Autosomal haplotypes as markers for the histories and structures of human populations

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

by
Suzanne P Lavelle (BA Hons, BSc Hons)
Department of Genetics
University of Leicester
2009

Autosomal haplotypes as markers for the histories and structures of human populations

Suzanne P Lavelle

Abstact

The demographic history of humans is very complex. Populations have undergone bottlenecks, isolation, migration, admixture and expansions. All of these have added to the complexity of what makes a population and how that population has changed over time. A record of these events can be found in our DNA.

This project used autosomal DNA to trace the histories and structures of human populations by using a combination of a SNP (single nucleotide polymorphism) and an STR (short tandem repeat) – SNPSTR (Mountain et al. 2002). Forensic STRs formed the basis for the SNPSTR systems because their allelic diversity is well characterised, their mutation rates have been reliably measured and they are robust in PCR amplification.

Four SNPSTR systems were found, using SNPs which had been verified by HapMap and/ or Perlegen and which were < 500 base pairs away from the forensic STRs.

These SNPSTRs were typed on DNAs from the HapMap project, the CEPH-HGDP, Cornwall, UK African Caribbeans, Danes and Greenland Inuit. They were analysed using an ABI3100 and GeneMapper software. Data from the combined SNPSTRs allowed inferences to be made about population structures, and also enabled the calculation of the TMRCA of the derived SNPs associated with the forensic STRs.

Population structure was evident in the MDS plots where rudimentary population groupings could be seen. The Americas were outliers, reflecting their later peopling some 15,000 years ago (Jobling et al. 2003). Haplogroup analysis highlighted population isolates, such as the Surui in Brazil.

The STRUCTURE analysis of the SNPSTR data has also provided some insights into the admixed nature of the autosomal DNA in known admixed populations such as the Greenland Inuit and to some extent, the African Caribbeans.

One SNPSTR was expanded into a larger haplotype block - a PHAX - Phylogeographically informative Haplotypes on the Autosomes and seX chromosomes, by means of a SNaPshot reaction. The preliminary data from this suggested that this would allow us to gain a more complete insight into the histories and structures of human populations.

Acknowledgements

My greatest thanks go to my partner Sara, who has been with me throughout this journey and has supported me mentally, financially and emotionally every step of the way. Her love and strength not only enabled me to start this PhD, but also, and more importantly, made it possible for me to finish it.

I also owe a great deal of gratitude to my late parents, Margaret and Gerry Lavelle. Their belief in me always, and their love and encouragement when I first made the move to change careers to a scientific one made the decision much easier. I also thank them for their DNA contribution – without which I wouldn't be here. Sadly neither lived to see the end of this great endeavour.

My next thanks are for my first tutors at Scarborough College, John Skinner for making chemistry not only interesting, but also totally fun, and Lesley Grey, who made me want to know more about everything. She encouraged me every step of the way and was a fantastic teacher and role model.

Thanks also go to Mark Jobling whose easy manner and approachability made it possible for me to ask lots of questions and whose help made the mysteries of DNA and population genetics understandable. He has provided encouragement, advice and friendship, which will always be with me.

It would be wrong also not to thank the members of Mark Jobling's group who have also helped out with this project. Special thanks are due to Stéphane for his wise counsel, help with data analysis and copious coffee, and also to Holly who showed me the delights of macros and provided spicy snacks.

Thanks must also be extended to the many members of the Genetics department and my cohort of PhD students who helped me find things, work things out and enabled me to learn from them. Special thanks here go to Rita without whom my PCR would never have got working again, and to Ruth for advice on how to get the best out of my DNA. Ed saved me weeks of work by explaining pivot tables, and Sue helped out with referencing software.

I also owe a great deal of thanks to Annette Cashmore, who has not only encouraged me to get this thesis finished and written, but has supported me whilst doing so. It was Annette who first offered me the opportunity to do the PhD at Leicester, and she has since continued to keep me here, which has been wonderful. The whole GENIE team have also made the writing up process much more fun than I thought it could be.

Finally, I also want to thank Richard, my dive buddy and instructor, who not only taught me to scuba dive when I first moved to Leicester, which proved to be a wonderful distraction, but he has also become an invaluable friend. His willingness to help out in the other things in life has made the writing-up experience much easier.

Table of Contents

1 Introduction

1.1 Why do we care about human histories and population structures?

As human beings, it is inherently interesting for us to explore our origins - where our ancestors came from, how long ago and what the processes of our evolution were. In the past, human population studies relied on the fossil record, archaeological research, anthropological information and historical linguistics to tell us about the origin and spread of humans across the globe. Now, however, with the development of genetic molecular methods we are able to gain a complementary insight into the past and hence our present.

As modern humans spread across the globe (see Figure 1-1) after their origins in Africa about 130,000 years ago, they formed into groups and became relatively isolated, either geographically, linguistically, or culturally. As they spread across the globe, they also entered diverse selective environments. Some were better adapted to deal with these environments and survived, others did not. This selection, coupled with the relative isolation of these groups led to non-random mating within the whole population, and this has created population structure.

Recent studies using dynamic population models to simulate the parameters of the colonization process using the Centre d'Etude du Polymorphisme Humain-Human Genome Diversity Project (CEPH-HGDP) DNAs have indicated that a

founding population of approximately 1000 effective individuals started to expand around 56,000 years ago (Liu et al. 2006), and these populations rapidly colonized new habitats.

Perhaps such a relatively low number of founding individuals could lie behind the reason why the differences between populations on a genetic level are quite small (between 3% – 5% of the total genetic variation only). Moreover, the major differences are within populations where approximately 85% of genetic variation is found (Lewontin 1972). However, since Lewontin's original work, further studies have shown that this may not be totally accurate because many more markers are needed to be able to correctly ascertain the genetic differences within and between populations (Witherspoon et al. 2007). One further study using microsatellite data also shows that geographic distance from east Africa explains 85% of the decrease in gene diversity within human populations (Prugnolle et al. 2005).

Having ascertained that most the genetic diversity lies within populations, this makes population structure of interest to us because it also affects the distribution of neutral variation, Mendelian disease alleles for diseases such as cystic fibrosis and also alleles influencing the susceptibility to complex disorders such as diabetes. In addition to this, knowledge of population structure is important for calculating the efficacy of some drugs and possibly the likelihood of adverse reactions, as it is known that there are inter-ethnic

differences in drug-metabolising enzyme allele frequencies, and that these affect drug response (Wilson et al. 2001).

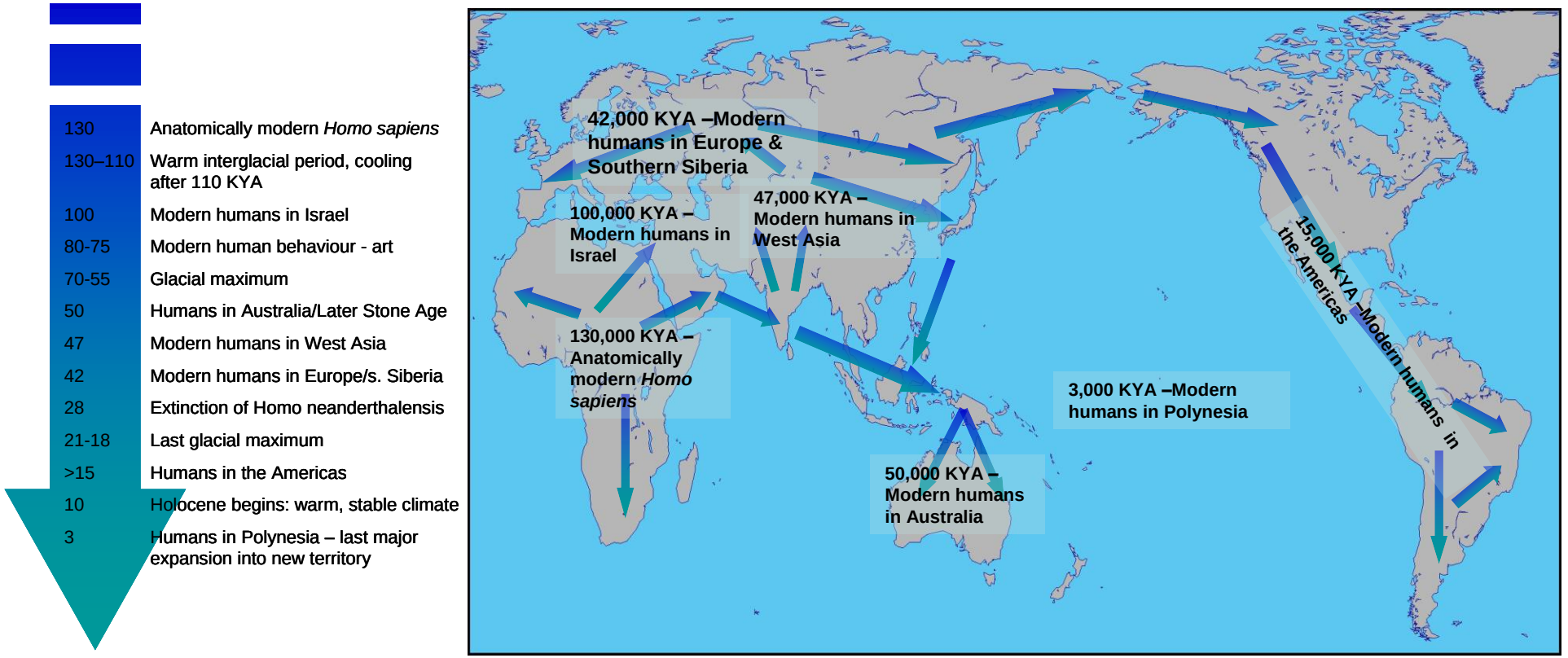


Figure 1-1. Map showing the spread of modern Homo sapiens across the globe with accompanying timeline
The timeline is in units of thousands of years ago (KYA), ranging from 130KYA to 3 KYA. Source: (Jobling et al. 2003)

A final reason why population histories and structures are of interest is because of the role they can play in forensic investigations and identification of human remains. This is of particular importance after mass mortality disasters such as 9/ 11 (Brenner and Weir 2003), natural disasters such as the Asian tsunami which hit Thailand and the rest of Indonesia in 2004 and also mass genocide during wars. The non-random distribution of alleles among different populations and sub-populations affects the matching probability of forensic profiles, a subject that was much debated in the early 1990s (Balding and Donnelly 1995). Also, it is now common practice to use small samples of DNA to predict the ethnic origin of a person (Lowe et al. 2001), and the reliability of this method depends on a clear understanding of population structure, and its relationship to ethnicity.

1.2 How have we studied human histories and populations to date?

1.2.1 Blood groups and HLA

Some of the earliest work to investigate population structure was carried out using classical markers such as blood groups and HLA (Human leukocyte antigen) alleles. The first human genetic polymorphism, the ABO blood group system, was described in 1900 (Landsteiner 1900). The blood groups were detected serologically, using antigens on the red blood cells and their reactions to specific antibodies in the blood serum. Using this system, Landsteiner classified individuals into four types: A, B, AB and O. The first use of these blood groupings for population studies was carried out in 1919 where the

frequencies of these blood groups were identified in different poplulations (Hirszfeld and Hirszfeld 1919). The HLA (Human leukocyte antigen) is the name of the human major histocompatibility complex (MHC) and the MHC loci are some of the most genetically variable coding loci in humans. They were first investigated due to their importance in tissue transplantation but were soon recognized as regions with a large genetic variation which could be used for disease association and population genetics studies. These types of markers were certainly influential and continue to be used today; however, they are prone to selective effects (HLA alleles in particular) and the evolutionary relationships between the alleles were unknown. What was needed was a way to investigate the DNA sequences themselves, and this started to occur after the discovery that our DNA was unique to each of us and that we all had our own genetic fingerprint.

1.2.2 RFLP analysis and Southern blotting

Detecting the differences in DNA sequencing and also being able to detect specific DNA sequencing in a DNA sample became possible in the 1970s with the discovery of both the Southern Blot technique and also the ability to break DNAs up into much smaller fragments and analyse them by detecting differences in size.

Restriction fragment length polymorphism analysis (RFLP) is a simple technique where sequence-specific restriction enzymes are used to break high molecular weight DNA up into smaller sections and these are then separated

according to their lengths by gel electrophoresis. RFLP analysis was the first DNA profiling technique used and was used in forensics and paternity testing (Jobling and Gill 2004)

Southern blotting, which was named after Edwin M Southern who developed this procedure in the 1970s (Southern 1975), is designed to locate a particular sequence of DNA within a complex mixture. It combines the transfer of electrophoresis-separated restriction-digested DNA fragments to a filter membrane and then subsequent fragment detection by probe hybridization. Hybridization of the probe to a specific DNA fragment on the filter membrane indicates that that particular fragment contains the DNA sequence complementary to the probe. These are then detected by autoradiography.

RFLP analysis combined with Southern blotting was the first DNA profiling technique used and was exploited in forensics and paternity testing (Jobling and Gill 2004). Although both Southern blots and RFLP analysis are still used today, they are both slow and cumbersome and also require a large amount (several micrograms) of sample DNA. This makes them unsuitable for use on larger numbers of DNA samples where it is necessary to process these more rapidly.

The more rapid detection of differences between DNA samples became possible after the invention of the polymerase chain reaction (PCR) technique in the 1980s.

1.2.3 PCR

Although Kary Mullis is credited with inventing the PCR technique in 1984 (Mullis 1990), a method which was very similar was first described in 1971 by Kleppe (Kleppe et al. 1971) where an enzymatic assay was used to replicate a short DNA template with primers *in vitro*.

The PCR method is based upon the use of a DNA polymerase which is able to withstand the high temperature of >90 °C needed for the separation of the two DNA strands. *Taq* polymerase had been discovered in 1976 (Chien et al. 1976) and this paved the way for improvements in the PCR method, making it faster (automated) and more reliable than when used with non-thermostable polymerases.

PCR enables the amplification of a small amount of DNA and can generate 100 billion similar molecules in the space of a couple of hours. This has not only obviated the need for large amounts of initial sample DNA, but it has also speeded up the process radically. Automated PCR thermocyclers are able to simultaneously amplify thousands of different DNA samples in one session. PCR is now a common technique used for a variety of applications including: DNA cloning for sequencing, the identification of genetic fingerprinting for paternity analysis or forensic applications, the detection and diagnosis of infectious diseases, and , of course, also for use in DNA-based phylogeny.

With the array of modern techniques available to us today, and primarily, PCR and DNA sequence analysis, it is now possible to use any section of DNA we choose to examine population histories and structures. So, the question is, what type of DNA markers should and could be used for population genetics studies? The main focus has been on using single nucleotide polymorphisms (SNPs) and also on short tandem repeats (STRs). See Box 1 for more information on these two markers.

Traditionally, uniparentally inherited sections of DNA such as the male-specific region of the Y chromosome or mitochondrial DNA (mtDNA) have been used in genetic analyses because to all intents and purposes, these two sections of DNA can each be considered as non-recombining. This makes them very useful for phylogenetic studies of human populations because it is simpler to trace single ancestors as any change in the DNA would only be due to a mutation (Jobling et al. 2003). The Y chromosome is paternally inherited and non-recombining (apart from the small pseudoautosomal regions at the ends of the chromosome) and there is now a robust phylogeny of Y-chromosomal haplogroups, defined by binary markers such as SNPs (see

Figure 1-2). Mitochondrial DNA is also non-recombining, but is maternally inherited, and produces a similarly simple phylogenetic tree (Figure 1-3).

Box 1. SNPs and STRs

SNPs – single nucleotide polymorphisms are binary DNA sequence variations that occur when a single nucleotide in the genome sequence is changed into another; and can also include the insertion or deletion of a nucleotide. There are variations between human populations, so a SNP allele that is common in one geographical range or ethnic group may be rarer in another. On their own, SNPs are not very informative; however, if many SNPs are combined, they become informative markers for population geneticists. There are currently over 6.5 million validated SNPs in the human genome (dbSNP, Build 130 – May 2009). The average mutation rate of SNPs is ~ 10^{-8} per base per generation (Nachman and Crowell 2000).

If a SNP occurs in a coding region it can be classed as either synonymous or nonsynonymous (with missense and nonsense effects). The effect of a single SNP on a gene may not be large – it may influence the activity of the encoded protein in a subtle way - but these subtle effects can influence susceptibility to common diseases, such as heart disease or Alzheimer's disease.

STRs – short tandem repeats occur when a pattern of two or more nucleotides is repeated and the repeated sequences are directly adjacent to each other. The repeat unit can range in length from 2 to 10 base pairs, and in polymorphic examples the number of repeats is typically between 8 and 30 repeats. By identifying a specific set of variable STR alleles it is possible to create a genetic profile of an individual and this makes STRs not only useful for population genetics studies, but also invaluable for forensic identification purposes.

Box 1. SNPs and STRs - continued

STR mutation rates in humans are known to be relatively high at ~10^{-3} per locus per generation (Brinkmann et al. 1998). STRBase (www.cstl.nist.gov/ biotech/ strbase/ index.htm) provides detailed information on all of the STRs currently used in forensic science work, including mutation rates and disease allele associations.

Most STRs studied in population genetics are outside coding regions and regarded as selectively neutral, however there are microsatellites present in coding regions or regulatory regions and some of these may influence transcription and hence can affect the gene expression or the function of gene products.

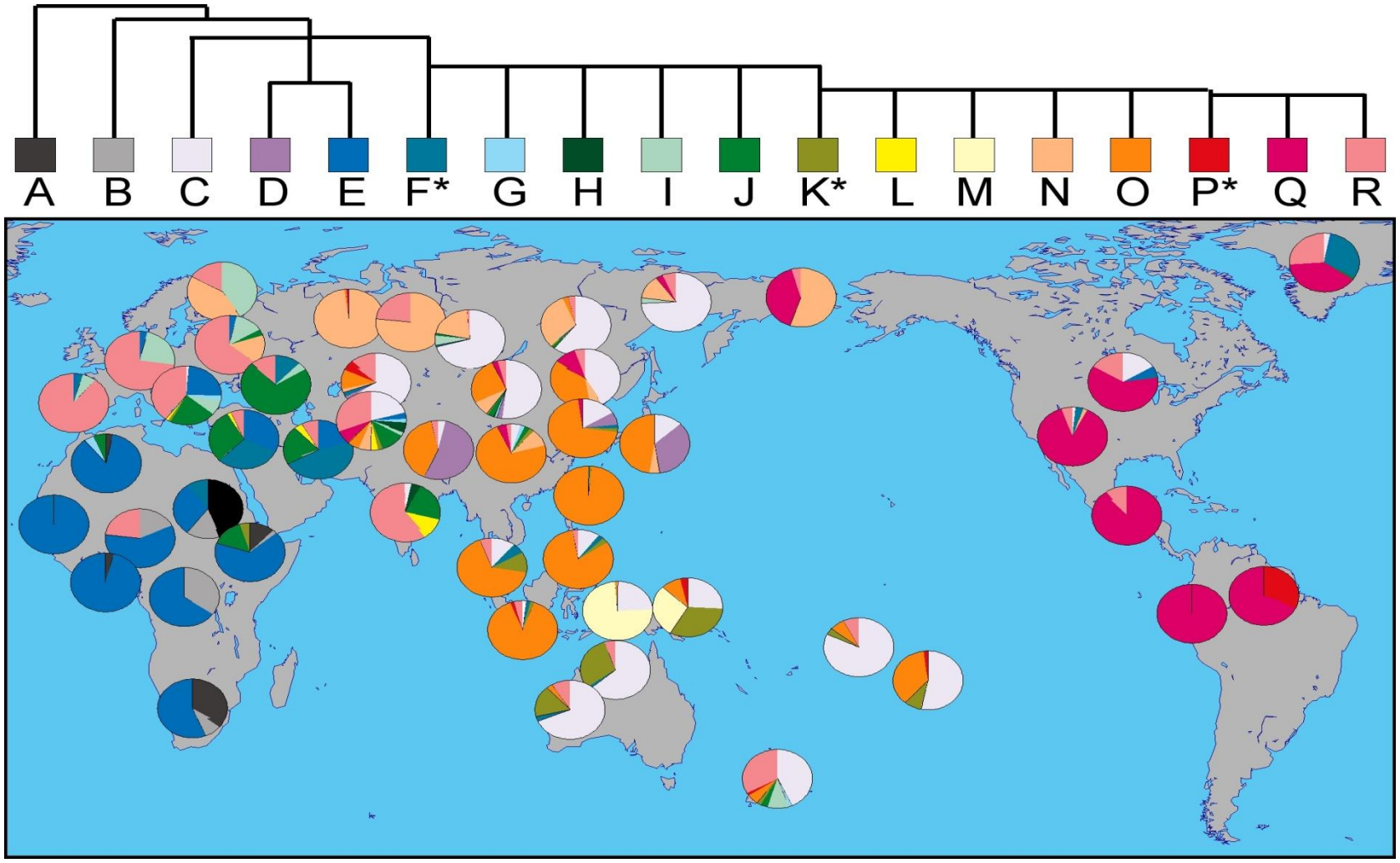


Figure 1-2. World-wide Y chromosome haplogroup map and tree (indigenous populations).
A distinct geographical distribution can be seen, where Haplogroups A and B are more common in Africa and P, Q and R are more frequent in the Americas. The pie charts represent populations and the sectors within each pie chart are proportional to the haplogroup frequencies which are shown in the simplified phylogeny at the top of the figure. Source: based on (Jobling and Tyler-Smith 2003).

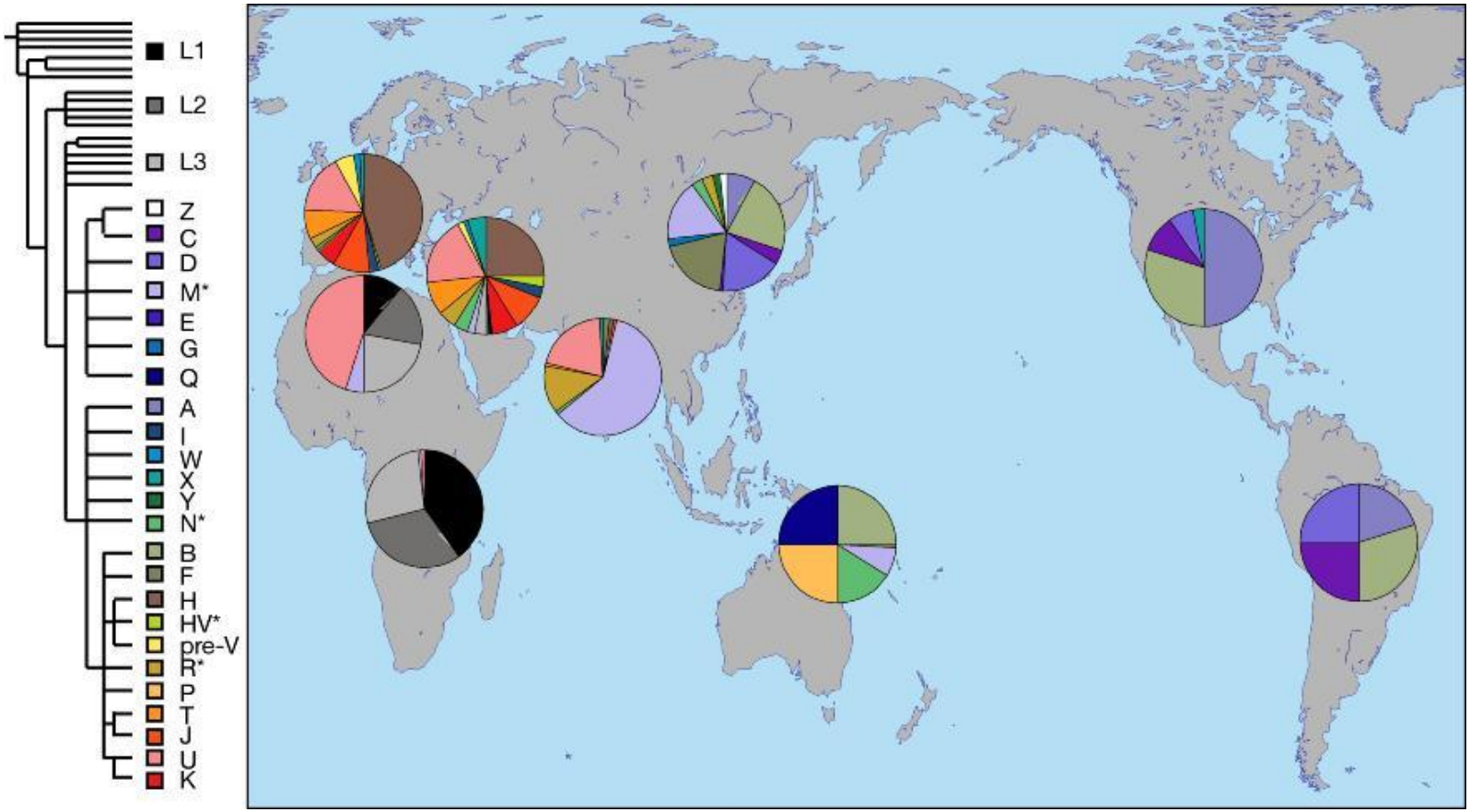


Figure 1-3. World-wide mtDNA haplogroup map and tree (indigenous populations).
mtDNA haplogroups show a distinct geographical distribution, with L1, L2 and L3 being the deeper rooting branches and also located in Africa, and with haplogroups B, C and D being more common in the Americas. The pie charts represent populations and the sectors within each pie chart are proportional to the haplogroup frequencies which are shown in the simplified phylogeny to the left of the figure. Source: (Jobling et al. 2003) .

The distributions of Y chromosome and mtDNA haplogroups are geographically differentiated and so can provide information on ancestry. The differences between the older, deep-rooting branches of the African populations compared to those of the "newer" populations such as the Americas can clearly be seen. The geographical structuring is due to migration history of human populations, amplified by the strong effects of genetic drift. Drift is accentuated on the Y chromosome because its effective population size is only 25% of that of any autosome (it is haploid and only present in males). Genetic drift will accelerate the differentiation between groups of Y chromosomes in different populations. In addition to this, the distribution of the Y chromosome is influenced by the mating behaviour of men. For example, in modern societies, 70% of cultures practice patrilocality where the men stay at their birthplace but the women move from their home to that of their new husbands (Burton et al. 1996; Seielstad et al. 1998). Although this does create an enhancement of differentiation in local Y chromosomes, it is unlikely to have any effects at the continental level (Wilder et al. 2004).

However, despite being very useful markers for population studies, the Y chromosome and mtDNA still only give us a very limited view of past events as they each rely on a single locus, and hence provide only a single realization of the evolutionary process. Furthermore, they may also be subject to possible problems from selection.

In principle, positive selection could be suspected to be acting on the Y if the TMRCA of all modern Y chromosomes was significantly younger than expected. The TMRCA is expected to be relatively young compared to the average age of autosomal segments, because the effective population size of the Y chromosome is lower. However, a departure from neutral expectation is hard to judge because of the likely effects of male behaviours (e.g. social selection leading to enhanced reproductive success for some men, patrilocality, wars). Social selection is so strong that natural selection may be impossible to detect, or is overridden. However, many different studies, and many estimates of TMRCA, have suggested that, as far as we can tell, the Y chromosome behaves like a neutral locus (Jobling and Tyler-Smith 2003). Also, it is important to note that there are no candidate Y-linked phenotypes that could be considered to be differentially adapted to different environments. One issue that is directly related to the Y chromosome, however, is that it is subject to purifying selection. An example of this is in the loss or inactivation of Y genes which can produce an XY female (Disteche et al. 1986) or male infertility (Vogt et al. 1996). The Y chromosome is very rich in low copy number repeated sequences, and male infertility is caused by non-allelic homologous recombination between these paralogues, through the deletion of the azoospermia factor regions *AZFa*, *AZFb* and *AZFc* (McElreavey et al. 2006). However, despite this, the evidence from population studies to date, does not indicate that the Y chromosome is under selection, and can therefore be considered a neutral locus (Jobling and Tyler-Smith 2003).

In contrast, mtDNA may also not be as "neutral" as was first thought. It may also be under pressure from positive selection and this would then reflect in the mtDNA haplogroups map and trees which are currently used to describe the evolution and spread of this marker. It has been suggested (Wallace 2005) that mtDNA has adapted according to climate change and that as humans moved northwards, uncoupling protein 1 (UCP1) induction in brown adipose tissue was not sufficient to regulate body temperature in the colder climates. Therefore, there was an acquisition of mtDNA mutations that partially uncoupled oxidative phosphorylation and this in turn led to perpetually increased mitochondrial heat production. In support of this theory, Wallace suggested that lineage J in mtDNA harbours specific "uncoupling DNA variants" that reduce mitochondrial ATP output in favour of heat production. He has shown a correlation between mtDNA lineages and geographic origins of indigenous populations. For example, there are two correlated non-synonymous point mutations on the ND3 and the ATP6 gene which are characterized by a clear association with temperature and which also appear to be targets of natural selection which have produced this association with climate (Balloux et al. 2009).

1.3 Ascertainment bias

At this stage that it is very important to consider just how the DNA samples for all these studies have been collected, and also to note something about ascertainment bias.

Ascertainment bias is the systematic distortion in a data set which is caused by the way in which the samples are collected or markers are selected. This is particularly important for SNPs, because STRs, which are usually variable in all populations, provide a less biased measure of diversity. One example of ascertainment bias would be, if for example, variants were discovered in a set of European Y chromosomes, then a Chinese population was tested for these markers and it was found that there was no variation present. It could be concluded that the Chinese Y chromosomes were of low diversity, but this conclusion would be wrong. An example of this (Su et al. 1999) was highlighted when high variability SNPs were ascertained in a southern Chinese population and then tested on northern Chinese populations. The findings suggested that there was much less variability in northern China. However, when a global set of binary markers was tested on the Chinese populations, more haplotypes were found in northern China.

Another way to describe ascertainment bias would in the terms of the "caveman effect". This refers to non-random sampling methods, which is also an issue when trying to decipher population structure from DNA samples which have been previously collected for different purposes and studies. A very simple example of the "caveman effect" can be seen in the way that modern societies almost always associate prehistoric people with caves and therefore assume that they lived in caves. However, this is a false conclusion, and simply because most of the data has survived in caves, does not necessarily mean that they lived in them.

1.4 What kind of biparental systems do we need to study human histories and populations?

It is now possible, due to the emergence of large-scale genotyping methods, to type over 1 million SNPs at one time on genome-wide typing analyses. Microarray technology is constantly improving in reliability and processing power. This makes it less expensive and faster to type more and more SNPs. This wealth of data has been very useful in defining more recent studies in population structure analysis. One example is the HapMap project (2003) and also the recent exploitation of the HGDP-CEPH panel (Cann et al. 2002). The downside of this new technology is that the SNPs being typed are distributed widely throughout the genome and are therefore quite far apart. This means that they are limited in usefulness for population history studies because the historical information which is found in haplotype data is not present in the multitude of unlinked or loosely linked SNPs. A haplotype is a combination of alleles at multiple loci on the same chromosome, which, if they are closely linked, are transmitted together.

To supplement data from the non-recombining haplotypes found in the Y chromosome and mtDNA, there is a need for informative systems, such as haplotype data, within the rest of the genome, and here there are two main options available. We could choose to look at a few large autosomal haplotype blocks, such as those which have been found by the HapMap project (2005), or alternatively we could look at many small haplotypes formed by pairs of very

closely linked polymorphic markers such as a combined SNP and STR markers (SNPSTRs) on the austosomes (Mountain et al. 2002).

Haplotype blocks are regions of low recombination on autosomes or X chromosomes and our genome is made up of a mosaic of such blocks which are between 5 -200 kb in length and where 3 – 7 haplotypes account for most of the genetic variation in modern humans (Paabo 2003). The HapMap project (Altshuler et al. 2005) - International HapMap Consortium 2003 and 2005 - took sets of samples from peoples across the world, including the Yoruba in Ibadan, Nigeria (abbreviated to YRI); , trios in Utah, USA, from the Centre d'Etude du Polymorphisme Humain collection (abbreviated to CEU); Han Chinese in Beijing, China (abbreviated to CHB); and Japanese in Tokyo, Japan (abbreviated to JPT). The aim of the HapMap project was to create a public, genome-wide database of common human sequence variation, in order to provide information as a guide to genetic studies of clinical phenotypes. As part of the work, the HapMap project also examined the extent to which haplotypes were shared across populations. A hidden Markov model was used, in which each haplotype was modelled in turn as an imperfect mosaic of other haplotypes. This enabled structures within the HapMap samples to be studied and additionally, the identification of regions in which there was no evidence of historical recombination (in those samples); these have been dubbed PHAXs (Phylogeographically informative haplotypes on the autosomes and sex chromosomes) by Mark Jobling (Ballereau et al. in preparation). These non-recombining haplotype blocks provide a useful resource to study human

genetic diversity in general, and work is underway to catalogue them and prioritise them for investigation. Many of these haplotype blocks contain many markers and give us an informative tree, but each also gives us a single evolutionary picture and therefore is not only somewhat biased, but also could be subject to selection acting upon it. It would certainly be possible to type several PHAXs, however, that would entail the custom typing of many SNPs and would therefore prove to be laborious and expensive.

The alternative, therefore, would be to use several independent SNPSTRs as a system to study the histories and structures of populations. Many loci may tell a story of the whole genome rather than their individual versions of it and the data they provide should be less likely to display the effects of selection, as it is very unlikely that selection will have been acting in the same way on all of these markers at any one time. SNPSTRs will, in theory, tell us about the histories of the autosomal DNA.

In order to set the work carried out in this project into context, it is useful to look at some of the studies which have been carried out to date with regards to population structure and histories. Almost on a daily basis, new papers are published detailing information about a particular population and what generalizations can be drawn from the study. For example, mtDNA and the Y chromosome have been used to analyse the peopling of Korea, with the conclusion that an early northern Asian settlement with at least one subsequent male-biased southern-to-northern migration occurred which was possibly

associated with the spread of rice agriculture (Jin et al. 2009). As it would be a very difficult and long task to take all of these into account with regards to writing an introduction to this project, which is primarily about SNPSTRs as markers, it is instead useful to focus on some of the more key papers published in this field within the last 5 years. In this way, it will enable much more focus to be placed on where SNPSTRs can be used for analysis of population histories and structures, and it will also provide a more comprehensible context for this work.

1.5 Autosomal haplotype blocks

Both the Y chromosome and mtDNA have proven to be very useful in helping to reconstruct our past. They provide very robust phylogenetic trees and also excellent information about haplogroups and both can also be dated to provide a TMRCA. However, they provide a somewhat limited view of human population structure as they do not include autosomal DNA, which is the DNA which is inherited from both parents, and as can be seen from the example of patrilocality, males and females do have different social behaviours, and this is reflected in our DNA. Autosomal DNA will give us a more complete and balanced view of the histories and structures of human populations, and typing of large numbers of independent loci would provide much more information about what has been happening in our autosomal DNA over time. Haplotype blocks are areas of DNA where markers show strong linkage disequilibrium (LD) (Daly et al. 2001). The boundaries between these blocks often correspond to recombination hotspots (Jeffreys et al. 2001), and the blocks themselves yield

informative haplotypes. However, a certain amount of caution is needed because the haplotype blocks could be the result of past recombination events and genetic drift, which would mean that they would differ substantially between populations and would therefore make it very difficult to construct phylogenetic trees using these blocks alone. Despite this caveat, autosomal haplotype blocks as well as large numbers of SNPs and different STRs, as well as a combination of SNPs and STRs – SNPSTRs (Mountain et al. 2002) have been used to investigate population histories and structures and the results have proven to be contentious, with some claiming that population structure is clinal, and others arguing for the case of a more clustered population structure. Two current figures in population structure, Rosenberg and Pritchard, have convincing arguments for both a clinal and a structured view of population histories. Because it is still unclear which of these two is the most likely, I will describe both theories here and cover each one individually so that both can be taken into account within the framework of the results achieved by this thesis.

1.6 The problem with population models

What is a "population"? And if we are to agree on the definition of what constitutes a population, then how can we use genetic analysis to reflect this population? There are many different components that make up a population; language, culture, physical characteristics and also, and perhaps most importantly, geographical location. In population genetics models, very often, geography has not been taken into account and this has lead to misleading conclusions in the past. Large landscape features such as a mountain range,

large bodies of water, or deserts are often barriers to the spread of humans and therefore impose a structure of their own upon the patterns we see in the DNA. Population models should therefore incorporate geography into their algorithms in order to gain a more realistic view of the spread of humans across the globe.

Recent studies using dynamic population models to simulate the parameters of the colonization process using a global panel of DNAs (the CEPH-HGDP DNAs, which will be described more fully later) have indicated that a founding population of approximately 1000 effective individuals started to expand around 56,000 years ago (Liu et al. 2006). These populations rapidly colonized new habitats.

As early as the 1970s when the relationship between the frequency of human blood groups and geographic locations were studied (Mourant et al. 1976), it was recognised that there were strong geographical patterns – which were clinal in nature. The presence of these clinal patterns led to the establishment of the principle of isolation by distance (IBD), and the general idea of distance-based methods (DBMs), which are incorporated into many population genetics models today. However, not everyone agrees that IBD should be taken into account, and certainly when looking at population structure, there are many different types of computer models and algorithms which are used to simulate population histories and structures.

There are basically two types of clustering methods which can be used to show whether or not populations are structured or are in clines; distance-based methods (DBMs) and model-based methods.

DBMs calculate a pairwise distance matrix between every pair of individuals. These are represented graphically, for example in multi-dimensional scaling (MDS) plots. Clusters can then be identified by eye. There are disadvantages to DBMs and these include the following:

- Clusters are very dependent on the distance measure and the graphical representation chosen.
- It is often difficult to assess the meaningfulness of any clusters obtained.
- It is not easy to incorporate additional information, such as geographical data.

On the other hand, there are model-based methods – these assume random observations are taken from a parametric model. Then, using standard statistical methods such as maximum likelihood or Bayesian methods, inference for the parameters corresponding to each cluster is done jointly with inference for the cluster membership of each individual.

One of the main arguments against incorporating IBD and also the clinal-based population structure comes from (Pritchard et al. 2000). In this population structure paper, these authors used a model-based clustering method with multilocus genotype data to assign individuals into populations. Individuals

can belong to two or more populations simultaneously (in the case of admixture), but no particular mutational processes are assumed in the model, which is then used to demonstrate population structure. The program assumes a model in which there are "K" number of populations, where "K" may be unknown. Within the model, each population is characterized by a set of allele frequencies at each locus, and individuals are assigned to populations on the basis of their genotypes, while at the same time, the population allele frequencies are estimated. Pritchard et al. (2000) used microsatellites, SNPs and Restriction Fragment Length Polymorphisms (RFLPs) as markers and made the assumption that the markers are unlinked and are at linkage disequilibrium (LD) with each other in the populations. The model also assumes that all markers used are at Hardy-Weinberg equilibrium within the populations. This population model is known as STRUCTURE, and has been used to investigate population admixture in Chapter 5 of this thesis, where STRUCTURE will be discussed in more detail.

Due to the disadvantages of DBMs, Pritchard et al. (2000) used the model-based method of a computer software program (STRUCTURE) to analyse 30 biallelic RFLPs in Africans and Europeans. Using a K=2 setting, (i.e. setting a limit of two possible populations to which STURCTURE could assign the RFLPs), they obtained two distinct clusters; one for the African population and another for the Europeans (see Figure 1-4).

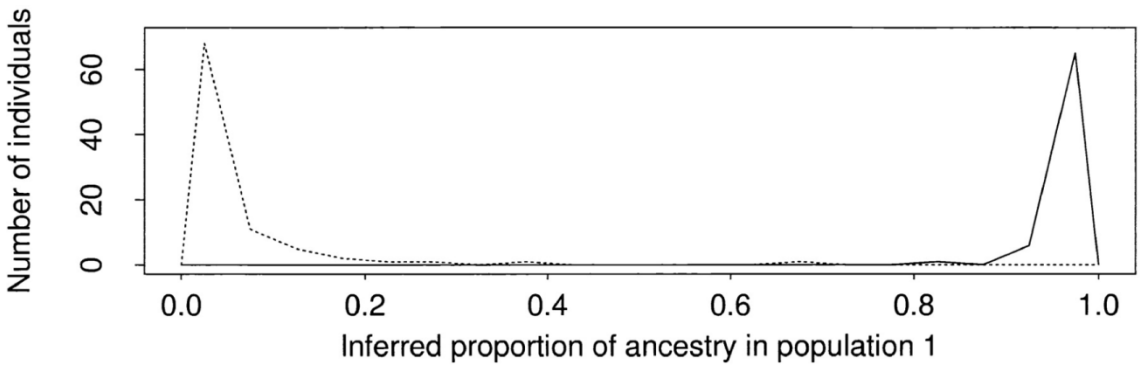


Figure 1-4. A summary of the clustering results for the data from the African and European populations. The dashed line represents the European samples and the solid line represents the African samples. Source: (Pritchard et al. 2000).

With settings of K=3, 4 and 5, the results obtained continued to reflect the population structure of two distinct populations. When geographical data was also included, there was much better clustering of the populations. The model used also allowed for the addition of information about the existing extent of admixture within a sample where the genetic makeup of individuals is drawn from more than one of the "K" populations.

STRUCTURE was also used to analyse 337 autosomal STR loci belonging to 1056 individuals from 52 different populations (Rosenberg et al. 2002) in the HGDP-CEPH panel (Chapter 4). Having already ascertained that 85% of genetic variation is within populations, and only 15% of genetic variation is between populations (Lewontin 1972), it was necessary to find a reliable method for identifying population structures from the 15% of between population variation. Rosenberg et al. (2002) identified 6 main genetic clusters, 5 of which corresponded to the major geographical regions, and when STRUCTURE was used with a K=6, some clusters were highlighted where there

were so-called “private” alleles; for example, one such cluster was made up of the Kalash people, an ethnic group of the Hindu Kush mountain range, in the North-West Frontier Province of Pakistan.

In several populations, there were individuals who were assigned to more than one cluster and this could either reflect a continuous gradation in allele frequencies, or it could be a signal of admixture between neighbouring groups. Using STRUCTURE, Rosenberg et al. (2002) found that self-reported affiliations to populations were highly informative and reliable, except for where there had been a recent history of population admixture.

Additionally, populations often clustered into families of language groups rather than purely geographical proximity, for example the Hazara (a Persian-speaking people from the central region of Afghanistan and northwestern Pakistan) and the Uygur (a Turkic ethnic group living in Eastern and Central Asia), clustered together, as did the Tu (they are distributed throughout the 31 provinces and regions of China, but with a higher concentration in the northwest) and the southern Chinese populations.

Europe had the smallest among-population variance (only 0.7%) and it proved to be more difficult to detect population structure in Europe. K=3 did not show up any structure, but in additional runs of STRUCTURE it was just possible to detect differences between the Basques and the Sardinians. The Russians grouped with the Adygei (from a republic of Russia, located in south-east

Europe in the north foothills of the Caucasus Mountains) and the Orcadians, which was possibly due to shared Viking ancestry (Haywood 1995).

Overall, using STRUCTURE, Rosenberg et al. (2002) found that world-level boundaries between the major clusters corresponded to the physical barriers such as oceans, the Himalayas and the large deserts, such as the Sahara. This led to the conclusion that populations were structured along the lines of geographical boundaries and language, and that there were no gradual clines present in the structure of human populations. The picture of population structure which emerged depended on the types of markers used, and if markers were particularly informative, then fewer were needed to get the same result.

Evolution, which is the change of the genetic material of a population over time, results in the accumulation of small differences and over a long period of time, these differences can result in substantial changes in the population. However, as the process of evolution is affected by a range of different factors, from environmental to cultural, the process is very difficult to re-create accurately in computer models. Some stochastic modelling processes attempt this, and certainly those which use the coalescent model (Kingman 2000) try to take as many random factors as possible into account. Therefore, it is not really surprising that there are alternative views to population structure, and one of the most convincing is the argument for a clinal structure to human population,

which has been proposed by others in the field (Bamshad et al. 2003; Serre and Paabo 2004).

Ever more complex geographical models are continually being developed and used to answer questions about human populations and structure. However, as there are often many interpretations to the answers received, it appears that currently, no one model is effective enough to provide definitive answers. We can not effectively take into account all the factors which may have affected our ancestors as they colonized new habitats. They were exposed to different environments to the ones we know today, and may also have encountered different pathogens – a signature of which can be seen in the distribution of the Hb^s mutation which causes sickle cell anaemia and which has a high frequency in Africa (Allison 1954). However, it is not so simple to test whether any one particular polymorphism (such as the Hb^s mutation) has been actively selected for. Given that all polymorphisms are affected by past demography (whether they have been selected or not), neutral genomic regions should fit into a general pattern, and those which are under selection will therefore deviate from this pattern (Biswas and Akey 2006). So, if we focus on the clinal pattern which is seen, then we should be able to infer some of the key parameter of human settlement histories.

1.7 The argument against a structured demographic

As has been stated previously, most human genetic diversity is between individuals rather than between populations and/ or continents. In 2003 DNA

polymorphisms were used to look for population structure (Bamshad et al. 2003) and the results suggested that genetic diversity was organised in a series of continental clades.

Historically, the DNA samples have been collected from defined "populations", for example, such as the Norwegians, the Yorubans, the San, etc. These populations are defined by a common language, common cultural practices, and sometimes by religion or by having a common myth of origin (for, example, the Dreamtime Stories of Australian Aboriginals). In these DNA samples, often admixed or individuals with mixed ancestry are omitted from the analysis. However, the problem arises in using the "population" as the unit because very often the cultural traits which have been used to define the population are much more recent (1 – 2 KYA) than the actual population structure. This introduces one level of bias into the possible results.

The second possible source of bias in DNA collection is that sometimes samples are collected on a geographical basis. This, however, has not been practiced widely due to logistical difficulties. Where samples have been collected from large numbers of individuals from a restricted geographical region, they have shown spatial gradients of allele frequencies which are only disrupted by linguistic or geographical barriers (Rosser et al. 2000; Karafet et al. 2001).

The presence of strong geographic patterns which showed up as clines within a population structure established the principle of Isolation by Distance (IBD)

models to describe the pattern of human genetic diversity. IBD models suggest decreasing gene flow with increasing distance. World-wide studies which have been based on DNA from populations have tended to find that individuals cluster depending on their continent of origin (Bamshad et al. 2003), and this is sometimes taken to imply that human genetic diversity is actually structured according to "race". Race is yet another term in population genetics which is very difficult to define and has been used in the past as a basis for biomedical research of ethic groups within populations (Risch et al. 2002; Burchard et al. 2003). Race may only be useful in biomedical research or even in clinical practice if the term "race" is used to define a set of environmental factors which define the background of an individual, and which may then subsequently assist in identifying rare disease alleles. However, if this is the case, then cultural factors should also be taken into account. Results do tend to differ between regional and global studies, and this may be due to gradients in allele frequencies being restricted to smaller geographic regions. The continents could therefore be distinguished by discontinuities in genetic diversity.

In order to ascertain whether the results achieved by Rosenberg et al. (2002), showing a highly structured population were real, or artefacts of ascertainment bias, or study design (Serre and Paabo 2004) carried out the same analyses, but changed the sampling methods. It was found that if the sampling was based on individuals and geography rather than on populations, then they tended to see a gradual variation and isolation by distance on a world-wide scale. (Rosenberg et al. 2002) did see some individual clustering using STRUCTURE, along

continental lines, as well as signals of admixture; however this does not provide any information on population histories. For example, when K=6, the Kalash population separated off from Pakistan, however, this is not a major subdivision of real populations.

In order to get a more realistic picture of population histories, the sampling needs to maximise the geographic distribution of the samples taken and also keep sample sizes similar for each geographical area. This would avoid the creation of false substructures. (Serre and Paabo 2004) tested this and they did achieve different results to those obtained by Rosenberg et al. (2002). They used a subset of the CEPH-HGDP samples and the STRUCTURE model parameter of "uncorrelated allele frequencies" (UAF). They were, however, able to replicate Rosenberg's findings, when the "correlated allele frequencies" parameter was used, despite using a smaller sample size than Rosenberg et al. (2002). Using the smaller sample size and the UAF, individuals tended to cluster according to continents, where K=4 separated the sample into Africa, Europe, American and Asia/ Oceania. When the whole CEPH-HGDP panel was analysed using STUCTURE, a world-wide pattern of admixture became apparent, as did the sampling gaps in this admixture pattern, such as the lack of Native American samples. The samples, however, do show a structure and diversity which is organised along continental lines, and this is especially the case when the sampling is sparse across the region. However, in cases where the sampling is more evenly distributed, then the populations to not form discrete clusters along the continental lines. Instead, "theoretical" populations group together,

with some individuals showing admixture with at least two of these populations.

These results indicate that the type of answers you obtain are largely dependent on the type of questions you are looking to answer. This is very nicely demonstrated by the fact that Rosenberg et al. (2002) set out to investigate whether individuals could be assigned to culturally pre-defined populations on the basis of their genotypes, whereas Serre and Pääbo (2004) investigated the patterns of relatedness across the human gene pool. The results of this showed a gradual variation, and an isolation by distance model rather than major genetic discontinuities was more typical of human diversity globally. There may still be some local discontinuities, for example, between people from different linguistic groups, and also between some continental groups, but these can only be discerned if enough markers are used. The overall pattern of human diversity which emerged was a clinal one. There was also no evidence that the genetic structure of human populations were divided along the lines of race. Figure 1-5 below illustrates the way in which sampling can present a false structured effect whereas the reality is a more clinal pattern.

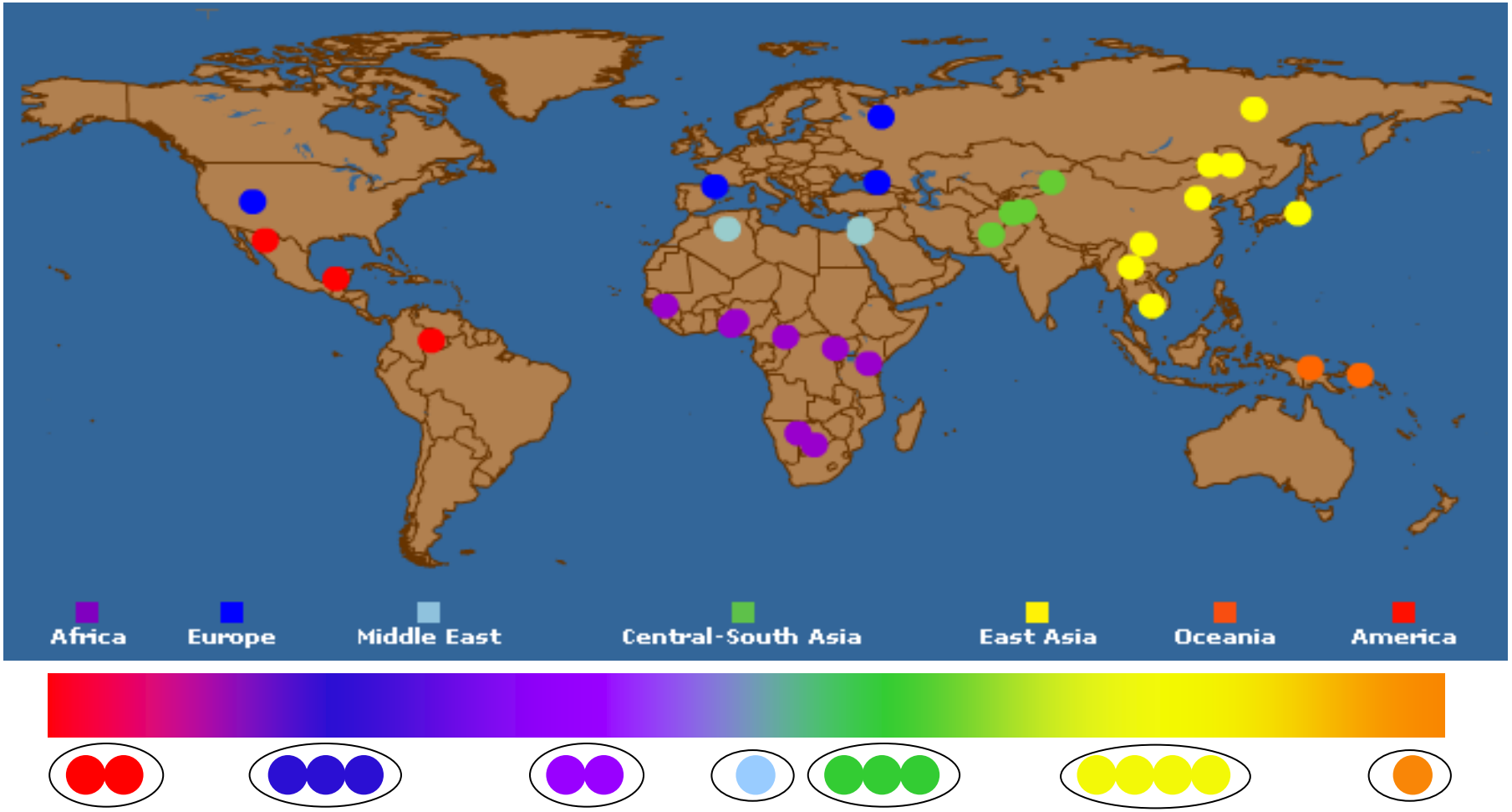


Figure 1-5. CEPH-HGDP populations showing sample collection locations and how sampling may give false clusters in analyses.

The CEPH-HGDP is a global panel of DNAs. Genetic clusters which are actually biologically meaningless are seen as a result of heterogeneous sampling. The gradation in colour from red to orange represents a hypothetical situation of continuous variation in allele frequencies. If sampling is heterogeneous (as can be seen by the small circles underneath the colour bar, each representing a population), then the clinal pattern of variation is lost and instead a clustered pattern is seen.

1.8 Combining different marker types - SNPSTRs

In 1996 Tishkoff and colleagues (Tishkoff et al. 1996), used a pair of linked genetic markers on chromosome 12 (an STR and a partial deletion of an ancient *Alu* retrotransposon insertion) and demonstrated their global pattern of haplotype frequency variation and linkage disequilibrium. The pattern, according to these authors, demonstrated evidence supporting a common and recent African origin for all non-African human populations. This work demonstrated the evolutionary information contained in the autosomal haplotypes composed of different classes of markers.

Following on from this work, Mountain and colleagues (Mountain et al. 2002) developed a method to enable the definition of autosomal haplotypes based on a pair of linked markers consisting of a SNP with a nearby STR polymorphism – a SNPSTR. Within each SNPSTR, the SNP and STR are physically linked.

With the mutation rates of the SNP and STR differing markedly, these two types of genetic markers can provide complementary evolutionary information. Some SNPSTRs may be globally informative, others may be regionally more informative. Also, each SNPSTR system, when unlinked to others on recombining chromosomes, is able to provide independent information about the evolutionary history of humans.

Additionally, SNPSTRs are relatively easy to ascertain in one single reaction by using techniques that will only amplify the section of DNA required and then

labelling each SNP with a different fluorescent dye, which can easily be detected. This makes it possible to carry out the typing of large numbers of DNA samples in a relatively short space of time.

As with other sets of polymorphisms (Pamilo and Nei 1988), using the information provided by several different SNPSTRs increases the power to distinguish between alternative theories and may enable us to gain a fuller picture of the history of our autosomal DNA.

1.9 Overview of this thesis

This thesis sets out to establish whether using SNPSTRs based on the well-documented forensic STRs, would provide a fuller insight into the histories and structures of human populations. Mountain et al. (2002) showed that even one SNPSTR system is enough to provide insight into population histories, and this was using STRs of di-nucleotides of more than 10 repeat units.

The advantage of using SNPSTRs based on the forensic STRs is that much data is already available about the distribution and mutation rates of these STRs, as well as information on variant alleles, and this data can be taken into account in any subsequent analysis of this work. Additionally, the forensic STRs are formed of tetra-nucleotide repeat units, and this make them more stable and them less liable to slippage.

After a Materials and Methods chapter, Chapter 3 of this thesis gives further details of SNPSTRs in general and describes the markers chosen for this work, and how the systems were established and validated.

Chapter 4 of this thesis uses the SNPSTRs to explore whether using autosomal DNA markers shows a different picture of population structures to that seen using mtDNA and the Y chromosome in general global samples.

Chapter 5 of this thesis then examines admixed populations and asks whether SNPSTRs can be used as markers to detect signals of admixture which had previously only been detected using Y chromosome or mtDNA analysis.

Finally, in Chapter 6, SNPSTRs are expanded into larger haplotype blocks to ask whether these larger haplotype blocks based on SNPSTRs are informative as markers for population histories and structures.

2 Materials and Methods

2.1 Materials

2.1.1 Chemical and molecular biology reagents

Materials were purchased from the following companies: Applied Biosystems (Applera) (Warrington, UK), Amersham pharmacia biotech Ltd. (Buckinghamshire, UK), Applied Biosystems (Foster City, USA), Fisher Scientific Limited (Loughborough, UK), New England Biolabs (Hertfordshire, UK), Sigma-Aldrich (Poole, UK), Stratagene (La Jolla, USA), Qiagen Ltd. (Crawley, UK), Fermentas Life Sciences (York, UK), FMC Bioproducts (Rockland, USA), National Diagnostics, (Hull, UK), Thermo Fisher Scientific (Ulm, Germany), Promega Corporation (Southampton, UK).

2.1.2 Oligonucleotides

Oligonucleotides, including fluorescently labelled oligonucleotides and SNaPshot oligonucleotides, were synthesized by Thermo Fisher Scientific and Sigma Aldrich and were provided after HPLC purification. Full information on the primers used will be given in relevant Chapters of this thesis.

2.1.3 Commonly used solutions

NDS	0.5 M EDTA, 10 mM Tris-HCl, 1%(w/ v) sodium lauroyl sarkosine, pH9.5
TE	10 mM Tris-HCl pH7.5, 1 mM EDTA (pH8.0)

10X TBE buffer	0.89 M Tris-borate, 2mM EDTA (pH8.3)
5X Bromophenol blue loading buffer	5X TBE, 0.25% (w/ v) bromophenol blue, 0.25% (w/ v) xylene cyanol, 15% (w/ v) Ficoll (type 400, Pharmacia) in H_2O
Shrimp Alkaline Phosphatase (SAP)	supplied in 25 mM Tris-HCl (pH 7.6 at 4 °C), 1 mM $MgCl_2$, 0.1 mM $ZnCl_2$ and 50% (v/ v) glycerol
10X SAP Reaction Buffer	50 mM Tris-HCl (pH 9.0 at 37 °C), 10 mM $MgCl_2$

2.1.4 DNA samples

DNA samples from the CEPH-HGDP panel (Cann et al. 2002) and the HapMap populations (HapMap 2003; HapMap 2005) were used at a concentration of ~5ng/ µl. These samples were available as pre-extracted DNA. Other populations sampled used were 70 unrelated Inuit males from six different locations in Greenland (Nanortalik, Nuuk, Ilulissat, Uummannaq, Upernavik and Ittoqqortoormiit) and 62 unrelated Danish males from Copenhagen, collected by Søren Nørby (Copenhagen). These samples were in the form of blood from which DNA was extracted using a Qiagen kit as outlined below in

section 2.2. The remaining DNA samples used were from Cornish and African-Caribbean DNA samples already held in the laboratory collections as pre-extracted DNA.

Samples for assay development came in the form of buccal swabs from fellow laboratory members and DNA was extracted from these as outlined below.

Full information on all DNA samples used will be proved in later Chapters.

2.2 Methods

2.2.1 Collection of buccal samples

The following were used for collection of buccal samples (King et al. 2006):

- A flat-bottomed, screw-top 2 ml tube (Sarstedt) containing 750 µl of NDS
- A sterile brush (Rocket Medical cervical cytology brush – R57.483)

DNA for controls and testing of methods was provided by laboratory members and self-sampled by brushing the inside of the cheek and re-suspending cellular material in the NDS buffer. This procedure yielded ~600-700 µl of NDS containing lysed buccal cells. DNA was then extracted from these buccal cell suspensions using the method outlined below.

2.2.2 Extraction of DNA from buccal and blood samples

Qiagen kits (QIAamp DNA Mini Kit - Cat no: 51306*)* were used for speed and ease of method. Buffers and solutions were all provided in the extraction kit supplied by Qiagen Ltd.

2.2.3 Extraction procedure for buccal cell lysates

20 µl of Qiagen protease/ proteinase K stock solution was placed in a 1.5 ml microcentrifuge tube. 200 µl of buccal cell lysate in NDS and 200 µl of buffer AL was added to this. This step is to ensure cell lysis and digestion of any proteins. The mixture was vortexed for 15 seconds and incubated at 56 ˚C for 10 minutes, and then centrifuged for 1 minute at 6000 x g (8000 rpm) in an Eppendorf centrifuge to consolidate the contents. 200 µl of 100% ethanol were added, and vortexed to mix. A QIAamp spin column was placed in a 2 ml collection tube, and the mix was added to the centre of the spin column, and centrifuged at 6000 x g (8000 rpm) in an Eppendorf centrifuge for 1 minute. During this step the DNA adsorbs to the silica matrix. The collection tube and filtrate were discarded. The spin column was placed into a new 2 ml collection tube for the subsequent washing steps. A 500 µl volume of wash buffer AW1 was added, and centrifuged at 6000 x g (8000 rpm) for 1 minute. The collection tube and filtrate were discarded, and replaced by another collection tube. A second 500 µl volume of wash buffer AW2 was added, and centrifuged at 20,000 x g (13,000 rpm) for 3 minutes. The collection tube and filtrate were discarded and replaced by another collection tube, and the column was

centrifuged at 13,000 rpm for 1 minute. The spin column was placed into a clean 1.5 ml microcentrifuge tube, and the DNA was eluted from the silica matrix by adding 100 µl of the low-salt buffer AE (or distilled H_2O in some cases). The mixture was incubated at room temperature for 1 minute, and then centrifuged at 6000 x g (8000 rpm) for 1 minute. In order to increase the yield by up to 15%, the 100 µl of elute was added back to the same spin column and passed through the same silica matrix. It was again centrifuged at 6000 x g (8000 rpm) for 1 minute.

This method resulted in ~100 µl of DNA solution of a concentration that was unknown, but estimated from gel electrophoresis of a small number of samples and from its performance in PCR reactions to be 2-10 ng/ µl, with the variability probably introduced by differences in use of the sampling brush, and perhaps 'shedder' status of buccal cells.

2.2.4 Extraction procedure for blood samples

This was carried out by the same method as for the buccal cell lysates above, except that all processes were carried out in a laminar flow hood and all used tips and tubes were disposed in a clinical waste bin.

This method also resulted in ~100 µl of DNA solution of a concentration that was unknown, and the variability in DNA concentration was most likely due to the degraded nature and length of storage of the blood.

2.2.5 Re-hydrating lyophilized oligonucleotide primer stocks

Lyophilized primers were routinely ordered from Thermohybaid and Sigma Aldrich. The lyophilized pellet was re-suspended in a volume of sterile distilled water calculated to create a 100 μM stock solution.

2.2.6 SNPSTR haplotyping

Each SNPSTR was haplotyped using a labelled SNP-specific primer and an unlabelled primer lying adjacent to the STR in a multiplex reaction. All analyses were based on the Polymerase Chain Reaction (PCR) (Saiki 1985).

2.2.7 General procedure for DNA amplification

Assays employed a final concentration of 1 μM for each primer for most of the assays undertaken, 1 X AmpliTaq Gold buffer, 1.75 mM Mg^{2+}, 300 μM dNTPs (Promega) and 0.1 U/ μl of AmpliTaq Gold polymerase enzyme (Applied Biosystems), 5-10 ng of DNA, and 5% (w/ v) glycerol (Sigma). PCR reactions were cycled using an MJ Research Tetrad 2 Peltier Thermal Cycler employing a number of different cycling conditions (see later for the details of different conditions relating to each assay). To minimize contamination, PCR reactions were set up in a laminar flow hood using pre-PCR dedicated pipettes and consumables, and zero-DNA controls were also included with each set of PCR reactions. Each PCR reaction was set up in 200 μl PCR tubes.

2.2.8 Primer annealing temperature

Initial primer annealing temperature was calculated using the approximation, A/T = 2 ˚C and C/G = 4 ˚C. This allows an estimation of the annealing temperature but if initial amplifications appeared to generate non-specific amplicons, the annealing temperature was varied to determine which gave the best amplification.

2.2.9 Agarose gel electrophoresis

Agarose gel electrophoresis was carried out to verify the success of PCR amplifications. This technique was carried out using varied wells as needed in horizontal submarine agarose gels containing ethidium bromide (0.5 µg/ml) (Sigma) in 1 X TBE buffer using electrophoresis tanks manufactured at the University of Leicester and Bio-RAD electrophoresis power supplies. DNA was visualized on a UV transilluminator, and photographs were taken with a Syngene GeneFlash Gel Documentation System (Syngene, Cambridge, UK) or saved to disk. The agarose gel (Seakem, FMC Bioproducts) concentrations varied (1-3% [w/v]) depending on the size of the DNA fragments being resolved, and were run at 8 or 4 volts per cm.

For the majority of gels a size marker of ϕX174 phage DNA digested with *Hae*III (Invitrogen) was used. This gives the following fragment sizes (in bp): 1353, 1078, 872, 603, 316, 281, 271, 234, 194, 118, 72. Occasionally, a higher molecular weight size marker was required, and λ phage DNA digested with *Hin*dIII

(Invitrogen) was used, which gave the following fragment sizes (in bp): 23130, 9436, 6557, 4361, 2322, 2027, 560, 125.

2.2.10 SNPSTR typing methods

SNPSTRs were typed by means of allele-specific amplification using one labelled allele-specific primer and one unlabelled universal primer (Chapters 3 - 5) and then analysed using the ABI 3100 Genetic Analyzer and GeneScan Analysis software (ABI). This provided phased information on the SNP and STR simultaneously.

2.2.11 ABI 3100 capillary electrophoresis apparatus

The ABI3100 Genetic Analyzer (Applied Biosystems) was used for sequencing, PowerPlex 16 (Promega) DNA profiling, SNPSTR ascertainment and SNaPshot (ABI) assays. This equipment carries out capillary electrophoresis of the DNA using a laser to detect fluorescently labelled DNA fragments as they pass by the capillary detection window. All results obtained were then analysed with ABI GeneScan or GeneMapper Software (versions 3.7 and 4). The run lengths, injection times, dyesets and size standards were adjusted to suit each application being run, and are discussed more fully in the relevant later Chapters.

2.2.12 SNaPshot assays

SNaPshot (ABI) is a minisequencing method of SNP detection, validation and scoring which can interrogate up to 10 SNP loci in one single-base extension reaction. Figure 2-1 illustrates how the method works. Binary markers were typed in one multiplex using the SNaPshot minisequencing procedure (Applied Biosystems), an ABI3100 Genetic Analyzer (Applied Biosystems) and Genemapper Software Version 4. Primer sequences for the first SNaPshot SNP assay are listed in Chapter 6.

