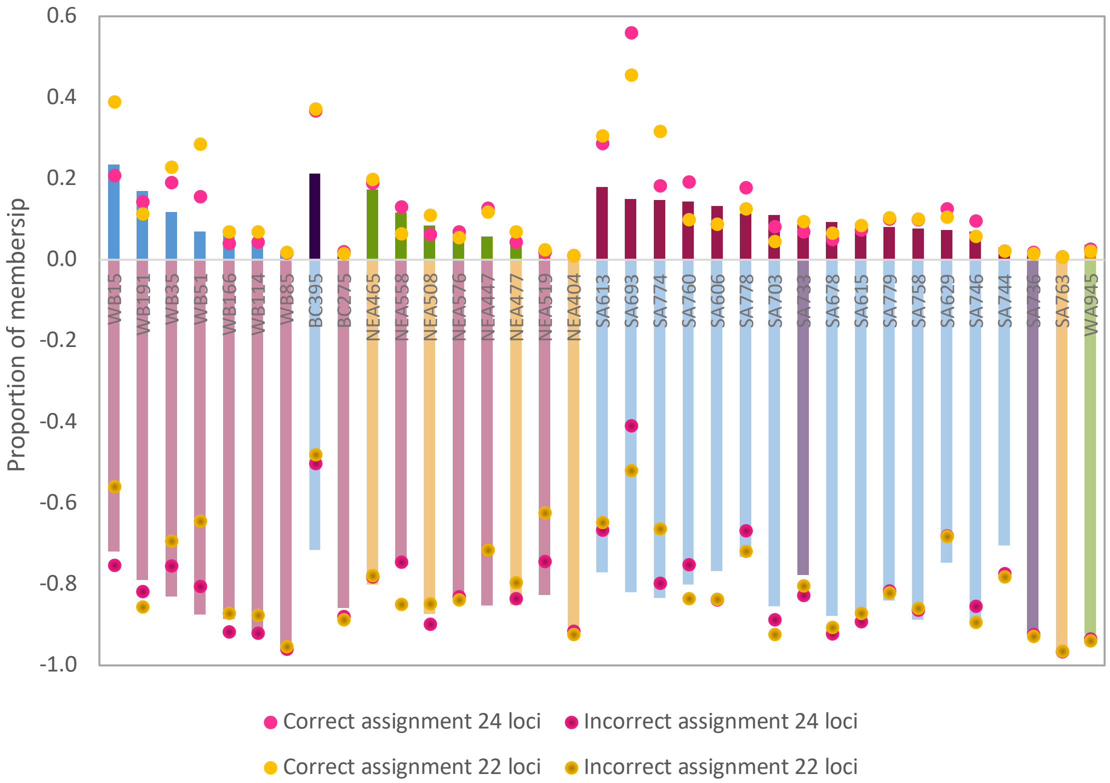


Figure 5.13: Proportion of membership for incorrectly assigned samples

This graph shows the proportion of membership for 35 samples which were assigned to the incorrect cluster in the 26 loci STRUCTURE run. Bars above 0 on the Y axis represent proportion of membership for the correct population cluster, whilst anything below 0 indicates proportion of membership for the incorrect population. Dots denote the proportion of membership for these samples for the same two populations when looking at the 24 and 22 loci STRUCTURE runs (pink and yellow, respectively). Colour coding is consistent with cluster colouring in previous graphs (White British: blue; British Chinese: dark purple; North East African: dark green; South Asian: plum and West African: orange)

Reducing the number of loci appears to improve ancestry inference for some of these samples. Notably, the proportion of assignment for sample SA693 to the correct, South Asian cluster goes from 0.15 to 0.56 when looking at the 24 loci run and inversely, the proportion of membership for the White British cluster falls below 0.5. None of the proportion of membership values for the correct cluster change enough to reach 0.7 however, making the results for these sample inconclusive. Using data for the top 24 and 22 most informative loci leads to the number of these particular samples assigned to the incorrect cluster dropping to 30 and 26, respectively. Whilst this seems promising initially, across the whole dataset, some samples which had previously not been incorrectly classified now have a proportion of membership of over 0.7 for the wrong cluster, as previously mentioned for the White British sample example. The total number of incorrectly classified samples therefore only drops to 34 for the 24 STR analysis and 32 for the 22 STR analysis, showing once again the limited value of reducing marker numbers.

5.2.3.2. Pairwise comparisons

The results presented so far indicate that certain populations are harder to differentiate than others. In particular, the White British and South Asian populations have the greatest number of samples misclassifying between the two clusters. I_n values were calculated between all pairs of populations to see if any of the loci score highly at ancestry informativeness between a pair of populations that does not score highly when comparing all five populations simultaneously. Table 5.4 shows the I_n values for all markers for all pairs of populations.

Table 5.4: I_n values for pairs of populations

Ancestry informativeness coefficient was calculated for the 26 autosomal STRs studied between pairs of populations. The colour coding was kept as in Table 5.3, with the markers in darker green corresponding to high I_n values when looking at all five groups.

White British & British Chinese		White British & North East African		White British & South Asian		White British & West African		British Chinese & North East African	
Locus	I_n	Locus	I_n	Locus	I_n	Locus	I_n	Locus	I_n
D13S317	0.188	D21S11	0.191	D2S1338	0.118	D21S11	0.220	D21S11	0.247
TH01	0.146	D2S1338	0.188	D1S1656	0.105	Penta D	0.217	D2S1338	0.229
D12S391	0.144	D12S391	0.167	D12S391	0.078	D12S391	0.186	D13S317	0.186
D1S1656	0.132	D8S1179	0.150	D16S539	0.067	D3S1358	0.171	D1S1656	0.172
D21S11	0.131	D1S1656	0.140	D21S11	0.065	D2S1338	0.165	D12S391	0.164
Penta E	0.127	vWA	0.122	D8S1179	0.060	D8S1179	0.164	D4S2408	0.163
D2S441	0.123	Penta D	0.107	Penta E	0.057	D1S1656	0.149	D2S441	0.159
D2S1338	0.117	D3S1358	0.093	D19S433	0.050	vWA	0.135	vWA	0.153
D7S820	0.112	D18S51	0.089	D13S317	0.046	D2S441	0.121	D5S818	0.122
D6S1043	0.091	FGA	0.086	D2S441	0.045	D19S433	0.119	D18S51	0.116
D8S1179	0.090	D4S2408	0.081	D18S51	0.037	Penta E	0.113	D3S1358	0.114
vWA	0.088	D2S441	0.080	TH01	0.033	D13S317	0.101	D7S820	0.113
D4S2408	0.073	D5S818	0.079	vWA	0.033	TPOX	0.095	Penta D	0.096
D16S539	0.066	Penta E	0.067	D3S1358	0.032	TH01	0.092	Penta E	0.089
D3S1358	0.064	TH01	0.062	FGA	0.029	D6S1043	0.092	D8S1179	0.089
D5S818	0.059	D6S1043	0.048	Penta D	0.027	D4S2408	0.092	D9S1122	0.085
D19S433	0.057	TPOX	0.047	D7S820	0.024	D18S51	0.087	D16S539	0.065
FGA	0.049	CSF1PO	0.041	D20S482	0.024	FGA	0.087	D6S1043	0.062
Penta D	0.049	D9S1122	0.039	D10S1248	0.023	D9S1122	0.076	TH01	0.060
D18S51	0.047	D13S317	0.039	D6S1043	0.021	D7S820	0.067	FGA	0.060
D9S1122	0.044	D19S433	0.031	D4S2408	0.020	D10S1248	0.058	D19S433	0.054
D20S482	0.031	D7S820	0.031	TPOX	0.020	D5S818	0.058	D20S482	0.049
D17S1301	0.026	D10S1248	0.027	D5S818	0.017	CSF1PO	0.052	TPOX	0.039
D10S1248	0.025	D16S539	0.019	D9S1122	0.014	D16S539	0.048	CSF1PO	0.038
CSF1PO	0.011	D20S482	0.018	D17S1301	0.014	D17S1301	0.031	D10S1248	0.022
TPOX	0.009	D17S1301	0.015	CSF1PO	0.009	D20S482	0.031	D17S1301	0.021

British Chinese & South Asian		British Chinese & West African		North East African & South Asian		North East African & West African		South Asian & West African	
Locus	I_n	Locus	I_n	Locus	I_n	Locus	I_n	Locus	I_n
D21S11	0.140	D13S317	0.383	D2S1338	0.217	D2S1338	0.176	D21S11	0.203
D7S820	0.120	D21S11	0.317	D12S391	0.178	D12S391	0.133	D12S391	0.196
D4S2408	0.106	vWA	0.224	D21S11	0.152	D21S11	0.129	Penta D	0.186
D2S1338	0.100	Penta E	0.223	D1S1656	0.148	Penta D	0.096	D13S317	0.171
D13S317	0.090	D3S1358	0.222	D2S441	0.140	Penta E	0.093	D2S1338	0.152
TH01	0.084	D2S1338	0.191	D18S51	0.137	D13S317	0.085	D3S1358	0.151
D1S1656	0.079	Penta D	0.188	vWA	0.103	D3S1358	0.084	vWA	0.147
D12S391	0.076	D4S2408	0.179	D5S818	0.093	vWA	0.076	Penta E	0.142
D2S441	0.070	D12S391	0.175	Penta D	0.083	D2S441	0.072	D2S441	0.141
D9S1122	0.067	D1S1656	0.153	D8S1179	0.081	D9S1122	0.063	D18S51	0.137
Penta E	0.064	D8S1179	0.138	FGA	0.073	D6S1043	0.062	D1S1656	0.136
vWA	0.062	D2S441	0.136	D3S1358	0.065	D18S51	0.062	D8S1179	0.108
D6S1043	0.060	D5S818	0.124	CSF1PO	0.058	D1S1656	0.056	D19S433	0.101
D3S1358	0.058	D7S820	0.093	D13S317	0.057	D19S433	0.049	D6S1043	0.098
D5S818	0.056	D18S51	0.091	D19S433	0.055	FGA	0.048	D9S1122	0.083
D10S1248	0.049	TPOX	0.087	D16S539	0.048	D20S482	0.048	TH01	0.077
D16S539	0.044	D6S1043	0.085	Penta E	0.042	D8S1179	0.045	FGA	0.072
Penta D	0.043	D19S433	0.081	D10S1248	0.039	D7S820	0.042	CSF1PO	0.068
D8S1179	0.041	TH01	0.079	D6S1043	0.035	D5S818	0.039	D10S1248	0.066
D18S51	0.040	FGA	0.062	D20S482	0.035	TPOX	0.036	D7S820	0.059
D19S433	0.038	D16S539	0.060	D7S820	0.035	TH01	0.030	D5S818	0.059
TPOX	0.027	CSF1PO	0.057	D4S2408	0.032	D4S2408	0.028	D4S2408	0.057
FGA	0.025	D9S1122	0.048	TH01	0.031	D10S1248	0.026	TPOX	0.056
D17S1301	0.018	D17S1301	0.031	D9S1122	0.024	D16S539	0.025	D16S539	0.041
D20S482	0.017	D10S1248	0.029	TPOX	0.017	CSF1PO	0.020	D20S482	0.022
CSF1PO	0.013	D20S482	0.023	D17S1301	0.013	D17S1301	0.010	D17S1301	0.014

STRUCTURE plots were also run using the top 6, 12 and 24 most informative loci for distinguishing the British Chinese and West African populations (**Figure 5.14**) as an example of easy to separate clusters, given this pairing had the highest cumulative I_n. Although using just the 6 most informative loci resulted in no cluster differentiation, using the 12 most informative loci was almost as good as using all 26 markers, showing how genetically different these two populations are when looking at the autosomal STR results. This also correlates some of the findings of other research groups, who were able to show clear clustering patterns when removing geo-graphically close populations such as the Middle East and Central-South Asia groups from the HGDP-CEPH reference population panel [152, 170].

Figure 5.14 shows how easy it would be to differentiate British Chinese from West African samples using autosomal STR data from this marker set. The same thing was done for the White British and South Asian populations, but with the top 6, 12, 20, 21, 22, 23, 24 most informative loci to try and identify the best number of markers (Figure 5.15). It is immediately obvious from these STRUCTURE plots that these two populations are much harder to differentiate than the British Chinese and West African populations, although it is hard to tell from them what number of informative loci offers the best chance to distinguish the two groups. Average proportion of membership was extracted for these STRUCTURE runs and are shown in Figure 5.16.

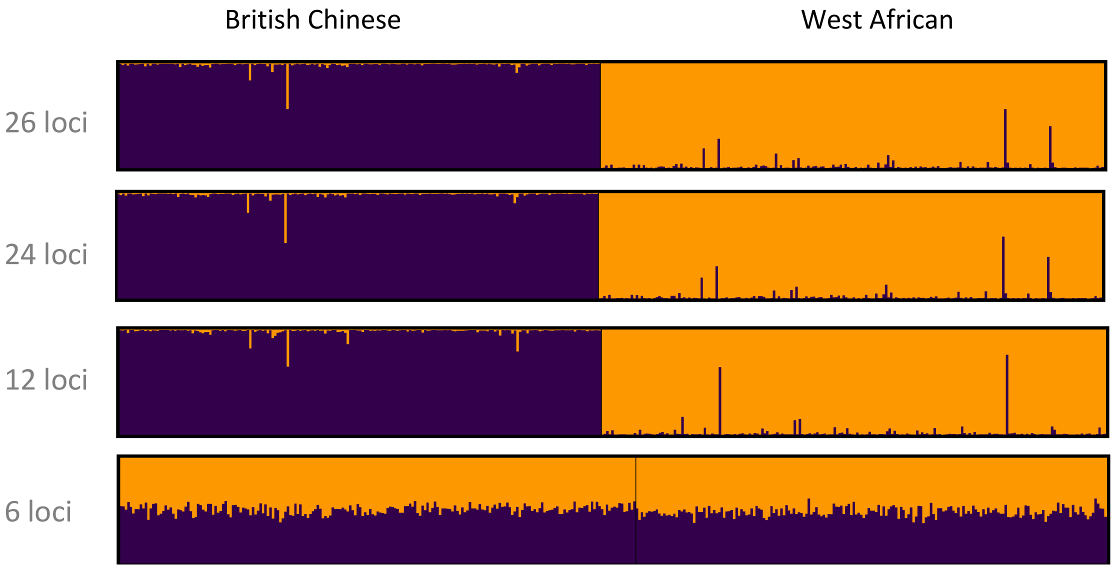


Figure 5.14: STRUCTURE plots for the British Chinese and West African populations using loci of most informativeness

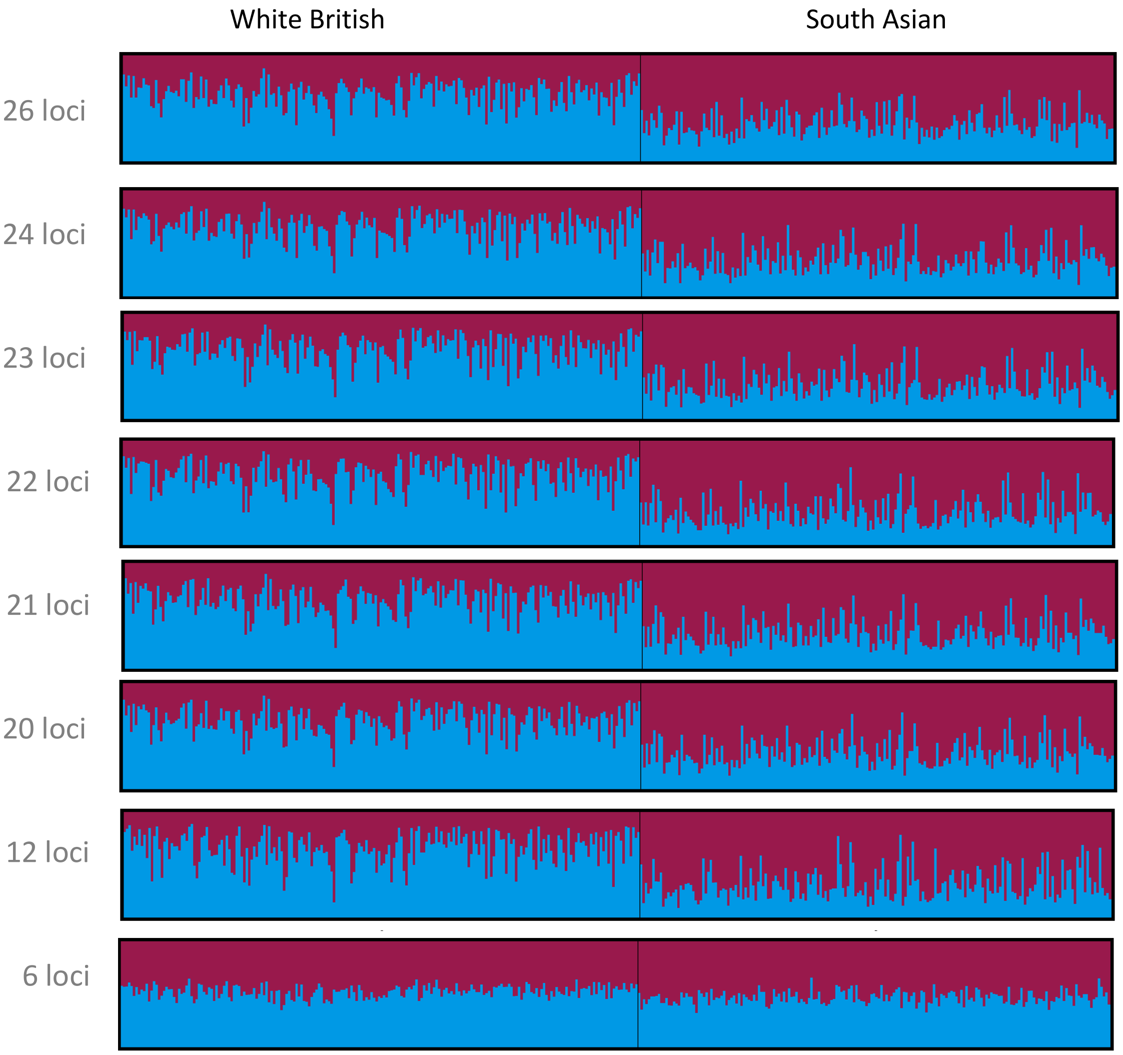


Figure 5.15: STRUCTURE plots for the White British and South Asian populations using loci of most informativeness

