ASSESSING THE EFFECTS OF SUBPOPULATIONS ON THE APPLICATION OF FORENSIC DNA PROFILING

By

Daniel John Clark

A thesis submitted in partial fulfilment for the requirements of the degree of Doctor of Philosophy at the University of Central Lancashire

February 2013

STUDENT DECLARATION

I declare that while registered as a candidate for the research degree, I have not been a registered candidate or enrolled student for another award of the University or other academic or professional institution.

I declare that no material contained in the thesis has been used in any other submission for an academic award and is solely my own work.

Signature of Candidate…….....

Type of Award – Doctor of Philosophy

School – Forensic & Investigative Sciences

ABSTRACT

Currently, UK forensic service providers (FSPs) tend to employ three geographically-broad databases when estimating profile frequencies based on a standard SGM Plus® DNA profile. These estimations will typically include correction factors to take into account issues such as substructuring of populations and sampling inefficiencies. It has been shown previously that regional genetic variation within the UK 'Caucasian' population is negligible but consideration has to be made for profiles which may originate from an individual of a more genetically isolated population.

Samples were collected from Indian, Pakistani and UK (white British) donors; as well as Kalash individuals, a small population from the Khyber Pakhtunkhwa region in the North West of Pakistan. These were profiled using the SGM Plus® and Identifiler® kits and databases for each population were compiled.

The greatest pairwise F_{ST} was seen between the Kalash and Indian population at 2.9 %. Allele frequency data were collected for each population and each sample's profile frequency was estimated against all other databases to see whether samples reported a more conservative profile frequency (higher match probability) in their cognate database or in that of another population. A combined database comprising the Indian, Pakistani and previously published Bangladeshi data was also formed and used to calculate the level of correction required to make all samples of a population report a more conservative profile frequency in this combined database as opposed to their cognates. At the standard F_{ST} correction of 3 % – the minimum correction used by some FSPs, 94 % of the UK samples reported a more conservative profile frequency in the South Asian database; the lowest proportion that did so from all four populations. The Kalash dataset required the highest correction factor at F_{ST} = 12 % to make 100 % of samples report more conservative match probabilities when measured against the combined database.

It was established that the current levels of correction applied to profile frequency calculations were more than sufficient; with random match probabilities remaining in the order of less than one in one billion for all samples in all databases with a correction of F_{ST} = 5 %. Although significant pairwise F_{ST} differences were observed as well as significant differentiation between populations across all SGM Plus® loci, no evidence of substructuring was detected using a program which employs a Bayesian probabilistic clustering approach, STRUCTURE, likely due to an insufficient number of samples and number of loci tested.

Marked differences were seen in allele frequencies of the Kalash population, which also exhibited the highest affiliation to their cognate database, at least 80 %, with or without correction. AMOVA analysis also confirmed the greatest variance between groups was seen when the Kalash were kept as a separate entity from the other South Asian populations.

Although current UK practice for applying F_{ST} correction prior to estimating STR match probabilities seems generous, there will be occasions when an estimation may appear less conservative when based on a broad database. Conversely, in this study, the one in one billion match probability ceiling threshold was not exceeded for any sample being compared to all databases. Therefore, although consideration should be given to a suspect's reference population prior to frequency estimation, the current correction factors applied should be sufficient in the vast majority of cases. In instances where partial profiles are obtained, this caused little effect on the estimation of geographic origin, compared to full profiles, with the populations used in this study.

CONTENTS

LIST OF TABLES, FIGURES AND GRAPHS

ACKNOWLEDGEMENTS

I would like to thank the staff and fellow students of the School of Forensic and Investigative Sciences for their support and help over the past few years. Whether it be advice, technical support or just a chat, someone was always on hand. I would also like to thank the School for the opportunities given to me to attend conferences, seminars and training; all of which helped to further this study in different ways.

My sincere thanks go to my supervisors, Dr William Goodwin and Dr Sibte Hadi. Without their encouragement and support, the project would not have got to this stage. I would also like to thank them for the samples they collected that formed the basis of this study. Special thanks to Dr Arati Iyengar and Dr Judith Smith for their continued support and advice during the years.

To my family and friends for all of their support, a very sincere thank you; not just with this study but with the things in between – I could not have got this far without you.

Finally a very special thank you to my wife, Nicola, who has ‘suffered’ with me during most of the time I have been working on this project. Your constant love and support have been invaluable and have kept me going when times seemed tough.

ABBREVIATIONS

AMOVA	Analysis of Molecular Variance
bp	base pair
DNA	Deoxyribonucleic acid
H_o	Observed heterozygosity
H_e	Expected heterozygosity
mtDNA	Mitochondrial DNA
NDNAD	National DNA Database
PCR	Polymerase Chain Reaction
PD	Power of Discrimination
PE	Power of Exclusion
PIC	Polymorphism Information Content
SNP	Single Nucleotide Polymorphism
STR	Short Tandem Repeat
TPI	Typical Paternity Index
VNTR	Variable Number Tandem Repeats
v/v	Volume/volume
w/v	Weight/volume
YBP	Years Before Present

1 INTRODUCTION

1.1 History of DNA Profiling

The first recorded use of a genetic polymorphism being used as an investigative tool was the discovery of differences between individuals in the ABO blood grouping system (Landsteiner, 1900). Although poor in terms of discriminatory power, any variable biological system bearing genetic variation can be used to eliminate someone from an investigation. By the 1980s, various serological systems were utilised to analyse variations between individuals which may lead to a genetic identification, albeit one of low informativeness with high match probabilities compared to today (Jobling & Gill, 2004).

In 1985, Professor Sir Alec Jeffreys showed that highly variable repeating segments of DNA (minisatellites) existed within the human genome. These minisatellites ranged in size, typically from 300 – 10,000 bp (Nakamura *et al.*, 1987), containing core units of repeated sequences approximately 10 – 100 bp in length (Jobling & Gill, 2004). Jeffreys found that the number of core repeats in these minisatellites varied from person to person across multiple loci and that the 'multi-locus probe' (MLP) technique could be used to identify a particular person from a sample of biological material (Jeffreys *et al.*, 1985).

Jeffreys' initial discovery employed the use of the Southern Blot technique (Southern, 1975). The technique relied on the migration of DNA fragments, having been previously digested by restriction enzymes and separated by weight, to a nitrocellulose membrane. From this a radioactively labelled probe, specific to the flanking region of the area of interest, would hybridize with the fragments to reveal the hypervariable multi-band patterning synonymous with early DNA fingerprinting.

1.2 Polymorphic Markers

1.2.1 Single-locus Probes (SLPs)

Following from Jeffreys' work on MLPs, in the late 1980s, minisatellites that were highly polymorphic were applied to criminal casework. Using single-locus probes (SLPs) a maximum of two alleles were seen in an individual which aided interpretation. When several SLP loci were analysed simultaneously, this allowed for the generation of a relatively simple genetic profile which could be stored on a database and compared to the profiles of other individuals. The development of SLP analysis ultimately secured the first criminal conviction based on DNA evidence; that of Colin Pitchfork, accused of a double rape and murder in Leicestershire in 1986. By the early 1990s, SLP analysis was in routine use in paternity testing and criminal casework, albeit mainly utilised for serious offences (Greenhalgh *et al.*, 1992). Generating a SLP profile was laborious and time consuming and the result offered low discriminatory power and mixtures were difficult to interpret (Jobling & Gill, 2004).

1.2.2 Short Tandem Repeats (STRs)

During the continued use of SLPs, work was on going into the characterisation of the next generation of highly polymorphic markers, STRs (Edwards *et al*., 1991). With the relatively recent discovery of the extremely-sensitive PCR amplification method (Mullis *et al*., 1986), minute amounts of DNA, including degraded DNA (Schmerer *et al.*, 1999), could be amplified relatively easily (Saiki *et al.,* 1988). This was not the first time PCR-based technology was employed in DNA profiling (discussed later), but combined with STRs, the potential of a highly-sensitive, highly-discriminating method of individual identification on a molecular level was being realised.

1.2.2.1 STR Structure

STRs are known as 'microsatellites'. Smaller than their 'minisatellite' counterparts, they also contain a variable number of tandem repeats, typically between 1 and 6 bp in size (Figure 1.1) and although initially reported as being less than 300 bp in total, per allele (Kimpton *et al.*, 1993), modern multiplex kits (discussed later), such as Promega's PowerPlex® 16, generate amplicons of just under 500 bp (Butler, *et al.*, 2003).

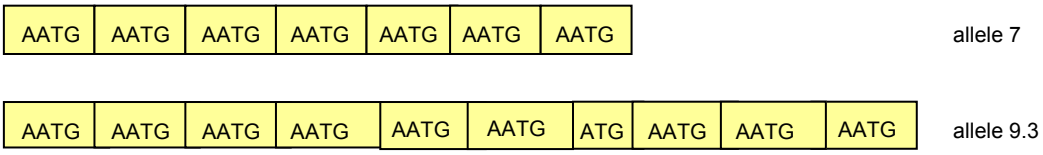


Figure 1.1: Schematic diagram of two alleles comprising repeat units at the TH01 locus

The variety and diversity of STR markers make them a desirable choice when it comes to DNA profiling. Ranging from simple core repeat units, such as TH01, (Edwards *et al.,* 1991) to complex ones such as D21 (Sharma & Litt, 1992), comprising several different core repeat units with intervening non-repeated sequences. STR markers are highly polymorphic and it is their susceptibility to mutation that maintains this.

1.2.2.2 STR Markers in Forensic Analysis

STR markers can be analysed from a variety of biological materials, including semen (Shewale *et al.,* 2003), blood (Gill, *et al.*, 1990), hair (including hair shafts) (Higuchi *et al.,* 1988; Barbaro, *et al.*, 2000) and skin cells (van Oorschot & Jones, 1997). Of the three-billion base pairs comprising the human genome, it was previously thought that approximately 90 % of them do not play a role in protein

synthesis and are essentially 'non-coding' and therefore considered to be 'junk' DNA. It was within this 90 % of the genome where the majority of the tandemly-repeated sequences lie that enable a unique profile to be established (Schneider, 1997). However, it has recently been reported that just over 80 % of the entire genome plays a functional role in the biochemistry of cells; for example, chromatin structure and transcription-factor binding sites (see section 1.2.3.4) (The Encode Project Consortium, 2012).

The Forensic Science Service developed the first multiplex systems for use in human identification in 1993 (Kimpton *et al.,* 1993). At the time, a DNA profile could be attained from as little as 100 picograms (Frégeau & Fourney, 1993). This was followed by the six loci (plus the sex-determining amelogenin locus [Sullivan *et al.,* 1993]) kit, known as Second Generation Multiplex (SGM) and manufactured by Applied Biosystems (Sparkes *et al.,* 1996) which became the kit of choice to instigate the UK's National DNA Database (NDNAD) (Werrett, 1997). In 1999, the kit currently used today as the industry standard was developed: SGM Plus®. This contained the same six loci and amelogenin components as the original kit plus four more loci to make a profile more discriminatory: D3S1358, D16S539, D2S1338 and D19S433 (Cotton *et al.,* 2000). In 2010, the NGM SElect™ kit was released which added an additional seven loci including the highly polymorphic SE33 locus (Green *et al.,* 2012). Loci concordance between samples profiled with this advanced kit and the same loci of the SGM Plus® kit is vital if NGM SElect™ is to ever succeed in becoming the new profiling standard in the UK.

By March 2012, the UK held the DNA profiles of over 5.9 million individuals (NPIA, 2012). Samples suitable for the database are held from anyone convicted of a crime and also, since 2004, anyone arrested for a recordable offence, even if subsequently released without charge. Conversely, in 2008, a European Court of Human Rights upheld an action brought by 'Mr S' and Michael Marper of the United

Kingdom in which they contested the retention of their DNA profiles on the NDNAD. Having been arrested but later acquitted or having charges dropped against them, they were successful in having their DNA profiles removed as it was deemed a violation of their privacy (S. and Marper v. The United Kingdom - 30562/04 [2008] ECHR 1581 [4 December 2008]).

The profiles held on the UK's NDNAD comprise STR markers included in the Applied Biosystems' AmpFℓSTR® SGM Plus® PCR Amplification Kit. Although no longer the most powerful kit for obtaining the most discriminatory match probabilities (Identifiler® also by Applied Biosystems contains 15 loci plus amelogenin), the SGM Plus® kit was a major advance in DNA profiling just a few years after the database was setup in 1995. Irrespective of this, it can provide match probabilities low enough to be admissible evidence in court, usually far exceeding the ultra-conservative 'ceiling principle' match probability of one in one billion when a complete or almost complete DNA profile is obtained.

Match probability calculations are based on allele and genotype frequencies. The allele frequencies at a particular locus are used to calculate genotype frequencies using Hardy-Weinberg predictions. The frequency of that particular genotype is then multiplied together with the frequencies of other loci comprising the particular multiplex kit employed to generate a match probability for the DNA profile. This method of obtaining a profile frequency is termed the 'product rule'. Typically, based on SGM Plus® loci, the most common DNA profile determined from genotypes with the greatest frequency of occurrence in a sample of 'UK Caucasians' would return a profile frequency of approximately 2.0×10^{-10} ($F_{ST} = 0$ and size bias correction applied [Balding & Nichols, 1994]) (Foreman & Evett, 2001).

The idea of the 'ceiling principle' was proposed by the National Research Council in 1992. It allowed for an extremely conservative estimate of match probabilities by using population data from whichever ethnic group reported the most common

allele frequency at a particular locus, regardless of the defendant's purported ethnic background. In theory, by choosing the maximum allele frequency for every marker analysed (hence the term 'ceiling-principle'), the benefit of the doubt was given to the defendant and this would rarely be challenged by defence counsel (Lander and Budowle, 1994).

Not everyone agreed though; challenges were made as to the validity of the assumptions made that small, or isolate, populations would be under-represented in general allele frequency databases (Devlin, *et al.*, 1993), particularly where rare alleles were concerned (NRC, 1996). Another study argues that the ceiling principle effectively means allele frequencies are so artificially inflated, they are essentially arbitrary numbers and makes massive assumptions about population structuring (Morton, *et al.*, 1993). However, the principle is still used in criminal courts in the UK as a way of demonstrating the magnitude of the DNA evidence without over-estimating it or confusing the court with match probabilities of inconceivable proportions (Foreman & Evett, 2001).

1.2.2.3 Allele Frequency Databases

Variation in STR allele frequencies between different groups of people has been previously reported (Bowcock *et al*., 1994; Lowe *et al*., 2001) and has been utilised in an attempt to draw inferences on genetic ancestry using clustering algorithms, specifically exploiting detectable differences in allele frequencies between different populations (Rosenberg *et al.,* 2002). This may be of interest in both criminal investigations and studies of population genetics; however, for it to be feasible, databases need to be created to analyse the effect this variation has on local gene pools. The forensic community is constantly updated with population data, of which allele frequency data of isolated populations may be of particular interest. From a criminal justice perspective, criminal reference databases of many countries have

seen increased growth year-on-year, particularly the UK: aided by the five year initiation by the Home Office of the DNA Expansion Programme in 2000 (Walsh *et al.*, 2008).

As DNA profiling becomes more commonplace and cost-effective, there has been a surge in published population databases over recent years, covering geographically-broad areas to small, isolate locations, for example – the tribal population of Orissa, India (Alshamali *et al*., 2005; Barni *et al*., 2007; Binda *et al*., 2000; Clark *et al*., 2009; Gehrig *et al*., 1999; Hadi *et al*., 2004; Havas *et al*., 2007; Junge, *et al.*, 2001; Kashyap *et al*., 2006a; Marian *et al*., 2007; Maruyama *et al*., 2008; Nussbaumer *et al*., 2001; Parson *et al*., 1998; Reichenpfader, *et al.*, 2003; Sahoo & Kashyap, 2002; Yong *et al*., 2007a; Yong *et al*., 2007b).

1.2.2.4 STR Variability and Mutation

Mutational events such as base insertion, deletion and substitution can all cause allele variants. A major cause of variation, slipped-strand mispairing, can increase the number of variants created and is one reason for the differences seen between simple and complex markers at different loci. This occurs when DNA strands denature as part of the replication process and then misalign upon annealing. As replication continues, additional repeat units may be added or deleted. The longer the repeated sequences get, the greater the chance of slipped-strand mispairing, potentially leading to additional allele variants (Levinson & Gutman, 1987). Recombination and errors in the DNA repair mechanisms can also lead to allele mutation (Jobling & Gill, 2004; Jarne & Lagoda, 1996).

The majority of markers used in forensic DNA profiling consist of tetranucleotide repeat units (Butler, 2006). Originally, a study by Weber and Wong (1993) suggested that tetranucleotide repeat units had a mutation rate four times higher than that of dinucleotide repeats. The average mutation rate across all STRs was

given as 1.2 x 10^{-3} per locus, per gamete, per generation, though only chromosome 19 was studied in this case. It is recognised that mutation rates can differ not only between loci, but that alleles on the same locus can mutate at different rates (Brinkmann *et al.,* 1998). Other studies disagree with Weber and Wong (1993) and argue that it is dinucleotides that have the highest mutation rate, then tri-, then tetranucleotides (Edwards *et al.,* 1992; Kruglyak *et al.,* 1998). Kelkar *et al.*, (2008) showed that the number of repeat motifs affected whether mono-, di-, tri- or tetra-nucleotide repeats induced the greatest mutability. At 15 repeat units, tetranucleotides showed the greatest mutability whereas, at five repeats, dinucleotides were more susceptible to mutation. Their findings complement those of Levinson and Gutman (1987) reporting that greater motif and overall microsatellite length lead to greater rates of mutation and express a practically exponential relationship. The view that mono- and di- nucleotide repeats have greater mutation rates may be due to the comparable lack of high-repeat tri- and tetra- nucleotide microsatellites (Kelkar *et al.*, 2008).

Regardless of mutation rate, there is evidence to suggest that longer microsatellites are more unstable and therefore highly likely to mutate to a shorter microsatellite (Lai & Sun, 2003). In contrast, shorter microsatellites are more likely to expand and add repeat units. Providing an appropriate and representative sample of the population has been taken, this is unlikely to have a significant adverse effect on population data. However, mutations may prove more problematic if those used in identity and relationship testing are affected.

With the current robust commercial kits available on the market, STR markers continue to be at the forefront of forensic DNA analysis. Their highly polymorphic nature, ability to amplify from minute template amounts, and high discriminatory power make them a reliable tool in human identity and paternity testing.

1.2.3 Single Nucleotide Polymorphisms (SNPs)

SNPs are mostly bi-allelic which means their basic discriminatory power is less than that of markers such as STRs (approximately 50 SNPs would be required to give a comparable discriminatory level equal to that of 12 STRs [Gill, 2001]), but there are many more in the genome to examine (The International HapMap Consortium, 2010). SNPs occur at a rate of around 1 per 1000 – 2000 bp (Barbujani & Colonna, 2010; The International SNP Map Working Group, 2001) and having a lower mutation rate than STRs (Nachman & Crowell, 2000), they are able to provide a more accurate historical record of human genetic diversity. Because of their short length, i.e. one base pair, SNPs are ideal for the analysis of degraded DNA as they can be amplified on short amplicons (Alaeddini, *et al.*, 2010; Dixon *et al.*, 2006).

Over the last decade, several studies have been dedicated to investigating the patterns of human genetic diversity and variation, of which SNPs play a major part in this. These, amongst others, include the HapMap, HGDP (Human Genome Diversity Project), 1000 genomes, Genographic Project and ENCODE (Encyclopedia of DNA Elements). Collectively, it is estimated these studies have analysed genotypes of 85,000 individuals from populations spanning the globe (Novembre & Ramachandran, 2010).

1.2.3.1 The International HapMap Project

The International HapMap project was launched in 2002 with the initial aim of genotyping over 1,000,000 SNPs to study patterns of variation, frequencies, relationships between SNPs and nearby variants as well as their role in identifying risks of disease through genome-wide association studies (of particular importance to facilitate advances in diagnostic tools and clinical research) (The International HapMap Consortium, 2003). Populations from Africa, Asia and Europe were selected to provide the 270 samples used in the study, all of whom gave informed consent prior to donating their DNA. Each of the six countries comprising institutions which formed The International HapMap Consortium had a different role to play, including sampling, genotyping and analysis.

SNPs were chosen based on those reported in previous studies, those known to appear in greater than one sample and those forming part of coding sequences. Variations in sequences are reported as haplotypes: the specific combination of alleles along a chromosome or a particular area of a chromosome (The International HapMap Consortium, 2003).

The project has advanced considerably over the years and the data derived from it have aided other studies in areas such as genetic distances between STRs (Phillips *et al.*, 2012) and the identification of recombination 'hotspots' – again, useful in disease studies and the characterisation of patterns seen among haplotypes (Li, *et al.*, 2006).

The first set of results (Phase I) were published in 2005, providing insight into patterns of linkage disequilibrium (LD), recombination hotspots and the relationships between neighbouring SNPs. It also provided the first estimates as to the genetic variation seen between the population panels; the most variation being seen between the Yoruba population from Ibadan, Nigeria, and the combined Han

Chinese from Beijing, China and Japanese from Tokyo, Japan panels. The Consortium do, however, give caution to comparisons with data derived from different marker types, for example, STRs (The International HapMap Consortium, 2005).

Phase II, published in 2007, characterised a further 2.1 million SNPs studied from the same initial set of volunteers. These data improved and fine-tuned the results from Phase I, providing greater haplotype resolution, and detection of recombinant hotspots (The International HapMap Consortium, 2007).

Phase III brought the total number of volunteers to 1,184 and included individuals from populations not previously included in Phases I and II of the study such as people with Chinese ancestry living in Denver, Colorado; Gujarati Indians based in Houston, Texas and Maasai from Kinyawa, Kenya. A principal part of this study was to include the previously omitted minor allele frequencies (SNP alleles with a frequency of 0.05 or less) with the aim of scanning across populations for signs of variation as opposed to interrogating individual populations. Perhaps not unexpectedly, informativeness of low-frequency variants was greater between closely related populations, for example, the Chinese and Japanese data panels in this case (The International HapMap 3 Consortium, 2010).

Although The International HapMap Project was set up to look into sequence variation and common patterns of genetic variation with the aim of clinical advances, it has also served to improve our understanding of human genetic diversity and variation both within and between populations. This is, potentially, a powerful tool in attempting to assign people to populations based on DNA evidence.

1.2.3.2 The Human Genome Diversity Project

The HGDP set out to genotype 650,000 SNPs in over 1,000 individuals representing 52 populations from around the world. The samples were collected in collaboration with the Centre d'Etude du Polymorphisme Humain (CEPH), the data collected from which being known as the 'HGDP-CEPH Human Genome Diversity Cell Line Panel' (Cann *et al.*, 2002).

The panel has been used in several studies as a reference base of genetic data, of particular interest are those which have shown a correlation between genetic diversity and geography (Ramachandran *et al.*, 2005; Liu, H *et al.*, 2006) and those which examine forensic core loci variation between populations and their usefulness in inferring genetic ancestry (Phillips, *et al.*, 2011). Use of the same reference data will inevitably make comparison of data easier between studies. Unlike The International HapMap Project, HGDP focussed on shared ancestry and the influences of geography; some of the populations included in the study were relatively isolated, which was not the remit of The International HapMap Project. As reported by Cavalli-Sforza in 2005 though, there was no reason why the two studies could not complement each other in terms of understanding human genetic variation from both a historical and geographical perspective.

A study by Rosenberg *et al.*, 2002, was one of the first to utilise the data collected by the HGDP. Although discussed later in more detail, they demonstrated the proportions of genetic variation accountable by differences among major population groups (identified as part of the study) and those within-populations. This was later developed in a study by Li *et al.*, 2008, who, using the same samples collected for the HGDP, studied 650,000 SNPs and reported a high correlation between donor ancestry and population substructure. Its results were largely concordant with that of Rosenberg *et al.*, 2002, but with a higher resolution of genetic differentiation

between populations, likely due to the high number of SNPs analysed compared to the 377 autosomal microsatellite markers utilised by Rosenberg *et al.*, 2002.

1.2.3.3 The 1000 Genomes Project

This project was setup with the aim of sequencing, identifying and cataloguing over 95 % of all forms of polymorphic genomic variation in humans. The dataset utilised as part of the HapMap project was used here and variants with lower frequencies (down to 10^{-3}) were also included in the results (The 1000 Genomes Project Consortium, 2010).

Similar to the HapMap project in its aim: quantifying the risks of genetic variants linked to disease susceptibility, the 1000 genome project concentrated more on the phenotypic consequences of such variants and for providing a reference database for genome-wide association studies.

Although not setup as a tool to investigate genetic diversity of the populations studied, the data naturally lend themselves to such interpretation. As with the HapMap project, genetic variation was at its greatest (although still relatively low, considering isolated populations were deliberately not sought) between the Yoruba and combined Han Chinese and Japanese panels.

1.2.3.4 Encyclopedia of DNA Elements (ENCODE)

ENCODE was setup in 2003 and its purpose was to examine and assign function to the 99 % of the genome which did not code for the approximately 20,000 genes comprising the remaining 1 %. It complements the work of the Human Genome Project which sought to identify and sequence these genes. ENCODE identified approximately 80 % of the genome had a functional role whether it be involved it gene expression or genomic structure (Maher, 2012; The Encode Project Consortium, 2012). The project has served to enhance previous knowledge

gathered in the genetics of diseases, such as cancer, and may, therefore, also offer a greater insight into more specific sequence variation between populations.

1.3 Human Evolution and Migration

To better understand the modern-day relationships between populations, it is important to acknowledge the underlying models of human evolution and migration which have led to the differences seen in allele frequencies. Geography is a key factor in understanding demographic expansion and the genetic diversity of human populations. There are two main models of human population expansion, each part-characterised by differing geographical migration patterns.

Until the late 1980s the prevailing model for the evolution of modern *Homo sapiens* was the multiregional model (Figure 1.2) (Thorne & Wolpoff, 1992). Under this model, *Homo erectus* migrated out of Africa approximately one million years ago and colonised large parts of Eurasia by 500,000 to 800,000 years before present (YBP). Different populations of *Homo erectus* then developed into modern humans with gene flow between the populations which prevented speciation (Turner & Chamberlain, 1989). Arguments against this theory include the high levels of gene flow between Eurasia and Africa which would be necessary for the multiregional model to be plausible (Harding & McVean, 2004).

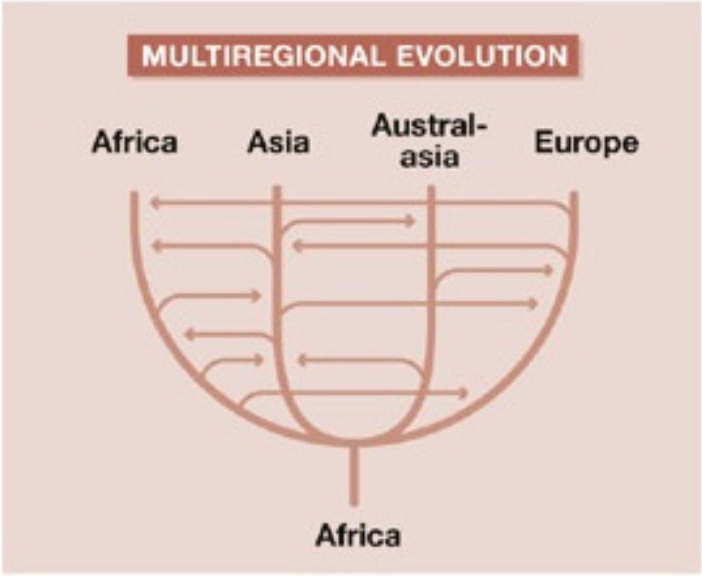


Figure 1.2: Diagram representing the multiregional model. Each population was established after *Homo erectus* left Africa and the horizontal arrows show migration and gene flow between them (taken from Stoneking, 2008)

This model has largely been replaced by the out of Africa model of human evolution, where modern *Homo sapiens* evolved within Africa around 150,000 YBP and moved from Africa around 70-80,000 YBP, colonising large parts of Asia by 60,000 YBP (Torroni *et al*., 2006) and Europe by 30-40,000 YBP. This model is supported by the highest levels of diversity often seen among sub-Saharan African populations compared with non-African populations (Excoffier, 2002). If high levels of gene flow did occur between continents, this level of diversity would not be as easily explained if favouring the multiregional model. It is believed that small, isolate populations were well established long before the expanse of modern humans from Africa which would have provided much of the diversity seen in Africa today (Behar *et al.,* 2008). In addition, studies have suggested that a serial founder effect model is responsible for the relationship between genetic variation and geographic distance (Ramachandran *et al.*, 2005; Deshpande *et al.*, 2009). This suggests that there was a single site of human expansion, based in Africa, where geographic progression occurred in small steps, with periods of settlement in between. Each progression carried with it a small subset of the previous population. With each

migration, genetic variation is reduced, hence the correlation between genetic and geographic distance.

Genetic differences amongst the major population groups can be explained by genetic drift and selection operating on the populations in different regions (Stringer & Gamble, 1994). The out of Africa model has been supported by both the archaeological record and several genetic studies that have concluded that phylogenetic trees constructed using human DNA data root predominantly in Africa, which is evidence that Africa was the original home of the modern human genome about 150-200,000 YBP (Cann, *et al.*, 1987; Hammer *et al*., 1998).

There are some that suggest the fossil evidence of East Asia supports multiregional evolution (Wolpoff, *et al.*, 2000). However, it is acknowledged that Western Europe did not fully comply with this theory: lacking in explanation for the fate of the Neanderthals. There is some evidence of admixture of modern humans with archaic hominins, such as the Neanderthals, though studies have taken different approaches and have come to different conclusions. Some have determined that Neanderthals are distinct from modern humans although sequencing was restricted to the hypervariable regions of mtDNA (Briggs *et al.*, 2009). Conversely, the possibility that Neanderthals interacted with anatomically modern humans has raised questions as to the possibility of gene flow between the two species sometime when they co-existed between 30-80,000 YBP (Green, *et al.*, 2010). This study made significant advances in attempts to sequence a whole Neanderthal genome. Interestingly, the study found that between 1-4 % of Eurasian genomes derived from those of the Neanderthals. Although this seems to contradict the out of Africa model, the authors note that the majority of genetic variation is still likely to have originated from the evolution of modern humans within Africa.

Regardless of which theory, or combination thereof, may have occurred, a major factor of human migration is the genetic diversity it facilitates. This, in turn, allows

for research to trace back the genetic relationships between and within populations, increasing the understanding of human evolution and the effects of factors such as drift, migration and selection.

1.4 Race and Genetic Variation

The terms ‘race’ and ‘ethnicity’ have provided much debate between sociologists and biologists regarding their definition and use. They are often used, perhaps incorrectly, as interchangeable terms and for many years social scientists and biologists have been debating the differences between the two. According to Hutchinson & Smith (1996), ‘ethnicity’ can mean ‘the quality of belonging to an ethnic community or group’, or ‘what it is you have if you are an “ethnic group”’. Essentially, ‘ethnicity’ allows for a self-defining classification based on an individual’s socio-cultural perceptions.

A group of people may call themselves a race or an ethnic group, but it does not mean genetic differences will be clearly distinguishable from one group to the next (Lee, *et al.*, 2001). Due to this ambiguity, race is a term rarely used in biological classification (Triggs, *et al.*, 2000), as little scientific evidence exists to support the theories that biological differences account for racial differences (Lillie-Blanton & Laveist, 1996).

It is for this reason that the term ‘geographical origin’ may be more appropriate; or a system based on socio-anthropological attributes when describing an individual rather than trying to assign a race or ethnic group (Schneider, 2007).

1.4.1 Early Racial Classifications

The early classical racial categories, Caucasian, Negroid and Mongoloid were first put forward by Georges Cuvier in 1828 and were based on skull morphology and skin colour (Barbujani and Colonna, 2010). The term ‘Caucasian’ as a classification of race was first suggested by Blumenbach, an 18th Century anthropologist, and was named after the peoples of the Southern Caucasus region. Today, it is synonymous with ‘White’ as a term for racial categorisation (Bhopal and Donaldson, 1998). Although such broad terms such as ‘White: British’ and ‘Afro-Caribbean’ are

still widely used today, for example, for research purposes or by Government agencies (census data, police records), they serve only as a proxy for some of the biological and environmental factors that lead to varying phenotypic characteristics (Bamshad *et al.*, 2004).

1.4.2 Race as a Biological Marker

The geneticist's approach to race and ethnicity is that biological classifications stem back to early population genetics and provide a means of distinguishing between two different groups of people. There is a presumed homogeneity of different groups or races, where genetic variation within them is greater than that between them (Barbujani *et al.*, 1997). This also enables scientists to categorise biological differences by use of these social classifications and makes comparison of data simpler to comprehend (Foster & Sharp, 2002).

However, as Dyson (1998) stated, Foster and Sharp comment that race and ethnicity is mainly attributed to cultural and historical aspects and has very little to do with biological patterns. Goodman (2000) says that race cannot be systematically classified because there is no race 'scale' to measure it against. As race continuously diversifies with time and place (due to virtually no geographical limit on where people may wish to settle and raise offspring), classifying people into a race becomes an increasingly unreliable task. Therefore, it is the opinion of some authors that there is no biological basis for race and that using biology as a 'marker' to determine race is inaccurate (Goodman, 2000; Risch *et al*., 2002).

1.4.3 Genetic Variation

As discussed previously, mutation is a key factor behind establishing new alleles and introducing genetic variation but it is genetic drift that has a particularly profound effect on small, isolate populations where gene flow between populations may also be limited. These differences play a major role in the study of genetic

ancestry and disease susceptibility among populations. It is reasonable to assume that patterns of genetic variation are closely linked to the geographical spread of populations and the distance between them.

1.4.3.1 Subpopulations

A subpopulation arises when there are sufficient recognisable differences between it and the larger population. An indication of relatedness between the people of a subpopulation can be estimated based on the premise that alleles in a particular subpopulation differ in frequency when compared to other subpopulations, and the population as a whole. Complications are introduced when you have to consider varying degrees of substructure.

Defining ‘the larger’ or ‘the whole’ population is difficult but there are different hierarchal levels which may be used as a starting point to examine genetic variability therein. For example, the UK ‘Caucasian’ and Indian populations are considered to be within the same metapopulation: a term established by Levins (1969) to define a population within a population that exhibited spatial separation, largely via geographical means. This does not mean that there is no migration, or gene flow, between populations within the same metapopulation but it may therefore be reasonable to treat them as subpopulations.

In addition, it is not unlikely that further substructuring will have occurred within the subpopulations, but these may not be as readily identifiable. For those that are known, and coupled with social and cultural differences, it is reasonable to allow the use of different allele frequency databases based on previous studies of genetic variability.

1.4.3.2 Measuring Genetic Variation

Population differentiation can be quantified by using a series of fixation, or *F*, statistics and can express relationships between alleles at different hierarchical levels including individuals relative to the total population (F_{IT}), individuals relative to a subpopulation (F_{IS}) and a subpopulation relative to the total population (F_{ST}) (Wright, 1965). They are essentially inbreeding coefficients so the higher the F_{ST} value, the greater the variance of allele frequencies between subpopulations within the total population due to the inbreeding occurring within them. Where allele frequencies are equal between populations, F_{ST} will be 0; if allele frequencies become fixed in subpopulations, F_{ST} will increase up to 1, showing less genetic diversity in the subpopulation compared to the total population (Bamshad *et al.*, 2004). Factors such as random mating, natural selection and mutation can all affect the level of F_{ST} and as it rises, the chance of assigning an individual to a particular subpopulation based on differences in allele frequencies increases, especially if an allele is fixed to just one subpopulation.

$F_{ST,}$ also described as θ, is of particular interest in forensic DNA profiling as it takes into consideration the effects of co-ancestry or the proportion of alleles sharing a common ancestor within a particular subpopulation (Balding & Nichols, 1994).

1.4.3.3 Difference in Genetic Diversity between Humans and Other Primates

F_{ST} levels in humans are relatively low (Li *et al.*, 2008), with comparisons between pairs of broad racial classifications at approximately 0.035 or lower (Foreman & Lambert, 2000). A study of samples stored on the NDNAD in the UK, originating from donors classed as 'Caucasian' reported regional pairwise F_{ST} differences of less than 0.005 (Foreman, *et al.*, 1998).

In comparison, a study looking at two gorilla species in equatorial Africa, western (*Gorilla gorilla*) and eastern (*Gorilla beringei*) showed significant variation at 0.38 with a distance of 1000 km between them (Thalmann *et al.*, 2006). Considering that humans cover a broader geographic area, we show one of the lowest levels of primate genetic diversity (Kaessmann *et al.*, 2001) suggesting a recent origin and that admixture helps maintain our relative homogeny compared with one of our closest genetic relatives.

1.4.3.4 Genetic Variation and Racial Classification

An early study into genetic variation and the classification of humans into broad racial groups reported that nearly 85 % of all human genetic variation occurred between individuals of the same population (Lewontin, 1972). Therefore, with just over 15 % being accounted for by differences between groups, Lewontin concluded that there was no genetic basis for the application of categorising individuals into groups and people appeared different because of the amount of their individual genetic variation; not their attribution to a specific race.

Nearly 20 years later, Lewontin's analysis was criticised for not considering the patterns of correlation in the data showing the differences between groups, which, according to Edwards (2003), is imperative to understanding population structuring and not simply relying on the level of allelic variation between populations at a number of loci. To show the effect of such a rudimentary approach, a Korean study (Ahn *et al.*, 2009) compared the entire genome sequence of a Korean male with two European male subjects. Each of the European males shared more SNPs with the Korean male than each other. This does not, however, suggest that Europeans are genetically more alike Koreans than each other, but highlights the need to look at where the data are different and any correlations that become apparent. It also reinforces the notion that broad categorisation of individuals does not always work

when you have individuals in one supposed ‘race’ genetically more distinct than those in another.

Edwards also commented on Lewontin’s analysis of just 17 polymorphic loci and that greater accuracy in group affiliation can be achieved when more loci are included in the classification. Although within-population variation may appear consistent once a certain number of loci have been analysed, increasing the number of markers serves to reduce the probability of a misclassification.

A subsequent study by Barbujani *et al.* (1997) demonstrated significant differences in diversity between individuals within the same population at all but two of the 109 DNA markers studied. They reported almost complete concordance with Lewontin’s (1972) work, concluding that 84.4 % of the genetic variation of human populations is due to between-individual differences in the same population. This further compounded Lewontin’s argument that there was no genetic basis for racial classification of humans as they found no discontinuities within their data to support this. They also agreed that genetic variation between individuals of different population groups only slightly exceeded the levels of variation between individuals within the same population. Barbujani *et al.* do, however, self-critically declare that their sample size was limited and may not have been sufficient to comment on the feasibility of racial classification.

A key study by Rosenberg *et al.*, (2002), one of the first to use the HGDP panel of samples, further increased the amount of genetic variation among within-population individuals to 93 – 95 %. Despite, therefore, reducing the estimate of diversity between groups, they were still able to show sample clustering to one of six broad geographical regions, without knowledge of the donor’s ancestral background using an algorithm-based program: STRUCTURE (Pritchard, *et al.*, 2000). This program seeks to identify the hidden correlation structure, described by Edwards (2003), within genotypic data and classifies samples into clusters. Of the 4199 alleles

studied, nearly 47 % appeared in each of the major geographical groups, with just over 7 % appearing in one group only. It is the distinctiveness of those rarer alleles that allow samples to be grouped into clusters based on a probabilistic approach. It also shows how each sample could be apportioned into alternative clusters if they exhibit allele frequencies akin to another cluster. With that in mind, STRUCTURE is most efficient in detecting relatively homogenous populations and those with rare alleles.

Levels of variation and diversity can also differ depending on which markers you assess. For example, Mitochondrial DNA (mtDNA) shows greater continent-specific variation than nuclear DNA: approximately 35 % compared to 12 % (Wallace, *et al.*, 1999). However, certain genetic traits are more prevalent than others in different populations. For example, mutations in the Melanocortin 1 Receptor (MC1R) gene which affects skin colour and is often associated with having red hair (Rees, 2000), vary between populations: largely through selective pressure. For example, eumelanin production (dark pigmentation) is particularly high in African populations where selective pressures have imposed a strong functional constraint on MC1R. Similarly, an increased risk of UV sensitivity, fair skin and red hair are detectable in certain genotypic variants of MC1R (Harding *et al.*, 2000). Other polymorphic loci within the genome have alleles at different frequencies in different populations through the effect of genetic drift.

1.4.3.5 Clines, Clusters and Sampling

As populations expand, gene flow between populations, selection and genetic drift will play pivotal roles in defining some of the characteristics of a population. Some of the resultant diversity will be observable phenotypically, such as skin colour. Other examples of selection include metabolic adaptation to diet, for example, lactase persistence is low in the Chinese population (Tishkoff *et al.*, 2007). In the

pastoralist Western world, it is comparatively high as the population has been more reliant on milk as an important part of their diet. This provides a very broad scale of differences between populations with varying gradients of change between whole countries and smaller populations or communities within a country (Barbujani *et al*., 1997).

Geography clearly plays a major role in human genetic variation, accounting for much of the genetic variance between populations (Manica, *et al.*, 2005), with genetic diversity increasing between populations over greater geographical space and within-population diversity decreasing the greater the over-ground distance from east Africa (Lawson Handley *et al.*, 2007). Similar findings were also reported of variation in genetic diversity of *Helicobacter pylori*, a bacterium over half of all humans carry which can cause peptic ulcers (Linz *et al.*, 2007). Most differences in gene frequencies are clinal, where they vary gradually between neighbouring populations – a premise backed up by the 'isolation by distance' (IBD) model: the idea that genetic similarities decrease with increased geographical separation, often in a linear manner (Cavalli-Sforza, *et al.*, 1994).

It is on this basis that the term 'race', usually encompassing the three categories 'Caucasian', 'Negroid' and 'Asian', is inaccurate when describing the genetics of a population. It is argued that to be able to categorise people into 'races', biologically, discrete clusters of human genetic variation would need to exist, with discontinuities in allele frequency data being diagnostic of such apportionment, and as such, a clinal gradation is a more likely explanation of diversity (Serre & Pääbo, 2004). Clines may be the result of the expansion of modern humans out of Africa, or the restricted levels of gene flow across regions (Novembre & Di Renzo, 2009). Conversely, clustering was observable in the study of Rosenberg *et al.*, (2002), where individuals were placed into six clusters, five corresponding to different continents.

Sampling was described by Serre and Pääbo, (2004), as the main reason why Rosenberg *et al.*, (2002), were able to visualise clustering of populations. If individuals were sampled over a large area based on geography, rather than sampling individuals on a, perhaps self-defined, population basis, this would highlight gradual changes rather than enhancing discontinuities in population sampling. For example, Figure 1.3 shows a coloured bar which represents continual allele frequency variation. By sampling a few select areas (coloured circles), data may falsely give the impression of clustering (black ellipses).

The authors also criticised the apparent ease with which a sample may alter its affiliation to a particular cluster in STRUCTURE based on the number of inferred populations. They cite the isolated Kalash population of the Pakistan clustering with major Eurasian populations until STRUCTURE is programmed to assume six populations are present. They state that this does not help understand the history of human populations when it is possible for a population such as the Kalash to be categorised in the same group as a major subdivision of human variation and puts the theory of population clustering into question.

Figure 1.3: The effects of sampling on clines and clusters (taken from Lawson Handley *et al.*, 2007)

A seemingly contradictory explanation to the above is that genetic variation expresses both signs of clinal gradation and clustering. In 2005, Rosenberg *et al*. responded to Serre and Pääbo, (2004), and expanded on their previous work of 2002 (by increasing the number of loci analysed nearly three-fold), which had already acknowledged that membership coefficients of samples were rarely

restricted to just one cluster and accepted gradations of diversity across regions. They argued that pockets of clustering could still be identified across a clinal scale, mainly due to geographical barriers, such as the Himalayas and oceans.

They demonstrated this by showing that the level of genetic diversity between populations within the same cluster increased as geographical separation increased; a finding consistent with a clinal scale of variation. Conversely, genetic diversity was greater between populations in different clusters, even though they were separated by the same geographical distance. These geographical barriers facilitate the discontinuities in allele frequency data that STRUCTURE then harvests to identify clusters.

Given that phenotypic traits may not always align with particular genetic traits, assigning individuals on the basis of race is near impossible, particularly when neither clines nor clusters can fully explain human diversity alone (Loring Brace *et al.*, 1993). Geography, on the other hand, recognises the effects of distance on diversity, whilst also acknowledging that discrete clusters are apparent along the way (Manica, *et al.*, 2005).

1.5 Population Assignment

Trying to distinguish between two people from the same or similar populations may be a precarious task due to how genetically related we are to each other (Brenner, 1998). Problems arise due to the diverse genetic composition of those who classify themselves as having a background that can present a lot of admixture, for example, a Hispanic population that is a mixture of Native Americans, Caucasians and Africans (Risch *et al.*, 2002).

1.5.1 Ancestry Informative Markers

Ancestry Informative Markers (AIMs) can aid in analysing population stratification and ancestry. AIMs are polymorphic loci where allele frequencies show distinct differences between populations (Jobling & Gill, 2004). This may help in not only determining the genetic ancestry of an individual but also in identifying the most likely geographical origin of an unknown donor sample. Clustering algorithms, such as those used by STRUCTURE, utilise the correlation patterns in the genetic differences between populations as discussed previously. Ideally, an allele will be fixed to one particular population (Rosenberg *et al.*, 2003), a "diagnostic" genotype, and this compared with other informative markers, will help build an estimate of most likely geographical origin.

Autosomal AIMs have an advantage over lineage studies of mtDNA or Y-chromosome DNA as they allow measuring of admixture caused by all of an individual's ancestors rather than just that of one family line (Halder *et al.*, 2008). mtDNA does not recombine with patrilineal DNA making it solely maternally inherited (Bender, *et al.*, 2000) and the Y chromosome is the largest non-recombining region of the human genome (Rebala & Szczerkowska, 2005) with 95 % not involved in any X-Y crossover (Skaletsky *et al.*, 2003). Excluding the pseudoautosomal regions, which are located at the terminals of the Y-chromosome,

there is no recombination with the X chromosome during meiosis (Gusmao *et al*., 1999; Iida & Kishi, 2005; Rebala & Szczerkowska, 2005), meaning mutation is pertinent to observing variation of the Y-chromosome. Despite this, the Y-chromosome does possess some useful forensic and population study qualities: in separating male-female STR mixtures and analysis of population structure, of particular interest is the association of paternal lineage and surnames (Jobling & Tyler-Smith, 1995; Jobling & Gill, 2004).

The restrictions in mtDNA diversity are based on similar principles that preclude greater variation of the Y-chromosome. One of its greatest advantages is its high copy number relative to nuclear DNA. The number of copies per cell varies according to the cell's requirements. For example, epithelial cells which are of particular use in forensic work, have approximately 5,000 mtDNA molecules per cell. Sperm cells have a few hundred and a single oocyte can contain 50,000 (Bender, *et al.*, 2000). This makes mtDNA useful in forensic casework even though it generally comprises less than 1 % of the total cellular DNA available. However, its high copy number makes it far easier to amplify, meaning it is a useful tool for genetic characterisation (Budowle *et al*., 1999) and the analysis of aged/degraded samples (Budowle *et al.*, 2003). This increased availability of mtDNA comes with a disadvantage: a lower power of discrimination (Divne & Allen, 2005; Liu, Y *et al*., 2006).

1.5.1.1 STR AIMs

Allele frequency differences within STRs, both inter and intra population, have been observed and have been used to assess the geographical origin of the sample (Bowcock *et al.*, 1994; Lowe *et al*., 2001; Rosenberg *et al*., 2003). Bowcock *et al.* (1994) studied 30 microsatellite markers (including D13S137 – used in Applied Biosystems' Identifiler® PCR amplification kit) in 148 individuals, comprising

samples from approximately 10 people from 14 indigenous populations. They reported that 87.8 % of the individuals sampled formed discrete clusters which coincided with their known geographical origin, albeit only continental specific. They admit that by choosing relatively isolated populations, the chance of successful population assignment was enhanced but it was an early indication as to the potential of geographical profiling.

Lowe *et al.* (2001) used calculations based on Bayes' Theorem to establish which ethnic group a sample belonged to using the six loci which made up the original Applied Biosystems' kit that saw the introduction of the NDNAD: SGM (HUMVWFA31/A [vWA], HUMTH01 [TH01], HUMFIBRA [FGA], D8S1179 [D8], D21S11 [D21] and D18S51 [D18]). The greatest genetic diversity was seen amongst Afro-Caribbean and Southeast Asian populations. The research used samples from five British ethnic groups which make up 99.7 % of the UK population: Caucasians, Afro-Caribbeans, Indian sub-continentals, Southeast Asians and Middle Easterners.

Studies such as Rosenberg *et al.*, (2003), which employ the HGDP Cell Line Panel use samples that were obtained from discrete populations such as Kalash, Surui and Melanesian. With European populations exhibiting the least among-population variation (0.7 %), this suggests that multiple databases would be needed to cover all regions of a particular country, particularly in Europe, to enable accurate classification of a sample.

Other studies include an attempt to infer ethnic origin using STR profiles. One study by Klintschar *et al*. (2003) used the Combined DNA Index System (CODIS) STR system which consists of 13 loci. Allele frequencies of samples were calculated and to test the usefulness of the CODIS loci in inferring ethnicity, they removed a sample from the population, recalculated the allele frequencies minus the single sample and then categorised it based on these new frequencies. Eight populations

were examined: Austrians; Egyptians; Hungarians from Budapest; Hungarian Romanies from Baranya County; and a further four populations, all resident in New York City - Caucasians, Afro-Americans, Asians and Hispanics. The two Hungarian populations provided the highest rate of correct identification, reaching almost 90 %. The lowest was Hispanics living in New York City with a 63.8 % success rate.

As expected, an increase in the number of examined loci led to an overall increase of success in the reliability of establishing ethnic origin. However, Klintschar *et al.* (2003) also noted that after analysing 18 STRs, the amount of information regarding group affiliation began to plateau showing that examining more STR loci does not necessarily mean more accurate data.

In the UK, it means that such databases may aid in distinguishing between UK (or white British) individuals and those with recent ancestral backgrounds from outside the country – a potentially useful tool for police intelligence.

1.5.1.2 SNP AIMs

As discussed previously, a greater number of SNPs are required to produce a comparable level of discriminatory power that an STR profile can provide. However, what gives autosomal STRs the advantage in human identity, their relatively high mutability, makes them less useful in retaining and revealing the history of ancestral genetics (Phillips *et al.*, 2007). SNPs, with their low mutation rate and increased stability, make them the preferred marker for measuring discontinuity between populations. A panel of SNPs with marked continental disparity will increase the chances of successful population assignment.

Some of the most useful SNPs will be those on genes that have exhibited a high degree of regional positive selection. This phenomenon, where resilience to major changes, for example, geographic barriers and rapid human population expansion,

generates genotypes to best cope with the changing demands of the environment (Voight *et al.*, 2006).

Rosenberg *et al.*, (2003), commented that dinucleotide microsatellites were the most informative of markers, being five to eight times that of SNPs. However, they note that there is a lack of correlation between the level of informativeness of SNPs in distinguishing between major continental groups and subpopulations. For that reason, the use of informative microsatellites or SNPs will depend on the level of resolution required.

Earlier studies on population assignment utilising SNPs attempted to assign samples based to one of three continental regions - European, African and Asian. One study looked at 211 SNPs mainly from pigmentation genes and 56 of those showed marked differences between samples from each of the three groups with 99 %, 98 % and 100 % accuracy in identifying an individual's descent from European, African and Asian ancestry, respectively (Frudakis *et al*., 2003). Pigmentation genes were targeted in this study due to the expected level of discontinuity between the three continental groups sampled. For example, the increased melanin levels in African individuals, compared with their European and Asian counterparts, will have been under greater selective pressure during the course of human evolution.

A study by Lao *et al*., (2006) reduced the number of SNPs required for continental identification to just ten. All were autosomal SNPs as the authors believed that this provided a more accurate estimation of ancestry rather than relying solely on mtDNA and Y-chromosome data as discussed previously. Using these 10 SNPs, they showed that they were able to identify almost the same level of population structuring in the HGDP samples as the 377 autosomal microsatellites used by Rosenberg *et al.*, (2002).

As Frudakis *et al.*, (2003), note, their work began to show the power of SNP analysis and its potential as a useful forensic tool. According to the authors, at that time, 70 % of samples in criminal cases are a mixture from two or more donors and trying to separate these if both have an unknown origin, when looking at potentially 56 SNPs, could prove complicated.

Studies such as those described rely on discrete differences of allele frequencies between populations. Depending on the level of resolution required, the required allele frequency difference (δ) typically ranges between 0.3 and 0.45 (Baye *et al.*, 2009; Halder *et al.*, 2008; Kidd *et al.*, 2012; Kosoy *et al.*, 2009). Similarly, studies look for AIMs which can separate continental populations and develop panels of SNPs of between 100 – 200 (Tian *et al.*, 2006).

Halder *et al.*, (2008), derived a panel of 176 autosomal SNPs which could apportion samples from four continental populations (European, West Africa, American and East Asian) with a high degree of accuracy. They comment that studies have suggested fewer SNPs can be used to identify population stratification, including the study of Lao *et al.*, (2006), but argue that such few markers are less informative increase the chance of error. However, Lao *et al.*, (2006), admit that their 10 SNPs work best on the population data supplied as part of the HGDP where the relationship between the populations is known. They agree that a more comprehensive panel of SNPs would be advantageous where less was known about the genetic ancestry of the samples.

Kosoy *et al.*, (2009), extended the development of SNPs AIMs by being one of the first to use the widely available TaqMan® SNP genotyping assay supplied by Applied Biosystems™. They developed a panel of 128 autosomal SNPs which looked for similar continental disparity of samples as that of Halder *et al.*, (2008). They too showed that reduced SNPs panels, as low as 24, could provide almost the same level of discrimination as the full 128 SNPs but that certain comparisons were

more affected than others. For example, the ability to distinguish between European and South Asian samples was greatly affected when less than 64 SNPs were used; though these two populations showed the lowest inter-population F_{ST} value (approximately 0.07).

Phillips *et al.*, (2007), devised a multiplex assay comprising 34 SNPs for continental separation of sub-Saharan Africans, European and East Asians using SNaPshot® technology developed by Applied Biosystems. They first used these three populations as a test of the informativeness of the 34 SNPs before using the HGDP samples to assign individuals without any prior knowledge of ancestry. As part of the 34 SNPs, three showed allelic-specificity to each of the three test populations and those three SNPs alone could apportion all 360 individuals in the test study to the correct population. When using all 34 SNPs on the HGDP samples, 1 % of samples were misclassified. This was accounted for by the portion of the HGDP samples originating from Sardinia: a population that showed affiliation to multiple population groups as defined by STRUCTURE analysis on a higher than average basis compared with other populations.

Recently, this 34-plex assay has been amended to remove one less informative SNP and replace it with one of near fixed East Asian origin (Fondevila *et al.*, 2012). This allowed for greater population assignment and clearer population stratification in STRUCTURE analyses than in the previous Phillips *et al.*, (2007), study.

Research is beginning to examine intra-country population stratification and the potential of SNPs in distinguishing samples from different regions. Qu *et al.*, (2012), showed, as studies before, that approximately 30 SNPs would be required for reasonable correct assignment of different populations, in this case, the northern and southern Han Chinese, separated by the Yangtze River. When 140 AIMs were used, the samples were unambiguously assigned to the correct population. A similar pattern has been seen between northern and southern Europeans (Bauchet

et al., 2007; Tian *et al.*, 2009) although to visualise population stratification over an area covering many diverse populations, many more SNPs required genotyping. Further population diversity was noted though when northern and southern Europe were analysed separately and additional AIMs were used, allowing for greater distinction of regional and ethnic groups along an east-west gradient.

1.5.2 Externally Visible Characteristics

AIMs may prove useful in police investigations where the person involved may not be known and any information which can narrow down the potential number of suspects will be welcomed. Although the studies discussed above did not explicitly set out to accomplish this, association and admixture mapping studies looking at genetic variation among and between populations will obviously have forensic application such as the MC1R test for red hair as discussed previously (Rees, 2000).

To complement the use of AIMs, Externally Visible Characteristics (EVCs) can further reduce the pool of individuals of interest (Kayser & Schneider, 2009). Eye, hair, skin colour and characteristics are all traits which have been investigated as part of genome-wide association studies, with eye colour being the most accurately predicted phenotype (Kayser & de Kniff, 2011) with over 90 % success of blue and brown eye colour distinction using 15 SNPs (Liu *et al.*, 2009).

Using over 6,000 Dutch European volunteers, this success rate was maintained when the most informative eye-colour SNPs were reduced to just six and developed into a single-plex assay: IrisPlex (Walsh *et al.*, 2011a). The system was later validated as a robust tool for use in forensic casework in accordance with the Scientific Working Group on DNA Analysis Methods (SWGDAM) guidelines (Walsh *et al.*, 2011b) and has now advanced to include accurate predictions of hair colour and shade; the new system known as HIrisplex (Walsh *et al.*, 2012a). IrisPlex was

then tested on seven populations across Europe (Norway, Estonia, UK, France, Spain, Italy and Greece) and showed a clear north-south divide when considering just blue and brown eye colour (Walsh *et al.*, 2012b).

Of course, eye colour is not finite and there are ranges of colours in between the two discussed here which may skew the ability to accurately predict phenotype. What it highlights is the potential of AIMs and EVC working in tandem to build the most accurate estimation of geographic origin. Ultimately, it comes down to getting the balance right between the need for information, cost and other evidence available. The use of AIMs and systems such as IrisPlex are not currently used routinely in forensic casework and further development is required before they become a viable option. Neither would work particularly well with mixed DNA samples and further work is required to assess their value when it comes to degraded DNA samples.

1.6 Project Background

The rationale for this study is to expand the role of forensic genetics – by helping to understand the differences between and within human populations and apply this knowledge to support police intelligence through routine DNA profiling. Specifically, this study will examine the effects on profile frequency estimations when considering populations with differing social and cultural aspects. In conjunction with the allele frequency data collected, the potential applications of identifying a geographical origin of a potential suspect or missing person will also be explored, as this is something that is not currently used as fully as it could be in the UK.

1.6.1 Population Structure of England

According to figures from the Government's National Statistics agency, in 2003, 49,138,831 people were living in England (Table 1.1). The following table summarises how most of this total is split between different ethnic backgrounds.

Table 1.1: Summary of main ethnic groups in England

Ethnicity	Number	Percentage (%)
White (British, Irish or 'other')	44,679,361	91.0
Asian/Asian British - Indian	1,028,546	2.1
Asian/Asian British - Pakistan	706,539	1.4
Bangladesh/'Other' Asian	513,204	1.0
All other classifications	2,211,181	4.5
TOTAL	**49,138,831**	**100**

In total, individuals classed as being of Asian descent, specifically from the Indian subcontinent, represent a total of 4.4 % of the UK population. Individuals with African ancestry, including those from the Caribbean, amount for 3 % of the population.