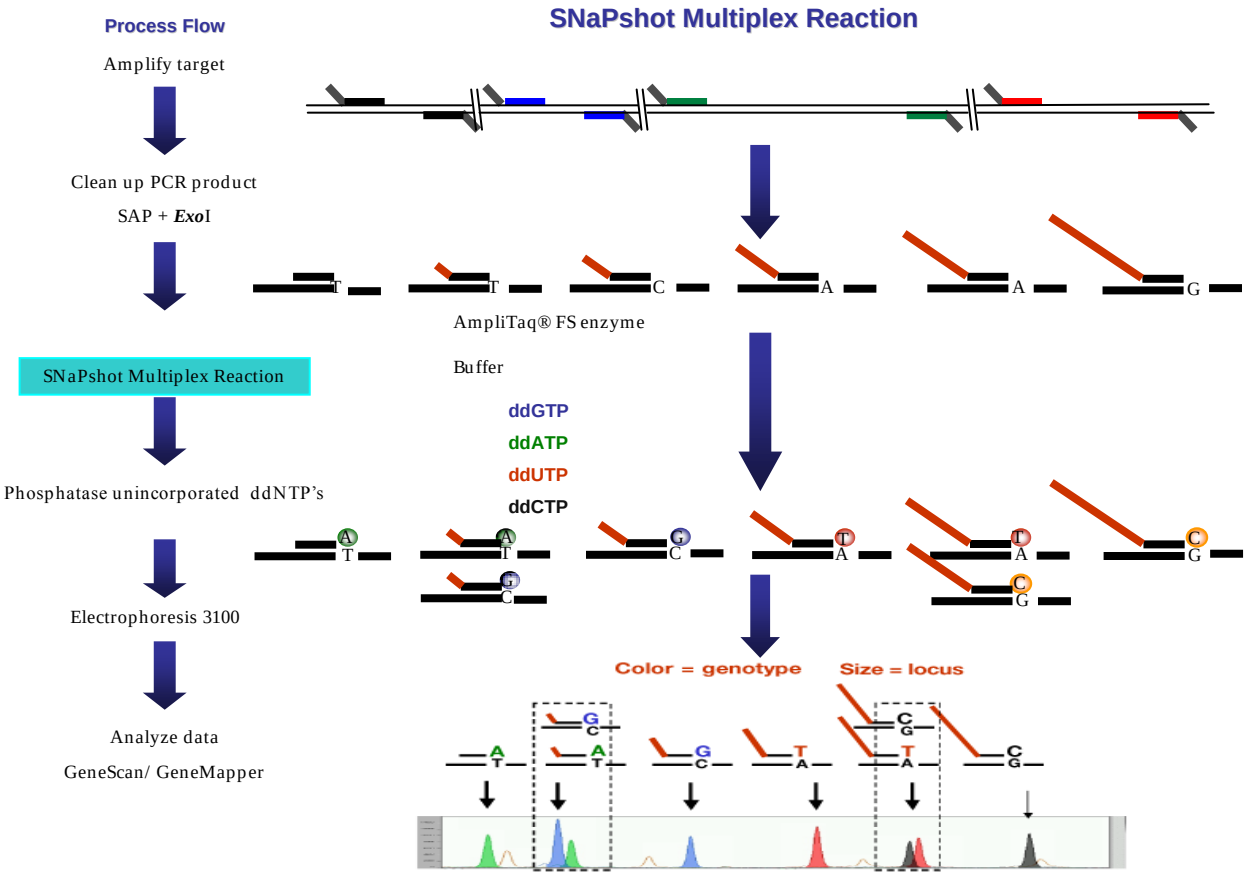


Figure 2-1. Work-flow chart of a SNaPshot Multiplex Reaction and results seen on electropherogram. Source Applied Biosystems.

2.2.13 SNaPshot flanking PCR

The initial stage of the SNaPshot reaction involved the PCR amplification of the flanking regions around each SNP being studied in a multiplex reaction. For each multiplex, each primer was used in the final reaction concentration of 1μM. These were used in a PCR reaction volume of 10μl, in conjunction with 1X AmpliTaq gold buffer, 1.75 mM Mg^{2+}, 300 μM dNTPs and 0.1 U/ μl of AmpliTaq Gold polymerase enzyme, 5-10 ng of DNA, and 5% (w/ v) glycerol. PCR programme "P7952-3" used the following conditions: 94 ºC for 10 minutes, then followed by 94 ºC for 1 minute, 55.8 ºC for 1 minute, 70 ºC for 2.5 minutes for a total of 30 cycles ending at 65 ºC for 10 minutes and 15 ºC for 5 minutes.

PCR products were treated with the enzymes Exonuclease I (ExoI; New England Biolabs) and shrimp alkaline phophatase (SAP; Promega) to digest excess primers and dephosphorylate excess dNTPs respectively. Failure to treat products in this way compromises the single-base extension reaction in the next phase of the procedure. A 5 μl aliquot of PCR reaction was incubated with 0.15 unit of ExoI and 1.5 units of SAP with 1.7 X ExoI buffer at 37 ºC for 1 hour, followed by heat inactivation at 80 ºC for 15 minutes and rapid cooling to 4 ºC for 15 minutes.

2.2.14 SNaPshot single-base primer extension

The SNaPshot primer extension uses a primer that anneals to the target so that its 3' end is immediately adjacent to the SNP of interest. The SNaPshot mix contains four fluorescently labelled ddNTPs, one of which is incorporated at the

3' end of the primer, as determined by the SNP sequence state. When the product is visualized on the ABI3100 following capillary electrophoresis the extended primer is represented by a peak on the electropherogram, the colour dependent on the base present at the SNP site (Figure 2-2).

A number of the SNP primers were given 5' polyA tails in order to alter their relative mobilities and retain an even distribution of the different extended products on the electropherogram as can be seen in the example below (Figure 2-2).

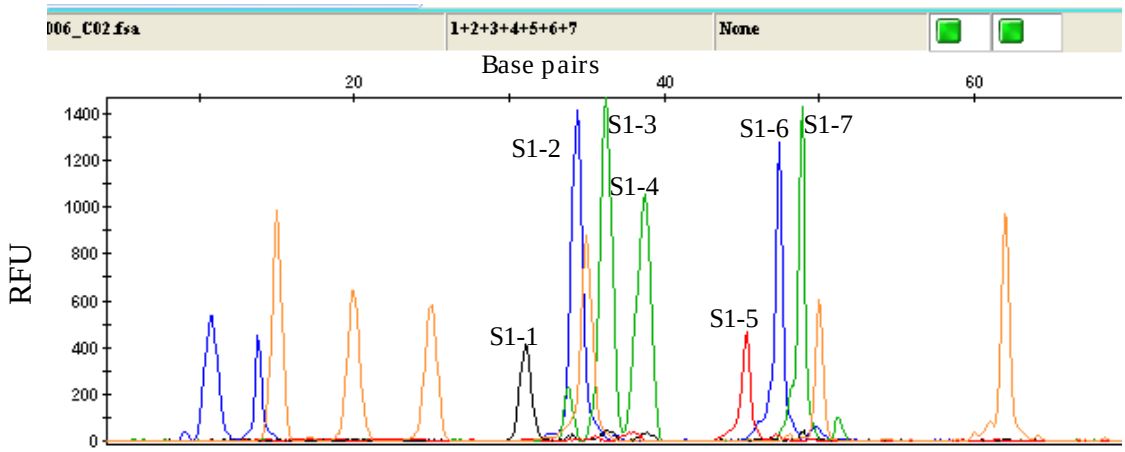


Figure 2-2. Electropherogram of 7 SNP specific primers . Showing additional polyA tails which has ensured an adequate size difference between fragments and hence an even distribution of products. The first two blue peaks are background noise and can be ignored. The orange peaks are the Liz120 size standards (15, 20, 25, 35, 50, and 62 nucleotides). Primers used are as follows; S1-1 -20mer, S1-2 -20mer$(A)_5$, S1-3 – 20mer $(A)_{10}$, S1-4 – 20mer $(A)_{15}$, S1-5 – 20mer $(A)_{20}$, S1-6 – 22mer $(A)_{23}$, S1-7 – 17mer $(A)_{33}$.

For the SNaPshot reaction, 2 µl of exonuclease-treated PCR product was added to 5 µl of SNaPshot mix (Applied Biosystems) and a 1 µl mix of primers (concentrations given in Chapter 6), made up to 10 µl with water. The extension reaction was carried out using a Thermocycler under the following conditions:

96 ºC for 10 seconds, 50 ºC for 5 seconds and 60 ºC for 30 seconds, for a total of 35 cycles, and then cooled to 15 ºC.

2.2.15 SNaPshot clean-up reaction

1 Unit of FastAp™ - Thermosensitive Alkaline Phosphatase (Fermentas) was added to each reaction to remove any remaining ddNTPs, and thus to prevent spurious peaks on the electropherograms. It does this by catalyzing the release of 5' and 3' phosphate groups from DNA, RNA and nucleotides. The reaction was incubated at 37 ºC for 10 minutes followed by 75 ºC for 5 minutes. FastAp™ was used as a final cleanup instead of using Shrimp Alkaline Phosphatase (SAP) because it cut the incubation down to 10 minutes instead of 2 hours. Trials carried out with FastAp proved that a ten-minute incubation was as effective as using SAP over a 2 hour incubation period.

2.2.16 Analysis of SNaPshot products

The ABI3100 requires a known size standard, therefore 25 µl of fluorescently labelled size standard (Liz120; Applied Biosystems) was added to 1ml of Hi-Di Formamide (National Diagnostics) and then 10 µl of this mix was aliquoted into the respective wells of an unskirted 96-well plate (Advanced Biotechnnologies), which was to contain a sample. A 1 µl aliquot of each cleaned SNaPshot reaction product was then added into these labelled wells. The sample was then denatured by heating to 95 ºC for 3 minutes and then cooled on ice for 5 minutes. The plates were spun in a centrifuge to consolidate the contents and then the plate loaded onto the ABI3100 capillary electrophoresis apparatus and

run under the default SNaPshot conditions (See Appendix 2 for ABI running conditions).

The SNaPshot assays and results are discussed in full in Chapter 6.

2.2.17 PCR for DNA sequencing

Following standard PCR amplification using double volumes (20 µl), PCR-products were purified using Qiagen columns (Qiaquick PCR Purification kit: cat. no. 28104).

2.2.18 Procedure for PCR product purification

This was carried out according to manufacturer's instructions using the Qiagen Qiaquick PCR Purification kit. Purified water was added to make the final PCR volume up to 30 µl. A 5:1 ratio of Buffer PB to volume was added (150 µl). The mixture was vortexed to mix and then spun to consolidate the contents before being added to the Qiagen spin column. It was spun at 20,000 x g (13,000rpm) for 1 minute and the supernatant was discarded. 0.75 ml of Buffer PE was added to the column and this was spun at 20,000 x g (13,000rpm) for another minute. Again the supernatant was discarded before being spun again for 1 minute to ensure that no Buffer PE remained in the spin column. The Qiagen spin column was then placed in a new clean 1 ml Eppendorf tube and 30 µl of Buffer EB was added to the column and left for 1 minute before it was spun at 20,000 x g (13,000rpm) for 1 minute. This cleaned PCR product was ready for use in a DNA sequencing reaction or stored at -20 ºC until needed.

2.2.19 DNA Sequencing

DNA sequencing was carried out using the chain termination procedure (Sanger et al. 1977). A 10 µl reaction contained 4 µl of BigDye terminator version 1.1 (Applied Biosystems), 0.37 µM primers (primer used depending on the direction of sequence required), and around 50 ng purified DNA, made up to volume with distilled water. The sequencing reaction was carried out under the following conditions: 1) 96 ºC for 30 seconds, 2) 96 ºC for 10 seconds 3) 50 ºC for 5 seconds and 4) 60 ºC for 4 minutes, repeating steps 2 – 4 for 25 cycles.

The reaction was cleaned using Qiagen Dye-Ex 2.0 Spin Kit (cat. No. 63206) to remove any excess BigDye. The columns were vortexed to mix up the gel and mesh contained within and then placed in a collection tube. The caps were loosened and the bottom sealing tabs were removed. They were then spun at 750 x g (2800 rpm) for 3 minutes. This left a gel matrix in the columns and the supernatant was discarded. 10 µl of distilled water was added to the sequenced PCR products and this final volume of 20 µl was gently added to the top of the gel matrix in the spin column. The column was placed in a clean, labelled collection tube and then spun again at 750 x g (2800 rpm) for 3 minutes. The collected, cleaned sequencing product was left to evaporate in a heating block (Techne Dry-Block DB.2A) until no liquid remained and the dried pellet was then re-suspended in 14 µl Hi-Di Formamide (National Diagnostics) prior to capillary electrophoresis on the ABI3100. The 14 µl samples were loaded into 96-well unskirted plates. Samples were denatured (heated at 95 ºC for 3 minutes and cooled on ice for 5 minutes) before being loaded on the ABI3100 capillary

electrophoresis apparatus and run using the default ABI3100 sequencing conditions (see Appendix 2).

Results were analysed using ABI Sequence Analysis Software (v. 3.7)

2.2.20 PowerPlex™ 16 System of DNA Profiling

A 10µl reaction contained 1 µl 10x Gold* Buffer (Promega), 1 µl 10 x primer mix (Promega) and 3 units *Taq* polymerase enzyme (Applied Biosystems). To this was added approximately 50 ng of DNA. The reaction was carried out in a thermocycler under the following conditions : "Power_16", 1) incubate 95 ºC for 11 mins., 2) incubate 96 ºC for 1 min., 3) incubate 94 ºC for 30 sec., 4) incubate 60 ºC for 3 sec – ramp to 60 ºC at 0.5 ºC per sec., 5) incubate 70 ºC for 45 sec – ramp to 70 ºC at 0.2 ºC per sec., 6) Cycle to step 3 for 9 more times, 7) incubate 90 ºC for 30 sec., 8) incubate at 60 ºC for 30 sec. – ramp to 60 ºC at 0.5 ºC per sec., 9) incubate 70 ºC for 45 sec – ramp to 70 ºC at 0.2 ºC per sec., 10) Cycle to step 7 for 20 more times, 11) incubate 60 ºC for 30 mins., 12) incubate at 15 ºC forever. In order to accurately size the fragments, the ABI3100 needs a size standard. 1 µl ILS600 (Promega) was added to 9 µl of Hi-Di Formamide (National Diagnostics) and 1 µl of the Powerplex PCR product. The samples were loaded into 96-well unskirted plates, denatured (heated at 95 ºC for 3 minutes and cooled on ice for 5 minutes) before being loaded on the ABI3100 capillary electrophoresis apparatus and run under the following conditions: Genescan, Dye Set Z, Red Dye, Project Name: 3100 project1, RunModule:

Genescan36pop4 default module. Results were analysed using ABI Genescan software.

2.2.21 Statistical analysis

F_{ST}

F_{ST} (Wright 1951) is a measurement of the mean amount of genetic diversity within subpopulations in comparison to that of the metapopulation. F_{ST} can also be used, therefore, as a method of estimating genetic distance between populations. It varies between 0 and 1, and where gene flow is high, and there is little differentiation between subpopulations F_{ST} will be close to zero. Conversely with subpopulations that are highly differentiated from one another the overall diversity in the metapopulation is therefore higher and much greater than that in any subpopulation and F_{ST} is closer to 1. An F_{ST} of 0.3 means that 70% of allele frequency exists within the subpopulations and that 30% of the total allele frequency variance is found between subpopulations (Jobling and Gill 2004). F_{ST} is therefore suitable for use with both haplogroup frequency data and STR frequency data, but note that it does not encapsulate any information about the molecular similarity or dissimilarity of alleles.

R_{ST}

The calculation of R_{ST} (Slatkin 1995) considers the proportion of STR diversity that occurs between subpopulations and builds in knowledge of the mechanism with which STRs mutate using the stepwise mutational model (SMM) (Ohta

1973) of STR mutation. It therefore contains information on the molecular distances between alleles.

2.2.22 Multidimensional Scaling

Multidimensional scaling (MDS) (Torgerson 1952) is a method of multivariate analysis which allows us to reduce multidimensional space to two dimensions while minimizing loss of information as a way of displaying genetic distances between populations in a graphical way.

2.2.23 Genetic Distances

Arlequin 2.00 (Schneider et al. 2000a) was used to calculate genetic distances, using haplogroup frequency data, in the form of pairwise F_{ST} values (with associated p-values generated by permutation) for the geographically defined populations. R_{ST} (Slatkin 1995) analysis, also using Arlequin 2.00, was utilized to calculate pairwise genetic distances using the corresponding STR data. In each case, these distances then formed the basis for a multidimensional scaling analysis carried out under PROXSCAL in SPSS 12.0.

2.2.24 Work carried out by other people during the course of this PhD research

It would be amiss not to acknowledge the contribution of others who have helped with data typing and analysis during the course of this PhD research project. With this in mind:

• The Cornish samples were typed by Paraskevi Christofidou, a final-year project student.

• Stéphane Ballereau located and derived the PHAXs which form part of Chapter 6 of this work.

3 Choice, Development and Validation of SNPSTRs

3.1 Introduction to SNPSTRs

A SNPSTR is a polymorphic STR (short tandem repeat) with a closely physically linked (≤ 500 bp) SNP (single nucleotide polymorphism). A SNP is a single base substitution, insertion or deletion of any one of the four bases (A, C, G or T), and an STR is a tandem repeat array with a 2 - 6 base pair (bp) repeat unit which is polymorphic in length (usually up to 30 repeats) due, most likely, to replication slippage (Schlotterer and Tautz 1992) or defects in mismatch repair. Analysing SNPSTRs using SNP-allele-specific PCR allows the typing of variation at both loci simultaneously and thereby provides information not only about genotypes, but also about phase and hence the direct determination of haplotypes. As the mutation rates of the SNP (~ 10^{-8}) (Nachman and Crowell 2000) and the STR (~10^{-3} per locus per gamete per generation) (Brinkmann et al. 1998) are very different, the derived SNP can provide a 'tag' for a subset of STR alleles which will provide information on population relationships. Individually, SNPSTRs are not very informative; however, the power comes from combining them with other SNPSTR systems or expanding the SNPSTR into a PHAX (Phylogeographically informative Haplotypes on Autosomes and X-chromosomes). PHAXs are discussed in Chapter 6 of this thesis. Additionally, because SNPSTRs are spread throughout the genome, selection is unlikely to be acting on all of them simultaneously at any one time. This reduces the effects of selection which may be present if one large autosomal block is used. Figure 3-1 below shows a simplified diagram of a SNPSTR

system and how the results would appear once analysed by capillary electrophoresis.

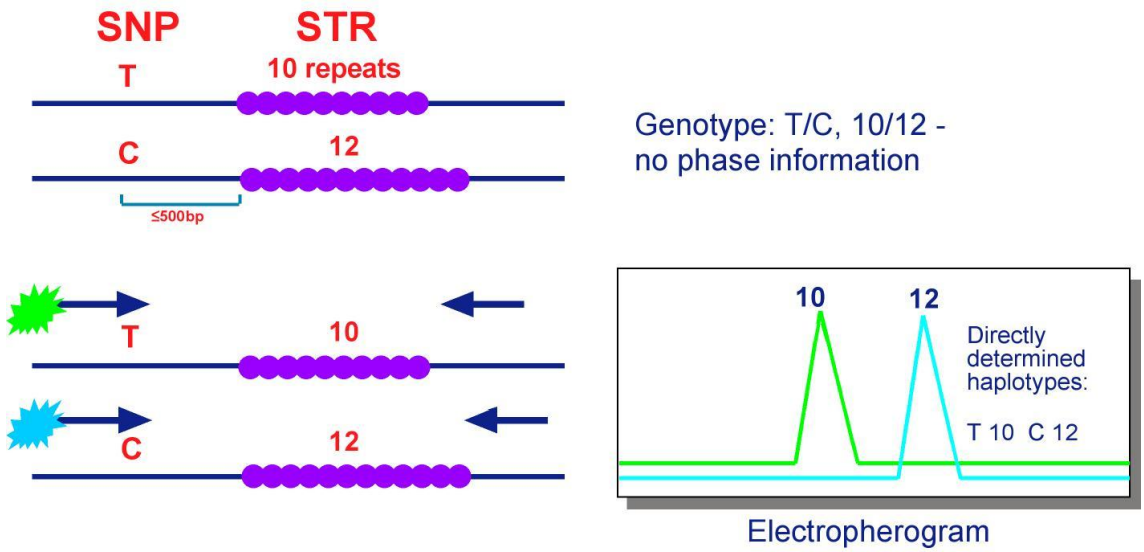


Figure 3-1. Diagram showing SNP alleles and the STR allele repeat numbers at a SNPSTR. Using labelled primers as seen in the lower part of the figure, it is possible to obtain phased data for both SNP and STR simultaneously.

3.2 Forensic STRs as suitable STRs for creating SNPSTRs

For this research, rather than use using the dinucleotide STR systems previously used (Mountain et al. 2002), or ascertaining novel STRs, it was decided to look for alternative ones which were already very well characterized. The following features have been listed as being desirable for any STR system (Butler 2003):

- A narrow size allele range that permits multiplexing
- High discriminating power
- A regular repeat unit
- High heterozygosity

- Robust PCR amplification with low stutter
- A known mutation rate

All of these features exist in the tetranucleotide and pentanucleotide repeat STRs which have been popular among forensic scientists for many years. See Figure 3-2 below for a list of the different forensic STRs in the Promega PowerPlex16 kit (an example of a commercial STR multiplex) and the respective investigation forces which use them for DNA analysis.

Table 1. The PowerPlex™ 16 System and STR Loci Selected for Databasing Standards.

PowerPlex™ 16	CODIS	Interpol	ENFSI	GITAD
Penta E	CSF1PO	FGA	FGA	CSF1PO
D18S51	FGA	D21S11	D21S11	TH01
D21S11	TH01	TH01	TH01	TPOX
TH01	TPOX	vWA	vWA	D16S539
D3S1358	vWA		D8S1179	D7S820
FGA	D3S1358		D18S51	D13S317
TPOX	D5S818		D3S1358	
D8S1179	D7S820			
vWA	D8S1179			
Amelogenin	D13S317			
Penta D	D16S539			
CSF1PO	D18S51			
D16S539	D21S11			
D7S820				
D13S317				
D5S818				

Figure 3-2. The STRs covered by the Promega PowerPlex 16 System.
CODIS - The Combined DNA Index System, used by the FBI, Interpol - The International Criminal Police Organization, ENFSI – European Network of Forensic Science Institutes, GITAD - Grupo Iberoamericano de Trabajo en Análisis de DNA. Source: The PowerPlex TM 16 System, Promega (Sprecher 2000)

The forensic STRs are very widely used by forensic laboratories world-wide, and much population data are available on them. In addition to this, due to having been used in paternity testing for many years, the mutation rate per generation (including sex-specific mutation parameters) is also known to a high degree of precision. It is important to know the mutation rate, as this enables us to calculate the TMRCA of the SNP-associated allele, and also provides information on losses and gains, and maternally versus paternally biased mutations. Many alleles have been sequenced in the forensic context and this has provided data about their internal structural variability. So, for these reasons and also because using these STRs would make this work of interest to forensic geneticists who use the STRs, it was decided to use the forensic STRs as a basis for novel SNPSTR systems.

3.2.1 Forensic STRs and Mutation

Some of the first forensic STR loci that are still used today were originally characterized in 1991. They were discovered "by accident" in many different laboratories whilst carrying out disease gene location studies, for example the forensic STR D8S1179 was used in the localization of a gene connected to Meckel-Gruber Syndrome (Morgan et al. 2002), as well as being the most commonly examined locus associated with the gene responsible for urinary microalbuminuria (Fox et al. 2005).

To date, approximately 40 different STRs have been used in forensic science; however, by consensus, between 10 and 13 common loci are now used in most

countries. In the USA, information on the 13 core CODIS (Combined DNA Index System) STRs has been collected for 10 years and these loci therefore dominate the genetic information that is maintained on DNA databases. In the UK, additional loci are used.

Figure 3-3 shows the locations of the forensic STRs in the commercial PowerPlex 16 kit and their positions on the chromosomes.

The first forensic STR kits contained only four loci; THO1, FES/ FPS, VWA and F13A1 (Kimpton et al. 1994) but today many different kits are on the market (an example is the AmpFℓSTR® SEfiler Plus™PCR Amplification Kit from Applied Biosystems). One that is commonly used in the UK is the PowerPlex 16 kit (Promega) which targets 15 STRs on the autosomes and includes the amelogenin sex test (Sullivan et al. 1993). It allows co-amplification and 3-colour detection of 16 loci. STR length variation is detected via electrophoretic separation and then fluorescent excitation. The three colours used are; fluorescein (FL) – blue, carboxy-tetramethyl rhodamine (TMR) – yellow, and JOE – green. In addition to this, there is a size standard present which is labelled with CXR – red. The amplified products are less than 500bp long and this enables degraded DNA to be used, and because the products are discrete and consist of separate allele lengths, it is possible to construct allelic ladders which enable easy comparisons between the results from a sample and the ladder.

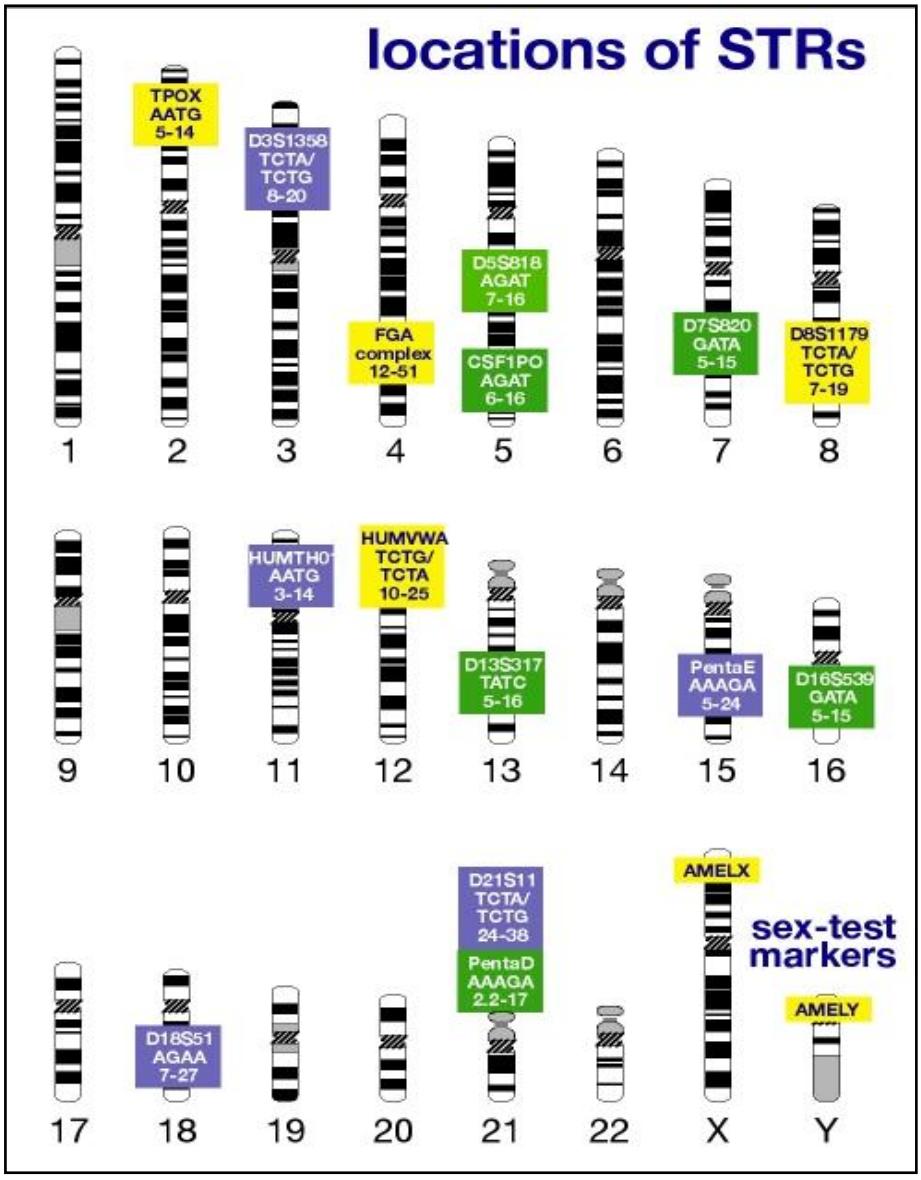


Figure 3-3. Diagram showing the locations of the 15 forensic STRs on the autosomes and the amelogenin sex test present in the Promega PowerPlex 16 Kit. The boxes contain the name of the STR, the repeat motif and the range of repeats possible. The colours indicate which dyes are used to detect the STRs Source: Adapted from (Jobling et al. 2003).

The core loci which are on separate chromosomes will segregate independently of one another during meiosis due to the independent assortment of chromosomes. This enables the product rule to be used to estimate random match probabilities with DNA profiles generated from multiple loci. Those markers that lie on the same chromosome, for example, D5S818 and CSF1PO on

chromosome 5, are far enough apart from each other (approx 26.3 megabases) to be separated by recombination events.

Each of the forensic STRs shows variation through mutation – theoretically all have resulted from one "founder" individual at some point in the past and all subsequent changes are due to mutations over time. STRs also show variation through independent assortment of chromosomes during meiosis, and variation is also achieved through recombination

As mentioned previously, a major advantage of using forensic STRs in this project is that their mutation rates are well known and through the course of paternity testing, a database of these rates has been kept. Knowing the rate will not only enable calculation of the TMRCA, but will also be of use in other population evolutionary studies where specific population genetic questions are being investigated. For example, we could envisage studying population isolates where a population had been founded by very few individuals, and estimating the number of founders using knowledge of the mutation rate. STRBase, created by John M. Butler, (*www.cstl.nist.gov/biotech/strbase*) contains up to date information on all of the forensic STRs as well as those which are not so commonly used today and is a valuable reference source. Figure 3-4 shows the mutation rates for the STRs used in the PowerPlex 16 kit. The PowerPlex 16 forensic STRs were used for this study because the kit was already available in the laboratory and this made it more economical to use this set of STRs rather than any other. From Figure 3-4 it is easy to see that the mutation rates are all

well below 1% per generation, and this is important for paternity testing (from which these data have been collated) because links are made between the child and the alleged father based on the assumption that the alleles will remain the same as they are passed from one generation to the next (Weir 1996). However, in the case of mismatches the likelihood of true paternity can still be assessed by incorporating the probability of mutation into the calculations (Brinkmann et al. 1998).

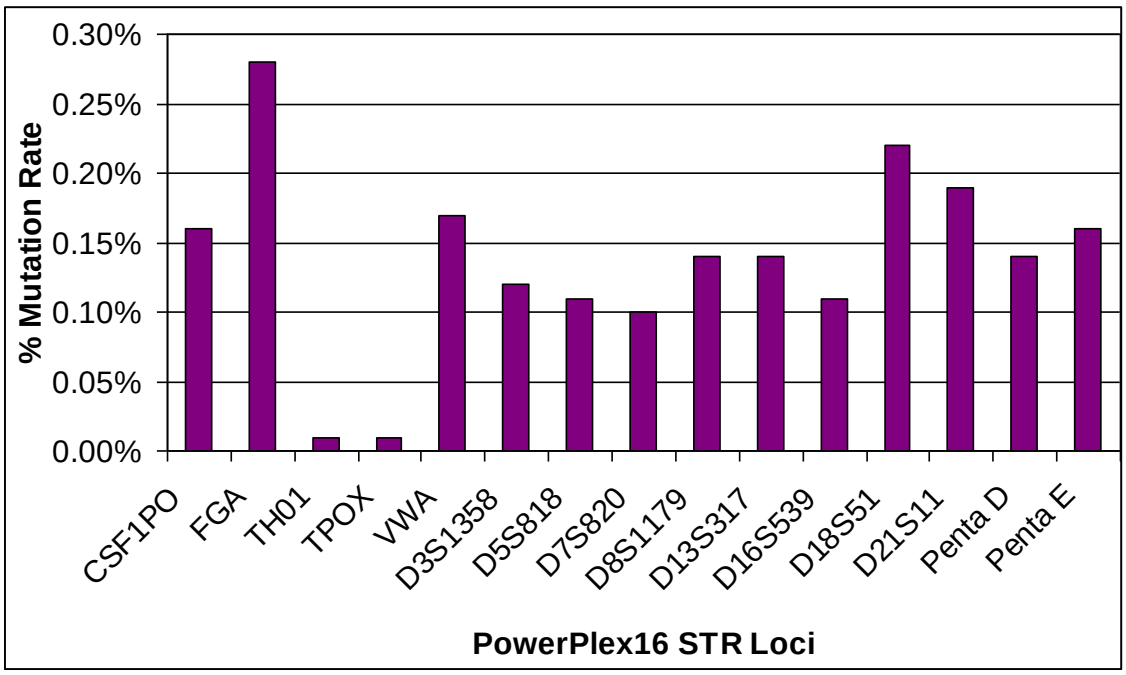


Figure 3-4 . PowerPlex16 STR loci mutation rates observed during the course of paternity testing. The rates are low, for example, the TH01 locus showed mutations in 83 out of 627,253 cases, and the D5S818 showed 924 mutations in 784,468 cases. Source: STRBase (*www.cstl.nist.gov/biotech/strbase*).

None of the 18 core STR loci are in gene coding regions (although some are close to them) and nor are they trinucleotide repeats which in some cases can be prone to expansions that cause genetic defects. To date, there has been only one call (Szibor et al. 2005) to remove a particular STR from human identity testing due to its proximity to a coding region which is directly linked to several genetic diseases. This is the HumARA CAG repeat which is located in a coding region (androgen receptor gene, exon 1) and which codes for a polyglutamine tract. X-linked spinal and bulbar muscular atrophy is attributed to a mutation at this locus where there are more than 43 trinucleotide repeats (Laspada et al. 1991). In addition to this, typing the HumARA locus can also detect other health risks such as impaired spermatogenesis (Tut et al. 1997), cardiovascular risk factors (Zitzmann et al. 2003) and several more. For these reasons, the

HumARA locus may not only be under active selection, but there are also ethical questions raised by typing this locus and revealing the number of repeats present in the STR alleles during routine forensic analysis. In any case, this particular STR was not widely used, and was not incorporated into the standard commercial kits.

The four forensic STRs used for this work were D5S818, D16S539, D3S1358 and CSF1PO (Table 3-5 below). These four were chosen because they were mostly located on separate chromosomes (except for D5S818 and CSF1PO), and also because they had suitable SNPs nearby (see Section 3.3 SNP Selection). D5S818 is close to a beta-globulin gene and an immune regulatory region, and it could have a role in malaria predisposition; however, it is considered selectively neutral (Sprecher 2000).

PowerPlex 16 STR	Other Names	LOCAT'N.	Chrom.	Ref and Source	STR ALLE INFO.	GENE PROXIMITY
CSF1PO	GDB:212649	Chr 5; 149.436	5q33.1	STRBase (May 2004, NCBI build 35) Ensembl Release 50 (July 2008)	AGAT (5-16)	Within the human c-fms proto-oncogene for CSF-1 receptor gene, 6th intron
D3S1358	GDB:196594	Chr 3; 45.557	3p21.3	STRBase (May 2004, NCBI build 35) Ensembl Release 50 (July 2008)	TCTA[TCTG]$_3$ [TCTA]$_{15}$, (8-20)	Within LARS2 - class 1 aminoacyl-tRNA synthetase, mitochondrial leucyl-tRNA synthetase
D5S818	CHLC.512 CHLC.GATA3F03 CHLC.GATA3F03.512 CHLC.GATA3F03.P6581 GATA-D5S818 GATA-P6581 GATA3F03 GDB:686610 SHGC-4079	Chr 5; 123.139	5q23.2	STRBase (May 2004, NCBI build 35) Ensembl Release 50 (July 2008)	AGAT (6-18)	> 1000 bp Casein Kinase I, Gamma-3; CSNK1G3, Gene map locus 5q23. Accession No. 604253 The casein kinase I (CKI) gene family encodes serine / threonine kinases that preferentially phosphorylate acidic substrates using ATP as a phosphate donor. CKI proteins are monomeric and range

						from 25 to 55 kD. They are ubiquitous, being found in the nuclei, cytoplasm, and membrane fractions of eukaryotic cells.
D16S539	CHLC.715 CHLC.GATA11C06 CHLC.GATA11C06.715 CHLC.GATA11C06.P6766 GATA-D16S539 GATA-P6766 GATA11C06 GDB:686448 SHGC-17627 SHGC-4544	Chr 16; 84.944	16q24.1	STRBase (May 2004, NCBI build 35) Ensembl Release 50 (July 2008)	GATA (4-16)	>3000 bp LOC732275 hypothetical protein similar to hCG1645603

Table 3-5 Detailed information on the four STRs used as part of the SNPSTR systems developed in this work.

3.3 SNP Selection

SNPs are usually unique mutational events and there exist several million SNPs in the human genome (Strachan and Read 2004) which on their own (and used in small numbers) are generally less informative than STRs. As previously mentioned SNPs show identity by descent as well as by state.

SNPs show variation through independent chromosomal assortment, recombination and also mutation, although the mutation rate is very low (Nachman and Crowell 2000). Their advantage over STRs is that there are many more of them in the genome, and they can also be typed in highly degraded DNA which is useful for forensic analysis. However, due to the fact that they are biallelic, on their own, they have a very low discrimination power and approximately 50 SNPs are needed to equal the low match probability of existing forensic STR multiplexes (Gill 2001). Nevertheless, work is ongoing to see if SNP profiling can perform to the same standard as STR profiling and work by Dixon et al (2006) on a 21 SNP profiling kit has shown some success when used for analysing highly degraded and low copy number DNA template. Notably, novel array-based SNP typing technologies that simultaneously type up to a million SNPs (such as the Affymetrix Genome-Wide Human SNP Array 6.0 which contains 909,622 SNPs), give individual identification and are potentially powerful tools in forensic analysis, including in the resolution of complex DNA mixtures (Homer et al. 2008).

The SNPs used in this project were selected from SNP databases typed by the HapMap project (Altshuler et al. 2005) and/ or Perlegen (a genetic tools company founded in 2000 within Affymetrix, and which typed the majority of SNPs within the HapMap Project). Those that showed a minor allele frequency of ≥ 0.1 in the HapMap populations were included in the selection and any below that value were not used as they were population-specific or fixed in certain populations and would therefore not be as informative in the first SNPSTR system trials. However, in the PHAX study (to be discussed in Chapter 6), lower-frequency population-specific SNPs were included in the analysis as they could be informative in non-HapMap populations.

A key requirement for the SNPSTR system was that each SNP was not more than 500bp away from the linked STR to minimize the possibility of recombination between the SNP and the STR during meiosis.

It was also decided to avoid any SNPs which were in LINEs or SINEs as this could have led to problems with specificity of SNPSTR amplification.

Table 3-6 lists the four SNPs finally chosen for inclusion in the SNPSTR systems which formed this volume of work.

SNP NO. And found on + or - DNA strand	ALLELES	ANCESTRAL STATE	LOCATION ON CHROMOSOME	DIST FROM STR	VALIDATED BY	MAF INFO	SOURCE	Local Recomb. Behaviour (Oxford Recomb. Map) Phax?	LOCAL REPEAT CONTEXT
rs2116791 +	G/ T - K	G	Chr 5: 149,435,793	299 Cent	Perlegen	Not fixed: 0.1 – 0.5 in Hapmap Pops.	Hapmap Jan. 2007 dbSNP build 125	Les than 1.5 cM/ Mb	Is in an INTRON of CSF1R
rs17077990 +	C/ G - S	C	Chr 3: 45556963	271 Tlm	Perlegen	Not fixed: 0.18 – 0.3 in Hapmap Pops.	Hapmap Jan. 2007 dbSNP build 125	Less than 0.13 cM/ Mb	Is in an INTRON of LARS 2.
rs25768 +	A/ G - R	G	Chr 5: 123139205	13bp Cent	Perlegen	Not fixed: 0.04 – 0.3 in Hapmap Pops.	Hapmap Jan. 2007 dbSNP build 125	0.3 cM/ Mb	No genes in the region
rs1728369 -	G/ T - K	T	Chr 16: 84943714	94bp Cent	Perlegen	Not fixed: 0.1 – 0.4 in Hapmap Pops	Hapmap Jan. 2007 dbSNP build 125	Less than 5 cM/ Mb	No genes in the region

Table 3-6 The 4 SNPs used in the SNPSTR system and their locations, minor allele frequency (MAF) information, and their local context.

The ancestral state of each SNP was determined using chimp sequence data from the Ensembl database. The SNP location with regard to the distance from the STR is shown as Cent (Centromeric) or Tlm (Telomeric). The genome build information comes from dbSNP build 125.

3.4 SNPSTR Systems

The SNPs were verified in available DNA samples by PCR amplification followed by sequencing (See Materials and Methods, Chapter 2). For the SNPSTR PCR, locus-specific primers were designed, as well as additional primers for the flanking regions downstream of the STR. A table of primers used in this study for SNPSTR amplification alone can be seen below in Table 3-7.

Primer	Sequence	Label at 5' end
D5S818F	agccacagtttacaacatttgtatct	None
D5R-A	ggtcctctttggtatccttat	HEX
D5R-G	ggtcctctttggtatccttac	FAM
D16R	gatacatgcttacagatgcacacacaaac	None
D16FA	agcactgaaagaagaatcca	HEX
D16FC	agcactgaaagaagaatccc	FAM
D3PPLEXF	actgcagtccaatctgggt	None
D3FC	actcagcttcagcccataccc	HEX
D3FG	actcagcttcagcccataccg	FAM
CSF1PO Green F	accctgtgtctcagttttccta	None
CSF1PO SNPR A	taaatgtctcagagcctgctca	FAM
CSF1PO SNPR C	taaatgtctcagagcctgctcc	HEX

Table 3-7. Primers used in SNPSTR amplification with their respective labels to enable detection in subsequent capillary electrophoresis analysis.

Following verification of the SNPs and the identification of SNP heterozygotes and homozygotes, the PowerPlex16® kit was used to type the STR alleles and thus to confirm STR homozygotes and heterozygotes. The DNAs used for this pilot study came from anonymous, healthy volunteers in the lab. Data were analysed using SeqA sequence analysis software from Applied Biosystems.

An example of an electropherogram result from the PowerPlex16® kit can be seen in Figure 3-8 below.

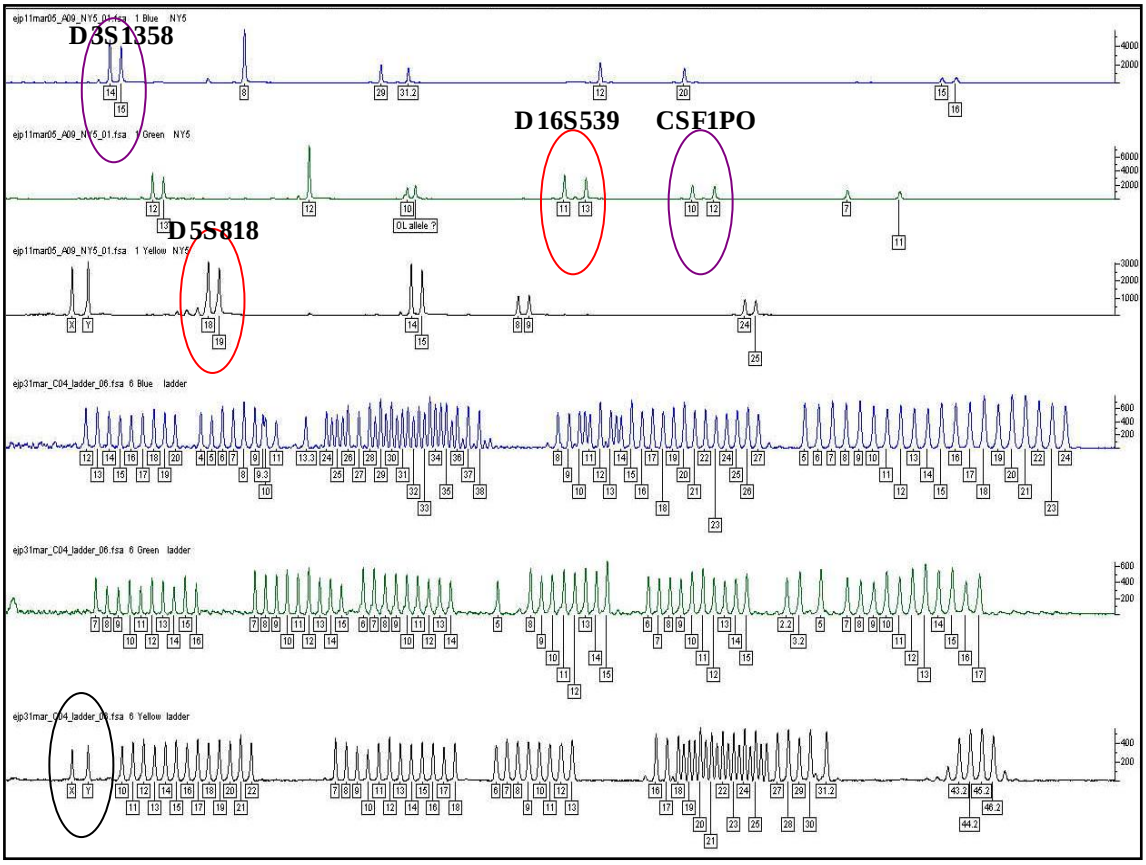


Figure 3-8. PowerPlex16® STR Profile
The D5S818 and D16S539 STRs have been circled in red, and the D3S1358 and CSF1PO STRs in purple. D5S818 in this sample has heterozygous STR alleles of 12 and 13 repeats. D16S539 has heterozygous STR alleles of 11 and 13 repeats. The XY Amelogenin test (circled in black) indicates that both X and Y are present and therefore this sample comes from a male. The allelic ladders (provided by the manufacturer) are indicated on the lower three sections of the figure. Source: Emma Parkin

The four SNPSTRs were amplified in two duplex reactions, one containing the forensic STR D5S818 and SNP rs25768, and which will be abbreviated as D5 from now on, with D16S539 and SNP rs1728369, and which will be abbreviated as D16. The other duplex reaction contained the forensic STR D3S1358 and SNP rs17077990 (D3) with CSF1PO and SNP rs2116791 (CSF).

3.4.1 D5 SNPSTR

This SNPSTR, as mentioned above, consists of the D5S818 STR and the SNP rs25768, both located on Chromosome 5, with the SNP only 13 base pairs centromeric of the STR. Figure 3-9 below shows a schematic of the D5 SNPSTR.

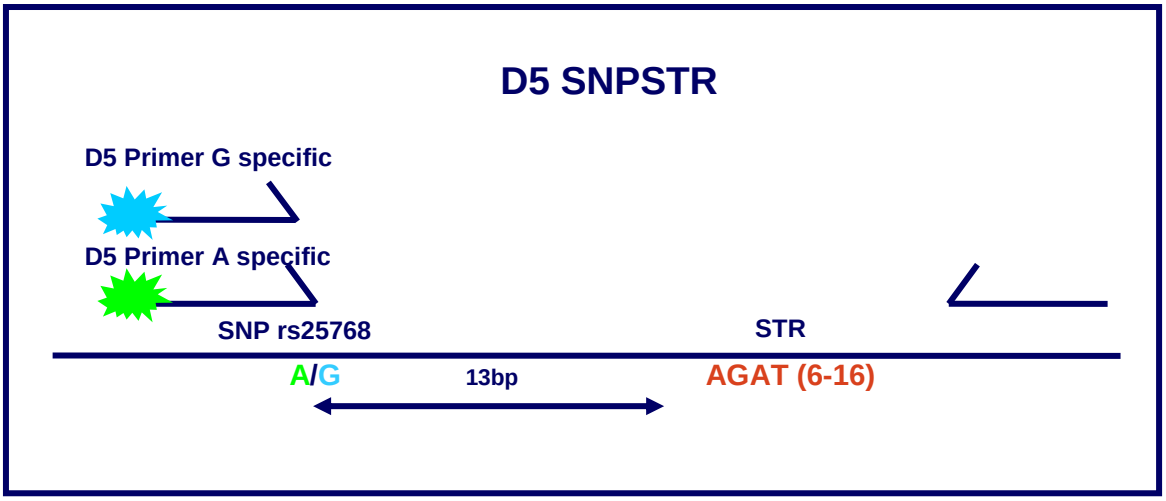


Figure 3-9. The D5S818 SNPSTR showing the STR repeat, the SNP accession number and the SNP-specific labelled primers.

The figure below shows the DNA sequence surrounding the D5 STR and the analysed SNP.

80761 ctgagacatg catatgcttt taaagcttct aattaaagtg gtgtcccaga taatctgtac

80821 taataaaagt atattttaat agcaagtatg tgacaagggt gattttcctc tttggtatcc

80881 ttaRgtaata ttttgaagat agatagatag atagatagat agatagatag atagatagat

80941 agaggtataa ataaggatac agataaagat acaaatgttg taaactgtgg ctatgattgg

81001 aatcacttgg ctaaaaagca ctaaagcatt cctctgagag agacaattac ttttttgctt

Figure 3-10. DNA sequence showing the location of the D5S818 STR which is highlighted in red, and the primers used for the SNPSTR amplification are highlighted in purple, with arrows indicating primer direction (GENBANK AC008512).

Initially the D5 SNPSTR locus was amplified by PCR as per Chapter 2 using the primers listed in Table 3-7. SNPSTR amplification was verified by means of gel electrophoresis as the sizes of the SNP and STR product varied and were able to be visualized on an agarose gel. An example of this can be seen in Figure 3-11.

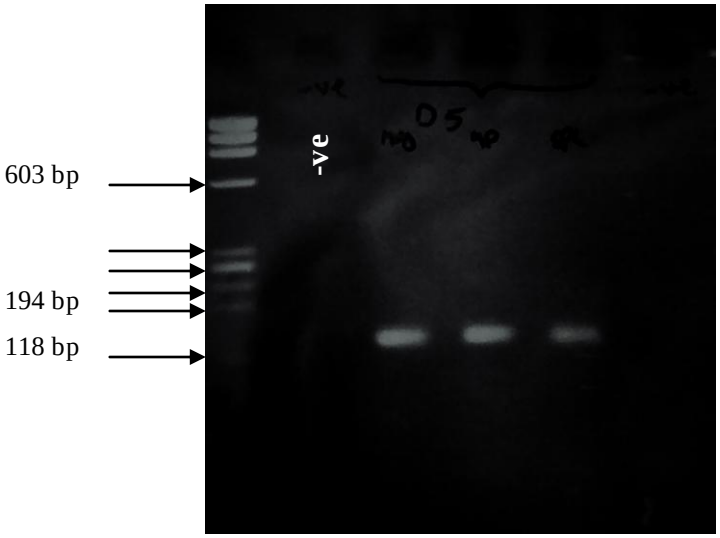


Figure 3-11. Amplification of the D5 SNPSTR locus using DNA from laboratory members and also a negative control. The Ø size marker used shows sizes of: 1353, 1-78, 872,603, 316,281-271,234,194,118 and 72 base pairs. The D5 SNPSTR products are approximately 134bp in size which is as expected from the genomic sequence data.

The SNPSTR PCR product was then sequenced using the ABI3100 as described in Chapter 2. This was carried out as a final check to ensure that the correct region had been amplified and also that the SNP-specific primers were indeed selecting the correct SNPs. An example of the resulting trace after analysis using SeqA, can be seen in Figure 3-12.

Figure 3-12. **SeqA trace of D5 SNPSTR showing both the SNP (highlighted in blue) and the STR with the repeat numbers ('reps') indicated.** Trace A shows a homozygote for the T allele of the SNP and 12 repeats of the STR. Trace B shows a homozygote for the C allele of the SNP with 12 repeats of the STR and Trace C shows a heterozygote with both T and C alleles of the SNP and 13 STR repeats. However, it is not possible to use this method to check for STR length as only the shortest repeat length is clearly discernible.

3.4.2 D16 SNPSTR

This SNPSTR (seen in Figure 3-13) consists of the D16S539 STR and the SNP rs1728369, both located on Chromosome 16, with the SNP located 95 base pairs centromeric to the STR.

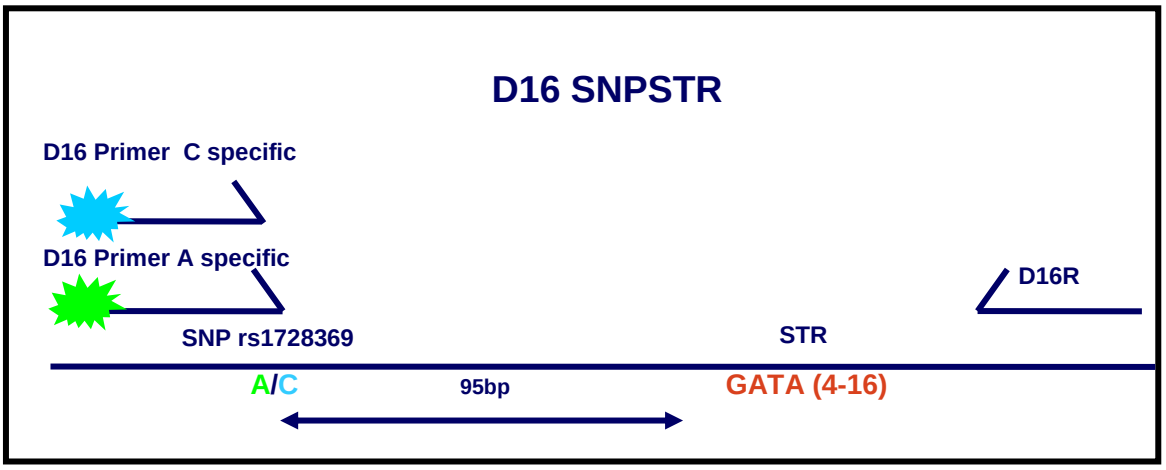


Figure 3-13. **Schematic of the D16 SNPSTR showing the relative locations of the STR, SNP and primers.**

The figure below shows the DNA sequence where the D16 STR and the SNP are located.

100741 gacagccctg cacccaggag ctgggggtc taagagcttg taaaaagtgt acaagtgcca

100801 gatgctcgtt gtgcacaaat ctaaatgcag aaaagcactg aaagaagaat ccKaaaacc

100861 acagttccca tttttatatg ggagcaaaca aaggcagatc ccaagctctt cctcttccct

100921 agatcaatac agacagacag acaggtggat agatagatag atagatagat agatagatag

100981 atagatagat atcattgaaa gacaaaacag agatggatga tagatacatg cttacagatg

101041 cacacacaaa cgctaaatgg tataaaaatg gaatcactct gtaggctgtt ttaccaccta

Figure 3-14. **DNA** sequence **showing the location of the D16 STR** which is highlighted in red, and the primers used for SNPSTR amplification are highlighted in purple, with arrows indicating primer direction (GENBANK AC092327).

SNPSTR amplification was verified by means of gel electrophoresis as the sizes of the SNP and STR products varied and were able to be visualized on an agarose gel. An example of this can be seen in Figure 3-15.

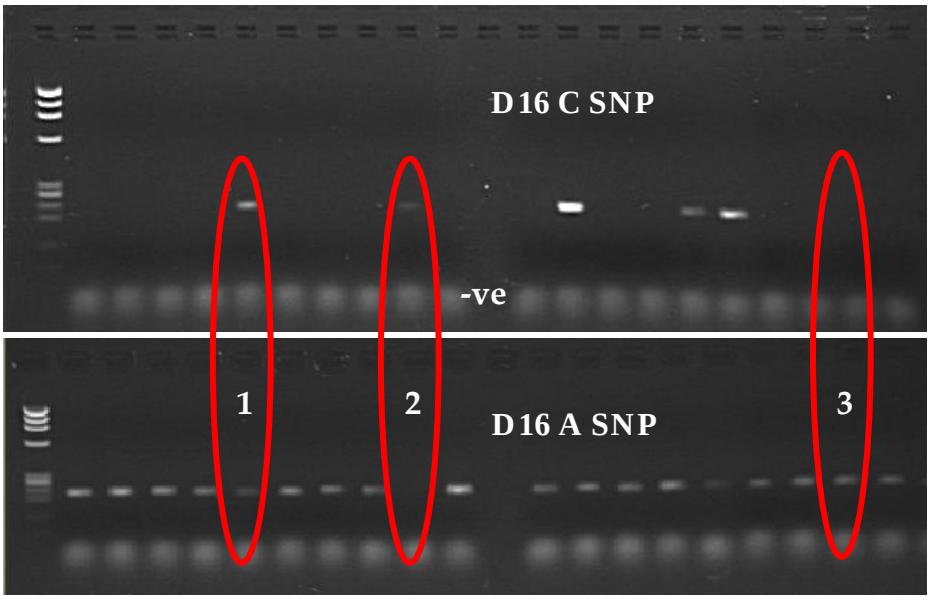


Figure 3-15. **Fragments resulting from the D16 SNPSTR PCRs using DNA from Himalayan samples held in the lab., and also a negative control.** The size marker used on the right hand side is the Ø marker with size fragments from top to bottom of: 1353, 1078, 872, 603, 316, 281-271, 234, 194, 118 and 72 base pairs. The D16 SNPSTR products are approximately 240bp in size which is as expected from genomic sequence data. Circled in red; 1 shows a heterozygote for the SNP A- and C-alleles, 2 shows a homozygote for the C-allele and 3 shows a homozygote for the A-allele.

Following the successful amplification of the D16 SNPSTR, the samples were then amplified using the labelled SNP-specific primers and then analysed on the ABI3100 and GeneScan Analysis Software from ABI in order to ascertain the SNP and STR lengths for each sample

An example of a successful electropherogram can be seen in Figure 3-16.

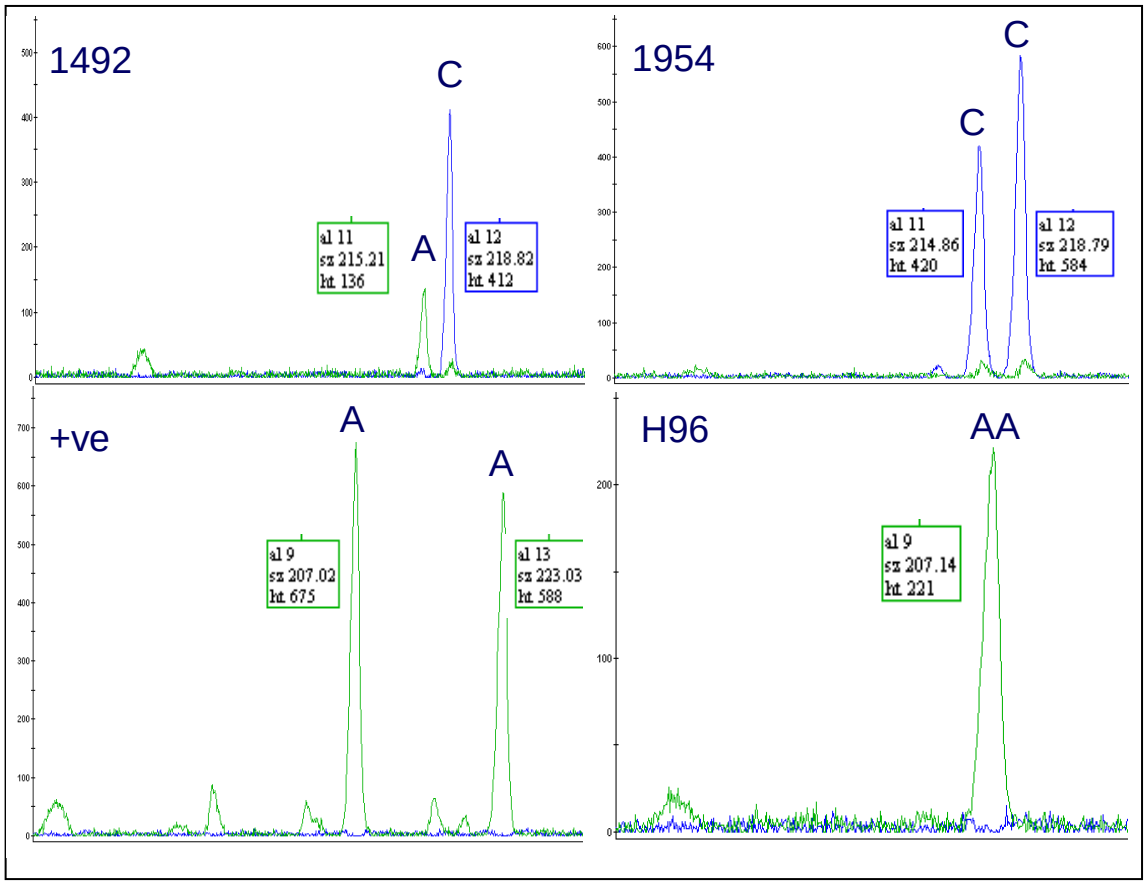


Figure 3-16. Electropherogram results of the D16 SNPSTR using labelled primers on 4 different samples. Sample 1492, 1954 and H96 are Himalayan DNA samples and the +ve control comes from a lab member. 1492 shows two peaks; a SNP A-allele associated with an STR of 11 repeats, and a SNP C-allele associated with an STR of 12 repeats. Sample 1954 shows two C-alleles, one associated with an STR of 11 repeats and the other with an STR of 12 repeats. H96 is homozygous for the SNP A-allele, and also homozygous for 9 STR repeats. The +ve control is homozygous for the SNP A-allele, but carries STR alleles of 9 and 13 repeats.

3.4.3 D3 SNPSTR

This SNPSTR consists of the D3S1358 STR and SNP rs17077990 , both located on Chromosome 3. The SNP is located 271 base pairs telomeric from the STR. Figure 3-17 shows a schematic of the D3 SNPSTR.

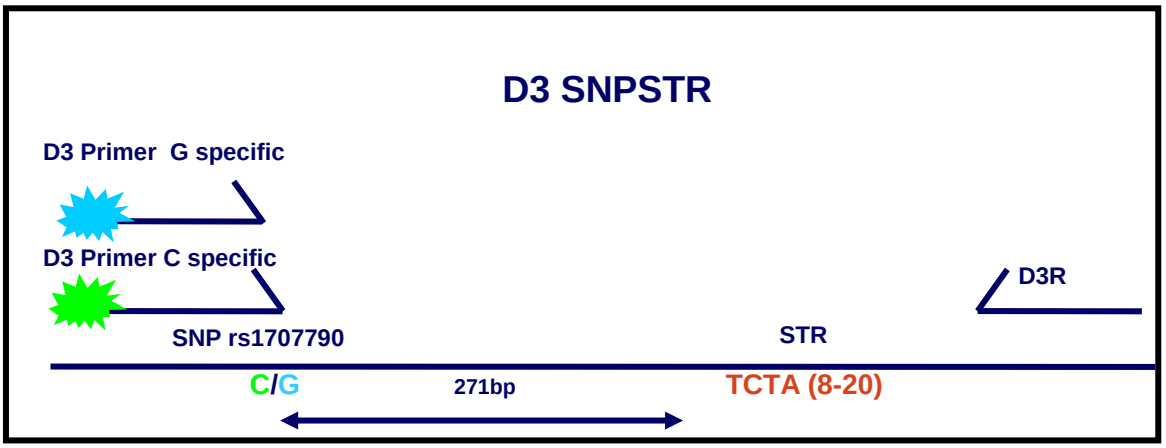


Figure 3-17. **Schematic of the D3 SNPSTR showing the relative locations of the STR, SNP and primers.**

Figure 3-18 shows the NCBI sequence of the D3 STR and the SNP.

45556921 gaattgactc cctctgtcac aaactcagct tcagcccata ccStgagcca tagacctatc

45556981 cctctaatgc attgtactag tctcagggct aataacaagg gagaggtgtc aaagggccag

45557041 ttccacctcc accaccagtg gaaaagctat tcccaggtga ggactgcagc tgccagggca

45557101 ctgctccaga atgggcatgc tggccatatt cacttgccca cttctgccca gggatctatt

45557161 tttctgtggt gtgtattccc tgtgcctttg ggggcatctc ttatactcat gaaatcaaca

45557221 gaggcttgca tgtatctatc tgtctatcta tctatctatc tatctatcta tctatctatc

45557281 tatctatcta tctatctatg agacagggtc ttgctctgtc acccagattg gactgcagtg

Figure 3-18. **DNA sequence showing the location of the D3 STR** which is highlighted in red, and the primers used for SNPSTR amplification are highlighted in purple, with arrows indicating primer direction (GENBANK AC099539).

This SNPSTR was successfully amplified by PCR and the results were then confirmed by sequencing. An example of the D3 SNPSTR PCR can be seen in Figure 3-19.

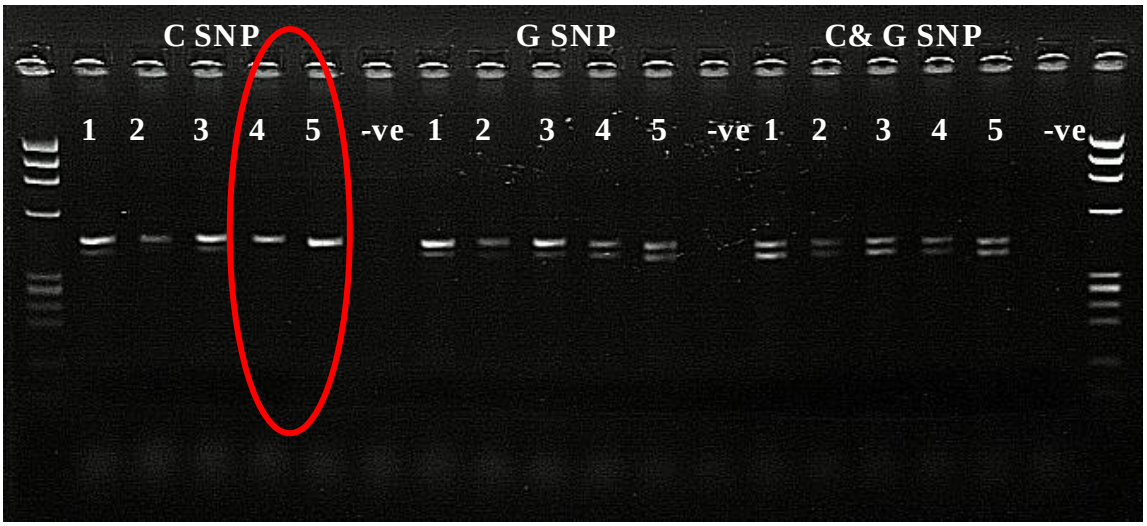


Figure 3-19. **Fragments resulting from the D3 SNPSTR PCRs using DNA from laboratory members, and also a negative control**. The size marker used on the right hand size is the Ø marker with size fragments from top to bottom of: 1353, 1078, 872,603, 316,281-271,234,194,118 and 72 base pairs. The D3 SNPSTR products are ± 375bp in size which is as expected. Circled in red; samples 4 and 5 show homozygotes for the C SNP, the rest are all heterozygotes for C and G SNPs.

These results were confirmed by sequencing and two examples can be seen in Figure 3-20.

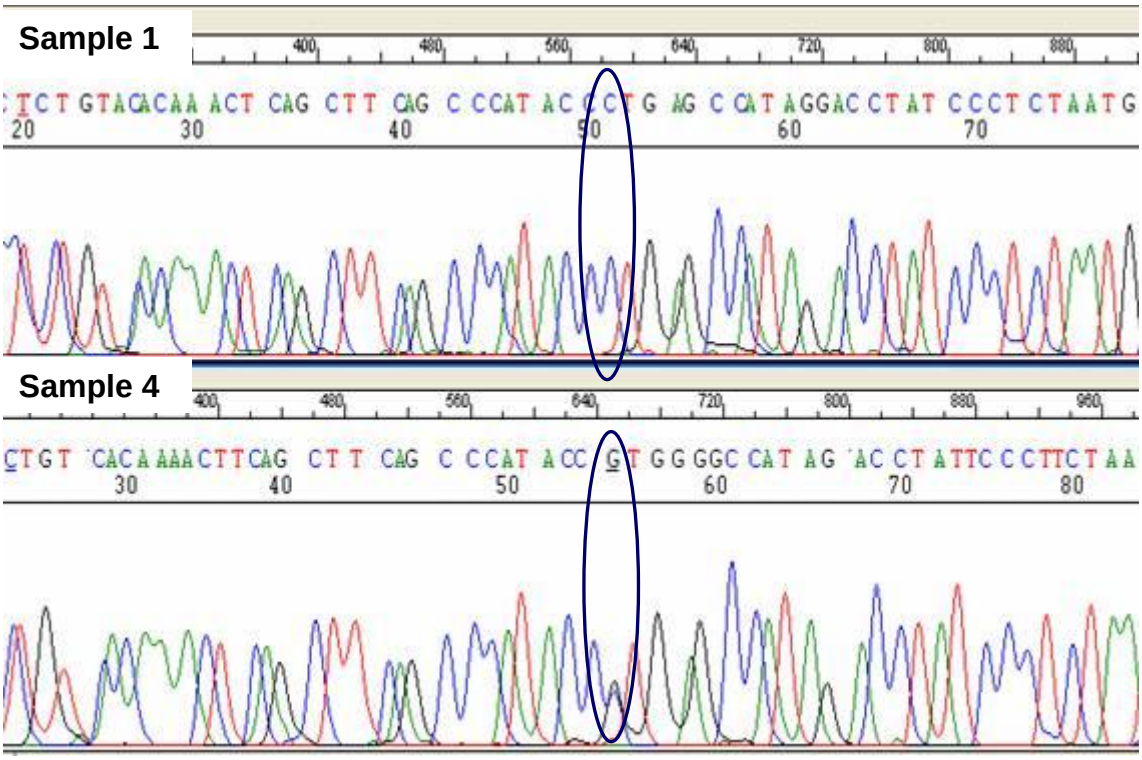


Figure 3-20. **Sequencing results of the D3 SNP.** Sample 1 shows a homozygote (CC) for the SNP and sample 4 shows a heterozygote (GC) for the SNP.