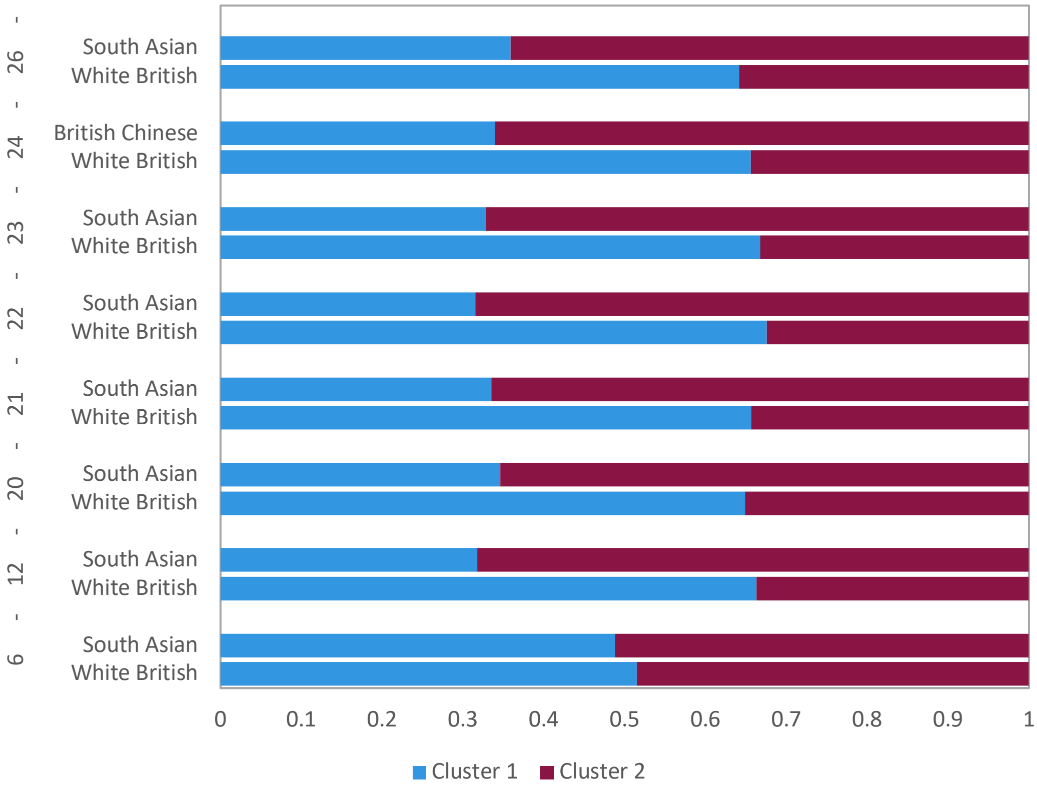


Figure 5.16: Proportion of membership for the White British and South Asian STRUCTURE runs

The number of loci (according to most informative, high I_n between the two groups) used for each STRUCTURE run is listed on the left-hand side of the figure.

Figure 5.16 suggests that the best number of loci for differentiating the White British and South Asian population groups, by a very narrow margin, is the 22 most informative loci. The difference between the runs using 12 or more loci is minimal, suggesting there is little added value with more loci. The average proportion of membership does remain less than that for either population when looking at all five groups simultaneously, and the STRUCTURE run indicates that individual ancestry inference is not improved for the samples which did not achieve a proportion of membership of over 0.7 for the correct cluster in the initial runs.

STRUCTURE plots were also re-run for all five populations using data from 16 markers: the top 12 most informative loci across all five populations, and an additional 4 markers with high I_n between specific pairs of populations (White British vs South Asian and British Chinese vs South Asian): D4S2408, TH01, D7S820, and D16S539. These markers are highlighted in Table 5.4 in red, and the STRUCTURE plot generated for all samples using these top 16 loci are shown with the plot for the 22 most informative loci in Figure

5.17. The aim of doing this was to observe whether using a reduced number of markers, but including loci that are specifically informative for these groups might help reduce noise whilst improving ancestry inference. The average proportion of assignment coefficients for this STRUCTURE plot confirm what can be seen in the figure, that using this subset of loci does not improve on the plot obtained using the top 22 most informative loci. The top 22 loci coincidentally already contain those four markers identified as potentially useful for distinguishing the South Asian population from the White British and British Chinese groups.

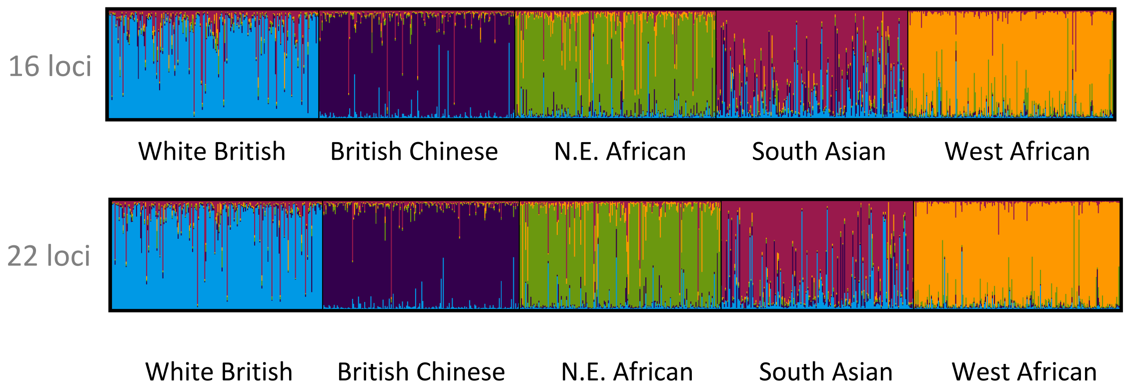


Figure 5.17: STRUCTURE plots for all five populations using 16 loci compared to 22

The rationale for reducing the number of loci used was that some noise may be removed from the STRUCTURE plots by removing markers that are not useful in terms of ancestry inference, but could be adding in unnecessary variation unrelated to population specific enrichment. Whilst an interesting concept, reducing the number of loci according to ancestry informativeness does not seem to greatly improve population inference in this dataset. Calculating I_n values helped confirm how much more useful sequence-level data is in terms of using the autosomal STR data for ancestry estimation, but the extent of what can be achieved using this STR data alone appears to reach its limit when using all data available: complete sequence-level allelic data for 26 autosomal STRs.

5.3. Ancestry informative SNPs

The ForenSeq DNA Signature Prep kit contains a primer mix (DNA Primer Mix B, DPMB) which consists of the same primers as those in DNA Primer Mix A, with the addition of primers for the amplification of 22 phenotype-informative SNPs, and 56 biogeographical ancestry-informative SNPs. It is currently the only commercial solution for the simultaneous amplification of core autosomal STR loci and ancestry-informative markers and showcases the multiplexing advantage of using MPS. Given the co-amplification of these SNPs with the previously discussed STRs when using DPMB, and the promising results from the sequence-based STR alleles, the combined value of these markers for ancestry estimation comes into question. The next goal of this research was to compare results obtained from the STRs to those of this specifically chosen ancestry-informative SNP set on the same samples, before ascertaining if the combination of both marker types provides a better population differentiation than either alone.

The 56 SNPs used for ancestry inference in DPMB were selected from two publications that focused on identifying a small panel of highly informative SNP markers for ancestry estimation [104, 227]. In 2014, Kidd et al. stated that a very large number of ancestry-informative markers could provide accurate discrimination of 6-7 geographic regions, but a small, efficient, and robust panel is more relevant for forensic applications. They went on to identify a small panel of SNPs which would be useful for global population differentiation [151]. The resulting panel of 55 SNPs is well characterised, has been broadly applied as a standalone panel and is commonly referred to as the "Kidd SNPs". The 56 SNPs amplified by the ForenSeq DNA Signature Prep DMPB correspond to these 55 Kidd SNPs, as well as SNP rs1919550 which appears to have limited global variability, but is useful at distinguishing native American individuals from other populations [228]. Presumably, the decision was made to incorporate this SNP to enhance the primer set's capability to separate American populations, although the fact that it is in full linkage disequilibrium with another SNP (rs12498138) in DPMB makes it redundant (C. Phillips, personal communication). Figure 5.18 and Figure 5.19 show examples of the allelic distribution at rs1919550 and another SNP, rs1042602 across the five populations studied, using the data for the 219-sample set. As suggested above, rs1919550 appears to show limited variability across the five populations, with the A allele being very common in all groups studied. rs1042602, however, is highly polymorphic in Europeans, where the A allele has been associated with light skin and eye colour [229, 230].

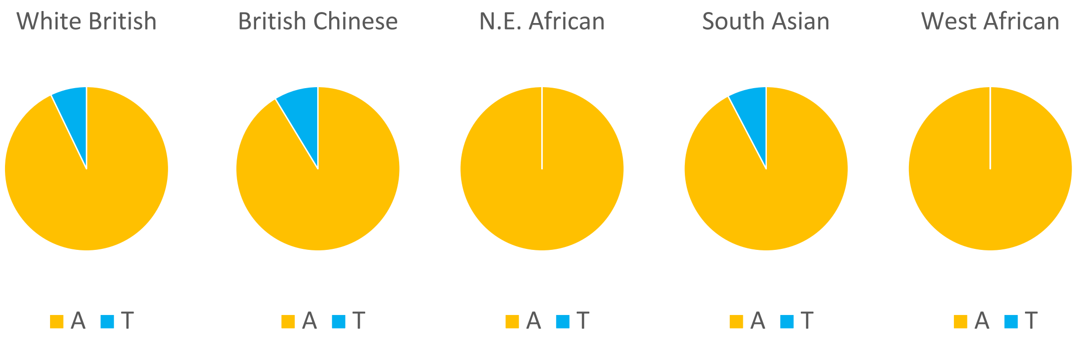


Figure 5.18: Allele frequency distribution for rs1919550

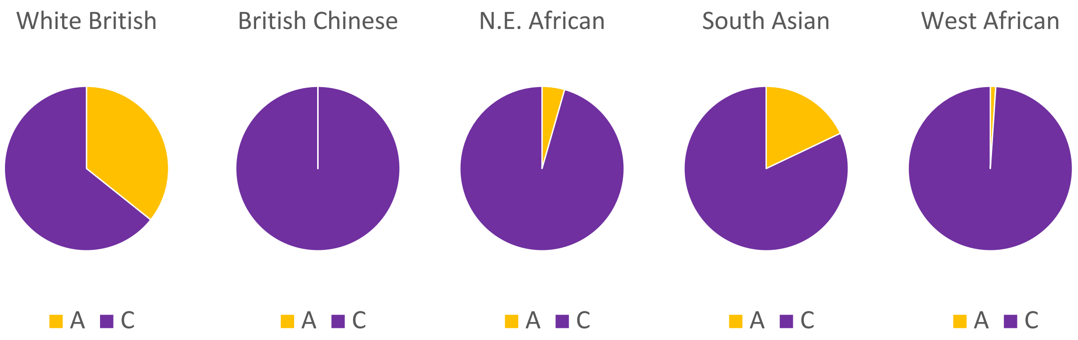
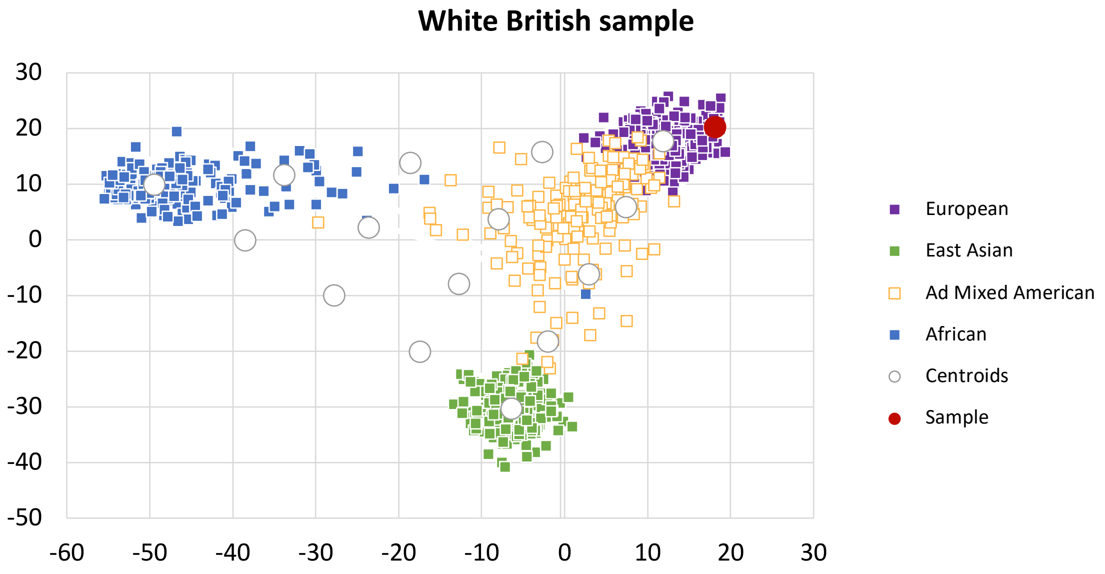
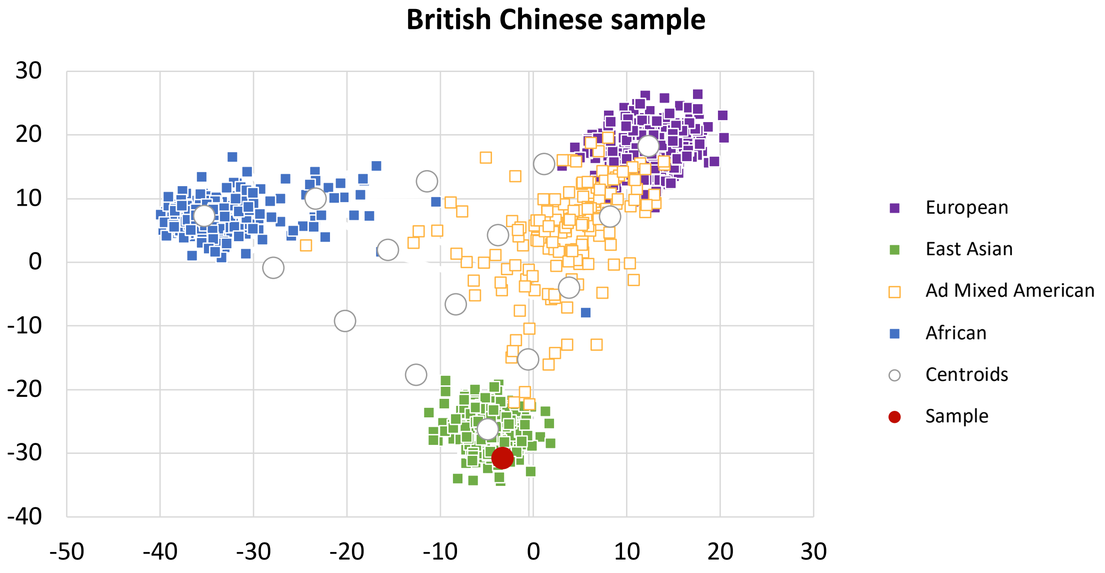


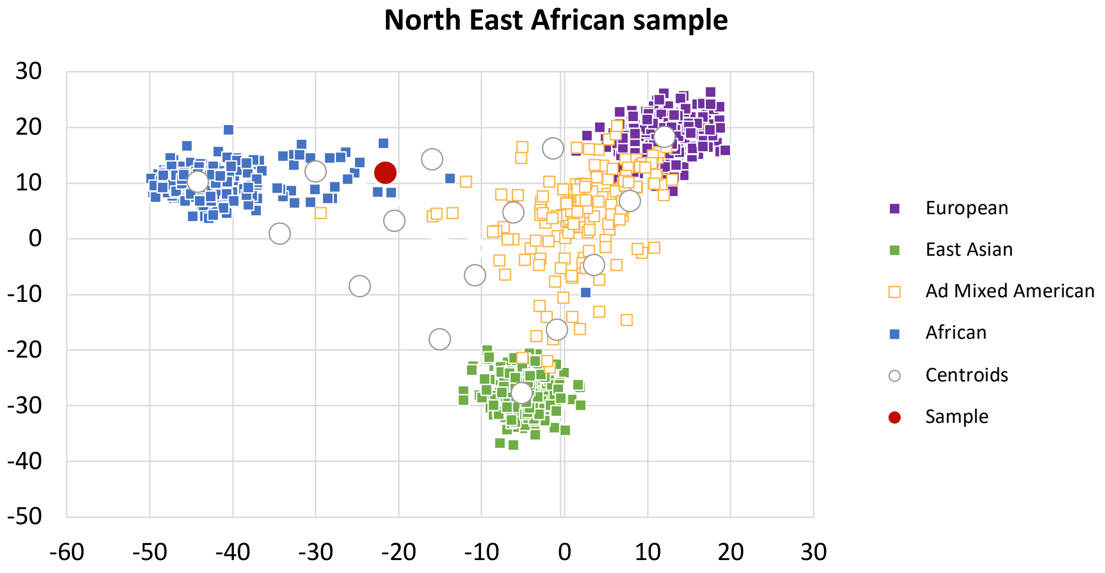
Figure 5.19: Allele frequency distribution for rs1042602

5.3.1. Universal Analysis Software estimation

In the UAS, the 56 aSNPs are analysed using principal component analysis (PCA). The model in UAS was trained on the European, East Asian and African (except for the ASW, “African Ancestry in Southwest US” group) super populations of the 1000 Genomes Phase I data [231]. The sample being tested is then projected based on its aSNP genotype calls onto the pre-trained components of the PCA plot, alongside data for the Ad-Mixed American super population for context. Although the estimation feature of UAS can be useful for a sample whose bio-geographical ancestry aligns with one of the three reference super populations, it is not effective for predicting the ancestry of an unrepresented group. Figure 5.20 shows an example of PCA plot generated by UAS for a sample from each population group studied.