The distribution of non-white British people is unevenly spread across the country. For example, in the North West of England (Cheshire, Cumbria, Greater Manchester, Lancashire and Merseyside) the number of individuals classed as Pakistani is 116,968 which is 16.6 % of the country's 706,539 Pakistanis, or one in 58. Compare this with the North East and the proportion of the country's Pakistani population there is only 2 % (14,079) or one in 179. The distribution is also uneven within cities; statistics from London show that generally speaking, people from the same ethnic background form one or more clusters around the city in which they live. For example, boroughs around the edge of London are densely populated with white British people – at least 75 % in most areas. People from Pakistan form two main large clusters – one just North East of the centre and another on the North West. Indians have settled in almost exactly the same areas, but there are a greater

number of them. As expected, this shows that people of the same ethnic background tend to form close-knit communities within areas and mainly within the inner city, having little representation in the outer boroughs.

These data highlight that there may be regional differences in genetic variation and therefore generalised broad population databases may not always be the most suitable, or conservative, when it comes to profile frequency estimation. Importantly, knowledge of regional genetic variation and estimation of a profile frequency from a DNA sample would only be of use to police if someone from that region were sought for an offence. Although this is impossible to guarantee, knowledge of where the majority of a particular ethnic group reside may be of use to the police.

1.6.2 Consanguineous Marriage

Marriage between closely related individuals is commonplace in many parts of South Asia and the term ‘consanguineous’ is usually used to denote a union between second cousins or closer (Woodley, 2008). It is not unusual for first-cousins in countries such as Pakistan to marry (Hussain & Bittles, 1999). This is predominantly in keeping with long traditions of consanguinity within families as well as the desire to maintain close family ties. Other factors such as religion and culture may also influence these unions.

However, with consanguinity comes the increased risk of genetic-related problems. By reducing the size of the gene pool, the likelihood of offspring inheriting previously rare recessive alleles is increased and this can lead to both physical and mental problems, such as deafness, congenital heart disease and reduced cognitive performance (Saggar & Bittles, 2008; Ropers, 2008). Some disorders studied show little difference in frequency of occurrence in children born of consanguineous parents compared with those from non-consanguineous parents. For example, no significant differences were noted between the two groups in

incidences of Down Syndrome, Sickle cell disease and Type I diabetes, however, congenital heart disease was significantly higher (p = 0.01) in children with parents who were first cousins (El Mouzan *et al.*, 2010).

By having parents that are related, the chances are that the progeny will have a larger proportion of a homozygous genome. The more closely-related the parents, the greater this effect is likely to be. Genetically, the measure of consanguinity is known as the kinship or inbreeding coefficient and is denoted by F_{IS} which considers the relationship between the probability of obtaining matching alleles between the individual and their subpopulation (Overall & Nichols, 2001). In an inbred population, the chance of sharing alleles by descent is increased. At the lowest level of the definition of consanguinity (i.e. a marriage between completely unrelated second cousins), a child would be expected to have 1/64 of its genome homozygous. This would give an inbreeding coefficient of 0.0156 (Woods *et al.*, 2006; Saggar & Bittles, 2008; Woodley, 2008).

As first cousin marriages are prevalent in areas with this practice of consanguineous marriages, each individual will share 1/8 of their genome meaning a child is likely to be homozygous at 1/16 of its genome (F = 0.0625). This suggests a level of homozygosity at 6.25 % higher than the basal rate.

In a forensic context, any generalised database used against a population with this level of increased homozygosity could potentially report highly inaccurate results in terms of profile frequencies. Genetic drift is likely to be more extreme in populations where genetic diversity is limited, usually in small populations, and this would need to be considered when choosing a suitable database to compare a profile against. Alleles measured as rare in a UK (white) population may not be in a more consanguineous one from Pakistan, thus potentially working against any defendant from such a population by reporting a lower profile frequency.

1.7 Aims of the Project

The aim of this study is to evaluate the effect that subpopulations have on DNA profiling. Previous studies have demonstrated that it is relatively easy to differentiate between major continental groups and as the UK is home to many diverse populations, differentiating between members of the different population groups may be of forensic importance to aid police investigations.

In order to assess the consequence of substructuring within populations, the first stage of the study will focus on profiling samples from South Asian and UK (white) populations. The UK (white) population will be referred to as the 'UK' or 'UK population' from hereon in. In order to do this, samples were collected as follows: - 252 UK individuals; 575 samples from five Pakistani populations - Baluchi, Makrani, Punjabi, Pushtoon and Sindhi; 172 UK-based Gujarati and 120 Kalash from Pakistan.

Statistical analyses will be performed using the data obtained. The allele frequencies within each population will be used to calculate profile frequencies for all samples across all populations within the study and note the effects subpopulation corrections have on profile frequency estimation and population assignments of individual samples.

2 MATERIALS AND METHODS

2.1 Materials

2.1.1 Enzymes and Reagents

2.1.1.1 ReddyMix™ PCR Master Mix

This was supplied by ABgene, UK. It contained all the components necessary to carry out PCR, apart from primers and template DNA. This included: 1.25U *Taq* DNA polymerase, 1.5 mM $MgCl_2$ and 0.2 mM each of dATP, dGTP, dCTP and dTTP. The mix was provided at 1.1 X concentration.

2.1.1.2 Thermo-Start® PCR Master Mix

This was supplied by ABgene, UK. It contained all the components necessary to carry out PCR reactions apart from primers and template DNA. This included: 1.25U Thermo-Start DNA polymerase; 1.5 mM $MgCl_2$ and 0.2 mM each of dATP; dGTP; dCTP and dTTP. The mix is provided at a 1.1 X concentration and required an initial, one off pre-incubation period at 95 °C for 15 min for the polymerase to activate. This helps reduce non-specific priming and primer-dimer formation.

2.1.2 Commercial Kits

2.1.2.1 QIAamp® DNA Blood Mini Kit

This kit was used to extract DNA from buccal swabs obtained. It is supplied by Qiagen, UK and utilises a spin column method. The kit contains cell lysis buffer, DNA binding buffer, wash buffer and elution buffer.

2.1.2.2 Quant-iT™ PicoGreen®

Supplied by Invitrogen, UK, this kit provided an accurate method of double-stranded DNA quantification. The DNA concentration was calculated based on emitted fluorescence of fluorescein which produces levels of fluorescence proportional to the quantity of DNA.

2.1.2.3 AmpFℓSTR® SGM Plus® PCR Amplification Kit

This is a human identification kit and was used to analyse 10 STR loci plus the sex-identification locus, amelogenin. It was provided by Applied Biosystems, UK.

2.1.2.4 AmpFℓSTR® Identifiler® PCR Amplification Kit

This is a human identification kit was and used to analyse 15 STR loci plus the sex-identification locus, amelogenin. It was provided by Applied Biosystems, UK.

2.1.3 Swabs

2.1.3.1 Sterilin®

These sterile swabs were used to collect buccal cell samples from individuals. They were supplied by Copan, Italy.

2.2 Methods

2.2.1 Sterilisation

Any equipment, including glass and plastics, that was required to be sterile was autoclaved at 20 psi for 15 min. 1.5 ml and 0.2 ml microfuge tubes were sterilised by exposure to UV light (λ = 250 nm) for at least 20 min, with the caps off. Pre-packaged and sterilised filter tips were used for all sample extraction and PCR setup procedures.

2.2.2 Contamination and Working Areas

2.2.2.1 Stored Samples

Many of the samples that were analysed had been collected from Southern Asia (Pakistan) for previous studies and stored for a number of years. As these samples were irreplaceable, preventing contamination was imperative. With this in mind, each step of the experiment was carried out in a different laboratory or area within a laboratory. Initial DNA extractions were carried out in a flow hood (Aura mini, BioAir, Italy) in a pre-PCR lab to avoid extraneous contamination. The flow hood was wiped clean with 95 % (v/v) ethanol before and after the extraction process.

2.2.2.2 New Samples

Samples from the UK and Indian Gujarati populations were collected for this study. Although easier to re-sample if necessary, prevention of sample contamination was still essential to avoid data errors and the need to remove samples available for data analysis, i.e. should a mixed profile be obtained in one of the samples.

2.2.2.3 Laboratory Setup

To minimise the risk of contamination, each stage of the DNA profiling process was conducted in different areas of the laboratories. Extraction and PCR setup were performed in one laboratory and then run on a thermal cycler in a post-PCR laboratory. Any further analysis, such as gel electrophoresis or sequencing using the ABI 310 Genetic Analyzer, were also carried out in the post-PCR laboratory.

A one-way transfer of samples throughout the process was also employed to ensure amplified template DNA did not find its way back to the sample extraction area: PCR products make ideal templates for further amplification if allowed to contaminate extracted samples.

2.2.3 DNA Extraction

For it to be efficiently analysed, DNA must be removed from the cellular material and other components that often surrounds it. Several methods have been devised that lyse cells, break down proteins and remove extraneous matter to leave a pure yield of DNA including Chelex® extraction (Walsh *et al.,* 1991). Each method has various pros and cons, some of which will affect the quality of the yield but there is often the need for balance between the quality required and the cost and labour/equipment needs.

DNA extraction for the majority of the samples in this study was carried out using the QIAamp® DNA Blood Mini Kit, manufactured by QIAGEN (see sections 2.1.2.1 and 2.2.3.3). Although relatively expensive per sample, it provided consistent, high quality yields in little time.

2.2.3.1 Stored Samples

The samples collected previous to this study were buccal swabs which had been soaked in cell lysis solution (Puregene DNA Extraction Kit, Gentra, Flowgen, Novara Group Ltd., UK) and stored in 1.5 ml sterile tubes in a -20 °C freezer. For the extraction, all samples were vortexed and 200 µl of the cell lysis solution was removed and placed into a fresh, sterile 1.5 ml tube. A further 200 µl of lysis buffer (Buffer AL) from the QIAamp® DNA Blood Mini Kit was then added to the new tube.

2.2.3.2 New Samples

The samples collected from the UK and Indian populations were collected as buccal swabs and were stored in the refrigerator until extracted. To perform the extraction 400 µl of lysis buffer (Buffer AL) from the QIAamp® DNA Blood Mini Kit was added along with sodium dodecyl sulphate (SDS) solution to aid breakdown of the cell

walls. From this stage onwards, stored samples and new samples were treated in the same way (as described in section 2.2.3.3) to complete the DNA extraction.

2.2.3.3 Extraction of DNA using QIAamp® DNA Blood Mini Kit

This kit works on a spin-column based method and is a safer alternative to phenol-chloroform, which also provides high yields of pure DNA. With each sample already containing lysis buffer, 20 μl of Proteinase K (20 mg/ml) was then added to each sample to aid in the denaturing of proteins, as well as the inactivation of nucleases, which may otherwise interfere with the extraction of the DNA. All samples were then vortexed and incubated at 56 °C for 30 min. Then, 200 μl of 95 % (v/v) ethanol was added and the samples were vortexed again. The swab heads were subsequently removed and the contents of the tubes added to the spin columns supplied with the QIAamp® kit. The columns were centrifuged at 13,000 x g for 1 min and the flow-through discarded. Samples were then washed using 500 μl of Buffer AW1 and centrifuged for another min at 13,000 x g. The flow-through was discarded and a new 2 ml collection tube was attached to the spin column. The samples were then washed with 500 μl of a second buffer, Buffer AW2 and centrifuged for 3 min at 13,000 x g. Once complete, the flow-through was discarded and the spin columns were placed inside sterile 1.5 ml tubes with the caps removed. Finally, 70 μl of elution buffer (Buffer AE) was added to the centre of the column membrane and incubated at room temperature for one min before being centrifuged for one min at 13,000 x g. The DNA was transferred from the elute to a new 1.5 ml tube and sealed. The extracted DNA was stored at -20 °C. Samples obtained were suitable for both mtDNA and STR analysis.

The method for samples collected from other populations were similar but as some samples were originally collected for other studies, these were already stored in lysis buffer from the Gentra® Puregene® Blood Kit. There is compatibility between

this kit and the QIAamp® DNA Blood Mini Kit as both use an anionic surfactant to initiate cell lysis. Following cell lysis, the QIAamp® DNA Blood Mini Kit was used to complete the extraction for all samples.

2.2.4 DNA Quantification

2.2.4.1 Agarose Gels

Determination of quantity of extracted DNA is essential to prepare for further analysis as too much DNA can produce undesirable results. Agarose gel electrophoresis was used to determine this. The majority of quantification was carried out on 1.5 % (w/v) gels. The range used throughout this study was between 1.0 % (w/v) and 3.0 % (w/v) depending on the resolution required. The size of the fragments, and the size difference between them, if expecting two or more, can also have an effect on the concentration of gel required: small fragments and those which may only be a few base pairs apart usually appeared clearer at higher concentrations of agarose.

Prior to quantification a 50 X stock solution of TAE (2 M Tris, 1 M acetate, 100 mM EDTA and water) was prepared. A working 1 X solution was prepared as required. Agarose gels were made depending on the resolution required and this was most commonly at 1.5 % (w/v) so 0.6 g of agarose was added to 40 ml of 1 X TAE solution and heated in the microwave until all agarose had dissolved. It was then allowed to cool to around 55 °C and then poured into a gel casting tray.

From each sample, 5 μl was taken and 2 μl of 6 X Bromophenol blue loading buffer (2.5 % Ficoll 400, 11 mM EDTA, 3.3 mM Tris-HCl, 0.017 % SDS and 0.015 % Bromophenol Blue) (Sambrook, *et al.*, 1989) added to it. The samples were run alongside a *Hin*d III digest of lambda DNA (New England BioLabs Inc., USA) (this contains eight fragments of known length – 23,130 bp; 9,416 bp; 6,557 bp; 4,361

bp; 2,322 bp; 2,027 bp; 564 bp and 125 bp) for quantification at 10 V/cm for 20 min. This is used to approximate the mass of DNA in each sample: important for effective PCR reactions. Once the dye in the loading buffer had migrated approximately three-quarters of the way down the gel, the gel was removed and stained in an ethidium bromide solution (10 mg/ml in 500 ml of water) for 7 or 8 min. The gel was then rinsed in water and placed in the UVP GelDocIt™ (UK) and visualised under UV light (λ = 570-640 nm). A photograph was then printed for records. An example of the results is shown in Figure 2.1.

Lane 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17

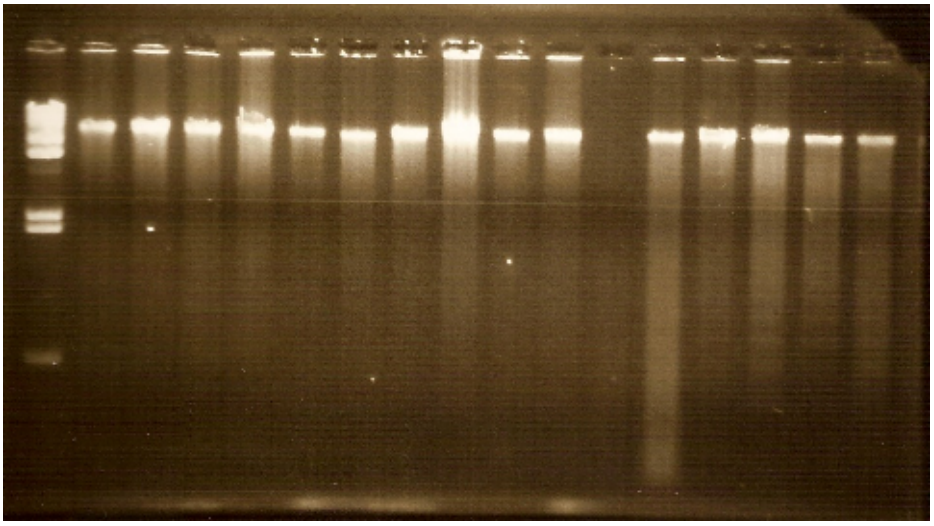


Figure 2.1: An example of a gel used to quantify some Pushtoon, Sindhi and Punjabi samples

Lane 1: Lambda *Hin*d III molecular weight marker (200 µl/ml) – 2 µl loaded

Lanes 2, 3 & 12: Pushtoon samples

Lanes 4 – 11: Sindhi samples

Lanes 13 – 17: Punjabi samples

All samples had 2 µl of sample added to 2 µl of 6 X loading buffer before loading onto gel. The samples were run on a 1.5 % (w/v) agarose gel at 5 V/cm for approximately 30 min using 1 X TAE buffer, then stained in ethidium bromide solution (50 µl [10 mg/ml] in 500 ml of water) for approximately 7 min before being viewed under UV light (λ = 570-640 nm).

2.2.4.2 PicoGreen® Quantification

When large batches of samples were prepared, agarose gel electrophoresis was not always the quickest method of assessing the quantity of DNA present. A more accurate method, though more expensive and time-consuming, was the Quant-iT™ PicoGreen® kit, supplied by Invitrogen, UK and used in tandem with a spectrofluorimeter. The kit was developed to quantify the amount of double-stranded DNA by measuring the level of fluorescence emitted by fluorescein which binds to double stranded DNA. There is negligible interference from extraneous components such as salts, detergents and proteins. The kit is ultra-sensitive: being able to detect DNA at concentrations as low as 25 pg/ml. Its main drawback is that it is not human-specific, but this was of little concern in this study considering the origin of the samples was known.

Two sets of standards prepared using control DNA to run alongside the sample: a high-range (1 ng/ml – 1 µg/ml) and low-range (25 pg/ml – 25 ng/ml) set of concentrations (Tables 2.1 & 2.2). These were altered depending on expected concentrations. As the samples collected for this study were expected to provide a high yield of DNA as all were buccal swabs, the two standard curves were adapted to produce one, broader range curve focussing more on the higher range (Table 2.3). The amount of reagent used was also reduced to save on resources. The PicoGreen® solution (containing the fluorescein) was also prepared by diluting the required amount in 20 X TE (200mM Tris-HCl, 20 mM EDTA) buffer.

With this method of quantification, time is important as the components are photosensitive. To ensure maximum accuracy, all tubes containing the reagents were kept in the dark prior to use and wrapped in tin foil during use to minimise light degradation. The reagents and standards were prepared according to manufacturer's

guidelines. Once prepared, they were placed in darkness in a refrigerator while the samples were prepared.

A standard 96-well plate can be used for the quantification. The samples were added to TE solution which is in each well. The first two rows of the plate were reserved for the standards. Once all samples and standards were ready, the final part to be added was the PicoGreen® solution. Light intervention needed to be minimised so this step was carried out quickly, without compromising accuracy.

The plates were analysed on the Tecan GENios Pro plate reader. The instrument shines light onto each sample at a wavelength which excites the fluorescein. This is then measured in relative fluorescence units (RFU) and recorded in an Excel spread sheet to allow for easy data analysis. If the standard curves are of sufficient quality, the samples can then be quantified quickly and accurately. An example of raw data obtained from PicoGreen® quantification is shown in Table 2.4.

Table 2.1: Recommended PicoGreen® setup for high-range curve

Volume of TE (µl)	Volume of 2 µg/ml control DNA (µl)	Volume of PicoGreen® reagent (µl)	Final DNA Concentration
0	1000	1000	1 µg/ml
900	100	1000	100 ng/ml
990	10	1000	10 ng/ml
999	1	1000	1 ng/ml
1000	0	1000	blank

Table 2.2: Recommended PicoGreen® setup for low-range curve

Volume of TE (µl)	Volume of 50 ng/ml control DNA (µl)	Volume of PicoGreen® reagent (µl)	Final DNA Concentration
0	1000	1000	25 ng/ml
900	100	1000	2.5 ng/ml
990	10	1000	250 pg/ml
999	1	1000	25 pg/ml
1000	0	1000	blank

Table 2.3: Adapted standard curve covering a broader range

Volume of TE (µl)	Volume of 2 µg/ml control DNA (µl)	Volume of PicoGreen® reagent (µl)	Final DNA Concentration (ng/ml)
75	75	150	500
93.75	56.25	150	375
112.5	37.5	150	250
120	30	150	200
127.5	22.5	150	150
135	15	150	100
138.75	11.25	150	75
142.5	7.5	150	50
146.25	3.75	150	25
148.5	1.5	150	10
149.25	0.75	150	5
150	0	150	blank

Table 2.4: An example of raw data after PicoGreen® quantification of Kalash samples

GENios Pro; Serial number: 12903500128; Firmware: V 2.30 01/04 GeniosPRO; XFLUOR4GENIOSPRO Version: V 4.53

Date:	13/10/07		
Time:	16:32		
Measurement mode:	Fluorescence Top		
Excitation wavelength:	485	nm	
Emission wavelength:	535	nm	
Gain (Optimal):	23		
Number of flashes:	10		
Lag time:	0	µs	
Integration time:	40	µs	
Mirror selection:	Dichroic 3 (e.g. FI)		
Plate definition file:	GRE96fb.pdf		

Rawdata (RFU) Temperature: 23.5 °C

<>	1	2	3	4	5	6	7	8	9	10	11	12
A	21370	17262	11854	9340	7394	5057	4159	3023	2673	1627	1765	3732
B	14506	3855	20341	18550	970	8670	20618	6075	17174	4161	6492	13374
C	17428	41214	6081	5862	7433	8379	5430	9121	14598	4520	21241	19684
D	33349	5140	5157	9403	6334	10497	11771	26152	15920	7488	23579	6353
E	4870	6655	5050	5069	5210	4267	6295	9918	12634	5402	10040	6363
F	2771	5419	4820	7058	3972	11302	6571	7025	5596	3605	6414	6596
G	22279	4514	5617	5614	3360	2893	8128	3708	3656	22071	4425	7436
H	2094	3500	2829	10605	22153	2883	2982	2790	2285	4255	1898	1265

The data are shown in relative fluorescence units (RFUs). Wells A1 – A9 contain the standards and are used calculate the quantity of DNA in each of the Kalash samples.

2.2.5 STR Analysis

2.2.5.1 STR Kits

Applied Biosystems have produced some of the most frequently used kits for forensic analysis. Two of their kits were used in this study: AmpFℓSTR® SGM Plus® and AmpFℓSTR® Identifiler® (used only on the Kalash population). The former analyses 10 loci plus the amelogenin locus, whilst the latter examines 15 loci plus amelogenin. The Identifiler® kit contains the same loci as the SGM Plus® kit in addition to D7S820, CSF1PO, D13S317, TPOX and D5S818.

2.2.5.2 Reduced Volume PCR

In an effort to save resources, the volumes of the kits used for PCR were reduced compared to the manufacturer's recommendations (Table 2.5). However, this makes the profiles more susceptible to peak imbalances, which may give false indications of a mixed profile, allelic drop-out or drop-in (Gaines *et al*., 2002). To validate this method, a selection of the samples were analysed in full-volume reactions. DNA concentration remained as recommended at 0.05 – 0.125 ng/µl, irrespective of the final volume being used. This amended procedure provided adequate profiles; in cases where a profile was off-scale, the sample was diluted and re-analysed. The thermal cycling procedure remained in accordance with the manufacturer's instructions regardless of the volume used (Table 2.6).

Table 2.5: Volumes used for PCR compared to manufacturer's instructions

Component	Recommended Volume (µl)	Adjusted Volume (µl)
PCR Reaction Mix	10.5	3
Primer Set	5.5	1.5
AmpliTaq Gold DNA Polymerase	0.5	1.0 (following a 1/5 dilution with dH_2O)
Water	0	0.5
Template DNA	10	1
TOTAL	**26.5**	**7**

Table 2.6: Thermal cycling parameters for SGM Plus® and Identifiler® kits

Stage	Temperature (ºC)	Time (min)	Number of cycles
Initial Denaturation	95	11	1
Denaturation	94	1	
Annealing	59	1	28
Extension	72	1	
Final Extension	60	45	1

2.2.5.3 Controls

To monitor and maintain the quality of results, positive and negative controls were used throughout the STR analysis procedure. A PCR positive is supplied with the STR kits and this was run alongside the samples and treated in exactly the same way before and during PCR. When analysed, the profile obtained was compared to the manufacturer's guidelines to monitor accuracy and reproducibility.

To monitor potential contamination, a negative control was introduced which contains deionised water in place of any sample, in the same reaction volumes. As with the positive, this went through the same PCR and electrophoresis procedures. If there was evidence of contamination, the samples amplified with the affected negative were examined for the possible contamination. Any samples in which contamination could not be ruled out were re-run.

2.2.5.4 Size Standards

Where the SGM Plus® kit (a four-dye system) was used, Applied Biosystems' 500-ROX™ (red dye) size standard was used. The five-dye Identifiler® kit used 500-LIZ™ (orange dye): both standards comprise 16 fragments of the same size (35, 50, 75, 100, 139, 150, 160, 200, 250, 300, 340, 350, 400, 450, 490, and 500 bp).

2.2.5.5 Allelic Ladders

As the Genetic Analyzer was often running over long periods of time, fluctuating external factors, e.g. temperature, can have an effect on allele designation when analysed with the size and allele-designating GeneMapper® ID software. To address this, an allelic ladder was run every six or seven samples to produce a composite ladder, thus taking into account any factors which may be having an effect on the run around that time. Applied Biosystem's Genotyper® software did not produce composite

ladders so in these cases, running a ladder with every few samples was essential to ensure allele designations were accurate or peaks were not labelled as off-ladder.

2.2.5.6 Sample Preparation

Prior to completion of PCR, a size standard and allelic ladder master mix was prepared. This consisted of 0.3 µl of 500-ROX™ and 10 µl of Hi-Di Formamide (Applied Biosystems).

This master mix was aliquoted into 0.2 ml PCR tubes and to these, 1 µl of PCR product was added, making a total volume of 11.3 µl available for electrophoresis. For the allelic ladder, 1 µl of PCR product is replaced with 1 µl of the relevant ladder. Once all samples had been prepared, they were all returned to the thermal cycler for 3 min at 95 °C to denature, followed by 3 min at 4 °C.

2.2.5.7 Sample Profiling

Samples from all four populations were genotyped using the Applied Biosystems' ABI PRISM® 310 Genetic Analyzer, following standard procedures except the run time which was increased from 24 min to 30 min to ensure a complete profile was obtained.

Prior to electrophoresis, the machine was calibrated to ensure accurate positioning of the autosampler, buffer was replaced and checks were made to ensure there was sufficient POP-4 (Performance Optimized Polymer-4) for the entire run. The module used, which determines the set of virtual filters, was GS STR POP 4 (1 ml) F.

2.2.5.8 Profile Analysis

Three software packages were used for the sizing of alleles and genotyping: Applied Biosystems' GeneScan® v3.1 for sizing alleles against the allelic ladder and internal size standard, in conjunction with Genotyper® v2.5.2 for allele calling and profile

comparisons; and the GeneMapper® *ID* v3.2 software which combines both former applications in one.

Following manufacturer's guidelines, samples were analysed and printed out as a record. This data was then inputted on the relevant database.

2.3 Statistical Analysis

Once a profile had been obtained, it was added to a database for that population. With over 600 samples genotyped across all four populations, statistical analysis was required in order to assess the population genetics, along with typical forensic parameters. Several different analyses were carried out, including tests for Hardy-Weinberg equilibrium and *F* statistics (Guo & Thompson, 1992).

A number of statistical packages, along with software add-ons in Microsoft® Excel were used to analyse the data.

2.3.1 Software for Statistical Analysis

2.3.1.1 PowerStats

This is a Microsoft® Excel add-on which provides a simple method for calculating typical forensic parameters (Tereba, 1999). It is available from the Promega Corporation, USA.

2.3.1.2 Arlequin

Arlequin v. 3.11 is a comprehensive statistical program used for analysing several areas relevant to population statistics such as Analysis of Molecular Variance (AMOVA), Wright's F-statistics, exact tests of Hardy-Weinberg equilibrium and population differentiation (Excoffier, *et al.*, 2005).

2.3.1.3 STRUCTURE

This programme employs a Bayesian probabilistic clustering approach to assigning individuals to a population based on genotypic data. It uses information from databases of populations (where the number of populations or clusters may, or may not be known) and attempts to group samples based on allele frequencies as it looks to maximise

Hardy-Weinberg equilibrium and look for population groupings exhibiting minimal linkage disequilibrium (Pritchard, *et al.*, 2000). This software package was utilised to assess whether any inference of population structure could be estimated based on STR profiles alone.

2.3.2 Data Analysis

2.3.2.1 Allele Frequencies

Allele frequencies were calculated by dividing the number of observations of a specific allele, by the total number of alleles within a population. This forms the basis for further calculations in population genetics and is shown for the four populations in Tables 4.1 to 4.4.

2.3.2.2 Typical Forensic Parameters

At the bottom of each of the allele frequency tables, there are further data showing typical forensic parameters. These give an indication as to the utility of each locus for identification purposes. The data includes: observed (H_o) and expected (H_e) heterozygosity; power of discrimination (PD) (Jones, 1972); probability of exclusion (PE) (Chakraborty, *et al.*, 1974); Hardy-Weinberg exact test (*p* value); polymorphic information content (PIC) (Botstein *et al*., 1980); probability of a match (MP) (Jones, 1972); and typical paternity index (TPI).

2.3.2.3 Minimum Allele Frequency

Every population studied showed alleles that appeared only once in that group. Although an allele frequency can be calculated, it can be inaccurate if that allele is under-represented in the population. To counteract this, a common approach is to use the formula 5 / 2*n*, where *n* is the number of individuals sampled in the population. This

means that every allele is counted at least five times, artificially increasing that allele frequency (National Research Council, 1996).

2.3.2.4 Hardy-Weinberg Tests

The Hardy-Weinberg principle states that, with random mating and no interference from factors including selection, migration, mutation, genetic drift and limited population size, genotype frequencies will remain constant from generation to generation (Stern, 1943). The genotype frequencies are calculated as follows:

Homozygote: A_1A_1: p^2 or q^2

Heterozygote: A_1A_2 or A_2A_1: $2pq$

In practice, the above criteria are impossible to achieve (for example, infinite population size) but if a large enough database exists and only a few generations are being examined, the effects should be negligible.

For each locus of a profile, the Hardy-Weinberg principle is employed to calculate the genotype frequencies. From the data, it is also possible to determine the number of heterozygotes observed (H_o) in the population. These genotype frequencies are then multiplied together, a process known as the 'product rule', and the result is the profile frequency.

The allele frequencies are used to calculate the expected heterozygosity (H_e). This calculation assumes no interference from any of the aforementioned criteria and in theory, should match the observed heterozygosity, though this is rarely the case.

2.3.2.5 Population Structure and *F* Statistics

As discussed in section 1.4.3.2, these data will provide inference regarding structuring between the populations studied and give an indication as to level of co-ancestry within a subpopulation.

2.3.2.6 Heterozygosity Test

If there is no significant difference between the observed and expected heterozygosity at a locus, the population can be deemed to be in Hardy-Weinberg equilibrium at that locus (Law *et al.*, 2003). Any deviation from Hardy-Weinberg equilibrium suggests that genotypic compositions of a population are not constant and something is acting on that population to cause changes in allele frequencies. In forensic terms, you may not have a true representative sample of the population if there is deviation away from the equilibrium. The effects of previous population substructuring may also affect sampling as well as more indirect factors such as genotyping errors.

The observed heterozygosity and majority of the forensic parameters were measured using PowerStats (Promega Corporation). The expected heterozygosity and Hardy Weinberg exact test were calculated using Arlequin v. 3.1 (Excoffier, *et al.*, 2005).

2.3.2.7 Exact Test

In the past, to determine if there was a significant difference between H_o and H_e, a chi-squared test was performed. The problem with this test is that it does not perform well with small numbers of alleles, so was inaccurate when considering rare alleles or alleles with low frequencies (Evett *et al*., 1996a; Guo & Thompson, 1992).

Fisher's exact test, developed by R.A Fisher, examines the significance of any non-random association between two variables; in this case, H_o and H_e. It works on data presented in a contingency table format and can be applied to allelic data. The method was later developed to include multiple alleles and Markov chain Monte Carlo (MCMC) sampling (Guo & Thompson, 1992).

The result of the test is a *p* value which measures any significant difference between H_o and H_e. Using the conventional 5 % significance level, a *p* value of less than 0.05

rejects the null hypothesis so that the difference seen between H_o and H_e is statistically significant.

2.3.2.8 Bonferroni Correction

The Bonferroni correction, developed by Carlo Emilio Bonferroni, adjusts the significance level of data based on the number of independent hypotheses being tested on a particular dataset. If $p < 0.05$, this suggests that a significant difference will be observed by chance every one in 20 tests on the same dataset – known as a Type I error (rejecting the null hypothesis when it is actually true).

In population genetics, each locus is counted as having an independent hypothesis. Therefore, using the SGM Plus® kit, with 10 loci, the new significance level would be 0.05 / 10 = 0.005. With the Identifiler® kit's 15 loci, this would equate to 0.003.

This method reduces the chance of a Type I error but must therefore increase the chance of Type II errors (accepting the null hypothesis when it is actually false) (Perneger, 1998) but is still used in the communication of population data and was applied to data in this study.

3 SAMPLE COLLECTION AND DNA EXTRACTION

3.1 Introduction

In order to assess the suitability of different databases when evaluating forensic evidence, samples must be collected from relevant populations. Databases need to be compiled to achieve reasonable representation across all populations being studied. Specifically, for STR databases, the NRC does not suggest a minimum number of contributors but recommends 'several hundred' (National Research Council, 1996). There are many publications containing over one thousand samples, conversely there are instances when smaller datasets of less than two hundred samples have been reported (Maruyama *et al*., 2008; Yong *et al*., 2007a; Yong *et al*., 2007b). The size of the dataset should take into consideration any statistical analyses that may be required. Certain analyses, such as the exact test (Guo & Thompson, 1992) were originally designed for use with small datasets, where loci had no more than 10 alleles. With the use of highly polymorphic loci with multiple alleles the exact test may not be the most appropriate method to use to assess departures from Hardy-Weinberg equilibrium. With datasets the size of those used in this study, the exact test would not be expected to detect small departures and any large deviations would need confirming in order to assess they are genuine and not caused by experimental factors such as the loci analysed or genotyping errors (Gill *et al.*, 2003).

In this study, it was not possible to collect over 1,000 samples for each of the populations analysed. This may affect the accuracy of some statistical analyses carried out on the data collected but there are practical limitations to consider when sampling, such as time and cost. Early work to assess how many samples would be required to provide a conservative estimate of allele and genotype frequencies based on VNTR analysis suggested 100-150 individuals would be adequate (Chakraborty, 1992). All populations studied herein meet or exceed that recommendation.

The databases compiled for this study will serve several purposes. Firstly, they allow for allele frequency data to be collected and analysed – assessing typical forensic parameters. Secondly, they will provide population-structuring information which will allow samples to be compared to non-cognate databases to assess whether substructuring within populations has a significant impact on the evaluation of forensic evidence. The data will also be used to assess whether the differences in allele frequencies enables inference as to which population a sample is likely to have originated from. Any patterns of homogeneity, within isolated populations in particular, would provide further evidence of genetic variation within and between populations and how this may affect forensic DNA profiling.

3.2 Rationale for Samples Collected

In England, people from the Indian subcontinent form the next largest proportion of the population (approximately 2.25 million or 4.5 %) after white, British people (National Statistics, 2003). In response to this, samples were collected from several different groups of individuals who were either living in areas that form part of the Indian subcontinent or those with ancestral links to the region (informed consent was obtained from all individuals):

- The Gujarat region of India – samples were taken from individuals living in Preston, UK who originate or have ancestral history to the Gujarat region.

- Pakistan – near the border with India and along the south coast. Samples were taken from caste groups including Punjab, Pushtoon, Sindhi, Makrani and Baluchi.

- The isolate Kalash in the Chitral region of the North West Frontier Province (now known as Khyber Pakhtunkhwa) of Pakistan.

- Samples were also taken from white, British students in Preston, England.

Table 3.1 shows the number of samples collected across all populations. Samples from Pakistan, including the Kalash were collected several years ago for use in a separate research study. The samples were stored frozen and then used for this study. The Gujarat and UK samples were collected specifically for this study. As well as analysing each population individually, the effect of combining databases from the Indian subcontinent was also investigated. The aim is to increase knowledge and awareness of how databases are applied in forensic cases and highlight the effects of the substructuring within populations. This is a factor which can greatly affect profile frequency calculations and must be taken into consideration along with the alleged

offender's ancestral background, in order to calculate a match probability which is both fair and avoids overstating the strength of the DNA evidence.

Table 3.1: Total number of samples collected across all populations

Population	Samples
Punjabi	200
Pushtoon	170
Sindhi	100
Baluchi	40
Makrani	65
Gujarati	172
Kalash	120
UK	252
TOTAL	**1119**

This represents all samples collected from previous studies and this study. Not all from each population were used as this was dependent on availability and quality of yield.

3.3 Geography of the Indian Subcontinent

The Indian subcontinent (Figure 3.1) is a term used to describe countries in South Asia from Pakistan on the west, to Nepal and Bhutan on the east, covering about 4,480,000 km² of the Asian continent. To the southwest is the Arabian Sea and to the southeast is the Bay of Bengal. South of Sri Lanka is the Indian Ocean. Figure 3.2 shows the approximate areas sampled were collected from.

Figure 3.1: Map of the countries that lie on the Indian subcontinent (http://www.worldatlas.com/webimage/countrys/asia/indiansub.jpg)

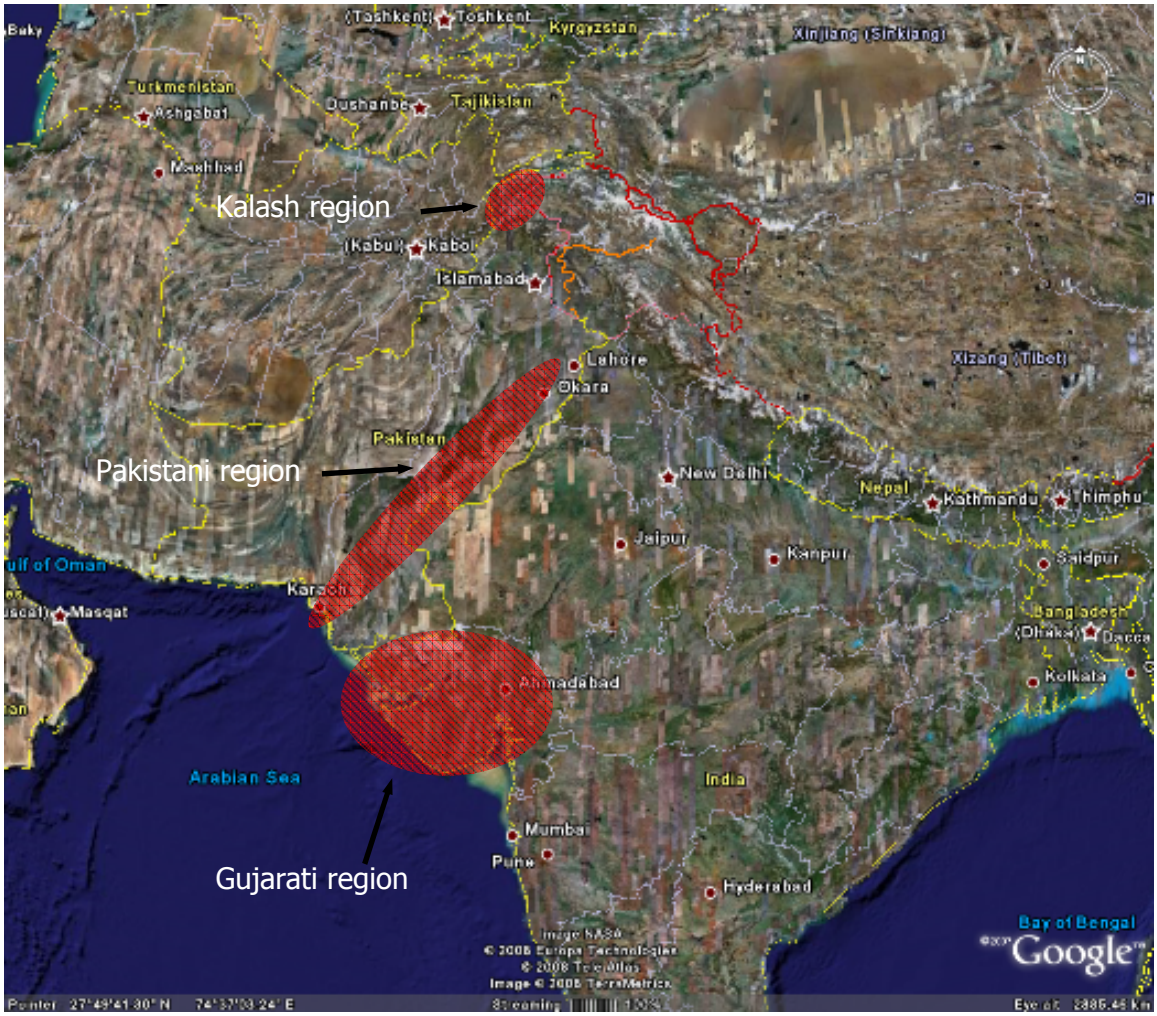


Figure 3.2: Map showing the regions of the Indian subcontinent from where populations used in this study originate (Google Earth)

3.4 Sampling of the UK Population

3.4.1 Location

All samples from the UK database were taken from students studying at the University of Central Lancashire in Preston, in the North West of England. While a large proportion of these students may be native to the North West, this was not a concerning factor prior to sample collection; participants could come from any part of the United Kingdom. Numerous studies have shown that many 'Caucasian' populations, including in the UK, do not show any discernible variation between observed groups in terms of STR allele frequencies and that substructuring within the white British population will be of negligible effect (Evett *et al.*, 1996b; Foreman, *et al.*, 1998 and Foreman & Lambert, 2000).

Further analyses of allele frequencies of UK-residing ethnic groups compared with their respective cognate populations also showed that the use of broad allele frequency databases was adequate and robust. In 97 % of cases assessed, match probabilities calculated using the broad database with F_{ST} = 2 % were the same or higher than when calculated against a cognate database: advantageous to a defendant (Foreman & Lambert, 2000). This is why the use of F_{ST} = 2 % or 3 % as a posterior correction to profile frequencies of the 'Caucasian' population to allow for inbreeding is seen as extremely generous (discussed later). In this respect the samples collected for this study, from the UK at least, should be representative of the population as a whole.

3.4.2 Sample Collection

Samples were taken from all individuals who understood and agreed to the informed consent which was obtained prior to sampling (Appendix I). A buccal swab was then taken and each individual was asked to best describe their ethnic background from a set list (Table 3.2) and apply the relevant code to their sample. This meant those who declared their ethnicity as 'white – British' could be used to build a UK database. The

only pre-requisite for participation in this database was that to the best of their knowledge, a participant's family and ancestors were from the UK. Those samples which could not be used for the UK database were stored for potential use in future databases.

Table 3.2: Codes used for sample donors to self-classify their ethnicity

Ethnic Origin	Code
White – British	11
White – Irish	12
White – Other	19
Black/Black British – Caribbean	21
Black/Black British – African	22
Black – Other	29
Asian/Asian British – Indian	31
Asian/Asian British – Pakistani	32
Asian/Asian British – Bangladeshi	33
Chinese	34
White and Black Caribbean	41
White and Black African	42
White and Asian	43
Other Mixed Background	49

3.5 Sampling of the Gujarat Population

3.5.1 Location

The Gujarati people live in the state of Gujarat which is in the North West of India and borders Pakistan.

3.5.2 Sample Collection

Preston is home to a large population who originate from Gujarat. Samples were collected by visiting homes in an area of Preston largely populated by people from the Gujarat region and permission sought and informed consent received from household members. Samples were taken from unrelated individuals. In a sampling strategy such as this, finding people in the same household who are unrelated may prove difficult, though there is also the possibility that people within the area are related to each other even though not residing at the same address. The alternative would be to collect samples from the Gujarat region itself but the Indian Government does not allow DNA samples to be taken out of the country.

3.6 Sampling of the Pakistani Population

3.6.1 Location

Samples were collected from five different populations in Pakistan. The Punjabi population, located over a vast area in the Punjab region to the North East of the country; the Pushtoons in the North West Frontier Province; Sindhis from the Sind Province and Makrani and Baluchi populations from towards the South coast – particularly within Karachi city.

3.6.2 Sample Collection

Buccal samples were obtained from individuals who had given consent (Makrani and Baluchi populations). Samples were collected mainly from staff and students at Army Medical Colleges except for the Makrani and Baluchi populations where requests were made through their chieftains to provide samples at collection centres in their neighbourhood.

3.7 Sampling of the Kalash Population

3.7.1 Location

The Kalash people are a small, close-knit tribal community in the North West Frontier Province of Pakistan. There are approximately 3,000 people and they speak Kalasha, an Indo-Aryan language which is a branch of the Indo-European class of languages. It is believed their origins are more strongly linked with Europe or the Middle East rather than Central or Eastern Asia as the Kalash people believe they descend from the Greek soldiers of Alexander the Great's army who invaded the North West of Pakistan in 327 – 323 BC (Firasat *et al.,* 2007). Qamar *et al.*, (2002), suggested Greek admixture of 20 – 40 % based on Y-chromosome analysis albeit there was no strong evidence of a Greek origin overall and the authors acknowledged genetic drift could be a contributing factor to the interpretation of the data. Subsequent studies have dismissed such a range for a Greek influence (Kivisild *et al.,* 2003) or significantly reduced any significance of Greek admixture and instead suggesting ancestral ties to Eurasian populations or Central Asia (Mansoor *et al.,* 2004). There is evidence to suggest they are practically genetically isolate: showing little affinity to neighbouring regions or groups and no association to East Asia, further supporting their claims of a Eurasian origin (Rosenberg *et al.,* 2002).

Prior to sample collection, permission was sought through the chieftains via Assistant Commissioner Chitral. Consent was then obtained from household seniors and individuals. Samples were not knowingly taken from individuals who had volunteered to form part of the 'HGDP-CEPH Human Genome Diversity Cell Line Panel' (Cann *et al.*, 2002).

3.7.2 Sample Collection

Buccal swabs were taken from several villages in the Bumboret Valley of the Kalash region. The use of buccal swabs provides a quick, inexpensive, non-intimate method of sample collection thus also more likely to increase donor participation.

3.8 Sample Storage and Handling

This study used samples collected via buccal swabs thus the quality of yield is expected to be relatively high. However, appropriate sample storage is imperative to maintain the integrity and quality of the sample. Buccal swabs taken from participants in the UK were kept frozen at -20 ºC until required. Samples which had been extracted were kept refrigerated for up to 24 hours before PCR. For longer-term storage, they were also frozen at -20 ºC. Samples from the Indian subcontinent were held in cell lysis solution from the Puregene® DNA Purification Kit. Data from Hadi (personal communication), show that there was no marked degradation of DNA held at 36 °C for five days when stored in lysis solution and so this was used as the storage method during the sample collection phase. Research by Graham *et al.*, (2008) also showed that DNA preservation of samples stored in buffer solutions at room temperature is achievable for up to 12 months. This DNA preservation method was adopted for this study but incorporated long-term freezing of samples instead of leaving them for excessive periods at room temperature.

3.9 Discussion

It was essential when gathering samples for a study of this kind that the sampling methods are representative of the population. Reliable estimation of profile frequencies relies upon fair sampling. The most accurate results would be obtained from sampling everyone within the population in question, but this is impractical, not least for the time and expense that would be incurred. If the data are not a true reflection of the population as a whole, any results obtained become questionable in their accuracy. In most studies of this kind, a 'convenience' approach has been employed – one which may be the most practical, yet there will be no reliable way of placing individual samples into distinct subpopulations (Foreman, *et al.*, 1997; National Research Council, 1996). This convenience approach to sampling should have little effect on studies utilising STRs of forensic relevance as the loci analysed are located in non-coding regions of the genome hence the effect of selection is negligible.

The number of samples taken in a group to deem it representative of a population is also important. The NRC recommends several hundred samples but the size of the population will have a clear impact on the proportionality of samples collected; for example, the size of the UK population compared with the Kalash. Again, given the loci analysed are not under selective pressure, this should have a negligible effect on the data. In studies such as this, a major limitation is always going to be the number of donors and how it would be useful to sample as many individuals as possible from each population of interest. As discussed by Foreman and Evett, (2001), a greater number of samples may seem beneficial but this would only serve to increase the precision of the database when estimating population parameters rather than strengthen its accuracy.

The 'clines vs. clusters' debate also affects how populations should be sampled (see section 1.4.3.5). In 2004, Serre and Pääbo argued that by sampling individuals on a self-defined ethnicity basis from within discrete population groups the chance of

observing clustering was increased because the sampling was effectively a snapshot of allele frequencies at certain points along a clinal scale. Instead, sampling should occur over large geographical areas to show the clinal relationship between distant populations. Rosenberg *et al.*, (2005), stated that even if sampling were even along a genetic cline, this approach would highlight clustering due to geographical barriers.

To gather samples from individuals of a particular ethnic group requires that you either know the history of the individual's family or you accept that what they tell you is an accurate and true description of their ethnic origin. They themselves may only know their family history going back a couple of generations. This has the ability to introduce anomalies into the data and highlights another reason why high sampling numbers are advantageous: it reduces the effect of any skew in data that may occur from a sample that has, knowingly or not, been misclassified into a population.

Recovery of DNA using buccal swabs is a method shown to produce high yields of DNA whilst maintaining the high-stability of buccal cells to allow for long-term storage, transportation, etc. From a group of 408 individuals, there was a 99 % success rate in successful amplification of six PCR amplicons ranging from 233 bp to 742 bp. In other studies, 100 % concordance in PCR amplification between blood and buccal cells has been reported (Richards *et al.,* 1993). This, coupled with the low PCR failure rate, buccal swabs are now routinely taken from subjects arrested for a recordable offence for inclusion on the NDNAD. Using a non-invasive method such as a buccal swab is also likely to increase the number of willing participants and also eliminate the need for a medical practitioner.

4 AUTOSOMAL STR ANALYSIS

4.1 Introduction

DNA profiling has advanced considerably since its introduction in 1985 when Professor Sir Alec Jeffreys demonstrated that the minisatellites (variable number tandem repeats - VNTR) he had been studying had the potential to produce a unique profile of an individual based on these highly polymorphic loci (Jeffreys *et al.*, 1985).

This was a costly, time-consuming and labour-intensive method of detecting highly-polymorphic loci. It was also not practical for many forensic applications as a relatively high amount (typically 100 ng) of good quality DNA was required; something not often available for analysis from a crime scene. Following the advent of PCR (Mullis *et al*., 1986), rapid development of DNA profiling techniques occurred with research into producing the most informative, cost-effective, PCR-based profiling kits. This also saw the utility of a new, highly-polymorphic genetic marker, the short tandem repeat (STR).

4.1.1 STR Amplification Kits

In 1999, Applied Biosystems™ released the AmpFℓSTR® SGM Plus® PCR amplification kit, utilising highly polymorphic STR regions and with greater discriminatory power than its predecessor, AmpFℓSTR® SGM. Samples were profiled with the new 10-locus, plus sex-determining, kit. A validation study confirmed that there was complete concordance when a selection of samples processed with the obsolete SGM kit were processed with the new SGM Plus® kit (Cotton *et al.,* 2000).

Applied Biosystems™ later developed the AmpFℓSTR® Identifiler® PCR amplification kit which added an additional five loci to those comprising the SGM Plus® kit and included all 13 CODIS (Combined DNA Indexing System) loci as used by the United States (Hoyle, 1998) plus D2S1338, D19S433 and the sex-determining amelogenin locus (Collins *et al.,* 2004).

4.1.1.1 Sex-determination

Sex is indicated by the amplification of two homologues of the amelogenin gene for the development of tooth enamel (Salido *et al.*, 1992), one on the X-chromosome and one on the Y-chromosome. Primers binding to intron 1 amplify a 106 bp region on the X-chromosome and a 112 bp region on the Y-chromosome therefore highlighting a 6 bp difference between the X- and Y-chromosomes allowing to differentiate between male and female donors were first described by Sullivan *et al.* (1993). The test is not conclusive though and mutations at primer-binding sites of deletions may suggest a male donor is actually female. This is particularly prevalent in males from the Indian subcontinent: one study reporting over 6 % of males from Kathmandu, Nepal with apparently female STR profiles (Cadenas *et al.*, 2007) and another reporting 0.23 % of males from India with Y-chromosome drop out (Kashyap *et al.*, 2006b). The differences seen here may be due to a common ancestor of the Kathmandu sample set and independent mutations in the Indian population. Regardless, sex indication of samples from the Indian subcontinent should be treated with caution and so this applies to the majority of samples used within this study. Due to the anonymity and destruction of 10 % of all samples following collection (see Appendix I) it is not possible to determine what proportion of each population sampled was male or female and therefore whether any of the male samples were affected by amelogenin drop out.

4.1.2 Profiling of Samples

The Indian (Preston Gujarati), Pakistani and UK populations were profiled using the AmpFℓSTR® SGM Plus® kit and the Kalash using the AmpFℓSTR® Identifiler® PCR kit to provide further allele frequency data for this isolate population. As the Identifiler® kit contained the same loci (and utilised the same primer sets) as the SGM Plus® kit, population comparisons were still possible with all four populations.

4.2 Statistical Analysis

4.2.1 Allele Frequencies

The allele frequencies in the different populations are listed in Tables 4.1 to 4.4. Some alleles were unique to particular populations. For example, alleles 5.3 and 7.3 of the TH01 were only apparent in the Kalash population and were not from the same individual. Allele 26.2 of the FGA locus was found only in the Indian population. Also, the UK population exhibited allele 36 on the FGA locus, not seen elsewhere. These were, however, all subjected to the minimum allele frequency correction for the purpose of match probability estimations. Nevertheless, it is these rare alleles which will add to the genetic variation between different populations and make profile frequencies more discriminating.

Although the Kalash population were profiled using the Identifiler® kit, only the loci contained within the SGM Plus® kit were used for comparative analyses with other populations. Furthermore, the data from the Indian and Pakistani populations have been published (Clark *et al.*, 2009).

4.2.2 Forensic Parameters

A range of parameters with typical forensic relevance were calculated using PowerStats (Tereba, 1999). A brief description of each metric follows.

4.2.2.1 Observed (H_o) and Expected (H_e) Heterozygosity

Represents the proportion of observed heterozygotes in the dataset (H_o) compared to the expected number of heterozygotes (H_e) based on the Hardy-Weinberg principle (see section 2.3.2.4).

4.2.2.2 Power of Discrimination (PD)

This is the probability that two individuals chosen at random will have different genotypes at the locus in question (Jones, 1972).

4.2.2.3 Probability of Exclusion (PE)

Primarily concerned with paternity cases, this is the probability that an individual; chosen at random will have a different genotype at a given locus when compared to another individual. It is also used as a method to exclude a male as being the biological father of a child (Chakraborty, *et al.*, 1974).

4.2.2.4 Polymorphism Information Content (PIC)

This shows the informativeness of a locus. When described by Botstein *et al*. (1980), it was developed to estimate the probability of correctly identifying a parental genotype based on the genotype of the offspring, particularly when considering rare alleles.

4.2.2.5 Match Probability (MP)

The match probability, often seen as providing the most essential information when comparing DNA profiles, is the probability that two randomly selected individuals will have matching genotypes at a given locus.

4.2.2.6 Typical Paternity Index (TPI)

The odds that an alleged father is the biological father assuming matching genotypes with a child at a given locus. Combining each locus will provide the combined paternity index.

4.2.2.7 Exact Test for Hardy-Weinberg equilibrium (*p*)

A measure of any significant differences between the observed and expected heterozygosity (see section 2.3.2.7).

Table 4.1: Allele frequencies of the Kalash population

The table on the following page shows the allele frequencies of the Kalash population profiled with the Identifiler® kit and typical forensic parameters at the end.

For Tables 4.1 to 4.4 the following abbreviations are used:

H_o = Observed heterozygosity

H_e = Expected heterozygosity

PD = Power of discrimination

PE = Power of exclusion

p = Exact test of Hardy Weinberg principle

PIC = Polymorphism information content

MP = Match probability

TPI = Typical paternity index

Where a *p* value was less than the conventional 0.05 (*), it was compared to the Bonferroni correction value (0.05 / 15 = 0.0033). Where the *p* value is higher than this corrected value, it is deemed there is no significant difference between H_o and H_e.

