Inference of Biogeographical Ancestry and Pigmentation Phenotype using Single Nucleotide Polymorphisms

Charmain Vanessa Castel

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

Centre for Forensic Science

University of Technology, Sydney, NSW, Australia

2014

Certificate of Original Authorship

I certify that work in this thesis has not previously been submitted for a degree nor has it been submitted as part of the requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged, In addition, I certify that all information sources and literature used are indicated in the thesis.

Charmain Vanessa Castel
July 2014

Acknowledgements

First and foremost, I would like to sincerely thank my supervisor Associate Professor Anita Piper for her ongoing support and invaluable guidance throughout my candidature. I have thoroughly enjoyed working with you over the years and have learnt so much, not only about the 'science' but also myself and my capabilities. Thank you for helping me to grow into an independent researcher and for your patience and continuous encouragement.

Thank you to Prof. Claude Roux (Centre for Forensic Science) and the Faculty of Science at the University of Technology, Sydney (UTS) for their funding and support of this project. I would also like to thank the UTS Graduate School who have awarded me with an Australian Postgraduate Award and have provided funding support to attend national and international conferences.

Thank you to Dr Runa Daniel (Victoria Police Forensic Services Centre) for the advice she provided in the early stages of this project, particularly for showing me the ins and outs of the ABI 310. Special thanks also extend to Connie Akratos and Zofia Winiarski from Technical Services (UTS) for kindly providing me with some of the materials and equipment I needed to complete the assay validation phase of this project. I would also like to thank all of the volunteers who provided DNA samples and without whom this project would not have been possible.

Thank you to all my friends and colleagues at UTS, past and present. In particular, Dr Michael Johnson, Dr Marilyn Katrib, Dr Rob Walker, Dr Michael Lees, Dr Nicholas Archer and Prue Gatt for convincing me to do a PhD. Whilst this has at times been a tough journey, I am so glad I took the bait and could not have done it without your encouragement, guidance and wisdom. Thank you all so much for the fun times and memories that I will cherish for a lifetime. Most importantly, I would like to thank Dr Stephanie Dowdell for being my office buddy, for all the teas and invaluable words of wisdom, for your kindness and generosity and for all the guidance, moral support and encouragement that you have given me over the years. I am so lucky to have made a lifetime friend like you!

Thank you to my dearest and oldest friend Mohini Saini for 15 years of friendship, laughter and fun. Despite the distance that now separates us, you have not ceased to provide me with your ongoing support and encouragement all the way from the other side of the World and I truly appreciate all that you have done.

I would especially like to thank (fellow soon to be Dr) Michael Strauss for sharing this incredible journey with me and for being by my side from day one. Thank you for being an amazing and understanding person, for your patience and for always supporting, and encouraging me through it all, particularly the difficult times. I would not have made it without you and I look forward to starting the next chapter of our lives together.

Finally, I would like to thank my family, especially my parents Yves Castel, Lou and Tony Doyle and my sister Karen for their wisdom and ongoing support and throughout my life. Each of you have inspired me in your own special way, but you have all taught me to never give up and that no challenge is too great. Your unwavering belief in me and my abilities astounds me and I would not have accomplished this, one of the greatest intellectual challenges of my life, without you all. I am truly grateful to have been blessed with a family like you.

Table of Contents

Table of Figures

Table of Tables

Publications

Castel C.V. and Piper A.A (2011) Development of a SNP multiplex assay for the inference of biogeographical ancestry and pigmentation phenotype. *Forensic Sci Int Genet Suppl Series,* **3***,* e411-412

Presentations

Castel, C.V. and Piper, A.A. Development of a SNP multiplex assay for the inference of biogeographical ancestry and pigmentation phenotype. Poster presented at the 24th World Congress of the International Society of Forensic Genetics, Vienna, Austria, August 2011. Abstract P014.

Castel, C.V. and Piper, A.A. Mitochondrial single nucleotide polymorphisms: Informative markers of biogeographical ancestry. Poster presented at the 20th International Symposium on the Forensic Sciences: Forensic Science on trial, Sydney, Australia, September 2010. Abstract 238.

List of Abbreviations

AIM	Ancestry Informative Marker
α-MSH	α-Melanocyte Stimulating Factor
ASCCEG	Australian Standard Classification of Cultural and Ethnic Groups
ASIP	Agouti Signalling Protein
ATPase	ATP Synthase
AUC	Area Under the receiver characteristic operating Curve
BGA	Biogeographical Ancestry
bp	Base Pair
cAMP	Cyclic Adenosine Monophosphate
CEPH	Centre d'Etude du Polymorphisme Humain
CI	Confidence Interval
CIP	Calf Intestinal Phosphatase
CO	Cytochrome C oxidase
Cytb	Cytochrome b
DARC	Duffy Antigen/Receptor for Chemokine
dbSNP	SNP database
DCT	Dopachrome Tautomerase
δ	Absolute Allele Frequency Difference
ddNTP	Dideoxyribonucleotide Triphosphate
dH_20	Deionised water
DHPLC	Denaturing High Performance Liquid Chromatography
DNA	Deoxyribonucleic Acid
dNTP	Deoxyribonucleotide Triphosphate
DTT	Dithiothreitol
EDTA	Ethylinediaminetetraacetic Acid
EVC	External Visible Characteristic
ExoI	Exonuclease I
FIC	Fisher Information Content
[F]ddNTP	Fluorescently labelled Dideoxyribonucleotide Triphosphate
F_{ST}	F-Statistic

FUT	Fucosyltransferase
GDA	Genetic Distance Analysis
gDNA	Genomic Deoxyribonucleic Acid
H_e	Expected Heterozygosity
HGDP	Human Genome Diversity Project
HLTF	Helicase-Like Transcription Factor
H_o	Observed Heterozygosity
HWE	Hardy Weinberg Equilibrium
I_n	Rosenberg's Informativeness for Assignment measure
IL	Interleukin
LD	Linkage Disequilibrium
LE	Linkage Equilibrium
LEF1	Lymphoid Enhancer Binding Factor 1
MA	Minor Allele
MALDI-TOF	Matrix Assisted Laser Desorption Ionisation-Time Of Flight
MATP	Membrane Associate Transporter Protein
MC1R	Melanocortin-1 Receptor
MCMC	Markov Chain Monte Carlo
$MgCl_2$	Magnesium Chloride
MITF	Microphthalmia Transcription Factor
MLE	Maximum Likelihood Estimate
mM	Millimolar
mtDNA	Mitochondrial Deoxyribonucleic Acid
NCBI	National Center for Biotechnology Information
ND	NADH dehydrogenase
nDNA	Nuclear Deoxyribonucleic Acid
ng	Nanogram
NJ	Neighbour-Joining
NRY	Non-Recombining region of the Y-Chromosome
nt	Nucleotide
OCA2	Oculocutaneous Albinism Type 2
OR	Odds Ratio

PAR	Pseudoautosomal Region
PCoA	Principal Co-ordinates Analysis
PCR	Polymerase Chain Reaction
pg	Picogram
RFLP	Restriction Fragment Length Polymorphism
RFU	Relative Fluorescence Unit
rRNA	Ribosomal Ribonucleic Acid
SAP	Shrimp Alkaline Phosphatase
SBE	Single Base Extension
SIC	Shannon Information Content
SLC24A5	Solute Carrier Family 24 member 5
SNP	Single Nucleotide Polymorphism
STR	Short Tandem Repeat
TBE	Tris Borate EDTA
TE	Tris EDTA
T_m	Theoretical Melting Temperature
TMRCA	Time of the Most Recent Common Ancestor
TYR	Tyrosine
TYRP1	Tyrosine Related Protein-1
µg	Microgram
µL	Microlitre
USA	United States of America
UTS	University of Technology, Sydney

Abstract

Conventional DNA profiling of Short Tandem Repeats (STR) provides little evidentiary value in the absence of reference profiles or in the case of a non-match. Recently, the forensic DNA intelligence field has flourished to provide investigators with valuable information from DNA samples that can narrow the collection of potential matches by identifying previously unknown reference individuals. Intelligence data of special interest includes the biogeographical ancestry (BGA) and external visible characteristics (EVC) such as the eye, hair and skin colour of unknown DNA samples donors.

Innovative technological advances like next-generation sequencing and microarrays have been crucial to the establishment of population diversity repositories comprising millions of DNA markers, the most abundant of which include single nucleotide polymorphisms (SNPs). Large-scale SNP studies of global populations have enabled reconstructions of mitochondrial (mtDNA) and non-recombining Y-chromosome (NRY) phylogenies, providing highly comprehensive population specific patterns of maternal and paternal genetic variation. Similarly, numerous patterns of autosomal genetic variation have been identified between different populations. These studies have culminated in panels of markers capable of resolving ancestry at the continental level. The identification of autosomal SNPs associated with human pigmentation variation has also resulted in the discovery of specific SNPs capable of predicting EVCs. Several DNA intelligence and phenotyping assays for the inference of BGA and for the prediction of eye, hair and skin colour have subsequently been developed. However, most of these intelligence tools have primarily focused on the analysis of one class of SNPs, hence limiting the amount of ancestry intelligence that could be obtained. The scarcity and often environmentally compromised nature of forensic biological evidence means that performing numerous individual intelligence tests is not optimal and a consolidated DNA intelligence diagnostic test is very much needed.

This study aimed to develop a SNP genotyping system that combined autosomal, NRY and mtDNA markers for comprehensive predictions of BGA and EVCs. Candidate SNPs were selected through literature and database searches to identify loci exhibiting skewed allele frequency differences between Sub-Saharan African, North African, Middle Eastern, European, South and East Asian populations. A hierarchical

arrangement comprising five separate multiplexes was implemented, in which SNP typing was performed by single-base extension assays. The haploid mtDNA and NRY SNPs were grouped into Multiplex 1 to 4, with SNPs defining maternal and paternal lineages (haplogroups) affiliated with the same geographic region grouped in the same reaction. The markers defining basal haplogroups were included in Multiplex 1, which is then used to identify the subsequent multiplex(es) required to achieve further haplogroup resolution and to minimise the number of tests required. The autosomal SNPs are typed separately in Multiplex 5.

A performance evaluation of the 5-multiplex SNP assay was undertaken on 146 individuals originating from the six major population groups of interest. Population genetic analyses of the mtDNA and NRY haplotypes and autosomal genotypes revealed that a greater degree of population differentiation was achieved with the selected NRY and autosomal SNPs than with the mtDNA SNPs. Moreover, the results indicated that the assay primarily allowed for the differentiation of continental ancestry, with populations in close proximity within continents, such as Europe, the Middle East and South Asia, often difficult to distinguish. However, the observed correlation between the declared and inferred geographic regions of maternal and paternal origin was high; 73-100% for maternal and 79-100% for paternal regional BGA. The bi-parental BGA predictions ranged from 85 to 95%, provided Middle Easterners and Europeans were grouped into a single Western Eurasian population. In 99% of cases, two of the three SNP classes correctly predicted the same ancestry from one of the five broad geographical regions (Sub-Saharan Africa, North Africa, Western Eurasia, South Asia and East Asia). High prediction accuracies were also observed for the inference of EVCs including hair (86-88%) and eye colour (81-95%). The DNA intelligence assay also demonstrated advanced performance with low starting amounts of genomic DNA, with full profiles observed for up to 100pg of template and for the analysis of routine casework biological samples. Consequently, this study presented the successful development of a novel, consolidated DNA intelligence tool that has displayed high performance for the inference of regional (continental) BGA and EVC in preliminary tests. Further validations of the assay are required; however the developed 5-multiplex SNP assay remains a valuable DNA intelligence diagnostic tool for the forensic science community.

CHAPTER 1

INTRODUCTION

Chapter 1. Introduction

1.1 Preface

Innovative technological advances in molecular biology have revolutionised the investigation of human genetic and biological variation and their potential relationship with individual ancestry. The correlation between genetics and ancestry has long been debated (Foster and Sharp, 2002, Burchard *et al.*, 2003, Royal and Dunston, 2004). However, the Human Genome Project has pioneered the study of the interrelationship between genetic variation, phenotype and biological function (Hood and Rowen, 2013). Human genetic variation is affected by the combined action of molecular factors such as mutation, recombination and random assortment, and population factors including inbreeding, selection, genetic drift and migration (Hedrick, 2011a). Single nucleotide polymorphisms (SNPs) are abundant markers of human genetic variation that occur in nuclear DNA approximately every 1900bp (Sachidanandam *et al.*, 2001) and comprise single base pair positions in DNA for which typically two sequence alternatives (alleles) exist between different individuals [For review see Brookes (1999)]. The collaborative efforts of the SNP Consortium (Sachidanandam *et al.*, 2001) and the Human Genome Project (International Human Genome Sequencing Consortium, 2001), as well as ongoing research, have currently identified over 60 million human SNPs, 44 million of which have been validated and are available in the National Center for Biotechnology Information (NCBI) public SNP database (dbSNP) [NCBI (2013), dbSNP build 138]. Similar large-scale genomic resources such as the HapMap (The International HapMap Consortium, 2010), Centre d'Etude du Polymorphisme Humain (CEPH) (Cann *et al.*, 2002) and 1000 Genomes (Abecasis *et al.*, 2010) databases are not only catalogues of human variability but also provide population information for the samples studied, thus facilitating the investigation of interpopulation differences in human genetic variation.

Nuclear DNA (nDNA) SNPs exhibit relatively low mutation rates of 2.5×10^{-8} nucleotides per generation compared to other DNA polymorphisms such as Short Tandem Repeats (STR) for which estimates range from $0.85\text{-}1.93\times10^{-3}$ depending on the length of the repeat motif (Zhivotovsky, 2001). This makes SNPs relatively stable and confers a reduced rate of recurrent mutation, such that SNPs have been used as novel

markers in population genetic analyses for studies of human evolution [For review see Cavalli-Sforza and Marcus (2003) and Morin *et al.* (2004)]. Population studies have revealed widely accepted patterns of genetic variation across major population groups and numerous genetic polymorphisms capable of identifying and distinguishing different global populations have been uncovered (Torroni *et al.*, 1996, Hammer *et al.*, 2001, Underhill *et al.*, 2001, Rosenberg *et al.*, 2002, Tanaka *et al.*, 2004, Jakobsson *et al.*, 2008, Lao *et al.*, 2008, Li *et al.*, 2008). Such polymorphisms, termed ancestry informative markers (AIMs), have demonstrated the ability to infer ancestry affiliations in diverse and admixed populations (Shriver *et al.*, 2003, Collins-Schramm *et al.*, 2004, Kim *et al.*, 2005, Yang *et al.*, 2005, Baye *et al.*, 2009, Kersbergen *et al.*, 2009, Kosoy *et al.*, 2009, Nassir *et al.*, 2009, Kidd *et al.*, 2011, Pereira *et al.*, 2012).

Nuclear SNPs have also been invaluable as genetic markers in association studies, assisting in the identification of causative genes for complex diseases such as diabetes (Willer *et al.*, 2007, Zhong *et al.*, 2010) and cancer (Zhu *et al.*, 2001, Easton *et al.*, 2007, Tomlinson *et al.*, 2007). Coding region SNPs have been associated with numerous inherited monogenic disorders (Kerem et al., 1989, Thyagarajan et al., 2008) whilst regulatory region SNPs have been linked with common diseases (Davignon et al., 1988). This causative association of specific SNPs with inherited disease or disease susceptibility has resulted in their application as diagnostic markers in medicine (Nishizuka *et al.*, 2003, Darabi *et al.*, 2012) [also reviewed by Schaaf *et al.* (2011)]. Furthermore, the association between ancestry and disease susceptibility has now been firmly established. The population frequencies of SNPs associated with susceptibility to inherited disorders such as sickle cell anaemia (Neel, 1953, Penaloza and Salamanca-Gomez, 1996) and cystic fibrosis (Kalman *et al.*, 1994, Bobadilla *et al.*, 2002) have been reported and have demonstrated significant links to particular ethnic groups.

1.2 Tracing the origins and phylogeny of modern humans using haploid markers

Prior to the development of major genotyping techniques, which facilitated the analysis of DNA sequence variation, archaeological fossil evidence provided the only source of information regarding the possible origins of *Homo sapiens*. Archaeological and

anthropological findings have generally shown support for the origin of modern humans in East Africa between 150 000 and 200 000 years ago and their initial spread to the rest of the continent, before expanding throughout the Old World approximately 50-60 000 years ago (Stringer, 2002, White *et al.*, 2003, McDougall *et al.*, 2005). The discovery of new molecular techniques in the early 1980's initiated the study of DNA variation (Brown, 1980, Johnson *et al.*, 1983) prompting interest in using genetic data to estimate the time for the emergence of modern humans from Africa, also commonly referred to as TMRCA (Time of the Most Recent Common Ancestor). Early studies investigating autosomal restriction length polymorphisms (RFLPs) were in accordance with previous archaeological estimates (Bowcock *et al.*, 1991, Mountain and Cavalli-Sforza, 1994) however the utility of haploid markers soon became apparent and the discovery of SNPs located in mitochondrial DNA (mtDNA) and the 'non-recombining region of the Y chromosome' (NRY) have since significantly advanced the reconstruction of human population history [reviewed in (Underhill and Kivisild, 2007)].

1.2.1 MtDNA and Y chromosomal variation:

Unlike autosomes, mtDNA and the Y-chromosome are haploid and demonstrate uniparental inheritance, maternal in the case of mtDNA and paternal in the case of the Y-chromosome (Figure 1.1) The Y-chromosome, depicted in Figure 1.2, expresses relatively few genes and contains two pseudoautosomal regions (PARs) comprising approximately 5% of the sequence which are situated at each end of the chromosome (Rappold, 1993). The PARS have been shown to undergo recombination with similar PARS on the X-chromosome (Henke *et al.*, 1993, Lien *et al.*, 2000) [also reviewed in Flaquer *et al.* (2008)]. The remainder of the Y-chromosome which does not recombine is referred to as the 'non-recombining region of the Y-chromosome' (NRY) and consists of highly repetitive sequences which are transposon rich (Skaletsky *et al.*, 2003). However, Skaletsky *et al.* (2003) also proposed that gene conversion (non-reciprocal recombination) between the palindrome arms was a frequent occurrence in both the human and chimpanzee lineages when sequencing Y-linked orthologues of human NRY palindromes. Despite these findings, no other evidence of recombination has since been put forward and an established route for human NRY recombination is yet to be proposed.

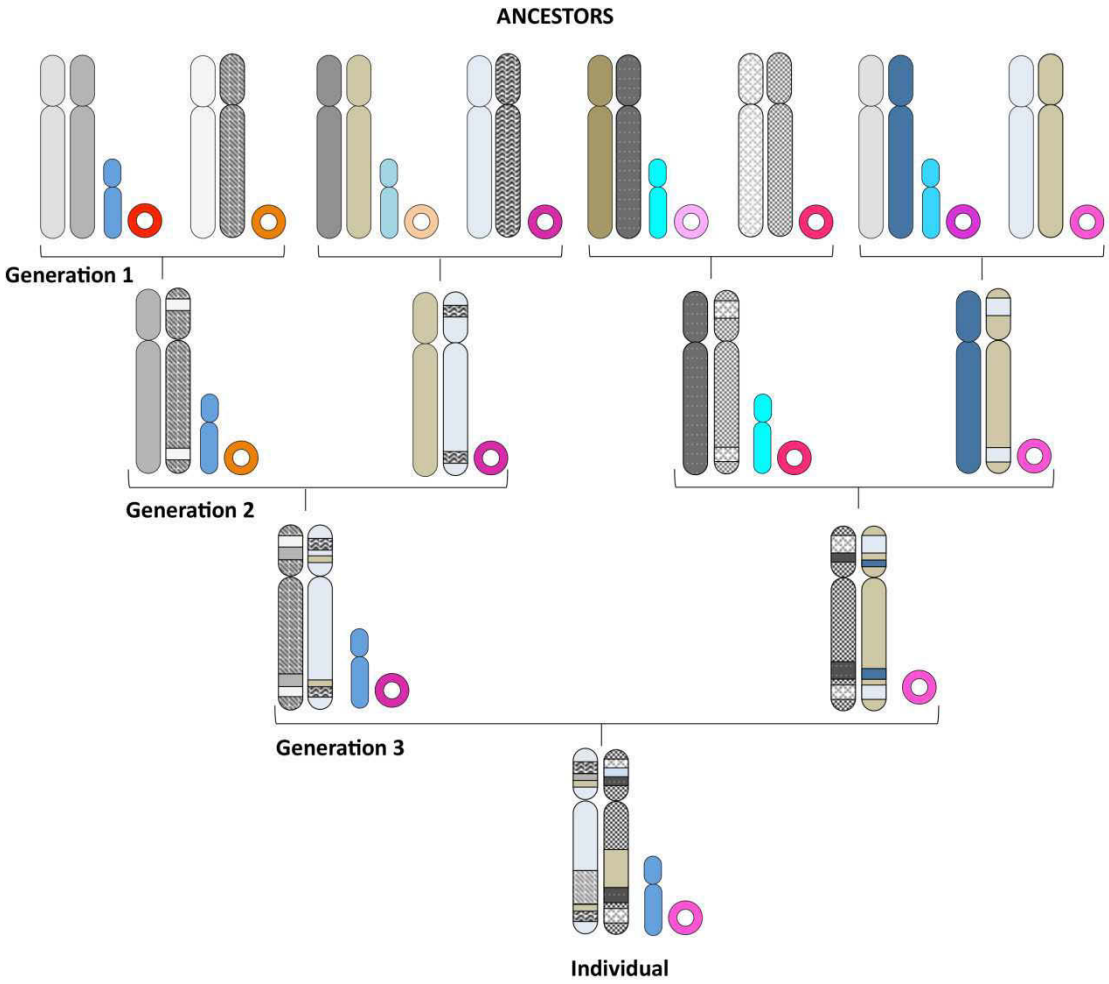


Figure 1.1: Model of autosomal, Y-chromosomal and mitochondrial DNA inheritance.

The large pair of chromosomes depicts autosomal sequences that undergo recombination and random assortment at each generation. In contrast, the non-recombining Y-chromosome sequences (small chromosome) and mitochondrial DNA sequences (small circle) are inherited relatively unchanged along the paternal and maternal lineages, respectively. Thus, an individual's autosomal DNA sequences may be traced back to multiple ancestors whilst the Y-chromosome (excluding the pseudoautosomal regions) and mitochondrial DNA sequences have descended from a single ancestor.

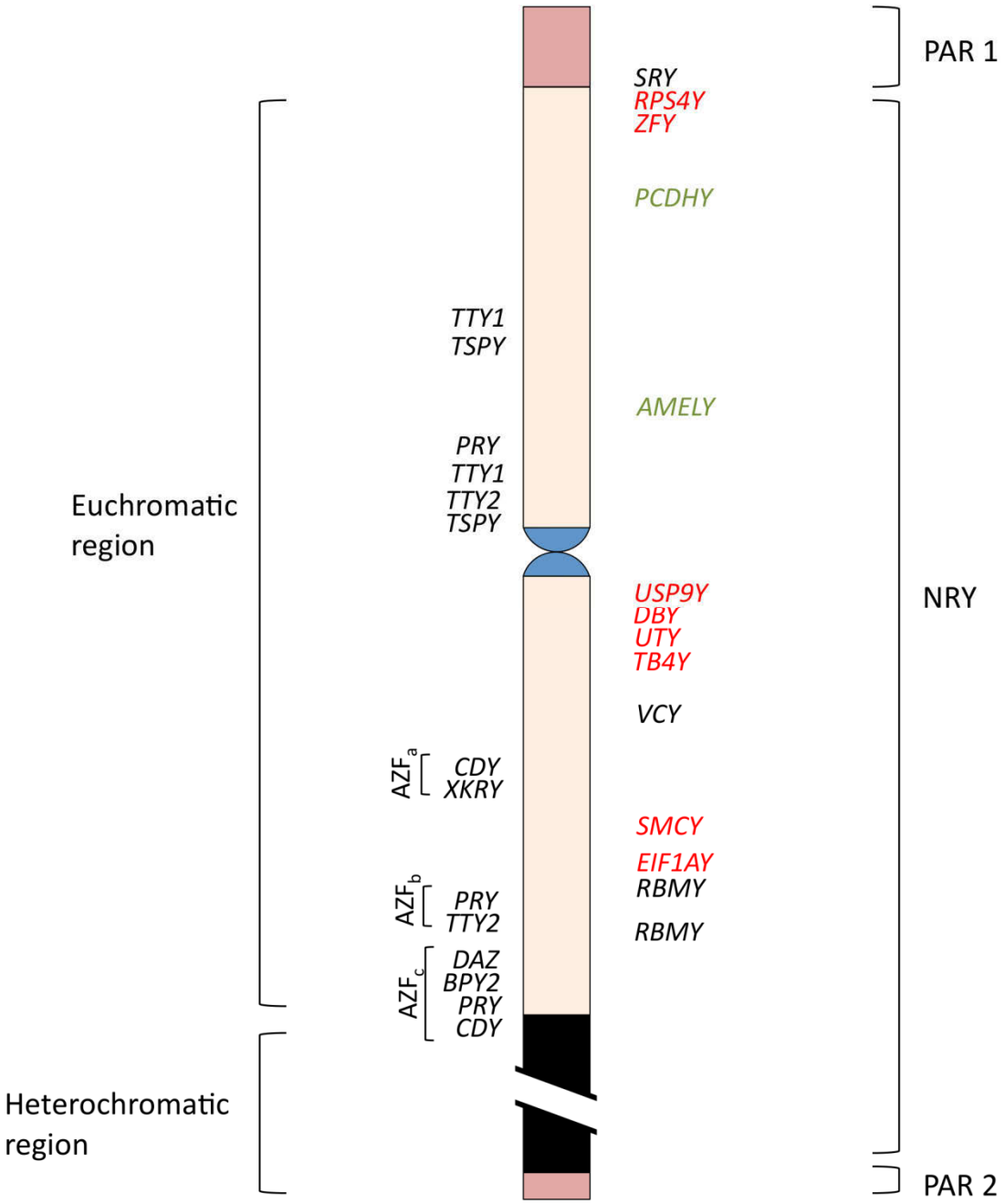


Figure 1.2: Schematic of the human Y chromosome, depicting major regions and active genes.

Shown on the far right, are the pseudoautosomal region 1 (PAR1), pseudoautosomal region 2 (PAR2) and non-recombining region (NRY). The location of genes is specified on the schematic with genes in red identifying housekeeping genes, genes in black denoting those only expressed in the testes whilst the green genes are neither widely expressed or testes specific.

Similar to the NRY, mtDNA, which is a small 16,569 base pair closed circular molecule abundant in cells (Figure 1.3) is considered to not undergo recombination (Moraes *et al.*, 2002). Reports of rare transmission of paternal mtDNA in mammalian cells have led to speculations of the possibility of limited recombination (Thyagarajan *et al.*, 1996). Mutational events characteristic of self-mtDNA recombination in the form of low frequency rearrangements of somatic mtDNA (replication intermediates) have been reported in human heart muscle (Kajander *et al.*, 2001). However paternal mtDNA inheritance has not been observed in humans either in pedigree studies (Jazin *et al.*, 1998, Sigurdardottir *et al.*, 2000) or cases of *in vitro* fertilization (Houshmand *et al.*, 1997, Torroni *et al.*, 1998, Marchington *et al.*, 2002). As such, it is still currently widely accepted that mtDNA and NRY are strictly uniparentally inherited without recombination.

The uniparental inheritance of mtDNA and NRY implies that all individuals have descended from a single ancestral mtDNA and Y type, with genetic variation arising from the sequential accumulation of polymorphisms, predominantly in the form of SNPs scattered throughout the mitochondrial and NRY genome and which remain linked within diverging maternal and paternal lineages [reviewed in (Underhill and Kivisild, 2007)]. This genetic drift or change in allele frequencies within a population over time eventually culminates in the sorting of SNPs among different populations. Inbreeding between individuals within the same population, restricts the shuffling of SNPs among population groups inhabiting different geographical areas. Notably, as humans dispersed, independent lineages known as haplotypes emerged. Each haplotype is defined by a specific motif of accumulated SNP alleles which have become fixed in these diverging populations because they are inherited as a 'block' through the generations, directly from mother to child for mtDNA and directly from father to son for NRY (Underhill and Kivisild, 2007). Consequently, haplotypes are commonly restricted to specific populations, with each population demonstrating a unique combination of haplotypes (Torroni and Wallace, 1994, Hammer *et al.*, 1997).

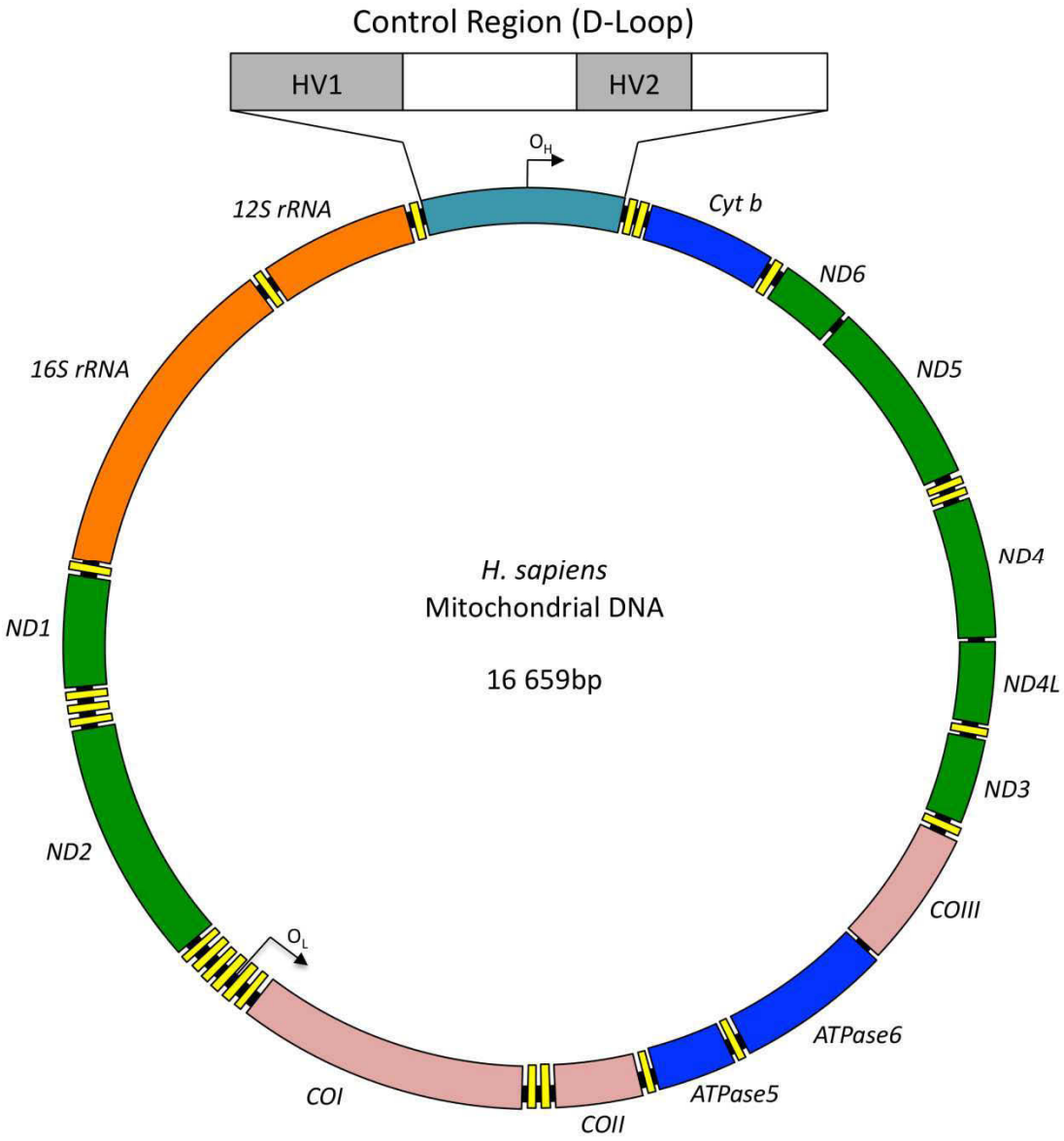


Figure 1.3: **Schematic of the human mitochondrial DNA genome, showing major regions and active genes**

Shown above, is an enlarged segment of the control region (D-loop) (light blue region), depicting hypervariable region 1 (HV1) and hypervariable region 2 (HV2). The location of genes is shown on the genome map with the orange regions specifying ribosomal RNA (rRNA) genes, green regions indicate NADH dehydrogenase (ND) genes, pink regions are those containing cytochrome C oxidase (CO) genes, dark blue regions represent ATP synthase (ATPase) genes, yellow regions are transfer RNA genes and the black region specifies ubiquinol: cytochrome oxidoreductase genes. The origins of heavy (O_H) and light (O_L) strand replication are also shown.

The extent of mtDNA and NRY diversity in different populations is also affected by migration which results in admixture and increases diversity, as admixed populations contain both ancestral alleles and those of new arrivals [reviewed in Wilkins (2006)]. Throughout history, migration events stemming from warfare, colonisation or trade generally involved men who admixed with native populations, resulting in the introgression of NRY instead of mtDNA from these male immigrants into the native population (Wilkins, 2006). Numerous studies of Native American populations have shown elevated frequencies of European NRY haplotypes but not mtDNA haplotypes suggesting predominant male contact in these regions (Carvajal-Carmona *et al.*, 2000, Mesa *et al.*, 2000, Carvajal-Carmona *et al.*, 2003).

The non-recombinant haploid characteristic of mtDNA and the NRY has been useful in evaluating contemporary population diversity and global migration patterns, where the mtDNA reflects the demographic history of females and the NRY reflects the demographic history of males. However, mtDNA haplotypes have been viewed as less informative indicators of geographic origin compared to the NRY haplotypes, which has been attributed to the nearly eight times greater migration rate for females than males (Seielstad *et al.*, 1998). The higher female migration rate has been associated with the widespread practice of patrilocality throughout history, where it is more common for women to leave their birth place and relocate to the homeland of their husbands than the opposite, resulting in more geographically structured Y-chromosomes compared to mtDNA (Seielstad *et al.*, 1998).

Interestingly, a study investigating global patterns of Y-chromosome and mtDNA variation for 389 individuals representing ten geographically sparse global populations, revealed that genetic diversity at the global and continental scale were not shaped by an elevated migration rate of females (Wilder *et al.*, 2004a). Patrilocality has been shown to have more significant effects at the local/regional scale, with higher Y chromosome differentiation observed between patrilocal Northern Thailand Hill tribes, whilst the opposite was observed in matrilocal groups (Besaggio *et al.*, 2007). Similarly, patrilocal Bedouin tribes of the Sinai Peninsula and populations of the Caucasus, Russia, Kurds and Kalmyks also demonstrated parallel genetic patterns consistent with patrilocality (Salem *et al.*, 1996, Malyarchuk *et al.*, 2004, Nasidze *et al.*, 2004, Nasidze *et al.*, 2005a, Nasidze *et al.*, 2005b). It has been suggested that genetic patterns observed within local

regions mainly reveal recent historical events whilst in contrast the genetic patterns of broader geographic areas are more reflective of ancient demographic histories coinciding with pre-agricultural human societies (Wilkins and Marlowe, 2006). As such, the effect of female migration due to patrilocality and relocation to the marital residence associated with the advent of sedentary agricultural practices would be more readily observed between regional populations rather than more geographically isolated populations. This has been proposed by Wilkins and Marlowe (2006) to be the likely reason for the lack of concordance between the Seielstad *et al.* (1998) study whose dense, global sampling scheme showed evidence of higher female migration rate whilst the more geographically sparse sampling of the Wilder *et al*. (2004a) study did not.

Social and cultural practices have also had major effects on mitochondrial and Y chromosomal diversity. Using a population genetic statistic measuring heterozygosity, it was demonstrated that some Central African populations have low Y-chromosome diversity caused by a reduction in the number of Y chromosomes in the population contributing genes to the next generation (Hammer *et al.*, 1997). It was suggested that this reduction is caused by the high rate of polygamy observed in Central African populations (Lucotte *et al.*, 1994), with similar polygamy-associated reduction in Y chromosome diversity also identified in the patrilocal Bedouin tribes of the northern Sinai peninsula (Salem *et al.*, 1996). Other gender specific activities such as war or hunting, as well as demographic differences between the sexes, including the higher mortality rate for males were proposed by Hammer and Zegura (1996) to also cause reductions in the Y-chromosome population. Nonetheless, differences between maternal and paternal lineages may provide the answer to understanding the origin of modern humans and the migration events that have left the specific genetic patterns, which characterise our modern global populations, particularly if cultural practices and local history are taken into account.

1.2.2 Uncovering the origin of modern humans using mtDNA and NRY SNPs

An initial population study analysing DNA sequence variation in human mtDNA by targeting RFLPs provided evidence for the correlation between mtDNA variation and geographical ancestry by demonstrating that RFLP patterns were shared across different racial groups (Brown, 1980). Two further RFLP studies, which explored mtDNA

variation in individuals representing a variety of geographical regions, identified significant mtDNA variation between major ethnic groups with the greatest variation occurring in Africa (Johnson *et al.*, 1983, Merriweather *et al.*, 1991). However, all mtDNA haplotypes were related under a single phylogenetic tree, with closely related mtDNA haplotypes forming clusters according to geographic origin (Johnson *et al.*, 1983, Merriweather *et al.*, 1991). This confirmed Brown's (1980) initial findings regarding the association of mtDNA variation with geographical origin and the African origin of humans, which culminated in the currently accepted 'Out of Africa' model (Cann *et al.*, 1987).

The utility of NRY SNPs in the investigation of human genetic variation became apparent much later. MtDNA is a small and abundant molecule and the genome can be entirely analysed at the molecular level relatively easily (Lansman *et al.*, 1981). However due to the much larger size of the Y chromosome the discovery of sequence polymorphisms such as RFLPS and SNPs was more difficult and it was not until the construction of Y-chromosome DNA libraries in the late 1980's that tools for the study of Y-DNA polymorphisms emerged. Numerous human Y specific DNA probes were isolated and used to identify the first NRY RFLPs and SNPs (Casanova, 1985, Lucotte and Ngo, 1985, Ngo *et al.*, 1986). This encouraged Y- chromosome applications in human evolution studies, although these earlier studies were aimed primarily at testing or complimenting inferences made using mtDNA and autosomal markers [For a review of early studies see Hammer and Zegura (1996)]

In 1995, Hammer conducted the first investigation of human NRY haplotype diversity across 16 global geographical regions (Hammer, 1995). A 2.6kb fragment referred to as the 'YAP' region that encompasses a 300bp Alu insertion was sequenced. This Alu element belongs to a family of short repeat sequences which contain the recognition site for the restriction enzyme *AluI* and which are polymorphic in diverse human genomes. The study uncovered three SNPs and a variable length poly(A) tail associated with the YAP element, which resulted in the identification of five YAP haplotypes (Hammer, 1995). These haplotypes showed the greatest frequency in African populations and were consistent with earlier mtDNA and autosomal propositions regarding the origin of humans in Africa, followed by dispersal to Europe and Asia. A subsequent study by Hammer *et al.* (1997) investigating these 5 aforementioned YAP haplotypes, along with

the DYS271 A-G transition (Seielstad *et al.*, 1994) and the DYS19 microsatellite (Roewer *et al.*, 1992) in 60 different global populations found similar results in support of the 'Out of Africa' model.

Other models of human origin have been proposed, including the multiregional continuity model which advocates that contemporary human populations originated on separate continents before evolving and diverging through a network of worldwide genetic exchanges into modern humans (Wolpoff *et al.*, 1984). However, there is currently little evidence to support such theories. In contrast, anthropological evidence (Stringer, 2002) compliments accumulated genetic data (Brown, 1980, Cann *et al.*, 1987, Seielstad *et al.*, 1994, Hammer, 1995) in support of the 'Out of Africa' model. Nonetheless, the successful extraction and sequencing of Neanderthal DNA has challenged our understanding of human origins. Neanderthals are a group of extinct Hominids believed to have coexisted with modern humans in Europe and parts of western Asia (Hublin, 2009). Early studies comparing Neanderthal and modern human DNA found no evidence of inbreeding (Krings *et al.*, 1997), however recent genetic studies have showed that a small percentage of Europeans, Central Asians and Siberians carry Neanderthal DNA whilst Africans do not, strongly suggesting the likelihood that inbreeding occurred between Neanderthals and non-African modern humans. (Krause *et al.*, 2007, Wall *et al.*, 2009, Green *et al.*, 2010, Sankararaman *et al.*, 2012, Wall *et al.*, 2013, Vernot and Akey, 2014). These novel findings indicate the strong possibility that a mixture of both the 'Out of Africa' and multiregional continuity model were plausible, where the majority of Modern Humans originated in parallel to Neanderthals and dispersed out of Africa but were subjected to complex interactions and inbreeding post-dispersal.

Current evidence for the origin of humans in Sub-Saharan Africa approximately 100,000 to 200,000 thousand years ago and subsequent dispersal throughout the World was recently reviewed by Beyin (Beyin, 2011). Several TMRCA estimates have been proposed using mtDNA and NRY data; however a specific estimate for the exact time since the emergence of modern humans remains to be identified. Interestingly, NRY estimates are more recent than mtDNA derived estimates and range from 46 to 91 thousand years ago (Pritchard *et al.*, 1999) to 60 to 130 thousand years ago (Tang *et al.*, 2002). In comparison, mtDNA estimates suggest a more ancient origin with estimates

ranging from 171±50 thousand years ago to 200-281 thousand years ago (Ingman *et al.*, 2000, Tang *et al.*, 2002). The twofold difference between mtDNA and NRY TMRCAs was proposed to have resulted from the reduced effective population size of males which arises from contrasting demographic factors between the sexes, including the common practice of patrilocal marriage, an increased rate of male mortality and the greater variability in reproductive success for males than females (Wilder *et al.*, 2004b).

Since the first formulation of the 'Out of Africa' model by Cann *et al*. (1987), much of the model has been elucidated using complementary studies of mtDNA and Y-chromosome diversity. The current model suggests that modern humans expanded throughout the African continent and dispersed out of East Africa into Eurasia about 100-150 thousand years ago (Beyin, 2011) (Figure 1.4). Two major dispersal routes have been proposed for the Africa-Eurasia migration which have been reviewed in Beyin (2011). The first supports an upward continental route through the Levantine corridor and the Near East, followed by dispersal into eastern and western Eurasia (Bar-Yosef and Belfer-Cohen, 2001, Maca-Meyer *et al.*, 2001, Derricourt, 2005, Olivieri *et al.*, 2006, Badro *et al.*, 2013), and the second, supports the dispersal of humans out of the Horn of Africa following a coastal route through to South Asia into eastern Eurasia and Oceania (Lahr and Foley, 1994, Macaulay *et al.*, 2005, Petraglia *et al.*, 2010, Armitage *et al.*, 2011, Mellars *et al.*, 2013). However, it appears likely that both routes were used concurrently (Luis *et al.*, 2004, Rowold *et al.*, 2007), although it has also been suggested that the Levantine corridor may have been the preferred route over the Horn of Africa (Rowold *et al.*, 2007). Dispersal then continued into northern Eurasia and east across the Americas (Torroni *et al.*, 1993, Bailliet *et al.*, 1994, Bianchi *et al.*, 1998, Hammer *et al.*, 1998, Bergen *et al.*, 1999, Richards *et al.*, 2000, Pakendorf *et al.*, 2006), followed by recent back migration of Eurasians from Europe and the Near East into North Africa (Hammer *et al.*, 1997, Richards *et al.*, 2000, Oliveri *et al.*, 2006). Polynesia was the last part of Oceania to be colonised about 2000-3000 years ago (Kayser *et al.*, 2006).

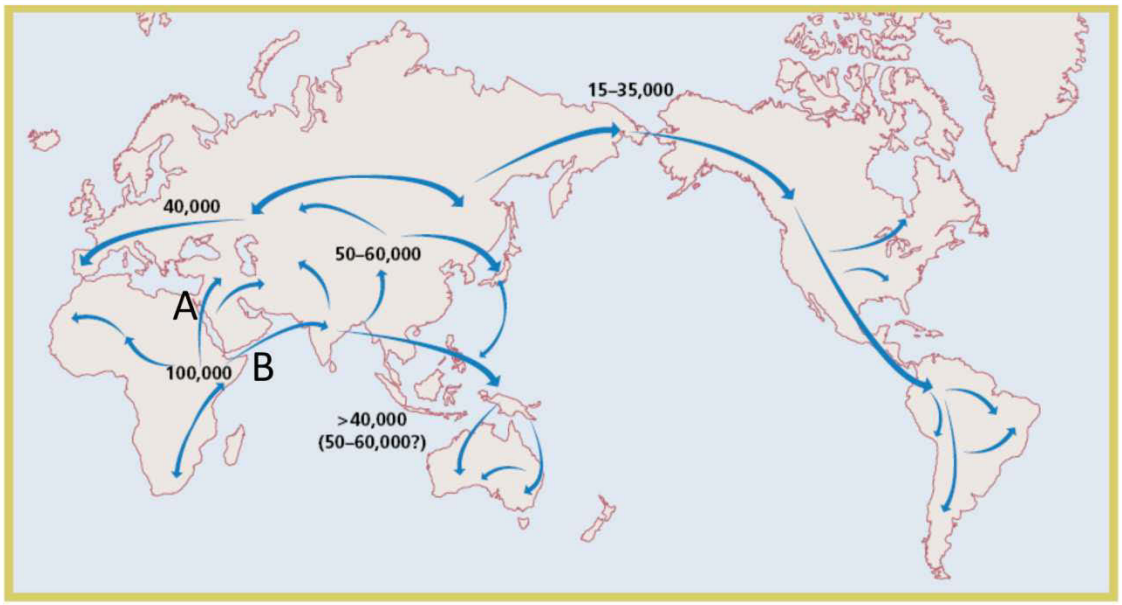


Figure 1.4: The migration of modern humans
The schematic depicts the 'Out of Africa 'origin of modern humans between 150,000 to 100,000 years ago followed by radiation from East Africa throughout the African continent. Dispersal and expansion occurred from the same region via a northern (A) and a southern (B) route between 60,000 to 40,000 years ago. Oceania, Europe and America were settled from Asia in this order. Image adapted from Cavalli-Sforza and Marcus (2003).

1.2.3 Reconstructing human mitochondrial and NRY phylogeny

In the 1990's, large-scale studies mtDNA of ethnically diverse subjects were carried out. These involved either RFLP survey of the complete mtDNA molecule, mainly targeting the coding region, or sequencing the HV1 control segment [reviewed in Torroni *et al.* (2006)]. The elevated mutation rate of the control region ($3x10^{-3}$ nucleotides/generation) (Sigurdardottir *et al.*, 2000) has been associated with relatively high levels of recurrent mutations that arise in different populations at the same site to create parallel haplotypes, resulting in a phenomenon known as homoplasy (Non *et al.*, 2007, Salas *et al.*, 2007). Contrastingly, the mtDNA coding region demonstrates a reduced frequency of homoplasies due to its much lower constant mutation rate of $2x\ 10^{-5}$ nucleotides/generation (Sigurdardottir *et al.*, 2000, Silva *et al.*, 2002). The abundance of HV1 homoplasies led to suggestions that using HV1 polymorphisms alone cannot adequately distinguish the basal structure of the mtDNA phylogenetic tree (Stewart, 1993, Finnila *et al.*, 2001). However, studies coupling high resolution RFLP

with HV1 sequencing produced good estimates of mtDNA phylogeny (Richards and Macaulay, 2001, Macaulay *et al.*, 2005). For instance, a study of Native American tribes demonstrated that HV1 SNPs complemented RFLPs to form complex haplotypes which could still be distinguished when reversions and/or parallel mutations occurred (Torroni *et al.*, 1993). Similarly, a study of Japanese mtDNA diversity using HV1 sequence polymorphisms identified 169 haplotypes and 147 polymorphic sites, which correlated with Japanese-specific haplotypes previously established using coding region polymorphisms (Maruyama *et al.*, 2003).

Population studies assessing mtDNA variation in different ethnic groups, using a variety of methods have identified countless numbers of mtDNA polymorphisms. This has resulted in the establishment of MITOMAP, a database of human mtDNA sequence variation which unifies data regarding 'mitochondrial genome structure and function, pathogenic mutations and their clinical characteristics, population associated variation, and gene-gene interactions' (Kogelnik *et al.*, 1996). This tool has allowed scientists to resolve the basal branching structure of mtDNA variation in most parts of the world, to produce a refined mtDNA phylogenetic tree. Torroni *et al.*'s (1993) analysis of Native American populations provided the framework for the current accepted mtDNA haplogroup nomenclature, by identifying four basal branches of the tree each representing separate groups of related haplotypes, termed haplogroups (A, B, C, D) (Torroni *et al.*, 1993). As the haplogroup structure of other continental populations became resolved, cladistic rules for the hierarchical organisation of haplogroups and sub haplogroups became firmly established (Richards, 1998). A complete mtDNA tree was constructed by Ruiz-Pesini *et al.* (2007) using over 3000 coding region SNPs and a small number of informative control region SNPs (Ruiz-Pesini *et al.*, 2007).

Despite the establishment of cladistic rules for mtDNA haplogroup designation (Richards, 1998) there still remained a lack of uniformity and consistency in mtDNA haplogroup nomenclature with different authors using the same name for different haplogroups or others using different names for the same haplogroup. A study investigating 55 publications describing mtDNA variation unified the nomenclature and produced an updated and complete mtDNA phylogeny using control and coding region data (van Oven and Kayser, 2009). A simplified form of the tree is shown in Figure 1.5. The updated mtDNA tree with universal nomenclature is now available online at

http://www.phylotree.org and is regularly updated, providing researchers with a valuable tool in their study of mtDNA diversity.

The complexity of the NRY and the initial difficulty encountered in finding polymorphisms masked its utility in human population studies and for the reconstruction of NRY phylogeny. Less than 60 NRY polymorphisms had been discovered by 1996, which included only 3 SNPs (The Y Chromosome Consortium, 2002). However use of the denaturing high performance liquid chromatography (DHPLC) technique for mutation detection enabled identification of 19 new SNPs in a sample of 718 males representing 54 global populations (Underhill *et al.*, 1997). A more complex phylogeny was reconstructed with each branch of the tree representing one haplogroup which are each defined by a specific SNP (Underhill *et al.*, 1997). Subsequently DHPLC has been used in various population studies investigating global NRY diversity, resulting in the discovery of over 200 NRY SNPs which have been used to produce good estimates of NRY phylogeny (Shen *et al.*, 2000, Underhill *et al.*, 2000, Hammer *et al.*, 2001, Underhill *et al.*, 2001). Whilst identification of NRY SNPs is more challenging than for mtDNA SNPs, the lower NRY mutation rate means that there is a reduced risk of homoplasy, which can be problematic when reconstructing phylogenies from the rapidly mutating mtDNA (Stewart, 1993, Finnila *et al.*, 2001). Thus, it has been suggested that phylogenetic reconstructions are facilitated when using NRY sequence polymorphisms and the derived phylogenetic relationships are more accurate, as the motif of accumulated sequence polymorphisms is more preserved in different paternal lineages (Jobling and Tyler-Smith, 2003).

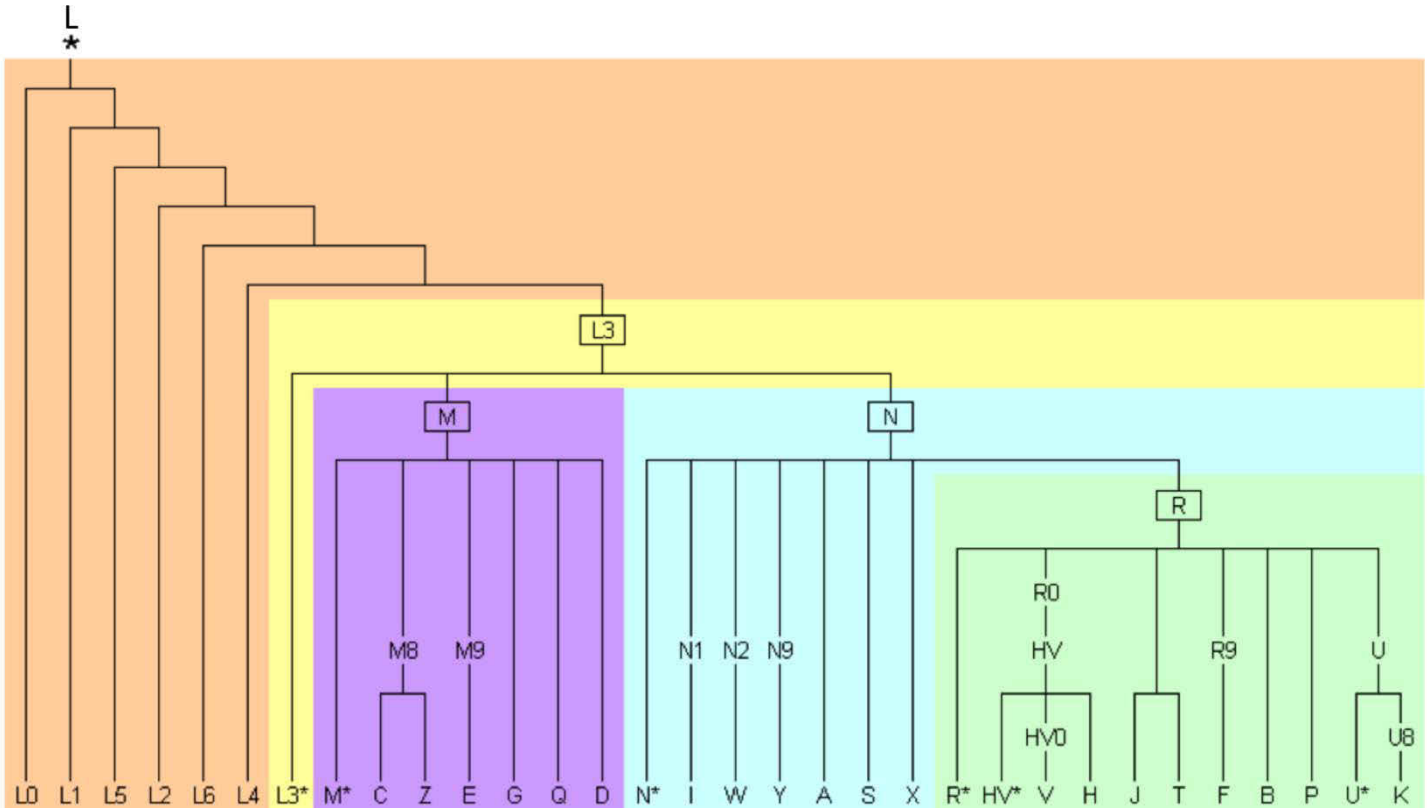


Figure 1.5: **Simplified schematic of the mtDNA phylogenetic tree**

The star at the root of the tree represents the most recent common maternal ancestor of modern humans. The African specific macro haplogroup L and its descendant haplogroups (orange) constitute the deep roots of the tree and represent the African origin of modern humans. Haplogroup L3 (yellow) gave rise to haplogroups M, N and R. All descendant haplogroups of L (orange), M (purple), N (blue) and R (green) are represented as alphabetical letters at the branch tips. Haplogroups accompanied by an asterisk (*) represent all other descendant lineages of a clade other than those already shown. Note that this tree is only a simplified depiction of the topology of major mtDNA haplogroups. For the complete global mtDNA phylogeny see the full tree at http://www.phylotree.org. Image modified from van Oven and Kayser (2009).

The explosion of NRY data was however problematic, as the increase in number of NRY polymorphisms became plagued by the establishment of various systems used to name newly discovered polymorphisms. By 2002, there were over seven different nomenclature systems in the literature, which complicated the comparison of results between different publications (The Y Chromosome Consortium, 2002). In 2002, the Y chromosome consortium (YCC) was established to gather and unify NRY data. The YCC genotyped 245 published markers in 74 males from different global populations and reconstructed a single phylogeny from the 153 haplogroups observed (The Y Chromosome Consortium, 2002). The tree produced showed various nested monophyletic clades (haplogroups), which make up the branches and consist of sets of paternal lineages which are related by a shared, derived state at one or several polymorphic sites. Interestingly, out of the 245 markers, there were only eight recurrent mutations which occurred across different haplogroups and as such could be easily distinguished (The Y Chromosome Consortium, 2002). The YCC also established a hierarchical nomenclature system for NRY haplogroups, which unified past systems. The cladistic rules for the hierarchical organisation of NRY haplogroups were based upon results from a previous study of 43 NRY SNPs in 50 human populations which identified hierarchical patterns of global NRY diversity (Hammer *et al.*, 2001). The YCC nomenclature system is now broadly accepted and has standardised the names of NRY haplogroups (Shen *et al.*, 2004, Pericic *et al.*, 2005, Kumar *et al.*, 2007, Rosa *et al.*, 2007, Battaglia *et al.*, 2009, Zhong *et al.*, 2011, Dulik *et al.*, 2012).

In 2008, Karafet *et al.* genotyped 599 NRY markers and produced a revised NRY tree following the YCC cladistic rules for haplogroup nomenclature and classification (Karafet *et al.*, 2008). The tree contains 311 distinct haplogroups including two new haplogroups, S and T (Karafet *et al.*, 2008). A simplified version of the tree is depicted in Figure 1.6. The selected markers included previously reported SNPs from a major study used to characterise DNA variation in 3 major populations which had identified 334 NRY SNPs in 33 males of African, European and Asian descent (Hinds *et al.*, 2005) but also included SNPs that had been previously mapped to the 2002 YCC tree and newly identified SNPs from RFLP or direct sequencing PCR (Karafet *et al.*, 2008). The major rearrangements to the 2002 YCC tree included new sub-groups of major previously identified haplogroups, which are defined by newly identified SNPs or the

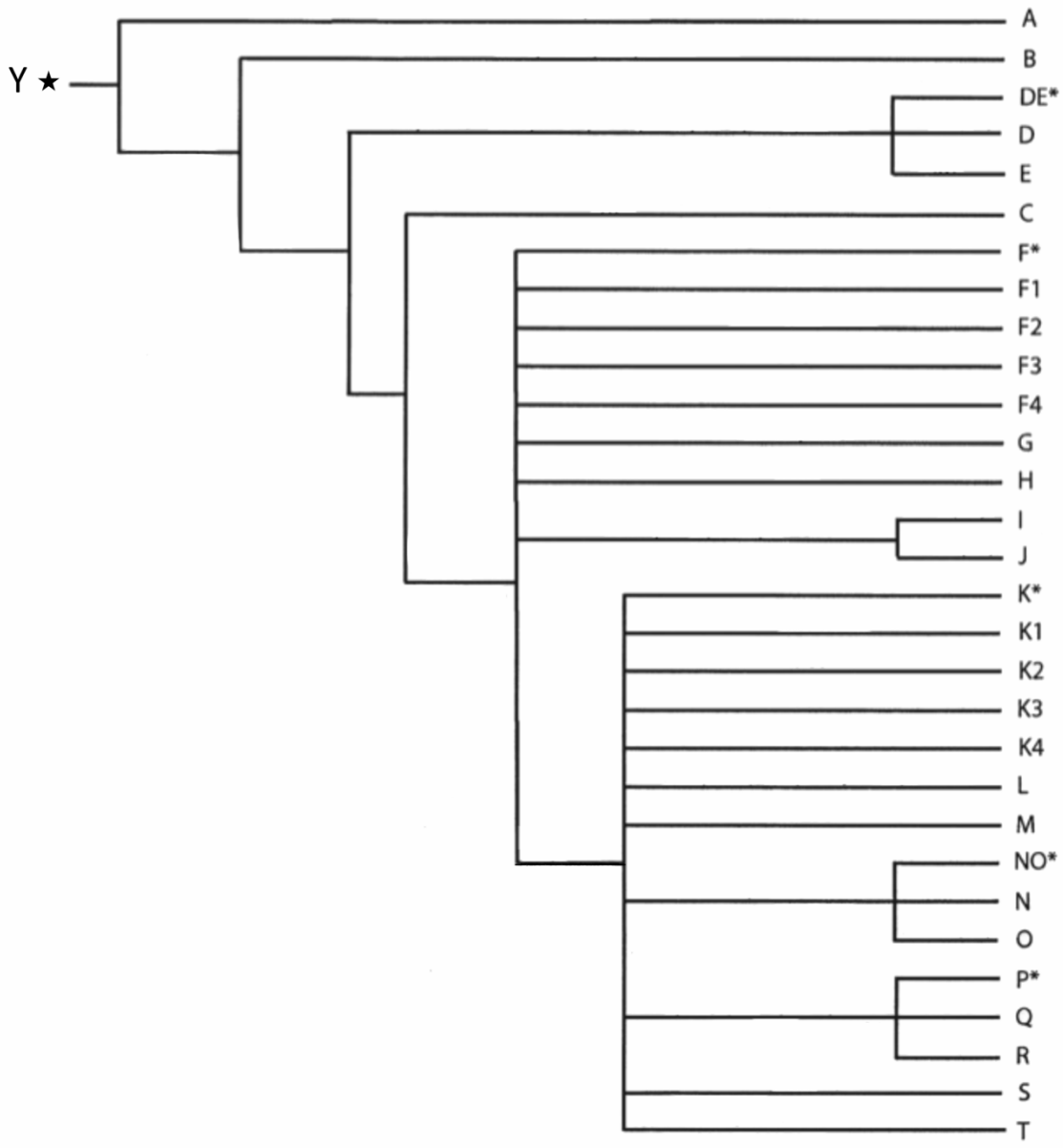


Figure 1.6: **Simplified schematic of the NRY phylogenetic tree.**

The star at the root of the tree represents the most recent common paternal ancestor of modern humans. The haplogroups are represented as capital letters at the branch tips. Defining mutations for each haplogroup are shown along the branches or in the box on the bottom left for those shorter branches of the tree. Haplogroups accompanied by an asterisk (*) represent all other descendant lineages of a clade other than those already shown. Note that this tree is only a simplified depiction of the topology of major NRY haplogroups. For the complete global NRY phylogeny see the full tree at http://www.isogg.org/tree/index.html. Image modified from Karafet *et al.* (2008).

revision and restructuring of several haplogroups and their branches, including haplogroup E, K, N and R (Karafet *et al.*, 2008). The International Society of Genetic Genealogy was established to create a web-based NRY phylogeny using this unified nomenclature and which is routinely updated to incorporate developments in the field. The most recent NRY tree is available online at http://www.isogg.org/tree/index.html (International Society of Genetic Genealogy, 2012).

1.2.4 Global distribution of mtDNA and NRY haplogroups

Extensive studies of mtDNA and NRY diversity in different populations have revealed correlations between the major branches of the trees, which represent major haplogroups and continental/regional populations, indicating that mtDNA and NRY haplogroup distributions can vary significantly at the sub-population level [reviewed in Underhill and Kivisild (2007)]. The distinct trend observed in population studies is the occurrence of haplogroups at high frequencies in particular geographical regions, coupled with either complete absence in other regions, or low frequencies usually in neighbouring regions that are attributed to migration or admixture. For example, mtDNA paragroup L and its haplogroup L3 (refer to Figure 1.5) are overwhelmingly predominant in Sub-Saharan populations, but with low frequencies detected in North Africa and the Middle East (Salas *et al.*, 2002, Gonder *et al.*, 2007).

Paragroup M and its various haplogroups predominate in eastern Eurasia, for example relatively high frequencies of mtDNA haplogroup D is found in all Asian regions except South Asia (Torroni *et al.*, 1993, Kivisild *et al.*, 2002, Metspalu *et al.*, 2004, Kong *et al.*, 2006, Sun *et al.*, 2006). Europeans are predominantly mtDNA haplogroups H, U, J and T however these haplogroups also predominate in North Africa and the Middle East (Richards *et al.*, 2000, Finnila *et al.*, 2001, Brandstatter *et al.*, 2003, Badro *et al.*, 2013). Haplogroup distributions frequently also vary at the regional level. For example the incidence of mtDNA haplogroup T varies in Western Mediterranean populations, with a lower incidence in Andalousians and Basques compared to other Spanish, Portuguese and Italian populations (Plaza *et al.*, 2003). Major differences in mtDNA haplogroup distribution between various regional districts of France have also been identified (Dubut *et al.*, 2004) and distinct differences in mtDNA haplogroup diversity in the Han

ethnic group, which constitutes 93% of the Chinese population, have been identified in different regions of China (Yao *et al.*, 2002, Gan *et al.*, 2008).

Similar trends are found for NRY haplogroups, where haplogroup E is predominant in Sub-Saharan Africa but a lower incidence is also observed in North Africa, the Middle East and Europe (Cruciani *et al.*, 2004, Luis *et al.*, 2004, Cruciani *et al.*, 2007, Badro *et al.*, 2013). The NRY haplogroups Y, BT and B also occur in relatively high frequencies in Sub-Saharan Africans (Luis *et al.*, 2004, Gomes *et al.*, 2010, de Filippo *et al.*, 2011, Badro *et al.*, 2013). Haplogroups R1b and I predominate in Europeans but lower incidences of other haplogroups, like G and J, which are typically prevalent in North Africa and the Middle East, are observed in Europe (Cinnioglu *et al.*, 2004, Battaglia *et al.*, 2009, Myres *et al.*, 2011, Badro *et al.*, 2013). Again regional variations exist, for example NRY haplogroup I occurs at a high frequency in Sardinia but is relatively low in the Basque region where haplogroup R1b predominates (Underhill *et al.*, 2000). Furthermore R1b is observed in greater frequencies in Western Europe compared to Eastern Europe where R1a is more prevalent (Kayser *et al.*, 2005). On the other hand, whilst NRY haplogroups C, K and O predominate in Asia (Kayser *et al.*, 2003, Sengupta *et al.*, 2006, Xue *et al.*, 2006) there exists differences in haplogroup frequencies between various Asian populations. NRY haplogroup C is found in high frequencies in North Asian and Native American populations (Bianchi *et al.*, 1998, Zerjal *et al.*, 2002, Nasidze *et al.*, 2005a, Dulik *et al.*, 2012) contrastingly to haplogroup O which is predominant in South, East and South-East Asia (Kayser *et al.*, 2006, Xue *et al.*, 2006, Zhong *et al.*, 2011). A moderate frequency of haplogroup O is also observed in South Asians and Oceanians, with sub-haplogroup O2a, occurring at high frequencies in central and east Indian populations (Kumar *et al.*, 2007, Sharma *et al.*, 2012).

It is evident that the mtDNA and NRY trees clearly depict continent and region specific variation and phylogeny. This has not only been relevant to trace human dispersal and evolution, but has also played a fundamental role in the advancement of modern forensic intelligence tools allowing for the prediction of biogeographical ancestry. In this project the distribution of major mtDNA and NRY haplogroups in continental populations of interest have been compiled from haplogroup frequencies reported in the literature and are described in detail in chapter 3.

1.3 Autosomal SNPs

Whilst the applications of haploid markers in population genetics flourished, autosomal markers have had a less prominent role in studies of human evolution. The main reason for this is that differences in autosomal polymorphisms between different populations can arise from both mutation, random assortment and recombination, unlike haploid polymorphisms, which solely arise due to mutational events and are inherited as a block by each generation within that specific lineage (Jorde *et al.*, 1998). Prior to the human genome project, only a limited number of autosomal polymorphisms exhibiting distinct population differences had been discovered. These markers primarily emerged from disease association studies, particularly those assessing differences in patterns of genetic variation in blood grouping genes such as the ABO genes (Iwasaki *et al.*, 2000) and the FUT-2 locus (Pang *et al.*, 2001), tumor necrosis factor genes (Bridges *et al.*, 2002), cytochrome genes (Wooding *et al.*, 2002) and the β-globin gene (Wainscoat *et al.*, 1986) in various populations. Autosomal polymorphisms associated with pigmentation genes have also been of interest for applications in population genetics, since phenotypic characteristics such as hair, eye and skin colour are distinctly different between various global populations (Alaluf *et al.*, 2003, Shriver *et al.*, 2003, Myles *et al.*, 2007, Soejima and Koda, 2007).

1.3.1 Autosomal SNPs: ancestry informative markers

Wainscoat *et al.* (1986) investigated five RFLPs in the human adult β-globin gene in individuals from eight diverse global populations and showed that African populations were characterised by a unique haplotype solely restricted to that group, whilst other global populations demonstrated diverse haplotypes which were common amongst them (Wainscoat *et al.*, 1986). This supported previous mtDNA evidence in favour of the proposed 'Out of Africa' hypothesis (Brown, 1980, Johnson *et al.*, 1983) and suggested that patterns of genetic diversity derived from autosomal polymorphisms could be used to reconstruct human phylogeny (Wainscoat *et al.*, 1986). A later study of β-globin gene RFLPs found similar results but advocated that polymorphisms located in small gene clusters, such as the β-globin gene, could not solely be used to calculate divergence and TMRCA estimates (Long *et al.*, 1990). This culminated in various studies investigating numerous autosomal RFLPs scattered throughout the human genome, from which were

reconstructed only simple human phylogenies (Bowcock *et al.*, 1991, Mountain and Cavalli-Sforza, 1994). However the few phylogenies derived from autosomal markers were not as informative as those obtained using haploid markers because autosomal polymorphisms are affected by recombination during meiosis. This has complicated the process of tracing genealogies and consistent TMRCA estimates from different autosomal genes have been more difficult to obtain, with most varying considerably between 500,000 to 1 million years ago (Mountain and Cavalli-Sforza, 1994, Harding *et al.*, 1997, Templeton, 2002). These estimates are approximately four to five times greater than those obtained using mtDNA and NRY data, which was attributed to the larger effective size of the autosomal DNA population relative to the mtDNA or Y chromosome DNA population (due to the random assortment of diploid maternal and paternal alleles) (Cavalli-Sforza and Marcus, 2003). Furthermore, heterozygote advantage, a form of balancing selection has also been suggested to increase TMRCA estimates (Cavalli-Sforza and Marcus, 2003).

Although recombination has disadvantaged the application of autosomal polymorphisms in human evolutionary studies, study of autosomal ancestry instead of uniparental ancestry has been of increasing interest in recent years due to its ability to capture the genetic diversity contributed by all ancestors rather than just a limited few within a uniparental lineage (Royal *et al.*, 2010). Thus, autosomal polymorphisms can provide useful data regarding more recent genetic events. Autosomal markers have been utilised to measure the degree of admixture and population structure within particular population groups, sometimes in conjunction with mtDNA and Y chromosome data, in order to investigate variations in the contributions of ancestral populations across the different classes of DNA (Jorde *et al.*, 1998, Jorde *et al.*, 2000, Mesa *et al.*, 2000, Collins-Schramm *et al.*, 2004, Yang *et al.*, 2005, Lind *et al.*, 2007, Li *et al.*, 2008, Watkins *et al.*, 2008, Corach *et al.*, 2010, Lao *et al.*, 2010). In recent years, informative sets of autosomal markers useful in admixture mapping have been compiled (Halder *et al.*, 2008, Baye *et al.*, 2009, Kosoy *et al.*, 2009, Galanter *et al.*, 2012, Nievergelt *et al.*, 2013). Such applications of autosomal polymorphisms have clarified certain aspects of human evolution such as migrations, bottlenecks and local demographic processes including sex-based gene flow, which may be responsible for differences in admixture proportions between different DNA classes (Lind *et al.*, 2007, Hunley *et al.*, 2009).

1.3.2 Autosomal markers of human pigmentation

One of the most notable examples of interpopulation differences in human phenotypic variation is reflected by the range of hair, eye and skin colour pigmentation. This has promoted the desire to identify key pigmentation genes and their polymorphic variants that characterise distinct phenotypic differences at the individual as well as interpopulation level. Human pigmentation is controlled by the combined effects of multiple genes (Sturm *et al.*, 2001). Many SNPs in genes associated with normal human pigmentation variation within or between populations have been identified to date and will be discussed in more detail in chapter 4. A number of these polymorphic variants have also been linked to disease states and are reviewed in Sturm (2006) and Tully (2007).

1.4 SNP Databases

The Human Genome Project has resulted in the identification of millions of autosomal polymorphisms and the collaborative International HapMap project was subsequently established to categorise human genetic variants shared across different population groups of African, Asian and European ancestry into a single database (The International HapMap Consortium, 2003). By the end of Phase II, a total of 3.5 million SNPs were identified (The International HapMap Consortium, 2007), with the final phase of the HapMap project, known as HapMap3 resulting in the creation of an integrated dataset containing 1.5 million common and rare SNP and copy number variant alleles (The International HapMap Consortium, 2010). The Human Genetic Diversity project was a parallel effort which developed a Human Genome Diversity Panel (HGDP) comprising a collection of cell lines from 1050 individuals from 52 global populations that was compiled from laboratories worldwide to provide a source of DNA and RNA for studies of modern human population history (Cavalli-Sforza, 2005). All genotyping and sequencing results generated using this panel are submitted to a central database (available at http://www.cephb.fr/en/hgdp/main.php) that currently contains 1 million SNPs as well as microsatellites and indels.

The emergence of next-generation sequencing platforms have contributed to the recent 1000 Genomes project which aims to identify genetic variants with a frequency of at

least 1% in the European, East and South Asian, West African and Native American populations studied (Abecasis *et al.*, 2010). The ultimate goal of the project is to sequence 2,500 samples from these five global population groups with the Phase 1 sequencing results for 1092 individuals now freely available at http://www.1000genomes.org/home. Not only can the sequences, genotypes and allele frequencies be accessed on each of these individual databases, the National Center for Biotechnology Information database of Single Nucleotide Polymorphisms (dbSNP) of nucleotide sequence variation is a public domain archive of SNPs, indels and microsatellites which also houses the HapMap, HGDP-CEPH and 1000 Genomes data (Sherry *et al.*, 2001). The dbSNP also accepts individual laboratory submissions of variations from any species, from any part of the genome. It is possible to search SNPs and other variants in the dbSNP using identifiers known as reference SNP (refSNP) cluster ID numbers (rs#) or submitted SNP accession numbers (ss#) (Sherry *et al.*, 2001).

As a result of such collaborative efforts, AIMs which include polymorphisms in the autosomes, mtDNA and sex chromosomes have been identified which exhibit distinct allele frequency differences between ancestral populations. From these databases, several panels of informative AIMs have since been reported (Shriver *et al.*, 2003, Collins-Schramm *et al.*, 2004, Smith *et al.*, 2004, Yang *et al.*, 2005, Kim and Borevitz, 2006, Halder *et al.*, 2008, Baye *et al.*, 2009, Kosoy *et al.*, 2009, Paschou *et al.*, 2010, Kidd *et al.*, 2011, Galanter *et al.*, 2012). These panels have revolutionised the applications of genetic polymorphisms in admixture mapping and association studies, particularly for the inference of physical appearance and biogeographical ancestry.

1.5 Applications of genetic polymorphisms in forensic casework:

Advances in DNA sequencing and typing have not only enabled the identification of distinct patterns of human genetic variation within and between various global populations, but have also contributed to the progress of forensic science capabilities. The forensic applications of genetic polymorphisms have come a long way since the initial use of DNA typing in forensic casework in 1986 (Gill *et al.*, 1985). At present, DNA evidence constitutes one of the primary sources of identification evidence and

DNA typing using individual variation in STR length is the commonest technique utilised for the analysis of biological forensic evidence (Butler, 2004). However, there has been an increased interest in establishing new DNA analysis systems targeting AIMs to obtain intelligence information about unknown DNA samples.

1.5.1 SNP analysis in forensic identification

DNA typing for forensic identification is traditionally performed using nuclear STRs due to the abundance of alleles per locus, which confers a higher discriminatory power than can be achieved with nuclear SNPs (Butler *et al.*, 2007). It is estimated that 50-100 SNPs are required to achieve the same degree of discrimination as current 9-15 STR loci multiplex DNA profiling kits (Gill, 2001, Gill *et al.*, 2004), which suggests that forensic SNP assays must demonstrate a high multiplexing capability as well as suitability for the analysis of forensic biological evidence. However a major difficulty encountered in forensic DNA profiling is that conventional STR typing requires amplicons that are too large for the analysis of nDNA extracted from trace amounts of biological sample or from samples that have been compromised by exposure to extreme environments. In such cases SNP typing can be valuable because these polymorphisms can be typically analysed on smaller amplicons. Presently, various multiplex autosomal SNP assays for forensic identification have been developed (Dixon *et al.*, 2005, Kidd *et al.*, 2006, Sanchez *et al.*, 2006, Phillips *et al.*, 2007a, Kim *et al.*, 2010, Pakstis *et al.*, 2010, Lou *et al.*, 2011).

The use of SNP microarrays as a supplement to standard STR typing to identify individuals in a mixture has also been demonstrated (Homer *et al.*, 2008). Alternatively, the application of insertion/deletion (INDEL) polymorphisms for forensic identification has been recently explored because these markers combine the desirable features of STRs and SNPs such as a widespread distribution throughout the genome, variable allele frequencies and analysis on short amplicons that is amenable to large-scale automated multiplexing (Weber *et al.*, 2002, Mills *et al.*, 2006). This has resulted in the development of several multiplex INDEL assays and panels (Pereira *et al.*, 2009, LaRue *et al.*, 2012, Li *et al.*, 2012, LaRue *et al.*, 2014), with the Investigator DIPlex® INDEL typing kit (Qiagen, USA) demonstrating suitability to the analysis of degraded samples (Hollard *et al.*, 2011). However, due to the greater discriminatory power achieved with

STRs (Butler *et al.*, 2007) compared to SNPs and INDELs (Neuvonen *et al.,* Amorim and Pereira, 2005), particularly for parentage and kinship analysis, recent developments have also investigated reduced sized STR amplicons incorporated in several commercially available 'mini-STR' assays that are specifically designed for the analysis of degraded material (Parsons *et al.*, 2007, Hill *et al.*, 2008).

Nonetheless, severely degraded samples or biological samples containing limited nDNA (shed hairs, bone, teeth) can still be problematic to analyse using autosomal SNP and mini-STR typing, as much lower quantities of nDNA is present in these tissues (Andréasson et al., 2006). More recently, specific SNPs located in nucleosomes, genomic regions that are protected from DNA degradation by the histone-DNA complex, have been incorporated in a nucleosome SNP multiplex assay that could achieve greater genotyping success rates for severely degraded DNA compared to existing autosomal SNP and mini-STR assays (Freire-Aradas et al., 2012). Whilst this research demonstrates a promising future for the application of novel autosomal markers, the use of mtDNA typing remains advantageous due to the high mtDNA copy number per cell (more than one thousand copies in certain cell types) (Miller *et al.*, 2003, Legros *et al.*, 2004).

MtDNA typing is usually performed for disaster victim identification (DVI) and missing person cases where the DNA evidence is usually compromised and nDNA recovery less likely (Anjos *et al.*, 2004, Sudoyo *et al.*, 2008, Coble *et al.*, 2009, Just *et al.*, 2009, Alvarez-Cubero *et al.*, 2012, Templeton *et al.*, 2013). Currently, mtDNA typing involves sequence analysis of the HV1 and HV2 region, targeting a total of 610bp of the control region (Figure 1.3), which encompasses the region containing the highest mutation rate and thus the greatest level of variation between individuals (Parson and Bandelt, 2007). However, mtDNA HV1/HV2 forensic typing has a lower discriminatory power than nDNA typing due to the non-recombining matrilineal mode of inheritance of the molecule which is inherited as a single locus and thus, results in no, or only slight differences in the mtDNA profile of members of the same maternal line [For review see Zietkiewicz *et al.* (2012)]. Despite the higher mitochondrial SNP mutation rates, nuclear SNP profiles are more individualized due to random assortment and recombination during meiosis (Kaestle *et al.*, 2006). The abundance of homoplasies due to recurrent mutations in the control region have also resulted in numerous common

HV1/HV2 haplotypes currently indistinguishable solely based on HV1/HV2 data (Stewart, 1993, Finnila *et al.*, 2001). This is problematic in forensic casework as there is a high probability that two unrelated individuals will match by chance due to identical HVI/HV2 alleles being shared as a result of parallel emergence in different lineages rather than inheritance from a common ancestor (Non *et al.*, 2007, Salas *et al.*, 2007). However, the discriminatory power of mtDNA typing has been increased using informative coding region SNPs in addition to HV1/HV2 sequencing (Brandstatter *et al.*, 2003, Coble *et al.*, 2004, Quintans *et al.*, 2004, Alvarez-Iglesias *et al.*, 2007, Andreasson *et al.*, 2007, Nilsson *et al.*, 2008, Kohnemann *et al.*, 2009).

The typing of Y-STRs and SNPs has also demonstrated applicability in forensic casework to characterise DNA from a male perpetrator in biological evidence recovered from sexual assault and rape cases (Honda *et al.*, 1999, Betz *et al.*, 2001, Prinz *et al.*, 2001, Prinz and Sansone, 2001, Sibille *et al.*, 2002, Maiquilla *et al.*, 2011). The primary form of evidence in such cases is a post-coital swab or semen sample that often contains a mixture of both the victim and male perpetrator's DNA. The two DNA profiles are usually able to be distinguished through the use of a differential DNA extraction method which separates the DNA from epithelial cells of the female victim from the more robust sperm cells of the male perpetrator (Butler, 2005). Y-chromosome typing is applied in such casework scenarios when the signal from the female DNA overwhelms the male DNA, or when it is difficult to separate male and female cell types, for example in vasectomised or azoospermic male perpetrators, where the only source of male DNA is epithelial cells from the exterior of the penis or those lining the urethra and which are shed during ejaculation (Honda *et al.*, 1999, Prinz *et al.*, 2001, Prinz and Sansone, 2001). More recently the applications of Y-chromosome typing have also been extended to the analysis of other mixed samples including saliva after kissing (Kamodyova *et al.*, 2013) and fingernail scrapings (Malsom *et al.*, 2009).

1.5.2 Applications of genetic polymorphisms in forensic intelligence

1.5.2.1 *Prediction of biogeographical ancestry and physical appearance*

STR DNA profiling for identification which is routinely used in forensic casework is of limited value when reference samples are not available, as these are required for

comparison and the calculation of match probabilities for unknown DNA profiles (Jobling and Gill, 2004). In cases where biological samples have been collected from a crime scene in the absence of eyewitness evidence, it is of great interest to obtain intelligence information to narrow the collection of potential matches and facilitate investigations. Information regarding an individual's physical characteristics (skin, hair, eye colour) or ancestry can provide informative sources of intelligence. For example, biogeographical ancestry (BGA) inferences would provide invaluable data in cases of mass casualties such as air crash disasters that usually involve compromised human remains, often of individuals from diverse geographic regions. Consequently, ancestry and phenotypic information would assist the identification process by reducing the search for reference samples or of potential matches to individuals of a specific ancestry/physical appearance [for review see Kayser and de Knijff (2011)]. The identification of patterns of genetic variation in the autosomes and haploid genomes that are characteristic of specific geographical regions (Torroni and Wallace, 1994, Underhill *et al.*, 2000, Hammer *et al.*, 2001, Pang *et al.*, 2001, Rosenberg *et al.*, 2002, Collins-Schramm *et al.*, 2004, Kayser *et al.*, 2006, Li *et al.*, 2008, Lao *et al.*, 2010, Howes *et al.*, 2011) and of EVCs (Frudakis *et al.*, 2003a, Graf *et al.*, 2005, Frudakis *et al.*, 2007, Sulem *et al.*, 2007, Han *et al.*, 2008, Kayser *et al.*, 2008b, Sturm *et al.*, 2008, Sulem *et al.*, 2008, Branicki *et al.*, 2009, Mengel-From *et al.*, 2010, Candille *et al.*, 2012, Zhang *et al.*, 2013, Durso *et al.*, 2014) indicates that such polymorphisms can be informative sources of forensic intelligence.

The use of genetic variation as markers of biogeographical ancestry was explored as early as 1994 through the use of sequence-specific oligonucleotide (SSO) hybridisation to the mitochondrial hypervariable regions for ancestry prediction of 525 individuals from 5 different ethnic groups (Connor and Stoneking, 1994). Both logistic regression and discriminant function analysis were applied and yielded similar classification accuracy results ranging from 31.1-81% and 36.5-80.3% respectively. The lower classification accuracy achieved for certain ethnic groups, such as South-East Asians and Japanese, was attributed to the fact that the haplotypes detected might be shared across ethnic groups, which may contain migrants from other mtDNA lineages because socially defined ethnic groupings do not always correlate with the mtDNA phylogeny (Connor and Stoneking, 1994). This study showed that there may be limitations to using mtDNA to classify ethnic groups, however though the classification accuracy achieved

was rather average, the study clearly demonstrated the ability to predict ancestry from DNA.

In 1997, STRs were used to estimate the ethnic affiliation of major US populations using log likelihood ratios as a classification approach (Shriver *et al.*, 1997). The study identified a series of informative STRs for the ethnic affiliation estimation of African, European and Hispanic ancestry and further demonstrated the utility of genetic markers in predicting BGA. In 2003, a linear discriminant classification method incorporating 56 autosomal pigmentation SNPs into a classifier model for the inference of ancestry was developed (Frudakis *et al.*, 2003b). The assay was able to correctly identify individuals of European, African and Asian ancestry with 99, 98 and 100% accuracy, respectively. The classification method showed comparable accuracy to the log likelihood ratio approach applied by Shriver *et al.* (1997) on STR data. This earlier log likelihood classification method was criticised by Brenner (1998) as being inappropriate in establishing classification accuracy and only useful in assigning a value to genetic markers for BGA inferences. This was supported by Frudakis *et al.* (2003b) who also claimed that though their linear classifier achieves similar classification accuracy to the log likelihood method, more SNPs are required than STRs to achieve the same resolution and that their technique is relevant only for inferences of major ancestry affiliations. Frudakis et al., (2003b) proposed that due to the underlying level of admixture in populations, a classification method which infers ancestry proportions (admixture) in individuals may be more accurate.

Interestingly, a study published a short time prior to Frudakis *et al.* (2003b) utilised admixture mapping to measure the effects of candidate genes on pigmentation as well as individual ancestry using a panel of 34 autosomal AIMs and demonstrated that it is possible to estimate the ancestry and physical appearance of an individual using DNA analysis and an appropriate number of informative markers (Shriver *et al.*, 2003). The incorporation of highly informative markers suggests that a reduced number of polymorphisms are required to be genotyped to obtain meaningful predictions, which has promoted the need to measure the extent at which specific markers contribute to ancestry inferences. A statistic, I_n, was proposed which measures the informativeness of markers for the inference of ancestry (Rosenberg *et al.*, 2003). However, this test was developed using autosomal markers and has not been applied to haploid markers. The

informativeness of haploid markers is reflected through the population frequencies and continental specificity of the haplogroups they identify, which are now widely accepted.

The Shriver *et al.* (2003) study thus paved the way for the development of assays for the inference of BGA and EVCs. Several multiplex assays for the inference of BGA have now been developed which target autosomal (Ray *et al.*, 2005, Phillips *et al.*, 2007b, Halder *et al.*, 2008, Daniel *et al.*, 2009, Kersbergen *et al.*, 2009, Nassir *et al.*, 2009, Lao *et al.*, 2010, Londin *et al.*, 2010, Paschou *et al.*, 2010, Pereira *et al.*, 2012, Nievergelt *et al.*, 2013, Phillips *et al.*, 2013) or NRY markers (Brion *et al.*, 2005, Wetton *et al.*, 2005, Onofri *et al.*, 2006, van Oven *et al.*, 2011a). More recently, autosomal SNP multiplex assays for the inference of physical characteristics such as eye, hair and skin colour have also been developed (Liu *et al.*, 2009, Valenzuela *et al.*, 2010, Branicki *et al.*, 2011, Spichenok *et al.*, 2011, Walsh *et al.*, 2011, Allwood and Harbison, 2013, Ruiz *et al.*, 2013, Walsh *et al.*, 2013). In 2009, 10 autosomal SNPs located within 6 pigmentation candidate genes were incorporated in an assay that was used to successfully infer both the biogeographical origin and pigment phenotype of ancient skeletal remains (Bouakaze *et al.*, 2009). Very recently a 50-SNP assay comprised of three multiplex SBE reactions was developed and demonstrated the accurate prediction of BGA (98.6%) and eye colour (for 61% of declared Europeans) among primary United States populations (Gettings *et al.,* 2014).

Although the utility of mtDNA for inference of BGA is clearly evident from population studies, mtDNA SNP assays for BGA prediction have been slower to develop. This may be attributed to the conviction that mtDNA polymorphisms demonstrate reduced geographical informativeness compared to those in the Y-chromosome due to the widespread practice of patrilocality, which has favoured the analysis of NRY SNPs as the preferred markers of BGA (Seielstad *et al.*, 1998). Nevertheless inferences based on mtDNA SNPs are also informative and may provide the sole source of information for compromised and highly degraded samples (Alvarez-Cubero *et al.*, 2012). Since the first applications of mtDNA to predict ancestry in 1994 (Connor and Stoneking, 1994), several successful multiplex assays have since been developed (Nelson *et al.*, 2007, Schlebusch *et al.*, 2009, Paneto *et al.*, 2011, van Oven *et al.*, 2011b, Ballantyne *et al.*, 2012). These assays have shown improved classification accuracy compared to those achieved by Connor and Stoneking (1994) as they have incorporated more informative

population specific markers identified through the newly refined mtDNA phylogeny. All of these autosomal, NRY and mtDNA SNP assays have incorporated a variety of different SNPs, genotyping techniques and assignment methods for prediction, yet all have shown a strong ability to successfully infer the BGA and EVCs of individuals from diverse and admixed populations.

1.5.2.2 *Genotyping strategies*

The current genotyping strategy of choice for most DNA intelligence assays has been the minisequencing technique using the SNaPshot™ kit (Life Technologies, USA). This genotyping strategy involves the single base extension of unlabelled primers with labelled dideoxyribonucleotide triphosphates (ddNTPs) (Budowle, 2004). The extended primers are separated using capillary electrophoresis and the allele is identified by detection of the incorporated fluorescent tag (Budowle, 2004). The SNaPshot™ kit has been widely used for the development of identification and intelligence SNP assays (Brandstatter *et al.*, 2003, Vallone *et al.*, 2004, Sanchez *et al.*, 2005, Onofri *et al.*, 2006, Phillips *et al.*, 2007b, Daniel *et al.*, 2009, Lao *et al.*, 2010, Lou *et al.*, 2011, Ballantyne *et al.*, 2012, Ruiz *et al.*, 2013, Walsh *et al.*, 2013) and has been identified as the leading technique for the analysis of forensic samples and for the validation of SNP allele frequencies (Amigo *et al.*, 2008a) as it is relatively cheap, robust, highly sensitive and has high multiplexing capacity with multiplexes simultaneously typing up to 42 SNPs being developed (Paneto *et al.*, 2011). The technique also only requires instrumentation (automated DNA sequencer) already commonly found in forensic laboratories (Budowle, 2004).

Matrix-assisted laser desorption ionisation time of flight (MALDI-TOF) mass spectrometry (MS) has also been used as an effective alternative to capillary electrophoresis to genotype multiplex single-base extension products (Bray *et al.*, 2001, Wise *et al.*, 2003, Mengel-Jorgensen *et al.*, 2004, Sun and Guo, 2006), with several human identification assays (Petkovski *et al.*, 2005, Johansen *et al.*, 2013) and patrilineage (Paracchini *et al.*, 2002, Lessig *et al.*, 2005) and matrilineage (Xiu-Cheng Fan *et al.*, 2008, Cerezo *et al.*, 2009) tracing assays designed specifically for the MALDI-TOF MS genotyping platform. Other genotyping strategies explored for the development of forensic identification SNP assays include pyrosequencing (Andreasson

et al., 2007) and microarray technology, which has demonstrated high throughput and high multiplexing capability with the analysis of 150 coding mtDNA SNPs reported (Sigurdsson *et al.*, 2006).

Numerous genotyping technologies have contributed to the establishment of population diversity repositories comprising millions of DNA markers (International Human Genome Sequencing Consortium, 2001, The International HapMap Consortium, 2003, 2007, 2010), which have been invaluable to the identification of panels of AIMs (Kersbergen *et al.*, 2009, Kosoy *et al.*, 2009, Nassir *et al.*, 2009, Paschou *et al.*, 2010, Galanter *et al.*, 2012). Genotyping technologies capable of simultaneously typing thousands of markers scattered throughout the human genome, such as the GeneChip® Human Mapping (Affymetrix Inc., USA) and HumanCore (Illumina Inc., USA) arrays have been developed, potentially opening the possibility of individual phenotypic determination. Advances in large-scale next generation sequencing technologies such as MiSeq (Illumina, Inc., USA) and SOLiD (Life Technologies, USA) have resulted in the increasingly widespread availability of this technology in research laboratories in recent years [reviewed in Kumar et al. (2012)] and is likely to prove an invaluable tool for forensic SNP detection.

The application of oligo-ligation assays, which utilise fluorescent labelled probes specific to each SNP/allele combination that are then identified using capillary sequencers, has also been explored for the development of both forensic identification and intelligence assays. One example of an oligo-ligation assay includes the GenPlex™ (Life Technologies, USA) system. This assay was adapted from the SNPlex™ (Life Technologies, USA) assay and was developed to genotype 49 of the 52 autosomal SNPs in Sanchez *et al.* (2006)'s forensic identification panel (Phillips *et al.*, 2007a). GenPlex™ (Life Technologies, USA) has been validated as a technique that provides comparable and possibly enhanced SNP genotyping than the SNaPshot™ assay (Life Technologies, USA) (Phillips *et al.*, 2007a, Tomas *et al.*, 2008). Several ancestry prediction multiplex assays have been developed using SNPlex™ (Life Technologies, USA) technology (Lee *et al.*, 2010, Nievergelt *et al.*, 2013) and other oligo-ligation techniques such as Taqman® (Life Technologies, USA) technology (Yang *et al.*, 2005, Kosoy *et al.*, 2009, Nassir *et al.*, 2009).

1.5.2.3 *Assignment methods*

Assays for inference of BGA or EVCs not only require an appropriate genotyping technique, but also mathematical models to compute statistics from the genotype data to provide probative values for the correlations observed between the genetic data and the ancestry/phenotypic information. There are two major methods used for the assignment of individuals to specifically defined populations (these can be defined for example as continental populations or groups of individuals which share similar physical characteristics) using multilocus genotypes generated from the analysis of multiple polymorphic (SNP) loci. Both involve classifications using maximum likelihood calculations, which are methods that find or approximate the probability that makes the observed data most likely (Manel *et al.*, 2005). The first is a frequentist assignment test that assumes random mating in the tested populations and involves the calculation of a log-likelihood value for the expected frequency of a particular genotype which is derived by multiplying the allele frequencies across all loci. This is calculated for each observed genotype, for each population, using the allele frequencies within the respective populations. The individual is then assigned to the population with the highest log likelihood (Paetkau *et al.*, 1995, Paetkau *et al.*, 2004). This model is available with several software for population genetic analyses such as GenAlEx 6.5 (Peakall and Smouse, 2012) and GeneClass2 (Piry *et al.*, 2004).

The alternative model utilises a Bayesian approach to calculate log likelihood values to assign multilocus genotypes to specified populations (Rannala and Mountain, 1997). The probability of a multilocus genotype occurring within a population is calculated using the allele frequencies in that population. Monte Carlo simulations are performed to generate a large number (up to 10,000) of random multilocus genotypes for each of the other candidate populations, using allele frequencies estimated from reference population samples to create a likelihood distribution (Rannala and Mountain, 1997). The likelihood of the observed genotype is compared to that of the simulated genotype likelihood distribution for each candidate population. A likelihood threshold is specified and any population for which the probability falls below this value is rejected (Rannala and Mountain, 1997). The population for which the likelihood of the observed genotype is highest is assigned as the population of origin. A version of this assignment test is also available in GeneClass2 which also incorporates a ‘leave one out’ option to exclude

the tested individual from its population sample in order to avoid bias when estimating allele frequencies (Piry *et al.*, 2004).

A revised version of this model which does not require predefined reference populations was proposed (Pritchard *et al.*, 2000) and a clustering algorithm was developed which has been incorporated in the software STRUCTURE (Falush *et al.*, 2003, 2007, Hubisz *et al.*, 2009). This software computes the posterior probability that an observed multilocus genotype originates from one of the candidate populations and compares this probability to a specified threshold. Any population for which the probability is less than the threshold is rejected as the population of origin for the observed multilocus genotype (Pritchard *et al.*, 2000). Thus it is possible to investigate population stratification, infer the presence of distinct populations identified as clusters based on similarities in multilocus genotype data, assign individuals to populations, identify migrants and admixed individuals and estimate population allele frequencies, particularly for populations which have increased incidence of migrants or admixture (Falush *et al.*, 2003, 2007, Hubisz *et al.*, 2009). It is also possible to specify the number of populations (K) with which to analyse the data sets and as such this allows for the identification and correction of population stratification caused by admixture (Falush *et al.*, 2003, 2007, Hubisz *et al.*, 2009).

Bayesian based assignments have been suggested to be less accurate than frequency based models (Cegelski *et al.*, 2003), however STRUCTURE has been the software of choice for the identification of AIMs and for the development of BGA prediction assays using autosomal markers (Rosenberg *et al.*, 2003, Shriver *et al.*, 2003, Ray *et al.*, 2005, Phillips *et al.*, 2007b, Halder *et al.*, 2008, Bouakaze *et al.*, 2009, Kersbergen *et al.*, 2009, Kosoy *et al.*, 2009, Londin *et al.*, 2010, Paschou *et al.*, 2010, Nievergelt *et al.*, 2013). These modern assays have demonstrated improved classification accuracies compared to the earlier methods used (Shriver *et al.*, 1997, Frudakis *et al.*, 2003b), which strongly suggest that the ability to measure admixture within populations, as is possible with STRUCTURE, can improve ancestry predictions.

More recently, a Bayesian classification system was applied for the inference of BGA from profiles obtained using a 34-Plex SNP assay and was incorporated in an open access web portal (http://mathgene.usc.es/snipper/) which can be used to analyse SNP

genotypes uploaded by the user (Phillips *et al.*, 2007b). Log likelihood values are calculated for the individual genotypes for each population group, using allele frequencies estimated from training set samples. These training sets encompass groups of samples representative of each ancestral population. All training set samples could be assigned to the correct ancestral origin, including when only 3 fixed difference SNPs were used, which are SNPs where one allele is found exclusively in one population and the other allele in the other populations. Samples of the CEPH-HGDP panel were tested to ensure that the training sets selected were representative of the populations studied and results showed a very low misclassification rate of 1%.

The study also showed the effect of partial profiles on classification probabilities. It was demonstrated that using less than 34 SNPs had minimal effects on the likelihood ratios and that half the SNPs in the assay were not essential to achieve correct ancestry assignment, reinforcing the concept that informativeness is more crucial than the number of SNPs used and the applicability of the assay for the analysis of compromised DNA samples where partial profiles may result (Phillips *et al.*, 2007b). The assay was successfully used to infer the European or North African ancestry of 4 out of 7 samples extracted from forensic exhibits recovered from the 11th March Madrid Commuter train bombings (Phillips *et al.*, 2009). Though the 34-plex SNP assay was initially designed to infer African, Asian and European ancestry, it was still possible to assign individuals a European or North African ancestry with the use of appropriate training sets. The geographical proximity of these populations increases the likelihood of admixture and was the likely cause for the failure to assign 3 samples with low likelihoods. It was further demonstrated that using smaller scale AIM panels which focus on more specific differentiations can improve the classification probabilities, thus demonstrating that multiplex BGA assays must be tailored to the populations to be investigated (Phillips *et al.*, 2009).

Most BGA prediction assays incorporating haploid markers have mainly involved the development of classification systems using the haplogroups as a geographic assignment method (Brion, 2005, Wetton *et al.*, 2005, Onofri *et al.*, 2006, Nelson *et al.*, 2007, van Oven *et al.*, 2011a, van Oven *et al.*, 2011b, Ballantyne *et al.*, 2012). Due to the uniparental and non-recombining mode of inheritance of mtDNA and Y chromosome markers, these markers are inherited as a block (haplotype) and remain

linked. As such, frequentist and Bayesian assignment tests are not applicable as these assume independence and linkage equilibrium between loci. However, there is now sufficient population data available which has identified haplogroups diagnostic of major global continental and regional areas and that can be directly used to determine an individual's a geographic origin.

Contrastingly, multinomial logistic regression has been the classification method of choice for the inference of EVCs using autosomal markers (Liu *et al.*, 2009, Branicki *et al.*, 2011, Walsh *et al.*, 2011, Walsh *et al.*, 2013). This statistical model estimates the probability of membership to more than two categories using maximum likelihood calculations. The initial prediction model based on multinomial logistic regression was developed by Liu *et al.* (2009) for the inference of eye colour. The model was modified slightly for the prediction of hair colour using 12 SNPs in pigmentation genes (Branicki *et al.*, 2011). The accuracy of prediction was evaluated through area under the receiver characteristic operating curve (AUC) measurements, which showed high values for differentiating between blonde, brown, red and black hair colour (Branicki *et al.*, 2011). Liu *et al* (2009)'s model was also later applied for the prediction of blue/brown eye colour using 6 pigmentation SNPs (Walsh *et al.*, 2011) and more recently 18 SNPs were added and incorporated in an updated HIrisplex assay for the inference of both eye and hair colour (Walsh *et al.*, 2013). A macro to calculate the probability of an individual having blue, brown or otherwise coloured eyes, as well as blonde, brown, black or red hair based on the alleles observed at each of the 24 pigmentation markers was generated (Walsh *et al.*, 2013). The macro incorporated all the necessary model parameters derived from Liu *et al.*'s (2009) model building data set of 3804 Dutch individuals.

Recently, the Snipper Bayesian classifier which was initially developed for ancestry inferences from multi-locus genotypes (Phillips *et al.*, 2007b) was modified for the prediction of eye, skin and hair colour using 37 pigmentation-associated SNPs (Ruiz *et al.*, 2013). The Bayesian classifier was observed to be comparable to the Walsh *et al.* classifier through AUC analyses and was demonstrated to be a simple and user friendly alternative to multinomial logistic regression capable of analysing partial profiles. Though there are a variety of classification methods for the prediction of BGA and pigmentation phenotype, an appropriate method is generally selected to suit the type of

data in order to identify the most informative markers and maximise prediction accuracy.

1.6 Project background

The development of new genotyping technologies such as microarrays has consolidated our knowledge of genetics and human origins and resulted in the identification of numerous DNA polymorphisms with widespread applications in medical, population and forensic research [for reviews see (Cavalli-Sforza and Marcus, 2003, Underhill and Kivisild, 2007, Kayser and de Knijff, 2011, Schaaf *et al.*, 2011)]. In particular, these advancements have strongly influenced the field of forensic science which has evolved dramatically over the past 15 years. Whilst STR typing is now widely used in forensic casework for identity testing and kinship analyses, it is primarily reliant on the availability of reference samples which are required to establish a link with DNA profiles extracted from forensic evidence samples (Jobling and Gill, 2004). Though STR markers are highly polymorphic and thus more suitable to applications in forensic identity testing (Moretti *et al.*, 2001a), the utility of SNPs has been widespread, from their applications as forensic identity markers to their use for intelligence gathering purposes [reviewed in Budowle and van Daal (2008)]. The low mutation rate and varied locations of SNPs in the genome across both non-coding and coding regions have contributed to their utility as candidate markers in various ancestry (Onofri *et al.*, 2006, Phillips *et al.*, 2007b, Halder *et al.*, 2008, Daniel *et al.*, 2009, Kersbergen *et al.*, 2009, Paschou *et al.*, 2010, Paneto *et al.*, 2011, van Oven *et al.*, 2011a, van Oven *et al.*, 2011b, Nievergelt *et al.*, 2013) and EVC prediction assays (Liu *et al.*, 2009, Valenzuela *et al.*, 2010, Branicki *et al.*, 2011, Walsh *et al.*, 2011, Ruiz *et al.*, 2013, Walsh *et al.*, 2013). These assays have demonstrated that SNPs located in the autosomes, mtDNA and NRY are each informative markers of BGA and/or EVCs.

Nonetheless, most of the currently developed diagnostic tools for DNA intelligence have predominantly been developed to analyse different DNA intelligence elements individually such as bi-parental BGA or eye colour only and have targeted only SNPs of the same class [for example Kersbergen *et al.*(2009), Nievergelt *et al.* (2013) and Walsh *et al.* (2011)]. Numerous autosomal multiplex SNP assays for the inference of ancestry

have investigated mtDNA or NRY typing in addition to the autosomal markers. In their analysis of forensic material recovered from the 11th March Madrid commuter train bombings using a 34-plex assay for the inference of ancestry, Philips *et al.* (2009) also performed additional NRY SNP and mtDNA typing to confirm ancestry predictions made using the autosomal 34-plex panel. In an earlier study, mtDNA typing had already successfully demonstrated the ability to identify and broadly infer the BGA from the degraded human remains of 9/11 terrorists (Edson *et al.*, 2004). Other studies have investigated the ancestry of self-declared US Americans using AIMs from all three genetic classes to detect the paternal, maternal and bi-parental ancestry of four major US populations (Lao *et al.*, 2010) and of Argentinians (Corach *et al.*, 2010) These studies have shown that the autosomes, mtDNA and NRY region may each reveal unique aspects of an individual's genetic history and that these markers are more informative in combination by allowing us to study different levels of admixture to provide a comprehensive overview of an individual's ancestry.

Many AIMs are located in pigmentation genes and have been shown to also be strongly associated with certain physical characteristics such as rs1426654 and rs16891982 (Nakayama *et al.*, 2002, Soejima and Koda, 2007), thus adding further valuable intelligence information that may be collected from the genotypes. The Philips *et al.* (2009), Corach *et al.* (2010) and Lao *et al.* (2010) studies combined 3 different genetic classes, however these required 3 separate large scale assays to be performed including sequencing of the entire mtDNA control region. Nevertheless, the broad range of SNP genotyping strategies currently available strongly favour the development of novel diagnostic DNA intelligence tools combining informative autosomal, NRY and mtDNA markers for comprehensive predictions of both BGA and EVCs and which is yet to be attempted.

The possibility of combining several types of SNPs in one assay was demonstrated by Divne *et al.* (2005) who developed a microarray for forensic identification of degraded DNA samples which successfully combined 12 nuclear and 21 HV1 mtDNA markers (Divne and Allen, 2005). In 2009, a novel multiplex PCR for the analysis of highly degraded samples was developed as a preliminary test to gather intelligence such as sex determination and to assess the applicability of a sample for further DNA typing using STR or mtDNA analysis (Von Wurmb- Schwark *et al.*, 2009). The assay was validated

to have a detection threshold of 25pg and successfully analysed two autosomal STRs, two Y-STRs, the gender determining amelogenin locus, two different sized fragments of the mtDNA HV1 region as well as a 344bp fragment of *Hydra vulgaris* actin gene which is included as a quality sensor for PCR efficiency and the presence of inhibitors in a sample (Von Wurmb- Schwark *et al.*, 2009). This study provided a solid framework from which more substantial multiplexes combining a larger collection of BGA and/or EVC informative autosomal, mtDNA and NRY markers could be developed and provides an insight into novel DNA typing strategies that may significantly improve the analysis of forensic DNA evidence.

1.7 Objectives

Throughout this introductory chapter, the utility of genetic polymorphisms as markers of BGA and pigmentation phenotype was established and the potential utility of combining different types of markers in forensic DNA typing to obtain a more comprehensive overview of the genetic history of an individual was demonstrated. The current doctoral project aims to select autosomal, mtDNA and NRY chromosome SNPs which exhibit strong associations to specific population groups and/or EVCs and to incorporate these candidate markers in a novel SNaPshot™ (Life Technologies, USA) multiplex assay for inferences of maternal, paternal and bi-parental BGA and EVCs of unknown individuals present at a crime or disaster scene. This particular SNP genotyping strategy was selected for its cost, multiplexing capability, sensitivity and equipment requirements, which are desirable for implementation of the assay in forensic casework (Budowle, 2004). A large panel of polymorphisms was selected for the prediction of both continental and more regional inferences of population groups comprising the broad Australian population, namely Sub-Saharan Africa, North Africa, the Middle East, Europe, South and East Asia.

An initial performance evaluation of the developed multiplex SNP assay was performed on 146 DNA samples to investigate the degree of correlation between the genetic data and the inferred BGA/EVC against the self-declared ancestry and phenotype of the DNA donors, in order to obtain a preliminary assessment of the predictive accuracy and reliability of the system. A sufficiently predictive multiplex SNP assay should be able to

clearly distinguish between different population groups and pigmentation phenotypes. However, since no polymorphism is wholly specific to a particular population or physical trait, various statistical tests including population genetic analyses and classification methods were applied on the multilocus haplotypes and genotypes to investigate the degree of population differentiation achieved with the candidate SNPs to achieve meaningful predictions.

Ultimately, the goal was to develop a rapid and robust assay that can provide comprehensive DNA intelligence regarding the ancestry and potential physical appearance of an unknown individual. A preliminary validation of the SNP genotyping system for applications in casework was also carried out to test assay robustness and sensitivity on different types of common forensic biological specimens. Furthermore, we anticipate that the observed distribution of the candidate SNPs across the different population groups will complement global population data concerning the specific ancestry and phenotypic associations of these markers. This study is the first of its kind and offers preliminary insights into the performance of a novel consolidated DNA intelligence diagnostic tool for the inference of BGA and EVCs from autosomal, mtDNA and NRY SNPs.

CHAPTER 2

MATERIALS AND METHODS

Chapter 2. Materials and Methods

2.1 Materials

All materials are listed in alphabetical order with the name of the supplier provided for each item.