White British sample
30
20
10
0
-10
-20
-30
-40
-50
-60
-50
-40
-30
-20
-10
0
10
20
30
European
East Asian
Ad Mixed American
African
Centroids
Sample


British Chinese sample
30
20
10
0
-10
-20
-30
-40
-50
-40
-30
-20
-10
0
10
20
30
European
East Asian
Ad Mixed American
African
Centroids
Sample


North East African sample
30
20
10
0
-10
-20
-30
-40
-50
-60
-50
-40
-30
-20
-10
0
10
20
30
European
East Asian
Ad Mixed American
African
Centroids
Sample

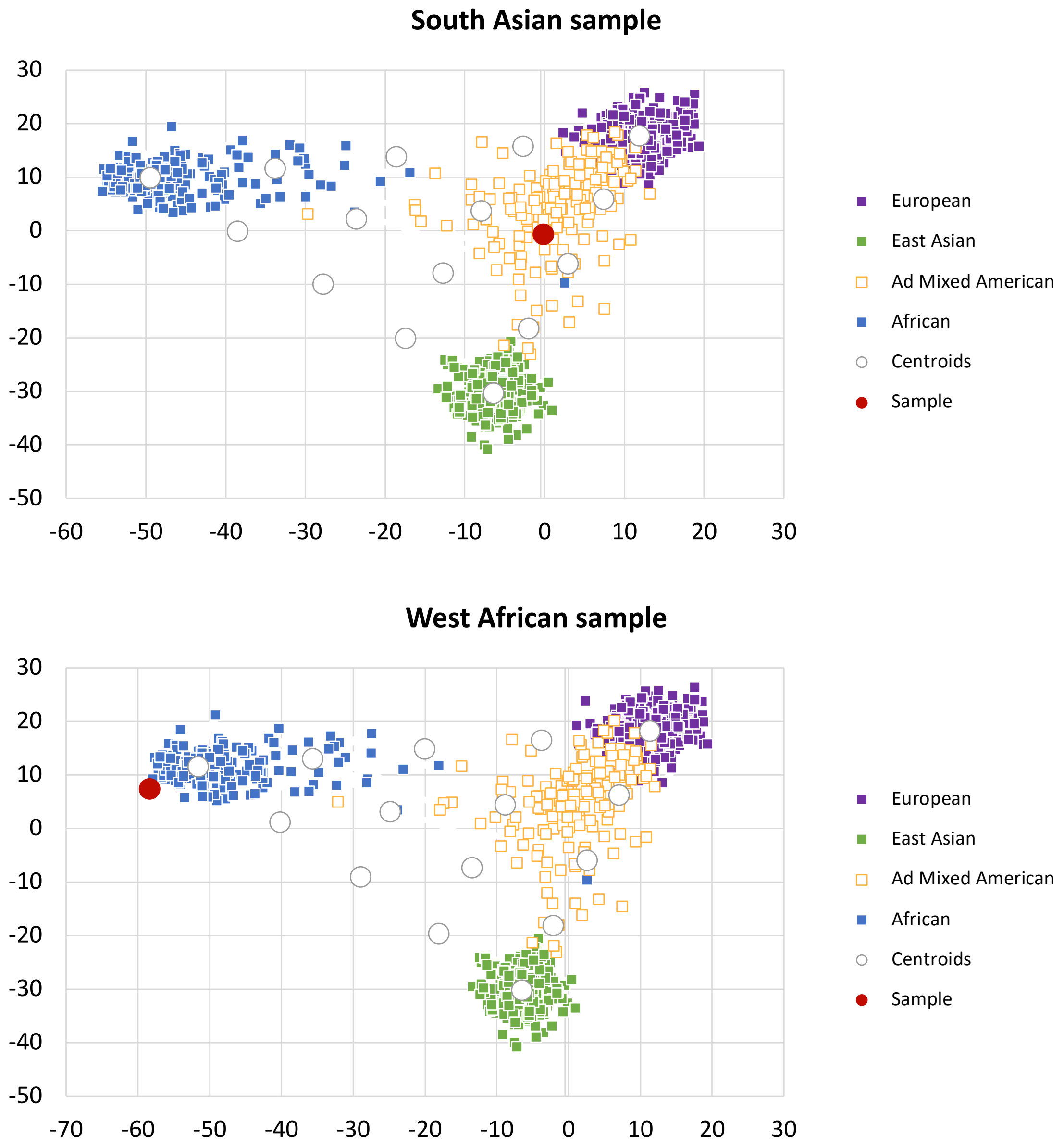


Figure 5.20: UAS bio-geographical ancestry estimation plots

Examples of the PCA plots generated using UAS for bio-geographical ancestry estimation for a sample from each population group. The populations on which the software was trained are shown in blue, purple and green for the African, European and Asian population, respectively. The admixed American populations are shown in orange, and are used for context rather than estimation. The dot representing the sample being tested is shown in red, and the user is able to make a general interpretation as to which meta-population the sample clusters closest to.

The White British, British Chinese and West African samples are projected close to the European, Asian and African samples respectively, as expected. The North East African and South Asian samples are less straightforward, with the South Asian sample falling half way between the Asian and European clusters. This is undoubtedly due to lack of reference data for these populations.

This shortcoming of the UAS software for bio-geographical ancestry estimation has already been highlighted by several publications, including Ramani et al. who genotyped 1030 unrelated individuals living in Singapore of Chinese, Malay and Indian origin [232]. Whilst the UAS software was able to accurately place the Chinese samples with the East Asian cluster on the PCA plot, the Malay and Indian samples clustered in between reference populations. Hussing et al. found that 22 out of 23 European samples projected close to the correct cluster, whereas the handful of Middle Eastern and North African samples they sequenced did not return a useable prediction [116]. Wendt et al. noted that their Yavapei Native American population samples clustered either with the East Asian cluster or in between reference populations [233]. Despite the presence of a SNP specifically chosen for differentiating American populations, the lack of reference data for Native American populations once again hinders any interpretation.

Although the UAS has limited use for ancestry determination due to lacking reference data for certain populations such as South Asian, North East African, Middle Eastern etc., it is expected that genotype results for the SNPs targeted in DPMB would still be useful for distinguishing these populations using other software.

4.1.1. STRUCTURE analysis

As with the autosomal STR data, aSNP genotypes were used to generate STRUCTURE plots. Figure 5.21 shows that the software was able to distinguish 5 distinct clusters under K=5, corresponding to the 5 population groups.

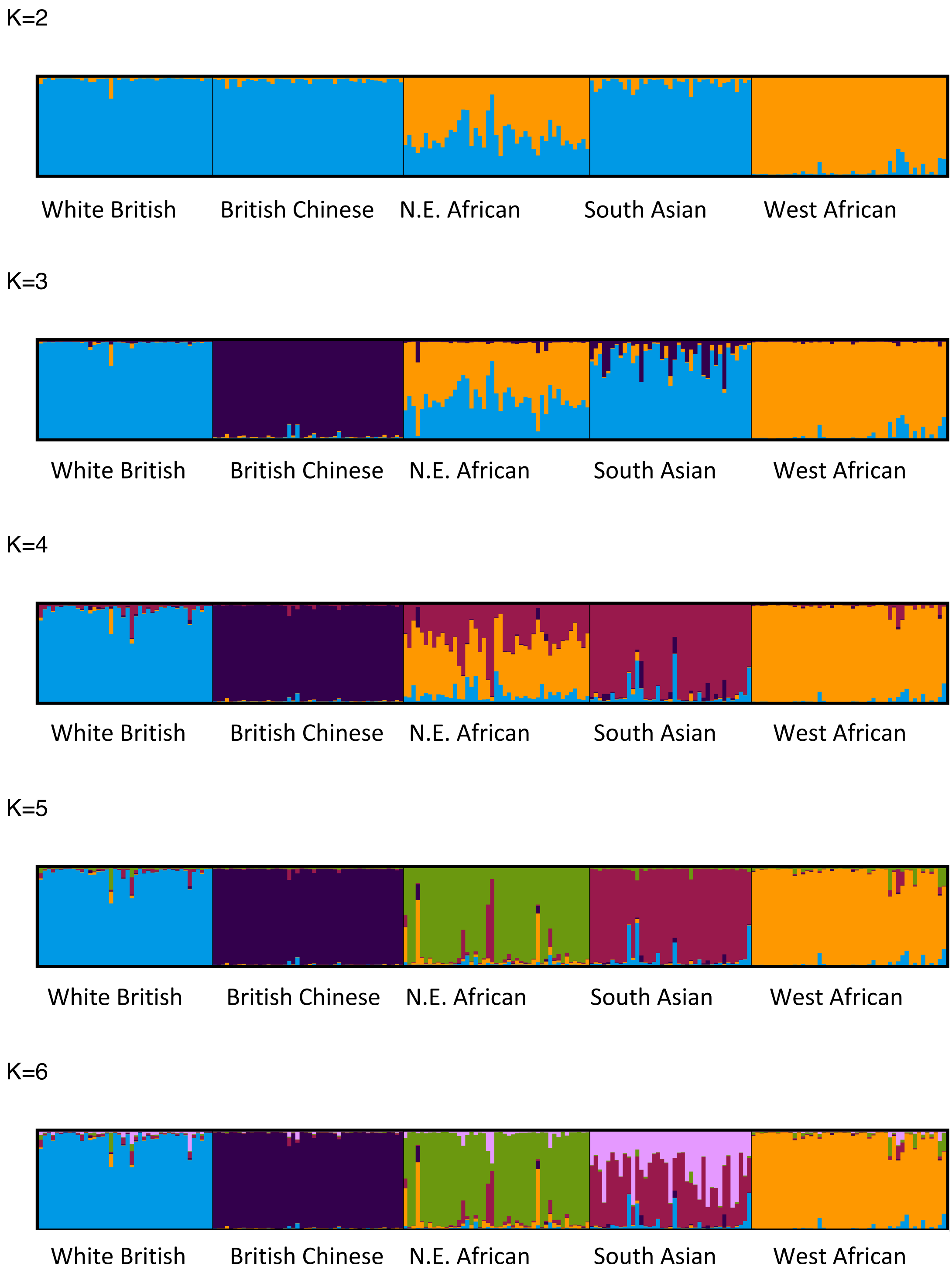


Figure 5.21: STRUCTURE plots for the five populations studied, generated using genotypes for the 56 aSNP markers targeted by DNA Primer Mix B in the ForenSeq DNA Signature Prep Kit.

5.3.1.1. Proportion of membership and incorrectly assigned samples

If a proportion of membership to the correct cluster of over 0.7 is taken as a correct population assignment, 94% of samples (205/219) are assigned to the correct group, and 98% of samples (214/219) if using a value of 0.5 or higher. Given that the samples re-analysed for the aSNPs were chosen at random from the 989 samples analysed for the autosomal STRs, aSNP data was not available for all the samples which were incorrectly assigned with aSTRs in section 5.2.2.3. Of the 67 samples which had a proportion of membership of over 0.5 for the wrong cluster using aSTR data (Table 5.2), 17 happened to be re-analysed using the primers for the aSNPs in DPMB. Of these, only 4 have a proportion of membership of over 0.5 for the wrong cluster using aSNP results, and only one of those had a coefficient of over 0.7. Table 5.5 shows the proportion of membership for these 4 samples, and also shows the only sample which did not have a proportion of membership of over 0.5 for any of the groups.

Table 5.5: Samples with a proportion of membership of >0.5 for the incorrect cluster using aSNP data

For each sample, the value in black corresponds to the proportion of membership for the correct cluster, and the value in red corresponds to the highest proportion of membership. All other values are shown in grey. WB= White British, BC= British Chinese, NEA= North East African, SA= South Asian, WA= West African.

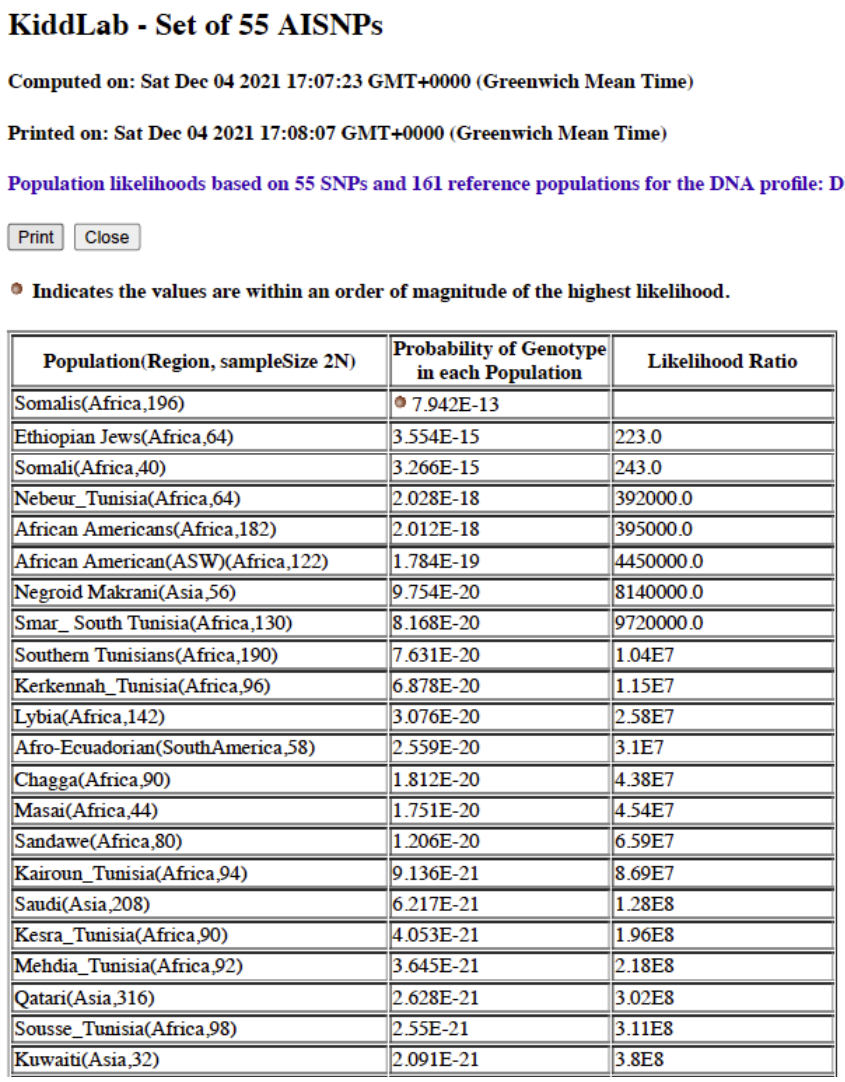
Sample	Population	1. WB	2. SA	3. BC	4. WA	5. NEA
NEA404	NEA	0.007	0.012	0.161	**0.666**	**0.155**
NEA438	NEA	0.025	**0.55**	0.004	0.046	**0.375**
NEA439	NEA	0.008	**0.855**	0.002	0.023	**0.111**
NEA465	NEA	0.03	0.014	0.075	**0.505**	**0.376**
SA654	SA	0.432	**0.384**	0.013	0.045	0.126

The four samples which had a proportion of membership of over 0.5 for the wrong cluster were North East African. Samples NEA404 and NEA465 have a proportion of membership of 0.66 and 0.505 respectively, for the West African cluster, despite self-declared North East African ancestry. In the aSTR STRUCTURE analysis, NEA404 and NEA465 also clustered with the West African population cluster, with a proportion of membership of 0.9 and 0.8 assignment, respectively, for this population. The reason for these samples misclassifying is uncertain, although both are known to have identity-informative SNP alleles only seen in the West African population (unpublished data). This could indicate incorrect self-declared ancestry, or even recent admixture.

Samples NEA438 and NEA439 both showed a proportion of membership of just over 0.5 for the South Asian cluster using aSTR data, but this increases to 0.85 for NEA439 when using aSNP genotypes, causing it to be mis-classified. Although very little is known about the donors of the samples, they do appear to have come from the same region of Somalia according to their self-declared ancestry. These results may indicate that STRUCTURE is picking up on population substructure, or a population that is more closely related genetically to the samples in the South Asian cluster than those in the North East African cluster.

5.3.2. FROG-KB

The 55 ancestry informative SNPs originally developed as a standalone panel for ancestry inference developed by the group at the university of Yale [105] can be used with the FROG-kb database [234]. This database contains data from populations not included in the UAS reference set, including from the Middle East, South and Central Asian, and Oceania. FROG-kb provides the ability to calculate relative likelihoods of ancestry from different reference populations for uploaded aSNP genotypes, derived from the ALlele FREquency Database (ALFRED http://alfred.med.yale.edu). Figure 5.22 shows an example of the type of results obtained when uploading aSNP genotypes for a North East African sample. This sample is the same as the one shown in Figure 5.20 which does not cluster with any of the three super populations present in UAS, which is expected as the UAS database does not contain North East African reference population data. This particular sample returns the highest probability of genotype for the Somali population, which is consistent with the self-declared ancestry. This highlights the need for appropriate reference population, with the FROG-kb algorithm using 161 populations, including a Somali population of 196 individuals.



KiddLab - Set of 55 AISNPs

Computed on: Sat Dec 04 2021 17:07:23 GMT+0000 (Greenwich Mean Time)

Printed on: Sat Dec 04 2021 17:08:07 GMT+0000 (Greenwich Mean Time)

Population likelihoods based on 55 SNPs and 161 reference populations for the DNA profile: DNAProfile_UId

● Indicates the values are within an order of magnitude of the highest likelihood.