Following in from this, the final SNPSTR was ascertained.

3.4.4 CSF1PO SNPSTR

This SNPSTR consists of the CSF1PO STR and SNP rs2116791, both located on Chromosome 5, with the SNP located 299 base pairs centromeric to the STR. Figure 3-21 below shows a schematic of the CSF1PO SNPSTR.

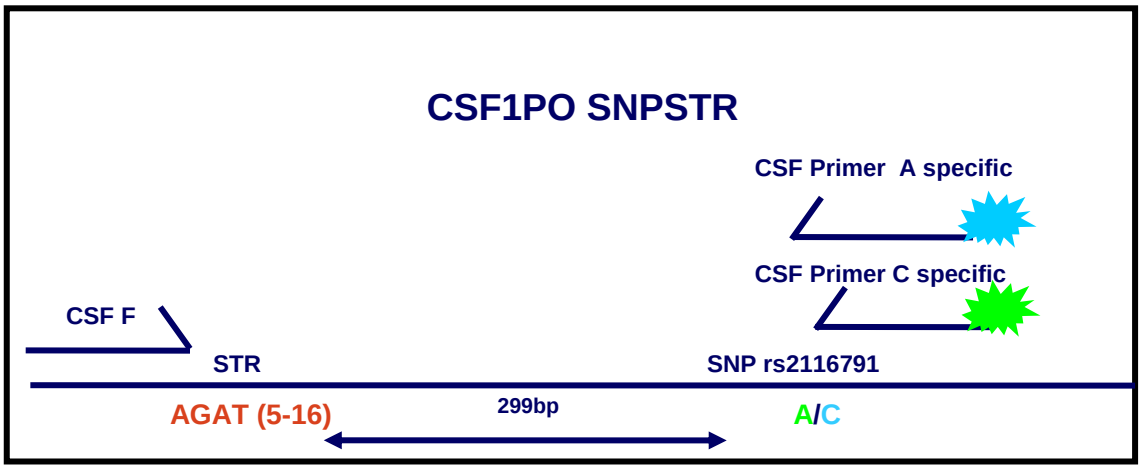


Figure 3-21. **Schematic of the CSF1PO SNPSTR showing the relative locations of the STR, SNP and primers.**

Figure 3-22 below shows the NCBI sequence where the CSF1PO STR and the SNP are located.

11881 accaccctgt gtctcagttt tcctacctgt aaaatgaaga tattaacagt aactgccttc

11941 atagatagaa gatagataga ttagatagat agatagatag atagatagat agatagatag

12001 atagatagat aggaagtact tagaacaggg tctgacacag gaaatgctgt ccaagtgtgc

12061 accaggagat agtatctgag aaggctcagt ctggcaccat gtgggttggg tgggaacctg

12121 gaggctggag aatgggctga agatggccag tggtgtgtgg aagagtctga gatgcaggga

12181 tgaggaagag aaaggagata aggatgacct ccaggtctct ggctatggtg attgggtgca

12241 ggcagtggca gtcactggac tcagaccctg aagcaaggca gcagctcatc ggagtgKgag

12301 caggctctga gacatttagg tctggccgtg cctcatgtgt tgaatgttat gggagatgga

Figure 3-22. **DNA sequence showing the location of the CSF1PO STR** which is highlighted in red, and the primers used for SNPSTR amplification are highlighted in purple, with arrows indicating primer direction (GENBANK U63963).

SNPSTR amplification was verified by means of gel electrophoresis as the sizes of the SNP and STR product varied and were able to be visualized on an agarose gel. An example of this can be seen in Figure 3-23. DNA samples came from the HapMap DNA collection which has already been typed for SNPs and

this therefore provided a positive control for my SNPSTR PCR products. The results from the electrophoresis gel mirrored those in the HapMap database. For this reason, it was not necessary to sequence these PCR products.

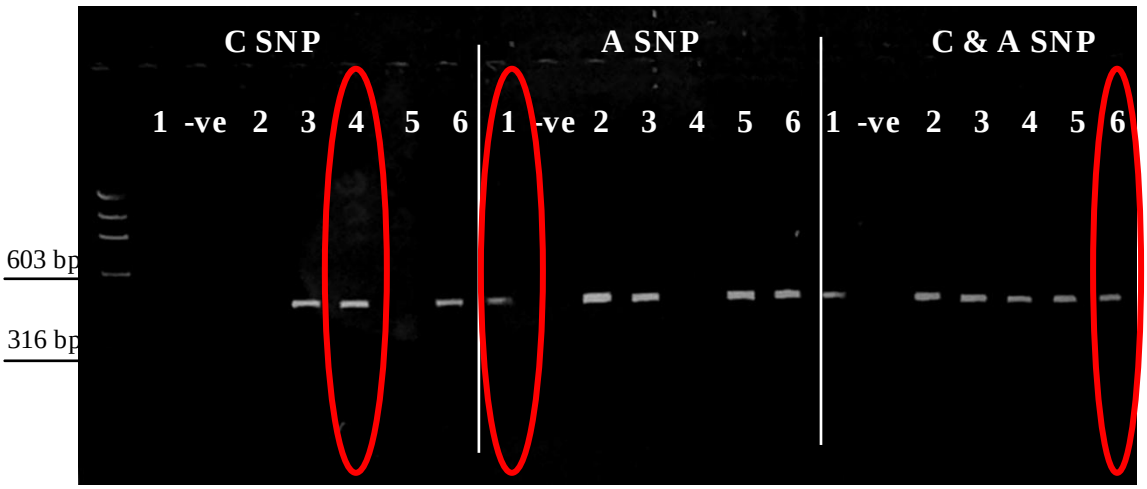


Figure 3-23. **Fragments resulting from the CSF1PO SNPSTR PCRs using DNA from HapMap samples held in the lab., and also a negative control**. The size marker used on the right hand size is the Ø marker with size fragments from top to bottom of: 1353, 1078, 872, 603, 316, 281-271, 234, 194, 118 and 72 base pairs. The CSF1PO SNPSTR products are ± 460bp in size which is as expected. Circled in red; sample No. 1 shows a homozygote for the A SNP, sample No. 4 shows a homozygote for the C SNP and sample No. 6 shows a heterozygote for the C and A SNPs.

This marks the end of the SNPSTR selection chapter. In the following chapters the analysis and typing of these four SNPSTRs in different global populations is discussed and it will be seen how SNPSTRs can be used to estimate TMRCA of the derived SNP as well as outline how the information provided by SNPSTRs ties in with known population migration histories.

4 Analysis of SNPSTRs in HapMap, CEPH-HGDP, and Cornish population samples

4.1 Introduction

This chapter focuses on the typing of the four SNPSTR systems described in Chapter 3 on two global population samples, as well as on a smaller sample of DNAs from Cornwall, UK.

The SNPSTRs were typed on the HapMap DNAs (HapMap 2003), the CEPH-HGDP DNAs (Cann et al. 2002) – samples which have been widely used previously, so that much data is already available on them, which makes them ideal for trialling new marker systems such as the SNPSTRs based on the forensic STRs. These samples are described more fully below. In addition, typing of a set of DNAs from Cornwall, UK (Richards et al. 2000) is also described here.

The typing was carried out on these populations in order to ascertain whether the SNPSTRs are able to provide data on migration patterns of populations and to see if these patterns are consistent with the current ideas on their migration histories. In addition to this, inferences were made about the TMRCAs of the derived SNPs.

For the CEPH-HGDP SNPSTRs typed here, the STR typing results were continually checked against those achieved by Peter de Knijff (2006 - personal

communication) when all the CEPH-HGDP DNAs were typed for not only the STRs used within this project, but were also typed for Y chromosome STRs. This independent source of STR typing, based upon the PowerPlex 16 typing kit, enabled a continual validation of the method used and of the results obtained. STR length results achieved by means of SNPSTR typing were compared with those from Peter the Knijff and if the STR repeat numbers did not match those of Peter de Knijff's, they were re-analysed. If the result achieved for SNPSTR typing was consistently different from Peter's but contained the same different STR length, then my result was taken to be the correct version. The difference in STR allele calling between my results and those obtained by Peter de Knijff was low, at between 1.83% and 2.75% for the STRs.

In order to assess the quality of typing, and to investigate evolutionary factors that might affect particular populations, Hardy-Weinberg equilibrium testing was carried out on all of the SNPSTRs.

4.1.1 The Hardy-Weinberg Equilibrium

The Hardy-Weinberg theorem explains how allele frequencies in one generation can be used to calculate the genotype proportions in the next generation. For example, in diploid organisms (such as humans), two alleles A and a at the same locus, with frequencies of *p* and *q* respectively can combine to create three genotypes: AA, Aa and aa. If the frequency of these two alleles is

know in a population, then it is possible to predict the proportion of the genotypes in the succeeding generations by combining gametes at random. This will produce genotypes in the next generation in the following proportions: AA = p^2, Aa = $2pq$ and aa = q^2.

The Hardy–Weinberg theorem states that both allele and genotype frequencies in a population of diploid organisms should remain constant, i.e. that they are in equilibrium from one generation to the next, provided that: a) selection is not acting on the population, b) there is no mutation , c) there is random mating, and d) there is no migration.

If populations are non-randomly mating, have undergone mutation and/ or selection, or if there is a limited population size, or random genetic drift and gene flow, then this might change the allele frequencies. The Hardy-Weinberg equation can also be used to pick up any unusual population history characteristics or errors in genotyping, for example, if there has been mistyping where heterozygotes have been wrongly called as homozygotes (or vice versa).

- **Random Mating**

  The Hardy-Weinberg principle states the population will have the given genotypic frequencies after a single generation of random mating within the population. When violations of this occur, the population will not have Hardy–Weinberg proportions. An example of possible violations is inbreeding, which causes an increase in homozygosity for all genes.

- **Assortative Mating**

  This causes an increase in homozygosity only for those genes involved in the trait that is assortatively mated (and genes in linkage disequilibrium with them).

- **Small Population Size**

  This causes a random change in genotypic frequencies, particularly if the population is very small. This is due to a sampling effect, and is called genetic drift.

The remaining assumptions affect the allele frequencies, but do not, in themselves, affect random mating. If a population violates one of these, the population will continue to have Hardy–Weinberg proportions each generation, but the allele frequencies will change.

- **Selection**

  Selection causes allele frequencies to change, often quite rapidly. While directional selection eventually leads to the loss of all alleles except the favoured one, some forms of selection, such as balancing selection, lead to equilibrium without loss of alleles.

- **Mutation**

  Mutation has a very subtle effect on allele frequencies. Recurrent mutation will maintain alleles in the population, even if there is strong selection against them.

- **Migration**

  Migration genetically links two or more populations together. In general, allele frequencies will become more homogeneous among the populations. Some models of migration inherently include non-random mating and so for those models, the Hardy–Weinberg proportions will normally not be valid.

If genotype proportions are in Hardy-Weinberg equilibrium, then it is assumed that selection is not acting upon any of the loci under observation and that they are in effect neutral loci.

However, if the genotype proportions are not in Hardy-Weinberg equilibrium, then one of the conclusions is that evolution is occurring and that one or more of the conditions of the Hardy-Weinberg principle have not been met.

4.2 Typing of SNPSTRs in population samples

The HapMap DNAs consist of 270 DNA samples from 4 populations as described in more detail below. Data were obtained for all 4 SNPSTRs, for all 4 HapMap populations and results were checked against the known HapMap pedigrees in order to verify that the SNPSTRs were called correctly. Some results could not be verified as not all SNPSTR typing was successful, which resulted in some cases of missing data. Where this occurred, the remaining results achieved for an individual were assumed to be correct and were

included in all of the final analyses. This decision was made because of the previous checks on SNPSTR typing carried out by comparing the CEPH-HGDP results with the known STR alleles which had previously been typed by Peter de Knijff's group (Department of Human Genetics, Leiden University Medical Centre), as described in paragraph 4.1 above.

The CEPH-HGDP panel consists of 1064 DNAs from groups of populations from all over the world, and is described in more detail below. There are known duplicates in the panel (Rosenberg 2006), and these were excluded from the final data, although there were typed and used as a typing check/ verification in order to ensure that the typing methods and allele calling using the GeneMapper software was consistent.

The CEPH-HGDP DNAs from each of 25/ 51 population samples represent unrelated individuals (no closer than first cousins); the remaining 26 populations contain at least one of 96 pairs of first and/ or second degree relatives (Cann et al. 2002).

The 67 samples from Cornwall are from male farmers, living in the county of Cornwall. The DNA was collected at cattle markets on two separate occasions, and there are no known relatives in the samples. Unfortunately there is very little other information available on these samples which could be used to verify the SNPSTR typing, however, as these samples were typed after the HapMap and CEPH-HGDP and the method had been proven to be reliable, all

results achieved were assumed to be correct and were included in the final analysis. These DNAs were analysed in order to supplement the number of samples from western Europe, as there are none from western Europe in the CEPH-HGDP panel (here the "Europeans" are all US citizens with self-proclaimed European ancestry), and only 16 samples from Northern Europe in the HapMap DNAs from the Orcadians.

The typing of the SNPSTRs on African Caribbean (Shriver et al. 2003) and Danish and Greenland Inuit DNAs (Bosch et al. 2003) will be discussed in Chapter 5 where these samples formed part of a study to use the SNPSTRs to examine population structure in admixed populations.

4.2.1 The International HapMap Project sample set of DNAs

The HapMap project (2003) was established in 2002 with the aim to determine the common patterns of DNA sequence variation (or haplotypes) in the human genome. The aim of this was to facilitate the discovery of sequence variants that affected common disease. The project was set up to create a public database of haplotypes which would provide information to guide genetic studies of clinically important phenotypes and find genes associated with human disease. The International HapMap Project (2005) is a partnership of scientists and funding agencies from around the world and has been possible due to the sequencing of the human genome and the availability of databases containing common SNPs and knowledge of LD. It is estimated that there is one SNP per kilobase and that there are between 9 – 10 million common SNPs (with a minor allele frequency ≥ 0.05) in the assembled human genome (Frazer

et al. 2007) with varying frequencies of common and rare alleles. Although most of the common haplotypes occur in all human populations, their frequencies differ among populations (HapMap 2005). For this reason, data from several populations were needed for the HapMap project in order for it to be able to ascertain the tag SNPs needed to give adequate coverage of those associated with disease. Tag SNPs are representative SNPs in a region of the genome where there is high LD (i.e. non-random association of alleles at two or more loci). By using tag SNPs, it is possible to assay genetic variation without having to type every SNP in a chromosomal region. Tag SNPs are particularly useful in whole-genome SNP association studies such as the Wellcome Trust Case Control Consortium study (Consortium 2007). Pilot studies carried out before the start of the project found that there were sufficient differences in the haplotype frequencies among population samples from Nigeria (Yoruba), Japan, China and the U.S. (residents with ancestry from northern and western Europe, collected in 1980 by the Centre d'Etude du Polymorphisme Humain – CEPH – and extensively used in genetic linkage studies).

The DNAs from the HapMap project are available from the Coriell Cell Repositories (New Jersey, USA), which provide the tools for research scientists by establishing, verifying, maintaining, and distributing cell lines and DNA derived from cell lines. The Coriell Cell Repositories are part of the Coriell Institute for Medical Research which was founded in 1953 and is a non-profit making, biomedical research institution.

The DNA samples for the HapMap project come from 270 people, broken down into populations as in the table below:

Country	Population	No. Individuals	Trios
Nigeria	Yoruba of Ibadan	90	Yes
Japan	Tokyo	45	No
China	Han from Beijing	45	No
Northern and western European	CEPH	90	Yes

Table 4-1. Summary of DNA samples in HapMap project.
All the samples used in the HapMap project were obtained with protocols approved by the appropriate ethics committees.

A full list of the DNA samples in the HapMap project and which were typed for the SNPSTRs are shown in Appendix 3.

The HapMap project is being continually updated with Phase II containing the typing of more than 2 million additional SNPs by Perlegen Sciences and an additional 500,000 by Affymetrix. Phase II also revealed new aspects of LD, and showed that 10-30% of pairs of individuals within a population shared at least one region of extended genetic identity which arose from a recent ancestor (Frazer et al. 2007). Phase III which was completed October 2008 saw the inclusion of Phase I and II data as well as corrections to previously mis-called SNPs.

4.2.2 The CEPH-HGDP DNAs

The Centre d'Etude du Polymorphisme Humain (CEPH), in collaboration with the Human Genome Diversity Project (HGDP) has made a collection of cultured

lymphoblastoid cell lines (LCLs) available from different laboratories, that represents individuals in different world populations (Cann et al. 2002). The collection was formed in order to supply an unlimited quantity of a standard set of DNAs for studies of sequence diversity and to aid in the study of the history of human populations. The collection is held at the Fondation Jean Dausset in Paris. The CEPH-HGDP panel of DNAs are produced from 1064 LCLs and represent 1050 individuals from 51 world-wide populations. Unfortunately there are no samples from either India or Australia, which would have provided information about the rich genetic diversity present on the Indian continent, as well as information about the founders of Australia. India has a rich population diversity (Bamshad et al. 2001) and Australia houses the founders of some of the earliest peoples of the world (Jones 1989). Each of the original blood samples used in the CEPH-HGDP was freely donated under conditions of informed consent. The information provided with the DNAs includes the geographic and population origin and also the sex. The CEPH-HGDP also contains more male samples than female samples, which is of benefit to those studying the Y chromosome.

Since the CEPH-HGDP panel was collated (Cann et al. 2002), more than 650,000 SNPS and 853 microsatellites as wel as 10 known indel loci and copy number varations (CNVs) have been typed on the panel (http:/ / www.cephb.fr/ en/ hgdp/ diversity.php). There are also 13 known duplicate samples within the collection, as well as pairs of close relatives, and some atypical samples were also identified (Rosenberg 2006).

Although it can not be denied that the CEPH-HGDP panel is a very useful resource for population genetics studies, there are some problems associated with using these samples for this type of work. The main issue about using these samples is that they are mostly from population isolates and therefore provide a somewhat biased sample which is not truly representative of a spread of human populations.

Below is a summary table of the DNA samples from the CEPH-HGDP. A table containing the CEPH-HGDP unique identification numbers can be seen with the results later in this section.

	Geographic Origin	Population	No. LCLs	♂	♀	Dupli-cates
SUB-SAHARAN AFRICA - 127 LCLs	Central African Republic	Biaka Pygmy	36	33	3	3
	Dem. Republic of Congo	Mbuti Pygmy	15	13	2	
	Senegal	Mandenka	24	16	8	
	Nigeria	Yoruba	25	13	12	
	Namibia	San	7	7	0	
	Kenya (All Bantu Speakers)	Bantu NE (12)	12	11	1	
	S. Africa Bantu S.E.	Bantu S.E. Pedi (1)	1	1	0	
	S. Africa Bantu S.E.	Bantu S.E. Sotho (1)	1	1	0	
	S. Africa Bantu S.E.	Bantu S.E. Tswana (2)	2	2	0	
	S. Africa Bantu S.E.	Bantu S.E. Zulu (1)	1	1	0	
	S. Africa Bantu S.W.	Bantu S.W. Herero (2)	2	2	0	
	S. Africa Bantu S.W.	Bantu S.W. Ovambo (1)	1	1	0	
N. AFR. 30 LCLs	Algeria (Mzab)	Mozabite	30	20	10	
MIDDLE EAST - 148 LCLs	Israel (Negev)	Bedouin	49	28	21	1
	Israel (Carmel)	Druze	48	14	34	1
	Israel (Central)	Palestinian	51	17	34	
ASIA - 451 LCLs	Pakistan	Balochi	25	25	0	
	Pakistan	Brahui	25	25	0	
	Pakistan	Hazara	25	25	0	1
	Pakistan	Makrani	25	20	5	
	Pakistan	Sindhi	25	21	4	
	Pakistan	Pathan	25	20	5	
	Pakistan	Kalash	25	20	5	
	Pakistan	Burusho	25	20	5	
	China	Han	45	24	21	1
	China	Yizu (Yi) (minority)	10	9	1	
	China	Miaozu (Miao) (minority)	10	7	3	
	China	Oroqen (minority)	10	7	3	
	China	Daur (minority)	10	7	3	
	China	Mongola (minority)	10	7	3	
	China	Hezhen (minority)	10	7	3	1
	China	Xibo (minority)	9	8	1	
	China	Uygur (minority)	10	8	2	
	China	Dai (minority)	10	7	3	
	China	Lahu (minority)	10	7	3	

	China	She (minority)	10	7	3	
	China	Tujia	10	9	1	
	China	Naxi (minority)	10	8	2	
	China	Tu (minority)	10	7	3	
	Siberia	Yakut	25	18	7	
	Japan	Japanese	31	23	8	1
	Cambodia	Cambodian	11	6	5	
OCEANIA - 39 LCLs	New Guinea	Papuan	17	13	4	
	Bougainville	NAN Melanesian	22	8	14	3
EUROPE - 161 LCLs	France	French (various regions)	29	12	17	
	France	Basque	24	16	8	
	Italy	Sardinian	28	16	12	
	Italy	from Bergamo	14	9	5	1
	Italy	Tuscan	8	6	2	
	Orkney Islands	Orcadian	16	7	9	
	Russia Caucasus	Adygei	17	7	10	
	Russia	Russian	25	16	9	
AMERICA - 108 LCLs	Mexico	Pima (relative pairs)	25	14	11	
	Mexico	Maya (relative pairs)	25	3	22	
	Colombia	Piapoco and Curripaco	13	5	8	
	Brazil	Karitiana (relative pairs)	24	10	14	
	Brazil	Surui (relative pairs)	21	11	10	
		TOTAL SAMPLES	**1064**			

Table 4-2. Summary table of the DNAs in the CEPH-HGDP panel. ♀-female, ♂-male. (Cann et al. 2002)

Rs1728369, which is the SNP that forms part of the D16 SNPSTR has been typed on the CEPH-HGDP, and the major allele frequencies for all of the CEPH-HGDP populations are available in the UCSC Genome Browser (http://genome.ucsc.edu/). The data for this comes from dbSNP build 130. The major allele frequencies for this SNP were also calculated as part of this project and the dbSNP allele frequencies were compared to the major allele

frequencies which were typed as part of the D16 SNPSTR. The results of this can be seen in later in this chapter.

4.3 Materials and Methods

4.3.1 Hardy-Weinberg

As noted previously, it was essential to carry out Hardy-Weinberg testing on all samples in order to assess the quality of typing, and to investigate evolutionary factors that might affect particular populations.

There are many different software packages available which will test for Hardy-Weinberg equilibrium within a sample. For this work, the software programme GenAlEx 6.2 (Peakall and Smouse 2006) was used. This calculated Hardy-Weinberg using the Chi-Square test as follows:

$$X^2 = \sum_{i=1}^{k} \frac{(O-E)^2}{E}$$

In this test, O_i is the observed number of individuals of the *i*-th genotype and E_i is the expected number. Degrees of freedom for the Chi-Square test (X^2) are calculated as DF = [Na(Na-1)]/ 2, where Na is the number of alleles at the locus. This is a goodness of fit test, but others are available, such as Arlequin (Schneider et al. 2000a), which uses a permutation test to check for Hardy-Weinberg equilibrium.

4.3.2 The HapMap DNAs

The HapMap DNA was supplied in three separate plates, each containing 5 duplicated samples and a buffer control as well as the sample DNA. The DNA concentration is 250 ng/ µl and there are 50 µg of DNA per well.

4.3.3 The CEPH-HGDP DNAs

The CEPH-HGDP DNAs were supplied in twelve separate plates, each containing one CEPH control sample and also space for an additional control sample. In practice, the concentration of the DNAs varied; however, the samples which we received from CEPH contained a notional 2µg of DNA per well, which was reported to be at 50ng/ µl.

4.3.4 The Cornish DNAs

The Cornish samples were pre-extracted DNA at unknown concentrations.

4.3.5 D5 and D16 Duplex PCR

The D5 - D16 SNPSTRs were co-amplified as a duplex PCR which used approximately 0.5ng/ µl of DNA per reaction. Each 10µl reaction also contained 1 x Amplitaq Gold® Buffer I (Applied Biosystems), 0.1U Amplitaq Gold® Taq polymerase (Applied Biosystems), 1.75mM $MgCl_2$, 300µM dNTPs, 5% (w/ v) glycerol, and 1µM of each primer as described in Chapter 3, section 3.4. The amplification conditions were as follows; 94ºC for 10 minutes, then 30 cycles of 90ºC for 1 minute, 58ºC for 1 minute and 65ºC for 2 minutes and a final step of

65ºC for 10 minutes. Some of the PCR products were visualized on a 2% (w/ v) agarose gel to verify successful amplification and ensure no contamination (see Chapter 3, sections 3.4.1 and 3.4.2). PCR products were diluted 1:500 using dH_2O (optimal dilution following serial dilution testing on the ABI3100).

4.3.6 D3 and CSF1PO Duplex PCR

The D3 - CSF duplex PCR used approximately 0.5ng/ µl of DNA per reaction. Each 10µl reaction also contained 1 x Amplitaq Gold® Buffer I (Applied Biosystems), 0.1U Amplitaq Gold® Taq polymerase (Applied Biosystems), 1.75mM $MgCl_2$, 300µM dNTPs, 5% (w/ v) glycerol, and 1µM of each primer as described in section 3.4. The amplification conditions were as follows; 94ºC for 10 minutes, then 30 cycles of 94ºC for 1 minute, 65.5ºC for 1 minute and 70ºC for 1 minute and a final step of 65ºC for 10 minutes. Some of the PCR products were visualized on a 2% (w/ v) agarose gel to verify successful amplification and ensure no contamination (see Chapter 3, sections 3.4.3 and 3.4.4). PCR products were diluted 1:400 using dH_2O (optimal dilution following serial dilution testing on the ABI3100).

4.3.7 ABI3100 analysis of D5 and D16 SNPSTRs

As the ABI3100 is designed to handle DNA sequences up to 800bp in length it is the ideal piece of equipment to use for typing the SNPSTRs, as they are only up to 500bp in length.

A size standard mix was required in order to size the SNPSTR fragments. The mix contained 0.035 µl of Rox400 size standard (Cambio Ltd) to 10µl of Hi-di-

formamide (Applied Biosystems). A 10µl aliquot of size standard mix was used in every well of a 96-well, non-skirted plate (ABgene). 2µl of the diluted PCR product was then added to each well. The plates were loaded on the ABI3100 and one programme condition was used for all SNPSTR typing. The ABI parameters for all SNPSTR typing remained the same and these are listed in the table below.

ABI Parameter	**SNPSTR Setting**
Run Temperature	60 ºC
Capillary fill volume	184 steps
Pre-run time	10 seconds
Pre-run voltage	15 kVolts
Injection time	22 seconds
Injection voltage	3 kVolts
Run voltage	15 kVolts
Data delay time	1500 seconds
Current tolerance	100 µAmps
Run current	100 µAmps
Voltage tolerance	0.6 kVolts
No. of steps	10 steps
Voltage step interval	60 seconds

Table 4-3. SNPSTR ABI3100 run conditions.

The SNPSTR program conditions on the ABI 3100 were as follows:

Dyeset D (Blue- 6FAM, Green – HEX, Yellow – NED, Red – ROX)

BioLIMS project - 3100 Project I

Analysis Module - GS400 HD

Run Module - Butler 22 Secs

The size standard was needed in order for the Genescan software to calibrate a size curve which was then used to define the sizes of unknown sample fragments. For the D5 and D16 SNPSTRs, the size fragments are in 50 – 400bp range.

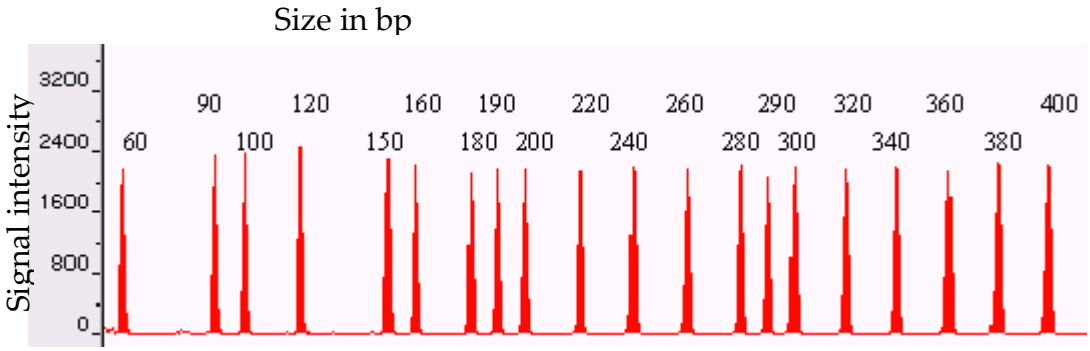


Figure 4-1. GeneScan Rox 400-HD size standard.

Output from the ABI3100 was analysed using GeneMapper software. V4 (ABI) and results were recorded.

4.3.8 ABI3100 analysis of D3 and CSF1PO SNPSTRs

A different size standard mix was required in order to size these SNPSTR fragments. For the D3 and CSF SNPSTRs, the size fragments are in 75 – 500bp range.

The mix contained 0.1 µl of Rox500 size standard (Cambio Ltd) to 10µl of Hi-di-formamide (Applied Biosystems). A 10µl aliquot of size standard mix was used in every well of a 96-well, non-skirted plate (ABgene). 2µl of the diluted PCR product was then added to each well. The plates were loaded on the ABI3100

and the ABI parameters from Table 4-3. **SNPSTR ABI3100 run conditions**) and run conditions as per section 4.3.7 were applied.

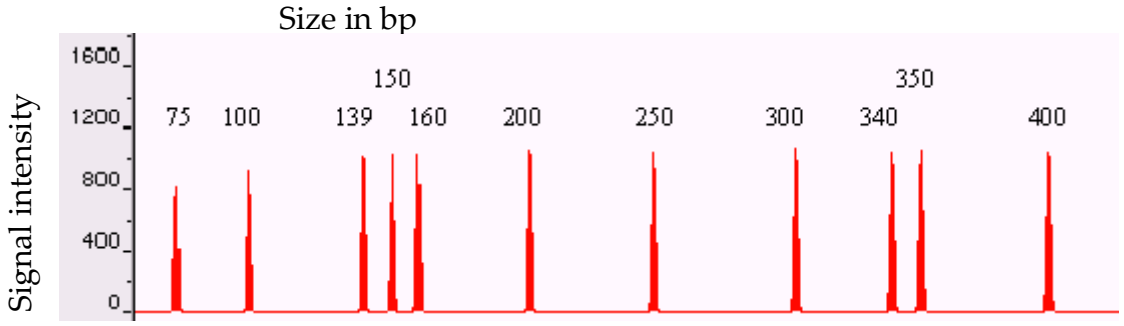


Figure 4-2 Cambio Rox 500 size standard..

Output from the ABI3100 was analysed using GeneMapper software (ABI) and results were recorded.

4.4 Results

4.4.1 Hardy-Weinberg Test

A full list of the results of applying this test to the distribution of the SNPs in the CEPH-HGDP, HapMap and Cornish DNA samples used is listed in Table 4-4.

Pop	Locus	DF	ChiSq	Prob	Signif
Cornish	rs2116791	1	0.193	0.660	ns
Cornish	rs1728369	1	0.367	0.545	ns
Cornish	rs17077990	1	0.129	0.719	ns
Cornish	rs25768	1	1.556	0.212	ns
HMYoruba	rs2116791	1	0.010	0.919	ns
HMYoruba	rs1728369	1	0.004	0.950	ns
HMYoruba	rs17077990	1	0.337	0.562	ns
HMYoruba	rs25768	1	2.515	0.113	ns
HMChina	rs2116791	1	0.428	0.513	ns
HMChina	rs1728369	1	0.301	0.584	ns
HMChina	rs17077990	1	0.077	0.781	ns
HMChina	rs25768	1	0.102	0.749	ns
HMJapan	rs2116791	1	0.666	0.414	ns
HMJapan	rs1728369	1	0.002	0.961	ns
HMJapan	rs17077990	1	0.090	0.765	ns
HMJapan	rs25768	1	1.035	0.309	ns
HM NW Europe	rs2116791	1	1.162	0.281	ns
HM NW Europe	rs1728369	1	0.615	0.433	ns
HM NW Europe	rs17077990	1	1.047	0.306	ns
HM NW Europe	rs25768	1	0.020	0.887	ns
N Africa	rs2116791	1	0.164	0.686	ns
N Africa	rs1728369	1	0.094	0.759	ns
N Africa	rs17077990	1	1.875	0.171	ns
N Africa	rs25768	1	0.833	0.361	ns
Sub Saharan Africa	rs2116791	1	3.451	0.063	ns
Sub Saharan Africa	rs1728369	1	3.161	0.075	ns
Sub Saharan Africa	rs17077990	1	1.543	0.214	ns
Sub Saharan Africa	rs25768	1	0.489	0.484	ns
N America	rs2116791	Monomorphic			
N America	rs1728369	1	0.005	0.941	ns
N America	rs17077990	1	0.208	0.648	ns
N America	rs25768	1	0.001	0.971	ns
S America	rs2116791	Monomorphic			
S America	rs1728369	Monomorphic			
S America	rs17077990	1	0.789	0.374	ns
S America	rs25768	1	5.298	0.021	*
NE Asia	rs2116791	Monomorphic			
NE Asia	rs1728369	1	0.259	0.611	ns
NE Asia	rs17077990	1	0.043	0.835	ns
NE Asia	rs25768	1	1.563	0.211	ns
S Asia	rs2116791	1	3.325	0.068	ns
S Asia	rs1728369	1	0.190	0.663	ns
S Asia	rs17077990	1	1.098	0.295	ns
S Asia	rs25768	1	1.444	0.229	ns
SE Asia	rs2116791	1	0.958	0.328	ns
SEAsia	rs1728369	1	2.224	0.136	ns
SEAsia	rs17077990	1	0.109	0.742	ns
SEAsia	rs25768	1	0.750	0.386	ns
E Europe	rs2116791	1	16.327	0.000	***
E Europe	rs1728369	1	0.156	0.693	ns
E Europe	rs17077990	1	0.034	0.853	ns
E Europe	rs25768	1	0.834	0.361	ns
N Europe	rs2116791	1	0.571	0.450	ns

N Europe	rs1728369	1	0.209	0.648	ns
N Europe	rs17077990	1	0.057	0.812	ns
N Europe	rs25768	1	0.092	0.761	ns
S Europe	rs2116791	1	3.095	0.079	ns
S Europe	rs1728369	1	0.091	0.762	ns
S Europe	rs17077990	1	1.445	0.229	ns
S Europe	rs25768	1	0.380	0.538	ns
Mid East	rs2116791	1	0.047	0.828	ns
Mid East	rs1728369	1	0.087	0.768	ns
Mid East	rs17077990	1	0.111	0.739	ns
Mid East	rs25768	1	0.311	0.577	ns
Oceania	rs2116791	Monomorphic			
Oceania	rs1728369	1	2.619	0.106	ns
Oceania	rs17077990	1	6.953	0.008	**
Oceania	rs25768	1	5.024	0.025	*
Key: ns=not significant, * P<0.05, ** P<0.01, *** P<0.001					
Bonferroni Correction P< 0.000625					

Table 4-4. Chi Square results for the SNPs from all DNA samples used which form part the SNPSTRs. The "populations" are divided as per the DNA samples supplied and named as per the CEPH-HGDP listings and the HapMap listings.

From these results it can be seen that several of the SNPs were monomorphic in particular populations, and hence it was not possible to carry out a Chi-Square test. There were two significant departures from Hardy-Weinberg, notably in Oceania for rs17077990, and in Eastern Europe for rs2116791. Applying the Bonferroni correction , however makes these results no longer significant. Overall, the results show that there was no significant difference in the majority of cases, which suggests that the typing method used was correct.

A full list of the results of applying this test to the distribution of the STRs in all of the DNA samples is listed in Table 4-5.

Pop	Locus	DF	ChiSq	Prob	Signif
Cornish	D5S818	15	13.188	0.588	ns
Cornish	D16S539	28	14.529	0.983	ns
Cornish	D3S1358	21	13.850	0.876	ns
Cornish	CSF1PO	10	16.391	0.089	ns
HMYoruba	D5S818	21	26.847	0.176	ns
HMYoruba	D16S539	21	19.044	0.582	ns
HMYoruba	D3S1358	15	22.623	0.092	ns
HMYoruba	CSF1PO	28	32.491	0.255	ns
HMChina	D5S818	21	14.127	0.864	ns
HMChina	D16S539	15	12.053	0.675	ns
HMChina	D3S1358	21	13.707	0.882	ns
HMChina	CSF1PO	21	18.312	0.629	ns
HMJapan	D5S818	21	20.655	0.480	ns
HMJapan	D16S539	15	6.794	0.963	ns
HMJapan	D3S1358	10	7.513	0.676	ns
HMJapan	CSF1PO	28	25.600	0.595	ns
HM NW Europe	D5S818	21	36.865	0.017	*
HM NW Europe	D16S539	21	19.654	0.543	ns
HM NW Europe	D3S1358	28	33.461	0.219	ns
HM NW Europe	CSF1PO	15	25.267	0.046	*
N Africa	D5S818	10	10.483	0.399	ns
N Africa	D16S539	10	11.136	0.347	ns
N Africa	D3S1358	10	6.392	0.781	ns
N Africa	CSF1PO	6	7.029	0.318	ns
Sub Saharan Africa	D5S818	28	68.918	0.000	***
Sub Saharan Africa	D16S539	28	17.872	0.929	ns
Sub Saharan Africa	D3S1358	21	17.429	0.685	ns
Sub Saharan Africa	CSF1PO	36	54.736	0.023	*
N America	D5S818	10	10.680	0.383	ns
N America	D16S539	10	8.567	0.574	ns
N America	D3S1358	15	11.310	0.730	ns
N America	CSF1PO	6	2.639	0.853	ns
S America	D5S818	10	7.567	0.671	ns
S America	D16S539	10	11.368	0.330	ns
S America	D3S1358	10	4.136	0.941	ns
S America	CSF1PO	3	7.836	0.050	*
NE Asia	D5S818	10	4.293	0.933	ns
NE Asia	D16S539	15	11.680	0.703	ns
NE Asia	D3S1358	21	54.124	0.000	***
NE Asia	CSF1PO	28	57.492	0.001	***
S Asia	D5S818	15	8.572	0.899	ns
S Asia	D16S539	21	27.089	0.168	ns
S Asia	D3S1358	21	104.420	0.000	***
S Asia	CSF1PO	15	5.755	0.984	ns
SE Asia	D5S818	28	23.112	0.727	ns
SEAsia	D16S539	21	22.539	0.369	ns
SEAsia	D3S1358	28	109.816	0.000	***
SEAsia	CSF1PO	45	459.228	0.000	***
E Europe	D5S818	21	35.104	0.028	*
E Europe	D16S539	28	29.438	0.391	ns
E Europe	D3S1358	15	6.266	0.975	ns
E Europe	CSF1PO	28	24.669	0.646	ns
N Europe	D5S818	10	11.921	0.290	ns

N Europe	D16S539	21	15.493	0.797	ns
N Europe	D3S1358	10	11.589	0.314	ns
N Europe	CSF1PO	15	38.020	0.001	***
S Europe	D5S818	15	17.135	0.311	ns
S Europe	D16S539	21	9.078	0.989	ns
S Europe	D3S1358	15	16.057	0.378	ns
S Europe	CSF1PO	15	16.359	0.359	ns
Mid East	D5S818	21	19.623	0.545	ns
Mid East	D16S539	28	34.533	0.184	ns
Mid East	D3S1358	15	12.685	0.627	ns
Mid East	CSF1PO	28	21.791	0.791	ns
Oceania	D5S818	15	16.465	0.352	ns
Oceania	D16S539	21	30.147	0.089	ns
Oceania	D3S1358	6	5.203	0.518	ns
Oceania	CSF1PO	15	10.008	0.819	ns
Key: ns=not significant, * P<0.05, ** P<0.01, *** P<0.001					
Bonferroni Correction P< 0.000625					

Table 4-5. Chi Square results for the STRs which form part the SNPSTRs. The "populations" are divided as per the DNA samples supplied and named as per the CEPH-HGDP listings and the HapMap listings.

From this table it can be seen that there is significant departure from Hardy-Weinberg, even after applying the Bonferroni correction for the following loci and samples:

Pop	Locus	DF	ChiSq	Prob	Signif
Sub Saharan Africa	D5S818	28	68.918	0.000	***
NE Asia	D3S1358	21	54.124	0.000	***
S Asia	D3S1358	21	104.420	0.000	***
SE Asia	D3S1358	28	109.816	0.000	***
SE Asia	CSF1PO	45	459.228	0.000	***

Table 4-6. STRs which did not conform to Hardy-Weinberg equilibrium.

Data from a recent paper also found non-conformity with Hardy-Weinberg equilibrium for the D5S818 STR within a population from Egypt (Omran et al. 2009), and the suggested reason for this was due to population admixture and substructure. These could also be the reason for the departure from Hardy-Weinberg observed in the Sub-Saharan populations tested here, however, it is

more likely due to the fact that the populations which are used, are not true representatives of real populations, but are instead, groups of individuals. These have then been pooled together into a non-panmictic group.

The departure from Hardy-Weinberg equilibrium observed in the North East Asian samples for the D3S1358 locus is most likely due to the relatively small sample size (22 samples only).

There remains significant departure from Hardy-Weinberg equilibrium for the D3S1358 locus in the South Asian samples. As all of the STR typing results were continually checked against those carried out by Peter de Knijff's laboratory (P. de Knijff (2006) personal communication), typing errors can be eliminated as an explanation and hence it must be concluded that either the sampling method of the South Asian DNAs has caused this, or that the observed departure is due to the action of selection.

The other departures from Hardy-Weinberg equilibrium are for the D3S1358 and CSF1PO loci in South East Asia. This could be caused by the relatively small sample size of only 17 individuals.

However, the main point to stress is that overall, for all SNPSTRs tested for Hardy-Weinberg, the results show that there was no significant difference in the majority of cases, which suggests that the SNPSTR typing method used throughout was correct.

4.4.2 Time to most recent common ancestor (TMRCA)

Diversity of STR alleles associated with the derived SNP at each SNPSTR was used to estimate the TMRCA of the SNP in each case. This was carried out in order to see if the estimated TMRCA of each SNP was consistent with what is known about human population history and the distribution of the SNP derived alleles observed here. In order to ascertain this, the TMRCA for each of the SNPSTRs needed to be estimated.

In theory, it should be possible to calculate a TMRCA figure for the derived SNP at each SNPSTR system to estimate the length of time since each of the STR alleles associated with the derived SNP allele shared a common ancestor.

TMRCA can be calculated by two main methods. The first method is one that requires a population model, and the second method uses the diversity already present in a group of chromosomes without the need for a population model. Those that use population models are based on coalescent simulations where the evolution of haplotypes are simulated backward in time until the haplotypes present in the SNPSTR being studied are incorporated into one genealogy. This type of method requires knowledge of the population demography and is prone to error if there is any deviation from the correct model (Tang et al. 2002) . Non model-based methods rely simply on knowledge of allelic diversity in order to date an allele. When a novel haplotype is created by a SNP mutation near an STR, the mutation will be present on only one

chromosome. Over time, descendent chromosomes will also acquire new STR diversity while still retaining the mutational marker SNP of the founding haplotype. The amount of diversity among these chromosomes can then be used to calculate the TMRCA if the mutation rate and generation time for these chromosomes are known. It is, however, very easy for error in the mutation rate to give very false estimates of the TMRCA (Tremblay and Vezina 2000).

Statistic	**Usage**
Rho (ρ) Measures mean number of mutations between the root and each individual in the sample.	DNA sequence and STR with haplotype root specified. Each haplotype is compared to the ancestral root and time is calculated using this as a foundation.
Variance For data where a smooth mismatch distribution of intra-allelic diversity is present. Variance = variance accumulated since the foundation.	Measures population expansions using microsatellite data – and calculates TMRCA as the time of initial expansion. There is no need for a root haplotype or phylogeny.
ASD (Average Squared Distance) Between the root and microsatellite data.	Knowledge of root haplotype is needed. Uses the stepwise mutation model to calculate TMRCA and thus can only be used for microsatellite data. However, does not need to construct a phylogeny as is calculated from data directly.

Table 4-7. TMRCA summary statistics.
This table shows the three main types of summary statistic which can be used to calculate TMRCA. (Di Rienzo et al. 1994; Goldstein et al. 1995; Forster et al. 1996).

Unfortunately none of these can be used with regards to calculating the TMRCA of a SNPSTR, which contains elements with two different mutation rates. The SNP has a slower mutation rate of ~ 10^{-8} per base per generation (Nachman and Crowell 2000) and the STR has a much faster mutation rate of ~10^{-3} per locus per generation (Brinkmann et al. 1998). To date, there is no software available which is able to take both mutation rates into account and for this reason the TMRCA of the SNPSTRs has been calculated using a method which was originally devised in 1943 for bacterial mutations (Luria and Delbruck 1943), and has been used subsequently to estimate SNP TMRCA based on flanking STR haplotype diversity (Underhill et al. 1996). The TMRCA has been calculated for the SNPSTRs using the ancestral STR allele which is associated with the derived SNP. The ancestral state of the STR is deduced by using the STR repeat number with the highest frequency. This was calculated using the data from all of the samples which were typed. For example, the D5S8181 STR range associated with the derived SNP for all of the populations can be seen in the figure below, where it can be seen that the STR repeat number with the highest frequency for the derived SNP is 13 repeats.

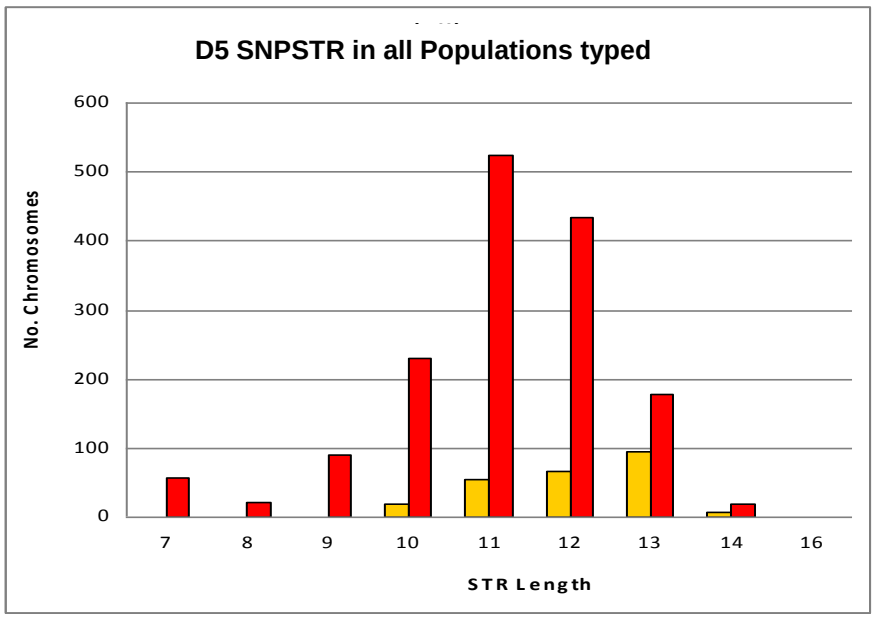


Figure 4-3. Chart showing the D5S818 SNPSTR data for all populations typed. The STR lengths associated with the ancestral SNP allele is shown in red, the derived is in yellow. Not visible on this chart is one chromosome with 9 STR repeats associated with the minor allele and one chromosome associated with 16 STR repeats associated with the major allele.

The Luria and Delbrook method used the following formula to calculate the TMRCA of the bacterial mutations:

(n/ m) / STR mutation rate x generation time, where (n) = the number of chromosomes in the data set which represent the ancestral state of STR repeats associated with the derived SNP allele, and (m) = the number of chromosomes remaining in the data set which are associated with the non-ancestral STR repeat numbers, associated with the derived SNP. So for the D5 SNPSTR this is calculated as below.

There are 95 chromosomes (n) in the data set which contain an STR of 13 repeats only associated with the derived SNP allele. As 13 repeats is the most common number of repeats, this is taken to be the ancestral state. There are 149

chromosomes (m) who share between them the 9 to 12 and 14 STR repeats. n/ m = 0.64. The mutation rate for the STR has been taken from STRbase and is 0.11% (http:/ / www.cstl.nist.gov/ strbase/ mutation.htm) and therefore time (t) is calculated as 0.64 / 0.11% = 579.6. The generation time used to make the calculations in this case is 30 years (Fenner 2005). So t x 30 = 17388.65 (approximately 17500 years before present [YBP]) This is the TMRCA for the derived SNP at the D5S818 SNPSTR.

Figure 4-4 shows the different ages for the derived SNPs for each of the four SNPSTRs studied, calculated in the manner stated above.

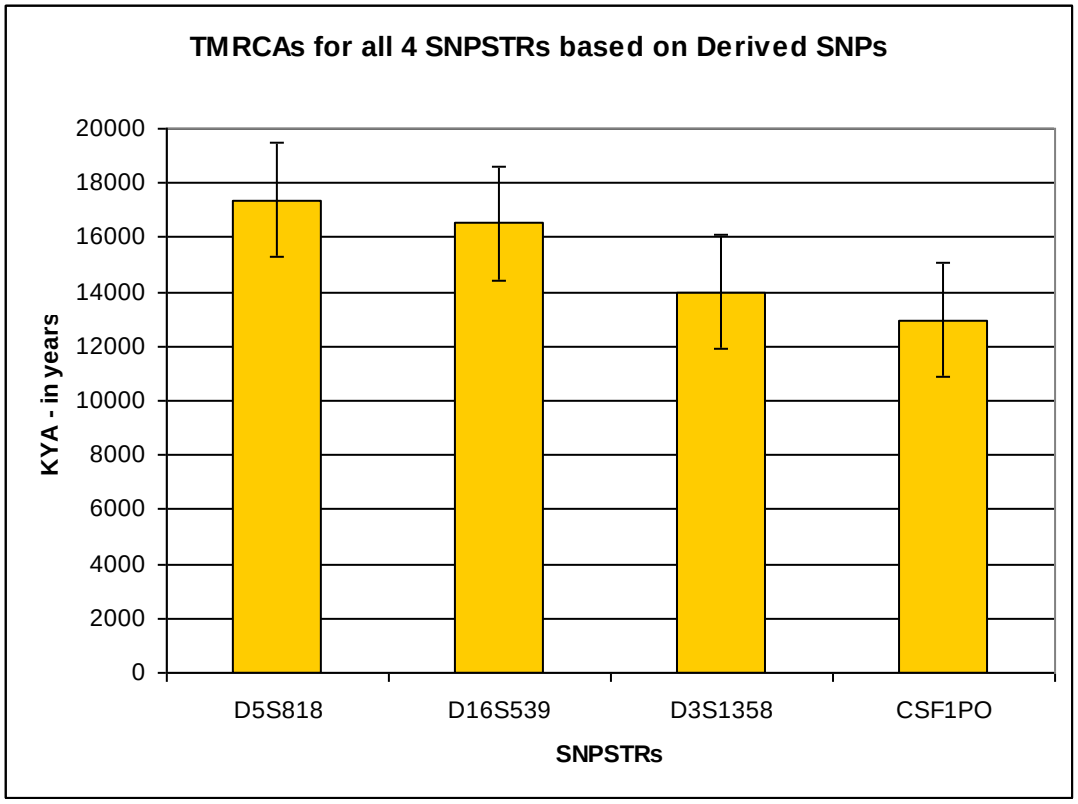


Figure 4-4. TMRCA of derived SNP alleles based on the ancestral STRs. This is taken from global populations which include the HapMap, CEPH-HGDP, Cornish, African-Caribbean, Danish and Greenland-Inuit DNA samples (data on the latter are described in Chapter 5). Error bars indicate standard deviation.

As can been seen from the figure above, all of these are between 13,000 and 17,500 years old, which places the origins of the derived SNPs after the last glacial maxim which occurred between 18 to 21 KYA, and also during the period of time when the first humans were colonising the Americas at around 15 KYA (Jobling et al. 2003)

4.4.3 Ancestral allele frequencies for rs1728369 - comparisons

As briefly discussed in the introduction, it was possible to compare the results from the D16 SNPSTR SNP (rs1728369) which has been typed by dbSNP (build 130) on the CEPH-HGDP, with the results achieved as part of the SNPSTR typing in this project. This provides another check on the accuracy of the typing methods, as it will highlight any significant differences, which would suggest typing errors. The results of performing the comparison are seen in the table below.

Population	dbSNP build 130	D16 SNPSTR
Adygei	0.8235	0.84375
Balochi	0.8333	0.81818
BantuKenya	0.4545	0.4375
BantuSouthAfrica	0.625	0.66
Basque	0.7917	0.791667
Bedouin	0.6444	0.720588
BiakaPygmy	0.5227	0.54
Brahui	0.86	0.86
Burusho	0.72	0.695
Cambodian	0.95	0.9
Colombian	1	1

Dai	0.85	0.83333
Daur	0.6111	0.6111
Druze	0.7024	0.720588
French	0.875	0.88
Han	0.8235	0.764706
Han-NChina	0.8	0.764706
Hazara	0.75	0.738095
Hezhen	0.7222	0.75
Italian	0.75	0.714286
Japanese	0.7857	0.788462
Kalash	0.9783	0.979167
Karitiana	1	1
Lahu	0.625	0.55
Makrani	0.86	0.854167
Mandenka	0.5227	0.416667
Maya	0.9524	0.978261
MbutiPygmy	0.7308	0.857143
Melanesian	0.8636	0.8636
Miao	0.9	0.9
Mongola	0.8	0.77778
Mozabite	0.6481	0.66667
Naxi	0.75	0.72222
Orcadian	0.7333	0.71875
Oroqen	0.6667	0.7
Palestinian	0.7283	0.734043
Papuan	0.8529	0.8529
Pathan	0.9318	0.869565
Pima	1	1
Russian	0.78	0.770833
San	0.6	0.58333
Sardinian	0.8393	0.8392

She	0.85	0.888889
Sindhi	0.9583	0.9565
Surui	1	1
Tu	0.85	0.8125
Tujia	0.9	0.9
Tuscan	0.7143	0.75
Uygur	0.9	0.9
Xibo	0.7778	0.75
Yakut	0.72	0.6875
Yi	0.8	0.8
Yoruba	0.4762	0.5

Table 4-8. Comparisons of the ancestral allele frequencies of rs1728369 between dbSNP and the same SNP typed as part of the D16 SNSPTR.

From the table above, it can be seen that the typing of the D16 SNP for this thesis was similar to that carried out by dbSNP, which is a validation for the results achieved for the SNPSTRs. Overall, the difference between the two is only 0.46%. The larger observed differences are most likely due to smaller population sizes.

4.4.4 CEPH-HGDP, HapMap and Cornish SNPs

The figures below show the world-wide locations of the four SNPSTRs typed, and the major and minor alleles found in each population.

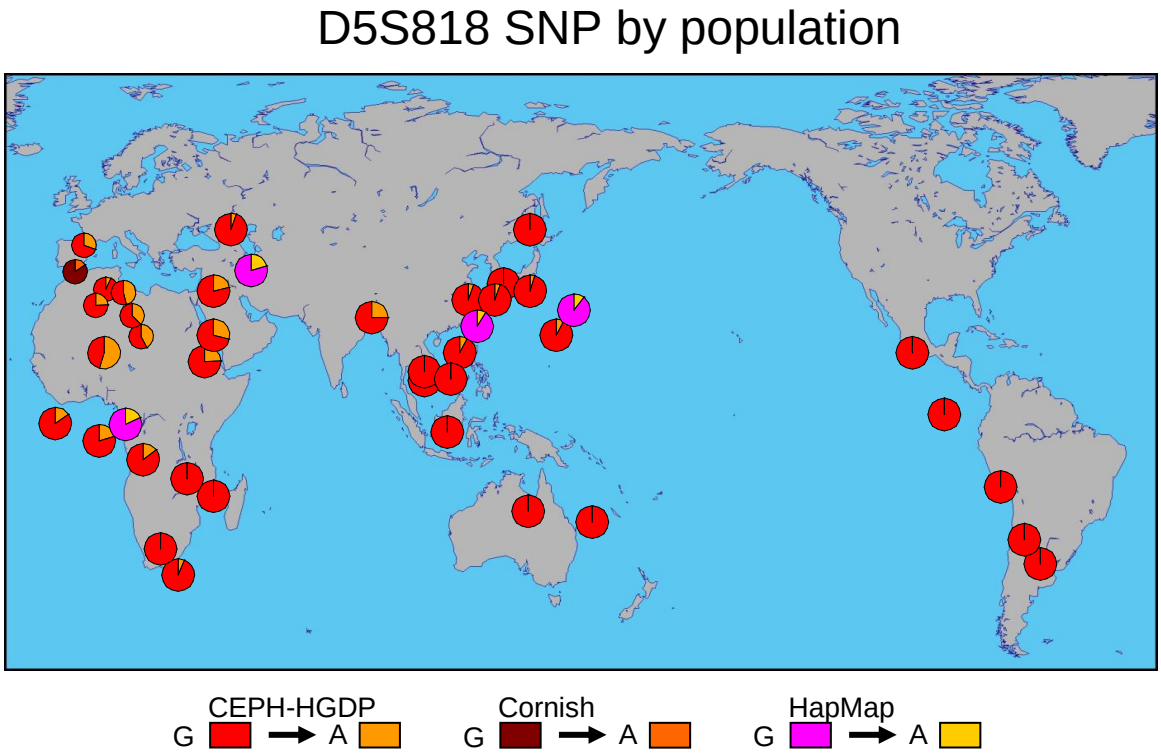


Figure 4-5. World map of the D5S818 SNPSTR SNP, shown by population. Red, brown and pink colours indicate major alleles; gold, orange and yellow indicate minor alleles.

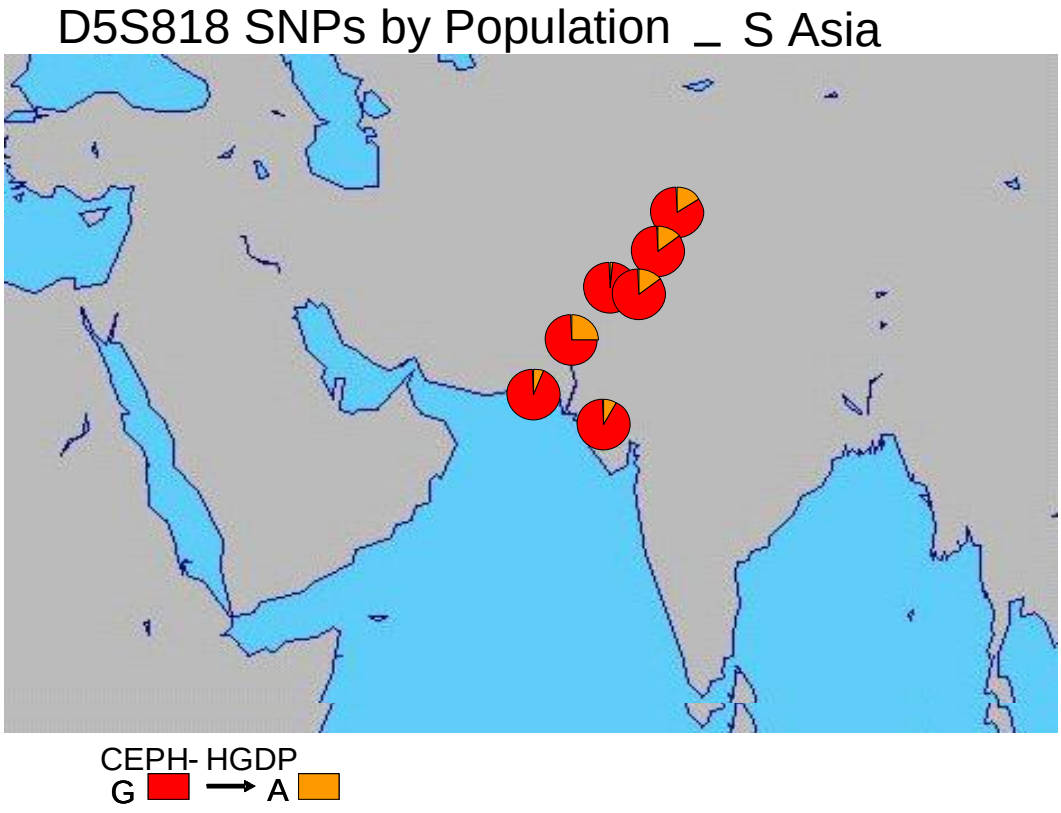


Figure 4-6. S. Asia map of the D5S818 SNPSTR SNP. Red indicates the major allele and gold indicates the minor allele.

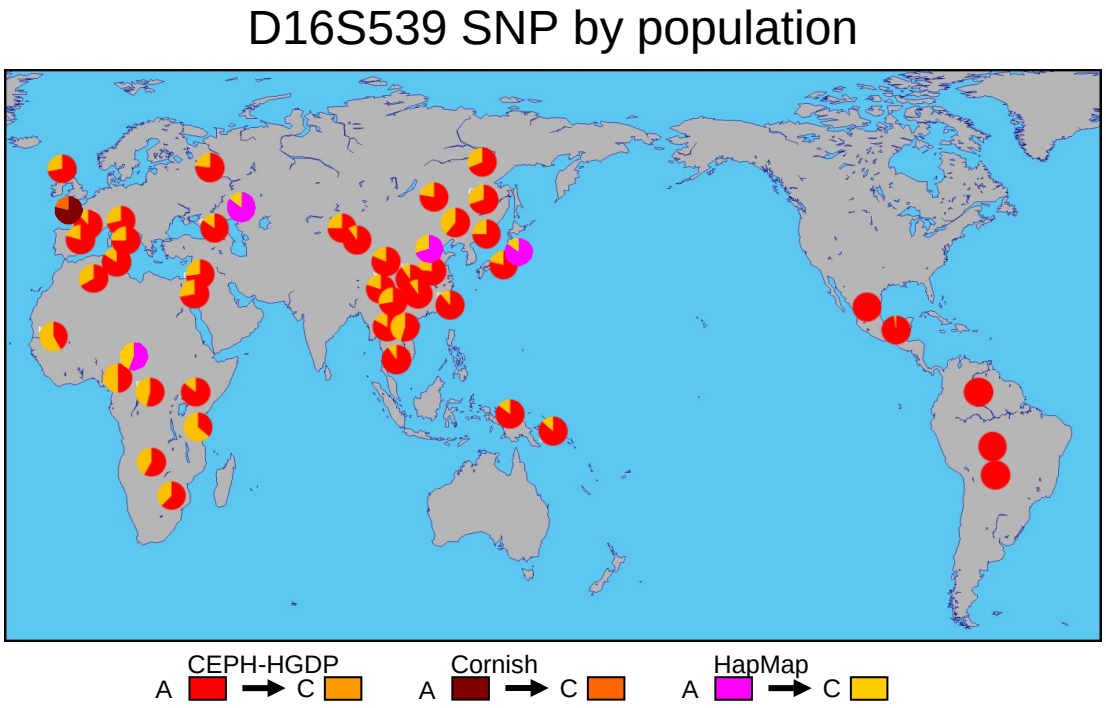


Figure 4-7. World map of the D16S539 SNPSTR SNP, shown by population. Red, brown and pink colours indicate major alleles; gold, orange and yellow indicate minor alleles.

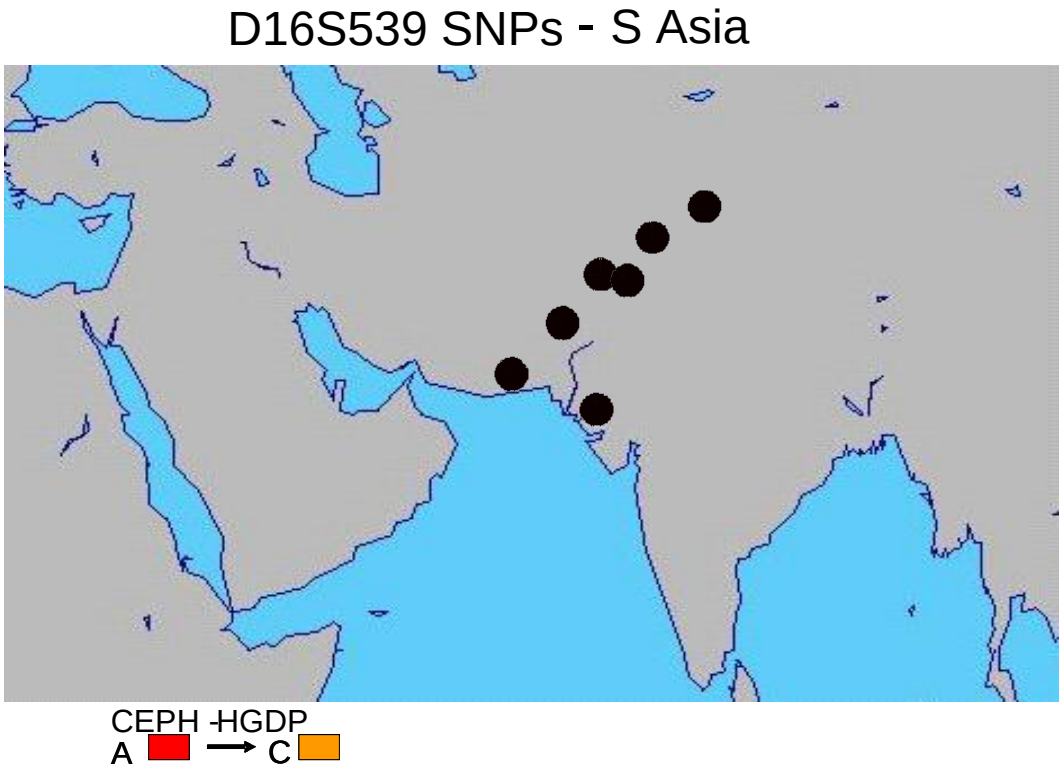


Figure 4-8. S. Asia map of the D16S539 SNPSTR SNP. Red indicates the major allele and gold indicates the minor allele.

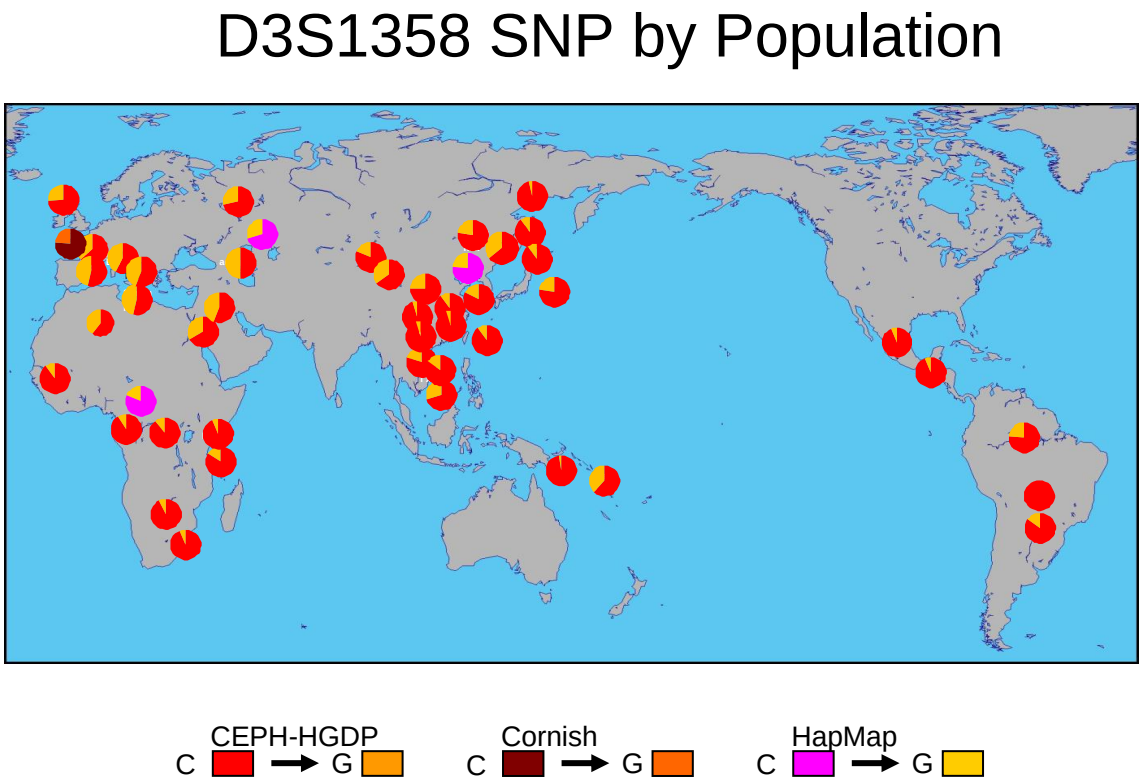


Figure 4-9. World map of the D3S1359 SNPSTR SNP, shown by population. Red, brown and pink colours indicate major alleles; gold, orange and yellow indicate minor alleles.

