Universidade de Santiago de Compostela
Facultade de Medicina e Odontoloxía
Instituto de Ciencias Forenses "Luís Concheiro"

Forensic Ancestry Analysis with Autosomal Polymorphisms

Memoria que presenta para optar al Grado de Doctor,

Christopher Paul Phillips

Santiago de Compostela, julio 2017

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

El Profesor Doctor **Ángel Carracedo Álvarez**, Catedrático de Medicina Legal del *Departamento de Ciencias Forenses, Anatomía Patolóxica, Xinecoloxía e Obstetricia e Pediatría* de la *Facultade de Medicina e Odontoloxía* de la *Universidade de Santiago de Compostela*,

CERTIFICA:

Que la presente memoria que lleva por título "*Forensic Ancestry Analysis with Autosomal Polymorphisms*" del Licenciado en *Biological Sciences (Genetics)* por la *University of Birmingham* y Máster en *Applied Genetics* por la *University of Birmingham* **Christopher Paul Phillips**, ha sido llevada a cabo bajo su dirección, considerándola en condiciones para optar al Grado de Doctor y autorizándola para su defensa ante el tribunal correspondiente.

De acuerdo con el artigo 41 del Regulamento de Estudos de Doutoramento, declara también que la presente tesis doctoral es idónea para ser defendida en base a la modalidad de **COMPENDIO DE ARTÍCULOS**, en los que la participación del doctorando fue decisiva para su ejecución. El resto de coautores, además, están en conocimiento de que ninguno de los trabajos aquí reunidos podrá ser presentado en ninguna otra tesis doctoral.

Y para que así conste, se expide el presente certificado en Santiago de Compostela, a 10 de julio de 2017.

Prof. Dr. Ángel Carracedo Álvarez

Christopher Paul Phillips

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Este trabajo ha sido parcialmente financiado por los proyectos de investigación Bio2013-42188-R "Desarrollo de técnicas cronométricas utilizando marcadores epigenéticos para el análisis en genética forense" (2013-2016) y Bio2016-78525-R "Análisis de polimofismos multialélicos. RETOS 2016" (2016-2019) del Ministerio de Economía, Industria y Competitividad

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Acknowledgements

I would like to dedicate this thesis to my partner *Maviky*, who has patiently waited for me to complete the writing up and analysis, while providing unwavering support and encouragement for all my scientific work for more than 25 years now.

I am greatly indebted to *Maripu*, who helped me so much to complete the thesis with enthusiasm and patience, when this was often lacking from me. The same applies to her dedication to the task of completing several papers that were gathering dust in the last three years – the fact that they turned out to be very nice publications reflects her care and attention to detail.

Of course, I wish to thank *Ángel* for supervising the work of the thesis and always allowing me to explore so many different avenues of research without ever saying I needed to stop looking at new things and do some real work. Likewise, *Toño* has always been encouraging and stopped his work to explain a concept or critically examine an idea, which has been so valuable to me.

The post-graduate students I have nurtured over the years have, in turn, helped me to perform often complex scientific analyses and also explore ideas. I am so grateful to *Fonde, Carliña* and *David* at King's College, London for this support. Each one of these accomplished scientists has about five topics that we are working on together at any one time.

Thanks to everyone in the laboratory, past and present that has tolerated my lack of Spanish and other English eccentricities, especially *Meli, Raqui, Vane, Anita, Statto, Lollipop, Maria C* and *Franci*.

I would particularly like to thank my good friends the *Maths team*: Prof. Antonio Tato, Dr. Angeles Casares de Cal and Dr. Jose Dios Alvarez for their dedication to the development of the Snipper SNP analysis portal and the gift of teaching me to look at human genetics problems in completely new ways.

Lastly, I thank *Marito, Bea* and *Ines* for our times together in the hospital labs and their genotyping work for me since then, and not least to thank *Jorge* for his outstanding work in collecting extensive human variant data and the development of the much valued SPSmart databases.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Abbreviations

ˆ - approximately
A - Adenine
ACB - African Carribbeans in Barbados
aDNA - ancient DNA
AFR - African
AIM - ancestry informative marker
ASW - Americans of African ancestry in SW USA
BEB - Bengali from Bangladesh
bp - base pairs
C - cytosine
CDX - Chinese Dai in Xishuangbanna, China
CE - capillary electrophoresis
CEPH - Centre de'Etude du Polymorphisme Humain
CEU - Utah residents N & W European ancestry
CHB - Han Chinese in Bejing, China
chr - chromosome
CLM - Colombians from Medellin, Colombia
cM - centiMorgan
dbSNP NCBI short genetic variation database
ddNTP - deoxyribonucleotide
Δ / δ - Delta allele differentiation metric
E ASN - East Asian
ESN - Esan in Nigeria
EVC - externally visible characteristics
FIN - Finnish in Finland
G - guanine
GBR - British in England and Scotland
GIH - Gujarati Indian from Houston, Texas
GWD - Gambian in Western Divisions in the Gambia MSL Mende in Sierra Leone
HGDP - Hardy Weinberg equilibrium
IBS - Iberian population in Spain
IGV - integrative genomics viewer
In - informativeness for assignment metric
Indel - insertion/deletion polymorphism
ISP - Ion Sphere™ particles
ITU - Indian Telugu from the UK
JPT - Japanese in Tokyo, Japan

Kb - kilobase
KHV - Kinh in Ho Chi Minh City, Vietnam
KYA - 1000 years ago
LD - linkage disequilibrium
LR - likelihood ratio
LWK - Luhya in Webuye, Kenya
Mb - megabase
MCMC - Markov chain Monte Carlo
min - minute
mL - millilitre
µl - microlitre
mM - millimolar
MPS - massively parallel sequencing
mtDNA - mitochondrial DNA
MXL - Mexican ancestry from Los Angeles, USA
NT - nucleotide
NCBI - National Center for Biotechnology Information
ng - nanograms
PC - principal component
PCA - principle component analysis
PCR - polymerase chain reaction
PEL - Peruvian s form Lima, Peru
PJL - Punjabi from Lahore, Pakistan
PNG - Papua New Guinea
PUR - Puerto Ricans from Puerto Rico
RFU - relative fluorescence units
S ASN - South Asian
SBE - single base extension
sec - second
SNP - single nucleotide polymorphism
STR - short tandem repeat
STU - Sri Lankan Tamil from the UK
T - thymine
TSI - Toscani in Italia
YRI - Yoruba in Ibadan, Nigeria

Resumen

La inferencia de la ancestralidad de un individuo a partir del material biológico hallado en la escena del crimen es una técnica instaurada desde hace tiempo en la comunidad forense, pero muy especializada y que a menudo carece del nivel de información adecuado para realizar inferencias fiables. Los ensayos iniciales basados en proteínas polimórficas fueron aplicados con éxito por el autor en investigaciones de casos criminales durante los años 80, pero dichos ensayos fueron abandonados cuando se desarrollaron las metodologías de obtención de perfiles de ADN. Esta tesis describe el desarrollo, la optimización y reintroducción de los ensayos forenses de predicción de ancestralidad a través del genotipado de marcadores autosómicos. Los primeros ensayos de ADN para ancestralidad se basan en los marcadores denominados polimorfismos de nucleótido único (SNPs) y fueron desarrollados por una parte, por la casa comercial DNAprint Genomics y, por la otra, por el autor en Santiago. Los ensayos desarrollados por el autor y basados en SNPs han sido aplicados para la inferencia de ancestralidad en investigaciones criminales y casos de desapariciones durante más de 10 años. Esta tesis describe los pasos fundamentales para desarrollar ensayos de predicción de ancestralidad con fines forenses que puedan ser implementados en todos aquellos laboratorios que dispongan de un secuenciador de electroforesis capilar: la optimización de la PCR tipo multiplex para detectar ADN a partir de muestras limitadas; el proceso de compilación de datos poblacionales a partir de los cuales inferir la ancestralidad más probable del individuo; la detección de patrones de co-ancestralidad en individuos con origen diverso; y el desarrollo de herramientas estadísticas online que permitan inferir el origen probable de un individuo a partir de un perfil de SNPs. Utilizando la infraestructura ampliamente establecida que se utiliza para obtener perfiles de ADN mediante sistemas de electroforesis capilar, se establecieron ensayos de otros tipos de marcadores autosómicos adicionales, tales como: polimorfismos de Inserción-Deleción (Indels); microsatélites (STRs) y SNPs multialélicos. Finalmente, las tecnologías de secuenciación masiva en paralelo permiten el desarrollo de conjuntos de marcadores de ancestralidad más extensos, beneficiándose de las capacidades de dichas plataformas: el aumento de la capacidad de multiplex y la posibilidad de conocer la fase en la que se encuentran los SNPs en cada una de las cadenas, lo que ha permitido introducir los microhaplotipos como nuevos loci informativos de ancestralidad. Esta tesis describe como los microhaplotipos han sido reducidos en tamaño sistemáticamente con el fin mejorar la sensibilidad forense y como han sido incluidos en los ensayos de ancestralidad para tecnologías de secuenciación masiva en paralelo.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Abstract

The inference of a person's ancestry from the biological material they leave at a crime-scene has been a long-standing but specialised forensic technique, which often lacks sufficient detail to make a reliable inference of ancestry. Initial tests using polymorphic proteins, were successfully applied by the author to criminal investigations in the eighties, but when DNA profiling was developed, such tests were abandoned. This thesis describes the development, optimisation and successful re-introduction of forensic ancestry analysis tests that type autosomal genetic markers. The first dedicated DNA-based forensic ancestry tests used single nucleotide polymorphisms (SNPs) and were developed by DNAprint Genomics as a commercial service, and at Santiago by the author. The SNP tests developed by the author have been used for more than 10 years to successfully identify the ancestry of individuals in criminal investigations and in tests aiming to identify the remains of missing persons. This thesis describes the key steps in developing a forensic ancestry test that can be adopted by any laboratory using capillary electrophoresis equipment: optimisation of a PCR multiplex to detect DNA markers from contact traces; compilation of population data from which to infer the likely population of origin of the person; detection of co-ancestry patterns in an individual with admixed backgrounds; and development of online statistical tools that calculate the probability of an individual's ancestry from a submitted SNP profile. Using the same well established DNA profiling infrastructure of optimised capillary electrophoresis systems, additional types of autosomal markers were compiled from Insertion-Deletion polymorphisms (Indels); short tandem repeats (STRs) and multiple-allele SNPs. Finally, expanded PCR multiplexes of ancestry markers have been developed for massively parallel sequencing which exploit both the increased multiplexing capacity of this technology and the ability to know the phase of SNPs on a sequence stand, which has enabled Microhaplotypes to be added as new ancestry informative loci. The thesis describes how Microhaplotypes have been systematically reduced in size to improve their forensic sensitivity and introduced into ancestry tests using massively parallel sequencing technology.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Introduction

1. The forensic context of bio-geographical ancestry analysis

In London seventeen years ago, upon seeing somebody acting suspiciously outside my neighbour's house, I contacted the police. They asked a simple question that often frames the eyewitness prompts made by UK police officers: "Was he White, Black or Asian". I said the person appeared White (resisting the temptation to correct these descriptions to the more neutral terminology of European, African, South Asian). As it was dark and I only had brief glimpses, it was impossible to provide a concrete description. This story exemplifies the well-known fact that eyewitness is notoriously unreliable and can be shaped by preconceptions or the traumatic circumstances of a crime [1]. Therefore, the inference of bio-geographical ancestry using markers with population-differentiated variation provides opportunities to strengthen eyewitness accounts or in their absence, gain key information about a suspect's origin.

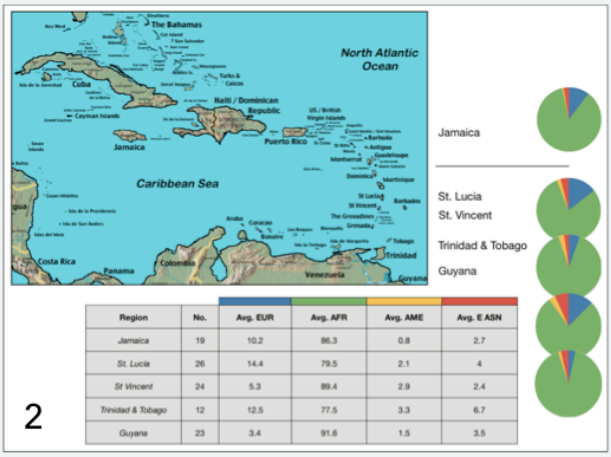
This thesis describes studies that explore the viability of forensic DNA tests estimating ancestry that can provide investigative leads when eyewitness testimony or a database hit are not available or reliable. **Box 1** describes a notable example of how bio-geographical ancestry estimation provided a key role in gaining knowledge about a serial offender who was absent from the UK national DNA database.

In simple terms, ancestry can be described as the genetic inheritance each individual carries from their ancestors, in the immediate past from their kinship, but over longer periods from population members that have occupied the same place of origin. Bio-geographical ancestry analysis focuses on population variation found in an individual that can signal their origin from a particular geographic region. Forensic bio-geographical ancestry testing exploits much of the recent advances in the understanding of human genomic variation, with the key factor that forensic DNA tests must be sensitive enough to successfully genotype contact traces or they will lack utility. Inference of ancestry in forensic analysis gives possibilities to substitute eyewitness testimony as described above—when descriptions are uncertain, unavailable or may misdirect investigators. Yet in forensic analysis, ancestry inference offers many other applications, including: (i) aiding cold case reviews with additional data on linked profiles; (ii) achieving more complete identifications of missing persons or disaster victims; (iii) confirming donor's self-declared ancestry and therefore maintaining the accuracy of databases for STRs, Y-markers and mitochondrial variation (mtDNA); (iv) refining familial search strategies highly dependent on STR allele frequency assumptions made prior to searching [2]; (v) assessing atypical combinations of physical characteristics in individuals with admixed parentage, e.g., using IrisPlex [3–5]); (vi) enhancing genetic studies where forensic sensitivity is necessary, e.g., testing medical archive material or archaeological DNA [6].

Box 1. Operation Minstead

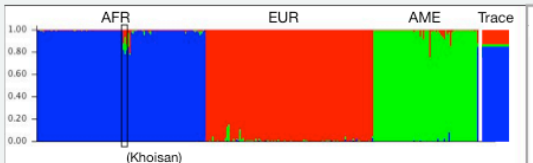
The largest and most expensive criminal investigation ever made in the UK, named Operation Minstead, tracked the attacks of a serial gerontophile rapist, active in SE London over an 18Y period (**Panel 1**). Without a database hit and with elderly victims lacking visual acuity, investigators had few leads. The US commercial company *DNAprint* performed an initial ancestry analysis on DNA from a nightdress using tests genotyping 166 SNPs. They estimated the donor to be admixed by identifying **64% African co-ancestry combined with 3% European and 33% Native American co-ancestry** - interpreted to be characteristic of the southern Caribbean (Windward Islands) region, closest to South America. The inference of likely region of origin was based on a small survey of 104 London police volunteers originating from Jamaica, St. Lucia, St. Vincent, Trinidad and Guyana (**Panel 2**). This directed police towards searches of flight records and mass DNA profiling of volunteer males in SE London with Windward Island origins. Following suggestions of unusual pigmentation from an eye-witness, USC then made their own ancestry analyses and typed SNPs associated with skin, and hair eye colour (applying the '**SHEP**' assays).

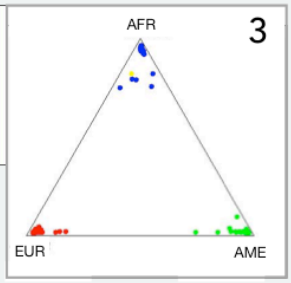
Several problems were encountered during the analyses performed at USC. First, the DNA available was from a Chelex extract made in 1997 which had degraded considerably, necessitating a fresh extract from the edges of a semen stain on the same nightdress *DNAprint* had used. Second, details of the ancestry inference systems used, the **SNP identifiers and the population survey data was not released** by *DNAprint* for our independent assessment. Furthermore, STRUCTURE analysis at USC based on 55 SNPs from 34-plex and PIMA SNaPshot assays revealed sharply contrasting co-ancestry ratios of: **African=85%, European=12%, American=3%** (**Panel 3**). While these results could be due to insufficiently differentiated SNP variants between Africa and Europe/America (compared to those of *DNAprint*), the SNPs of the PIMA set were chosen to be most informative for AME ancestry. Additionally, doubts had been raised about the accuracy of *DNAprint*'s ancestry tests, since their genealogy analyses regularly reported **paradoxical American co-ancestry** for individuals with no personal histories to suggest this was possible (*Bolnick et al., 2007, Science 318, 399-400*). Lastly, the population samples were too small and differences in allele frequency too slight between locations to allow strong inferences to be based on minor likelihood differences amongst the regions tested. Therefore, the USC analysis concluded there was **insufficient evidence for Windward Island origins** and the donor could originate from any part of mainland America or the West Atlantic (although a West African origin, e.g. Nigeria - a sizeable proportion of London's Africans, was considered unlikely).



Region	No.	Avg. EUR	Avg. AFR	Avg. AME	Avg. E ASN
Jamaica	19	10.2	86.3	0.8	2.7
St. Lucia	26	14.4	79.5	2.1	4
St Vincent	24	5.3	89.4	2.9	2.4
Trinidad & Tobago	12	12.5	77.5	3.3	6.7
Guyana	23	3.4	91.6	1.5	3.5

Pigmentation tests made at USC were prompted by a witness description of **red hair and pale skin**. The SHEP tests detected the **V60L MC1R** variant (rs1805005-GT) and other non-European skin tone variants. As only some MC1R SNPs were detected by SHEP assays, the possibility of a red-haired Afro-Caribbean individual due to a compound heterozygote of V60L with another rare MC1R variant could not be ruled out (*McKenzie et al. 2003, J. Inv. Derm. 207*). Consequently, USC sequenced the whole MC1R gene with no second variant detected; indicating the DNA donor **lacked red hair** but was very likely to show **intermediate-dark skin** and **brown eyes**.





Cluster membership proportions for 'Trace' DNA:
AFR=85% EUR=12% AME=3%
(yellow point in triangle plot and rightmost column in cluster plot of STRUCTURE output)

Eventually, **Delroy Grant** was apprehended by conventional investigative procedures using CCTV evidence. He was revealed to be **Jamaican** in origin with intermediate-dark skin tone and black hair (**Panel 1**). Lessons learned from these very challenging genetic analyses were: (1) Commercial DNA test suppliers may not respect requirements for independent scrutiny; (2) Police can 'jump' at the idea of geographic precision, even when this may lack genetic credibility and is not properly validated; (3) The notion of an unusual combination of physical characteristics (red-haired African-Caribbeans) has considerable appeal and may occasionally allow the suspect pool to be narrowed to a very small group.

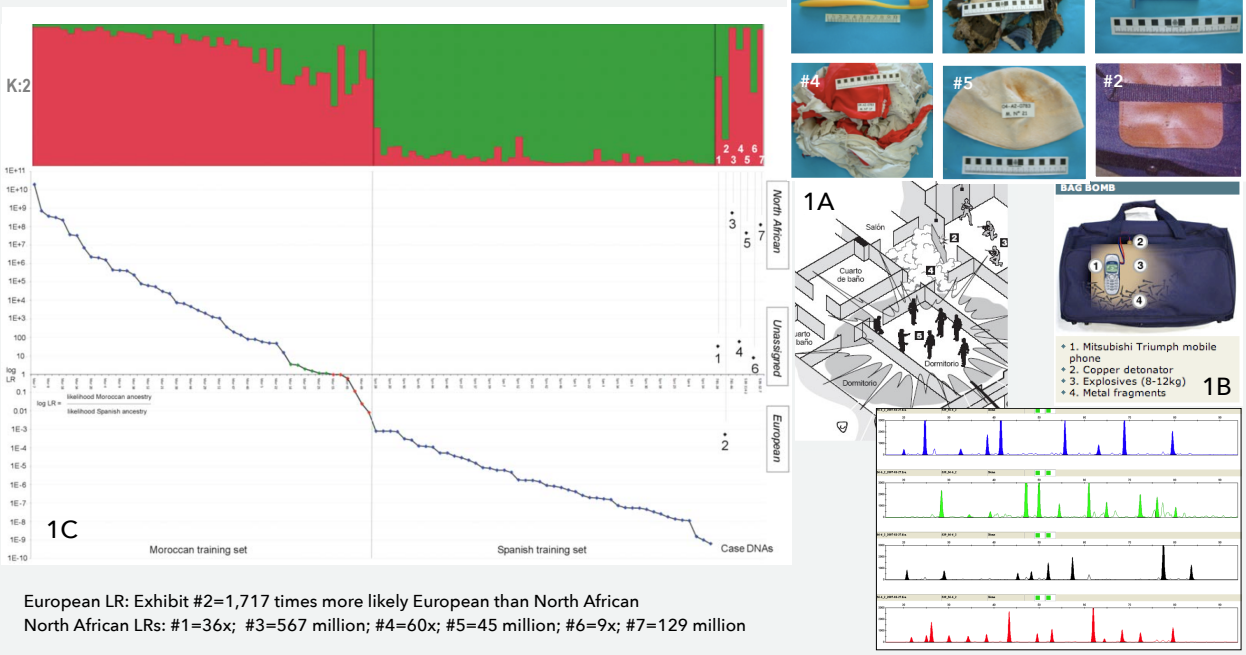
The studies described here centre on autosomal markers, despite Y and mtDNA uni-parental variation being highly differentiated geographically and therefore often forming the first and only step in forensic ancestry inference. Y and mtDNA variation is undisrupted by recombination, so is preserved in both lineages and correlates strongly with continental regions. However, Y and mtDNA variants collectively form single markers that can misrepresent an individual's overall ancestry when distant male/female lineages are inherited that have atypical ancestry. A notable example of this risk of misinterpretation was detection of African Y-chromosomes in a North Yorkshire kinship group that shared the same surname [7].

As co-ancestry in an individual indicates population admixture, increasingly common in modern urban demographics, the probability of detecting atypical lineages and misinterpreting an individual's overall ancestry rises markedly. Another advantage of recombining autosomal loci compared to Y and mtDNA is the relative ease with which population data is obtained, with as few as 30–40 samples providing adequate population allele frequency estimates. In the 11-M Madrid bomb investigation [8], discrepancies between ancestry inferences from autosomal markers and both Y and mtDNA were seen (detailed in **Box 2**). These stemmed from limited database coverage of North African populations, hampering interpretation of Y and mtDNA data based on very limited surveys of this region. The need for much larger databases to measure haplotype variation impacts reliable interpretation of uni-parental variation in many less well-studied regions and has prompted the YHRD/EMPOP forensic-community databases [9,10].

Lastly, it is important to remember forensic estimation of bio-geographical ancestry is not confined to genetic analysis, nor is it unique to the DNA profiling age. Analysis of skeletal biometrics is used to estimate ancestry with statistical classification approaches (e.g., canonical plots) similar to principal component analysis applied to genetic data. Early forensic ancestry tests used the Duffy marker in the DARC gene (rs2814778) 20 years before DNA profiling and it remains the most differentiated locus (for a brief survey of forensic ancestry analysis with classical markers, see **Thesis Paper #9, Phillips, 2015** [11]).

Box 2. The contribution of ancestry analysis to the 11-M Madrid bomb investigation

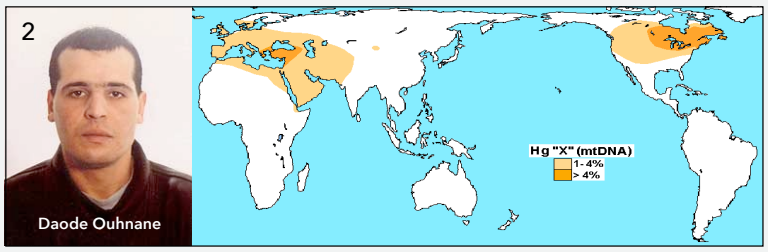
The investigation of the Madrid bomb attack of 11th March 2004 was the most extensive forensic case ever conducted in Spain and the second largest terrorist investigation in Europe after Lockerbie. Identifiler and Y-Filer STR profiling of 600 exhibits left seven key incriminatory samples recovered from personal items in the Legánes 'cell' (**Panel 1A**) plus a handprint contact trace on an unexploded bag-bomb (**Panel 1B**), **unmatched** to any known suspects. USC performed AIM-SNP analysis of these unmatched DNA extracts to **differentiate European and North African ancestry** to progress the investigation in its long search for further suspects.



European LR: Exhibit #2=1,717 times more likely European than North African
North African LRs: #1=36x; #3=567 million; #4=60x; #5=45 million; #6=9x; #7=129 million

Available **DNA was scant** from all seven exhibits (range: 0.07 - 3.3 ng/µl), so a decision to genotype the 34-plex autosomal SNPs rather than mt-DNA coding SNPs or Y-SNPs hinged on assessing the likely differentiation power of the chosen marker set. To gauge how well the 34 SNPs would distinguish European and North African populations, two **training sets of 48 Spanish and 48 Moroccans** resident in Madrid were genotyped then **cross-validated to estimate assignment error**. **Panel 1C** shows ranked likelihood ratios (LR) obtained from cross-validating each training set, with increasing distance from the 'balanced odds' midline of LR=1 indicating higher assignment likelihoods. Six Moroccan samples (red points) on or below the midline were mis-assigned as Spanish or had LR=1, giving an assignment **error rate of 12%** for North African. Another six Moroccans had LRs <10 times more likely North African which allowed **a likelihood threshold** to be set for non-assignment of **≤100 times**. Full SNP profiles were obtained. DNAs #1, #4 & #6 were below threshold LRs and so unassigned; DNAs #3, #5 & #7 gave high LRs to be North African; DNA #2 of the bag handle was >1,700 times more likely European.

The final outcome of these analyses was **four unequivocal ancestry assignments**, one for a European handprint on the bomb-bag handle, and three samples left unassigned. There was sufficient DNA left to complete Y-SNP analysis and ancestry assignments from Y-Filer, Y-SNP and/or mt-DNA tests agreed with those from SNPs. However, in the case of DNA extracted from a toothbrush found in Legánes (exhibit #3), Y-markers indicated **R1b1** haplotype and mt-DNA **X1** haplogroup; both with distributions of variation overlapping between Europe and North Africa, or lacked specificity to one particular geographic region (both factors applying to the mt-DNA X1 haplogroup distribution shown in **Panel 2**). The toothbrush had given the highest assignment likelihood from SNP analysis of **567 million times more likely North African**, so the uni-parental data's indications of European origins were ignored in favour of the contrary assignment from autosomal SNPs. One other problem was the **much lower reference sample sizes** for North Africa compared to Europe which made the uni-parental data LR difficult to interpret. The reported North African assignment from the unidentified toothbrush DNA prompted a **familial search** to be made of particular profiles and this culminated in the eventual identification of the **Algerian terrorist Daode Ouhnane** (**Panel 2**). Lessons learned from these crucial extended DNA analyses were: (1) Careful assessments of the informativeness expected from new tests are a necessary preamble when evidential material is extremely limited; (2) The scope of uni-parental marker database coverage has a bearing on how reliable the likelihood calculations will be. In the end, USC chose to compare, not to combine the LRs obtained from SNPs with those of Y/mt-DNA markers; (3) When low-level DNA extracts have been expertly prepared, the subsequent SNP tests can have extremely high forensic sensitivity.



2. Patterns of human population structure

Any concise overview of human population structure, as it is currently understood, will be an oversimplification. However, before ancestry can be inferred from small sets of forensically viable markers it is necessary to attempt a definition of human population groups based on the most strongly differentiated patterns of global genetic structure. The worldwide human population is clearly not a single entity, nor is it always appropriate to define small populations confined to narrow regions. The constraints of forensic PCR multiplex sizes and the difficulties of collecting sufficient reference data means that an over-simplified description of complex human population structure is a necessary compromise.

Human populations are not fully interbreeding, since geographic distance by itself can create a strong constraint on random mating. In addition to this, geophysical barriers such as oceans, deserts and mountains have restricted the free movement of people away from their regions-of-origin which are defined by such barriers. Therefore, population structure in early human groups became established as they continued to mate with immediate neighbours that shared their ancestry. This means forensic tests estimating ancestry might expect some success, depending on the distribution of human population structure remaining intact today. Studies of population variation in the pre-genomics age, starting with Lewontin [12], attempted to measure what genetic structure existed in modern human population groups using fairly limited numbers of polymorphic markers. Despite variation in loci and populations analysed, later studies with the same approach obtained very similar findings for apportionment of population variation (Table 10.2, [13]). Lewontin estimated within-population differences between individuals describe ~85% of autosomal variation and between-group differences ~10% (with ~5%, between-population differences within each group). As a simple example, a study comparing Pacific Islanders to Europeans would find that 10% of genetic differences between them result from their contrasted geographic origins and 90% would be found comparing individuals within each group.

A more comprehensive study by Rosenberg et al. was accomplished in 2002 [14] and used the STRUCTURE genetic-similarity clustering algorithm developed by the same group [15] to measure population structure in the now widely-used Human Genome Diversity Project-Centre Etude Polymorphism Humain sample set (herein: HGDP-CEPH). The HGDP-CEPH panel comprises 1064 individuals from 52 populations (51 when excluding the Southern Bantu singletons) and remains the most comprehensive global population survey [16], despite certain very small sample sizes and significant gaps in geographic coverage. Subsequent removal of related-pair individuals identified in many populations have since reduced the sample number to 952 (termed H952) and the HGDP-CEPH H952 panel has formed the core global human population sample set for all studies reported in this thesis.

Rosenberg's analyses used 377 STRs with high levels of polymorphism but subsequent studies of American and Oceanian populations by Wang et al. and Friedlaender et al., respectively [17,18], increased the STRs used and focused on many more populations originating from single continents. These studies established the first detailed assessments of worldwide population structure and its distribution [14,17,18]. Rosenberg estimated human diversity apportionment to be 93.2 / 94.1% within-population (the paired values corresponding to definitions of five and seven world groups or regions, respectively); 4.3 / 3.6% between-groups (2.5 / 2.4% between-population, within-group), lowering Lewontin's original estimate of between-group diversity to ~4%. Analysis with STRUCTURE consistently identifies genetic clusters based on each individual's similarity or dissimilarity to others in the sample set, (the cluster number with maximum likelihood is herein termed 'K'). Rosenberg identified continentally-defined clusters at K:5, consisting of Eurasia, sub-Saharan Africa, East Asia, America and Oceania. The seven region K:7 division assigned populations to Europe, Middle East and Central/South Asia within the broader Eurasian region. These results suggest the STRs used can separate a worldwide sample set into five groups that follow continental definitions, with evidence Eurasia separates into three further subdivisions also broadly matching the geographic distribution of populations.

If Rosenberg's results indicated Eurasian populations show less divergence than the broad division of groups into five continents, but more than between individual populations, then expanding the genetic variants used is likely to produce stable and reproducible K:7 clustering patterns. The study of Li et al. in 2008 [19] used 650,000 SNPs to analyse the same samples. Assessing the cluster plots of Rosenberg and Li (Fig. 1, [14] and [19]) indicates highly comparable patterns. Li identified seven clusters but few South Asian and no Middle East populations showed exclusive membership to one cluster (i.e., 100% proportions). Despite the greater detail obtained by Li, the pattern of human population structure and diversity is unchanged: a well-defined continental division of clusters with three Eurasian sub-groups more weakly differentiated. Therefore, it is appropriate for forensic ancestry testing to aim to assign individuals to five groups in the first instance. Rosenberg's findings led to criticism that the study chose mid-continent populations avoiding marginal zones where populations meet. This approach overlooks the clinal, continuous gradients of variation that reflect the true global patterns of population structure [20]. In response, Rosenberg's group re-analysed the HGDP-CEPH panel with more markers ([21], 377 up to 933 STRs) and demonstrated clusters are robust to sampling location. Therefore, genetic clusters represent underlying patterns of human variation and are not artefacts from uneven sampling along clines. Across the globe, allele frequency differences do increase with geographic distance in generally smooth gradients but small discontinuities remain and these create the clusters identified by STRUCTURE.

The study of American populations by Wang et al. [17] found decreasing genetic variation along the Africa–Asia/Eurasia–Oceania–America chain, explained by successive splits of populations whose small size reduced population variability each time. This serial founder model explains a successive reduction

in genetic variation with distance from the theoretical geographic focus of Addis Ababa. These patterns have a bearing on forensic ancestry testing, as American and Oceanian populations are more likely to show low heterozygosity variants with higher variability in Africans, Eurasians and East Asians. Additionally, African: non-African population divergence will generally be greater than other group comparisons; so fewer markers are well differentiated between Eurasians and East Asians. These characteristics of human variation indicate a repeated pattern of small-group migration into new regions, separation from the ancestral population group then rapid expansion. This process has allowed genetic drift to form a significant force in shaping contemporary human population structure. Three additional factors also partly explain the distribution of human diversity: regional variation in selection, migration and admixture (a sudden increase in gene flow between two differentiated populations), with the fourth most recently recognised phenomenon of archaic introgression.

Natural selection can vary according to bio-geographical factors such as climate, disease and diet [22]. Well-documented examples of these factors include genes: SLC24A5 producing de-pigmentation in Europeans; DARC producing resistance to malaria in Africans and LCT-MCM6 in three separate geographic regions as adaptation to milk consumption [23–25]; each creating strong discontinuities in variant allele frequencies,. Other equally strong allele frequency discontinuities occur from regional selection but the importance of the phenotypic change and its link to a bio-geographical factor is not apparent, e.g., EDAR and ABCC11 variants confined to much of East Asia [26,27]. So selection can lead to alleles reaching very high frequencies or even fixation in specific groups, but this process is rare [28]. The predominant mode for allele frequency differentiation to occur is more likely to be soft sweeps, where allele frequencies change more moderately and diversity between groups shows slight discontinuities [29]. Although loci near fixation are too rare to make a full set, the genes described harbour specific coding SNPs that remain the most powerful ancestry markers, with many now adopted for forensic use.

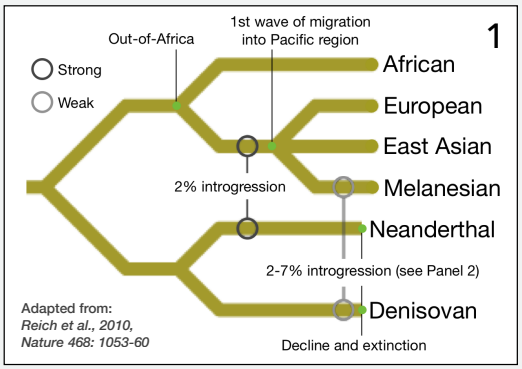
Mass movement of peoples followed by admixture is also a major influence on contemporary population diversity. The effects of slave trading and colonisation are well documented, but for more comprehensive insights into population movement predating historic record, very dense genetic data is needed. A study by Hellenthal et al. [30] used fineSTRUCTURE analysing recombinational decay of short segments containing SNP haplotypes. The same approach enabled the fine-scale analysis of UK population structure aiming to reconstruct demographic events in the peopling of the British Isles since the last Ice Age [31]. Lastly, Pickrell and Reich provide a comprehensive and informative review of the currently understood geography of human migration [32]. Their review summarises major population movements in the last 20 KY, from prehistory to recent colonialism. Lastly, the characterisation of Neanderthal–Denisovan genomes and the discovery of gene flow between these hominins and early humans has prompted much research, and Pickrell and Reich's review covers this and the most recent archaic genome analyses [32]. Studies indicate an average 2% of Neanderthal genetic ancestry is present in modern non-Africans (from introgression events 37-85 thousand years ago) and ~7%

Denisovan genetic ancestry is detected in modern Oceanians, located to smaller scale admixture events in SE Asia [33,34] - as detailed in **Box 3**.

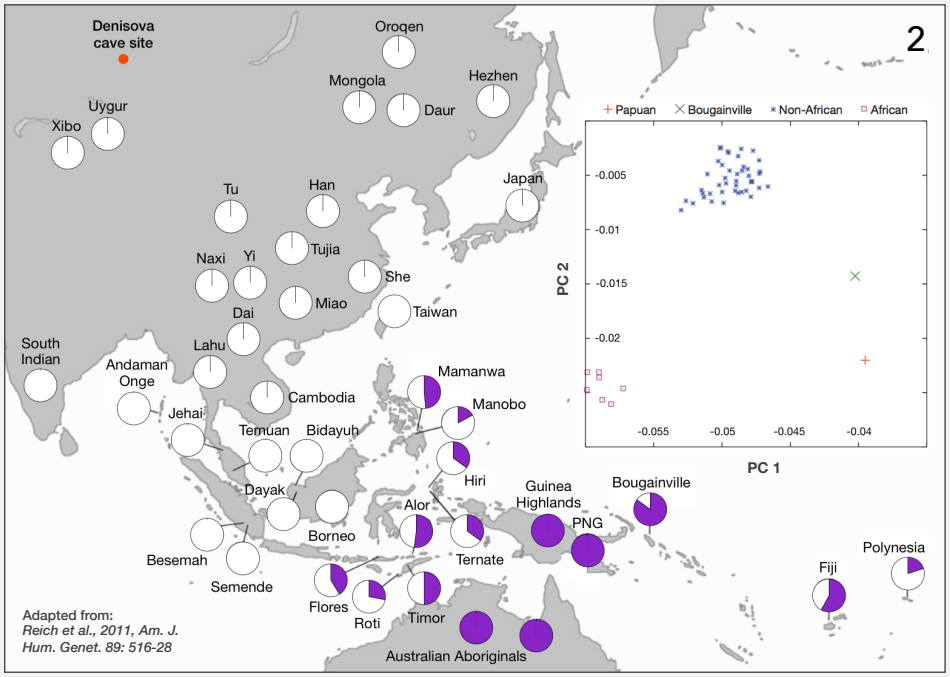
Summarising this dynamic and constantly revised field, ideas about human population history are undergoing further refinement as whole genome sequencing replaces SNP microarrays as the method of choice. Contemporary human population diversity is likely to have been shaped by past drift, selection and migration-mediated admixture, but archaic introgression has also contributed significantly to human population variation outside of Africa. A continental division of human population groups based on STRUCTURE cluster patterns provides a robust model that can form a suitable basis for ancestry assignment within the constraints of forensic testing, which necessitates simplification of complex human divergence patterns. There is not complete consensus about how STRUCTURE patterns can be interpreted, as more complete sampling of the globe, if it were possible, would be certain to reveal clinal patterns with just small discontinuities across continental divides. For forensic ancestry analysis, a five-group differentiation is a reasonable objective using compact marker sets selected to have strong allele frequency differentiation. Li's SNP analyses [19] indicate K:6, subdividing Eurasians into Europeans and South Asian groups is also feasible, while a K:7 division, differentiating Middle East Eurasians, will be much more challenging but a worthwhile goal. In practice, investigators see more value in fine-scale continental subdivisions (e.g., West vs. East European) and while forensic tests have limited marker numbers, investigator's expectations need careful handling by scientists. There is a tendency to conflate the high statistical power of DNA identity tests with the lower likelihoods in DNA ancestry tests, and what might be described as 'the illusion of geographic precision' can become established thinking. Such a misconception about the specificity of population analyses occurred with the UK Border Authority plans in 2009 to use forensic ancestry tests to distinguish Ethiopian, Somali, Kenyan and Sudanese asylum seekers [35]. This plan proceeded from the perceived success of an ancestry analysis of the 'Thames torso' case, where unidentified remains were assigned to a relatively small West African region, despite the approach used lacking proper validation or peer review [36]. This jump to over-interpret limited genetic data has similarities to current genetic genealogy analyses, which apply the very cautious academic studies of human populations to create "implausibly specific" individual histories [37,38]. The issues around both forms of misappropriation of forensic ancestry tests for unintended purposes that lack scientific credibility, are discussed in **Box 4**. Thus, forensic and population genetics specialists must guard against a desire to make overly detailed reconstructions of a person's ancestry, particularly when it has little or no relation to what is currently understood about human diversity.

