UNIVERSITY OF LEICESTER

Genetic Diversity and Population Structure of Saudi Arabia

By

Yahya M. Y. Khubrani MSc

Department of Genetics and Genome Biology

College of Life Sciences

University of Leicester

September 2019

Thesis submitted for the degree of Doctor of Philosophy at the University of Leicester

Genetic Diversity and Population Structure of Saudi Arabia

Yahya Khubrani

Abstract

Saudi Arabia is the largest country in the Arabian Peninsula, yet its genetic diversity and population structure has been little studied. This project has used analysis of forensic DNA markers, typed both conventionally (by capillary electrophoresis, CE) and by massively parallel sequencing (MPS), to characterise its genetic diversity. Y-chromosomal haplotypes based on 27 STRs (short tandem repeats) were generated in 597 unrelated Saudi males, classified into five geographical sub-regions. This showed marked population structure with low diversity in the Central and Northern regions, and high diversity in the East and West. Haplogroup J1 was very predominant and showed signals of recent expansion. Comparing geographically-matched males recruited in the UK with those recruited in Saudi Arabia showed significant haplotype differences, pointing to social structure. Variation in 21 autosomal STRs was also investigated. As in the Y chromosome study, this revealed population structure, but with a different geographical pattern. Heterozygote deficiency was observed at nearly all loci, probably as a consequence of high levels of consanguineous marriage. MPS analysis (via the Verogen ForenSeq™ DNA Signature Prep Kit) of 27 autosomal STRs and 91 identity-informative single nucleotide polymorphisms (iiSNPs) revealed sequence variation that reduced both the STR-based and iiSNP-based random match probabilities. As with the autosomal CE data, evidence of consanguinity was apparent from both marker types. A global comparison showed that the Saudi sample was typical of Middle Eastern populations, with higher inbreeding than is seen in most European, African and Central/South Asian populations. Y-STR and X-STR sequence data from the same experiments revealed sequence variation in both types of markers. A population-level comparison of the Saudi Arabian X-STRs with a global sample demonstrated affinity with other Middle Eastern populations. This study has revealed population structure and the influence of consanguinity in Saudi Arabia, and provides valuable reference datasets for forensic analysis.

Acknowledgements

I would first like to thank the Saudi Arabian Ministry of Interior who gave me the opportunity to do my PhD. I am grateful for having this chance. I would like also to thank all my work colleagues in the General Administration of Forensic Evidence, especially my colleagues at the Forensic Genetics lab, Riyadh and Najran. Thanks also to the Saudi Cultural Bureau for their support during my PhD. I also had continuous training and education at different institutions: Forensic Science Services (UK), University of Central Lancashire (UK), and the General Administration of Forensic Evidence (KSA). Without having support from such great people, it would have been impossible to make it this far; thanks are due to all of them.

Words cannot describe my good experience at Leicester with Prof Mark Jobling and Dr Jon Wetton. Their support, motivation, and friendly relationship has been a truly unforgettable experience. I would like also to thank all members of Mark Jobling's group and the Alec Jeffreys Forensic Genomics Unit for their encouragement and friendship: Dr Celia May, Dr Nitikorn Poriswanish, Jodie Lampert, Marwan El Khoury, Tunde Huszar, Jordan Beasley, Margherita Colucci, and Ettore Fedele. Thanks also to my panel review committee (Prof Yuri Dubrova and Dr Richard Badge) for their useful comments on my work and to the Bioinformatics and Biostatistics Analysis Support Hub (BBASH) especially Dr Chiara Batini and Dr Matthew Blades (for opening the door to the bioinformatic world). Also, thanks to all members of Leicester Arabic Genetics Club; we had really nice discussions.

I would like to thank all collaborators. Wellcome Sanger Institute (Chris Tyler-Smith, Yali Xue and Pille Hallast) for sharing data with us and having me at Sanger during the summer of 2018; indeed it was another nice memory during my PhD. Also thanks are due to many people from our commercial partners: ThermoFisher Scientific and Verogen and not forgetting Chris Phillips for sharing HGDP data with us, and all of our visitors such as Giordano Bruno, Dr Mohammad Alenizi, Hussain Alsafiah, and Mahdi Haidar. Thanks also for all participants who gave DNA for this project.

I acknowledge all family members and friends (it is really a very big list; I'm sure they know who they are). In short special thanks to my parents, my wife, my son Mohammed, my two lovely daughters, Orjwan and Bisan, and all my brothers and sisters for their support and love.

List of abbreviations

A	Adenine
AIM	Ancestry Informative Marker
AMOVA	Analysis of Molecular Variance
BAM	Binary Alignment Map
bp	Base pairs
C	Cytosine
CCD	Charge-coupled device
CE	Capillary electrophoresis
CEPH	Centre d'Étude du Polymorphisme Humain
CI	Confidence Interval
CODIS	Combined DNA Index System
dbSNP	SNP database
ddNTP	Dideoxy nucleotide triphosphate
dH_20	Deionised water
DNA	Deoxyribonucleic acid
dNTP	Deoxyribonucleotide triphosphate
DoC	Depth of Coverage
EDTA	Ethylenediaminetetraacetic Acid
emPCR	Emulsion PCR
EMPOP	EDNAP mitochondrial DNA population database
EPG	Electropherogram
EtBr	Ethidium Bromide
ExoI	Exonuclease I
F_{ST}	Fixation Index
FTDNA	FamilyTreeDNA
G	Guanine
H	Heterozygosity
HCl	Hydrochloric Acid

HD	Haplotype Diversity
He	Expected Heterozygosity
Hg	Haplogroup
HGDP	Human Genome Diversity Project
HGDP-CEPH	Human Genome Diversity Project - Centre d'Etude du Polymorphism Humain
HMP	Haplotype Match Probability
Ho	Observed Heterozygosity
HWE	Hardy-Weinberg Equilibrium
IBD	Isolation by Distance, or Identity-by-Descent
IBS	Identity-by-State
iiSNP	Identity-informative single nucleotide polymorphism
Indel	Insertion/Deletion Polymorphism
IPC	Internal PCR control
ISFG	International Society for Forensic Genetics
KSA	Kingdom of Saudi Arabia
KYA	Thousand Years Ago
LD	Linkage Disequilibrium
MCMC	Markov Chain Monte Carlo
MDS	Multi-Dimensional Scaling
MJ	Median-joining
MPS	Massively parallel sequencing
MSY	Male-specific region of the Y chromosome
mtDNA	Mitochondrial DNA
NCBI	National Center for Biotechnology Information
ng	Nanogram
NGS	Next-Generation Sequencing
NRY	Non-recombining portion of the Y chromosome
PAR	Pseudoautosomal Region
PCA	Principal Component Analysis

PCR	Polymerase Chain Reaction
PNACL	Protein and Nucleic Acid Chemistry Laboratory of the University of Leicester
RFLP	Restriction Fragment Length Polymorphism
RFU	Relative Fluorescence Unit
RM	Rapidly mutating
RMP	Random Match Probability
SAP	Shrimp Alkaline Phosphatase
SBS	Sequencing by synthesis
SD	Standard Deviation
SDS	Sodium Dodecyl Sulphate
SNP	Single Nucleotide Polymorphism
SRY	Sex-determining Region, Y
STR	Short Tandem Repeat
SV	Structural Variation
SWGDAM	Scientific Working Group on DNA Analysis Methods
T	Thymine
TAE	Tris-acetate-EDTA buffer
TBE	Tris-borate-EDTA buffer
TE	Tris-EDTA
TMRCA	Time to most recent common ancestor
Tris	Tris(hydroxymethyl)aminomethane
UV	Ultraviolet
YHRD	Y Chromosome Haplotype Reference Database
µl	microlitre

Table of Contents

Table of Figures

Table of Tables

Chapter 1 Introduction

1.1 Preface

This thesis involves the use of both conventional and massively parallel sequencing (MPS) forensic tools to study the genetic diversity of the Saudi Arabian population. The tools used include highly variable short tandem repeats (STRs) on the autosomes, Y chromosome and X chromosomes, as well as targeted autosomal single nucleotide polymorphisms (SNPs), and also some analysis of the sequence variation in mitochondrial DNA (mtDNA). In this introductory chapter, the general properties of sequence variants including STRs and SNPs are reviewed, followed by a consideration of specific forensic marker sets. The use of MPS as a method for high-resolution analysis of such loci is described, and the introduction closes with an account of the history, geography and social structure of Saudi Arabia, followed by a statement of aims and objectives for the study.

1.2 The human genome

Normal humans carry 46 chromosomes (as 23 pairs) which constitute the nuclear genome. These chromosomes are categorised into 22 autosomal chromosomes and two sex chromosomes, namely X and Y. In addition, a small portion of the genome exists as circular double-stranded DNA within mitochondria (Figure 1.1). The completion of the 3.2 billion base pairs (bp) of the human genome sequence is one of the greatest scientific achievements of the recent era (Lander et al. 2001; Venter et al. 2001). Subsequent work such as the 1000 Genomes Project has catalogued human sequence variation and found that nucleotide diversity is low across all populations, and on average copies of the genome differ in alignable sequence by <0.1% (1000 Genomes Project Consortium et al. 2010; Consortium 2007). However, given the size of the genome this small percentage encompasses a very large number of variants that can be used in genetic studies, making a profound impact in many disciplines including medicine, the understanding of human history, and in serving justice.

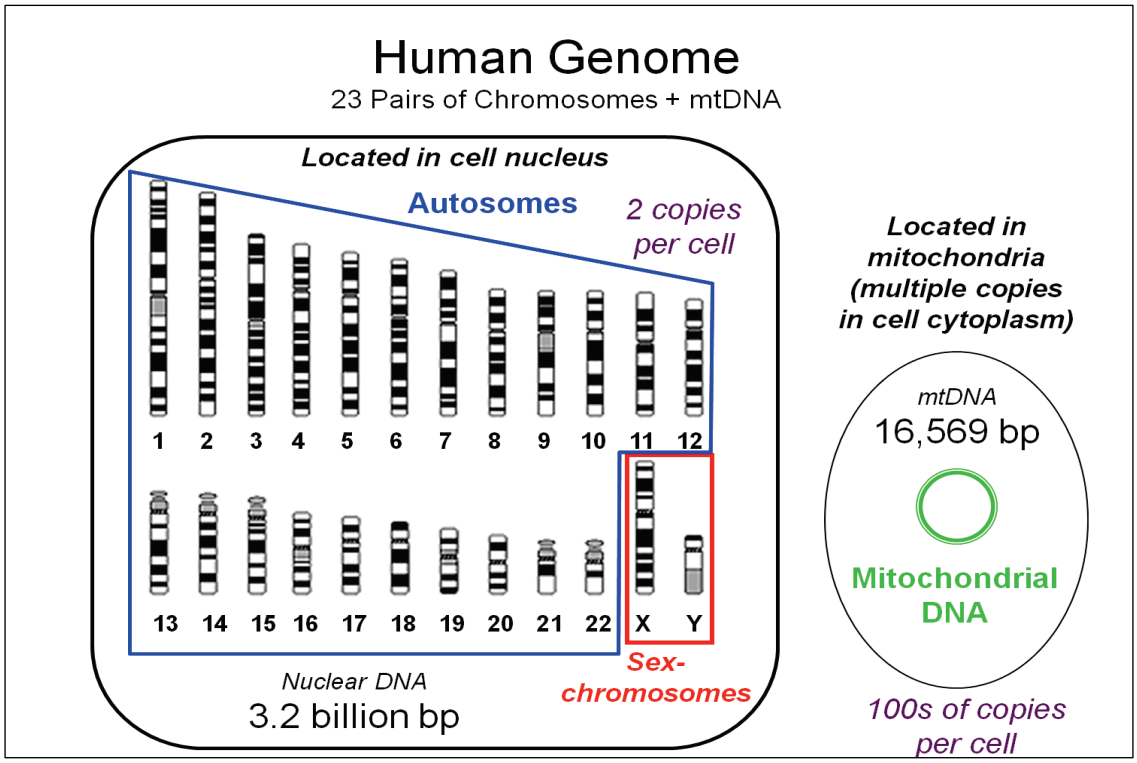


Figure 1.1: Illustration of the human genome. Chromosomes 1-22, sex chromosomes and mtDNA (Butler 2005).

1.3 Patterns of inheritance of the human genome

Genetic information within autosomes is inherited equally from both parents. However, an exception applies to the sex chromosomes and mtDNA, where the sex-determining Y chromosome is inherited through the paternal line, and mtDNA is inherited through the maternal line. The X chromosome has a different mode of inheritance where a son will inherit his X chromosome from his mother, while a daughter inherits one copy from each parent (Figure 1.2). These varied modes of inheritance for distinct segments of the genome make them useful for different applications in genetic studies, and lead to differences in the means of detecting and interpreting their patterns of DNA polymorphism.

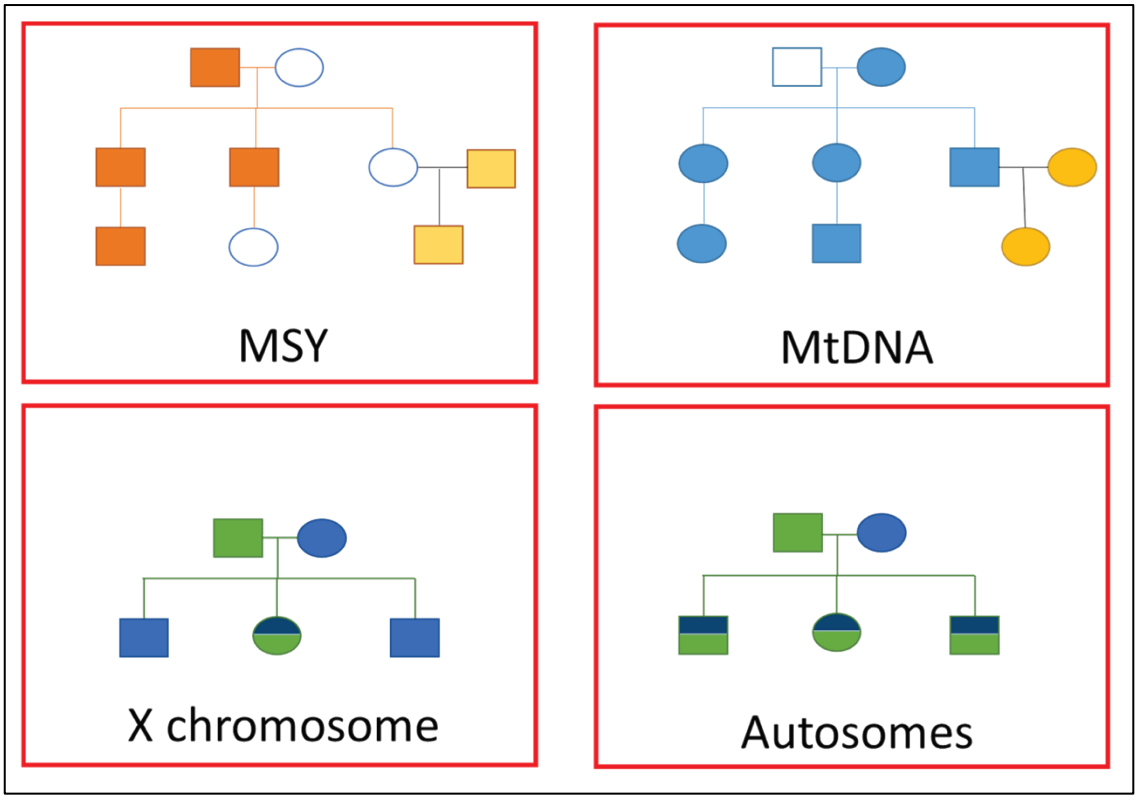


Figure 1.2: Inheritance of different parts of the human genome; The autosomes, male-specific region of the Y (MSY), X chromosome and mitochondrial DNA. Squares indicate males, circles indicate females and fill colour represents inheritance of genetic components.

1.4 Types of sequence variants in the human genome

Mutation and recombination are dynamic processes resulting in variation in the genome at both small and large scales. The commonly studied DNA polymorphisms are categorised as SNPs, STRs, insertion-deletions (indels) and structural variants, which include copy number variants (CNVs). Variants are usually considered by comparison with a genome reference. The human genome reference (currently GRCh38), is of good quality and completeness compared to the sequences of most other higher organisms and it continues to be improved and updated. Studies of genetic diversity with human populations, today via whole-genome sequencing approaches (1000 Genomes Project Consortium et al. 2015), have discovered very large numbers of variants, which can be explored in the context of the reference sequence via open access platforms such as the UCSC Genome Browser (Kent et al. 2002). The focus in this thesis is on two types of genetic marker, namely STRs and SNPs.

1.4.1 Short tandem repeats as genetic markers

STRs, also known as microsatellites, are made up of short sequences usually 2-8 nucleotides in length which are tandemly repeated and vary in their number at a specific locus. The number of repeats in a typical polymorphic STR ranges from ~10-30. STRs are abundant and distributed throughout the human genome (Subramanian et al. 2003): it is estimated that there are about 700,000 STRs in total, representing about 1% of the total genome by length (Willems et al. 2016). STRs can be found on the autosomes as well as the sex chromosomes (but not in human mtDNA) and are used widely in different forensic applications (Sibille et al. 2002).

A high proportion of STRs are found adjacent to retrotransposon sequences such as *Alu* or L1, and are likely to arise from mutation in the poly(A) tracts associated with these sequences (Nadir et al. 1996). The mechanism of STR growth and mutation is believed to be a result of slippage during DNA replication, resulting in either a gain or loss of repeats (Schlötterer and Tautz 1992). Mutation rates for STRs commonly used in forensic analysis and paternity testing (mostly tetranucleotide repeats) range from 10E-3 to 10E-4 per generation, mostly involving single step mutations (consistent with the Stepwise Mutation Model (Ohta and Kimura 1973)), with only 4.5% of mutations showing two-step changes (Brinkmann et al. 1998). Many factors play a role in determining STR mutation rates, including length of repeat, length of repeat array, sequence structure including the flanking region, recombination, sex and age (reviewed by (Fan and Chu 2007)).

According to their repeat region structures, STRs can be classified into three main types, namely: simple (all repeat units have an identical sequence, e.g. $(AATG)_n$), compound (two or more repeat types, e.g. $TCTA(TCTG)_m(TCTA)_n$), complex (multiple repeat types with intervening non-repeat sequences, e.g. $(TCTA)_m(TCTG)_n(TCTA)_3TA$ $(TCTA)_3TCA(TCTA)_2TCCATA(TCTA)_pTC$) (Butler 2009; Goodwin 2007) and hypervariable complex repeats which include variable and constant repeat-array elements such as SE33 (Buroker et al. 1987).

1.4.2 Single nucleotide polymorphisms as genetic markers

A SNP is the most simple and smallest-scale variant between two DNA sequences. It can be classified into three categories: base substitution, a single nucleotide insertion and a single nucleotide deletion (indel). In a base substitution a single nucleotide is replaced by another. When this substitution happens between the same nitrogenous base types (e.g. a pyrimidine [C or T] for another pyrimidine, or a purine [A or G] for another purine) this is called a transition (ts), while a pyrimidine/purine exchange is called a transversion (tv) as summarised in Table 1.1 below. Despite the fact that any nucleotide has only one possible transition, but two possible transversions, transitions are three times more frequent than transversions (reviewed by (Hallast et al. 2013)). SNPs are generally biallelic, though examples are known where three (Westen et al. 2009), or even four (Phillips et al. 2015), alleles are present in the population at an appreciable frequency.

Some individual nucleotides in the nuclear genome are more mutable than others. In particular, the rates of transition and transversion at the 5′-CpG-3′ dinucleotide (where "p" signifies a phosphate group) are about ten times higher than the rates at other dinucleotides (Nachman and Crowell 2000). This is because the C on both strands in these dinucleotides tends to be methylated (as 5-methylcytosine), and following spontaneous deamination, this modified base becomes thymine, which can escape DNA repair more effectively than uracil (the product of deamination at an unmethylated C) (Nachman and Crowell 2000). Whole-genome sequencing shows that each copy of the human genome contains about 4-5 million SNPs, and the number of independently discovered SNPs now exceeds 85 million (1000 Genomes Project Consortium et al. 2015).

Table 1.1: Types of SNP variation in a DNA sequence. Base at position of alteration is highlighted in red and hyphen represents a deletion.

Types of SNP variation	sequence examples
Ancestral sequence	GATACCTTTACGGTGAACA
Base substitution (ts)	GATACCTTTGCGGTGAACA
Base substitution (tv)	GATACCTTTCCGGTGAACA
Deletion	GATACCTTT-CGGTGAACA
Insertion	GATACCTTTAACGGTGAACA

1.5 Autosomal markers in forensic genetics

During the last three decades, DNA profiling has revolutionised the field of forensic investigation (Butler 2009). Differentiation between two individuals' DNAs can be conducted using autosomal markers which are highly variable between individuals with the exception of identical twins. The first DNA-based method utilised autosomal minisatellite variation, developed by Alec Jeffreys at the University of Leicester (Jeffreys et al. 1985). However, such analysis was time consuming, and required a large amount of good-quality DNA. Analysis of minisatellites was therefore replaced by PCR (polymerase chain reaction) amplification of STRs (Butler and Hill 2012). Forensic applications of autosomal STRs include individual identification, paternity testing, linking a crime scene sample to a suspect, linking different crime scenes, and disaster victim identification.

1.5.1 Autosomal STRs in forensic genetics

STRs are the preferred markers in forensic applications due to the multiple different-length alleles at each locus which yield a high power of discrimination. Since STR alleles can be simply described by the number of repeats, genotypes or haplotypes for multiple STRs are represented by lists of numbers, which are ideal for databasing and rapid searching. Finally, since most STRs are contained within a space of <150 bp, they can be amplified by PCR from fragmented DNA. A low rate of 'stutter' (due to strand slippage during PCR) is an important selection criterion for

forensic STRs, and therefore tetranucleotide loci are preferred to dinucleotides. An example of a tetranucleotide STR is illustrated in Figure 1.3, showing five different alleles at the TH01 locus.

Although an individual autosomal STR can have many different-length alleles, multiple STRs are required for individual identification. Provided STRs are independently inherited, allele frequencies can be multiplied (the 'product rule') in calculating the random match probability (RMP), the chance that two randomly selected individuals from a population will share identical profiles (Lewontin and Hartl 1991). In designing STR multiplexes, loci have been selected to lie on different chromosomes, or physically widely separated on the same chromosome, to avoid linkage between loci.

Figure 1.3: Examples of alleles at an autosomal STR. Five different alleles at the STR TH01 are shown, containing tetranucleotide repeats AATG. Currently 21 different alleles are recorded in STRBase (https://strbase.nist.gov/), spanning the range of 3-14 repeats and including 10 microvariant "intermediate alleles", which contain altered length repeat units (such as the ATG unit shown in allele 9.3, above). Fragment sizes range between 160 and 204 bp when using the Identifiler® kit primers.

The development of autosomal STR systems has gone through multiple stages, starting with four STRs by the UK Forensic Science Service (FSS) (Kimpton et al. 1994), to the current generation of commercial multiplexes that encompass up to 24 loci. The widely used commercial kits for autosomal loci include: GlobalFiler® (Applied Biosystems®) (21/24 are autosomal), PowerPlex® Fusion (Promega Corporation) (23/27 are autosomal) and Investigator 24plex QS Kit (QIAGEN)

(23/24 are autosomal). The relative sizes of STRs in these kits are illustrated in Figure 1.4 .

Initial analysis of STRs relied on radiolabelling of primers, evolving to silver staining, the use of fluorescent primers and slab polyacrylamide gels, and finally to capillary electrophoresis (CE). The principle of separation of loci is based on fragment size, where shorter fragments travel faster through the matrix in the capillary; multiple fluorescent dye colours are also used to distinguish similarly sized fragments of DNA. An example of an electropherogram (EPG) is shown in Figure 1.5. Given the numbers of STRs in these kits, RMPs are very low; for example RMP values for GlobalFiler® range from 2.21E-26 to 5.21E-25 in five South African populations (Ristow et al. 2016). The drive to increase the number of STRs has come from a need to allow comparisons between profiles generated by forensic laboratories worldwide, with different traditions of STR use, and also to maximise utility in matching partial profiles (Hennessy et al. 2014). Capillary Electrophoresis (CE) is now conducted semi-automatically using genetic analyzer machines.

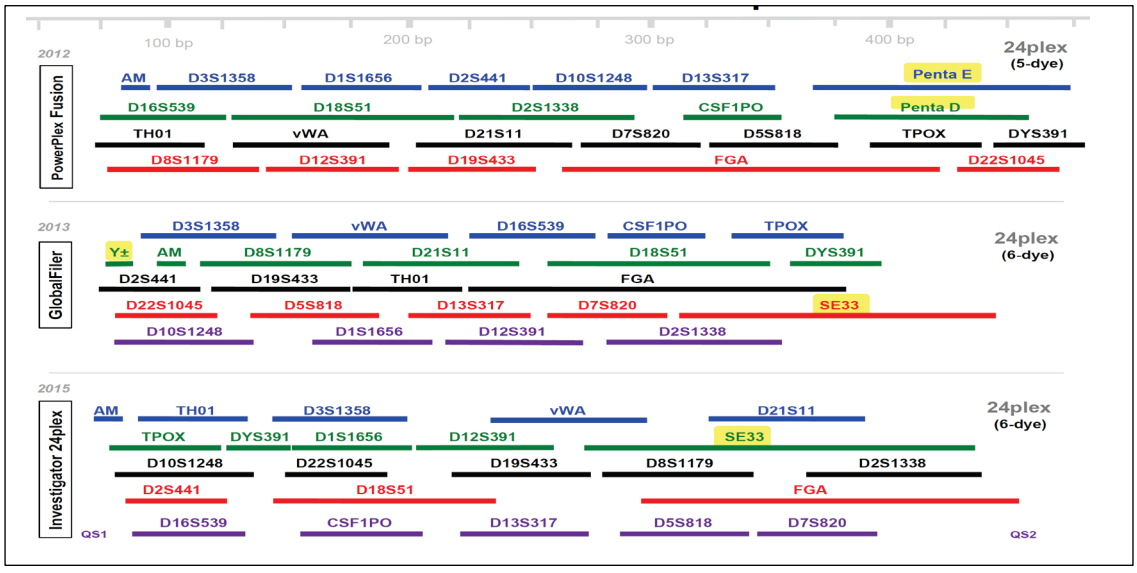


Figure 1.4: Relative sizes of STRs and fluorescent dye colours used in three commercial autosomal kits: PowerPlex® Fusion, GlobalFiler® and Investigator 24plex QS Kit (Butler 2012). All markers give fragment sizes between ~75 bp and 475 bp.

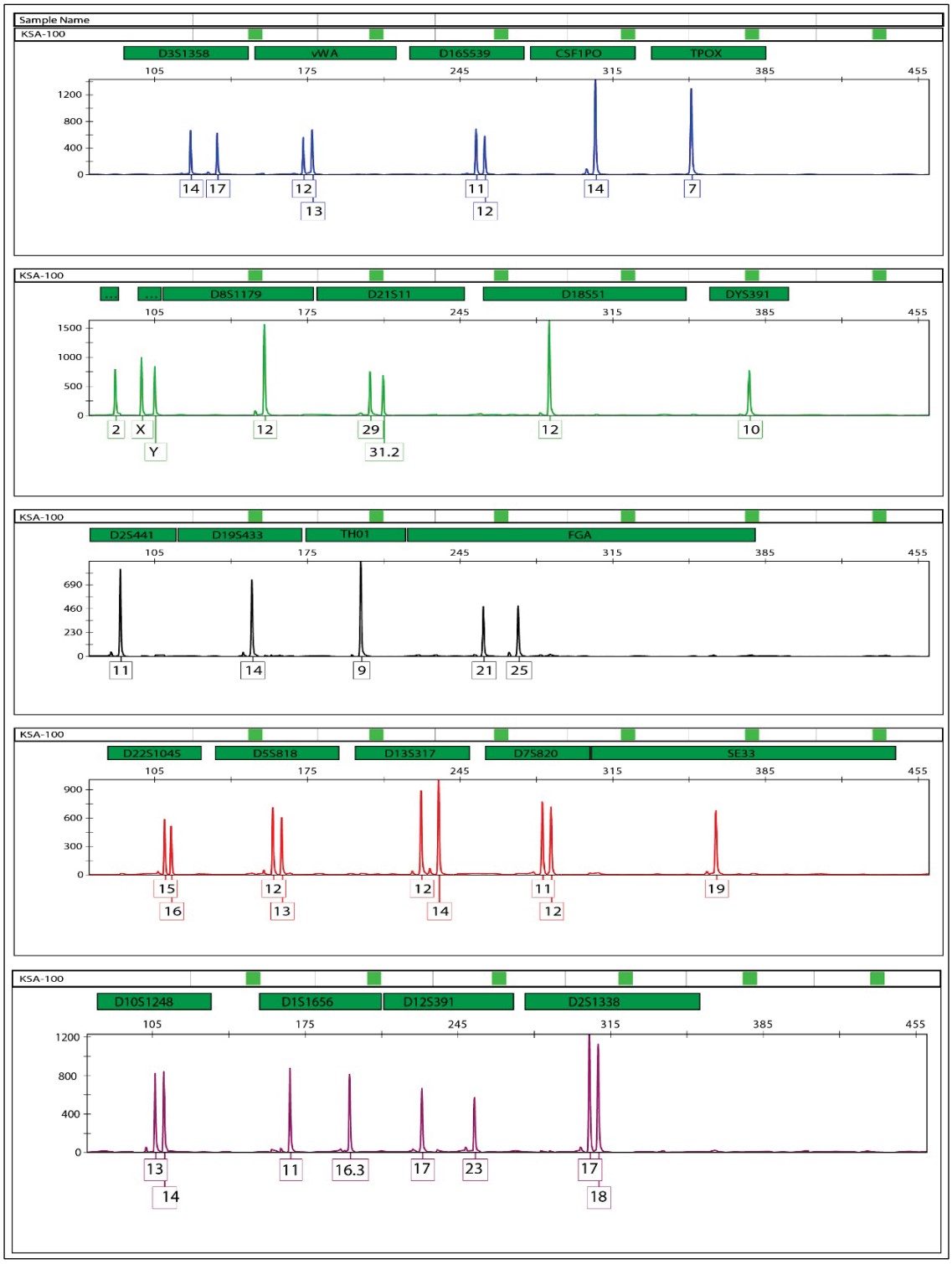


Figure 1.5: An example of an electropherogram (EPG) for a sample amplified with the GlobalFiler® kit and analysed with GeneMapper® ID-X v1.4 software. The y-axis shows relative fluorescent units (RFU), and the x-axis shows length in bp.

1.5.2 Autosomal SNPs in forensic genetics

Forensically useful targeted autosomal SNPs can be categorised based on their application, including identity informative SNPs (iiSNPs), ancestry informative SNPs and phenotype informative SNPs (Butler 2009). Focusing on iiSNPs, as mentioned above, autosomal SNPs are usually biallelic, while STRs are multiallelic, and therefore a larger number of SNPs is required to achieve equivalent discrimination power. It has been estimated that around 50-100 SNPs are required to provide a similar power of discrimination compared to typical sets of STRs (Chakraborty et al. 1999). Another issue is that allele frequencies of SNPs are more likely to be population specific (and so they can suffer from ascertainment bias), and therefore this makes the construction of population-specific databases particularly important. A final disadvantage is that mixtures are more difficult to resolve with SNPs than with STRs. SNPs, however, do have some advantages. The fact that they can be amplified on very small PCR amplicons (60-80 bp) is useful in degraded samples (Budowle and Van Daal 2008); SNPs are also stutter-free. Finally, in a short fragment of DNA <300 bp it is possible to capture multiple SNPs that are inherited together as one block known as a "microhaplotype", which can be highly informative (Oldoni et al. 2019).

Many SNP genotyping methods have been developed, including hybridisation methods, TaqMan® assay, SNaPshot™ and microarrays; however, these methods either have limited capacity for multiplexing, or the amount of DNA needed for typing is not suitable for forensic applications (Sobrino et al. 2005). It is also worth noting that CE and the other methods mentioned above do not provide any information about microhaplotypes (because the data are unphased). A good alternative may be MPS-based methods which can overcome the multiplexing issue and also provide haplotype sequences directly.

1.6 Properties and diversity of the Y chromosome

The Y chromosome at 60 Mb is one of the smallest human chromosomes, and makes up only about 2% of the genome (Jobling et al. 2014). Its most obvious function is in sex determination, reflected in the fact that normal males (46, XY) have a copy of

the Y chromosome and normal females do not (46, XX). The Y chromosome bears relatively few genes (about 80) for its size (Skaletsky et al. 2003). DNA polymorphisms on the Y chromosome have attracted a significant amount of interest for researchers focused on many different applications (Jobling and Tyler-Smith 2003). In the medical field for instance, the Y chromosome has been used to investigate susceptibility to diseases such as coronary artery disease (Maan et al. 2017), and to understand structural variation associated with male infertility (Krausz and Casamonti 2017). Furthermore, the Y chromosome has been used in evolutionary studies (Hughes et al. 2012), the analysis of human migration and admixture (Wilkins 2006), the linking of genealogy and surnames (Calafell and Larmuseau 2017) and forensic genetics (Kayser 2017).

Most of the Y chromosome is haploid and passed down from father to son without recombination (crossover) with the exception of two segments known as the Pseudoautosomal Regions (PAR1 and 2) which do cross over with the X chromosome (Rappold 1993). These regions are located at the ends of the chromosome and together represent only 3 Mb of the total chromosomal length. As illustrated in Figure 1.6, around half of the structure of Y chromosome is constituted by heterochromatin, with around 23 Mb of euchromatin (Jobling and Tyler-Smith 2017) The male-specific region of the Y (MSY), also referred to as the non-recombining portion of the Y (NRPY or NRY), is the largest portion of about 95% of the chromosome, and passes through generations without the reshuffling of recombination. However, accumulated mutations (polymorphisms) differentiate Y-chromosome lineages through time, and differentiating between such lineages has been useful in a variety of genetic studies (Jobling and Tyler-Smith 2003), including forensic analysis, which will be described below.

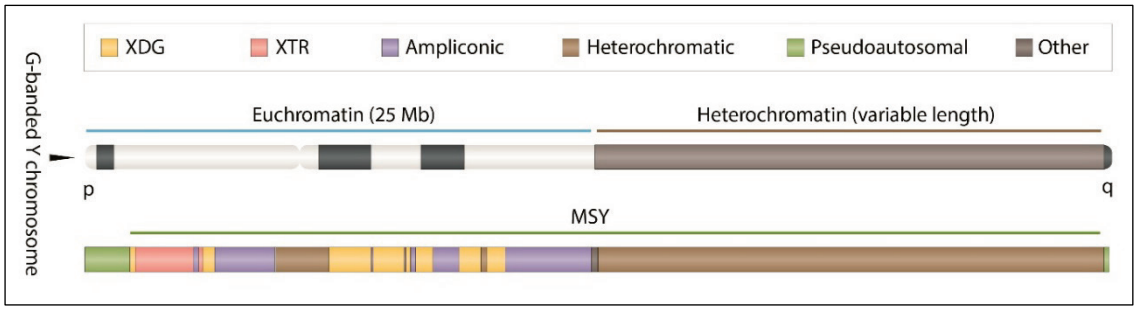


Figure 1.6: Y-Chromosome physical structure. The top bar shows the G-banding pattern of the Y chromosome while, different sequence classes shown below. The key above defines these sequence classes. XDG: X-degenerate region; XTR: X-transposed region. Modified from (Jobling and Tyler-Smith 2017)

1.6.1 Y chromosome STRs

The first Y-STRs were identified and used in the early 1990s (Roewer et al. 1992), and since then systematic searches (Kayser et al. 2004) have resulted in several hundred useful markers on the MSY. These define highly variable haplotypes, but it is important to note that profiles based on Y-STRs can never be as informative as autosomal STR profiles, because the Y-STRs are not independently inherited and therefore the product rule cannot be used to estimate RMPs (Jobling et al. 1997). In practice, methods such as haplotype counting must be used to assess the significance of a Y-profile match. In addition, haplotypes are shared among patrilineally related males (e.g. fathers and sons, brothers, paternal uncles etc.), and this has to be taken into account.

Knowledge of Y-STR mutation rates allows estimation of time-to-most-recent-common-ancestor (TMRCA) of sets of Y-STR haplotypes. In forensic terms, the mutation rate is also important in understanding the significance of a match or near-match. Several studies have compared Y-STR haplotypes in fathers and sons, allowing direct observation of mutation events. The average mutation rate of 15 Y-STRs was estimated as 2.80E-3 per locus per generation when 4999 DNA samples of confirmed father-son pairs were used (Kayser et al. 2000; Kayser and Sajantila 2001). However, this study covered only 15 commonly used Y-STR markers which will not result in an accurate overview of mutation rate for all STRs on the Y chromosome. A larger-scale investigation of 186 Y-STRs in 2010 among confirmed father-son pairs has also been conducted (Ballantyne et al. 2010). This study

estimated the mutation rate for 90% of the 186 Y-STRs as 1E-4 to 1E-3 per locus per generation (this set includes all the markers included in commercially available Y-chromosome kits at the time of conducting the study).

An even larger survey of Y-STR mutation rates (Willems et al. 2016) used high-coverage whole-genome sequencing data from around 1300 individuals, and employed an algorithm (MUTEA) to infer rates from population-scale data via a high-resolution SNP-phylogeny. The study investigated 702 Y-STRs which were selected from 4500 Y-STRs bioinformatically detectable in the available sequence data of the 1000 Genomes Project and Simons Genome Project samples (Mallick et al. 2016). This study also suggested a new collection of Y-STRs with higher mutation rates as shown in Table 1.2; however, since it was based on 75-100-bp Illumina sequencing reads, the study was not able to detect particularly long Y-STR alleles, and notably failed to accurately call variation in most of the rapidly mutating Y-STRs (see following section), many of which have complex internal structures (Willems et al. 2016).

1.6.2 Rapidly Mutating (RM) Y-STRs

A remarkable finding of the mutation study described above (Ballantyne et al. 2010) was the much higher mutation rate observed for 13 markers out of the total set of 186. Rates for these markers were estimated to be between 1.19E-2 and 7.73E-2 per locus per generation. Consequently, these markers were named “Rapidly Mutating” (RM) Y-STRs (Ballantyne et al. 2010).

RM Y-STRs have great forensic potential. These markers (namely: DYF387S1, DYF399S1, DYF403S1, DYF404S1, DYS449, DYS518, DYS526, DYS547, DYS570, DYS576, DYS612, DYS626, and DYS627) result in high discrimination power which can be applied in forensic casework to distinguish between male-line relatives. To address this potential, 604 unrelated males from 51 populations in the Human Genome Diversity Panel (HGDP) (Cann et al. 2002) were typed to compare RM Y-STRs to the set of 17 Y-STRs in the commercial Yfiler® (Life Technologies) kit (Ballantyne et al. 2012). This showed that RM markers differentiated all haplotypes with the exception of eight individuals who between them shared three similar

haplotypes. By contrast, 85 individuals shared 33 haplotypes based on the Yfiler® kit (Ballantyne et al. 2012). A larger-scale study has been conducted with a larger database of 14,644 related and unrelated individuals from 111 populations globally (Ballantyne et al. 2014). This study, with others, has concluded that RM Y-STRs can be valuable in forensic applications with much greater potential for paternally related male differentiation than in existing commercial Y-STR kits (Oh et al. 2015; Turrina et al. 2016; Zhang et al. 2016). While many RM Y-STR studies conducted the PCR in three separate multiplexes to type the 13 markers, a single multiplex reaction has also been proposed (Alghafri et al. 2015; Hadi 2016). Recently, apart from these markers, a new collection of 20 Y-STRs with high mutation rates has been suggested (Willems et al. 2016). This collection contains both di- and tetranucleotide markers with sequence length range of 40-72 bp, some of which were announced for the first time.

1.6.3 Y chromosome SNPs and haplogroups

Y-SNPs are not generally used in forensic analysis, but are useful because they define a robust phylogeny of 'haplogroups' within which Y-STR haplotypes can be considered (de Knijff 2000). Haplogroups show a high degree of geographical specificity (Jobling and Tyler-Smith 2003), and their global distribution can contribute to the understanding of human history, including past migrations, colonisations and admixture events. Comparison with maternally inherited mitochondrial DNA can illuminate sex-biased processes (Wilkins 2006).

Y chromosome haplogroups are defined by the presence of particular alleles of binary markers, almost all of which are SNPs. These markers are slowly evolving with low mutation recurrence rates, so SNP-defined haplotypes (haplogroups) can be used to construct a maximum-parsimony tree, rooted by determining ancestral states from the chimpanzee genome sequence (Jobling and Tyler-Smith 2017). Over time, this tree has developed in complexity as more Y chromosomes are analysed, and more SNPs discovered. In 2002, the Y Chromosome Consortium (Y Chromosome Consortium 2002) established an alphanumeric haplogroup classification (based on 245 SNPs) which was updated in 2008 (using ~600

markers) and persists today (Karafet et al. 2008). Current trees are based on next-generation sequencing data from various collections of Y chromosomes, including the 1000 Genomes Project panel. Here, sequence data from 1244 human Y chromosomes belonging to 26 populations identified 65,000 different polymorphisms (Poznik et al. 2016). The phylogenetic tree in Figure 1.7 was constructed using binary nucleotide variations based on 10.3 Mb of accessible Y-DNA sequence (Poznik et al. 2016), and shows the importance of recent Y lineage expansions.

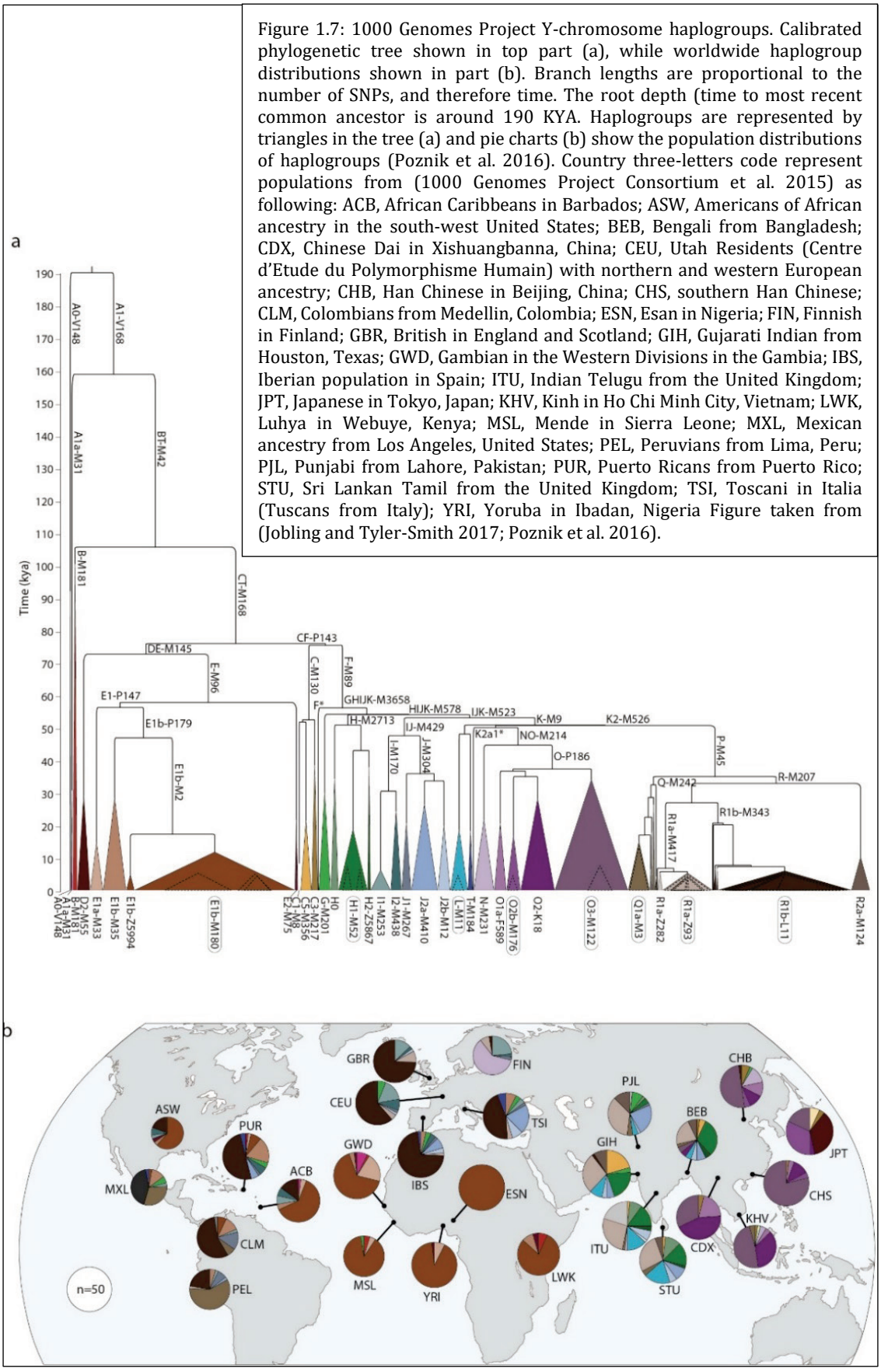


Figure 1.7: 1000 Genomes Project Y-chromosome haplogroups. Calibrated phylogenetic tree shown in top part (a), while worldwide haplogroup distributions shown in part (b). Branch lengths are proportional to the number of SNPs, and therefore time. The root depth (time to most recent common ancestor is around 190 KYA. Haplogroups are represented by triangles in the tree (a) and pie charts (b) show the population distributions of haplogroups (Poznik et al. 2016). Country three-letters code represent populations from (1000 Genomes Project Consortium et al. 2015) as following: ACB, African Caribbeans in Barbados; ASW, Americans of African ancestry in the south-west United States; BEB, Bengali from Bangladesh; CDX, Chinese Dai in Xishuangbanna, China; CEU, Utah Residents (Centre d'Etude du Polymorphisme Humain) with northern and western European ancestry; CHB, Han Chinese in Beijing, China; CHS, southern Han Chinese; CLM, Colombians from Medellin, Colombia; ESN, Esan in Nigeria; FIN, Finnish in Finland; GBR, British in England and Scotland; GIH, Gujarati Indian from Houston, Texas; GWD, Gambian in the Western Divisions in the Gambia; IBS, Iberian population in Spain; ITU, Indian Telugu from the United Kingdom; JPT, Japanese in Tokyo, Japan; KHV, Kinh in Ho Chi Minh City, Vietnam; LWK, Luhya in Webuye, Kenya; MSL, Mende in Sierra Leone; MXL, Mexican ancestry from Los Angeles, United States; PEL, Peruvians from Lima, Peru; PJL, Punjabi from Lahore, Pakistan; PUR, Puerto Ricans from Puerto Rico; STU, Sri Lankan Tamil from the United Kingdom; TSI, Toscani in Italia (Tuscans from Italy); YRI, Yoruba in Ibadan, Nigeria Figure taken from (Jobling and Tyler-Smith 2017; Poznik et al. 2016).

1.6.4 Haplogroup prediction

While Y-SNPs provide the best way for defining a phylogeny, Y-STRs can be used to predict the haplogroup (Jobling 2001). The diversity of allele length at STR markers is usually restricted within the same haplogroup. Many haplogroup predictor tools are available and several have been used in scientific papers. Examples of widely used predictors include Vadim Urasin's YPredictor (predictor.ydna.ru), the NevGen Y-DNA Haplogroup Predictor (www.nevgen.org) and Whit Athey's HAPEST (Athey 2005); www.hprg.com/hapest5/). Many factors could influence the prediction accuracy including haplotype similarity between haplogroups, the number of loci typed, and the number and haplogroup diversity of the datasets underlying the predictor models. Consequently, high resolution SNP typing is required to improve predictions and to provide reliable haplogroup assignment (Emmerova et al. 2017).

1.6.5 Time to most recent common ancestor

A good estimation of mutation rates for Y-chromosome SNPs is important for estimating TMRCA in evolutionary studies, is essential to interpret haplotype relationships between putative relatives in paternity testing, and allows better interpretation of Y chromosome DNA evidence in forensics. MSY mutation rates have been estimated directly using genealogically-based approaches by whole-genome sequencing in families, for up to 15 Mb of the chromosome. These studies have given a rate of about 3E-10 mutations per position per year (PPPY) (Helgason et al. 2015; Xue et al. 2009). An evolutionary approach has also been used, analyzing around 11 Mb of the chromosome in 367 individuals who share the same Y-haplogroup, yielding a rate of 0.78E-9 per bp per year (Balanovsky et al. 2015). The difference between the genealogical and evolutionary rate could be due to lineage-specific effects that have also been observed in population-based sequencing studies (Hallast et al. 2015; Scozzari et al. 2014)

1.6.6 Male-specific expansions

Y-chromosomal haplotypes are particularly susceptible to changes in population

frequency through drift, because of the low effective population size of the chromosome and the variance in offspring number of males. In some societies this effect can be strengthened by rules governing marriage and patrilineal descent, and this could lead to low Y diversity. Patterns of Y-STR haplotype diversity have been used to deduce such male-biased social structuring (reviewed by (Batini and Jobling 2017)); for example, a common haplotype seen in Asia was interpreted as the effect of descent from Genghis Khan (Zerjal et al. 2003), based on its TMRCA and geographical distribution. More generally, Y-STR and sequence-based analyses of some haplogroups demonstrates that they have expanded very recently and it has been systematically analysed in European populations, where three haplogroups have expanded rapidly in the last 3-5 thousand years (Batini et al. 2015). Such recent expansions and social structure effects have forensic relevance, because they give rise to clusters of closely related Y-STR haplotypes that can complicate the interpretation of matched haplotypes (Larmuseau et al. 2014).

1.7 Y chromosome analysis in forensic cases

Y chromosome analysis using Y-STRs has also become a fundamental pillar of DNA profiling, together with analysis of autosomal STRs. The important role of Y-STR analysis in forensic genetics is due to the unique inheritance pattern of the chromosome, which provides an informative result in many forensic cases (Goedbloed et al. 2009; Mulero et al. 2006). Crime statistics show that males are responsible for the highest percentage of violent crime and most sexual offences, and therefore Y-STR analysis can be usefully applied in these cases (Jobling et al. 1997). It is estimated that as many as 700 Y-STRs have the potential for use in forensic applications (Hanson and Ballantyne 2006; Willems et al. 2016). Y-STR analysis has been widely used in many cases including paternity testing, sexual assault (to amplify the male DNA component in male/female DNA mixtures), missing person identification and Disaster Victim Identification (DVI). Figure 1.8 shows the Y-STR markers commonly used in forensic analysis and their approximate locations on the Y chromosome.

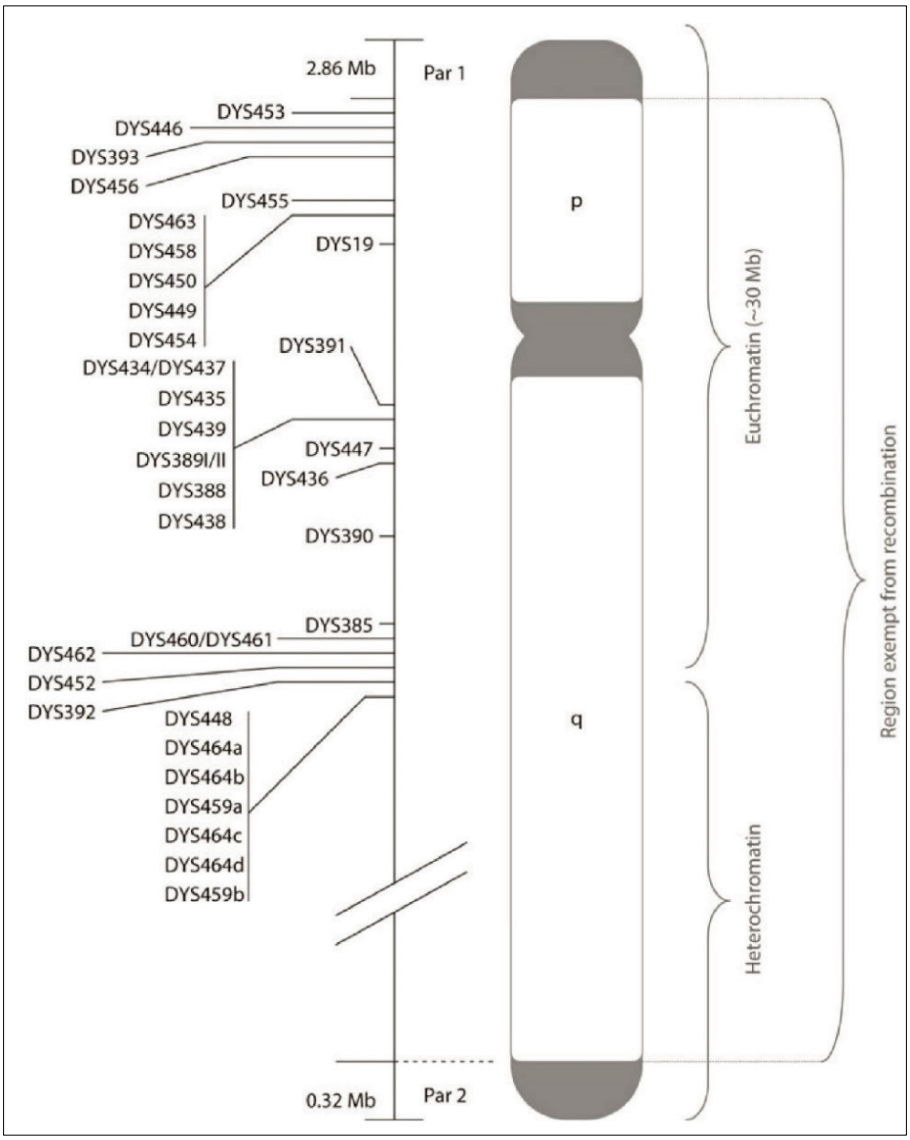


Figure 1.8: Commonly used Y-STR markers used in forensic analysis and their approximate locations on the Y chromosome (Buckleton et al. 2016).

1.7.1 Y chromosome results interpretation in forensics

As mentioned above, unlike autosomal STRs, the product rule cannot be used to calculate the significance of Y-chromosomal DNA evidence matches because Y-STRs are inherited together as one block, which has a major impact on the strength of evidence contributed by the Y result. However, the Y can still be informative in terms of exclusion. Using Y-STRs, inclusions can be problematic, and many factors can impact the strength of a result including numbers of Y-STRs used and their mutation rates, and population structure and relatedness (Jobling et al. 1997). As

discussed earlier, RM Y-STRs can to some extent help to increase the strength of evidence contributed by a Y haplotype match (Ballantyne et al. 2014). However, many challenges come with interpretation of RM Y-STR results. These include issues with reporting matches and non-matches. The fact that patrilineal relatives of a suspect are likely to be geographically close and socially similar affects even the most discriminating Y profiling method, and needs to be communicated in reporting (Andersen and Balding 2017)

1.7.2 Development of Y-STR Multiplex kits

The development of Y-STR analysis in forensic genetics has continued to progress by adopting new Y-STR markers. Starting with the initial suggested set of markers (DYS19, DYS389I+II, DYS390, DYS391, DYS392 and DYS393) known now as the minimal haplotype (de Knijff et al. 1997; Kayser et al. 1997), commercial companies have contributed to accelerated advances in Y-STR analysis in forensics. For example, by 2006 commercial Y-STR kits were available with varying numbers of markers: 17 STRs in the AmpFlSTR Yfiler® kit, 12 STRs in the PowerPlex® Y system and 11 STRs in the Y-Plex 12 kit (Johns et al. 2006). Recently, the number of Y-STR markers utilised in commercial kits has expanded, and the commercial companies have also started including some of the RM Y-STR markers in their kits. For example, PowerPlex® Y23 System contains 23 markers (Butler et al. 2012; Thompson et al. 2013) which includes two RM markers, namely DYS570 and DYS576 (Thompson et al. 2013). Lastly, Yfiler® Plus includes 27 Y-STR markers, Seven of which are RM markers (Phillips et al. 2014). Figure 1.9 shows relative fragment size of commonly used Y-STRs in three common commercial kits while Table 1.2 lists STRs markers including proposed RM markers.

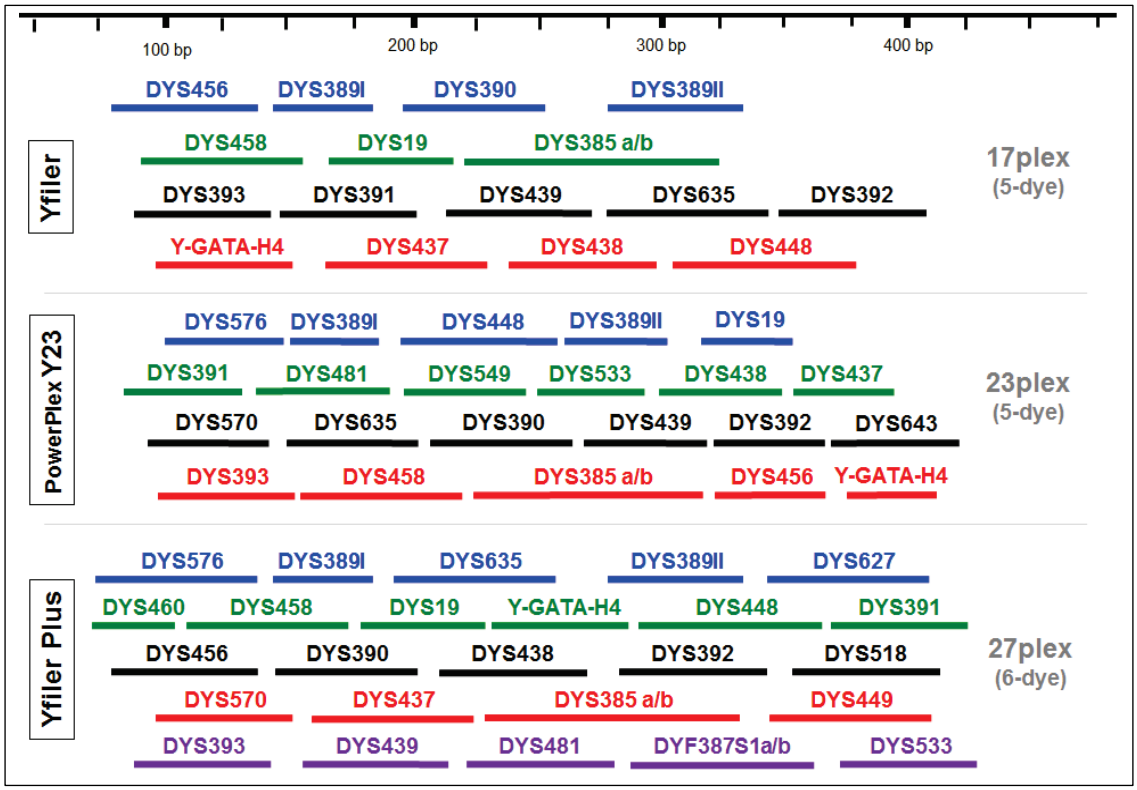


Figure 1.9: Relative scales for Y-STRs markers used in Yfiler®, Yfiler® Plus and PowerPlex® Y23 kit (Butler 2012). All markers give fragment sizes between ~75 bp and 425 bp.

Table 1.2: Comparison between Y-STRs in the commercial kits and proposed RM markers set (Ballantyne et al. 2010) and RM2 set (Willems et al. 2016). Same colours represent shared Y-STR markers between sets.

Yfiler® Plus	RM-1 set	PowerPlex® Y23	RM-2 set
DYS570 and DYS576	DYS570 and DYS576	DYS570 and DYS576	DYS570 and DYS576
DYS627 and DYS518	DYS627 and DYS518	DYS549 and DYS643	DYS549 and DYS543
DYS449	DYS449	DYS460	DYS442
DYF387S1	DYF387S1		DYS548
DYS389I	DYF403S1	DYS389I	Tetranucleotide 1
DYS635	DYF404S1	DYS635	Tetranucleotide 2
DYS389II	DYS526	DYS389II	Tetranucleotide 3
DYS19	DYS547	DYS19	Dinucleotide 1
DYS448	DYS612	DYS448	Dinucleotide 2
DYS391	DYS626	DYS391	Dinucleotide 3
DYS390	DYF399S1	DYS390	Dinucleotide 4
DYS438		DYS438	Dinucleotide 5
DYS392		DYS392	DYS461
DYS437		DYS437	DYS467
DYS385, DYS393 and DYS533		DYS385, DYS393 and DYS533	
Y-GATA-H4		Y-GATA-H4	Y-GATA-H4
DYS456, DYS439 and DYS481		DYS456, DYS439 and DYS481	DYS456, DYS439 and DYS481
DYS458 and DYS460			

1.8 The X chromosome as a forensic marker

The human reference X chromosome is 156,040,895 bp in length (GRCh38.p13), and can undergo recombination along its full length in females, who possess two X chromosome copies (homologous pair). In contrast, males possess only one copy of the X inherited from their mother, and in males the X chromosome undergoes crossing over only in the PARs shared with the Y chromosome. As a consequence, the X chromosome has a relatively low recombination rate. Because X chromosomes reside in males for a third of the time and in females for the other two thirds, the distribution of X chromosome types in the population is affected more by female histories than male histories (Schaffner 2004). As a result of this unique pattern of inheritance X-STRs have potential applications in forensic cases, although they are not widely used. Among these are complex parentage and missing persons investigations where X-chromosome analysis can help to establish a link between a male and his daughter (Szibor et al. 2003).

Because of the low recombination rate of the X chromosome, and because some X-STRs are physically close, it has been suggested that these close loci be considered as a "Linkage Group" (Szibor et al. 2003), rather than as independent markers, therefore, each LG is considered as a haplotype (similar to MSY and mtDNA). The product rule cannot be applied to loci within a LG when calculating RMP, but is still applicable between independent LGs.

Commonly used X-STRs are grouped into four linkage groups (Figure 1.10). For example, the loci in the Argus X-12 commercial typing system are classified within linkage groups as follows: LG 1 (DXS10135, DXS10148 and DXS8378), LG 2 (DXS7132, DX10074 and DXS10079), LG 3 (DXS10103, DXS10101 and HPRTB) and LG 4 (DXS7423, DXS10134 and DXS10146) (Nothnagel et al. 2012).

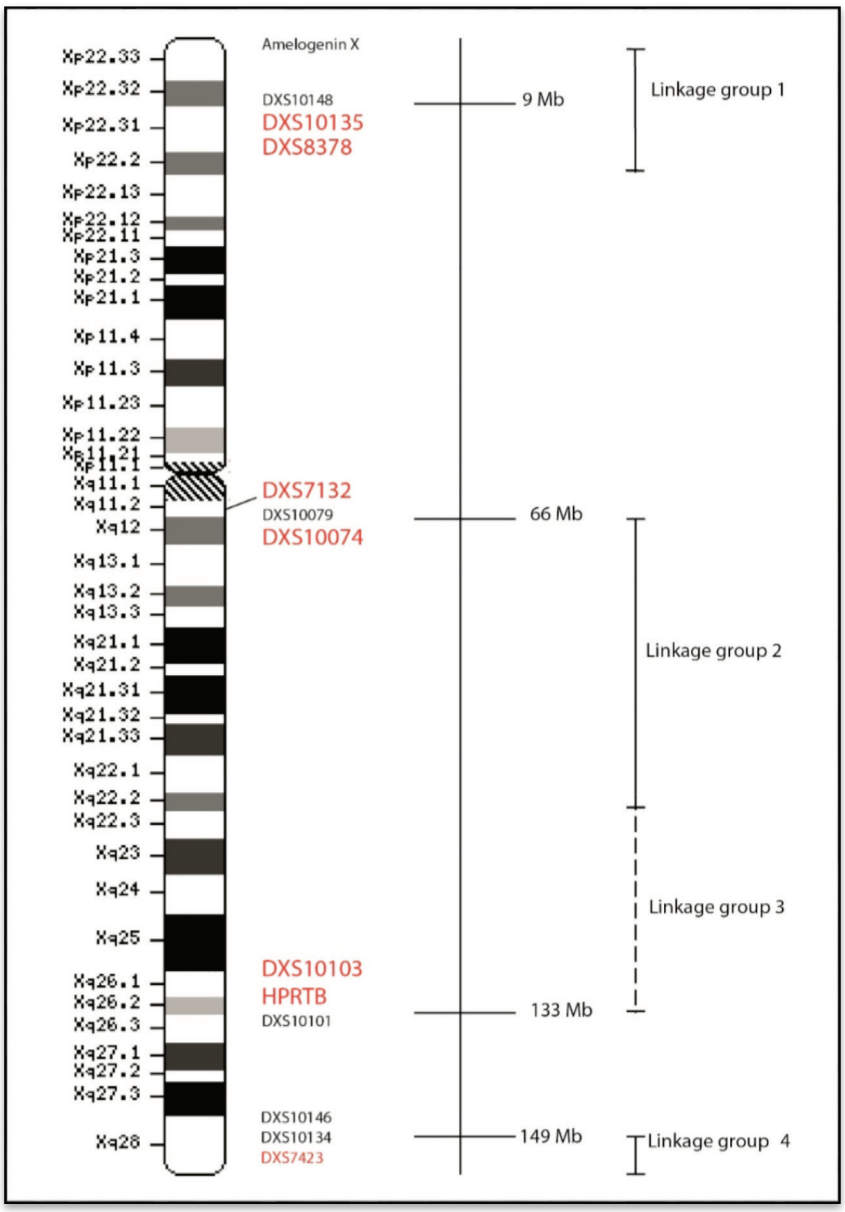


Figure 1.10: X-STR markers used in the Verogen ForenSeq™ DNA Signature Prep Kit along with other commonly used X-STRs with their approximate locations on the X chromosome (QIAGEN 2015).

1.9 Mitochondrial DNA in forensic genetics

Mitochondria are essential organellar structures that provide the cell with energy, and are abundant (100s - >1000 copies) in most cells (Yao et al. 2015). Mitochondria contain their own small genomes in the form of mitochondrial DNA (mtDNA), a circular double-stranded DNA molecule of about 16,569 base pairs (Figure 1.11) which encodes two ribosomal RNA (rRNAs), 22 transfer RNA (tRNAs) and 13 proteins (Anderson et al. 1981). Each mitochondrion contains several mtDNAs,

therefore mtDNA is many times more abundant overall that nuclear DNA, which means that mtDNA-based methods are highly sensitive.

Overall, mtDNA mutation rate is about ten times higher than that of the nuclear genome (Brown et al. 1979), probably because of reduced repair systems and the proximity to the mutagenic electron transport chain. The mutation rate also varies across the sequence (Soares et al. 2009), with a ten times higher rate still in the ~1.2-kb non-coding control region (D-loop). Although sperm carry small numbers of paternal mitochondria in the mid-piece, during fertilisation these are diluted by the tens of thousands of maternal mitochondria in the egg, and then apparently targeted and destroyed (Ankel-Simons and Cummins 1996). This means that mtDNA is inherited only through the maternal line with no recombination; hence, shared haplotypes will be found in matrilineages, giving mtDNA its low discrimination power. The primary forensic use is in highly degraded or trace (hair/bone) cases such as human remains (Holland and Parsons 1999).

Diversity of mtDNA is generally analysed by Sanger sequencing of PCR products, which can target the most variable control region, or encompass the whole molecule. As with the Y chromosome, mtDNA SNPs define a phylogeny of lineages known as haplogroups. These haplogroups are non-randomly distributed geographically, but do not show such high geographical specificity as is seen for the Y. Because of the female-line inheritance, the distribution of mtDNA types in populations reflects female history, and provides a counterpart to the male history reflected in the Y chromosome (Underhill and Kivisild 2007). Comparisons of the two can reveal past episodes of sex-biased admixture (Wilkins 2006).

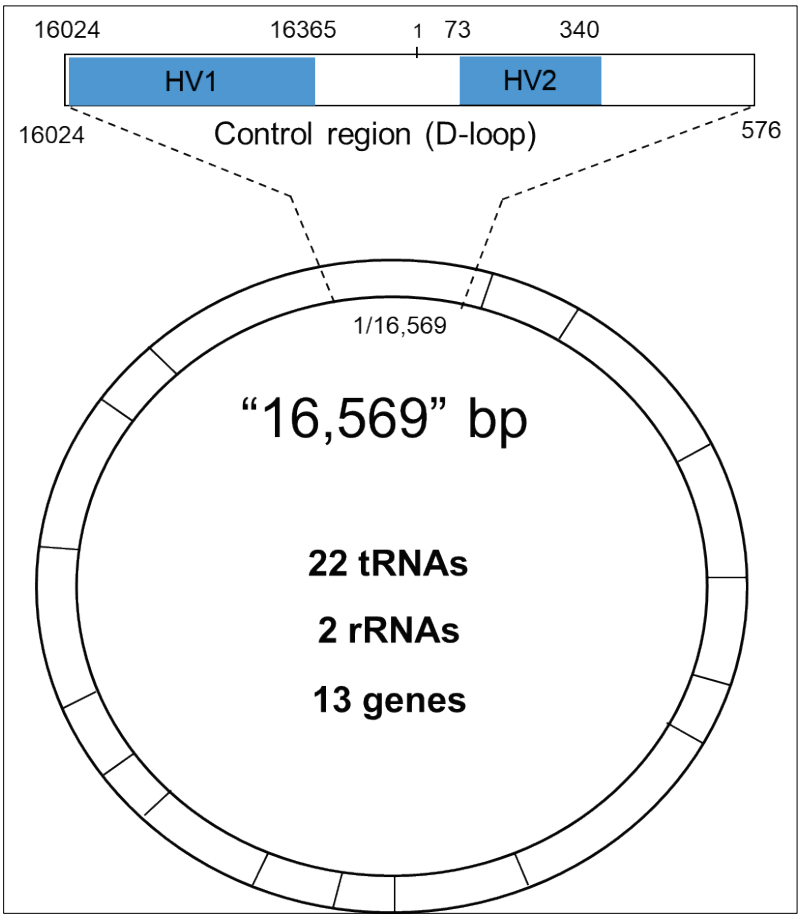


Figure 1.11: The human mitochondrial DNA molecule, showing the control region (D-loop) (Butler 2012). The 16,569 bp of sequence contains genes encoding two rRNAs, 22 tRNAs and 13 proteins.

1.10 Detection methods for STRs

Capillary Electrophoresis (CE) is the widely-used technique that is used to separate and detect amplified DNA fragments (van Oorschot and Ballantyne 2013). Applied Biosystems has provided different generations of Genetic Analyzer CE apparatus, starting from the ABI 310 single-capillary 4-dye-colour system to the most modern CE instrument (ABI 3500), which is capable of 6-dye detection with a multi-capillary system (Butler and Hill 2012). DNA fragment detection is based on two principles, separation by fragment size and differentiation via different fluorescent dyes attached to specific PCR primers. Negatively charged DNA fragments migrate through a narrow capillary filled with polymer (polydimethyl acrylamide, e.g. POP4, POP7) which works as a sieving environment. This allows the smaller DNA fragments to travel faster than the larger ones. DNA fragments move towards the

positive electrode (anode) under a high voltage. When DNA fragments pass the genetic analyzer window detector, a laser beam excites the fluorescently-labelled fragments. Light emission is captured by a CCD camera. Automatically, the genetic analyzer software will separate DNA fragments based on colour and size (Figure 1.12).

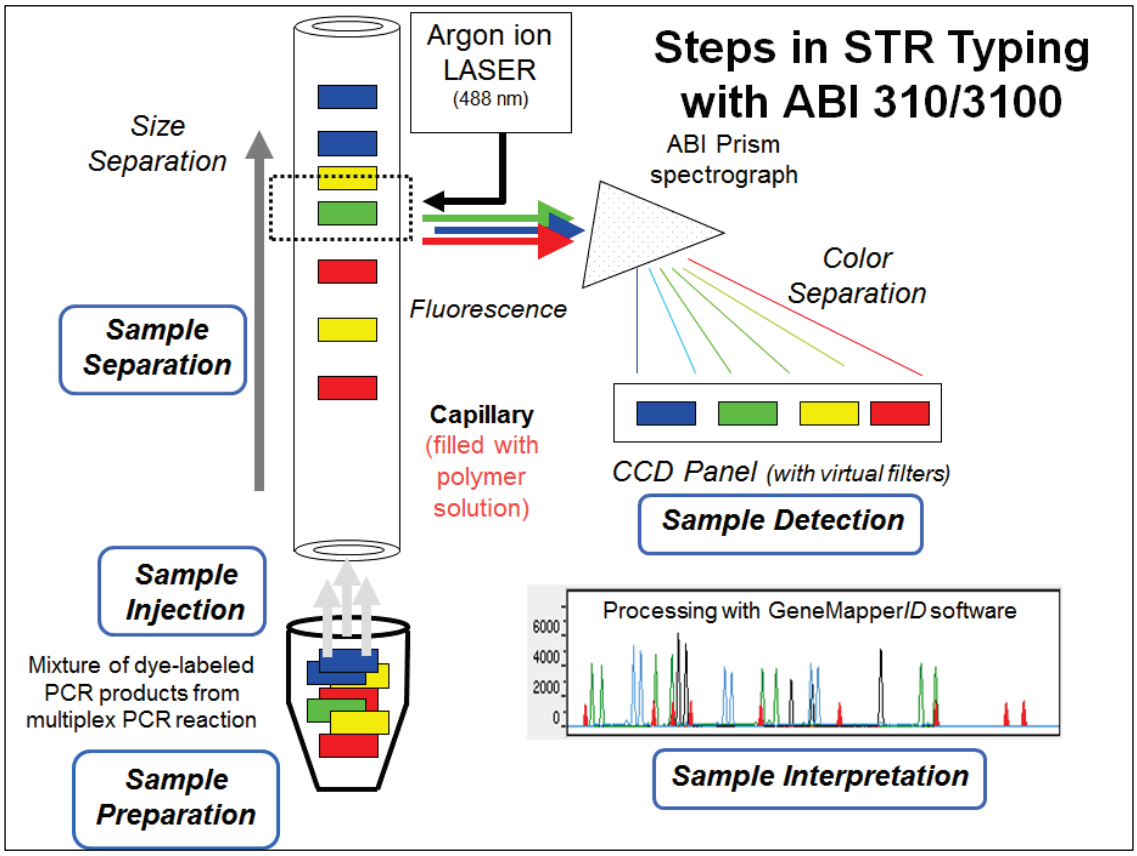


Figure 1.12: Illustration of the genetic analyzer capillary electrophoresis process for DNA fragment detection (Butler 2012). Separation is by fragment size and colour using a 5-dye-colour system. A CCD camera captures the emitted light from excited fragments and uses analytical software for reading the raw data.

Accurate genotyping in STR analysis requires use of an 'allelic ladder' (Butler 2007), which is a mixture of DNA fragments that includes a collection of pre-amplified and labelled known alleles. The allelic ladder is used to aid accurate genotyping of unknown samples. The alleles in the allelic ladder and in the sample are all sized using the same size standard and run in the same experiment, to match similar conditions (Butler 2012; Hartzell et al. 2003; Raziel et al. 2012), as shown in Figure 1.13

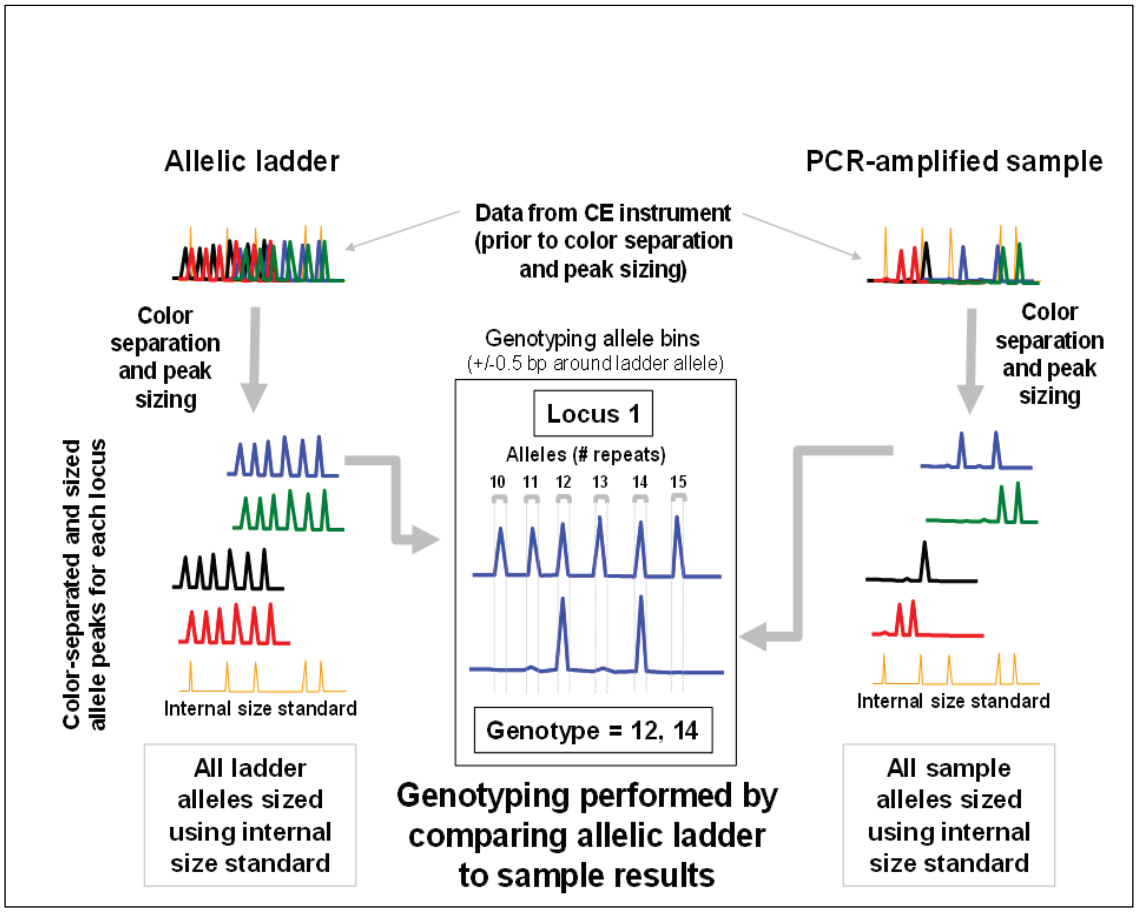


Figure 1.13: STR calling by capillary electrophoresis. The internal size standard allows sizing of DNA fragments both for the unknown samples and the allelic ladder alleles. Following that, alleles in an unknown sample are compared to an allelic ladder for designation (Butler 2012).