2.1.1 Reagents, kits and consumables

Boric Acid- MP Biomedicals LLC, USA
Dithiothreitol (DTT)- Thermo Scientific, USA
DNA Grade Agarose – Progen Industries Ltd, Australia
50bp DNA ladder- New England Biolabs, USA
2mM dNTP mix- Fisher Biotec, Australia
10mM dNTP mix- Fisher Biotec, Australia
EDTA disodium salt, Analar grade- Sigma-Aldrich Co.LLC, USA
FTA® Classic card or Mini Card- Whatman International Ltd, UK
Hi-Di™ Formamide- Life Technologies, USA
Harris Micro-Punch™ 2.0mm (with cutting mat)- Whatman International Ltd, UK
GelRed™ Nucleic Acid Gel stain (10000× in water)- Biotium Inc., USA
Genescan-120 LIZ® size standard- Life Technologies, USA
310 Genetic Analyzer capillary, 47cm- Life Technologies, USA
1.5mL graduated reaction tube- Greiner Bio-One, Germany
6× Loading Buffer, Blue- New England Biolabs, USA
25mM Magnesium chloride ($MgCl_2$)- Fisher Biotec, Australia
MicroAmp® 8-tube strip (0.2mL)- Life Technologies, USA
10×PCR reaction buffer- Fisher Biotec, Australia
0.2mL PCR tubes- Astral Scientific, Australia
0.5mL PCR tubes- Astral Scientific, Australia
POP-4™ polymer for the 310 Genetic Analyzer, 5mL- Life Technologies, USA
POP-6™ polymer for the 310 Genetic Analyzer, 3mL- Life Technologies, USA
PureLink™ Genomic DNA Mini kit- Life Technologies, USA
Qubit™ dsDNA HS Assay- Life Technologies, USA
310 10× Running buffer (with EDTA), 25mL- Life Technologies, USA

SNaPshot® Primer Focus kit- Life Technologies, USA
SNaPshot® matrix standard set DS-02- Life Technologies, USA
ABI PRISM® SNaPshot™ Multiplex kit 100 reactions- Life Technologies, USA
Sterile foam tipped applicator- Whatman International Ltd, UK
Sterile soft rayon swab –Interpath Services, Australia
Tris, ultra pure grade- Amresco LLC, USA

2.1.2 Enzymes

Calf-intestinal alkaline phosphatase (CIP) 1000U (10000U/mL)- Thermo Scientific, USA
ExoSAP-It 100 reactions (200µL)- USB Corporation, USA
Illustra ExoStar 1 Step 100 reactions (200 µL)- GE Healthcare, USA
Taq Polymerase 500U (5.5$^{\varphi}$U/mL)- Fisher Biotec, Australia

2.1.3 Solutions

All solutions were prepared using Analar reagents and deionised water (dH_20) and were sterilised by autoclaving:

5× TBE electrophoresis buffer; 0.45M Tris-Borate, 0.01M EDTA, pH 8.0
TE buffer pH 8.0; 10mM Tris-HCl (pH7.5), 1mM EDTA (pH 8.0)
TE 0.1 buffer pH 8.0; 10mM Tris-HCl (pH 7.5), 0.1mM EDTA (pH 8.0)
Tris-HCl buffer (pH 7.5); 1M Tris, pH 7.5

2.1.4 Synthetic Oligonucleotides

All oligonucleotides were synthesised by Invitrogen (Life Technologies, USA) with polymerase chain reaction (PCR) and single-base extension (SBE) primer sequences for each multiplex assay shown in Table 2.1 and Table 2.2 respectively. The PCR primers were supplied dry and desalted and the SBE primers were cartridge purified. All primers were hydrated with sterile TE buffer (pH 8.0) to 100µM based on the quantity (nmoles) stated by the manufacturer on the Certificate of Analysis. The concentration of the final

Table 2.1: Designed (a) mtDNA, (b) NRY and (c) autosomal PCR primers

(a)	MtDNA SNP		PCR Primer Sequences (5'-3') (F: forward, R:Reverse)	T_m (°C)	Amplicon size (bp)
MULTIPLEX 1	C10400T	F	CCACAACTCAACGGCTACATAGAA	59	426
		R	GGAGGATATGAGGTGTGAGCGATA	60	
	T10238C		Same amplicon as C10400T		
	C8281G	F	CATGAGCTGTCCCCACATTAGG	60	268-359
		R	GGTGATGAGGAATAGTGTAAGGAGTATGG	61	
	C10873T	F	TGGCCTAGACTACGTACATAACCTAAA	58	308
		R	CGTGATAGTGGTTCACTGGATAAGT	57	
	C3970T*	F	CCTTCGACCTTGCCGAAG	58	301
		R	GTGAGTGGTAGGAAGTTTTTTCATAGG	59	
	C3970T	F	CTTGCCGAAGGGGAGTCC	60	297
		R	GTGAGTGGTAGGAAGTTTTTTCATAGG	59	
	T12705C	F	TATTCATGTGCCTAGACCAAGAAGT	57	273
		R	AGCCGATGAACAGTTGGAATAG	57	
	A1736G	F	ACAAGTCGTAACATGGTAAGTGTACTGG	60	246
		R	CATCTTTCCCTTGCGGTACTATATCTA	59	
MULTIPLEX 2	T7028C	F	GAGCATATTTCACCTCCGCTACC	60	333
		R	GGATTTTGGCGTAGGTTTGGTCTA	61	
	A12308G	F	GAAAGCTCACAAGAACTGCTAACTCA	59	323
		R	GTTCGAGATAATAACTTCTTGGTCTAGGC	60	
	T4216C	F	GACCCTACTTCTAACCTCCCTGTTCTT	61	297
		R	TGATAGGTGGCACGGAGAATTT	59	
	A15607G	F	CCACATCAAGCCCGAATGA	59	270
		R	ACGGATGCTACTTGTCCAATGA	58	
MULTIPLEX 3	G3010A	F	GATCCAATAACTTGACCAACGG	60	287
		R	TATCATTTACGGGGGAAGGCG	61	
	T980C	F	GCAATGCAGCTCAAAACGCTTAG	62	267
		R	AAAGCCACTTTCGTAGTCTATTTTGTGT	60	
	C12633A	F	CCTGTAGCATTGTTCGTTACATGGT	60	245
		R	ATTGCTTGAATGGCTGCTGTG	60	
	A3348G	F	CAAGAACAGGGTTTGTTAAGATGG	61	204
		R	TTGGGGCCTTTGCGTAGTTG	61	
MULTIPLEX 4	C4140T	F	CTTCATAGCCGAATACACAAACA	59	360
		R	TCCTAGAAATAAGGGGGTTTAAG	60	
	G4491A	F	AGAATCGAACCCATCCCTGAG	61	326
		R	CGGTTGCTTGCGTGAGGAAA	61	
	A4715G	F	CCGCATCCATAATCCTTCTAATA	60	312
		R	CACCTCAACTGCCTGCTATGA	61	
	C4883T		Same amplicon as A4715G		
	T9824C	F	ACCCCGCTAAATCCCCTAGA	60	290
		R	GCAGATAGTGAGGAAAGTTGAG	60	
	C12403T	F	ACAGCTATCCATTGGTCTTAGG	60	265
		R	TGGCTCAGTGTCAGTTCGAGATA	60	

* Original C3970T PCR primer pair excluded from the final Multiplex 1 assay

(b)	NRY SNP		PCR Primer Sequences (5'-3') (F: forward, R:Reverse)	T_m (°C)	Amplicon size (bp)
MULTIPLEX 1	M9	F	CGGCGTTTAAAACAATCGAAA	59	192
		R	CATTGAACGTTTGAACATGTCTAAAT	57	
	M207	F	GGGCAAATGTAAGTCAAGCAAGA	59	163
		R	TTTCTAGGCTGTTCGCTGCTAC	58	
	M89	F	GAGTCTTATCCTATGAGGTGCCATGA	61	148
		R	CCAGGATCACCAGCAAAGGTAG	60	
	M201	F	AGTTTCTCAGATCTAATAATCCAGTATCAAC	57	101
		R	AGGTTCAAATCCCATATCCAGC	58	
	M45	F	AACCTTTTAACAGTAACTCTAGGAGAGAGGA	60	92
		R	GGCCTGGACCTCAGAAGGAG	60	
MULTIPLEX 2	M174	F	CTGACCGTTAATTGGAGAGAACATCT	60	259
		R	GCACCCCTCACTTCTGCACT	59	
	M180	F	TGCAAGACAGTCTTCATCAACCAC	60	249
		R	TCTCTACATGAACAGCACATGCATT	60	
	M130	F	CTCAGGAGACATGTATAAGGAAGTGTTTA	58	230
		R	GCAACAGTAAGTCGAATGCCCT	59	
	M181	F	GGCAGGAAAACATGAAGCCAT	60	208
		R	GCACACTAGCTATAAGCAAAAGAAATCT	58	
	M40	F	TTGTATTTTTAGTAGAGACCCAGTTTCACC	60	192
		R	GATGTGATTAATTGACCTACTGATAAGACTC	58	
	M217	F	AGGCCAGTATCTCCAAAATCCTCT	59	181
		R	AAAGCTGCTGTGGCTTTCATC	58	
	M38	F	AGGAGATCTGTTGGCTACTGTTACCC	61	171
		R	CATCTGTCATAATAAACTTTTTCATACTTCCAC	61	
	M91	F	GCAAAAATCCCCCTACATTGCTATT	61	150
		R	TTCTGGCAGTGCCCTTCCA	61	
	M215	F	CAAACTGTTGGTAAATTTTAGAG	60	140
		R	ACTTGGAACAACTGCGAGCA	60	
MULTIPLEX 3	M343	F	TTCTTGCCTTGCTTAGAATTAGG	59	355
		R	ACCCCCACATATCTCCAGG	60	
	M170	F	ATGGGTTCCAGATTTTGCAT	59	233
		R	TGAGACACAACCCACACTGAA	60	
	M124	F	AGCAAAGTTGAGGTTGCACA	60	218
		R	ATTCATGCTCTCAGGGGAAA	60	
	M172	F	TTTATCCCCCAAACCCATTTTG	61	151
		R	GGATCCATCTTCACTCCATGTTG	59	
	M69	F	CTCTGCAAACTGTACTCCTG	59	136
		R	ATCTCCCCTTAGCTCTCCTG	61	
	SRY 10831.2	F	ATAGCAAAAAATGACACAAGGCACC	61	125
		R	TGCCTTTCCTGGATATTTCATATACA	60	
MULTIPLEX 4	M175	F	TTGAGCAAGAAAAATAGTACCCAA	58	231-226
		R	GCATTTTCAGTTAGCCTTGATTGA	60	
	M70	F	TTCTGTAGGCACCATCTGTGAA	60	208
		R	ACCCAGGAGTACAGTTTGCAG	61	
	M147	F	GGATTGTTTAAGGCCAGGAGG	61	197-198
		R	CAAAATACCTGAATAAGCTGGTG	60	
	P31	F	TTTTGGGGAACAGGTAGGTG	58	179
		R	ACTTGGGAGGCTGATACACG	60	
	M119	F	GTTATGGGTTATTCCAATTCAGCA	60	153
		R	ACCCTAAGGAAGTCACGAAG	58	
	M22	F	TCCTGGTTTCCCTATTTGTTTTAAG	61	141
		R	CCATGTTTCGTGCAGATTAAGG	58	
	P256	F	AGGCCTTTCTGCCCTACACTAGAT	60	99
		R	TGTCATAAAGCAGAAGTTCTGAGGA	58	
	M122	F	TTGCTCTGTGTTAGAAAAGATAG	58	82
		R	CAAGGTAGAAAAGCAATTGAGATA	58	

(c)	Autosomal SNP		PCR Primer Sequences (5'-3') (F: forward, R:Reverse)	T_m (°C)	Amplicon size (bp)
MULTIPLEX 5	*OCA2-3*	F	TAAACATACTTCCAAATAATCCAGG	60	173
	rs4778138	R	ACTTGCTCTCCTTTTGATACCAG	61	
	SLC24A5	F	CATACTCTTTCACTTTATTAGGCA	58	164
	rs1426654	R	AGTAAGCAAGAAGTATAAGGAGCA	60	
	OCA2-4	F	TGGAATTGGATACTGACAATGG	58	150
	rs1545397	R	TGAAATCATGGGGGAGAGAG	58	
	ASIP	F	GTGCCGCTTCTTCCGCAG	61	141
	rs6058017	R	GGAAGCCGCCCTGTTAGG	61	
	DARC	F	CAGTGGGCGTGGGGTAAG	61	121
	rs2814778	R	ACAGAAGGGCTGGGACGG	61	
	MATP	F	TGCACACAACTCCACAGAGT	58	120
	rs16891982	R	CTTAGGAGAGAGAAAGACTTACA	60	
	FUT2	F	CCATGCTGGTCGTTCAGATG	60	116
	rs1800021	R	CATGGCTTGAATCTTCGCTAG	60	
	OCA2-1	F	GACTCAGCCGCGCTAGGT	61	115
	rs7495174	R	GCAGGGAGGGTTTACACAG	60	
	F13B	F	AATGCACTAAGCCTGACCTGAGT	58	110
	rs6003	R	TCCAGTGGTTTTGTACCCTGAA	58	
	IL-6	F	GCCTCAATGACGACCTAAGC	60	105
	rs1800795	R	GGGTGGGGCTGATTGGAAA	60	
	OCA2-2	F	ACTCCTGCATCACCAATGGGA	61	103
	rs4778241	R	AATTGTTGGCTGGTAGTTGCA	58	
	MC1R	F	CGCTACCACAGCATCGTGA	60	97
	rs885479	R	AGTAGGCGATGAAGAGCGTG	61	
	HERC2	F	GGCTCTCTGTGTCTGATCCAAGA	59	85
	rs12913832	R	GCCCCTGATGATGATAGCG	60	

Table 2.2: Designed (a) mtDNA and NRY and (b) autosomal SBE primers

(a)		SNP	St$^{\Phi}$	*5'aspecific tail*-Target specific primer sequence (5'-3')	Size (nt)	Alleles & colour
MULTIPLEX 1	MTDNA	C10400T*	s	*ggtgccacgtcgtgaaagtctgacaa*GACTACAAAAAGGATTAGACTGAAC	52	C/T
		C10400T	a	*taccgaaac*CGTTTTGTTTAAACTATATACCAATTC	36	G/A
		T10238C	a	*(gact)$_3$ga*CAATTTCTAGATCAAATAATAAGAAGGT	43	A/G
		C8281G	s	*gtgccacgtcgtgaaagtctgacaa*ACCCTATAGCACCCCCTCTA	46	C^{a}/G^{d}
		C10873T	s	*(gact)$_3$gagact*GCCTAATTATTAGCATCATCCC	41	C/T
		C3970T*	a	*gtgaaagtctgacaa*TGTGTATTCGGCTATGAAGAATA	39	G/A
		C3970T$^{\#}$	a	*gtgaaagtctgacaa*TGTGGATTCGGCTATGAAGAATA	39	G/A
		T12705C*	a	*acgtcgtgaaagtctgacaa*GTAACTAAGATTAGTATGGTAATTAGGAA	50	A/G
		T12705C	a	GGTAACTAAGATTAGTATGGTAATTAGGAA	31	A/G
		A1736G	s	*gtagccaa*TTAGCCAAACCATTTACCCA	28	A/G
	NRY	M9	s	*atata*CGGCCTAAGATGGTTGAAT	25	C/G
		M207	a	*gact*TCAAAAGGTATTGTTATTCTCTTT	29	T/C
		M89	s	*(gact)$_2$at*CTTCCTAAGGTTATGTACAAAAATCT	37	C/T
		M201	a	*acagtct*CAACACTAAGTACCTATTACGAAAA	33	C/A
		M45	a	TCAGAAGGAGCTTTTTGC	19	C/T
MULTIPLEX 2	MTDNA	T7028C	s	ACGACACGTACTACGTTGTAGC	22	T/C
		A12308G	s	*g*ATTGGTCTTAGGCCCCAA	19	A/G
		T4216C	s	*cacttagta*ACTCACCCTAGCATTACTTATATGA	35	T/C
		A15607G	a	GCAAGGACGCCTCCTAG	18	T/C
	NRY	M174	s	*(gact)$_2$ga*GAATACCTTCTGGAGTGCCC	31	T/C
		M180	a	*gactg*CAGCACATGCATTAGTAAAAGTC	29	A/G
		M130	s	*(gact)$_4$ga*AGGGCAATAAACCTTGGATTTC	41	C/T
		M181	a	*gactga*CTATTCAAGCTACTACCTTAAAGCA	32	A/G
		M40	s	*gactgac*TCACCCTGTGATCCGCT	25	C/T
		M217	a	*(gact)$_3$g*ATGTATTTTTCCTTCTGAAGAGTT	38	T/G
		M38	a	*(gact)$_4$g*ACAGGGGAATGCTTACTGAATAAAA	43	A/C
		M91	s	*(gact)$_3$gacgact*CCCTACATTGCTATTCTGTTTTTTTT	46	T/A
		M215	a	*gac*CATCAGCTGGAACAGTTAGAAAG	27	T/C
MULTIPLEX 3	MTDNA	G3010A	a	*gactg*TAATAGCGGCTGCACCAT	24	C/T
		T980C	a	*gact*CCCTCCCCAATAAAGCTAAAAC	27	T/C
		C12633A	a	CACAGTGAGAATTCTATGAT	21	G/T
		A3348G	s	*gactg*CTACTCCTCATTGTACCCATTCT	29	A/G
	NRY	M343	s	AGTGCCCTCGTGTTCCA	18	C/A
		M170	a	*(gact)$_4$gac*GACACAACCCACACTGAAAAAAA	43	T/G
		M124	s	*(gact)$_4$ga*GTTGCACAAACTCAGTATTATTAAACC	46	G/A
		M172	s	*(gact)$_3$g*CCCAAACCCATTTTGATGCTT	35	T/G
		M69	s	*(gact)$_2$g*AGGCTGTTTACACTCCTGAAA	31	T/C
		SRY10831.2	s	*(gact)$_3$ga*CATAGGTGAACCTTGAAAATGTTA	39	C/T
MULTIPLEX 4	MTDNA	C4140T	s	*(gact)$_4$ga*GAATTCGAACAGCATACCC	38	C/T
		G4491A	a	*gact*CAAAGATGGTAGAGTAGATGA	26	C/T
		A4715G	a	AGTATTGGTTATGGTTCATTG	22	T/C
		C4883T	a	*(gact)$_5$*GAGATTTGGTATATGATTGAGAT	44	G/A
		T9824C	a	TTGAGCCAATAATGACGTG	20	A/G
		C12403T	s	CCCCATCCTTACCACC	17	C/T
	NRY	M175	s	*(gact)$_3$*ACATGCCTTCTCACTTCTC	32	T^{a}/A^{d}
		M70	a	*(gact)$_3$ga*TTCTGTTGTGGTAGTCTTAG	35	T/G
		M147	s	*(gact)$_4$gac*CCCTGTCTCTGAAAGAAAAAAA	42	C^{a}/T^{i}
		P31	a	*(gact)$_4$g*ATCGCAAAAAAACCCCAAAAAA	40	A/G
		M119	s	*(gact)$_6$g*TTCCAATTCAGCATACAGGC	46	T/G
		M22	a	t*(gact)$_3$*CACTGGGGGAAGGAGG	30	T/C
		P256	s	*gac*GCCCTACACTAGATAGAAAG	24	G/A
		M122	s	*(gact)$_2$g*AGATTTTCCCCTGAGAGC	28	A/G

$^{\Phi}$ DNA strand homologous to the SBE primer, where s =sense strand and a= antisense strand

*SBE primers excluded from final multiplex assays

$^{\#}$ redesigned C3970T SBE primer showing inserted mismatch

^{a}= ancestral allele, ^{d}= allele detected when deletion is present, ^{i}= allele detected when insertion is present

(b)	SNP	St$^{\Phi}$	*5'aspecific tail*- Target specific primer sequence (5'-3')	Size (nt)	Alleles & colour
MULTIPLEX 5	*OCA2-3 rs4778138*	a	*(gact)$_5$*CACTGATTTAGCTGTGTTCTG	42	C/T
	SLC24A5 rs1426654	s	*(gact)$_3$ga*TCTCAGGATGTTGCAGGC	33	A/G
	OCA2-4 rs1545397	s	*(gact)$_4$gac*CAACTTTGTGAATATACTAAAATAC	45	A/T
	ASIP rs6058017*	a	*gactgac*TCCCCGAAGCCCTGCC	24	C/T
	ASIP$^{\#}$ rs6058017	s	ACTCCCGTCCGCGAGC	17	A/G
	DARC rs2814778	a	*gac*GCGCCTGTGCTTCCAAG	21	A/G
	MATP rs16891982*	s	*gactgac*GGTTGGATGTTGGGGCTT	26	C/G
	MATP rs16891982	s	*(gact)$_2$ga*GGTTGGATGTTGGGGCTT	29	C/G
	FUT2 rs1800021	s	*(gact)$_4$g*CTTTGTCTTTACGGTTTCCACT	40	A/G
	OCA2-1 rs7495174	a	*gactgac*AGGCAAGTTCCCCTAAAGGT	28	T/C
	F13B rs6003	s	*(gact)$_3$*GTATAAAATTCAAGAGAACATGC	36	G/A
	IL-6 rs1800795	s	*(gact)5gac*CCCCCTAGTTGTGTCTTGC	47	G/C
	OCA2-2 rs4778241*	a	*(gact)$_4$*TGTTGGCTGGTAGTTGCAATT	38	T/G
	OCA2-2 rs4778241	a	*(gact)$_3$gac*TGTTGGCTGGTAGTTGCAATT	37	T/G
	MC1R rs885479	a	*a*ATGGCCGCAACGGCT	17	C/T
	HERC2 rs12913832	s	*gactgac*GCCAGTTTCATTTGAGCATTAA	30	A/G

$^{\Phi}$ DNA strand homologous to the SBE primer, where s =sense strand and a= antisense strand
*SBE primers excluded from final multiplex 5 assay
$^{\#}$ redesigned ASIP SBE primer showing inserted mismatch

stocks was verified using the ND-1000 Spectrophotometer (NanoDrop Technologies, USA) and primer stocks were then diluted as appropriate and aliquoted into multiple 2μM and 20μM working solutions with dH_20 and stored at -20°C. For each multiplex assay, the relevant PCR or SBE primers were combined in a single primer mix to be added to each sample. The concentration of each primer in the multiplex assays was modified during assay optimisation to achieve relatively balanced amplification and allelic signals for all SNPs in the multiplex.

2.1.5 Population samples

DNA samples were from the University of Technology, Sydney (UTS) DNA database which was compiled from a sampling regime conducted at the university in accordance with UTS Human Research Ethics Committee approval (03/76). DNA samples were collected from consenting participants and included staff and students of the science faculty as well as other individuals in the greater Sydney region. Participants rubbed the interior of each cheek for 15 seconds, using each side of a foam tip swab (Interpath, Australia), to collect saliva and buccal cells. Both sides of the swab were then pressed onto the circular sample areas of a numbered FTA® Custom mini card or classic card (Whatman, United Kingdom). The FTA® cards were air dried at room temperature and were then returned to their correspondingly labelled envelopes by each participant, to avoid cross contamination. Over 300 FTA® bound DNA samples are contained in the UTS DNA Database and remain securely stored at room temperature. Each DNA donor completed an informed consent and ancestry and phenotype declaration form providing details of their biogeographical ancestry (BGA) over 3 generations and external visible characteristics (EVC) including hair (blonde, light brown, red, dark brown, black), eye (blue, green, hazel, dark brown) and skin (fair, olive and dark) colour. The classification of BGA groups used on the ancestry declaration form was derived from the Australian standard classification of cultural and ethnic groups (ASCCEG) defined by the Australian Bureau of Statistics (2001).

A total of 146 samples (116 males and 30 females) whose parents and maternal and paternal grandparents originated from the same population were selected from the UTS DNA database for analysis. All 146 samples were divided into six major sub-population

groups according to the declared population of origin (Table 2.3) and were genotyped using the developed 5-multiplex SNP assay (refer to section 2.2.2).

Table 2.3: Distribution of the Australian population samples genotyped using the developed 5-multiplex SNP genotyping system

Sub-population	Country	Number of individuals (N)
Sub-Saharan Africa (N=9)	Cameroon	3
	Uganda	2
	Eritrea	1
	Ghana	1
	Nigeria	1
	West Africa	1
North Africa (N=21)	Egypt	21
Middle East (N=15)	Lebanon	7
	Iran	3
	Iraq	2
	Kurdistan	2
	Afghanistan	1
Europe (N=41)	Australian Caucasian	15
	Greece	6
	England	5
	Italy	5
	Ireland	3
	Croatia	2
	Sweden	2
	Malta	1
	Scotland	1
	Poland	1
South Asia (N=21)	India	12
	Pakistan	4
	Sri Lanka	4
	Bangladesh	1
East Asia (N=39)	China	27
	Philippines	5
	Korea	4
	Taiwan	2
	Vietnam	1
TOTAL		**146**

2.1.6 Software

Arlequin version 3.5.1.3- Excoffier & Lischer, 2010, at
http://cmpg.unibe.ch/software/arlequin35/Arl35Downloads.html
AutoDimer version 1.0- Vallone & Butler, 2004, at
http://www.cstl.nist.gov/div831/strbase/AutoDimerHomepage/DownloadPage.htm
CLUMPP version 1.1.2- Jakobsson & Rosenberg, 2007, at
http://www.stanford.edu/group/rosenberglab/clumpp.html
Distruct version 1.1- Rosenberg, 2004, at
http://www.stanford.edu/group/rosenberglab/distruct.html
FigTree version 1.4- Rambaut, 2012, at http://tree.bio.ed.ac.uk/software/figtree/
GenAlEx version 6.501- Peakall & Smouse, 2006, 2012, at
http://biology.anu.edu.au/GenAlEx/Download.html
GeneScan® Analysis software version 3.7- Life Technologies, USA
GeneScan® Data Collection software version 1.1- Life Technologies, USA
Genetic Distance Analysis (GDA)- Lewis and Zaykin, 2002 at
http://hydrodictyon.eeb.uconn.edu/people/plewis/software.php
IBM® SPSS® Statistics for Windows version 19- IBM Corp., USA
OligoExplorer® 1.5- Genelink, USA, at
http://www.genelink.com/tools/gl-downloads.asp
PrimerBLAST- NCBI, USA at http://www.ncbi.nlm.nih.gov/tools/primer-blast/
Primer Express® version 2.0- Life Technologies, USA
Snipper App Suite 2.0- University of Santiago de Compostela, Spain, hosted at
http://mathgene.usc.es/snipper/
STRUCTURE version 2.3.4- Pritchard *et al.*, 2000; Falush *et al.*, 2003, 2007; Hubisz *et al.*, 2009, at http://pritchardlab.stanford.edu/structure.html
Structure Harvester web version 6.93- Earl & vonHoldt, 2012, at
http://taylor0.biology.ucla.edu/structureHarvester/
TreeView version 1.6.6- Page, 1996, at
http://taxonomy.zoology.gla.ac.uk/rod/treeview.html

2.2 Methods

2.2.1 Multiplex SNP Assay design

Extensive literature and database searches were conducted to select mtDNA, NRY and autosomal SNPs which demonstrate skewed allele frequencies in the Sub-Saharan African, North African, Middle Eastern, European, South Asian and East Asian populations or which have previously been identified as ancestry informative markers (AIM). Several autosomal SNPs showing major allelic associations to pigmentation phenotype were also selected to allow for inference of EVC such as hair, eye and skin colour. The literature and database investigations performed and the SNP selection criteria used are described in detail in chapters 3 and 4.

A total of 21 coding region mtDNA, 28 NRY SNPs and indels and 13 autosomal SNPs were selected. Due to the large number of loci investigated, a hierarchical multiplex design approach using five separate multiplex assays was implemented to facilitate genotyping. All haploid SNPs were grouped within Multiplexes 1 to 4. The development of each multiplex assay involved (1) the development of a primary multiplex PCR containing all desired SNP-containing amplicons and (2) the development of a multiplex SNaPshot® reaction to simultaneously genotype all SNP loci. Multiplex assay development and optimisation is discussed in detail in chapters 3 and 4.

2.2.1.1 Multiplex polymerase chain reaction (PCR) primer design

For successful multiplexing, PCR primers must be designed to enable specific amplification of desired target sequences in the presence of unrelated combinations of primer sets and under the same cycling conditions. In order to facilitate this, all PCR primers were designed to have theoretical melting temperatures (T_m) of approximately 60±3°C (180mM salt concentration) with all primers in the same multiplex within 4°C of each other and a GC content of 35-65%. Primer length was restricted to 18-35 bases and if possible, the last 4 bases on the 3'end were designed to contain the same base composition of mostly G and A bases and a low number of C and T bases, to prevent primer dimer formation between two related or unrelated primers. Regions surrounding

the SNP of interest were screened and long runs of repeats or palindromes were excluded if possible, to ensure low complexity amplicons with a GC content of 30-60% (T_m 73-85°C) and to reduce the formation of template secondary structure.

Amplicon sizes were restricted to 80-430bp, ideally with a minimum 7bp spacing between products to allow for visualisation in agarose gels and with the SNP site located at least 30bp from either amplicon end to allow for sufficient room to design the SBE primers. In general, primers for the NRY and autosomal SNPs were designed to produce the smallest amplicons (<200bp) to allow for the analysis of degraded/trace DNA samples. The mtDNA amplicons were designed to be larger (250-430bp) due to the larger copy number of mtDNA in contrast to the Y-chromosome and autosomes. Whenever possible, SNPs located in close proximity were amplified together on the same amplicon to reduce the number of primer pairs in the multiplex.

Primer design was conducted using Primer Express version 2.0 (Life Technologies, USA) and OligoExplorer 1.5 (Genelink, USA, available at http://www.genelink.com/tools/gl-downloads.asp). The PCR primers were screened for hairpin formation and primer dimer interactions using AutoDimer version 1.0 [(Vallone and Butler, 2004), available at http://www.cstl.nist.gov/div831/strbase/AutoDimerHomepage/DownloadPage.htm]. The PCR primers were also checked for template specificity against other regions of the human genome using PrimerBLAST (NCBI, USA, available at http://www.ncbi.nlm.nih.gov/tools/primer-blast/). Primers with more than 5 consecutive matches at the 3' end to any non-target DNA sequence were excluded if the amplified product of this non-specific template interaction was predicted to be less than 500bp. The PCR primers designed for each multiplex are shown in Table 2.1.

2.2.1.2 Multiplex single base extension (SBE) primer design

The design of SBE primers for use in multiplex SNaPshot® assays requires that all alleles within a mixture be individually distinguishable electrophoretically through a combination of size and allele colour. This was primarily achieved by altering the primer lengths through the addition of non-homologous 5'tails that typically consist of poly-GACT or polymers of A, T, C or G bases. In most cases, the length of the tail was

kept smaller than that of the homologous portion of the primer, however, for those primers where the melting temperature and GC content of the homologous sequence was high (>54°C), the tails were permitted to be larger. The final length of the extended SBE primers was maintained between 17 to 50 nucleotides with SNPs in the same multiplex sharing at least one polymorphic nucleotide (i.e. colour) having a minimum separation of 3 nucleotides.

The homologous portion of the primer was designed from either the sense or the antisense strand and comprised 16-30 nucleotides directly upstream of the SNP loci (5' to 3'). Since the recommended annealing temperature of the SNaPshot® multiplex kit is 50°C, the melting temperature of the primer was maintained between 50-56°C (180mM salt concentration) with all primers within 2-3°C of each other. OligoExplorer 1.5 (Genelink, USA) was used to design the homologous sequences and the tails were manually selected to minimise primer dimer interactions and non-specific template interactions. The SBE primers were also tested for hairpin and primer dimer (or self-extension) formation using AutoDimer version 1.0 (Vallone and Butler, 2004).The SBE primers designed for each multiplex are shown in Table 2.2.

2.2.1.3 SNaPshot® Primer Focus reaction

It is not possible to completely predict the exact size of SBE products since the electrophoretic mobility of extended primers shorter than 35bp is strongly affected by the nucleotide composition and attached fluorophore. Hence the electrophoretic mobility of all SBE primers was checked for potential overlap using the SNaPshot® Primer Focus™ reaction (Life Technologies, USA). This reaction involves the single base extension of a primer with fluorescently labelled dideoxyribonucleotide triphosphates ([F]ddNTPs), however, no template DNA is added and the Primer Focus™ enzyme extends each SNaPshot® primer by attaching all four nucleotides randomly so that 4 coloured peaks are visualised for each primer.

The SBE primers were tested at a final concentration of 0.2µM in a total reaction volume of 10µL as described in Table 2.4 and incubated at 37°C for 60 minutes followed by enzyme deactivation at 70°C for 10 minutes. Primer Focus™ products were stored in the dark at 4°C (overnight) or -20°C (long term) until post-reaction clean-up

by treatment with 1U calf-intestinal alkaline phosphatase (CIP) at 37°C for 60 minutes followed by enzyme deactivation at 72°C for 15 minutes. 1µL of purified sample was then added to 8.5µL of Hi-Di® Formamide (Life Technologies, USA) and 0.5µL of GeneScan®-120 LIZ size standard (Life Technologies, USA), denatured at 95°C for 5 minutes and chilled on ice prior to electrophoresis on the ABI PRISM® 310 Genetic Analyzer (Life Technologies, USA) and data analysis using GeneScan® Analysis software v3.7 (Life Technologies, USA), as detailed in section 2.2.2.7.

Table 2.4: SNaPshot® Primer Focus™ reaction preparation

Kit Reagent	Volume (µL)
SNaPshot® Primer Focus™ ddNTP mix	4.0
SNaPshot® Primer Focus™ 10× reaction mix	1.0
SNaPshot® Primer Focus™ Co-Factor	1.0
SNaPshot® Primer Focus™ Enzyme	0.1
SNaPshot® primer (2µM)	1.0
Sterile dH_20	2.9

2.2.2 Multiplex Assay optimisation and genotyping

2.2.2.1 Isolation and purification of buccal cells

A Harris Micro-Punch™ with 2mm tip (Whatman, United Kingdom) was used to extract 2mm discs from buccal swab-containing FTA® cards. Discs were placed in microfuge tubes, washed twice in 500µL 10mM sodium hydroxide (NaOH) for 5 minutes at room temperature on a rotating wheel, followed by two washes in 500µL TE 0.1 buffer (pH 8.0) for 5 minutes. Disc were then dried at 55°C for 15 minutes and stored at -20°C for up to one month prior to PCR amplification.

2.2.2.2 Polymerase Chain Reaction (PCR)

Fragments of DNA encompassing the selected SNP loci were amplified by PCR in a total reaction volume of 25µL containing a single buccal swab disc sample as template. For each of the five multiplex assays, optimisation of the PCR reaction commenced with the singleplex amplification of each primer pair. For Multiplex 1, the singleplex PCR reactions contained 3mM magnesium chloride ($MgCl_2$), 0.2mM dinucleotide triphosphate (dNTP) mix, 1× reaction buffer, 0.75U *Taq* Polymerase and the relevant PCR primer pair at 0.1µM for mtDNA primers or 0.2µM for NRY and autosomal primers. PCR reactions were carried out using an Eppendorf MasterCycler gradient PCR machine (Eppendorf, Germany) with an initial denaturation at 95°C for 3 minutes followed by 35 cycles of denaturation at 95°C for 1 minute, annealing at 57°C for 1 minute and extension at 72°C for 1 minute, with a final extension at 72°C for 7 minutes. These cycling parameters were used for all five multiplexes throughout the optimisation process. The samples were then held at 4°C and stored at -20°C long term prior to subsequent electrophoresis or SNaPshot® reactions.

Upon confirmation of correct sized products in singleplex amplifications, the relevant PCR primer pairs were then grouped in small multiplex PCR reactions and then combined in the final multiplex reaction for each assay. For example, Multiplex 1 PCR primers were tested in 5-Plex, 6-Plex and 11-plex PCR reactions amplifying 5, 6 and 11 loci respectively. When optimising the Multiplex 1 PCR, the same reagent concentrations as for singleplex PCR reactions were used initially, except that all PCR primer concentrations were 0.2µM. Individual reagent concentrations including $MgCl_2$, dNTP mix, reaction buffer, *Taq* Polymerase and PCR primer concentration were then varied as described in chapters 3 and 4 until relatively equalised product yields were achieved. The final optimised PCR reagent concentrations for Multiplex 1 contained 4mM $MgCl_2$, 0.4mM dNTP mix, 1.75U *Taq* polymerase, 2× PCR buffer and these concentrations were also used for the singleplex and multiplex amplification of multiplex 2 to 5 PCR reactions, excluding the PCR primer concentrations. Final concentrations of PCR primers for each multiplex SNaPshot® assay were achieved by adjustment of both individual PCR and SBE primers as discussed in chapters 3 and 4. PCR negative controls containing no FTA® disc were included with every PCR reaction.

2.2.2.3 Agarose gel electrophoresis

To confirm that the correct sized fragments were amplified, the singleplex and multiplex PCR products were electrophoresed in 4% (w/v) agarose/0.5× TBE gels containing 0.5× GelRed™ (Biotium Inc., USA) nucleic acid stain alongside a 50bp DNA ladder (New England Biolabs, USA), which includes 17 fragments ranging from 50 to 1350bp at 50bp intervals. Samples were prepared for electrophoresis by combining 2µL of PCR product with 2µL 6× gel loading dye, blue (New England Biolabs, USA) and 6µL 0.5× TBE buffer to a final loading volume of 10µL. Samples were electrophoresed in 0.5× TBE buffer at 85V for 120-150 minutes then visualised on a UV transilluminator and photographed using a Canon G4 Digital Camera and Ultra Cam Digital Imaging Hood (SciTECH, Australia).

2.2.2.4 Post-PCR purification of DNA template

Prior to performing the SNaPshot® reaction, the primary PCR product must be purified to avoid the participation of PCR primers and residual dNTPs in the primer extension reaction. *Exonuclease I* (*ExoI*) degrades excess single stranded PCR primers and Shrimp Alkaline Phosphatase (SAP) catalyses the removal of 5'phosphate groups from residual dNTPs. 1µL of the commercially available ExoSAP-It kit (USB Corporation, USA) or the Illustra™ ExoStar kit (GE, Healthcare, USA) which incorporate these enzymes, was added to 1.5µL PCR product and incubated at 37°C for 30 minutes, followed by enzyme deactivation at 80°C for 15 minutes. The purified PCR products were then stored at 4°C (24 hours) or -20°C (long term). It was found that more than 4 freeze/thaw cycles substantially decreased ExoSAP-It and ExoStar reagent activity, evident through an increase in artefact peaks, thus the ExoSAP-It and ExoStar stocks were aliquoted into small 20µL working stocks and stored at -20°C prior to their use.

2.2.2.5 SNaPshot® reaction

SNP genotyping was performed using a commercially available SNaPshot® multiplex kit (Life Technologies, USA) in which unlabelled SBE primers anneal to purified PCR templates one base short of the SNP site and are then extended by the addition of a single [F]ddNTP that is complementary to the base on the template strand at the SNP

site (as depicted in Figure 2.1). Extension is then terminated as the ddNTP lacks a hydroxyl group, preventing further phosphate bond formation. Each [F]ddNTP is labelled with a different fluorophore (Table 2.5).

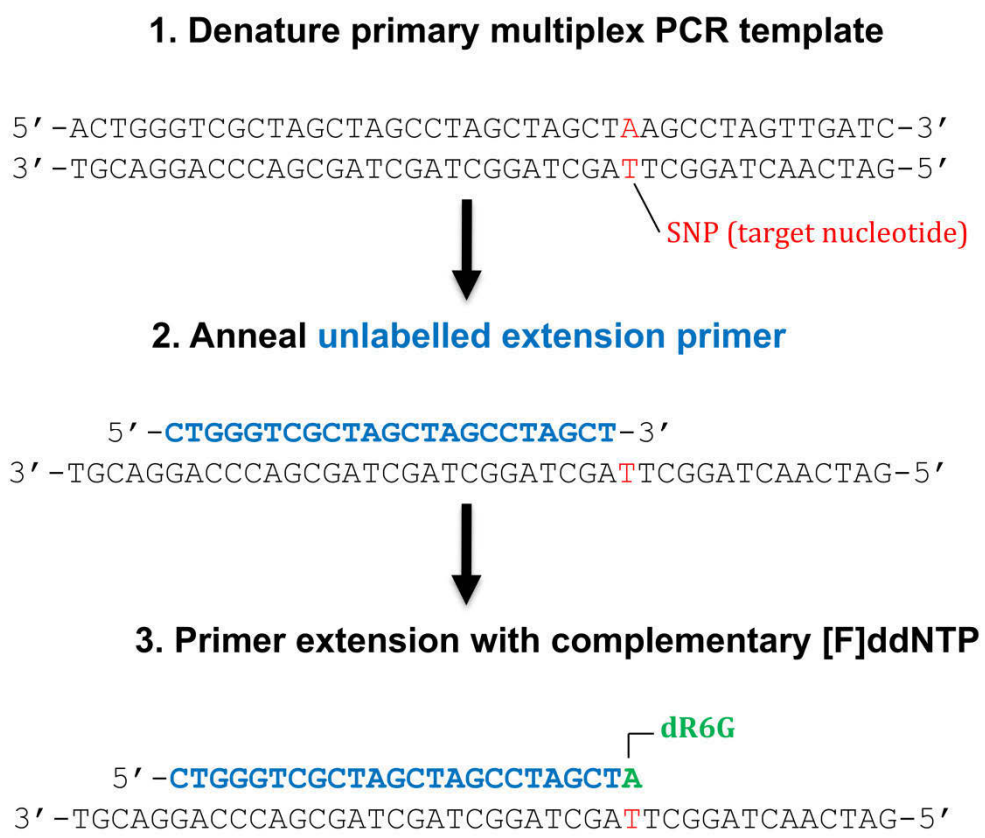


Figure 2.1: **Schematic of the SNaPshot® reaction**

Table 2.5: Dye labels assigned to individual ddNTPs in the SNaPshot® reaction

ddNTP	Dye label (Fluorophore)	Peak colour
A	dR6G	Green
C	dTAMRA™	Black
G	dR110	Blue
T	dROX™	Red

The SNaPshot® reactions were initially performed in a total reaction volume of 5µL containing 2.5µL SNaPshot® Multiplex Ready Reaction Mix, 1µL SBE primer mix, and 1.5µL purified multiplex PCR product. Optimisation of each multiplex assay commenced with the singleplex extension of individual SBE primers at a final concentration of 0.2µM to confirm that the expected alleles are correct. The SBE primers were then tested in small multiplex SNaPshot® reactions and finally combined in a single multiplex SNaPshot® reaction for each assay. For example, Multiplex 1 SBE primers were tested in 5Plex, 6Plex and 11plex SNaPshot® reactions genotyping 5, 6 and 11 loci respectively using the purified Multiplex 1 PCR product as template. All SBE primers of the multiplex were pooled into a primer mix and 1µL was added such that all SBE primers had a final concentration of 0.2µM in a total reaction volume of 5µL.

The concentration of individual SBE primers in the pooled primer mixes were modified during the optimisation phase of each assay to achieve relatively equalised allele peak heights, as described in detail in chapters 3 and 4, with all final optimised SBE primer concentrations summarised in Table 3.4 and Table 4.3. Positive and negative SNaPshot® (no template) controls were routinely performed (Table 2.6). The PCR negative (no template) controls were also used as template in SNaPshot® reactions to monitor contamination at the PCR stage. Thermal cycling of the SNaPshot® reactions was performed on an Eppendorf MasterCycler gradient PCR machine using 25 cycles of 96°C for 10 seconds, 50°C for 5 seconds and 60°C for 30 seconds. Samples were stored at -20°C in complete darkness to prevent continual loss of fluorescence from the light sensitive [F]ddNTPs.

Table 2.6: Preparation of SNaPshot® reaction controls

Kit Reagent	Positive control (µL)	Negative control (µL)
SNaPshot® Multiplex Ready Reaction mix	2.5	2.5
SNaPshot® Control Primer mix	0.5	0.5
SNaPshot® Control Template	1.0	1.0
Sterile dH20	1.0	2.0

2.2.2.6 Post- extension treatment

The SNaPshot® products were purified with CIP to remove the fluorescent label from the 5' phosphoryl groups of unincorporated [F]ddNTPs because these have been shown to co-migrate with the extended primers during capillary electrophoresis (ABI PRISM® SNaPshot™ Multiplex kit Protocol, Life Technologies, USA). The CIP enzyme was diluted with 1×NEBuffer 3 (5mM KCl, 1mM Tris-HCl (pH 8.2), 0.1mM $MgCl_2$. 0.01mM $ZnCl_2$ and 5% glycerol) (Thermo Scientific, USA) to a concentration of 1U/µL and one unit of CIP was added directly to 5µL of SNaPshot® products. Samples were thoroughly mixed and incubated at 37°C for 60 minutes, followed by CIP deactivation at 75°C for 15 minutes. Treated SNaPshot® products were stored in the dark at 4°C (2 days) or -20°C (long term) prior to electrophoresis.

2.2.2.7 Capillary electrophoresis and Genescan® Analysis

Prior to analysing the SBE products with capillary electrophoresis on the ABI PRISM® 310 Genetic Analyzer, spatial and spectral calibrations were performed according to the manufacturer's instructions provided with the SNaPshot® multiplex kit. To prepare samples for electrophoresis, 0.5µL of treated SNaPshot® products was added to 9µL of Hi-Di™ Formamide and 0.5µL of GeneScan -120 LIZ® size standard, then heat denatured at 95°C for 5 minutes. The samples were then transferred to MicroAmp® 8-tube strips, loaded onto an automated ABI PRISM® 310 Genetic Analyzer and electrophoresed in a single 36cm×50µm capillary, filled with POP-6 polymer (Life Technologies, USA). The SBE primers separate according to their size and are detected when the fluorescently labelled extended primer is excited by a laser beam at the detector window of the instrument. A spectrophotometer separates the light emitted by the different fluorophores and a CCD camera processes the data. The run module GS POP-4 (1mL) E5.md5 specified by the manufacturer was used for analysis of SNaPshot® products. The electrophoresis conditions encoded by this module are outlined in Table 2.7. A slight modification to the sample preparation and run module were used for Multiplex 5 where 0.5µL of each of the treated ASIP singleplex and Multiplex 5 12-Plex SNaPshot® products were added to 8.5µL of Hi-Di™ Formamide and 0.5µL of GeneScan -120 LIZ® size standard and the injection time was increased from 5 seconds to 10 seconds.

Although, the manufacturer recommends that the SNaPshot® products are separated using POP-4 polymer, it has been shown that POP-6, which is usually used for sequencing applications, increases resolution and discrimination of SNaPshot® products by increasing sizing accuracy (Vallone *et al.*, 2004, Brion, 2005). For this reason, POP-6 polymer was implemented instead of POP-4 during capillary electrophoresis whilst still maintaining the GS POP-4 (1mL) E5.md5 run module.

The raw data generated by the GeneScan® Data Collection software version 1.1 (Life Technologies, USA) was analysed using GeneScan® Analysis software version 3.7. The analysed data is presented in a profile (electropherogram) as a series of coloured peaks representative of each extended SNaPshot® primer and incorporated [F]ddNTPs. The following default analysis parameters were used: matrix standard set DS-02, filter set E5, minimum peak size of 50 relative fluorescent units (RFU) and Local Southern Method for size calling.

Table 2.7: Run parameters encoded by module GS POP-4 (1mL) E5.md5 used for electrophoresis of SNaPshot® products on the ABI PRISM® 310 Genetic Analyzer

Parameter	Module GS POP-4 (1mL) E5.md5
Capillary array	36cm (distance to detector)
	POP-6 polymer*
Injection time	5 seconds
Electrophoresis voltage	15 kV
Run time	24 minutes
Run voltage	15 kV
Heat plate temperature	60°C

*module GS POP-4 (1mL) E5.md5 is designed to be used with POP-4 polymer however for our applications POP-6 polymer was used instead.

2.2.2.8 Haplogroup and genotype determination

Alleles at each locus were manually called following analysis of the electropherograms. A minimum peak height of 80 RFU was used for allele calling of the haploid mtDNA and NRY SNPs. The final mtDNA and NRY haplogroups were assigned based on the haplotypes derived from the results of Multiplex 1 to 4 using the phylogenies shown in Figure 3.3 to Figure 3.6. For the autosomal alleles in Multiplex 5, a minimum peak height ratio of 4:2:1:1 was applied for the interpretation of heterozygote genotypes to account for the differences in fluorophore emissions between the dR110 (G-blue), dR6G (A-green), TAMRA (C-yellow) and dROX (T-red) dyes and a minimum peak height ratio of 3:1 was used for accepting a heterozygote genotype at any particular locus as recommended by Sanchez *et al.* (2006). One exception was the implementation of a modified 4:1 ratio for heterozygotes at the rs12913832 (HERC2) locus due to the strong imbalance observed between the A and G alleles, which was also encountered by Phillips *et al.* (2012). Thus a minimum peak height of 400RFU (blue), 200RFU (green) and 100 RFU (yellow and red) was applied for allele calling of autosomal SNPs.

2.2.3 Statistical Analysis

Genotyping results for the 146 population samples tested were recorded in Microsoft Excel worksheets as haplotypes for the mtDNA and NRY SNPs (one column for each locus) and as genotypes for the autosomal SNPs where each locus is represented as two columns, one for each of the two possible alleles. Missing data was recorded as '?'. Three separate worksheets were maintained for the mtDNA, NRY and autosomal data. These Excel files were then modified accordingly to suit the requirements of the different statistical programs used.

2.2.3.1 Calculating haplogroup, allele and genotype frequencies

MtDNA and NRY haplogroup frequencies in the Sub-Saharan African, North African, Middle Eastern, European, South Asian and East Asian populations were calculated using the direct counting method. Allele and genotype frequencies were calculated for all 13 candidate autosomal SNPs in each of the six populations using GenAlEx v6.501

(Peakall and Smouse, 2012). The 'Frequency' option for co-dominant data was used and the 'Frequency by pop' parameter was selected.

2.2.3.2 Assessing mtDNA and NRY population diversity

Arlequin v3.5 (Excoffier and Lischer, 2010) was used to calculate various diversity measures and their associated standard deviations from the mtDNA and NRY haplotype data for the Sub-Saharan African, North African, Middle Eastern, European, South Asian and East Asian populations. The 'standard diversity indices' and 'molecular diversity indices' options under the Molecular Diversity setting were selected to calculate the following diversity measures:

1) Number of haplotypes (k)

   This refers to the number of different combinations of alleles at multiple loci identified in each population.

2) Gene diversity (H)

   A diversity index that is equivalent to the expected heterozygosity for diploid data. It calculates the probability that two randomly selected haplotypes are different within a population group according to the statistic proposed by Nei (1987) (page 180).

3) Mean number of pairwise differences (P)

   A molecular index which refers to the mean number of differences between all pairs of haplotypes in a population group, calculated according to the methods of Tajima (1983, 1993) which assume no recombination between sites and selective neutrality.

4) Nucleotide diversity (π)

   This statistic measures the probability that two randomly selected nucleotide sites are different and is equivalent to the gene diversity at the nucleotide level. The calculation of nucleotide diversity in Arlequin is derived from the methods of Nei (1987) (page 257) and Tajima (1983).

2.2.3.3 Assessing population diversity using autosomal SNPs

The observed and expected heterozygosity were calculated for all the genotypes at each of the thirteen autosomal SNP loci for all six populations using GenAlEx v.6.501 (Peakall and Smouse, 2012). The observed heterozygosity (H_o) is derived from the number of heterozygotes in a given population of sample size, N which is determined by direct counting and the expected heterozygosity (H_e) describes the genetic diversity within the population which is calculated as 1 minus the sum of the squared allele frequencies (Hartl and Clark, 2006). Both H_o and H_e were calculated by selecting the 'Frequency' menu for co-dominant data followed by the 'Het, Fstat and Poly by pop' parameter in GenAlEx.

2.2.3.4 Analysis of Molecular Variance (AMOVA)

An Analysis of Molecular Variance (AMOVA) was performed in GenAlEx v.6.501 (Peakall and Smouse, 2012) on the mtDNA and NRY haplotype data and the autosomal genotype data of the 146 individuals tested using the 5-multiplex SNP assay. GenAlEx calculates the Φ_{PT} statistic for haploid (mtDNA and NRY) data and F_{ST} for diploid (autosomal) data. Both statistics estimate the amount of genetic variation that exists among population groups from the molecular variance of the gene frequencies as defined by Excoffier *et al.* (1992). The Φ_{PT} statistic is simply a standardised version of F_{ST} for haploid data derived according to the methods of Meirmans (2006).

For all analyses, the population samples were grouped into six population groups (Sub-Saharan African, North African, Middle Eastern, European, South Asian and East Asian) according to the declared maternal and paternal ancestry of the donors (see Table 2.3). The 'AMOVA' menu was selected and the 'Haploid' option was specified for the analysis of mtDNA and NRY haplotypes, whilst the 'Codom-Allelic' option was used for the autosomal genotypes. Under the 'Total Data Options' tab the following parameters were also selected for all analyses: 999 permutations and suppress within individual analysis.

2.2.3.5 Genetic Distance and Number of Migrants

GenAlEx v.6.501 (Peakall and Smouse, 2012) was used to calculate population pairwise Φ_{PT} and F_{ST} distances for the haploid (mtDNA and NRY) data and autosomal genotypes respectively. This is also available under the 'AMOVA' option and the same parameters described in section 2.2.3.4 were applied except additional parameters were specified under the 'Pairwise Population Options' tab including 'Output pairwise F_{ST} or Φ_{PT}' (depending on whether haploid or diploid data is analysed) and 999 permutations. The option to 'include Nm matrix' was also selected to calculate the number of migrants (gene flow) between populations according to the method of Wright (1969) (page 312-317).

2.2.3.6 Principal Co-ordinates Analysis (PCoA)

A Principal Co-ordinates Analysis (PCoA) was performed in GenAlEx v.6.501 (Peakall and Smouse, 2012) on the population pairwise Φ_{PT} and F_{ST} distances derived from the haploid (mtDNA and NRY) data and autosomal genotypes as described in section 2.2.3.5. This multivariate technique implemented in GenAlEx is used to identify major axes of variation in a multi-dimensional dataset that can be used to visualise the genetic separation between populations and is based on an algorithm published by Orloci (1978). The 'PCoA' menu was selected and the input data type specified as 'Tri Distance Matrix'. The 'Covariance-Standardised method' was applied to produce a plot of the first two axes (co-ordinate 1 v co-ordinate 2) to visualise the mtDNA, NRY and autosomal genetic distances.

2.2.3.7 Neighbour-joining (NJ) trees

Neighbour-Joining (NJ) trees were produced from the derived population pairwise Φ_{PT} (mtDNA and NRY) and F_{ST} (autosomal) distance matrices (refer to section 2.2.3.5) using the Genetic Data Analysis (GDA) program (Lewis and Zaykin, 2002). The matrices were first converted into the appropriate format and were then imported into the program. The 'Show Phenogram' option was selected from the 'Distance' menu to generate Neighbour-Joining (NJ) trees. The output trees produced by GDA were saved as .TRE files and opened using TreeView version 1.6.6 (Page, 1996) to produce an

output file suitable for the FigTree version 1.4 program (Rambaut, 2012) which was then used to format the trees and save them as .pdf files.

2.2.3.8 STRUCTURE analysis

STRUCTURE v2.3.4 (Pritchard *et al.*, 2000, Falush *et al.*, 2003, 2007, Hubisz *et al.*, 2009) was used to detect population sub-structure within our autosomal genotype dataset of 146 individuals originating from six major sub-population groups (Sub-Saharan African, North African, Middle Eastern, European, South Asian and East Asian populations) in Australia. The program applies clustering algorithms to analyse multi-locus genotype data to identify sub-groups (clusters) with similar allele frequencies under the assumptions of Hardy-Weinberg equilibrium (HWE) and linkage equilibrium (LE) (Pritchard *et al.*, 2000, Falush *et al.*, 2003, 2007). In this way, the uppermost hierarchical level of population structure can be detected and individuals probabilistically assigned to the estimated ancestral populations (i.e. clusters), without the need to specify population of origin or sampling location prior to the analysis (Pritchard *et al.*, 2000, Falush *et al.*, 2003, 2007). Several models are available, however the admixture model which assumes that each individual has inherited ancestry from each of the *K* ancestral populations has been recommended for the inference of population stratification (Falush *et al.*, 2003, Evanno *et al.*, 2005). Recent developments to the STRUCTURE program have seen the implementation of new models in version 2.3.4 that are particularly beneficial for the analysis of small or medium sized datasets and that allow for the proportion of individuals assigned to a specific cluster to vary according to the sampling location or *a priori* population of origin (Hubisz *et al.*, 2009).

Population structure was estimated for our Australian population dataset from the 13 autosomal SNP loci genotypes of 146 individuals typed with Multiplex 5 and using the admixture model and the correlated allele frequency option available in STRUCTURE v2.3.4. Two separate analyses were performed, the first without considering prior population information (Analysis 1) and the second using the *a priori* population of origin (LOCPRIOR option, Analysis 2). Parameters used for both analyses included specifying the 'infer alpha' option to allow the degree of admixture (α) to be inferred from the data, setting the value for lambda (λ) to 1, using a burn-in period of 100,000

iterations and a Markov Chain Monte Carlo (MCMC) length equal to 50,000. For each analysis 10 independent runs for each *K*=1 to *K*=8 were performed.

The STRUCTURE results for all 10 runs at each K were converted into a single .zip archive and uploaded into the STRUCTURE Harvester program (Earl and vonHoldt, 2012) available at http://taylor0.biology.ucla.edu/structureHarvester/. This program implements the Evanno method (Evanno *et al.*, 2005) and was used to estimate the most likely number of population clusters (*K*), which is identified as the highest value of ΔK. Refer to Evanno (2005) for details of the method including the calculation of ΔK. The program also produces the output files for the program CLUMPP V1.1.2 (Jakobsson and Rosenberg, 2007), which aligns the multiple replicate analyses of the same dataset to sort the cluster labels, that can then be used to produce graphical displays of the original STRUCTURE results. The .ind and .pop files for the desired value of *K* to be plotted were extracted from the STRUCTURE Harvester output and were then analysed in CLUMPP V1.1.2 using the 'Greedy' algorithm (Option 2) with 10000 repeats and all other parameters maintained at the default value. The CLUMPP output file was then imported directly into the Distruct V1.1 (Rosenberg, 2004) program to visualise the individual cluster membership proportions for each individual, specify cluster colours and labels and produce resulting bar plots in PostScript (.ps) format.

A STRUCTURE analysis of the 1000 Genomes (Phase I) genotypes that were used as reference samples for the inference of BGA using the Snipper (Phillips *et al.*, 2007b) portal (see section 2.2.3.10) was also performed to assess the differentiation achieved with the candidate autosomal SNPs between the Sub-Saharan African, European and East Asian populations. The multilocus genotypes for the 912 individuals were analysed using the admixture model without considering prior population information and the correlated allele frequency option. All other parameters were the same as those previously described for the analysis of our dataset of 146 genotypes, with the exception of a burn-in period of 50,000 iterations and MCMC length equal to 30,000. A total of 10 independent runs for each *K*=1 to *K*=6 were performed. Bar plots depicting each individual's ancestry proportions were produced using CLUMPP V1.1.2 and Distruct V1.1 as described previously.

2.2.3.9 Investigating departures from independence

Departures from independence (HWE and LE) in the autosomal data were assessed using Fisher's exact test for allelic association (Guo and Thompson, 1992, Zaykin *et al.*, 1995) available in GDA (Lewis and Zaykin, 2002). Independence testing was performed on each of the six major populations by selecting the 'Disequilibrium' menu, followed by the 'Exact test' option as the analysis type. Under the 'Options' tab, the measure parameter was set to 'Fisher', the 'Discard' missing data option was selected and the number of shufflings set at 10000. Depending on whether HWE or LE was tested, the following parameters were then also selected according to the analysis performed:

1) Hardy-Weinberg Equilibrium: Analyse subsets 'up to 1 locus' and shuffle method 'break up all loci'.

2) Linkage Equilibrium: Analyse subsets 'up to 2 loci' and shuffle method 'preserve all loci'.

The *p*-values resulting from Fisher's exact tests are considered significant when their value exceeds a set level of significance (typically 0.05 or 5%). This refers to the proportion of genotypic arrays generated by permutation, whose probability is less than the observed probability in the sample that would lead to the rejection of the null hypothesis that the populations are in HWE or that the pair of loci is in LE (Zaykin *et al.*, 1995). When a large number of tests are performed it is expected that at a significant level of 5%, there will be an increased risk of making a type I error, that is that a result is significant by chance given that the null hypothesis is true (Miller, 1981). To correct for the effect of multiple testing encountered in independence testing of multilocus genotype data, the Bonferroni method (Miller, 1981) was applied whereby the significance level, α (0.05), is divided by n, the number of simultaneous tests performed. This ensures that the probability of making a type I error is no greater than α/n and is a conservative way of reducing the risk of incorrectly rejecting the null hypothesis when it is true. Thus, the significance thresholds of $p<0.0038$ (13 tests) and $p<0.0006$ (78 tests) were applied for HWE and LE testing respectively, in this study.

2.2.3.10 Inference of bi-parental BGA using the Snipper App Suite 2.0

The Snipper App Suite 2.0 (Phillips *et al.*, 2007b), which is hosted on a web portal available at http://mathgene.usc.es/snipper/, is a maximum likelihood estimate method (MLE) method that was tested for the inference of the bi-parental BGA of 146 individuals genotyped using Multiplex 5. The test is derived from a Bayesian classification system (Duda *et al.*, 2001) that allows for unknown multilocus genotypes to be assigned to populations using training sets of reference samples for each ancestral population (Phillips *et al.*, 2007b). For each unknown profile, log-likelihoods are estimated for each reference population using the allele frequencies in the respective training sets under the assumption of HWE and LE (Phillips *et al.*, 2007b). Individuals are assigned to the population with the highest log-likelihood value, in this case the least negative. The assigned population is regarded as the inferred bi-parental BGA of the individual.

Multilocus genotypes for each of the 13 candidate autosomal SNP loci were retrieved for 1000 Genomes (Phase I) samples from the SPSmart v5.1.1 database (available at http://spsmart.cesga.es/). These genotypes for 912 individuals of Sub-Saharan African, European and East Asian ancestry were uploaded into the Snipper App Suite as reference populations to be used as training sets. The genotypes for the 146 individuals tested in this study were entered as 'unknowns' and assigned to one of the three ancestral populations. The 'Classification of multiple profiles using a custom Excel file of populations' option was selected to perform the assignments.

2.2.3.11 Inference of bi-parental BGA using GenAlEx v6.501

The population assignment test available in the GenAlEx v6.501 program (Peakall and Smouse, 2012) was also used for the inference of bi-parental BGA. This frequency-based test (Paetkau *et al.*, 1995, Paetkau *et al.*, 2004) assigns individuals to populations using MLE. These MLE are derived for each individual by calculating the expected genotype frequency at each locus, assuming random mating in the population in question. The frequencies are then multiplied across all loci and log transformed to give a log-likelihood value (i.e. MLE). Log-likelihood values are calculated for each individual, in each population using the allele frequencies for the relevant population so

that individuals are assigned to the population with the highest log-likelihood (Paetkau *et al.*, 2004).

To perform the test, the multilocus genotypes for each individual were organised according to their declared ancestry, which represented six population groups (Sub-Saharan African, North African, Middle Eastern, European, South Asian and East Asian populations). The 'Assignment' menu was selected in GenAlEx, followed by the 'Pop Assign' option from the sub-menu. The 'Assign All Pops' and 'Likelihood Positive' options were chosen in the 'Population Assignment Options' menu. The 'Likelihood Positive' option is used to convert the negative log-likelihoods to positive values. The default 'Leave One Out' option for frequency estimates and a 0.01 value for allele frequencies of zero encountered for any particular alleles were maintained. The 'Leave One Out' option provides a bias correction for population frequency by excluding the individual to be assigned before calculating the adjusted allele frequencies used to estimate the log-likelihoods (Paetkau *et al.*, 2004).

The bi-parental BGA was predicted using population assignment tests conducted on the six population groups. Further tests were performed where individuals were also assigned to five populations comprising a Western Eurasia population in which were grouped Europeans and Middle Easterners, followed by assignments to four populations, with further grouping of the Europeans, Middle Easterners and South Asians in a single Eurasia population. A bi-plot of the assignment indices for the four populations was also produced by selecting the 'Assignment Graph' option under 'Graph Options'.

2.2.3.12 Assessing allelic associations of nine pigmentation SNPs with external visible characteristics

The allelic association of the nine pigmentation SNPs (rs885479, rs16891982, rs1426654, rs1545397, rs6058017, rs7498174, rs4778241, rs4778138, rs12913832) with hair, eye and skin colour was determined through the calculation of Fisher's exact test (2-tailed) and odds ratios (OR) with associated 95% confidence intervals (CI) using IBM® SPSS® Statistics for Windows version 19 (IBM Corp., USA). The associations were evaluated from the nine pigmentation SNP genotypes of the 40 individuals with

declared ancestry from Europe (Table 2.3). Individuals were grouped according to their declared hair, eye and skin colour as shown in Table 2.8. To perform Fisher's exact tests, the genotypes at each locus in each hair, eye or skin colour category were designated according to the presence or absence of the minor allele (MA) which was assumed to be dominant using the value 1 or 0 respectively. For example if the MA was T and the alternate allele was G, the genotypes TT and TG would be designated a value of 1 and the genotype GG, a value of 0. The association of the OCA2 haplotype for SNPs rs7495174, rs4778241 and rs4778138 with hair, eye and skin colour was also tested where individuals heterozygous or homozygous for the TGT haplotype were designated a value of 1 and those homozygous for any other alternative haplotype were given a value of 0. The null hypothesis that there is no relationship between the presence of the minor allele/haplotype and the pigmentation phenotype (e.g. dark hair) in question was rejected when the resultant *p*-value was less than 0.05. The odds of demonstrating a specific phenotype whilst having a particular genotype was assessed from the calculated OR for the positively associated pigmentation SNPs (when $p<0.05$).

Table 2.8: Phenotype categories used in Fisher's Exact tests of allelic association on 40 Europeans.
Individuals were grouped according to their declared hair, eye and skin colour.

External visible characteristic	Declared shade	N*	Test category
Hair	Black	7	Dark hair
	Dark brown	19	Dark hair
	Light brown	9	Light hair
	Blonde	5	Light hair
Eyes	Blue	13	Blue or Non-brown
	Hazel/Light brown	16	Non-blue or Non-brown
	Green	1	Non-blue or Non-brown
	Dark brown	10	Brown or Non-blue
Skin	Fair	26	Fair
	Olive	14	Dark

*N: number of individuals

2.2.3.13 Inference of pigmentation phenotype and external visible characteristics

Population assignment tests were performed using GenAlEx v6.501 program (Peakall and Smouse, 2012) to infer the external visible characteristics of the 146 individuals genotyped using multiplex 5. The same procedure and parameters as those applied for the inference of BGA were used (refer to section 2.2.3.11), however instead of grouping the samples according to the declared population of origin, the individuals were grouped into various phenotype categories according to their declared hair, eye and skin colour as shown in Table 2.9. The phenotype category that the individual was assigned to was regarded as the inferred EVC. Two different population assignments were performed, the first using only the genotypes for SNPs showing significant associations with hair, eye and/or skin colour in the exact tests (refer to section 5.2.6 for allelic association results of individual SNPs) and the second used the genotypes for all nine candidate pigmentation SNPs (rs885479, rs16891982, rs1426654, rs1545397, rs6058017, rs7498174, rs4778241, rs4778138, rs12913832).

Table 2.9 Hair, eye and skin colour categories used in population assignment tests for the inference of the pigmentation phenotype of 146 individuals using nine pigmentation SNPs.
Individuals were grouped according to their declared hair, eye and skin colour.

External visible characteristic	Declared shade	N*	Assignment category
Hair	Black	79	Dark hair
	Dark brown	50	Dark hair
	Light brown	11	Light hair
	Blonde	5	Light hair
	Red	1	Light hair
Eyes	Blue	16	Blue/Green
	Green	3	Blue/Green
	Hazel/Light brown	25	Hazel/Dark brown
	Dark brown	102	Hazel/Dark brown
Skin	Fair	46	Fair
	Olive	73	Olive
	Dark	27	Dark

*N: number of individuals

2.2.4 Sensitivity testing

The sensitivity of the forensic intelligence 5-multiplex SNP genotyping system was tested on genomic DNA extracted from a sample of whole blood from an individual of European descent using the PureLink Genomic DNA Mini kit (Life Technologies, USA) according to the manufacturer's instructions. The genomic DNA (gDNA) concentration was quantitated using the Qubit™ dsDNA HS Assay (Life Technologies, USA) which involves the use of a fluorescent dye that is highly selective for double stranded DNA and measurement of the fluorescent signal to calculate the DNA concentration using the Qubit® 2.0 Fluorometer (Life Technologies, USA). Quantitation was performed according to the manufacturer's instructions and three separate readings of the sample were taken and the average used as the final gDNA concentration, which was determined to be 15 ng/µL. A dilution series of the gDNA extract was prepared to produce 2, 0.6, 0.2, 0.06, 0.02, 0.006 and 0.002ng/µL stocks. A 5µL aliquot of each diluted stock solution was added to seven individual PCR reactions such that the total amount added to each reaction was 10, 3, 1, 0.3, 0.1, 0.03 and 0.01ng. These sensitivity PCR and subsequent SNaPshot® reactions were prepared for each of multiplex 1 to 5 and genotyped, as described in sections 2.2.2.2 to 2.2.2.8. A PCR negative (no template) control as well as positive and negative SNaPshot® reaction controls were also prepared in addition to the sensitivity samples.

2.2.5 Preliminary validation of the forensic intelligence 5-multiplex SNP genotyping system for forensic casework

2.2.5.1 Mock casework DNA samples

A preliminary validation of the applicability of the 5-multiplex SNP assay in forensic casework was performed on the following commonly collected biological samples from four donors of different biogeographical origin:

Donor 1: Individual of declared European ancestry provided a sample of whole blood and liquid semen. A 50µL aliquot of blood was deposited on a fragment of sterile cotton cloth to simulate blood stained clothing and was left to dry at room temperature

overnight prior to DNA extraction. The liquid blood and semen samples used had been stored at 4°C for approximately 12 months before casework validation was performed.

Donor 2: Individual of declared European ancestry provided a sample of liquid semen which was stored at 4°C overnight post collection. The next day, a 50µL aliquot was deposited on a fragment of sterile cotton cloth to simulate semen stained clothing and was left to dry at room temperature. The cloth sample was then stored in a paper envelope at room temperature for approximately 3 weeks prior to DNA extraction.

Donor 3: Individual of declared East Asian ancestry provided a smoked cigarette butt which was stored in a paper envelope at 4°C until DNA extraction was carried out.

Donor 4: Individual of declared East Asian ancestry. The donor consumed a can of soft drink and provided it for testing. The can was stored in a large paper envelope at 4°C until DNA extraction was carried out, at which time a wet swab of the rim of the can was taken to simulate a trace saliva sample.

2.2.5.2 Genomic DNA extraction of casework DNA samples

Genomic DNA was extracted from the casework samples using the PureLink™ Genomic DNA Mini kit (Life Technologies, USA) according to the manufacturer's instructions, except for the following modifications to the lysate preparation protocol depending on the type of biological sample extracted:

Blood or semen stained cloth: A 5mm by 5mm square was removed from the blood stained or semen stained cloth and was cut into smaller pieces using sterile scissors and placed in sterile 1.5mL microcentrifuge tube. Note that the preparation of the bloodstained or semen stained sample lysates is performed separately ensuring that different sterile scissors are used for each sampling to avoid cross-contamination.

Cigarette butt: A 5mm strip of the filter paper was removed, cut into small pieces using sterile scissors and placed in a sterile 1.5mL microcentrifuge tube.

Soft drink can swab: The swab was cut into half using sterile scissors and was placed in a sterile 1.5mL microcentrifuge tube.

Liquid semen: 5µL was placed in a sterile 1.5mL microcentrifuge tube.

To each sample, 180µL of PureLink™ Genomic digestion buffer and 20µL of Proteinase K were added, ensuring that all substrate pieces were submerged. For the semen cloth and liquid semen samples, an additional 20µL of 0.39M dithiothreitol (DTT) was also added. The samples were incubated at 56°C for 1 hour or 90 minutes in the case of the semen cloth and liquid semen samples, with occasional vortexing. After the incubation time, all lysates were centrifuged at 13,000 revolutions per minute (rpm) for 3 minutes to pellet any of the paper/cloth/swab material. The samples were then transferred to new sterile 1.5mL microcentrifuge tubes, to which was added 20µL of RNAse A. The samples were vortexed and left to incubate at 37°C for 5 minutes, followed by the addition of 200µL of PureLink™ Genomic Lysis/Binding buffer. The samples were vortexed to achieve a homogenous solution and incubated for 5 minutes at room temperature. Then, 200µL of absolute ethanol was added to the lysates, which were left to incubate at room temperature for 2 minutes. The 'Purification protocol' was then carried out according to the manufacturer's instructions. The gDNA was eluted through two repeat elution steps in 50µL of PureLink™ Genomic Elution Buffer, resulting in a final DNA extract volume of 100µL.

2.2.5.3 Genomic DNA quantitation

The gDNA extracted from the various mock casework samples was quantitated using the Qubit™ dsDNA HS Assay (Life Technologies, USA) and the fluorescence measured using the Qubit® 2.0 Fluorometer (Life Technologies, USA) according to the manufacturer's instructions. Three separate readings were taken for each sample and the average of the triplicates was calculated as the final gDNA concentration, which was used to determine the volume of DNA extract required to add 3ng of gDNA to the PCR reactions. The gDNA yield of some extracts was too low to add 3ng, in which case the maximum volume of 7µL was added. For gDNA yields and volume of DNA extracts added to the PCR reactions for each mock casework sample refer to Table 2.10.

Table 2.10: Concentration of gDNA extracted from mock casework biological samples determined using the Qubit™ dsDNA HS Assay.

The concentration of the gDNA extracts was used to determine the volume (µL) of extract required to add approximately 3ng to the PCR reaction. The final amount (µg) of gDNA template that was added for the amplification of each casework sample is also shown.

Casework sample	Concentration (ng/µL)	Volume of extract added to PCR (µL)	Amount of gDNA added to PCR (µg)
Blood cloth	2.90	1.00	2.90
Semen cloth	4.69	0.64	3.00
Liquid semen	0.46	6.50	3.00
Can swab	0.15	7.00	1.05
Cigarette butt	0.39	7.00	2.73

2.2.5.4 Genotyping mock casework samples

The mock casework samples were genotyped using only the multiplexes within the hierarchical set (Multiplex 1 to 4) which were required to derive the final mtDNA and NRY haplogroups based on the result of Multiplex 1. All samples were then also genotyped with Multiplex 5. Genotyping including the preparation of PCR and SNaPshot® reactions was performed as described in sections 2.2.2.2 to 2.2.2.8 except for the volume of DNA extract added as template for each sample which was as depicted in Table 2.10. Positive controls of buccal swab samples immobilised on FTA® cards, which were provided by each donor, were also performed to confirm the profiles. The maternal and paternal BGA were inferred from the geographical origin of the derived mtDNA and NRY haplogroup respectively, according to the key specified in Figure 3.1 and Figure 3.2. Bi-parental BGA was inferred as described in section 2.2.3.11, except that the autosomal genotypes of the casework samples were added to the excel worksheet as 'unknowns', after the 'reference' genotypes of the 146 population samples grouped into five groups (Sub-Saharan Africa, North Africa, Western Eurasia, South Asia and East Asia) according to their declared BGA. The casework samples were assigned to one of the five populations by selecting the 'Last Pop Unknown' option in the 'Population Assignment Options' menu. The hair, eye and skin colour of the donors was inferred from the genotypes for the nine autosomal pigmentation SNPs (rs885479, rs16891982, rs1426654, rs1545397, rs6058017, rs7498174, rs4778241, rs4778138, rs12913832) as described in section 2.2.3.13.

CHAPTER 3

MULTIPLEX SNP ASSAY DEVELOPMENT STAGE 1:

MULTIPLEX 1 TO 4 FOR THE INFERENCE OF MATERNAL AND PATERNAL ANCESTRY

Chapter 3. Multiplex SNP Assay Development (Stage 1): Multiplex 1 to 4 for the Inference of Maternal and Paternal Ancestry

3.1 Introduction

Determining the biogeographical ancestry (BGA) and external visible characteristics (EVCs) of an unknown DNA sample from a crime scene is a useful source of intelligence information when, in the absence of a reference sample, conventional short tandem repeat (STR) DNA typing fails to identify the unknown donor. Assays developed for the inference of BGA and EVC usually target single nucleotide polymorphisms (SNPs) since their low mutation rate makes them stable markers. Furthermore, SNPs can be typed on short fragments of DNA, making them more applicable to the analysis of low DNA template samples, which constitute a large proportion of forensic biological evidence (Jobling and Gill, 2004).

Studies of haploid mitochondrial DNA (mtDNA) and non-recombining Y chromosome (NRY) SNPs in population genetics have revealed distinct patterns of SNP motifs termed haplotypes that are related under a single phylogenetic tree [reviewed in Underhill and Kivisild (2007)]. The branches of the tree represent groups of related haplotypes termed haplogroups that demonstrate geographic specificity. Many autosomal SNPs have also been identified to be ancestry informative markers (AIMs) capable of identifying and distinguishing different population groups (Collins-Schramm *et al.*, 2004, Yang *et al.*, 2005, Halder *et al.*, 2008, Kersbergen *et al.*, 2009, Kosoy *et al.*, 2009, Nassir *et al.*, 2009, Kidd *et al.*, 2011, Galanter *et al.*, 2012, Nievergelt *et al.*, 2013). Autosomal SNPs in pigmentation genes that have strong associations with human pigmentation variation have also been found to be useful markers of EVCs (Sulem *et al.*, 2007, Han *et al.*, 2008, Kayser *et al.*, 2008b, Sturm *et al.*, 2008, Branicki *et al.*, 2009). Numerous forensic intelligence SNP assays for the prediction of continental BGA or EVC have been reported, some of which have been validated for use in casework (Phillips *et al.*, 2007b, Kersbergen *et al.*, 2009, Londin *et al.*, 2010, Paneto *et al.*, 2011, van Oven *et al.*, 2011a, van Oven *et al.*, 2011b, Walsh *et al.*, 2011, Pereira *et al.*, 2012, Phillips *et al.*, 2013, Ruiz *et al.*, 2013, Walsh *et al.*, 2013). However these existing BGA and EVC prediction assays have predominantly targeted only SNPs

of the same class (i.e. mitochondrial, NRY or autosomal), although autosomal multiplex SNP assays for the inference of BGA have sometimes been supplemented by mtDNA or NRY typing to support inferences made using the autosomal markers (Phillips *et al.*, 2009, Corach *et al.*, 2010, Lao *et al.*, 2010). An assay, that combines SNPs from each of these three classes in a logical decision tree, is expected to provide enhanced and mutually supportive predictions as to maternal, paternal and bi-parental ancestry, as well as the EVCs of an unknown DNA sample. Development of such a novel forensic intelligence tool focusing on six population groups of interest (Sub-Saharan Africa, North Africa, Middle East, Europe, South Asia, East/South-East Asia) was the primary objective explored in this thesis. This chapter describes the design and optimisation of a hierarchical multiplex assay typing haploid mtDNA and NRY chromosome SNPs. The development of a multiplex assay typing diploid autosomal SNPs follows in chapter 4.

3.2 SNP Assay development- Stage 1

3.2.1 Haploid SNP loci selection

In order for a forensic intelligence assay to achieve high predictive ability it is important to consider the informativeness as well as the number of SNP loci to be targeted. To select population informative candidate SNPs, extensive literature reviews were conducted to identify mtDNA and NRY haplogroups that are characteristic of Sub-Saharan African, North African, Middle Eastern, European, South Asian, East/South East Asian populations. Haplogroup databases were compiled by pooling various published frequencies and from these, several mtDNA and NRY haplogroups that occurred at high frequencies and showed specific associations with our global regions of interest were identified. The global distribution of these candidate mtDNA and NRY haplogroups and their estimated frequencies are illustrated in Table 3.1 and Table 3.2. The SNP assay was not primarily intended to discriminate Oceanian populations, however some of the candidate mtDNA markers were later identified to be recurrent and inadvertently define other haplogroups such as P, which is frequent in Oceania. Thus, frequency estimates for the targeted mtDNA and NRY haplogroups in Oceania were also provided in Table 3.1 and Table 3.2. It is evident from these estimated frequencies that no one haplogroup is 100% specific to a particular population, however

Table 3.1: Global distribution (%) of candidate mtDNA haplogroups.

It must be noted that the frequencies shown are estimates only and were derived by combining reported frequencies from multiple studies that did not always investigate all of the same haplogroups.

MtDNA haplogroup[§]	Sub-Saharan Africa[a]	North Africa/ Middle East[b]	Europe[c]	South Asia[d]	East/South East Asia[e]	Oceania[f]
	N= 2513	N=3517	N=4355	N=2798	N=3912	N=2198
L	82.05		0.92	1.6	0.03	
M*	0.04	1.14	0.30	61.47	13.06	42.36
M1	9.03	0.88	0.11		0.03	
M7				0.07	10.40	0.18
M8		0.11		1.47	5.80	
M9				0.39	3.02	3.87
M10				0.04	1.53	
M80'D		0.31	0.11	1.64	24.28	0.77
N*	0.40	7.25	5.76	0.39	5.01	
N1	1.47	1.71		0.32		
A		0.14	0.09	1.29	6.06	
R*	3.30	11.80	4.68	12.08	2.66	
B4'5		0.20	0.02	0.93	13.52	35.99
B4a1					1.20	3.09
H*	0.24	25.39	47.07	2.61	0.33	
H1		0.14	3.10			
P						13.69
R2'JT*	0.40	7.28	7.65	0.50	0.10	
J1		3.38	2.59	0.29	0.05	
T*	0.64	5.94	7.87	0.93	0.10	
T1		2.81	1.33		0.05	
R9					11.96	
U*	1.31	18.45	17.84	8.94	0.77	
U6	1.11	2.87	0.11			
U7		1.76	0.41	2.14	0.03	

[§] Van Oven *et al*, (2009) nomenclature corresponding to the mtDNA tree (Build 15, updated 30th September 2012) available at http://www.phylotree.org.

* All other descendant lineages of the haplogroup excluding those already shown.

[a] Ethiopia, Eritrea, Mozambique, Khoisan, Bantu, Central, East and West Africa, Somalia (Cinnioglu *et al.*, 2004, Gonzalez *et al.*, 2007, Carina *et al.*, 2009, Mikkelsen *et al.*, 2012)

[b] Nubia, Egypt, Bedouins, Yemeni Jews, Iraq, Iran, Syria, Jordan, Palestine, Druze, Turkey, Kurdistan, Armenia, Azerbaijan, Uzbekistan, Morocco, Libya, Afghanistan, Algeria, Tunisia, Mauritania, Berbers, Persia, Tajikistan (Richards *et al.*, 2000, Gonzalez *et al.*, 2003, Plaza *et al.*, 2003, Cinnioglu *et al.*, 2004, Kivisild *et al.*, 2004, Derenko *et al.*, 2007, Behar *et al.*, 2008)

[c] Switzerland, Austria, France, Italy, Germany, Iceland, Scotland, Spain, England, Wales, Portugal, Sicily (Helgason *et al.*, 2001, Plaza *et al.*, 2003)

[d] India, Bangladesh, Pakistan, Sri Lanka (Kivisild *et al.*, 2004, Behar *et al.*, 2008)

[e] China, Thailand, Korea, Manchuria, Vietnam, Japan, Philippines, Indonesia, Malaysia (Kivisild *et al.*, 2004, Tanaka *et al.*, 2004, Trejaut *et al.*, 2005)

[f] Papua New Guinea, Louisiades, New Britain, New Hanover, New Ireland, Bougainville, Solomons, Santa Cruz, Vanuatu, New Caledonia, Fiji, Australia, Cook, Samoa, Tonga, Futuna, Tuvalu, Trobriand (Merriwether *et al.*, 2005, Kayser *et al.*, 2006, Hudjashov *et al.*, 2007)

Table 3.2: Global distribution (%) of candidate NRY haplogroups.

It must be noted that the frequencies shown are estimates only and were derived by combining reported frequencies from multiple studies that did not always investigate all of the same haplogroups.

NRY haplogroup[§]	Sub-Saharan Africa[a]	North Africa/ Middle East[b]	Europe[c]	South Asia[d]	East/South east Asia[e]	Oceania[f]
	N= 2062	N=2115	N=1961	N=2200	N=2330	N=2055
Y	6.01			0.05	0.39	
BT	4.56	0.24				
B	4.07		0.05			
C*		0.28	0.31	3.86	3.39	6.81
C2					0.56	12.85
C3			0.05	0.55	4.89	
D	1.99				4.98	0.05
E*	16.10	0.47	0.20	0.32		0.05
E1b1a	52.18	2.36	0.46	0.05		
E1b1b	4.95	32.20	8.47	0.09		
F*	0.29	22.70	1.58	11.95	1.29	0.54
G	0.29	4.78	4.49	1.14		
H		0.09	0.92	18.05	0.04	
I	1.65	2.70	20.81	0.05		
J2	0.24	17.92	6.83	8.23	0.39	
K*	0.73	1.65	2.29	3.05	7.60	33.19
K1				1.36		
L1		2.32		9.36		
M			0.05		0.60	32.99
O*		0.05	0.05	0.18	3.61	0.34
O1a					13.00	1.70
O2	0.10		0.05	6.91	28.54	0.19
O3			0.15	2.95	26.87	9.34
P	0.34	0.95	0.31	4.68	0.56	0.10
R*	0.10	0.76	0.15	1.68	1.97	0.97
R1a1	0.53	2.08	14.79	17.55	0.73	
R1b	5.63	5.25	37.68	0.86	0.56	0.88
R2a		0.14		7.09		
T	0.24	3.07	0.25		0.04	

[§] Y Chromosome Consortium (2002) nomenclature corresponding to the NRY tree (v8.64, updated 19th July 2013) available at http://www.isogg.org/tree/index.html.

* All other descendant lineages of the haplogroup excluding those already shown.

[a] African Americans, Khoisan, Bantu, Pygmy, Uganda, Kenya, Tanzania, Benin, Rwanda, Burkina Faso, Ethiopia (Hammer *et al.*, 2001, Bortolini *et al.*, 2003, Luis *et al.*, 2004, Hammer *et al.*, 2006, Badro *et al.*, 2013)

[b] Egypt, Saudi Arabia, Tunisia, Morocco, Oman, Lebanon, Yemen, Palestine (Luis *et al.*, 2004, Zalloua *et al.*, 2008, Badro *et al.*, 2013)

[c] Italy, Balkans, Slovakia, European Americans (Hammer *et al.*, 2006, Badro *et al.*, 2013)

[d] Sri Lanka India, Pakistan (Hammer *et al.*, 2001, Ramana *et al.*, 2001, Cordaux *et al.*, 2004, Sengupta *et al.*, 2006)

[e] Vietnam, China, Mongolia, Siberia, Korea, Japan, Taiwan, Cambodia, Philippines, Malaysia, Bali, Sumatra, Borneo, Moluccas (Hammer *et al.*, 2001, Scheinfeldt *et al.*, 2006, Sengupta *et al.*, 2006, Xue *et al.*, 2006, Kayser *et al.*, 2008a)

[f] Papua New Guinea, Australian Aborigines, Micronesians, Fiji, Vanuatu, Cook islands, Tonga, New Britain, New Ireland, Tokelau, Futuna, Tuvalu, Niue, Trobriand (Hammer *et al.*, 2001, Scheinfeldt *et al.*, 2006, Kayser *et al.*, 2008a)

the global distribution of the most basal haplogroups (paragroups) correlates strongly with broad geographical regions, such as mtDNA paragroup L which predominates in Sub-Saharan Africa, paragroups N, H and U which are prevalent in North Africa, Europe and the Middle East and paragroup M which occurs primarily in Asia and Oceania (Table 3.1). Similarly, NRY paragroups Y, B and E occur at high frequencies in Sub-Saharan Africa, paragroup F in North Africa and the Middle East, paragroup R in Europe and South Asia and paragroup K and O in Asia and Oceania (Table 3.2).

To further discriminate between populations that share geographic borders and which are more closely genetically related, haplogroups and sub-haplogroups were also investigated. MtDNA haplogroups R9, B4'5 and M80'D as well as sub-haplogroups M7, M8, M9 and M10 are found in much greater frequency in East and South-East Asian populations compared to South Asia and were subsequently selected to distinguish these populations. Sub-haplogroup M1 occurs in 10-15% of Africans (Gonzalez *et al.*, 2007, Mikkelsen *et al.*, 2012) and was also chosen to distinguish these populations from the South and East Asian groups, in which most lineages of mtDNA paragroup M predominate (Kivisild *et al.*, 2002, Kong *et al.*, 2006, Chandrasekar *et al.*, 2009). Sub-haplogroups N1, U6 and U7 were selected as these occurred in higher frequencies in North African and Middle Eastern populations compared to Europe where in contrast, haplogroups H*, T* and sub-haplogroup H1 predominate (Table 3.1). Likewise, though NRY paragroup E is found predominantly in Sub-Saharan Africa, it also occurs in a large proportion of North Africans and Middle Eastern individuals (Table 3.2). However, at the sub-haplogroup level, E1b1a predominates in Sub-Saharan Africa (52%) compared to E1b1b, which is found at higher frequencies in North Africa and the Middle East (32%), such that these sub-haplogroups were also incorporated to distinguish these populations.

Descending from NRY paragroup F, which shows a wide distribution throughout Europe, the Middle East and Asia, are haplogroups G, J2 and T, which were also chosen for their high occurrence in North African and Middle Eastern populations, compared to European and Asian populations (Table 3.2). To distinguish individuals originating from Europe, prevalent haplogroups in this population such as I (21%), R1a1 (15%) and R1b (38%) were targeted, whilst haplogroups H and L1 were targeted for their high frequencies in South Asia (Table 3.2). Haplogroups D, O and its subgroups O1a, O2

and O3 as well as sub-haplogroup C3 that occur at high frequencies in East and South East Asian populations were included to identify these groups (Table 3.2). Some NRY haplogroups including M and sub-haplogroup C2 that are prevalent in Oceania were also selected to complement the data provided by the unplanned inclusion of the Oceania specific mtDNA haplogroup P (Table 3.2).

The defining SNPs for the candidate haplogroups and sub-haplogroups that were identified to be prevalent in our populations of interest were selected from the mtDNA and Y-DNA phylogenetic trees available at http://www.phylotree.org (van Oven and Kayser, 2009) and http://www.isogg.org/tree/index.html (International Society of Genetic Genealogy, 2012) respectively. Only SNPs located in the mtDNA coding region were selected due to the higher mutation rate of the control region which results in recurrent mutations that are homoplastic between different population groups (Non *et al.*, 2007). For those mtDNA and NRY haplogroups that had multiple defining SNPs, the SNP that showed the lowest level of homoplasy within the mtDNA/NRY tree at the time was selected. Some recurrent mutations were however also selected if these could be used to identify different population specific haplogroups, as this would minimise the number of loci to be included in the multiplex. An example is mtDNA SNP T10238C, which defines both sub-haplogroups N1 and B4a1. A total of 21 mtDNA SNPs and 28 NRY SNPs were selected for incorporation into the assay. Diagrammatic representations of the mtDNA and NRY phylogenies for the selected SNP loci showing the global geographical distribution of the mtDNA and NRY haplogroups are provided in Figure 3.1 and Figure 3.2 respectively.

As new genetic data becomes increasingly available, the mtDNA and NRY haplogroup phylogeny becomes more refined, resulting in the re-organisation of markers and the progressive identification of recurrent SNPs. For example NRY marker M180 which initially defined haplogroup E1b1a (Karafet *et al.*, 2008) now defines sub-haplogroup E1b1a1a1 according the most current NRY tree (v8.64, updated 19 July 2013) available at http://www.isogg.org/tree/index.html). In addition, several candidate mtDNA SNPs were also recently identified to be recurrent in multiple unintended lineages such as the 8281-8289 deletion which defines the East Asian/Oceanian haplogroups B6 and B4'5 (Kivisild *et al.*, 2006, Kong *et al.*, 2006) that are indistinguishable in this assay, as well as African sub-haplogroup M1a1d (Olivieri *et al.*, 2006). Consequently, the haplogroup

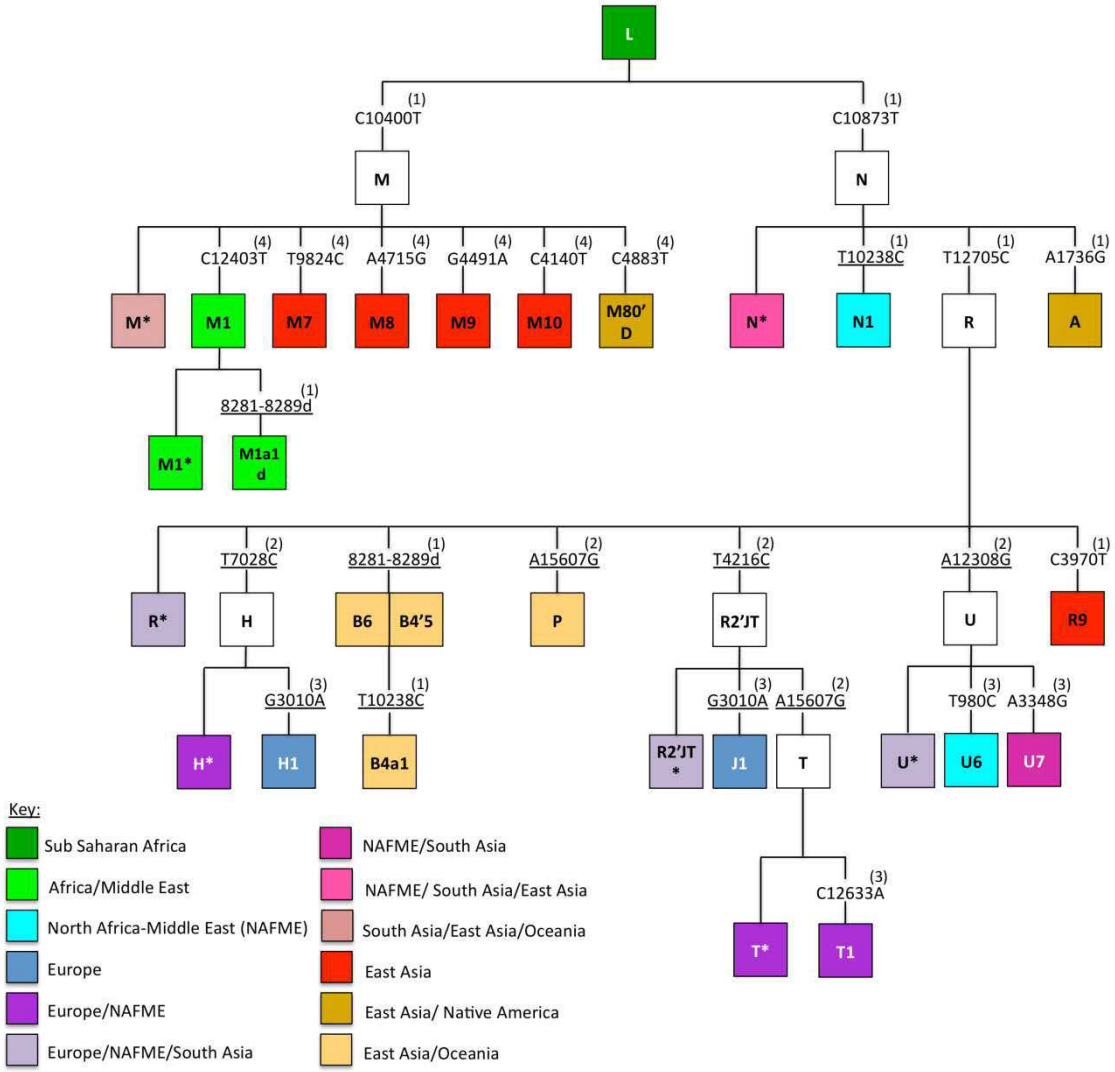


Figure 3.1: Phylogenetic tree of the 21 selected mtDNA SNPs showing the geographical affiliation of the defining haplogroups.

The haplogroup phylogeny follows the van Oven *et al*. (2009) nomenclature corresponding to the most updated mtDNA tree (Build 15, updated 30th September 2012, available at http://www.phylotree.org).The coloured boxes represent the geographical region of origin of the haplogroup as specified in the key.

The candidate SNPs for this research are shown along the branches as nucleotide positions and substitutions observed within the revised Cambridge Reference Sequence (rCRS). The suffix 'd' denotes a deletion and the underlined SNPs indicate recurrent mutations. Haplogroups with an asterisk (*) identify individuals that belong to one of the other descendant lineages of the haplogroup not targeted in the assay.

The bracketed numbers (1, 2, 3 and 4) associated with each SNP represent the multiplex in which the SNP is grouped in this project.

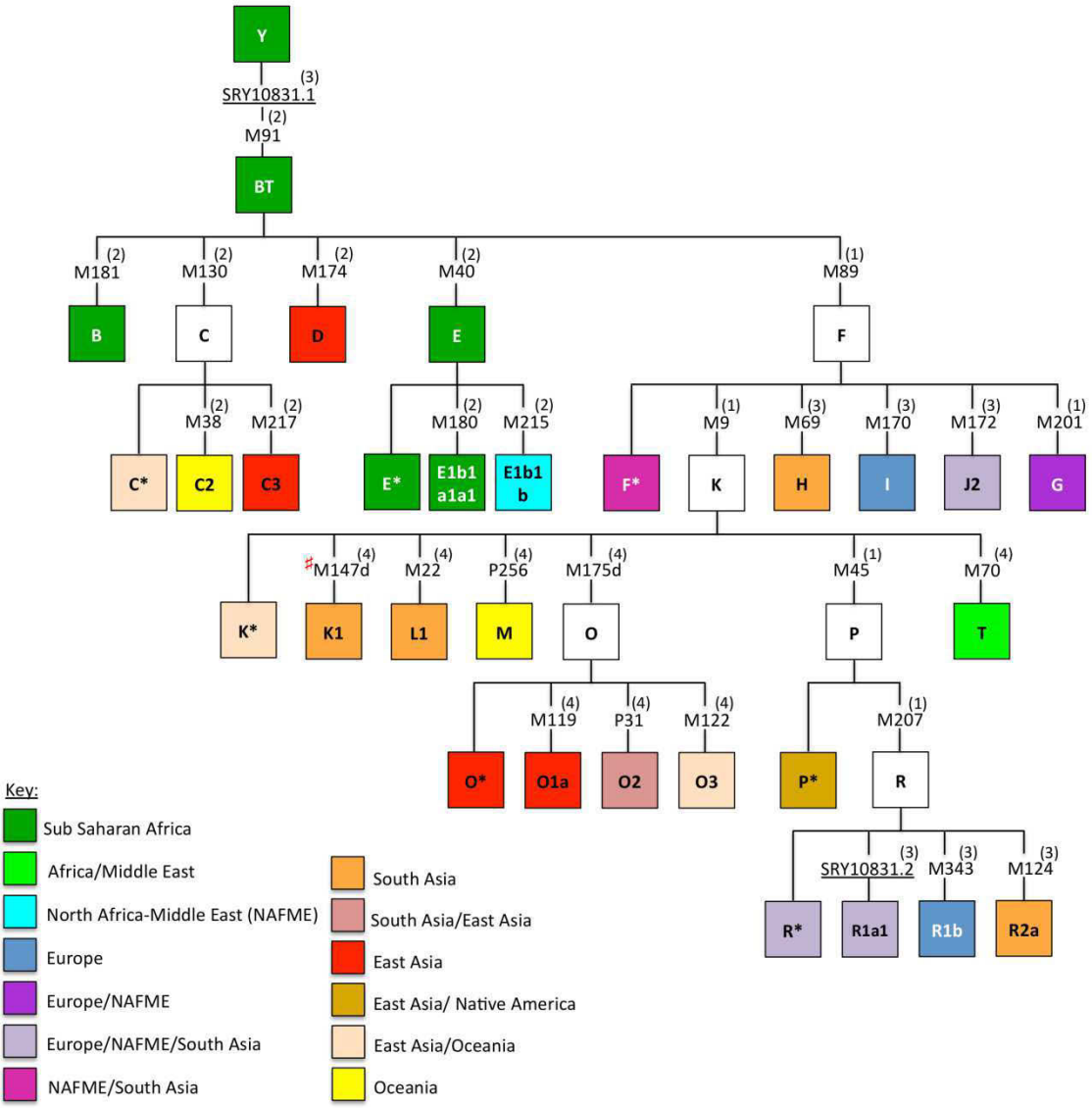


Figure 3.2: Phylogenetic tree of the 28 selected NRY SNPs showing the geographical affiliation of the defining haplogroups.

The haplogroup phylogeny follows the Y Chromosome Consortium (2002) nomenclature corresponding to the most updated NRY tree (v8.64, updated 19th July 2013, available at http://www.isogg.org/tree/index.html). The coloured boxes represent the geographical region of origin of the haplogroup as specified in the key.

The candidate SNPs for this research are shown along the branches. The suffix 'd' denotes a deletion and the underlined SNPs indicate recurrent mutations. Haplogroups with an asterisk (*) identify individuals that belong to one of the other descendant lineages of the haplogroup not targeted in the assay. The # symbol indicates a SNP which was recently identified as a private SNP and has been removed from the 2013 Y-DNA haplogroup tree (v8.64). Investigations revealed that M147d did not meet the population distribution criteria for placement on the tree.

The bracketed numbers (1, 2, 3 and 4) associated with each SNP represent the multiplex in which the SNP is grouped.

nomenclature and phylogeny for the candidate SNPs selected at the start of this project in 2009 have been updated in Figure 3.1 and Figure 3.2 to correspond with the most current mtDNA (Build 15, 30 September 2012) and Y DNA haplogroup trees (v8.64, 19 July 2013).

3.2.2 Hierarchical multiplex assay design

A total of 49 haploid SNPs were selected to be assayed using the SNaPshot® kit (Life Technologies, USA), a simple, sensitive and robust genotyping technique (see Section 2.2.2.5) which utilises equipment already available in most forensic laboratories. Although, single-base extension (SBE) multiplex assays simultaneously typing up to 35 SNPs have been developed (Sanchez *et al.*, 2003), the number of SNPs required for our project (including the diploid autosomal SNPs) is large and would be challenging to type simultaneously using a single multiplex assay.

A widely used strategy for assays targeting haploid SNPs involves the hierarchical arrangement of markers within small multiplexes which parallel the hierarchical structure of the mtDNA and NRY phylogeny (Paracchini *et al.*, 2002, Brion, 2005, Onofri *et al.*, 2006, Wiesbauer *et al.*, 2006, van Oven *et al.*, 2011b, Ballantyne *et al.*, 2012). This hierarchical strategy requires the preliminary determination of the most basal branches of the tree, then proceeding along the branches to the shallowest markers to assign the final haplogroup. Hierarchical multiplex assays can reduce cost since irrelevant SNPs do not need to be typed and are also more time effective, particularly when a large number of loci is to be analysed, as is usually the case with mtDNA and Y chromosome lineage studies. The hierarchical strategy is also fairly simple and easily reproducible in a casework setting. Consequently a hierarchical design strategy was adopted to facilitate the grouping of the mtDNA and NRY SNPs within our DNA intelligence assay.

The 49 selected haploid SNPs were grouped within four multiplexes, with those SNPs defining basal mtDNA and NRY haplogroups arranged together in Multiplex 1. The result of this primary multiplex is then used to determine which of Multiplexes 2, 3 or 4 is to be performed next. These multiplexes contain the shallowest haplogroups, which mainly stem from the same clade (branch) and are used to achieve further haplogroup

resolution. MtDNA and NRY SNPs defining haplogroups with similar geographical origins are typed within the same multiplex, to reduce the number of tests required and to decrease analysis time. The arrangement of all the selected mtDNA and NRY SNPs across the four multiplexes is demonstrated in Figure 3.1 and Figure 3.2 respectively. The haplogroup defining haplotypes and phylogenies of markers incorporated in each individual multiplex are shown in Figure 3.3 to Figure 3.6. As will be discussed in chapter 4, 13 autosomal SNPs demonstrating associations with BGA and EVCs were also selected. Since these markers are diploid, a maximum of two allele peaks are expected per SNP loci and thus the autosomal SNPs were grouped separately in Multiplex 5 to facilitate primer design and allele calling/genotype interpretation.

3.2.3 Multiplex primer design

In a SNaPshot® multiplex SNP genotyping assay all regions containing the target SNPs are amplified in one multiplex PCR reaction and all single base extension (SBE) reactions for each SNP are carried out in one multiplex SNaPshot® reaction. Thus, multiplex PCR primers must be able to produce similar quantities of specific products under the same reaction conditions with minimal interactions between the various primer sets present. Each multiplex contains high copy number mtDNA and single copy number NRY DNA, thus to promote balanced amplification, mtDNA amplicons were designed to be larger (250-430bp) than NRY SNP amplicons (less than 200bp) because small amplicons amplify more efficiently than larger ones. Candidate SNPs that are located in close proximity such as T10238C/ C10400T and A4715G/C4883T were amplified on the same amplicon using the same primer pair. Primer design programs such as Primer Express (Life Technologies, USA), PrimerPlex (Premier BioSoft, USA) and MP Primer (Shen *et al.*, 2010) can be used to design multiplex PCR primer sets. However, these programs utilise stringent parameters and it is often difficult to achieve large multiplex PCR primer sets. For this reason, each primer pair was designed individually using the criteria described in Section 2.2.1.1 and the multiplex PCR primer sets were then tested for any non-specific interactions using AutoDimer version 1.0 (Vallone and Butler, 2004) and Primer BLAST (NCBI, USA).

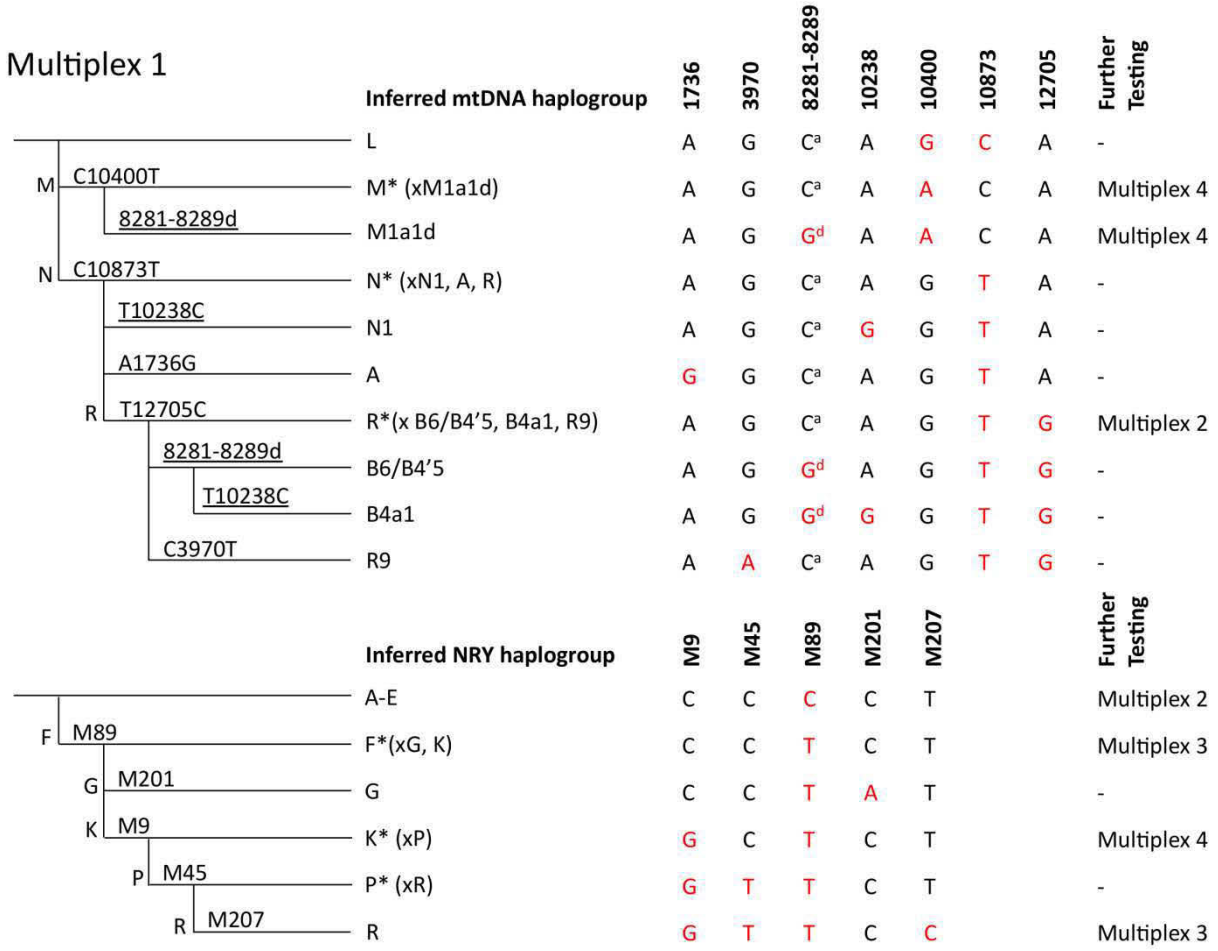


Figure 3.3: MtDNA and NRY SNP phylogenies and haplogroup defining haplotypes for Multiplex 1.

Recurrent SNPs are underlined. The alleles of deletion 8281-8289 are denoted as 'a' (ancestral) and 'd' (deletion). Alleles shown in red are the minimal SNPs required for haplogroup assignment. Additional multiplexes required for further haplogroup resolution are noted.

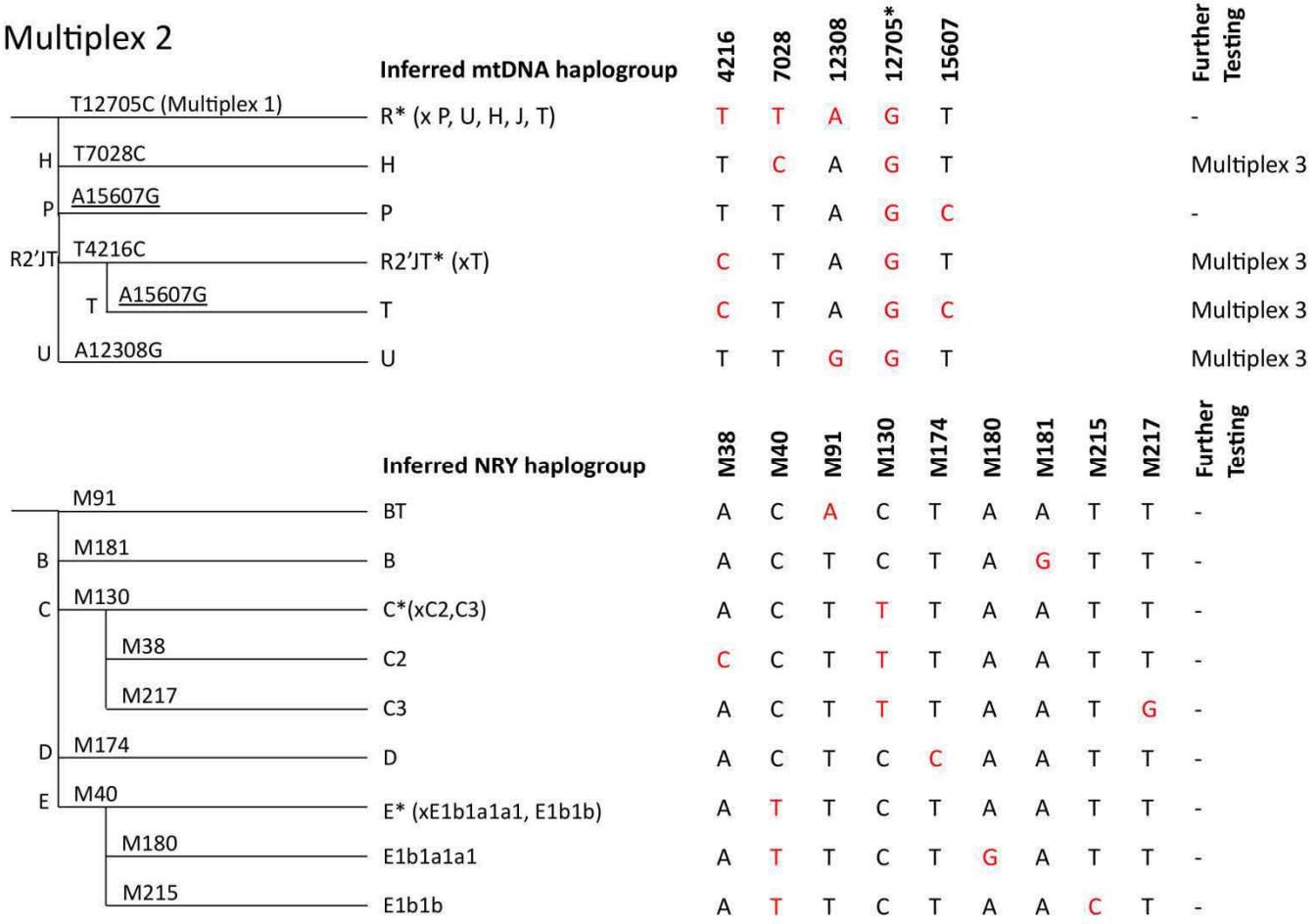


Figure 3.4: MtDNA and NRY SNP phylogenies and haplogroup defining haplotypes for Multiplex 2.