Box 3. Archaic hominin introgression shaped modern human variation outside Africa

The sequencing of Neanderthal and Denisovan genomes is the latest development in a very active area of research into ancient DNA that seeks to add key details to the population histories of present-day humans, as well as discovering what makes the hominin branches distinct from each other and from Apes (using the Chimp genome as the comparison point). What is now certain from DNA studies is that Neanderthals, Denisovans and the early human groups that formed or contributed to current worldwide populations, occupied the same Eurasian territorial ranges during the Late Pleistocene period. Although the Denisovan (East Eurasia) and Neanderthal (West) ranges did not overlap and dental morphology suggests these two hominin lineages may have diverged >300,000 years ago, with support for this from genome comparisons. More importantly, genome reconstructions show significant amounts of gene-flow between human groups and Neanderthals across west Eurasia; and between ancestors of present-day Melanesians and Denisovans in Asia (**Panel 1**). The exact relationship between Neanderthals and Denisovans is unresolved, as the former has little preserved DNA per fossil while the latter has DNA of much higher quantity but from one exceptional phalanx bone (plus a molar of a second individual from the same cave, at much lower quantity). Ratios of human- vs. microbial-matched DNA reach 70% in the phalanx; a maximum 5%, average <1% in Neanderthal sources and 0.17% in the molar (two mtDNA substitutions distinguish it from the phalanx). Modern DNA contamination was estimated to be <1% from mtDNA, Y-markers (the phalanx is female) plus Chimp-Human SNP ancestral alleles shared with Denisova vs. fully fixed derived alleles in modern humans. Therefore, a previously unknown archaic population that was neither Neanderthal nor modern human was present in Siberia before 50,000 years ago. The Denisovans: a reconstructed population from the Denisova cave discoveries shows how **ancient DNA can reveal 'genomes in search of a fossil'** - i.e. ancient hominins whose existence had not been expected from the archaeological or fossil records.



Two key findings from the detection of this gene-flow impact the current understanding of present-day worldwide patterns of variation. Firstly, there was archaic Neanderthal gene flow into human groups outside Africa some 37,000–85,000 years ago. This **gene flow contributed ~2% of the genetic ancestry of non-Africans**. Earlier evidence of ancient admixture in non-Africans had been proposed from analysis of modern humans, but there was general acceptance of admixture occurring only when ancient DNA evidence revealed that the deeply divergent genome segments in present-day non-Africans are related to those of Neanderthals. Secondly, there was gene flow from a population related to the Denisovans into the ancestors of present-day aboriginal people from Papua New Guinea, Australia, and the Philippines. These populations all live in Oceania, far from the Denisova cave in Altai, Siberia, but no Denisovan DNA is detected in individuals from mainland East Asia. Therefore, the **admixture events must have occurred in Southeast Asia** since no equivalent gene flow is found in the genomes of mainland East Asians *nor* in any descendant population of the earliest Southeast Asian settlers (white charts, **Panel 2**, *Reich, et al., 2011, Am. J. Hum. Genet. 89: 516-28*). It also reveals Denisovans occupied a vast geographic and ecological range, from Siberia to tropical Asia, far exceeding that of Neanderthals. This **gene flow is estimated to have contributed up to 7% of Denisovan genetic ancestry in modern Oceanian populations** (see **Panel 2** for range).



Proportions of Denisovan DNA detected in an extended set of Oceanian & Asian populations, expressed as a fraction of Papua New Guinea (PNG) proportions (set to 100%). Data indicates PNG and Australian Aboriginals are more closely related to first wave settlers of Oceania than other populations of the region. White pie charts reveal all modern mainland E Asian and most SE Asian populations lack Denisova admixture. Charts with lines indicate locations of HGDP-CEPH Asian population samples.

PCA plot upper right shows the positions of 53 HGDP-CEPH population mean principal components (PCs) in relation to PC1-PC2 defined by Denisova-Neanderthal-Chimp variation in 255,077 SNPs. Non-Africans form a distinct cluster (which includes all the East Asian populations on the map, left) while both the Oceanian populations are obvious outliers.

Box 4. Misconceptions and misappropriation - DTC ancestry tests and the UKBA Human Provenance pilot

SNP tests to infer bio-geographical ancestry have formed part of direct-to-consumer (DTC) commercial genealogy services and have been subject to much criticism due to a **lack of proper validation**: the carefully organised assessments of the error-rates of a test's statistical inferences. It can be argued that such ancestry services offer customers a unique opportunity to find out more about their origins, which had not been possible before the genomics era. However, there is little opportunity for customers to critically appraise the test result's scientific validity or geographic accuracy. It is regularly the case that DTC reports give a variety of claims about the precise geographic locations of a person's forebears. If they tell an acceptable and often exciting personal history then the customer is generally happy and does not challenge the accuracy of the claims. However, paradoxical ancestry inferences can be made when the markers used have not been properly assessed in a sufficient range of populations. The *Science* commentary shown in **Panel 1** by Bolnick (see *Box 1* for it's relevance to the Minstead case) highlighted the consistently incorrect interpretation of certain SNP alleles to be indicative of Native American ancestry in the *DNAprint* DTC ancestry tests. This reveals major misconceptions due to lack of validation and error estimation in the AIMs used. This led to the problem of their application unmodified to forensic cases that could have caused inaccurate ancestry data to be given to investigators - in turn this can undermine confidence in the quality of data from other forensic ancestry tests that *have* had their error/success carefully checked.

Most DTC tests rely on Y and mtDNA data rather than autosomal loci as they are easier tests to perform and their variation databases have sufficient worldwide scope to give good geographic data. Much of this data has proved very useful to the forensic genetics community in the form of much improved global sampling and, in the case of Y variation, the opportunity to **match rare surnames to the Y-lineages** of male DTC customers from countries with paternal isonomy. Another positive aspect of improved knowledge of genomic variation is the future prospects to utilise carefully constructed genome-wide SNP panels specifically designed to analyse fine-scale human variation patterns, such as the Genographic SNP array (*Genochip*) or the Affymetrix *Human Origins* chip. It seems certain these will find their way into DTC ancestry testing soon and this will improve reference data coverage to the benefit of population and forensic genetics. They also have obvious applicability when DNA is not in scant quantities, e.g. from unidentified, but recently deceased individuals.

2

ScienceInsider

Breaking news and analysis from the world of science policy

Roundup 9/28: Shock and Awe Edition | Main | Key Questions on Nationality Testing

SEPTEMBER 29, 2009

Scientists Decry "Flawed" and "Horrifying" Nationality Tests

Scientists are greeting with surprise and dismay a project to use DNA and isotope analysis of tissue from asylum seekers to evaluate their nationality and help decide who can enter the United Kingdom. "Horrifying," "naïve," and "flawed" are among the adjectives geneticists and isotope specialists have used to describe the "Human Provenance pilot project," launched quietly in mid-September by the U.K. Border Agency. Their consensus: The project is not scientifically valid–or even sensible.

Another geneticist says the UK Forensic Science Service requested his opinion earlier this year on how to develop a genetic assay to distinguish East African populations. "I thought it was for forensic purposes, not border control," says Christopher Phillips of the University of Santiago de Compostela in Spain, who with colleagues recently used a DNA sample to correctly infer the ancestry of a suspect in the 2004 train bombings in Madrid. After expressing skepticism about the goal, he suggested some research the FSS could conduct but says he heard no more from them.

Having their fate rest on unproven methods is particularly dangerous for asylum-seekers in the United Kingdom, notes Phillips, because unlike criminal defendants, they have limited or no rights to challenge evidence or appeal. "You can't parachute in a technique if it isn't properly validated," he says.

Genomics Law Report

News and analysis from the intersection of genomics, personalized medicine and the law

Why the Errors of the Human Provenance Project Will Echo Beyond the U.K.'s Borders

Posted by Dan Vorhaus on September 29, 2009

To the list of iniquitous uses of genomic data we may now add "national origin." By mis-appropriating genomic technologies in a questionable attempt at border control, the Human Provenance project has shifted the conversation from a prospective discussion of risk avoidance and mitigation to a present debate over whether the Border Agency's use of genomic and isotope data analysis is ethically or legally appropriate, let alone scientifically valid.

From the Genetic Information Nondiscrimination Act (GINA) in the United States to the EU's Directive 95/46/EC on data protection to Germany's recently enacted Human Genetic Examination Act, where governments have tackled the issue of genomic data availability and usage, the results to date have typically been restrictive and proscriptive. This is a trend that, if allowed to proceed unchecked, threatens the future of genomic research and one which many scientists and policymakers vigorously oppose.

Worse than scientific misconceptions from ancestry tests sold to an uncritical public, is the misappropriation of existing forensic tests for improper and scientifically unsound use. The **UK Border Authority Human Provenance Project** - proposed to differentiate the origins of UK asylum seekers from Sudan, Somalia, Kenya and Ethiopia - rightly provoked a storm of protest from geneticists (**Panel 2**) on its flawed thinking, lack of scientific validity and sinister overtones. It revealed policy makers can be ill-informed and make poor judgements, in the absence of expert input to discussions about the precision with which a person's ancestry can be inferred. In the case of the Provenance Pilot, this mainly stemmed from police investigators incorrectly judging the ancestry analysis of a boy's torso found in the Thames (the 'Adam torso' case) to have been geographically accurate. This is an example of **confirmation bias** - when results seem to fit the supposition and so are deemed correct. Yet, no error rate estimation was made of the torso analyses and they were never known to be geographically correct.

GENETICS

The Science and Business of Genetic Ancestry Testing

Commercially available tests of genetic ancestry have significant scientific limitations, but are serious matters for many test-takers.

Deborah A. Bolnick,[1]* Duana Fullwiley,[2] Troy Duster,[3,4] Richard S. Cooper,[5] Joan H. Fujimura,[6] Jonathan Kahn,[7] Jay S. Kaufman,[8] Jonathan Marks,[9] Ann Morning,[3] Alondra Nelson,[10] Pilar Ossorio,[11] Jenny Reardon,[12] Susan M. Reverby,[13] Kimberly TallBear[14,15]

At least two dozen companies now market "genetic ancestry tests" to help consumers reconstruct their family histories and determine the geographic origins of their ancestors. More than 460,000 people have purchased these tests over the past 6 years (*1*), and public interest is still skyrocketing (*1–4*). Some scientists support this enterprise because it makes genetics accessible and relevant; others view it with indifference, seeing the tests as merely "recreational." However, both scientists and consumers should approach genetic ancestry testing with caution because (i) the tests can have a profound impact on individuals and communities, (ii) the assum-

The Impact of "Recreational Genetics"
Although genetic ancestry testing is often described as "recreational genetics," many consumers do not take these tests lightly. Each test costs $100 to $900, and consumers often have deep personal reasons for purchasing these products. Many indi-

African communities. Other Americans have taken the tests in hope of obtaining Native American tribal affiliation (and benefits like financial support, housing, education, health care, and affirmation of identity) or to challenge tribal membership decisions (*7*).

The problems described here are likely responsible for the most paradoxical results of this test. For instance, the AncestryByDNA test suggests that most people from the Middle East, India, and the Mediterranean region of Europe have Native American ancestry (*15*). Because no archaeological, genetic, or historical evidence supports this suggestion, the test probably considers some markers to be diagnostic of Native American ancestry when, in fact, they are not.

1

3. Choosing ancestry informative markers

3.1. Measures of locus divergence and the first forensic ancestry panel

Early forensic ancestry tests of autosomal ancestry informative marker (AIM) SNPs were based on admixture mapping (MALD) panels that had in turn used the 2001 Human Genome Mapping Project SNP map [39]. The first AIM-SNP panel specifically for forensic use was launched in 2003 and comprised 178 SNPs detected in multiple PCR multiplexes using the now defunct SNPstream system. This ancestry test was run for seven years by the DNAprint Company as the 'Ancestry-By-DNA' service. Therefore, during that period data about the SNPs (their identifiers, population frequencies and genotyping performance with forensic DNA) were not available for independent review by the forensic and legal communities. Eventually the component SNP details were published in 2008 [40] just before DNAprint ceased operations. The original selection of SNPs for the DNAprint panels had followed the framework developed by Shriver et al. for identifying ancestry informative variation [41,42]. Shriver proposed the genetic distance between populations for any one marker could be estimated from the δ metric: the allele frequency differential, as the absolute value of p_x-p_y (comparing allele frequency p in populations X and Y). The δ value is very simply calculated in binary loci but has a more complex derivation in multiple allele systems such as STRs (estimated from the genetic distance matrix of individual $\delta\mu^2$ values [43]). Shriver demonstrated that SNPs sorted by δ produced a ranked list of ancestry markers that maximise the collective divergence amongst the population group comparisons they are selected for. Population differentiation is more commonly measured by the fixation index F_{ST}, while δ is further refined by calculating the informativeness-for-assignment metric $\boldsymbol{I_n}$ derived from Jensen–Shannon's Divergence measure [44,45]. In practice, all four values are closely related measurements of degrees of population differentiation. For example, $F_{ST} \approx \delta^2$ or $F_{ST} \approx \delta/(2-\delta)$, and $\boldsymbol{I_n}$ is Divergence x 0.693 (i.e., converting the natural log to $\boldsymbol{\ln_{(2)}}$). All have maximum values of 1 in pairwise population comparisons, denoting complete divergence and zero for no discernible divergence. Divergence values can be automatically estimated for up to 2000 SNPs and 200 populations when their genotypes are obtained from SPSmart then uploaded to the Snipper websites (http:// spsmart.cesga.es and http://mathgene.usc.es/snipper/index.php, respectively), as described in [46]. The relationship of F_{ST}, δ and Shannon's or Rosenberg's Divergence is explored further in **Thesis Paper #9, Phillips, 2015** [11]).

Before taking the obvious step of selecting the topmost SNPs from a ranked list and bringing them together into a compact test, there are several factors needing consideration: the balance of divergence the AIM set shows amongst population groups; the availability and scope of population data; and SNP acquisition bias. The distribution of human diversity has led to strong divergence between African and other populations followed by that between Eurasians and other populations, with East Asians showing the lowest divergence with Oceanians and Americans due to recent founding events in these regions. Therefore, selection of forensic AIM-SNPs tends to find many more African-

informative loci than for other group comparisons. Americans as a population group with only 15 thousand years of separation has the least divergence from the closely related East Asians [47]. This means that divergence values need careful consideration for the population comparisons that a test seeks to make. As well as being easier to find, African-informative AIM-SNPs also show higher average levels of differentiation. If a reasonable goal of a compact forensic ancestry test is to differentiate Africa, Europe and East Asia, it is harder to find markers distinguishing Europeans and East Asians. Furthermore, very similar divergence values can be obtained from differing allele frequency distributions. To illustrate this principal, **Figure 1** shows several highly informative AIM-SNPs with different patterns of divergence between the above three groups. SNPs rs12075 and rs4988235 show contrasting population specific divergences (***I_n POP***) between Europe and the other two (which can be put as ***I_n EUR*** vs. ***I_n AFR*** and ***I_n E ASN***). If a forensic test only used SNPs like rs12075 it would have less power to differentiate Europeans, so SNPs such as rs4988235 are required to redress the balance. It is also possible to obtain In and Divergence estimates from the online SPSmart and Snipper portals respectively. The other three SNPs detailed in **Figure 1** are close to allelic fixation in their respective groups (frequencies of 0 or 1) and provide arguably the best three binary AIMs in the human genome. The final cumulative ***I_n POP*** values are reasonably equilibrated in the range 1.35-1.48, indicating these five SNPs offer some balance in their capacity to differentiate the three target population groups with equal power. However, when forensic ancestry tests grow to 30 or more loci, maintaining a balance of population-specific ***I_n*** divergence values becomes more difficult. Another challenge to maintaining balanced divergences is the paucity of fixed SNPs with maximum differentiation in certain pairs of groups.

Fixed-allele frequency distributions originate from favourable coding SNP substitutions creating hard sweeps from very strong positive selection. However, soft sweeps are more common, while rapidly evolving traits under strong selection such as hypolactasia (the underlying SNP for this trait is rs4988235 in **Figure 1**) usually fail to replace all existing variation in a region [48]. Lastly, AIMs not at fixation but showing allele frequency differences have varying divergence values in each population comparison so each new marker added produces imbalance. Undue divergence imbalance in an AIM set can bias the estimation of co-ancestry in individuals from admixed populations, as illustrated in the analysis of Bolivians by Taboada-Echalar et al. [49]. This study compared co-ancestry proportion estimates from a 46 AIM-indel set [50] with a much larger genomics AIM set of 446 SNPs (the 'LACE' panel [51]). The admixed Bolivian's AIM-indel data consistently under-estimated Native American ancestry and over-estimation European ancestry compared to the 446 LACE SNPs. The indels have less divergence for Americans than Europeans (***I_n AME*** < ***I_n EUR***), whereas the LACE panel is more successfully balanced between these groups, suggesting small-scale marker sets appropriate for forensic analysis are prone to biased estimation of co-ancestry proportions in individuals from admixed populations. Population-specific divergence (PSD or ***I_n POP***) was previously recognised by Shriver et al. [52] and termed the locus-specific branch length (LBSL). LSBLs for the above groups can be estimated by calculating three I_n divergences for: African vs. the other two populations (***I_n AFR***);

European vs. the other two; East Asian vs. the other two. In **Figure 1**, rs2814778 shows lower I_n ***EUR*** and I_n ***E ASN*** values, as all these group's divergence is with Africans. Providing that the cumulative PSD values (obtained from addition) in each population group are comparable, the AIM set can be considered to have balanced differentiation of those groups and this can minimise the admixture estimation bias prevalent in the small-scale AIM sets necessary for forensic ancestry analyses.

3.2. Availability of reference population data and SNaPshot-based forensic ancestry panels

Although an ancestry test may intend to target the differentiation of other population groups besides Africans, Europeans and East Asians, marker variation data is not always available to allow selection of AIMs informative for Oceanian, unadmixed American or South Asian/Middle East Eurasian populations. However, access to detailed SNP data from Li's HGDP-CEPH study of 650,000 loci [19] and the 1000 Genomes project [53,54] is straightforward. In 2014, 1000 Genomes published a final comprehensive catalog of human variants with SNP numbers expanded from the initial Phase I list of ~28 million variants in 629 individuals from 12 populations, to ~79 million variants (77,520,219 single nucleotide SNPs comprising simple A/C/G/T substitutions) in 2,504 individuals from 26 populations [55]. The 1000 Genomes SNP data is particularly detailed and almost all SNPs with minor allele frequencies ≥1% have been characterised in the populations detailed in **Figure 2**. Although Li's HGDP-CEPH data surveys much less loci in comparison, the inclusion of two Oceanian, five American, nine Central-South Asian and four Middle East populations addresses other worldwide regions. Although SNP data must be collected locus-by-locus in the 1000 Genomes website, a simpler approach uses the SPSmart ENGINES portal [56] (Phase I data), which accepts queries that comprise lists of SNPs, chromosome segments or gene symbols. The genotypes obtained can then be downloaded to Excel and when populations are labeled as African, non-African; European, non-European, etc., cross-validation can be performed in Snipper (http://mathgene.usc.es/snipper/analysispopfile2_new. html) to obtain their PSD/I_n ***POP*** values.

Two forensic AIM panels were developed shortly after the DNAprint set, using SNaPshot primer extension chemistry: a 34-plex SNP assay from the SNP*for*ID Consortium [57,58] (herein '34-plex') and a set of 47 SNPs developed in Holland [59]. Both sets have subsequently been adapted: the 34-plex with a single SNP swap-out (rs727811 > rs3827760 [58]); and the Dutch panel condensed by Lao et al. into two 12-plex assays [60]. Selection of 47 Dutch SNPs involved screening 8,474 candidates in the Affymetrix 10K genome-wide array using 74 Y-Chromosome Consortium samples analysed with STRUCTURE (optimum K:4 clusters of African–American–Asian–Eurasian groups). The SNPs were assessed with F_{ST} and I_n4 i.e. Divergence comparing the four regions defined by STRUCTURE, and I_n6; which compared six regions dividing Asia into Asia or Northern Asia (Russia and Siberia) plus Africa into Central or South Africa. The best match of STRUCTURE cluster patterns was obtained with AIMs selected with highest pairwise F_{ST} values (distinct from classical F_{ST} looking at all groups together). This shows that aiming for balanced divergence selects the most informative set. Analysing the HGDP-CEPH

panel with the best 47 SNPs gave optimum cluster patterns at K:4 (Oceanians and East Asians not separated). In contrast, the 34-plex selection process used reduced population data available from the MALD panels published at the time [61,62]. Therefore, 34-plex development was too early to properly evaluate non-European Eurasian, American or Oceanian variation. Nevertheless, using the 34-plex set to analyse HGDP-CEPH samples with STRUCTURE produced cluster patterns reasonably well matched to Rosenberg's (Fig. 1, [14]; Fig. 3, [57]; Fig. 4A, [58]). Considerations of ascertainment bias are illustrated by the selection processes applied to both of these forensic AIM sets. First, many population surveys used to select AIMs for MALD and CCAS applications are very limited in sample size and geographic scope. The HGDP-CEPH panel was used to ensure 34-plex had low within-group divergence, but this may not apply to a continent as diverse as Africa. Second, the 650,000 SNPs typed for HGDP-CEPH with the Illumina 650K set were mainly selected from European and African American population data [63], therefore many loci with low allele frequencies in either group were excluded but could prove useful for other population differentiations. Notably, loci close to fixation are the best AIMs, but have no value for mapping or association studies as they lack statistical power and are consequently excluded from genome-wide SNP arrays.

Finally, two other SNaPshot-based forensic ancestry panels have been published by Gettings et al. [64], genotyping 50 AIM SNPs in three multiplexed assays and by Daniel et al. [65], genotyping 14 AIM SNPs in two multiplexed assays.

3.3. Large-scale genomics ancestry panels and forensic SNP genotyping with NGS

Since most of the SNP sets described above were developed, larger panels have been compiled to provide statistical adjustments for genomics studies. These sets provide powerful AIMs that are well worth consideration for forensic use. In order-of-publication they are: Paschou et al. of 50 SNPs [66]; Kosoy et al. of 128 SNPs [67] (often named Seldin's AIM panel), and; Galanter et al. of 446 SNPs [51]. All three studies have focussed on African, European and Native American SNP variation (i.e. not East Asian), but no studies developed optimised PCR multiplexes. Two recently published forensic AIM sets from Kidd et al. [68] and **Phillips et al.** [69] combined 55 and 128 SNPs, respectively. Both anticipate the expanded multiplexing scales offered by next generation sequencing (NGS). The study of Kidd assessed the Kosoy AIMs [67] with a large set of new populations and indicated they are 'transportable' to East Asians or other groups such as South Asians, not originally targeted in the SNP selection process.

	SNP	Chr:position	Gene	RA	RA AFR (85 YRI)	RA EUR (85 CEU)	RA E ASN (85 CHB)	I_{n3}	$I_{n\ AFR}$	$I_{n\ EUR}$	$I_{n\ E\ ASN}$	AFR	EUR	E ASN
1	**rs4988235**	2:136608646	MCM6*	C	1	0.276	1	*0.335*	*0.131*	*0.349*	*0.131*			
2	**rs12075**	1:159175354	DARC	G	0	0.435	0.947	*0.379*	*0.318*	*0.001*	*0.313*			
3	**rs2814778**	1:159174683	DARC	A	1	0	0	*0.599*	*0.662*	*0.200*	*0.200*			
4	**rs1426654**	15:48426484	SLC24A5	A	0.018	1	0.029	*0.557*	*0.188*	*0.617*	*0.169*			
5	**rs3827760**	2:109513601	EDAR	T	1	1	0.053	*0.531*	*0.186*	*0.186*	*0.576*			
								Cumulative $I_{n\ POP}$:	***1.48***	***1.35***	***1.39***			

*Promotor SNP for LCT

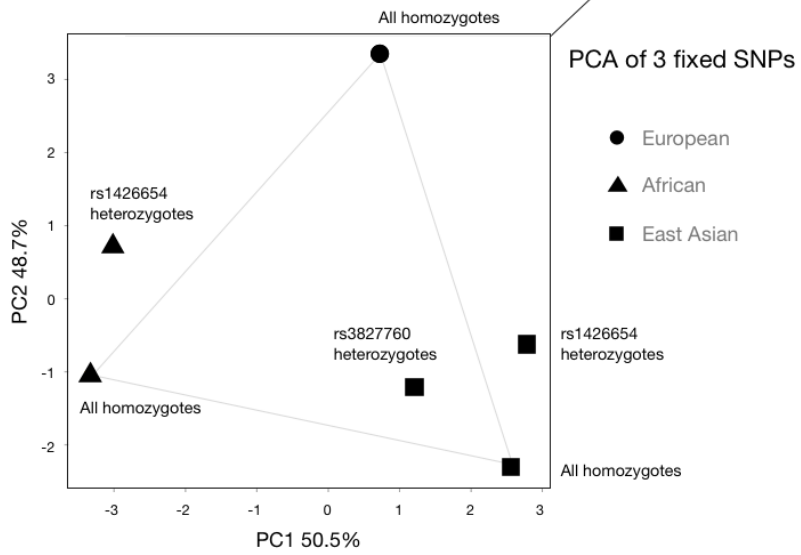


Figure 1. Five examples of AIM-SNPs. SNP 1 shows a population group-specific allele, SNP 2 has near-fixed variation between Africans and East Asians. SNPs 3–5 are the most informative (reflected in the In values listed) with fixed alleles in each group. Combined InPOP divergences reach a reasonably comparable level of balance as SNP properties compensate for the distribution of variation amongst the groups analysed. The PCA plot shows analysis of genotypes for SNPs 3–5, where a perfect triangle indicates genetic data was almost completely transformed to two PC axes. Genotypes are from sample size-adjusted 1000 Genomes Phase III data. RA, reference allele; Chr, chromosome. The promotor SNP for LCT, rs4988235 is sited in MCM6.

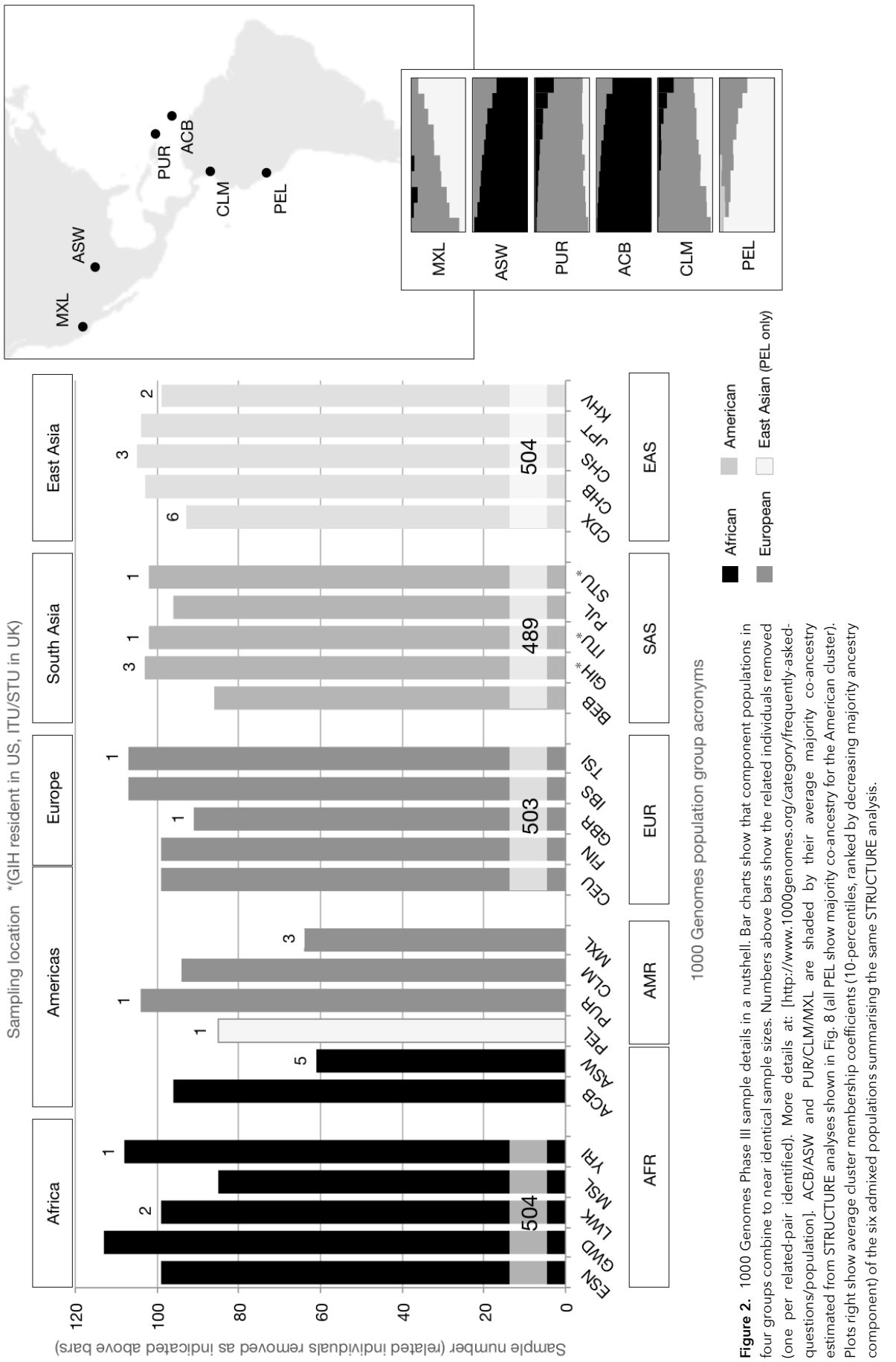


Figure 2. 1000 Genomes Phase III sample details in a nutshell. Bar charts show that component populations in four groups combine to near identical sample sizes. Numbers above bars show the related individuals removed (one per related-pair identified). More details at: [http://www.1000genomes.org/category/frequently-asked-questions/population]. ACB/ASW and PUR/CLM/MXL are shaded by their average majority co-ancestry estimated from STRUCTURE analyses shown in Fig. 8 (all PEL show majority co-ancestry for the American cluster). Plots right show average cluster membership coefficients (10-percentiles, ranked by decreasing majority ancestry component) of the six admixed populations summarising the same STRUCTURE analysis.

To adjust further for SNP ascertainment bias and add more highly differentiated AIMs, Kidd developed a non-overlapping set of 55 AIMs listed in the FROG-kb website [70]. The combination of 128 plus 55 AIMs (165 novel loci discounting overlapping loci), forms the HID-Ion Precision ID Ancestry Panel optimised for the Ion PGM™ system [71], while the 55 Kidd AIMs alone form the ancestry informative portion of the Illumina MiSeq ForenSeq® DNA Signature Kit system [72]. The study of **Phillips** selected 128 'Global' AIMs from several sources including the Kiddlab 55, but the highest proportion were taken from Galanter's LACE panel [51]. The two main objectives of this study were to incorporate new AIMs that differentiated Native American and Oceanian ancestry and to balance the PSD/I_n ***POP*** values as fully as possible. Five PSD values were collected for each SNP and the composition carefully adjusted so each group's cumulative divergence reached near-identical levels of differentiation (I_n ***EUR***: 14.56, I_n ***E ASN***: 14.23, I_n ***OCE***: 14.71, I_n ***AME***: 14.82, I_n ***AFR***: 14.84).

With so many SNPs now available to choose and scope for re-combining different sets into larger PCR multiplexes for NGS [73], it is instructive to compare the top AIMs. To do this, 1000 Genomes data were collated from SPSmart then individual PSD values were estimated for the component SNPs of the main sets described above, comparing standard African, European and East Asian populations (YRI, CEU, CHB). The twenty most informative SNPs for each group differentiation from seven AIM panels are shown in **Figure 3** and the top 24 across all panels are listed in **Table 1**. Certain patterns are discernible, notably the higher overall ancestry informativeness of the top Global and Kiddlab SNPs, particularly for European/East Asian and African comparisons, respectively. Common AIMs in Global/34-plex and Kiddlab sets of African rs2814778, rs1871534, European rs16891982, rs1426654, and East Asian rs3827760, rs4918664 (3rd vs. 2nd best) are also evident, with rs2814778 in *DNAprint* and rs16891982 in Lao sets. These SNPs therefore represent a core set and would form the first step in building all forensic ancestry tests, as do many other Global and Kiddlab SNPs. **Table 2** lists the ten AIMs common to four or five panels and it is noteworthy all but one are coding SNPs. The EDAR SNPs rs260690 and rs3827760 highlight the issue of using linked loci from the same divergent genes in forensic ancestry sets. The SNPs are not closely sited (109,579,738-109,513,601=61 kb), but multiple markers from one genomic region add a degree of bias in assessing admixed individuals. For example, an individual may be predominantly European with minor East Asian co-ancestry but inherits alleles rs260690-C and rs3827760-G, adding twice the indicative data from one genomic segment. This factor is worth considering carefully as, e.g., SLC45A2—rs26722/rs35395 have been proposed as potential AIMs for sets including rs16891982 from the same gene [74]. Any gene variation under strong selection can affect a large array of SNPs to create similar levels of divergence, while regions without recombination tend to show identical allelic patterns across quite large spans [75].

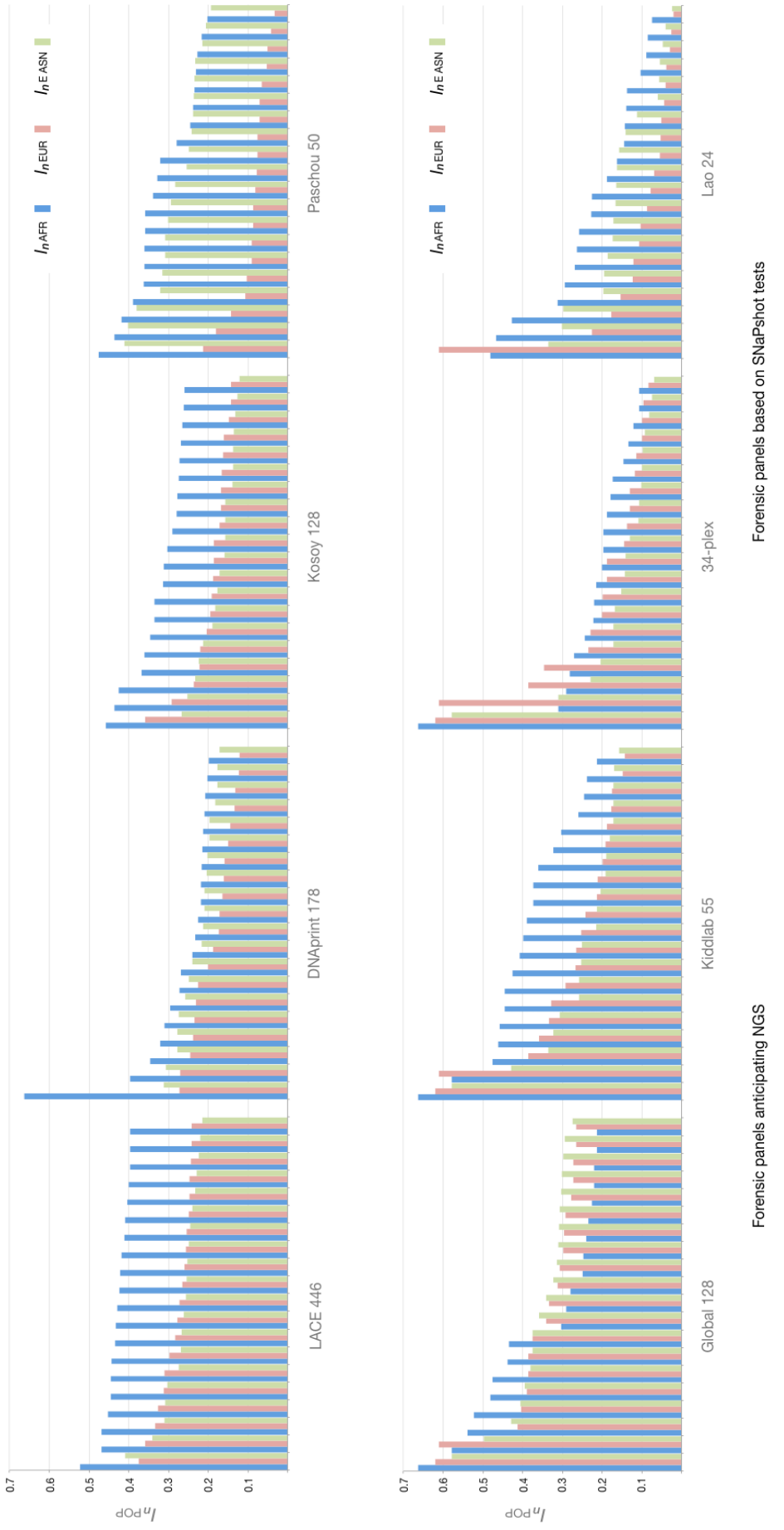


Figure 3. Distribution of population specific divergence values in the best 20 SNPs of eight AIM panels applicable to forensic ancestry analysis. Lower charts show panels with multiplexed assays in use (*Kiddlab* 55+Kosoy 128 in *Ion PGM*™ ancestry panel). Most panels show higher African informativeness amongst their most powerful markers. For marker commonality between these panels and overall best markers, see Tables 1-2.