1.11 DNA sequencing approaches to STR diversity

As described above, the traditional approach to STR diversity analysis is via CE. This approach considers only allele length, ignoring potentially informative internal sequence information. This, for example, could include different repeat unit types within an array, plus indels, SNPs and also variation in the flanking region. Early attempts to capture the DNA sequence information within STRs used Sanger sequencing, which, for diploid autosomal loci, requires the initial separation of the two alleles. Such work was needed for the designation of repeat counts for alleles of particular lengths.

1.11.1 Sanger sequencing

Sequencing DNA using chain-termination (Sanger method) was introduced in 1977 by Frederick Sanger. The principle is based on incorporation of a ddNTP which has no hydroxyl group in the 3´ position which prevents further addition of nucleotides (Sanger et al. 1977). The sequencing reaction starts with annealing a primer to the target fragment. A DNA polymerase such as *Taq* polymerase starts the synthesis of a new complementary sequence by incorporating one of the four complementary dNTPs. However, once a ddNTP is incorporated from a low-concentration pool, the chain terminates, leading to the sequencing reaction containing a mix of terminated products of different fragment sizes (Figure 1.14). Separation of these fragments used to be carried out using polyacrylamide gels which required nucleotide-specific radio labeling through a time-consuming process. However, currently sequencing can be carried out using fluorescently labelled ddNTPs which facilitate automation of the sequencing process. Currently detection is carried out in the same genetic analyzer machines that are used for STR profiling. Sanger sequencing was the method used for the Human Genome Project, and still has applications for small fragment sequencing.

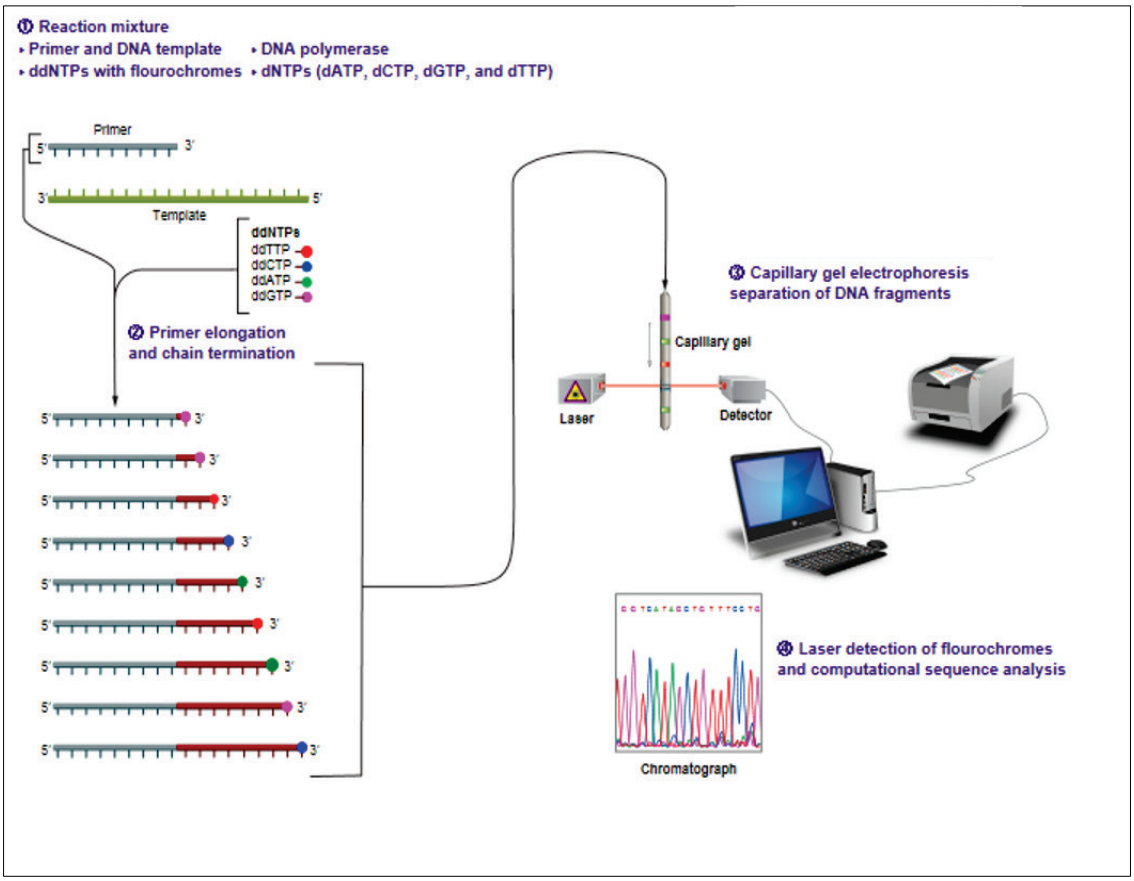


Figure 1.14: Principle of Sanger sequencing. Incorporation of ddNTPs results in differently- labelled fragment sizes. Separation is based on fragment sizes, and a CCD camera captures emissions from fluorescently labelled incorporated ddNTPs. Image modified from "Sanger sequencing," by Estevezj (CC BY-SA 3.0).

1.12 Massively parallel sequencing

A further limitation of CE in STR typing is that it also limits the number of loci that can be simultaneously analysed, because of the limited numbers of dyes that can be detected (currently six), and also the problem of size overlap in amplified fragments within the same colour channel, which needs to be avoided. This limitation has recently been overcome by the introduction of MPS. Technologies based on MPS arose as a result of demand for large-scale genome sequence data that was more cost and time effective. They have now revolutionised medical genetics and population genetics, and are making an impact in forensic genetics.

Since early 2005 several technologies have been introduced, involving different

chemistries (Goodwin et al. 2016). However, as the name implies, the methods have in common the key feature that many sequencing reactions are carried out in parallel. Most technologies have also seen an increase in sequence read length over time (Mardis 2013); for example, the most popular technology, Illumina's sequencing-by-synthesis method, began with ~35-bp reads but now can yield reads of up to 500 bp.

1.12.1 MPS in forensic analysis

In forensic genetics, MPS can provide high throughput and discover all the sequence variation in STR amplicons, and also analyse numerous fragments simultaneously. Thus an approach based on MPS can in principle overcome traditional and Sanger sequencing limitations of low throughput (Churchill et al. 2016). Also, MPS approaches have the advantage that primers can be placed very close to STR repeat arrays, because there is no need to space out amplicon sizes across the 120-400 bp range as in CE. This provides more chances for successful analysis of forensically challenging samples such as degraded DNA (Iozzi et al. 2015). In addition to amplifying many STRs as short fragments, MPS can also potentially analyse other DNA fragments, e.g. containing SNPs, or from mtDNA (Churchill et al. 2016).

1.12.2 Forensic MPS commercial solution

Commercial MPS platforms are now available for forensic applications, and provide analysis of different forensic markers as summarised in Table 1.3. ThermoFisher Scientific commercialised two SNP typing kits in 2014 designed for the Ion PGM System released 2012 (Børsting and Morling 2016). These were followed by a 10-plex for sequencing ten autosomal STRs (Fordyce et al. 2015) and finally the Precision ID GlobalFiler® NGS STR panel which simultaneously amplifies 30 STRs. Currently commercially available ThermoFisher Scientific kits allow the sequencing of autosomal STRs, mtDNA and a SNP panel; however, there is still no kit available for the analysis of X and Y-STR markers. ThermoFisher also does not combine different markers (e.g. aSTRs and iiSNPs) in the same kit, but provides these markers separately. PowerSeq™ Systems by Promega is another platform initially released and validated for forensic application in 2015 (Gettings et al. 2016; Zeng et

al. 2015a; Zeng et al. 2015b). Currently two commercial kits are available: the PowerSeq™ 46GY System kit for autosomal and Y-STRs, and the PowerSeq™ CRM Nested System kit for the mtDNA control region including HVI, HVII and HVIII analysis (Holland et al. 2019). Multiple prototypes include the PowerSeq™ Auto/Mito/Y System kit which targets autosomal STRs, Y-STRs and the mtDNA control region (Huszar et al. 2019) but these are still not commercially available. In addition, to date, there is no Promega kit for SNP typing available on the market and Promega has not developed its own software for MPS DNA analysis yet.

The MiSeq FGx™ platform was launched by Illumina in January 2015 as a new validated MPS system designed for forensic applications (Caratti et al. 2015). Currently there are two kits available from Verogen: one is the ForenSeq™ DNA Signature Prep Kit and the other is the more recent ForenSeq™ mtDNA Control Region Solution Prep Kit, with a new mtDNA Whole Genome kit planned for the end of 2019. Verogen provides an analytical software (i.e. ForenSeq™ Universal Analysis Software) for the analysis and visualisation of the STR and SNP sequencing results. The software also enables statistical analysis for ancestry and phenotype based on a set of selected SNPs. Also, it allows users to carry out comparison analysis of the sequencing results between two samples.

Table 1.3: Comparison between available MPS systems for forensic analysis.

MPS systems	Autosomal STRs	Y-STRs	X-STRs	SNPs	MtDNA	Analytical software	Commercial multiplexing kit
ThermoFisher	✓	✗	✗	✓	✓	✓	✗
Promega	✓	✓	✗	✗	✓	✗	✓
Verogen	✓	✓	✓	✓	✓	✓	✓

1.12.3 Overall workflow and library preparation

Different MPS methods follow overall similar steps, which include genomic DNA preparation, library preparation, amplification, sequencing and data analysis as illustrated in Figure 1.15. Sequencing can be conducted upon platforms that utilise different chemistries for example: Illumina, Ion Torrent and Pacific BioSciences and 10X Genomics (van Dijk et al. 2018). While the following sections describe in detail the MPS workflow and sequencing methods related to the ForenSeq™ kit on the MiSeq FGx™ platform, the principles outlined below are not specific to this platform.

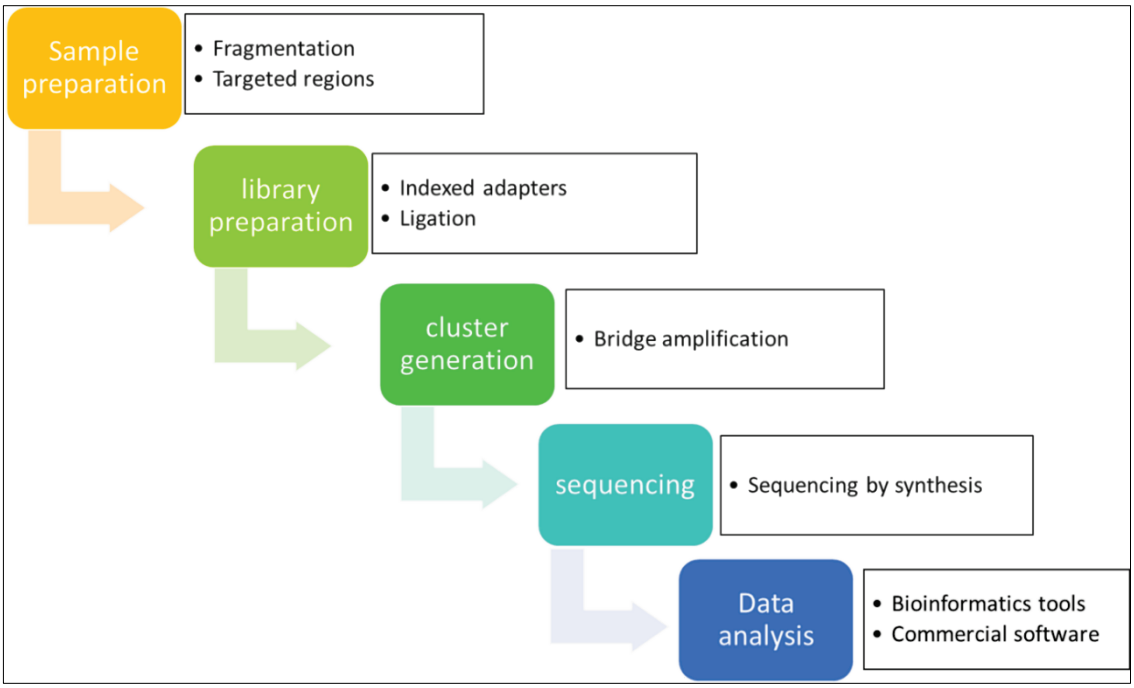


Figure 1.15: Generic MPS workflow. In this project, template DNA was prepared by amplifying target regions; indexed adapters were used in library preparation, bridge amplifications and sequencing by synthesis were used in the sequencing process.

The workflow involves four main steps, namely: library preparation, cluster generation, sequencing, and data analysis. The practical work can be completed in three days and sequencing takes 27 hours on the MiSeq FGx™. ForenSeq library preparation starts from either extracted genomic DNA, or samples such as saliva. In addition, samples on FTA cards can be processed after a wash step. Targeted genome regions are amplified in PCR1 using ForenSeq™ DNA Signature primers

which have tags (or tails) that allow ligation of unique indexed adapters that are added in the second PCR. Index sequences are used to identify each sample amplicon, while adapter sequences complementary to the immobilised flow-cell oligos are added to allow hybridization to the flow-cell surface Figure 1.16. The flow-cell is a glass slide with lanes (8 channels) that are coated with two types of oligo. A wash step then purifies the libraries and gets rid of residual primers, while a normalisation step using magnetic beads aims to achieve equal concentration of each individual library in the final pool.

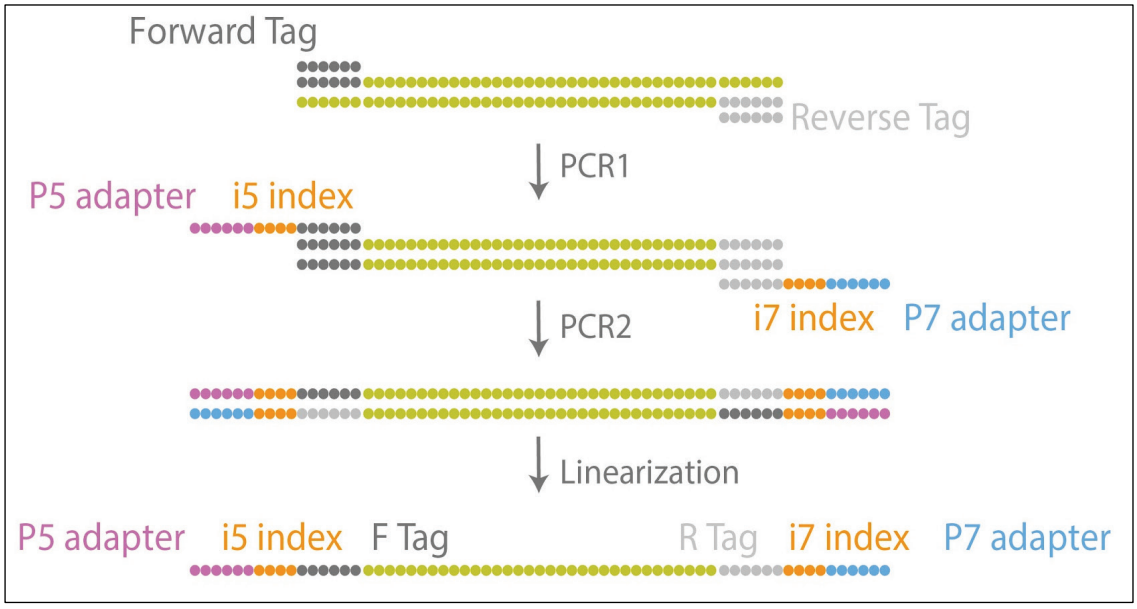


Figure 1.16: ForenSeq library preparation Target regions were amplified in PCR1 using tagged primers, followed by addition of indexed adaptors to tagged regions in PCR2. Final fragment must have i5 and i7 indices. Modified from (Illumina 2015a).

1.12.4 Cluster Generation

Amplified fragments are now attached to one of the two oligos on the surface of the flow cell (Figure 1.17). Each DNA fragment is amplified isothermally in a process called bridge amplification. First, a polymerase makes complementary DNA sequences from hybridised fragments. Generated double-stranded DNA fragments are then denatured and the original template is washed away. The remaining strand folds over allowing the second adapter to attach to the other oligo on the flow-cell

surface. DNA polymerase generates the complementary strand of this bridge to produce two complementary single-strands following a denaturation step. In this way the process of bridge amplification will continue simultaneously for all different fragments to make thousands of copies, which form a cluster. Finally, one oligo sequence will remain and the other is cleaved and washed away.

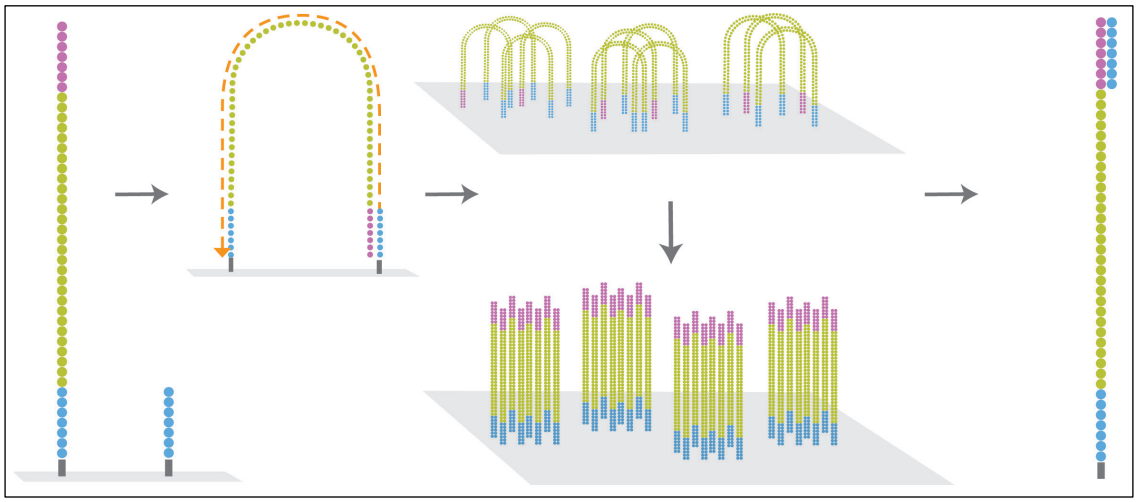


Figure 1.17: Cluster Generation. Fragments are hybridised to the flow-cell surface. Bridge amplification is carried out to generate a cluster for each fragment. Finally, one oligo sequence will remain and the other is cleaved and washed away. Modified from (Illumina 2015b).

1.12.5 Sequencing by synthesis (SBS)

Specific to Illumina®, an important feature of SBS is the 'cyclic reversible termination' reaction. Incorporated nucleotides have a blocked hydroxyl group as in Sanger sequencing, but a key difference is that this is reversible. All four fluorescently-labelled nucleotides are added in each cycle. Following the incorporation and excitation of nucleotides an image is taken to record the fluorophore. After imaging, the blocking group is removed and washed away, and the next cycle can then start again until the first read is complete (Figure 1.18). ForenSeq carries out 301 cycles for the first read but only 31 cycles for the second read. Sequence can either be read from one end only of a fragment (single-end sequencing), or from both ends (paired-end).

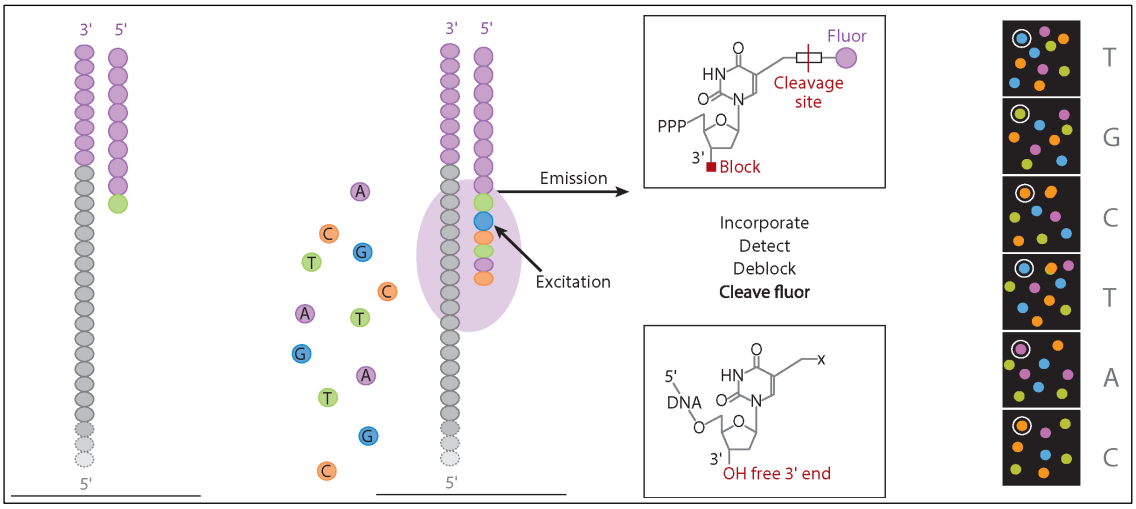


Figure 1.18: Sequencing by synthesis (Illumina 2015b) All four fluorescently-labelled nucleotides are added in each sequencing cycle. Imaging takes place to record emission of incorporated nucleotides.

1.12.6 Multiplexing and de-multiplexing

In targeted sequencing studies where sequences are determined in many individuals, DNA 'bar-coding' or 'indexing' can be carried out. Each sequencing library made from the DNA of a given individual has a short (e.g. 6 bp) specific sequence incorporated after the adapter primer. After sequencing, this motif can be recognised bioinformatically, and allows sequence reads to be assigned to particular individuals. The Illumina technology described above allows the simultaneous shotgun sequencing of millions or even billions of clusters. In order to sequence multiple samples in a single run (96 samples) combinations of unique indices of A5 and R7 are added for each sample. These indices are 6 bp in length and there are 12 different R7 sequences and 8 different A5 sequences, enough to differentiate up to 96 samples Figure 1.19.

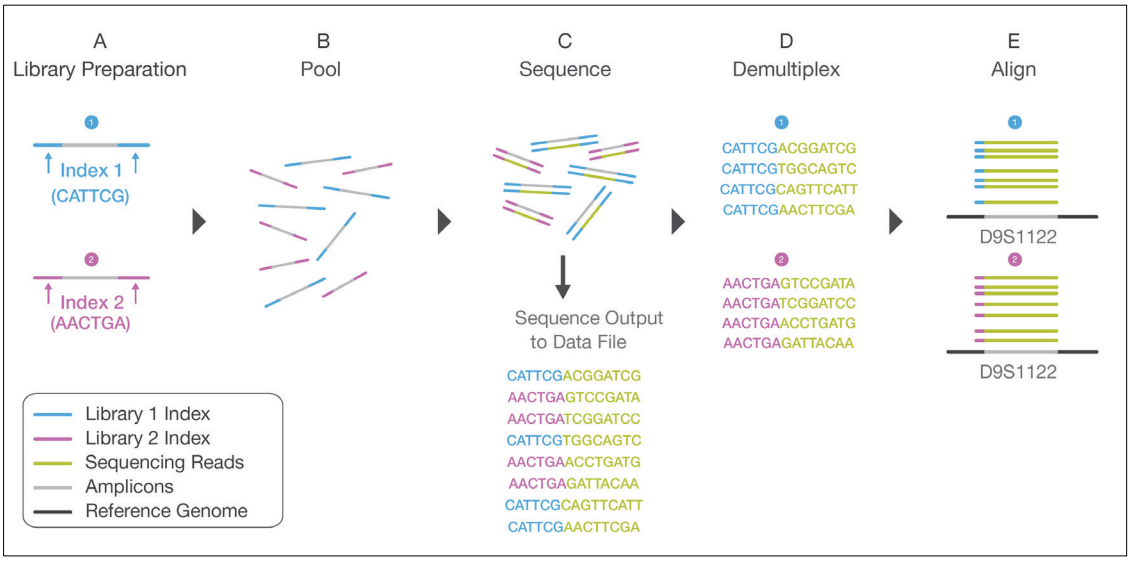


Figure 1.19: Multiplexing and de-multiplexing Each sequencing library has a unique combination of short index sequences (6 bp); after sequencing these indices are used to distinguish individual source DNAs in the “de-multiplexing” process.

1.12.7 The ForenSeq™ DNA Signature Prep Kit as a forensic tool

The ForenSeq™ DNA Signature Prep Kit (Verogen), released in 2015, exemplifies the advantages of an MPS approach by allowing simultaneous amplification of either >150 loci including the standard autosomal and Y-STRs plus X-STRs and identity-informative SNPs (iiSNPs), or >230 loci which also include biogeographical ancestry- and phenotypically-informative SNPs. The kit has been validated through a range of performance tests, including robustness, reproducibility, concordance with CE and sensitivity of detection (Almalki et al. 2017; Churchill et al. 2016; Guo et al. 2017; Iozzi et al. 2015; Jäger et al. 2017; Just et al. 2017; Köcher et al. 2018; Sharma et al. 2017; Silvia et al. 2017; Xavier and Parson 2017). The kit has also been used on challenging samples to evaluate its applicability to real forensic cases (Almohammed et al. 2017; Bodner et al. 2016; Köcher et al. 2018): in these tests, it compared favourably with CE, for example on formalin-fixed paraffin-embedded tissue, as well as on ancient bone samples, detecting a greater number of informative markers in seven out of ten cases (Churchill et al. 2016). Even though

the CE-based approach produced marginally more data from STRs, the combined data from both STRs and iiSNPs in the DNA Signature Prep Kit provided greater resolution in challenging samples (Votrubova et al. 2017).

The DNA Signature Prep Kit has also improved resolution of both male/male and male/female mixtures involving minor contributors as low as 1:20 (Churchill et al. 2016; Jäger et al. 2017; Köcher et al. 2018; Xavier and Parson 2017) and this capability has proved useful in the first sexual assault court case using MPS in the Netherlands. The kit has also been implemented in casework by the INPS (Institut National de Police Scientifique) laboratory in Lyon, with MPS profiles uploaded to the French national DNA database, and the FBI has approved the kit, the MiSeq FGx™ System and the UAS for the US National DNA Index System within the terms of newly published SWGDAM guidelines for MPS (SWGDM 2019).

1.13 The need for population studies and local reference databases

In forensic genetics, the introduction of new methods and new loci requires extensive validation, and a key part of that is exploring how new systems handle genotypes that were not encountered in their developmental validation. As Middle East populations are not usually used in the developmental validation process it is vital that they are thoroughly examined before implementation in this region. Furthermore, in order to evaluate evidence in the Middle East, population reference data are needed.

1.14 Saudi Arabia

This thesis will focus on the Kingdom of Saudi Arabia, located in the southwest of Asia, and a strategic junction between the three old-world continents of Asia, Africa and Europe. It provides an important junction between Africa and Asia (Petraglia et al. 2015). Saudi Arabia constitutes more than 70% of the Arabian Peninsula and is the second largest Arabic country in terms of land area. The country is bordered by Iraq and Jordan to the north, Kuwait, Qatar and the United Arab Emirates to the east,

Yemen and Oman to the south, and The Red Sea to the west. Egypt, Sudan and Eritrea face Saudi Arabia over the Red Sea, and Iran on the other side of the Arabian Gulf.

Administratively, Saudi Arabia is divided into 13 regions (Figure 1.20) which are distributed into five geographical areas namely: Central (C) (Riyadh, Al-Qassim), Northern (N) (Northern borders region, Tabuk, Al-Jawf and Hail), Southern (S) (Asir, Jazan, Bahah and Najran), Eastern (E) (Eastern province) and Western region (W) (Makkah and Madinah) as shown in the map below (Figure 1.20).

Figure 1.20: Saudi Arabia administrative map (Modified from (Memish et al. 2012). Five main geographical regions are highlighted using different colour codes.

1.14.1 Historical importance of Saudi Arabia

Beside its geographical location, Saudi Arabia provides interesting historical reasons for population genetic exploration. Firstly, it is the home of an Arabic population. Secondly, Saudi Arabia is the birthplace of Islam which started in the 7th century (Hourani 1991), and had important demographic consequences later on for North Africa and Iberia. Furthermore, many studies show significant records of ancient human habitation and civilization in the Arabian Peninsula (Groucutt et al. 2015). Saudi Arabia lies close to the ancestral homeland of the human species within Africa and of its proximity to both of the routes thought to have been used as humans migrated out of Africa either via the strait of Bab-el-Mandeb (southern route) or the northern route makes it a possible reservoir of genetic variants that were carried by these migrants. Recent discoveries of fossil bones (estimated as 85,000 years old) in Saudi Arabia add supporting evidence to the prehistoric presence of *Homo sapiens* in the region (Groucutt et al. 2018). Such findings might be surprising if we consider only the current desert climate. However, several studies show that the Arabian Peninsula's climate in the past provided a suitable environment for human habitation (Jennings et al. 2015; Parton et al. 2015). Figure 1.21 shows an illustration of current annual rainfall compared to that in the last interglacial period ~130 KYA (thousand years ago) which suggests that Saudi Arabia was much wetter than at present (Jennings et al. 2015).

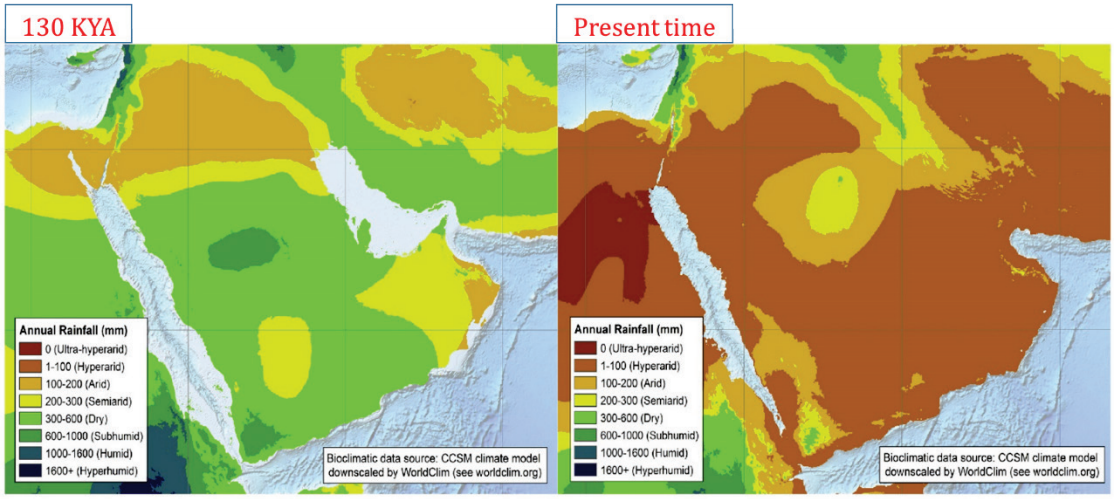


Figure 1.21: Illustration of rainfall at the present time and 130 KYA (Jennings et al. 2015).

The tribal system was and remains the commonest form of governance in traditional populations within Saudi Arabia, in which every independent tribe has its own leader, although, many rulers have established a form of rule over these tribal systems in the past. Previous populations and civilisations have inhabited Arabia for many thousands of years, however, the province of Arabia was first unified beginning in the 7th century AD (called the Islamic era) with the empire ruling over both Saudi Arabia and other nations. The first Saudi kingdom was established in the 18th century, leading to the current Saudi kingdom (third) which was established only about 100 years ago (Al-Rasheed 2010).

The discovery of oil in the last century led to a rapid population boom, as increased wealth improved health care and provided the infrastructure to support large populations even in the harsh environment (Al-Rasheed 2010; Wynbrandt and Gerges 2010). The original population has increased rapidly and has been supplemented by new immigrants which currently account for around a third of the total population (Figure 1.22).

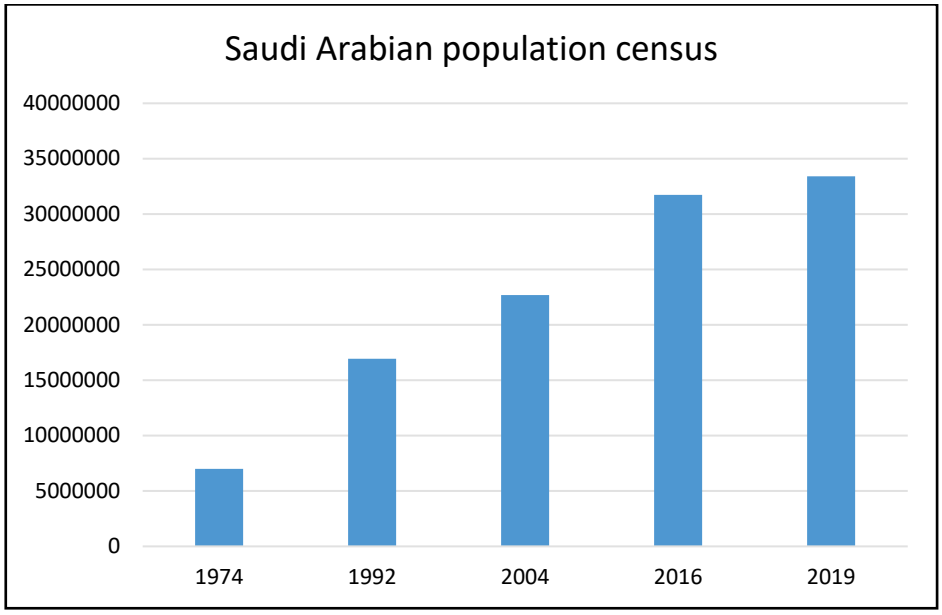


Figure 1.22: Rapid Saudi population growth between1974-2019.Data obtained from Saudi Arabian General Authority for Statistics (www.stats.gov.sa, accessed 2/09/19).

1.14.2 Consanguinity and population structure

Culture and customs in human populations differ and can have a major impact on genetic diversity. This has implications for understanding history, for medical genetics, and for forensic data interpretation. In many human populations, marriage between related individuals is frequent, and is encouraged for socioeconomic reasons. The Middle East is known from ethnographic (Bittles and Black 2010) and genetic studies (Leutenegger et al. 2011; Monies et al. 2019) to have high consanguinity (defined as marriages between second-cousins or closer (Bittles 2001). Saudi Arabia has among the highest frequencies of consanguinity, with reported rates between 55 and 69% (Al-Abdulkareem and Ballal 1998; El-Hazmi et al. 1995; El-Mouzan et al. 2007). Such practices are expected to reduce genetic diversity, and there may be a sex-biased effect here. In turn these effects could vary from region to region, leading to population structure. It is important to understand these phenomena from the forensic perspective, since there are implications for the significance of matching profiles, and the appropriateness of population databases.

1.15 Aims and objectives of the current study

This thesis will investigate the genetics of the Saudi population by applying current DNA-variant analysis techniques, including Y- and autosomal STR typing using CE-based kits, and also MPS-based approaches that have the potential to reveal a finer-grained picture of diversity.

Aim 1: To assess the pattern of Y-chromosomal diversity across Saudi Arabia and ask if there is geographical population structuring within the country, and to assess differences with other neighbouring populations.

Objectives:

- To type the CE Y-STR multiplex Yfiler® Plus (27 Y-STRs) in a large sample of Saudi Arabian males, sampled to represent the different geographical areas
- To use population genetic methods to compare regions within Saudi Arabia
- To use population genetic methods to compare Saudi Arabia with other neighbouring populations
- To ask if mode of recruitment (either within Saudi Arabia, or among expatriates in the UK) provides consistent results

Aim 2: To assess pattern of autosomal STR genotype diversity across Saudi Arabia and address issues of consanguinity, geographical population structuring and differentiation with other neighbouring populations

Objectives:

- To type the CE autosomal STR multiplex GlobalFiler® (24 autosomal STRs) in a large sample of Saudi Arabian individuals
- To use population genetic methods to ask if there is any signal, in the form of reduced heterozygosity, of consanguinity
- To use population genetic methods to compare regions within Saudi Arabia

based on autosomal data

• To use population genetic methods to compare Saudi Arabia with other neighbouring populations, based on autosomal data

Aim 3: To assess patterns of diversity at the sequence level among Saudi individuals

Objectives:

• To apply the ForenSeq™ DNA Signature Prep Kit multiplex, which includes autosomal and Y-STRs, plus X-STRs and autosomal SNPs, to a sample of Saudi males

• To catalogue the observed diversity and assess forensic statistics based on sequenced markers

• To compare Y- and autosomal STRs at the sequence level with the same markers at the CE level

• To ask if signals of consanguinity are revealed by sequencing autosomal STRs and SNPs

• To assess diversity based on X-STRs

• To use population genetic methods to compare the Saudi sample with populations from neighbouring countries

Chapter 2 Materials and methods

This Chapter will describe the experimental and analytical methodology used in this thesis. Materials including, chemical reagents and kit contents are listed in Appendix (Materials 9.2). All experiments were conducted following good laboratory practice to prevent the occurrence of contamination. Protective clothing (laboratory coats, gloves, face masks and hair cover) were used when appropriate. In addition, positive and negative controls were used to test for contamination. Laboratory waste disposal was conducted following University of Leicester disposal routes.

2.1 DNA sampling

Ethical review for recruitment of human DNA donors and subsequent analysis was provided by the Saudi General Administration for Forensic Evidence and the University of Leicester Research Ethics Committee (ref: 4945-maj4-genetics). Informed consent was provided by all participants. Sampling was either in the form of blood spots on Flinders Technology Associates (FTA) cards (Whatman Bioscience, UK) (Nuchprayoon et al. 2007), buccal swabs (King et al. 2006), or from saliva samples via the Oragene•DNA (OG-500) kit (DNA Genotek). Details of sampling strategy and individuals sampled for each experiment are given in the corresponding chapters.

2.2 DNA extraction

DNAs were extracted and purified from blood spots on FTA cards using a fully automated forensic DNA processing STARlet workstation (Hamilton) using the PrepFiler® Forensic DNA Extraction Kit (ThermoFisher Scientific). A 1.2-mm diameter punch, made using the BSD Punching System (ThermoFisher Scientific), was used in this reaction. Automated DNA extraction with QIAamp DNA Mini Kits and the QIAcube (QIAGEN) method was used for saliva samples collected via the Oragene•DNA kit (200 µl) and for 200-µl buccal swab washes in NDS buffer (Anand 1986) (0.5M EDTA, 10mM Tris-HCl, 1% [w/v] sodium lauroyl sarcosine, pH9.5). The

principle of the QIAamp method is based upon DNA binding to a silica membrane under high-salt conditions. Following washing, the genomic DNA is recovered from the silica membrane using a low-salt elution buffer. Extracted samples were stored at -20°C for subsequent analysis.

2.3 DNA quantification

DNA was quantified either by spectrophotometry at 260 nm in a NanoDrop 2000c (ThermoFisher Scientific) device, or by Real-Time PCR using the Quantifiler® kit in an Applied Biosystems® 7500 Real-Time PCR Machine.

2.3.1 Real-Time PCR quantification

The principle of Real-Time PCR quantification is based on the masking effect of a quencher dye at the 3´ end of a probe, which prevents any fluorescence emission from the reporter dye at the 5´ end. When the PCR process starts, the *Taq* polymerase extending from the forward primer cleaves this attachment and frees the reporter dye, allowing for emission (Figure 2.1). Measuring the emissions for test samples, and comparison with a standard curve of emission against PCR cycle, allows DNA quantity to be estimated. In the PCR process two specific primers are used to amplify a specific region in the human genome as well as using an internal PCR control (IPC) for measuring any inhibition effect in the samples.

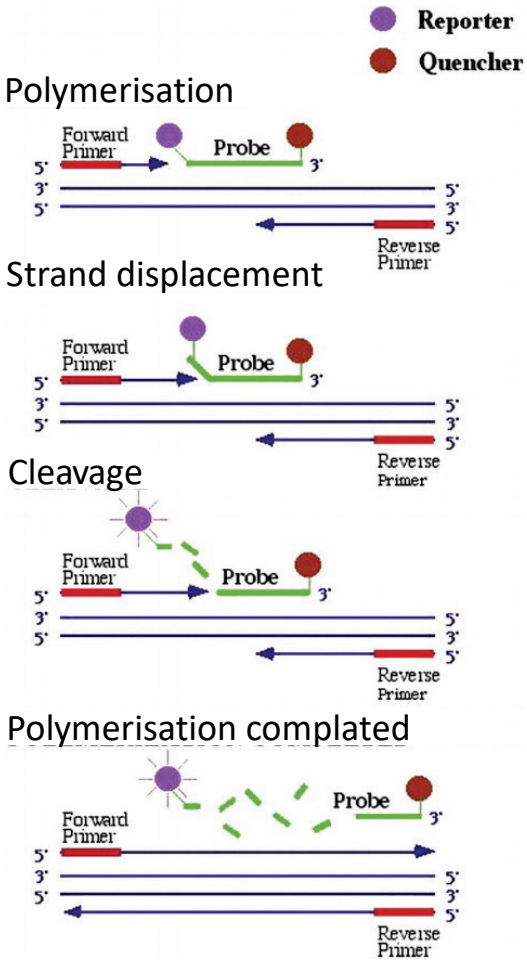


Figure 2.1: Quantification assay using Real-Time PCR (Schaad et al. 2003). The quencher blocks fluorescence emission from the reporter dye when attached.

2.3.2 Real-Time PCR using the Quantifiler® kit

The process starts with preparation of the DNA standard: eight serial dilutions are prepared between 50 ng/µl and 0.023 ng/µl. Reactions are prepared from 23 µl PCR mix and 2 µl of DNA sample, whether standard or test, in 96-well optical plates. Plates are sealed prior to centrifugation, followed by loading into the Applied Biosystems 7500 following the manufacturer's recommendations. During exponential amplification (40 PCR cycles), the system records the fluorescence after each cycle. Quantification of the unknown samples is carried out by comparison of the cycle number (Ct) at which the exponential amplification results in fluorescence exceeding the threshold value in comparison to the standard curve. Accuracy of quantification of the standards can be assessed from the correlation coefficient R^2 (ideally = 1) and slope (~ -3.32) of the standard curve. The run input and output are controlled by Applied Biosystems SDS Software v1.2.3.

2.3.3 NanoDrop

The NanoDrop (Desjardins and Conklin 2010) was used according to manufacturer's standard protocols, and required 1 µl DNA in solution. To test the purity of DNA, the ratio of absorbance at 260 nm to 280 nm was used, where a ratio of 1.8 indicates ideal sample purity.

2.4 Targeted amplification of human loci

The polymerase chain reaction (PCR) is one of the fundamental techniques in molecular biology. The steps of a PCR cycle are denaturation to obtain single-stranded DNA, followed by an annealing step which allows primers to attach to the target region, then an extension step in which polymerase synthesizes the complementary strand. PCR-based methods were used to target a number of loci within the human genome.

2.4.1 Y-chromosomal STR amplification with the Yfiler® Plus PCR Amplification Kit

Samples were amplified using the six-dye Yfiler® Plus kit (6-FAM™/Blue, VIC™/Green, NED™/Yellow, TAZ/™Red, SID™/Purple and LIZ™/Orange for the GeneScan™ 600 LIZ™ Size Standard (ThermoFisher)) following the manufacturer's protocol. PCRs contained 10 µl Yfiler® Plus master mix, 5 µl Yfiler® Plus primer set (including fluorescently-labelled primers), and 1 ng DNA template, for a final volume of 25 µl. For the positive control, 1 ng 007 control DNA (supplied by ThermoFisher) was used. PCRs were run in a Veriti® 96-Well Fast Thermal Cycler (Applied Biosystems) starting with 1 min at 95°C (hot start, to activate the polymerase); this was followed by 29 cycles of denaturation (4 s at 94°C), annealing and extension 1 min at 61.5°C, ending with a hold at 4°C.

2.4.2 Autosomal STR amplification with the GlobalFiler® PCR Amplification Kit

The GlobalFiler® kit was used to amplify autosomal STRs. It is also based on a six-dye system similar to Yfiler® Plus. Following the manufacturer's protocol, PCRs were prepared by combining 7.5 µl of Master Mix and 2.5 µl Primer Set Mix with 15 µl for each sample (1 ng DNA template in the final reaction volume of 25 µl) and for the positive control 10 µl of control DNA (0.1 ng/µl) with 5 µl of low-TE buffer. The Veriti thermal cycler was used to perform 29 PCR cycles as follows: initial hold 1 min at 95°C, 29 cycles of 4 s at 94°C (denaturation) and 1 min at 61.5°C (annealing

and extension). Final extension was performed at 60°C for 20 min.

2.4.3 DNA amplification with the ForenSeq™ DNA Signature Prep Kit

ForenSeq library preparation starts from either extracted genomic DNA, or samples such as saliva. In addition, samples on FTA cards can be processed after a wash step. Targeted genome regions are amplified in PCR1 using ForenSeq primers which have tags (or tails) that allow ligation of unique indexed adapters that are added in the second PCR. Index sequences are used to identify the source individual, while adapter sequences complementary to the immobilised flow-cell oligos are added to allow hybridisation to the flow cell.

2.4.3.1 Library preparation

The first PCR amplifies and tags target loci. A master mix is prepared by combining 5.0 µl DNA Primer Mix A (DPMA), 4.7 µl PCR1 Reaction Mix and 0.3 µl Enzyme Mix. 10 µl per well was distributed in a 96-well plate. Reactions included either 5 µl of test DNA sample (1 ng/µl), a diluted positive control 2800M (from 2 µl 2800M stock with 98 µl nuclease-free water (NFW)), or NFW as a Negative Control. The plate was then loaded into the Veriti thermal cycler with the following setup: Initial hold for 3 min at 98°C followed by two rounds of PCR cycles. In the first round, eight cycles were performed as follows: 96°C for 45 s, 80°C for 30 s, 54°C for 2 min and 68°C for 2 min. This was followed by the second round as follows: ten cycles of 96°C for 30 s and 68°C for 3 min. The procedure ended with a 10 min incubation at 68°C.

The indexing process takes place in a second PCR. A unique combination of the Index Adapters (short DNA sequences) (i5 and i7) must be added to all reactions. PCR is performed as follows: 98°C for 30 s, followed by 15 cycles of 98°C for 20 s, 66°C for 30 s, 68°C for 90 s; ending with a 68°C incubation for 10 min.

2.4.3.2 Library purification

45 µl of Sample Purification Beads was added to a new midi plate prior adding 45 µl from the previous PCR. The plate was shaken at 1800 revolutions per minute (rpm) for 2 min, then incubated at room temperature for 5 min. Supernatants were

discarded after placing the plate on a magnetic stand (Invitrogen™). Two washes with 200 µl freshly prepared 80% (v/v) EtOH were carried out. Finally, 52.5 µl Resuspension Buffer (RSB) was added, and the plate shaken at 1800 rpm for 2 min, followed by a 2 min room temperature incubation. The plate was placed on the magnetic stand, then 50 µl supernatant transferred to a new PCR plate for normalisation.

2.4.3.3 Library normalisation

The normalisation step aims to achieve uniform targeted cluster density across DNA libraries. 45 µl from a mixture of Library Normalization Additives (LNA1) (46.8 µl) and Library Normalization Beads 1 (LNB1) (8.5 µl) were transferred to a new midi plate prior to adding 20 µl of DNA library. The plate was shaken for 30 min at 1800 rpm. Three wash steps using 30 µl of Library Normalization Storage buffer were carried out. 32 µl newly prepared 0.1 N HP3 (NaOH) were added to each library. Following a shaking step for 5 min at 1800 rpm, 30 µl supernatant was transferred to a new PCR plate.

2.4.3.4 Denaturation

5 µl of each library were pooled in a 1.5-ml microcentrifuge tube. 7 µl from this pooled library was transferred to a new microcentrifuge tube containing 591 µl Hybridization Buffer, 2 µl from a mixture which prepared from 2 µl of Human Sequencing Control and 2 µl HP3 (2N NaOH) and 36 µl NFW. The tube was then placed in a heating block for denaturation at 96°C for 2 min, then immediately transferred to ice. The whole contents were then loaded into a reagent cartridge for sequencing on a Verogen MiSeq FGx™ device in Forensic Genomics mode, following the manufacturer's protocol.

2.4.4 PCR amplification of the mtDNA control region

Amplification of the mtDNA control region using the primers L15999 (5´-CACCATTAGCACCCAAAGCT-3´) and H409 (5′-CTGTTAAAAGTGCATACCGCC-3′) as one amplicon (Sigurðardóttir *et al,* 2000) was performed in a reaction containing:

11x PCR buffer (0.90 µl) (Jeffreys et al. 1990), 1M Tris base (0.125 µl), 10µM primer 1 (0.3 µl), 10µM primer 2 (0.3 µl), 20:1 *Taq*:*Pfu* polymerase mix (0.06 µl), prepared from 5 µl of 2.5U/µl Pfu added to a 50 µl tube of *Taq* polymerase at 5 U/µl, NFW and 10 ng genomic DNA. The PCR included a 95°C hot-start step of 3 min followed by 31 cycles of 94°C for 30 s, 60°C for 30 s and 70°C for 30 s.

ExoI/SAP purification was used to remove excess single-stranded DNA (primers) and deactivate excess unincorporated dNTPs. 5 µl of PCR product were combined with 0.5 µl ExoI (20 U/µl), 1.5 µl recombinant Shrimp Alkaline Phosphatase (1 U/µl) and 1.3 µl ExoI buffer (10 x) (New England Biolabs). This purification mixture was then incubated in a PCR machine as follows: 37°C for 1 h; 80 °C for 15 min; 4°C for 15 min.

To confirm presence of the targeted PCR product and estimate relative amounts of product, agarose gel electrophoresis was performed. 1 µl of DNA sample, 2 µl loading dye and 2 µl water were run on a 1.5% (w/v) agarose gel in 1 x TBE, with 2 µl of HyperLadder™ 1 kb as a size marker. 0.5 µg/ml ethidium bromide was used to stain the products. Gels were run in 1× TBE at 10 V/cm and products viewed and pictured using a UV transilluminator and GeneSnap image acquisition software (Syngene). The relative intensities of product were estimated compared to the size marker, to determine the required input for the sequencing reaction.

Sanger sequencing reactions were prepared as follows: Big Dye Terminator mix v 3.1 (1 µl), 5 x Big Dye Terminator Buffer (3.5 µL), 3.2 µM sequencing primer (1 µl) and 20-30 ng/kb PCR product were combined with NFW to take the final volume to 20 µl PCR was performed as follows: 95°C for 3 min followed by 25 cycles of 96°C for 10 s, 50°C for 5 s and 60°C for 4 min. Removal of excess dye from sequencing product was carried out by adding 2 µl of 2.2% (w/v) sodium dodecyl sulfate (SDS) followed by incubation for 5 min at 98°C and 10 min at 25°C. Purification was performed using a QIAGEN DyeEx 2.0 spin column following the manufacturer's protocol. Samples were then sent for sequencing using the University of Leicester's sequencing facility PNACL (Protein and Nucleic Acid Chemistry Laboratory).

2.4.4.1 Run parameters of STRs within GlobalFiler® and Yfiler® Plus PCR Amplification Kits

A master mix for capillary electrophoresis injection was prepared comprising 9.6 µl formamide and 0.4 µl ILS 600 (internal lane standard) per well. Following this 1 µl of amplified DNA was added to each well, plus 1 µl of allelic ladder. DNA was denatured at 95°C for 3 min, followed by cooling on ice. Following the manufacturer's recommendations for electrophoresis under conditions described in Table 2.1, amplified samples were run on an ABI 3500 or 3130xl Genetic Analyzers (Applied Biosystems).

Table 2.1: Capillary electrophoresis run parameters ABI 3500 Genetic Analyzers.

Injection Parameters	**GlobalFiler®**	**Yfiler® Plus**
Polymer type	POP 4	POP 4
Capillary	36 cm	36 cm
Injection time	15 s	16 s
Injection voltage	1.2 kV	1.2 kV
Run condition	13 kV/1550 s	13 kV/1550 s
Run Temperature	60°C	60°C
Dye set	J6	J6

2.4.5 mtDNA

2.4.5.1 Sanger sequencing

The electrophoretic separation stage of mtDNA sequencing was carried out by (PNACL) on an Applied Biosystems 3730 Genetic Analyzer. For samples that showed low quality sequencing results because of the presence of uninterrupted polyC tracts, resequencing was carried out as illustrated in Figure 2.2. in order to capture missing sequences The two additional primers (5′-CATGCTTACAAGCAAGTACAGC-3′ and 5′-GCTGTGCAGACATTCAATTG-3′) were designed using Primer3Plus (Rozen and Skaletsky 2000). PCR primers were produced by Sigma-Aldrich at 0.025 mMol scale, and dissolved in NFW to make 10 µM stocks.

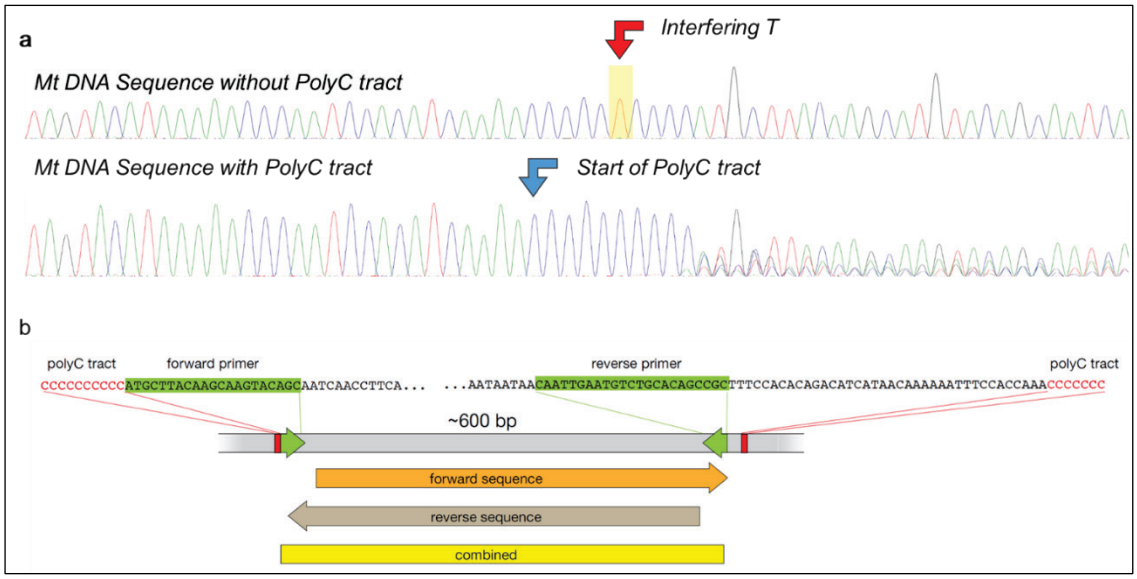


Figure 2.2: Resequencing of samples with uninterrupted polyC tracts a) shows an example of two samples with and without polyC tracts. b) scale of the mtDNA sequence showing two region of polyC tracts and position of the additional primers highlighted in green.

2.4.5.2 Sequencing procedure for the ForenSeq™ DNA Signature Prep Kit

In Illumina sequencing on the MiSeq platform, DNA fragments are immobilised on the surface of a flow-cell. Bridging amplification then generates separate clusters which each contain ~1000 identical molecules, within which sequencing will take place. Sequencing primers are annealed to each template molecule in the clusters, and a specialised DNA polymerase incorporates a fluorescently labelled terminating nucleotide complementary to the template base. Fluorescence imaging with a camera identifies the incorporated nucleotide. A cleavage step then removes the chain terminator and the fluorescent dye before the next incorporation step. This process continues cyclically, allowing the sequence of the molecule in each cluster to be read. The sequence can be read from both ends (paired-end).

Before performing the run, the reagent cartridge was kept in a water bath at room temperature for about 1 h to completely defrost, and was then inverted to resuspend and dissolve any precipitates. 600 µl pooled denatured library was

loaded into the sample well in the reagent cartridge. A flow cell was removed from storage buffer, rinsed thoroughly with NFW then dried, making sure it was clear of any streaks, fingerprints or fibres before loading it into the MiSeq FGx™. SBS Solution, a waste bottle and the reagent cartridge were loaded into the sequencer when prompted by the UAS software. The run was then performed following the manufacturer's protocol.

2.5 Data analysis

2.5.1 Analysis of STR data

STR data were analysed using GeneMapper® ID-X v1.4 software. This uses a matrix file to separate and size the STR fragment peaks based on the dye colour and an internal standard. Fragments in the unknown samples and DNA controls were also compared to an allelic ladder (a mixture of DNA fragments representing the common alleles in the general population) to provide accurate allele designations. Standard criteria for interpretation of forensic STRs were applied including, for example: stutter (additional PCR products resulting from polymerase slippage, usually one repeat unit smaller than the true allele) below 15%, minimum peak height 75 relative fluorescence units (RFU) for heterozygous alleles, 150 RFU for homozygous alleles, and balance between heterozygous alleles ≥60%.

2.5.2 Analysis of mtDNA sequence data

Sequence Scanner (Applied Biosystems) was used for initial interpretation of sequence data and checking data quality. Following that, BioEdit (Hall 1999) and CodonCode Aligner were used to visualise and interpret DNA sequences collectively; some 'N' nucleotides were resolved manually after visual inspection of the sequence traces. The ClustalW alignment program within MEGA6 (Tamura et al. 2013) was used to align mitochondrial control region sequences, applying default parameters. The beginning and the ends of sequences were trimmed to remove low-quality sequence. The mtDNA nomenclature tool within the mtDNAprofiler analysis software (http://mtprofiler.yonsei.ac.kr:8080/index.php?cat=1) was used to

prepare input files for haplogroup predictions, which were done using HaploGrep 2 (v2.1.19) (https://haplogrep.uibk.ac.at/) (Weissensteiner et al. 2016). The output of this software also provided an index of the quality of prediction based on typed variants within each sequence. True distances between haplotypes based on sequence variation were visualised through a median-joining network using Network 5.0 (Bandelt et al. 1999) applying a transition: transversion ratio of 1:10), and annotated using Network Publisher (http://www.fluxus-engineering.com/nwpub.htm).

2.5.3 Analysis of MPS data from the ForenSeq™ DNA Signature Prep Kit

2.5.3.1 ForenSeq™ Universal Analysis Software (UAS)

In targeted sequencing studies where sequences are determined in many individuals, DNA 'bar-coding' or 'indexing' can be carried out. Each sequencing library made from the DNA of a given individual has a short (e.g. 6-bp) specific sequence incorporated after the adapter primer. After sequencing, this motif can be recognised bioinformatically, and allows sequence reads to be assigned to particular individuals. The UAS is an analysis software suite provided for analysis of ForenSeq data. It allows run setup, live run monitoring, sequence data analysis, and reporting. In run setup, i5 index and i7 index are assigned to each sample. During the run, UAS presents a monitoring screen which shows cluster density, cluster passing filter, phasing and pre-phasing which appears after cycle 26. All three runs in this project passed the quality metrics recommendations and per run results. UAS also performs sequence alignment, allele calling, and genotyping displaying the result in a user-friendly interface. Genotype calls are generated using pre-defined analytical thresholds.

The user can access the UAS interface through a web browser and undertake the interpretation process (Liu and Harbison 2018). Negative controls should show no signs of any product while positive controls should show sequencing results for all markers in the primer mix. One feature of the software is a display showing the sample representation for the run. Total reads per sample of less than 85,000 will be highlighted in orange. Such measurements provide a global picture of the run;

however, interpretation should be conducted by analysis of each sample, in which flagged loci require particular attention. UAS can generate three different reports: sample level genotype, project level genotype and flanking region. Both sample and genotype reports provide genotype, depth of coverage (DoC), and repeat region sequence for STRs. Flanking region reports, which are automatically produced prior to user interpretation, provide sequence information for both STRs and iiSNPs; however, these reports provide all raw sequence that has DoC of ≥10 and no user interpretation can be involved prior to report generation.

2.5.3.2 Use of the Short Tandem Repeat allele identification tool STRait Razor

STRait Razor (Woerner et al. 2017) is a program to call and characterise STRs, indels and SNPs in MPS data. It can be run in common operating systems such as Microsoft Windows. The process starts by running a STRait Razor Perl script for haplotype calling for a set of pre-defined loci using the matching of 5´ and 3´ anchor sequences. This is followed by an annotation and calling step using the Strait Razor Excel workbook. This provides length- and sequence-based genotyping, locus haplotype data and nomenclature using a nomenclature database which encompasses more than 2700 haplotypes; any allele not in the list is highlighted as 'Novel'.

2.5.3.3 Bioinformatic analysis of MPS from raw data

As well as considering the UAS report, bioinformatic analysis for the raw data was also carried out in some cases where result confirmation was needed. Analysis of the raw data quality was checked using FastQC (Andrews 2010). Adaptors and poor-quality bases were removed using Trimmomatic (Bolger et al. 2014). The Burrows-Wheeler Alignment tool (BWA) (Li 2013) was used to map reads to the human genome reference sequence GRCh38. BAM files were visualized via IGV 2.3 (Robinson et al. 2011). BCFtools (v1.8) was used for iiSNP calling from Human Genome Diversity Project (HGDP) data, with minimum base quality 20 and mapping quality 20 (Li 2011).

Chapter 3 Analysis of 27 Y-STRs in Saudi Arabian males

3.1 Introduction

Saudi Arabia is the largest country in the Arabian Peninsula. Its population of ~33 million people is distributed highly non-uniformly (Figure 3.1), with very low densities in its large desert areas, but high densities concentrated around a small number of cities. Its indigenous Arab people (~62% of the population; www.stats.gov.sa, accessed 2/09/19) are historically organised into geographically-differentiated patrilineal descent groups, or tribes (Al-Rasheed 2010), with a tradition of consanguinity (Al-Gazali et al. 2006). This geographical and social organisation might be expected to have an effect on patterns of genetic diversity, particularly regarding the male-specific region of the Y chromosome (MSY), which in turn could have implications in interpretation of DNA profiles.

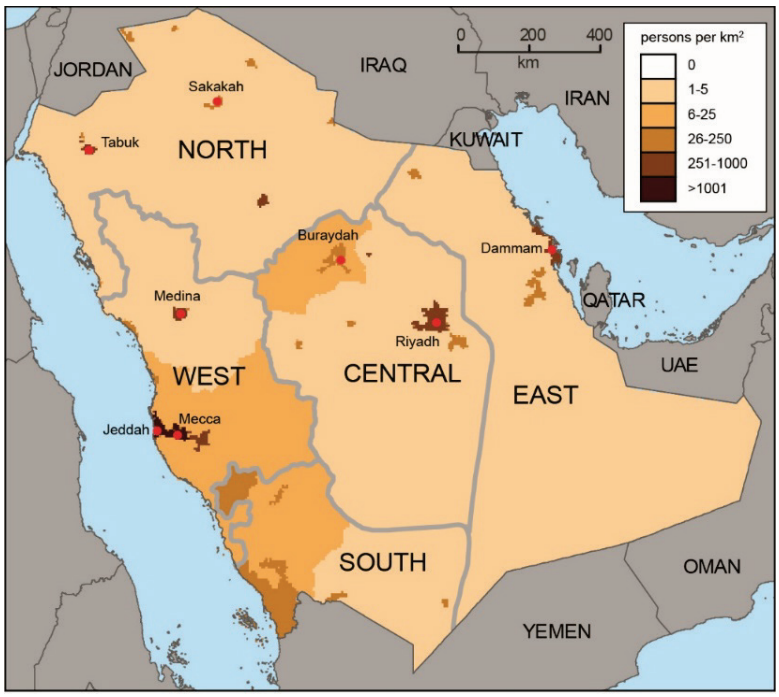


Figure 3.1: Map of Saudi Arabia, showing population density and sub-regional divisions used in this study. Population density is indicated by shading as shown in the key, top right, and locations of some major cities are shown. Adapted from Global Rural-Urban Mapping Project (sedac.ciesin.columbia.edu/gpw/), under a Creative Commons 3.0 Attribution License. Administratively, Saudi Arabia is divided into 13 regions which here are considered as five larger geographical areas, namely: Central (Riyadh, Al-Qassim), Northern (Northern Borders, Tabuk, Al-Jawf and Hail), Southern (Asir, Jazan, Bahah and Najran), Eastern (Eastern Province) and Western (Mecca and Medina).

Population genetic studies on Saudi Arabia to date are limited. Exome sequencing of a set of samples from the Arabian Peninsula including Saudi individuals demonstrated relatively high inbreeding coefficients (Scott et al. 2016), consistent with a history of consanguineous marriage. A general analysis of Saudi Arabian mitochondrial DNA (mtDNA) diversity (Abu-Amero et al. 2007) showed a pattern of haplogroups similar to that of other Arabian Peninsula samples. In another mtDNA-based study (Abu-Amero et al. 2008) - the only example to divide Saudi Arabia sub-regionally - central, northern, western and southeastern sub-groups formed a single cluster in a multi-dimensional scaling (MDS) analysis when compared to other Arabian Peninsula samples, but also presented significant inter-group differences. Y-chromosome studies have analysed the seven Y-STRs defining the minimal haplotype (Alshamali et al. 2009), or haplogroup-defining SNPs together with 17 Y-STRs (Yfiler®) for one specific haplogroup (Abu-Amero et al. 2009). The first of these (Alshamali et al. 2009) revealed lower diversity in Saudi Arabia than in populations from outside the Arabian Peninsula, and affinity between Saudi Arabia and Yemen, which together were strongly differentiated from Oman and Dubai. It was speculated that this might be due to the influence of patrilineal descent and polygyny. The second study (Abu-Amero et al. 2009) showed that haplogroup J1 was the most prominent lineage (42%) in the Saudi Arabian sample studied, and that genetic distances based on haplogroup frequencies were relatively small among Arabian Peninsula samples. The focus of Y-STR typing on one lineage precludes any population-based conclusions on haplotype diversity from this study.

To date, therefore, while some general studies have been carried out, little has been done to characterise population structure within Saudi Arabia. Knowledge of any such structure is important in the interpretation of the significance of DNA-based forensic evidence, and in the construction of appropriate databases. The aims of this Chapter are:

(i) to use the 27 Y-chromosomal short-tandem repeats (Y-STRs) in the Yfiler® Plus kit to characterise haplotypes in 597 Saudi males sub-divided by geographical region.

(ii) to consider the relationships of Y-chromosome diversity between

regions within the country and also between Saudi Arabia and other surrounding populations;

(iii) to compare the spectrum of Y-chromosome types in males recruited within Saudi Arabia with that of regionally-matched males recruited in the United Kingdom, to ask if social structuring also influences patterns of Y-haplotype diversity.

3.2 Materials and Methods

3.2.1 DNA sampling

Five hundred and ninety-seven DNA samples were collected from indigenous Saudi Arabian males. Of these, 503 were extracted from blood spots on FTA cards (Whatman, UK), sampled from individuals recruited within Saudi Arabia itself. The remaining 94 were extracted from buccal swabs (King et al. 2006), or from saliva samples via the Oragene•DNA kit (DNA Genotek), from Saudi males resident within the UK. In each case, males with ancestry (to the level of paternal great-grandfather) from five geographical subdivisions of the country shown in Figure 3.1 (Central, Northern, Southern, Eastern, and Western) were sampled, and consideration of relatedness ensured that all sampled males were separated by at least three generations. Ethical review for recruitment and analysis was provided by the Saudi General Administration for Forensic Evidence and the University of Leicester Research Ethics Committee. Informed consent was provided by all participants.

3.2.2 DNA Extraction and Quantification

DNAs were extracted and purified from FTA blood-spot samples using a fully automated STARlet workstation (Hamilton) and the PrepFiler® Forensic DNA Extraction Kit (ThermoFisher Scientific), starting from 1.2-mm diameter punches produced using the BSD100 Punching System (Microelectronic Systems). Buccal samples were extracted via QIAamp DNA Mini Kits on a QIAcube robotic workstation (QIAGEN). All DNA samples were quantified using the Quantifiler® Human DNA Quantification Kit (ThermoFisher Scientific) on an Applied

Biosystems® 7500 Real-Time PCR System.

3.2.3 DNA amplification and fragment detection

The Yfiler® Plus PCR Amplification Kit was used to generate Y-chromosome haplotypes for the 27 STRs DYS19, DYS385a, DYS385b, DYF387S1/S2, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439, DYS448, DYS449, DYS456, DYS458, DYS460, DYS481, DYS518, DYS533, DYS570, DYS576, DYS627, DYS635, and Y-GATA-H4. PCRs were conducted as recommended by the manufacturer on a Veriti (ThermoFisher Scientific). Fragments were detected using an ABI3500 or ABI3130xl Genetic Analyzer (ThermoFisher Scientific) using the manufacturer's recommended protocols. GeneMapper® ID-X software v1.4 was used for allele calling and interpretation.

3.2.4 Haplogroup prediction and assessment of accuracy

Y-SNP haplogroups were predicted from Y-STR haplotypes using the online Y-DNA Haplogroup Predictor NevGen (http://www.nevgen.org), which is based on a previously-implemented Bayesian approach (Athey 2006), with the additional consideration of pairwise correlation of alleles between different Y-STRs in the calculation of haplogroup probabilities.

3.2.5 Median-joining networks

Median-joining networks (Bandelt et al. 1999) were constructed using the software Network 5.0 and Network Publisher (http://www.fluxus-engineering.com/sharenet.htm). In the case of intermediate alleles, repeat numbers were rounded to the nearest integer or down if the intermediate allele was half the length of a full repeat; constitutively duplicated loci (DYS385a,b; DYF387S1,S2) were removed for network construction. Deletion alleles were coded '99' in input files, and thereby considered as missing data.

3.2.6 Forensic and population genetic parameters

For each sample or sub-sample, haplotype diversity was calculated using the

formula $n(1-\Sigma p_i^2)/(n-1)$, where n is the sample size and p_i the frequency of the ith haplotype. Haplotype match probability (HMP) was estimated as the sum of squares of the haplotype frequencies. Discrimination capacity (DC) was calculated as the ratio between the number of distinct haplotypes and the total number of haplotypes in the sample.

Rst calculations based on Y-STR data and multi-dimensional scaling (MDS) plots (Kruskal 1964) were carried out using comparative population data and the calculation tool within the online Y Haplotype Reference Database (https://yhrd.org/amova). The square plots produced by this approach were graphically adjusted to Euclidean space for display purposes.

Population differentiation tests based on predicted haplogroup frequencies were carried out within Arlequin (Excoffier and Lischer 2010) based on predicted haplogroups, using a method analogous to Fisher's exact test. Haplogroup-based gene diversity was also calculated in Arlequin.

Based on published comparisons of the performance of sequence- and STR-based dating (Hallast et al. 2015), time-to-most-recent-common-ancestor (TMRCA) was estimated by the average-squared distance (ASD) method (Goldstein et al. 1995a, b), using 23 STRs omitting the duplicated STRs (DYS385a,b; DYF387S1,S2), and also using a reduced set of 18 omitting the Rapidly Mutating STRs (DYS449, DYS518, DYS570, DYS576 & DYS627). The modal haplotype was used as a root, and the mean pedigree mutation rate across STRs was used as measured in father-son pairs and available from yhrd.org/pages/resources/mutation rates.

3.3 Results

The 27 Y-STRs contained in the Yfiler® Plus kit were amplified in DNAs from 597 Saudi Arabian males (Appendix, Table 9.1) contains a full list of haplotypes, as well as other sample information. Haplogroups were also predicted from each STR haplotype, using the prediction tool NevGen, and tested prediction accuracy based on a large independent set of Y-STR data and known Y-SNP haplogroups.

3.3.1 Y-STR allele and haplotype diversity within Saudi Arabia

Considering allelic diversity, the dataset is characterised by a very high proportion of individuals (424/597; 71%) carrying intermediate alleles, in particular .2 alleles at DYS458; this is a known characteristic of haplogroup J1 (Ferri et al. 2008), and immediately suggests that this haplogroup strongly predominates in the sample. In addition, twenty-four haplotypes carry .2 alleles at DYS627, and 22 of these are also predicted to belong to hg J1. The allele 23.2, present in a single individual of 'unpredicted' haplogroup, is not yet catalogued in YHRD (Release 54, June 2017). One copy of the duplicated STR DYF387S1,S2 carries a .2 allele in 14 haplotypes, and of these haplotypes, ten are predicted to belong to the generally rare hg B, and are therefore likely identical by descent - these have been observed previously in the same haplogroup (Iacovacci et al. 2017). Sporadic examples of other intermediate alleles are observed at DYS390, DYS392, DYS448, DYS449 (including allele 31.2, not listed in YHRD) and DYS570.

The dataset contains one allele duplication - a tri-allelic pattern at DYF387S1,S2 - and also one example of a deleted allele at DYS448 in a haplotype predicted to belong to haplogroup J1. DYS448 deletions have been described previously (Balaresque et al. 2008), though not (to the author's knowledge) in this haplogroup; deletions are recurrent, and driven by unequal recombination between flanking large repeats. Deletion is also observed in a single case for both copies of DYF387S1,S2; five such cases are included in the 18,921 haplotypes in the YHRD.

Turning to haplotypes, Table 3.1 lists diversity summary statistics for the whole dataset, and for the five geographical subdivisions. The 597 males carry 543 distinct

haplotypes, including 25 identical pairs, and two trios, providing a discrimination capacity of 95.3%. However, when the sub-set of 17 Y-STRs making up the Yfiler® haplotype is considered, this shows a much higher level of haplotype sharing in the dataset: one haplotype, for example, is represented 22 times, and the discrimination capacity is only 74.7%; this compares, for example, with discrimination capacities of 98.5% and 95.7% for Yfiler® Plus and Yfiler® respectively in a US Caucasian sample (Applied Biosystems 2016).

Considering the five geographical subdivisions, comparisons of discrimination capacity at the levels of Yfiler® Plus and Yfiler® reveal striking differences in diversity between regions (Table 3.1). Values for Yfiler® Plus range from ~94% to 100%, but the range of values for Yfiler® haplotypes is much broader; in particular, the Central sample shows a discrimination capacity of only 72%, while the corresponding value for the Eastern sample is over 95%. In general, this points to relatively low diversity in the Central and Northern samples, with relatively high diversity in the East and West.

Table 3.1: Diversity summary statistics for Y-STR haplotypes in the entire sample set, and by geographical subdivisions. pop: population; n: number of individuals; ht: haplotype; HMP: haplotype match probability; DC: discrimination capacity; C: central; E: eastern; N: northern; S: southern; W: western. Yfiler® comparison: lists statistics considering only those STRs included in the 17-STR Yfiler® kit.

		Yfiler® plus							Yfiler® comparison			
pop	n	No. unique hts	No. pair hts	No. trio hts	HMP	Haplotype Diversity	% unique hts	DC	HMP	Haplotype Diversity	% unique hts	DC
All	597	541	25	2	0.0018	0.9998	90.6	95.3	0.0048	0.9969	63.0	74.7
All-SA	503	454	23	1	0.0022	0.9998	90.3	95.0	0.0052	0.9967	61.4	74.0
All-UK	94	92	1		0.0109	0.9998	97.9	98.9	0.0129	0.9977	88.3	92.6
C	125	117	4		0.0085	0.9995	93.6	96.8	0.0245	0.9834	62.4	72.0
C-SA	107	101	3		0.0099	0.9995	94.4	97.2	0.0263	0.9829	59.8	71.0
C-UK	18	18			0.0556	1.0000	100.0	100.0	0.0617	0.9935	88.9	94.4
E	110	110			0.0091	1.0000	100.0	100.0	0.0099	0.9992	90.9	95.5
E-SA	93	93			0.0108	1.0000	100.0	100.0	0.0117	0.9991	91.4	95.7
E-UK	17	17			0.0588	1.0000	100.0	100.0	0.0657	0.9926	88.2	94.1
N	106	92	7		0.0107	0.9987	86.8	93.4	0.0176	0.9917	57.5	74.5
N-SA	96	84	6		0.0117	0.9987	87.5	93.8	0.0202	0.9901	55.2	72.9
N-UK	10	10			0.1000	1.0000	100.0	100.0	0.1000	1.0000	100.0	100.0
S	140	125	6	1	0.0081	0.9991	89.3	94.3	0.0106	0.9965	72.1	83.6
S-SA	114	99	6	1	0.0102	0.9986	86.8	93.0	0.0132	0.9955	67.5	81.6
S-UK	26	26			0.0385	1.0000	100.0	100.0	0.0414	0.9969	92.3	96.2
W	116	106	5		0.0094	0.9993	91.4	95.7	0.0113	0.9973	79.3	87.9
W-SA	93	85	4		0.0119	0.9988	91.4	94.6	0.0135	0.9972	77.4	88.2
W-UK	23	23			0.0435	1.0000	100.0	100.0	0.0435	1.0000	100.0	100.0

3.3.2 Saudi Y-chromosome diversity compared with nearby regions

The total Saudi Arabian population sample was compared with other samples from the Arabian Peninsula, using MDS based on Rst distances calculated from Y-STR haplotypes (Figure 3.2a). In the first dimension of the plot, Iraq and Qatar lie at the extremes, with a cluster of other populations between them. The Saudi Arabian sample from this study (KSA) lies midway between this cluster and Qatar, and close to a previously published Saudi Arabian sample (Alshamali et al. 2009). However, when the KSA sample is subdivided into its five geographically-defined subsamples, two of these (Northern and Central) overlap with each other and cluster with Qatar, whereas the Eastern and Western subsamples overlap with each other and show affiliation with the major cluster of populations in the middle of the plot (Jordan, Kuwait, Oman, UAE and Yemen). The Southern sample lies at a similar first-dimension position to these, but is shifted in the second dimension of the plot, suggesting a distinct haplotype distribution to that of the Eastern and Western samples.