Allele	D8S1179 *n*=115	D21S11 *n*=115	D7S820 *n*=115	CSF1PO *n*=115	D3S1358 *n*=115	TH01 *n*=115	D13S317 *n*=115	D16S539 *n*=115	D2S1338 *n*=115	D19S433 *n*=115	vWA *n*=115	TPOX *n*=115	D18S51 *n*=114	D5S818 *n*=115	FGA *n*=115
5.3	-	-	-	-	-	0.0217	-	-	-	-	-	-	-	-	-
6	-	-	-	-	-	0.1783	-	-	-	-	-	-	-	-	-
7	-	-	0.0217	-	-	0.1043	-	-	-	-	-	-	-	-	-
7.3	-	-	-	-	-	0.0217	-	-	-	-	-	-	-	-	-
8	0.0217	-	0.1652	-	-	0.1217	0.0870	0.0478	-	-	-	0.3174	-	-	-
8.3	-	-	-	-	-	0.0217	-	-	-	-	-	-	-	-	-
9	-	-	0.0217	0.0217	-	0.3000	0.0217	0.1783	-	-	-	0.1522	-	0.0217	-
9.3	-	-	-	-	-	0.2783	-	-	-	-	-	-	-	-	-
10	0.2478	-	0.2565	0.3087	-	0.0217	0.0391	0.0696	-	-	-	0.0217	-	0.0217	-
10.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
11	0.1174	-	0.3957	0.2000	-	-	0.2826	0.2478	-	-	-	0.5261	-	0.0609	-
12	0.0217	-	0.1739	0.4478	-	-	0.4435	0.2870	-	0.0217	-	-	0.1491	0.4217	-
12.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
13	0.2783	-	-	0.0391	-	-	0.1043	0.1696	-	0.0348	0.0217	-	0.0482	0.4826	-
13.2	-	-	-	-	-	-	-	-	-	0.0478	-	-	-	-	-
14	0.1609	-	-	-	0.0217	-	0.0348	-	-	0.4913	0.0565	-	0.2719	0.0217	-
14.2	-	-	-	-	-	-	-	-	-	0.2087	-	-	-	-	-
15	0.0739	-	-	-	0.3522	-	-	-	-	0.1043	0.2435	-	0.1711	-	-
15.2	-	-	-	-	-	-	-	-	-	0.0217	-	-	-	-	-
16	0.1130	-	-	-	0.3565	-	-	-	0.0217	0.0652	0.1217	-	0.1447	-	-
16.2	-	-	-	-	-	-	-	-	-	0.0217	-	-	-	-	-
17	-	-	-	-	0.2217	-	-	-	0.0696	0.0217	0.2087	-	0.0307	-	-
17.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
18	-	-	-	-	0.0522	-	-	-	0.0739	-	0.1739	-	0.0833	-	0.0565
18.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
19	-	-	-	-	0.0217	-	-	-	0.1217	-	0.1652	-	0.0395	-	0.0217
20	-	0.0217	-	-	-	-	-	-	0.0870	-	0.0217	-	0.0217	-	0.1826
21	-	-	-	-	-	-	-	-	0.0217	-	0.0217	-	0.0217	-	0.1304
21.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
22	-	-	-	-	-	-	-	-	0.1000	-	-	-	0.0307	-	0.2957
22.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.0217

23	-	-	-	-	-	-	-	-	0.2435	-	-	-	0.0217	-	0.0739
23.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
24	-	-	-	-	-	-	-	-	0.1783	-	-	-	-	-	0.0696
24.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.1000
25	-	-	-	-	-	-	-	-	0.0696	-	-	-	-	-	0.0696
26	-	0.0217	-	-	-	-	-	-	0.0391	-	-	-	-	-	-
26.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
27	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
28	-	0.0217	-	-	-	-	-	-	-	-	-	-	-	-	-
29	-	0.3565	-	-	-	-	-	-	-	-	-	-	-	-	-
29.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
30	-	0.2174	-	-	-	-	-	-	-	-	-	-	-	-	-
30.2	-	0.0391	-	-	-	-	-	-	-	-	-	-	-	-	-
31	-	0.0217	-	-	-	-	-	-	-	-	-	-	-	-	-
31.2	-	0.0435	-	-	-	-	-	-	-	-	-	-	-	-	-
32	-	0.0870	-	-	-	-	-	-	-	-	-	-	-	-	-
32.2	-	0.0261	-	-	-	-	-	-	-	-	-	-	-	-	-
33.2	-	0.1565	-	-	-	-	-	-	-	-	-	-	-	-	-
34.2	-	0.0217	-	-	-	-	-	-	-	-	-	-	-	-	-
35.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
36	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
H_o	0.843	0.843	0.678	0.704	0.687	0.835	0.739	0.843	0.870	0.661	0.765	0.626	0.868	0.557	0.861
H_e	0.807	0.769	0.723	0.666	0.700	0.609	0.705	0.792	0.863	0.485	0.825	0.602	0.844	0.588	0.813
PD	0.920	0.925	0.868	0.811	0.851	0.897	0.866	0.915	0.961	0.869	0.940	0.761	0.948	0.745	0.947
PE	0.682	0.682	0.395	0.435	0.408	0.665	0.491	0.682	0.734	0.370	0.536	0.323	0.731	0.242	0.716
p	0.239	0.882	0.039*	0.257	0.305	0.409	0.818	0.672	0.404	0.360	0.252	0.325	0.549	0.213	0.620
PIC	0.780	0.760	0.670	0.600	0.640	0.897	0.660	0.760	0.840	0.660	0.800	0.530	0.820	0.500	0.820
MP	0.080	0.075	0.132	0.189	0.149	0.103	0.134	0.085	0.039	0.131	0.060	0.239	0.052	0.255	0.053
TPI	3.190	3.190	1.550	1.690	1.600	3.030	1.920	3.190	3.830	1.470	2.130	1.340	3.800	1.130	3.590

* Bonferroni Correction (0.05 / 15 = 0.0033)

Table 4.2: Allele frequencies of the Preston Gujarati population profiled with the SGM Plus® kit

Allele	D3S1385 *n*=172	vWA *n*=172	D16S539 *n*=172	D2S1338 *n*=172	D8S1179 *n*=172	D21S11 *n*=172	D18S51 *n*=172	D19S443 *n*=172	TH01 *n*=172	FGA *n*=172
6	-	-	-	-	-	-	-	-	0.2471	-
7	-	-	-	-	-	-	-	-	0.1890	-
7.3	-	-	-	-	-	-	-	-	-	-
8	-	-	0.0552	-	0.0145	-	-	-	0.1221	-
8.3	-	-	-	-	-	-	-	-	-	-
9	-	-	0.1483	-	0.0174	-	-	0.0145	0.2762	-
9.3	-	-	-	-	-	-	-	-	0.1599	-
10	-	-	0.0843	-	0.1686	-	0.0145	-	0.0145	-
10.2	-	-	-	-	-	-	-	-	-	-
11	-	-	0.3459	-	0.0756	-	0.0145	-	-	-
12	-	-	0.2035	-	0.1308	-	0.0930	0.1163	-	-
12.2	-	-	-	-	-	-	-	0.0145	-	-
13	0.0145	0.0145	0.1424	-	0.2500	-	0.1279	0.2907	-	-
13.2	-	-	-	-	-	-	-	0.0145	-	-
14	0.0785	0.1337	0.0174	-	0.1628	-	0.2820	0.2471	-	-
14.2	-	-	-	-	-	-	-	0.0436	-	-
15	0.2820	0.0581	0.0145	-	0.1308	-	0.1802	0.1366	-	-
15.2	-	-	-	-	-	-	-	0.0698	-	-
16	0.2733	0.2558	-	0.0145	0.0436	0.0145	0.1192	0.0349	-	-
16.2	-	-	-	-	-	-	-	0.0174	-	-
17	0.2529	0.2820	-	0.0465	0.0145	-	0.0872	0.0145	-	-
17.2	-	-	-	-	-	-	-	-	-	-
18	0.1076	0.1860	-	0.1919	-	-	0.0291	-	-	0.0145
18.2	-	-	-	-	-	-	-	0.0145	-	-
19	-	0.0698	-	0.1599	-	-	0.0291	-	-	0.1047
20	-	0.0145	-	0.1221	-	0.0145	0.0145	-	-	0.0988
21	-	0.0145	-	0.0291	-	-	0.0145	-	-	0.1076
21.2	-	-	-	-	-	-	-	-	-	0.0145
22	-	-	-	0.0959	-	-	-	-	-	0.1221

22.2	-	-	-	-	-	-	-	-	-	0.0233
23	-	-	-	0.1715	-	-	0.0145	-	-	0.1541
23.2	-	-	-	-	-	-	-	-	-	0.0145
24	-	-	-	0.0814	-	-	0.0145	-	-	0.2006
24.2	-	-	-	-	-	-	-	-	-	0.0145
25	-	-	-	0.0669	-	-	-	-	-	0.1047
26	-	-	-	0.0233	-	-	-	-	-	0.0523
26.2	-	-	-	-	-	-	-	-	-	0.0145
27	-	-	-	0.0145	-	0.0145	-	-	-	0.0145
28	-	-	-	-	-	0.1279	-	-	-	0.0145
29	-	-	-	-	-	0.2122	-	-	-	-
29.2	-	-	-	-	-	0.0145	-	-	-	-
30	-	-	-	-	-	0.1890	-	-	-	-
30.2	-	-	-	-	-	0.0262	-	-	-	-
31	-	-	-	-	-	0.0262	-	-	-	-
31.2	-	-	-	-	-	0.1308	-	-	-	-
32	-	-	-	-	-	0.0145	-	-	-	-
32.2	-	-	-	-	-	0.1599	-	-	-	-
33.2	-	-	-	-	-	0.0843	-	-	-	-
34.2	-	-	-	-	-	0.0145	-	-	-	-
35.2	-	-	-	-	-	0.0145	-	-	-	-
H_o	0.744	0.855	0.791	0.814	0.872	0.855	0.837	0.814	0.785	0.890
H_e	0.766	0.797	0.788	0.872	0.843	0.834	0.841	0.764	0.700	0.868
PD	0.905	0.908	0.921	0.967	0.952	0.957	0.954	0.933	0.915	0.963
PE	0.500	0.704	0.582	0.625	0.739	0.704	0.670	0.625	0.571	0.774
p	0.831	0.082	0.878	0.312	0.987	0.966	0.932	0.03484*	0.364	0.01327*
PIC	0.725	0.764	0.757	0.856	0.821	0.830	0.821	0.790	0.750	0.860
MP	0.095	0.092	0.079	0.033	0.048	0.043	0.046	0.067	0.085	0.037
TPI	1.950	3.440	2.390	2.690	3.910	3.440	3.070	2.690	2.320	4.530

* Bonferroni Correction (0.05 / 10 = 0.005)

Table 4.3: Allele frequencies of the combined Pakistani population profiled with the SGM Plus® kit

Allele	D3S1385 *n*=157	vWA *n*=157	D16S539 *n*=157	D2S1338 *n*=148	D8S1179 *n*=157	D21S11 *n*=157	D18S51 *n*=157	D19S443 *n*=157	TH01 *n*=157	FGA *n*=157
6	-	-	-	-	-	-	-	-	0.2387	-
7	-	-	-	-	-	-	-	-	0.2161	-
7.3	-	-	-	-	-	-	-	-	-	-
8	-	-	0.0581	-	0.0161	-	-	-	0.1516	-
8.3	-	-	-	-	-	-	-	-	0.0161	-
9	-	-	0.1581	-	0.0161	-	-	-	0.2161	-
9.3	-	-	-	-	-	-	-	-	0.1613	-
10	-	-	0.0839	-	0.1194	-	0.0161	-	0.0161	-
10.2	-	-	-	-	-	-	-	-	-	-
11	0.0161	-	0.3355	-	0.1097	-	0.0161	0.0161	-	-
12	0.0161	-	0.2452	-	0.0968	-	0.0903	0.0419	-	-
12.2	-	-	-	-	-	-	-	0.0161	-	-
13	0.0161	0.0161	0.1032	-	0.2032	-	0.1129	0.2548	-	-
13.2	-	-	-	-	-	-	-	0.0194	-	-
14	0.0677	0.1258	0.0129	-	0.1806	-	0.2323	0.2452	-	-
14.2	-	-	-	-	-	-	-	0.1097	-	-
15	0.2806	0.0968	-	-	0.1839	-	0.1677	0.1387	-	-
15.2	-	-	-	-	-	-	-	0.0839	-	-
16	0.3161	0.2742	-	0.0137	0.0742	-	0.1419	0.0323	-	-
16.2	-	-	-	-	-	-	-	0.0548	-	-
17	0.1774	0.2323	-	0.0753	0.0161	-	0.1065	0.0161	-	-
17.2	-	-	-	-	-	-	-	0.0161	-	-
18	0.1419	0.1903	-	0.1507	0.0161	-	0.0548	-	-	0.0161
18.2	-	-	-	-	-	-	-	-	-	-
19	0.0161	0.0710	0.0161	0.1781	-	-	0.0419	-	-	0.0452
20	-	0.0161	-	0.1370	-	-	0.0161	-	-	0.0742
21	-	-	-	0.0342	-	-	0.0161	-	-	0.1742
21.2	-	-	-	-	-	-	-	-	-	0.0161
22	-	-	-	0.0616	-	-	-	-	-	0.2032

22.2	-	-	-	-	-	-	-	-	-	0.0161
23	-	-	-	0.1849	-	-	0.0161	-	-	0.1323
23.2	-	-	-	-	-	-	-	-	-	-
24	-	-	-	0.0788	-	-	-	-	-	0.1516
24.2	-	-	-	-	-	-	-	-	-	0.0161
25	-	-	-	0.0719	-	-	-	-	-	0.1258
26	-	-	-	0.0137	-	-	-	-	-	0.0484
26.2	-	-	-	-	-	-	-	-	-	-
27	-	-	-	-	-	0.0161	-	-	-	0.0161
28	-	-	-	-	-	0.1516	-	-	-	0.0161
29	-	-	-	-	-	0.1903	-	-	-	-
29.2	-	-	-	-	-	0.0161	-	-	-	-
30	-	-	-	-	-	0.1581	-	-	-	-
30.2	-	-	-	-	-	0.0419	-	-	-	-
31	-	-	-	-	-	0.0387	-	-	-	-
31.2	-	-	-	-	-	0.1516	-	-	-	-
32	-	-	-	-	-	0.0161	-	-	-	-
32.2	-	-	-	-	-	0.1548	-	-	-	-
33.2	-	-	-	-	-	0.0935	-	-	-	-
34.2	-	-	-	-	-	-	-	-	-	-
35.2	-	-	-	-	-	0.0161	-	-	-	-
H_o	0.735	0.787	0.742	0.772	0.884	0.806	0.826	0.710	0.684	0.826
H_e	0.767	0.807	0.784	0.874	0.854	0.832	0.863	0.742	0.732	0.854
PD	0.908	0.933	0.922	0.968	0.955	0.960	0.963	0.949	0.930	0.963
PE	0.485	0.575	0.496	0.549	0.763	0.635	0.648	0.587	0.518	0.660
p	0.103	0.888	0.520	0.391	0.684	0.530	0.784	0.268	0.930	0.172
PIC	0.730	0.780	0.750	0.860	0.830	0.840	0.850	0.810	0.770	0.850
MP	0.092	0.067	0.078	0.032	0.045	0.040	0.037	0.051	0.070	0.037
TPI	1.890	2.350	1.940	2.200	4.310	2.770	2.870	2.420	2.040	2.980

Table 4.4: Allele frequencies of the UK population profiled with the SGM Plus® kit

Allele	D3S1385 *n*=252	vWA *n*=252	D16S539 *n*=252	D2S1338 *n*=252	D8S1179 *n*=252	D21S11 *n*=252	D18S51 *n*=252	D19S443 *n*=252	TH01 *n*=252	FGA *n*=252
6	-	-	-	-	-	-	-	-	0.2202	-
7	-	-	-	-	-	-	0.0099	-	0.2083	-
7.3	-	-	-	-	-	-	-	-	-	-
8	-	-	0.0218	-	0.0218	-	-	-	0.0893	-
8.3	-	-	-	-	-	-	-	-	-	-
9	-	-	0.1270	-	0.0119	-	0.0099	-	0.1270	-
9.3	-	-	-	-	-	-	-	-	0.3433	-
10	-	-	0.0575	-	0.1071	-	0.0218	-	0.0119	-
10.2	-	-	-	-	-	-	-	-	-	-
11	-	-	0.2798	-	0.0655	-	0.0139	0.0099	-	-
12	-	-	0.2917	-	0.1409	-	0.1528	0.0873	-	-
12.2	-	-	-	-	-	-	-	-	-	-
13	0.0099	-	0.1944	-	0.3552	-	0.1111	0.2321	-	-
13.2	-	-	-	-	-	-	-	0.0099	-	-
14	0.1528	0.0992	0.0278	-	0.2063	-	0.1825	0.3889	-	-
14.2	-	-	-	-	-	-	0.0099	0.0278	-	-
15	0.2540	0.0913	-	0.0099	0.0754	-	0.1409	0.1647	-	-
15.2	-	-	-	-	-	-	-	0.0218	-	-
16	0.2599	0.2302	-	0.0516	0.0159	-	0.1230	0.0496	-	-
16.2	-	-	-	-	-	-	-	0.0099	-	-
17	0.1806	0.2679	-	0.1885	-	-	0.0913	0.0099	-	0.0099
17.2	-	-	-	0.0099	-	-	-	-	-	-
18	0.1389	0.2242	-	0.0714	-	-	0.0694	-	-	0.0139
18.2	-	-	-	-	-	-	-	0.0099	-	-
19	0.0099	0.0655	-	0.1190	-	-	0.0496	-	-	0.0615
20	-	0.0179	-	0.1369	-	0.0099	0.0179	-	-	0.1766
21	-	0.0099	-	0.0258	-	-	0.0139	-	-	0.1964
21.2	-	-	-	-	-	-	-	-	-	-
22	-	-	-	0.0456	-	-	0.0099	-	-	0.1548

22.2	-	-	-	-	-	-	-	-	-	0.0099
23	-	-	-	0.1131	-	-	-	-	-	0.1329
23.2	-	-	-	-	-	-	-	-	-	0.0099
24	-	-	-	0.1032	-	-	-	-	-	0.1151
24.2	-	-	-	-	-	0.0099	-	-	-	-
25	-	-	-	0.1190	-	-	0.0099	-	-	0.0933
26	-	-	-	0.0218	-	0.0099	-	-	-	0.0298
26.2	-	-	-	-	-	-	-	-	-	-
27	-	-	-	-	-	0.0238	-	-	-	0.0099
28	-	-	-	-	-	0.1548	-	-	-	-
29	-	-	-	-	-	0.2143	-	-	-	-
29.2	-	-	-	-	-	0.0099	-	-	-	-
30	-	-	-	-	-	0.2520	-	-	-	-
30.2	-	-	-	-	-	0.0218	-	-	-	-
31	-	-	-	-	-	0.0833	-	-	-	-
31.2	-	-	-	-	-	0.0972	-	-	-	-
32	-	-	-	-	-	0.0198	-	-	-	-
32.2	-	-	-	-	-	0.0774	-	-	-	-
33	-	-	-	-	-	0.0099	-	-	-	-
33.2	-	-	-	-	-	0.0337	-	-	-	-
34.2	-	-	-	-	-	0.0099	-	-	-	-
35	-	-	-	-	-	0.0099	-	-	-	-
35.2	-	-	-	-	-	-	-	-	-	-
H_o	0.806	0.837	0.750	0.817	0.786	0.782	0.794	0.762	0.714	0.877
H_e	0.794	0.804	0.780	0.884	0.791	0.812	0.880	0.724	0.680	0.859
PD	0.921	0.927	0.914	0.972	0.927	0.957	0.971	0.903	0.904	0.961
PE	0.609	0.670	0.510	0.632	0.573	0.647	0.587	0.587	0.573	0.757
p	0.938	0.786	0.269	0.038*	0.097	0.770	0.006*	0.148	0.026*	0.491
PIC	0.760	0.770	0.740	0.870	0.760	0.820	0.870	0.720	0.730	0.850
MP	0.079	0.073	0.086	0.028	0.073	0.043	0.029	0.097	0.096	0.039
TPI	2.570	3.070	2.000	2.740	2.330	2.860	2.420	2.420	2.330	4.200

* Bonferroni Correction (0.05 / 10 = 0.005)

Graph 4.1: Allele frequency distribution at the D3 locus across all populations

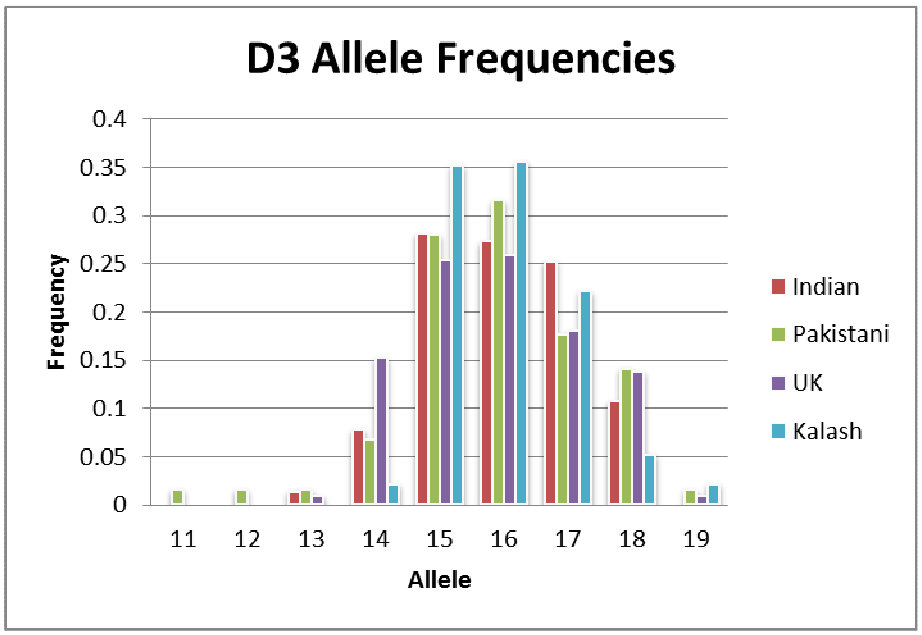
Frequencies of the Indian, Pakistani and UK populations appear fairly uniform with the exception of allele 14 where the UK population shows approximately double the frequency. The Kalash have a higher count of alleles 15 and 16 in particular with a considerably smaller frequency of allele 14 than the UK.

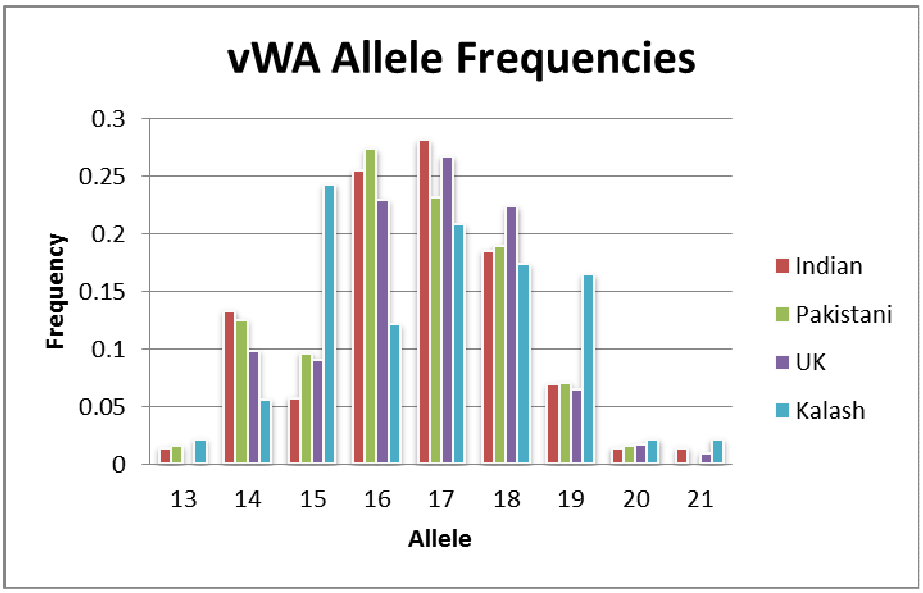
Graph 4.2: Allele frequency distribution at the vWA locus across all populations

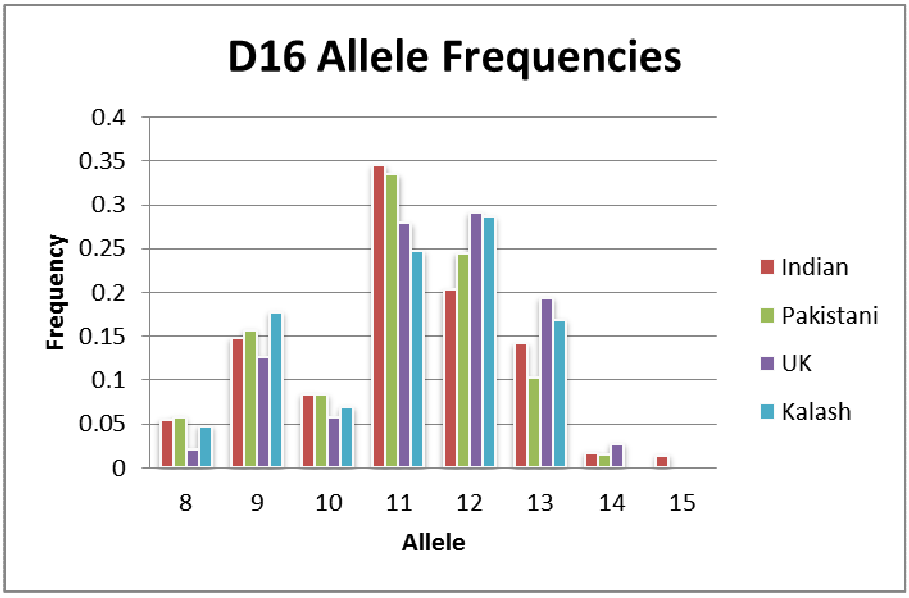
The Kalash show considerable increase in frequency at allele 15 and 19 compared with the other populations with approximately 24 % and 17 % respectively of the population carrying these alleles compared with less than 10 % for both alleles across the remaining populations.

Graph 4.3: Allele frequency distribution at the D16 locus across all populations

The Kalash show no incidence of alleles at the higher molecular weight, with only one sample from the Indian population comprising allele 15.


D3 Allele Frequencies
Frequency
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
11 12 13 14 15 16 17 18 19
Allele
Indian
Pakistani
UK
Kalash


vWA Allele Frequencies
Frequency
0.3
0.25
0.2
0.15
0.1
0.05
0
13 14 15 16 17 18 19 20 21
Allele
Indian
Pakistani
UK
Kalash


D16 Allele Frequencies
Frequency
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
8 9 10 11 12 13 14 15
Allele
Indian
Pakistani
UK
Kalash

Graph 4.4: Allele frequency distribution at the D2 locus across all populations

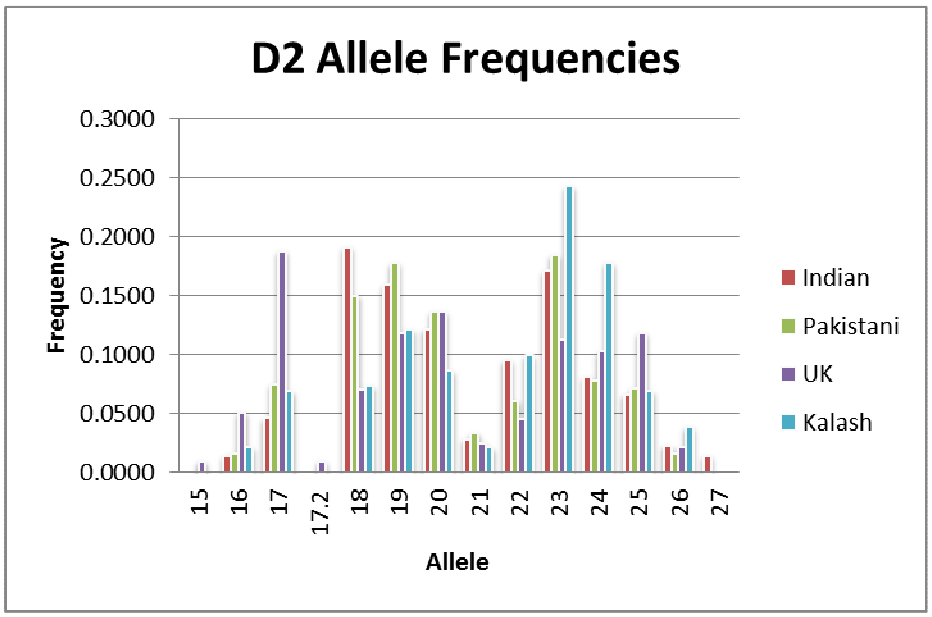
D2 shows an uneven distribution of alleles across the four populations, particularly at allele 17 where it is more prevalent in the UK population, with allele 18 being approximately twice as common in the Indian and Pakistani populations. The Kalash show a clear rise in frequency at alleles 23 and 24, with an increase of over 100 % compared with the other Indian subcontinent populations.

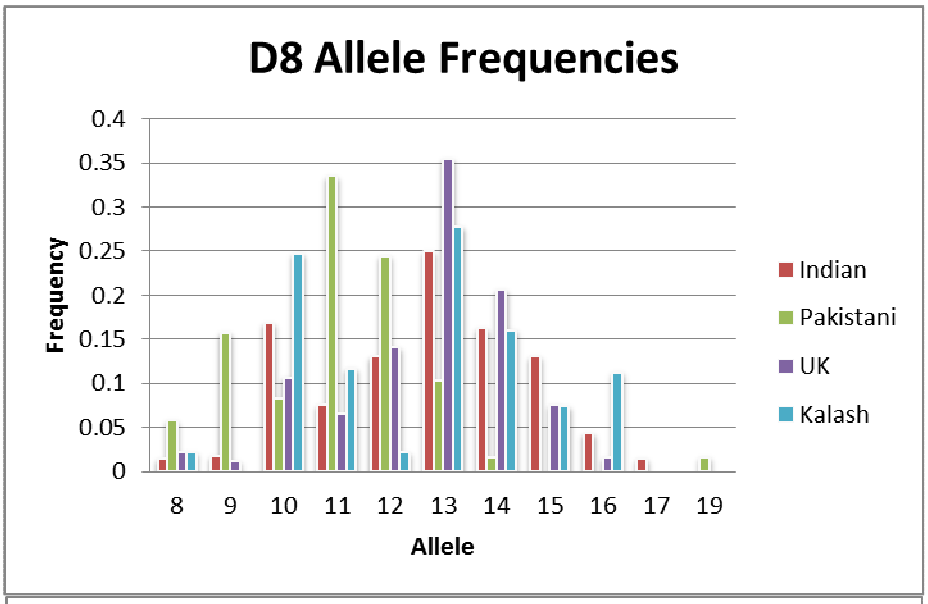
Graph 4.5: Allele frequency distribution at the D8 locus across all populations

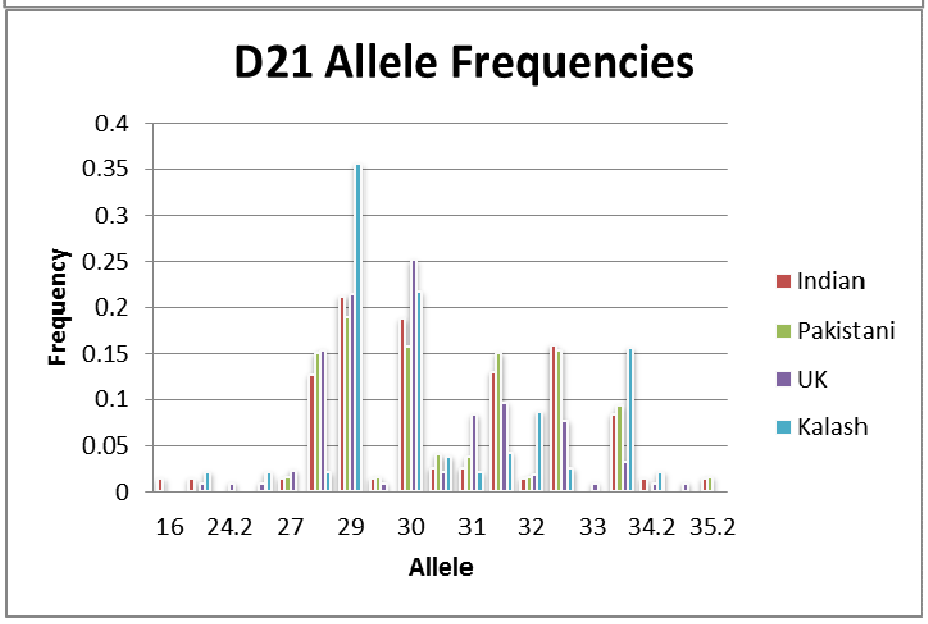
At D8, the Pakistani population appears to dominate the lower molecular weight allele frequencies, with the remaining three populations relatively balanced around allele 13 where the Pakistani frequency is decreased; clear distinction in the Kalash population at alleles 10 and 16.

Graph 4.6: Allele frequency distribution at the D21 locus across all populations

Allele frequencies at D21 appear relatively balanced across all populations apart from the Kalash where there alleles 29, 32 and 33.2 are more common. The UK population stands out at allele 31 with the two Indian subcontinent populations in close proximity across the most common alleles.


D2 Allele Frequencies
0.3000
0.2500
0.2000
0.1500
0.1000
0.0500
0.0000
Frequency
15
16
17
17.2
18
19
20
21
22
23
24
25
26
27
Allele
Indian
Pakistani
UK
Kalash


D8 Allele Frequencies
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Frequency
8
9
10
11
12
13
14
15
16
17
19
Allele
Indian
Pakistani
UK
Kalash

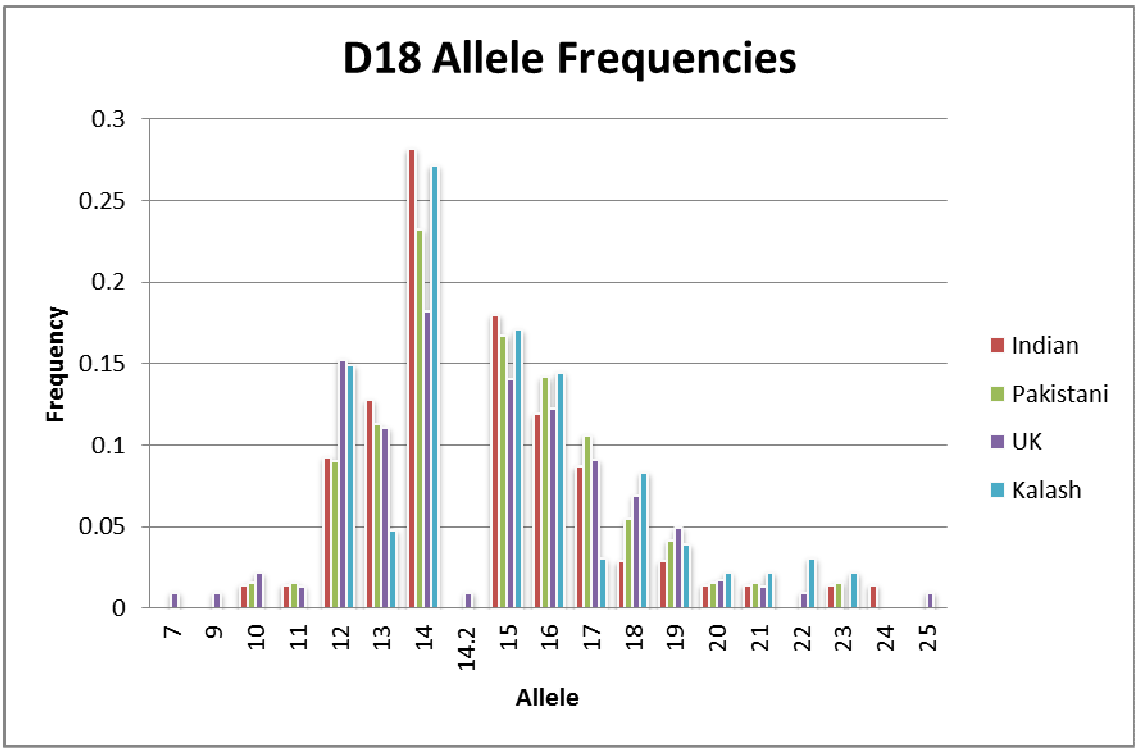

D21 Allele Frequencies
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Frequency
16
24.2
27
29
30
31
32
33
34.2
35.2
Allele
Indian
Pakistani
UK
Kalash

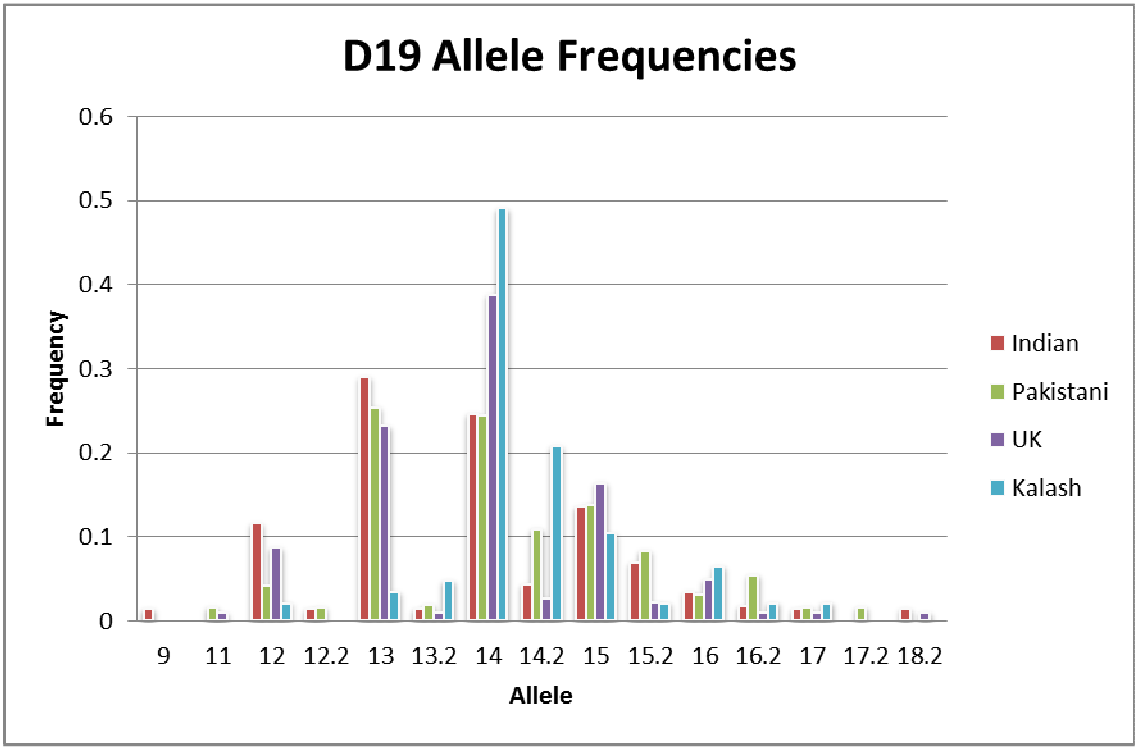
Graph 4.7: Allele frequency distribution at the D18 locus across all populations

Alleles 7, 9, 14.2 and 25 are represented within the UK population only at D18, though all at low frequencies. Similar patterns visible looking at alleles 13 and 17 with the Kalash having a much lower frequency here than the other three populations. Also, the UK population is represented at all alleles profiled at D18 apart from 23 and 24.

Graph 4.8: Allele frequency distribution at the D19 locus across all populations

Allele frequencies at allele 13 show a considerable difference between the Kalash population when compared to the other three. However, at alleles 14 and 14.2 the Kalash shows the greatest frequency of occurrence.


D18 Allele Frequencies
0.3
0.25
0.2
0.15
0.1
0.05
0
Frequency
7
9
10
11
12
13
14
14.2
15
16
17
18
19
20
21
22
23
24
25
Allele
Indian
Pakistani
UK
Kalash

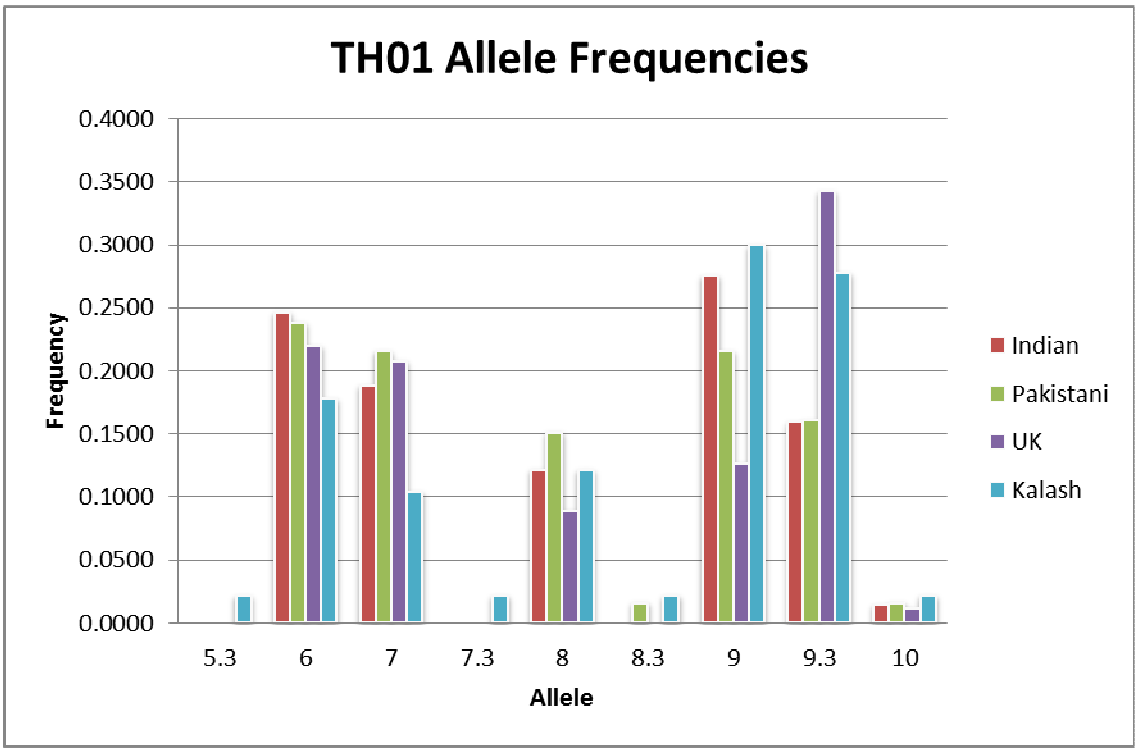

D19 Allele Frequencies
0.6
0.5
0.4
0.3
0.2
0.1
0
Frequency
9
11
12
12.2
13
13.2
14
14.2
15
15.2
16
16.2
17
17.2
18.2
Allele
Indian
Pakistani
UK
Kalash

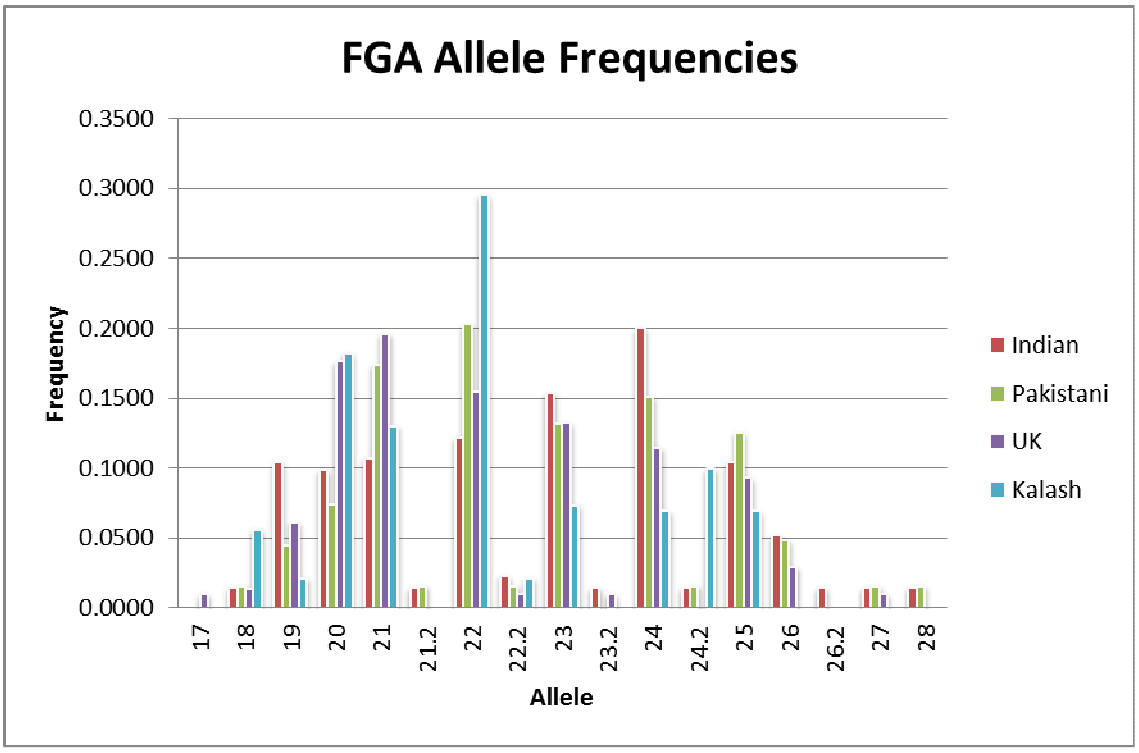
Graph 4.9: Allele frequency distribution at the TH01 locus across all populations

Although low in frequency, only samples from the Kalash carry alleles 5.3 and 7.3. At allele 7, there is a similar distribution across all population apart from the Kalash which is approximately 50 % lower. The UK and Kalash populations have the highest incidence of allele 9.3 at approximately 30 %, compared with the Indian and Pakistani populations, comprising just over 15 % of individuals with allele 9.3.

Graph 4.10: Allele frequency distribution at the FGA locus across all populations

Clear differences between populations at some alleles associated with FGA: the Kalash predominantly at allele 22 as well as 24.2, which had no occurrence in the UK population. The UK samples do not tend to mirror any of the other populations in terms of allele frequencies. For example, at allele 20, the UK population has a similar prevalence to the Kalash but at allele 21, it is more similar with the Pakistani population.


TH01 Allele Frequencies
0.4000
0.3500
0.3000
0.2500
0.2000
0.1500
0.1000
0.0500
0.0000
Frequency
5.3
6
7
7.3
8
8.3
9
9.3
10
Allele
Indian
Pakistani
UK
Kalash


FGA Allele Frequencies
0.3500
0.3000
0.2500
0.2000
0.1500
0.1000
0.0500
0.0000
Frequency
17
18
19
20
21
21.2
22
22.2
23
23.2
24
24.2
25
26
26.2
27
28
Allele
Indian
Pakistani
UK
Kalash

4.2.3 Electropherograms

Samples were genotyped and analysed as per sections 2.2.5.7 and 2.2.5.8. Examples of electropherograms produced following electrophoresis on the Applied Biosystems'™ ABI PRISM® 310 Genetic Analyzer are shown in Figures 4.1 and 4.2.

Figure 4.1 shows a sample from the Sindh population having been amplified using the AmpFℓSTR® SGM Plus® PCR amplification kit. It shows a male STR profile, heterozygous at all loci except D3. The number directly below each peak (not including amelogenin) indicates the number of repeat units at that specific allele of a locus. The number below that is the size of the amplicon in bp. This is determined by direct comparison to the size standard, in this case 500-ROX™ shown as red peaks at the bottom of the electropherogram. A DNA profile is typically reported as the combination of the number of repeat units at all loci, plus amelogenin.

Figure 4.2 shows a sample from the Kalash population having been amplified using the AmpFℓSTR® Identifiler® PCR amplification kit. Although all loci of the AmpFℓSTR® SGM Plus® PCR amplification kit are included in this kit, they are shown in a different order to allow for the clearest resolution of each locus to ensure, for example, minimal spectral overlap. If the same sample were amplified using both kits, the corresponding loci would report the same number of repeat units, regardless of locus order.

Figure 4.1: Electropherogram of a sample from the Sindh population after amplification with the AmpFℓSTR® SGM Plus® PCR amplification kit

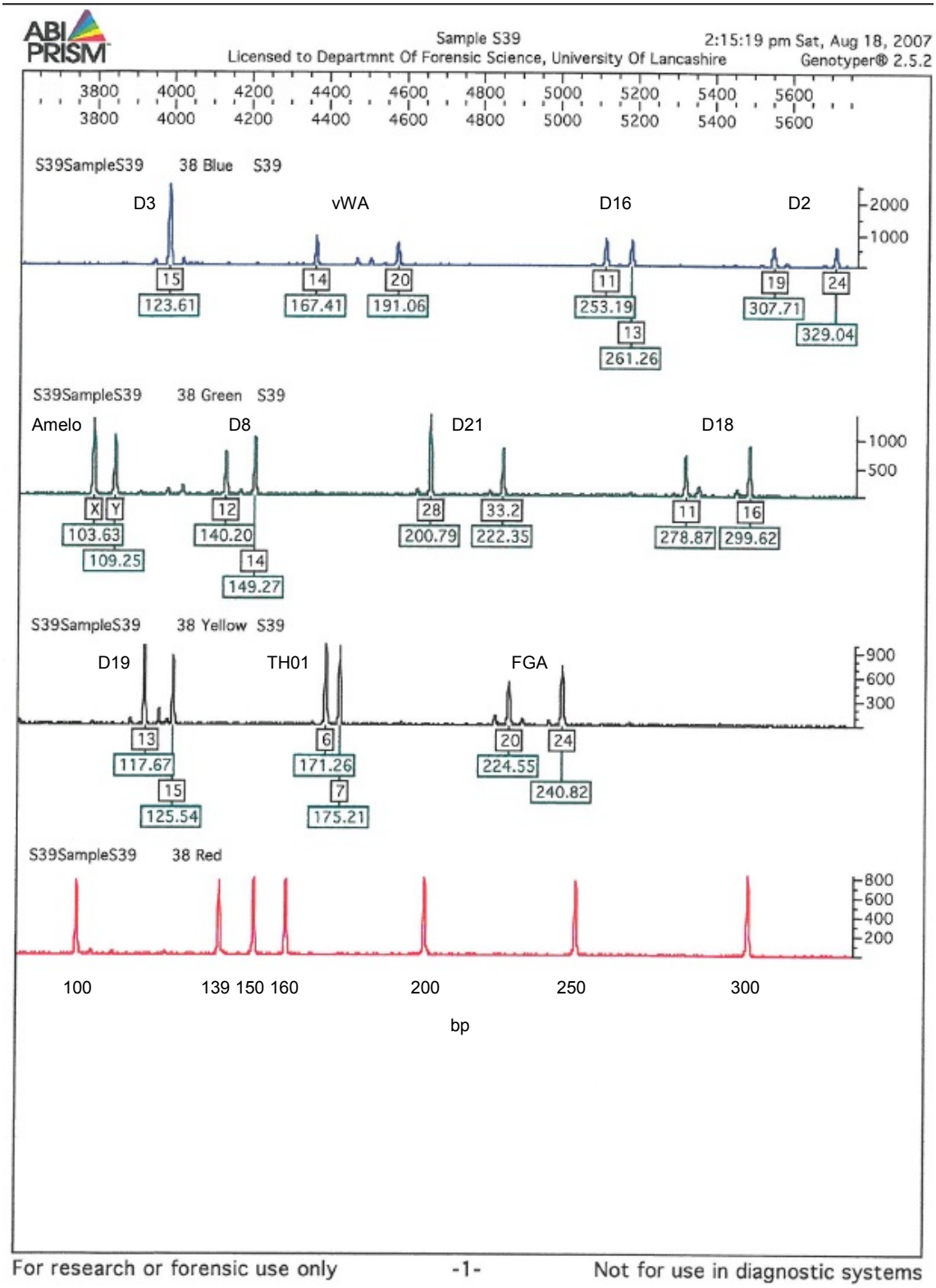
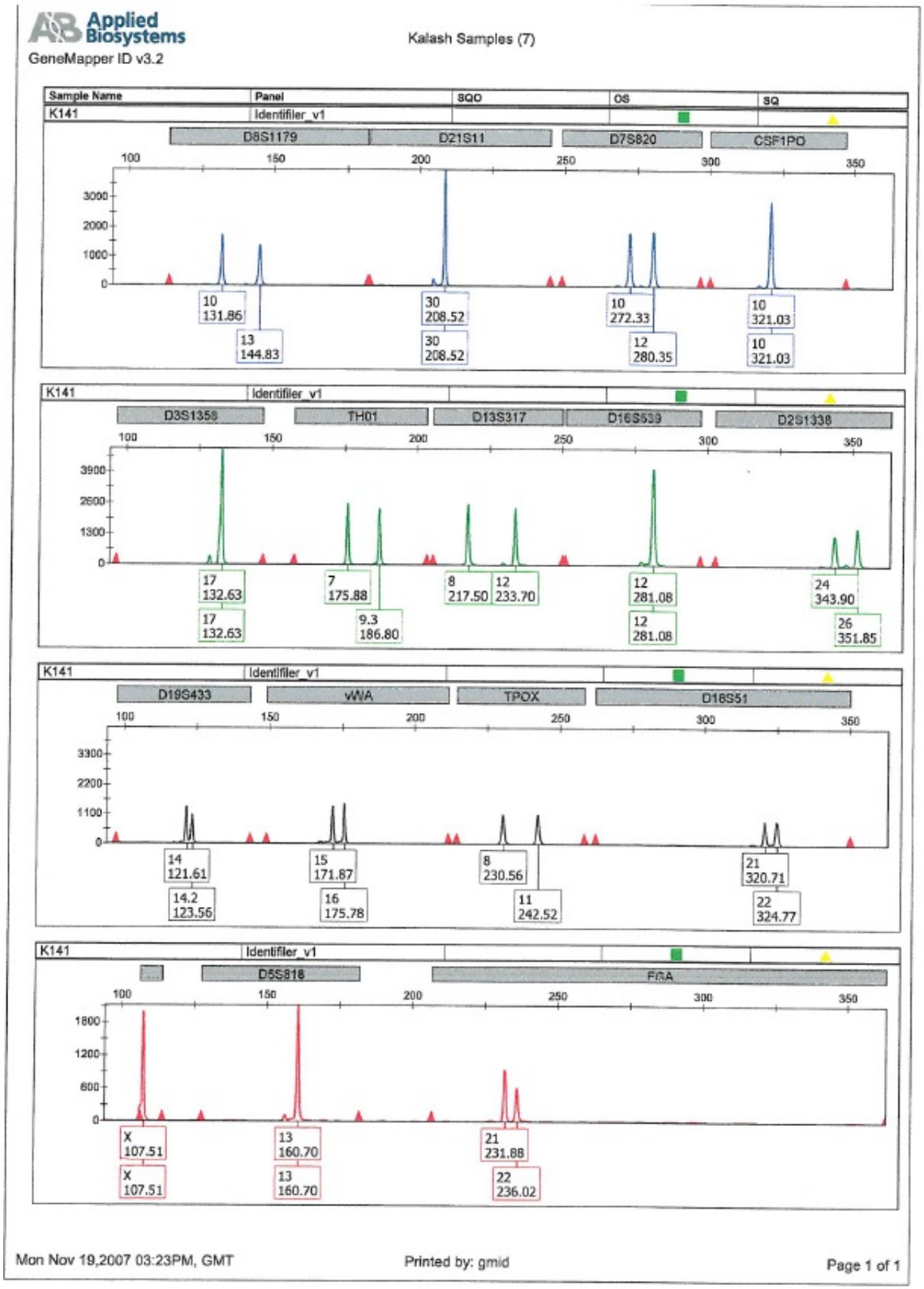


Figure 4.2: Electropherogram of a sample from the Kalash population after amplification with the AmpFℓSTR® Identifiler® PCR amplification kit



4.2.4 Exact Test of Population Differentiation

Population differentiation allows the identification of populations which may be genetically isolated from other tested populations and if so, to what extent (Balloux & Lugon-Moulin, 2002). The test is based on Fisher's R x C contingency table where the number of copies of a particular allele are recorded for each population under consideration (Raymond & Rousset, 1995). The null hypothesis to test in this case is that there are no significant differences in allele frequencies between pairs of populations at each locus (p = <0.05).

Table 4.5 shows the p values for the exact test of population differentiation across individual loci for all populations, note the Kalash showing significant differences in comparison with each population; the only exception being D16 when compared to the Pakistani data.

Table 4.6 summarises the significance of differentiation across all loci between all populations. There are few significant differences between the Indian and Pakistani populations which is not wholly unexpected given their geography though when considering all loci collectively, the null hypothesis that there is no significant difference between the allele frequencies of the Indian and Pakistani populations is rejected.

The effect isolated populations such as the Kalash have on profile frequency estimates are discussed further in Chapter 6. Chapter 5 will focus on the UK and larger Asian populations which make up a considerable proportion of the UK's minority population.

Table 4.5: *p* values for the exact test of population differentiation across individual loci of each population

	D3	vWA	D16	D2	D8	D21	D18	D19	TH01	FGA
In/Ka	**0.0001**	**0.0000**	**0.0291**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0011**	**0.0000**
In/Pa	0.1202	0.6205	0.7250	0.6301	0.0853	0.6384	0.7477	**0.0002**	0.4045	**0.0040**
In/UK	**0.0169**	0.2990	**0.0089**	**0.0000**	**0.0003**	**0.0000**	**0.0010**	**0.0000**	**0.0000**	**0.0000**
Ka/Pa	**0.0001**	**0.0000**	0.0557	**0.0001**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0001**	**0.0000**
Ka/UK	**0.0000**	**0.0000**	**0.0061**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**
Pa/UK	**0.0187**	0.3343	**0.0009**	**0.0000**	**0.0000**	**0.0000**	0.2837	**0.0000**	**0.0000**	**0.0006**

Table 4.6: *p* values for the exact test of population differentiation across all loci of each population

	India	Kalash	Pakistan
Kalash	**0.0000**	-	**0.0000**
Pakistan	**0.0028**	**0.0000**	-
UK	**0.0000**	**0.0000**	**0.0000**

In = Indian, Ka = Kalash, UK = UK, Pa = Pakistani

Values in bold show significant differences ($p < 0.05$) in population differentiation at individual loci and across all loci.

4.3 Discussion

4.3.1 Population Databases

In order to establish an accurate allele frequency table for the analysis of DNA profiles and perform robust, statistical analyses on those data, a good representation of a population is required for a database. To say that the datasets used have been of sufficient size to illustrate substructuring and the inherent differences between populations may be optimistic but this is a problem faced by many studies of this nature. If a database were to be inclusive of all members of the population/subpopulation of interest, confidence in that data would be greater. Similarly, if the entire population was included on a database, match probabilities would be superfluous and it would be simply a question of identity. Due to the practical limitations of a such a database, if a system were employed for forensic application to estimate the geographical origin of a perpetrator based on DNA evidence, as large a database as practicable would be advantageous to get as near true representation as possible.

When compiling a database, individuals are categorised based on where they say they come from, or what ethnicity they belong to and the belief that they are not related to other individuals being sampled. An ideal database would be one that closely matches the suspect's historical geographical origin and not, for example, a general 'Caucasian' database (Foreman, *et al.*, 1998). However, a deep understanding of an individual's ancestral history may not be available and there may be recent admixture which skews the allele frequency data of a population; to what extent, and whether it is significant depends on the level of admixture. The advantages of isolate populations, such as the Kalash, with regard to accurately assigning unknown DNA profiles to a population or geographical region are shown in this study, but in the more developed-world, people will continue to travel and emigrate and introduce new alleles to populations. In this study, Table 4.6 shows significant population differentiation between all populations

suggesting that, even with a limited selection of loci compared to studies such as Rosenberg *et al.*, (2002), and Rosenberg *et al.*, (2005), these differences may be utilised to assign unknown samples to the correct population.

4.3.2 Allele Frequencies

Considering all loci, no allele appeared at a substantially higher frequency than any other across all populations. The only exception is variant 24.2 at locus FGA which is more prevalent in the Kalash population with a frequency of 0.100 compared with frequencies of no higher than 0.016 in the Indian and Pakistani databases and not seen at all in the UK population (Graph 4.10).

4.3.2.1 Exact Test for Hardy-Weinberg

There were no significant departures from Hardy-Weinberg equilibrium once the Bonferroni correction had been applied which suggests that there is no significant difference between H_o and H_e for any of the loci analysed across all populations. All of the differences that would have been significant prior to the application of the Bonferroni correction, for example locus D7 of the Kalash database, show a lower observed heterozygosity than expected. This may be due to the presence of null alleles or suggests a system of mating where inbreeding is practised (Loeschcke, *et al.*, 1994).

One problem with the Bonferroni correction is that it increases the chance of Type II errors, that is, actual significant differences in the data are mistakenly accepted as not significant and so the null hypothesis is falsely accepted. Another is that the level of Bonferroni correction applied is decided by the number of tests (in this case equal to the number of loci) being carried out, yet the loci are independent of one another (Perneger, 1998). The Bonferroni correction is an ultra-conservative method of avoiding Type I errors – suggesting a significant difference between H_o and H_e when, in fact, there is not which is why it is necessary to repeat tests to ensure consistency.

5 EFFECT OF USING UK OR ASIAN POPULATION DATABASES ON PROFILE FREQUENCIES

5.1 Population Substructuring

In its simplest form, profile frequencies are calculated using a method known as the product rule (National Research Council, 1996). Using the allele frequencies obtained from the sampled population, genotype frequencies are calculated for each locus and then multiplied together to give a profile frequency. However, when trying a suspect in a court of law, questions have been raised as to how relevant the population sampled is to a) the suspect and b) the location of the crime (Hunter, 1998).

The UK is home to many diverse cultures and it has been argued successfully that to compare the DNA profile of a white British suspect to a database comprising samples from communities originating from the Indian subcontinent would seem unfair. If the degree of substructuring between the suspect's population and the database population is large, it will decrease the match probability thus making that profile appear rarer than it actually is (Foreman & Evett, 2001; Foreman, *et al.*, 1998). This would be deemed unfair and would work in favour of the prosecution. Restricting calculations to one broad database may only have a small effect on the match probability but it introduces bias against innocent suspects.