Table 1 SNPs common to three or four established ancestry panels. Bold rs-numbers indicate the most informative markers differentiating Africans, Europeans and East Asians. Note rs260690 and rs3827760 are AIMs from the same divergent EDAR gene.

SNP	LACE 446	DNAprint 178	Kosoy 128	Global 128	Kiddlab 55	Paschou 50	34-plex	Lao 24
rs16891982	–	–	–	√	√	–	√	√
rs2814778	–	√	–	√	√	–	√	–
rs12913832	–	–	–	√	√	–	√	–
rs1426654	–	–	–	√	√	–	√	–
rs174570	√	–	–	√	√	–	–	–
rs260690[a]	√	–	√	–	√	–	–	–
rs310644	–	–	–	√	√	√	–	–
rs3827760[a]	–	–	–	√	√	–	√	–
rs730570	–	√	–	√	–	–	√	–
rs9522149	–	–	√	√	√	–	–	–

[a] Both SNPs sited in EDAR.

Table 2 The 24 most informative AIM-SNPs. **(A)** African-informative markers, using divergence In values calculated from genotypes of 1000 Genomes Yoruba in Ibadan, Nigeria (YRI). **(B)** European-informative markers, using 1000 Genomes CEPH Utah residents with N & W European ancestry (CEU). **(C)** East Asian-informative markers, using Han Chinese in Beijing, China (CHB).

(A) Rank	AIM	I_nAFR	Kiddlab 55	Global 128	Other sets
1	rs2814778	0.663	√	√	DNAprint/34-plex
2	rs1871534	0.579	√	√	
3	rs2789823	0.540		√	
4	rs6875659	0.523		√	LACE 446
5	rs1369290	0.481		√	Lao 24
6	rs310644	0.476	√	√	Paschou 50
7	rs11051	0.469			LACE 446
8	rs10258063	0.469			LACE 446
9	rs1448484	0.468			Lao 24
10	rs3916235	0.461	√		
11	rs4891825	0.457	√		Kosoy 128
12	rs4598087	0.452			LACE 446
13	rs4789193	0.446			LACE 446
14	rs3823159	0.445	√		
15	rs10497191	0.445	√		LACE 446
16	rs7752055	0.442			LACE 446
17	rs6034866	0.439		√	
18	rs10007810	0.436			Kosoy 128
19	rs387098	0.436			Paschou 50
20	rs1197062	0.435		√	LACE 446
21	rs10848765	0.432			LACE 446
22	rs6866970	0.429			LACE 446
23	rs1478785	0.426			Lao 24
24	rs11652805	0.426			Kosoy 128
		11.274 (cumulative I_nAFR)			

(B)

Rank	AIM	I_nEUR	Kiddlab 55	Global 128	Other
1	rs1426654	0.620	√	√	34-plex
2	rs16891982	0.611	√	√	Lao/34-plex
3	rs8072587	0.414		√	
4	rs7531501	0.404		√	
5	rs12142199	0.389		√	
6	rs12913832	0.386	√	√	34-plex
7	rs7084970	0.386		√	
8	rs820371	0.375		√	LACE 446
9	rs260690	0.359	√		LACE/Kosoy
10	rs182549	0.346			34-plex
11	rs1592672	0.340		√	
12	rs1924381	0.333		√	LACE 446
13	rs6754311	0.333	√		
14	rs2196051	0.329	√		
15	rs1453858	0.326			LACE 446
16	rs4791868	0.311		√	LACE 446
17	rs1419138	0.309			LACE 446
18	rs634392	0.307		√	
19	rs1486341	0.298		√	LACE 446
20	rs930072	0.295		√	
21	rs9522149	0.292	√	√	Kosoy 128
22	rs8068853	0.283			LACE 446
23	rs4787040	0.278		√	LACE 446
24	rs595961	0.273		√	DNAprint 178
		8.522 (cumulative I_nEUR)			

(C)

Rank	AIM	I_nE ASN	Kiddlab 55	Global 128	Other
1	rs3827760	0.578	√	√	34-plex
2	rs17822931	0.498		√	
3	rs4918664	0.428	√	√	
4	rs6583859	0.411			Paschou 50
5	rs4244304	0.409			LACE 446
6	rs6437783	0.406		√	
7	rs10882168	0.402			Paschou 50
8	rs12594144	0.394		√	
9	rs9809818	0.380		√	Paschou 50
10	rs4935501	0.375		√	
11	rs4657449	0.375		√	
12	rs2180052	0.359		√	
13	rs10079352	0.341		√	LACE 446
14	rs1876482	0.334	√		Lao 24
15	rs1229984	0.323	√	√	
16	rs9388489	0.321			Paschou 50
17	rs2572450	0.316			Paschou 50
18	rs17544484	0.314		√	
19	rs830599	0.312			DNAprint 178
20	rs1586861	0.311			LACE 446
21	rs881929	0.310		√	34-plex
22	rs4841527	0.309			Paschou 50
23	rs1366220	0.307		√	LACE 446
24	rs1560971	0.307			Paschou 50
		8.820 (cumulative I_nE ASN)			

4. Population data analysis systems

4.1. Publicly available SNP data

SNP genotype data is easily obtained from SPSmart, but it currently only queries 1000 Genomes Phase I populations (14 rather than the 26 of the final release), while the FROGkb and ALFRED allele frequency sites also provide comprehensive data [76,77]. Once variant data has been acquired, three statistical systems of population comparison are applicable to analysis of bio-geographical ancestry: Bayes analysis, principal component analysis (PCA) and STRUCTURE, itself using Bayesian analyses. Each analysis system uses reference population data and makes inferences from the comparative patterns of variation detected. A profile of AIM genotypes of unknown ancestry is analysed at the same time and compared to reference data. Therefore, a key factor needing careful consideration in forensic ancestry inference is the relevance, quality and scope of the population data available. Although genetic data is extensive and freely available from open-access portals, there are significant gaps in population coverage in both 1000 Genomes and HGDP-CEPH sampling. The HGDP-CEPH panel lacks data for SNPs outside the Illumina 650K genome-wide array, suffers from very small sample sizes for many populations and has ascertainment bias issues previously discussed. There are also coverage gaps for Native North American, Native Australian, Micronesian, Polynesian, North Asian, Southeast Asian, North African, hunter-gatherer African and East African populations, not filled by 1000 genomes. The forensic SPSmart browsers have been set up to accept population data and maintains dedicated pages for the 34-plex [78] and 46 AIM-indel [79] panels that already have optimised CE genotyping systems fully described [57,58,50], while larger NGS AIM panels are now ready to use. These factors are important because ancestry analysis is most effective when the population reference data has maximum scope. Therefore, a worthwhile goal would be to characterise a large collection of populations for a small number of ancestry panels using manageable sample sizes (samples of ~50 per population are sufficient). The growing interest in forensic NGS analysis is likely to make such a program easier to establish.

4.2. Bayes analysis

Lowe et al. developed the first DNA-era forensic ancestry test in 2001 using six STRs then in routine use in the UK [80]. Lowe's study was the first to propose Bayes analysis to assign an STR profile of unknown ancestry to the most likely population of origin in a simple and intuitive way. Bayes analysis uses the combined genotype frequencies estimated for each population to calculate their likelihood and assigns a probability of ancestry from the ratio of the two highest likelihoods. Although ancestry assignment error rates were high compared to later SNP analysis levels, this was partly due to reliance on police descriptions to label DNA samples as belonging to five different populations. This highlights the problem of potential mismatches between population genetics and the public understanding of what is commonly termed 'ethnicity'. For example, police and the public often fail to distinguish

between South and East Asians or sub-Saharan and North Africans. Nevertheless, Lowe's study set the direction for future development of SNP-based ancestry tests by applying a simple Bayesian approach to forensic data.

The 34-plex SNP ancestry test previously described, organised Bayes analysis in the online portal named Snipper – a web-based likelihood calculator. The *Snipper* site holds training sets, providing the reference data from which allele frequencies are calculated, although users can upload their own SNP data. The original 34-plex training sets comprised HGDP-CEPH genotypes plus in-house populations from Mozambique, Somalia, Taiwan, Mainland China, NW Spain and Denmark. The dual sampling allowed a swapped test set-training set analysis, i.e., one population acts as training set for the other, treated as 'unknown', and vice versa. The *Snipper* site also allows a crosscheck of novel training set data by one-out cross validation (http://mathgene.usc.es/snipper/analysispopfile2_new.html). Custom training set data uploaded to Snipper, are potentially most useful, since genotypes generated by a laboratory or collected from online/published data can be applied to any AIM set of interest. The steps for manipulating SPSmart 1000 Genomes or HGDP-CEPH SNP genotypes then creating custom training sets are detailed in Ref. [46].

To illustrate analysis of a SNP profile, 34-plex genotypes for control DNA 9947A can be uploaded to Snipper. 34-plex SNP profiles can be assessed with the fixed training set page (http://mathgene.usc.es/snipper/popchoosing5groups.html), comprising HGDP-CEPH training sets for 34 SNPs and/or 46 indels. Users can opt for three, four or five group reference data allowing selection of African–European–East Asian genotypes for 34-plex profiles; four groups, adding Americans, for 46 AIM-indel data and five, adding Oceanians, for combined 80-marker profiles. Uploading the 9947A profile returns a European vs. East Asian likelihood ratio (LR) of 2.58E+21; a very high likelihood to be European rather than East Asian or African. However, such a high value is difficult to interpret directly and needs some qualification. First, only three possible population groups were compared for this assignment, if more are included the LRs drop, since other groups can be less divergent from Europeans than East Asians. Choosing the option to upload five group training sets with the same profile returns an LR to be European of 5.18E+18 compared to American, as this group replaces East Asian as the second highest likelihood. Second, the profile analysed can be correctly assigned to a group but the donor originates from a divergent population. However, this is conservative in effect, reducing the probability obtained, as allele frequencies match less well with the profile. The HGDP-CEPH San samples from South Africa are all rs2814778-T homozygotes, reducing this SNP's African likelihood down to a very low value, but the remaining 33 SNPs produce likelihoods almost identical to other Africans. Third, it is difficult to obtain a sufficient spread of population data to properly represent the full range of within-group variation. It is important to ensure the AIMs used show much lower within-group than between-group divergence. SNPs rs12913832 and rs182549 (associated with blue eyes and hypolactasia, respectively) are the only 34-plex AIMs with significant within-group variation. Within-group divergence is particularly relevant to Eurasia, where populations occupy a large and

varied geographic area. The approach adopted for the Eurasiaplex SNP panel, differentiating Europeans from South Asians [81], was to set a threshold probability. Establishing a realistic minimum threshold for Snipper, below which no assignment is made, can help minimise error if carefully balanced against a reasonable non-classification rate. This approach was also used in the 11-M ancestry analyses to define the range of Snipper probabilities that were considered unreliable [8]. Lastly, the effect of partial data on Bayes ancestry assignment probabilities can be explored by uploading a progressively deficient profile to Snipper. **Figure 4** shows decreasing European assignment probabilities as SNPs are removed from the 9947A profile (marked NN), starting with the best marker, rs1426654 and working down the InEUR ranked list. Although likelihoods eventually reach uninformative levels, when 25% of markers are missing the LR exceeds 10 million, and a profile of the 19 least informative SNPs still exceeds 1000.

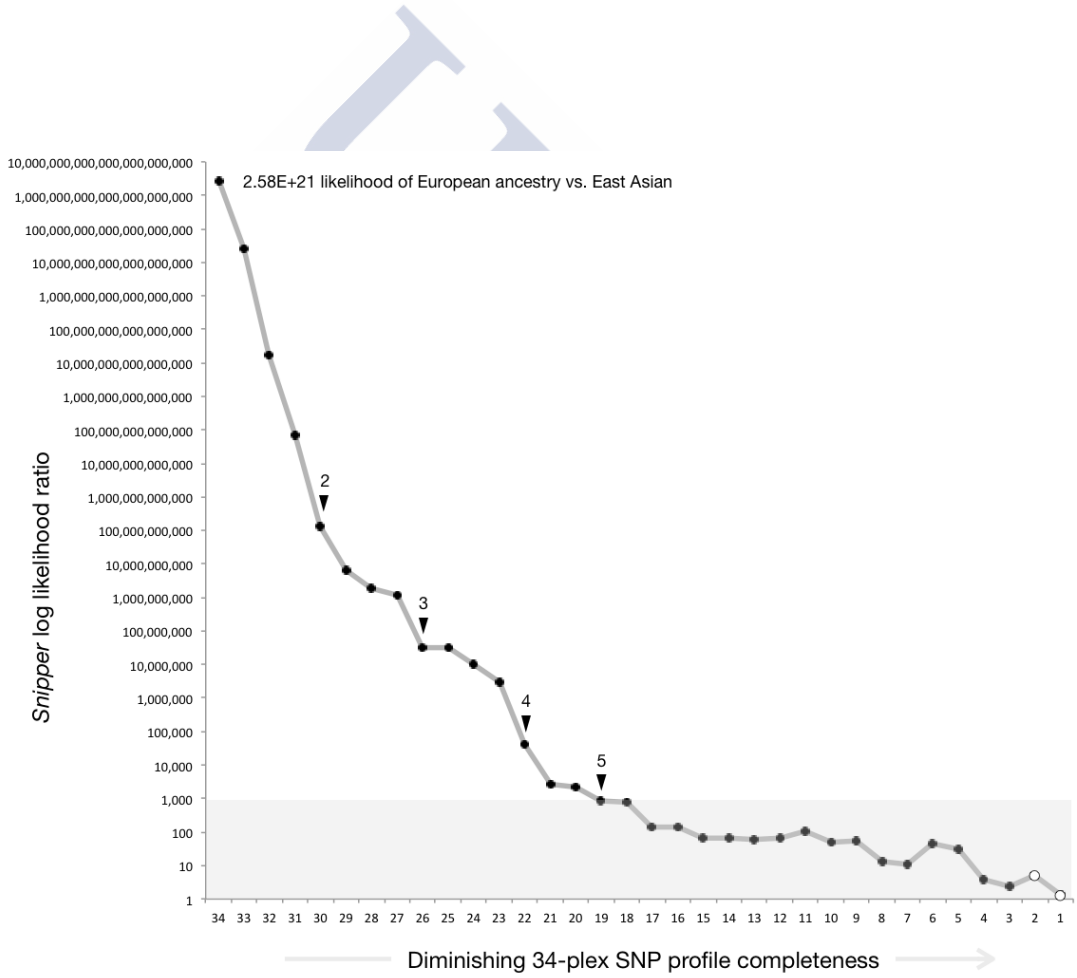


Figure 4. Bayes analysis with Snipper of progressively degraded 34-plex 9947A profile, removing the most informative AIM-SNPs first. Log_{10} likelihood ratio (LR) probabilities are for European vs. East Asian ancestry likelihoods, apart from the 1-2 SNP profiles showing African vs. European LRs. Numbers refer to points shown on the 34-plex PCA of Figure 5A. The shaded portion indicates a suggested minimum LR threshold below which no inference is made, but see the position of points 3–5 in the PCA of Figure 5A.

4.3. Principal component analysis

PCA tests were first proposed in the nineties by Cavalli-Sforza et al., in order to summarise complex population data from multiple loci, in a worldwide study of the geographic distribution of classical marker variation [82]. PCA is the most widely used type of multi-dimensional scaling (MDS) analyses that reduce the dimensionality of data while keeping the largest possible portion of it variability. PCA calculates a new set of uncorrelated variables: the principal components (PCs), made from a linear combination of the original variables ('R' dimensions). Each new PC captures only a proportion of variance, but is estimated sequentially, i.e., the first PC captures the largest proportion, then the second PC, etc. The combined PCs define a sample's eigenvector [83,84]. When analysing population genetic data from simple SNP tests such as those already described, the condensation of total variance follows an approximate route of $R^{20\sim200}$ into R^3, i.e., extracting ~3 PCs sequentially from allelic data that has high dimensionality. Therefore, three PCs commonly account for a large percentage of total variation and efficiently represent the main patterns of genetic divergence found in the SNP data [83,84]. PCA plots display the PCs as X–Y–Z axes with their proportions of variance and any one sample's position defined by its eigenvector. However, 3D plots are not easily displayed 'on paper', so publications tend to show PC1–PC2; PC1–PC3; PC2–PC3 individually or more often 2D PC1–PC2 plots containing most information in the simplest space.

The review of new developments in forensic genetics by Kayser and de Knijff [85] contains a number of definitions of terms and a good set of examples of SNP-based multidimensional scaling plots (MDS, distinct from PCA, as plots showed Laplacian eigenvector analyses [86]). Before describing how simple 2D PCA plots can be generated from forensic SNP data in Snipper, it is worthwhile highlighting benefits and shortcomings of this type of analysis applied to populations. The spatial arrangement of population clusters in PCA specifically, and MDS in general, can be a product of the geometric transformations used as much as the divergence patterns amongst the populations. Kayser and de Knijff highlight that Laplacian eigenvector analysis benefits from comparing each sample only to its immediate neighbours [85]. Therefore, the inference of past population events such as random genetic drift, or migration history from directly comparing MDS plots to geography, remains controversial [82–84,87–89]. Similarly, the tendency to superimpose PCA distributions directly onto geographic space may create close matches that are persuasive, but cannot properly define the relative degrees of divergence amongst the populations compared. Nevertheless, fine-scale population differentiations have been successfully obtained in step-wise sampling of Western Europe from two parallel studies [90,91]. Superimposition of 2D PCAs and maps is particularly good in each analysis and reflects the detail and geographic resolution achievable from half a million SNPs ([90] suggests ~800 km). A very good overview of the pitfalls of over-interpretation of PCA analysis is provided in opinion box 6.3 in Ref. [13].

Although caution is necessary, PCAs actually provide an intuitive and simply understood way to interpret patterns of divergence amongst sets of populations. If populations are sufficiently diverse and the markers well differentiated, individuals form discrete clusters of points with distributions in 2D space that reflect their genetic differentiation. To illustrate this, a PCA made from just three SNPs is shown in **Figure 1**. Because the three markers have fixed alleles, the linear combination of their data should be perfect, i.e., forming an equilateral triangle. However, the effect of a small number of heterozygotes in Africans and East Asians is clearly shown in the displaced points (all points are multiple samples with identical eigenvalues). When many more loci are used, sample's PCA positions disperse to mainly unique points in a plot. An informative approach for forensic ancestry analysis is to overlay a sample directly onto a set of reference population PCA clusters and assess its relative position. This prompted development of a PCA module in Snipper that makes the Bayes analysis and simultaneously generates a PCA plot marking the novel profile positions. Reference genotypes are uploaded in the same Excel file as the profile data. Reference genotype rows are marked with '1' and unknown profiles with '0' so their eigenvalues are calculated individually and they can be positioning directly over reference clusters. The *Snipper* analysis returns a PC1–PC2 plot with accompanying Bayes likelihoods below. **Figure 5** shows two PCAs made by *Snipper* analysis of 34-plex (plot A) and Global SNP data. Each analysis used the same 1000 Genomes reference data for three groups with the Global plot adding 1000 Genomes South Asians, HGDP-CEPH Americans and Oceanians. Although individual populations are not marked, there is no discernible substructure within any one cluster suggesting both panels have minimised within-group divergence. The 9947A positions are shown in the same way forensic SNP profiles would be marked, with the grey points in plot A showing PCA positions for the 34-plex profile with reducing profile completeness (**Figure 4**). Lastly, point M shows an artificial 3:1 mixed DNA sample combining Chinese and European donors. As the 34-plex SNaPshot test makes little distinction between imbalanced and normal heterozygote peak pairs, the genotypes show a comparable number of East Asian- and European-informative alleles and mimics a PCA distribution seen in individuals with co-ancestry.

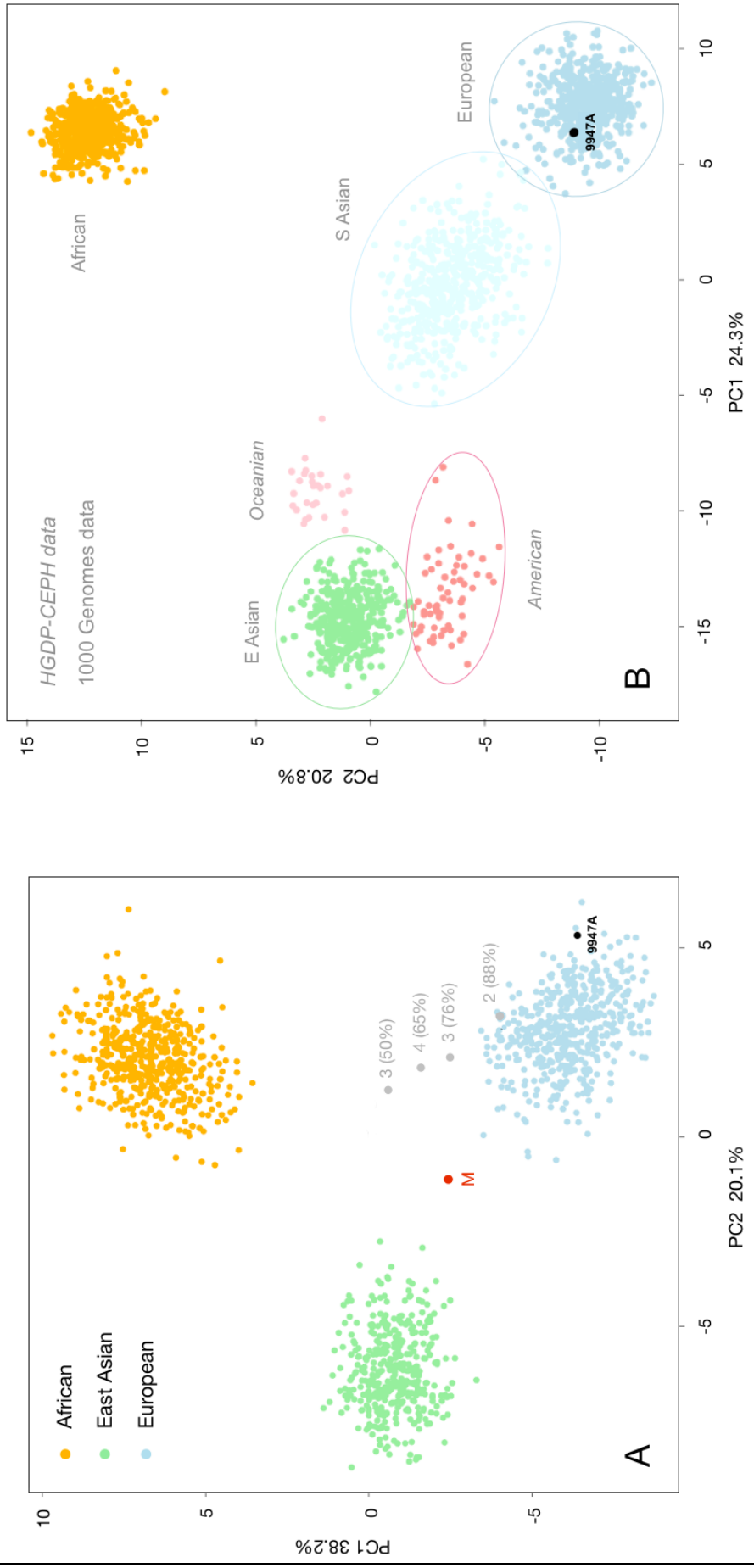


Figure 5. Snipper generated PCA plots showing the system of overlaying an unknown profile (9947A) onto reference data clusters. **(A)** 34-plex genotypes from 1000 Genomes. Grey points are points of reduced 9947A profile shown in Figure 4. Point M is an artificial 3:1 mixture of European-East Asian donors. **(B)** Global 128 AIM panel genotypes from 1000 Genomes and HGDP-CEPH Native American/Oceanian samples. Borders added to show marginal overlap between two sets of clusters.

4.4. STRUCTURE analysis

The most widely used population analysis program STRUCTURE [15,92] applies a systematic Bayesian clustering approach that can handle both SNP and STR genotypes simultaneously, offering more flexibility than Bayes analysis with Snipper or PCA. The graphical processing of STRUCTURE output is enhanced with DISTRUCT [93], making it straightforward to create the cluster plots typically seen in many published studies. More importantly, the robustness of the population cluster number (K) estimation which lies at the core of STRUCTURE analyses is now measurable using CLUMPP [94]. Once cluster analysis has been made for reference populations, individual ancestry can be inferred from cluster membership observed in novel samples. A matrix of cluster membership coefficients is produced from each STRUCTURE run, allowing comparisons between reference and unknown sample coefficients. Cluster analysis is normally unsupervised, so samples are not labeled by region and consequently clusters are assigned solely on the patterns of genetic similarity detected amongst the samples. This process is often used to test the efficiency of a marker set to differentiate particular group comparisons, i.e., if clusters match the region of origin well then the panel can be considered informative for the groups that have been analysed.

While it is easy to be persuaded by a good fit of clusters to population data, important caveats apply to analyses relying on STRUCTURE to infer ancestry. The estimation of K that best fits the data is not necessarily easy to achieve, nor is it always straightforward to interpret the relationship of K to the actual genetic structure in the populations analysed [95]. Often several K values give near identical likelihoods-of-data in CLUMPP, when it is best to take the first stable probability and smallest K, not let prior assumptions about the sampled populations influence interpretation. The study by Kidd of the Kosoy AIM panel with a very extensive set of populations [68] raised two important issues on the use of STRUCTURE. First, heterogeneous sample sizes and the distribution of sampled populations can strongly influence the formation of clusters and best-fit probabilities of K. Second, STRUCTURE uses analyses that are stochastic, often leading to different outcomes between runs. A statement in [87] summarises this effect well: 'the point of using STRUCTURE is not the single best run or the most common pattern seen, but the stability of aspects of the patterns (obtained)'. Lastly, STRUCTURE analysis is often used to assess admixture and can provide clearly delineated clusters that are easy to compare to individuals with known admixture [58]. Again, Kidd discussed the dangers of this approach in [68], highlighting the fact that the original populations contributing to admixture cannot be efficiently extrapolated from modern samples. Furthermore, populations on the margins of the main continents can often mimic the patterns seen in samples of admixed individuals.

5. The complexities of population admixture

Section 2 described population admixture as a dominant characteristic of populations on continental margins and a regular occurrence ever since small human groups first migrated. Populations have

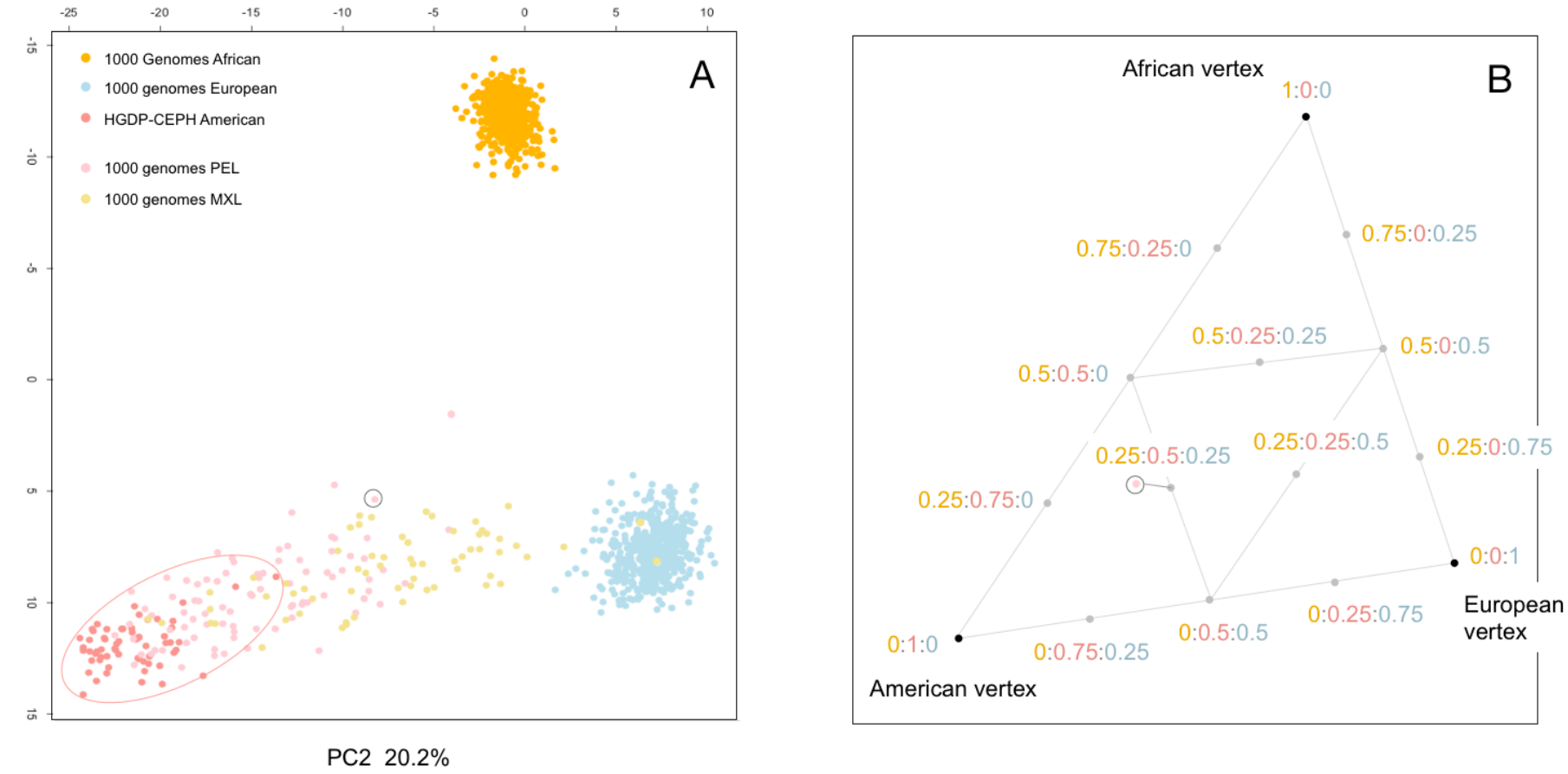
continued to meet with increasing frequency across 2500 years of trade, conquest and slavery (Fig. 2, [32]). Two centuries of urbanisation and mass movement have since removed the cultural and social barriers that previously substituted for geographic separation. Consequently, forensic ancestry analyses can expect to see a large proportion of admixture patterns amongst tested individuals. Investigators also have particular interest in admixture because it suggests the possibility of unusual combinations of physical characteristics in a suspect. The author's laboratory sequenced the MC1R gene in a DNA sample, as it gave strong indications of mainly African co-ancestry in the donor plus an MC1R V60L 'r' variant (rs1805005-T) suggesting a possible combination of red hair and dark skin [96]. Therefore, it is instructive to assess how the three analytical approaches to forensic ancestry inference outlined above each deal with admixture. If a suitable detection framework can be established this can prompt follow up tests to increase the genetic differentiation of the contributor populations, improving estimation of co-ancestry components, particularly when Y and mtDNA data can be added.

Bayes analysis is the most limited approach for admixture detection as an LR is largely quantitative; it makes an inference based on the two highest likelihoods, which can have much reduced values while still providing a number. Although the lowest value from a set of LRs gives indications of the likely contributor populations, it is not easy to arrange comparisons of expected likelihood ranges from unadmixed vs. admixed reference individuals. The pairwise ranked log LR charts used in the 11-M analyses (Fig. 1 in [8] and **Box 2**) can be applied to public genomic data from admixed populations such as Mexicans vs. appropriate contributor ancestries (European and Native American). The steps needed to produce these charts with Snipper output are described in [46]. Depending on the populations compared, a pattern often seen in the data consists of a flat distribution of log LR values showing minor differences on both sides of the midline, then a gradient of values in between from admixed samples that may be steep when proportions of individuals have recent admixture. With the obvious risk of over-interpretation of complex patterns from limited genetic data, such a comparative analysis can only realistically provide a way to set LR thresholds to minimise assignment error. This was the process applied to the 11-M data, where ~10% of the Moroccans tested gave European ancestry assignments but with LRs below 100. Setting a threshold of 100 allowed more secure interpretation of the much higher LRs obtained in four of the seven 34-plex profiles from the investigation [8].

PCA can provide a simple system for identifying admixed individuals that the analysis may position between reference clusters in simple 2D plots. The caveat applies that partial data or undetected mixed DNA genotypes will displace the true position of an individual towards clusters that are not necessarily related to their ancestry. More broadly based comparisons with PCA also need an efficient way to view 3D plots to ensure separation of contributor population clusters in PC3 are detected. Therefore, it is best to compare three admixture contributor populations per analysis. This can be arranged from the known divergence in the AIMs used. For example, applying the Kosoy AIM panel to a PCA comparison of American and East Asian reference data would show limited divergence that Kiddlab SNPs could address [67,68]. Therefore, a forensic sample analysed in a US laboratory might

consider a PCA of Africans/Europeans with Native American or East Asian data in separate analyses. This approach has been adopted by Illumina as an automated analysis of AIMs data from the Illumina ForenSeq forensic marker panel [72]. The PCA plot generated also calculates centroids that place a series of points scaled to the eigenvectors of the reference cluster centres (the triangle vertices formed by three clusters in simple PC1–PC2 plots). The distance to the closest centroid is reported for the forensic sample's position to help interpretation of points outside a reference cluster. The concept is illustrated in **Figure 6** with a three-way PCA of African–American–European plus PEL and MXL admixed populations (Global AIMs genotypes). An example PEL point is shown closest to the 0.25–0.5–0.25 centroid (above reference group order), suggesting this sample has majority American co-ancestry with detectable European and African components.

STRUCTURE has been the most widely used approach for analysing admixture patterns but coefficients of cluster member- ship taken from the output matrix do not necessarily provide a definitive picture of a person's likely admixture, given all the caveats listed in Section 4. It is also a mistake to interpret membership coefficients below 10% as meaningful. Attempts have been made to address the variance in cluster membership estimates that will be useful to explore further, but these have been developed with large marker sets in mind [97]. Although running STRUCTURE for each new profile is cumbersome, using cluster plots to assess joint memberships and possible admixture may be required to give a complimentary approach to PCA. Therefore, STRUCTURE provides a follow up strategy for complete, single-source profiles tested with PCA-Bayes analysis that show displacement outside the reference clusters and/or low LRs. To illustrate patterns laboratories could expect to find, the six admixed populations from 1000 Genomes were analysed population-by-population, in parallel PCA-STRUCTURE runs (Global AIMs). The patterns in each paired analysis match well. PCA plot outliers correspond to samples with the highest ratios of joint cluster membership and instances of three-contributor admixture show displacement towards mid-plot positions. Lastly, it is worthwhile to gain knowledge of the admixture profile of a population sample, even though this is highly variable, an idea of the range and limits can help in the interpretation of American 'Hispanic' population data in particular. Cluster plots in **Figure 7** show quite flat sigmoid distributions helping to define the range extremes for the two major contributor groups in each case. However, average values have limited value in such varied cluster proportions, so the 10-percentiles were calculated from the STRUCTURE output and plotted in **Figure 2**. These indicate Mexicans have a balanced range of European and American co-ancestry contributions. African Americans/African Caribbeans show European co-ancestry ranging from 0 to ~40/20%. Interestingly, Puerto Ricans, Colombians and Peruvians all show a third co-ancestry contributor to varying degrees (East Asian proportions in PEL were close to 10% but consistent). Colombians show the most heterogeneous patterns, exemplifying the more challenging type of population that forensic ancestry tests will need to analyse.



PCA

Figure 6. **(A)** PCA plot of African, European and American reference groups compared to PEL and MXL (1000 Genomes+HGDP-CEPH Americans), Global AIM panel. **(B)** Map of centroids based on the geometric distribution of the three reference group mid-cluster vertices in plot A. Numbers denote ratios of approximate admixture proportions for each centroid. One PEL sample indicated is closest to the centroid of 25% African, 50% American, 25% European admixture proportions.