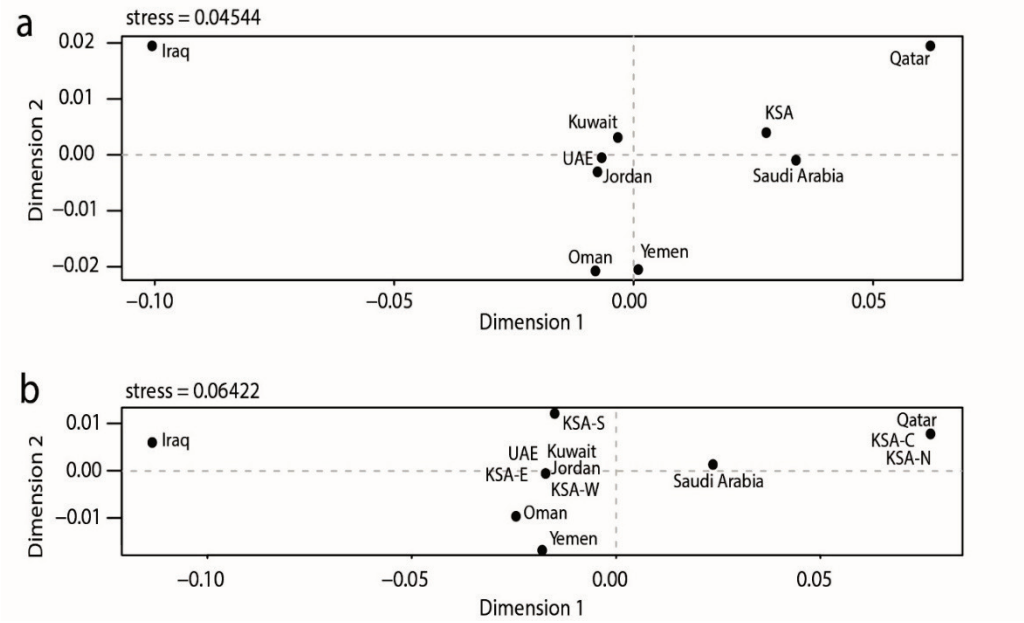


Figure 3.2: Multidimensional scaling (MDS) plots of Arabian Peninsula populations based on Y-STR haplotypes. Comparison with other datasets required reduction of the number of STRs to a shared set of nine. Comparison of the total dataset (KSA) with an independent dataset of 106 Y-STR haplotypes from the same country ('Saudi Arabia' (Alshamali et al. 2009)), and other datasets from the Arabian peninsula: Iraq (n=249; YHRD data), Jordan (n=254 (El-Sibai et al. 2009)), Kuwait (n=645 (Taqi et al. 2015; Triki-Fendri et al. 2010)), Oman (n=262 (Alshamali et al. 2009)), Qatar (n=46 (Cadenas et al. 2008)), UAE (n=684 (Alshamali et al. 2009; Nazir et al. 2016)), Yemen (n=375 (Alshamali et al. 2009), plus YHRD data). b) Comparison of the current dataset divided into five geographical sub-groups (KSA-N, -S, -E, -W, -C) with other Arabian Peninsula datasets as in part (a).

Inclusion of population samples (Hallenberg et al. 2005; Iacovacci et al. 2017; Manni et al. 2002; Nasidze et al. 2003; Piatek et al. 2012; Roewer et al. 2009) from a wider surrounding geographical region (Figure 3.3) does not change these relationships substantively.

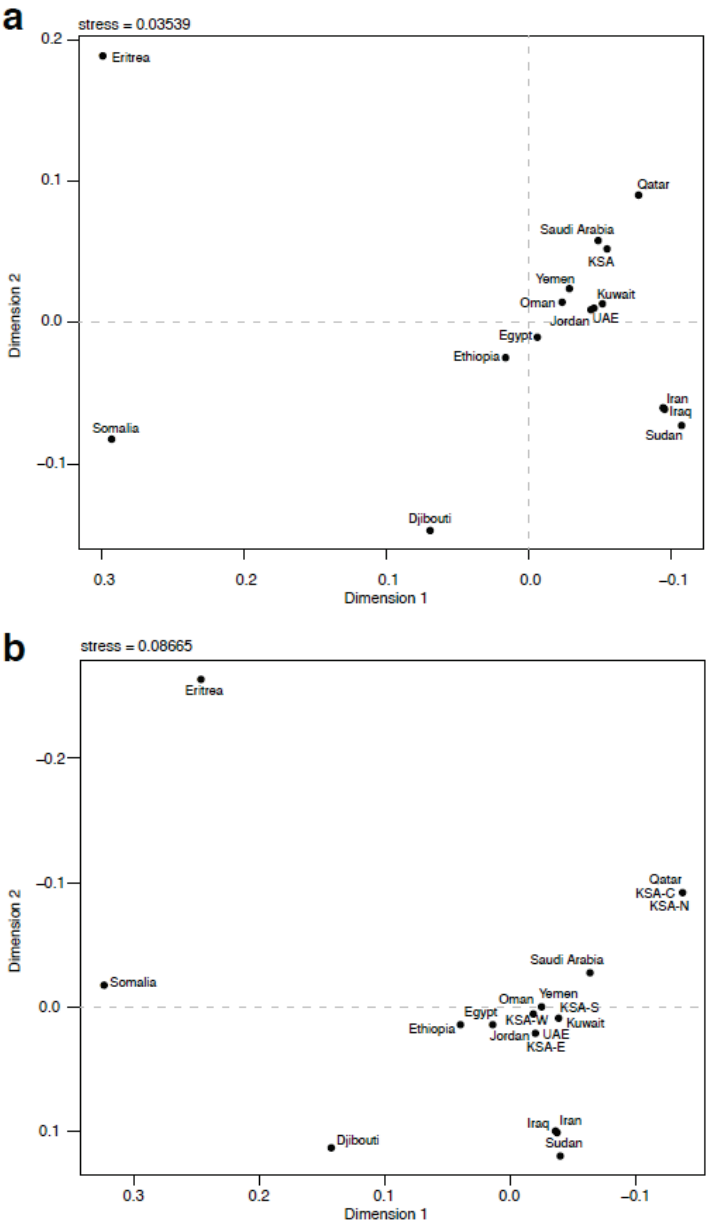


Figure 3.3: Multidimensional scaling (MDS) plots based on Y-STR haplotypes, including regional populations. Comparison with other datasets required reduction of the number of STRs to a shared set of nine.

a) Comparison of the total dataset (KSA) with an independent dataset of 106 haplotypes from the same country ('Saudi Arabia' - [1]), and other datasets from the Arabian Peninsula: Iraq (n=249; YHRD data), Jordan (n=254 [2]), Kuwait (n=645 [3, 4]), Oman (n=262 [1]), Qatar (n=46 [5]), UAE (n=684 [1, 6]), Yemen (n=375 [1], plus YHRD data), plus a geographically broad dataset from the Arabian Peninsula and additional surrounding nations (Djibouti (n=54; YHRD data), Egypt (n=289 [7]), Eritrea (n=161 [8]), Ethiopia (n=275 [8]), Iran (n=1887 [1, 9, 10]), Somalia (n=201 [11]), Sudan (n=64 [12]).

b) Comparison of the current dataset divided into five geographical sub-groups (KSA-N, -S, -E, -W, -C) with other datasets as in part (a).

3.3.3 Analysis of diversity via network analysis and haplogroup prediction

Before use on the current dataset, the accuracy of the haplogroup prediction method chosen in this study was assessed by using an independent dataset of Y-STR and Y-SNP haplotypes from 743 self-declared Saudi Arabian males downloaded from FamilyTreeDNA (www.familytreedna.com). The NevGen method requires input of a standard set of Y-STRs, but as yet this does not include the Yfiler® Plus set. The PowerPlex® Y23 (PPY23) set of 23 Y-STRs was therefore input with the two markers DYS549 and DYS643 marked as missing data, since these are present in PPY23 but not in Yfiler® Plus. Y-SNP resolution varied in the FamilyTreeDNA dataset, depending on whether custom Y-SNP typing or Y-chromosome resequencing had been carried out, and the level of haplogroup definition was standardised to a broad resolution of the thirteen haplogroups, A, B, E1b1b, E1b1a, G, H, J1, J2, L, Q, R1a, R1b and T. For the set of 743 FamilyTreeDNA Y chromosomes, NevGen predicted a compatible haplogroup in 738 cases (99.3%). The five incompatible predictions all involved mis-prediction of haplogroup E sublineages as either hg I2a2a, or D1a. In applying NevGen to the current dataset, all predictions were therefore accepted except those for hgs I (n=5) and D (n=1), and in addition predictions were rejected for haplogroups (hg N; n=1, and hg O; n=2) for which examples were not found in the FamilyTreeDNA dataset. This total set of nine haplotypes was defined as 'unpredicted'. To further understand the relationship between Y-STR haplotypes and predicted haplogroups, a median-joining network was constructed combining haplotypes both from our dataset and from the FamilyTreeDNA dataset Figure 3.4; this demonstrates coherence of haplogroup prediction and haplotype clustering.

In order to understand the relationships between Y-STR haplotypes in the dataset, a median-joining network was constructed (Figure 3.5a). Based on NevGen predictions, haplogroups were assigned to haplotypes within the network (Appendix, Table 9.1). Most predicted haplogroups form coherent clusters, with the exception of haplogroup E1b1b, which forms two well-separated clusters, likely indicating distinct sub-lineages that cannot be reliably distinguished by the prediction method used; the same split of haplogroup E1b1b is seen in a network containing the FamilyTreeDNA haplotypes of

known haplogroup (Figure 3.5a). Network sub-structures for most haplogroups are generally extended, although a cluster of related haplotypes exists among the predicted haplogroup E1b1b chromosomes. However, the network's major feature is a central star-like cluster of closely related haplotypes assigned to haplogroup J1 (71% of the total sample), suggesting a recent expansion for this set of lineages. The TMRCA of this cluster was estimated using the average-squared distance method and a mean pedigree-based STR mutation rate. Considering the total set of 23 non-duplicated Y-STRs, this yields a TMRCA of 2494±487 years; removal of rapidly-mutating markers, which might be expected to bias the estimate, reduces the number of STRs to 18 and increases the age slightly to 2754±389 years. Application of the same methods to the FamilyTreeDNA dataset of confirmed haplogroup J1 haplotypes yields very similar estimates, for example 2783±394 years for the reduced set of 18 Y-STRs. It is worth noting that use of the so-called 'evolutionary' average mutation rate of 6.9E-4 per STR per generation (Zhivotovsky et al. 2004) yields greatly elevated and very divergent TMRCA estimates for 23 and 18 Y-STRs in the current dataset of 19,835±3874 years and 9809±1916 years respectively.

Figure 3.6 shows the same network, but with haplotypes coloured by region of origin. There is little evidence from inspecting this network of geographical substructuring, although the haplogroup E1b1b sub-cluster mentioned above is mostly formed by Western samples. Table 3.2 and Figure 3.5b present predicted haplogroup distributions for the geographically defined sub-samples. One striking feature is the difference in the frequency of predicted haplogroup J1 in the Northern+Central+Southern samples (93% collectively) than that in the Eastern+Western pair (50%; exact test $p<0.001$); on the other hand, the latter pair has a significantly higher frequency of predicted haplogroup E1b1b (19% vs 6%; exact test $p<0.001$). Considering gene diversity values from haplogroup frequencies (Table 3.2), the Northern+Central pair shows significantly lower diversity than the Southern sample, which in turn is significantly lower than the Eastern+Western pair ($p<0.05$; 2σ).

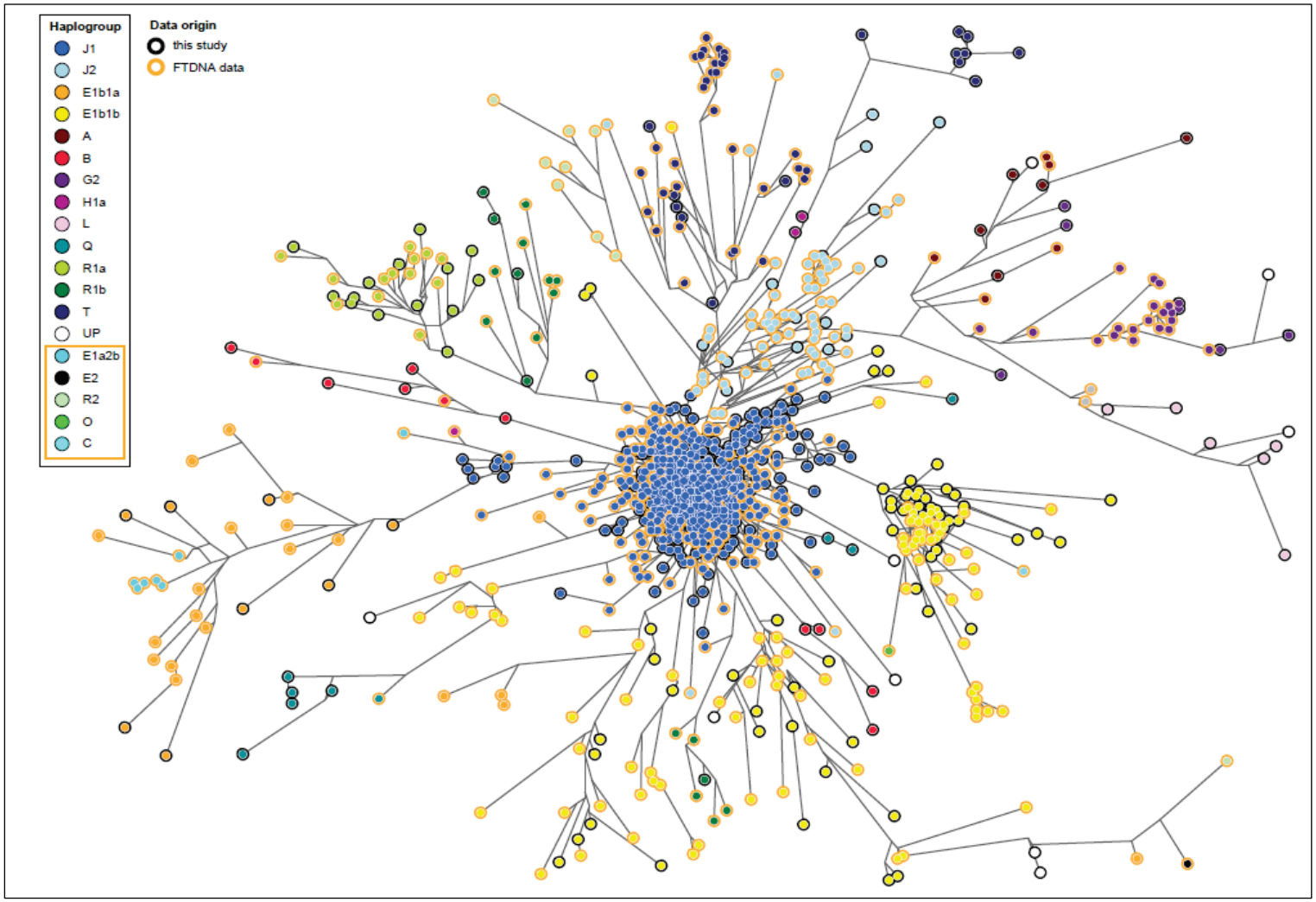


Figure 3.4: Median-joining network of Y-STR haplotypes, including FTDNA samples and samples described in this study. Median-joining network constructed from 23-Y-STR haplotypes in the 597 males of the current study, combined with 743 haplotypes from the Saudi Arabian FamilyTreeDNA (FTDNA) dataset. Circles represent haplotypes, with area proportional to sample size, and lines between them proportional to the number of mutational steps. Circle fill colours represent haplogroups, and circle edge colours indicate the two different sample origins, as shown in the key, top left. Within the key panel, the orange box highlights the five haplogroups found only in the FamilyTreeDNA dataset

.

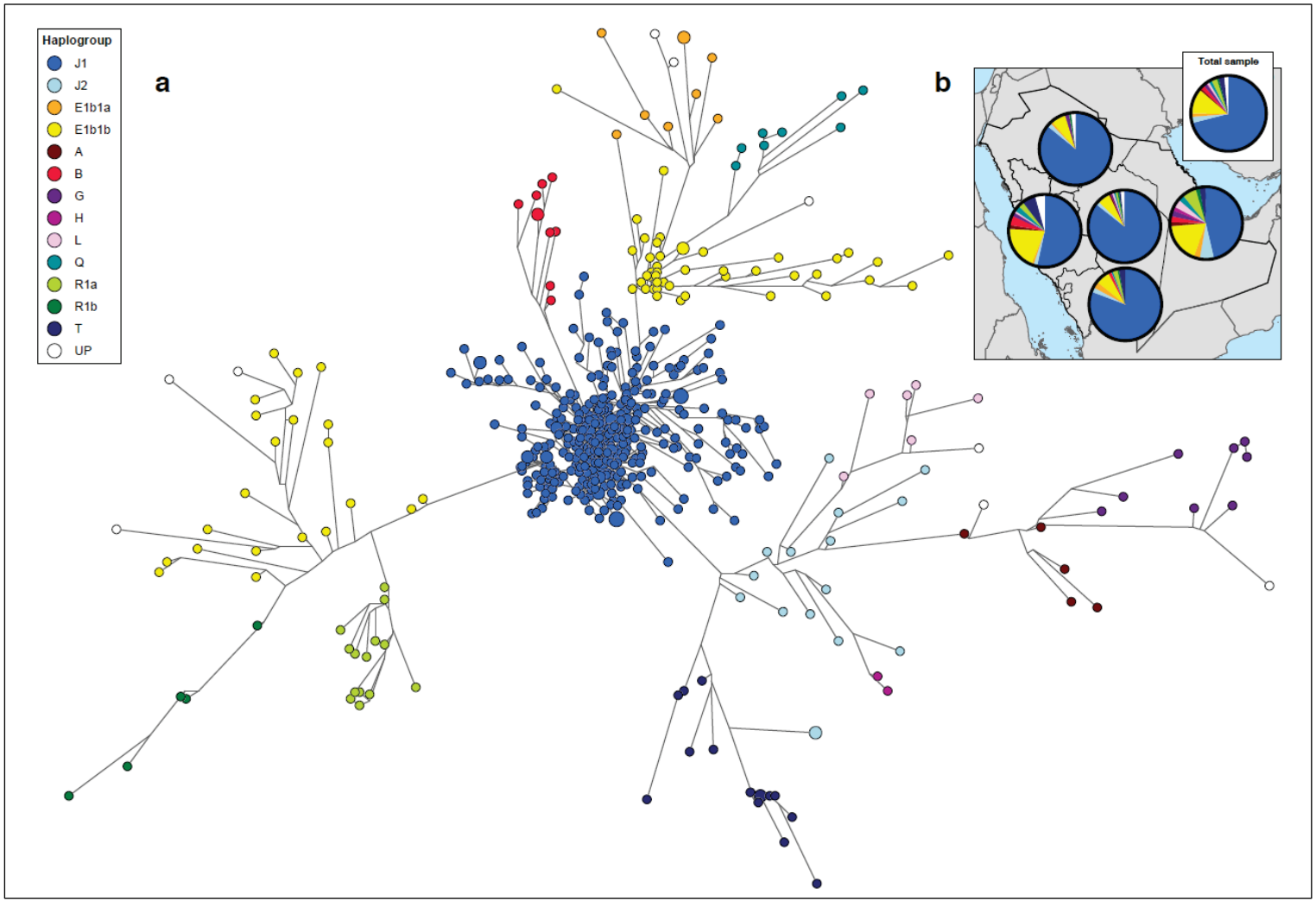


Figure 3.5: Median-joining network of Y-STR haplotypes, and geographical distribution of predicted haplogroups. a) Median-joining network for 597 Saudi Arabian haplotypes, constructed from data on 23 Y-STRs. Circles represent haplotypes, with area proportional to sample size, and lines between them proportional to the number of mutational steps. Colours represent haplogroups given in the key, top left. b) Map showing distributions of predicted haplogroups in five regional samples as pie-charts, not to scale; haplogroup distribution in the overall sample is represented in the pie-chart inset top right. Colours of sectors indicate haplogroups as shown in the key.

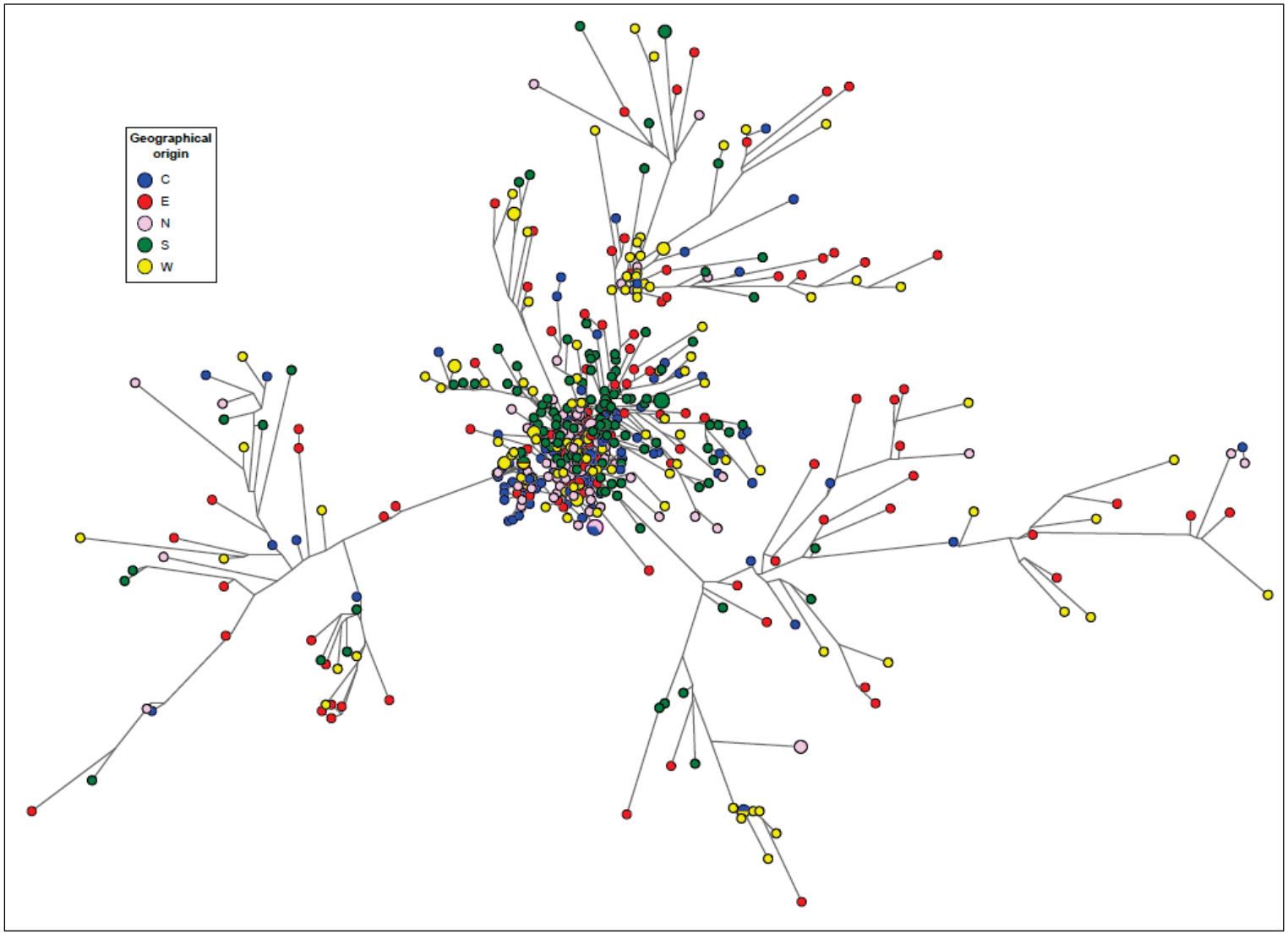


Figure 3.6: Median-joining network of Y-STR haplotypes, showing sub-regions of origin. Median-joining network for 597 Saudi Arabian haplotypes, constructed from data on 23 Y-STRs. Circles represent haplotypes, with area proportional to sample size, and lines between them proportional to the number of mutational steps. Colours represent the five different geographical sub-regions of origin, as given in the key, top left.

Table 3.2: Predicted haplogroup distributions and diversities in SA. This include the entire sample set considering geographical subdivisions and country of recruitment. pop: population; n: number of individuals; UP: unpredicted haplogroup; h: gene diversity; s.d.: standard deviation; C: central; E: eastern; N: northern; S: southern; W: western. SA: Saudi Arabia-recruited; UK: UK-recruited. a: Abu-Amero et al. (2009) (ref) b: Haplogroups are known from SNP typing for the 8 samples of Abu-Amero placed here in the 'UP' category, so not truly 'unpredicted'.

		Predicted haplogroup														
pop	n	A	B	E1b1a	E1b1b	G	H	J1	J2	L	Q	R1a	R1b	T	UP	h ± s.d.
All	597	5	10	9	66	8	2	424	16	6	8	14	5	15	9	0.481 ± 0.024
All-SA	503	4	10	6	48	5	2	374	9	5	7	10	5	13	5	0.436 ± 0.027
All-UK	94	1		3	18	3		50	7	1	1	4		2	4	0.676 ± 0.046
C	125	1			7	1		107	2	1	1	1	1	1	2	0.265 ± 0.052
C-SA	107	1			3	1		96	1	1	1		1	1	1	0.195 ± 0.052
C-UK	18				4			11	1			1			1	0.601 ± 0.113
E	110	2	3	3	20	3	2	51	7	4	3	7	2	3		0.745 ± 0.037
E-SA	93	2	3	2	15	2	2	43	5	4	3	7	2	3		0.752 ± 0.041
E-UK	17			1	5	1		8	2							0.713 ± 0.083
N	106			1	7	2		91	2				1		2	0.260 ± 0.056
N-SA	96			1	7	1		84	1				1		1	0.231 ± 0.056
N-UK	10					1		7	1						1	0.533 ± 0.180
S	140		2	4	9			113	3		1	3	1	4		0.344 ± 0.052
S-SA	114		2	3	7			97	1			1	1	2		0.273 ± 0.054
S-UK	26			1	2			16	2		1	2		2		0.618 ± 0.106
W	116	2	5	1	23	2		62	2	1	3	3		7	5	0.671 ± 0.041
W-SA	93	1	5		16	1		54	1		3	2		7	3	0.629 ± 0.050
W-UK	23	1		1	7	1		8	1			1			3	0.802 ± 0.060
Abu-Amero et al. (2009)[a]	157	0	3	12	12	5	3	63	25	3	4	8	3	8	8[b]	0.796 ± 0.026

3.3.4 Comparison of cohorts recruited in Saudi Arabia and the UK

Of the 597 collected samples, 94 (~16%) were recruited not in Saudi Arabia itself, but in the UK. To ask whether place of recruitment influenced the spectrum of haplotypes observed, Y-STR haplotype diversities were compared between the two differently recruited samples considering the same parameters as in Table 3.1. In the Saudi-recruited sample, both the proportion of unique haplotypes (~90%) and the discrimination capacity (~95%) are similar to the corresponding values in the cohort as a whole (~91% and ~95%). However, the UK-recruited sample shows much higher values - ~98% and 99% respectively, indicating that this mode of recruitment is sampling a more diverse subset of the Saudi Arabian population, despite the two groups being geographically matched. Predicted haplogroup frequencies in the Saudi- and UK-recruited samples (Figure 3.7a) are significantly different (exact test $p<0.001$), including a much higher frequency of predicted haplogroup J1 in the former. Similar comparisons at the sub-regional level (Figure 3.7a) show significant differences between predicted haplogroup frequencies (Table 3.2) for the Saudi- and UK-recruited samples from Central, Western, and Southern regions ($p\leq0.02$). For the North, the difference is not significant, probably due to the small sample size (n=10) of the UK-recruited sample. For the East, however, the corresponding sample size is larger (n=23), and the lack of significant difference ($p=0.795$) probably indicates true homogeneity of the Saudi- and UK-recruited samples for this region.

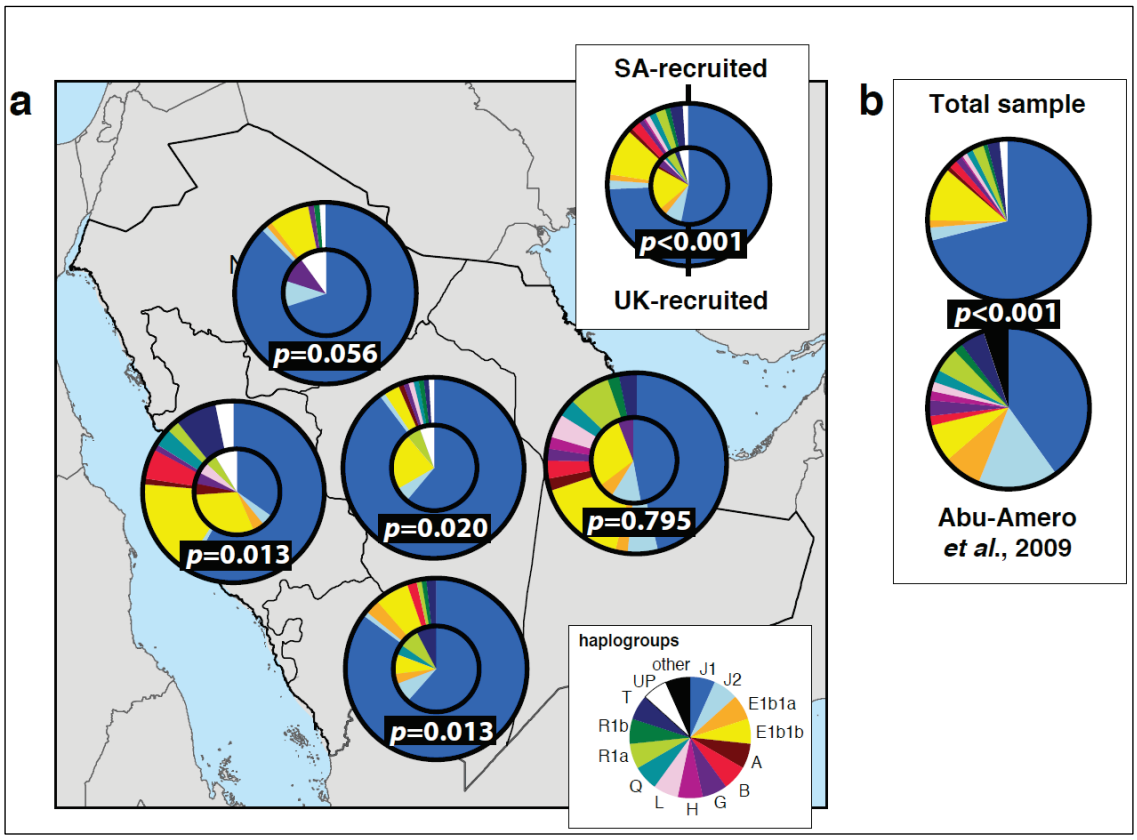


Figure 3.7: Comparison of Saudi- and UK-recruited cohorts by frequency of predicted haplogroups. a) Map showing distributions of predicted haplogroups in the five geographical regions, each divided into Saudi-recruited (outer pie-chart) and UK-recruited (inner pie-chart) samples. Pie-charts are not to scale. Haplogroup distributions in Saudi-recruited and UK-recruited samples for the total dataset are shown in the pie-chart inset top right. In each case, the p-value of a population differentiation test between Saudi- and UK-recruited samples is given. Colours of sectors indicate haplogroups as show in the key bottom right. b) Comparison of the haplogroup distribution in the total dataset (top) with that in published data (Abu-Amero et al. 2009) (bottom), with the p-value of the population differentiation test given.

3.4 Discussion

In this Chapter, the Yfiler® Plus haplotypes of a set of 597 Saudi Arabian males have been determined. Also, the effect on haplotype composition of dividing the sample into five geographically-defined sub-groups, and by two different countries of recruitment (Saudi Arabia itself, and the United Kingdom), has been considered.

The Yfiler® Plus system provides a discrimination capacity of 95.3% in the overall sample, which, while lower than that for US Caucasian, US Hispanic and African-American samples (Applied Biosystems 2016), exceeds that for an US Asian sample (94.4%). However, the added value of using the extended set of 27 Y-STRs contained in Yfiler® Plus is clearly demonstrated by a comparison with the 17 STRs defining

the Yfiler® system – discrimination capacity in the Saudi Arabian sample falls much more markedly than in the US samples, to only 74.7%. This suggests that, despite the care taken to avoid sampling related males, there are many individuals in the sample whose haplotypes are similar because of deeper patrilineal descent from shared ancestors. This probably reflects a general property of many Middle Eastern populations: of all global regions, the Middle East was previously shown to exhibit the greatest difference between diversity assessed by Yfiler® STRs and RM-YSTRs (Ballantyne et al. 2012), and the lowest Yfiler® discrimination capacity (~84%).

As well as determining Y-STR haplotypes, Y-SNP haplogroups were predicted from these haplotypes. A number of prediction methods exist, taking different approaches including phylogenetic trees with STR mutation rates (YPredictor; http://predictor.ydna.ru/), machine learning (Schlecht et al. 2008), partitional clustering (Seman et al. 2012), simple allele frequencies (Athey 2005), and Bayesian allele frequencies (Athey 2006); NevGen: http://www.nevgen.org). Evaluating prediction methods is not straightforward because some have been produced by the genetic genealogy community and are therefore not published in mainstream journals (Athey 2005, 2006) or peer reviewed (YPredictor; NevGen); exact methodology is sometimes unclear. There has been debate about the accuracy of prediction; for example, criticism (Muzzio et al. 2011) of one widely used method (Athey 2006), was counter-criticised (Athey 2011) for using only 7 Y-STRs in evaluation. It seems clear that the larger the number of STRs, the better, and here 21 STRs from the Yfiler® Plus set were used. In addition, any prediction method is only as good as the Y-SNP+Y-STR comparative datasets it uses for training or classification, and sometimes these are not well described – comparative datasets that are too small, or that do not include populations appropriate to the sample being predicted, may give unreliable results. In order to address this problem, a large set of independent Y-SNP+Y-STR data was used from the same population (Saudi Arabia) as that under study to test prediction performance. The chosen method, NevGen, performs well, and provides a >99% accuracy in the current sample; however, it should be recognized that SNP typing is required for definitive haplogroup determination (Gusmao et al. 2017).

The median-joining network of haplotypes (Figure 3.5a) exhibits a large central star-like cluster that corresponds to predicted haplogroup J1, and contains many of the identical or highly similar haplotypes. Such features are commonly interpreted as past male-lineage expansions (Batini and Jobling 2017). Star-like features of haplotypes comprising haplogroup J1 have been reported before in specific Arabian populations (Mohammad et al. 2009) and in broader Middle Eastern samples (Chiaroni et al. 2010). Interpretations of its origins initially focused on the 7th-century Muslim expansion (Nebel et al. 2002), and were supported by some later studies (Zalloua et al. 2008), but some other authors have interpreted it in terms of much earlier spread in the Neolithic (Chiaroni et al. 2010; Platt et al. 2017) followed by Bronze Age expansion in the Arabian Peninsula (Chiaroni et al. 2010). The age of the expansion is clearly crucial, and this in turn is affected by the method chosen, but most strongly by the choice of mutation rate. The mean 'evolutionary' rate of 6.9E-4 mutations per generation (Zhivotovsky et al. 2004) has been widely used (Abu-Amero et al. 2009; Chiaroni et al. 2010; Platt et al. 2017), and has been reported as performing better the directly-determined 'pedigree' rate for the dating of ancient events such as the coalescence of the whole Y phylogeny (Hallast et al. 2015; Wei et al. 2013). However, this rate was estimated from a relatively small set of 7-10 STRs (Zhivotovsky et al. 2004), not including RM-YSTRs, so certainly cannot be universally applied. Furthermore, for haplogroups showing star-like patterns in networks, and for which Y-chromosome resequencing data indicate recent TMRCAs (<10 thousand years), the pedigree mutation rate has been shown to perform best (Hallast et al. 2015). This rate was therefore used, and yielded a TMRCA for predicted haplogroup J1 of around 2800 years. This is several-fold younger than published estimates (Abu-Amero et al. 2009; Chiaroni et al. 2010; Platt et al. 2017), and if correct, suggests late Bronze Age dispersion, possibly followed by later spread during the Islamic expansion.

Aside from the dominant predicted haplogroup J1, the range of other predicted lineages (Table 3.2) is similar to that seen in a previous Y-SNP-based study of Saudi Arabia (Abu-Amero et al. 2009). However, the frequency of these haplogroups differs very significantly between the sample collected here and the published sample (Abu-Amero et al. 2009) (Figure 3.7b, Table 3.2). This suggests that there is

considerable Saudi Arabian heterogeneity, and that the sub-populations sampled in these two studies are very different. Such heterogeneity is confirmed when our sample is subdivided into five geographical regions. In both Y-STR- (Figure 3.2b, Table 3.1) and predicted-haplogroup-based (Figure 3.5b, Table 3.2) comparisons, the Central and Northern regions are highly similar in composition, and, surprisingly, the Eastern and Western samples are also highly similar. The Central and Northern pair is highly diverged from the Eastern and Western pair. The Southern region is somewhat distinct from all other regions with respect to both its spectrum of predicted haplogroups (Figure 3.5b, Table 3.2) and Y-STR haplotypes (Figure 3.2b). Considering discrimination capacity at the level of Yfiler® (Table 3.1), the Central and Northern samples show similarly low values, with higher and increasing values in the Southern, Western and Eastern samples respectively. Similar results are shown by comparing predicted haplogroup distributions (Table 3.2). The low diversity and similarity of Central and Northern areas reflect their relative geographical isolation within the desert heartland of the country, and possible bottleneck associated with the onset of desertification around 3000 years ago (Groucutt and Petraglia 2012). By contrast, the relatively high diversity of the Eastern and Western areas reflects their closeness to the sea and outside influences from other populations that may historically have brought in migrants.

While most of the 597 participants were contacted and recruited within Saudi Arabia itself, about 16% were recruited in the UK. Despite attempting to match the two groups by geographical sub-region, there are significant differences in the haplotype and predicted haplogroup constitutions of these two groups overall (Figure 3.7a, Table 3.1, Table 3.2), and for the Central, Southern and Western sub-groups in particular. The UK-recruited sample size for the Northern region is very small so it is hard to draw any conclusion, while for the Eastern region the two differently-recruited sub-samples seem genuinely similar in composition. Taken together, this suggests that there is social structure in the country, which influences the probability of males leaving Saudi Arabia and undertaking study in the UK, and that this structure correlates with different sub-groups having different haplotype compositions. This social structuring appears to be less marked in the East of the country than elsewhere. It means that the UK-recruited sample is an inappropriate

proxy for Saudi Arabia generally, and indicates that caution is needed when considering expatriate groups as representative of their country of origin.

The strong geographical and social structure observed in Saudi Arabia has important implications for the interpretation of Y-STR profiles in casework. Geographically appropriate databases must be used for assessment of evidential weight, and more work should be done to understand the social structuring reflected in the two differently recruited cohorts. Given the patrilineal descent structures of the tribal system, analysis of tribal names and surnames together with Y haplotypes should be illuminating. Other tribally-based Middle Eastern countries may also show marked population structure.

It will also be important to ask whether autosomal STR diversity is affected by the same factors that give rise to low diversity and a high degree of population structure among Y-haplotypes in Saudi Arabia. Sex-bias in population structure also be addressed using maternally-inherited mtDNA; inspection of published data in which the country is subdivided (Abu-Amero et al. 2008) shows that the mtDNA haplogroup spectrum of the Central region differs significantly from other regions except the Northern ($p \leq 0.05$, Bonferroni-corrected; our analysis). This indication of structuring among maternal lineages, as well as paternal lineages, suggests that further analysis in the same samples would be worthwhile.

Chapter 4 Analysis of 21 autosomal STRs in a Saudi Arabian sample

4.1 Introduction

Several previous studies have investigated autosomal allele frequencies in Saudi Arabia, initially covering just eight STRs among 207 individuals (Sinha et al. 1999), then 190 individuals from the Riyadh region analysed with the 15-STR Identifiler® multiplex (Osman et al. 2015), and 500 individuals from six cities spread across the five regions with the 21 STRs of the GlobalFiler® system (Alenizi et al. 2013; Alsafiah et al. 2017; Osman et al. 2015; Sinha et al. 1999). Further studies examined variation at 13 (Profiler Plus) and 15 STRs (Identifiler®) in Saudis resident in the bordering countries of Dubai (Alshamali et al. 2005) and Kuwait (Alenizi et al. 2013) respectively. Whilst these studies determined autosomal STR allele frequencies in Saudi nationals, they did not explore population structure within KSA. Analysis of the sample-set analysed here, which includes approximately equal representation of the five regions of the Kingdom, has in the previous Chapter revealed striking population structure in Y-chromosome variation, with greater diversity in the East and West and relative homogeneity within the North to South central axis of the country (Khubrani et al. 2018). The aim of this Chapter is to ask whether population structure can also be detected with autosomal markers, and if so whether it presents a similar geographical pattern to that of the patrilineal variation.

4.2 Materials and Methods

4.2.1 DNA sampling, extraction and quantification

Samples were collected and DNAs extracted and analysed as described in Chapter 3.

4.2.2 DNA amplification and fragment detection

The GlobalFiler® PCR amplification kit was used to generate profiles based on 21 autosomal STRs: D3S1358, vWA, D16S539, CSF1PO, TPOX, D8S1179, D21S11, D18S51, D2S441, D19S433, TH01, FGA, D22S1045, D5S818, D13S317, D7S820,

SE33, D10S1248, D1S1656, D12S391 and D2S1338, along with three additional markers used in sex determination (DYS391, Y indel and Amelogenin). Amplification was performed on a Veriti PCR machine (ThermoFisher Scientific) and fragment detection on an ABI3500 Genetic Analyzer (ThermoFisher Scientific) in accordance with the manufacturer's recommended protocols. GeneMapper® ID-X software v1.4 was used for allele calling and interpretation.

This work followed the guidelines of *FSI:Genetics* for publication of population genetic data (Carracedo et al. 2013; Carracedo et al. 2010; Gusmao et al. 2017) and for allele nomenclature (Schneider 2007). The dataset has been QC checked via STRidER (Bodner et al. 2016), with the dataset reference STR000119.

4.2.3 Forensic and statistical analysis

PowerStats v1.2 software (Promega Corporation, Madison, WI, USA) (Tereba 1999) was used to calculate allele frequencies, Random Match Probability (PM), Power of Discrimination (PD), Power of Exclusion (PE), Typical Paternity Index (TPI), observed homozygosity and observed heterozygosity. Arlequin v 3.5 (Excoffier and Lischer 2010) was used to test Hardy-Weinberg equilibrium, calculate expected heterozygosity, perform AMOVA to investigate genetic diversity within and between the five geographical regions, and undertake population differentiation tests for KSA and neighbouring countries. Additional measures of diversity (F_{IS} and F_{IT}) were calculated using FSTAT (Goudet 1995) for the five regions. Average pairwise F_{ST} values for 13 loci (CSF1PO, D13S317, D16S539, D18S51, D21S11, D3S1358, D5S818, D7S820, D8S1179, FGA, TH01, TPOX, vWA) shared with other studies of Saudi Arabians and neighbouring populations were used to generate multidimensional scaling (MDS) plots using the (MASS) package (Venables and Ripley 2002) in the R library.

4.3 Results

4.3.1 Data description and forensic statistics

The 21 autosomal STRs targeted by the GlobalFiler® kit were amplified from 523 Saudi Arabian males. Table 4.1 presents allele frequency data and forensic statistics for the whole KSA dataset; these measures are also provided for each of the five geographical subdivisions (Figure 4.1a) in Appendix (Table 9.2)

The least variable loci are TPOX and D22S1045, each with seven allelic variants in the KSA dataset, and the most variable locus was SE33 with 42 alleles. These respective loci had the lowest and highest Probabilities of Discrimination (0.820, 0.837 and 0.993 respectively), and conversely the highest and lowest major allele frequencies: TPOX allele 8 had a frequency of 0.552 in the total KSA dataset, being most frequent in the North (0.596) and rarest in the Central region (0.517), whilst SE33 allele 18, the most common variant at that locus, was detected at a frequency of just 0.107 in the KSA dataset overall. The combined power of discrimination (PD) and the combined power of exclusion (PE) for all loci in the KSA data are 0.99999999999999999999999999705 (1-(2.95E-26)) and 0.9999999563, respectively.

Table 4.1: KSA allele frequencies and forensic statistics MP: Random Match Probability, PE: Power of Exclusion, O-Het: observed heterozygosity, E-Het: expected heterozygosity, HWE p-value: probability of deviation from Hardy-Weinberg equilibrium, F_{IS} p- value: significance of F_{IS}.

Allele	D3S1358	vWA	D16S539	CSF1PO	TPOX	D8S1179	D21S11	D18S51	D2S441	D19S433	TH01	FGA	D22S1045	D5S818	D13S317	D7S820	SE33	D10S1248	D1S1656	D12S391	D2S1338
4											0.001										
6					0.002						0.339										
6.3																	0.005				
7				0.001	0.001						0.162				0.001	0.012					
7.3																	0.005				
8			0.031	0.005	0.552	0.002					0.110			0.011	0.105	0.165			0.001		
9			0.113	0.011	0.163	0.007			0.011		0.282			0.076	0.033	0.111	0.004	0.018			
9.3											0.096										
10			0.128	0.304	0.113	0.038		0.001	0.124	0.001	0.010		0.003	0.123	0.062	0.361	0.001		0.002		
11			0.359	0.294	0.157	0.137		0.021	0.358	0.011	0.001		0.107	0.262	0.255	0.183	0.001	0.009	0.042		
11.2																	0.002				
11.3									0.077												
12			0.212	0.345	0.012	0.188		0.144	0.080	0.095			0.011	0.346	0.341	0.146	0.006	0.037	0.131		
12.2										0.009											
13	0.007	0.002	0.137	0.033		0.185		0.210	0.010	0.168				0.171	0.147	0.020	0.015	0.167	0.099		
13.2										0.046							0.001				
13.3			0.001						0.001								0.001				
14	0.060	0.029	0.017	0.006		0.192		0.122	0.300	0.209			0.058	0.011	0.054	0.001	0.052	0.362	0.125	0.002	0.004
14.2										0.048											
15	0.239	0.157	0.002			0.201		0.119	0.033	0.141			0.508	0.001	0.001		0.030	0.257	0.136	0.025	
15.2	0.001							0.003		0.126											
15.3																			0.033		
16	0.314	0.282				0.043		0.118	0.006	0.074			0.258				0.058	0.105	0.251	0.011	0.077
16.2	0.001							0.006		0.053											
16.3																			0.054		
17	0.267	0.272				0.008		0.094		0.007			0.054				0.075	0.043	0.074	0.110	0.225
17.2								0.002		0.012											
17.3																			0.022		
18	0.105	0.179						0.073				0.007					0.107	0.001	0.003	0.149	0.098
18.3																			0.020	0.004	
19	0.007	0.066						0.041				0.064					0.064		0.001	0.133	0.127
19.3																			0.006	0.005	
20		0.012						0.026				0.070					0.044			0.100	0.239
20.2																	0.007				
21		0.001						0.008				0.139					0.018			0.089	0.047
21.2												0.001					0.076				
22								0.009				0.122					0.002			0.123	0.014
22.2												0.002					0.018				
22.3												0.002									
23								0.004				0.185								0.140	0.068
23.2												0.002					0.020				
24								0.001				0.226					0.001			0.057	0.064

Allele	D3S1358	vWA	D16S539	CSF1PO	TPOX	D8S1179	D21S11	D18S51	D2S441	D19S433	TH01	FGA	D22S1045	D5S818	D13S317	D7S820	SE33	D10S1248	D1S1656	D12S391	D2S1338
24.2																	0.028				
25												0.112								0.041	0.030
25.2																	0.027				
26												0.052								0.007	0.006
26.2																	0.050				
27							0.011					0.008								0.004	0.001
27.2																	0.050				
28							0.131					0.006									
28.2																	0.042				
29							0.270					0.003									
29.2																	0.036				
30							0.256					0.001									
30.2							0.011										0.043				
31							0.046										0.001				
31.2							0.103					0.001					0.049				
32							0.007														
32.2							0.112										0.038				
33																	0.005				
33.2							0.039										0.007				
34																	0.005				
34.2							0.007										0.005				
35							0.005										0.002				
35.2							0.001										0.001				
36							0.001														
36.2																	0.001				
37							0.001										0.001				

	D3S1358	vWA	D16S539	CSF1PO	TPOX	D8S1179	D21S11	D18S51	D2S441	D19S433	TH01	FGA	D22S1045	D5S818	D13S317	D7S820	SE33	D10S1248	D1S1656	D12S391	D2S1338
MP	0.097	0.078	0.081	0.147	0.180	0.052	0.056	0.029	0.100	0.030	0.097	0.038	0.163	0.092	0.081	0.081	0.007	0.096	0.032	0.024	0.040
PE	0.474	0.471	0.490	0.385	0.287	0.644	0.556	0.685	0.443	0.656	0.486	0.659	0.291	0.461	0.452	0.490	0.832	0.566	0.634	0.659	0.570
O-Het.	0.729	0.727	0.738	0.671	0.597	0.824	0.778	0.845	0.709	0.830	0.736	0.832	0.600	0.721	0.715	0.738	0.918	0.782	0.818	0.832	0.784
E-Het	0.759	0.785	0.777	0.701	0.632	0.832	0.818	0.876	0.754	0.871	0.759	0.857	0.658	0.762	0.779	0.775	0.948	0.761	0.864	0.890	0.850
P-value	0.264	**<0.0001**	0.019	0.141	0.444	0.944	0.637	0.226	0.004	0.874	0.092	0.130	0.029	0.158	0.020	0.374	0.053	0.364	0.559	0.068	**<0.0001**
FIS	0.037	0.072	0.050	0.043	0.058	0.010	0.047	0.032	0.064	0.045	0.027	0.027	0.086	0.049	0.078	0.049	0.032	-0.036	0.055	0.065	0.076
P-value	0.072	**0.001**	0.014	0.070	0.020	0.321	0.011	0.029	**0.005**	**0.005**	0.135	0.067	**0.001**	0.022	**0.001**	0.020	**0.002**	0.952	**0.001**	**<0.0001**	**<0.0001**
Alleles	9	9	9	8	7	10	15	18	10	14	8	18	7	8	9	8	42	9	16	16	14

4.3.2 Rare variants, off-ladder and null alleles

Thirty-five alleles were each observed only once in the entire dataset, and have been designated "rare" among Saudis. Of these, 13 were also globally uncommon, and were among the 26 off-ladder alleles recorded at these loci: D3S1358 (15.2 [N_{obs}=1] and 16.2 [1]), D16S539 (13.3 [1]), D18S51 (15.2, [3], 16.2 [6], 17.2 [2]), D2S441 (13.3 [1]), FGA (21.2 [1], 22.2 [2], 22.3 [2], 23.2 [2]), SE33 (7.3 [5], 10 [1], 11.2 [2], 13.2 [1], 13.3 [1], 22 [2], 24 [1], 31 [1], 33 [5], 34 [5], 36.2 [1]), D1S1656 (8 [1], 18 [3], 19 [1]) and D12S391 (18.3 [4]). All 26 off-ladder alleles have been described previously in STRBase (http://strbase.nist.gov/index.htm) (Ruitberg et al. 2001). No peak was detected at D2S1338 in one individual, despite good signal strength at all other loci, and despite retyping the sample twice. This individual appears to be a null homozygote at this locus, and he was therefore excluded from further analyses, as recommended (Bodner et al. 2016).

4.3.3 Genetic Structure

All but one locus showed a deficiency of heterozygotes against expectation in the whole KSA dataset, and this was also apparent within all five regions for between 16 and 20 loci (Table 2 for KSA, and Appendix, Table 9.3 for each region).

Table 4.2: Pairwise F_{ST} between regional sub-populations below the diagonal and p value above the diagonal.

Region	C	E	N	S	W
C		0.0322	0.0020	<0.0001	0.5664
E	0.0017		0.1006	0.1084	0.8291
N	0.0028	0.0013		<0.0001	0.0010
S	0.0041	0.0012	0.0053		0.0166
W	0.0001	0	0.0040	0.0021	

The deviation from Hardy-Weinberg equilibrium was significant following Bonferroni correction ($p \leq 0.00001$) for D2S1338 and vWA. Heterozygote deficiency was also evident from AMOVA analysis, with an inbreeding coefficient (F_{IS}) of 0.0476

representing 4.71% of variation for the KSA dataset, whilst F_{ST} was 0.0021 (F_{IS} values for the regions are shown in Appendix, Table 9.3). F_{ST} values between the five regions (Table 4.2) show that the greatest differentiation is between the North and South, with West, Central and East being less differentiated, and the East region being most similar to the other regions, as reflected in an MDS plot (Figure 4.1b).

Whilst the entire KSA dataset broadly clustered with the previously published Saudi datasets in MDS analysis (Figure 4.1c), division according to sub-regional origin showed that the North and South sub-regions were also the most differentiated from most other nearby populations (Figure 4.1d). The results of per-locus population differentiation tests between the KSA and regional datasets, previously published Saudi datasets and neighbouring countries are presented in Appendix, Table 9.4.

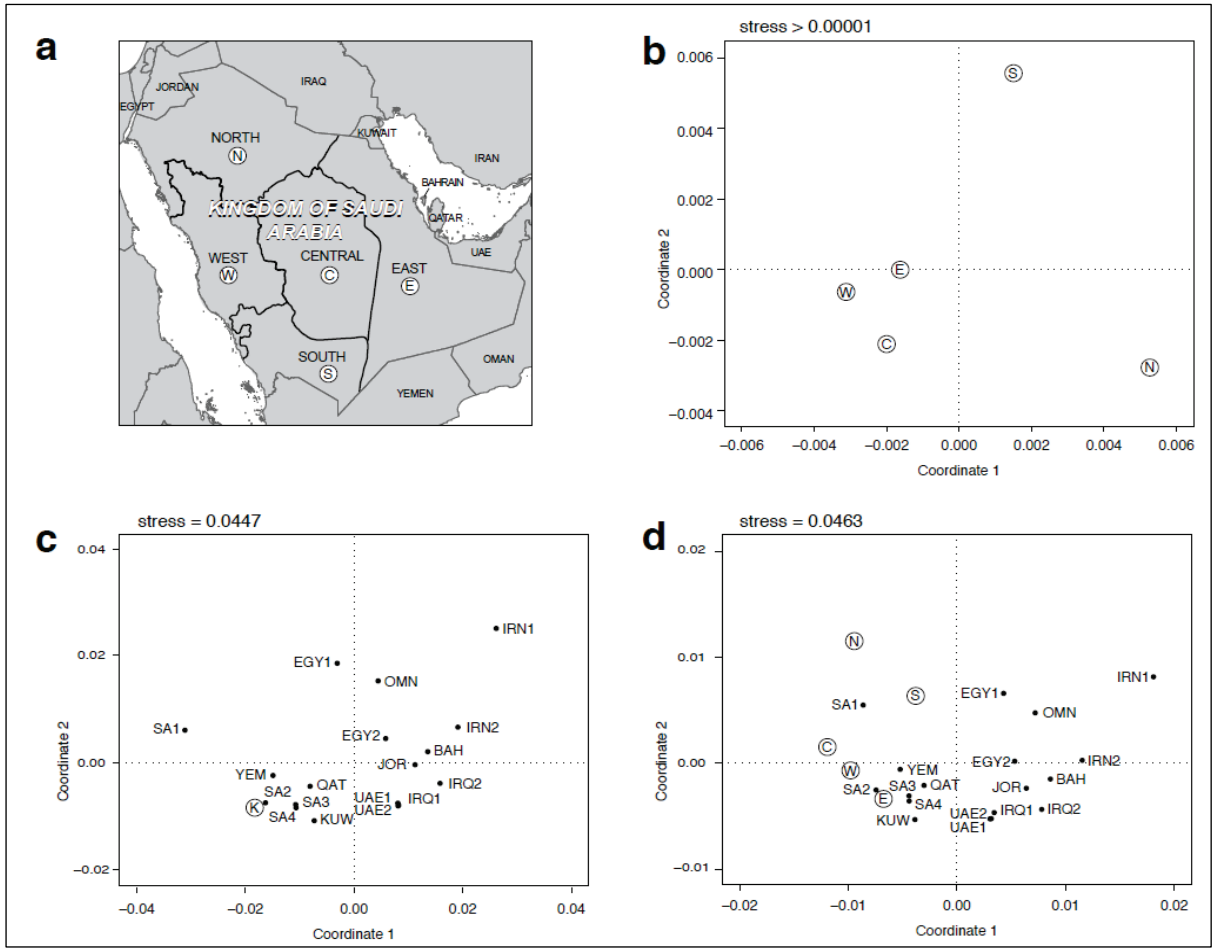


Figure 4.1: Map of sample locations, and multidimensional scaling (MDS) plots based on pairwise F_{ST} values derived from autosomal STR data a) Map of Saudi Arabia, showing location of the five geographical sub-regions; b) MDS plot comparing the five KSA sub-regions; c) MDS comparison of combined KSA dataset (encircled K) compared to previously published Saudi Arabian datasets and neighbouring countries; c) Five KSA sub-regions (encircled N - northern; S - southern; E - eastern; W - western; C- central) compared to previously published Saudi Arabian datasets and neighbouring countries. Comparative data sources and population abbreviations are as follows: BAH - Bahrain (Abuidrees et al. 2014), EGY1, EGY2 - Egypt, respectively (Alenizi et al. 2013), (Omran et al. 2009); IRN1, IRN2 - Iran, respectively (Alenizi et al. 2013), (Hedjazi et al. 2013) ; IRQ1, IRQ2 - Iraq, respectively (Alenizi et al. 2013),(Farhan et al. 2016); JOR - Jordan (Al-Eitan and Tubaishat 2018); KUW - Kuwait (Alenizi et al. 2013); OMN - Oman (Alshamali et al. 2005); QAT - Qatar (Pérez-Miranda et al. 2006); SA1, SA2, SA3, SA4 - Saudi Arabia, respectively (Osman et al. 2015), (Alsafiah et al. 2017), (Alshamali et al. 2005), ((Alenizi et al. 2013); UAE1, UAE2 - United Arab Emirates, respectively (Jones et al. 2017) and (Ali Alhmoudi et al. 2015); YEM - Yemen (Alshamali et al. 2005).

4.4 Discussion

In this analysis of diversity at the 21 autosomal STR loci of the GlobalFiler® multiplex no previously unreported alleles were found among 523 indigenous Saudi Arabian individuals and no evidence was obtained of genetic differentiation between the combined KSA dataset and previously published Saudi autosomal datasets, through exact tests of allele frequency. In common with previous studies (Alenizi et al. 2013; Alsafiah et al. 2017; Osman et al. 2015; Sinha et al. 1999), a tendency towards heterozygote deficiency was observed affecting almost all loci, although only two retained significance following Bonferroni correction. This observation, and the elevated F_{IS} values, are likely to reflect historical marriage practices in KSA in which partners are usually from within the same tribal group, and rates of first-cousin marriages are high, at around 30% (Al-Gazali et al. 2006). One of the two loci that showed significant deviation from Hardy-Weinberg equilibrium due to heterozygote deficiency (D2S1338) also showed evidence of the presence of a null allele, identified through a null homozygote. It seems reasonable to assume that the combined effect of consanguinity (Al-Gazali et al. 2006) and presence of null alleles in the heterozygous state produced the apparent excess of CE length "homozygotes" ($p= 0.00001$). It is unclear whether both factors, or only inbreeding, contributed to the significant but slightly weaker distortion observed at vWA ($p= 0.00006$).

Interestingly, a null has previously been reported at D2S1338 with the Identifiler® multiplex, which could be weakly detected as a visible allele with the NGM kit (Westen et al. 2014). Identifiler®, NGM and GlobalFiler® are all ThermoFisher multiplexes and share exactly the same primers at this locus (Matt Phipps, ThermoFisher personal communication). The same SNP (rs567937457) which lies 174 bp downstream of the repeat array was also identified by NIST as causing discordance between Identifiler® and Promega PP18 kits (http://strbase.nist.gov/pub_pres/NIST-Update-EDNAP-Apr2011.pdf). It has been suggested that the discrepancy between Identifiler® and NGM is related to the longer annealing/extension time (3 min vs. 1 min) in the latter, which may permit amplification from the poorly matched primer (Westen et al. 2014); however, in the

homozygous null individual described here, extending the annealing/extension time to 3 minutes did not yield a detectable peak. Subsequent amplification with the MiniFiler kit which has a 183 bp shorter amplicon also yielded no result. The positions of the MiniFiler primers are proprietary information but can encompass at most 60 bp of flanking region which must exclude the aforementioned SNP as the cause, and would imply either a substantial deletion affecting both downstream primers or a polymorphism upstream of the repeat, where the primer is already close to the repeat array. Unfortunately, it was not possible to explore the nature of this polymorphism any further due to the limitations of the DNA donor consent.

This Chapter represents the first study to specifically address the question of sub-structure within the indigenous Saudi Arabian population using autosomal STR markers. The studied population sample, which has approximately equal representation of the five geographic regions of the Kingdom, was shown in the previous Chapter to display striking differentiation in Y-chromosomal haplogroup (Y-SNP) and haplotype (Y-STR, Yfiler® Plus) distributions (Khubrani et al. 2018), suggesting that the country is genetically substructured at least with respect to male-specific markers.

A different picture emerged when the genetic diversity of the geographic sub-regions was explored with GlobalFiler®. Significantly different pairwise F_{ST} values were noted between the South, North and Central/West regions, with the East appearing intermediate, and not significantly differentiated from any other region (Figure 4.1b). The combined KSA dataset is virtually indistinguishable from sets of Saudi donors sampled in Dubai and Kuwait (Alenizi et al. 2013; Alshamali et al. 2005) and the previous multi-city GlobalFiler® survey within Saudi Arabia (Alsafiah et al. 2017), but somewhat different from the Identifiler® dataset derived from bone marrow donors at a Riyadh hospital (Osman et al. 2015), which is an outlier (Figure 4.1c). However, after division of the current dataset into the five regions (Figure 4.1d) sample-sets from the previously published population surveys now cluster most closely with the Eastern region. Also, the North and South are the most differentiated from each other and from the other regions, with the exception of the bone marrow donor set (Osman et al. 2015), which shows similarity with the South.

Autosomal differentiation between North and South may be a consequence of historically limited migration between these regions. Migrants are now primarily attracted by the oil industry in the East, the holy cities of the West and the capital in the Central region (Al-Rasheed 2010). Prior to the establishment of the Kingdom of Saudi Arabia the interior was sparsely populated by sedentary farmers or nomads who moved within their tribal areas, reflected by some of the current administrative boundaries. The rapid growth of the Saudi population, stimulated by the discovery of oil, has involved movement into the cities from the surrounding areas, but the rural populations have not generally moved between regions. The regional patrilineal criterion for inclusion in the dataset (Khubrani et al. 2018) meant that the results of Y-chromosome tests reflect the historic boundaries of tribal groups, and these are strongest and most stable in the North and Central regions. By contrast, the East and West received the greatest inward migration in recent centuries, as they were outward facing and the destinations of traders and pilgrims (Al-Hathloul and Edadan 1993). The autosomal results also reflect the movement of women, who are likely to move to their husband's home. No restriction was applied on the origin of the maternal line, and as a consequence a different pattern reflecting more recent migration patterns is unsurprising. In conclusion, this Chapter demonstrates the high discrimination power of GlobalFiler® in the Saudi population, which makes it suitable for the purposes of forensic DNA identification and paternity testing; the allele frequencies derived here can be utilised by Saudi forensic providers for DNA interpretation purposes. While the observed allele frequency variation between regions will have limited influence on interpretation issues, the high level of consanguinity and presence of null alleles should be taken into account.

Chapter 5 Massively parallel sequencing of autosomal STRs and identity-informative SNPs in Saudi Arabia

5.1 Introduction

Traditionally, forensic analysis of short tandem repeat (STR) diversity has been performed via capillary electrophoresis (CE), which considers only amplicon length and overlooks potentially informative sequence variation. Additional polymorphisms could include different numbers of diverged repeat units, together with indels and SNPs in both repeat arrays themselves and flanking regions: such variation can be accessed through massively parallel sequencing (MPS). MPS approaches not only provide increased (sequence-level) resolution of individual loci, but can also simultaneously analyse many diverse loci in a single test, thus simplifying analysis of different marker types and making the best use of limited amounts of casework material (de Knijff 2019).

The ForenSeq™ DNA Signature Prep Kit from Verogen exemplifies the advantages of the MPS approach by simultaneously amplifying more than 150 loci including the standard autosomal and Y-STRs plus X-STRs and identity-informative SNPs (iiSNPs). Although the primary interface with the manufacturer's ForenSeq™ Universal Analysis Software (UAS) currently focuses on only a single iiSNP within each amplicon, additional variation exists in the flanking regions of many target SNPs (King et al. 2018), which could be useful in further increasing discriminating power and interpreting mixed stains. For several of the targeted SNPs the additional variant sites result in multiple alleles that will tend to be co-inherited as a microhaplotype, giving some iiSNP amplicons a resolving power approaching those of traditional simple-sequence STR loci. The DNA Signature Prep Kit has also improved resolution of both male/male and male/female mixtures involving minor contributors as low as 1:20 (Churchill et al. 2016; Jäger et al. 2017; Köcher et al. 2018; Xavier and Parson 2017) and this capability has proved useful in the first sexual assault court case using MPS in the Netherlands. The kit has also been implemented in casework by the INPS (Institut National de Police Scientifique)

laboratory in Lyon, with MPS profiles uploaded to the French national DNA database, and the FBI has approved the kit itself, the MiSeq FGx™ System and the UAS for the US National DNA Index System within the terms of newly published SWGDAM guidelines for MPS (SWGDM 2019).

The advantages conferred by MPS approaches could be particularly relevant to forensic science in the Middle East, where high average annual temperatures favour the use of forensic markers such as short-amplicon STRs and iiSNPs that provide robust discrimination from degraded DNA. In addition, populations in the region tend to exhibit endogamy and population structure (Al-Gazali et al. 2006), leading to clusters of individuals that share a common heritage and reducing the discrimination of forensic multiplexes relative to more diverse and exogamous societies. Previous Chapters have shown by conventional CE typing of both Y- (Khubrani et al. 2018) and autosomal STRs (Khubrani et al. 2019b) that the indigenous population of Saudi Arabia is highly structured. This is apparent between the five regions of the country (North, South, East, West and Central), which have different tribal compositions and historical exposures to immigration.

Here the DNA Signature Prep Kit is applied to a sample dataset of Saudi males currently residing in the UK, with the following aims:

(i) to determine whether this sample of males reflects both the genetic composition of their home nation and the expected high rates of consanguinity, which could impact on the observed homozygosity of the autosomal markers;

(ii) to ask whether analysing sequence variation of autosomal markers (aSTRs and iiSNPs) within the DNA Signature Prep Kit significantly enhances resolution;

(iii) to report novel variants within this population of the Middle East, a region from which little MPS data has so far been published (Phillips et al. 2018).

5.2 Materials and Methods

5.2.1 DNA sampling, extraction and quantification

Samples were collected from 89 indigenous Arab males residing in the United Kingdom as described in Chapter 3. DNA was extracted either from buccal swabs (King et al. 2006), or from saliva samples using the Oragene•DNA (OG-500) kit (DNA Genotek), and quantified as previously described (Khubrani et al. 2018).

5.2.2 Library preparation and sequencing

Sequencing libraries were prepared using the ForenSeq™ DNA Signature Prep Kit according to the manufacturer's recommendations (Verogen Inc., San Diego, CA, USA). Primer Mix A was used to amplify 58 STRs (27 autosomal STRs discussed in this Chapter, as well as 7 X-STRs and 24 Y-STRs which will be described in Chapter 6) and 94 identity-informative SNPs (iiSNPs) from 1 ng of template DNA. Steps for library preparation include target-specific amplification, target enrichment including incorporation of indexed adapters, purification, bead-normalisation and pooling, prior to sequencing on a Verogen MiSeq FGx™ device, all of which were performed in accordance with the manufacturer's recommended protocols. Either 96 or 32 libraries were analysed in each sequencing run, which included one positive and one negative control.

5.2.3 Calling of iiSNPs from HGDP sequence data

The iiSNPs were jointly called from the publicly available cram files of whole-genome sequenced (mean coverage 35 ×) Human Genome Diversity Project (HGDP) – CEPH (Centre d'Etude du Polymorphisme Humain) samples (Cann et al. 2002), mapped to GRCh38 using BCFtools (v1.8)(Li 2011) with minimum base quality 20 and mapping quality 20. This work was carried out by Dr Pille Hallast at the Wellcome Sanger Institute.

5.2.4 Data analysis

Sequence data were analysed using Verogen's default settings for the Analytical

Threshold (AT), Interpretation Threshold (IT), Stutter Filter and Intra-Locus Balance in the ForenSeq™ Universal Analysis Software (UAS). Any sequence detected above the analytical threshold of 10 reads is reported to the user for their consideration, while above the 30-read interpretation threshold the UAS automatically reports the presence of an allele if the overall read depth for the locus is <650. When ≥650 reads are collected for a locus, the AT and IT defaults are set at 1.5% and 4.5% of reads respectively. Between the AT and IT, the result is flagged by the UAS and the user can determine whether the sequence is a true variant. User interpretation is also required when the Intra-locus Balance (equivalent to heterozygote balance) falls below 60%, or the level of stutter exceeds the default Stutter Filter value, which varies between STR loci. As all of the samples studied here were good quality single-source reference DNAs, as demonstrated by earlier Yfiler® Plus profiling (Khubrani et al. 2018), the interpretation was relatively straightforward.

The UAS provides a visual interface which displays each STR locus for an individual as a histogram arranged according to conventional CE allele length, with isometric heterozygotes (alleles of the same length but different sequence) shown as stacked bars. For STRs, the visual UAS interface displays only repeat region sequence variants, and for each of the iiSNP amplicons, only the target SNP. However, it is also possible to view "Flanking Region Reports" that show the flanking regions of both aSTR and iiSNP amplicons (which for Penta E is limited to a maximum of 197 bp to ensure sequencing data integrity). These files highlight variation at some, but not all, additional polymorphic sites within the amplicons. STRait Razor v3.0 (Woerner et al. 2017) was used to check bioinformatic concordance of allele calls and to clarify appropriate nomenclature in line with ISFG considerations (Parson et al. 2016). In the following sections STR sequence variation is described at three levels: length - a measure solely of allele length for compatibility with conventional CE methods; repeat region sequence - sequence variation within the repeat array as reported by UAS; and repeat plus flanking sequence, including all polymorphisms within the reported region of the amplicon, as obtained from the additional Flanking Region Reports.

5.2.5 Population, forensic and statistical analysis

Arlequin v 3.5 (Excoffier and Lischer 2010) was used to test Hardy-Weinberg equilibrium, and to calculate expected heterozygosity and pairwise linkage disequilibrium (LD). It was also used for performing AMOVA to investigate genetic diversity, to calculate fixation indices (F_{IS}) and to undertake population differentiation tests. STRAF (Gouy and Zieger 2017) was used to calculate forensic statistics including: genotype count (N), allele count based on sequence (N_{all}), genetic diversity (GD), polymorphism information content (PIC), random match probability (PM), power of discrimination (PD), observed and expected heterozygosity (H_{obs} and H_{exp}), power of exclusion (PE), and typical paternity index (TPI). Allele frequencies were calculated in Excel (allele count/total).

5.3 Results

DNA profiles of 89 Saudi males generated with the ForenSeq™ DNA Signature Prep Kit Primer Mix A gave a mean total read count per sample of 34,821 reads for the complete complement of autosomal STR alleles, and 41,650 for the set of iiSNPs. Following visual checks of individual loci flagged by the default settings of the ForenSeq™ UAS, genotypes were called for all 27 autosomal STRs and 91 of the 94 iiSNPs. The three lowest-performing iiSNPs (rs1736442, rs2920816 and rs719366) were dropped from the analysis due to increased observations of sub-threshold calls (15, 9 and 5 respectively). The SNP rs1031825 was retained in the analysis but was below threshold in three individuals. In addition, two individuals had sub-threshold calls at the STR locus Penta E. Average read depths for the analysed autosomal STRs ranged from 3979 at TH01 to 285 at vWA, and for the iiSNPs from 1710 at rs1109037 to 64 at rs1031825.

5.3.1 Autosomal STR sequence variation and impact on discrimination

Among the 4804 STR alleles typed in the 89 individuals, there were 238 distinct length variants and 340 repeat sequence sub-variants identified by the UAS across the 27 loci (Appendix, Table 9.5; Figure 5.1a). The loci D17S1301 and D4S2408 presented the lowest diversity, with five alleles each; together with ten other STRs, these showed only length variation when visualised solely with the UAS, thus providing the same discrimination power as a conventional CE approach. The UAS highlighted additional sequence variation within the repeat regions of the remaining 15 loci, contributing between one extra variant (at D19S433, TH01 and CSF1PO) up to 25 additional alleles at D12S391 (Figure 5.1a). Examination of the Flanking Region Report revealed a further 17 distinct variants including one, three, four and six additional alleles at D22S1045, D20S482, D16S539 and D7S820 respectively, whereas the visual interface of the UAS displayed only length variation. Two loci, D1S1656 and D2S441, each showed a total of six extra alleles as a result of sequence variation both within and flanking the repeat region, with SNPs in the flanking region contributing one and two of these extra alleles respectively. The sequences and frequencies of all autosomal STR alleles are shown in Appendix,

Table 9.5. The number of additional alleles created by detection of sequence variation either within the repeat or flanking regions differs considerably between loci, but a more important parameter is how this affects the power of discrimination (PD). Figure 5.1b shows PD for each of the autosomal STRs, subdividing this into the proportions due to repeat number (length), repeat region sequence and flanking region variation.

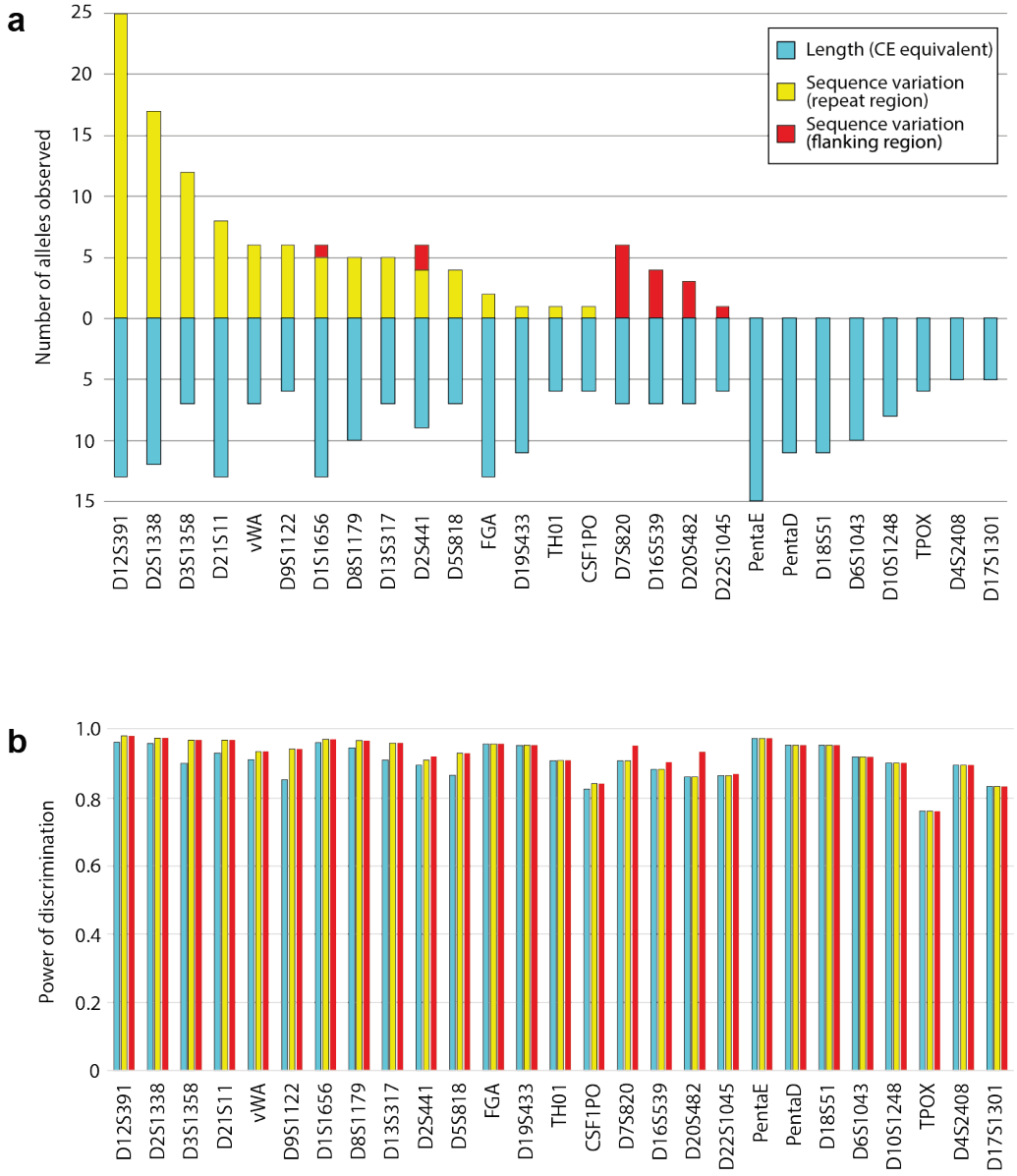


Figure 5.1: Counts of distinguishable alleles by STR locus, and per-locus increment of discriminatory power due to sequence variants.. a) The observed number of length variants among the 89 Saudi males is shown in blue below the x-axis, and the number of additional alleles resulting from sequence variation within and flanking the repeat array are shown above in yellow and red respectively. b) The power of discrimination resulting from length variation alone (equivalent to CE typing) is shown in blue and the additional contributions made by sequence variation within and flanking the repeat array are shown in yellow and red respectively.