5.1.1 Balding and Nichols Correction

In 1994, Balding and Nichols described a method which was primarily developed to counteract the effects of kinship on match probability calculations: the idea that a criminal and an innocent person may share alleles from a common ancestor (identity by descent). Balding and Nichols interpreted F_{ST}, originally defined by Wright (1965), as the proportion of alleles within a subpopulation that share a common ancestor, though they do note that this interpretation is not akin to the original interpopulation definition.

It is important that DNA evidence is not overstated and, ideally, a database associated with a suspect's ancestral origins would be the most appropriate to use. Balding and Nichols' method was devised to incorporate the degree of population substructuring into the calculation of the match probability. Although F_{ST} values among populations may be known, this method allows for an increase in the correction applied to account for potential sampling error or correlations within and between loci tested.

Higher F_{ST} values caused by consanguineous marriages in cultures resident in the UK would skew match probability estimates for a white British individual. Balding and Nichols came up with two formulae (known as the 'Balding & Nichols correction') which work on the fundamentals of the Hardy-Weinberg principle but allow for population substructure (see section 2.3.2.4). To calculate the genotype frequency at each locus, the following formulae would be employed, the results of which are multiplied together to obtain a match probability:

$$\text{Homozygote: } \frac{[2\theta + (1 - \theta)p_i][3\theta + (1 - \theta)p_i]}{(1 + \theta)(1 + 2\theta)}$$

$$\text{Heterozygote: } \frac{2[\theta + (1 - \theta)p_i][\theta + (1 - \theta)p_j]}{(1 + \theta)(1 + 2\theta)}$$

Where p_i and p_j are the allele frequencies of the profile obtained as determined by the population database and θ (also described as F_{ST} when considering the Balding and Nichols definition) is the degree of substructuring exhibited within that population, if known. If θ is set at 0, these formulae provide the same results as if using the conventional 'product rule'.

Empirical studies have shown that actual values of θ tend to be low (National Research Council, 1996). Following on from the work of Balding and Nichols, the National

Research Council (1996) suggested that by routinely incorporating a value of F_{ST} = 1 % into the correction formulae, this will be adequate in addressing any uncertainty the effect of subpopulations may have had on allele frequencies. For those more isolated populations, a more conservative value of F_{ST} = 3 % may be used.

In the UK, the Forensic Science Service (FSS) adopted an overly conservative approach whereby F_{ST} = 3 % would be used when analysing UK 'Caucasian' and UK Afro-Caribbean populations and F_{ST} = 5 % for UK Indo-Pakistani populations (Foreman & Lambert, 2000). Although more favourable to the defendant, studies have since shown that these F_{ST} estimates may be a little extreme and that F_{ST} = 2 %, which equates to a 2 % differentiation between the suspect's subpopulation and the database population, is more accurate and still very generous in some cases (Foreman & Evett, 2001; Foreman, *et al.*, 1998).

5.1.2 Substructure in the South Asian and UK Populations

Table 5.1 shows the pairwise F_{ST} values between populations and their significance. Performing pairwise testing allows for a simple method of comparing genetic differentiation and geographical distance between the populations. The pairwise F_{ST} calculations were performed using Arlequin v. 3.1 (Excoffier, *et al.*, 2005). The greatest difference is observed between the Kalash and Indian populations at just under 3 %. Being the most isolated population sampled for this study, the Kalash showed the greatest differentiation between all populations. The level of correction applied to the Balding and Nichols formulae varies between forensic service providers (usually based on internal validation studies) but a minimum allowance of F_{ST} = 3 % is still used by some. This would therefore compensate for any substructuring between populations. Using a value F_{ST} = 5 % would be highly conservative even when comparing to a population as remote as the Kalash.

As expected, little difference is observed between the Indian and Pakistani populations, for example, only two loci (D19 and FGA) showed a significant difference between the

Indian and Pakistani populations (Table 4.5) and this is reflected in the low measure of F_{ST} pairwise differentiation at approximately 0.1 %. It is interesting to note that the pairwise difference between the Kalash and UK populations is not as great as that between the Kalash and Indian population. Based on geography alone, it would be an obvious assumption to make that greater geographic distance was linked to greater genetic differentiation (Manica, et al., 2005; Ramachandran et al., 2005). As Rosenberg et al., (2005), discuss, genetic differentiation was greater between populations in different clusters (see section 1.4.3.4) (perhaps more likely to be separated by geographical barriers) than those intracluster populations just separated by distance, even if the distance separating the pairs of populations is the same. Given the distinctiveness of the Kalash population, however, it is perhaps irrelevant to compare them to other populations in terms of geography as they would skew any correlation between genetic variance and geographical distance.

In this study, one may expect sampling bias to have an effect on the data collected. The Kalash samples were collected from three different areas and with a total of 115 samples, the number from each sampling area would have been relatively small. The effect of genetic drift on this small, isolated population is likely to be more pronounced due to this.

Table 5.1: Pairwise F_{ST} values between populations (lower diagonal) and their respective significance levels (upper diagonal)

	Indian	Kalash	Pakistani	UK
Indian		*	**	*
Kalash	0.02936		*	*
Pakistani	0.00146	0.02568		*
UK	0.01221	0.02709	0.01156	

* = p = 0.00000 ± 0.000 ** = p = 0.02010 ± 0.0012

5.1.3 STRUCTURE Analysis

To further assess evidence of population substructuring, an alternative Bayesian, probabilistic approach was taken using the STRUCTURE software, developed in 1999 (Pritchard, *et al.*, 2000). This allows samples to be assigned to *K* populations or clusters based on allele frequencies as it looks to maximise Hardy-Weinberg equilibrium and look for population groupings exhibiting minimal linkage disequilibrium (Pritchard, *et al.*, 2000). *K* may be assumed prior to analysis or calculated based on log likelihood Ln P(*X*|*K*) where *X* denotes the data of the individual samples and *K* represents the number of populations.

For this analysis, genotypic data are required and in this section the focus is on the UK, Indian and Pakistani data. The effect of the Kalash population on STRUCTURE estimates will be considered in the next chapter.

The software allows for various population models to be assumed prior to analysis. All populations in this study were tested against the admixture model with correlated allele frequencies which means any individual may have inherited some of its genetic ancestry from each population. This is based on the assumption that genetic drift has occurred since the evolution of modern humans and that we all stem back to this original population, thus sharing our ancestry (Rosenberg *et al*., 2005)

A priori knowledge of the number of populations (*K*) can also be set but this was not done here and the assignment of samples to populations was based purely on genetic structure. The program was asked to calculate structuring of populations based on *K* = 1 to *K* = 4. This was because it was not expected that the data would split into greater than four clusters and that the maximum log likelihood (Ln P[*X*|*K*]) will be achieved by *K* = 4; this indicates the most likely number of populations within the complete dataset. Runs of 20,000 iterations were used, preceded by a burn-in period of 10,000 iterations for each *K* as per Rosenberg *et al.*, (2002). Each simulation was also run 10 times to

check for consistency. Once analysed, each sample was assigned a probability of apportionment to each of the available clusters.

Table 5.2: Results of STRUCTURE analysis showing maximum log likelihood for *K* = 1 – 4, mean proportion of samples added to each cluster and in parentheses, the number of samples from each population showing greater than 75 % assignment to a cluster (if applicable)

K =	Ln P(X|K)	Cluster 1	Cluster 2	Cluster 3	Cluster 4
1	-22366				
2	-22443	0.500 (14 UK, 7 IN, 2 PA)	0.500 (0 UK, 9 IN, 19 PA)		
3	-22707	0.333	0.333	0.333	
4	-23015	0.250	0.250	0.250	0.250

The results in Table 5.2 show that there is inadequate genetic differentiation between the populations as clusters share an equal proportion of all samples depending on *K*. Also, the maximum likelihood calculated from an average of all 10 runs was found when *K* = 1. This is perhaps to be expected as it is well documented that most genetic diversity stems from within-populations rather than between them (Barbujani *et al*., 1997; Rosenberg *et al.,* 2002). This study is also based on using a common forensic DNA profiling kit, the AmpFℓSTR® SGM Plus® kit, which contains only 10 autosomal loci. Although this has been sufficient to show significant differences in the pairwise F_{ST} values (Table 5.1), it uses far fewer loci than the seminal studies on population structuring (Rosenberg *et al.,* 2002; Rosenberg *et al.,* 2005). Although there may be a genuine lack of substructuring in the sample set, sample size is also likely to affect the ability to measure any apparent substructuring. If the chosen loci are particularly informative, it may be sufficient to use less loci or a smaller sample size. Conversely, to obtain an accurate measure, greater loci or samples would be required (Rosenberg *et al*., 2001). In this study, sample sizes were based on those recommended for constructing representative databases (Chakraborty, *et al.*, 1974), not for performing STRUCTURE analysis.

However, the table also shows the number of samples from each population which show a 75 % or greater affinity to a particular cluster. As can be seen, all those from the UK that meet this criterion (14) are in Cluster 1 with fewer South Asian samples present (seven Indian and two Pakistani). The second cluster though shows an increase to nine Indian and 19 Pakistani samples assigned to it suggesting that all three populations contain a few individuals with extreme concordance to its particular population with an apparent divide between the UK and South-Asians.

At *K* = 3, no sample shows a great skew of over 75 % towards any particular cluster. This reinforces the inadequate level of substructuring reported by the clustering proportions and the relatively low (compared to 'standards' used by FSPs) maximum pairwise differentiation between the three populations: 2.9 % (Table 5.1). As *K*

increases, samples are apportioning to clusters more evenly which suggests there is no significant discontinuity in allele frequencies with regard to geographic difference. Although this appears to contradict the significance of differentiation shown in Tables 4.5 and 4.6 and the pairwise differentiation values in Table 5.1, the number of loci used in this study represent just under 3 % of those used by Rosenberg *et al.*, (2002), and, as discussed, could be having an adverse effect measuring substructure.

The Europe – South Asian divide in affiliation to clusters also highlights the separation of these populations following the African migration. A study by Ramachandran *et al.*, (2005), considered almost 1,000 points of origin in Africa for a serial-founder scenario to explain the source of human migration and diversity. Correlation coefficient values (R^2) of expected heterozygosity given geographical distance were greatest in Africa. Furthermore, the level of genetic diversity seen within Africa further supports this claim of it being the single point of human population expansion. From there, the population diverged into what is the modern day sub-Saharan population which subsequently split again and led to migrations to Eurasia, Oceania, East Asia and America (Zhivotovsky, *et al.*, 2003).

5.1.4 Cognate and Combined Databases

The forensic community has long encouraged the use of cognate databases to avoid the overstatement of DNA evidence (Gill *et al*., 2003). This has manifested in the development of a large number of databases representing not only countries but also regions and populations within those countries. However, in regions of the world with numerous recognised subpopulations, it would be implausible to have a separate database for each of them. In certain investigations it may not be possible to select a single database to represent the suspect as their origin may not be known and, therefore, it may be acceptable and more appropriate to use a combined database compiled from representative data of a broader geographical region.

As previously discussed, the FSS used a F_{ST} correction value of 2 % in conjunction with the Balding and Nichols correction. Other forensic service providers use different values depending on broad racial categories. These values are based on literature discussed within this study. However, it cannot be assumed that a value of F_{ST} = 2 % is representative of all populations, particularly those more isolated. If this study were to use this correction value as a standard then it would still leave some DNA evidence overstated, i.e. unfavourable towards the defendant, albeit perhaps not having much effect on the final match probability. In the UK, providers tend to use three or more databases, calculate match probabilities against each database, apply F_{ST} corrections and then quote the most conservative figure. This may still occur when the ancestral origin of an individual is known and fits into one of the broad categories of databases. To the defendant, the advantage of this method is that it practically negates issues over the 'wrong' database being used (because the highest match probability is being quoted). However, for an almost-complete DNA profile or better, the eventual interpretation of the DNA evidence is likely to be unaffected as a very small match probability would still be expected, regardless of the database used.

Due to the 'ceiling principle', where match probabilities lower than one in one billion are generally not quoted, it is highly likely that a match probability will be much lower when using an individual's 'correct' database and is therefore not something a defence team will usually seek knowing the most conservative figure has already been quoted.

To assess the effects of cognate versus combined databases, a method described in Gill *et al*., (2003), was adopted. The study examined the difference between match probabilities when a sample was analysed in its cognate database and then again in the combined database. The resulting parameter, *d*, then gives an indication as to whether the combined database is conservative or not and if so, to what extent.

In this study, the same method was utilised where:

$$d = Freq_{cg} / Freq_{cb}$$

Where $Freq_{cg}$ is the match probability of a sample calculated from its cognate database and $Freq_{cb}$ is the match probability of the same sample but calculated using the combined database, both incorporating the Balding and Nichols correction.

Based on this approach, a value of $d < 1$ would indicate that the match probability of a sample was more conservative in the combined database that in its cognate one. Conversely, where $d > 1$, this shows that a sample is more conservative in its cognate database. In a criminal case, this would therefore be more favourable to the defendant as it would make their profile appear more common than if it was analysed in a combined database comprising people from a broad range of locations that can only be assumed to be similar to where the defendant may be from.

To examine the differences in genotype frequency estimation and, therefore, match probability, a spread sheet was developed that could calculate profile frequencies for all samples in a population and automatically re-calculate those frequencies to incorporate any F_{ST} correction. Additionally, the sample's match probability was also calculated for each of the other population datasets and adjusted accordingly to any F_{ST} correction.

A combined south Asian database was constructed using data from the Indian (n = 172) and Pakistani (n = 157) datasets as well as a Bangladeshi population (n = 156) (Alshamali *et al*., 2005). Each sample from the Indian, Pakistani, UK and the genetic isolate Kalash population was then compared to this new database. To avoid bias, each sample from the Indian and Pakistani databases were removed from the combined database before being analysed with the new allele frequencies. To further test the effect of F_{ST} corrections on databases, these values were altered in both the cognate and combined databases.

Based on the work of Gill *et al*., (2003), Graphs 5.1 and 5.2 that follow show the level of conservativeness of the combined databases against the log of the cognate profile frequency of the dataset being analysed. A data point above zero indicates how many

times more conservative the calculated profile frequency is in the combined database compared to the cognate one. In contrast a negative value indicates how many more times the cognate database is more conservative than the combined one. F_{ST} corrections for the cognate and combined databases were set at F_{ST} = 2 % based on the recommendations of Foreman & Evett (2001) and Foreman, *et al.*, (1998).

Table 5.3 shows the percentage of samples from each database that appear more conservative in the combined database at varying corrections of F_{ST}. Table 5.4 illustrates an approximation of how high the F_{ST} correction would have to be in order for all samples in each database to become more conservative in the combined database.

Graph 5.1: Indian samples (*n* = 172) analysed against a cognate and combined South Asian database (*n* = 484 taking into account that each Indian samples is removed prior to comparison with the combined database)

a) No correction applied to either the cognate or combined databases.

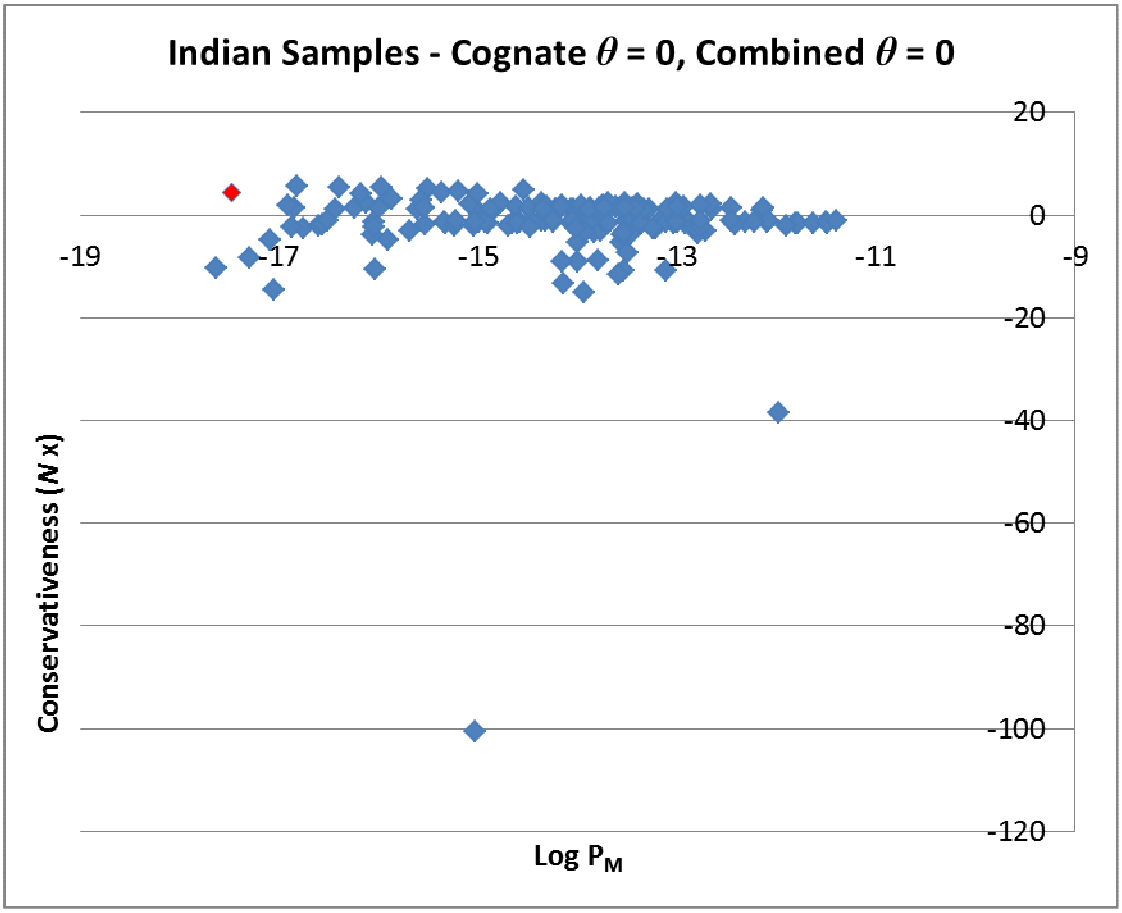
Percent > 0 = 42 %

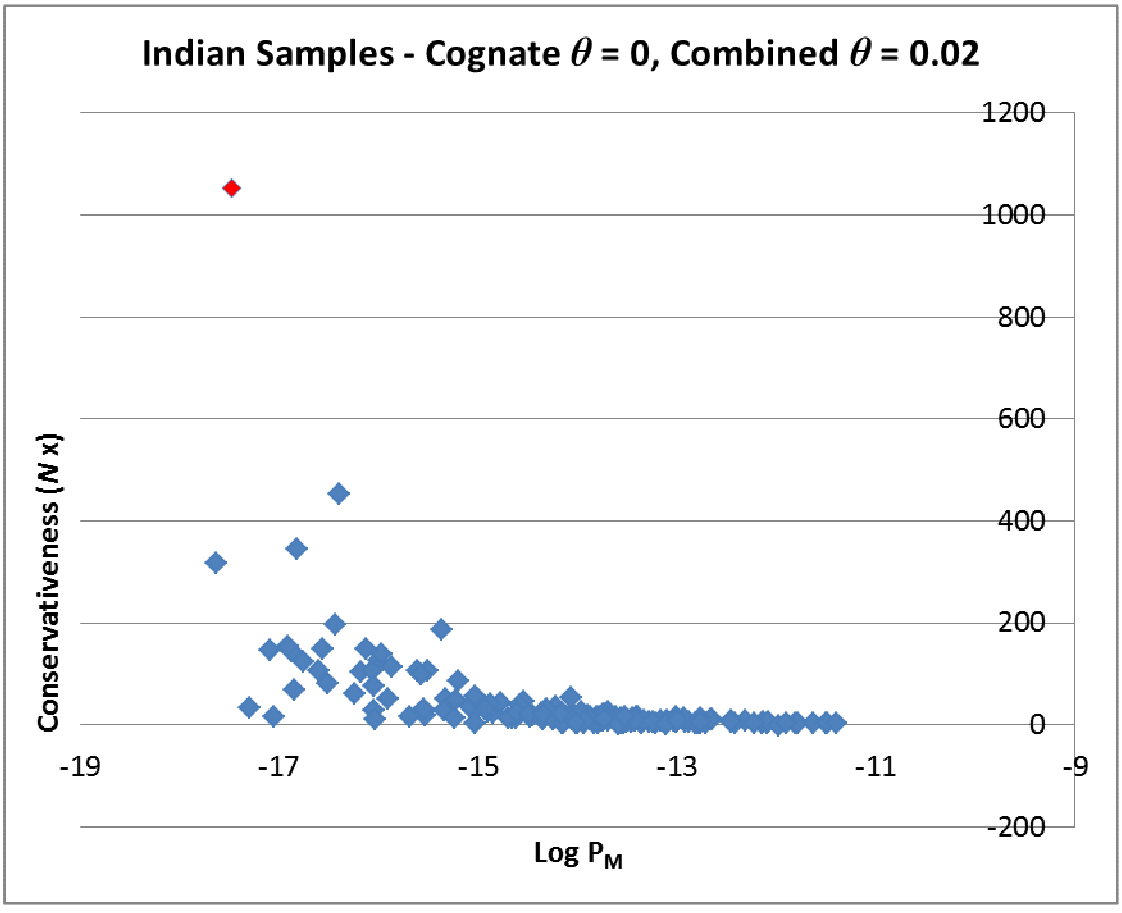
Each point on the scatter plot opposite represents a sample from which a complete DNA profile has been obtained. Less than half of the samples comprising the Indian database provide a more conservative (higher) match probability when compared to the relevant allele frequencies of the combined database. One sample has been highlighted in red to show how it is affected by the changes in F_{ST} that follow.

b) F_{ST} value of 2 % applied to the combined database.

Percent > 0 = 99 %

A F_{ST} correction of 2 % makes nearly all match probabilities more conservative when applied to the combined database only. This level of correction is in agreement with the findings of Foreman & Lambert, (2000), who state that F_{ST} = 2 % should be sufficient for a robust match probability calculation regardless of whether the database is representative of the suspect's true population. The sample highlighted in red is greater than 1000 times more conservative against the combined database with this level of correction.

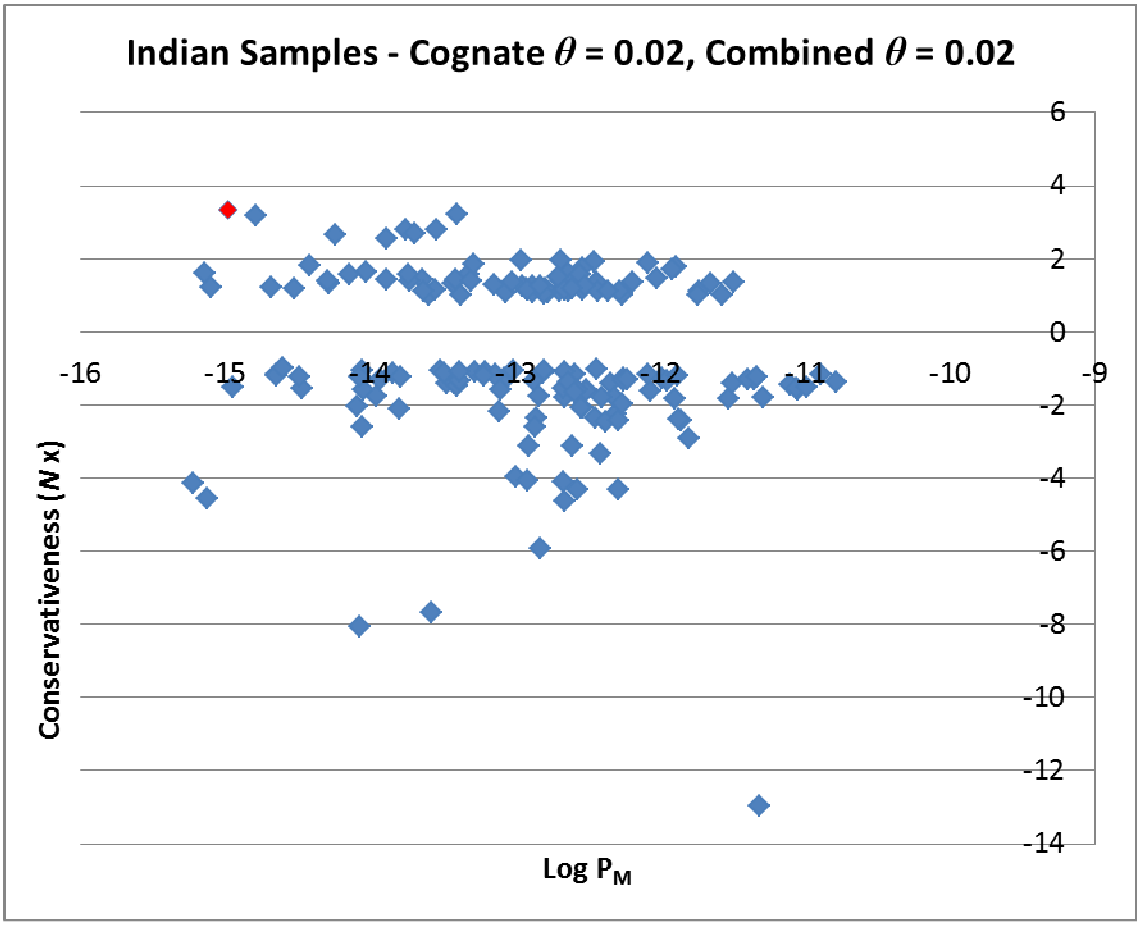

Indian Samples - Cognate θ = 0, Combined θ = 0
20
0
-20
-40
-60
-80
-100
-120
-19
-17
-15
-13
-11
-9
Conservativeness (N x)
Log P_M


Indian Samples - Cognate θ = 0, Combined θ = 0.02
1200
1000
800
600
400
200
0
-200
-19
-17
-15
-13
-11
-9
Conservativeness (N x)
Log P_M

c) F_{ST} value of 2 % applied to both the combined and cognate database.

Percent > 0 = 43 %

The samples cluster more tightly together (note smaller y-axis) around 0 for conservativeness (i.e. 0 would represent no difference in match probability between cognate and combined database). The red sample is more comparable with the other samples when F_{ST} = 2 % is applied to both databases and does not show such compelling conservativeness to the combined database as shown in Graph 5.1b.


Indian Samples - Cognate θ = 0.02, Combined θ = 0.02
6
4
2
0
-2
-4
-6
-8
-10
-12
-14
-16
-15
-14
-13
-12
-11
-10
-9
Conservativeness (N x)
Log P_M

Graph 5.2: UK samples (*n* = 252) analysed against its cognate and the combined South Asian database (*n* = 485)

a) No correction applied to either the cognate or combined databases.

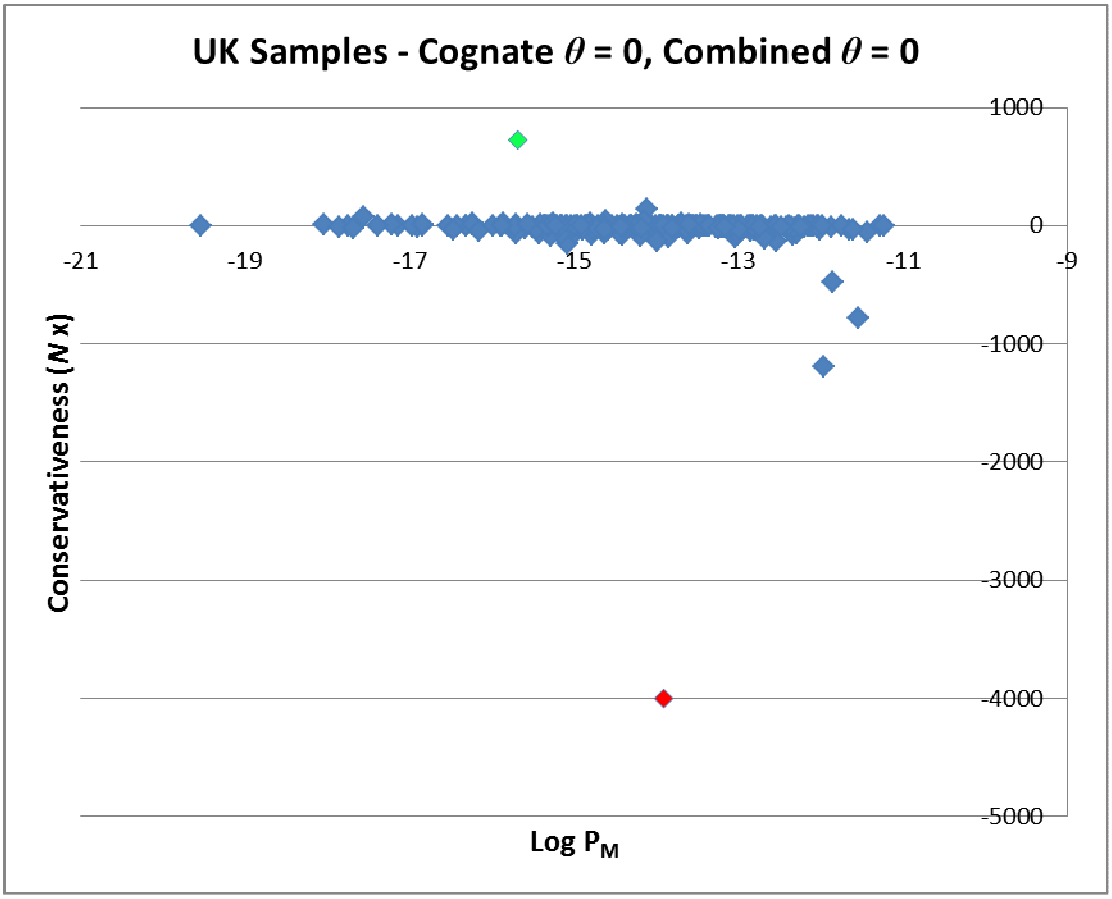
Percent > 0 = 21 %

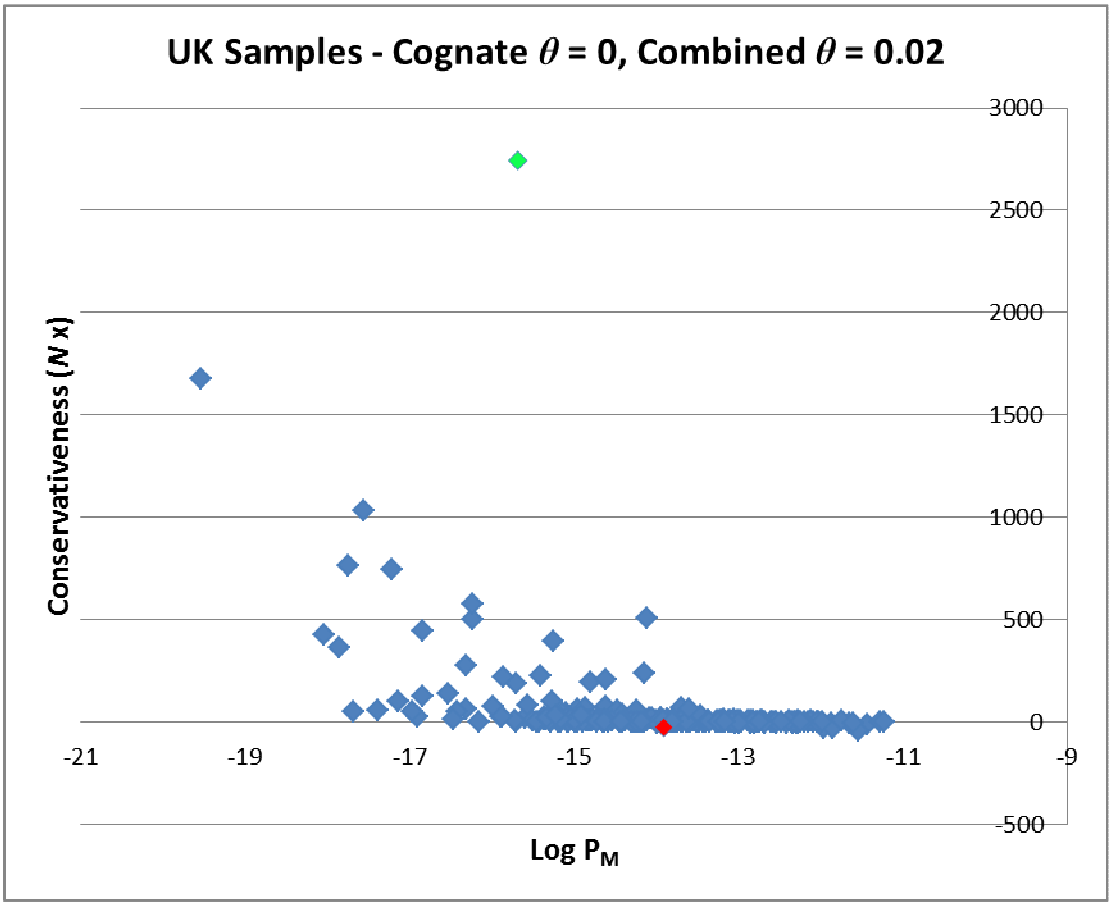
Fewer samples result in a more conservative match probability when calculated against the combined database which is expected given the greater genetic differentiation seen between the UK and other populations sampled in this study (Table 5.1). One sample (highlighted in red) shows approximately 4000 times greater affinity to the UK database. Another sample (highlighted in green) already shows over 700 times greater affinity to the combined database.

b) F_{ST} value of 2 % applied to the combined database.

Percent > 0 = 80 %

A similar pattern is observed as that of the Indian samples at this level of correction (Graph 5.1b): much greater clustering around the x-axis and the sample in red now shows approximately 25 times more conservativeness in its cognate database. The affinity of the green sample to the combined database has increased nearly four-fold.


UK Samples - Cognate θ = 0, Combined θ = 0
1000
0
-1000
-2000
-3000
-4000
-5000
-21
-19
-17
-15
-13
-11
-9
Conservativeness (N x)
Log P_M


UK Samples - Cognate θ = 0, Combined θ = 0.02
3000
2500
2000
1500
1000
500
0
-500
-21
-19
-17
-15
-13
-11
-9
Conservativeness (N x)
Log P_M

c) F_{ST} value of 2 % applied to both the combined and cognate database

Percent > 0 = 18 %

As seen with the Indian samples (Graph 5.1c), the application of F_{ST} = 2 % to both databases has reduced the extremity of any affinity to either database and clustered the samples nearer to the point of no difference between cognate or combined database. This is despite a reduction in the number of samples reporting a more conservative match probability in the combined database. Although not as great as when no correction is applied to either database, the sample in red has again shown a distinct affinity towards the UK database.

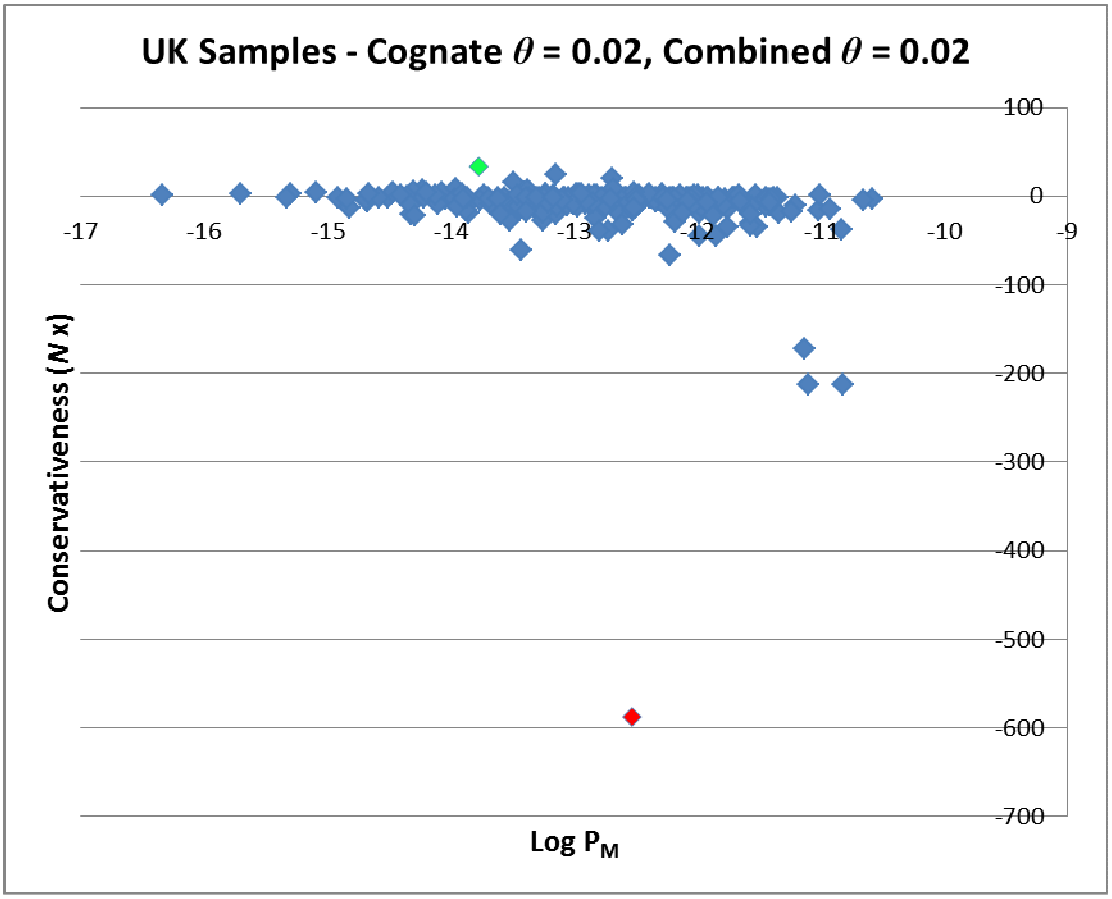

UK Samples - Cognate θ = 0.02, Combined θ = 0.02
100
0
-100
-200
-300
-400
-500
-600
-700
-17
-16
-15
-14
-13
-12
-11
-10
-9
Conservativeness (N x)
Log P_M

Table 5.3: Percentage of samples from each database which appear more conservative in the combined South Asian database with varying F_{ST} corrections applied to the combined database

F_{ST} Correction (%)	UK	Indian	Pakistani
0	21	41	46
1	57	95	95
2	80	99	98
3	94	99	100

Table 5.4: Approximate F_{ST} correction required in order to make all samples within a database more conservative in the combined South Asian database

Population	F_{ST} Correction (%)
UK	6.4
Kalash	12.0
Indian	1.7
Pakistani	2.7

What is evident looking at the Indian samples is that by using the previous FSS standard of F_{ST} = 2 %, almost all samples are more conservative when analysed against a combined South Asian database compared with the cognate data (Graph 5.1b). Indeed, the recommended F_{ST} value of 2 % is more than adequate when considering a suitable correction as 99 % of samples reported more conservative match probabilities when calculated against the combined database.

When no correction is made to either database, 42 % of the Indian sample match probabilities remain more conservative in the combined database (Graph 5.1a). This is not to be wholly unexpected as the combined database is made of South Asian populations, including the Indian samples taken for the purpose of this study. However, there is one sample with a profile frequency over 100 times more conservative in the cognate database. This is exceptional compared to the other samples in which the majority lie between approximately -20 and +15 and suggests the presence of at least one allele with a lower frequency than those in other samples.

When a correction of 0.02 is applied to the combined database only, the other extreme can be seen where one sample now reports a profile frequency over 1000 times more conservative in the combined database (highlighted in red – Graph 5.1b). The same sample showed approximately 4.4 times more conservativeness in the combined database when no correction is applied to either (Graph 5.1a). This difference may be explained by the low match probability (1.46×10^{-17}) evident when no correction is applied to either database; any correction applied thereafter will have a more profound effect on rare alleles. This can also be seen in the distribution pattern of the samples as, generally speaking, affinity to the combined database increases as match probability decreases. The fact that the reported match probabilities are low indicates the presence of alleles with lower frequencies in these Indian samples which are therefore showing greater affinity to those neighbouring populations included in the combined database.

When a F_{ST} correction of 2 % is made to the cognate database also, the number of samples reporting more conservative match probabilities in the combined database resembles that of the uncorrected data (Graph 5.1c). The difference is that the data are more tightly clustered; there are no such extremes of affinity as with the uncorrected.

With the UK samples, there is a much stronger initial affinity to the cognate database with some samples when compared to a South Asian database: one at -4000 (highlighted in red – Graph 5.2a). This is not altogether unexpected given the greater pairwise difference seen between the UK and other populations comprising the combined database (Table 5.1) but interesting: it takes a relatively modest correction of F_{ST} = 2 % in the combined database to make 80 % of UK samples more conservative in the combined database (Graph 5.2b).

Conversely, one sample (highlighted in green – Graph 5.2a) already shows an initial affinity to the combined database, increasing to nearly 2,750 times more conservative in the combined database following correction of F_{ST} = 2 % (Graph 5.2b). This result may be due to a sampling error: either the sample was initially labelled incorrectly, or; the individual was not aware of any admixture of previous mixed ancestry. As discussed by Serre & Pääbo (2004), samples are often dismissed from studies such as this if they do not fit into a population-specific category. Those that are believed to 'fit in' are based on cultural traits which may not truly define the 'population' or may be relatively recent in terms of human evolution. This is why they advocate a sampling across distance method as opposed to pre-defined conceptions of what constitutes a 'population'. In this case, it is possible that the green sample is exhibiting a degree of admixture; with allele frequencies more akin to those comprising the South Asian database.

Finally, a similar pattern to those seen with the Indian samples emerges when the correction is applied to both databases; the percentage of samples more conservative in the combined database returns to a figure similar to that when both databases are

uncorrected. However, the major difference between the Indian and UK samples is that the UK covers a much broader range in terms of affinity for one database or the other (Graph 5.1c and Graph 5.2c).

5.2 Prediction of Ancestry based on DNA Profile Frequency Estimation

As the F_{ST} and population differentiation statistics have shown, there are varying levels of substructuring between the populations observed. This means that calculating a profile frequency in a database based purely on HWE and the product rule could be inaccurate if no consideration is given to the inbreeding / substructuring effect (Balding & Nichols, 1994; Foreman & Lambert, 2000; Foreman & Evett, 2001; Gill *et al.*, 2003).

As well as generating profile frequencies for each sample within its cognate database, the genotypes of each sample were also compared to the allele frequencies of the other three populations and a profile frequency generated from each of them. This was to test which population each sample would be assigned to had its origin been unknown (see section 6.5.1). A sample was assigned to the population which reported the highest frequency, i.e. more likely to appear in that population than any other. This was based purely on the frequency of occurrence of the alleles at each locus; no other genotypic or phenotypic factors were considered at this stage.

The analysis was performed for each database and in order to visualise the effect of differentiation between the populations, the natural log (ln) of each sample's profile frequency was taken for all four populations that it had been compared to. Each population pair was then plotted on a scatter graph with each point representing the natural log of the profile frequency calculated from its origin population against that of the profile frequency calculated against the 'wrong' database.

To observe what effect varying F_{ST} values had on profile frequencies, F_{ST} values of 0, 2 and 5 % were used to alter the genotype frequencies accordingly between a population pair. Graph 5.3 shows the effect that substructuring has between the UK and Indian populations and Graph 5.4 shows how little genetic difference there is between the Indian and Pakistani populations in terms of making an attempt to assign a sample to

one or the other and what little difference it would make if an Indian dataset were used where a Pakistani suspect were concerned or *vice-versa*.

Graph 5.3: Effect of substructuring on the UK and Indian populations with varying F_{ST} values. Profile frequencies for each sample are calculated using both the UK and Indian databases

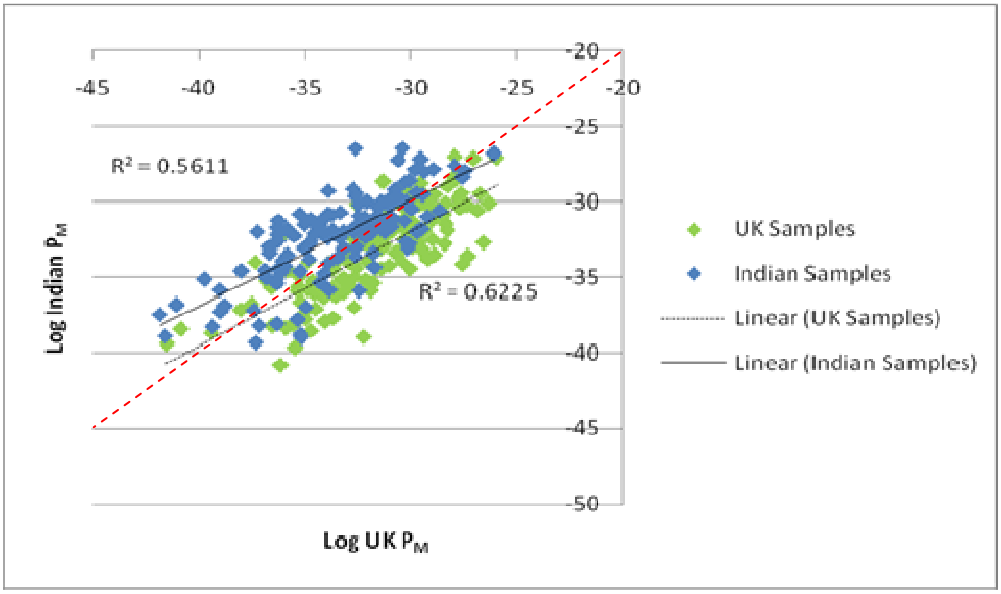
a) F_{ST} = 0 %

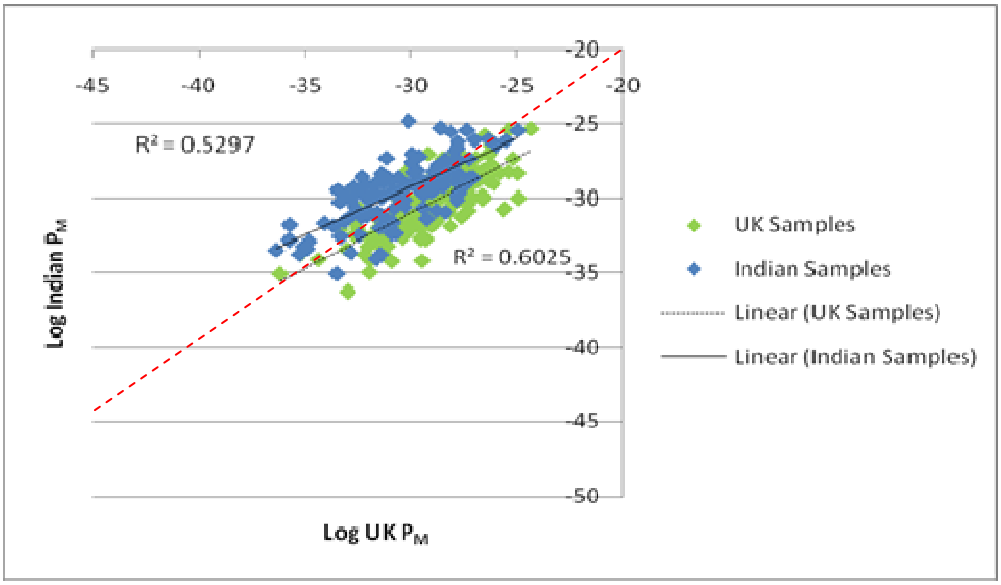
With no correction applied, there is evidence of differentiation with the samples from each database mostly falling either side of the x = y line according to population.

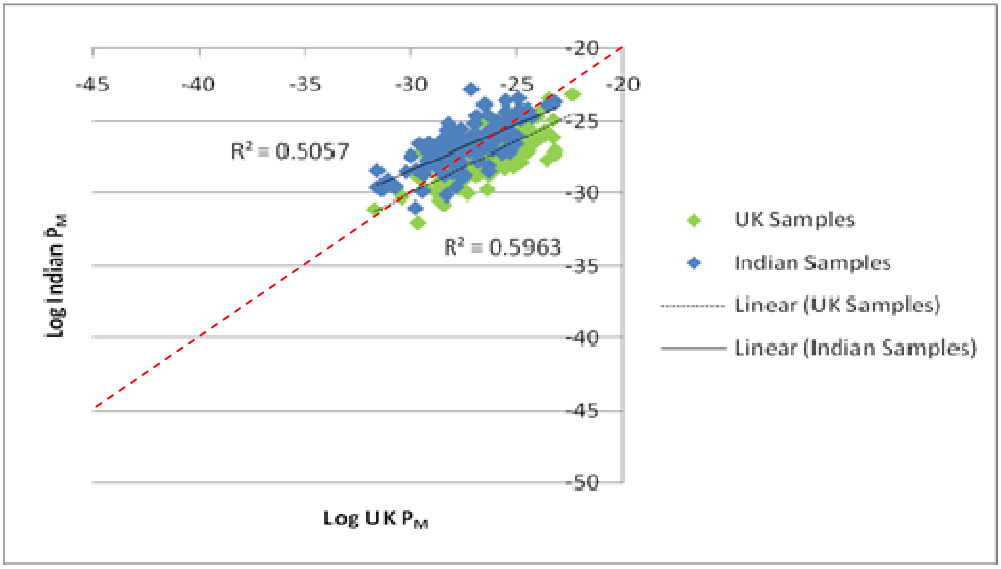
b) F_{ST} = 2 %

c) F_{ST} = 5 %

As F_{ST} increases, the differentiation becomes less apparent, reiterated by a reduction in correlation coefficient for each population. This suggests the level of correction applied is accounting for the effects described previously (see section 5.1.2) and reducing the disparity seen between what may be the offender's population and the 'wrong' population.


-45
-40
-35
-30
-25
-20
-20
-25
-30
-35
-40
-45
-50
R² = 0.5611
R² = 0.6225
Log Indian P_M
Log UK P_M
UK Samples
Indian Samples
Linear (UK Samples)
Linear (Indian Samples)


-45
-40
-35
-30
-25
-20
-20
-25
-30
-35
-40
-45
-50
R² = 0.5297
R² = 0.6025
Log Indian P_M
Log UK P_M
UK Samples
Indian Samples
Linear (UK Samples)
Linear (Indian Samples)


-45
-40
-35
-30
-25
-20
-20
-25
-30
-35
-40
-45
-50
R² = 0.5057
R² = 0.5963
Log Indian P_M
Log UK P_M
UK Samples
Indian Samples
Linear (UK Samples)
Linear (Indian Samples)

Graph 5.4: Effect of substructuring on the Indian and Pakistani populations with varying F_{ST} values. Profile frequencies for each sample are calculated using both the Indian and Pakistani databases

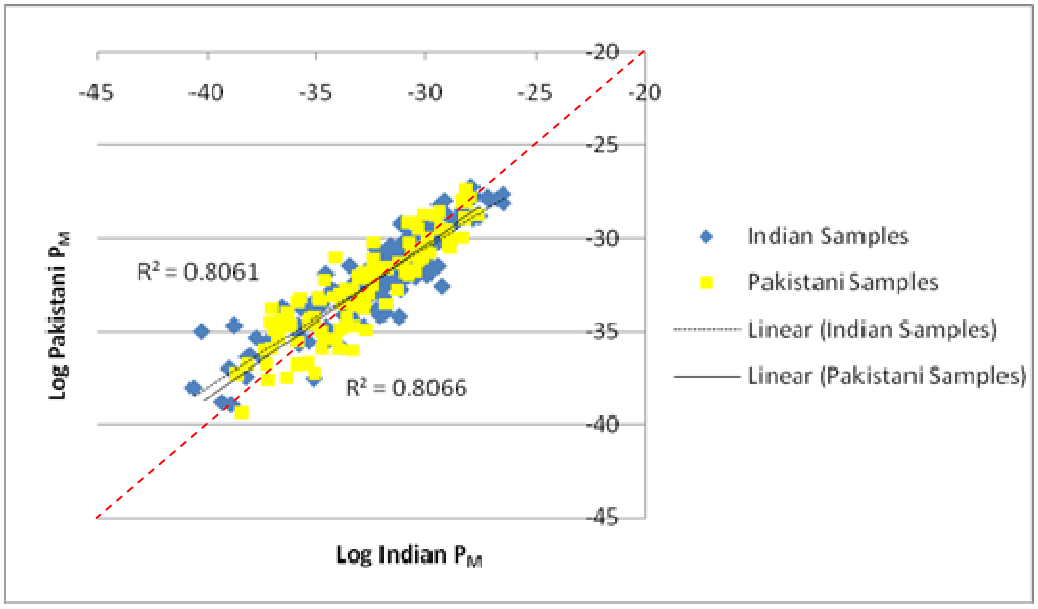
a) F_{ST} = 0 %

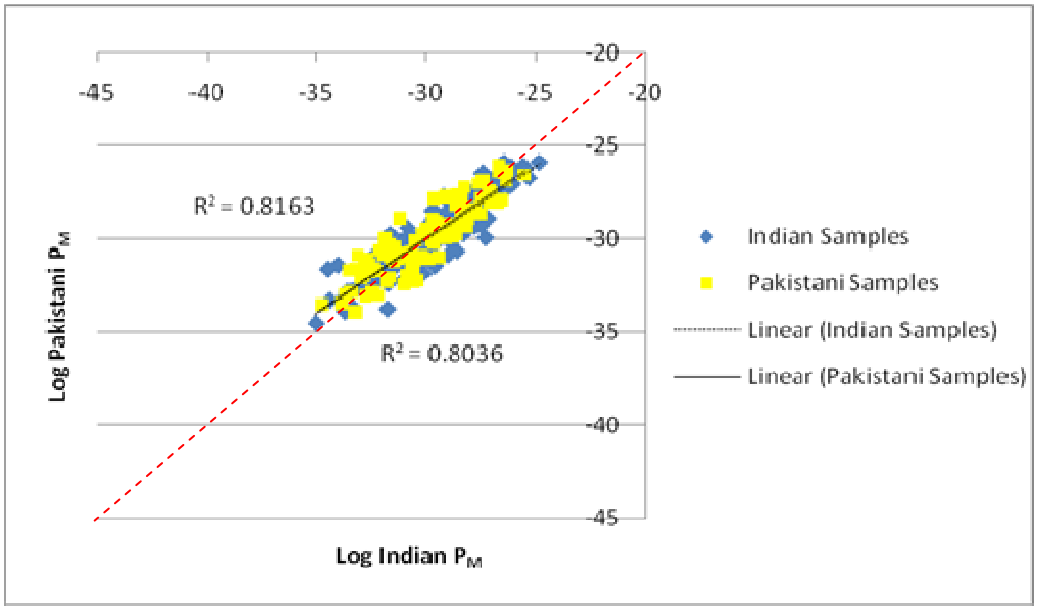
With no correction applied, there is a much closer relationship apparent between the Indian and Pakistani samples, clustering closely together around x = y. The higher correlation coefficients show the similarity between the Indian and Pakistani allele frequencies.

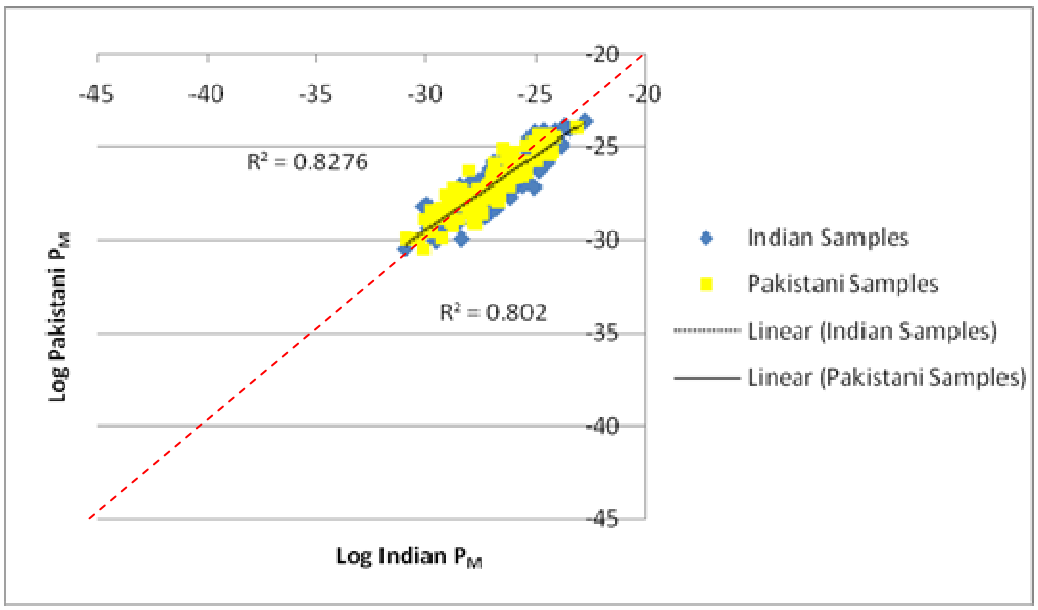
b) F_{ST} = 2 %

c) F_{ST} = 5 %

Again, the increase in F_{ST} correction to both databases serves to reduce the match probability of each sample and, in doing so, makes using the 'wrong' database less problematic.


-20
-45
-40
-35
-30
-25
-20
-25
-30
-35
-40
-45
Log Pakistani P_M
R^2 = 0.8061
R^2 = 0.8066
Log Indian P_M
Indian Samples
Pakistani Samples
Linear (Indian Samples)
Linear (Pakistani Samples)


-20
-45
-40
-35
-30
-25
-20
-25
-30
-35
-40
-45
Log Pakistani P_M
R^2 = 0.8163
R^2 = 0.8036
Log Indian P_M
Indian Samples
Pakistani Samples
Linear (Indian Samples)
Linear (Pakistani Samples)


-20
-45
-40
-35
-30
-25
-20
-25
-30
-35
-40
-45
Log Pakistani P_M
R^2 = 0.8276
R^2 = 0.802
Log Indian P_M
Indian Samples
Pakistani Samples
Linear (Indian Samples)
Linear (Pakistani Samples)

Comparing the UK and Indian populations, there is evidence of substructuring, albeit subtle. As the value of F_{ST} increase, the points on the graph gather closer together as the natural log of the frequencies increases. This is a result of the sample becoming seemingly more common in a population as the level of differentiation between populations is taken into consideration. The effect of population comparisons with the more isolated Kalash population will be examined in Chapter 6.

5.2.1 Geographical Assignment of Individuals

What can be observed in all six graphs representing substructuring is that there are several samples in each that appear as though they should belong to the population they are being compared to rather than their own. This can have implications when trying to assign samples to the correct population based on the profile frequency alone, without prior knowledge of their origin.

To test this, each sample was assigned to a population based on which database reported the highest frequency. The results are shown in Chapter 6 to include the Kalash population but in an ideal situation, each sample would be assigned to the population of origin, but as Graphs 6.3 and 6.4 indicate, there is overlap between populations which will affect the success of matching a sample to the correct population; the effects of substructuring will also be discussed.