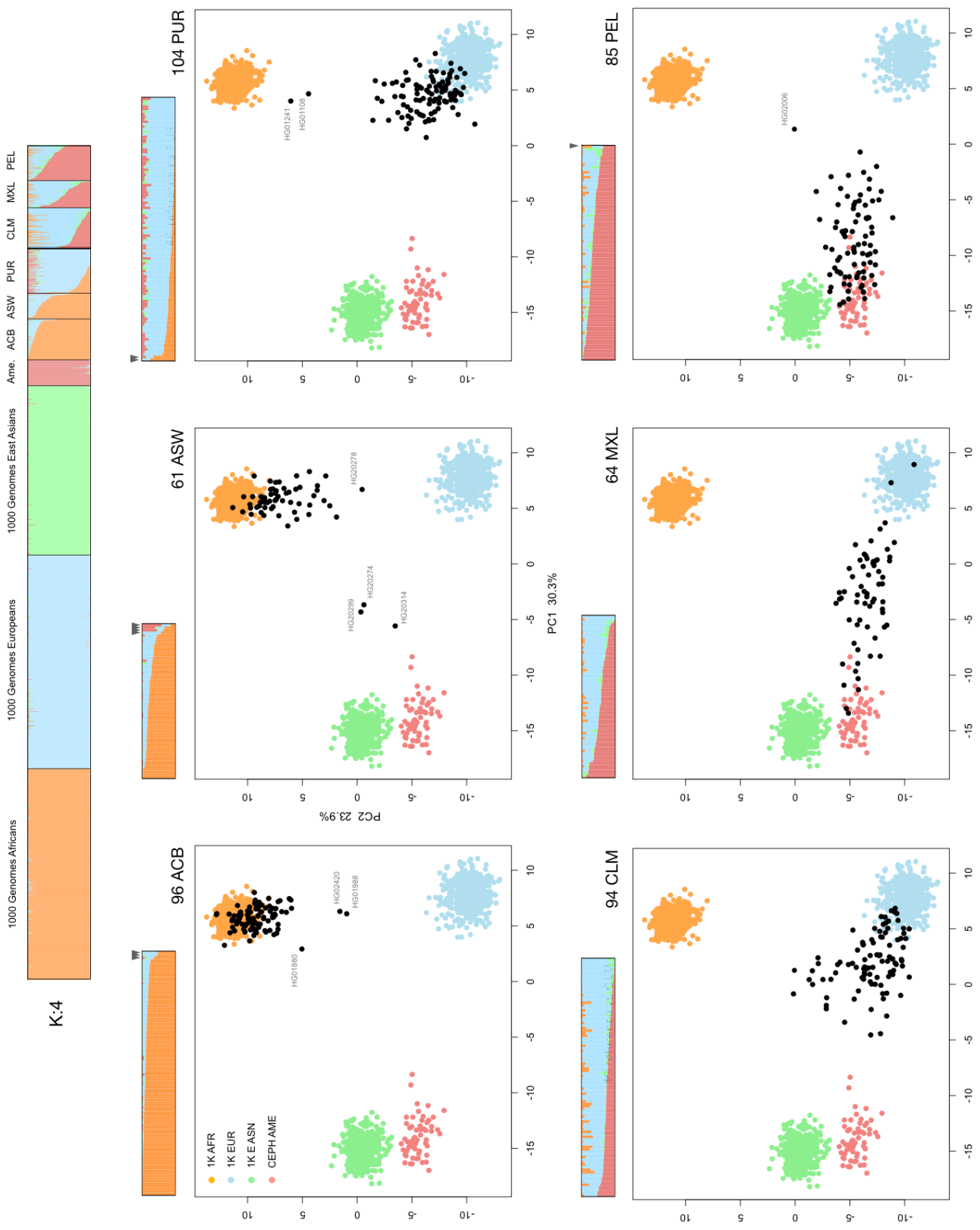


Figure 7. STRUCTURE and PCA analysis with 128 Global AIM-SNPs of the six admixed populations of the 1000 Genomes final variant catalog. Upper cluster plot shows all populations (East Asians exclude 1000 Genomes CHS population), individual plots show admixed populations in more detail, ranked by decreasing African or American cluster memberships. Outlier points in each PCA are numbered and indicated on the matched cluster plots. Summarising 10-percentile cluster plots (average in each 10% 'bin' of STRUCTURE membership proportions) are shown in Figure 2 for these six American region populations.

6. Beyond binary AIM-SNP panels

6.1. Indels

Indel variation mirrors that of SNPs as they are binary loci that can often provide ancestry information. Indels keep the simplicity, multiplexing scale and capacity for very short amplicon PCR of SNPs. The Marshfield linkage marker sets [98] include extensive numbers of short binary indels and several AIM-indel panels were sourced from these sets. In order of publication date, studies are Santos et al., 2010: 48 indels, three multiplexes [99,100]; Pereira et al., 2012: 48 indels, one multiplex [50]; and Zaumsegel et al., 2013: 21 indels, one multiplex [101]. Although AIM-indels are not as informative as the best AIM-SNPs, they utilise the same dye-linked primer system applied to identification indels [102]. Forensic SNP genotyping with SNaPshot does not efficiently distinguish the peak height skews of heterozygotes from patterns seen in mixed DNA. (see **Box 5**) All three AIM-indel assays detect dye-labeled PCR products sent directly from PCR amplifications to capillary electrophoresis (PCR-to-CE). Hence, peak pairs within any one locus are much more balanced and mixtures can be identified from imbalanced signal ratios [50,102]. Ability to detect mixed DNA is an important consideration for forensic ancestry tests as individuals with co-ancestry can be indistinguishable from mixed DNA genotype patterns. It is noteworthy that mixed DNA sample 'M' shown in the Figure 5 PCA, is positioned halfway between two clusters corresponding to the ancestries of the samples combined, which mimics admixture. Pereira's 46-plex AIM-indel panel has equivalent forensic sensitivity to the 38-plex ID-indel test from the same group [102] and gives comparable $\boldsymbol{I_n 3}$ Divergence (Africa–Europe–East Asia) to the 34-plex SNPs, while adding differentiation of Native Americans. Therefore, this panel provides a simple option for laboratories interested in forensic ancestry inference from a single test, with the key feature of detecting mixed DNA. The SPSmart browser lists HGDP-CEPH 46-plex genotypes using the same framework as SNPs (http://spsmart.cesga.es/search.php? dataSet=forindel46) and Snipper includes HGDP-CEPH training sets as stand-alone data or combined with 34-plex. In each case the allele description format is A=short, C=long and G=third alleles.

6.2. Autosomal STRs

Two approaches can be used for ancestry inference with autosomal STRs: applying a large panel of existing markers or adopting specialist STRs with strong population differentiation. The study in 2003 by Rosenberg et al. [44] looked in detail at the 377 STRs used by the same group to analyse worldwide population structure [14] and compared their ancestry informativeness to SNPs. Rosenberg's key findings were that randomly chosen STRs were more informative for ancestry than random SNPs and a greater proportion of STRs were considered highly informative compared to SNPs. This is not surprising; given the original 377 STRs had so effectively identified the principal genetic clusters. However, the right hand tail of the distribution of SNP $I_n 3$ values crossed those of STRs, so finding and developing the most population-differentiated SNPs is the best approach for building the most

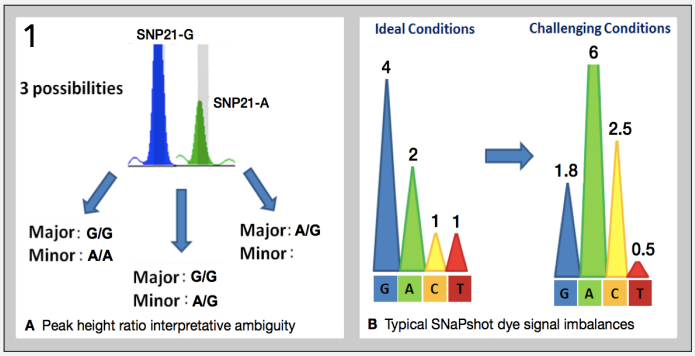
ancestry-informative panels. Another finding with consequences for assessment of forensic STRs as AIMs, was that di-nucleotide repeat STRs were much more differentiated across population groups than tri-/tetra-nucleotide repeat loci. Di-nucleotide STRs are impractical for forensic use but established STRs are unlikely to provide the best information for ancestry inference. Despite these results, it is important to explore how effectively core STRs can infer ancestry as DNA profiling data is generated in almost all routine forensic tests.

Other studies have assessed STR ancestry-informativeness since Lowe's study [80], including: Londin et al. in 2010 [103], **Phillips et al.** in 2011 [104] and Pereira et al. in 2012 [105]. Londin assessed the ancestry informativeness of *Identifiler* plus four other STRs but failed to differentiate a global sample set (7 groups including Middle East). Consequently, the 19 were replaced with 36 novel STRs, 33 being dinucleotide-repeat STRs, confirming Rosenberg's findings about these loci [44]. **Phillips** assessed 15 *Identifiler* and 5 Extended-ESS STRs with the HGDP-CEPH sample set, using STRUCTURE to gauge these STR's ability to infer ancestry (the HGDP-CEPH set excluded Middle East/Central South Asians). **Phillips** used STRUCTURE membership coefficients to accomplish ancestry assignments, as Snipper did not then handle multi-allele data. Average membership proportions and cluster plots indicated genetic data from 20 STRs could differentiate most HGDP-CEPH samples into four groups, with Oceanians only formed a fifth cluster at K:5 when 34-plex SNPs were added to the analysis. Although the study compared Identifiler and ESS 15-STR sets, the lowest assignment error rates for five group comparisons were ~15% with 20 STRs. This ancestry inference performance is not particularly encouraging but several positive outcomes need mentioning. First, Snipper was modified to accommodate STR profiles by using frequency-based custom training set input (http://mathgene.usc.es/snipper/frequencies_new.html) with HGDP-CEPH frequencies generated from the study and now listed in a dedicated STR browser called pop.STR: http://spsmart.cesga.es/popstr.php). Second, the assignment error rate dropped to 4–10% for a four group comparison by assigning ancestry based on membership coefficients greater than 0.5. Lastly, combining 34-plex SNP plus 20 STR genotypes led to all samples in the reduced HGDP-CEPH set being successfully assigned, improving the performance of SNPs alone. In the third study of STRs as AIMs, Pereira used a very large dataset of 54,000 17-STR profiles for three, five and seven regional divisions. Despite the size of the database there were certain problems: only about 1.5% of the profiles were African and 90% of profiles lacked Penta D/E genotypes. Nevertheless, the data was used to train a machine learning system based on decision tables and Bayes analysis producing ~14% error in three region comparisons (i.e., the three main population groups) – comparable to that found in [104]. The machine learning system was placed in a web-based calculator: *Pop Affiliator*, where genotypes can be input for each STR and assignment probabilities returned. It is not clear from [105] what the output probabilities mean, but they appear to be akin to STRUCTURE membership coefficients, so values below 50% suggest non-assignment and if close to this value are likely to be unreliable indicators of ancestry. The *Pop Affiliator* site has recently been upgraded (http://cracs.fc.up.pt/~nf/popaffiliator2) with modified choices of three or five group comparisons (see **Box 6**).

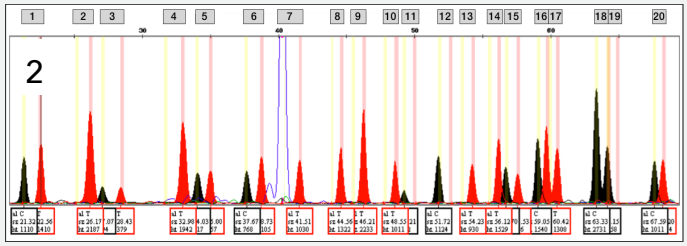
Alternatively, *Snipper* offers Bayes analysis of allele frequency data identical to the algorithm for binary SNPs/indels (http:// mathgene.usc.es/snipper/frequencies_new.html). A 32 STR frequency-based training set template file is provided that is adaptable to cover the combinations of recently expanded STR sets such as Life Technologies' *GlobalFiler*, Promega *Fusion* and Qiagen *HDplex* (the latter two combined providing the 32 non-overlapping STRs listed in *Snipper*). In a recent STR review by **Phillips, et al, 2014** [106] the expanded 32-STR dataset was formally evaluated for ancestry inference performance using *Snipper* STR frequency input and gives much improved ancestry inference rates. For the same reduced HGDP-CEPH sample set used to assess 20 STRs in [104], error rates were 0.8% for Americans and East Asians; 0.6% for Europeans; and 1.9% for Africans. However, applying an LR threshold of 100 led to just one American sample misclassifying and the reasonable non-classification rate of 5–15% sub-threshold probabilities (Fig. 6, [106]). The review of **Phillips, et al,** also highlighted presence of population-specific alleles in certain STRs (Fig. 5, [106]), with the most marked specificity found in the 9-repeat allele of D9S1120 by **Phillips, et al** [107]. This STR differentiates 53% of Native Americans, making it worth consideration by forensic laboratories in the Americas. Unfortunately, other instances of population specificity are less frequent and informative, comprising D18S51-16.2 to 19.2 alleles (6% of Africans); Penta D-2.2 (22% of Europeans); Penta D-3.2 (8% of East Asians); and D21S2055-19.1 (25% of Europeans) - as described in **Box 6**. Finally, novel ancestry-informative tetra-nucleotide repeat STRs were developed by **Phillips, et al, 2013** [108] combined in a 12-plex assay. Ancestry inference performance was good for all groups (assessed with the reduced HGDP-CEPH set) when combined with 20 established STRs, but showed poorer success in Africans: error rates were typically 2–8%, but reached 18% for African assignments.

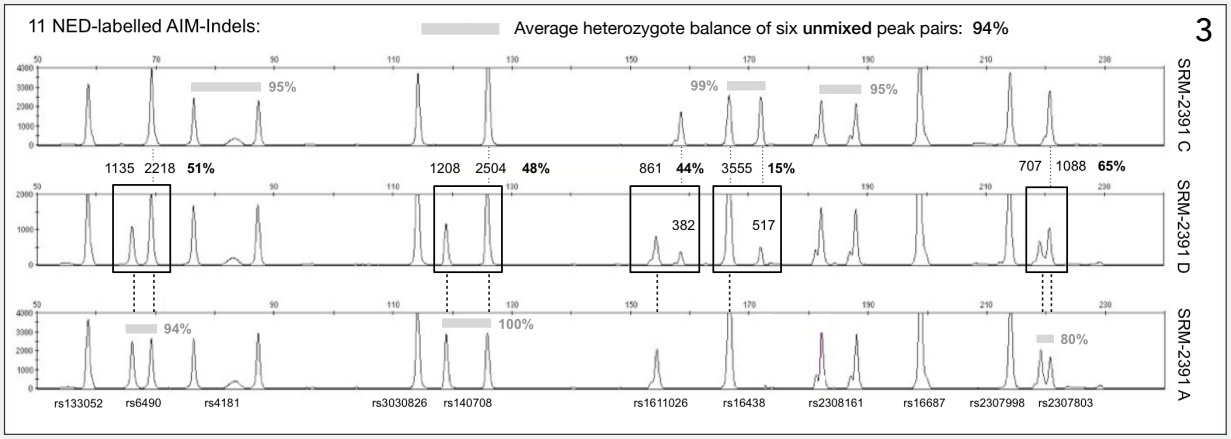
Box 5. Detection of mixed DNA with SNPs and Indels

Although SNaPshot assays allow straightforward detection of SNPs using a forensic lab's well-validated CE regimes, its ability to detect mixed DNA is hampered by **significant signal peak imbalances** in normal DNA that can mask indications of a mixture. Since most SNPs are binary loci, they cannot indicate mixed DNA by the presence of more than two extension product peaks, unless multiple-allele SNPs such as rs4540055 and rs5030240 are included in the multiplex (e.g. in the 34-plex AIMs). SNaPshot signal imbalances arise from three factors: (1) **Stochastic effects** from the need for two separate cycling reactions of an initial capture PCR and subsequent single base extension to label the SNP site with the dyes detecting the allele(s) present; (2) Slight **differences in extension efficiency** between the alleles in any one SNP (these can indicate clustering SNPs where a non-consensus allele may be linked to one of the target SNP alleles); (3) **Marked disparities in signal intensity between SNaPshot dyes**. This last factor is by far the most influential and allows a degree of predictability and reproducibility in the expected signal ratios, permitting the application of interpretative rules for any one peak combination. So, with optimum reaction conditions and DNA, signal ratios follow a **4:2:1:1 ratio** for dyes associated with G:A:C:T alleles, as shown in **Panel 1**. However, signal ratios become accentuated and much less predictable when analysing low-level DNA, typical of forensic material. Therefore, detecting mixtures with SNaPshot is not a reliable process and in the context of forensic ancestry inference the raised heterozygosity seen in mixed DNA give identical patterns to those seen for AIM-SNPs in individuals with co-ancestry due to admixture. SNaPshot profiles with a large number of heterozygotes and imbalanced peak height ratios (PHRs) can be interpreted as typical of mixed DNA or equally, could indicate a pattern of poor quality DNA from an individual that has co-ancestry.



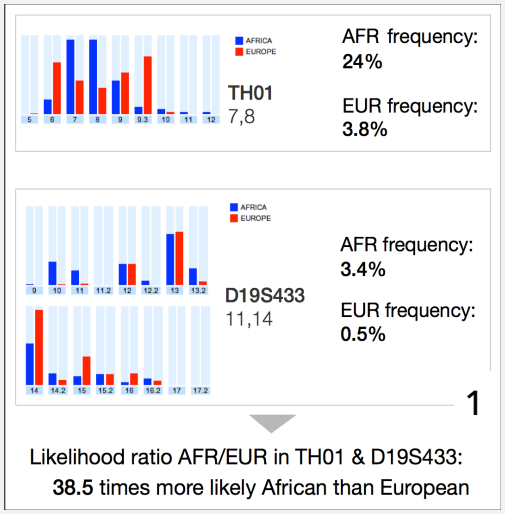
Two solutions can be used to reduce signal imbalance in CE-based detection: (1) To develop multiplexes of **CT-allele SNPs** where the red and yellow dye labels used have much more balanced relative signal strengths per peak pair (example 20-plex in **Panel 2**); (2) To develop sets of **Indels** where CE separation is based on the presence/absence of short (1-5 bp) insertions. Using the **same dye** label for the PCR primers of any one Indel creates allelic products with **identical signal strengths** (and one cycling reaction helps to control stochastic effects). An initial multiplex of 38 ID-Indels led to much better balanced PHRs than those of SNaPshot and proven capability to detect mixed DNA. The simplicity of this direct **PCR-to-CE** system was then extended to a single PCR multiplex of 46 ancestry informative Indels. This test has proved to be robust, sensitive and an efficient system to detect mixtures. The overall level of population differentiation obtained is similar to that of the 34-plex SNP test, but improved for AME. Subsequent experiments with artificial mixtures such as NIST's SRM 2391-D (A:C combined 3:1) underline the efficiency of Indels to distinguish imbalanced peak pairs resulting from two components combined compared to normal PHR variation. The average **PHR of 94%** seen in the NED-labelled Indel peak pairs of **unmixed 2391-A/C DNAs** is shown in **Panel 3** is (i.e. lowest peaks are ≥94% of the highest, with rs2307803 an outlier at 80%). This contrasts with a **PHR range of 15%-65%** in five peak pairs of the **mixed DNA 2391-D** profile in the centre. Most recently, the sensitivity of Indels to mixed DNA was confirmed by the fact that all 19 participants of the EDNAP AIMs exercise detected the 3:1 mixture forming one of the six controls, in addition to a very high overall genotyping accuracy of **99.8%**.



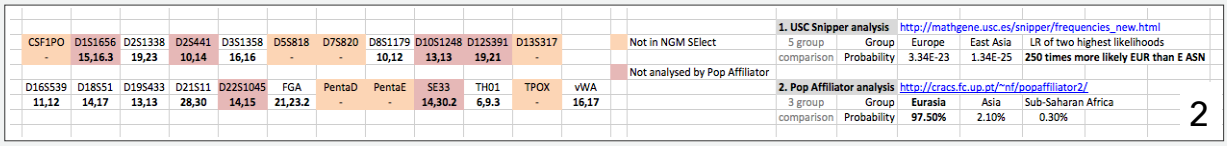


Box 6. How informative are STR genotypes for the inference of ancestry?

As STR profiling is successfully completed in nearly all forensic DNA analyses, it is important to assess how well micro-satellite allelic variation can aid ancestry inference in its own right. It is also often the case that **the only DNA data available is an STR profile** and it can be asked if these genotypes alone provide investigative information about a likely donor ancestry when extra tests are not possible.

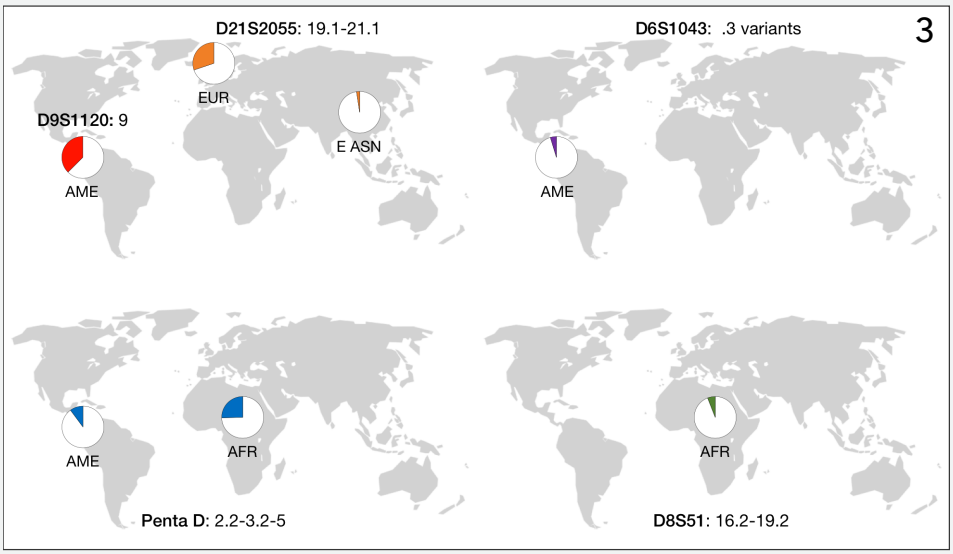


The principle of likelihood ratio comparisons, used routinely in SNP-based ancestry inference, evolved from STR profile frequency calculations made for different populations relevant to a lab's surrounding region and included in most identification reports given to court. If a profile's STR alleles show strong differences in frequency these begin to accumulate as increasing contrasts in the profile frequency for each population compared. This principle is illustrated in **Panel 1**, where the Minstead suspect, Delroy Grant's TH01 and D19S433 alleles are listed with AFR and EUR frequencies (see ***Box 1***). When combined, these produce an LR of **38.5 times more likely African** then European for the DNA donor. Therefore without SNP analysis there was already statistical indications of Grant's ancestry. This process has been explored in several ways but the most practical current systems for analysing forensic STR profiles are provided by a frequency-based analyser in *Snipper* and an alternative online STR analyser: *Pop Affiliator*. The *Snipper* system compares a single profile to a user-uploaded file of allele frequencies for each STR - data obtainable for populations or their groups directly from the *pop.STR* browser (*pop.STR* charts shown in **Panel 1**). Then, LRs are prepared in identical fashion to SNP analyses and the highest LR denotes the most likely population of origin. The *Pop Affiliator* system is **fixed to the Promega ESX-17 loci** although users can opt for a 3- or 5-group comparison (Eurasia-Asia-Africa or Eurasia-Asia-Africa-Near East-North Africa). Results are given as percentage probabilities and based on a machine learning algorithm that accesses a very large database of STR allele frequency distributions from the five regions described above.



In the casework example given in **Panel 2**, an unidentified skeleton found at a waste site was typed with ABI *NGM SElect*. This full profile can be analysed with *Snipper* (*pop.STR* lists HGDP-CEPH data for all current CE- or MPS-based forensic STRs) and this provides an LR of **250 times more likely European** than East Asian ancestry for the unidentified individual. Two problems arise with *Pop Affiliator*: six STRs in *NGM SElect* have no reference data, diminishing the analysis scope; likewise, six STRs with population data in *Pop Affiliator* must be entered as blanks as they are not part of this multiplex. Despite the limitations of fixing the STRs that can analysed, the inference is the same, with a **97.5% probability to be Eurasian** in origin. In practice, each system can be utilised and their ancestry inferences compared for any one profile, although as STR data expands (e.g. ABI *GlobalFiler* types 23 STRs) *Pop Affiliator's* usefulness becomes further limited.

Lastly, rare instances, of **population-specific STR alleles** can add very high individual likelihoods to a particular ancestry. In addition to the well-documented example of **D9S1120 9-repeat** with American specificity, four other less informative but useful STR allele frequency distributions are shown in **Panel 3**. These are not always specific to one group and D21S2055 is not a core STR; however they illustrate the increased ancestry informativeness possible when expanding the range of STRs applied.



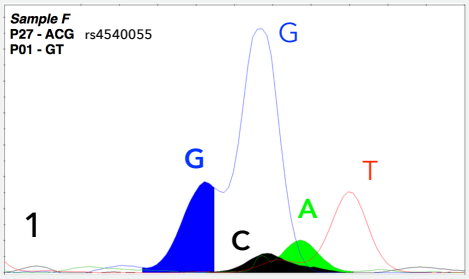
6.3. Microhaplotypes and multiple-allele SNPs

NGS will improve forensic ancestry analysis in other ways besides enlarging SNP multiplexes to increase an AIM panel's informativeness. Massively parallel sequencing of short fragments genotypes all other SNPs amplified alongside the targeted variant. Therefore, SNPs embedded in STRs, as well as multiple SNPs forming haplotypes are genotyped simultaneously and many show ancestry-informative allele distributions. Kidd's group have been the first to identify and catalog haplotypes of potential use in forensic analysis, terming them: Minihaplotypes (1–10 kilobase spans) and microhaplotypes (≤200 bp) [109,110]. Since these show loose and tight physical linkage respectively, the key to finding ancestry-informative haplotypes is careful gauging of recombination rates in the region of interest. Although very low recombination rates help preserve SNP combinations across kilobase spans, some recombination is required to generate informative haplotype frequencies amongst populations. Likewise, very short spans need recombination activity to generate new allele combinations. Two examples illustrate typical informative haplotypes: a 3-SNP Minihaplotype in PAH (Fig. 1, [110]), and a 3-SNP microhaplotype in EDAR (Fig. 4, [110]). The PAH rs869916–rs1722383–rs1042503 haplotype spans 2687 bp with average haplotype heterozygosity (AHH) of 0.51, with GAA a high frequency haplotype in East Asians. The EDAR rs260694–rs11123719–rs11691107 haplotype spans 125 bp with AHH = 0.41, but with informative haplotypes in several populations (GTC: Africans; TCC; East Asians, Americans; TTC: Eurasians; TTT: Africans, Oceanians). Lastly, it is worth noting that autosomal SNP haplotypes will be highly informative for identifying lineage groups within populations identical by descent across many loci, potentially aiding familial searching and complex kinship analysis as well as improving geographic resolution.

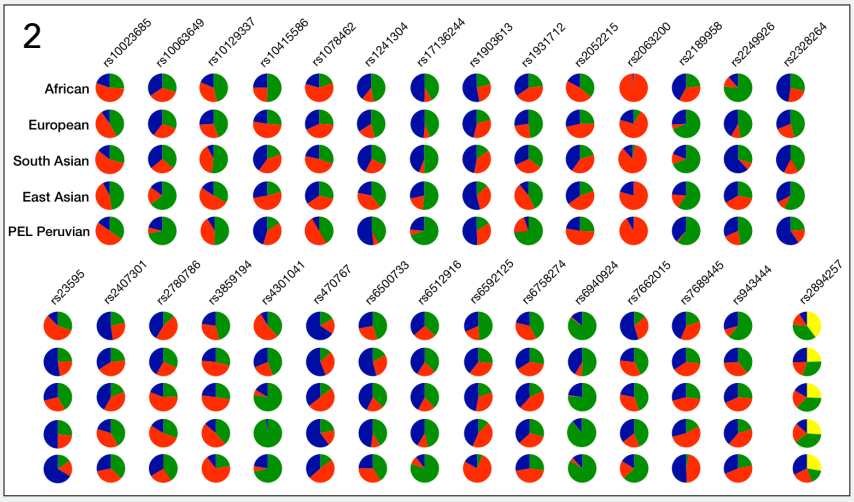
Multiple-allele SNPs were initially considered rare or anachronistic, then went undetected by genome-wide SNP arrays used by HapMap and were removed from 1000 Genomes first variant catalog. Now they have been fully characterised and make up 1 in 300 of the final catalog (259,370 of 78,136,341 variants). Two tri-allelic SNPs are in the 34-plex set as they show marked population differentiation while providing the means to detect third alleles in simple mixed DNA (Fig. 6, [58]). The Global AIM panel includes 6/128 tri-allelic SNPs, adding mixture detection capabilities to NGS tests. This useful feature also motivated the study by Westen et al. in 2009 [111] that developed 16 tri-allelic SNPs, several showing high African–European differentiation. Therefore, many multiple-allele SNPs will have high ancestry informativeness through increased opportunity for drift to influence the geographic distributions of six or ten genotypes (in tetra-allelic SNPs) compared to binary loci. The best forensic multiple-allele SNPs are shown in **Box 7**.

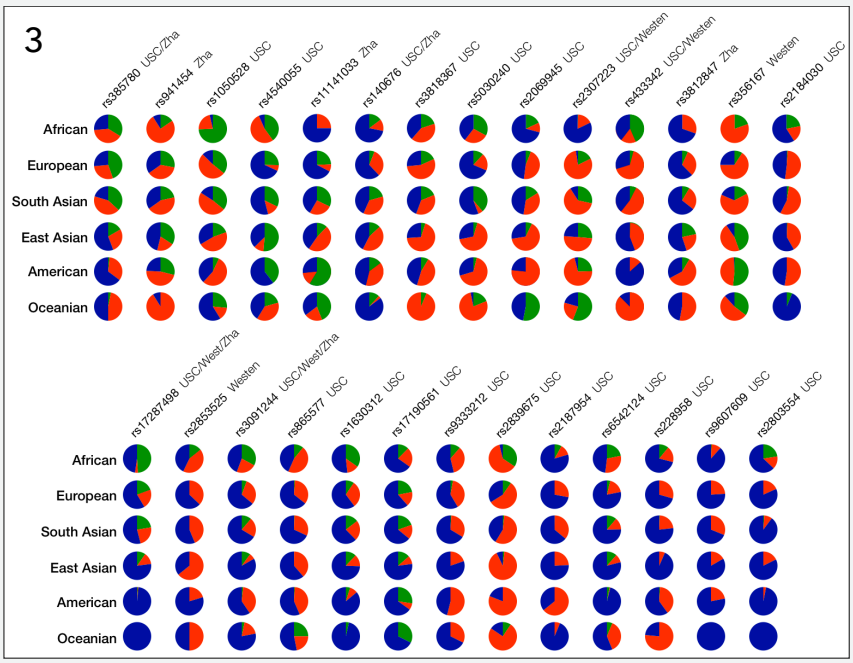
Box 7. Multiple-allele SNPs as ancestry-informative markers

Following the chance discovery of a SNP with three common alleles (rs385780), the author began to compile examples of multiple nucleotide substitutions at single SNP sites. Nine tri-allelic SNPs with high levels of polymorphism were reported at ISFG Arcachon 2003, these loci had been characterised by using SNaPshot multiplexes tested with artificially created mixed DNA. Follow-up studies at USC with Sequenom iPLEX, also using single base extension chemistry, validated a total of 19 SNPs by genotyping the HGDP-CEPH sample panel. The original motivation for developing multiple-allele SNPs for forensic analysis was to provide a **better mixture detection** system than was possible by comparing peak height ratios in SNaPshot electropherograms which tend to show irregular skews in signal strength (see *Box 5*). **Panel 1** shows the patterns obtained with 34-plex tri-allelic SNP rs4540055, genotyping an artificial mixture of two donors at a 3:1 ratio. However, a potentially more useful property of multiple-allele SNPs is **highly informative contrasts in allele frequencies amongst populations** seen in a large proportion of loci and often differentiating populations with much lower divergence in their binary SNP variation. This pattern of variation is likely to be due in part to the increased opportunity for genetic drift to occur when SNPs have twice the genotypes. This feature creates more opportunity for random changes to occur between diverging populations.



Two factors have favoured the forensic development of multiple-allele SNPs: the straightforward detection of their alleles by CE using SNaPshot and the recent emergence of re-sequencing to characterise human SNP variation rather than hybridisation-based arrays (i.e. the whole-genome scan arrays used by Hapmap). Re-sequencing detects multiple-allele SNPs when arrays cannot, as they depend on pairs of allele-specific capture probes so **a third allele is unanticipated**. The final variant catalog of 1000 Genomes lists 508,917 multiple-allele SNPs amongst a total of 77 million SNPs (1 in 152). The author has collected the most polymorphic 1000 Genomes multiple-allele SNPs and constructed a 29-plex of the best loci, suitable for **identity and mixture detection** (**Panel 2**). Note that a single tetra-allelic SNP, rs2894257, was also included. Previously, 27 of the most informative multiple-allele SNPs for **ancestry analysis** were compiled and genotyped with the HGDP-CEPH panel. Some of these loci were also identified in two published studies dedicated to the development of forensic mixed DNA detection panels (*Westen et al., 2009, Forensic Sci. Int. Genet. 3: 233-241* and *Zha et al., 2012, Electrophoresis, 33: 841-848*) and these are detailed in **Panel 3**.





Since multiple-allele SNPs give much more discrimination power per locus than binary SNPs, they have obvious application for kinship analyses when ability to successfully genotype highly degraded DNA is advantageous, e.g. identification of missing persons. With this goal, the author is now developing a panel of **several hundred multiple-allele SNPs** to create large-scale multiplexes applicable to massively parallel sequencing (MPS) systems suitable for forensic analysis. These will be applied to the challenging analysis of the missing in regions of conflict.

Objectives

1. To compile SNP allele frequency data from online sources for a range of worldwide populations. To curate this data as SNP database depth and coverage progresses from the initial map of 1.42 million human SNPs. To build stand-alone allele frequency browsers for HapMap, Perlegen, CEPH Foundation and 1000 Genomes SNP data that allow full genotype downloads for large numbers of SNPs and/or populations.

2. To evaluate SNaPshot as a forensic SNP genotyping system from its initial availability in 2001. To develop novel SNP sets specifically for forensic use, comprising: ancestry-informative SNPs (AIM-SNPs); identification SNPs (ID-SNPs); coding-region SNPs for prediction of common-variation externally visible characteristics (EVC-SNPs); mitochondrial DNA variants (mt-SNPs); and Y-chromosome SNPs (Y-SNPs).

3. To develop a complete forensic ancestry inference solution based on AIM-SNP typing with established forensic capillary electrophoresis platforms, comprising: a highly multiplexed PCR and SNaPshot assay; open-access genotype analysis tools using Bayes likelihood calculations (*Snipper*); comprehensive population surveys for the SNPs of the test (the SNP*for*ID SPSmart pages).

4. To enhance the 34-plex AIM-SNP assay, routinely applied to the differentiation of Europeans, Africans and East Asians, by developing novel SNP sets designed to be run alongside the core 34 SNPs that focus on South Asian, Native American and Oceanian populations-of-origin (*Eurasiaplex*, *PIMA* and *Pacifiplex*). To develop additional sets focussed on improved SNaPshot peak balance (CT-only SNPs); European-African co-ancestry patterns in admixed individuals (*Admixplex*) and X-chromosome AIM-SNPs.

5. To analyse the sensitivity of SNP tests optimised for forensic use by assessing genotyping performance of the PCR multiplex with a wide range of skeletal material. To apply the SNP test assessments to crime-scene contact trace analysis in order to enhance investigative leads. To promote the concept of ancestry inference in a forensic context to progress criminal investigations lacking database hits or eye-witness. To progress cold-case review approaches by using ancestry inference data.

6. To identify and collect population variation data and genomic details of non-binary SNPs, comprising tri-allelic and tetra-allelic single nucleotide variation with multiple base substitutions detected by whole genome re-sequencing. To develop a panel of multiple allele SNPs for MPS of 250-300 markers obtained from screening the complete human variant catalog published by 1000 Genomes.

7. To optimise and extend the practicality of ancestry analysis using Indels and STRs to establish mixed-marker approaches that allow ancestry inference from standard DNA profiling data (when

evidential material is no longer available for additional DNA tests), or when a more secure system is required for the analysis of mixed DNA than is possible with SNaPshot SNP genotyping. To develop a frequency-based classifier in Snipper applicable to forensic STR data, but also extending the scope of ancestry inference to haplotype data such as Y-SNPs and autosomal SNP microhaplotypes. To enhance *Snipper* with fixed training set data for 46 Indels alongside 34 SNPs and to extend SPSmart to forensic Indel sets.

8. To rebuild smaller AIM sets into enlarged single 130 to 160-plex panels applicable to compact massively parallel sequencing (MPS) platforms, increasingly being adopted for forensic DNA analysis. To validate the resulting SNP panels in terms of sequence balance, sensitivity and genotyping concordance.

9. To identify and catalog the collateral variants associated with the core STR and SNP markers genotyped by forensic MPS assays. To analyse their ancestry-informative potential starting with the HGDP-CEPH panel genotyped by the Illumina MiSeq-based Forenseq system.

10. To build sets of Microhaplotype loci, comprising sets of closely linked SNP alleles in haplotype combinations, that show strongly population-differentiated patterns of variation. To develop forensic MPS assays that comprise panels of Microhaplotype loci informative for ancestry or applicable to identification (e.g. missing person identification) from amplification of Microhaplotypes in very short fragment lengths.

Thesis papers

1. Building a forensic ancestry panel from the ground up: The EUROFORGEN Global AIM-SNP set.

Reference: Forensic Sci Int Genet (2014) 11:13-25

ISSN: 1872-4973

DOI: 10.1016/j.fsigen.2014.02.012

Web link: http://www.fsigenetics.com/article/S1872-4973(14)00040-4/fulltext

Authors: Phillips C[(1)], Parson W[(2)], Lundsberg B[(3)], Santos C[(4)], Freire-Aradas A[(4)],Torres M[(5)], Eduardoff M[(6)], Børsting C[(3)], Johansen P[(3)], Fondevila M[(4)], Morling N[(3)], Schneider P[(7)]; EUROFORGEN-NoE Consortium, Carracedo A[(8)], Lareu MV[(4)].

Author information:

(1) Forensic Genetics Unit, Institute of Legal Medicine, Faculty of Medicine,University of Santiago de Compostela, ES-15705 Santiago de Compostela, Galicia, Spain. Electronic address: c.phillips@mac.com.

(2) Institute of Legal Medicine, Innsbruck Medical University, Müllerstrasse 44, A-6020 Innsbruck, Austria; Penn State Eberly College of Science, University Park, PA, USA.

(3) Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Frederik V's Vej 11, DK-2100 Copenhagen, Denmark.

(4) Forensic Genetics Unit, Institute of Legal Medicine, Faculty of Medicine, University of Santiago de Compostela, ES-15705 Santiago de Compostela, Galicia, Spain.

(5) Grupo de Medicina Xenómica (GMX), Faculty of Medicine, University of Santiago de Compostela, Spain.

(6) Institute of Legal Medicine, Innsbruck Medical University, Müllerstrasse 44, A-6020 Innsbruck, Austria.

(7) Institute of Legal Medicine, University Hospital Cologne, D-50823 Cologne, Germany.