Twenty-three ‘novel’ alleles, summarised in Table 5.1 (and more fully described in Appendix, Table 9.6), were absent from the STRait Razor v3.0 default allele list. However, fifteen had been reported previously either in ForenSeq™ DNA Signature Prep Kit-related literature (Phillips et al. 2018) or in the STRSeq database (Gettings et al. 2017) (accessed via GenBank, July 2019); of these, seven sequences were previously seen in the Middle East, with two of them not yet reported outside of the region. Of the 23 ‘novel’ alleles, two are rare short simple repeat-number variants; fifteen are compound STR alleles with novel numerical combinations of repeat blocks; three have single-base insertions within a repeat array producing intermediate alleles; and the remaining three have SNP variants in their flanking regions. Of the six SNP variants, only two have existing entries in dbSNP (rs563997442, rs554502154).

Table 5.1: Novel STR alleles found in this study. Alleles are missing from the STRait Razor v3.0 default database. The nature of its novelty (repeat length [RL], repeat region sequence [RS] or flanking sequence [FS]); and the geographical distribution of matching alleles in HGDP metapopulations: MEA: Middle East; AFR: sub-Saharan Africa; EUR: Europe: OCE: Oceania; SAS: South Asia; EAS: East Asia.

Nomenclature	Type	Observations	HGDP occurrence
CSF1PO [CE 12]-GRCh38-Chr5-150076318-150076389 ATCT ACCT (ATCT)10	RS	3	MEA
D10S1248 [CE 9]-GRCh38-Chr10 129294226-129294318 (GGAA)9	RL	2	MEA/AFR
D12S391 [CE 23]-GRCh38-Chr12-12296981-12297168 (AGAT)12 (AGAC)11	RS	1	EUR/OCE/SAS
D12S391 [CE 23]-GRCh38-Chr12-12296981-12297168 (AGAT)12 (AGAC)10 AGAT	RS	1	EUR
D12S391 [CE 26]-GRCh38-Chr12-12296981-12297168 (AGAT)17 (AGAC)8 AGAT	RS	1	EAS
D13S317 [CE 9]-GRCh38-Chr13-82147986-82148107 (TATC)8 AATC	RS	1	not observed
D16S539 [CE 8]-GRCh38-Chr16-86352664-86352781 (GATA)8 86352692-G (rs563997442)	FS	1	n/a
D1S1656 [CE 17]-GRCh38-Chr1-230769555-230769682 CCTA (TCTA)16 230769682-G (rs NA)	FS	1	n/a
D21S11 [CE 36]-GRCh38-Chr21-19181939-19182111 (TCTA)11 (TCTG)6 (TCTA)3 TA (TCTA)3 TCA (TCTA)2 TCCA TA (TCTA)11	RS	1	not observed
D22S1045 [CE 15]-GRCh38-Chr22-37140181-37140357 (ATT)12 ACT (ATT)2 37140182-A (rs554502154)	FS	1	n/a
D2S1338 [CE 15]-GRCh38-Chr2-218014856-218014964 (GGAA)9 (GGCA)6	RS	1	EUR
D2S441 [CE 9]-GRCh38-Chr2-68011918-68012017 (TCTA)9	RL	1	MEA/EUR/AFR/EAS
D3S1358 [CE 13]-GRCh38-Chr3-45540691-45540820 TCTA (TCTG)3 (TCTA)9	RS	1	not observed
D3S1358 [CE 17]-GRCh38-Chr3-45540691-45540820 TCTA (TCTG)2 TCTC (TCTA)13	RS	2	MEA/SAS
D3S1358 [CE 18]-GRCh38-Chr3-45540691-45540820 TCTA (TCTG)2 TCTC (TCTA)14	RS	1	not observed
D3S1358 [CE 18]-GRCh38-Chr3-45540691-45540820 TCTA (TCTG)4 (TCTA)13	RS	1	MEA/SAS
D3S1358 [CE 19]-GRCh38-Chr3-45540691-45540820 TCTA (TCTG)4 (TCTA)14	RS	1	MEA
D6S1043 [CE 25]-GRCh38-Chr6-91740160-91740292 (ATCT)6 ATGT (ATCT)4 ATGT (ATCT)13	RS	1	not observed
D8S1179 [CE 17.1]-GRCh38-Chr8-124894867-124894921 (TCTA)2 (TCTG)2 (TCTA)12 TCTTA	RS	1	not observed
D9S1122 [CE 13]-GRCh38-Chr9-77073809-77073880 TAGA TCGA (TAGA)8 CAGA (TAGA)2	RS	1	not observed
FGA [CE 16.1]-GRCh38-Chr4-154587713-154587823 (GGAA)2 GGAG (AAAG)3 A (AAAG)5 AGAA AAAA (GAAA)3 154587760-A	RS	1	not observed
PentaE [CE 16.4]-GRCh38-Chr15-96830996-96831114 (TCTTT)16 TCTT	RS	2	MEA/SAS
TH01 [CE 6]-GRCh38-Chr11-2171056-2171127 (AATG)3 AATA (AATG)2	RS	1	not observed

In terms of simple repeat number nomenclature for comparison with CE data, no unusual genotypes were observed. However, some loci displayed read counts at stutter positions above the stutter threshold that could represent potential somatic triallelic patterns. The allele counts that exceeded the stutter threshold by the greatest margin were at D8S1179, where three alleles were called by the UAS at 12,15,16 with read depths of 247, 86 and 295 respectively. The lowest of these contributions was at a stutter position and would equate to a stutter proportion of 0.29 of allele 16, whereas the next strongest stutter observed at this locus in this study was 0.26, only fractionally higher than the recommended UAS default permissible stutter ratio of 0.25 for the locus. Similar triallelic patterns generated by CE have been reported previously at this locus [https://strbase.nist.gov and (Mertens et al. 2009). While imbalanced "triallelic" patterns can result from somatic mutations it is also possible that they are simply the result of unusually prominent stutter. It is unclear which explanation is appropriate in these two instances as the profiles are otherwise of good quality and do not show excessive stutter at other loci, which might have been indicative of overamplification.

5.3.2 Sequence variation at autosomal iiSNPs and in their flanking regions

Fifty-four of the 91 iiSNP amplicons included in the analysis showed no additional variation within the regions covered by the Flanking Region Report, while 27 have a second polymorphic site, eight have a third and two amplicons show variation at four positions (see Appendix, Table 9.7). Combinations of alleles at two pairs and one trio of linked SNPs showed perfect associations (AT & TA at rs279844 & rs279845, AT & GC at rs6950322 & rs6955448, ATT & CCC at rs409820, rs430044 & rs430046 [target iiSNP underlined]) resulting in just two distinct microhaplotypes per amplicon in our population sample. The lack of perfect association between polymorphic sites within the remaining 34 amplicons added to the diversity with 25, six, two and one amplicons displaying three, four, five and six microhaplotypes respectively. As with the autosomal STRs, the number of additional alleles observed through sequencing (Figure 5.2a) is compared with the increase in power of discrimination that they confer (Figure 5.2b).

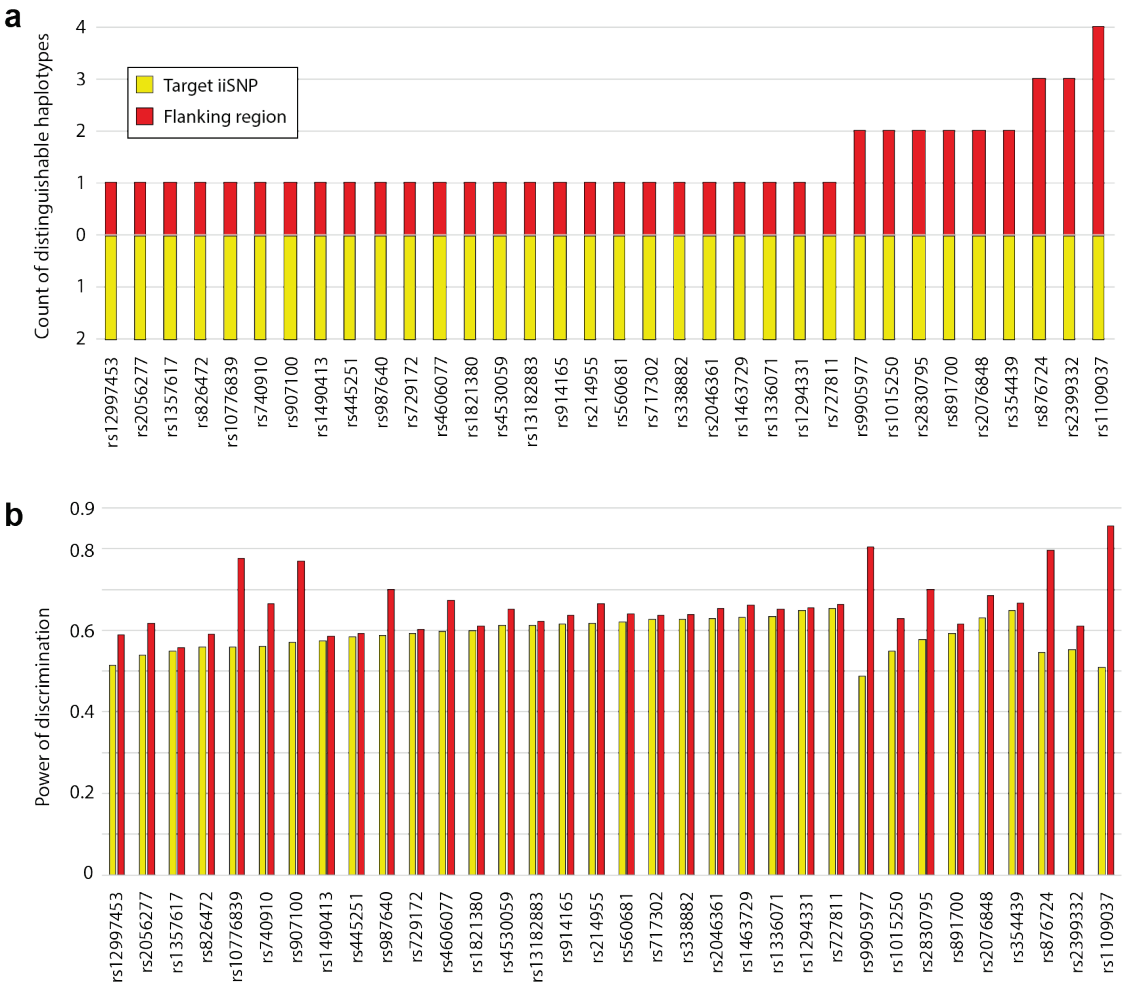


Figure 5.2: Counts of distinguishable haplotypes by SNP locus, and increment of discriminatory power due to sequence variants. a) The observed number of haplotypes defined by the targeted SNP base alone among the 89 Saudi males is shown in yellow below the x-axis, and the number of additional haplotypes resulting from sequence variation due to flanking SNPs is shown above the axis in red. b) The probability of discrimination resulting from target SNP variation alone is shown in yellow and the additional contributions made by flanking SNPs are shown in red.

The UAS highlights the target SNP in each amplicon, and also some - though not all - additional variants in the flanking regions, in its Flanking Region Report. Some of these additional variants are already described in the online database dbSNP (build 151), while others are novel. Here, eight amplicons displayed variation at positions not highlighted by the UAS. Two of these additional SNPs were associated with the rs1109037 amplicon, namely rs999755320, a rare C>T variant 1 bp upstream (at GRCh37 chr2:10085721 with a frequency of 2/30,976 in GnomAD) and a previously unrecorded T>C variant 4 bp downstream (@ 10085726). The other loci each displayed one additional rare or previously unrecorded variant site - six base

substitutions and one indel; for these, dbSNP provides some information about previous observations within large genome-wide sequencing datasets. These studies include **TOPMed** - comprising participants in US National Heart, Lung, and Blood Institute studies (http://dx.doi.org/10.1101/563866), **GnomAD** - the Genome Aggregation Database led by the Broad Institute (http://gnomad-old.broadinstitute.org/about), and an Estonian dataset comprising genetic variation from a pharmacogenomics study of adverse drug effects using electronic health records (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA489787).

- rs1015250 (+ rs1307278892 G/C, 9 bp downstream @ 1823783, TOPMed freq 5/125,568);
- rs1294331 (+ rs92280418 A/G, 4 bp upstream @ 233448409 [UAS reports reverse strand], GnomAD freq 1/30,950);
- rs13182883 (+ novel A/G SNP, 21 bp upstream @ 136633317),
- rs338882 (+ rs527589535 G/A SNP, 100 bp upstream @ 178890625 [UAS reports reverse strand], no frequency data available);
- rs354439 (+ novel A/C SNP, 30 bp upstream @ 106938381 [UAS reports reverse strand]);
- rs729172 (+ rs556717752 G/A, 32 bp upstream @ 233448409 [UAS reports reverse strand], GnomAD freq 2/30,984);
- rs891700 (+ rs543563536, T insertion in run of 10 Ts, 8 bp downstream @ chr1:239881909-239881918, freq 8/4480 among Estonians, also reported in https://doi.org/10.1016/j.fsigen.2017.09.003).

All of these were confirmed as true alleles displaying heterozygote balance between 52-98%. The sequences and frequencies of all iiSNP alleles are shown in Appendix, Table 9.8, and those instances described above where the UAS flanking sequence report does not draw attention to nine additional polymorphic sites are included.

5.3.3 Possible recombination and recurrent mutation within SNP microhaplotypes

All four possible combinations of variants at co-amplified pairs of biallelic SNPs were observed within four amplicons (target iiSNP underlined): rs2830795 & rs12626695 (separated by 43 bp), rs2399332 & rs2399334 (separated by 100 bp), rs1109037 & rs1109038 (separated by 63 bp) and rs2076848 & rs7947725 (separated by 21 bp). Although recombination could explain these, the alternative explanation of recurrent mutation at a hypermutable CpG site is applicable in the last two cases and would seem most probable given the close proximity of two SNPs. The likely product of a CpG to CpA transformation at rs7947725 was seen in one and five individuals linked to the minor (A) and major (T) allele at rs2076848, respectively. At rs1109037 it is possible that the target SNP itself may have experienced recurrent mutation also resulting in a change from CpG to CpA, as a result of methylcytosine conversion to thymine on the opposing strand (see Appendix, Table 9.7 for frequencies and sequence context).

5.3.4 Effect of combining autosomal STR and SNP sequence variants

One of the major potential benefits of typing with the DNA Signature Prep Kit is the increase in discrimination power when genotype data from independently inherited autosomal SNPs and STRs are combined. As there was no significant deviation from LD either among or between STRs and SNPs following Bonferroni correction, it is possible to combine the two marker types using the product rule. Forensic statistics for the autosomal STRs and SNPs are presented in Appendix, Table 9.9 and Table 9.10. In the Saudi Arabian dataset, the RMP decreased by three orders of magnitude due to sequence variation within the STR repeat region itself compared to length variation alone, whilst sequence polymorphisms in the flanking regions of the STRs resulted in less than one order of magnitude further reduction (Figure 5.3). The Saudi Arabian data were compared with data from four comparator populations from the USA (King et al. 2018; Novroski et al. 2016): the increases in resolution offered by sequencing in the Saudi Arabian sample are comparable with the other four groups. Notably, the difference between STR-based and iiSNP-based resolution

is smaller in the African-Americans than in the other groups, suggesting that iiSNPs may be underestimating diversity in this particular population.

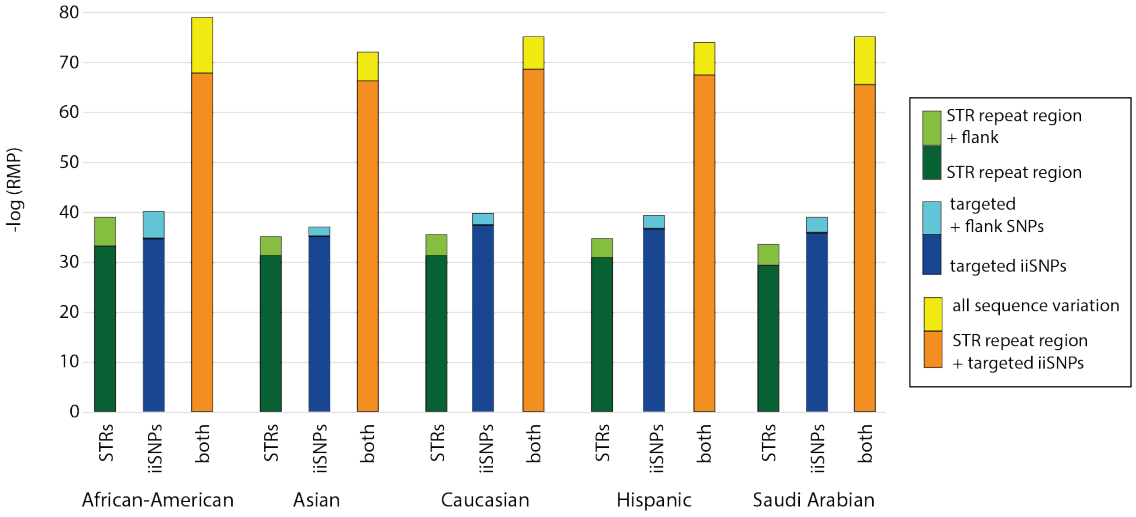


Figure 5.3: Increase in random match probability (RMP) offered by flanking sequence variation for both STRs and iiSNPs. Data are shown for the Saudi Arabians studied here, and for four comparator populations from (King et al. 2018; Novroski et al. 2016)

5.3.5 Evidence of consanguinity from patterns of STR and SNP diversity

In Chapter 4, evidence was provided from CE-based autosomal STR genotype data of the impact of consanguinity in the Saudi Arabian population (Khubrani et al. 2019b). Does the sequence-based analysis of both STRs and SNPs show a similar impact? Among STRs, testing for fit to Hardy-Weinberg expectations revealed significant heterozygote deficiency only at TPOX ($p<0.005$ after Bonferroni correction). TPOX shows no internal sequence variation, simply mirroring the observed CE length variability, and was completely concordant with the results obtained from these individuals (data not shown) with the AmpFLSTR GlobalFiler® Kit (ThermoFisher Scientific). At the CE length level of comparison, 23 out of the 27 STR loci showed an excess of homozygotes ($p=0.0002$, binomial distribution, http://onlinestatbook.com/2/calculators/binomial_dist.html; see Appendix, Table 9.11). This is consistent with 20 out of 21 GlobalFiler® Kit loci displaying heterozygote deficiency ($p<0.0001$, binomial distribution) in the CE dataset of 523

indigenous males resident in Saudi Arabia (designated KSA) (Khubrani et al. 2019b). F_{ST} between the UK-based Saudi and KSA datasets at both regional (N, E, S, W and Central) and combined levels showed no evidence of significant differences between corresponding samples, and thus the UK-based Saudi Arabian individuals can be considered to be a representative sample of autosomal variation within the indigenous population of the country as a whole (F_{ST} = 0.00086; p-value 0.09791). Heterozygote deficiency was also evident from AMOVA analysis, with an inbreeding coefficient (F_{IS}) of 0.04131 representing 4.13% of variation among our UK donors at the CE level, very close to the value obtained with GlobalFiler® from the KSA dataset (F_{IS}=0.0476; (Khubrani et al. 2019b). None of the iiSNPs showed a significant deviation from Hardy-Weinberg equilibrium after Bonferroni correction, but 63 of the 91 iiSNPs included in the analysis showed a deficiency of heterozygotes (p=0.0002, binomial distribution) mirroring the effect seen with autosomal STRs (see Appendix, Table 9.12).

5.3.6 Comparing Saudi Arabian STR and SNP sequence diversity patterns with a set of global population samples

In order to set the observed Saudi Arabian population diversity and apparent effects of consanguinity in a broader context, data were extracted from the well-characterised HGDP samples (Cann et al. 2002): STR sequence data were available from Phillips et al. (2018) and iiSNP data were extracted from publicly available whole-genome sequences (see Appendix, Table 9.13). Here, only the target iiSNPs (excluding flanking variants) were considered, as these are most likely to be reported in casework, and only the 27 HGDP populations that contain 15 or more individuals were analysed. In Figure 5.4a F_{IS} values are compared between the Saudi dataset and these HGDP populations. European, African and East Asian populations generally show wide variation in F_{IS}, with a tendency to low values; a particularly broad range is seen in Central and South Asian populations. However, the five populations from the Middle East, including the Saudi Arabians studied here, show consistently raised and highly similar levels indicative of widespread and frequent endogamy. In Figure 5.4b is also shown, for the HGDP samples only, the proportion of specific mating types (unrelated, first cousin, second cousin, avuncular) estimated

by a maximum-likelihood method from the distribution of homozygous-by-descent segments from high-density genome-wide SNP data by Leutenegger et al. (2011). This illustrates the high levels of consanguineous mating types in the Middle Eastern samples, though it also shows that some Central and South Asian samples exceed these levels.

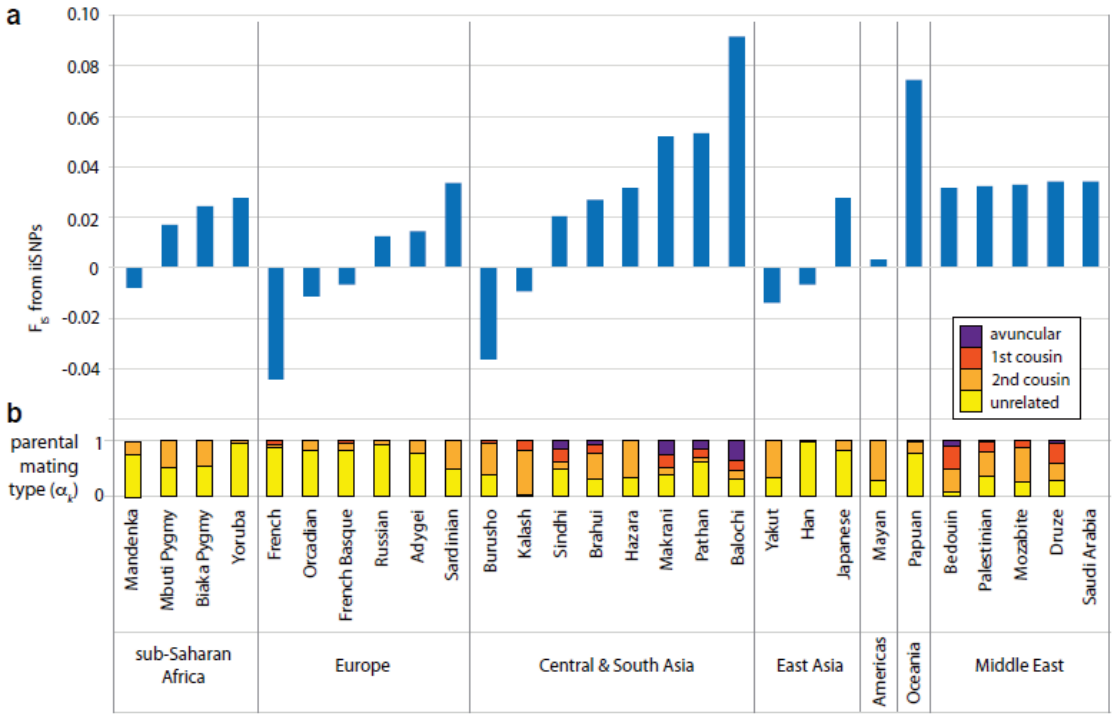


Figure 5.4: F_{IS} for autosomal iiSNPs in Saudi Arabian and HGDP populations, compared with consanguineous mating type frequencies. a) Mean F_{IS} is shown (target iiSNPs only) for HGDP populations with n≥15 and for Saudi Arabian data, clustered by continental origins; b) The coloured bars show genome-wide estimates of parental mating types for the HGDP samples, taken from (Leutenegger et al. 2011). Different mating types are shown in the inset key.

For the full set of HGDP populations the level of homozygosity observed from the iiSNPs was also compared with an estimate based on haplotype homozygosity along chromosome 16 from genome-wide SNP data previously published by Li et al. (2008) (Figure 5.5a). Again, the iiSNP-based estimates are similar for all Middle Eastern populations (including the Saudi sample), and this is also true for the haplotype homozygosity estimates for the HGDP Middle Eastern samples. The haplotype homozygosity shows a trend to increase from Africa to the rest of the Old World and to the New World, previously noted both from genome-wide SNP (Li et al. 2008) and STR data (Ramachandran et al. 2005). However, it is striking that the

iiSNP-based homozygosity values do not follow this trend, with marked underestimation of both African and Middle Eastern diversity (Figure 5.5b), probably reflecting ascertainment bias in the original choice of SNPs for individual identification.

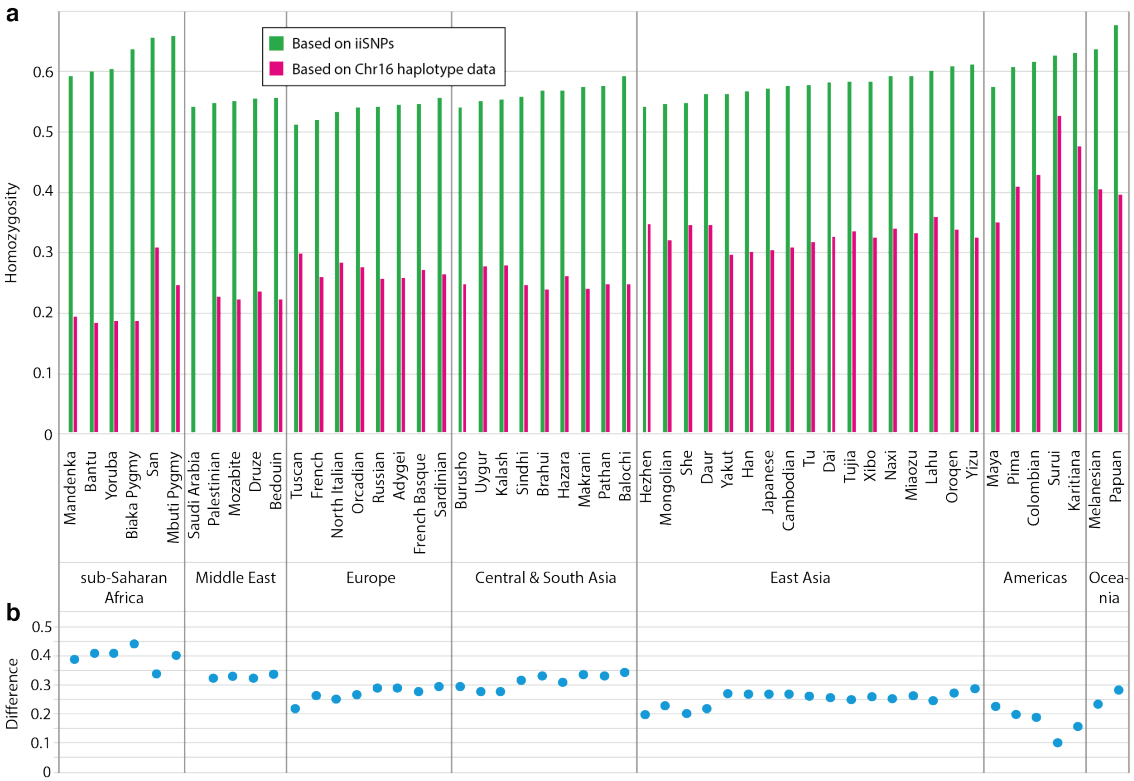


Figure 5.5: Observed homozygosity in HGDP populations based on iiSNPs, compared with a genomic estimate a) Observed homozygosity based on 91 iiSNPs, including the value for the Saudi Arabian sample studied here (red bars), and also (for the HGDP samples only) homozygosity based on chromosome 16 haplotypes (green bars) estimated from genome-wide SNP chip data and taken from (Li et al. 2008). Here, we follow (Li et al. 2008) by combining the North and South Bantu samples into a single Bantu classification. (b) The differences between iiSNP- and chromosome 16-based values, highlighting the underestimation of true homozygosity by iiSNPs in the African populations.

5.4 Discussion

A global survey of consanguinity has shown that the Middle East, North and sub-Saharan Africa and Western, Central and South Asia have consanguineous marriage levels between 20% and 50% (Bittles 2007). Arabic-speaking countries showed the highest rates in the Middle East (Tadmouri et al. 2009). Indeed first-cousin marriages are particularly common and preferred in many regional communities, including the socially conservative regions of Saudi Arabia (El-Mouzan et al. 2007) where they comprise up to 33% of marriages. Saudi Arabia is also remarkable for its very rapid expansion in population size from 3.9 million in 1950 prior to the oil boom, to 20.8 million indigenous Saudi citizens in 2018, largely as a result of decreased child mortality and increasing life expectancy. Consequently, it is expected that clusters of closely related and genetically similar individuals will be commonplace, requiring more discriminating tests to generate the same confidence of individual resolution. Here, the high (sequence-based) resolution of MPS analysis, and the large number of independent autosomal loci (27 STRs and 91 iiSNPs) were used to analyse a sample of Saudi Arabians in order to address this question.

The lack of significant differentiation between the UK-based sample of Saudi Arabians analysed here and the previously described (Khubrani et al. 2019b) combined KSA dataset (F_{ST} =0.00072), suggests that the autosomal STR sequence dataset here is representative of autosomal variation in the indigenous Saudi population as a whole. This study also replicates previous findings of heterozygote deficiency in Saudi populations (Alsafiah et al. 2017; Khubrani et al. 2019a; Khubrani et al. 2019b) likely due to high levels of endogamy (F_{IS}=0.04131 for autosomal STRs based on CE length and 0.04201 for iiSNPs) although the loci showing significant deviations differ between studies. The previous observation (Chapter 4) of an apparent significant deficiency of heterozygotes in the KSA population at D2S1338, thought to be partly due to a null allele, was not replicated in this Chapter. However, this is unsurprising as the primer positions differ between the GlobalFiler® and DNA Signature Prep kits, the latter of which has an amplicon approximately 180 bp shorter. In contrast the apparent homozygous excess seen at TPOX here was clearly not due to null heterozygotes caused specifically by the DNA

Signature Prep Kit primers, as all homozygous individuals were also scored as homozygotes with GlobalFiler®, and no significant excess of such "homozygotes" was seen in a much larger KSA dataset typed with the GlobalFiler® kit (Chapter 4; Khubrani et al. 2019). It seems likely that TPOX by chance showed the greatest excess of homozygosity due to consanguinity. Significant homozygous excess has been detected in many populations in the Middle East affecting a variety of loci: Iraq: D21S11 , FGA (Farhan et al. 2016), Qatar - D13S317, D19S433 and vWA (Perez-Miranda et al. 2006), Kuwait - vWA (Alenizi et al. 2014) and Tunisia - CSF1PO, Penta D and TPOX (Brandt-Casadevall et al. 2003). The lack of consistency across loci suggests that null alleles are not the general cause, and that demographic and stochastic factors are more likely.

The detection of sequence variation within autosomal STR amplicons enhances resolution by distinguishing among isoalleles that appear identical in CE analysis. Previous surveys of diverse ethnic groups have revealed differences of several orders of magnitude in random match probability (RMP) between length- and sequence-based estimates. Novroski et al. (2016) obtained the following length- and sequence-RMP estimates for African-Americans (8.5410E-34, 1.31E-39), Asians (6.37E-32, 8.66E-36), Caucasians (6.28E-32, 3.63E-36) and Hispanics (1.51E-31, 1.23E-35), whilst the levels of resolution were lower within our Saudi population (2.61E-30- length, 2.07E-33 - repeat region sequence & 3.49E-34 - including both repeat and flanking sequence). This reflects their more restricted geographic origins and therefore expected lower levels of diversity.

In previous studies, although sequencing of some STRs provided up to two-fold increased resolution per locus, other loci such as D10S1248, Penta D, Penta E and TPOX showed no improvement, as observed in this dataset (Churchill et al. 2016; Gettings et al. 2016; Just et al. 2017). This is heavily influenced by the complexity of the STR repeat structure, but novel variants can arise in other STRs with a typically simple repeat structure as demonstrated by CSF1PO and TH01, both of which were invariant in previous studies but here displayed rare or regionally restricted variants that may be seen more frequently as datasets of Middle Eastern populations grow. Nine of the STR loci showed no improvement with sequencing, but six showed

a greater than two-fold increase within the Saudi population. These include D12S391, where 13 length variants increased to 38 sequence variants, accompanied by a two-fold improvement in RMP from 0.0357 to 0.0173; and D3S1358, which saw a three-fold improvement linked to an increase from seven length variants to 13 sequence variants. Whereas most gains were due to novel arrangements of repeat variants in compound STRs, a greater than two-fold improvement resulted at D20S482 due solely to the presence of SNPs flanking the repeat array.

As with the STRs, sequencing of the iiSNP amplicons revealed additional sequence variants within the regions flanking the target SNP. Notably, not all of these variants were highlighted in the UAS Flanking Region Report, so care is needed in analysing data to ensure that all variation is recorded. Of the 91 iiSNPs analysed here, 34 showed additional sequence variation. This led to the RMP of 9.97E-37 for the target iiSNPs alone decreasing to 8.88E-40 when the flanking sequence variants were included. Most additional variants create novel rare microhaplotypes within the Saudi sample. The microhaplotypes created by the flanking sequence variation deserve further investigation in larger and more diverse samples.

It has been reported that the iiSNPs alone in the DNA Signature Prep Kit provide greater discrimination power than do available commercial STR kits (King et al. 2018), and this was certainly true within the Saudi dataset generated here: taking sequence variation into account, the RMP of the autosomal STRs was 3.49E-34 compared with 8.88E-40 for the iiSNP amplicons. Overall, this study demonstrates that, while the additional alleles revealed by MPS substantially improve discrimination, the greatest contribution comes from the ability to analyse many independent loci simultaneously. The added power offered by combining STRs and SNPs will be beneficial in challenging cases, such as mixture analysis, complex kinship cases and where degradation results in partial profiles.

In conclusion, MPS analysis of autosomal loci in a representative sample of Saudi males demonstrated that profiling can significantly increase the discrimination power of forensic DNA testing through the simultaneous amplification of both STRs and identity-informative SNPs. This small-scale survey has uncovered a number of previously unobserved and rare alleles that may be present at significant

frequencies within the Arabian Peninsula, supporting the value of larger-scale surveys of this region. A striking feature of the data was a general deficiency of heterozygosity, and comparison with HGDP samples supports the idea that this reflects the practice of consanguinity in Saudi Arabia and in other Middle Eastern populations.

Chapter 6 Massively Parallel Sequencing of X- and Y-STRs in a Saudi Arabian population sample

6.1 Introduction

Whilst multiplexing of STR loci for capillary electrophoresis (CE) is severely constrained by the need to avoid overlapping size ranges, massively parallel sequencing (MPS) allows the simultaneous analysis of many more markers from a single forensic sample (de Knijff 2019). Today's MPS tests are limited only by the considerable complexities of efficiently and uniformly amplifying a much larger number of loci. A leading example is Verogen's ForenSeq™ DNA Signature Prep Kit (Churchill et al. 2016) capable of detecting more than 230 loci including the standard autosomal and Y-STRs plus X-STRs as well as >90 identity-informative SNPs (iiSNPs) and >80 biogeographical ancestry- and phenotypically-informative SNPs. The inclusion of X- and Y-chromosomal markers along with the standard panel of autosomal STRs opens other investigative avenues if a simple autosomal match is not obtained. Y-STRs can detect the male component of mixtures in sexual assault cases where the autosomal profile of the offender may be lost among the background of female DNA (Purps et al. 2015). They also provide highly informative lineage markers for tracing relatedness among individuals sharing a common male ancestor (King and Jobling 2009). Over a few generations this can provide links to surnames or social groupings based on common patrilineal descent, and over longer timescales an indication of biogeographic ancestry. X-STRs are also useful in male: female mixtures where conversely the male component overwhelms the female. As males are hemizygous for the X chromosome, traces of the two female X chromosomes are more likely to be interpretable than the female's autosomes which have a greater chance of being masked (Szibor 2007; Szibor et al. 2003). X-STRs can also play a key role in establishing close relationships when key relatives are unavailable for testing (Tillmar et al. 2017).

Most MPS studies have concentrated on the autosomal STR markers due to their predominant role in current forensic casework. With the obvious exception of ancestry-informative markers, most markers on the autosomes display low levels of

population differentiation partly because their exposure to genetic bottlenecks is reduced; even a single pair of parents possess four potentially different copies of every autosomal locus. By contrast, uniparentally inherited markers (the single Y passed from father to son, and mitochondrial DNA from mother to child), have a much lower effective population size and are thus more susceptible to drift resulting in local changes of frequency (Jorde et al. 2000). Even the X chromosome (passed from father to daughter, and mother to child) has an effective population size just three quarters that of the autosomes and so is expected to show an intermediate level of differentiation (Charlesworth 2009). Populations which have persisted in small numbers for many generations and have then experienced rapid growth can show particularly high levels of differentiation from others, due to founder effects. Consequently there is a requirement for studies to provide population-specific allele frequency data and to identify any interpretation issues that might arise from rare alleles that may not have been detected in the large European, American and East Asian datasets that are routinely used to validate emerging technologies and set reporting guidelines.

In previous Chapters, conventional CE methods were used to characterise levels of autosomal and Y-STR diversity in indigenous males sampled across five regions of the Kingdom of Saudi Arabia (Khubrani et al. 2018; Khubrani et al. 2019b). The country is a vast and, until recently, very sparsely populated region characterised largely by inhospitable desert which was capable of supporting only small nomadic groups. With the discovery of oil less than 100 years ago a rapid population boom was fueled as increased wealth improved health care and provided the infrastructure to support large populations even in the harsh environment (Al-Rasheed 2010; Wynbrandt and Gerges 2010). The nomadic tribal groups settled and expanded within their own regions leading to a highly structured population based on a small number of patrilineal descent groups in each town (Al-Hathloul and Edadan 1993). Differentiation was also enhanced by the tradition of polygamy (Panter-Brick 1991) which is likely to have further restricted Y-chromosome diversity and allowed certain lineages to predominate in local settlements. Although first-cousin marriages are still very common, wives tend to move to their husband's home (Galaty and Salzman 1981) and so a greater homogenisation of mitochondrial

and autosomal sequences is expected, with X-STRs likely to reflect the pattern of greater homogeneity demonstrated by mitochondrial markers, due to their residence in women for two generations out of three on average.

In this Chapter, sequence variation within Y- and X-STRs in a Saudi Arabian population is investigated using the DNA Signature Prep Kit. The population sample comprises 89 Saudi Arabian males currently resident in the UK for whom Chapter 5 reported the autosomal STR and iiSNP data obtained using the same kit. The aims of this Chapter are:

(i) to further explore whether this sample faithfully reflects Saudi genetic diversity;
(ii) to measure the increase in genetic discrimination provided by the Y-STR sequences and how they correlate with Y-SNP haplogroups;
(iii) to ask whether the X-STRs show regional differentiation;
(iv) to ask, for both marker types, whether local variants exist and how they are handled by the currently available bioinformatic tools; and
(v) to compare levels of diversity with other global populations.

6.2 Materials and methods

6.2.1 Samples

As described in Chapter 3, samples were collected from 89 indigenous Arab males residing in the United Kingdom whose continuous paternal line ancestry can be traced back to a great-grandfather within one of the five geographical subdivisions of Saudi Arabia (Khubrani et al. 2018) (Central N = 19, Northern N = 9, Southern N = 26, Eastern N = 16, and Western N = 19). Although recruitment was based on paternal ancestry, in all but seven cases the birthplaces of participants' paternal and maternal ancestors come from the same geographical region.

6.2.2 Y and X-STR profiling with the DNA Signature Prep Kit

DNA was extracted and quantified as previously described (Khubrani et al. 2018).

The ForenSeq™DNA Signature Prep Kit was used for library preparation for 24 Y-STRs (DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS437, DYS438, DYS439, DYS448, DYS460, DYS481, DYS505, DYS522, DYS533, DYS549, DYS570, DYS576, DYS612, DYS635, DYS643, DYS385a-b, DYF387S1, Y-GATA-H4) and 7 X-STRs (DXS10074, DXS10103, DXS10135, DXS7132, DXS7423, DXS8378, HPRTB). Massively Parallel Sequencing (MPS) was carried out on a Verogen MiSeq FGx™ device following the manufacturer's protocol. Data analysis followed the same approach as in Chapter 5 (Khubrani *et al.* 2019b) using both the Verogen ForenSeq™ Universal Analysis Software (UAS) and STRait Razor v3 (Woerner et al. 2017).

6.2.3 Median-joining networks and Y haplogroup assignment

A median-joining network (Bandelt et al. 1999) based solely on Y-STR allele length was created with Network 5.0 and Network Publisher (http://www.fluxus-engineering.com/sharenet.htm). Intermediate alleles were rounded to the nearest integer and constitutively duplicated loci (DYS385a,b; DYF387S1,S2) were excluded from network construction. The network was coloured according to previously reported Y-SNP haplogroup predictions made using the NevGen online Y-DNA Haplogroup Predictor (Khubrani et al. 2018). Assignments were originally based on Yfiler® Plus Y-STR profiles using a previously-implemented Bayesian approach (Athey 2006) with the additional consideration of pairwise correlation of alleles between different Y-STRs in the calculation of haplogroup probabilities (http://www.nevgen.org). Haplogroup definition was standardised to a broad resolution of ten haplogroups (A, E1b1b, E1b1a, G, J1, J2, L, Q, R1a and T) for which the NevGen software correctly predicted (Chapter 3) the known haplogroups among 743 self-declared SNP-typed Saudi Arabian males downloaded from FamilyTreeDNA (www.familytreedna.com). As Yfiler® Plus lacks two Y-STRs (DYS549 and DYS643) used by the predictor, the predictions were repeated using the updated NevGen Genealogy Tools v1.1 desktop software package incorporating the missing STRs from the DNA Signature Prep Kit profiles: the old and new predictions are concordant. To be conservative the six samples flagged with a probability of inappropriate prediction greater than 95% were denoted as unpredicted. To these were added the single individual with a NevGen prediction of

haplogroup N, even though the predictor had calculated a 99.6% likelihood of correct assignment, because there were no examples of this haplogroup in the Saudi FTDNA database to confirm the accuracy of prediction.

6.2.4 Sanger sequencing of mtDNA

A section of the mtDNA control region was amplified with primers L15999 (5′-CACCATTAGCACCCAAAGCT-3′) and H409 (5′-CTGTTAAAAGTGCATACCGCC-3′) as one amplicon (Sigurðardóttir et al. 2000) with 11x PCR buffer (0.90 µl in a 10 µl reaction) (Kauppi et al. 2009), 12.5 mM Tris base, 0.3 µM of each primer, 0.015 U Pfu and 0.3 U *Taq* polymerases. PCR commenced with 95°C for 3 min, followed by 31 cycles at 94°C for 30 s, 60°C for 30 s and 70°C for 30 s, then ExoI/SAP purification to remove single-stranded DNA and deactivate unincorporated dNTPs during incubation at 37°C for 1 h, 80 °C for 15 min and 4°C for 15 min. The yield of the targeted PCR product was estimated by agarose gel electrophoresis against 2 µl of 1 kb HyperLadder™ (Bioline). Sequencing reactions of 20-30 ng/kb PCR product were performed with Big Dye Terminator mix v 3.1 and excess dye removed with QIAGEN DyeEx 2.0 spin columns. Samples that showed low quality following long uninterrupted poly-C tracts (>9 bp) were re-sequenced with additional primers 5′-CATGCTTACAAGCAAGTACAGC-3′ and 5′-GCTGTGCAGACATTCAATTG-3′.

Sequence Scanner (Applied Biosystems) was used for initial sequence data interpretation and checking data quality. BioEdit (Hall 1999) and CodonCode Aligner (CodonCode Corporation) were used to visualise and interpret DNA sequences collectively. Sequences were aligned with ClustalW within MEGA6 (Tamura et al. 2013) using default parameters. The mtDNA nomenclature tool within mtDNAprofiler (http://mtprofiler.yonsei.ac.kr:8080/index.php?cat=1) was used to prepare input files (.hsd) for haplogroup predictions using HaploGrep 2.0 (https://haplogrep.uibk.ac.at/) which provided an index of prediction quality.

6.2.5 Population, forensic and statistical analysis

Arlequin v 3.5 (Excoffier and Lischer 2010) was used to investigate pairwise linkage disequilibrium (LD) and pairwise genetic distance (F_{ST}). F_{ST} comparisons were made

with multidimensional scaling (MDS) plots created with the (MASS) package in the R Library (Venables and Ripley 2013). STRAF (Gouy and Zieger 2017) was used to calculate forensic statistics including: genotype count (N), allele count based on sequence (N_{all}), genetic diversity (GD), polymorphism information content (PIC), random match probability (PM) and power of discrimination (PD). Allele frequencies were calculated in Excel (allele count/total). Discrimination capacity (DC) was calculated as the number of unique haplotypes divided by the total number of haplotypes in the dataset. Compound power of discrimination (CPD) was calculated from the formula (1- product [1-PM]). The StatX package (Guo 2017; Lang et al. 2019) was used to calculate X-STR forensic parameters for all loci and for the two pairs of loci in the previously defined linkage groups (Szibor et al. 2003) LG1 (DXS8378 and DXS1035) and LG2 (DXS10074 and DXS7132).

6.3 Results

The DNA Signature Prep Kit Primer Mix A gave an average depth of coverage per locus per individual of 902 reads for the 24 Y-STRs and 1004 reads for six of the seven X-STRs among the 89 Saudi males. Following visual checks of individual loci flagged by the default settings of the UAS, genotypes were called for all 24 Y-STRs and for six X-STRs. DXS10103 was the lowest performing marker with 26 sub-threshold samples, and was removed from the analysis; DYS392 and DYS522 were sub-threshold in three and two cases respectively, but were retained. Average read depths for the Y-STRs ranged from 137 at DYS522 to 3474 at DYS438, and from 641 at DXS8378 to 1350 at DXS10074 at the X-STRs.

6.3.1 X and Y-STR sequence variation and impact on discrimination

Among 534 X-STR alleles typed in the 89 Saudi males, the UAS identified 56 length variants (equivalent to CE alleles) and 75 repeat sequence sub-variants across the six loci, and amongst the 2295 Y-STR alleles there were 147 length variants and 192 repeat sequence sub-variants. The sequences and frequencies of all sex-chromosomal STR alleles are shown in Appendix, Table 9.14. Three X-STR loci showed no additional sequence variants (HPRTB, DXS7423 and DXS8378) with the latter two showing the lowest diversity with just five distinct allele lengths. Of the

Y-STR markers DYS389I, DYS437, DYS438 and DYS460 each displayed just three length variants, and along with eight other loci provided the same discrimination power as would be seen with a conventional CE approach. However, additional sequence variation within the repeat regions of eight Y-STRs contributed up to 17 additional alleles (at DYS389II) with similarly varied contributions at the X-STRs, with one extra variant at DXS10074, up to fourteen additional alleles at DXS10135. The Flanking Region Report revealed further alleles differentiated by sequence polymorphisms outside the repeat region at five Y-STRs contributing a single distinct variant at each of DYS612, DYS437, DYS533 and DYS481, and two variants at DYS438. DXS7132 and DXS10074 each show one additional allele due to flanking region variation (Figure 6.1a).

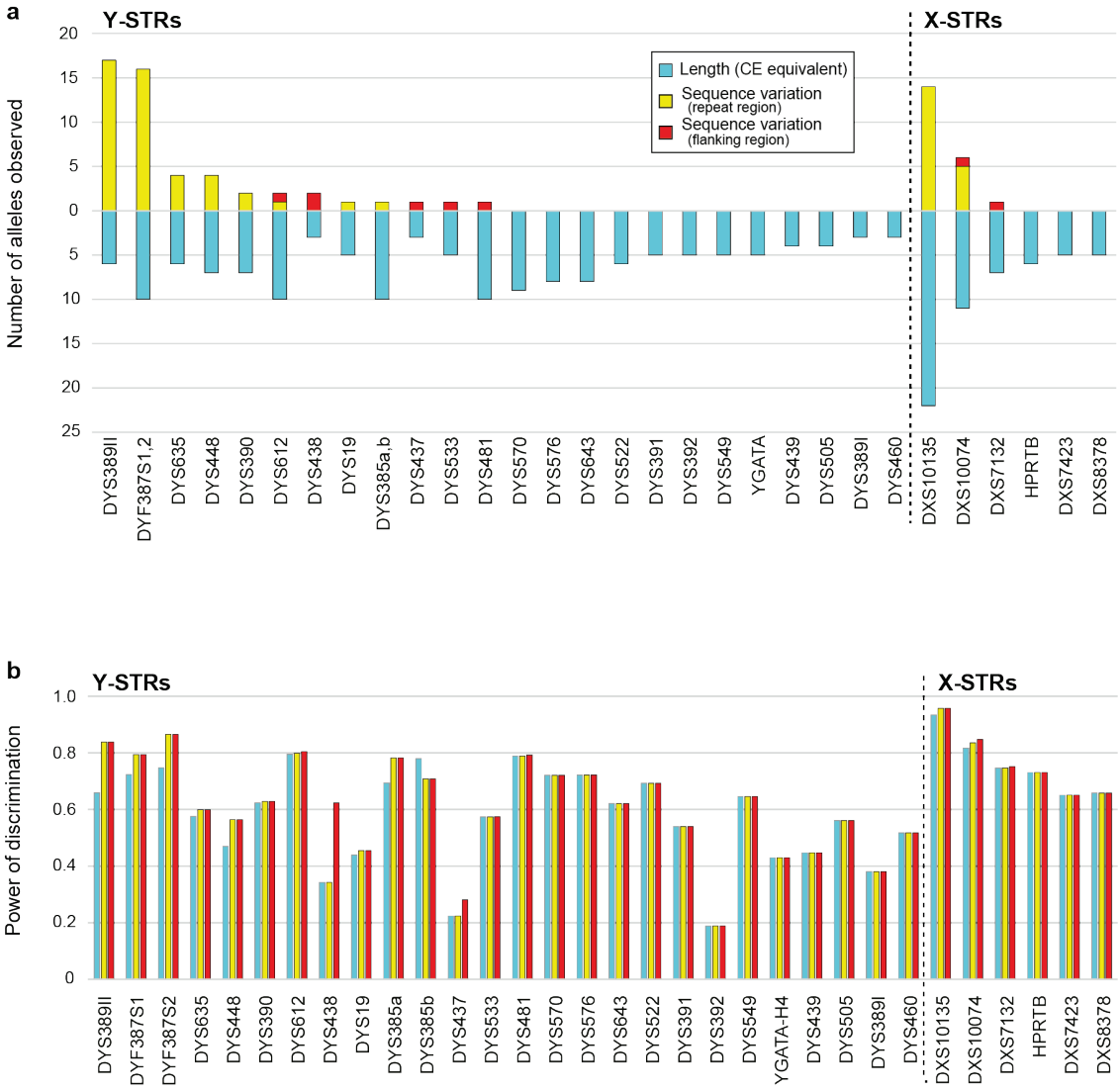


Figure 6.1: Numbers of distinguishable alleles by STR locus, and per-locus increase of discriminatory power due to sequence variants. a) The number of observed length variants among 89 Saudi males is shown in blue below the x-axis, and the number of additional alleles resulting from sequence variation within and flanking the repeat array are shown above in yellow and red respectively. b) The power of discrimination resulting from length variation only (equivalent to CE analysis) is shown in blue and the contributions made by sequence variation within and flanking the repeat array are shown in yellow and red respectively

6.3.2 Y and X STR sequences

The analysis identified 27 Y-STR alleles and six X-STR alleles which were absent from the STRait Razor v3.0 default allele list, these are summarised in Appendix, Table 9.15. These variants were searched for in GenBank (July 2019) and among the profiled individuals (Phillips et al. 2018) from the Human Genome Diversity Project (HGDP) panel (Cann et al. 2002) to determine the likely geographic distribution of these Y-STR and X-STR alleles; a tendency was noted for them to occur in the Middle East, Africa and South Asia.

6.3.3 Concordance and interpretation issues

The MPS samples had previously been CE typed with the Yfiler® Plus multiplex (Chapter 3) permitting a direct comparison at the following loci: DYF387S1a/b, DYS19, DYS385a/b, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS437, DYS438, DYS439, DYS448, DYS460, DYS481, DYS533, DYS570, DYS576, DYS635 and Y-GATA-H4. All but one allele length designation was concordant between the two methods at these loci. Concordance tests could not be performed for the remaining five Y-STR loci (DYS505, DYS522, DYS549, DYS612 and DYS643) which are not included in the Yfiler® Plus multiplex, nor for the X-STRs. Compared with autosomal STRs, those on the X and Y have been less intensively studied using MPS approaches, and may be more affected by population sub-structuring. Consequently, it is likely that the discovery of genetic variants strongly associated with under-studied ethnic groups will highlight more issues with interpretation, bioinformatic analysis and cross-platform concordance than autosomal STRs. Here some of the issues noted in this relatively small Saudi dataset are described.

6.3.3.1 *Discordance between ForenSeq™ Universal Analysis Software and CE result*

A single length discordance was observed between CE and MPS data, which was probably a result of bioinformatic treatment of the MPS data. In one individual UAS called a DYF387S1 allele 41, whereas STRait Razor designated it as 40.1, which is concordant with both the CE result and a visual examination of the sequence. Upon examination of the FASTQ files a 3-bp deletion at a flanking site was found where there are normally three consecutive repeats of GAA, but in this donor just two. STRait Razor uses the anchor sequence 5´-TAACAAGAAAAAAAGAAGAA-3´ and starts calculating allele length with the third GAA repeat, so would report a .1 allele if the third GAA was missing; however, though it is not clear what criteria UAS uses, it can be speculated that calling commences after the last detected GAA thereby failing to take into account the 3-bp deletion (Figure 6.2). Further examples of this deletion were sought in the HGDP panel by selecting three samples (HGDP00382, HGDP00267 and HGDP00277) that share the same repeat region sequence, but CE genotyping did not find this deletion among these samples.

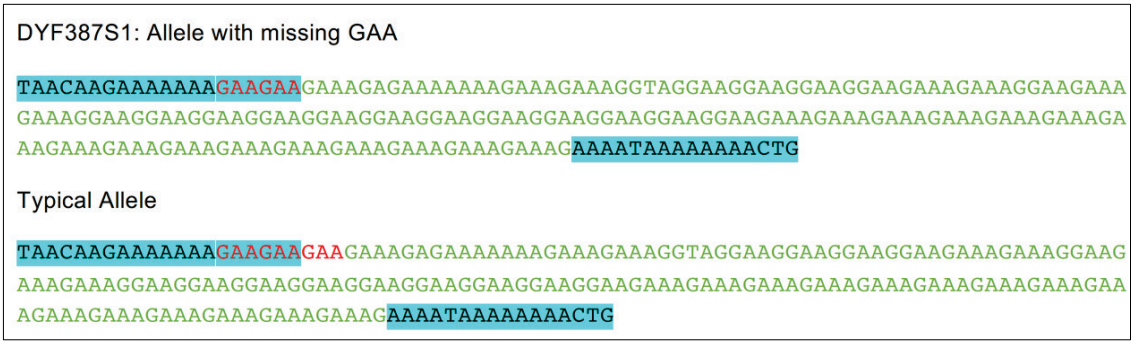


Figure 6.2: Discordance between ForenSeq™ Universal Analysis Software and CE result for DYF387S1.STRait Razor anchor positions are highlighted in light blue and the consecutive repeats of GAA are shown in red.

6.3.3.2 *Bioinformatic null allele at DXS7132*

The UAS did not output any data for DXS7132 in one individual, despite an examination of the raw FASTQ files with STRait Razor revealing 1240 reads from this locus. The “missing” allele was identical to a common variant at this locus (CE

15) except for a previously unreported G>C change 7 bp downstream of the last pure GATA repeat. This rare base substitution likely disrupts the anchor sequence used by the UAS to identify the repeat region resulting in no matching reads being returned by the UAS interface (see Figure 6.3); this issue will be corrected in a future UAS update (Richard Kessell, Verogen, personal communication).

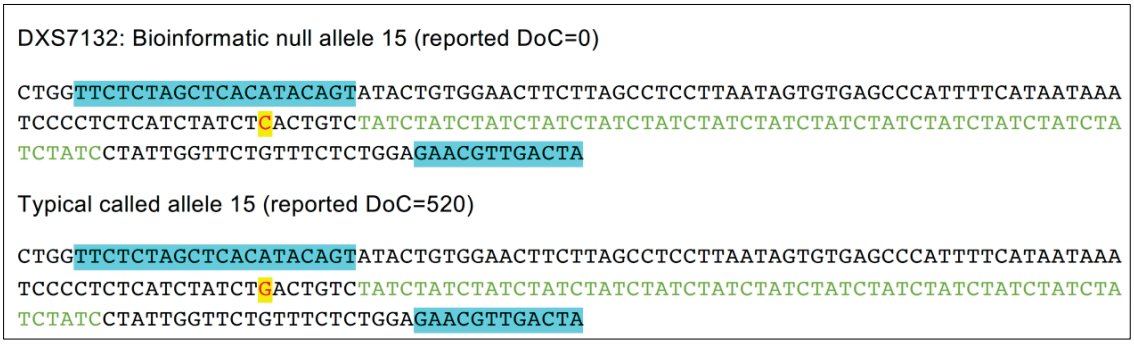
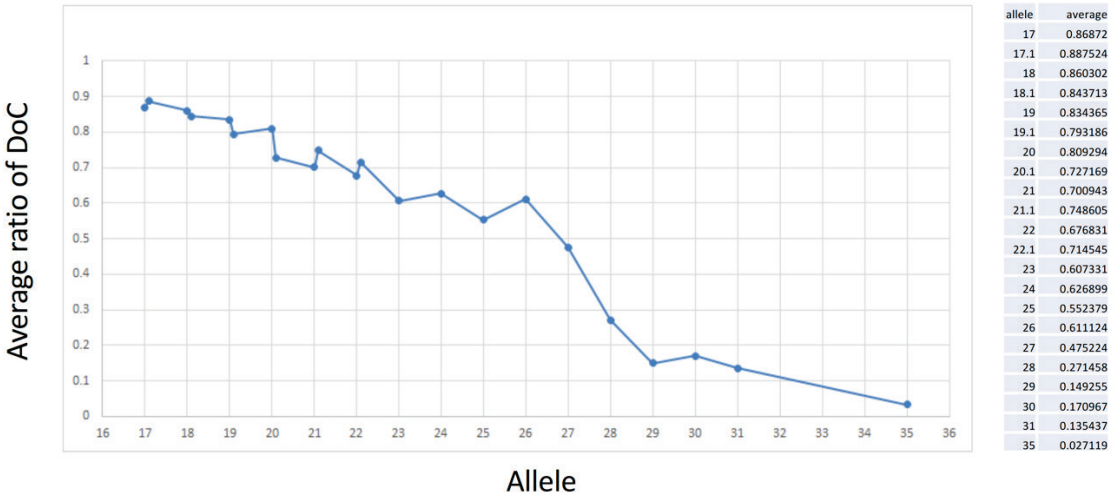


Figure 6.3: Bioinformatic null allele at DXS7132 due to interruption of an anchor sequence by a base substitution. STRait Razor anchor positions are highlighted in light blue and SNPs for the null allele and an example of a typical allele are highlighted in yellow.

6.3.3.3 Depth of coverage at DXS10135

Although it did not result in any miscalls, a striking difference in the depth of coverage was observed between the Project Report and Flanking Region Report for DXS10135 in almost half of the samples. The difference increased with the length of the allele at this particularly long STR, presumably as quality-control filters are increasingly likely to reduce reported reads as the allele length increases. For the longest alleles the depth of coverage in the Flanking Report is only a tenth of that shown in the Project Report (Figure 6.4). This raises the possibility of loss of flanking data in low-input samples.



allele	average
17	0.86872
17.1	0.887524
18	0.860302
18.1	0.843713
19	0.834365
19.1	0.793186
20	0.809294
20.1	0.727169
21	0.700943
21.1	0.748605
22	0.676831
22.1	0.714545
23	0.607331
24	0.626899
25	0.552379
26	0.611124
27	0.475224
28	0.271458
29	0.149255
30	0.170967
31	0.135437
35	0.027119

Figure 6.4: Difference in the depth of coverage between the Project Report and Flanking Region Report at DXS10135. DoC is negatively correlated with allele length.

6.3.3.4 Nomenclature discordance between the ForenSeq™ Universal Analysis Software and STRait Razor at DYS612

Currently the UAS reports allele length as six repeats shorter than STRait Razor at DYS612, a rapidly mutating Y-STR which is not present in the mainstream CE multiplexes or reported in YHRD. This difference has already been highlighted in the literature and arises from early genetic surveys suggesting that the initial $(CCT)_5(CTT)$ motif of the complex DYS612 locus (normally represented by the structure: $(CCT)_5(CTT)$ $(TCT)_4(CCT)(TCT)_n$) was invariant (Li et al. 2019; Novroski et al. 2016; Wendt et al. 2017). However, the allele in HGDP00775 starts $(CCT)_4(TCT)_4CCT(TCT)_{24}$, and another 37 HGDP samples (Phillips et al. 2018) have structures in the range $(CCT)_4TTT$ CCT $(TCT)_4CCT(TCT)_n$. These data strongly argue in favour of inclusion of the extended motif in order to capture all variation in MPS studies.

6.3.4 Y-chromosome haplogroup-associated SNPs and repeat motifs

A median-joining network of the Y-STR haplotypes based on allele length data was used to visualise any correlation between sequence variants such as SNPs and repeat structure motifs and haplogroups. Y haplogroups were provisionally

assigned by the NevGen Genealogy Tools v1.1 prediction software based on the 23 Y-STRs in the PPY23 multiplex. This was achieved by combining allele-length data from the 20 overlapping loci detected with the Signature DNA Prep Kit with the remaining three Y-STRs typed previously with Yfiler® Plus (Khubrani et al. 2018). Examination of the flanking sequences revealed five SNPs (namely, L255, M4790, BY7692, Z16708 and S17543) which had been previously linked to specific haplogroups by sequencing of Y chromosomes (YFull Tree). L255, within the flanking region of DYS438, is a defining SNP for haplogroup J1, which accounts for 55% of the current data set and is one of the most common Y haplogroups in Middle Eastern populations (Chiaroni et al. 2010). The DYS437 amplicon of three samples displayed the derived allele at M4790, a defining SNP for hg E1b1a (equivalent to E-M2), the commonest haplogroup in sub-Saharan Africa (Wood et al. 2005), the other three SNPs each seen in single individuals define two distinct sub-haplogroups in G and one in E1b1b. In total these five SNPs provide haplogroup information for 62% of the dataset and were completely concordant with NevGen-based predictions. Table 6.1 describes the location and haplogroup association of each of these SNPs.

Table 6.1: Location and haplogroup association of Y-SNPs detected by the DNA Signature Prep Kit

Y-STR	Location and rs Number	predicted-hg	ISOGG Ybrowse	No. with SNP motif	No. with haplogroup	Associated ISOGG haplogroup	Associated Yfull haplogroup
DYS437	DYS437 [CE 14] 12346264-T (rs9786886)	E1b1a	M4790	3	3	N/A	E-M2
DYS438	DYS438 [CE 10] 12825955-C (rs760613324)	J1	L255,PF4706	48	49	J1	J1
DYS438	DYS438 [CE 11] 12825955-C (rs760613324)	J1	L255,PF4706	1	49	J1	J1
DYS481	DYS481 [CE 23] 8558404-C (rs769329768)	E1b1b	BY7692	1	15	E	E-V3262
DYS533	DYS533 [CE 9] 16281313-G (rs NA)	G	Z16708	1	3	G	G-Y36006
DYS612	DYS612 [CE 36] 13640936-T (rs574356875)	G	S17543, 3315	1	3	G2a2b1	G-M406

In addition to SNP variants it has been demonstrated that repeat arrays within complex STRs which are too short to undergo frequent replication slippage can be extremely stable and persist for thousands of generations (Huszar et al. 2018). The repeat structures of all Y-STRs were examined, and mapped onto a standard median-joining network based solely on allele length data (such data are also the basis of NevGen haplogroup prediction) (Figure 6.5). A combination of SNPs and stable STR motifs is enough to characterise most (91%) haplotypes into the concordant predicted haplogroup.

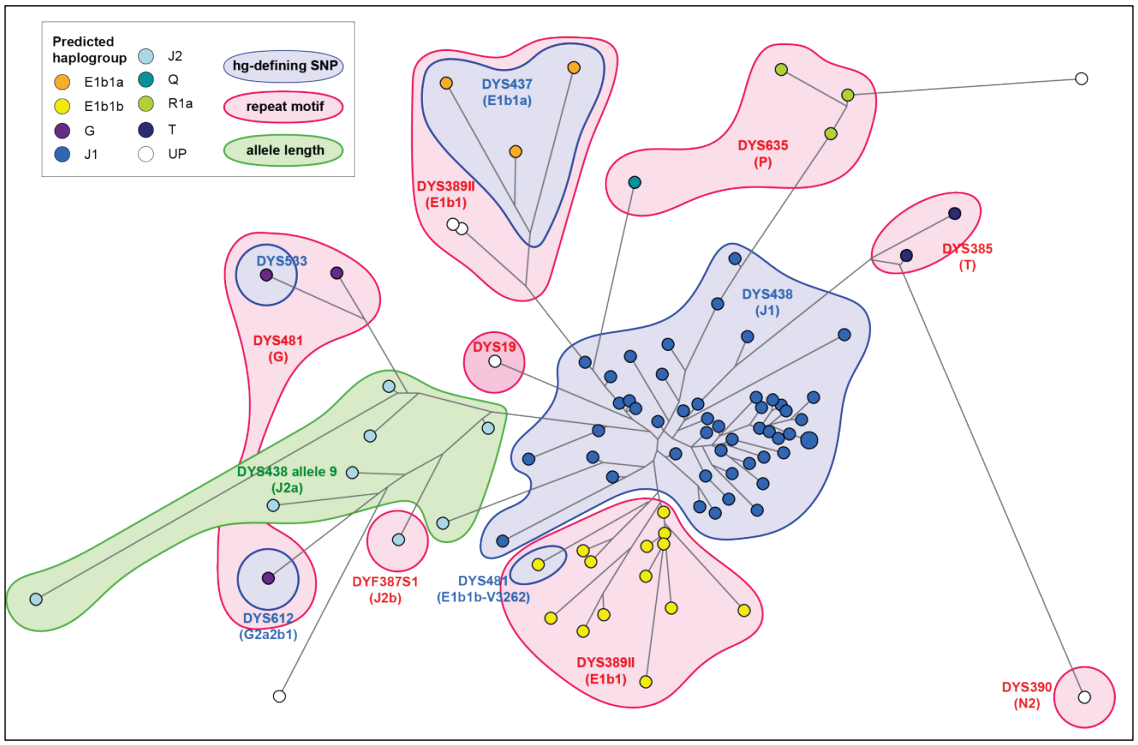


Figure 6.5: Median-joining network showing the association of SNPs and structural variants with haplogroups. The network of haplotypes in 89 Saudi males is based solely on allele length. Nodes represent haplotypes, with the single larger node indicating one shared by two men. Lines connecting nodes have lengths proportional to the number of mutational steps. Each node is coloured according to its NevGen haplogroup prediction as shown in the key, including six haplotypes flagged as having a probability of inappropriate prediction greater than 95% which are shown as unpredicted (UP). Shapes enclose groups of haplotypes sharing a distinctive characteristic at a particular locus - blue indicates a defining flanking SNP, pink a defining sequence motif and green a characteristic allele length. Where the phylogenetic range of the SNP allele or motif has been previously established this is shown in brackets after the locus.

6.3.5 Association of Y-STR repeat structure motifs with specific Y-haplogroups

Some structures were only seen in a single individual, such as the bearer of a haplotype weakly linked by NevGen to the predominantly African (King et al. 2007) haplogroup A who had a terminal $(TCTA)_2$ rather than the $(TCTA)_3$ seen in all other DYS19 alleles. Other structures are shared by many individuals, including the terminal $(CAGA)_6$ or $(CAGA)_7$ at DYS389II which has previously been shown to be characteristic of haplogroup E (Huszar et al. 2018). The $(CAGA)_{6,7}$ motif is seen in all 18 predicted haplogroup E individuals but also one J1 individual which reflects the increasing likelihood of recurrence as repeat blocks become longer and more likely to experience replication slippage. This motif links two disparate branches of the network and is further subdivided by the previously mentioned SNP which distinguishes between the two main E lineages, E1b1a and E1b1b. Similarly, the three predicted haplogroup G individuals occur on two separate network branches but are linked by a duplication of the single CTG trinucleotide which normally precedes the $(CTT)_n$ array at DYS481, a locus usually considered to comprise a single variable trinucleotide block.

Duplications can involve more than one repeat block. Super-haplogroup P, which encompasses haplogroups Q and R, is characterised by duplication of the entire terminal $(TACA)_2(TAGA)_4$ motif of DYS635. This structure was associated with all haplogroup Q and R individuals (Huszar et al. 2018) but was also seen in two haplogroup A2 individuals, and so cannot be taken as definitive evidence of haplogroup P membership. In this study it linked a weakly predicted haplogroup Q with two R1a individuals.

Y chromosomes in two individuals were predicted as haplogroup T, and shared a SNP within the DYS385 $(TTTC)_n$ repeat array, producing the terminal motif TTTA $(TTTC)_2$; this variant was observed once by Huszar et al. (2018) in one of four haplogroup T1a individuals, and so this might be indicative of the same haplogroup in the Saudi individuals studied here. Falling as it does within one of the constitutively duplicated Y-STRs, there is also the possibility that a variant in one copy of DYS385 might be either lost or duplicated by gene conversion events

(Balaresque et al. 2014) involving the other, homogeneous, copy. This could lead to a patchy distribution of the variant within a haplogroup if it is spontaneously lost from different branches. The Y-STR structural motifs which could aid haplogroup prediction are listed in Table 6.2.

Table 6.2: Y-STR repeat structure motifs which are associated with specific Y-haplogroups

Locus	Structure	A	E1b1a	E1b1b	G	J1	Q	R1a	T
DYS19	(TCTA)12 CCTA (TCTA)2	1							
DYS389II	(TAGA)10 (CAGA)3 N48 (TAGA)11 (CAGA)6			6					
DYS389II	(TAGA)11 (CAGA)3 N48 (TAGA)10 (CAGA)6			3					
DYS389II	(TAGA)10 (CAGA)3 N48 (TAGA)12 (CAGA)6		1	5		1			
DYS389II	(TAGA)11 (CAGA)3 N48 (TAGA)11 (CAGA)6		1						
DYS389II	(TAGA)10 (CAGA)3 N48 (TAGA)13 (CAGA)6			1					
DYS389II	(TAGA)10 (CAGA)3 N48 (TAGA)11 (CAGA)7			1					
DYS389II	(TAGA)11 (CAGA)3 N48 (TAGA)12 (CAGA)6		1						
DYS533	(TATC)9 16281313-G (rs NA)				1				
DYS635	(TACA)2 (TAGA)2 (TACA)2 (TAGA)2 (TACA)2 (TAGA)4						1		
DYS635	(TAGA)9 (TACA)2 (TAGA)2 (TACA)2 (TAGA)2 (TACA)2 (TAGA)4							3	
DYS385	(TTTC)13 TTTA (TTTC)2								2
DYS390	(TAGA)4 CAGA (TAGA)5 TGA(TAGA)6 (CAGA)8							3	
DYS481	[CE21]-GRCh38-ChrY-8558313-8558408 (CTG)2 (CTT)20				3				
Total Count		1	3	16	3	49	1	3	2

Haplogroup J1, the commonest in this study, is characterised by the possession of an intermediate allele at DYS458. This results from another intra-repeat polymorphism, a partial deletion of an internal repeat which changes the array from a single $(GAAA)_n$ block to $(GAAA)_nAA(GAAA)_2$. Unfortunately this locus is not analysed in the Signature DNA Prep Kit, but J1 is easily identified by the aforementioned L255 SNP flanking the DYS438 repeat array, a SNP marker amplified but not accessible to analysis in all Y-STR CE multiplexes. Haplogroup J2, the second most common lineage in the Arabian Peninsula, does not have a characteristic repeat structure in the MPS kit tested here and so assignment to this lineage cannot be confirmed by sequence data alone; however, straightforward

allele length (as determined by CE) clearly resolves J2 from its sister group J1, e.g. at DYS438, where all eight J2 individuals have 9 repeats whilst all 49 J1 individuals have 10 or more repeats. For reference, all Y-STR sequences are listed in Appendix, Table 9.16 along with the predicted haplogroup of the bearer.

6.3.6 Forensic statistics and population comparisons

Among Y-STR loci the greatest genetic diversity (GD) in terms of allele length was seen at DYS481 and DYS612 (both 0.80); however, when sequence variants were included the complex loci DYS389II and DYF387S2 showed higher diversity at 0.849 and 0.875 respectively (Figure 6.1b). At the other end of the scale DYS392, a simple $(ATA)_n$ repeat locus, had the lowest GD at 0.192 and was not subdivided by sequencing. For the X-STRs, DXS10135 showed the highest GD (0.945) which increased to 0.968 when sequence variation was included, while DXS7423 had the lowest GD (0.659) and showed no improvement through sequencing. Haplotype diversity for LG1 (0.981) was increased to 0.987 when repeat region sequence was taken into account and for LG2 the value increased from 0.959 to 0.964 (repeat region sequence only) to 0.967 (including flanking SNPs) provides forensic statistical parameters for individual loci and for the two linkage groups. Discrimination capacity for the 24 Y-STR haplotype was 98.9%, with only one incidence of a shared haplotype seen in the dataset. The combined power of discrimination (CPD) for the X-STRs in males increased from 0.99990 when considering only length variation to 0.99995 once sequence variation was included, and from 0.99986 to 0.99990 when considering the LG CPD. Forensic statistics for both STRs are listed (Appendix, Table 9.17).

6.3.7 Population comparisons based on X-STR data

In Chapter 4 it was shown that the sample of 89 UK-recruited males was representative of Saudi Arabian autosomal diversity, by comparison with a much larger set of samples. Before analysing the X-STR data in the same samples, it is worth asking if they are also likely to reflect a fair representation of X diversity. Because the pattern of X-chromosomal diversity is relatively heavily influenced by the history and behaviours of females, an analysis of maternally inherited mtDNA

may be informative here. Sanger sequences of the mitochondrial DNA control region were determined in 87 males and used to predict mitochondrial haplogroups using HaploGrep 2.0. Comparison with previously published frequencies in a large sample from Saudi Arabia (Abu-Amero et al. 2008) as shown in Figure 6.6 revealed no significant difference (Exact p-value = 0.24036). Haplogroup predictions are summarised in Appendix, Table 9.18, and are typical of those that might be expected given Saudi Arabia's geographic location and the origins of its populace. These findings therefore give no reason to suspect that X-STR data in the current sample are biased by an unrepresentative matrilineal history.

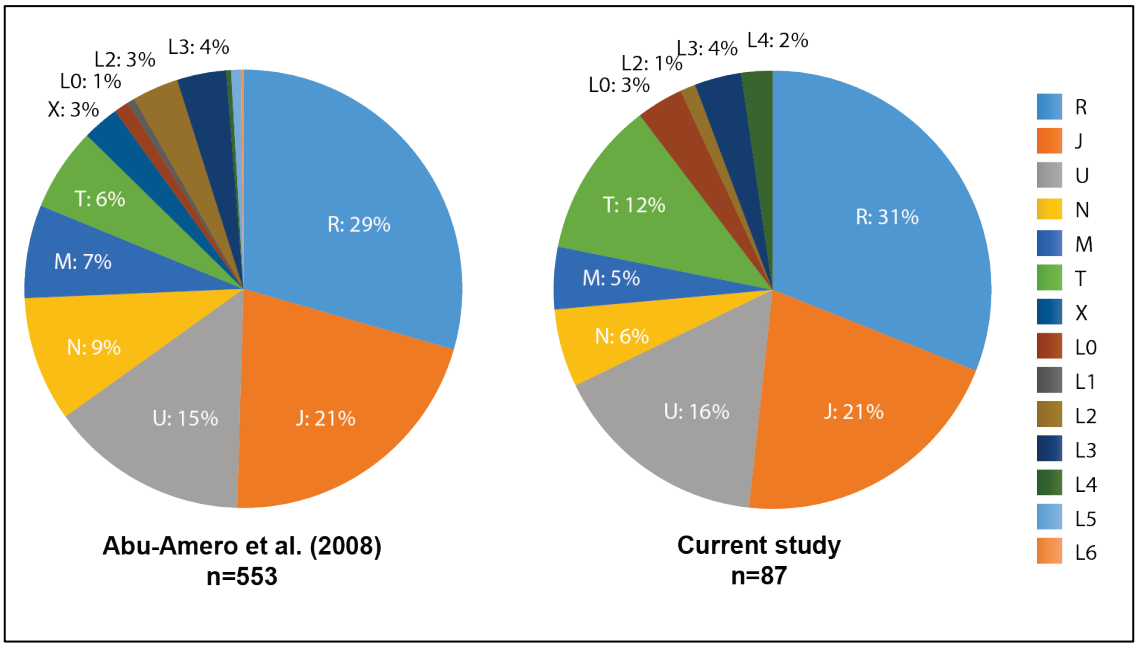


Figure 6.6: Saudi Arabian Mitochondrial haplogroup haplogroups. a) Haplogroup frequencies in current study b) Previously published Saudi Arabian data (Abu-Amero et al. 2008)

Because there are very little published data on the sequence structures of X- and Y-STR alleles from the Middle East, comparisons were made where possible at the CE level. For Y-STRs a high degree of differentiation has been shown using 597 samples spanning the five regions of Saudi Arabia (Khubrani et al. 2018). The 89 samples collected in the UK and available for X-STR typing were too few to test for differences between the five regions, but could be compared with neighbouring countries and

relevant ethnic groups including Bedouin, Palestinian, Mozabite (Phillips et al. 2018), Iraqi (Poulsen et al. 2016), UAE (Almarri and Lootah 2018), Egyptian (Elakkary et al. 2014) and Algerian (Bekada et al. 2010), as well as European, Asian, African and admixed populations (Almarri and Lootah 2018) to provide a wider context. MDS plots based on F_{ST} values (Figure 6.7) show the Saudi dataset to be closest to the HGDP Bedouin and place the Saudis along with other Middle Eastern populations between European, African and South Asian populations, reflecting their geographic position.