Population(Region, sampleSize 2N)	Probability of Genotype in each Population	Likelihood Ratio
Somalis(Africa,196)	● 7.942E-13	
Ethiopian Jews(Africa,64)	3.554E-15	223.0
Somali(Africa,40)	3.266E-15	243.0
Nebeur_Tunisia(Africa,64)	2.028E-18	392000.0
African Americans(Africa,182)	2.012E-18	395000.0
African American(ASW)(Africa,122)	1.784E-19	4450000.0
Negroid Makrani(Asia,56)	9.754E-20	8140000.0
Smar_ South Tunisia(Africa,130)	8.168E-20	9720000.0
Southern Tunisians(Africa,190)	7.631E-20	1.04E7
Kerkennah_Tunisia(Africa,96)	6.878E-20	1.15E7
Lybia(Africa,142)	3.076E-20	2.58E7
Afro-Ecuadorian(SouthAmerica,58)	2.559E-20	3.1E7
Chagga(Africa,90)	1.812E-20	4.38E7
Masai(Africa,44)	1.751E-20	4.54E7
Sandawe(Africa,80)	1.206E-20	6.59E7
Kairoun_Tunisia(Africa,94)	9.136E-21	8.69E7
Saudi(Asia,208)	6.217E-21	1.28E8
Kesra_Tunisia(Africa,90)	4.053E-21	1.96E8
Mehdia_Tunisia(Africa,92)	3.645E-21	2.18E8
Qatari(Asia,316)	2.628E-21	3.02E8
Sousse_Tunisia(Africa,98)	2.55E-21	3.11E8
Kuwaiti(Asia,32)	2.091E-21	3.8E8

Figure 5.22: Results obtained from FROG-kb for a North East African sample

The four samples listed in Table 5.5 which returned a proportion of assignment of over 0.5 for the incorrect population, were also run through FROG-kb. Samples NEA438 and NEA439 clustered with the South Asian samples despite having a North East African self-declared ancestry. Figure 5.23 shows that sample NEA438 returns probabilities of genotypes for multiple populations that are all within an order of magnitude of the highest probability of genotype, which is for the Negroid Makrani population, an ethnic group of Pakistan and Indian with African heritage. Sample NEA439 returned the highest probability of genotype for the Saudi and Quatari populations (5.338E-13 and 6.586E-14, respectively). As stated previously, these individuals are of Somali self-declared ancestry, but analysis may be picking up on population substructure, or highlighting a small population with genetic similarities to Middle Eastern or South Asian groups, for which reference data is perhaps not yet available.

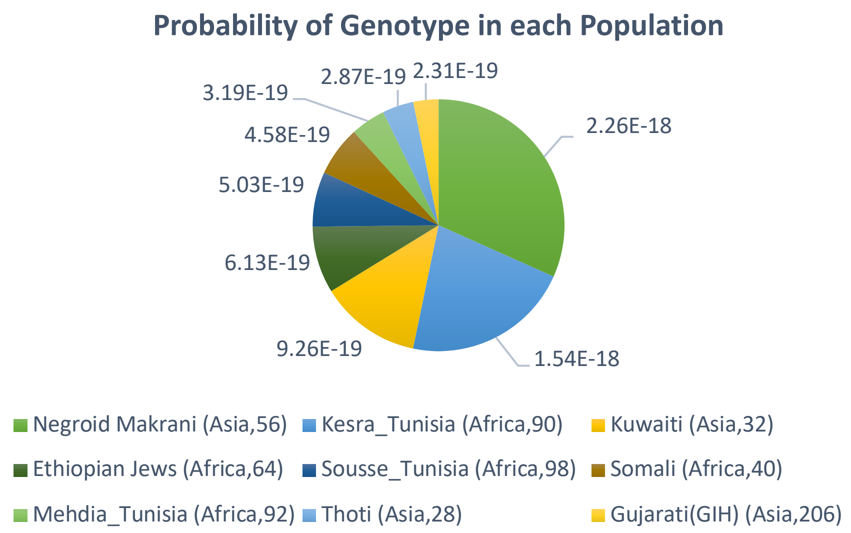


Figure 5.23: FROG-kb results for sample NEA438

Probability of genotype for each of the populations where the values are within an order of magnitude of the highest likelihood.

Samples NEA404 and NEA465 clustered with the West African samples in STRUCTURE, despite self-declared North East African ancestry. Figure 5.24 shows the FROG-kb results for these two samples, showing they both have the highest probability of genotype in the Sandawe population of Tanzania, which is geographically closer to North East Africa than West Africa. If these persons are decended from individuals recently emigrated from Tanzania, this could explain this phenomenom and the incorrect classification using STRUCTURE due to lack of other Tanzanian samples in the dataset.

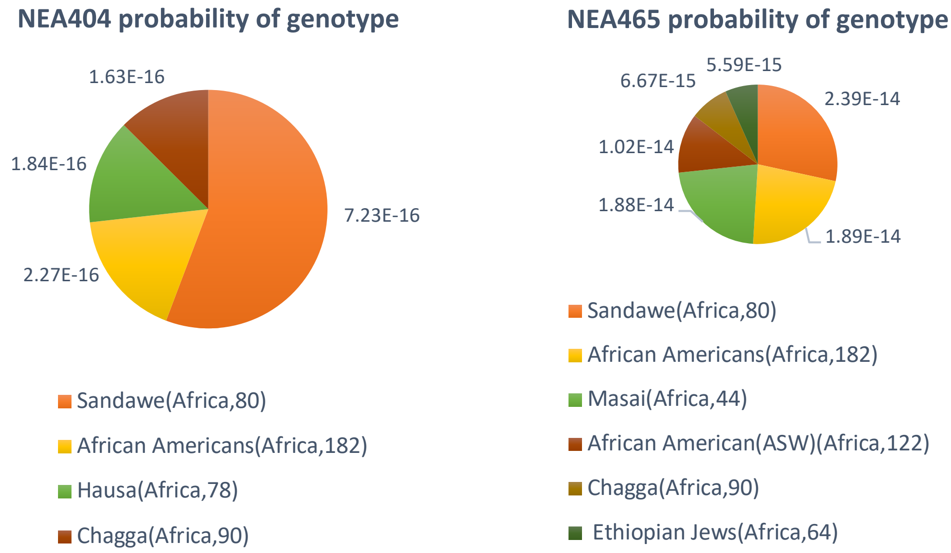


Figure 5.24: FROG-kb results for samples NEA404 and NEA465

Probability of genotype for each of the populations where the values are within an order of magnitude of the highest likelihood.

Results obtained by targeting the 56 aSNPs in DPMB confirm that this panel can reliably be used for estimating ancestry. The lack of reference populations in UAS make its utility limited, but genotypes can easily be extracted for use with other, third-party tools. Individual sample results can be uploaded to websites such as FROG-kb, which have their own reference populations for ancestry estimation, enabling the probability of any given profile within each population to be calculated. STRUCTURE plots enable the results for an entire dataset to be visualised, showing in this work how the different samples can clearly be separated into 5 clusters, due to the genetic similarities within a population, and the genetic differences between separate global populations.

5.4. Combining ancestry informative SNP and autosomal STR data

In their 2011 study, Phillips et al. concluded that the highest classification success could be obtained by combining the genotypes from forensic STRs with ancestry-informative SNPs [170]. This resulted in error free assignments, and group membership proportions of above 0.7 for five global population groups (African, East Asian, American, European and Oceanian). This represented an improvement on both the use of aSTRs alone, but also on their 34plex SNP assay which had led to a small number of European samples being misclassified. Given that the ForenSeq DNA Signature Prep kit DPMB is used to amplify autosomal STRs and SNPs simultaneously, the next step was to identify whether combining data for both marker types led to a more powerful level of population differentiation for the samples studied.

5.4.1. STRUCTURE Analysis

As with the aSTR and aSNP data, STRUCTURE plots were run for K=2 to K=6 for the combined genotypes for the 219 samples analysed using both markers types. These plots are shown in Figure 5.25.

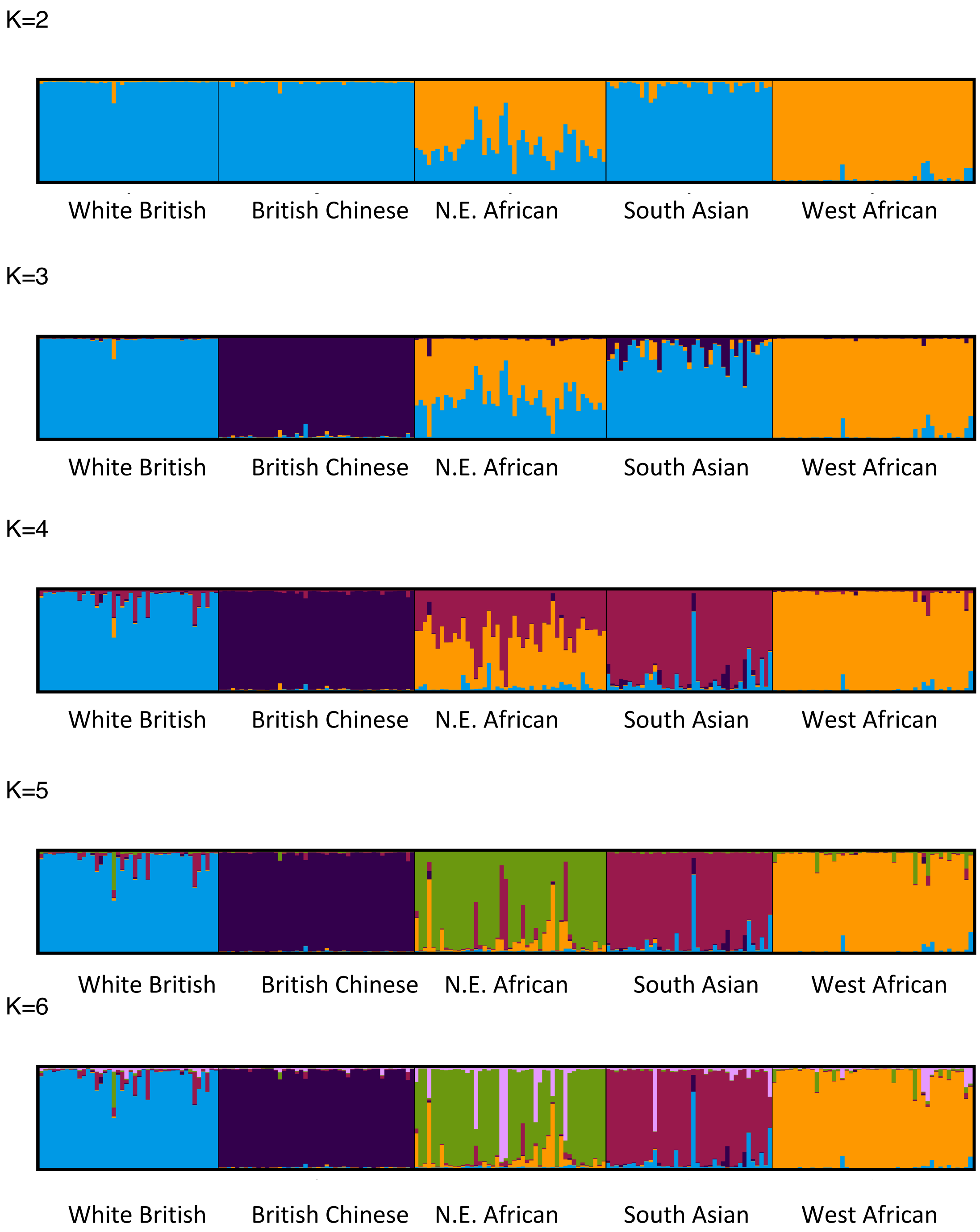


Figure 5.25: STRUCTURE plots for the five populations studied, generated using aSTR and aSNP genotypes for 219 samples

5.4.1.1. Proportion of membership

Average proportion of membership was extracted from the STRUCTURE analyses (K=5) run with autosomal STR data alone, ancestry-informative SNP data alone and combined aSTR and aSNP data. Figure 5.26 shows the proportion of membership to the different clusters for each population. Overall, the aSNP and combined runs appear to provide the highest average proportion of membership to the correct population for all groups. This improvement compared to the aSTR run is more visible for the White British and South Asian populations. The difference between the combined run and the aSNP data alone run appears negligible, except for the North East African population where the combined run is more similar to the aSTR run.

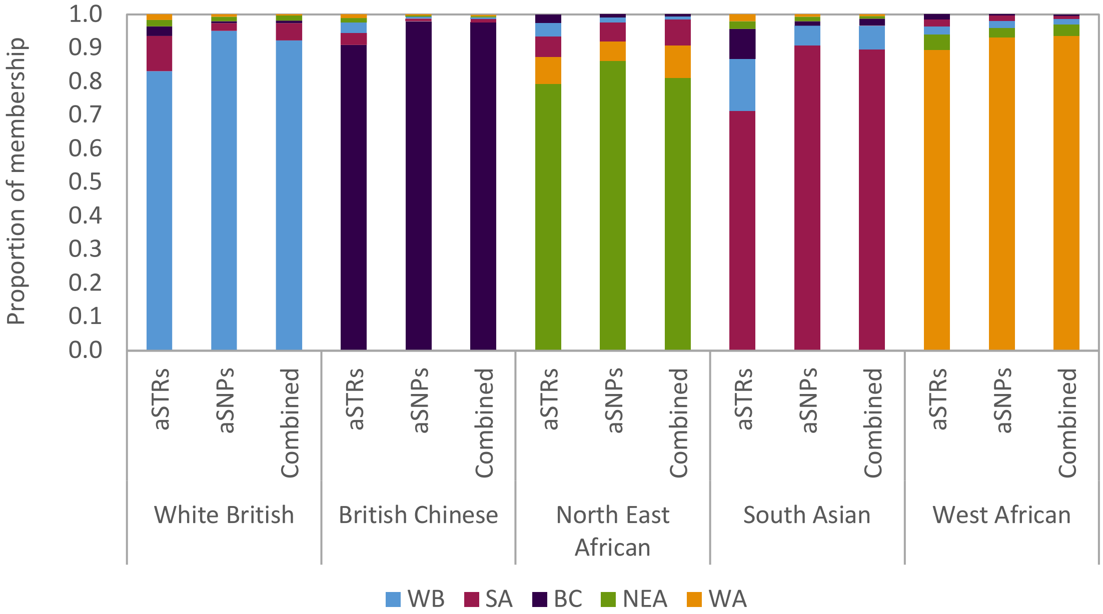


Figure 5.26: Proportion of membership obtained from STRUCTURE for runs generated using aSTR, aSNP, and combined aSTR and aSNP genotypes

In order to delve further into the difference between the STRUCTURE runs generated using aSTR data alone and the combined runs, Figure 5.27 shows the proportion of assignment for the individual samples that were analysed with K=5 for just autosomal STR data, and for the combined STR and ancestry-informative SNP data, i.e. the same samples as used for the STRUCTURE runs presented in Figure 5.25. Each line represents a sample, and the gain in proportion of assignment to the correct cluster when looking at the combined run is shown in pink, whilst the reverse (loss in proportion of assignment to the correct cluster) is shown in grey. All of the British Chinese samples had been correctly assigned using STR data alone, so although the addition of SNP genotypes did improve the average assignment coefficient for the correct cluster, the difference on an individual sample level is negligible. A similar observation can be made for the West African population, where the majority of samples classified correctly using aSTR data alone, although the addition of SNP data does push 2 previously inconclusive samples to be correctly classified (proportion of membership > 0.7 for the West African cluster).

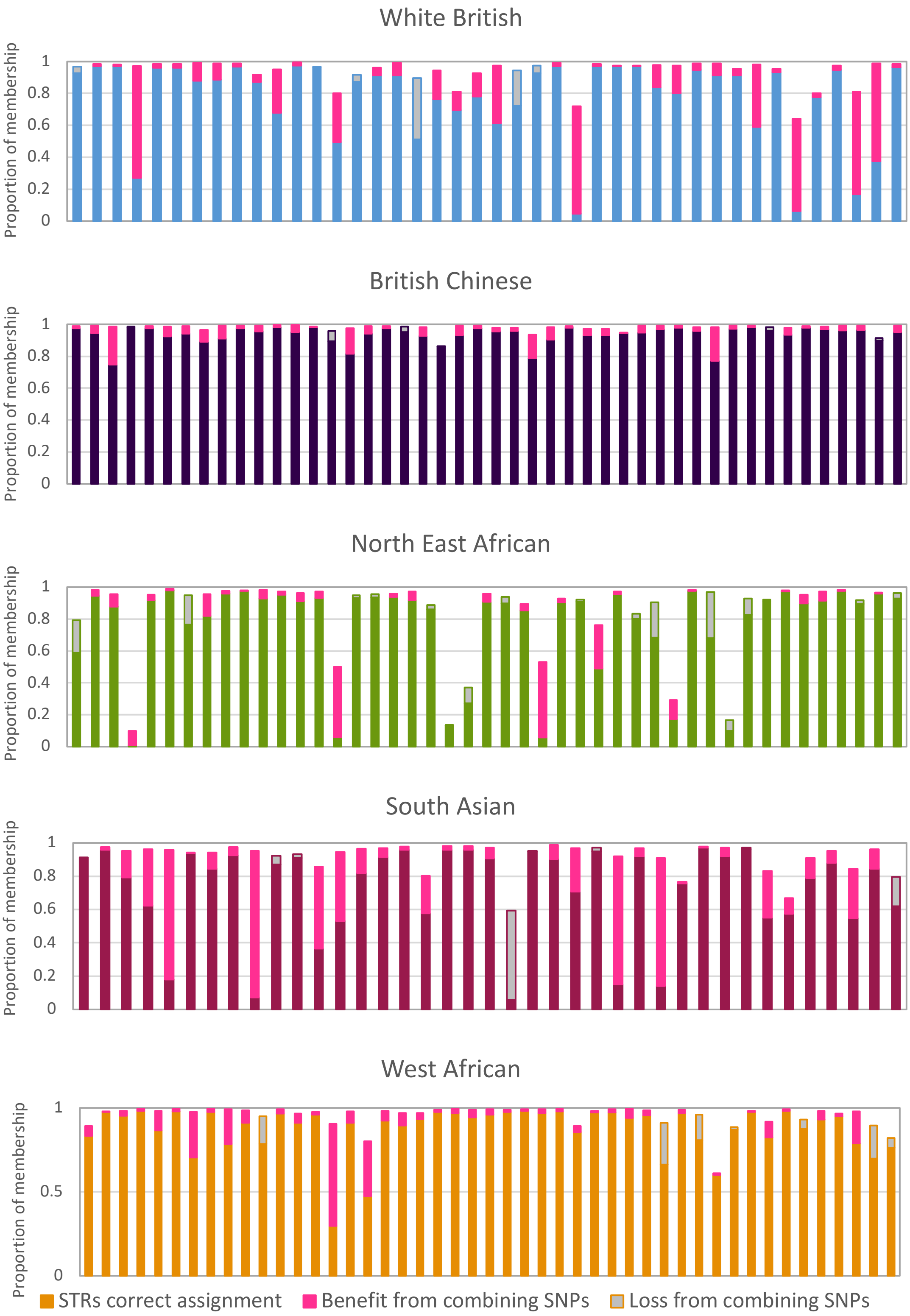


Figure 5.27: Individual proportion of membership to the correct cluster

Each bar represents an individual, with the proportion of membership to the correct population cluster coloured accordingly in each panel (White British: blue; British Chinese: dark purple; North East African: dark green; South Asian: plum and West African: orange). Gain in correct proportion of membership when adding aSNP data is shown in pink, whilst decrease in correct proportion of membership is shown in grey.