(8) Forensic Genetics Unit, Institute of Legal Medicine, Faculty of Medicine, University of Santiago de Compostela, ES-15705 Santiago de Compostela, Galicia,Spain; Grupo de Medicina Xenómica (GMX), Faculty of Medicine, University of Santiago de Compostela, Spain; Center of Excellence in Genomic Medicine Research, King Abdulaziz University, Jeddah, Saudi Arabia.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

2. Tetra-allelic SNPs: Informative forensic markers compiled from public whole-genome sequence data.

Reference: Forensic Sci Int Genet. (2015) 19:100-106

ISSN: 1872-4973

DOI: 10.1016/j.fsigen.2015.06.011

Web link: http://www.fsigenetics.com/article/S1872-4973(15)30027-2/fulltext

Authors: Phillips C[(1)], Amigo J[(2)], Carracedo Á[(3)], Lareu MV[(4)].

Author information:

(1) Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Spain. Electronic address: c.phillips@mac.com.

(2) Grupo de Medicina Xenómica (GMX), Faculty of Medicine, University of Santiago de Compostela, Spain.

(3) Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Spain; Grupo de Medicina Xenómica (GMX), Faculty of Medicine, University of Santiago de Compostela, Spain; Center of Excellence in Genomic Medicine Research, King Abdulaziz University, Jeddah, Saudi Arabia.

(4) Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiagode Compostela, Spain.; Center of Excellence in Genomic Medicine Research, King Abdulaziz University, Jeddah, Saudi Arabia.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

3. Revision of the SNP*for*ID 34-plex forensic ancestry test: Assay enhancements, standard reference sample genotypes and extended population studies.

Reference: Forensic Sci Int Genet. (2013) 7(1):63-74

ISSN: 1872-4973

DOI: 10.1016/j.fsigen.2012.06.007

Web link: http://www.fsigenetics.com/article/S1872-4973(12)00140-8/fulltext

Authors: Fondevila M[(1)], Phillips C[(1)], Santos C[(1)], Freire-Aradas A[(1)], Vallone PM[(2)], Butler JM[(2)], Lareu MV[(1)], Carracedo Á[(1)].

Author information:

(1) Forensic Genetics Unit, Institute of Legal Medicine, University of Santiago de Compostela, Spain

(2) U.S. National Institute of Standards and Technology, Biochemical Science Division, Gaithersburg, MD, USA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

4. Nonbinary single-nucleotide polymorphism markers.

Reference: International Congress Series (2004) 1261:27-29

ISSN: 0531-5131

DOI: 10.1016/j.ics.2003.12.008

Web link: http://www.sciencedirect.com/science/article/pii/S0531513103019484

Authors: Phillips C[(1)], Lareu MV[(1)], Salas A[(1)], Carracedo Á[(1)].

Author information:

(1) Institute of Legal Medicine, University of Santiago de Compostela, Calle San Francisco s/n 15705 Santiago de Compostela, Galicia, Spain.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

5. Forensic ancestry analysis with two capillary electrophoresis ancestry informative marker (AIM) panels: Results of a collaborative EDNAP exercise.

Reference: Forensic Sci Int Genet (2015) 19:56-67

ISSN: 1872-4973

DOI: 10.1016/j.fsigen.2015.06.004

Web link: http://www.fsigenetics.com/article/S1872-4973(15)30024-7/fulltext

Authors: Santos C[(1)], Fondevila M[(1)], Ballard D[(2)], Banemann R[(3)], Bento AM[(4)], Børsting C[(5)], Branicki W[(6)], Brisighelli F[(7)], Burrington M[(8)], Capal T[(9)], Chaitanya L[(10)], Daniel R[(11)], Decroyer V[(12)], England R[(13)], Gettings KB[(14)], Gross TE[(15)], Haas C[(16)], Harteveld J[(17)], Hoff-Olsen P[(18)], Hoffmann A[(3)], Kayser M[(10)], Kohler P[(18)], Linacre A[(19)], Mayr-Eduardoff M[(20)], McGovern C[(13)], Morling N[(21)], O'Donnell G[(8)], Parson W[(22)], Pascali VL[(7)], Porto MJ[(4)], Roseth A[(18)], Schneider PM[(15)], Sijen T[(17)], Stenzl V[(9)], Court DS[(2)], Templeton JE[(19)], Turanska M[(23)], Vallone PM[(14)], van Oorschot RA[(11)], Zatkalikova L[(23)], Carracedo Á[(1)], Phillips C[(1)]; EUROFORGEN-NoE Consortium.

Author information:

(1) Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Santiago de Compostela, Spain.

(2) Department of Forensic and Analytical Science, Faculty of Life Science, King's College London, UK.

(3) Federal Criminal Police Office, Wiesbaden, Germany.

(4) Forensic Genetic and Biology Service, Centre Branch, National Institute of Legal Medicine and Forensic Sciences, Coimbra, Portugal.

(5) Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Frederik V's Vej 11, Copenhagen, Denmark.

(6) Section of Forensic Genetics, Institute of Forensic Research, Kraków, Poland.

(7) Forensic Genetics Laboratory, Institute of Legal Medicine, Università Cattolica del Sacro Cuore, Rome, Italy.

(8) Forensic Science Laboratory, Dublin, Ireland.

(9) Department of Forensic Genetics, Institute of Criminalistics, Prague, Czech Republic.

(10) Department of Forensic Molecular Biology, Erasmus MC University Medical Centre Rotterdam, Rotterdam, The Netherlands.

(11) Office of the Chief Forensic Scientist, Forensic Services Department, Victoria Police, Australia.

(12) National Institute of Criminalistics and Criminology, Chaussée de Vilvoorde 100, Brussels, Belgium.

(13) ESR, Private Bag 92021, Auckland, New Zealand.

(14) Material Measurement Laboratory, National Institute of Standards and Technology, Gaithersburg, MD, USA.

(15) Institute of Legal Medicine, Faculty of Medicine, University of Cologne, Cologne, Germany.

(16) Zurich Institute of Forensic Medicine, University of Zurich, Zurich, Switzerland.

(17) Department of Human Biological Traces, Netherlands Forensic Institute, The Hague, The Netherlands.

(18) Department of Forensic Biology, Norwegian Institute of Public Health, Oslo, Norway.

(19) School of Biological Sciences, Flinders University, Adelaide, South Australia 5042, Australia.

(20) Institute of Legal Medicine, Medical University of Innsbruck, Innsbruck, Austria.

(21) Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Frederik V's Vej 11, Copenhagen, Denmark; National Institute of Criminalistics and Criminology, Chaussée de Vilvoorde 100, Brussels, Belgium.

(22) Institute of Legal Medicine, Medical University of Innsbruck, Innsbruck, Austria; Forensic Science Program, The Pennsylvania State University, University Park, PA, USA.

(23) Institute of Forensic Science, Ministry of the Interior, Department of Biology and DNA Analysis, Slovenská Lupca, Slovakia.

6. A 34-plex autosomal SNP single base extension assay for ancestry investigations

Reference: DNA Electrophoresis Protocols for Forensic Genetics. Methods in Molecular Biology (2012) 830:109-126.

ISBN: 978-1-61779-460-5

DOI: 10.1007/978-1-61779-461-2_8

Web link: https://link.springer.com/protocol/10.1007%2F978-1-61779-461-2_8

Authors: Phillips C[(1)], Fondevila M[(1)], Lareu MV[(1)].

Author information:
(1) Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Santiago de Compostela, Spain.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

7. Online resources for SNP analysis: a review and route map.

Reference: Mol Biotechnol (2007) 35(1):65-97

ISSN: 1559-0305

DOI: 10.1385/mb:35:1:65

Web link: https://link.springer.com/article/10.1385/MB:35:1:65

Authors: Phillips C[(1)].

Author information:

(1) The Spanish National Genotyping Centre CeGen, Santiago node, Genomic Medicine Group, University of Santiago de Compostela, Galicia, Spain.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

8. Inference of ancestry in forensic analysis I: autosomal ancestry-informative marker sets.

Reference: Forensic DNA typing protocols. Methods Mol Biol (2016) 1420:233-253

ISBN: 987-1-4939-3595-6

DOI: 10.1007/978-1-4939-3597-0_18

Web link: https://link.springer.com/protocol/10.1007%2F978-1-4939-3597-0_18

Authors: Phillips C[(1)], Santos C[(1)] , Fondevila M[(1)], Carracedo Á[(1)(2)], Lareu MV[(1)]

Author information:

(1) Forensic Genetics Unit, Luis Concheiro Institute of Forensic Sciences, Genomic Medicine Group, University of Santiago de Compostela, Galicia, 15782, Spain.

(2) Center of Excellence in Genomic Medicine Research, King Abdulaziz University, Jeddah, Saudi Arabia.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

9. Ancestry informative markers

Reference: Encyclopedia of Forensic Sciences (2013).

ISBN: 978-0-12-382166-9

Web link: http://www.sciencedirect.com/science/referenceworks/9780123821669 - ancv0035

Authors: Phillips C[(1)].

Author information:
(1) University of Santiago de Compostela, Galicia, Spain.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

10. Application of Autosomal SNPs and Indels in Forensic Analysis.

Reference: Forensic Sci Rev (2012) 24(1):43-62

ISSN: 1042-7201

Authors: Phillips C[(1)].

Author information:
(1) Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Santiago de Compostela, Galicia, Spain.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

11. Ancestry informative markers: inference of ancestry in aged bone samples using an autosomal AIM-Indel multiplex.

Reference: Forensic Sci Int Genet (2015) 16:58-63

ISSN: 1872-4973

DOI: 10.1016/j.fsigen.2014.11.025

Web link: http://www.fsigenetics.com/article/S1872-4973(14)00275-0/fulltext

Authors: Romanini C[(1)], Romero M[(1)], Salado Puerto M[(1)], Catelli L[(1)], Phillips C[(2)], Pereira R[(3)], Gusmão L[(4)], Vullo C[(1)].

Author information:

(1) Forensic DNA Laboratory, Argentinean Forensic Anthropology Team (EAAF) Independencia 644,3A, 5000 Cordoba, Argentina.

(2) Forensic Genetics Unit, Institute of Legal Medicine, Faculty of Medicine, University of Santiago de Compostela, ES-15705 Santiago de Compostela, Galicia, Spain.

(3) Institute of Molecular Pathology and Immunology of the University of Porto (IPATIMUP), Porto, Portugal.

(4) Institute of Molecular Pathology and Immunology of the University of Porto (IPATIMUP), Porto, Portugal; DNA Diagnostic Laboratory (LDD), State University of Rio de Janeiro (UERJ), Rio de Janeiro, Brazil.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

12. Inter-laboratory evaluation of SNP-based forensic identification by massively parallel sequencing using the Ion PGM™.

Reference: Forensic Sci Int Genet (2015) 17:110-121

ISSN: 1872-4973

DOI: 10.1016/j.fsigen.2015.04.007

Web link: http://www.fsigenetics.com/article/S1872-4973(15)00077-0/fulltext

Authors: Eduardoff M[(1)], Santos C[(2)], de la Puente M[(2)], Gross TE[(3)], Fondevila M[(2)], Strobl C[(1)], Sobrino B[(4)], Ballard D[(5)], Schneider PM[(3)], Carracedo Á[(6)], Lareu MV[(2)], Parson W[(7)], Phillips C[(2)].

Author information:

(1) Institute of Legal Medicine, Innsbruck Medical University, Innsbruck, Austria.

(2) Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Santiago de Compostela, Spain.

(3) Institute of Legal Medicine, Faculty of Medicine, University of Cologne, Cologne, Germany.

(4) Grupo de Medicina Xenómica (GMX), Faculty of Medicine, University of Santiago de Compostela, Spain.

(5) Department of Forensic and Analytical Science, Faculty of Life Science, King's College, London, UK.

(6) Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Santiago de Compostela, Spain; Grupo de Medicina Xenómica (GMX), Faculty of Medicine, University of Santiago de Compostela, Spain.

(7)Institute of Legal Medicine, Innsbruck Medical University, Innsbruck, Austria; Forensic Science Program, The Pennsylvania State University, PA, USA.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

13. Pacifiplex: an ancestry-informative SNP panel centred on Australia and the Pacific region.

Reference: Forensic Sci Int Genet (2016) 20:71-80

ISSN: 1872-4973

DOI: 10.1016/j.fsigen.2015.10.003

Web link: http://www.fsigenetics.com/article/S1872-4973(15)30079-X/fulltext

Authors: Santos C[(1)], Phillips C[(1)], Fondevila M[(1)], Daniel R[(2)], van Oorschot RA[(2)], Burchard E[(3)], Schanfield MS[(4)], Souto L[(5)], Uacyisrael J[(6)], Via M[(7)], Carracedo Á[(8)], Lareu MV[(1)].

Author information:

(1) Forensic Genetics Unit, Institute of Legal Medicine, University of Santiago de Compostela, Spain.

(2) Office of the Chief Forensic Scientist, Victoria Police Forensic Services Department, Victoria, Australia.

(3) University of California San Francisco, San Francisco, California, USA.

(4) Department of Forensic Science, George Washington University, Mount Vernon College Campus, Washington, USA.

(5) Department of Biology, University of Aveiro, Aveiro, Portugal.

(6) Fiji Police Forensic Biology and DNA Laboratory, Nasova, Suva, Fiji.

(7) Universitat de Barcelona, Barcelona, Spain.

(8) Forensic Genetics Unit, Institute of Legal Medicine, University of Santiago de Compostela, Spain; Galician Foundation of Genomic Medicine (SERGAS), CIBERER (University of Santiago de Compostela), Sanatiago de Compostela, Spain.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

14. The Global AIMs Nano set: A 31-plex SNaPshot assay of ancestry-informative SNPs.

Reference: Forensic Sci Int Genet (2016) 22:81-88

ISSN: 1872-4973

DOI: 10.1016/j.fsigen.2016.01.015

Web link: http://www.fsigenetics.com/article/S1872-4973(16)30015-1/fulltext

Authors: de la Puente M[(1)], Santos C[(1)], Fondevila M[(1)], Manzo L[(1)], EUROFORGEN-NoE Consortium, Carracedo Á[(2)], Lareu MV[(1)],Phillips C[(1)].

Author information:

(1) Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiagode Compostela, Spain.

(2) Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Spain; Center of Excellence in Genomic Medicine Research, King Abdulaziz University, Jeddah, Saudi Arabia.Medicine (SERGAS), CIBERER (University of Santiago de Compostela), Sanatiago de Compostela, Spain.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

15. Inter-laboratory evaluation of the EUROFORGEN Global ancestry-informative SNP panel by massively parallel sequencing using the Ion PGM™.

Reference: Forensic Sci Int Genet (2016) 23:178-189

ISSN: 1872-4973

DOI: 10.1016/j.fsigen.2016.04.008

Web link: http://www.fsigenetics.com/article/S1872-4973(16)30064-3/fulltext

Authors: Eduardoff M[(1)], Gross TE[(2)], Santos C[(3)], de la Puente M[(3)], Ballard D[(4)], Strobl C[(1)], Børsting C[(5)], Morling N[(5)], Fusco L[(5)], Hussing C[(5)], Egyed B[(6)], Souto L[(7)], Uacyisrael J[(8)], Syndercombe Court D[(4)], Carracedo Á[(9)], Lareu MV[(3)], Schneider PM[(2)], Parson W[(10)], Phillips C[(11)], EUROFORGEN-NoE Consortium.

Author information:

(1) Institute of Legal Medicine, Medical University of Innsbruck, Innsbruck, Austria.

(2) Institute of Legal Medicine, Faculty of Medicine, University of Cologne, Cologne, Germany.

(3) Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Santiago de Compostela, Spain.

(4) Faculty of Life Sciences and Medicine, King's College, London, UK.

(5) Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark.

(6) Department of Genetics, Faculty of Science, Eötvös Loránd University Budapest, Hungary.

(7) Department of Biology, University of Aveiro, Aveiro, Portugal.

(8) Fiji Police Forensic Biology and DNA Laboratory, Nasova, Suva, Fiji.

(9) Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Santiago de Compostela, Spain; Center of Excellence in Genomic Medicine Research, King Abdulaziz University, Jeddah, Saudi Arabia.

(10) Institute of Legal Medicine, Medical University of Innsbruck, Innsbruck, Austria; Forensic Science Program, The Pennsylvania State University, PA, USA.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Discussion

1. To compile SNP allele frequency data from online sources for a range of worldwide populations. To curate this data as SNP database depth and coverage progresses from the initial map of 1.42 million human SNPs. To build stand-alone allele frequency browsers for HapMap, Perlegen, CEPH Foundation and 1000 Genomes SNP data that allow full genotype downloads for large numbers of SNPs and/or populations.

When the first comprehensive human SNP map was published in 2001, as one of the principle research goals of the human genome mapping project (HGMP), it became possible to start accessing this huge data resource for forensic purposes. The two main SNP variant databases that became established based on the genomic information from the HGMP were NCBI´s dbSNP and The SNP Consortium (TSC). While dbSNP was part of a bigger collection of integrated databases at NCBI that sought to compile all relevant genetic information for: scientific publications (Pubmed); DNA sequences (Genbank); genes (Gene); nucleotide variants (dbSNP); and phenotypes for medically important traits or common human variation (OMIM). While initially, TSC formed the main source of human population variation data for the 1.42 million variants of HGMP´s first SNP map. For this reason, the TSC database was of prime importance for the selection of SNPs for forensic use, having sufficient levels of polymorphism in two or more population groups. Therefore, all SNPs initially identified as showing high levels of forensic discrimination in different populations came from TSC allele frequency reports, with the supporting information on context sequence, map position and proximity of genes or other SNPs, from dbSNP. The steps for the handling of SNP genetic data from online resources and the processing of all information relevant to analysing SNPs in medical studies, as well as for forensic use, are reviewed in **Thesis Paper #7, C. Phillips, 2007** [112]**.**

In total, four different human variant databases were used for the first selections of forensic ID-SNP candidates, just ahead of the start of the SNP*for*ID Consortium - a collective with the aim to develop SNP analysis for forensic purposes. These SNP databases were:
(i) The Celera Discovery System database (CDS) comprising 4,802,233 SNPs discovered from the Celera private human genome mapping initiative plus the best validated SNPs from public databases (www.celeradiscoverysystem.com).
(ii) The Applied Biosystems Assays-on-Demand (ABI) database comprising 146,636 SNPs with the best allele frequency information available at the time - derived from 46 Africans and Europeans (http://www.appliedbiosystems.com/products/assays).
(iii) The NCBI dbSNP database comprising 2,243,761 unique SNPs discovered from the major genomics centres that contributed to HGMP (http://www.ncbi.nlm.nih.gov/entrez/).

(iv) The SNP Consortium (TSC) database comprising 1,255,326 SNPs with high quality validation taken from HGMP and dbSNP, comprising loci with allele frequency data from up to three population groups (http://snp.cshl.org/).

The detailed steps taken in the forensic SNP selection processes, accessing the above databases are described in **Phillips, et al, 2004** [113] and **Phillips, et al, 2005** [114].

Access to all four main human SNP databases listed above (principally TSC), allowed identification of almost 350 SNPs that were suitably discriminatory for forensic identification purposes, as they showed reasonably balanced levels of polymorphism comparing European and African allele frequencies. Data from East Asian populations (Han Chinese and Japanese) was added at a much slower pace, so a large proportion of the 350 forensic SNP candidates lacked population variation data for East Asia. A simple rule-of-thumb ensured SNPs had a comparable level of discrimination, within an informative range of values, amongst the two or three major population groups (herein, European=EUR; African=AFR; East Asian=E ASN). Candidate SNPs should have a minimum 30% heterozygosity in one population (0.28 minor allele frequency) and a minimum 20% heterozygosity in all three (0.17 minor allele frequency). These may appear at first sight to be quite low overall levels of polymorphism, but the standard forensic measurement of differentiation of random pairs from the same population: Discrimination Power (Dp) does not fall significantly in value between minor allele frequencies of 0.5 and 0.3. As shown in **Figure D1**, the Dp only drops by 6.5% between these limits, so they are inclusive enough values for the minor allele frequency to provide an unrestricted group of candidate loci for forensic identification use. SNP frequency differences between populations can be highly contrasted at any one locus, but once 24 or more SNPs are collected together, then the final cumulative Dp obtained from all the variant data tends to become balanced between populations.

When compiling markers for the first SNP identification panel, the cumulative Dp in each population group exceeded those of 16 STRs when approximately 30-32 loci had been combined. As a key marker selection criterion was at least one SNP positioned on the p-arm and q-arm of each chromosome, to create a set of 44 loci, the minor allele frequency limits described above could be relaxed sufficiently after 75% of the loci had been compiled to then begin to include SNPs with lower levels of polymorphism in some populations. It also allowed inclusion of a small proportion of SNPs at the end of the selection process with more informative variation in East Asians, where data had not always been available in initial selection from the TSC database. These SNPs were mainly taken from the ABI assays on demand database.

The criteria used to select the final forensic identification panel of 52 SNPs described in **Sanchez, Phillips et al, 2006A** and **2006B** [115,116] from the candidate pool of 350 required detailed analysis of the online databases outlined above and centred on:

(i) amplicon sizes generated from optimum primer designs less than 120 bp; (ii) reported minimum 30% heterozygosity (0.28 minor allele frequency) in at least one population, and minimum 20% heterozygosity (0.17 minor allele frequency) in all three populations - relaxed in the final phase, as described; (iii) a freely assorting marker set using SNPs from the distal parts of the p- and q-arms of each autosome; (iv) a minimum distance of 100 kb between candidate SNPs and neighbouring genes; (v) no likely association with the STR loci most commonly used in forensic analysis; and (vi) unique, high quality (i.e. free from low complexity DNA) flanking sequence that did not have interfering polymorphisms, such as common SNPs/Indels in potential primer binding sites.

Suitable genomic regions were chosen for SNP searches using NCBI Map Viewer (http://www.ncbi.nlm.nih.gov/mapview/maps.cgi/) with alignments of the gene and variation maps following the general guidelines described by **Phillips, 2005** [114]. Three searches were performed using dbSNP builds 112 (p-arm loci), 115 (q-arm loci) and 118 (supplementary loci).

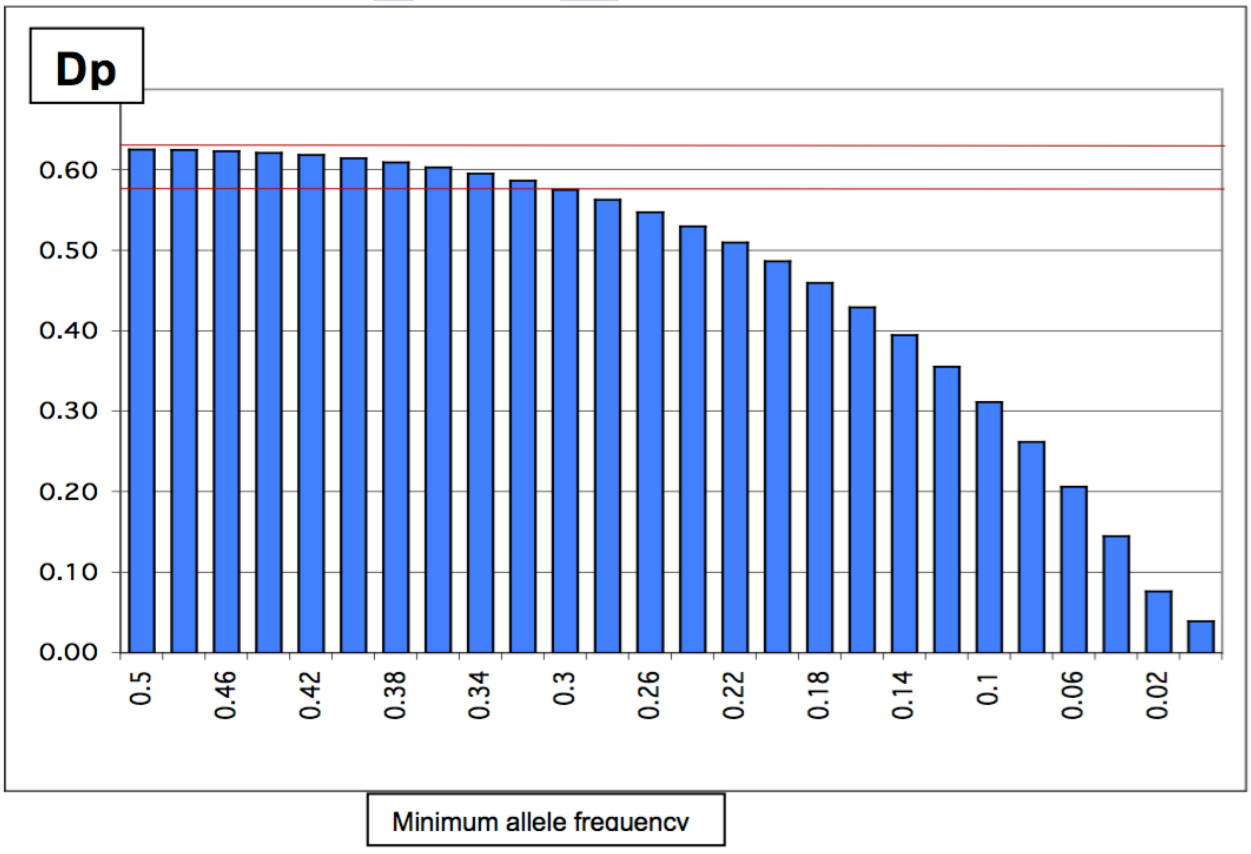


Figure D1. Dp values for different SNP minimum allele frequencies. Between the frequency values 0.5 to 0.3, the Dp value only drops from 62 to 58%, or just 6.5% of the highest Dp obtainable with use of SNPs for forensic identification purposes (red lines).

From these searches, sets of 46, 67 and 25 SNPs were selected giving a median two loci per p-arm and four loci per q-arm from each autosome. In the case of the third SNP selection of 25 loci, several markers were included with more limited variability in one of the three populations, in order to add predictive power for population-of-origin; a selection process that eventually led to dedicated ancestry informative SNP sets complimentary to the 52 ID-SNPs. The final screening of SNPs before assimilation into candidate pools involved careful examination of the flanking sequence to ensure that the region available for primer design (approximately 100 bp on each side of the SNP) was free from low complexity sequence (repetitive sequence, poly-base tracts and sequence found in multiple genomic locations) and had zero or a low number of clustering SNPs with high allele frequencies. Furthermore, sequence features and interactions such as stable loops, complimentary nucleotide tracts within the pool of candidate SNP´s flanking sequences and irregular %GC levels, led to rejection of otherwise suitable SNPs. This last sequence scrutiny step delivered a candidate pool of SNPs with the highest quality of context sequence and involved the loss of many markers that often had better quality levels of polymorphism. This has led to criticism of the 52 SNP set for lacking many of the best loci in terms of polymorphism and best possible population balance (i.e. low between-population Fst, maximum heterozygosity in all populations; see Pakstis, et al, 2010 [117]). However, the final Dp values from the 52 SNPs far exceed those of the largest STR multiplexes and since approximately 24-28 SNPs in any one population reach Dp values equivalent to an inferred global *uniqueness* (i.e. a probability of 1 in 9 billion to find an identical SNP profile in other unrelated individuals), half of the markers in the 52-plex only provide greater data depth in cases where profiles are incomplete. Therefore, once 75% of the 52 SNPs had been assembled population balance became irrelevant. A formal comparison of Dps from different ID-SNP sets is made in **Thesis Paper #10, C. Phillips, 2012** [118], which contrasted marker sets from SNP*for*ID, and *Kiddlab* selections plus two smaller sets developed for forensic identification purposes; comprising 38 Indels or 18 nucleosome binding site-located SNPs (**Pereira, Phillips, et al, 2009; Freire, et al, 2011** [102,119]). The cumulative Dp values show a minor bias towards EUR discriminatory SNPs in the 52 SNPs compared to AFR and E ASN, due to retention of SNPs that originally just had EUR frequency data in TSC. Apart from the small-scale nucleosome SNPs, all the other three sets gave random match probabilities that far exceeded the world population of 1 in 9 billion, indicating that full or even marginally partial profiles of any set would uniquely identify an individual.

When considering the application of SNPs to kinship analysis, where short amplified fragments would give their use better chances of success compared to STRs, i.e. in the identification of missing persons or mass disaster victims; it is important to consider that the number of SNPs required to match the informativeness of STRs is much higher in relationship testing than in identification. This is because relationship testing compares the alleles shared by two individuals, whereas identification uses the whole genotype, so at any one locus only half the genetic information is used to establish a statistical likelihood of relatedness. Charles Brenner has published an elegant "thought experiment" that provides a simple framework for comparing the information content of SNPs and STRs in both

applications [120]. If the SNPs are perfect (0.5:0.5 allele frequency distributions), then one STR is equivalent to 2.6 SNPs for identification and 4 SNPs for relationship testing. Therefore, while approximately 40 SNPs are sufficient to match data from STRs for identification purposes, more than 60 are necessary for relationship testing, not allowing for the slightly reduced power of most SNPs having less informative allele frequencies than a perfect 0.5:0.5.

The final 52-SNP forensic panel created from the genome data scrutiny processes described in this section is outlined in detail in **Sanchez, Phillips, et al, 2006** [115,116]. The SNP selection pipeline developed for the 52-SNP test created a robust marker set that has had little modification since its publication ten years ago. The set has been reduced from 52 to 48 SNPs and various subsets of these 48 loci have been created for different detection platforms. However, the PCR primer designs remain unchanged in these ten years. The strict sequence quality criteria established by SNP*for*ID have also led to similar predictable performance in PCR multiplexes developed for a range of other SNP sets for forensic application. As forensic DNA analysis has renewed interest in SNP genotyping due to the availability of massively parallel sequencing systems, the 52 SNPs have been combined with another set of 85 ID-SNPs as outlined in **Thesis Paper #12, M. Eduardoff, et al, 2015** [121], and continue to offer a robust system for identification of degraded DNA that would otherwise give partial or failed STR profiles. This indicates that the careful checks of flanking sequence quality made during SNP selection and rejection of otherwise highly suitable markers was a worthwhile step to ensure long-term multiplex performance for the analysis of the types of extracted DNA commonly seen in forensic testing. Finally, mention should be made of the contrast between SNP*for*ID's strategy of due regard for a SNP candidate's genomic position (i.e., if two perfect candidates occupied closely sited positions on the same chromosome, only one was chosen, developed and described in the publication of the set); and that adopted by *Kiddlab* to publish all potential candidates that met a single population variation criterium. The final total of 88 candidate SNPs from *Kiddlab* included several sets of markers that were closely linked. Their adoption for forensic SNP tests using MPS has therefore led to the problem of linked SNPs forming haplotypes that are little different from the genetic variation seen in one of the SNPs of the pair. In particular, very close linkage occurs in SNP pairs: rs10768550-rs10500617 (separated by 679 nucleotides) and rs9606186-rs5746846 (287 nucleotides). Each linked SNP pair is present in the HID-Ion and Qiagen MPS SNP sets, but absent from Illumina's smaller MPS set of 95 SNPs. This linkage creates almost complete association of alleles in most populations and as it requires the handling of both SNP pairs as haplotypes makes two of the SNPs in each MPS set largely redundant. We compiled haplotype frequencies for the 2504 samples of the 1000 Genomes project and this data complements a recent evaluation of the Qiagen SNP-ID panel which examined its performance in the Illumina MiSeq MPS system and compiled relevant haplotype frequencies from a Swedish population sample in Grandell et al, 2016 [122]. Nevertheless, some statistical adjustment is necessary within the SNP sets that include *Kiddlab* loci when applied to relationship testing, where haplotypes are much more likely to be shared amongst related individuals and require adjustment for allele frequency estimation bias when these are based on population samples of unrelated individuals.

Luckily, detailed analyses have been made of recombination frequencies amongst commonly used forensic markers **(Phillips, et al, 2011, Fondevila, et al, 2012** [123-124]). The recombination rate estimates generated from the forensic genetic map data published since 2011-12 has also recently been adapted into a program that takes account of close linkage as multiplex scales become ever larger from developments in MPS technologies. Specifically, the statistical program ***ILIR*** (Impact of Linkage on forensic markers for Identity and Relationship tests) was developed to handle linked and unlinked genetic data for forensic relationship testing scenarios that involve expanded marker sets to meet the needs of such testing when pedigrees of surviving relatives are incomplete and/or the tested relationships are distant within that pedigree (**Tillmar and Phillips, 2016** [125]).

The second part of the genomic SNP data objective was to build online browsers for the markers selected for forensic use. By the time this initiative was started there were two forensic SNP multiplexes developed by the author: the above 52-plex ID-SNP set and the 34 ancestry-informative SNPs (AIM-SNPs) collected into an optimised single PCR test described in **Phillips, et al, 2007** [57] and revised slightly in **Thesis Paper #3, M. Fondevila, C. Phillips, et al, 2012**. [8] The first SNP browser developed at USC compiled population data from USC and SNP*for*ID partner laboratories. The novel characteristic of the browser was the ability to choose groups of individual populations (e.g. NW Spanish with Danish) and re-calculate SNP allele frequencies from the combination of their data. This provided the option to combine data into continent-wide summary frequencies, or to allow for significant divergence of frequencies due to differing population histories (e.g. Nigerian vs. Somali AFR populations). The SNP*for*ID browser retains two main databases of 52 ID-SNPs and 34 AIM-SNPs and continues to expand the scope of populations for which data is combinable. Another key feature of this browser is the ability to download the frequency estimates or the genotypes directly to the user´s PC. The SNP*for*ID browser is described in [78].

This same data-marting model of combining allele frequencies based on a user´s choice of populations and the ability to download subsets of genotypes, was next extended to 640,000 SNPs characterised for HGDP-CEPH panel samples by independent but parallel Stanford and Michigan University genotyping studies. The browser that was developed, termed *SPSmart* (SNPs for Population Studies) also included the Perlegen and HapMap SNP genotype repositories. The *SPSmart* browser is described in **Amigo, Salas, Phillips, Carracedo, 2008** [126,127]. *SPSmart* has proved to be a key tool in the selection, assessment and compilation of forensic SNP sets by the author. One limitation of accessing CEPH, Perlegen and HapMap data is their restriction to whole-genome scan (WGS) systems for SNP genotyping. This means many SNPs not located near genes or which are not polymorphic in European populations were not selected for inclusion in WGS panels. This shortfall has now been addressed with the final phase of *SPSmart* browser development, with its extension to include 1000 Genomes data in a browser termed ENGINES (ENtire Genome Interface for Exploring SNPs). The ENGINES browser is described in [128]. While it is now possible to interrogate data for 28 million SNPs in 1000 Genomes by rs-number, gene or chromosome segment using ENGINES, the 1000 Genomes data has undergone a

final large-scale expansion of markers, samples and populations (28>79 million loci, 629>2504 individuals, 12>26 populations). Therefore, when this data has been uploaded to ENGINES it will provide the most convenient tool for accessing 1000 Genomes when multiple SNP data is required and individual genotypes must be scrutinised.

One future goal for the USC SNP frequency browsers that is worth pursuing, is the ability to select SNPs, then automatically convert their genotypes into training sets for use in the forensic SNP analysis portal developed by USC as the *Snipper* suite of classification tools. This would lead to an easier process of selecting and utilising population data for AIM-SNPs that are part of custom forensic ancestry sets. All population data for HapMap, Perlegen, HGDP-CEPH 650K analyses (both Stanford and Michigan University studies) and 1000 Genomes projects (the latter, up to Phase III data releases), are freely available from the central *SPSmart* portal. SPS SNP browser undergoes periodic revision and renewal when necessary and enjoys a high worldwide hit rate (rates of web access overview shown in **Figure D2**). In addition, dedicated forensic data pages have been established for forensic 52 and 34 SNPs originally chosen for ID and ancestry purposes by SNP*for*ID (the SNP*for*ID browser) and Indels chosen for ancestry analysis (see later discussions).

Lastly, complete HGDP-CEPH genotype data for 60 forensic STRs is also held in the SPS framework in a browser named ***pop.STR***. An initial description of ***pop.STR*** was made in **Amigo, Phillips, et al, 2009** [129]. The ***pop.STR*** website adapted the *SPSmart* data framework and algorithms to create a similar browser which allows the analysis of forensic STR allele frequencies in identical fashion to *SPSmart*. The first STR data was built on in-house frequency studies of the 15 STRs of *Identifiler* plus the five ESS STRs in the 52 CEPH-HGDP populations. The ***pop.STR*** database can also combine the STR genotypes based on user-defined population groupings - as with SNP data, but genotypes are not shown or available to download, as this functionality could lead to knowledge of the STR profiles of the HGDP-CEPH donors and raises privacy concerns for individuals from countries with national DNA database regimes in place (e.g. Orcadians that may be listed on the UK DNA database). Summary allele frequency data is presented in tab-delimited format, applicable to transfer to Excel (the semicolon delimited data is converted to a table using the 'text to columns' function). As new STRs are introduced to forensic MPS multiplexes, the allele frequency data in ***pop.STR*** will become increasingly important to help assess the extra discrimination power brought by novel STRs - in the main these will be so-called Mini STRs which have formed the bulk of new data added to pop.STR during 2014-2016. The importance of assessing new Mini STRs for MPS is discussed in: **Parson, et al, 2016** [130] and it is likely that a range of new STRs will be adopted for MPS analysis as the upper limits of PCR multiplexing of forensic STRs is pushed higher. The ***pop.STR*** data handling framework, the population variation characteristics of core forensic STRs and their compiled data in ***pop.STR***; plus analysis of forensic STR population data are covered by publications: **Amigo, Phillips, et al, 2009** [129], **Phillips, et al, 2011** [104], **Phillips, et al, 2013** [108], **Phillips, et al, 2014** [106] and **Phillips 2016** [131].