Alleles shown in red are the minimal SNPs required for haplogroup assignment. SNPs marked with an asterisk (*) indicate SNPs typed in a previous multiplex. Additional multiplexes required for further haplogroup resolution are noted.

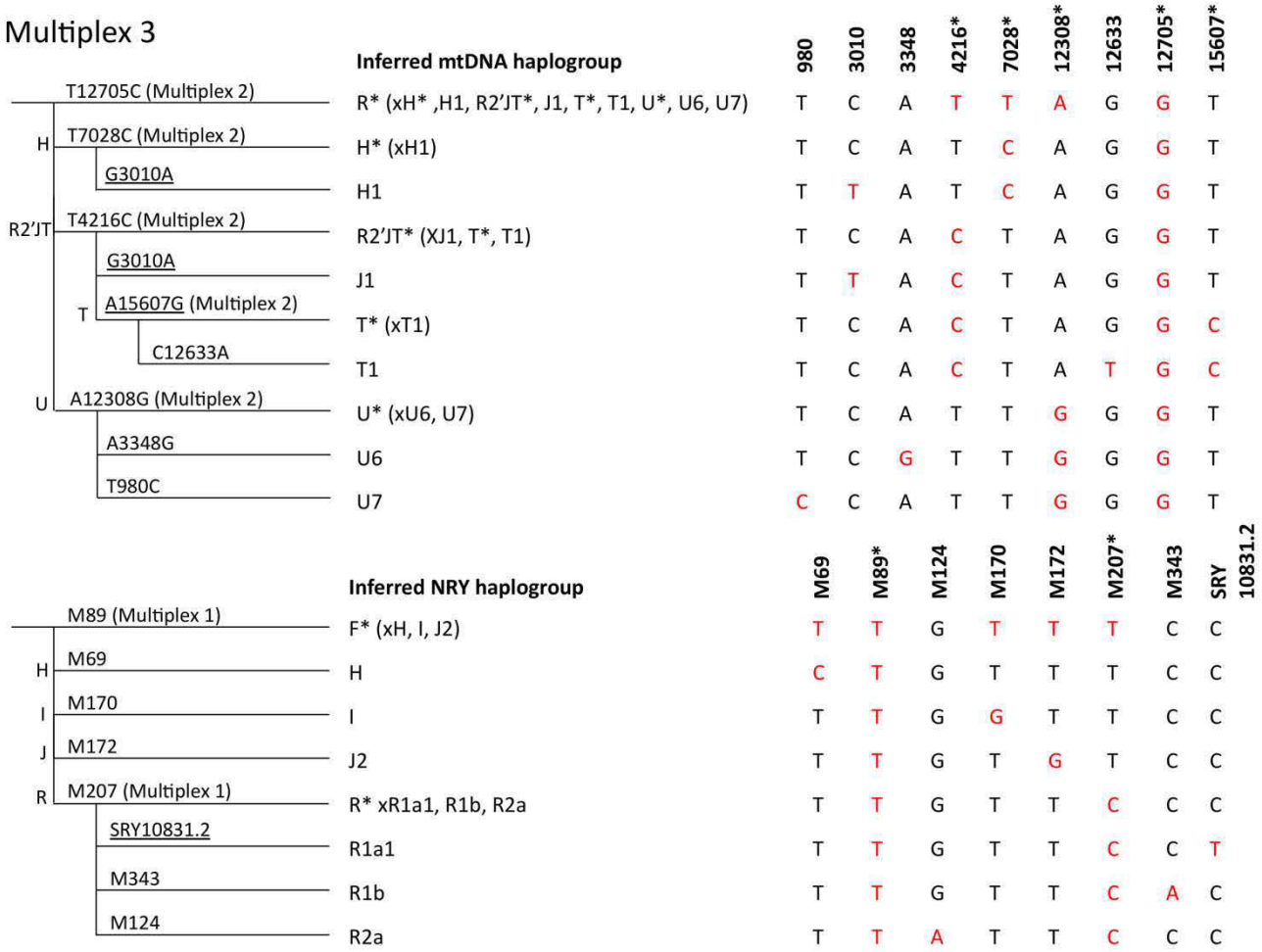


Figure 3.5: MtDNA and NRY SNP phylogenies and haplogroup defining haplotypes for Multiplex 3.

Recurrent SNPs are underlined. Alleles shown in red are the minimal SNPs required for haplogroup assignment. SNPs marked with an asterisk (*) indicate SNPs typed in a previous multiplex.

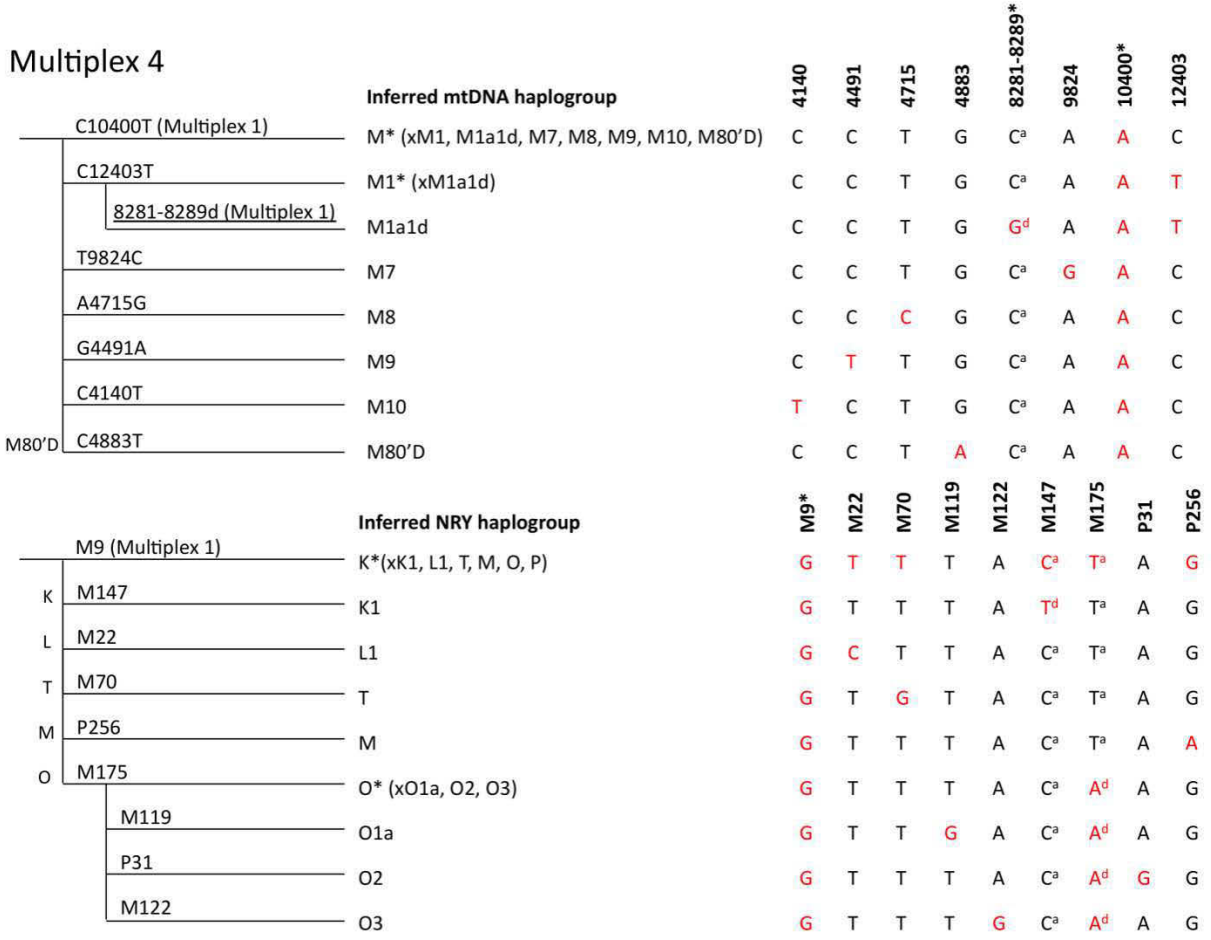


Figure 3.6: MtDNA and NRY SNP phylogenies and haplogroup defining haplotypes for Multiplex 4.

Alleles shown in red are the minimal SNPs required for haplogroup assignment. SNPs marked with an asterisk (*) indicate SNPs typed in a previous multiplex. The alleles of deletion M147 and M175 are denoted as 'a' (ancestral) and 'd' (deletion).

As described in section 2.2.1.2, the design of SBE primers for use in multiplex SNaPshot™ reactions must ensure that all SNaPshot™ products can be resolved and detected without any overlap in size and colour and that the primers are specific and efficiently extended. The DNA region used to design the SBE primers is restricted, as the primer must anneal one base short of the SNP site on either the sense or antisense strand, making it more difficult to design primers which abide to the same criteria, particularly in terms of GC content. All multiplex SBE primers were designed manually and non-homologous 5' tails that would allow for allele discrimination based on size differences with minimal primer interactions were added according to the criteria outlined in section 2.2.1.2. The designed SBE primer sets were also analysed for non-specific interactions using AutoDimer version 1.0 (Vallone and Butler, 2004) and Primer BLAST (NCBI, USA). The selected PCR and SBE primer sequences designed for multiplex 1 to 4 are provided in Tables 2.1 to 2.4.

3.3 Results and Discussion

Multiplex SNaPshot® reactions require extensive optimisation of the preliminary multiplex PCR and SBE reactions to achieve simultaneous and balanced amplification of multiple DNA fragments and SBE products that can be unambiguously detected. Low sensitivity, poor specificity and preferential amplification are common problems, thus in addition to careful primer design it is necessary to optimise PCR conditions by varying reagent concentrations such as $MgCl_2$, dNTP, primer, template and *Taq* polymerase and altering annealing temperatures to maximise amplification success (Markoulatos *et al.*, 2002). Each of the Multiplex 1 to 4 PCR and SBE reactions was optimised as outlined in Section 2.2.2. Since all primers were designed according to the same criteria, the optimised Multiplex 1 PCR conditions were used as the starting conditions for the optimisation of the Multiplex 2, 3 and 4 PCR reactions.

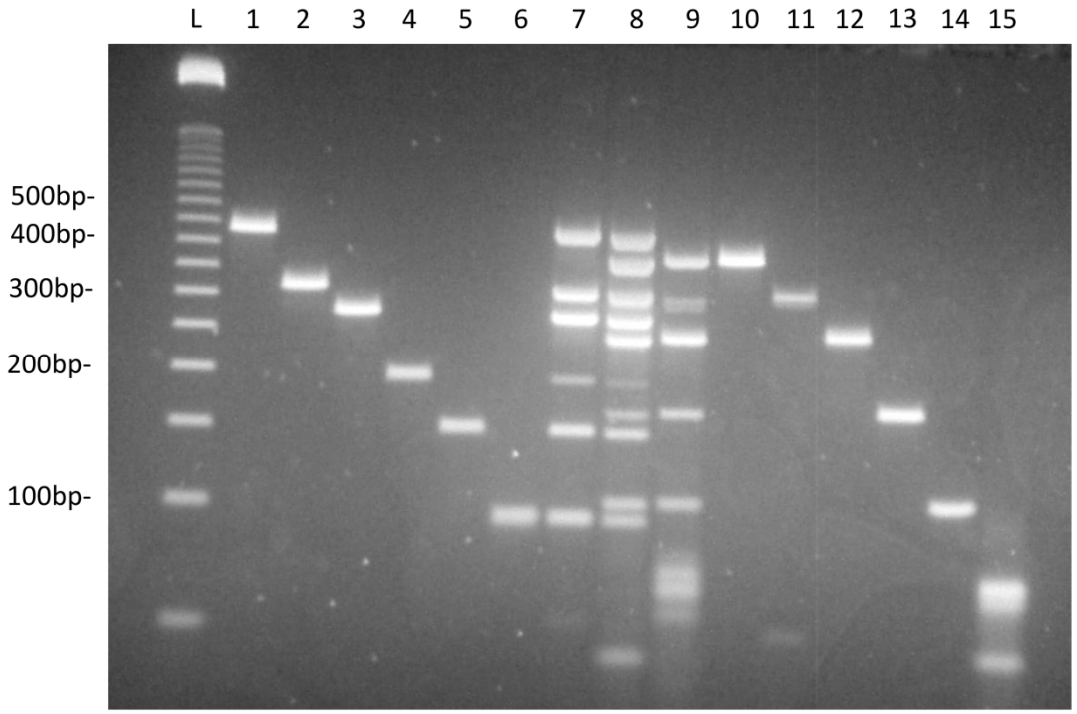
3.3.1 Multiplex 1 assay optimisation

3.3.1.1 *Singleplex and multiplex amplification of Multiplex 1 PCR primers*

To test PCR primer specificity and efficacy Multiplex 1 PCR primer pairs were amplified individually in singleplex PCR reactions (25µL) containing one 2mm purified

FTA disc, 3mM $MgCl_2$, 0.2mM dNTP mix, 1× reaction buffer, 0.75U *Taq* Polymerase, PCR forward and reverse primer concentrations of 0.2µM for Y-SNP primers and 0.1µM for mtDNA SNP primers and an annealing temperature of 57°C. Successful amplification of all individual Multiplex 1 PCR primer pairs was confirmed by agarose gel electrophoresis, where similar yields of all PCR products of expected size were visualised in the absence of non-specific products (Figure 3.7) An exception was the C3970T singleplex product, which showed the lowest PCR product yield and a faint primer dimer band indicating that some non-specific interactions were occurring between the forward and reverse PCR primers (Figure 3.7, lane 11). Note that the mtDNA SNPs for T10238C and C10400T were amplified by the same PCR primer pair and are located on the same amplicon.

To test whether all multiplex 1 PCR products could be amplified simultaneously, the PCR primers were then grouped into a 5-Plex containing PCR primers for C8281G, C3970T, A1736G, M207, M201, a 6-Plex which contained PCR primers for T10238C/ C10400T, C10873T, T12705C, M9, M89, M45 and an 11-Plex combining all multiplex 1 PCR primers. The 5-plex and 6-plex combinations were performed to enable unambiguous identification of amplicons of similar sizes, which are difficult to identify in the 11-plex PCR product. The same conditions as those used in the singleplex reactions were applied, except that all PCR primers in each of the 5-Plex, 6-Plex and 11-Plex had an equimolar final concentration of 0.2µM. Agarose gel electrophoresis confirmed the presence of five PCR products in the 5-Plex and six PCR products in the 6-Plex, but only ten PCR products were identified in the 11-Plex (Figure 3.7). Since the C10873T (308bp) and the C3970T (301bp) products were visible separately in the 6-Plex and 5-Plex respectively, it was assumed that they were present in the 11-Plex reaction but could not be resolved in the 4% agarose gel due to the difference of only 8bp in their size. The minor primer dimers visualised in the 11-Plex are indicative of minimal non-specific interactions occurring between Multiplex 1 PCR primer pairs. The Multiplex 1 PCR product yields were slightly reduced in the multiplex reactions compared to the individual singleplex reactions due to competition between primer pairs and depletion of reagents. There were also noticeable differences in amplification efficiencies between PCR products. Overall the mtDNA PCR products (with the exception of C3970T) were more abundant compared to the NRY PCR products, especially M9 which showed the lowest yield in the 11-Plex (Figure 3.7, Lane 8). This



Legend:
L: 50bp DNA ladder
1: T10238C/ C10400T (426bp)
2: C10873T (308bp)
3: T12705C (273bp)
4: M9 (192bp)
5: M89 (148bp)
6: M45 (92bp)
7: 6-Plex: T10238C/ C10400T, C10873T, T12705C, M9, M89, M45
8: 11-plex: All multiplex 1 PCR primers
9: 5-Plex: C8281G, C3970T, A1736G, M207, M201
10: C8281G (359-368bp)
11: C3970T (301bp)
12: A1736G (246bp)
13: M207 (163bp)
14: M201 (101bp)
15: 11-plex negative control

Figure 3.7: Multiplex 1 singleplex and multiplex PCR products.

Multiplex 1 PCR primer pairs were amplified individually and in 6-Plex, 7-Plex and 13-Plex reactions using standard PCR conditions. The NRY loci (M9, M89, M45, M207, M201) were amplified at 0.2µM and the mtDNA loci (T10238C/C10400T, C10873T, T12705C, C8281G, C3970T, A1736G) at 0.1µM in singleplex. All PCR primers were amplified in the 6-Plex, 7-Plex and 13-Plex at an equimolar concentration of 0.2µM.

is presumably due to the higher copy number of these targets compared to the Y-chromosome.

3.3.1.2 *Multiplex 1 PCR optimisation- Adjusting magnesium and dNTP concentration*

When amplifying numerous loci simultaneously, the more efficiently amplified loci will have a negative effect on the product yield of less efficient targets as there is a limited pool of reagents for which all the PCR products compete thus the reagent concentrations in multiplex PCR reactions are usually greater than those applied for singleplex reactions. Since *Taq* polymerase is magnesium dependent, magnesium concentration optimisation is essential. However dNTPs also bind magnesium, so an optimum magnesium/dNTP balance is also important in achieving optimal PCR product yields.

A variety of magnesium/dNTP concentrations ranging from 1mM $MgCl_2$/0.2mM dNTP to 5mM $MgCl_2$/0.5mM dNTP were tested on 11-Plex Multiplex 1 PCR reactions with all other reagent concentrations maintained as per the original 11-Plex reaction described in section 3.3.1.1. The lowest PCR product yields were observed when using 1mM $MgCl_2$/0.2mM dNTP, with some PCR products such as M9 (192bp) amplifying very little or not at all (Figure 3.8, lane 1). Amplification success was improved as the magnesium/dNTP concentration was increased, with the greatest yields achieved when using 5mM $MgCl_2$/0.5mM dNTP (Figure 3.8, lane 8). However, though the yield for M9 was greatest at this concentration, there were striking amplification imbalances between the mtDNA and NRY loci. Generally, more balanced amplification between the mtDNA and the NRY loci were observed when using $MgCl_2$ concentrations greater than 3mM and dNTP concentrations greater than 0.4mM (Figure 3.8). Despite the overall reduced product yield and substantial primer dimers observed when using 4mM $MgCl_2$/0.4mM dNTP, these concentrations were tested against variable *Taq* polymerase concentrations to identify the optimum concentration that would improve product yield and reduce non-specific interactions, whilst maintaining balanced amplification between the mtDNA and NRY loci.

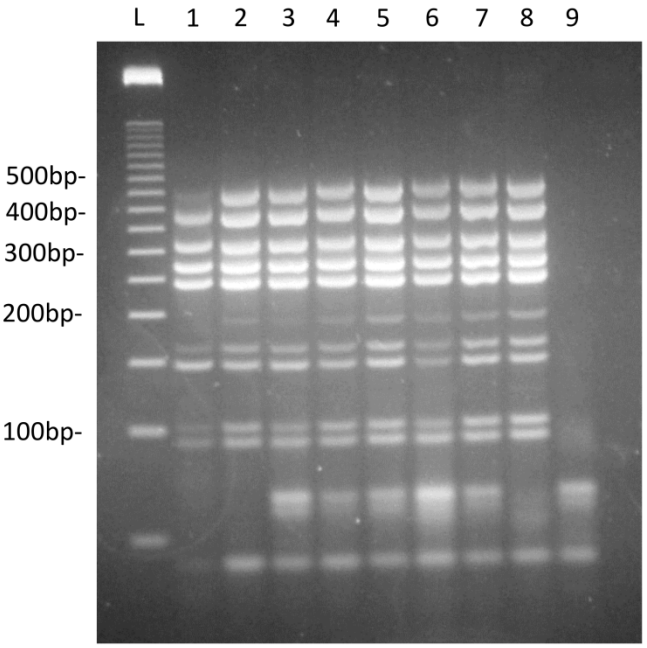


Figure 3.8: Multiplex 1 PCR optimisation of magnesium and dNTP concentration. Multiplex 1 PCR primer pairs were simultaneously amplified at an equimolar concentration of 0.2µM in 11-Plex reactions using standard PCR conditions and variable magnesium and dNTP concentrations as indicated.

3.3.1.3 *Multiplex 1 PCR optimisation- Adjusting Taq polymerase concentration*

To optimise *Taq* polymerase concentration a range from 0.75U to 2U of *Taq* polymerase per 25µL reactions was tested on Multiplex 1 11-plex reactions using one 2mm purified FTA disc, 4mM $MgCl_2$, 0.4mM dNTP mix, 1× reaction buffer and equimolar multiplex PCR primer concentrations of 0.2µM (Figure 3.9). A concentration of 1.75U *Taq* was judged to be the optimal enzyme concentration as it showed the most balanced yields between the mtDNA and NRY loci, whilst also demonstrating minimal primer dimer formation. When using 2U *Taq* polymerase the PCR product yield, particularly for the Y loci, began to decrease and there was an increase in primer dimer formation. This was attributed to the high glycerol content in the stock solution due to excessive enzyme likely inhibiting amplification.

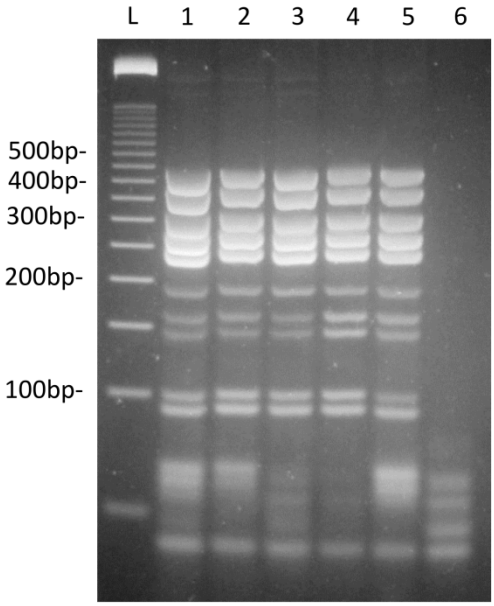


Figure 3.9: Multiplex 1 PCR optimisation of *Taq* polymerase concentration.
Multiplex 1 PCR primer pairs were simultaneously amplified at an equimolar concentration of 0.2µM in 11-Plex reactions using standard PCR conditions including 4mM $MgCl_2$, 0.4mM dNTP mix and **(1)** 0.75U, **(2)** 1U, **(3)** 1.5U, **(4)** 1.75U and **(5)** 2U *Taq* polymerase. The 50bp DNA ladder **(L)** and the 11-Plex negative control **(6)** are also shown.

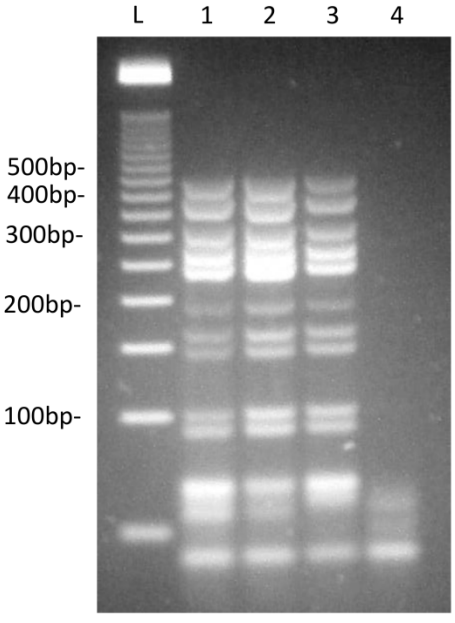


Figure 3.10: Multiplex 1 PCR optimisation of PCR buffer concentration.
Multiplex 1 PCR primer pairs were simultaneously amplified at an equimolar concentration of 0.2µM in 11-Plex reactions using standard PCR conditions including 4mM $MgCl_2$, 0.4mM dNTP mix, 1.75U *Taq* polymerase and **(1)** 1×buffer, **(2)** 1.5×buffer, **(3)** 2×buffer. The 50bp DNA ladder **(L)** and the 11-Plex negative control **(4)** are also shown.

3.3.1.4 *Multiplex 1 PCR optimisation- Adjusting PCR buffer concentration:*

Though the mtDNA and the NRY loci show relatively balanced amplification after the optimisation of magnesium, dNTP and *Taq* concentrations, the mtDNA PCR product yields remained much stronger compared to the NRY PCR products yields (Figure 3.8 and Figure 3.9). It was been reported that large PCR products amplify better at lower salt concentrations whilst short PCR products amplify better at higher salt concentrations (Henegariu *et al.*, 1997). Thus, in an attempt to further equalise the amplification between the mtDNA and NRY loci, the PCR buffer concentration of the Multiplex 1 11-plex reactions was increased. As shown in Figure 3.10, increasing the PCR buffer concentration to 1.5× resulted in an overall increase in yield across all PCR products. Increasing the buffer concentration to 2× reduced the mtDNA product yields, but to a greater extent than the NRY products, giving better equalisation between the mtDNA and NRY loci (Figure 3.10, Lane 3). Hence, the 11-Plex PCR product using 4mM $MgCl_2$, 0.4mM dNTP mix, 1.75U *Taq* polymerase, 2× PCR buffer and a final equimolar PCR primer concentration of 0.2μM was used as template for preliminary SBE singleplex and multiplex reactions, with the possibility for further optimisation of PCR primer concentrations if suggested during Multiplex 1 SBE optimisation.

3.3.1.5 *Singleplex extension of Multiplex 1 SBE primers using Multiplex 1 PCR template*

Following preliminary Multiplex 1 PCR optimisation, Multiplex 1 SNaPshot™ optimisation commenced. Prior to multiplexing it is important to verify that all SNaPshot™ products can be resolved and detected without any overlap in size and colour. The electrophoretic mobility of Multiplex 1 SBE primers was tested using the SNaPshot® Primer Focus™ kit (Life Technologies, USA) as described in section 2.2.1.3, before undertaking any SNaPshot™ reactions. The results indicated that the multiplex 1 SBE primers showed sufficient spatial separation and unambiguous detection of all extension products (data not shown) allowing for the commencement of Multiplex 1 SNaPshot™ optimisation.

The Multiplex 1 SBE primers were first extended individually using as template the 11-plex Multiplex 1 PCR product depicted in Figure 3.10 (Lane 3), which was amplified

with the optimised conditions summarised in Section 3.3.1.4. Singleplex SNaPshot™ reactions were prepared as per section 2.2.2.5, with each SBE primer at a final concentration of 0.2µM. SNaPshot™ products were purified prior to capillary electrophoresis as described in sections 2.2.2.6 and 2.2.2.7. Electropherograms for these singleplex products showed unsatisfactory results, with the majority of alleles demonstrating signals less than 1000RFU with the exception of A1736G, which showed the strongest signal (>3000RFU). The SBE primers for C3970T, T12705C and C10400T failed to amplify at all (Figure 3.11). The alleles that were detected were the correct size and expected colour and the majority of missing alleles were those of the longest primers (>50bp).

The Multiplex 1 PCR product used as template for these singleplex SNaPshot™ reactions showed strong PCR product yields (Figure 3.10, Lane 3), thus it was assumed that the low signals observed were due to inadequate SBE primer concentrations and the singleplex SNaPshot™ reactions were repeated using 2µM SBE primer. With this 10-fold increase in SBE primer concentration, the signal of most alleles increased to greater than 2000RFU and the SBE primer for T12705C that initially failed to amplify at 0.2µM was now visible (Figure 3.12). Though an increase in primer concentration appeared to have a positive effect, SBE primers greater than 45bp still showed signals averaging 500-1000RFU. However, the longest primer C10400T as well as the C3970T primer still failed to amplify when using 2µM.

3.3.1.6 *Multiplex 1 SBE reactions using Multiplex 1 PCR product as template.*

Despite several problematic SBE primers failing to be detected in singleplex SNaPshot™ reactions, the Multiplex 1 SBE set was also tested in multiplex SNaPshot™ reactions to investigate potential inhibitory non-specific primer interactions. The primer set was divided into two separate 6-Plex SBE reactions, 6-Plex A (M45, M207, M89, C10873T, T12705C, C10400T) and 6-Plex B (M9, A1736G, M201, C3970T, C10238T, C8281G) and a 12–plex reaction containing all SBE primers. Each SBE primer in these reactions was extended at a final equimolar concentration of 0.2µM. All Multiplex 1 SBE primers were also grouped in two 12-Plex reactions. An additional 12-plex2 reaction was tested to compare the effect of varying the SBE primer concentrations in

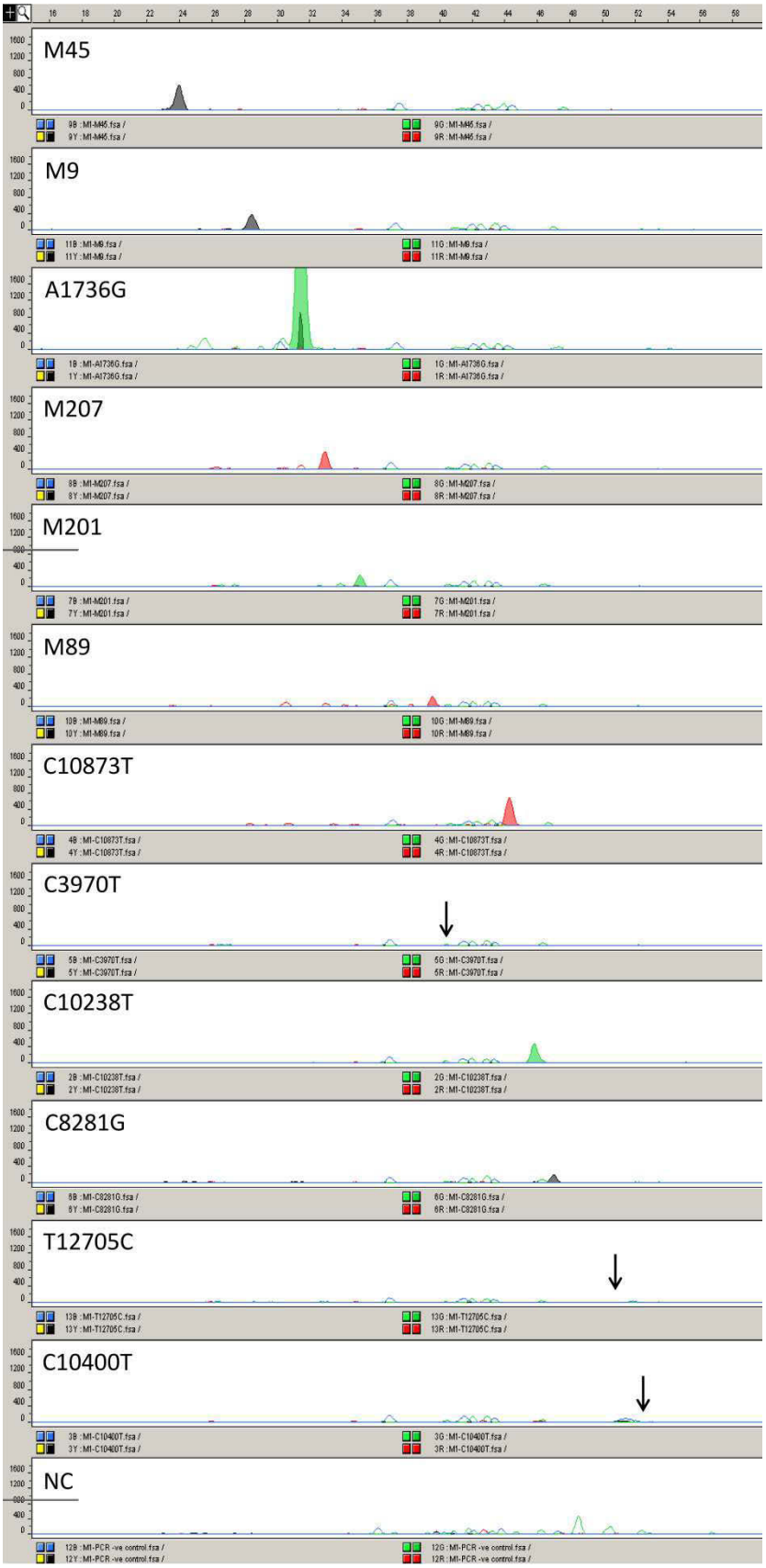


Figure 3.11: Capillary electropherograms of singleplex extensions of 0.2µM Multiplex 1 SBE primers using Multiplex 1 PCR product as template.

Black arrows indicate the expected position for alleles C3970T and C10400T that have failed to extend. The 11-Plex PCR negative control is also shown (NC).

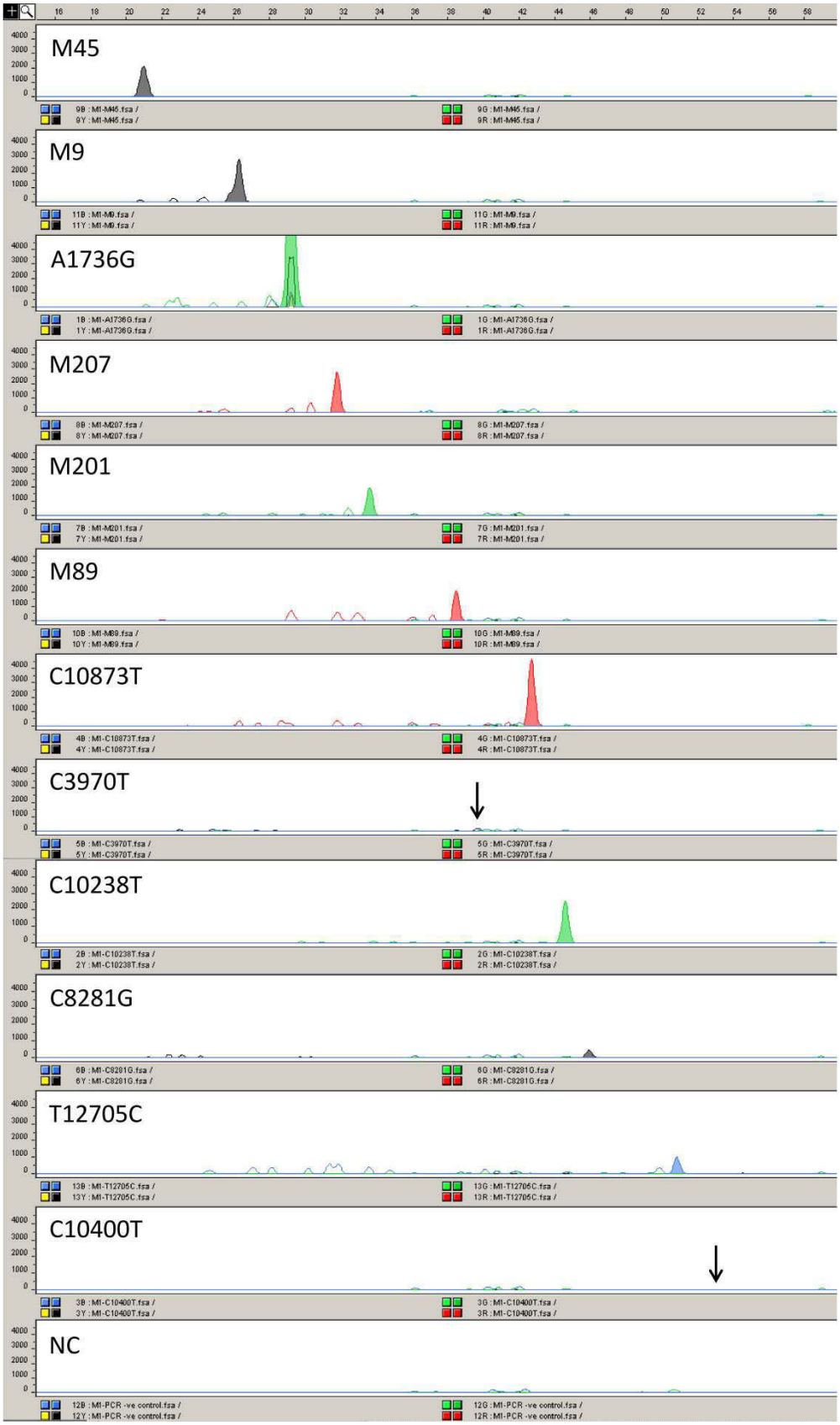


Figure 3.12: Capillary electropherograms of singleplex extensions of 2µM Multiplex 1 SBE primers using Multiplex 1 PCR product as template.

Black arrows indicate the expected position for alleles C3970T, T12705C and C10400T that have failed to extend. The 11-Plex PCR negative control is also shown (**NC**).

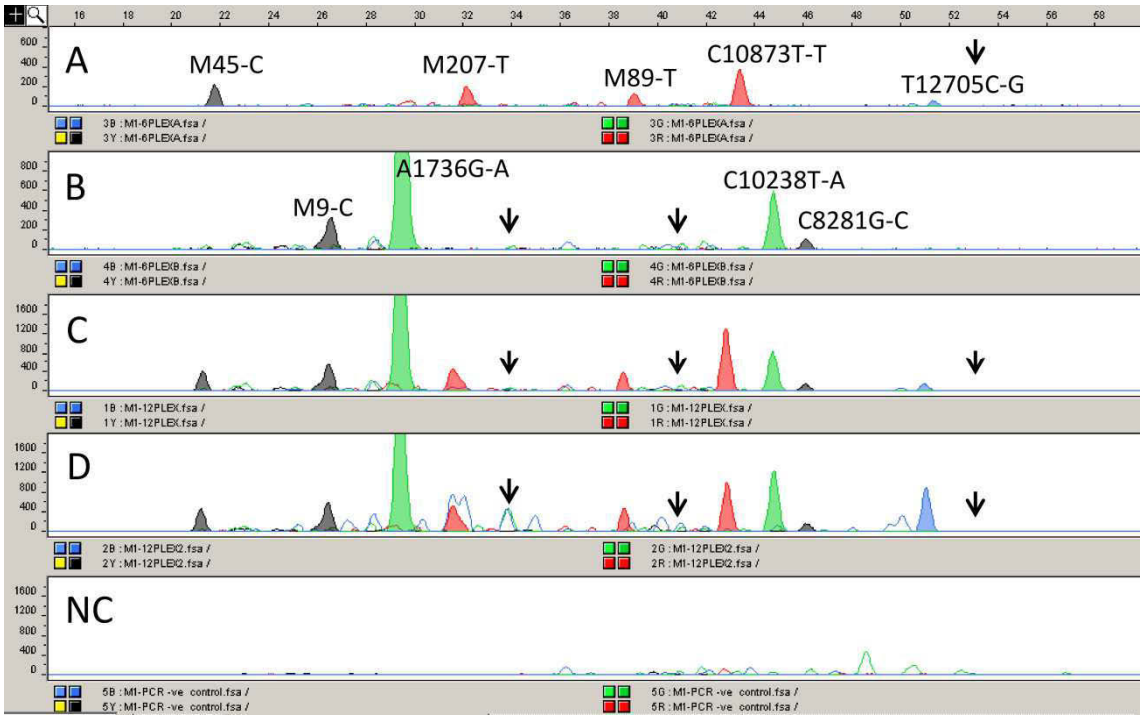
line with the strength of extension seen in the singleplex reactions. In the 12-Plex2 multiplex, the concentration of SBE primers was set at 0.4µM except for T12705C, C3970T and C10400T, which were set at 2µM and C10873T and A1736G, which were the strongest alleles in the singleplex SNaPshot™ reactions and were maintained at 0.2µM. The same Multiplex 1 PCR product used for the singleplex SBE reactions was used as template for these reactions.

Electropherograms for the separation of these Multiplex 1 SNaPshot™ products (Figure 3.13A-C) showed successful extension of all Multiplex 1 SBE primers when using a concentration of 0.2µM except M201 (34bp), C3970T (40bp) and C10400T (53bp), the latter two of which had also failed in the singleplexes. However, the signal of the extended alleles was low, ranging from 200RFU to 1200RFU, with the exception of A1736G, which was the most efficiently amplified at over 2000RFU. Modifications in SBE primer concentrations in 12-Plex2 had minimal effects on the signal of alleles compared to equimolar SBE concentration of 0.2µM, except for T12705C, which showed roughly a four-fold increase in signal when 2µM of SBE primer was used. (Figure 3.13D). The M201 allele also appeared to have extended at an SBE primer concentration of 0.4µM however a blue artefact peak masked the green allele peak.

3.3.1.7 *Investigating the failed extension of the C10400T, C3970T, and T12705C alleles*

Since the C10400T allele failed to extend in the multiplex and the singleplex SNaPshot™ reactions despite the PCR product for this locus showing a strong yield, it was deduced that the failed extension resulted from an inefficient SBE primer. Though T12705C could be extended, the signal of this allele was still low when using a SBE primer concentration of 2µM, which is considerably high. These inefficient SBE primers were the longest and had 5' aspecific tail lengths larger than the specific homologous primer sequence, which also had low GC content (~30%). It was hypothesised that these long tails were too 'floppy' and prevented the primer from strongly annealing to the complimentary target sequence, and that this was exacerbated by the low GC content of the homologous portion, thereby decreasing SBE primer amplification efficiency. For this reason, the SBE primer for C10400T was redesigned to be homologous to the alternate strand and the final expected size was reduced to

36bp. The SBE primer for T12705C was also modified and reduced in size to 31bp by removing the 5' aspecific tail and adding an extra homologous 'G' base. The modified SBE primer sequences for these alleles are provided in Table 2.2.



Legend:
(A) 6-plexA: M45, M207, M89, C10873T, T12705C, C10400T SBE primers = 0.2µM.
(B) 6-PlexB: M9, A1736G, M201, C3970T, C10238T, C8281G SBE primers = 0.2µM.
(C) 12-Plex: All Multiplex 1 SBE primers =0.2µM.
(D) 12-Plex2: All SBE primers = 0.4µM except C10873T and A1736G=0.2µM and T12705C, C3970T and C10400T=2µM.
(NC) 11-Plex PCR negative control with SBE primer concentrations used in (D).

Figure 3.13: Capillary electropherograms of Multiplex 1 SBE primer optimisations using Multiplex 1 PCR product as template.

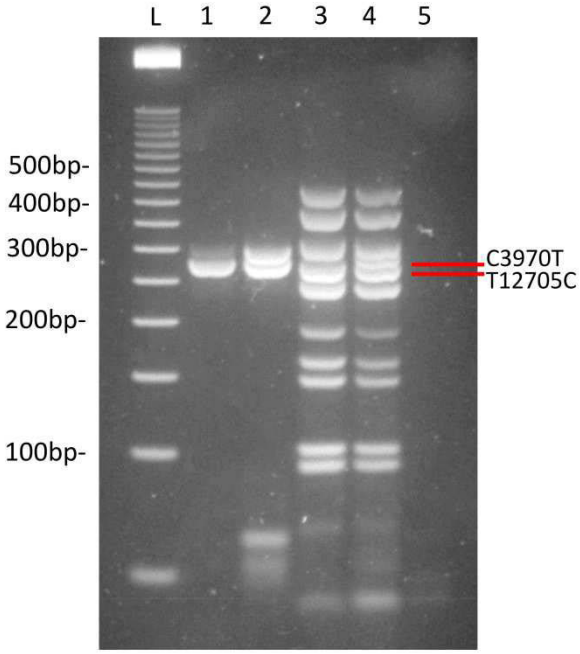
Multiplex 1 PCR product was prepared using the optimised conditions summarised in Section 3.3.1.4 and an equimolar PCR primer concentration of 0.2µM. The SBE primer concentrations were modified as detailed in the legend. The black arrows indicate the position for alleles M201 (34bp), C3970T (40bp) and C10400T (53bp) that have failed to amplify.

The C3970T allele could not be detected in either singleplex or multiplex SBE reactions using multiplex 1 PCR product as template, despite increases in SBE primer concentration (Figure 3.11-Figure 3.13). Failure to observe a blue or green allele of expected size for C3970T indicated that the specific template containing the SNP was likely to be absent and that the SBE primer was only able to self-extend. The C3970T PCR product was initially assumed to have successfully amplified since ten correct sized bands were resolved when 11-Plex PCR products were visualised in 4% agarose gels and non-specific interactions between multiplex PCR primer sets were thus excluded, as two products should be affected if this was the case (Figure 3.7, lane 8). The reason for only observing ten out of the expected eleven bands was attributed to the likelihood that the C3970T amplicon could not be resolved from the C10873T amplicon, which is only just 7bp larger and that these PCR products were visualised as a doublet band. However, failure of the C3970T allele to be detected raised the possibility that the C3970T PCR product may not be amplifying efficiently in the multiplex PCR as it is difficult to ascertain the yield within the doublet. Furthermore, the singleplex amplification of the C3970T PCR primer pair was the only Multiplex 1 PCR singleplex to show primer dimers and the lowest singleplex PCR product yield (Figure 3.7, Lane 11) indicating the possibility of non-specific interactions between the C3970T primers.

The autodimer results for the C3970T PCR primer pair were re-evaluated and a primer dimer interaction between the forward PCR primer was identified and is depicted in Figure 3.14. An alternate forward primer for C3970T was designed which produced a 297bp amplicon (Table 2.1). The original and modified C3970T PCR primer pairs were tested at a concentration of 0.3µM in a duplex PCR reaction also containing T12705C (0.1µM) and in an 11-Plex PCR reaction containing all other Multiplex 1 PCR primers at 0.2µM except M9, which was increased to 0.4µM, as this product showed the lowest yield after preliminary multiplex 1 PCR optimisation (Figure 3.10). All other optimised PCR conditions were as summarised at the end of section 3.3.1.4. Use of the modified C3970T primer pair resulted in considerably greater C3970T PCR product in both duplex and 11-plex reactions, compared to use of the original PCR primer pair (Figure 3.15). Since the new C3970T amplicon is smaller than the original amplicon it could now be resolved from C10873T, allowing visualisation of all 11 Multiplex 1 PCR products. The replacement of the forward C3970T primer at a concentration of 0.3µM

```
3970T-F CCTTCGACCTTGCCGAAG versus 3970T-F CCTTCGACCTTGCCGAAG
Matches = 12
Score = 7
CTTCGNCNNNGNCGAAG
est. tm = 8.6 °C
DeltaG 37 degrees = -2.65 kcal/mole
                  3'-GAAGCCGTTCCAGCTTCC-5'
                     |||||x|xxx|x|||||
                 5'-CCTTCGACCTTGCCGAAG-3
```

Figure 3.14: Autodimer result for C3970T PCR forward primer dimer screen



Legend:

L: 50bp DNA ladder
1: T12705C /original C3970T duplex
2: T12705C /modified C3970T duplex
3: Multiplex 1 11-Plex with original C3970T PCR primers
4: Multiplex 1 11-Plex with modified C3970T PCR primer
5: 11-plex negative control

Figure 3.15: Amplification of the original and modified C3970T PCR primer pair in duplex and Multiplex 1 PCR reactions.

Multiplex 1 duplex and 11-Plex PCR products were prepared using the optimised conditions described in Section 3.3.1.4. The original and modified C3970T PCR primers were amplified at 0.3µM in duplex with T12705C at 0.1µM. An equimolar PCR primer concentration of 0.2µM was used for the 11-Plex PCR, except for C3970T and M9, which were amplified at 0.3µM and 0.4µM respectively.

had little effect on the yield of other multiplex 1 PCR products, with relatively balanced amplification remaining.

The efficacy of the SBE primer for the C3970T allele was also tested (0.2µM) in a singleplex SNaPshot™ reaction using singleplex C3970T PCR product generated using standard amplification conditions. As shown in Figure 3.16, a black peak was observed at the C3970T allele position instead of the expected blue or green allele. Re-review of the autodimer results for the C3970T SBE primer showed a notable hairpin involving the 3' end of the primer (Figure 3.17), suggesting that the observed black peak is an artefact arising from the 3'end extension of the hairpin depicted in Figure 3.17 with a C base. Numerous smaller black peaks were also observed that were deduced to be products arising from the extension of truncated SBE primers from low purity primer batches and were only mainly visible in singleplex SNaPshot™ reactions. The SBE primer could not be redesigned to be homologous to the alternative sense strand as the 5' to 3' sequence adjacent to the SNP site is very GC rich (>80%) and showed strong self-annealing for up to 20 bps from the C3970T SNP site. As an alternative solution the original C3970T SBE primer homologous to the antisense strand was modified by the addition of a mismatch. As shown in Figure 3.17, a 'G' nucleotide was substituted for the 'T' nucleotide to reduce the strength of the hairpin at the 3' end. The modified C3970T primer sequence is provided in Table 2.2.

3.3.1.8 *Singleplex extension of modified multiplex 1 SBE primers using the modified multiplex 1 PCR product as template*

The modified SBE primers for C3970T, T12705C and C10400T were tested in singleplex SNaPshot™ reactions using Multiplex 1 PCR product incorporating the modified C3970T PCR primer pair as template. Multiplex 1 PCR reactions were prepared using one 2mm purified FTA disc and the optimised PCR conditions summarised at the end of section 3.3.1.4 were applied. All PCR primers were kept at an equimolar concentration of 0.2µM, except for the modified C3970T PCR primers at a concentration of 0.3µM and M9 at concentration of 0.4µM. All modified SBE primers were extended at a concentration of 0.2µM.

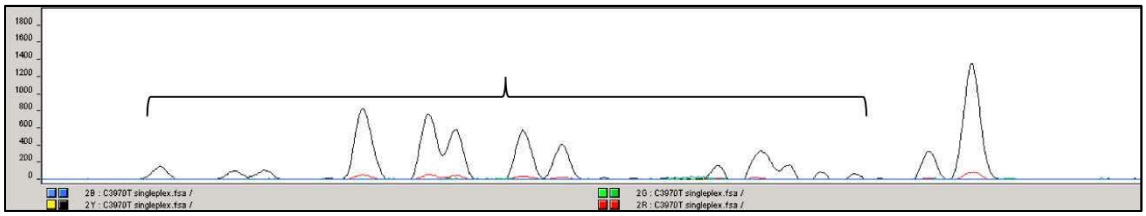


Figure 3.16: Capillary electropherograms of the singleplex extension of 0.2µM C3970T SBE primer using singleplex PCR product as template.
Extended brackets highlight non-specific extension products.

```
C3970T
Matches = 5
Score = 3
NNGAATA
est. tm = less than zero
DeltaG 37.0 degrees = -0.93 kcal/mole

 TCGGCTTATGTGTAACAGTCTGAAAGTG-5'
A  XX|||||
 TGAAGAATA-3'
```

Figure 3.17: Autodimer result for C3970T SBE primer hairpin screen.
The 'G' nucleotide coloured in red denotes the template nucleotide for the extension of the above hairpin which will result in the single base extension of the C3970T SBE primer with a 'C' nucleotide that will be visualised as a black non-specific peak. The underlined 'T' nucleotide denotes the site where the mismatch was inserted within the modified C3970T SBE primer via substitution with a 'G' nucleotide to minimise the impact of the hairpin amplification.

As evident in Figure 3.18, separation of the singleplex SNaPshot™ products using capillary electrophoresis demonstrated successful extension of all modified SBE primers and much improved signals compared to when using the original SBE primers which required concentrations of up to 2µM to amplify (Figure 3.12). The signals for alleles T12705C and C10400T were greater than 2000RFU when using an SBE primer concentration of only 0.2µM. Extension of the original C3970T SBE primer was now detectable in a singleplex SNaPshot™ reaction using multiplex 1 PCR product containing the modified C3970T PCR primer pair, suggesting that the template

insufficiency had initially contributed to the failed extension of this allele. However, an approximately tenfold increase in signal was observed for the modified C3970T SBE primer with the inserted mismatch (Figure 3.18B) compared to the original SBE primer (Figure 3.18A) thus confirming that the inefficient original SBE primer also contributed significantly to the failure of detecting this allele.

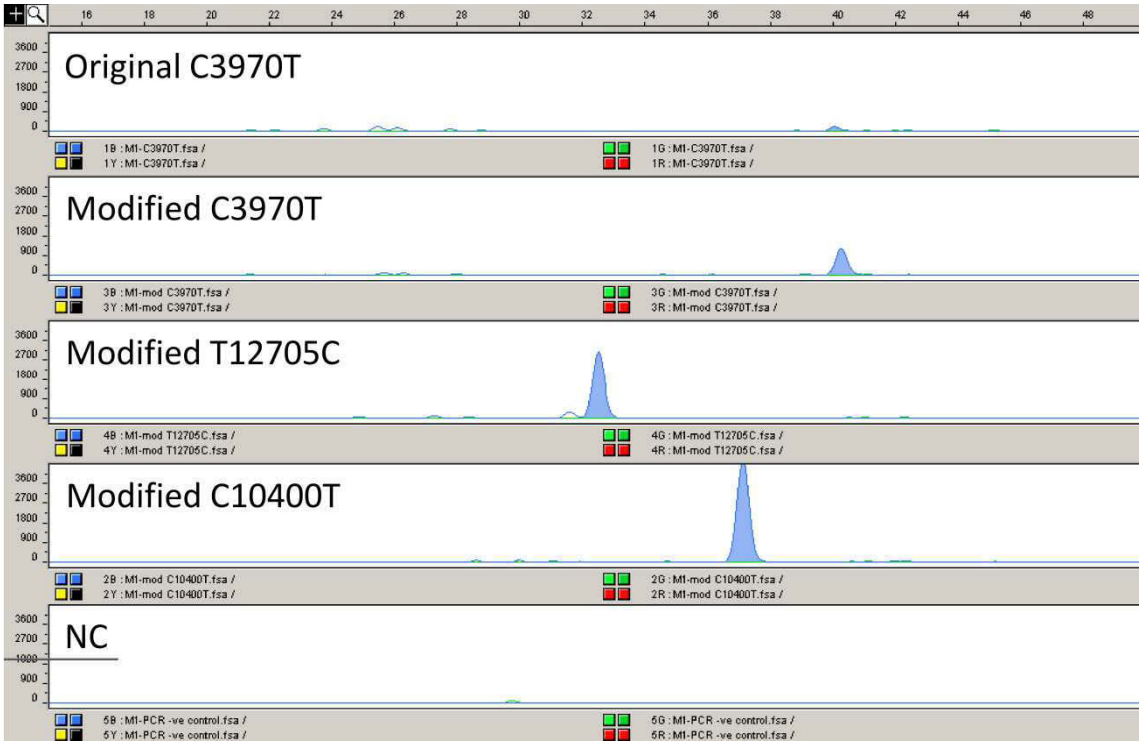


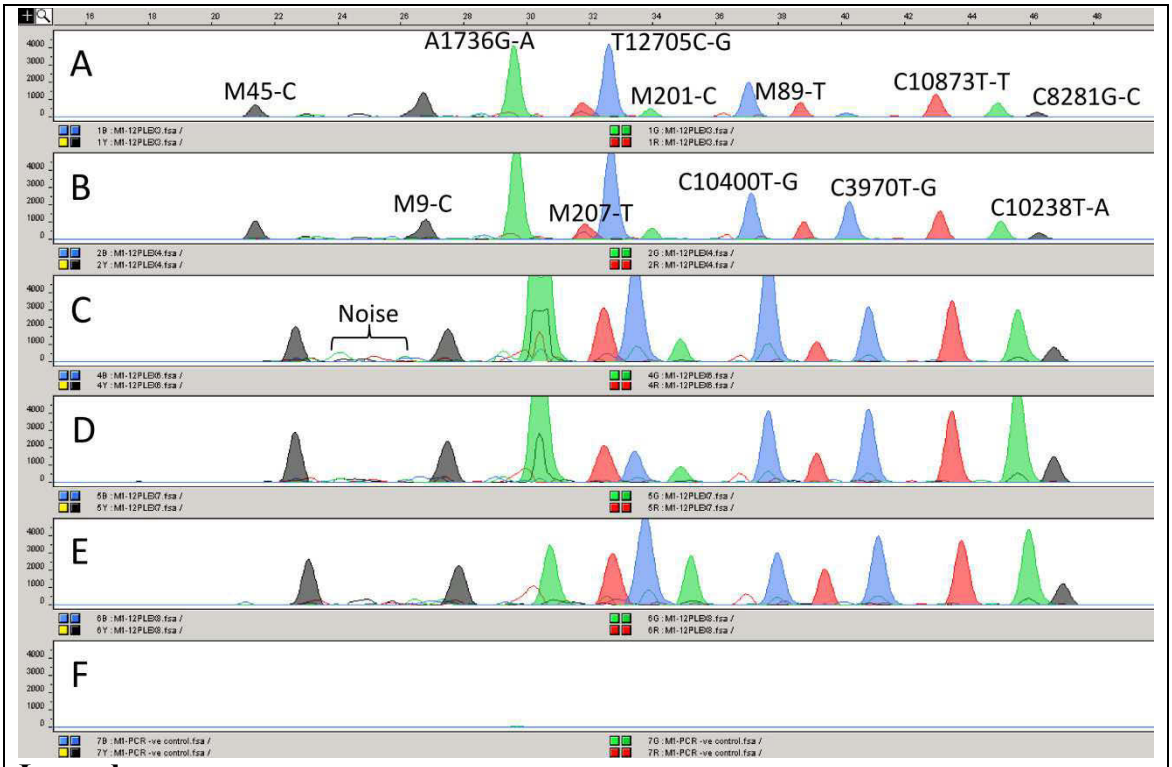
Figure 3.18: Capillary electropherograms of singleplex extensions of 0.2µM of the modified C3970T, T12705C and C10400T Multiplex 1 SBE primers using Multiplex 1 PCR product as template.

The Multiplex 1 PCR product was prepared using the optimised conditions described in Section 3.3.1.4 and an equimolar PCR primer concentration of 0.2µM, except for the modified C3970T PCR primers at 0.3µM and M9 at 0.4µM. The original C3970T SBE primer was also extended at 0.2µM for comparison. All SBE primers were extended at a concentration of 0.2µM. The 11-Plex PCR negative control is also shown (**NC**).

3.3.1.9 *Equalisation of allele signals through the optimisation of Multiplex 1 PCR and SBE primer concentrations*

Initially, the modified Multiplex 1 SBE primers were grouped and tested in 12-Plex SNaPshot™ reactions with the same Multiplex 1 PCR product used for the singleplex SNaPshot™ reactions as template (see section 3.3.1.5). Most of the original SBE primers that showed lower signals in previous optimisation experiments were tested at a concentration of 0.4µM except for C10873T and A1736G and the modified T12705C, C3970T and C10400T SBE primers, which showed higher signals and so were tested at a lower concentration of 0.2µM. Another 12-Plex SNaPshot™ reaction was also tested on an alternative Multiplex 1 PCR product that incorporated all the same PCR primer and reagent concentrations, except that the original C3970T PCR primer pair was included. Capillary electrophoresis of the Multiplex 1 SBE products showed successful extension of all twelve alleles with correct size and colour extension products with either of the Multiplex 1 PCR products as template (Figure 3.19A & B). However, the signal for the modified C3970T SBE primer increased to approximately 2500RFU when using the Multiplex 1 PCR product containing the modified C3970T PCR primer pair as template (Figure 3.19B) compared to the PCR product with the original PCR primer pair where amplification of this allele was rather minimal (Figure 3.19A). This confirmed that the failure of this allele to amplify in the past was caused by both inefficient PCR and SBE primers.

Whilst all alleles could now be successfully simultaneously detected in a single 12-Plex SNaPshot™ reaction with the modified SBE primers, noticeable differences in peak heights required further optimisation (Figure 3.19B). The majority of alleles tested at 0.4µM SBE primer concentration showed signals averaging about 1000RFU. The A1736G and the modified C10400T and T12705C SBE primers, which were at the lower concentration of 0.2µM, showed the strongest signals (2000-4000RFU). In order to increase the signals of weak loci, the PCR primer concentrations were further investigated since contrastingly to base extension reactions, PCR is an exponential rather than linear process. Thus, the greatest effect on allele signal is likely to be achieved by modifying PCR primer concentration, which influences the amount of template available for the single base extension of each allele.



Legend:

(A) Multiplex 1 11-plex PCR template was prepared using PCR conditions described in Section 3.3.1.5 and incorporating the original C3970T PCR primers. All SBE primers were tested at 0.4µM, except for C10873T, A1736G and the modified T12705C, C3970T and C10400T SBE primers= 0.2µM.

(B) Same PCR conditions as (A) but includes the modified C3970T PCR primers.

(C) Optimisation of PCR primer concentration was performed by reducing mtDNA PCR primer concentrations (except C3970T) to 0.1µM and increasing M9 to 0.5µM and M207 and M89 to 0.3µM. All SBE primers maintained at an equimolar concentration of 0.2µM.

(D) Optimisation of SBE primer concentrations using the same PCR product as in (C) as template. All SBE primers were maintained at a concentration of 0.2µM except for T12705C, A1736G and C10400T =0.1µM and M45, M9, M89, C8281G and C10238T= 0.4µM.

(E) The final optimised Multiplex 1 SNP assay derived from further PCR and SBE modifications to (D) with M201 and M89 PCR primers increased to 0.4µM and C8281G to 0.3µM. The M9 and C8281G SBE primers were also increased to 0.5µM, M201 to 0.4µM, M207 to 0.3µM and T12705C to 0.15µM.

(F) The 11-Plex PCR negative control with SBE primer concentrations used in (E).

Figure 3.19: Capillary electropherograms for the optimisation of Multiplex 1 12-Plex SNaPshot™ reactions incorporating the redesigned SBE primers and using Multiplex 1 PCR product as template.

Optimisations involved the modification of PCR and/or SBE primer concentrations as detailed in the legend.

As evident in Figure 3.15, the majority of mtDNA PCR products showed strong yields and demonstrated relatively stronger allele signals than the NRY SNPs (Figure 3.19B). The PCR primer concentration for the mtDNA loci A1736G, C10238T/ C10400T, C10873T and T12705C was reduced from 0.2μM to 0.1μM to test if the yields of NRY products would increase by reducing competition for reagents. The PCR primer concentrations were also increased for alleles showing the lowest signals, with primers for M9, increased from 0.4μM to 0.5μM and for M207 and M89 increased from 0.2μM to 0.3μM. These PCR primer concentration modifications tested using equimolar SBE primer concentrations of 0.2μM showed generalised signal increases across all alleles (Figure 3.19C), including for mitochondrial alleles for which the PCR primer concentrations had been reduced. However, fluorescence intensity decreases with the age of SNaPshot™ kit reagents and with increased freeze/thaw cycles. A new kit had been used in Figure 3.19C, which most likely contributed to the over amplified mtDNA alleles that showed evidence of pull up, such as A1736G (>4000RFU), even though PCR primer concentrations were reduced.

The effect of modifying both PCR and SBE primer concentrations was then tested on the same PCR product as was used in Figure 3.19C as template. The SBE primer concentration for T12705C, A1736G and C10400T was reduced from 0.2μM to 0.1μM, whilst M45, M9, M89, C8281G and C10238T were increased from 0.2μM to 0.4μM, with all other alleles maintained at a concentration of 0.2μM (Figure 3.19D). Reductions in SBE primer concentration diminished the signal of the previously over amplified A1736G, C10400T and T12705C alleles though the latter showed dramatic reduction in signal at a SBE concentration of 0.1μM which was not deemed suitable. Increases in both PCR and SBE primer concentration showed improvements in the signal of weaker alleles, including M45, M9, C10238T and C8281G. Other alleles which were not directly modified, such as C3970T and C10873T, also showed a slight increase in signal arising from the PCR and SBE primer modifications of other alleles in the Multiplex 1 set.

To further improve the low signal of alleles M207, M201, M89 and C8281G the PCR primer concentrations for M201 and M89 were increased to 0.4μM and for C8281G to 0.3μM. The SBE primer concentration was increased for M9 and C8281G to 0.5μM, M201 to 0.4μM and M207 to 0.3μM. The SBE primer concentration for T12705C was

also adjusted to an intermediate 0.15µM as 0.1µM had decreased the signal too dramatically, whilst 0.2µM caused over amplification. As shown in Figure 3.19E, these PCR and SBE primer modifications achieved fairly balanced amplification of all Multiplex 1 alleles, with all peaks averaging between 2000-3000RFU. The PCR and SBE primer concentrations used in Figure 3.19E were retained as the final optimised conditions for Multiplex 1 and were applied in the preliminary testing of population samples.

The presence of a noisy baseline between the 20nt to 30nt position was also commonly observed and did not appear to diminish at the conclusion of Multiplex 1 SNaPshot™ optimisation. It was found that the noise became more prominent when the Multiplex 1 PCR template was inadequately purified using older aliquots ExoSAP-It (USB Corporation, USA) or ExoStar (GE Healthcare, USA) (data not shown). The reagent activity was observed to dramatically decrease after more than 4 freeze/thaw cycles so it is recommended that the stock be stored in smaller aliquots to ensure optimum purification of the PCR templates and minimise noisy baselines, which can interfere with allele detection.

3.3.1.10 *Effect of polymer concentration on the electrophoretic mobility of Multiplex 1 SBE primers*

POP-6™ polymer was used for capillary electrophoresis instead of the POP-4™ polymer recommended by the manufacturers of the SNaPshot™ kit because better resolution and size discrimination of SBE primers, particularly for similar sized fragments, have been reported for POP-6™ polymer (Vallone *et al.*, 2004, Brion, 2005). The higher 6% polydimethylacrylamide concentration of POP-6™ polymer is typically applied for the separation of sequencing products, which differ by only 1 nucleotide, whilst the lower 4% polydimethylacrylamide concentration of POP-4™ polymer has been used for fragment analysis. Though the separation of SNaPshot™ extension products is typically classified as a form of fragment analysis, the difference in size between products is usually quite small and in some cases differs by only 1-2bp. It is therefore reasonable to propose that POP-6™ polymer can be applied instead of POP-4™ polymer. Though the higher polymer concentration will delay the migration of the fragments, usually requiring longer run times, no changes are required to the capillary

length or run module for products up to 65 nucleotides which can still be resolved during the 24 minute run time (Vallone *et al.*, 2004).

To confirm the suitability of POP-6™ polymer over POP-4™ polymer for our applications, final optimised Multiplex 1 extension products were separated by capillary electrophoresis using both POP-4™ and POP-6™ polymer and the run parameters specified by the SNaPshot™ GS POP-4 (1mL).E5.md5 module (Table 2.7). As shown in Figure 3.20, visible differences in the migration of alleles and in peak morphology were observed with POP-6™ polymer, providing greater spatial resolution and discrimination of alleles compared to POP-4™ polymer. This was most evident when comparing alleles M207, T12705C and M201, which overlapped in POP-4™ but could be discriminated in POP-6™. The peak width in POP-6™ polymer was also approximately half that observed in POP-4™ and contributed to the greater resolution achieved when using this higher polymer concentration.

When comparing the sizing of Multiplex 1 extension products in both polymers (Table 3.3), it was found that the most accurate sizing (defined by the expected size of the primer matching the observed size of the product) was achieved when using POP-6™ polymer, where the observed size of the multiplex alleles was only about 1 to 3 nucleotides larger than the expected size. In contrast, much greater differences were observed in POP-4™ polymer where the observed allele size was 3 to 5 nucleotides larger than the expected size particularly for alleles less than 35 nucleotides. However, it was generally found that the sizing accuracy improved as the length of the primer increased when using both POP-4™ and POP-6™ polymer, with the greatest mobility differences observed for primers less than 45 nucleotides. The spatial resolution and sizing accuracy differences observed between the POP-4™ and POP-6™ polymers for the Multiplex 1 SBE alleles paralleled those from previous studies in the literature (Vallone *et al.*, 2004, Brion, 2005). This confirmed that using POP-6™ polymer to separate SNaPshot™ extension products is more beneficial than POP-4™ polymer particularly for similar sized products, as the higher concentration improves the resolution and discrimination of alleles. Thus, this supported the decision to use POP-6™ polymer for the capillary electrophoresis of our multiplex assays.

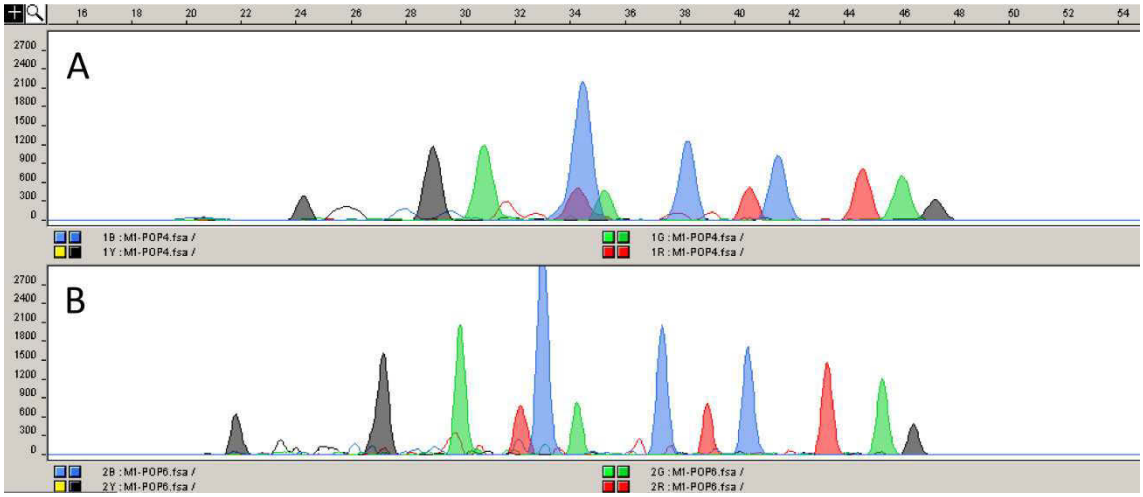


Figure 3.20: Capillary electropherograms of multiplex 1 SNaPshot™ reactions separated in (A) POP-4™ and (B) POP-6™ polymer.

Table 3.3: Comparison of expected and observed sizes of multiplex 1 SBE primers in POP-4™ and POP-6™ polymer.

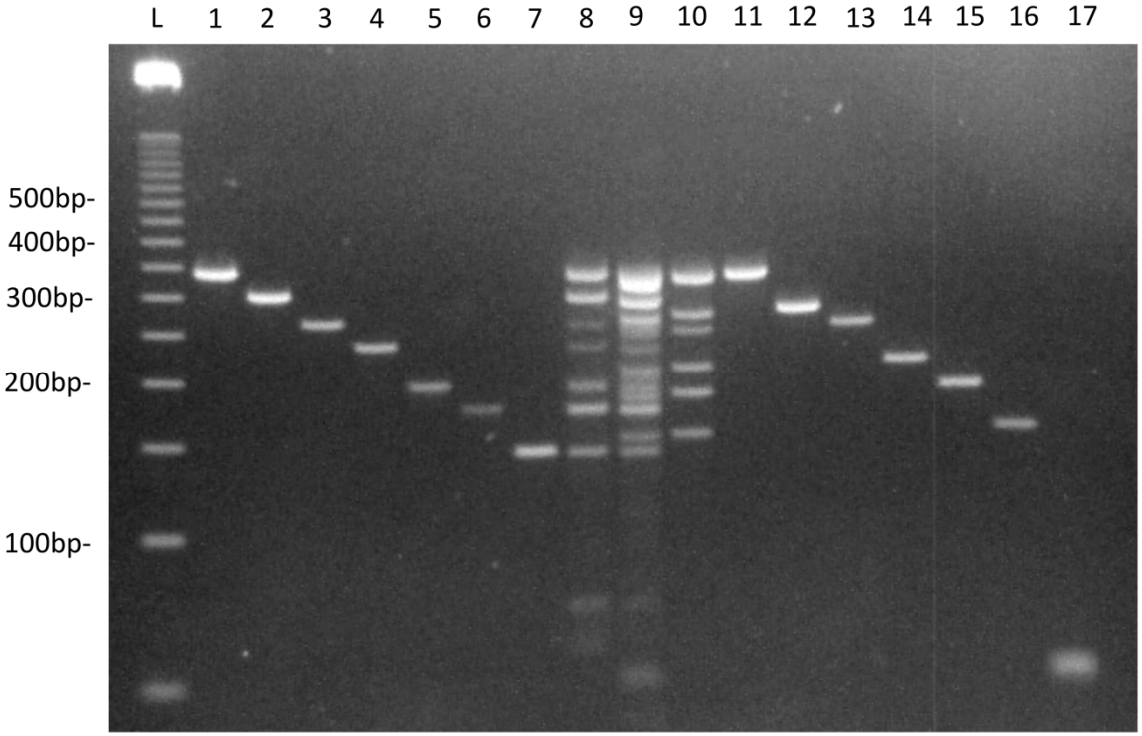
SNP	Expected Size (nt)	Observed Size (nt)	
		POP-4™	POP-6™
M45	19	24.26	21.63
M9	25	29.04	27.18
A1736G	28	30.83	29.98
M207	29	34.31	32.17
T12705C	31	34.51	32.99
M201	33	35.29	34.26
C10400T	36	38.31	37.39
M89	37	40.56	39.01
C3970T	39	41.59	40.49
C10873T	41	44.64	43.38
C10238T	43	46.10	45.37
C8281G	46	47.29	46.54

3.3.2 Multiplex 2 assay optimisation

3.3.2.1 *Preliminary optimisation of the Multiplex 2 PCR reaction*

As for Multiplex 1, each Multiplex 2 PCR primer pair was first amplified in singleplex PCR reactions and then grouped in 6-Plex (A12308G, A15607G, M180, M181, M217, M91), 7-Plex (T7028C, T4216C, M174, M130, M40, M38, M215) and 13-Plex PCR reactions. The optimised Multiplex 1 PCR conditions of 4mM $MgCl_2$, 0.4mM dNTP mix, 1.75U *Taq* polymerase and 2× PCR buffer were used. In the singleplex PCR reactions the NRY primers were tested at 0.2µM and the mtDNA primers at 0.1µM, whilst in the multiplex PCR reactions all primers were maintained at 0.2µM. As shown in Figure 3.21, all loci, both individually and in combination, were successfully amplified without major non-specific interactions. A total of twelve PCR products were detected in the 13-Plex reaction, with the T7028C (333bp) and A12308G (323bp) amplicons assumed to appear as a doublet band due to the 10bp size difference that could not be resolved in the 4% (w/v) agarose gel.

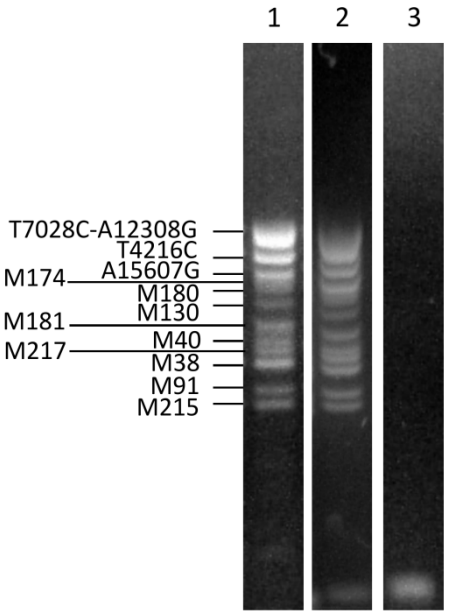
Minor differences in yield were observed between different Multiplex 2 PCR products thus PCR primer concentrations were optimised from a starting concentration of 0.2µM to achieve balanced amplification of all loci. The mtDNA loci that demonstrated the highest yields at 0.2µM were reduced to 0.1µM whilst less efficiently amplified products such as M174, M180 and M130 were increased to 0.3µM until relatively equalised Multiplex 2 PCR product yields were achieved (Figure 3.22). This optimised Multiplex 2 PCR product was then used as template for the optimisation of Multiplex 2 SBE reactions.



Legend:
L: 50bp DNA ladder
1: T7028C (333bp)
2: T4216C (297bp)
3: M174 (259bp)
4: M130 (230bp)
5: M40 (192bp)
6: M38 (171bp)
7: M215 (140bp)
8: 7-Plex: T7028C, T4216C, M174, M130, M40, M38, M215
9: 13-Plex: All multiplex 2 PCR primers
10: 6-Plex: A12308G, A15607G, M180, M181, M217, M91
11: A12308G (323bp)
12: A15607G (270bp)
13: M180 (249bp)
14: M181 (208bp)
15: M217 (181bp)
16: M91 (150bp)
17: 13-plex PCR negative control

Figure 3.21: Multiplex 2 singleplex and multiplex PCR products.

Multiplex 2 PCR primer pairs were amplified at equimolar concentrations of 0.2µM in 6-Plex, 7-Plex and 13-Plex reactions using optimised Multiplex 1 PCR conditions summarised in section 3.3.1.4. The PCR primers for the NRY loci (M174, M130, M40, M38, M215, M180, M181, M217, M91) were also amplified individually at 0.2µM and for the mtDNA loci (T7028C, T4216C, A12308G, A15607G) at 0.1µM in singleplex.



Legend:
1: All PCR primers=0.2µM
2: All PCR primers= 0.2µM except T4216C, T7028C, A12308G and C15607G= 0.1µM and M174, M180 and M130= 0.3µM
3: 13-plex PCR negative control

Figure 3.22: Optimisation of Multiplex 2 PCR primer concentrations

3.3.2.2 *Optimisation of the multiplex 2 SNaPshot™ reaction*

Use of the SNaPshot® Primer Focus™ kit (Life Technologies, USA) as described in section 2.2.1.3 confirmed that the Multiplex 2 SBE primers provided sufficient spatial separation with no alleles overlapping in both size and colour upon separation in POP-6™ polymer (data not shown). Optimisation of the Multiplex 2 SNaPshot™ assay involved the preliminary extension of Multiplex 2 SBE primers at 0.2µM in singleplex SNaPshot™ reactions using the optimised Multiplex 2 PCR product described in section 3.3.2.1. All thirteen alleles of expected size and colour were detected (Figure 3.23) confirming the specificity of the SBE primers. The allele signals averaged 2000-3000RFU with the exception of A12308G, which showed over amplification and pull-up, and of M180 and M130, which had the lowest signals at 1000-1500RFU. Overall, singleplex amplification of the Multiplex 2 SBE primers achieved generally balanced signals and was less problematic than for Multiplex 1.

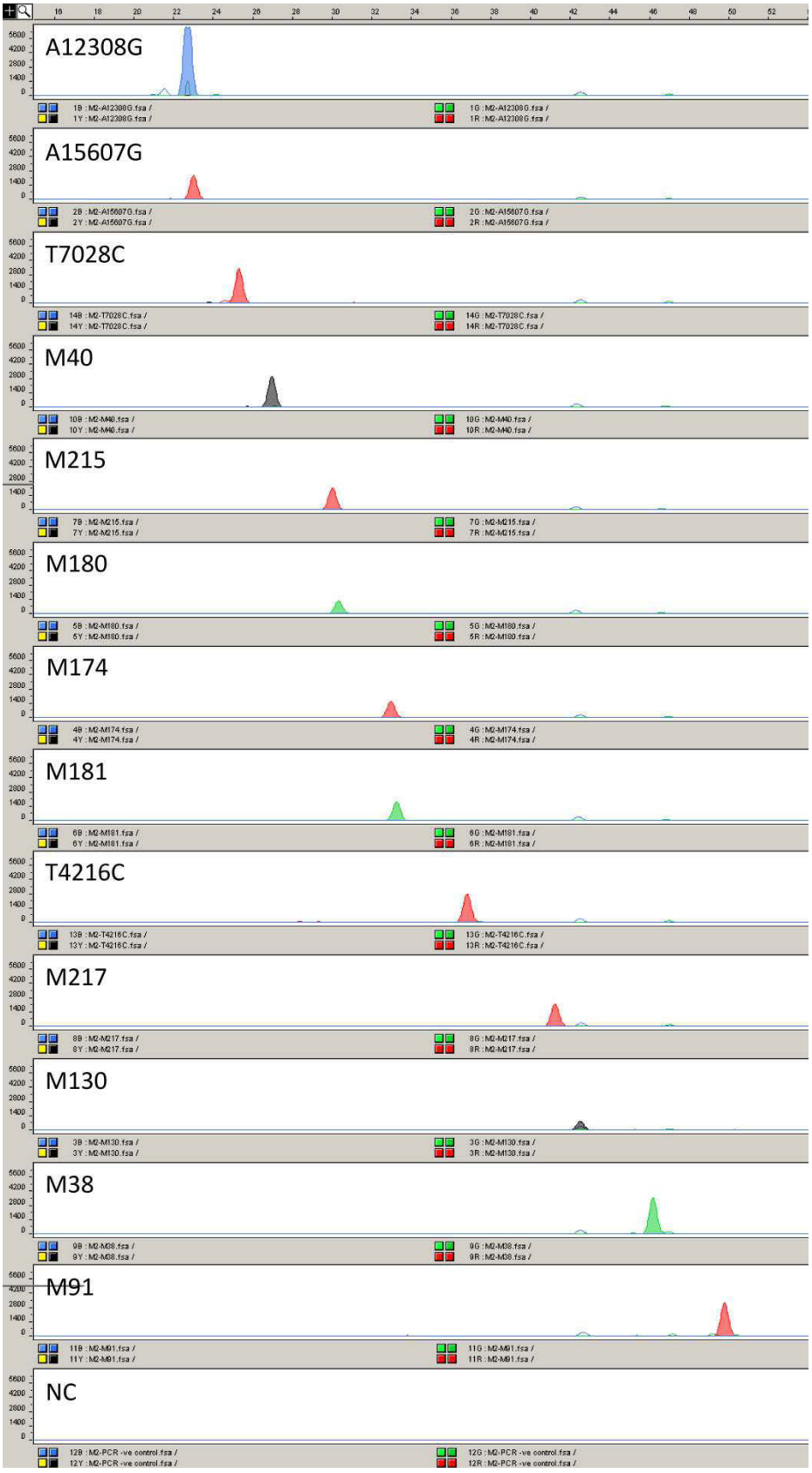
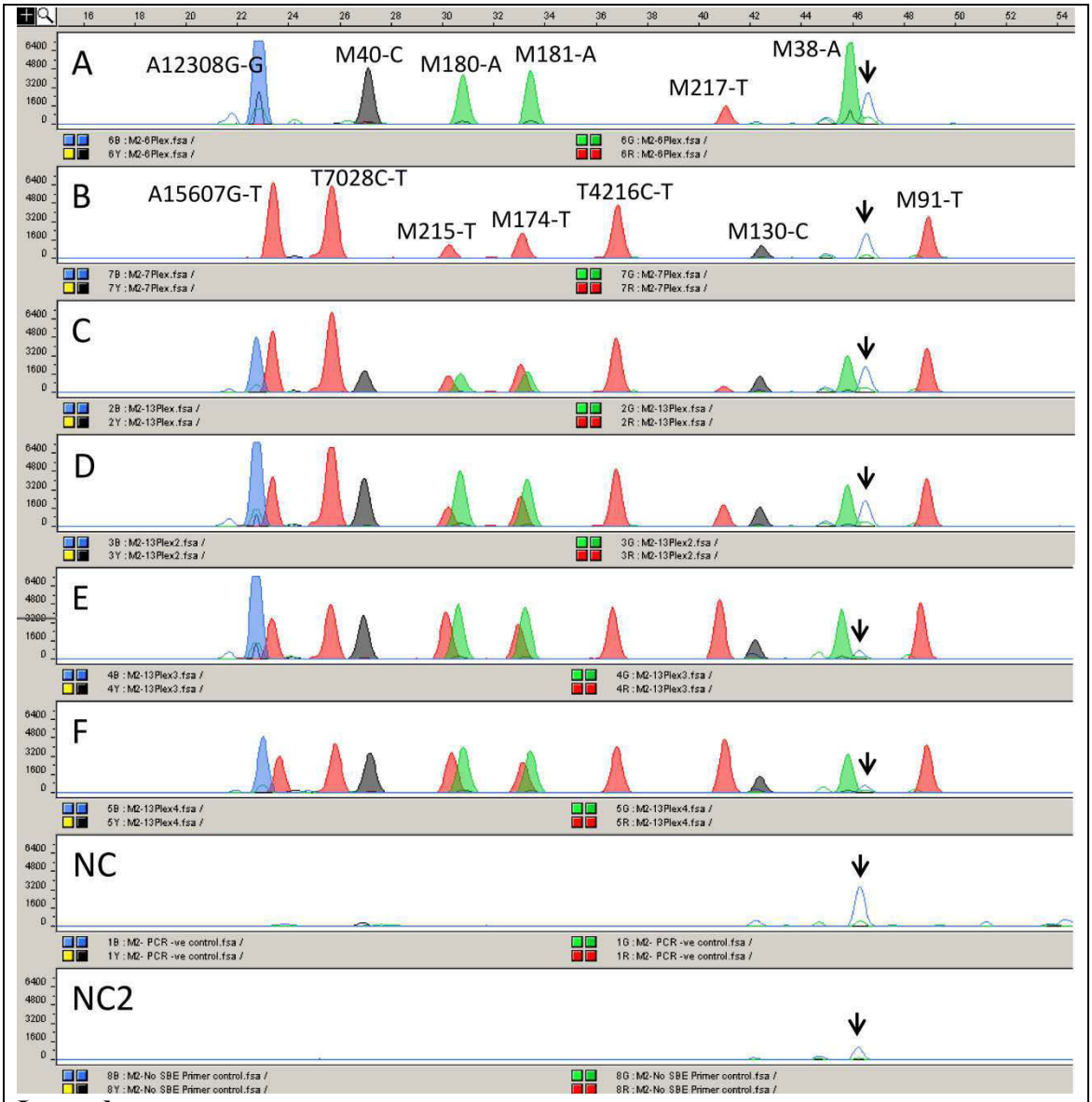


Figure 3.23: Capillary electropherograms of singleplex extensions of 0.2µM multiplex 2 SBE primers using optimised Multiplex 2 PCR product as template.

Multiplex 2 PCR product was prepared using the optimised conditions described in section 3.3.2.1. The 13-Plex PCR negative control is also shown (**NC**).

The SBE primers were then tested in 6-Plex (A12308G, M40, M180, M181, M217, M38), 7-Plex (A15607G, T7028C, M215, M174, T4216C, M130, M91) and 13-Plex SNaPshot™ reactions at an equimolar concentration of 0.2µM, using the same optimised multiplex 2 PCR product as template. As shown in Figure 3.24A-C, all expected alleles were detected in each of these multiplex reactions, though marked differences in signal were observed between loci, with the stronger alleles detected at greater than 3000RFU and the weaker alleles averaging 1000-2000RFU. The signal of alleles decreased between the smaller multiplexes (Figure 3.24A-B) and the 13-Plex (Figure 3.24C), most likely due to competition between greater numbers of SBE primers in the larger multiplex. To achieve more balanced allele signals across all loci, SBE primer concentrations of the weaker loci M215, M180, M217 and M130 were increased to 0.4µM and of M174 to 0.3µM. Though A15607G and M38 were not overextended in the 13-Plex depicted in Figure 3.24C, these alleles showed strong signals in the 6-Plex (Figure 3.24A). Thus, the SBE primer concentration for these loci was decreased to 0.1µM to observe if this would have a positive effect on the other loci. As shown in Figure 3.24D, the reduction in A15607G and M38 SBE primer concentrations had minimal effect on their individual signals, but increased signals were observed for other unmodified alleles such as A12308G, T7028C and M40, indicating that the A15607G and M38 primers were likely in excess and competing with these primers. Weaker alleles for which the SBE primer concentrations were increased also showed improved signals, with the exception of M215.

Further optimisation experiments were carried out to test the effect of increasing the PCR primer concentration for M180, M174 and M130 alleles to 0.4µM and for M217 allele to 0.3µM, whilst maintaining the SBE primer concentrations used in Figure 3.24D. These changes resulted in relatively balanced extension of all loci to 3000-4000RFU, with the greatest improvement observed for M217 (Figure 3.24E). Additionally, though M215 PCR primer concentrations were not altered, the signal for this allele also improved under these modified PCR and SBE conditions. However as the A12308G allele still appeared slightly over extended in Figure 3.24E, the SBE primer concentration for this allele was further decreased to 0.1µM to finalise assay optimisation. The final optimised multiplex 2 assay is depicted in Figure 3.24F and the optimised PCR and SBE conditions used for this reaction were applied in the preliminary testing of population samples.



Legend:

(A), (B) and (C) 6-Plex, 7-Plex and 13-Plex SBE reactions respectively, using Multiplex 2 13-plex PCR template shown in Figure 3.22 (Lane 2). All SBE primers were extended at equimolar SBE primer concentrations of 0.2µM.

(D) Same PCR product as (C) but the SBE primer concentrations were modified with M215, M180, M217, M130 =0.4µM, M174= 0.3µM and A15607G, M38=0.1µM and all other primers at 0.2µM.

(E) The PCR primer concentrations were modified from (D) with M180, M174, M130=0.4µM and M217=0.3µM. The SBE primer concentrations were as in (D).

(F) Final optimised Multiplex 2 SNP assay using the same conditions as (E) with minor reduction of A12308G SBE primer to 0.1µM.

(NC) 13-plex PCR negative control with SBE primer concentrations used in (F).

(NC2) 13-plex PCR negative control with no SBE primers confirms the non-specific extension of multiplex PCR products visualised as a 46.5nt 'G' peak (black arrow).

Figure 3.24: Capillary electropherograms for the optimisation of Multiplex 2 SNaPshot™ reactions using Multiplex 2 PCR product.

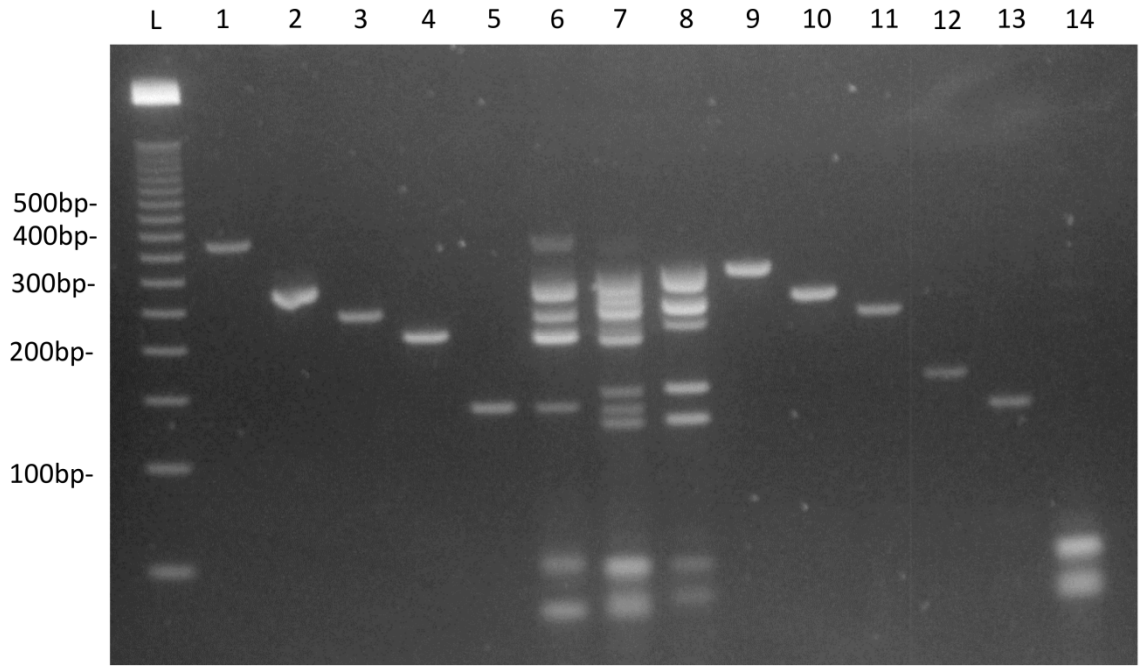
Optimisations involved the modification of PCR and/or SBE primer concentrations as detailed in the legend.

A blue artefact peak appearing at the 46.5nt position (denoted by a black arrow) was observed in all multiplex reactions (Figure 3.24A-F). The presence of this peak in the PCR negative control (Figure 3.24, NC) indicated that the artefact was likely to be the product of non-specific SBE primer interactions or arising from the PCR template. To investigate this, another negative control was tested which included Multiplex 2 PCR template as well as all SNaPshot™ reagents, with the exception of the SBE primers. The blue artefact peak was also detected in this negative control (Figure 3.24, NC2) and indicated the likelihood that non-specific extension of multiplex 2 PCR products or undigested PCR primers was occurring. However, though close to the M38 allele, the signal for this artefact peak decreased following optimisation of the Multiplex 2 SNaPshot™ reaction and the colour did not interfere with the detection of the alternative C allele (black) at the M38 locus. Thus, no further attempt was made to eliminate it.

3.3.3 Multiplex 3 assay optimisation

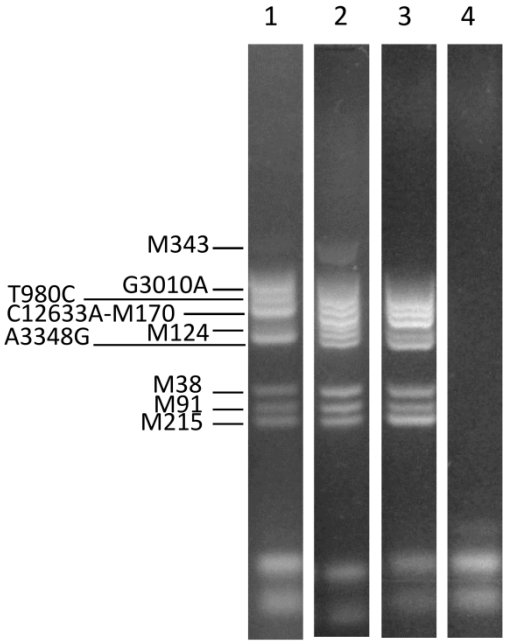
3.3.3.1 *Preliminary optimisation of the Multiplex 3 PCR reaction*

Initially, Multiplex 3 PCR primers were amplified in singleplex PCR reactions and then grouped in two 5-Plex PCR reactions, 5-Plex1 (M343, T980C, M170, A3348G, M69) and 5-Plex2 (G3010A, C12633A, M124, M172, SRY10831.2) and then combined in a single 10-Plex reaction. The reactions were prepared using the same reagent and PCR primer concentrations as those applied for the Multiplex 2 singleplex PCR reactions (section 3.3.2.1). All loci could be individually amplified and showed the expected sized product (Figure 3.25). All expected PCR products were also visualised in each of the 5-Plex reactions, however only nine bands were observed in the 10-Plex reaction, with C12633A (245bp) and M170 (233bp) appearing as a doublet band that could not be resolved in the 4% (w/v) agarose gel due to the 12bp difference (Figure 3.25).



Legend:
L: 50bp DNA ladder
1: M343 (355bp)
2: T980C (267bp)
3: M170 (233bp)
4: A3348G (204bp)
5: M69 (136bp)
6: 5-Plex1: M343, T980C, M170, A3348G, M69
7: 10-plex: All Multiplex 3 PCR primers
8: 5-Plex2: G3010A, C12633A, M124, M172, SRY10831.2
9: G3010A (287bp)
10: C12633A (245bp)
11: M124 (218bp)
12: M172 (151bp)
13: SRY10831.2 (125bp)
14: 10-plex PCR negative control

Figure 3.25: Multiplex 3 singleplex and multiplex PCR products.
Multiplex 3 PCR primer pairs were amplified at equimolar concentrations of 0.2µM in 5-Plex and 10-Plex reactions using optimised Multiplex 1 PCR conditions summarised in section 3.3.1.4. The PCR primers for the NRY loci (M343, M170, M69, M124, M172, SRY10831.2) were also amplified individually at 0.2µM and for the mtDNA loci (T980C, A3348G, G3010A, C12633A) at 0.1µM in singleplex.



Legend:
1: All PCR primers=0.2µM
2: All PCR primers= 0.2µM except T980C, A3348G, G3010A, C12633A= 0.1µM and M124, M343=0.3µM
3: Same as lane 2 except M170=0.3µM
4: 13-plex PCR negative control

Figure 3.26: Optimisation of Multiplex 3 PCR primer concentrations.

Preliminary optimisation of PCR primer concentrations from a starting concentration of 0.2µM was required to balance the visible PCR products yield differences in the 10-Plex, particularly for the M343 product (355bp) which was almost absent and the poorly amplifying M124 (218bp) loci. The PCR primer concentration for these loci was increased to 0.3µM, whilst the PCR primer concentration for the high yield mtDNA loci was reduced to 0.1µM. As shown in Lane 2 of Figure 3.26, relatively balanced yields were obtained using these concentrations, with a visible increase in the M343 yield. The yield for M170 appeared lower using these conditions so an increased PCR primer concentration of 0.3µM was also tested (Figure 3.26, Lane 3). However, this PCR primer modification had a negative effect on M343, which showed a similarly reduced yield to when the M170 PCR primer concentration was 0.2µM. Consequently, the PCR conditions used in Figure 3.26 (Lane 2) appeared to be the most suitable template for the optimisation of Multiplex 3 SBE reactions.

3.3.3.2 *Optimisation of the Multiplex 3 SNaPshot™ reaction*

Use of the SNaPshot® Primer Focus™ kit (Life Technologies, USA) as described in section 2.2.1.3 confirmed that all Multiplex 3 SBE primers could be resolved and unambiguously detected with no overlapping peaks (data not shown). Following the same optimisation process as for Multiplexes 1 and 2, the Multiplex 3 SBE primers were first extended at a concentration of 0.2µM in singleplex SNaPshot™ reactions using the Multiplex 3 PCR product shown in Figure 3.26 (Lane 2) as template. As shown in Figure 3.27, all alleles were detected at the expected size and colour and showed strong signals ranging from 3000-6000RFU, with the exception of M343, which appeared much weaker at 500RFU. This was likely caused by the lower yield of the M343 PCR product compared to the other products in the multiplex even though PCR primer concentration for this locus was increased to 0.3µM during preliminary PCR optimisation (Figure 3.26, Lane 3).

Further testing of the Multiplex 3 SBE primers at an equimolar concentration of 0.2µM in separate 5-Plex and 10-Plex SNaPshot™ reactions showed successful amplification of all alleles (Figure 3.28A-C). As anticipated, allele signals were slightly decreased compared to those achieved in the singleplex SNaPshot™ reactions due to the increased number of competing primer sets. The minor signal imbalances observed between alleles were optimised by first modifying PCR primer concentrations. Since the preliminary Multiplex 3 PCR optimisations showed that increasing the PCR primer concentration for M170 had a negative effect on the yield of the M343 product (Figure 3.26, Lane 3), the PCR product in Lane 2 with M170 at 0.2µM was used as template for 10-Plex SNaPshot™ reactions with SBE primers at an equimolar concentration of 0.2µM. Major improvements were observed in the signal of all alleles, especially M343, under these modified PCR conditions (Figure 3.28D). However, over amplification of M124 was observed, so the PCR primer concentration for this locus was also reduced back to 0.2µM. Unfortunately, when using an equimolar SBE primer concentration of 0.2µM this PCR primer reduction had little effect on the M124 signal which remained over amplified (Figure 3.28E).

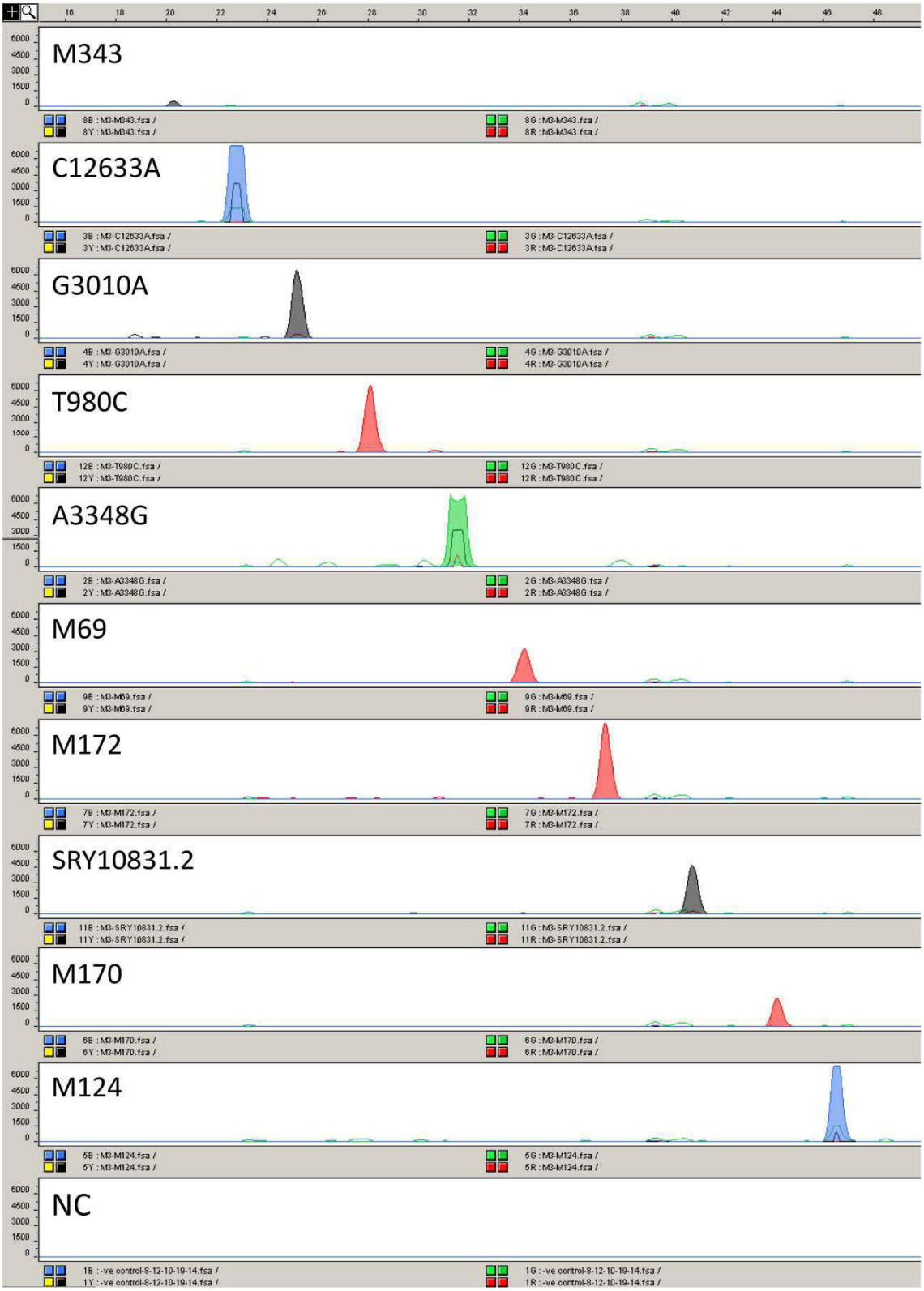
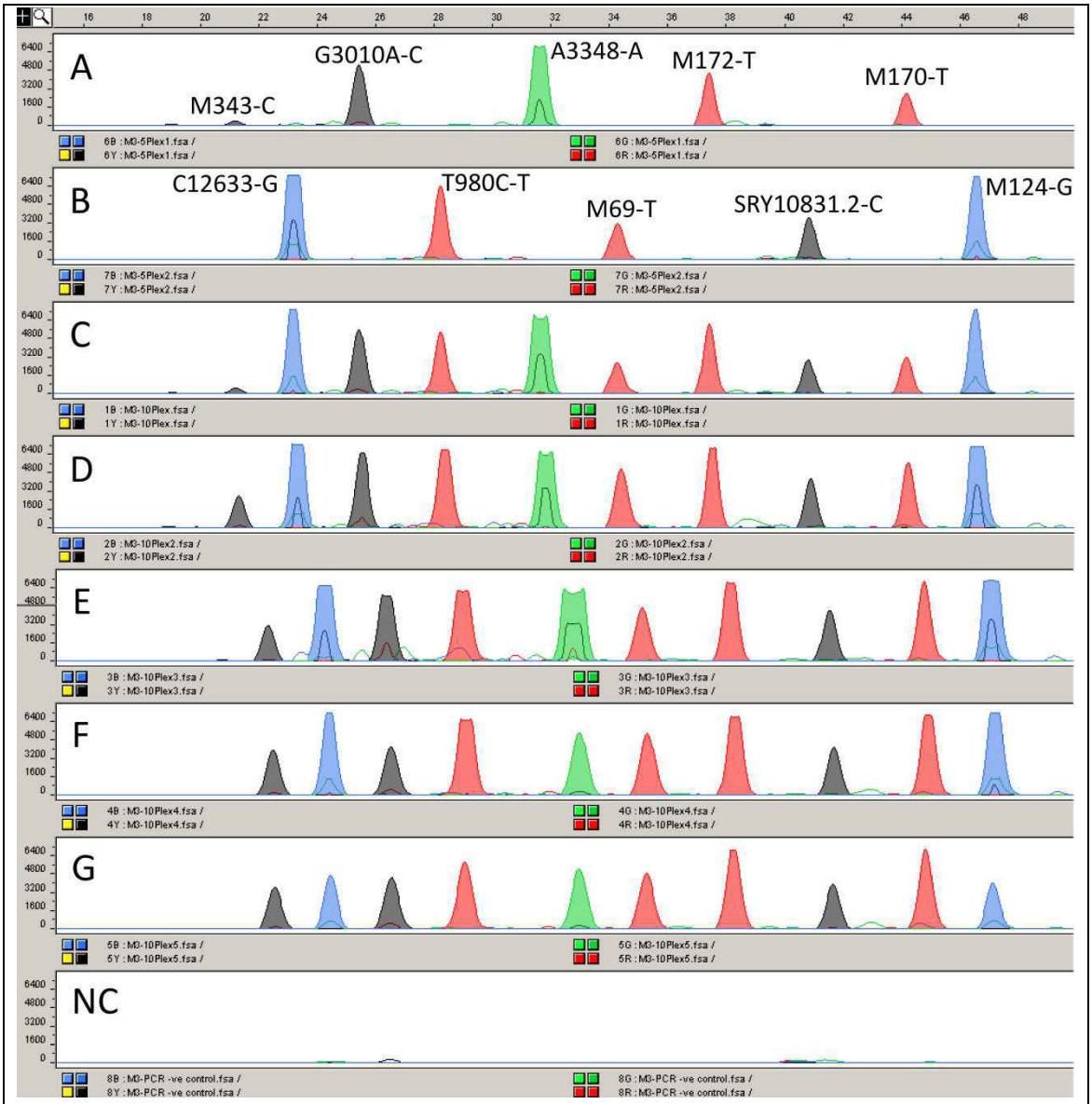


Figure 3.27: Electropherograms of the singleplex amplification of 0.2µM multiplex 3 SBE primers using optimised Multiplex 3 PCR product as template.

The multiplex 3 PCR product depicted in Figure 3.26 (Lane 3) was used as template. The 13-Plex PCR negative control is also shown (**NC**).



Legend:

(A), (B) and (C) 5-Plex1, 5-Plex2 and 10-Plex SBE reactions respectively, prepared using Multiplex 3 10-plex PCR template shown in Figure 3.26 (Lane 3). All SBE primers were extended at equimolar SBE primer concentrations of 0.2µM.

(D) 10-Plex SBE reaction prepared using Multiplex 3 10-plex PCR template shown in Figure 3.26 (Lane 2). All SBE primers= 0.2µM.

(E) Same PCR primer concentrations as (D) except M124= 0.2µM. All SBE primers= 0.2µM.

(F) The same PCR primer product as (E) but SBE primer concentrations were modified with C12633A, G3010A, M124 = 0.1µM, A13348G = 0.075µM and all others= 0.2µM

(G) Final optimised Multiplex 3 SNP assay with the same conditions as (F) except SBE primer concentrations were modified with T980C= 0.1µM, M172, M170= 0.15µM and M124= 0.05µM

(NC) 13-plex PCR negative control with SBE primer concentrations used in (G).

Figure 3.28: Capillary electropherograms for the optimisation of Multiplex 3 SNaPshot™ reactions using Multiplex 3 PCR product as template.

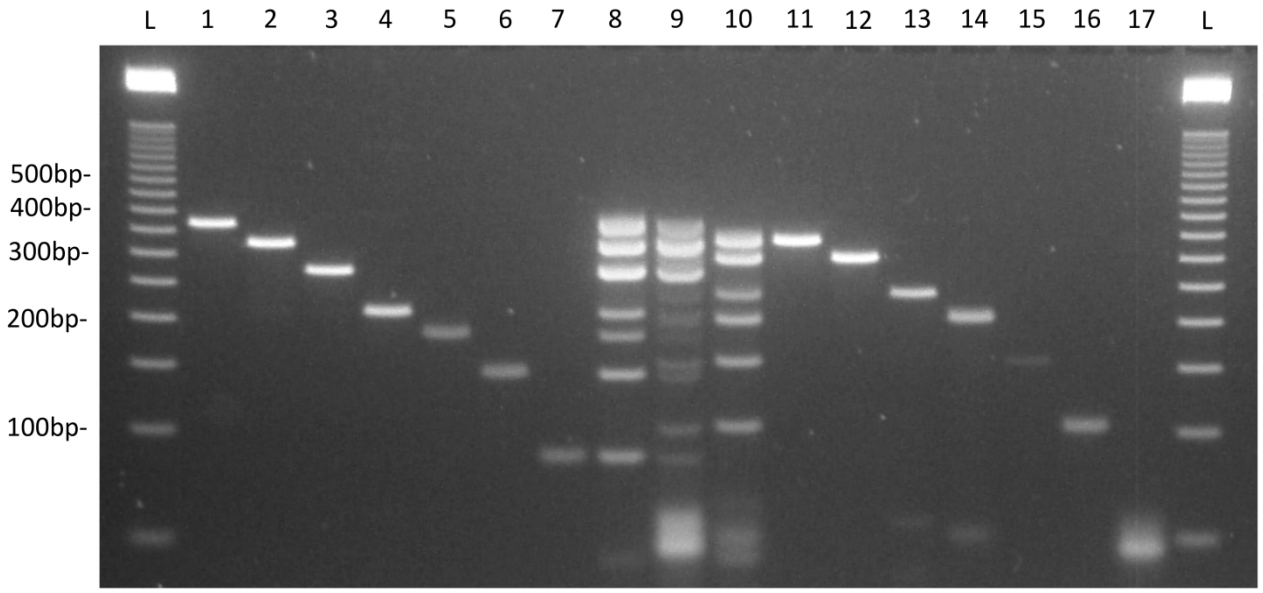
Optimisations involved the modification of PCR and/or SBE primer concentrations as detailed in the legend.