The benefit of combining the two sets of markers is most apparent in the White British and South Asian populations. These two populations were the hardest to separate when looking at aSTR data alone, and the addition of 56 ancestry-informative SNPs here clearly helps to push the individual proportion of membership to the correct cluster. Six samples in the White British population were inconclusive using aSTR results (no proportion of membership > 0.7), and 3 misclassified as South Asian. All cluster correctly in the combined run, bar 1 which moved from incorrectly classified to inconclusive. Of the 9 inconclusive samples and 3 samples with incorrect cluster assignment when only looking at STR data for the South Asian population, adding SNPs leads to a correct assignment for all bar 2 samples. One remains inconclusive with a proportion of membership below 0.7 for any cluster (albeit at 0.67 for the South Asian cluster), whilst one of the samples goes from inconclusive to incorrectly assigned to the White British cluster. This sample, shown with a largely grey bar in the middle of the South Asian panel in Figure 5.27, has a proportion of membership for the South Asian cluster of 0.59 within the aSTR run, 0.71 within the aSNP run but of 0.06 in the combined run – clustering much better with the White British group instead (Sample SA683, discussed further). Another notable sample in the South Asian group is one which shows a proportion of membership increase from 0.37 to 0.86 when comparing the aSTR only run to the combined run. This sample was the only sample to be inconclusive with aSNP data alone (SA654 in Table 5.5), suggesting the combined data is more accurate for ancestry inference of this sample. Finally, one South Asian sample which had correctly clustered in the aSTR run only is inconclusive in the combined run.

The results for the North East African group are varied. Where 5 samples were inconclusive with STR data alone, 2 remain inconclusive in the combined run and 2 are assigned to the incorrect, South Asian cluster with a proportion of over 0.7. Of the 3 samples which had a proportion of above 0.7 for the incorrect cluster, two are inconclusive with the combined run, and one remains incorrectly assigned. Additionally, two samples which had the correct assignment with STR data alone are inconclusive in the combined run. The three samples with a proportion of membership of above 0.7 for the wrong cluster will be discussed in the following section.

5.4.1.2. Incorrectly assigned samples

Of the 35 samples which have a proportion of membership of above 0.7 for the incorrect cluster when using aSTR data, 9 were re-analysed for the DPMB aSNPs. Of these, 1 has a proportion of membership of over 0.7 for the wrong population using aSNP results (as discussed in section 5.3.1.1, Table 5.5), which is still the case in the combined graph. In total, 4 samples were incorrectly assigned in the combined STRUCTURE plots. Table 5.6 shows the proportion of membership values for these samples, compared to that of the aSTR and aSNP only STRUCTURE runs. Sample NEA404 clusters with the West African group despite self-declared North East African ancestry, and assigned Tanzanian ancestry by FROG-kb. Sample SA683 was correctly assigned to the South Asian cluster using aSNP data alone, and had a proportion of membership for this cluster of 0.59 with aSTR data, but surprisingly is incorrectly assigned to the White British group when looking at the combined plot.

Table 5.6: Proportion of membership for individual samples which were incorrectly assigned in the combined aSNP and aSTR STRUCTURE run

The proportion of membership for the correct cluster is highlighted in bold for each line, and the values are colour coded (smaller to largest = lightest to darker blue).

Sample	Pop	Run	WB	BC	NEA	SA	WA
NEA404	NEA	aSTRs	0.026	0.012	**0.011**	0.041	0.910
		aSNPs	0.007	0.161	**0.155**	0.012	0.666
		Combined	0.002	0.086	**0.100**	0.089	0.723
NEA438	NEA	aSTRs	0.131	0.012	**0.134**	0.557	0.167
		aSNPs	0.025	0.004	**0.375**	0.550	0.046
		Combined	0.011	0.002	**0.133**	0.741	0.113
NEA439	NEA	aSTRs	0.019	0.073	**0.372**	0.512	0.024
		aSNPs	0.008	0.002	**0.111**	0.855	0.023
		Combined	0.010	0.006	**0.272**	0.706	0.006
SA683	SA	aSTRs	0.373	0.016	0.012	**0.592**	0.007
		aSNPs	0.233	0.049	0.003	**0.714**	0.001
		Combined	0.775	0.164	0.003	**0.057**	0.002

Samples NEA438 and NEA439 are the two North East African samples which consistently appear to be assigned a higher proportion of membership for the South Asian cluster, and which gave mixed results when uploaded to FROG-kb. When looking at the run for k=6, Figure 5.28 shows that STRUCTURE does appear to be picking up a "new" group for those two samples amongst the other North East African samples, supporting the theory that these samples may be more genetically related to a different, under-represented population in this study. These two individuals may self-report as of Somali origin but in fact have recent Middle Eastern ancestry for example. Future work could look at adding Middle Eastern sample data in the STRUCTURE analyses to see if these particular samples cluster with a higher proportion of membership to this population group.

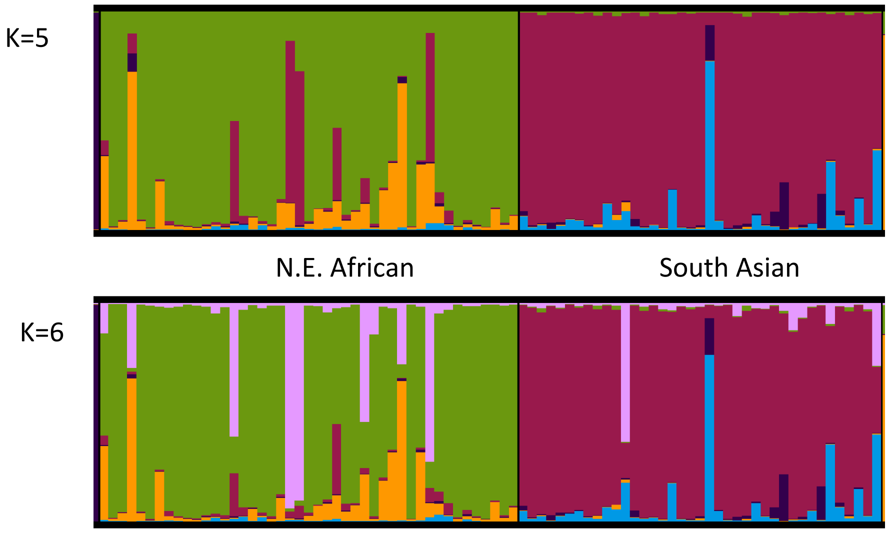


Figure 5.28: STRUCTURE results for K=5 and K=6 for the North East African and South Asian samples with combined aSNP and aSTR data

Zoomed in view from the STRUCTURE plots in Figure 5.25.

5.5. Discussion on ancestry estimation

In 2015, Chris Phillips described being asked to define a person of interest as "White, Black or Asian" by police in London some 17 years prior, following the witnessing of a crime [148]. Aside from the lack of more neutral terminology, he goes on to depict the situation as taking place at night, making his own evidence likely to be unreliable. Bio-geographical ancestry estimation from DNA recovered from a scene can provide a more reliable and unbiased alternative to eyewitness testimony, and has been used as a tool for investigative intelligence by police forces. There are a number of well characterised markers in our DNA which show enrichment in certain populations and are used specifically for ancestry estimation. Ancestry informative SNPs and non-core STR loci were discussed in this context, with the majority of marker panels for forensic ancestry estimation relying on the former.

The results presented from STRUCTURE analyses and group membership proportions suggest that data from the aSTRs present in the ForenSeq DNA Signature Prep kit have the potential to be used for ancestry estimation for five global populations. With a strict group membership proportion of 0.7 or above, 84% of samples were grouped correctly, with a general error rate of 3.5%. This is a major improvement on a previous publication looking at length-based data for 20 autosomal markers, which had error averages of 12-15% [170]. It is also an advance on results obtained using length-based data for the 26 ForenSeq STRs, where 72% of samples were classified correctly with an error rate of 5.7%. These results once again highlight the value added by massively parallel sequencing. An interesting point for future work would be to genotype a further set of samples from each of the global populations studied and blindly assign them using the data from this project as a training set in STRUCTURE, which would further test these findings.

Autosomal STRs are and likely always will be the first loci targeted for forensic DNA analysis, as they offer the highest probability of individual identification. An ancestry-informative marker panel can then be used in "no hit, no suspect" cases, but this requires an additional time and cost investment, and relies on the presence of enough sample. As routine DNA testing of autosomal STRs progresses to MPS, the results from this project show that there is now the very real possibility of getting both an individual's DNA profile and an estimation of their bio-geographic origin from one test. Combining aSNP and aSTR data didn't show any improvement on using a dedicated aSNP panel alone, but the ancestry inference potential of the STRs in the ForenSeq DNA Signature Prep kit is almost as good as an aSNP panel. It is likely that this ancestry prediction may also improve with wider population data sets, as suggested by the two North East African samples that showed membership to a sixth cluster, suggesting we are looking at sub-populations within the data.

6. Conclusions

When this PhD project started, very few forensic laboratories were equipped with massively parallel sequencing technology, and most projects relating to implementation were still in their infancy. The surge in literature in recent years highlights that the move from traditional DNA typing methods to MPS ones is happening on a global scale. The results presented early in this thesis demonstrate that genotypes obtained using sequencing are concordant with the capillary electrophoresis results for the same samples, ensuring back-compatibility with established DNA databases. One major barrier to implementation was described as a lack of available sequence-based allelic frequencies [129]. The output of this PhD complements other global databases to overcome such barriers by providing comprehensive, curated frequencies. At the time of writing, the article published in 2018 (Appendix I) has been cited over 50 times, further demonstrating the demand for such information. Research into MPS feasibility is the process driving adoption in forensic casework laboratories.

The huge increase in allelic diversity witnessed when sequencing STRs, rather than separating them based on length, brought about an additional set of questions and possible barriers regarding the number of samples needed to generate representative population frequencies, as well as how to name the sequence-based alleles identified. Nomenclature was discussed throughout the initial chapters of this thesis, and the field will continue to narrow down on the best system through collaborative discussions and increased availability of large-scale datasets. The usefulness of flanking regions of STRs was less than hoped in the context of power of discrimination, and the added complexity from a nomenclature perspective is substantial, but it is still vital to take these regions into account for proper compatibility with CE and between different MPS kits.

While the first two results chapters focussed on characterising five UK-relevant populations using MPS, the final chapter looked at expanding how we utilise sequence diversity to differentiate them. Whilst it is highly unlikely that bio-geographic ancestry estimation will completely move away from SNP panels, results show that there is useable, ancestry informative data in the sequenced genotypes of aSTR markers. Removing the need to return to the sample, or extract, to obtain more information is an attractive option for forensic scientists as it can reduce cost and time constraints.

The shift to MPS in forensics is still meeting some resistance for autosomal STR analysis due to the well-established, routine use of capillary electrophoresis for DNA identification, but an increase in validation and accreditation of MPS protocols is steadily breaking down the "it's how we've always done it" frame of mind. As the barriers to implementation come down, the advantages of sequencing continue to be highlighted. This PhD project has shown that an STR profile generated using the ForenSeq DNA Signature Prep kit is considerably more discriminatory than a CE one, and could feasibly also be used to predict ancestry for five global populations. If MPS does become the routine method for DNA typing in forensics, there is no doubt that further research, including an expansion in the number of global populations where high quality STR sequence data is available, will lead to more robust and reliable results for ancestry estimation.

Appendix I – Devesse *et al*. 2018, Forensic Science International: Genetics (34) 57–61

Appendix II – Devesse *et al*. 2020, Forensic Science International: Genetics (48) 102356

Appendix III – List of associated publications

Gettings KB, Borsuk LA, Ballard D, Bodner M, Budowle B, **Devesse L**, King J, Parson W, Phillips C and Vallone PM (2017), *'STRSeq: A catalog of sequence diversity at human identification Short Tandem Repeat loci*', Forensic Science International-Genetics, vol. 31, pp. 111-117. https://doi.org/10.1016/j.fsigen.2017.08.017

Devesse LA, Ballard DJ, Davenport LB, Gettings KB, Borsuk LA, Vallone PM and Syndercombe Court, D (2017), '*The tao of MPS: Common novel variants'*, Forensic Science International: Genetics Supplement Series. https://doi.org/10.1016/j.fsigss.2017.09.222

Devesse L, Ballard D, Davenport L, Riethorst I, Mason-Buck G and Court, DS (2018), *'Concordance of the ForenSeq™ system and characterisation of sequence-specific autosomal STR alleles across two major population groups*', Forensic Science International-Genetics, vol. 34, pp. 57-61. https://doi.org/10.1016/j.fsigen.2017.10.012

Phillips C, **Devesse L**, Ballard D, van Weert L, de la Puente M, Melis S, Álvarez Iglesias V, Freire-Aradas A, Oldroyd N, Holt C, Syndercombe Court D, Carracedo Á and Lareu MV (2018), *'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. https://doi.org/10.1002/elps.201800117

Devesse L, Davenport L, Borsuk L, Gettings K, Mason-Buck G, Vallone PM, Syndercombe Court D and Ballard D (2020), *'Classification of STR allelic variation using massively parallel sequencing and assessment of flanking region power*', Forensic Science International: Genetics, vol. 48, 102356. https://doi.org/10.1016/j.fsigen.2020.102356

References

1. Gill, P., A. J. Jeffreys, and D. J. Werrett, *Forensic application of DNA 'fingerprints'.* Nature, 1985. **318**(6046): p. 577-9.
2. International HapMap, Consortium, *The International HapMap Project.* Nature, 2003. **426**(6968): p. 789-96.
3. Albright, T. D., *Why eyewitnesses fail.* Proc Natl Acad Sci U S A, 2017. **114**(30): p. 7758-7764.
4. Garrett, Brandon, *Convicting the innocent*, in *Convicting the Innocent*. 2011, Harvard University Press.
5. Ambrose, Tom, *Child killer Colin Pitchfork released from prison*, in *The Guardian*. 2021: https://www.theguardian.com/law/2021/sep/01/child-killer-colin-pitchfork-released-from-prison.
6. Cobain, Ian, *Killer breakthrough – the day DNA evidence first nailed a murderer*, in *The Guardian*. 2016: https://www.theguardian.com/uk-news/2016/jun/07/killer-dna-evidence-genetic-profiling-criminal-investigation.
7. *DNA Resource - Forensic DNA Policy*. Available from: https://www.dnaresource.com/resources.
8. Jeffreys, A. J., V. Wilson, and S. L. Thein, *Hypervariable 'minisatellite' regions in human DNA.* Nature, 1985. **314**(6006): p. 67-73.
9. Butler, John M., *Advanced Topics in Forensic DNA Typing : Methodology*. 3rd ed. 2011, Burlington: Elsevier Science. 1 online resource (699 pages).
10. Jobling, M. A. and P. Gill, *Encoded evidence: DNA in forensic analysis.* Nature Reviews Genetics, 2004. **5**(10): p. 739-751.
11. Butler, J. M., *Short tandem repeat typing technologies used in human identity testing.* Biotechniques, 2007. **43**(4): p. ii-v.
12. Butler, J. M. and C. R. Hill, *Biology and Genetics of New Autosomal STR Loci Useful for Forensic DNA Analysis.* Forensic Sci Rev, 2012. **24**(1): p. 15-26.
13. Butler, John M. and John M. Butler, *Fundamentals of Forensic DNA Typing*. 2009, Elsevier Science,: San Diego. p. 1 online resource (519 p.).
14. Bassam, B. J., G. Caetano-Anolles, and P. M. Gresshoff, *Fast and sensitive silver staining of DNA in polyacrylamide gels.* Anal Biochem, 1991. **196**(1): p. 80-3.
15. Edwards, A., A. Civitello, H. A. Hammond, and C. T. Caskey, *DNA typing and genetic mapping with trimeric and tetrameric tandem repeats.* Am J Hum Genet, 1991. **49**(4): p. 746-56.
16. Frazier, R. R., E. S. Millican, S. K. Watson, N. J. Oldroyd, R. L. Sparkles, K. M. Taylor, S. Panchal, L. Bark, C. P. Kimpton, and P. D. Gill, *Validation of the Applied Biosystems Prism 377 automated sequencer for the forensic short tandem repeat analysis.* Electrophoresis, 1996. **17**(10): p. 1550-2.
17. Ruitberg, C. M., D. J. Reeder, and J. M. Butler, *STRBase: a short tandem repeat DNA database for the human identity testing community.* Nucleic Acids Res, 2001. **29**(1): p. 320-2.
18. Bar, W., B. Brinkmann, B. Budowle, A. Carracedo, P. Gill, P. Lincoln, W. Mayr, and B. Olaisen, *DNA recommendations. Further report of the DNA Commission of the ISFH regarding the use of short tandem repeat systems. International Society for Forensic Haemogenetics.* Int J Legal Med, 1997. **110**(4): p. 175-6.
19. in *The Evaluation of Forensic DNA Evidence*. 1996: Washington (DC).
20. Butler, John M., *Advanced topics in forensic DNA typing : methodology*. [3rd ed. 2012, Amsterdam ; Boston: Elsevier/Academic Press. xvii, 680 p.
21. Chakraborty, R., *Sample size requirements for addressing the population genetic issues of forensic use of DNA typing.* Hum Biol, 1992. **64**(2): p. 141-59.
22. *Forensic Science Regulator Guidance: Allele Frequency Databases and Reporting Guidance for the DNA (Short Tandem Repeat) Profiling*, FSR-G-213, Editor. 2020.
23. Martin, P. D., H. Schmitter, and P. M. Schneider, *A brief history of the formation of DNA databases in forensic science within Europe.* Forensic Sci Int, 2001. **119**(2): p. 225-31.
24. Gill, P., L. Fereday, N. Morling, and P. M. Schneider, *The evolution of DNA databases--recommendations for new European STR loci.* Forensic Sci Int, 2006. **156**(2-3): p. 242-4.
25. Gill, P., L. Fereday, N. Morling, and P. M. Schneider, *New multiplexes for Europe-amendments and clarification of strategic development.* Forensic Sci Int, 2006. **163**(1-2): p. 155-7.
26. Ge, J., R. Chakraborty, A. Eisenberg, and B. Budowle, *Comparisons of familial DNA database searching strategies.* J Forensic Sci, 2011. **56**(6): p. 1448-56.