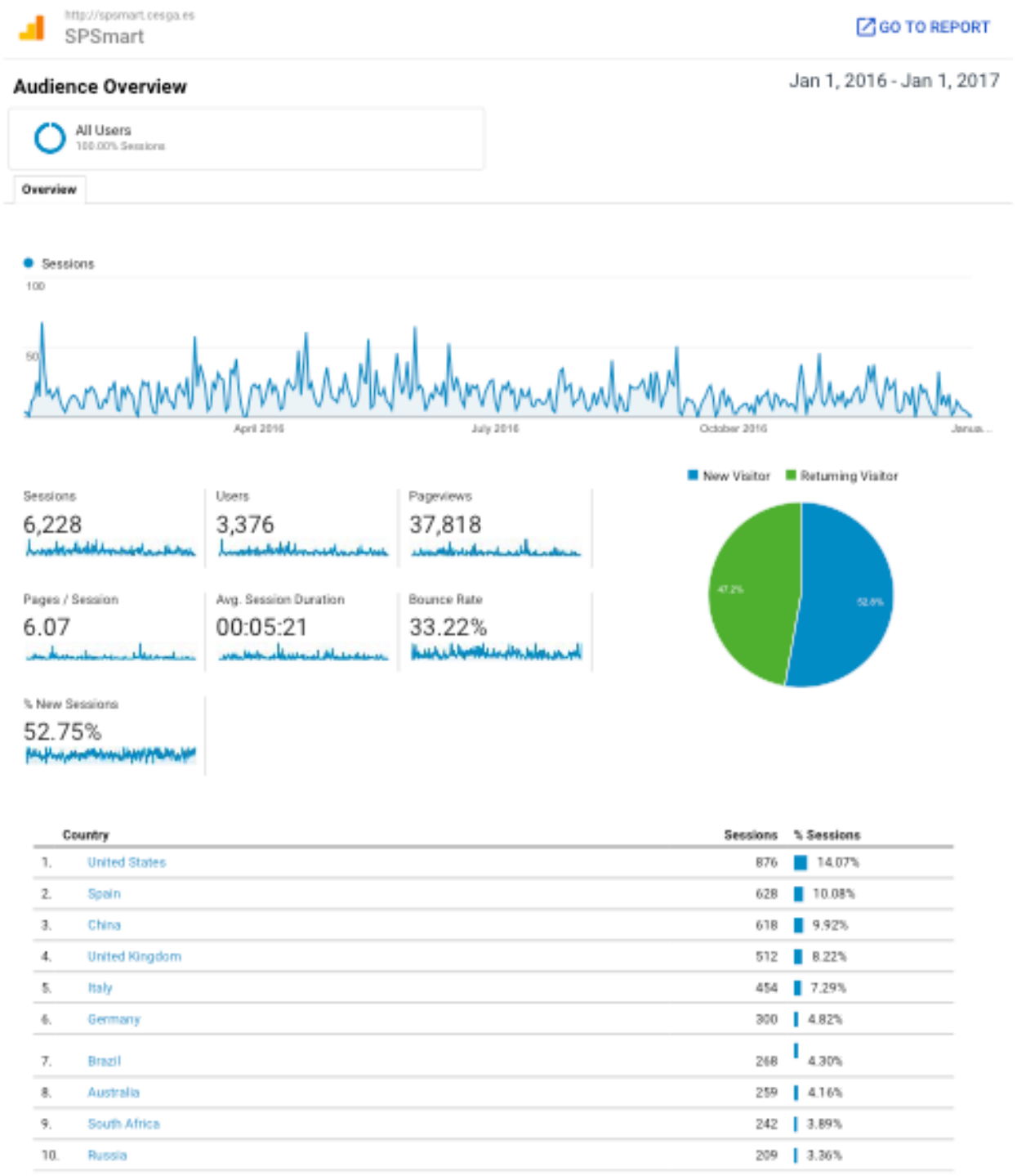


Figure D2. Web access rates for the *SPSmart* suite of SNP, forensic STR and forensic Indel databases for the year of 2016. About half of the 3,376 visitors were new to the site and were predominantly from US, Spain, China, UK and Italy. Average pages generated per visit were 6, so multiple SNP queries were common amongst scientists visiting the database and collecting genotype data.

This section finishes with a cautionary tale about being thorough and following the correct process of using online genomic data to properly identify and characterise forensic markers not in common use. In early 2016, while preparing sequence variant nomenclature guidelines (in anticipation of expanded forensic multiplexes for MPS), it was discovered that STR D5S2500 had multiple positions and genomic characteristics reported. This ambiguity occurred because the D5S2500 locus consists of two different microsatellites forming separate components in the capillary electrophoresis multiplexes of Qiagen's HDplex and AGCU ScienTech's recently introduced non-CODIS 21+1 STR multiplex. Detailed studies of the surrounding sequence of the HDplex locus and the AGCU locus [132] revealed the HDplex D5S2500 has the correctly assigned STR name, while that of the AGCU 21plex, closely positioned a further 1643 nucleotides in the human reference sequence, is an unnamed microsatellite, which was given the unique identifier of D5S2800. The fact that D5S2500 had existed as two distinct STR loci undetected for almost ten years underlines the need for careful scrutiny of the genomic properties of forensic STRs, as they become adapted for sequence analysis with MPS. A recommendation was made that precise chromosome location data must be reported for any forensic marker under development but not in common use, so that the genomic characteristics of the locus are validated to the same level of accuracy as its allelic variation and forensic performance. In fact, the detection of the D5S2500 ambiguity prompted detailed analysis of D6S477 and D15S659 - two further recently adopted STRs in another recently introduced capillary electrophoresis multiplex called the Goldeneye® 22NC DNA identification system [133]. The primer pair listed for D6S477 anneal internally to the D15S659 primer pair producing a smaller amplicon (~142 bp) but at the same site. Therefore, the Goldeneye multiplex has two primers pairs amplifying the same locus. The author points out that despite this, it is concerning that the 9947A DNA genotypes listed for D6S477 (10.2,13) and D15S659 (14,16) are incommensurate. Therefore, a much more thorough genomic audit is necessary for forensic STRs not in common use, to bring them up to the same standard of sequence knowledge that is enjoyed by SNPs. The ***STRidER*** database has been established in early 2016 to begin the process of applying a rigorous data-quality check of STR allele recognition and nomenclature, population analysis, and allele frequency estimation (**Bodmer, et al, 2016** [134]). It seems apparent that the ***STRidER*** framework will also need to apply similar rigour to the genomic identification of newly adopted STRs. To date, the STR multiplexes that have uncommonly used STRs are: Qiagen HDplex (**Phillips, et al, 2014** [135]); Promega CS7 (**Phillips, et al, 2013** [136]); Goldeneye® 22NC (Zhang, et al, 2013 [137], but multiplex modified since this report); and the Scientech AGCU 21+1 set mentioned above (Zhu et al 2015 [138]). Several of these uncommonly used STRs have also been adopted for extended MPS STR kits which at the time of writing comprised: Illumina ForenSeq Signature DNA kit comprising the five extra STRs of D1S1627; D4S2408; D9S1122; D17S1301; D20S482, and Thermo Fisher Scientific Precision ID kit comprising the eight of D1S1677; D2S1776; D3S4529; D4S2408 (in common with the ForenSeq kit); D5S2800 (see above); D6S474; D12ATA63; D14S1434 (all twelve of these markers are NIST Mini STRs in AGCU 21+1).

2. To evaluate SNaPshot as a forensic SNP genotyping system from its initial availability in 2001. To develop novel SNP sets specifically for forensic use, comprising: ancestry-informative SNPs (AIM-SNPs); identification SNPs (ID-SNPs); coding-region SNPs for prediction of common-variation externally visible characteristics (EVC-SNPs); mitochondrial DNA variants (mt-SNPs); and Y-chromosome SNPs (Y-SNPs).

Throughout the period of study for this thesis, SNaPshot single base extension chemistry has been the system of choice for forensic SNP genotyping. The main reason for this is the sensitivity that is offered by its use of dye-based analysis based on laser excitation of labelled DNA products that are run past a static scanning point by capillary electrophoresis (CE) in automated detectors (the injection of samples, buffer and cleaning flushes from micro-titre plates being robotised). This approach relies on automated CE systems already optimised to detect the same dyes attached to PCR products from the amplification of standard forensic STR sets, so this levers the optimised PCR regimes and electrophoresis protocols focussed on CE sensitivity: maximising the signal strength of forensic DNA profiling made by the same automated CE detectors. The other advantage of SNaPshot is that the system avoids the need to commit to expensive and complex new technologies specifically for SNP analysis - the established detectors and PCR reaction conditions can be used with little or no adaptation. In fact, any requirement to run forensic SNP genotyping as a parallel approach to mainstream STR analysis or Sanger sequencing is very easy to accommodate because similar or identical CE regimes can be used for all three forms of forensic DNA analysis on the same detector.

Several alternative approaches to SNP genotyping have been used successfully in the genomics and medical genetics fields, offering much larger multiplex scales and higher sample throughput. However, they require input DNA in the range 10-100 ng and to work at the highest multiplex levels for WGS analyses, the input DNA requirements extend to micrograms of DNA. This precludes such technologies from forensic use. It was not until the emergence of compact massively parallel sequencing (MPS) systems in the last two years, that the genomics field was able to provide a SNP genotyping approach that could rival the sensitivity of SNaPshot for analysing forensic material.

Therefore, constructing small-scale PCR multiplexes as the target-enrichment preamble to single base extension has been the standard procedure for developing SNP sets applicable to particular forensic applications, namely: ID-SNPs; Y-SNPs; mt-SNPs; Nucleosome SNPs (suggested to be sited in genomic regions that are protected from the effects of severe DNA degradation); tri-allelic SNPs; AIM-SNPs (with different sets designed to address particular population differentiations) and forensic phenotyping / EVC informative SNPs.

Although the haplotypic SNPs of Y-Chromosome and mitochondrial DNA are easier to genotype with SNaPshot than equivalent autosomal SNPs - because single peaks are obtained per SNP, near-identical

procedures has been used to develop all forensic SNP tests at USC. The development of a 17-marker mitochondrial coding SNP set was described in **Quintáns, et al, 2004** [139]. This SNP set was developed by identifying the most informative SNPs, combining them in single target-capture PCR and SBE reactions and ensuring the electrophoretic separation was sufficient to unequivocally call the alleles from the peak patterns in CE. The analysis of 52 ID-SNPs in one PCR amplification step required the division of the SBE reactions into two parallel SBE reactions of 23 and 29 SNPs. Although CE detectors have a default capillary polymer fill stage between each injection, the development of the 52-SNP test also led to a convenient, improvised workaround of the double analysis: by creating a new GeneScan method file that allowed sequential injection into the same capillary. In this way, both SBE reaction products could be loaded into adjacent cells in successive 2x8 column arrays (for 16 capillary detectors) and a 'composite' profile constructed where the fragments of the 23 SNPs of SBE-1 are followed, with a minimal electrophoretic gap, by the 29 SNPs of SBE-2.

A key additional refinement added to forensic SNaPshot tests since their first development, has been the use of redundant primer sets to allow for variant nucleotides sited in primer binding sites (making the distinction between SNPs and variants, when the latter can be at much lower frequencies or in certain populations only). Perhaps the best method for maximising the PCR yield with redundant primer sets is to make use of two PCR primer pairs to create four different amplification products from their combinations, so that any binding inefficiency or full non-binding is compensated by the other three pairings. Such an approach has been developed by Andreas Tillmar for the development of a 130-SNP MPS forensic identification panel that uses all 52 of the SNP*for*ID set [122]. In the assessments of this 130-plex PCR with the Ion PGM™ MPS system, no discernible difference was seen in the PCR yield compared to single-pair PCR primer sets, but this system of multiple redundant primers is primarily designed to reduce allele drop-out from closely sited flanking region SNPs, not to improve the sensitivity or PCR yield of the amplification reaction as a whole (Fig. 2 in **de la Puente, Phillips et al, 2016** [140]).

Another less commonly used refinement is the adoption of CT-only SNP sets that can help to achieve optimum peak balance by avoiding over-signalling in the blue and green dyes used in SNaPshot. This approach was applied in the development of one early ID-SNP set at NIST [141], and was explored in a CT-SNP only ancestry set, where the focus was on the distinction of mixed DNA patterns (where more peak pairs are seen in binary markers but they tend to be imbalanced) from the higher heterozygosity that occurs in admixed individuals, this AIM-SNP set is discussed in more detail in the next section that focusses on complete-solution forensic ancestry analysis approaches. Luckily, the identification of CT SNPs suitable for either identification or ancestry analysis purposes

The development and optimisation of a full range of forensic SNaPshot tests for several different applications are described in: **Quintans, et al, 2004** [139]; **Weiler, et al, 2016** (mtDNA coding SNPs [142]); **Sanchez, Phillips, et al, 2006** [115]; Lou, et al, 2011 (ID SNPs [143]); **Freire-**

Aradas, et al, 2011 (degraded DNA [119]); Walsh, et al, 2011, **Ruiz, Phillips, et al, 2013, Chaitanya, et al, 2014** (eye colour predictive SNPs); **Maroñas, Phillips, et al, 2014** (skin tone predictive SNPs); Walsh, et al, 2013, **Söchtig, Phillips, et al, 2015** (hair colour predictive SNPs); **Marcínska, et al, 2015** (male pattern baldness predictive SNPs) **Póspiech, et al, 2015** (hair morphology predictive SNPs) [144-151]. This list of forensic SNaPshot assays does not include those developed for ancestry analysis, which are discussed in the next section.

As MPS is set to gain increasing traction in forensic analysis in the coming years, it is important to highlight the critical role all of the above SNP tests, plus many other forensic SNaPshot assays, have had in accelerating the development of the larger target capture PCR multiplexes used by MPS. A recent study of the 'portability' of established SNaPshot PCR reactions to MPS indicate that five different amplifications can be made independently then their amplified products combined ahead of the MPS library preparation stage. Apart from a PCR amplicon purification step that used QIAquick centrifuge columns to remove unincorporated primers, the pooled amplicons were unmodified before the MPS libraries were prepared. The MPS results indicated some sequence coverage imbalance in the SNPs from one SNaPshot PCR that was still in development so still had incomplete primer-ratio optimisation, but overall coverage was remarkably balanced for most of the 136 SNPs. This study indicates that an optimised target capture PCR previously developed for SNaPshot represents an optimised PCR for MPS. It is apparent that SNP sets of differing scales will be easy to reconfigure into expanded SNP multiplexes that compile the most informative forensic markers and have scope for accommodating more as these are identified. Two studies assessing the direct porting of SNaPshot-based PCR products to MPS regimes are detailed in: **Daniel, Santos, Phillips, et al, 2015**; and **Mehta, Daniel, Phillips and McNevin, 2016** [73,152]**.** These assess the adoption of SNaPshot-based PCR multiplexes for the Ion PGM™ and Illumina MiSeq platforms respectively.

Two reviews of the development of SNaPshot assays and their application to forensic SNP genotyping have been published and these provide complimentary overviews of the wide-scale use of SNaPshot to develop SNPs for forensic application ahead of their adoption in MPS tests: **Fondevila, Børsting, Phillips, et al, 2016** [153] and **Mehta, Daniel, Phillips, et al, 2016** [154].

Finally, mention should be made of a realistic CE-based alternative to SNaPshot, called Genplex, that was explored by the author in detail in 2006-2008. Genplex used oligo-ligation chemistry to identify the SNP site with a locus specific DNA fragment bearing a dye label ligated to a pair of allele-specific DNA fragments, each with a slightly different mobility modifier - enabling electrophoretic separation of each SNP as a set sharing the same dye and run in that dye channel in CE. Multiplexes were built up to 48 SNPs of ID-SNPs and AIM-SNPs, but each just used the blue and green dyes, so multiplexes up to 96 loci would have been possible incorporating the red and yellow dyes also, although this scale of multiplexing was never explored. The technique was poorly supported by Applied Biosystems, who developed the technology to run on their CE detectors and using their proprietary dyes and mobility

modifying agents, and was eventually withdrawn as a commercial product. Nevertheless, a successful system was developed and optimised using 48 of the 52 SNP*for*ID identification SNPs, with particular sensitivity to mixed DNA, as described by **Phillips, et al, 2007** [155] and **Musgrave Brown, et al, 2008** [156]. An ancestry-informative set of 48 SNPs was also designed but never reached the stage of multiplex production. Genplex technology is described in **Box 8** with examples of mixed DNA detection during the evaluation of the system.

3. To develop a complete forensic ancestry inference solution based on AIM-SNP typing with established forensic capillary electrophoresis platforms, comprising: a highly multiplexed PCR and SNaPshot assay; open-access genotype analysis tools using Bayes likelihood calculations (*Snipper*); comprehensive population surveys for the SNPs of the test (the SNP*for*ID *SPSmart* pages).

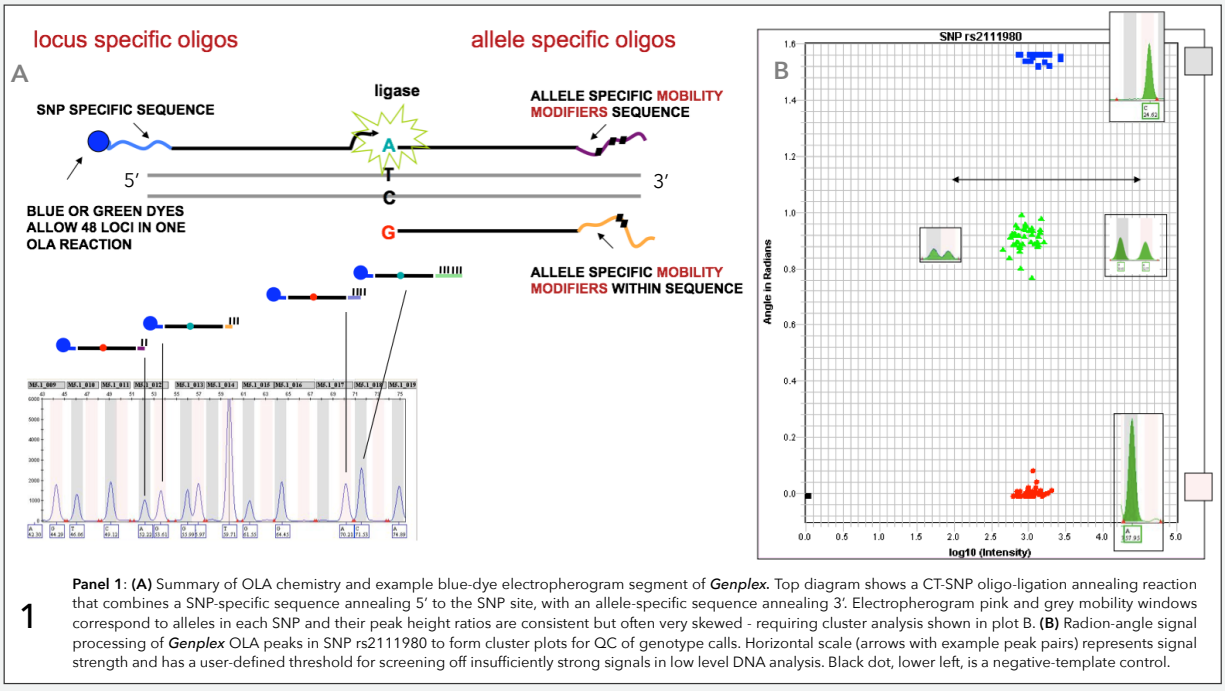
A comprehensive forensic ancestry analysis system must aim to combine the key elements of: a sensitive series of SNP genotyping tests, preferably applicable to the CE regimes available to the bulk of forensic laboratories from existing use of automated sequencers; a robust statistical analysis toolbox for inferring the geographic origin of the profiles obtained from SNPs or other types of variant, preferably with real-time analyses; and allele frequency databases for the markers used that can allow a customised approach to the statistical analyses using a series of reference populations relevant to the circumstances of the investigation.

The USC "ancestry analysis pipeline" attempts to address each of the above stages in the inference of the geographic origin of unknown DNA donors detected as contact traces from criminal investigations or presented as unidentified human remains - i.e. evidential material that requires identification of an individual in cases without significant investigative leads, whether a lack of eyewitness testimony, a DNA database hit or other circumstantial evidence that can point to a likely identity.

The SNaPshot PCR amplification multiplexes have proved to be sensitive enough to provide full profiles from sub-nanogram input DNA, as evidenced from the 11-M profiling results - see **Box 2**. An initial first test of 34 SNPs can be used as the universal initial step in a nested approach of dual SNP tests where the second SNaPshot test that can follow is directed by the likelihood values seen from the genotypes of the 34 SNPs. This nested analysis process is outlined in more detail below.

Box 8. Genplex

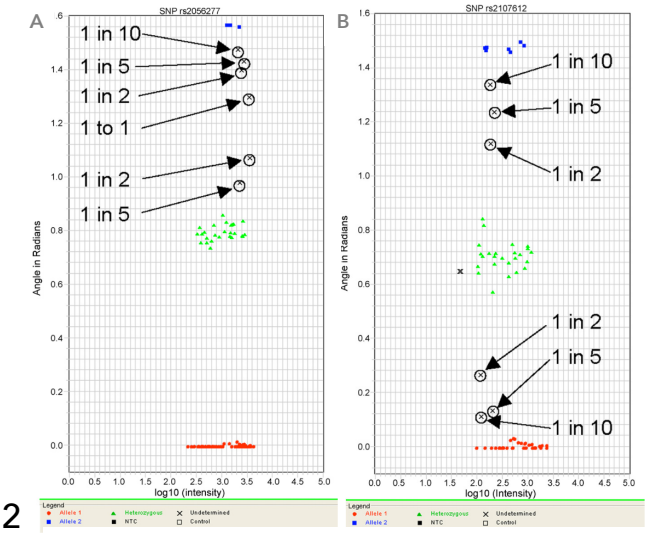
Genplex technology was launched as a CE-based SNP typing system from Applied Biosystems (AB) in 2005. SNP sites are genotyped by an oligo-ligation annealing (**OLA**) step centred on binding to the SNP site: a 3′ allele-specific oligonucleotide with differentiated mobility (precisely set by modifying strand inter-chelators) and a matched 5′ dye-linked oligonucleotide to label the products. As **OLA efficiency differs** between both alleles, software assigns allele calls based on **cluster analysis** to aid accuracy when OLA signal ratios are very skewed (i.e. the green mid-plot heterozygote cluster shown for rs2111980 in **panel 1B** can be much closer to the 0 or 1.6 radion-angle baselines). Compared to SNaPshot CE analysis the use of one dye per SNP and cluster analysis in ***Genplex*** gave a more reliable SNP genotyping system sensitive to mixed DNA, as shown in **Panel 2A/B**. Here outlier signal ratio points (black crosses) from known mixtures correspond well to the expected positions and can be detected down to 1 in 10 ratios. Experiments with degraded DNA also also indicated a more sensitive genotyping system than STR genotyping with CE from the shorter amplified fragments (**Panel 2C**). Finally, the use of just blue and green dyes to genotype 48 SNPs indicated much larger multiplexes would have been analysable with ***Genplex***.



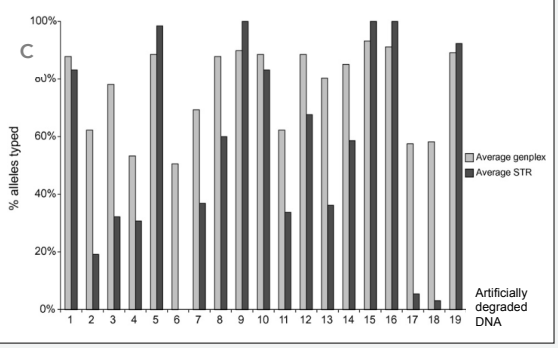
Panel 1: **(A)** Summary of OLA chemistry and example blue-dye electropherogram segment of ***Genplex***. Top diagram shows a CT-SNP oligo-ligation annealing reaction that combines a SNP-specific sequence annealing 5′ to the SNP site, with an allele-specific sequence annealing 3′. Electropherogram pink and grey mobility windows correspond to alleles in each SNP and their peak height ratios are consistent but often very skewed - requiring cluster analysis shown in plot B. **(B)** Radion-angle signal processing of ***Genplex*** OLA peaks in SNP rs2111980 to form cluster plots for QC of genotype calls. Horizontal scale (arrows with example peak pairs) represents signal strength and has a user-defined threshold for screening off insufficiently strong signals in low level DNA analysis. Black dot, lower left, is a negative-template control.

1

After much effort in SNP*for*ID to develop ***Genplex*** fully and create the 48-SNP ID set plus a custom 48-SNP ancestry set (with two tri-allelic SNPs), as well as the forensic validation of mixtures, dilutions and degraded DNA shown, it was withdrawn from the market by AB.



Panel 2: Forensic utility of ***Genplex*** as a CE-based analysis system of short-amplicon binary SNPs. **(A)** Genotype cluster plots showing positions of a homozygote1:heterozygote mixture series (**x**) in SNP rs20562777; indicating well defined deflections from unmixed clusters that match their mixture ratios well (though 1 in 10 heterozygote not detected). **(B)** Patterns from homozygote1:homozygote2 mixtures in rs2107612 are generally even more well defined, although 1:1 clusters with heterozygotes (black cross left). **(C)** Genotyping success with ***Genplex*** SNP analysis was better than STR analysis in 14 of 19 UV-degraded DNAs.



2

The development of open-access statistical tools embedded in a single web-based portal was a key step in providing a complete solution to forensic ancestry inference with SNPs. The *Snipper* portal carries a number of analysis algorithms focussed on naïve Bayes analysis using the likelihood ratio of the two highest ancestry assignment likelihoods. The population group likelihoods being simply calculated from the product of individual SNP allele frequencies (of those seen in the submitted SNP profile). Initially, *Snipper* provided fixed training sets from which the allele frequencies were derived, but as more SNP markers were chosen to enhance the first 34 AIM-SNPs there was a need to accommodate custom genotype data. This was achieved by allowing user-uploaded data for their own choice of SNPs and populations. The most efficient way to obtain such data is to use *SPSmart* and ENGINES to access HGDP-CEPH and 1000 Genomes data - thus providing a complete solution of: sensitive forensic tests; data analysis; and custom data collection for tailored training sets. The analysis of custom data is also possible by reclassification and cross-validation analysis steps that can be applied to a custom training set file to assess how successfully the markers make particular population differentiations (see: **Thesis Paper #6, C. Phillips, et al, 2012** [157]; **Thesis Paper #8, C. Phillips, et al, 2016** [158]; **Santos, Phillips et al 2016** [159]). This can provide useful assessments of the range of likelihoods that might be expected from analysing unknown samples with ancestries similar to the custom populations used to construct the tailored training sets. This form of analysis was informative in the 11-M investigation by creating a ´profile´ of ranked pairwise likelihoods that illustrated different degrees of divergence (from Spanish Europeans) within the Moroccan samples that formed the training set. Such an analysis approach can also help to set a threshold likelihood value, below which no assignments are made. This follows the assumption that erroneous ancestry assignments due to a lack of sufficient divergence tend to produce very low assignment likelihoods. Therefore, a high value is reliable but sub-threshold values below those seen in erroneous assignments can be treated as unreliable. This sets the overall classification success rate and cross-validation can therefore set a balance between an aggressive threshold, where a large proportion of samples are not classified, and a relaxed threshold where a degree of assignment error will occur. Setting a balance between these two extremes is not always easy, but a threshold set by assessments of classification error, in the above way, can guide interpretation of low likelihoods, particularly when these are the result of partial profiles from poor DNA rather than a lack of population divergence. Lastly, a good set of example pairwise likelihood plots is found in Supplementary Fig. S3 of **Thesis Paper #15, M. Eduardoff, et al, 2016** [160].

The *Snipper* suite has also been enhanced by the provision of PCA analysis modules. these have been a valuable visual aid to help interpret the likelihood values - particularly when they are lower than those of the training set due to admixture or partial data. The 2015 EDNAP inter-laboratory exercise of CE-based forensic ancestry tests indicated the PCA module of *Snipper* gave an intuitive and complementary form of data analysis to the Bayes analysis made in parallel in the same *Snipper* analyses. At this point, *Snipper* provides two of the three main ancestry analysis approaches in wide-scale use in forensic and population genetics, namely: likelihood ratio analyses (Bayes analysis); PCA;

and genetic cluster detection algorithms (STRUCTURE, ADMIXTURE, ADMIXMAP). The last of these three analysis regimes takes a disproportionate amount of time to complete as they rely on repeated randomised rearrangements of genetic data and resampling (MCMC steps). Therefore, genetic cluster detection algorithms are not suited to real-time analysis of forensic data made within a web-based portal such as *Snipper*, and have not been considered viable. Recently, it has been proposed to offer a data- input option for user-uploaded training set files in Excel format that can be automatically reformatted for input into STRUCTURE. As this can be the most time consuming and error-prone step in the handling of data with STRUCTURE, this would be useful functionality, particularly for occasional forensic users. The value, relevance and potential pitfalls of using the most commonly applied STRUCTURE cluster detection program to forensic ancestry analysis are reviewed in **Porras, Ruiz, Santos, Phillips, et al, 2013** [161]. The ADMIXTURE program is also gaining popularity and offers the fastest approach of the three regimes for the identification of population clusters.

4. To enhance the 34-plex AIM-SNP assay, routinely applied to the differentiation of Europeans, Africans and East Asians, by developing novel SNP sets designed to be run alongside the core 34 SNPs that focus on South Asian, Native American and Oceanian populations-of-origin (*Eurasiaplex*, *PIMA* and *Pacifiplex*). To develop additional sets focussed on improved SNaPshot peak balance (CT-only SNPs); European-African co-ancestry patterns in admixed individuals (*Admixplex*) and X-chromosome AIM-SNPs.

The first 34-SNP ancestry test was designed to differentiate EUR, AFR and E ASN geographic origins and at the time of its publication, no data was available for SNP variation in other population groups. When the HGDP-CEPH SNP data was accessible from *SPSmart* it then became viable to design supplementary forensic ancestry tests centred on Native American-diagnostic population variation; Pacific region-diagnostic variation for the populations of the fifth major continental group of Oceania; and the differentiation of South Asian (S ASN) from European population variation, i.e. comparing east and west Eurasia. To develop the first two multiplexes, this involved the compilation of mainly fixed-difference SNPs, where the allele frequency of many of the markers was close to zero in Native Americans (AME) and Oceanians (OCE) and close to one in the other population groups (but to a reduced extent in E ASN populations). In the latter multiplex there was much less population differentiation between EUR and S ASN populations due to an absence of geographic barriers and three millennia of trade and population movement along the silk route and northern Steppe. Tests were named PIMA (Population Informative Multiplex for the Americas); *Pacifiplex* and *Eurasiaplex* respectively. Development and validation of *Pacifiplex* is described in **Thesis Paper #13, Santos, et al, 2016** [162], and the *Eurasiaplex* panel in **Phillips, et al, 2013** [81]. PIMA has not been published to date, but has been adopted in a range of analyses where Native American co-ancestry is detected and requires better differentiation from EUR and E ASN than many existing ancestry SNP sets

can provide. One example of the value of PIMA SNPs was the Minstead investigation ancestry analyses, where the markers used by *DNAprint* lacked sufficient divergence from EUR to allow accurate estimation of co-ancestry proportions in the DNA donor, who had origins from the Caribbean (see **Box 1**).

The above specialised ancestry tests were designed to provide a nested approach from the initial 34-SNP analysis: i.e., if a European-indicative likelihood was obtained but below the range of values typically found in cross validation of the training set samples with this ancestry, then Eurasiaplex can provide a second strike test capable of confirming European or South Asian ancestry with a much more definitive likelihood than those obtained with 34 SNPs alone. Likewise, a weak likelihood from 34 SNPs but indicating East Asian ancestry could signal origins for the contact trace donor from one of the two groups of AME or OCE that show reduced divergence from E ASN populations. Therefore, the second strike tests appropriate from such findings would be the application of PIMA and *Pacifiplex* tests to obtain a more strongly indicative likelihood for origins in one of these three closely related groups.

CT substitution SNPs produce equivalent SNaPshot signals for the dye-labeled extension fragments corresponding to each base and therefore have better peak characteristics for the detection of mixtures. A 22-plex (termed *CT-plex*) that contained some of the best ancestry informative SNPs was developed and showed reasonably well balanced profiles of allele peaks. This permitted the mapping of peak height ratios in a simple 2D plot with the 45° line denoting perfect peak balance (in heterozygotes). This project also showed that sufficient CT substitution AIM-SNPs (or AG sites typed on the reverse strand) exist as a major proportion of best ancestry markers to provide a large enough candidate pool without compromising informativeness. Typical *CT-plex* SNaPshot profiles are shown in panel 2 of **Box 5**. Peak height ratios in this multiplex are sufficiently balanced to identify mixed source DNA with SNaPshot, even though this process is usually hampered by reliance on the identification of blue and green dye-labelled extension products that have much less regularity in signal strength - their elimination from the *CT-plex* ancestry set meant it was possible to approach forensic mixtures with SNaPshot with as much reliability for their detection as STR and Indel analysis.

In addition, a SNP set for the differentiation of European and African populations was developed called *Admixplex*, as these two co-ancestries are the most commonly encountered in admixed individuals from the US, South America, South Africa and the UK. Furthermore, there are a much larger proportion of SNPs showing well differentiated allele frequency distributions, so it is easy to choose SNPs with the most extreme allele frequencies between EUR and AFR. The *Admixplex* set comprised 28 extreme-difference SNPs, where the allele frequency distributions in AFR and EUR were as close to fixation as possible (i.e. close to 0 for any one allele in Africans and close to 1 in Europeans). The principle underlying use of extreme-difference SNPs here is to derive the number of heterozygotes in the SNaPshot profile, as this should be proportional to an individual's co-ancestry proportions if admixture is a recent event in their pedigree (i.e. parental or grand-parental). In a set of 28 SNPs that each have

completely fixed alleles between AFR and non-AFR or specifically, EUR populations; the number of heterozygotes that can be expected in individuals with parental co-ancestry will be ~14, and if one grandparent is African the number of heterozygotes will be ~7. Although this exact ratio of heterozygotes to co-ancestral contribution obviously varies to some extent in practice and not all SNPs are completely fixed in terms of their allele frequency distributions, this set of AIMs can provide a clear distinction between parental admixture and grandparental or greater admixture and this can help to re-define a suspect pool towards those that may have evidence of two different parental ancestries in their birth record - with this SNP set detecting African and European co-ancestries. The *Admixplex* set also allows a more precise mapping of parental co-ancestry indicated by the analysis of mtDNA and Y-chromosome markers, since they should reveal a match to e.g. an AFR mtDNA pattern and a EUR Y-chromosome pattern, allowing detection of a grandfather vs a grandmother or a father vs mother, depending on the heterozygote proportions observed.

The average European co-ancestry contribution in the genomes of African Americans varies between 5% to 35% in the bulk of individuals, and the same kind of proportions exist in UK African Caribbeans (see the two STRUCTURE plots labelled ASW and ACB in **Figure 2** of the Introduction, for samples of these populations respectively). This means that forensic samples from African-origin populations such as those of the US or the UK will show between 5% (1 or 2) and 35% (about 5-9) European allelic variation for the SNPs of *Admixplex*.

During the periods of ancestry marker compilation described above, X-SNPs were considered to be potentially highly informative ancestry markers for two reasons. Firstly, they show very high levels of differentiation across global-scale population comparisons (the X-chromosome F_{ST} average is 0.21 compared to an autosomal F_{ST} average of 0.12), due largely to the X-chromosome's lower effective population size of 75% of that of autosomes. Secondly, their presence as a single set of markers in admixed males allows more detail to be obtained about such individual's co-ancestry patterns. Although the analysis of mt-DNA and Y-chromosome loci provides similar detail, the X-chromosome can act as a substitute Y in females and adds extra detail to inferences made from mt-DNA data. From an initial plan to collect just one SNP close to, or at allelic fixation (i.e. frequencies of 0 or 1) in each population group, the ancestry-informative X-SNP set was expanded to 17 markers, chosen to have an even distribution on the X-chromosome, so that loci did not cluster markedly in few linkage blocks in a similar way to the X-STRs used in forensic analysis. The location of suitable SNPs in relation to the distribution of linkage on the X-chromosome was informed by detailed studies of recombination rates made previously by **Phillips et al**, 2011 [123]. Attempts were made to develop a system for estimating the likely time-point of admixture (in the first instance, parental, grand-parental or deeper rooted) from inferring recombination creating disrupted combinations of X-SNP alleles (as female meioses have an obligatory cross-over event). However, this project did not progress sufficiently to arrive at a workable tool, although a 17-plex SNaPshot assay was optimised, comprising three African-, European-, and American-informative SNPs, plus four East Asian- and Oceanian-informative SNPs. These showed good

levels of allelic fixation (see panel 1 in **Box 9**) and applied as a single SNP set , were sufficient to differentiate HGDP-CEPH individuals. X-SNPs are worth exploring further as they provide informative detail for the inference of admixture patterns, particularly when X genotype data is compared to the genetic variation in autosomes, mt-DNA and Y markers.

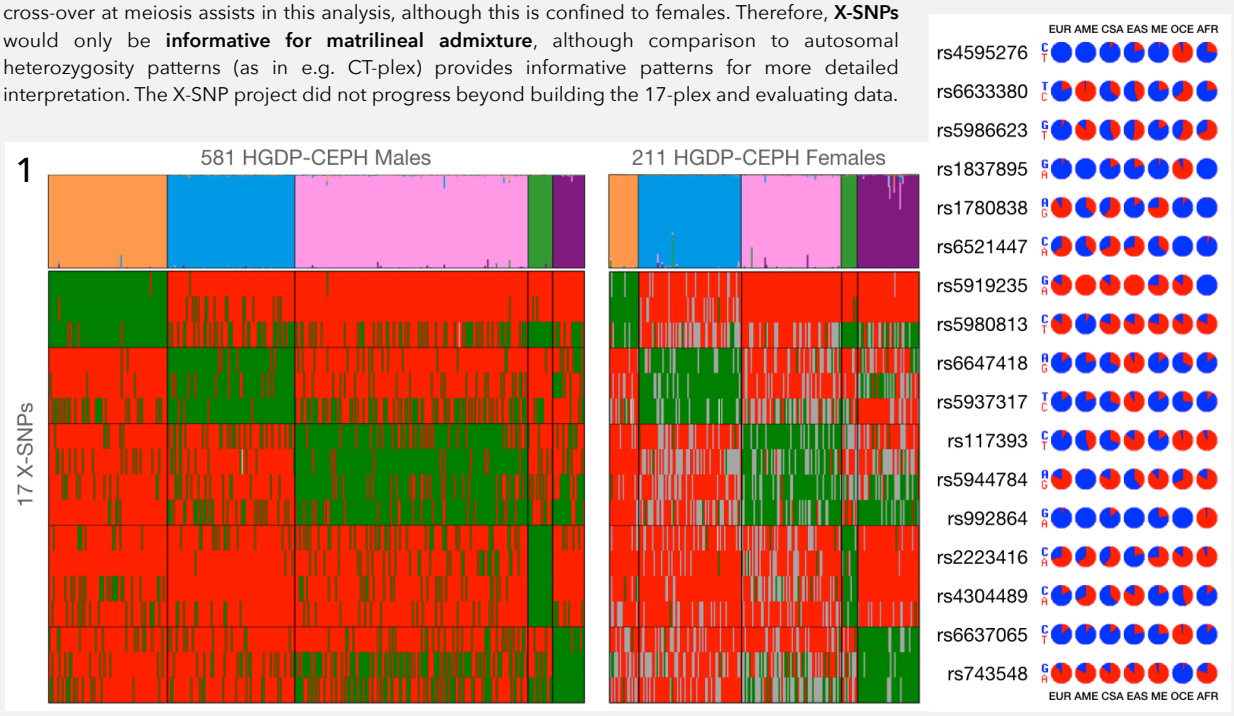
When it was the right moment to compile all the best AIM SNPs for a single 128 marker MPS-based forensic multiplex, *CT-plex*, *Pacifiplex*, PIMA and *Admixplex* forensic sets were mined for best loci. In addition, various case-control association study sets developed for genotyping with Sequenom (e.g. as described in: **Borel, et al, 2012** [163]) that seek to maximise the population differentiation detected but remain compact, were also mined for suitable candidate AIM-SNPs. The initiative to rebuild ancestry marker sets for forensic MPS sequence analysis is described in the discussion of Objective 8.