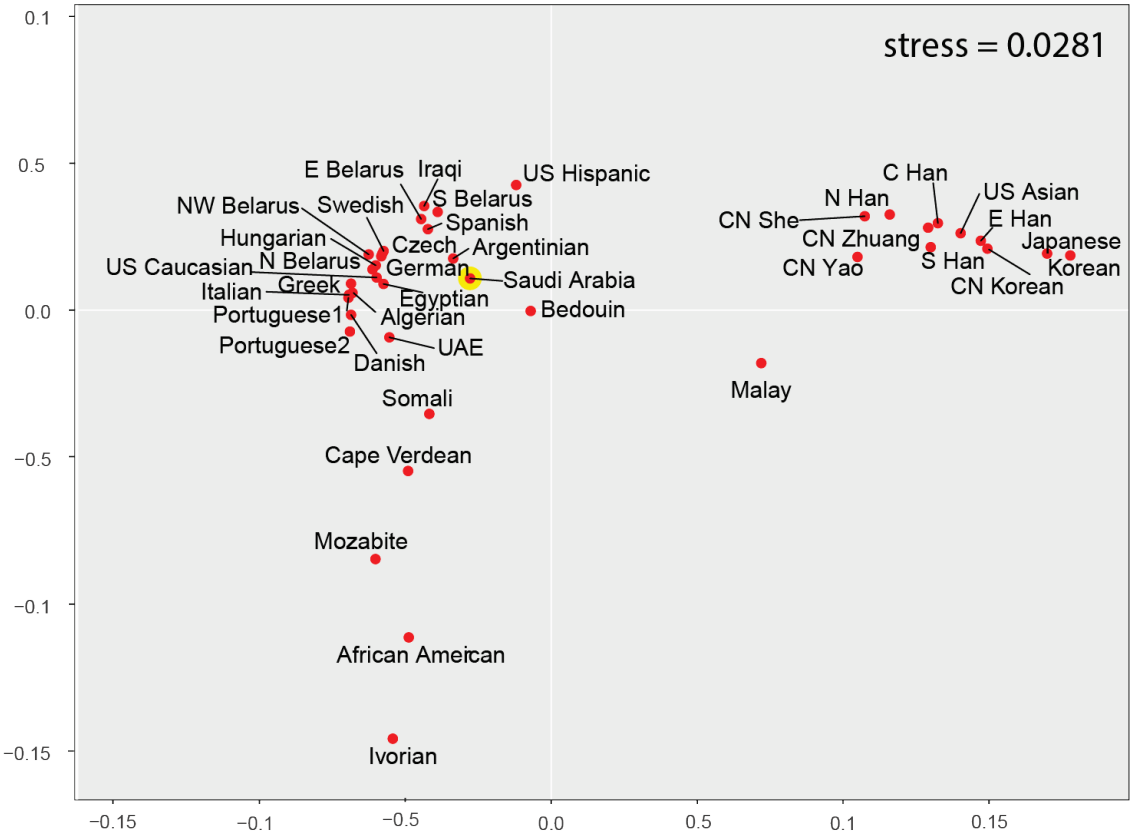


Figure 6.7: MDS plots based on F_{ST} values for X-STRs data. Comparative data are from (Almarri and Lootah 2018; Bekada et al. 2010; Elakkary et al. 2014; Phillips et al. 2018; Poulsen et al. 2016). The Saudi sample from the current study is highlighted with a yellow ring. CN [Chinese], S [South], N [North], W [West], E [East], US [United States]

Average diversity at the six X-STRs was calculated for Saudi Arabia, the UAE and those HGDP populations for which at least 15 males had been typed (See Figure 6.8). Saudi Arabia along with other Middle Eastern populations showed intermediate levels of genetic diversity. Overall diversity was highest among the Biaka Pygmies reflecting the great genetic diversity retained in Africa, while lower values were seen among the Japanese as previously noted (Arbiza et al. 2014) and particularly so in the Kalash which show signs of an earlier genetic bottleneck (Ayub et al. 2015) and may be further influenced by their unusual marriage practices which allow women freedom to divorce and remarry.

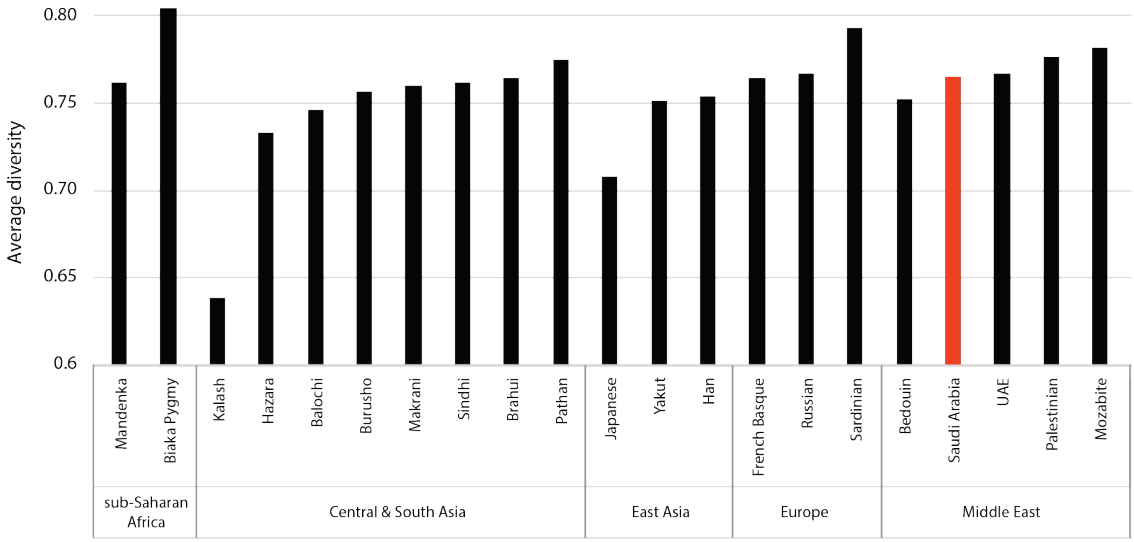


Figure 6.8: Average diversity at the six X-STRs. Comparative data are from HGDP (Phillips et al. 2018), UAE (Almarri and Lootah 2018). Data from the current study are highlighted in red.

6.4 Discussion

Analysis in Chapters 3 and 4 of the Saudi Arabian population has shown that the five regions of the country are strongly differentiated with respect to Y-STR haplotype frequency (Khubrani et al. 2018), with weaker structure seen at autosomal loci (Khubrani et al. 2019b). This reflects the patrilineal nature of Saudi tribal society which has a profound influence on settlement and movement within the country: tribal names are strongly associated with certain regions in which related males are concentrated. This is particularly true of central axis of Northern, Central and Southern regions which are more socially conservative. The Y-STR dataset obtained from Saudi students studying in the UK (Khubrani et al. 2018) was more closely aligned with the haplotype frequencies of the Eastern and Western regions which have historically been more open to migration and less conservative. In the Saudi-resident population, differentiation in autosomal markers showed a different geographic pattern suggested to be mediated by the movement of women (Khubrani et al. 2019b). It was therefore of interest to establish whether the genetic makeup of the UK-recruited Saudi sample matched that of the resident Saudi population at other markers. Since there are no published data on Saudi X-chromosomes, Sanger sequencing of the mitochondrial control region was used to compare the 89 males with existing mitochondrial data. The very close correspondence in terms of haplogroup frequency (F_{ST} < 0.0001, p= 0.6728) suggests that the UK-recruited males are a broadly representative dataset. This concurs with previous comparisons of autosomal STR allele length frequencies between the 89 UK Saudis and 509 indigenous Saudi Arabian males (Chapter 4;(Khubrani et al. 2019a)). Unfortunately, it was not possible to make sequence-level comparisons with the large Saudi-resident male dataset.

Sequencing revealed a considerable increase in the number of detectable alleles at some loci, primarily those with complex structures comprising different repeat array blocks, each of which could vary in the number of tandem repeats. While the number of allelic variants was increased the dataset was already characterised at the CE level by unique X-STR haplotypes and just one pair of shared Y-STR haplotypes that remained identical even after sequencing. One benefit of the

sequencing approach was the identification of haplogroup-defining SNPs and repeat motifs within the STR amplicons. Haplogroups defined by SNPs can often be reliably predicted using CE data on multiple Y-STRs (Khubrani et al. 2018), because the Y-STR allele lengths are closely correlated with Y-SNP haplogroups having diversified from the Y-STR haplotype of the male who carried the initial Y-SNP mutation (de Knijff 2000). In some cases, such as the link between J1 and DYS458 .2 intermediate alleles resulting from a 2-bp indel, there is a strong correlation with a single locus even using CE data alone. More generally, the characteristic tendency of alleles in restricted size ranges to be linked to certain haplogroups could aid mixture deconvolution if, for example, the observed structures indicated that there was both a J1 and E-M2 male present in the mixture. Haplogroup prediction is not without error as the first man to bear a haplogroup-defining SNP probably shared his entire Y-STR haplotype with his father and underived brothers, but lineage sorting over time eliminates most of those ancestral profiles and descendants of both the ancestral and derived variant bearing males diversify through the stepwise acquisition of several *de novo* mutations becoming distinguished by suites of variants that can be used to accurately classify their extant descendants. The increasing availability of STR sequence data from more Y-STR loci and a wide range of populations, combined with a greater use of Y-SNP typing is expected to provide the resources necessary to improve haplogroup prediction tools by establishing the frequency with which each allele length and structure is associated with unique event polymorphisms, and by filling the knowledge gaps relating to haplogroups from under-sampled populations.

A number of allele structures were detected which are not currently included in the STRait Razor v3.0 default database, indicating that they occur at low frequency in the better studied European, American and East Asian populations. It is likely that some of these variants may reach high frequencies in certain localities due to the concentration of close relatives within tribal communities. While these globally rare alleles can be very useful for elucidating biogeographic ancestry, they can present forensic interpretation issues when their structures differ from the norm, as the software used to call alleles requires landmarks within the sequence to establish how the sequence length should be designated. Rare variants which disrupt these

"anchor" sequences can result in both mis-designation of allele length for comparison against CE data or even failure to call the allele at all as in the case of bioinformatic nulls. Further population surveys and collaboration with software developers is required to establish appropriate nomenclature for the loci and rules on allele designation.

Whilst it is apparent that the UK-recruited Saudi dataset has a mitochondrial and autosomal (though not Y-chromosomal (Khubrani et al. 2018)) composition that is broadly representative of the country, it was necessary to rely on comparison with neighbouring populations for which X-STR data are available to ask whether the same is true for the X-chromosome. Overall the Saudi dataset shows an intermediate level of X-STR diversity, and MDS analysis places it between European, African and Asian populations, as would be expected given its geographic location. It is likely that larger surveys are required to truly capture diversity at these under-studied markers.

In conclusion, massively parallel sequencing of X- and Y-STRs within a Saudi population has demonstrated a significant increase in allele diversity over CE methods which, combined with the convenience of performing the tests at the same time as autosomal DNA profiling and reduced sample consumption, suggests it could play a significant role in challenging cases.

Chapter 7 General Discussion and Future Work

This thesis has used analysis of forensic markers, analysed both conventionally and by massively parallel sequencing, to characterise the genetic diversity of samples from Saudi Arabia. This has provided useful forensic reference data, and revealed evidence of geographical and social structure in the population, and of the genetic impact of consanguinity.

In Chapter 3, the 27-STR Yfiler® Plus kit was used to generate haplotypes in 597 unrelated Saudi males, classified into five geographical sub-regions. Yfiler® Plus provided a good discrimination capacity of 95.3%, which was greatly reduced (74.7%) when considering the reduced Yfiler® set of 17 Y-STRs, justifying the use of the expanded marker set in this population. The five geographical divisions were strikingly different in haplotype composition, with low diversity and similar compositions in the Central and Northern regions, and high diversity and similar haplotype compositions in the East and West. These patterns probably reflect the geographical isolation of the desert heartland of the peninsula, and the proximity to external influences of the Eastern and Western areas. Haplogroups predicted from Y-STR haplotypes showed the predominance (71%) of haplogroup J1, which was significantly more common in Central, Northern and Southern groups than in East and West, and formed a star-like expansion cluster in a median-joining network with an estimated age of ~2800 years. Most of the 597 males were sampled within Saudi Arabia, but ~16% were sampled in the UK. Despite matching these two groups by home sub-region, they were significantly different in haplotype and predicted haplogroup constitutions. This suggests social structure influencing the probability of leaving Saudi Arabia to study abroad, correlated with different Y-chromosome compositions. The UK-recruited sample is an inappropriate proxy for Saudi Arabia generally, and this demonstrates that caution is needed when considering expatriate groups as representative of a country of origin. This study shows the importance of geographical and social structuring that may affect the utility of forensic databases and the interpretation of Y-STR profiles.

In Chapter 4, variation in the 21 autosomal STRs detected by the GlobalFiler®

multiplex was investigated in a sample of 523 indigenous males from five geographic regions of Saudi Arabia. As in the Y chromosome study, this revealed population structure, but with a different geographical pattern. Allele frequencies for the entire dataset were found to be broadly similar to those determined in previous studies of Saudis, but significant differences were found among regions. Genetic distances were greatest between the Northern and Southern regions, while the West, Central and East appeared most similar to each other, and to previously published surveys. Differences between autosomal and Y-chromosomal patterns probably reflect genetic drift on the Y chromosome, exacerbated by prevalent patrilineal descent groups in different regions. Heterozygote deficiency was observed at nearly all loci in all regions, probably as a consequence of high levels of consanguineous marriage.

In Chapters 5 and 6, MPS was applied to the UK-recruited sample of 89 Saudi males, in the form of the Verogen ForenSeq™ DNA Signature Prep Kit. Chapter 5 described the autosomal data from these experiments, comprising MPS analysis of 27 autosomal STRs and 91 iiSNPs. This revealed sequence variation in the composition of complex STR arrays, and SNPs in their flanking regions. Similarly, additional polymorphic sites were observed within the amplicons of 37 of the 91 iiSNPs, forming up to six microhaplotypes per locus. Sequencing reduced both the STR-based and iiSNP-based random match probabilities, and the lack of significant LD between STRs and target iiSNPs allowed the two marker types to be combined using the product rule, yielding a RMP of 2.39E-73. As with the autosomal CE data, evidence of consanguinity was apparent from both marker types: the majority of STRs and iiSNPs showed fewer than expected heterozygotes, demonstrating an overall homozygote excess probably reflecting the high frequency of first-cousin marriages in Saudi Arabia. A global comparison with the HGDP panel showed that the Saudi sample was typical of Middle Eastern populations, with a higher level of inbreeding than is seen in most European, African and Central/South Asian populations, correlating with known patterns of endogamy. Given reduced levels of diversity within endogamous groups, the ability to combine both STRs and SNPs offers significant benefits in the analysis of forensic evidence in Saudi Arabia and the Middle East region more generally.

Chapter 6 reported the Y- and X-STR sequence data from the same MPS experiments. This work identified repeat sequence and flanking sequence variation in both X- and Y-STRs. Examination of Y-STR flanking sequences revealed defining SNPs, which together with repeat sequence motifs and length variants comprised haplogroup-specific features for 91% of the sample. Although the males in this study were recruited based on paternal ancestry, sequencing of the mtDNA control region showed a spectrum of mtDNA haplogroups not significantly different from previously published data. A population-level comparison of the Saudi Arabian X-STRs was made with a global sample, demonstrating affinity with previous data on other Middle Eastern populations.

7.1 Limitations and caveats of the study

Appropriate sample sizes in population studies are important if robust and useful conclusions are to be obtained; although the statistical methods of comparison used can account for small sample sizes in significance testing, nevertheless real differences could be missed. The 503 participants sampled within Saudi Arabia itself represent a good sample size compared to other studies in the field (Gusmao et al. 2017), and the per-region sub-samples are also similarly sized and reasonable. Recruitment of Saudi males in the UK was done to provide DNA samples that could be used for MPS analysis, and while the overall sample size (93) was respectable, the per-region sub-samples were unequally sized and small. This could be improved by further sampling.

The expectation from the UK-recruited sample was that it would be representative of Saudi diversity, and this proved to be the case for autosomal and mtDNA markers, but interestingly, not for the Y, and this revealed the influence of social structure linked to Y lineage, and therefore probably tribal origin. The influence of tribal identity on patterns of Y diversity generally is an interesting subject but unfortunately could not be explored in this thesis due to reasons of confidentiality. In practice, a forensic service might be interested in prediction of tribal identity (or surname) from a Y-STR haplotype, as has been suggested for the UK, for example (King et al. 2006).

Future work

The sampling strategy here, whether in Saudi Arabia itself or the UK, focused on indigenous individuals of Arabic ancestry. However, around 37% of the population (www.stats.gov.sa) is comprised of groups that originate outside Saudi Arabia, and these might be expected to carry their own haplotypes, genotypes and allele frequencies. Studies of such groups are necessary if a fully representative reference database for forensic interpretation is to be constructed. Future studies should therefore include such samples and generally increase sample sizes, and also improve sampling for the Middle East generally - a part of the world that has been understudied compared to other regions.

The Y chromosome analysis performed here highlighted the relatively low genetic diversity characteristic of a patrilineal society that has expanded recently. Although the 27 Y-STRs of Yfiler® Plus perform well in discrimination, there are still identical haplotypes, so there is scope for usefully applying a greater number of RM Y-STRs in this population. In addition, haplogroup prediction indicates a very high frequency of haplogroup J1. At the SNP level, it is likely that this lineage could be subdivided by SNPs that could show sub-lineages that are region- or tribe-specific, which might have forensic application.

The ideal way to analyse Y diversity is arguably to sequence the 'callable' region of the MSY (Jobling and Tyler-Smith 2017). This is done commercially through the 'BigY' test offered by FamilyTreeDNA, and a project has been initiated to harvest (with informed consent) the 10-Mb BigY sequences of hundreds of men from Saudi Arabia and other Middle Eastern nations to form a high-resolution tree that would allow haplogroups to be dissected and tribal and regional associations to be defined. Many customers of FamilyTreeDNA also have Y-STR data based on as many as 111 STRs, which is also being made available; this will allow a very high-resolution comparison of STR- and SNP-based haplotypes, and permit haplogroup prediction with better resolution. The whole dataset should illuminate the timing and dynamics of past male-mediated expansions in the region (Batini and Jobling 2017).

The findings here provide strong evidence for an effect of patrilineal descent groups, in the diversity of the Y chromosome. More work could be done on the X chromosome and mtDNA to support and understand the history of sex-biased behaviours. Such work is ideally sequencing-based (including whole mitogenomes and large segments of the Y and autosomes), since this allows unbiased comparisons of different parts of the genome at the nucleotide diversity level (Wilson Sayres et al. 2014).

Sequencing whole genomes in Saudi Arabia would provide further insights into population history, and in particular the extent and influence of consanguinity. This is likely to arise through medical studies, which aim to produce a catalogue of normal genetic variation and the allele frequency spectrum in the Saudi population, valuable for subsequent medical genomic studies of individuals with specific disorders. Much work has already been done at the level of exome sequencing, which reveals the prevalence of autozygous mutations in individuals from consanguineous pedigrees (Monies et al. 2019). Whole genome sequences could then be integrated into wider datasets such as those from the 1000 Genomes Project (1000 Genomes Project Consortium et al. 2015) and HGDP panel (Bergström et al. 2019), for a greater understanding of global genome variation and the role of the Middle East in this, and in the history and diversity of humans.

Chapter 8 References

1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR (2015) A global reference for human genetic variation. Nature 526: 68-74.

1000 Genomes Project Consortium, Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, Gibbs RA, Hurles ME, McVean GA (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061-1073.

Abu-Amero KK, Gonzalez AM, Larruga JM, Bosley TM, Cabrera VM (2007) Eurasian and African mitochondrial DNA influences in the Saudi Arabian population. BioMed Central Evolutionary Biology 7: 32.

Abu-Amero KK, Hellani A, Gonzalez AM, Larruga JM, Cabrera VM, Underhill PA (2009) Saudi Arabian Y-Chromosome diversity and its relationship with nearby regions. BioMed Central Genetics 10: 59.

Abu-Amero KK, Larruga JM, Cabrera VM, Gonzalez AM (2008) Mitochondrial DNA structure in the Arabian Peninsula. BioMed Central Evolutionary Biology 8: 45.

Abuidrees AS, Ishaq MJ, En Pu C, Alhamad NA, Alnafea HA, Almehaizea AM (2014) A Globally used 15 Short Tandem Repeats (STR) Loci in Forensic Human Identification, with their Allele Frequencies and Statistical Values in the Population of Bahrain. Arab Gulf Journal of Scientific Research 32.

Al-Abdulkareem AA, Ballal SG (1998) Consanguineous marriage in an urban area of Saudi Arabia: rates and adverse health effects on the offspring. Journal of Community Health 23: 75-83.

Al-Eitan LN, Tubaishat RR (2018) Evaluation of forensic genetic efficiency parameters of 22 autosomal STR markers (PowerPlex® Fusion system) in a population sample of Arab descent from Jordan. Australian Journal of Forensic Sciences 50: 97-109.

Al-Gazali L, Hamamy H, Al-Arrayad S (2006) Genetic disorders in the Arab world. British Medical Journal (Clinical research ed.) 333: 831-834.

Al-Hathloul S, Edadan N (1993) Evolution of settlement pattern in Saudi Arabia: A historical analysis. Habitat International 17: 31-46.

Al-Rasheed M (2010) A History of Saudi Arabia. Cambridge University Press, Cambridge

Alenizi M, Ge J, Ismael S, Al-Enezi H, Al-Awadhi A, Al-Duaij W, Al-Saleh B, Ghulloom Z, Budowle B (2013) Population genetic analyses of 15 STR loci from seven forensically-relevant populations residing in the state of Kuwait. Forensic

Science International: Genetics 7: e106-e107.

Alenizi M, Ge J, Salih A, Alenizi H, Al jabber J, Ziab J, Al harbi E, Isameal S, Budowle B (2014) Population data on 25 autosomal STRs for 500 unrelated Kuwaitis. Forensic Science International: Genetics 12: 126-127.

Alghafri R, Goodwin W, Ralf A, Kayser M, Hadi S (2015) A novel multiplex assay for simultaneously analysing 13 rapidly mutating Y-STRs. Forensic Science International: Genetics 17: 91-98.

Ali Alhmoudi O, Jones RJ, Tay GK, Alsafar H, Hadi S (2015) Population genetics data for 21 autosomal STR loci for United Arab Emirates (UAE) population using next generation multiplex STR kit. Forensic Science International: Genetics 19: 190-191.

Almalki N, Chow HY, Sharma V, Hart K, Siegel D, Wurmbach E (2017) Systematic assessment of the performance of Illumina's MiSeq FGx™ forensic genomics system. Electrophoresis 38: 846-854.

Almarri MA, Lootah RA (2018) Allelic and haplotype diversity of 12 X-STRs in the United Arab Emirates. Forensic Science International: Genetics 33: e4-e6.

Almohammed E, Zgonjanin D, Iyengar A, Ballard D, Devesse L, Sibte H (2017) A study of degraded skeletal samples using ForenSeq DNA Signature™ Kit. Forensic Science International: Genetics Supplement Series 6: e410-e412.

Alsafiah HM, Goodwin WH, Hadi S, Alshaikhi MA, Wepeba P-P (2017) Population genetic data for 21 autosomal STR loci for the Saudi Arabian population using the GlobalFiler® PCR amplification kit. Forensic Science International: Genetics 31: e59-e61.

Alshamali F, Alkhayat AQ, Budowle B, Watson ND (2005) STR population diversity in nine ethnic populations living in Dubai. Forensic Science International 152: 267-279.

Alshamali F, Pereira L, Budowle B, Poloni ES, Currat M (2009) Local population structure in Arabian Peninsula revealed by Y-STR diversity. Human Heredity 68: 45-54.

Anand R (1986) Pulsed field gel electrophoresis: a technique for fractionating large DNA molecules. Trends in Genetics 2: 278-283.

Andersen MM, Balding DJ (2017) How convincing is a matching Y-chromosome profile? PLoS Genetics 13: e1007028.

Anderson S, Bankier AT, Barrell GB, de Bruijn MHL, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJH, Staden R, Young IG (1981) Sequence and organisation of the human mitochondrial genome. Nature 290: 457-465.

Andrews S (2010) FastQC: FastQC: a quality control tool for high throughput sequence data [online] Available at https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Ankel-Simons F, Cummins JM (1996) Misconceptions about mitochondria and mammalian fertilization: implications for theories on human evolution. Proceedings of the National Academy of Sciences of the United States of America 93: 13859-13863.

Applied Biosystems (2016) Yfiler™ Plus PCR Amplication Kit User Guide [online] Available at https://assets.thermofisher.com/TFS-Assets/LSG/manuals/4485610_YfilerPlus_UG.pdf. Life Technologies, Carlsbad, CA

Arbiza L, Gottipati S, Siepel A, Keinan A (2014) Contrasting X-linked and autosomal diversity across 14 human populations. American Journal of Human Genetics 94: 827-844.

Athey TW (2005) Haplogroup prediction from Y-STR values using an allele frequency approach. Journal of Genetic Genealogy 1: 1-7.

Athey TW (2006) Haplogroup prediction from Y-STR values using a Bayesian-allele-frequency approach. Journal of Genetic Genealogy 2: 34-39.

Athey W (2011) Comments on the article, "Software for Y haplogroup predictions, a word of caution". International Journal of Legal Medicine 125: 901-903.

Ayub Q, Mezzavilla M, Pagani L, Haber M, Mohyuddin A, Khaliq S, Mehdi Syed Q, Tyler-Smith C (2015) The Kalash Genetic Isolate: Ancient Divergence, Drift, and Selection. American Journal of Human Genetics 96: 775-783.

Balanovsky O, Zhabagin M, Agdzhoyan A, Chukhryaeva M, Zaporozhchenko V, Utevska O, Highnam G, Sabitov Z, Greenspan E, Dibirova K, Skhalyakho R, Kuznetsova M, Koshel S, Yusupov Y, Nymadawa P, Zhumadilov Z, Pocheshkhova E, Haber M, Zalloua PA, Yepiskoposyan L, Dybo A, Tyler-Smith C, Balanovska E (2015) Deep phylogenetic analysis of haplogroup G1 provides estimates of SNP and STR mutation rates on the human Y-chromosome and reveals migrations of Iranic speakers. PLoS One 10: e0122968.

Balaresque P, Bowden GR, Parkin EJ, Omran GA, Heyer E, Quintana-Murci L, Roewer L, Stoneking M, Nasidze I, Carvalho-Silva DR, Tyler-Smith C, de Knijff P, Jobling MA (2008) Dynamic nature of the proximal AZFc region of the human Y chromosome: multiple independent deletion and duplication events revealed by microsatellite analysis. Human Mutation 29: 1171-1180.

Balaresque P, King TE, Parkin EJ, Heyer E, Carvalho-Silva D, Kraaijenbrink T, de Knijff P, Tyler-Smith C, Jobling MA (2014) Gene conversion violates the stepwise mutation model for microsatellites in Y-chromosomal palindromic repeats. Human Mutation 35: 609-617.

Ballantyne KN, Goedbloed M, Fang R, Schaap O, Lao O, Wollstein A, Choi Y, van Duijn K, Vermeulen M, Brauer S, Decorte R, Poetsch M, von Wurmb-Schwark N, de Knijff P, Labuda D, Vezina H, Knoblauch H, Lessig R, Roewer L, Ploski R, Dobosz T, Henke L, Henke J, Furtado MR, Kayser M (2010) Mutability of Y-chromosomal microsatellites: rates, characteristics, molecular bases, and forensic implications. American Journal of Human Genetics 87: 341-353.

Ballantyne KN, Keerl V, Wollstein A, Choi Y, Zuniga SB, Ralf A, Vermeulen M, de Knijff P, Kayser M (2012) A new future of forensic Y-chromosome analysis: Rapidly mutating Y-STRs for differentiating male relatives and paternal lineages. Forensic Science International: Genetics 6: 208-218.

Ballantyne KN, Ralf A, Aboukhalid R, Achakzai NM, Anjos MJ, Ayub Q, Balažic J, Ballantyne J, Ballard DJ, Berger B, Bobillo C, Bouabdellah M, Burri H, Capal T, Caratti S, Cárdenas J, Cartault F, Carvalho EF, Carvalho M, Cheng B, Coble MD, Comas D, Corach D, D'Amato ME, Davison S, de Knijff P, De Ungria MCA, Decorte R, Dobosz T, Dupuy BM, Elmrghni S, Gliwiński M, Gomes SC, Grol L, Haas C, Hanson E, Henke J, Henke L, Herrera-Rodríguez F, Hill CR, Holmlund G, Honda K, Immel UD, Inokuchi S, Jobling MA, Kaddura M, Kim JS, Kim SH, Kim W, King TE, Klausriegler E, Kling D, Kovačević L, Kovatsi L, Krajewski P, Kravchenko S, Larmuseau MHD, Lee EY, Lessig R, Livshits LA, Marjanović D, Minarik M, Mizuno N, Moreira H, Morling N, Mukherjee M, Munier P, Nagaraju J, Neuhuber F, Nie S, Nilasitsataporn P, Nishi T, Oh HH, Olofsson J, Onofri V, Palo JU, Pamjav H, Parson W, Petlach M, Phillips C, Ploski R, Prasad SPR, Primorac D, Purnomo GA, Purps J, Rangel-Villalobos H, Reogonekbała K, Rerkamnuaychoke B, Gonzalez DR, Robino C, Roewer L, Rosa A, Sajantila A, Sala A, Salvador JM, Sanz P, Schmitt C, Sharma AK, Silva DA, Shin KJ, et al. (2014) Toward Male Individualization with Rapidly Mutating Y-Chromosomal Short Tandem Repeats. Human Mutation 35: 1021-1032.

Bandelt H-J, Forster P, Röhl A (1999) Median-joining networks for inferring intraspecific phylogenies. Molecular Biology and Evolution 16: 37-48.

Batini C, Hallast P, Zadik D, Delser PM, Benazzo A, Ghirotto S, Arroyo-Pardo E, Cavalleri GL, de Knijff P, Dupuy BM, Eriksen HA, King TE, de Munain AL, Lopez-Parra AM, Loutradis A, Milasin J, Novelletto A, Pamjav H, Sajantila A, Tolun A, Winney B, Jobling MA (2015) Large-scale recent expansion of European patrilineages shown by population resequencing. Nature Communications 6: 7152.

Batini C, Jobling MA (2017) Detecting past male-mediated expansions using the Y chromosome. Human Genetics 136: 547-557.

Bekada A, Benhamamouch S, Boudjema A, Fodil M, Menegon S, Torre C, Robino C (2010) Analysis of 21 X-chromosomal STRs in an Algerian population sample. International Journal of Legal Medicine 124: 287-294.

Bergström A, McCarthy SA, Hui R, Almarri MA, Ayub Q, Danecek P, Chen Y, Felkel S, Hallast P, Kamm J (2019) Insights into human genetic variation and

population history from 929 diverse genomes. bioRxiv: 674986.

Bittles A (2007) A global overview on consanguinity. EGF [en línea].

Bittles AH (2001) Consanguinity and its relevance to clinical genetics. Clinical Genetics 60: 89-98.

Bittles AH, Black M (2010) Consanguinity, human evolution, and complex diseases. Proceedings of the National Academy of Sciences of the United States of America 107: 1779-1786.

Bodner M, Bastisch I, Butler JM, Fimmers R, Gill P, Gusmao L, Morling N, Phillips C, Prinz M, Schneider PM, Parson W (2016) 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 Science International: Genetics 24: 97-102.

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114-2120.

Børsting C, Morling N (2016) Chapter 22 - Genomic Applications in Forensic Medicine. In: Kumar D, Antonarakis S (eds) Medical and Health Genomics. Academic Press, Oxford, pp 295-309

Brandt-Casadevall C, Ben Dhiab M, Taroni F, Gehrig C, Dimo-Simonin N, Zemni M, Mangin P (2003) Tunisian population allele frequencies for 15 PCR-based loci. International Congress Series 1239: 113-116.

Brinkmann B, Klintschar M, Neuhuber F, Hühne J, Rolf B (1998) Mutation Rate in Human Microsatellites: Influence of the Structure and Length of the Tandem Repeat. American Journal of Human Genetics 62: 1408-1415.

Brown WM, George M, Wilson AC (1979) Rapid evolution of animal mitochondrial DNA. Proceedings of the National Academy of Sciences of the United States of America 76: 1967-1971.

Buckleton JS, Bright J-A, Taylor D (2016) Forensic DNA evidence interpretation. CRC press, Boca Raton

Budowle B, Van Daal A (2008) Forensically relevant SNP classes. BioTechniques 44: 603-610.

Buroker N, Bestwick R, Haight G, Magenis R, Litt M (1987) A hypervariable repeated sequence on human chromosome 1p36. Human Genetics 77: 175-181.

Butler J, Hill CR, Coble M (2012) Variability of new STR loci and kits in US population groups [online] Available at http://www.promega.co.uk/resources/profiles-in-dna/2012/variability-of-new-str-loci-and-kits-in-us-population-groups/.

Butler JM (2005) Forensic DNA typing: biology, technology, and genetics of STR markers. Elsevier, New York

Butler JM (2007) Short tandem repeat typing technologies used in human identity testing. BioTechniques 43: Sii-Sv.

Butler JM (2009) Fundamentals of forensic DNA typing. Academic Press, London

Butler JM (2012) Advanced Topics in Forensic DNA Typing: Methodology, 2nd edn. Elsevier, Amsterdam

Butler JM, Hill CR (2012) Biology and genetics of new autosomal STR loci useful for forensic DNA analysis. Forensic Science Review 24: 15-26.

Cadenas AM, Zhivotovsky LA, Cavalli-Sforza LL, Underhill PA, Herrera RJ (2008) Y-chromosome diversity characterizes the Gulf of Oman. European Journal of Human Genetics 16: 374-386.

Calafell F, Larmuseau MHD (2017) The Y chromosome as the most popular marker in genetic genealogy benefits interdisciplinary research. Human Genetics 136: 559-573.

Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, Bodmer J, Bodmer WF, Bonne-Tamir B, Cambon-Thomsen A, Chen Z, Chu JY, Carcassi C, Contu L, Du RF, Excoffier L, Ferrara GB, Friedlaender JS, Groot H, Gurwitz D, Jenkins T, Herrera RJ, Huang XY, Kidd J, Kidd KK, Langaney A, Lin AA, Mehdi SQ, Parham P, Piazza A, Pistillo MP, Qian YP, Shu QF, Xu JJ, Zhu S, Weber JL, Greely HT, Feldman MW, Thomas G, Dausset J, Cavalli-Sforza LL (2002) A human genome diversity cell line panel. Science 296: 261-262.

Caratti S, Turrina S, Ferrian M, Cosentino E, De Leo D (2015) MiSeq FGx sequencing system: A new platform for forensic genetics. Forensic Science International: Genetics Supplement Series 5: e98-e100.

Carracedo A, Butler JM, Gusmao L, Linacre A, Parson W, Roewer L, Schneider PM (2013) New guidelines for the publication of genetic population data. Forensic Science International: Genetics 7: 217-220.

Carracedo A, Butler JM, Gusmao L, Parson W, Roewer L, Schneider PM (2010) Publication of population data for forensic purposes. Forensic Science International: Genetics 4: 145-147.

Chakraborty R, Stivers DN, Su B, Zhong Y, Budowle B (1999) The utility of short tandem repeat loci beyond human identification: Implications for development of new DNA typing systems. Electrophoresis 20: 1682-1696.

Charlesworth B (2009) Effective population size and patterns of molecular evolution and variation. Nature Reviews Genetics 10: 195-205.

Chiaroni J, King RJ, Myres NM, Henn BM, Ducourneau A, Mitchell MJ, Boetsch G,

Sheikha I, Lin AA, Nik-Ahd M, Ahmad J, Lattanzi F, Herrera RJ, Ibrahim ME, Brody A, Semino O, Kivisild T, Underhill PA (2010) The emergence of Y-chromosome haplogroup J1e among Arabic-speaking populations. European Journal of Human Genetics 18: 348-353.

Churchill JD, Schmedes SE, King JL, Budowle B (2016) Evaluation of the Illumina® Beta Version ForenSeq™ DNA Signature Prep Kit for use in genetic profiling. Forensic Science International: Genetics 20: 20-29.

Consortium IH (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851.

de Knijff P (2000) Messages through bottlenecks: on the combined use of slow and fast evolving polymorphic markers on the human Y chromosome. American Journal of Human Genetics 67: 1055-1061.

de Knijff P (2019) From next generation sequencing to now generation sequencing in forensics. Forensic Science International: Genetics 38: 175-180.

de Knijff P, Kayser M, Caglià A, Corach D, Fretwell N, Gehrig C, Graziosi G, Heidorn F, Herrmann S, Herzog B, Hidding M, Honda K, Jobling M, Krawczak M, Leim K, Meuser S, Meyer E, Oesterreich W, Pandya A, Parson W, Penacino G, Perez-Lezaun A, Piccinini A, Prinz M, Schmitt C, Schneider PM, Szibor R, Teifel-Greding J, Weichhold G, Roewer L (1997) Chromosome Y microsatellites: population genetic and evolutionary aspects. International Journal of Legal Medicine 110: 134-140.

Desjardins P, Conklin D (2010) NanoDrop microvolume quantitation of nucleic acids. Journal of Visualized Experiments: e2565.

El-Hazmi M, Al-Swailem A, Warsy A, Al-Swailem A, Sulaimani R, Al-Meshari A (1995) Consanguinity among the Saudi Arabian population. Journal of Medical Genetics 32: 623-626.

El-Mouzan MI, Al-Salloum AA, Al-Herbish AS, Qurachi MM, Al-Omar AA (2007) Regional variations in the prevalence of consanguinity in Saudi Arabia. Saudi Medical Journal 28: 1881-1884.

El-Sibai M, Platt DE, Haber M, Xue Y, Youhanna SC, Wells RS, Izaabel H, Sanyoura MF, Harmanani H, Bonab MA, Behbehani J, Hashwa F, Tyler-Smith C, Zalloua PA (2009) Geographical structure of the Y-chromosomal genetic landscape of the Levant: a coastal-inland contrast. Annals of Human Genetics 73: 568-81.

Elakkary S, Hoffmeister-Ullerich S, Schulze C, Seif E, Sheta A, Hering S, Edelmann J, Augustin C (2014) Genetic polymorphisms of twelve X-STRs of the investigator Argus X-12 kit and additional six X-STR centromere region loci in an Egyptian population sample. Forensic Science International: Genetics 11: 26-30.

Emmerova B, Ehler E, Comas D, Votrubova J, Vanek D (2017) Comparison of Y-chromosomal haplogroup predictors. Forensic Science International: Genetics Supplement Series 6: e145-e147.

Excoffier L, Lischer HE (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Molecular Ecology Resources 10: 564-567.

Fan H, Chu J-Y (2007) A brief review of short tandem repeat mutation. Genomics, Proteomics & Bioinformatics 5: 7-14.

Farhan MM, Hadi S, Iyengar A, Goodwin W (2016) Population genetic data for 20 autosomal STR loci in an Iraqi Arab population: Application to the identification of human remains. Forensic Science International: Genetics 25: e10-e11.

Ferri G, Robino C, Alu M, Luiselli D, Tofanelli S, Caciagli L, Onofri C, Pelotti S, Di Gaetano C, Crobu F, Beduschi G, Capelli C (2008) Molecular characterisation and population genetics of the DYS458 .2 allelic variant. Forensic Science International: Genetics Supplement Series 1: 203–205.

Fordyce SL, Mogensen HS, Børsting C, Lagacé RE, Chang C-W, Rajagopalan N, Morling N (2015) Second-generation sequencing of forensic STRs using the Ion Torrent™ HID STR 10-plex and the Ion PGM™. Forensic Science International: Genetics 14: 132-140.

Galaty JG, Salzman PC (1981) Change and development in nomadic and pastoral societies. Brill, Leiden

Gettings KB, Borsuk LA, Ballard D, Bodner M, Budowle B, Devesse L, King J, Parson W, Phillips C, Vallone PM (2017) STRSeq: A catalog of sequence diversity at human identification Short Tandem Repeat loci. Forensic Science International: Genetics 31: 111-117.

Gettings KB, Kiesler KM, Faith SA, Montano E, Baker CH, Young BA, Guerrieri RA, Vallone PM (2016) Sequence variation of 22 autosomal STR loci detected by next generation sequencing. Forensic Science International: Genetics 21: 15-21.

Goedbloed M, Vermeulen M, Fang RN, Lembring M, Wollstein A, Ballantyne K, Lao O, Brauer S, Krüger C, Roewer L (2009) Comprehensive mutation analysis of 17 Y-chromosomal short tandem repeat polymorphisms included in the AmpFlSTR® Yfiler® PCR amplification kit. International Journal of Legal Medicine 123: 471-482.

Goldstein DB, Linares AR, Cavalli-Sforza LL, Feldman MW (1995a) An evaluation of genetic distances for use with microsatellite loci. Genetics 139: 463-471.

Goldstein DB, Linares AR, Cavalli-Sforza LL, Feldman MW (1995b) Genetic absolute dating based on microsatellites and the origin of modern humans.

Proceedings of the National Academy of Sciences of the United States of America 92: 6723-6727.

Goodwin S, McPherson JD, McCombie WR (2016) Coming of age: ten years of next-generation sequencing technologies. Nature Reviews Genetics 17: 333-351.

Goodwin W (2007) An introduction to forensic genetics. Wiley, Chichester

Goudet J (1995) FSTAT (Version 1.2): A Computer Program to Calculate F-Statistics. Journal of Heredity 86: 485–486.

Gouy A, Zieger M (2017) STRAF—A convenient online tool for STR data evaluation in forensic genetics. Forensic Science International: Genetics 30: 148-151.

Groucutt HS, Grün R, Zalmout IA, Drake NA, Armitage SJ, Candy I, Clark-Wilson R, Louys J, Breeze PS, Duval M (2018) Homo sapiens in Arabia by 85,000 years ago. Nature Ecology & Evolution 2: 800.

Groucutt HS, Petraglia MD (2012) The prehistory of the Arabian peninsula: deserts, dispersals, and demography. Evolutionary Anthropology 21: 113-25.

Groucutt HS, Scerri EM, Lewis L, Clark-Balzan L, Blinkhorn J, Jennings RP, Parton A, Petraglia MD (2015) Stone tool assemblages and models for the dispersal of Homo sapiens out of Africa. Quaternary International 382: 8-30.

Guo F (2017) Population genetic data for 12 X-STR loci in the Northern Han Chinese and StatsX package as tools for population statistics on X-STR. Forensic Science International: Genetics 26: e1-e8.

Guo F, Yu J, Zhang L, Li J (2017) Massively parallel sequencing of forensic STRs and SNPs using the Illumina® ForenSeq™ DNA Signature Prep Kit on the MiSeq FGx™ Forensic Genomics System. Forensic Science International: Genetics 31: 135-148.

Gusmao L, Butler JM, Linacre A, Parson W, Roewer L, Schneider PM, Carracedo A (2017) Revised guidelines for the publication of genetic population data. Forensic Science International: Genetics 30: 160-163.

Hadi S (2016) Analysis of Rapidly Mutating Y Chromosome Short Tandem Repeats (RM Y-STRs). In: Goodwin W (ed) Forensic DNA Typing Protocols. Springer New York, New York, pp 201-211

Hall TA BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT Nucleic acids symposium series 1999. [London]: Information Retrieval Ltd., c1979-c2000., pp 95-98

Hallast P, Balaresque P, Bowden GR, Ballereau SJ, Jobling MA (2013) Recombination dynamics of a human Y-chromosomal palindrome: rapid GC-biased gene conversion, multi-kilobase conversion tracts, and rare inversions. PLoS Genetics 9: e1003666.

Hallast P, Batini C, Zadik D, Maisano Delser P, Wetton JH, Arroyo-Pardo E, Cavalleri GL, de Knijff P, Destro Bisol G, Dupuy BM, Eriksen HA, Jorde LB, King TE, Larmuseau MH, Lopez de Munain A, Lopez-Parra AM, Loutradis A, Milasin J, Novelletto A, Pamjav H, Sajantila A, Schempp W, Sears M, Tolun A, Tyler-Smith C, Van Geystelen A, Watkins S, Winney B, Jobling MA (2015) The Y-chromosome tree bursts into leaf: 13,000 high-confidence SNPs covering the majority of known clades. Molecular Biology and Evolution 32: 661-673.

Hallenberg C, Simonsen B, Sanchez J, Morling N (2005) Y-chromosome STR haplotypes in Somalis. Forensic Science International 151: 317-321.

Hanson EK, Ballantyne J (2006) Comprehensive annotated STR physical map of the human Y chromosome: Forensic implications. Legal Medicine 8: 110-120.

Hartzell B, Graham K, McCord B (2003) Response of short tandem repeat systems to temperature and sizing methods. Forensic science international 133: 228-234.

Hedjazi A, Nikbakht A, Hosseini M, Hoseinzadeh A, Hosseini SMV (2013) Allele frequencies for 15 autosomal STR loci in Fars province population, southwest of Iran. Legal Medicine 15: 226-228.

Helgason A, Einarsson AW, Guethmundsdottir VB, Sigurethsson A, Gunnarsdottir ED, Jagadeesan A, Ebenesersdottir SS, Kong A, Stefansson K (2015) The Y-chromosome point mutation rate in humans. Nature Genetics 47: 453-457.

Hennessy LK, Mehendale N, Chear K, Jovanovich S, Williams S, Park C, Gangano S (2014) Developmental validation of the GlobalFiler® express kit, a 24-marker STR assay, on the RapidHIT® System. Forensic Science International: Genetics 13: 247-258.

Holland MM, Bonds RM, Holland CA, McElhoe JA (2019) Recovery of mtDNA from unfired metallic ammunition components with an assessment of sequence profile quality and DNA damage through MPS analysis. Forensic Science International: Genetics 39: 86-96.

Holland MM, Parsons TJ (1999) Mitochondrial DNA Sequence Analysis-Validation and Use for Forensic Casework. Forensic Science Review 11: 21-50.

Hourani A (1991) A. History if the Arab Peoples. Cambridge: Harvard University Press, Cambridge

Hughes JF, Skaletsky H, Brown LG, Pyntikova T, Graves T, Fulton RS, Dugan S, Ding Y, Buhay CJ, Kremitzki C, Wang Q, Shen H, Holder M, Villasana D, Nazareth LV, Cree A, Courtney L, Veizer J, Kotkiewicz H, Cho TJ, Koutseva N, Rozen S, Muzny DM, Warren WC, Gibbs RA, Wilson RK, Page DC (2012) Strict evolutionary conservation followed rapid gene loss on human and rhesus Y chromosomes. Nature 483: 82-86.

Huszar TI, Jobling MA, Wetton JH (2018) A phylogenetic framework facilitates Y-

STR variant discovery and classification via massively parallel sequencing. Forensic Science International: Genetics 35: 97-106.

Huszar TI, Wetton JH, Jobling MA (2019) Mitigating the effects of reference sequence bias in single-multiplex massively parallel sequencing of the mitochondrial DNA control region. Forensic Science International: Genetics 40: 9-17.

Iacovacci G, D'Atanasio E, Marini O, Coppa A, Sellitto D, Trombetta B, Berti A, Cruciani F (2017) Forensic data and microvariant sequence characterization of 27 Y-STR loci analyzed in four Eastern African countries. Forensic Science International: Genetics 27: 123-131.

Illumina (2015a) ForenSeq DNA Signature Prep Reference Guide [online] Available at https://verogen.com/wp-content/uploads/2018/08/ForenSeq-DNA-Prep-Guide-VD2018005-A.pdf.

Illumina (2015b) Targeted Next-Generation Sequencing for Forensic Genomics [online] Available at https://verogen.com/wp-content/uploads/2018/08/app_spotlight_forensics-1.pdf.

Iozzi S, Carboni I, Contini E, Pescucci C, Frusconi S, Nutini AL, Torricelli F, Ricci U (2015) Forensic genetics in NGS era: New frontiers for massively parallel typing. Forensic Science International: Genetics Supplement Series 5: e418-e419.

Jäger AC, Alvarez ML, Davis CP, Guzmán E, Han Y, Way L, Walichiewicz P, Silva D, Pham N, Caves G, Bruand J, Schlesinger F, Pond SJK, Varlaro J, Stephens KM, Holt CL (2017) Developmental validation of the MiSeq FGx Forensic Genomics System for Targeted Next Generation Sequencing in Forensic DNA Casework and Database Laboratories. Forensic Science International: Genetics 28: 52-70.

Jeffreys AJ, Neumann R, Wilson V (1990) Repeat unit sequence variation in minisatellites: a novel source of DNA polymorphism for studying variation and mutation by single molecule analysis. Cell 60: 473-485.

Jeffreys AJ, Wilson V, Thein SL (1985) Hypervariable 'minisatellite' regions in human DNA. Nature 314: 67-73.

Jennings RP, Singarayer J, Stone EJ, Krebs-Kanzow U, Khon V, Nisancioglu KH, Pfeiffer M, Zhang X, Parker A, Parton A (2015) The greening of Arabia: Multiple opportunities for human occupation of the Arabian Peninsula during the Late Pleistocene inferred from an ensemble of climate model simulations. Quaternary International 382: 181-199.

Jobling MA (2001) Y-chromosomal SNP haplotype diversity in forensic analysis. Forensic Science International 118: 158-162.

Jobling MA, Hollox EJ, Hurles ME, Kivisild T, Tyler-Smith C (2014) Human

Evolutionary Genetics, 2nd edn. Garland Science, New York and London

Jobling MA, Pandya A, Tyler-Smith C (1997) The Y chromosome in forensic analysis and paternity testing. International Journal of Legal Medicine 110: 118-124.

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

Jobling MA, Tyler-Smith C (2017) Human Y-chromosome variation in the genome-sequencing era. Nature Reviews Genetics 18: 485–497.

Johns L, Burton R, Thomson J Study to compare three commercial Y-STR testing kits International Congress Series 2006. Elsevier, pp 192-194

Jones RJ, Al Tayaare W, Tay GK, Alsafar H, Goodwin WH (2017) Population data for 21 autosomal short tandem repeat markers in the Arabic population of the United Arab Emirates. Forensic Science International: Genetics 28: e41-e42.

Jorde LB, Watkins WS, Bamshad MJ, Dixon ME, Ricker CE, Seielstad MT, Batzer MA (2000) The Distribution of Human Genetic Diversity: A Comparison of Mitochondrial, Autosomal, and Y-Chromosome Data. American Journal of Human Genetics 66: 979-988.

Just RS, Moreno LI, Smerick JB, Irwin JA (2017) Performance and concordance of the ForenSeq™ system for autosomal and Y chromosome short tandem repeat sequencing of reference-type specimens. Forensic Science International: Genetics 28: 1-9.

Karafet TM, Mendez FL, Meilerman M, Underhill PA, Zegura SL, Hammer MF (2008) New binary polymorphisms reshape and increase resolution of the human Y-chromosomal haplogroup tree. Genome Research 18: 830-838.

Kauppi L, May CA, Jeffreys AJ (2009) Analysis of meiotic recombination products from human sperm. Meiosis. Springer, Totowa, pp 323-355

Kayser M (2017) Forensic use of Y-chromosome DNA: a general overview. Human Genetics 136: 621-635.

Kayser M, Caglià A, Corach D, Fretwell N, Gehrig C, Graziosi G, Heidorn F, Herrmann S, Herzog B, Hidding M, Honda K, Jobling M, Krawczak M, Leim K, Meuser S, Meyer E, Oesterreich W, Pandya A, Parson W, Penacino G, Perez-Lezaun A, Piccinini A, Prinz M, Schmitt C, Schneider PM, Szibor R, Teifel-Greding J, Weichhold G, de Knijff P, Roewer L (1997) Evaluation of Y-chromosomal STRs: a multicenter study. International Journal of Legal Medicine 110: 125-133.

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

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

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

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D (2002) The human genome browser at UCSC. Genome Research 12: 996-1006.

Khubrani YM, Hallast P, Jobling MA, Wetton JH (2019a) Massively parallel sequencing of autosomal STRs and identity-informative SNPs highlights consanguinity in Saudi Arabia. Forensic Science International: Genetics: 102164.

Khubrani YM, Wetton JH, Jobling MA (2018) Extensive geographical and social structure in the paternal lineages of Saudi Arabia revealed by analysis of 27 Y-STRs. Forensic Science International: Genetics 33: 98-105.

Khubrani YM, Wetton JH, Jobling MA (2019b) Analysis of 21 autosomal STRs in Saudi Arabia reveals population structure and the influence of consanguinity. Forensic Science International: Genetics 39: 97-102.

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

King JL, Churchill JD, Novroski NMM, Zeng X, Warshauer DH, Seah L-H, Budowle B (2018) Increasing the discrimination power of ancestry- and identity-informative SNP loci within the ForenSeq™ DNA Signature Prep Kit. Forensic Science International: Genetics 36: 60-76.

King TE, Ballereau SJ, Schürer K, Jobling MA (2006) Genetic signatures of coancestry within surnames. Current Biology 16: 384-388.

King TE, Jobling MA (2009) What's in a name? Y chromosomes, surnames and the genetic genealogy revolution. Trends in Genetics 25: 351-360.

King TE, Parkin EJ, Swinfield G, Cruciani F, Scozzari R, Rosa A, Lim SK, Xue Y, Tyler-Smith C, Jobling MA (2007) Africans in Yorkshire? The deepest-rooting clade of the Y phylogeny within an English genealogy. European Journal of Human Genetics 15: 288-293.

Köcher S, Müller P, Berger B, Bodner M, Parson W, Roewer L, Willuweit S (2018) Inter-laboratory validation study of the ForenSeq™ DNA Signature Prep Kit. Forensic Science International: Genetics 36: 77-85.

Krausz C, Casamonti E (2017) Spermatogenic failure and the Y chromosome. Human Genetics 136: 637-655.

Kruskal JB (1964) Multidimensional scaling by optimizing a goodness of fit test to a nonmetric hypothesis. Psychometrika 19: 1-27.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860-921.

Lang Y, Guo F, Niu Q (2019) StatsX v2.0: the interactive graphical software for population statistics on X-STR. International Journal of Legal Medicine 133: 39-44.

Larmuseau MH, Vanderheyden N, Van Geystelen A, van Oven M, de Knijff P, Decorte R (2014) Recent radiation within Y-chromosomal haplogroup R-M269 resulted in high Y-STR haplotype resemblance. Annals of Human Genetics 78: 92-103.

Leutenegger A-L, Sahbatou M, Gazal S, Cann H, Génin E (2011) Consanguinity around the world: what do the genomic data of the HGDP-CEPH diversity panel tell us? European Journal of Human Genetics 19: 583-587.

Lewontin RC, Hartl DL (1991) Population genetics in forensic DNA typing. Science 254: 1745-1750.

Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27: 2987-2993.

Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:1303.3997.

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, Myers RM (2008) Worldwide human relationships inferred from genome-wide patterns of variation. Science 319: 1100-1104.

Li R, Li H, Peng D, Hao B, Wang Z, Huang E, Wu R, Sun H (2019) Improved pairwise kinship analysis using massively parallel sequencing. Forensic Science International: Genetics 38: 77-85.

Liu Y-Y, Harbison S (2018) A review of bioinformatic methods for forensic DNA analyses. Forensic Science International: Genetics 33: 117-128.

Maan AA, Eales J, Akbarov A, Rowland J, Xu X, Jobling MA, Charchar FJ, Tomaszewski M (2017) The Y chromosome: a blueprint for men's health? European Journal of Human Genetics 25: 1181-1188.

Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, Zhao M, Chennagiri N, Nordenfelt S, Tandon A, Skoglund P, Lazaridis I, Sankararaman S, Fu Q, Rohland N, Renaud G, Erlich Y, Willems T, Gallo C, Spence JP, Song YS, Poletti G, Balloux F, van Driem G, de Knijff P, Romero IG, Jha AR, Behar DM, Bravi CM, Capelli C, Hervig T, Moreno-Estrada A, Posukh OL, Balanovska E, Balanovsky O, Karachanak-Yankova S, Sahakyan H, Toncheva D, Yepiskoposyan L, Tyler-Smith C, Xue Y, Abdullah MS, Ruiz-Linares A, Beall CM, Di Rienzo A, Jeong C, Starikovskaya EB, Metspalu E, Parik J, Villems R, Henn BM, Hodoglugil U, Mahley R, Sajantila A, Stamatoyannopoulos G, Wee JT, Khusainova R, Khusnutdinova E, Litvinov S, Ayodo G, Comas D, Hammer MF, Kivisild T, Klitz W, Winkler CA, Labuda D, Bamshad M, Jorde LB, Tishkoff SA, Watkins WS, Metspalu M, Dryomov S, Sukernik R, Singh L, Thangaraj K, Paabo S, Kelso J, Patterson N, Reich D (2016) The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature 538: 201-206.

Manni F, Leonardi P, Barakat A, Rouba H, Heyer E, Klintschar M, McElreavey K, Quintana-Murci L (2002) Y-chromosome analysis in Egypt suggests a genetic regional continuity in Northeastern Africa. Human Biology 74: 645-658.

Mardis ER (2013) Next-generation sequencing platforms. Annual Review of Analytical Chemistry 6: 287-303.

Memish ZA, Shibl AM, Kambal AM, Ohaly YA, Ishaq A, Livermore DM (2012) Antimicrobial resistance among non-fermenting Gram-negative bacteria in Saudi Arabia. Journal of antimicrobial chemotherapy 67: 1701-1705.

Mertens G, Rand S, Jehaes E, Mommers N, Cardoen E, De Bruyn I, Leijnen G, Van Brussel K, Jacobs W (2009) Observation of tri-allelic patterns in autosomal STRs during routine casework. Forensic Science International: Genetics Supplement Series 2: 38-40.

Mohammad T, Xue Y, Evison M, Tyler-Smith C (2009) Genetic structure of nomadic Bedouin from Kuwait. Heredity 103: 425-433.

Monies D, Abouelhoda M, Assoum M, Moghrabi N, Rafiullah R, Almontashiri N, Alowain M, Alzaidan H, Alsayed M, Subhani S (2019) Lessons Learned from Large-Scale, First-Tier Clinical Exome Sequencing in a Highly

Consanguineous Population. American Journal of Human Genetics 104: 1182-1201.

Mulero JJ, Chang CW, Calandro LM, Green RL, Li Y, Johnson CL, Hennessy LK (2006) Development and validation of the AmpFℓSTR® Yfiler™ PCR amplification kit: a male specific, single amplification 17 Y-STR multiplex system. Journal of Forensic Sciences 51: 64-75.

Muzzio M, Ramallo V, Motti JM, Santos MR, Lopez Camelo JS, Bailliet G (2011) Software for Y-haplogroup predictions: a word of caution. International Journal of Legal Medicine 125: 143-147.

Nachman MW, Crowell SL (2000) Estimate of the mutation rate per nucleotide in humans. Genetics 156: 297-304.

Nadir E, Margalit H, Gallily T, Ben-Sasson SA (1996) Microsatellite spreading in the human genome: evolutionary mechanisms and structural implications. Proceedings of the National Academy of Sciences of the United States of America 93: 6470-6475.

Nasidze I, Schadlich H, Stoneking M (2003) Haplotypes from the Caucasus, Turkey and Iran for nine Y-STR loci. Forensic Science International 137: 85-93.

Nazir M, Alhaddad H, Alenizi M, Alenizi H, Taqi Z, Sanqoor S, Alrazouqi A, Hassan A, Alfalasi R, Gaur S, Al Jaber J, Ziab J, Al-Harbi E, Moura-Neto RS, Budowle B (2016) A genetic overview of 23Y-STR markers in UAE population. Forensic Science International: Genetics 23: 150-152.

Nebel A, Landau-Tasseron E, Filon D, Oppenheim A, Faerman M (2002) Genetic evidence for the expansion of Arabian tribes into the Southern Levant and North Africa. American Journal of Human Genetics 70: 1594-1596.

Nothnagel M, Szibor R, Vollrath O, Augustin C, Edelmann J, Geppert M, Alves C, Gusmão L, Vennemann M, Hou Y (2012) Collaborative genetic mapping of 12 forensic short tandem repeat (STR) loci on the human X chromosome. Forensic Science International: Genetics 6: 778-784.

Novroski NMM, King JL, Churchill JD, Seah LH, Budowle B (2016) Characterization of genetic sequence variation of 58 STR loci in four major population groups. Forensic Science International: Genetics 25: 214-226.

Nuchprayoon S, Saksirisampant W, Jaijakul S, Nuchprayoon I (2007) Flinders technology associates (FTA) filter paper–based DNA extraction with polymerase chain reaction (PCR) for detection of Pneumocystis jirovecii from respiratory specimens of immunocompromised patients. Journal of Clinical Laboratory Analysis 21: 382-386.

Oh YN, Lee HY, Lee EY, Kim EH, Yang WI, Shin K-J (2015) Haplotype and mutation analysis for newly suggested Y-STRs in Korean father–son pairs. Forensic Science International: Genetics 15: 64-68.

Ohta T, Kimura M (1973) A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genetics Research 22: 201-204.

Oldoni F, Kidd KK, Podini D (2019) Microhaplotypes in forensic genetics. Forensic Science International: Genetics 38: 54-69.

Omran GA, Rutty GN, Jobling MA (2009) Genetic variation of 15 autosomal STR loci in Upper (Southern) Egyptians. Forensic Science International: Genetics 3: e39-e44.

Osman AE, Alsafar H, Tay GK, Theyab JBJM, Mubasher M, Sheikh NE-E, AlHarthi H, Crawford MH, Gehad El Ghazali G (2015) Autosomal short tandem repeat (STR) variation based on 15 loci in a population from the Central Region (Riyadh Province) of Saudi Arabia. Journal of Forensic Research 6: 1000267.

Panter-Brick C (1991) Parental responses to consanguinity and genetic disease in Saudi Arabia. Social Science & Medicine 33: 1295-1302.

Parson W, Ballard D, Budowle B, Butler JM, Gettings KB, Gill P, Gusmão L, Hares DR, Irwin JA, King JL, Knijff Pd, Morling N, Prinz M, Schneider PM, Neste CV, Willuweit S, Phillips C (2016) Massively parallel sequencing of forensic STRs: Considerations of the DNA commission of the International Society for Forensic Genetics (ISFG) on minimal nomenclature requirements. Forensic Science International: Genetics 22: 54-63.

Parton A, White TS, Parker AG, Breeze PS, Jennings R, Groucutt HS, Petraglia MD (2015) Orbital-scale climate variability in Arabia as a potential motor for human dispersals. Quaternary International 382: 82-97.

Perez-Miranda AM, Alfonso-Sanchez MA, Pena JA, Herrera RJ (2006) Qatari DNA variation at a crossroad of human migrations. Human Heredity 61: 67-79.

Pérez-Miranda AM, Alfonso-Sánchez MA, Peña JA, Herrera RJ (2006) Qatari DNA variation at a crossroad of human migrations. Human Heredity 61: 67-79.

Petraglia MD, Parton A, Groucutt HS, Alsharekh A (2015) Green Arabia: Human prehistory at the crossroads of continents. Quaternary International 382: 1-7.

Phillips C, Amigo J, Carracedo A, Lareu M (2015) Tetra-allelic SNPs: informative forensic markers compiled from public whole-genome sequence data. Forensic Science International: Genetics 19: 100-106.

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 (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 39: 2708-2724.

Phillips C, Gelabert-Besada M, Fernandez-Formoso L, García-Magariños M, Santos C, Fondevila M, Ballard D, Syndercombe Court D, Carracedo Á, Victoria Lareu M (2014) "New turns from old STaRs": Enhancing the capabilities of forensic short tandem repeat analysis. Electrophoresis 35: 3173-3187.

Piatek J, Ossowski A, Parafiniuk M, Pudlo A, Jasionowicz K, Jalowinska K, Niemcunowicz-Janica A, Konarzewska M, Pepinski W (2012) Y-chromosomal haplotypes for the AmpFlSTR Yfiler PCR amplification kit in a population sample of Bedouins residing in the area of the Fourth Nile Cataract. Forensic Science International: Genetics 6: e176-e177.

Platt DE, Haber M, Dagher-Kharrat MB, Douaihy B, Khazen G, Ashrafian Bonab M, Salloum A, Mouzaya F, Luiselli D, Tyler-Smith C, Renfrew C, Matisoo-Smith E, Zalloua PA (2017) Mapping post-glacial expansions: the peopling of Southwest Asia. Scientific Reports 7: 40338.

Poulsen L, Tomas C, Drobnič K, Ivanova V, Mogensen HS, Kondili A, Miniati P, Bunokiene D, Jankauskiene J, Pereira V (2016) NGMSElect™ and Investigator® Argus X-12 analysis in population samples from Albania, Iraq, Lithuania, Slovenia, and Turkey. Forensic Science International: Genetics 22: 110-112.

Poznik GD, Xue Y, Mendez FL, Willems TF, Massaia A, Wilson Sayres MA, Ayub Q, McCarthy SA, Narechania A, Kashin S, Chen Y, Banerjee R, Rodriguez-Flores JL, Cerezo M, Shao H, Gymrek M, Malhotra A, Louzada S, Desalle R, Ritchie GR, Cerveira E, Fitzgerald TW, Garrison E, Marcketta A, Mittelman D, Romanovitch M, Zhang C, Zheng-Bradley X, Abecasis GR, McCarroll SA, Flicek P, Underhill PA, Coin L, Zerbino DR, Yang F, Lee C, Clarke L, Auton A, Erlich Y, Handsaker RE, 1000 Genomes Project Consortium, Bustamante CD, Tyler-Smith C (2016) Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences. Nature Genetics 48: 593-599.

Purps J, Geppert M, Nagy M, Roewer L (2015) Validation of a combined autosomal/Y-chromosomal STR approach for analyzing typical biological stains in sexual-assault cases. Forensic Science International: Genetics 19: 238-242.

QIAGEN (2015) Investigator® Argus X-12 QS Handbook [online] Available at: https://www.qiagen.com/us/resources/download.aspx?id=d3ab4ec3-33ec-4be3-8a88-6b66ebebdcf0&lang=en.

Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL (2005) Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proceedings of the National Academy of Sciences of the United States of America 102: 15942-15947.

Rappold GA (1993) The pseudoautosomal regions of the human sex chromosomes.

Human Genetics 92: 315-324.

Raziel A, Oz C, Carmon ADA, Ilsar R, Zamir A (2012) Discordance at D3S1358 locus involving SGM Plus™ and the European new generation multiplex kits. Forensic Science International: Genetics 6: 108-112.

Ristow PG, Cloete KW, D'Amato ME (2016) GlobalFiler® Express DNA amplification kit in South Africa: Extracting the past from the present. Forensic Science International: Genetics 24: 194-201.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP (2011) Integrative genomics viewer. Nature Biotechnology 29: 24-26.

Roewer L, Arnemann J, Spurr NK, Grzeschik KH, Epplen JT (1992) Simple repeat sequences on the human Y chromosome are equally polymorphic as their autosomal counterparts. Human Genetics 89: 389-394.

Roewer L, Willuweit S, Stoneking M, Nasidze I (2009) A Y-STR database of Iranian and Azerbaijanian minority populations. Forensic Science International: Genetics 4: e53-55.

Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. Bioinformatics methods and protocols. Springer, Totowa, pp 365-386

Ruitberg CM, Reeder DJ, Butler JM (2001) STRBase: a short tandem repeat DNA database for the human identity testing community. Nucleic Acids Research 29: 320-322.

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America 74: 5463-5467.

Schaad NW, Frederick RD, Shaw J, Schneider WL, Hickson R, Petrillo MD, Luster DG (2003) Advances in molecular-based diagnostics in meeting crop biosecurity and phytosanitary issues. Annual Review of Phytopathology 41: 305-324.

Schaffner SF (2004) The X chromosome in population genetics. Nature Reviews Genetics 5: 43-51.

Schlecht J, Kaplan ME, Barnard K, Karafet T, Hammer MF, Merchant NC (2008) Machine-learning approaches for classifying haplogroup from Y chromosome STR data. PLOS Computational Biology 4: e1000093.

Schlötterer C, Tautz D (1992) Slippage synthesis of simple sequence DNA. Nucleic Acids Research 20: 211-215.

Schneider PM (2007) Scientific standards for studies in forensic genetics. Forensic

Science International 165: 238-243.

Scott EM, Halees A, Itan Y, Spencer EG, He Y, Azab MA, Gabriel SB, Belkadi A, Boisson B, Abel L, Clark AG, Greater Middle East Variome C, Alkuraya FS, Casanova JL, Gleeson JG (2016) Characterization of Greater Middle Eastern genetic variation for enhanced disease gene discovery. Nature Genetics 48: 1071-1076.

Scozzari R, Massaia A, Trombetta B, Bellusci G, Myres NM, Novelletto A, Cruciani F (2014) An unbiased resource of novel SNP markers provides a new chronology for the human Y chromosome and reveals a deep phylogenetic structure in Africa. Genome Research 24: 535-544.

Seman A, Bakar ZA, Isa MN (2012) An efficient clustering algorithm for partitioning Y-short tandem repeats data. BioMed Central Research Notes 5: 557.

Sharma V, Chow HY, Siegel D, Wurmbach E (2017) Qualitative and quantitative assessment of Illumina's forensic STR and SNP kits on MiSeq FGx™. PLoS One 12: e0187932.

Sibille I, Duverneuil C, de la Grandmaison GL, Guerrouache K, Teissiere F, Durigon M, de Mazancourt P (2002) Y-STR DNA amplification as biological evidence in sexually assaulted female victims with no cytological detection of spermatozoa. Forensic Science International 125: 212-216.

Sigurðardóttir S, Helgason A, Gulcher JR, Stefansson K, Donnelly P (2000) The mutation rate in the human mtDNA control region. American Journal of Human Genetics 66: 1599-1609.

Silvia AL, Shugarts N, Smith J (2017) A preliminary assessment of the ForenSeq™ FGx System: next generation sequencing of an STR and SNP multiplex. International Journal of Legal Medicine 131: 73-86.

Sinha S, Amjad M, Rogers C, Hamby JE, Tahir UA, Balamurugan K, al-Kubaidan NA, Choudhry AR, Budowle B, Tahir MA (1999) Typing of eight short tandem repeat (STR) loci in a Saudi Arabian population. Forensic Science International 104: 143-146.

Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, Repping S, Pyntikova R, Ali J, Bieri T, Chinwalla A, Delehaunty A, Delehaunty K, Du H, Fewell G, Fulton L, Fulton R, Graves T, Hou S-F, Latrielle P, Leonard S, Mardis E, Maupin R, McPherson J, Miner T, Nash W, Nguyen C, Ozersky P, Pepin K, Rock S, Rohlfing T, Scott K, Schultz B, Strong C, Tin-Wollam A, Yang S-P, Waterston RH, Wilson RK, Rozen S, Page DC (2003) The male-specific region of the human Y chromosome: a mosaic of discrete sequence classes. Nature 423: 825-837.

Soares P, Ermini L, Thomson N, Mormina M, Rito T, Rohl A, Salas A, Oppenheimer S, Macaulay V, Richards MB (2009) Correcting for purifying selection: an improved human mitochondrial molecular clock. American Journal of

Human Genetics 84: 740-759.

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

Subramanian S, Mishra RK, Singh L (2003) Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. Genome Biology 4: R13.

SWGDM (2019) Addendum to the SWGDAM Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories to Address Next Generation Sequencing [online] Available at https://docs.wixstatic.com/ugd/4344b0_91f2b89538844575a9f51867def7be85.pdf.

Szibor R (2007) X-chromosomal markers: Past, present and future. Forensic Science International: Genetics 1: 93-99.

Szibor R, Krawczak M, Hering S, Edelmann J, Kuhlisch E, Krause D (2003) Use of X-linked markers for forensic purposes. International Journal of Legal Medicine 117: 67-74.

Tadmouri GO, Nair P, Obeid T, Al Ali MT, Al Khaja N, Hamamy HA (2009) Consanguinity and reproductive health among Arabs. Reproductive Health 6: 17.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: molecular evolutionary genetics analysis version 6.0. Molecular Biology and Evolution 30: 2725-2729.

Taqi Z, Alenizi M, Alenizi H, Ismael S, Dukhyil AA, Nazir M, Sanqoor S, Al Harbi E, Al-Jaber J, Theyab J, Moura-Neto RS, Budowle B (2015) Population genetics of 23 Y-STR markers in Kuwaiti population. Forensic Science International: Genetics 16: 203-204.

Tereba A (1999) Tools for analysis of population statistics. Profiles DNA 2: 14-16.

Thompson JM, Ewing MM, Frank WE, Pogemiller JJ, Nolde CA, Koehler DJ, Shaffer AM, Rabbach DR, Fulmer PM, Sprecher CJ (2013) Developmental validation of the PowerPlex® Y23 System: a single multiplex Y-STR analysis system for casework and database samples. Forensic Science International: Genetics 7: 240-250.

Tillmar AO, Kling D, Butler JM, Parson W, Prinz M, Schneider PM, Egeland T, Gusmão L (2017) DNA Commission of the International Society for Forensic Genetics (ISFG): Guidelines on the use of X-STRs in kinship analysis. Forensic Science International: Genetics 29: 269-275.

Triki-Fendri S, Alfadhli S, Ayadi I, Kharrat N, Ayadi H, Rebai A (2010) Genetic structure of Kuwaiti population revealed by Y-STR diversity. Annals of

Human Biology 37: 827-35.

Turrina S, Caratti S, Ferrian M, De Leo D (2016) Are rapidly mutating Y-short tandem repeats useful to resolve a lineage? Expanding mutability data on distant male relationships. Transfusion 56: 533-538.

Underhill PA, Kivisild T (2007) Use of Y chromosome and mitochondrial DNA population structure in tracing human migrations. Annual Review of Genetics 41: 539-564.

van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C (2018) The third revolution in sequencing technology. Trends in Genetics 34: 666-681.

van Oorschot R, Ballantyne K (2013) Capillary electrophoresis in forensic biology. In: J.A. Siegel PJS (ed) Encyclopedia of Forensic Sciences (second ed.). Academic Press, Waltham, pp 560-566

Venables WN, Ripley BD (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York

Venables WN, Ripley BD (2013) Modern applied statistics with S-PLUS. Springer Science & Business Media, New York

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XQH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang JH, Miklos GLG, Nelson C, Broder S, Clark AG, Nadeau C, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng ZM, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge WM, Gong FC, Gu ZP, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke ZX, Ketchum KA, Lai ZW, Lei YD, Li ZY, Li JY, Liang Y, Lin XY, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue BX, Sun JT, Wang ZY, Wang AH, Wang X, Wang J, Wei MH, Wides R, Xiao CL, Yan CH, et al. (2001) The sequence of the human genome. Science 291: 1304-1351.

Votrubova J, Ambers A, Budowle B, Vanek D (2017) Comparison of standard capillary electrophoresis based genotyping method and ForenSeq DNA Signature Prep kit (Illumina) on a set of challenging samples. Forensic Science International: Genetics Supplement Series 6: e140-e142.

Wei W, Ayub Q, Xue Y, Tyler-Smith C (2013) A comparison of Y-chromosomal lineage dating using either resequencing or Y-SNP plus Y-STR genotyping. Forensic Science International: Genetics 7: 568-72.

Weissensteiner H, Pacher D, Kloss-Brandstätter A, Forer L, Specht G, Bandelt H-J,

Kronenberg F, Salas A, Schönherr S (2016) HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. Nucleic Acids Research 44: W58-W63.

Wendt FR, King JL, Novroski NMM, Churchill JD, Ng J, Oldt RF, McCulloh KL, Weise JA, Smith DG, Kanthaswamy S, Budowle B (2017) Flanking region variation of ForenSeq™ DNA Signature Prep Kit STR and SNP loci in Yavapai Native Americans. Forensic Science International: Genetics 28: 146-154.

Westen AA, Kraaijenbrink T, Robles de Medina EA, Harteveld J, Willemse P, Zuniga SB, van der Gaag KJ, Weiler NE, Warnaar J, Kayser M, Sijen T, de Knijff P (2014) Comparing six commercial autosomal STR kits in a large Dutch population sample. Forensic Science International: Genetics 10: 55-63.

Westen AA, Matai AS, Laros JF, Meiland HC, Jasper M, de Leeuw WJ, de Knijff P, Sijen T (2009) Tri-allelic SNP markers enable analysis of mixed and degraded DNA samples. Forensic Science International: Genetics 3: 233-241.

Wilkins JF (2006) Unraveling male and female histories from human genetic data. Current Opinion in Genetics & Development 16: 611-617.

Willems T, Gymrek M, Poznik GD, Tyler-Smith C, Erlich Y (2016) Population-Scale Sequencing Data Enable Precise Estimates of Y-STR Mutation Rates. American Journal of Human Genetics 98: 919-933.

Wilson Sayres MA, Lohmueller KE, Nielsen R (2014) Natural selection reduced diversity on human Y chromosomes. PLoS Genetics 10: e1004064.