27. Schneider, Peter M., *Expansion of the European Standard Set of DNA Database Loci—the Current Situation.* Profiles in DNA, 2009(March 2009).

28. Hares, D. R., *Expanding the CODIS core loci in the United States.* Forensic Sci Int Genet, 2012. **6**(1): p. e52-4.

29. Hares, D. R., *Selection and implementation of expanded CODIS core loci in the United States.* Forensic Sci Int Genet, 2015. **17**: p. 33-34.

30. Ludeman, M. J., C. Zhong, J. J. Mulero, R. E. Lagace, L. K. Hennessy, M. L. Short, and D. Y. Wang, *Developmental validation of GlobalFiler PCR amplification kit: a 6-dye multiplex assay designed for amplification of casework samples.* Int J Legal Med, 2018. **132**(6): p. 1555-1573.

31. Life_Technologies, *GlobalFiler™ Express PCR Amplification Kit User Guide.* Publication Part Number 4477672 Rev. A, 2012.

32. Promega, *PowerPlex® ESI 17 Pro System Technical Manual.* 2017.

33. QIAgen, *Investigator® 24plex QS Handbook.* 2016.

34. Wang, D. Y., S. Gopinath, R. E. Lagace, W. Norona, L. K. Hennessy, M. L. Short, and J. J. Mulero, *Developmental validation of the GlobalFiler((R)) Express PCR Amplification Kit: A 6-dye multiplex assay for the direct amplification of reference samples.* Forensic Sci Int Genet, 2015. **19**: p. 148-155.

35. Martin, P., L. F. de Simon, G. Luque, M. J. Farfan, and A. Alonso, *Improving DNA data exchange: validation studies on a single 6 dye STR kit with 24 loci.* Forensic Sci Int Genet, 2014. **13**: p. 68-78.

36. Oostdik, K., K. Lenz, J. Nye, K. Schelling, D. Yet, S. Bruski, J. Strong, C. Buchanan, J. Sutton, J. Linner*, et al.*, *Developmental validation of the PowerPlex((R)) Fusion System for analysis of casework and reference samples: A 24-locus multiplex for new database standards.* Forensic Sci Int Genet, 2014. **12**: p. 69-76.

37. Promega, *Spectrum CE System*. 2020.

38. Hughes-Stamm, S. R., K. J. Ashton, and A. van Daal, *Assessment of DNA degradation and the genotyping success of highly degraded samples.* Int J Legal Med, 2011. **125**(3): p. 341-8.

39. van Oorschot, R. A., K. N. Ballantyne, and R. J. Mitchell, *Forensic trace DNA: a review.* Investig Genet, 2010. **1**(1): p. 14.

40. Butler, J. M., Y. Shen, and B. R. McCord, *The development of reduced size STR amplicons as tools for analysis of degraded DNA.* J Forensic Sci, 2003. **48**(5): p. 1054-64.

41. Burrill, J., B. Daniel, and N. Frascione, *A review of trace "Touch DNA" deposits: Variability factors and an exploration of cellular composition.* Forensic Sci Int Genet, 2019. **39**: p. 8-18.

42. John M. Butler, Hari Iyer, Rich Press, Melissa K. Taylor, Peter M. Vallone, Sheila Willis, *DNA Mixture Interpretation: A NIST Scientific Foundation Review*, U.S.D.o. Commerce, Editor. 2021.

43. Isaacson, J., E. Schwoebel, A. Shcherbina, D. Ricke, J. Harper, M. Petrovick, J. Bobrow, T. Boettcher, B. Helfer, C. Zook*, et al.*, *Robust detection of individual forensic profiles in DNA mixtures.* Forensic Sci Int Genet, 2015. **14**: p. 31-7.

44. Novroski, N. M. M., F. R. Wendt, A. E. Woerner, M. M. Bus, M. Coble, and B. Budowle, *Expanding beyond the current core STR loci: An exploration of 73 STR markers with increased diversity for enhanced DNA mixture deconvolution.* Forensic Sci Int Genet, 2019. **38**: p. 121-129.

45. Genomes Project, Consortium, A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean*, et al.*, *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.

46. Gill, P., D. J. Werrett, B. Budowle, and R. Guerrieri, *An assessment of whether SNPs will replace STRs in national DNA databases--joint considerations of the DNA working group of the European Network of Forensic Science Institutes (ENFSI) and the Scientific Working Group on DNA Analysis Methods (SWGDAM).* Sci Justice, 2004. **44**(1): p. 51-3.

47. Kidd, K. K., A. J. Pakstis, W. C. Speed, E. L. Grigorenko, S. L. Kajuna, N. J. Karoma, S. Kungulilo, J. J. Kim, R. B. Lu, A. Odunsi*, et al.*, *Developing a SNP panel for forensic identification of individuals.* Forensic Sci Int, 2006. **164**(1): p. 20-32.

48. Yousefi, S., T. Abbassi-Daloii, T. Kraaijenbrink, M. Vermaat, H. Mei, P. van 't Hof, M. van Iterson, D. V. Zhernakova, A. Claringbould, L. Franke*, et al.*, *A SNP panel for identification of DNA and RNA specimens.* BMC Genomics, 2018. **19**(1): p. 90.

49. Sobrino, B., M. Brion, and A. Carracedo, *SNPs in forensic genetics: a review on SNP typing methodologies.* Forensic Sci Int, 2005. **154**(2-3): p. 181-94.

50. Sanchez, J. J., C. Phillips, C. Borsting, K. Balogh, M. Bogus, M. Fondevila, C. D. Harrison, E. Musgrave-Brown, A. Salas, D. Syndercombe-Court*, et al.*, *A multiplex assay with 52 single nucleotide polymorphisms for human identification.* Electrophoresis, 2006. **27**(9): p. 1713-24.

51. Daniel, R., C. Santos, C. Phillips, M. Fondevila, R. A. van Oorschot, A. Carracedo, M. V. Lareu, and D. McNevin, *A SNaPshot of next generation sequencing for forensic SNP analysis.* Forensic Sci Int Genet, 2015. **14**: p. 50-60.

52. Sanger, F., *Sequences, sequences, and sequences.* Annu Rev Biochem, 1988. **57**: p. 1-28.

53. Margulies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y. J. Chen, Z. Chen*, et al.*, *Genome sequencing in microfabricated high-density picolitre reactors.* Nature, 2005. **437**(7057): p. 376-80.

54. Goodwin, S., J. D. McPherson, and W. R. McCombie, *Coming of age: ten years of next-generation sequencing technologies.* Nat Rev Genet, 2016. **17**(6): p. 333-51.

55. Kayser, M. and W. Parson, *Transitioning from Forensic Genetics to Forensic Genomics.* Genes (Basel), 2017. **9**(1).

56. de Knijff, P., *From next generation sequencing to now generation sequencing in forensics.* Forensic Sci Int Genet, 2019. **38**: p. 175-180.

57. Dalsgaard, Sigrun; Rockenbauer, Eszter; Gelardi, Chiara; Børsting, Claus; Fordyce, Sarah Louise; Morling, Niels, *Characterization of mutations and sequence variations in complex STR loci by second generation sequencing.* Forensic Science International: Genetics. Supplement Series, 2013. **4**(1): p. e218-e219.

58. Fordyce, S. L., M. C. Avila-Arcos, E. Rockenbauer, C. Borsting, R. Frank-Hansen, F. T. Petersen, E. Willerslev, A. J. Hansen, N. Morling, and M. T. Gilbert, *High-throughput sequencing of core STR loci for forensic genetic investigations using the Roche Genome Sequencer FLX platform.* Biotechniques, 2011. **51**(2): p. 127-33.

59. Friis, S. L., A. Buchard, E. Rockenbauer, C. Borsting, and N. Morling, *Introduction of the Python script STRinNGS for analysis of STR regions in FASTQ or BAM files and expansion of the Danish STR sequence database to 11 STRs.* Forensic Sci Int Genet, 2016. **21**: p. 68-75.

60. Rockenbauer, E., S. Hansen, M. Mikkelsen, C. Borsting, and N. Morling, *Characterization of mutations and sequence variants in the D21S11 locus by next generation sequencing.* Forensic Sci Int Genet, 2014. **8**(1): p. 68-72.

61. Gelardi, C., E. Rockenbauer, S. Dalsgaard, C. Borsting, and N. Morling, *Second generation sequencing of three STRs D3S1358, D12S391 and D21S11 in Danes and a new nomenclature for sequenced STR alleles.* Forensic Sci Int Genet, 2014. **12**: p. 38-41.

62. Gettings, K. B., K. M. Kiesler, S. A. Faith, E. Montano, C. H. Baker, B. A. Young, R. A. Guerrieri, and P. M. Vallone, *Sequence variation of 22 autosomal STR loci detected by next generation sequencing.* Forensic Sci Int Genet, 2016. **21**: p. 15-21.

63. Novroski, N. M., J. L. King, J. D. Churchill, L. H. Seah, and B. Budowle, *Characterization of genetic sequence variation of 58 STR loci in four major population groups.* Forensic Sci Int Genet, 2016. **25**: p. 214-226.

64. Wendt, F. R., J. D. Churchill, N. M. Novroski, J. L. King, J. Ng, R. F. Oldt, K. L. McCulloh, J. A. Weise, D. G. Smith, S. Kanthaswamy*, et al.*, *Genetic analysis of the Yavapai Native Americans from West-Central Arizona using the Illumina MiSeq FGx forensic genomics system.* Forensic Sci Int Genet, 2016. **24**: p. 18-23.

65. Gettings, K. B., R. A. Aponte, P. M. Vallone, and J. M. Butler, *STR allele sequence variation: Current knowledge and future issues.* Forensic Sci Int Genet, 2015. **18**: p. 118-30.

66. Guo, F., J. Yu, L. Zhang, and J. Li, *Massively parallel sequencing of forensic STRs and SNPs using the Illumina((R)) ForenSeq DNA Signature Prep Kit on the MiSeq FGx Forensic Genomics System.* Forensic Sci Int Genet, 2017. **31**: p. 135-148.

67. Walsh, P. S., N. J. Fildes, and R. Reynolds, *Sequence analysis and characterization of stutter products at the tetranucleotide repeat locus vWA.* Nucleic Acids Res, 1996. **24**(14): p. 2807-12.

68. Bright, J. A., D. Taylor, J. M. Curran, and J. S. Buckleton, *Developing allelic and stutter peak height models for a continuous method of DNA interpretation.* Forensic Sci Int Genet, 2013. **7**(2): p. 296-304.

69. Borsting, C. and N. Morling, *Next generation sequencing and its applications in forensic genetics.* Forensic Sci Int Genet, 2015. **18**: p. 78-89.

70. Churchill, J. D., J. Chang, J. Ge, N. Rajagopalan, S. C. Wootton, C. W. Chang, R. Lagace, W. Liao, J. L. King, and B. Budowle, *Blind study evaluation illustrates utility of the Ion PGM system for use in human identity DNA typing.* Croat Med J, 2015. **56**(3): p. 218-29.

71. Zeng, X., J. King, S. Hermanson, J. Patel, D. R. Storts, and B. Budowle, *An evaluation of the PowerSeq Auto System: A multiplex short tandem repeat marker kit compatible with massively parallel sequencing.* Forensic Sci Int Genet, 2015. **19**: p. 172-179.

72. Fordyce, S. L., H. S. Mogensen, C. Borsting, R. E. Lagace, C. W. Chang, N. Rajagopalan, and N. Morling, *Second-generation sequencing of forensic STRs using the Ion Torrent HID STR 10-plex and the Ion PGM.* Forensic Sci Int Genet, 2015. **14**: p. 132-40.

73. Churchill, J. D., S. E. Schmedes, J. L. King, and B. Budowle, *Evaluation of the Illumina((R)) Beta Version ForenSeq DNA Signature Prep Kit for use in genetic profiling.* Forensic Sci Int Genet, 2016. **20**: p. 20-9.

74. Just, R. S., L. I. Moreno, J. B. Smerick, and J. A. Irwin, *Performance and concordance of the ForenSeq system for autosomal and Y chromosome short tandem repeat sequencing of reference-type specimens.* Forensic Sci Int Genet, 2017. **28**: p. 1-9.

75. Van Neste, C., F. Van Nieuwerburgh, D. Van Hoofstat, and D. Deforce, *Forensic STR analysis using massive parallel sequencing.* Forensic Sci Int Genet, 2012. **6**(6): p. 810-8.

76. Ballantyne, K. N., V. Keerl, A. Wollstein, Y. Choi, S. B. Zuniga, A. Ralf, M. Vermeulen, P. de Knijff, and M. Kayser, *A new future of forensic Y-chromosome analysis: rapidly mutating Y-STRs for differentiating male relatives and paternal lineages.* Forensic Sci Int Genet, 2012. **6**(2): p. 208-18.

77. Definis Gojanovic, M. and D. Sutlovic, *Skeletal remains from World War II mass grave: from discovery to identification.* Croat Med J, 2007. **48**(4): p. 520-7.

78. Purps, J., M. Geppert, M. Nagy, and L. Roewer, *Validation of a combined autosomal/Y-chromosomal STR approach for analyzing typical biological stains in sexual-assault cases.* Forensic Sci Int Genet, 2015. **19**: p. 238-242.

79. Khan, K., M. H. Siddiqi, M. Abbas, M. Almas, and M. Idrees, *Forensic applications of Y chromosomal properties.* Leg Med (Tokyo), 2017. **26**: p. 86-91.

80. Butler, Katherine, Michelle Peck, Jessica Hart, Moses Schanfield, and Daniele Podini, *Molecular "eyewitness": Forensic prediction of phenotype and ancestry.* Forensic Science International: Genetics Supplement Series, 2011. **3**(1): p. e498-e499.

81. Kayser, M., *Forensic DNA Phenotyping: Predicting human appearance from crime scene material for investigative purposes.* Forensic Sci Int Genet, 2015. **18**: p. 33-48.

82. Hollard, C., C. Keyser, T. Delabarde, A. Gonzalez, C. Vilela Lamego, V. Zvenigorosky, and B. Ludes, *Case report: on the use of the HID-Ion AmpliSeq Ancestry Panel in a real forensic case.* Int J Legal Med, 2017. **131**(2): p. 351-358.

83. Ambers, A. D., J. D. Churchill, J. L. King, M. Stoljarova, H. Gill-King, M. Assidi, M. Abu-Elmagd, A. Buhmeida, M. Al-Qahtani, and B. Budowle, *More comprehensive forensic genetic marker analyses for accurate human remains identification using massively parallel DNA sequencing.* BMC Genomics, 2016. **17**(Suppl 9): p. 750.

84. Barlow, Vicky, *The development of enhanced experimental strategies for the DNA analysis of low-template or compromised forensic sample types*, in *Health and Life Sciences*. 2015, University of Northumbria: Newcastle.

85. Chaitanya, L., I. Z. Pajnic, S. Walsh, J. Balazic, T. Zupanc, and M. Kayser, *Bringing colour back after 70 years: Predicting eye and hair colour from skeletal remains of World War II victims using the HIrisPlex system.* Forensic Sci Int Genet, 2017. **26**: p. 48-57.

86. Draus-Barini, J., S. Walsh, E. Pospiech, T. Kupiec, H. Glab, W. Branicki, and M. Kayser, *Bona fide colour: DNA prediction of human eye and hair colour from ancient and contemporary skeletal remains.* Investig Genet, 2013. **4**(1): p. 3.

87. Eduardoff, M., T. E. Gross, C. Santos, M. de la Puente, D. Ballard, C. Strobl, C. Borsting, N. Morling, L. Fusco, C. Hussing*, et al.*, *Inter-laboratory evaluation of the EUROFORGEN Global ancestry-informative SNP panel by massively parallel sequencing using the Ion PGM.* Forensic Sci Int Genet, 2016. **23**: p. 178-189.

88. Li, C. X., A. J. Pakstis, L. Jiang, Y. L. Wei, Q. F. Sun, H. Wu, O. Bulbul, P. Wang, L. L. Kang, J. R. Kidd*, et al.*, *A panel of 74 AISNPs: Improved ancestry inference within Eastern Asia.* Forensic Sci Int Genet, 2016. **23**: p. 101-110.

89. Silvia, A. L., N. Shugarts, and J. Smith, *A preliminary assessment of the ForenSeq FGx System: next generation sequencing of an STR and SNP multiplex.* Int J Legal Med, 2017. **131**(1): p. 73-86.

90. Warshauer, D. H., C. P. Davis, C. Holt, Y. Han, P. Walichiewicz, T. Richardson, K. Stephens, A. Jager, J. King, and B. Budowle, *Massively parallel sequencing of forensically relevant single nucleotide polymorphisms using TruSeq forensic amplicon.* Int J Legal Med, 2015. **129**(1): p. 31-6.

91. Dixon, L. A., A. E. Dobbins, H. K. Pulker, J. M. Butler, P. M. Vallone, M. D. Coble, W. Parson, B. Berger, P. Grubwieser, H. S. Mogensen*, et al.*, *Analysis of artificially degraded DNA using STRs*

and SNPs--results of a collaborative European (EDNAP) exercise. Forensic Sci Int, 2006. **164**(1): p. 33-44.