5. To analyse the sensitivity of SNP tests optimised for forensic use by assessing genotyping performance of the PCR multiplex with a wide range of skeletal material. To apply the SNP test assessments to crime-scene contact trace analysis in order to enhance investigative leads. To promote the concept of ancestry inference in a forensic context to progress criminal investigations lacking database hits or eye-witness. To progress cold-case review approaches by using ancestry inference data.

The single most important characteristic of SNaPshot tests is their forensic sensitivity. Despite many alternative SNP genotyping systems offering faster throughput and automated processes of allele calling; both of these factors well suited to genomic applications, their requirement to use as much as 100 ng of input DNA represents a significant drawback for most forensic analyses and hampers the adoption of less sensitive but otherwise appropriate SNP genotyping systems. At the other extreme, The Taqman real-time PCR analysis system is sensitive enough for forensic purposes but can only genotype one or two SNPs per reaction, so uses a disproportionate amount of material to achieve the same levels of genotyping coverage as SNaPshot multiplexes of 20-30 markers in one PCR amplification.

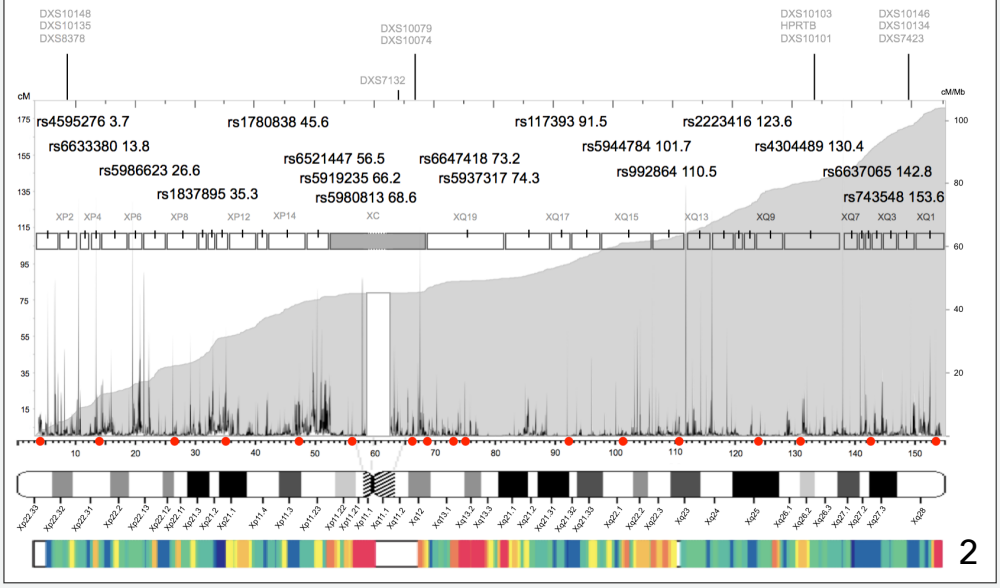
Box 9. X-SNPs as ancestry-informative markers

The X-chromosome shows considerable geographic stratification due to its lower effective population size (75%) than autosomes. During the compilation of autosomal ancestry-informative SNPs, X-SNPs were also observed to have potential as AIMs that could provide particular properties in the **genetic analysis of admixed individuals**. In the same way mt-DNA can indicate the ancestry of a person's matrilineage, X-SNPs can analyse this in males (where the Y-chromosome data indicates the patrilineage), whereas in females, X-SNP data can provide information on both lineages. Therefore, X-SNPs make important additional markers for the analysis of admixture. This property proved useful in the analysis of the **Minstead DNA**, as detection of the African-informative allele in one X-SNP in the PIMA multiplex revealed the matrilineage. In combination with the presence of the R1b Y-chromosome haplotype provided better detail about the balance of EUR-AFR admixture contributions in the suspect's DNA. This prompted the development of a mini-plex of five X-SNPs - one marker at, or close to, fixed allele frequencies in each population. The set was expanded to 17 X-SNPs to allow a degree of co-ancestry monitoring in male DNA by estimating the likely generation (i.e. parental or grandparental) that the admixture event occurred by analysis of recombination creating rearranged sets of X-SNP alleles. An obligatory X-chromosome cross-over at meiosis assists in this analysis, although this is confined to females. Therefore, **X-SNPs** would only be **informative for matrilineal admixture**, although comparison to autosomal heterozygosity patterns (as in e.g. CT-plex) provides informative patterns for more detailed interpretation. The X-SNP project did not progress beyond building the 17-plex and evaluating data.



Panel 1: Allele frequency pie charts for 17 X-SNPs in 7 population groups plus STRUCTURE cluster plots arranged next to CEPH sample raster plots, where the population-informative genotypes are marked in green (heterozygote female genotypes in grey). Populations comprise: AFR=orange, EUR=blue, EAS=pink, OCE=green, AME=purple.

Map shows local recombination activity (cM/Mb, black peaks) and cumulative (cM, solid grey plot) genetic distance plus colour-coded r^2 heat map (red=high values, blue=low values). Values after rs-numbers indicate Mb positions. Four X-STR linkage groups indicated by long lines.



Panel 2: Map of the X-chromosome showing the position of the selected X-SNP ancestry markers (red dots) in relation to the distribution of X recombination rates. SNPs were chosen to have optimally spaced distributions in order to maximise the chance of recombination. Multiple SNPs in single LD blocks (notably the centromeric block at 62 Mb-65.5 Mb) would have reduced allelic reassortment.

For forensic analysis of very degraded DNA, SNP genotyping offers the most sensitive type of tests for characterising autosomal DNA (mitochondrial sequencing also benefits from amplifying a circular DNA molecule present at high copy numbers per cell). This is because the amplified fragments for SNPs can be much shorter than those carrying STRs - in theory as small as 41 nucleotides in length (using two 20-mer PCR primers to amplify the target SNP site). This ability to be genotyped from smaller DNA fragments tends to translate into much better performance with the most degraded DNA, where STR analysis fails completely or gives almost fully incomplete profiles. An early proof of concept for the expected sensitivity of forensic SNP genotyping with SNaPshot, was the analysis of badly burnt skeletal material recovered from the floor of a forest in Ourense, South Galicia in 2006, following a major fire in the area. By applying reduced-length amplicon STR sets and SNP multiplexes, there was an opportunity to assess the ability of each approach to successfully type highly degraded material from a very challenging case, since the most likely origin of the skeleton was from a person missing more than ten years previously. Therefore, the DNA was not only subject to putrefaction, but also exposed to the very high temperatures typical of forest fires. This case therefore represents the most degraded type of target DNA possible. Results indicated that only miniaturised STRs gave results for micro-satellites from DNA extracted from the femur of the skeleton, with some extra peaks and some loci failing. However, SNPs produced full profiles and there was little or no distinction in profile quality from analysis with smaller PCRs of 23 and 29 SNPs and the combined single-tube amplification with 52 SNPs. Lastly, the 34plex AIM SNP profile was also complete and gave a likelihood to be European of 164 Billion times more likely than African and 44 billion times more likely than East Asian. These analyses were described in **Fondevila, Phillips, et al, 2008A** [164]. The number of SNPs in the PCR multiplex does not appear to have a direct influence on the chances of success (e.g. a 52-plex was as successful as a 23-plex and 29-plex made individually) and this depends more on the quality of DNA obtained or adequate control of inhibition (**Fondevila, Phillips, et al, 2008B** [165]).

This first proof of forensic SNP analysis sensitivity was followed by the successful analysis of seven contact trace DNAs from the 11-M Madrid bomb investigation [8], with extracts giving quantities of 0.07 - 0.11 - 0.19 - 0.3 - 2.0 - 3.29 - 12.7 ng/ul, below optimum amounts in the majority of DNAs. All extracts gave full 34plex AIM SNP profiles and indicates that low level DNA can provide sufficient target for successful amplification if there is no degradation.

Since the above two pilot analyses were made, numerous identification cases have been successfully completed at USC by applying SNaPshot SNP genotyping tests. Three of these cases are shown in **Figure D3**, illustrating the range of samples that form the DNA source material for SNP-based identification, where STRs have failed or would likely to give poor profiles**.** Dentine provides the best source of DNA from skeletal material and usually multiple tests can be run to maximise the genotype data depth in any one analysis. The paper describing *Eurasiaplex* [81] gives a good example of a SNP-based analysis of ancestry that was able to give a clearer idea of the population of origin of skeletal remains washed ashore at La Coruña, than canonical analysis of cranial dimensions. Although it was not

known how long the skull had been in the sea, the successful analysis of the DNA in this case again underlines the increased chances of success that short-amplicon analysis of degraded brings to forensic identification. The identification of World War II soldier remains in a mass grave in Papua New Guinea was important in order to complete their successful repatriation to Japan or Australia. The Japanese soldier remains would be cremated once repatriated, emphasising the importance of secure ancestry inference, as the Australian relatives eventually linked to the soldiers remains could have different wishes for their burial. Population admixture was another complicating factor - as East Asian-European admixture had begun to feature in the Australian demographic profile well before the generation of soldiers were born. In this identification program, both 34 AIM-SNPs and 46 AIM-Indels were applied and the quality of results were comparable. Within a range of samples from the same burial site, a proportion worked well and gave sufficient genotypes and others did not. There was no way to predict the degree of genotyping success from the extracted DNA, but clearly it was important to gauge the reliability of the ancestry inference when much of the genotyping was incomplete. Luckily, a direct relationship exists between the number of genotypes obtained and the output from *Snipper* in the form of Bayes likelihoods and PCA positions. This data also matches well with STRUCTURE cluster membership estimates, although interpretation of these results depends on the assumption that only European or East Asian ancestry (or their proportions as an admixture ratio) applies to the analysis of these skeletal remains. **Figure D4** demonstrates the principle of correlation of profile completeness with the statistical reliability of ancestry inference data of *Snipper*-derived Bayes/PCA data and the cluster plots from *STRUCTURE* analyses. Samples A, D, E all give unequivocal ancestry assignment statistics from 34-plex SNP test data completeness of 88%, 88% and 97%. In the case of sample C reaching 56% genotype data completeness, the Bayes likelihood is still very high, although reduced compared to the values for A, D, E; and the *STRUCTURE* cluster plot shows only European ancestry. But the sample's PCA position on the edge of the European cluster indicates the effect of losing almost half the genotype data. In contrast to the other four, sample B only produced 4/34 reliable SNP genotypes, so is consequently positioned in the middle of the PCA plot, gives uninformative Bayes likelihoods and the *STRUCTURE* cluster plots indicate African co-ancestry components. This DNA sample was not successfully assigned to an East Asian or European ancestry. It is relevant to emphasise that the tendency with very degraded DNA is for SNPs to show random drop-out, rather than failing systematically (e.g. linked to amplicon size). Therefore, many cases will have a significant proportion of markers failing, but when they are relatively uninformative for the ancestry differentiation being sought, the results can be as reliable as those from more complete genotype data and the ranked population divergence list in *Snipper* with failing SNPs highlighted in red becomes an important aid to interpretation of the ancestry inference statistics.

Lastly, as MPS is increasingly being assessed as a sensitive and broadly-based forensic SNP genotyping system, it is appropriate to describe the first ID-SNP analysis made with the LT HID-Ion Identity Panel - the first commercially available forensic MPS kit. The sample of degraded DNA analysed was extracted from a 12th Century skeleton found in a medieval mass grave at Volders, Austria. **Figure D5**

summarises the distribution of sequence coverage obtained from analyses made at the Institute of Legal Medicine, Innsbruck Medical University, Austria, of this ancient DNA sample and reported in **Thesis Paper #12, M. Eduardoff, et al, 2016** [121]. The sequence coverage data obtained is shown as a ranked plot of the 169 SNPs analysed with the prototype HID-Ion panel. Over 80% of the SNPs analysed gave coverage above a threshold for reliable genotyping of 20 sequence reads, and altogether, about half of these markers gave coverage of more than 100 sequence reads. Just taking those 82 SNPs that had more than 100 sequence reads produced a cumulative random match probability of 1.2E-33, some five orders of magnitude higher than an expected level of discrimination possible with the 21-STR GlobalFiler multiplex. Given the likely quality of the DNA extracted from a skeleton more than 800 years old, this initial trial of MPS technology for forensic SNP analysis suggests its sensitivity will exceed that already provided by SNaPshot tests and will be the system of choice for such challenging analyses. Only 32 SNPs had coverage below 20 sequence reads and four more had no-calls in both replicates, further suggesting a robust forensic assay. The informativeness obtained from 82 SNPs reveals the power of large panels of SNPs compared to even the largest forensic identification STR sets. **Figure D5** also highlights the continuing efforts to reduce the size of amplified fragments towards shorter lengths and this has been successfully achieved in the final commercial SNP identification kit released by LT for forensic MPS analysis. For 57 SNPs of the 124 loci retained in this identification SNP set the average amplicon length has been reduced to an average 57.5 nucleotides.

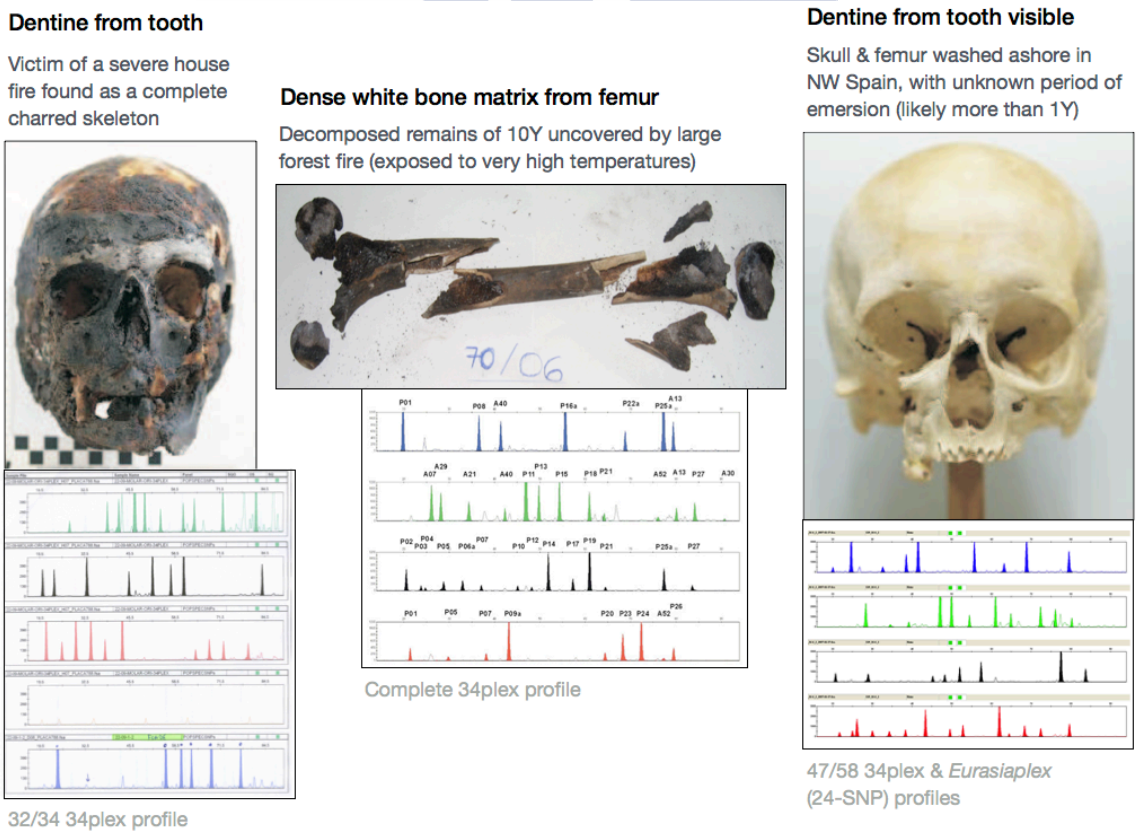
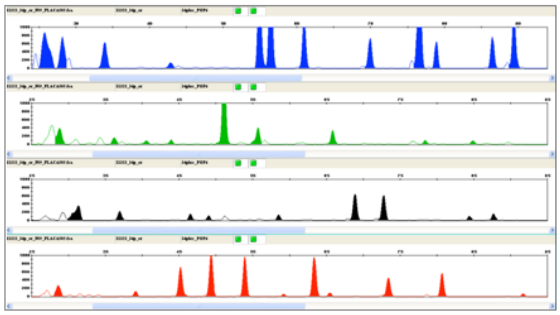


Figure D3. Examples of challenging SNP analyses to identify missing persons with successful ancestry SNP genotyping of DNA subject to high temperatures or degradative processes (or both).

34-SNP profile of sample A (dentine from tooth shown left) completeness was 88% =30/34 SNPs)



Bayes analysis with *Snipper*

Binary AIM classification of multiple individuals
with an Excel file of populations (.xlsx format)

A: *88% completeness* 9,355,471,189,361,381,376 times more likely E ASN than EUR

B: *12% completeness* 47 times more likely EUR than AFR

C: *56% completeness* 69,184,156,498,170 times more likely EUR than E ASN

D: *88% completeness* 237,256,787,732,999,681,605,632 times more likely EUR than E ASN

E: *97% completeness* 2,795,277,788,041,023,852,877,185,024 times more likely EUR than E ASN

Identifying genetic clusters with STRUCTURE

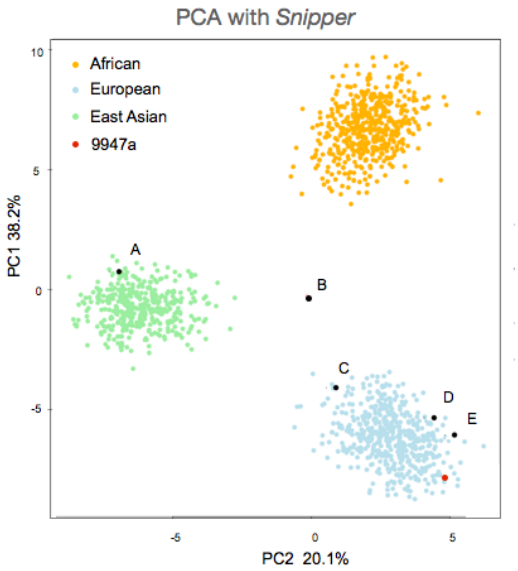


Figure D4. (Above) Summary details of a World War II soldier identification program of a mass grave site in Papua New Guinea. Five examples of 34-SNP analyses are shown that successfully identified one Japanese and three Australian soldiers. Profile completeness varied and needed to be taken into account when interpreting data, e.g. sample B was not classified and gave an unreliable Bayes likelihood, PCA position and STRUCTURE cluster membership proportion plot due to largely incomplete genotype data.

Figure D5. (Next page) MPS analysis of an archeological DNA sample extracted from skeletal remains of the 5th/6th and 12th/13th Century Volders burial site, Tyrol, Austria. Extract yielded 450 pg quantified with Quantifiler Duo, amplified with 25 PCR cycles (with or without 5 library amplifications). Average coverage (between replicates) exceeded 100 sequence reads in 48.5% of SNPs and was between a minimum 20 to 100 reads in 32.5%. 36 SNPs had coverage below 20 reads or no genotype calls in both replicates. The 82 SNPs with coverage >100 reads gave a cumulative random match probability (RMP) of 1.2E-33, five orders of magnitude better than the expected values from the 21 STRs of GlobalFiler. A relationship was discernible between a SNP's coverage and amplicon length indicated by the six average amplicon length values shown next SNPs ranked by coverage - from 99 to 137 nucleotides. The re-designed SNP panel removed 45 SNPs with genotyping problems and reduced the amplicon lengths of 57 to an average of 57.5 nucleotides. Although 67 SNPs kept the same primer designs, this panel is likely to provide improved sensitivity for the analysis of highly degraded DNA. The SNP composition of the final commercial release of the HID-Ion panel (the Precision ID set) is shown bottom right. MPS SNP analyses made at IMU, Innsbruck; data analysis made jointly between IMU and USC.

Volders DNA run in duplicate
gave above-average no calls
N / NN
QUAL=0
SNPs with No Calls or Dropouts
DNA from skeletal remains at Volders
archeological site, Tyrol, Austria
34
34 upper
Y-clade SNPs
43
43 K. Kidd
SNPs
48
48 SNPforID
SNPlex
(1 shared SNP)
124
124 SNPs
90 autosomal SNPs
The Applied Biosystems Precision
ID human identification SNP set
(commercially released May 2016)
Prototype HID-Ion
SNP panel amplicons
Applied Biosystems
Precision ID amplicons
99
117
119
120
123
137
47
68
108
130
45 SNPs removed from prototype
57 SNPs with reduced amplicon sizes (av. 57.5 nt)
67 SNPs retain primer designs of prototype
RMP
1.2E-33
GlobalFiler
1.2E-28
48.5% of SNPs: >100 x coverage
32.5% of SNPs: 20-100 x coverage
19% of SNPs: <20 x coverage
N / NN
QUAL=0

6. To identify and collect population variation data and genomic details of non-binary SNPs, comprising tri-allelic and tetra-allelic single nucleotide variation with multiple base substitutions detected by whole genome re-sequencing. To develop a panel of multiple allele SNPs for MPS of 250-300 markers obtained from screening the complete human variant catalog published by 1000 Genomes.

The initial chance discovery of tri-allelic SNPs at USC prompted further efforts to collect and validate any such markers that came to light, that showed informative allelic variability in the three alleles. This was less easy to achieve than the recognition of the best bi-allelic ancestry-informative SNPs, as tri-allelic variation is not detected using the dual-dye (Cy3-Cy5) system adopted for whole-genome SNP arrays that have been in widespread use since 2005. This precluded the population analysis of HGDP-CEPH variation generated by the Illumina 650,000 SNP array, which had helped identify so many binary ancestry-informative SNPs. The Hapmap project did not list any tri-allelic SNPs as it was based on use of array technology to identify SNP variation in the study populations chosen.

The first forensic set of tri-allelic SNPs developed for the purpose (**Thesis Paper #4, C. Phillips, et al, 2003** [166]) and the sets that followed (**Musgrave Brown, et al, 2006** [167]; Westen, et al, 2009 [112]; Zha, et al, 2012 [168]) were mainly focussed on the additional benefit of multiple-allele SNPs to provide indications of DNA mixtures from detection of three different alleles at any one SNP position in a mixed profile. In a parallel study of markers termed "SWaP" SNPs (G/C=S or A/T=W Amid Palindromes), only the rs5030240 tri-allelic SNP was included as a proof-of-concept to demonstrate their effectiveness for mixture detection (the underlying principle of SWaP SNPs is explained in **Figure D6**). The study of Westen that followed in 2009 also concentrated on mixture detection by developing a SNaPshot multiplex of 16 tri-allelic SNPs (although one was actually a binary SNP), from a candidate pool of 31, that was able to detect the minor mixture component at ratios as low as 1:8. Westen also noted the increased discrimination power of tri-allelic SNPs and this factor has informed more recent studies at USC aimed at creating a missing persons MPS panel that maximises the relationship testing capacity of the PCR multiplex while keeping the characteristic of very short amplicons more readily achieved in SNP amplification than STRs. Finally, Zha, et al, reported 20 tri-allelic SNPs, discovered by Pyrosequencing and compiled into a 20-plex forensic SNaPshot assay, also with emphasis on mixture detection. The best tri-allelic SNPs for ancestry inference from the studies of USC, Westen and Zha are summarised in the allele frequency pie charts of **Box 7**, panel 3. In the same box, Panel 2 shows 28 multiple-allele SNPs (only one a tetra-allelic SNP) that were selected specifically for maximum heterozygosity in Europeans, in order to maximise the chances of three-allele profile patterns for mixture detection and to raise relationship testing statistics when analysing highly degraded DNA. Note that not all of the 27 tri-allelic SNPs in Panel 2 have high polymorphism levels in the other population groups (e.g. rs2052215 and rs6940924). For the 27 markers in panel 3, ancestry-informative allele frequency patterns are seen most frequently in Africans and Oceanians, with these two population groups representing the extremes of distance from a hypothetical human migration focus-

of-initiation from Addis Ababa. Individuals from African populations tend to show all three alleles, often with the third allele at a low frequency, and as populations are positioned further from Africa (along an idealised axis of human migration that equates to genetic distance), the variability decreases in the form of a reducing frequency for one or two of the alleles, or their complete disappearance in Native Americans and Oceanians - as exemplified by rs17287498, where two alleles have been lost in these population groups and a sharp contrast exists for their frequency distributions in Eurasians and Africans. This trend indicates a more fluid evolution of tri-allelic SNP allele frequency distributions in human populations compared to binary SNPs, although the tri-allelic loci here have been selected for their high population divergence. Loss of heterozygosity in genetic markers with increasing 'distance from Addis Ababa' is a widely recorded phenomenon (Li, et al, 2008, Fig. 3B [19]). However, in general, tri-allelic SNPs are more divergent than their bi-allelic counterparts and this is most likely to be a direct result of the increased chances for random genetic drift influencing allele frequencies amongst six genotypes, compared to its influence when only three genotypes occur. Similarly, allelic variation can be lost in small founding groups when at low frequencies, and this is more likely to happen in one of the alleles of SNPs with three alleles than in SNPs with two, hence, the noticeably reduced variation levels seen in Native Americans and Oceanians in *all* the multiple-allele SNPs shown in **Box 7**.

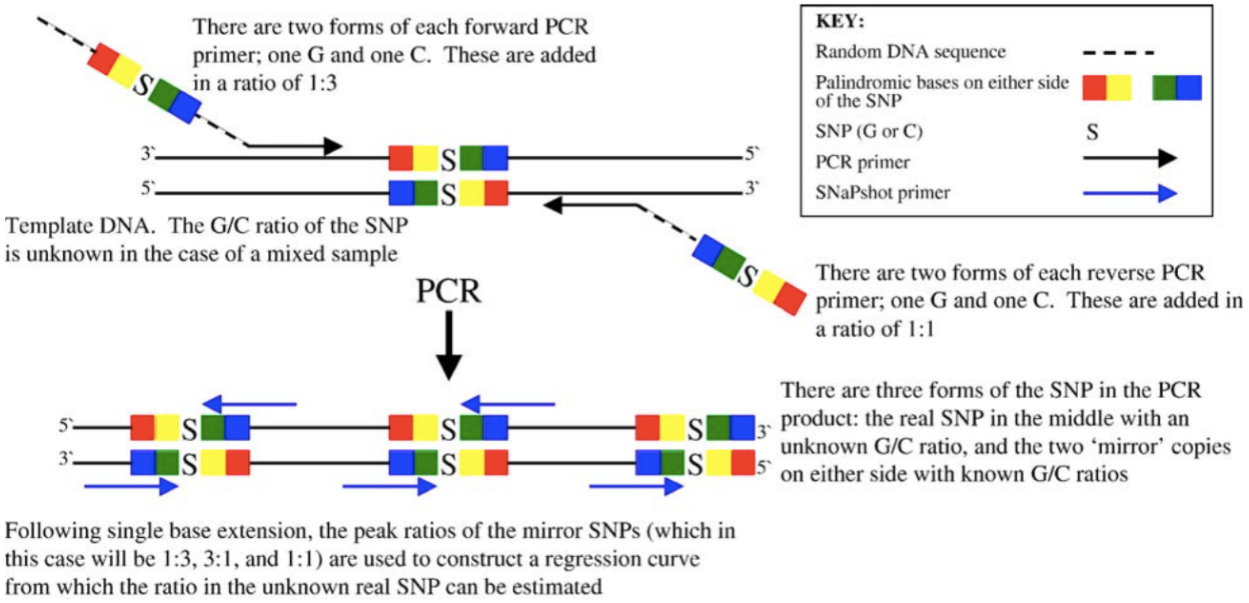


Figure D6. The principle of SWaP SNPs developed during SNPforID mid-phase activities. Each SNP is amplified using PCR primers with modified tails consisting of a random sequence that contains a replicate of the SNP under examination, (the 'real' SNP), and the two bases of the palindromic sequence on either side. So two forms of each primer are used, representing the two alleles of the locus. Primers are added to the PCR in known ratios to create amplicons containing the real SNP with an unknown allele ratio, flanked by two 'mirror' copies of the SNP with known allele ratios. Each of the three SNPs is then interrogated in a single-base extension (SBE) reaction. As the SNPs are located within palindromic sequences, the 3' environment for the SBE primers is very similar. The peak ratios from the known mirror SNPs can be used to construct a standard curve from which the allele ratio in the real SNP is estimated. The tri-allelic SNP rs5030240 was used as an additional mixture detection system in the same SWaP SNP multiplex.

The much higher PCR multiplex scales possible with MPS prompted the final phase of the development of forensic tri-allelic SNP panels. These panels will be primarily developed for enhanced relationship testing of degraded DNA for the identification of missing persons, but also SNPs with strong population divergence will be compiled to establish much larger ancestry panels that could potentially differentiate more closely related populations with better likelihood ratios than possible with bi-allelic SNPs (for the reasons of increased effects of genetic drift outlined above). Selection of suitable SNPs involved a first step of compiling all multiple-allele single nucleotide variants identified by 1000 Genomes (SNVs: SNPs plus variant sites where one or more variants were at <1% frequency in one or multiple populations). Tri-allelic SNPs identified in 1000 Genomes Phase I variant site analyses were initially removed from data releases under the premise that they were possible sequence mis-alignment or artefacts; until late 2014, when the sites that had been confirmed to have three or four alleles at one position were re-instated into the variant catalog [169]. A systematic analysis of the final fully-curated Phase III variant data release revealed a total of 274,425 multi-allelic SNVs of which 15,055 are X-chromosome and none have so far been compiled for the Y. On the 22 autosomes, 1,625 CNVs are tetra-allelic and these loci in the 1000 Genomes variant catalog have been studied in detail at USC and the characteristics of the most forensically informative for identification and ancestry inference were reported in **Thesis Paper #2, C. Phillips et al, 2015** [170]. This leaves 272,800 tri-allelic SNVs, of which most will have one or two alleles below 1% population frequencies, and consequently have no utility for forensic analysis. Nevertheless, such a large catalog requires a systematic method to identify the most informative SNPs for each forensic purpose (**Phillips et al, *in preparation*** [171]). The best tri-allelic SNPs for identification are those with the highest heterozygosity levels, whether applied to the analysis of degraded criminal casework material, where SNPs would offer better success than STRs but cannot easily detect mixtures; or applied to relationship testing where the presence of up to six different genotypes in a complex pedigree is desirable. Therefore, making simple population group allele frequency estimates from the 1000 Genomes genotype data and deriving the estimated heterozygosity per SNP, per population, is relatively straightforward. It is preferential to have high heterozygosity levels in all population groups (so called 'universality'; which though desirable, should not be the defining characteristic for choosing the best forensic SNPs - see **Fig. D1**). Therefore, simple heterozygosity range limits were set at the maximum 0.66 value down to 0.55 or 0.5 minimum average heterozygosity for European, African and East Asian population group allele frequencies, defining *all* of the best tri-allelic SNPs as more informative per marker than the best bi-allelic loci. South Asian and Peruvian (the proxy population for gauging Native American variability) data was ignored, but generally their heterozygosity values for the top SNPs were close to those found in the three main population groups. Some SNPs had population-specific heterozygosity values below 0.5 and a small proportion were below 0.35. However, the value of the occasional rare allele in relationship testing should not be overlooked, and low heterozygosity was not confined to any one population group, so the effect of below-average informativeness in a proportion of loci per population was balanced out across the whole candidate SNP set. Relaxing the minimum three-population average heterozygosity to 0.5 allowed many more candidates to be identified that could provide informative markers but would

also make it easier to apply strict selection criteria for other key characteristics, such as optimum amplicon length, and reject SNPs failing such requirements. These genomic characteristics for rejection of candidates were considered to be: poor context sequence making the SNP unsuitable for MPS due to the risk of misalignment problems; limited ability to design very short amplicons in the PCR; and the need to avoid clusters of SNPs very close together in a genomic position, thus likely to be linked.

One final critical step is to check each SNP position with the NCBI BLAST sequence homology search tool in order to ensure candidates occupy unique positions in the genome. It is likely that a proportion of SNPs have three alleles because their flanking sequence is found in multiple positions and the 'third' allele is a single nucleotide difference at one position combined with a binary SNP in another, or two positions having SNPs with a total of three different alleles. Although these would be discernible in MPS sequence coverage values, the design of unique primers is obviously compromised. This may or may not be a common explanation for tri-allelic patterns. In a systematic search of 41 tetra-allelic SNPs informative for identification purposes, 16 were found in non-unique regions, with BLAST analysis reporting multiple hits on different chromosomes for the sequences around these loci. The identification of multiple-allele SNPs is likely to positively select sites in non-unique sequence tracts that could carry divergent SNPs (having different binary allele combinations as described above). A 40% loss of suitable candidates in this case still leaves a sufficient number to merit further analysis and would leave more than 2,000 tri-allelic candidates to assess for forensic purposes. The final multiplex size envisaged for an MPS test also has a bearing on compiling candidate loci, as it is very likely that more than 10% of the 2000 candidates would meet the requirements of good identification loci, while the upper limits of PCR multiplexing for MPS target capture may be far higher than 200 in the near future. Despite the emphasis on MPS-scale multiplexes, USC have finally compiled a set of 29 multiple-allele SNPs into a SNaPshot test that can be used for forensic identification, but allows the inference of ancestry and detection of mixtures (**Fondevila, et al, *in preparation*** [172]).

Simulations of genotype combinations in different related pairs: parent/child; full sibs; half sibs; first cousins; demonstrate a clear benefit from using tri-allelic SNPs that can analyse six genotypes in relationship test scenarios. The distribution of relationship likelihood ratios (comparing the claimed relationship hypothesis to the unrelated pair hypothesis) are shown in **Figure D7-A**, for three close-relative-pair scenarios that suggest tri-allelic SNPs and Microhaplotypes in sufficient numbers (in the simulations 1400 loci were used with 1000 Genomes allele/haplotype frequencies) achieve clear separations of the distribution of likelihood ratio values for the claimed relationship (pink LR distributions) and unrelated pairs (green) in all cases up to first cousins and in about 35% of cases in second cousins (LR ranges in the non overlapping distributions). **Figure D7-B** compares 140 bi-allelic SNPs and the equivalent number of tri-allelic SNPs. In each case there is a shift to the right (better likelihoods of the claimed relationship) using tri-allelic SNPs, with a reducing effect on the likelihood ranges as the relationship becomes more distant and the 'unrelated pair hypothesis' likelihoods show increasing overlap with those of the claimed relationship. Finally, the value of a rare allele-3 in a small

proportion of tri-allelic SNPs, as should be expected to happen frequently (or in many populations), should not be overlooked and may be lost with a focus on highest possible heterozygosities from 'ideal' 0.33-0.33-0.34 frequency splits. For example, if tri-allelic SNPs with 10% allele-3 frequencies were chosen, a multiplex of 200 would find this third allele present in the analysed sample and surviving relatives in about 20 loci. However, the value of finding such rare alleles in the statistical inference of relatedness would be much higher than the presence of a common allele and relationship statistics consequently benefit from these allelic patterns disproportionately. To explore this idea, simulations can also be made with theoretical tri-allelic SNPs that have different allele frequency characteristics. These are summarised in **Figure D7-C**, where it can be seen that use of SNPs with allele-3 frequencies of 5% (0.05) will not reduce the relationship likelihoods markedly. This data suggests that for those SNPs with much lower population-specific heterozygosities compared to the three-population average values, their reduced variability does not unduly affect the value of the final SNP set for relationship testing applications. Note that the half-sib simulations in Fig D5-C indicate the small amount of overlapping likelihoods almost disappear when increasing multiplex levels from 140 to 300 tri-allelic SNPs Therefore, tri-allelic SNPs are particularly beneficial for relationship testing in enlarged multiplexes (compared to STRs in much smaller numbers of loci), while they preserve the value of short amplicons for analysing degraded DNA. An over-emphasis on high heterozygosity values may overlook the value of the rare allele-3 in a proportion of SNPs in any relationship testing scenario, so it is worthwhile to concentrate on obtaining the shortest possible amplicon lengths and finding candidate loci with good quality flanking sequences, rather than perfect frequency characteristics when building a large-scale MPS multiplex for missing persons identification.