To further optimise the Multiplex 3 SNaPshot™ reaction, the PCR conditions used in Figure 3.28E were maintained and tested on a series of SBE primer concentration modifications. The SBE primer concentration for over amplified alleles C12633A, G3010A and M124 was halved to 0.1µM and A13348G was slightly reduced to 0.075µM. The SBE primer concentration for the weaker M343 allele was also increased to 0.3µM. These adjustments were able to achieve mostly balanced signals, though some alleles remained high and showed pull up and split peaks (Figure 3.28F). The SBE primer concentration for T980C was then halved to 0.1µM, M172 and M170 were slightly reduced to 0.15µM and M124 was reduced further to 0.05µM. As depicted in Figure 3.28G, these conditions finalised Multiplex 3 SNaPshot™ reaction optimisation, as balanced amplification of all alleles within a signal range of 3000-4000RFU was achieved. These final PCR and SBE concentrations were applied in the testing of initial population samples.

3.3.4 Multiplex 4 assay optimisation

3.3.4.1 *Preliminary optimisation of the Multiplex 4 PCR reaction*

Multiplex 4 PCR primers were initially tested in singleplex reactions using the same reagent and PCR primer concentrations used for Multiplex 2 described in Section 3.3.2.1. As shown in Figure 3.29, all Multiplex 4 PCR products could be individually amplified and were the expected size when visualised in a 4% (w/v) agarose gel. The PCR primers were then tested in 6-Plex (G4491A, T9824C, M175, M147, M11 9, P256), 7-Plex (C4140T, A4715G/4883T, C12403T, M70, P31, M22, M122) and 13-Plex PCR reactions at equimolar concentrations of 0.2µM. All expected products were visualised in the 5- or 6-Plex multiplexes but only twelve product bands were visible in the 13-Plex, presumably due to the co-migration of the G4491A and A4715G/C4883T products, which are only separated by a 14bp difference and could not be resolved in the 4% (w/v) agarose gel. The results indicated that all primers could work together under the conditions, in the absence of detrimental non-specific interactions (Figure 3.29, Lane 8-10). However as observed for multiplexes 2 and 3, there were noticeable differences in product yields with the mtDNA loci amplifying much more efficiently than the NRY loci.



Legend:
L: 50bp DNA ladder
1: C4140T (362bp)
2: A4715G/C4883T (312bp)
3: C12403T (265bp)
4: M70 (208bp)
5: P31 (179bp)
6: M22 (141bp)
7: M122 (82bp)
8: 7-Plex: C4140T, A4715G/4883T, C12403T, M70, P31, M22, M122
9: 13-plex: All Multiplex 4 PCR primers
10: 6-Plex: G4491A, T9824C, M175, M147, M119, P256
11: G4491A (326bp)
12: T9824C (290bp)
13: M175 (226-231bp)
14: M147 (197-198bp)
15: M119 (153bp)
16: P256 (99bp)
17: 10-plex PCR negative controls

Figure 3.29: Multiplex 4 singleplex and multiplex PCR products.

Multiplex 4 PCR primer pairs were amplified at equimolar concentrations of 0.2µM in 6-Plex, 7-Plex and 13-Plex reactions using optimised Multiplex 1 PCR conditions summarised in section 3.3.1.4. The PCR primers for the NRY loci (M70, P31, M22, M122, M175, M147, M119, P256) were also amplified individually at 0.2µM and for the mtDNA loci (C4140T, A4715G/C4883T, C12403T, G4491A, T9824C) at 0.1µM in singleplex.

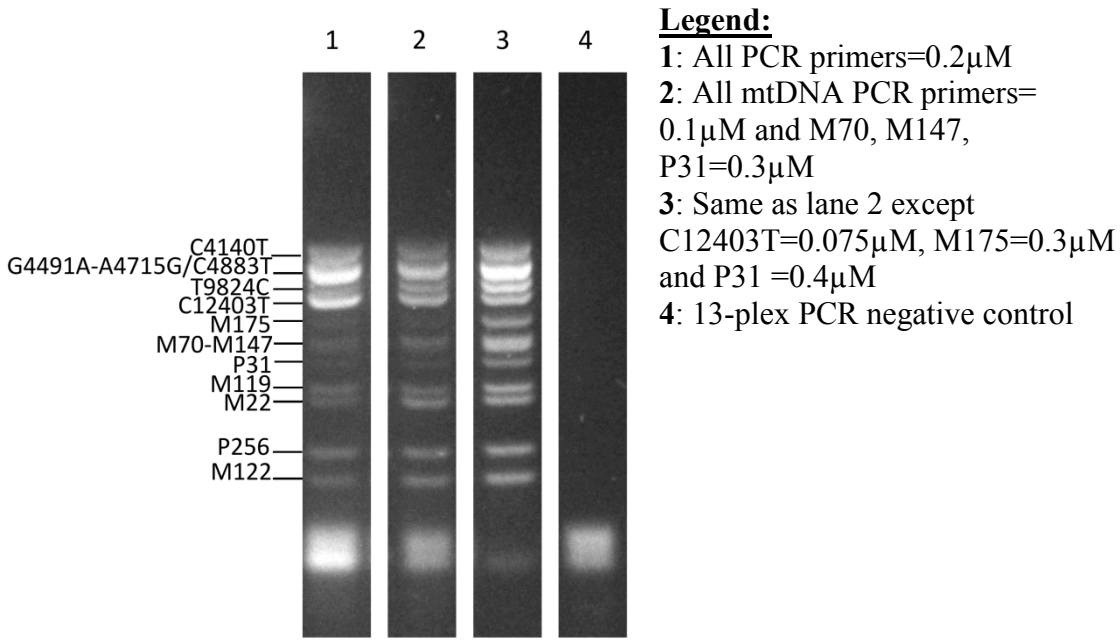


Figure 3.30: Optimisation of Multiplex 4 PCR primer concentrations.

To achieve more balanced amplification the PCR primer concentration for the stronger mtDNA loci (C4140T, G4491A, A4715G/C4883T, T9824C and C12403T) were reduced to 0.1µM, whilst those for the weaker amplifying M70, M147 and P31 loci were increased to 0.3µM. These modifications successfully reduced the yield of the mtDNA loci, although C12403T still demonstrated a higher yield whilst the yield for P31 only showed minor improvement (Figure 3.30, Lane 2). Further decrease of C12403T primer concentration to 0.075µM and increase in M175 and P31 primer concentrations to 0.3µM and 0.4µM respectively, resulted in an optimised preliminary Multiplex 4 PCR product which showed relatively balanced amplification of all PCR products (Figure 3.30, lane 3).

3.3.4.2 *Optimisation of the multiplex 4 SNaPshot™ reaction*

Use of the SNaPshot® Primer Focus™ kit prior to commencing Multiplex 4 SNaPshot™ optimisation had demonstrated adequate spatial separation of all Multiplex 4 SBE

primers except M22 and M175 in POP-6™ polymer (data not shown). The M22 and M175 loci share a 'T' allele that is detected at the 33.80nt and 34.70nt position respectively. A separation of only one nucleotide between the same coloured peaks of neighbouring loci is not recommended, however it was decided to examine the mobility of Multiplex 4 SBE primers extended in the 14-Plex SNaPshot™ reaction prior to any decision to alter the size of these primers.

The optimised Multiplex 4 PCR product depicted in Figure 3.30 (Lane 3) was used as starting template for the optimisation of the Multiplex 4 SNaPshot™ reaction. The SBE primers were first extended individually at a concentration of 0.2μM in singleplex SNaPshot™ reactions. All alleles were found to be the correct size and colour within a satisfactory signal range of 2000-3000RFU (Figure 3.31). Multiplex SNaPshot™ reactions were then tested by grouping primers in two separate 7-Plex reactions and a single 14-Plex reaction with all SBE primers at a concentration of 0.2μM. All alleles were successfully detected in each of these multiplexes, in the absence of problematic non-specific primer interactions (Figure 3.32A-C). The M22 and M175 'T' alleles, which showed only one nucleotide separation with Primer Focus™ appeared merged at the base but could both be resolved as separate peaks. This highlighted that altering the size of these primers was not necessary.

Minor signal differences were observed between alleles within the 14-Plex reaction requiring only increase of SBE concentration to 0.3μM for the slightly less efficiently extending loci M175, C4140T and M147 to produce relatively balanced signals across all alleles (Figure 3.32D). The peak heights achieved with these optimised SBE primer concentrations only ranged between 900-1500RFU, which was lower than that observed for other multiplexes. However the peak heights observed in Figure 3.32D were within the recommended 1000-2000RFU range and increased when a new SNaPshot™ kit was used (data not shown). Consequently, the PCR and SBE primer concentrations of Figure 3.32D were selected as the final optimised conditions for the Multiplex 4 SNaPshot™ reaction and were applied during initial population testing.

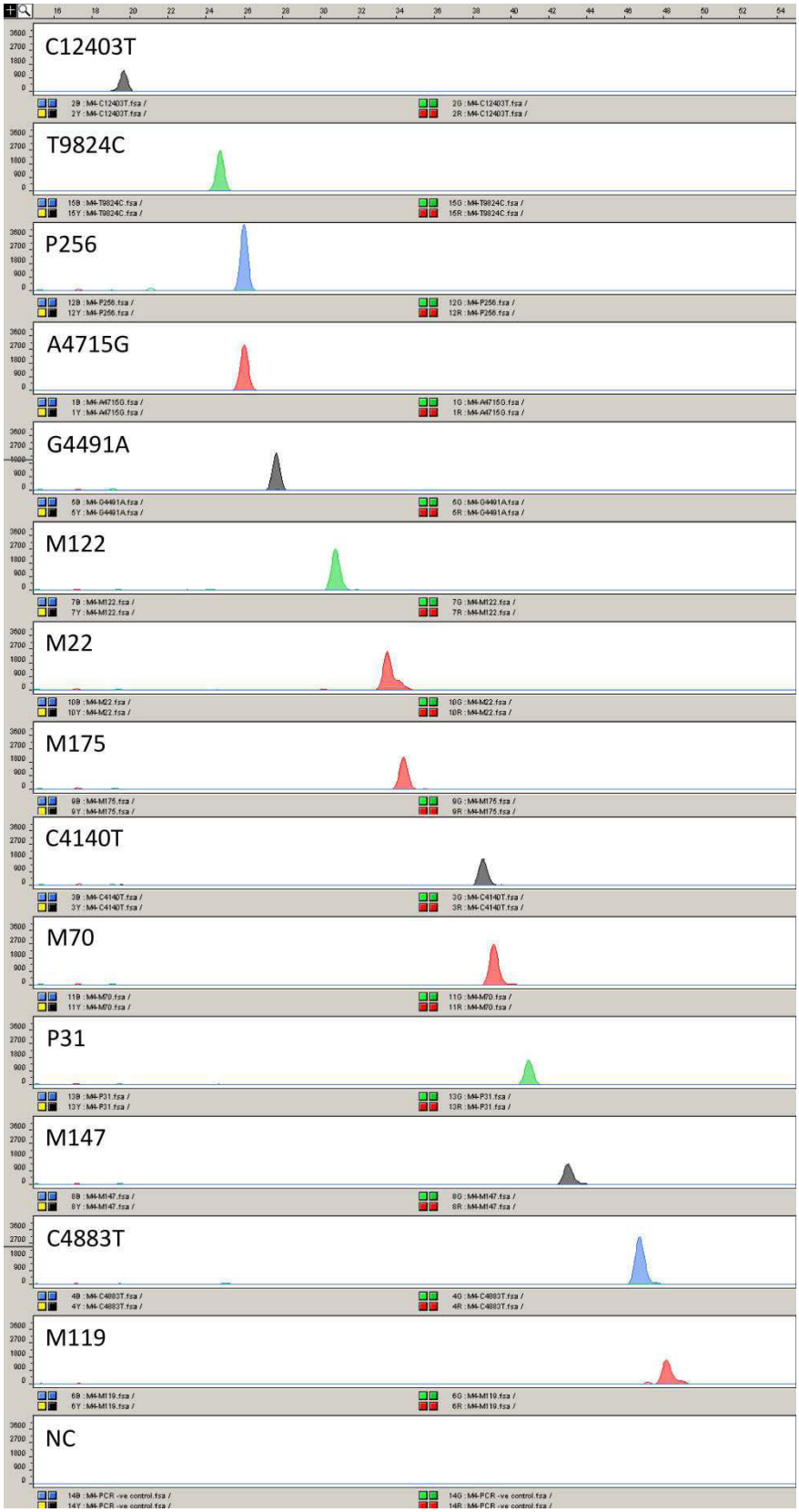
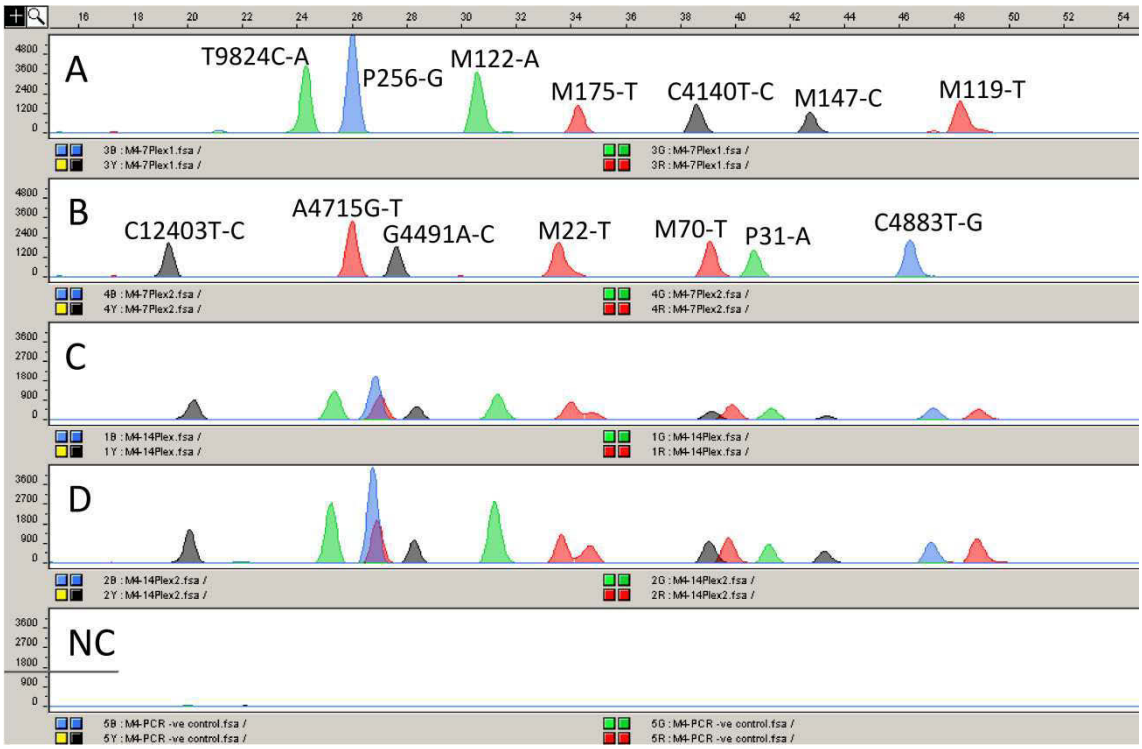


Figure 3.31: Capillary electropherograms of singleplex extensions of 0.2µM Multiplex 4 SBE primers using optimised Multiplex 4 PCR product as template.

The optimised Multiplex 4 PCR product depicted in Figure 3.29 (Lane 3) was used as template. The 13-Plex PCR negative control is also shown (**NC**).



Legend:
(A), (B) and (C) 7-Plex1, 7-Plex2 and 14-Plex SBE reactions respectively, prepared using Multiplex 4 13-plex PCR template shown in Figure 3.29 (Lane 3). All SBE primers= 0.2µM.
(D) Same conditions as (C) except that SBE primer concentrations M175, C4140T and M147 were increased to 0.3µM. All other SBE primers maintained at 0.2µM.
(NC) 14-plex PCR negative control with the SBE primer concentrations used in (D).

Figure 3.32: Capillary electropherograms for the optimisation of Multiplex 4 SNaPshot™ reactions using Multiplex 4 PCR product as template.
Optimisations involved the modification of SBE primer concentrations as detailed in the legend.

3.3.5 Initial population testing

Prior to performing large-scale genotyping of population samples, the optimised haploid SNP assays were tested on five DNA samples originating from donors with different declared biogeographical ancestries. One donor with the same declared maternal and paternal grandparental BGA was selected for each population group, which included individuals of English, Chinese, Cameroonian, Egyptian and Indian origin. The final optimised PCR and SBE primer concentrations for Multiplexes 1 to 4 are summarised in

Table 3.4 and were used to genotype these population samples. The haploid SNP multiplex assays developed in this chapter follow a hierarchical design, which does not normally require the samples to be tested using all four multiplexes to obtain the final genotype. However for quality control purposes, the five population samples were genotyped using each of the Multiplex 1 to 4 SNP assays. This was done to ensure that there is no ambiguity in the genotypes across the four multiplexes for the haplogroups detected. Preliminary population testing is also important to monitor any variations in peak heights between alternate alleles at different locus that were not previously detected in the sample used to optimise the assays.

The electropherograms of the Multiplex 1 to 4 profiles for the five population samples tested are depicted in Figure 3.33 and Figure 3.34. Variations in peak heights were observed between the samples due to differences in the amount of starting DNA template contained on the sampled discs. Despite this, all alleles were detected in each of the multiplexes, including the alternate alleles not previously visualised in the DNA sample used during the optimisation experiments. The only exception was the Chinese DNA sample for which the P31 allele failed to amplify to a detectable level in multiplex 4 (Figure 3.34B). This sample also demonstrated lower peak heights in the other multiplexes, which suggested that the FTA® card for this sample contained less DNA than the others. Overall, the peak heights within each multiplex also remained relatively equalised regardless of these variations in template DNA concentration and no major differences were observed between the peak heights of two alternate alleles at a given loci.

There were no ambiguities in the Multiplex 1 to 4 profiles of all five individuals, which indicated that the hierarchical strategy was effective. For example, considering the Cameroonian sample, the individual was deduced to have mtDNA haplogroup L and NRY haplogroup A-E from Multiplex 1 (Figure 3.33C). No further testing was required for mtDNA haplogroup assignation but further testing using Multiplex 2 was necessary to identify the final NRY haplogroup (refer to Figure 3.3), which was deduced to be E1b1a (Figure 3.33C). The Multiplex 3 and 4 profiles showed no other mtDNA or NRY SNP alleles, only ancestral alleles (Figure 3.34C). This result was expected since the haplogroups were already previously identified from the other multiplexes. The mtDNA and NRY haplogroup status of the five individuals genotyped using the optimised

Table 3.4: Final optimised PCR and SBE primer concentrations for Multiplex 1- 4

	SNP	PCR primer concentration (µM)	SBE primer concentration (µM)
MULTIPLEX 1	M45	0.2	0.4
	M9	0.5	0.5
	M207	0.3	0.3
	M201	0.4	0.4
	M89	0.4	0.4
	C8281G	0.3	0.5
	T12705C	0.1	0.15
	C3970T	0.3	0.2
	C10873T	0.1	0.2
	C10238T*	0.1	0.4
	A1736G	0.1	0.1
	C10400T*	0.1	0.1
MUTLIPLEX 2	T7028C	0.1	0.2
	A12308G	0.1	0.1
	T4216C	0.1	0.2
	T15607G	0.1	0.1
	M174	0.4	0.3
	M180	0.4	0.4
	M130	0.4	0.4
	M181	0.2	0.2
	M40	0.2	0.2
	M217	0.3	0.4
	M38	0.2	0.1
	M91	0.2	0.2
	M215	0.2	0.4
MULTIPLEX 3	M343	0.3	0.3
	G3010A	0.1	0.1
	T980C	0.1	0.1
	C12633A	0.1	0.05
	M170	0.2	0.15
	M124	0.2	0.05
	A3348G	0.1	0.05
	M172	0.2	0.1
	M69	0.2	0.2
	SRY10831.2	0.2	0.2
MULTIPLEX 4	C4140T	0.1	0.3
	G4491A	0.1	0.2
	A4715G*	0.1	0.2
	C4883T*	0.1	0.2
	T9824C	0.1	0.2
	C12403T	0.075	0.2
	M175	0.3	0.3
	M70	0.3	0.2
	M147	0.3	0.3
	P31	0.4	0.2
	M119	0.2	0.2
	M22	0.2	0.2
	P256	0.2	0.2
	M122	0.2	0.2

*Amplified by the same PCR primer pair within their respective multiplex.

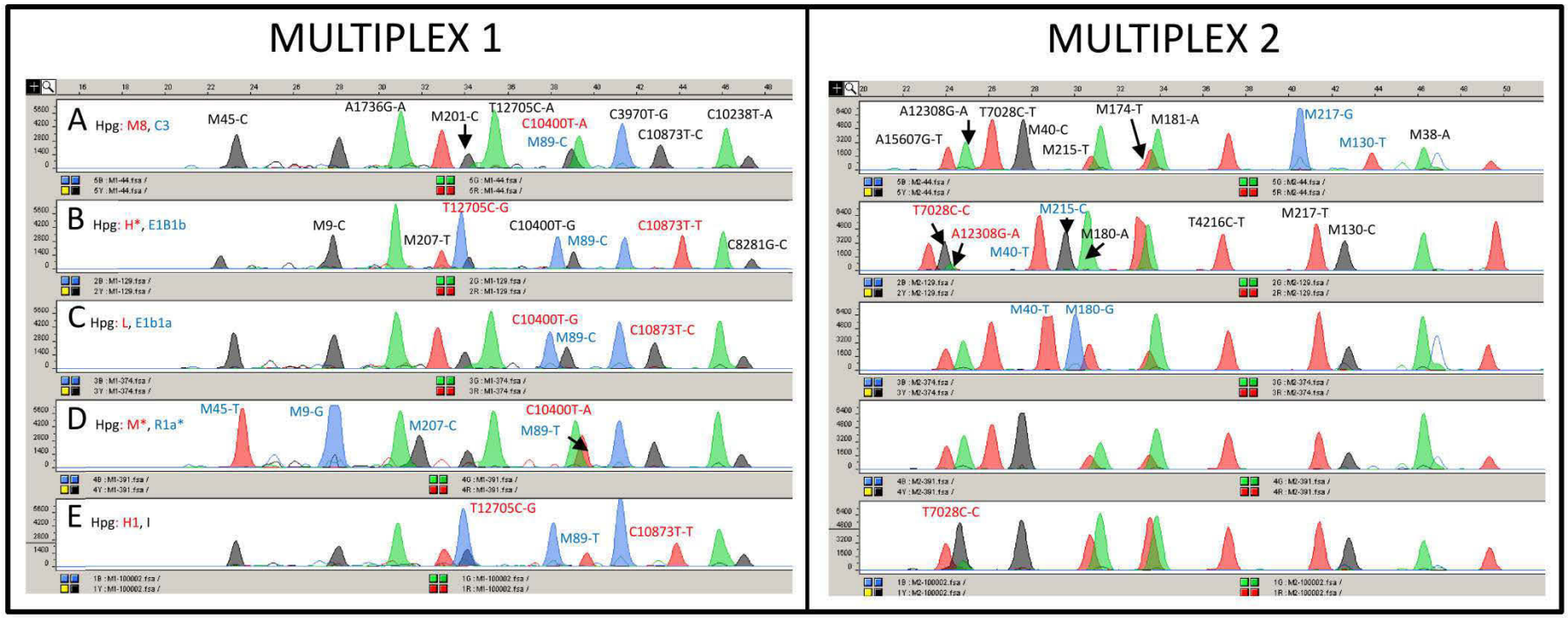


Figure 3.33: Initial population testing electropherograms of the Multiplex 1 and 2 profiles for five individuals with differing BGA.

(A) Chinese individual with mtDNA haplogroup (hpg) M8 and NRY hpg C3, (B) Egyptian individual with mtDNA hpg H*and NRY hpg E1b1b, (C) Cameroonian individual with mtDNA hpg L and NRY hpg E1b1a, (D) Indian individual with mtDNA hpg M* and NRY hpg R1a*, (D) English individual with mtDNA hpg H1 and NRY hpg I. The minimal SNPs required for mtDNA and NRY haplogroup assignation are shown in red and blue respectively.

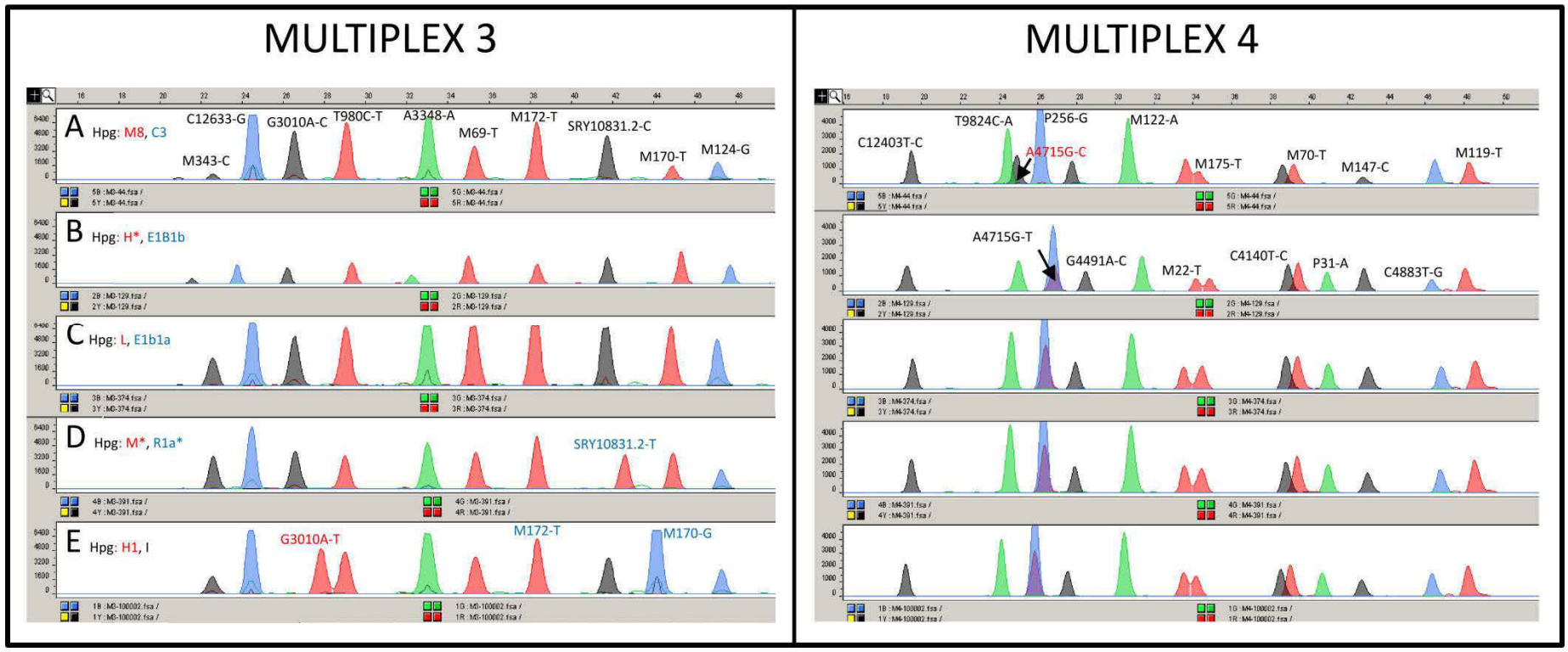


Figure 3.34: Initial population testing electropherograms of the Multiplex 3 and 4 profiles for five individuals with differing BGA.

(A) Chinese individual with mtDNA haplogroup (hpg) M8 and NRY hpg C3, (B) Egyptian individual with mtDNA hpg H*and NRY hpg E1b1b, (C) Cameroonian individual with mtDNA hpg L and NRY hpg E1b1a, (D) Indian individual with mtDNA hpg M* and NRY hpg R1a*, (D) English individual with mtDNA hpg H1 and NRY hpg I. The minimal SNPs required for mtDNA and NRY haplogroup assignation are shown in red and blue respectively.

haploid SNP assays was found to correlate with their declared maternal and paternal grandparental BGA (Figure 3.1 and Figure 3.2). These preliminary population results provided a promising example of the ability of the multiplex haploid SNP assays to successfully infer maternal and paternal BGA. Evidently, the initial population testing results have demonstrated that Multiplex 1 to 4 were adequately optimised and that the hierarchical design is operational. This concluded the first part of the development of our multiplex SNP assay, with the Multiplex 1 to 4 haploid SNP assays ready for validation through large scale population sample testing.

3.4 Summary

This chapter described the first stage of development of a multiplex genotyping assay based on single base extension technology targeting three different classes of SNPs (mtDNA, NRY and autosomal) for inferences of BGA and EVCs. Extensive literature searches resulted in the selection of 49 haploid SNPs (21 mtDNA and 28 NRY SNPs) that define haplogroups, which show distinct frequency differences between major global populations. A hierarchical strategy was used to design the assay to accommodate for the large number of SNPs selected. The haploid mtDNA and NRY SNPs were grouped into four multiplexes, with SNPs defining basal mtDNA and NRY haplogroups arranged together in Multiplex 1. The results from Multiplex 1 determine which subsequent multiplex is to be performed to achieve further haplogroup resolution. For each of Multiplex 1 to 4, PCR and SBE primers were designed. Preliminary optimisation of the PCR reactions was performed by initially testing the PCR primers in singleplex reactions prior to combining them in a single multiplex reaction. Reagent and PCR primer concentrations were altered for Multiplex 1 until relatively equalised PCR product yields were obtained. The optimised reagent concentrations determined from Multiplex 1 were then applied to the other multiplexes, for which only PCR primer concentrations were varied to achieve balanced amplification of all loci.

The SBE reactions were then optimised using the optimised multiplex PCR products as template. The specificity of the designed SBE primers was first verified through extension in singleplex SNaPshot™ reactions, followed by detection in multiplex SNaPshot™ reactions. Allele peak heights for each of the multiplexes were optimised

through the adjustment of PCR primer and/or SBE primer concentrations until all alleles could be unambiguously detected and peak heights ranging from 2500-4000RFU were obtained. Though some Multiplex 1 PCR and SBE primers required redesign, the issues encountered in the development of this multiplex culminated in improved primer design guidelines and facilitated the development of subsequent multiplexes. Following the completion of multiplex assay optimisations, allele specificity and satisfactory allele spacing were achieved in all 4 multiplexes, which confirmed good primer design. The decision to use POP-6™ polymer for capillary electrophoresis instead of the recommended POP-4™ polymer also played a role in achieving adequate spacing between alleles as greater spatial resolution and sizing accuracy were observed in POP-6™ compared to POP-4™ polymer.

Initial population testing was performed to evaluate the four optimised multiplex haploid SNP assays. All four multiplexes were used to genotype five individuals with different declared BGA to detect the alternate alleles at each locus. Initial population testing results were promising and demonstrated that the hierarchical system was effective as no ambiguous genotypes were detected across the four multiplexes. The peak heights detected were satisfactory and remained adequately equalised despite variations in the DNA content of FTA® discs from different donor samples. The haplogroup status of the five genotyped individuals demonstrated correlations between the geographic affiliation of these mtDNA and NRY haplogroups as reported in the literature and each individual's declared BGA. The first stage of multiplex assay development was thus completed, as all four haploid SNP multiplexes were confirmed to be adequately optimised and ready for large-scale genotyping of population samples, the results of which are discussed in chapter 5. The successful development of the haploid multiplexes explored in this chapter instigated the commencement of the next stage of hierarchical multiplex assay development, namely that of the fifth multiplex autosomal SNP assay, which is detailed in chapter 4.

CHAPTER 4

MULTIPLEX SNP ASSAY DEVELOPMENT STAGE 2:

MULTIPLEX 5 FOR THE INFERENCE OF BI-PARENTAL ANCESTRY AND PIGMENTATION PHENOTYPE

Chapter 4. Multiplex SNP Assay Development (Stage 2): Multiplex 5 for the Inference of Bi-parental Ancestry and Pigmentation Phenotype

4.1 Introduction

As discussed in chapter 1, ancestry informative markers (AIMs) which are genetic polymorphisms that demonstrate considerable allele frequency differences between global populations, have been successfully used to infer the biogeographical ancestry (BGA) of individuals originating from diverse and admixed populations (Shriver *et al.*, 2003, Collins-Schramm *et al.*, 2004, Kim *et al.*, 2005, Yang *et al.*, 2005, Baye *et al.*, 2009, Kersbergen *et al.*, 2009, Kosoy *et al.*, 2009, Nassir *et al.*, 2009, Kidd *et al.*, 2011, Pereira *et al.*, 2012). The primary objective of this study was to develop a forensic intelligence assay genotyping three separate classes of SNPs, including AIMs, for the prediction of maternal, paternal and bi-parental biogeographical ancestry (BGA) and external visible characteristics (EVC). The first stage of assay development discussed in chapter 3 involved the selection of 21 mitochondrial DNA (mtDNA) and 28 non-recombining Y-chromosome (NRY) SNPs defining population-specific haplogroups. The selected haploid SNPs were separated and grouped into four multiplexes following a hierarchical design, in which the result of Multiplex 1, which types SNPs defining basal mtDNA and NRY haplogroups, is used to determine which subsequent multiplex(es) is required to assign the final haplogroup. The maternal and paternal BGA of the individual is assigned to be the population of origin defined by their final mtDNA and NRY haplogroups.

Uniparentally inherited haploid markers represent ancient lineage based ancestries and are frequently asymmetric, reflecting the likelihood of maternal and paternal lineage ancestors from different geographic locations (Royal *et al.*, 2010). However, mtDNA and NRY SNPs only reflect part of the total genetic ancestry of an individual. The bulk of the genetic history is found in the autosomes, with bi-parentally inherited autosomal SNPs constituting the most abundant class of AIMs. Due to the shuffling of autosomal DNA through random assortment and recombination, only a small fraction of ancestors are represented by any given genetic loci in an individual and so, in contrast to haploid

markers, the ancestry information obtained from an individual may reflect quite recent ancestry contributions (Royal *et al.*, 2010). Nonetheless autosomal SNPs that show strong associations with specific biogeographical origins have been identified (Collins-Schramm *et al.*, 2004, Yang *et al.*, 2005, Kersbergen *et al.*, 2009, Kosoy *et al.*, 2009, Nassir *et al.*, 2009, Paschou *et al.*, 2010, Kidd *et al.*, 2011, Phillips *et al.*, 2013). Autosomal SNPs can also be used to investigate admixture and the effects of migration in populations (Shriver *et al.*, 2003, Sankararaman *et al.*, 2008, Baye *et al.*, 2009, Galanter *et al.*, 2012, Nievergelt *et al.*, 2013). Moreover, since genes determine phenotypic characteristics, autosomal SNPs can exhibit strong associations with specific traits (Kim and Borevitz, 2006) and several have been associated with pigmentation variation in humans (Sulem *et al.*, 2007, Han *et al.*, 2008, Kayser *et al.*, 2008b, Sturm *et al.*, 2008, Branicki *et al.*, 2009). The final stage of assay development thus involved the selection of thirteen autosomal SNPs that demonstrate strong associations with BGA and/or EVCs. These autosomal SNPs were grouped separately in a fifth multiplex to facilitate primer design and genotype interpretation. The design and optimisation of the preliminary Multiplex 5 PCR and SNaPshot™ reactions are discussed in detail in this chapter.

4.2 SNP Assay development- Stage 2

4.2.1 Autosomal SNP selection

The selection of appropriate autosomal SNP loci for applications in ancestry investigations requires SNPs that demonstrate low heterozygosity and thus are not highly polymorphic, as well as SNPs that demonstrate high inter-population diversity among and between global population groups (Budowle and van Daal, 2008). The SNP loci that have been subjected to genetic drift and are progressing towards fixation due to positive selection in these population groups are considered the most informative loci (Budowle and van Daal, 2008). Allele frequencies are used to examine the population associations of SNP loci, with ideal AIMs demonstrating fixed difference allele frequencies, where one allele is exclusive to one population and the alternative allele is found exclusively in the other(s) (Rosenberg *et al.*, 2003). The rs1426654 and rs16891982 SNPs located in the *SLC24A5* and *SLC45A2* pigmentation genes respectively, as well as rs2814778 in the blood group DARC gene, are fixed difference

markers that have previously been incorporated in a 34-Plex ancestry prediction assay (Phillips *et al.*, 2007b). However, due to the nature of human genetics, these fixed difference alleles are rare, as most alleles are generally shared among populations. For this reason, skewed allele frequencies where one allele is prevalent in one population but extremely rare in the other can also identify SNP loci capable of distinguishing different population groups. A SNP with an allele frequency differential between populations of at least 0.6 has been suggested to qualify as a skewed frequency marker (Shriver *et al.*, 1997). Alternatively, population-specific markers where one allele occurs in one or two populations but is absent from the others also provide good candidates for ancestry estimation (Phillips *et al.*, 2004).

Various statistics such as Absolute Allele Frequency Difference (δ), Shannon Information Content (SIC), Fisher Information Content (FIC), F-statistics (F_{ST}) and the Informativeness for Assignment measure (I_n) have been proposed for the selection of the most informative ancestry SNP loci and have been described in detail in Rosenberg *et al.* (2003). A study comparing the performance of these statistics revealed that I_n was the most effective at selecting the set of AIMs with the lowest bias and mean square error in ancestry estimations (Ding *et al.*, 2011). Several panels of AIMs have been identified using these statistics (Shriver *et al.*, 2003, Collins-Schramm *et al.*, 2004, Phillips *et al.*, 2004, Smith *et al.*, 2004, Yang *et al.*, 2005, Halder *et al.*, 2008, Baye *et al.*, 2009, Kosoy *et al.*, 2009, Kidd *et al.*, 2011) and numerous markers have been incorporated in autosomal ancestry prediction assays (Ray *et al.*, 2005, Yang *et al.*, 2005, Alloco *et al.*, 2007, Phillips *et al.*, 2007b, Halder *et al.*, 2008, Kersbergen *et al.*, 2009, Lao *et al.*, 2010, Londin *et al.*, 2010, Paschou *et al.*, 2010, Pereira *et al.*, 2012, Phillips *et al.*, 2013). However, the AIMs incorporated in these assays were selected for their associations to the target populations investigated by the authors of these studies. Thus, these previously developed assays were merely consulted and only appropriate markers demonstrating associations with our populations of interest were chosen for incorporation into our multiplex assay.

Literature and database reviews were conducted to identify population specific autosomal SNPs, fixed difference and skewed allele frequency markers that distinguish Sub-Saharan African, European, Middle Eastern, South Asian and East Asian populations. The databases from which genotype and allele frequency data can be

retrieved include the HapMap, CEPH, 1000 GENOMES and NCBI dbSNP databases (refer to Section 1.4). The web interface, SPSmart v5.1.1 (available at http://spsmart.cesga.es/), can be used to search for SNP data across these major databases and can also provide statistical summaries including allele frequencies, heterozygosity and F_{ST} (Amigo *et al.*, 2008b). However, the main issue with using these public SNP databases to search for suitable AIMs is that SNP data may not always be available for all populations of interest, such as the Middle East and South Asian populations, for which SNP data is currently limited to the CEPH database. For this reason, the literature was reviewed first to identify appropriate SNP loci and the SPSmart v5.1.1 interface was then used to retrieve SNP frequency data for the selected SNPs to confirm population associations from the literature.

From the literature reviews, allele and genotype frequency data was gathered and used to select autosomal SNPs demonstrating strong population specific associations. A total of thirteen candidate autosomal SNPs located in blood group, cytokine and pigmentation genes were chosen to differentiate the major population groups in Australia. Some autosomal SNPs were also selected for their associations with EVCs. The candidate SNPs for Multiplex 5 are summarised in Table 4.1, which also shows the SNP position, nucleotide substitution and amino acid substitutions for candidate SNPs arising in the coding region acquired from the NCBI dbSNP database (Sherry *et al.*, 2001). The allele frequencies for each candidate SNP were retrieved from the 1000 Genomes Database (Phase I) using SPSmart v5.1.1 and are provided in Table 4.2. The characteristics of the candidate SNPs that contributed to their incorporation within Multiplex 5 are discussed in the following section.

4.2.1.1 *AIMs in blood grouping and cytokine genes*

The Duffy blood group is one of the most commonly investigated blood groups to reconstruct global population genetic relationships and migrations (Mourant and Domaniewska-Sobczak, 1962). The Duffy Antigen/Receptor for Chemokine (*DARC*) gene encodes the Duffy glycoprotein found in the erythrocyte membrane which functions as a chemokine receptor as well as a binding site for the malaria parasites *Plasmodium vivax* and *Plasmodium knowlesi* (Miller *et al.*, 1975, Horuk *et al.*, 1993).

Table 4.1: Multiplex 5 candidate autosomal SNPs

SNP Name	dSNP ID number	Gene	Chromosome	Nucleotide Position*	Base Change (Ancestral>Variant)	Protein Accession[#], Amino acid substitution
MC1R	rs885479	*MC1R*	16	89986154	G>A	NP_002377.4, R163Q
MATP	rs16891982	*SLC45A2*	5	33951693	C>G	NP_057264.3, F374L
SLC24A5	rs1426654	*SLC24A5*	15	48426484	G>A	NP_995322.1, A111T
OCA2-1	rs7495174	*OCA2*	15	28344238	G>A	-
OCA2-2	rs4778241	*OCA2*	15	26012308	A>C	-
OCA2-3	rs4778138	*OCA2*	15	26009415	G>A	-
OCA2-4	rs1545397	*OCA2*	15	28187772	A>T	-
DARC	rs2814778	*DARC*	1	159174683	T>C	-
F13B	rs6003	*F13B*	1	197031021	G>A	NP_001985.2, R115H
HERC2	rs12913832	*HERC2*	15	28365618	A>G	-
ASIP	rs6058017	*ASIP*	20	32856998	G>A	-
FUT2	rs1800021	*FUT2*	19	49206286	A>G	NP_000502.4, I25V
IL-6	rs1800795	*IL-6*	7	22766645	G>C	-

*Reference assembly (build 36.3) and [#]Protein accession taken from the NCBI dbSNP database (available at http://www.ncbi.nlm.nih.gov/projects/SNP/).

Table 4.2: Allele frequencies for Multiplex 5 candidate autosomal SNPs retrieved from the 1000 Genomes Database (Phase I) using SPSmart v5.1.1.

SNP	Alleles	Sub-Saharan Africa[a] N=246	Europe[b] N=380	East Asia[c] N= 286
rs885479	G	0.970	0.913	0.353
	A	0.030	0.087	0.647
rs16891982	C	0.941	0.029	0.983
	G	0.059	0.971	0.017
rs1426654	G	0.917	0.004	0.986
	A	0.083	0.996	0.014
rs7495174	G	0.177	0.070	0.647
	A	0.823	0.930	0.353
rs4778241	A	0.598	0.218	0.766
	C	0.402	0.782	0.234
rs4778138	G	0.709	0.142	0.717
	A	0.291	0.858	0.283
rs1545397	A	0.961	0.921	0.086
	T	0.039	0.079	0.914
rs2814778	C	0.943	0.003	0.000
	T	0.057	0.997	1.000
rs6003	C	0.646	0.079	0.014
	T	0.354	0.921	0.986
rs12913832	A	0.961	0.289	0.998
	G	0.039	0.711	0.002
rs6058017	G	0.762	0.104	0.198
	A	0.238	0.896	0.802
rs1800021	A	0.896	1.000	1.000
	G	0.104	0.000	0.000
rs1800795	C	0.026	0.417	0.000
	G	0.974	0.583	1.000

[a] Yoruba, Lunya, Gambians, Mende, Esan

[b] Utah residents with Northern and Western European ancestry, Tuscans, Brits from England and Scotland, Iberians

[c] Han Chinese from Beijing and South China, Japanese, Vietnamese Kinh

The *DARC* gene has two co-dominant alleles, FY*A and FY*B which are differentiated by a single G125A nucleotide substitution that results in a Asp to Gly amino acid change at position 42 and produces two antigenic protein variants Fy^a and Fy^b (Iwamoto *et al.*, 1995). There are four Duffy phenotypes, Fy(a+b+), Fy(a+b-), Fy(a-b+) and Fy(a-b-). The expression of the Duffy antigens is affected by a T to C substitution at position -33 of the promoter region of the *DARC* gene. This SNP (rs2814778) disrupts the binding site for the erythroid transcription factor and prevents expression of the Duffy antigens in erythrocytes, resulting in the Fy(a-b-) phenotype (Tournamille *et al.*, 1995).

The Duffy alleles show distinct geographical patterns of genetic variation, with the Fy(a-b-) phenotype observed at high frequencies in individuals of African descent compared to other population groups where this phenotype is practically non-existent (Sanger *et al.*, 1955, Welch *et al.*, 1977, Cavalli-Sforza *et al.*, 1994, Howes *et al.*, 2011). The frequency of the rs2814778*C allele which contributes to the Fy(a-b-) phenotype ranges from 80-100% in African American and Sub-Saharan African individuals. However, this allele is almost absent from European and Asian populations where the rs2814778*T allele is partially fixed (Cavalli-Sforza *et al.*, 1994, Olsson *et al.*, 1998, Nickel *et al.*, 1999, Halder *et al.*, 2008, Howes *et al.*, 2011). These allele frequencies reported in the literature complement those retrieved from the 1000 Genome Database (Table 4.2). Individuals of African ancestry who are homozygous for the rs2814778*C allele are completely resistant to *P. knowlesi* and *P. vivax* as these parasites require the expression of the DARC protein in order to infect erythrocytes, which suggests positive selection of this locus in the Sub Saharan African population where Malaria is endemic (Miller *et al.*, 1975, Miller *et al.*, 1976, Hedrick, 2011b). Due to the partial fixation of the rs2814778 polymorphism in the African population, this AIM has also demonstrated one of the highest level of population differentiation as measured by F_{ST} (Cavalli-Sforza *et al.*, 1994). Accordingly, the rs2814778 was selected for its potential to distinguish Sub-Saharan African individuals from those of European and Asian ancestry.

AIMs have also been identified in the Fucosyltransferase 2 (*FUT2*) gene, which is one of the major components of the ABO blood group system. The ABO blood group antigens consist of α(1,2)-linked fucose containing glycans which are found on the glycoproteins and glycolipids of erythrocyte membranes and epithelial cells (Oriol *et al.*, 1986). Whilst the *FUT1* gene is responsible for the expression of the precursor H

antigen on erythrocytes, the *FUT2* gene encodes the α(1,2) fucosyltransferase 2 enzyme which generates the expression of H antigens in mucus and other secretions such as saliva (Oriol *et al.*, 1981, Le Pendu *et al.*, 1985, Clausen and Hakomori, 1989). There are two enzyme variants, which are expressed by a functional *Se* allele and a non-functional *se* allele. Individuals who are able to present ABO antigens in secretions must have at least one functional *Se* allele and those who fail to express ABO antigens are homozygous for the non-functional *se* allele (Kelly *et al.*, 1995). These non-secretor phenotypes arise from mutations in the second exon of the *FUT2* gene, with the most frequently observed including the se^{428} allele in Europeans, Iranians and Africans (Kelly *et al.*, 1995, Liu *et al.*, 1998, Koda *et al.*, 2001, Ferrer-Admetlla *et al.*, 2009) as well as se^{385} in South East and East Asians (Koda *et al.*, 2001, Soejima *et al.*, 2007, Ferrer-Admetlla *et al.*, 2009).

Numerous *FUT2* polymorphisms have been associated with the secretor phenotype and many of these are population specific (Koda *et al.*, 2001). The Se^{40}, Se^{375} and Se^{481} polymorphisms are predominant in Xhosas from South Africa (Liu *et al.*, 1998, Koda *et al.*, 2001), whilst the Se^{357} and Se^{480} variants are observed in African, European and Middle Eastern populations (Kelly *et al.*, 1995, Liu *et al.*, 1998, Koda *et al.*, 2001, Soejima *et al.*, 2007). More recently, the Se^{375} polymorphism has demonstrated a higher allele frequency in Oceanian populations (0.2830) compared to Sub-Saharan African populations (0.0120) where this SNP was initially detected but has not been observed in other populations (Ferrer-Admetlla *et al.*, 2009). Of particular interest, the functional Se^{40} (rs1800021) variant is a non-synonymous A to G substitution which causes the I25V amino acid change in the FUT2 protein and is detected in 10-28% of African individuals, particularly the Xhosa population where the allele occurs at a frequency of 0.21 but is practically absent from the European and Asian populations (Liu *et al.*, 1998, Koda *et al.*, 2001, Soejima *et al.*, 2007). As a result, the Se^{40} (rs1800021) variant was selected as a candidate allele as the marker is capable of distinguishing the African population from other populations.

Population genetic variations have also been identified in cytokine genes particularly those of the interleukin (IL) family including IL-4, IL-6 and IL-10 (Eskdale *et al.*, 1998, Mok *et al.*, 1998, Cox *et al.*, 2001, Wu *et al.*, 2001, Hoffmann *et al.*, 2002, Meenagh *et al.*, 2002, Hassan *et al.*, 2003). Interleukins play a central role in the modulation of

immune responses and IL-6, a pleiotropic cytokine expressed by many cell types in response to inflammatory stimuli, not only regulates the immune response but also has roles in acute inflammation and haematopoiesis particularly in the differentiation of B and T cells (Kishimoto *et al.*, 1995). This cytokine has been associated with various inflammatory, autoimmune and malignant diseases such as juvenile rheumatoid arthritis, neuroblastoma, breast cancer, susceptibility to Alzheimer's, atherosclerosis and cardiovascular disease (Fishman *et al.*, 1998, DeMichele *et al.*, 2003, Licastro *et al.*, 2003, Tzoulaki *et al.*, 2005, Lagmay *et al.*, 2009, Yin *et al.*, 2012).

One IL-6 polymorphism of interest, rs1800795, is a G to C substitution arising at position -174 of the promoter region of the *IL-6* gene that results in decreased production of the IL-6 cytokine (Fishman *et al.*, 1998, Rivera-Chavez *et al.*, 2003, Lagmay *et al.*, 2009). The rs1800795 SNP, which is also known in the literature as IL6-174, shows heterogeneity across global populations. The C allele is polymorphic in Europeans where it occurs at frequencies ranging from 0.30 to 0.40 but is rare in Asian and African populations where the G allele predominates (0.90-1.0) (Cox *et al.*, 2001, Hoffmann *et al.*, 2002, Lim *et al.*, 2002, Meenagh *et al.*, 2002, Hassan *et al.*, 2003, Lei *et al.*, 2003). Similar allele frequencies retrieved from the 1000 Genomes Database are reflected in Table 4.2. The CC genotype is absent from Chinese, Korean and Japanese populations but occurs in 10-20% of Europeans (Lim *et al.*, 2002, Lei *et al.*, 2003). The CG and GG genotypes occur at intermediate frequencies in Europeans but Middle East and South Asian populations show a higher incidence of CG genotypes (~0.25) compared to Africans and East Asians where this genotype is relatively not detected (Fishman *et al.*, 1998, Meenagh *et al.*, 2002). Accordingly, the rs1800795 polymorphism was included as a candidate SNP as it has the ability to distinguish Europeans from Asians and Africans.

Another variant, which has been identified as an AIM is the rs6003 polymorphism. This SNP encodes a C to T substitution in the *F13B* gene, which causes a R115H amino acid change in the coagulation factor XIII B subunit. The rs6003 variant was identified through studies compiling ancestry informative marker panels which demonstrated δ greater than 0.3 between population groups through extensive SNP database searches (Shriver *et al.*, 2003, Smith *et al.*, 2004, Salari *et al.*, 2005, Choudhry *et al.*, 2006, Ziv *et al.*, 2006, Halder *et al.*, 2008). The ancestral C allele occurs at high frequencies ranging

from 0.65 to 0.70 in the African population but is found at low frequencies in European (~0.08) and Asian (~0.03) populations where the T allele (>0.9) predominates (Shriver *et al.*, 2003, Smith *et al.*, 2004, Salari *et al.*, 2005, Choudhry *et al.*, 2006, Ziv *et al.*, 2006, Halder *et al.*, 2008). The allele frequencies shown in Table 4.2 retrieved from the 100 Genome Database parallel those reported in the literature and the calculated δ value comparing Africans and Europeans was in the range of 0.62-0.64 (Shriver *et al.*, 2003, Salari *et al.*, 2005, Choudhry *et al.*, 2006, Halder *et al.*, 2008). This confirmed that the rs6003 polymorphism is a suitable skewed frequency AIM that has the potential to distinguish the African population from other populations and resulted in the incorporation of this marker in Multiplex 5.

4.2.1.2 *SNPs in Human pigmentation genes*

Human pigmentation variation arises due to differences in the number, size and distribution of melanosomes, the organelles in which the pigment melanin accumulates as well as the type of melanin pigment they produce, either eumelanin (brown/black) or phaeomelanin (yellow/red), with the ratio of the two giving rise to a variety of pigmentation colours (Sturm *et al.*, 1998). As illustrated in Figure 4.1, melanin production is controlled by the combined effects of multiple genes. The Melanocortin-1 Receptor (*MC1R*) gene encodes the MC1R-G protein coupled membrane receptor for α-melanocyte stimulating factor (α-MSH) which is found in the membrane of melanosomes (Donatien *et al.*, 1992). Activation of this receptor through the binding of α-MSH, triggers an increase in intracellular cyclic adenosine monophosphate (cAMP), which promotes the production of eumelanin (Hunt *et al.*, 1995, Busca and Ballotti, 2000). The microphthalmia transcription factor (MITF) is activated by cAMP and stimulates the production of tyrosinase (TYR), tyrosine related protein-1 (TYRP1) and dopachrome tautomerase (DCT) which catalyse the transformation of tyrosine to eumelanin under optimum conditions (Yasumoto *et al.*, 1997, Bertolotto *et al.*, 1998, Smith *et al.*, 1998, Ludwig *et al.*, 2004). In contrast to activation of MC1R by α-MSH, the Agouti Signalling Protein (ASIP) antagonistically binds to the MC1R receptor, blocking the α-MSH induced signalling pathway within melanocytes and effectively inhibits eumelanin production in favour of phaeomelanogenesis (Lu *et al.*, 1994, Suzuki *et al.*, 1997, Yang *et al.*, 1997).

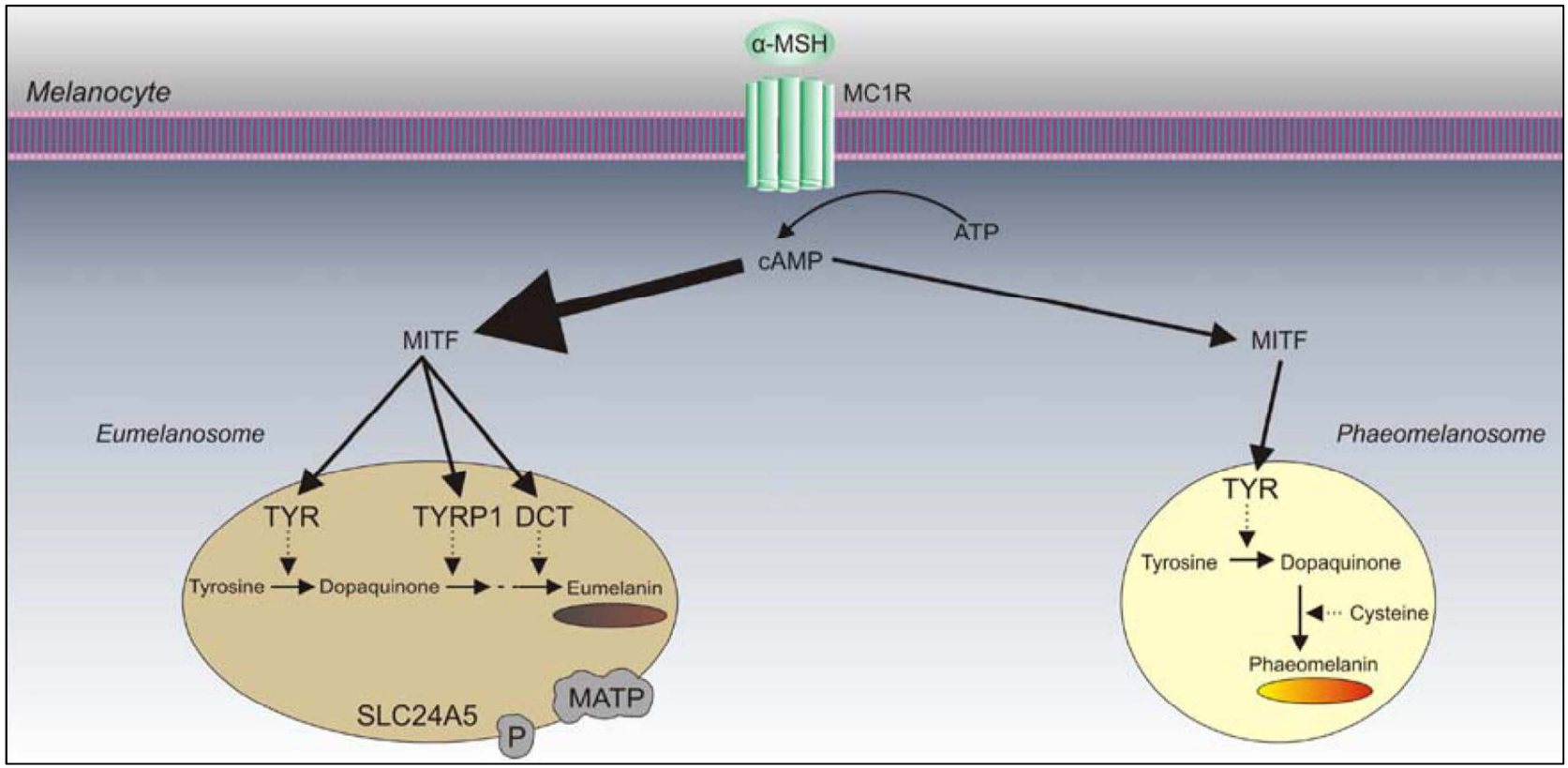


Figure 4.1: Overview of melanogenesis

Binding of α-melanocyte stimulating factor (α-MSH) to the MC1R membrane receptor triggers the production of intracellular cAMP and activates the MITF. MITF activates tyrosinase (TYR), tyrosine related protein-1 (TYRP1) and dopachrome tautomerase (DCT), which catalyse the transformation of tyrosine to eumelanin under optimum conditions. MATP, P protein and SLC24A5 are also essential and produce the conditions, which favour eumelanosome maturation. When cAMP levels are low, such as during antagonism of MC1R by Agouti Signalling Protein (ASIP), phaeomelanosome formation is favoured through the activation of tyrosinase, which converts tyrosine to dopaquinone, to which cysteine is added and this product is then oxidised to produce phaeomelanin. Image acquired from Bouakaze et al. (2009).

Other proteins that favour eumelanosome maturation include the membrane associated transporter protein (MATP) encoded by the *SLC45A2* gene (Du and Fisher, 2002); the trans-melanosomal P protein, encoded by the Oculocutaneous Albinism 2 (*OCA2*) gene (Rebbeck *et al.*, 2002) and the solute carrier family 24 member 5 (SLC24A5) which is a member of the potassium-dependent sodium/calcium exchanger family (Lamason *et al.*, 2005). The role of the latter three proteins remains unclear, though it has been proposed that the P protein may act as an anion transporter to regulate melanosomal pH and subsequently the activity of the TYR enzyme, the rate of melanin production and the ratio of eumelanin/phaeomelanin (Ancans *et al.*, 2001, Brilliant, 2001). Alternatively the P-protein may be involved in trafficking of the tyrosinase enzyme during melanosome maturation (Toyofuku *et al.*, 2002).

The *HERC2* gene, though not classified as a pigmentation gene, has been suggested to contain regulatory elements for the *OCA2* gene (Eiberg *et al.*, 2008, Kayser *et al.*, 2008c, Sturm *et al.*, 2008, Visser *et al.*, 2012) and is located 11.7kb downstream of the *OCA2* gene (Ji et al., 1999). A conserved region in intron 86 of *HERC2* contains a binding site for the helicase-like transcription factor (HLTF) (Sturm *et al.*, 2008), in which recruits the SWI-SNF chromatin remodelling complex that catalyses the conversion of the initially closed heterochromatin state of the *HERC2/OCA2* locus to a more relaxed and open euchromatin state (Visser *et al.*, 2012). This allows association of MITF and Lymphoid enhancer-binding factor 1 (LEF1) to this locus control region in HERC2, which enhance assembly of RNA Polymerase 2 to the *OCA2* promoter and hence initiation of *OCA2* transcription (Visser *et al.*, 2012), thereby favouring eumelanosome production and brown/black pigmentation. A *HERC2* variant G allele at rs12913832 located in intron 86 results in the inhibited expression of a luciferase reporter gene by 60-70% in *OCA2* mRNA expressing Caco2 cells (Eiberg et al., 2008). It was proposed that the rs12913832*G allele changes the HLTF binding site so that the chromatin remains closed, resulting in decreased *OCA2* expression and light pigmentation (Visser *et al.*, 2012). As outlined in the subsequent sections, many SNPs in genes associated with normal human pigmentation variation have been identified which are associated with specific phenotypes and/or populations, some of which have been chosen for incorporation into Multiplex 5.

4.2.1.2.1 *Fixed difference markers in the SLC24A5 and SLC45A2 genes*

The rs16891982 C>G SNP in the *SLC45A2* gene which encodes the MATP and results in a non-synonymous mutation (Phe373Leu), has been shown to be associated with pigmentation variation within and between populations (Nakayama *et al.*, 2002). The ancestral C allele was found to be fixed in Europeans but relatively absent from the other more darkly pigmented populations, including Africans, Asians and Oceanians where the G substitution is more predominant (Nakayama *et al.*, 2002, Yuasa *et al.*, 2004, Graf *et al.*, 2005). The low frequency of European individuals who carry the G allele tend to have darker pigmentation, with the rs16891982*G substitution strongly correlated with black hair, olive skin and brown eyes in Europeans (Graf *et al.*, 2005).

Similarly, the rs1426654 G>A (Ala111Thr) variant causes a missense mutation in the *SLC24A5* cation exchanger gene has also demonstrated distinct population pigmentation differences. The ancestral G allele predominates in African, Native American and Asian populations, in contrast to the A allele, which is almost fixed in Europeans, but has also been observed at high frequencies of 50-90% in South Asian populations (Smith *et al.*, 2004, Yang *et al.*, 2005, Phillips *et al.*, 2007b, Soejima and Koda, 2007). This AIM has also demonstrated strong associations with the fair skin and light hair phenotype observed in European populations (Soejima and Koda, 2007).

The two SNPs, rs1426654 and rs16891982 have been previously incorporated in an autosomal ancestry prediction assay and may be considered as fixed difference markers (Phillips *et al.*, 2007b). Consequently both the rs1426654 and rs16891982 SNPs were selected here for incorporation into Multiplex 5 for their ability to distinguish Europeans from other population groups as well as the correlations of these markers with pigmentation variation in the European population.

4.2.1.2.2 *Polymorphisms in the MC1R gene*

Nine SNPs in MC1R were initially found to be linked with red hair, fair skin and increased sensitivity to UV radiation (Valverde *et al.*, 1995). Subsequent studies of these SNPs in Europeans confirmed the Arg151Cys, Arg160Trp and Asp294His variants to be most strongly associated with red hair and fair skin, (Valverde *et al.*,

1995, Box *et al.*, 1997, Palmer *et al.*, 2000, Bastiaens *et al.*, 2001, Grimes *et al.*, 2001, Sulem *et al.*, 2007) Further, homozygotes or compound heterozygotes for these three variants were found to be more likely to develop this phenotype. These MC1R receptor variants were still capable of binding α-MSH but had decreased ability to activate cAMP (Frandberg *et al.*, 1998), thereby contributing to the switch from eumelanin to phaeomelanin production and hence to lighter pigmentation (Cone *et al.*, 1996). Over 30 other variant MCR1 alleles that show some associations with hair and skin colour in Caucasian populations comprising British, Irish and Australian subjects have been identified (Valverde *et al.*, 1995, Box *et al.*, 1997, Rana *et al.*, 1999, Harding *et al.*, 2000, Grimes *et al.*, 2001).

Studies investigating the frequencies of *MC1R* variants in Asian and African populations have identified Arg67Gln, Val92Met and Arg163Gln variants at high frequencies in Asian populations (Box *et al.*, 1997, Rana *et al.*, 1999, Harding *et al.*, 2000, Grimes *et al.*, 2001, Yuasa *et al.*, 2007, Masui *et al.*, 2009, Spichenok *et al.*, 2011) but an overall limited variability in African populations (Rana *et al.*, 1999, Harding *et al.*, 2000). For example the allelic frequencies for the Arg163Gln (rs885479 G>A) variant show skewed interpopulation differences with the A allele occurring in 50-80% of East Asians including individuals from China, Japan and Indonesia whilst the ancestral G allele is partially fixed in Europeans and Africans (Valverde *et al.*, 1995, Box *et al.*, 1997, Rana *et al.*, 1999, Harding *et al.*, 2000, Grimes *et al.*, 2001, Yuasa *et al.*, 2007, Masui *et al.*, 2009, Spichenok *et al.*, 2011). Interestingly, the A allele is found at much lower frequencies (0.07-0.1) in South Asian populations where G allele is prevalent (Rana *et al.*, 1999, Yuasa *et al.*, 2007).

The three variants identified in Asian populations did not show associations with particular hair or skin colours (Box *et al.*, 1997, Palmer *et al.*, 2000, Grimes *et al.*, 2001). A Thr314Thr *MC1R* variant that is found at high frequency in African populations (Harding *et al.*, 2000) was also observed to not play a significant role in distinguishing pigmentation differences between ethnic groups (Shriver *et al.*, 2003).
However, the Arg163Gln and other *MC1R* polymorphisms have been associated with variations in eye colour (Frudakis *et al.*, 2003a), particularly when combined with SNPs in other pigmentation genes (Branicki *et al.*, 2009, Spichenok *et al.*, 2011, Pneuman *et al.*, 2012). In particular the rs885479 (G>A) SNP shows strong associations with brown

eye, fair skin and brown hair colour when combined with the SNPs in the OCA2 (rs4778138) and HERC2 (rs12913832) pigmentation genes (Branicki *et al.*, 2009, Spichenok *et al.*, 2011, Pneuman *et al.*, 2012).which have also been selected as candidate SNPs in this assay. Thus, the rs885479 marker was incorporated into the Multiplex 5 assay as it can distinguish East Asians from other population groups and has been identified as a useful predictor for eye, hair and skin colour when analysed with other pigmentation SNPs.

4.2.1.2.3 *SNPs in the OCA2 gene*

The *OCA2* gene has been identified as a major contributor of human eye colour (Frudakis *et al.*, 2003a, Sturm and Frudakis, 2004, Duffy *et al.*, 2007, Frudakis *et al.*, 2007) and many SNPs with associations to specific eye colours have been identified. For example, individuals who carry the Arg305Trp or Arg419Gln variants or a combination of both, were less likely to have blue or grey eyes (Rebbeck *et al.*, 2002). Using Fisher's exact test and multiple regression analysis to investigate the correlation between 271 *OCA2* polymorphisms and eye colour variation (measured using iris colour as the melanin indicator), Frudakis *et al.* (2007) confirmed 6 previously described associations (Frudakis *et al.*, 2003a) and identified 27 new associations with eye colour (Frudakis *et al.*, 2007). Of particular interest, three SNPs, rs7495174 (G>A), rs4778241 (A>C) and rs4778138 (G>A) located in intron 1 of the *OCA2* gene have shown the strongest association with blue/non-blue eye colour when grouped together as a haplotype block, with the ACA/ACA genotype occurring in approximately 90% of individuals with blue or green eyes (Duffy *et al.*, 2007).

Frequency analysis of the Duffy *et al.* (2007) *OCA2* intron 1 haplotype using the HapMap database have revealed interpopulation differences, with the highest frequency (0.835) for the ACA haplotype in Europeans (Duffy *et al.*, 2007). Minor frequencies were observed in Asian (0.121) and Africans (0.074) suggesting positive selection of this haplotype in European populations (Duffy *et al.*, 2007). Retrieval of individual allele frequencies for the rs7495174, rs4778241 and rs4778138 using the 1000 Genomes database also demonstrate the population specific associations for these markers. The rs74955174 ancestral G allele is found in 65% of East Asians whilst the A allele predominates and European populations and is found at intermediate frequencies

in Africans (Table 4.2). The rs4778241 C and rs4778138 A alleles are found in approximately 80% of Europeans whilst the alternate alleles predominate in the other populations (Table 4.2). These three SNPs were therefore selected for Multiplex 5 to distinguish European and Asian populations and for their contributions to the prediction of blue/non-blue eye colour.

Another *OCA2* SNP, rs1545397 A>T (His615Arg) has been identified as a population specific marker which occurs in 85-90% of East Asians, particularly those of Japanese and Chinese ancestry with the ancestral allele predominating in European and African populations (Edwards *et al.*, 2010, Spichenok *et al.*, 2011). This SNP has also been identified as a predictor of brown eye and hair colour and fair skin when combined with the rs12913832 in the *HERC2* gene (Spichenok *et al.*, 2011, Pneuman *et al.*, 2012). The rs1545397 marker was also incorporated into the Multiplex 5 assay as it is an AIM capable of distinguishing East Asian populations from other populations and may additionally reinforce EVC associations of other pigmentation SNPs incorporated in the multiplex such as rs12913832, which is discussed in the following section.

4.2.1.2.4 *Polymorphisms in the HERC2 gene*

The rs12913832 (A>G) SNP located in intron 86 of the *HERC2* gene has demonstrated very strong correlations with blue/brown eye colour (Eiberg *et al.*, 2008, Sturm *et al.*, 2008). The G substitution which disrupts the HLTF transcription binding site for *OCA2* thereby decreasing *OCA2* gene expression and results in lighter pigmentation (Sturm *et al.*, 2008, Sturm and Larsson, 2009, Visser *et al.*, 2012), was shown to have almost complete association with blue eye colour in individuals from Denmark, Turkey and Jordan (Eiberg *et al.*, 2008). An alternative study identified the *HERC2* SNP rs1916977 as the strongest determinant of eye colour, with the C allele associated with blue iris colour and the T allele with brown iris colour in Europeans (Kayser *et al.*, 2008c). A different study reported conflicting results suggesting that rs12913832 was the most informative predictor of eye colour, however also demonstrated that rs1916977, amongst other HERC2 SNPs such rs1129038 and rs1667394, showed significant associations to eye colour when the SNPs were analysed together as a haplotype (Sturm *et al.*, 2008). Both the Kayser *et al.* (2008c) and Sturm *et al.* (2008) studies showed that combining these *HERC2* genotypes with *OCA2* genotypes such as the three SNP-

haplotype in intron 1 of the *OCA2* gene (Duffy *et al.*, 2007) and *OCA2* SNP Arg419Gln (Rebbeck *et al.*, 2002) improved eye colour prediction compared to when only the *OCA2* SNPs were used (Kayser *et al.*, 2008c, Sturm *et al.*, 2008).

Though the HERC2 gene appears to play a central role in eye colour pigmentation, it has also been linked with other pigmentation phenotypes such as hair (Han *et al.*, 2008, Branicki *et al.*, 2009, Branicki *et al.*, 2011) and skin (Branicki *et al.*, 2009, Valenzuela *et al.*, 2010, Spichenok *et al.*, 2011) colour. In particular, the rs12913832 substitution has demonstrated the strongest association to both hair and skin colour compared to other HERC2 variants (Han *et al.*, 2008, Branicki *et al.*, 2009, Valenzuela *et al.*, 2010). This *HERC2* SNP has been incorporated in several EVC prediction assays, including Irisplex and HIrisplex, to complement other pigmentation SNPs such as rs1545397 (*OCA2*), rs1426654 (*SLC24A5*), rs16891982 (*SLC45A2*) and rs885479 (*MC1R*) (Spichenok *et al.*, 2011, Walsh *et al.*, 2011, Pneuman *et al.*, 2012, Allwood and Harbison, 2013, Ruiz *et al.*, 2013, Walsh *et al.*, 2013) to achieve high eye and hair colour prediction accuracies (>90%), though less effective skin colour predictions (~70-80%).

The allele frequencies of rs12913832 derived from the HGDP-CEPH database also reflect the observed eye colour distribution across global populations, demonstrating the potential utility of this polymorphism for ancestry predictions. The rs12913832 G allele is predominantly restricted to Europe where blue and intermediate eye colours are expected and which correlates with the high incidence of GG/AG genotypes observed in these populations (Walsh *et al.*, 2011). In contrast, the rs12913832 ancestral A allele which is associated with brown eye colour, is found almost exclusively in African, East Asian and Native American populations where this phenotype is expected (Walsh *et al.*, 2011). The rs12913832 ancestral A allele occurs at lower frequencies in Middle Eastern and South Asian populations where a moderate incidence of AG genotypes and low incidence of GG genotypes is observed. This correlates with the low occurrence of intermediate and blue eye colours in some individuals originating from these populations and which are relatively absent from African, East Asian and Native American populations. Similar allele distributions are observed for the 1000 Genomes database (Table 4.2). The rs12913832 SNP was selected as a candidate SNP as it can distinguish Europeans from other populations and is a good predictor of eye colour,

which can also be used to infer hair and skin colour when combined with other Multiplex 5 candidate SNPs.

4.2.1.2.5 *Agouti Signalling Protein polymorphisms*

A G8818A (rs6058017 G>A) substitution in the 3' untranslated region of the Agouti Signalling Protein (ASIP) gene has been proposed to increase the antagonist action of ASIP to α-MSH by binding to MC1R and blocking cAMP signalling within the melanocytes (Kanetsky *et al.*, 2002) This favours phaeomelanin production and explains the lighter skin pigmentation observed in individuals carrying the substitution (Kanetsky *et al.*, 2002, Bonilla *et al.*, 2005). In contrast, Europeans carrying the ancestral G allele demonstrated a higher likelihood of having brown hair and brown eyes (Kanetsky *et al.*, 2002). A population study of the rs6058017 polymorphism revealed skewed allele frequencies between global populations, where the ancestral G allele occurred in 80% of Africans, 28% of Asians and 12% of Europeans (Zeigler-Johnson *et al.*, 2004). These allele frequencies are similar to those acquired from the 1000 Genomes database (Table 4.2) and complement the observed global pigmentation distribution, as the A substitution associated with lighter hair and eye colour shades were observed at higher frequencies in Europeans where these phenotypes are prevalent. The ancestral G allele was also found in 66% of African Americans and was associated with higher melanin indexes, suggesting a likely correlation between the rs6058017 SNP and skin pigmentation (Bonilla *et al.*, 2005). Whilst other polymorphisms of the ASIP gene have also been associated with eye and hair colour variation (Frudakis *et al.*, 2003a, Sulem *et al.*, 2007, Valenzuela *et al.*, 2010, Walsh *et al.*, 2013), the rs6058017 was incorporated into Multiplex 5 due to the skewed allele frequencies which can distinguish Europeans and Africans as well as the pigmentation phenotypes associated with these population groups.

4.2.2 Multiplex primer design

Similarly to Multiplex 1 to 4, PCR and SBE primers were designed for each of the Multiplex 5 candidate SNPs using the design criteria outlined in sections 2.2.1.1 and 2.2.1.2. The final Multiplex 5 PCR and SBE primer sequences are provided in Table

2.1c and Table 2.2b respectively. The PCR primers were designed so that all amplicons are less than 200bp in size to ensure that the assay would be applicable to the analysis of degraded and compromised DNA samples. The efficiency of multiplex PCR reactions is improved when amplicons are similar sizes as these products will generally require similar amplification conditions, thereby increasing the likelihood of obtaining balanced amplification of all PCR products in the multiplex (Sanchez *et al.*, 2005). For this reason, the minimum 7bp spacing requirement between PCR products was not maintained and a few Multiplex 5 PCR products overlap in size. Whilst this prevents the assessment of individual PCR products yields in the multiplex as the amplicons cannot be resolved in a 4% agarose gel, preliminary optimisation of PCR primer concentrations is achievable by grouping the PCR primers in smaller multiplexes. The final PCR primer concentration optimisations for the 13-Plex PCR reaction can then conducted in conjunction with the optimisation of the 13-Plex SNaPshot™ reaction. The SBE primers were manually designed and both the Multiplex 5 PCR and SBE primer sets were tested for non-specific interactions including hairpin and dimers in the same way as for Multiplex 1 to 4 (section 3.2.3).

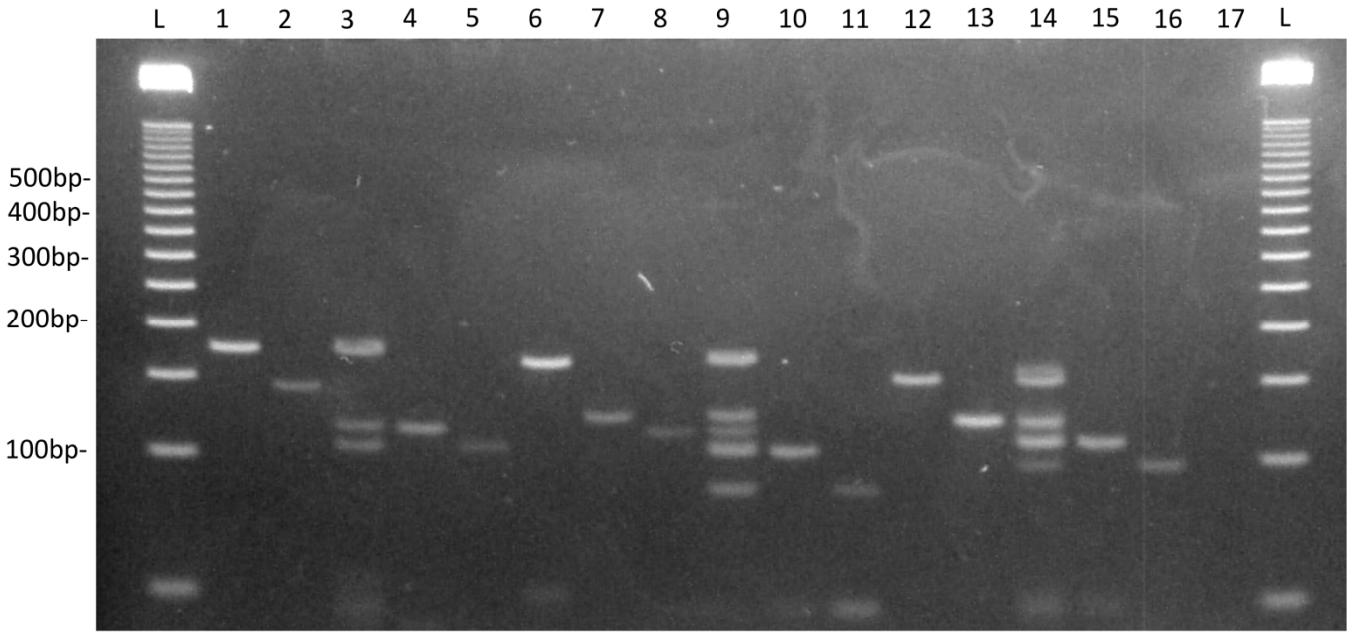
4.3 Results and Discussion

4.3.1 Multiplex 5 assay optimisation

4.3.1.1 *Preliminary optimisation of the Multiplex 5 PCR reaction*

A similar optimisation process to that used for Multiplexes 1 to 4 was employed for Multiplex 5. Each Multiplex 5 PCR primer pair was individually amplified in singleplex PCR reactions to test the specificity of the designed PCR primers. The PCR reactions were carried out using the optimised PCR conditions and cycling parameters established for Multiplex 1 (section 2.2.2.2), except that initial PCR primer concentration were 0.2µM.

Visualisation of the singleplex PCR products in a 4% (w/v) agarose gel confirmed the successful and specific amplification of each Multiplex 5 PCR primer pair (Figure 4.2) with all products of expected size and generally in the absence of non-specific products. However, SLC24A5, HERC2, OCA2-4 and F13B showed some evidence of primer



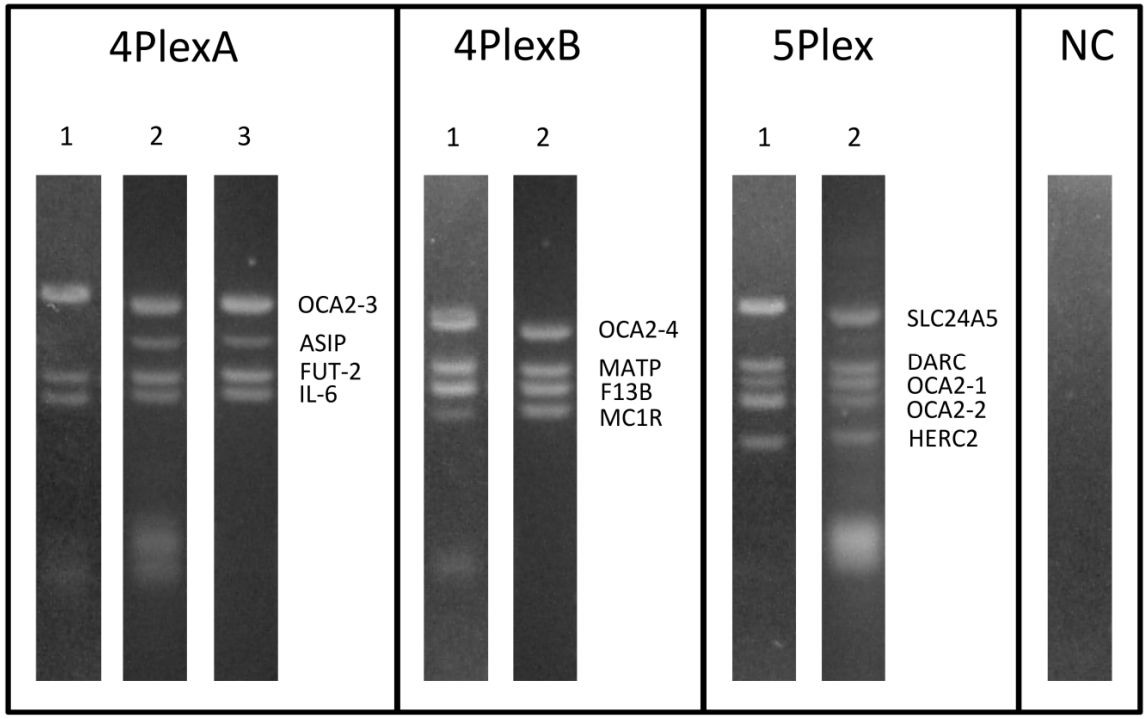
Legend:
L: 50bp ladder
1: OCA2-3 (173bp)
2: ASIP (141bp)
3: 4PlexA (OCA2-3, ASIP, FUT2, IL-6)
4: FUT2 (116bp)
5: IL-6 (105bp)
6: SLC24A5 (164bp)
7: DARC (121bp)
8: OCA2-1 (115bp)
9: 5Plex (SLC24A5, DARC, OCA2-1, OCA2-2, HERC2)
10: OCA2-2 (103bp)
11: HERC2 (85bp)
12: OCA2-4 (150bp)
13: MATP (120bp)
14: 4PlexB (OCA2-4, MATP, F13B, MC1R)
15: F13B (110bp)
16: MC1R (97bp)
17:13-plex PCR negative control

Figure 4.2: Multiplex 5 singleplex and multiplex PCR products
Multiplex 5 PCR primer pairs were amplified at equimolar concentrations of 0.2µM in singleplex, 4-Plex, and 5-Plex reactions using optimised Multiplex 1 PCR conditions summarised in section 2.2.2.2.

dimers at the 50bp position, indicating the presence of mild non-specific interactions between the forward and reverse primer in these pairs. Nevertheless, these products showed strong yields so primer dimer formation was attributed to an excess primer concentration in the singleplex PCR reaction and were not considered to be problematic.

The Multiplex 5 PCR primers were then grouped in small multiplexes to test for non-specific interactions between primer pairs in the set. Since the Multiplex 5 PCR products were designed to be less than 200bp, some products overlap in size and cannot be resolved through agarose gel electrophoresis. For this reason, the PCR primers of the complete Multiplex 5 13-Plex PCR were separated and amplified in 4-PlexA (OCA2-3, ASIP, FUT-2, IL-6), 4-PlexB (OCA2-4, MATP, F13B, MC1R) and 5-Plex (SLC24A5, DARC, OCA2-1, OCA2-2, HERC2) PCR reactions using an equimolar PCR primer concentration of 0.2µM and the optimised Multiplex 1 reagent concentrations and cycling parameters described in section 2.2.2.2. All expected sized products were observed in each of these multiplexes, with the exception of ASIP, which failed to amplify in 4-PlexA at a concentration of 0.2µM (Figure 4.2).

As depicted in Figure 4.2, mild variability in PCR product yields was also observed within each of these multiplexes. Preliminary optimisations of the PCR product yields were carried out by adjusting PCR primer concentrations to obtain relatively balanced amplification of Multiplex 5 PCR products within the 4-PlexA, 4-PlexB and 5-Plex PCR reactions. The failure of ASIP to amplify at 0.2µM in 4-PlexA, as shown in Figure 4.2, was overcome by increasing the PCR primer concentration for this locus to 0.3µM. The ASIP loci could be successfully amplified at this higher concentration, however, the yield for this PCR product was lower than that of the OCA2-3, FUT2 and IL-6 products in 4-PlexA (Figure 4.3, 4-PlexA- Lane 2). The ASIP PCR primer concentration was increased further to 0.4µM, which effectively equalised the yields of all four PCR products in 4-PlexA (Figure 4.3, 4-PlexA-Lane 3). Only minor PCR primer optimisations were required for the 4-PlexB and 5-Plex, as MC1R and OCA2-1 only showed slightly reduced yields respectively. The PCR primer concentration for these loci was increased to 0.3µM and achieved relatively balanced amplification of all products within both the 4-PlexB and 5-Plex PCR reactions (Figure 4.3, 4-PlexB/ 5-Plex-Lane 2). These preliminary optimised PCR primer concentrations determined for



Legend:

4-PlexA:
1: Equimolar PCR primer concentration = 0.2µM
2: All PCR primers = 0.2µM except ASIP= 0.3µM
3: All PCR primers = 0.2µM except ASIP=0.4µM

4-PlexB:
1: Equimolar PCR primer concentration = 0.2µM
2: All PCR primers = 0.2µM except MC1R= 0.3µM

5-Plex:
1: Equimolar PCR primer concentration = 0.2µM
2: All PCR primers = 0.2µM except OCA2-1= 0.3µM

NC: 13-Plex PCR negative control

Figure 4.3: Preliminary optimisation of Multiplex 5 PCR primer concentrations.

4-PlexA, 4-PlexB and 5-Plex were then tested as a 13-plex PCR reaction, which was used as template for the optimisation of the Multiplex 5 SNaPshot™ reaction.

4.3.1.2 *Optimisation of the Multiplex 5 SNaPshot™ reaction*

4.3.1.2.1 *Singleplex extension of Multiplex 5 SBE primers using preliminary Multiplex 5 PCR product as template.*

As for Multiplex 1 to 4, the electrophoretic mobility of the Multiplex 5 SBE primers was initially tested using the SNaPshot® Primer Focus™ kit as described in section 2.2.1.3 prior to commencing optimisation of the SNaPshot™ reaction. Overall, adequate spatial separation of the Multiplex 5 alleles was observed with no obvious overlap in size and colour (data not shown). However, a potential overlap between the OCA2-1*C and MATP*C allele was identified, with less than 1nt difference detected through the Primer Focus™ results for these alleles. Prior to redesigning these SBE primers, their mobility within the multiplex reaction was first examined to see if the alleles could be resolved as separate peaks despite the minute sizing difference.

The first stage of Multiplex 5 SNaPshot™ optimisation involved the preliminary extension of Multiplex 5 SBE primers in singleplex SNaPshot™ reactions. These and subsequent SNaPshot™ reactions in this chapter were prepared as detailed in section 2.2.2.5 and purified and separated using capillary electrophoresis as outlined in sections 2.2.2.6 and 2.2.2.7. The SBE primers were individually extended at a concentration of 0.2µM using Multiplex 5 PCR product as template. The 13-Plex PCR product template was prepared by combining the optimised PCR primer concentrations for 4-PlexA, 4-PlexB and 5-plex in a single reaction described in section 4.3.1.1. All other PCR reagent concentrations and cycling parameters used were the same as for the optimised Multiplex 1 PCR described in section 2.2.2.2. Visualisation of the singleplex SNaPshot™ products in POP-6™ polymer confirmed the specificity of the SBE primers as all alleles were detected to be the expected size and colour except for ASIP, which failed to extend (Figure 4.4). Variability in peak heights was observed between alleles, particularly for the heterozygous loci such as OCA2-1, OCA2-3 and OCA2-4 where both alleles were detected at about half the average 1500RFU peak height (Figure 4.4). The OCA2-2, FUT-2 and IL-6 alleles showed the strongest signals.

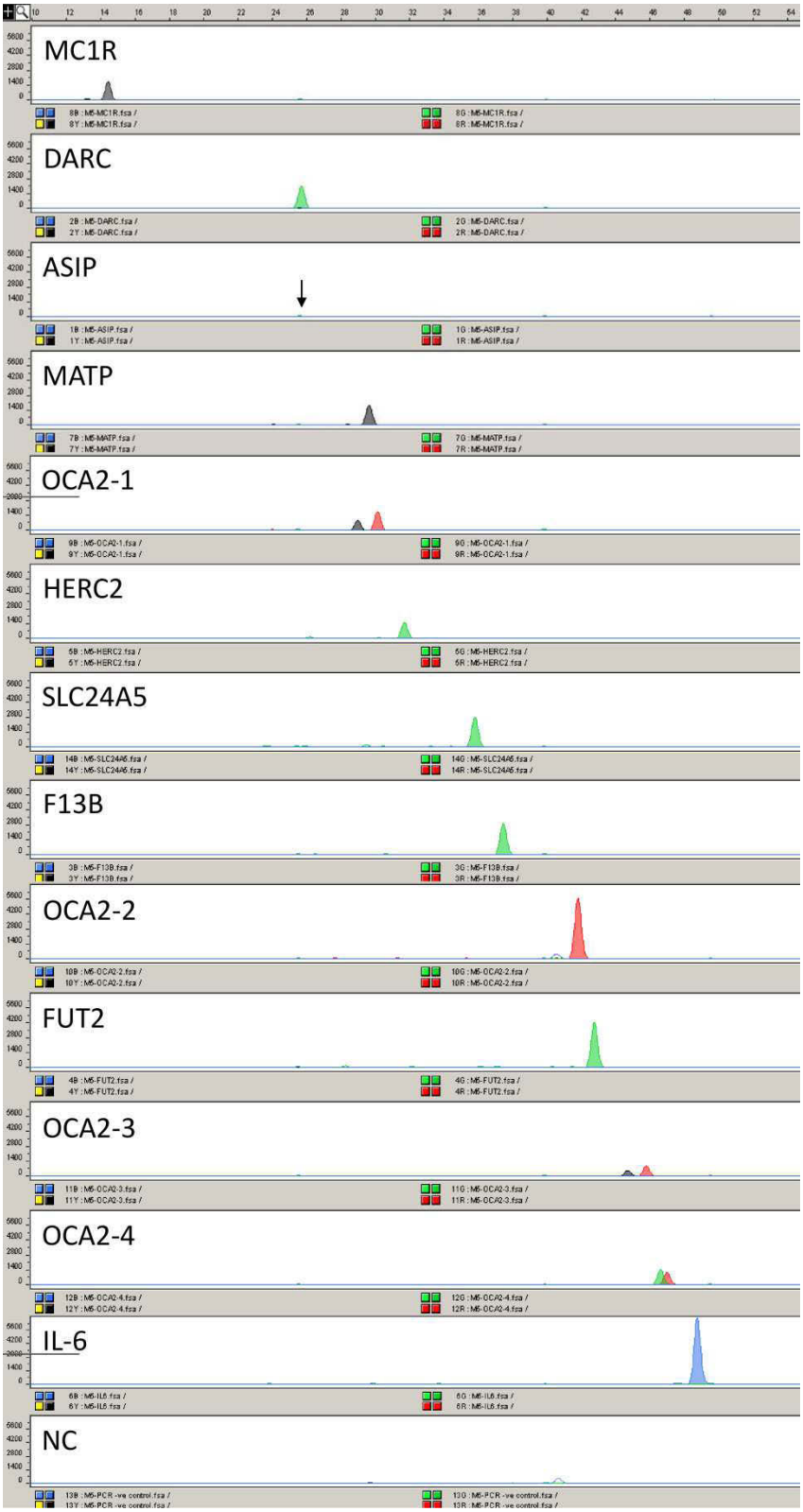


Figure 4.4: Capillary electropherograms of the singleplex extension of 0.2µM Multiplex 5 SBE primers using Multiplex 5 PCR product as template.

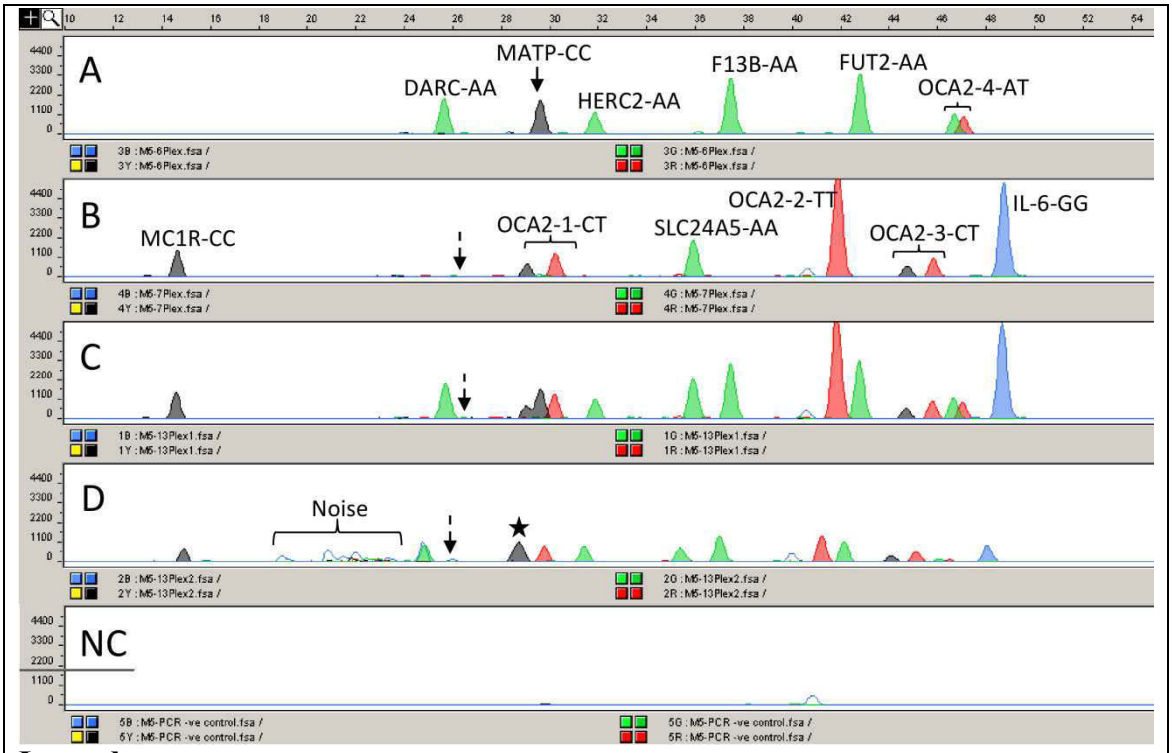
Multiplex 5 13-Plex PCR product was prepared by combining the optimised conditions and PCR primer concentrations for 4-PlexA, 4-PlexB and 5-plex described in Section 4.3.1.1. Black arrows indicate the expected position for allele ASIP that failed to extend. The 13-Plex PCR negative control is also shown (**NC**).

4.3.1.2.2 *Multiplex extension of Multiplex 5 SBE primers using preliminary Multiplex 5 PCR product as template*

The Multiplex 5 SBE primers were tested in 6-Plex, 7-Plex and 13-Plex SNaPshot™ reactions at an equimolar concentration of 0.2µM. The same Multiplex 5 PCR product for the singleplex SNaPshot™ reactions was also used as a template (refer to Section 4.3.1.2.1). All expected alleles were detected in these multiplex SNaPshot™ reactions in the absence of problematic non-specific interactions (Figure 4.5A-C) with the exception of ASIP which again failed to extend in both the 7-Plex (Figure 4.5B) and 13-Plex (Figure 4.5C). The OCA2-1*C and MATP*C alleles, which were identified as potentially overlapping through the Primer Focus™ results, could be resolved as separate peaks, with the MATP-C peak detected between the OCA2-1*C and OCA2-1*T peaks (Figure 4.5C). However, the Multiplex 5 allele signals were observed to vary dramatically between different loci. The heterozygous alleles showed the lowest peak heights (1000RFU) whilst homozygous loci signals varied between 1600RFU and 5000RFU.

The Multiplex 5 SBE primer concentrations in the 13-Plex were modified with the aim of achieving balanced allele signals and the successful extension of the ASIP allele(s). The failure of the ASIP SNP to extend was hypothesised to have resulted from using an inadequate SBE primer concentration and was consequently increased to 0.4µM. The SBE primer concentration for other low signal loci including OCA2-1, HERC2- OCA2-3 and OCA2-4 was also increased to 0.3µM, whilst the higher signal OCA2-2 and IL-6 alleles were reduced to 0.1µM. These SBE primer modifications did not appear to have a positive effect on ASIP, which still failed to extend at 0.4µM (Figure 4.5D). Although, relatively balanced peak heights were observed when using these modified SBE primer concentrations (Figure 4.5D), the overall allele signals were lower than achieved when using an equimolar concentration of 0.2µM (Figure 4.5C). A considerable amount of noise was detected in the 18-24nt region, indicative of inadequate purification of the PCR product used as template for the SNaPshot™ reaction. This was likely caused by an old ExoSAP-It/ExoStar aliquot, which would have contributed to the decreased signals observed. Furthermore, loss of spatial resolution could also be seen in Figure 4.5D with the OCA2-1*C and MATP*C alleles merging into a single black peak (identified as a star) despite having been adequately resolved in Figure 4.5C. The loss of

resolution is typically observed when using older POP-6™ polymer, which may also result in a reduction in allele signal. Consistent with this, the OCA2-1*C and MATP*C alleles could be spatially resolved when using fresh polymer (data not shown). However to achieve better spacing between the OCA2-1 and MATP alleles and to allow these peaks to be distinguished even if the polymer begins to age, another 3nt were added to the aspecific 5' tail of the MATP SBE primer.



Legend:

(A), (B) and (C) 6-Plex, 7-Plex and 13-Plex SBE reactions respectively, using Multiplex 5 13-plex PCR template prepared as described in section 4.3.1.2.1. All SBE primers were extended at equimolar SBE primer concentrations of 0.2µM.

(D) Same PCR product as (C) but the SBE primer concentrations were modified with ASIP=0.4µM, OCA2-1, HERC2, OCA2-3 and OCA2-4= 0.3µM, OCA2-2 and IL-6= 0.1µM and all other primers at 0.2µM.

(NC) 13-plex PCR negative control with SBE primer concentrations used in (D).

Figure 4.5: Capillary electropherograms for the optimisation of Multiplex 5 SNaPshot™ reactions using Multiplex 5 PCR product as template.

Optimisations involved the modification of SBE primer concentrations as detailed in the legend. Dashed arrows show the position of the missing ASIP allele and the black star shows the merged black peak for OCA2-1*C and MATP*C alleles due to sizing overlap.

A small blue artefact peak at the 41nt position preceding the OCA2-2*T allele peak was also widely observed in the singleplex OCA2-2 and multiplex SNaPshot™ reactions as well as in the negative PCR control. It was proposed that this artefact peak arose from the extension of non-specific SBE primer interactions or of PCR template products such as primer dimers. The blue artefact peak was detected in close proximity to the FUT2*G allele (42nt) and at about the same position as the OCA2-2*G allele, as determined from the Primer Focus™ results. To avoid the artefact overlapping in size and colour with surrounding SNaPshot™ products, 1nt was removed from the 5'aspecific tail of the OCA2-2 SBE primer. The modified MATP and OCA2-2SBE primer sequences are provided in Table 2.2b. Prior to the continuation of Multiplex 5 SNaPshot™ reaction optimisation to incorporate the modified MATP and OCA2-2 SBE primers, further troubleshooting experiments (detailed in section 4.3.1.2.3) were also performed to investigate the failed extension of the ASIP allele(s).

4.3.1.2.3 *Investigation of the failed extension of the ASIP SBE primer*

The ASIP allele(s) failed to extend in both singleplex and multiplex SNaPshot™ reactions using Multiplex 5 PCR product as template, despite increasing the SBE primer concentration for this locus (Figure 4.4C and Figure 4.5D). It was hypothesised that insufficient ASIP PCR product within the 13-Plex PCR template was preventing the extension of the ASIP SNP. The various Multiplex 5 PCR products were originally optimised in smaller multiplex PCR reactions (4PlexA, 4PlexB and 5Plex in Figure 4.3) because some PCR products overlapped in size and could not be distinguished in a complete 13-Plex PCR reaction. Though the ASIP PCR product yield appeared satisfactory in the optimised 4-PlexA at a PCR primer concentration of 0.4µM (Figure 4.3, 4-PlexA-Lane 3), non-specific interactions between ASIP and other PCR primers, which were optimised separately within the 4-PlexB or 5-Plex, may have decreased the ASIP product yield within the 13-Plex PCR and would have contributed to the failed extension of the ASIP marker.

To investigate the effect of using variable primer concentrations on the yield of the ASIP product, the ASIP PCR primers were amplified at 0.2µM in singleplex and at 0.4µM and 1µM in 4-PlexA and 13-Plex PCR reactions. The optimised Multiplex 1 reagent concentrations (excluding PCR primer concentrations) and cycling parameters

described in section 2.2.2.2 were used to prepare these PCR reactions. Whilst not all Multiplex 5 PCR products can be resolved in a 4% (w/v) agarose gel due to size overlap, ASIP is sufficiently different to other products to be identified as a single band. ASIP was successfully amplified in singleplex, 4-plexA and 13-Plex PCR reactions, with variations in the ASIP product yield observed when using different PCR primer concentrations (Figure 4.6). The singleplex ASIP yield was fairly moderate at a primer concentration of 0.2µM (Figure 4.6A) but greater yields were observed in the 4-PlexA and 13-Plex PCR reactions using ASIP PCR primer concentrations of 0.4µM and 1µM respectively. The fact that the ASIP product could be visualised in the 13-Plex when using both the preliminary optimised 0.4µM PCR primer concentration and the increased 1µM concentration, indicated that the failed extension of the ASIP loci in SNaPshot™ reactions was not caused by insufficient ASIP PCR product but rather by problems with the ASIP SBE primer itself. This troubleshooting experiment did however demonstrate that an ASIP PCR primer concentration of 1µM should be used instead of 0.4µM to improve the chance of amplifying this locus.

Since the ASIP locus failed to extend in all experiments to date, the ASIP genotype of the donor DNA sample needed to first be confirmed. The ASIP singleplex PCR product was used as template for the singleplex extension of the ASIP SBE primer at a concentration of 0.2µM. A single red peak corresponding to a homozygous ASIP-TT genotype was visualised at the 27nt position (Figure 4.7A) confirming the genotype of the donor and that the ASIP SBE primer could be extended from an ASIP singleplex PCR product. The peak height observed was satisfactory at 2400RFU and the allele detected was within the expected size range and colour and could be discriminated from other Multiplex 5 alleles.

The finding that ASIP could be extended in singleplex products but not in the 13-Plex product suggested the possibility of detrimental interactions between the ASIP SBE primer and one or more of the other SBE primers. When reviewing the Multiplex 5 Autodimer Primer Screen results, the strongest interaction observed between ASIP and all other SBE primers was with the MATP SBE primer (dG (37 ºC)= -2.07 kCal/mole), although stronger interactions have been noted in previous multiplexes without any observable extension problems. To investigate this possibility and the effect of

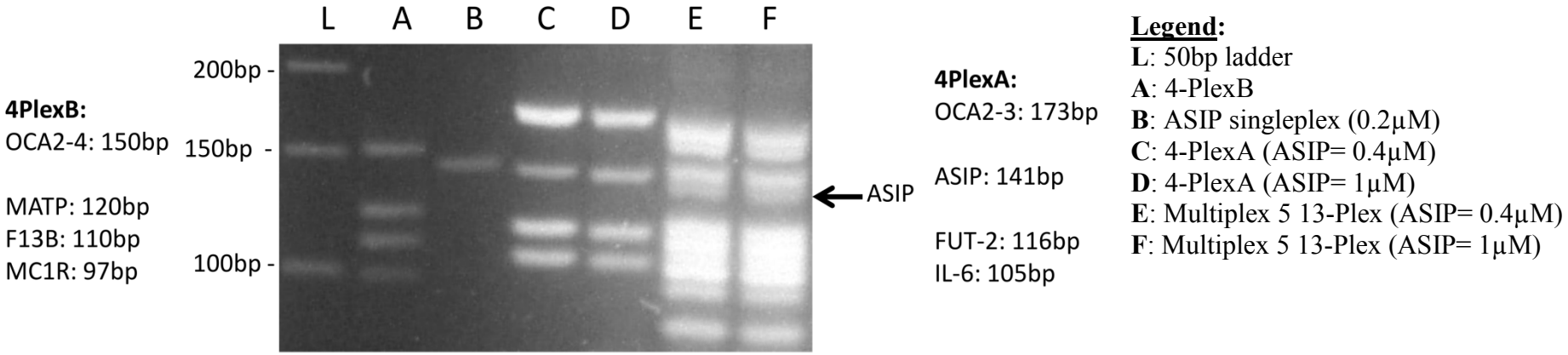
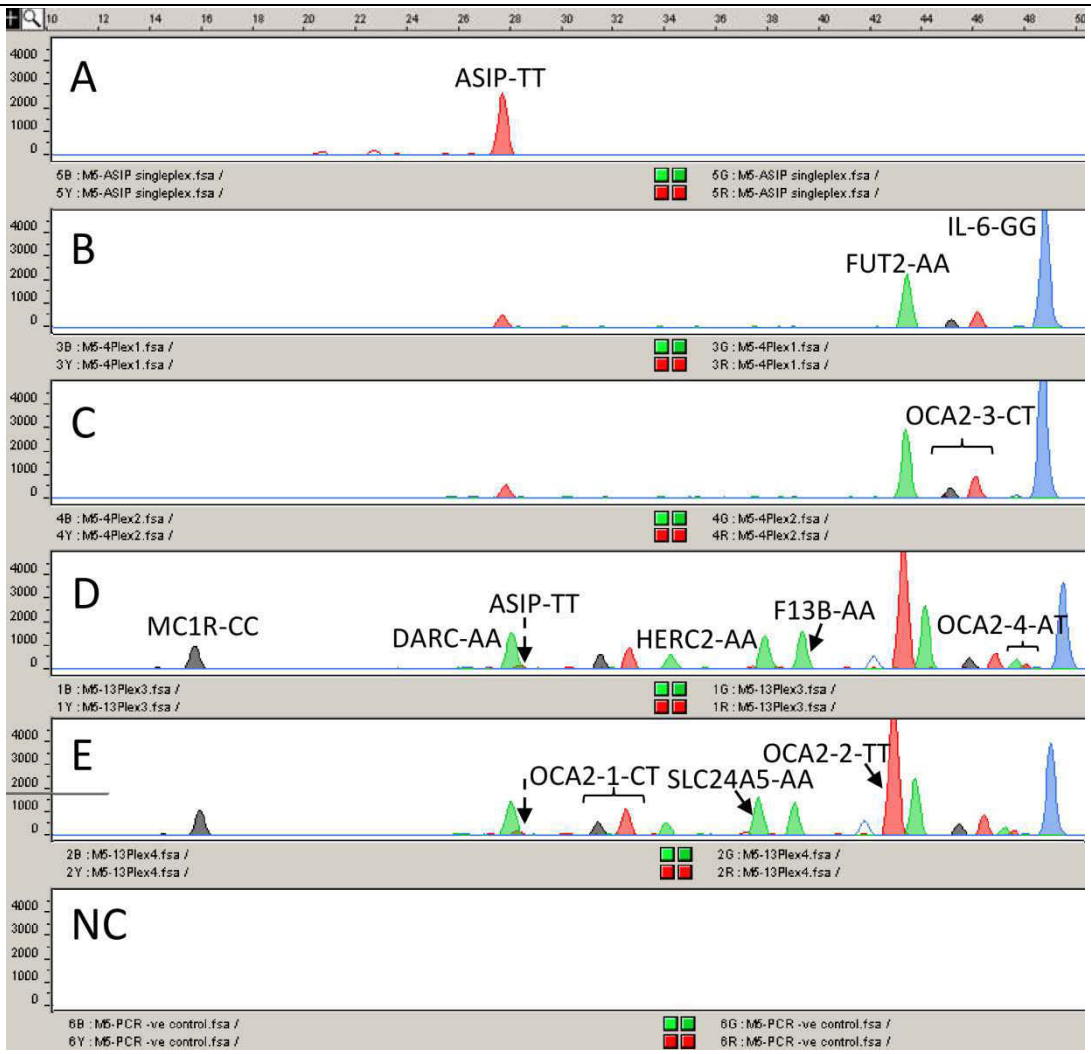


Figure 4.6: Amplification of ASIP PCR primers in singleplex, 4-PlexA and 13-Plex PCR reactions at variable PCR primer concentrations.

The ASIP PCR primers were amplified using the optimised Multiplex 1 PCR conditions summarised in section 2.2.2.2 at variable concentrations as indicated in the legend. The PCR primers in 4-PlexB (OCA2-4, MATP, F13B and MC1R) as well as all other primers in 4-PlexA (OCA2-3, FUT-2, IL-6) and the 13-Plex were amplified at an equimolar concentration of 0.2µM, except for OCA2-1 and MC1R which were amplified at 0.3µM in 4-PlexB and the 13-Plexes.



Legend:

(A) Singleplex ASIP SBE reaction using singleplex ASIP PCR product as template. Both PCR and SBE primer concentrations=0.2µM.

(B) 4-PlexA SBE reaction using 4-PlexA PCR template with ASIP PCR primers at 0.4µM and FUT-2, OCA2-3 and IL-6 at 0.2µM. All SBE primers extended at 0.2µM.

(C) Same as (B) except the ASIP PCR primers in the 4-PlexA PCR template were amplified at 1µM.

(D) 12-Plex SBE reaction excluding the MATP SBE primer extending all other SBE primers at an equimolar concentration of 0.2µM using the Multiplex 5 13-Plex PCR template with OCA2-1 and MC1R PCR primers= 0.3µM and ASIP PCR primers= 0.4µM.

(E) Same as (D) except the ASIP PCR primers in the 13-Plex PCR template were amplified at 1µM.