5.3 Discussion

Match probability calculations are often based on data collected from three very broad population categories, but in some countries, the chances of the true perpetrator of an offence fitting perfectly within one of those categories may be slim. Corrections are applied to take into account differences between the offender's true population and the database in which their profile is actually compared to. For the purpose of match probability calculations, even if the true population is known, it is often treated as unknown to provide the most conservative outcome.

By not including frequency data based on the offender's true population, this instantly introduces bias and needs to be corrected (Triggs, *et al.*, 2000). This is linked to the effect of substructuring in populations which is evident within the three populations considered here. As expected, no significant differences are seen between the Indian and Pakistani populations if the profile frequency of a sample from one population is inadvertently calculated based on the dataset of another; in effect, they are the same population. This is supported by the pairwise difference reported in Table 5.1 and the close fit of samples with the alternate population in Graph 5.4. With the UK population, even if a match probability is calculated using a South Asian database and taking into account the corrections applied by forensic service providers, it is unlikely to have a significant overall effect on the how this match probability is perceived by a court.

As a general rule, the most conservative, corrected match probability estimated from the in-house databases held by the forensic service providers will be the one quoted for court proceedings; even if the ethnic origin of a suspect is known. This negates the need for discussion as to the best database to use with the exception of when the suspect may have come from a particularly isolated population.

This does not mean to say though that substructuring is not an important factor that requires consideration – it would depend on which populations were being studied, how

broad they were and how similar or dissimilar they may be to the 'true' origin of a sample if they were to be used for profile frequency estimation.

A study of autosomal markers in India (Kashyap, *et al.*, 2006a), including the 13 CODIS loci (Butler, 2006), showed that there was no evidence of population grouping based on language, socio-cultural practices or geography bar two areas sampled which showed greater affiliation to clusters not seen at such high levels with neighbouring populations. The authors expressed caution at the implications of broad forensic databases for calculating match probabilities. Variation between groups was low and the highest recorded F_{ST} was just over 3 %, therefore, the correction previously employed by the FSS (2 %) would risk overstating the strength of the DNA evidence. This difference of 1 % may only affect the match probability by one or two orders of magnitude and therefore have little adverse effect on the evidential value of the evidence (Balding & Nichols, 1994). Other forensic providers take a far more conservative approach to the use of Asian databases, with corrections of up to F_{ST} = 5 % and this would more than suffice in the majority of cases.

Within the UK, work has also been conducted at a more regional level, which shows that there is very little difference, or, at least, none of any practical importance, when measuring population differentiation on a more discrete geographical basis, i.e. counties or towns (Evett *et al*., 1996b; Foreman, *et al.*, 1998). Indeed, in this study alone, 94 % of the 252 samples taken from volunteers from all over the UK appeared more conservative in a South Asian database with a correction of F_{ST} = 3 % (Table 5.3). This is not something that would be expected if there were regional isolates within the UK.

It is important to differentiate between the reasons for applying correction factors in order to avoid overstating the value of DNA evidence and taking into account actual substructuring. In the UK, for example, a correction factor of at least F_{ST} = 2 % may be used regardless of the suspect's presumed origin, even if known to be a white British

individual. This is to avoid overstating the strength of the DNA evidence rather than take into account substructuring within the UK, which has already been shown to be at a much lower level than F_{ST} = 2 %. If the origin of a sample was known and this was compared to a corresponding dataset, there would be no need to take into account substructuring as the profile frequency would be calculated based on representative allele frequencies where the suspect's profile should appear relatively common compared to other databases. This is why the Balding and Nichols correction was formulated: by applying θ, this compensates for any uncertainty concerning the allele frequencies which are being used to estimate the profile frequency instead of those which come from the suspect's true (but potentially unknown) population.

With genuine population substructuring, F_{ST} may be applied to take into account the decrease in heterozygosity, where consanguinity and geographical barriers have perhaps led to increased homozygosity and, subsequently, genetically distinct subpopulations (Overall, 2009) which may, or may not, be in Hardy-Weinberg equilibrium themselves. As will be discussed in Chapter 6, using a convenient database based on geographical factors will not necessarily be enough to compensate for true substructuring within a population.

The effects and incorporation of substructure parameters in genotype frequency estimations had little effect on the magnitude of the potential evidential value of DNA evidence, even when applied at higher than recommended levels. This is despite significant differences between pairwise population F_{ST} comparisons (Table 5.1) and pairwise population differentiation at all loci (Table 4.6). Although there appears to be a lack of substructuring considering the results of the STRUCTURE analysis (Table 5.2), the number of loci analysed in this study is likely to have been insufficient for STRUCTURE to detect disparity between the populations.

Undoubtedly a larger sample size for each database with an increase in the number of loci examined may assist in detecting substructuring. Although knowledge of population

structure is most useful when comparing a population to a more isolated one, there are clear differences between the UK and Pakistani databases and the UK and Indian databases (samples of which were actually collected in Preston, England) despite approximately 250 samples or fewer being collected from each population.

The Pakistani and Indian populations showed the least divergence from each other with a difference of approximately 0.1 % between them. This is despite the Indian samples being taken from a Gujarati community within the UK and the Pakistani samples being taken from within Pakistan. So although these two datasets have come from two geographically-distinct locations, it is difficult to distinguish between them.

The UK database showed a significant pairwise F_{ST} difference to both the Indian and Pakistani databases of 1.2 % which would be sufficiently compensated for using the Balding and Nichols correction should a UK sample be compared to a database comprising the two Asian datasets. By using correction factors greater than those that might be necessary, it allows for additional uncertainties: those which may be realised and evident as well as those which are not. If the standard corrections used by forensic service providers can allow for differences between such diverse populations such as the UK and South Asians, then it should also be sufficient to correct for differences within populations which are not obvious.

The major databases used for profile frequency estimations are deemed to exhibit minimal levels of substructure (Foreman & Lambert, 2000), so providing the appropriate correction is applied, the database used to calculate the match probability of any profile should not unfairly overstate the DNA evidence.

6 EFFECT OF AN ISOLATED POPULATION ON PROFILE FREQUENCY ESTIMATIONS

6.1 Introduction

The previous chapter has highlighted that although significant differentiation was seen between loci and the population datasets collected for this study, the effect this had on match probabilities is relatively small in terms of the weight of DNA evidence, particularly when a complete DNA profile has been obtained. Despite the inability to detect any substructuring, the application of an appropriate level of F_{ST} correction was required to remove the majority of bias caused by the use of a non-cognate database relative to the 'offender population'.

The current practices employed by forensic service providers of applying correction of at least F_{ST} = 3 % to allow for population substructuring and co-ancestry appears to be more than adequate to compensate for any unknown substructuring as well as sampling inadequacies. Indeed, it may be deemed too great a correction as the maximum differentiation seen between the UK and Gujarati populations in this study was approximately 1.2 %. This shows that were a non-cognate database used as a reference for an 'unknown' profile, differences between them will likely be accounted for.

Those samples which continued to show greatest affinity to their cognate database still reported match probabilities in the order of less than one in one billion. Although a comparison of these samples to a combined database would therefore be less conservative, the match probability will still have been increased and applying a correction factor at a level where it is practically certain that all samples compared against it would appear more conservative runs the risk of vastly understating the value of the DNA evidence (Gill *et al.*, 2003).

The aim of this part of the study is to examine the effect a genetically isolated population may have on match probability calculations of individuals from other, perhaps more genetically diverse, populations when calculated using a database which may not be representative of their own. In such a scenario, it may be argued in court that the quoted match probability overstates the power of the evidence as it is not based on a suitable dataset most likened to the suspect's true population of origin.

In contrast to this, it must also be noted that specifying the actual population most akin to that of the suspect will be practically impossible (Foreman, *et al.*, 1998). Another important factor to consider is location: should a population dataset based on that of the suspect or on the location of the crime be used, based on the premise that somebody local committed it? Databases tend to consist of "convenience" samples: populations covering large geographical distances or particular ethnic groups. The correction factor incorporated into the product rule by Balding and Nichols (1994) allows not only for variation in allele frequencies seen within the suspect's subpopulation, but other uncertainties such as sampling error, for instance, it allows for the possibility that someone in the database may not be a 'true' member of that population/subpopulation. The higher the correction factor applied, the more conservative the reported genotype frequency and, subsequently, match probability.

6.2 Isolated Populations

Genetically isolated populations can further hinder attempts to identify the true effect of substructuring if endogamy is also a factor. Many Asian and African populations who practise consanguinity have the potential to disproportionately elevate F_{ST} values perhaps beyond those regularly accounted for in DNA profiling (Zhivotovsky *et al.*, 2001).

When Curran and Buckleton (2007) examined this further, they noted that sampling of families had induced higher values of F_{ST}. The risk, they state, is that calculations of substructuring can differ depending on the sample dataset and particular tests employed to analyse that data. What is sometimes overlooked is the level of relatedness between people within the same community and in these situations, the conventional formulae proposed by Balding and Nichols may not be the most appropriate.

The wider implications highlighted here are that population databases used by forensic service providers need to be made up of random, unrelated people who represent a good cross-section of the population. Also, when sampling a population and estimating a value to define genetic differentiation, caution must be applied with regards to how much of the observed differentiation is due to close relatedness and how much is due to actual substructuring. Curran and Buckleton (2007) conclude that there are formulae which may be better placed at calculating match probabilities when the effect of relatedness and sub-structuring are better known within a population.

Earlier work by Curran, *et al.*, (2003) attempted to examine the effect of substructuring on match probability calculations and found that although small, an effect was present. There is a chance that a match probability could be biased towards the prosecution if substructuring had not been taken into consideration. However, there are no set rules for the level of substructuring that may be exhibited by a population thus highlighting the benefit of using a population comparable to a suspect where practicable.

In this study, the Kalash population of the North West Frontier Province of Pakistan were sampled. They were selected for their perceived cultural and genetic isolation: some showing observable phenotypic differences to their Pakistani neighbours and having stronger ancestral links with Central or Eastern Asia rather than Europe or the Middle East (Mansoor *et al.,* 2004).

6.2.1 Estimation of pairwise F_{ST} values

Table 5.1 shows the pairwise F_{ST} values including the Kalash population. The pairwise F_{ST} comparisons show a greater than 2.5 % variation when compared to all other populations ($p = 0.0000$). A clear differentiation is evident in all standard SGM Plus® loci when each population is compared to the Kalash ($p = 0.0000$ for each comparison). At each locus, the Kalash show significant differences compared with each population; the only exception being D16 when compared to the Pakistani database (Tables 4.5 and 4.6).

6.3 Statistical Analyses

6.3.1 STRUCTURE Analysis

As discussed in the previous chapter, STRUCTURE analysis was performed again but this time to include the Kalash population. To account for the expected increase in estimated population number by the software, the range of *K* was increased to cover up to six populations.

A priori knowledge of *K* was again withheld to allow for population-classification of each sample based on the genetic data alone. Except for the number of assumed populations to test for (K), the same settings were used as for the previous experiment: runs of 20,000 iterations, preceded by a burn-in period of 10,000 iterations for each *K* = *n*. Each simulation was run 10 times to ensure consistency.

As Table 6.1 shows, STRUCTURE reports the highest log likelihood at *K* = 3 and the number of samples from each population apportioned to a cluster where affinity is 75 % and over. Although there are clear differences in the number of samples from each population showing greater than 75 % affiliation to each of the three clusters, overall sample membership is roughly symmetrical (~1 / *K* to each cluster). Given the symmetric apportionment of samples to clusters, this suggests a lack of population structuring (Pritchard, *et al.*, 2010). However, the differences seen between populations with regard to individual sample assignment suggests a level of allelic differentiation great enough for STRUCTURE to detect. The roughly symmetrical cluster assignment indicates that too few samples from each population show a great enough affinity to a cluster to detect overall structuring.

Interestingly, when the UK, Indian and Pakistani populations were analysed with STRUCTURE alone, the analysis estimated that *K* = 1 was the most likely proposition as it could not differentiate between the three datasets (Table 5.2). With the addition of the Kalash, it now appears to have recognised a distinction, albeit perhaps slight,

between the UK data and the Indian and Pakistani data by estimating that three populations are present. Apportionment shows that cluster 1 contains the largest number of Kalash individuals, cluster 2 the largest number of Indian and Pakistani individuals and cluster 3, the largest number of UK individuals. Initial evaluation may lead to the suggestion that it is assumed that the Indian and Pakistani data are more closely related to the Kalash than the UK. However, Table 5.1 contradicts this: showing a greater pairwise difference between the Kalash and UK populations than the Kalash and Pakistani populations. Pritchard, *et al.*, (2010) note that STRUCTURE is most effective with small, discrete populations and in those analyses including admixed populations, estimation of the 'correct' *K* may essentially be arbitrary. This may explain the difference seen between Tables 5.2 and 6.1. It is worthy to note that the estimated probability of the collective data comprising two genetically-distinct populations (Ln P[X|K]) is only slightly less than that calculated for three.

Table 6.1: Results of STRUCTURE analysis showing maximum log likelihood for *K* = 1 – 6, mean proportion of samples added to each cluster and in parentheses, the number of samples from each population showing greater than 75 % assignment to a cluster (if applicable)

K =	Ln P(X|K)	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
1	-26891						
2	-26793	0.510	0.490				
3	**-26769**	**0.302**	**0.347**	**0.349**			
		(6UK, 4IN, 3PA, 76KA)	(0UK, 23IN, 28PA, 11KA)	(31UK, 18IN, 6PA, 2KA)			
4	-27093	0.250	0.249	0.255	0.246		
5	-27090	0.201	0.200	0.201	0.200	0.198	
6	-27494	0.168	0.166	0.165	0.167	0.168	0.166

6.3.1.1 Affiliation to Clusters

Figure 6.1 shows estimated population assignment proportions for each sample. Each vertical line represents one individual and the coloured segments indicate an individual's affiliation to a particular cluster. The bar plots are sorted into populations as labelled at the bottom of the figure.

At *K* = 2, there is little obvious distinction between samples belong to the UK, Indian and Pakistani populations with samples showing an approximate 60/40 split between clusters. The Kalash, however, show approximately 80 % association with one particular cluster and differentiation is clear at this estimation of *K*.

When three populations are assumed, the affinity of Kalash samples to one cluster increases to approximately 90 %, however several samples from the other populations also show partial membership to this Kalash-dominated cluster. There is also a clearer difference at the point where the UK samples end and the Pakistani start in the cluster depicted by green, with both the Pakistani and Indian populations showing similar memberships in each cluster.

There is no single population showing exclusivity to one particular cluster and this suggests a clinal gradation between populations or admixture between neighbouring populations. The fact that the Kalash show some association with the Indian and Pakistani populations supports the proposition of their European or Middle Eastern Origin as no cluster affiliation would be expected to be seen between the Kalash and East Asian populations (Rosenberg *et al.,* 2002).

One limitation of these analyses are that the data are based on loci contained within the SGM Plus® loci only which were not chosen for how informative they are when it comes to distinguishing between regional groups. According to some studies, dinucleotides may provide greater population resolution given their higher mutation rate (Weber & Wong, 1993; Rosenberg *et al.*, 2003). The SGM Plus® kit consists of

tetranucleotide loci so further analysis of these samples using a dinucleotide multiplex spanning a greater range of alleles may increase the precision of membership assignment to clusters. Although tetranucleotides are deemed the least informative when it comes to population assignment (Rosenberg *et al.* 2003), this is advantageous in forensic DNA profiling where a lack of differentiation between populations allows for highly diverse loci across all populations and therefore more robust match probability estimates.

Figure 6.2 that follows shows a cluster plot at $K = 3$ (as determined by the highest log likelihood) but in this instance, STRUCTURE is setup to allow for the four inferred populations (UK, Indian, Pakistani and Kalash). The nearer an individual sample to a corner then the stronger the affiliation to that particular cluster. In this case, there appears to be considerable admixture between all populations but there is some distinction between the Kalash and UK samples compared to the Indian and Pakistani individuals. As stated before, the lack of clear distinction is most likely due to the limited loci tested rather than truly admixed populations.

Figure 6.1: Bar plot to show distribution of samples between clusters with increasing number of assumed populations, *K*

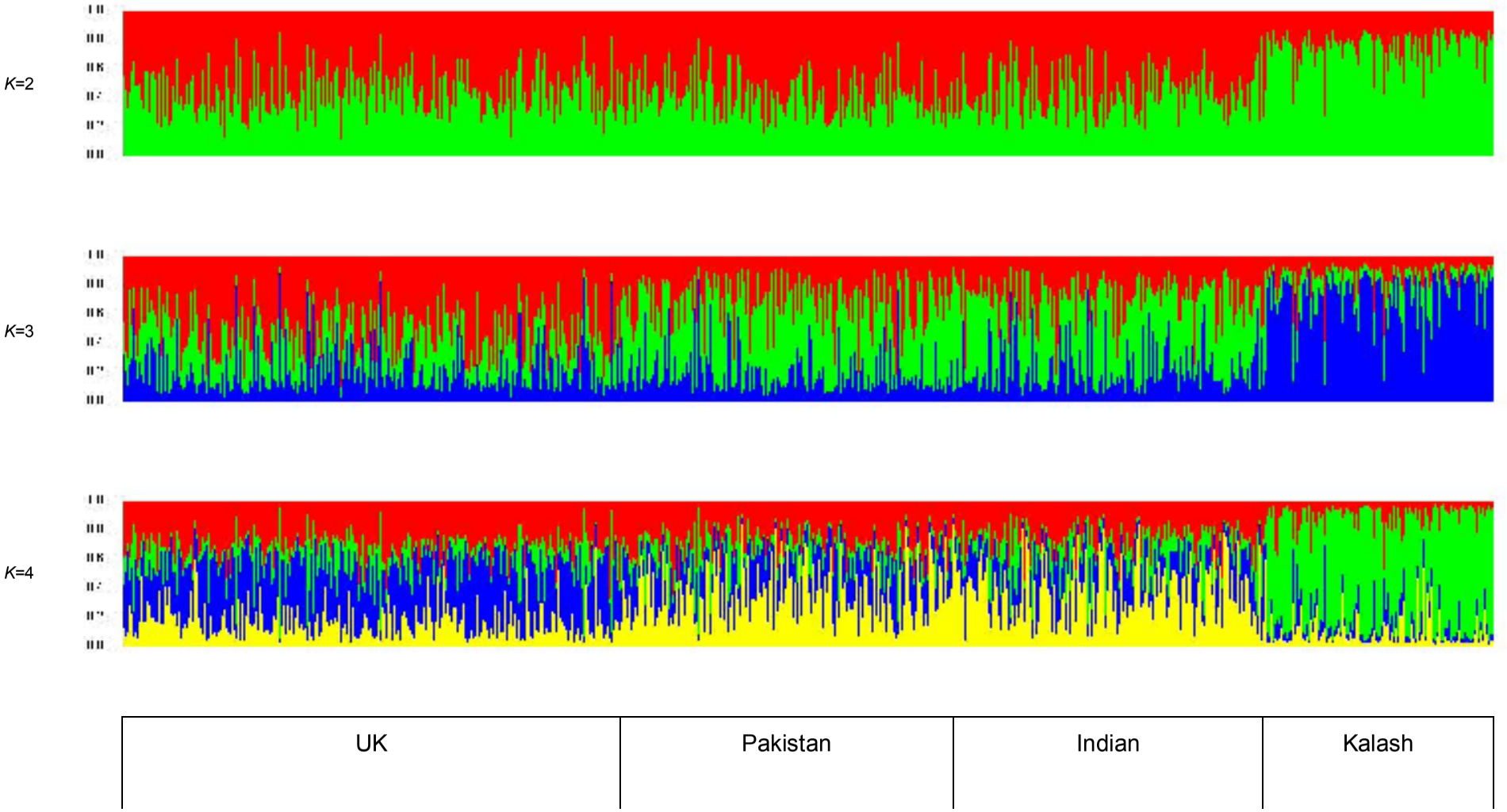
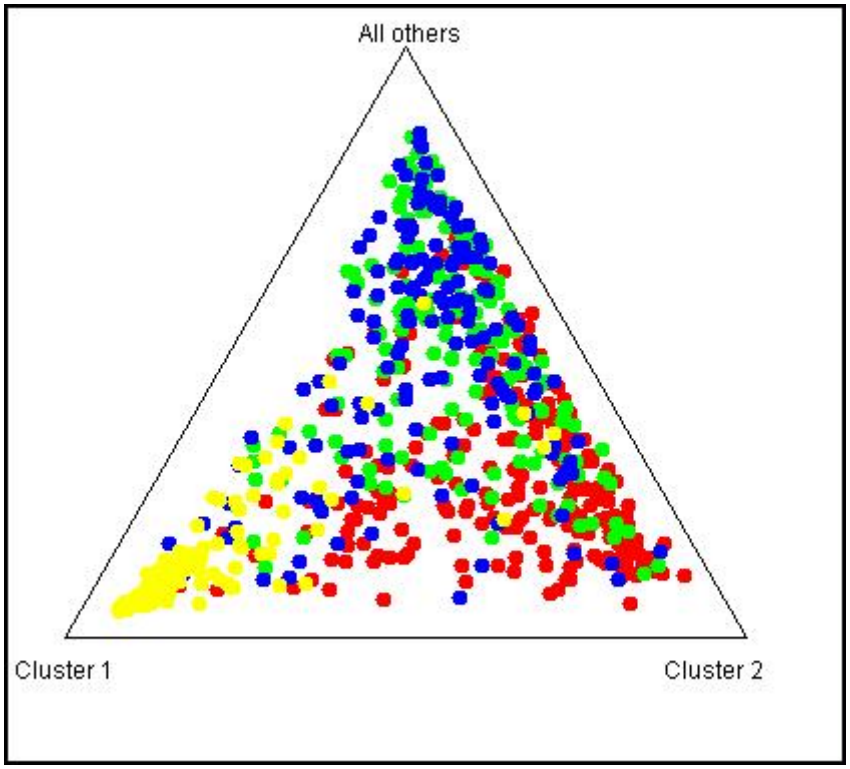


Figure 6.2: Cluster plot to show distribution of samples between clusters at *K* = 3



Each red dot represents a UK sample; green dots Indian; blue dots Pakistani and yellow dots Kalash.

6.3.2 Analysis of Molecular Variance

Analysis of Molecular Variance (AMOVA) provides an estimation of population structure at different levels of hierarchy. It considers differences among groups, among populations present within groups and among individuals within populations (Excoffier, *et al.*, 1992). AMOVA analysis is capable of testing the partitioning of genetic variation within and among pre-defined population groups.

To perform the AMOVA analysis, a genetic structure has to be established which places populations into groups which can then be tested (Excoffier, *et al.*, 2005). Here, four different population scenarios were pre-defined and are shown below in Figure 6.3. An ellipse that surrounds two or more populations indicates a group. A population by itself can also be considered a group. As the Pakistani samples comprise individuals from the Punjabi, Pushtoon and Sindhi regions of Pakistan, a test was also performed to assess whether there were any significant differences between each of these three subgroups and the remaining three main populations (Figure 6.3 – Test 2).

Table 6.2 shows the results of the AMOVA analyses for the various grouping scenarios. As expected, genetic variability among individuals within populations was greatest, accounting for over 95 % of variation in each test and has not been shown in the table.

The results show that variation amongst groups was greatest when the populations were placed into the most logical groups: Indian and Pakistani data together, the UK and Kalash kept as separate entities. When the Pakistani samples were broken down into the three regions they were collected from, this highlighted a slightly greater variation among individuals within each population. This may be because when combined, the subtle differences between the three regions are more difficult to realise and an apparent homogeneity exists. By separating the populations prior to analysis, in effect showing *a priori* knowledge, these differences are easier to observe.

Alternatively, the sampling of the groups within the Pakistani dataset is having an effect on the apparent variation among the individuals within each group. Sample sizes of caste groups in Pakistan ranged from 100 to 200 (Table 3.1) and this may not be sufficient for the results of AMOVA to accurately report differentiation between the three sampled regions of Pakistani as truly representative (Fitzpatrick, 2009).

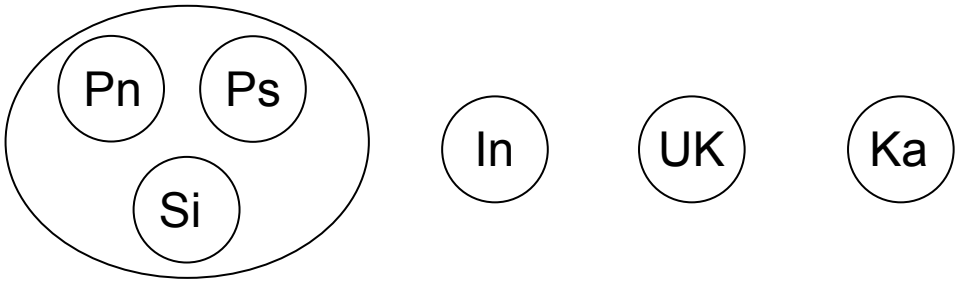
The negative figure for variance ‘among populations within groups’ in Test 2 suggests a lack of population substructure and can effectively be treated as zero. Nonetheless, it provides some reassurance that the Pakistani database is representative of some of the main population groups within the country.

Figure 6.3: Configuration of groups for AMOVA analysis

Test 1)

In = Indian, Pa = Pakistani, Ka = Kalash

Test 2)



Pn = Punjabi, Ps = Pushtoon, Si = Sindhi

Test 3)

Test 4)

Table 6.2: AMOVA results for each scenario tested (significant P-values in bold)

Test 1): Indian and Pakistani in one group, UK and Kalash separate

Hierarchy	d.f.	Sum of squares	Variance	% variance	P-value
Among groups	2	74.481	0.07236	1.74	0.15934 ± 0.01003
Among populations within groups	1	6.071	0.00600	0.14	**0.00000** ± 0.00000
Among individuals within populations	690	2839.293	0.03765	0.91	**0.04790** ± 0.00522

Test 2): Punjabi, Pushtoon and Sindhi in one group, UK, Indian and Kalash separate

Hierarchy	d.f.	Sum of squares	Variance	% variance	
Among groups	3	56916.575	50.22179	1.54	0.10307 ± 0.00317
Among populations within groups	2	5730.492	-4.59904	-0.14	1.0000 ± 0.00000
Among individuals within populations	688	2293582.380	111.63945	3.42	**0.02485** ± 0.00172

Test 3): Indian and UK in one group, Pakistani and Kalash separate

Hierarchy	d.f.	Sum of squares	Variance	% variance	
Among groups	2	54.763	0.01908	0.46	0.50050 ± 0.00521
Among populations within groups	1	25.788	0.05300	1.28	**0.00000** ± 0.00000
Among individuals within populations	690	2839.293	0.03765	0.91	0.05317 ± 0.00210

Test 4): Pakistani and UK in one group, Indian and Kalash separate

Hierarchy	d.f.	Sum of squares	Variance	% variance	
Among groups	2	57.937	0.02467	0.59	0.33366 ± 0.00432
Among populations within groups	1	22.614	0.04819	1.16	**0.00000** ± 0.00000
Among individuals within populations	690	2839.293	0.03765	0.91	0.05990 ± 0.00237

6.4 Kalash and Combined Databases

6.4.1 Database Selection

In the unlikely event that a population database most akin to the defendant is unavailable for estimation of a match probability compared to a profile obtained from a crime stain, more generic databases have to be used. These are most likely to comprise data from the most prevalent ethnic groups within the country. The rationale for this is that it is quite reasonable to assume that the perpetrator of a crime is likely to fall within one of the main populations sampled for databases held by forensic service providers.

Potential pitfalls of such databases become apparent when someone deposits a crime stain at a scene but it is not obvious which dataset would be best placed to provide a fair estimation of match probability. There may be a concern that the match probability cannot be estimated because it is based on inaccurate data or the information gathered about the perpetrator's ancestry is inaccurate. Therefore, consideration is required when determining whether one database would be better suited than another to avoid overstating (or unduly understating) the weight of any DNA evidence.

6.4.1.1 Corrections for Isolated Populations

The likelihood of someone from the Kalash region of Pakistan committing an offence within the UK is small, but not impossible. What may be more problematic is where a member of the Kalash population is linked to an offence in a neighbouring region of Pakistan – it may not be evident which database to use for match probability calculations. A Kalash dataset would be the correct one to use but an appropriately corrected 'general' Pakistani database could suffice, or perhaps be more conservative. Forensic service providers do not hold information on every population

in the world so certain assumptions and allowances have to be made to compensate for this. Balding and Nichols (1994) suggest that applying a correction factor of F_{ST} = 5 % is sufficient to allow for most isolated, genetically distinct populations. However, they note that this may need to be increased for particularly small populations where consanguinity is commonplace. The Kalash, with a population of approximately 3,000 people, could be considered to fit these criteria. Foreman and Lambert (2000) make reference to studies showing that communities on the Indian subcontinent typically exhibit F_{ST} measures ranging from 2.4 – 3.3 %; the highest being 3.7 % in an unnamed community in Southern India where uncle-niece marriages were common.

A correction of F_{ST} = 5 % is considered to be extremely generous and is used by some forensic service providers today for use with Afro-Caribbean and Indo-Pakistani match probability estimations. The counter-argument of taking this approach is that the match probability may be vastly underestimated as a consequence of using a high correction factor even if calculated with a general database. This would be favourable to the defendant but an issue may arise when a database is used based on a population in stark contrast to that of the defendant and a match probability of greater than one in one billion was estimated – the generally accepted 'ceiling' figure used in UK courts. This is used as a 'fair and reasonable' assessment of the value of the evidence and avoids possible confusion when attempting to convey minute match probabilities (Foreman & Evett, 2001). This may happen when the profile being used to estimate match probability is not complete, say, only 16 out of 20 alleles (for SGM Plus®) are present. In addition, the level of correction and the database used can have a profound effect on whether a match probability meets the 'less than one in one billion' criterion used in UK courts to denote the rarity of the match observed.

6.4.2 Kalash Match Probabilities based on the Combined Database

As with the UK, Indian and Pakistani populations, the match probabilities of the Kalash samples were estimated based on the South Asian combined database encompassing Pakistani, Indian and Bangladeshi data. The same correction factors were applied to the Kalash as per those tested previously, including the F_{ST} = 5 % correction currently used as the upper limit for match probability estimates in the UK.

Graph 6.1: Kalash samples analysed against a cognate and combined South Asian database

a) No correction applied to either the cognate or combined databases.

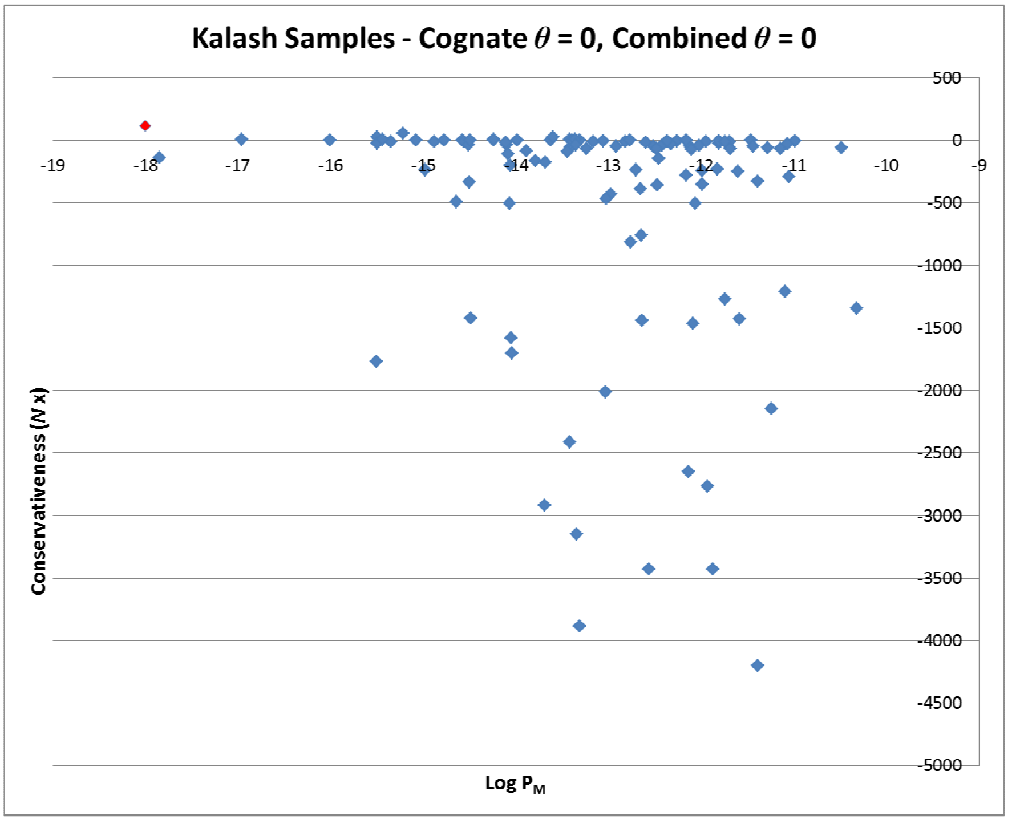
Percent > 0 = 10 %

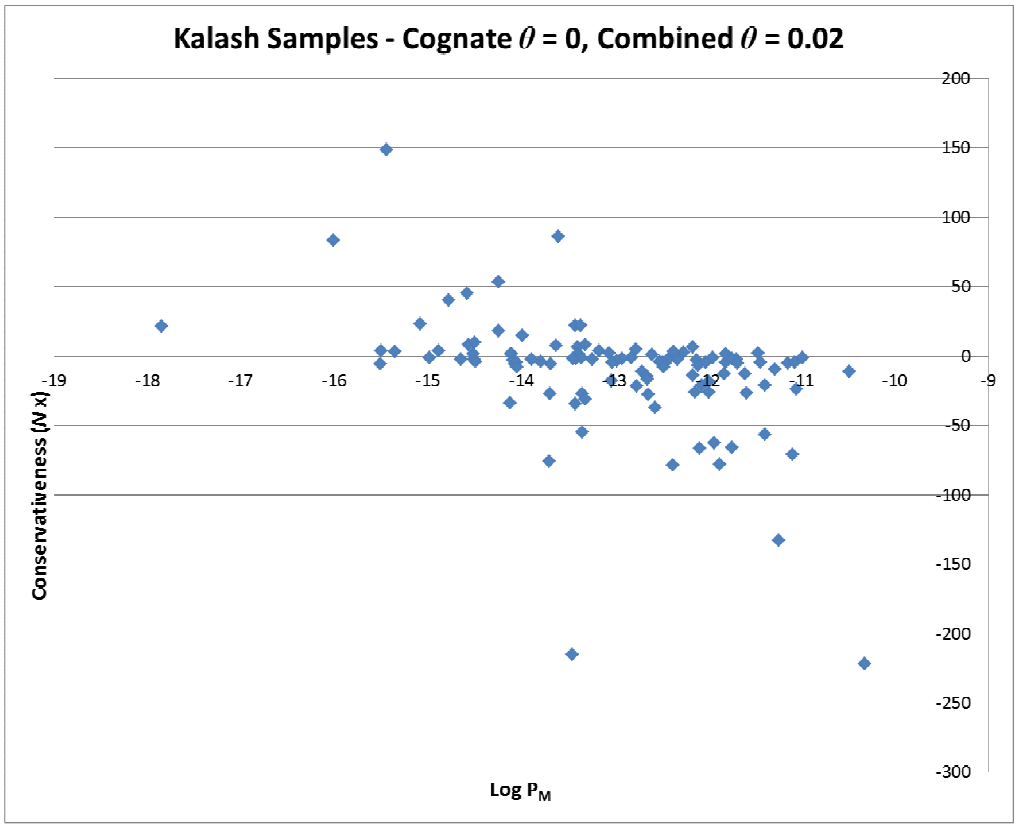
As expected from this genetically isolated population, only one in 10 samples appear more conservative in the combined South Asian database. One, (K089 – highlighted in red), shows the greatest affinity to the combined database as nearly 120 times more conservative than when compared to its cognate database. Another (not shown but referred to as sample K126), reports considerable affinity to its cognate database being over 228,000 more conservative than when compared with the combined database.

b) F_{ST} value of 2 % applied to the combined database.

Percent > 0 = 33 %

The axis have been adjusted to better visualise the dispersal of the samples but this has meant sample K089 is not shown because it reported a match probability over 4,700 times greater in the combined database at this level of correction. Sample K126 now shows a reduced level of conservativeness to its cognate database, but still over 1,100 times more than the combined database.


Kalash Samples - Cognate θ = 0, Combined θ = 0
500
0
-500
-1000
-1500
-2000
-2500
-3000
-3500
-4000
-4500
-5000
-19
-18
-17
-16
-15
-14
-13
-12
-11
-10
-9
Conservativeness (N x)
Log P_M


Kalash Samples - Cognate θ = 0, Combined θ = 0.02
200
150
100
50
0
-50
-100
-150
-200
-250
-300
-19
-18
-17
-16
-15
-14
-13
-12
-11
-10
-9
Conservativeness (N x)
Log P_M

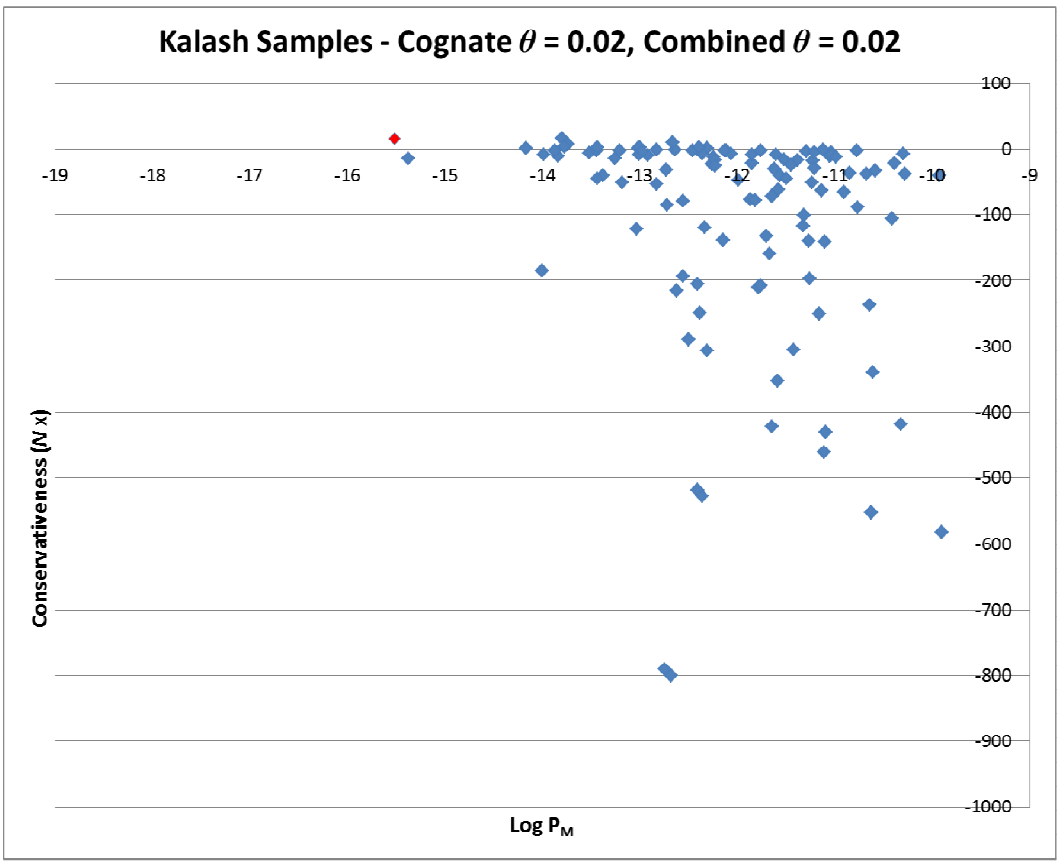
c) F_{ST} value of 2 % applied to the combined database and 2 % applied to the cognate database.

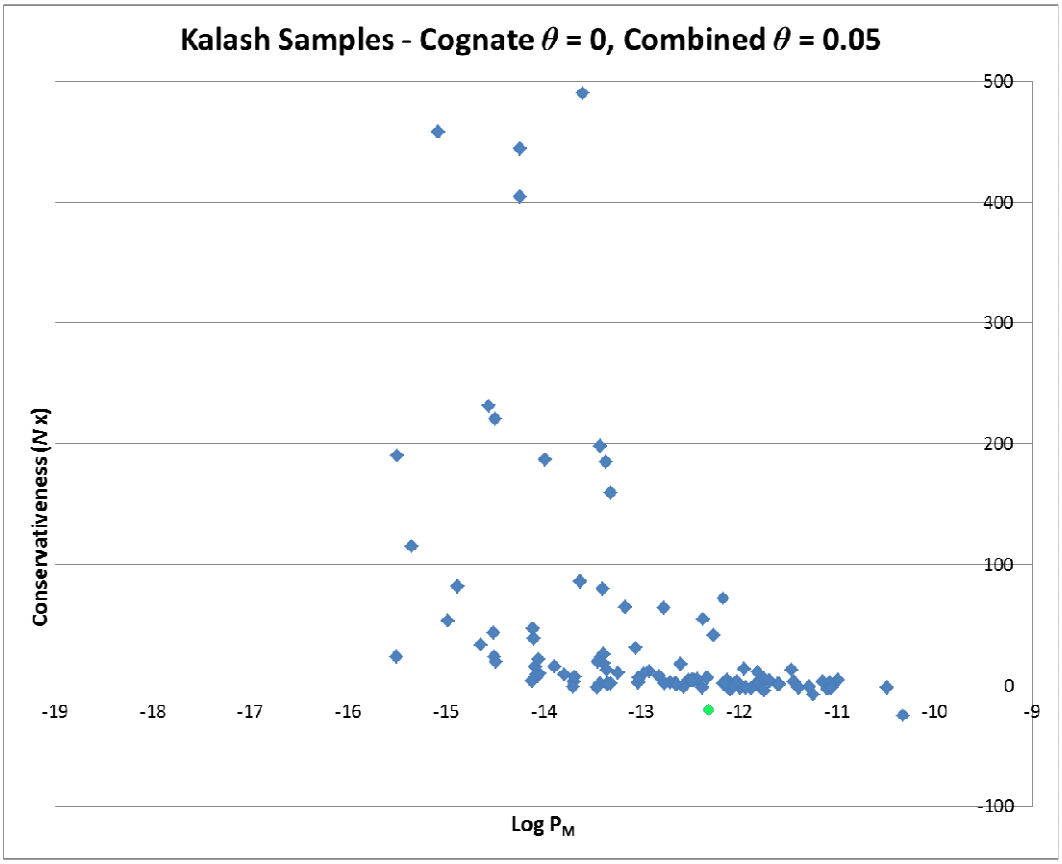
Percent > 0 = 10 %

d) F_{ST} value of 5 % applied to the combined database and no correction to the cognate database.

Percent > 0 = 82 %

Even at this high level of correction, nearly a fifth of Kalash samples return more conservative match probabilities in their cognate database. As before, with correction only applied to the combined database, the sample previously highlighted in red reports a match probability nearly 119,000 times more conservative in the combined database in this scenario. The sample referred to as K126 can now be seen (in green) and would still report a more conservative match probability when compared to its cognate database; though much less so than before at nearly 20 times more conservative.


Kalash Samples - Cognate θ = 0.02, Combined θ = 0.02
Conservativeness (N x)
Log P_M
100
0
-100
-200
-300
-400
-500
-600
-700
-800
-900
-1000
-19
-18
-17
-16
-15
-14
-13
-12
-11
-10
-9


Kalash Samples - Cognate θ = 0, Combined θ = 0.05
Conservativeness (N x)
Log P_M
500
400
300
200
100
0
-100
-19
-18
-17
-16
-15
-14
-13
-12
-11
-10
-9

Graph 6.1 shows the effect of applying no F_{ST} correction, the blanket 2 % previously used by the FSS for all populations and 5 %. The major difference in this scenario is the increase required in the y-axis to be able to plot most samples. The most extreme example of non-conservativeness was seen in scenario 6.1a) where no correction was applied to either the cognate or combined database. Sample K126 returned a cognate / combined ratio of approximately 228,000: a match probability five orders of magnitude greater in the cognate database than the combined one. Although still reporting a match probability in the order of less than one in one billion ('ceiling principle' [National Research Council, 1992]), it highlights that caution should be applied when considering to use a particular database for a match probability estimate – particularly when concerned with an isolated population. Studies by Curran, *et al.*, (2003) showed that application of the product rule can introduce a mild bias towards the prosecution when used on subdivided populations. Although present, it was stressed that it should not be overstated and that match probability estimations were usually only affected by a factor of 10. Even with the apparent bias, match probabilities were still in the order of less than one in one billion for a full SGM Plus® profile.

Where F_{ST} = 2 % is applied to the combined database only, as in scenario 6.1b), one sample, K089, has reported an approximate 4,750 times more conservative match probability in the combined database than the cognate database; with the second most conservative sample, K122, at approximately 550. Interestingly, looking at scenario 6.1a), the most conservative figure is provided by sample K089 but this time at approximately 120 and as many of the samples lie around the baseline; no one sample stands out as already being highly conservative in the combined database. When the correction is applied in scenario 6.1b), there are still many samples situated around the base line but it has had a much greater effect on sample K089. The two extremes seen in samples K089 and K126 suggest the allele frequencies of these samples were overall less concordant with those making up the rest of the

Kalash database. Applying a correction factor has amplified these differences, changing each sample's affinity to a particular database.

In scenario 6.1c), where the same correction of F_{ST} = 2 % is applied to both the cognate and combined datasets, the extreme conservativeness exhibited by sample K126 reduces to approximately 7,250 times more conservative in the cognate database. Table 5.4 shows a F_{ST} correction of approximately 12 % would be required to make all samples more conservative in the combined South Asian database; without this anomalous sample, that figure may be reduced but incorporating all samples gives the most conservative estimate of the required correction.

With a correction of F_{ST} = 5 %, 82 % of Kalash samples showed match probabilities that appear more conservative when estimated using the combined South Asian database. Even with this extremely generous allowance for shared ancestry, there is still close to 20 % of Kalash samples which appear less conservative when their match probabilities are estimated using the combined South Asian database. This is in contrast with the Indian and Pakistani samples where 99 % and 100 % of all samples, respectively, reported more conservative profile frequencies in the combined database with a correction of F_{ST} = 3 % (Table 5.3).

Using the standard of F_{ST} = 2 %, only 33 % of the Kalash samples gave a higher match probability when estimated using the South Asian Database compared to 80 % of UK samples (Graph 5.2b). As shown previously in Table 5.4, a correction of F_{ST} = 6.4 % would be required to make all the UK samples appear more conservative in the combined database.

As described by Balding and Nichols (1994) a F_{ST} correction of 5 % should be sufficient for genetically differentiated populations. They go on to say that a higher correction may be required for more isolated populations where consanguinity is commonplace. However, as discussed by Foreman *et al.,* (1998), from a criminal proceedings perspective, a match probability should ideally be calculated from a

database made up of those who had opportunity to commit the offence in question. A reasonable defence argument would be that the databases used in the UK for example, do not reflect their client's 'target population', i.e. one composed of individuals from the same area. Of course, the suspect may not always be willing to provide their ancestral history, hence why a correction factor of 0.05 for populations such as the Kalash is deemed reasonable; indeed in this study, it would favour 82 % of the individuals in the dataset to have their match probability estimated against the combined database. However, in terms of using STR profiles as a utility to estimating geographic origin, this can cause issues as will be discussed later with the potential for an investigation to be misled or focussed in the wrong area.

To further evaluate conservativeness in combined databases compared with cognate sets, another method employed by Gill *et al*., (2003) was utilised. Every population from the South Asian datasets, including the Bangladeshi data (Alshamali *et al.*, 2005), were pooled in a variety of ways to produce new combined databases. As in section 5.1.4, *d* was calculated for all samples and an average calculated ($\overline{d}$). The F_{ST} correction was then altered in the combined database only to give a new estimate of $\overline{d}$.

If *d* is positive, a sample's calculated match probability is more conservative in its cognate database than in the combined one; hence in criminal proceedings, the use of a general database would be considered detrimental rather than if a database more akin to their client's genetic ancestry was used. In addition, the frequency of samples where *d* <0, >0, >1, etc. are shown to highlight more extreme examples of profile frequencies that require a greater correction factor to be applied before appearing more conservative in a combined database.

Four combined databases were created:

1) Indian with Pakistani
2) Indian, Pakistani and Bangladeshi
3) Indian, Pakistani and Kalash
4) Indian, Pakistani, Bangladeshi and Kalash

Table 6.3 shows the results of varying F_{ST} corrections on the above combined database scenarios. The first point to note with all scenarios is that at the first correction level of F_{ST} = 0.3 %, $\bar{d}$ falls below zero indicating that most samples will now appear more conservative in the combined databases rather than their own. The standard deviation is high across all scenarios but this can be explained by the relatively small databases used in this study; these values are in concordance with those reported by Gill *et al.*, (2003) who populations of nearly 6,000 samples would be required to bring the standard deviation to around 0.1.

Overall, the combined databases which do not include the Kalash samples show a higher level of conservatism at each correction factor than those which do include them. This is also supported by the observation that a higher proportion of samples show $d < 0$ at each correction factor when the Kalash data are omitted suggesting that the combined databases would give a higher match probability.

Graph 6.2 summarises $\bar{d}$ as calculated for each scenario of the combined databases. An immediate trend can be observed where the two databases which include the Kalash population are distinct from the two databases which do not. If the Kalash are not included in the combined database, the match probabilities of samples estimated using it are more conservative at a given F_{ST} value than when the Kalash are included. As expected, the Kalash data skew what may be considered to be nominal levels of inbreeding but suggests that the effect is not one which will

significantly affect match probability estimations assuming sufficient allowances are made.

Table 6.3: The effect of varying F_{ST} levels on combined databases

1) Combined database of Indian and Pakistani samples. *d* is calculated for each sample of the Indian and Pakistani populations at each F_{ST} correction and an average calculated.

2) Results for a combined database comprising Indian, Pakistani and Bangladeshi samples.

F_{ST}	d (av)	SD	P(d < 0)	P(d > 0)	P(d > 1)	P(d > 2)	P(d > 3)
0	0.1000	0.4535	0.4322	0.4732	0.0473	0.0063	0.0000
0.3	-0.1613	0.4271	0.7224	0.2145	0.0221	0.0000	0.0000
0.5	-0.3118	0.4320	0.7918	0.1546	0.0126	0.0000	0.0000
1.0	-0.6453	0.4663	0.8991	0.0568	0.0032	0.0000	0.0000
2.0	-1.2064	0.5507	0.9401	0.0189	0.0000	0.0000	0.0000
3.0	-1.6828	0.6265	0.9558	0.0032	0.0000	0.0000	0.0000

F_{ST}	d (av)	SD	P(d < 0)	P(d > 0)	P(d > 1)	P(d > 2)	P(d > 3)
0	0.1047	0.4941	0.4159	0.4557	0.0489	0.0092	0.0000
0.3	-0.1561	0.4573	0.6850	0.2202	0.0214	0.0031	0.0000
0.5	-0.3064	0.4573	0.7706	0.1437	0.0153	0.0000	0.0000
1.0	-0.6396	0.4824	0.8777	0.0459	0.0061	0.0000	0.0000
2.0	-1.2012	0.5569	0.9144	0.0153	0.0000	0.0000	0.0000
3.0	-1.6782	0.6281	0.9266	0.0031	0.0000	0.0000	0.0000

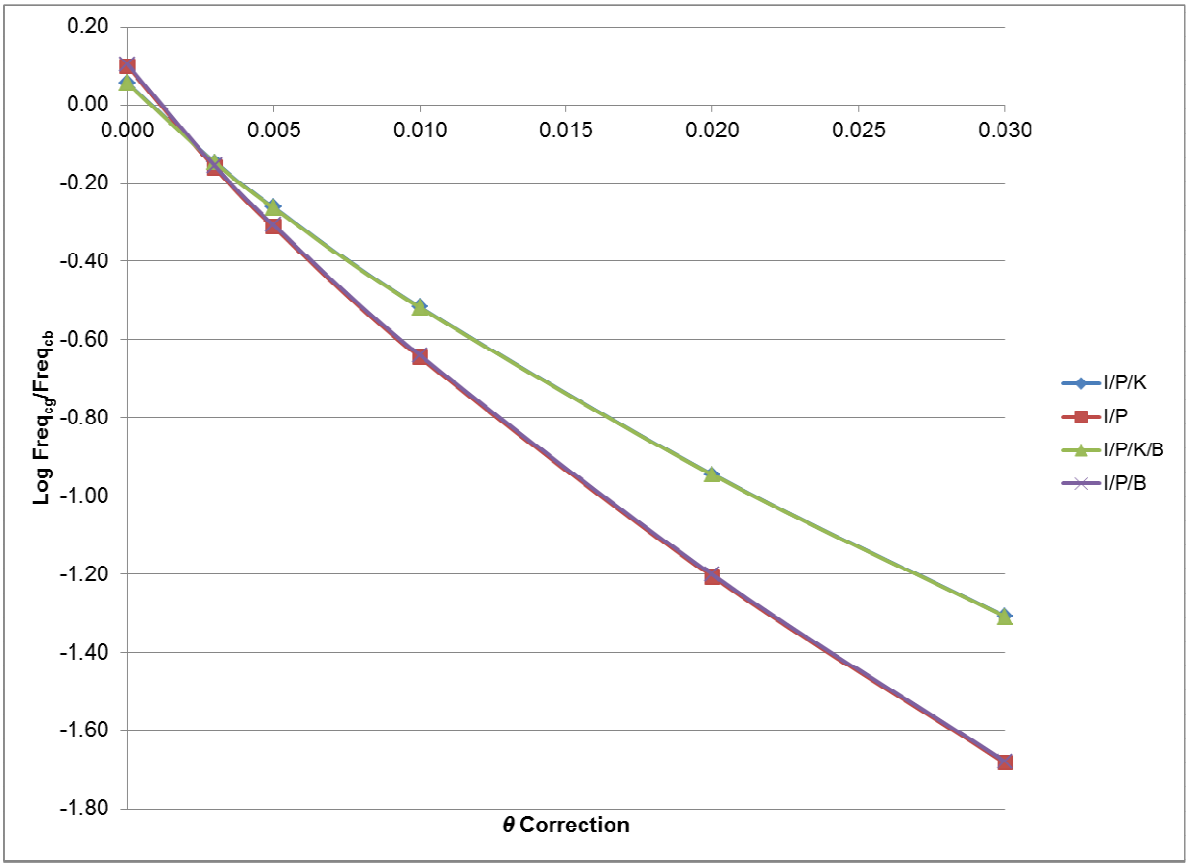
3) Results for a combined database which includes Indian, Pakistani and Kalash samples.

4) Results for a combined database including all South Asian populations used in this study.

F_{ST}	d (av)	SD	P(d < 0)	P(d > 0)	P(d > 1)	P(d > 2)	P(d > 3)
0	0.0573	0.5745	0.2483	0.3968	0.0510	0.0093	0.0000
0.3	-0.1458	0.5397	0.3898	0.2877	0.0255	0.0023	0.0000
0.5	-0.2615	0.5380	0.4687	0.2158	0.0209	0.0000	0.0000
1.0	-0.5166	0.5567	0.6056	0.0928	0.0070	0.0000	0.0000
2.0	-0.9447	0.6194	0.6821	0.0232	0.0000	0.0000	0.0000
3.0	-1.3078	0.6821	0.6984	0.0070	0.0000	0.0000	0.0000

F_{ST}	d (av)	SD	P(d < 0)	P(d > 0)	P(d > 1)	P(d > 2)	P(d > 3)
0	0.0558	0.5807	0.2575	0.3944	0.0441	0.0093	0.0000
0.3	-0.1473	0.5382	0.4153	0.2645	0.0232	0.0023	0.0000
0.5	-0.2629	0.5338	0.4872	0.1949	0.0209	0.0023	0.0000
1.0	-0.5178	0.5489	0.6172	0.0812	0.0070	0.0000	0.0000
2.0	-0.9456	0.6089	0.6868	0.0186	0.0000	0.0000	0.0000
3.0	-1.3087	0.6708	0.7030	0.0023	0.0000	0.0000	0.0000

Graph 6.2: Effect of varying F_{ST} levels on average of *d* across four combined databases



I = Indian, P = Pakistani, K = Kalash, B = Bangladeshi

6.5 Effect of Substructuring on DNA Profile Frequency Estimation

In contrast to the graphs shown in section 5.2, where there is no visible distinction between the Indian and Pakistani samples, Graphs 6.3 and 6.4 show clear differences when the Kalash are compared to both the UK and Indian samples. Due to the similarities between the Indian and Pakistani samples, the comparison between Kalash and Pakistani samples has not been shown.

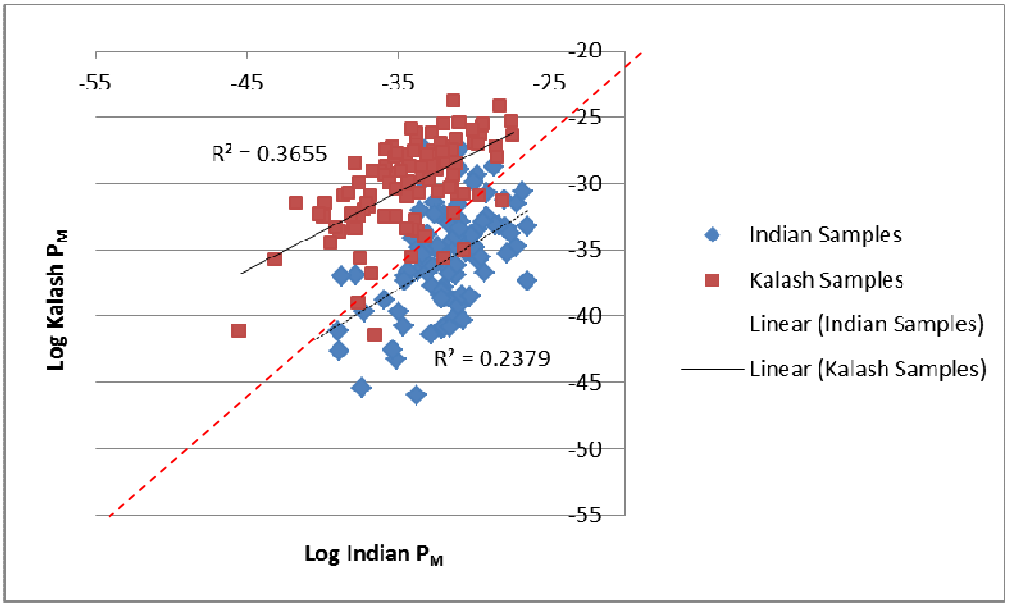
What is also evident is that as the F_{ST} correction applied increases, although the match probabilities increase, they are still below the one in one billion ceiling figure (10^{-9}) quoted routinely in the UK criminal justice system with regard to forensic DNA evidence.