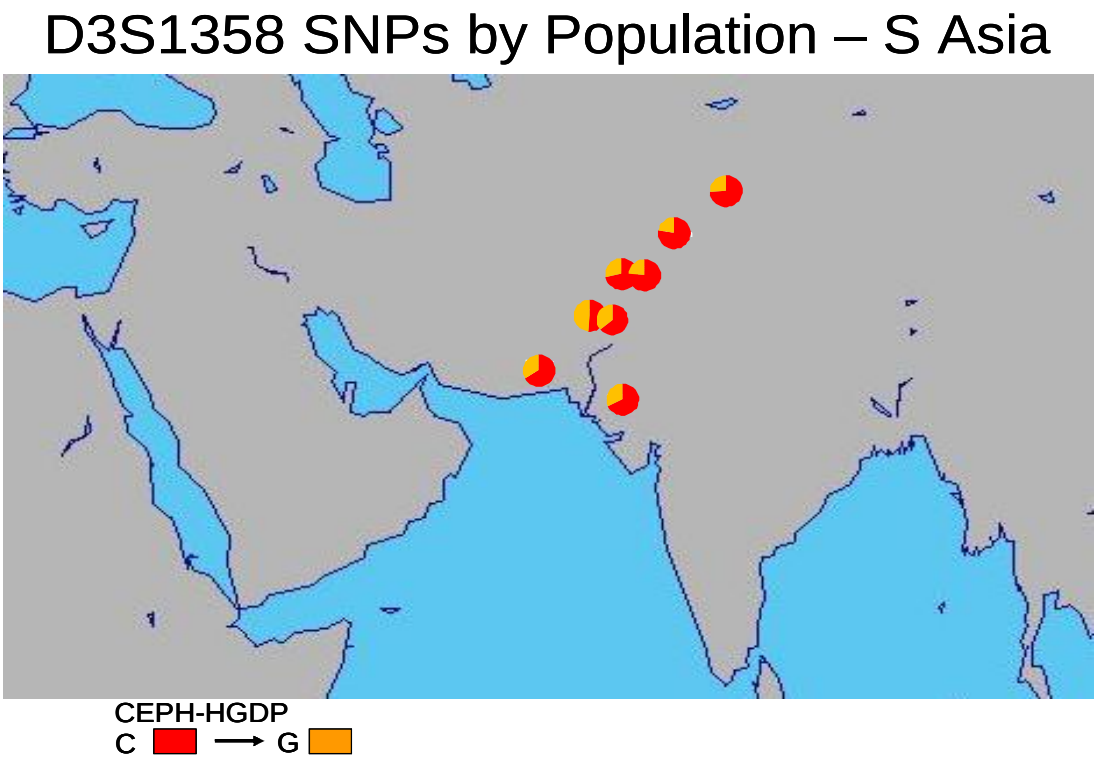


Figure 4-10. S. Asia map of the D3S1358 SNPSTR SNP. Red indicates the major allele and gold indicates the minor allele.

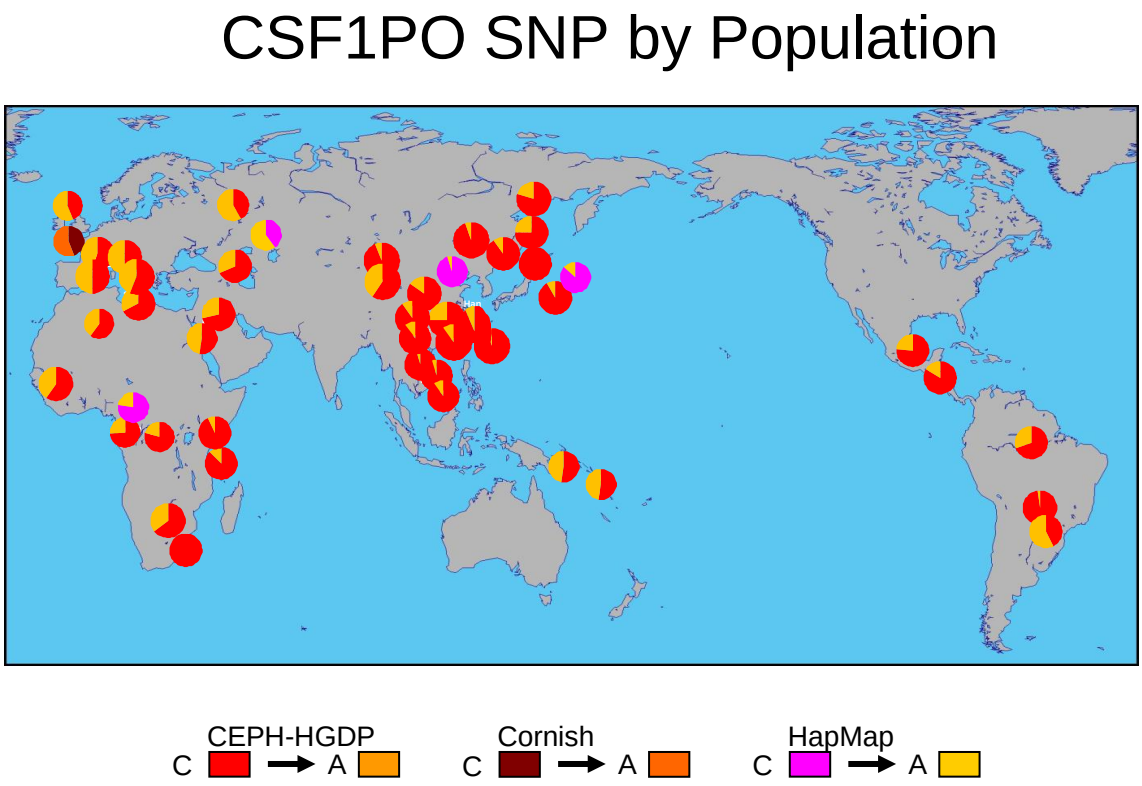


Figure 4-11. World map of the CSF1PO SNPSTR SNP, shown by population. Red, brown and pink colours indicate major alleles; gold, orange and yellow indicate minor alleles.

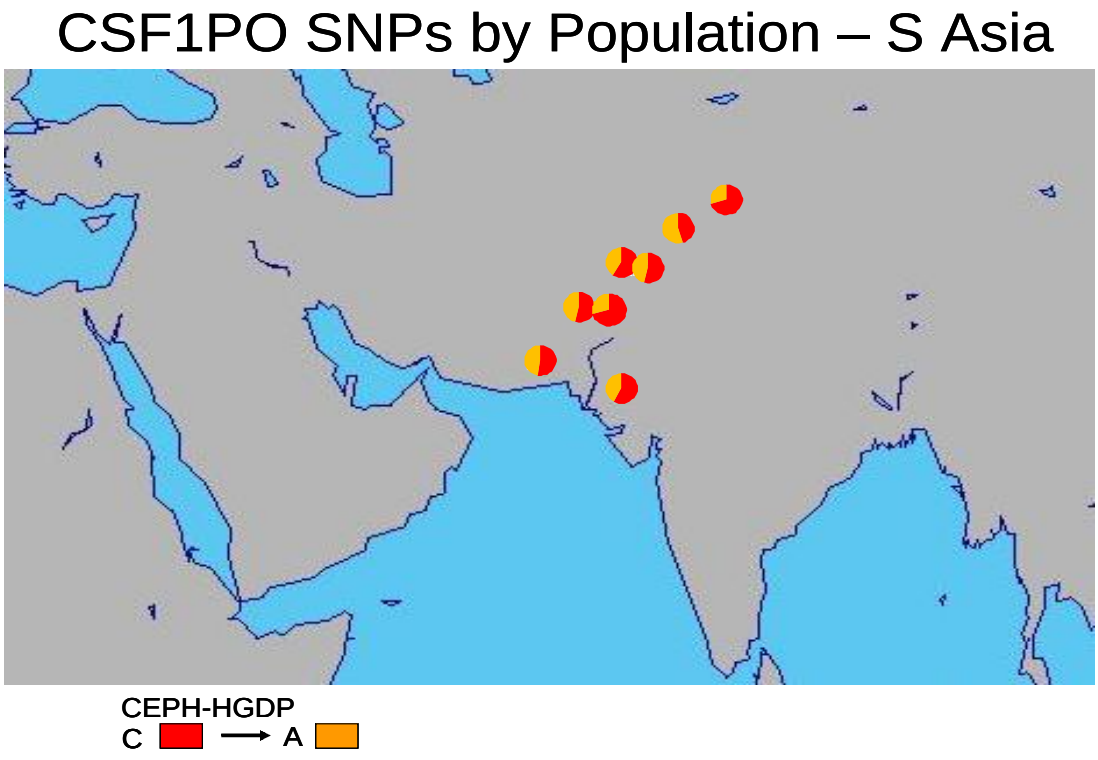


Figure 4-12. S. Asia map of the D3S1358 SNPSTR SNP. Red indicates the major allele and gold indicates the minor allele.

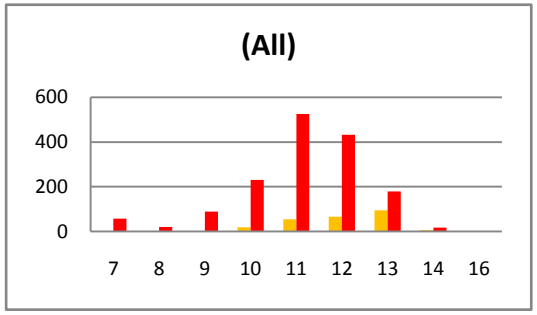
The figures above only show the distribution of the SNPs and do not take into account the added information which is provided by the forensic STRs.

4.4.5 CEPH-HGDP, HapMap and Cornish SNPSTRs

In order to be able to obtain a much fuller picture of the world-wide distribution of the SNPSTRs, the figures below show the combined SNP and STR data for all subpopulations of the CEPH-HGDP DNA panel as well as the HapMap and Cornish DNA samples.

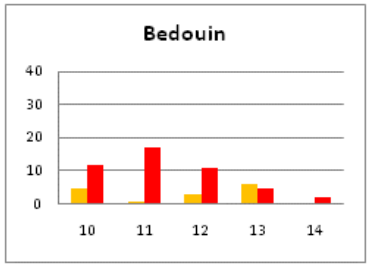
Section A - D5S818 SNPSTRs in all populations typed

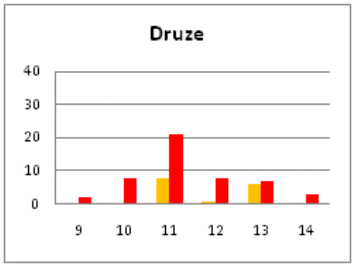
G → A

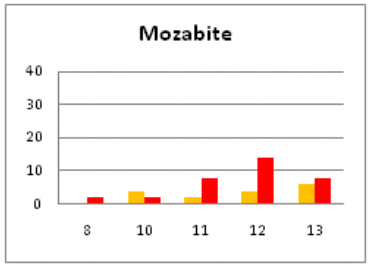


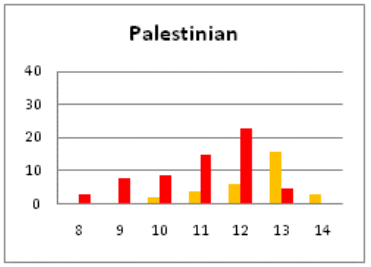
From this chart, it is possible to see that the full range of STR repeats in all populations typed is between 7 and 16 repeats. The smaller charts below do not show the full range possible.

D5S818 SNPSTRs in the Middle East

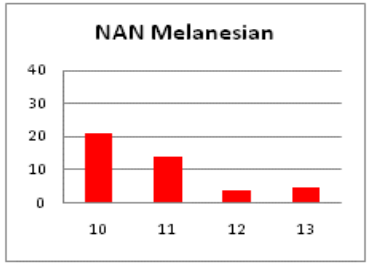


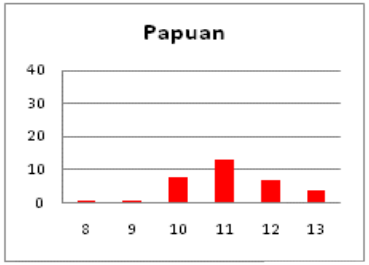






D5S818 SNPSTRs in Oceania

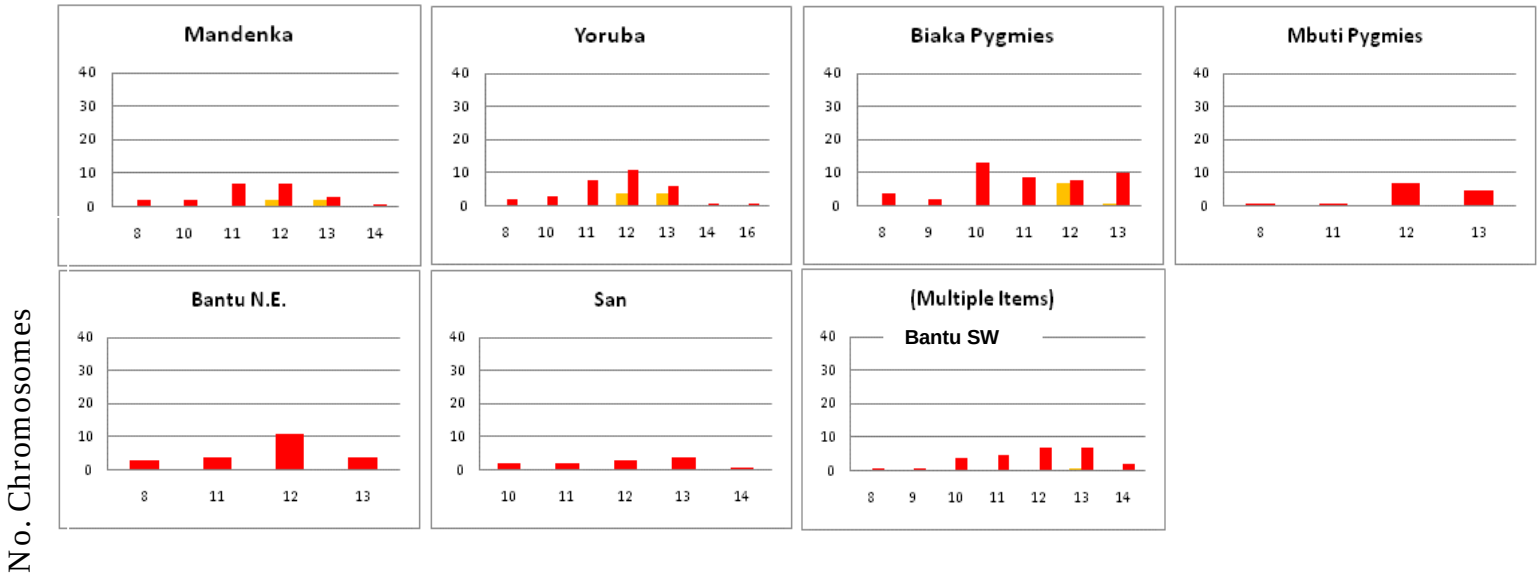




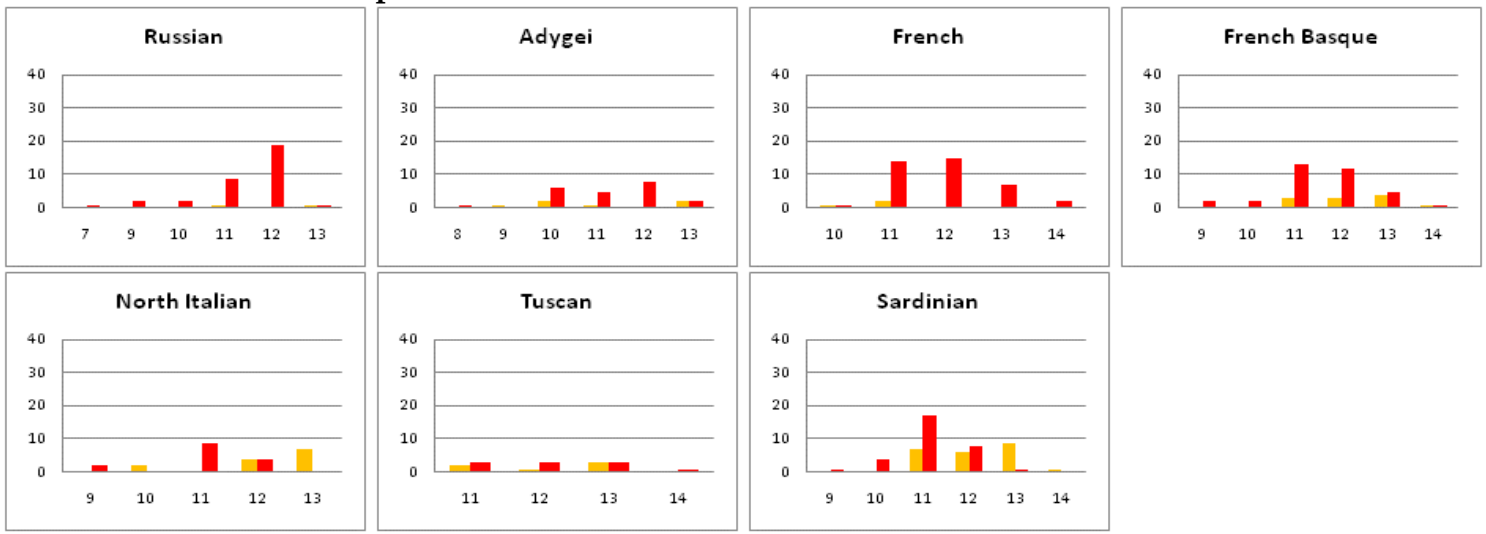
No. Chromosomes

STR Length/ repeats

D5S818 SNPSTRs in Africa

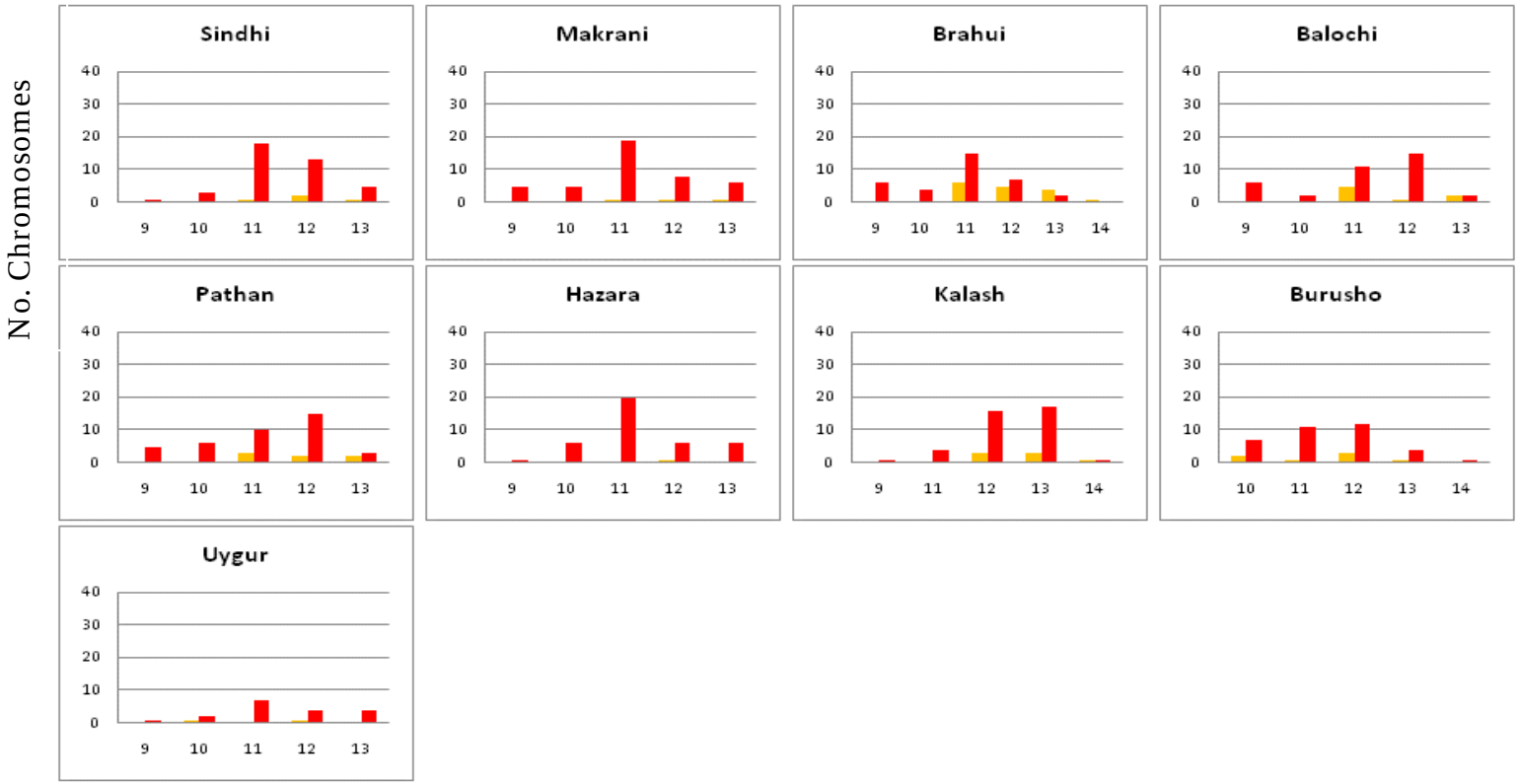


D5S818 SNPSTRs in Europe

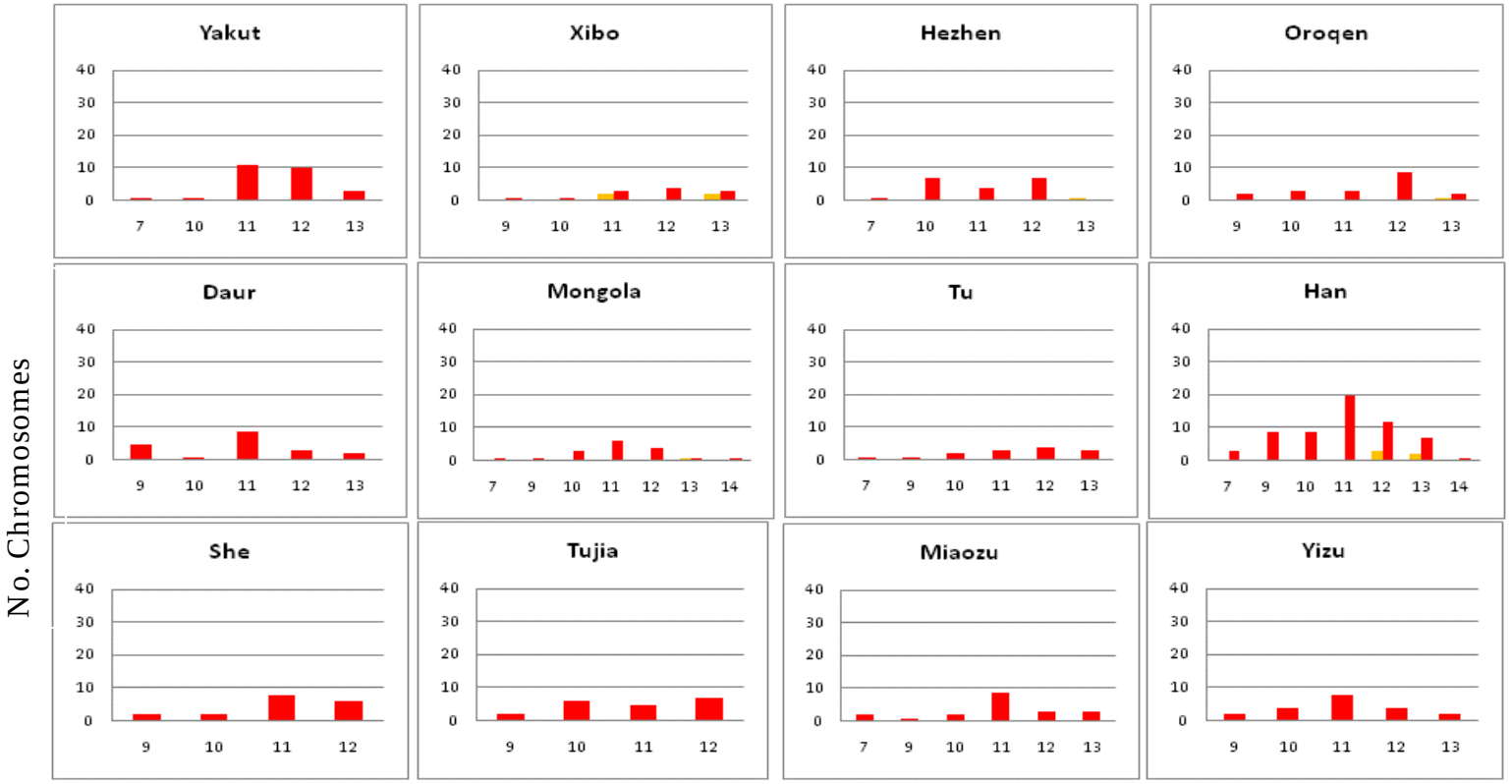


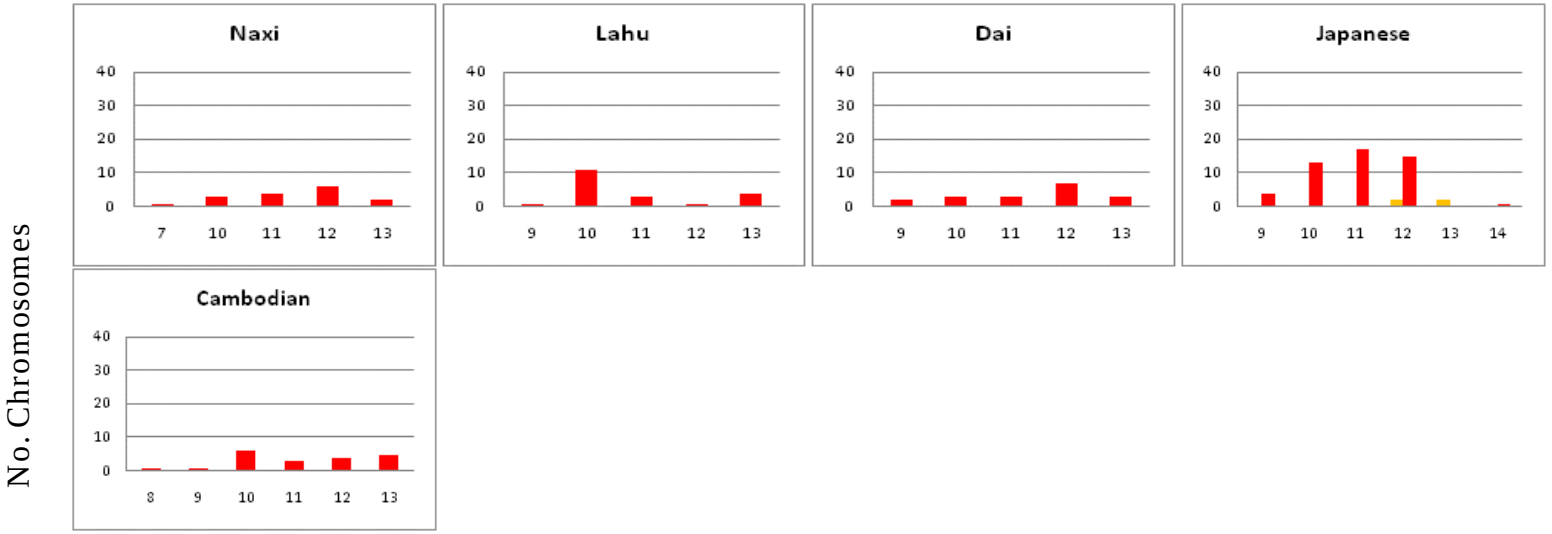
STR Length/ repeats

D5S818 SNPSTRs in Central and South Asia

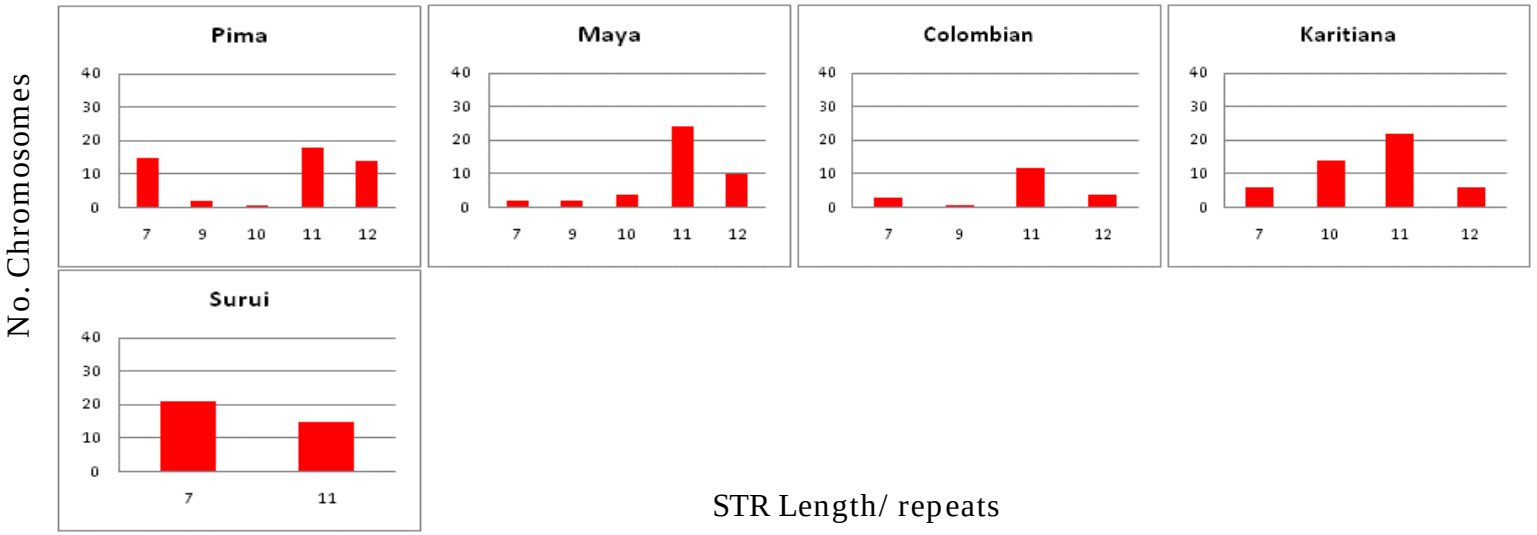


D5S818 SNPSTRs in East Asia



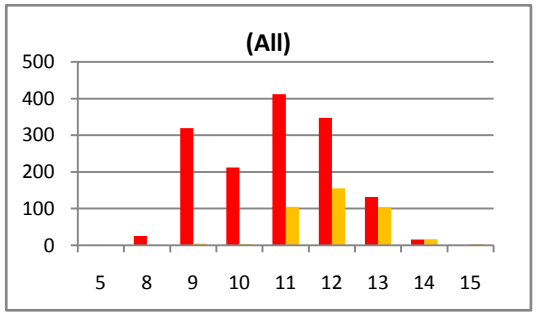

Naxi
40
30
20
10
0
7
10
11
12
13
Lahu
40
30
20
10
0
9
10
11
12
13
Dai
40
30
20
10
0
9
10
11
12
13
Japanese
40
30
20
10
0
9
10
11
12
13
14
Cambodian
40
30
20
10
0
8
9
10
11
12
13
No. Chromosomes
STR Length/ repeats

D5S818 SNPSTRs in America



Section B - D16S539 SNPSTRs in all populations typed

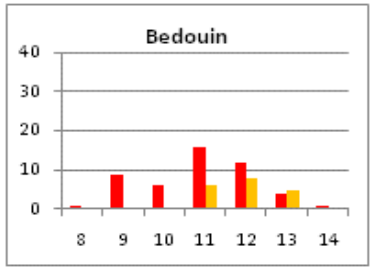
A → C

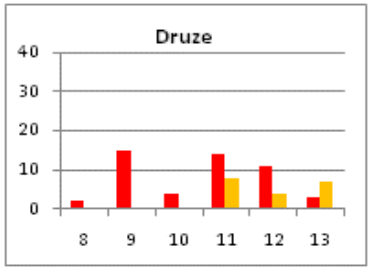


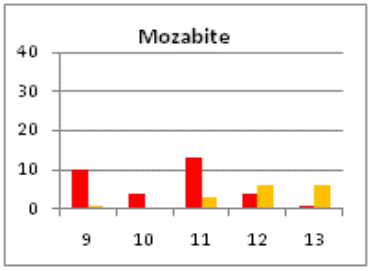
From this chart, it is possible to see that the full range of STR repeats in all populations typed is between 5 and 15 repeats. The smaller charts below do not show the full range possible.

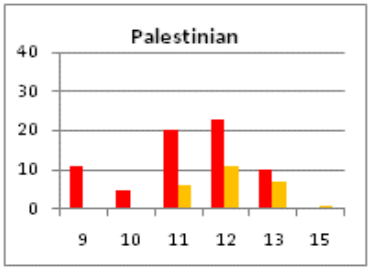
No. Chromosomes

D16S539 SNPSTRs in the Middle East

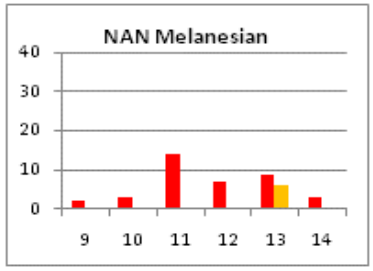


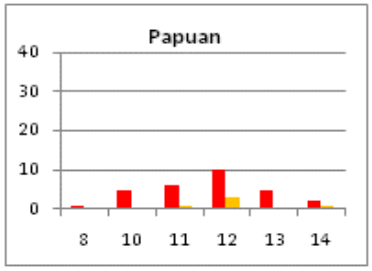






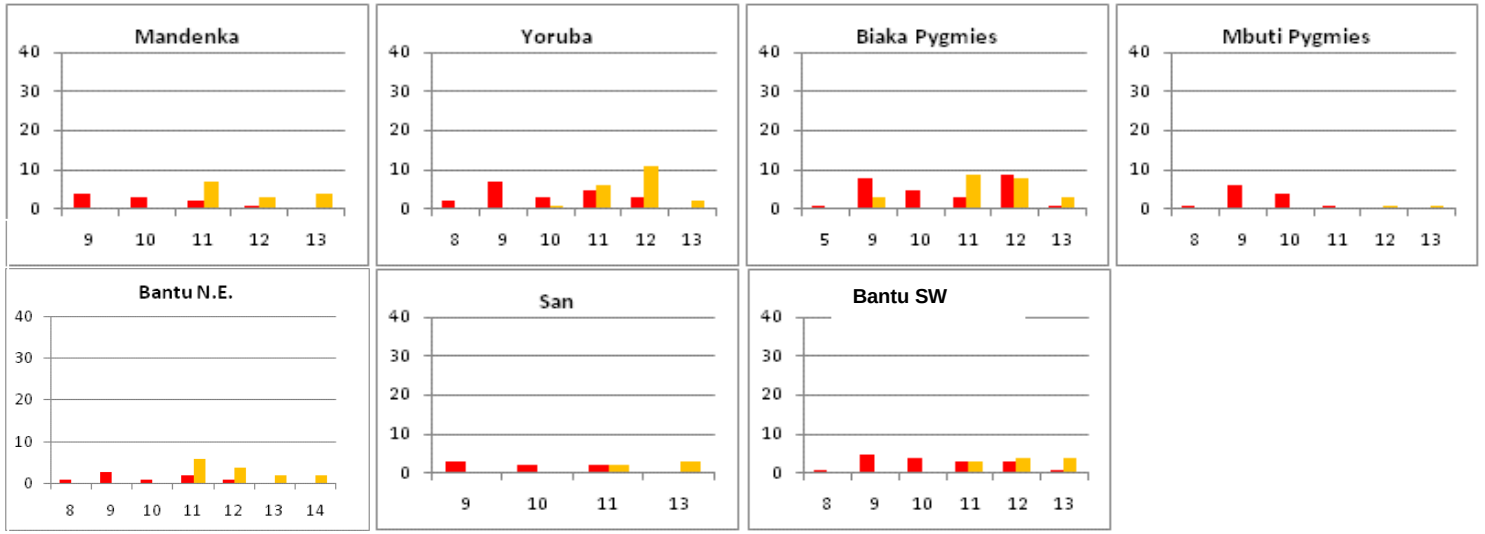
D16S539 SNPSTRs in Oceania



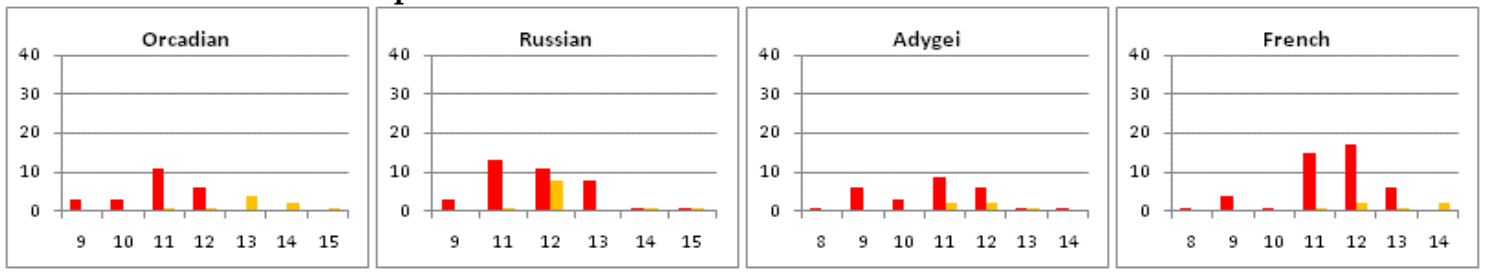


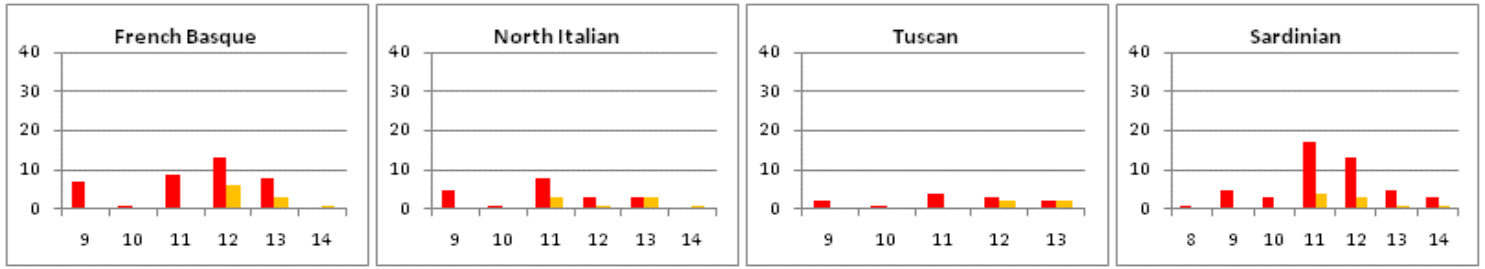
STR Length/ repeats

D16S539 SNPSTRs in Africa

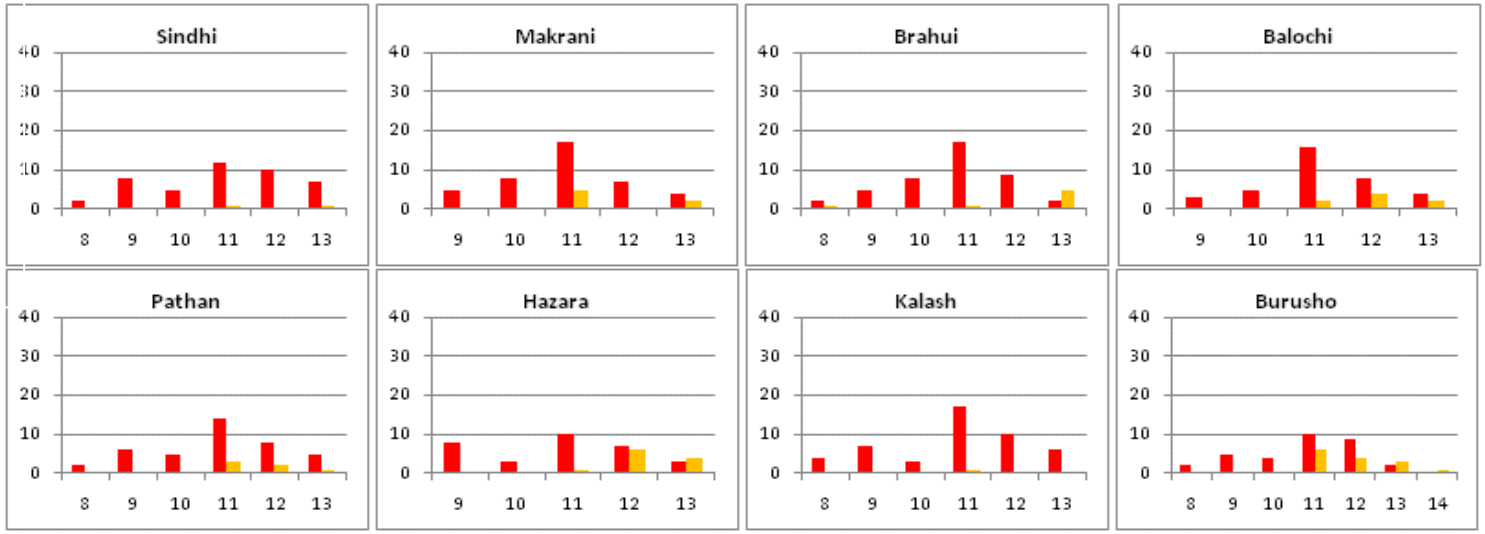


D16S539 SNPSTRs in Europe



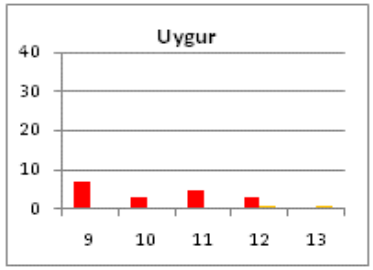


D16S539 SNPSTRs in Central and South Asia

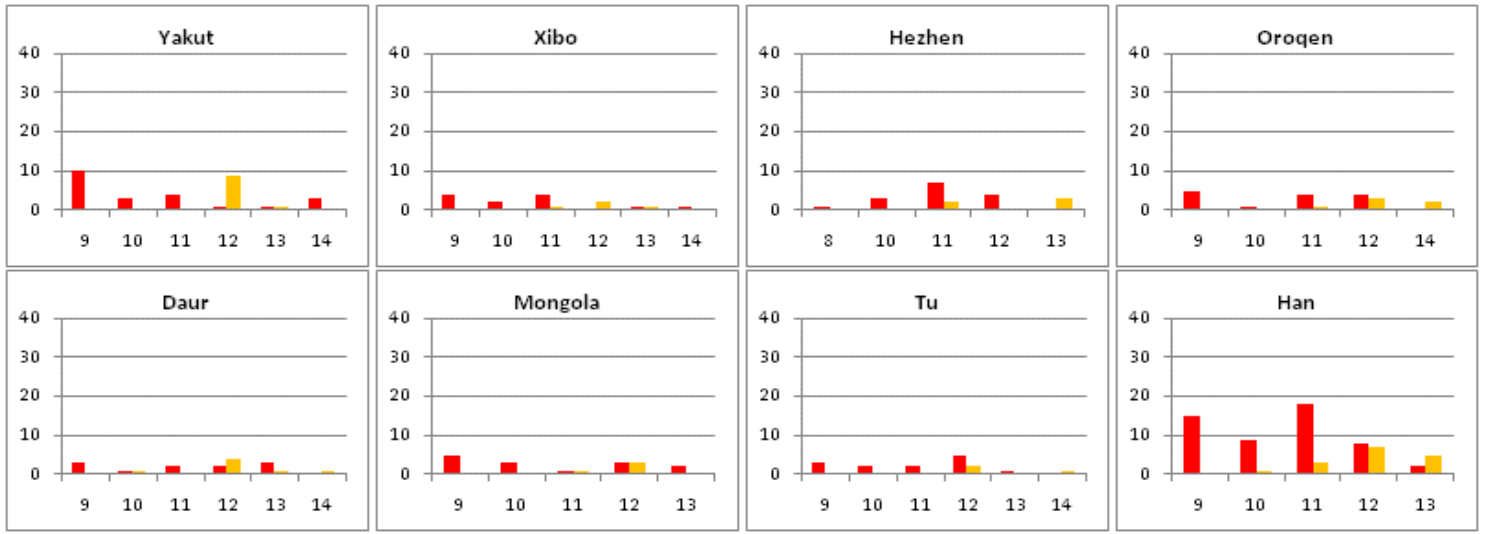


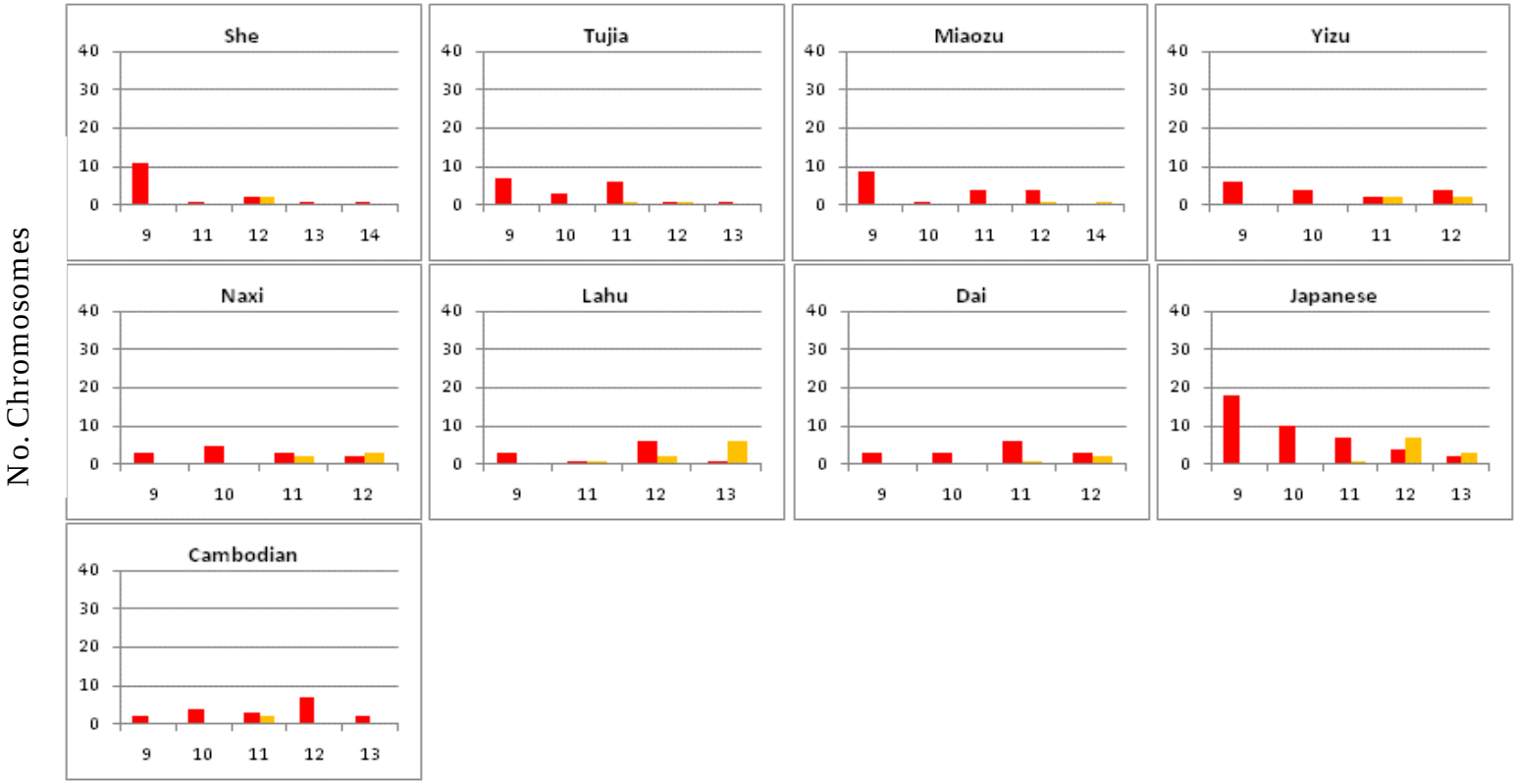
No. Chromosomes

STR Length/ repeats

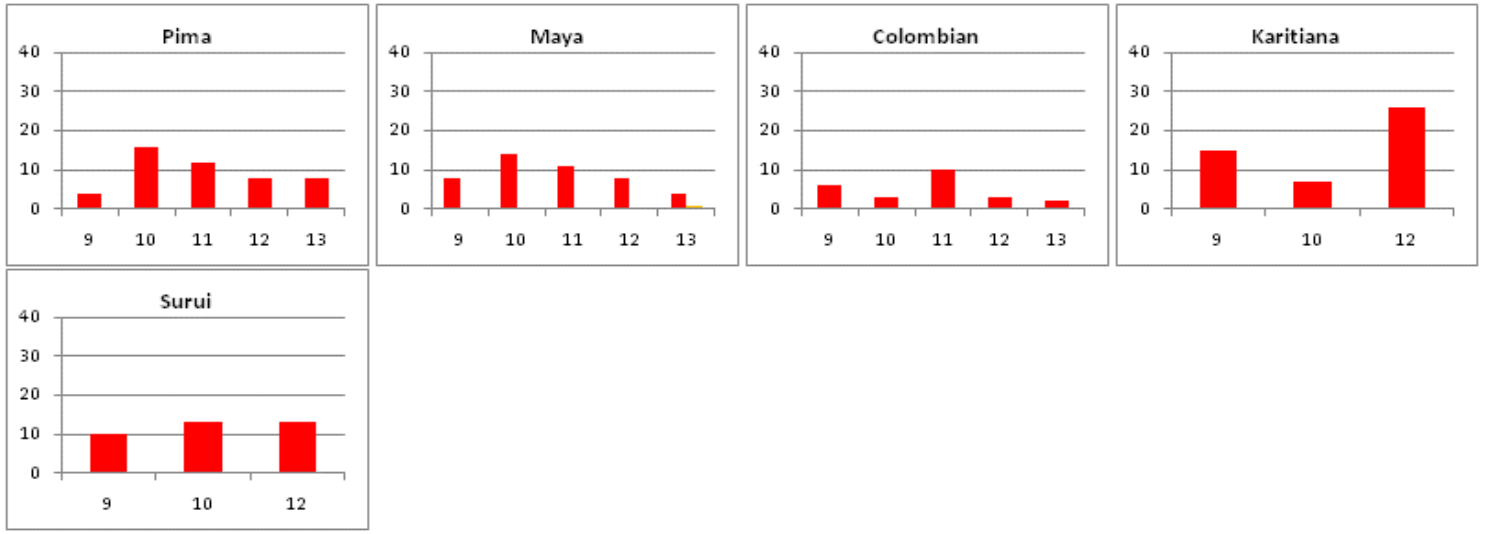


D16S539 SNPSTRs in South East Asia




No. Chromosomes
She
Tujia
Miaozu
Yizu
Naxi
Lahu
Dai
Japanese
Cambodian
0
10
20
30
40
STR Length/ repeats

D16S539 SNPSTRs in America

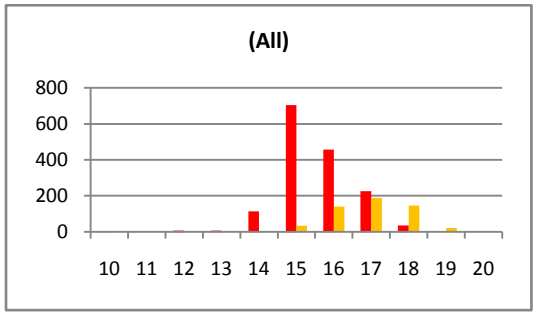


No. Chromosomes

STR Length/ repeats

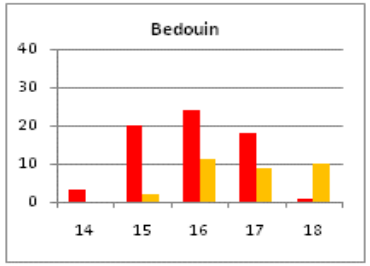
Section C - D3S1358 SNPSTRs in all populations typed

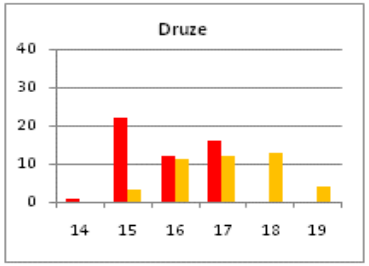
C → G

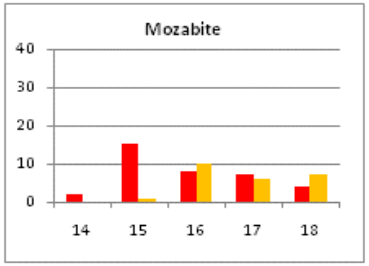


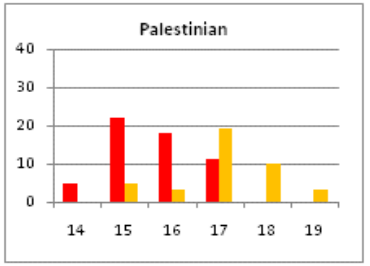
From this chart, it is possible to see that the full range of STR repeats in all populations typed is between 10 and 20 repeats. The smaller charts below do not show the full range possible.

D3S1358SNPSTRs in the Middle East



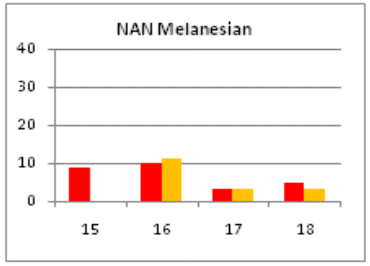


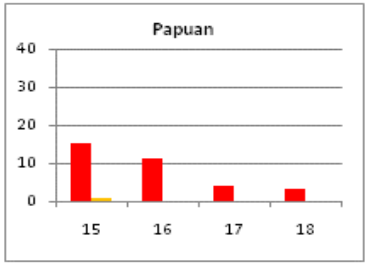




No. Chromosomes

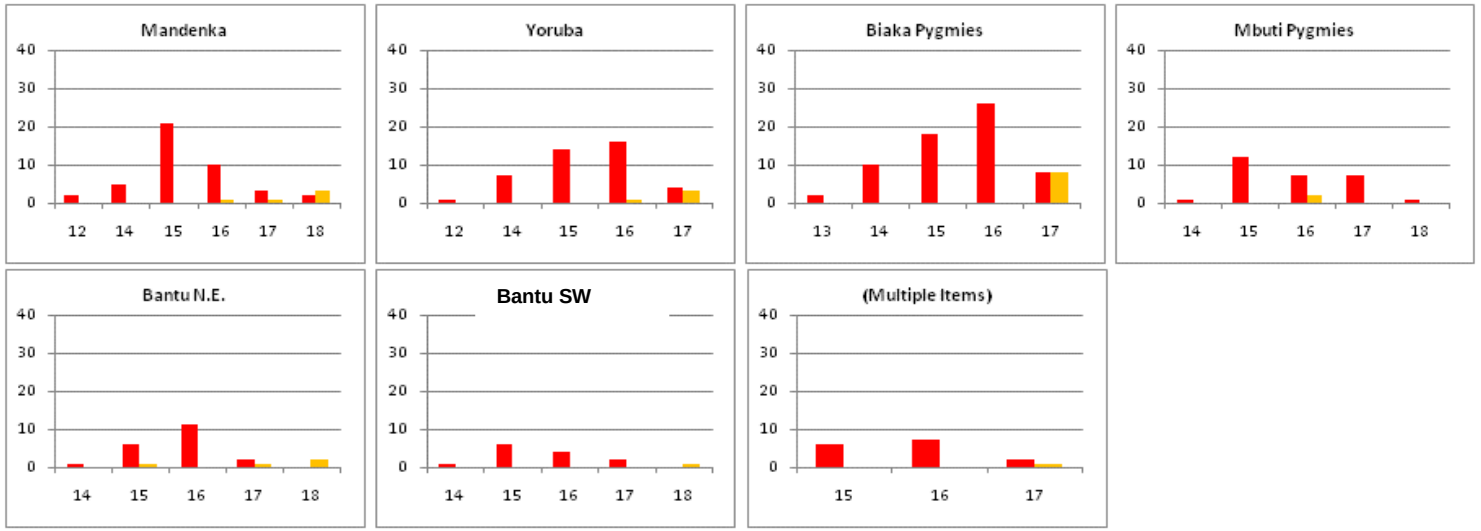
D3S1358SNPSTRs in Oceania



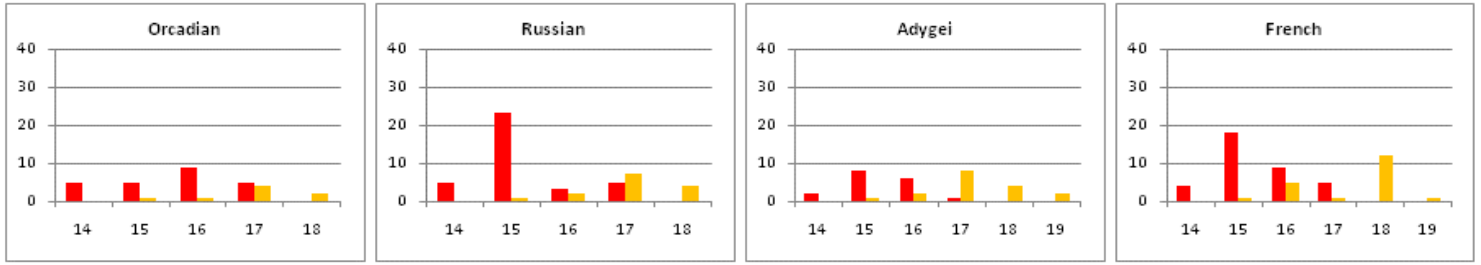


STR Length/ repeats

D3S1358 SNPSTRs in Africa

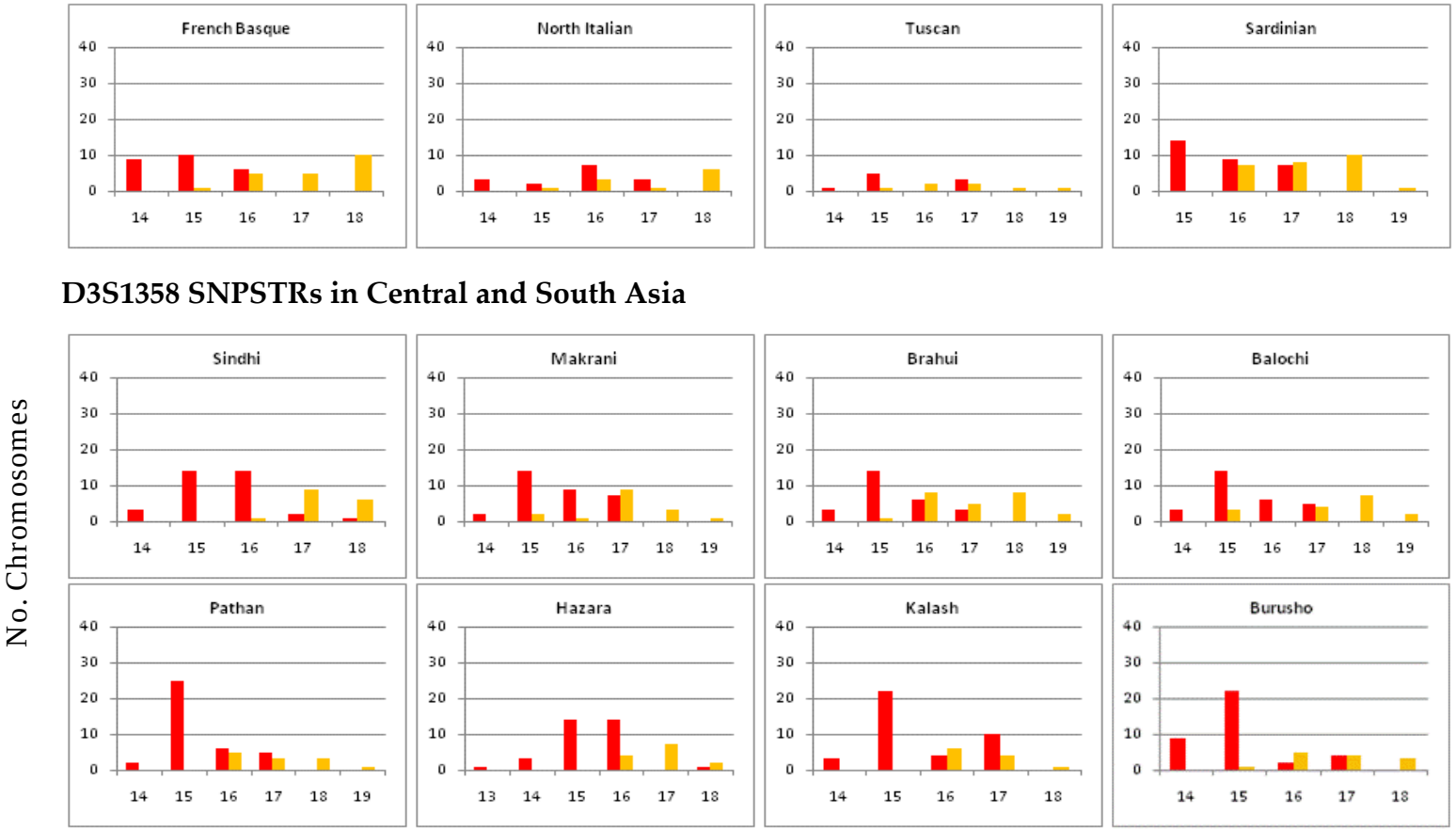


D3S1358 SNPSTRs in Europe

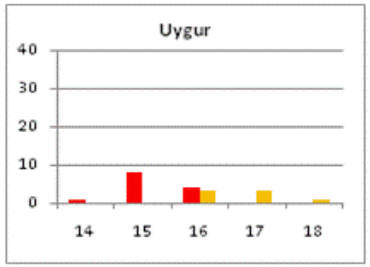


No. Chromosomes

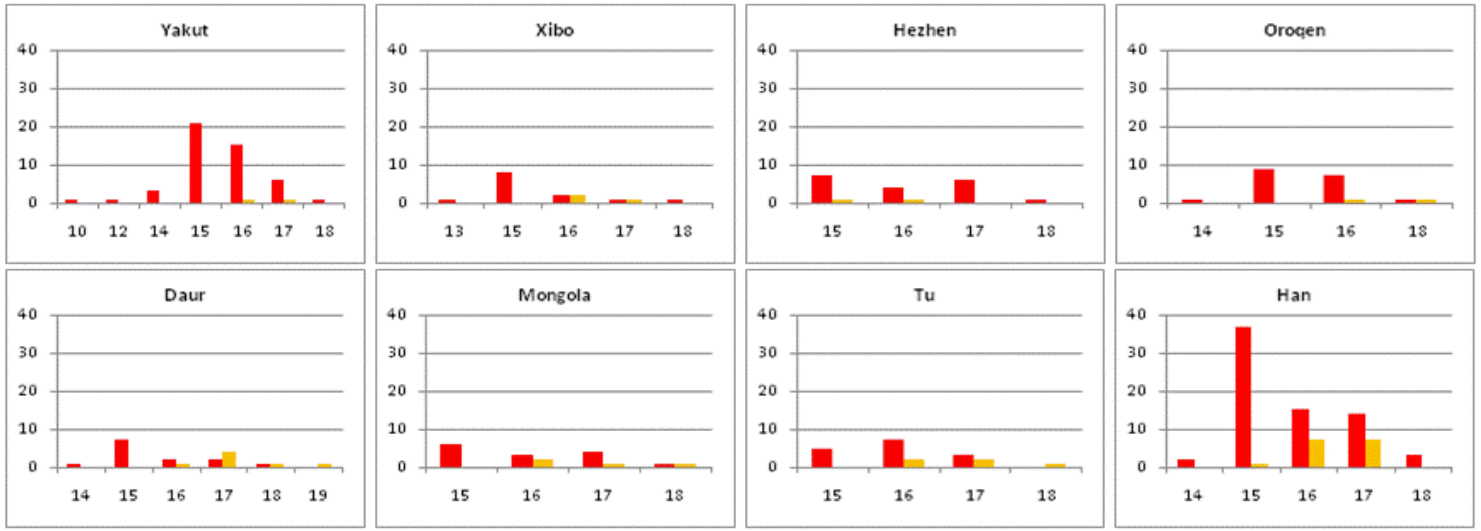
STR Length/ repeats

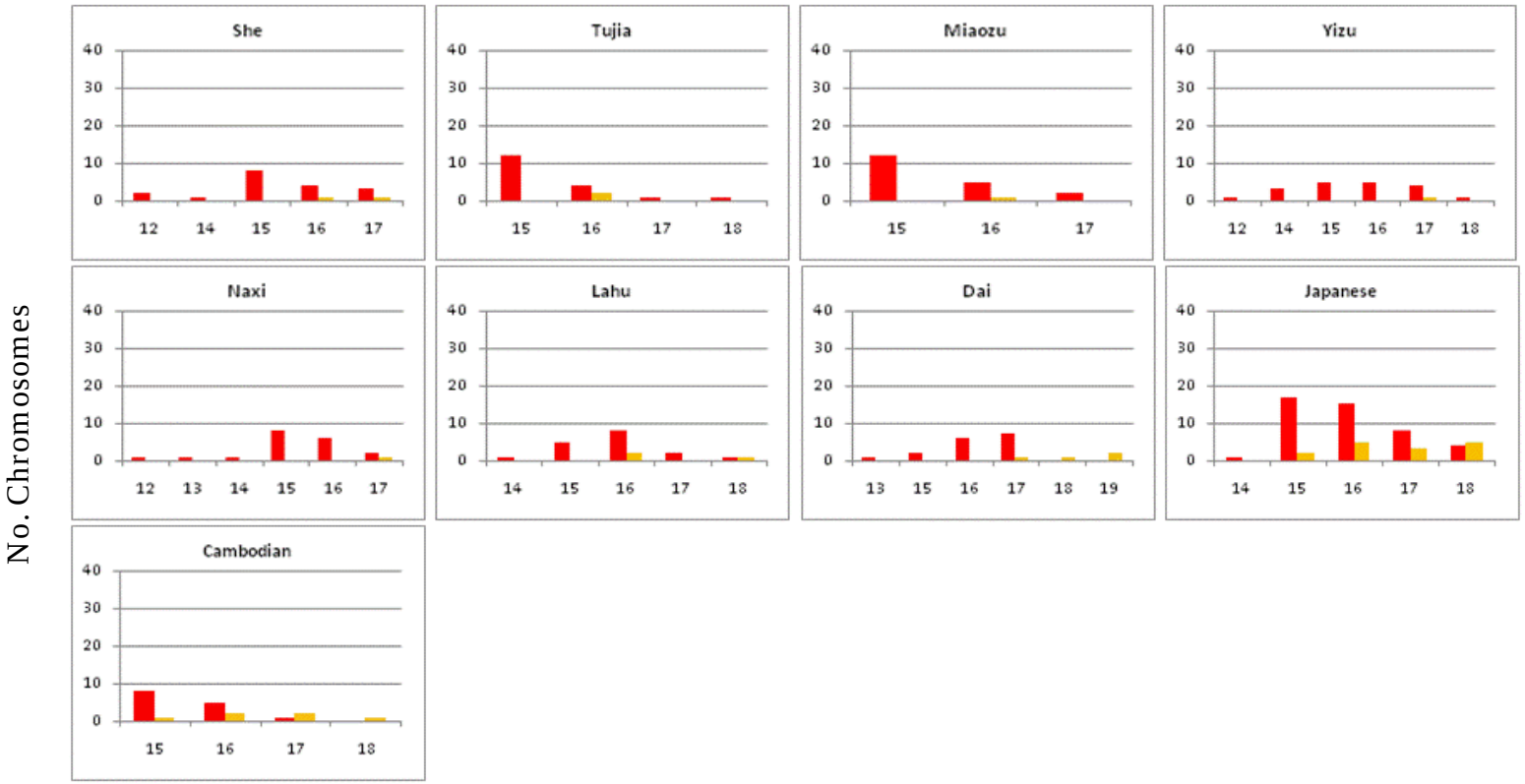


D3S1358 SNPSTRs in Central and South Asia

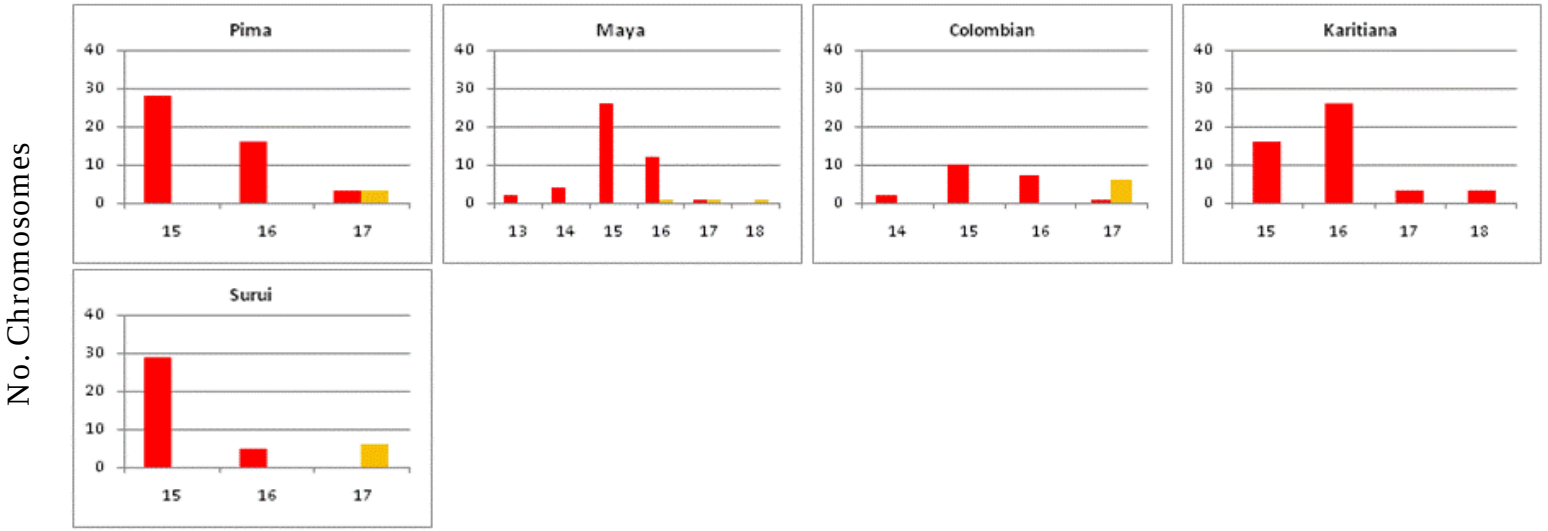


D3S1358 SNPSTRs in South East Asia



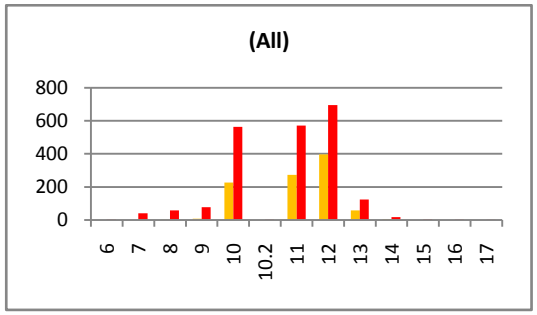

No. Chromosomes
She
Tujia
Miaozu
Yizu
Naxi
Lahu
Dai
Japanese
Cambodian
STR Length/ repeats

D3S1358 SNPSTRs in America



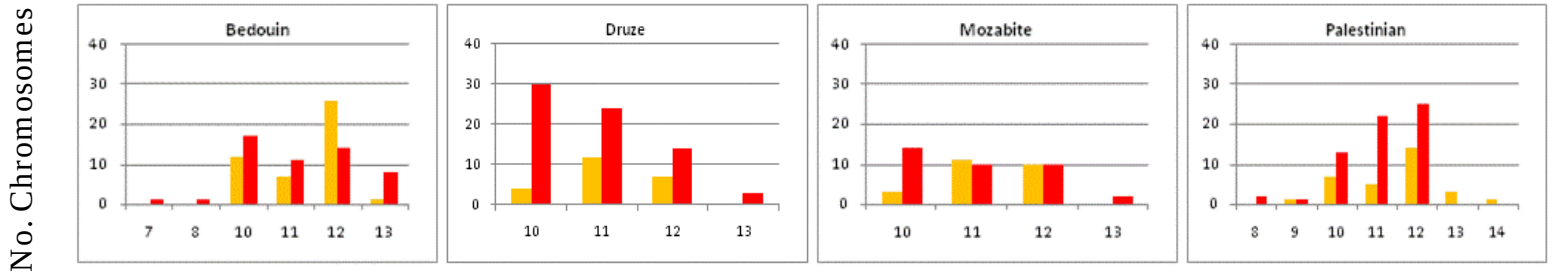
Section D - CSF1PO SNPSTRs in all populations typed

C → A



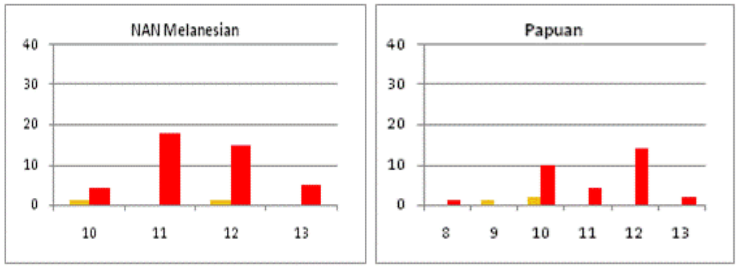
From this chart, it is possible to see that the full range of STR repeats in all populations typed is between 6 and 17 repeats. The smaller charts below do not show the full range possible.