92. Calafell, F., R. Anglada, N. Bonet, M. Gonzalez-Ruiz, G. Prats-Munoz, R. Rasal, C. Lalueza-Fox, J. Bertranpetit, A. Malgosa, and 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, 2016. **37**(21): p. 2841-2847.
93. Thanakiatkrai, P., S. Phetpeng, S. Sotthibandhu, W. Asawutmangkul, Y. Piwpankaew, J. E. Foong, J. Koo, and T. Kitpipit, *Performance comparison of MiSeq forensic genomics system and STR-CE using control and mock IED samples.* Forensic Science International Genetics Supplement Series, 2017. **6**: p. E320-E321.
94. Borsting, C., H. S. Mogensen, and N. Morling, *Forensic genetic SNP typing of low-template DNA and highly degraded DNA from crime case samples.* Forensic Sci Int Genet, 2013. **7**(3): p. 345-52.
95. Devesse, Laurence, David Ballard, Lucinda Davenport, Immy Riethorst, Gabriella Mason-Buck, and Denise Syndercombe Court, *Concordance of the ForenSeq™ system and characterisation of sequence-specific autosomal STR alleles across two major population groups.* Forensic Science International-Genetics, 2017.
96. Wendt, F. R., 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*, et al.*, *Flanking region variation of ForenSeq DNA Signature Prep Kit STR and SNP loci in Yavapai Native Americans.* Forensic Sci Int Genet, 2017. **28**: p. 146-154.
97. Wu, R., D. Peng, H. Ren, R. Li, H. Li, N. Wang, X. Shen, E. Huang, Y. Zhang, and H. Sun, *Characterization of genetic polymorphisms in Nigerians residing in Guangzhou using massively parallel sequencing.* Forensic Sci Int Genet, 2020. **48**: p. 102323.
98. Gettings, K. B., L. A. Borsuk, C. R. Steffen, K. M. Kiesler, and P. M. Vallone, *Sequence-based U.S. population data for 27 autosomal STR loci.* Forensic Sci Int Genet, 2018. **37**: p. 106-115.
99. Illumina, *Illumina Sequencing Platforms Brochure*. 2019.
100. Verogen, *ForenSeqTM DNA Signature Prep Reference Guide.* Document #VD2018005 Rev. A, 2018.
101. McLaren, R. S., J. Patel, M. M. Ewing, D. R. Storts, F. Noel, S. Dognaux, C. R. Hill, M. C. Kline, and J. M. Butler, *Developmental validation of the PowerPlex(R) ESI 17 Pro System.* Forensic Sci Int Genet, 2013. **7**(3): p. e69-73.
102. Dixon, L. A., C. M. Murray, E. J. Archer, A. E. Dobbins, P. Koumi, and P. Gill, *Validation of a 21-locus autosomal SNP multiplex for forensic identification purposes.* Forensic Sci Int, 2005. **154**(1): p. 62-77.
103. Kidd, K. K., J. R. Kidd, W. C. Speed, R. Fang, M. R. Furtado, F. C. Hyland, and A. J. Pakstis, *Expanding data and resources for forensic use of SNPs in individual identification.* Forensic Sci Int Genet, 2012. **6**(5): p. 646-52.
104. Nievergelt, C. M., A. X. Maihofer, T. Shekhtman, O. Libiger, X. Wang, K. K. Kidd, and J. R. Kidd, *Inference of human continental origin and admixture proportions using a highly discriminative ancestry informative 41-SNP panel.* Investig Genet, 2013. **4**(1): p. 13.
105. Kidd, K. K., W. C. Speed, A. J. Pakstis, M. R. Furtado, R. Fang, A. Madbouly, M. Maiers, M. Middha, F. R. Friedlaender, and J. R. Kidd, *Progress toward an efficient panel of SNPs for ancestry inference.* Forensic Sci Int Genet, 2014. **10**: p. 23-32.
106. Walsh, S., L. Chaitanya, L. Clarisse, L. Wirken, J. Draus-Barini, L. Kovatsi, H. Maeda, T. Ishikawa, T. Sijen, P. de Knijff*, et al.*, *Developmental validation of the HIrisPlex system: DNA-based eye and hair colour prediction for forensic and anthropological usage.* Forensic Sci Int Genet, 2014. **9**: p. 150-61.
107. Walsh, S., F. Liu, K. N. Ballantyne, M. van Oven, O. Lao, and M. Kayser, *IrisPlex: a sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information.* Forensic Sci Int Genet, 2011. **5**(3): p. 170-80.
108. Jager, A. C., M. L. Alvarez, C. P. Davis, E. Guzman, Y. Han, L. Way, P. Walichiewicz, D. Silva, N. Pham, G. Caves*, et al.*, *Developmental validation of the MiSeq FGx Forensic Genomics System for Targeted Next Generation Sequencing in Forensic DNA Casework and Database Laboratories.* Forensic Sci Int Genet, 2017. **28**: p. 52-70.
109. Illumina, *ForenSeq™ DNA Signature Prep Reference Guide.* Document #15049528 v01, 2015.
110. Zhang, Q., Z. Zhou, Q. Liu, L. Liu, L. Shao, M. Zhang, X. Ding, Y. Gao, and S. Wang, *Evaluation of the performance of Illumina's ForenSeq system on serially degraded samples.* Electrophoresis, 2018. **39**(21): p. 2674-2684.

111. Carrasco, P., C. Inostroza, M. Didier, M. Godoy, C. L. Holt, J. Tabak, and A. Loftus, *Optimizing DNA recovery and forensic typing of degraded blood and dental remains using a specialized extraction method, comprehensive qPCR sample characterization, and massively parallel sequencing.* Int J Legal Med, 2020. **134**(1): p. 79-91.

112. Hill, C. R., M. C. Kline, J. J. Mulero, R. E. Lagace, C. W. Chang, L. K. Hennessy, and J. M. Butler, *Concordance study between the AmpFlSTR MiniFiler PCR amplification kit and conventional STR typing kits.* J Forensic Sci, 2007. **52**(4): p. 870-3.

113. Mulero, J. J., C. W. Chang, R. E. Lagace, D. Y. Wang, J. L. Bas, T. P. McMahon, and L. K. Hennessy, *Development and validation of the AmpFlSTR MiniFiler PCR Amplification Kit: a MiniSTR multiplex for the analysis of degraded and/or PCR inhibited DNA.* J Forensic Sci, 2008. **53**(4): p. 838-52.

114. Verogen, *MiSeq FGx™ Instrument Reference Guide.* Document # VD2018006 Rev. A, 2018.

115. Xavier, C. and W. Parson, *Evaluation of the Illumina ForenSeq DNA Signature Prep Kit - MPS forensic application for the MiSeq FGx benchtop sequencer.* Forensic Sci Int Genet, 2017. **28**: p. 188-194.

116. Hussing, C., C. Borsting, H. S. Mogensen, and N. Morling, *Testing of the Illumina (R) ForenSeq (TM) kit.* Forensic Science International Genetics Supplement Series, 2015. **5**: p. E449-E450.

117. Almalki, N., H. Y. Chow, V. Sharma, K. Hart, D. Siegel, and E. Wurmbach, *Systematic assessment of the performance of illumina's MiSeq FGx forensic genomics system.* Electrophoresis, 2017. **38**(6): p. 846-854.

118. Kocher, S., P. Muller, B. Berger, M. Bodner, W. Parson, L. Roewer, S. Willuweit, and D. NASeqEx Consortium, *Inter-laboratory validation study of the ForenSeq DNA Signature Prep Kit.* Forensic Sci Int Genet, 2018. **36**: p. 77-85.

119. Hollard, C., L. Ausset, Y. Chantrel, S. Jullien, M. Clot, M. Faivre, E. Suzanne, L. Pene, and F. X. Laurent, *Automation and developmental validation of the ForenSeq() DNA Signature Preparation kit for high-throughput analysis in forensic laboratories.* Forensic Sci Int Genet, 2019. **40**: p. 37-45.

120. Moreno, L. I., M. B. Galusha, and R. Just, *A closer look at Verogen's Forenseq DNA Signature Prep kit autosomal and Y-STR data for streamlined analysis of routine reference samples.* Electrophoresis, 2018. **39**(21): p. 2685-2693.

121. Bentley, D. R., S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, D. J. Evers, C. L. Barnes, H. R. Bignell*, et al.*, *Accurate whole human genome sequencing using reversible terminator chemistry.* Nature, 2008. **456**(7218): p. 53-9.

122. Verogen, *ForenSeq™ Universal Analysis Software Guide.* Document #VD2018007 Rev. A, 2018.

123. Ravi, R. K., K. Walton, and M. Khosroheidari, *MiSeq: A Next Generation Sequencing Platform for Genomic Analysis.* Methods Mol Biol, 2018. **1706**: p. 223-232.

124. Schirmer, M., U. Z. Ijaz, R. D'Amore, N. Hall, W. T. Sloan, and C. Quince, *Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform.* Nucleic Acids Res, 2015. **43**(6): p. e37.

125. Smith, T. F. and M. S. Waterman, *Identification of common molecular subsequences.* J Mol Biol, 1981. **147**(1): p. 195-7.

126. Needleman, S. B. and C. D. Wunsch, *A general method applicable to the search for similarities in the amino acid sequence of two proteins.* J Mol Biol, 1970. **48**(3): p. 443-53.

127. Warshauer, David H., Jonathan L. King, and Bruce Budowle, *STRait Razor v2.0: The improved STR Allele Identification Tool – Razor.* Forensic Science International: Genetics, 2015. **14**(Supplement C): p. 182-186.

128. Hoogenboom, Jerry, Titia Sijen, and Kristiaan J. van der Gaag, *STRNaming: Generating simple, informative names for sequenced STR alleles in a standardised and automated manner.* Forensic Science International: Genetics, 2021. **52**: p. 102473.

129. Alonso, A., P. Muller, L. Roewer, S. Willuweit, B. Budowle, and W. Parson, *European survey on forensic applications of massively parallel sequencing.* Forensic Sci Int Genet, 2017. **29**: p. e23-e25.

130. Alonso, A., P. A. Barrio, P. Muller, S. Kocher, B. Berger, P. Martin, M. Bodner, S. Willuweit, W. Parson, L. Roewer*, et al.*, *Current state-of-art of STR sequencing in forensic genetics.* Electrophoresis, 2018. **39**(21): p. 2655-2668.

131. Hill, C. R., D. L. Duewer, M. C. Kline, C. J. Sprecher, R. S. McLaren, D. R. Rabbach, B. E. Krenke, M. G. Ensenberger, P. M. Fulmer, D. R. Storts*, et al.*, *Concordance and population studies along with stutter and peak height ratio analysis for the PowerPlex (R) ESX 17 and ESI 17 Systems.* Forensic Sci Int Genet, 2011. **5**(4): p. 269-75.

132. Hill, C. R., M.C Kline, D.L. Duewer, and J.M. Butler, *Concordance testing comparing STR multiplex kits with a standard data set.* Forensic Science International: Genetics Supplement Series Supplement Series, 2011. **3**(1): p. e188-e189.

133. Desjardins, Jeff. *The World's 7.5 Billion People, in One Chart*. Visual Capitalist 2019 [cited 2021 11/10]; Available from: https://www.visualcapitalist.com/worlds-7-5-billion-people-chart/.

134. Hussing, C., C. Huber, R. Bytyci, H. S. Mogensen, N. Morling, and C. Borsting, *Sequencing of 231 forensic genetic markers using the MiSeq FGx forensic genomics system - an evaluation of the assay and software.* Forensic Sci Res, 2018. **3**(2): p. 111-123.

135. Parson, W., D. Ballard, B. Budowle, J. M. Butler, K. B. Gettings, P. Gill, L. Gusmao, D. R. Hares, J. A. Irwin, J. L. King*, et al.*, *Massively parallel sequencing of forensic STRs: Considerations of the DNA commission of the International Society for Forensic Genetics (ISFG) on minimal nomenclature requirements.* Forensic Sci Int Genet, 2016. **22**: p. 54-63.

136. Gettings, K. B., D. Ballard, M. Bodner, L. A. Borsuk, J. L. King, W. Parson, and C. Phillips, *Report from the STRAND Working Group on the 2019 STR sequence nomenclature meeting.* Forensic Sci Int Genet, 2019. **43**: p. 102165.

137. Young, B., T. Faris, and L. Armogida, *A nomenclature for sequence-based forensic DNA analysis.* Forensic Sci Int Genet, 2019. **42**: p. 14-20.

138. Just, R. S. and J. A. Irwin, *Use of the LUS in sequence allele designations to facilitate probabilistic genotyping of NGS-based STR typing results.* Forensic Sci Int Genet, 2018. **34**: p. 197-205.

139. Gettings, K. B., L. A. Borsuk, D. Ballard, M. Bodner, B. Budowle, L. Devesse, J. King, W. Parson, C. Phillips, and P. M. Vallone, *STRSeq: A catalog of sequence diversity at human identification Short Tandem Repeat loci.* Forensic Sci Int Genet, 2017. **31**: p. 111-117.

140. Londin, E. R., M. A. Keller, C. Maista, G. Smith, L. A. Mamounas, R. Zhang, S. J. Madore, K. Gwinn, and R. A. Corriveau, *CoAIMs: a cost-effective panel of ancestry informative markers for determining continental origins.* PLoS One, 2010. **5**(10): p. e13443.

141. Tishkoff, S. A. and K. K. Kidd, *Implications of biogeography of human populations for 'race' and medicine.* Nat Genet, 2004. **36**(11 Suppl): p. S21-7.

142. Fejerman, L. and E. Ziv, *Population differences in breast cancer severity.* Pharmacogenomics, 2008. **9**(3): p. 323-33.

143. Davis, T. M., *Ethnic diversity in type 2 diabetes.* Diabet Med, 2008. **25 Suppl 2**: p. 52-6.

144. Pritchard, J. K. and N. A. Rosenberg, *Use of unlinked genetic markers to detect population stratification in association studies.* Am J Hum Genet, 1999. **65**(1): p. 220-8.

145. Kidd, J. R., F. R. Friedlaender, W. C. Speed, A. J. Pakstis, F. M. De La Vega, and K. K. Kidd, *Analyses of a set of 128 ancestry informative single-nucleotide polymorphisms in a global set of 119 population samples.* Investig Genet, 2011. **2**(1): p. 1.

146. Phillips, C., A. Salas, J. J. Sanchez, M. Fondevila, A. Gomez-Tato, J. Alvarez-Dios, M. Calaza, M. C. de Cal, D. Ballard, M. V. Lareu*, et al.*, *Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs.* Forensic Sci Int Genet, 2007. **1**(3-4): p. 273-80.

147. Walsh, S., F. Liu, A. Wollstein, L. Kovatsi, A. Ralf, A. Kosiniak-Kamysz, W. Branicki, and M. Kayser, *The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA.* Forensic Sci Int Genet, 2013. **7**(1): p. 98-115.

148. Phillips, C., *Forensic genetic analysis of bio-geographical ancestry.* Forensic Sci Int Genet, 2015. **18**: p. 49-65.

149. Bulbul, O. and G. Filoglu, *Development of a SNP panel for predicting biogeographical ancestry and phenotype using massively parallel sequencing.* Electrophoresis, 2018. **39**(21): p. 2743-2751.

150. Bulbul, O., A. J. Pakstis, U. Soundararajan, C. Gurkan, J. E. Brissenden, J. M. Roscoe, B. Evsanaa, A. Togtokh, P. Paschou, E. L. Grigorenko*, et al.*, *Ancestry inference of 96 population samples using microhaplotypes.* Int J Legal Med, 2018. **132**(3): p. 703-711.

151. Kidd, K. K., A. J. Pakstis, W. C. Speed, R. Lagace, J. Chang, S. Wootton, E. Haigh, and J. R. Kidd, *Current sequencing technology makes microhaplotypes a powerful new type of genetic marker for forensics.* Forensic Sci Int Genet, 2014. **12**: p. 215-24.

152. Moriot, A., C. Santos, A. Freire-Aradas, C. Phillips, and D. Hall, *Inferring biogeographic ancestry with compound markers of slow and fast evolving polymorphisms.* Eur J Hum Genet, 2018. **26**(11): p. 1697-1707.

153. Phillips, C., L. Fernandez-Formoso, M. Gelabert-Besada, M. Garcia-Magarinos, C. Santos, M. Fondevila, A. Carracedo, and M. V. Lareu, *Development of a novel forensic STR multiplex for ancestry analysis and extended identity testing.* Electrophoresis, 2013. **34**(8): p. 1151-62.

154. Messina, F., T. Di Corcia, M. Ragazzo, C. Sanchez Mellado, I. Contini, P. Malaspina, B. M. Ciminelli, O. Rickards, and C. Jodice, *Signs of continental ancestry in urban populations of Peru through autosomal STR loci and mitochondrial DNA typing.* PLoS One, 2018. **13**(7): p. e0200796.

155. Algee-Hewitt, B. F., M. D. Edge, J. Kim, J. Z. Li, and N. A. Rosenberg, *Individual Identifiability Predicts Population Identifiability in Forensic Microsatellite Markers.* Curr Biol, 2016. **26**(7): p. 935-42.

156. Norton, H. L., R. A. Kittles, E. Parra, P. McKeigue, X. Mao, K. Cheng, V. A. Canfield, D. G. Bradley, B. McEvoy, and M. D. Shriver, *Genetic evidence for the convergent evolution of light skin in Europeans and East Asians.* Mol Biol Evol, 2007. **24**(3): p. 710-22.

157. Tian, C., R. Kosoy, R. Nassir, A. Lee, P. Villoslada, L. Klareskog, L. Hammarstrom, H. J. Garchon, A. E. Pulver, M. Ransom*, et al.*, *European population genetic substructure: further definition of ancestry informative markers for distinguishing among diverse European ethnic groups.* Mol Med, 2009. **15**(11-12): p. 371-83.

158. Kidd, J. R., F. Friedlaender, A. J. Pakstis, M. Furtado, R. Fang, X. Wang, C. M. Nievergelt, and K. K. Kidd, *Single nucleotide polymorphisms and haplotypes in Native American populations.* Am J Phys Anthropol, 2011. **146**(4): p. 495-502.

159. Paschou, P., J. Lewis, A. Javed, and P. Drineas, *Ancestry informative markers for fine-scale individual assignment to worldwide populations.* J Med Genet, 2010. **47**(12): p. 835-47.

160. Nassir, R., R. Kosoy, C. Tian, P. A. White, L. M. Butler, G. Silva, R. Kittles, M. E. Alarcon-Riquelme, P. K. Gregersen, J. W. Belmont*, et al.*, *An ancestry informative marker set for determining continental origin: validation and extension using human genome diversity panels.* BMC Genet, 2009. **10**: p. 39.

161. Phillips, C., A. Freire Aradas, A. K. Kriegel, M. Fondevila, O. Bulbul, C. Santos, F. Serrulla Rech, M. D. Perez Carceles, A. Carracedo, P. M. Schneider*, et al.*, *Eurasiaplex: a forensic SNP assay for differentiating European and South Asian ancestries.* Forensic Sci Int Genet, 2013. **7**(3): p. 359-66.

162. *AncestryDNA® – Frequently Asked Questions*. [cited 2021 12/10]; Available from: www.Ancestry.com.