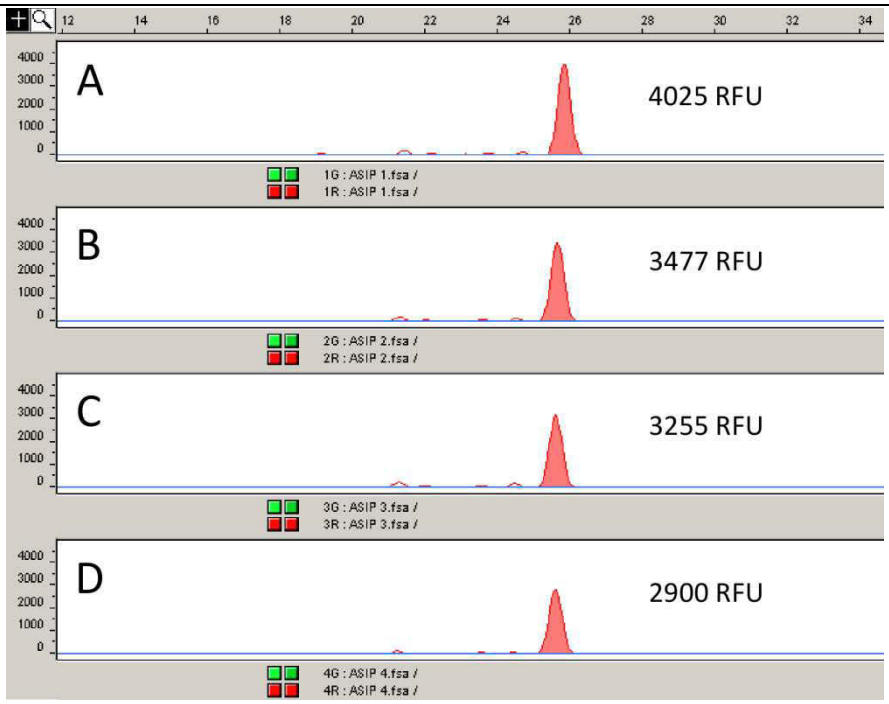
(NC) 13-Plex PCR negative control with SBE primer concentrations used in (E).

Figure 4.7: Capillary electropherograms of the extension of the ASIP SBE primer in singleplex, 4-PlexA and 12-Plex SNaPshot™ reactions using singleplex and multiplex PCR products as template.

Dashed arrows indicate the position of the small ASIP-TT allele peak.

modifying ASIP PCR primer concentration on the extension success of the ASIP locus in multiplex PCR products, the 4-PlexA, (which doesn't contain the MATP amplicon) and the 13-Plex PCR products amplified with ASIP PCR primers at 0.4µM and 1µM were used as template for a 4-PlexA SNaPshot™ reaction and a 12-Plex SNaPshot™ reaction in which the MATP SBE primer was excluded. All SBE primers were tested at an equimolar concentration of 0.2µM. The ASIP allele was successfully extended in 4-PlexA SNaPshot™ reactions at a peak height of 600RFU, with minimal differences in signal observed when using an ASIP PCR primer concentration of 0.4µM (Figure 4.7B) or 1µM (Figure 4.7C). Similar results were observed for the 12-Plex SNaPshot™ reactions irrespective of the ASIP PCR primer concentration used, though the ASIP allele peak heights were now dramatically decreased to less than 300RFU (Figure 4.7D and E) suggesting that problems additional to possible interactions between the ASIP and MATP SBE primers may be present.

The high GC content (73%) of the ASIP PCR amplicon was proposed to be another contributor to the poor extension of the ASIP allele. As the SNaPshot™ cycling protocol does not include an initial denaturation step it was hypothesised that incomplete denaturation of the ASIP DNA template was occurring, thereby preventing the ASIP SBE primer from annealing and reducing the extension efficiency of this locus. A series of singleplex ASIP SNaPshot™ reactions prepared with singleplex ASIP PCR product as template was tested using numerous variations of the SNaPshot™ cycling protocol incorporating an initial denaturation step. However, the highest ASIP peak height was achieved when using the standard SNaPshot™ cycling protocol without a denaturation step (Figure 4.8A). An approximately 600RFU decrease in signal was observed when an initial denaturation of 3 minutes at 96ºC was used (Figure 4.8B). A larger decrease of approximately 800RFU was observed when using an initial denaturation step of 3 minutes at 96ºC plus an extended denaturation of 30s per cycle (Figure 4.8C). The use of PCR adjuvants to facilitate DNA denaturation was also tested by adding 5% DMSO to a singleplex SNaPshot™ reaction which was performed using the standard SNaPshot™ cycling protocol. Thus, though the high GC content of the ASIP PCR amplicon may also be contributing to the reduced extension of this allele, little could be done to improve this.



Legend:
(A) Standard SNaPshot™ protocol (96º- 10s, 50º- 5s, 60º- 30s).
(B) Initial denaturation at 96º for 3 min followed by the standard SNaPshot™ cycling protocol (96º- 10s, 50º- 5s, 60º- 30s).
(C) Initial denaturation at 96º for 3 min followed by 96º for 30s, 50º for 5s, 60º for 30s.
(D) 5% DMSO was added to the singleplex SNaPshot™ reaction and run using the standard SNaPshot™ protocol (96º- 10s, 50º- 5s, 60º- 30s)

Figure 4.8: Optimisation of singleplex ASIP SNaPshot™ reactions using modifications to the SNaPshot™ cycling protocol and the addition of PCR adjuvants.

The ASIP SBE primer was extended in singleplex SNaPshot™ reactions at a concentration of 0.2µM using as template the singleplex PCR product derived from the amplification of 0.2µM of ASIP PCR primers. Modifications to the SNaPshot™ cycling protocol and the addition of PCR adjuvants were tested as described in the legend.

It appears that a combination of factors may be preventing the successful extension of the ASIP SNP in the multiplex SNaPshot™ reaction. These may include non-specific interactions between the ASIP SBE primer and other Multiplex 5 SBE primers such as MATP as well as the high GC content of the ASIP PCR product. Furthermore, potentially problematic non-specific binding between the ASIP SBE primer and the DARC and IL-6 PCR products were identified when both amplicon strands were aligned against the ASIP SBE primer sequence (data not shown). This suggests the

possibility that the ASIP SBE primer is annealing non-specifically to other PCR templates in the Multiplex 5 PCR product (including the DARC and IL-6 PCR amplicons). Whilst these non-specific interactions could not be extended as there was no homology at the 3' end, these would still reduce the amount of ASIP SBE primer available to bind to the correct template and thus, would reduce the extension success of the ASIP SNP. Multiplexing introduces competition for reagents including SBE primers and the multiplex PCR product provides numerous amplicon templates, which also vary in their base composition and thus, extension efficiency. This can result in unbalanced SNP extension with more efficient templates dominating the SNaPshot™ reaction at the detriment of more problematic templates. Non-specific binding of the SBE primer to other templates can also exacerbate the unbalanced extension of loci by reducing the available SBE primer concentration.

4.3.1.2.4 *Redesigning the ASIP SBE primer*

Typing ASIP separately in a singleplex SNaPshot™ reaction using singleplex ASIP PCR product as template would diminish the practicality of our final multiplex SNP genotyping assay to forensic casework. The decision was made to redesign the ASIP SBE primer to attempt to extend this SNP within the Multiplex 5 SNaPshot™ reaction. The current ASIP SBE primer was designed to be homologous to the antisense strand. A study investigating the role of MC1R and ASIP polymorphisms in melanoma and carcinoma development, designed a SBE primer homologous to the sense strand to successfully genotype the ASIP 8818A>G (rs6058017) polymorphism in a singleplex SNaPshot™ reaction (Brudnik *et al.*, 2009). However, this ASIP SBE primer was 32nt in size and could not be incorporated into our Multiplex 5 assay as it would overlap in size and colour with the detection of the SLC24A5 and F13B alleles. The Brudnik *et al.* (2009) primer was used as model to design another ASIP SBE primer 17nt in length which was also homologous to the sense strand except that no 5' aspecific tail was added. However, Autodimer results for this alternate ASIP SBE primer demonstrated major self-primer interactions (Figure 4.9A), hence to reduce the severity of the primer self-annealing, a mismatch was inserted by the substitution of a 'G' nucleotide located at the centre of the primer sequence with a 'T' nucleotide (Figure 4.9B). Autodimer results for the hairpin screen showed no major hairpin formations (data not shown). The redesigned ASIP SBE primer was also aligned against all Multiplex 5 PCR amplicon

sequences, with no potentially problematic non-specific binding to other templates observed (data not shown). The modified ASIP SBE primer sequence can be found in Table 2.2c.

(A) Autodimer Primer screen of ASIP-Brudnik 2009 SBE primer

```
Matches = 12
Score = 8
NCTCNCGGCCGNGAGN
est. tm =  37.9°C
DeltaG 37 degrees = -9.60 kcal/mole

                5'-ACTCCCGGCCGCGAGC-3'
                   x|||x|||||x|||x
                3'-CGAGCGCCGGCCCTCA-5'
```

(B) Autodimer Primer screen of modified ASIP-Brudnik 2009 SBE primer:

```
Matches = 10
Score = 4
NCTCNCGNNCGNGAGN
est. tm =  1.8°C
DeltaG 37 degrees = -2.84 kcal/mole

                  5'-ACTCCCGTCCGCGAGC-3'
                     x|||x||xx||x|||x
                  3'-CGAGCGCCTGCCCTCA-5'


Matches = 6
Score = 2
GNNCGCGNNC
est. tm =  less than zero
DeltaG 37 degrees = -0.06 kcal/mole

                  3'-CGAGCGCCTGCCCTCA-5'
                      |xx||||xx|
            5'-ACTCCCGTCCGCGAGC-3'
```

Figure 4.9: Autodimer primer screen results for (A) the unmodified ASIP SBE primer and (B) the modified ASIP SBE primer with inserted mismatch adapted from Brudnik *et al.* 2009.

The red coloured and underlined 'G' nucleotide in (A) denotes the site of the inserted mismatch through substitution of this base with a 'T' nucleotide (also coloured red and underlined) in the modified ASIP SBE primer shown in (B).

4.3.1.2.5 *Singleplex extension of modified Multiplex 5 SBE primers using Multiplex 5 PCR product as template*

The modified ASIP, MATP and OCA2-2 SBE primers were extended at a concentration of 0.2µM in singleplex SNaPshot™ reactions using Multiplex 5 PCR product as template. The modified ASIP SBE primer was also extended at a concentration of 1µM. The multiplex 5 PCR product template was prepared using the optimised Multiplex 1 reagent concentrations and cycling parameters described in section 2.2.2.2. Multiplex 5 PCR primers were amplified at an equimolar concentration of 0.2µM, except for OCA2-1 and MC1R at 0.3µM and ASIP at 0.4µM. As shown in Figure 4.10, all modified SBE primers could be successfully extended at a concentration of 0.2µM in singleplex SNaPshot™ reactions on multiplex PCR product including ASIP, suggesting that the new ASIP primer sequence with the inserted mismatch improved the extension success of this SNP. The ASIP peak height achieved when using a concentration of 0.2µM was unsatisfactorily low at 200RFU (Figure 4.10A) however, increasing the ASIP SBE primer concentration to 1µM dramatically improved the extension success of the SNP, with a signal of 800RFU observed (Figure 4.10B).

The peak heights of MATP and OCA2-2 averaged more than 4000RFU with the latter showing mild over-amplification with a slightly split peak (Figure 4.10D). It was considered highly likely that the SBE primer concentration for these SNPs would need to be reduced during optimisation of the multiplex SNaPshot™ reaction. The modifications to the 5' aspecific tail of these two primers successfully altered the observed size of the detected alleles so that they are now sufficiently distant from surrounding SNaPshot™ products to be unambiguously detected. However, this was also further verified through incorporation of the modified SBE primers in Multiplex 5 SNaPshot™ reactions (see Section 4.3.1.2.6).

The blue artefact peak that was previously detected at the 41nt position, preceding the OCA2-2*T allele, also shifted and was now detected at the 40nt position (Figure 4.10D). This strongly indicated that the artefact peak arose from the non-specific extension of the OCA2-2 SBE primer, which had been modified to include 1nt less in the 5'aspecific tail. Autodimer results did not show any major hairpin or primer dimer formation for this primer (data not shown), however, since this artefact was also

detected in the 13-Plex PCR negative control (Figure 4.10E), it was believed that the OCA2-2 SBE primer was extending after non-specifically binding to primer dimer products in the Multiplex 5 PCR product used as template. The formation of this artefact is problematic as it is the same colour and position as the alternate OCA2-2 allele. Nonetheless, the artefact occurred at a much lower signal compared to the OCA2-2*T allele peak so it was possible to genotype the donor of the DNA sample as being homozygous at this locus. It was also proposed that the optimisation of the OCA2-2 SBE primer concentration to reduce over-amplification observed in the singleplex reaction would also reduce the incidence of the artefact peak, which was likely arising as a result of excess OCA2-2 SBE primer.

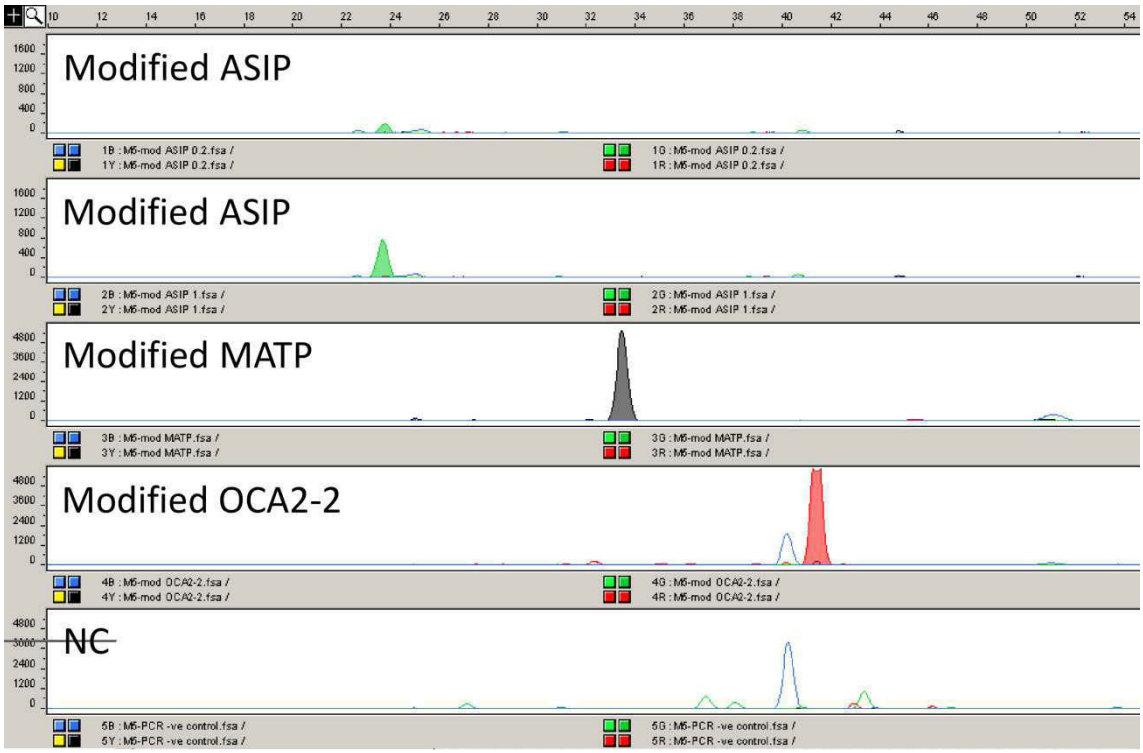


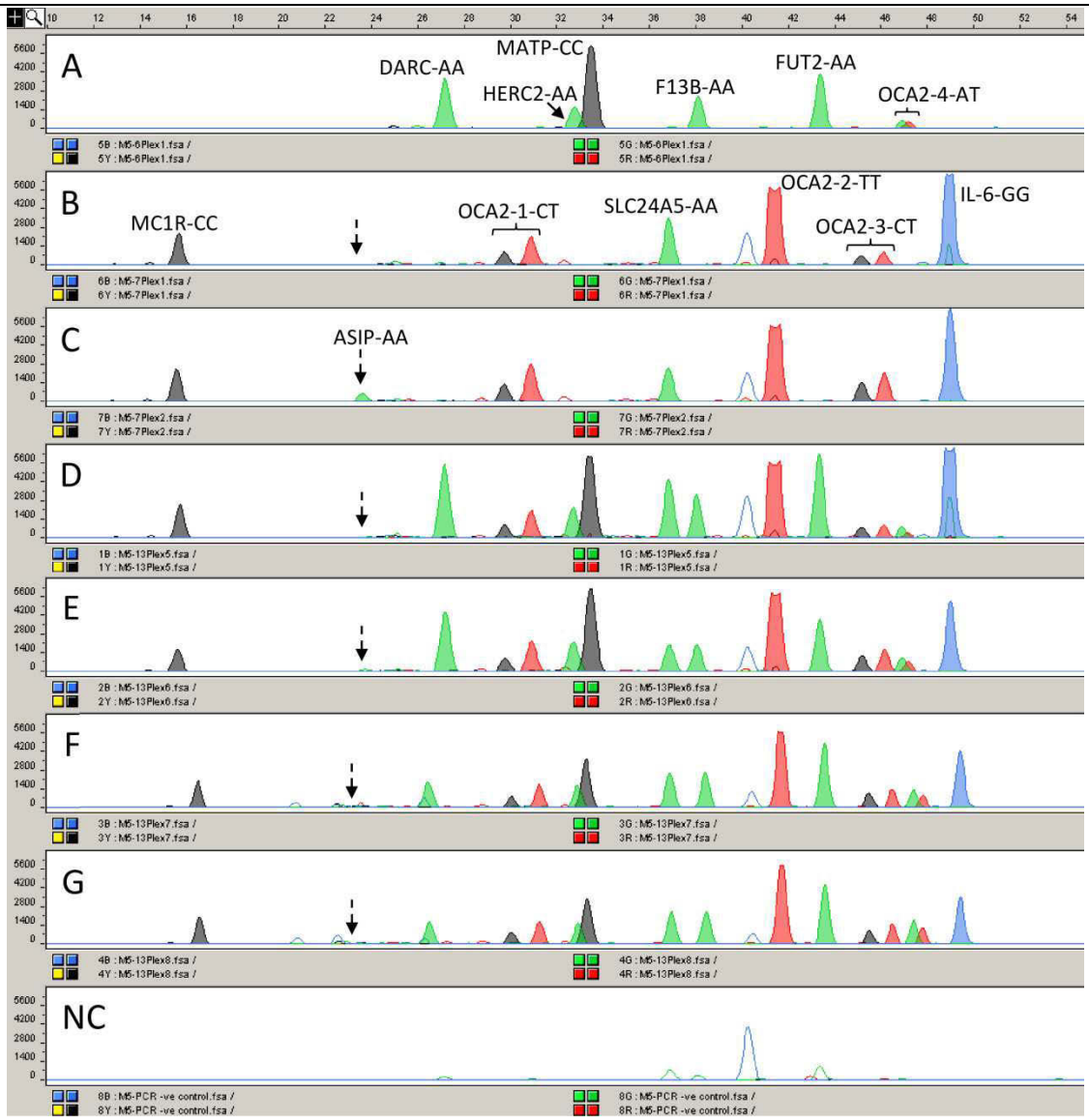
Figure 4.10: Capillary electropherograms of the singleplex amplification of the modified Multiplex 5 SBE primers using Multiplex 5 PCR product as template.

The Multiplex 5 PCR product was prepared using the optimised conditions described in section 2.2.2.2 and an equimolar PCR primer concentration of 0.2µM with the exception of OCA2-1 and MC1R at 0.3µM and ASIP at 0.4µM. All modified SBE primers were extended at a concentration of 0.2µM. The modified ASIP SBE primer was also extended at a concentration of 1µM. The 13-Plex PCR negative control is also shown (**NC**).

4.3.1.2.6 *Incorporation of the modified Multiplex 5 SBE primers into the Multiplex 5 SNaPshot™ reaction*

The modified ASIP, MATP and OCA2-2 SBE primers were grouped into 6-Plex (DARC, HERC2, MATP, F13B, FUT2, OCA2-4), 7-Plex (MC1R, ASIP, OCA2-1, SLC24A5, OCA2-2, OCA2-3, IL-6) and 13-Plex SNaPshot™ reactions using the same Multiplex 5 PCR product as template as was used for the singleplex SNaPshot™ reactions (Figure 4.10). The SBE primers in the multiplex SNaPshot™ reactions were initially tested at an equimolar concentration of 0.2µM and all alleles were detected in the 6-Plex, 7-Plex and 13-Plex SNaPshot™ reactions using Multiplex 5 PCR product as template, with the exception of ASIP (Figure 4.11A, B, D). This was expected as ASIP already showed very low signal when extended in singleplex SNaPshot™ reaction at a SBE primer concentration of 0.2µM (Figure 4.10A). However the modified MATP and OCA2-2 SBE primers demonstrated improved peak resolution and spacing from surrounding allele peaks, in particular MATP could now be unambiguously distinguished from OCA2-1. In the 13-Plex, the peak height of all alleles detected varied across the loci with MATP-CC, FUT2-AA, IL-6-GG and OCA2-2-TT allele peaks showing the strongest signals and the latter demonstrating a split peak, a sign of over extension (Figure 4.11D). The heterozygous OCA2-3-CT and OCA2-4-AT loci showed the lowest peak heights (~700RFU) and the homozygous HERC2-AA locus was detected at less than half the peak height (2100RFU) of the other stronger homozygous peaks, which ranged from 4200-7000RFU.

To achieve more balanced amplification of all loci, SBE primer concentration modifications were performed. It must be noted however that because Multiplex 5 is typing diploid autosomal SNPs, differences in peak height between homozygous (2 copies of the same allele) and heterozygous loci (1 copy of each allele) are to be expected. Thus, the aim was to achieve adequate peak heights where the homozygous allele peaks were roughly double the height of the heterozygous allele peaks. The SBE primer concentration for OCA2-3 and OCA2-4, which showed the lowest peak height, was doubled to 0.4µM and the concentration for OCA2-1 and HERC2 was slightly increased to 0.3µM in the 13-Plex. The concentration for ASIP was also increased to 1µM in both 7-Plex and 13-Plex SNaPshot™ reactions, as this SBE concentration achieved a moderate peak height in singleplex SNaPshot™ reaction (Figure 4.10B).



Legend:

(A) and **(B)** 6-Plex and 7-Plex SBE reactions respectively, using Multiplex 5 13-plex PCR template prepared as described in section 4.3.1.2.5. All SBE primers = 0.2µM.

(C) Same as (B) except for the following SBE primer modifications: ASIP= 1µM, OCA2-1 and HERC2= 0.3µM, OCA2-3 and OCA2-4= 0.4µM and all others= 0.2µM.

(D) 13-Plex SBE reaction prepared using the same 13-Plex PCR template in (A). All SBE primers= 0.2µM.

(E) Same as (D) except for the following SBE primer modifications: ASIP= 1µM, OCA2-1 and HERC2= 0.3µM, OCA2-3 and OCA2-4= 0.4µM and all others= 0.2µM.

(F) Same as (E) except DARC, MATP, OCA2-2 and IL-6 SBE primers= 0.1µM.

(G) PCR primer concentrations were modified with MC1R and HERC2= 0.3µM, OCA2-1 and OCA2-4=0.4µM. The same SBE concentrations as (F) were used.

(NC) 13-Plex PCR negative control with SBE primer concentrations used in (G).

Figure 4.11: Capillary electropherograms for the optimisation of the Multiplex 5 13-Plex SNaPshot™ reaction incorporating the modified SBE primers and using Multiplex 5 PCR product as template.

Optimisations involved the modification of PCR and/or SBE primer concentrations as detailed in the legend

The effect of these SBE primer modifications is evident in both the 7-Plex and 13-Plex SNaPshot™ reactions where increases in peak heights were observed for both OCA2-1 and OCA2-3 (Figure 4.11C and Figure 4.11E). However, the modified ASIP SBE primer could only be extended at a concentration of 1µM in the 7-Plex SNaPshot™ reaction, where it was detected at a low peak height of approximately 700RFU compared to the other alleles (Figure 4.11C). The ASIP primer failed to extend in the 13-Plex even at this elevated concentration. As observed in the singleplex SNaPshot™ reaction (Figure 4.10B), this SNP locus is much less efficiently extended compared to other SNPs, even when using the modified SBE primer. This low efficiency is likely to be exacerbated in the 13-Plex where there is a large number of SBE primers working under the same conditions (Figure 4.11E). Increasing HERC2 to 0.3µM appeared to have minimal effects on the peak height however increasing OCA2-4 also improved the signal for these heterozygous peaks in the 13-Plex (Figure 4.11E).

As evident in Figure 4.11E, the DARC, MATP, OCA2-2 and IL-6 allele peaks showed strong peak heights greater than 4200RFU. Consequently the SBE primer concentration for these loci was reduced to 0.1µM to determine whether this modification would have a positive effect on other less efficiently amplifying loci including ASIP. The peak heights for IL-6 and OCA2-2 satisfactorily decreased to 4200-6000RFU, though the latter still showed slight over-amplification and could be reduced further (Figure 4.11F). In contrast, the MATP and DARC allele peak heights decreased too much, indicating that the higher SBE primer concentration might need to be maintained for these loci. The ASIP allele peak still failed to be detected when the DARC, MATP, OCA2-2 and IL-6 SBE primer concentrations in Figure 4.11F were reduced. The heterozygous OCA2-1 and OCA2-4 allele peaks were observed at approximately 1400RFU, whilst the homozygous MC1R and HERC2 peaks were slightly low with a peak height averaging 1500RFU, compared to other homozygous peaks (>3000RFU).

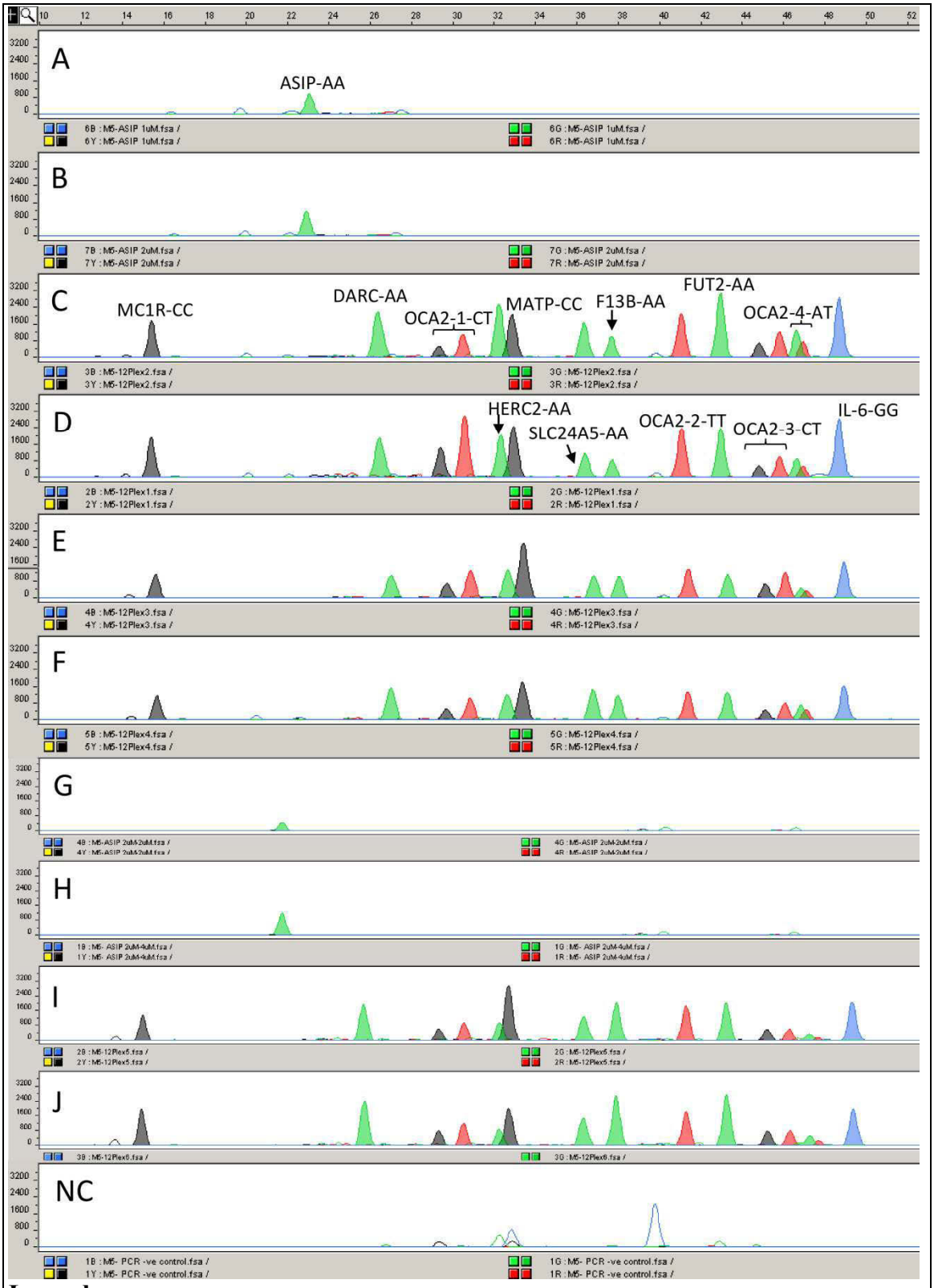
Since the OCA2-1, OCA2-4, MC1R and HERC2 extensions were already being performed at moderately high SBE primer concentrations of 0.3-0.4µM the PCR primer concentration for these loci was increased in an attempt to slightly improve the peak height of the alleles. The MC1R and HERC2 PCR primers were increased to 0.3µM whilst the OCA2-1 and OCA2-4 PCR primers were increased to 0.4µM. These PCR primer modifications achieved only minimal improvements in the peak height of

OCA2-1 and OCA2-4 compared to MC1R and HERC2, which showed no difference (Figure 4.11G). Improvement of the homozygous peak heights was still required for MC1R, DARC and HERC2, whilst the OCA2-2 signal remained slightly elevated compared to other homozygous peaks.

Consistent failure of the modified ASIP SBE primer to be extended in 13-Plex SNaPshot™ reactions even at a concentration of 1µM (Figure 4.11E,F) led to the decision to exclude this SBE primer from the Multiplex 5 SNaPshot™ reaction. The ASIP SBE primer will instead be extended individually in a singleplex SNaPshot™ reaction using Multiplex 5 PCR product as template, (see section 4.3.1.2.8) as this was the only way the allele could be detected (Figure 4.10B). The Multiplex 5 12-Plex SNaPshot™ reaction containing all other SBE primers except for ASIP also required optimisation to ensure that balanced amplification between heterozygous and homozygous peaks was ultimately achieved.

4.3.1.2.7 *Optimisation of 12-Plex SNaPshot™ reactions excluding ASIP*

The Multiplex 5 PCR product tested in Figure 4.11G was used as template for 12-Plex SNaPshot™ reactions containing all Multiplex 5 SBE primers except ASIP. The SBE primer concentrations were mainly the same as those tested in Figure 4.11F and Figure 4.11G with the exception of MC1R, which was increased from 0.2µM to 0.3µM and HERC2, which was increased from 0.3µM to 0.4µM. The SBE primer concentration for the high signal OCA2-2 locus was also reduced from 0.1µM to 0.05µM. As shown in Figure 4.12C, these SBE primer concentrations showed improved relative peak heights, with MC1R, HERC2 and OCA2-2 demonstrating balanced signals compared to other previous concentrations (see Figure 4.11G).



Legend:

(A) and **(B)** ASIP singleplex SBE reactions using the same Multiplex 5 13-plex PCR product shown in Figure 4.11G as template. An ASIP SBE primer concentration of 1µM and 2µM was used respectively.

(C) Multiplex 5 12-Plex SBE reaction (excluding ASIP) using the same 13-Plex PCR template as in (A) and (B). All SBE primers at 0.2µM except DARC, MATP, OCA2-2 and IL-6 = 0.1µM, OCA2-1 and HERC2= 0.3µM, OCA2-3 and OCA2-4= 0.4µM.

(D) Same as (C) except SBE primers were modified with MC1R, OCA2-1= 0.4µM, SLC24A5 and F13B=0.25µM, MATP and IL-6= 0.15µM.
(E) Same as (D) except SBE primers were modified with OCA2-1, SLC24A5 and F13B = 0.3µM.
(F) Same as (E) except SBE primers were modified with OCA2-1= 0.25µM, DARC =0.3µM and SLC24A5 and F13B= 0.4µM.
(G) and **(H)** ASIP singleplex SBE reactions using a modified Multiplex 5 13-plex PCR product with ASIP PCR primers= 2µM as template. An ASIP SBE primer concentration of 2µM and 4µM was used respectively.
(I) Multiplex 5 12-Plex SBE reaction (excluding ASIP) using the same PCR template as in (G) and (H) and the same SBE primer concentrations used in (F).
(J) Same as (I) except SBE primers were modified with MATP=0.1µM and OCA2-2= 0.075µM
(NC) 13-plex PCR negative control with SBE primer concentrations used in (J).

Figure 4.12: Capillary electropherograms for the separate optimisation of ASIP singleplex and Multiplex 5 12-Plex SNaPshot™ reactions using Multiplex 5 PCR product as template.
Optimisations involved the modification of PCR and/or SBE primer concentrations as detailed in the legend.

The homozygous peaks were detected at approximately 1600-2500RFU with the exception of the SLC24A5 and F13B peaks, which showed slightly lower signals. To ensure that heterozygous alleles for these loci can be unambiguously detected, the peaks heights of the two expected alleles must be sufficiently high enough. This means that the observed homozygous peak height needs to be a minimum of 2000-3000 RFU to ensure that heterozygous peaks are at least 1000 RFU. The SBE primer concentration for the observed homozygous loci was further increased to achieve slightly greater peak heights. The MC1R SBE primer was increased to 0.4µM, SLC24A5 and F13B to 0.25µM as well as MATP and IL-6 to 0.15µM. The SBE primer concentration for the heterozygous OCA2-3 and OCA2-4 peaks, which showed adequate peak heights relative to the homozygous peaks in Figure 4.12C, were not modified. However, for the OCA2-1 peak, which demonstrated a slightly weaker signal, the SBE primer concentration was increased from 0.3µM to 0.4µM. These SBE primer modifications further improved the relative peak height. All homozygous peaks were approximately 2000RFU, except for the SLC24A5 and F13B peaks, which were still much lower than the other homozygous peaks and required further optimisation (Figure 4.12D). Increasing the SBE primer concentration for OCA2-1 had an adverse effect and increased the peak height too dramatically. The heterozygote imbalance between the

two alleles was exacerbated, with the T allele detected at a greater peak height compared to other homozygous loci in the 12-Plex (Figure 4.12D).

To further optimise on the conditions used in Figure 4.12D, the primer concentration for OCA2-1 was reduced to 0.3µM and primer concentrations of SLC24A5 and F13B increased to 0.3µM, with all other SBE primers in the 12-Plex maintained as in Figure 4.12D. Improvements were observed in relative peak heights, with most homozygous peaks detected at the same level, though the heterozygous OCA2-1 and OCA2-3 peaks still demonstrated higher peak heights relative to other heterozygous loci (Figure 4.12E). The OCA2-1 SBE primer concentration was further reduced to 0.25µM, whilst the concentration for DARC was increased to 0.3µM and those of SLC24A5 and F13B increased to 0.4µM. These SBE primer concentrations showed good results, with the equalisation of the relative peak height for the observed homozygous and heterozygous peaks achieved (Figure 4.12F). The observed peak heights were lower than in Figure 4.12C and D but this was attributed to a loss of fluorescence intensity, typically observed when using an old SNaPshot™ ready reaction mix aliquot. Overall, the PCR and SBE primer concentrations tested in Figure 4.12F were deemed adequate as the peak height proportions were satisfactory and the blue 'G' artefact peak (40nt) associated with the OCA2-2 peak had greatly diminished.

4.3.1.2.8 *Optimisation of the ASIP singleplex SNaPshot™ reaction*

Though a low ASIP peak height of ~800RFU was observed in the singleplex SNaPshot™ reaction when using a 1µM SBE primer concentration (Figure 4.12A), this is not a significant issue as there are no surrounding non-specific artefact peaks that will affect the detection of the ASIP alleles. Further optimisation of the ASIP singleplex SNaPshot™ reaction was attempted by increasing the SBE primer concentration to 2µM using the Multiplex 5 PCR product tested in Figure 4.11G as template. As shown in Figure 4.12A and B, no major differences were observed between these two SBE primer concentrations, with a homozygous peak height of approximately 800-900RFU detected in each case, which was deemed satisfactory.

4.3.1.2.9 *Fine tuning of the Multiplex 5 ASIP singleplex and 12-Plex SNaPshot™ reactions*

The current optimised singleplex ASIP (Figure 4.12B) and 12-Plex SNaPshot™ reactions Figure 4.12F) were tested on five DNA samples of differing biogeographical ancestries however due to likely variations in amount of DNA deposited on the FTA® disc sampled, the ASIP allele failed to be detected or was observed at a very low signal for some of the DNA samples tested (data not shown). For this reason, the Multiplex 5 PCR template was further optimised by increasing the ASIP PCR concentration from 1µM to 2µM to enhance ASIP PCR product formation for lower starting DNA concentration. This Multiplex 5 PCR product was used as template in singleplex ASIP SNaPshot™ reactions and both 2µM and 4µM concentrations of the ASIP SBE primer were tested. These increased PCR and SBE ASIP primer concentrations had a positive effect on the peak height of the ASIP allele which showed a greater peak height of 1100RFU when using 4µM (Figure 4.12H) compared to the previously optimised 2µM SBE concentration (Figure 4.12G).

The Multiplex 5 PCR product with an ASIP PCR primer concentration of 2µM was also used as template in 12-plex SNaPshot™ reactions, applying the same previously optimised SBE primer concentrations as in Figure 4.12F. The increased ASIP PCR primer concentration had a negligible effect on the peak heights of alleles in the 12-Plex, with a slight increase in MATP and a minor imbalance between the SLC24A5 and F13B peaks observed (Figure 4.12I). In testing of different samples using the preliminary optimised Multiplex 5 conditions depicted in Figure 4.12F it was noticed that excessively high signals for the G alleles coupled with pull-up in comparison to other peaks, particularly the C alleles at the MATP locus was sometimes observed (data not shown). Consequently, the SBE primer concentration for MATP was slightly reduced from 0.15µM to 0.10µM to minimise over extension of the G allele peaks and to prevent potential pull up from masking the C allele, which is detected at the same position. Furthermore, the T allele at the OCA2-2 locus showed major heterozygote imbalance compared to the G allele especially in low template DNA samples during preliminary testing (data not shown) and as such the SBE primer concentration for this locus was slightly increased from 0.05µM to 0.075µM. As evident in Figure 4.12G, these minor SBE primer modifications did not have significant effects on the overall

peak heights which still remained relatively equalised, except for the HERC2-AA and F13B-AA homozygote peaks which showed weaker signals compared to other homozygote loci. However, since these loci have an alternate G allele that was detected at a higher signal compared to other nucleotides, the PCR and SBE primer concentrations used in Figure 4.12G were selected as the final optimised conditions for the Multiplex 5 SNaPshot™ reaction. These PCR and SBE primer concentrations are summarised in Table 4.3.

Table 4.3: Final optimised PCR and SBE primer concentrations for Multiplex 5

	SNP	dbSNP ID number	PCR primer concentration (µM)	SBE primer concentration (µM)
MULTIPLEX 5	OCA2-3	rs4778138	0.20	0.40
	SLC24A5	rs1426654	0.20	0.40
	OCA2-4	rs1545397	0.40	0.40
	ASIP	rs6058017	2.0	4.0
	DARC	rs2814778	0.20	0.30
	MATP	rs16891982	0.20	0.10
	FUT2	rs1800021	0.20	0.20
	OCA2-1	rs7495174	0.40	0.25
	F13B	rs6003	0.20	0.40
	IL-6	rs1800795	0.20	0.15
	OCA2-2	rs4778241	0.20	0.075
	MC1R	rs885479	0.40	0.40
	HERC2	rs12913832	0.30	0.40

4.3.2 Initial population testing

The same five DNA samples originating from donors of English, Chinese, Cameroonian, Egyptian and Indian origin that were used for the initial population testing of the haploid SNP assays (See section 3.3.5) were genotyped using the developed Multiplex 5 SNP assay to evaluate peak height variations, which may arise from differing amounts of DNA deposited on the FTA® cards by the volunteer donors.

The Multiplex 5 profiles for the initial population testing consisted of the ASIP singleplex and the 12-Plex SNaPshot™ reactions and are shown in Figure 4.13. As

anticipated, differences in peak heights were evident at all loci especially for the alternate alleles not visualised during optimisation experiments. All five individuals were homozygous for the A allele at the ASIP locus, with the allele peaks detected between 300-700RFU (Figure 4.13). Whilst this peak height range is considered low, these alleles were visualised in singleplex in the absence of any major artefact peaks that would potentially hinder allele detection at this locus. Consequently, this detection range was deemed acceptable for allele calling.

The electropherograms for the 12-Plex SNaPshot™ reactions clearly demonstrate the disparity in the peak heights of each candidate allele at different loci. The G alleles showed greater signals compared to the A alleles at the DARC, F13B and HERC2 loci and to the C alleles at the MATP and IL-6 loci (Figure 4.13). The peak height of the G allele was also greater in comparison to the T allele at the OCA2-2 loci, which is particularly evident through the substantial heterozygote imbalance observed at this locus in Figure 4.13D. Standardised rules for the interpretation of heterozygote imbalances when using SNaPshot™ dyes were previously established for the 52-Plex SNP for ID assay (Sanchez *et al.*, 2006). The signal variations observed for Multiplex 5 in Figure 4.13 are comparable to the reported ratio of 4:2:1:1 for the differences in fluorophore emissions between the dR110 (G-blue), dR6G (A-green), TAMRA (C-yellow) and dROX (T-red) dyes (Sanchez *et al.*, 2006).

However, a noteworthy signal difference that did not correspond with these standardised heterozygote patterns was seen for HERC2, where a strong green/blue imbalance was identified (Figure 4.13). This was attributed to the poor extension of the A allele that was widely observed during optimisation experiments and was further confirmed in these initial population testing results (Figure 4.13). The same HERC2 SNP (rs12913832) was also incorporated in a 34-Plex autosomal SBE assay for ancestry investigations which reported similar observations regarding the poor extension of the A allele compared to the G allele (Phillips *et al.*, 2012). In accordance with the interpretation rules proposed by Sanchez *et al*. (2006) to account for variations in the fluorophore emissions of [F]ddNTPs in the SNaPshot™ kit, the suggested minimum peak height of 400RFU (G), 200RFU (A) and 100RFU (C and T) were implemented for allele calling and a maximum peak height ratio of 3:1 was accepted as heterozygote.

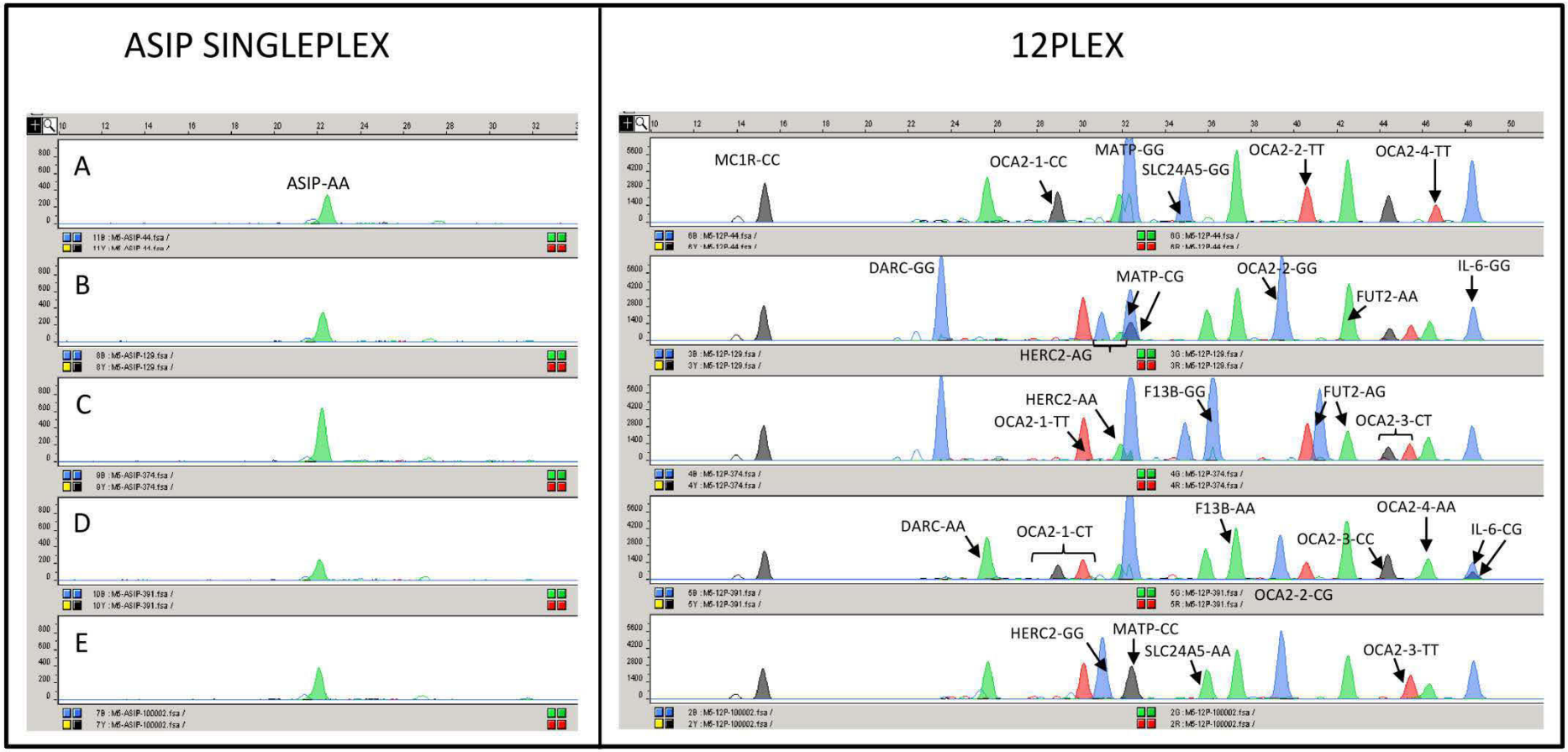


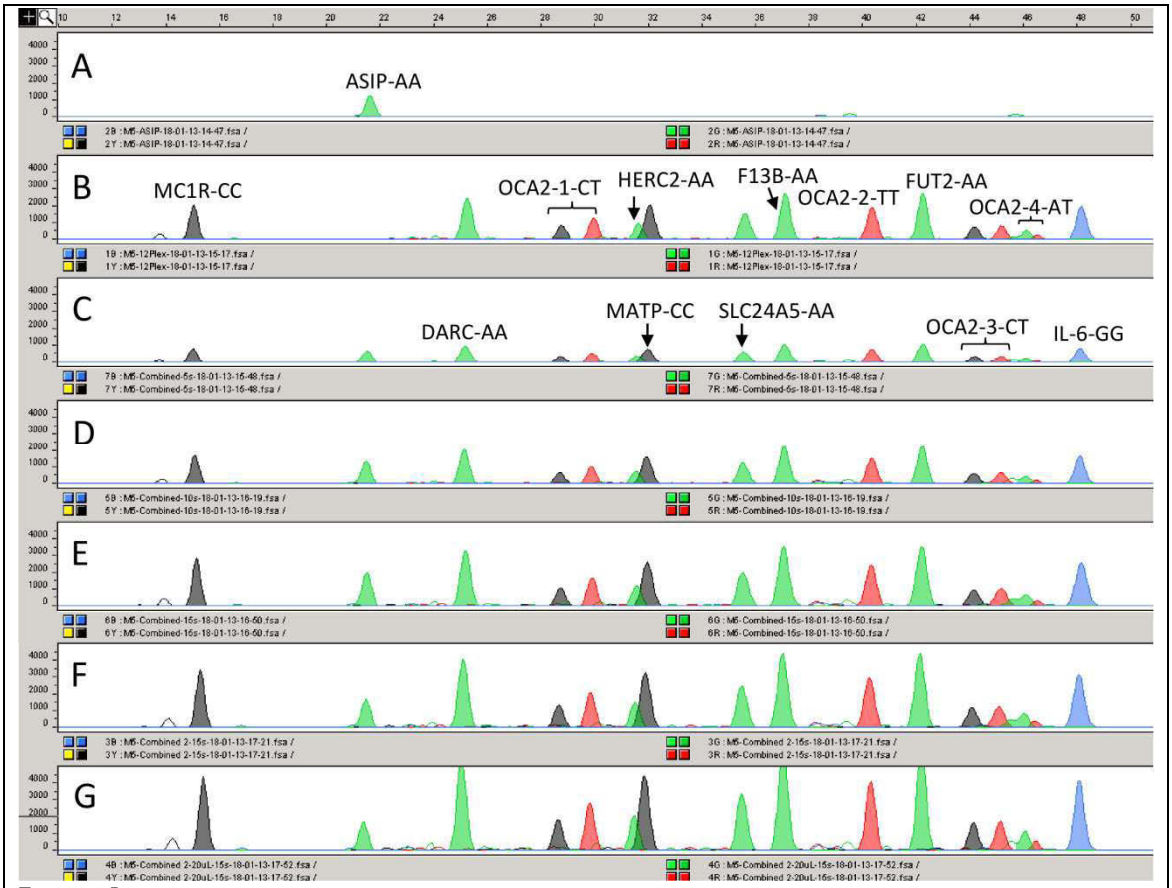
Figure 4.13: Initial population testing electropherograms of the Multiplex 5 profiles for five individuals with differing BGA.

(A) Chinese, (B) Egyptian, (C) Cameroonian, (D) Indian and (E) English donor. The DNA profiles depicted for each individual consist of the ASIP singleplex and the Multiplex 5 12-Plex SNaPshot™ reaction both prepared using Multiplex 5 PCR product as template. The Multiplex 5 PCR and SBE reactions were prepared using the optimised primer concentrations summarised in Table 4.3.

A ratio of up to 4:1 was accepted as heterozygote for the HERC2 locus to account for the poor extension of the A allele. Despite fluctuations in allele peak height caused by the variable emissions of the different SNaPshot™ dyes and the variable DNA concentration on the sampled FTA® cards, the initial population testing results have confirmed that the Multiplex 5 autosomal assay has been sufficiently optimised and have identified guidelines for allele calling. This concluded the final phase of development of our multiplex assay, with the complete 5-multiplex SNP genotyping system ready for validation through large scale population testing.

4.3.3 Simultaneous separation and detection of the ASIP singleplex and multiplex 5 12-Plex SNaPshot™ reactions using capillary electrophoresis

Due to the inability to extend the ASIP SBE primer in the Multiplex 5 13-Plex SNaPshot™ reaction, this SNP was required to be typed in a separate singleplex reaction. Though the Multiplex 5 PCR product could be used as template for both ASIP singleplex and 12-Plex SNaPshot™ reactions, both of these products must be separated in individual capillary electrophoresis runs, which is time consuming and consumes a greater volume of reagents. Since the ASIP SBE primer was designed to be unambiguously spatially resolved from the other Multiplex 5 SBE primers, it was hypothesised that the two SNaPshot™ products could be combined and separated in the same sample using capillary electrophoresis, to minimise analysis time and cost. To investigate this possibility, 0.5µL aliquots of ASIP singleplex and Multiplex 5 12-Plex SNaPshot™ reactions, were combined with 0.5µL of Genescan 120-LIZ® internal size standard in a final injection mix volume of 10µL. As shown in Figure 4.14C, the ASIP allele could be distinguished in combined samples. However, an approximately two fold decrease in peak height across all alleles was observed for the combined sample compared to when the ASIP singleplex (Figure 4.14A) and the Multiplex 5 12-Plex (Figure 4.14B) SNaPshot™ reactions were separated individually.



Legend:

(A) and **(B)** singleplex ASIP and 12-Plex SNaPshot™ reactions respectively, separated using standard 5 second injection.
(C) Combined sample containing 0.5µL of each of the singleplex ASIP and 12-Plex SNaPshot™ products in a final volume of 10µL. Injection time= 5 seconds.
(D) Same as (C), except the injection time was increased to 10 seconds.
(E) Same as (C), except the injection time was increased to 15 seconds.
(F) Combined sample containing 0.75µL of singleplex ASIP and 0.5µL of 12-Plex SNaPshot™ products in a final volume of 10µL. Injection time= 15 seconds.
(G) Same as (C), except the final combined sample volume was increased to 20µL.

Figure 4.14: Electropherograms of the final ASIP singleplex and Multiplex 5 12-Plex SNaPshot™ reactions in separate and combined capillary electrophoresis runs using variable injection times.

Capillary electrophoresis utilises electrokinetic injection to move the negatively charged DNA molecules from the sample into the capillary by applying a voltage over a defined time (Butler *et al.*, 2004). However, small interfering ions in the sample such as salts, dNTPs and primers increase the sample conductivity (ionic strength) and are preferentially injected due to their higher mobility compared to the larger DNA molecules (Butler *et al.*, 2004). It has been shown that DNA injection efficiency is

reduced in the presence of high concentrations of low molecular weight ionic species and that sample purification prior to capillary electrophoresis dramatically improves preferential injection of DNA products and detected peak signals (Schwartz *et al.*, 1991, Ruiz-Martinez *et al.*, 1998, Salas-Solano *et al.*, 1998). The decrease in peak height for the combined sample observed in Figure 4.14C was attributed to a reduction in the amount of DNA injected into the capillary due to increased sample ionic strength caused by pooling the two SNaPshot™ products.

Pre-electrophoresis purification and dilution of pooled DNA samples increases the processing time, which is not ideal for forensic purposes. Alternatively, increased injection settings such as injection time and voltage have been successfully used to improve the efficiency of STR typing for low template DNA samples (Westen *et al.*, 2009). To investigate whether increased signals can be achieved for the combined ASIP singleplex and 12-Plex SNaPshot™ reactions by modifying injection settings, the same pooled sample in Figure 4.14C was also electrophoresed using an increased injection time of 10 and 15 seconds. The run voltage was not modified as the GS POP-4 (1 mL) E5.md5 module used for the separation of SNaPshot™ products already specifies the maximum setting of 15 kV.

As evident in Figure 4.14D and Figure 4.14E, increasing the injection time from 5 seconds improves the peak heights of the combined sample. At an injection time of 10 seconds (Figure 4.14D) the allele peaks in the combined sample have returned to approximately the same levels as detected in the individual ASIP singleplex (Figure 4.14A) and 12-Plex (Figure 4.14B) SNaPshot™ reactions runs. The highest peak heights were observed for the combined sample electrophoresed when using an injection time of 15 seconds where the homozygous peaks averaged about 3000RFU (Figure 4.14E). Injection settings affect peak morphology and increased injection times have been associated with peak broadening (Moretti *et al.*, 2001b). However, the most dramatic effects were reported for 15 and 20 second injection times and affected larger sized peaks (>300 bases) (Moretti *et al.*, 2001b). As the Multiplex 5 SNaPshot™ products are all less than 50 bases in size, the effect on peak width and resolution was negligible with very minor broadening was observed when using a 15 second injection time (Figure 4.14E).

A slight increase in signal to noise ratio was detected when the injection time was more than 5 seconds, with the greatest effect observed for the 15 second injection (Figure 4.14E). These areas of noise were detected away from any allele peaks, such as in between the ASIP and DARC and F13B and OCA2-2 alleles, so this was not considered problematic and the detection threshold for allele peaks could be maintained at 400RFU (G), 200RFU (A) and 100RFU (C and T). However, increased injection times unfortunately increase the peak height of both alleles and artefact peaks. This was seen in Figure 4.14D and E with the small non-specific 'A' peak preceding the OCA2-4*A allele peak detected at similar peak heights when using a 10 and 15 second injection time respectively. This was not considered a major issue, as the non-specific 'A' peak did not overlap with any expected allele peaks.

Whilst, satisfactory results were obtained for the combined sample using increased injection times, the ASIP SNaPshot™ product extends less efficiently than other alleles in the 12-Plex SNaPshot™ reaction (Figure 4.14D and E). To investigate whether the ASIP signal in the combined sample could be increased to achieve a more balanced profile, the amount of ASIP SNaPshot™ product incorporated into the combined sample was increased to 0.75µL, with all other volumes maintained as per the original combined sample, in a final injection mix volumes of 10µL and 20µL and were electrophoresed using an injection time of 15 seconds.

Increasing the amount of ASIP SNaPshot™ product did not improve the signal of the ASIP allele irrespective of the combined sample volume (Figure 4.14F and G). However, higher peak heights with homozygous peaks averaging about 4000RFU were observed for the samples prepared in a larger final volume of 20µL (Figure 4.14G). This was expected as the larger injection volume diluted the ASIP and 12-Plex SNaPshot™ products to reduce the ionic strength of the solution and improve injection efficiency of the combined sample. Overall, the best results were achieved for the original combined sample incorporating 0.5µL of each of the products with 0.5µL of Genescan 120-LIZ® internal size standard in a final injection mix volume of 10µL, using an injection volume of 10 seconds (Figure 4.14D). These conditions were selected as the most appropriate for simultaneously separating and detecting the ASIP and 12-Plex SNaPshot™ products. This is since these conditions achieved similar peak heights to

those observed when the products were electrophoresed individually and showed the most balanced signals between the ASIP and 12-Plex peaks. Though the 15 second injection increased the peak height, the signal to noise ratio and artefact peak detection were more pronounced.

4.4 Summary

This chapter discussed the final stage of development of an autosomal multiplex SNP assay comprising the last of five multiplex SBE assays targeting three different classes of SNPs for the inference of BGA and EVCs. The first stage of assay development, explored in chapter 3, involved the categorisation of selected haploid mtDNA and NRY SNPs within four hierarchical multiplex SNaPshot™ assays, which were optimised for the inference of maternal and paternal BGA. A similar assay optimisation process was followed for the autosomal SNaPshot™ assay. Extensive literature and database searches were performed to select 13 autosomal SNPs that demonstrated skewed allele frequencies in Sub-Saharan African, European, Middle Eastern, South and East Asian populations and/or associations with different hair, eye and skin phenotypes.

Similarly to the hierarchical haploid multiplex assay, PCR and SBE primers were designed to genotype the autosomal SNPs. The Multiplex 5 PCR reaction was initially optimised by testing the PCR primers in a singleplex reaction before combining them in a 13-Plex PCR reaction. The same PCR reagent concentrations optimised for Multiplex 1 were used for Multiplex 5 with the exception of PCR primer concentrations, which were varied to achieve balanced amplification of all loci. The preliminary optimised Multiplex 5 PCR product was then used as template for the optimisation of the Multiplex 5 SBE reaction. Initially, the specificity of the designed SBE primers was verified through singleplex SNaPshot™ reactions prior to grouping all the Multiplex 5 SBE primers in a single 13-plex SNaPshot™ reaction.

Several issues with the designed SBE primers were identified including poor spatial resolution of OCA2-18*C and MATP*C as well as poor extension of the ASIP SBE primer, which was attributed to a variety of causes (refer to 4.3.1.2.3). These SBE primers were redesigned and the modified primers reincorporated into the Multiplex 5

13-plex SNaPshot™ reaction. However, consistent failure of the ASIP SBE primer to extend in the multiplex SNaPshot™ reaction resulted in the optimisation of separate ASIP singleplex and 12-plex SNaPshot™ reactions, which both use the Multiplex 5 PCR product as template. Allele peak heights for these SNaPshot™ reactions were optimised by modifying PCR and SBE primer concentrations until all alleles could be unambiguously detected and homozygous peak heights ranged between 2000-2500RFU, in the absence of over-amplification. Due to the diploid nature of autosomal SNPs, heterozygote alleles were optimised until they were detected at approximately half the peak height of homozygous allele peaks.

The optimised Multiplex 5 autosomal SNP assay was used to perform initial population testing on the same five volunteer DNA donors typed using the four optimised multiplex haploid SNP assays. These initial population testing results confirmed that the Multiplex 5 assay was sufficiently optimised and ready for large-scale population testing (discussed in chapter 5), as all peak heights detected were satisfactory despite variations in the DNA content of the FTA discs used across the five donors tested. However, inconsistencies in the allele peak heights at different loci were observed, which were attributed to the reported ratio of 4:2:1:1 for the differences in fluorophore emissions between the dR110 (G-blue), dR6G (A-green), TAMRA (C-yellow) and dROX (T-red) dyes (Sanchez *et al.*, 2006). These standardised rules proposed for the interpretation of heterozygote imbalances when using SNaPshot™ genotyping technology were implemented for allele calling in Multiplex 5, including a minimum peak height ratio of 3:1 for accepting a heterozygote at any particular locus. In addition, the strong green/blue imbalance at the HERC2 locus (Figure 4.13) which was attributed to the poor extension of the A allele resulted in a modified 4:1 ratio for heterozygotes at the HERC2 locus to account for the imbalanced extension between the A and G alleles.

The ASIP singleplex and 12-plex SNaPshot™ reactions were also simultaneously separated and detected when combined in equal proportions within the same injection sample to minimise sample processing time. Increasing the injection time to 10 seconds achieved appropriate signal levels, with minimal effect on peak morphology and signal to noise ratio. This combined sample preparation and the modified injection settings were implemented for Multiplex 5 when performing large- scale population genotyping.

CHAPTER 5

PRELIMINARY VALIDATION OF THE 5-MULTIPLEX SNP ASSAY FOR FORENSIC INTELLIGENCE

Chapter 5. Preliminary validation of the 5-multiplex SNP assay for forensic intelligence

5.1 Introduction

The objective of this study was to develop a multiplex SNP genotyping assay combining mtDNA, NRY and autosomal SNPs, with the intention of creating a novel forensic intelligence tool for the inference of biogeographical ancestry (BGA) and external visible characteristics (EVC) from unknown DNA samples. The assay was divided into five separate multiplex SNaPshot™ (Life Technologies, USA) assays: Multiplex 1 to 4 grouping the haploid mtDNA and NRY SNPs according to a hierarchical design and Multiplex 5 containing the ancestry and pigmentation informative autosomal SNPs. The development and optimisation of these multiplex assays was discussed in chapters 3 and 4. This chapter describes the preliminary validation of the 5-multiplex SNP genotyping system to assess its utility for the inference of ancestry and pigmentation phenotype.

The validation study involved genotyping 146 population samples with declared BGAs from Sub-Saharan and North Africa, the Middle East, Europe and South and East Asia. Only individuals for whom both maternal and paternal grandparents were declared to be of the same BGA were utilised for this study. Maternal and paternal BGA was inferred from the geographic region of origin of the derived mtDNA and NRY haplogroups respectively, whilst for autosomal genotypes, maximum likelihood estimates (MLE) were calculated to assign each individual to a population of origin and/or a specific phenotype. The degree of correlation between the inferred and declared BGA and EVC was assessed to determine the predictive power of the 5-multiplex SNP assay. A preliminary sensitivity study and genotyping of simulated casework samples were also performed to further investigate the applicability of the developed assay in forensic casework.

5.2 Results

5.2.1 Large- scale genotyping of population samples

The final optimised conditions for the Multiplex 1 to 5 assays (refer to Table 3.4 and Table 4.3) were used to genotype 116 male samples originating from six major global population groups prevalent in Australia including Sub-Saharan Africa (N=7), North Africa (N=15), Middle East (N=10), Europe (N=36), South Asia (N=14) and East Asia (N=34). Refer to section 2.1.5 and Table 2.3 for details regarding the declared country of origin of selected samples that were included in each population group. Of the 116 male samples, 85 were genotyped through all five multiplexes, with no ambiguous haplogroup designations observed, confirming the efficacy of the hierarchical multiplex design. The remaining samples were genotyped using only the appropriate multiplex(es) required, as identified through the result of Multiplex 1. An additional 30 female samples were also genotyped to further assess the specificity of the developed 5-multiplex SNP assay and to increase the sample size for population genetic analysis of the autosomal data.

The majority of samples were successfully typed and full profiles obtained. A few samples occasionally demonstrated low signals and allele drop out caused by low DNA template concentrations on the donor FTA® cards. These samples were re-typed using a new disc, which typically improved the result, though in a few instances the addition of a second disc to the PCR reaction was required to obtain a complete profile (data not shown). It was also found that the P31 NRY SNP typed in Multiplex 4 consistently failed to extend in 9 out of 34 individuals of East Asian origin despite repeating the PCR and SBE multiplex reaction several times and observing strong signals for other alleles in the same multiplex. This SNP defines NRY haplogroup O2, which is prevalent in East Asia. All nine individuals for which P31 could not be detected in multiplex SBE reactions were re-typed in singleplex SBE reactions with the P31 SBE primer at a concentration of 4μM and using the original Multiplex 4 PCR product as template. All nine of these East Asian males were identified as having the G allele, indicative of the O2 NRY haplogroup. Review of the DNA sequence surrounding the SNP revealed two repeat regions of Ts interspersed with G nucleotides, thus it is possible that the presence of the G substitution in individuals belonging to the O2 haplogroup may be causing the

SBE primer to bind to non-specific templates, resulting in failed extension. Consequently it is suggested that the P31 SBE primer be redesigned on the alternate strand.

5.2.2 Sensitivity testing

To investigate the sensitivity of the five multiplexes, genomic DNA (gDNA) was extracted from whole blood obtained from a European individual and different starting amounts of DNA template ranging from 10ng to 0.01ng were added to each optimised multiplex reaction as described in section 2.2.4. As shown in Figure 5.1 to Figure 5.3, full profiles were obtained for Multiplex 1 to 5 with as little as 0.1ng, except for ASIP which could not extend at 0.1ng (Figure 5.3). For Multiplex 1, full profiles were obtained for as little as 0.03ng of gDNA template, however at lower starting template concentrations, the M201, C10400T, M89 and C10238T alleles could not be called (Figure 5.1) Alleles such as M45 and C8281G, which demonstrated the lowest peak heights even at high starting gDNA template concentrations, showed consistent signals despite reductions in DNA template. Multiplex 2 showed full profiles only up to a starting concentration of 0.1ng with a loss of NRY loci at lower concentrations (Figure 5.1). Interestingly, the A12308G*A allele, which had consistently shown low peak heights in comparison to the A12308G*G allele, demonstrated slightly increased peak heights at concentrations 0.3ng or less.

The highest sensitivity was observed for Multiplexes 3 and 4, with complete profiles obtained for starting DNA template concentrations as low as 0.01ng (Figure 5.2). The only exception was for the Multiplex 3 M70 allele, which could not be called when using 0.01ng of template DNA. In both multiplexes, the mtDNA loci showed the highest signals, with the NRY loci decreasing dramatically at starting template concentrations below 0.1ng (Figure 5.2). Some artefacts, identified by asterisks in Figure 5.1 to Figure 5.3, were also evident at the lowest concentrations of starting gDNA template. In Multiplex 5, allele drop-out was detected when DNA template concentrations fell below 0.1ng for the 12-Plex, with loss of one allele at the MATP and

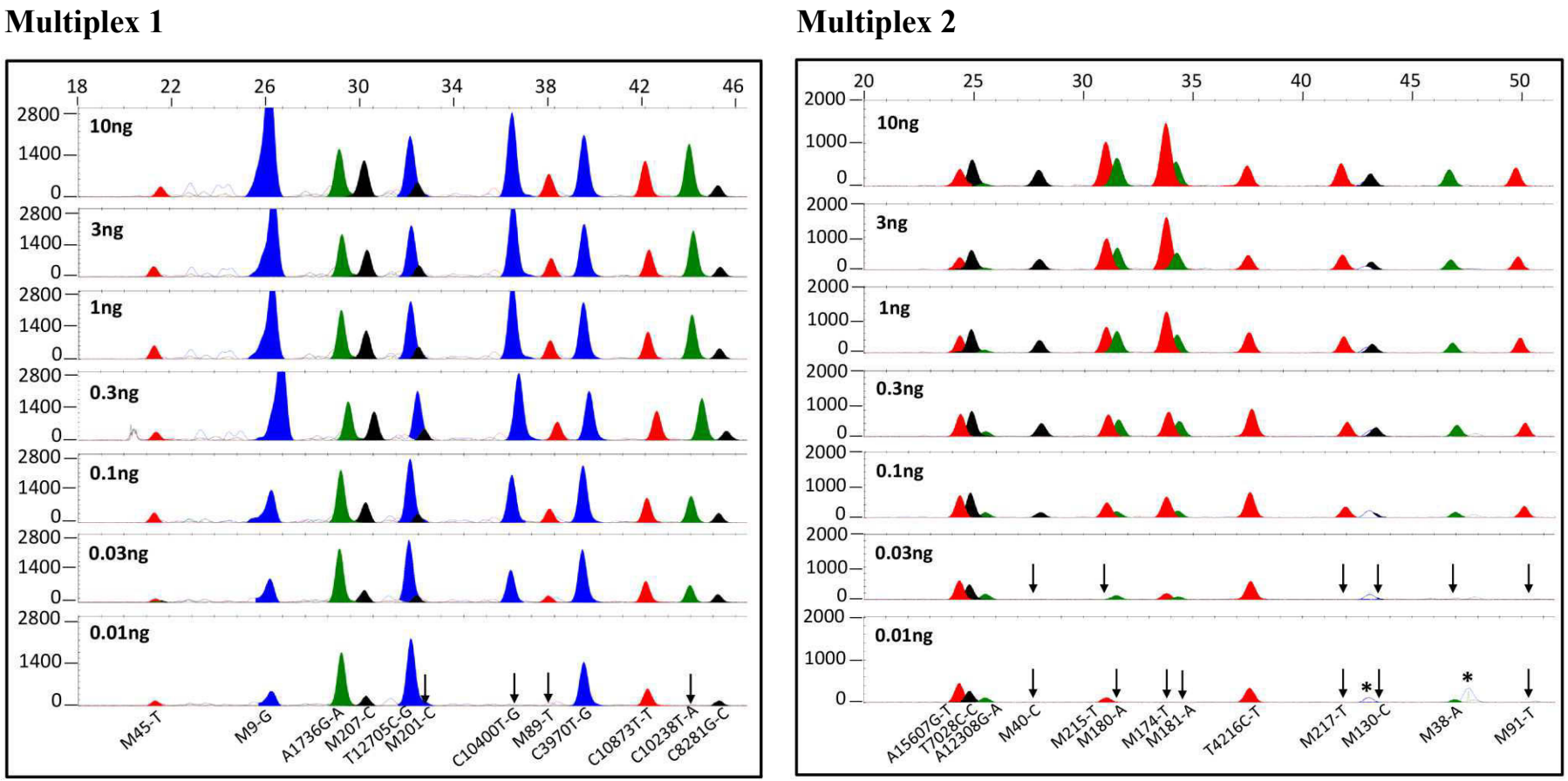


Figure 5.1: Multiplex 1 and 2 sensitivity profiles for a European individual genotyped using various amounts of starting DNA template.

The European individual was identified to belong to mtDNA haplogroup H* and NRY haplogroup R1a1. Each multiplex assay was performed on different amounts of starting DNA template ranging from 10ng to 0.01ng. The black arrows indicate allele dropout at lower DNA concentrations and black '*' indicate artefacts that become more prominent and may affect allele calling at very low DNA template concentrations.

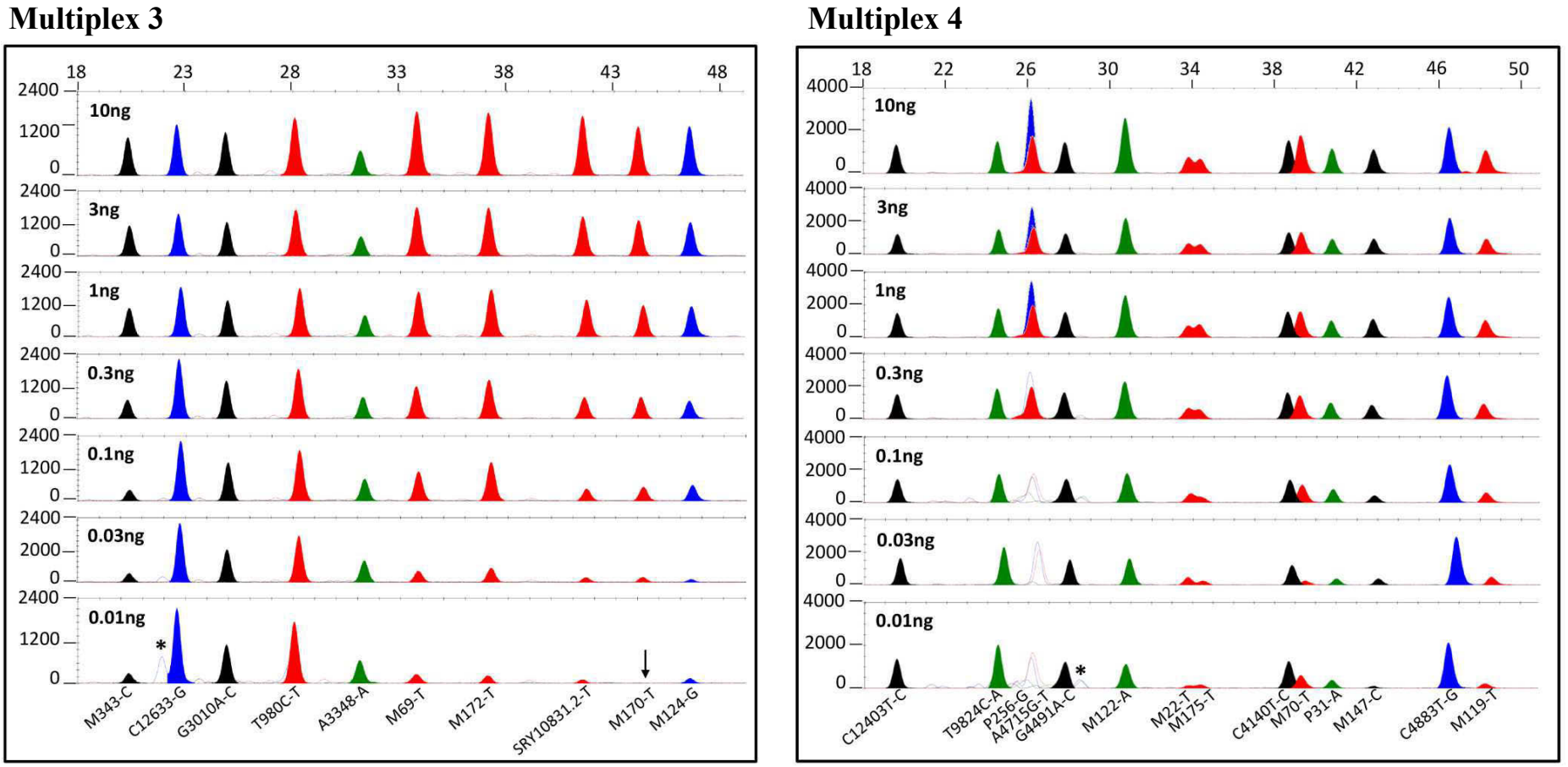


Figure 5.2: Multiplex 3 and 4 sensitivity profiles for a European individual genotyped using various amounts of starting DNA template.

The European individual was identified to belong to mtDNA haplogroup H* and NRY haplogroup R1a1. Each multiplex assay was performed on different amounts of starting DNA template ranging from 10ng to 0.01ng. The black arrows indicate allele dropout at lower DNA concentrations and black '*' indicate artefacts that become more prominent and may affect allele calling at very low DNA template concentrations. Note the P256 and A4715G alleles in Multiplex 4 could not be coloured because the peak heights overlap.

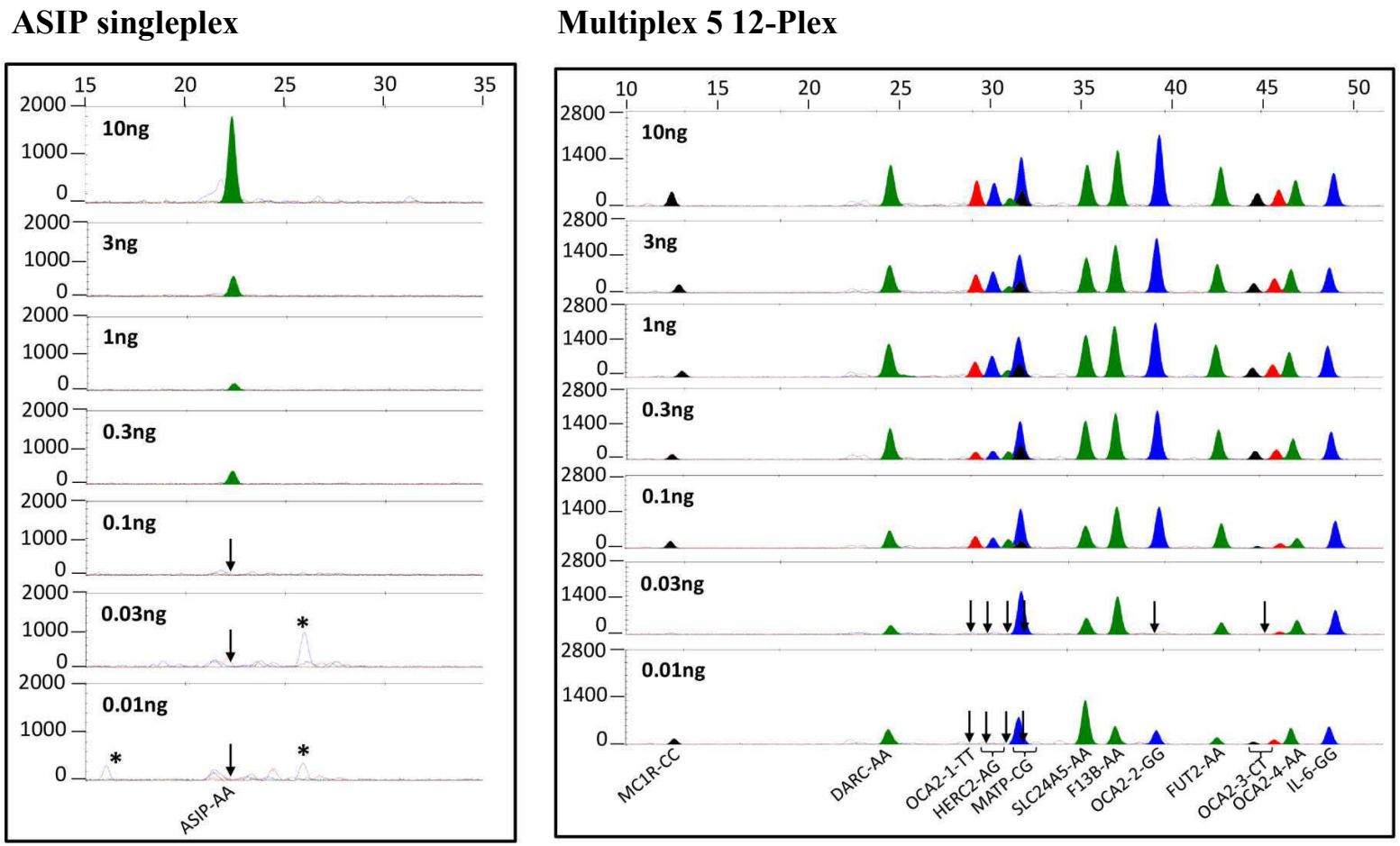


Figure 5.3: Multiplex 5 sensitivity profiles for a European individual genotyped using various amounts of starting DNA template.

The multiplex 5 PCR was performed on different amounts of starting DNA template ranging from 10ng to 0.01ng. The ASIP singleplex and Multiplex 5 SBE reactions were then performed separately. The black arrows indicate allele dropout at lower DNA concentrations and black '*' indicate artefacts that become more prominent and may affect allele calling at very low DNA template concentrations.

and OCA2-3 alleles thereby altering the genotype from heterozygous to homozygous (Figure 5.3). The ASIP locus, which is extended separately in a singleplex SBE reaction, could not be called below 0.3ng of starting template and an increase in artefact peaks was observed at the lowest genomic DNA template concentrations (Figure 5.3).

Generally, mtDNA loci could be detected at lower starting genomic DNA concentrations due to the higher copy number of mtDNA compared to NRY and autosomal DNA. Furthermore, the mtDNA loci that failed to be detected below 0.1ng of starting template were those with the largest PCR and/or SBE products such as C10400T and C10238T in Multiplex 1 (Figure 5.1). Overall, a detection limit of 0.1ng of starting genomic DNA template was identified to be the sensitivity threshold of the 5-multiplex SNP assay for the production of full profiles.

5.2.3 Distribution of mtDNA, NRY and autosomal SNP alleles in test population groups

5.2.3.1 *mtDNA and NRY haplogroups*

A total of 23 mtDNA and 22 NRY haplogroups were detected for the 116 males genotyped and haplogroup frequencies in each declared population group were determined as described in section 2.2.3.1. As shown in Figure 5.4, distinct mtDNA and NRY haplogroup frequency distributions were observed for the six different declared populations of origin. Nine mtDNA and twelve NRY haplogroups were found to be population specific, including mtDNA haplogroups A, B4'5/B6, B4a1, M7, M8, M9, M10, M80'D which were only observed in East Asians and M1 in North Africans (Figure 5.4A). The NRY haplogroups C3 and O* including subgroups O1a, O2 and O3 were also confined to East Asian individuals, whilst BT and Y only occurred in the North African population (Figure 5.4B).

All declared Sub-Saharan African individuals were typed as mtDNA haplogroup L and the majority were NRY haplogroups E1b1a1a1 (71.4%) and E* (14.3%), which were solely observed in this population (Figure 5.4). North Africans demonstrated dramatically different mtDNA and NRY haplogroup distributions to Sub-Saharan

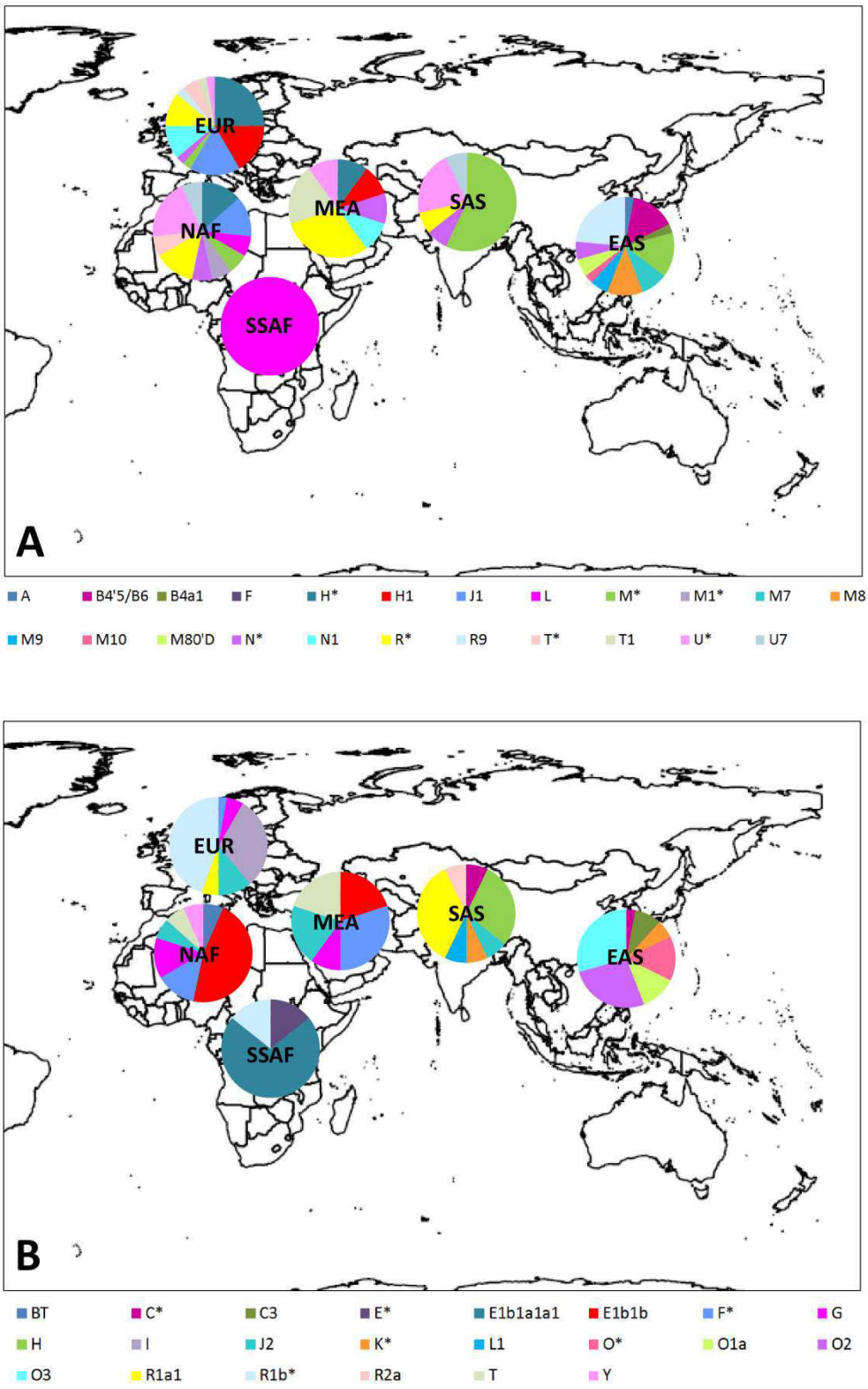


Figure 5.4: Geographic distribution of observed (A) mtDNA and (B) NRY haplogroups.

Haplogroup frequency (%) distribution for 116 males originating from Sub-Saharan Africa (**SSAF**), North Africa (**NAF**), Europe (**EUR**), Middle East (**MEA**), South Asia (**SAS**) and East Asia (**EAS**). For details regarding the declared country of origin of the individuals tested that are included in each population group refer to Table 2.3.

Africans yet similar distributions to individuals of Middle Eastern origin, with mtDNA haplogroups H*, R* and U* and NRY haplogroups E1b1b, F*, G, J1 and T observed in both populations (Figure 5.4). However, NRY haplogroup E1b1b was over represented in North Africans, with a frequency (46.7%) greater than two-fold that observed in Middle Eastern samples (20%), whilst the opposite trend was seen for haplogroups F* and T (Figure 5.4B). The mtDNA haplogroup distribution of individuals of European origin greatly overlapped that of North African and Middle Eastern populations, however, H* was over represented in Europeans at a frequency of 25%, twice as frequent in this population compared to North Africans and Middle Easterners (Figure 5.4A). The NRY haplogroup distribution of Europeans appeared more distinct, with 44.4% of individuals belonging to haplogroup R1b* and 30.6% to haplogroup I, which was specific to the European population (Figure 5.4B).

In contrast, the majority of males of declared South Asian BGA were predominantly haplogroup M* (57.1%), with the remainder belonging to mtDNA haplogroups also observed in North African and Middle Eastern populations, including N*, R*, U* and U7 (Figure 5.4A). Individuals of South Asian origin demonstrated more defined NRY haplogroup distribution distinctions with 35.7% and 28.6% of males observed to be haplogroups R1a1 and H respectively and others belonging to NRY haplogroups L1 and R2a, which were specific to the South Asian population (Figure 5.4B). The mtDNA and NRY haplogroup distribution of East Asians appeared to be the most distinct compared to other populations and showed the greatest number of population specific haplogroups (Figure 5.4). The majority of East Asian individuals belonged to mtDNA haplogroups R9 (23.5%), M* (14.7%) and B4’5/B6 (14.7%) and NRY haplogroups O2 (26.5%) and O3 (29.4%).

5.2.3.2 *Autosomal allele and genotype frequencies*

The allele and genotype frequencies at each autosomal SNP locus were calculated for the declared Sub-Saharan Africa, North Africa, Middle East, Europe, South Asia and East Asia populations using GenAlEx v.6.501 (Peakall and Smouse, 2012) as detailed in section 2.2.3.1. Frequencies were derived from the analysis of 146 genotypes comprising 116 males and 30 females. The frequency data was initially used to evaluate

the ability of the selected autosomal markers to differentiate between our populations of interest prior to performing further population genetic analyses. Variations in the allele and genotype frequency distribution of the autosomal loci across the six populations were observed and are discussed in detail below.

MC1R (rs885479), DARC (rs2814778) and OCA2-1 (rs7495174)

The MC1R rs885479 (C>T) allele and genotype frequencies showed distinct differences between East Asian and all other populations, with the highest incidence of the T allele (51.3%) and CT/TT genotypes detected in individuals of declared East Asian BGA (Figure 5.5A&B). A complete absence of the T allele was observed in Sub-Saharan African, North African and Middle Eastern populations whilst very low frequencies of the T allele were identified in the European and South Asian populations, with all these individuals presenting the CT genotype.

At the DARC rs2814778 (A>G) locus, the G allele strongly correlated with African populations (Figure 5.5A&B) where it occurred at frequencies >90% in both Sub-Saharan and North African groups, with the majority of individuals in these populations demonstrating the GG genotype. In contrast, the A allele was predominant in all non-African populations where a very low incidence (< 3%) of G was observed.

Allele and genotype frequency differences between the East Asian and other populations were also identified at the OCA2-1 rs7495174 (C>T) locus (Figure 5.5A&B). The highest frequency of the C allele (68%) and CC genotype (49%) was observed in East Asians compared to other populations where the C allele frequency ranged from 13 to 28%. In contrast, the African, Middle Eastern, European and South Asian populations showed similar frequencies where the T allele and TT genotypes predominated. Interestingly, the highest frequency of CT genotypes was observed in the Middle Eastern population.

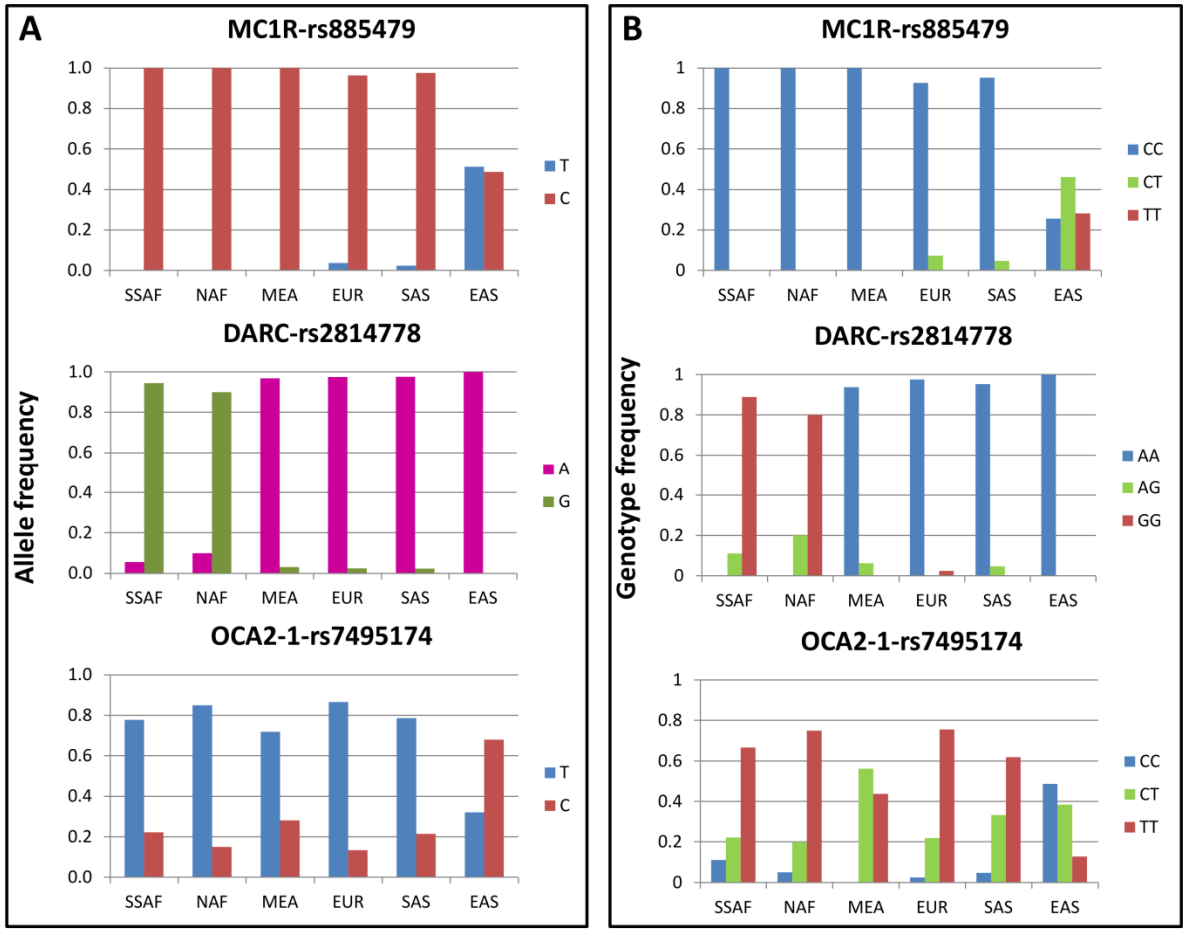


Figure 5.5: Allele and genotype frequencies for the MC1R (rs885479), DARC (rs2814778) and OCA2-1 (rs7495174) loci.

Allele (**A**) and genotype (**B**) frequencies for 146 individuals originating from Sub-Saharan Africa (**SSAF**), North Africa (**NAF**), Middle East (**MEA**), Europe (**EUR**), South Asia (**SAS**) and East Asia (**EAS**).

HERC2 (rs12913832), MATP (rs16891982) and SLC24A5 (rs1426654)

The North African, Middle Eastern and European populations showed concordant allele frequencies at the HERC2 rs12913832 (A>G) locus and could be clearly distinguished from the other populations (Figure 5.6A). Europeans demonstrated the highest frequency of the G allele (57%) followed by North Africans (38%) and Middle Easterners (25%). In contrast, the G allele was practically absent from the Sub-Saharan African, South and East Asian populations where it was detected at very low frequencies (2.5-5.6%). Although the G allele occurred at high frequency in North Africans, Middle Easterners and Europeans, the genotype frequencies varied across the populations (Figure 5.6B). Europeans demonstrated relatively equal AG and GG frequencies (~40% each), whilst North Africans showed a higher incidence of AG genotypes (45%) compared to GG genotypes (15%).

At the MATP rs16891982 (G>C) locus, the observed allele and genotype frequencies also indicated distinct differences between the North African, Middle Eastern and European populations and other population groups (Figure 5.6A&B). The European population showed high frequencies of the C allele (89%) and CC genotype (80.5%), whilst the North African and Middle Eastern populations showed similar intermediate frequencies for the C allele and CC genotype. In contrast, the Sub-Saharan African, South and East Asian populations demonstrated high frequencies for the G allele (90-94%) and GG genotype (81-90%).

At the SLC24A5 rs1426654 (G>A) locus, distinguishing differences in allele and genotype frequencies were identified between the Sub-Saharan African and East Asian populations and the remaining populations (Figure 5.6A&B). In both populations, the G allele occurred at high frequency (78-95%) and the GG genotype predominated. However, in the European and North African populations the G allele was almost never detected, with the majority of individuals from these populations showing the AA genotype. Though low frequencies of the G allele were observed in the Middle Eastern and South Asian population (~19%), the A allele was mostly predominant in these populations as well.

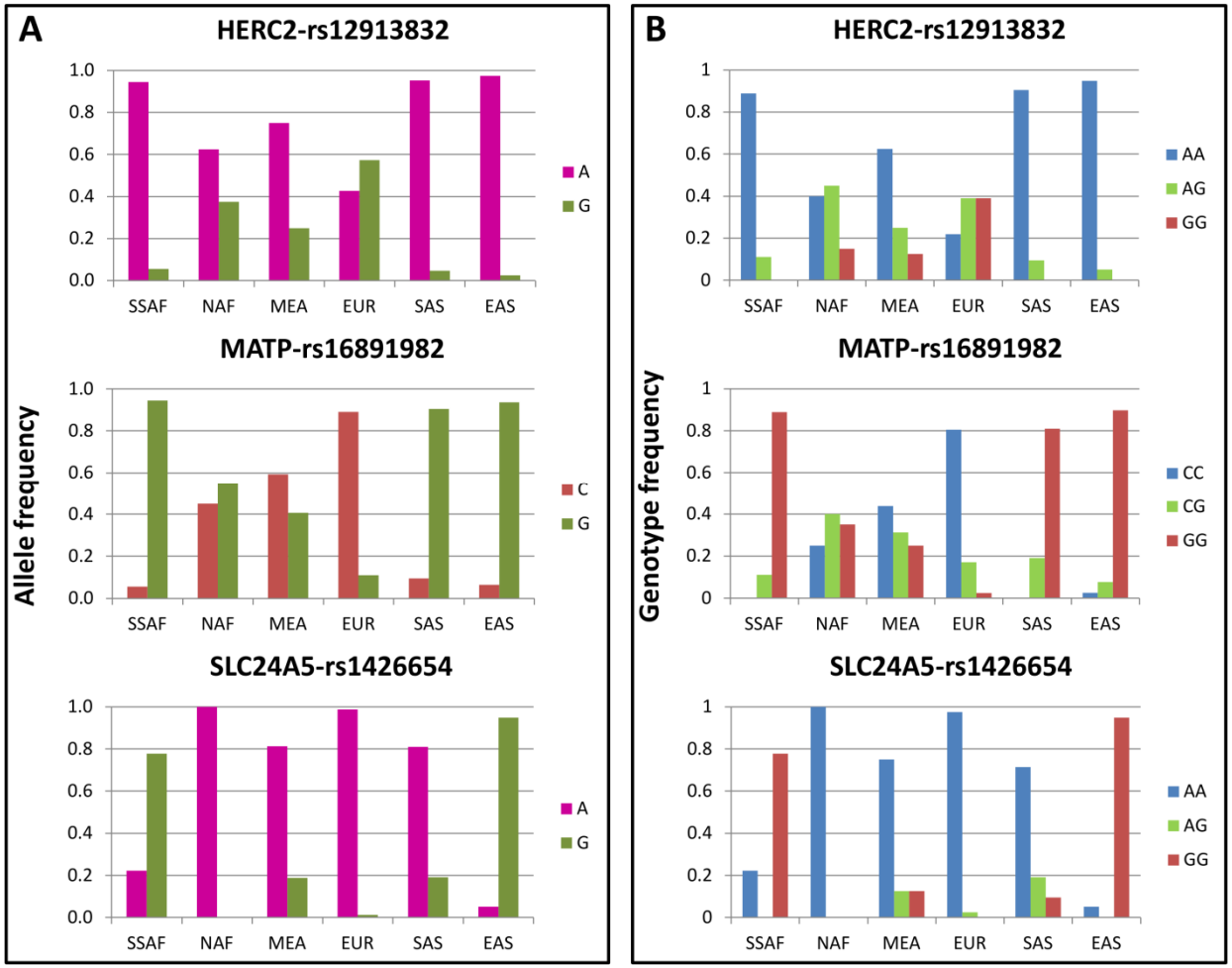


Figure 5.6: Allele and genotype frequencies for the HERC2 (rs12913832), MATP (rs16891982) and SLC24A5 (rs1426654) loci.

Allele (**A**) and genotype (**B**) frequencies for 146 individuals originating from Sub-Saharan Africa (**SSAF**), North Africa (**NAF**), Middle East (**MEA**), Europe (**EUR**), South Asia (**SAS**) and East Asia (**EAS**).

F13B (rs6003), OCA2-2 (rs4778241) and FUT2 (rs1800021)

The allele and genotype frequencies observed for the F13B rs6003 (G>A) locus clearly differentiated the African populations from the remaining populations (Figure 5.7A&B). Sub-Saharan Africans showed the highest frequency for the G allele (50%) and GG genotype (33%), followed by the North African population, which showed the second highest frequency for the G allele (25%). Contrastingly, the A allele predominated (>86%) in the other non-African populations where the G allele was relatively absent.

Minor differences in allele and genotype frequencies were observed for the OCA2-2 rs477824 (T>G) locus between the six populations (Figure 5.7A&B). The T allele (74%) and TT genotype (59%) were most frequent in East Asia whilst the G allele and GG genotype showed the highest frequency in Europeans. However, intermediate allele and genotype frequencies were identified for the other populations except for North Africa and South Asia where interestingly, the highest incidence of GT genotypes was seen.

On the other hand, the Sub-Saharan African population could be clearly differentiated from the other populations through the allele and genotype frequencies for the FUT2 rs1800021 (A>G) locus (Figure 5.7A&B). The G allele and AG genotype were only detected within the Sub-Saharan African population and were absent from the other populations where the A allele and AA genotype predominated, with the exception of one South Asian individual who was identified to have the GG genotype.

OCA2-3 (rs4778138), OCA2-4 (rs1545397), IL-6 (rs1800795) and ASIP (rs6058017)

Similarities in allele and genotype frequencies were observed at the OCA2-3 rs4778138 (C>T) locus between the Sub-Saharan African and East Asian populations (Figure 5.8A&B). In these populations, the C allele (~76%) and CC genotypes (56-59%) occurred at high frequencies, such that Sub-Saharan Africans and East Asians could be differentiated from European population where the T allele and TT genotype predominated. Conversely, similar intermediate C and T allele frequencies were observed between the North African, Middle Eastern and South Asian populations (Figure 5.8A). However these three populations could be differentiated according to the

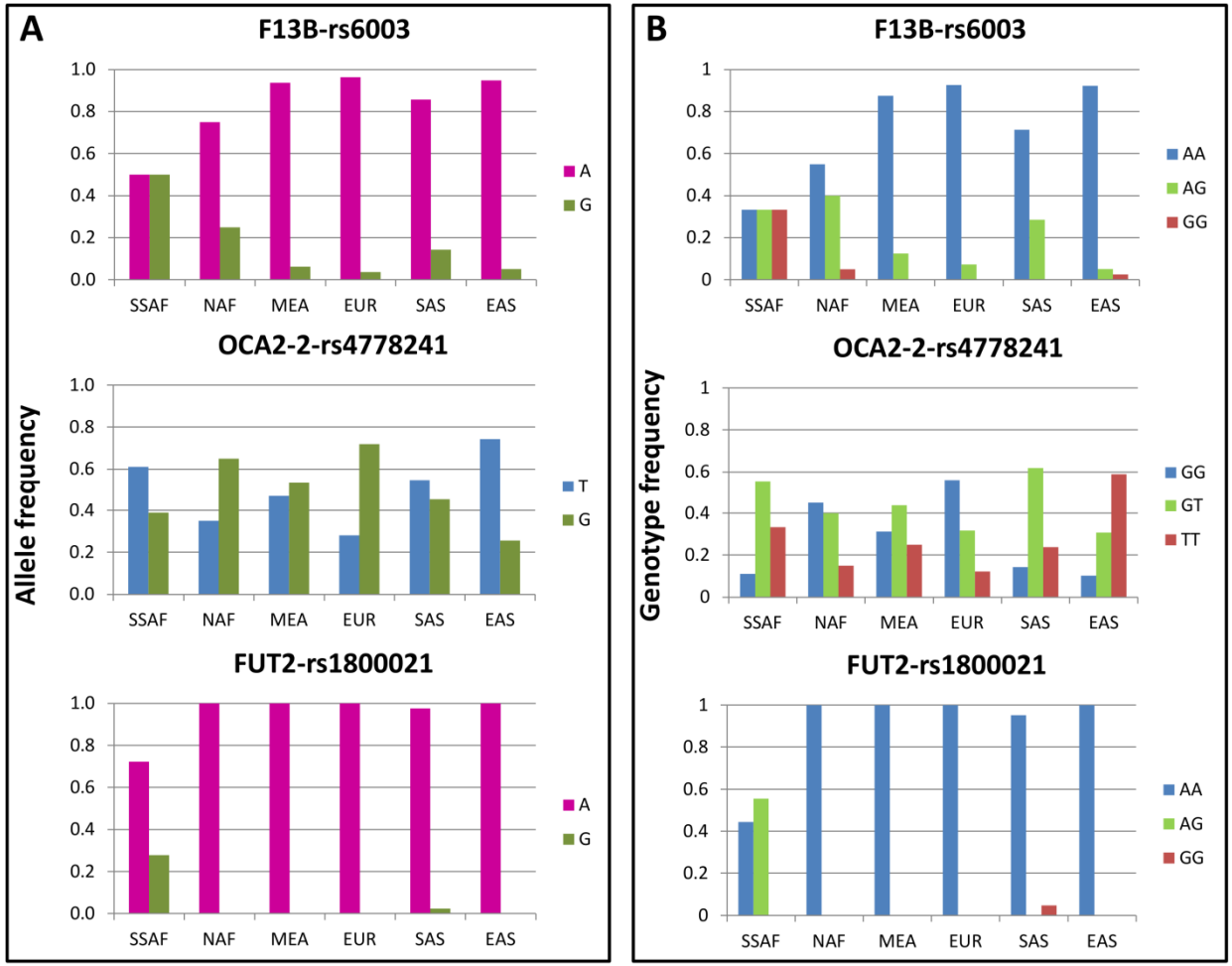


Figure 5.7: Allele and genotype frequencies for the F13B (rs6003), OCA2-2 (rs4778241) and FUT2 (rs1800021) loci.

Allele (**A**) and genotype (**B**) frequencies for 146 individuals originating from Sub-Saharan Africa (**SSAF**), North Africa (**NAF**), Middle East (**MEA**), Europe (**EUR**), South Asia (**SAS**) and East Asia (**EAS**).

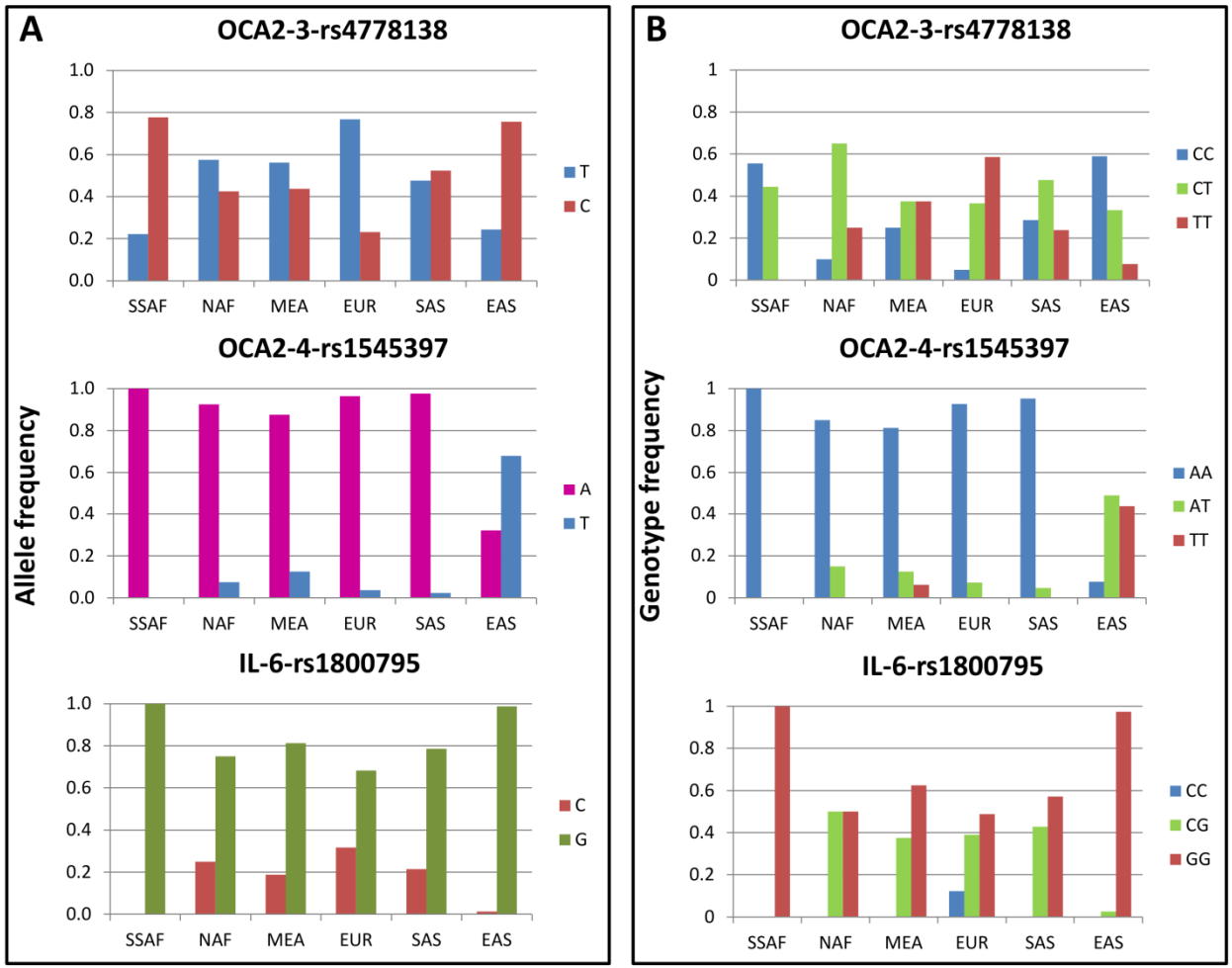


Figure 5.8: Allele and genotype frequencies for the OCA2-3 (rs4778138), OCA2-4 (rs1545397) and IL-6 (rs1800795) loci.

Allele (**A**) and genotype (**B**) frequencies for 146 individuals originating from Sub-Saharan Africa (**SSAF**), North Africa (**NAF**), Middle East (**MEA**), Europe (**EUR**), South Asia (**SAS**) and East Asia (**EAS**).

genotype frequencies, with the North Africans showing the highest frequency of CT genotypes (65%); almost double which observed in Middle Eastern and South Asians.

From the allele and genotype frequencies observed at the OCA2-4 rs1545397 (A>T), the East Asian population could be distinguished from the other five population groups (Figure 5.8A&B). East Asians showed the highest incidence of the T allele (68%) and TT genotype (44%), compared to the other populations where the A allele and AA genotype predominated.

The IL-6 rs1800795 (G>C) locus also demonstrated similar allele and genotype frequencies between the Sub-Saharan African and East Asian populations, distinguishing these groups from the other populations (Figure 5.8A&B). The C allele was found to be relatively absent (<1.3%) from Sub-Saharan African and East Asian populations but was detected at high frequencies (>19%) in North Africa, the Middle East, Europe and South Asia, with the highest incidence of the C allele (32%) and CC genotype (12%) observed in Europeans.

Differences in allele and genotype frequencies at the ASIP rs6058017 (G>A) locus were identified that differentiated the Sub-Saharan African population from other populations (Figure 5.9A&B). The G allele was detected in 50% of individuals of Sub-Saharan African descent and 44% demonstrated the GG genotype. Contrastingly, in the other populations the G allele occurred at much lower frequencies ranging from 7 to 23%. Minor differences in the allele and genotype frequencies between North Africans, the Middle Easterners, Europeans and South Asians were observed, with Europeans showing the highest A allele (94%) and AA genotype (88%) frequencies (Figure 5.9B).