CSF1PO SNPSTRs in the Middle East

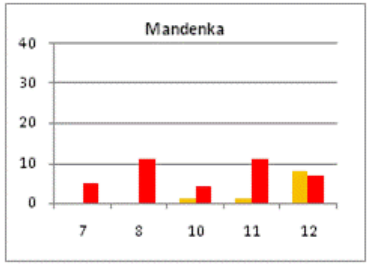


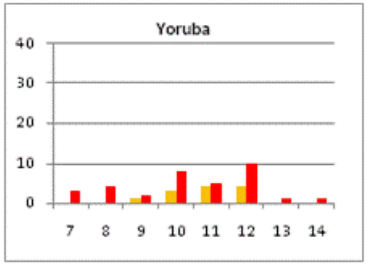
STR Length/ repeats

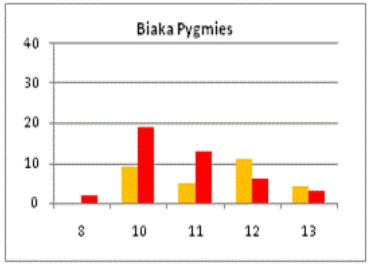
CSF1PO SNPSTRs in Oceania

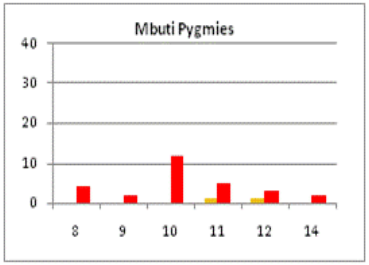


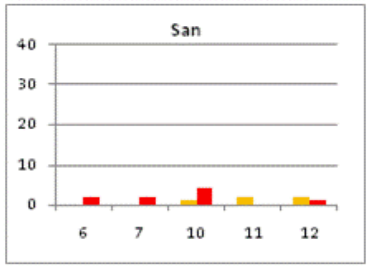
CSF1PO SNPSTRs in Africa

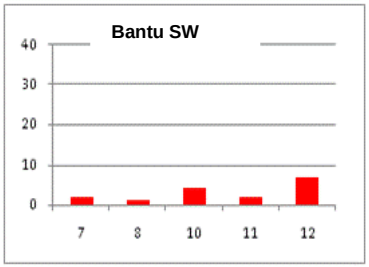


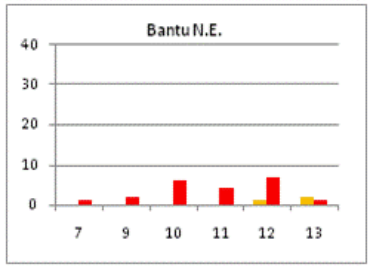








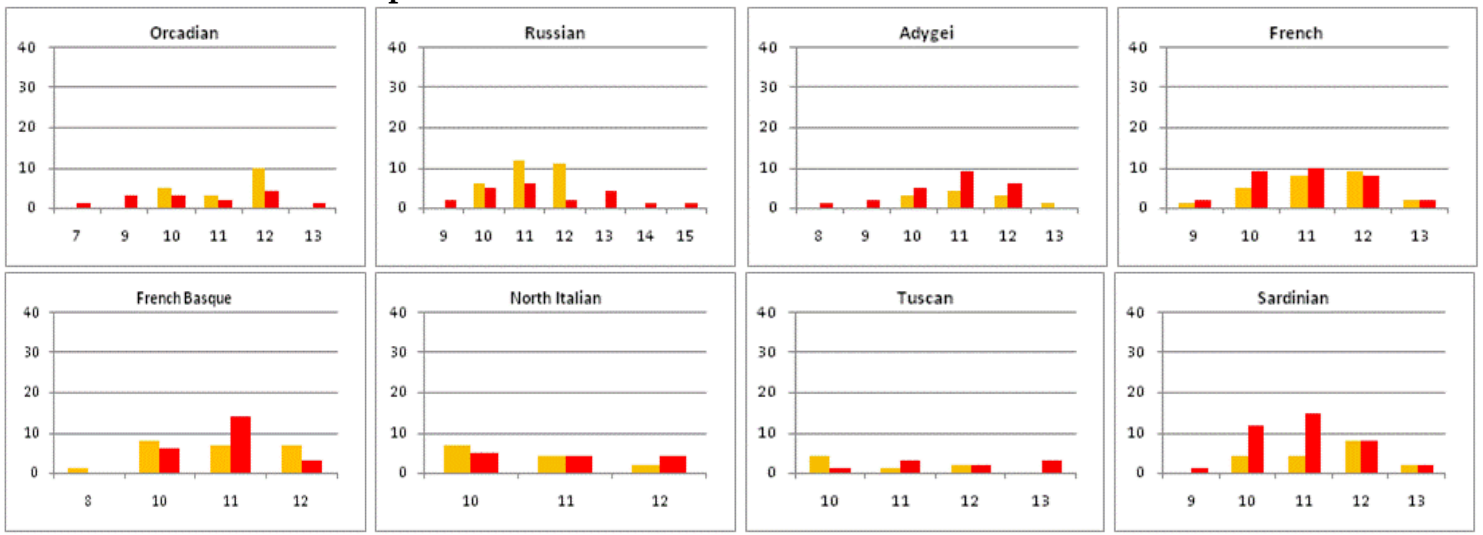




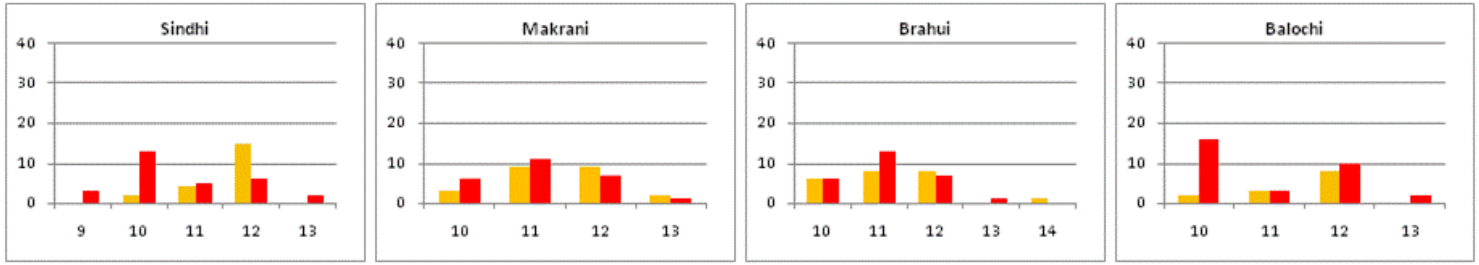
No. Chromosomes

STR Length/ repeats

CSF1PO SNPSTRs in Europe

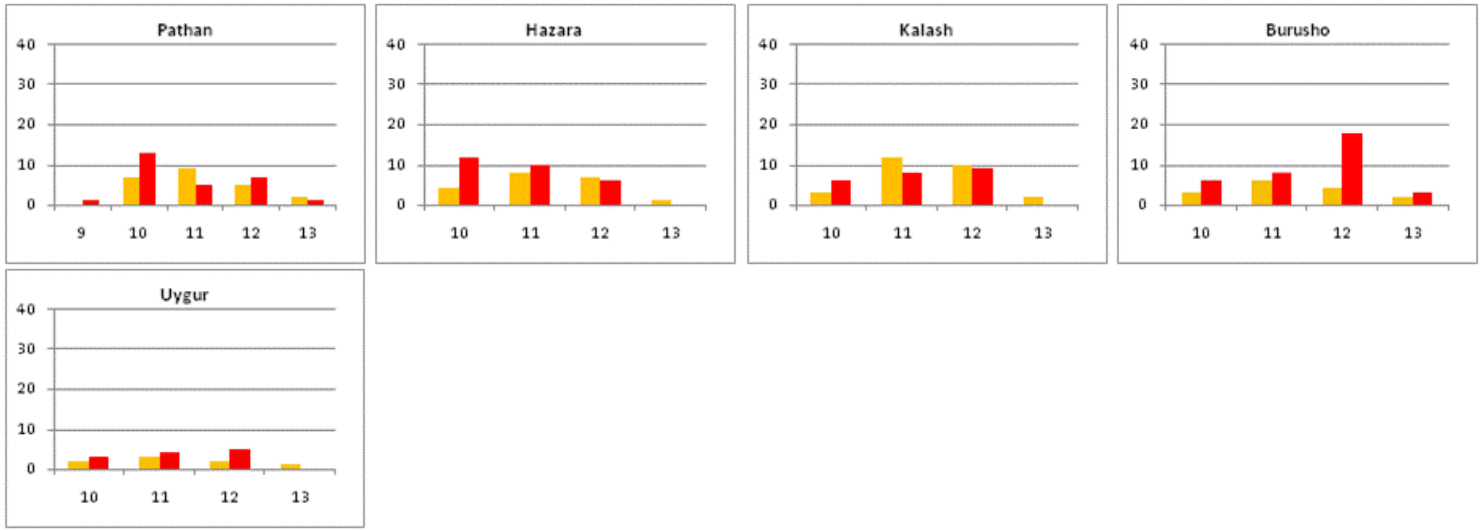


CSF1PO SNPSTRs in Central and South Asia

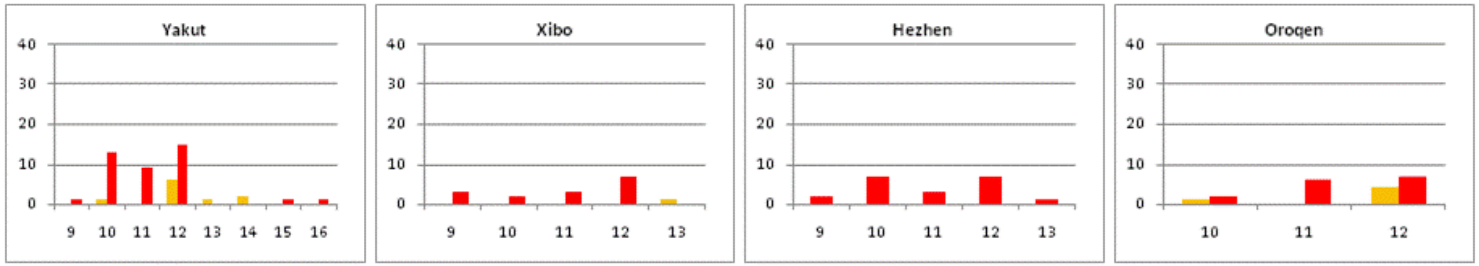


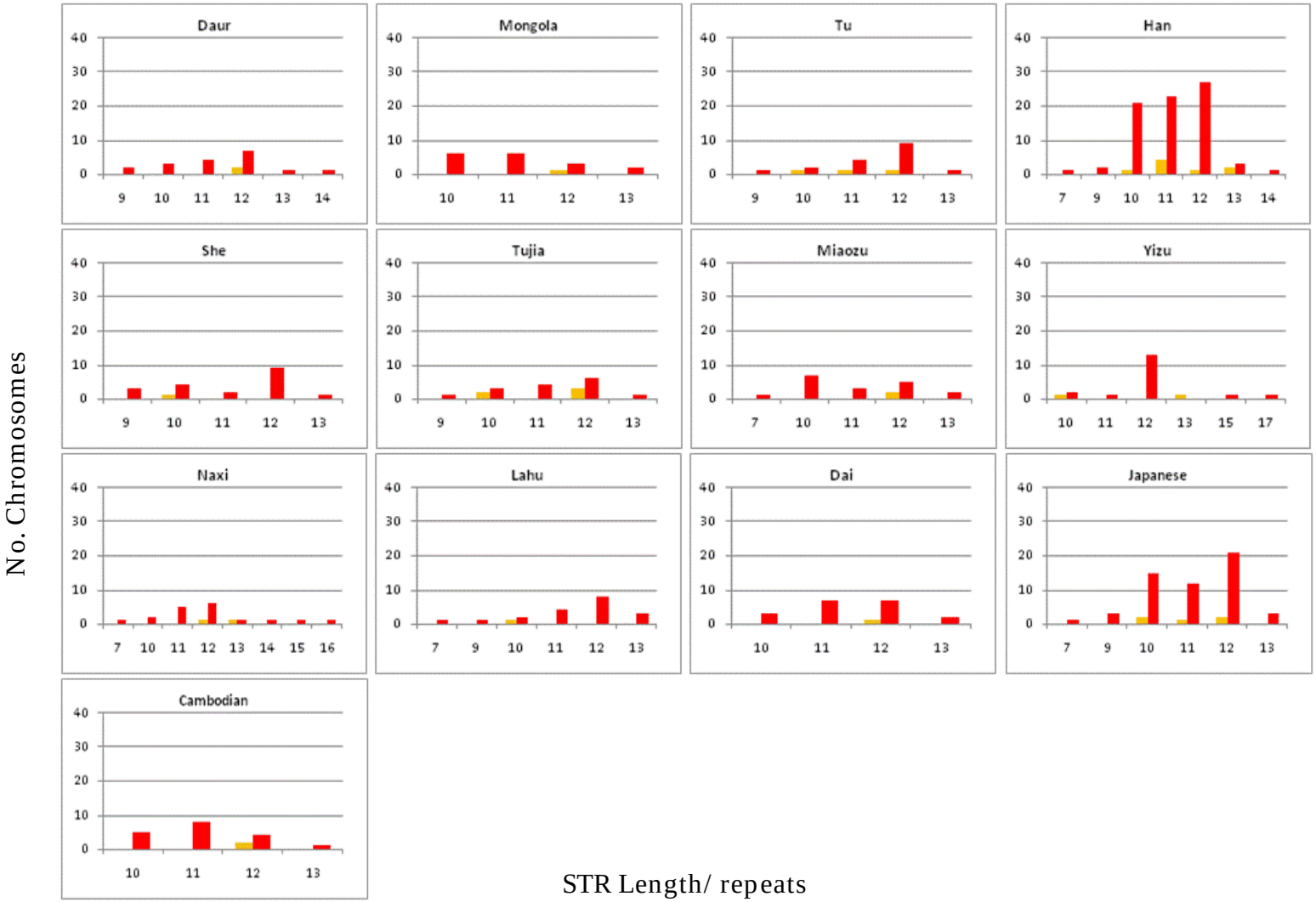
No. Chromosomes

STR Length/ repeats



CSF1PO SNPSTRs in South East Asia




Daur
Mongola
Tu
Han
She
Tujia
Miaozu
Yizu
Naxi
Lahu
Dai
Japanese
Cambodian
No. Chromosomes
STR Length/ repeats

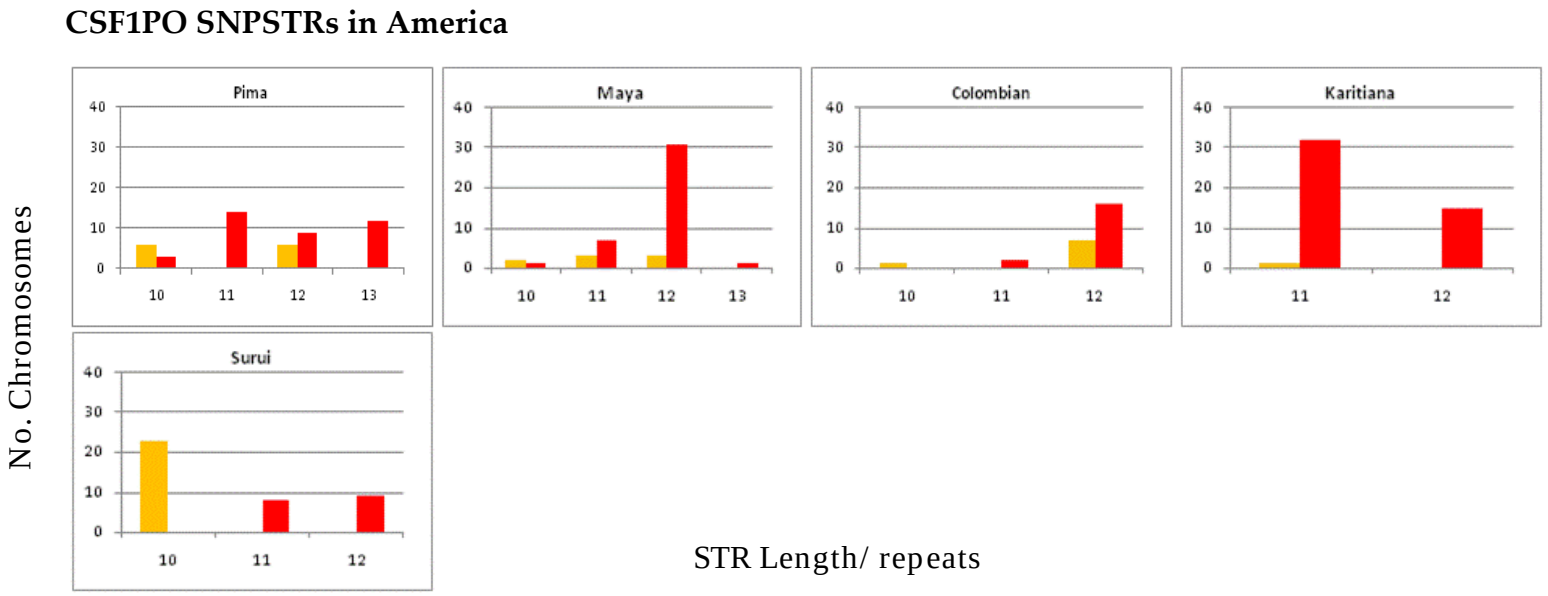


Figure 4-13. Charts showing SNPSTR distributions in global population samples from the CEPH-HGDP. Section A shows D5S818 SNPSTRs, Section B shows D16S539 SNPSTRs. Section C shows D3S1358 SNPSTRs. Section D shows CSF1PO SNPSTRs. In each case, the red bars indicate the STR lengths based on the ancestral state of the SNP, and the orange bars indicate the STR lengths based on the derived state of the SNPs.

4.4.6 Cornish SNPSTRs

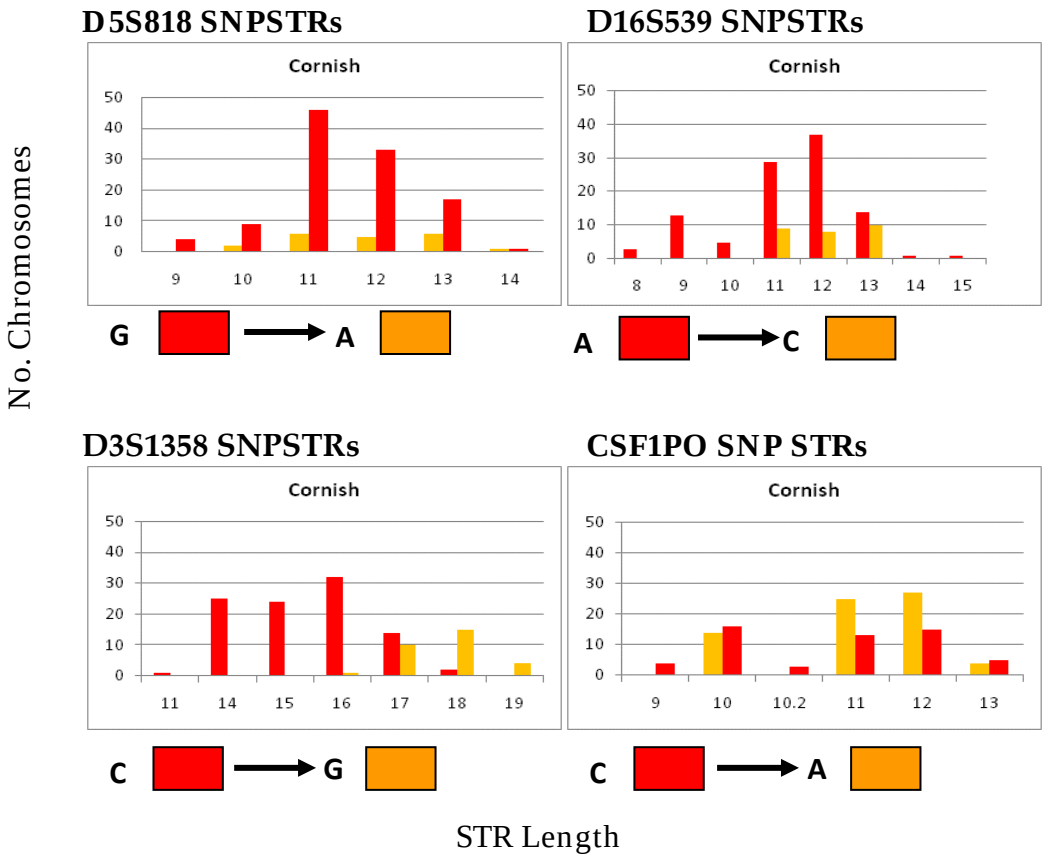
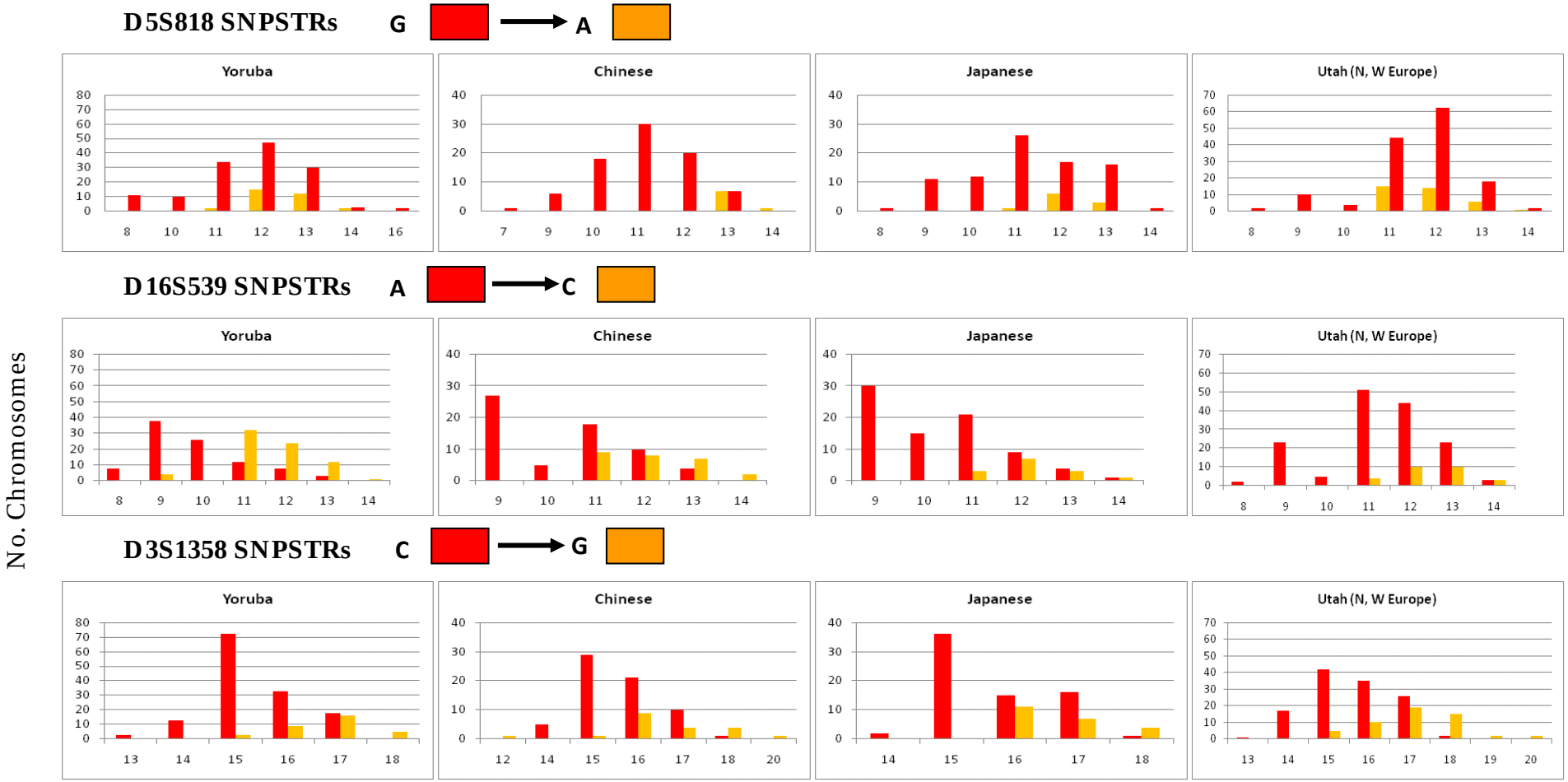


Figure 4-14. Charts showing SNPSTR distributions in the Cornish population samples. In each case, the red bars indicate the STR lengths based on the ancestral state of the SNP, and the orange bars indicate the STR lengths based on the derived state of the SNPs.

4.4.7 HapMap Population SNPSTRs



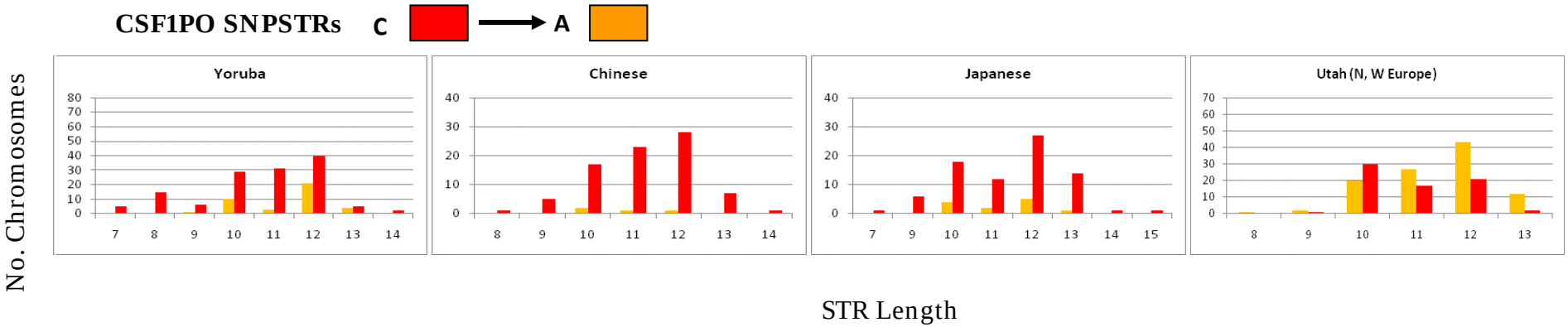


Figure 4-15. Charts showing SNPSTR distributions in global population samples from the HapMap DNA samples. In each case, the red bars indicate the STR lengths based on the ancestral state of the SNP, and the orange bars indicate the STR lengths based on the derived state of the SNPs.

Following on from these simple plots which show SNPSTR variation, both F_{ST} and R_{ST} values were calculated for all of the SNPSTRs in all of the population samples and MDS plots were created to see if there were any particular groupings of populations which stood out for each of the four SNPSTRs typed. These can be seen in the figures below.

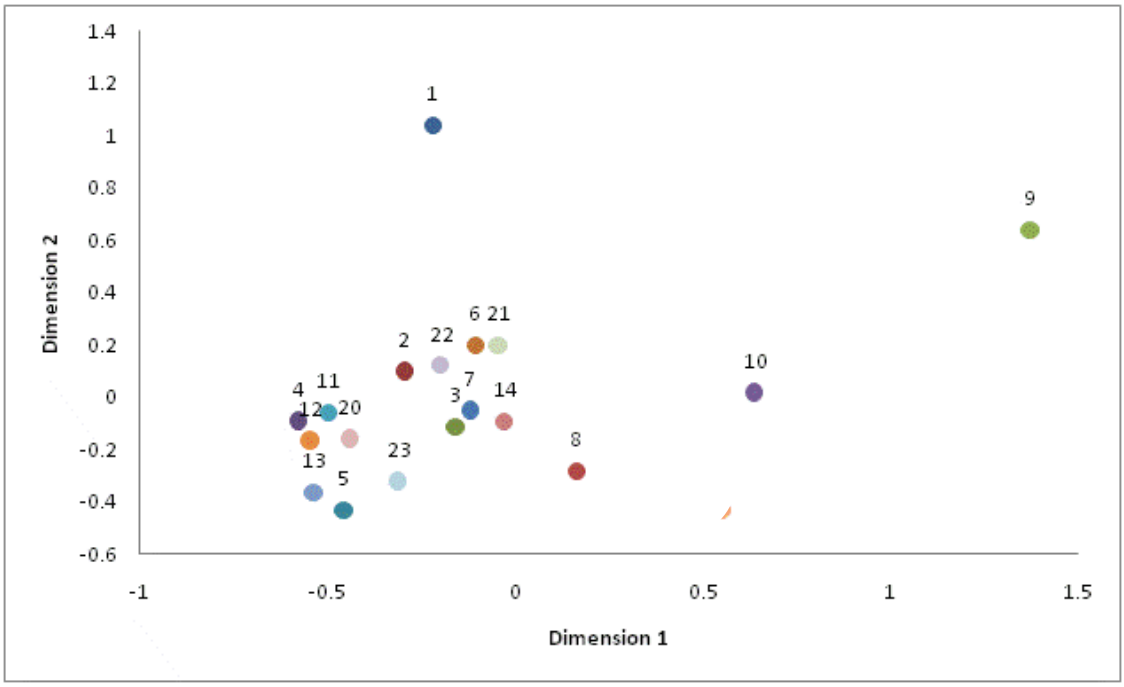


Figure 4-16. F_{ST} MDS plot for D5S818 SNPSTR showing all populations typed. The numbers represent the following populations: 1= Oceania, 2= Middle East, 3=Southern Europe, 4=Northern Europe, 5=Eastern Europe, 6=South East Asia, 7=South Asia, 8=North East Asia, 9=South America, 10=North America, 11=Sub Saharan Africa, 12=North Africa, 13=African Caribbean, 14=Cornwall, 20=Yoruba, 21=Chinese, 22=Japan, 23= Utah (Western European).

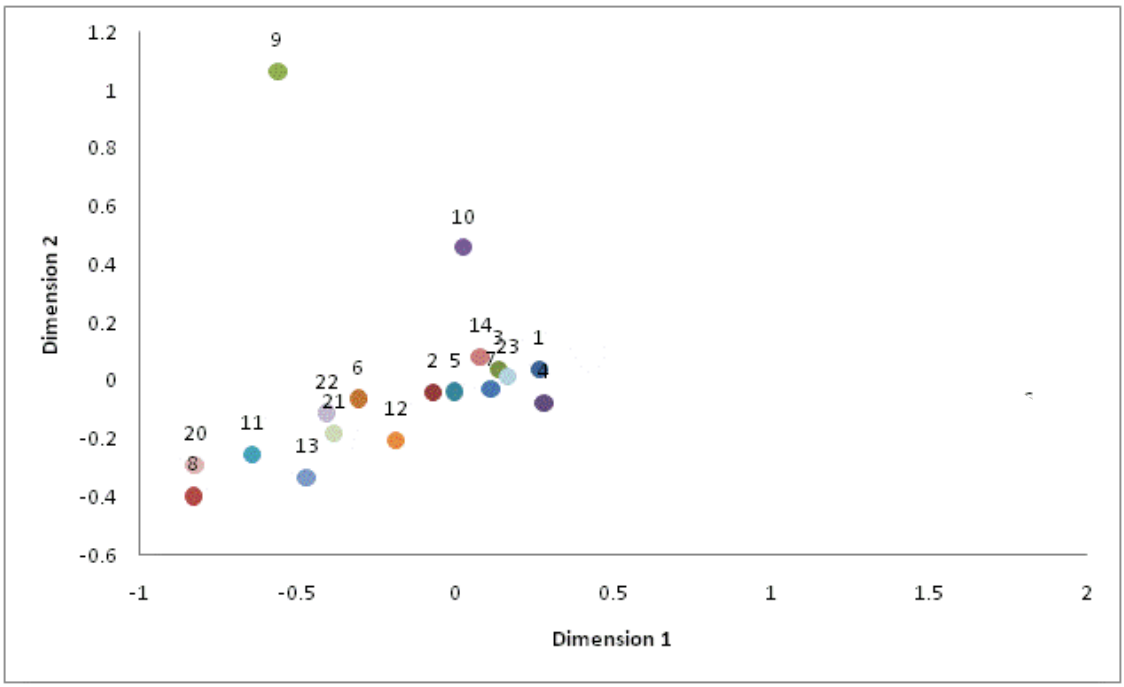


Figure 4-17. F_{ST} MDS plot for D16S539 SNPSTR showing all populations typed. The numbers represent the following populations: 1= Oceania, 2= Middle East, 3=Southern Europe, 4=Northern Europe, 5=Eastern Europe, 6=South East Asia, 7=South Asia, 8=North East Asia, 9=South America, 10=North America, 11=Sub Saharan Africa, 12=North Africa, 13=African Caribbean, 14=Cornwall, 20=Yoruba, 21=Chinese, 22=Japan, 23= Utah (Western European).

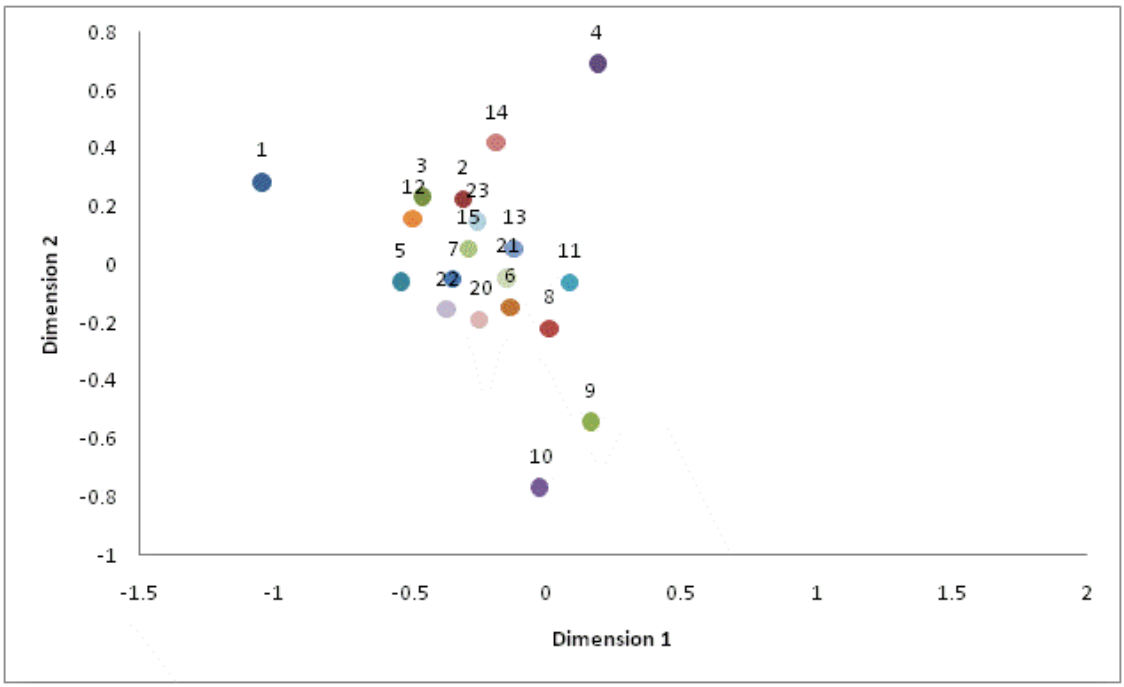


Figure 4-18. F_{ST} MDS plot for D3S1358 SNPSTR showing all populations typed. The numbers represent the following populations: 1= Oceania, 2= Middle East, 3=Southern Europe, 4=Northern Europe, 5=Eastern Europe, 6=South East Asia, 7=South Asia, 8=North East Asia, 9=South America, 10=North America, 11=Sub Saharan Africa, 12=North Africa, 13=African Caribbean, 14=Cornwall, 15=Denmark, 20=Yoruba, 21=Chinese, 22=Japan, 23= Utah (Western European).

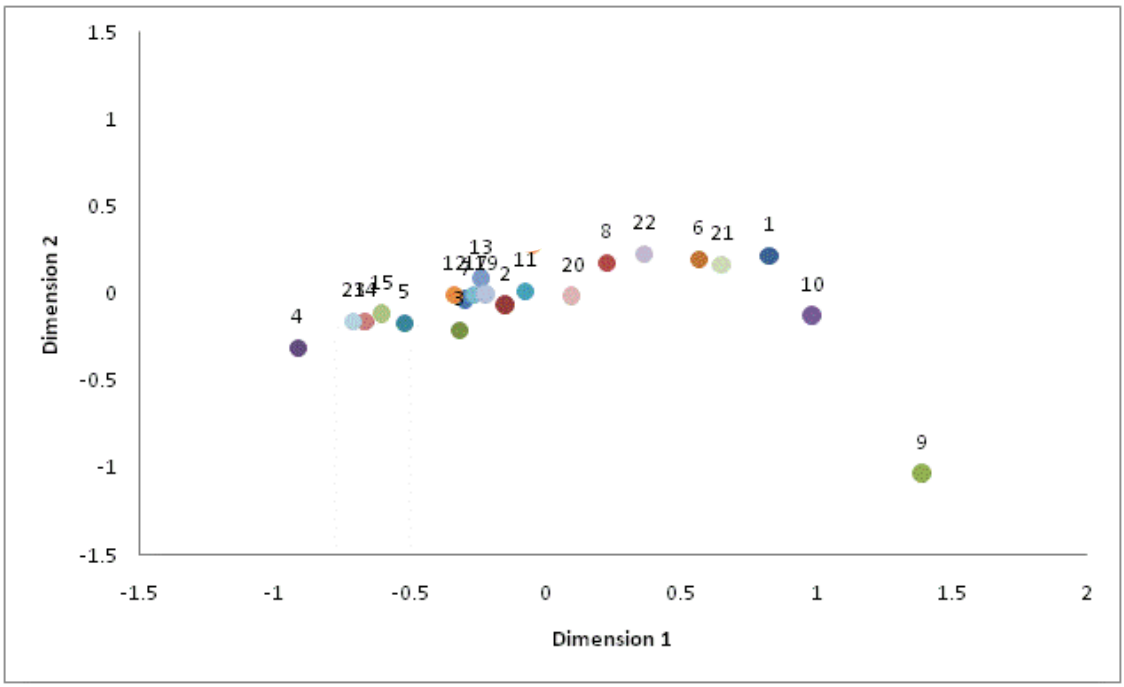


Figure 4-19. F_{ST} MDS plot for CSF1PO SNPSTR showing all populations typed. The numbers represent the following populations: 1= Oceania, 2= Middle East, 3=Southern Europe, 4=Northern Europe, 5=Eastern Europe, 6=South East Asia, 7=South Asia, 8=North East Asia, 9=South America, 10=North America, 11=Sub Saharan Africa, 12=North Africa, 13=African Caribbean, 14=Cornwall, 15=Denmark, 20=Yoruba, 21=Chinese, 22=Japan, 23=Utah (Western European).

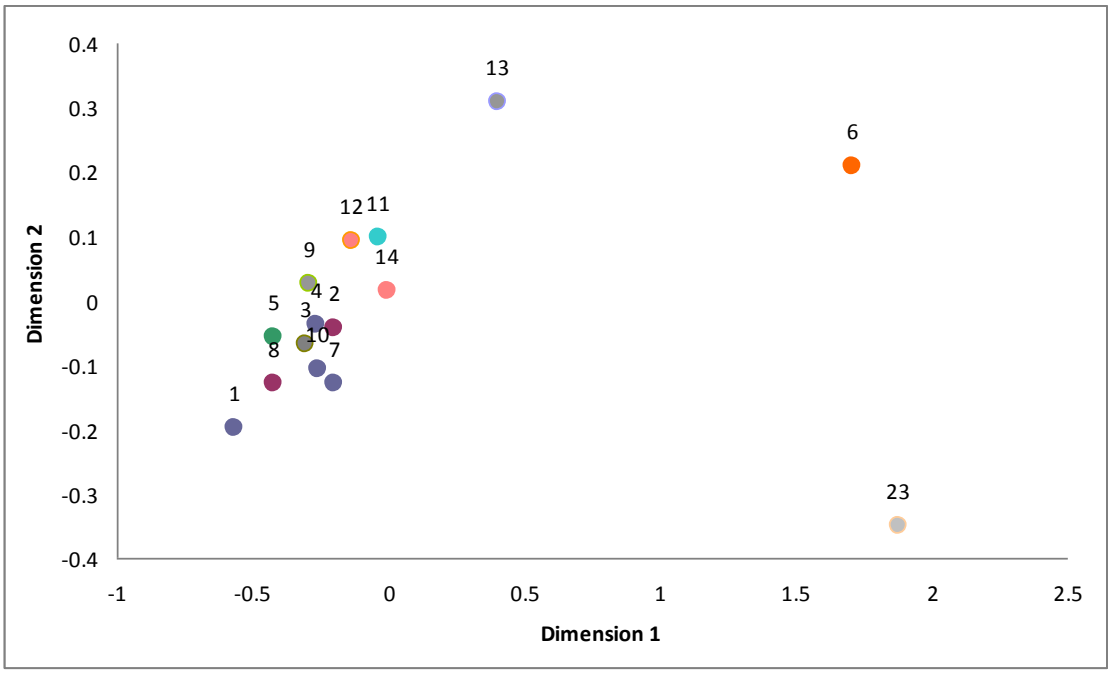


Figure 4-20. R_{ST} MDS plot for D5S818 SNPSTR showing all populations typed. The numbers represent the following populations: 1= North Africa, 2= Sub Saharan Africa, 3=North America, 4=South America, 5=North East Asia, 6=South Asia, 7=South East Asia, 8=Eastern Europe, 9=Northern Europe, 10=Southern Europe, 11=Middle East, 12=Oceania, 13=Cornwall, 14=Denmark, 23= African-Caribbean.

In the R_{ST} plot for the D5 SNPSTR above, the HapMap samples were omitted since the populations they contain are well covered by HGDP-CEPH (He et al. 2009a).

Population gene diversity was also calculated for each SNPSTR using Arlequin (Schneider et al. 2000a) and the results are shown in the graph below.

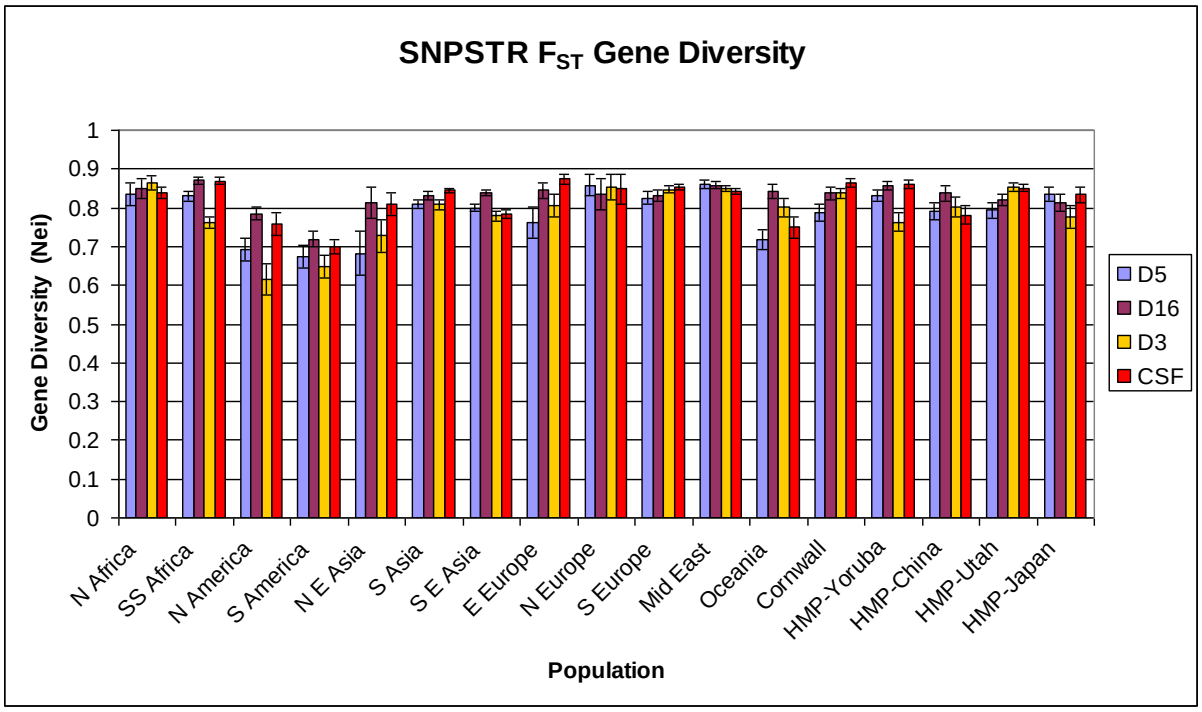


Figure 4-21. Nei's gene diversity calculated using Arlequin (Schneider et al. 2000a) for all 4 SNPSTRs, using the F_{ST} results rather than the molecular R_{ST} results. The error bars represent the error for Nei's gene diversity as calculated by Arlequin. This graph includes all the DNAs from the CEPH-HGDP, HapMap and Cornwall.

4.5 Discussion

Generally, from the charts in Figure 4-13 it can be seen that the STR allele ranges for the minor and major alleles always do overlap, for those populations where one or other allele is not fixed. There is one exception where this is not the case, and this is for the Surui, where for the D3 SNPSTR, the minor allele range starts after the major allele range, and also for the same population in the CSF1PO SNPSTR, where the minor allele occurs before the major allele range starts.

Overall, the minor SNP allele range is associated with a narrower range of STR allele length than the major allele, but there are quite a few exceptions to this, and these can be seen in the table below.

Population	SNPSTR allele	Observation
Adygei	D3	Minor allele range is wider than the major allele range
Balochi	D3	Minor allele range is wider than the major allele range
Bantu NE	D16	Minor and major allele ranges are equal
Brahui	CSF	Minor allele range is wider than the major allele range
Brahui	D3	Minor allele range is wider than the major allele range
Burusho	CSF	Minor and major allele ranges are equal
Cambodian	D3	Minor allele range is wider than the major allele range
Druze	D3	Minor allele range is wider than the major allele range
French	CSF	Minor and major allele ranges are equal
French	D3	Minor allele range is wider than the major allele range
French Basque	CSF	Minor allele range is wider than the major allele range
French Basque	D3	Minor allele range is wider than the major allele range
Hapmap CEU	CSF	Minor and major allele ranges are equal
Hapmap CEU	D3	Minor and major allele ranges are equal
Hapmap CHB	D3	Minor and major allele ranges are equal
Hazara	CSF	Minor allele range is wider than the major allele range
Kalash	CSF	Minor allele range is wider than the major allele range
Makrani	CSF	Minor and major allele ranges are equal
Makrani	D3	Minor allele range is wider than the major allele range
North Italian	CSF	Minor and major allele ranges are equal
Orcadian	D16	Minor allele range is wider than the major allele range
Palistinian	CSF	Minor allele range is wider than the major allele range
Palistinian	D3	Minor allele range is wider than the major allele range
Sardianian	D3	Minor allele range is wider than the major allele range
Tuscan	D3	Minor allele range is wider than the major allele range
Uygur	CSF	Minor allele range is wider than the major allele range

Table 4-9. Observations from Figure 4-13 where the minor SNP allele is associated with an equal or wider range of STR allele length variation than the major allele.

There are also quite a few populations, where the major allele is fixed, for example in many of the East Asian populations.

In addition to this, some of the smaller STR alleles can be seen to be "frozen". Some examples of this can be seen in the Yakut and Mayan populations for the D5 SNPSTR (Fig 4-13, Section A), the Cambodian and N.E. Bantu populations for the D3 SNPSTR, and the Nan Melanesian and Biaka Pygmy populations for the CSF1PO SNPSTR.

From Figure 4-13, Section A, (D5S818 SNPSTR) it can also be seen that Oceania only has the ancestral SNP and that this is mainly associated with STR allele repeats of 10 or 11, which is most likely the ancestral STR repeat number. This suggests that the Oceania D5S818 SNPSTR came from a few founder individuals and either there has not been sufficient time for a mutation to occur at the SNP, suggesting that this is a more recent arrival, or that if there was initially a derived SNP, that this has been lost, perhaps through drift. As Australia and the surrounding islands were inhabited some 50,000 years ago (Jobling et al. 2003), which is just before modern humans appeared in Europe, it would suggest that drift was the more likely cause of the loss of the derived SNP in these populations. Of course, it could also simply be a case that the sample size was too small and individuals carrying the derived SNP allele were not sampled.

The Americas also only have the ancestral SNP, and in addition the Surui only have two STR alleles (7 and 11). The Surui are a very isolated population and this probably accounts for the lack of diversity of the STR. The Americas were the last major continent to be peopled by modern humans at around 15,000 years ago, (Jobling et al. 2003). If the Americas were peopled by North Eastern Europeans who crossed over via a land bridge in the Bering Strait, then it may well be that that the derived SNP was not present in these populations in the first place. Relevant founding populations for the Americas were not analysed here, but certainly, the Russian DNAs analysed also did not contain any of the derived SNPs, and these only start to occur in the Southern Europeans (French and French Basque and North Italians) – see Figure 4-13 – section A. This would suggest that Northern European populations are more closely related to the founders of the Americas. The Y chromosome haplogroups in the Americas are predominantly Q and R, and certainly R is common in N. Europe and both Q and R are part of the same superhaplogroup P*, . Haplogroups Q and R are not found in the Southern Europeans. The results could suggest that the population of the Americas underwent a bottleneck after crossing the Bering Strait which caused the lack of diversity at the D5S818 SNPSTR.

East Asian populations also lack the derived SNP (with the exception of the Japanese, Xibo, Mongolian and Han). However, the Y chromosome haplogroups and mitochondrial haplogroups (see Chapter 1, figures 1-2 and 1-3) are very

diverse for these populations. In this case, the SNPSTRs may be reflecting the sampling methods more than any real population structure, and certainly a recent paper indicated that the HapMap CHB samples are not suited to standard population genetic studies due to the way in which the samples were collected (He et al. 2009b).

Figure 4-13 – Section B shows the D16S539 SNPSTR charts. From these it can be seen that the Americas again only contain the ancestral SNP (with the exception of one individual in the Mayan population with the derived SNP), and both the Karitiana and Surui once again are seen to be population isolates with very low STR diversity. However, for the D16S539 SNPSTR, the Russian graph does not show the same as for the D5S818 SNPSTR, as there are derived SNP alleles present in the Russian DNAs. This suggests that the Americas may indeed have undergone a bottleneck rather than drift being the case of the reduced SNPSTR diversity seen here.

The D3S1358 SNPSTR pattern shows much more diversity in the European and Central and South Asian DNAs than there is in the African DNAs, where the San and Bantu mostly carry the ancestral SNP. This is somewhat reflected in the mtDNA where the African haplogroups consist of mostly L1, L2 and L3, which are the deeper rooting branches (Chapter 1, Figure 1-3). These mtDNA haplogroups are absent in Central and South Asia, and where there is much more mtDNA

diversity. There is a similar pattern for the Y chromosome too (Chapter 1, Figure 1-2). Anatomically modern humans first appeared in Africa around 130,000 years ago, before they spread across the globe. The D3S1358 SNPSTR is less diverse in Africa and this reflects the data seen in the mtDNA and Y chromosome.

The D3S1358 SNPSTR also highlights the population isolate of the Surui, where there is only one derived SNP allele which is only associated with a 17-repeat STR allele. This is the same for the Pima and the Colombian DNAs. The Karitiana display features of an isolate in that they do not have any derived SNPs.

A similar picture is seen with the CSF1PO SNPSTR, where there is also less diversity in the Bantu populations of Africa than in the rest of the African samples and in Europe. Additionally, the CSF1PO SNPSTR also highlights the Karitiana and the Surui as population isolates.

The Cornish samples in
Figure 4-14 are very similar to the European DNAs which means that they can effectively be included as part of the European DNA samples in the CEPH-HGDP. The results from the F_{ST} analysis carried out using Arlequin (Schneider et al. 2000a) show that they are statistically not significantly different from the Northern European DNAs, with p-values ranging between 0 and 0.0186.

The Utah HapMap samples in

Figure **4-15** also largely resemble the data from the European CEPH-HGDP and Cornish samples, which shows that the SNPSTRs are effective for identifying population structure which is similar. Again, the Arlequin F_{ST} values for comparisons of the Utah HapMap samples with the Cornish and N. European samples are not significantly different.

The one figure which stands out, however, is the low diversity of derived SNPs seen in the D5S818 SNPSTR in the HapMap Chinese sample. A similar picture is seen in the CEPH-HGDP data Figure 4-13 – Section A.

From the SNPSTR charts alone, already some population structure can be seen and some theories about the spread of human populations can be made. This would suggest that these SNPSTRs are useful for identifying basic population structures. However, more SNPSTR markers are needed to gain a fuller picture as well as further analysis of the date from the four SNPSTRs analysed in this case. Further data analysis would include using software such as STRUCTURE (Pritchard et al. 2000) or SIMCOAL (Excoffier et al. 2000).

From Figure 4-16, which shows the F_{ST} data for the D5S818 SNPSTR, the main outliers are Oceania (1), S. America (9) and N. America (10). The data seen here reflects that seen in the graphs and fits with the assumptions made above about the peopling of Oceania and the Americas. Figure 4-17 also shows S. America and N.

America as outliers and the rest of the populations grouping together. Figure 4-18 showing the F_{ST} data for the D3S1358 SNPSTR shows N. America and S. America grouping together, and Oceania and N. Europe are also seen to be outliers here. This Northern Europe outlier cannot be seen in the SNPSTR graphs (Figure 4-13 – Section C), nor can it be seen in Figure 4-19, which does not show the N. European population as being any less diverse than the S. European populations. However, as MDS plots are in essence two-dimensional representations of a multi-dimensional image, it is often difficult to ascertain using these plots alone whether a particular population is really a true outlier or whether it is merely an artefact of the MDS plot.

The MDS plot for the CSF1PO F_{ST} data again shows S. America as an outlier and the rest of the populations group together. This also reflects the data seen in the graphs above.

The R_{ST} data produced slightly different results, showing the Cornish (13) and South Asian (6) and African Caribbean (23) populations as outliers. The result is most likely different because of the fact that Arlequin (Schneider et al. 2000a) is not designed to handle SNPSTR data and that the SNP was treated as an STR and given an artificial weighting of 1 repeat for the ancestral state and 99 repeats for the derived state. The rationale for this is described more fully in Chapter 5.

The SNPSTR gene diversity graph, Figure 4-21, shows North and South America as being less diverse than the majority of the other populations tested. This is certainly seen in the data from the SNPs shown in, Figure 4-5, Figure 4-7, and Figure 4-9, as well as in all of the SNPSTR graphs in Figure 4-13 (Sections A-D).

The data from this initial SNPSTR analysis would seem to suggest that SNPSTRs are indeed useful as markers for population structures and also to some extent, for population histories, if used in conjunction with other markers, such as the Y chromosome or mtDNA.

In the next chapter, the issue of using SNPSTRs as possible makers for assessing the structures of admixed populations, and it is also assed whether or not they are informative if used in this context. The DNAs used for this include the Greenland Inuit, Danish and African Caribbean DNAs.

5 Assessing the Structure of Admixed Populations using SNPSTRs

In this chapter the use of SNPSTRs as possible makers for assessing the structures of admixed populations is discussed and it is asked whether or not they are informative if used in this context. The introduction covers population admixture and asks why it is of interest to us today. This is then followed by background information on the Greenland Inuit people and an analysis of the DNA samples from this population. Finally, there follows background information on the African-Caribbean people and an analysis of the DNA samples from this population.

5.1 Introduction: What is population admixture and why is it of interest?

Populations are not static entities: they move and change over time because they are made up of lots of separate individuals. Because of this, two or more ancestral populations of individuals can mix and generate hybrid populations. Another type of admixture is when a new hybrid population is formed due to the mixing of individuals who would not normally mix – for example, during forced mass migrations (such as the African, Egyptian and Greek slave trades) or if one population moves into the territory of another, for example during wars. These last two events usually occur or start at a specific point in time and therefore the population admixture can be seen as a historical event or process within a particular time-frame.

Many of the populations that make up the world today can be considered in terms of admixture. However, this is much easier said than done because of the difficulty of identifying appropriate parental populations. In addition, not only do we detect the possible proportions of genetic contribution of each parental population at the start of the admixing, but we will also see all the accumulated changes that have happened to the admixed population over time. Finally, the parental populations are necessarily modern proxies for past populations which themselves have undergone changes since the admixture events. It may therefore prove to be impossible to differentiate between the consequences of admixture and gene flow, genetic drift, selection and mutation.

In addition to this, over time, barriers to population mixing have come and gone. For example, the last ice age which ended some 14 KYA provided a barrier to populations mixing in Europe, but after this the more northern areas of land were slowly re-populated by inhabitants from refugia in more southerly regions (Jobling et al. 2003). This saw the mixing of populations which had previously been separated by ice to the north, or the Mediterranean between refugial peninsulas to the south.

On a more temporary basis, there are also political barriers to mixing – such as the Berlin Wall, and currently the separation between North and South Korea.

Cultural and linguistic barriers also separate populations, but these are not as static as geographical barriers such as mountain ranges, seas and oceans.

Some of the earliest molecular work on genetic admixture involved the comparison of blood groups between African Americans and European Americans and Africans (Glass and Li 1953) and skin colour was used as the phenotypic evidence of admixture.

Today, we use markers at the DNA level rather than at the phenotypic level to study admixture; however, our ability to do so depends on how differentiated the source populations were from one another in the first place. If populations are very similar, it is much more difficult to detect signals of admixture. In these cases, often historical, linguistic and archaeological records can help to build up a fuller picture of the past.

Another factor in determining how easy it is to detect admixture is the choice of markers which are used. In population genetics, markers which are thought to be selectively neutral are generally chosen because they will reflect population processes better than ones which may have been subject to selection due to different phenotypic or metabolic effects. However, it is not possible to ignore the fact that disease prevalences are sometimes different between ancestral populations. When admixture occurs between two populations with different

disease prevalences, then the hybrid population would display prevalences that are intermediate between the two ancestral populations, for Mendelian disorders at least. This then enables the correlation between the degree of admixture and disease susceptibility in one or other of the ancestral populations. Certainly much of today's research into population admixture has a bio-medical application where research into projects related to health and genetic epidemiology are frequently carried out. One example is the 2007 study carried out to investigate admixture of Hispanic populations in the USA and how this may bias estimates of associations between candidate cancer susceptibility genes and breast cancer (Sweeney et al. 2007). It is important to understand the population structure because if the structure is not properly taken account, spurious associations could be made when disease genes are being sought and also it could lead to unreliable match probabilities in a forensic context (Parra et al. 1998; Pfaff et al. 2001).

In addition to the different methods for detecting admixture, there are also many different types of admixed populations. Some examples of these are outlined below.

5.1.1 Sex-biased admixture

This is admixture where there are unequal contributions of the different sexes from an ancestral population. It can readily be detected using uniparentally inherited markers. One example of a male-biased admixture is in the Greenland Inuit population of today, where the Y chromosomes are predominantly (~60%) of

European origin (Bosch et al. 2003) and the mitochondrial DNA showed a complete absence of any European mtDNAs, all falling into Native American lineages (Saillard et al. 2000).

5.1.2 Transnational isolates

These are populations which are isolated by means of geography, culture, history or language and as a result have very little contact with surrounding populations, yet have a widespread geographical distribution. Because of their isolation, they often have unusually high frequencies of normally rare genetic diseases. Transnational isolates may be subdivided into smaller grouping even with a large isolate, such as occurs in the European Roma. These isolates sometimes make it very difficult to perform quantitative admixture analyses because it sometimes is not possible to ascertain parental populations. This is certainly the case for the European Roma (Jobling et al. 2003).

5.2 Materials and methods

The SNPSTRs were typed on Greenland Inuit, Danish, and Northern European DNAs, as well as on African Caribbean DNAs collected in the UK as part of a project associated with a television programme called “Motherland”, and they were compared to DNAs from Yoruba and Europe from the HGDP-CEPH and HapMap panels. The Greenland Inuit DNA samples were kindly supplied by Søren Nørby of the University of Copenhagen.

SNPSTR typing methods and materials are all as described in Chapter 2 (Materials and Methods).

The SNPs were also tested to ascertain whether they were in Hardy-Weinberg equilibrium (as described in Chapter 4). As the populations are admixed, it could have an impact on the results, depending on time since the admixture event, and also depending on the pattern of admixture. The Greenland Inuit are known to be a male-biased admixed population and it is known that population admixture can lead to deviations from Hardy-Weinberg (Crow and Kimura 1970). The results of this are in the table below.

Pop	Locus	DF	ChiSq	Prob	Signif
Afro-Caribbean	rs2116791	1	1.133	0.287	ns
Afro-Caribbean	rs1728369	1	1.849	0.174	ns
Afro-Caribbean	rs17077990	1	2.507	0.113	ns
Afro-Caribbean	rs25768	1	0.301	0.583	ns
Denmark	rs2116791	Monomorphic			
Denmark	rs1728369	1	3.296	0.069	ns
Denmark	rs17077990	1	0.489	0.484	ns
Denmark	rs25768	1	0.032	0.858	ns
Greenland	rs2116791	Monomorphic			
Greenland	rs1728369	1	0.004	0.952	ns
Greenland	rs17077990	1	0.373	0.541	ns
Greenland	rs25768	1	0.036	0.850	ns
Key: ns=not significant, * P<0.05, ** P<0.01, *** P<0.001					
Bonferroni Correction	**P< 0.00416**				

Table 5-1. Chi-Square results for the SNPs from all DNA samples used which form part the SNPSTRs.