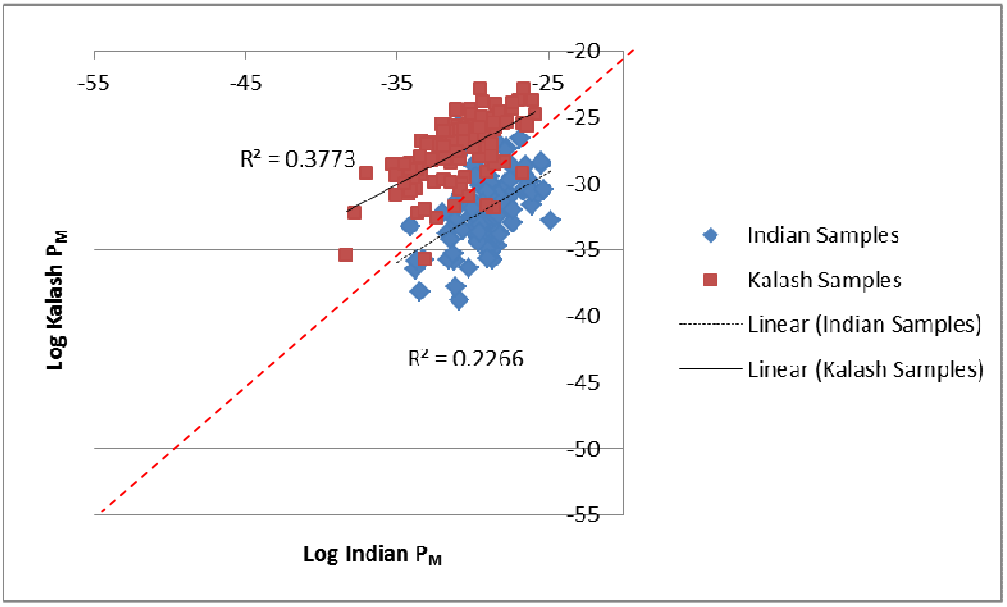
Graph 6.3: Effect of substructuring on the Kalash and Indian populations with varying F_{ST} values. Profile frequencies for each sample are calculated using both the Kalash and Indian databases

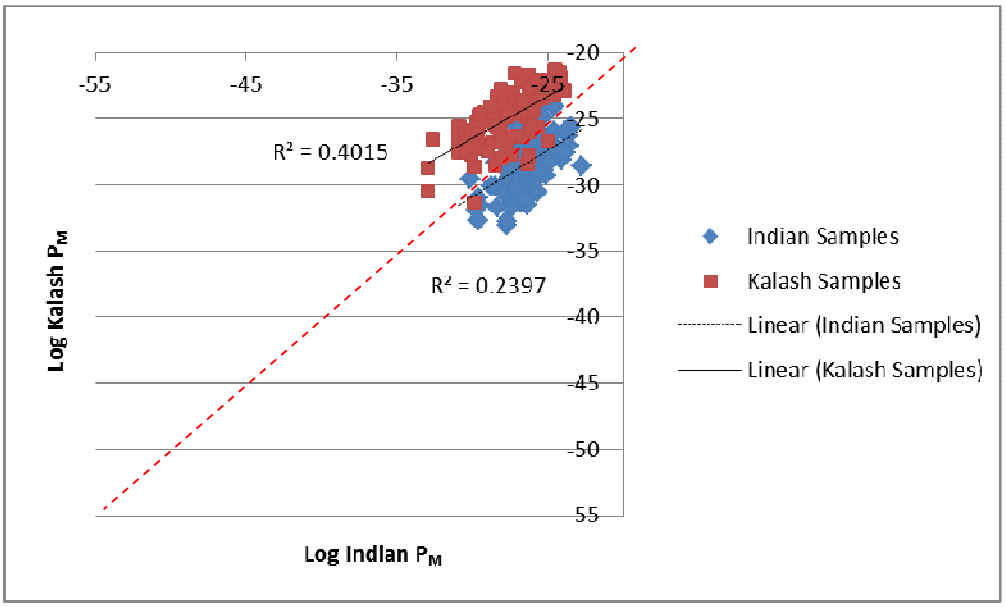
a) F_{ST} = 0 %

b) F_{ST} = 2 %

c) F_{ST} = 5 %


-55
-45
-35
-25
-20
-25
-30
-35
-40
-45
-50
-55
R² = 0.3655
R² = 0.2379
Log Kalash P_M
Log Indian P_M
Indian Samples
Kalash Samples
Linear (Indian Samples)
Linear (Kalash Samples)


-55
-45
-35
-25
-20
-25
-30
-35
-40
-45
-50
-55
R² = 0.3773
R² = 0.2266
Log Kalash P_M
Log Indian P_M
Indian Samples
Kalash Samples
Linear (Indian Samples)
Linear (Kalash Samples)

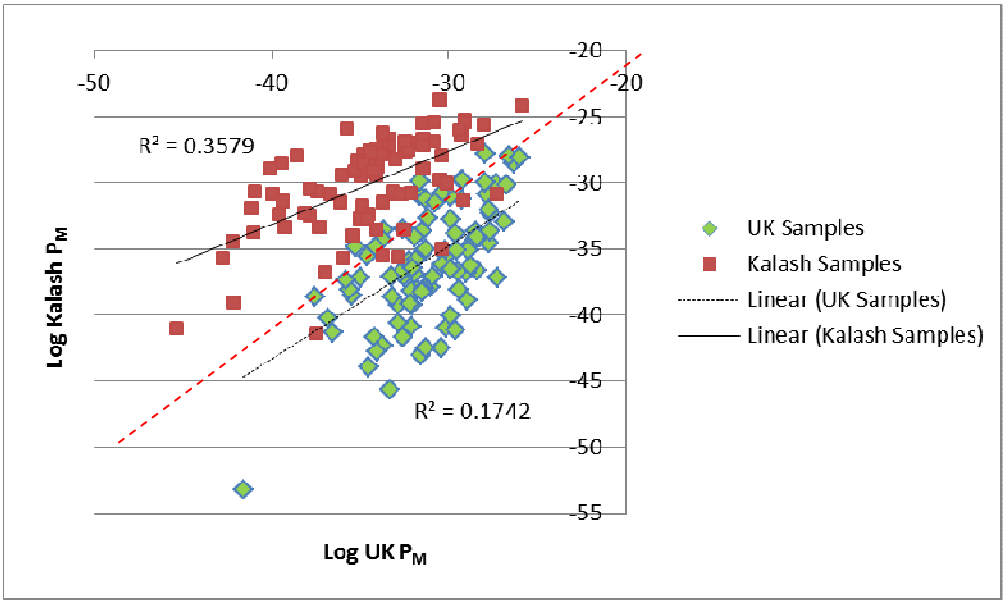

-55
-45
-35
-25
-20
-25
-30
-35
-40
-45
-50
55
R² = 0.4015
R² = 0.2397
Log Kalash P_M
Log Indian P_M
Indian Samples
Kalash Samples
Linear (Indian Samples)
Linear (Kalash Samples)

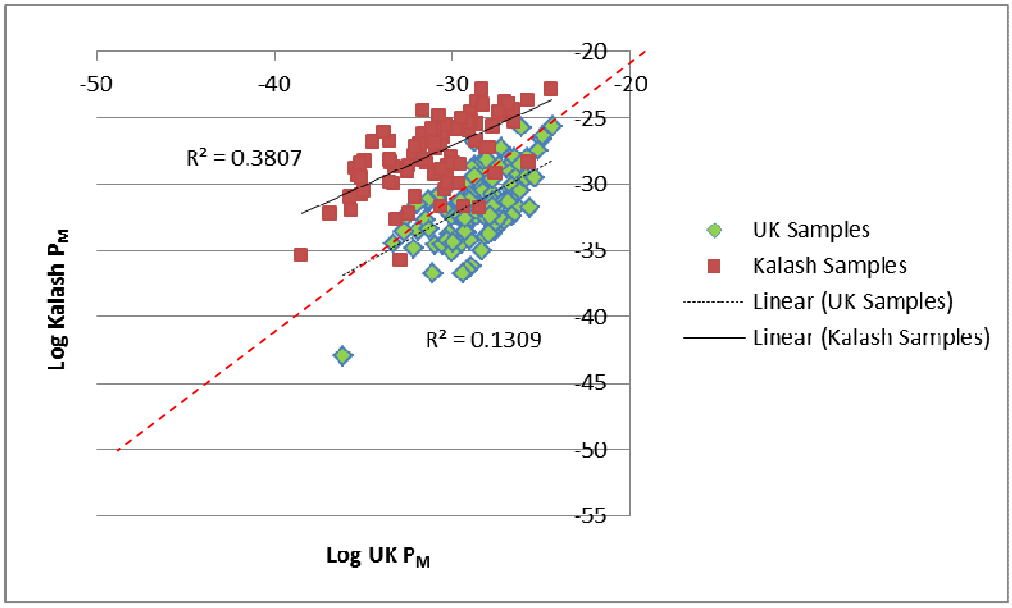
Graph 6.4: Effect of substructuring on the UK and Kalash populations with varying F_{ST} values. Profile frequencies for each sample are calculated using both the Kalash and UK databases

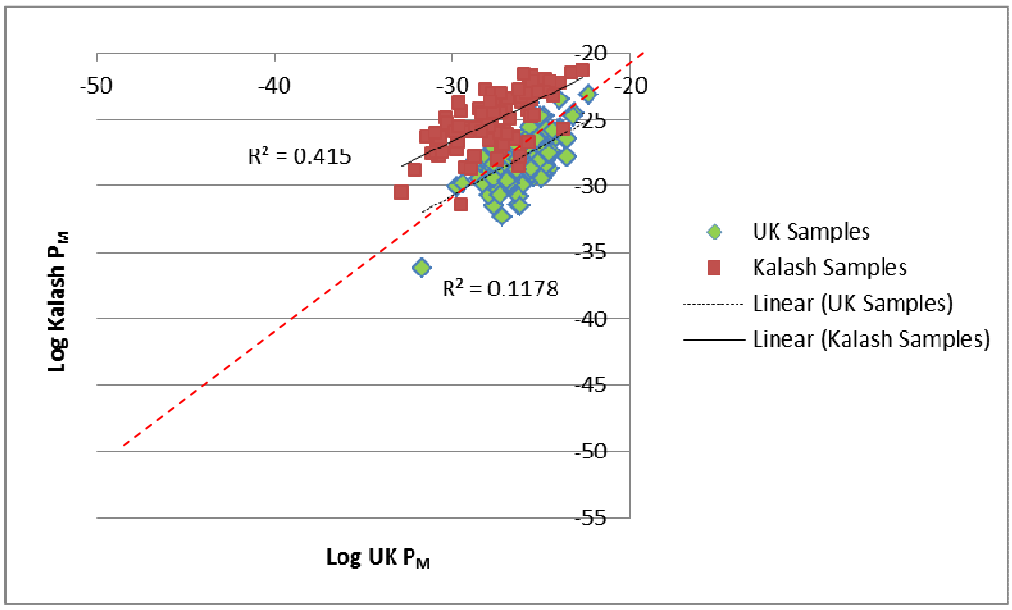
a) F_{ST} = 0 %

b) F_{ST} = 2 %

c) F_{ST} = 5 %


-50
-40
-30
-20
-20
-25
-30
-35
-40
-45
-50
-55
R² = 0.3579
R² = 0.1742
Log Kalash P_M
Log UK P_M
UK Samples
Kalash Samples
Linear (UK Samples)
Linear (Kalash Samples)


-50
-40
-30
-20
-20
-25
-30
-35
-40
-45
-50
-55
R² = 0.3807
R² = 0.1309
Log Kalash P_M
Log UK P_M
UK Samples
Kalash Samples
Linear (UK Samples)
Linear (Kalash Samples)


-50
-40
-30
-20
-20
-25
-30
-35
-40
-45
-50
-55
R² = 0.415
R² = 0.1178
Log Kalash P_M
Log UK P_M
UK Samples
Kalash Samples
Linear (UK Samples)
Linear (Kalash Samples)

Comparing the UK and Indian populations (Graph 5.3), although subtle, there is evidence of substructuring. As the value of F_{ST} increase, the samples cluster closer together as the natural log of the frequencies also increases. This is a result of the sample becoming seemingly more common in a population as the co-ancestry is taken into consideration. With the more isolated Kalash population, the difference is more apparent when compared to the UK population (Graph 6.4). Even when F_{ST} = 2 % and 5 %, a clear separation can be seen.

6.5.1 Geographical Assignment of Individuals

As in section 5.2.1 there are again instances where several samples in each population show a greater affinity to the population to which they are being compared to rather than their own. However, with the inclusion of the Kalash, a clearer distinction is visible. This will have implications when trying to estimate the geographical assignment of samples to the correct population, without prior knowledge of their origin, based on the profile frequency alone.

To test this, each sample was assigned to a population based on which database reported the highest frequency. Given the relatively low number of loci included in the AmpFℓSTR® SGM Plus® kit, it may not be expected to be sufficient to successfully identify the population of unknown samples. However, given the level of some of the pairwise F_{ST} values (Table 5.1) and the comparison of sample profiles against non-cognate databases (Graphs 5.3, 5.4, 6.3 and 6.4), there is evidence of population differentiation, even if not detectable by the Bayesian clustering method utilised in STRUCTURE.

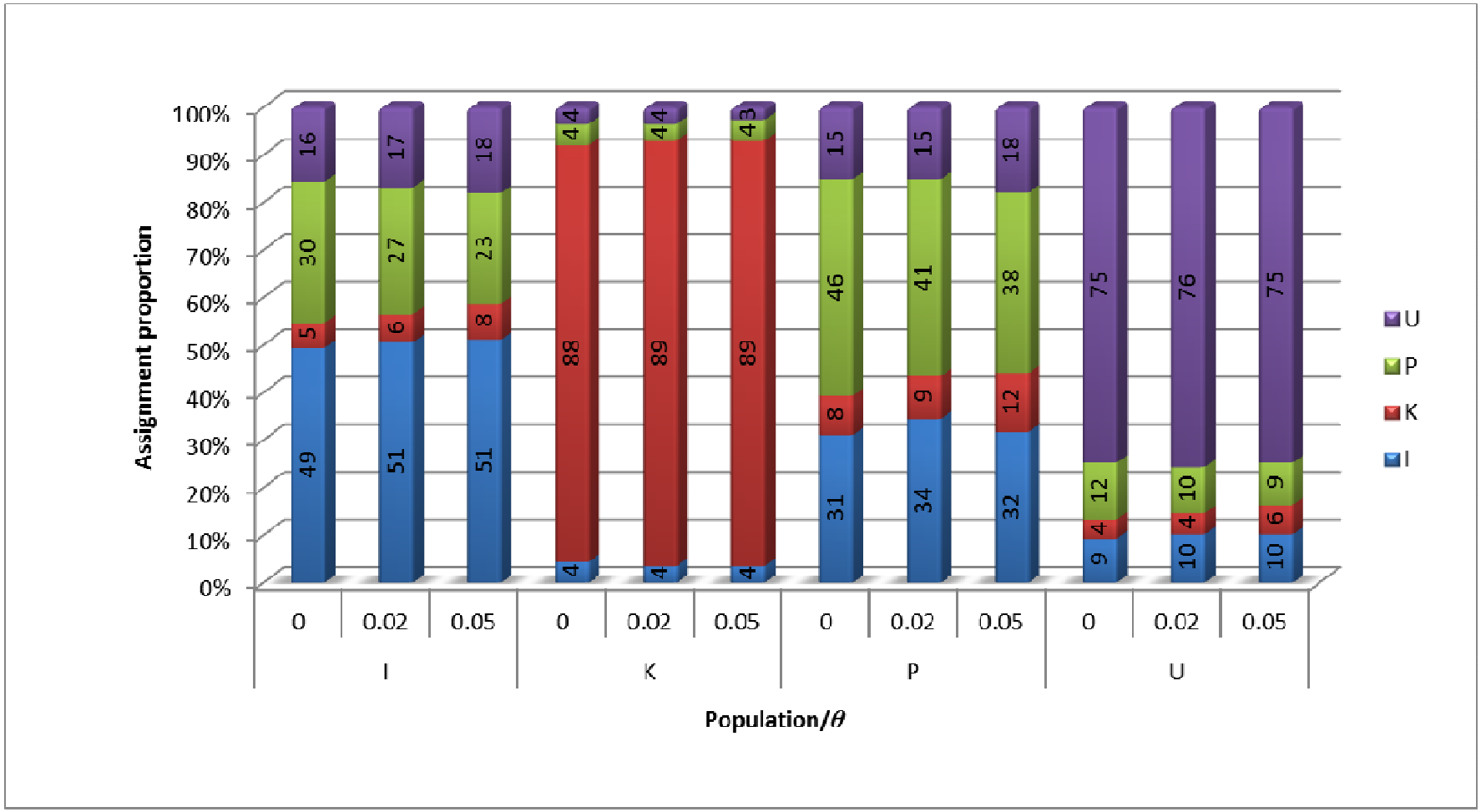
Taking substructuring into consideration also, the assignment test was repeated for three F_{ST} values (0 %, 2 % and 5 %), to observe whether the level of correction affects the proportion of individuals deemed to originate from one population rather than another when considering profile frequency alone (Table 6.4 and Graph 6.5).

Table 6.4: Geographical assignment of each sample to a population based on profile frequency

		Database																					
		Indian (F_{ST})						Kalash (F_{ST})						Pakistani (F_{ST})						UK (F_{ST})					
		0 %		2 %		5 %		0 %		2 %		5 %		0 %		2 %		5 %		0 %		2 %		5 %	
		n	%	n	%	n	%	n	%	n	%	n	%	n	%	n	%	n	%	n	%	n	%	n	%
Sample	Indian	85	49	87	51	88	51	5	4	4	4	4	4	45	31	50	34	46	32	21	9	23	10	23	10
	Kalash	9	5	10	6	13	8	100	88	102	89	102	89	12	8	13	9	18	12	9	4	10	4	14	6
	Pakistani	51	30	46	27	40	23	5	4	4	4	5	4	66	46	60	41	55	38	27	12	22	10	20	9
	UK	27	16	29	17	31	18	4	4	4	4	3	3	22	15	22	15	26	18	169	75	171	76	169	75

n = Number of individuals assigned to a particular population.

Graph 6.5: Proportion of samples assigned to each database at varying F_{ST} values



U = UK, P = Pakistani, K = Kalash, I = Indian

Graph 6.5 shows that varying F_{ST} values have little effect on sample assignment. As expected, almost 90 % of Kalash samples were assigned correctly at all F_{ST} values due to their greater genetic differentiation compared to the other populations.

Looking at the UK database, approximately 75 % of samples were correctly assigned, with samples from the Pakistani dataset forming the largest proportion of incorrect allocations; between 9 % and 12 %. In addition, between 15 % and 18 % of UK samples showed a higher affinity for the Indian or Pakistani databases; and between 3 % and 4 % for the Kalash database.

As the data have shown, how useful this tool might be in police investigations depends entirely on the populations the 'unknown' sample is compared to. The UK samples were assigned correctly at least three out of four times, with approximately one in 10 Indian and Pakistani samples being incorrectly assigned as UK. Although the pairwise F_{ST} values between the UK population and both the Indian and Pakistani population is just over 1 % (Table 5.1), there is potential to provide limited intelligence information as to a possible geographic origin of an offender.

Given the little differentiation seen between the samples of the Indian and Pakistani databases (Table 5.1), it is not unexpected that there is little difference in the proportions of Indian and Pakistani samples wrongly affiliated with the opposing population's database. That said, cognate samples from each database do form the highest proportion of correctly assigned samples in each, though little reliability may be placed upon such an estimation.

Graph 6.5 correlates with the results of the scatter plots with regard to the genetic distance of the populations. For example, the Pakistani and Indian populations are virtually indistinguishable (Graph 5.4); while Graph 6.5 shows that these two populations have the largest misallocation of samples in each other's database. The Kalash, showing greatest genetic differentiation with all populations, reports the highest

success rate of assigning its own samples to the correct database. Considering the geographical locale of the Kalash and other populations that comprise the combined South Asian database, based on STR data alone, it would be difficult to estimate geographical origin assuming no prior knowledge of the sample donor.

Changes in F_{ST} value had the greatest effect on Pakistani samples in the Pakistani database. At F_{ST} = 0 %, 2 % and 5 %, the correct allocation of Pakistani samples was 46 %, 41 % and 38 % respectively. This reduction in correct assignment at F_{ST} = 2 % was due mainly to more Indian samples being inadvertently assigned as Pakistani instead. At F_{ST} = 5 %, more UK and Kalash samples are incorrectly assigned to the Pakistani database though not at sufficient levels provide a realistic estimation that those samples truly originate from such a population.

6.6 Discussion

Populations such as the Kalash have the potential to confuse estimations of geographical origin of a stain donor based on STR profiling alone. As Graph 6.5 shows, although a DNA profile may be unique, profile frequencies cannot provide a reliable assessment of correct geographical assignment. Even if a profile frequency is compared to the correct population and the perpetrator in question actually comes from a subpopulation that has been tested, estimations for assignment to that population may come out as low as 38 % such as with the Pakistani population, yet be the correct one. Compared with the Kalash, where estimations for assignment may reach almost 90 %, yet for this to happen, the Kalash would have to be included in the databases being used to compare the 'unknown' sample too. Additional databases of isolated populations are being published continuously and to include them all in such an assessment is impractical.

There are instances where match probabilities of some Kalash samples appear more conservative in the UK population even if no F_{ST} correction is applied. Clearly in this study, that estimation would be wrong and therefore, from a criminal trial perspective, non-DNA evidence must be used in conjunction with DNA evidence and it is important that the weight of any DNA evidence is not overstated as to potentially mislead the jury. The uniqueness of a DNA profile cannot truly be established; uniqueness must be true or false but to establish this it would involve sampling the entire population, hence why a probability is assigned as to the likelihood of obtaining a matching DNA profile (SGM Plus® or Identifiler®, for example) in someone other than, and unrelated to a potential suspect is given. This is when non-DNA evidence plays a key role but is an area that a scientific expert should not impinge (Balding, 1999).

One of the greatest setbacks faced when trying to establish a model for accurate match probability calculations is sampling. Ideally, every population, subpopulation and any

additional hierarchal levels would need to be sampled. Although Chakraborty (1992) suggested that sampling between 100 – 150 individuals from each population may be sufficient to use for profile frequency estimations, it is difficult to establish how many populations and subpopulations there are. How these populations are sampled is also of importance; sampling individuals who all declare the same self-defined ethnicity may allow for accentuated discontinuities between populations (Serre & Pääbo, 2004). This may lead to inaccurate estimations of population assignment.

One question raised by Buckleton, *et al.*, (2006) was that even though the model devised by Balding and Nichols (1994) was brought about to deal with substructuring within the population, what if the subpopulations themselves departed from Hardy-Weinberg expectations; in effect creating sub-subpopulations? Their experiments showed that even with inbreeding within the sub-subpopulation, in over 99 % of cases, match probability estimates were still in favour of the defendant. Therefore, the Balding and Nichols model can still be seen as a conservative tool but with populations such as the Kalash, the effect of such genetically isolated populations must be taken into consideration.

The Kalash stands out as a distinct population in many of the analyses conducted here which is in concordance with results obtained by Rosenberg *et al.*, (2002 & 2005). The STRUCTURE analyses have shown that there is little differentiation between the UK, Pakistani and Indian populations yet the genetic variation with the Kalash population enable early identification of an isolate group. There are also instances where certain allele frequencies in the Kalash greatly exceed those of the other three populations such as allele 15 at the vWA locus and allele 29 at D21 (Graphs 4.2 and 4.6, respectively).

Phillips *et al.*, (2011) examined a selection of populations from the CEPH-HGDP (though not including the Kalash) and concluded that a F_{ST} correction of 10 % was

considered highly conservative based on the range of populations they had studied. It was, on average, more than three times the level of correction required to account for any population substructure. They reported a maximum pairwise difference of F_{ST} = 2 % between studied Europeans and Aboriginals from the Northern Territories of Australia, deemed the most isolated population including in their study.

The F_{ST} value of 10 % would not wholly capture all of the Kalash samples used in this study when their match probabilities were calculated against the South Asian database (Table 5.4). Yet this is the figure Phillips *et al.*, (2011) provide to be sufficiently conservative to use with the populations of the CEPH-HGDP. Two of which, they say, are the most stratified of the collection; the Karitiana (a small, Amazonian population of less than 200 individuals) and the Surui (approximately 800 people spread amongst villages between the Brazilian provinces of Mato Grosso and Rondonia). Any population sampled may have samples reporting outlying match probabilities such that the removal of one or two samples showing extreme affiliation to a cognate database may mean the F_{ST} value suggested by Phillips *et al.*, (2011) is adequate. As stated previously, there is a risk of greatly underestimating the weight of DNA evidence as well as overstating it (Gill *et al.*, 2003). A balance must be struck between what is deemed fair and reasonable to a defendant whilst also realising the true discriminatory power of DNA evidence.

6.6.1 The Assumption of Independence when using the Product Rule

Following from that, there is contention as to the use of the product rule for calculating match probabilities; however, this forms the basis of the research for this study and calculations of match probabilities performed by many forensic service providers regularly. Triggs and Buckleton (2002) highlighted that as there is no accurate way of measuring independence of a database, should the model upon which the product rule is based be used for calculating match probabilities? Knowledge about the dynamics of

the populations sampled is vital (such as migration and occurrences of consanguinity) and validation of the subsequent database is required in order to assess the use of the product rule in calculations.

Where evidence of subdivision is shown to be minimal, in that the use of the product rule will have negligible effect on the match probability estimation in a forensic context, then this may be an acceptable method for match probability estimations. However, this would technically only hold true if studies of genetic variation across whole populations had been conducted. In reality, sample databases are much smaller and represent only a fraction of a population. This is why correction factors such as F_{ST}, as described by Balding and Nichols (1994), are employed in a forensic context to account for most, if not all, of the substructure of a population. By using what are believed to be hyper-conservative allowances for inter-relatedness, F_{ST} corrections can also incorporate questions of uncertainty over sampling and independence issues which may underlie the database used for match probability estimations. The NRC report (1996) suggested a F_{ST} correction of 1 % should be applied in instances where an appropriate database is not available and the use of a general one has to be relied upon. This correction may be increased to 3 % for those believed to be from more isolated populations. Despite the possibility of understating the DNA evidence (Gill *et al.*, 2003), it could be argued that to be truly conservative, a figure should be used which exceeds the levels of subdivision recorded for any population studied.

It is populations such as the Kalash that are potentially at the greatest disadvantage where match probability estimations are made. These estimations are often given on a 'fair and reasonable' basis of the weight of the evidence and use of the product rule alone may introduce slight bias to the prosecution. Fair and reasonable will ideally be a figure which closely relates to the actual match probability: something which cannot be calculated without using a database comprising the entire relevant population, if the offender's true population is known. In a forensic context, providing a highly

conservative estimation is likely to make little difference to how the DNA evidence is interpreted in court proceedings. To a jury, the use of a 'ceiling' figure should be enough to convey the impression that the DNA profile being presented as evidence is rare (Foreman & Evett, 2001).

Databases such as those used in this study may be deemed too small to assess if disequilibrium is present or not, thus affecting the validity of a database due to being unable to truly test its independence. In small isolated populations the potential magnitude of any effect this may have needs to be taken into account. Use of the product rule alone may only serve to increase any bias to the prosecution. The corrections and assumptions of database independence used with sound, robust knowledge of population genetics should alleviate any concerns with regard to the use of DNA evidence in court (Triggs & Buckleton, 2002).

6.6.2 Population Assignment

Although the presence of substructuring failed to be recognised in the STRUCTURE tests (Tables 5.2 and 6.1), significant differences were observed in pairwise F_{ST} and genetic differentiation tests (Table 5.1 and 4.6, respectively). This would go some way to explaining why, overall, sample assignment to the correct population was poor but, as an intelligence-based tool, there is potential to aid an investigation. As seen, this depends on whether the 'true' offender population is included in any comparisons and what other populations are compared alongside it. That said, two genetically similar populations (Indian and Pakistani databases in this case) may only serve to hinder such intelligence by not providing a credible lead worthy of investigative resources. This is despite both databases reporting the highest proportion of correct assignments from their own samples, albeit 51 % or less (Table 6.4, Graph 6.5).

Studies have been carried out previously which attempted to assign geographical origin to samples based on STR profiles using up to 19 loci (Klintschar *et al*., 2003). They

took samples from Austrians, Egyptians, two Hungarian populations and four populations from New York: US Caucasian; Afro-Caribbean; Asian; and Hispanic. Of the four US populations, there was a less than 70 % success rate in correctly assigning US Caucasians and Hispanics; although they did report at 93 % success rate at determining whether a sample belonged to one of the US or non-US groups. Had they taken substructuring into account, these figures may have been reduced, particularly as all four US populations were taken from one city. With a greater chance of migration between populations living in close proximity, this is perhaps a flaw of studies such as this but it highlights the potential of a system not originally designed for its ability to segregate DNA profiles. Indeed, by their very nature, the STRs selected are chosen based on their high polymorphism rate and discriminatory power.

Work has been conducted at a more regional level in the UK, which shows that there is very little difference (or at least none of practical importance) when measuring population differentiation on a more distinct geographical basis, i.e. counties or towns (Evett *et al*., 1996b; Foreman, *et al.*, 1998). In a region such as the North West province of Pakistan, where the Kalash reside, it may be easier to observe any regional differentiation compared to other parts of the country. In practice, however, no system of classification or 'grouping' will be 100 % accurate, particularly when based on highly polymorphic STR markers and so information provided by such a system would be for police intelligence purposes only.

6.6.3 Substructuring of Populations

The UK database showed almost equal F_{ST} values to both the Indian and Pakistani samples at 1.2 % (Table 5.1) which further suggests that there is little difference between these Asian populations. The Kalash, however, show the greatest effect of substructuring amongst all populations with a pairwise-difference of at least 2.6 % from all other populations studied. This divergence is clear even though the geographic

distance between the Kalash and Pakistani populations is very small. This may highlight the level of consanguinity within the Kalash as a close-knit community as well as other genetic phenomena such as genetic drift and the founder effect which will have a greater effect on smaller populations.

As seen with the Kalash data, inclusion of such an isolated population can affect profile frequency estimations and correct assignment of samples to populations. With the application of generous correction factors, much of the genetic variability between populations should be accounted for. Generally speaking, forensic STRs are of little use in terms of aiding geographical assignment of a sample. However, when considering populations such as the Kalash, although geographical assignment may be improved, higher correction factors may be required in whichever database is used for profile frequency estimations to avoid overstating the strength of the DNA evidence greater. The effects and incorporation of substructure parameters in profile frequency estimation (as well as size bias and minimum allele frequency corrections) have ensured that match probabilities are not overestimated. If they were ever thought to be because, essentially, a F_{ST} value is arbitrary and only a snapshot of a representative proportion of the population at that time, the magnitude of any effect should be minimal and not of any detriment to a defendant (Foreman & Evett, 2001; Triggs & Buckleton, 2002).

7 GENERAL DISCUSSION

The rationale for this study was to examine the effects of cognate and non-cognate databases on population isolates and to consider their application in forensic DNA profiling. The feasibility of estimating geographical origin based on familiar and common approaches to DNA profiling was also considered.

Comparisons of profiles between populations showed marked differences in profile frequency calculations when calculated against alternative databases compiled from the other populations. This led to the investigation of combined databases and exploring the effect that population substructuring can have on profile frequencies and hence, match probability calculations.

7.1 Autosomal STR Analysis

Statistical analyses revealed geographical separation of populations cannot be assumed to indicate greater genetic variability between them. Socio-cultural factors within countries may account for some of the greater differences seen, particularly within this study, though should not be relied upon when considering genetic variance and differentiation between populations (Manica, *et al.*, 2005).

The inclusion of F_{ST} as a co-ancestry correction factor can be adjusted to account for both expected and unexpected variability. Including a population such as the Kalash into a combined population database has shown that it can skew allele frequencies sufficiently to have an adverse effect on the match probabilities of the profiles compared against it. Although expected in this study, it highlights a need to monitor populations being grouped together to form a broad database. The magnitude of any such error may only be of importance when dealing with a particularly isolated population with perhaps several rare alleles. These are the occasions when a standard F_{ST} value may be inadequate.

From a criminal justice perspective, to remove any ambiguity as to the database and/or level of correction being employed, the simplest option would be to have a specific database compiled from each population. The data from this study suggest that this may not be necessary when considering large populations from neighbouring geographical areas, for example, India and Pakistan. However, defining a population in this respect is difficult: it cannot be based purely on geographical boundaries as data collected and analysed from the Kalash region has shown and is contrary to the 'isolation by distance' (IBD) model (Cavalli-Sforza, *et al.*, 1994). Although a database from each population may reduce uncertainty in criminal proceedings, obtaining sufficient, representative data from every population and genetic isolate would be impractical, hence the significance of using the appropriate level of correction.

7.1.1 Effect of Profile Completeness

One major factor of this study is that the data have relied on complete DNA profiles, comprising 20 alleles (plus amelogenin) as defined by the AmpFℓSTR® SGM Plus® PCR Amplification Kit (Applied Biosystems, UK). To allow for comparisons between populations and attempt to apportion unknown individuals to them, greater precision is achieved with more loci available to profile. Not surprisingly, not all samples retrieved from crime scenes yield complete DNA profiles but the use of SGM Plus® allows for highly discriminating match probabilities – even for non-complete DNA profiles.

Evidence type invariably effects the quality of a DNA profile, for instance, blood is more likely to give a complete profile than, for example, epithelial or 'touch' DNA samples (Cotton *et al.*, 2000). However, regardless of the source, DNA from aged samples can show signs of degradation; in particular, amplicons of high-molecular weight (Golenberg, *et al.*, 1996), which become fragmented over time due to environmental factors and bacterial action (Coble & Butler, 2005). This results in an incomplete profile

where allelic drop-out has likely occurred as a result of the fragmentation of template DNA (Miller, *et al.*, 2002).

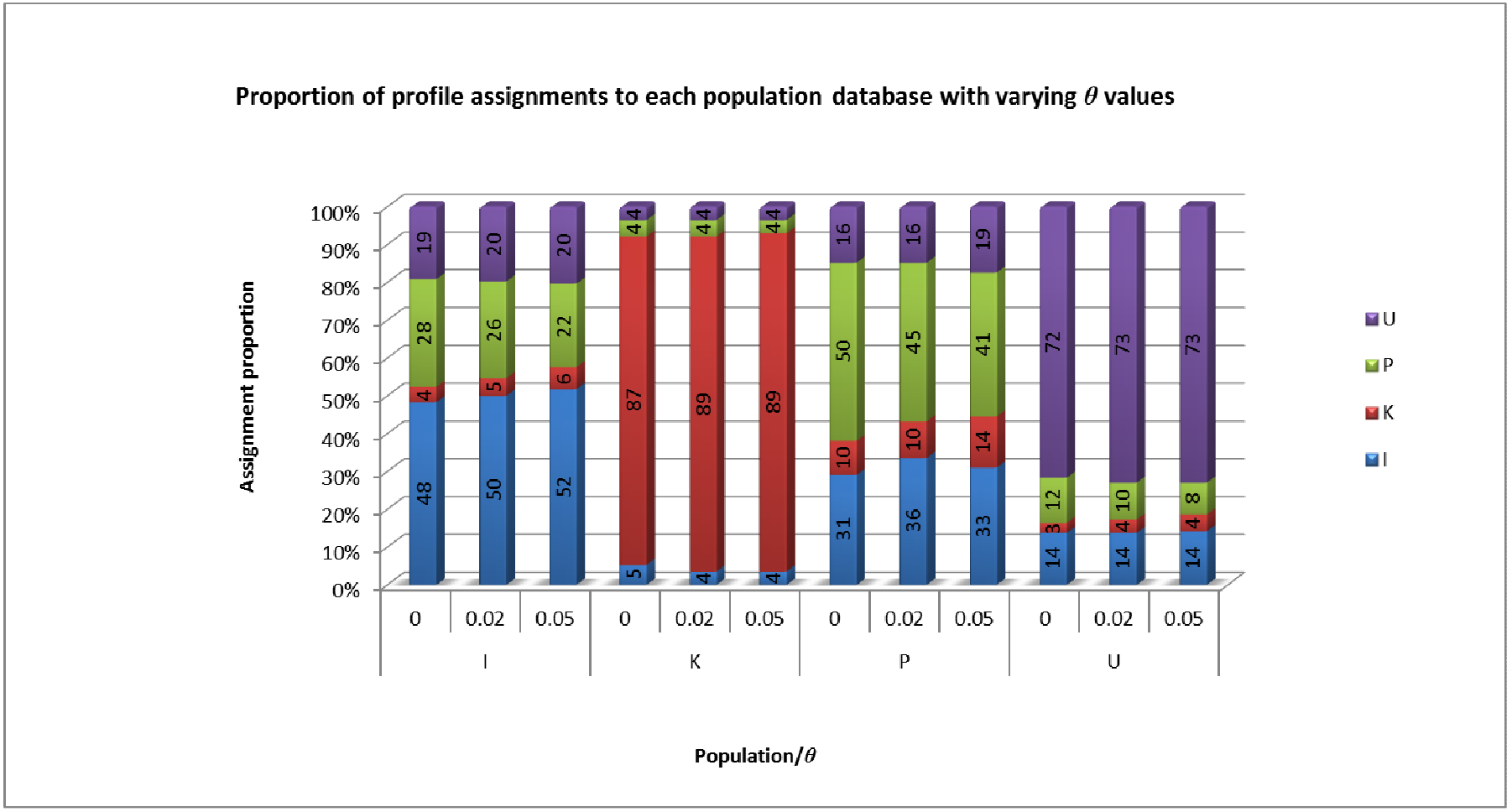
To simulate this effect, the most discriminatory locus (PD [Jones, 1972]) in each database was removed; in all cases, this was locus D2 (Tables 4.1-4.4). The profile frequencies of each sample were then recalculated, providing a higher match probability. In addition, the four additional loci included in the SGM Plus® kit were removed (D3S1385, D16S539, D2S1338 and D19S443), leaving the original six loci comprising the original SGM kit where adventitious matches were recorded (Goodwin, *et al.*, 2011).

Graph 7.1 shows the effect removal of D2 has on sample assignment and Graph 7.2 shows assignment based on the original SGM loci. There is little difference, as perhaps expected, with the removal of the D2 genotype frequency from each sample. However, the proportion of Pakistani samples being correctly assigned has increased at each level of F_{ST} correction as well as the number of Indian samples in the Indian database at F_{ST} = 5 % when compared to the data from full profiles (Graph 6.5). Interestingly, a similar effect occurs when the additional four SGM Plus® loci are removed (Graph 7.2). For the Indian database, the proportion of correct assignments increases where F_{ST} = 2 % or 5 %. Conversely, the greatest reduction of correct assignments (8 %) is apparent in the UK population at F_{ST} = 0 % and 2 % when compared with full profile comparisons (Graph 6.5).

Overall, the precision of correct individual assignment may increase with the addition of further loci but, as shown, the removal of some clearly make some samples show an even greater affinity for their own cognate database. With the Indian and Pakistani databases, this may be due to the lack of significant differentiation across many of the loci tested (Table 4.5). Therefore, there are generally more samples from the opposing population being incorrectly assigned to either the Indian or Pakistani database.

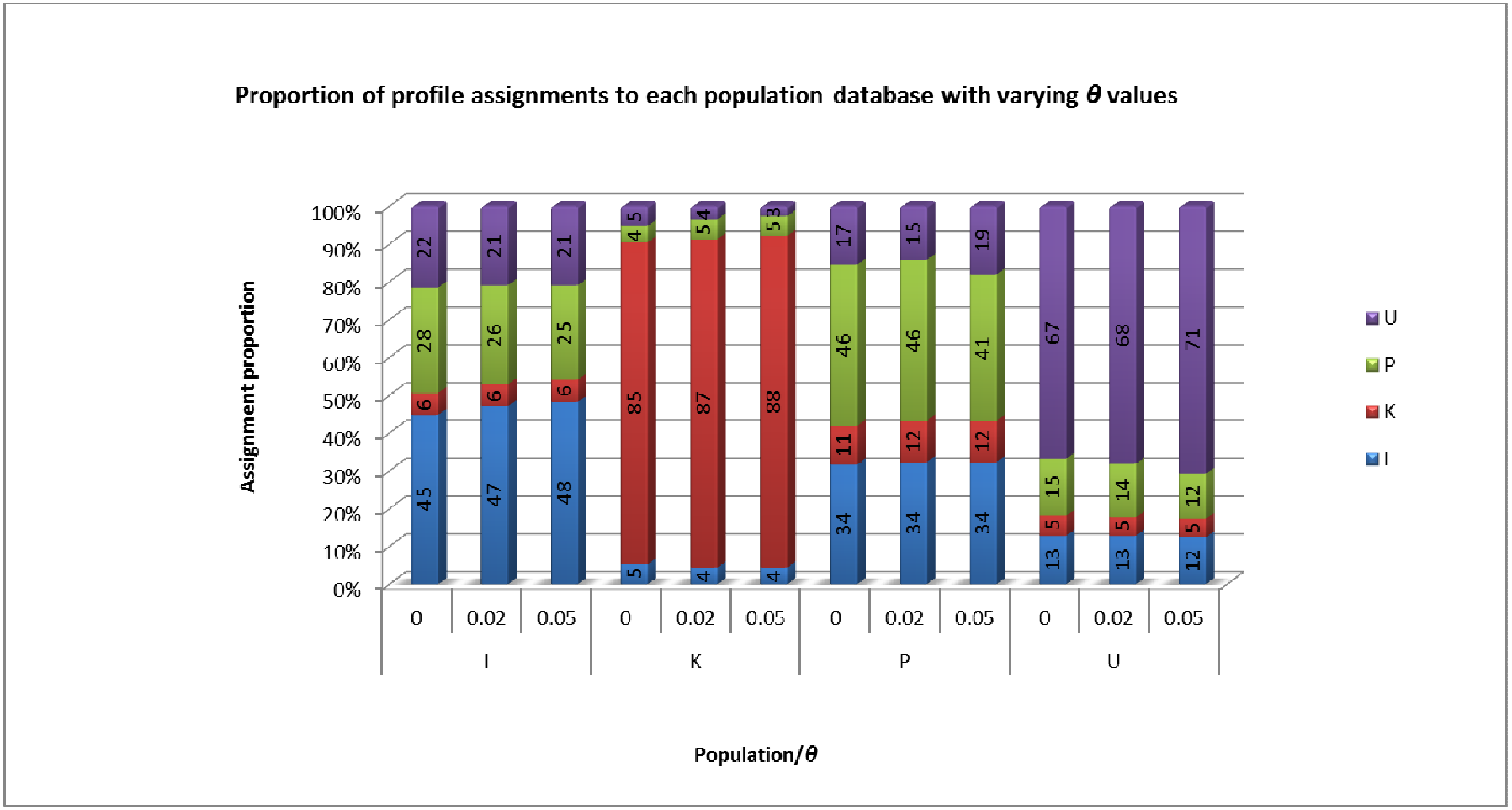
With the UK data, the four loci that have been removed all differentiated significantly with each of the other three populations (Table 4.5). This may have caused the reduction in correct population assignment though it still shows at least a 67 % success rate.

Graph 7.1: Proportion of sample assignments to each database at varying F_{ST} values with locus D2 removed



U = UK, P = Pakistani, K = Kalash, I = Indian

Graph 7.2: Proportion of sample assignments to each database at varying F_{ST} values with SGM loci only



U = UK, P = Pakistani, K = Kalash, I = Indian

7.2 Future Work

New population databases are published regularly and these aid the points raised above with regard to appreciating genetic variability between populations and need to use databases most akin to someone's genetic background if it is to be used in criminal proceedings.

Substructuring within populations must also continue to be explored so that it can be shown that the correction factors currently applied by forensic service providers are accurate and do not overstate the strength any DNA evidence presented to a court.

There is also potential in the recent advances in whole genome sequencing that have formed the basis of projects such as the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2010). This may further improve the resolution of genetic variation both within and between populations but are not suitable, or really necessary, for routine DNA forensic investigations. Efforts continue to identify informative SNPs which characterise strong population substructuring across the genome (Li *et al.*, 2008). The rapid expansion of this field continues to improve the accuracy of ancestry-based genome studies. Although there is incredible potential in such a tool, the practical element to forensic application also has to be considered; key factors including cost and timeliness of results. There are currently methods which could be employed to aid geographical origin on a continental basis which may be more appropriate (Phillips *et al.*, 2007).

The use of externally visible characteristics has been discussed previously (section 1.5.2) (Kayser & Schneider, 2009; Kayser & de Kniff, 2011). The purpose is to interrogate genes which may be used to accurately predict certain phenotypic features, such as eye, hair and skin colour. Just a few select SNPs can already provide accurate predictions of phenotypic characteristics. In 2007, Lao *et al.* reported that over 80 % of skin colour variation can be predicted using just five SNPs. The 'HIrisplex' system,

developed by Walsh *et al.*, (2012a), reports at least a 69.5 % success rate in predicting hair colour of Europeans which builds upon the work conducted to accurately predict eye colour in 90 % of individuals using just six SNPs (Walsh *et al.*, 2011a).

Combined with conventional DNA analysis, there is potential for a useful intelligence tool to narrow down a search for an individual based on a DNA profile alone. What must be remembered throughout is that the use of EVCs results in a prediction and not a definitive lead and appropriate caveats should be in place regarding their use. As DNA population databases using common multiplex kits continue to emerge, this should build upon the knowledge of population genetic variation and increase the possibility of predicting ancestral/geographical origin. In tandem with the power of EVCs, routine DNA profiling may soon be able to provide more information in the early stages of an investigation, rather than currently where its value is only realised once a suspect is identified.

APPENDIX I

UNIVERSITY OF CENTRAL LANCASHIRE

CONSENT FORM

ASSESSING THE EFFECT OF SUBPOPULATIONS ON THE APPLICATION OF FORENSIC DNA PROFILING

Researchers: Dan Clark – Department of Forensic & Investigative Science

Supervisors: Dr William Goodwin and Dr Sibte Hadi

RESEARCHER'S STATEMENT
We are asking you to be in a research study. The purpose of this consent form is to give you the information you will need to understand about this study. Please read this form carefully. You may ask any questions about any sample you give and how it will be used, or anything else about the research or this form that is not clear. When all of your questions have been answered, you can decide if you want to be in this study or not. This process is called "informed consent." A copy of your form will be kept in MB130 and can be seen upon request.

PURPOSE AND PROCEDURE
As part of the study, we will be taking DNA samples from you.

If you agree, we would like to keep samples of your DNA. The samples will be kept here at the University of Central Lancashire and will be used for research and teaching.

The research that is done with your samples will not be of any direct use to you. It is being done for use as a Forensic tool and no information about your sample or anyone else's can be given. The results may be shared with collaborating institutions / companies but no personal data will be revealed.

THINGS TO THINK ABOUT
We will label your samples with a code that will indicate to us which ethnic origin you consider yourself to belong to. This will be asked when you provide the sample. This code will not be printed on your consent form. We will not keep any record that would link the code with any information that could identify you. Before work commences, 10 % of all samples collected will be destroyed to ensure anonymity.

Sometimes DNA is used for genetic research (about diseases that are passed on in families). Because we will not be able to link the test results with your name, we will not be able to give you any test results.

OTHER INFORMATION
Your sample will be kept until it is used up or destroyed. The samples will be used only for research and teaching. Your name will not be used in any published reports about this study.

QUESTIONS
If you have questions about this research or about this study, please contact one of the people listed on this form.

Dan Clark – dclark@uclan.ac.uk

William Goodwin – whgoodwin@uclan.ac.uk

Sibte Hadi – shadi@uclan.ac.uk

_____ _____

Signature of person obtaining consent Date

_____ _____

Printed name of person obtaining consent Date

SUBJECT'S STATEMENT
I agree to allow the University of Central Lancashire to store my DNA sample for future research about inferring geographical origin. I understand that no information can be given to me or anyone else about specific samples as there is no way of tracing them.

_____ _____

Signature of person obtaining consent Date

_____ _____

Printed name of person obtaining consent Date

APPENDIX II

STR data, genotype and profile frequencies for all samples profiled using the AmpFℓSTR® SGM Plus® PCR Amplification Kit (UK, Indian and Pakistani) and the Kalash population, profiled with the AmpFℓSTR® Identifiler® PCR Amplification Kit

Table A2.1 STR data for the UK population

Sample Name	D3		vWA		D16		D2		D8		D21		D18		D19		TH01		FGA	
UK1	15	16	15	19	11	11	17	24	12	13	30	30	14	14	12	16	6	6	23	24
UK3	14	16	16	17	11	13	20	24	11	12	29	31	12	12	14	14.2	6	7	20	23
UK5	14	17	19	20	11	12	20	25	13	14	29	31.2	12	15.0	13	14	7	8	23	23
UK8	13	15	14	16	9	12	24	24	12	12	29	32	14	17	12	16.2	6	9	21	23
UK10	15	16	15	18	14	14	17	23	13	16	29	31	14	15	13	14	6	9	21	23
UK12	16	16	14	17	11	11	17	20	11	14	30.2	32	14.0	15	14	14	7	9.3	20	21
UK14	17	17	17	18	9	13	17	20	13	14	29	30	15	17	14	15	6	9.3	20	22
UK16	16	17	17	18	9	11	16	20	9	12	29	30	11	13	13.2	16	9	9.3	19	22
UK18	16	16	15	15	10	11	17	23	14	14	30	32.2	18	20	14	15	6	7	23	25
UK20	14	15	16	19	11	11	19	22	10	12	30	32.2	14	16	12	13	9	9.3	18	26
UK21	14	15	17	18	9	13	17	20	13	15	29	30	14	16	12	14	7	7	22	25
UK23	17	17	17	18	11	12	23	23	10	13	28	30	14	17	16	16	7	9.3	21	23
UK25	17	17	18	19	9	9	19	24	13	15	29	30	13	18	13	14	9.3	9.3	21	26
UK26	13	14	18	18	9	13	18	18	11	14	29	29	12	17	13	15	9	9.3	20	25
UK28	17	18	18	19	10	13	20	22	13	13	30	33.2	12	13	13	14.2	7	9.3	21	22
UK30	15	17	15	17	13	13	23	25	10	13	32.2	32.2	14	18	12	14	7	9.3	21	25
UK32	17	17	16	18	9	12	23	25	13	13	30	31.2	12	14	14	16.2	8	9.3	20	24
UK34	14	18	16	17	9	11	17	24	12	15	29	29.2	13	21	13	14	6	7	22	26
UK36	15	16	16	17	11	12	19	22	13	15	30	33.2	18	18	13	14	9	9.3	23	23

Sample Name	D3		vWA		D16		D2		D8		D21		D18		D19		TH01		FGA	
UK38	16	18	16	17	11	12	17	26	12	14	28	28	14	15	15	15.2	6	7	20	23
UK39	16	16	16	17	9	11	17	24	11	13	29	30	12	13	13	14	8	9.3	21	25
UK42	16	18	14	17	12	12	19	19	14	15	30	32.2	13	17	14	15	6	9.3	24	26
UK44	16	17	15	16	8	13	23	24	14	14	30	31.2	17	17	13	16	8	9	19	20
UK45	14	17	18	19	10	10	17	21	10	12	29	32.2	15	21	13	16.2	7	9.3	19	24
UK48	16	18	16	18	11	13	20	25	14	15	28	30	16	18	13	14	6	9.3	22	24
UK49	15	16	18	18	11	13	17	23	10	14	30	31.2	15	18	14	15	6	6	24	26
UK52	14	15	17	18	13	13	18	22	10	14	28	31.2	7	14	14	14	6	6	22	24
UK53	17	17	17	18	11	12	22	23	10	13	28	29	14	18	14	14	6	9.3	21	24
UK54	16	17	16	17	9	12	18	25	13	14	30	30	12	16	12	14	7	9	21	23
UK56	14	16	17	18	9	11	23	25	13	13	30	31	12	12	13	15	8	9.3	21	23
UK62	15	17	17	18	9	11	17	17	12	12	30	32.2	15	21	13	17	7	9.3	20	22
UK65	14	18	16	16	8	11	22	25	8	8	29	31	15	17	13	16	6	9.3	20	22
UK68	15	16	15	17	11	12	19	23	10	12	27	31	13	13	13	14	6	9.3	21	22
UK71	14	18	16	18	11	11	23	24	13	13	28	32.2	14	18	14	15	6	9.3	19	22
UK72	16	17	16	17	11	13	23	24	10	14	30	32.2	13	14	14	15	6	7	21	25
UK73	15	15	17	18	8	11	16	22	11	13	30	30	14	15	13	15	6	9.3	24	27
UK74	16	17	16	16	9	13	21	25	15	16	29	31.2	15	16	12	14	7	7	20	21
UK75	15	16	16	16	11	14	19	19	11	12	28	28	16	16	13	15	8	9.3	22	23
UK77	15	18	16	18	12	13	19	22	11	13	28	28	10	18	13	15	7	9	20	23
UK80	14	15	16	17	10	13	19	22	12	13	28	31	12	15	13	14	6	8	20	22
UK81	15	17	18	19	13	13	16	16	13	14	30	32.2	15	15	13.2	15	7	9.3	22	23
UK82	14	14	17	19	11	13	18	23	11	12	30	30	16	16	13	14.2	9.3	9.3	21	23
UK83	16	18	17	18	11	12	20	20	12	13	31.2	32.2	12	12	14	15	7	9.3	21	21
UK84	17	17	14	20	13	13	17	24	12	13	30	30	13	16	14	14	8	9.3	18	20
UK86	16	17	16	18	9	13	18	25	12	14	31	31	13	14	14	16.2	6	9.3	22	22
UK87	14	16	16	19	9	12	16	25	13	15	30	30	17	20	13	14	9	9.3	20	24
UK88	14	17	16	18	12	13	17	25	12	12	30	30	14	15	14	16	7	9	21	24
UK89	14	15	14	16	8	12	16	26	12	13	29	29	15	17	15	15	9.3	9.3	20	24
UK90	15	17	16	19	11	11	17	17	10	10	31	31	12	12	13	15	9	9.3	19	21

Sample Name	D3		vWA		D16		D2		D8		D21		D18		D19		TH01		FGA	
UK91	17	18	17	19	9	11	17	19	12	13	28	30	20	20	13	15	6	9.3	20	24
UK92	16	17	16	16	9	13	17	17	12	12	29	30	14	14	14	14	9	9	24	25
UK95	13	18	16	17	9	11	20	21	8	8	28	33.2	13	17	12	12	9.3	9.3	24	26
UK97	16	17	17	19	11	13	22	26	10	14	28	31	13	14	14	16	6	9	23	36
UK98	15	16	17	18	11	13	16	20	9	14	33.2	33.2	12	12	13	14	6	7	19	21
UK101	16	17	17	21	10	12	17	20	12	14	31	31.2	14	14	15	16	6	6	22	26
UK102	18	18	14	18	9	11	20	24	14	15	27	30	15	15	13	13	9.3	9.3	21	21
UK106	14	15	15	18	8	13	20	20	14	15	28	31	18	18	12	14	6	9.3	22	22
UK107	15	16	14	14	10	12	19	23	10	14	29	30	12	12	14	14	7	9	20	25
UK108	14	15	18	18	10	11	19	24	12	13	30	33.2	13	16	12	15	6	7	21	21
UK109	15	15	17	20	12	12	17	23	12	13	29	29	14	15	14	16	9.3	9.3	23	24
UK110	14	16	17	17	11	13	18	25	14	14	30.2	31.2	14	14	12	13	7	8	22	24
UK111	15	18	15	17	9	11	19	20	12	14	28	28	10	10	14	14	9.3	9.3	23	24
UK115	15	17	16	19	11	11	20	20	13	15	28	31	14	14	12	15	7	9.3	21	23
UK116	16	17	16	18	9	13	18	18	12	12	28	28	13	13	14	15.2	9.3	9.3	20	21
UK117	15	16	15	18	11	12	17	17	12	14	28	29	12	17	14	15.2	6	8	21	25
UK121	16	18	15	16	11	12	17	25	13	13	31	31.2	15	15	12	15	8	9	24	24
UK122	15	16	15	17	10	12	19	24	13	16	29	30	13	13	13	14	6	8	20	21
UK125	14	16	15	16	10	14	20	22	12	15	29	29	16	17	12	14	7	9.3	20	22
UK126	15	17	17	17	12	13	16	23	13	16	30	30.2	15	15	12	12	6	9.3	18	23
UK129	17	17	16	17	11	12	17	17	13	15	32.2	32.2	15	15	14	15	7	7	21	21
UK130	16	17	18	18	11	12	23	25	11	12	29	30	14	14	13	13	7	9.3	21	22
UK132	16	18	16	18	11	12	24	26	13	13	29	32.2	14	20	13	15	6	9.3	20	21
UK133	16	18	15	17	11	12	23	25	14	16	28	31	11	17	13	13	6	9.3	20	24
UK134	14	18	14	18	13	13	21	25	8	10	29	29	12	15	12	15	6	9.3	21	23
UK135	14	14	14	18	9	13	20	23	12	13	28	30	18	21	12	13	7	9.3	22	22
UK136	16	16	14	15	11	11	19	24	12	13	28	29	12	16	14	15	9.3	9.3	20	21
UK137	16	16	14	16	11	11	20	23	10	12	30	34.2	12	15	13	14	6	6	19	23
UK138	16	18	15	17	11	14	18	25	14	16	28	30	12	12	13	14	6	7	20	23
UK139	17	18	16	19	11	11	17	23	13	14	28	32.2	15	16	13	14	7	7	20	21

Sample Name	D3		vWA		D16		D2		D8		D21		D18		D19		TH01		FGA	
UK140	15	19	16	18	11	11	23	23	10	14	32	32	18	18	14	16	8	9.3	20	21
UK141	14	16	16	17	11	13	20	20	13	13	29	29	14	15	13	14	7	9	22	27
UK143	15	17	14	17	11	12	19	25	10	14	28	31	12	16	13	15	7	9	19	21
UK144	15	17	16	19	9	11	18	24	10	14	29	29	10	12	14	15	6	9.3	20	22
UK145	17	18	15	16	11	12	17	19	11	13	28	29	15	19	12	12	6	9	22	24
UK146	16	17	14	16	12	12	17	23	12	14	29	29	17	20	14	14	7	9.3	22.2	25
UK147	14	18	14	18	11	12	17	17	10	13	29	32.2	13	14	14	15.2	7	8	22	24
UK148	14	17	16	18	10	14	16	17	13	13	30	31	17	19	13	14.2	8	9.3	24	25
UK149	16	17	17	19	12	12	20	25	11	13	30	31.2	15	16	14	15	7	9.3	19	19
UK150	14	18	14	17	10	12	18	25	13	15	28	30	15	16	13	15	7	9.3	22	24
UK151	14	15	16	17	12	13	17	21	14	14	27	29	14	14	13	16	8	9.3	21	25
UK152	15	17	14	16	9	9	16	25	11	14	28	30	13	18	14	14	6	9.3	22	25
UK153	16	17	17	19	11	12	16	19	13	14	28	29	12	19	13	15	6	7	21	24
UK154	14	18	17	18	11	12	17	25	12	15	27	30	13	16	14.2	15	9.3	9.3	19	22
UK155	17	19	16	17	11	12	22	23	13	14	30	30	12	14	14	14	6	9.3	20	22
UK156	17	18	17	19	9	11	23	24	13	13	29	32.2	15	15	14	15	6	9.3	20	24
UK157	14	15	16	18	11	13	20	24	12	13	28	28	12	12	14	14	9	9.3	23	27
UK158	15	16	16	17	12	13	17	24	13	14	29	32	12	18	13	14	6	9.3	20	21
UK159	14	15	16	17	11	14	17	18	13	15	28	31.2	12	15	14	15	6	9.3	20	23
UK160	16	18	16	19	11	12	17	23	11	14	29	30	15	21	13	15	8	9	21	23
UK161	16	17	17	17	12	12	17	24	10	13	28	28	12	14	14	14	7	8	20	23
UK162	15	17	14	18	12	13	17	20	12	13	28	28	15	19	14	14	6	7	22	25
UK163	14	15	17	18	12	12	19	22	13	14	30	33.2	15	16	14	16	7	7	19	20
UK164	14	16	14	16	11	13	19	20	14	15	28	33.2	13	15	14	15	9	9.3	20	22
UK165	15	18	14	17	9	9	17	19	10	14	29	30	16	16	14	15	9.3	9.3	21	21
UK166	16	16	17	17	13	14	18	25	13	13	31.2	32	13	18	12	15	7	9.3	21	24
UK167	17	19	14	17	10	11	24	24	12	12	29	31.2	16	16	14	15.2	9.3	9.3	20	25
UK168	15	16	16	18	13	13	18	25	13	13	28	29	12	16	12	14	7	9.3	24	25
UK169	15	16	17	17	12	13	23	25	10	11	30	31.2	16	16	13	14	6	9.3	22	25
UK170	16	16	14	18	11	12	18	19	13	14	26	32.2	13	14	14	14	6	9.3	20	21