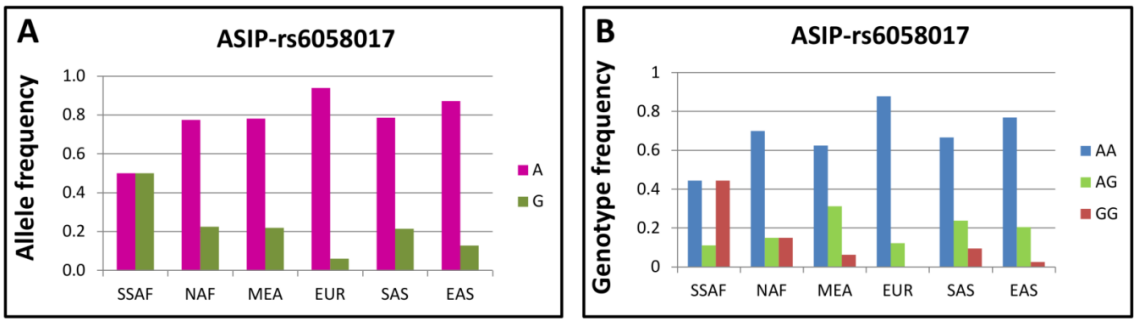


Figure 5.9: Allele and genotype frequencies for the ASIP (rs6058017) locus.
Allele (**A**) and genotype (**B**) frequencies for 146 individuals originating from Sub-Saharan Africa (**SSAF**), North Africa (**NAF**), Middle East (**MEA**), Europe (**EUR**), South Asia (**SAS**) and East Asia (**EAS**).

5.2.4 Genetic Diversity in six major Australian sub-populations

5.2.4.1 *MtDNA and NRY diversity*

The mtDNA and NRY diversity within each of the six population groups was investigated using the haplotype data for 116 males. The number of haplotypes (k), gene diversity (H), mean number of pairwise differences (P) and nucleotide diversity (π) were calculated using Arlequin v3.5 (Excoffier and Lischer, 2010) as described in section 2.2.3.2, with the derived diversity indices shown in Table 5.1. The number of mtDNA and NRY haplotypes observed was variable across all populations, with the lowest number identified in individuals with declared Sub-Saharan Africa BGA (mtDNA k=1, NRY k=3) and the most number of haplotypes occurring in individuals with European BGA (mtDNA k=18, NRY k=11) and East Asia (mtDNA k=19, NRY k=17).

For Sub-Saharan Africans, the gene diversity, mean number of pairwise differences and nucleotide diversity deduced from the mtDNA genotypes were all estimated to be zero, as all individuals in this group belonged to the same haplogroup L (Figure 5.4). The mtDNA gene diversity in the other populations ranged from H=0.63 to H=0.96 with the lowest diversity observed in South Asians, where only 5 haplotypes were observed compared to other groups where k ranged from 8 to 19. The mtDNA gene diversity was comparably high for North Africa, the Middle East, Europe and East Asia (H=0.94-

0.96) and overall was higher than the NRY gene diversity across all 6 populations which ranged from H=0.52 (Sub-Saharan Africa) to H=0.93 (East Asia).

Table 5.1: Diversity indices deduced from the analysis of (A) 21 mtDNA and (B) 28 NRY SNPs in 116 males from six populations.

The sample size (**n**), number of haplotypes (**k**), gene diversity (**H**), mean number of pairwise differences (**P**) and nucleotide diversity (**π**) including standard deviation of the mean are shown for each population group.

(A) mtDNA

Population	n	k	H±SD	P±SD	π ±SD
Sub Saharan Africa	7	1	0.0000±0.0000	0.0000±0.0000	0.0000±0.0000
North Africa	15	10	0.9429±0.0403	2.629±1.486	0.1252±0.0793
Middle East	10	8	0.9556±0.0594	2.311±1.379	0.1100±0.0742
Europe	36	18	0.9381±0.0224	2.222±1.254	0.1481±0.0929
South Asia	14	5	0.6293±0.1227	2.132±1.260	0.1015±0.0674
East Asia	34	19	0.9572±0.0159	2.963±1.589	0.2279±0.1359

(B) NRY

Population	n	k	H±SD	P±SD	π ±SD
Sub Saharan Africa	7	3	0.5238±0.2086	2.476±1.516	0.0884±0.0620
North Africa	15	8	0.8381±0.0852	2.467±1.411	0.0881±0.0566
Middle East	10	6	0.8889±0.0754	2.289±1.368	0.0817±0.0553
Europe	36	11	0.8460±0.0418	2.906±1.562	0.1453±0.0868
South Asia	14	7	0.8242±0.0781	3.231±1.772	0.1154±0.0710
East Asia	34	17	0.9251±0.0294	1.636±0.9899	0.1259±0.0847

The mean number of pairwise differences deduced from the mtDNA genotype data was comparable (P=2.13 to P=2.96) but more variable for the NRY data (P=1.64 to P=3.23) where the lowest P was determined for the East Asian population, which showed the highest gene diversity. The NRY nucleotide diversity was low and ranged from π=0.08 to π=0.15, in contrast, the mtDNA nucleotide diversity was comparable for all populations excluding Sub- Saharan Africa (π=0.10 to π=0.15), but was highest for the East Asian population (π= 0.22).

5.2.4.2 *Autosomal diversity*

Diversity within the declared Sub-Saharan African, North African, Middle Eastern, European, South and East Asian populations was investigated using genotype data from 146 individuals typed with Multiplex 5. Heterozygosity is a commonly applied measure of genetic diversity which describes the probability that an individual will carry two different alleles at a given locus (Hartl and Clark, 2006). Low heterozygosity at a locus is characterised by lower observed heterozygosity (H_o) values than expected heterozygosity (H_e) values and indicates a reduction in diversity at that locus within a population, most likely due to geographical and/or cultural isolation, which promotes inbreeding and non-random mating (Hartl and Clark, 2006).

For each of the 13 autosomal SNP loci, the H_o and H_e values were calculated in each of the six populations using GenAlEx v.6.501 (Peakall and Smouse, 2012) as described in section 2.2.3.3. As evident in Table 5.2, low heterozygosity was observed at multiple autosomal loci and in numerous populations, however in most cases the differences were minimal. Locus OCA2-1 (rs7495174) and ASIP (rs6058017) demonstrated low heterozygosity in all populations except for the Middle East and Europe respectively. Lower H_o than H_e values were observed in three of the six populations for the HERC2 (rs12913832), MATP (rs16891982) and OCA2-3 (rs4778138) loci and in four of the six populations for the OCA2-2 (rs4778241) locus (Table 5.2). The greatest differences (d) between H_o and H_e values were identified at the ASIP (rs6058017) locus in the Sub-Saharan African (d=0.39) and North African populations (d=0.20), the SLC24A5 (rs1426654) locus in the Sub-Saharan African (d=0.35) and Middle Eastern populations (d=0.18) and at the F13B (rs6003) locus in the Sub-Saharan African population

Table 5.2: Analysis of heterozygosity at 13 autosomal SNP loci in six major Australian sub-populations

Observed (H_o) and expected heterozygosity (H_e) was calculated at each locus and for each population from genotype data of 146 individuals obtained using Multiplex 5. Bold values indicate loci where H_o is less than H_e.

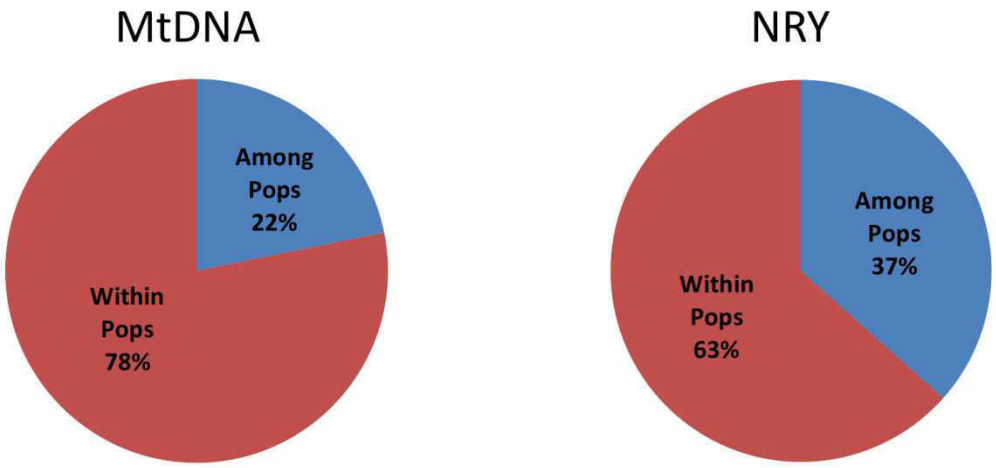
Population		MC1R	DARC	OCA2-1	HERC2	MATP	SLC24A5	F13B	OCA2-2	FUT2	OCA2-3	OCA2-4	IL-6	ASIP
Sub-Saharan Africa	N	9	9	9	9	9	9	9	9	9	9	9	9	9
	H_o	0.000	0.111	**0.222**	0.111	0.111	**0.000**	**0.333**	0.556	0.556	0.444	0.000	0.000	**0.111**
	H_e	0.000	0.105	**0.346**	0.105	0.105	**0.346**	**0.500**	0.475	0.401	0.346	0.000	0.000	**0.500**
North Africa	N	20	20	20	20	20	20	20	20	20	20	20	20	20
	H_o	0.000	0.200	**0.200**	**0.450**	**0.400**	0.000	0.400	**0.400**	0.000	0.650	0.150	0.500	**0.150**
	H_e	0.000	0.180	**0.255**	**0.469**	**0.495**	0.000	0.375	**0.455**	0.000	0.489	0.139	0.375	**0.349**
Middle East	N	16	16	16	16	16	16	16	16	16	16	16	16	16
	H_o	0.000	0.063	0.563	**0.250**	0.313	**0.125**	0.125	**0.438**	0.000	**0.375**	**0.125**	0.375	**0.313**
	H_e	0.000	0.061	0.404	**0.375**	0.482	**0.305**	0.117	**0.498**	0.000	**0.492**	**0.219**	0.305	**0.342**
Europe	N	41	41	41	41	41	41	41	41	41	41	41	41	41
	H_o	0.073	**0.000**	**0.220**	**0.415**	**0.171**	0.024	0.073	**0.317**	0.000	0.366	0.073	**0.390**	0.122
	H_e	0.070	**0.048**	**0.232**	**0.489**	**0.195**	0.024	0.070	**0.404**	0.000	0.356	0.070	**0.433**	0.115
South Asia	N	21	21	21	21	21	21	21	21	21	21	21	21	21
	H_o	0.048	0.048	**0.333**	0.095	0.190	**0.190**	0.286	0.619	0.048	**0.476**	0.048	0.429	**0.238**
	H_e	0.046	0.046	**0.337**	0.091	0.172	**0.308**	0.245	0.495	0.046	**0.499**	0.046	0.337	**0.337**
East Asia	N	39	39	39	39	39	39	39	39	39	39	39	39	39
	H_o	**0.462**	0.000	**0.385**	0.051	**0.077**	**0.000**	**0.051**	**0.308**	0.000	**0.333**	0.487	0.026	**0.205**
	H_e	**0.500**	0.000	**0.436**	0.050	**0.120**	**0.097**	**0.097**	**0.381**	0.000	**0.369**	0.436	0.025	**0.224**

(d=0.17) (Table 5.2). Evidently, many of the candidate autosomal loci have demonstrated low heterozygosity or conversely high homozygosity in one or more populations, which is desirable when distinguishing populations of interest.

5.2.5 Analysis of population structure using mtDNA and NRY SNPs

An Analysis of Molecular Variance (AMOVA) was performed on the 21 mtDNA and 28 NRY SNP haplotype data of 116 males from six populations using GenAlEx v.6.501 (Peakall and Smouse, 2012) as outlined in section 2.2.3.4. For haploid data, the GenAlEx program calculates the fixation index Φ_{PT} which is a variant of the standardised F_{ST} index and provides a measure of the genetic differentiation among populations (Meirmans, 2006). As shown in Figure 5.10, the AMOVA results demonstrated that a greater percentage of genetic variation occurred among populations for the NRY haplotypes (37%) compared to the mtDNA haplotypes (22%). For both sets of haploid data the *p*-value was less than 0.05, resulting in the rejection of the null hypothesis and indicating that there is a significant genetic difference for the mtDNA and NRY genotypes among the Sub-Saharan African, North African, Middle Eastern, European, South Asian and East Asian populations (Figure 5.10).

Genetic distance measures can be used to investigate the degree of genetic relationship that exists between populations, with the most common measures of genetic distance including F_{ST} (or other analogous fixation indices) (Wright, 1965, Excoffier *et al.*, 1992) and Nei's genetic distance (Nei, 1987). Genetic distance matrices of pairwise Φ_{PT} and M values (measures of migrant exchange) were calculated from the mtDNA and NRY data for the six populations using GenAlEx v.6.501 (Peakall and Smouse, 2012), as detailed in section 2.2.3.5 and results are provided in Table 5.3. The majority of pairwise genetic distances were significant ($p<0.05$) with the exception of East Asia/South Asia, Europe/Middle East and North Africa/Middle East for the 21 SNP mtDNA haplotype data (Table 5.3A) and North Africa/Middle East for the 28 SNP NRY haplotype data (Table 5.3A). These population pairs also demonstrated the lowest pairwise Φ_{PT} values indicating that they are closely genetically related ($\Phi_{PT}<0.06$) (Table 5.3). In particular, the Europe/Middle East and North Africa/Middle East pairs showed very low pairwise Φ_{PT} values for the mtDNA data ($\Phi_{PT=}$ 0.005 and $\Phi_{PT=}$ 0.000 respectively) (Table 5.3A).



		df*	Sum of Squares	Mean Square	Est. Variance	% Variation	Φ_{PT}	P
mtDNA	Among Pops	5	36.659	7.332	0.340	22	0.218	**0.001**
	Within Pops	110	134.099	1.219	1.219	78		
	Total	115	170.758		1.559	100		
NRY	Among Pops	5	76.286	15.257	0.774	37	0.366	**0.001**
	Within Pops	110	147.514	1.341	1.341	63		
	Total	115	223.800		2.115	100		

*df= degrees of freedom

Figure 5.10: AMOVA results for the analysis of six Australian sub-populations using 21 mtDNA and 28 NRY SNPs.

The mtDNA and NRY haplotypes of 116 males originating from South Asia, East Asia, Europe, Middle East, North Africa and Sub-Saharan Africa were analysed. Significant *p*-values (<0.05) are shown in bold.

Table 5.3: Genetic distance matrices of population pairwise Φ_{PT} and M values from the analysis of 21 mtDNA and 28 NRY SNPs in six Australian sub-populations.

Population pairwise Φ_{PT} values are shown above the diagonal and population pairwise M values are shown below the diagonal. The **(A)** mtDNA and **(B)** NRY haplotypes of 116 males originating from **SAS**=South Asia, **EAS**=East Asia, **EUR**= Europe, **MEA**=Middle East, **NAF**=North Africa, **SSAF**=Sub-Saharan Africa were analysed. Significant *p*-values are denoted as '**' if $p<0.01$ or '*' if $p<0.05$.

(A) mtDNA

	SAS	EAS	EUR	MEA	NAF	SSAF
SAS		0.028	0.341**	0.282**	0.139*	0.272*
EAS	17.213		0.281**	0.212**	0.139**	0.185*
EUR	0.697	1.278		0.005	0.056*	0.472**
MEA	1.276	1.856	91.193		0.000#	0.510**
NAF	3.107	3.101	8.380	N/A		0.364**
SSA	1.338	2.204	0.560	0.481	0.874	

negative Φ_{PT}= -0.015 converted to 0

(B) NRY

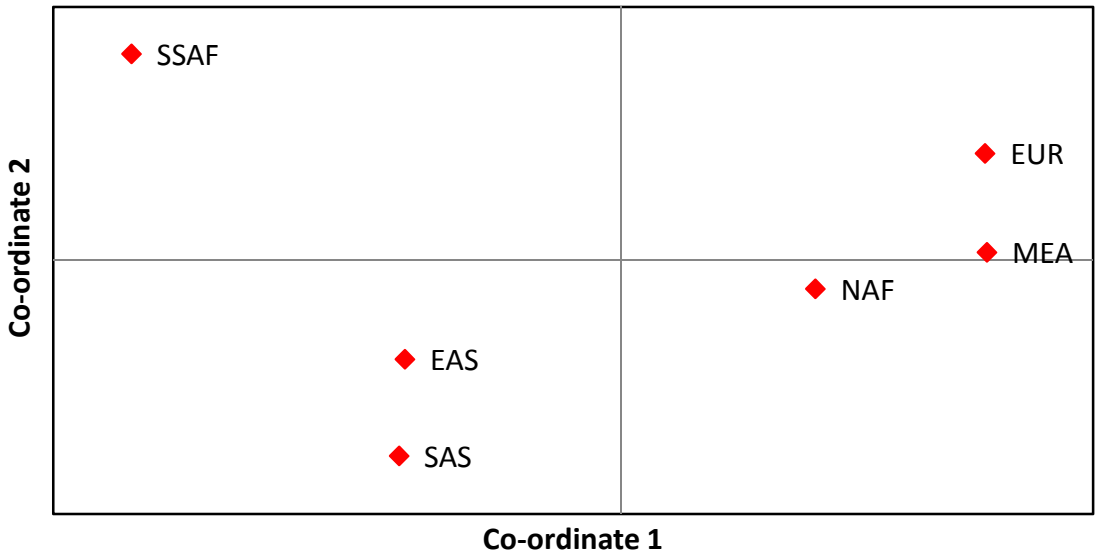
	SAS	EAS	EUR	MEA	NAF	SSAF
SAS		0.363**	0.103*	0.184**	0.319**	0.425**
EAS	0.879		0.407**	0.378**	0.474**	0.570**
EUR	4.364	0.728		0.234**	0.367**	0.447**
MEA	2.212	0.823	1.639		0.062	0.361**
NAF	1.065	0.555	0.863	7.621		0.247**
SSA	0.676	0.378	0.620	0.885	1.522	

From the mtDNA haplotypes, the most differentiated pairs of populations were Sub-Saharan Africa/ Middle East (51%), Sub-Saharan Africa/Europe (47.2%), Sub-Saharan Africa/North Africa (36.4%) and Europe/South Asia (34.1%) (Table 5.3A). On the other hand from the NRY haplotypes, the greatest genetic differentiation was observed between Sub- Saharan Africa/East Asia (57%), North Africa/East Asia (47.4%), Sub-Saharan Africa/Europe (44.7%) and Sub-Saharan Africa/South Asia (42.5%) (Table 5.3B).

The distance matrix of pairwise M values also correlates with the genetic relationship of the six populations inferred from the pairwise Φ_{PT} values (Table 5.3). The M value provides a measure of the absolute number of migrants exchanged between two populations. As evident in Table 5.3, a greater number of migrants (larger M value) were observed for population pairs showing low Φ_{PT} values and which shared close genetic relationships. For example for the mtDNA data, the Middle East and Europe populations appeared very closely genetically related ($\Phi_{PT=}$ 0.005) with M= 91.193.

Principal Coordinates Analysis (PCoA) was used to produce a graphical display of the complex mtDNA and NRY genetic distance matrices. This multivariate technique identifies the major axes of variation within a multidimensional data set, such that the first two or three axes will usually reveal the greatest separation between groups, with each successive axis explaining proportionately less of the total variation. In essence, by comparing the first two axes, the separation among populations can be visualised. As described in section 2.2.3.6, a PCoA was conducted on the mtDNA and NRY pairwise population Φ_{PT} distance matrices shown above the diagonal in Table 5.3 using GenAlEx v.6.501 (Peakall and Smouse, 2012). The PCoA plot produced from the analysis of the mtDNA distance matrix revealed that the six populations are separated into three distinct geographical clusters (Figure 5.11A). The European, Middle Eastern and North African populations are grouped separately into a Western Eurasia cluster, in contrast with the East and South Asia populations, which are grouped into an Eastern Eurasia cluster (Figure 5.11A). The Sub-Saharan African population can be distinctly separated from the other six populations and the two Western and Eastern Eurasia clusters (Figure 5.11A).

(A) mtDNA



(B) NRY

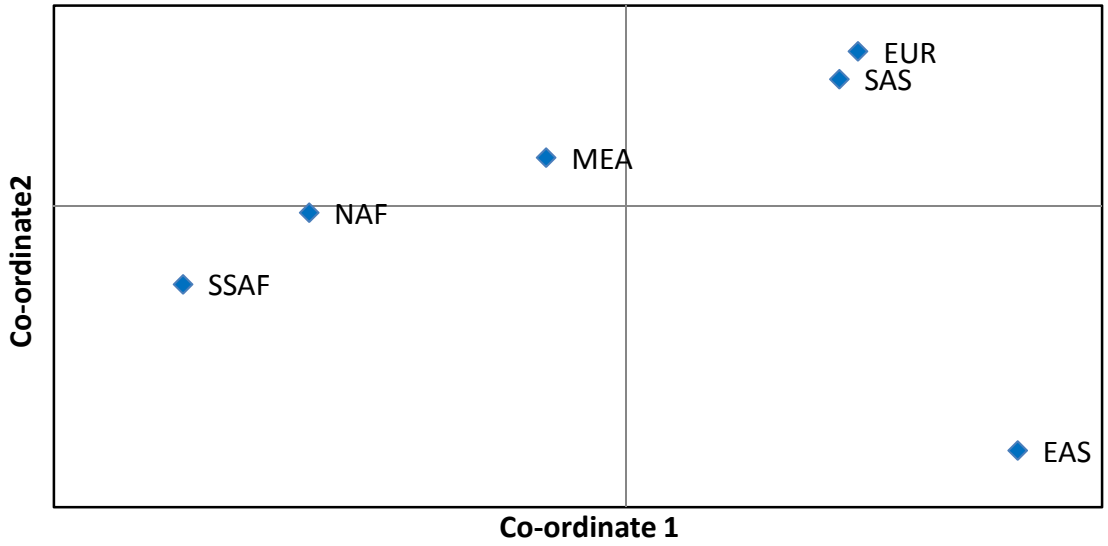


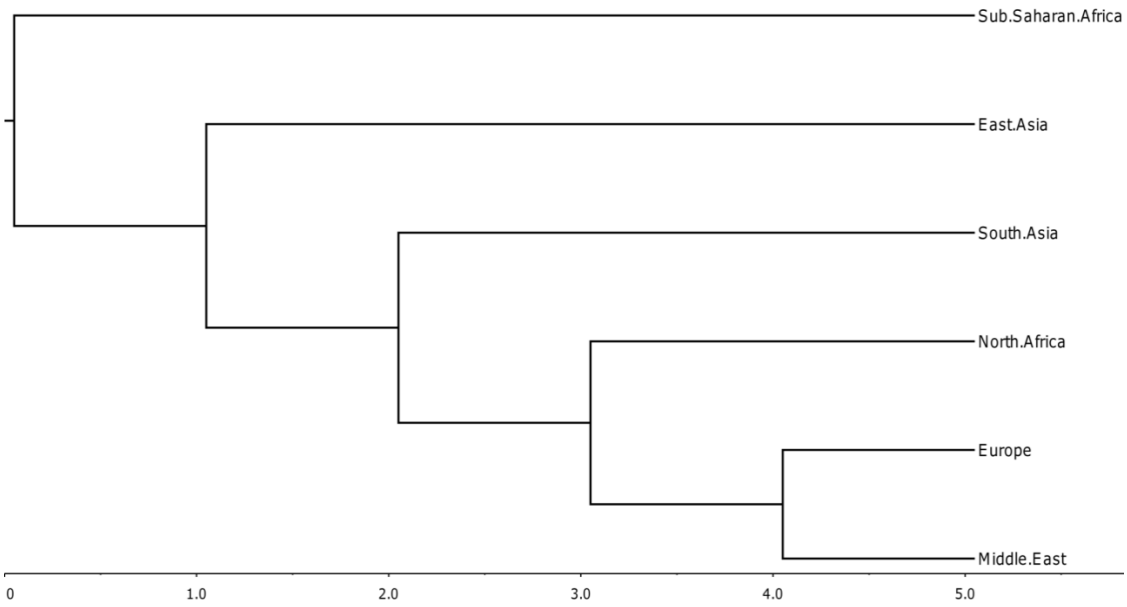
Figure 5.11: Principal Coordinates Analysis of mtDNA and NRY genetic distances for six Australian sub-populations.

The pairwise population Φ_{PT} distance matrices shown above the diagonal in Table 5.3A & B and derived from the **(A)** 21 mtDNA and **(B)** 28 NRY SNP haplotypes of 116 males originating from **SAS**=South Asia, **EAS**=East Asia, **EUR**= Europe, **MEA**=Middle East, **NAF**=North Africa, **SSAF**=Sub-Saharan Africa were analysed using the covariance-standardised method.

On the other hand, four population clusters were identified from the PCoA analysis of the calculated NRY distance matrix (Figure 5.11B). The East Asian population appears isolated, whilst the Europe and South Asia populations cluster together. The Sub-Saharan African and Middle Eastern populations are separated from each other and other populations; however, the North African individuals appear to cluster more closely with the Sub-Saharan Africa group (Figure 5.11B). The six populations are grouped into four clusters somewhat also correlated with broad geographical regions-Africa (Sub-Saharan and North Africa), Western Eurasia (Europe and South Asia), Central Eurasia (Middle East) and Eastern Eurasia (East Asia). Superior population differentiation is apparent for the NRY markers, particularly when distinguishing Middle Eastern individuals from other populations, although more difficult differentiation of South Asian and European individuals using the candidate NRY SNPs was also observed (Figure 5.11B).

The calculated mtDNA and NRY distance matrices shown in Table 5.3 were also used to construct neighbour-joining (NJ) trees for all six populations through the TREEVIEW v1.6.6 and FigTree v1.4 programs, as described in section 2.2.3.7. The mtDNA NJ tree depicted in Figure 5.12A highlights the close genetic relationship of the North Africa, Middle East and Europe populations as well as the East Asia and South Asia populations, which correlates with the PCoA results (Figure 5.11A). Surprisingly, the Sub-Saharan Africa population appears to be the most differentiated from the other populations but is most closely related to the East Asia population (Figure 5.12A). The most distant relationship was observed between Sub-Saharan Africa and the Middle East, followed by Europe and North Africa (Figure 5.12A). In contrast, different genetic relationships between the six populations were inferred for the NRY NJ tree. As shown in Figure 5.12B, the Sub-Saharan Africa population appears to be most closely related to the North Africa and Middle East populations and the most genetically distant from Europe, South Asia and East Asia. The close genetic relationship between the Europe and South Asia populations is also highlighted. The genetic relationship for the six populations inferred from the NRY haplotypes shows a distinct correlation with geographical location, as populations in geographical proximity demonstrated closer genetic relationships, except, for Europe and South Asia (Figure 5.12B). The NRY NJ

(A) mtDNA



(B) NRY

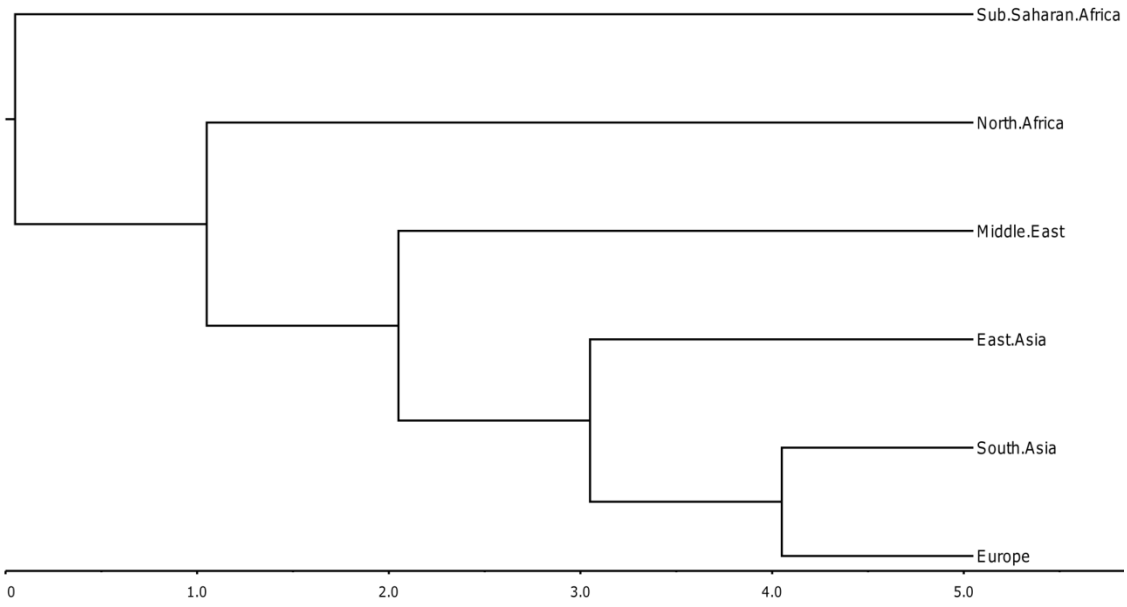


Figure 5.12: Neighbour-joining tree for six Australian sub-populations based on calculated mtDNA and NRY genetic distances.

The tree was constructed from the Φ_{PT} distance matrix shown above the diagonal in Table 5.3A (mtDNA) and Table 5.3B (NRY) derived from the **(A)** 21 mtDNA and **(B)** 28 NRY SNP haplotypes respectively of 116 males originating from the six populations indicated in the cladogram.

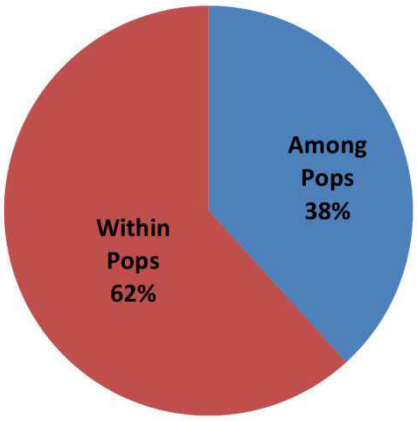
tree results were also analogous with the population clustering observed through PCoA of NRY Φ_{PT} distances (Figure 5.11B).

5.2.6 Analysis of population structure using autosomal SNPs

As for the mtDNA and NRY markers, an AMOVA was also performed on the autosomal dataset using GenAlEx v.6.501 (Peakall and Smouse, 2012) as described in section 2.2.3.4. The fixation index F_{ST} was calculated to measure the degree of genetic differentiation among the six populations. As shown in Figure 5.13, 38% of the genetic variation was found to occur among populations. A *p*-value of 0.001 confirmed that a significant difference exists between the 13 autosomal SNP genotypes among the Sub-Saharan Africa, North Africa, Middle East, Europe, South Asia and East Asia populations.

Additionally, genetic distance matrices of population pairwise F_{ST} and M values were calculated for the six populations using GenAlEx v.6.501 (Peakall and Smouse, 2012) as detailed in section 2.2.3.5. As depicted in Table 5.4, highly significant pairwise F_{ST} distances were observed between all population pairs tested (p <0.01). The Middle East and South Asia populations demonstrated the lowest pairwise F_{ST} value (F_{ST}=0.065) followed by Europe and the Middle East (F_{ST}=0.098), indicating that these populations were the most closely genetically related based on the autosomal genotypes. The greatest genetic differentiation was identified between Europe and East Asia (55.8%), Sub-Saharan Africa and Europe (55.2%), North Africa and East Asia (51%) and Sub-Saharan Africa and East Asia (43.2%).

The distance matrix of pairwise M values shown below the diagonal in Table 5.4, correlates with the genetic relationships inferred from pairwise F_{ST} values derived from the autosomal genotypes. As was observed for the mtDNA and NRY haplotype data (Table 5.3), the largest absolute number of migrants exchanged (M value) was identified between closely genetically related populations (Table 5.4). These populations are typically in geographical proximity, which promotes gene flow due to facilitated migration, as is exemplified by the Middle Eastern and South Asian



	df*	Sum of Squares	Mean Square	Est. Variance	% Variation	F_{ST}	P
Among Pops	5	223.911	44.782	0.933	38	0.382	**0.001**
Within Pops	286	430.825	1.506	1.506	62		
Total	291	654.736		2.439	100		

*df= degrees of freedom

Figure 5.13: AMOVA results for the analysis of six Australian sub-populations using 13 autosomal SNPs.

The autosomal genotypes of 146 individuals originating from South Asia, East Asia, Europe, Middle East, North Africa and Sub-Saharan Africa were analysed. Significant *p*-values ($p<0.05$) are shown in bold.

Table 5.4: Genetic distance matrix of population pairwise F_{ST} and M values from the analysis of 13 autosomal SNPs in six Australian sub-populations.

Population pairwise F_{ST} values are shown above the diagonal and population pairwise M values are shown below the diagonal. The genotypes of 146 individuals originating from **SAS**=South Asia, **EAS**=East Asia, **EUR**= Europe, **MEA**=Middle East, **NAF**=North Africa, **SSAF**=Sub-Saharan Africa were analysed. Significant *p*-values are denoted as '**' if $p<0.01$.

	SAS	EAS	EUR	MEA	NAF	SSAF
SAS		0.351**	0.288**	0.065**	0.233**	0.314**
EAS	0.461		0.558**	0.370**	0.510**	0.432**
EUR	0.618	0.198		0.098**	0.269**	0.552**
MEA	3.584	0.425	2.296		0.177**	0.344**
NAF	0.822	0.240	0.680	1.160		0.250**
SSAF	0.545	0.329	0.203	0.477	0.748	

populations, which were found to be the most closely related (F_{ST}=0.065) and showed the highest M value of 3.584 (Table 5.4).

Furthermore, PCoA was performed to graphically visualise the autosomal pairwise population F_{ST} distance matrix shown above the diagonal in Table 5.4 (refer to section 2.2.3.6). The PCoA plot shown in Figure 5.14 revealed that based on the autosomal genotype data obtained from Multiplex 5, the six populations are divided into at least four distinct clusters. The Sub-Saharan African, North African, East Asian and European populations can be distinctly separated from each other (Figure 5.14). However, the Middle Eastern and South Asian populations cluster closely between Europe and North Africa and may be more difficult to distinguish from these populations, but are separated enough from Sub-Saharan Africa and East Asia to be distinguished from these groups (Figure 5.14).

A NJ tree was also constructed from the autosomal pairwise population F_{ST} distance matrix using the TREEVIEW v1.6.6 and FigTree v1.4 programs as described in section 2.2.3.7. The autosomal NJ tree produced is depicted in Figure 5.15 and demonstrates clustering of the six populations into two groups; with the Sub-Saharan and North African populations closely related and forming one cluster that can be distinctly separated from other non-African populations (Figure 5.15). The remaining populations clustered together, with East Asia observed to be more closely related to South Asia, which was most closely related to the European population (Figure 5.15). Europe was found to share a close genetic relationship with the Middle Eastern population, in accordance with PCoA results (Figure 5.14). The greatest separation was identified between Africa, East Asia and Europe (Figure 5.15) and these results along with the PCoA plot shown in Figure 5.14, provide a good overview of the extent of differentiation that can be achieved between the six populations using the 13 autosomal SNPs incorporated in Multiplex 5, which appears to be greatest for continental populations rather than those that are in geographical proximity within continents.

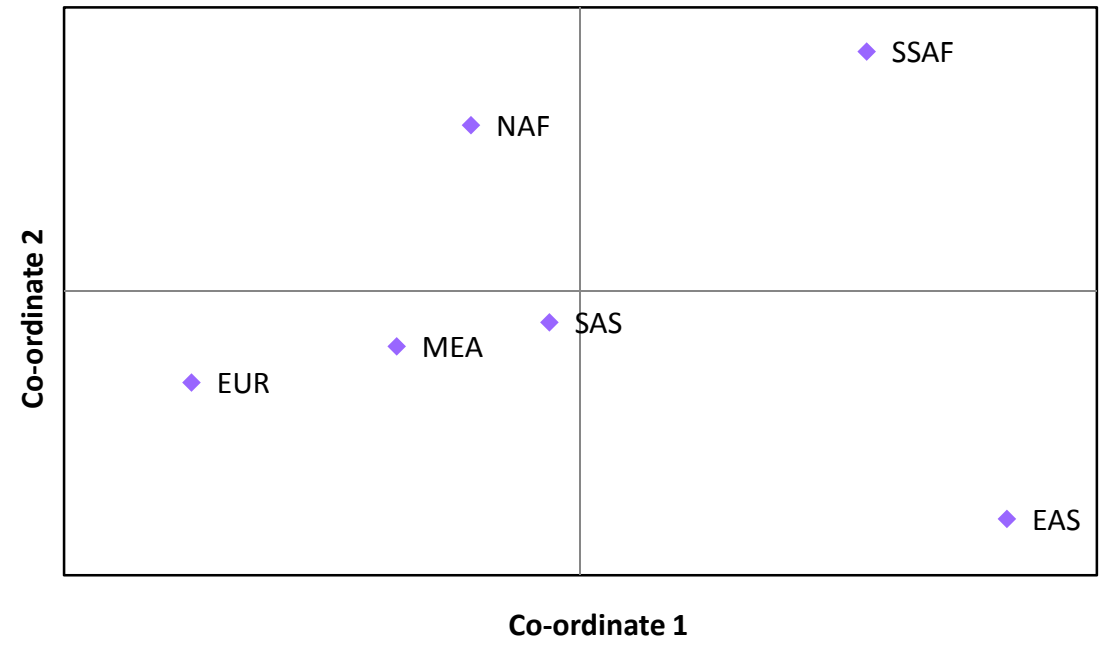


Figure 5.14: Principal Coordinates Analysis of autosomal genetic distances for six Australian sub-populations.

The pairwise population F_{ST} distance matrix shown above the diagonal in Table 5.4 and derived from the 13 autosomal SNP genotypes of 146 individuals originating from **SAS**=South Asia, **EAS**=East Asia, **EUR**= Europe, **MEA**=Middle East, **NAF**=North Africa, **SSAF**=Sub-Saharan Africa were analysed using the covariance-standardised method.

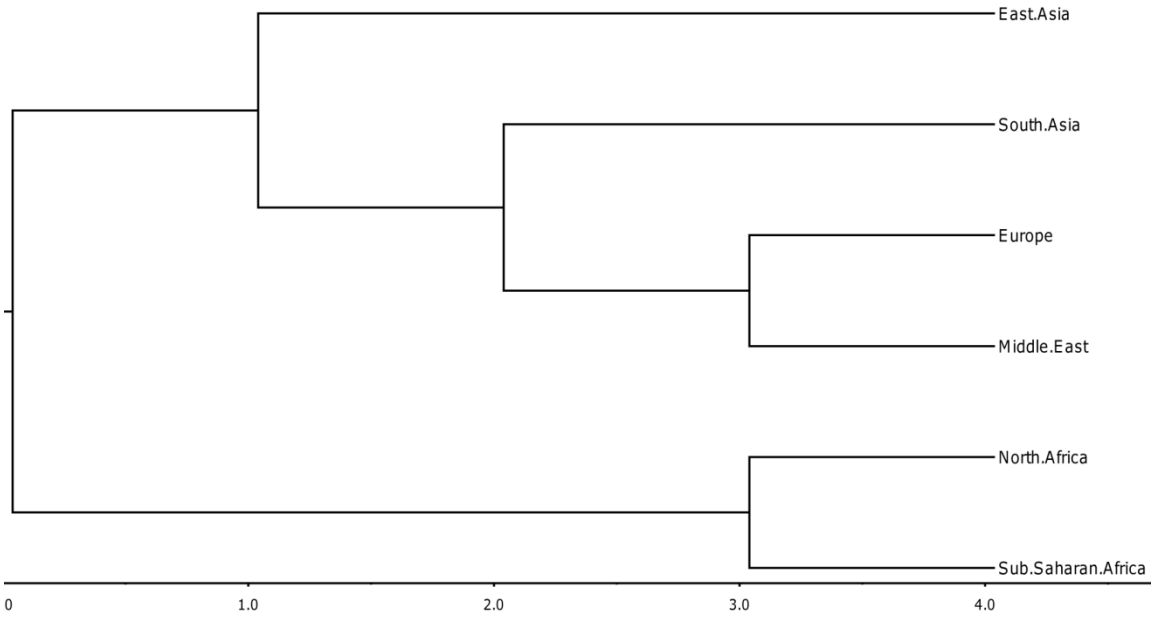


Figure 5.15: Neighbour-joining tree for six Australian sub-populations based on calculated autosomal genetic distances.

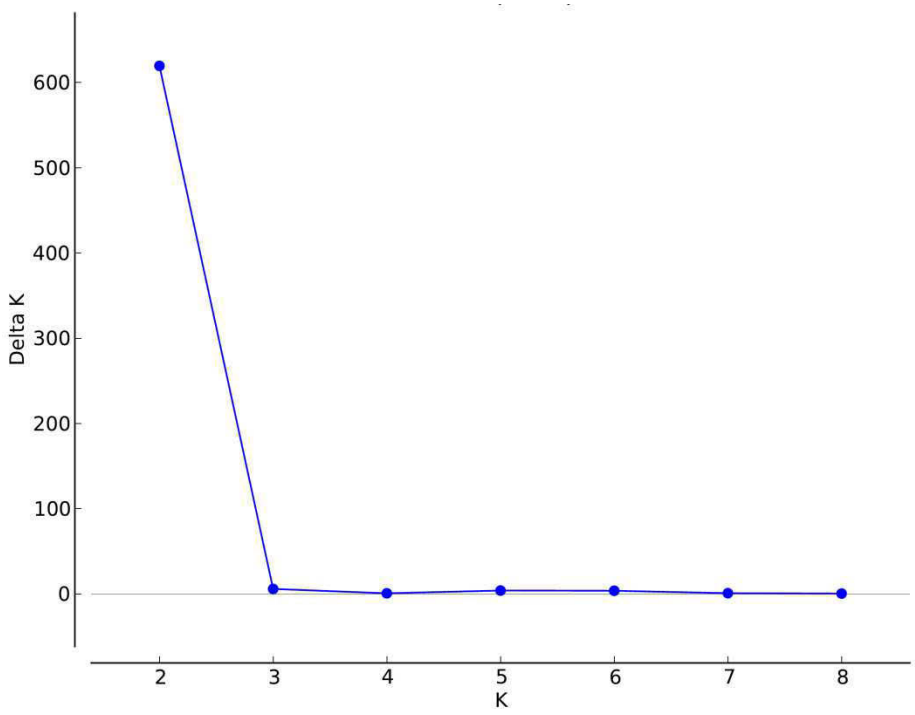
The tree was constructed from the F_{ST} distance matrix shown above the diagonal in Table 5.4 and derived from the 13 autosomal SNP genotypes of 146 males originating from the six populations indicated in the cladogram.

A further analysis of population sub-structure within the autosomal dataset was performed using STRUCTURE v2.3.4 (Pritchard *et al.*, 2000, Falush *et al.*, 2003, 2007, Hubisz *et al.*, 2009) as outlined in section 2.2.3.8 This program uses clustering methods to identify sub-groups with similar allele frequencies within a large population and can estimate shared ancestry. The admixture model and the correlated allele frequency option recommended in cases when mild population sub-structure is expected were used to investigate the presence of sub-structure within our Australian dataset through two separate analyses (Falush *et al.*, 2003, Evanno *et al.*, 2005). The first analysis does not consider prior population information when assigning individuals to ancestral populations and the second analysis based on the LOCPRIOR model considers the pre-specified sampling location (or population of origin) when estimating ancestry proportions. For each analysis, 10 independent runs for each K=1 to K=8 were performed using the parameters detailed in section 2.2.3.8. The STRUCTURE results were analysed using STRUCTURE Harvester (Earl and vonHoldt, 2012) to collate the data at each K over the multiple runs and produce the necessary output files required to create bar plots using CLUMPP v1.1.2 and DISTRUCT v1.1 (refer to section 2.2.3.8).

The Evanno method (Evanno *et al.*, 2005) available through the STRUCTURE Harvester program was used to estimate the most likely number of population clusters (K) within our dataset representing the greater Australian population. As shown in Figure 5.16, the highest ΔK was observed at K=2, indicating that the dataset clusters into two genetic entities irrespective of whether prior population information is considered or not. However, these results are conflicting as the dataset is composed of individuals known to have originated from different sub-populations and who have been shown to cluster into at least 3 separate population groups through PCoA (Figure 5.14).

The bar plots of individual ancestry proportions for each K provided a better representation of the most likely number of population clusters (Figure 5.17). In Analysis 1, where prior population information was not considered for clustering analysis, the most likely number of populations was identified as two; where individuals from Sub-Saharan Africa and East Asia clustered together whilst those originating from North Africa, Europe, the Middle East and South Asia formed a second cluster (Figure 5.17). Contrastingly, when using the LOCPRIOR model (Analysis 2) greater levels of

(A) Analysis 1: No prior population information



(B) Analysis 2: LOCPRIOR model

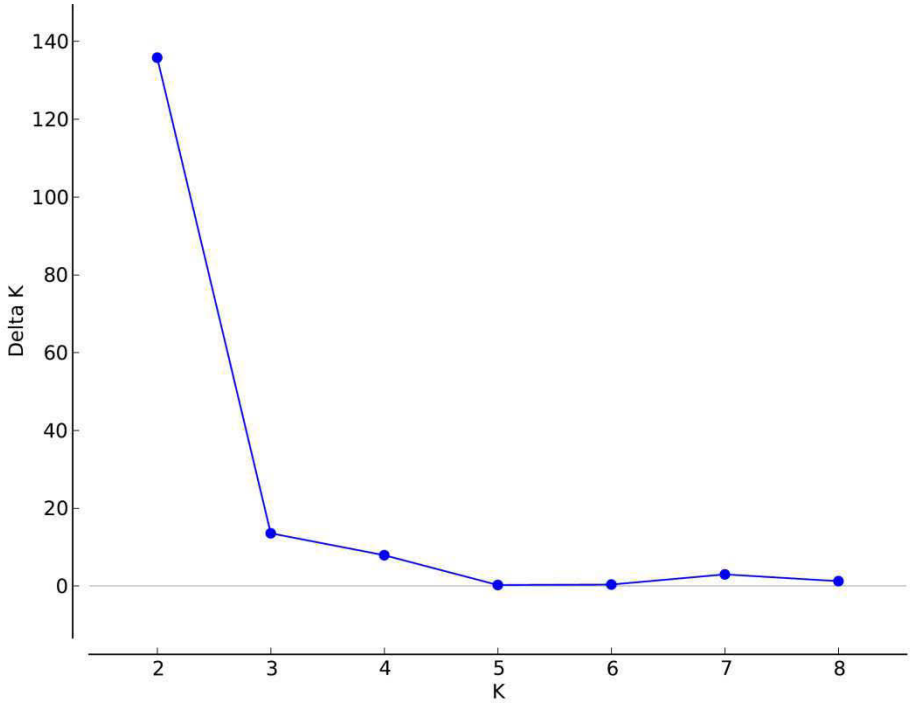


Figure 5.16: Estimation of the number of *K* clusters in the autosomal dataset using the Evanno method.

STRUCTURE Harvester plots of Δ*K* at *K*=1 to *K*=8 over 10 independent runs following the Evanno *et al.* (2005) method for identifying the true number of clusters, *K* when (A) prior population information is not considered and (B) the LOCPIROR model is applied. For both analyses, *K*=2 was identified as the most likely number of clusters.

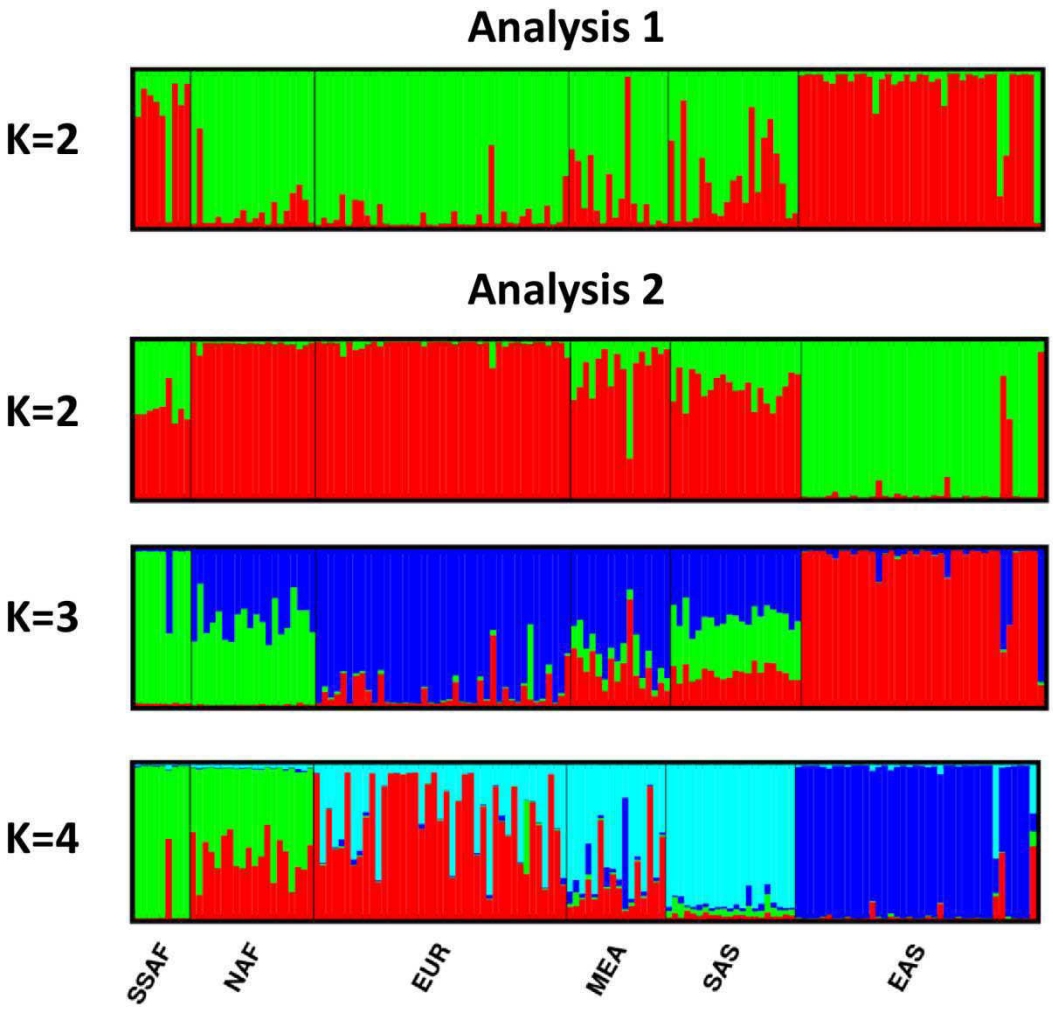


Figure 5.17: Estimation of population sub-structure within the autosomal dataset using STRUCTURE v2.3.4.

Two separate clustering analyses were performed: Analysis 1 did not consider the *a priori* population of origin and Analysis 2 applied the LOCPRIOR model to consider the population of origin when estimating the ancestry proportions. Bar plots were produced using DISTRUCT v1.1. Each of the 146 individuals in the dataset is depicted as a single vertical line, which is partitioned into coloured segments that represent an individual's proportion of membership in each of the ancestral population clusters. The red, green, dark and light blue colours distinguish the clusters. Black lines are used to separate individuals according to their population of origin which are identified in the plots as **SSAF**= Sub-Saharan Africa, **NAF**= North Africa, **EUR**= Europe, **MEA**= Middle East, **SAS**= South Asia, **EAS**= East Asia. The most likely number of ancestral populations for Analysis 1 is two but this increases to four when the *a priori* population of origin is considered (Analysis 2).

population structure were detected at *K*=2. Individuals from North Africa and Europe formed one cluster, East Asia clustered separately and Sub-Saharan Africans showed approximately 50% membership to the two clusters (Figure 5.17). On the other hand, individuals from the Middle East and South Asia demonstrated greater proportions of ancestry to the North African/European cluster compared to the East Asian cluster (Figure 5.17).

As evident in Figure 5.17, increasing the number of clusters to *K*=3 showed improved partitioning of individuals according to their population of origin when using the LOCPRIOR model. The Sub-Saharan Africans, Europeans and East Asians formed separate clusters, whilst North Africans demonstrated shared ancestry between the Sub-Saharan and European groups. Middle Eastern individuals showed a greater proportion of membership with the European cluster, whilst South Asians showed equal ancestry proportions with the Sub-Saharan African, European and East Asian populations. For our dataset, the most likely number of ancestral populations was identified to be four with further separation of the South Asian individuals achieved at this level (Figure 5.17).

The STRUCTURE results revealed the presence of population sub-structure within our autosomal dataset as was expected based on PCoA results (Figure 5.14) and provide a summary of the population differentiation that may be achieved using the 13 autosomal SNPs incorporated in Multiplex 5. Using the LOCPRIOR model to consider the *a priori* sampling location increases the number of ancestral populations, which were found to be mostly analogous with major sub-populations in Australia. Four major clusters were identified comprising individuals of Sub-Saharan African, European, South Asian and East Asian origin. Middle Eastern individuals appeared to cluster more closely with Europeans but drew a large proportion of ancestry from South Asia. In contrast, the North Africans appeared to form a separate group drawing 50% membership with the Sub-Saharan African and European ancestral populations (Figure 5.17).

5.2.7 Assessing departures from independence at autosomal SNP loci

Departures from Hardy-Weinberg (HWE) and Linkage equilibrium (LE) present in the genotype data from 146 individuals typed using Multiplex 5 were investigated using Fisher's exact test for allelic independence available in the Genetic Data Analysis (GDA) software v1.1 (Lewis and Zaykin, 2002) (for details refer to section 2.2.3.9). HWE assumes random mating within a population of infinite size where migration, mutation and natural selection are non- existent (Hartl and Clark, 2006). LE is used to describe the random association of genes or alleles at a particular locus which is affected by the rate of recombination and is attained through HWE, namely random mating and its other assumptions (Hartl and Clark, 2006). Independence testing was performed prior to inferring bi-parental BGA from the autosomal genotypes using maximum likelihood estimates (see section 5.2.9) because these methods assume HWE and random mating. The *p*-values derived from the HWE and LE analysis are shown in Table 5.5 and Table 5.6 respectively. As explained in section 2.2.3.9, the Bonferroni method was applied to correct for multiple testing and a significance threshold of $p<0.0038$ (13 tests) and $p<0.0006$ (78 tests) was set for the HWE and LE analysis respectively. Only the LE *p*-values for locus combinations that demonstrated significant results (Bonferroni $p<0.0006$) are depicted in Table 5.6.

As shown in Table 5.5, all six populations demonstrated no departures from HWE for any of the 13 autosomal SNPs after Bonferroni correction (all $p>0.0038$). Contrastingly, departures from LE were observed for four pairs of loci, which demonstrated significant *p*-values ($p<0.0006$) after Bonferroni correction (Table 5.6). These departures were identified in the East Asian population (3 departures) followed by the European population (4 departures). No locus combinations showed departures from LE in the remaining four populations (Table 5.6). The locus pairs that were in linkage disequilibrium (LD) were those located within the same pigmentation gene (i.e. OCA2-1, OCA2-2 and OCA2-3) or which are located in close proximity, on the same chromosome, as was the case for the OCA2-2 and HERC2 loci.

Table 5.5: Hardy-Weinberg equilibrium analysis of 13 autosomal SNP loci in six major Australian sub-populations

The *p*-values were calculated using Fisher's exact test and 10000 permutations for each locus in the Sub-Saharan Africa (**SSAF**), North Africa (**NAF**), Middle East (**MEA**), Europe (**EUR**), South Asia (**SAS**) and East Asia (**EAS**) populations. All *p*-values were not significant following Bonferroni correction for 13 tests ($p<0.0038$).

Locus	SSAF	NAF	MEA	EUR	SAS	EAS
MC1R	1.0000	1.0000	1.0000	1.0000	1.0000	0.7478
DARC	1.0000	1.0000	1.0000	0.0132	1.0000	1.0000
OCA2-1	0.3351	0.3493	0.2553	0.5359	1.0000	0.4634
HERC2	1.0000	1.0000	0.1990	0.3434	1.0000	1.0000
MATP	1.0000	0.4023	0.2858	0.3843	1.0000	0.1299
SLC24A5	0.0121	1.0000	0.0447	1.0000	0.1200	0.0003
F13B	0.5032	1.0000	1.0000	1.0000	1.0000	0.0805
OCA2-2	1.0000	0.6325	0.6394	0.2404	0.3879	0.2248
FUT2	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
OCA2-3	1.0000	0.3589	0.3425	1.0000	1.0000	0.6622
OCA2-4	1.0000	1.0000	0.1888	1.0000	1.0000	0.7116
IL-6	1.0000	0.2832	1.0000	0.4933	0.5384	1.0000
ASIP	0.0356	0.0235	1.0000	1.0000	0.1930	0.4861

Table 5.6: Linkage equilibrium analysis of 13 autosomal SNP loci in six major Australian sub-populations

The *p*-values were calculated for each locus pair using Fisher's exact test and 10000 permutations in each population. The population codes used are the same as in Table 5.5. Only locus pairs demonstrating significant results are shown. Significant *p*-values ($p<0.0006$) indicating departures from LE following Bonferroni correction for 78 pairwise comparisons are shown in bold.

Locus combination	SSAF	NAF	MEA	EUR	SAS	EAS
OCA2-1/OCA2-2	0.7598	0.0084	0.2356	**0.0000**	0.3419	**0.0000**
OCA2-1/OCA2-3	0.2809	0.0909	0.0805	**0.0001**	0.0212	**0.0000**
HERC2/OCA2-2	0.1102	0.0252	0.0673	**0.0000**	0.0950	0.0075
OCA2-2/OCA2-3	1.0000	0.0009	0.2297	**0.0002**	0.4418	**0.0000**

5.2.8 Inference of maternal and paternal BGA

The maternal and paternal regional BGA of each individual was inferred from the geographical region of origin of their derived mtDNA and NRY haplogroup respectively, according to the key specified in Figure 3.1 and Figure 3.2. Prediction correlation was determined as the percentage of males for whom the declared BGA correlated with the geographic region of origin of the derived mtDNA and NRY haplogroup. Since many of the SNPs selected define haplogroups that are prevalent in multiple populations in close geographical proximity, prediction was considered correct if one of the continental regions of origin for the haplogroup matched the declared BGA of the individual. For example, for an East Asian individual belonging to mtDNA haplogroup M* which is prevalent in South and East Asia the inferred maternal BGA would be considered as correct.

Overall, higher prediction correlations were observed for the NRY SNPs compared to the mtDNA SNPs, with 100% of East Asian and Middle Eastern individuals, 97.2% of Europeans, 86.7% of North Africans and 85.7% of Sub-Saharan Africans and South Asian males showing correlations between the declared BGA and geographical region of origin of the derived NRY haplogroups (Table 5.7). On the other hand, though the maternal regional BGA correlation was lower for declared North African (73.3%), Middle Eastern (90%) and European (77.8%) males, the mtDNA SNPs were more successful in correctly inferring the maternal BGA of individuals of Sub-Saharan African (100%) and South Asian (100%) descent compared to the NRY SNPs (Table 5.7).

To investigate the overall prediction correlations between the declared and the inferred regional BGA that was achieved with the hierarchical multiplex SNP assay (Multiplex 1 to 4), the degree of equivalence between the inferred maternal and paternal population or geographic region of origin was determined. Note that only individuals for whom both maternal and paternal grandparents were declared to have the same BGA were chosen for genotyping, thus we would expect a reasonably high degree of correlation between maternal and paternal BGA. As shown in Table 5.8, a high degree of prediction correlation was achieved for East Asian (100%), Middle Eastern (90%), Sub-Saharan African (85.7%), South Asian (85.7%) and European (77.8%) males when considering

both maternal and paternal regional BGA. An intermediate prediction correlation of 60% was observed for the inference of North African ancestry. Whilst some of these percentage correlations are lower than those achieved for the separate maternal and paternal BGA inferences (Table 5.7), utilising both sets of data can assist in narrowing down the population or geographic region of origin for haplogroups that are prevalent in multiple populations; particularly since the global distribution for the selected NRY haplogroups is more specific and overlaps less than that of the mtDNA haplogroups (Figure 3.1 and Figure 3.2).

Table 5.7: Prediction correlation of the Multiplex 1 to 4 assays for the inference of maternal and paternal regional BGA

The correlation between declared and inferred regional BGA for the Multiplex 1 to 4 assays was calculated for 116 males and refers to the frequency (%) of males for whom the declared BGA correlates with the region of geographic origin of their derived mtDNA and NRY haplogroup.

Declared BGA (N)	Correlation (%)	
	Maternal BGA	Paternal BGA
Sub Saharan Africa (7)	100.0	85.7
North Africa (15)	73.3	86.7
Middle East (10)	90.0	100.0
Europe (36)	77.8	97.2
South Asia (14)	100.0	85.7
East Asia (34)	100.0	100.0

Table 5.8: Overall prediction correlation of Multiplex 1 to 4 for the inference of regional BGA

The overall prediction correlation of the Multiplex 1 to 4 assays was calculated for 116 males and refers to the frequency (%) of males for whom the inferred maternal and paternal regional BGA correlates with their declared BGA.

Declared BGA (N)	Overall prediction correlation (%)
Sub Saharan Africa (7)	85.7
North Africa (15)	60.0
Middle East (10)	90.0
Europe (36)	77.8
South Asia (14)	85.7
East Asia (34)	100.0

These results indicate that the selected mtDNA and NRY SNPs incorporated in the hierarchical multiplex SNP assay comprising Multiplex 1 to 4 can be successfully used to infer the maternal and paternal regional BGA of males of Sub-Saharan African, North African, Middle Eastern, European, South and East Asian origin with a generally high degree of correlation when considered separately and in combination (mtDNA: 73.3-100%, NRY: 85.7-100%, Combined: 60-100%). However, the closely related populations such as North Africa and South Asia demonstrated less accurate maternal and paternal BGA inferences based on the selected mtDNA and NRY SNPs respectively, suggesting that further differentiation of these populations with additional markers is required.

An additional 30 female samples was typed using Multiplex 1 to 4. The mtDNA haplogroups were derived and the maternal BGA was inferred for each individual as for the male samples. In all female samples the NRY markers were not detected which confirmed the specificity of the PCR and SBE primers (data not shown). The distribution of the derived mtDNA haplogroups for the female samples is shown in Table 5.9. Perfect correlations between the declared BGA and region of geographic origin of the derived mtDNA haplogroups were observed for females of Sub-Saharan African, North African, European and East Asian origin. Only one Middle Eastern and one South Asian female belonged to mtDNA haplogroups that are not typically prevalent in their declared population of origin. Overall, the mtDNA results for the females were consistent with observations for the male samples (Table 5.7).

Table 5.9: Derived mtDNA haplogroups and prediction correlation for a sample of 30 females typed using Multiplex 1 to 4

The prediction correlation refers to the frequency (%) of females for whom the declared BGA correlates with the geographical region of origin of their derived mtDNA haplogroup. Females highlighted in red are individuals for which the derived mtDNA haplogroups are not typically prevalent in the declared population of origin.

Derived mtDNA haplogroup	**Declared population of origin***					
	SSAF N=2	**NAF** N=5	**MEA** N=6	**EUR** N=5	**SAS** N=7	**EAS** N=5
H*				1		
H1				2		
J1			**1**			
L	1					
M*					4	1
M1	1					
M7						1
M9					**1**	1
M80'D						1
N*						
N1			1			
R*		2	2		1	
R9						1
T*		1	1	2		
T1		1				
U*		1			1	
U7			1			
Prediction correlation (%)	100	100	83.3	100	85.7	100

*SSAF: Sub-Saharan Africa, NAF: North Africa, MEA: Middle East, EUR: Europe, SAS: South Asia, EAS: East Asia

5.2.9 Inference of bi-parental BGA using Multiplex 5

Two statistical methods based on the calculation of MLE were used to infer the BGA of the 146 individuals genotyped using Multiplex 5 and are discussed below.

5.2.9.1 *Inference of bi-parental BGA using the Snipper App Suite 2.0*

The Snipper App Suite 2.0 hosted on a web portal (available at http://mathgene.usc.es/snipper/) was used to calculate MLE for the 146 individuals typed using Multiplex 5 as described in section 2.2.3.10. The genotypes of 912 individuals originating from Sub-Saharan Africa, Europe and East Asia typed as part of Phase I of the 1000 Genomes project were used as reference populations. The differentiation achieved with the candidate autosomal SNPs between the Sub-Saharan Africa, Europe and East Asia populations for the 1000 Genome samples was initially assessed using STRUCTURE (section 2.2.3.8). As evident in Figure 5.18, clustering results for the analysis of the 1000 Genome (Phase I) samples without any prior specification of the ancestral groups showed that the individuals were separated into three clusters according to their population of origin at K=3. This confirmed that the 13 candidate SNPs were capable of differentiating these three major continental populations and that the 1000 Genome samples provided a suitable reference dataset.

The frequency of individuals assigned to the Sub-Saharan Africa, Europe and East Asia reference populations was deduced and is shown in Table 5.10. High prediction accuracies were observed for the inference of Sub-Saharan African (89%), European (98%) and East Asian (95%) ancestry when using the Multiplex 5 13-Plex assay. The majority (55%) of North African individuals were predicted to be of Sub-Saharan African origin, whilst the majority of Middle Eastern (75%) and South Asian (86%) individuals were inferred to be of European origin. These results parallel the PCoA, NJ and STRUCTURE analysis results of the autosomal genotype data, which demonstrated a close genetic relationship between Sub-Saharan and North Africans as well as Europeans, Middle Easterners and South Asians (refer to section 5.2.6).

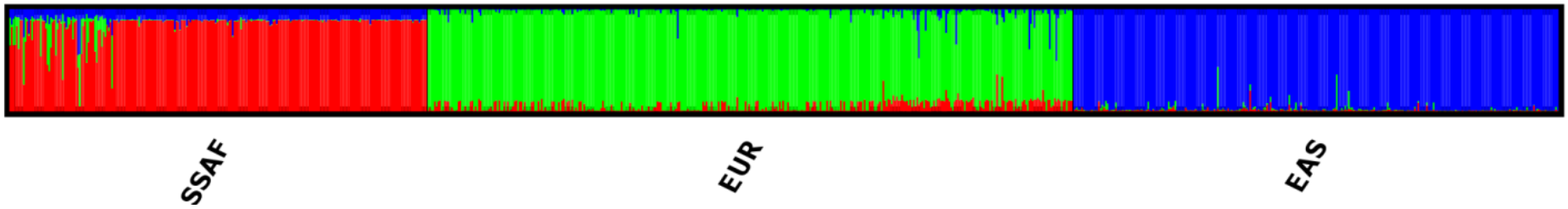


Figure 5.18: Individual genetic ancestry at *K*=3 in the 1000 Genomes (Phase I) dataset estimated by STRUCTURE v2.3.4 from 13 autosomal SNPs.

The bar plot was produced using DISTRUCT v1.1. Each of the 912 individuals in the 1000 Genomes (Phase I) dataset is depicted as a single vertical line partitioned into coloured segments that represent an individual's proportion of membership in each of the ancestral population clusters. The red, green and dark blue colours distinguish the clusters. Black lines are used to separate individuals according to their population of origin which are identified in the plots as **SSAF**= Sub-Saharan Africa, **EUR**= Europe and **EAS**= East Asia.

Table 5.10: Results for the inference of bi-parental BGA using SNIPPER

MLE were calculated using SNIPPER for 146 individuals typed using Multiplex 5 and were used to infer the population of origin of each sample against a reference dataset comprising 1000 Genomes (Phase I) population samples (N=912) originating from Sub-Saharan Africa, Europe and East Asia. The percentage of individuals assigned to each reference population is provided below and prediction accuracies (%) are shown in red.

Declared population of origin (N)	**Frequency (%) of individuals assigned to each reference population**		
	Sub Saharan Africa	**Europe**	**East Asia**
Sub Saharan Africa (9)	**88.89**	11.11	0.00
North Africa (20)	55.00	45.00	0.00
Middle East (16)	6.25	75.00	18.75
Europe (41)	2.44	**97.56**	0.00
South Asia (21)	14.29	85.71	0.00
East Asia (39)	2.56	2.56	**94.87**

5.2.9.2 *Inference of bi-parental BGA using GenAlEx v6.5*

The ability of the Multiplex 5 13-Plex assay to correctly infer the population of origin of 146 individuals was also assessed using GenAlEx v6.5 (Peakall and Smouse, 2012) as described in section 2.2.3.11. The population assignment test provided by this program calculates log-likelihood values for each sample in each ancestral population using the allele frequencies of the respective populations so that individuals are assigned to the population with the highest log-likelihood value.

Initially, the 146 individuals were assigned to one of six population groups including the Sub-Saharan Africa, North Africa, Middle East, Europe, South Asia and East Asia populations. The assay demonstrated a high degree of accuracy for the assignment of individuals originating from Sub-Saharan Africa (89.9%), North Africa (90%) and East Asia (94.9%) (Table 5.11A). However, intermediate prediction accuracies were observed for Europeans (75.6%) and South Asians (66.7%), with the Middle Eastern population showing the lowest percentage of correct assignment (50%). The majority of Middle Eastern individuals were incorrectly assigned to the European population whilst most of the incorrectly assigned European and South Asian individuals were classified as originating from the Middle East (Table 5.11A). These results confirm that the European, Middle Eastern and South Asian population cannot be sufficiently differentiated by the 13 candidate SNPs as previously observed through PCoA, NJ and STRUCTURE analyses (refer to section 5.2.6)

Analyses of population sub-structure on the autosomal dataset revealed the presence of four major ancestral populations comprising Sub-Saharan Africa, South and East Asia and a cluster combining European and Middle Eastern individuals (Figure 5.17). For this reason, the 146 individuals were then assigned to one of five population groups including Sub-Saharan Africa, North Africa, South Asia, East Asia and Western Eurasia in which were grouped individuals of European and Middle Eastern origin. Reducing the number of populations to five improved the prediction accuracy for South Asians (90.5%) and Western Eurasians (86%) (Table 5.11B). A large proportion (12.3%) of incorrectly assigned Western Eurasians was inferred to be from South Asia. This further confirmed that the close genetic relationship between the Europe, Middle East and

Table 5.11: Results for the inference of bi-parental BGA using GenAlEx v6.5.

The MLE calculated using GenAlEx v6.5 for 146 individuals typed with Multiplex 5 were used to assign each individual to **(A)** six population groups: **SSAF**= Sub-Saharan Africa, **NAF**= North Africa, **EUR**= Europe, **MEA**= Middle East, **SAS**= South Asia, **EAS**= East Asia, **(B)** five population groups where the European and Middle Eastern populations were combined into Western Eurasia (WEURA) and **(C)** four population groups where Europe, Middle East and South Asia were analysed together as Eurasia (EURA). The percentage of individuals assigned to each population is provided in the tables above, with prediction accuracies (%) shown in red.

(A) 6 Populations

Declared population of origin (N)	**Frequency (%) of genotyped individuals assigned to each population**					
	SSAF	**NAF**	**MEA**	**EUR**	**SAS**	**EAS**
Sub Saharan Africa (9)	**89.89**	11.11	0.00	0.00	0.00	0.00
North Africa (20)	5.00	**90.00**	0.00	0.00	5.00	0.00
Middle East (16)	0.00	0.00	**50.00**	25.00	18.75	0.00
Europe (41)	0.00	0.00	21.95	**75.61**	2.44	0.00
South Asia (21)	0.00	4.76	19.05	4.76	**66.67**	4.76
East Asia (39)	0.00	0.00	5.13	0.00	0.00	**94.87**

(B) 5 Populations

Declared population of origin (N)	**Frequency (%) of genotyped individuals assigned to each population**				
	SSAF	**NAF**	**WEURA**	**SAS**	**EAS**
Sub Saharan Africa (9)	**89.89**	11.11	0.00	0.00	0.00
North Africa (20)	10.00	**85.00**	0.00	5.00	0.00
Western Eurasia (57)	0.00	0.00	**85.96**	12.28	1.75
South Asia (21)	0.00	0.00	4.76	**90.48**	9.52
East Asia (39)	0.00	0.00	5.13	0.00	**94.87**

(C) 4 Populations

Declared population of origin (N)	**Frequency (%) of genotyped individuals assigned to each population**			
	SSAF	**NAF**	**EURA**	**EAS**
Sub Saharan Africa (9)	**89.89**	11.11	0.00	0.00
North Africa (20)	5.00	**95.00**	0.00	0.00
Eurasia (78)	1.28	2.56	**93.59**	2.56
East Asia (39)	0.00	0.00	5.13	**94.87**

South Asian populations did not allow for adequate differentiation of these three groups based solely on the Multiplex 5 candidate SNPs.

The ability for the Multiplex 5 13plex assay to correctly infer bi-parental ancestry was further investigated by combining individuals of European, Middle Eastern and South Asian origin into the Eurasia group. When the number of populations was reduced to four, high prediction accuracies were observed for the assignment of Sub-Saharan African (89.9%), North African (95%), Eurasian (94%) and East Asian (95%) individuals (Table 5.11C). These results suggest that the Multiplex 5 13-Plex assay is capable of inferring bi-parental BGA with greater than 90% accuracy for major continental groups only. As depicted in the pairwise population assignment plots in Figure 5.19, a high degree of differentiation evidenced through minimal overlap between data points for the Sub-Saharan African, North African and East Asian populations can be achieved using the 13 autosomal candidate SNPs. However, Europeans, Middle Easterners and South Asians cannot be adequately differentiated from each other but are sufficiently separated from the other major continental populations when combined in the Eurasian group (Figure 5.19).

5.2.10 Predictive power of the 5-multiplex SNP genotyping system for the inference of regional biogeographical ancestry

To investigate the predictive accuracy of the 5-multiplex SNP genotyping system for inferences of BGA, the degree of correlation between the correctly predicted maternal, paternal and bi-parental regional BGA for 116 males was determined. As shown in Table 5.12A, the inferred ancestry of 84% of individuals originating from Sub-Saharan Africa, Europe and East Asia correlated across all three SNP classes when comparing the inferred maternal and paternal BGA with the bi-parental BGA deduced from Snipper using a reference dataset of 1000 Genomes samples (refer to Table 5.10). The BGA of all 77 males could be accurately predicted when at least two inferred regions of geographic origin were correct and matching. The East Asian population showed the greatest accuracy of prediction, where all three predicted regional ancestries were correct and matched each other (97.1%) whilst the European population demonstrated the lowest predictive accuracy (72.2%) when comparing all three inference

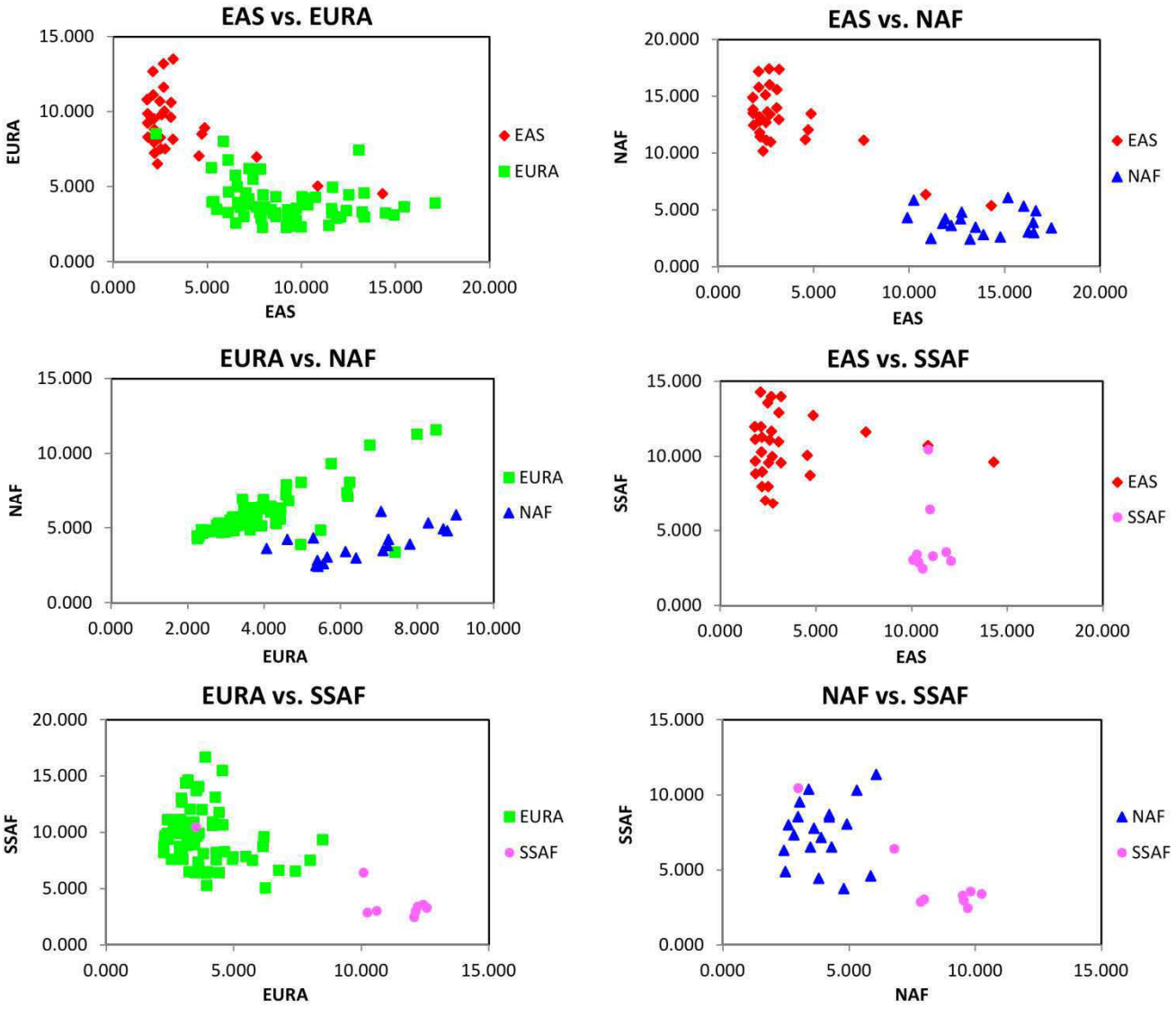


Figure 5.19: GenAlEx Pairwise population assignment plots for four major continental population groups.

Genotype data of 146 individuals typed using Multiplex 5 were used to assign each individual to one of four continental populations, representing major global geographic regions: **EAS**= East Asia, **EURA**= Eurasia (Europe, Middle East and South Asia), **NAF**= North Africa and **SSAF**= Sub-Saharan Africa. The plots produced for pairwise population assignment comparisons demonstrate a high degree of differentiation with minimal overlap between data points of population pairs.

Table 5.12: Summary of the predictive power of the complete multiplex SNP assay (Multiplex 1 to 5) for the inference of regional BGA.

The overall predictive accuracy of the 5-multiplex SNP assay was estimated for 116 males as the percentage of individuals for whom the predicted maternal, paternal and bi-parental regional BGA were complementary across two or all three SNP classes and correlated with the declared BGA. In Table **(A)**, the predictive accuracy was deduced for 77 out of 116 males originating from Sub-Saharan Africa, Europe or East Asia based on the bi-parental BGA derived from Snipper (Table 5.10) with 1000 Genomes samples as reference populations. Table **(B)** shows the predictive accuracy for all 116 males based on the inferred bi-parental regional BGA derived from GenAlEx (5 Populations- Table 5.11B).

(A)

Declared BGA (N)	**Predictive accuracy (%)**				
	All inferred BGA match	**Only 2 inferred BGA match**			**At least 2 inferred BGA match**
		mtDNA/NRY	MtDNA/Autosomal	NRY/Autosomal	
Sub Saharan Africa (7)	85.7	0.0	14.3	0.0	100.0
Europe (36)	72.2	2.8	5.6	19.4	100.0
East Asia (34)	97.1	2.9	0.0	0.0	100.0
TOTAL (77)	84.4	2.6	3.9	7.8	100.0

(B)

Declared BGA (N)	**Predictive accuracy (%)**				
	All inferred BGA match	**Only 2 inferred BGA match**			**At least 2 inferred BGA match**
		mtDNA/NRY	MtDNA/Autosomal	NRY/Autosomal	
Sub Saharan Africa (7)	85.7	0.0	14.3	0.0	100
North Africa (15)	53.3	6.7	13.3	26.7	100
Middle East (10)	60.0	30.0	0.0	10.0	100
Europe (36)	75.0	2.8	0.0	19.4	97.2
South Asia (14)	64.3	21.4	14.3	0.0	100
East Asia (34)	94.1	5.9	0.0	0.0	100
TOTAL (116)	75.9	10.3	2.6	10.3	99.1

results for the mtDNA, NRY and autosomal SNPs. However, 19.4% of Europeans showed matching inferred paternal and bi-parental regional BGA, suggesting that the NRY and autosomal SNPs are the most effective markers in differentiating this population.

When comparing the inferred maternal and paternal geographic region of origin against the bi-parental BGA deduced from GenAlEx for the assignment of individuals to one of five regional populations (Sub-Saharan Africa, North Africa, Europe, East Asia and Central Eurasia) (refer to Table 5.11B), a lower overall predictive accuracy of 75.9% was observed, where all inferred ancestries were correct and correlated across all three SNP classes (Table 5.12B). However, 0.9% of individuals demonstrated no correlations between all three inferred regions of origin but in these cases, one of the SNPs was still capable of correctly inferring the population or geographic region of origin. This indicates that the regional BGA of at least 99% of tested males can be accurately predicted when at least two of the correctly inferred BGA matched.

East Asians (94.1%) and Sub-Saharan Africans (85.7%) showed the highest predictive accuracy when all three SNP classes correctly inferred the regional BGA and were concordant, whilst the remaining populations demonstrated intermediate predictive accuracies ranging from 53 to 75%. The regional BGA of most remaining Middle Eastern individuals (30%) was accurately predicted from complementary maternal and paternal ancestries suggesting that the mtDNA and NRY SNPs are the better markers to distinguish the Middle Eastern population. On the other hand, 21.4% of South Asians showed correlations between the inferred maternal and bi-parental regional BGA or the maternal and paternal regional BGA, suggesting that this population is most effectively differentiated using the mtDNA and autosomal or the NRY and autosomal SNPs. Contrastingly, a predictive accuracy of 26.7% was observed for North Africans when the inferred paternal and bi-parental regional BGA were complementary, indicating that the North Africa population is best distinguished using the NRY and autosomal markers. Combining the inference results obtained using mtDNA, NRY and/or autosomal SNPs improves the strength of the BGA predictions compared to using only one SNP class (Table 5.7, Table 5.10 and Table 5.11). Furthermore, different populations were more clearly distinguished using specific pairs of SNPs such as South

Asian individuals who were more efficiently differentiated using mtDNA and autosomal markers (Table 5.12B).

Though the 5-multiplex SNP genotyping system was primarily developed to genotype male individuals, the amount of informative ancestry data that can be inferred for females was also investigated. Only the maternal and bi-parental BGA of females can be predicted from the mtDNA and autosomal SNPs. Thus, the predictive power for Multiplex 1 to 5 was also assessed for 146 individuals (116 males and 30 females) by determining the extent of correlation between the inferred maternal and bi-parental regional BGA. A much lower overall predictive accuracy of 81% was observed when only the maternal and bi-parental BGA are available, with 19% of individuals showing no match between both inferred ancestries (Table 5.13). At the population level, the Sub-Saharan African and East Asian populations demonstrated high predictive accuracies (97-100%) when both maternal and bi-parental regional BGA were concordant (Table 5.13). However, the North African, European and South Asian populations showed intermediate predictive accuracies (70-78%) whilst the Middle Eastern population demonstrated the lowest predictive accuracy (50%) (Table 5.13). The majority of the 50% of Middle Eastern individuals who did not show correlations between the maternal and bi-parental inferences were however, correctly assigned to the appropriate region of geographic origin using mtDNA data but incorrectly assigned to another closely related population such as South Asia or North Africa based on the autosomal data.

Overall, the mtDNA SNPs were found to predict regional BGA less accurately than the NRY SNPs, particularly for the North African population but with the exception of the Middle East and South Asian populations (Table 5.7). Contrastingly, autosomal SNPs demonstrated relatively high accuracies for the prediction of North African and Western Eurasian (Europe and Middle East) ancestry when assigning individuals to five population groups (Table 5.11B) but favoured the grouping of these two populations into a single Eurasian group (Table 5.11C). These results indicate that the 5-multiplex SNP assay can be used to predict the BGA of females originating from Sub-Saharan Africa and East Asia with a high degree of accuracy but cannot adequately distinguish European, Middle Eastern, North African and South Asian females. In these cases,

looking at the markers that are most informative for a given population, such as mtDNA SNPs for the Middle East/South Asia and autosomal SNPs for North Africa/Western Eurasia, is more accurate when both maternal and bi-parental regional BGA inferences are conflicting. The 5-multiplex SNP genotyping system is most effective for male individuals because the NRY SNPs can assist in distinguishing specific population groups that are more difficult to differentiate solely based on mtDNA and autosomal data, particularly closely genetically related populations such as Europe, North Africa, the Middle East and South Asia (Table 5.7).

Table 5.13: Assessing the predictive power of the 5-multiplex SNP genotyping system for the inference of BGA of female individuals.

The predictive accuracy of the 5-multiplex SNP assay was estimated as the percentage correlation between declared BGA and the inferred maternal and bi-parental BGA derived from GenAlEx (5 Populations- Table 5.11B) because the mtDNA and autosomal SNPs are the only useable markers when genotyping female samples. Due to the small number of females genotyped (N=30), the inferred maternal and bi-parental BGA of 116 males was also included in analysis.

Declared BGA (N)	Predictive accuracy (%)-matching maternal and bi-parental BGA
Sub Saharan African (9)	100.0
North Africa (20)	70.0
Middle East (16)	50.0
Europe (41)	78.0
South Asia (21)	76.2
East Asia (39)	97.4
TOTAL (146)	80.8

5.2.11 Inference of pigmentation phenotype and EVCs using nine SNPs in pigmentation genes

As discussed in chapter 4, whilst most autosomal SNPs were selected for their ability to distinguish the six populations of interest for the inference of BGA, many of the candidate SNPs are located in pigmentation genes and also demonstrate associations with EVCs, as discussed in section 2.2.3.12. The utility of nine pigmentation SNPs

incorporated in Multiplex 5 (rs885479, rs16891982, rs1426654, rs1545397, rs6058017, rs7498174, rs4778241, rs4778138, rs12913832) for the prediction of EVCs was investigated. Initially, the associations of the nine individual SNPs with hair, eye and skin colour variation in a sample of 40 Europeans were determined using Fisher's exact test (2-tailed) in SPSS® v19 (IBM Corp., USA) as described in section 2.2.3.12. Only Europeans were selected for analysis as this population shows the most diverse range of phenotypes and to exclude any effect of ancestry on the association tests. Since only one European individual had declared their hair colour to be red, this sample was excluded from the association tests. Odds ratios (OR) with 95% confidence intervals (CI) and associated *p*-values were calculated using a dominant model of allele classification where the genotypes were classified according to the presence or absence of the minor allele (MA), which was assumed to be dominant (1, 0 respectively) (Table 5.14).

Only two SNPs, rs12913832 (OR= 7.000, CI=1.300-37.704, *p*= 0.020) and rs4778138 (OR=10.500, CI=2.308-47.777, *p*= 0.002) were found to be associated with hair colour in separate tests. Four SNPs (rs7498174, rs4778241, rs4778138, rs12913832) were associated with eye colour variation, with the rs12913832*A allele demonstrating the strongest association with brown eye colour in Europeans (OR=156.000, CI=12.888-188.233, *p*= <0.000). The rs12913832 SNP was also the only marker that showed associations with skin colour (OR=15.167, CI=1.723-133.526, *p*=0.005). Additionally, the TGT haplotype combining SNPs rs7495174, rs4778241 and rs4778138 which has previously been linked with blue eye colour (Duffy *et al.*, 2007) was investigated. The TGT haplotype was also found to be strongly associated with blue eye colour in our European sample (OR=11.000, CI=1.998-60.572, *p*=0.006).

Due to the small sample size of individuals genotyped using Multiplex 5 in the preliminary validation phase of this study, the predictive utility of the pigmentation SNPs for inference of EVCs could not be assessed using logistic regression models and thus MLE were derived instead. The EVCs of 146 individuals originating from six populations were predicted using GenAlEx v6.5 (Peakall and Smouse, 2012) as described in section 2.2.3.13. The genotypes for the SNPs which showed significant associations with hair, eye and skin colour (refer to Table 5.14 for results) were used to assign each individual to the dark (black/dark brown) or light (light brown/blonde)

Table 5.14: Association of nine pigmentation SNPs with hair, eye and skin colour variation in Europeans (N=40).

Odds ratios (OR) with 95% confidence intervals (CI) and associated *p*-values were calculated using Fisher's exact test (2-sided) and a dominant model of allele classification where the genotypes were classified according to the presence or absence of the minor allele (MA), which was assumed to be dominant. Significant values are shown in bold. N/A indicates that the OR could not be computed due to a denominator value of zero.

SNP-MA	Hair colour[♯]		Eye colour				Skin colour	
	Dark v Light		Blue v Non-blue		Brown v Non-brown		Dark v Fair	
	OR (95% CI)	P	OR (95% CI)	P	OR (95% CI)	P	OR (95% CI)	P
rs885479-T	N/A	1.000	N/A	0.538	N/A	0.539	0.923 (0.076-11.173)	1.000
rs1426654-G	N/A	1.000	N/A	1.000	N/A	1.000	N/A	1.000
rs1545397-T	N/A	1.000	N/A	0.538	N/A	0.539	4.167 (0.343-50.618)	0.276
rs6058017-G	0.783 (0.115-5.341)	1.000	N/A	0.154	N/A	0.143	3.273 (0.477-22.463)	0.322
rs7498174-C	6.882 (0.771-61.406)	0.070	N/A	**0.016**	N/A	**0.007**	1.333 (0.305-5.830)	0.718
rs4778241-T	2.500 (0.622-10.049)	0.315	0.145 (0.027-0.786)	**0.020**	8.182 (1.514-44.211)	**0.017**	1.023 (0.275-3.804)	1.000
rs4778138-C	7.000 (1.300-37.704)	**0.020**	0.067 (0.008-0.588)	**0.005**	17.727 (2.009-156.457)	**0.002**	3.000 (0.780-11.536)	0.176
rs12913832-A	10.500 (2.308-47.777)	**0.002**	0.010 (0.001-0.111)	**<0.000**	156.000 (12.888-188.233)	**<0.000**	15.167 (1.723-133.526)	**0.005**
rs16891982-G	4.789 (0.525-43.697)	0.222	0.238 (0.026-2.180)	0.236	4.789 (0.525-43.697)	0.222	2.200 (0.174-2.381)	0.416
OCA2-TGT*	0.580 (0.109-3.082)	0.662	11.000 (1.998-60.572)	**0.006**	0.074 (0.013-0.411)	**0.002**	0.643 (0.174-2.381)	0.741

* OCA2 haplotype grouping rs7495174, rs4778241 and rs4778138 respectively.

♯ Hair colour classification- Dark= black or dark brown, Light= light brown or blonde

hair group, blue/green or light/dark brown eye group and fair, olive or dark skin group.

As shown in Table 5.15A, high predictive accuracies were observed for the inference of dark (black/dark brown) (86%) and light (light brown/red/blonde) (88%) hair when using the rs4778138 and rs12913832 SNPs. Eye colour was also predicted with a high degree of accuracy using the rs7498174, rs4778241, rs4778138 and rs12913832 with 90% of individuals with blue/green eyes and 83% of individuals with light/dark brown eyes correctly distinguished (Table 5.15B). Skin colour prediction was less effective (Table 5.15C), with only one SNP, rs12913832 showing significant associations with skin colour (Table 5.14). No individuals with olive skin were correctly assigned using only the rs12913832 genotype, with 63% assigned as having dark skin; however a larger proportion of individuals with dark (89%) compared to light (52%) skin were correctly assigned (Top Table in Table 5.15C). When individuals were grouped into two skin colour groups instead- those with fair skin and those with olive or dark skin, low predictive accuracies were observed for fair skinned (52%) and olive/dark skinned (69%) individuals (Bottom Table in Table 5.15C). All fair skinned individuals incorrectly assigned as having dark skin were those with dark pigmentation, such as dark hair and/ or eyes. Out of the 31% of individuals in the olive/dark skinned group who were incorrectly inferred to have fair skin, 38% were individuals with olive skin compared to only 11% of individuals with dark skin.

Despite no associations with hair, eye and skin colour detected for the remaining five pigmentation SNPs most likely due to the small sample size, these markers were previously linked with pigmentation variation as discussed in section 4.2.1.2. For this reason, the effect of combining all nine pigmentation SNPs included in Multiplex 5 (rs885479, rs1426654, rs1545397, rs6058017, rs7498174, rs4778241, rs4778138, rs12913832 and rs16891982) on the inference of EVCs was also investigated. The genotypes of 146 individuals were used to calculate MLE using GenAlEx to assign each individual to the dark (black/dark brown) or light (light brown/red/blonde) hair group, blue/green or light/dark brown eye group and fair, olive or dark skin group.

Table 5.15: Inference of EVCs using only pigmentation SNPs demonstrating positive associations to hair, eye and skin colour variation.

The rs7498174, rs4778241, rs4778138 and rs12913832 SNPs which have demonstrated associations with pigmentation variation in Europeans (refer to Table 5.14) were used to predict **(A)** hair, **(B)** eye and **(C)** skin colour. Maximum likelihood estimates were calculated from the genotypes of 146 individuals from various populations using GenAlEx v6.5. Each individual was assigned to the dark (black/dark brown) or light (light brown/red/blonde) hair group, blue/green or light/dark brown eye group and fair, olive or dark skin group. The percentage of individuals assigned to each phenotypic group is provided in the tables above, with prediction accuracies (%) shown in red.

(A) Hair colour- rs12913832 and rs4778138

Declared hair colour (N)	Frequency (%) of genotyped individuals assigned:	
	Dark	Light
Dark (129)	**86.0**	14.0
Light (17)	11.8	**88.2**

(B) Eye colour- rs7498174, rs4778241, rs4778138, rs12913832

Declared eye colour (N)	Frequency (%) of genotyped individuals assigned:	
	Blue/Green	Light/Dark Brown
Blue/Green (19)	**89.5**	10.5
Light/Dark Brown (127)	17.3	**82.7**

(C) Skin colour- rs12913832

Declared skin colour (N)	Frequency (%) of genotyped individuals assigned:		
	Fair	Olive	Dark
Fair (46)	**52.5**	0.0	47.8
Olive (73)	37.0	**0.0**	63.0
Dark (27)	11.1	0.0	**88.9**

Declared skin colour (N)	Frequency (%) of genotyped individuals assigned:	
	Fair	Olive/Dark
Fair (46)	**52.2**	47.8
Olive/Dark (100)	31.0	**69.0**

As evident in Table 5.16A, a much higher predictive accuracy (94%) was observed for the inference of dark hair (black/dark brown) when using all nine pigmentation SNPs, however the percentage of light hair (light brown/red/blonde) individuals correctly assigned decreased to 78%. Similarly, the eye colour of 94% of individuals with blue/green eyes was correctly inferred based on the nine pigmentation SNP genotypes despite the predictive accuracy for the inference of light/dark brown eye colour decreasing slightly to 81% (Table 5.16B). Thirteen out of nineteen individuals with light/dark brown eye colour who were incorrectly assigned to the blue/green group had declared their eye colour to be hazel/light brown. Some shades of hazel are lighter and appear more similar to green, which may explain why the eye colour of these individuals was predicted to be blue/green. Overall it appears that hair colour prediction is most accurate when using only SNPs rs4778138 and rs12913832 but eye colour prediction is more successful when using the genotypes for all nine pigmentation SNPs

Skin colour prediction was improved when using all nine pigmentation SNPs compared to only the rs12913832 genotype, with a high predictive accuracy of 81.5% observed for the inference of dark skin (Table 5.16B). An intermediate predictive accuracy of 61% was observed for fair skinned individuals but a very low percentage of olive skinned individuals were correctly assigned (26%) (Table 5.16C). The majority of individuals with olive skin were assigned to either the fair (36%) or dark (38%) skin groups. Skin colour assessment is subjective to individual perception and individuals who declared their skin to be ‘fair’ might in fact be ‘olive’, whilst a declared ‘olive’ person may really be ‘fair’ or ‘dark’. Based on the subjective criteria used to collect EVC data of participants in this study, the true predictive accuracy of the assay for the inference of skin colour could not be precisely determined.

Table 5.16: Inference of EVCs using nine pigmentation SNPs.

The genotypes of 146 individuals from various populations for SNPs rs885479, rs1426654, rs1545397, rs6058017, rs7498174, rs4778241, rs4778138, rs12913832 and rs16891982 were used to calculate maximum likelihood estimates using GenAlEx v6.5. **(A)** Hair, **(B)** eye and **(C)** skin colour were inferred by assigning each individual to the dark (black/dark brown) or light (light brown/red/blonde) hair group, blue/green or light/dark brown eye group and fair, olive or dark skin group. The percentage of individuals assigned to each phenotypic group is provided in the tables above, with prediction accuracies (%) shown in red.

(A) Hair colour

Declared hair colour (N)	Frequency (%) of genotyped individuals assigned:	
	Dark	Light
Dark (129)	**77.5**	22.5
Light (17)	5.9	**94.1**

(B) Eye colour

Declared eye colour (N)	Frequency (%) of genotyped individuals assigned:	
	Blue/Green	Light/Dark Brown
Blue/Green (19)	**94.7**	5.3
Light/Dark Brown (127)	18.9	**81.1**

(C) Skin colour

Declared skin colour (N)	Frequency (%) of genotyped individuals assigned:		
	Fair	Olive	Dark
Fair (46)	**60.9**	8.7	30.4
Olive (73)	35.6	**26.0**	38.4
Dark (27)	3.7	14.8	**81.5**

5.2.12 Genotyping mock forensic casework samples

To investigate the applicability of the developed forensic intelligence 5-multiplex SNP assay on the analysis of casework samples, a variety of DNA samples were collected from four donors (section 2.2.5.1). Genomic DNA was extracted from each casework sample, quantitated and amplified as described in sections 2.2.5.2 to 2.2.5.4. Only multiplexes within the hierarchical set (Multiplex 1 to 4) that were required to derive the final mtDNA and NRY haplogroup based on the result of Multiplex 1 were performed

(refer to Figure 3.3 to Figure 3.6). For each sample, positive controls consisting of buccal swabs taken from each donor were also amplified to confirm results. As shown in Figure 5.20 to Figure 5.23, full profiles were obtained for all samples, with the exception of the ASIP and P31 loci that were problematic in two samples as discussed below.

Donor 1 who provided a whole blood sample that was deposited on cloth and liquid semen sample (refer to section 2.2.5.1 for details) was genotyped using Multiplexes 1-3 and 5 (Figure 5.20). From the results of Multiplex 1-3, donor 1 was identified to belong to mtDNA haplogroup H* and NRY haplogroup R1a1. Full profiles were obtained for both blood and semen samples for all multiplexes, with the exception of the ASIP (Multiplex 5) locus, which could not be detected in the liquid semen sample. Only 0.46ng/µL of genomic DNA was extracted from the semen sample (Table 2.10) and despite 3ng of starting template DNA added to the Multiplex 5 PCR, these alleles could not be called even after repeating the reaction several times (data not shown). Whilst ASIP could be detected with starting genomic DNA concentrations of at least 0.3ng (Figure 5.3), the semen sample used was at least 12 months old and though stored at 4°C there is a possibility that slight degradation may have occurred. This is evident as lower signals were observed for all multiplexes compared to the blood sample, despite the same starting template concentration used for both samples (Figure 5.20).

A semen sample from Donor 2 was deposited on cloth and stored at room temperature for 3 weeks prior to extraction (section 2.2.5.1). The extracted genomic DNA was genotyped using Multiplexes 1-3 and 5 with full profiles obtained for all multiplexes so that donor 2 was identified to be of mtDNA haplogroup U* and NRY haplogroup G (Figure 5.21). Donor 3 provided a smoked cigarette butt from which genomic DNA was extracted and typed using Multiplexes 1, 2, 4 and 5 which all showed full profiles and identified donor 3 as belonging to mtDNA haplogroup M8 and NRY haplogroup C3 (Figure 5.22). On the other hand, a trace saliva sample taken from a swab of a soft drink can consumed by donor 4 was genotyped using Multiplex 1, 4 and 5 and showed full profiles for all multiplexes except for P31 in Multiplex 4 and ASIP in Multiplex 5, which could not be called even after repeating the multiplexes several times (Figure 5.23). From the result of Multiplex 1 and 4, donor 4 was deduced to belong to mtDNA

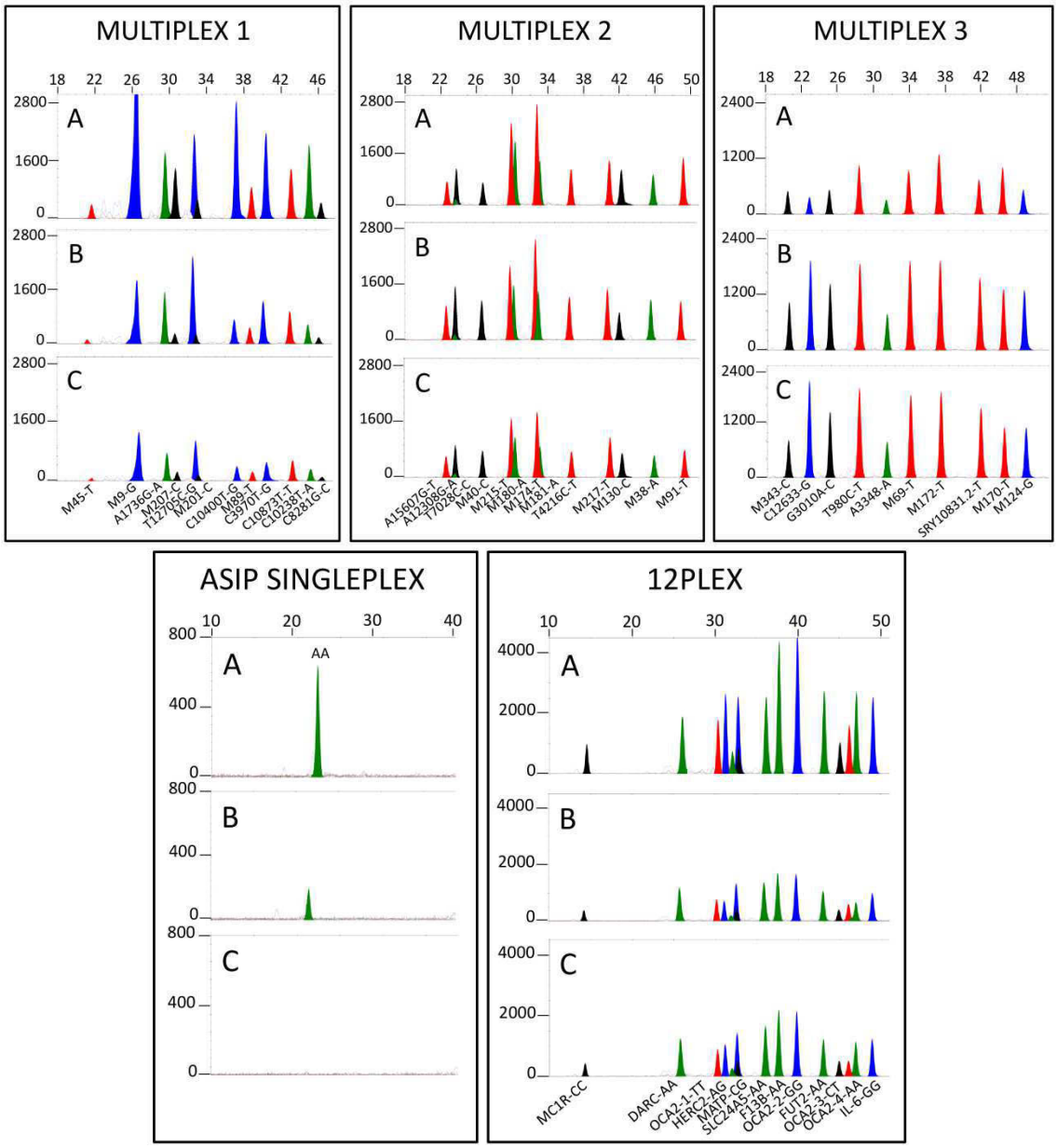


Figure 5.20: Preliminary validation of the 5-multiplex SNP genotyping system on forensic casework samples from Donor 1.

Donor 1 positive control profiles from buccal swabs stored on FTA® cards are shown in (A). DNA was isolated from (B) whole blood deposited on cloth and (C) liquid semen from Donor 1. Profiles were obtained for each sample using only Multiplex 1, 2, 3 and 5 as Multiplex 4 was not required to derive the mtDNA and NRY haplogroups based on the result of Multiplex 1. Donor 1 was identified to belong to mtDNA haplogroup H* and NRY haplogroup R1a1.

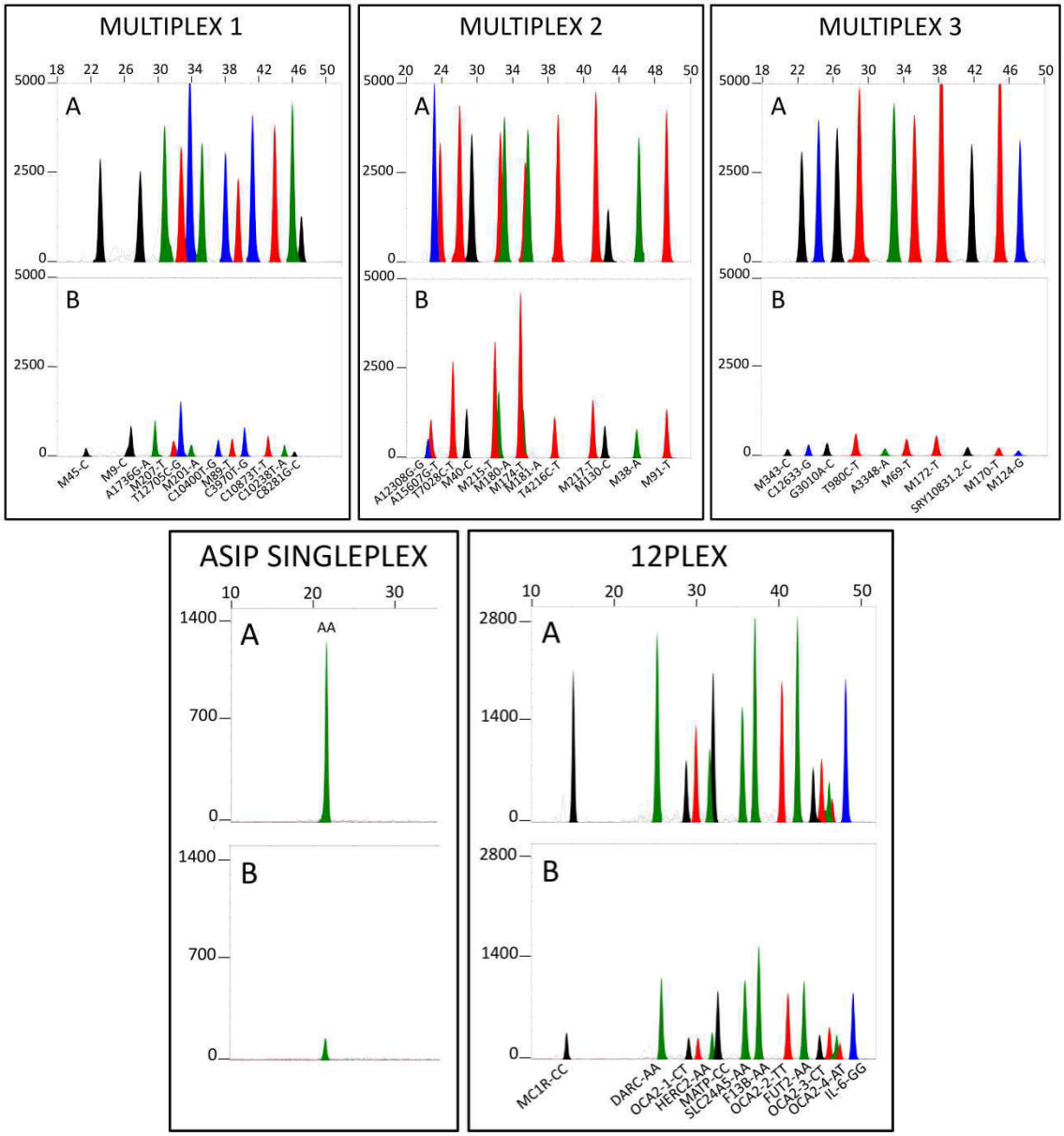


Figure 5.21: Preliminary validation of the 5-multiplex SNP genotyping system on forensic casework samples from Donor 2.

Donor 2 positive control profiles from buccal swabs stored on FTA® cards are shown in (A). DNA was isolated from (B) liquid semen deposited on cloth from Donor 2. Profiles were obtained for each sample using only Multiplex 1, 2, 3 and 5 as Multiplex 4 was not required to derive the mtDNA and NRY haplogroups based on the result of Multiplex 1. Donor 2 was identified to belong to mtDNA haplogroup U* and NRY haplogroup G.

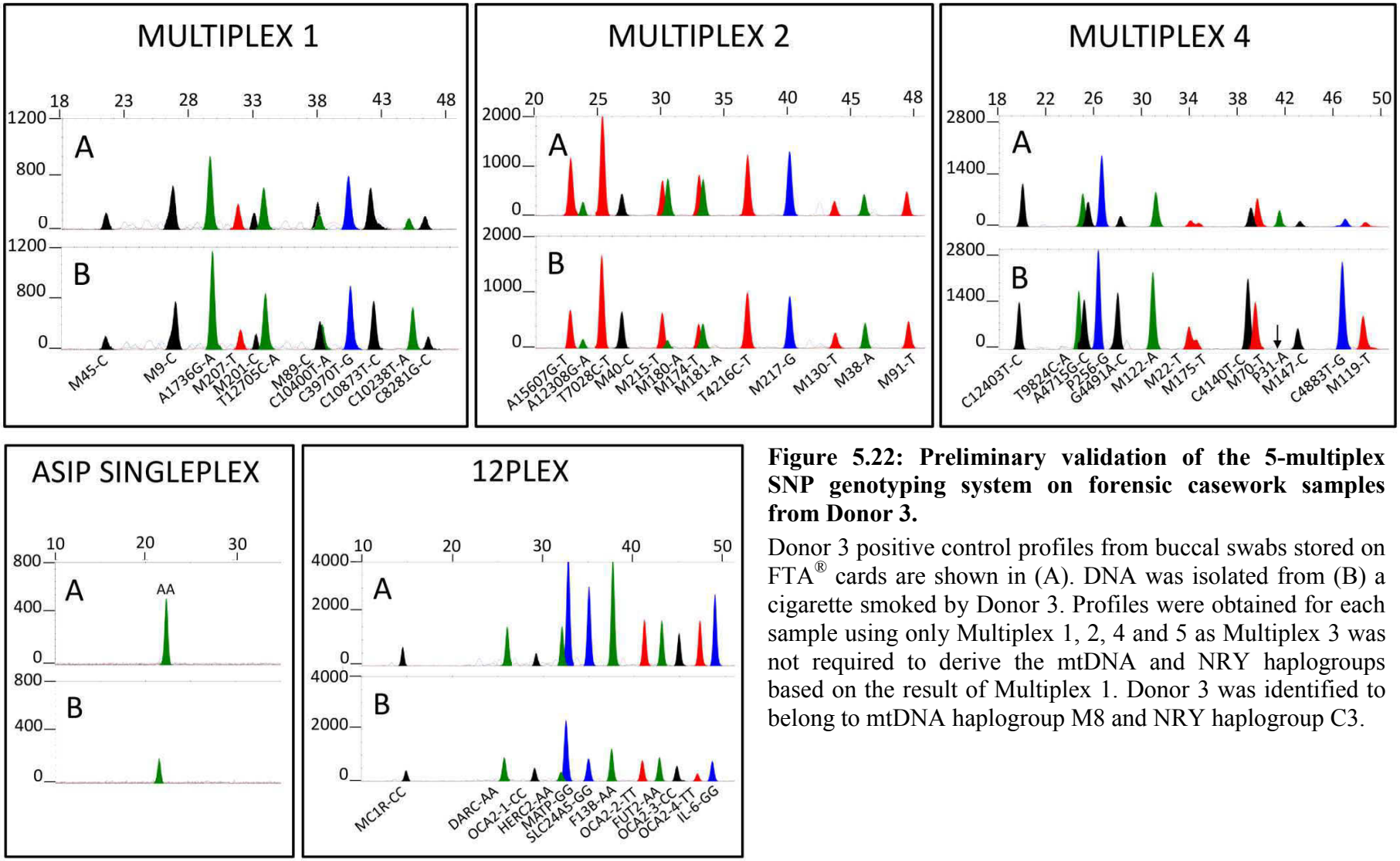


Figure 5.22: Preliminary validation of the 5-multiplex SNP genotyping system on forensic casework samples from Donor 3.