From these results it can be seen two of the SNPs were monomorphic in Denmark and Greenland Inuit, and hence it was not possible to carry out a Chi-Square test

for these. A full list of the results of applying this test to the distribution of the STRs in all of the DNA samples is listed in Table 5-2.

Pop	Locus	DF	ChiSq	Prob	Signif
Afro-Caribbean	D5S818	28	14.418	0.984	ns
Afro-Caribbean	D16S539	28	24.779	0.640	ns
Afro-Caribbean	D3S1358	28	25.369	0.608	ns
Afro-Caribbean	CSF1PO	28	14.010	0.987	ns
Denmark	D5S818	10	15.070	0.130	ns
Denmark	D16S539	15	14.956	0.455	ns
Denmark	D3S1358	15	11.249	0.735	ns
Denmark	CSF1PO	15	19.566	0.189	ns
Greenland	D5S818	21	18.061	0.645	ns
Greenland	D16S539	21	22.451	0.374	ns
Greenland	D3S1358	15	12.443	0.645	ns
Greenland	CSF1PO	15	9.398	0.856	ns
Key: ns=not significant, * P<0.05, ** P<0.01, *** P<0.001					
Bonferroni Correction	**P< 0.00416**				

Table 5-2. Chi Square results for the STRs which form part the SNPSTRs.

There are a number of different methods for detecting admixture and estimating admixture proportions of a population. These are able to take a lot of different factors, such as drift, into account, and give results in terms of admixture proportions. Table 5-3 lists some of the software available for estimating admixture. Unfortunately these methods deal either with SNPs or STRs, but not when both are used together as a marker. For this reason, formal population admixture assessment using SNPSTRs was not carried out, but instead population structure was analysed using the program STRUCTURE (Pritchard et al. 2000). This method allows individuals to be clustered on the basis of their genetic information and has previously been applied to genome-wide STR and SNP datasets from the HGDP-CEPH panel where clusters of individuals were identified

corresponding to specific geographical regions (Rosenberg et al. 2002). STRUCTURE is discussed in more detail later in this chapter.

Method	Sofware	URL	References
LEA	LEA	http:/ / www.rubic.rdg.ac.uk/ ~mab/ software. html	(Chikhi et al. 2001)
m_y (pair of ancestral populations) m_r	ADMIX1.0	http:/ / web.unife.it/ progetti/ genetica/ Giorgi o/ giorgio_soft.html	(Bertorelle and Excoffier 1998)
M_y (multiple ancestral populations)	ADMIX2.0	http:/ / web.unife.it/ progetti/ genetica/ Isabell e/ admix2_0.html	(Dupanloup and Bertorelle 2001)
Gene identity	ADMIX95	http:/ / www.genetica.fmed.edu.uy/ software. htm	
Bayesian individual admixture	ADMIXMAP	http:/ / homepages.ed.ac.uk/ pmckeigu/ admi xmap/ index.html	
Model-based clustering	STRUCTURE	http:/ / pritch.bsd.uchicago.edu/ structure.htm l	(Pritchard et al. 2000)

Table 5-3. List of some of the software which can be used for the estimation of population admixture. (Jobling et al. 2003)

Statistical analysis of the data was carried out using Arlequin (Schneider et al. 2000b) with data analysed using SPSS (SPSS 2001), and STRUCTURE (Pritchard et al. 2000). As none of this software is currently adapted to use combined SNP and STR (SNPSTR) data, where necessary, the SNP was treated as an STR, but was given it a value of either 1 or 99, where 1 represents the ancestral state of the SNP and 99 represents the derived state. 1 and 99 represent very widely separated mutational steps in a stepwise mutational model. The figure of 99 was chosen after trials with step numbers differing between 5 and 999, where it was found that 99 offered the most consistent results.

5.3 Admixture in the Greenland Inuit population

5.3.1 Historical background

Human populations have inhabited the Arctic regions for at least 5000 years and archaeological data show that Greenland has been inhabited sporadically for 4 – 4.5 thousand years, with the first settlers arriving via Ellesmere Island from the Canadian peninsula (probably during a warm interglacial period).

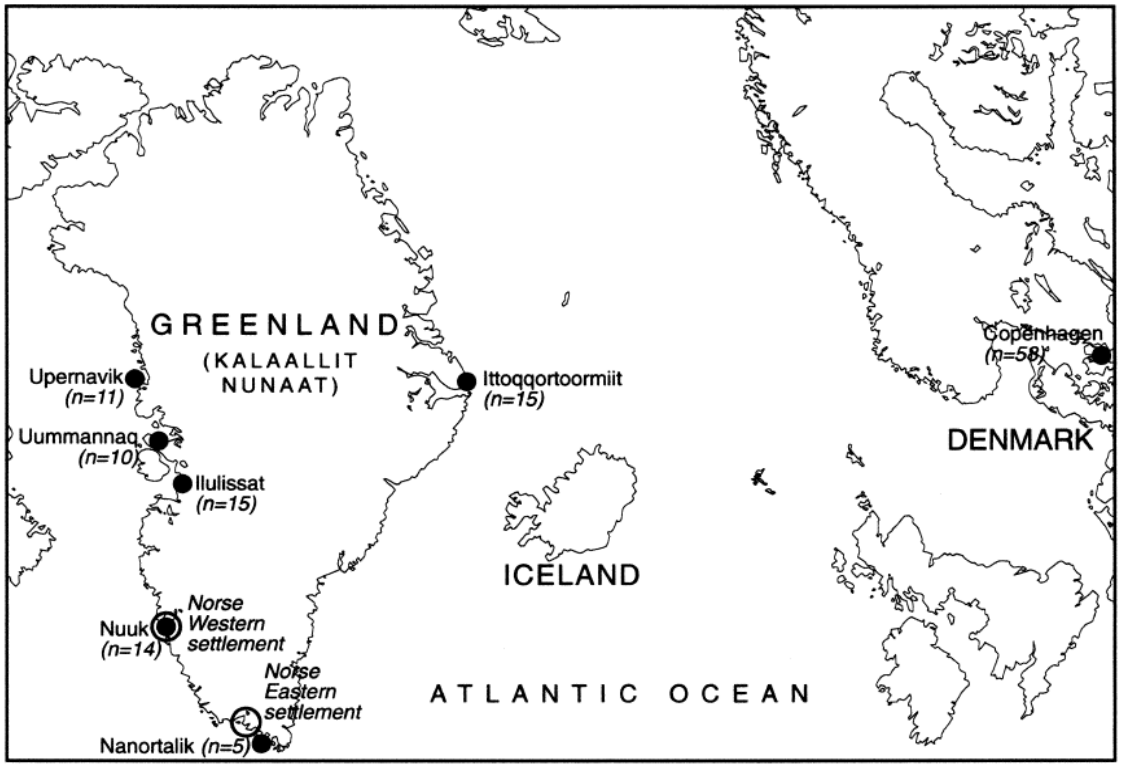


Figure 5-1. Map showing the relative location and size of Greenland and its proximity to Ellesmere Island, which is located at the top, left hand edge of Greenland. Also indicated are the major settlement areas and the number of samples from each site. Taken from (Bosch et al. 2003).

Archaeological evidence shows that Greenland housed resident groups of so-called Dorset populations in the north (Fitzhugh 1984). These people were well adapted to living in the cold climates and hunted seals on the ice, musk ox and caribou. When the Thule, a more technically advanced group (who arose in Alaska around 1000 years ago) spread east, they reached north-west Greenland approximately 800 years ago (McGhee 2000). This spread took place during a cooling of the climate in what is known as the Little Ice Age. And it is this climate change which may go some way to explaining the geographical location of the mtDNA haplotypes which are currently found in Greenland Inuit. The Thule (who came from a warmer

climate and specialized in open water hunting for seals and whales) spread south and congregated in the south and east of Greenland. When these events occurred, one of three possible scenarios could have happened to the existing populations:

- The new tribes could have replaced the existing/ previous populations
- The new tribes could have interbred with the existing populations, or
- Only the culture of the Thule spread east and this alone was transmitted

Past work on mitochondrial DNA (mtDNA) has shown a close similarity between the mtDNA from Siberian Eskimos, Chukchi and Greenland Inuit (Starikovskaya et al. 1998) (Saillard et al. 2000). However, exactly what the mtDNA haplotypes of the ancestral populations were (either Thule or Dorset) from over 1000 years ago, remains unclear. In order to try and establish some background information on the maternal lineages, work was carried out using 395 new mtDNA control regions from Greenland Inuit and Canadian Kitikmeot Inuit (Helgason et al. 2006). The results of this work showed a complicated pattern of regional stratification which is not accounted for by the hypothesis that all Inuit populations in Canada and Greenland are only descended from Thule ancestors. The haplotype frequencies among Canadian Inuit are very different to those from frequencies observed in Greenland or Siberia. In Greenland, there are three distinct groups between the South, West and East and it is unlikely that these differences are the result of a single wave of advance from Alaska only 800 to 1000 years ago. The reason for this is because it would mean that only 27 – 33 generations of drift and a small scale

migration have created a wide pattern of diversity. The alternative is that the Thule encountered existing Dorset populations and interbred with them. This altered the pre-existing geographic pattern of genetic variation and this explanation also gives more time for the complex structural pattern to emerge. There is some direct archaeological evidence to back this theory which has shown that there was interaction between the Dorset and Thule cultures on Victoria Island (Friesen 2004).

The main limitation of this study into mtDNA variation in Inuit populations of Greenland (Helgason et al. 2006) was the lack of DNA available from north American Inuit populations. Instead Siberian populations were used and they are not the best representatives of the Thule mtDNA source pool.

A different perspective on admixture of Inuit populations in Greenland was provided by a study carried out which used the Y chromosome as a marker instead of mtDNA (Bosch et al. 2003). Here, binary markers on the Y chromosome were used to analyse diversity. However, the timescales for this admixture event were more recent than those being investigated with mtDNA.

It is known that in AD 985 ships from Iceland, carrying Icelandic Norse who were originally from the Scandinavian coast and the north west British Isles, sailed to south west Greenland and settled there. At that time, there were no other

populations present there. They established two settlements; one around present day Nuuk (on the west coast) and another north of Nanortalik (which was known as the Eastern settlement), see Figure 5-1 (Kleivan 1984). These two settlements were inhabited by ~1400 people for a period of 500 years (Lynnerup 1998). The last reliable record of contact between the Icelanders and the Greenland Norse was documented in 1414 (Kleivan 1984). Later in the 14th Century the Norse were attacked by what were documented as being "hostile barbarians" (Kleivan 1984) and by the 16th Century, when Europeans visited Greenland they only encountered Inuit people - the Norse had disappeared. This started a quest to find out what had happened to the Norse and in 1721 there was a further concerted effort to colonize Greenland by the Danes and Norwegians and to convert the indigenous people to Lutheranism (Gad 1984). There have been several theories about what happened to the Norse. Their disappearance is linked to a periodic cold phase (Dansgaard et al. 1975), so they could have died out, emigrated back to Iceland or remained and integrated with the indigenous Inuit (Pringle 1997). Lines of evidence were pursued to see if there were any that would indicate contact between the Norse and the Inuit. These included artefacts, shared mythologies and behavioural practices, however, all results were inconclusive and alternative explanations were available. In 1969 evidence based on blood-groups was claimed to support interbreeding (Persson 1969) but detailed analysis of skeletal remains did not support these findings (Scott and Alexandersen 1991; Lynnerup 1998). However, a study carried out which investigated lactose malabsorption in

Greenland Inuit (Gudmand-Hoyer and Jarnum 1969) shows some evidence of autosomal admixture between the Greeland Inuit and the Danes. In a sample of 32 Greenland Inuit, which included 7 who had known Danish ancestry, the lactose tolerance test was performed and the results showed very clearly that on the whole, those with Danish ancestry did not suffer from lactose malabsorption, whereas the indigenous Greenland Inuit did. This early work inadvertently shows indications of autosomal admixture within the Greenland Inuit, with those with Danish ancestry having required enzyme lactase in their digestive system, and the indigenous Inuit lacking the enzyme.

Additionally, further work had been carried out using mtDNA markers from 82 Greenland Inuit, finding no European mtDNA lineage; instead, a spectrum of lineages which were consistent with descent from Alaskan Inuit were found (Saillard et al. 2000). The analysis of the Y chromosome was therefore essential to find out if perhaps there had been a case of sex-biased admixture in Greenland, as had been shown to be the case in Polynesia (Hurles et al. 1998) and in the Americas (Alves-Silva et al. 2000). Because of the robust phylogeny which exists for the Y chromosome, it should be possible to distinguish European from non-European lineages. In order to try and find out what had happened to the Norse, DNA from 69 Greenland Inuit males was sampled from 6 different locations, along with 58 DNA samples from males in Copenhagen (Denmark). 16 Y chromosome binary markers were analysed, along with 7 Y-specific microsatellites. All markers were

compared with microsatellite haplotypes in the European Y STR Haplotype Reference Database (Roewer et al. 2001). The results suggest that there had been unidirectional European admixture into the Inuit. There were a set of ancestral lineages present in the Inuit which were absent from the Europeans and additionally, a set of complementary lineages which were common in Europeans but rare in Native Americans were also identified. The one issue here is that using non-Inuit Native American populations as ancestral reference sources is not ideal because there are different origins for the Inuit and the Amerindians. There is also some European admixture present in modern Native North Americans (Chakraborty 1986) which could confound the results achieved. However, the overall results achieved show that almost 60% of Y chromosomes from the Inuit Greenland males sampled had European origin (Bosch et al. 2003). In contrast, the was a complete absence of European mtDNA (Saillard et al. 2000) which shows that there has been a highly sex-biased admixture event into a non-European population.

The historical records provide two possible sources for the European Y chromosome; one is the original Norse settlers from Iceland who had vanished by the 16th century but who may have been assimilated, and the other source could be the Danish-Norwegian colonizers who came to Greenland after 1727. A comparison of the pooled European component of the Inuit samples with data from European populations show that the Scandinavians are the most likely

source, however, there is also contradictory evidence when it comes to trying to distinguish between the Icelandic Norse and the Danish. If diversity is considered, then the Danes appear to be the most likely source, however, if F_{ST} within haplogroups is considered, then there is a clear relationship with the Icelanders. It is currently not possible to distinguish between these two separate sources due to the large error in dating methods. However, the high degree of sex bias does favour the Danish/ Norwegian source and historical evidence backs this up (Persson 1969). The earlier Norse settlers would have included women, and if they had been assimilated, then there would have been some European mtDNA lineages present. It may be that these simply haven't been sampled yet, or that they have been lost due to drift. However, the European Y chromosomes are distributed widely throughout Greenland and are not concentrated around the former Norse settlements.

5.4 Results of SNPSTR typing on Greenland Inuit DNAs

The question arises of the impact of the sex-biased European admixture into the Greenland Inuit on their autosomal genomes. This could be addressed using ancestry-informative markers (AIMs), or whole-genome analysis, and a formal admixture analysis. Here, however SNPSTRs are used to assess population structure within the Inuit and to ask if a sign of European admixture could be detected.

One very important issue to bear in mind is that the populations used as ancestral population references for the Greenland Inuit were probably not truly representative of the parental population. In this instance, the limits were the available DNAs, which were of Native, North American populations (namely the Maya and the Pima) from the CEPH-HGDP. These two populations reside in Mexico. More appropriate DNAs to use would have been Canadian Inuit, however, none of these were available for analysis at this time.

The charts in Figure 5-2 show the results of the SNPSTRs typed on the Danish (from Copenhagen) and the Greenland Inuit (all samples pooled) DNAs provided by Soren Nørby. These allow differences between distributions of alleles between the populations to be seen by eye.

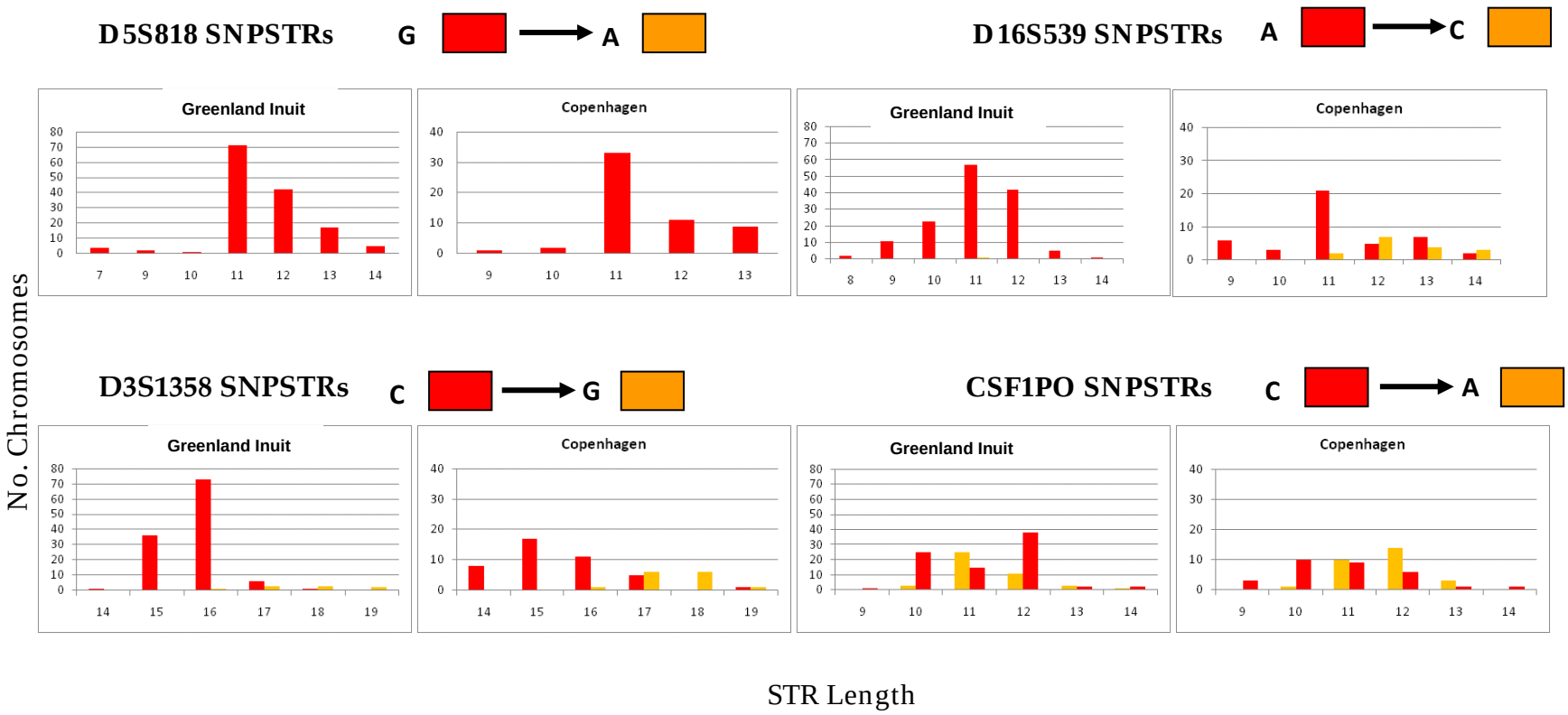


Figure 5-2. SNPSTR charts showing SNPSTR distributions in population samples from the Danish and Greenland Inuit DNA samples. In each case, the red bars indicate the STR lengths based on the ancestral state of the SNP, and the orange bars indicate the STR lengths based on the derived state of the SNPs.

Multi-dimensional Scaling (MDS) plots based on population-pairwise F_{ST} values

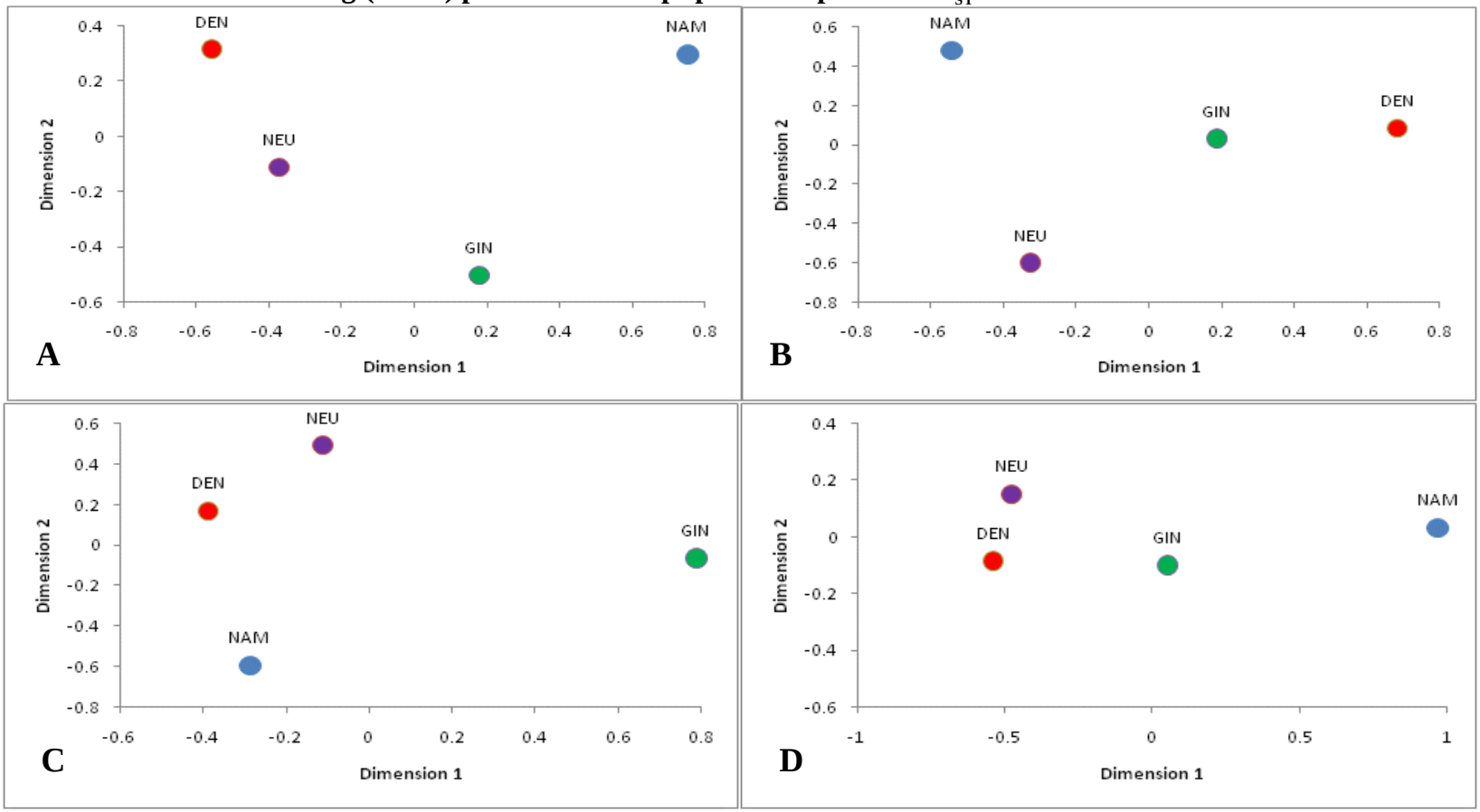


Figure 5-3.. MDS plots created using SPSS. DEN = Denmark (from Søren Nørby), NEU = North East Europe (from the CEPH-HGDP), NAM = Native North American (Maya and Pima), GIN = Greenland Inuit (from Søren Nørby). Plot A shows the D5S818 SNPSTR data, plot B shows the D16S539 SNPSTR, plot C shows the D3S1356 SNPSTR and plot D shows the CSF1PO SNPSTR.

F_{ST} considers different categories where each SNPSTR combination is one category, and does not consider the molecular distance between alleles. In order to be able to compare the parental populations, it must first be ascertained if they are significantly different from one another in the first place. This test was carried out by calculating the F_{ST} P-values using Arlequin for each of the parental populations for each of the SNPSTRs. The results can be seen in Table 5-4 below.

	GI vs DNK	GI vs NAM	GI vs NEU	DNK vs NAM	DNK vs NEU	NEU vs NAM
D5S818	0.15315 (+ - 0.0385)	0	0	0	0	0
D16S539	0	0	0	0	0	0
D3S1356	0	0	0	0	0.3333 (+ - 0.408)	0
CSF1PO	0	0	0	0	0.44144 (+ - 0.0515)	0

Table 5-4. F_{ST} P-values of the 4 SNPSTRs for each of the parental populations, calculated by Arlequin. Significance level = 0.0500. GI = Greenland Inuit, DNK = Denmark, NAM = North American, NEU = Northern Europe.

The significant difference between the most relevant parental populations, i.e. DNK and NEU vs NAM means that it is valid to use them in admixture queries as it provides some power to detect admixture and hence validates the results achieved. The lack of significant differences between Danish and Northern European samples for two of the SNPSTRs can be seen in the MDS plots in Figure 5-3, where, for at least three SNPSTRs, Denmark and Northern Europe tend to group together. The fact that there is no real significant difference between the

Greenland Inuit and the Danish samples for the D5S818 SNPSTR is, however, not reflected in the MDS plot (A). The plots do show that the North American sample tends to be well differentiated from the other three samples for all 4 SNPSTRs. For plots A, B and D, the Greenland Inuit samples tend to lie closer to Denmark and Northern Europe than to North America. We already know that there is male-biased admixture in the Greenland Inuit population (Bosch et al. 2003), but the results achieved by examining the autosomal DNA do not suggest a high degree of European admixture. This could be due to a lack of discriminating power, and therefore the same analysis was carried out using R_{ST}, which considers molecular distances between alleles rather than simply the population categories. As we know that there was male-biased admixture, we assume that there would also be an impact on the autosomal DNA as well. If, however, the male-biased admixture was due to a relatively small number of very successful males where the Y chromosome had propagated well, then the admixture may be very difficult to detect in the autosomal DNA. However, if the admixture was due to a large number of males, then perhaps there may be significant autosomal impact which can be detected by the SNPSTR analysis. In fact, the Greenland Inuit Y chromosome diversity is relatively high, and this is as a result of past admixture events (Bosch et al. 2003) The R_{ST} statistic is more likely to detect this. However, the main problem with using this statistic is that it is primarily designed for STR data only and therefore in order to combine this with SNP data, the values of 1 and 99 were used for the SNPs – treating them as biallelic STRs with maximized allele

length differences between alleles. Although this is not truly representative of the different mutational properties of the SNP and STR, it is currently the best solution available if this particular statistic is needed for the analysis.

As with the F_{ST} statistic, in order to be able to compare the parental populations, it must first be ascertained if they are significantly different from one another in the first place. This test was carried out by calculating the R_{ST} P-values using Arlequin for each of the parental populations for each of the SNPSTRs. The results can be seen in Table 5-5 below.

	GI vs DNK	GI vs NAM	GI vs NEU	DNK vs NAM	DNK vs NEU	NEU vs NAM
D5S818	0.86486 (+-0.0412)	0	0	0.11712 (+-0.273)	0.78378 (+-0.0334)	0
D16S539	0	00.74775 (+-0.0471)	0	0	0.54955 (+-0.0550)	0
D3S1356	0	0.53153 (+-0.0438)	0	0	0.99099 (+-0.0030)	0
CSF1PO	0.09009 (+-0.0271)	0	0	0	0.29730 (+-0.0490)	0

Table 5-5. R_{ST} P values of the 4 SNPSTRs for each of the parental populations, calculated by Arlequin. Significance level = 0.0500. GI = Greenland Inuit, DNK = Denmark, NAM = North American, NEU = Northern Europe.

Because the R_{ST} values consider the molecular differences between populations, it is perhaps not surprising that here, there are fewer significant differences between

the parental populations for each of the SNPSTRs. The lack of significant differences between the Danish and Northern European DNA samples in all 4 SNPSTRs is seen by the clustering of these two groups in the in the MDS plots below. The lack of significant difference between Greenland Inuit and North America for the D3S1358 SNPSTR can also clearly be seen in Plot C, where these two cluster together. Figure 5-4 shows the MDS plots based on R_{ST} data for all 4 SNPSTRs.

Multi-dimensional Scaling (MDS) plots based on population-pairwise R_{ST} values

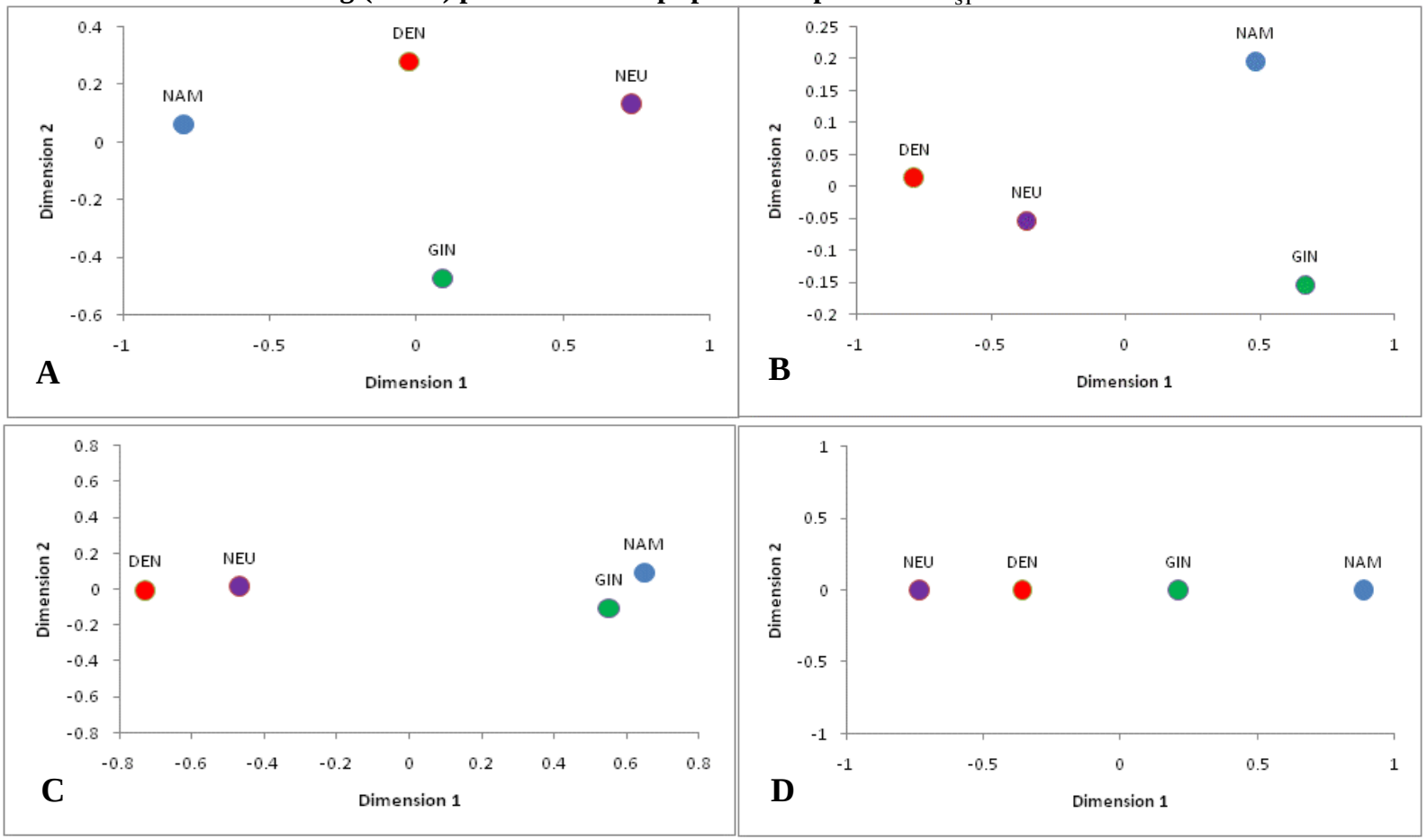


Figure 5-4. MDS plots created using SPSS. DEN = Denmark (from Søren Nørby), NEU = North East Europe (from the CEPH-HGDP), NAM = Native North American (Maya and Pima) , GIN = Greenland Inuit (from Søren Nørby) . Plot A shows the D5S818 SNPSTR data, plot B shows the D16S539 SNPSTR, plot C shows the D3S1356 SNPSTR and plot D shows the CSF1PO SNPSTR.

The MDS plots above are based on the R_{ST} values. These charts are different to those from the F_{ST} data, and distinctly different patterns can be seen for each of the four SNPSTRs. Plot A showing D5S818 does not show up any groupings at all, which is not surprising as three of the parental populations are not significantly different from each other (see Table 5-5) and therefore any clustering seen would not be "real". However, Plots B and C do appear to show two distinct groupings, one containing Denmark and Northern Europe, and the other containing North America and Greenland Inuit – but this is again shown to be lacking in significance because of the fact that the parental populations are not significantly different from one another (

Table 5-5).

As can be seen from the results above, these locus-by-locus analyses are of inherently low power, and give inconsistent results. For this reason, it was decided to use STRUCTURE (Pritchard et al. 2000) to assess population structure for all SNPSTRs simultaneously, and to ask if a signal of admixture in the autosomal DNA could be observed.

STRUCTURE implements a model-based clustering method for inferring population structure using genotype data consisting of unlinked markers. In STRUCTURE a model is assumed in which there are *K* populations (where *K* may be unknown), and each population is characterized by a set of allele frequencies at

each locus. Individuals in the sample being tested are assigned (probabilistically) to populations. One of the assumptions which is made in STRUCTURE is that within the populations, the loci are at Hardy-Weinberg equilibrium and in linkage equilibrium. If they are not, then STRUCTURE will assign individuals in such a way that they achieve this. The important fact to note is that STRUCTURE is not designed to handle markers that are extremely close together, which means that it is not the ideal software for analyzing SNPSTRs; however, as each of the 4 SNPSTRs are not linked to each other, the data achieved have some value. Therefore, despite some suitability issues, STRUCTURE is currently the best option available for analyzing SNPSTRs, and for this reason it was decided to use it on the Greenland Inuit DNA data. Figure 5-5 shows the results achieved using STRUCTURE. In this analysis, STR data alone from CEPH-HGDP populations, along with the Greenland Inuit samples were compared with SNPSTR data to see whether the SNPSTRs were more effective than STRs alone at detecting population admixture in the Greenland Inuit.

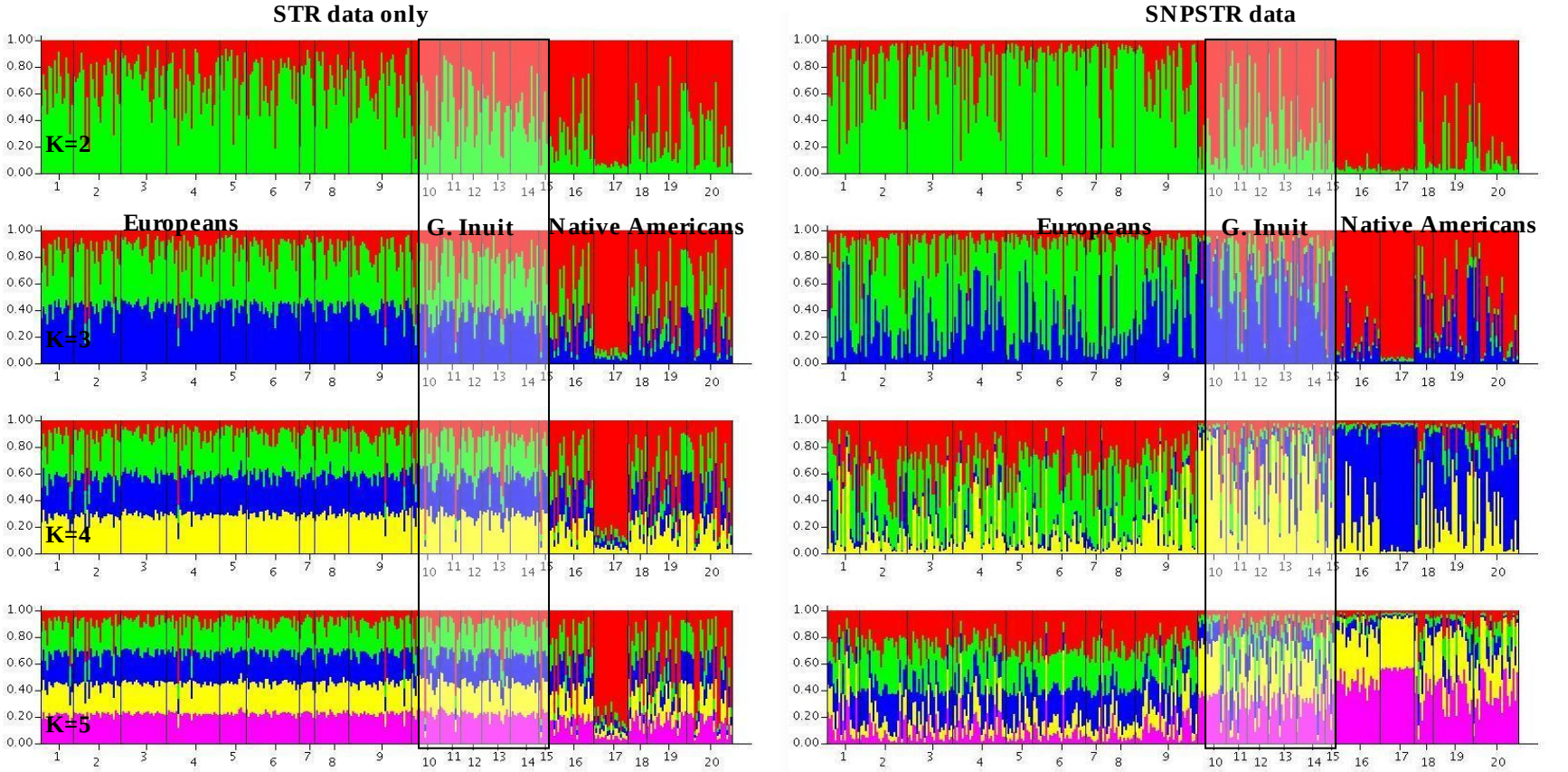


Figure 5-5. STRUCTURE **analysis of STRs and SNPSTRs as viewed with the CLUMPP and *distruct* software (http://rosenberglab.bioinformatics.med.umich.edu/software.html) of the Greenland Inuit DNAs along with those from Native America and Europe.** The numbers underneath the blocks (1 – 20) represent the pre-defined populations. Numbers 1 – 9 are European samples, numbers 10-15 are Greenland Inuit DNAs and numbers 16-20 are Native American DNAs (see Appendix 1 for a full population list). Each individual is represented by a single vertical line broken into *K* coloured segments, with lengths which are proportional to each of the *K* inferred clusters.

5.5 Discussion of Greenland Inuit and Danish DNA results

As mentioned previously, Greenland was first inhabited about 4,500 years ago, which is relatively late compared to other countries, such as Australia, which was colonized as early as 50,000 years ago (Jobling et al. 2003), which is before modern humans appeared in Europe. This late colonization, followed by isolation could lead to the development of private alleles present in the Greenland Inuit population. However, if there has been admixture with the Norse and later the Danes who colonized Greenland much later, then these private alleles may well have been lost over time.

From Figure 5-2, which shows the SNPSTRs, it can be seen that the D5S818 SNPSTR only has the ancestral SNPs in both Greenland Inuit populations and also in Denmark, while the STR repeats show a wider range in the Greenland Inuit than in Denmark. Both populations show that 11 STR repeats have the highest frequency. As the Greenland Inuit have a higher range of STR repeats, it suggests that the D5S818 SNPSTR was already present in the Greenland Inuit when the Danish settled there later. As both populations have similar profiles for this SNPSTR, it is not possible to detect any admixture by means of this SNPSTR.

For the D16S539 SNPSTR, the Danish samples have both ancestral and derived SNPs, but the Greenland Inuit sample only contains one allele with a derived SNP (which is only just visible in the Figure 5-2). This SNPSTR is more informative

than the D5S818 SNPSTR as it could possibly show the effects of a single founder for this SNPSTR in the Greenland Inuit, possibly from the Danes, but this is actually impossible to tell.

The D3S1358 SNPSTR has an interesting distribution pattern in that the ancestral SNPs are mainly present on the lower STR repeats, both in the Greenland Inuit and in the Danish samples. Additionally, in both populations, STR alleles associated with the derived SNPs cover an identical range of STR allele repeat numbers.

Finally, the CSF1PO SNPSTR is interesting in that the derived SNP alleles associated with 11, 12 and 13 STR repeats are in the majority for the Danish DNA samples, whereas only the derived SNP associated with the 11 STR repeats is in the majority for the Greenland Inuit samples. The profiles for the different populations are interesting as it could suggest the same founder individual for both populations. There is no way to tell whether the DNA travelled from Denmark to Greenland or vice versa.

The caveat with all of these results is that what is seen may simply be due to the fact that not enough DNA samples were taken and that these are not truly representative of the Danish and Greenland Inuit populations.

From the STRUCTURE results in Figure 5-5 it can be seen that using the K = 2 population data setting is the most reasonable setting to use to detect structure due to the European/ Native American admixture. Other settings for K can reveal some other structural aspects, such as the singular nature of population 17 (the Surui – an indigenous people living in the Pará region of Brazil who are very isolated).

From the STR data only, it is not possible to tell if there is any clear signal of population admixture in the Greenland Inuit. However, when the SNPSTR data are considered, there is a clearer signal of admixture in the Greenland Inuit. It is possible to use SNPs alone to detect admixture, however, as only 4 SNPs were considered, they would not have had sufficient discriminating power to detect admixture.

From the STRUCTURE analysis, the results suggest that the Greenland Inuit are a hybrid of Europeans and Native Americans at the SNPSTR loci. However, as the representatives of the parental populations (namely Native Americans and Northern Europeans) are few and also, in the case of the Native American populations used, not truly representative of the parentals, STRUCTURE produces very biased proportions of the ancestry estimates. This occurs because in the absence of any non-admixed individuals, there is probably some non-identifiability where the allele frequencies are possibly pushed further apart, and

the admixture proportions are squeezed together. Despite this, the same degree of model fit is achieved, and STRUCTURE is not able to overcome this issue. For this reason, admixture estimates produced by STRUCTURE can not be relied upon to be truly representative (Hubisz et al. 2009).

Following on from this analysis, similar studies were carried out using African Caribbean DNAs. Again, the materials and methods were as described in Chapter 2, and the DNA analysis was also carried out as with the Greenland Inuit, with the exception of the STRUCTURE analysis. This was not carried out for the African Caribbean DNAs because there were so few relevant population DNAs available to which they could be compared.

5.6 Admixture in the African-Caribbean population

5.6.1 Background

The Caribbean consists of four main geographical regions comprising the Bahamas, the Greater Antilles (which includes Jamaica, Cuba, Hispaniola and Puerto Rico), the Lesser Antilles (including the Eastern chain of islands and the group of islands to the north of the Venezuelan coast), and the Islands of Trinidad, Tobago and Barbados. It extends 300km east from the Yucatán peninsula and 1500km south from Cuba to the South American continent and includes more than 7000 smaller islands. Figure 5-6 shows a map of the Caribbean.

Humans have occupied the Caribbean for over 8000 years, with the first inhabitants coming from the South American continent and the Yucatán Peninsula (Weeks and Ferbel 1994). The original native inhabitants of some of these islands were known as the Ciboney, the Arawak and the Carib (Rouse 1992).

In 1492 the Spaniards reached the Caribbean Islands and it was only Spanish men who made the journey (Rouse 1992). The indigenous Caribbean populations were decimated by disease, warfare and forced labour (Sued-Badillo 2003). In 1518, 4,000 indigenous Africans were transported to the Antilles to act as a labour force. These were mostly males, but did include some females at a slightly later date. Then again in the 17th century, more enslaved Africans from the sub-Saharan states of West and Central Africa were imported to work on the plantations (Rogoziânski 1999). The transatlantic slave trade lasted for three centuries and was one of the largest forced migration events in human history.

Figure 5-6. . Map of the Caribbean showing the islands and their relative locations. (Source: Wikimedia Commons)

Because of this mixing of peoples, the Caribbean islands have been the focus of many studies carried out by population geneticists. Some have focused on the search for ancestry informative makers (AIMs) to aid biomedical research where racial categories are deconstructed into useful variables. One such study (Benn-Torres et al. 2008) used 28 AIMs (SNPs) to study the genetic ancestry of 298 individuals of African descent from the Caribbean islands of Jamaica, St. Thomas and Barbados. They found that Jamaica had the highest levels of European (12.4%) and Native American (3.2%) admixture, St. Thomas had very similar levels of admixture to African Americans in continental USA (86.8% West African, 10.6% European and 2.6% Native American), and Barbados had the highest levels of West African ancestry (89.6%) and the lowest European (10.2%) and Native American (0.2%) ancestry. There were definite differences in population substructure across the three Caribbean islands, where there was significant structure in Jamaica and St. Thomas but not in Barbados. These differences stem from the diverse colonial and historical experiences and the subsequent evolutionary processes. Native populations did survive in the more inaccessible regions and it was shown that location also influenced the admixture pattern. One example can be seen in Barbados, which is relatively isolated and this deterred the European colonists and other migrants. This resulted in lower levels of non-African admixture on Barbados. It is very important that these differences are acknowledged as they may have implications for case-control studies of complex diseases in Caribbean populations (Hirschhorn et al. 2002). In this study, the addition of more markers

may have enabled a better estimate of admixture, however, the AIMs did prove to be useful markers for admixture mapping.

In a second study, mitochondrial DNA and Y chromosome diversity was investigated in the English-speaking Caribbean (Benn Torres et al. 2007). In this study 501 individuals from 8 different locations (Dominica, Grenada, Jamaica, St. Kitts, St. Lucia, St. Thomas, St. Vincent and Trinidad) were sampled and then nucleotide and gene diversity (Nei 1987) were examined in order to investigate the within- and between-population diversity for the mitochondrial locus (using DNASp (Rozas et al. 2003)). For the Y chromosome, haplotype diversity was estimated using Arlequin (Schneider et al. 2000a). Following on from this, AMOVA was used to explore the population structures within the Caribbean samples and between the Caribbean and African populations. Four potential parent populations were identified: Africans, South Asians, Europeans and Native Americans. The small sample size could have limited the analysis, but the results achieved did match other studies previously carried out (Parra et al. 1998; Miljkovic-Gacic et al. 2005). The similar range in gene, nucleotide and haplotype diversity between the Caribbean and the African populations suggests that for the Y chromosome and mtDNA loci, there has not been any loss of genetic diversity as a result of the forced migration of subsets of African populations to the Caribbean. However, having said that, the results could have been improved by the inclusion

of the appropriate Native American parental data, which had not been used for this study.

AMOVA analysis of the Y chromosome data and mtDNA data on the Caribbean islands showed that most variation was within the island and that only a small, but significant amount of variation occurred between the islands. The substructure found could be due to genetic drift, isolation or other factors. The AMOVA analysis of the Y chromosome data showed variation between the African and Caribbean populations and this is attributable to the non-African admixture in the Caribbean groups. MtDNA analysis showed that a large proportion of the mtDNA was of African origin. There was no substantial contribution from indigenous Caribbean or Eurasian females. This was confirmed by the historical records which stated that most of the indigenous population was eliminated before the islands were re-colonized with the enslaved Africans (Rouse 1992). The conclusion of this work was that the distribution of the African genetic contribution to each of the island populations was tightly linked to the respective colonial histories as well as each island's demographic composition and this switched depending on whether they were under British, French or Spanish control.

5.6.2 The African-Caribbean populations in the UK

Most African-Caribbean people who reside in the UK have come from Jamaica. They were originally of West African origin. A minority come from the other

islands which make up the Caribbean, and include Guyana, which though located on the South American mainland, is very culturally similar to the Caribbean, and was historically considered to be part of the British West Indies, and Belize (formerly British Honduras).

Migration from the Caribbean to the UK was rare before World War II and there is little historical information available on those that did make it over before then. The earliest record dates back to 1837 where a Jamaican was the minister of Cradley Heath Baptist Church (Britain 2008). In the mid 19th Century, there were small Caribbean communities living in the areas of the large ports (Cardiff, Liverpool and South Shields) but they were mainly formed of freed slaves. After World War II many African-Caribbean people migrated to North America and Europe. This was primarily as a result of the losses during the war, when the British government encouraged mass immigration from the countries of the British Empire and Commonwealth to fill shortages in the labour market.

In the UK Census of 2001, 565,876 people classified themselves in the category 'Black Caribbean'. The Census also recorded the respondents' countries of birth and the 2001 Census recorded 146,401 people from Jamaica, 21,601 from Barbados, 21,283 from Trinidad and Tobago, 20,872 from Guyana, 9,783 from Grenada, 8,265 from St. Lucia, 7,983 from Montserrat, 7,091 from St. Vincent and the Grenadines,

6,739 from Dominica, 6,519 from St. Kitts and Nevis, 3,891 from Antigua and Barbuda, and 498 from Anguilla.

5.7 Results of SNPSTR typing on African-Caribbean DNA samples

Knowing that the current UK population of African Caribbean people originated from various different islands of the Caribbean, and are likely to have primarily West African descent with some sex-biased European admixture, SNPSTRs were used to assess population structure. In the MDS plots which follow, the populations which were used from Central and Western Africa include the Yoruba, Mandenka and Biaka DNAs from the CEPH-HGDP. The North Eastern Europeans include the Orcadian, Danish and Cornish populations.

Statistical analysis of the data was carried out using Arlequin (Schneider et al. 2000b) and SPSS (SPSS 2001).

Table 5-6. below shows the F_{ST} P values as calculated by Arlequin for the clusters which are used to ascertain whether or not, the different parental DNAsare significantly different to each other in the first instance. If they are not, then any groupings seen in the MDS plots below, and any other analyses, cannot be used as useful indication of any population structure. A study using skin pigmentation , biogeographical ancestry and admixture mapping (Shriver et al. 2003) compared markers African Americans, African Caribbeans from the UK and European

Americans from Pennsylvania and modeled population admixture using STRUCTURE (Pritchard et al. 2000). This concluded that the UK African Caribbean population had darker skin pigmentation than the African Americans due to lower levels of European Admixture and that the average European ancestry in the African Caribbeans was 10.2% compared to 18.6% in the African Americans.

From Table 5-6, it can be seen that for the majority of SNPSTRs, there is no significant difference between the African-Caribbean DNAs and those from Central and Western Africa. This is reflected in the MDS plots, where these two populations appear to group closer to each other than to the European samples.

	CW AF vs NEU	CW AF vs AFCAR	NEU vs AFCAR
D5S818	0	0	0
D16S539	0	0.29730 (+-0.0430)	0
D3S1356	0	0.21622 (+-0.0411)	0.09009 (+-0.0303)
CSF1PO	0	0.55856 (+-0.0633)	0

Table 5-6. F_{ST} P values of the 4 SNPSTRs for each of the parental populations, calculated by Arlequin. Significance level = 0.0500. CW AF = Central and Western Africa, AFCAR = African-Caribbean, NEU = North Eastern Europe.

The F_{ST} figure considers different categories where each SNPSTR combination is one category. The F_{ST} figure does not consider the molecular distance between alleles, instead it is used to distinguish haplogroups from each other.

MDS plots are in essence, a two dimensional representation of a multi-dimensional image, and therefore distances are not representative as true distances between populations.

Multi-dimensional Scaling Plots (MDS) Showing F_{ST} Statistics

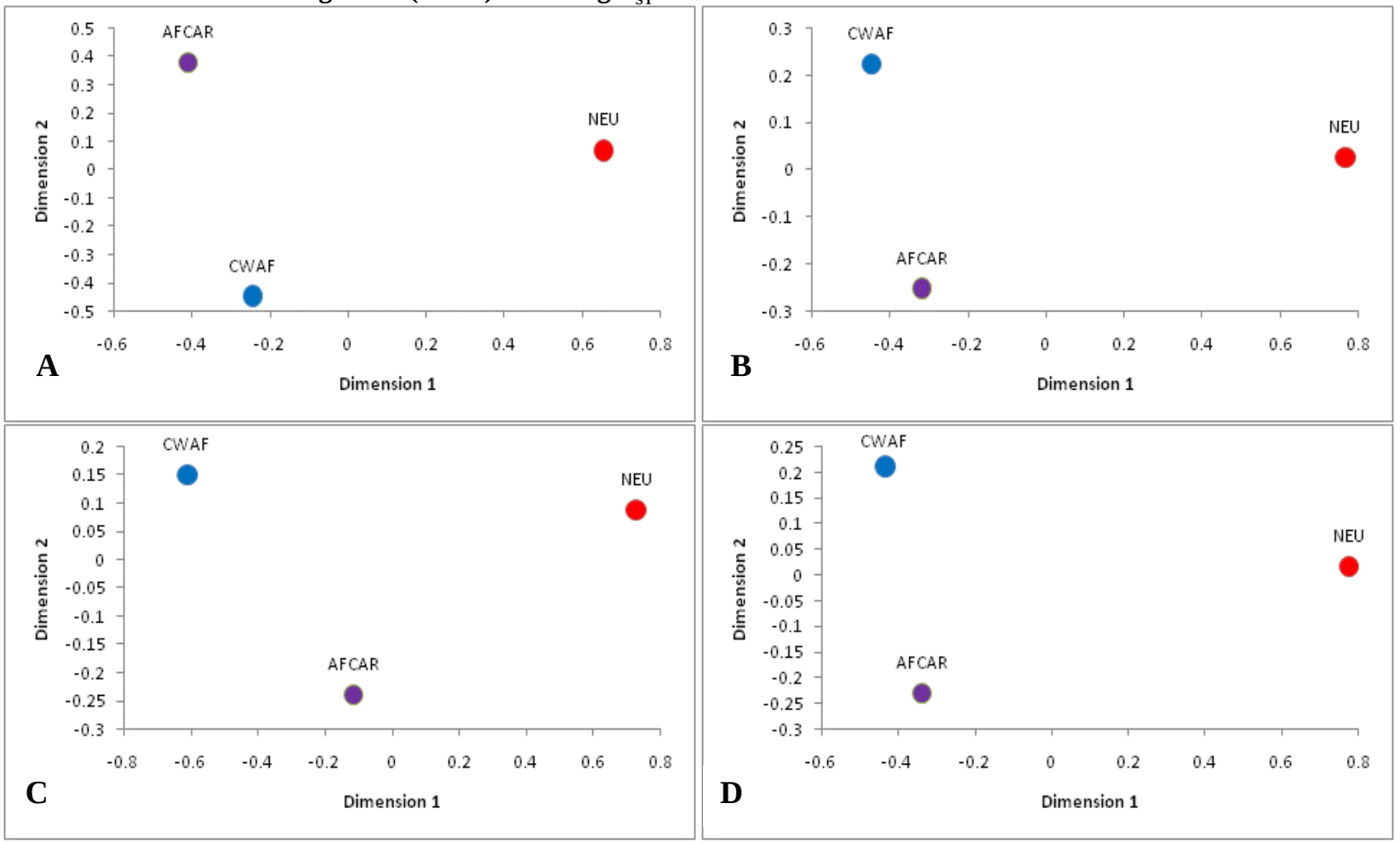


Figure 5-7. MDS plots created using SPSS. CWAF = Central and Western Africa, NEU = North East Europe, AFCAR = African Caribbean. Plot A shows the D5S818 SNPSTR data, plot B shows the D16S539 SNPSTR, plot C shows the D3S1356 SNPSTR and plot D shows the CSF1PO SNPSTR.

However, from the plots above, which are limited as only 3 populations are included, it is not possible to see if there is any evidence of any population groupings and for this reason, the same analysis was carried out using R_{ST}, which considers molecular distances between alleles. The only problem with using this statistic is that it is primarily designed for STR data only and therefore in order to combine this with SNP data, the values of 1 and 99 were used for the SNPs. Although this is not truly representative of the different mutational properties of the SNP and STR, it is currently the best solution available if this particular statistic is needed for the analysis.

	CW AF vs NEU	CW AF vs AFCAR	NEU vs AFCAR
D5S818	0	0	0
D16S539	0	0	0
D3S1356	0	0	0
CSF1PO	0	0	0

Table 5-7. R_{ST} P values of the 4 SNPSTRs for each of the parental populations, calculated by Arlequin. Significance level = 0.0500. CW AF = Central and Western Africa, AFCAR = African-Caribbean, NEU = North Eastern Europe.

Thankfully, with respect to the molecular distances between the parental populations and the test population, all are significantly different to each other, which means that they can be compared directly in the MDS plots below, and any clustering observed are valid.

Multi-dimensional Scaling Plots (MDS) Showing R_{ST} Statistics

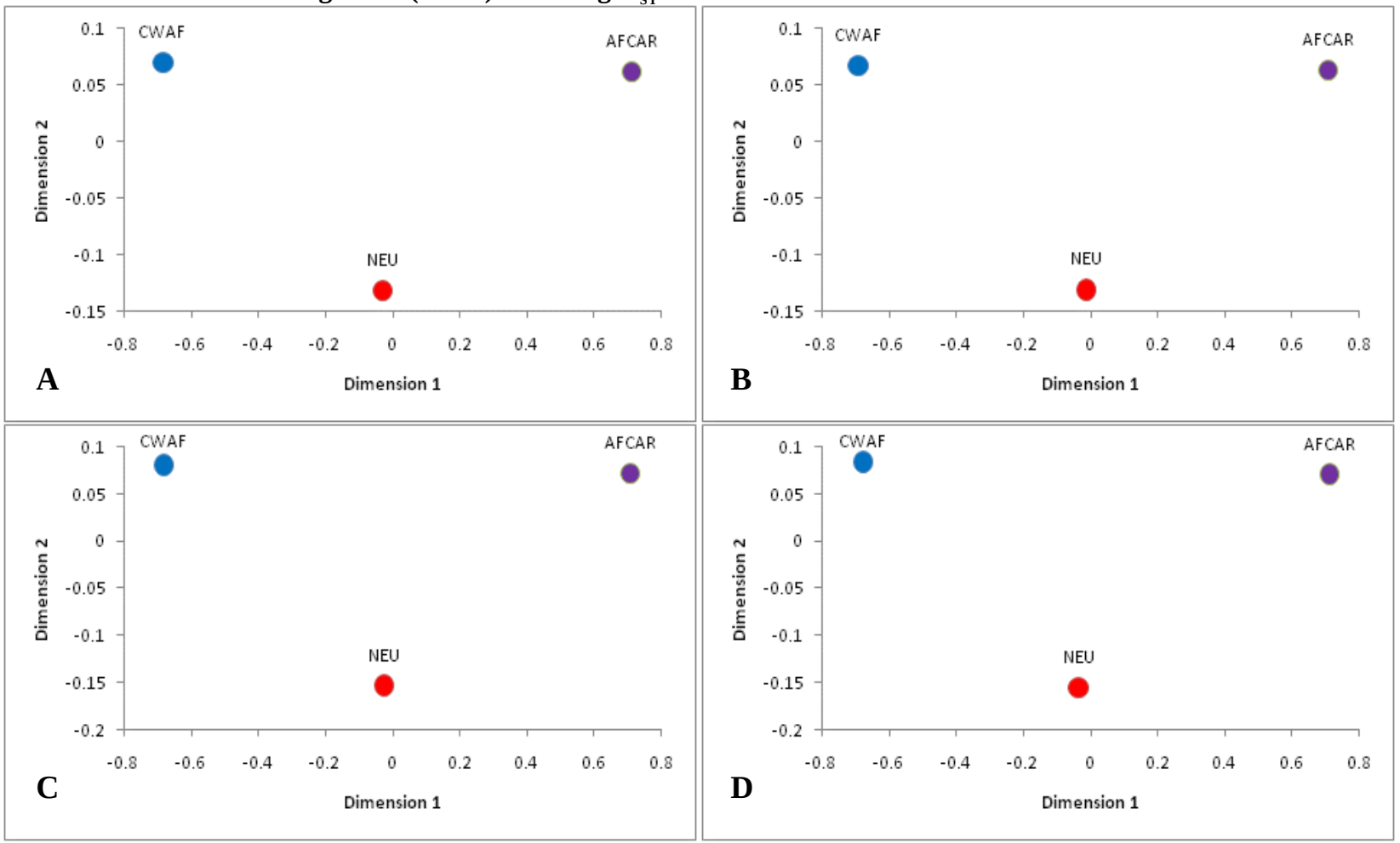
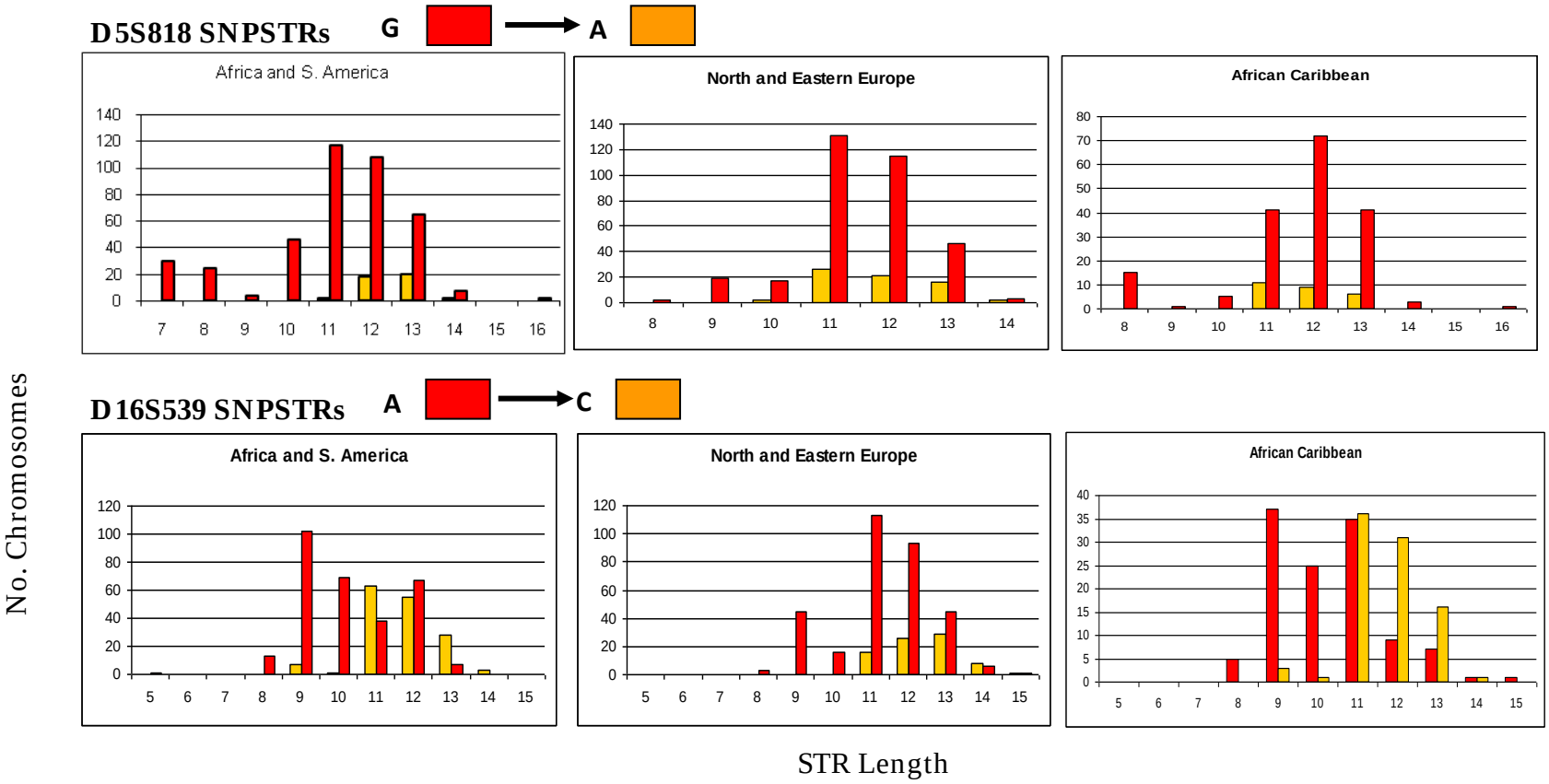


Figure 5-8. MDS plots created using SPSS. CWAF = Central and Western Africa, NEU = North East Europe, AFCAR = African Caribbean. Plot A shows the D5S818 SNPSTR data, plot B shows the D16S539 SNPSTR, plot C shows the D3S1356 SNPSTR and plot D shows the CSF1PO SNPSTR.

These charts are slightly different to those from the F_{ST} data; however, they are all exactly the same as each other. All three populations are equidistant from each other, which does not provide any information about possible population groupings or admixture. This result is somewhat unexpected, and perhaps the inclusion of additional populations in the analysis may show different groupings which would be more informative.

The charts in Figure 5-9 below, show the results of the SNPSTRs typed on the African Caribbean, African and Northern and Eastern European DNAs. These allow any differences between the distributions of alleles for each SNPSTR to be seen by eye.


D5S818 SNPSTRs
G
A
Africa and S. America
North and Eastern Europe
African Caribbean
D16S539 SNPSTRs
A
C
Africa and S. America
North and Eastern Europe
African Caribbean
No. Chromosomes
STR Length

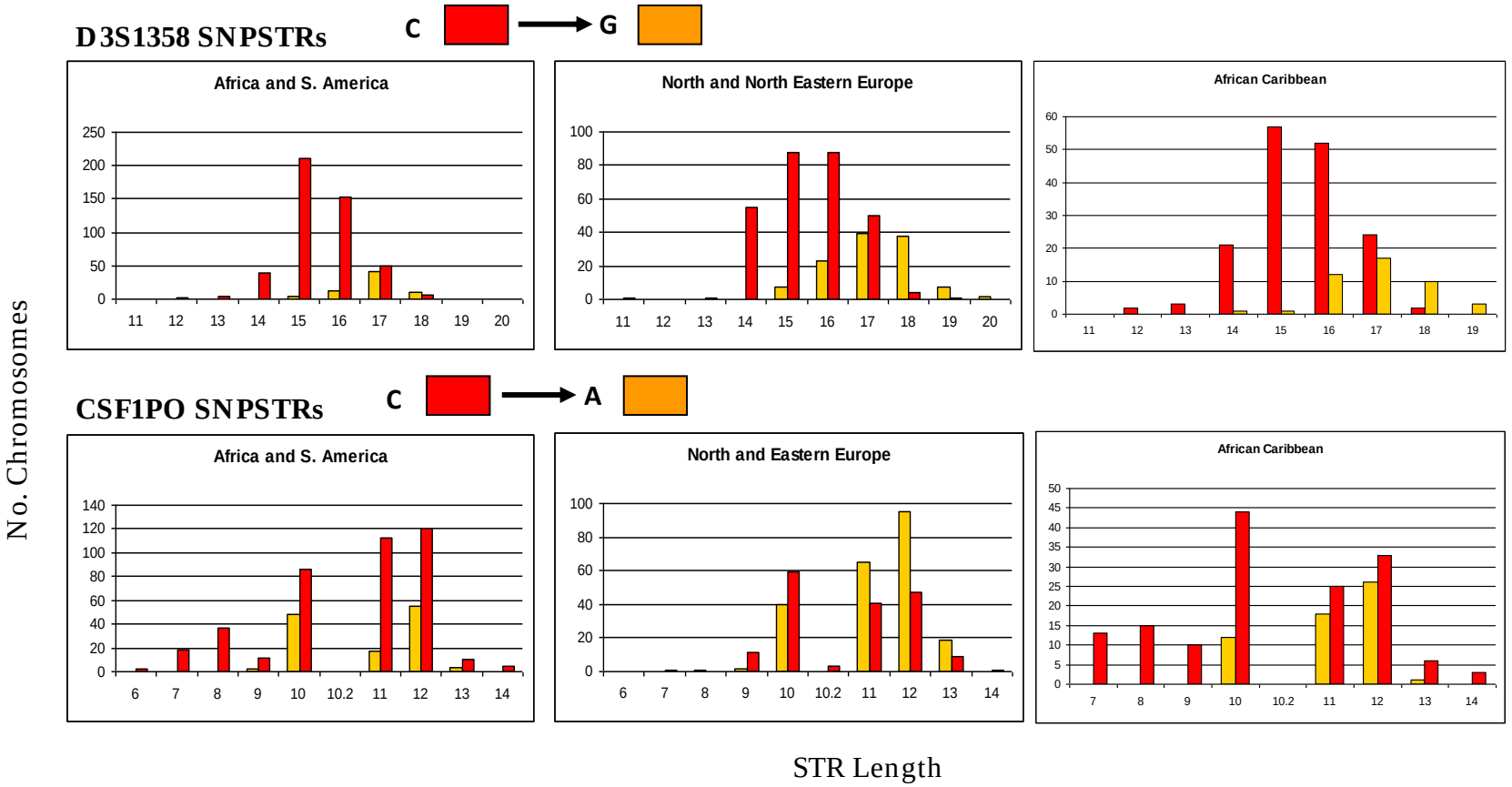


Figure 5-9. SNPSTR charts of all 4 SNPSTRs typed on the African Caribbean DNAs. These charts show the SNPSTR distributions. In each case, the red bars indicate the STR lengths based on the ancestral state of the SNP, and the orange bars indicate the STR lengths based on the derived state of the SNPs.