Woerner AE, King JL, Budowle B (2017) Fast STR allele identification with STRait Razor 3.0. Forensic Science International: Genetics 30: 18-23.

Wood ET, Stover DA, Ehret C, Destro-Bisol G, Spedini G, McLeod H, Louie L, Bamshad M, Strassmann BI, Soodyall H, Hammer MF (2005) Contrasting patterns of Y chromosome and mtDNA variation in Africa: evidence for sex-biased demographic processes. European Journal of Human Genetics 13: 867-876.

Wynbrandt J, Gerges FA (2010) Brief History of Saudi Arabia : Brief History of Saudi Arabia (2nd Edition). Infobase Learning, New York

Xavier C, Parson W (2017) Evaluation of the Illumina ForenSeq™ DNA Signature Prep Kit – MPS forensic application for the MiSeq FGx™ benchtop sequencer. Forensic Science International: Genetics 28: 188-194.

Xue Y, Wang Q, Long Q, Ng BL, Swerdlow H, Burton J, Skuce C, Taylor R, Abdellah Z, Zhao Y, Macarthur DG, Quail MA, Carter NP, Yang H, Tyler-Smith C (2009) Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. Current Biology 19: 1453-1457.

Y Chromosome Consortium (2002) A nomenclature system for the tree of human Y-chromosomal binary haplogroups. Genome Research 12: 339-348.

Yao Y-G, Kajigaya S, Young NS (2015) Mitochondrial DNA mutations in single human blood cells. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis 779: 68-77.

Zalloua PA, Xue Y, Khalife J, Makhoul N, Debiane L, Platt DE, Royyuru AK, Herrera RJ, Hernanz DF, Blue-Smith J, Wells RS, Comas D, Bertranpetit J, Tyler-Smith C (2008) Y-chromosomal diversity in Lebanon is structured by recent historical events. American Journal of Human Genetics 82: 873-882.

Zeng X, King J, Hermanson S, Patel J, Storts DR, Budowle B (2015a) An evaluation of the PowerSeq™ Auto System: A multiplex short tandem repeat marker kit compatible with massively parallel sequencing. Forensic Science International: Genetics 19: 172-179.

Zeng X, King JL, Stoljarova M, Warshauer DH, LaRue BL, Sajantila A, Patel J, Storts DR, Budowle B (2015b) High sensitivity multiplex short tandem repeat loci analyses with massively parallel sequencing. Forensic Science International: Genetics 16: 38-47.

Zerjal T, Xue Y, Bertorelle G, Wells RS, Bao W, Zhu S, Qamar R, Ayub Q, Mohyuddin A, Fu S, Li P, Yuldasheva N, Ruzibakiev R, Xu J, Shu Q, Du R, Yang H, Hurles ME, Robinson E, Gerelsaikhan T, Dashnyam B, Mehdi SQ, Tyler-Smith C (2003) The genetic legacy of the Mongols. American Journal of Human Genetics 72: 717-721.

Zhang W, Xiao C, Wei T, Pan C, Yi S, Huang D (2016) Haplotype diversity of 13 RM Y-STRs in Chinese Han population and an update on the allele designation of DYF403S1. Forensic Science International: Genetics 23: e1-e9.

Zhivotovsky LA, Underhill PA, Cinnioglu C, Kayser M, Morar B, Kivisild T, Scozzari R, Cruciani F, Destro-Bisol G, Spedini G, Chambers GK, Herrera RJ, Yong KK, Gresham D, Tournev I, Feldman MW, Kalaydjieva L (2004) The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. American Journal of Human Genetics 74: 50-61.

Chapter 9 Appendices

9.1 List of supplementary result tables

Table 9.1 Sub-regional affiliations, and predicted Y-haplogroups for 597 Saudi Arabian males. (Text)

Table 9.2: Regional allele frequencies and forensic statistics.

Table 9.3: Regional per-locus observed (O-het) and expected heterozygosity (E-het) values

Table 9.4: Population differentiation exact tests

Table 9.5: Autosomal STR variants.

Table 9.6: Autosomal STR variants absent from STRait Razor database

Table 9.7: iiSNP Flanking Region Report with novel SNP variants

Table 9.8: iiSNP variants (not highlighted in Flanking Region Report).

Table 9.9: STR forensic statistics (allele length, repeat region sequence & flanking region)

Table 9.10: iiSNP forensic statistics (target SNP only & whole amplicon)

Table 9.11: Deviation from expected heterozygosity under HWE at each of the autosomal STR loci.

Table 9.12: Deviation from expected heterozygosity under HWE at each of the autosomal iiSNP loci.

Table 9.13: Table S7: HGDP iiSNP genotypes

Table 9.14: Sex-chromosomal STR variants with frequencies.

Table 9.15: Sex chromosomal STR variants absent from STRait Razor database.

Table 9.16: Y-STR structure linked to haplogroup.

Table 9.17: STR forensic statistics.

Table 9.18: Haplogrep 2 mitochondrial haplogroup predictions

Note: All appendix tables are provided on a CD; Tables 9.1, 9.3, 9.6, 9.8, 9.15 9.16 and 9.18 are also provided in text form here.

Table 9.1 Sub-regional affiliations, and predicted Y-haplogroups for 597 Saudi Arabian males.

Sample	Recruitment	NevGen predicted haplogroup	Haplogroup probability	Haplogroup fitness	Sample	Recruitment	NevGen predicted haplogroup	Haplogroup probability	Haplogroup fitness	Sample	Recruitment	NevGen predicted haplogroup	Haplogroup probability	Haplogroup fitness	Sample	Recruitment	NevGen predicted haplogroup	Haplogroup probability
C1	SA	UP	90.06	8.03	C72	SA	J1	100	34.16	E65	SA	J1	100	65.58	N53	SA	J1	100
C10	SA	J1	99.42	49	C73	SA	J1	100	50.47	E66	SA	UP	100	41.05	N54	SA	J1	99.99
C100	SA	J1	100	55.64	C74	SA	J1	100	57.07	E67	SA	UP	100	30.65	N55	SA	J1	100
C101	SA	A	98.83	25.64	C75	SA	E1b1b	100	29.36	E68	SA	J1	100	53.48	N56	SA	J1	100
C104	SA	J1	100	45.24	C76	SA	J1	100	50.47	E69	SA	E1b1a	99.99	29.83	N57	SA	J1	100
C105	SA	J1	100	64	C77	SA	J1	99.98	33.88	E70	SA	E1b1b	100	21.35	N59	SA	J1	100
C106	SA	J1	100	58.17	C78	SA	J1	100	52.97	E71	SA	J2	99.97	53.28	N6	SA	J1	100
C107	SA	J1	100	44.87	C8	SA	J1	100	47.05	E72	SA	J1	100	47.16	N60	SA	J1	100
C108	SA	J1	100	57.07	C80	SA	J1	99.99	40.87	E73	SA	J1	99.95	38.6	N61	SA	J1	100
C109	SA	J1	100	62.07	C81	SA	J1	100	45.01	E74	SA	J1	100	40.08	N62	SA	J1	100
C11	SA	J1	100	53.01	C82	SA	G	100	49.93	E75	SA	J1	100	59.25	N63	SA	J1	100
C110	SA	J1	100	51.73	C83	SA	J1	99.95	36.24	E76	SA	A	100	24.54	N64	SA	J1	97.31
C111	SA	J1	100	49.81	C84	SA	J1	100	35.98	E77	SA	Q	100	30.99	N65	SA	J1	100
C112	SA	J1	100	46.78	C85	SA	J1	100	55.32	E78	SA	E1b1b	98.13	20.35	N67	SA	J1	100
C113	SA	J1	100	46.49	C86	SA	J1	100	58.87	E79	SA	T	100	38.49	N69	SA	J1	100
C114	SA	J1	100	52.97	C87	SA	Q	100	43.58	E80	SA	J1	100	48.17	N7	SA	J1	99.98
C115	SA	J1	100	51.03	C88	SA	J1	100	53.47	E81	SA	J1	99.99	44.33	N70	SA	J1	100
C116	SA	J1	100	54.27	C89	SA	J1	100	46.49	E82	SA	R1a	100	43.94	N71	SA	J1	100
C117	SA	J1	100	39.32	C9	SA	J1	99.97	31.65	E83	SA	J1	100	54.59	N72	SA	J1	100
C118	SA	J1	99.86	43.41	C90	SA	J1	100	49.53	E84	SA	J1	99.92	47.07	N73	SA	J1	100
C119	SA	J1	100	56.7	C94	SA	J1	100	62.07	E86	SA	J1	100	41.49	N74	SA	J1	100
C12	SA	J1	100	62.79	C95	SA	J1	99.99	35.07	E87	SA	R1a	100	44.73	N75	SA	J2	99.98
C120	SA	J1	100	53.3	C96	SA	J1	99.54	28.49	E88	SA	H	93.88	23.37	N76	SA	J1	100
C13	SA	T	100	37.81	C97	SA	J1	100	36.64	E89	SA	J1	100	38.75	N77	SA	J1	100
C14	SA	J1	100	45.39	E100	SA	J2	99.94	56.41	E90	SA	R1a	100	42.91	N78	SA	J1	100
C15	SA	J1	99.99	37.23	E101	SA	J1	99.65	35.73	E91	SA	Q	100	41.93	N79	SA	J1	100
C16	SA	J1	100	57.24	E102	SA	E1b1a	100	44.36	E92	SA	J1	99.84	44.41	N8	SA	E1b1b	100
C17	SA	J1	100	58.87	E103	SA	J2	90.55	38.36	E93	SA	J1	100	60.24	N80	SA	J1	100
C18	SA	J1	100	56.04	E104	SA	UP	100	40.63	E94	SA	J2	68.3	33.08	N81	SA	J1	100
C2	SA	J1	99.94	50.93	E105	SA	Q	100	46.77	E95	SA	J1	99.96	50.03	N82	SA	J1	100
C20	SA	J1	99.86	40.62	E106	SA	J1	99.97	37.74	E96	SA	E1b1b	100	37.07	N83	SA	J1	99.85
C21	SA	J1	99.99	45.76	E107	SA	R1a	100	39.22	E97	SA	R1a	100	41.44	N84	SA	E1b1b	100
C22	SA	J1	100	37.63	E108	SA	J1	100	58.8	E98	SA	J1	100	44.81	N85	SA	J1	100
C23	SA	J1	100	56.7	E109	SA	R1a	100	47.17	E99	SA	J1	99.99	45.21	N86	SA	E1b1b	100
C24	SA	J1	100	58.09	E110	SA	J1	100	37.74	N100	SA	E1b1b	99.89	13.95	N87	SA	E1b1b	100
C25	SA	J1	100	53.88	E111	SA	E1b1b	100	37.4	N101	SA	J1	100	58.03	N88	SA	R1b	98.92
C26	SA	J1	100	61.81	E113	SA	A	100	46.47	N102	SA	J1	100	57.03	N89	SA	J1	97.98
C28	SA	J1	100	51.46	E114	SA	T	100	28.58	N103	SA	J1	100	53.22	N90	SA	J1	100
C29	SA	J1	99.45	42.06	E115	SA	J1	100	60.56	N104	SA	J1	100	55.43	N91	SA	J1	100
C30	SA	J1	99.43	31.52	E116	SA	J1	100	55.33	N105	SA	J1	100	45.66	N92	SA	J1	100
C31	SA	J1	99.66	24.06	E117	SA	R1b	100	37.25	N107	SA	J1	100	58.14	N93	SA	J1	97.85
C32	SA	J1	100	41.61	E118	SA	B	100	26.29	N108	SA	J1	100	55.54	N94	SA	J1	100
C33	SA	UP	100	35.13	E119	SA	J1	100	41.58	N11	SA	J1	100	57.99	N95	SA	J1	99.79
C34	SA	J1	100	50.6	E120	SA	J2	98.97	33.23	N110	SA	J1	100	45.87	N96	SA	E1b1b	65.58
C35	SA	J1	100	57.24	E13	SA	R1a	99.99	29.73	N111	SA	J1	100	49	N97	SA	J1	100
C36	SA	E1b1b	100	33.3	E17	SA	J1	99.99	37.85	N113	SA	J1	100	48.04	N98	SA	J1	100
C37	SA	J1	100	51.92	E19	SA	E1b1b	99.99	36.24	N114	SA	E1b1b	100	40.98	N99	SA	E1b1a	100
C38	SA	J1	100	52.12	E20	SA	J1	100	43.79	N115	SA	J1	100	49.78	S10	SA	E1b1b	99.41
C39	SA	E1b1b	100	34.14	E22	SA	J1	99.89	51.61	N119	SA	J1	100	44.87	S100	SA	J1	99.31
C4	SA	J1	100	44.7	E24	SA	J1	99.99	42.74	N12	SA	G	100	55.05	S101	SA	J1	99.98
C40	SA	J1	100	66.19	E25	SA	J1	100	57.66	N13	SA	J1	100	57.99	S102	SA	J1	100
C42	SA	J1	100	47	E27	SA	J1	100	52.77	N15	SA	J1	99.99	53.74	S103	SA	J1	99.98
C43	SA	J1	100	52.66	E29	SA	J1	99.59	28.42	N19	SA	J1	100	53.29	S104	SA	E1b1a	100
C44	SA	J1	100	54.88	E30	SA	E1b1b	99.98	24.09	N20	SA	UP	71.45	23.23	S105	SA	J1	100
C45	SA	J1	100	59.77	E32	SA	B	100	16.79	N21	SA	J1	100	54.18	S106	SA	J1	99.97
C46	SA	J1	100	59.22	E34	SA	E1b1b	99.87	24.46	N22	SA	J1	99.96	51.71	S107	SA	J1	100
C47	SA	J1	100	58.87	E35	SA	H	90.08	21.78	N23	SA	J1	99.99	38.35	S108	SA	R1b	97.24
C48	SA	J1	99.83	29.6	E36	SA	J1	100	54.78	N24	SA	J1	100	41.22	S109	SA	J1	100
C5	SA	J1	100	57.07	E37	SA	J1	99.95	48.41	N25	SA	J1	98.14	39.86	S11	SA	E1b1b	100
C50	SA	J1	100	27.54	E38	SA	J1	100	43.86	N26	SA	J1	99.87	39.42	S110	SA	J1	99.91
C51	SA	J1	100	48.37	E39	SA	E1b1b	100	36.61	N28	SA	J1	100	57.96	S111	SA	J2	99.29
C52	SA	J1	100	57.24	E40	SA	G	50.57	12.88	N29	SA	J1	99.98	53.39	S112	SA	J1	99.72
C53	SA	J1	100	45.89	E41	SA	B	100	44.1	N30	SA	J1	100	50.86	S113	SA	J1	99.91
C54	SA	J2	99.86	49.64	E43	SA	J1	100	45.04	N31	SA	J1	99.83	51.42	S114	SA	J1	100
C55	SA	J1	99.99	43.56	E44	SA	J1	100	52.17	N32	SA	J1	100	51.39	S115	SA	J1	99.99
C56	SA	J1	100	44.29	E45	SA	E1b1b	87.08	22.14	N33	SA	J1	99.93	50.29	S116	SA	J1	99.93
C57	SA	J1	100	58.87	E46	SA	E1b1b	100	28.75	N34	SA	J1	100	45.41	S117	SA	J1	99.99
C58	SA	J1	100	55.32	E47	SA	J1	99.93	33.42	N35	SA	J1	100	46.46	S118	SA	J1	100
C59	SA	J1	100	34.16	E49	SA	G	99.42	41.89	N36	SA	J1	100	56.63	S119	SA	J1	99.96
C6	SA	J1	100	41.24	E50	SA	J1	99.87	45.84	N38	SA	J1	100	48.81	S12	SA	J1	100
C60	SA	J1	100	46.49	E51	SA	E1b1b	67.83	18.71	N39	SA	J1	100	48.81	S120	SA	J1	100
C61	SA	J1	100	48.71	E53	SA	J1	100	42.37	N40	SA	J1	99.98	39.29	S16	SA	J1	98.57
C62	SA	J1	100	66.19	E54	SA	T	99.99	33.44	N41	SA	J1	99.98	45.64	S17	SA	J1	100
C63	SA	J1	100	45.41	E55	SA	J1	100	50.9	N42	SA	J1	100	52.3	S18	SA	J1	100
C64	SA	R1b	99.74	27.1	E56	SA	J1	100	39.75	N43	SA	J1	100	64.97	S19	SA	J1	100
C65	SA	J1	99.99	36.87	E57	SA	R1b	99.98	31.53	N44	SA	J1	100	50.52	S2	SA	J1	99.85
C66	SA	J1	100	59.95	E58	SA	E1b1b	99.72	30.38	N45	SA	J1	100	60.48	S20	SA	J1	100
C67	SA	J1	99.99	45.54	E59	SA	UP	100	41.89	N46	SA	J1	98.69	39.26	S21	SA	J1	100
C68	SA	J1	100	52.38	E60	SA	J1	100	55.51	N47	SA	J1	99.98	39.29	S22	SA	J1	100
C69	SA	J1	100	59.91	E61	SA	E1b1b	100	28.75	N48	SA	J1	100	40.64	S23	SA	J1	100
C7	SA	J1	99.99	36.17	E62	SA	E1b1b	100	39.91	N49	SA	J1	99.97	37.95	S24	SA	J1	100
C70	SA	J1	99.99	29.61	E63	SA	E1b1b	99.99	17.87	N51	SA	J1	100	58.03	S25	SA	J1	99.95
C71	SA	J1	99.95	31.06	E64	SA	J1	99.99	46.3	N52	SA	J1	100	63.46	S26	SA	J1	100

Sample	Recruitment	NevGen predicted haplogroup	Haplogroup probability	Haplogroup fitness	Sample	Recruitment	NevGen predicted haplogroup	Haplogroup probability	Haplogroup fitness	Sample	Recruitment	NevGen predicted haplogroup	Haplogroup probability	Haplogroup fitness	Sample	Recruitment	NevGen predicted haplogroup	Haplogroup probability
S27	SA	J1	100	52.66	W103	SA	UP	51.19	10.58	W95	SA	J1	99.99	65.58	W16	UK	J1	100
S28	SA	J1	99.9	27.75	W104	SA	J1	99.93	29.9	W96	SA	J1	100	41.05	W17	UK	G	100
S29	SA	J1	99.98	38.31	W105	SA	J1	100	54.59	W97	SA	J1	98.51	30.65	W18	UK	E1b1b	100
S3	SA	J1	99.96	39.67	W106	SA	J1	100	56.41	W98	SA	E1b1b	100	53.48	W19	UK	UP	99.92
S30	SA	B	100	15.34	W107	SA	E1b1b	100	41.66	W99	SA	J1	100	29.83	W2	UK	E1b1b	100
S31	SA	J1	100	54.59	W108	SA	J2	99.48	26.92	C1	UK	J1	100	21.35	W20	UK	UP	95.7
S32	SA	J1	100	37.65	W109	SA	E1b1b	100	33.24	C10	UK	E1b1b	100	53.28	W21	UK	A	100
S33	SA	J1	100	54.31	W11	SA	J1	100	54.05	C11	UK	J1	100	47.16	W22	UK	R1a	100
S34	SA	J1	100	45.43	W110	SA	J1	99.97	36.51	C12	UK	J1	100	38.6	W23	UK	J1	99.2
S35	SA	J1	100	45.99	W111	SA	Q	100	22.47	C14	UK	J1	100	40.08	W3	UK	J1	100
S36	SA	J1	99.96	36.73	W112	SA	J1	100	50.06	C15	UK	J1	100	59.25	W4	UK	E1b1b	99.98
S37	SA	B	100	11.5	W113	SA	J1	100	57.86	C16	UK	J1	100	24.54	W5	UK	J1	100
S38	SA	J1	100	37.65	W114	SA	J1	97.98	43.27	C17	UK	E1b1b	100	30.99	W6	UK	E1b1b	99.99
S39	SA	J1	100	48.1	W115	SA	E1b1b	100	33.24	C18	UK	J1	100	20.35	W7	UK	J1	99.98
S4	SA	J1	100	55.99	W12	SA	J1	100	57.55	C19	UK	E1b1b	100	38.49	W8	UK	J1	100
S40	SA	J1	99.97	42	W14	SA	J1	98.6	32.78	C2	UK	J1	99.96	48.17	W9	UK	J1	100
S41	SA	J1	100	49.23	W15	SA	B	99.95	20.4	C20	UK	R1a	100	44.33				
S42	SA	R1a	100	43.68	W16	SA	J1	100	36.14	C3	UK	J1	100	43.94				
S43	SA	J1	100	42.09	W19	SA	J1	100	52.79	C4	UK	J2	99.04	54.59				
S44	SA	J1	99.98	38.31	W2	SA	J1	100	57.93	C5	UK	J1	100	47.07				
S45	SA	J1	100	41.77	W20	SA	J1	100	52.25	C6	UK	E1b1b	100	41.49				
S46	SA	J1	99.17	37.06	W21	SA	T	100	37.81	C7	UK	J1	100	44.73				
S47	SA	J1	99.99	46.65	W22	SA	T	100	38.25	C9	UK	UP	83.68	23.37				
S48	SA	J1	100	48.05	W23	SA	T	100	33.94	E1	UK	J1	100	38.75				
S49	SA	J1	100	44.72	W25	SA	T	100	28.8	E10	UK	J1	100	42.91				
S5	SA	J1	100	52.66	W28	SA	J1	100	54.57	E11	UK	J2	98.78	41.93				
S50	SA	J1	99.85	38.92	W36	SA	J1	100	54.15	E12	UK	E1b1b	99.49	44.41				
S51	SA	E1b1a	100	29.66	W37	SA	B	100	33.11	E13	UK	J1	100	60.24				
S52	SA	E1b1b	99.94	18.63	W38	SA	J1	100	38.87	E14	UK	J1	99.2	33.08				
S53	SA	T	99.99	22.17	W39	SA	J1	100	55.24	E15	UK	J1	99.95	50.03				
S54	SA	J1	100	46.12	W40	SA	J1	100	46.03	E16	UK	J1	99.99	37.07				
S55	SA	J1	100	45.91	W41	SA	J1	100	42.49	E17	UK	G	69.18	41.44				
S56	SA	J1	99.97	42	W42	SA	J1	100	38.87	E2	UK	E1b1b	95.57	44.81				
S57	SA	E1b1a	100	29.66	W43	SA	J1	99.86	26.16	E3	UK	J1	99.82	45.21				
S58	SA	J1	99.98	49.56	W44	SA	J1	100	65.58	E4	UK	E1b1b	100	13.95				
S59	SA	T	97.25	20.21	W45	SA	J1	100	54.15	E5	UK	E1b1b	99.64	58.03				
S60	SA	J1	99.31	41.91	W46	SA	B	100	33.11	E6	UK	E1b1b	98.98	57.03				
S61	SA	J1	100	56.07	W47	SA	T	100	32.99	E7	UK	J1	100	53.22				
S62	SA	J1	99.99	40.36	W48	SA	J1	100	53.61	E8	UK	E1b1a	100	55.43				
S63	SA	J1	100	61.35	W49	SA	J1	100	53.61	E9	UK	J2	100	45.66				
S64	SA	J1	99.96	39.67	W50	SA	J1	100	45.27	N1	UK	J1	99.54	58.14				
S65	SA	J1	100	37.2	W51	SA	J1	51.33	33.78	N10	UK	J1	99.86	55.54				
S66	SA	J1	99.95	45.19	W52	SA	J1	78.27	27.55	N11	UK	J2	99.98	57.99				
S67	SA	J1	100	35.6	W53	SA	E1b1b	100	32.11	N2	UK	J1	100	45.87				
S68	SA	J1	96.05	30.18	W54	SA	Q	100	50.02	N3	UK	J1	99.98	49				
S69	SA	E1b1b	99.97	36.02	W55	SA	J1	100	53.17	N4	UK	J1	100	48.04				
S7	SA	J1	100	50.94	W56	SA	E1b1b	99.98	30.81	N5	UK	UP	100	40.98				
S70	SA	J1	99.95	40.79	W57	SA	J1	100	56.41	N7	UK	J1	100	49.78				
S71	SA	J1	100	60.88	W58	SA	J1	100	58.87	N8	UK	J1	100	44.87				
S72	SA	J1	99.96	39.67	W59	SA	G	99.46	17.88	N9	UK	G	100	55.05				
S73	SA	J1	99.93	48.84	W60	SA	E1b1b	97.79	28.37	S1	UK	R1a	100	57.99				
S74	SA	J1	100	51.71	W61	SA	T	100	27.04	S10	UK	J1	100	53.74				
S75	SA	J1	100	46.39	W62	SA	J1	100	50.4	S11	UK	J1	100	53.29				
S76	SA	J1	100	41.97	W63	SA	UP	99.86	15.47	S12	UK	J2	99.84	23.23				
S77	SA	E1b1b	100	16.08	W64	SA	J1	99.97	40.82	S13	UK	R1a	100	54.18				
S78	SA	J1	100	47.14	W65	SA	B	100	40.87	S14	UK	J1	100	51.71				
S79	SA	E1b1b	100	31.51	W66	SA	E1b1b	100	32.71	S15	UK	J1	100	38.35				
S8	SA	J1	100	45.13	W67	SA	J1	98.6	32.78	S16	UK	J1	78.42	41.22				
S80	SA	J1	100	43.1	W68	SA	J1	94.23	29.98	S17	UK	J1	100	39.86				
S81	SA	J1	99.98	39.32	W69	SA	J1	100	58.87	S18	UK	J1	100	39.42				
S82	SA	J1	63.44	42.29	W70	SA	E1b1b	71.34	16.47	S19	UK	J2	99.86	57.96				
S84	SA	J1	99.96	40.52	W71	SA	J1	100	53.98	S2	UK	J1	99.99	53.39				
S85	SA	J1	95.78	22.53	W72	SA	J1	99.99	39.14	S21	UK	E1b1b	100	50.86				
S86	SA	J1	99.79	45.53	W73	SA	J1	100	37.29	S22	UK	Q	57.03	51.42				
S87	SA	J1	100	51.84	W74	SA	E1b1b	100	41.51	S23	UK	J1	99.77	51.39				
S88	SA	J1	99.51	26.57	W75	SA	J1	99.71	24.57	S24	UK	E1b1a	99.97	50.29				
S89	SA	J1	63.59	38.16	W76	SA	E1b1b	100	20.61	S25	UK	J1	100	45.41				
S9	SA	J1	76.07	41.3	W78	SA	R1a	100	37.67	S26	UK	J1	100	46.46				
S90	SA	E1b1b	100	30.98	W79	SA	E1b1b	100	40.68	S27	UK	T	100	56.63				
S91	SA	J1	99.99	44.18	W80	SA	E1b1b	100	35.37	S3	UK	J1	100	48.81				
S92	SA	J1	99.95	50.44	W81	SA	A	89.74	18.57	S4	UK	J1	100	48.81				
S93	SA	J1	100	52.26	W82	SA	J1	100	48.15	S5	UK	T	100	39.29				
S94	SA	J1	100	62.12	W83	SA	E1b1b	96.74	30.53	S6	UK	J1	100	45.64				
S95	SA	J1	95.1	28.96	W84	SA	T	100	39.24	S7	UK	J1	100	52.3				
S96	SA	J1	99.98	44.17	W85	SA	Q	54.12	18.44	S8	UK	E1b1b	100	64.97				
S97	SA	J1	98.42	31.56	W86	SA	J1	100	38.74	S9	UK	J1	99.21	50.52				
S98	SA	J1	100	48.54	W87	SA	E1b1b	100	40.24	W1	UK	E1b1b	100	60.48				
S99	SA	J1	100	45.38	W89	SA	J1	100	59.95	W10	UK	E1b1b	100	39.26				
W1	SA	J1	100	48.73	W9	SA	UP	50.91	18.64	W11	UK	E1b1b	100	39.29				
W10	SA	B	100	19.98	W90	SA	J1	100	51.96	W12	UK	J1	92.29	40.64				
W100	SA	J1	100	59.3	W91	SA	J1	100	60.14	W13	UK	UP	45.73	37.95				
W101	SA	R1a	100	52.1	W92	SA	J1	100	46.02	W14	UK	J2	66.52	58.03				
W102	SA	E1b1b	99.85	22.5	W94	SA	J1	99.95	29.96	W15	UK	E1b1a	100	63.46				

Table 9.3: Regional per-locus observed (O-het) and expected heterozygosity (E-het) values.

C						N					E					S					W				
Locus	Genotype	alleles	O-Het.	E-Het	P-value	Genotype	alleles	O-Het.	E-Het	P-value	Genotype	alleles	O-Het.	E-Het	P-value	Genotype	alleles	O-Het.	E-Het	P-value	Genotype	alleles	O-Het.	E-Het	P-value
D3S1358	230	7	0.774	0.770	0.051	208	7	0.644	0.728	0.077	176	6	0.773	0.767	0.997	206	7	0.718	0.758	0.410	226	8	0.735	0.756	0.153
vWA	230	7	0.730	0.784	0.119	208	7	0.692	0.782	**0.049**	176	8	0.727	0.793	0.053	206	8	0.767	0.768	0.195	226	7	0.717	0.788	0.511
D16S539	230	9	0.713	0.753	0.079	208	8	0.683	0.776	**0.021**	176	6	0.739	0.783	0.873	206	7	0.816	0.809	0.199	226	7	0.743	0.764	0.816
CSF1PO	230	6	0.713	0.694	0.313	208	7	0.692	0.692	0.873	176	7	0.682	0.720	0.733	206	5	0.631	0.699	0.192	226	7	0.637	0.701	0.385
TPOX	230	5	0.678	0.663	0.074	208	5	0.587	0.594	0.757	176	6	0.511	0.602	0.365	206	6	0.612	0.646	0.380	226	5	0.575	0.645	0.210
D8S1179	230	9	0.826	0.829	0.609	208	9	0.827	0.840	0.217	176	10	0.795	0.830	0.811	206	10	0.845	0.842	0.433	226	9	0.823	0.819	0.684
D21S11	230	11	0.800	0.820	0.832	208	12	0.740	0.789	0.163	176	11	0.750	0.819	0.473	206	10	0.796	0.802	0.925	226	14	0.796	0.847	0.784
D18S51	230	15	0.843	0.871	0.093	208	14	0.808	0.855	0.663	176	16	0.886	0.891	0.823	206	12	0.835	0.880	0.733	226	15	0.858	0.872	0.152
D2S441	230	8	0.765	0.753	**0.012**	208	10	0.712	0.763	0.556	176	8	0.705	0.739	**0.029**	206	9	0.650	0.769	**0.033**	226	9	0.708	0.748	0.490
D19S433	230	14	0.809	0.880	0.492	208	11	0.856	0.851	0.577	176	12	0.841	0.867	0.645	206	13	0.854	0.872	0.528	226	13	0.796	0.880	0.161
TH01	230	6	0.652	0.725	0.239	208	7	0.750	0.781	0.553	176	6	0.761	0.763	0.464	206	7	0.816	0.782	0.882	226	5	0.717	0.741	0.968
FGA	230	14	0.826	0.839	0.694	208	12	0.846	0.864	0.927	176	13	0.784	0.856	0.269	206	13	0.874	0.866	0.072	226	15	0.823	0.858	0.079
D22S1045	230	7	0.626	0.657	**0.039**	208	6	0.567	0.620	0.627	176	6	0.614	0.668	0.412	206	6	0.602	0.637	0.729	226	7	0.593	0.698	**0.022**
D5S818	230	7	0.652	0.738	0.152	208	7	0.788	0.792	0.947	176	8	0.682	0.747	0.240	206	7	0.748	0.753	0.089	226	7	0.735	0.754	0.805
D13S317	230	8	0.661	0.770	**0.031**	208	8	0.760	0.809	0.250	176	7	0.784	0.792	0.875	206	7	0.709	0.779	0.614	226	7	0.681	0.738	**0.010**
D7S820	230	7	0.704	0.769	0.099	208	7	0.692	0.775	**0.028**	176	7	0.795	0.771	0.313	206	8	0.738	0.763	0.700	226	7	0.770	0.786	0.293
SE33	230	32	0.939	0.950	0.338	208	30	0.913	0.947	0.312	176	26	0.898	0.954	0.099	206	31	0.932	0.940	0.763	226	28	0.903	0.948	0.182
D10S1248	230	7	0.696	0.749	0.703	208	8	0.788	0.741	0.556	176	7	0.898	0.762	0.335	206	8	0.796	0.764	0.390	226	8	0.761	0.769	0.117
D1S1656	230	14	0.861	0.873	0.759	208	15	0.846	0.846	0.506	176	11	0.773	0.879	**0.022**	206	11	0.835	0.852	0.955	226	12	0.770	0.867	0.088
D12S391	230	15	0.809	0.890	0.720	208	14	0.817	0.878	0.124	176	14	0.852	0.896	0.723	206	14	0.835	0.889	0.719	226	13	0.850	0.890	0.120
D2S1338	230	13	0.783	0.834	0.108	208	10	0.740	0.852	**0.007**	176	10	0.795	0.855	0.081	206	11	0.777	0.852	**0.031**	226	12	0.823	0.856	0.377

Table 9.6: Autosomal STR variants absent from STRait Razor database: Nomenclature, allele length, repeat and flanking sequences with previous observations. AFR: AFRICAN, EAS: EAST ASIAN, EUR: EUROPEAN, MEA: MIDDLE EAST, OCE: OCEANIAN, SAS: SOUTH ASIAN, RL: Repeat Length, RS: Repeat Sequence, FS: Flank Sequence; n/a: not available as flanking SNP data not in Phillips et al. 2018

Locus	CE allele	Seq	Nomenclature	Type	Obs	HGDP occurrence	GenBank	Novel
CSF1PO	12	AAGATAGATAGATTAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGGTAGATAGGAAG	CSF1PO [CE 12]-GRCh38-Chr5-150076318-150076389 ATCT ACCT (ATCT)10	RS	3	MEA	Missing	
D10S1248	9	TTGAACAAATGAGTGAGTGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAATGAAGACAATACA ACCAGAGTT	D10S1248 [CE 9]-GRCh38-Chr10 129294226-129294318 (GGAA)9	RL	2	MEA/AFR	MH167056.1	
D12S391	23	CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGTAGATAGATAGATAGATAGATAGATAGATAGATAG ATAGATAGATAGATAGATAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGAGGG GATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCT GGAGACCCAGGGACACTA	D12S391 [CE 23]-GRCh38-Chr12-12296981-12297168 (AGAT)12 (AGAC)11	RS	1	EUR/OCE/SAS	MH167177.1	
D12S391	23	CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGTAGATAGATAGATAGATAGATAGATAGATAGATAG ATAGATAGATAGATAGATAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGATGAGAGGG GATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTCCATCTGCAAGCT GGAGACCCAGGGACACTA	D12S391 [CE 23]-GRCh38-Chr12-12296981-12297168 (AGAT)12 (AGAC)10 AGAT	RS	1	EUR	Missing	
D12S391	26	CAGAGAGAAAGAATCAACAGGATCAATGGATGCATAGGTAGATAGATAGATAGATAGATAGATAGATAGATAG ATAGATAGATAGATAGATAGATAGATAGATAGATAGACAGACAGACAGACAGACAGACAGACAGACAGA CAGATGAGAGGGGATTTATTAGAGGAATTAGCTCAAGTGATATGGAGGCTGAAAAATCTCATGACAGTC CATCTGCAAGCTGGAGACCCAGGGACACTA	D12S391 [CE 26]-GRCh38-Chr12-12296981-12297168 (AGAT)17 (AGAC)8 AGAT	RS	1	EAS	Missing	
D13S317	9	TCTGACCCATCTAACGCCTATCTGTATTTACAAATACATTATCTATCTATCTATCTATCTATCTATCTA TCAATCAATCAATCATCTATCTATCTTTCTGTCTGTCTTTTTGGG	D13S317 [CE 9]-GRCh38-Chr13-82147986-82148107 (TATC)8 AATC	RS	1	Missing	MH167205.1	
D16S539	8	TCCTCTTCCCTAGATCAATACAGACAGAGAGACAGGTGGATAGATAGATAGATAGATAGATAGATAGAT ATCATTGAAAGACAAAACAGAGATGGATGATAGATAC	D16S539 [CE 8]-GRCh38-Chr16-86352664-86352781 (GATA)8 86352692-G (rs563997442)	FS	1	No-Flank	MK570018.1	
D1S1656	17	TAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGGT GTGTGTGTGTTTAATTGTATGTATATATATTTGGTTCCCTAGTGATTCTATTTCTCTGAG	D1S1656 [CE 17]-GRCh38-Chr1-230769555-230769682 CCTA (TCTA)16 230769682-G (rs NA)	FS	1	n/a	Missing	Novel
D21S11	36	AAATATGTGAGTCAATTCCCCAAGTGAATTGCCTTCTATCTATCTATCTATCTATCTATCTATCTATCTATCT ATCTATCTATCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCT ATCCATATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCGTCTATCTAT	D21S11 [CE 36]-GRCh38-Chr21-19181939-19182111 (TCTA)11 (TCTG)6 (TCTA)3 TA (TCTA)3 TCA (TCTA)2 TCCA TA (TCTA)11	RS	1	Missing	Missing	Novel
D22S1045	15	CATTGGAATTCCCCAAACTGGCCAGTTCCTCTCCACCCTATAGACCCTGTCCTAGCCTTCTTATAGCTG CTATGGGGGCTAGATTTTCCCCGATGATAGTAGTCTCATTATTATTATTATTATTATTATTATTATT TATTACTATTATTGTTATAAAAATATTGCCAAT	D22S1045 [CE 15]-GRCh38-Chr22-37140181-37140357 (ATT)12 ACT (ATT)2 37140182-A (rs554502154)	FS	1	n/a	Missing	Novel
D2S1338	15	AAATGGCTTGGCCTTGCCTGCCTGCCTGCCTGCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTC CTTCCCTC	D2S1338 [CE 15]-GRCh38-Chr2-218014856-218014964 (GGAA)9 (GGCA)6	RS	1	EUR	Missing	
D2S441	9	CCAGGAACTGTGGCTCATCTATGAAAACTTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATC ATAACACCACAGCCACTTA	D2S441 [CE 9]-GRCh38-Chr2-68011918-68012017 (TCTA)9	RL	1	MEA/EUR/AFR/EAS	MH167314.1	
D3S1358	13	TTTGGGGGCATCTCTTATACTCATGAAATCAACAGAGGCTTGCATGTATCTATCTGTCTGTCTGTCTAT CTATCTATCTATCTATCTATCTATCTATCTATGAGACAGGGTCTTGCTC	D3S1358 [CE 13]-GRCh38-Chr3-45540691-45540820 TCTA (TCTG)3 (TCTA)9	RS	1	Missing	Missing	Novel
D3S1358	17	TTTGGGGGCATCTCTTATACTCATGAAATCAACAGAGGCTTGCATGTATCTATCTGTCTGTCTCTCTAT CTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATGAGACAGGGTCTTGCTC	D3S1358 [CE 17]-GRCh38-Chr3-45540691-45540820 TCTA (TCTG)2 TCTC (TCTA)13	RS	2	MEA/SAS	MK990348.1	
D3S1358	18	TTTGGGGGCATCTCTTATACTCATGAAATCAACAGAGGCTTGCATGTATCTATCTGTCTGTCTCTCTAT CTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATGAGACAGGGTCTTGCTC	D3S1358 [CE 18]-GRCh38-Chr3-45540691-45540820 TCTA (TCTG)2 TCTC (TCTA)14	RS	1	Missing	MK990350.1	
D3S1358	18	TTTGGGGGCATCTCTTATACTCATGAAATCAACAGAGGCTTGCATGTATCTATCTGTCTGTCTGTCTGT CTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATGAGACAGGGTCTTGCTC	D3S1358 [CE 18]-GRCh38-Chr3-45540691-45540820 TCTA (TCTG)4 (TCTA)13	RS	1	MEA/SAS	MH166987.1	
D3S1358	19	TTTGGGGGCATCTCTTATACTCATGAAATCAACAGAGGCTTGCATGTATCTATCTGTCTGTCTGTCTGT CTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATGAGACAGGGTCTTGCTC	D3S1358 [CE 19]-GRCh38-Chr3-45540691-45540820 TCTA (TCTG)4 (TCTA)14	RS	1	MEA	Missing	
D6S1043	25	GATCAATAGATTGATAGATCAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATA GATACATAGATAGATAGATAGATACATAGATAGATAGATAGATAGATAGATAGGATTTATTATGGGAAG TGGCTCATACAATTATGGAAGCTGAGAAATTTCACAATATGCCATCT	D6S1043 [CE 25]-GRCh38-Chr6-91740160-91740292 (ATCT)6 ATGT (ATCT)4 ATGT (ATCT)13	RS RL	1	Missing	MH166958.1	
D8S1179	17.1	TCTATCTATCTGTCTGTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTTA TTCCC	D8S1179 [CE 17.1]-GRCh38-Chr8-124894867-124894921 (TCTA)2 (TCTG)2 (TCTA)12 TCTTA	RS	1	Missing	Missing	Novel
D9S1122	13	AGATAACTGTAGATAGGTAGATCGATAGATAGATAGATAGATAGATAGATAGATAGATAGACAGATAGATAGA TATTAAT	D9S1122 [CE 13]-GRCh38-Chr9-77073809-77073880 TAGA TCGA (TAGA)8 CAGA (TAGA)2	RS	1	Missing	Missing	Novel
FGA	16.1	GCATATTTACAAGCTAGTTTCTTTCTTTCTTTTTTCTCTTTCTTTCTTTCTTTCTTTTCTTTCTTTCTT TCTCCTTCCTTCCTTTCTTCCTTTCTTTTTTGCTGG	FGA [CE 16.1]-GRCh38-Chr4-154587713-154587823 (GGAA)2 GGAG (AAAG)3 A (AAAG)5 AGAA AAAA (GAAA)3 154587760-A	RS	1	Missing	Missing	Novel
PentaE	16.4	AAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGA AAAGAAAAGAAAAGAAAATTGTAAGGAGTTTTCT	PentaE [CE 16.4]-GRCh38-Chr15-96830996-96831114 (TCTTT)16 TCTT	RS	2	MEA/SAS	Missing	
TH01	6	TGCAGGTCACAGGGAACACAGACTCCATGGTGAATGAATGAATGAATAAATGAATGAGGGAAATAAGG	TH01 [CE 6]-GRCh38-Chr11-2171056-2171127 (AATG)3 AATA (AATG)2	RS	1	Missing	Missing	Novel

Table 9.8: iiSNP variants (not highlighted in Flanking Region Report).

locus	iSNP & Variant Reference SNP	iSNP & Variant GRCh37 Position	Detected Bases	Format of UAS Flanking Region Report with new SNPs highlighted in red and bold font (in plain text in the original report)	UAS strand	Unreported SNP	GRCh37 SNV (Ref>Alt)	GRCh37 Position
rs1015250	rs6475200_NA_rs1015250_rs145984676	1823749_1823767_1823774_1823792	ATCC	AAAGGTTACTAAGTGATGGAGTTAGGAA AAGAACCCAGGTGTTTTATTCCTGTCCC AGTGATTTTTCA	plus			
rs1015250	rs6475200_NA_rs1015250_rs145984676	1823749_1823767_1823774_1823792	ATGC	AAAGGTTACTAAGTGATGGAGTTAGGAA AAGAACCCAGGTGTTTTATTGCTGTCCC AGTGATTTTTCA				
rs1015250	rs6475200_NA_rs1015250_rs145984676	1823749_1823767_1823774_1823792	GTCC	AAAGGTTACTAAGTGATGGAGTTAGGAA AAGAACCCAGGTGTTTTATTCCTGTCCC ACTGATTTTTCA		rs1307278892	G>C	1823783
rs1015250	rs6475200_NA_rs1015250_rs145984676	1823749_1823767_1823774_1823792	GTCC	AAAGGTTACTAAGTGATGGAGTTGGGAA AAGAACCCAGGTGTTTTATTCCTGTCCC AGTGATTTTTCA				
rs1109037	rs1109037_rs183533496_ NA_NA_NA_rs1109038	10085722_10085753_10085760_1008576 8_10085785_10085786	ACGGGA	CCAGTTTCTCCAGAGTGGAAAGACTTTC ATCTCGCACTGGCACGACCTTGAGACCC CGGGTTCTGATGAACTGGGAGG	plus			
rs1109037	rs1109037_rs183533496_ NA_NA_NA_rs1109038	10085722_10085753_10085760_ 10085768_10085785_10085786	ACGGGG	CCAGTTTCTCCAGAGTGGAAAGACTTTC ATCTCGCACTGGCACGACCTTGAGACCC CGGGTTCTGATGAACTGGGGGG				
rs1109037	rs1109037_rs183533496_ NA_NA_NA_rs1109038	10085722_10085753_10085760_ 10085768_10085785_10085786	GCGGGG	CCAGTTTCTCCGGAGCGGAAAGACTTTC ATCTCGCACTGGCACGACCTTGAGACCC CGGGTTCTGATGAACTGGGGGG		NA	T>C	10085726
rs1109037	rs1109037_rs183533496_ NA_NA_NA_rs1109038	10085722_10085753_10085760_ 10085768_10085785_10085786	GCGGGG	CCAGTTTCTCCGGAGTGGAAAGACTTTC ATCTCGCACTGGCACGACCTTGAGACCC CGGGTTCTGATGAACTGGGGGG				
rs1109037	rs1109037_rs183533496_ NA_NA_NA_rs1109038	10085722_10085753_10085760_ 10085768_10085785_10085786	GCGGGA	CCAGTTTCTCCGGAGTGGAAAGACTTTC ATCTCGCACTGGCACGACCTTGAGACCC CGGGTTCTGATGAACTGGGAGG				
rs1109037	rs1109037_rs183533496_ NA_NA_NA_rs1109038	10085722_10085753_10085760_ 10085768_10085785_10085786	GCGGGA	CCAGTTTCTCTGGAGTGGAAAGACTTTC ATCTCGCACTGGCACGACCTTGAGACCC CGGGTTCTGATGAACTGGGAGG		rs999755320	C>T	10085721
rs1294331	NA_rs1294331	233448415_233448413	GA	AGTATAGCTATGGATTTTTATTGAATTT TTG	minus	rs92280418	A>G	233448409
rs1294331	NA_rs1294331	233448415_233448413	GA	AGTATAGTTATGGATTTTTATTGAATTT TTG				
rs1294331	NA_rs1294331	233448415_233448413	GG	AGTGTAGTTATGGATTTTTATTGAATTT TTG				

locus	iSNP & Variant Reference SNP	iSNP & Variant GRCh37 Position	Detected Bases	Format of UAS Flanking Region Report with new SNPs highlighted in red and bold font (in plain text in the original report)	UAS strand	Unreported SNP	GRCh37 SNV (Ref>Alt)	GRCh37 Position
rs13182883	rs13182883	136633338	A	TGAGGGGAGGGGTCCCTTCTGGCCTAGT AGAGGGCCTGGCCT**A**CAGTGAGCATTCA AATCCTC**A**AGGAACAGGGTGGGGAGGTG GGACAAAGGCAGGAAGAAAGTAACGGAGA GCCTGGGGAGACA	plus	NA	G>A	136633317
rs13182883	rs13182883	136633338	A	TGAGGGGAGGGGTCCCTTCTGGCCTAGTA GAGGGCCTGGCCT**G**CAGTGAGCATTCAAA TCCTC**A**AGGAACAGGGTGGGGAGGTGGGA CAAAGGCAGGAAGAAAGTAACGGAGAGCC TGGGGAGACA				
rs13182883	rs13182883	136633338	G	TGAGGGGAGGGGTCCCTTCTGGCCTAGTA GAGGGCCTGGCCT**G**CAGTGAGCATTCAAA TCCTC**G**AGGAACAGGGTGGGGAGGTGGGA CAAAGGCAGGAAGAAAGTAACGGAGAGCC TGGGGAGACA				
rs338882	rs338882_rs746755126_ NA_NA	178690725_178690719_ 178690640_178690620	CCGC	GCCTGTGCA**C**ACACA**C**GTTTGGGACAAGG GCTGGATTCTTCGGCTGGGATGTCTCTCA GAGCTCTTGACTTGGTCCCTTTGGCTGGG GCTTGCC**G**TGAGGTGTGGGCTG**C**GCCA**C**G	minus			
rs338882	rs338882_rs746755126_ NA_NA	178690725_178690719_ 178690640_178690620	TCGC	GCCTGTGCA**T**ACACA**C**GTTTGGGACAAGG GCTGGATTCTTCGGCTGGGATGTCTCTCA GAGCTCTTGACTTGGTCCCTTTGGCTGGG GCTTGCC**G**TGAGGTGTGGGCTG**C**GCCA**C**G				
rs338882	rs338882_rs746755126_ NA_NA	178690725_178690719_ 178690640_178690620	TCGC	GCCTGTGCA**T**ACACA**C**GTTTGGGACAAGG GCTGGATTCTTCGGCTGGGATGTCTCTCA GAGCTCTTGACTTGGTCCCTTTGGCTGGG GCTTGCC**G**TGAGGTGTGGGCTG**T**GCCA**C**G		rs527589535	G>A	178890625
rs354439	NA_rs354439_rs564750466_ rs144284297_NA	106938442_106938411_ 106938406_106938390_106938362	CAGAC	TGTTCTGGTGGCTTCTCTTTCC**C**TTATGT ATCTCTCTCATGTATCACATTCCT**A**TTAA **G**CACAATATTCTGAAT**A**TCATTCAC**G**GTT TTCTATCGCAACCTG**C**AATTTGAGAGTTA AGAA	minus	NA	A>C	106938381
rs354439	NA_rs354439_rs564750466 _rs144284297_NA	106938442_106938411_ 106938406_106938390_106938362	CAGAC	TGTTCTGGTGGCTTCTCTTTCC**C**TTATGT ATCTCTCTCATGTATCACATTCCT**A**TTAA **G**CACAATATTCTGAAT**A**TCATTCAC**T**GTT TTCTATCGCAACCTG**C**AATTTGAGAGTTA AGAA				
rs354439	NA_rs354439_rs564750466_ rs144284297_NA	106938442_106938411_ 106938406_106938390_106938362	CAGCC	TGTTCTGGTGGCTTCTCTTTCC**C**TTATGT ATCTCTCTCATGTATCACATTCCT**A**TTAA **G**CACAATATTCTGAAT**C**TCATTCAC**T**GTT TTCTATCGCAACCTG**C**AATTTGAGAGTTA AGAA				
rs354439	NA_rs354439_rs564750466_ rs144284297_NA	106938442_106938411_ 106938406_106938390_106938362	CTGAC	TGTTCTGGTGGCTTCTCTTTCC**C**TTATGT ATCTCTCTCATGTATCACATTCCT**T**TTAA **G**CACAATATTCTGAAT**A**TCATTCAC**T**GTT TTCTATCGCAACCTG**C**AATTTGAGAGTTA AGAA				

locus	iSNP & Variant Reference SNP	iSNP & Variant GRCh37 Position	Detected Bases	Format of UAS Flanking Region Report with new SNPs highlighted in red and bold font (in plain text in the original report)	UAS strand	Unreported SNP	GRCh37 SNV (Ref>Alt)	GRCh37 Position
rs729172	rs729172	5606197	A	AGC**C**TCATTAATATGACCAAGGCTCCTCT GCAGAC**A**GAATGTATGTAACCG	minus			
rs729172	rs729172	5606197	C	AGC**C**TCATTAATATGACCAAGGCTCCTCT GCAGAC**C**GAATGTATGTAACCG				
rs729172	rs729172	5606197	C	AGC**T**TCATTAATATGACCAAGGCTCCTCT GCAGAC**C**GAATGTATGTAACCG		rs556717752	G>A	233448409
rs891700	rs12047255_rs891700	239881878_239881926	AG	GTGTTA**A**CAGTAAAACATTTTCATCAAAT TTCCATTCTTTTTTTTTTGAAGCCT**G**CTT GCATAGTTCTAAGG	plus			
rs891700	rs12047255_rs891700	239881878_239881926	GG	GTGTTA**G**CAGTAAAACATTTTCATCAAAT TTCCATTCTTTTTTTTTTGAAGCCT**G**CTT GCATAGTTCTAAGG				
rs891700	rs12047255_rs891700	239881878_239881926	GA	GTGTTA**G**CAGTAAAACATTTTCATCAAAT TTCCATTCTTTTTTTTTTGAAGCCT**A**CTT GCATAGTTCTAAGG				
rs891700	rs12047255_rs891700	239881878_239881926	GA	GTGTTA**G**CAGTAAAACATTTTCATCAAAT TTCCATTCTTTTTTTTTT**T**GAAGCCT**A**CT TGCATAGTTCTAAGG		rs543563536	T insertion	239881918
UAS Flanking Region Report					**Additional data on novel SNPs**			

Table 9.15: Sex chromosomal STR variants absent from STRait Razor database

locus	allele	Type	Obs	HGDP occurrence	GenBank	Novel
DYF387S1	35	RS	2	Missing	Missing	DYF387S1 [CE 35] -GRCh38-ChrY-23785347-23785500 (AAAG)3 GTAG (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)11 (AAAG)11
DYF387S1	35.1	RS	1	Missing	Missing	DYF387S1 [CE 35.1]-GRCh38-ChrY-23785347-23785500 (AAAG)3 GTAG (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)9 (AAAG)9 A (AAAG)4
DYF387S1	39	RS	1	Missing	Missing	DYF387S1 [CE 39] -GRCh38-ChrY-23785347-23785500 (AAAG)3 GTAG (GAAG)4 (AAAG)2 GAAG AAAG (GAAG)11 (AAAG)16
DYF387S1	40.1	FS	1	Missing	Missing	DYF387S1 [CE 41] -GRCh38-ChrY-23785347-23785500 (AAAG)3 GTAG (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)12 (AAAG)16 Del GAA
DYF387S1	43	RS	1	SAS(1)	Missing	DYF387S1 [CE 43] -GRCh38-ChrY-23785347-23785500 (AAAG)3 GTAG (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)13 (AAAG)17
DYS19	14	RS	1	SAS(1)	Missing	DYS19 [CE 14]-GRCh38-ChrY-9684267-9684443 (TCTA)12 CCTA (TCTA)2
DYS389II	27	RS	1	Missing	Missing	DYS389II [CE 27]-GRCh38-ChrY:12500448-12500633 (TAGA)10 (CAGA)3 N48 (TAGA)9 (CAGA)5
DYS389II	30	RS	3	MEA (14)	Missing	DYS389II [CE 30]-GRCh38-ChrY:12500448-12500633 (TAGA)11 (CAGA)3 N48 (TAGA)10 (CAGA)6
DYS389II	31	RS	1	Missing	Missing	DYS389II [CE 31]-GRCh38-ChrY:12500448-12500633 (TAGA)9 (CAGA)3 N48 (TAGA)15 (CAGA)4
DYS389II	31	RS	1	Missing	Missing	DYS389II [CE 31]-GRCh38-ChrY:12500448-12500633 (TAGA)10 (CAGA)3 N48 (TAGA)11 (CAGA)7
DYS389II	32	RS	1	AFR (4)	Missing	DYS389II [CE 32]-GRCh38-ChrY:12500448-12500633 (TAGA)11 (CAGA)3 N48 (TAGA)12 (CAGA)6
DYS390	24.3	RS	1	Missing	Missing	DYS390 [CE24.3]-GRCh38-ChrY-15162996-15163167(TAGA)4CAGA(TAGA)5 TGA (TAGA)6(CAGA)8
DYS391	8	RS	1	AFR (2)	Missing	DYS391 [CE 8] -GRCh38-ChrY-11982074-11982165 (TCTA)8
DYS391	13	RS	1	Missing	Missing	DYS391 [CE 13] -GRCh38-ChrY-11982074-11982165 (TCTA)13
DYS438	11	FS	1	n/a	MK990419.1	DYS438 [CE 11] -GRCh38-ChrY-12825876-12825984 (TTTTC)11 12825955-C (rs760613324)
DYS448	17.4	RS	1	Missing	Missing	DYS448 [CE 17.4]-GRCh38-ChrY:22218904-22219083 (AGAGAT)11 (ATAGAG)2 (AGATAG)3 ATAGAT AGAGAA (AGAGAT)5 AGAT AGAGAT
DYS481	19	RS	1	SAS(2)	Missing	DYS481 [CE 19]-GRCh38-ChrY-8558313-8558408 (CTT)19
DYS481	23	FS	1	n/a	Missing	DYS481 [CE 23]-GRCh38-ChrY-8558313-8558408 (CTT)23 8558404-C (rs rs769329768)
DYS481	31	RS	1	Missing	Missing	DYS481 [CE31]-GRCh38-ChrY-8558313-8558408(CTT)31
DYS522	9	RS	1	AFR (2) SAS(4)	Missing	DYS522 [CE 9]-GRCh38-ChrY-7547442-7547646 (ATAG)9
DYS533	9	FS	1	n/a	Missing	DYS533 [CE 9]-GRCh38-ChrY-16281285-16281417(TATC)9-16281313-G (rs NA)
DYS570	22	RS	1	MEA (6)	Missing	DYS570 [CE 22]-GRCh38-ChrY-6993158-6993271 (TTTC)5 TCTC (TTTC)16

locus	allele	Type	Obs	HGDP occurrence	GenBank	Novel
DYS612	30	FS	1	n/a	Missing	DYS612 [CE 36]-GRCh38-ChrY-13640705-13640861 (CCT)5 CTT (TCT)4 CCT (TCT)25 13640936-T (rs574356875)
DYS612	32	RS	1	Missing	Missing	DYS612 [CE 38]-GRCh38-ChrY-13640705-13640861 (CCT)5 CTT (TCT)4 CCT (TCT)4 CCT (TCT)22
DYS635	20	RS	1	EUR (1)	Missing	DYS635 [CE 20]-GRCh38-ChrY-12258829-12258975 (TAGA)9 (TACA)3 (TAGA)2 (TACA)2 (TAGA)4
DYS643	15	RS	2	AFR (2)	Missing	DYS643 [CE 15]-GRCh38-ChrY-15314082-15314203 (CTTTT)15
Y-GATA-H4	9	RS	1	AFR (1) SAS(1)	MK990441.1	Y-GATA-H4 [CE 9]-GRCh38-ChrY:16631636-16631759 (TCTA)9
DXS10074	15	RS	1	Missing	Missing	DXS10074 [CE 15]-GRCh38-ChrX-67757300-67757464 (AAGA)13 AAGG AAGA AAGG AAGA
DXS10074	16	RS	3	MEA (1)	Missing	DXS10074 [CE 16]-GRCh38-ChrX-67757300-67757464 (AAGA)15 (AAGG)2 AAGA
DXS10074	16	RS	1	Missing	Missing	DXS10074 [CE 16]-GRCh38-ChrX-67757300-67757464 (AAGA)14 AAGG AAGA AAGG AAGA
DXS10135	21	RS	1	MEA (1)	Missing	DXS10135 [CE 21]-GRCh38-ChrX-9338302-9338453 (AAGA)3 GAAA GGA (AAGA)11 AAGG (AAGA)3 AAGG AAGA AAAG
DXS10135	22	RS	1	MEA (4) EUR (1)	Missing	DXS10135 [CE 22]-GRCh38-ChrX-9338302-9338453 (AAGA)3 GAAA GGA (AAGA)12 AAGG (AAGA)3 AAGG AAGA AAAG
DXS7132	15	FS	1	n/a	Missing	DXS7132 [CE 15]-GRCh38-ChrX-65435623-65435778 (TAGA)15 65435707-G

Table 9.16: Y-STR structure linked to haplogroup.

Nomenclature	A	E1b1a	E1b1b	G	J1	J2	Q	R1a	T	N	L1a	O
DYS19 [CE 13]-GRCh38-ChrY-9684267-9684443 (TCTA)10 CCTA (TCTA)3			4									
DYS19 [CE 14]-GRCh38-ChrY-9684267-9684443 (TCTA)11 CCTA (TCTA)3			10	1	46	5			1		1	
DYS19 [CE 14]-GRCh38-ChrY-9684267-9684443 (TCTA)12 CCTA (TCTA)2	1											
DYS19 [CE 15]-GRCh38-ChrY-9684267-9684443 (TCTA)12 CCTA (TCTA)3		2	2	1	3	3	1			1		
DYS19 [CE 16]-GRCh38-ChrY-9684267-9684443 (TCTA)13 CCTA (TCTA)3				1				3	1			
DYS19 [CE 17]-GRCh38-ChrY-9684267-9684443 (TCTA)14 CCTA (TCTA)3		1										1
DYS437 [CE 14]-GRCh38-ChrY-12346255-12346431 (TCTA)8 (TCTG)2 (TCTA)4			16		49	2	1	3	2	1		1
DYS437 [CE 14]-GRCh38-ChrY-12346255-12346431 (TCTA)8 (TCTG)2 (TCTA)4 12346264-T (rs9786886)		3										
DYS437 [CE 15]-GRCh38-ChrY-12346255-12346431 (TCTA)9 (TCTG)2 (TCTA)4				1		4						
DYS437 [CE 16]-GRCh38-ChrY-12346255-12346431 (TCTA)10 (TCTG)2 (TCTA)4	1			2		2					1	
DYS389II [CE 27]-GRCh38-ChrY:12500448-12500633 (TAGA)10 (CAGA)3 N48 (TAGA)9 (CAGA)5						1						
DYS389II [CE 28]-GRCh38-ChrY:12500448-12500633 (TAGA)9 (CAGA)3 N48 (TAGA)11 (CAGA)5						1					1	
DYS389II [CE 29]-GRCh38-ChrY:12500448-12500633 (TAGA)10 (CAGA)3 N48 (TAGA)11 (CAGA)5				2	9	2	1					
DYS389II [CE 29]-GRCh38-ChrY:12500448-12500633 (TAGA)10 (CAGA)3 N48 (TAGA)12 (CAGA)4									1			
DYS389II [CE 29]-GRCh38-ChrY:12500448-12500633 (TAGA)11 (CAGA)3 N48 (TAGA)10 (CAGA)5						1						
DYS389II [CE 29]-GRCh38-ChrY:12500448-12500633 (TAGA)11 (CAGA)3 N48 (TAGA)11 (CAGA)4										1		
DYS389II [CE 29]-GRCh38-ChrY:12500448-12500633 (TAGA)9 (CAGA)3 N48 (TAGA)12 (CAGA)5				1								
DYS389II [CE 30]-GRCh38-ChrY:12500448-12500633 (TAGA)10 (CAGA)3 N48 (TAGA)11 (CAGA)6			6									
DYS389II [CE 30]-GRCh38-ChrY:12500448-12500633 (TAGA)10 (CAGA)3 N48 (TAGA)12 (CAGA)5	1				26			3				
DYS389II [CE 30]-GRCh38-ChrY:12500448-12500633 (TAGA)11 (CAGA)3 N48 (TAGA)10 (CAGA)6			3									
DYS389II [CE 30]-GRCh38-ChrY:12500448-12500633 (TAGA)11 (CAGA)3 N48 (TAGA)11 (CAGA)5					2	3						
DYS389II [CE 30]-GRCh38-ChrY:12500448-12500633 (TAGA)9 (CAGA)3 N48 (TAGA)13 (CAGA)5					1							
DYS389II [CE 31]-GRCh38-ChrY:12500448-12500633 (TAGA)10 (CAGA)3 N48 (TAGA)12 (CAGA)6		1	5		1							
DYS389II [CE 31]-GRCh38-ChrY:12500448-12500633 (TAGA)10 (CAGA)3 N48 (TAGA)13 (CAGA)5					6							
DYS389II [CE 31]-GRCh38-ChrY:12500448-12500633 (TAGA)11 (CAGA)3 N48 (TAGA)11 (CAGA)6		1										
DYS389II [CE 31]-GRCh38-ChrY:12500448-12500633 (TAGA)11 (CAGA)3 N48 (TAGA)12 (CAGA)5					2							
DYS389II [CE 32]-GRCh38-ChrY:12500448-12500633 (TAGA)10 (CAGA)3 N48 (TAGA)13 (CAGA)6			1									
DYS389II [CE 32]-GRCh38-ChrY:12500448-12500633 (TAGA)10 (CAGA)3 N48 (TAGA)14 (CAGA)5					1							
DYS389II [CE 32]-GRCh38-ChrY:12500448-12500633 (TAGA)11 (CAGA)3 N48 (TAGA)13 (CAGA)5					1							
DYS389II [CE 32]-GRCh38-ChrY:12500448-12500633 (TAGA)11 (CAGA)3 N48 (TAGA)14 (CAGA)4									1			
DYS389II [CE 31]-GRCh38-ChrY:12500448-12500633 (TAGA)10 (CAGA)3 N48 (TAGA)11 (CAGA)7			1									
DYS389II [CE 31]-GRCh38-ChrY:12500448-12500633 (TAGA)9 (CAGA)3 N48 (TAGA)15 (CAGA)4												1
DYS389II [CE 32]-GRCh38-ChrY:12500448-12500633 (TAGA)11 (CAGA)3 N48 (TAGA)12 (CAGA)6		1										
DYS481 [CE 19]-GRCh38-ChrY-8558313-8558408 (CTG)1 (CTT)19			1									
DYS481 [CE 21]-GRCh38-ChrY-8558313-8558408 (CTG)2 (CTT)20				3								
DYS481 [CE 22]-GRCh38-ChrY-8558313-8558408 (CTG)1 (CTT)22						3			1			
DYS481 [CE 23]-GRCh38-ChrY-8558313-8558408 (CTG)1 (CTT)23					4	4		2			1	
DYS481 [CE 23]-GRCh38-ChrY-8558313-8558408 (CTG)1 (CTT)23 8558404-C (rs769329768)			1									
DYS481 [CE 24]-GRCh38-ChrY-8558313-8558408 (CTG)1 (CTT)24		1	1		9	1		1	1	1		
DYS481 [CE 25]-GRCh38-ChrY-8558313-8558408 (CTG)1 (CTT)25			7		22							1
DYS481 [CE 26]-GRCh38-ChrY-8558313-8558408 (CTG)1 (CTT)26		1	4		14							
DYS481 [CE 27]-GRCh38-ChrY-8558313-8558408 (CTG)1 (CTT)27			2				1					

Nomenclature	A	E1b1a	E1b1b	G	J1	J2	Q	R1a	T	N	L1a	O
DYS481 [CE 28]-GRCh38-ChrY-8558313-8558408 (CTG)1 (CTT)28	1											
DYS481 [CE 31]-GRCh38-ChrY-8558313-8558408 (CTG)1 (CTT)31		1										
DYS533 [CE 9]-GRCh38-ChrY-16281285-16281417 (TATC)9			1	1								
DYS533 [CE 9]-GRCh38-ChrY-16281285-16281417 (TATC)9 16281313-G (rs NA)				1								
DYS533 [CE 10]-GRCh38-ChrY-16281285-16281417 (TATC)10	1		6	1		1						
DYS533 [CE 11]-GRCh38-ChrY-16281285-16281417 (TATC)11		3	5		36	5	1		1			1
DYS533 [CE 12]-GRCh38-ChrY-16281285-16281417 (TATC)12			4		12	2		3	1	1	1	
DYS533 [CE 13]-GRCh38-ChrY-16281285-16281417 (TATC)13					1							
DYS612 [CE 31]-GRCh38-ChrY-13640705-13640861 (CCT)5 CTT (TCT)4 CCT (TCT)20						1						
DYS612 [CE 32]-GRCh38-ChrY-13640705-13640861 (CCT)5 CTT (TCT)4 CCT (TCT)21		1						1				
DYS612 [CE 33]-GRCh38-ChrY-13640705-13640861 (CCT)5 CTT (TCT)4 CCT (TCT)22						1						
DYS612 [CE 34]-GRCh38-ChrY-13640705-13640861 (CCT)5 CTT (TCT)4 CCT (TCT)23			1		2	1						
DYS612 [CE 35]-GRCh38-ChrY-13640705-13640861 (CCT)5 CTT (TCT)4 CCT (TCT)24		2	5		4	1						
DYS612 [CE 36]-GRCh38-ChrY-13640705-13640861 (CCT)5 CTT (TCT)4 CCT (TCT)25	1		5		13	2	1	2				1
DYS612 [CE 36]-GRCh38-ChrY-13640705-13640861 (CCT)5 CTT (TCT)4 CCT (TCT)25 13640936-T (rs574356875)				1								
DYS612 [CE 37]-GRCh38-ChrY-13640705-13640861 (CCT)5 CTT (TCT)4 CCT (TCT)26			3	1	14	2			1	1		
DYS612 [CE 38]-GRCh38-ChrY-13640705-13640861 (CCT)5 CTT (TCT)4 CCT (TCT)27			1		13				1		1	
DYS612 [CE 38]-GRCh38-ChrY-13640705-13640861 (CCT)5 CTT (TCT)4 CCT (TCT)4 CCT (TCT)22			1	1								
DYS612 [CE 39]-GRCh38-ChrY-13640705-13640861 (CCT)5 CTT (TCT)4 CCT (TCT)28					2							
DYS612 [CE 40]-GRCh38-ChrY-13640705-13640861 (CCT)5 CTT (TCT)4 CCT (TCT)29					1							
DYS438 [CE 9]-GRCh38-ChrY-12825876-12825984 (TTTTC)9						6			2			
DYS438 [CE 9]-GRCh38-ChrY-12825876-12825984 (TTTTC)9						1						
DYS438 [CE 10]-GRCh38-ChrY-12825876-12825984 (TTTTC)10	1	1	15	3		1				1		1
DYS438 [CE 10]-GRCh38-ChrY-12825876-12825984 (TTTTC)10 12825955-C (rs760613324)					48							
DYS438 [CE 11]-GRCh38-ChrY-12825876-12825984 (TTTTC)11		2	1				1	3			1	
DYS438 [CE 11]-GRCh38-ChrY-12825876-12825984 (TTTTC)11 12825955-C (rs760613324)					1							
DYS635 [CE 19]-GRCh38-ChrY-12258829-12258975 (TAGA)9 (TACA)2 (TAGA)2 (TACA)2 (TAGA)4	1											
DYS635 [CE 20]-GRCh38-ChrY-12258829-12258975 (TAGA)10 (TACA)2 (TAGA)2 (TACA)2 (TAGA)4					8	2						1
DYS635 [CE 20]-GRCh38-ChrY-12258829-12258975 (TAGA)9 (TACA)3 (TAGA)2 (TACA)2 (TAGA)4					1							
DYS635 [CE 21]-GRCh38-ChrY-12258829-12258975 (TAGA)11 (TACA)2 (TAGA)2 (TACA)2 (TAGA)4		2	9	1	37	3			1	1		
DYS635 [CE 21]-GRCh38-ChrY-12258829-12258975 (TAGA)12 (TACA)1 (TAGA)2 (TACA)2 (TAGA)4					1							
DYS635 [CE 22]-GRCh38-ChrY-12258829-12258975 (TAGA)12 (TACA)2 (TAGA)2 (TACA)2 (TAGA)4		1	2	2	1	2			1			
DYS635 [CE 22]-GRCh38-ChrY-12258829-12258975 (TAGA)8 (TACA)2 (TAGA)2 (TACA)2 (TAGA)2 (TACA)2 (TAGA)4							1					
DYS635 [CE 23]-GRCh38-ChrY-12258829-12258975 (TAGA)13 (TACA)2 (TAGA)2 (TACA)2 (TAGA)4			4		1	1						
DYS635 [CE 23]-GRCh38-ChrY-12258829-12258975 (TAGA)9 (TACA)2 (TAGA)2 (TACA)2 (TAGA)2 (TACA)2 (TAGA)4								3				
DYS635 [CE 24]-GRCh38-ChrY-12258829-12258975 (TAGA)14 (TACA)2 (TAGA)2 (TACA)2 (TAGA)4			1								1	
DYS385 [CE 11]-GRCh38-ChrY-18639701-18639898 (TTTC)11	1							2		1		
DYS385 [CE 12]-GRCh38-ChrY-18639701-18639898 (TTTC)12	1			1	3	1		1				1
DYS385 [CE 13]-GRCh38-ChrY-18639701-18639898 (TTTC)13			2		37	5		1	1			
DYS385 [CE 14]-GRCh38-ChrY-18639701-18639898 (TTTC)14			2	1	8	3	1	2	1		1	
DYS385 [CE 15]-GRCh38-ChrY-18639701-18639898 (TTTC)15			2	2		1						1
DYS385 [CE 16]-GRCh38-ChrY-18639701-18639898 (TTTC)13 TTTA (TTTC)2									2			
DYS385 [CE 16]-GRCh38-ChrY-18639701-18639898 (TTTC)16		2	6	1	1	3	1			1		
DYS385 [CE 17]-GRCh38-ChrY-18639701-18639898 (TTTC)17		2	9		9	2					1	
DYS385 [CE 18]-GRCh38-ChrY-18639701-18639898 (TTTC)18		1	7		23	1						

Nomenclature	A	E1b1a	E1b1b	G	J1	J2	Q	R1a	T	N	L1a	O
DYS385 [CE 19]-GRCh38-ChrY-18639701-18639898 (TTTC)19		1	1		15							
DYS385 [CE 20]-GRCh38-ChrY-18639701-18639898 (TTTC)20					1							
DYF387S1 [CE 35]-GRCh38-ChrY-23785347-23785500 (AAAG)3 GTAG (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)11 (AAAG)11			2									
DYF387S1 [CE 35]-GRCh38-ChrY-23785347-23785500 (AAAG)3 GTAG (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)8 (AAAG)14			2									
DYF387S1 [CE 35]-GRCh38-ChrY-23785347-23785500 (AAAG)3 GTAG (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)9 (AAAG)13				1	2		1					
DYF387S1 [CE 35]-GRCh38-ChrY-23785347-23785500 (AAAG)3 GTAG (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)9 (AAAG)9 A (AAAG)4					1							
DYF387S1 [CE 36]-GRCh38-ChrY-23785347-23785500 (AAAG)3 GTAG (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)10 (AAAG)13	1				5							
DYF387S1 [CE 36]-GRCh38-ChrY-23785347-23785500 (AAAG)3 GTAG (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)11 (AAAG)12			1									
DYF387S1 [CE 36]-GRCh38-ChrY-23785347-23785500 (AAAG)3 GTAG (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)8 (AAAG)15			2								1	
DYF387S1 [CE 36]-GRCh38-ChrY-23785347-23785500 (AAAG)3 GTAG (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)9 (AAAG)14		1			15	1	1		1			
DYF387S1 [CE 37]-GRCh38-ChrY-23785347-23785500 (AAAG)3 GTAG (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)10 (AAAG)14	1			2	22	1			1			
DYF387S1 [CE 37]-GRCh38-ChrY-23785347-23785500 (AAAG)3 GTAG (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)11 (AAAG)13			4									
DYF387S1 [CE 37]-GRCh38-ChrY-23785347-23785500 (AAAG)3 GTAG (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)9 (AAAG)15		1	1		27	2		2		1		
DYF387S1 [CE 38]-GRCh38-ChrY-23785347-23785500 (AAAG)3 GTAG (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)10 (AAAG)15			2	2	12							
DYF387S1 [CE 38]-GRCh38-ChrY-23785347-23785500 (AAAG)3 GTAG (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)11 (AAAG)14			2		2	1		1				
DYF387S1 [CE 38]-GRCh38-ChrY-23785347-23785500 (AAAG)3 GTAG (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)9 (AAAG)16			2		4	1				1		1
DYF387S1 [CE 39]-GRCh38-ChrY-23785347-23785500 (AAAG)3 GTAG (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)10 (AAAG)16		2			2	1						
DYF387S1 [CE 39]-GRCh38-ChrY-23785347-23785500 (AAAG)3 GTAG (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)11 (AAAG)15		1	6		1	3		2				
DYF387S1 [CE 39]-GRCh38-ChrY-23785347-23785500 (AAAG)3 GTAG (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)13 (AAAG)13									1			
DYF387S1 [CE 39]-GRCh38-ChrY-23785347-23785500 (AAAG)3 GTAG (GAAG)4 (AAAG)2 GAAG (AAAG)1 (GAAG)11 (AAAG)16						1						
DYF387S1 [CE 40]-GRCh38-ChrY-23785347-23785500 (AAAG)3 GTAG (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)10 (AAAG)17		1	1		1							
DYF387S1 [CE 40]-GRCh38-ChrY-23785347-23785500 (AAAG)3 GTAG (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)11 (AAAG)16			2			2					1	
DYF387S1 [CE 40]-GRCh38-ChrY-23785347-23785500 (AAAG)3 GTAG (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)12 (AAAG)15						2						
DYF387S1 [CE 41]-GRCh38-ChrY-23785347-23785500 (AAAG)3 GTAG (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)10 (AAAG)18					1							
DYF387S1 [CE 41]-GRCh38-ChrY-23785347-23785500 (AAAG)3 GTAG (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)11 (AAAG)17												1
DYF387S1 [CE 41]-GRCh38-ChrY-23785347-23785500 (AAAG)3 GTAG (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)12 (AAAG)16 Del GAA						1						
DYF387S1 [CE 43]-GRCh38-ChrY-23785347-23785500 (AAAG)3 GTAG (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)13 (AAAG)17							1					
DYS448 [CE 17.4]-GRCh38-ChrY:22218904-22219083 (AGAGAT)11 (ATAGAG)2 (AGATAG)3 ATAGAT AGAGAA (AGAGAT)5 AGAT AGAGAT						1						
DYS448 [CE 17] -GRCh38-ChrY:22218904-22219083 (AGAGAT)10 (ATAGAG)2 (AGATAG)3 ATAGAT AGAGAA (AGAGAT)7												1
DYS448 [CE 18] -GRCh38-ChrY:22218904-22219083 (AGAGAT)10 (ATAGAG)2 (AGATAG)3 ATAGAT AGAGAA (AGAGAT)8											1	
DYS448 [CE 19] -GRCh38-ChrY:22218904-22219083 (AGAGAT)11 (ATAGAG)2 (AGATAG)3 ATAGAT AGAGAA (AGAGAT)8			2		5	1		1	2			
DYS448 [CE 20] -GRCh38-ChrY:22218904-22219083 (AGAGAT)11 (ATAGAG)2 (AGATAG)3 ATAGAT AGAGAA (AGAGAT)9	1			1			1					
DYS448 [CE 20] -GRCh38-ChrY:22218904-22219083 (AGAGAT)12 (ATAGAG)2 (AGATAG)3 ATAGAT AGAGAA (AGAGAT)8			12	1	41	1		2				
DYS448 [CE 20] -GRCh38-ChrY:22218904-22219083 (AGAGAT)13 (ATAGAG)2 (AGATAG)3 ATAGAT AGAGAA (AGAGAT)7		1	2									
DYS448 [CE 21] -GRCh38-ChrY:22218904-22219083 (AGAGAT)12 (ATAGAG)2 (AGATAG)3 ATAGAT AGAGAA (AGAGAT)9					3							
DYS448 [CE 21] -GRCh38-ChrY:22218904-22219083 (AGAGAT)13 (ATAGAG)2 (AGATAG)3 ATAGAT AGAGAA (AGAGAT)8		2				4				1		
DYS448 [CE 22] -GRCh38-ChrY:22218904-22219083 (AGAGAT)12 (ATAGAG)2 (AGATAG)3 ATAGAT AGAGAA (AGAGAT)10				1								
DYS448 [CE 22] -GRCh38-ChrY:22218904-22219083 (AGAGAT)13 (ATAGAG)2 (AGATAG)3 ATAGAT AGAGAA (AGAGAT)9						1						
DYS390 [CE 21]-GRCh38-ChrY-15162996-15163167 (TAGA)4 CAGA (TAGA)8 (CAGA)8		3		1								
DYS390 [CE 21]-GRCh38-ChrY-15162996-15163167 (TAGA)4 CAGA (TAGA)9 (CAGA)7	1											
DYS390 [CE 22]-GRCh38-ChrY-15162996-15163167 (TAGA)4 CAGA (TAGA)9 (CAGA)8				1	4	2					1	
DYS390 [CE 23]-GRCh38-ChrY-15162996-15163167 (TAGA)4 CAGA (TAGA)10 (CAGA)8			3	1	38	4	1		1			1
DYS390 [CE 24]-GRCh38-ChrY-15162996-15163167 (TAGA)4 CAGA (TAGA)10 (CAGA)9					1							
DYS390 [CE 24]-GRCh38-ChrY-15162996-15163167 (TAGA)4 CAGA (TAGA)11 (CAGA)8			13		5	2			1			
DYS390 [CE 25]-GRCh38-ChrY-15162996-15163167 (TAGA)4 CAGA (TAGA)12 (CAGA)8								3				
DYS390 [CE 26]-GRCh38-ChrY-15162996-15163167 (TAGA)4 CAGA (TAGA)13 (CAGA)8					1							
DYS390[CE24.3]-GRCh38-ChrY-15162996-15163167(TAGA)4CAGA(TAGA)5 TGA (TAGA)6(CAGA)8										1		

Table 9.18: HaploGrep 2 mitochondrial haplogroup predictions

SampleID	Haplogroup	short	Overall Rank	SampleID	Haplogroup	short	Overall Rank
C1	L4a1	L4	1	S11	R0a1a	R0	0.8801
C10	J1d1a	J	1	S12	H2a2a	H	1
C11	K1a4a1f	U	1	S13	H2a3	H	1
C12	U7a2a	U	0.8029	S14	J1b2a	J	0.9376
C13	R0a1a	R0	0.8784	S15	J2a2	J	0.8963
C14	J1b2a	J	0.983	S16	H2a2a	H	1
C15	H6	H	1	S17	R0a1a	R0	0.8556
C16	N1a3a	N	1	S18	U9a	U	0.8528
C17	J1b	J	1	S19	T2c1c1	T	0.8127
C18	J1d1a	J	1	S2	U3	U	0.9158
C19	T1a+152	T	0.9627	S21	L3i1a	L3	0.8916
C2	U7a2a	U	0.7538	S22	J1b8	J	0.8988
C20	J1b	J	1	S23	G1a	U	0.9713
C3	J1d1a	J	1	S24	H6b	H	0.956
C4	T1a+152	T	0.9442	S25	H1c+152	H	0.7737
C5	N1b1	N	0.904	S26	J2a2a1a	J	1
C7	J2a2a1+16311	J	1	S27	H57	H	0.6397
C9	K	U	0.735	S3	U3a	U	0.8154
E1	J1b2a	J	0.9376	S4	H6b	H	0.8236
E10	T1a	T	1	S5	U2c'd	U	0.7668
E11	L3e3	L3	1	S6	J1b	J	1
E12	H57	H	0.7482	S7	M1a1	M	0.8115
E13	H1e1a1	H	0.7854	S8	J2b	J	1
E15	HV1	H	0.7628	S9	H6b	H	0.956
E16	L0a1b2	L0	0.9434	W11	R0a2c	R0	0.7911
E17	T1a2b	T	0.8921	W12	J2a2b1	J	0.9179
E2	L0a1a	L0	0.9242	W13	I1a1	N	0.9493
E3	R0a1a	R0	0.8784	W14	L2a1b+143	L2	0.9723
E4	T2c1+146	T	0.9405	W15	R0a	R0	0.8553
E5	T1a	T	1	W16	U2d	U	0.6511
E6	H57	H	0.7482	W17	J2a2	J	1
E7	H57	H	0.7482	W18	J2a2a1+16311	J	1
E8	M57a	M	1	W19	L0a1c1	L0	0.9796
E9	H13b1+200	H	0.9375	W20	K1b2	U	0.965
N1	H6b	H	0.956	W21	H2a1	H	0.8647
N11	U5a1	U	0.901	W23	M6a1a	M	0.974
N2	L3x2a	L3	0.8736	W3	T2c1+146	T	0.9122
N3	T1a1'3	T	1	W4	N1b1	N	0.9132
N4	H57	H	0.6518	W5	N1a1'2	N	0.6937
N5	M1a1	M	0.8777	W6	R0a	R0	0.8553
N7	L4a1	L4	1	W7	R30b2	R0	0.8735
N9	J1b	J	1	W8	H2a1	H	0.8416
S1	U3	U	0.9158	W9	T1b	T	0.8097
S10	U8b1a2a	U	0.6936				

9.2 Materials Appendix

9.2.1 Kits

PrepFiler® Forensic DNA Extraction Kit (ThermoFisher Scientific)

PrepFiler®Lysis Buffer

PrepFiler® Magnetic Particles

PrepFiler® Wash Buffer A Concentrate

PrepFiler® Wash Buffer B Concentrate

PrepFiler® Elution Buffer

Oragene•DNA (OG-500) kit

QIAamp DNA Mini Kits (QIAGEN)

QIAamp Mini Spin Columns

Collection Tubes (2 ml)

Buffer AL, ATL, AW1, AW and AE

Proteinase K

GlobalFiler® PCR amplification kit

GlobalFiler® Master Mix

GlobalFiler® Primer Set

GlobalFiler® Allelic Ladder

DNA Control 007 (0.1 ng/µl)

Quantifiler® Human DNA Quantification Kit

Quantifiler® Human Primer Mix

Quantifiler® Human DNA Standard

Quantifiler® PCR Reaction Mix

Yfiler® Plus PCR Amplification Kit

Yfiler® Plus Master Mix
DNA Control 007 (2 ng/µl)
Yfiler® Plus Primer Set
Yfiler® Plus Allelic Ladder

ForenSeq™ DNA Signature Prep Kit

Box 1 contents:

Control DNA 2800M (2800M)
PCR1 Reaction Mix (PCR1)
Enzyme Mix (FEM)
DNA Primer Mix A (DPMA)
Library Normalization Additives 1 (LNA1)

Box 2 contents:

Library Normalization Additives 1 (LNA1)
Library Normalization Storage Buffer 2 (LNS2)
Library Normalization Wash 1 (LNW1)
HP3 2N-NaOH (HP3)
PCR2 Reaction Mix (PCR2)
Human Sequencing Control (HSC)
A5 Index Adapter(A501- A508)
R7 Index Adapter(R701- R712)

Box 3 contents:

Library Normalization Beads 1 (LNB1)
Resuspension Buffer (RSB)
Sample Purification Beads (SPB)

MiSeq FGxTM Reagent Kit

Box 1 contents:

Reagent cartridge
HT1

Box 2 contents:

SBS Solution (PR2) Bottle
MiSeq FGx Flow Cell

9.2.2 Chemical reagents and enzymes

Ethanol - absolute

Molecular Biology Grade water (Sigma-Aldrich)

Agarose (Sigma-Aldrich, powder form)

Tris/Borate/EDTA buffer

Ethidium Bromide (Sigma, 10 mg/mL)

Proteinase K (20 mg/mL)

EDTA (0.5 M EDTA)

TE buffer (10 mM Tris-HCL, pH 7.5: 0.1 mM EDTA)

11xPCR buffer

5 U/µl *Taq* DNA polymerase (Bioline)

2.5 U/µl *Pfu* DNA polymerase (ThermoFisher Scientific)

1M Tris base

11.1× PCR buffer, recipe as described in [Kauppi et al., 2009]

10% (w/v) SDS

1× Tris-Borate EDTA (TBE) with 0.5 µg/ml Ethidium Bromide (EtBr)

SeaKem® LE Agarose (Lonza)

Loading dye: Bromophenol Blue (Bioline)

BigDye® Terminator v3.1 Ready Reaction Mix (Life Technologies)

BigDye® Terminator v3.1 5× Sequencing Buffer (Life Technologies)

Hi-Di formamide (Applied Biosystems)

Exonuclease I (ExoI) (New England Biolabs)

Shrimp Alkaline Phosphatase (SAP) (New England Biolabs)

9.3 List of publications

1- Extensive geographical and social structure in the paternal lineages of Saudi Arabia revealed by analysis of 27 Y-STRs
2- Analysis of 21 autosomal STRs in Saudi Arabia reveals population structure and the influence of consanguinity
3- Massively parallel sequencing of autosomal STRs and identity-informative SNPs highlights consanguinity in Saudi Arabia [Accepted for publication]

PhD Thesis: Genetic Diversity and Population Structure of Saudi Arabia
Yahya M. Y. Khubrani

Contribution statements for published work in this thesis

Because all of the results chapters of this thesis are either published or under review as multi-author papers that do not carry author contribution statements, a contribution statement is provided here for clarity:

Yahya M. Khubrani carried out recruitment of Saudi- and UK-based DNA donors; all laboratory work; all data analysis; and contributed to research strategy, data interpretation, preparation of figures and tables, and writing of all the published work.