Sample Name	D3		vWA		D16		D2		D8		D21		D18		D19		TH01		FGA	
UK171	15	16	17	17	11	13	18	21	13	13	29	31.2	12	22	13	13	9	9.3	21	25
UK172	14	18	18	18	10	12	17	20	12	12	28	30	10	16	13	15	10	10	19	22
UK173	15	15	17	17	12	13	17	18	10	13	30	31.2	16	17	13	14	9.3	9.3	20	23
UK174	15	15	17	19	13	13	17	19	12	14	28	29	12	18	13	15.2	6	9.3	18	24
UK175	15	18	15	15	9	13	16	25	13	15	28	32.2	15	19	14	14	7	9.3	18	22
UK176	14	18	16	17	11	13	17	23	14	15	29	29	13	16	14	14	6	9	24	25
UK177	15	18	15	16	12	12	16	19	13	13	30	30	12	14	14	14	9.3	9.3	20	22
UK178	15	17	15	18	9	11	22	25	10	14	30	31.2	13	17	12	14.2	9.3	9.3	20	24
UK179	16	18	18	18	11	12	17	17	8	10	28	30	14	15	14	14	6	9.3	20	24
UK180	14	15	16	17	12	13	17	20	8	13	29	30	14	19	14	14.2	6	6	20	21
UK181	14	16	14	17	10	13	22	23	11	12	31	34.2	12	15	14	14	6	6	22	23.2
UK182	16	17	16	17	10	12	20	25	13	14	30	32.2	11	19	13	14	6	8	20	22
UK183	16	17	16	18	12	13	20	26	14	15	28	32	14	19	14	14.2	6	9.3	19	22
UK184	15	16	14	18	11	12	23	24	12	13	30	32.2	9	17	12	14.2	6	9	23	24
UK185	15	18	15	17	11	11	19	20	12	14	27	31.2	13	15	13	13	8	8	23	23.2
UK186	18	18	17	18	13	13	18	23	13	14	30	30	14	19	13	14	6	6	20	23
UK187	15	16	16	18	11	13	24	24	13	13	30	30.2	12	14	13	14	7	8	23	24
UK188	15	18	16	17	11	12	17	23	10	13	30	31	12	14	12	14	6	7	21	21
UK189	16	18	17	17	11	12	19	25	10	13	29	34.2	12	18	12	15	7	9.3	18	22
UK190	17	18	17	17	9	11	17	25	8	13	30	30.2	12	14	13	15	9.3	9.3	23	23
UK191	16	16	14	16	8	9	19	20	13	13	30	31.2	14	14	14	14	6	9.3	21	23
UK192	15	16	14	14	12	13	24	24	10	13	29	29	14	19	12	14	8	9.3	21	22
UK195	14	16	15	16	8	11	17	19	10	14	28	31	14	14	13	14	6	9	21	24
UK196	14	15	16	16	11	13	17	24	10	13	29	30	16	20	13	13	6	6	21	23
UK197	15	15	15	19	11	12	19	25	13	14	30	33	11	16	11	14	6	6	24	24
UK198	15	16	16	17	11	12	19	22	13	15	30	33.2	18	25	13	14	9	9.3	23	23
UK199	17	18	17	18	9	13	18	24	14	14	30	31.2	12	17	13	16	6	9	19	20
UK200	14	15	17	18	12	12	20	24	12	13	30	31	13	16	14	14	9.3	9.3	21	25
UK201	15	15	16	17	10	11	16	19	13	15	29	30	13	16	13	16	7	9	20	22
UK202	16	17	16	18	10	11	20	25	13	14	31	31.2	16	21	14	16	7	7	20	24

Sample Name	D3		vWA		D16		D2		D8		D21		D18		D19		TH01		FGA	
UK203	15	17	16	18	9	12	17	25	11	13	27	31	14	15	14	15	6	9	21	25
UK204	15	16	15	18	10	13	19	19	11	13	30	31	12	19	13	13	9.3	9.3	20	23.2
UK205	17	18	16	18	10	11	21	25	10	13	29	30	17	18	14	15	7	9.3	19	24
UK206	15	16	16	18	10	14	23	25	10	13	28	29	14	14	14.2	15	6	7	20	25
UK207	15	15	17	18	8	12	24	24	10	13	28	32.2	16	19	14	15	7	9.3	24	25
UK208	16	16	17	18	11	11	20	23	13	13	28	29	13	14	14	14	6	9.3	22	23
UK209	16	16	15	19	12	12	17	19	13	13	28	32	14	19	15	15.2	7	7	19	20
UK210	15	16	18	20	11	12	19	23	13	14	29	20	13	13	12	16	8	9.3	20	25
UK211	15	18	14	20	9	12	17	23	13	13	27	28	12	17	14	15	7	7	21	24
UK212	14	14	16	18	9	12	17	19	13	14	29	29	13	15	14	15	7	8	19	23
UK213	15	17	14	16	11	13	17	19	12	15	28	30	11	19	13	14	6	9	20	22
UK214	16	17	16	17	9	11	17	18	13	13	24.2	29	13	14	14	14	6	6	20	23
UK215	15	15	14	16	13	13	20	20	13	13	29	31.2	14	19	13	14	6	9	21	21
UK216	14	15	15	18	11	12	20	23	13	14	31.2	32.2	15	16	13	14	7	7	19	20
UK217	16	18	18	18	11	12	20	23	12	12	28	31	14	14	14	15	7	9	21	25
UK218	16	17	17	18	11	14	18	19	11	11	28	30	15	16	13	15	7	9.3	20	25
UK219	13	17	16	18	12	12	17	22	14	15	28	30	12	16	12	15	6	7	22	22
UK220	14	16	18	18	11	11	22	25	14	14	28	30	16	19	13	14	7	9.3	20	22
UK221	15	18	15	18	11	13	17	26	9	14	29	31	14	15	15	16	6	8	20	22
UK222	15	17	16	18	11	13	18	25	13	14	29	31	15	19	13	15	9.3	9.3	21	26
UK223	15	16	16	17	9	9	20	23	8	15	32.2	33.2	15	17	13	13	9.3	9.3	23	25
UK224	14	16	14	17	12	13	19	24	10	11	30	32.2	11	14	14	15	9.3	9.3	25	25
UK225	15	16	16	18	11	12	20	25	11	12	31	32.2	17	18	13	15	7	9	21	25
UK226	15	18	17	18	9	12	19	23	10	16	30	32.2	13	16	13	16	7	9	21	24
UK228	16	17	14	17	9	13	15	19	11	13	29	31	15	15	13	14	6	8	21	23
UK229	14	16	14	16	9	11	19	25	13	13	30	32.2	14	16	13	14	6	9	20	21
UK230	14	17	16	17	9	13	17	20	12	15	29	30	13	16	12	14	6	9.3	22	23
UK231	14	17	18	18	12	12	25	25	13	14	29	30	15	17	14	14.2	7	9.3	19	26
UK233	16	17	16	17	11	14	23	23	10	13	29	31.2	15	16	12	14	7	9.3	20	21
UK235	16	18	15	16	11	11	18	20	10	13	30	31.2	14	19	13	14	6	9.3	24	25

Sample Name	D3		vWA		D16		D2		D8		D21		D18		D19		TH01		FGA	
UK236	16	16	16	19	12	13	19	23	13	14	30	32.2	14	19	14	14	8	9.3	19	20
UK237	15	17	17	18	11	12	16	20	11	13	29	30	13	16	12	14	9	9.3	21	23
UK238	15	18	14	19	12	12	22	24	14	15	29	31	16	17	14	14.2	9.3	9.3	21	24
UK240	15	17	16	18	9	11	19	24	13	14	28	30	14	15	14	15	7	9.3	21	21
UK241	15	15	16	16	12	13	16	16	13	13	28	29	12	14	14	14	7	9.3	19	19
UK242	14	15	14	17	12	13	16	17	9	13	29	29	13	19	14	15	9.3	9.3	20	20
UK243	16	18	14	17	11	12	17	26	10	15	31	31	18	20	14	14	6	9.3	23	27
UK244	17	18	17	17	12	12	18	25	11	13	30	31	18	19	12	14	7	9	23	24
UK245	16	16	14	18	10	12	16	21	13	14	27	29	12	19	14	15.2	8	10	22	24
UK246	15	15	17	20	12	13	17	19	13	14	31	31.2	13	17	14	15	6	8	21	24
UK247	18	18	16	16	11	13	17	24	10	14	30	32	17	17	14	16.2	7	9	21	25
UK248	15	18	17	18	11	12	17	17	14	14	27	29	12	14	14	14	9	9.3	23	23
UK249	14	16	17	19	9	12	19	20	13	15	28	30	12	17	13	14	7	10	20	26
UK250	14	16	16	17	9	12	17	21	11	12	31.2	33.2	12	16	14	14	6	7	19	20
UK251	16	17	16	17	12	12	20	25	13	14	28	32.2	12	17	12	14	6	9	22	25
UK252	15	16	14	17	9	13	25	25	13	14	30.2	31.2	12	17	15	15	6	8	22	24
UK254	15	15	16	18	9	10	17	20	9	12	27	30	12	19	13	14	7	9	20	21
UK255	14	15	15	17	11	13	17	21	10	12	29	30	9	17	12	16	9	9.3	19	22
UK256	14	17	16	16	11	11	17	26	12	13	29	31.2	10	13	14	14.2	9.3	9.3	22	23
UK257	16	17	16	17	13	13	20	23	12	15	28	31.2	14	16	12	14	6	9	20	22
UK260	16	18	17	18	11	12	20	20	12	13	31.2	32.2	12	12	14	15	7	9.3	21	22.2
UK262	14	15	17	18	9	13	17	20	13	15	29	30	14	16	12	14	7	9	22	25
UK263	17	17	15	15	9	11	17	22	12	13	30	30	12	17	12	14	9.3	9.3	22	23
UK264	16	18	17	17	9	14	17	20	12	13	29	29	12	17	12	14	7	8	21	25
UK265	15	16	16	18	9	11	17	20	13	14	28	29	14	15	13	14	6	9	25	25
UK266	15	18	14	18	11	12	19	25	8	13	30.2	31.2	12	12	13.2	14	9.3	9.3	20	24
UK267	15	16	14	16	11	12	17	24	13	14	30	31.2	13	15	11	13	6	7	19	20
UK268	18	18	16	21	11	12	18	18	13	15	28	30	12	17	13	15	6	7	24	26
UK269	15	16	16	18	12	12	17	17	11	11	30	32.2	16	16	14	15	7	9.3	20	21
UK270	15	15	16	17	11	12	18	21	13	13	28	29	13	20	14	14	8	9.3	21	23

Sample Name	D3		vWA		D16		D2		D8		D21		D18		D19		TH01		FGA	
UK271	18	18	17	18	9	12	19	26	8	14	29	31	14	16	14	14	6	9.3	20	20
UK272	15	18	16	18	8	11	24	24	13	16	29	33.2	15	21	13	14	7	9.3	21	22
UK273	15	16	15	16	13	13	17	23	13	13	29	30	12	19	12	14	7	8	20	22
UK274	15	16	16	20	12	13	25	25	10	14	28	28	14	17	12	15	8	9.3	20	21
UK275	14	15	18	19	12	12	17.2	18	12	13	30	30.2	11	14	14	15	7	9.3	20	22
UK276	15	18	16	17	12	12	19	24	13	14	30	31.2	15	17	13	15.2	7	9	23	25
UK277	15	18	17	17	11	12	20	25	11	13	28	29	18	18	13	13	7	9	23	25
UK278	14	14	17	19	12	12	16	20	10	13	27	28	13	15	15	16	9	9.3	23	24
UK279	14	16	18	18	12	13	17	20	13	13	30	30	13	13	13.2	14	7	9	20	20
UK280	15	17	18	18	10	12	18	19	13	14	30	31.2	13	14	14	15	6	7	20	22
UK281	15	17	14	17	12	12	16	23	13	13	29	30	15	17	13	14	9	9.3	22	25
UK282	16	16	18	18	8	12	17	22	10	13	29	31	12	18	12	15	8	9	22	24
UK283	15	15	16	17	9	12	17	20	13	14	30	31.2	13	13	13	14	7	8	25	26
UK284	15	17	14	17	11	13	24	25	11	13	30	31.2	14	15	14	15.2	9	9.3	21	21
UK285	15	16	15	20	12	13	22	26	11	13	30	32.2	17	18	14	15	6	7	22	25
UK286	14	18	14	16	12	12	23	23	10	14	31.2	33.2	10	12	13	14	6	9.3	23	25
UK287	15	18	16	17	11	12	21	21	12	12	29	29	12	18	15	15	9.3	9.3	20	21
UK288	15	16	15	17	11	13	20	25	13	13	28	29	13	15	13	14	9	9.3	20	20
UK289	16	17	14	16	12	13	18	23	12	13	32.2	33.2	14	18	14	14	6	9.3	21	23
UK290	16	17	14	18	11	13	18	18	12	13	30	31	14	15	13	15	7	9.3	20	26
UK291	14	17	15	18	12	12	17	19	10	14	31	31.2	14.2	16	13	14	6	7	19	21
UK292	14	16	14	16	9	12	19	23	14	14	29	32.2	14	15	13	14.2	8	9	21	22
UK293	15	18	16	17	11	11	16	17	14	14	30	33.2	14	17	13	15	6	9.3	19	21
UK294	14	16	16	19	13	13	17	20	10	13	29	29	14	14	13	15	7	9	21	22
UK295	15	16	14	18	12	13	16	18	13	14	28	29	13	14	14	17	6	9.3	22	23
UK296	14	15	16	17	9	12	23	25	13	14	28	33.2	12	13	13	16	6	9.3	20	22
UK297	18	18	17	19	12	13	17	17	13	14	31	31.2	14	18	13	16	6	9.3	22	22
UK298	15	16	18	18	12	13	16	20	14	15	28	31.2	15	15	13	16	9.3	9.3	20	21
UK333	14	17	18	19	9	11	18	20	11	13	30.2	31	16	18	14	15	6	9	19	24
UK334	16	18	15	19	11	12	23	25	13	14	30	32.2	10	12	14	16	9.3	9.3	17	21

Sample Name	D3		vWA		D16		D2		D8		D21		D18		D19		TH01		FGA	
UK335	16	16	15	17	10	11	16	25	13	14	29	31.2	14	14	13	14	9.3	9.3	19	22
UK336	17	18	16	17	10	12	20	24	10	15	30	30	16	18	14	15	8	9.3	20	21
UK337	14	15	17	20	12	12	20	25	13	13	29	32.2	10	17	13	14	7	8	22	23
UK338	14	16	15	19	8	12	23	24	12	13	27	29	12	13	14	14	8	9.3	21	25
UK339	15	15	17	18	12	12	24	24	13	13	26	30	14	15	13	14	8	9	18	22
UK340	14	15	15	16	9	9	22	24	14	15	29	30.2	13	19	13	14	6	9	23	25
UK341	14	17	18	19	11	13	23	26	13	14	29	32	13	16	13	15	7	9	20	23
UK342	14	17	14	18	9	13	24	24	10	13	30	31	12	13	13	14	7	9.3	22	24
UK343	14	14	17	17	12	14	23	25	10	13	30	31.2	13	17	14	14	6	9.3	23	25
UK344	15	16	14	15	11	12	16	24	11	14	30	31.2	12	16	13	14	8	10	20	21
UK345	14	15	16	17	10	11	18	19	10	15	28	29	15	15	13	15	6	9.3	23	25
UK346	15	17	14	17	10	10	17	17	10	13	29	31.2	12	16	13	15	8	9.3	21	23
UK347	16	17	17	18	11	13	17	19	12	15	30	32.2	12	16	14	15	6	7	21	22
UK348	16	16	17	18	13	14	17	17	9	10	30	33.2	14	16	14	15.2	6	7	20	23
UK349	14	18	15	17	13	13	17	20	13	14	28	32.2	12	17	13	13	7	10	19	23.2
UK351	15	16	17	18	11	11	20	20	10	12	28	30	12	18	12	14	7	9.3	22	24
UK352	14	15	15	15	12	13	25	25	14	14	28	31.2	14	15	13	14	6	9.3	21	26
UK353	15	16	17	18	11	12	19	19	13	14	31.2	35	10	17	15	18.2	6	9.3	21	21
UK354	16	16	15	18	9	12	19	24	11	13	29	30.2	12	14	14	15	9.3	9.3	20	24
UK355	14	17	16	17	11	12	24	25	10	12	29	31.2	10	14	13	13	9	9.3	20	21
UK356	14	17	15	16	12	12	20	25	14	15	28	30	14	16	13	14	6	9.3	21	26
UK357	16	17	17	17	9	9	17	19	13	13	29	30	16	17	14	15	7	9	20	21
UK358	17	18	16	18	11	12	17	19	13	14	30	34.2	12	16	13	13	8	9	23.2	24

Table A2.2 STR data for the Indian population

Sample Name	D3		vWA		D16		D2		D8		D21		D18		D19		TH01		FGA	
1	17	18	14	15	9	12	20	23	10	12	30	33.2	12	14	14	15	6	9	23	25
2	15	16	16	17	9	13	18	25	13	17	31.2	33.2	16	18	13	14	8	9	23	24
3	16	17	16	17	13	13	20	20	12	13	30	30.2	17	17	12.2	13	7	9	22	24
4	16	17	16	17	9	12	18	20	12	13	31	33.2	14	14	12	14	9	9.3	19	22
5	16	16	14	16	8	12	20	23	11	13	28	32.2	14	15	14	17	9	9	23	24
6	15	15	16	17	8	12	23	26	10	13	30.2	31	12	14	12	13	9	9.3	23	23
7	15	17	16	17	11	12	19	19	12	13	30	31.2	12	12	13	14	6	6	22	23
8	16	16	17	18	11	11	17	23	12	12	31.2	32.2	15	20	14	15	6	9	24	25
9	15	16	14	17	11	13	22	23	10	13	31.2	32.2	12	20	13	16	6	9.3	22	24
10	15	15	17	18	8	11	19	24	11	16	29	29	12	13	12	15.2	7	8	20	26
11	15	16	17	18	9	11	18	26	15	15	27	33.2	13	15	12.2	13	8	8	21	24
12	16	18	14	17	8	13	20	23	13	15	30	33.2	16	17	12	14.2	8	9	21.2	26
13	16	17	16	18	11	13	23	23	10	10	29	30	12	18	14	14	6	8	19	23
14	16	17	16	18	9	13	18	23	14	15	28	32.2	12	15	13	14	6	9	20	21
16	17	17	14	16	8	11	18	23	10	11	27	32.2	13	13	13	14	6	9	21	22
17	16	16	17	18	9	12	19	23	12	13	30	30	15	15	12	15	7	9.3	19	25
18	16	17	16	17	10	10	19	19	10	11	29	31.2	12	13	13	15	9	9	21	24
19	15	16	15	19	10	12	20	25	10	16	27	32.2	14	16	12	12	9.3	9.3	20	21
20	14	14	16	20	11	13	20	23	13	16	30	32	14	16	13	14	9	9	21	24
21	16	16	16	16	11	13	18	18	11	16	30	31.2	11	14	14	15	9	9.3	19	27
22	15	18	16	18	11	12	17	24	12	13	28	31.2	14	15	13	14.2	9.3	9.3	19	25
24	14	15	17	19	11	12	21	23	11	15	29	33.2	14	15	13	13	7	9	22	23
25	15	16	14	14	11	13	24	25	11	13	32.2	32.2	14	18	14	14	9	9	23	24
26	15	15	16	16	11	11	18	23	10	13	30	32.2	14	19	14	15.2	7	9	22	25
27	15	16	17	19	11	12	23	24	10	10	28	34.2	13	16	14	15	8	9	22	23
28	15	17	14	17	12	13	22	23	9	12	28	31.2	12	13	14	15	6	9	19	23
29	17	18	15	17	8	11	17	24	11	12	30	32.2	15	21	13	18.2	6	6	21	21

Sample Name	D3		vWA		D16		D2		D8		D21		D18		D19		TH01		FGA	
30	17	17	16	18	8	13	23	23	10	12	29	29	13	14	15	15	6	9	22	27
31	17	17	16	19	12	13	17	20	11	15	29	32.2	15	16	12	15	8	9.3	23	25
32	13	18	16	18	11	11	19	24	12	13	30	34.2	16	19	13	14	6	8	20	26
33	15	17	16	17	9	11	17	19	12	14	29	33.2	14	14	13	14	6	8	19	23
34	16	17	15	17	11	11	18	19	13	13	30.2	31.2	13	14	13	14	6	7	20	24
35	17	18	16	17	11	13	23	26	11	13	31.2	31.2	15	18	14	15	6	6	24	25
36	15	16	17	18	9	11	18	23	14	15	28	32.2	14	19	14.2	15	9	9.3	22.2	25
37	15	18	14	19	11	12	22	22	14	15	28	30	14	16	13	14	6	9.3	24	25
38	15	18	18	21	8	12	23	24	10	10	32.2	33.2	15	17	12.2	15	8	9	22	22
39	15	18	16	18	8	12	24	25	12	12	30	32.2	16	16	13	13	7	8	19	24
40	16	17	14	15	11	12	22	24	10	13	30	32	14	15	13	14	9	9.3	24	26
41	15	16	17	18	10	11	18	20	9	14	31.2	34.2	13	17	12	15	8	9	22	24
42	17	18	15	18	11	12	23	24	13	16	29	33.2	14	14	14.2	16	6	9	22	23
43	15	15	18	19	9	11	20	22	13	13	32.2	32.2	14	17	14	14	8	9	20	26
44	14	17	14	15	9	11	18	19	13	13	31	32.2	12	16	15	15.2	9	9	22.2	22.2
45	15	16	16	17	10	13	20	25	13	14	30	30	10	14	12	15.2	7	8	20	26
46	16	17	17	17	11	11	18	22	13	15	29	30.2	14	16	14.2	15.2	6	8	20	23
47	16	17	14	14	9	12	19	23	10	14	30	32.2	15	19	13	15.2	6	9	21	25
48	15	16	14	16	11	13	23	23	12	13	29	29	14	19	16	16	6	9	20	21
49	17	17	14	18	8	12	17	23	13	14	28	29	13	13	12	14	7	9	22	26.2
50	14	18	16	17	11	13	18	18	11	14	33.2	33.2	14	15	12	13	6	9	21	24
51	15	17	16	18	10	11	18	26	10	14	29	32.2	14	14	12	13	9	9.3	19	25
52	16	16	14	17	9	11	18	22	14	14	30	33.2	15	17	13	13	9	9.3	21	24
53	17	17	16	16	9	10	20	24	10	14	29.2	32.2	13	14	13	15	7	10	20	26
54	16	17	17	17	11	12	20	23	13	14	28	32.2	14	17	14	15.2	6	9	24	24
55	15	15	16	17	11	14	19	25	13	14	28	30	14	16	14	16	8	9.3	22	24
56	17	17	15	18	11	11	18	18	11	13	31.2	32.2	14	17	12	14	6	9	19	19
57	16	17	16	17	12	12	19	25	10	14	30	33.2	14	15	14	16	9.3	9.3	22	23
58	16	17	14	17	10	11	19	24	13	14	29	29	16	16	14	15	8	9	21	24
59	16	18	16	16	9	9	23	25	13	14	28	30	14	18	13	13	6	9	24	24

Sample Name	D3		vWA		D16		D2		D8		D21		D18		D19		TH01		FGA	
60	15	18	16	17	10	12	19	23	12	13	30	31	15	17	14	15	6	9.3	24	25
61	15	17	17	18	11	11	18	22	13	13	29	31	13	13	12	14	7	9	21	23
62	15	15	14	19	9	11	17	22	13	16	29	29	15	15	13	14	8	9	19	22
63	15	17	17	17	9	13	22	24	14	14	31	33.2	13	14	13	14	6	7	20	22
64	15	15	16	16	11	11	18	23	10	15	29	32.2	13	15	13	13	6	9	23	24
66	14	17	16	19	10	11	19	23	10	16	29	30	17	24	15	16.2	7	7	19	26
67	14	15	16	19	11	11	19	23	13	14	29	30	12	19	12	13	6	9	22	26
68	16	17	15	17	11	11	18	22	10	13	30	33.2	16	23	13	15.2	9	9.3	22	24
69	16	18	18	18	12	13	18	21	12	14	28	31.2	13	14	14	14	9.3	9.3	21	24
70	16	16	17	18	12	14	22	23	12	16	29	30	13	17	15	15	6	9.3	22	25
71	16	18	14	17	9	11	19	22	9	14	30	30	13	16	12	12	7	8	23	24
72	14	18	15	18	11	12	19	20	13	16	31.2	33.2	14	15	13	14	6	9.3	22	23
74	15	17	18	19	8	9	18	20	15	17	29	29	14	16	14	16.2	6	9	20	21
75	15	15	16	17	12	12	18	21	10	14	29	31.2	12	18	13	14	9	9	21	23
76	17	18	16	18	11	11	18	25	13	15	28	30	14	15	14	14	6	8	22.2	23
77	16	17	16	17	10	13	18	20	8	13	31.2	31.2	15	16	13	13	9	9.3	19	24
78	14	15	14	19	12	15	19	20	10	14	29	30	13	16	13	14.2	9	9	24	24
79	16	18	14	17	12	12	16	20	12	17	28	31.2	15	17	13	14	7	9.3	23	25
80	15	17	16	18	11	13	18	25	12	15	29	32.2	14	17	14	17	7	8	19	24
81	15	17	14	14	11	14	22	24	11	15	29	33.2	13	13	12	15	6	9	19	24
82	15	15	16	18	11	12	20	20	10	15	28	29	12	18	12	12	6	9.3	20	20
83	14	17	14	19	11	12	18	19	15	16	29	30	17	17	15.2	16.2	6	7	19	19
84	17	18	17	18	8	9	20	24	13	14	29	20	15	15	13	16	7	7	19	24
85	15	16	14	17	11	11	20	25	13	15	29	34.2	14	16	15	15.2	6	9.3	24	24
86	16	17	17	18	9	11	18	22	13	13	32.2	33.2	15	23	12	15.2	7	9	24	24
87	15	18	14	17	11	13	24	25	13	14	31.2	32.2	14	20	13	13	8	9	22	23
88	14	18	17	18	11	12	18	18	11	14	29	33.2	12	13	14	15	6	6	23	25
89	14	17	17	18	9	13	18	22	13	14	32.2	33.2	15	15	14	15.2	7	9	22	25
90	16	18	17	18	11	12	18	19	15	16	29	29	12	14	14	15	6	7	23	25
91	15	17	14	19	9	14	22	23	13	15	29	30	13	14	15	15	9	9.3	19	23

Sample Name	D3		vWA		D16		D2		D8		D21		D18		D19		TH01		FGA	
92	15	16	14	17	11	12	20	23	13	15	32.2	32.2	14	16	15	16	9.3	9.3	20	21
93	14	16	16	17	9	9	19	20	8	15	30	32.2	13	16	13	13	7	9	18	24
94	15	15	16	16	10	13	19	20	10	13	28	33.2	12	21	13	14	8	9	20	22
95	16	16	14	18	8	11	17	24	14	16	32.2	33.2	14	15	13	14.2	7	7	24	25
96	16	16	14	16	11	12	16	25	9	10	29	29	15	16	13	14.2	9	9	20	26
97	15	16	15	16	11	11	22	22	11	12	29	31.2	14	15	12	15	6	6	24	24
98	15	16	14	14	11	14	18	19	13	16	28	29	14	14	15.2	16.2	7	8	20	26
99	15	15	18	19	10	11	21	21	14	15	28	28	13	15	13	14	6	8	23	25
100	16	16	18	18	9	13	22	23	13	15	29	30	14	16	13	14	9	9	21	23
101	15	16	17	18	10	12	24	25	10	15	30	32.2	14	14	14	15	6	9	21	21
103	16	18	18	19	9	11	18	19	14	16	30	31.2	12	15	12	13	6	7	19	24
104	14	16	16	19	10	13	17	20	14	15	30	31.2	14	18	13	14	9	9	24	25
105	15	17	16	18	12	13	19	19	13	13	28	28	12	14	14.2	15.2	7	9.3	23	24
106	17	17	14	19	11	13	22	24	10	14	28	32.2	14	16	14	15.2	7	9.3	21	26
107	14	16	16	17	11	12	22	25	14	15	29	30	14	17	13	14	6	9	21	22
108	17	18	14	19	11	12	17	20	11	13	29	31.2	12	15	12	14	7	8	20	21
109	16	16	16	17	11	11	18	19	13	15	29	32.2	14	20	12	13	7	9	21	26.2
110	15	16	16	17	11	12	18	21	12	13	28	31.2	13	14	14	15	6	6	22	24
111	15	16	15	16	9	13	18	18	13	14	28	28	14	15	12	12	7	7	21	24
112	14	16	14	16	11	12	19	21	13	16	28	32.2	13	14	13	15	6	9	19	22
113	13	14	17	17	11	13	23	23	13	15	29	30	11	17	13	15	6	7	22.2	24
114	15	15	18	18	12	12	20	20	12	14	30	31.2	12	17	14	14	7	8	23	24
115	15	16	15	16	10	13	20	23	12	13	28	29	13	13	12	13	6	9	21	24
116	15	15	16	18	8	11	20	22	11	13	30	30.2	14	14	13	13	7	9.3	22	24.2
117	15	17	16	17	11	11	18	22	13	15	30	31.2	15	15	13	13	9.3	9.3	19	26
118	16	18	17	18	10	12	19	19	10	15	29	32.2	12	14	14	15.2	6	7	20	24
119	17	17	17	18	11	13	20	23	12	14	28	33.2	14	16	12	13	9	9.3	22	26
120	16	18	13	17	11	13	17	17	11	13	33.2	33.2	15	15	13	13	6	7	26	27
122	15	15	18	18	9	11	20	23	12	14	31.2	32.2	14	15	14	15	6	6	21	22
123	15	17	19	20	8	10	18	20	10	13	30.2	32.2	15	19	13	14	7	8	21	24

Sample Name	D3		vWA		D16		D2		D8		D21		D18		D19		TH01		FGA	
124	15	16	14	19	11	12	18	25	12	17	29	31.2	12	17	12	13	7	9.3	21	22.2
125	15	17	14	16	11	13	23	23	10	14	30	35.2	12	15	13	14.2	6	9	24	25
126	14	15	16	18	9	13	18	23	11	14	29	30	14	16	13	14.2	6	6	21	25
127	14	15	17	17	11	13	21	26	13	14	28	32.2	12	14	13	14	9.3	10	22	23
128	16	17	15	17	11	12	23	23	13	15	30	32.2	14	15	13	14	6	6	19	23
129	15	16	16	18	10	10	19	26	10	14	29	30	13	14	13	14	9.3	9.3	23	25
130	15	16	14	16	8	11	18	23	10	14	28	29	17	20	13	13	6	7	19	20
131	15	17	17	17	9	10	18	25	11	15	30.2	31.2	14	17	13	13	7	9	19	20
132	16	18	17	18	9	9	18	25	12	14	30.2	31.2	13	14	13	13.2	8	9	22	23
133	14	16	16	17	11	11	18	21	13	13	31.2	32.2	12	15	14	14	6	9.3	23	24
135	14	16	16	18	9	11	23	24	12	13	28	20	10	16	12	14	6	7	22	25
136	16	17	16	17	9	13	18	18	13	14	32.2	33.2	15	23	12	14	7	9	24	25
137	15	16	17	17	12	13	18	19	10	10	28	30	14	19	13	13.2	7	9.3	21	25
139	15	17	16	17	8	10	19	22	10	14	28	30	13	19	12.2	13	6	9	23	24
141	16	17	17	18	9	9	23	24	12	13	30.2	32.2	12	15	13	14	6	8	20	23
142	15	16	17	19	11	12	18	24	13	13	28	29	13	14	12	13	6	7	21	23
146	15	15	14	17	9	9	19	23	10	13	30	33.2	14	15	15	17	8	9	19	21
147	14	16	16	19	12	12	18	18	12	13	28	30	11	14	13.2	15	9	9.3	20	22
149	17	18	17	17	11	12	23	23	12	15	27	31.2	16	17	13	13	8	9	21	26
151	15	15	14	18	9	12	19	19	10	13	30	31.2	14	14	16	17	8	9	19	24
153	18	18	16	17	10	12	20	20	14	14	30	32.2	14	14	12	14	6	9	22.2	26
154	16	18	16	17	12	13	19	20	8	14	31	31.2	14	18	14	14.2	6	6	21	22.2
155	14	15	18	20	9	12	23	25	13	13	31.2	32.2	12	21	15	15	7	9.3	24	24
156	15	16	14	15	9	12	24	24	13	15	30	32.2	12	13	14	16	6	9	20	22
157	16	16	16	17	9	10	22	25	10	15	29	30	15	17	12	14	7	9	21	25
158	14	15	16	17	9	12	23	26	10	11	31	32.2	10	15	14	15.2	9.3	9.3	24	25
160	16	16	15	16	9	14	19	19	12	13	29	31	15	16	14	15.2	6	9.3	19	20
161	15	17	16	17	11	13	19	22	15	15	29	30	16	16	15	15.2	8	9	25	25
163	15	16	18	18	11	13	23	23	10	11	32.2	32.2	13	15	15	15.2	8	9	19	23
164	14	16	15	18	11	11	17	19	12	14	28	30	14	14	13	16	9	9	24	25

Sample Name	D3		vWA		D16		D2		D8		D21		D18		D19		TH01		FGA	
165	15	15	16	17	9	10	25	26	10	13	30	31.2	14	16	15	15.2	6	7	20	23
166	15	16	16	18	11	11	16	19	9	10	16	29	14	17	12	13	9	9.3	22	23
168	15	17	16	17	11	13	17	23	12	14	29	30	14	15	13	14	7	8	20	24
169	16	17	16	19	13	13	19	23	13	13	28	28	13	16	15.2	16.2	6	7	24	25
170	14	17	15	17	11	13	18	22	10	12	29	29	14	17	13	16	7	8	24	24
171	18	18	14	18	9	12	19	19	11	11	29	30	16	18	13	14.2	6	9	25	28
172	17	17	17	18	12	13	19	20	12	15	30	31.2	12	17	9	13.2	6	6	19	24
173	16	17	14	17	11	12	20	25	11	15	29	32.2	12	19	13	15	7	9.3	23	25
174	16	17	16	17	10	13	18	27	10	12	30	31.2	13	14	14	14	6	7	20	23
175	15	17	16	18	9	11	18	19	10	15	29	31.2	13	15	13	14	9	9.3	19	23
176	16	17	18	19	8	11	18	19	10	13	28	29	13	16	12	13	7	7	22	23
177	16	17	17	18	9	11	17	18	12	13	29	29	15	16	13	15.2	6	7	23	24
178	16	17	15	16	12	12	19	22	10	12	28	29	14	15	13	14.2	8	9	20	23.2
179	14	17	16	17	11	12	18	23	9	10	29	31.2	13	14	12	13	6	9.3	22	23
180	16	17	16	17	9	11	18	24	10	12	29	30	14	15	13	15	7	9	22	25
181	15	18	16	17	12	12	18	22	14	15	28	30	14	17	13	14	6	7	19	22
182	15	17	16	16	10	10	17	19	10	12	31.2	32.2	14	15	14	15	7	8	20	23
183	17	17	14	15	11	12	18	21	12	15	29	32.2	13	16	13	13	7	8	23	23
184	15	18	16	17	10	11	22	24	10	14	28	31.2	14	17	14	17	7	8	19	28
185	17	18	16	17	11	13	18	25	10	12	28	32.2	14	14	14	14.2	7	9.3	22	23
187	15	17	14	16	11	12	24	24	10	15	28	29	14	17	14	15	6	9.3	24	25
188	15	17	16	18	9	11	18	19	10	15	29	31.2	13	15	12	13	9	9	19	23
189	17	17	17	17	8	13	19	20	14	15	32.2	33.2	15	16	13	14	6	9.3	24	26
193	17	17	14	17	11	12	19	22	10	11	30	33.2	12	16	15	15.2	6	7	19	20
194	17	18	17	18	11	11	18	20	10	14	31.2	32.2	14	15	15	16.2	6	9	20	20

Table A2.3 STR data for the Pakistani population (J = Punjabi, P = Pushtoon, S = Sindhi)

Sample Name	D3		vWA		D16		D2		D8		D21		D18		D19		TH01		FGA	
J007	16	17	15	16	11	11	17	17	10	13	31.2	32.2	18	19	13	13	6	9	19	19
J008	15	15	14	15	11	11	22	23	10	14	31.2	32.2	13	15	14	14	9	9	22	25
J014	15	16	16	19	12	13	23	25	13	14	28	31	12	21	11	13	8	10	22	22
J020	16	16	16	17	10	11	24	24	10	11	30	32.2	13	17	13	15	6	7	23	24
J030	17	18	15	16	11	12	23	25	8	15	29	32.2	15	15	14	16.2	9	9.3	19	22
J031	15	18	17	18	10	11	18	20	13	16	28	31.2	15	19	13	14	6	9	22	25
J034	14	16	15	16	11	11	18	18	12	15	29	30	14	14	13	15	8	9.3	18	24
J038	16	18	14	17	9	10	18	23	13	15	30	30	13	14	14	14	7	8	23	23
J054	15	18	16	16	11	13	20	23	12	15	29	30	14	15	14	14	9	8	21	22
J055	15	18	15	16	11	11	20	21	12	17	30	31.2	14	16	15.2	16.2	6	9.3	21	22
J056	13	15	15	16	9	9	18	18	10	15	28	29	13	13	15	15.2	6	8	21	22
J057	15	16	17	18	11	11	23	23	10	14	31.2	31.2	14	16	13	15	9	9.3	24	24
J058	16	16	14	19	11	11	18	25	13	16	27	33.2	13	16	13	14	9	9.3	21	24
J059	15	17	18	18	12	12	19	19	10	11	28	30	15	17	13	15.2	6	7	23	24
J061	15	15	16	17	9	12	19	23	13	16	30	31.2	13	15	13.2	15.2	7	9.3	18	22
J063	15	16	15	15	9	12	19	22	8	13	29	31	14	15	14	17.2	7	9	21	23
J064	14	18	16	17	9	12	18	23	11	12	30	32.2	12	15	13	15	6	9	23	24
J065	14	17	17	19	9	10	20	20	13	15	30	33.2	14	14	13	16.2	6	8	21	23
J067	15	17	14	18	12	13	18	20	11	14	29	31.2	14	14	14.2	15	9	9.3	23	26
J071	16	17	16	17	11	14	16	19	13	15	32.2	32.2	11	14	14	15	7	9	20	24
J075	15	15	14	15	9	12	17	19	14	15	30	32.2	14	16	14	15	6	9	21	23
J076	16	17	18	18	11	12	19	19	14	16	32.2	33.2	17	18	14	14.2	6	7	21	25
J077	18	18	14	16	9	9	17	18	12	15	28	31.2	15	17	14	15	7	8	21	22
J078	14	15	17	18	11	12	23	25	9	12	32.2	35.2	14	15	14	14.2	6	8	20	21
J089	15	16	14	16	8	10	22	23	10	15	30	33.2	12	15	13.2	14.2	6	9	21	25
J090	16	17	16	16	12	13	20	23	10	12	31	33.2	15	16	15	15	6	6	19	22
J098	15	18	15	18	9	11	22	23	14	14	29	32.2	16	18	14.2	14.2	9	9.3	23	25

Sample Name	D3		vWA		D16		D2		D8		D21		D18		D19		TH01		FGA	
J103	15	17	17	18	13	13	17	23	10	14	29	32.2	14	17	13	14	6	7	20	22
J105	16	17	15	18	11	14	20	25	15	15	30	32.2	14	19	13	14	6	9.3	23	24
J110	17	18	16	18	8	11	17	24	13	14	31	32.2	12	18	13	13	8	9	25	27
J112	15	17	15	16	8	10	19	21	16	15	29	32.2	13	15	14	16.2	8	9.3	22	23
J113	15	15	17	18	8	12	19	24	13	16	31.2	32.2	13	17	14.2	16.2	6	9	22	26
J117	16	18	14	16	8	11	18	26	13	15	31.2	31.2	12	20	13	15	6	7	22	25
J119	15	18	14	17	12	13	17	19	11	13	29	32	15	18	11	14	6	9	21	22
J121	16	17	16	16	13	13	18	19	13	14	29	29	12	16	15	15.2	7	9.3	23	23
J122	14	15	15	16	9	12	17	23	13	14	29	29	12	14	15	15	6	9.3	21	21
P002	17	18	14	14	8	12	19	24	12	13	32.2	32.2	13	17	14	14	7	9.3	25	26
P003	17	18	14	16	11	13	22	25	15	15	29	29	13	17	13	13	6	9	22	23
P004	14	16	15	17	10	13	18	19	13	13	28	32.2	14	16	13	15.2	6	8	21.2	22
P006	15	15	14	16	11	11	18	25	10	13	30	32.2	14	16	13	13	7	8	19	21
P007	16	18	14	16	12	12	17	25	14	16	28	33.2	14	15	12	14	6	9	23	26
P008	14	15	14	17	10	11	19	20	14	15	30.2	31.2	12	17	14.2	15.2	6	7	22	24.2
P009	15	16	14	17	8	12	17	18	9	13	29	30	12	19	12	14.2	6	6	22	24
P014	16	17	17	17	9	11	23	24	13	14	28	32.2	14	17	13	14	7	7	21	25
P015	18	18	16	18	11	12	23	25	12	13	30	31.2	11	20	13	15.2	7	9.3	20	20
P016	16	18	16	17	10	12	17	22	10	13	28	29	13	14	14	15.2	9.3	10	23	24
P018	17	17	17	19	11	12	23	24	11	13	31	33.2	15	16	13	14	6	9	20	23
P019	14	16	15	16	9	12	19	24	11	15	28	32.2	12	14	12	16	8	9	21	21
P020	16	16	18	19	12	12	18	18	10	15	30	32.2	14	17	14	15	6	9	20	22
P026	16	17	17	18	9	12	19	23	13	14	30	30.2	12	13	14	14	7	9.3	25	28
P027	14	16	14	18	10	13	19	25	14	14	29	31.2	13	14	13	15.2	6	6	19	19
P032	15	15	15	17	9	10	16	20	10	13	29	30	17	17	14	14	8	9	22	22
P034	16	18	16	16	8	9	23	23	13	15	31.2	31.2	14	14	14	14.2	6	7	22	25
P035	14	15	16	18	9	11	19	22	10	13	28	33.2	15	17	14.2	14.2	6	7	24	25
P036	17	19	16	17	13	13	20	23	11	12	29	32.2	10	14	13	15.2	7	9.3	23	23
P038	17	18	15	17	11	12	18	19	14	15	31.2	33.2	15	16	13	15.2	6	8	22	25
P040	15	17	18	19	10	11	16	22	13	14	32.2	32.2	14	17	14.2	14.2	9	9.3	24	25

Sample Name	D3		vWA		D16		D2		D8		D21		D18		D19		TH01		FGA	
P042	14	17	16	17	11	11	23	23	10	10	29	32.2	12	17	13	14	6	6	22	24
P044	16	16	16	18	13	13	18	24	13	15	28	32.2	12	13	14	15.2	7	9.3	22	25
P046	15	16	16	16	12	13	18	21	12	15	32.2	32.2	17	17	15.2	15.2	9.3	9.3	22	22
P047	16	16	14	15	9	11	18	23	13	13	29	29	13	19	14	14	6	8	20	20
P049	15	17	17	18	9	10	17	26	13	14	28	31.2	16	16	13	14.2	7	7	21	22
P051	15	16	17	18	12	12	19	19	13	14	30	33.2	12	14	14	15	8	9.3	20	24
P052	18	18	16	16	12	12	25	25	10	14	32.2	33.2	12	17	12.2	16.2	6	8	22	23
P053	16	16	17	18	10	11	20	20	13	14	29.2	32.2	14	14	14	15	6	7	23	26
P054	14	18	14	17	11	13	18	20	15	18	29	31.2	14	15	13	14	7	9	19	22
P055	15	16	16	18	11	12	18	22	14	15	28	32.2	12	15	13	14.2	6	9	22	23
P056	16	16	16	18	13	13	19	20	10	16	31.2	31.2	15	16	14	15.2	6	9	21	25
P057	15	15	16	16	11	12	18	22	15	15	28	31.2	13	14	13	15	7	9	19	22
P058	16	17	18	20	8	12	19	19	10	11	29	31.2	12	17	13	15	9	9	18	20
P060	14	16	17	17	9	12	18	18	12	13	29	32.2	13	16	14	15	6	6	22	25
P062	17	18	16	16	11	11	24	25	11	13	30	33.2	13	15	13	13	8	9	24	24
P065	15	18	17	18	10	12	22	23	10	16	28	29	15	16	12	13	6	7	22	24
P066	14	16	16	18	11	11	19	23	13	14	28	29	14	16	14	15	9.3	9.3	24	24
P067	14	16	16	17	9	11	19	20	11	13	29	30.2	14	14	13	13	7	9	21.2	22
P068	15	16	18	19	12	13	23	26	10	12	28	31	16	17	14	16	8	9.3	24	26
P069	16	18	17	18	8	9	19	23	11	13	31	32.2	19	20	15	15	7	8	22	26
P071	16	18	17	17	9	13	19	22	14	15	28	28	16	16	14.2	16.2	8	9.3	20	21
P072	16	18	15	17	9	10	20	24	10	13	27	28	16	17	13	14	7	9	21	23
P073	16	17	16	18	11	11	25	25	11	12	30	32.2	13	20	13	15.2	6	6	20	22
P074	15	18	16	19	12	13	16	18	13	16	31.2	33.2	15	18	13	14	8	8	20	22
P075	15	16	16	19	9	9	17	21	11	13	30.2	31.2	13	15	12	16	6	9.3	22	25
P078	15	17	14	19	12	12	18	20	11	12	28	30	14	17	13	13.2	7	9	22	24
P079	16	18	16	19	10	12	18	19	13	15	29	31.2	12	12	14	14.2	8.3	9	21	26
P081	14	18	16	16	11	12	19	20	10	14	30	33.2	13	14	15	15.2	7	8	21	23
P082	15	17	14	16	11	11	17	20	14	15	30	33.2	15	16	13	15.2	7	9	21	22
P083	15	17	16	16	10	11	19	19	13	16	28	28	16	16	15.2	15.2	6	9	23	26

Sample Name	D3		vWA		D16		D2		D8		D21		D18		D19		TH01		FGA	
P086	14	15	16	16	11	11	18	18	14	14	32.2	35.2	14	19	14	14	6	7	20	22
P087	15	17	14	16	11	11	23	23	11	13	30.2	31.2	15	18	14	14.2	7	7	22	24
P088	16	16	18	18	10	12	20	20	12	17	28	33.2	14	17	12	13	7	9	23	25
P090	15	17	14	14	11	12	17	19	13	15	29	31.2	16	18	14	14.2	8	9	21	22.2
P092	17	18	14	17	11	12	18	19	12	13	30	30.2	14	15	14	16	6	9	21.2	24
P093	15	16	16	17	11	12	22	25	10	10	29	30	14	19	13	14	6	6	22	24
P094	14	15	16	18	9	11	18	24	11	15	31	33.2	16	18	13	15	9	9.3	23	25
S001	16	17	16	17	12	12	23	23	10	12	28	29	13	18	13	13	7	9	21	23
S002	15	16	14	14	11	12	17	24	11	14	29	30	14	14	13	15.2	7	7	25	25
S004	15	15	18	19	10	11	20	24	11	14	29	31.2	14	14	12	16	7	8	22	23
S005	15	17	17	19	8	11	20	22	10	15	29	30.2	15	17	14	14	9	9	22.2	25
S006	15	18	15	17	9	12	18	20	10	14	31.2	33.2	13	15	13	16	6	6	21	25
S007	16	17	14	14	9	12	22	23	15	16	28	30	14	15	15	16	7	9.3	24	26
S008	16	17	16	17	11	11	18	23	14	15	29	33.2	16	18	13	14	6	9	20	21
S009	16	16	14	14	9	12	24	25	13	14	30	31.2	14	15	13	16.2	7	9	21	25
S012	15	15	16	18	11	11	20	20	14	15	28	30	12	15	15	15	8	8	23	26
S013	17	18	15	17	11	13	17	25	14	14	29	30	12	14	12	13	9	9	20	22
S014	16	18	14	18	11	14	22	23	12	15	30	31.2	14	18	13	14.2	7	8	22	23
S015	15	18	16	18	11	13	18	23	13	15	28	28	15	15	13.2	14.2	7	9.3	24	24
S017	16	17	17	17	9	9	19	20	15	15	30	33.2	18	21	14.2	15	6	9.3	22	24
S019	16	17	16	16	8	8	19	20	11	15	28	30	13	16	14.2	16	7	8	24	25
S020	15	15	17	18	9	13	20	23	10	11	28	30	14	15	14.2	15	6	9.3	24	25
S022	15	16	14	17	11	12	25	25	11	15	30	31	17	21	14	15	9.3	9.3	20	21
S023	17	18	14	19	9	11	19	23	11	12	30.2	31.2	12	14	12	14	6	6	19	24
S024	15	16	16	17	9	12	19	19	11	14	29	32.2	15	16	13	16	6	9.3	22	23
S026	16	17	16	18	10	11	17	23	11	12	32.2	32.2	13	15	13	13	8	9	21	22
S027	16	16	15	18	9	11	17	19	12	16	28	29	14	18	13	14	8	8	24	24.2
S030	16	16	16	17	8	11	22	23	11	13	29	29	12	17	15	16	8	8	22.2	25
S032	15	17	17	17	9	11	20	21	12	14	32.2	33.2	14	16	14	14.2	7	9	22	22
S034	16	18	17	19	12	13	20	24	10	14	33.2	33.2	15	15	13	15.2	7	9.3	19	24

Sample Name	D3		vWA		D16		D2		D8		D21		D18		D19		TH01		FGA	
S036	15	15	16	17	11	12	23	23	14	16	30	31.2	14	15	15	15.2	6	8	21	24
S037	11	17	13	16	9	12	19	19	10	15	29	29	13	18	14.2	15	6	7	19	23
S038	14	16	15	18	11	13	19	22	11	15	29	32.2	12	16	14	16.2	7	7	19	25
S039	15	15	14	20	11	13	19	24	12	14	28	33.2	11	16	13	15	6	7	20	24
S040	15	17	16	18	11	11	17	19	14	16	30	31.2	11	15	13	13.2	6	9.3	21	26
S042	15	17	17	18	11	12	18	21	11	13	29	29	13	18	13	16.2	7	7	20	21
S043	16	16	17	18	10	11	24	26	11	11	30	30	10	14	13	13	6	7	24	25
S046	15	15	16	19	8	11	18	19	15	15	28	30	13	17	13	15	6	9	23	24
S048	15	17	18	18	8	11	19	23	10	15	29	31.2	15	15	12.2	16.2	7	9	21	26
S050	15	17	16	17	11	12	18	18	13	16	31.2	32.2	14	17	14	16.2	7	9.3	24	24
S052	16	16	17	18	9	12	18	20	13	13	29	31.2	14	17	13	13	7	8	21	21
S054	14	16	14	15	12	12	21	21	11	13	29	30	16	21	13	14.2	6	8	23	26
S055	16	16	17	17	10	12	20	20	14	16	28	30	15	15	13	14	7	9.3	22	25
S056	16	16	18	18	11	11	21	23	12	16	28	32.2	14	16	14.2	16.2	8	8	22	23
S058	15	17	15	16	9	11	19	23	10	12	30	31.2	14	14	13	14	9.3	9.3	21	21
S059	16	17	17	19	11	12	17	21	10	12	28	32.2	14	17	14.2	14.2	9	9.3	21	25
S060	16	17	18	18	10	11	20	23	14	15	28	31.2	12	16	12	14	6	8	20	21
S061	15	15	17	17	10	11	18	24	14	15	28	33.2	13	18	12	15	7	7	22.2	25
S062	15	15	16	18	9	11	19	24	14	17	30	31.2	15	23	13	14	6	6	20	21
S068	18	18	15	19	9	12	18	20	15	15	28	30.2	14	16	13	14	6	10	19	23
S069	16	16	14	17	11	12	23	25	8	12	29	30.2	14	16	14	17	9	9.3	23	24
S072	15	15	16	17	12	13	23	24	11	14	28	30.2	13	13	13	14	7	7	20	22
S074	12	17	17	18	19	12	24	24	11	15	28	30.2	13	16	14	15	9	9	22	22
S075	18	19	15	18	10	12	19	23	13	14	31.2	33.2	12	19	14	15	8	9	24	25
S079	16	17	16	18	11	12	17	20	15	16	28	29	13	15	14	15.2	6	6	25	26
S084	15	16	18	19	8	9	23	23	9	13	31	31	19	19	14	14	8	9.3	21	25
S085	14	16	14	17	9	11	20	20	14	15	29	30	12	14	14	14	7	7	24	24
S091	15	16	16	17	9	11	18	23	14	16	30.2	32.2	14	17	13	14	9	9	22	24

Table A2.4 STR data for the Kalash population

Sample Name	D3		vWA		D16		D2		D8		D21		D18		D19		TH01		FGA	
K001	15	17	15	17	11	12	23	24	10	11	30	31	12	14	13.2	16	8	9.3	20	20
K003	16	17	15	17	9	13	20	22	11	13	29	33.2	15	18	14	15	9	9.3	18	20
K004	15	17	14	17	11	12	18	22	10	14	29	29	14	15	14	14	8	9.3	22	23
K005	16	16	15	19	9	13	19	22	10	11	29	29	14	18	13.2	15	7	9	18	20
K006	15	17	15	15	9	12	17	18	11	13	30	33.2	12	12	13.2	14	5.3	8.3	21	25
K007	15	15	16	18	10	12	23	26	10	11	30	34.2	14	15	15	16	7.3	9.3	21	22
K008	16	17	15	16	11	12	23	26	11	13	31.2	33.2	18	21	14	15	9.3	9.3	24	24.2
K009	15	17	15	16	11	13	19	20	10	10	29	30	14	16	14	14.2	6	8	20	23
K010	16	17	14	18	11	11	22	22	10	13	29	30	12	14	14	14	8	9.3	20	22
K011	15	16	17	17	12	12	18	23	10	11	29	29	16	22	14	14	9.3	9.3	20	22
K012	15	15	18	19	12	13	23	23	10	13	29	33.2	12	13	14	14	9	9.3	20	20
K013	15	16	15	16	9	12	20	23	10	13	30	33.2	12	13	14	14.2	9	9.3	20	22.2
K021	15	16	16	19	9	11	17	23	11	13	29	33.2	12	15	13	14	9	9.3	20	24.2
K023	15	18	17	19	11	12	17	18	10	14	29	32.2	14	16	12	15	9.3	9.3	24	25
K024	15	17	16	19	9	12	20	23	11	13	29	33.2	14	15	14.2	15	8	9.3	22	25
K025	15	15	17	17	9	11	20	24	10	14	28	30	12	16	14	15	6	9.3	21	23
K026	16	17	15	18	12	13	18	20	10	13	29	30	18	18	14	15	9	9	21	23
K028	17	17	16	17	9	12	18	22	10	13	32	32.2	14	16	14	14.2	6	9.3	21	23
K030	15	18	17	19	10	11	22	23	13	16	29	32	14	15	14	14	8	9	22	23
K031	16	17	17	19	9	12	22	24	10	13	29	32	14	16	14	14	6	6	21	24
K033	15	15	18	19	11	13	20	24	11	14	30	34.2	14	16	13.2	15	8	9	22	25
K034	15	17	18	19	9	12	19	23	13	16	29	32	14	16	14.2	14.2	6	8	18	20
K035	16	17	16	16	9	10	16	23	10	13	30	32.2	16	18	16	16	9	9	20	23
K036	15	16	18	18	11	12	22	23	11	13	29	33.2	14	15	13	14	9	9.3	21	22
K038	16	16	14	15	9	13	21	22	15	15	31.2	32	16	20	14	14.2	7	9.3	18	18
K039	16	16	17	18	12	12	23	23	11	15	29	32.2	16	16	14	16	7	9	21	22
K044	15	17	18	18	10	12	20	24	10	13	29	30.2	14	16	14	14	6	9.3	24	24.2