Donor 3 positive control profiles from buccal swabs stored on FTA® cards are shown in (A). DNA was isolated from (B) a cigarette smoked by Donor 3. Profiles were obtained for each sample using only Multiplex 1, 2, 4 and 5 as Multiplex 3 was not required to derive the mtDNA and NRY haplogroups based on the result of Multiplex 1. Donor 3 was identified to belong to mtDNA haplogroup M8 and NRY haplogroup C3.

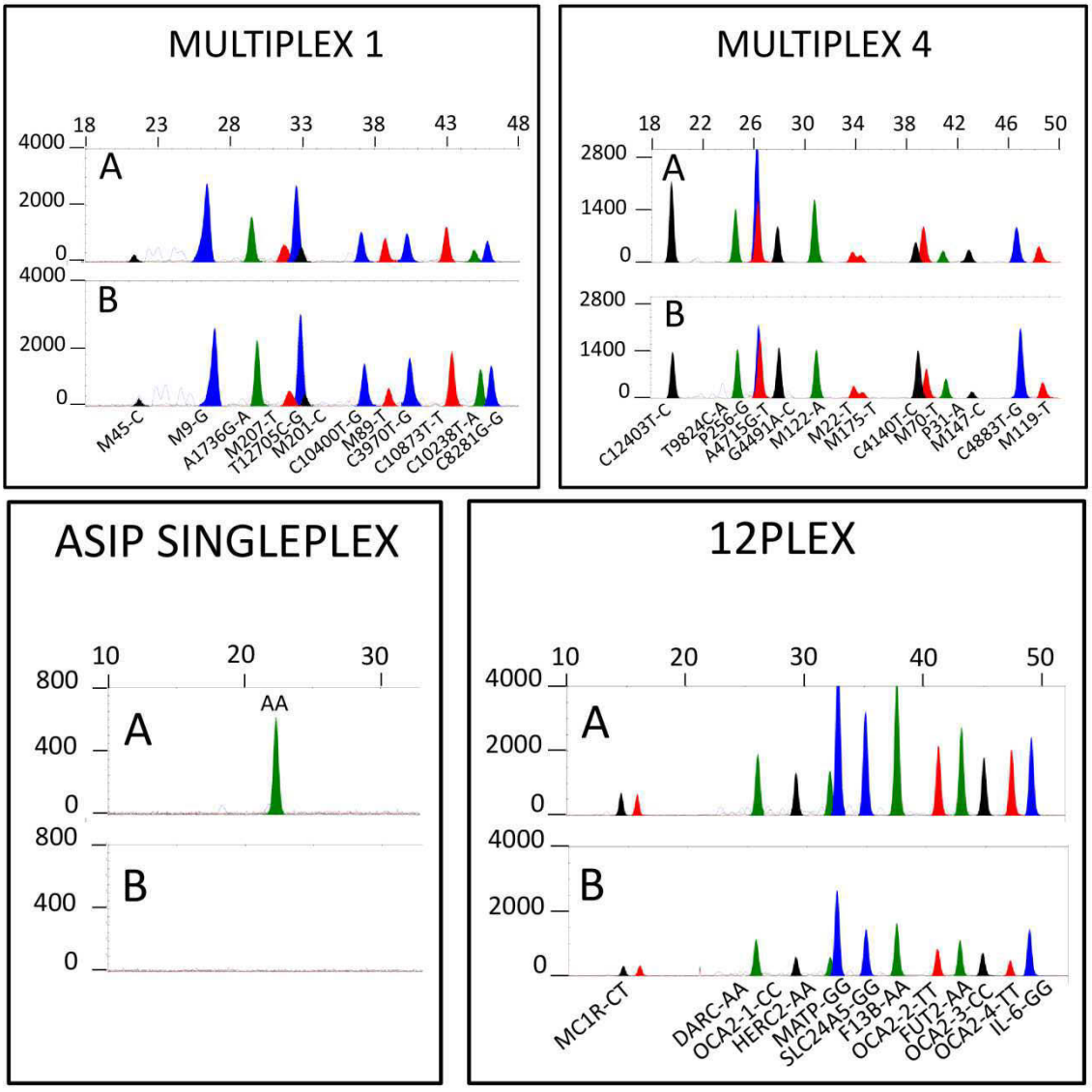


Figure 5.23: Preliminary validation of the 5-multiplex SNP genotyping system on forensic casework samples from Donor 4.

Donor 4 positive control profiles from buccal swabs stored on FTA® cards are shown in (A). DNA was isolated from (B) a swab of a soft drink can consumed by Donor 4. Profiles were obtained for each sample using only Multiplex 1, 4 and 5 as Multiplex 2 and 3 were not required to derive the mtDNA and NRY haplogroups based on the result of Multiplex 1. Donor 4 was identified to belong to mtDNA haplogroup B4'5 and NRY haplogroup K*.

haplogroup B4'5 and NRY haplogroup K* despite the inability to call P31. With only 0.15ng/μL of genomic DNA extracted from the trace saliva sample, such that only 1.05ng of starting template DNA could be added to the Multiplex 5 PCR (Table 2.10), the nature of the sample as a trace DNA swab suggests that degradation is likely to have occurred which may have prevented amplification of the already problematic ASIP and P31 loci.

The maternal, paternal and bi-parental BGA of each sample was predicted from the derived mtDNA and NRY haplogroups and autosomal genotypes. For all donors, inferred regional BGA correlated with the declared ancestry for at least two SNP classes (Table 5.17). Donor 2 was correctly identified as being of Western Eurasian origin and donors 3 and 4 as originating from East Asia from all three types of markers (Table 5.17). However, the declared European donor 1 demonstrated only correlations between the inferred maternal and bi-parental ancestry (Table 5.17). The hair, eye and skin colour of the donors was also inferred from the genotypes for nine pigmentation SNPs included in Multiplex 5, despite missing data for the ASIP locus for donors 1 (semen) and 4 (can swab). The hair and eye colour of all donors was correctly inferred, except for Donor 1 who was predicted to have light hair and blue/green eyes (Table 5.18). Skin colour prediction was less accurate, however the inferred skin shades correlated with expectations based on the predicted BGA of the donors, with European donors (1 and 2) predicted as having fair skin and more darkly pigmented donors of East Asian origin predicted as having dark skin (Table 5.18). In summary, the preliminary validation of the five multiplexes on simulated casework DNA samples has demonstrate that the developed forensic intelligence assay can be used to successfully genotype a variety of biological DNA samples for the prediction of BGA and EVCs.

Table 5.17: Inferred regional BGA of simulated casework sample donors.
Note that the population designation 'NAFME' refers to North African or Middle Eastern ancestry.

Sample	Declared BGA	Inferred BGA		
		Maternal	Paternal	Bi-parental*
Donor 1	Europe	Europe/NAFME	Europe/NAFME/South Asia	W.Eurasia
Donor 2	Europe	Europe/NAFME	Europe/NAFME	W.Eurasia
Donor 3	East Asia	East Asia	East Asia	East Asia
Donor 4	East Asia	East Asia	East Asia/Oceania	East Asia

*from GenAlEx MLE for assignment of individuals to five populations: Sub-Saharan Africa, Europe, Western Eurasia (Europe and Middle East), South Asia and East Asia.

Table 5.18: Inferred hair, eye and skin colour of simulated casework sample donors.

Sample	Hair		Eye		Skin	
	Declared	Inferred*	Declared	Inferred*	Declared	Inferred*
Donor 1	Black	Light	Hazel	Blue/Green	Olive	Fair
Donor 2	D.brown	Dark	Hazel	Hazel/D.brown	Fair	Fair
Donor 3	Black	Dark	D.brown	Hazel/D.brown	Olive	Dark
Donor 4	Black	Dark	D.brown	Hazel/D.brown	Fair	Dark

*from GenAlEx MLE calculated from genotypes for nine pigmentation SNPs (rs885479, rs1426654, rs1545397, rs6058017, rs7498174, rs4778241, rs4778138, rs12913832 and rs16891982) in Multiplex 5.

5.3 Summary

This chapter discussed the preliminary validation of a novel 5-multiplex SNP genotyping system for forensic intelligence capable of simultaneously genotyping mtDNA, NRY and autosomal SNPs. The selection of candidate SNP loci and development of the five multiplex assays were previously discussed in chapter 3 and 4. As part of the preliminary validation of the SNP genotyping system, a pilot population genetic analysis of the haplotypes and genotypes of 146 individuals was carried out. Samples were divided into the Sub Saharan African, North African, Middle Eastern, European, South and East Asian groups according to their self-declared ancestry over two paternal and maternal generations. Overall, the five optimised multiplexes performed well during population testing and could be efficiently used to obtain full profiles for all individuals tested. The efficiency of the hierarchical arrangement of the candidate haploid SNPs in Multiplex 1 to 4 was confirmed through the absence of ambiguous haplotypes for all 85 out of the 146 individuals tested using all four multiplexes. However, population testing also revealed technical modifications required including re-design of the P31 SBE primer in Multiplex 4, which failed to extend in individuals carrying the G allele.

Population genetic analyses were performed to assess the expected population differentiation that could be achieved with the candidate mtDNA, NRY and autosomal SNPs. The derived haplogroup, allele and genotype frequencies for the candidate markers revealed differences between the six population groups, whilst diversity and AMOVA results suggested that a greater degree of population differentiation was achieved with the NRY and autosomal candidate SNPs than with the mtDNA markers. This was further demonstrated by the mtDNA, NRY and autosomal genetic distances and their subsequent analysis using PCoA, NJ trees and STRUCTURE, which identified clustering of populations according to geographical location, with reduced differentiation of populations in geographical proximity such as the Middle East and South Asia. Overall, the population genetic analyses revealed that the 5-multiplex SNP genotyping system allowed for greater differentiation between continents rather than of populations within continents, especially for the mtDNA SNPs.

The performance of the developed DNA intelligence assay for the inference of BGA was investigated using MLE derived using the GenAlEx v6.5 (Peakall and Smouse, 2012) and the Snipper App Suite 2.0 (Phillips *et al.*, 2007b). The results revealed that the 5-multiplex SNP genotyping system could infer the maternal and paternal region of geographic origin with a relatively high degree of accuracy (73- 100%), particularly for Sub-Saharan African, European and East Asian ancestry (81-100%), based on the geographic affiliation of the derived mtDNA and NRY haplogroups respectively. Sub-Saharan African (89%), European (98%) and East Asian (95%) bi-parental ancestry could also be predicted very accurately (89-98%) from the autosomal SNP genotypes when using the 1000 Genomes (Phase I) samples as a reference database and when assigning individuals to four population groups representing broad geographic regions in GenAlex: Sub-Saharan Africa (90%), North Africa (95%), Eurasia (Europe, Middle East and South Asia) (94%) and East Asia (95%). Considering the maternal, paternal and bi-parental BGA inferences in combination provided a comprehensive overview of each individual's genetic history and increased the overall predictive power. The 5-multiplex SNP genotyping system was observed to correctly predict the regional BGA of least 99% of males, where at least two predicted geographic regions of origin were correct and matching.

Performance testing of the candidate pigmentation SNPs for the inference of EVC demonstrated the highly accurate (86-88%) differentiation of light (light brown, red, blonde) versus dark (black, dark brown) hair based on the genotypes for SNPs rs12913832 and rs4778138 as well as of blue/green (95%) versus hazel/dark brown (81%) eyes when using the genotypes for all nine pigmentation SNPs. Prediction of skin colour was generally observed to be less accurate except for dark skin which was inferred with 82% accuracy. In addition, sensitivity testing of the five multiplex SNP assays also demonstrated advanced performance with low starting amounts of genomic DNA down to 0.1ng and for the analysis of routine casework biological samples. A thorough discussion of the results presented throughout this thesis as well as the conclusions and future directions for this research are provided in the final chapter.

CHAPTER 6

GENERAL DISCUSSION

Chapter 6. General Discussion

6.1 Overview of the project

The primary objective of this study was to develop a novel forensic intelligence test comprising five separate multiplex SNaPshot™ (Life Technologies, USA) assays capable of simultaneously genotyping mitochondrial DNA (mtDNA), non-recombining Y-chromosome (NRY) and autosomal ancestry and pigmentation phenotype informative SNPs. A total of 49 haploid markers comprising 21 mtDNA and 28 NRY SNPs that differentiate diagnostic haplogroups for major continental populations were selected for inclusion in the DNA intelligence assay. In addition, 13 autosomal ancestry informative markers (AIM) were also selected, of which nine SNPs also have known associations with human pigmentation variation. The SNaPshot® single-base extension (SBE) kit (Life Technologies, USA) was chosen as the SNP genotyping technique as it has been used for the development of numerous multiplex assays targeting different types of markers (Paneto *et al.*, 2011, van Oven *et al.*, 2011a, Ballantyne *et al.*, 2012) and is simple, sensitive and robust whilst utilising equipment already available in most forensic laboratories (Sobrino *et al.*, 2005). The assay development phase including the selection of candidate loci, PCR and extension primer design as well as the optimisation of Multiplex 1 to 5 were previously discussed in chapters 3 and 4. Following the completion of multiplex assay development, an initial performance evaluation of the 5-multiplex SNP genotyping system for the inference of maternal, paternal and bi-parental biogeographical ancestry (BGA) and external visible characteristics (EVC) was undertaken through large-scale population testing, population genetic and statistical analyses. Preliminary casework validations of the forensic intelligence assay including an assessment of assay predictive power, robustness and sensitivity for the analysis of different types of commonly encountered forensic biological specimens were also performed.

6.2 Performance of the 5-multiplex SNP assay during population testing

The genotyping performance of the 5-multiplex SNP assay comprising Multiplex 1 to 5 was investigated through population testing of 116 males and 30 females (section 2.1.5). Donors were grouped into six major population groups (Sub-Saharan Africa, North Africa, Middle East, Europe, South Asia and East Asia) based on their self-declared ancestry tracing back to two preceding generations. For each selected individual, both parents, as well as maternal and paternal grandparents, were required to have originated from the same population. As discussed in chapter 3, the haploid mtDNA and NRY markers were grouped into Multiplex 1 to 4 according to a hierarchical scheme, where the result of Multiplex 1 is used to identify the next multiplex to perform. This hierarchical design strategy was selected as it is widely applied for assays typing uni-parentally inherited markers and facilitates the large number of loci required to be analysed in such lineage studies (Paracchini *et al.*, 2002, Brion, 2005, Onofri *et al.*, 2006, Wiesbauer *et al.*, 2006, van Oven *et al.*, 2011b, Ballantyne *et al.*, 2012). The mtDNA and NRY haplogroups arranged together in Multiplex 1 were somewhat similar to those studied by Van Oven *et al.* (2011b) and Onofri *et al.* (2006), however this hierarchical assay was the first of its kind to apply a hierarchical arrangement to simultaneously type two different classes of markers.

To confirm the efficiency of the hierarchical grouping of our 21 mtDNA and 28 NRY candidate SNPs, 85 out of the 146 population samples were genotyped using all four multiplexes. The absence of ambiguous haplotypes confirmed that the hierarchical arrangement of the haploid markers in Multiplex 1 to 4 was correct. The remaining 61 population samples were genotyped using only the necessary multiplex(es) required to derive the final mtDNA and NRY haplogroups from the hierarchical assay, but all 146 individuals were genotyped using Multiplex 5, in which were grouped the autosomal markers. Our decision to group mtDNA and NRY SNPs that define haplogroups with similar geographical origins in the same multiplex, as previously implemented by Van Oven *et al.* (2011a) and (2011b), was also efficient. This is because although five multiplexes were developed, only a limited number of assays are required to obtain complete mtDNA and NRY haplotypes and autosomal genotypes, which minimises

processing time and amount of DNA sample used. In general, individuals of Sub-Saharan African, South and East Asian origin could be genotyped using approximately two to three out of the five multiplexes, whilst Europeans, North Africans and Middle Eastern individuals required up to four multiplexes.

During the course of this study, new emerging population studies and sequencing data have resulted in the refinement of the global mtDNA phylogeny. At the start of this project, some of the selected mtDNA SNPs already demonstrated homoplasy within the phylogeny and allowed for the unintentional detection of some additional haplogroups that are relatively rare such as M1a1d. However, the very recent revision of the updated mtDNA phylogeny led to the discovery that a large proportion of the candidate mtDNA SNPs are now recurrent and can be used to identify a larger number of rare haplogroups. For this reason, the phylogeny for the 21 candidate mtDNA SNPs was refined and is shown in Figure 6.1. It must be emphasised that these recent homoplasies were not known at the time of hierarchical assay development. However, no ambiguous haplotypes were observed when genotyping individuals with the same declared maternal and paternal ancestry with the currently implemented hierarchical grouping of haplogroups of similar geographic origin within the same multiplex reaction. Hence, the proposed order of multiplexes to be performed based on the result of Multiplex 1 is still recommended (refer to Figure 3.3 to Figure 3.6). The updated phylogeny is still however provided because these new haplogroups may be encountered when typing admixed individuals for whom the conventional multiplex genotyping order may not apply.

Overall, the five optimised multiplexes performed well during population testing and could be efficiently used to derive the mtDNA and NRY haplogroups as well as autosomal genotypes of all individuals tested. Allele drop-out of larger loci such as T10238C and C10400T in Multiplex 1 and M343 in Multiplex 3 occurred on rare occasions, due to low DNA template concentration on the donor FTA® cards, however repeating these samples with one to two FTA® discs achieved full profiles. However, the G allele at the P31 locus, which defines the East Asian O2 NRY haplogroup, was observed to be a null allele in 9 out of 34 males of declared East Asian

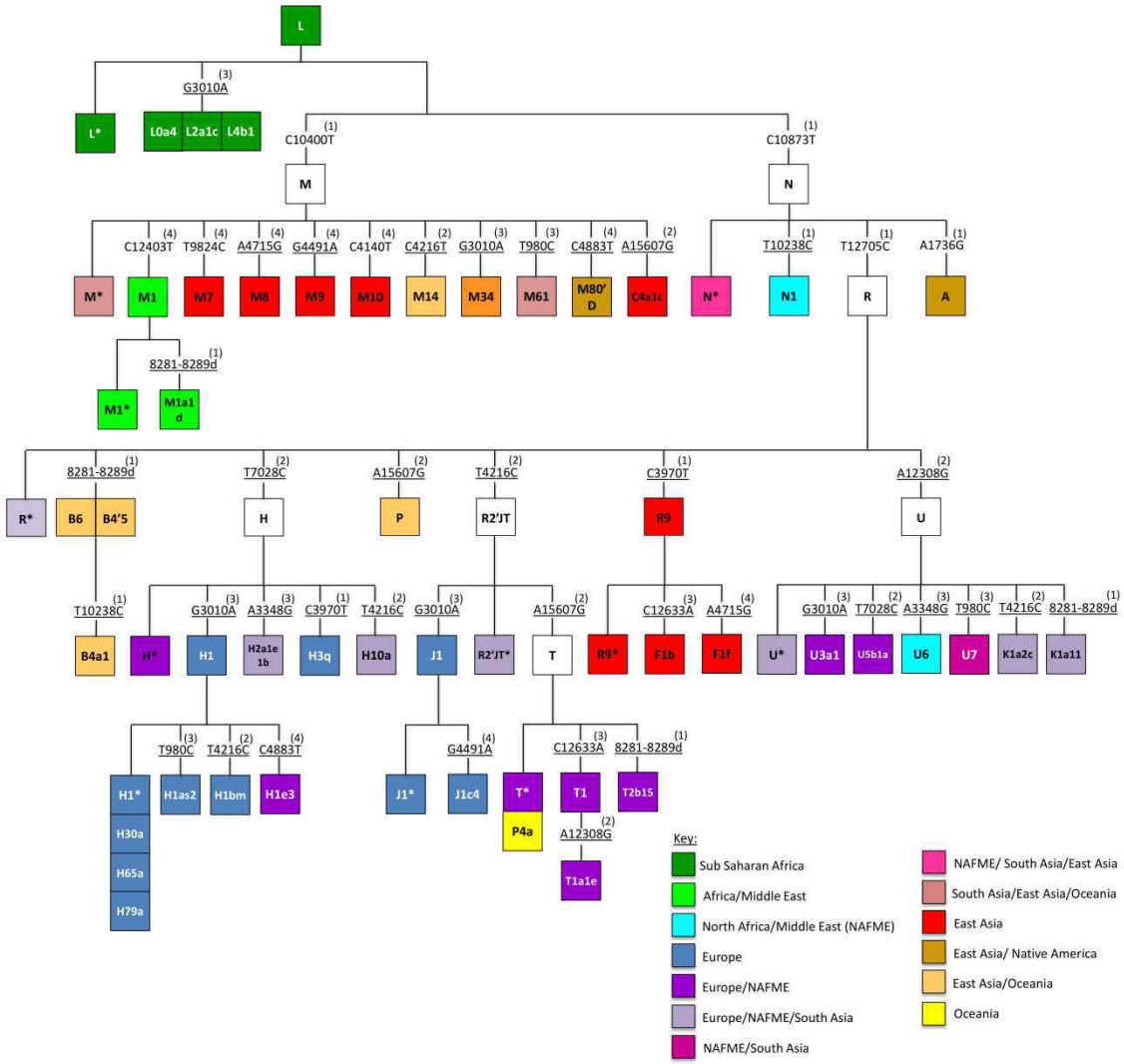


Figure 6.1: Updated Phylogenetic tree of the 21 selected mtDNA SNPs showing the geographical origin of the defining haplogroups.

Several of the candidate SNPs have been recently identified as recurrent and now define other rare haplogroups to produce this updated phylogeny. The phylogeny follows the van Oven *et al.* (2009) nomenclature corresponding to the most updated mDNA tree (Build 15, updated 30th September 2012, available at http://www.phylotree.org). The coloured boxes represent the region of geographical origin of the haplogroup as specified in the key. The candidate SNPs for this research are shown along the branches as nucleotide positions and substitutions observed within the revised Cambridge Reference Sequence (rCRS). The suffix 'd' denotes a deletion and the underlined SNPs indicate recurrent mutations. The bracketed numbers (1, 2, 3 and 4) associated with each SNP represent the multiplex in which the SNP is grouped in this project.

origin and could only be typed in singleplex using 4μM of P31 single-base extension (SBE) primer. The null G allele was only encountered in East Asian males and no issue with the detection of the ancestral A allele were noted for any of the remaining 107 males tested. It was deduced that the G substitution caused the P31 SBE primer to anneal non-specifically to surrounding regions upstream and downstream of the SNP site, which were observed to contain strings of T nucleotides interspersed with G bases. Due to the discovery of this issue during the late stages of this project, the SBE primer was not redesigned however it is recommended that a replacement P31 SBE primer is designed from the alternate stand for future applications.

6.3 Population genetic analyses of mtDNA, NRY and autosomal haplotype and genotype data

To investigate the degree of population differentiation that could be achieved using the 21 mtDNA, 28 NRY and 13 autosomal candidate SNPs, a pilot population genetic analysis was undertaken on the mtDNA and NRY haplotypes obtained for 116 males and the genotypes of 146 individuals originating from six major populations prevalent in Australia (section 5.2.1). Unfortunately, due to the wide scope of this project and the significant length of time invested in the development of five multiplex assays, time and resource constraints resulted in a smaller number of samples collected and genotyped than preferred. Some population groups such as Sub-Saharan Africa are very small ($N<10$), thus, frequency based analyses must be interpreted with caution. However the general population genetic trends observed in this pilot study of the Sub-Saharan African, North African, Middle Eastern, European, South and East Asian populations are presented and discussed in this section.

Analysis of the frequencies of the tested SNPs in the six major sub-population groups prevalent in Australia revealed distinct differences between the distribution of mtDNA, NRY and autosomal SNPs in these population groups (section 5.2.3). The mtDNA and NRY haplogroup distributions obtained correlated well with previous reported results in the literature, as compiled in Table 3.1 and Table 3.2. The geographic distribution of the mtDNA and NRY haplogroups in the six populations of interest demonstrated that the selected haploid SNPs could differentiate Sub-Saharan African, European and East

Asian populations well (Figure 5.4). However mtDNA haplogroups H*, H1, J1, N1 U* and U7 as well as NRY haplogroups E1b1b, F*, G, J2 and T demonstrated overlapping distributions between individuals of declared European, North African, Middle Eastern and South Asian origin (Figure 5.4). Generally, a greater number of mtDNA haplogroups was shared between these populations compared to NRY haplogroups, which suggests that these populations are more difficult to distinguish from each other using the selected mtDNA SNPs (Figure 5.4). However, the presence of high frequency and population specific NRY haplogroups such as E1b1b, F*, H, and R1a1 (Figure 5.4B) confirms that the NRY markers are more informative when distinguishing individuals of North African, Middle Eastern and South Asian origin.

As expected, the calculated diversity measures depicted in Table 5.1 correlated with the observed mtDNA and NRY haplogroup distribution (Figure 5.4), with populations demonstrating the least number of haplotypes such as Sub-Saharan Africa and South Asia showing the lowest gene diversities. Overall, higher gene (H) and nucleotide (π) diversity measures were observed for mtDNA haplotypes compared to NRY haplotypes, though the opposite was found for the mean number of pairwise differences (Table 5.1). This is due to a greater number of mtDNA sub-haplogroups targeted in this assay, which are all part of the same lineage and thus show only minimal sequence differences between them. In comparison, a greater number of NRY haplogroups (as opposed to sub-haplogroups) that identify different lineages and which demonstrate greater sequence differences were included in this assay. The analysis of molecular variance (AMOVA) results also showed that the degree of diversity within populations was greater for mtDNA (78%) compared to the NRY (63%) haplotypes (Table 5.1), whilst larger pairwise Φ_{PT} values were observed for the NRY haplotypes compared to the mtDNA haplotypes (Table 5.3), further suggesting that greater population differentiation is achieved with the NRY candidate SNPs than the mtDNA markers.

The mtDNA and NRY diversity results correlate with widespread observations that genetic differences between human populations are greater for the Y-chromosome compared to mtDNA (Seielstad *et al.*, 1998, Jorde *et al.*, 2000, Kayser *et al.*, 2003). These genetic differences have been attributed to a higher female migration rate due to patrilocality (Seielstad *et al.*, 1998, Hamilton *et al.*, 2005), a cultural practice ubiquitous in 70% of human populations (Murdoch, 1981), where it is more common for a woman

to migrate to the birthplace of her husband than the opposite. In patrilocal societies, the genetic variation within the population is expected to be greater for mtDNA than the Y-chromosome and greater genetic differentiation is expected for the Y-chromosome compared to the mtDNA. Despite the small sample size of this pilot study, the diversity (Table 5.1), AMOVA (Figure 5.10) and pairwise Φ_{PT} (Table 5.3) estimates for the Sub-Saharan African, North African, Middle Eastern, European, South and East Asian populations all showed genetic differences characteristic of patrilocal societies, including higher diversity for mtDNA data and greater population differentiation using the NRY SNPs, similar to previous observations for the hill tribes of Thailand (Oota *et al.*, 2001, Hamilton *et al.*, 2005, Besaggio *et al.*, 2007) and the Semende and Besemah people of Indonesia (Gunnarsdottir *et al.*, 2011).

Migration, which promotes gene flow and admixture between groups, is not only affected by culture and language but is also facilitated by geographical proximity. For all populations, the pairwise M values which provide an estimate of the number of migrants exchanged (gene flow) between two populations, were also found to be greater for mtDNA than the Y-chromosome (Table 5.3). Furthermore, all population pairs that demonstrated the lowest degree of genetic differentiation (Φ_{PT} <6%) and the highest M values were populations that shared geographical borders such as Europe/Middle East/North Africa and South Asia/East Asia (Table 5.3). In conclusion, the observed genetic differences between the mtDNA and NRY data for the six global populations support previous findings that geographical location and cultural practices such as patrilocality affect migration patterns and gene flow and contribute to the observed differences in between-population variation for mtDNA and the Y-chromosome (Jorde *et al.*, 2000, Oota *et al.*, 2001, Kayser *et al.*, 2003, Hamilton *et al.*, 2005, Besaggio *et al.*, 2007, Gunnarsdottir *et al.*, 2011).

The mtDNA and NRY are respectively solely maternally and paternally inherited in the absence of recombination and represent ancient lineage based ancestries which capture the generational history of an individual and reflect the geographical origin of maternal and paternal lineage ancestors (Royal *et al.*, 2010). On the other hand, the autosomes, which are shuffled through random assortment and recombination, only represent a portion of an individual's genetic history at any given locus and reflect more recent ancestry contributions. Population genetic analyses of the 13 autosomal SNP genotypes

of 146 individuals originating from Sub-Saharan Africa, North Africa, Middle East, Europe, South and East Asia obtained with Multiplex 5 showed both similarities and differences to the mtDNA and NRY results. The Sub-Saharan African, European and East Asian populations demonstrated contrasting autosomal allele and genotype frequencies to one another at each of the 13 candidate SNPs (Figure 5.5 to Figure 5.9) which were comparable to the literature (refer to section 4.2.1 for citations) and the frequencies retrieved from the 1000 Genomes (Phase I) database (Table 4.2).

The Sub-Saharan African population was distinguishable from other populations at the DARC, F13B, FUT2 and ASIP loci, whilst distinctive allele and genotype frequencies were observed for the East Asian population at the MC1R, OCA2-1, OCA2-2 and OCA2-4 loci. In contrast, Europeans demonstrated distinct frequency differences at the HERC2, MATP, SLC24A5, OCA2-2 and IL-6 loci. Several of the autosomal SNPs selected were also identified as fixed difference markers in the Sub-Saharan African, European and East Asian populations, where one allele is exclusive to one population and the alternate allele is found exclusively in the others. For example the SLC24A5*A and MATP*C alleles are fixed in Europeans and the DARC*G allele is fixed in Africans as previously reported (Phillips *et al.*, 2007b).

As shown in Figure 5.5 to Figure 5.9, populations that share geographical borders and are closely genetically related, such as North/Sub-Saharan Africa, Europe/North Africa/Middle East and South/East Asia, generally demonstrated similar autosomal allele and genotype frequencies. This was expected as the candidate SNPs were primarily selected to differentiate major continental groups (Africa, East Asia and Europe). However, closely genetically related populations could also be differentiated at various loci. For example, the Sub-Saharan and North African populations could be distinguished at the HERC2, MATP, SLC24A5, F13B, FUT2, OCA2-3, IL-6 and ASIP loci and North Africa and the Middle East at the DARC, F13B and SLC24A5 loci. Europe and the Middle East were distinguishable at the HERC2, MATP, SLC24A5, OCA2-3 and ASIP loci whilst South and East Asian individuals showed contrasting allele and genotype frequencies at the MC1R, OCA2-1, SLC24A5, OCA2-4 and IL-6 loci. This suggested that Multiplex 5 would also be able to differentiate the three major continental groups from these other inter-continental populations. Estimating the level of significance for the observed frequency differences between the six population

groups would have been beneficial, nonetheless the trends observed in Figure 5.5 to Figure 5.9 are reasonably distinctive for the Sub-Saharan African, European and East Asian populations and it is clearly evident that the 13 candidate SNPs can successfully differentiate these continental populations in the absence of statistical significance. However validations with larger data sets would enable calculation of the level of significance for close genetically related populations such as North Africa, Middle East and South Asia.

Heterozygosity which measures the probability that an individual will carry two different alleles (Hartl and Clark, 2006) was also calculated to assess the degree of diversity in each of the six populations of interest. For intelligence purposes, markers that demonstrate low heterozygosity (high homozygosity), which is characterised by lower observed heterozygosity (H_o) compared to expected heterozygosity (H_e) are preferred, as this favours marker specificity to a given population. In these cases, the diversity is low at that locus within a specific population (i.e. the Fixation index or inbreeding coefficient is high) and generally not in others, in order to achieve adequate differentiation between the population groups of interest (Budowle and van Daal, 2008). This is since SNPs that exhibit low heterozygosity are less polymorphic within a population but can exhibit variation between populations.

Low heterozygosity was observed at multiple autosomal loci including OCA2-1, which showed increased homozygosity in five of the six populations as well as for ASIP, which demonstrated similar results except in the European population (Table 5.2). The HERC2, MATP, OCA2-2, SLC24A5 and OCA2-3 loci showed high homozygosity in up to four populations, typically North Africa, Europe, the Middle East and East Asia (Table 5.2). The greatest difference between H_o and H_e was detected for ASIP and F13B in Sub-Saharan and SLC24A5 in Sub-Saharan Africa, the Middle East and South Asia (Table 5.2). These results suggest that these candidate markers were successfully selected and that these loci appear to be most useful in distinguishing the six populations of interest. However, in many cases the differences observed were minimal and determining whether these differences were significant would enable for more definite conclusions to be drawn from these results.

The autosomal AMOVA results shown in Figure 5.13 were similar to those obtained for the NRY data (Figure 5.10B). A global F_{ST} of 0.38 indicated that 38% of the genetic variation occurred among the six populations. This suggests that the autosomal and NRY SNPs can overall, differentiate the populations of interest more efficiently than the mtDNA markers, for which a large proportion of the genetic variation (78%) occurred within the populations compared to only 62% and 63% of the variation for the NRY and autosomal SNPs respectively (Figure 5.10B and Figure 5.13). The population pairwise F_{ST} values were significant between all population pairs tested for the autosomal data (p<0.01) and most populations were well differentiated (generally 20% or higher) (Table 5.4). However, the number of migrants (M) exchanged was generally lower for the autosomal SNPs (Table 5.4) compared to the haploid markers (Table 5.3). Conversely, the South Asia/Middle East and Europe/Middle East pairings showed closer genetic relationships than observed for the mtDNA (Table 5.3A) and NRY results (Table 5.3B), which demonstrated closer genetic relationships between East/South Asia, Middle East/North Africa and/or Europe/South Asia. However, generally, the candidate NRY markers appeared to be the most informative compared to the mtDNA and autosomal SNPs when distinguishing populations sharing geographic borders such as the Middle Eastern, South Asian and European populations.

Visualisation of the mtDNA, NRY and autosomal genetic distances using principal co-ordinates analysis (PCoA) and neighbor-joining (NJ) trees identified distinct differences in the degree of population differentiation achieved with each set of candidate markers. The general trend observed was clustering of populations according to geographical location with varying degrees of differentiation achieved between the haploid and autosomal SNPs. Three distinct clusters were identified from the mtDNA haplotypes which identify three continental regions- a Western Eurasian cluster comprising individuals originating from Europe, the Middle East and North Africa, an Eastern Eurasia cluster in which were grouped South and East Asian individuals and the Sub-Saharan African individuals clustering separately (Figure 5.11A). In contrast, analysis of the NRY data demonstrated four geographical clusters representing Africa (Sub-Saharan and North), Western Eurasia (Europe and South Asia), the Middle East and East Asia (Figure 5.11B). Greater population differentiation was apparent for the NRY markers compared to the mtDNA SNPs, with a clear separation of the Middle Eastern population achieved. However, clustering of the South Asian population with Europe

indicated that these two populations might not be adequately differentiated using the candidate NRY markers. The PCoA results for the autosomal SNPs showed that the Sub-Saharan African, North African, European and East Asian populations could be well differentiated, although clustering of the Middle Eastern and South Asian populations between North Africa and Europe was evident (Figure 5.14).

The NJ tree for all three sets of markers (Figure 5.12 and Figure 5.15) reflected similar genetic relationships to those observed with PCoA (Figure 5.11 and Figure 5.14) and paralleled the trends inferred from the frequency and diversity results. As expected, close genetic relationships were mostly observed between populations with common geographical borders, with Sub-Saharan Africa well differentiated from other populations. However, the mtDNA NJ tree showed a close genetic relationship between Sub-Saharan Africa, East and South Asia (Figure 5.12A). This was a clear contrast to the NRY and autosomal NJ trees, which depicted close relationships between Sub-Saharan Africa and North Africa (Figure 5.12B and Figure 5.14 respectively). Whilst Sub-Saharan Africa is not in geographical proximity to East Asia, the observed mtDNA haplogroup distribution, which shows a high incidence of haplogroup M and its sub-groups in South and East Asia (Figure 5.4A), is the likely reason for the close genetic relationship observed between these populations. Haplogroup M is one of two primary branches that split from African haplogroup L3 which encompasses all mtDNAs outside of Africa (Chen *et al.*, 1995).

Two dispersal routes for the migration of Humans out of Africa into Eurasia have been proposed-one through Africa, the Levant and North Asia (Lahr and Foley, 1994, Quintana-Murci *et al.*, 1999) and the other through Ethiopia, South Asia and onwards to East Asia and Oceania (Kivisild *et al.*, 1999). Evidence in support of the Southern route was gained through the study of mtDNA (Macaulay *et al.*, 2005, Sun *et al.*, 2006, Chandrasekar *et al.*, 2009) due to the ubiquitous nature of haplogroup M in India where it comprises 70% of lineages (Kivisild *et al.*, 2003, Metspalu *et al.*, 2004). Failure to identify basal mtDNA haplogroups M, N and R in North Asia (Torroni *et al.*, 1993, Derenko *et al.*, 2003, Macaulay *et al.*, 2005, Sun *et al.*, 2006, Derenko *et al.*, 2007) diminished the likelihood of a North Asian route and it was proposed that the presence of mtDNA haplogroups M1 and U6 in North Africa was the result of back migration (Olivieri *et al.*, 2006).

Although we observed a small incidence of haplogroups M, N and R and their sub-groups in our North African and Middle Eastern populations (Figure 5.4A), these populations showed more similar mtDNA haplogroup distributions to Europe, which explains why these groups appear to branch closely together yet more distantly from Sub-Saharan Africa in the NJ tree (Figure 5.12A). No set conclusions can be made from these results in support of either dispersal routes due to the small sample size and the limitations of the number of candidate markers and study populations of this study. However the PCoA and NJ tree results have provided a useful insight into the potential usefulness of the candidate mtDNA, NRY and autosomal SNPs in differentiating the six populations of interest for effective inferences of BGA. It appears that the developed forensic intelligence assay allows for greater differentiation between continents rather than differentiation of populations within continents, particularly for the mtDNA SNPs.

An additional investigation of population sub-structure performed on the autosomal genotypes using STRUCTURE v2.3.4 (Pritchard *et al.*, 2000, Falush *et al.*, 2003, 2007, Hubisz *et al.*, 2009) revealed that the most likely number of population clusters was four when using the admixture model and the LOCPRIOR function to consider the *a priori* population of origin when estimating ancestry proportions, as identified through the bar plots (Figure 5.17). The Evanno method (Evanno *et al.*, 2005) was initially used to determine the true number of clusters (K) and involves the calculation of the rate of change in the log probability of data between successive K values (ΔK) to estimate the uppermost hierarchical level of structure, which is identified as the K with the highest value of ΔK. However, when using the admixture model with and without considering the population of origin of individuals (LOCPRIOR), K=2 was identified as the most likely number of clusters for both analyses (Figure 5.16). This was not expected as the individuals were known to have originated from different sub-populations and were observed to cluster into at least 3 separate groups through PCoA (Figure 5.17).

It has been proposed that the Evanno method (Evanno *et al.*, 2005) does not perform as efficiently when the sample size and number of loci is small and underestimates the number of clusters when there is hierarchical population structure (Waples and Gaggiotti, 2006, Coulon *et al.*, 2008) because the ΔK approach can only identify the K for the uppermost hierarchical level (Evanno *et al.*, 2005). The autosomal NJ tree (Figure 5.15) clearly demonstrates hierarchical structuring of our six populations into

two major groups separating African from non-African populations, which coincide with the *K*=2 results obtained using the Evanno method (Evanno *et al.*, 2005) (Figure 5.16). Therefore, the bar plots of the individual ancestry proportions at each *K*, shown in Figure 5.20, demonstrated a more realistic representation of the most likely number of clusters and were accordingly used in preference to the Evanno graphs.

It has been suggested that the LOCPRIOR function produces more accurate ancestry estimates for small and medium datasets particularly when individuals are sampled from various locations (Hubisz *et al.*, 2009). This allows the proportion of individuals assigned to a specific cluster to vary according to their sampling location and allows for lower levels of sub-structure to be detected at reduced levels of divergence or with minimal data. Accordingly, the result depicted in Figure 5.17 demonstrated an improvement in the number of clusters inferred, which increased from *K*=2 when using the admixture model without considering prior population information to *K*=4 when using the LOCPRIOR function. It is evident that considering the *a priori* population of origin is able to identify lower levels of sub-structure in this small dataset, with four clusters identified which are mostly analogous with the major sub-population groups in Australia. The four major clusters clearly distinguished individuals of Sub-Saharan African, European, South Asian and East Asian origin. However, Middle Eastern individuals showed a high proportion of European ancestry, whilst North Africans showed 50% membership to the Sub-Saharan African and European clusters (Figure 5.17). The STRUCTURE results reinforced previous observations from PCoA and NJ tree regarding the close genetic relationship between Europeans, Middle Easterners and South Asians and the difficulty in differentiating these populations using the 13 candidate autosomal SNPs.

The autosomal data was further assessed for departures from Hardy-Weinberg equilibrium (HWE) and linkage equilibrium (LE) using exact tests prior to applying inference methods, which assume independence and random mating when calculating allele frequencies. To account for the increased probability of observing a false positive significant result (Type I error) when performing multiple hypothesis tests, the Bonferroni correction was applied, which involves adjusting the significance level α to α/n where n refers to the number of independent tests performed (Miller, 1981). No departures from HWE were seen following Bonferroni correction (all $p>0.0038$ for 13

tests) (Table 5.5). This is particularly interesting as departures from HWE are somewhat expected because the assumptions that a population is randomly mating with no sub-structure, that is of infinite size and which is not influenced by migration, mutation or natural selection are most often violated in modern human populations, depending on the population or markers investigated.

Four locus pairs out of 78 pairwise comparisons demonstrated departures from LE in the European and East Asian populations following Bonferroni correction ($p<0.0006$) (Table 5.6). Only loci located in the OCA2 gene (rs7495174, rs4778241, rs4778138) as well as the HERC2 locus (rs12913832) were found to be in linkage disequilibrium (LD). This finding is not surprising as the HERC2 gene is only 11.7kb upstream of the OCA2 gene and contains regulatory elements for the OCA2 pigmentation gene such as the transcription binding site for helicase-like transcription factor (HLTF), an enhancer which was shown to directly regulate transcription of the OCA2 gene (Sturm *et al.*, 2008, Sturm and Larsson, 2009, Visser *et al.*, 2012). The physical proximity of loci increases the likelihood that they will be inherited together however other forces may also contribute to LD including population sub-division and natural selection (Slatkin, 2008). Previous findings have also identified strong linkage disequilibrium within the OCA2-HERC2 region as well as large differences in allele frequencies between major global populations, which were proposed to be indicators of positive selection in Europeans (Voight *et al.*, 2006, Donnelly *et al.*, 2012). Loci that are in LD are not desirable when using inference methods based on maximum likelihood estimates (MLE), which assume independence and may introduce a slight bias when inferring ancestry. However, these loci may still be useful to differentiate populations in which the SNPs have been positively selected, with LD mapping recently applied for the successful inference of admixture in various populations (Moorjani *et al.*, 2011, Loh *et al.*, 2013). In the future, it may be beneficial to analyse the three OCA2 loci (rs7495174, rs4778241, rs4778138) and HERC2 (rs12913832) locus as a haplotype when calculating MLE in order to minimise bias.

It must be noted that small sample sizes are problematic for hypothesis testing as these often result in low power which decreases the chance that a statistically significant result reflects a true effect and thus results obtained from small data sets must be interpreted with caution (Button *et al.*, 2013). A recent study has shown that for

multiple tests, increases in sample size are required to improve power; however the increase in sample size necessary was greater for smaller numbers of tests performed on small datasets than the opposite (Lazzeroni and Ray, 2012). Failure to detect departures from HWE or LE in this dataset is most likely a reflection of the sample size and does not imply that these departures are non-existent. Furthermore, any future validations of the assay using a larger number of population samples should consider performing LD analyses using the program Haploview (Barrett *et al.*, 2005) that is specifically designed for SNPs to obtain more meaningful LD estimations. It would also be beneficial to investigate whether any associations exist between the mtDNA, NRY and/or autosomal markers and specific population groups.

6.4 Inference of biogeographical ancestry using the 5-multiplex SNP genotyping system

As discussed previously in section 6.3, the pilot population genetic analyses of the mtDNA/NRY haplotypes and autosomal genotypes revealed that the 5-multiplex SNP genotyping system could successfully differentiate between continents but that differentiation of inter-continental populations was more difficult, particularly with the candidate mtDNA SNPs. To investigate the performance of the developed forensic intelligence test for the inference of BGA, the maternal and paternal BGA of each individual was inferred from the geographic affiliation of the derived mtDNA and NRY haplogroups and the bi-parental BGA was predicted from MLE calculated from the autosomal genotypes using GenAlEx v6.5 (Peakall and Smouse, 2012) and the Snipper App Suite 2.0 (Phillips *et al.*, 2007b).

The geographic designations for the mtDNA and NRY haplogroups that were used to assign individuals (refer to Figure 3.1 and Figure 3.2) are only representative descriptors, which are based on the high prevalence of these haplogroups in the most probable population of origin. However numerous haplogroups are widespread in multiple populations and sometimes only combined regions such as Europe/Middle East/South Asia can be inferred. This is because there are no sharp genetic borders and gene flow occurs between populations inhabiting the same continent as a result of migration, cultural practices and colonisation. This is particularly evident for inter-

continental populations that are in geographical proximity such as North Africa, Middle East and South Asia which are situated along the migration route of humans as they dispersed from Africa (Luis *et al.*, 2004, Olivieri *et al.*, 2006, Rowold *et al.*, 2007, Armitage *et al.*, 2011). A large proportion of African mtDNA and NRY lineages have been identified in Middle Eastern populations, whilst North Africans have been shown to carry a considerable portion of European mtDNA and NRY lineages (Olivieri *et al.*, 2006, Armitage *et al.*, 2011, Badro *et al.*, 2013). Thus, inferences made from haploid markers, particularly those defining basal haplogroups as is the case for numerous of the markers selected in this study, need to be interpreted carefully whilst considering the local demographic history.

A relatively high percentage of individuals demonstrated prediction correlations between the declared BGA and the inferred maternal and paternal geographic region of origin, which ranged from 73% to 100% and 86% to 100% respectively (Table 5.7). In general, higher prediction correlations were observed for the NRY SNPs compared to the mtDNA SNPs, but both sets of markers inferred continental BGA, namely Sub-Saharan African, European and East Asian ancestry very accurately (78-100%). However, the inference of North African, Middle Eastern and European regional BGA showed a higher degree of correlation with declared BGA for the NRY markers, whilst inferences for South Asian ancestry were enhanced when using the mtDNA haplotypes. These results support the population genetic analyses (refer to section 5.2.4 to 5.2.6), which showed greater differentiation of continental populations and clustering of genetically related populations, which were better differentiated using NRY markers, except for South Asia. The addition of extra mtDNA markers that define less shallow haplogroups is recommended to improve the differentiation of such populations that lack gene flow barriers.

Since the geographic designations for the mtDNA and NRY haplogroups can sometimes only infer regions rather than specific populations it is possible that an individual may belong to a haplogroup which occurs at very low frequencies in a specific population and which is not included in the geographic designations specified in Figure 3.1 and Figure 3.2, but this does not necessarily imply that the prediction is inaccurate. For example, two males (14%) of declared South Asian origin were found to be NRY haplogroup K* (excluding K1, L1, M, O, P and T), which is specified as identifying

East Asian/Oceanian ancestry in Figure 3.2. However haplogroup K* (excluding K2, L, N, M, N, O and P) has been found to occur in approximately 3% of Indians (Sahoo *et al.*, 2006). As such, predictions may be facilitated by modifying the geographic designations used in this project to the broad regional labels applied in previous maternal and paternal ancestry inference genotyping systems (van Oven *et al.*, 2011a, van Oven *et al.*, 2011b, Ballantyne *et al.*, 2012). For example, instead of using the East Asian/Oceania designation for K*, one could designate the geographical region of origin as Western Eurasia/Oceania to account for small frequencies of this haplogroup in South Asia and the Middle East.

Similarly, the autosomal markers demonstrated high accuracies for the prediction of Sub-Saharan African (89%), European (98%) and East Asian (95%) ancestry according to MLE calculated using the Snipper App Suite 2.0 (Phillips *et al.*, 2007b) and the 1000 Genomes (Phase I) dataset as reference sample (Table 5.10) and were predominantly similar to those observed for the mtDNA and NRY SNPs (Table 5.7). STRUCTURE Analysis of the 1000 Genomes dataset further confirmed that the candidate autosomal SNPs were adequately selected to differentiate these three major continental populations (K=3) (Figure 5.18), as previously observed through PCoA (Figure 5.14) and NJ analyses (Figure 5.15). A large proportion of North African individuals were predicted to be of Sub-Saharan African origin, whilst most Middle Eastern and South Asian individuals were inferred to be of European origin. These results can be explained by the close genetic relationship between Sub-Saharan and North Africans as well as Europeans, Middle Easterners and South Asians previously inferred from the PCoA, NJ and STRUCTURE analysis results of the autosomal genotype data (Figure 5.14, Figure 5.15 and Figure 5.17 respectively).

Due to the lack of adequate reference datasets containing complete genotypes for all of the candidate autosomal markers for the North African, Middle Eastern and South Asian populations, MLE calculated using GenAlEx v6.5 (Peakall and Smouse, 2012) were used to investigate the predictive accuracy of the developed DNA intelligence assay for the inference of these ancestral populations. As recommended, the 'leave one out' option was used to exclude each individual to be assigned to an ancestral population when calculating the log likelihood values from the allele frequencies in the respective populations in order to reduce bias (Paetkau *et al.*, 2004). High predictive accuracies

were observed for the inference of Sub-Saharan African (89%), North African (90%) and East Asian (95%) bi-parental ancestry; however more intermediate accuracies were obtained for Europeans (76%) and South Asians (67%), with the lowest degree of success observed for the Middle Eastern populations (Table 5.11A).

Improvements in prediction accuracy were identified when the Middle Eastern and European populations were grouped into a larger Western Eurasian group (Table 5.11B), however the highest prediction accuracies ranging from 90 to 95% were observed when the European, Middle Eastern and South Asian populations were grouped together within a single Eurasia population (Table 5.11C). These results further confirmed that the selected autosomal SNPs could only differentiate major continental populations and thus are most effective for the inference of regional BGA, which was additionally reinforced by the pairwise population assignment plots depicted in Figure 5.19. Nevertheless, the BGA inference results obtained in this study were comparable to those of other autosomal ancestry tests, which used a greater number of autosomal SNPs but also struggled to adequately distinguish Eurasian populations lacking barriers to gene flow including Europe, the Middle East and Central/South Asia (Phillips *et al.*, 2007b, Kersbergen *et al.*, 2009). The addition of other autosomal AIMs capable of achieving within-continent resolution is recommended to improve predictions, with several such markers incorporated in the recently developed Eurasiaplex assay to achieve high predictive accuracies for the differentiation of European and South Asian ancestries (Phillips *et al.*, 2013).

Although self-declared ancestry can be reliably used to assign population samples into groups, individuals may not be aware of admixture events that may have occurred in their genetic history, especially ancient events. For example some European individuals have been identified to carry African mtDNA haplogroups despite not being aware of any African ancestry within their family history (Malyarchuk *et al.*, 2008). This possibility must be considered when studying uniparentally inherited markers such as mtDNA and NRY SNPs, which can only individually provide details of the maternal and paternal lineage as a whole and thus, should be combined with each other and/or autosomal markers to obtain a more comprehensive overview of an individual's ancestry. When combining the mtDNA and NRY inferences, the degree of correlation between the correctly inferred maternal and paternal region of geographic origin

remained high and ranged from 78 to 100% for most populations, yet for North African ancestry only 60% of individuals showed correlations between the declared BGA and the inferred maternal and paternal regional BGA (Table 5.8). Evidently, combining both inference results confirms the likelihood that the person originates from a specific geographic region/continental population. However, using mtDNA data to infer maternal North African ancestry did not improve results and more accurate regional BGA inferences were still obtained when using only the NRY markers for this population. This further confirmed that the selected mtDNA SNPs were more suited to the inference of broad geographical regions (i.e. continental or regional BGA) and could not adequately distinguish North Africans from Sub-Saharan Africans, Europeans and South Asians as deduced from PCoA and NJ analyses (refer to section 5.2.6).

Considering the inferred bi-parental BGA derived from the thirteen autosomal SNP genotypes in addition to the maternal and paternal regional BGA inferred from the mtDNA and NRY haplotypes, increases the overall predictive power of the multiplex assay (Table 5.12A). As shown in Table 5.12A, a high percentage of Sub-Saharan African (86%), European (72%) and East Asian (97%) males genotyped showed correct and matching predicted ancestries for all three SNP classes when comparing the bi-parental BGA deduced from Snipper (Table 5.10). Furthermore, the regional BGA of all 77 males could be accurately predicted when at least two inferred geographic regions of origin were matching. Similar overall predictive accuracies were also observed for these major continental populations when comparing the bi-parental BGA deduced from GenAlex (Table 5.11B). Only 0.9% of individuals demonstrated no correlations between all three inferred regional BGA but in these cases, one of the SNPs was still capable of correctly inferring the continental population or geographic region of origin (Table 5.12). This indicates that the regional BGA of at least 99% of tested males can be accurately predicted when at least two of the correctly inferred BGA match.

The preliminary validation results presented in this thesis confirm that combining evidence from multiple types of genetic markers provides a more comprehensive overview of a person's genetic ancestry. Although the three classes of markers may not be capable of adequately distinguishing closely genetically related populations when considered separately, in combination the ancestry information provided by the different SNPs is complementary and can assist in narrowing down the likely BGA of an

unknown DNA sample to one particular population or geographic region. Particularly, ancestry in several Eurasian sub-populations was more accurately predicted using specific SNP combinations. For example, South Asian origin which was most efficiently identified using mtDNA and autosomal markers, Middle Eastern ancestry which was most accurately predicted using mtDNA and NRY SNPs and North African and European ancestry which were best predicted through correlating paternal (NRY) and bi-parental (autosomal) ancestry (Table 5.12B).

Additionally, although the DNA intelligence assay was primarily designed for applications on male samples, the NRY SNPs also serve as a sex-determining marker to provide additional intelligence data and were not detected in any of the 30 females test samples confirming specificity of the primers. A comparison of the degree of correlation between the inferred maternal and bi-parental region of geographic origin demonstrated that the assay could also be efficiently used to infer the regional BGA of female samples. As evident in Table 5.13, the predictive accuracy for the inference of Sub-Saharan African (100%) and East Asian (97%) ancestry was very high and comparable to that achieved for males with the additional NRY markers (Table 5.12). However, more intermediate predictive accuracies were observed for European (78%), North African (70%) and South Asian (76%) ancestry, with Middle Easterners (50%) demonstrating the lowest accuracy (Table 5.13). This effectively confirms that only broad continental regions (Sub-Saharan Africa, Eurasia and East Asia) can be distinguished using mtDNA and autosomal SNPs. Thus the 5-multiplex genotyping system is most effective for male individuals because the NRY SNPs can assist in distinguishing specific population groups that are more difficult to differentiate solely based on mtDNA and autosomal data (see Figure 5.11 and Figure 5.12 for mtDNA, Figure 5.14 and Figure 5.15 for autosomal), particularly closely genetically related Eurasian populations such as Europe, North Africa, the Middle East and South Asia. Nevertheless, anticipated future developments of the current assay aiming to incorporate additional mtDNA SNPs defining deeper haplogroups to achieve greater differentiation of Eurasian populations is expected to increase the applicability of this multiplex SNP genotyping system for the analysis of female samples.

At the start of this project, conventional methods used to develop forensic intelligence tools focused on genotyping one specific type of DNA marker and preliminary

applications combining genetic evidence from multiple markers were typically performed separately. Accordingly, the premise of this study was to develop a novel forensic intelligence assay, which could be used to simultaneously genotype three different classes of SNPs for more accurate inferences of BGA. Towards the completion of this project, an all-in one diagnostic chip for forensic intelligence was reported which is capable of simultaneously genotyping 192,658 autosomal SNPs, 3,012 Y-chromosomal, 5,075 X/XY chromosomal and 428 mtDNA SNPs for kinship and BGA inference (Keating *et al.*, 2013). Ongoing technological and scientific advances are inevitable, particularly during lengthy doctoral research projects. However the recent development of the Identitas v1 Forensic Chip (Keating *et al.*, 2013) confirms the rationale for undertaking this study.

The developed 5-multiplex SNP genotyping system demonstrated comparable prediction accuracies to the Identitas v1 Forensic Chip (Keating *et al.*, 2013) for the inference of bi-parental ancestry of major continental groups (Sub-Saharan Africa, Eurasia and East Asia) (Table 5.10 and Table 5.11) and also showed high predictive power (~99%) when bi-parental ancestry inferences were supported by matrilineal and patrilineal regional ancestry, despite using a much smaller set of markers (Table 5.12). However, the chip was capable of inferring Middle Eastern and South Asian ancestry more accurately due to greater number of mtDNA and NRY markers which allowed for the assignment of deeper haplogroups, in contrast to our more basal haplogroups, which could only differentiate major continental regions. Furthermore, whilst a single multiplex typing tool is more favourable over SNaPshot™ technology, which requires more test reactions due to its limited multiplexing capability, chip technology is still comparatively costly, as it requires the purchase of specialist equipment not typically used in forensic laboratories. The developed 5-multiplex SNP genotyping system remains an effective preliminary forensic intelligence tool that can be of utility to laboratories with limited resources.

6.5 Inference of external visible characteristics using nine pigmentation associated SNPs

In addition to selecting autosomal SNPs that could differentiate major global populations, nine of the candidate autosomal SNPs (rs885479, rs1426654, rs1545397, rs6058017, rs7498174, rs4778241, rs4778138, rs12913832 and rs16891982) located in pigmentation genes were also selected for their known associations with EVCs (see section 4.2.1.2). Phenotype associations are usually identified using a dataset of individuals originating from the same population, typically from Europe where the greatest amount of hair, eye and skin colour variation exists to ensure that these associations are true and do not arise due to continental population stratification. As shown in Table 5.14, only the rs12913832 and rs4778138 SNPs demonstrated associations with dark (black/dark brown) versus light (light brown/blonde) hair colour and four SNPs (rs4798174, rs4778241, rs4778138 and rs12913832) were associated with brown versus non-brown and blue versus non-blue eye colour in our European sample in accordance with previous findings (Duffy *et al.*, 2007, Eiberg *et al.*, 2008, Han *et al.*, 2008, Sturm *et al.*, 2008, Branicki *et al.*, 2009). These results also confirmed an association between the TGT haplotype for SNPs rs7495174, rs4778241 and rs4778138 and blue eye colour in Europeans as previously reported by Duffy *et al.* (2007). Only the rs12913832 SNP was observed to be associated with skin pigmentation variation in Europeans, consistent with other studies (Han *et al.*, 2008, Branicki *et al.*, 2009).

It must be noted that the size of the European sample used to test these associations was small (N=40), which explains the large confidence intervals observed most SNPs tested (Table 5.14) and the inability in some cases to calculate the odds ratio due to a zero value for the denominator because various SNPs showed monomorphic genotypes in some phenotype categories. An example is the rs1545397 (OCA2-4) locus for which no European individual with declared light hair carried the AA or AT genotype. The small sample size imposed limitations on the statistical tests that could be used and logistic regression models which have been widely applied to investigate genotype/EVC associations (Kanetsky *et al.*, 2002, Rebbeck *et al.*, 2002, Sturm *et al.*, 2008, Branicki *et al.*, 2009, Branicki *et al.*, 2011) would have been preferred. Thus, whilst the majority of candidate pigmentation SNPs did not show significant associations with pigmentation

variation in Europeans, which contrasts with previous published findings detailed in section 4.2.1.2, this does not necessarily imply that the associations do not exist. More extensive validations of the assay using a larger dataset of European individuals are recommended to obtain a more accurate assessment of the associations of the nine candidate pigmentation SNPs with hair, eye and skin colour. Furthermore, the application of a recessive model of inheritance rather the dominant model tested in this preliminary validation has been shown to produce stronger associations and is also recommended (Branicki *et al.*, 2009).

Due to the sample size limitations of the preliminary assay validation phase, MLE were used to initially assess the predictive ability of the genotyping system for the inference of broad EVC categories of individuals originating from Sub-Saharan Africa, North Africa, Europe, Middle East and East Asia. Hair colour prediction to differentiate dark (black/dark brown) hair from light (light brown/red/blonde) hair was high (86-88%) and overall more accurate when using only the genotypes for SNPs rs12913832 and rs4778138 that showed associations with hair colour using exact tests (refer to Table 5.14) rather than when all nine pigmentation SNPs were analysed (Table 5.15A). Nonetheless, the predictive accuracy for light hair colour increased from 88 to 94% when assigning individuals based on the genotypes for the nine pigmentation SNPs, however only 78% of individuals with declared dark hair were correctly assigned to the dark hair group (Table 5.16A). The majority of dark haired individuals who were incorrectly assigned to the light hair group were individuals with light (blue/green/hazel) eyes and fair skin, which suggests that the remaining six pigmentation SNPs are likely to be more strongly correlated with eye and/or skin colour variation. Indeed, the prediction of eye and skin colour was observed to be more accurate based on the genotypes for all nine pigmentation SNPs rather than just those which showed associations in exact tests. The findings are consistent with previous reports that multiple combinations of the rs4778138, rs12913832, rs16891982, rs885479 and rs1426654 SNPs can enhance the prediction of eye and skin colour (Valenzuela *et al.*, 2010, Spichenok *et al.*, 2011, Walsh *et al.*, 2011, Pneuman *et al.*, 2012).

The nine pigmentation SNPs could predict blue/green eye colour with 95% accuracy though light/dark brown eye colour prediction was less accurate, with only 81% of

individuals correctly assigned Table 5.16B). Thirteen out of the nineteen individuals with light/dark brown eye colour who were incorrectly assigned to the blue/green group had declared their eye colour to be hazel/light brown. Some shades of hazel are lighter and appear more similar to green, which may explain why the eye colour of these individuals was predicted to be blue/green. Various eye prediction assays also reported problems with the prediction of intermediate eye colours such as green and hazel due to the continuous nature of the trait and the tendency to employ classification systems encompassing broad groups which over or under simplify the quantitative nature of eye colour variation (Liu *et al.*, 2009, Spichenok *et al.*, 2011, Pospiech *et al.*, 2012).
Similarly, dark skin was predicted with 82% accuracy from the nine pigmentation SNP genotypes but fair skin was predicted less efficiently (61%) and olive skin with a very low degree of accuracy (26%) (Table 5.16C). The majority of declared olive skin individuals were assigned to either the fair (36%) or dark (38%) skin groups suggesting that the assay cannot differentiate intermediate skin shades and was also evident for assignments based on the rs12913832 genotype only (Table 5.15). Grouping of samples into appropriate categories can affect the degree of correct assignment and the eye and skin colour self-declaration criteria implemented in this study is highly subjective to individual interpretation. It is proposed that future validation studies on larger datasets incorporate a visual criteria of different eye colour shades that participants can select from or alternatively collecting photographic evidence of the iris for each volunteer, as previously done by Liu *et al.* (2009) and Walsh *et al.* (2011), in order to improve classification of samples into more accurate groups. To improve the classification accuracy of skin colour groups, previous methods including assessment by a single observer (e.g. dermatologist) (Branicki *et al.*, 2009, Walsh *et al.*, 2011) or measurement of the melanin index/skin reflectance (Duffy *et al.*, 2004, Bonilla *et al.*, 2005) is recommended.

The predictive power achieved with the candidate pigmentation SNPs for the inference of hair, eye and skin colour in these preliminary validation stages was fairly comparable to previous EVC prediction assays (Liu *et al.*, 2009, Valenzuela *et al.*, 2010, Branicki *et al.*, 2011, Spichenok *et al.*, 2011, Walsh *et al.*, 2011, Pneuman *et al.*, 2012, Pospiech *et al.*, 2012, Ruiz *et al.*, 2013), even when only two out of the nine SNPs were used, as was the case for eye colour prediction (Table 5.15A). This confirms that typing only a few informative SNPs can also produce accurate inferences and that the candidate

pigmentation SNPs were successfully selected to differentiate major pigmentation phenotypes. Additionally, the prediction of hair and eye colour also demonstrated a relationship with declared BGA, with blue/green eye colour and light hair (light brown, blonde or red) colour predominantly only predicted in Europeans and minor predictions in Middle Eastern and North African individuals (data not shown). On the other hand, dark (black/dark brown) hair and hazel/dark brown eyes were the only predicted pigmentation phenotypes in East and South Asia and Sub-Saharan Africa, with the exception of a single East Asian individual who was predicted to have light hair despite declaring their hair to be dark brown. Most individuals originating from East and South Asia and Sub-Saharan Africa were predicated to have a dark skin shade, with Europeans, Middle Easterners and North Africans predicted to have fair or olive skin. In general the prediction results for the six populations of interest were consistent with general knowledge of the global distribution of phenotypic variation and other EVC predictions performed on a range of populations (Spichenok *et al.*, 2011, Walsh *et al.*, 2011, Pneuman *et al.*, 2012) and indicate that inferences of EVCs can also further complement BGA inferences for more comprehensive forensic intelligence, as initially expected.

Curiously, a number of European, Middle Eastern and North African individuals with declared dark brown eyes were incorrectly predicted to have blue/green eyes. These individuals demonstrated heterozygote genotypes (AG) at rs12913832 suggesting that as carriers of the rs12913832*G allele they were incorrectly assigned to the blue/green eye colour group. This is most likely arising because MLE are calculated from allele frequencies and do not take into account the genotype (i.e. the number of minor alleles), thereby resulting in prediction errors for these heterozygote individuals. The rs12913832 has shown the strongest associations with blue/brown eye colour variation in this (Table 5.14) and other studies (Eiberg *et al.*, 2008, Sturm *et al.*, 2008, Branicki *et al.*, 2009) with blue eye colour determined by the rs12913832 GG genotype in contrast to the AG and AA genotypes which have been associated with non-blue eye colour (intermediate or brown) (Eiberg *et al.*, 2008, Sturm *et al.*, 2008). Several currently available EVC prediction assays applied multinomial logistic regression which allows for the specification of the number of minor alleles (additive model) or inheritance status (dominant or recessive model) to establish a prediction model from genotypes obtained from large datasets of European samples (Liu *et al.*, 2009, Valenzuela *et al.*,

2010, Branicki *et al.*, 2011, Walsh *et al.*, 2011). This was attempted in this study, however the small sample size of the European sample could not produce an accurate model. Future multiplex assay validation studies would benefit from a much larger European dataset that may be used for multinomial logistic regression to identify and further confirm phenotypic associations and potentially develop a prediction model for our candidate pigmentation SNPs.

The preliminary validation results confirmed that the developed forensic intelligence assay could not only infer regional (continental) BGA but also hair and eye colour with a relatively high degree of accuracy, similar to that achieved by the Identitas v1 Forensic Chip (Keating *et al.*, 2013). Although only broad phenotype categories could be inferred rather than specific hair and eye colour categories due to the small sample size, the predictive power of our assay was higher for the inference of dark (86%) and light (88%) hair when using only two SNPs (Table 5.15A) as well as for blue/green (95%) and light/dark brown (81%) eyes when using all nine pigmentation SNPs (Table 5.16B) than reported for the Identitas v1 Forensic Chip (Keating *et al.*, 2013) for each separate hair (black, brown, red and blonde) and eye (brown, blue and intermediate) colour category.

The nine candidate pigmentation SNPs included in Multiplex 5 were also valuable for the prediction of darker skin colour which was not tested using the Identitas v1 Forensic Chip (Keating *et al.*, 2013), with 82% of declared dark skinned individuals correctly assigned to this group (Table 5.16C). Conversely, due to the limitations of the self-declaration criteria available to the donors in the consent forms used, the true predictive ability of the SNPs for the inference of fair and intermediate skin colour could not be adequately determined. Nonetheless, it must be emphasized that skin colour prediction remains the most difficult and least effective in comparison to hair and eye colour, with current assays only still capable of achieving intermediate prediction accuracies (Branicki *et al.*, 2009, Valenzuela *et al.*, 2010, Spichenok *et al.*, 2011).

6.6 Assay sensitivity and application in forensic casework

The 5-multiplex SNP genotyping system for forensic intelligence demonstrated a high degree of sensitivity in preliminary tests, with full profiles obtained for all five multiplexes with as little as 0.1ng of starting genomic DNA (Figure 5.1 to Figure 5.3). The only exception was the ASIP rs6058017 locus, which could only be detected up to 0.3ng of DNA (Figure 5.3). The ASIP marker was highly problematic throughout assay development and required typing in a separate singleplex SNaPshot™ reaction (refer to section 4.3.1.2.3). Whilst this locus showed distinguishing allele and genotype frequencies capable of differentiating Sub-Saharan Africans from other populations (Figure 5.9), numerous other candidate SNPs such as rs2814778, rs6003 and rs1800021 were capable of achieving better differentiation, suggesting that this locus could be removed (Figure 5.5 and Figure 5.7). Furthermore, the ASIP rs6058017 locus did not demonstrate any associations with hair, eye and skin colour in contrast to previous reports (Kanetsky *et al.*, 2002, Bonilla *et al.*, 2005), although this may be due to the sample size limitations of our preliminary validation study as discussed previously in section 6.5. Thus, to facilitate genotyping and improve EVC prediction it is recommended that the rs6058017 locus is replaced with an alternative ASIP marker, rs2424984 which has also demonstrated associations with hair, eye and skin colour and was observed to be a better predictor of skin colour than rs6058017 particularly when associated with SNPs rs16891982 and rs1426654 (Valenzuela *et al.*, 2010).

Whilst a preliminary sensitivity threshold of 0.1ng of starting genomic DNA template was observed for the complete 5-multiplex SNP genotyping system, different multiplexes showed varying sensitivity limits (Figure 5.1 to Figure 5.3). Generally, the NRY and autosomal loci were the first to drop out at low DNA template concentrations. This confirmed that maintaining larger PCR fragment lengths for the mtDNA loci to compensate for the greater copy number was beneficial in ensuring the amplification success of the smaller and less abundant NRY and autosomal loci. Despite the large mtDNA fragment sizes, full profiles for these markers could be detected across Multiplex 1 to 4 with starting genomic DNA template concentrations down to 0.03ng (Figure 5.1 and Figure 5.2).

Overall, the preliminary sensitivity of the developed assay was greater than that reported for autosomal SNP multiplexes designed for identification purposes which required 500pg of starting template (Dixon *et al.*, 2005, Sanchez *et al.*, 2006), the Identitas v1 Forensic Chip (Keating *et al.*, 2013) as well as commercially available STR multiplexes (Greenspoon *et al.*, 2004, Wang *et al.*, 2012, Green *et al.*, 2013) including the AmpFlSTR Minifiler kit (Applied Biosystems) recommended for degraded DNA typing (Mulero *et al.*, 2008). The sensitivity achieved was also relatively comparable to previous NRY SNP multiplexes for inferences of BGA (Sanchez *et al.*, 2003, Onofri *et al.*, 2006), however, the sensitivity for the mtDNA markers was much lower than previously reported in similar mtDNA multiplex assays (Brandstatter *et al.*, 2003, Vallone *et al.*, 2004, van Oven *et al.*, 2011b). Although concentrations below 10pg were not tested, the mtDNA loci C10400T and C10238T, which are amplified on the same fragment (426bp), demonstrated allele drop-out at starting DNA template concentrations below 30pg in Multiplex 1 (Figure 5.1). It is highly likely that the reduced sensitivity observed for the mtDNA SNPs is the result of using much larger fragment sizes (~200-426bp) compared to previously published mtDNA multiplex SNP assays (Brandstatter *et al.*, 2003, Vallone *et al.*, 2004, van Oven *et al.*, 2011b) which were maintained below 150bp in size. To improve the sensitivity of the assay, particularly for the analysis of degraded samples, future developments should consider reducing amplicon sizes especially for larger mtDNA fragments (e.g. C10400T and C10238T) to less than 200bp if possible.

The preliminary sensitivity results indicate that the developed DNA intelligence assay would meet the requirements of routine forensic casework. This was further confirmed by testing the five-multiplex SNP genotyping system on simulated casework samples including blood, semen and trace saliva extracted from a cigarette butt and drink can. Full profiles were obtained for all samples with the exception of the NRY P31 (Multiplex 4) and the autosomal ASIP rs6058017 (Multiplex 5) loci, which failed to amplify in the neat semen and trace saliva sample extracted from a swab of soft drink can (Figure 5.20 to Figure 5.23). These markers were problematic throughout assay development and large-scale genotyping of population samples and as discussed previously it is recommended that the P31 primer is redesigned on the alternate strand (section 6.2) and the ASIP locus replaced with the an alternative SNP. Overall, the maternal, paternal and bi-parental regional BGA as well as hair, eye and skin colour of

each sample donor was successfully inferred and correlated with declared ancestry and EVCs (Table 5.17 and Table 5.18), confirming the applicability of the developed assay in forensic casework.

It must be emphasised that the spin column based DNA extraction technique applied in this study may not be suitable for the optimum recovery of DNA, particularly for low DNA template samples such as trace saliva swabs. A variety of currently available DNA extraction kits such as the PrepFiler™ (Life Technologies, US), DNA IQ™ (Promega, USA), the EZ1 Investigator Kit (Qiagen, USA), and QIAamp Investigator BioRobot Kit (Qiagen, USA) can be used with robotic systems to facilitate automation and have been widely validated for use in forensic casework (Greenspoon *et al.*, 2006, Frégeau *et al.*, 2008, Brevnov *et al.*, 2009, Davis *et al.*, 2012, Liu *et al.*, 2012b). DNA extraction techniques utilising magnetic beads such as the DNA IQ™ (Promega, USA) and PrepFiler™ (Life Technologies, USA) kits have been reported to recover higher DNA yields from a variety of common forensic DNA compared to spin columns (Brevnov *et al.*, 2009). Thus, testing the developed 5-multiplex SNP assay on genomic DNA extracted using alternative, more sensitive techniques is expected to increase the yield of extracted genomic DNA for improved genotyping results.

Furthermore, fluorometric quantitation and not PCR-based sample quantitation was used to quantitate DNA in this study prior to genotyping the extracted DNA samples. Quantitative PCR techniques such as the Quantifiler® kit (Life Technologies, USA) are human DNA specific and very sensitive with reliable detection of up to 16pg of DNA reported (Green *et al.*, 2005). In contrast, fluorometric quantitation methods which incorporate dyes that bind specifically to double stranded DNA such as the PicoGreen® DS reagent (Life Technologies, USA) or the Quant-iT™ HS reagent (Life Technologies, USA) are less sensitive when measuring very low DNA concentrations due to stochastic effects with reported detection limits of 50pg and 200pg of DNA respectively according to manufacturer's specifications. Additionally, these reagents are not specific to human DNA, thus can be problematic for use on forensic samples, which are often contaminated with non-Human biological material. Thus, it is anticipated that greater multiplex assay sensitivity thresholds than those observed for the preliminary sensitivity testing and casework validation will be achieved when using improved DNA extraction

and quantitation techniques that are widely available in laboratories and have already been validated for forensic casework.

Although a preliminary sensitivity threshold of 100ng of starting genomic DNA template was identified, further sensitivity studies using DNA extracted from donors originating from different populations are also recommended to investigate the detection limit of alternate alleles at each locus in the various multiplexes. Additional testing of assay performance on a larger and more varied pool of simulated casework samples from a greater selection of substrates, including degraded DNA and testing on admixed individuals for successful inference of BGA and EVCs is also recommended. Assay performance when genotyping DNA mixtures should also be considered since the signals of the fluorescent dyes used for SBE are imbalanced and would render the identification of mixtures difficult. Whilst the signal ratio of the dyes remains relatively consistent and could be used to derive the major and minor contributors in a mixture when using diploid autosomal SNPs, however this may become more complex and less apparent for haploid SNPs because only one peak is detected. Notably, it is evident that more extensive validation studies should be pursued in order to identify clearer sensitivity and mixture thresholds.

6.7 Conclusions and Future Directions

This thesis has presented the successful development of a novel forensic intelligence assay capable of simultaneously genotyping mtDNA, NRY and autosomal SNPs for the inference of multiple intelligence elements including regional BGA, EVCs and sex determination. The 5-multiplex SNP genotyping system has revealed relatively high prediction correlations between declared BGA and inferred maternal (73-100%), paternal (86-100%) and bi-parental (85-95%) geographic region of origin, with greater than 99% of tested individuals demonstrating correct and matching regional BGA inferences from all three DNA types. High prediction accuracies were also observed for the inference of EVCs including hair (86-88%) and eye colour (81-95%). The DNA intelligence assay has also demonstrated advanced performance with low starting amounts of genomic DNA and for the analysis of routine casework biological samples.

Nonetheless, a few technical modifications of the multiplex SNP intelligence assay are suggested for further assay optimisations. Instrument and resource limitations led to the development and optimisation of the 5-multiplex assay using the outdated ABI310 capillary electrophoresis system (Life Technologies, USA). Whilst POP-6 was used instead of POP-4, as this increased the resolution of SBE fragments (see section 3.3.1.10), this polymer is no longer commonly available because the new generation of capillary sequencers such as the ABI 3500 (Life Technologies, USA) only run with either POP-4 or POP-7. The latter consists of a higher concentration polymer, which theoretically should increase the resolution achieved with POP-6. Thus, it is anticipated that the current DNA intelligence assay would perform well using the same run parameters (Table 2.7) on newer instruments, which also have improved fluorescence detection sensitivities. However, testing the developed 5-Multiplex assay on newer capillary instruments is recommended to properly assess assay performance, particularly if amendments to run time are required for the 36cm capillary array.

Future developments that aim to increase the population resolution achieved with each set of markers are recommended as well as a revision of the current selected SNP loci; especially the mtDNA markers. More informative SNPs should be selected to differentiate closely genetically related Eurasian (inter-continental) populations. In particular, since the most recent updated mtDNA phylogeny has identified that a large proportion of the candidate mtDNA SNPs are in fact recurrent, it is proposed that many of these markers be replaced with less homoplastic SNPs. Additional autosomal SNPs associated with hair, eye and skin colour are also recommended for incorporation into future developments of the assay to improve prediction of these traits. Notably, the recent discovery of markers associated with body height, age and facial morphology (Lango Allen *et al.*, 2010, Teschendorff *et al.*, 2010, Zubakov *et al.*, 2010, Liu *et al.*, 2012a, Paternoster *et al.*, 2012) provides the potential for such markers to be included in future versions of the DNA intelligence multiplex SNP assay. These associations are only preliminary and only low predictive values have been currently achieved, however incorporation of these SNPs should be contemplated in the future once more data becomes available.

Whilst SNP genotyping using SBE technology is robust and cost effective and has a high multiplexing capability, the development of a comprehensive and all-inclusive

DNA intelligence tool requires a very large number of candidate SNPs. While several measures were taken in this study to minimise the number of test reactions required such as grouping SNPs associated with similar geographic regions together, performing only one test would ultimately be most advantageous. The rapid development of high-resolution SNP microarrays and next-generation sequencing is fast becoming the future of DNA intelligence tools. Ultimately, it would be more practical to consolidate the mtDNA, NRY and autosomal candidate SNPs in this study as well as any additional markers selected in future developments within a single next-generation genotyping platform. This would not only allow for a larger multiplexed panel, but would also provide high-throughput analysis of a larger number of samples that can be analysed simultaneously and would provide more meaningful population data about the candidate mtDNA, NRY and autosomal SNPs. Although, the DNA intelligence SNP assay presented in this thesis still necessitates minor technical improvements and more extensive testing and validation, it is anticipated that future adaptations of the assay will also prove to be a valuable intelligence tool to the forensic science community to be used in conjunction with or in the absence of eye witness evidence in routine investigations, as well as for national security, missing person and mass disaster situations. It may be especially useful at the present time for laboratories or research groups with limited equipment and resources and for which the similar, recently developed Identitas v1 Forensic Chip (Keating *et al.*, 2013) is not currently accessible.

References

Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., Hurles, M. E. & McVean, G. A. 2010. A map of human genome variation from population-scale sequencing. *Nature,* 467**,** 1061-73.

Alaluf, S., Barrett, K., Blount, M. & Carter, N. 2003. Ethnic variation in tyrosinase and TYRP1 expression in photoexposed and photoprotected human skin. *Pigment Cell Res,* 16**,** 35-42.

Alloco, D., Song, Q., Gibbons, G., Ramoni, M. & Kohane, I. 2007. Geography and genography: prediction of continental origin using randomly selected single nucleotide polymorphisms. *BMC Genomics,* 8.

Allwood, J. S. & Harbison, S. 2013. SNP model development for the prediction of eye colour in New Zealand. *Forensic Sci Int Genet,* 7**,** 444-52.

Alvarez-Cubero, M. J., Saiz, M., Martinez-Gonzalez, L. J., Alvarez, J. C., Eisenberg, A. J., Budowle, B. & Lorente, J. A. 2012. Genetic Identification of Missing Persons: DNA Analysis of Human Remains and Compromised Samples. *Pathobiology,* 79**,** 228-238.

Alvarez-Iglesias, V., Jaime, J., Carracedo, A. & Salas, A. 2007. Coding region mitochondrial DNA SNPs: Targeting East Asian and Native American haplogroups. *Forensic Sci Int Genet,* 1**,** 44-55.

Amigo, J., Phillips, C., Lareu, M. & Carracedo, A. 2008a. The SNPforID browser: an online tool for query and display of frequency data from the SNPforID project. *Int J Legal Med,* 122**,** 435-40.

Amigo, J., Salas, A., Phillips, C. & Carracedo, A. 2008b. SPSmart: adapting population based SNP genotype databases for fast and comprehensive web access. *BMC Bioinformatics,* 9**,** 428.

Amorim, A. & Pereira, L. 2005. Pros and cons in the use of SNPs in forensic kinship investigation: a comparative analysis with STRs. *Forensic Sci Int,* 150**,** 17-21.

Ancans, J., Tobin, D. J., Hoogduijn, M. J., Smit, N. P., Wakamatsu, K. & Thody, A. J. 2001. Melanosomal pH controls rate of melanogenesis, eumelanin/phaeomelanin ratio and melanosome maturation in melanocytes and melanoma cells. *Exp Cell Res,* 268**,** 26-35.

Andréasson, H., Nilsson, M., Budowle, B., Lundberg, H. & Allen, M. 2006. Nuclear and mitochondrial DNA quantification of various forensic materials. *Forensic Sci Int,* 164**,** 56-64.

Andreasson, H., Nilsson, M., Styrman, H., Pettersson, U. & Allen, M. 2007. Forensic mitochondrial coding region analysis for increased discrimination using pyrosequencing technology. *Forensic Sci Int Genet,* 1**,** 35-43.

Anjos, M., Carvalho, M., Andrade, L., Lopes, V. & Serra, A. 2004. Individual genetic identification of biological samples: a case of an aircraft accident. *Forensic Sci Int,* 146**,** s115-117.

Armitage, S. J., Jasim, S. A., Marks, A. E., Parker, A. G., Usik, V. I. & Uerpmann, H.-P. 2011. The Southern Route "Out of Africa": Evidence for an Early Expansion of Modern Humans into Arabia. *Science,* 331**,** 453-456.

Badro, D. A., Douaihy, B., Haber, M., Youhanna, S. C., Salloum, A., Ghassibe-Sabbagh, M., Johnsrud, B., Khazen, G., Matisoo-Smith, E., Soria-Hernanz, D. F., Wells, R. S., Tyler-Smith, C., Platt, D. E., Zalloua, P. A. & The Genographic, C. 2013. Y-Chromosome and mtDNA Genetics Reveal Significant Contrasts in

Affinities of Modern Middle Eastern Populations with European and African Populations. *PLoS ONE,* 8**,** e54616.
Bailliet, G., Rothhammer, F., Carnese, F. R., Bravi, C. M. & Bianchi, N. O. 1994. Founder mitochondrial haplotypes in Amerindian populations. *Am J Hum Genet,* 55**,** 27-33.
Ballantyne, K. N., van Oven, M., Ralf, A., Stoneking, M., Mitchell, R. J., van Oorschot, R. A. & Kayser, M. 2012. MtDNA SNP multiplexes for efficient inference of matrilineal genetic ancestry within Oceania. *Forensic Sci Int Genet,* 6**,** 425-36.
Bar-Yosef, O. & Belfer-Cohen, A. 2001. From Africa to Eurasia — early dispersals. *Quatern Int,* 75**,** 19-28.
Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. 2005. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics,* 21**,** 263-5.
Bastiaens, M., ter Huurne, J., Kielich, C., Gruis, N., Westendorp, R. & Vermeer, B. 2001. Melanocortin-1 receptor gene variants determine the risk of non-melanoma skin cancer independently of fair skin and red hair. *Am J Hum Genet,* 68**,** 884-94.
Battaglia, V., Fornarino, S., Al-Zahery, N., Olivieri, A., Pala, M., Myres, N. M., King, R. J., Rootsi, S., Marjanovic, D., Primorac, D., Hadziselimovic, R., Vidovic, S., Drobnic, K., Durmishi, N., Torroni, A., Santachiara-Benerecetti, A. S., Underhill, P. A. & Semino, O. 2009. Y-chromosomal evidence of the cultural diffusion of agriculture in Southeast Europe. *Eur J Hum Genet,* 17**,** 820-30.
Baye, T., Tiwari, H., Allison, D. & Go, R. 2009. Database mining for selection of SNP markers useful in admixture mapping. *BioData Min,* 2**,** 1.
Behar, D. M., Metspalu, E., Kivisild, T., Rosset, S., Tzur, S., Hadid, Y., Yudkovsky, G., Rosengarten, D., Pereira, L., Amorim, A., Kutuev, I., Gurwitz, D., Bonne-Tamir, B., Villems, R. & Skorecki, K. 2008. Counting the Founders: The Matrilineal Genetic Ancestry of the Jewish Diaspora. *PLoS ONE,* 3**,** e2062.
Bergen, A., Wang, C., Tsai, J., Jefferson, K., Dey, C., Smith, K. & Park, S. 1999. An Asian-Native American paternal lineage identified by RPS4Y resequencing and by microsatellite haplotyping. *Ann Hum Genet,* 63**,** 63-80.
Bertolotto, C., Abbe, P., Hemesath, T. J., Bille, K., Fisher, D. E., Ortonne, J.-P. & Ballotti, R. 1998. Microphthalmia Gene Product as a Signal Transducer in cAMP-Induced Differentiation of Melanocytes. *J Cell Biol,* 142**,** 827-835.
Besaggio, D., Fuselli, S., Srikummool, M., Kampuansai, J. & Bertorelle, G. 2007. Genetic variation in Northern Thailand Hill tribes: origins and relationships with social structure and linguistic differences. *BMC Evol Biol,* 7 (suppl.2)**,** S12.
Betz, A., Bassler, G., Dietl, G., Steil, X., Weyermann, G. & Pflug, W. 2001. DYS STR analysis with epithelial cells in a rape case. *Forensic Sci Int,* 118**,** 126-30.
Beyin, A. 2011. Upper Pleistocene Human Dispersals out of Africa: A Review of the Current State of the Debate. *Int J Evol Biol,* 2011**,** 1-17.
Bianchi, N. O., Catanesi, C. I., Bailliet, G., Martinez-Marignac, V. L., Bravi, C. M., Vidal-Rioja, L. B., Herrera, R. J. & López-Camelo, J. S. 1998. Characterization of Ancestral and Derived Y-Chromosome Haplotypes of New World Native Populations. *Am J Hum Genet,* 63**,** 1862-1871.
Bobadilla, J., Macek, M., Fine, J. & Farrell, P. 2002. Cystic Fibrosis: A Worldwide Analysis of CFTR Mutations—Correlation With Incidence Data and Application to Screening. *Hum Mutat,* 19**,** 575-606.
Bonilla, C., Boxill, L., Donald, S., Williams, T., Sylvester, N., Parra, E. & Dios, S. 2005. The 8818G allele of the agouti signaling protein (ASIP) gene is ancestral

and is associated with darker skin color in African Americans. *Hum Genet,* 116**,** 402-6.

Bortolini, M. C., Salzano, F. M., Thomas, M. G., Stuart, S., Nasanen, S. P., Bau, C. H., Hutz, M. H., Layrisse, Z., Petzl-Erler, M. L., Tsuneto, L. T., Hill, K., Hurtado, A. M., Castro-de-Guerra, D., Torres, M. M., Groot, H., Michalski, R., Nymadawa, P., Bedoya, G., Bradman, N., Labuda, D. & Ruiz-Linares, A. 2003. Y-chromosome evidence for differing ancient demographic histories in the Americas. *Am J Hum Genet,* 73**,** 524-39.

Bouakaze, C., Keyser, C., Crubezy, E., Montagnon, D. & Ludes, B. 2009. Pigment phenotype and biogeographical ancestry from ancient skeletal remains: inferences from multiplexed autosomal SNP analysis. *Int J Legal Med,* 123**,** 315-325.

Bowcock, A., Kidd, J., Mountain, J., Hebert, J., Carotenuto, L., Kidd, K. & Cavalli-Sforza, L. L. 1991. Drift, admixture and selection in human evolution: a study with DNA polymorphisms. *Proc Natl Acad Sci USA,* 88**,** 839-43.

Box, N., Wyeth, J., O'Gorman, L., Martin, N. & Sturm, R. 1997. Characterization of melanocyte stimulating hormone receptor variant alleles in twins with red hair. *Hum Mol Genet,* 6**,** 1891-7.

Brandstatter, A., Parsons, T. J. & Parson, W. 2003. Rapid screening of mtDNA coding region SNPs for the identification of west European Caucasian haplogroups. *Int J Legal Med,* 117**,** 291-298.

Branicki, W., Brudnik, U. & Wojas-Pelc, A. 2009. Interactions Between HERC2, OCA2 and MC1R May Influence Human Pigmentation Phenotype. *Ann Hum Genet,* 73**,** 160-170.

Branicki, W., Liu, F., van Duijn, K., Draus-Barini, J., Pośpiech, E., Walsh, S., Kupiec, T., Wojas-Pelc, A. & Kayser, M. 2011. Model-based prediction of human hair color using DNA variants. *Hum Genet,* 129**,** 443-454.

Bray, M. S., Boerwinkle, E. & Doris, P. A. 2001. High-throughput multiplex SNP genotyping with MALDI-TOF mass spectrometry: practice, problems and promise. *Hum Mutat,* 17**,** 296-304.

Brenner, C. H. 1998. Difficulties in the estimation of ethnic affiliation. *Am J Hum Genet,* 62**,** 1558-60.

Brevnov, M. G., Pawar, H. S., Mundt, J., Calandro, L. M., Furtado, M. R. & Shewale, J. G. 2009. Developmental Validation of the PrepFiler™ Forensic DNA Extraction Kit for Extraction of Genomic DNA from Biological Samples*. *Journal of Forensic Sciences,* 54**,** 599-607.

Bridges, S. J., Jeng, J., Moran, M., Kuffner, T. & Whitworth, W. 2002. Single-nucleotide polymorphisms in tumor necrosis factor receptor genes: definition of novel haplotypes and racial/ethnic differences. *Arthritis Rheum,* 46**,** 2045-50.

Brilliant, M. H. 2001. The mouse p (pink-eyed dilution) and human P genes, oculocutaneous albinism type 2 (OCA2), and melanosomal pH. *Pigment Cell Res,* 14**,** 86-93.

Brion, M. 2005. Y chromosome SNP analysis using the single-base extension: a hierarchical multiplex design. *Methods Mol Biol,* 297**,** 229-242.

Brion, M., Sanchez, J. J., Balogh, K., Thacker, C., Blanco-Verea, A. & Borsting, C. 2005. Introduction of a single nucleotide polymorphism-based "Major Y-chromosome haplogroup typing kit" suitable for predicting the geographical origin of male lineages. *Electrophoresis,* 26**,** 4411-4420.

Brookes, A. J. 1999. The essence of SNPs. *Gene,* 234**,** 177-186.

Brown, W. 1980. Polymorphism in mitochondrial DNA of humans as revealed by restriction endonuclease analysis. *Proc Natl Acad Sci USA,* 77**,** 3605-3609.

Brudnik, U., Branicki, W., Wojas-Pelc, A. & Kanas, P. 2009. The contribution of melanocortin 1 receptor gene polymorphisms and the agouti signalling protein gene 8818A>G polymorphism to cutaneous melanoma and basal cell carcinoma in a Polish population. *Exp Dermatol,* 18**,** 167-174.

Budowle, B. 2004. SNP typing strategies. *Forensic Sci Int,* 146**,** S139-S142.

Budowle, B. & van Daal, A. 2008. Forensically relevant SNP classes. *BioTechniques,* 44**,** 603-8, 610.

Burchard, E., Ziv, E., Coyle, N., Gomez, S., Tang, H., Karter, A. & Mountain, J. 2003. The importance of race and ethnic background in biomedical research and clinical practice. *N Engl J Med,* 348**,** 1170-5.

Busca, R. & Ballotti, R. 2000. Cyclic AMP a Key Messenger in the Regulation of Skin Pigmentation. *Pigment Cell Res,* 13**,** 60-69.

Butler, J., Coble, M. & Vallone, P. 2007. STRs vs. SNPs: thoughts on the future of forensic DNA testing. *Forensic Science, Medicine, and Pathology,* 3**,** 200-205.

Butler, J. M. 2004. Short tandem repeat analysis for human identity testing. *Curr Protoc Hum Genet,* Chapter 14**,** Unit 14.8.

Butler, J. M. 2005. *Forensic DNA typing: Biology, Technology and Genetics of STR markers,* 2nd ed*,* Elsevier Academic Press, London, UK.

Butler, J. M., Buel, E., Crivellente, F. & McCord, B. R. 2004. Forensic DNA typing by capillary electrophoresis using the ABI Prism 310 and 3100 genetic analyzers for STR analysis. *Electrophoresis,* 25**,** 1397-1412.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J. & Munafo, M. R. 2013. Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci,* 14**,** 365-376.

Candille, S. I., Absher, D. M., Beleza, S., Bauchet, M., McEvoy, B., Garrison, N. A., Li, J. Z., Myers, R. M., Barsh, G. S., Tang, H. & Shriver, M. D. 2012. Genome-wide association studies of quantitatively measured skin, hair, and eye pigmentation in four European populations. *PLoS ONE,* 7**,** e48294.

Cann, H. M., de Toma, C., Cazes, L., Legrand, M.-F., Morel, V., Piouffre, L., Bodmer, J., Bodmer, W. F., Bonne-Tamir, B., Cambon-Thomsen, A., Chen, Z., Chu, J., Carcassi, C., Contu, L., Du, R., Excoffier, L., Ferrara, G. B., Friedlaender, J. S., Groot, H., Gurwitz, D., Jenkins, T., Herrera, R. J., Huang, X., Kidd, J., Kidd, K. K., Langaney, A., Lin, A. A., Mehdi, S. Q., Parham, P., Piazza, A., Pistillo, M. P., Qian, Y., Shu, Q., Xu, J., Zhu, S., Weber, J. L., Greely, H. T., Feldman, M. W., Thomas, G., Dausset, J. & Cavalli-Sforza, L. L. 2002. A Human Genome Diversity Cell Line Panel. *Science,* 296**,** 261-262.

Cann, R., Stoneking, M. & Wilson, M. R. 1987. Mitochondrial DNA and human evolution. *Nature,* 325**,** 31-36.

Carina, M. S., Thijessen, N. & Himla, S. 2009. SNaPshot minisequencing to resolve mitochondrial macro-haplogroups found in Africa. *Electrophoresis,* 30**,** 3657-3664.

Carvajal-Carmona, L. G., Ophoff, R., Service, S., Hartiala, J., Molina, J., Leon, P., Ospina, J., Bedoya, G., Freimer, N. & Ruiz-Linares, A. 2003. Genetic demography of Antioquia (Colombia) and the Central Valley of Costa Rica. *Hum Genet,* 112**,** 534-41.

Carvajal-Carmona, L. G., Soto, I. D., Pineda, N., Ortíz-Barrientos, D., Duque, C., Ospina-Duque, J., McCarthy, M., Montoya, P., Alvarez, V. M., Bedoya, G. & Ruiz-Linares, A. 2000. Strong Amerind/White Sex Bias and a Possible

Sephardic Contribution among the Founders of a Population in Northwest Colombia. *Am J Hum Genet,* 67**,** 1287-1295.
Casanova, M. 1985. A human Y-linked DNA polymorphism and its potential for estimating genetic and evolutionary distance. *Science,* 230**,** 1403-6.
Cavalli-Sforza, L. L. 2005. The Human Genome Diversity Project: past, present and future. *Nat Rev Genet,* 6**,** 333-340.
Cavalli-Sforza, L. L. & Marcus, W. 2003. The application of molecular genetic approaches to the study of human evolution. *Nat Genet,* 33**,** 266-75.
Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. 1994. *The history and geography of human genes,* Princeton University Press, Princeton, NJ.
Cegelski, C., Waits, L. & Anderson, N. 2003. Assessing population structure and gene flow in Montana wolverines (Gulo gulo) using assignment based approaches. *Mol Ecol,* 12**,** 2907-2918.
Cerezo, M., Cerny, V., Carracedo, A. & Salas, A. 2009. Applications of MALDI-TOF MS to large-scale human mtDNA population-based studies. *Electrophoresis,* 30**,** 3665-73.
Chandrasekar, A., Kumar, S., Sreenath, J., Sarkar, B. N., Urade, B. P., Mallick, S., Bandopadhyay, S. S., Barua, P., Barik, S. S., Basu, D., Kiran, U., Gangopadhyay, P., Sahani, R., Prasad, B. V. R., Gangopadhyay, S., Lakshmi, G. R., Ravuri, R. R., Padmaja, K., Venugopal, P. N., Sharma, M. B. & Rao, V. R. 2009. Updating Phylogeny of Mitochondrial DNA Macrohaplogroup M in India: Dispersal of Modern Human in South Asian Corridor. *PLoS ONE,* 4**,** e7447.
Chen, Y.-S., Torroni, A., Excoffier, L., Santachiara-Benerecetti, S. & Wallace, D. 1995. Analysis of mtDNA variation reveals the most ancient of all human continent specific haplogroups. *Am J Hum Genet,* 57**,** 133-149.
Choudhry, S., Coyle, N. E., Tang, H., Salari, K., Lind, D., Clark, S. L., Tsai, H.-J., Naqvi, M., Phong, A., Ung, N., Matallana, H., Avila, P. C., Casal, J., Torres, A., Nazario, S., Castro, R., Battle, N. C., Perez-Stable, E. J., Kwok, P.-Y., Sheppard, D., Shriver, M. D., Rodriguez-Cintron, W., Risch, N., Ziv, E. & Burchard, E. G. 2006. Population stratification confounds genetic association studies among Latinos. *Hum Genet,* 118**,** 652-664.
Cinnioglu, C., King, R., Kivisild, T., Kalfoglu, E., Atasoy, S. & Cavalleri, G. 2004. Excavating Y-chromosome haplotype strata in Anatolia. *Hum Genet,* 114**,** 127-148.
Clausen, H. & Hakomori, S.-i. 1989. ABH and Related Histo-Blood Group Antigens; Immunochemical Differences in Carrier Isotypes and Their Distribution1. *Vox Sang,* 56**,** 1-20.
Coble, M. D., Just, R. S., O'Callaghan, J. E., Letmanyi, I. H., Peterson, C. T., Irwin, J. A. & Parsons, T. J. 2004. Single nucleotide polymorphisms over the entire mtDNA genome that increase the power of forensic testing in Caucasians. *Int J Legal Med,* 118**,** 137-146.
Coble, M. D., Loreille, O. M., Wadhams, M. J., Edson, S. M., Maynard, K., Meyer, C. E., Niederstätter, H., Berger, C., Berger, B., Falsetti, A. B., Gill, P., Parson, W. & Finelli, L. N. 2009. Mystery Solved: The Identification of the Two Missing Romanov Children Using DNA Analysis. *PLoS ONE,* 4**,** e4838.
Collins-Schramm, H., Chima, B., Morii, T., Wah, K. & Figueroa, Y. 2004. Mexican American ancestry-informative markers: examination of population structure and marker characteristics in European Americans, Mexican Americans, Amerindians and Asians. *Hum Genet,* 114**,** 263-71.

Cone, R., Koppula, S., Vage, D., Klungland, H., Boston, B. & Chen, W. 1996. The melanocortin receptors: agonists, antagonists, and the hormonal control of pigmentation. *Recent Prog Horm Res,* 51**,** 287-317.

Connor, A. & Stoneking, M. 1994. Assessing ethnicity from human mitochondrial DNA types determined by hybridization with sequence-specific oligonucleotides. *J Forensic Sci,* 39**,** 1360-71.

Corach, D., Lao, O., Bobillo, C., Van Der Gaag, K., Zuniga, S., Vermeulen, M., Van Duijn, K., Goedbloed, M., Vallone, P. M., Parson, W., De Knijff, P. & Kayser, M. 2010. Inferring Continental Ancestry of Argentineans from Autosomal, Y-Chromosomal and Mitochondrial DNA. *Ann Hum Genet,* 74**,** 65-76.

Cordaux, R., Aunger, R., Bentley, G., Nasidze, I., Sirajuddin, S. M. & Stoneking, M. 2004. Independent origins of Indian caste and tribal paternal lineages. *Curr Biol,* 14**,** 231-5.

Coulon, A., Fitzpatrick, J. W., Bowman, R., Stith, B. M., Makarewich, C. A., Stenzler, L. M. & Lovette, I. J. 2008. Congruent population structure inferred from dispersal behaviour and intensive genetic surveys of the threatened Florida scrub-jay (Aphelocoma cœrulescens). *Mol Ecol,* 17**,** 1685-1701.

Cox, E. D., Hoffmann, S. C., DiMercurio, B. S., Wesley, R. A., Harlan, D. M., Kirk, A. D. & Blair, P. J. 2001. Cytokine Polymorphic Analyses Indicate Ethnic Differences in the Allelic Distribution of Interleukin-2 and Interleukin-61. *Transplantation,* 72**,** 720-6.

Cruciani, F., La Fratta, R., Santolamazza, P., Sellitto, D., Pascone, R., Moral, P., Watson, E., Guida, V., Colomb, E. B., Zaharova, B., Lavinha, J., Vona, G., Aman, R., Cali, F., Akar, N., Richards, M., Torroni, A., Novelletto, A. & Scozzari, R. 2004. Phylogeographic analysis of haplogroup E3b (E-M215) Y chromosomes reveals multiple migratory events within and out of Africa. *Am J Hum Genet,* 74**,** 1014-22.

Cruciani, F., La Fratta, R., Trombetta, B., Santolamazza, P., Sellitto, D., Colomb, E. B., Dugoujon, J.-M., Crivellaro, F., Benincasa, T., Pascone, R., Moral, P., Watson, E., Melegh, B., Barbujani, G., Fuselli, S., Vona, G., Zagradisnik, B., Assum, G., Brdicka, R., Kozlov, A. I., Efremov, G. D., Coppa, A., Novelletto, A. & Scozzari, R. 2007. Tracing Past Human Male Movements in Northern/Eastern Africa and Western Eurasia: New Clues from Y-Chromosomal Haplogroups E-M78 and J-M12. *Mol Biol Evol,* 24**,** 1300-1311.

Daniel, R., Sanchez, J. J., Nassif, N. T., Hernandez, A. & Walsh, S. J. 2009. Partial forensic validation of a 16plex SNP assay for the inference of biogeographical ancestry. *Forensic Sci Int Genet Suppl Series,* 2**,** 477-478.

Darabi, H., Czene, K., Zhao, W., Liu, J., Hall, P. & Humphreys, K. 2012. Breast cancer risk prediction and individualised screening based on common genetic variation and breast density measurement. *Breast Cancer Res,* 14**,** R25.

Davignon, J., Gregg, R. & Sing, C. 1988. Apoliporotein E polymorphism and atherosclerosis. *Atherosclerosis,* 8**,** 1-21.

Davis, C. P., King, J. L., Budowle, B., Eisenberg, A. J. & Turnbough, M. A. 2012. Extraction platform evaluations: A comparison of Automate Express™, EZ1® Advanced XL, and Maxwell® 16 Bench-top DNA extraction systems. *Legal Med (Tokyo),* 14**,** 36-39.

de Filippo, C., Barbieri, C., Whitten, M., Mpoloka, S. W., Gunnarsdóttir, E. D., Bostoen, K., Nyambe, T., Beyer, K., Schreiber, H., de Knijff, P., Luiselli, D., Stoneking, M. & Pakendorf, B. 2011. Y-Chromosomal Variation in Sub-Saharan

Africa: Insights Into the History of Niger-Congo Groups. *Mol Biol Evol,* 28**,** 1255-1269.

DeMichele, A., Martin, A.-M., Mick, R., Gor, P., Wray, L., Klein-Cabral, M., Athanasiadis, G., Colligan, T., Stadtmauer, E. & Weber, B. 2003. Interleukin-6 −174G→C Polymorphism Is Associated with Improved Outcome in High-Risk Breast Cancer. *Cancer Res,* 63**,** 8051-8056.

Derenko, M., Grzybowski, T., Malyarchuk, B., Dambueva, M. & Denisova, G. 2003. Diversity of mitochondrial DNA lineages in South Siberia. *Ann Hum Genet,* 67**,** 391-411.

Derenko, M., Malyarchuk, B., Grzybowski, T., Denisova, G. & Dambueva, M. 2007. Phylogeographic analysis of mitochondrial DNA in Northern Asian Populations. *Am J Hum Genet,* 81**,** 1025-1041.

Derricourt, R. 2005. Getting "Out of Africa": Sea Crossings, Land Crossings and Culture in the Hominin Migrations. *Journal of World Prehistory,* 19**,** 119-132.

Ding, L., Wiener, H., Abebe, T., Altaye, M., Go, R., Kercsmar, C., Grabowski, G., Martin, L., Khurana Hershey, G., Chakorborty, R. & Baye, T. 2011. Comparison of measures of marker informativeness for ancestry and admixture mapping. *BMC Genomics,* 12**,** 622.

Divne, A.-M. & Allen, M. 2005. A DNA microarray system for forensic SNP analysis. *Forensic Sci Int,* 154**,** 111-121.

Dixon, L., Murray, C., Archer, E., Dobbins, A. & Koumi, P. 2005. Validation of a 21 locus autosomal SNP multiplex for forensic identification purposes. *Forensic Sci Int,* 154**,** 62-77.

Donatien, P. D., Hunt, G., Pieron, C., Lunec, J., Taieb, A. & Thody, A. J. 1992. The expression of functional MSH receptors on cultured human melanocytes. *Arch Dermatol Res,* 284**,** 424-6.

Donnelly, M. P., Paschou, P., Grigorenko, E., Gurwitz, D., Barta, C., Lu, R. B., Zhukova, O. V., Kim, J. J., Siniscalco, M., New, M., Li, H., Kajuna, S. L., Manolopoulos, V. G., Speed, W. C., Pakstis, A. J., Kidd, J. R. & Kidd, K. K. 2012. A global view of the OCA2-HERC2 region and pigmentation. *Hum Genet,* 131**,** 683-96.

Du, J. & Fisher, D. 2002. Identification of Aim-1 as the underwhite mouse mutant and its Transcriptional regulation by MITF. *J Biol Chem,* 277**,** 402-6.

Dubut, V., Chollet, L., Murail, P., Cartault, F. & Beraud-Colomb, E. 2004. MtDNA polymorphisms in five French groups: importance of regional sampling. *Eur J Hum Genet,* 12**,** 293-300.

Duda, R. P., Hart, P. & Stork, D. 2001. *Pattern Classification,* Wiley, London.

Duffy, D., Montgomery, G., Chen, W., Zhao, Z., Le, L. & James, M. 2007. A three-single-nucleotide polymorphism haplotype in intron 1 of OCA2 explains most human eye-color variation. *Am J Hum Genet,* 80**,** 241-52.

Duffy, D. L., Box, N. F., Chen, W., Palmer, J. S., Montgomery, G. W., James, M. R., Hayward, N. K., Martin, N. G. & Sturm, R. A. 2004. Interactive effects of MC1R and OCA2 on melanoma risk phenotypes. *Hum Mol Genet,* 13**,** 447-461.

Dulik, Matthew C., Zhadanov, Sergey I., Osipova, Ludmila P., Askapuli, A., Gau, L., Gokcumen, O., Rubinstein, S. & Schurr, Theodore G. 2012. Mitochondrial DNA and Y Chromosome Variation Provides Evidence for a Recent Common Ancestry between Native Americans and Indigenous Altaians. *Am J Hum Genet,* 90**,** 229-246.

Durso, D. F., Bydlowski, S. P., Hutz, M. H., Suarez-Kurtz, G., Magalhaes, T. R. & Pena, S. D. 2014. Association of genetic variants with self-assessed color categories in Brazilians. *PLoS ONE,* 9**,** e83926.

Earl, D. & vonHoldt, B. 2012. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv Genet Resour,* 4**,** 359-361.

Easton, D. F., Pooley, K. A., Dunning, A. M., Pharoah, P. D. P., Thompson, D., Ballinger, D. G., Struewing, J. P., Morrison, J., Field, H., Luben, R., Wareham, N., Ahmed, S., Healey, C. S. & Bowman, R. 2007. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature,* 447**,** 1087-93.

Edson, S., Ross, J., Coble, M. D., Parsons, T. J. & Barritt, S. M. 2004. Naming the dead- Confronting the realities of rapid identification of degraded skeletal remains. *Forensic Sci Rev,* 16**,** 63-90.

Edwards, M., Bigham, A., Tan, J., Li, S., Gozdzik, A., Ross, K., Jin, L. & Parra, E. J. 2010. Association of the *OCA2* Polymorphism His615Arg with Melanin Content in East Asian Populations: Further Evidence of Convergent Evolution of Skin Pigmentation. *PLoS Genet,* 6**,** e1000867.