Pille Hallast carried out SNP calling and extraction from the HGDP whole-genome sequence data (Chapter 6).

Jon Wetton and Mark Jobling contributed to research strategy, data interpretation, preparation of figures and tables, and writing of the published work.

Forensic Science International: Genetics 33 (2018) 98–105

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsigen

Research paper

Extensive geographical and social structure in the paternal lineages of Saudi Arabia revealed by analysis of 27 Y-STRs

Yahya M. Khubrani[a,b], Jon H. Wetton[a,*], Mark A. Jobling[a,*]

[a] Department of Genetics & Genome Biology, University of Leicester, University Road, Leicester, UK
[b] Forensic Genetics Laboratory, General Administration of Criminal Evidence, Public Security, Ministry of Interior, Saudi Arabia

ARTICLE INFO

Keywords:
Saudi Arabia
Y-STRs
Haplogroup
Haplotype
Population structure

ABSTRACT

Saudi Arabia's indigenous population is organized into patrilineal descent groups, but to date, little has been done to characterize its population structure, in particular with respect to the male-specific region of the Y chromosome. We have used the 27-STR Yfiler® Plus kit to generate haplotypes in 597 unrelated Saudi males, classified into five geographical regions (North, South, Central, East and West). Overall, Yfiler® Plus provides a good discrimination capacity of 95.3%, but this is greatly reduced (74.7%) when considering the reduced Yfiler® set of 17 Y-STRs, justifying the use of the expanded set of markers in this population. Comparison of the five geographical divisions reveals striking differences, with low diversity and similar haplotype spectra in the Central and Northern regions, and high diversity and similar haplotype spectra in the East and West. These patterns likely reflect the geographical isolation of the desert heartland of the peninsula, and the proximity to the sea of the Eastern and Western areas, and consequent historical immigration. We predicted haplogroups from Y-STR haplotypes, testing the performance of prediction by using a large independent set of Saudi Arabian Y-STR + Y-SNP data. Prediction indicated predominance (71%) of haplogroup J1, which was significantly more common in Central, Northern and Southern groups than in East and West, and formed a star-like expansion cluster in a median-joining network with an estimated age of ~2800 years. Most of our 597 participants were sampled within Saudi Arabia itself, but ~16% were sampled in the UK. Despite matching these two groups by home sub-region, we observed significant differences in haplotype and predicted haplogroup constitutions overall, and for most sub-regions individually. This suggests social structure influencing the probability of leaving Saudi Arabia, correlated with different Y-chromosome compositions. The UK-recruited sample is an inappropriate proxy for Saudi Arabia generally, and caution is needed when considering expatriate groups as representative of country of origin. Our study shows the importance of geographical and social structuring that may affect the utility of forensic databases and the interpretation of Y-STR profiles.

1. Introduction

Saudi Arabia is the largest country in the Arabian Peninsula. Its population of ~32 million people is distributed highly non-uniformly (Fig. 1), with very low densities in its large desert areas, but high densities concentrated around a small number of cities. Its indigenous Arab people (~63% of the population; www.stats.gov.sa, accessed 12/07/17) are historically organized into geographically-differentiated patrilineal descent groups, or tribes [1], with a tradition of consanguinity [2]. This geographical and social organization might be expected to have an effect on patterns of genetic diversity, particularly regarding the male-specific region of the Y chromosome (MSY), which in turn could have implications in interpretation of DNA profiles.

Genetic studies on Saudi Arabia to date are limited. Exome sequencing of a set of samples from the Arabian Peninsula including Saudi individuals demonstrated relatively high inbreeding coefficients [3], consistent with a history of consanguineous marriage. A general analysis of Saudi Arabian mitochondrial DNA (mtDNA) diversity [4] showed a pattern of haplogroups similar to that of other Arabian Peninsula samples. In another mtDNA-based study [5] – the only example to divide Saudi Arabia sub-regionally – central, northern, western and southeastern sub-groups formed a single cluster in a multi-dimensional scaling (MDS) analysis when compared to other Arabian Peninsula samples, but also presented significant inter-group differences. Y-chromosome studies have analysed the seven Y-STRs defining the minimal haplotype [6], or haplogroup-defining SNPs together with 17 Y-STRs (Yfiler®) for one specific haplogroup [7]. The first of these [6] revealed lower diversity in Saudi Arabia than in populations from

* Corresponding authors at: Department of Genetics & Genome Biology, University of Leicester, University Road, Leicester LE1 7RH, UK.
E-mail addresses: jw418@le.ac.uk (J.H. Wetton), maj4@le.ac.uk (M.A. Jobling).

https://doi.org/10.1016/j.fsigen.2017.11.015
Received 21 July 2017; Received in revised form 20 November 2017; Accepted 24 November 2017
Available online 02 December 2017
1872-4973/

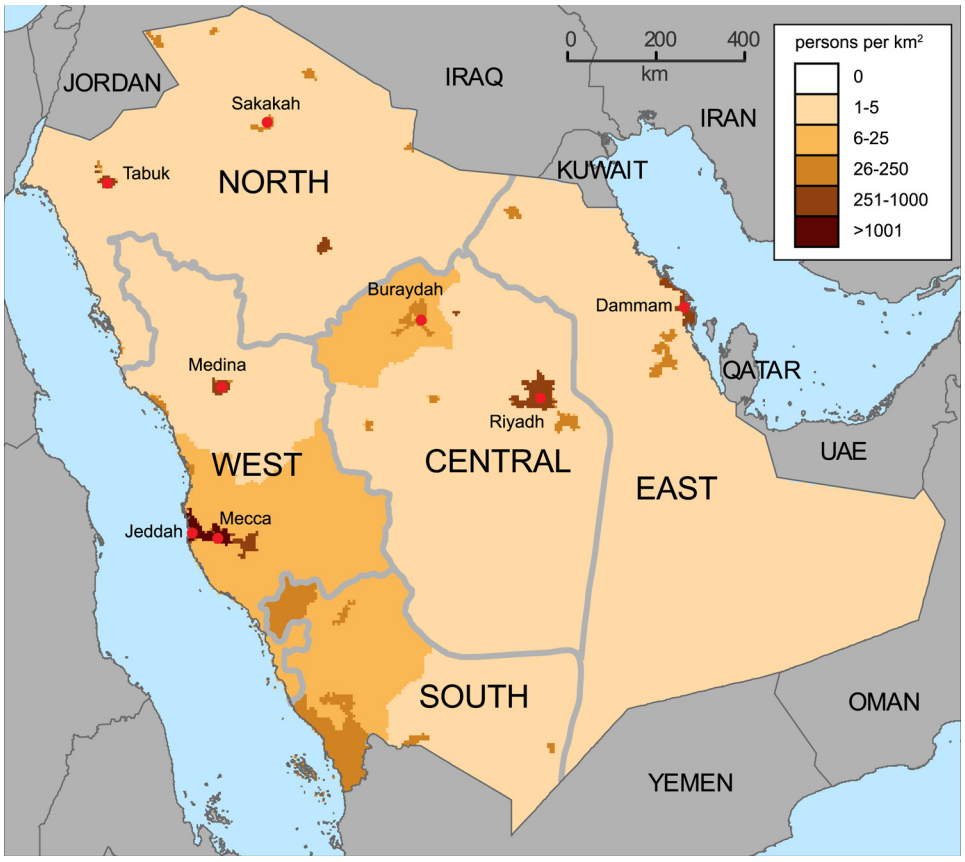


Fig. 1. Map of Saudi Arabia, showing population density and sub-regional divisions used in this study. Population density is indicated by shading as shown in the key, top right, and locations of some major cities are shown. Adapted from Global Rural-Urban Mapping Project (sedac.ciesin.columbia.edu/gpw/), under a Creative Commons 3.0 Attribution License. Administratively, Saudi Arabia is divided into 13 regions which we here consider as five larger geographical areas, namely: Central (Riyadh, Al-Qassim), Northern (Northern borders region, Tabuk, Al-Jawf and Hail), Southern (Asir, Jazan, Bahah and Najran), Eastern (Eastern province) and Western (Mecca and Medina).

outside the Arabian Peninsula, and affinity between Saudi Arabia and Yemen, which together were strongly differentiated from Oman and Dubai. It was speculated that this might be due to the influence of patrilineal descent and polygyny. The second study [7] showed that haplogroup J1 was the most prominent lineage (42%) in the Saudi Arabian sample studied, and that genetic distances based on haplogroup frequencies were relatively small among Arabian Peninsula samples. The focus of Y-STR typing on one lineage precludes any population-based conclusions on haplotype diversity from this study.

To date, therefore, while some general studies have been carried out, little has been done to characterize population structure within Saudi Arabia. Knowledge of any such structure is important in the interpretation of the significance of DNA-based forensic evidence, and in the construction of appropriate databases. Here, we use the 27 Y-chromosomal short-tandem repeats (Y-STRs) in the Yfiler® Plus kit to characterize haplotypes in 597 Saudi males sub-divided by geographical region. We consider the relationships of Y-chromosome diversity between regions within the country and also between Saudi Arabia and other surrounding populations. Finally, we compare the spectrum of Y-chromosome types in males recruited within Saudi Arabia with that of regionally-matched males recruited in the United Kingdom, to ask if social structuring also influences patterns of Y-haplotype diversity.

2. Materials and methods

2.1. DNA sampling

Five hundred and ninety-seven DNA samples were collected from indigenous Saudi Arabian males who were ethnically and linguistically Arabic. Of these, 503 were extracted from blood spots on FTA cards (Whatman, UK), sampled from individuals recruited within Saudi Arabia itself. The remaining 94 were extracted from buccal swabs [8], or from saliva samples via the Oragene kit (DNA Genotek), from Saudi males resident within the UK. In each case, males with ancestry (to the level of paternal great-grandfather) from five geographical subdivisions of the country shown in Fig. 1 (Central, Northern, Southern, Eastern, and Western) were sampled, and consideration of relatedness ensured that all sampled males were separated by at least three generations. Ethical review for recruitment and analysis was provided by the Saudi General Administration for Forensic Evidence and the University of Leicester Research Ethics Committee. Informed consent was provided by all participants.

2.2. DNA extraction and quantification

DNAs were extracted and purified from FTA blood-spot samples using a fully automated STARlet workstation (Hamilton) and the PrepFiler® Forensic DNA Extraction Kit (Thermo Fisher Scientific), starting from 1.2-mm diameter punches produced using the BSD100 Punching System (Microelectronic Systems). Buccal samples were extracted via QIAamp DNA Mini Kits on a QIAcube robotic workstation (Qiagen). All DNA samples were quantified using the Quantifiler® Human DNA Quantification Kit (Thermo Fisher Scientific) on an Applied Biosystems® 7500 Real-Time PCR System.

2.3. DNA amplification and fragment detection

The Yfiler® Plus PCR Amplification Kit was used to generate Y-chromosome haplotypes for the 27 STRs DYS19, DYS385a, DYS385b, DYF387S1a, DYF387S1b, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439, DYS448, DYS449, DYS456, DYS458, DYS460, DYS481, DYS518, DYS533, DYS570, DYS576, DYS627, DYS635, and Y-GATA H4. PCRs were conducted as recommended by the manufacturer on a Veriti (Thermo Fisher Scientific). Fragments were detected using an ABI3500 or ABI3130xl

Genetic Analyzer (Thermo Fisher Scientific) using the manufacturer's recommended protocols. GeneMapper IDX software V1.4 was used for allele calling and interpretation.

2.4. Haplogroup prediction and assessment of accuracy

Y-SNP haplogroups were predicted from Y-STR haplotypes using the online Y-DNA Haplogroup Predictor NevGen (http://www.nevgen.org), which is based on a previously-implemented Bayesian approach [9], with the additional consideration of pairwise correlation of alleles between different Y-STRs in the calculation of haplogroup probabilities.

We assessed the accuracy of this method by using a dataset of Y-STR and Y-SNP haplotypes from 743 self-declared Saudi Arabian males downloaded from FamilyTreeDNA (www.familytreedna.com). The NevGen method requires input of a standard set of Y-STRs, but as yet this does not include the Yfiler® Plus set. We therefore input the PowerPlex Y23 (PPY23) set of 23 Y-STRs with the two markers DYS549 and DYS643 marked as missing data, since these are present in PPY23 but not in Yfiler® Plus. Y-SNP resolution varied in the FamilyTreeDNA dataset, depending on whether custom Y-SNP typing or Y-chromosome resequencing had been carried out, and we standardized the level of haplogroup definition to a broad resolution of the thirteen haplogroups, A, B, E1b1b, E1b1a, G, H, J1, J2, L, Q, R1a, R1b and T (nomenclature according to [10]). For the set of 743 FamilyTreeDNA Y chromosomes, NevGen predicted a compatible haplogroup in 738 cases (99.3%). The five incompatible predictions all involved mis-prediction of haplogroup E sublineages as either hg I2a2a, or D1a. In applying NevGen to our own dataset, we therefore accepted all predictions except those for hgs I (n = 5) and D (n = 1), and in addition rejected predictions of haplogroups (hg N; n = 1, and hg O; n = 2) for which examples were not found in the FamilyTreeDNA dataset. We defined this total set of nine haplotypes as 'unpredicted'. To further understand the relationship between Y-STR haplotypes and predicted haplogroups, we also constructed a median-joining network combining haplotypes both from our dataset and from the FamilyTreeDNA dataset (Fig. S1); this demonstrates coherence of haplogroup prediction and haplotype clustering.

2.5. Median-joining networks

Median-joining networks [11] were constructed using the software Network 5.0 and Network Publisher (http://www.fluxus-engineering.com/sharenet.htm). In the case of intermediate alleles, repeat numbers were rounded to the nearest integer; constitutively duplicated loci (DYS385a,b; DYF387S1 a/b) were removed for network construction. Deletion alleles were coded '99' in input files, and thereby considered as missing data.

2.6. Forensic and population genetic parameters

For each sample or sub-sample, haplotype diversity was calculated using the formula $n(1-\Sigma p_i^2)/(n-1)$, where n is the sample size and p_i the frequency of the ith haplotype. Haplotype match probability (HMP) was estimated as the sum of squares of the haplotype frequencies. Discrimination capacity (DC) was calculated as the ratio between the number of distinct haplotypes and the total number of haplotypes in the sample.

Rst calculations based on Y-STR data and multi-dimensional scaling (MDS) plots [12] were carried out using comparative population data and the calculation tool within the online Y Haplotype Reference Database ([13]; https://yhrd.org/amova). The square plots produced by this approach were graphically adjusted to Euclidean space for display purposes.

Population differentiation tests based on predicted haplogroup frequencies were carried out within Arlequin [14] based on predicted haplogroups, using a method analogous to Fisher's exact test. Haplogroup-based gene diversity was also calculated in Arlequin.

Based on published comparisons of the performance of sequence- and STR-based dating [15], time-to-most-recent-common-ancestor (TMRCA) was estimated by the average-squared distance (ASD) method [16,17], using 23 STRs omitting the duplicated STRs (DYS385a,b; DYF387S1 a/b), and also using a reduced set of 18 omitting the Rapidly Mutating STRs (DYS449, DYS518, DYS570, DYS576 & DYS627). We used the modal haplotype as a root, and the mean pedigree mutation rate across STRs as measured in father-son pairs and available from yhrd.org/pages/resources/mutation_rates.

3. Results

The 27 Y-STRs contained in the Yfiler® Plus kit were amplified in DNAs from 597 Saudi Arabian males. Table S1 contains a full list of haplotypes, as well as other sample information; data are also available from YHRD, release 55 (YA004270 – Central; YA004271 – East; YA004272 – North; YA004273 – South; YA004274 – West). We also predicted haplogroups from each STR haplotype, using the prediction tool NevGen, and tested prediction accuracy based on a large independent set of Y-STR data and known Y-SNP haplogroups, as described in Materials and Methods.

3.1. Y-STR allele and haplotype diversity within Saudi Arabia

Considering allelic diversity, our dataset is characterized by a very high proportion of individuals (424/597; 71%) carrying intermediate alleles, in particular .2 alleles at DYS458; this is a known characteristic of haplogroup J1 [18], and immediately suggests that this haplogroup strongly predominates in the sample. In addition, twenty-four haplotypes carry .2 alleles at DYS627, and 22 of these are also predicted to belong to hg J1. The allele 23.2, present in a single individual of 'unpredicted' haplogroup, is not yet catalogued in YHRD (Release 54, June 2017). One copy of the duplicated STR DYF387S1 a/b carries a .2 allele in 14 haplotypes, and of these haplotypes, ten are predicted to belong to the generally rare hg B, and are therefore likely identical by descent – these have been observed previously in the same haplogroup [19]. Sporadic examples of other intermediate alleles are observed at DYS390, DYS392, DYS448, DYS449 (including allele 31.2, not listed in YHRD) and DYS570.

We observe one allele duplication among our data – a tri-allelic pattern at DYF387S1 a/b. We also observe one example of a deleted allele at DYS448 in a haplotype predicted to belong to haplogroup J1. DYS448 deletions have been described previously [20], though not (to our knowledge) in this haplogroup; deletions are recurrent, and driven by unequal recombination between flanking large repeats. Deletion is also observed in a single case for both copies of DYF387S1 a/b; five such cases are included in the 18,921 haplotypes in the YHRD.

Turning to haplotypes, Table 1 lists diversity summary statistics for the whole dataset, and for the five geographical subdivisions. The 597 males carry 543 distinct haplotypes, including 25 identical pairs, and two trios, providing a discrimination capacity of 95.3%. However, when we consider the sub-set of 17 Y-STRs making up the Yfiler® haplotype, we see a much higher level of haplotype sharing in the dataset: one haplotype, for example, is represented 22 times, and the discrimination capacity is only 74.7%; this compares, for example, with discrimination capacities of 98.5% and 95.7% for Yfiler® Plus and Yfiler® respectively in a US Caucasian sample [21].

Considering the five geographical subdivisions, comparisons of discrimination capacity at the levels of Yfiler® Plus and Yfiler® reveal striking differences in diversity between regions (Table 1). Values for Yfiler® Plus range from ~94% to 100%, but the range of values for Yfiler® haplotypes is much broader; in particular, the Central sample shows a discrimination capacity of only 72%, while the corresponding value for the Eastern sample is over 95%. In general, this points to relatively low diversity in the Central and Northern samples, with relatively high diversity in the East and West.

Table 1
Diversity summary statistics for Y-STR haplotypes in the entire sample set, and by geographical subdivisions.

		Yfiler® plus							Yfiler® comparison			
pop	n	No. unique hts	No. pair hts	No. trio hts	HMP	Haplotype Diversity	% unique hts	DC	HMP	Haplotype Diversity	% unique hts	DC
All	597	541	25	2	0.0018	0.9998	91.0	95.3	0.0048	0.9969	63.0	74.7
C	125	117	4		0.0085	0.9995	93.6	96.8	0.0245	0.9834	62.4	72.0
E	110	110			0.0091	1.0000	100.0	100.0	0.0099	0.9992	90.9	95.5
N	106	92	7		0.0107	0.9987	86.8	93.4	0.0176	0.9917	57.5	74.5
S	140	125	6	1	0.0081	0.9991	89.3	94.3	0.0106	0.9965	72.1	83.6
W	116	106	5		0.0094	0.9993	91.4	95.7	0.0113	0.9973	79.3	87.9

pop: population; n: number of individuals; ht: haplotype; HMP: haplotype match probability; DC: discrimination capacity; C: central; E: eastern; N: northern; S: southern; W: western. Yfiler® comparison: lists statistics considering only those STRs included in the 17-STR Yfiler® kit.

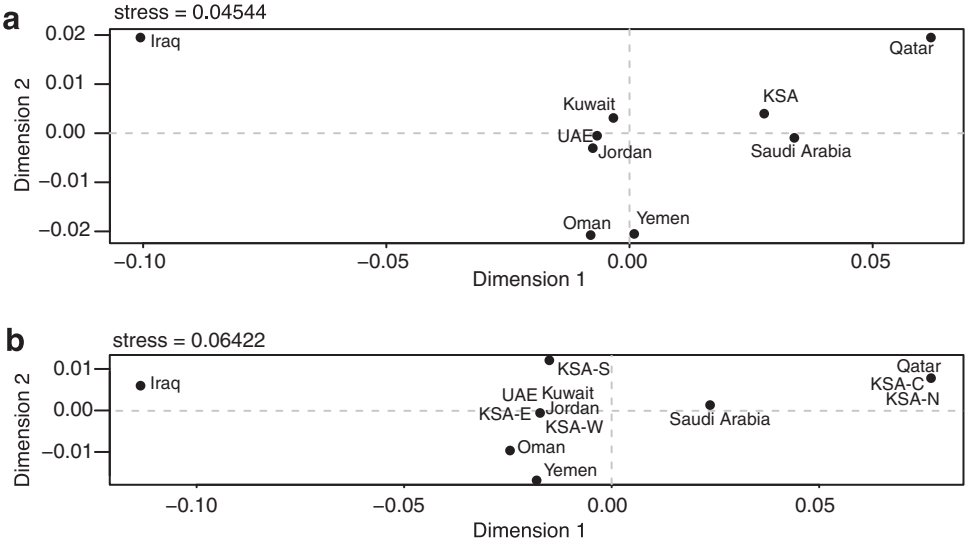


Fig. 2. Multidimensional scaling (MDS) plots of Arabian Peninsula populations based on Y-STR haplotypes.
Comparison with other datasets required reduction of the number of STRs to a shared set of nine.
a) Comparison of our total dataset (KSA) with an independent dataset of 106 Y-STR haplotypes from the same country ('Saudi Arabia' [6]), and other datasets from the Arabian peninsula: Iraq (n = 249; YHRD data), Jordan (n = 254 [43]), Kuwait (n = 645 [44,45]), Oman (n = 262 [6]), Qatar (n = 46 [46]), UAE (n = 684 [6,47]), Yemen (n = 375 [6], plus YHRD data).
b) Comparison of our dataset divided into five geographical sub-groups (KSA-N, -S, -E, -W, -C) with other Arabian Peninsula datasets as in part (a).

We compared our total Saudi Arabian population sample with other samples from the Arabian Peninsula, using multi-dimensional scaling based on Rst distances calculated from Y-STR haplotypes (Fig. 2a). In the first dimension of the plot, Iraq and Qatar lie at the extremes, with a cluster of other populations between them. Our Saudi Arabian sample (KSA) lies midway between this cluster and Qatar, and close to a previously published Saudi Arabian sample [6]. However, when we subdivide our sample into its five geographically-defined subsamples, two of these (Northern and Central) overlap with each other and cluster with Qatar, whereas the Eastern and Western subsamples overlap with each other and show affiliation with the major cluster of populations in the middle of the plot (Jordan, Kuwait, Oman, UAE and Yemen). The Southern sample lies at a similar first-dimension position to these, but is shifted in the second dimension of the plot, suggesting a distinct haplotype distribution to that of the Eastern and Western samples. Inclusion of population samples [19,22–26] from a wider surrounding geographical region (Fig. S2) does not change these relationships substantively.

3.2. Analysis of diversity via network analysis and haplogroup prediction

In order to understand the relationships between Y-STR haplotypes in the dataset, we constructed a median-joining network (Fig. 3a). Based on NevGen predictions (Materials & Methods), haplogroups were assigned to haplotypes within the network (Table S1). Most predicted haplogroups form coherent clusters, with the exception of haplogroup E1b1b, which forms two well-separated clusters, possibly indicating distinct sub-lineages that cannot be reliably distinguished by the prediction method used; the same split of haplogroup E1b1b is seen in a network containing the FamilyTreeDNA haplotypes of known haplogroup (Fig. S1). Network sub-structures for most haplogroups are generally extended, although a cluster of related haplotypes exists among the predicted haplogroup E1b1b chromosomes. However, the network's major feature is a central star-like cluster of closely related haplotypes assigned to haplogroup J1 (71% of the total sample), suggesting a recent expansion for this set of lineages. We estimated the TMRCA of this cluster using the average-squared distance method and a mean pedigree-based STR mutation rate. Considering the total set of 23 non-duplicated Y-STRs, this yields a TMRCA of 2494 ± 487 years; removal of rapidly-mutating markers, which might be expected to bias the estimate, reduces the number of STRs to 18 and increases the age slightly to 2754 ± 389 years. Application of the same methods to the FamilyTreeDNA dataset of confirmed haplogroup J1 haplotypes yields very similar estimates, for example 2783 ± 394 years for the reduced set of 18 Y-STRs. It is worth noting that use of the so-called 'evolutionary' average mutation rate of 6.9×10^{-4} per STR per generation [27] yields greatly elevated and very divergent TMRCA estimates for 23 and 18 Y-STRs in our own dataset of 19,835 ± 3874 years and 9809 ± 1916 years respectively.

Fig. S3 shows the same network, but with haplotypes coloured by region of origin. There is little evidence from inspecting this network of geographical substructuring, although the haplogroup E1b1b subcluster mentioned above is mostly formed by Western samples. Table 2 and Fig. 3b present predicted haplogroup distributions for the geographically defined sub-samples. One striking feature is the difference in the frequency of predicted haplogroup J1 in the Northern + Central + Southern samples (93% collectively) than that in the Eastern + Western pair (50%; exact test $p < 0.001$); on the other hand, the latter pair has a significantly higher frequency of predicted haplogroup E1b1b (19% vs 6%; exact test $p < 0.001$). Considering gene

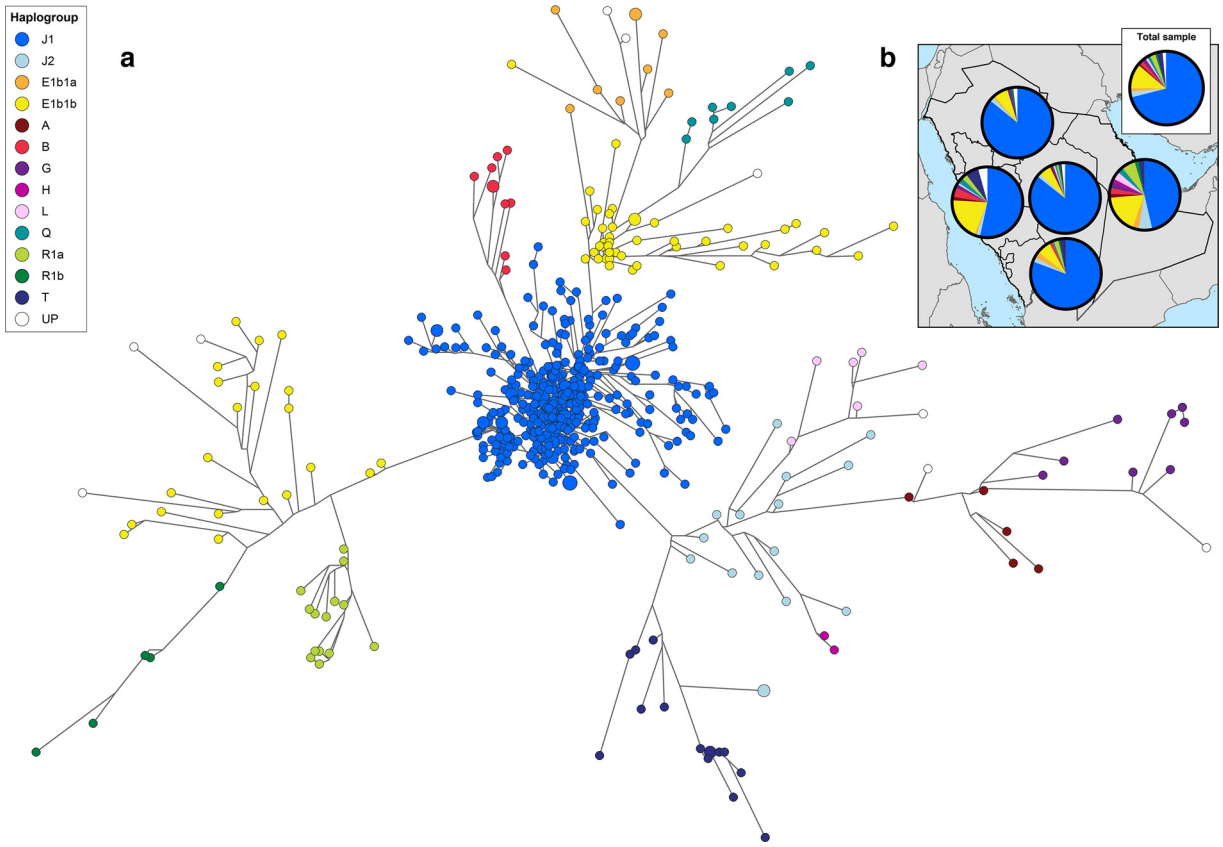


Fig. 3. Median-joining network of Y-STR haplotypes, and geographical distribution of predicted haplogroups.
a) Median-joining network for 597 Saudi Arabian haplotypes, constructed from data on 23 Y-STRs. Circles represent haplotypes, with area proportional to sample size, and lines between them proportional to the number of mutational steps. Colours represent haplogroups given in the key, top left. b) Map showing distributions of predicted haplogroups in five regional samples as pie-charts, not to scale; haplogroup distribution in the overall sample is represented in the pie-chart inset top right. Colours of sectors indicate haplogroups as shown in the key.

diversity values from haplogroup frequencies (Table 2), the Northern + Central pair shows significantly lower diversity than the Southern sample, which in turn is significantly lower than the Eastern + Western pair ($p < 0.05$).

3.3. Comparison of cohorts recruited in Saudi Arabia and the UK

Of our 597 samples, 94 (~16%) were recruited not in Saudi Arabia itself, but in the UK. To ask whether place of recruitment influenced the spectrum of haplotypes observed, we compared Y-STR haplotype diversities in the two differently recruited samples (Table S2), considering the same parameters as in Table 1. In the Saudi-recruited sample, both the proportion of unique haplotypes (~90%) and the discrimination capacity (~95%) are similar to the corresponding values in the cohort as a whole (~91% and ~95%). However, the UK-recruited sample shows much higher values – ~98% and 99% respectively, indicating that this mode of recruitment is sampling a more diverse subset of the Saudi Arabian population, despite the two groups being geographically matched. Predicted haplogroup frequencies in the Saudi- and UK-recruited samples (Fig. 4a) are significantly different (exact test $p < 0.001$), including a much higher frequency of predicted haplogroup J1 in the former. Similar comparisons at the sub-regional level (Fig. 4a) show significant differences between predicted haplogroup frequencies (Table S3) for the Saudi- and UK-recruited samples from Central, Western, and Southern regions ($p \leq 0.02$). For the North, the difference is not significant, probably due to the small sample size (n = 10) of the UK-recruited sample. For the East, however, the corresponding sample size is larger (n = 17), and the lack of significant difference ($p = 0.795$) probably indicates true homogeneity of the Saudi- and UK-recruited samples for this region.

4. Discussion

In this study, we have determined the Yfiler® Plus haplotypes of a set

Table 2
Predicted haplogroup distributions and diversities in the entire sample set, and by geographical subdivisions.

		Predicted haplogroup													
pop	n	A	B	E1b1a	E1b1b	G	H	J1	J2	L	Q	R1a	R1b	T	UP	h ± s.d.
All	597	5	10	9	66	8	2	424	16	6	8	14	5	15	9	0.481 ± 0.024
C	125	1	0	0	7	1	0	107	2	1	1	1	1	1	2	0.265 ± 0.052
E	110	2	3	3	20	3	2	51	7	4	3	7	2	3	0	0.745 ± 0.037
N	106	0	0	1	7	2	0	91	2	0	0	0	1	0	2	0.260 ± 0.056
S	140	0	2	4	9	0	0	113	3	0	1	3	1	4	0	0.344 ± 0.052
W	116	2	5	1	23	2	0	62	2	1	3	3	0	7	5	0.671 ± 0.041

pop: population; n: number of individuals; UP: unpredicted haplogroup; h: gene diversity; s.d.: standard deviation; C: central; E: eastern; N: northern; S: southern; W: western.

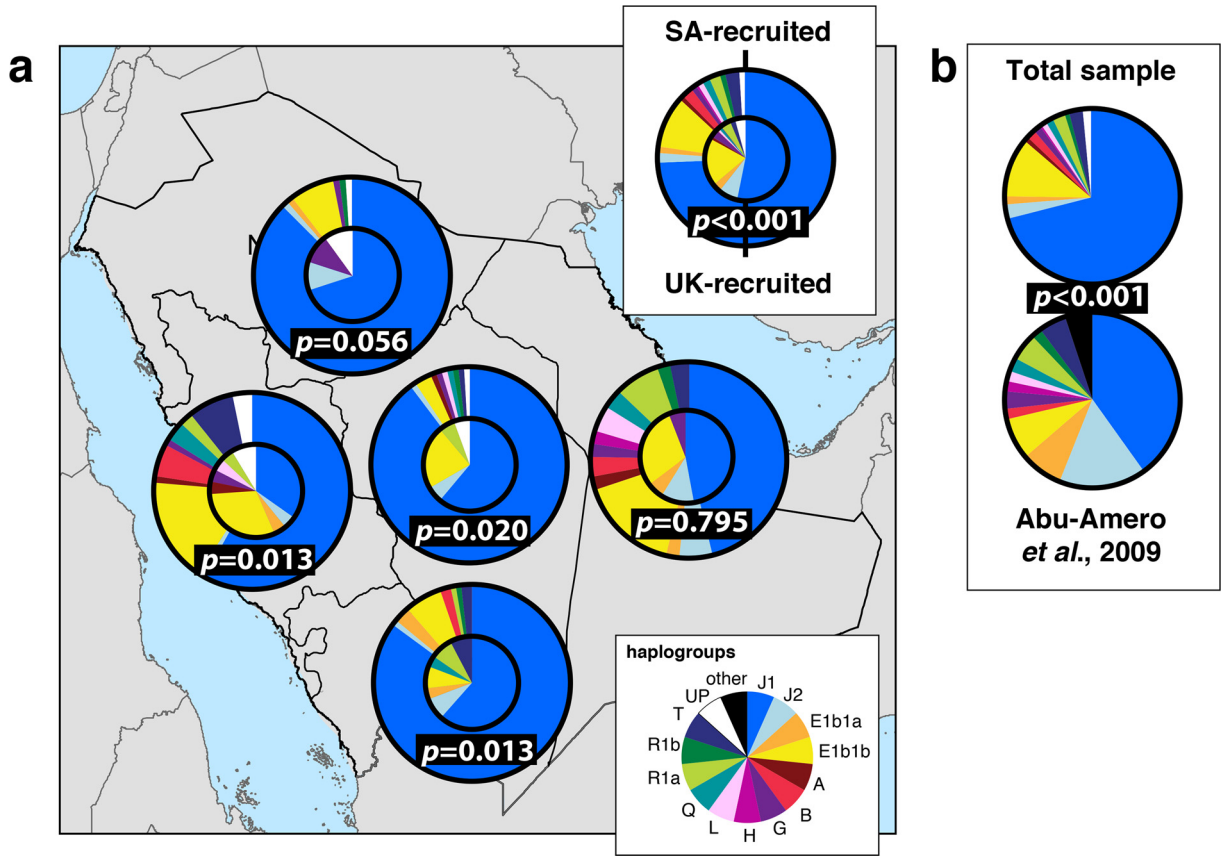


Fig. 4. Comparison of Saudi- and UK-recruited cohorts by frequency of predicted haplogroups.
a) Map showing distributions of predicted haplogroups in the five geographical regions, each divided into Saudi-recruited (outer pie-chart) and UK-recruited (inner pie-chart) samples. Pie-charts are not to scale. Haplogroup distributions in Saudi-recruited and UK-recruited samples for the total dataset are shown in the pie-chart inset top right. In each case, the *p*-value of a population differentiation test between Saudi- and UK-recruited samples is given. Colours of sectors indicate haplogroups as show in the key bottom right. b) Comparison of the haplogroup distribution in the total dataset (top) with that in published data [7] (bottom), with the *p*-value of the population differentiation test given.

of 597 Saudi Arabian males, and also considered how haplotype composition is affected by division into five geographically-defined subgroups, and by two different countries of recruitment (Saudi Arabia itself, and the United Kingdom).

The Yfiler® Plus system provides a discrimination capacity of 95.3% in the overall sample, which, while lower than that for US Caucasian, US Hispanic and African-American samples [21], exceeds that for an US Asian sample (94.4%). However, the added value of using the extended set of 27 Y-STRs contained in Yfiler® Plus is clearly demonstrated by a comparison with the 17 STRs defining the Yfiler® system – discrimination capacity in the Saudi Arabian sample falls much more markedly than in the US samples, to only 74.7%. This suggests that, despite our care in avoiding related males, there are many individuals in the sample whose haplotypes are similar because of deeper patrilineal descent from shared ancestors. This probably reflects a general property of many Middle Eastern populations: of all global regions, the Middle East was previously shown to exhibit the greatest difference between diversity assessed by Yfiler® STRs and RM-YSTRs [28], and the lowest Yfiler® discrimination capacity (~84%). We note that the benefits of using Yfiler Plus compared to Yfiler have also been reported in low-diversity patrilineal groups in East Africa [19].

As well as determining Y-STR haplotypes, we have predicted Y-SNP haplogroups from these haplotypes. A number of prediction methods exist, taking different approaches including phylogenetic trees with STR mutation rates (YPredictor; http://predictor.ydna.ru/), machine learning [29], partitional clustering [30], simple allele frequencies [31], and Bayesian allele frequencies ([9]; NevGen: http://www.nevgen.org). Evaluating prediction methods is not straightforward because some have been produced by the genetic genealogy community and are therefore not published in mainstream journals [9,31] or peer reviewed (YPredictor; NevGen); exact methodology is sometimes unclear. There has been debate about the accuracy of prediction; for example, criticism [32] of one widely used method [9], was counter-criticized [33] for using only 7 Y-STRs in evaluation. It seems clear that the larger the number of STRs, the better, and here we have used 21 STRs from the Yfiler® Plus set. In addition, any prediction method is only as good as the Y-SNP + Y-STR comparative datasets it uses for training or classification, and sometimes these are not well described – comparative datasets that are too small, or that do not include populations appropriate to the sample being predicted, may give unreliable results. In order to address this problem, we used a large set of independent Y-SNP + Y-STR data from the same population (Saudi Arabia) as that under study to test prediction performance. The chosen method, NevGen, performs well, and provides a > 99% accuracy in our sample, but we recognize the need to undertake SNP typing for definitive haplogroup determination [34].

The median-joining network of haplotypes (Fig. 3a) exhibits a large central star-like cluster that corresponds to predicted haplogroup J1, and contains many of the identical or highly similar haplotypes. Such features are commonly interpreted as past male-lineage expansions [35]. Star-like features of haplotypes comprising haplogroup J1 have been reported before in specific Arabian populations [36] and in broader Middle Eastern samples [37]. Interpretations of its origins initially focused on the 7th-century Muslim expansion [38], and were supported by some later studies [39], but some other authors have interpreted it in terms of much earlier spread in the Neolithic [37,40] followed by Bronze Age expansion in the Arabian Peninsula [37]. The age of the expansion is clearly crucial, and this in turn is affected by the

method chosen, but most strongly by the choice of mutation rate. The mean 'evolutionary' rate of 6.9×10^{-4} mutations per generation [27] has been widely used [7,37,40], and has been reported as performing better the directly-determined 'pedigree' rate for the dating of ancient events such as the coalescence of the whole Y phylogeny [15,41]. However, this rate was estimated from a relatively small set of 7–10 STRs [27], not including RM-YSTRs, so certainly cannot be universally applied. Furthermore, for haplogroups showing star-like patterns in networks, and for which Y-chromosome resequencing data indicate recent TMRCAs (< 10 thousand years), the pedigree mutation rate has been shown to perform best [15]. We therefore used this rate, and obtain a TMRCA for predicted haplogroup J1 of around 2800 years. This is several-fold younger than published estimates [7,37,40], and if correct, suggests late Bronze Age dispersion, possibly followed by later spread during the Islamic expansion.

Aside from the dominant predicted haplogroup J1, the range of other predicted lineages (Tables 2, and S3) is similar to that seen in a previous Y-SNP-based study of Saudi Arabia [7]. However, the frequency of these haplogroups differs very significantly between our sample and the published sample [7] (Fig. 4b, Table S3). This suggests that there is considerable Saudi Arabian heterogeneity, and that the sub-populations sampled in these two studies are very different. Such heterogeneity is confirmed when we subdivide our sample into five geographical regions. In both Y-STR- (Fig. 2b, Table 1) and predicted-haplogroup-based (Fig. 3b, Table 2) comparisons, the Central and Northern regions are highly similar in composition, and, surprisingly, the Eastern and Western samples are also highly similar. The Central + Northern pair is highly diverged from the Eastern + Western pair. The Southern region is somewhat distinct from all other regions with respect to both its spectrum of predicted haplogroups (Fig. 3b, Table 2) and Y-STR haplotypes (Fig. 2b). Considering discrimination capacity at the level of Yfiler® (Table 1), the Central and Northern samples show similarly low values, with higher and increasing values in the Southern, Western and Eastern samples respectively. Similar results are shown by comparing predicted haplogroup distributions (Table 2). The low diversity and similarity of Central and Northern areas reflect their relative geographical isolation within the desert heartland of the country, and possible bottleneck associated with the onset of desertification around 3000 years ago [42]. By contrast, the relatively high diversity of the Eastern and Western areas reflects their closeness to the sea and outside influences from other populations that may historically have brought in migrants.

While most of our 597 participants were contacted and recruited within Saudi Arabia itself, about 16% were recruited in the UK. Despite attempting to match the two groups by geographical sub-region, there are significant differences in the haplotype and predicted haplogroup constitutions of these two groups overall (Fig. 4a, Tables S2 and S3), and for the Central, Southern and Western sub-groups in particular. The UK-recruited sample size for the Northern region is very small so it is hard to draw any conclusion, while for the Eastern region the two differently-recruited sub-samples seem genuinely similar in composition. Taken together, this suggests that there is social structure in the country, which influences the probability of males leaving Saudi Arabia and undertaking study in the UK, and that this structure correlates with different sub-groups having different haplotype compositions. This social structuring appears to be less marked in the East of the country than elsewhere. It means that the UK-recruited sample is an inappropriate proxy for Saudi Arabia generally, and indicates that caution is needed when considering expatriate groups as representative of their country of origin.

The strong geographical and social structure we have observed in Saudi Arabia has important implications for the interpretation of Y-STR profiles in casework. Geographically appropriate databases must be used for assessment of evidential weight, and more work should be done to understand the social structuring reflected in the two differently recruited cohorts. Given the patrilineal descent structures of the tribal system, analysis of tribal names and surnames together with Y haplotypes should be illuminating. Other tribally-based Middle Eastern countries may also show marked population structure.

It will also be important to ask whether autosomal STR diversity is affected by the same factors that give rise to low diversity and a high degree of population structure among Y-haplotypes in Saudi Arabia. Sex-bias in population structure could also be addressed using maternally-inherited mtDNA. Indeed, inspection of a published study in which the country is subdivided [5] shows that the mtDNA haplogroup spectrum of the Central region differs significantly from those of other regions except the North ($p \leq 0.05$, Bonferroni-corrected; our analysis). This indication of structuring among maternal lineages, as well as paternal lineages, suggests that further analysis in our samples would be worthwhile.

Conflicts of interest

None.

Acknowledgments

YMK was supported by the Saudi Arabian Ministry of Interior, and by a PhD studentship grant from the Saudi Arabian Cultural Bureau, London. We thank members of the Forensic Genetics Laboratory, General Administration of Criminal Evidence, Riyadh, for assistance, in particular Ahmed Z. Asiri, Mohammed S. Asiri, Rashed H. AlSheal, Obaid G. AlAsaadi, Khalid Y. AlZahrani and Fahad S. AlRakaf. We also thank Milos Cetkovic Gentula and Aco Nevski of NevGen for running batch haplogroup predictions for us, and Jon Kyte and Thermo Fisher for their support.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.fsigen.2017.11.015.

References

[1] M. Al-Rasheed, A History of Saudi Arabia, Cambridge University Press, Cambridge, 2010.
[2] L. Al-Gazali, H. Hamamy, S. Al-Arrayad, Genetic disorders in the Arab world, BMJ 333 (2006) 831–834.
[3] E.M. Scott, A. Halees, Y. Itan, E.G. Spencer, Y. He, M.A. Azab, S.B. Gabriel, A. Belkadi, B. Boisson, L. Abel, A.G. Clark, C. Greater Middle East Variome, F.S. Alkuraya, J.L. Casanova, J.G. Gleeson, Characterization of Greater Middle Eastern genetic variation for enhanced disease gene discovery, Nat. Genet. 48 (2016) 1071–1076.
[4] K.K. Abu-Amero, A.M. Gonzalez, J.M. Larruga, T.M. Bosley, V.M. Cabrera, Eurasian and African mitochondrial DNA influences in the Saudi Arabian population, BMC Evol. Biol. 7 (2007) 32.
[5] K.K. Abu-Amero, J.M. Larruga, V.M. Cabrera, A.M. Gonzalez, Mitochondrial DNA structure in the Arabian Peninsula, BMC Evol. Biol. 8 (2008) 45.
[6] F. Alshamali, L. Pereira, B. Budowle, E.S. Poloni, M. Currat, Local population structure in Arabian Peninsula revealed by Y-STR diversity, Hum. Hered. 68 (2009) 45–54.
[7] K.K. Abu-Amero, A. Hellani, A.M. Gonzalez, J.M. Larruga, V.M. Cabrera, P.A. Underhill, Saudi Arabian Y-Chromosome diversity and its relationship with nearby regions, BMC Genet. 10 (2009) 59.
[8] T.E. King, S.J. Ballereau, K. Schürer, M.A. Jobling, Genetic signatures of coancestry within surnames, Curr. Biol. 16 (2006) 384–388.
[9] T.W. Athey, Haplogroup prediction from Y-STR values using a Bayesian-allele-frequency approach, J. Genet. Geneal. 2 (2006) 34–39.
[10] T.M. Karafet, F.L. Mendez, M. Meilerman, P.A. Underhill, S.L. Zegura, M.F. Hammer, New binary polymorphisms reshape and increase resolution of the human Y-chromosomal haplogroup tree, Genome Res. 18 (2008) 830–838.
[11] H.-J. Bandelt, P. Forster, A. Röhl, Median-joining networks for inferring intraspecific phylogenies, Mol. Biol. Evol. 16 (1999) 37–48.
[12] J.B. Kruskal, Multidimensional scaling by optimizing a goodness of fit test to a nonmetric hypothesis, Psychometrika 19 (1964) 1–27.
[13] S. Willuweit, L. Roewer, The new Y chromosome haplotype reference database, Forensic Sci. Int. Genet. 15 (2015) 43–48.
[14] L. Excoffier, 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. 10 (2010) 564–567.

[15] P. Hallast, C. Batini, D. Zadik, P. Maisano Delser, J.H. Wetton, E. Arroyo-Pardo, G.L. Cavalleri, P. de Knijff, G. Destro Bisol, B.M. Dupuy, H.A. Eriksen, L.B. Jorde, T.E. King, M.H. Larmuseau, A. Lopez de Munain, A.M. Lopez-Parra, A. Loutradis, J. Milasin, A. Novelletto, H. Pamjav, A. Sajantila, W. Schempp, M. Sears, A. Tolun, C. Tyler-Smith, A. Van Geystelen, S. Watkins, B. Winney, M.A. Jobling, The Y-chromosome tree bursts into leaf: 13,000 high-confidence SNPs covering the majority of known clades, Mol. Biol. Evol. 32 (2015) 661–673.
[16] D.B. Goldstein, A.R. Linares, L.L. Cavalli-Sforza, M.W. Feldman, An evaluation of genetic distances for use with microsatellite loci, Genetics 139 (1995) 463–471.
[17] D.B. Goldstein, A.R. Linares, L.L. Cavalli-Sforza, M.W. Feldman, Genetic absolute dating based on microsatellites and the origin of modern humans, Proc. Natl. Acad. Sci. U. S. A. 92 (1995) 6723–6727.
[18] G. Ferri, C. Robino, M. Alu, D. Luiselli, S. Tofanelli, L. Caciagli, C. Onofri, S. Pelotti, C. Di Gaetano, F. Crobu, G. Beduschi, C. Capelli, Molecular characterisation and population genetics of the DYS458 .2 allelic variant, Forensic Sci. Int. Genet. Suppl. Ser. 1 (2008) 203–205.
[19] G. Iacovacci, E. D'Atanasio, O. Marini, A. Coppa, D. Sellitto, B. Trombetta, A. Berti, F. Cruciani, Forensic data and microvariant sequence characterization of 27 Y-STR loci analyzed in four Eastern African countries, Forensic Sci. Int. Genet. 27 (2017) 123–131.
[20] P. Balaresque, G.R. Bowden, E.J. Parkin, G.A. Omran, E. Heyer, L. Quintana-Murci, L. Roewer, M. Stoneking, I. Nasidze, D.R. Carvalho-Silva, C. Tyler-Smith, P. de Knijff, M.A. Jobling, Dynamic nature of the proximal AZFc region of the human Y chromosome: multiple independent deletion and duplication events revealed by microsatellite analysis, Hum. Mutat. 29 (2008) 1171–1180.
[21] Applied Biosystems, YfilerTM Plus PCR Amplication Kit User Guide, Life Technologies Corporation, Carlsbad, CA, 2016.
[22] F. Manni, P. Leonardi, A. Barakat, H. Rouba, E. Heyer, M. Klintschar, K. McElreavey, L. Quintana-Murci, Y-chromosome analysis in Egypt suggests a genetic regional continuity in Northeastern Africa, Hum. Biol. 74 (2002) 645–658.
[23] I. Nasidze, H. Schadlich, M. Stoneking, Haplotypes from the Caucasus, Turkey and Iran for nine Y-STR loci, Forensic Sci. Int. 137 (2003) 85–93.
[24] L. Roewer, S. Willuweit, M. Stoneking, I. Nasidze, A Y-STR database of Iranian and Azerbaijanian minority populations, Forensic Sci. Int. Genet. 4 (2009) e53–5.
[25] C. Hallenberg, B. Simonsen, J. Sanchez, N. Morling, Y-chromosome STR haplotypes in Somalis, Forensic Sci. Int. 151 (2005) 317–321.
[26] J. Piatek, A. Ossowski, M. Parafiniuk, A. Pudlo, K. Jasionowicz, K. Jalowinska, A. Niemcunowicz-Janica, M. Konarzewska, W. Pepinski, Y-chromosomal haplotypes for the AmpFlSTR Yfiler PCR amplification kit in a population sample of Bedouins residing in the area of the Fourth Nile Cataract, Forensic Sci. Int. Genet. 6 (2012) e176–e177.
[27] L.A. Zhivotovsky, P.A. Underhill, C. Cinnioglu, M. Kayser, B. Morar, T. Kivisild, R. Scozzari, F. Cruciani, G. Destro-Bisol, G. Spedini, G.K. Chambers, R.J. Herrera, K.K. Yong, D. Gresham, I. Tournev, M.W. Feldman, L. Kalaydjieva, The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time, Am. J. Hum. Genet. 74 (2004) 50–61.
[28] K.N. Ballantyne, V. Keerl, A. Wollstein, Y. Choi, S.B. Zuniga, A. Ralf, M. Vermeulen, P. de Knijff, 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. 6 (2012) 208–218.
[29] J. Schlecht, M.E. Kaplan, K. Barnard, T. Karafet, M.F. Hammer, N.C. Merchant, Machine-learning approaches for classifying haplogroup from Y chromosome STR data, PLoS Comput. Biol. 4 (2008) e1000093.
[30] A. Seman, Z.A. Bakar, M.N. Isa, An efficient clustering algorithm for partitioning Y-short tandem repeats data, BMC Res. Notes 5 (2012) 557.
[31] T.W. Athey, Haplogroup prediction from Y-STR values using an allele frequency approach, J. Genet. Geneal. 1 (2005) 1–7.
[32] M. Muzzio, V. Ramallo, J.M. Motti, M.R. Santos, J.S. Lopez Camelo, G. Bailliet, Software for Y-haplogroup predictions: a word of caution, Int. J. Legal Med. 125 (2011) 143–147.
[33] W. Athey, Comments on the article, Software for Y haplogroup predictions, a word of caution, Int. J. Legal Med. 125 (2011) 901–903 author reply 905-6.
[34] L. Gusmao, J.M. Butler, A. Linacre, W. Parson, L. Roewer, P.M. Schneider, A. Carracedo, Revised guidelines for the publication of genetic population data, Forensic Sci. Int. Genet. 30 (2017) 160–163, http://dx.doi.org/10.1016/j.fsigen.2017.06.007.
[35] C. Batini, M.A. Jobling, Detecting past male-mediated expansions using the Y chromosome, Hum. Genet. 136 (2017) 547–557.
[36] T. Mohammad, Y. Xue, M. Evison, C. Tyler-Smith, Genetic structure of nomadic Bedouin from Kuwait, Heredity (Edinb) 103 (2009) 425–433.
[37] J. Chiaroni, R.J. King, N.M. Myres, B.M. Henn, A. Ducourneau, M.J. Mitchell, G. Boetsch, I. Sheikha, A.A. Lin, M. Nik-Ahd, J. Ahmad, F. Lattanzi, R.J. Herrera, M.E. Ibrahim, A. Brody, O. Semino, T. Kivisild, P.A. Underhill, The emergence of Y-chromosome haplogroup J1e among Arabic-speaking populations, Eur. J. Hum. Genet. 18 (2010) 348–353.
[38] A. Nebel, E. Landau-Tasseron, D. Filon, A. Oppenheim, M. Faerman, Genetic evidence for the expansion of Arabian tribes into the Southern Levant and North Africa, Am. J. Hum. Genet. 70 (2002) 1594–1596.
[39] P.A. Zalloua, Y. Xue, J. Khalife, N. Makhoul, L. Debiane, D.E. Platt, A.K. Royyuru, R.J. Herrera, D.F. Hernanz, J. Blue-Smith, R.S. Wells, D. Comas, J. Bertranpetit, C. Tyler-Smith, Y-chromosomal diversity in Lebanon is structured by recent historical events, Am. J. Hum. Genet. 82 (2008) 873–882.
[40] D.E. Platt, M. Haber, M.B. Dagher-Kharrat, B. Douaihy, G. Khazen, M. Ashrafian Bonab, A. Salloum, F. Mouzaya, D. Luiselli, C. Tyler-Smith, C. Renfrew, E. Matisoo-Smith, P.A. Zalloua, Mapping post-glacial expansions: the peopling of Southwest Asia, Sci. Rep. 7 (2017) 40338.
[41] W. Wei, Q. Ayub, Y. Xue, C. Tyler-Smith, A comparison of Y-chromosomal lineage dating using either resequencing or Y-SNP plus Y-STR genotyping, Forensic Sci. Int. Genet. 7 (2013) 568–572.
[42] H.S. Groucutt, M.D. Petraglia, The prehistory of the Arabian peninsula: deserts, dispersals, and demography, Evol. Anthropol. 21 (2012) 113–125.
[43] M. El-Sibai, D.E. Platt, M. Haber, Y. Xue, S.C. Youhanna, R.S. Wells, H. Izaabel, M.F. Sanyoura, H. Harmanani, M.A. Bonab, J. Behbehani, F. Hashwa, C. Tyler-Smith, P.A. Zalloua, Geographical structure of the Y-chromosomal genetic landscape of the Levant: a coastal-inland contrast, Ann. Hum. Genet. 73 (2009) 568–581.
[44] Z. Taqi, M. Alenizi, H. Alenizi, S. Ismael, A.A. Dukhyil, M. Nazir, S. Sanqoor, E. Al Harbi, J. Al-Jaber, J. Theyab, R.S. Moura-Neto, B. Budowle, Population genetics of 23 Y-STR markers in Kuwaiti population, Forensic Sci. Int. Genet. 16 (2015) 203–204.
[45] S. Triki-Fendri, S. Alfadhli, I. Ayadi, N. Kharrat, H. Ayadi, A. Rebai, Genetic structure of Kuwaiti population revealed by Y-STR diversity, Ann. Hum. Biol. 37 (2010) 827–835.
[46] A.M. Cadenas, L.A. Zhivotovsky, L.L. Cavalli-Sforza, P.A. Underhill, R.J. Herrera, Y-chromosome diversity characterizes the Gulf of Oman, Eur. J. Hum. Genet. 16 (2008) 374–386.
[47] M. Nazir, H. Alhaddad, M. Alenizi, H. Alenizi, Z. Taqi, S. Sanqoor, A. Alrazouqi, A. Hassan, R. Alfalasi, S. Gaur, J. Al Jaber, J. Ziab, E. Al-Harbi, R.S. Moura-Neto, B. Budowle, A genetic overview of 23Y-STR markers in UAE population, Forensic Sci. Int. Genet. 23 (2016) 150–152.

Forensic Science International: Genetics 39 (2019) 97–102

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsigen

Research paper

Analysis of 21 autosomal STRs in Saudi Arabia reveals population structure and the influence of consanguinity

Yahya M. Khubrani[a,b], Jon H. Wetton[a,*], Mark A. Jobling[a,*]

[a] Department of Genetics & Genome Biology, University of Leicester, University Road, Leicester, United Kingdom
[b] Forensic Genetics Laboratory, General Administration of Criminal Evidence, Public Security, Ministry of Interior, Saudi Arabia

ARTICLE INFO

Keywords:
Saudi Arabia
Autosomal STRs
Population structure
Heterozygosity
Consanguinity
GlobalFiler

ABSTRACT

Variation in the 21 autosomal STRs detected by the GlobalFiler multiplex was investigated in a sample of 523 indigenous male Arabs from five geographic regions of Saudi Arabia. Although allele frequencies for the entire dataset were found to be broadly similar to those determined in previous studies of Saudi citizens, significant differences were found among regions. Heterozygote deficiency was observed at nearly all loci in all regions, probably as a consequence of high levels of consanguineous marriage; in the case of D2S1338, which showed the largest deviation from Hardy-Weinberg equilibrium, the presence of a null allele also played a part. Genetic distances were greatest between the Northern and Southern regions, whilst the West, Central and East appeared most similar to each other, and to previously published surveys. This contrasts with previously described variation among paternal lineages in the same sample-set: Y-chromosome variation was limited within the North/Central/South core compared with the more diverse East and West. Differences between autosomal and Y-chromosomal patterns may reflect genetic drift on the Y chromosome, exacerbated by prevalent patrilineal descent groups in different regions.

1. Introduction

The Kingdom of Saudi Arabia (KSA) is located in the southwest corner of Asia at the junction between the three old-world continents of Asia, Africa and Europe. It constitutes the majority of the Arabian Peninsula along with Kuwait, Qatar and the United Arab Emirates to the east, Yemen and Oman to the south, and both Jordan and Iraq to the north. It also faces Egypt, Sudan and Eritrea over the Red Sea to the west, and Iran over the Arabian Gulf to the east. The country is populated by approximately 32.5 million people of whom 20.4 million are Saudi nationals according to Saudi Arabian General Authority for Statistics (last accessed 01/10/2018). Administratively, the country is divided into 13 regions across five geographical areas namely: Central (Riyadh, Al-Qassim), Northern (Northern borders region, Tabuk, Al-Jawf and Hail), Southern (Asir, Jazan, Al-Bahah and Najran), Eastern (Eastern province) and Western (Makkah and Al-Madinah) (Fig. 1a).

Several previous studies have investigated autosomal allele frequencies in Saudi Arabia, initially covering just eight STRs among 207 individuals [1], then 190 individuals from the Riyadh region analysed with the 15-STR Identifiler multiplex [2], and 500 individuals from six cities spread across the five regions with the 21 STRs of the GlobalFiler system [3]. Further studies examined variation at 13 (Profiler Plus) and 15 STRs (Identifiler) in Saudis resident in the bordering countries of Dubai [4] and Kuwait [5] respectively. Whilst these studies determined autosomal STR allele frequencies in Saudi nationals, they did not explore population structure within KSA. Analysis of the sample-set described here, which includes approximately equal representation of the five regions of the Kingdom, has previously revealed striking population structure in Y-chromosome variation, with greater diversity in the East and West and relative homogeneity within the North to South central axis of the country [6]. Here we ask whether population structure can also be detected with autosomal markers, and if so whether it presents a similar geographical pattern to that of the patrilineal variation.

2. Materials and methods

2.1. DNA sampling

Samples were collected from indigenous Saudi Arabian males [6] with continuous paternal ancestry back to their great-grandfather within each of the five geographical subdivisions of the country (Fig. 1a; Central N = 115, Northern N = 104, Southern N = 103, Eastern N = 88, and Western N = 113) ensuring that all donors were

* Corresponding authors at: Department of Genetics & Genome Biology, University of Leicester, University Road, Leicester, LE1 7RH, United Kingdom.
E-mail addresses: jw418@le.ac.uk (J.H. Wetton), maj4@le.ac.uk (M.A. Jobling).

https://doi.org/10.1016/j.fsigen.2018.12.006
Received 7 October 2018; Received in revised form 12 December 2018; Accepted 18 December 2018
Available online 21 December 2018
1872-4973/

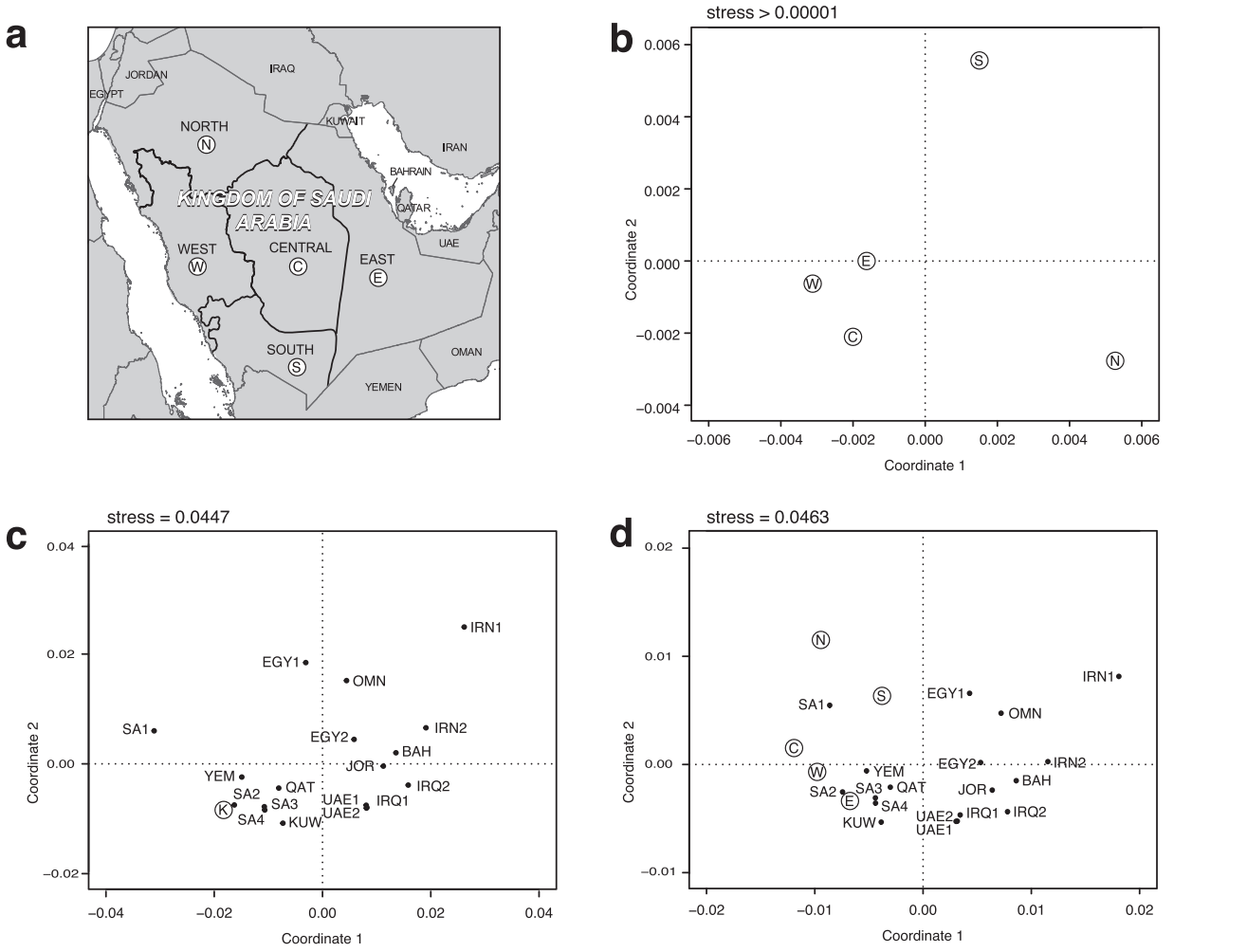


Fig. 1. Map of sample locations, and multidimensional scaling (MDS) plots based on pairwise F_{ST} values derived from autosomal STR data.
a) Map of Saudi Arabia, showing location of the five geographical sub-regions; b) MDS plot comparing the five KSA sub-regions; c) MDS comparison of combined KSA dataset (encircled K) compared to previously published Saudi Arabian datasets and neighbouring countries; d) Five KSA sub-regions (encircled N - northern; S - southern; E - eastern; W - western; C- central) compared to previously published Saudi Arabian datasets and neighbouring countries. Comparative data sources and population abbreviations are as follows: BAH - Bahrain [20], EGY1, EGY2 - Egypt, respectively [5,21]; IRN1, IRN2 - Iran, respectively [5,22]; IRQ1, IRQ2 - Iraq, respectively [5,23]; JOR - Jordan [24]; KUW - Kuwait [5]; OMN - Oman [4]; QAT - Qatar [25]; SA1, SA2, SA3, SA4 - Saudi Arabia, respectively [2–5]; UAE1, UAE2 - United Arab Emirates, respectively [26,27]; YEM - Yemen [4].

separated by at least three generations. Ethical review for recruitment and analysis was provided by the Saudi General Administration for Forensic Evidence and the University of Leicester Research Ethics Committee. Informed consent was provided by all participants.

2.2. DNA extraction and quantification

DNA from FTA blood-spot samples (Whatman, UK) was extracted, purified and quantified as previously described [6].

2.3. DNA amplification and fragment detection

The GlobalFiler® PCR amplification kit was used to generate profiles based on 21 autosomal STRs: D3S1358, vWA, D16S539, CSF1PO, TPOX, D8S1179, D21S11, D18S51, D2S441, D19S433, TH01, FGA, D22S1045, D5S818, D13S317, D7S820, SE33, D10S1248, D1S1656, D12S391 and D2S1338, along with three additional markers used in sex determination (DYS391, Y indel and Amelogenin). Amplification was performed on a Veriti PCR machine (Thermo Fisher Scientific) and fragment detection on an ABI3500 Genetic Analyzer (Thermo Fisher Scientific) in accordance with the manufacturer's recommended protocols. GeneMapper IDX software V1.4 was used for allele calling and interpretation.

This work followed the guidelines of *FSI:Genetics* for publication of population genetic data [7–9] and for allele nomenclature [10]. The dataset has been QC checked via STRidER [11], with the dataset reference STR000119.

2.4. Forensic and statistical analysis

PowerStats v1.2 software (Promega Corporation, Madison, WI, USA) [12] was used to calculate allele frequencies, Random Match Probability (PM), Power of Discrimination (PD), Power of Exclusion (PE), Typical Paternity Index (TPI), observed homozygosity and observed heterozygosity. Arlequin v 3.5 [13] was used to test Hardy-Weinberg equilibrium, calculate expected heterozygosity, perform AMOVA to investigate genetic diversity within and between the five geographical regions, and undertake population differentiation tests for KSA and neighbouring countries. F_{IS} was calculated using FSTAT [14]. Average pairwise F_{ST} values for 13 loci (CSF1PO, D13S317, D16S539, D18S51, D21S11, D3S1358, D5S818, D7S820, D8S1179, FGA, TH01, TPOX, vWA) shared with other studies of Saudi Arabians and neighbouring populations were used to generate multidimensional scaling (MDS) plots using the (MASS) package [15] in the R library.