Sample Name	D3		vWA		D16		D2		D8		D21		D18		D19		TH01		FGA	
K047	16	17	16	16	11	12	17	23	10	10	29	30	14	16	14	14	9	9.3	20	20
K048	15	15	15	17	11	13	23	24	12	15	30	33.2	14	15	14.2	16	6	9.3	22	22
K050	16	17	17	17	10	11	20	20	13	16	29	30	14	17	12	14	9.3	10	23	24
K053	15	15	15	16	10	11	17	23	10	11	30	30.2	14	16	14	16	9	9.3	18	24.2
K054	15	16	16	18	12	12	22	24	13	14	29	30.2	14	15	14	14	9	9.3	22	23
K055	16	17	15	15	11	13	18	25	14	16	29	30	12	19	14	14.2	7	9	20	21
K057	15	16	14	19	9	9	19	23	10	14	26	29	14	15	14	14	6	9	21	24.2
K060	16	16	15	19	11	12	22	25	10	14	30	33.2	15	16	14.2	14.2	6	8	22.2	23
K061	16	16	15	17	8	10	18	23	10	15	30	30	14	19	14.2	14.2	9	9.3	22	22
K062	15	16	17	18	11	12	23	23	13	13	29	31.2	15	15	14.2	14.2	7	8	21	22
K063	16	17	16	17	8	13	19	23	13	16	33.2	33.2	12	15	14	14	6	9	20	21
K064	15	15	16	18	11	12	17	22	16	16	29	29	12	13	14	15.2	9	9.3	21	23
K066	15	15	15	17	9	12	19	24	13	16	30	30	15	15	14	16.2	6	8	20	22
K067	16	17	15	15	11	11	19	25	14	14	30	32	15	16	14	14.2	8	9	21	22
K068	16	17	15	17	11	12	23	24	11	13	29	32	12	15	14.2	15	7	9.3	20	25
K070	15	17	17	17	10	11	22	23	10	13	29	30	14	14	14	14.2	8	9.3	24	25
K071	15	17	17	20	9	11	22	25	10	14	26	29	14	18	13	14	9	9	24.2	25
K072	15	16	17	19	12	13	17	21	10	13	30	34.2	12	18	14.2	15	8	9.3	21	22
K074	16	17	15	19	11	11	19	25	10	15	30	33.2	15	16	14.2	15	6	8	21	22
K075	16	16	15	19	12	13	19	19	13	14	28	29	13	15	15	15	7	8	22	22
K076	17	18	17	18	11	12	23	24	14	14	26	32	14	16	14.2	14.2	9	9.3	22	25
K077	15	16	17	19	9	13	24	26	14	14	29	33.2	12	12	14	14.2	7	8	21	22
K078	15	16	14	19	9	9	23	23	10	13	29	32	14	15	14	14	7	9	22	24.2
K079	15	17	15	18	9	11	23	23	11	16	29	33.2	12	14	13.2	14	6	9.3	20	22
K080	15	16	15	15	10	12	17	22	13	14	29	33.2	12	14	14	14	7	9	18	20
K081	15	15	15	19	12	12	24	24	13	14	28	32	14	19	14.2	16	6	9.3	20	22
K082	15	15	14	17	8	11	22	25	11	14	29	33.2	14	14	12	14.2	6	9.3	22	24.2
K083	16	18	15	19	11	13	18	20	8	13	32	33.2	12	17	14	14	7	7	20	24
K084	16	17	15	15	9	13	23	24	11	13	29	31.2	12	16	14	14.2	6	9.3	21	24
K085	17	17	18	19	9	13	24	26	10	10	29	30	18	21	14	14	9.3	9.3	24	24.2

Sample Name	D3		vWA		D16		D2		D8		D21		D18		D19		TH01		FGA	
K087	16	17	15	17	11	13	24	24	13	16	30	32	15	17	14	14.2	6	9.3	22	22
K088	16	16	15	15	11	13	22	24	13	15	20	33.2	12	18	13	15	6	9	18	22
K089	14	15	14	21	10	12	18	23	13	15	30	31	17	18	14	14.2	6	9.3	23	24
K090	16	18	15	16	11	11	17	23	14	16	29	29	13	18	14	16	9	9.3	22	24
K092	14	16	15	17	11	12	19	23	14	15	29	30	12	14	12	15	9	9.3	23	24.2
K093	16	16	18	19	9	10	17	24	13	16	29	29	16	17	14	16	7	9.3	20	20
K094	15	16	17	19	12	13	19	19	10	14	32	34.2	13	15	13	14	7	9.3	20	25
K097	15	15	18	19	11	12	23	25	13	14	29	33.2	14	16	14	14.2	8	9	18	21
K098	16	16	16	17	9	9	23	25	11	13	26	30	13	14	14	15	7	8	22	25
K102	15	15	18	19	11	13	23	24	10	16	29	30	14	15	14	14	6	7	20	20
K103	15	16	17	19	12	12	20	23	15	15	31.2	33.2	14	15	14	14	9	9	22	22
K104	15	16	14	14	11	12	19	24	13	15	29	30.2	12	13	15	16	7	7	21	22
K107	16	17	15	15	10	13	23	26	13	13	30	33.2	12	14	14	14.2	6	6	20	25
K109	15	17	14	19	12	12	23	24	14	14	29	34.2	12	16	13.2	14.2	6	7	22	22.2
K110	15	15	15	17	8	13	19	23	11	11	31.2	33.2	12	19	14	16	9	9	20	22
K111	15	17	15	19	11	12	20	25	15	16	32	33.2	15	15	12	14.2	6	6	21	22
K112	15	15	17	18	9	11	17	23	10	13	29	31.2	17	19	14	14	9	9.3	18	20
K113	15	17	18	19	10	11	24	24	13	16	26	30	14	19	14	14	6	9	20	24
K114	16	17	18	18	9	11	18	21	10	10	32.2	33.2	16	18	14	14.2	9	9.3	20	25
K116	16	16	19	20	13	13	19	24	13	16	30	33.2	12	14	14	15.2	6	9.3	22	24.2
K117	14	17	15	18	12	13	22	24	11	13	30	30	14	14	14	14	6	9	21	23
K118	15	17	15	16	11	13	18	20	10	14	29	30	13	15	13	15	9	9.3	21	25
K119	15	15	14	15	9	13	19	24	14	16	29	29	14	17	14	14	8	9	20	22
K120	15	18	17	18	9	11	18	23	13	16	29	29	13	16	14	14.2	9	9	21	21
K122	15	15	16	20	8	12	19	25	10	13	29	33.2	14	21	15.2	15.2	9	9.3	22	22
K123	15	16	18	18	11	13	19	20	14	16	30	31.2	15	16	13	14.2	6	9	20	22
K124	16	18	18	18	9	12	19	23	13	14	29	32	12	14	13.2	14.2	6	8	22	24.2
K126	15	17	15	17	11	13	24	26	10	16	30	33.2	12	22	14	14.2	9	9.3	20	24.2
K127	16	17	17	18	9	11	22	24	10	14	29	29	14	23	13.2	14	7	9	22	24.2
K128	16	17	16	16	8	12	18	23	14	14	29	30	14	14	14	14	9	9.3	22	23

Sample Name	D3		vWA		D16		D2		D8		D21		D18		D19		TH01		FGA	
K129	15	17	15	15	9	12	19	19	10	13	29	30.2	19	19	14.2	15	6	9	22	24.2
K130	16	16	15	18	8	10	17	24	10	16	29	30.2	14	16	14	14.2	8	9	18	24.2
K131	15	17	19	19	9	12	22	23	10	13	29	32	16	18	14	14	6	9.3	19	23
K132	15	16	17	19	10	11	24	25	13	14	30	33.2	12	15	14	14	9	9.3	22	24
K133	15	17	17	17	11	12	23	25	10	13	29	30	12	15	14	14	8	9.3	21	22
K134	15	19	15	18	8	9	19	25	10	13	29	32	15	18	14	14	8	9	20	21
K135	16	17	15	17	9	12	17	24	11	13	29	33.2	15	22	14	17	8	9	22	22
K136	15	18	15	15	9	12	23	25	10	10	30.2	32	16	18	15	16	9	9.3	21	22.2
K137	16	16	14	18	11	12	20	23	13	13	29	30.2	14	14	14	14.2	6	8	22	22
K138	16	17	17	17	12	13	17	22	13	15	29	29	12	14	14	15	9	9.3	22	25
K139	16	18	13	18	12	12	19	23	13	14	29	30.2	12	14	14	14.2	6	9	22	24.2
K140	15	15	16	19	8	13	23	24	10	13	29	30	14	18	14.2	14.2	9	9	22	25
K141	17	17	15	16	12	12	24	26	10	13	30	30	21	22	14	14.2	7	9.3	21	22
K142	16	18	13	17	12	13	19	24	10	14	29	29	14	19	13.2	15	9.3	9.3	22	24.2
K143	16	17	18	19	11	11	19	20	10	16	29	30	12	22	14	14	9	9	22	24.2
K144	15	15	15	16	11	13	23	25	11	16	29	32	15	22	14	16	6	9	18	22
K145	16	18	14	18	9	12	17	24	10	16	29	29	14	16	14	14.2	6	9	22	24.2
K147	15	16	13	16	9	13	19	22	10	16	29	30	15	16	14	16	9	9.3	21	25
K148	16	17	16	19	12	12	20	26	11	13	32.2	33.2	13	15	13	14	6	9	18	22
K149	16	16	15	19	8	9	20	23	13	14	29	33.2	14	14	13.2	14	7	8	22	24.2
K150	15	16	15	17	12	13	19	25	13	16	33.2	33.2	16	23	14	14	6	9	20	24
K152	15	16	15	18	12	13	18	24	13	14	30	33.2	12	15	14	14.2	9	9.3	20	22
K153	16	17	16	17	9	11	18	20	13	15	28	32	15	16	14	14	9	9.3	22	24.2
K154	16	16	18	19	11	13	23	23	10	15	30	31.2	14	15	14	14.2	9.3	9.3	22	22
K155	16	16	15	15	12	13	24	26	10	11	29	30	18	22	14	15	6	7	20	24.2
K156	16	18	18	19	10	12	24	24	10	11	29	33.2	14	14	13.2	14	7	9	20	23
K157	15	17	17	17	8	13	23	24	10	10	29	30	14	18	14	14.2	6	8	22	24

REFERENCES

Ahn, S.-M., Kim, T.-H., Lee, S., Kim, D., Ghang, H., Kim, D.-S., Kim, B.-C., Kim, S.-Y., Kim, W.-Y., Kim, C., Park, D., Lee, Y. S., Kim, S., Reja, R., Jho, S., Kim, C. G., Cha, J.-Y., Kim, K.-H., Lee, B., Bhak, J., & Kim, S.-J. (2009) 'The first Korean genome sequence and analysis: Full genome sequencing for a socio-ethnic group'. Genome Res., 19, (9), p.1622-1629

Alaeddini, R., Walsh, S. J., & Abbas, A. (2010) 'Forensic implications of genetic analyses from degraded DNA – A review'. Forensic Sci. Int. Gen., 4, p.148-157

Alshamali, F., Alkhayat, A. Q., Budowle, B., & Watson, N. D. (2005) 'STR population diversity in nine ethnic populations living in Dubai'. Forensic Sci. Int., 152, (2-3), p.267-279

Balding, D. J. (1999) 'When can a DNA profile be regarded as unique?'. Sci. Justice, 39, (4), p.257-260

Balding, D. J., & Nichols, R. A. (1994) 'DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands'. Forensic Sci. Int., 64, (2-3), p.125-140

Balloux, F., & Lugon-Moulin, N. (2002) 'The estimation of population differentiation with microsatellite markers'. Mol. Ecol., 11, p.155-165

Bamshad, M., Wooding, S., Salisbury, B. A., & Claiborne Stephens, J. (2004) 'Deconstructing the relationship between genetics and race'. Nature, 5, p.598-609

Barbaro, A., Falcone, G., & Barbaro, A., (2000) 'DNA typing from hair shaft'. Prog. For. Genet., 8, p.523.525

Barbujani, G., & Colonna, V. (2010) 'Human genome diversity: frequently asked questions'. Trends Genet., 26, p.285-295

Barbujani, G., Magagni, A., Minch, E., & Cavalli-Sforza, L. L. (1997) 'An apportionment of human DNA diversity'. Proc. Natl. Acad. Sci. USA, 94, (9), p.4516-4519

Barni, F., Berti, A., Pianese, A., Boccellino, A., Miller, M. P., Caperna, A., & Lago, G. (2007) 'Allele frequencies of 15 autosomal STR loci in the Iraq population with comparisons to other populations from the middle-eastern region'. Forensic Sci. Int., 167, (1), p.87-92

Bauchet, M., McEvoy, B., Pearson, L. N., Quillen, E. E., Sarkisian, T., Hovhannesyan, K., Deka, R., Bradley, D. G., & Shriver, M. D., (2007) 'Measuring European population stratification with Microarray Genotype Data'. Am. J. Hum. Genet., 80, (5), p.948-956

Baye, T. M., Tiwari, H. K., Allison, D. B., & Go, R. C. (2009) 'Database mining for selection of SNP markers useful in admixture mapping'. BioData Min., 2, (1). p.1-8

Behar, D. M., Villems, R., Soodyall, H., Blue-Smith, J., Pereira, L., Metspalu, E., Scozzari, R., Makkan, H., Tzur, S., Comas, D., Bertranpetit, J., Quintana-Murci, L., Tyler-Smith, C., Spencer Wells, R., & Rosset, S., (2008) 'The Dawn of Human Matrilineal Diversity'. Am. J. Hum. Genet., 82, (5), p.1130-1140

Bender, K., Schneider, P. M., & Rittner, C. (2000) 'Application of mtDNA sequence analysis in forensic casework for the identification of human remains'. Forensic Sci. Int., 113, (1-3), p.103-107

Bhopal, R. & Donaldson, L. (1998) 'White, European, Western, Caucasian, or what? Inappropriate labelling in research on race, ethnicity, and health'. Am. J. Public Health, 88, p.1303-1307

Binda, S., Borer, U. V., Gehrig, C., Hochmeister, M., & Budowle, B. (2000) 'Swiss Caucasian population data for the STR loci D2S1338 and D19S433 using the AmpFℓSTR SGM plus PCR amplification kit'. Forensic Sci. Int., 108, (2), p.117-120

Botstein, D., White, R. L., Skolnick, M., & Davis, R. W. (1980) 'Construction of a genetic linkage map in man using restriction fragment length polymorphisms'. Am. J. Hum. Genet., 32, (3), p.314-331

Bowcock, A. M., Ruiz-Linares, A., Tomfohrde, J., Minch, E., Kidd, J. R., & Cavalli-Sforza, L. L. (1994) 'High resolution of human evolutionary trees with polymorphic microsatellites'. Nature, 368, (6470), p.455-457

Brenner, C. H. (1998) 'Difficulties in the estimation of ethnic affiliation'. Am. J. Hum. Genet., 62, (6), p.1558-1560

Briggs, A. W., Good, J. M., Green, R. E., Krause, J., Maricic, T., Stenzel, U., Lalueza-Fox, C., Rudan, P., Brajković, D., Kućan, Ž., Gušic, I., Schmitz, R., Doronichev, V. B., Golovanova, L. V., de la Rasilla, M., Fortea, J., Rosas, A., & Pääbo, S., (2009) 'Targeted retrieval and analysis of five Neandertal mtDNA genomes'. Science, 325, p.318-321

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'. Am. J. Hum. Genet., 62, p.1408-1415

Buckleton, J. S., Curran, J. M., & Walsh, S. J. (2006) 'How reliable is the subpopulation model in DNA testimony?' Forensic Sci. Int., 157, p.144-148

Budowle, B., Allard, M. W., Wilson, M. R., & Chakraborty, R. (2003) 'Forensics and Mitochondrial DNA: Applications, Debates, and Foundations'. Annu. Rev. Genomics Hum. Genet. 4, p.119-141

Budowle, B., Wilson, M. R., DiZinno, J. A., Stauffer, C., Fasano, M. A., Holland, M. M., & Monson, K. L. (1999) 'Mitochondrial DNA regions HVSI and HVSII population data'. Forensic Sci. Int., 103, (1), p.23-35

Butler, J. 2001, Forensic DNA Typing Academic Press, London.

Butler, J. (2006) 'Genetics and Genomics of Core Short Tandem Repeat Loci Used in Human Identity Testing'. J. For. Sci., 51, (2), p.253-265

Butler, J., Shen, Y., & McCord, B. R. (2003) 'The development of reduced size STR amplicons as tools for analysis of degraded DNA'. J. For. Sci., 48, (5), p.1054-1064

Cadenas, A. M., Regueiro, M., Gayden, T., Singh, N., Zhivotovsky, L. A., Underhill, P. A., & Herrera, R. J. (2007) 'Male amelogenin dropouts: phylogenetic context, origins and implications'. Forensic Sci. Int., 166, p.155-163

Cann, H.M., de Toma, C., Cazes, L., Legrand, M-F., Morel, V., Piouffre, L., Bodmer, J., Bodmer, W. F., Bonne-Tamir, B., Cambon-Thomsen, A., Chen, Z., Chu, J., Carcassi, C., Contu, L., Du, R., Excoffier, L., Ferrara, G. B., Friedlaender, J. S., Groot, H., Gurwitz, D., Jenkins, T., Herrera, R. J., Huang, X., Kidd, J., Kidd, K. K., Langaney, A., Lin, A. A., Qasim Mehdi, S., Parham, P., Piazza, A., Pia Pistillo, M., Qian, Y., Shu, Q., Xu, J., Zhu, S., Weber, J. L., Greely, H. T., Feldman, M. W., Thomas, G., Dausset, J., & Cavalli-Sforza, L. L. (2002) 'A human genome diversity cell line panel'. Science, 296, (5566), p.261-262

Cann, R. L., Stoneking, M., & Wilson, A. C. (1987) 'Mitochondrial DNA and human evolution'. Nature, 325, (6099), p.31-36

Cavalli-Sforza, L. L., Menozzi, P., & Piazza, A. 1994, The History and Geography of Human Genes Princeton University Press, Princeton.

Chakraborty, R. (1992) 'Sample size requirements for addressing the population genetic issues of forensic use of DNA typing'. Hum. Biol., 64, (2), p.141-159

Chakraborty, R., Shaw, M., & Schull, W. J. (1974) 'Exclusion of paternity: the current state of the art'. Am. J. Hum. Genet., 26, (4), p.477-488

Clark, D., Hadi, S., Iyengar, A., Smith, J., Garg, V & Goodwin, W. (2009) 'STR data for the AmpFℓSTR® SGM Plus® loci from two South Asian populations', Legal Med., 11, (2), p.97-100

Coble, M. D. & Butler, J. M. (2005) 'Characterization of new MiniSTR Loci aid analysis of degraded DNA'. J. Forensic. Sci., 50, (1), p.43-53

Collins, P. J., Hennessy, L. K., Leibelt, C. S., Roby, R. K., Reeder, D. J., & Foxall, P. A., (2004) 'Developmental validation of a single-tube amplification of the 13 CODIS STR loci, D2S1338, D19S433, and Amelogenin: The AmpFℓSTR® Identifiler® PCR amplification kit', J. Forensic Sci., 49, (6), p.1-13

Cotton, E. A., Allsop, R. F., Guest, J. L., Frazier, R. R. E., Koumi, P., Callow, I. P., Seager, A., & Sparkes, R. L., (2000) 'Validation of the AmpFℓSTR® SGM Plus® system for use in forensic casework'. Forensic Sci. Int., 112, p.151-161

Curran, J. M., Buckleton, J. S. (2007) 'The appropriate use of subpopulation corrections for differences in endogamous communities'. Forensic Sci. Int., 168, p.106-111

Curran, J. M., Buckleton, J. S., Triggs, C. M., (2003) 'What is the magnitude of the subpopulation effect?' Forensic Sci. Int., 135, p.1-8

Deshpande, O., Batzoglou, S., Feldman, M. W., & Cavalli-Sforza, L. L. (2009) 'A serial founder effect model for human settlement out of Africa'. Proc. R. Soc. B., 276, p.291-300

Devlin, B., Risch, N., & Roeder, K. (1993) 'Statistical evaluation of DNA fingerprinting: a critique of the NRC's report'. Science, 259, (5096), p.748-749

Divne, A. M. & Allen, M. (2005) 'A DNA microarray system for forensic SNP analysis'. Forensic Sci. Int., 154, (2-3), p.111-121

Dixon, L. A., Dobbins, A. E., Pulker, H. K., Butler, J. M., Vallone, P. M., Coble, M. D., Parson, W., Berger, B., Grubwieser, P., Mogensen, H. S., Morling, N., Nielsen, K., Sanchez, J. J., Petkovski, E., Carracedo, A., Sanchez-Diz, P., Ramos-Luis, E., Briōn, M., Irwin, J. A., Just, R. S., Loreille, O., Parsons, T. J., Syndercombe-Court, D., Schmitter, H., Stradmann-Bellinghausen, B., Bender, K., & Gill, P. (2006) 'Analysis of artificially degraded DNA using STRs and SNPs--results of a collaborative European (EDNAP) exercise'. Forensic Sci. Int., 164, (1), p.33-44

Dyson, S. M. (1998) '"Race", ethnicity and haemoglobin disorders'. Soc. Sci. Med., 47, (1), p.121-131

Edwards, A. W. F. (2003) 'Human genetic diversity: Lewontin’s fallacy'. BioEssays, 25, p.798-801

Edwards, A., Civitello, A., Hammond, H. A. & Caskey, C. T. (1991) 'DNA Typing and Genetic Mapping with Trimeric and Tetrameric Tandem Repeats'. Am. J. Hum. Genet., 49, p.746-756

Edwards, A., Hammond, H. A., Jin, L., Caskey, C. T., & Chakraborty, R. (1992) 'Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups'. Genomics, 12, (2), p.241-253

El Mouzan, M. I., Al Salloum, A. A., Al Herbish, A. S., Qurachi, M. M., & Al Omar, A. A. (2010) 'Consanguinity and major genetic disorders in Saudi children: a community-based cross-sectional study'. Ann. Saudi Med., 28, (3), p.169-173

Evett, I. W., Gill, P. D., Scrange, J. K., & Weir, B. S. (1996a) 'Establishing the robustness of short-tandem-repeat statistics for forensic applications'. Am. J. Hum. Genet., 58, (2), p.398-407

Evett, I. W., Lambert, J. A., Buckleton, J. S., & Weir, B. S. (1996b) ‘Statistical analysis of a large file of data from STR profiles of British Caucasians to support forensic casework’, Int. J. Legal Med., 109, p.173-177

Excoffier, L. (2002) 'Human demographic history: refining the recent African origin model'. Curr. Opin. Genet. Dev., 12, p.675-682

Excoffier, L., Laval, G., & Schneider, S. (2005) 'Arlequin ver. 3.1: An integrated software package for population genetics data analysis'. Evol. Bioinform. Online, 1, p.47-50

Excoffier, L., Smouse, P. E., & Quattro, J. M. (1992) 'Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data'. Genetics, 131, (2), p.479-491

Firasat, S., Khaliq, S., Mohyuddin, A., Papaioannou, M., Tyler-Smith, C., Underhill, P. A., & Ayub, Q. (2007) ‘Y-Chromosomal evidence for a limited Greek contribution to the Pathan population of Pakistan’, Eur. J. Hum. Genet., 15, p.121-126

Fitzpatrick, B. M. (2009) 'Power and sample size for nested analysis of molecular variance'. Mol. Ecol., 18, p.3961-3966

Fondevila, M., Phillips, C., Santos, C., Freire Aradas, A., Vallone, P. M., Butler, J. M., Lareu, M. V., & Carracedo, A. (2012) 'Revision of the SNPforID 34-plex forensic ancestry test: Assay enhancements, standard reference sample genotypes and extended population studies'. Forensic Sci. Int. Genet., (IN PRESS) http://dx.doi.org/10.1016/ j.fsigen.2012.06.007

Foreman, L. A. & Evett, I. W. (2001) 'Statistical analyses to support forensic interpretation for a new ten-locus STR profiling system'. Int. J. Legal Med., 114, (3), p.147-155

Foreman, L. A., & Lambert, J. A., (2000) 'Genetic differentiation within and between four UK ethnic groups’. Forensic Sci. Int., 114, (1), p.7-20

Foreman, L. A., Lambert, J. A., & Evett, I. W. (1998) 'Regional genetic variation in Caucasians'. Forensic Sci. Int., 95, (1), p.27-37

Foreman, L. A., Smith, A. F. M., & Evett, I. W. (1997) 'A Bayesian approach to validating STR multiplex databases for use in forensic casework', Int. J. Legal Med., 110, p.244-250

Foster, M. W. & Sharp, R. R. (2002) 'Race, ethnicity, and genomics: social classifications as proxies of biological heterogeneity'. Genome Res., 12, (6), p.844-850

Frégeau, C. J., & Fourney, R. M. (1993) 'DNA typing with fluorescently tagged short tandem repeats: a sensitive and accurate approach to human identification'. Biotechniques, 15, (1), p.100-119

Frudakis, T., Venkateswarlu, K., Thomas, M. J., Gaskin, Z., Ginjupalli, S., Gunturi, S., Ponnuswamy, V., Natarajan, S., & Nachimuthu, P. K. (2003) 'A classifier for the SNP-based inference of ancestry'. J. Forensic Sci., 48, (4), p.771-782

Gaines, M. L., Wojtkiewicz, P. W., Valentine, J. A., & Brown, C. L. (2002) 'Reduced volume PCR amplification reactions using the AmpFℓSTR Profiler Plus kit'. J. Forensic Sci., 47, (6), p.1224-1237

Gehrig, C., Hochmeister, M., Borer, U. V., Dirnhofer, R., & Budowle, B. (1999) 'Swiss Caucasian population data for 13 STR loci using AmpFℓSTR profiler plus and cofiler PCR amplification kits'. J. Forensic Sci., 44, (5), p.1035-1038

Gill, P. (2001) 'An assessment of the utility of single nucleotide polymorphisms (SNPs) for forensic purposes'. Int. J. Leg. Med., 114, p.204-210

Gill, P., Foreman, L., Buckleton, J. S., Triggs, C. M., & Allen, H. (2003) 'A comparison of adjustment methods to test the robustness of an STR DNA database comprised of 24 European populations'. Forensic Sci. Int., 131, (2-3), p.184-196

Gill, P., Sullivan, K., & Werrett, D. J. (1989) 'The analysis of hypervariable DNA profiles: problems associated with the objective determination of the probability of a match'. Hum. Genet., 85, p.75-79

Golenberg, E. M., Bickel, A. & Weihs, P. (1996) 'Effect of highly fragmented DNA on PCR'. Nucleic Acids Res., 24, (24), p.5026-5033

Goodman, A. H. (2000) 'Why genes don't count (for racial differences in health)'. Am. J. Public Health, 90, (11), p.1699-1702

Goodwin, W., Linacre, A., & Hadi, S. (2011) 'An introduction to forensic genetics'. 2nd edition, Chichester: Wiley.

Graham, E. A. M., Turk, E. E., & Rutty, G. N. (2008) 'Room temperature DNA preservation of soft tissue for rapid DNA extraction: an addition to the disaster victim identification investigators toolkit?' Forensic Sci. Int. Gen., 2, p.29-34

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H., Hansen, N. F., Durand, E. Y., Malaspinas, A. S., Jensen, J. D., Marques-Bonet, T., Alkan, C., Prüfer, K., Meyer, M., Burbano, H. A., Good, J. M., Schultz, R., Aximu-Petri, A., Butthof, A., Höber, B., Höffner, B., Siegemund, M., Weihmann, A., Nusbaum, C., Lander, E. S., Russ, C., Novod, N., Affourtit, J., Egholm, M., Verna, C., Rudan, P., Brajkovic, D., Kucan, Z., Gusic, I., Doronichev, V. B., Golovanova, L. V., Lalueza-Fox, C., de la Rasilla, M., Fortea, J., Rosas, A., Schmitz, R. W., Johnson, P. L., Eichler, E. E., Falush, D., Birney, E., Mullikin, J. C., Slatkin, M., Nielsen, R., Kelso, J., Lachmann, M., Reich, D., & Pääbo, S. (2010) 'A draft sequence of the Neandertal genome'. Science, 328, (5979), p.710-722

Green, R. L., Lagacé, R. E., Oldroyd, N. J., Hennessy, L. K., & Mulero, J. J. (2012) 'Developmental validation of the AmpFℓSTR® NGM SElect™ PCR Amplification Kit: A next-generation STR multiplex with the SE33 locus'. Forensic. Sci. Int. Genet., (in press), http://dx.doi.org/10.1016/j.fsigen.2012.05.012

Greenhalgh, M., Burridge, F., & Willott, G. (1992) 'Experiences with single locus DNA probes in casework'. For. Sci. Int., 57, p.29-37

Guo, S. W. & Thompson, E. A. (1992) 'Performing the exact test of Hardy-Weinberg proportion for multiple alleles'. Biometrics, 48, (2), p.361-372

Gusmao, L., Brion, M., Gonzalez-Neira, A., Lareu, M., & Carracedo, A. (1999) 'Y chromosome specific polymorphisms in forensic analysis'. Leg. Med. (Tokyo), 1, (2), p.55-60

Hadi, C. A., Simsek, F., Katyrcy, N., & Tasdelen, B. (2004) 'STR data for the AmpFℓSTR SGM Plus from the eastern and western sections of Mediterranean region of Turkey'. Forensic Sci. Int., 142, (1), p.55-57

Halder, I., Shriver, M., Thomas, M., Fernadez, J. R., & Frudakis, T. (2008) 'A Panel of Ancestry Informative Markers for Estimating Individual Biogeographical Ancestry and Admixture From Four Continents: Utility and Applications'. Hum. Mutat., 29, (5), p.648-658

Hammer, M. F., Karafet, T., Rasanayagam, A., Wood, E. T., Altheide, T. K., Jenkins, T., Griffiths, R. C., Templeton, A. R., & Zegura, S. L. (1998) 'Out of Africa and back again: nested cladistic analysis of human Y chromosome variation'. Mol. Biol. Evol., 15, (4), p.427-441

Harding, R. M., Healy, E., Ray, A. J., Ellis, N. S., Flanagan, N., Todd, C., Dixon, C., Sanjantila, A., Jackson, I. J., Birch-Machin, M. A., & Rees, J. L., (2000) 'Evidence for Variable Selective Pressures at MC1R'. Am. J. Hum. Genet., 66, p.1351-1361

Harding, R. M. & McVean, G. (2004) 'A structured ancestral population for the evolution of modern humans'. Curr. Opin. Genet. Dev., 14, (6), p.667-674

Havas, D., Jeran, N., Efremovska, L., Dordevic, D., & Rudan, P. (2007) 'Population genetics of 15 AmpFℓSTR Identifiler loci in Macedonians and Macedonian Romani (Gypsy)'. Forensic Sci. Int., 173, (2-3), p.220-224

Higuchi, R., von Beroldingen, C. H., Sensabaugh, G.F., & Erlich, H. A., (1988) 'DNA typing from single hairs'. Nature, 332, (6164), p.543-546

Hoyle, R. (1998) 'Forensics. The FBI's national DNA database', Nat. Biotechnol., 16, (11), p.987

Hunter, K. (July 1998) 'A new direction in DNA?', Crim. Law Rev., p.478-480

Hussain, R. & Bittles, R.J., (1999) 'Consanguineous marriage and differentials in age at marriage, contraceptive use and fertility in Pakistan', J. Biosoc. Sci., 31, 121-138

Hutchinson, J. & Smith, A. 1996, Ethnicity Oxford University Press, Oxford.

Iida, R. & Kishi, K. (2005) 'Identification, characterization and forensic application of novel Y-STRs'. Leg. Med. (Tokyo), 7, (4), p.255-258

Jarne, P. & Lagoda, P. J. L. (1996) 'Microsatellites, from molecules to populations and back'. Trends Ecol. Evol., 11, (10), p.424-429

Jeffreys, A. J., Wilson, V., & Thein, S. L. (1985) 'Individual-specific 'fingerprints' of human DNA'. Nature, 316, (6023), p.76-79

Jobling, M. A. & Gill, P. (2004) 'Encoded Evidence: DNA in Forensic Analysis'. Nature, 5, p.739-751

Jobling, M. A. & Tyler-Smith, C. (1995) ' Fathers and sons: the Y chromosome and human evolution'. Trends Genet., 11, (11), p.449-456

Jones, D. A. (1972) 'Blood samples: probability of discrimination'. J. Forensic Sci. Soc., 12, (2), p.355-359

Junge, A., Verheesen, M., & Madea, B. (2001) 'Genetic variation and population genetic data of the short tandem repeat locus D8S320'. Forensic Sci. Int., 119, (1), p.11-16

Kaessman, H., Wiebe, V., Weiss, G., & Pääbo, S. (2001) 'Great ape DNA sequences reveal a reduced diversity and an expansion in humans'. Nature Genetics, 27, p.155-156

Kashyap, V. K., Guha, S., Sitalaximi, T., Bindu, G. H., Hasnain, S. E., & Trivedi, R. (2006a) 'Genetic structure of Indian populations based on fifteen autosomal microsatellite loci'. BMC. Genet., 7, p.28

Kashyap, V. K., Sahoo, S., Sitalaximi, T., & Trivedi, R. (2006b) 'Deletions in the Y-derived amelogenin gene fragment in the Indian population'. BMC. Med. Genet., 7, (37), doi:10.1186/1471-2350-7-37

Kayser, M. & de Kniff, P. (2011) 'Improving human forensics through advances in genetics, genomics and molecular biology'. Nat. Genet., 12, p.179-192

Kayser, M. & Schneider, P. M. (2009) 'DNA-based prediction of human externally visible characteristics in forensics: Motivations, scientific challenges, and ethical considerations'. Forensic Sci. Int. - Genetics, 3, p.154-161

Kelkar, Y. D., Tyekucheva, S., Chiaromonte, F., & Makova, K. D. (2008) 'The genome-wide determinants of human and chimpanzee microsatellite evolution'. Genome Res., 18, (1), p.30-38

Kidd, K. K., Kidd, J. R., Speed, W. C., Fang, R., Furtado, M. R., Hyland, F. C., & Pakstis, A. J. (2012) 'Expanding data and resources for forensic use of SNPs in individual identification'. Forensic Sci. Int. Genet., 6, (5), p.646-652

Kimpton, C. P., Gill, P., Walton, A., Urquhart, A., Millican, E. S., & Adams, M. (1993) 'Automated DNA profiling employing multiplex amplification of short tandem repeat loci'. Genome Res., 3, p.13-22

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

Kivisild, T., Rootsi, S., Metspalu, M., Mastana, S., Kaldma, K., Parik, J., Metspalu, E., Adojaan, M., Tolk, H. V., Stepanov, V., Golge, M., Usanga, E., Papiha, S. S., Cinnioglu, C., King, R., Cavalli-Sforza, L., Underhill, P. A., & Villems, R. (2003) 'The genetic heritage of the earliest settlers persists both in Indian tribal and caste populations'. Am. J. Hum. Genet., 72, (2), p.313-332

Klintschar, M., Furedi, S., Egyed, B., Reichenpfader, B., & Kleiber, M. (2003) 'Estimating the ethnic origin (EEO) of individuals using short tandem repeat loci of forensic relevance'. International Congress Series (1239), p.53-56

Kosoy, R., Nassir, R., Tian, C., White, P. A., Butler, L. M., Silva, G., Kittles, R., Alarcon-Riquelme, M. E., Gregersen, P. K., Belmont, J. W., De La Vega, F. M., & Seldin, M. F. (2009) 'Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America'. Hum. Mutat., 30, (1), p.69-78

Kruglyak, S., Durrett, R. T., Schug, M. D., & Aquadro, C. F. (1998) 'Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations'. Proc. Natl. Acad. Sci., 95, p.10774-10778

Lai, Y. & Sun, F. (2003) 'The relationship between microsatellite slippage mutation rate and the number of repeat units'. Mol. Biol. Evol., 20, (12), p.2123-2131

Lander, E. S., & Budowle, B. (1994) 'DNA fingerprinting, dispute laid to rest'. Nature, 371, p.735-738

Landsteiner, K. (1900) 'Zur Kenntnis der antifermentativen, lytisichen und agglutinierenden Wirkungen des Blutserums und der Lymphe'. Zentralbl. Bakteriol., 27, p.357-362

Lao, O., Duijn, K., Kersbergen, P., de Knijff, P., & Kayser, M. (2006) 'Proportioning whole-genome single-nucleotide-polymorphism diversity for the identification of geographic population structure and genetic ancestry'. Am. J. Hum. Genet., 78, (4), p.680-690

Lao, O., de Gruijter, J. M., Duijn, K., Navarro, A., & Kayser, M. (2007) 'Signatures of positive selection in genes associated with human skin pigmentation as revealed from analyses of single nucleotide polymorphisms'. Ann. Hum. Genet., 71, (3), p.354-369

Law, B., Buckleton, J. S., Triggs, C. M., & Weir, B. S. (2003) 'Effects of population structure and admixture on exact tests for association between loci'. Genetics, 164, (1), p.381-387

Lawson Handley, L. J., Manica, A., Goudet, J. & Balloux, F. (2007) 'Going the distance: human population genetics in a clinal world'. Trends Genet., 23, (9), p.432-439

Lee, S. S., Mountain, J., & Koenig, B. A. (2001) 'The meanings of "race" in the new genomics: implications for health disparities research'. Yale J. Health Policy Law Ethics, 1, p.33-75

Levins, R. (1969) 'Some demographic and genetic consequences of environmental heterogeneity for biological control' Bull. Entomol. Soc. Am., 15, p.237-240

Levinson, G., & Gutman, G. A., (1987) 'Slipped-strand mispairing: a major mechanism for DNA sequence evolution'. Mol. Biol. Evol., 4, (3), p.203-221

Lewontin, R. C., (1972) 'The apportionment of human diversity'. Evol. Biol., 6, p.381-398

Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., Cann, H. M., Barsh, G. S., Feldman, M., Cavalli-Sforza, L. L., & Myers, R. M. (2008) 'Worldwide human relationships inferred from genome-wide patterns of variation'. Science, 319, p.1100-1104

Li, J., Zhang, M. Q., & Zhang, X. (2006) 'A New Method for Detecting Human Recombination Hotspots and Its Applications to the HapMap ENCODE Data'. Am. J. Hum. Genet., 79, p.628-639

Lillie-Blanton, M., & Laveist, T. (1996) 'Race/ethnicity, the social environment, and health'. Soc. Sci. Med., 43, (1), p.83-91

Linz, B., Balloux, F., Moodley, Y., Manica, A., Liu, H., Roumagnac, P., Falush, D., Stamer, C., Prugnolle, F., van der Merwe, S. W., Yamaoka, Y., Graham, D. Y., Perez-Trallero, E., Wadstrom, T., Suerbaum, S. & Achtman, M. (2007) 'An African origin for the intimate association between humans and *Helicobacter pylori*'. Nature, 445, p.915-918

Liu, F., van Duijn, K., Vingerling, J. R., Hofman, A., Uitterlinden, A. G., Janssens, A. C. J. W., & Kayser, M. (2009) 'Eye color and the prediction of complex phenotypes from genotypes'. Curr. Biol., 19, (5), p.R192-R193

Liu, H., Prugnolle, F., Manica, A., & Balloux, F. (2006) 'A geographically explicit genetic model of worldwide human-settlement history'. Am. J. Hum. Genet., 79, p.230-237

Liu, Y. C., Hao, J. P., Yan, J. G., Tang, H., Wang, J., Ren, H., & Ren, J. C. (2006) 'Polymorphisms analysis of mitochondrial DNA in coding area'. Fa. Yi. Xue. Za Zhi., 22, (1), p.45-47

Loeschcke, V., Tomiuk, J., & Jain, S. K. (1994) 'Conservation Genetics', 1st ed., Birkhäuser, Basel

Loring Brace, C., Tracer, D. P., Yaroch, L. A., Robb, J., Brandt, K. & Nelson, A. R. (1993) 'Clines and clusters versus “Race:” a test in Ancient Egypt and the case of a death on the Nile'. Yearb. Phys. Anthropol., 36, p.1-31

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

Maher, B. (2012) 'The human encyclopaedia'. Nature, 489, p.46-48

Manica, A., Prugnolle, F., & Balloux, F. (2005) 'Geography is a better determinant of human genetic differentiation than ethnicity'. Hum. Genet., 118, (3-4), p.366-371

Mansoor, A., Mazhar, K., Khaliq, S., Hameed, A., Rehman, S., Siddiqi, S., Papaioannou, M., Cavalli-Sforza, L. L., Qasim Mehdi, S., & Ayub, Q. (2004) 'Investigation of the Greek ancestry of populations from Northern Pakistan'. Hum. Genet., 114, p.484-490

Marian, C., Anghel, A., Bel, S. M., Ferencz, B. K., Ursoniu, S., Dressler, M., Popescu, O., & Budowle, B. (2007) 'STR data for the 15 AmpFℓSTR identifiler loci in the Western Romanian population'. Forensic Sci. Int., 170, (1), p.73-75

Maruyama, S., Minaguchi, K., Takezaki, N., & Nambiar, P. (2008) 'Population data on 15 STR loci using AmpFℓSTR Identifiler kit in a Malay population living in and around Kuala Lumpur, Malaysia'. Leg. Med. (Tokyo)

Miller, C. R., Joyce, P., & Waits, L. P. (2002) 'Assessing allelic dropout and genotype reliability using maximum likelihood'. Genetics., 160, p.357-366

Morton, N. E., Collins, A., & Balazs, I. (1993) 'Kinship bioassay on hypervariable loci in Blacks and Caucasians'. Proc. Natl. Acad. Sci., 90, p.1892-1896

Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., & Erlich, H., (1986) 'Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction'. Cold Spring Harb. Symp. Quant. Biol., 51, (1), p.263-273

Nachman, M. W., & Crowell, S. L., (2000) 'Estimate of the Mutation Rate per Nucleotide in Humans'. Genetics, 156, p.297-304

Nakamura, Y., Leppert, M., O'Connell, P., Wolff, R., Holm, T., Culver, M., Martin, C., Fujimoto, E., Hoff, M., Kumlin, E., & White, R. (1987) 'Variable number of tandem repeat (VNTR) markers for human gene mapping'. Science, 235, p.1616-1622

National Policing Improvement Agency (NPIA), (2012) 'Statistics'. [online] Available at: http://www.npia.police.uk/en/13338.htm [Accessed 2 September 2012]

National Research Council (1992), 'DNA Technology in Forensic Science', Washington D. C.: National Academy Press

National Research Council (1996), 'The Evaluation of Forensic DNA Evidence', Washington D. C.: National Academy Press

Novembre, J., & Di Renzo, A. (2009) 'Spatial patterns of variation due to natural selection in humans'. Genetics, 10, p.745-755

Novembre, J., & Ramachandran, S. (2011) 'Perspectives on human population structure at the cusp of the sequencing era'. Annu. Rev. Genomics Hum. Genet., 12, p.245-274

Nussbaumer, C., Hanslik, S., Fichtinger, M., & Bauer, G. (2001) 'STR data for the AmpFℓSTR SGM plus from a regional population of Austria'. Forensic Sci. Int., 122, (2-3), p.181-183

Overall, A. D. J., (2009) 'The influence of the Wahlund Effect on the consanguinity hypothesis: consequences for recessive disease incidence in a socially structured Pakistani population'. Hum. Hered., 67, p.140-144

Overall, A. D. J., & Nichols, R. A. (2001) 'A method for distinguishing consanguinity and population substructure using multilocus genotype data'. Mol. Biol Evol., 18, (11), p.2048-2056

Parson, W., Parsons, T. J., Scheithauer, R., & Holland, M. M. (1998) 'Population data for 101 Austrian Caucasian mitochondrial DNA d-loop sequences: application of mtDNA sequence analysis to a forensic case'. Int. J. Legal Med., 111, (3), p.124-132

Perneger, T. V. (1998) 'What's wrong with Bonferroni adjustments'. BMJ, 316, (7139), p.1236-1238

Phillips, C., Ballard, D., Gill, P., Syndercombe-Court, D., Carracedo, Á., & Lareu, M. V. (2012) 'The recombination landscape around forensic STRs: Accurate measurement of genetic distances between syntenic STR pairs using HapMap high density SNP data'. Forensic Sci. Int. Gen., 6, p.354-365

Phillips, C., Fernandez-Formoso, L., Garcia-Magariños, M., Porras, L., Tvedebrink, T., Amigo, J., Fondevila, M., Gomez-Tato, A., Alvarez-Dios, J., Freire-Aradas, A., Gomez-Carballa, A., Mosquera-Miguel, A., Carracedo, A., & Lareu, M. V. (2011) 'Analysis of global variability in 15 established and 5 new European Standard Set (ESS) STRs using the CEPH human genome diversity panel'. Forensic Sci. Int. Gen., 5, (3), p.155-169

Phillips, C., Salas, A., Sánchez, J. J., Fondevila, M., Gómez-Tato, A., Alvarez-Dios, J., Calaza, M., de Cal, M. C., Ballard, D., Lareu, M. V., Carracedo, A. & SNPforID Consortium (2007) 'Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs'. Forensic Sci. Int. Genet., 1, p.273-280

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000) 'Inference of population structure using multilocus genotype data'. Genetics, 155, (2), p.945-959

Pritchard, J. K., Wen, X., & Falush, D. (2010) 'Documentation for STRUCTURE software: version 2.3'. [online] Available at: http://pritch.bsd.uchicago.edu/structure.html

Qamar, R., Ayub, Q., Mohyuddin, A., Helgason, A., Mazhar, K., Mansoor, A., Zerjal, T., Tyler-Smith, C., & Mehdi, S. Q. (2002) 'Y-chromosomal DNA variation in Pakistan'. Am. J. Hum. Genet., 70, (5), p.1107-1124

Qu, H.-Q., Li, Q., Xu, S., McCormick, J. B., Fisher-Hoch, S. P., Xiong, M., Qian, J., & Jin, L. (2012) 'Ancestry Informative Marker Set for Han Chinese Population'. G3, 2, (3), p.339-341

Ramachandran, S., Deshpande, O., Roseman, C. C., Rosenberg, N. A., Feldman, M. W., & Cavalli-Sforza, L. L. (2005) '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, p15942–15947

Raymond, M. & Rousset, F. (1995) 'An exact test for population differentiation'. Evolution, 49, (6), p.1280-1283

Rebala, K. & Szczerkowska, Z. (2005) 'Polish population study on Y chromosome haplotypes defined by 18 STR loci'. Int. J. Legal Med., 119, (5), p.303-305

Rees, J. L. (2000) 'The Melanocortin 1 Receptor (MC1R): more than just red hair'. Pigm. Cell Res., 13, p.135-140

Reichenpfader, B., Immel, U., & Klintschar, M. (2003) 'Population data on the AmpFℓSTR SGM plus PCR amplification kit in Germans and Austrians'. Forensic Sci. Int., 132, (1), p.84-86

Richards, B., Skoletsky, J., Shuber, A. P., Balfour, R., Stern, R. C., Dorkin, H. L., Parad, R. B., Witt, D., & Klinger, K. W. (1993) 'Muliplex PCR amplification from the CFTR gene using DNA prepared from buccal brushes/swabs', Hum. Mol. Genet., 2, (2), p.159-163

Risch, N., Burchard, E., Ziv, E., & Tang, H. (2002) 'Categorization of humans in biomedical research: genes, race and disease'. Genome Biol., 3, (7), p.comment2007

Ropers, H H., (2008), 'Genetics of intellectual disability', Curr. Opin. Genet. Dev, 18:241–250

Rosenberg, N.A., Burke, T., Elo, K., Feldman, M. W., Friedlin, P., Groenen, M. A. M., Hillel, J., Mäki-Tanila, A., Tixier-Boichard, M., Vignal, A., Wimmers, K., & Weigend, S. (2001) 'Empirical evaluation of genetic clustering methods using multilocus genotypes from 20 chicken breeds'. Genetics, 159, p.699-713

Rosenberg, N. A., Li, L. M., Ward, R., & Pritchard, J. K. (2003) 'Informativeness of genetic markers for inference of ancestry'. Am. J. Hum. Genet., 73, (6), p.1402-1422

Rosenberg, N. A., Mahajan, S., Ramachandran, S., Zhao, C., Pritchard, J. K., (2005) 'Clines, clusters, and the effect of study design on the inference of human population structure'. PLoS Genet., 1(6): e70

Rosenberg, N. A., Pritchard, J. K., Weber J. L., Cann, H. M., Kidd, K. K., Zhivotovsky L. A., Feldman, M. W., (2002) 'Genetic structure of human populations'. Science, 298, p.2381-2385

S. and Marper v. The United Kingdom - 30562/04 [2008] ECHR 1581 (4 December 2008)

Saggar, A & Bittles, A, (2008) 'Consanguinity and child health'. J. Paediatr. Child Health, 18, (5), p.244-249

Sahoo, S. & Kashyap, V. K. (2002) 'Genetic variation at 15 autosomal microsatellite loci in the three highly endogamous tribal populations of Orissa, India'. Forensic Sci. Int., 130, (2-3), p.189-193

Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K. B., & Erlich, H. A., (1988) 'Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase'. Science, 239, p.487-491

Salido, E. C., Yen, P. H., Koprivnikar, K., Yu, L., & Shapiro, L. J. (1992) 'The human enamel protein gene amelogenin is expressed from both the X and the Y chromosomes'. Am. J. Hum. Genet., 50, p.303-316

Sambrook, J., Fritsch, E.F., & Maniatis, T. (1989) in: Molecular Cloning: A Laboratory Manual, Cold Springs Harbor Laboratory, New York

Schmerer, W. M., Hummel, S., & Herrmann, B., (1999) 'Optimized DNA extraction to improve reproducibility of short tandem repeat genotyping with highly degraded DNA as target'. Electrophoresis, 20, p.1712-1716

Schneider, P. M. (1997) 'Basic issues in forensic DNA typing'. Forensic Sci. Int., 88, (1), p.17-22

Schneider, P. M. (2007) 'Scientific standards for studies in forensic genetics'. Forensic Sci. Int., 165, (2-3), p.238-243

Serre, D. & Pääbo, S. (2004) 'Evidence for gradients of human genetic diversity within and among continents'. Genome Res., 14, p1679-1685

Sharma, V. & Litt, M., (1992) 'Tetranucleotide repeat polymorphism at the D21S11 locus'. Hum. Mol. Genet., 1, (1), p.67

Shewale, J. G., Sikka, S. C., Schneida, E., & Sinha, S. K., (2003) 'DNA Profiling of Azoospermic Semen Samples from Vasectomized Males by Using Y-PLEX™6 Amplification Kit'. J. For. Sci., 48, (1), p.127-129

Skaletsky, H., Kuroda-Kawaguchi, T., Minx, P. J., Cordum, H. S., Hillier, L., Brown, L. G., Repping, S., Pyntikova, T., Ali, J., Bieri, T., Chinwalla, A., Delehaunty, A., Delehaunty, K., Du, H., Fewell, G., Fulton, L., Fulton, R., Graves, T., Hou, S., 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., Waterston, R. H., Wilson, R. K., Rozen, S., & Page, D. C. (2003) 'The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes'. Nature, 423, p.825-837

Southern, E. M. (1975) 'Detection of specific sequences among DNA fragments separated by gel electrophoresis'. J. Mol. Biol., 98, p.503-517

Sparkes, R., Kimpton, C., Watson, S., Oldroyd, N., Clayton, T., Barnett, L., Arnold, J., Thompson, C., Hale, R., Chapman, J., Urquhart, A., & Gill, P. (1996) 'The validation of a 7 locus multiplex STR test for use in Forensic casework. I. Mixtures, ageing, degradation and species studies'. Int. J. Legal Med., 109, p.186–194

Stern, C. (1943) 'The Hardy-Weinberg Law'. Science, 97, (2510), p.137-138

Stoneking, M. (2008), 'Human origins. The molecular perspective', EMBO reports, 9, p.S46-S50

Stringer, C. & Gamble, C. (1994) 'In Search of the Neanderthals' Thames & Hudson, London.

Sullivan, K. M., Mannucci, A., Kimpton, C.P., & Gill, P. (1993) 'A rapid and quantitative DNA sex test: fluorescence-based PCR analysis of X-Y homologous gene amelogenin'. BioTechniques, 15, (4), p.636-638

Tereba, A. (1999) 'Tools for Analysis of Population Statistics: PowerStats', Profiles in DNA, 2, p.14-16.

Thalmann, O., Fischer, A., Lankester, F., Pääbo, S., & Vigilant, L. (2007) 'The Complex Evolutionary History of Gorillas: Insights from Genomic Data'. Mol. Biol. Evol., 24, (1), p.146-158

The 1000 Genome Project Consortium (2010) 'A map of human genome variation from population-scale sequencing'. Nature, 467, p.1061-1073

The ENCODE Project Consortium (2012) 'An integrated encyclopedia of DNA elements in the human genome'. Nature, 489, p.57-74

The International HapMap Consortium (2003) 'The International HapMap Project'. Nature, 426, p.789-796

The International HapMap Consortium (2005) 'A haplotype map of the human genome'. Nature, 437, p.1299-1320

The International HapMap Consortium (2007) 'A second generation human haplotype map of over 3.1 million SNPs'. Nature, 449, p.851-861

The International HapMap Consortium (2010) 'Integrating common and rare genetic variation in diverse human populations'. Nature, 467, p.52-58

The International SNP Map Working Group – Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Marth, G., Sherry, S., Mullikin, J. C., Mortimore, B. J., Willey, D. L., Hunt, S. E., Cole, C. G., Coggill, P. C., Rice, C. M., Ning, Z., Rogers, J., Bentley, D. R., Kwok, P. Y., Mardis, E. R., Yeh, R. T., Schultz, B., Cook, L., Davenport, R., Dante, M., Fulton, L., Hillier, L., Waterston, R. H., McPherson, J. D., Gilman, B., Schaffner, S., Van Etten, W. J., Reich, D., Higgins, J., Daly, M. J., Blumenstiel, B., Baldwin, J., Stange-Thomann, N., Zody, M. C., Linton, L., Lander, E. S., & Altshuler, D. (2001) 'A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms'. Nature, 15, (409), p.928-933

Thomson, J. A., Pilotti, V., Stevens, P., Ayres, K. L., & Debenham, P. G. (1999) 'Validation of short tandem repeat analysis for the investigation of cases of disputed paternity'. Forensic Sci. Int., 100, (1-2), p.1-16

Thorne, A. G. & Wolpoff, M. H. (1992) 'The multiregional evolution of humans'. Sci. Am., 266, (4), p.76-3

Tian, C., Hinds, D., Shigeta, R., Kittles, R., Ballinger, D. G., & Seldin, M. F. (2006) 'A Genomewide Single-Nucleotide–Polymorphism Panel with High Ancestry Information for African American Admixture Mapping'. Am. J. Hum. Genet., 79, p.640-649

Tian, C., Kosoy, R., Nassir, R., Lee, A. Villoslada, P., Klareskog, L., Hammarstr□m, L., Garchon, H. J., Pulver, A. E., Ransom, M., Gregersen, P. K., & Seldin, M. F. (2009) 'European Population Genetic Substructure: Further Definition of Ancestry Informative Markers for Distinguishing among Diverse European Ethnic Groups'. Mol. Med., 15, (11-12), p.371-383

Tishkoff, S. A., Reed, F. A., Ranciaro, A., Voight, B. F., Babbitt, C. C., Silverman, J. S., Powell, K., Mortensen, H. M., Hirbo, J. B., Osman, M., Ibrahim, M., Omar, S. A., Lema, G., Nyambo, T. B., Ghori, J., Bumpstead, S., Pritchard, J. K., Wray, G. A., & Deloukas, P. (2007) 'Convergent adaptation of human lactase persistence in Africa and Europe'. Nat. Genet., 39, (1), p.31-40

Torroni, A., Achilli, A., Macaulay, V., Richards, M., & Bandelt, H. J. (2006) 'Harvesting the fruit of the human mtDNA tree'. Trends Genet., 22, (6), p.339-345

Triggs, C. M., Buckleton, J. S., (2002) ‘Logical implications of applying the principles of population genetics to the interpretation of DNA profiling evidence’. Forensic Sci. Int., 128, p.108-114

Triggs, C., Harbison, S. A. & Buckleton, J. (2000) 'The calculation of DNA match probabilities in mixed race populations', Sci. Justice, 40, (1), p.33-38

Turner, A. & Chamberlain, A. (1989) 'Speciation, morphological change and the status of African Homo erectus'. J. Hum. Evol., 18, (2), p.115-130

van Oorschot, R. A. H., & Jones, M. K., (1997) 'DNA fingerprints from fingerprints'. Nature, 387, p.767

Voight, B. F., Kudaravalli, S., Wen, X., & Pritchard, J. K. (2006) 'A map of recent positive selection in the human genome'. PLoS Biol., 4, (3), p.446-458

Wallace, D. C., Brown, M. D., & Lott, M. T. (1999) 'Mitochondrial DNA variation in human evolution and disease'. Gene, 238, (1), p.211-230

Walsh, P. S., Metzger, D. A., & Hiquchi, R. (1991) ' Chelex 100 as a medium for simple extraction of DNA for PCR-based typing from forensic material', Biotechniques, 10, (4), p.506-513

Walsh, S. J. (2004) 'Recent advances in forensic genetics'. Expert. Rev. Mol. Diagn., 4, (1), p.31-40

Walsh, S. J., Buckleton, J. S., Ribaux, O., Roux, C., & Raymond, T. (2008) 'Comparing the growth and effectiveness of forensic DNA databases', Forensic Sci. Int. Genetics Supplement Series, 1, (1), p.667-668

Walsh, S., Liu, F., Ballantyne, K. N., van Oven, M., Lao, O., & Kayser, M. (2011a) 'IrisPlex: A sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information'. Forensic Sci. Int. Genet., 5, p.170-180

Walsh, S., Lindenburgh, A., Zuniga, S. B., Sijen, T., de Kniff, P., Kayser, M., & Ballantyne, K. N., (2011b) 'Developmental validation of the IrisPlex system: Determination of blue and brown iris colour for forensic intelligence'. Forensic Sci. Int. Genet., 5, p.464-471

Walsh, S., Liu, F., Wollstein, A., Kovatsi, L., Ralf, A., Kosiniak-Kamysz, A., Branicki, W., & Kayser, M. (2012a) 'The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA'. Forensic Sci. Int. Genet., [in-press]

Walsh, S., Wollstein, A., Liu, F., Chakravarthy, U., Rahu, M., Seland, J. H., Soubrane, G., Tomazzoli, L., Topouzis, F., Vingerling, J. R., Vioque, J., Fletcher, A. E., Ballantyne, K. N., & Kayser, M. (2012b) 'DNA-based eye colour prediction across Europe with the IrisPlex system'. Forensic Sci. Int. Genet., 6, p.330-340

Weber, J. L., & Wong, C., (1993) 'Mutation of human short tandem repeats'. Hum. Mol. Genet., 2, (8), p.1123-1128

Werrett, D. J., (1997) 'The National DNA Database'. Forensic Sci. Int. 88, p.33-42

Wolpoff, M., Hawks, J., Caspari, R., (2000) 'Multiregional, not multiple origins'. Am. J. Phys. Anthropol., 112, p.129-136

Woodley, M. A., (2008) 'Inbreeding depression and IQ in a study of 72 countries', Intelligence, doi:10.1016/j.intell.2008.10.007

Woods, C. G., Cox, J., Springell, K., Hampshire, D.J., Mohamed, M.D., McKibbin, M., Stern, R., Raymond, F.L., Sandford, R., Malik Sharif, S., Karbani, G., Ahmed, M., Bond, J., Clayton, D., Inglehearn, C.F. (2006) 'Quantification of homozygosity in consanguineous individuals with autosomal recessive disease', Am. J. Hum. Genet., 78, (5), p.889-96

Wright, S. (1965) 'The interpretation of population structure by F-statistics with special regard to systems of mating', Evolution, 19, (3), p.395-420

Yong, R. Y., Gan, L. S., Coble, M. D., & Yap, E. P. (2007a) 'Allele frequencies of six miniSTR loci of three ethnic populations in Singapore'. Forensic Sci. Int., 166, (2-3), p.240-243

Yong, R. Y., Gan, L. S., Coble, M. D., & Yap, E. P. (2007b) 'Polymorphism studies of four miniSTR loci for three ethnic populations in Singapore'. Leg. Med. (Tokyo), 9, (5), p.278-281

Zhivotovsky, L. A., Ahmed, S., Wang, W., & Bittles, A. H. (2001) 'The forensic DNA implications of genetic differentiation between endogamous communities'. Forensic Sci. Int., 119, (3), p.269-272

Zhivotovsky, L. A ., Rosenberg, N. A., & Feldman, M. W. (2003) 'Features of evolution and expansion of modern humans, inferred from genomewide microsatellite markers'. Am. J. Hum. Genet., 72, p.1171-1186