Eiberg, H., Troelsen, J., Nielsen, M., Mikkelsen, A., Mengel-From, J., Kjaer, K. & Hansen, L. 2008. Blue eye color in humans may be caused by a perfectly associated founder mutation in a regulatory element located within the *HERC2* gene inhibiting *OCA2* expression. *Hum Genet,* 123**,** 177-187.

Eskdale, J., McNicholl, J., Wordsworth, P., Jonas, B., Huizinga, T., Field, M. & Gallagher, G. 1998. Interleukin-10 microsatellite polymorphisms and IL-10 locus alleles in rheumatoid arthritis susceptibility. *The Lancet,* 352**,** 1282-1283.

Evanno, G., Regnaut, S. & Goudet, J. 2005. Detecting the number of clusters of individuals using the software structure: a simulation study. *Mol Ecol,* 14**,** 2611-2620.

Excoffier, L. & Lischer, H. E. L. 2010. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol Resour,* 10**,** 564-567.

Excoffier, L., Smouse, P. E. & Quattro, J. M. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics,* 131**,** 479-91.

Falush, D., Stephens, M. & Pritchard, J. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics,* 164**,** 1567-87.

Falush, D., Stephens, M. & Pritchard, J. 2007. Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Mol Ecol Notes,* 7**,** 574-578.

Ferrer-Admetlla, A., Sikora, M., Laayouni, H., Esteve, A., Roubinet, F., Blancher, A., Calafell, F., Bertranpetit, J. & Casals, F. 2009. A Natural History of FUT2 Polymorphism in Humans. *Mol Biol Evol,* 26**,** 1993-2003.

Finnila, S., Lehtonen, M. & Majamaa, K. 2001. Phylogenetic network for European mtDNA. *Am J Hum Genet,* 68**,** 1475-1484.

Fishman, D., Faulds, G., Jeffery, R., Mohamed-Ali, V., Yudkin, J. S., Humphries, S. & Woo, P. 1998. The effect of novel polymorphisms in the interleukin-6 (IL-6) gene on IL-6 transcription and plasma IL-6 levels, and an association with systemic-onset juvenile chronic arthritis. *J Clin Invest,* 102**,** 1369-76.

Flaquer, A., Rappold, G. A., Wienker, T. F. & Fischer, C. 2008. The human pseudoautosomal regions: a review for genetic epidemiologists. *Eur J Hum Genet,* 16**,** 771-9.

Foster, M. & Sharp, R. 2002. Race, ethnicity, and genomics: social classifications as proxies of biological heterogeneity. *Genome Res,* 12**,** 844-50.

Frandberg, P.-E., Doufexis, M., Kapas, S. & Chhajlani, V. 1998. Human pigmentation phenotype: a point mutation generates nonfunctional MSH receptor. *Biochem Biophys Res Commun,* 245**,** 490-92.

Frégeau, C. J., Marc Lett, C., Elliott, J., Yensen, C. & Fourney, R. M. 2008. Automated Processing of Forensic Casework Samples Using Robotic Workstations Equipped with Nondisposable Tips: Contamination Prevention. *J Forensic Sci,* 53**,** 632-651.

Freire-Aradas, A., Fondevila, M., Kriegel, A. K., Phillips, C., Gill, P., Prieto, L., Schneider, P. M., Carracedo, A. & Lareu, M. V. 2012. A new SNP assay for identification of highly degraded human DNA. *Forensic Sci Int Genet,* 6**,** 341-9.

Frudakis, T., Terravainen, T. & Thomas, M. 2007. Multilocus OCA2 genotypes specify human iris colors. *Hum Genet,* 122**,** 311-326.

Frudakis, T., Thomas, M., Gaskin, Z., Venkateswarlu, K. & Chandra, K. 2003a. Sequences associated with human iris pigmentation. *Genetics,* 165**,** 2071-2083.

Frudakis, T., Venkateswarlu, K., Thomas, M., Gaskin, Z., Ginjupalli, S., Gunturi, S. & Ponnuswamy, V. 2003b. A classifier for the SNP based inference of ancestry. *J Forensic Sci,* 48**,** 771-782.

Galanter, J. M., Fernandez-Lopez, J. C., Gignoux, C. R., Barnholtz-Sloan, J., Fernandez-Rozadilla, C., Via, M., Hidalgo-Miranda, A., Contreras, A. V., Figueroa, L. U., Raska, P., Jimenez-Sanchez, G., Silva Zolezzi, I., Torres, M., Ponte, C. R., Ruiz, Y., Salas, A., Nguyen, E., Eng, C., Borjas, L., Zabala, W., Barreto, G., RondÃ³n GonzÃ¡lez, F., Ibarra, A., Taboada, P., Porras, L., Moreno, F. n., Bigham, A., Gutierrez, G., Brutsaert, T., LeÃ³n-Velarde, F., Moore, L. G., Vargas, E., Cruz, M., Escobedo, J., Rodriguez-Santana, J., Rodriguez-CintrÃ³n, W., Chapela, R., Ford, J. G., Bustamante, C., Seminara, D., Shriver, M., Ziv, E., Gonzalez Burchard, E., Haile, R., Parra, E., Carracedo, A. & for the, L. C. 2012. Development of a Panel of Genome-Wide Ancestry Informative Markers to Study Admixture Throughout the Americas. *PLoS Genet,* 8**,** e1002554.

Gan, R.-J., Pan, S.-L., Mustavich, L., Qin, Z., Cai, X. & Qian, J. 2008. Pinghua population as an exception of Han Chinese's coherent genetic structure. *Hum Genet,* 53**,** 303-313.

Gettings, K. B., Lai, R., Johnson, J. L., Peck, M. A., Hart, J. A., Gordish-Dressman, H., Schanfield, M. S. & Podini, D. S. 2014. A 50-SNP assay for biogeographic ancestry and phenotype prediction in the U.S. population. *Forensic Sci Int Genet,* 8**,** 101-8.

Gill, P. 2001. An assessment of the utility of single nucleotide polymorphisms (SNPs) for forensic purposes. *Int J Legal Med,* 114**,** 204-10.

Gill, P., Jeffreys, A. J. & Werrett, D. J. 1985. Forensic application of DNA 'fingerprints'. *Nature,* 318**,** 577-9.

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

Gomes, V., Sánchez-Diz, P., Amorim, A., Carracedo, Á. & Gusmão, L. 2010. Digging deeper into East African human Y chromosome lineages. *Hum Genet,* 127**,** 603-613.

Gonder, C., Mortensen, H., Reed, F., De Sousa, A. & Tishkoff, S. 2007. Whole mtDNA genome sequence analysis of ancient African lineages. *Mol Biol Evol,* 24**,** 757-768.

Gonzalez, A. M., Brehm, A., Perez, J. A., Maca-Meyer, N., Flores, C. & Cabrera, V. M. 2003. Mitochondrial DNA affinities at the Atlantic fringe of Europe. *Am J Phys Anthropol,* 120**,** 391-404.

Gonzalez, A. M., Larruga, J. M., Abu-Amero, K., Shi, Y., Pestano, J. & Cabrera, V. M. 2007. Mitochondrial lineage M1 traces an early human backflow to Africa. *BMC Genomics,* 8.

Graf, J., Hodgson, R. & van Daal, A. 2005. Single nucleotide polymorphisms in the MATP gene are associated with normal human pigmentation variation. *Hum Mutat,* 25**,** 278-284.

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H., Hansen, N. F., Durand, E. Y., Malaspinas, A. S., Jensen, J. D., Marques-Bonet, T., Alkan, C., Prufer, K., Meyer, M., Burbano, H. A., Good, J. M., Schultz, R., Aximu-Petri, A., Butthof, A., Hober, B., Hoffner, B., Siegemund, M., Weihmann, A., Nusbaum, C., Lander, E. S., Russ, C., Novod, N., Affourtit, J., Egholm, M., Verna, C., Rudan, P., Brajkovic, D., Kucan, Z., Gusic, I., Doronichev, V. B., Golovanova, L. V., Lalueza-Fox, C., de la Rasilla, M., Fortea, J., Rosas, A., Schmitz, R. W., Johnson, P. L., Eichler, E. E., Falush, D., Birney, E., Mullikin, J. C., Slatkin, M., Nielsen, R., Kelso, J., Lachmann, M., Reich, D. & Paabo, S. 2010. A draft sequence of the Neandertal genome. *Science,* 328**,** 710-22.

Green, R. L., Lagacé, R. E., Oldroyd, N. J., Hennessy, L. K. & Mulero, J. J. 2013. Developmental validation of the AmpFℓSTR® NGM SElect™ PCR Amplification Kit: A next-generation STR multiplex with the SE33 locus. *Forensic Sci Int Genet,* 7**,** 41-51.

Green, R. L., Roinestad, I. C., Boland, C. & Hennessy, L. K. 2005. Developmental validation of the quantifiler real-time PCR kits for the quantification of human nuclear DNA samples. *J Forensic Sci,* 50**,** 809-25.

Greenspoon, S. A., Ban, J. D., Pablo, L., Crouse, C. A., Kist, F. G., Tomsey, C. S., Glessner, A. L., Mihalacki, L. R., Long, T. M., Heidebrecht, B. J., Braunstein, C. A., Freeman, D. A., Soberalski, C., Nathan, B., Amin, A. S., Douglas, E. K. & Schumm, J. W. 2004. Validation and implementation of the PowerPlex 16 BIO System STR multiplex for forensic casework. *J Forensic Sci,* 49**,** 71-80.

Greenspoon, S. A., Sykes, K. L. V., Ban, J. D., Pollard, A., Baisden, M., Farr, M., Graham, N., Collins, B. L., Green, M. M. & Christenson, C. C. 2006. Automated PCR setup for forensic casework samples using the Normalization Wizard and PCR Setup robotic methods. *Forensic Sci Int,* 164**,** 240-248.

Grimes, E., Noake, P., Dixon, L. & Urquhart, A. 2001. Sequence polymorphism in the human melanocortin 1 receptor gene as an indicator of the red hair phenotype. *Forensic Sci Int,* 122**,** 124-129.

Gunnarsdottir, E. D., Nandineni, M. R., Li, M., Myles, S., Gil, D., Pakendorf, B. & Stoneking, M. 2011. Larger mitochondrial DNA than Y-chromosome differences between matrilocal and patrilocal groups from Sumatra. *Nat Commun,* 2.

Guo, S. W. & Thompson, E. A. 1992. Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics,* 48**,** 361-72.

Halder, I., Shriver, M., Thomas, M., Fernandez, J. & Frudakis, T. 2008. A panel of ancestry informative markers for estimating individual biogeographical ancestry and admixture from four continents: utility and applications. *Hum Mutat,* 29**,** 648-658.

Hamilton, G., Stoneking, M. & Excoffier, L. 2005. Molecular analysis reveals tighter social regulation of immigration in patrilocal populations than in matrilocal populations. *Proc Natl Acad Sci USA,* 102**,** 7476-7480.

Hammer, M. 1995. A recent common ancestry for human Y chromosomes. *Nature,* 378**,** 376-8.

Hammer, M., Chamberlain, V., Kearney, V., Stover, D. & Zhang, G. 2006. Population structure of Y chromosome SNP haplogroups in the United States and forensic implications for constructing Y chromosome STR databases. *Forensic Sci Int,* 164**,** 45-55.

Hammer, M., Karafet, T., Rasanayagam, A., Wood, E., Altheide, T. & Jenkins, T. 1998. Out of Africa and back again: nested cladistic analysis of human Y chromosome variation. *Mol Biol Evol,* 15**,** 427-441.

Hammer, M., Karafet, T., Redd, A., Jarjanazi, H., Santachiara-Benerecetti, S. & Soodyall, H. 2001. Hierarchical patterns of global human Y-chromosome diversity. *Mol Biol Evol,* 18**,** 1189-203.

Hammer, M., Spurdle, A., Karafet, T., Bonner, M., Wood, E., Novelletto, A. & Malaspina, P. 1997. The Geographic Distribution of Human Y Chromosome Variation. *Genetics,* 145**,** 787-805.

Hammer, M. & Zegura, S. 1996. The Role of the Y Chromosome in Human Evolutionary Studies. *Evol Anthropol,* 5**,** 116-134.

Han, J., Kraft, P., Nan, H., Guo, Q. & Chen, C. 2008. A genome-wide association study identifies novel alleles associated with hair color and skin pigmentation. *PLoS Genet,* 4**,** e1000074.

Harding, R., Fullerton, S., Griffiths, R., Bond, J., Cox, M., Schneider, J. & Moulin, D. 1997. Archaic African and Asian lineages in the genetic ancestry of modern humans. *Am J Hum Genet,* 60**,** 772-89.

Harding, R., Healy, E., Ray, A., Ellis, N., Flanagan, N., Todd, C. & Dixon, C. 2000. Evidence for variable selective pressures at MC1R. *Am J Hum Genet,* 66**,** 1351-61.

Hartl, D. & Clark, A. G. 2006. *Principles of population genetics,* 4th*,* Sinauer Associates, Inc., Sunderland, Massachussets.

Hassan, M. I., Aschner, Y., Manning, C. H., Xu, J. & Aschner, J. L. 2003. Racial differences in selected cytokine allelic and genotypic frequencies among healthy, pregnant women in North Carolina. *Cytokine,* 21**,** 10-16.

Hedrick, P. W. 2011a. *Genetics of populations,* 4th Edn*,* Jones & Bartlett Publishers, Sudburry.

Hedrick, P. W. 2011b. Population genetics of malaria resistance in humans. *Heredity,* 107**,** 283-304.

Helgason, A., Hickey, E., Goodacre, S., Bosnes, V., Stefansson, K. & Ward, R. 2001. mtDNA and the islands of the North Atlantic: estimating the proportions of Norse and Gaelic ancestry. *Am J Hum Genet,* 68**,** 723-737.

Henegariu, O., Heerema, N. A., Dlouhy, S. R., Vance, G. H. & Vogt, P. H. 1997. Multiplex PCR: critical parameters and step-by-step protocol. *BioTechniques,* 23**,** 504-11.

Henke, A., Fischer, C. & Rappold, G. 1993. Genetic map of the human pseudoautosomal region reveals a high rate of recombination in female meiosis at the Xp telomere. *Genomics,* 18**,** 478-85.

Hill, C. R., Kline, M. C., Coble, M. D. & Butler, J. M. 2008. Characterization of 26 MiniSTR Loci for Improved Analysis of Degraded DNA Samples. *J Forensic Sci,* 53**,** 73-80.

Hinds, D., Stuve, L., Nilsen, G., Halperin, E., Eskin, E., Ballinger, D., Frazer, K. & Cox, D. 2005. Whole-genome patterns of common DNA variation in three human populations. *Science,* 307**,** 1072-9.

Hoffmann, S. C., Stanley, E. M., Cox, E. D., DiMercurio, B. S., Koziol, D. E., Harlan, D. M., Kirk, A. D. & Blair, P. J. 2002. Ethnicity Greatly Influences Cytokine Gene Polymorphism Distribution. *Am J Transplant,* 2**,** 560-567.

Hollard, C., Mendisco, F., Keyser, C., Crubézy, E. & Ludes, B. 2011. First application of the Investigator DIPplex indels typing kit for the analysis of ancient DNA samples. *Forensic Sci Int Genet Suppl Series,* 3**,** e393-e394.

Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., Pearson, J. V., Stephan, D. A., Nelson, S. F. & Craig, D. W. 2008. Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays. *PLoS Genet,* 4**,** e1000167.

Honda, K., Roewer, L. & de Knijff, P. 1999. Male DNA typing from 25-year-old vaginalswabs using Y chromosomal STR polymorphisms in a retrial request case. *J Forensic Sci,* 44**,** 868-72.

Hood, L. & Rowen, L. 2013. The Human Genome Project: big science transforms biology and medicine. *Genome Med,* 5**,** 79.

Horuk, R., Chitnis, C. E., Darbonne, W. C., Colby, T. J., Rybicki, A., Hadley, T. J. & Miller, L. H. 1993. A receptor for the malarial parasite Plasmodium vivax: the erythrocyte chemokine receptor. *Science,* 261**,** 1182-1184.

Houshmand, M., Holme, E., Hanson, C., Wennerholm, U.-B. & Hamberger, L. 1997. Is paternal mitochondrial DNA transferred to the offspring following intracytoplasmic sperm injection? *Journal of Assisted Reproduction and Genetics,* 14**,** 223-227.

Howes, R. E., Patil, A. P., Piel, F. B., Nyangiri, O. A., Kabaria, C. W., Gething, P. W., Zimmerman, P. A., Barnadas, C., Beall, C. M., Gebremedhin, A., Menard, D., Williams, T. N., Weatherall, D. J. & Hay, S. I. 2011. The global distribution of the Duffy blood group. *Nat Commun,* 2**,** 266.

Hubisz, M. J., Falush, D., Stephens, M. & Pritchard, J. K. 2009. Inferring weak population structure with the assistance of sample group information. *Mol Ecol Resour,* 9**,** 1322-32.

Hublin, J. J. 2009. The origin of Neandertals. *Proc Natl Acad Sci USA,* 106**,** 16022-16027.

Hudjashov, G., Kivisild, T., Underhill, P. A., Endicott, P., Sanchez, J. J., Lin, A. A., Shen, P., Oefner, P., Renfrew, C., Villems, R. & Forster, P. 2007. Revealing the prehistoric settlement of Australia by Y chromosome and mtDNA analysis. *Proc Natl Acad Sci USA,* 104**,** 8726-8730.

Hunley, K., Healy, M. & Long, J. 2009. The Global Pattern of Gene Identity Variation Reveals a History of Long-Range Migrations, Bottlenecks, and Local Mate Exchange: Implications for Biological Race. *Am J Phys Anthropol,* 139**,** 35-46.

Hunt, G., Kyne, S., Wakamatsu, K., Ito, S. & Thody, A. J. 1995. Nle4DPhe7 α-Melanocyte-Stimulating Hormone Increases the Eumelanin:Phaeomelanin Ratio in Cultured Human Melanocytes. *J Investig Dermatol,* 104**,** 83-85.

Ingman, M., Kaessmann, H., Paabo, S. & Gyllensten, U. 2000. Mitochondrial genome variation and the origin of modern humans. *Nature,* 408**,** 708-13.

International Human Genome Sequencing Consortium 2001. Initial sequencing and analysis of the human genome. *Nature,* 409**,** 860-921.

International Society of Genetic Genealogy. 2012. *Y-DNA Haplogroup tree 2012: Version 7.45, 9 August 2012* [Online]. Available: http://www.isogg.org/tree/

Iwamoto, S., Omi, T., Kajii, E. & Ikemoto, S. 1995. Genomic organization of the glycoprotein D gene: Duffy blood group Fy^a/Fy^b alloantigen system is associated with a polymorphism at the 44- amino acid residue. *Blood,* 85**,** 622-626.

Iwasaki, M., Kobayashi, K., Suzuki, H., Anan, K. & Ohno, S. 2000. Polymorphism of the ABO blood group genes in Han, Kazak and Uygur populations in the Silk Route of northwestern China. *Tissue Antigens,* 56**,** 136-142.

Jakobsson, M. & Rosenberg, N. A. 2007. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics,* 23**,** 1801-1806.

Jakobsson, M., Scholz, S. W., Scheet, P., Gibbs, J. R., VanLiere, J. M., Fung, H.-C., Szpiech, Z. A., Degnan, J. H., Wang, K., Guerreiro, R., Bras, J. M., Schymick, J. C., Hernandez, D. G., Traynor, B. J., Simon-Sanchez, J., Matarin, M., Britton, A., van de Leemput, J., Rafferty, I., Bucan, M., Cann, H. M., Hardy, J. A., Rosenberg, N. A. & Singleton, A. B. 2008. Genotype, haplotype and copy-number variation in worldwide human populations. *Nature,* 451**,** 998-1003.

Jazin, E., Soodyall, H., Jalonen, P., Lindholm, E., Stoneking, M. & Gyllensten, U. 1998. Mitochondrial mutation rate revisited: hot spots and polymorphism. *Nat Genet,* 18**,** 109-110.

Ji, Y., Walkowicz, M. J., Buiting, K., Johnson, D. K., Tarvin, R. E., Rinchik, E. M., Horsthemke, B., Stubbs, L. & Nicholls, R. D. 1999. The Ancestral Gene for Transcribed, Low-Copy Repeats in the Prader-Willi/Angelman Region Encodes a Large Protein Implicated in Protein Trafficking, Which Is Deficient in Mice with Neuromuscular and Spermiogenic Abnormalities. *Hum Mol Genet,* 8**,** 533-542.

Jobling, M. & Tyler-Smith, C. 2003. The human Y chromosome: an evolutionary marker comes of age. *Nat Rev Genet,* 4**,** 598-612.

Jobling, M. A. & Gill, P. 2004. Encoded evidence: DNA in forensic analysis. *Nat Rev Genet,* 5**,** 739-51.

Johansen, P., Andersen, J. D., Borsting, C. & Morling, N. 2013. Evaluation of the iPLEX(R) Sample ID Plus Panel designed for the Sequenom MassARRAY(R) system. A SNP typing assay developed for human identification and sample tracking based on the SNPforID panel. *Forensic Sci Int Genet,* 7**,** 482-7.

Johnson, M., Wallace, D. C., Ferris, S., Rattazi, M. & Cavalli-Sforza, L. L. 1983. Radiation of human mitochondrial types analysed by restriction endonuclease cleavage pattern. *J Mol Evol,* 19**,** 255-271.

Jorde, L., Bamshad, M. & Rogers, A. 1998. Using mitochondrial and nuclear DNA markers to reconstruct human evolution. *Bioessays,* 20**,** 126-136.

Jorde, L., Watkins, W., Dixon, M., Ricker, C., Seielstad, M. T. & Batzer, M. 2000. The distribution of human genetic diversity: a comparison of mitochondrial, autosomal and Y-chromosomal data. *Am J Hum Genet,* 66**,** 978-988.

Just, R. S., Leney, M., Barritt, S. M., Los, C., Smith, B. C., Holland, T. & Parsons, T. J. 2009. The Use of Mitochondrial DNA Single Nucleotide Polymorphisms to

Assist in the Resolution of Three Challenging Forensic Cases. *J Forensic Sci,* 54**,** 887-891.

Kaestle, F. A., Kittles, R. A., Roth, A. L. & Ungvarsky, E. J. 2006. Database limitations of the evidentiary value of forensic mitochondrial DNA evidence. *Am Crim L Rev,* 43**,** 53-88.

Kajander, O., Karhunen, P., Holt, I. & Jacobs, H. 2001. Prominent mitochondrial DNA recombination intermediates in human heart muscle. *EMBO Rep,* 2**,** 1007-1012.

Kalman, Y., Kerem, E., Darvasi, A., DeMarchi, J. & Kerem, B. 1994. Difference in frequencies of the cystic fibrosis alleles, delta F508 and W1282X, between carriers and patients. *Eur J Hum Genet,* 2**,** 77-82.

Kamodyova, N., Durdiakova, J., Celec, P., Sedlackova, T., Repiska, G., Sviezena, B. & Minarik, G. 2013. Prevalence and persistence of male DNA identified in mixed saliva samples after intense kissing. *Forensic Sci Int Genet,* 7**,** 124-8.

Kanetsky, P., Swoyer, J., Panossian, S., Holmes, R. & Guerry, D. 2002. A polymorphism in the Agouti signaling protein gene is associated with human pigmentation. *Am J Hum Genet,* 70**,** 770-775.

Karafet, T., Mendez, F., Meilerman, M., Underhill, P., Zegura, S. & Hammer, M. 2008. New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. *Genome Res,* 18**,** 830-838.

Kayser, M., Brauer, S., Cordaux, R., Casto, A. & Lao, O. 2006. Melanesian and Asian origins of Polynesians: mtDNA and Y-chromosome gradients across the pacific. *Mol Biol Evol,* 23**,** 2234-2244.

Kayser, M., Brauer, S., Weiss, G., Schiefenhovel, W., Underhill, P., Shen, P., Oefner, P., Tommaseo-Ponzetta, M. & Stoneking, M. 2003. Reduced Y-chromosome, but not mitochondrial DNA, diversity in human populations from West New Guinea. *Am J Hum Genet,* 72**,** 281-302.

Kayser, M., Choi, Y., Van Oven, M., Mona, S., Brauer, S., Trent, R. & Suarkia, D. 2008a. The impact of the Austronesian expansion: evidence from mtDNA and Y chromosome diversity in the Admiralty Islands of Melanesia. *Mol Biol Evol,* 25**,** 1362-1374.

Kayser, M. & de Knijff, P. 2011. Improving human forensics through advances in genetics, genomics and molecular biology. *Nat Rev Genet,* 12**,** 179-192.

Kayser, M., Lao, O., Anslinger, K., Augustin, C., Bargel, G., Edelmann, J., Elias, S., Heinrich, M., Henke, J., Henke, L., Hohoff, C., Illing, A., Jonkisz, A., Kuzniar, P., Lebioda, A., Lessig, R., Lewicki, S., Maciejewska, A., Monies, D. M., Pawlowski, R., Poetsch, M., Schmid, D., Schmidt, U., Schneider, P. M., Stradmann-Bellinghausen, B., Szibor, R., Wegener, R., Wozniak, M., Zoledziewska, M., Roewer, L., Dobosz, T. & Ploski, R. 2005. Significant genetic differentiation between Poland and Germany follows present-day political borders, as revealed by Y-chromosome analysis. *Hum Genet,* 117**,** 428-43.

Kayser, M., Liu, F., Janssens, A. C. J. W., Rivadeneira, F., Lao, O., van Duijn, K., Vermeulen, M., Arp, P., Jhamai, M. M., van Ijcken, W. F. J., den Dunnen, J. T., Heath, S., Zelenika, D., Despriet, D. D. G., Klaver, C. C. W., Vingerling, J. R., de Jong, P. T. V. M., Hofman, A., Aulchenko, Y. S., Uitterlinden, A. G., Oostra, B. A. & van Duijn, C. M. 2008b. Three Genome-wide Association Studies and a Linkage Analysis Identify HERC2 as a Human Iris Color Gene. *Am J Hum Genet,* 82**,** 411-423.

Kayser, M., Liu, F., Janssens, A. C. J. W., Rivadeneira, F., Lao, O., van Duijn, K., Vermeulen, M., Arp, P., Jhamai, M. M., van Ijcken, W. F. J., den Dunnen, J. T.,

Heath, S., Zelenika, D., Despriet, D. D. G., Klaver, C. C. W., Vingerling, J. R., de Jong, P. T. V. M., Hofman, A., Aulchenko, Y. S., Uitterlinden, A. G., Oostra, B. A. & van Duijn, C. M. 2008c. Three Genome-wide Association Studies and a Linkage Analysis Identify HERC2 as a Human Iris Color Gene. *American Journal of Human Genetics,* 82**,** 411-423.

Keating, B., Bansal, A., Walsh, S., Millman, J., Newman, J., Kidd, K., Budowle, B., Eisenberg, A., Donfack, J., Gasparini, P., Budimlija, Z., Henders, A., Chandrupatla, H., Duffy, D., Gordon, S., Hysi, P., Liu, F., Medland, S., Rubin, L., Martin, N., Spector, T. & Kayser, M. 2013. First all-in-one diagnostic tool for DNA intelligence: genome-wide inference of biogeographic ancestry, appearance, relatedness, and sex with the Identitas v1 Forensic Chip. *Int J Legal Med,* 127**,** 559-572.

Kelly, R. J., Rouquier, S., Giorgi, D., Lennon, G. G. & Lowe, J. B. 1995. Sequence and Expression of a Candidate for the Human Secretor Blood Group α(1,2)Fucosyltransferase Gene (FUT2): Homozygosity for an enzyme-inactivating nonsense mutation commonly correlates with the non-secretor phenotype. *J Biol Chem,* 270**,** 4640-4649.

Kerem, B., Rommens, J., Buchanan, J., Markiewicz, D., Cox, T., Chakravarti, A. & Buchwald, M. 1989. Identification of the cystic fibrosis gene: genetic analysis. *Science,* 245**,** 1073-1080.

Kersbergen, P., van Duijn, K., Kloosterman, A. D., den Dunnen, J. T., Kayser, M. & de Knijff, P. 2009. Developing a set of ancestry-sensitive DNA markers reflecting continental origins of humans. *BMC Genet,* 10**,** 69.

Kidd, J., Friedlaender, F., Speed, W., Pakstis, A., De La Vega, F. & Kidd, K. 2011. Analyses of a set of 128 ancestry informative single-nucleotide polymorphisms in a global set of 119 population samples. *Investig Genet,* 2**,** 1.

Kidd, K., Pakstis, A., Speed, W., Grigorenko, E., Kajuna, S., Karoma, N., Kungulilo, S., Kim, J.-J., Lu, R.-B., Odunsi, A., Okonofua, J., Schulz, L., Zhukova, O. & Kidd, J. 2006. Developing a SNP panel for forensic identification. *Forensic Sci Int,* 164**,** 20-32.

Kim, J. J., Han, B. G., Lee, H. I., Yoo, H. W. & Lee, J. K. 2010. Development of SNP-based human identification system. *Int J Legal Med,* 124**,** 125-31.

Kim, J. J., Verdu, P., Pakstis, A. J., Speed, W. C., Kidd, J. R. & Kidd, K. K. 2005. Use of autosomal loci for clustering individuals and populations of East Asian origin. *Hum Genet,* 117**,** 511-9.

Kim, S. & Borevitz, J. 2006. Mining the HapMap to dissect complex traits. *Genome Biol,* 7**,** 310.

Kishimoto, T., Akira, S., Narazaki, M. & Taga, T. 1995. Interleukin-6 family of cytokines and gp130. *Blood,* 86**,** 1243-1254.

Kivisild, T., Bamshad, M. J., Kaldma, K., Metspalu, M., Metspalu, E., Reidla, M., Laos, S., Parik, J., Watkins, W. S., Dixon, M. E., Papiha, S. S., Mastana, S. S., Mir, M. R., Ferak, V. & Villems, R. 1999. Deep common ancestry of indian and western-Eurasian mitochondrial DNA lineages. *Curr Biol,* 9**,** 1331-4.

Kivisild, T., Reidla, M., Metspalu, E., Rosa, A., Brehm, A., Pennarun, E., Parik, J. r., Geberhiwot, T., Usanga, E. & Villems, R. 2004. Ethiopian Mitochondrial DNA Heritage: Tracking Gene Flow Across and Around the Gate of Tears. *Am J Hum Genet,* 75**,** 752-770.

Kivisild, T., Rootsi, S., Metspalu, M., Mastana, S., Kaldma, K. & Parik, J. 2003. The Genetic Heritage of the Earliest Settlers Persists Both in Indian Tribal and Caste Populations. *Am J Hum Genet,* 73**,** 313-332.

Kivisild, T., Shen, P., Wall, D. P., Do, B., Sung, R., Davis, K., Passarino, G., Underhill, P. A., Scharfe, C., Torroni, A., Scozzari, R., Modiano, D., Coppa, A., de Knijff, P., Feldman, M., Cavalli-Sforza, L. L. & Oefner, P. J. 2006. The Role of Selection in the Evolution of Human Mitochondrial Genomes. *Genetics,* 172**,** 373-387.

Kivisild, T., Tolk, H., Parik, J., Wang, C., Papiha, S. & Bandelt, H. J. 2002. The emerging limbs and twigs of the East Asian mtDNA tree. *Mol Biol Evol,* 19**,** 1737-1751.

Koda, Y., Tachida, H., Pang, H., Liu, Y., Soejima, M., Ghaderi, A. A., Takenaka, O. & Kimura, H. 2001. Contrasting Patterns of Polymorphisms at the ABO-Secretor Gene (FUT2) and Plasma α(1,3)Fucosyltransferase Gene (FUT6) in Human Populations. *Genetics,* 158**,** 747-756.

Kogelnik, A., Lott, M. T., Brown, M., Navathe, S. & Wallace, D. 1996. MITOMAP: a human mitochondrial genome database. *Nucleic Acids Res,* 24**,** 177-9.

Kohnemann, S., Hohoff, C. & Pfeiffer, H. 2009. An economical mtDNA SNP assay detecting different mitochondrial haplogroups in identical HVR 1 samples of Caucasian ancestry. *Mitochondrion,* 9**,** 370-5.

Kong, Q.-P., Bandelt, H. J., Sun, C., Yao, Y.-G., Salas, A., Achilli, A. & Wang, C. 2006. Updating the East Asian mtDNA phylogeny: a prerequisite for the identification of pathogenic mutations. *Hum Mol Genet,* 15**,** 2076-2086.

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

Krause, J., Orlando, L., Serre, D., Viola, B., Prufer, K., Richards, M. P., Hublin, J. J., Hanni, C., Derevianko, A. P. & Paabo, S. 2007. Neanderthals in central Asia and Siberia. *Nature,* 449**,** 902-4.

Krings, M., Stone, A., Schmitz, R. W., Krainitzki, H., Stoneking, M. & Paabo, S. 1997. Neandertal DNA sequences and the origin of modern humans. *Cell,* 90**,** 19-30.

Kumar, S., Banks, T. W. & Cloutier, S. 2012. SNP Discovery through Next-Generation Sequencing and Its Applications. *Int J Plant Genomics,* 2012**,** 15.

Kumar, V., Reddy, A., Babu, J., Rao, T., Langstieh, B., Thangaraj, K., Reddy, A., Singh, L. & Reddy, B. 2007. Y-chromosome evidence suggests a common paternal heritage of Austro-Asiatic populations. *BMC Evol Biol,* 7**,** 47.

Lagmay, J. P., London, W. B., Gross, T. G., Termuhlen, A., Sullivan, N., Axel, A., Mundy, B., Ranalli, M., Canner, J., McGrady, P. & Hall, B. 2009. Prognostic significance of interleukin-6 single nucleotide polymorphism genotypes in neuroblastoma: rs1800795 (promoter) and rs8192284 (receptor). *Clin Cancer Res,* 15**,** 5234-9.

Lahr, M. & Foley, R. 1994. Multiple dispersals and modern human origins. *Evol Anthropol,* 3**,** 48-60.

Lamason, R. L., Mohideen, M. A., Mest, J. R., Wong, A. C., Norton, H. L., Aros, M. C., Jurynec, M. J., Mao, X., Humphreville, V. R., Humbert, J. E., Sinha, S., Moore, J. L., Jagadeeswaran, P., Zhao, W., Ning, G., Makalowska, I., McKeigue, P. M., O'Donnell, D., Kittles, R., Parra, E. J., Mangini, N. J., Grunwald, D. J., Shriver, M. D., Canfield, V. A. & Cheng, K. C. 2005. SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science,* 310**,** 1782-6.

Lango Allen, H., Estrada, K., Lettre, G., Berndt, S. I., Weedon, M. N., Rivadeneira, F., Willer, C. J., Jackson, A. U., Vedantam, S., Raychaudhuri, S., Ferreira, T.,

Wood, A. R., Weyant, R. J., Segrè, A. V., Speliotes, E. K., Wheeler, E., Soranzo, N., Park, J.-H., Yang, J., Gudbjartsson, D., Heard-Costa, N. L., Randall, J. C., Qi, L., Vernon Smith, A., Mägi, R., Pastinen, T., Liang, L., Heid, I. M., Luan, J. a., Thorleifsson, G., Winkler, T. W., Goddard, M. E., Sin Lo, K., Palmer, C., Workalemahu, T., Aulchenko, Y. S., Johansson, A., Zillikens, M. C., Feitosa, M. F., Esko, T., Johnson, T., Ketkar, S., Kraft, P., Mangino, M., Prokopenko, I., Absher, D., Albrecht, E., Ernst, F., Glazer, N. L., Hayward, C., Hottenga, J.-J., Jacobs, K. B., Knowles, J. W., Kutalik, Z., Monda, K. L., Polasek, O., Preuss, M., Rayner, N. W., Robertson, N. R., Steinthorsdottir, V., Tyrer, J. P., Voight, B. F., Wiklund, F., Xu, J., Zhao, J. H., Nyholt, D. R., Pellikka, N., Perola, M., Perry, J. R. B., Surakka, I., Tammesoo, M.-L., Altmaier, E. L., Amin, N., Aspelund, T., Bhangale, T., Boucher, G., Chasman, D. I., Chen, C., Coin, L., Cooper, M. N., Dixon, A. L., Gibson, Q., Grundberg, E., Hao, K., Juhani Junttila, M., Kaplan, L. M., Kettunen, J., König, I. R., Kwan, T., Lawrence, R. W., Levinson, D. F., Lorentzon, M., McKnight, B., Morris, A. P., Müller, M., Suh Ngwa, J., Purcell, S., Rafelt, S., Salem, R. M., Salvi, E., *et al.* 2010. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature,* 467**,** 832-838.

Lansman, R., Shade, R., Shapira, J. & Avise, J. 1981. The use of restriction endonucleases to measure mitochondrial DNA sequence relatedness in natural populations.III. Techniques and potential applications. *J Mol Evol,* 17**,** 214-226.

Lao, O., Lu, T. T., Nothnagel, M., Junge, O., Freitag-Wolf, S., Caliebe, A., Balascakova, M., Bertranpetit, J., Bindoff, L. A., Comas, D., Holmlund, G., Kouvatsi, A., Macek, M., Mollet, I., Parson, W., Palo, J., Ploski, R., Sajantila, A., Tagliabracci, A., Gether, U., Werge, T., Rivadeneira, F., Hofman, A., Uitterlinden, A. G., Gieger, C., Wichmann, H.-E., Rüther, A., Schreiber, S., Becker, C., Nürnberg, P., Nelson, M. R., Krawczak, M. & Kayser, M. 2008. Correlation between Genetic and Geographic Structure in Europe. *Curr Biol,* 18**,** 1241-1248.

Lao, O., Vallone, P. M., Coble, M. D., Diegoli, T. M., van Oven, M., van der Gaag, K. J., Pijpe, J., de Knijff, P. & Kayser, M. 2010. Evaluating self-declared ancestry of U.S. Americans with autosomal, Y-chromosomal and mitochondrial DNA. *Hum Mutat,* 31**,** E1875-E1893.

LaRue, B. L., Ge, J., King, J. L. & Budowle, B. 2012. A validation study of the Qiagen Investigator DIPplex(R) kit; an INDEL-based assay for human identification. *Int J Legal Med,* 126**,** 533-40.

LaRue, B. L., Lagace, R., Chang, C. W., Holt, A., Hennessy, L., Ge, J., King, J. L., Chakraborty, R. & Budowle, B. 2014. Characterization of 114 insertion/deletion (INDEL) polymorphisms, and selection for a global INDEL panel for human identification. *Leg Med (Tokyo),* 16**,** 26-32.

Lazzeroni, L. C. & Ray, A. 2012. The cost of large numbers of hypothesis tests on power, effect size and sample size. *Mol Psychiatry,* 17**,** 108-114.

Le Pendu, J., Cartron, J. P., Lemieux, R. U. & Oriol, R. 1985. The presence of at least two different H-blood-group-related beta-D-gal alpha-2-L-fucosyltransferases in human serum and the genetics of blood group H substances. *Am J Hum Genet,* 37**,** 749-60.

Lee, Y., Teitelbaum, S., Wolff, M., Wetmur, J. & Chen, J. 2010. Comparing genetic ancestry and self-reported race/ethnicity in a multiethnic population in New York City. *Journal of Genetics,* 89**,** 417-423.

Legros, F., Malka, F., Frachon, P., Lombès, A. & Rojo, M. 2004. Organization and dynamics of human mitochondrial DNA. *J Cell Sci,* 117**,** 2653-2662.

Lei, S.-f., Deng, F.-y., Liu, X.-h., Huang, Q.-r., Qin, Y., Zhou, Q., Jiang, D.-k., Li, Y.-m., Mo, X.-y., Liu, M.-y., Chen, X.-d., Wu, X.-s., Shen, H., Dvornyk, V., Zhao, L., Recker, R. R. & Deng, H.-w. 2003. Polymorphisms of four bone mineral density candidate genes in Chinese populations and comparison with other populations of different ethnicity. *J Bone Miner Metab,* 21**,** 34-42.

Lessig, R., Zoledziewska, M., Fahr, K., Edelmann, J., Kostrzewa, M., Dobosz, T. & Kleemann, W. J. 2005. Y-SNP-genotyping - a new approach in forensic analysis. *Forensic Sci Int,* 154**,** 128-36.

Lewis, P. & Zaykin, D. 2002. Genetic Data Analysis: Computer program for the analysis of allelic data. version 1.1 ed.

Li, C., Zhang, S., Li, L., Chen, J., Liu, Y. & Zhao, S. 2012. Selection of 29 highly informative InDel markers for human identification and paternity analysis in Chinese Han population by the SNPlex genotyping system. *Mol Biol Rep,* 39**,** 3143-52.

Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., Cann, H. M., Barsh, G. S., Feldman, M., Cavalli-Sforza, L. L. & Myers, R. M. 2008. Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation. *Science,* 319**,** 1100-1104.

Licastro, F., Grimaldi, L. M. E., Bonafè, M., Martina, C., Olivieri, F., Cavallone, L., Giovanietti, S., Masliah, E. & Franceschi, C. 2003. Interleukin-6 gene alleles affect the risk of Alzheimer's disease and levels of the cytokine in blood and brain. *Neurobiol Aging,* 24**,** 921-926.

Lien, S., Szyda, J., Schechinger, B., Rappold, G. & Arnheim, N. 2000. Evidence for heterogeneity in recombination in the human pseudoautosomal region: high resolution analysis by sperm typing and radiation-hybrid mapping. *Am J Hum Genet,* 66**,** 557-66.

Lim, C. S., Zheng, S., Kim, Y. S., Ahn, C., Han, J. S., Kim, S., Lee, J. S. & Chae, D.-W. 2002. The −174 G to C polymorphism of Interleukin-6 gene is very rare in Koreans. *Cytokine,* 19**,** 52-54.

Lind, J., Hutcheson-Dilks, H., Williams, S., Moore, J. & M, E. 2007. Elevated male European and Female African contributions to the genomes of African American individuals. *Hum Genet,* 120**,** 713-722.

Liu, F., van der Lijn, F., Schurmann, C., Zhu, G., Chakravarty, M. M., Hysi, P. G., Wollstein, A., Lao, O., de Bruijne, M., Ikram, M. A., van der Lugt, A., Rivadeneira, F., Uitterlinden, A. G., Hofman, A., Niessen, W. J., Homuth, G., de Zubicaray, G., McMahon, K. L., Thompson, P. M., Daboul, A., Puls, R., Hegenscheid, K., Bevan, L., Pausova, Z., Medland, S. E., Montgomery, G. W., Wright, M. J., Wicking, C., Boehringer, S., Spector, T. D., Paus, T. Å., Martin, N. G., Biffar, R. & Kayser, M. 2012a. A Genome-Wide Association Study Identifies Five Loci Influencing Facial Morphology in Europeans. *PLoS Genet,* 8**,** e1002932.

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

Liu, J. Y., Zhong, C., Holt, A., Lagace, R., Harrold, M., Dixon, A. B., Brevnov, M. G., Shewale, J. G. & Hennessy, L. K. 2012b. AutoMate Express™ Forensic DNA Extraction System for the Extraction of Genomic DNA from Biological Samples*. *Journal of Forensic Sciences,* 57**,** 1022-1030.

Liu, Y., Koda, Y., Soejima, M., Pang, H., Schlaphoff, T., du Toit, E. D. & Kimura, H. 1998. Extensive polymorphism of the FUT2 gene in an African (Xhosa) population of South Africa. *Hum Genet,* 103**,** 204-210.

Loh, P.-R., Lipson, M., Patterson, N., Moorjani, P., Pickrell, J. K., Reich, D. & Berger, B. 2013. Inferring Admixture Histories of Human Populations Using Linkage Disequilibrium. *Genetics,* 193**,** 1233-1254.

Londin, E. R., Keller, M. A., Maista, C., Smith, G., Mamounas, L. A., Zhang, R., Madore, S. J., Gwinn, K. & Corriveau, R. A. 2010. CoAIMs: A Cost-Effective Panel of Ancestry Informative Markers for Determining Continental Origins. *PLoS ONE,* 5**,** e13443.

Long, J., Chakravarti, A., Boehm, C., Antonarakis, S. & Kazazian, H. 1990. Phylogeny of human beta-globin haplotypes and its implications for recent human evolution. *Am J Phys Anthropol,* 81**,** 113-30.

Lou, C., Cong, B., Li, S., Fu, L., Zhang, X., Feng, T., Su, S., Ma, C., Yu, F., Ye, J. & Pei, L. 2011. A SNaPshot assay for genotyping 44 individual identification single nucleotide polymorphisms. *Electrophoresis,* 32**,** 368-78.

Lu, D., Willard, D., Patel, I. R., Kadwell, S., Overton, L., Kost, T., Luther, M., Chen, W., Woychik, R. P., Wilkison, W. O. & Cone, R. D. 1994. Agouti protein is an antagonist of the melanocyte-stimulating-hormone receptor. *Nature,* 371**,** 799-802.

Lucotte, G., Gerard, N., Krishnamoorthy, R. & David, F. 1994. Reduced variability in Y-chromosome-specific haplotypes for some central African populations. *Hum Biol,* 66**,** 519-26.

Lucotte, G. & Ngo, N. Y. 1985. p49f, A highly polymorphic probe, that detects Taq1 RFLPs on the human Y chromosome. *Nucleic Acids Res,* 13**,** 8285.

Ludwig, A., Rehberg, S. & Wegner, M. 2004. Melanocyte-specific expression of dopachrome tautomerase is dependent on synergistic gene activation by the Sox10 and Mitf transcription factors. *FEBS Lett,* 556**,** 236-244.

Luis, J., Rowold, D., Regueiro, M., Caeiro, B. & Cinnioglu, C. 2004. The Levant versus the Horn of Africa: evidence for bidirectional corridors of human migration. *Am J Hum Genet,* 74**,** 532-544.

Maca-Meyer, N., Gonzales, A. M., Larruga, J. M., Flores, C. & Cabrera, V. M. 2001. Major genomic mitochondrial lineages delineate early human expansions. *BMC Genet,* 2**,** 13.

Macaulay, V., Hill, C., Achilli, A., Chiara, R., Douglas, C. & Meehan, W. 2005. Single, rapid, coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. *Science,* 308**,** 1034-1036.

Maiquilla, S. M., Salvador, J. M., Calacal, G. C., Sagum, M. S., Dalet, M. R., Delfin, F. C., Tabbada, K. A., Franco, S. A., Perdigon, H. B., Madrid, B. J., Tan, M. P. & De Ungria, M. C. 2011. Y-STR DNA analysis of 154 female child sexual assault cases in the Philippines. *Int J Legal Med,* 125**,** 817-824.

Malsom, S., Flanagan, N., McAlister, C. & Dixon, L. 2009. The prevalence of mixed DNA profiles in fingernail samples taken from couples who co-habit using autosomal and Y-STRs. *Forensic Sci Int Genet,* 3**,** 57-62.

Malyarchuk, B., Derenko, M., Grzybowski, T., Lunkina, A., Czarny, J., Rychkov, S., Morozova, I., Denisova, G. & Miscicka-Sliwka, D. 2004. Differentiation of mitochondrial DNA and Y chromosomes in Russian populations. *Hum Biol,* 76**,** 877-900.

Malyarchuk, B. A., Derenko, M., Perkova, M., Grzybowski, T., Vanecek, T. & Lazur, J. 2008. Reconstructing the phylogeny of African mitochondrial DNA lineages in Slavs. *Eur J Hum Genet,* 16**,** 1091-1096.

Manel, S., Gaggiotti, O. E. & Waples, R. S. 2005. Assignment methods: matching biological questions with appropriate techniques. *Trends Ecol Evol,* 20**,** 136-42.

Marchington, D. R., Scott Brown, M. S. G., Lamb, V. K., van Golde, R. J. T., Kremer, J. A. M., Tuerlings, J. H. A. M., Mariman, E. C. M., Balen, A. H. & Poulton, J. 2002. No evidence for paternal mtDNA transmission to offspring or extra-embryonic tissues after ICSI. *Mol Hum Reprod,* 8**,** 1046-1049.

Markoulatos, P., Siafakas, N. & Moncany, M. 2002. Multiplex Polymerase Chain reaction: A practical approach. *J Clin Lab Anal,* 16**,** 47-51.

Maruyama, S., Minaguchi, K. & Saitou, N. 2003. Sequence polymorphisms of the mitochondrial DNA control region and phylogenetic analysis of mtDNA lineages in the Japanese populations. *Int J Legal Med,* 117**,** 218-225.

Masui, S., Nakatome, M. & Matoba, R. 2009. Variants of the Melanocortin 1 receptor gene (MC1R) and P gene would be effective for estimating the population origin of an individual. *Leg Med (Tokyo),* 11 S463-5.

McDougall, I., Brown, F. H. & Fleagle, J. G. 2005. Stratigraphic placement and age of modern humans from Kibish, Ethiopia. *Nature,* 433**,** 733-6.

Meenagh, A., Williams, F., Ross, O. A., Patterson, C., Gorodezky, C., Hammond, M., Leheny, W. A. & Middleton, D. 2002. Frequency of cytokine polymorphisms in populations from western Europe, Africa, Asia, the Middle East and South America. *Hum Immunol,* 63**,** 1055-1061.

Meirmans, P. G. 2006. Using the AMOVA framework to estimate a standardized genetic differentiation measure. *Evolution,* 60**,** 2399-402.

Mellars, P., Gori, K. C., Carr, M., Soares, P. A. & Richards, M. B. 2013. Genetic and archaeological perspectives on the initial modern human colonization of southern Asia. *Proc Natl Acad Sci USA,* 110**,** 10699-10704.

Mengel-From, J., Borsting, C., Sanchez, J. J., Eiberg, H. & Morling, N. 2010. Human eye colour and HERC2, OCA2 and MATP. *Forensic Sci Int Genet,* 4**,** 323-8.

Mengel-Jorgensen, J., Sanchez, J. J., Borsting, C., Kirpekar, F. & Morling, N. 2004. MALDI-TOF mass spectrometric detection of multiplex single base extended primers. A study of 17 y-chromosome single-nucleotide polymorphisms. *Anal Chem,* 76**,** 6039-45.

Merriweather, D., Clark, A. G., Ballinger, S., Schurr, T., Soodyall, H., Jenkins, T., Sherry, S. & Wallace, D. C. 1991. The structure of human mitochondrial DNA variation. *J Mol Evol,* 33**,** 543-555.

Merriwether, D. A., Hodgson, J. A., Friedlaender, F. R., Allaby, R., Cerchio, S., Koki, G. & Friedlaender, J. S. 2005. Ancient mitochondrial M haplogroups identified in the Southwest Pacific. *Proc Natl Acad Sci USA,* 102**,** 13034-13039.

Mesa, N. R., MondragÃ³n, M. a. C., Soto, I. n. D., Parra, M. a. V., Duque, C., OrtÃ-z-Barrientos, D., GarcÃ-a, L. F., Velez, I. n. D., Bravo, M. a. L., MÃºnera, J. G., Bedoya, G., Bortolini, M.-C. t. & Ruiz-Linares, A. s. 2000. Autosomal, mtDNA, and Y-Chromosome Diversity in Amerinds: Pre- and Post-Columbian Patterns of Gene Flow in South America. *Am J Hum Genet,* 67**,** 1277-1286.

Metspalu, M., Kivisild, T., Metspalu, E., Parik, J., Hudjashov, G., Kaldma, K. & Serk, P. 2004. Most of the extant mtDNA boundaries in South and Southwest Asia were likely shaped during the initial settlement of Eurasia by anatomically modern humans. *BMC Genet,* 5**,** 26.

Mikkelsen, M., Fendt, L., Rock, A. W., Zimmermann, B., Rockenbauer, E., Hansen, A. J., Parson, W. & Morling, N. 2012. Forensic and phylogeographic characterisation of mtDNA lineages from Somalia. *Int J Legal Med,* 126**,** 573-579.

Miller, F. J., Rosenfeldt, F. L., Zhang, C., Linnane, A. W. & Nagley, P. 2003. Precise determination of mitochondrial DNA copy number in human skeletal and cardiac muscle by a PCR-based assay: lack of change of copy number with age. *Nucleic Acids Res,* 31**,** e61.

Miller, L. H., Mason, S. J., Clyde, D. F. & McGinniss, M. H. 1976. The Resistance Factor to Plasmodium vivax in Blacks. *N Engl J Med,* 295**,** 302-304.

Miller, L. H., Mason, S. J., Dvorak, J. A., McGinniss, M. H. & Rothman, I. K. 1975. Erythrocyte receptors for (*Plasmodium knowlesi*) malaria: Duffy blood group determinants. *Science,* 189**,** 561-563.

Miller, R. 1981. *Simultaneous Statistical Inference,* 2nd ed*,* Springer Verlag, New York.

Mills, R. E., Luttig, C. T., Larkins, C. E., Beauchamp, A., Tsui, C., Pittard, W. S. & Devine, S. E. 2006. An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res,* 16**,** 1182-90.

Mok, C. C., Lanchbury, J. S., Wai Chan, D. & Lau, C. S. 1998. Interleukin-10 promoter polymorphisms in Southern Chinese patients with systemic lupus erythematosus. *Arthritis Rheum,* 41**,** 1090-1095.

Moorjani, P., Patterson, N., Hirschhorn, J. N., Keinan, A., Hao, L., Atzmon, G., Burns, E., Ostrer, H., Price, A. L. & Reich, D. 2011. The History of African Gene Flow into Southern Europeans, Levantines, and Jews. *PLoS Genet,* 7**,** e1001373.

Moraes, C. T., Srivastava, S., Kirkinezos, I., Oca-Cossio, J., vanWaveren, C., Woischnick, M. & Diaz, F. 2002. Mitochondrial DNA structure and function. *In:* SHAPIRA, A. (ed.) *Int Rev Neurobiol.* Academic Press.

Moretti, T., Baumstark, D., Defenbaugh, K., Keys, J., Smerick, J. & Budowle, B. 2001a. Validation of short tandem repeats (STRs) for forensic usage: performance testing of fluorescent multiplex STR systems and analysis of authentic and simulated forensic samples. *J Forensic Sci,* 46**,** 647-660.

Moretti, T. R., Baumstark, A. L., Defenbaugh, D. A., Keys, K. M., Brown, A. L. & Budowle, B. 2001b. Validation of STR typing by capillary electrophoresis. *J Forensic Sci,* 46**,** 661-76.

Morin, P. A., Luikart, G., Wayne, R. K. & the, S. N. P. w. g. 2004. SNPs in ecology, evolution and conservation. *Trends Ecol Evol,* 19**,** 208-216.

Mountain, J. & Cavalli-Sforza, L. L. 1994. Inference of human evolution through cladistic analysis of nuclear DNA restriction polymorphisms. *Proc Natl Acad Sci USA,* 91**,** 6515-9.

Mourant, A. E. & Domaniewska-Sobczak, K. 1962. The Use in Anthropology of Blood Groups and Other Genetical Characters. *J Afr Hist,* 3**,** 291-296.

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

Murdoch, G. 1981. *Atlas of World Cultures,* University Pittsburgh Press, Pittsburgh.

Myles, S., Somel, M., Tang, K., Kelso, J. & Stoneking, M. 2007. Identifying genes underlying skin pigmentation differences among human populations. *Hum Genet,* 120**,** 613-21.

Myres, N. M., Rootsi, S., Lin, A. A., Jarve, M., King, R. J., Kutuev, I., Cabrera, V. M., Khusnutdinova, E. K., Pshenichnov, A., Yunusbayev, B., Balanovsky, O.,

Balanovska, E., Rudan, P., Baldovic, M., Herrera, R. J., Chiaroni, J., Di Cristofaro, J., Villems, R., Kivisild, T. & Underhill, P. A. 2011. A major Y-chromosome haplogroup R1b Holocene era founder effect in Central and Western Europe. *Eur J Hum Genet,* 19**,** 95-101.

Nakayama, K., Fukamachi, S., Kimura, H., Koda, Y., Soemantri, A. & Ishida, T. 2002. Distinctive distribution of AIM1 polymorphism among major human populations with different skin color. *J Hum Genet,* 47**,** 92-94.

Nasidze, I., Ling, E., Quinque, D., Dupanloup, I., Cordaux, R. & Rychkov, S. 2004. Mitochondrial DNA and Y-chromosome variation in the caucasus. *Ann Hum Genet,* 68**,** 205-221.

Nasidze, I., Quinque, D., Dupanloup, I., Cordaux, R., Kokshunova, L. & Stoneking, M. 2005a. Genetic evidence for the Mongolian ancestry of Kalmyks. *Am J Phys Anthropol,* 128**,** 846-54.

Nasidze, I., Quinque, D., Ozturk, M., Bendukidze, N. & Stoneking, M. 2005b. MtDNA and Y-chromosome variation in Kurdish groups. *Ann Hum Genet,* 69**,** 401-12.

Nassir, R., Kosoy, R., Tian, C., White, P., Butler, L., Silva, G. & Kittles, R. 2009. An ancestry informative marker set for determining continental origin: validation and extension using human genome diversity panels. *BMC Genet,* 10**,** 39.

NCBI, 2013, *Database of Single Nucleotide Polymorphisms (dbSNP)*, dbSNP build ID:138, Bethesda, MD: National Center for Biotechnology Information, National Library of Medicine, available at http://www.ncbi.nlm.nih.gov/SNP/.

Neel, J. 1953. Data pertaining to the population dynamics of sickle cell disease. *Am J Hum Genet,* 5**,** 154-157.

Nei, M. 1987. *Molecular Evolutionary Genetics,* Columbia University Press, New York, NY, USA.

Nelson, T. M., Just, R. S., Loreille, O., Schanfield, M. S. & Podini, D. 2007. Development of a multiplex single base extension assay for mitochondrial DNA haplogroup typing. *Croat Med J,* 48**,** 460-72.

Neuvonen, A. M., Palo Ju Fau - Hedman, M., Hedman M Fau - Sajantila, A. & Sajantila, A. Discrimination power of Investigator DIPplex loci in Finnish and Somali populations.

Ngo, K., Vergnaud, G., Johnsson, C., Lucotte, G. & Weissenbach, J. 1986. A DNA probe detecting multiple haplotypes of the human Y chromosome. *Am J Hum Genet,* 38**,** 407-18.

Nickel, R. G., Willadsen, S. A., Freidhoff, L. R., Huang, S.-K., Caraballo, L., Naidu, R. P., Levett, P., Blumenthal, M., Banks-Schlegel, S., Bleecker, E., Beaty, T., Ober, C. & Barnes, K. C. 1999. Determination of Duffy genotypes in three populations of African descent using PCR and sequence-specific oligonucleotides. *Hum Immunol,* 60**,** 738-742.

Nievergelt, C. M., Maihofer, A. X., Shekhtman, T., Libiger, O., Wang, X., Kidd, K. K. & Kidd, J. R. 2013. Inference of human continental origin and admixture proportions using a highly discriminative ancestry informative 41-SNP panel. *Investig Genet,* 4**,** 13.

Nilsson, M., Andreasson-Jansson, H., Ingman, M. & Allen, M. 2008. Evaluation of mitochondrial DNA coding region assays for increased discrimination in forensic analysis. *Forensic Sci Int Genet,* 2**,** 1-8.

Nishizuka, S., Chen, S.-T., Gwadry, F. G., Alexander, J., Major, S. M., Scherf, U., Reinhold, W. C., Waltham, M., Charboneau, L., Young, L., Bussey, K. J., Kim, S., Lababidi, S., Lee, J. K., Pittaluga, S., Scudiero, D. A., Sausville, E. A., Munson, P. J., Petricoin, E. F., Liotta, L. A., Hewitt, S. M., Raffeld, M. &

Weinstein, J. N. 2003. Diagnostic Markers That Distinguish Colon and Ovarian Adenocarcinomas: Identification by Genomic, Proteomic, and Tissue Array Profiling. *Cancer Res,* 63**,** 5243-5250.

Non, A. L., Kitchen, A. & Mulligan, C. J. 2007. Identification of the most informative regions of the mitochondrial genome for phylogenetic and coalescent analyses. *Mol Phylogenet Evol,* 44**,** 1164-1171.

Oliveri, A., Achilli, A., Pala, M., Battaglia, V. & Formarino, S. 2006. The mtDNA legacy of the Levantine early Upper Paleolithic in Africa. *Science,* 314**,** 1767-1770.

Olivieri, A., Achilli, A., Pala, M., Battaglia, V., Fornarino, S., Al-Zahery, N., Scozzari, R., Cruciani, F., Behar, D. M., Dugoujon, J. M., Coudray, C., Santachiara-Benerecetti, A. S., Semino, O., Bandelt, H. J. & Torroni, A. 2006. The mtDNA legacy of the Levantine early Upper Palaeolithic in Africa. *Science,* 314**,** 1767-70.

Olsson, M. L., Hansson, C., Avent, N. D., Akesson, I. E., Green, C. A. & Daniels, G. L. 1998. A clinically applicable method for determining the three major alleles at the Duffy (FY) blood group locus using polymerase chain reaction with allele-specific primers. *Transfusion,* 38**,** 168-173.

Onofri, V., Alessandrini, F., Turchi, C., Pesaresi, M., Buscemi, L. & Tagliabracci, A. 2006. Development of multiplex PCRs for evolutionary and forensic applications of 37 human Y chromosome SNPs. *Forensic Sci Int,* 157**,** 23-35.

Oota, H., Settheetham-Ishida, W., Tiwawech, D., Ishida, T. & Stoneking, M. 2001. Human mtDNA and Y-chromosome variation is correlated with matrilocal versus patrilocal residence. *Nat Genet,* 29**,** 20-21.

Oriol, R., Danilovs, J. & Hawkins, B. R. 1981. A new genetic model proposing that the Se gene is a structural gene closely linked to the H gene. *Am J Hum Genet,* 33**,** 421-31.

Oriol, R., Le Pendu, J. & Mollicone, R. 1986. Genetics of ABO, H, Lewis, X and Related Antigens. *Vox Sang,* 51**,** 161-171.

Orloci, L. 1978. *Multivariate analysis in vegetation research,* Dr W. Junk B. V, The Hague, Netherlands.

Paetkau, D., Calvert, W., Stirling I & C, S. 1995. Microsatellite analysis of population structure in Canadian polar bears. *Mol Ecol,* 4**,** 347-354.

Paetkau, D., Slade, R., Burdens, M. & A, E. 2004. Genetic assignment methods for the direct, real-time estimation of migration rate: a simulation based exploration of accuracy and power. *Mol Ecol,* 13**,** 55-65.

Page, R. D. M. 1996. Tree View: An application to display phylogenetic trees on personal computers. *CABIOS,* 12**,** 357-358.

Pakendorf, B., Novgorodov, I. N., Osakovskij, V. L., Danilova, A. P., Protod'jakonov, A. P. & Stoneking, M. 2006. Investigating the effects of prehistoric migrations in Siberia: genetic variation and the origins of Yakuts. *Hum Genet,* 120**,** 334-53.

Pakstis, A., Speed, W., Fang, R., Hyland, F. L., Furtado, M., Kidd, J. & Kidd, K. 2010. SNPs for a universal individual identification panel. *Hum Genet,* 127**,** 315-324.

Palmer, J., Duffy, D., Box, N., Aitken, J., O'Gorman, L., Green, A. & Hayward, N. 2000. Melanocortin-1receptor polymorphisms and risk of melanoma: is the association explained solely by pigmentation phenotype? . *Am J Hum Genet,* 66**,** 176-86.

Paneto, G. G., Köhnemann, S., Martins, J. A., Cicarelli, R. M. B. & Pfeiffer, H. 2011. A single multiplex PCR and SNaPshot minisequencing reaction of 42 SNPs to

classify admixture populations into mitochondrial DNA haplogroups. *Mitochondrion,* 11**,** 296-302.

Pang, H., Koda, Y., Soejima, M., Fujitani, N., Ogaki, T. & Saito, A. 2001. Polymorphism of the human ABO-Secretor locus (FUT2) in four populations in Asia: indication of distinct Asian subpopulations. *Ann Hum Genet,* 65**,** 429-437.

Paracchini, S., Arredi, B., Chalk, R. & Tyler-Smith, C. 2002. Hierarchical high-throughput SNP genotyping of the human Y-chromosome using MALDI-TOF mass spectrometry. *Nucleic Acids Res,* 30**,** e27.

Parson, W. & Bandelt, H. J. 2007. Extended guidelines for mtDNA typing of population data in forensic science. *Forensic Sci Int Genet,* 1**,** 13-19.

Parsons, T. J., Huel, R., Davoren, J., Katzmarzyk, C., Miloš, A., Selmanović, A., Smajlović, L., Coble, M. D. & Rizvić, A. 2007. Application of novel "mini-amplicon" STR multiplexes to high volume casework on degraded skeletal remains. *Forensic Sci Int Genet,* 1**,** 175-179.

Paschou, P., Lewis, J., Javed, A. & Drineas, P. 2010. Ancestry informative markers for fine-scale individual assignment to worldwide populations. *J Med Genet,* 47**,** 835-847.

Paternoster, L., Zhurov, Alexei I., Toma, Arshed M., Kemp, John P., St. Pourcain, B., Timpson, Nicholas J., McMahon, G., McArdle, W., Ring, Susan M., Smith, George D., Richmond, S. & Evans, David M. 2012. Genome-wide Association Study of Three-Dimensional Facial Morphology Identifies a Variant in PAX3 Associated with Nasion Position. *Am J Hum Genet,* 90**,** 478-485.

Peakall, R. & Smouse, P. E. 2012. GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research—an update. *Bioinformatics,* 28**,** 2537-2539.

Penaloza, R. & Salamanca-Gomez, F. 1996. Anemia of falciform cells and the gentics of the mexican population. *Gac Med Mex,* 132**,** 337-8.

Pereira, R., Phillips, C., Alves, C., Amorim, A., Carracedo, A. & Gusmao, L. 2009. A new multiplex for human identification using insertion/deletion polymorphisms. *Electrophoresis,* 30**,** 3682-90.

Pereira, R., Phillips, C., Pinto, N., Santos, C., Santos, S. E. B. d., Amorim, A., Carracedo, Á. & Gusmão, L. 2012. Straightforward Inference of Ancestry and Admixture Proportions through Ancestry-Informative Insertion Deletion Multiplexing. *PLoS ONE,* 7**,** e29684.

Pericic, M., Lauc, L. B., Klaric, I. M., Janicijevic, B. & Rudan, P. 2005. Review of Croatian genetic heritage as revealed by mitochondrial DNA and Y chromosomal lineages. *Croat Med J,* 46**,** 502-513.

Petkovski, E., Keyser-Tracqui, C., Hienne, R. & Ludes, B. 2005. SNPs and MALDI-TOF MS: tools for DNA typing in forensic paternity testing and anthropology. *J Forensic Sci,* 50**,** 535-41.

Petraglia, M. D., Haslam, M., Fuller, D. Q., Boivin, N. & Clarkson, C. 2010. Out of Africa: new hypotheses and evidence for the dispersal of Homo sapiens along the Indian Ocean rim. *Ann Hum Biol,* 37**,** 288-311.

Phillips, C., Aradas, A. F., Kriegel, A. K., Fondevila, M., Bulbul, O., Santos, C., Rech, F. S., Carceles, M. D., Carracedo, A., Schneider, P. M. & Lareu, M. V. 2013. Eurasiaplex: A forensic SNP assay for differentiating European and South Asian ancestries. *Forensic Sci Int Genet,* 7**,** 359-66.

Phillips, C., Fang, R., Ballard, D., Fondevila, M., Harrison, C. & Hyland, F. 2007a. Evaluation of the Genplex SNP typing system and a 49plex forensic marker panel. *Forensic Sci Int Genet,* 1**,** 180-185.

Phillips, C., Fondevila, M. & Lareau, M. V. 2012. A 34-plex autosomal SNP single base extension assay for ancestry investigations. *In:* ALONSO, A. (ed.) *DNA Electrophoresis Protocols for Forensic Genetics*. Humana Press.

Phillips, C., Lareu, M., Salas, A., Fondevila, M., Berniell Lee, G., Carracedo, A., Morling, N., Schneider, P. & Syndercombe Court, D. 2004. Population specific single nucleotide polymorphisms. *Int Congr Ser,* 1261**,** 233-235.

Phillips, C., Prieto, L., Fondevila, M., Salas, A., Gomez-Tato, A., Alvarez-Dios, J., Alonso, A., Blanco-Verea, A., Brion, M., Montesino, M., Carracedo, A. & Lareu, M. V. 2009. Ancestry analysis in the 11-M Madrid bomb attack investigation. *PLoS ONE,* 4**,** e6583.

Phillips, C., Salas, A., Sanchez, J. J., Fondevila, M. & Gomez-Tato, A. 2007b. Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. *Forensic Sci Int Genet,* 1**,** 273-280.

Piry, S., Alapetite, A., Cornuet, J., Paetkau, D. & Estoup, A. 2004. GeneClass2: A software for genetic assignment and first geenration migrant detection. *J Hered,* 95**,** 536-9.

Plaza, S., Calafell, F., Helal, A., Bouzerna, N. & Lefranc, G. 2003. Joining the pillars of Hercules: mtDNA sequences show multidirectional gene flow in the Western Mediterranean. *Ann Hum Genet,* 67**,** 312-328.

Pneuman, A., Budimlija, Z. M., Caragine, T., Prinz, M. & Wurmbach, E. 2012. Verification of eye and skin color predictors in various populations. *Legal Med (Tokyo),* 14**,** 78-83.

Pospiech, E., Draus-Barini, J., Kupiec, T., Wojas-Pelc, A. & Branicki, W. 2012. Prediction of eye color from genetic data using Bayesian approach. *J Forensic Sci,* 57**,** 880-6.

Prinz, M., Ishii, A., Coleman, A., Baum, H. & Shaler, R. 2001. Validation and casework application of a Y chromosome specific STR multiplex. *Forensic Sci Int,* 120**,** 177-88.

Prinz, M. & Sansone, M. 2001. Y chromosome-specific short tandem repeats in forensic casework. *Croat Med J,* 42**,** 288-91.

Pritchard, J., Seielstad, M. T., Perez-Lezaun, A. & Feldman, M. 1999. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol Biol Evol,* 16**,** 1791-8.

Pritchard, J., Stephens, M. & Donnelly, P. 2000. Inference of population structure using multilocus genotype data. *Genetics,* 155**,** 945-59.

Quintana-Murci, L., Semino, O., Bandelt, H. J., Passarino, M., McElreavey & Santachiara-Benerecetti, S. 1999. Genetic evidence of an early exit of homo sapiens sapiens from Africa through eastern Africa. *Nat Genet,* 23**,** 437-441.

Quintans, B., Alvarez-Iglesias, V., Salas, A., Phillips, C., Lareu, M. & Carracedo, A. 2004. Typing of mitochondrial DNA coding region SNPs of forensic and anthropological interest using SNaPshot minisequencing. *Forensic Sci Int,* 140**,** 251-257.

Ramana, G. V., Su, B., Jin, L., Singh, L., Wang, N., Underhill, P. & Chakraborty, R. 2001. Y-chromosome SNP haplotypes suggest evidence of gene flow among caste, tribe, and the migrant Siddi populations of Andhra Pradesh, South India. *Eur J Hum Genet,* 9**,** 695-700.

Rambaut, A. 2012. FigTree version 1.4: Tree Figure drawing tool.

Rana, B., Hewett-Emmett, D., Jin, L., Chang, B., Sambuughin, N., Lin, M., Watkins, S., Bamshad, M. & Jorde, L. 1999. High polymorphism at the human melanocortin 1 receptor locus. *Genetics,* 151**,** 1547-57.

Rannala, B. & Mountain, J. 1997. Detecting immigration by using multilocus genotypes. *Proc Natl Acad Sci USA,* 94**,** 9197-9291.

Rappold, G. 1993. The pseudoautosomal regions of the human sex chromosomes. *Hum Genet,* 92**,** 315-324.

Ray, D., Walker, J., Hall, A., Llewellyn, B., Ballantyne, J., Christian, A., Turtletaub, K. & Batzer, M. 2005. Inference of human origins using *Alu* insertion polymorphisms. *Forensic Sci Int,* 153**,** 117-124.

Rebbeck, T., Kanetsky, P., Walker, A., Holmes, R. & Halpern, A. 2002. P gene as an inherited biomarker of human eye color. *Cancer Epidemiol Biomarkers Prev,* 11**,** 782-784.

Richards, M. 1998. Phylogeography of mitochondrial DNA in Western Europe. *Ann Hum Genet,* 62**,** 241-260.

Richards, M. & Macaulay, V. 2001. The mitochondrial gene tree comes of age. *Am J Hum Genet,* 68**,** 1315-1320.

Richards, M., Macaulay, V., Hickey, E., Vega, E. & Sykes, B. 2000. Tracing European founder lineages in the Near Eastern mtDNA pool. *Am J Hum Genet,* 67**,** 1251-1276.

Rivera-Chavez, F. A., Peters-Hybki, D. L., Barber, R. C. & O'Keefe, G. E. 2003. Interleukin-6 promoter haplotypes and interleukin-6 cytokine responses. *Shock,* 20**,** 218-23.

Roewer, L., Arnemann, J., Spurr, N., Grzeschik, K. & Epplen, J. 1992. Simple repeat sequences on the human Y chromosome are equally polymorphic as their autosomal counterparts. *Hum Genet,* 89**,** 389-94.

Rosa, A., Ornelas, C., Jobling, M., Brehm, A. & Villems, R. 2007. Y-chromosomal diversity in the population of Guinea-Bissau: a multiethnic perspective. *BMC Evol Biol,* 7**,** 124.

Rosenberg, N., Li, L., Ward, R. & Pritchard, J. 2003. Informativeness of genetic markers for inference of ancestry. *Am J Hum Genet,* 73**,** 1402-1422.

Rosenberg, N. A. 2004. distruct: a program for the graphical display of population structure. *Mol Ecol Notes,* 4**,** 137-138.

Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A. & Feldman, M. W. 2002. Genetic structure of human populations. *Science,* 298**,** 2381-5.

Rowold, D., Luis, J., Terreros, M. & Herrera, R. 2007. Mitochondrial DNA geneflow indicates preferred usage of the Levant corridor over the Horn of Africa passageway. *Hum Genet,* 52**,** 436-447.

Royal, C. & Dunston, G. 2004. Changing the paradigm from 'race' to human genome variation. *Nat Genet,* 36**,** S5-S7.

Royal, C. D., Novembre, J., Fullerton, S. M., Goldstein, D. B., Long, J. C., Bamshad, M. J. & Clark, A. G. 2010. Inferring Genetic Ancestry: Opportunities, Challenges, and Implications. *Am J Hum Genet,* 86**,** 661-673.

Ruiz, Y., Phillips, C., Gomez-Tato, A., Alvarez-Dios, J., Casares de Cal, M., Cruz, R., Maroñas, O., Söchtig, J., Fondevila, M., Rodriguez-Cid, M. J., Carracedo, Á. & Lareu, M. V. 2013. Further development of forensic eye color predictive tests. *Forensic Sci Int Genet,* 7**,** 28-40.

Ruiz-Martinez, M. C., Salas-Solano, O., Carrilho, E., Kotler, L. & Karger, B. L. 1998. A Sample Purification Method for Rugged and High-Performance DNA Sequencing by Capillary Electrophoresis Using Replaceable Polymer Solutions. A. Development of the Cleanup Protocol. *Anal Chem,* 70**,** 1516-1527.

Ruiz-Pesini, E., Lott, M. T., Procaccio, V., Poole, J., Brandon, M., Mishmar, D., Yi, C. & Kreuziger, J. 2007. An enhanced MITOMAP with a global mtDNA mutational phylogeny. *Nucleic Acids Res,* 35**,** D823-D828.

Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Marth, G., Sherry, S., Mullikin, J. C., Mortimore, B. J., Willey, D. L., Hunt, S. E., Cole, C. G., Coggill, P. C., Rice, C. M., Ning, Z., Rogers, J., Bentley, D. R., Kwok, P. Y., Mardis, E. R., Yeh, R. T., Schultz, B., Cook, L., Davenport, R., Dante, M., Fulton, L., Hillier, L., Waterston, R. H., McPherson, J. D., Gilman, B., Schaffner, S., Van Etten, W. J., Reich, D., Higgins, J., Daly, M. J., Blumenstiel, B., Baldwin, J., Stange-Thomann, N., Zody, M. C., Linton, L., Lander, E. S. & Altshuler, D. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature,* 409**,** 928-33.

Sahoo, S., Singh, A., Himabindu, G., Banerjee, J., Sitalaximi, T., Gaikwad, S., Trivedi, R., Endicott, P., Kivisild, T., Metspalu, M., Villems, R. & Kashyap, V. K. 2006. A prehistory of Indian Y chromosomes: Evaluating demic diffusion scenarios. *Proc Natl Acad Sci USA,* 103**,** 843-848.

Salari, K., Choudhry, S., Tang, H., Naqvi, M., Lind, D., Avila, P. C., Coyle, N. E., Ung, N., Nazario, S., Casal, J., Torres-Palacios, A., Clark, S., Phong, A., Gomez, I., Matallana, H., Pérez-Stable, E. J., Shriver, M. D., Kwok, P.-Y., Sheppard, D., Rodriguez-Cintron, W., Risch, N. J., Burchard, E. G. & Ziv, E. 2005. Genetic admixture and asthma-related phenotypes in Mexican American and Puerto Rican asthmatics. *Genet Epidemiol,* 29**,** 76-86.

Salas, A., Bandelt, H.-J., Macaulay, V. & Richards, M. 2007. Phylogeographic investigations: The role of trees in forensic genetics. *Forensic Sci Int,* 168**,** 1-13.

Salas, A., Richards, M., De la Fe, T., Lareu, M. & Sobrino, B. 2002. The making of the African mtDNA landscape. *Am J Hum Genet,* 71**,** 1082-1111.

Salas-Solano, O., Ruiz-Martinez, M. C., Carrilho, E., Kotler, L. & Karger, B. L. 1998. A Sample Purification Method for Rugged and High-Performance DNA Sequencing by Capillary Electrophoresis Using Replaceable Polymer Solutions. B. Quantitative Determination of the Role of Sample Matrix Components on Sequencing Analysis. *Anal Chem,* 70**,** 1528-1535.

Salem, A., Badr, F., Gaballah, M. & Paabo, S. 1996. The genetics of traditional living: Y-chromosomal and mitochondrial lineages. *Am J Hum Genet,* 59**,** 741-743.

Sanchez, J. J., Borsting, C., Hallenberg, C., Buchard, A., Hernandez, A. & Morling, N. 2003. Multiplex PCR and minisequencing of SNPs- a model with 35 Y chromosome SNPs. *Forensic Sci Int,* 137**,** 74-84.

Sanchez, J. J., Borsting, C. & Morling, N. 2005. Typing of Y chromosome SNPs with multiplex PCR methods. *Methods Mol Biol,* 297**,** 209-228.

Sanchez, J. J., Phillips, C., Borsting, C., Balogh, K., Bogus, M. & Fondevila, M. 2006. A multiplex assay with 52 single nucleotide polymorphisms for human identification. *Electrophoresis,* 27**,** 1713-1724.

Sanger, R., Race, R. R. & Jack, J. 1955. The Duffy Blood Groups of New York Negroes: The Phenotype Fy (a−b−). *Brit J Haematol,* 1**,** 370-374.

Sankararaman, S., Patterson, N., Li, H., Paabo, S. & Reich, D. 2012. The date of interbreeding between Neandertals and modern humans. *PLoS Genet,* 8**,** e1002947.

Sankararaman, S., Sridhar, S., Kimmel, G. & Halperin, E. 2008. Estimating local ancestry in admixed populations. *Am J Hum Genet,* 82**,** 290-303.

Schaaf, C. P., Wiszniewska, J. & Beaudet, A. L. 2011. Copy Number and SNP Arrays in Clinical Diagnostics. *Annu Rev Genomics Hum Genet,* 12**,** 25-51.

Scheinfeldt, L., Friedlaender, F., Friedlaender, J., Latham, K., Koki, G., Karafet, T., Hammer, M. & Lorenz, J. 2006. Unexpected NRY Chromosome Variation in Northern Island Melanesia. *Mol Biol Evol,* 23**,** 1628-1641.

Schlebusch, C. M., Naidoo, T. & Soodyall, H. 2009. SNaPshot minisequencing to resolve mitochondrial macro-haplogroups found in Africa. *Electrophoresis,* 30**,** 3657-3664.

Schwartz, H. E., Ulfelder, K., Sunzeri, F. J., Busch, M. P. & Brownlee, R. G. 1991. Analysis of DNA restriction fragments and polymerase chain reaction products towards detection of the AIDS (HIV-1) virus in blood. *J Chromatogr A,* 559**,** 267-283.

Seielstad, M. T., Hebert, J., Lin, A., Underhill, P., Ibrahim, M., Vollrath, D. & Cavalli-Sforza, L. L. 1994. Construction of human Y-chromosomal haplotypes using a new polymorphic A to G transition. *Hum Mol Genet,* 3**,** 2159-61.

Seielstad, M. T., Minch, E. & Cavalli-Sforza, L. L. 1998. Genetic evidence for a higher female migration rate in humans. *Nat Genet,* 20**,** 278-280.

Sengupta, S., Zhivotovsky, L. A., King, R., Mehdi, S. Q., Edmonds, C. A., Chow, C. E., Lin, A. A., Mitra, M., Sil, S. K., Ramesh, A., Usha Rani, M. V., Thakur, C. M., Cavalli-Sforza, L. L., Majumder, P. P. & Underhill, P. A. 2006. Polarity and temporality of high-resolution y-chromosome distributions in India identify both indigenous and exogenous expansions and reveal minor genetic influence of Central Asian pastoralists. *Am J Hum Genet,* 78**,** 202-21.

Sharma, G., Tamang, R., Chaudhary, R., Singh, V. K., Shah, A. M., Anugula, S., Rani, D. S., Reddy, A. G., Eaaswarkhanth, M., Chaubey, G., Singh, L. & Thangaraj, K. 2012. Genetic Affinities of the Central Indian Tribal Populations. *PLoS ONE,* 7**,** e32546.

Shen, P., Lavi, T., Kivisild, T., Chou, V. & Sengun, D. 2004. Reconstruction of patrilineages and matrilineages of Samaritans and other Israeli populations from Y-chromosome and mitochondrial DNA sequence variation. *Hum Mutat,* 24**,** 248-260.

Shen, P., Wang, F., Underhill, P., Franco, C., Yang, W., Roxas, A. & Sung, R. 2000. Population genetic implications from sequence variation in four Y chromosome genes. *Proc Natl Acad Sci USA,* 97**,** 7354-9.

Shen, Z., Qu, W., Wang, W., Lu, Y., Wu, Y., Li, Z., Hang, X., Wang, X., Zhao, D. & Zhang, C. 2010. MPprimer: a program for reliable multiplex PCR primer design. *BMC Bioinformatics,* 11**,** 143.

Sherry, S., Ward, M., Kholodov, M., Baker, J., Phan, L., Smigielski, E. & Sirotkin, K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res,* 29**,** 308-11.

Shriver, M., Parra, E., Bonilla, C., Norton, H., Jovel, C. & Pfaff, C. 2003. Skin pigmentation, biogeographical ancestry and admixture mapping. *Hum Genet,* 112**,** 387-399.

Shriver, M. D., Smith, M. W., Jin, L., Marcini, A., Akey, J. M., Deka, R. & Ferrell, R. E. 1997. Ethnic-affiliation estimation by use of population-specific DNA markers. *Am J Hum Genet,* 60**,** 957-64.

Sibille, I., Duverneuil, C., Lorin de la Grandmaison, G., Guerrouache, K., Teissière, F., Durigon, M. & de Mazancourt, P. 2002. Y-STR DNA amplification as biological evidence in sexually assaulted female victims with no cytological detection of spermatozoa. *Forensic Sci Int,* 125**,** 212-216.

Sigurdardottir, S., Helgason, A., Gulcher, J., Stefansson, K. & Donnelly, P. 2000. The mutation rate in the human mtDNA control region. *Am J Hum Genet,* 66**,** 1599-1609.

Sigurdsson, S., Hedman, M., Sistonen, P., Sajantila, A. & Syvanen, A.-C. 2006. A microarray system for genotyping 150 single nucleotide polymorphisms in the coding region of human mitochondrial DNA. *Genomics,* 87**,** 534-542.

Silva, J., Bonatto, S., Holanda, A., Ribeiro-Dos-Santos, A., Paixao, B., Goldman, G. & Abe-Sandes, K. 2002. Mitochondrial genome diversity of Native Americans supports a single early entry of founder populations into America. *Am J Hum Genet,* 71**,** 187-192.

Skaletsky, H., Kuroda-Kawaguchi, T., Minx, P., Cordum, H., Hillier, L., Brown, L. & Repping, S. 2003. The male- specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature,* 423**,** 825-37.

Slatkin, M. 2008. Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. *Nat Rev Genet,* 9**,** 477-485.

Smith, M., Patterson, N., Lautenberger, J., Truelove, A., McDonald, G. & Waliszewska, A. 2004. A high-density admixture map for disease gene discovery in african americans. *Am J Hum Genet,* 74**,** 1001-13.

Smith, S. B., Zhou, B. K. & Orlow, S. J. 1998. Expression of tyrosinase and the tyrosinase related proteins in the Mitfvit (vitiligo) mouse eye: implications for the function of the microphthalmia transcription factor. *Exp Eye Res,* 66**,** 403-10.

Sobrino, B., Brion, M. & Carracedo, A. 2005. SNPs in forensic genetics: a review on SNP typing methodologies. *Forensic Sci Int,* 154**,** 181-194.

Soejima, M. & Koda, Y. 2007. Population differences of two coding SNPs in pigmentation-related genes SLC24A5 and SLC45A2. *Int J Legal Med,* 121**,** 36-39.

Soejima, M., Pang, H. & Koda, Y. 2007. Genetic variation of FUT2 in a Ghanaian population: identification of four novel mutations and inference of balancing selection. *Annals of Hematology,* 86**,** 199-204.

Spichenok, O., Budimlija, Z. M., Mitchell, A. A., Jenny, A., Kovacevic, L., Marjanovic, D., Caragine, T., Prinz, M. & Wurmbach, E. 2011. Prediction of eye and skin color in diverse populations using seven SNPs. *Forensic Sci Int Genet,* 5**,** 472-8.

Stewart, C. 1993. The power and pitfalls of parsimony. *Nature,* 361**,** 603-607.

Stringer, C. 2002. Modern Human origins: progress and prospects. *Philos Trans R Soc B,* 357**,** 563-579.

Sturm, R. 2006. A golden age of human pigmentation genetics. *Trends Genet,* 22**,** 464-8.

Sturm, R., Box, N. & Ramsay, M. 1998. Human pigmentation genetics: the difference is only skin deep. *Bioessays,* 20**,** 712-21.

Sturm, R., Duffy, D., Zhao, Z., Leite, F. & Stark, M. 2008. A single SNP in an evolutionary conserved region within intron 86 of the HERC2 gene determines human blue-brown eye color. *Am J Hum Genet,* 82**,** 424-31.

Sturm, R., Teasdale, R. & Box, N. 2001. Human pigmentation genes: identification, structure and consequences of polymorphic variation. *Gene,* 277**,** 49-62.

Sturm, R. A. & Frudakis, T. N. 2004. Eye colour: portals into pigmentation genes and ancestry. *Trends Genet,* 20**,** 327-332.

Sturm, R. A. & Larsson, M. 2009. Genetics of human iris colour and patterns. *Pigment Cell Melanoma Res,* 22**,** 544-62.

Sudoyo, H., Widodo, P. T., Suryadi, H., Lie, Y. S., Safari, D., Widjajanto, A. & Kadarmo, D. A. 2008. DNA analysis in perpetrator identification of terrorism-

related disaster: Suicide bombing of the Australian Embassy in Jakarta 2004. *Forensic Sci Int Genet,* 2**,** 231-237.

Sulem, P., Gudbjartsson, D., Stacey, S., Helgason, A. & Rafnar, T. 2007. Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nat Genet,* 39**,** 1443-5.

Sulem, P., Gudbjartsson, D. F., Stacey, S. N., Helgason, A., Rafnar, T., Jakobsdottir, M., Steinberg, S., Gudjonsson, S. A., Palsson, A., Thorleifsson, G., Palsson, S., Sigurgeirsson, B., Thorisdottir, K., Ragnarsson, R., Benediktsdottir, K. R., Aben, K. K., Vermeulen, S. H., Goldstein, A. M., Tucker, M. A., Kiemeney, L. A., Olafsson, J. H., Gulcher, J., Kong, A., Thorsteinsdottir, U. & Stefansson, K. 2008. Two newly identified genetic determinants of pigmentation in Europeans. *Nat Genet,* 40**,** 835-837.

Sun, C., Kong, Q.-P., Palanichamy, M., Agrawal, S., Bandelt, H. J. & Yao, Y.-G. 2006. The dazzling array of basal branches in mtDNA macrohaplogroups M from India as inferred from complete genomes. *Mol Biol Evol,* 23**,** 683-690.

Sun, X. & Guo, B. 2006. Genotyping single-nucleotide polymorphisms by matrix-assisted laser desorption/ionization time-of-flight-based mini-sequencing. *Methods Mol Med,* 128**,** 225-30.

Suzuki, I., Tada, A., Ollmann, M. M., Barsh, G. S., Im, S., Lamoreux, M. L., Hearing, V. J., Nordlund, J. J. & Abdel-Malek, Z. A. 1997. Agouti signaling protein inhibits melanogenesis and the response of human melanocytes to alpha-melanotropin. *J Investig Dermatol,* 108**,** 838-842.

Tajima, F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics,* 105**,** 437-460.

Tajima, F. 1993. Measurement of DNA polymorphism. *In:* TAKAHATA, N. & CLARK, A. G. (eds.) *Mechanisms of Molecular Evolution. Introduction to molecular Paleopopulation biology.* Tokyo, Sunderland: MA: Japan Scientific Societies Press, Sinauer Associates, Inc.

Tanaka, M., Cabrera, V. M., Gonzales, A. M., Larruga, J. M., Takeyasu, T., Fuku, N. & Guo, L. 2004. Mitochondrial genome variation in eastern Asia and the peopling of Japan. *Genome Res,* 14**,** 1832-1850.

Tang, H., Siegmund, D., Shen, P., Oefner, P. & Feldman, M. 2002. Frequentist estimation of coalescence times from nucleotide Sequence data using a treebased partition. *Genetics,* 161**,** 447-59.

Templeton, A. 2002. Out of Africa again and again. *Nature,* 416**,** 45-51.

Templeton, J., Brotherton, P., Llamas, B., Soubrier, J., Haak, W., Cooper, A. & Austin, J. 2013. DNA capture and next-generation sequencing can recover whole mitochondrial genomes from highly degraded samples for human identification. *Investig Genet,* 4**,** 1-13.

Teschendorff, A. E., Menon, U., Gentry-Maharaj, A., Ramus, S. J., Weisenberger, D. J., Shen, H., Campan, M., Noushmehr, H., Bell, C. G., Maxwell, A. P., Savage, D. A., Mueller-Holzner, E., Marth, C., Kocjan, G., Gayther, S. A., Jones, A., Beck, S., Wagner, W., Laird, P. W., Jacobs, I. J. & Widschwendter, M. 2010. Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome Res,* 20**,** 440-6.

The International HapMap Consortium 2003. The International HapMap Project. *Nature,* 426**,** 789-796.

The International HapMap Consortium 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature,* 449**,** 851-861.

The International HapMap Consortium 2010. Integrating common and rare genetic variation in diverse human populations. *Nature,* 467**,** 52-58.

The Y Chromosome Consortium 2002. A nomenclature system for the tree of human Y-chromosomal binary haplogroups. *Genome Res,* 12**,** 339-348.

Thyagarajan, B., Bower, M., Berger, M., Jones, S., Dolan, M. & Wang, X. 2008. A novel polymorphism in the FMR1 gene: implications for clinical testing of fragile X syndrome. *Arch Pathol Lab Med,* 132**,** 95-98.

Thyagarajan, B., Padua, R. & Campbell, C. 1996. Mammalian mitochondria possess homologous DNA recombination activity. *J Biol Chem,* 271**,** 27536-43.

Tomas, C., Stangegaard, M., Borsting, C., Hansen, A. & Morling, N. 2008. Typing of 48 autosomal SNPs and amelogenin with GenPlex SNP genotyping system in forensic genetics. *Forensic Sci Int Genet,* 3**,** 1-6.

Tomlinson, I., Webb, E., Carvajal-Carmona, L., Broderick, P., Kemp, Z., Spain, S., Penegar, S., Chandler, I., Gorman, M., Wood, W., Barclay, E., Lubbe, S., Martin, L., Sellick, G., Jaeger, E., Hubner, R., Wild, R., Rowan, A., Fielding, S., Howarth, K., Silver, A., Atkin, W., Muir, K., Logan, R., Kerr, D., Johnstone, E., Sieber, O., Gray, R., Thomas, H., Peto, J., Cazier, J. B. & Houlston, R. 2007. A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nat Genet,* 39**,** 984-8.

Torroni, A., Achilli, A., Macaulay, V., Richards, M. & Bandelt, H. J. 2006. Harvesting the fruit of the human mtDNA tree. *Trends Genet,* 22**,** 339-345.

Torroni, A., D'Urbano, L., Rengo, C., Scozzari, R., Sbracia, M., Manna, C., Cavazzini, C. & Sellito, D. 1998. Intracytoplasmic injection of spermatozoa does not appear to alter the mode of mitochondrial inheritance. *Hum Reprod,* 13**,** 1747-1749.

Torroni, A., Huoponen, K., Francalacci, P., Petrozzi, M., Morelli, L. & Scozzari, R. 1996. Classification of European mtDNAs from an analysis of three European populations. *Genetics,* 144**,** 1835-1850.

Torroni, A., Schurr, T., Cabell, M., Brown, M., Neel, J., Larsen, M., Smith, D., Vullo, C. & Wallace, D. 1993. Asian affinities and continental radiation of the four founding Native American mtDNAs. *Am J Hum Genet,* 53**,** 563-590.

Torroni, A. & Wallace, D. C. 1994. Mitochondrial DNA variation in human populations and implications for detection of mitochondrial DNA mutations of pathological significance. *J Bioenerg Biomembr,* 26**,** 261-270.

Tournamille, C., Colin, Y., Cartron, J. P. & Le Van Kim, C. 1995. Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy-negative individuals. *Nat Genet,* 10**,** 224-8.

Toyofuku, K., Valencia, J. C., Kushimoto, T., Costin, G. E., Virador, V. M., Vieira, W. D., Ferrans, V. J. & Hearing, V. J. 2002. The etiology of oculocutaneous albinism (OCA) type II: the pink protein modulates the processing and transport of tyrosinase. *Pigment Cell Res,* 15**,** 217-24.

Trejaut, J. A., Kivisild, T., Loo, J. H., Lee, C. L., He, C. L., Hsu, C. J., Li, Z. Y. & Lin, M. 2005. Traces of Archaic Mitochondrial Lineages Persist in Austronesian-Speaking Formosan Populations. *PLoS Biol,* 3**,** e247.

Tully, G. 2007. Genotype versus phenotype: Human pigmentation. *Forensic Sci Int Genet,* 1**,** 105-110.

Tzoulaki, I., Murray, G. D., Lee, A. J., Rumley, A., Lowe, G. D. O. & Fowkes, F. G. R. 2005. C-Reactive Protein, Interleukin-6, and Soluble Adhesion Molecules as Predictors of Progressive Peripheral Atherosclerosis in the General Population: Edinburgh Artery Study. *Circulation,* 112**,** 976-983.

Underhill, P., Jin, L., Lin, A., Mehdi, S., Jenkins, T. & Vollrath, D. 1997. Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. *Genome Res,* 7**,** 996-1005.

Underhill, P., Passarino, G., Lin, A., Shen, P., Mirazon Lahr, M., Foley, R. & Oefner, P. 2001. The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. *Ann Hum Genet,* 65**,** 43-62.

Underhill, P., Shei, P., Lin, A., Jin, L., Passarino, G., Yang, W. & Kauffman, E. 2000. Y chromosome sequence variation and the history of human populations. *Nat Genet,* 26**,** 358-361.

Underhill, P. A. & Kivisild, T. 2007. Use of Y chromosome and mitochondrial DNA population structure in tracing human migrations. *Annu Rev Genet,* 41**,** 539-64.

Valenzuela, R. K., Henderson, M. S., Walsh, M. H., Garrison, N. A., Kelch, J. T., Cohen-Barak, O., Erickson, D. T., John Meaney, F., Bruce Walsh, J., Cheng, K. C., Ito, S., Wakamatsu, K., Frudakis, T., Thomas, M. & Brilliant, M. H. 2010. Predicting Phenotype from Genotype: Normal Pigmentation. *J Forensic Sci,* 55**,** 315-22.

Vallone, P. M. & Butler, J. M. 2004. AutoDimer: A screening tool for primer- dimer and hairpin structures. *BioTechniques,* 37**,** 226-231.

Vallone, P. M., Just, R. S., Coble, M. D., Butler, J. M. & Parsons, T. J. 2004. A multiplex allele-specific primer extension assay for forensically informative SNPs distributed throughout the mitochondrial genome. *Int J Legal Med,* 118**,** 147-157.

Valverde, P., Healy, E., Jackson, I., Rees, J. & Thody, A. 1995. Variants of the melanocyte-stimulating hormone receptor gene are associated with red hair and fair skin in humans. *Nat Genet,* 11**,** 328-30.

van Oven, M. & Kayser, M. 2009. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum Mutat,* 30**,** E386-E394.

van Oven, M., Ralf, A. & Kayser, M. 2011a. An efficient multiplex genotyping approach for detecting the major worldwide human Y-chromosome haplogroups. *Int J Legal Med,* 125**,** 879-85.

van Oven, M., Vermeulen, M. & Kayser, M. 2011b. Multiplex genotyping system for efficient inference of matrilineal genetic ancestry with continental resolution. *Investig Genet,* 2**,** 6.

Vernot, B. & Akey, J. M. 2014. Resurrecting Surviving Neandertal Lineages from Modern Human Genomes. *Science,* 343**,** 1017-1021.

Visser, M., Kayser, M. & Palstra, R. J. 2012. HERC2 rs12913832 modulates human pigmentation by attenuating chromatin-loop formation between a long-range enhancer and the OCA2 promoter. *Genome Res,* 22**,** 446-55.

Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. 2006. A Map of Recent Positive Selection in the Human Genome. *PLoS Biol,* 4**,** e72.

Von Wurmb- Schwark, N., Preusse-Prange, A., Heinrich, A., Simeoni, E., Bosch, T. & Schwark, T. 2009. A new multiplex-PCR comprising autosomal and Y-specific STRs and mitochondrial DNA to analyze highly degraded material. *Forensic Sci Int Genet,* 3**,** 96-103.

Wainscoat, J., Hill, A., Boyce, A., Flint, J., Hernandez, M., Thein, S. & Old, J. 1986. Evolutionary relationships of human populations from an analysis of nuclear DNA polymorphisms. *Nature,* 319**,** 491-3.

Wall, J. D., Lohmueller, K. E. & Plagnol, V. 2009. Detecting Ancient Admixture and Estimating Demographic Parameters in Multiple Human Populations. *Mol Biol Evol,* 26**,** 1823-1827.

Wall, J. D., Yang, M. A., Jay, F., Kim, S. K., Durand, E. Y., Stevison, L. S., Gignoux, C., Woerner, A., Hammer, M. F. & Slatkin, M. 2013. Higher Levels of Neanderthal Ancestry in East Asians than in Europeans. *Genetics,* 194**,** 199-209.

Walsh, S., Liu, F., Ballantyne, K. N., van Oven, M., Lao, O. & Kayser, M. 2011. IrisPlex: a sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information. *Forensic Sci Int Genet,* 5**,** 170-80.

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

Wang, D. Y., Chang, C.-W., Lagacé, R. E., Calandro, L. M. & Hennessy, L. K. 2012. Developmental Validation of the AmpFℓSTR® Identifiler® Plus PCR Amplification Kit: An Established Multiplex Assay with Improved Performance. *J Forensic Sci,* 57**,** 453-465.

Waples, R. S. & Gaggiotti, O. 2006. What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. *Mol Ecol,* 15**,** 1419-1439.

Watkins, W., Thara, R., Mowry, B., Zhang, Y., Witherspoon, D. & Tolpinrud, W. 2008. Genetic variation in South Indian castes: evidence from Y-chromosome, mitochondrial, and autosomal polymorphisms. *BMC Genet,* 9.

Weber, J. L., David, D., Heil, J., Fan, Y., Zhao, C. & Marth, G. 2002. Human diallelic insertion/deletion polymorphisms. *Am J Hum Genet,* 71**,** 854-62.

Welch, S. G., McGregor, I. A. & Williams, K. 1977. The Duffy blood group and malaria prevalence in Gambian West Africans. *Trans R Soc Trop Med Hyg,* 71**,** 295-296.

Westen, A. A., Nagel, J. H., Benschop, C. C., Weiler, N. E., de Jong, B. J. & Sijen, T. 2009. Higher capillary electrophoresis injection settings as an efficient approach to increase the sensitivity of STR typing. *J Forensic Sci,* 54**,** 591-8.

Wetton, J., Tsang, K. & Khan, H. 2005. Inferring the population of origin of DNA evidence within the UK by allele specific hybridization of Y-SNPs. *Forensic Sci Int,* 152**,** 45-53.

White, T. D., Asfaw, B., DeGusta, D., Gilbert, H., Richards, G. D., Suwa, G. & Howell, F. C. 2003. Pleistocene Homo sapiens from Middle Awash, Ethiopia. *Nature,* 423**,** 742-7.

Wiesbauer, M., Meierhofer, D., Mayr, J. A., Sperl, W., Paulweber, B. & Kofler, B. 2006. Multiplex primer extension analysis for rapid detection of major European mitochondrial haplogroups. *Electrophoresis,* 27**,** 3864-3868.

Wilder, J., Kingan, S., Mobasher, Z., Pilkington, M. & Hammer, M. 2004a. 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.

Wilder, J. A., Mobasher, Z. & Hammer, M. F. 2004b. Genetic Evidence for Unequal Effective Population Sizes of Human Females and Males. *Mol Biol Evol,* 21**,** 2047-2057.

Wilkins, J. F. 2006. Unraveling male and female histories from human genetic data. *Curr Opin Genet Dev,* 16**,** 611-617.

Wilkins, J. F. & Marlowe, F. W. 2006. Sex-biased migration in humans: what should we expect from genetic data? *Bioessays,* 28**,** 290-300.

Willer, C. J., Bonnycastle, L. L., Conneely, K. N., Duren, W. L., Jackson, A. U., Scott, L. J., Narisu, N., Chines, P. S., Skol, A., Stringham, H. M., Petrie, J., Erdos, M. R., Swift, A. J., Enloe, S. T., Sprau, A. G., Smith, E., Tong, M., Doheny, K. F.,

Pugh, E. W., Watanabe, R. M., Buchanan, T. A., Valle, T. T., Bergman, R. N., Tuomilehto, J., Mohlke, K. L., Collins, F. S. & Boehnke, M. 2007. Screening of 134 Single Nucleotide Polymorphisms (SNPs) Previously Associated With Type 2 Diabetes Replicates Association With 12 SNPs in Nine Genes. *Diabetes,* 56**,** 256-264.

Wise, C. A., Paris, M., Morar, B., Wang, W., Kalaydjieva, L. & Bittles, A. H. 2003. A standard protocol for single nucleotide primer extension in the human genome using matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Rapid Commun Mass Spectrom,* 17**,** 1195-202.

Wolpoff, M. H., Wu, X. & Thorne, A. G. 1984. Modern Homo sapiens Origins: A General Theory of Hominid Evolution Involving the Fossil Evidence from East Asia. *In:* SMITH, F. H. & SPENCER, F. (eds.) *The Origins of modern humans: a world survey of the fossil evidence.* New York: A.R. Liss.

Wooding, S., Watkins, S., Bamshad, M., Dunn, D., Weiss, R. & Jorde, L. 2002. DNA sequence variation in a 3.7-kb noncoding sequence 5' of the CYP1A2 gene: implications for human population history and natural selection. *Am J Hum Genet,* 71**,** 528-42.

Wright, S. 1965. The Interpretation of Population Structure by F-Statistics with Special Regard to Systems of Mating. *Evolution,* 19**,** 395-420.

Wright, S. 1969. *Evolution and the genetics of populations,* University of Chicago Press, Chicago.

Wu, X., Di Rienzo, A. & Ober, C. 2001. A population genetics study of single nucleotide polymorphisms in the interleukin 4 receptor alpha (IL4RA) gene. *Genes Immun,* 2**,** 128-34.

Xiu-Cheng Fan, A., Garritsen, H. S., Tarhouny, S. E., Morris, M., Hahn, S., Holzgreve, W. & Zhong, X. Y. 2008. A rapid and accurate approach to identify single nucleotide polymorphisms of mitochondrial DNA using MALDI-TOF mass spectrometry. *Clin Chem Lab Med,* 46**,** 299-305.

Xue, Y., Zerjal, T., Bao, W., Zhu, S., Shu, Q., Xu, J., Du, R., Fu, S., Li, P., Hurles, M. E., Yang, H. & Tyler-Smith, C. 2006. Male demography in East Asia: a north-south contrast in human population expansion times. *Genetics,* 172**,** 2431-9.

Yang, N., Li, H., Criswell, L., Gregersen, P., Alarcon-Riquerline, M., Kittles, R., Shigeta, R., Silva, G., Pragna, P., Belmont, J. & Seldin, M. 2005. Examination of ancestry and ethnic affiliation using highly informative diallelic DNA markers: application to diverse and admixed populations and implications for clinical epidemiology and forensic medicine. *Hum Genet,* 118**,** 382-392.

Yang, Y.-K., Ollmann, M. M., Wilson, B. D., Dickinson, C., Yamada, T., Barsh, G. S. & Gantz, I. 1997. Effects of Recombinant Agouti-Signaling Protein on Melanocortin Action. *Mol Endocrinol,* 11**,** 274-280.

Yao, Y.-G., Kong, Q.-P., Bandelt, H. J., Kivisild, T. & Zhang, P. 2002. Phylogeographic differentiation of mitochondrial DNA in Han Chinese. *Am J Hum Genet,* 70**,** 635-651.

Yasumoto, K.-i., Yokoyama, K., Takahashi, K., Tomita, Y. & Shibahara, S. 1997. Functional Analysis of Microphthalmia-associated Transcription Factor in Pigment Cell-specific Transcription of the Human Tyrosinase Family Genes. *J Biol Chem,* 272**,** 503-509.

Yin, Y.-w., Hu, A.-m., Sun, Q.-q., Liu, H.-l., Wang, Q., Zeng, Y.-h., Xu, R.-j., Hou, Z.-z., Zhang, Z.-d., Zhang, S.-j. & Zhou, M.-q. 2012. Association between interleukin-6 gene −174 G/C polymorphism and the risk of coronary heart

disease: A meta-analysis of 20 studies including 9619 cases and 10,919 controls. *Gene,* 503**,** 25-30.

Yuasa, I., Umetsu, K., Harihara, S., Kido, A., Miyoshi, A., Saitou, N., Dashnyam, B. & Jin, F. 2007. Distribution of two Asian-related coding SNPs in the MC1R and OCA2 genes. *Biochem Genet,* 45**,** 535-542.

Yuasa, I., Umetsu, K., Watanabe, G., Nakamura, H., Endoh, M. & Irizawa, Y. 2004. MATP polymorphisms in Germans and Japanese: the L374F mutation as a population marker for Caucasoids. *Int J Legal Med,* 118**,** 364-6.

Zalloua, P., Xue, Y., Khalife, J., Makhoul, N., Debiane, L. & Platt, D. 2008. Y-chromosomal diversity in Lebanon is structured by recent historical events. *Am J Hum Genet,* 82**,** 873-882.

Zaykin, D., Zhivotovsky, L. & Weir, B. S. 1995. Exact tests for association between alleles at arbitrary numbers of loci. *Genetica,* 96**,** 169-178.

Zeigler-Johnson, C., Panossian, S., Gueye, S. M., Jalloh, M., Ofori-Adjei, D. & Kanetsky, P. A. 2004. Population Differences in the Frequency of the Agouti Signaling Protein g.8818A>G Polymorphism. *Pigment Cell Res,* 17**,** 185-187.

Zerjal, T., Wells, R. S., Yuldasheva, N., Ruzibakiev, R. & Tyler-Smith, C. 2002. A Genetic Landscape Reshaped by Recent Events: Y-Chromosomal Insights into Central Asia. *Am J Hum Genet,* 71**,** 466-482.

Zhang, M., Song, F., Liang, L., Nan, H., Zhang, J., Liu, H., Wang, L. E., Wei, Q., Lee, J. E., Amos, C. I., Kraft, P., Qureshi, A. A. & Han, J. 2013. Genome-wide association studies identify several new loci associated with pigmentation traits and skin cancer risk in European Americans. *Hum Mol Genet,* 22**,** 2948-59.

Zhivotovsky, L. A. 2001. Estimating Divergence Time with the Use of Microsatellite Genetic Distances: Impacts of Population Growth and Gene Flow. *Mol Biol Evol,* 18**,** 700-709.

Zhong, H., Beaulaurier, J., Lum, P. Y., Molony, C., Yang, X., MacNeil, D. J., Weingarth, D. T., Zhang, B., Greenawalt, D., Dobrin, R., Hao, K., Woo, S., Fabre-Suver, C., Qian, S., Tota, M. R., Keller, M. P., Kendziorski, C. M., Yandell, B. S., Castro, V., Attie, A. D., Kaplan, L. M. & Schadt, E. E. 2010. Liver and Adipose Expression Associated SNPs Are Enriched for Association to Type 2 Diabetes. *PLoS Genet,* 6**,** e1000932.

Zhong, H., Shi, H., Qi, X.-B., Duan, Z.-Y., Tan, P.-P., Jin, L., Su, B. & Ma, R. Z. 2011. Extended Y Chromosome Investigation Suggests Postglacial Migrations of Modern Humans into East Asia via the Northern Route. *Mol Biol Evol,* 28**,** 717-727.

Zhu, Y., Spitz, M. R., Lei, L., Mills, G. B. & Wu, X. 2001. A Single Nucleotide Polymorphism in the Matrix Metalloproteinase-1 Promoter Enhances Lung Cancer Susceptibility. *Cancer Res,* 61**,** 7825-7829.

Ziętkiewicz, E., Witt, M., Daca, P., Żebracka-Gala, J., Goniewicz, M., Jarząb, B. & Witt, M. 2012. Current genetic methodologies in the identification of disaster victims and in forensic analysis. *J Appl Genet,* 53**,** 41-60.

Ziv, E., John, E. M., Choudhry, S., Kho, J., Lorizio, W., Perez-Stable, E. J. & Burchard, E. G. 2006. Genetic Ancestry and Risk Factors for Breast Cancer among Latinas in the San Francisco Bay Area. *Cancer Epidemiol Biomarkers Prev,* 15**,** 1878-1885.

Zubakov, D., Liu, F., van Zelm, M. C., Vermeulen, J., Oostra, B. A., van Duijn, C. M., Driessen, G. J., van Dongen, J. J. M., Kayser, M. & Langerak, A. W. 2010. Estimating human age from T-cell DNA rearrangements. *Curr Biol,* 20**,** R970-R971.