Table 1
KSA allele frequencies and forensic statistics.

Allele	D3S1358	vWA	D16S539	CSF1PO	TPOX	D8S1179	D21S11	D18S51	D2S441	D19S433	TH01	FGA	D22S1045	D5S818	D13S317	D7S820	SE33	D10S1248	D1S1656	D12S391	D2S1338
4											0.001										
6					0.002						0.339										
6.3																	0.005				
7				0.001	0.001						0.162				0.001	0.012					
7.3																	0.005				
8			0.031	0.005	0.552	0.002					0.110			0.011	0.105	0.165			0.001		
9			0.113	0.011	0.163	0.007			0.011		0.282			0.076	0.033	0.111	0.004	0.018			
9.3											0.096										
10			0.128	0.304	0.113	0.038		0.001	0.124	0.001	0.010		0.003	0.123	0.062	0.361	0.001		0.002		
11			0.359	0.294	0.157	0.137		0.021	0.358	0.011	0.001		0.107	0.262	0.255	0.183	0.001	0.009	0.042		
11.2																	0.002				
11.3									0.077												
12			0.212	0.345	0.012	0.188		0.144	0.080	0.095			0.011	0.346	0.341	0.146	0.006	0.037	0.131		
12.2										0.009											
13	0.007	0.002	0.137	0.033		0.185		0.210	0.010	0.168				0.171	0.147	0.020	0.015	0.167	0.099		
13.2										0.046							0.001				
13.3			0.001						0.001								0.001				
14	0.060	0.029	0.017	0.006		0.192		0.122	0.300	0.209			0.058	0.011	0.054	0.001	0.052	0.362	0.125	0.002	0.004
14.2										0.048											
15	0.239	0.157	0.002			0.201		0.119	0.033	0.141			0.508	0.001	0.001		0.030	0.257	0.136	0.025	
15.2	0.001							0.003		0.126											
15.3																			0.033		
16	0.314	0.282				0.043		0.118	0.006	0.074			0.258				0.058	0.105	0.251	0.011	0.077
16.2	0.001							0.006		0.053											
16.3																			0.054		
17	0.267	0.272				0.008		0.094		0.007			0.054				0.075	0.043	0.074	0.110	0.225
17.2								0.002		0.012											
17.3																			0.022		
18	0.105	0.179						0.073				0.007					0.107	0.001	0.003	0.149	0.098
18.3																			0.020	0.004	
19	0.007	0.066						0.041				0.064					0.064		0.001	0.133	0.127
19.3																			0.006	0.005	
20		0.012						0.026				0.070					0.044			0.100	0.239
20.2																	0.007				
21		0.001						0.008				0.139					0.018			0.089	0.047
21.2												0.001					0.076				
22								0.009				0.122					0.002			0.123	0.014
22.2												0.002					0.018				
22.3												0.002									
23								0.004				0.185								0.140	0.068
23.2												0.002					0.020				
24								0.001				0.226					0.001			0.057	0.064
24.2																	0.028				
25												0.112								0.041	0.030
25.2																	0.027				
26												0.052								0.007	0.006
26.2																	0.050				
27							0.011					0.008								0.004	0.001
27.2																	0.050				
28							0.131					0.006									
28.2																	0.042				
29							0.270					0.003									
29.2																	0.036				
30							0.256					0.001									

(*continued on next page*)

Table 1 (*continued*)

Allele	D3S1358	vWA	D16S539	CSF1PO	TPOX	D8S1179	D21S11	D18S51	D2S441	D19S433	TH01	FGA	D22S1045	D5S818	D13S317	D7S820	SE33	D10S1248	D1S1656	D12S391	D2S1338
30.2							0.011										0.043				
31							0.046										0.001				
31.2							0.103					0.001					0.049				
32							0.007														
32.2							0.112										0.038				
33																	0.005				
33.2							0.039										0.007				
34																	0.005				
34.2							0.007										0.005				
35							0.005										0.002				
35.2							0.001										0.001				
36							0.001														
36.2																	0.001				
37							0.001										0.001				
	D3S1358	vWA	D16S539	CSF1PO	TPOX	D8S1179	D21S11	D18S51	D2S441	D19S433	TH01	FGA	D22S1045	D5S818	D13S317	D7S820	SE33	D10S1248	D1S1656	D12S391	D2S1338
MP	0.097	0.078	0.081	0.147	0.180	0.052	0.056	0.029	0.100	0.030	0.097	0.038	0.163	0.092	0.081	0.081	0.007	0.096	0.032	0.024	0.040
PE	0.474	0.471	0.490	0.385	0.287	0.644	0.556	0.685	0.443	0.656	0.486	0.659	0.291	0.461	0.452	0.490	0.832	0.566	0.634	0.659	0.570
O-Het	0.729	0.727	0.738	0.671	0.597	0.824	0.778	0.845	0.709	0.830	0.736	0.832	0.600	0.721	0.715	0.738	0.918	0.782	0.818	0.832	0.784
E-Het	0.759	0.785	0.777	0.701	0.632	0.832	0.818	0.876	0.754	0.871	0.759	0.857	0.658	0.762	0.779	0.775	0.948	0.761	0.864	0.890	0.850
P-value	0.264	< **0.0001**	0.019	0.141	0.444	0.944	0.637	0.226	0.004	0.874	0.092	0.130	0.029	0.158	0.020	0.374	0.053	0.364	0.559	0.068	< **0.0001**
Alleles	9	9	9	8	7	10	15	18	10	14	8	18	7	8	9	8	42	9	16	16	14
F_{IS}	0.037	0.072	0.050	0.043	0.058	0.010	0.047	0.032	0.064	0.045	0.027	0.027	0.086	0.049	0.078	0.049	0.032	−0.036	0.055	0.065	0.076
P-value	0.072	**0.001**	0.014	0.070	0.020	0.321	0.011	0.029	**0.005**	**0.005**	0.135	0.067	**0.001**	0.022	**0.001**	0.020	**0.002**	0.952	**0.001**	< **0.0001**	< **0.0001**

MP: Random Match Probability, PE: Power of Exclusion, O-Het: observed heterozygosity, E-Het: expected heterozygosity, HWE p-value: probability of deviation from Hardy Weinberg equilibrium, F_{IS} p-value: significance of F_{IS}.

3. Results

3.1. Data descriptions and forensic statistics

The 21 autosomal STRs targeted by the GlobalFiler kit were amplified from 523 Saudi Arabian males. Table 1 presents allele frequency data and forensic statistics for the whole KSA dataset; these measures are also provided for each of the five geographical subdivisions (Fig. 1a) in Table S1.

The least variable loci are TPOX and D22S1045, each with seven allelic variants in the KSA dataset, and the most variable locus was SE33 with 42 alleles. These respective loci had the lowest and highest Probabilities of Discrimination (0.820, 0.837 and 0.993 respectively), and conversely the highest and lowest major allele frequencies: TPOX allele 8 had a frequency of 0.552 in the total KSA dataset, being most frequent in the North (0.596) and rarest in the Central region (0.517), whilst SE33 allele 18, the most common variant at that locus, was detected at a frequency of just 0.107 in the KSA dataset overall. The combined power of discrimination (PD) and the combined power of exclusion (PE) for all loci in the KSA data are 0.99999999999999999999999999705 ($PI = 2.95 \times 10^{-26}$) and 0.9999999563, respectively.

3.2. Rare variants, off-ladder and null alleles

Thirty-five alleles were each observed only once in the entire dataset, and have been designated “rare” among Saudis. Of these, 13 were also globally uncommon, and were among the 26 off-ladder alleles recorded at these loci: D3S1358 (15.2 [$N_{obs} = 1$] and 16.2 [1]), D16S539 (13.3 [1]), D18S51 (15.2 [3], 16.2 [6], 17.2 [2]), D2S441 (13.3 [1]), FGA (21.2 [1], 22.2 [2], 22.3 [2], 23.2 [2]), SE33 (7.3 [5], 10 [1], 11.2 [2], 13.2 [1], 13.3 [1], 22 [2], 24 [1], 31 [1], 33 [5], 34 [5], 36.2 [1]), D1S1656 (8 [1],18 [3],19 [1]) and D12S391 (18.3 [4]). All 26 off-ladder alleles have been described previously in STRBase (http://strbase.nist.gov/index.htm) [16]. No peak was detected at D2S1338 in one individual, despite good signal strength at all other loci, and despite retyping the sample twice. We assume this individual is a null homozygote at this locus and have excluded him from further analyses, as recommended [11].

3.3. Genetic structure

All but one locus showed a deficiency of heterozygotes against expectation in the whole KSA dataset, and this was also apparent within all five regions for between 16 and 20 loci (Table 2 for KSA, and Table S2 for each region). The deviation from Hardy-Weinberg equilibrium was significant following Bonferroni correction ($P \leq 0.00001$) for D2S1338 and vWA. Heterozygote deficiency was also evident from AMOVA analysis, with an inbreeding coefficient (F_{IS}) of 0.0476 representing 4.71% of variation for the KSA dataset, whilst F_{ST} was 0.0021. F_{ST} values between the five regions (Table 2) show that the greatest differentiation is between the North and South, with West, Central and East being less differentiated, and the East region being most similar to the other regions, as reflected in an MDS plot (Fig. 1b).

Whilst the entire KSA dataset broadly clustered with the previously published Saudi datasets in MDS analysis (Fig. 1c), division according to sub-regional origin showed that the North and South sub-regions were also the most differentiated from most other nearby populations (Fig. 1d). The results of per-locus population differentiation tests between the KSA and regional datasets, previously published Saudi datasets and neighbouring countries are presented in Table S3.

Table 2
Pairwise F_{ST} between regional sub-populations.

Region	C	E	N	S	W
C	*	0.0322	**0.0020**	< **0.0001**	0.5664
E	0.0017	*	0.1006	0.1084	0.8291
N	0.0028	0.0013	*	< **0.0001**	**0.0010**
S	0.0041	0.0012	0.0053	*	0.0166
W	0.0001	0	0.0040	0.0021	*

Figures upper-right are p-values, with significant values in bold.

4. Discussion

In our analysis of diversity at the 21 autosomal STR loci of the GlobalFiler multiplex we found no previously unreported alleles among 523 indigenous Saudi Arabian individuals and no evidence of genetic differentiation between the combined KSA dataset and previously published Saudi autosomal datasets, through exact tests of allele frequency. In common with previous studies [1–3,5], a tendency towards heterozygote deficiency was observed affecting almost all loci, although only two retained significance following Bonferroni correction. This observation, and the elevated F_{IS} values, are likely to reflect historical marriage practices in KSA in which partners are usually from within the same tribal group, and rates of first-cousin marriages are high, at around 30% [17]. One of the two loci that showed significant deviation from Hardy-Weinberg equilibrium due to heterozygote deficiency (D2S1338) also showed evidence of the presence of a null allele, identified through a null homozygote. We assume that the combined effect of consanguinity [17] and presence of null alleles in the heterozygous state produced the apparent excess of CE length “homozygotes” ($P = 0.00001$). It is unclear whether both factors, or only inbreeding, contributed to the significant but slightly weaker distortion observed at vWA ($P = 0.00006$).

Interestingly, a null has previously been reported at D2S1338 with the Identifiler multiplex, which could be weakly detected as a visible allele with the NGM kit [18]. Identifiler, NGM and GlobalFiler are all Thermo Fisher multiplexes and share exactly the same primers at this locus (Matt Phipps, Thermo Fisher, personal communication). The same SNP (rs567937457) which lies 174 bp downstream of the repeat array was also identified by NIST as causing discordance between Identifiler and Promega PP18 kits (http://strbase.nist.gov/pub_pres/NIST-Update-EDNAP-Apr2011.pdf). It has been suggested that the discrepancy between Identifiler and NGM is related to the longer annealing/extension time (3 min vs. 1 min) in the latter, which may permit amplification from the poorly matched primer [18]; however, in our homozygous null individual, extending the annealing/extension time to 3 min did not yield a detectable peak. Subsequent amplification with the MiniFiler kit which has a 183 bp shorter amplicon also yielded no result. The positions of the MiniFiler primers are proprietary information but can encompass at most 60 bp of flanking region which must exclude the aforementioned SNP as the cause, and would imply either a substantial deletion affecting both downstream primers or a polymorphism upstream of the repeat, where the primer is already close to the repeat array. Unfortunately we are unable to explore the nature of this polymorphism any further due to the limitations of the DNA donor consent.

This study is the first to specifically address the question of substructure within the indigenous Saudi Arabian population using autosomal STR markers. Our population sample, which has approximately equal representation of the five geographic regions of the Kingdom, has previously been shown to display striking differentiation in Y-chromosomal haplogroup (Y-SNP) and haplotype (Y-STR, Yfiler Plus) distributions [6], suggesting that the country is genetically substructured at least with respect to male-specific markers.

A different picture emerged when we explored the genetic diversity of the geographic sub-regions with GlobalFiler. Significantly different pairwise F_{ST} values were noted between the South, North and Central/West regions, with the East appearing intermediate, and not significantly differentiated from any other region (Fig. 1b). The combined KSA dataset is virtually indistinguishable from sets of Saudi donors sampled in Dubai and Kuwait [4,5] and the previous multi-city

GlobalFiler survey within Saudi Arabia [3], but somewhat different from the Identifiler dataset derived from bone marrow donors at a Riyadh hospital [2], which is an outlier (Fig. 1c). However, after division of our dataset into the five regions (Fig. 1d) we see that sample-sets from the previously published population surveys now cluster most closely with the Eastern region. Also, the North and South are the most differentiated from each other and from the other regions, with the exception of the bone marrow donor set [2], which shows similarity with the South.

Autosomal differentiation between North and South may be a consequence of historically limited migration between these regions. Migrants are now primarily attracted by the oil industry in the East, the holy cities of the West and the capital in the Central region [19]. Prior to the establishment of the Kingdom of Saudi Arabia the interior was sparsely populated by sedentary farmers or nomads who moved within their tribal areas, reflected by some of the current administrative boundaries. The rapid growth of the Saudi population, stimulated by the discovery of oil, has involved movement into the cities from the surrounding areas, but the rural populations have not generally moved between regions. Our regional patrilineal criterion for inclusion in the dataset [6] meant that the results of Y-chromosome tests reflect the historic boundaries of tribal groups, and these are strongest and most stable in the North and Central regions. By contrast, the East and West received the greatest inward migration in recent centuries, as they were outward facing and the destinations of traders and pilgrims. The autosomal results also reflect the movement of women, who are likely to move to their husband's home. We applied no restriction on the origin of the maternal line, and as a consequence a different pattern reflecting more recent migration patterns is unsurprising.

5. Conclusion

This study is in concordance with the high discrimination power of GlobalFiler in the Saudi population, which makes it suitable for the purposes of forensic DNA identification and paternity testing; the allele frequencies derived in this study can be utilised by Saudi forensic laboratories for DNA interpretation purposes. Whilst the observed allele frequency variation between regions will have limited influence on interpretation issues, the high level of consanguinity and presence of null alleles should be taken into account.

Conflicts of interest

None.

Acknowledgments

YMK was supported by the Saudi Arabian Ministry of Interior, and by a PhD studentship from the Saudi Arabian Cultural Bureau, London. We thank members of the Forensic Genetics Laboratory, General Administration of Criminal Evidence, Riyadh, for assistance, and Jon Kyte & Matt Phipps from Thermo Fisher for their support.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.fsigen.2018.12.006.

References

[1] S. Sinha, M. Amjad, C. Rogers, J.E. Hamby, U.A. Tahir, K. Balamurugan, N.A. al-Kubaidan, A.R. Choudhry, B. Budowle, M.A. Tahir, Typing of eight short tandem repeat (STR) loci in a Saudi Arabian population, Forensic Sci. Int. 104 (1999) 143–146.
[2] A.E. Osman, H. Alsafar, G.K. Tay, J.B.J.M. Theyab, M. Mubasher, N.E.-E. Sheikh, H. AlHarthi, M.H. Crawford, G. Gehad El Ghazali, Autosomal short tandem repeat (STR) variation based on 15 loci in a population from the Central Region (Riyadh Province) of Saudi Arabia, J. Forensic Res. 6 (2015) 1000267.
[3] H.M. Alsafiah, W.H. Goodwin, S. Hadi, M.A. Alshaikhi, P.P. Wepeba, Population genetic data for 21 autosomal STR loci for the Saudi Arabian population using the GlobalFiler® PCR amplification kit, Forensic Sci. Int. Genet. 31 (2017) e59–e61.
[4] F. Alshamali, A.Q. Alkhayat, B. Budowle, N.D. Watson, STR population diversity in nine ethnic populations living in Dubai, Forensic Sci. Int. 152 (2005) 267–279.
[5] M. Al-Enizi, J. Ge, S. Ismael, H. Al-Enezi, A. Al-Awadhi, W. Al-Duaij, B. Al-Saleh, Z. Ghulloom, B. Budowle, Population genetic analyses of 15 STR loci from seven forensically-relevant populations residing in the state of Kuwait, Forensic Sci. Int. Genet. 7 (2013) e106–7.
[6] Y.M. Khubrani, J.H. Wetton, M.A. Jobling, Extensive geographical and social structure in the paternal lineages of Saudi Arabia revealed by analysis of 27 Y-STRs, Forensic Sci. Int. Genet. 33 (2018) 98–105.
[7] A. Carracedo, J.M. Butler, L. Gusmao, W. Parson, L. Roewer, P.M. Schneider, Publication of population data for forensic purposes, Forensic Sci. Int. Genet. 4 (2010) 145–147.
[8] A. Carracedo, J.M. Butler, L. Gusmao, A. Linacre, W. Parson, L. Roewer, P.M. Schneider, New guidelines for the publication of genetic population data, Forensic Sci. Int. Genet. 7 (2013) 217–220.
[9] L. Gusmao, J.M. Butler, A. Linacre, W. Parson, L. Roewer, P.M. Schneider, A. Carracedo, Revised guidelines for the publication of genetic population data, Forensic Sci. Int. Genet. 30 (2017) 160–163.
[10] P.M. Schneider, Scientific standards for studies in forensic genetics, Forensic Sci. Int. 165 (2007) 238–243.
[11] M. Bodner, I. Bastisch, J.M. Butler, R. Fimmers, P. Gill, L. Gusmao, N. Morling, C. Phillips, M. Prinz, P.M. Schneider, W. Parson, 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. 24 (2016) 97–102.
[12] A. Tereba, Tools for Analysis of Population Statistics, Profiles in DNA 3, Available from (1999), pp. 14–16 http://www.promega.com/geneticidtools/powerstats.
[13] L. Excoffier, 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. 10 (2010) 564–567.
[14] J. Goudet, FSTAT (version 1.2): a computer program to calculate F-statistics, J. Hered. 86 (1995) 485–486.
[15] W.N. Venables, B.D. Ripley, Modern Applied Statistics With S, fourth edition, Springer, New York, 2002.
[16] C.M. Ruitberg, D.J. Reeder, J.M. Butler, STRBase: a short tandem repeat DNA database for the human identity testing community, Nucleic Acids Res. 29 (2001) 320–322.
[17] L. Al-Gazali, H. Hamamy, S. Al-Arrayad, Genetic disorders in the Arab world, BMJ 333 (2006) 831–834.
[18] A.A. Westen, T. Kraaijenbrink, E.A. Robles de Medina, J. Harteveld, P. Willemse, S.B. Zuniga, K.J. van der Gaag, N.E. Weiler, J. Warnaar, M. Kayser, T. Sijen, P. de Knijff, Comparing six commercial autosomal STR kits in a large Dutch population sample, Forensic Sci. Int. Genet. 10 (2014) 55–63.
[19] M. Al-Rasheed, A History of Saudi Arabia, Cambridge University Press, Cambridge, 2010.
[20] A.S. Abuidrees, M.J. Ishaq, C.-E. Pu, N.A. Alhamad, H.A. Alnafea, A.M. Almehaizea, A globally used 15 short tandem repeats (STR) loci in forensic human identification, with their allele frequencies and statistical values in the population of Bahrain, Arab Gulf J. Sci. Res. 32 (2014) 177–182.
[21] G.A. Omran, G.N. Rutty, M.A. Jobling, Genetic variation of 15 autosomal STR loci in Upper (Southern) Egyptians, Forensic Sci. Int. Genet. 3 (2009) e39–44.
[22] A. Hedjazi, A. Nikbakht, M. Hosseini, A. Hoseinzadeh, S.M. Hosseini, Allele frequencies for 15 autosomal STR loci in Fars province population, southwest of Iran, Leg. Med. Tokyo (Tokyo) 15 (2013) 226–228.
[23] M.M. Farhan, S. Hadi, A. Iyengar, W. Goodwin, Population genetic data for 20 autosomal STR loci in an Iraqi Arab population: application to the identification of human remains, Forensic Sci. Int. Genet. 25 (2016) e10–e11.
[24] L.N. Al-Eitan, R.R. Tubaishat, Evaluation of forensic genetic efficiency parameters of 22 autosomal STR markers (PowerPlex® Fusion system) in a population sample of Arab descent from Jordan, Austin J. Forensic Sci. Criminol. 50 (2016) 97–109.
[25] A.M. Perez-Miranda, M.A. Alfonso-Sanchez, J.A. Pena, R.J. Herrera, Qatari DNA variation at a crossroad of human migrations, Hum. Hered. 61 (2006) 67–79.
[26] R.J. Jones, W.A. Tayyare, G.K. Tay, H. Alsafar, W.H. Goodwin, Population data for 21 autosomal short tandem repeat markers in the Arabic population of the United Arab Emirates, Forensic Sci. Int. Genet. 28 (2017) e41–e42.
[27] O. Ali Alhmoudi, R.J. Jones, G.K. Tay, H. Alsafar, S. Hadi, Population genetics data for 21 autosomal STR loci for United Arab Emirates (UAE) population using next generation multiplex STR kit, Forensic Sci. Int. Genet. 19 (2015) 190–191.

Forensic Science International: Genetics 43 (2019) 102164

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsigen

Research paper

Massively parallel sequencing of autosomal STRs and identity-informative SNPs highlights consanguinity in Saudi Arabia

Yahya M. Khubrani[a,b], Pille Hallast[c,d], Mark A. Jobling[a,*], Jon H. Wetton[a,*]

[a] Department of Genetics & Genome Biology, University of Leicester, Leicester, UK
[b] Forensic Genetics Laboratory, General Administration of Criminal Evidence, Public Security, Ministry of Interior, Saudi Arabia
[c] Institute of Biomedicine and Translational Medicine, University of Tartu, Tartu, Estonia
[d] Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, UK

ARTICLE INFO

Keywords:
Saudi Arabia
Massively parallel sequencing (MPS)
DNA Signature Prep Kit
Autosomal STRs
Identity-informative SNPs
HGDP
Population structure
Heterozygosity
Consanguinity
Inbreeding

ABSTRACT

While many studies have been undertaken of Middle Eastern populations using autosomal STR profiling by capillary electrophoresis, little has so far been published from this region on the forensic use of massively parallel sequencing (MPS). Here, we carried out MPS of 27 autosomal STRs and 91 identity-informative SNPs (iiSNPs) with the Verogen ForenSeq™ DNA Signature Prep Kit on a representative sample of 89 Saudi Arabian males, and analysed the resulting sequence data using Verogen's ForenSeq Universal Analysis Software (UAS) v1.3 and STRait Razor v3.0. This revealed sequence variation in the composition of complex STR arrays, and SNPs in their flanking regions, which raised the number of STR alleles from 238 distinct length variants to 357 sequence sub-variants. Similarly, between one and three additional polymorphic sites were observed within the amplicons of 37 of the 91 iiSNPs, forming up to six microhaplotypes per locus. These further enhance discrimination compared to the biallelic target SNP data presented by the primary UAS interface. In total, we observed twenty-two STR alleles previously unrecognised in the STRait Razor v3.0 default allele list, along with nine SNPs flanking target iiSNPs that were not highlighted by the UAS. Sequencing reduced the STR-based random match probability (RMP) from 2.62E-30 to 3.49E-34, and analysis of the iiSNP microhaplotypes reduced RMP from 9.97E-37 to 6.83E-40. The lack of significant linkage disequilibrium between STRs and target iiSNPs allowed the two marker types to be combined using the product rule, yielding a RMP of 2.39E-73. Evidence of consanguinity was apparent from both marker types. While TPOX was the only locus displaying a significant deviation from Hardy-Weinberg equilibrium, 23 out of 27 STRs and 63 out of 91 iiSNPs showed fewer than expected heterozygotes, demonstrating an overall homozygote excess probably reflecting the high frequency of first-cousin marriages in Saudi Arabia. We placed our data in a global context by considering the same markers in the Human Genome Diversity Panel (HGDP), revealing that the Saudi sample was typical of Middle Eastern populations, with a higher level of inbreeding than is seen in most European, African and Central/South Asian populations, correlating with known patterns of endogamy. Given reduced levels of diversity within endogamous groups, the ability to combine the discrimination power of both STRs and SNPs offers significant benefits in the analysis of forensic evidence in Saudi Arabia and the Middle East region more generally.

1. Introduction

Traditionally, forensic analysis of short tandem repeat (STR) diversity has been performed via capillary electrophoresis (CE), which considers only amplicon length and overlooks potentially informative sequence variation. Additional polymorphisms could include different numbers of diverged repeat units, together with indels and SNPs in both repeat arrays themselves and flanking regions: such variation can be accessed through massively parallel sequencing (MPS). MPS approaches not only provide increased (sequence-level) resolution of individual loci, but can also simultaneously analyse many diverse loci in a single test, thus simplifying analysis of different marker types and making the best use of limited amounts of casework material [1].

The ForenSeq™ DNA Signature Prep Kit (Verogen), formally released in 2015, exemplifies the advantages of an MPS approach by allowing the simultaneous amplification of either > 150 loci including the standard autosomal and Y-STRs plus X-STRs and identity-informative SNPs (iiSNPs), or > 230 loci which also include

* Corresponding authors at: Department of Genetics & Genome Biology, University of Leicester, University Road, Leicester LE1 7RH, UK.
E-mail addresses: maj4@le.ac.uk (M.A. Jobling), jw418@le.ac.uk (J.H. Wetton).

https://doi.org/10.1016/j.fsigen.2019.102164
Received 8 July 2019; Received in revised form 19 September 2019; Accepted 20 September 2019
Available online 22 September 2019

biogeographical ancestry- and phenotypically-informative SNPs; the choice of which to use depends on local legislation, sample type and/or application considerations. The kit has been extensively validated through a wide range of performance tests, including robustness, reproducibility, concordance with CE and sensitivity of detection [2–11]. Furthermore, the DNA Signature Prep Kit has been used on challenging samples to evaluate its applicability to real forensic cases [10,12,13]: in these tests, the kit as a whole compared favourably with CE, for example on formalin-fixed paraffin-embedded tissue, as well as on bone samples dating from the 7th to 18th centuries, detecting a greater number of informative markers in seven out of ten cases [2]. While optimised mini-STR multiplexes for CE and MPS SNP typing kits may perform better in detecting each type of marker on their own, the DNA Signature Prep Kit provided greater resolution in challenging samples as the simultaneous amplification of both STR and SNP markers allows the most discriminating markers to be detected from limited samples, despite the level of DNA degradation [14].

Although the primary interface with the manufacturer's ForenSeq™ Universal Analysis Software (UAS) currently focuses on only a single iiSNP within each amplicon, additional variation exists in the flanking regions of many target SNPs [15], which could be useful in further increasing discriminating power and interpreting mixed stains. For several of the targeted SNPs the additional variant sites result in multiple alleles that will tend to be co-inherited as a microhaplotype, giving some iiSNP amplicons a resolving power approaching those of traditional simple-sequence STR loci. The DNA Signature Prep Kit has also improved resolution of both male/male and male/female mixtures involving minor contributors as low as 1:20 [2,6,8,10] and this capability has proved useful in the first sexual assault court case using MPS in the Netherlands. The kit has also been implemented in casework by the INPS (Institut National de Police Scientifique) laboratory in Lyon, with MPS profiles uploaded to the French national DNA database, and the FBI has approved the kit itself, the MiSeq FGx System and the UAS for the US National DNA Index System within the terms of newly published SWGDAM guidelines for MPS [16].

The advantages conferred by MPS approaches could be particularly relevant to forensic science in the Middle East, where high average annual temperatures favour the use of forensic markers such as short-amplicon STRs and iiSNPs that provide robust discrimination from degraded DNA. In addition, populations in the region tend to exhibit endogamy and population structure [17], leading to clusters of individuals that share a common heritage and reducing the discrimination of forensic multiplexes relative to more diverse and exogamous societies. We have previously shown by conventional CE typing of both Y- [18] and autosomal STRs [19] that the indigenous population of Saudi Arabia is highly structured. This is apparent between the five regions of the country (North, South, East, West and Central), which have different tribal compositions and historical exposures to immigration.

Here we apply the DNA Signature Prep Kit typing to a sample of Saudi males currently residing in the UK, to determine whether they reflect both the genetic composition of their home nation and the expected high rates of consanguinity, which could impact on the observed homozygosity of the autosomal markers. We ask whether analysing sequence variation of autosomal markers (aSTRs and iiSNPs) within the DNA Signature Prep Kit significantly enhances resolution, and report novel variants within this population of the Middle East, a region from which little MPS data has so far been published [20].

2. Materials and methods

2.1. DNA sampling, extraction and quantification

Samples were collected from 89 indigenous Arab males residing in the United Kingdom whose continuous paternal line ancestry can be traced back to a great-grandfather within one of the five geographical subdivisions of Saudi Arabia [18] (Central N = 19, Northern N = 9, Southern N = 26, Eastern N = 16, and Western N = 19). Ethical review for recruitment, sampling and analysis was provided by the University of Leicester Research Ethics Committee. Informed consent was provided by all participants, and using genealogical data obtained from the donors we confirmed that all paternal lineages are unconnected within the last three generations.

DNA was extracted either from buccal swabs [21], or from saliva samples using the Oragene DNA (OG-500) kit (DNA Genotek), and quantified as previously described [18].

2.2. Library preparation and sequencing

Sequencing libraries were prepared using the ForenSeq™ DNA Signature Prep Kit according to the manufacturer's recommendations (Verogen Inc., San Diego, CA, USA). Primer Mix A was used to amplify 58 STRs (27 autosomal STRs discussed in this paper, as well as 7 X-STRs and 24 Y-STRs which will be reported separately) and 94 identity-informative SNPs (iiSNPs) from 1 ng of template DNA. Steps for library preparation include target-specific amplification, target enrichment including incorporation of indexed adapters, purification, bead-normalisation and pooling, prior to sequencing on a Verogen MiSeq FGx device, all of which were performed in accordance with the manufacturer's recommended protocols. Either 96 or 32 libraries were analysed in each sequencing run, which included one positive and one negative control.

2.3. Calling of iiSNPs from HGDP sequence data

The iiSNPs were jointly called from the cram files of whole-genome sequenced (mean coverage 35 ×) Human Genome Diversity Project (HGDP) – CEPH (Centre d'Etude du Polymorphisme Humain) samples [22], mapped to GRCh38 using BCFtools (v1.8) [23] with minimum base quality 20 and mapping quality 20.

2.4. Data analysis

Sequence data were analysed using Verogen's default settings for the Analytical Threshold (AT), Interpretation Threshold (IT), Stutter Filter and Intra-Locus Balance in the ForenSeq™ Universal Analysis Software (UAS). Any sequence detected above the analytical threshold of 10 reads is reported to the user for their consideration, while above the 30-read interpretation threshold the UAS automatically reports the presence of an allele if the overall read depth for the locus is < 650. When ≥650 reads are collected for a locus, the AT and IT defaults are set at 1.5% and 4.5% of reads respectively. Between the AT and IT, the result is flagged by the UAS and the user can determine whether the sequence is a true variant. User interpretation is also required when the Intra-locus Balance (equivalent to heterozygote balance) falls below 60%, or the level of stutter exceeds the default Stutter Filter value, which varies between STR loci. As all of our samples were good quality single-source reference DNAs, as demonstrated by earlier Yfiler Plus profiling [18], the interpretation was relatively straightforward.

The UAS provides a visual interface which displays each STR locus for an individual as a histogram arranged according to conventional CE allele length, with isometric heterozygotes (alleles of the same length but different sequence) shown as stacked bars. For STRs, the visual UAS interface displays only repeat region sequence variants, and for each of the iiSNP amplicons, only the target SNP. However, it is also possible to view "Flanking Region Reports" that show the flanking regions of both aSTR and iiSNP amplicons (which for Penta E is limited to a maximum of 197 bp to ensure sequencing data integrity). These files highlight variation at some, but not all, additional polymorphic sites within the amplicons. We used STRait Razor v3.0 [24] to check bioinformatic concordance of allele calls and to clarify appropriate nomenclature in line with ISFG considerations [25]. In the following sections we

describe STR sequence variation at three levels: length - a measure solely of allele length for compatibility with conventional CE methods; repeat region sequence - sequence variation within the repeat array as reported by UAS; and repeat plus flanking sequence, including all polymorphisms within the reported region of the amplicon, as obtained from the additional Flanking Region Reports.

2.5. Population, forensic and statistical analysis

Arlequin v3.5 [26] was used to test Hardy-Weinberg equilibrium, and to calculate expected heterozygosity and pairwise linkage disequilibrium (LD). It was also used for performing AMOVA to investigate genetic diversity, to calculate fixation indices (F_{IS}) and to undertake population differentiation tests. STRAF [27] was used to calculate forensic statistics including: genotype count (N), allele count based on sequence (N_{all}), genetic diversity (GD), polymorphism information content (PIC), random match probability (PM), power of discrimination (PD), observed and expected heterozygosity (H_{obs} and H_{exp}), power of exclusion (PE), and typical paternity index (TPI). Allele frequencies were calculated in Excel (allele count/total).

3. Results

DNA profiles of 89 Saudi males generated with the ForenSeq™ DNA Signature Prep Kit Primer Mix A gave a mean total read count per sample of 34,821 reads for the complete complement of autosomal STR alleles, and 41,650 for the set of iiSNPs. Following visual checks of individual loci flagged by the default settings of the ForenSeq™ UAS, we called genotypes for all 27 autosomal STRs and 91 of the 94 iiSNPs. The three lowest-performing iiSNPs (rs1736442, rs2920816 and rs719366) were dropped from the analysis due to increased observations of sub-threshold calls (15, 9 and 5 respectively). The SNP rs1031825 was retained in the analysis but was below threshold in three individuals. In addition, two individuals had sub-threshold calls at the STR locus Penta E. Average read depths for the analysed autosomal STRs ranged from 3979 at TH01 to 285 at vWA, and for the iiSNPs from 1710 at rs1109037 to 64 at rs1031825.

3.1. Autosomal STR sequence variation and impact on discrimination

Among the 4804 STR alleles typed in the 89 individuals, there were 238 distinct length variants and 340 repeat sequence sub-variants identified by the UAS across the 27 loci (Table S1; Fig. 1a). The loci D17S1301 and D4S2408 presented the lowest diversity, with five alleles each; together with ten other STRs, these showed only length variation when visualised solely with the UAS, thus providing the same discrimination power as a conventional CE approach. The UAS highlighted additional sequence variation within the repeat regions of the remaining 15 loci, contributing between one extra variant (at D19S433, TH01 and CSF1PO) up to 25 additional alleles at D12S391 (Fig. 1a). Examination of the Flanking Region Report revealed a further 17 distinct variants including one, three, four and six additional alleles at D22S1045, D20S482, D16S539 and D7S820 respectively, whereas the visual interface of the UAS displayed only length variation. Two loci, D1S1656 and D2S441, each showed a total of six extra alleles as a result of sequence variation both within and flanking the repeat region, with SNPs in the flanking region contributing one and two of these extra alleles respectively. The sequences and frequencies of all autosomal STR alleles are shown in Table S1.

The number of additional alleles created by detection of sequence variation either within the repeat or flanking regions differs considerably between loci, but a more important parameter is how this affects the power of discrimination (PD). Fig. 1b shows PD for each of the autosomal STRs, subdividing this into the proportions due to repeat number (length), repeat region sequence and flanking region variation.

Twenty-three 'novel' alleles, summarised in Table 1 (and more fully described in Table S1), were absent from the STRait Razor v3.0 default allele list. However, fifteen had been reported previously either in ForenSeq™ DNA Signature Prep Kit-related literature [20] or in the STRSeq database [28] (accessed via GenBank, July 2019); of these, seven sequences were previously seen in the Middle East, with two of them not yet reported outside of the region. Of the 23 'novel' alleles, two are rare short simple repeat-number variants; fifteen are compound STR alleles with novel numerical combinations of repeat blocks; three have single-base insertions within a repeat array producing intermediate alleles; and the remaining three have SNP variants in their flanking regions. Of the six SNP variants, only two have existing entries in dbSNP (rs563997442, rs554502154).

In terms of simple repeat number nomenclature for comparison with CE data, we did not observe any unusual genotypes. However, some loci displayed read counts at stutter positions above the stutter threshold that could represent potential somatic triallelic patterns. The allele counts that exceeded the stutter threshold by the greatest margin were at D8S1179, where three alleles were called by the UAS at 12, 15, 16 with read depths of 247, 86 and 295 respectively. The lowest of these contributions was at a stutter position and would equate to a stutter proportion of 0.29 of allele 16, whereas the next strongest stutter observed at this locus in this study was 0.26, only fractionally higher than the recommended UAS default permissible stutter ratio of 0.25 for the locus. Similar triallelic patterns generated by CE have been reported previously at this locus [https://strbase.nist.gov and 29]. While imbalanced "triallelic" patterns can result from somatic mutations it is also possible that they are simply the result of unusually prominent stutter. It is unclear which explanation is appropriate in these two instances as the profiles are otherwise of good quality and do not show excessive stutter at other loci, which might have been indicative of overamplification.

3.2. Sequence variation at autosomal iiSNPs and in their flanking regions

Fifty-four of the 91 iiSNP amplicons included in the analysis showed no additional variation within the regions covered by the Flanking Region Report, while 27 have a second polymorphic site, eight have a third and two amplicons show variation at four positions (see Table S2). Combinations of alleles at two pairs and one trio of linked SNPs showed perfect associations (AT & TA at **rs279844** & rs279845, AT & GC at rs6950322 & **rs6955448**, ATT & CCC at rs409820, rs430044 & **rs430046** [target iiSNP in bold text]) resulting in just two distinct microhaplotypes per amplicon in our population sample. The lack of perfect association between polymorphic sites within the remaining 34 amplicons added to the diversity with 25, six, two and one amplicons displaying three, four, five and six microhaplotypes respectively. As with the autosomal STRs, we compare the number of additional alleles we observe through sequencing (Fig. 2a) with the increase in power of discrimination that they confer (Fig. 2b).

The UAS highlights the target SNP in each amplicon, and also some - though not all - additional variants in the flanking regions, in its Flanking Region Report. Some of these additional variants are already described in the online database dbSNP (build 151), while others are novel. Here, eight amplicons displayed variation at positions not highlighted by the UAS. Two of these additional SNPs were associated with the **rs1109037** amplicon, namely rs999755320, a rare C > T variant 1 bp upstream (at GRCh37 chr2:10085721 with a frequency of 2/30,976 in GnomAD) and a previously unrecorded T > C variant 4 bp downstream (@ 10085726). The other loci each displayed one additional rare or previously unrecorded variant site - six base substitutions and one indel; for these, dbSNP provides some information about previous observations within large genome-wide sequencing datasets. These studies include **TOPMed** - comprising participants in US National Heart, Lung, and Blood Institute studies (https://doi.org/10.1101/563866), **GnomAD** - the Genome Aggregation Database led by the Broad Institute (http://gnomad-old.broadinstitute.org/about), and an

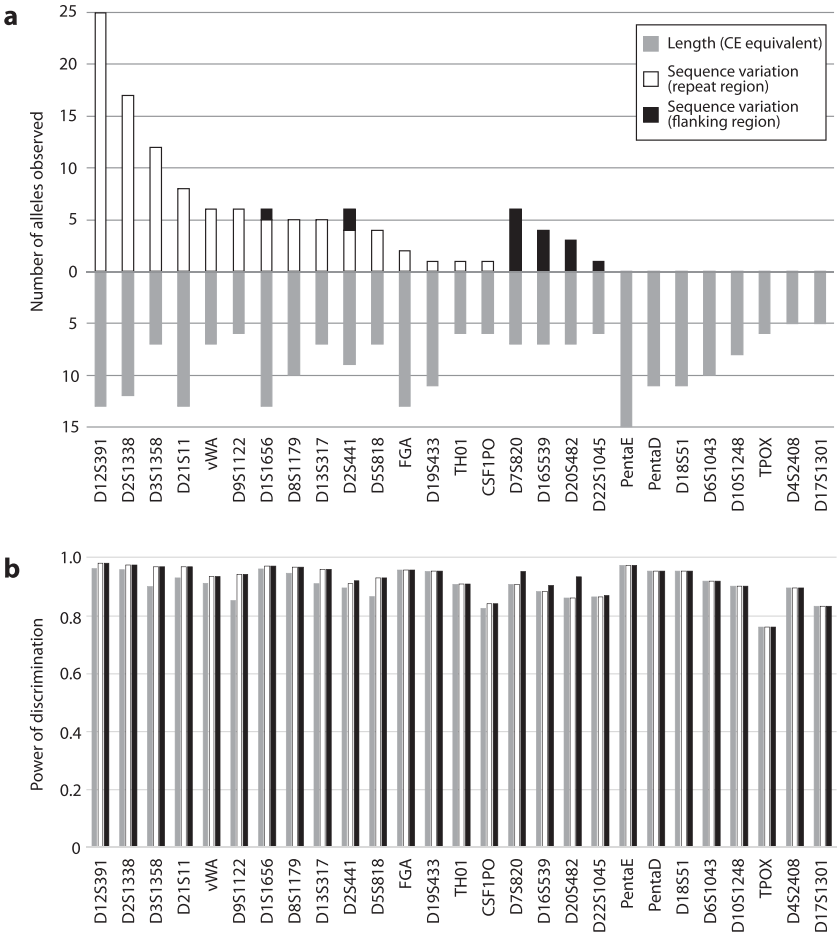


Fig. 1. Counts of distinguishable alleles by STR locus, and per-locus increment of discriminatory power due to sequence variants.

a) The observed number of length variants among the 89 Saudi males is shown in grey below the x-axis, and the number of additional alleles resulting from sequence variation within and flanking the repeat array are shown above in white and black respectively. b) The power of discrimination resulting from length variation alone (equivalent to CE typing) is shown in grey and the additional contributions made by sequence variation within and flanking the repeat array are shown in white and black respectively (see also Table S3).

Table 1

Novel STR alleles found in this study. Twenty-two STR alleles not previously recorded in the STRait Razor v3.0 default allele database are summarised here together with the allele structure as suggested by Parson et al. 2016 [25] the nature of its novelty (repeat length [RL], repeat region sequence [RS] or flanking sequence [FS]), its occurrence within the Saudi dataset (Obs.) and the geographical distribution of matching alleles in HGDP metapopulations [20]. AFR: African; EUR: European; MEA: Middle East; OCE: Oceanian; EAS: East Asian; SAS: South Asian; n/a: not available for search, since flanking variants are not listed in the literature [20].

Nomenclature	Type	Observations	HGDP occurrence
CSF1PO [CE 12]-GRCh38-Chr5-150076318-150076389 ATCT ACCT (ATCT)10	RS	3	MEA
D10S1248 [CE 9]-GRCh38-Chr10 129294226-129294318 (GGAA)9	RL	2	MEA/AFR
D12S391 [CE 23]-GRCh38-Chr12-12296981-12297168 (AGAT)12 (AGAC)11	RS	1	EUR/OCE/SAS
D12S391 [CE 23]-GRCh38-Chr12-12296981-12297168 (AGAT)12 (AGAC)10 AGAT	RS	1	EUR
D12S391 [CE 26]-GRCh38-Chr12-12296981-12297168 (AGAT)17 (AGAC)8 AGAT	RS	1	EAS
D13S317 [CE 9]-GRCh38-Chr13-82147986-82148107 (TATC)8 AATC	RS	1	not observed
D16S539 [CE 8]-GRCh38-Chr16-86352664-86352781 (GATA)8 86352692-G (rs563997442)	FS	1	n/a
D1S1656 [CE 17]-GRCh38-Chr1-230769555-230769682 CCTA (TCTA)16 230769682-G (rs NA)	FS	1	n/a
D21S11 [CE 36]-GRCh38-Chr21-19181939-19182111 (TCTA)11 (TCTG)6 (TCTA)3 TA (TCTA)3 TCA (TCTA)2 TCCA TA (TCTA)11	RS	1	not observed
D22S1045 [CE 15]-GRCh38-Chr22-37140181-37140357 (ATT)12 ACT (ATT)2 37140182-A (rs554502154)	FS	1	n/a
D2S1338 [CE 15]-GRCh38-Chr2-218014856-218014964 (GGAA)9 (GGCA)6	RS	1	EUR
D2S441 [CE 9]-GRCh38-Chr2-68011918-68012017 (TCTA)9	RL	1	MEA/EUR/AFR/EAS
D3S1358 [CE 13]-GRCh38-Chr3-45540691-45540820 TCTA (TCTG)3 (TCTA)9	RS	1	not observed
D3S1358 [CE 17]-GRCh38-Chr3-45540691-45540820 TCTA (TCTG)2 TCTC (TCTA)13	RS	2	MEA/SAS
D3S1358 [CE 18]-GRCh38-Chr3-45540691-45540820 TCTA (TCTG)2 TCTC (TCTA)14	RS	1	not observed
D3S1358 [CE 18]-GRCh38-Chr3-45540691-45540820 TCTA (TCTG)4 (TCTA)13	RS	1	MEA/SAS
D3S1358 [CE 19]-GRCh38-Chr3-45540691-45540820 TCTA (TCTG)4 (TCTA)14	RS	1	MEA
D6S1043 [CE 25]-GRCh38-Chr6-91740160-91740292 (ATCT)6 ATGT (ATCT)4 ATGT (ATCT)13	RS	1	not observed
D8S1179 [CE 17.1]-GRCh38-Chr8-124894867-124894921 (TCTA)2 (TCTG)2 (TCTA)12 TCTTA	RS	1	not observed
D9S1122 [CE 13]-GRCh38-Chr9-77073809-77073880 TAGA TCGA (TAGA)8 CAGA (TAGA)2	RS	1	not observed
FGA [CE 16.1]-GRCh38-Chr4-154587713-154587823 (GGAA)2 GGAG (AAAG)3 A (AAAG)5 AGAA AAAA (GAAA)3 154587760-A	RS	1	not observed
PentaE [CE 16.4]-GRCh38-Chr15-96830996-96831114 (TCTTT)16 TCTT	RS	2	MEA/SAS
TH01 [CE 6]-GRCh38-Chr11-2171056-2171127 (AATG)3 AATA (AATG)2	RS	1	not observed

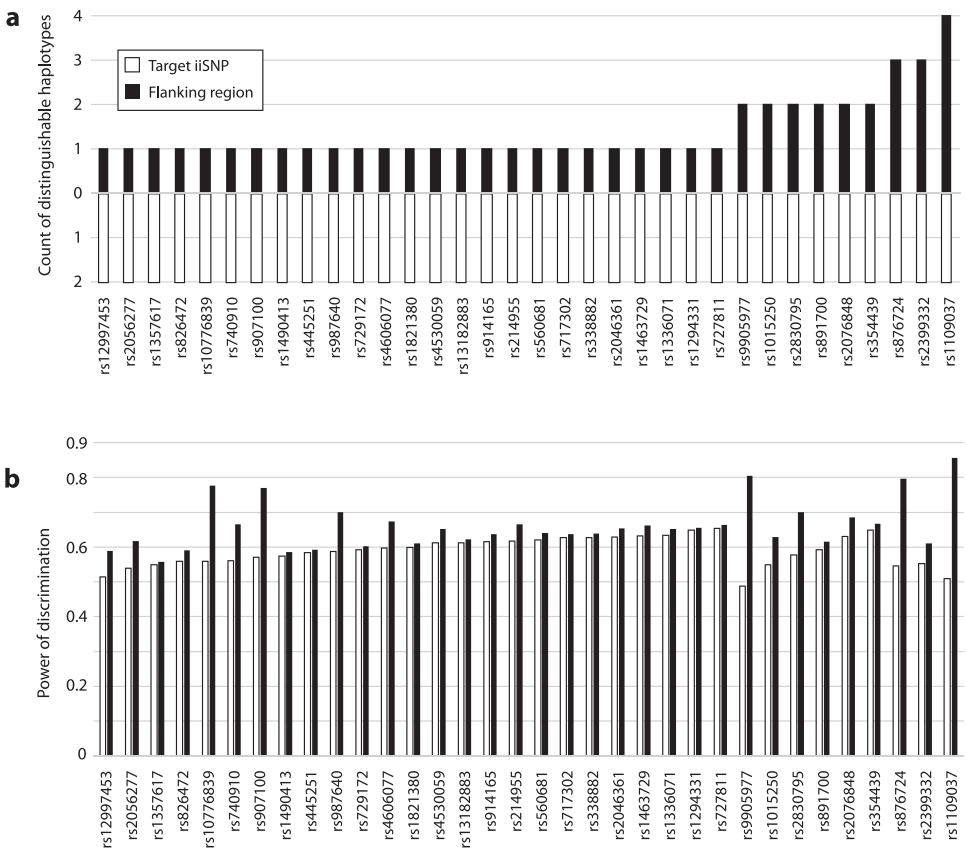


Fig. 2. Counts of distinguishable haplotypes by SNP locus, and increment of discriminatory power due to sequence variants.

a) The observed number of haplotypes defined by the targeted SNP base alone among the 89 Saudi males is shown in white below the x-axis, and the number of additional haplotypes resulting from sequence variation due to flanking SNPs is shown above the axis in black. b) The probability of discrimination resulting from target SNP variation alone is shown in white and the additional contributions made by flanking SNPs are shown in black (also Table S4).

Estonian dataset comprising genetic variation from a pharmacogenomics study of adverse drug effects using electronic health records (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA489787).

- **rs1015250** (+ rs1307278892 G/C, 9 bp downstream @ 1823783, TOPMed freq 5/125,568);
- **rs1294331** (+ rs92280418 A/G, 4 bp upstream @ 233448409 [UAS reports reverse strand], GnomAD freq 1/30,950);
- **rs13182883** (+ novel A/G SNP, 21 bp upstream @ 136633317),
- **rs338882** (+ rs527589535 G/A SNP, 100 bp upstream @ 178890625 [UAS reports reverse strand], no frequency data available);
- **rs354439** (+ novel A/C SNP, 30 bp upstream @ 106938381 [UAS reports reverse strand]);
- **rs729172** (+ rs556717752 G/A, 32 bp upstream @ 233448409 [UAS reports reverse strand], GnomAD freq 2/30,984);
- **rs891700** (+ rs543563536, T insertion in run of 10 Ts, 8 bp downstream @ chr1:239881909-239881918, freq 8/4480 among Estonians, also reported in https://doi.org/10.1016/j.fsigen.2017.09.003).

All of these were confirmed as true alleles displaying heterozygote balance between 52–98%. The sequences and frequencies of all iiSNP alleles are shown in Table S2, and those instances described above where the UAS flanking sequence report does not draw attention to nine additional polymorphic sites are included.

3.3. Possible recombination and recurrent mutation within SNP microhaplotypes

All four possible combinations of variants at co-amplified pairs of biallelic SNPs were observed within four amplicons (target iiSNP in bold text): **rs2830795** & rs12626695 (separated by 43 bp), **rs2399332** & rs2399334 (separated by 100 bp), **rs1109037** & rs1109038 (separated by 63 bp) and **rs2076848** & rs7947725 (separated by 21 bp). Although recombination could explain these, the alternative explanation of recurrent mutation at a hypermutable CpG site is applicable in the last two cases and would seem most probable given the close proximity of two SNPs. The likely product of a CpG to CpA transformation at rs7947725 was seen in one and five individuals linked to the minor (A) and major (T) allele at rs2076848, respectively. At rs1109037 it is possible that the target SNP itself may have experienced recurrent mutation also resulting in a change from CpG to CpA, as a result of methylcytosine conversion to thymine on the opposing strand (see Table S1 for frequencies and sequence context).

3.4. Effect of combining autosomal STR and SNP sequence variants

One of the major potential benefits of typing with the DNA Signature Prep Kit is the increase in discrimination power when genotype data from independently inherited autosomal SNPs and STRs are combined. As there was no significant deviation from LD either among or between STRs and SNPs following Bonferroni correction, it is possible to combine the two marker types using the product rule. Forensic statistics for the autosomal STRs and SNPs are presented in Tables S3 and S4. In our dataset, we found that the RMP decreased by three orders of magnitude due to sequence variation within the STR repeat region itself compared to length variation alone, whilst sequence polymorphisms in the flanking regions of the STRs resulted in less than one order of magnitude further reduction (Fig. 3). We compared the Saudi Arabian data with data from four comparator populations from the USA [15,30]: the increases in resolution offered by sequencing in the Saudi Arabian sample are comparable with the other four groups. Notably, the difference between STR-based and iiSNP-based resolution is smaller in the African-Americans than in the other groups, suggesting that iiSNPs may be underestimating diversity in this particular population.

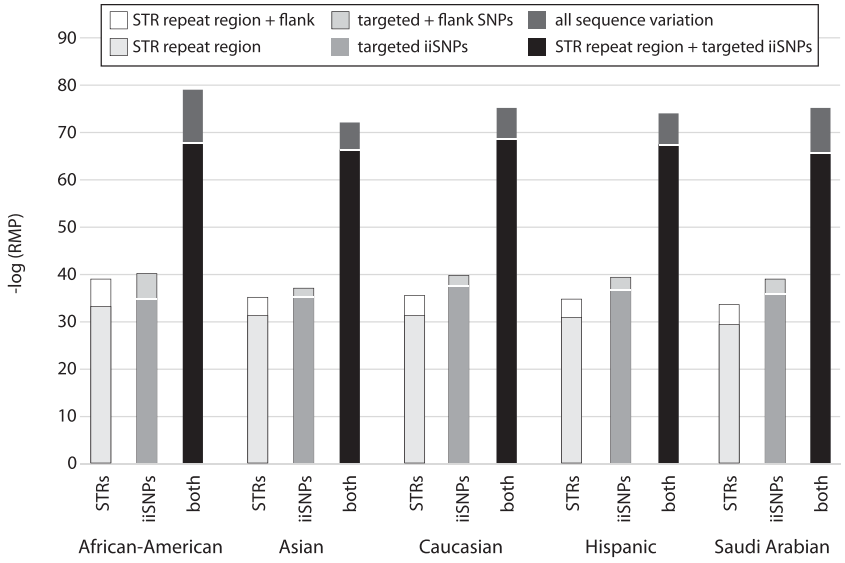


Fig. 3. Increase in random match probability (RMP) offered by flanking sequence variation for both STRs and iiSNPs. Data are shown for the Saudi Arabians studied here, and for four comparator populations from the literature [30].

3.5. Evidence of consanguinity from patterns of STR and SNP diversity

Previously, we provided evidence from CE-based autosomal STR genotype data of the impact of consanguinity in the Saudi Arabian population [19]. Here, we ask if the sequence-based analysis of both STRs and SNPs showed a similar impact.

Among STRs, testing for fit to Hardy-Weinberg expectations revealed significant heterozygote deficiency only at TPOX ($P < 0.005$ after Bonferroni correction). TPOX shows no internal sequence variation, simply mirroring the observed CE length variability, and was completely concordant with the results obtained from these individuals (data not shown) with the AmpFLSTR® GlobalFiler Kit (Thermo Fisher Scientific). At the CE length level of comparison, 23 out of the 27 STR loci showed an excess of homozygotes ($P = 0.0002$, binomial distribution, http://onlinestatbook.com/2/calculators/binomial_dist.html; see Table S5). This is consistent with 20 out of 21 GlobalFiler Kit loci displaying heterozygote deficiency ($P < 0.0001$, binomial distribution) in our CE dataset of 523 indigenous males resident in Saudi Arabia (designated KSA) [19]. F_{ST} between the UK-based Saudi and KSA datasets at both regional (N, E, S, W and Central) and combined levels showed no evidence of significant differences between corresponding samples, and thus we consider our UK-based Saudi Arabian individuals to be a representative sample of autosomal variation within the indigenous population of the country as a whole ($F_{ST} = 0.00086$; P-value 0.09791). Heterozygote deficiency was also evident from AMOVA analysis, with an inbreeding coefficient (F_{IS}) of 0.04131 representing 4.13% of variation among our UK donors at the CE level, very close to the value we obtained with GlobalFiler from our KSA dataset ($F_{IS} = 0.0476$; [19]). None of the iiSNPs showed a significant deviation from Hardy-Weinberg equilibrium after Bonferroni correction, but 63 of the 91 iiSNPs included in the analysis showed a deficiency of heterozygotes ($P = 0.0002$, binomial distribution) mirroring the effect seen with autosomal STRs (see Table S6).

3.6. Comparing Saudi Arabian STR and SNP sequence diversity patterns with a set of global population samples

In order to set the observed Saudi Arabian population diversity and apparent effects of consanguinity in a broader context, we extracted data from the well-characterised HGDP samples [22]: STR sequence data were available from Phillips et al. (2018) [20] and iiSNP data were extracted from publicly available whole-genome sequences (see Table S7). Here we considered only the target iiSNPs rather than including flanking variants, as these are most likely to be reported in casework, and analysed only the 27 HGDP populations that contain 15 or more individuals. In Fig. 4a we compare F_{IS} values between our Saudi dataset and these HGDP populations. European, African and East Asian populations generally show wide variation in F_{IS}, with a tendency to low values; a particularly broad range is seen in central South Asian populations. However, the five populations from the Middle East, including the Saudi Arabians studied here, show consistently raised and highly similar levels indicative of widespread and frequent endogamy. In Fig. 4b we also show, for the HGDP samples only, the proportion of specific mating types (unrelated, first cousin, second cousin, avuncular) estimated by a maximum-likelihood method from the distribution of homozygous-by-descent segments from high-density genome-wide SNP data by Leutenegger et al. (2011) [31]. This illustrates the high levels of consanguineous mating types in the Middle Eastern samples, though it also shows that some Central and South Asian samples exceed these levels.

For the full set of HGDP populations we also compared the level of homozygosity observed from the iiSNPs with an estimate based on haplotype homozygosity along chromosome 16 from genome-wide SNP data previously published by Li et al. (2008) [32] (Fig. S1). Again, the iiSNP-based estimates are similar for all Middle Eastern populations (including our Saudi sample), and this is also true for the haplotype homozygosity estimates for the HGDP Middle Eastern samples. The haplotype homozygosity shows a trend to increase from Africa to the rest of the Old World and to the New World, previously noted both from genome-wide SNP [32] and STR data [33]. However, it is striking that the iiSNP-based homozygosity values do not follow this trend, with marked underestimation of both African and Middle Eastern diversity (Fig. S1b), probably reflecting ascertainment bias in the original choice of SNPs for individual identification.

4. Discussion

A global survey of consanguinity has shown that the Middle East, North and sub-Saharan Africa and Western, Central and South Asia have consanguineous marriage levels between 20% and 50% [34]. Arabic-speaking countries showed the highest rates in the Middle East [35]. Indeed first-cousin marriages are particularly common and preferred in many regional communities, including the socially conservative regions of Saudi Arabia [36] where they comprise up to 33%

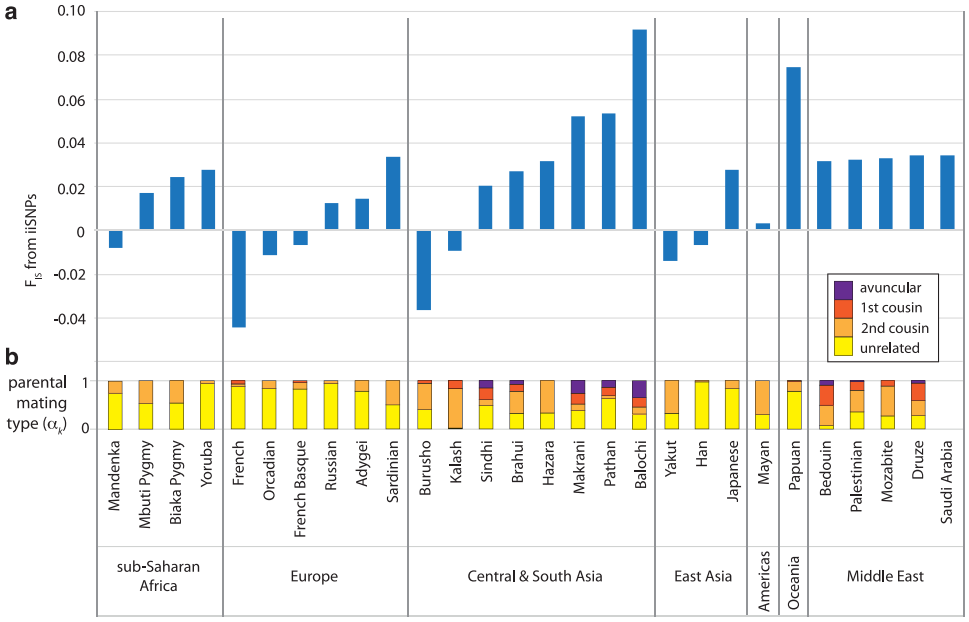


Fig. 4. F_{IS} for autosomal iiSNPs in Saudi Arabian and HGDP populations, compared with consanguineous mating type frequencies.

a) Mean F_{IS} is shown (target iiSNPs only) for HGDP populations with n ≥ 15 and for Saudi Arabian data, clustered by continental origins; b) The coloured bars show genome-wide estimates of parental mating types for the HGDP samples, taken from Leutenegger et al. (2011) [31]. Different mating types are shown in the inset key.

of marriages. Saudi Arabia is also remarkable for its very rapid expansion in population size from 3.9 million in 1950 prior to the oil boom, to 20.8 million indigenous Saudi citizens in 2018, largely as a result of decreased child mortality and increasing life expectancy. Consequently, it is expected that clusters of closely related and genetically similar individuals will be commonplace, requiring more discriminating tests to generate the same confidence of individual resolution. Here, we used the high (sequence-based) resolution of MPS analysis, and the large number of independent autosomal loci (27 STRs and 91 iiSNPs) to analyse a sample of Saudi Arabians in order to address this question.

The lack of significant differentiation between the UK-based sample of Saudi Arabians analysed here and our previously reported [19] combined KSA dataset (F_{ST} = 0.00072), suggests that our autosomal STR sequence dataset is representative of autosomal variation in the indigenous Saudi population as a whole. This study also replicates previous findings of heterozygote deficiency in Saudi populations [19,37] likely due to high levels of endogamy (F_{IS} = 0.04131 for autosomal STRs based on CE length and 0.04201 for iiSNPs) although the loci showing significant deviations differ between studies. Our earlier observation of an apparent significant deficiency of heterozygotes in the KSA population at D2S1338, thought to be partly due to a null allele, was not replicated in this study. However, this is unsurprising as the primer positions differ between the GlobalFiler and DNA Signature Prep kits, the latter of which has an amplicon approximately 180 bp shorter. In contrast the apparent homozygous excess seen at TPOX in this study was clearly not due to null heterozygotes caused specifically by the DNA Signature Prep Kit primers, as all homozygous individuals were also scored as homozygotes with GlobalFiler, and we did not see a significant excess of such “homozygotes” in a much larger KSA dataset typed with the GlobalFiler kit (Khubrani et al. 2019). It seems likely that TPOX by chance showed the greatest excess of homozygosity due to consanguinity. Significant homozygous excess has been detected in many populations in the Middle East affecting a variety of loci: Iraq - D21S11, FGA [38], Qatar - D13S317, D19S433 and vWA [39], Kuwait - vWA [40] and Tunisia - CSF1PO, Penta D and TPOX [41]. The lack of consistency across loci suggests that null alleles are not the general cause, and that demographic and stochastic factors are more likely.

The detection of sequence variation within autosomal STR amplicons enhances resolution by distinguishing among isoalleles that appear identical in CE analysis. Previous surveys of diverse ethnic groups have revealed differences of several orders of magnitude in random match probability (RMP) between length- and sequence-based estimates. Novroski et al. (2016) [30] obtained the following length- and sequence-RMP estimates for African-Americans (8.54E-34, 1.31E-39), Asians (6.37E-32, 8.66E-36), Caucasians (6.28E-32, 3.63E-36) and Hispanics (1.51E-31, 1.23E-35), whilst the levels of resolution were lower within our Saudi population (2.61E-30- length, 2.07E-33 - repeat region sequence & 3.49E-34 - including both repeat and flanking sequence). This reflects their more restricted geographic origins and therefore expected lower levels of diversity.

In previous studies, although sequencing of some STRs provided up to two-fold increased resolution per locus, other loci such as D10S1248, Penta D, Penta E and TPOX showed no improvement, as we observe in this dataset [2,7,42]. This is heavily influenced by the complexity of the STR repeat structure, but novel variants can arise in other STRs with a typically simple repeat structure as demonstrated by CSF1PO and TH01, both of which were invariant in previous studies but here displayed rare or regionally restricted variants that may be seen more frequently as datasets of Middle Eastern populations grow. Nine of the STR loci showed no improvement with sequencing, but six showed a greater than two-fold increase within the Saudi population. These include D12S391, where 13 length variants increased to 38 sequence variants, accompanied by a two-fold improvement in RMP from 0.0357 to 0.0173; and D3S1358, which saw a three-fold improvement linked to an increase from 7 length variants to 13 sequence variants. Whereas most gains were due to novel arrangements of repeat variants in compound STRs, a greater than two-fold improvement resulted at D20S482 due solely to the presence of SNPs flanking the repeat array.

As with the STRs, sequencing of the iiSNP amplicons revealed additional sequence variants within the regions flanking the target SNP. Notably, not all of these variants were highlighted in the UAS Flanking Region Report, so care is needed in analysing data to ensure that all variation is recorded. Of the 91 iiSNPs analysed here, 34 showed additional sequence variation. This led to the RMP of 9.97E-37 for the target iiSNPs alone decreasing to 8.88E-40 when the flanking sequence variants were included. Most additional variants create novel rare microhaplotypes within our sample. The microhaplotypes created by the flanking sequence variation deserve further investigation in larger and more diverse samples.

It has been reported that the iiSNPs alone in the DNA Signature Prep Kit provide greater discrimination power than do available commercial STR kits [15], and this was certainly true within our Saudi dataset: taking sequence variation into account, the RMP of the autosomal STRs

was 3.49E-34 compared with 8.88E-40 for the iiSNP amplicons. Overall, our study demonstrates that, while the additional alleles revealed by MPS substantially improve discrimination, the greatest contribution comes from the ability to analyse many independent loci simultaneously. The added power offered by combining STRs and SNPs will be beneficial in challenging cases, such as mixture analysis, complex kinship cases and where degradation results in partial profiles.

5. Conclusion

MPS analysis of autosomal loci in a representative sample of Saudi males demonstrated that profiling can significantly increase the discrimination power of forensic DNA testing through the simultaneous amplification of both STRs and identity-informative SNPs. This small-scale survey has uncovered a number of previously unobserved and rare alleles that may be present at significant frequencies within the Arabian Peninsula, supporting the value of larger-scale surveys of this region. A striking feature of the data was a general deficiency of heterozygosity, and comparison with HGDP samples supports the idea that this reflects the practice of consanguinity in Saudi Arabia and in other Middle Eastern populations.

Declaration of Competing Interest

None.

Acknowledgments

YMK was supported by the Saudi Arabian Ministry of Interior, and by a PhD studentship grant from the Saudi Arabian Cultural Bureau, London. PH was supported by Estonian Research Council Grants PUT1036 and IUT34-12. We thank all DNA donors, NUCLEUS Genomic Services at the University of Leicester for access to FGx sequencing, and Tunde Huszar and Verogen (Nicola Oldroyd-Clark, Sarah Naif, Richard Kessell and Cydne Holt) for useful discussions. We also thank Chris Phillips for sharing STR sequence data on the HGDP samples, and Chris Tyler-Smith for allowing the extraction of iiSNP genotypes from the HGDP WGS data. This research used the SPECTRE High Performance Computing Facility at the University of Leicester for data analysis.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.fsigen.2019.102164.

References

[1] P. de Knijff, From next generation sequencing to now generation sequencing in forensics, Forensic Sci. Int. Genet. 38 (2019) 175–180.
[2] J.D. Churchill, S.E. Schmedes, J.L. King, B. Budowle, Evaluation of the Illumina Beta Version ForenSeq DNA Signature Prep KIit for use in genetic profiling, Forensic Sci. Int. Genet. 20 (2016) 20–29.
[3] N. Almalki, H.Y. Chow, V. Sharma, K. Hart, D. Siegel, E. Wurmbach, Systematic assessment of the performance of Illumina's MiSeq FGx™ forensic genomics system, Electrophoresis 38 (2017) 846–854.
[4] S. Iozzi, I. Carboni, E. Contini, C. Pescucci, S. Frusconi, A.L. Nutini, et al., Forensic genetics in NGS era: new frontiers for massively parallel typing, Forensic Sci. Int. Genet. Suppl. Ser. 5 (2015) e418–e419.
[5] A.L. Silvia, N. Shugarts, J. Smith, A preliminary assessment of the ForenSeq™ FGx System: next generation sequencing of an STR and SNP multiplex, Int. J. Legal Med. 131 (2017) 73–86.
[6] C. Xavier, W. Parson, Evaluation of the Illumina ForenSeq DNA Signature Prep Kit – MPS forensic application for the MiSeq FGx™ benchtop sequencer, Forensic Sci. Int. Genet. 28 (2017) 188–194.
[7] R.S. Just, L.I. Moreno, J.B. Smerick, J.A. Irwin, Performance and concordance of the ForenSeq™ system for autosomal and Y chromosome short tandem repeat sequencing of reference-type specimens, Forensic Sci. Int. Genet. 28 (2017) 1–9.
[8] A.C. Jäger, M.L. Alvarez, C.P. Davis, E. Guzmán, Y. Han, L. Way, 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. 28 (2017) 52–70.
[9] V. Sharma, H.Y. Chow, D. Siegel, E. Wurmbach, Qualitative and quantitative assessment of Illumina's forensic STR and SNP kits on MiSeq FGx™, PLoS One 12 (2017) e0187932.
[10] S. Köcher, P. Müller, B. Berger, M. Bodner, W. Parson, L. Roewer, et al., Inter-laboratory validation study of the ForenSeq™ DNA Signature Prep Kit, Forensic Sci. Int. Genet. 36 (2018) 77–85.
[11] F. Guo, J. Yu, L. Zhang, J. Li, Massively parallel sequencing of forensic STRs and SNPs using the Illumina ForenSeq DNA Signature Prep Kit on the MiSeq FGx Forensic Genomics System, Forensic Sci. Int. Genet. 31 (2017) 135–148.
[12] Y. Ma, J.-Z. Kuang, T.-G. Nie, W. Zhu, Z. Yang, Next generation sequencing: improved resolution for paternal/maternal duos analysis, Forensic Sci. Int. Genet. 24 (2016) 83–85.
[13] E. Almohammed, D. Zgonjanin, A. Iyengar, D. Ballard, L. Devesse, H. Sibte, A study of degraded skeletal samples using ForenSeq DNA Signature™ Kit, Forensic Sci. Int. Genet. Suppl. Ser. 6 (2017) e410–e412.
[14] J. Votrubova, A. Ambers, B. Budowle, D. Vanek, Comparison of standard capillary electrophoresis based genotyping method and ForenSeq DNA Signature Prep kit (Illumina) on a set of challenging samples, Forensic Sci. Int. Genet. Suppl. Ser. 6 (2017) e140–e142.
[15] J.L. King, J.D. Churchill, N.M.M. Novroski, X. Zeng, D.H. Warshauer, L.-H. Seah, et al., Increasing the discrimination power of ancestry- and identity-informative SNP loci within the ForenSeq™ DNA Signature Prep Kit, Forensic Sci. Int. Genet. 36 (2018) 60–76.
[16] Scientific Working Group on DNA Analysis Methods (SWGDAM), Addendum to the SWGDAM Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories to Address Next Generation Sequencing, (2019) https://www.swgdam.org/publications.
[17] L. Al-Gazali, H. Hamamy, S. Al-Arrayad, Genetic disorders in the Arab world, Br. Med. J. 333 (2006) 831–834.
[18] Y.M. Khubrani, J.H. Wetton, M.A. Jobling, Extensive geographical and social structure in the paternal lineages of Saudi Arabia revealed by analysis of 27 Y-STRs, Forensic Sci. Int. Genet. 33 (2018) 98–105.
[19] Y.M. Khubrani, J.H. Wetton, M.A. Jobling, Analysis of 21 autosomal STRs in Saudi Arabia reveals population structure and the influence of consanguinity, Forensic Sci. Int. Genet. 39 (2019) 97–102.
[20] C. Phillips, L. Devesse, D. Ballard, L. van Weert, M. de la Puente, S. Melis, 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 39 (2018) 2708–2724.
[21] T.E. King, S.J. Ballereau, K.E. Schürer, M.A. Jobling, Genetic signatures of coancestry within surnames, Curr. Biol. 16 (2006) 384–388.
[22] H.M. Cann, C. De Toma, L. Cazes, M.-F. Legrand, V. Morel, L. Piouffre, et al., A human genome diversity cell line panel, Science 296 (2002) 261–262.
[23] H. Li, A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data, Bioinformatics 27 (2011) 2987–2993.
[24] A.E. Woerner, J.L. King, B. Budowle, Fast STR allele identification with STRait Razor 3.0, Forensic Sci. Int. Genet. 30 (2017) 18–23.
[25] W. Parson, D. Ballard, B. Budowle, J.M. Butler, K.B. Gettings, P. Gill, 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. 22 (2016) 54–63.
[26] L. Excoffier, 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. 10 (2010) 564–567.
[27] A. Gouy, M. Zieger, STRAF—a convenient online tool for STR data evaluation in forensic genetics, Forensic Sci. Int. Genet. 30 (2017) 148–151.
[28] K.B. Gettings, L.A. Borsuk, D. Ballard, M. Bodner, B. Budowle, L. Devesse, et al., STRSeq: a catalog of sequence diversity at human identification Short Tandem Repeat loci, Forensic Sci. Int. Genet. 31 (2017) 111–117.
[29] G. Mertens, S. Rand, E. Jehaes, N. Mommers, E. Cardoen, I. De Bruyn, et al., Observation of tri-allelic patterns in autosomal STRs during routine casework, Forensic Sci. Int. Genet. Suppl. Ser. 2 (2009) 38–40.
[30] N.M.M. Novroski, J.L. King, J.D. Churchill, L.H. Seah, B. Budowle, Characterization of genetic sequence variation of 58 STR loci in four major population groups, Forensic Sci. Int. Genet. 25 (2016) 214–226.
[31] A.-L. Leutenegger, M. Sahbatou, S. Gazal, H. Cann, E. Génin, Consanguinity around the world: what do the genomic data of the HGDP-CEPH diversity panel tell us? Eur. J. Hum. Genet. 19 (2011) 583.
[32] J.Z. Li, D.M. Absher, H. Tang, A.M. Southwick, A.M. Casto, S. Ramachandran, et al., Worldwide human relationships inferred from genome-wide patterns of variation, Science 319 (2008) 1100–1104.
[33] S. Ramachandran, O. Deshpande, C.C. Roseman, N.A. Rosenberg, M.W. Feldman, L.L. Cavalli-Sforza, Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa, Proc. Natl. Acad. Sci. U. S. A. 102 (2005) 15942–15947.
[34] A.H. Bittles, Consanguinity in Context, Cambridge University Press, Cambridge, 2012.
[35] G.O. Tadmouri, P. Nair, T. Obeid, M.T. Al Ali, N. Al Khaja, H.A. Hamamy, Consanguinity and reproductive health among Arabs, Reprod. Health 6 (2009) 17.
[36] M.I. El-Mouzan, A.A. Al-Salloum, A.S. Al-Herbish, M.M. Qurachi, A.A. Al-Omar, Regional variations in the prevalence of consanguinity in Saudi Arabia, Saudi Med. J. 28 (2007) 1881–1884.
[37] H.M. Alsafiah, W.H. Goodwin, S. Hadi, M.A. Alshaikhi, P.-P. Wepeba, Population genetic data for 21 autosomal STR loci for the Saudi Arabian population using the GlobalFiler® PCR amplification kit, Forensic Sci. Int. Genet. 31 (2017) e59–e61.

[38] M.M. Farhan, S. Hadi, A. Iyengar, W. Goodwin, Population genetic data for 20 autosomal STR loci in an Iraqi Arab population: application to the identification of human remains, Forensic Sci. Int. Genet. 25 (2016) e10–e11.
[39] A.M. Perez-Miranda, M.A. Alfonso-Sanchez, J.A. Pena, R.J. Herrera, Qatari DNA variation at a crossroad of human migrations, Hum. Hered. 61 (2006) 67–79.
[40] M. Al-enizi, J. Ge, A. Salih, H. Alenizi, J. Al jabber, J. Ziab, et al., Population data on 25 autosomal STRs for 500 unrelated Kuwaitis, Forensic Sci. Int. Genet. 12 (2014) 126–127.
[41] C. Brandt-Casadevall, M. Ben Dhiab, F. Taroni, C. Gehrig, N. Dimo-Simonin, M. Zemni, et al., Tunisian population allele frequencies for 15 PCR-based loci, Int. Congr. Ser. 1239 (2003) 113–116.
[42] K.B. Gettings, K.M. Kiesler, S.A. Faith, E. Montano, C.H. Baker, B.A. Young, et al., Sequence variation of 22 autosomal STR loci detected by next generation sequencing, Forensic Sci. Int. Genet. 21 (2016) 15–21.