163. Brown, Kristen. *How DNA Testing Botched My Family's Heritage, and Probably Yours, Too*. 2018 [cited 2021 12/10]; Available from: https://gizmodo.com/how-dna-testing-botched-my-familys-heritage-and-probab-1820932637.

164. Letzter, Rafi. *I took 9 different DNA tests and here's what I found*. 2018; Available from: https://www.livescience.com/63997-dna-ancestry-test-results-explained.html.

165. Charlsie Agro, Luke Denne. *Twins get some 'mystifying' results when they put 5 DNA ancestry kits to the test*. 2019 [cited 2021 12/10].

166. Krjutskov, K., T. Viltrop, P. Palta, E. Metspalu, E. Tamm, S. Suvi, K. Sak, A. Merilo, H. Sork, R. Teek*, et al.*, *Evaluation of the 124-plex SNP typing microarray for forensic testing.* Forensic Sci Int Genet, 2009. **4**(1): p. 43-8.

167. Rosenberg, N. A., J. K. Pritchard, J. L. Weber, H. M. Cann, K. K. Kidd, L. A. Zhivotovsky, and M. W. Feldman, *Genetic structure of human populations.* Science, 2002. **298**(5602): p. 2381-5.

168. Zhao, Z., W. Wen, K. Michailidou, M. K. Bolla, Q. Wang, B. Zhang, J. Long, X. O. Shu, M. K. Schmidt, R. L. Milne*, et al.*, *Association of genetic susceptibility variants for type 2 diabetes with breast cancer risk in women of European ancestry.* Cancer Causes Control, 2016. **27**(5): p. 679-93.

169. Lowe, A. L., A. Urquhart, L. A. Foreman, and I. W. Evett, *Inferring ethnic origin by means of an STR profile.* Forensic Sci Int, 2001. **119**(1): p. 17-22.

170. Phillips, C., L. Fernandez-Formoso, M. Garcia-Magarinos, L. Porras, T. Tvedebrink, J. Amigo, M. Fondevila, A. Gomez-Tato, J. Alvarez-Dios, A. Freire-Aradas*, et al.*, *Analysis of global variability in 15 established and 5 new European Standard Set (ESS) STRs using the CEPH human genome diversity panel.* Forensic Sci Int Genet, 2011. **5**(3): p. 155-69.

171. Pritchard, J. K., M. Stephens, and P. Donnelly, *Inference of population structure using multilocus genotype data.* Genetics, 2000. **155**(2): p. 945-59.

172. West, F. L. and B. F. B. Algee-Hewitt, *Cadaveric blood cards: Assessing DNA quality and quantity and the utility of STRs for the individual estimation of trihybrid ancestry and admixture proportions.* Forensic Sci Int, 2020. **2**: p. 114-122.

173. Pereira, L., F. Alshamali, R. Andreassen, R. Ballard, W. Chantratita, N. S. Cho, C. Coudray, J. M. Dugoujon, M. Espinoza, F. Gonzalez-Andrade*, et al.*, *PopAffiliator: online calculator for individual affiliation to a major population group based on 17 autosomal short tandem repeat genotype profile.* Int J Legal Med, 2011. **125**(5): p. 629-36.

174. Gusmao, L., J. M. Butler, A. Linacre, W. Parson, L. Roewer, P. M. Schneider, and A. Carracedo, *Revised guidelines for the publication of genetic population data.* Forensic Sci Int Genet, 2017. **30**: p. 160-163.

175. Bodner, M., I. Bastisch, J. M. Butler, R. Fimmers, P. Gill, L. Gusmao, N. Morling, C. Phillips, M. Prinz, P. M. Schneider*, et al.*, *Recommendations of the DNA Commission of the International Society for Forensic Genetics (ISFG) on quality control of autosomal Short Tandem Repeat allele frequency databasing (STRidER).* Forensic Sci Int Genet, 2016. **24**: p. 97-102.

176. Walsh, P. S., D. A. Metzger, and R. Higuchi, *Chelex 100 as a medium for simple extraction of DNA for PCR-based typing from forensic material.* Biotechniques, 1991. **10**(4): p. 506-13.

177. Ensenberger, M. G., J. Thompson, B. Hill, K. Homick, V. Kearney, K. A. Mayntz-Press, P. Mazur, A. McGuckian, J. Myers, K. Raley*, et al.*, *Developmental validation of the PowerPlex 16 HS System: an improved 16-locus fluorescent STR multiplex.* Forensic Sci Int Genet, 2010. **4**(4): p. 257-64.

178. Verogen, *Introducing the MiSeq FGx Reagent Micro Kit for Forensic Genomics*. 2021.

179. Untergasser, A., I. Cutcutache, T. Koressaar, J. Ye, B. C. Faircloth, M. Remm, and S. G. Rozen, *Primer3--new capabilities and interfaces.* Nucleic Acids Res, 2012. **40**(15): p. e115.

180. Hill, C. R., M. C. Kline, M. D. Coble, and J. M. Butler, *Characterization of 26 miniSTR loci for improved analysis of degraded DNA samples.* J Forensic Sci, 2008. **53**(1): p. 73-80.

181. Churchill, J. D., N. M. M. Novroski, J. L. King, L. H. Seah, and B. Budowle, *Population and performance analyses of four major populations with Illumina's FGx Forensic Genomics System.* Forensic Sci Int Genet, 2017. **30**: p. 81-92.

182. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.

183. Robinson, J. T., H. Thorvaldsdottir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, and J. P. Mesirov, *Integrative genomics viewer.* Nat Biotechnol, 2011. **29**(1): p. 24-6.

184. Phillips, C., K. B. Gettings, J. L. King, D. Ballard, M. Bodner, L. Borsuk, and W. Parson, *"The devil's in the detail": Release of an expanded, enhanced and dynamically revised forensic STR Sequence Guide.* Forensic Sci Int Genet, 2018. **34**: p. 162-169.

185. *Database of Single Nucleotide Polymorphisms (dbSNP)*. [cited 2019 9th of October]; Available from: https://www.ncbi.nlm.nih.gov/snp/rs11642858.

186. Kling, D., A. O. Tillmar, and T. Egeland, *Familias 3 - Extensions and new functionality.* Forensic Sci Int Genet, 2014. **13**: p. 121-7.

187. Egeland, T., P. F. Mostad, B. Mevag, and M. Stenersen, *Beyond traditional paternity and identification cases. Selecting the most probable pedigree.* Forensic Sci Int, 2000. **110**(1): p. 47-59.

188. Excoffier, L. and H. E. Lischer, *Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows.* Mol Ecol Resour, 2010. **10**(3): p. 564-7.

189. Laurent Excoffier, Heidi Lischer, *Arlequin version 3.5 User Manual - An Integrated Software Package for Population Genetics Data Analysis*. 2015, Swiss Institute of Bioinformatics.

190. Haynes, Winston, *Bonferroni Correction*, in *Encyclopedia of Systems Biology*, W. Dubitzky, O. Wolkenhauer, K.-H. Cho, and H. Yokota, Editors. 2013, Springer New York: New York, NY. p. 154-154.

191. Ristow, Peter G.; D'amato, Maria E., *Forensic statistics analysis toolbox (FORSTAT): A streamlined workflow for forensic statistics.* Forensic Science International: Genetics Supplement Series, 2017: p. e52-e54.

192. Raymond, M. and F. Rousset, *GENEPOP (Version 1.2): Population Genetics Software for Exact Tests and Ecumenicism.* Journal of Heredity, 1995. **86**(3): p. 248-249.

193. RStudio_Team. *RStudio: Integrated Development for R. RStudio, PBC, Boston, MA*. 2020; Available from: http://www.rstudio.com/.

194. Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and Subgroup Genome Project Data Processing, *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-9.

195. McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly*, et al.*, *The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.* Genome Res, 2010. **20**(9): p. 1297-303.

196. Peakall, R. and P. E. Smouse, *GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research--an update.* Bioinformatics, 2012. **28**(19): p. 2537-9.

197. Kopelman, N. M., J. Mayzel, M. Jakobsson, N. A. Rosenberg, and I. Mayrose, *Clumpak: a program for identifying clustering modes and packaging population structure inferences across K.* Mol Ecol Resour, 2015. **15**(5): p. 1179-91.

198. Rosenberg, N. A., L. M. Li, R. Ward, and J. K. Pritchard, *Informativeness of genetic markers for inference of ancestry.* Am J Hum Genet, 2003. **73**(6): p. 1402-22.

199. Churchill, J. D., N. M. M. Novroski, J. L. King, L. H. Seah, and B. Budowle, *Erratum to "Population and performance analyses of four major populations with Illumina's FGx Forensic Genomics System" [Forensic Sci. Int.: Genet. 30 (2017) 81-92].* Forensic Sci Int Genet, 2018. **33**: p. e17.

200. Kline, M. C., C. R. Hill, A. E. Decker, and J. M. Butler, *STR sequence analysis for characterizing normal, variant, and null alleles.* Forensic Sci Int Genet, 2011. **5**(4): p. 329-32.

201. Allor, C., D. D. Einum, and M. Scarpetta, *Identification and characterization of variant alleles at CODIS STR loci.* J Forensic Sci, 2005. **50**(5): p. 1128-33.

202. Krenke, B. E., A. Tereba, S. J. Anderson, E. Buel, S. Culhane, C. J. Finis, C. S. Tomsey, J. M. Zachetti, A. Masibay, D. R. Rabbach*, et al.*, *Validation of a 16-locus fluorescent multiplex system.* J Forensic Sci, 2002. **47**(4): p. 773-85.

203. Lane, A. B., *The nature of tri-allelic TPOX genotypes in African populations.* Forensic Sci Int Genet, 2008. **2**(2): p. 134-7.

204. Picanco, J. B., P. E. Raimann, Chasd Motta, R. Rodenbusch, L. Gusmao, and C. S. Alho, *Identification of the third/extra allele for forensic application in cases with TPOX tri-allelic pattern.* Forensic Sci Int Genet, 2015. **16**: p. 88-93.

205. King, J. L., A. E. Woerner, S. N. Mandape, K. B. Kapema, R. S. Moura-Neto, R. Silva, and B. Budowle, *STRait Razor Online: An enhanced user interface to facilitate interpretation of MPS data.* Forensic Sci Int Genet, 2021. **52**: p. 102463.

206. Hoogenboom, J., K. J. van der Gaag, R. H. de Leeuw, T. Sijen, P. de Knijff, and J. F. Laros, *FDSTools: A software package for analysis of massively parallel sequencing data with the ability to recognise and correct STR stutter and other PCR or sequencing noise.* Forensic Sci Int Genet, 2017. **27**: p. 27-40.

207. Zhang, S., Y. Niu, Y. Bian, R. Dong, X. Liu, Y. Bao, C. Jin, H. Zheng, and C. Li, *Sequence investigation of 34 forensic autosomal STRs with massively parallel sequencing.* Sci Rep, 2018. **8**(1): p. 6810.

208. Muller, P., C. Sell, T. Hadrys, J. Hedman, S. Bredemeyer, F. X. Laurent, L. Roewer, S. Achtruth, M. Sidstedt, T. Sijen*, et al.*, *Inter-laboratory study on standardized MPS libraries: evaluation of performance, concordance, and sensitivity using mixtures and degraded DNA.* Int J Legal Med, 2020. **134**(1): p. 185-198.

209. Zignol, M., A. M. Cabibbe, A. S. Dean, P. Glaziou, N. Alikhanova, C. Ama, S. Andres, A. Barbova, A. Borbe-Reyes, D. P. Chin*, et al.*, *Genetic sequencing for surveillance of drug resistance in tuberculosis in highly endemic countries: a multi-country population-based surveillance study.* Lancet Infect Dis, 2018. **18**(6): p. 675-683.

210. Barrio, P. A., P. Martin, A. Alonso, P. Muller, M. Bodner, B. Berger, W. Parson, B. Budowle, and Dnaseqex Consortium, *Massively parallel sequence data of 31 autosomal STR loci from 496 Spanish individuals revealed concordance with CE-STR technology and enhanced discrimination power.* Forensic Sci Int Genet, 2019. **42**: p. 49-55.

211. Peng, D., Y. Zhang, H. Ren, H. Li, R. Li, X. Shen, N. Wang, E. Huang, R. Wu, and H. Sun, *Identification of sequence polymorphisms at 58 STRs and 94 iiSNPs in a Tibetan population using massively parallel sequencing.* Sci Rep, 2020. **10**(1): p. 12225.

212. Delest, A., D. Godfrin, Y. Chantrel, A. Ulus, J. Vannier, M. Faivre, C. Hollard, and F. X. Laurent, *Sequenced-based French population data from 169 unrelated individuals with Verogen's ForenSeq DNA signature prep kit.* Forensic Sci Int Genet, 2020. **47**: p. 102304.

213. Phillips, C., L. Devesse, D. Ballard, L. van Weert, M. de la Puente, S. Melis, V. Alvarez Iglesias, A. Freire-Aradas, N. Oldroyd, C. Holt*, et al.*, *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, 2018.

214. Bredemeyer, Steffi, Lutz Roewer, and Sascha Willuweit, *Next generation sequencing of Y-STRs in father-son pairs and comparison with traditional capillary electrophoresis.* Forensic Sciences Research, 2021: p. 1-6.

215. Hill, C. R., D. L. Duewer, M. C. Kline, M. D. Coble, and J. M. Butler, *U.S. population data for 29 autosomal STR loci.* Forensic Sci Int Genet, 2013. **7**(3): p. e82-3.

216. NIST, *STRSeq BioProject (Award # 2016-DNR-6150)*. 2021.

217. Devesse, L., L. Davenport, L. Borsuk, K. Gettings, G. Mason-Buck, P. M. Vallone, D. Syndercombe Court, and D. Ballard, *Classification of STR allelic variation using massively parallel sequencing and assessment of flanking region power.* Forensic Sci Int Genet, 2020. **48**: p. 102356.

218. Excoffier, L., G. Laval, and S. Schneider, *Arlequin (version 3.0): an integrated software package for population genetics data analysis.* Evol Bioinform Online, 2007. **1**: p. 47-50.

219. Casals, F., R. Anglada, N. Bonet, R. Rasal, K. J. van der Gaag, J. Hoogenboom, N. Sole-Morata, D. Comas, and F. Calafell, *Length and repeat-sequence variation in 58 STRs and 94 SNPs in two Spanish populations.* Forensic Sci Int Genet, 2017. **30**: p. 66-70.

220. Butler, John M., *Forensic DNA typing : biology, technology, and genetics of STR markers*. 2nd ed. 2005, London ; Burlington, MA: Elsevier Academic Press. xvii, 660 p.

221. Butler, John M. and John M. Butler, *Fundamentals of forensic DNA typing*. 2010, Academic Press/Elsevier,: Amsterdam ; Boston. p. 1 online resource (xviii, 500 p.).

222. Novroski, N. M. M., J. L. King, J. D. Churchill, L. H. Seah, and B. Budowle, *Characterization of genetic sequence variation of 58 STR loci in four major population groups.* Forensic Sci Int Genet, 2016. **25**: p. 214-226.

223. Phillips, C., *Online resources for SNP analysis: a review and route map.* Mol Biotechnol, 2007. **35**(1): p. 65-97.

224. Phillips, C., M. Gelabert-Besada, L. Fernandez-Formoso, M. Garcia-Magarinos, C. Santos, M. Fondevila, D. Ballard, D. Syndercombe Court, A. Carracedo, and M. V. Lareu, *"New turns from old STaRs": enhancing the capabilities of forensic short tandem repeat analysis.* Electrophoresis, 2014. **35**(21-22): p. 3173-87.

225. Earl, Dent A. and vonHoldt, Bridgett M., *STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method.* Conservation Genetics Resources, 2012. **4**: p. 359-361.

226. Xu, S., W. Huang, J. Qian, and L. Jin, *Analysis of genomic admixture in Uyghur and its implication in mapping strategy.* Am J Hum Genet, 2008. **82**(4): p. 883-94.

227. Sampson, J. N., K. K. Kidd, J. R. Kidd, and H. Zhao, *Selecting SNPs to identify ancestry.* Ann Hum Genet, 2011. **75**(4): p. 539-53.

228. Yaeger, R., A. Avila-Bront, K. Abdul, P. C. Nolan, V. R. Grann, M. G. Birchette, S. Choudhry, E. G. Burchard, K. B. Beckman, P. Gorroochurn*, et al.*, *Comparing genetic ancestry and self-described race in african americans born in the United States and in Africa.* Cancer Epidemiol Biomarkers Prev, 2008. **17**(6): p. 1329-38.

229. Sulem, P., D. F. Gudbjartsson, S. N. Stacey, A. Helgason, T. Rafnar, K. P. Magnusson, A. Manolescu, A. Karason, A. Palsson, G. Thorleifsson*, et al.*, *Genetic determinants of hair, eye and skin pigmentation in Europeans.* Nat Genet, 2007. **39**(12): p. 1443-52.

230. Wilde, S., A. Timpson, K. Kirsanow, E. Kaiser, M. Kayser, M. Unterlander, N. Hollfelder, I. D. Potekhina, W. Schier, M. G. Thomas*, et al.*, *Direct evidence for positive selection of skin, hair, and eye pigmentation in Europeans during the last 5,000 y.* Proc Natl Acad Sci U S A, 2014. **111**(13): p. 4832-7.

231. Genomes Project, Consortium, G. R. Abecasis, D. Altshuler, A. Auton, L. D. Brooks, R. M. Durbin, R. A. Gibbs, M. E. Hurles, and G. A. McVean, *A map of human genome variation from population-scale sequencing.* Nature, 2010. **467**(7319): p. 1061-73.

232. Ramani, A., Y. Wong, S. Z. Tan, B. H. Shue, and C. Syn, *Ancestry prediction in Singapore population samples using the Illumina ForenSeq kit.* Forensic Sci Int Genet, 2017. **31**: p. 171-179.

233. Wendt, F. R., 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*, et al.*, *Genetic analysis of the Yavapai Native Americans from West-Central Arizona using the Illumina MiSeq FGx forensic genomics system.* Forensic Sci Int Genet, 2016. **24**: p. 18-23.

234. Rajeevan, H., U. Soundararajan, A. J. Pakstis, and K. K. Kidd, *Introducing the Forensic Research/Reference on Genetics knowledge base, FROG-kb.* Investig Genet, 2012. **3**(1): p. 18.