5.8 Discussion of African Caribbean DNA results

In general, the graphs in Figure 5-9 all show that the STRs associated with the derived SNP allele have a smaller range than those associated with the major SNP allele. The major and minor STR allele ranges overlap each other in all cases. There are also examples of smaller "frozen" STR ranges in all of the populations.

The two charts which stand out are the D16 SNPSTR one for the African Caribbean DNAs where the minor SNP allele STR frequencies are higher than the STR frequencies associated with the major SNP allele, and also the CSF1PO SNPSTR for the North and Eastern European DNAs where a similar phenomenon occurs.

These results are also reflected in the data achieved with the SNPSTR analysis, however without rare or variant alleles, it is very difficult to be able to distinguish which section of DNA has come from which ancestral population. Luckily, the CSF1PO SNPSTR does contain one such variant – the 10.2 STR repeat allele. This allele is present in the European DNAs but not in the African nor the African Caribbean DNAs. However, this one allele does not really provide us with much insight into the admixed nature of these populations. On a simple, by eye, comparison level, the African Caribbean SNPSTR patterns tend to be more similar to those of Africa than those from Northern and Eastern Europe. As the UK African Caribbean inhabitants only arrived recently

(around 70 years ago), there has not been sufficient time for changes in the SNPSTRs to become apparent. Therefore what would be picked up in these populations would be traces of the more ancient admixture events which occurred before the current inhabitants arrived in the UK. It would have been interesting to see if there were traces of the 4,000 indigenous Africans who were transported to the Caribbean Islands in 1518, or of the later 17th century enslaved Africans from West and Central Africa. As the SNPSTRs do show more similarity with the African ones, this is perhaps a signal of one or other of these two forced migrations. Unfortunately though, with only four markers, there is not sufficient data available to be able to distinguish between these two events.

Perhaps, using larger haplotype blocks, such as the PHAXs described in Chapter 6, may be more informative in describing the admixed nature of the autosomal DNA present in the UK African Caribbean population.

6 PHAXs – Phylogeographically informative Haplotypes on Autosomes and X-chromosomes

6.1 Introduction

This chapter is based on the bioinformatics work carried out in our laboratory by Dr. Stéphane Ballereau. 'PHAXs' (a name coined by Mark Jobling and as yet unpublished) represent extended haplotypes showing no evidence of historical recombination, and as such have the potential to turn the relatively small SNPSTRs with their limited information into larger haplotype blocks, which may be able to provide more insight into the histories and structures of populations. Using a list of the locations of the PHAXs as defined by Stéphane Ballereau as a basis, the SNPSTRs, (which had already been typed on the HapMap, CEPH-HGDP, Cornish, African-Caribbean, Danish and Inuit DNA samples) were used as a starting point to see if one or more lay within the boundaries of a PHAX. If they did, then the possible haplotypes were extracted and phylogenetic trees were produced to see whether the PHAX in combination with the SNPSTR were useful as a tool for population genetics.

6.2 Definition of a PHAX

Definitions of PHAXs have at their basis the publicly available polymorphism data from the HapMap project (2003) (http://www.hapmap.org). At the time when the PHAX data was being analysed for this thesis, the PHAXs had been defined using release 21 for the four population samples as follows: 60 individuals with European ancestry (CEU parents), 60 individuals with African

ancestry (YRI parents), and 90 individuals with East Asian ancestry: Japanese (JPT) and Chinese (CHB). Haplotypes were inferred using PHASE (Stephens et al. 2001; Stephens and Donnelly 2003) PHASE implements a Bayesian statistical method for reconstructing haplotypes from population genetic data. The software is able to deal with SNP, STR and other multi-allelic loci such as tri-allelic SNPs, as well as missing data. These phased haplotypes were downloaded from the HapMap web site and then used to derive linkage disequilibrium measures using Haploview (Barrett et al. 2005). Non-recombining regions were identified as non-overlapping series of at least three adjacent SNPs where each pair of SNPs had a D' value of 1 in each of the three HapMap populations (CEU, YRI and JPT+CHB), and for which only 3 of the 4 possible 2-allele haplotypes were observed in the entire sample set, including the ancestral haplotypes, whether these were observed or not. The D' value of 1 meant that the SNPs were in linkage disequilibrium and therefore defined the haplotype block.

Where several such regions overlapped, the series spanning the largest physical distance was selected and any other regions overlapping it were discarded. This process was repeated until no regions were left to chose. The ancestral states of the SNPs were obtained from the UCSC Genome Browser, March 2006 assembly, using the table browser function and the snp126OrthoPanTro2RheMac2 table. For any given SNP, the chimpanzee and/ or macaque sequence was taken to be representative of the ancestral state. If the SNP was not represented in the chimpanzee or macaque sequences, then

the ancestral state was taken to be the major allele in the global human sample (CEU+YRI+JPT+CHB). SNPs were filtered out if they were in a series of several SNPs where the ancestral state was not able to be derived from the chimpanzee or macaque sequences. SNPs were also filtered out if they formed a series of several SNPs which came exclusively from several different chromosomes in the chimpanzee or macaque. This was done in order to avoid using unreliable alignments of human to chimpanzee or macaque sequence. Finally, haplotypes accounting for the ancestral state of SNP alleles were derived for each PHAX, using SNPs for which genotypes were available in all three populations (CEU, YRI and JPT+CHB).

PHAXs, are, of course not totally reliable as haplotype blocks, because they can be broken up, not only by recombination, but also if any of the SNPs used to define the PHAX have been mistyped. If SNP mistyping was the cause of the break-up a PHAX, it would be possible to test for this. If the DNA segments either side of the PHAX remained in regions of low linkage disequilibrium (LD), then there may be a recombination hotspot responsible for the PHAX break-up. If, however, the areas either side of the PHAX definition were in high LD, then it is possible that SNP mistyping had broken up the PHAX. Much would also depend upon the model assumed for SNP mutation. Most models assume that each site evolves independently and that substitutions are time reversible such that the total overall rate of conversion into a certain nucleotide is equal to the total rate of loss of that same nucleotide at equilibrium. However, if a PHAX were defined using SNPs that did not fit this model, then

this would result in an unreliable definition of a PHAX. Finally CpG hypermutation would also interfere with the defining of a PHAX. In mammals and other organisms, CpG hypermutability is one of the major causes of nucleotide mutations because CpG dinucleotides are often methylated at C, and the methyl-C mutation spontaneously deaminates to yield T about 3 times more rapidly than other types of point mutations (Misawa and Kikuno 2009). It would, in theory, be possible to see if this was breaking up the PHAX by examining the DNA sequences where the PHAX is located. If the PHAX was in a region which was rich in repetitive DNA elements, such as SINEs or ALUs, then CpG hypermutability may be the cause, and examination of individual SNPs and their immediate context would be worthwhile.

On some occasions, however, it might be useful to consider adjacent PHAXs. This could be carried out where the recombinant haplotypes that separate the PHAXs are very rare. It depends upon the stringency of the PHAX definition protocol. The adjacent PHAXs were considered in all cases for this project, but in all of the cases, the PHAXs were broken up due to recombination occurring between adjacent PHAXs. Due to the stringency applied in the PHAX definition, the most likely cause in this case would be recombination rather than mistyping errors or CpG hypermutation.

6.3 Materials and methods

Table 6-1 lists the PHAXs which are located either next to, or encompassing, one of the four SNPSTRs. Based on the data contained in the table, the

haplotypes were obtained from the HapMap website (Release 21), and haplotype groups and trees were drawn. Figures 6-1, 6-2, 6-3, and 6-4 show the PHAXs, their locations on the chromosomes, the local recombination environment, as well as the haplotype groups and possible trees.

Chrom.	PHAX ID	Start (bp)	End (bp)	Length (bp)	Up Dist	Down Dist	Microsat	No. H.groups	Microsat Info.	Notes
5	19228	123136371	123138050	1680	974	2930	0	4	0	
5	19229	123138809	123139317	509	293	429	1	4	$(TCTA)_{11}$	Contains D5S818
5	19230	123148004	123148781	778	9757	8703	0	2	0	
16	11547	84939833	84942438	2606	1097	4096	0	6	0	
16	1548	84943868	84944468	601	933	61	0	4	0	D16S539 overlaps PHAX 11548
16	11549	84945596	84945785	190	2250	1667	0	4	0	
3	7951	45546084	45556890	10807	319	11255	3	8	$(AAAAG)_8$ $(AC)_{22}$ $(TG)_{15}$	
3	7952	45556963	45562987	6025	5778	376	1	5	$(TATC)_{16}$	Contains D3S1358
3	7953	45564415	45569126	4712	11917	7076	0	9	0	
5	23342	149434635	149435222	588	706	1611	0	4	0	
5	23343	149435808	149437004	1197	1076	438	1	2	$(CTAT)_{19}$	Contains CSF1PO
5	24444	149438896	149439574	679	3646	2650	0	5	0	

Table 6-1. List of PHAXs containing or located next to an SNPSTR.
A PHAX overlaps an STR if both 'Up Dist' and 'Down Dist' are shorter than the PHAXs length. This table is based on data contained within Release 21 of the HapMap project. The PHAXs containing the forensic STRs are shown above, as well as the PHAXs on both flanking regions. The 'Microsat Info.' column shows which microsatellites are found within the PHAX. The 'No. H.groups' = the numbers of haplogroups found within each PHAX based on dbSNP and genotyped SNPs.

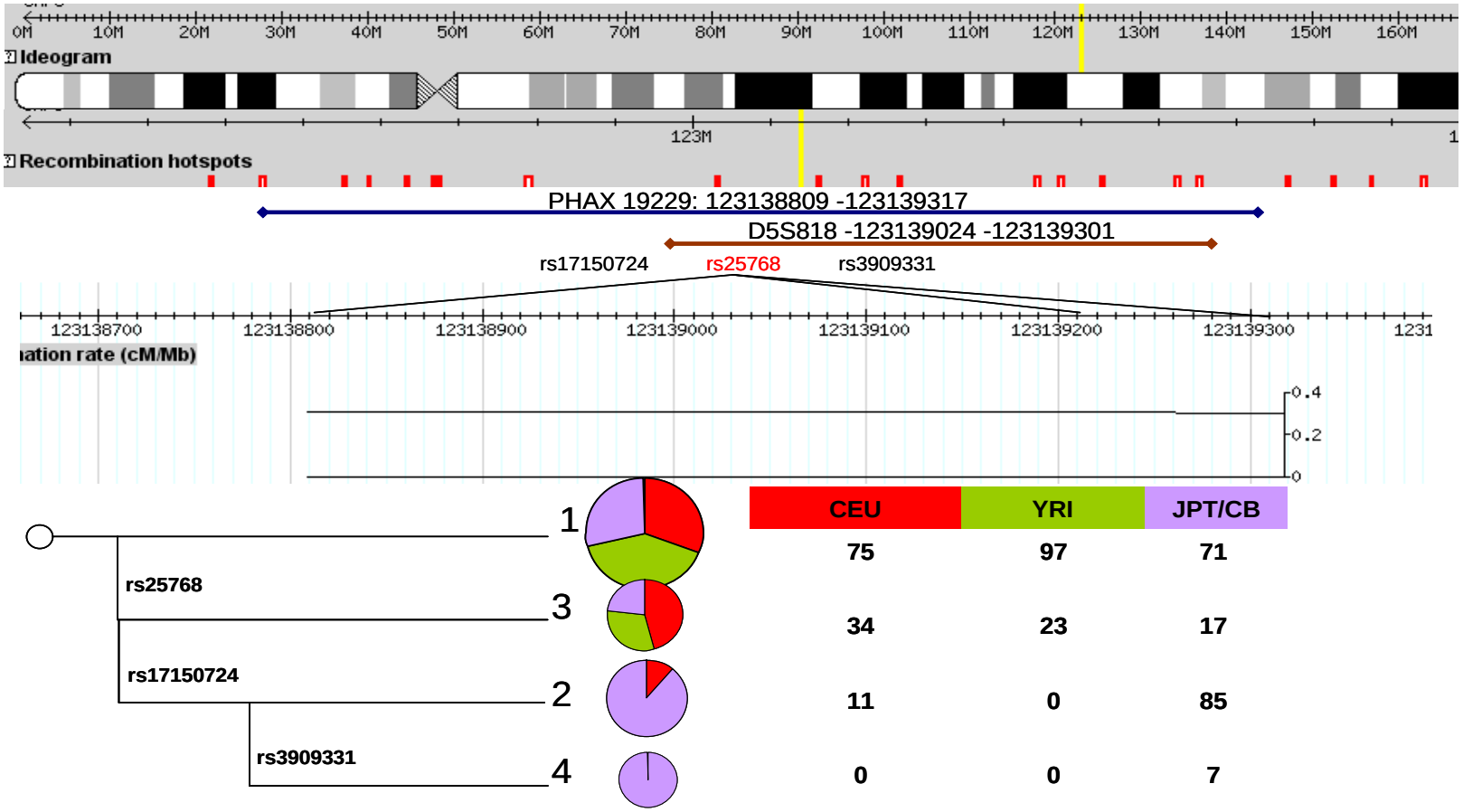


Figure 6-1. PHAX 19229 on Chromosome 5, containing the D5S818 SNPSTR.
rs25768 (in red), was typed as part of the SNPSTR assays; the other SNPs were not typed. The haplotype tree and corresponding pie charts show possible haplogroups. The pie chart sizes are relative to the numbers of individuals forming that particular haplogroup in the PHAX. The blue line indicates the position of the PHAX, the brown line indicates the position of the D5S818 STR. The average recombination rate in this area is very low (approximately 0.3 cM/ Mb).

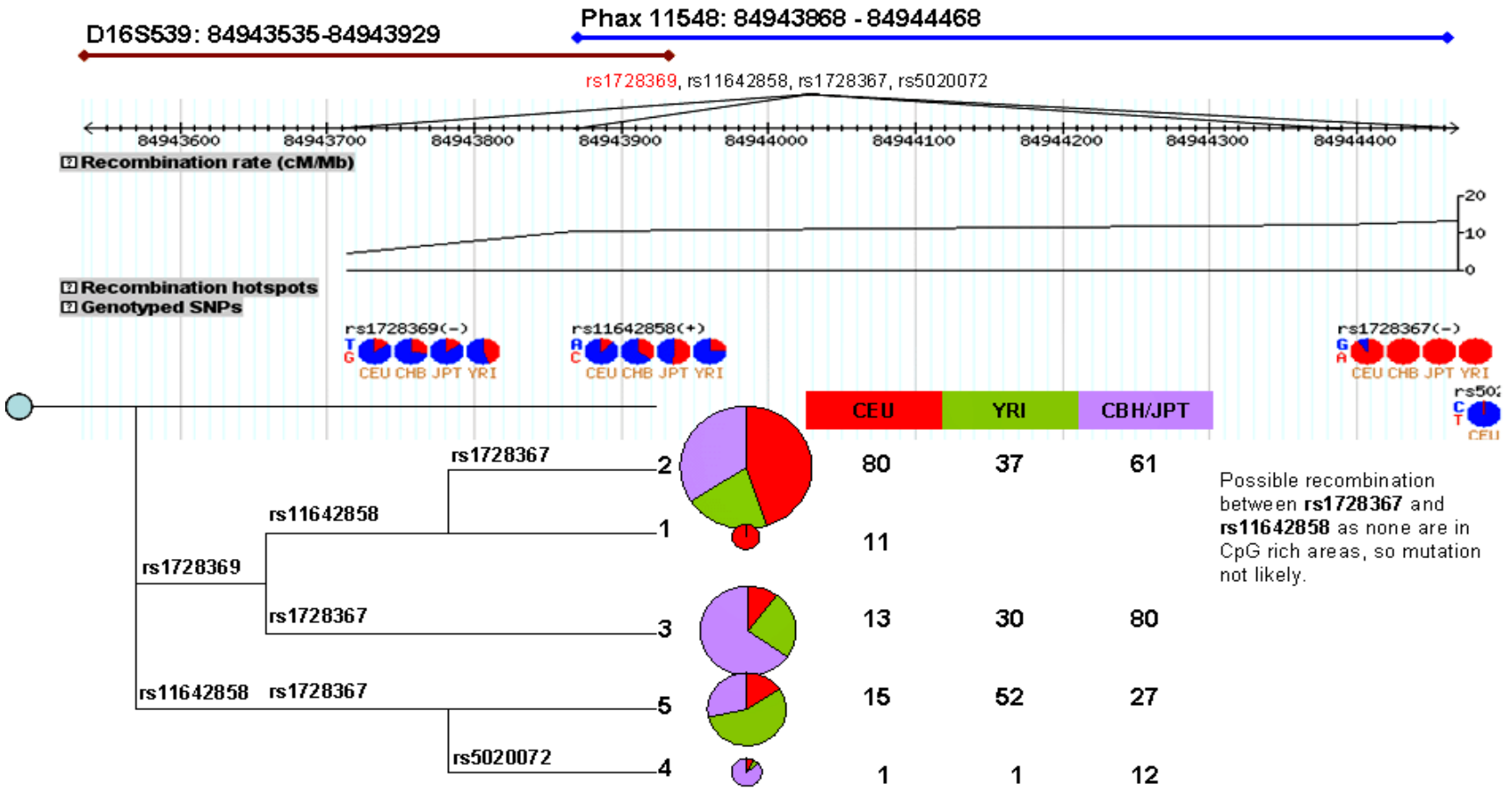


Figure 6-2. PHAX 11548 on Chromosome 16, flanking the D16S539 SNPSTR.
SNP rs1728369 (in red), was typed as part of the SNPSTR assays; the other SNPs were not typed. The haplotype tree and corresponding pie charts show the possible haplogroups. The pie chart sizes are relative to the numbers of individuals forming that particular haplogroup in the PHAX. The blue line indicates the position of the PHAX, and the brown line indicates the position of the D16S539 STR. The average recombination rate in this area is relatively high (approximately 12cM/ Mb).

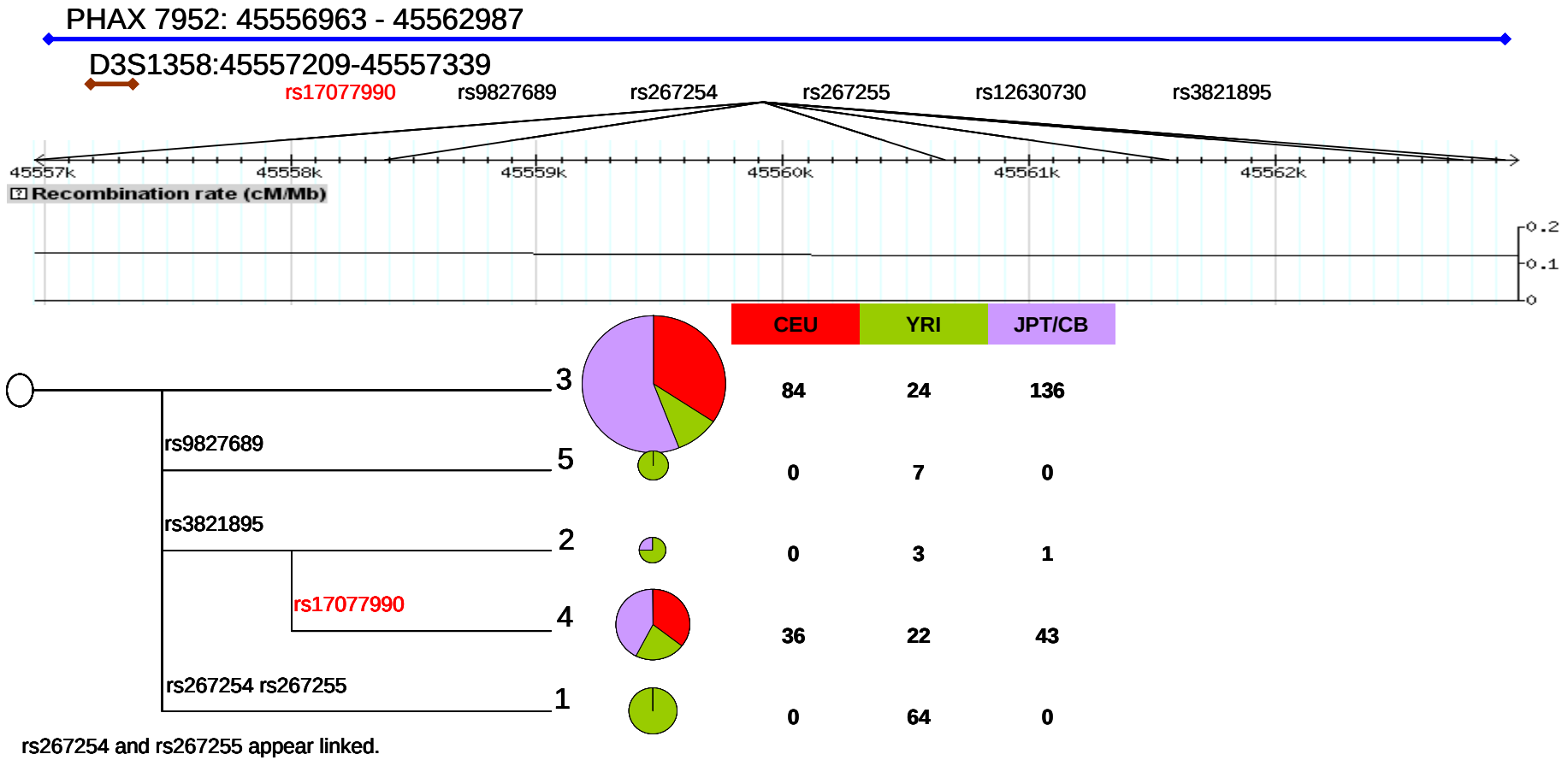


Figure 6-3. PHAX 7952 on Chromosome 3, containing the D3S1358 SNPSTR.

SNP rs17077990 (in red), was typed as part of the SNPSTR assays; the other SNPs were not typed. The haplotype tree and corresponding pie charts show the possible haplogroups. The pie chart sizes are relative to the numbers of individuals forming that particular haplogroup in the PHAX. The blue line indicates the position of the PHAX, and the brown line indicates the position of the D3S1358 STR. The average recombination rate in this area is very low (approximately 0.1cM/ Mb).

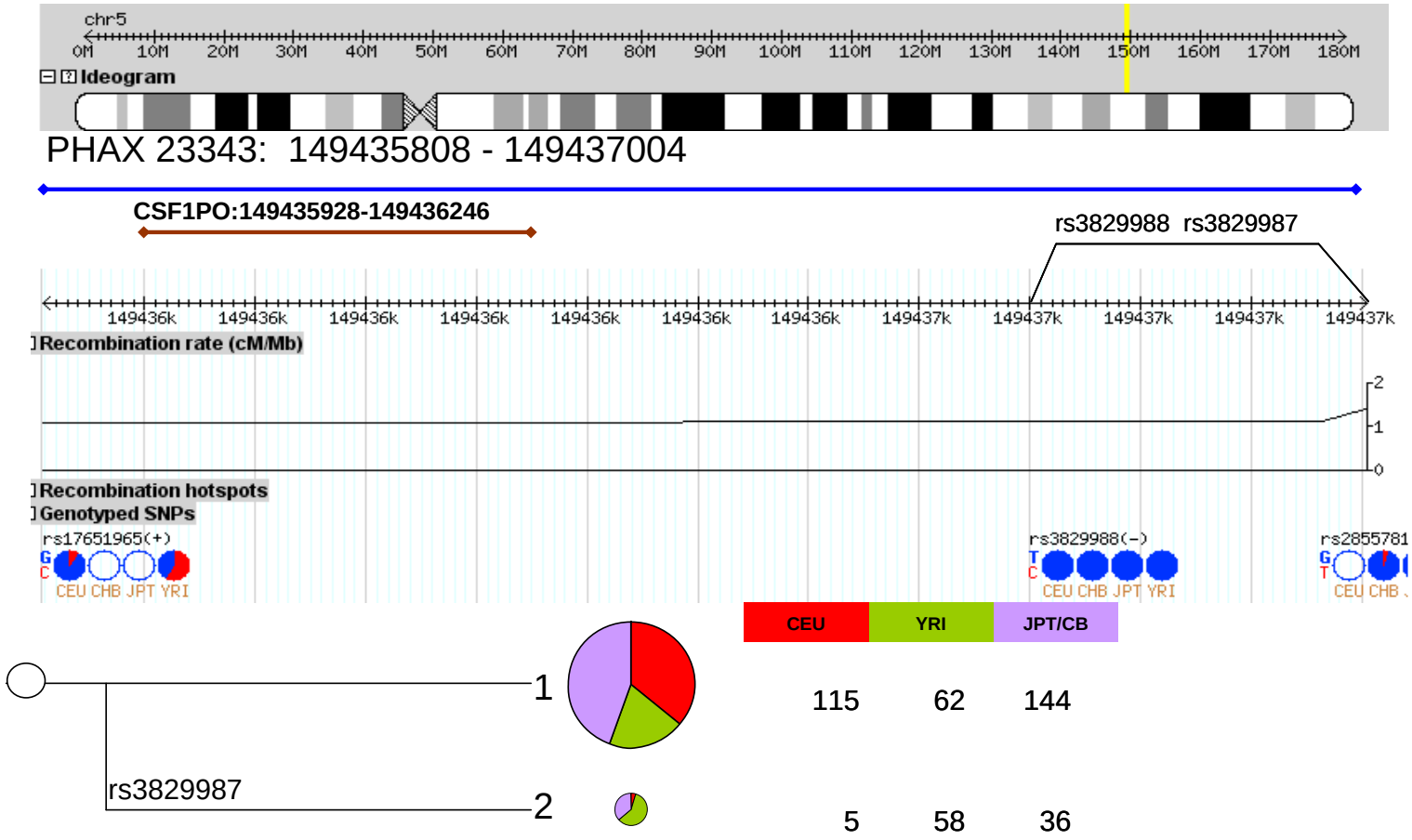


Figure 6-4. PHAX 23343 on Chromosome 5, containing the CSF1PO STR. SNP rs2116791 which was typed as part of the SNPSTR assays falls just outside of the PHAX boundary. The haplotype tree and corresponding pie charts show the possible haplogroups. The pie chart sizes are relative to the numbers of individuals forming that particular haplogroup in the PHAX. The blue line indicates the position of the PHAX, and the brown line indicates the position of the CSF1PO STR. The average recombination rate in this area is low (approximately 1.5cM/ Mb).

From the data obtained in the trees and the haplogroups, a decision was made whether to further investigate a particular PHAX or not. Further investigation would include typing of the SNPs contained within the PHAXs on the Greenland Inuit, Danish, Cornish and African Caribbean DNA samples. The SNPs would be typed by means of a SNaPshot reaction, method as outlined in Chapter 2, Materials and methods.

From the figures above and the resulting haplogroup blocks and trees, it was decided to concentrate solely on PHAX 7952 which contained the D3S1358 SNPSTR which had already been typed on the populations listed in Chapter 4. This PHAX was chosen because it contained the greatest number of SNPs which had already been genotyped, including one which had already been typed for this project. This typed SNP, along with the other genotyped ones would therefore provide positive controls for any further assays. However, it did allow sufficient scope for typing of not only HapMap genotyped SNPs but also of the additional dbSNP SNPs.

6.3.1 SNaPshot flanking PCR

As the region covered by PHAX 7952 was approximately 6025bp long, it was not possible to amplify this whole PHAX reliably in one PCR reaction. For this reason, the total area was divided into three sections, each containing several of the SNPs to be typed. The initial step was to reliably amplify each of these three sections and Figures 6-5, 6-6, and 6-7 show each of these three sections of

DNA in full detail, including the SNPs present and the primers used for the flanking PCR reactions.

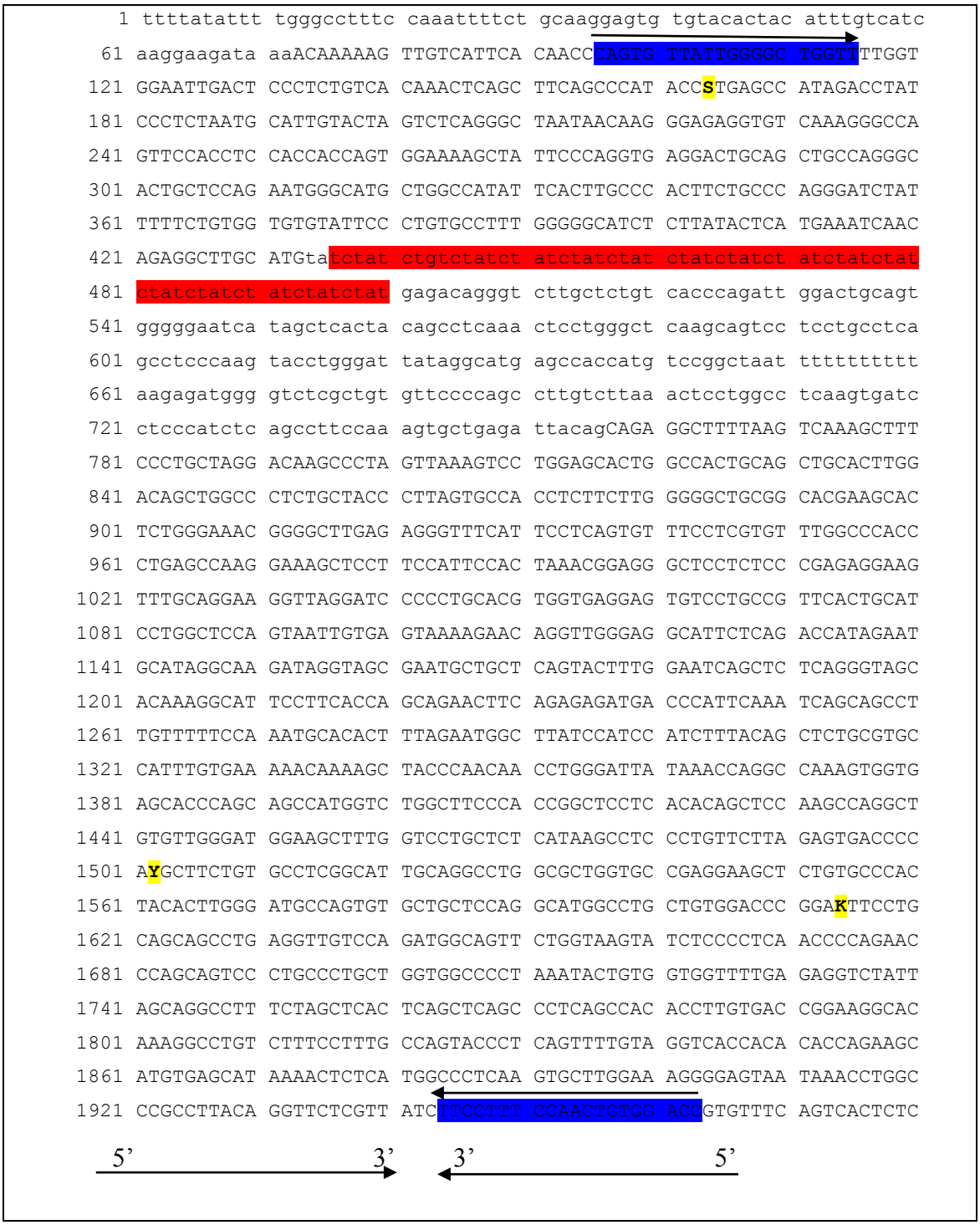


Figure 6-5. DNA sequence from NCBI build 36 showing SNaPshot flanking PCR using Blue2 Primer Set.

SNPS are highlighted in yellow and are from top to bottom: rs17077990, rs34191038 and rs9827689. DNA primers are shown in blue with 5’ to 3’ indicated by directional arrows.

Figure 6-6. DNA sequence from NCBI build 36 showing SNaPshot flanking PCR using Green Primer Set.

SNPS are highlighted in yellow and are from top to bottom: rs35703033, rs35568132, rs35744531, rs35015312, rs267254. DNA primers are shown in green with 5’ to 3’ indicated by directional arrows.

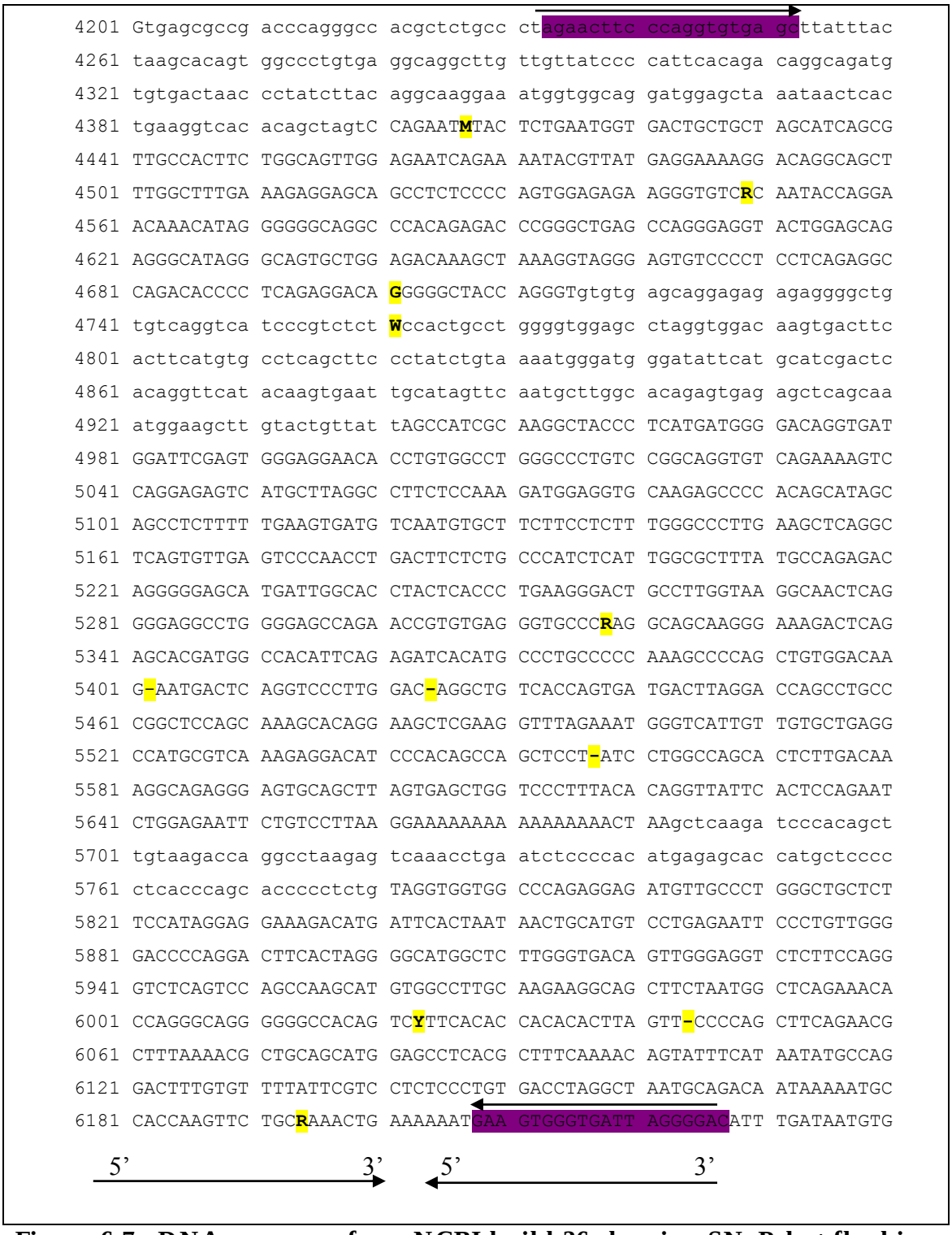


Figure 6-7. DNA sequence from NCBI build 36 showing SNaPshot flanking PCR using Purple Primer Set.

SNPS are highlighted in yellow and are from top to bottom: rs34058890, rs17077992, rs36095084, rs267255, rs267256, rs35383157, rs34581796, rs10700258, rs12630730, rs35129241, rs3821895. DNA primers are shown in purple with 5' to 3' indicated by directional arrows.

Table 6-2 below provides more detailed information on the SNPs within PHAX 7952

SNP	Alleles	Ambiguity Code	Ancestral Allele	Notes and validation status
rs17077990	C/ G	S	C	Included in SNPSTR assay. Genotyped by HapMap.
rs34191038	C/ T	Y	Unknown	Unvalidated
rs9827689	G/ T	K	G	Genotyped by HapMap.
rs35703033	C/ A	M	Unknown	Unvalidated
rs35568132	-/ T in-del	-	Unknown	Unvalidated
rs35744531	C/ A	M	Unknown	Unvalidated
rs35015312	-/ C in-del	-	Unknown	Unvalidated
rs267254	G/ A	R	T	Genotyped by HapMap.
rs34058890	C/ A	M	Unknown	Unvalidated
rs17077992	G/ A	R	G	Genotyped by HapMap.
rs36095084	G/ - in-del	-	Unknown	Unvalidated
rs267255	T/ A	W	A	Genotyped by HapMap.
rs267256	G/ A	R	G	Genotyped by HapMap.
rs35383157	-/ T in-del	-	Unknown	Unvalidated
rs34581796	-/ A in-del	-	Unknown	Unvalidated
rs10700258	-TCT in-del	-	Unknown	Unvalidated
rs12630730	C/ T	Y	C	Genotyped by HapMap.
rs35129241	-/ C in-del	-	Unknown	Unvalidated
rs3821895	A/ G	R	T	Genotyped by HapMap. This SNP marks the end of the PHAX

Table 6-2. SNPs within PHAX 7952. Data from NCBI website, build 129.
Two 'SNPs' which were within the D3S1358 STR were not included because one was simply an additional 4 base-pair repeat at the end of the STR, and the other was also a variation in the STR repeat at the start of the STR. As can be seen, the ancestral state for most of the unvalidated SNPs is unknown

A triplex PCR was carried out on the flanking regions of the PHAX, using the primers as listed in the table below:

Primer	Length	Sequence	5' Mdfctn
PHAX7952 Blue F-2	20	CAGTGTTATTGGGGCTGGTT	None
PHAX7952 Blue R-2	20	GGTCCACAGTTGGAAAGGAA	None
PHAX7952 Green F	20	GCATAGTGCTCTGGGGATAC	None
PHAX7952 Green R	20	GCATCCAGAAGCAGGCTAAG	None
PHAX7952 Purple F	20	AGAACTTCCCAGGTGTGAGC	None
PHAX7952 Purple R	20	GTCCCCTAATCACCCACTTC	None

Table 6-3. The sequences of primers used in PHAX 7952 triplex PCR. Mdfctn = modification.

Triplex PCR conditions were as follows: PCR programme "P7952-3" – 94 ºC for 10 minutes, followed by 94 ºC for 1 minute, 55.8 ºC for 1 minute, 70 ºC for 2.5 minutes for a total of 30 cycles ending at 65 ºC for 10 minutes and 15 ºC for 5 minutes. The product sizes for each of these PCR reactions are: Blue2 primers - 1868 bp, Green primers - 1683 bp, and Purple primers - 1994 bp. See Chapter 2 for a full description of the methods used for the SNaPshot triplex PCR.

Following on from the triplex PCR, SNaPshot primers were designed for a smaller subset of all the possible SNPs. This was done firstly to prove the method, but also because there is a limit on the numbers of SNPs which can be amplified and typed reliably in a single SNaPshot reaction. Therefore, the SNPs which were included in the first SNaPshot reaction were those that were validated by HapMap, and which would therefore provide a positive control

for the method, as they could all be verified online for the HapMap DNA samples.

6.4 SNaPshot primers – set 1

The first set of primers were directed at SNPs from within each of the three sections of DNA (as seen in Figure 6-5, Figure 6-6, and Figure 6-7), and each primer had to be optimized individually for temperature and concentration, before being pooled and then tested as part of a pooled set. The SNP-specific primers within each flanking primer set each had poly-A tails of different lengths added to make each SNaPshot product a different size and also to space them out sufficiently from each other so that the resulting peaks could be distinguished from each other when read on GeneMapper software (v3).

The SNPs within set 1 comprised of the following:

SNP	Primer Name	Primer Sequence	Length	Opt. TM/ °C	Final Primer [ ]
rs17077990	P7952-S1-1	ACTCAGCTTCAGC CCATACC	20bp	63.1	30µM
rs9827689	P7952-S1-2	$(A)_5$TGGCCTGCTGT GGACCCGGA	25bp	78.3	20 µM
rs267254	P7952-S1-3-40bp	$(A)_{20}$TGCCCATGTTC TTTGTGCCT	40bp	76.1	10 µM
rs267255	P7952-S1-4	$(A)_{15}$TCCAAGGGACC TGAGTCATT	35bp	73.5	20 µM
rs267256	P7952-S1-5	$(A)_{21}$CATCCCACAGC CAGCTCCT	40 bp	74.9	30 µM
rs12630730	P7952-S1-6-25bp	AAAGGAACTAAGT GTGTGGTGTGAA	25 bp	64.5	20 µM
rs3821895	P7952-S1-7	$(A)_{34}$TGCCACCAAGT TCTGC	50 bp	76.7	30 µM

Table 6-4. SNPs typed in the first SNaPshot assay.
Although some of the SNaPshot primers are the same length as each other, they were not used together in the same reaction. The purple primer set was used as one reaction, and the blue and green primer sets together in another reaction.

6.5 Sequencing of SNPs for verification of non-HapMap SNPs

In preparation for typing the remaining SNPs in PHAX 7952 which had not been typed by HapMap and for which there were, in effect, no positive controls, PCR products from a selection of DNAs from the CEPH-HGDP DNA panel were sent for sequencing by the Protein Nucleic Acid Chemistry Laboratory (PNACL), based at the University of Leicester.

One DNA sample from each subcontinent was chosen. Additionally, only those that were heterozygous at the D3S1358 SNPSTR for both alleles were chosen because of the fact that PHAX 7952 contains this STR.

CEPH No	Sub Continent	Population	Allele 1 SNPSTR	Allele 2 SNPSTR
HGDP00808	N. Europe	Orcadian	C16	G18
HGDP01384	E. Europe	Adygei	C15	G17
HGDP01374	S. Europe	French Basque	C15	G16
HGDP00858	N. America	Maya	C15	G17
HGDP00706	S. America	Colombian	C16	G17
HGDP01408	Sub Saharan Africa	Bantu, N.E.	C16	G18
HGDP01279	N. Africa	Mozabite	C15	G17
HGDP00631	Middle East	Bedouin	C16	G18
HGDP00150	S. Asia	Makrani	C15	G19
HGDP01334	S.E. Asia	She	C15	G17
HGDP00969	N.E Asia	Yakut	C16	G17
HGDP00657	Oceania	Nan Melanesian	C18	G16

Table 6-5. CEPH DNAs sent to PNACL for sequencing analysis.

Sequencing primers used were as follows:

P7952-Seq-F: CTGAAGGGACTGCCTTGGTA (20 bp) - Tm 64.1°C

P7952-Seq-R: CTAAGCTGCACTCCCTCTGC (20 bp) – Tm 64.2°C

The section of DNA sequenced was between the purple primer set (as seen in Table 6-4 above). PCR product size is 350bp long and contains 4 unverified SNPs. Figure 6-8 below shows the DNA segment. PCR conditions were as follows: PCR programme "PNACL1" – 96 ºC for 1 minute, followed by 96 ºC for 20 seconds, 68 ºC for 30 seconds, 68 ºC for 30 seconds for a total of 30 cycles ending at 66 ºC for 1 minute and 68 ºC for 5 minutes.

Figure 6-8. . Sequenced DNA segment.
Primers are highlighted in grey, SNPs are highlighted in yellow and are as follows: rs267256, rs35383157, rs34581796, and rs10700258.

The PCR products were cleaned with a Qiagen Qiaquick PCR Purification kit (see Chapter 2, Materials and Methods) and DNA was run on a gel in order to quantify the amount of DNA present. Cleaned PCR product was sent to PNACL at a concentration between 5 – 10 ng/ µl. The sequencing primers as above were sent to PNACL at a concentration of 100 µM.

6.5.1 Populations typed

It was intended to type all the SNPs within PHAX 7952 on the Danish and Inuit DNA samples provided by Søren Nørby, on the African-Caribbean DNA samples which were already held in the laboratory, and also on the Cornish samples provided by Andy Demaine. Please see Chapter 2 – Materials and methods for full details on these DNA samples.

6.5.2 Problems encountered

Due to significant problems encountered when trying to get the SNaPshot to work reliably on the triplex PCR DNA products, it was decided to carry out the initial PCR reactions separately and carry out the SNaPshot PCRs solely on the SNPs contained within each of those DNA amplicons. Although this would initially take longer, it would provide a more accurate set of results. It was hoped that the individual PCR products could be pooled together for the SNaPshot reaction, however, this proved to be impossible to do as the individual SNaPshot products were not able to be resolved clearly enough to be able to discern the different SNP states using GeneMapper software.

The decision was therefore made to treat each of the initial PCR products separately and carry out separate SNaPshot reactions on each of these with the hope that this would yield useful results.

6.6 Results

Only PHAX 7952 (as can be seen in Figure 6-9 below) contained a forensic STR which sat entirely within the PHAX and which produced a haplotype with a low recombination rate and sufficient typed SNPs to warrant further investigation and typing of the additional SNPs.

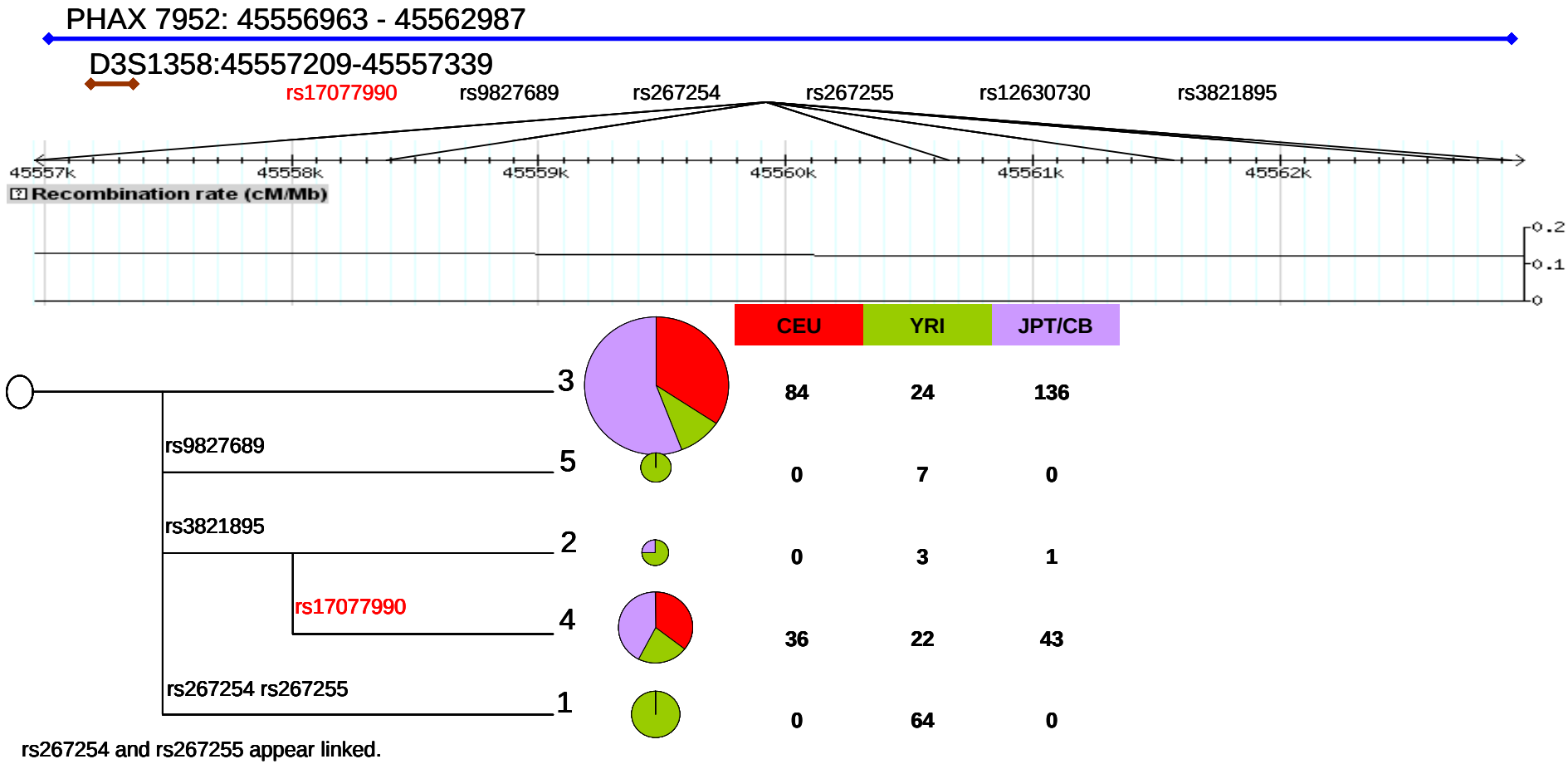


Figure 6-9. PHAX 7952 containing D3S1355

The HapMap SNPs are shown, as is the low average recombination rate (approx. 0.15 cM/ Mb) that covers this region. The pie charts show the haplogroups and their relative population frequencies. Haplogroup 3, which represents the ancestral state is the commonest haplogroup.

6.6.1 Triplex PCR

The initial triplex PCR using the primers from table 6-3 was successful and produced three distinct bands of DNA when ran on a 0.8% (w/ v)agarose gel .

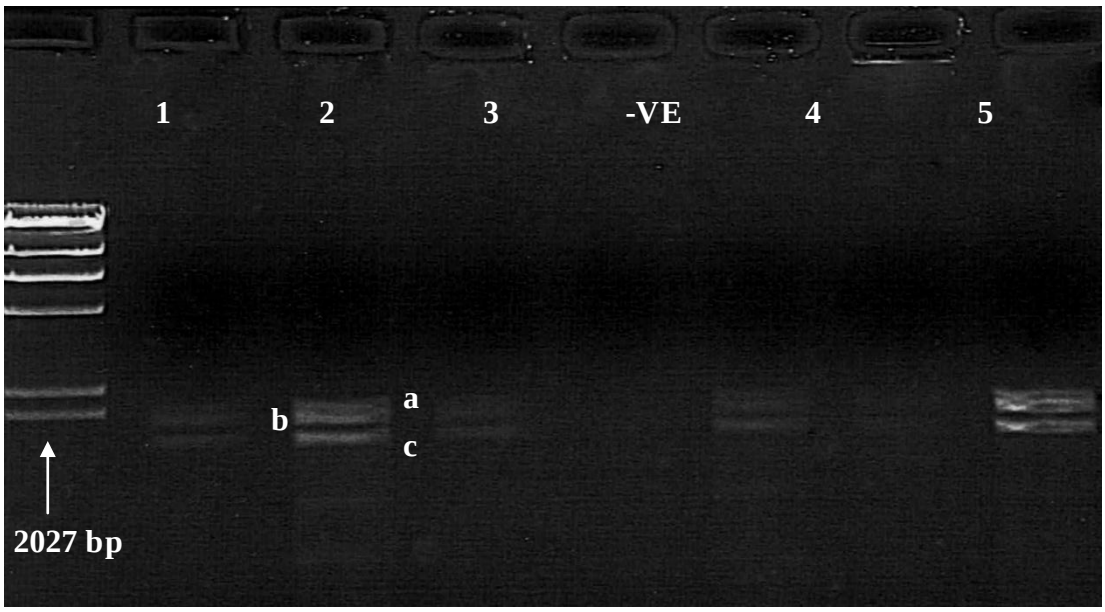


Figure 6-10. PHAX 7952 Triplex PCR .
The products produced from top to bottom show the DNA from: a) the Purple primers - 1994 bp, b) The Blue2 primers - 1868 bp and, c) the Green primers - 1683 bp. The DNA size marker used is the Lambda DNA/ *Hin*dIII Marker, with sizes from top to bottom: 23130, 9416, 6557, 4361, 2322, 2027, (and not visible; 564, 125). The DNA used (No's 1 – 5) were controls from laboratory members.

In order to be certain that the top band in the image above really represented two different bands, the same PCR was carried out but not as a triplex PCR and the resulting products were run on a 0.8% agarose gel. The results are seen in Figure 6-11, below.

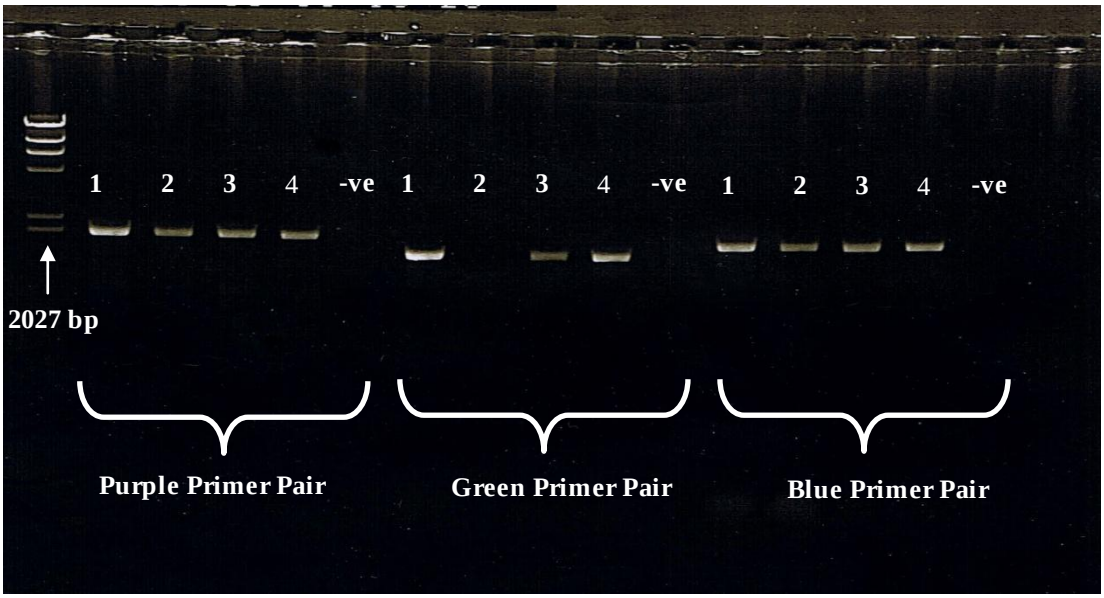


Figure 6-11. PHAX 7952 SNaPshot flanking PCR using Purple, Green and Blue Primer sets. The DNA samples No's 1 -4 are from laboratory members**.**

6.6.2 SNaPshot PCR

Following cleanup of the triplex and/ or individual PCR products with SAP and Exo1 (as per Materials and Methods, Chapter 2), the SNaPshot PCR reaction was carried out with the primers from set 1 (Table 6-4 above). Unfortunately, it proved to be impossible to get this reaction to work on the full primer set, and therefore, the initial PCRs were carried out individually using the blue, green and purple primer sets and then the set 1 SNaPshot primers were also separated and run only on the appropriate DNAs. This meant that it would be easier to distinguish the resulting DNA fragments when ran on the ABI3100 and analysed in GeneMapper software.

Examples of the resulting traces can be seen in Figures 6-12 and 6-13.

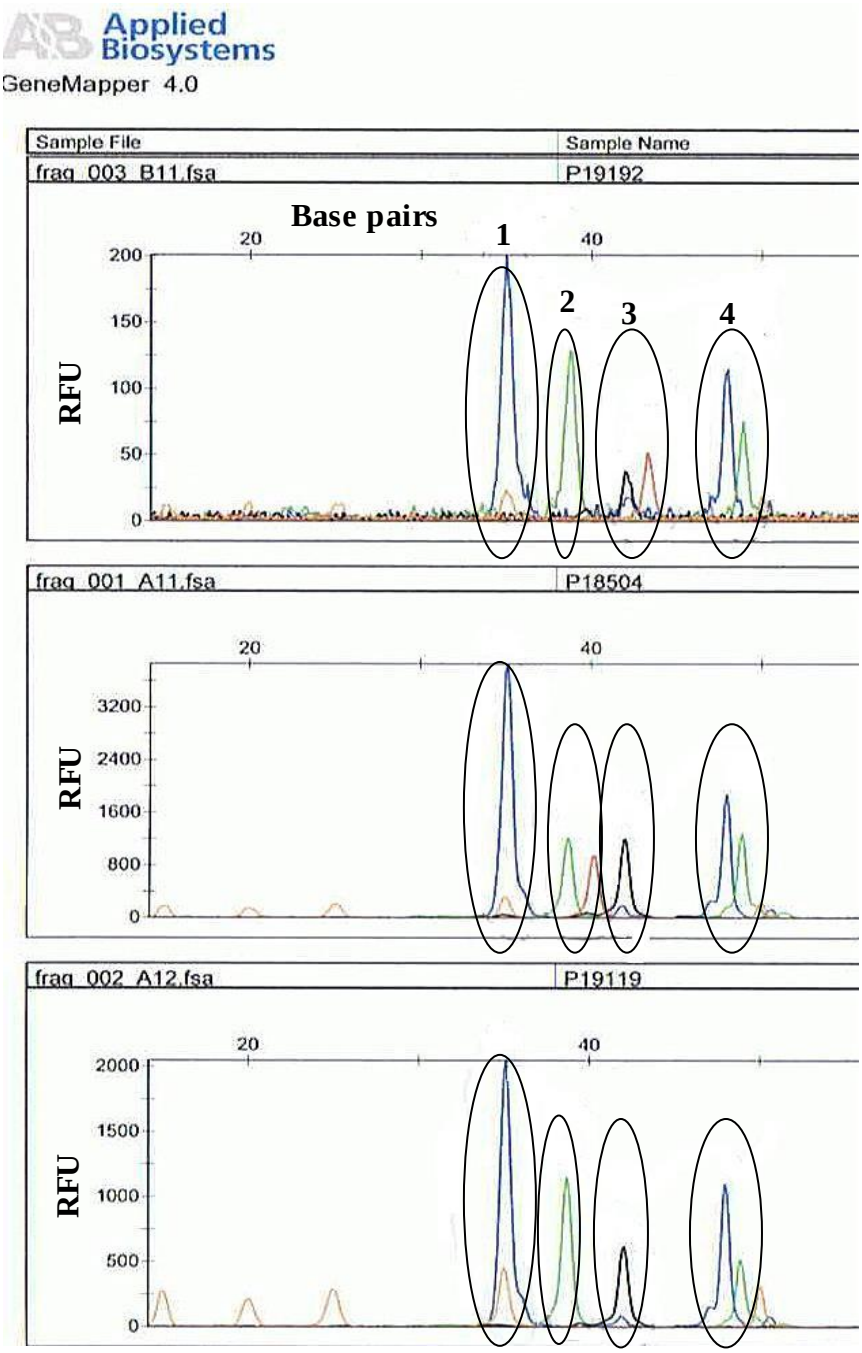


Figure 6-12. GeneMapper trace of the SNaPshot SNPs present in the DNA resulting from a PCR using the purple primer pair.

Trace colours represent the following SNP sequence states: black – C, red – T, green – A, blue – G. The DNAs used in this assay are: 19192, 18504 and 19119 from the HapMap samples. The first SNP circled is rs12630730, using primer P7952-S1-6-25, and all samples are homozygous for the G-allele. The second SNP circled is rs267255, using primer P7952-S1-4. Samples P19192 and P19119 are homozygous for the A-allele, P18504 is heterozygous for A- and T-alleles. The third SNP circled is rs267256, using primer P7952-S1-5. Sample P19192 is heterozygous for the C- and T-alleles, and samples 18504 and 19119 are homozygous for the C-allele. The fourth SNP circled is rs3821895, using primer P7952-S1-7. All three of the samples are heterozygous for the G- and A-alleles.

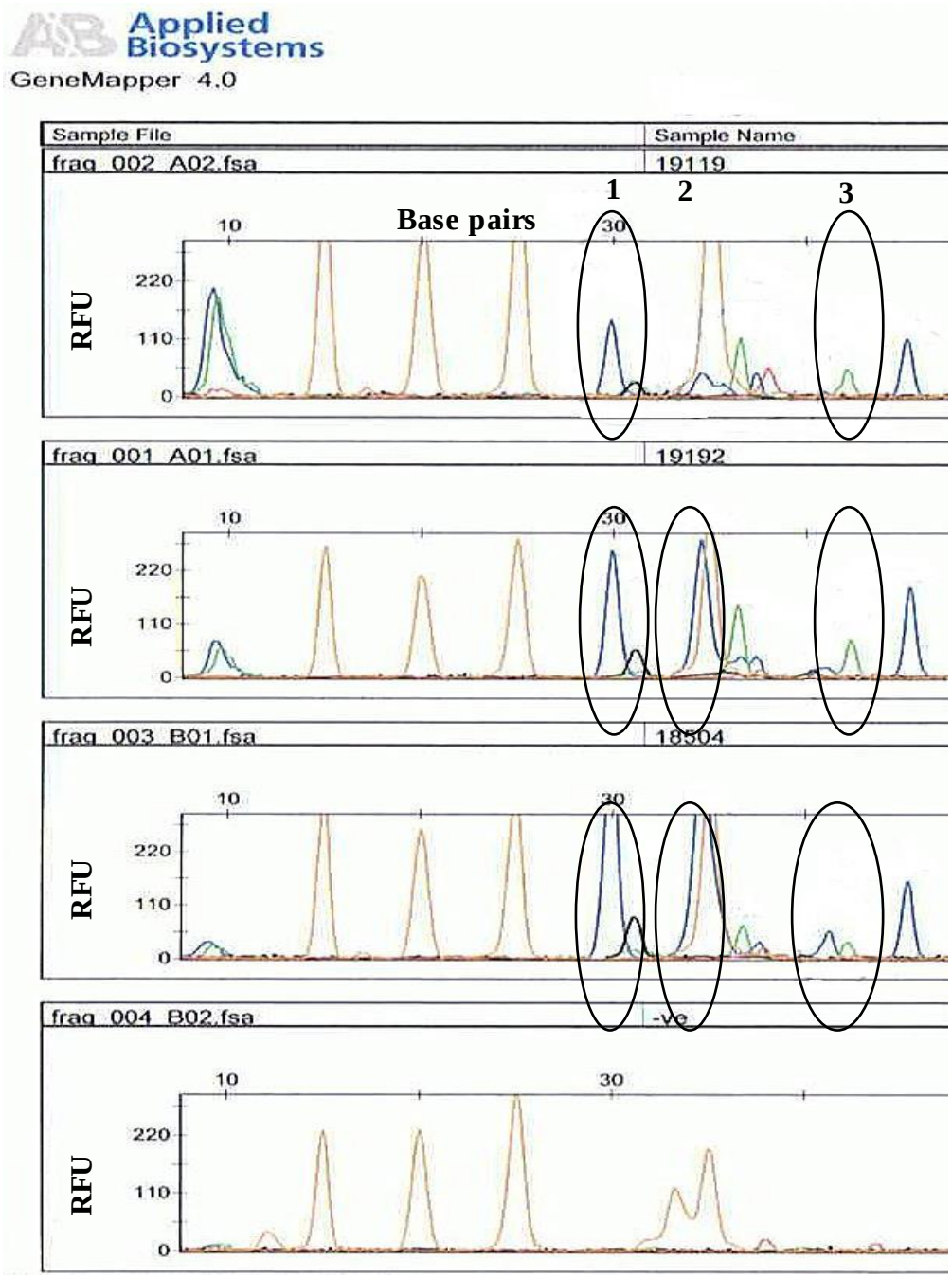


Figure 6-13. GeneMapper trace of the SNaPshot SNPs present in the DNA resulting from a PCR using the blue and green primer pairs.
SNP sequence state trace colours are: black – C, red – T, green – A, blue – G. The DNAs used are: 19119,19192 and18504 and from the HapMap samples, and a negative control. The first SNP circled: rs17077990, primer P7952-S1-1-20, all samples heterozygous for the G- and C-alleles. The second SNP circled: rs9827689, primer P7952-S1-2. Sample 191199 did not show any significant peaks, but P19192 and P18504 are homozygous for the G-allele. The third SNP circled: rs267254, primer P7952-S1-3-40bp. P19119 and P19192 are homozygous for the A-allele, and sample 18504 is heterozygous for the G- and A-allele. The peaks shown on the fourth trace for the Liz120 size standard are placed at 15, 20, 25 and 35 base pairs.

From Figure 6-12 and Figure 6-13, it can be seen that the results obtained were not reliable, as there were many additional peaks which would be easy to misinterpret if a section of DNA with unknown SNPs was interrogated. Partly for this reason, and after many trials and testing of different combinations of SNaPshot primers, it was decided to run each set separately.

6.6.3 SNaPshot on African Caribbean DNA using the purple primer set

Only 42 out of 111 DNA samples actually produced results that were able to be read with confidence. These results are shown in the table below.

Sample No.	SNP rs12630730 Primer: S1-6-25 Alleles: **C**/ T	SNP rs267255 Primer: S1-4-35 Alleles: **A**/ T	SNP rs267256 Primer: S1-5-40 Alleles: **G**/ A	SNP rs3821895 Primer: S1-7-50 Alleles: **T**/ C
28	CC	AT	GG	CT
4	CC	AT	GG	TT
122	CC	AT	GG	CT
3	CC	AT	GG	CT
29	CC	AA	AA	TT
30	CC	AT	GA	TT
6	CC	AA	GG	TT
162	CC	TT	GG	TT
164	CC	AT	GA	TT
165	CC	AA	GG	CT
74	CC	AT	GG	TT
73	CC	AT	GG	TT
9	CC	TT	GG	TT
22	CC	AA	GA	TT
23	CC	TT	GG	TT
36	CC	AT	GG	CT
10	CC	AA	GG	CT
24	CC	AA	GG	TT
11	CC	AT	GA	TT
155	CC	AT	GG	CT
26	CC	AA	AA	TT
38	CC	AT	GG	TT
354	CC	AA	GG	CT

185	CC	AT	GG	TT
186	CC	AA	GG	CT
232	CC	AT	GG	CT
234	CC	AA	GG	CT
188	CC	AA	GG	CT
189	CC	AA	GA	CT
356	CC	TT	GG	TT
230	CC	AA	AA	TT
187	CC	AT	GG	TT
-ve				

Table 6-6. SNPs in the African Caribbean DNA samples using the primers within the purple primer set only.

The data in Table 6-6 was used to investigate the haplogroups formed. These can be seen in Figure 6-14.

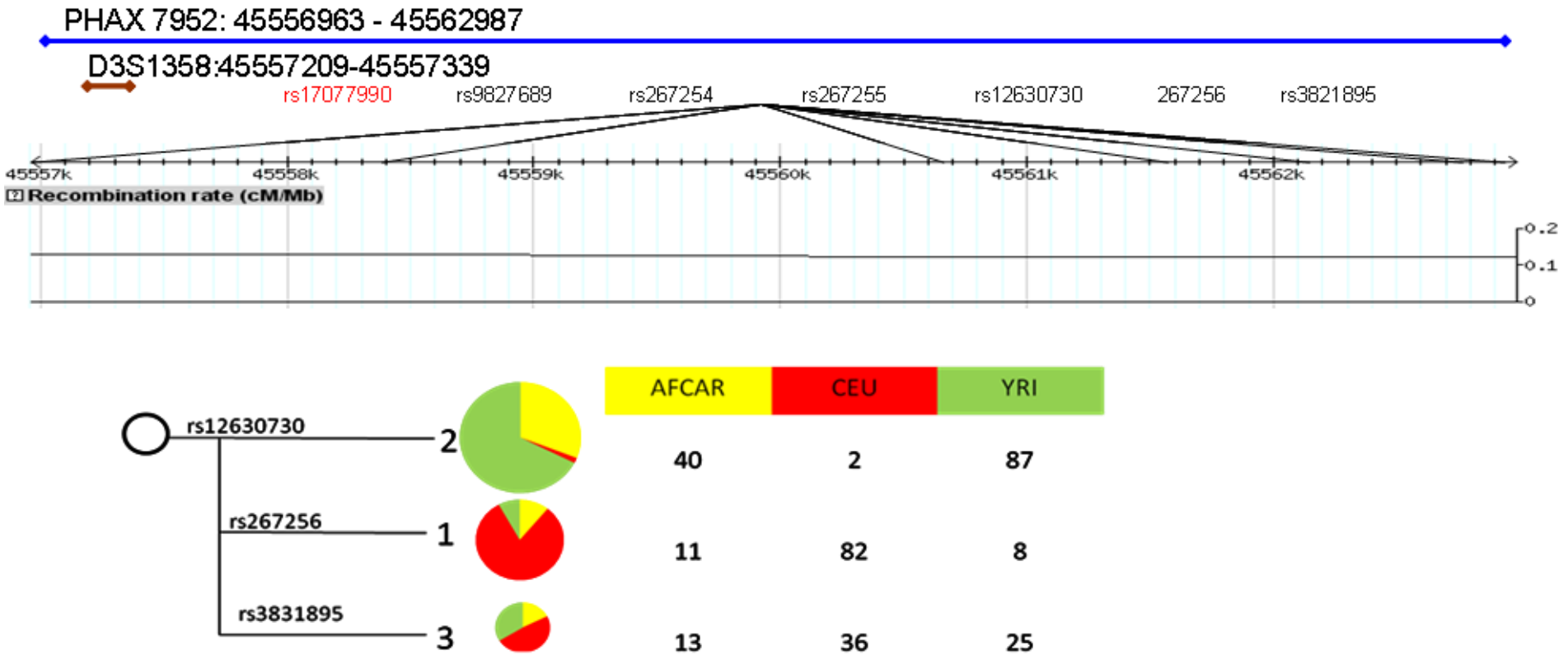


Figure 6-14. Haplogroups formed using the African Caribbean SNaPshot results. This shows comparisons with HapMap European and Yoruban SNP data. The haplotype tree and corresponding pie charts show the possible haplogroups. The pie chart sizes are relative to the numbers of individuals forming that particular haplogroup in the PHAX. The blue line indicates the position of the PHAX, and the brown line indicates the position of the D3S1358 STR. The average recombination rate in this area is low (approximately 0.15cM/ Mb).

From the figure above, it can be seen that the ancestral haplogroup predominantly comprises of the African Caribbean and Yoruban populations. This haplogroup accounts for 42% of the total DNAs sampled. The second commonest haplogroup is formed by the addition of SNP rs267256 and this one comprises mostly of the European DNAs (approximately 80%). The African Caribbean and Yoruban DNAs form roughly an equal share of the remainder of this haplogroup. The rarest haplogroup, comprising of only 24% of all the DNAs was split 50% belonging to the European DNAs and the remaining 50% split almost equally between the African Caribbean and Yoruban DNAs.

The indications from these very limited results set suggest that the African Caribbean DNAs show a high affinity with those from Yoruba. This can even be seen in rarer haplogroups, such as haplogroup 3, where the split between Yoruban and African Caribbean DNAs is almost equal. It would have been more interesting to carry this out using some of the CEPH-HGDP populations, as these more closely reflect the haplotypes present in European and African Caribbean peoples. However, at this time, the information on those DNAs is not yet available.

Unfortunately, it proved to be impossible to get any further results after this. This was due to a serious contamination problem which arose in the initial PCR, followed by a complete inability to get the PCR to work at all for a period of about 6 weeks. When the PCR problem did start to yield results again, and the product was cleaned, it failed to work when the SNaPshot PCR was carried out

on it, even when this had proved to be working properly initially. At this point, time ran out for the completion of laboratory work and it was unable to be completed.

6.6.4 Sequenced DNA by PNACL

The results from the DNA sent to PNACL for sequencing are listed in Table 6-7 below.

CEPH No	Sub Continent	rs267256	rs35383157	rs34581796	rs10700258
HGDP00808	N. Europe	G	Deletion (-)	Deletion (-)	T
HGDP01384	E. Europe	G	Deletion (-)	Deletion (-)	T
HGDP01374	S. Europe	G	Deletion (-)	Deletion (-)	T
HGDP00858	N. America	G	Deletion (-)	Deletion (-)	T
HGDP00706	S. America	G	Deletion (-)	Deletion (-)	T
HGDP01408	Sub Saharan Africa	G	Deletion (-)	Deletion (-)	T
HGDP01279	N. Africa	G	Deletion (-)	Deletion (-)	T
HGDP00631	Middle East	G	Deletion (-)	Deletion (-)	T
HGDP00150	S. Asia	G	Deletion (-)	Deletion (-)	T
HGDP01334	S.E. Asia	G	Deletion (-)	Deletion (-)	T
HGDP00969	N.E Asia	G	Deletion (-)	Deletion (-)	T
HGDP00657	Oceania	G	Deletion (-)	Deletion (-)	T

Table 6-7. Sequencing results of SNPs on populations from the CEPH-HGDP panel.

Although the sequencing was successful and the data were obtained for the SNPs which had not been typed by HapMap in the above populations, it was

not possible to ascertain whether the SNPs were homozygous or heterozygous due to possible heterozygosity for single base deletions and also a higher background noise on the sequencing trace.

An example of the sequencing trace achieved can be seen in Figure 6-15 below. This trace shows the last three SNPs as the first SNP was located outside of where the forward primer bound to the DNA supplied to PNACL. This trace clearly shows the deletion SNPs which are the first two shown below. As can be seen for the third SNP, it appears to be homozygous for the T allele, which represents the ancestral state for this allele.

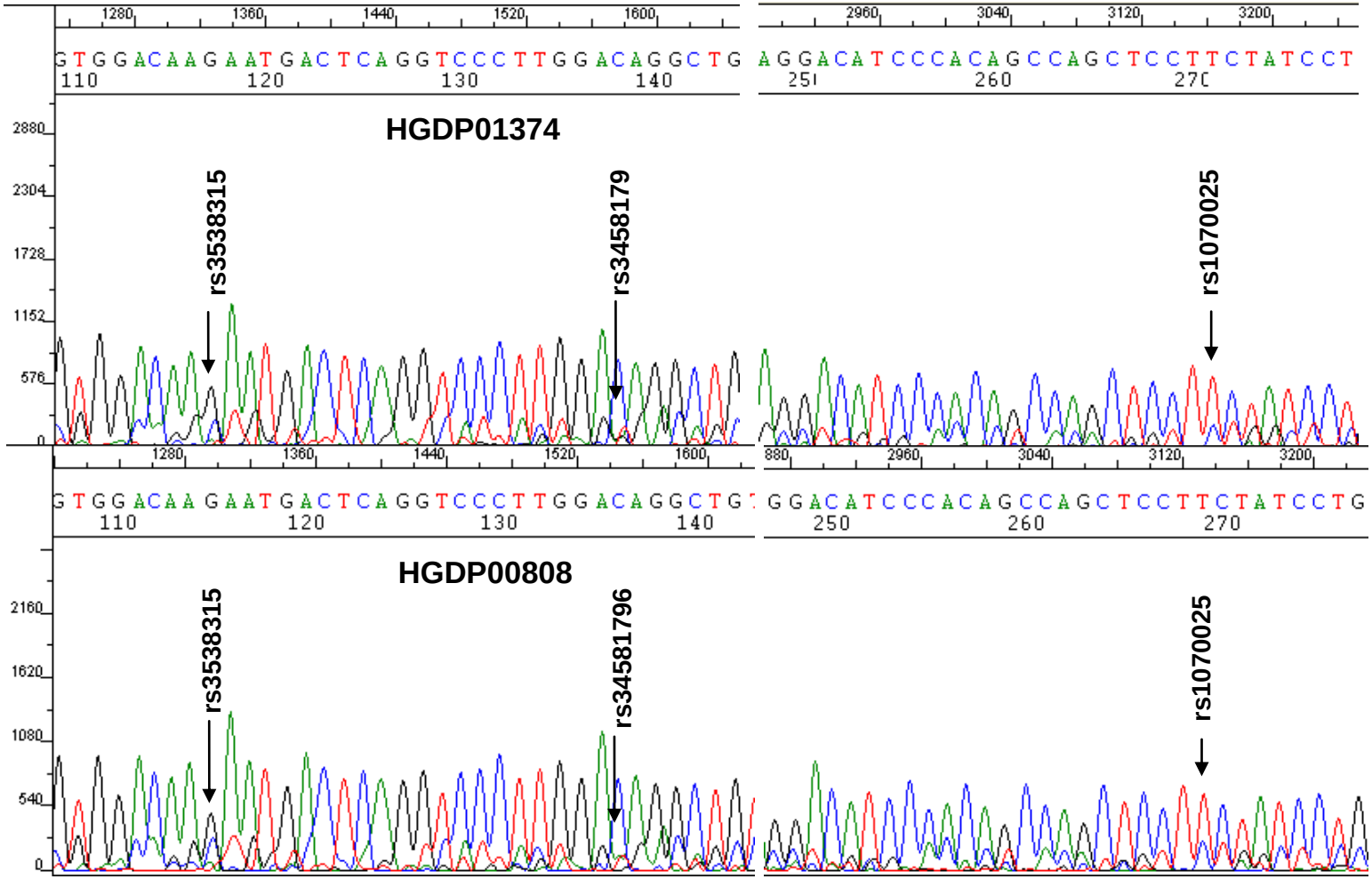


Figure 6-15. Sequence trace of three SNPs within PHAX7952 as a result of PNACL sequencing. Only the forward trace is shown here for both samples.

6.7 Discussion

Although the initial work on the PHAXs looked very promising and it certainly would be possible, in theory, to turn the SNPSTRs into larger haplotype blocks using the forensic STRs and the SNaPshot approach, however, in practice this did not work as planned in this case. The small number of samples which did work from the African Caribbean DNAs yielded insufficient data to analyse in any meaningful way, as there was no other equivalent population data to compare them to.

In the end, this approach takes a great deal of time and effort to get all of the SNPs separated equally and amplified in such a way as to be able to be resolved using GeneMapper software.

A much more suitable, and in the end, reliable method to obtain this type of population data for SNPs would be make use of some of the new SNP typing technologies present today, for example, a Gene Chip. And this is currently being done for a number of SNPs (including those in PHAX 7952) on the CEPH-HGDP DNA panel which will be analysed on a GoldenGate® Illumina® (Illumina Inc, San Diego, USA) 1536 custom SNP technology platform (Lynch et al. 2009). This SNP typing will be carried out in-house later this year. It would have been interesting to get this data in time for this project, however, as this has not been possible, the question about whether PHAXs would be informative population makers remains unanswered for now.

What is currently available is the data for the some of the SNPs within PHAX 7952 (see Figure 6-16). Even though it was not possible to type the SNPs using the SNaPshot reaction as had been hoped, it is possible to examine the STR diversity based on the HapMap SNP data for this PHAX.

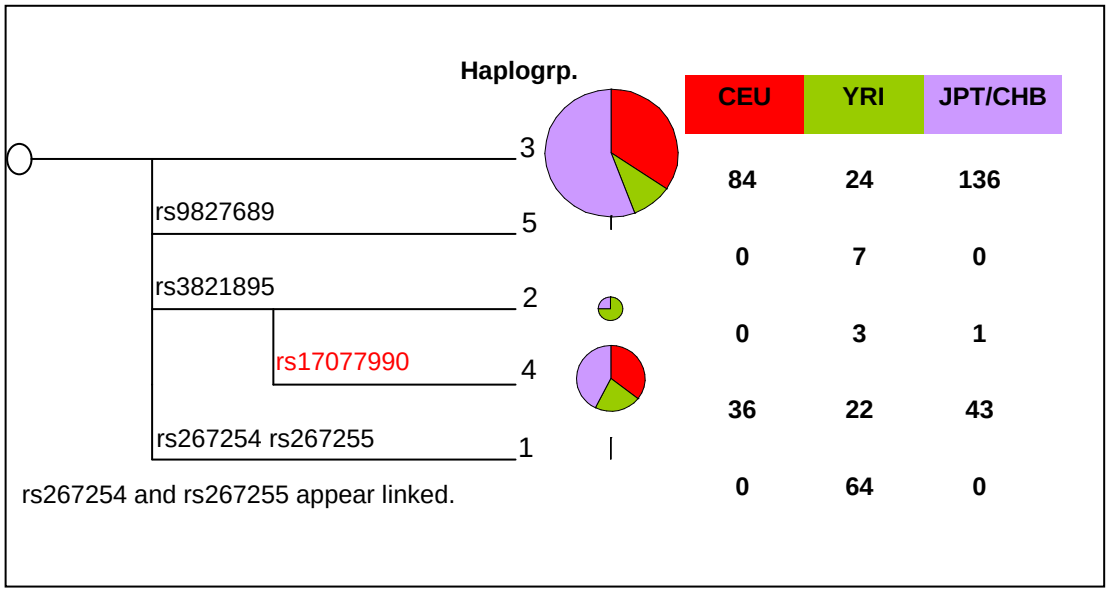


Figure 6-16. PHAX 7952 containing D3S1355
The pie charts show the haplogroups and their relative population frequencies. Pie chart sizes are relative to the numbers of individuals forming that particular haplogroup in the PHAX**.** Haplogroup 3, (the ancestral state) is the commonest haplogroup. Haplogroup 4 contains the SNP which was typed as part of the D3S1355 SNPSTR.

The figure above shows the haplogroups that are formed by the HapMap SNPs within PHAX 7952. As the STR variation in combination with rs17077990 is already known because it was typed as part of the D3 SNPSTR, the remaining haplogroups which are formed by the different SNPs can be further subdivided by examining STR variation. However, only those for which the phasing was known were included. Those individuals who did not belong to haplogroup 4, and where the SNP and STR were both heterozygous were therefore excluded from the results. After this, there remained no individuals which could be

further subdivided for Haplogroup 2. The resulting bar charts can be seen in the figures below.

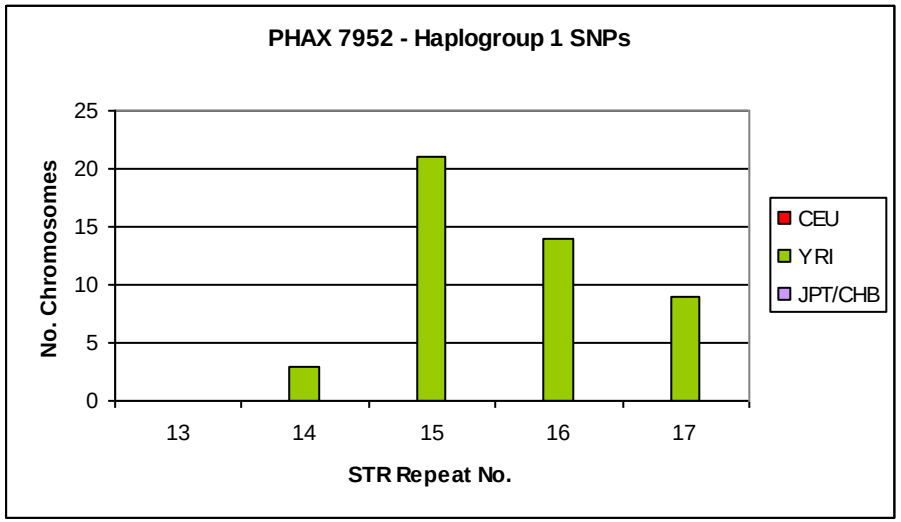
The phasing of the SNPs was already carried out by the HapMap project and data on the phased SNPs is available. Phasing was carried out by the HapMap project using the PHASE software (Stephens and Donnelly 2003), where the diploid data was converted into haplotypes. These phased haplotypes are downloadable from the HapMap website, and form the basis of the PHAXs discussed in this chapter.

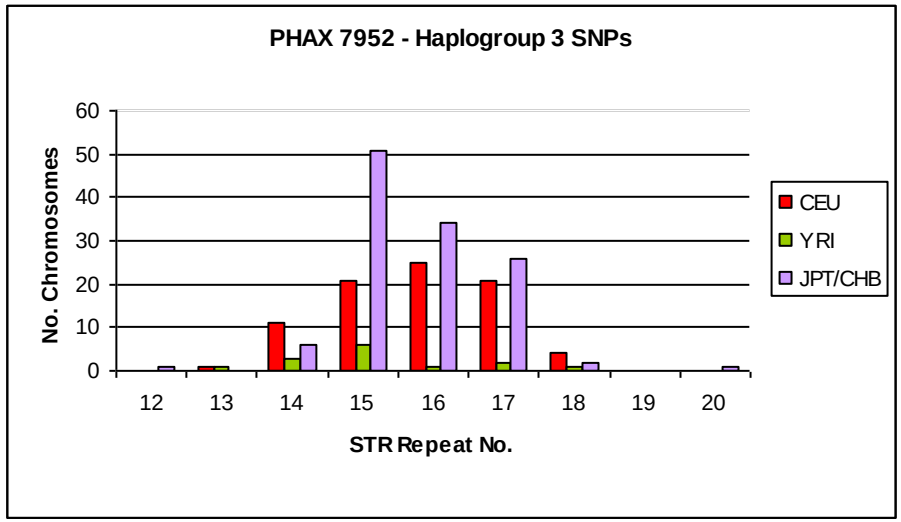
The issue becomes more complex when the STR data is also taken into consideration because it is not possible to phase STR data in the same way as SNPs based on allele frequencies for example, because the mutation rate of STRs is too high.

It is only possible to fully phase the SNPs and the STR within a PHAX when;

- the STR is homozygous, or
- the STR is heterozygous, but the SNP is also heterozygous.

It is not possible to phase the SNPs with the STR within the PHAX when the SNP is homozygous.





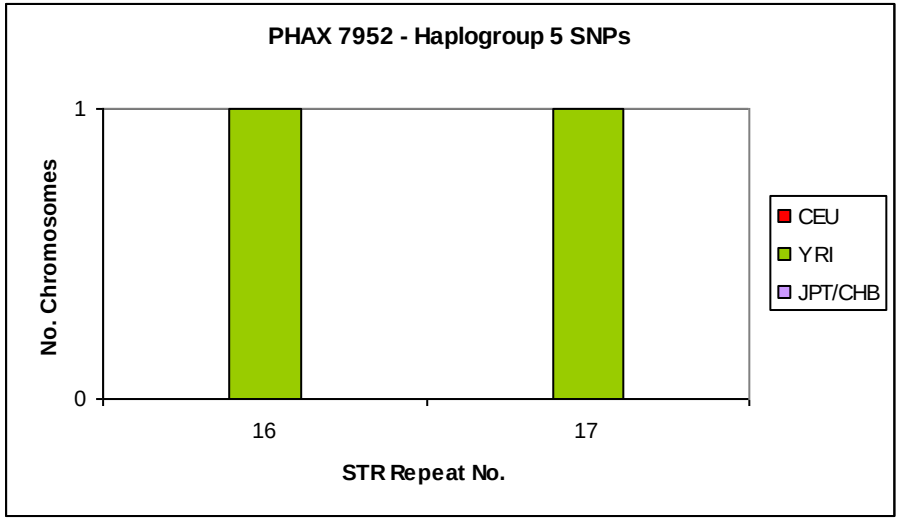


Figure 6-17. STR variation on the HapMap SNPs for PHAX 7952.

From the charts above, it can be seen that the distribution of STR alleles for haplogroups 1 and 5 are not even, despite the fact that both haplogroups are only present in the YRI samples. For haplogroup 5, there were only 2 chromosomes which does not provide sufficient details for meaningful comparisons or conclusions. Haplogroup 1 does show more variation around the STR, which may suggest that the linking of SNPs rs267254 and rs267255 which form this haplogroup, and which only occurs in the YRI, is older than rs9827689 which forms haplogroup 5. If the SNP determining haplogroup 5 is more recent, then this would also explain the lack of STR variation and low sample numbers.

Haplogroup 3, which represents the ancestral state, perhaps not surprisingly, shows more diversity around the STR and also includes all of the HapMap DNA populations. What is interesting is that the largest STR range (12 to 20 repeats) is found in the CHB/ JPT samples and that the YRI samples are not well represented here, although the range of the YRI samples is between 13 and 18 repeats, which is the same as the CEU samples. As there are many more individuals in the CHB/ JPT samples, and a wider STR range, it suggests that ancestral state of this haplogroup originated in China or Japan, or it may also be as a result of the way in which the DNA samples were collected.

It would have been useful to have the data available from the gene chip analysis of the CEPH-HGDP as this may have provided more insight into the subdivisions of the haplogroups within PHAX 7952.

7 Discussion

This thesis set out to establish whether SNPSTRs could be useful as markers for population structure and population histories. This was investigated by basing the SNPSTRs on the well characterized forensic STRs in the PowerPlex 16 kit and then finding nearby (within 500bp) SNPs which all had minor allele frequencies of ≥ 0.1 in the HapMap populations. Any SNPs below that value were not used as they were population-specific or fixed in certain populations and would therefore not be as informative in the first SNPSTR system trials. However, population-specific SNPs were included in the analysis in the PHAX work (Chapter 6), as they would be informative in non-HapMap populations.

The SNPSTR systems which fitted all the criteria were based on only four of the forensic STRs: D5S818, D16S539, D3S1358 and CSF1PO.

These four SNPSTRs were then typed on DNAs from the CEPH-HGDP (Cann et al. 2002), HapMap (HapMap 2003), Cornwall (Richards et al. 2000), UK African Caribbean , Greenland Inuit and Danish.

As there is already much population data available for the CEPH-HGDP and the HapMap, this would provide a basis for the evaluation of the actual SNPSTR typing methodology as well as provide additional data about the autosomal DNA.

During the course of this thesis, SNPSTRs were evaluated on two different fronts. The first was to see if they could be useful as informative markers of population histories, and secondly to ascertain if they could provide structural information about populations, such as additional information on known admixed populations, such as the Greenland Inuit.

7.1 SNPSTRs as markers for population histories

After typing the SNPSTRs on the CEPH-HGDP, and the HapMap populations, it has been shown that even with a limited number of SNPSTRs, (in this thesis, there were only four), they were still able to provide quite a lot of data about population histories, such as the TMRCA for the SNPSTRs and also data about population diversity.

The presence of only ancestral SNP alleles in some populations is also able to tell us information about past events in that population's history. For example, Oceania has only got the ancestral SNP and this SNP is mainly associated with STR allele repeats of 10 or 11, which is most likely the ancestral STR repeat number. This suggests that the Oceania D5S818 SNPSTR came from a few founder individuals and either there has not been sufficient time for a mutation to occur at the SNP, suggesting that this is a more recent arrival, or that if there initially a derived SNP, that this has been lost, perhaps through drift.

It is difficult to ascertain detailed diversity information from the SNPSTRs alone. The Y chromosome data show that Africa contains haplogroups with the

deepest rooting branches (see Figure 1-2), suggesting the more ancient lineages. Africa is dominated by haplogroup E and also contains haplogroup A, both of which are only found in Africa.

The mtDNA haplogroup tree (Figure 1-3) also shows Africa with the deeper rooting branches, containing haplogroups L1, L2 and L3 which are only found in Africa. However, unlike the Y chromosome, North Africa also contains much more recent haplogroups, in the form of haplogroup U, which is found in Europe and South East Asia as well.

If SNPSTR data is used in conjunction with data from other markers, such as the Y chromosome or mtDNA, then they become even more informative as markers and can be used to tell us about the history of autosomal DNA. In the case of Oceania, it mirrors the data from the Y chromosome, but this is not always the case. One good example where this may not be the case is in sex-biased admixed populations. Often in sex-biased admixed populations, we see the history of the Y chromosome and the mtDNA, for example in the Greenland Inuit, where the Y chromosome is typical of those from European populations (Bosch et al. 2003), but the mtDNA is more typical of that from Inuit populations.

7.2 SNPSTRs as markers for population structures

Because of the presence of many physical barriers to migration, along with a limited ability to travel, and the practice of many people of remaining in or near

their birthplace, human populations are rarely panmictic. There is usually a geographic range within which individuals are more closely related to one another than those randomly selected from the general population, and it is these behaviours which result in the genetic structure of populations.

The question arises of whether SNPSTRs are useful markers for population structures and whether this can be seen in the results achieved in this thesis.

Certainly the F_{ST} and R_{ST} data which produced the MDS plots have provided rudimentary population groupings, as can be seen in Chapter 4. Further analysis of the SNPSTR haplogroups using STRUCTURE (Pritchard et al. 2000) has enabled the detection of population isolates, such as the Surui. In this instance, the SNPSTR data provided by the D3S1358 SNPSTR showed that there is only one derived SNP allele which is only associated with a 17-repeat STR allele. (See Chapter 4).

The STRUCTURE analysis of the SNPSTR data has also provided some insights into the admixed nature of the autosomal DNA in known admixed populations such as the Greenland Inuit and to some extent, the African Caribbeans.

Extending the SNPSTRs into larger haplotype blocks such as PHAXs, as was undertaken in Chapter 6 of this thesis (and can be seen in Figure 7-1 below), provided additional population structure information.

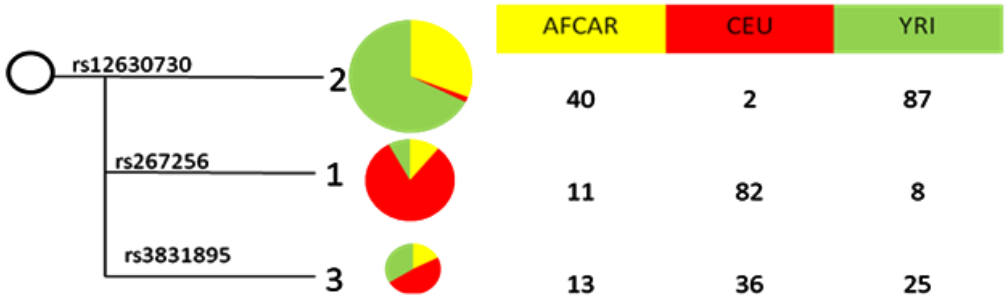


Figure 7-1. Haplogroups formed using the African Caribbean SNaPshot results. This shows comparisons with HapMap European and Yoruban SNP data. The haplotype tree and corresponding pie charts show the possible haplogroups. The pie chart sizes are relative to the numbers of individuals forming that particular haplogroup in the PHAX.

From this figure it can be seen that the PHAX data suggests that the African Caribbean DNAs show a high affinity with those from Yoruba, even in rarer haplogroups formed by the PHAX 7952 where the split between Yoruban and African Caribbean DNAs is almost equal.

Figure 7-2 highlights additional analysis of the SNPs, available from HapMap, within PHAX 7952. This shows the linking of two SNPs which only occurred in the Yorubans and also provided a timeframe in that these linked SNPs were older than SNP rs9827689 which formed haplogroup 5, which could possibly explain the lack of STR variation.

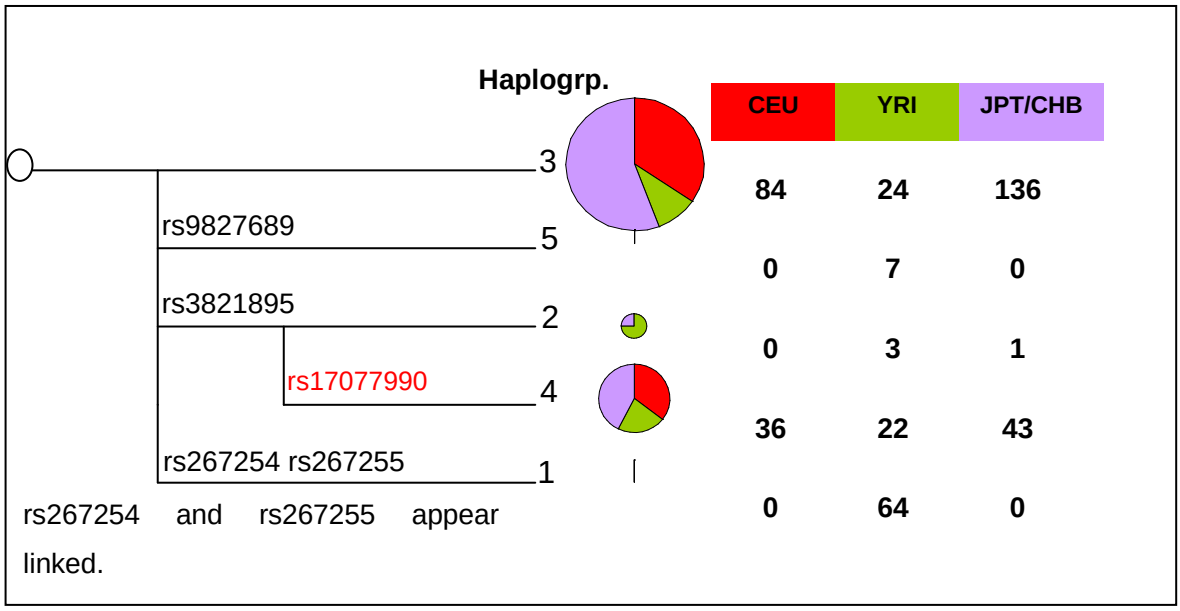


Figure 7-2. PHAX 7952 containing D3S1355
The pie charts show the haplogroups and their relative population frequencies. The pie chart sizes are relative to the numbers of individuals forming that particular haplogroup in the PHAX. Haplogroup 4 contains the SNP which was typed as part of the D3S1355 SNPSTR.

It is encouraging to see that even using only 4 SNPSTRs, there is already evidence of population structure. When one of these SNPSTRs is expanded into a larger haplotype block, further population structure information is revealed.

7.3 Future work and possible applications of SNPSTRs

7.3.1 The use of non-validated SNPs in forming SNPSTRs

It would have been interesting to look for SNPSTRs based on the forensic STRs using SNPs which were population specific, rather than those with minor allele frequencies >0.1. These SNPSTRs may have been able to provide more specific insights into intra-population structures.

Some of these alternative SNPs were intended to be included as part of the SNaPshot work in Chapter 6, however, unfortunately there was not enough time to complete this. The indications are that including both population-specific SNPs and also non-validated SNPs does provide both useful and interesting data which would have added a great deal to this thesis. It is hoped that the data provided by the GoldenGate Illumina analysis, which includes the SNPs already typed in this thesis, as well as others, will provide further insights into the potential of SNPSTRs as the basis for expanded haplotype blocks.

7.3.2 The use of alternative SNPSTRs

Although the aim of this thesis was to use the forensic STRs because of their known mutation rates, thus enabling the calculation of TMRCAs, it would be interesting to use some alternative SNPSTRs, such as those found on the SNPSTR Database for example, (http:/ / www.sbg.bio.ic.ac.uk/ ~ino/ cgi-bin/ SNPSTRdatabase.html) (Agrafioti and Stumpf 2007). This database contains SNPs and STRs which are less than 250 base pairs apart, and contains all inferable human SNPSTRs.

It would also have been interesting to develop multiplex systems in order to test for several SNPSTRs at once. Using more SNPSTRs would have provided a greater coverage and therefore more data which may have highlighted additional population structures and insights into more ancient admixture events.

7.3.3 Alternative applications for SNPSTRs

In this thesis, the emphasis was placed on SNPSTRs as markers for population histories and structures, however, there are other potential uses for SNPSTRs which were not investigated in the scope of this project.

One potential use would be as markers for disease susceptibility in much the same way that tag SNPs are used. The HapMap project (HapMap 2003), is currently working on making information on tag SNPs from the HapMap available, and this will enable researchers to use them to locate genes involved in medically important traits such as high blood pressure, or diabetes. SNPSTRs may be able to provide additional population data once a tag SNP has identified such a trait. It is already recognised that haplotype information improves the power to map disease genes (Akey et al. 2001; Ardlie et al. 2002).

Another possibility would be to use SNPSTRs for identification purposes. If more SNPSTRs were typed in a single reaction, it may be possible to use them as an alternative method for forensic or genealogical research. This is feasible because like current DNA fingerprinting, SNPSTRs only require a small amount of DNA and they have the potential to provide additional population-specific data as well as person-identification data.

Despite the plethora of SNPs that can be readily typed in a single experiment (currently ~1,000,000 for about £200), there is still promise in combining together these slow-mutating binary markers with linked rapidly mutating

STRs. If SNPSTRs are to be used in the future, the development of analytical methods that account for the different properties of the two constituent marker types is a priority. Here, ad hoc adaptations of existing methods have been used.

Abbreviations 1

List of Abbreviations

µl	microlitre
µM	micro Molar
µg	microgram
AD	Anno Domini
AMOVA	Analysis of Molecular Variance
ATP	Adenosine Tri Phospate
bp	base pairs
CEPH	Centre d'Etude du Polymorphisme Humain
CEU	CEPH European
CHB	Chinese
CNVs	Copy Number Variations
DBMs	Distance-Based Methods
ddNTPs	Dideoxy nucleoside triphosphates
DF	Degrees of Freedom
DNA	Deoxyribonucleic acid
dNTP	deoxyribonucleotide triphosphate
EDTA	Ethylenediaminetetraacetic Acid
HCl	Hydrochloric Acid
HGDP	Human Genome Diversity Project
HLA	Human Leukocyte Antigen
IBD	Isolation By Distance
JPT	Japanese
KYA	Thousand Years Ago
LCL	Lymphoblastoid Cell Line
LD	Linkage Disequilibrium
M	Molar
MDS	Multi-Dimensional Scaling
MHC	Major Histocompatibility Complex
ml	millilitre
mM	mili Molar
mtDNA	Mitochondrial DNA
NCBI	National Center for Biotechnology Information
NDS	Normal Donkey Serum
ng	nanogram

PCR	Polymerase Chain Reaction
PHAX	Phylogeographically informative Haplotypes on the Autosomes and seX chromosomes
PNACL	Protein Nucleic Acid Chemistry Laboratory
RFLP	Restriction Length Fragment Polymorphism
RNA	Ribonucleic acid
SAP	Shrimp Alkaline Phosphatase
SMM	Stepwise Mutational Model
SNP	Single Nucleotide Polymorphism
STR	Short Tandem Repeat
TBE	Tris/ Borate/ EDTA
TMRCA	Time to Most Recent Common Ancestor
Tris	Tris(hydroxymethyl)aminomethane
U	Unit
UAF	Uncorrelated Allele Frequencies
UCP	Uncoupling Protein
UCSC	University of California, Santa Cruz
UK	United Kingdom
US(A)	United States (of America)
UV	Ultra Violet
w/ v	Weight to volume
YBP	Years Before Present
YRI	Yoruban from Ibadan

Appendix 1

List of populations included in the Structure analysis of the Greenland Inuit admixture in Chapter 5.

Number	Subpopulation	Population	Continent
1	Adygei	E_Europe	Europe
2	Russian	E_Europe	Europe
3	French_Basque	S_Europe	Europe
4	French	S_Europe	Europe
5	North_Italian	S_Europe	Europe
6	Sardinian	S_Europe	Europe
7	Tuscan	S_Europe	Europe
8	Orcadian	N_Europe	Europe
8	UK	N_Europe	Europe
9	Copenhagen	N_Europe	Europe
10	Scorebysund	Greenland	Europe
11	Uummannaq	Greenland	Europe
12	Upernavik	Greenland	Europe
13	Ilulisaat	Greenland	Europe
14	Nuuk	Greenland	Europe
15	Nanortalik	Greenland	Europe
16	Karitiana	S_America	Americas
17	Surui	S_America	Americas
18	Colombian	S_America	Americas
19	Maya	N_America	Americas
20	Pima	N_America	Americas

Appendix 1. Table 7-3 List of populations included in the STRUCTURE analysis for Chapter 5.

Appendix 2

Sample Type	Sample
Panel	DS-30-3130XL-Pop7
Analysis Method	SNaPshot Default
Instrument Protocol	Snapshot-protocol-Suzanne
Size Standard	Liz120

Appendix 2. Table 7-4. ABI3100 SNaPshot default run conditions – used in Chapter 6

Dye Set	E
Mobility	DT3100 Pop4LR (BD) V1.Mob.
Project Name	3100 project1
Run Module	Ultraseq36pop4
Analysis	BC-3100 Pop4 UR

Appendix 2. Table 7-5. ABI3100 default run conditions for sequencing – used in Chapter 6

Appendix 3

YORUBA				JPT AND CHB			EUROPEAN			
Sample #	**Sex**	**Fam**	**Trio**	**Sample #**	**Sex**		**Sample #**	**Sex**	**Fam**	**Trio**
NA19238	♀	Y117	M	NA18526	♀	Chinese	NA12878	♀	1463	M
NA19240	♀	Y117	C	NA18529	♀	Chinese	NA12892	♀	1463	MgM
NA19239	♂	Y117	F	NA18532	♀	Chinese	NA12891	♂	1463	MgF
NA19193	♀	Y112	M	NA18537	♀	Chinese	NA12873	♀	1459	FgM
NA19192	♂	Y112	F	NA18540	♀	Chinese	NA12865	♀	1459	M
NA19194	♂	Y112	C	NA18542	♀	Chinese	NA12875	♀	1459	MgM
NA19099	♀	Y105	M	NA18545	♀	Chinese	NA12872	♂	1459	FgF
NA19100	♀	Y105	C	NA18547	♀	Chinese	NA12874	♂	1459	MgF
NA19098	♂	Y105	F	NA18550	♀	Chinese	NA12864	♂	1459	F
NA19131	♀	Y101	M	NA18552	♀	Chinese	NA12813	♀	1454	FgM
NA19132	♀	Y101	C	NA18555	♀	Chinese	NA12802	♀	1454	M
NA19130	♂	Y101	F	NA18564	♀	Chinese	NA12815	♀	1454	MgM
NA19127	♀	Y077	M	NA18566	♀	Chinese	NA12812	♂	1454	FgF
NA19129	♀	Y077	C	NA18570	♀	Chinese	NA12814	♂	1454	MgF
NA19128	♂	Y077	F	NA18571	♀	Chinese	NA12801	♂	1454	F
NA19143	♀	Y074	M	NA18573	♀	Chinese	NA12761	♀	1447	FgM
NA19144	♂	Y074	F	NA18576	♀	Chinese	NA12753	♀	1447	M
NA19145	♂	Y074	C	NA18577	♀	Chinese	NA12763	♀	1447	MgM
NA19152	♀	Y072	M	NA18579	♀	Chinese	NA12760	♂	1447	FgF
NA19153	♂	Y072	F	NA18582	♀	Chinese	NA12762	♂	1447	MgF
NA19154	♂	Y072	C	NA18592	♀	Chinese	NA12752	♂	1447	F
NA19140	♀	Y071	M	NA18593	♀	Chinese	NA12740	♀	1444	M
NA19141	♂	Y071	F	NA18594	♀	Chinese	NA12751	♀	1444	MgM
NA19142	♂	Y071	C	NA18942	♀	Japanese	NA12750	♂	1444	MgF
NA19116	♀	Y060	M	NA18947	♀	Japanese	NA12004	♀	1420	FgM
NA19119	♂	Y060	F	NA18949	♀	Japanese	NA12006	♀	1420	MgM
NA19120	♂	Y060	C	NA18951	♀	Japanese	NA10839	♀	1420	M
NA19222	♀	Y058	M	NA18956	♀	Japanese	NA12003	♂	1420	FgF
NA19221	♀	Y058	C	NA18964	♀	Japanese	NA12005	♂	1420	MgF
NA19223	♂	Y058	F	NA18968	♀	Japanese	NA10838	♂	1420	F
NA19159	♀	Y056	M	NA18969	♀	Japanese	NA12249	♀	1416	FgM
NA19160	♂	Y056	F	NA18972	♀	Japanese	NA12248	♂	1416	FgF
NA19161	♂	Y056	C	NA18973	♀	Japanese	NA10835	♂	1416	F
NA19206	♀	Y051	M	NA18975	♀	Japanese	NA12236	♀	1408	FgM
NA19207	♂	Y051	F	NA18976	♀	Japanese	NA12156	♀	1408	MgM
NA19208	♂	Y051	C	NA18978	♀	Japanese	NA10831	♀	1408	M
NA19209	♀	Y050	M	NA18980	♀	Japanese	NA12154	♂	1408	FgF
NA19210	♂	Y050	F	NA18981	♀	Japanese	NA12155	♂	1408	MgF
NA19211	♂	Y050	C	NA18987	♀	Japanese	NA10830	♂	1408	F
NA19204	♀	Y048	M	NA18991	♀	Japanese	NA12234	♀	1375	MgM
NA19203	♂	Y048	F	NA18992	♀	Japanese	NA10863	♀	1375	M

NA19205	♂	Y048	C	NA18997	♀	Japanese	NA12264	♂	1375	MgF
NA19172	♀	Y047	M	NA18998	♀	Japanese	NA11995	♀	1362	MgM
NA19171	♂	Y047	F	NA18999	♀	Japanese	NA11993	♀	1362	FgM
NA19173	♂	Y047	C	NA19003	♀	Japanese	NA10861	♀	1362	M
NA19201	♀	Y045	M	NA18524	♂	Chinese	NA11992	♂	1362	FgF
NA19202	♀	Y045	C	NA18558	♂	Chinese	NA11994	♂	1362	MgF
NA19200	♂	Y045	F	NA18561	♂	Chinese	NA10860	♂	1362	F
NA19137	♀	Y043	M	NA18562	♂	Chinese	NA12717	♀	1358	FgM
NA19138	♂	Y043	F	NA18563	♂	Chinese	NA12716	♂	1358	FgF
NA19139	♂	Y043	C	NA18572	♂	Chinese	NA12707	♂	1358	F
NA19102	♀	Y042	M	NA18603	♂	Chinese	NA11830	♀	1350	FgM
NA19101	♂	Y042	F	NA18605	♂	Chinese	NA11832	♀	1350	MgM
NA19103	♂	Y042	C	NA18608	♂	Chinese	NA10855	♀	1350	M
NA19093	♀	Y040	M	NA18609	♂	Chinese	NA11831	♂	1350	MgF
NA19094	♀	Y040	C	NA18611	♂	Chinese	NA11829	♂	1350	FgF
NA19092	♂	Y040	F	NA18612	♂	Chinese	NA10856	♂	1350	F
NA18912	♀	Y028	M	NA18620	♂	Chinese	NA11840	♀	1349	MgM
NA18913	♂	Y028	F	NA18621	♂	Chinese	NA10854	♀	1349	M
NA18914	♂	Y028	C	NA18622	♂	Chinese	NA11839	♂	1349	MgF
NA18861	♀	Y024	M	NA18623	♂	Chinese	NA11882	♀	1347	MgM
NA18862	♂	Y024	F	NA18624	♂	Chinese	NA10859	♀	1347	M
NA18863	♂	Y024	C	NA18632	♂	Chinese	NA11881	♂	1347	MgF
NA18855	♀	Y023	M	NA18633	♂	Chinese	NA12044	♀	1346	FgM
NA18856	♂	Y023	F	NA18635	♂	Chinese	NA12043	♂	1346	FgF
NA18857	♂	Y023	C	NA18636	♂	Chinese	NA10857	♂	1346	F
NA18852	♀	Y018	M	NA18637	♂	Chinese	NA07345	♀	1345	MgM
NA18853	♂	Y018	F	NA18940	♂	Japanese	NA07348	♀	1345	M
NA18854	♂	Y018	C	NA18943	♂	Japanese	NA07357	♂	1345	MgF
NA18870	♀	Y017	M	NA18944	♂	Japanese	NA12057	♀	1344	FgM
NA18871	♂	Y017	F	NA18945	♂	Japanese	NA12056	♂	1344	FgF
NA18872	♂	Y017	C	NA18948	♂	Japanese	NA10851	♂	1344	F
NA18523	♀	Y016	M	NA18952	♂	Japanese	NA07055	♀	1341	FgM
NA18522	♂	Y016	F	NA18953	♂	Japanese	NA06985	♀	1341	MgM
NA18521	♂	Y016	C	NA18959	♂	Japanese	NA06991	♀	1341	M
NA18517	♀	Y013	M	NA18960	♂	Japanese	NA06993	♂	1341	MgF
NA18516	♂	Y013	F	NA18961	♂	Japanese	NA07034	♂	1341	FgF
NA18515	♂	Y013	C	NA18965	♂	Japanese	NA07048	♂	1341	F
NA18858	♀	Y012	M	NA18966	♂	Japanese	NA07000	♀	1340	FgM
NA18859	♂	Y012	F	NA18967	♂	Japanese	NA07056	♀	1340	MgM
NA18860	♂	Y012	C	NA18970	♂	Japanese	NA07019	♀	1340	M
NA18508	♀	Y009	M	NA18971	♂	Japanese	NA06994	♂	1340	FgF
NA18507	♂	Y009	F	NA18974	♂	Japanese	NA07022	♂	1340	MgF
NA18506	♂	Y009	C	NA18990	♂	Japanese	NA07029	♂	1340	F
NA18505	♀	Y005	M	NA18994	♂	Japanese	NA12145	♀	1334	FgM
NA18504	♂	Y005	F	NA18995	♂	Japanese	NA12239	♀	1334	MgM
NA18503	♂	Y005	C	NA19000	♂	Japanese	NA10847	♀	1334	M

NA18502	♀	Y004	M	NA19005	♂	Japanese		NA12144	♂	1334	FgF
NA18501	♂	Y004	F	NA19007	♂	Japanese		NA12146	♂	1334	MgF
NA18500	♂	Y004	C	NA19012	♂	Japanese		NA10846	♂	1334	F

All DNA samples in the HapMap project on which all four SNPSTRs were typed. Key: ♀-female, ♂-male, F-father, M-mother, C-child, MgM-maternal grandmother, MgF-maternal grandfather, FgM-paternal grandmother, FgF-paternal grandfather

Bibliography

Agrafioti I, Stumpf MP (2007) SNPSTR: a database of compound microsatellite-SNP markers. Nucleic Acids Res 35:D71-75

Akey J, Jin L, Xiong M (2001) Haplotypes vs single marker linkage disequilibrium tests: what do we gain? Eur J Hum Genet 9:291-300

Allison AC (1954) Protection afforded by sickle-cell trait against subtertian malareal infection. Br Med J 1:290-294

Altshuler D, Brooks LD, Chakravarti A, Collins FS, Daly MJ, Donnelly P (2005) A haplotype map of the human genome. Nature 437:1299-1320

Alves-Silva J, da Silva Santos M, Guimaraes PE, Ferreira AC, Bandelt HJ, Pena SD, Prado VF (2000) The ancestry of Brazilian mtDNA lineages. Am J Hum Genet 67:444-461

Ardlie KG, Kruglyak L, Seielstad M (2002) Patterns of linkage disequilbrium in the human genome (vol 3, pg 299, 2002). Nature Reviews Genetics 3:566-566

Balding D, Donnelly P (1995) Inferring identity from DNA profile evidence. Proc Natl Acad Sci USA 92:11741-11745

Balloux F, Handley LJ, Jombart T, Liu H, Manica A (2009) Climate shaped the worldwide distribution of human mitochondrial DNA sequence variation. Proc Biol Sci 276:3447-3455

Bamshad M, Kivisild T, Watkins WS, Dixon ME, Ricker CE, Rao BB, Naidu JM, Prasad BV, Reddy PG, Rasanayagam A, Papiha SS, Villems R, Redd AJ, Hammer MF, Nguyen SV, Carroll ML, Batzer MA, Jorde LB (2001) Genetic evidence on the origins of Indian caste populations. Genome Res 11:994-1004

Bamshad MJ, Wooding S, Watkins WS, Ostler CT, Batzer MA, Jorde LB (2003) Human population genetic structure and inference of group membership. American Journal of Human Genetics 72:578-589

Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21:263-265

Benn-Torres J, Bonilla C, Robbins CM, Waterman L, Moses TY, Hernandez W, Santos ER, Bennett F, Aiken W, Tullock T, Coard K, Hennis A, Wu S, Nemesure B, Leske MC, Freeman V, Carpten J, Kittles RA (2008) Admixture and population stratification in African Caribbean populations. Ann Hum Genet 72:90-98

Benn Torres J, Kittles RA, Stone AC (2007) Mitochondrial and Y chromosome diversity in the English-speaking Caribbean. Ann Hum Genet 71:782-790

Bertorelle G, Excoffier L (1998) Inferring admixture proportions from molecular data. Mol Biol Evol 15:1298-1311

Biswas S, Akey JM (2006) Genomic insights into positive selection. Trends Genet 22:437-446

Bosch E, Calafell F, Rosser ZH, Norby S, Lynnerup N, Hurles ME, Jobling MA (2003) High level of male-biased Scandinavian admixture in Greenlandic Inuit shown by Y-chromosomal analysis. Hum Genet 112:353-363

Brenner CH, Weir BS (2003) Issues and strategies in the DNA identification of World Trade Center victims. Theor Popul Biol 63:173-178

Brinkmann B, Klintschar M, Neuhuber F, Huhne J, Rolf B (1998) Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. Am J Hum Genet 62:1408-1415

Britain BUoG (2008) Baptist History

Burchard EG, Ziv E, Coyle N, Gomez SL, Tang H, Karter AJ, Mountain JL, Perez-Stable EJ, Sheppard D, Risch N (2003) The importance of race and ethnic background in biomedical research and clinical practice. N Engl J Med 348:1170-1175

Burton M, Moore C, Whiting J, Romney A (1996) Regions based on social structure. Curr Anthropol 37:87-123

Butler JM (2003) Forensic DNA Typing. Elsevier Academic Press, London

Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, Bodmer J, et al. (2002) A human genome diversity cell line panel. Science 296:261-262

Chakraborty R (1986) Gene admixture in human populations: Models and predictions. American Journal of Physical Anthropology 29:1-43

Chien A, Edgar DB, Trela JM (1976) Deoxyribonucleic acid polymerase from the extreme thermophile Thermus aquaticus. J Bacteriol 127:1550-1557

Chikhi L, Bruford MW, Beaumont MA (2001) Estimation of admixture proportions: a likelihood-based approach using Markov chain Monte Carlo. Genetics 158:1347-1362

Consortium TWTCC (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661-678

Crow J, Kimura M (1970) An Introduction to Population Genetics Theory. Harper and Row, New York

Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES (2001) High-resolution haplotype structure in the human genome. Nat Genet 29:229-232

Dansgaard W, Johnsen SJ, Reeh N, Gundestrup N, Clausen HB, Hammer CU (1975) Climatic changes, Norsemen and modern man. Nature 255:24-28

Di Rienzo A, Peterson AC, Garza JC, Valdes AM, Slatkin M, Freimer NB (1994) Mutational processes of simple-sequence repeat loci in human populations. Proc Natl Acad Sci U S A 91:3166-3170

Disteche CM, Casanova M, Saal H, Friedman C, Sybert V, Graham J, Thuline H, Page DC, Fellous M (1986) Small deletions of the short arm of the Y chromosome in 46,XY females. Proc Natl Acad Sci U S A 83:7841-7844

Dupanloup I, Bertorelle G (2001) Inferring admixture proportions from molecular data: extension to any number of parental populations. Mol Biol Evol 18:672-675

Excoffier L, Novembre J, Schneider S (2000) SIMCOAL: a general coalescent program for the simulation of molecular data in interconnected populations with arbitrary demography. J Hered 91:506-509

Fenner JN (2005) Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. Am J Phys Anthropol 128:415-423

Fitzhugh W (1984) Paleo-Eskimo cultures of Greenland. Vol 5 - The Arctic. Smithsonian Institution, Washington DC

Forster P, Harding R, Torroni A, Bandelt HJ (1996) Origin and evolution of Native American mtDNA variation: a reappraisal. Am J Hum Genet 59:935-945

Fox CS, Yang Q, Guo CY, Cupples LA, Wilson PWF, Levy D, Meigs JB (2005) Genome-wide linkage analysis to urinary microalbuminuria in a community-based sample: The Framingham Heart Study. Kidney International 67:70-74

Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851-861

Friesen TM (2004) Contemporaneity of Dorset and Thule Cultures in the North American Arctic: New Radiocarbon Dates from Victoria Island, Nunavut. Current Anthropology 45:685-691

Gad F (1984) History of Colonial Greenland. Vol 5 - The Arctic. Smithsonian Institution, Washington DC

Gill P (2001) An assessment of the utility of single nucleotide polymorphisms (SNPs) for forensic purposes. International Journal of Legal Medicine 114:204-210

Glass B, Li CC (1953) The dynamics of racial intermixture; an analysis based on the American Negro. Am J Hum Genet 5:1-20

Goldstein DB, Linares AR, Cavalli-Sforza LL, Feldman MW (1995) An Evaluation of Genetic Distances for Use With Microsatellite Loci. Genetics 139:463-471

Gudmand-Hoyer E, Jarnum S (1969) Lactose malabsorption in Greenland Eskimos. Acta Med Scand 186:235-237

HapMap TIC (2003) The International HapMap Project. Nature 426:789-796

HapMap TIC (2005) A haplotype map of the human genome. Nature 437:1299-1320

Haywood J (1995) The Penguin Historical Atlas of the Vikings. Penguin Books, London

He M, Gitschier J, Zerjal T, de Knijff P, Tyler-Smith C, Xue Y (2009a) Geographical Affinities of the HapMap Samples. PLoS ONE 4:e4684

He M, Gitschier J, Zerjal T, de Knijff P, Tyler-Smith C, Xue Y (2009b) Geographical affinities of the HapMap samples. PLoS One 4:e4684. Epub 2009 Mar 4684.

Helgason A, Palsson G, Pedersen HS, Angulalik E, Gunnarsdottir ED, Yngvadottir B, Stefansson K (2006) mtDNA variation in Inuit populations of Greenland and Canada: migration history and population structure. Am J Phys Anthropol 130:123-134

Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K (2002) A comprehensive review of genetic association studies. Genet Med 4:45-61

Hirszfeld L, Hirszfeld H (1919) Essai d'application des methods au probléme des races. Anthropologie 29:505-537

Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, Pearson JV, Stephan DA, Nelson SF, Craig DW (2008) Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genet 4:e1000167

Hubisz M, Falush D, Stephens M, Pritchard JK (2009) Inferring weak population structure with the assistance of sample group information Molecular Ecology Resources

Hurles ME, Irven C, Nicholson J, Taylor PG, Santos FR, Loughlin J, Jobling MA, Sykes BC (1998) European Y-chromosomal lineages in Polynesians: a contrast to the population structure revealed by mtDNA. Am J Hum Genet 63:1793-1806

Jeffreys AJ, Kauppi L, Neumann R (2001) Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nat Genet 29:217-222

Jin HJ, Tyler-Smith C, Kim W (2009) The peopling of Korea revealed by analyses of mitochondrial DNA and Y-chromosomal markers. PLoS ONE 4

Jobling MA, Gill P (2004) Encoded evidence: DNA in forensic analysis (vol 5, pg 739, 2004). Nature Reviews Genetics 6:246-246

Jobling MA, Hurles ME, Tyler-Smith C (2003) Human Evolutionary Genetics: origins, peoples and disease. Garland Science, New York

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

Jones R (1989) The Human Revolution: Behavioural and Biological Perspectives on the Origin of Modern Humans. Princetown University Press, Princetown, NJ

Karafet T, Xu L, Du R, Wang W, Feng S, Wells RS, Redd AJ, Zegura SL, Hammer MF (2001) Paternal population history of East Asia: sources, patterns, and microevolutionary processes. Am J Hum Genet 69:615-628

Kimpton C, Fisher D, Watson S, Adams M, Urquhart A, Lygo J, Gill P (1994) Evaluation of an automated DNA profiling system employing multiplex

amplification of four tetrameric STR loci. International Journal of Legal Medicine 106:302-311

King TE, Ballereau SJ, Schurer KE, Jobling MA (2006) Genetic signatures of coancestry within surnames. Curr Biol 16:384-388

Kingman JF (2000) Origins of the coalescent. 1974-1982. Genetics 156:1461-1463

Kleivan I (1984) History of Norse in Greenland. In: Damas D (ed) Handbook or North American Indians. Vol 5 - The Arctic. Smithsonian Institution, Washington DC, pp 549 - 555

Kleppe K, Ohtsuka E, Kleppe R, Molineux I, Khorana HG (1971) Studies on polynucleotides. XCVI. Repair replications of short synthetic DNA's as catalyzed by DNA polymerases. J Mol Biol 56:341-361

Landsteiner K (1900) Zur Kenntnis des antifermentativen, lytischen und agglutinierenden Wirkungen des Blutserums und der Lymphe. Zbl Bakt 27:357−363

Laspada AR, Wilson EM, Lubahn DB, Harding AE, Fischbeck KH (1991) Androgen Receptor Gene-Mutations in X-Linked Spinal and Bulbar Muscular-Atrophy. Nature 352:77-79

Lewontin R (1972) Evolutionary Biology. Appleton-Century-Crofts, New York

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

Lowe AL, Urquhart A, Foreman LA, Evett IW (2001) Inferring ethnic origin by means of an STR profile. Forensic Science International 119:17-22

Luria SE, Delbruck M (1943) Mutations of bacteria from viris sensitivity to virus resistance. Genetics 28:491-511

Lynch AG, Dunning MJ, Iddawela M, Barbosa-Morais NL, Ritchie ME (2009) Considerations for the processing and analysis of GoldenGate-based two-colour Illumina platforms. Stat Methods Med Res

Lynnerup N (1998) The Greenland Norse: a biological-anthropological study. Commission for Scientific Research in Greenland, Copenhagen

McElreavey K, Ravel C, Chantot-Bastaraud S, Siffroi JP (2006) Y chromosome variants and male reproductive function. Int J Androl 29:298-303; discussion 304-296

McGhee R (2000) Radiocarbon dating and the timing of the Thule migration. Danish Polar Center, Copenhagen

Miljkovic-Gacic I, Ferrell RE, Patrick AL, Kammerer CM, Bunker CH (2005) Estimates of African, European and Native American ancestry in Afro-Caribbean men on the island of Tobago. Hum Hered 60:129-133

Misawa K, Kikuno RF (2009) Evaluation of the effect of CpG hypermutability on human codon substitution. Gene 431:18-22

Morgan NV, Gissen P, Malik-Sharif S, Bennett CP, Woods CG, Trembath RC, Maher ER, Johnson CA (2002) A novel locus for Meckel-Gruber

syndrome, MKS3, maps to chromosome 8q24. Journal of Medical Genetics 39:S69-S69

Mountain JL, Knight A, Jobin M, Gignoux C, Miller A, Lin AA, Underhill PA (2002) SNPSTRs: Empirically derived, rapidly typed, autosomal Haplotypes for inference of population history and mutational processes. Genome Research 12:1766-1772

Mourant AE, Kopec AC, Domaniewska-Sobczak K (1976) The distribution of the human blood groups, and other polymorphisms. Oxford University Press, London

Mullis KB (1990) The unusual origin of the polymerase chain reaction. Sci Am 262:56-61, 64-55

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

Nei M (1987) Molecular Evolutionary Genetics. Columbia University Press, New York

Ohta TK, M (1973) A model of mutation appropriate to estimate the number of electropherically detectable molecules in a finite population. Genet Res 22:201-204

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

Paabo S (2003) The mosaic that is our genome. Nature 421:409-412

Pamilo P, Nei M (1988) Relationships between gene trees and species trees. Mol Biol Evol 5:568-583

Parra EJ, Marcini A, Akey J, Martinson J, Batzer MA, Cooper R, Forrester T, Allison DB, Deka R, Ferrell RE, Shriver MD (1998) Estimating African American admixture proportions by use of population-specific alleles. Am J Hum Genet 63:1839-1851

Peakall R, Smouse P (2006) genalex 6: genetic analysis in Excel. Population genetic software for teaching and research. Molecular Ecology Notes 6:288-295

Persson I (1969) The Fate of the Icelandic Vikings in Greenland. Man 4:620-628

Pfaff CL, Parra EJ, Bonilla C, Hiester K, McKeigue PM, Kamboh MI, Hutchinson RG, Ferrell RE, Boerwinkle E, Shriver MD (2001) Population structure in admixed populations: effect of admixture dynamics on the pattern of linkage disequilibrium. Am J Hum Genet 68:198-207

Pringle H (1997) Archaeology: Death in Norse Greenland. Science 275:924-926

Pritchard J, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945-959

Prugnolle F, Manica A, Balloux F (2005) Geography predicts neutral genetic diversity of human populations. Curr Biol 15:R159-160

Richards M, Macaulay V, Hickey E, Vega E, Sykes B, Guida V, Rengo C, et al. (2000) Tracing European founder lineages in the Near Eastern mtDNA pool. Am J Hum Genet 67:1251-1276

Risch N, Burchard E, Ziv E, Tang H (2002) Categorization of humans in biomedical research: genes, race and disease. Genome Biol 3:comment2007

Roewer L, Krawczak M, Willuweit S, Nagy M, Alves C, Amorim A, Anslinger K, et al. (2001) Online reference database of European Y-chromosomal short tandem repeat (STR) haplotypes. Forensic Sci Int 118:106-113

Rogoziânski J (1999) A brief history of the Caribbean: from the Arawak and the Carib to the present, New York

Rosenberg N, Pritchard J, Weber JL, Cann H, Kidd K, Zhivotovsky L, Feldman M (2002) Genetic Structure of Human Populations. Science 298:2381-2385

Rosenberg NA (2006) Standardized Subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, Accounting for Atypical and Duplicated Samples and Pairs of Close Relatives. Annals of Human Genetics 70:1-7

Rosser ZH, Zerjal T, Hurles ME, Adojaan M, Alavantic D, Amorim A, Amos W, et al. (2000) Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. Am J Hum Genet 67:1526-1543

Rouse I (1992) The Tainos: rise and decline of the people who greeted Columbus. Yale University Press, New Haven

Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19:2496-2497

Saiki RK (1985) Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. Science v230:p1350(1355)

Saillard J, Forster P, Lynnerup N, Bandelt HJ, Norby S (2000) mtDNA variation among Greenland Eskimos: the edge of the Beringian expansion. Am J Hum Genet 67:718-726

Sanger F, Nicklen S, Coulson A (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74:5463-5467

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

Schneider S, Roessli D, Excoffier L (2000a) Arlequin: A software for population genetics data analysis. Ver 2.000.

Schneider S, Roessli D, Excoffier L (2000b) Arlequin: A software for population genetics data analysis. Ver 2.000. Genetics and Biometry Lab, Dept. of Anthropology, University of Geneva.

Scott G, Alexandersen V (1991) Dental morphological variation among medieval Greenlanders, Icelanders and Norwegians. Freund Scientific, Jerusalem

Seielstad MT, Minch E, Cavalli-Sforza LL (1998) Genetic evidence for a higher female migration rate in humans. Nature Genetics 20:278-280

Serre D, Paabo S (2004) Evidence for gradients of human genetic diversity within and among continents. Genome Res 14:1679-1685

Shriver MD, Parra EJ, Dios S, Bonilla C, Norton H, Jovel C, Pfaff C, Jones C, Massac A, Cameron N, Baron A, Jackson T, Argyropoulos G, Jin L, Hoggart CJ, McKeigue PM, Kittles RA (2003) Skin pigmentation, biogeographical ancestry and admixture mapping. Hum Genet 112:387-399

Slatkin M (1995) A measure of population subdivision based on microsatellite allele frequencies. Genetics 139:457-462

Southern E (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. J Mol Biol 98:503-517

Sprecher C, Krenke, B., Amiott, B. & Rabbach, D.,K (2000) The PowerPlex 16 System.

SPSS I (2001) SPSS for Windows release Rel. 11.0.1, Chicago

Starikovskaya YB, Sukernik RI, Schurr TG, Kogelnik AM, Wallace DC (1998) mtDNA diversity in Chukchi and Siberian Eskimos: implications for the genetic history of Ancient Beringia and the peopling of the New World. Am J Hum Genet 63:1473-1491

Stephens M, Donnelly P (2003) A comparison of bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet 73:1162-1169

Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68:978-989

Strachan T, Read A (2004) Genetic Markers in Human Molecular Genetics 3. Garland Science, London

Su B, Xiao J, Underhill P, Deka R, Zhang W, Akey J, Huang W, Shen D, Lu D, Luo J, Chu J, Tan J, Shen P, Davis R, Cavalli-Sforza L, Chakraborty R, Xiong M, Du R, Oefner P, Chen Z, Jin L (1999) Y-Chromosome evidence for a northward migration of modern humans into Eastern Asia during the last Ice Age. Am J Hum Genet 65:1718-1724

Sued-Badillo J (2003) Autochthonous Societies. UNESCO Publishing/Macmillan Publishers Ltd, Paris

Sullivan K, Mannucci A, Kimpton C, Gill P (1993) A rapid and quantitative DNA sex test - fluorescence-based PCR analysis of X-Y homologous gene amelogenin. Biotechniques 15

Sweeney C, Wolff RK, Byers T, Baumgartner KB, Giuliano AR, Herrick JS, Murtaugh MA, Samowitz WS, Slattery ML (2007) Genetic admixture

among Hispanics and candidate gene polymorphisms: potential for confounding in a breast cancer study? Cancer Epidemiol Biomarkers Prev 16:142-150

Szibor R, Hering S, Edelmann J (2005) The HumARA genotype is linked to spinal and bulbar muscular dystrophy and some further disease risks and should no longer be used as a DNA marker for forensic purposes. International Journal of Legal Medicine 119:179-180

Tang H, Siegmund DO, Shen P, Oefner PJ, Feldman MW (2002) Frequentist estimation of coalescence times from nucleotide sequence data using a tree-based partition. Genetics 161:447-459

Tishkoff SA, Dietzsch E, Speed W, Pakstis AJ, Kidd JR, Cheung K, BonneTamir B, SantachiaraBenerecetti AS, Moral P, Krings M, Paabo S, Watson E, Risch N, Jenkins T, Kidd KK (1996) Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. Science 271:1380-1387

Torgerson W (1952) Multidimensional scaling: I. theory and method. Psychometrika 17:401-419

Tremblay M, Vezina H (2000) New estimates of intergenerational time intervals for the calculation of age and origins of mutations. Am J Hum Genet 66:651-658

Tut TG, Ghadessy FJ, Trifiro MA, Pinsky L, Yong EL (1997) Long polyglutamine tracts in the androgen receptor are associated with reduced trans-activation, impaired sperm production, and male infertility. Journal of Clinical Endocrinology and Metabolism 82:3777-3782

Underhill PA, Jin L, Zemans R, Oefner PJ, Cavalli-Sforza LL (1996) A pre-Columbian Y chromosome-specific transition and its implications for human evolutionary history. Proc Natl Acad Sci U S A 93:196-200

Vogt PH, Edelmann A, Kirsch S, Henegariu O, Hirschmann P, Kiesewetter F, Kohn FM, Schill WB, Farah S, Ramos C, Hartmann M, Hartschuh W, Meschede D, Behre HM, Castel A, Nieschlag E, Weidner W, Grone HJ, Jung A, Engel W, Haidl G (1996) Human Y chromosome azoospermia factors (AZF) mapped to different subregions in Yq11. Hum Mol Genet 5:933-943

Wallace DC (2005) A MITOCHONDRIAL PARADIGM OF METABOLIC AND DEGENERATIVE DISEASES, AGING, AND CANCER: A Dawn for Evolutionary Medicine. Annual Review of Genetics 39:359-407

Weeks J, Ferbel P (1994) Ancient Caribbean. Garland Publishing, New York

Weir BS (1996) Genetic Data Analysis II. Sinauer Associates, Sunderland, MA

Wilder JA, Kingan SB, Mobasher Z, Pilkington MM, Hammer MF (2004) Global patterns of human mitochondrial DNA and Y-chromosome structure are not influenced by higher migration rates of females versus males. Nat Genet 36:1122-1125

Wilson JF, Weale ME, Smith AC, Gratrix F, Fletcher B, Thomas MG, Bradman N, Goldstein DB (2001) Population genetic structure of variable drug response. Nat Genet 29:265-269

Witherspoon D, Wooding S, Rogers A, Marchani E, Watkins WS, Batzer MA, Jorde LB (2007) Genetic similarities within and between human populations. Genetics 176:351-359

Wright S (1951) The genetical structure of populations. Ann Eugen 15:323-354

Zitzmann M, Depenbusch M, Gromoll J, Nieschlag E (2003) Prostate volume and growth in testosterone-substituted hypogonadal men are dependent on the CAG repeat polymorphism of the androgen receptor gene: A longitudinal pharmacogenetic study. Journal of Clinical Endocrinology and Metabolism 88:2049-2054