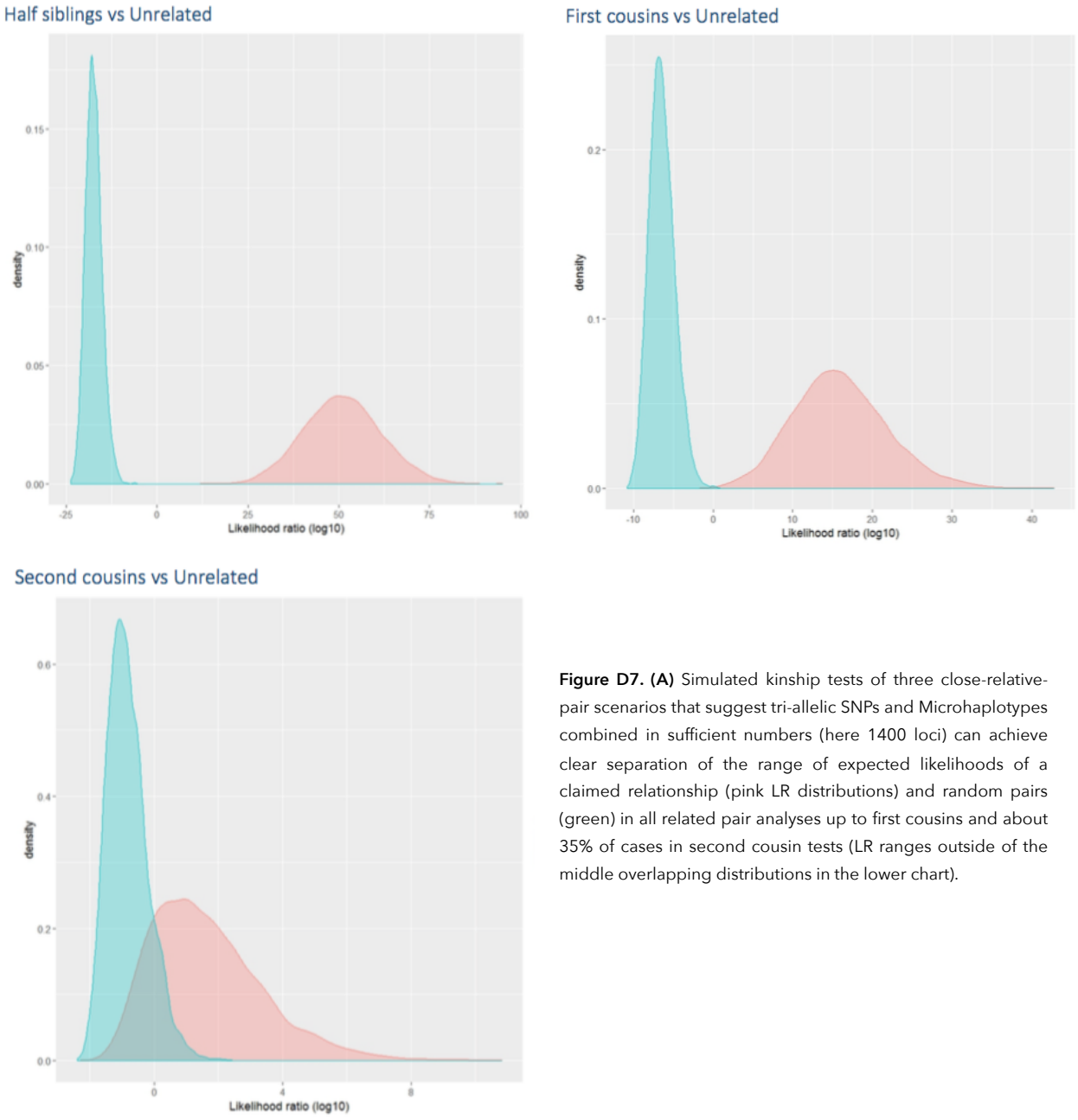


Figure D7. (A) Simulated kinship tests of three close-relative-pair scenarios that suggest tri-allelic SNPs and Microhaplotypes combined in sufficient numbers (here 1400 loci) can achieve clear separation of the range of expected likelihoods of a claimed relationship (pink LR distributions) and random pairs (green) in all related pair analyses up to first cousins and about 35% of cases in second cousin tests (LR ranges outside of the middle overlapping distributions in the lower chart).

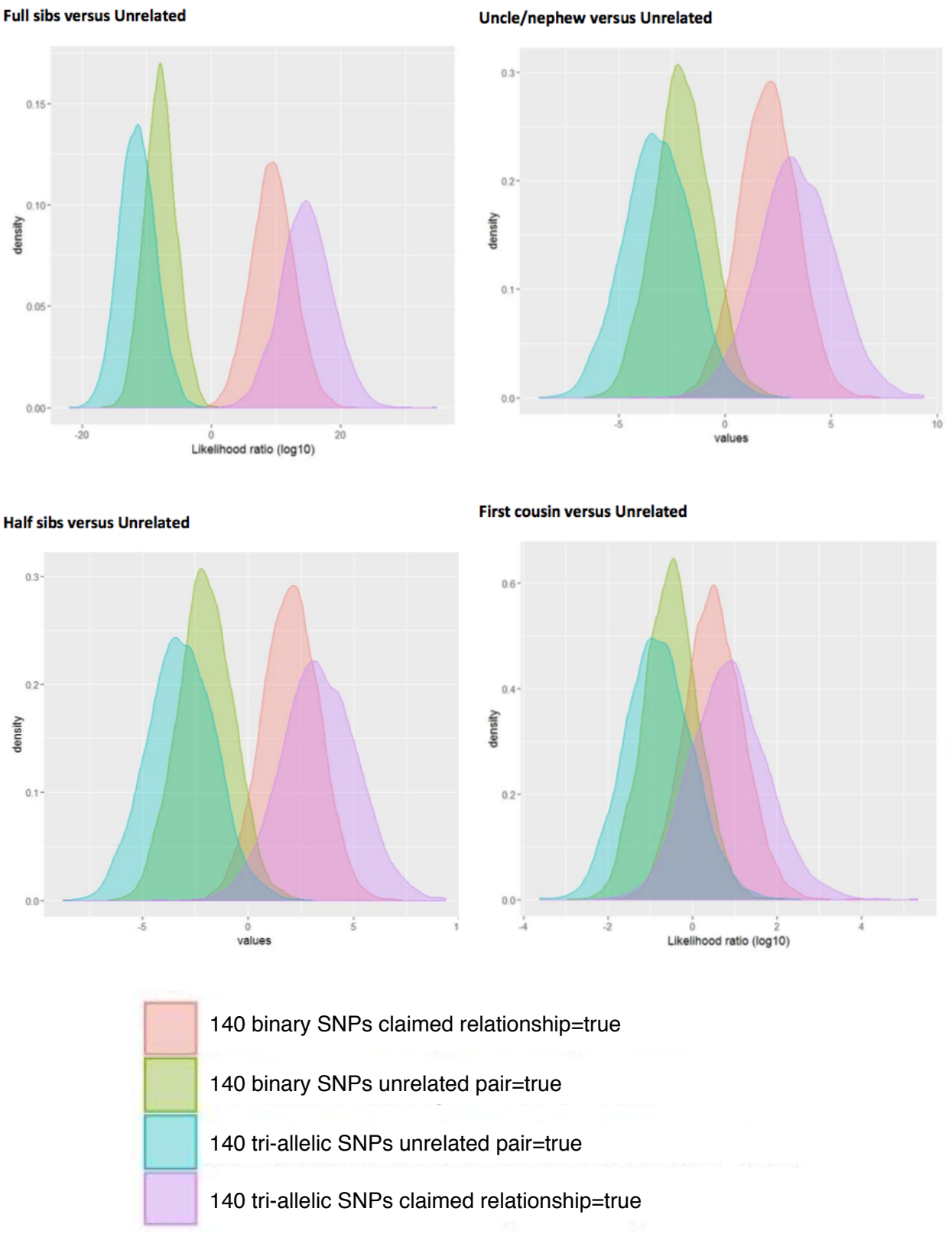


Figure D7. (B) Comparison of informativeness of binary and tri-allelic SNPs in numbers of 140 loci, to differentiate closely related pairs (pink shades) vs. unrelated pairs (green shades). Tri-allelic SNPs achieve better separations in each case.

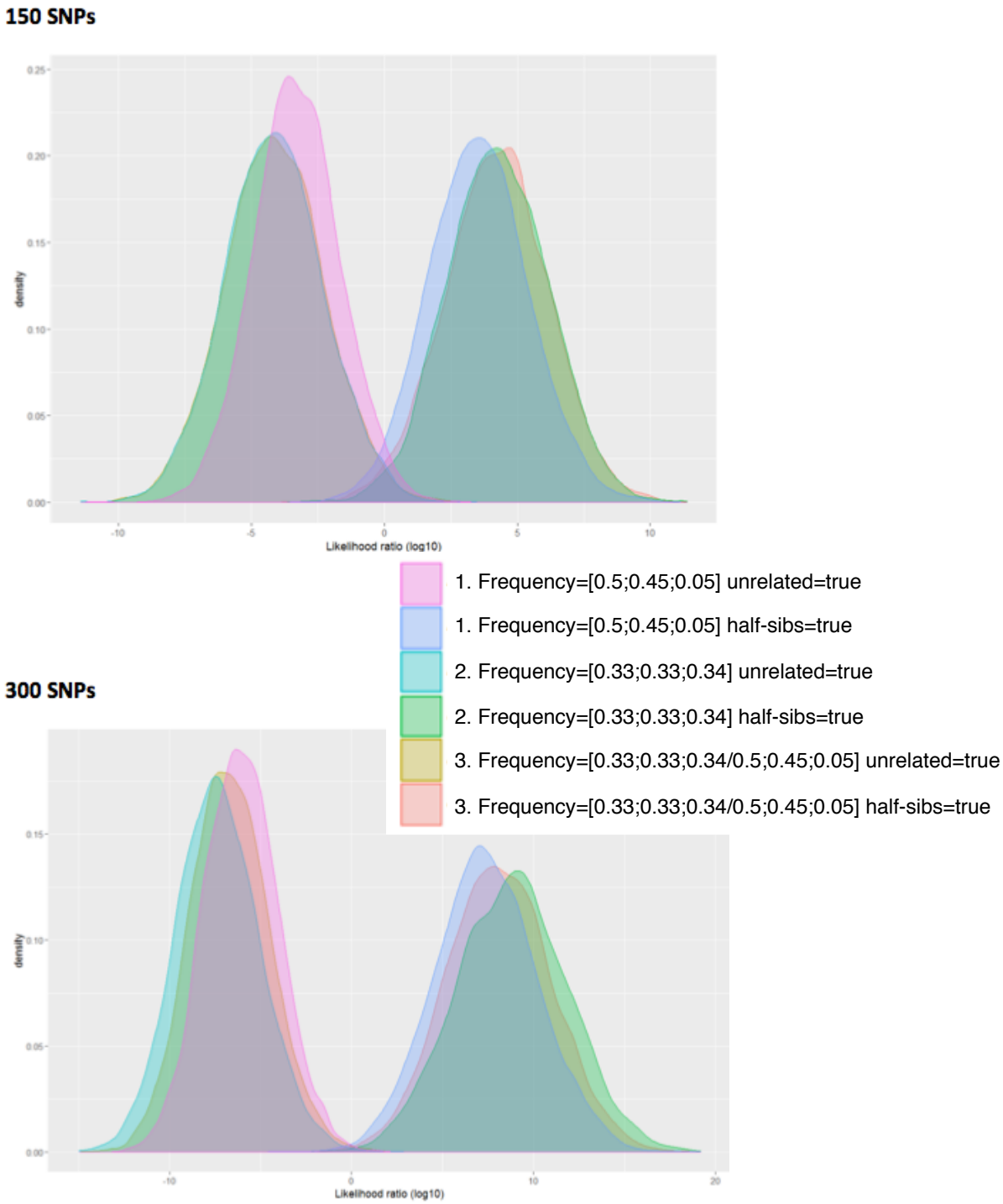


Figure D7. (C) The effect of allele frequency distributions in the component loci in a set of 150 or 300 tri-allelic SNPs, using just half-sib vs unrelated pair scenarios. Three distributions of allele frequencies comprised: 1. One minor [0.50, 0.45, 0.05], 2. Equal [0.33, 0.33, 0.34], 3. a mix of 1and 2 (half had "Equal", the other half "One minor"). Differences in LR value distributions are marginal, even when all loci have one allele at only 5% frequencies, suggesting not all SNPs have to be perfectly equal to be informative enough.

Building a forensic ancestry analysis panel with the tri-allelic SNPs identified from 1000 Genomes requires more carefully gauged selection criteria that are based on contrasts in allele frequency distributions between populations across all three alleles per SNP. **Figure D8** illustrates the problem of constructing appropriate criteria by showing four different tri-allelic SNPs with ancestry-informative allelic distributions for three hypothetical populations A-C. SNP 1 is clearly the most informative ancestry marker with a different allele at fixation in each population. Were such an extreme allele frequency SNP to exist in reality, it would have heterozygosities of 0 in each population and would not be detected by ranking heterozygosity values: as was done to find the best identification SNPs. SNPs 2 and 3 have very similar heterozygosities but SNP 3 could be considered an 'ideal' ancestry-indicative marker, as each population has a different major allele (green in A, blue in B, red in C). The last SNP 4 is tri-allelic in population A, but bi-allelic in B and C, so it lacks power to differentiate the last two population and the population-specific allele-3 in A may not be frequent enough to be informative compared to the best bi-allelic AIM-SNPs, although often this type of pattern can be useful when the population-specific allele is found in a population that is not readily differentiated form others such as South Asians. Therefore, a combination of ranking candidate loci on their heterozygosity levels, the presence of three alleles in most populations and allele frequency divergence is necessary. From the candidate list of all 1000 Genomes tri-allelic SNPs with allele-3 frequencies greater than 10%, SNPs were ranked by average heterozygosity with less stringent lower limits down to 0.25. However, to ensure sufficient power in the best AIMs to differentiate multiple populations, four-five of five populations groups (South Asian and PEL Americans included) need to show three allele variation. From the resulting ranked list, the allele frequency standard deviations (across five population groups) were calculated - these are easier to estimate than Shannon's Divergence and their values are highly correlated with Divergence for any one SNP. This allowed a sufficient number of tri-allelic SNPs informative for ancestry to be compiled. Thirty are now part of a single 165-marker ancestry multiplex developed for forensic ancestry inference using MPS.

The most recent application of ancestry-informative multiple-allele SNPs has been the construction of an MPS forensic panel dedicated to the differentiation of East Asian from European, South Asian and Oceanian population groups. The represents a revision of the Global AIMs panel described above and in other sections to preserve the balanced ancestry-differentiation power of the original panel, by condensing down the core set of binary AIMs to 84, but keeping the genotype data from these SNPs for the analysis of admixture ratios from the alleles detected in the analysed individual. In this way, a further 81 ancestry-informative loci were added, comprising 22 Microhaplotypes, 35 multiple-allele SNPs (including all 8 of those in the original Global panel) and 25 SNPs that consist of most of the original *Eurasiaplex* markers plus additional South Asian-informative loci. This panel is currently undergoing evaluation and optimisation for the newly launched S5 Ion Torrent MPS chemistry in USC and Innsbruck laboratories. The allele frequency characteristics of 32 of the 35 tri-allelic SNPs successfully incorporated into this forensic 165-marker multiplex (see section 8) are shown in **Figure D9.**

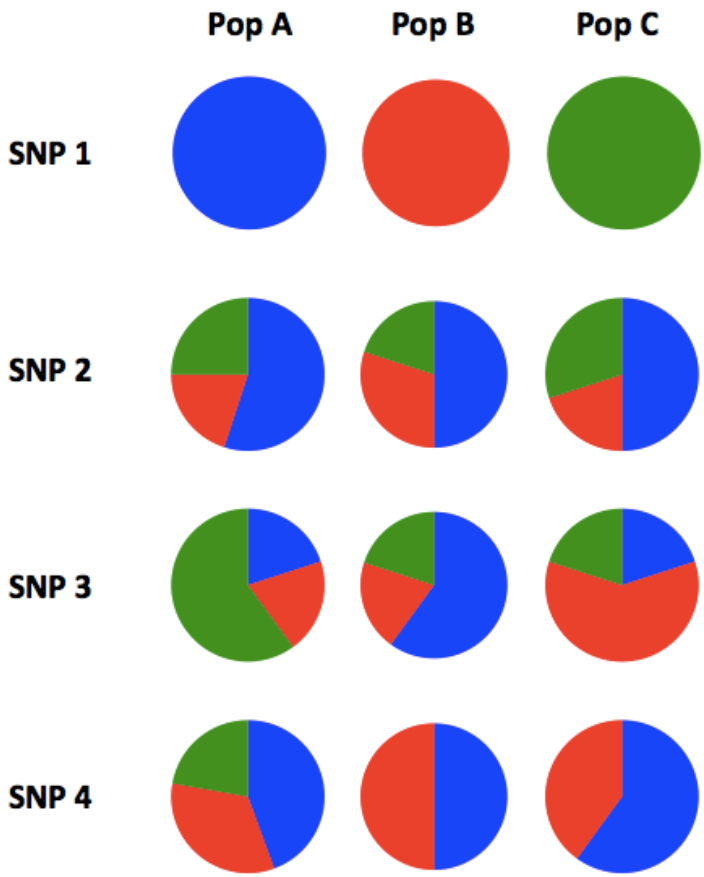


Figure D8. Contrasting population-indicative allele frequency distributions observed across three populations in four tri-allelic SNPs.

SNP 1: a perfect ancestry informative tri-allelic SNP.

SNP 2: a SNP with the least informative allele frequency distribution for ancestry purposes, but typical of loci chosen for identity purposes.

SNP 3: a typical optimum ancestry informative tri-allelic SNP with a different major allele in each population compared.

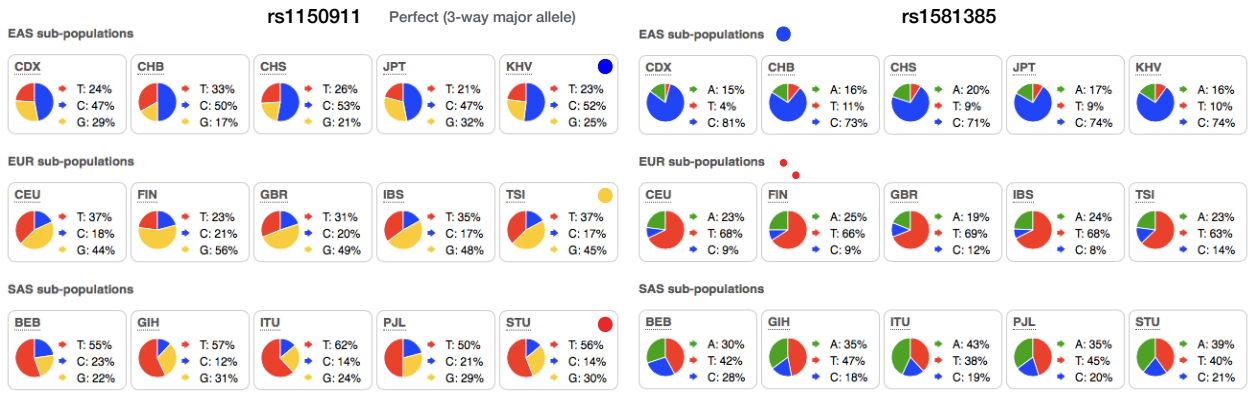
SNP 4: a typical ancestry informative tri-allelic SNP with a population-specific allele. This type of marker can be useful for analysing specific populations that may not always be readily differentiated from each other, e.g. if Pop A=South Asians, Pop B=Europeans.

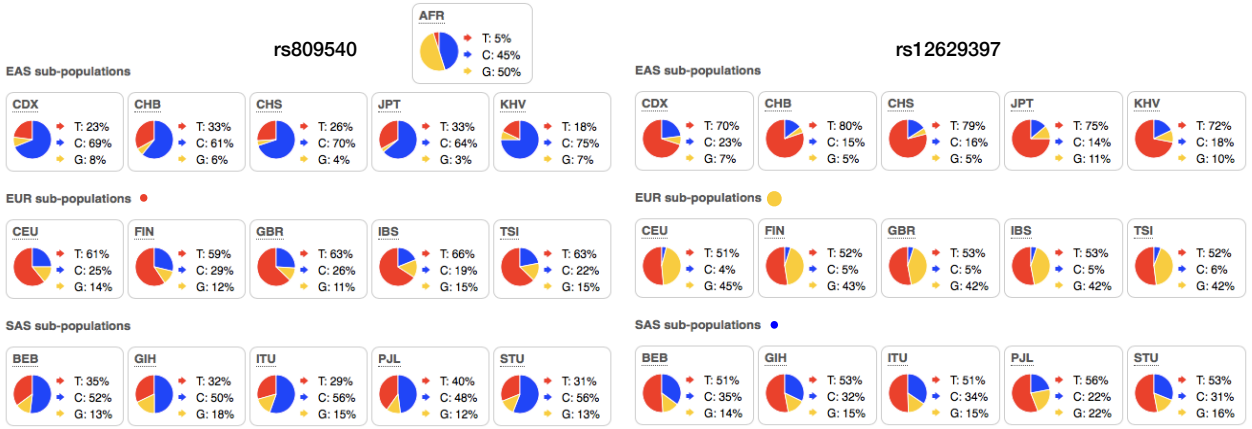
7. To optimise and extend the practicality of ancestry analysis using Indels and STRs to establish mixed-marker approaches that allow ancestry inference from standard DNA profiling data (when evidential material is no longer available for additional DNA tests), or when a more secure system is required for the analysis of mixed DNA than is possible with SNaPshot SNP genotyping. To develop a frequency-based classifier in Snipper applicable to forensic STR data, but also extending the scope of ancestry inference to haplotype data such as Y-SNPs and autosomal SNP microhaplotypes. To enhance Snipper with fixed training set data for 46 Indels alongside 34 SNPs and to extend SPSmart to forensic Indel sets.

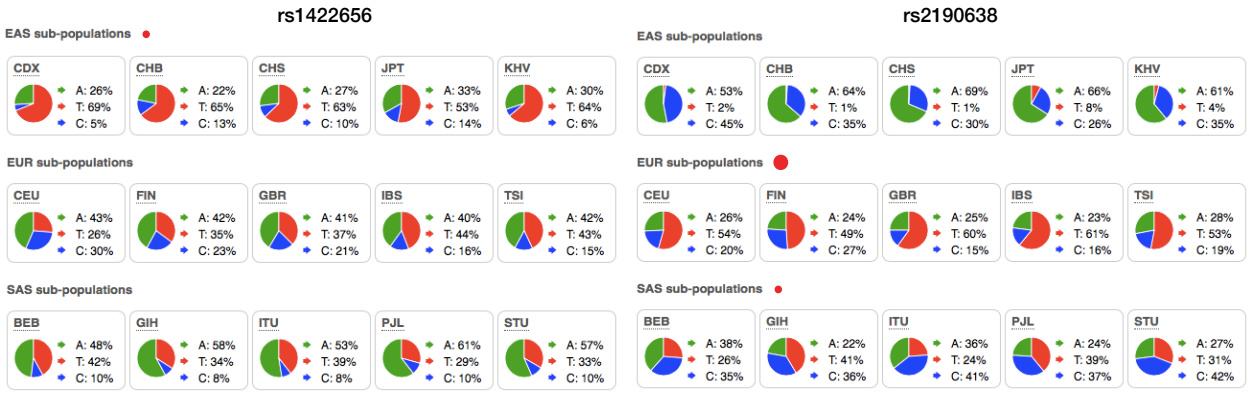
For forensic ancestry analysis, SNPs are the most informative markers, but they suffer from an impaired ability to detect mixed DNA; because there are marked differences in the strength of dye signals linked to each allele detected with SNaPshot. These contrasting signal strengths lead to imbalance amongst heterozygote peak pairs indistinguishable from the differences in signal ratios seen in mixed DNA. This characteristic of SNaPshot-based SNP genotyping prompted the development of two alternative ways to infer ancestry from CE-based forensic tests: use of ancestry-informative STRs and use of ancestry-informative Indels. Both techniques rely on dye-liked primers to label the PCR products that after

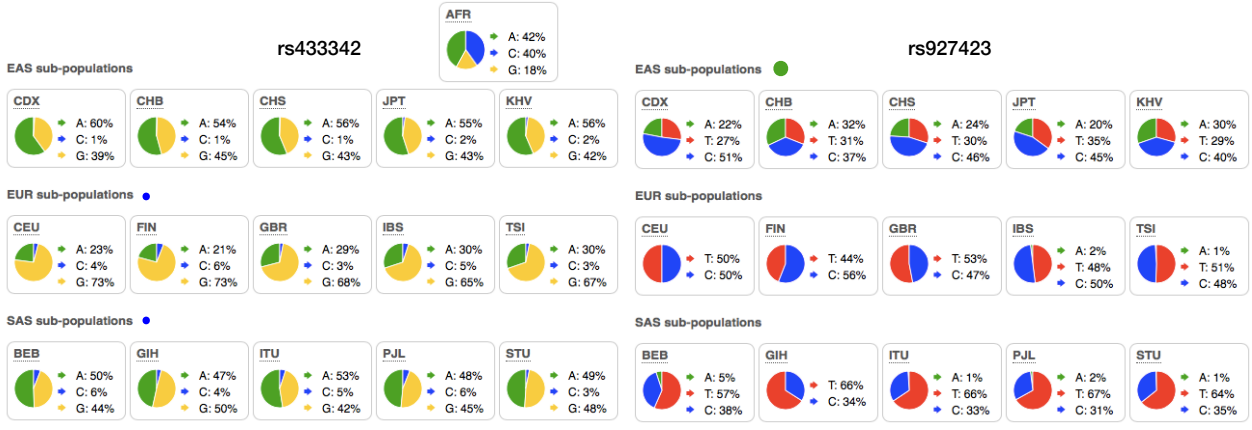
amplification go directly into CE (as with all forensic STR genotyping systems) - thus preserving the direct relationship between input DNA and signal strength. When the relationship between input DNA and CE signals is directly correlated in this way, the peaks in normal unmixed heterozygotes are much more balanced and have a distinct pattern of peak balance compared to ratios of the components in mixed DNA (except from ratios close to 1:1 mixtures of two homozygote genotypes). The key factor in dye-linked PCR is the use of a single dye to label the allelic products of each locus, so regardless of the differences in signal strength that exist between 6-FAM and NED, for example, each allele pair in the STR or Indel locus have comparable peak heights, as they are not compared across two dyes.

Figure D9. (This legend placed above graphics for clarity) Individual 1000 Genomes population allele frequency distributions for 32 of 35 multiple-allele SNPs of the new gAIMs2 forensic MPS panel for the three main population groups of East Asia (EAS); Europe (EUR) and South Asia (SAS) it is designed to analyse in detail. Where marked contrasts exist between African populations and these three population groups, summary pie charts for that group are shown (AFR). So-called "perfect" multiple-allele AIM-SNPs (loci with three different major alleles in different population groups) are shown with coloured dots on the right, the informative alleles are indicated with large dots (significant data) or small dots on the left. Some promising patterns are already evident: 1) rs65004633 and rs1612734 are tetra-allelic SNPs with population specific alleles in Africans and Africans plus East Asians respectively; 2) the presence of a South Asian-specific allele not found in European populations is very rare, and rs914468 was selected for this property of its third allele; 3) the well-established SNPs of rs4540055 and rs5030240 are amongst the least informative loci; 4) although rs17287498 is not currently mapped to a set location in the human genome, it was successfully sequenced and aligned in gAIMs1.

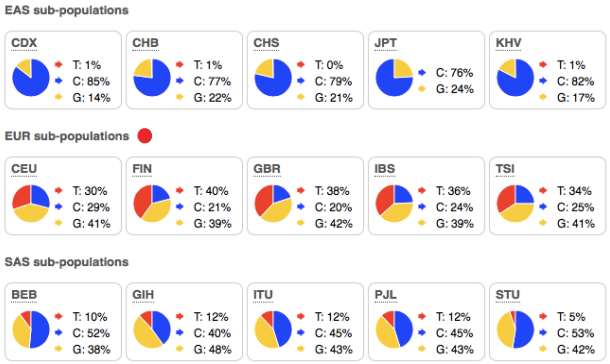

rs1150911
Perfect (3-way major allele)
EAS sub-populations
CDX T: 24% C: 47% G: 29%
CHB T: 33% C: 50% G: 17%
CHS T: 26% C: 53% G: 21%
JPT T: 21% C: 47% G: 32%
KHV T: 23% C: 52% G: 25%
EUR sub-populations
CEU T: 37% C: 18% G: 44%
FIN T: 23% C: 21% G: 56%
GBR T: 31% C: 20% G: 49%
IBS T: 35% C: 17% G: 48%
TSI T: 37% C: 17% G: 45%
SAS sub-populations
BEB T: 55% C: 23% G: 22%
GIH T: 57% C: 12% G: 31%
ITU T: 62% C: 14% G: 24%
PJL T: 50% C: 21% G: 29%
STU T: 56% C: 14% G: 30%
rs1581385
EAS sub-populations
CDX A: 15% T: 4% C: 81%
CHB A: 16% T: 11% C: 73%
CHS A: 20% T: 9% C: 71%
JPT A: 17% T: 9% C: 74%
KHV A: 16% T: 10% C: 74%
EUR sub-populations
CEU A: 23% T: 68% C: 9%
FIN A: 25% T: 66% C: 9%
GBR A: 19% T: 69% C: 12%
IBS A: 24% T: 68% C: 8%
TSI A: 23% T: 63% C: 14%
SAS sub-populations
BEB A: 30% T: 42% C: 28%
GIH A: 35% T: 47% C: 18%
ITU A: 43% T: 38% C: 19%
PJL A: 35% T: 45% C: 20%
STU A: 39% T: 40% C: 21%


rs809540
AFR T: 5% C: 45% G: 50%
EAS sub-populations
CDX T: 23% C: 69% G: 8%
CHB T: 33% C: 61% G: 6%
CHS T: 26% C: 70% G: 4%
JPT T: 33% C: 64% G: 3%
KHV T: 18% C: 75% G: 7%
EUR sub-populations
CEU T: 61% C: 25% G: 14%
FIN T: 59% C: 29% G: 12%
GBR T: 63% C: 26% G: 11%
IBS T: 66% C: 19% G: 15%
TSI T: 63% C: 22% G: 15%
SAS sub-populations
BEB T: 35% C: 52% G: 13%
GIH T: 32% C: 50% G: 18%
ITU T: 29% C: 56% G: 15%
PJL T: 40% C: 48% G: 12%
STU T: 31% C: 56% G: 13%
rs12629397
EAS sub-populations
CDX T: 70% C: 23% G: 7%
CHB T: 80% C: 15% G: 5%
CHS T: 79% C: 16% G: 5%
JPT T: 75% C: 14% G: 11%
KHV T: 72% C: 18% G: 10%
EUR sub-populations
CEU T: 51% C: 4% G: 45%
FIN T: 52% C: 5% G: 43%
GBR T: 53% C: 5% G: 42%
IBS T: 53% C: 5% G: 42%
TSI T: 52% C: 6% G: 42%
SAS sub-populations
BEB T: 51% C: 35% G: 14%
GIH T: 53% C: 32% G: 15%
ITU T: 51% C: 34% G: 15%
PJL T: 56% C: 22% G: 22%
STU T: 53% C: 31% G: 16%


rs1422656
EAS sub-populations
CDX A: 26% T: 69% C: 5%
CHB A: 22% T: 65% C: 13%
CHS A: 27% T: 63% C: 10%
JPT A: 33% T: 53% C: 14%
KHV A: 30% T: 64% C: 6%
EUR sub-populations
CEU A: 43% T: 26% C: 30%
FIN A: 42% T: 35% C: 23%
GBR A: 41% T: 37% C: 21%
IBS A: 40% T: 44% C: 16%
TSI A: 42% T: 43% C: 15%
SAS sub-populations
BEB A: 48% T: 42% C: 10%
GIH A: 58% T: 34% C: 8%
ITU A: 53% T: 39% C: 8%
PJL A: 61% T: 29% C: 10%
STU A: 57% T: 33% C: 10%
rs2190638
EAS sub-populations
CDX A: 53% T: 2% C: 45%
CHB A: 64% T: 1% C: 35%
CHS A: 69% T: 1% C: 30%
JPT A: 66% T: 8% C: 26%
KHV A: 61% T: 4% C: 35%
EUR sub-populations
CEU A: 26% T: 54% C: 20%
FIN A: 24% T: 49% C: 27%
GBR A: 25% T: 60% C: 15%
IBS A: 23% T: 61% C: 16%
TSI A: 28% T: 53% C: 19%
SAS sub-populations
BEB A: 38% T: 26% C: 35%
GIH A: 22% T: 41% C: 36%
ITU A: 36% T: 24% C: 41%
PJL A: 24% T: 39% C: 37%
STU A: 27% T: 31% C: 42%

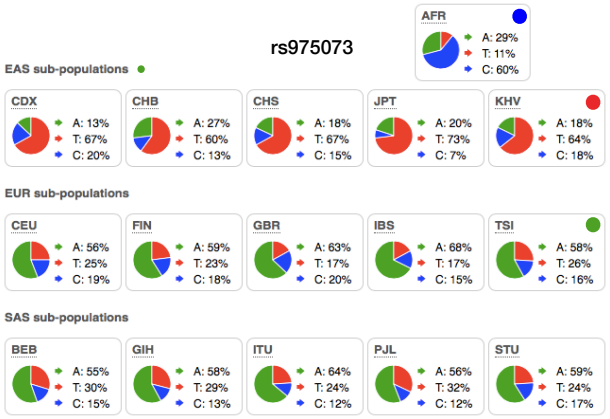

rs433342
AFR A: 42% C: 40% G: 18%
EAS sub-populations
CDX A: 60% C: 1% G: 39%
CHB A: 54% C: 1% G: 45%
CHS A: 56% C: 1% G: 43%
JPT A: 55% C: 2% G: 43%
KHV A: 56% C: 2% G: 42%
EUR sub-populations
CEU A: 23% C: 4% G: 73%
FIN A: 21% C: 6% G: 73%
GBR A: 29% C: 3% G: 68%
IBS A: 30% C: 5% G: 65%
TSI A: 30% C: 3% G: 67%
SAS sub-populations
BEB A: 50% C: 6% G: 44%
GIH A: 47% C: 4% G: 50%
ITU A: 53% C: 5% G: 42%
PJL A: 48% C: 6% G: 45%
STU A: 49% C: 3% G: 48%
rs927423
EAS sub-populations
CDX A: 22% T: 27% C: 51%
CHB A: 32% T: 31% C: 37%
CHS A: 24% T: 30% C: 46%
JPT A: 20% T: 35% C: 45%
KHV A: 30% T: 29% C: 40%
EUR sub-populations
CEU T: 50% C: 50%
FIN T: 44% C: 56%
GBR T: 53% C: 47%
IBS A: 2% T: 48% C: 50%
TSI A: 1% T: 51% C: 48%
SAS sub-populations
BEB A: 5% T: 57% C: 38%
GIH T: 66% C: 34%
ITU A: 1% T: 66% C: 33%
PJL A: 2% T: 67% C: 31%
STU A: 1% T: 64% C: 35%

rs1169671

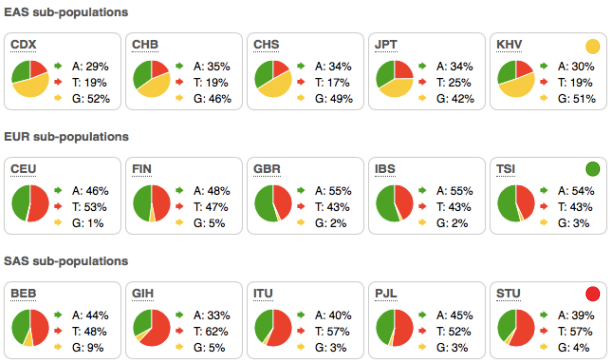


rs4540055

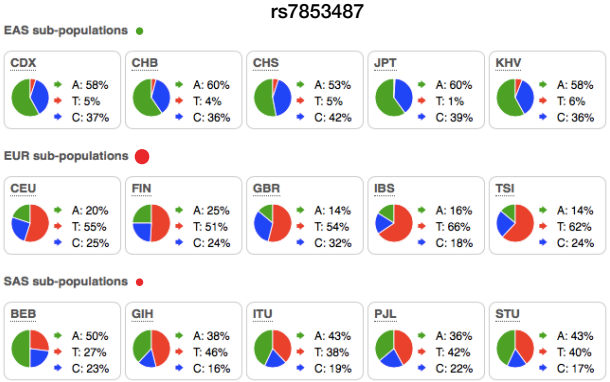
rs975073



rs7736783 Perfect (3-way major allele)



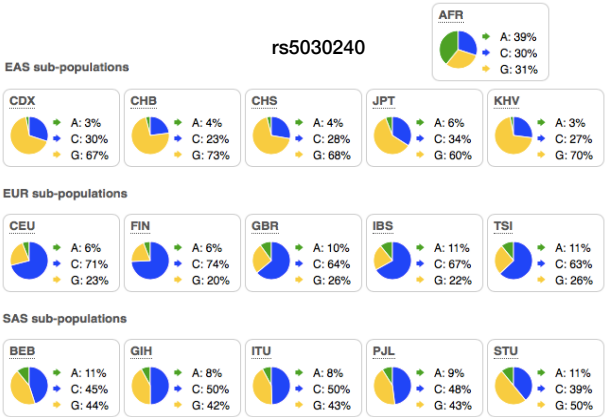
rs7853487



rs17287498 successfully genotyped in gAIMs1

Not mapped in 1000 Genomes

rs5030240



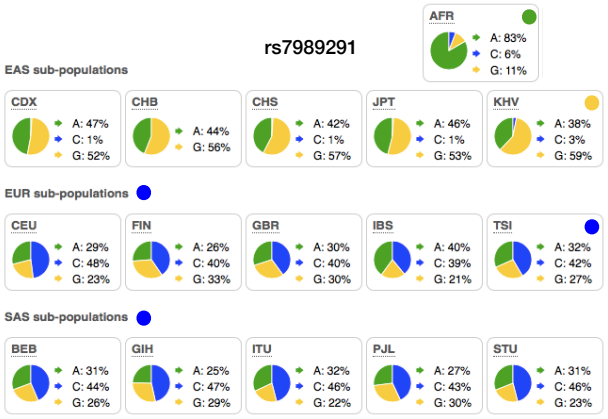
rs2387842 DataSlicer lists all Phase III genotypes

No variation in 1000 Genomes

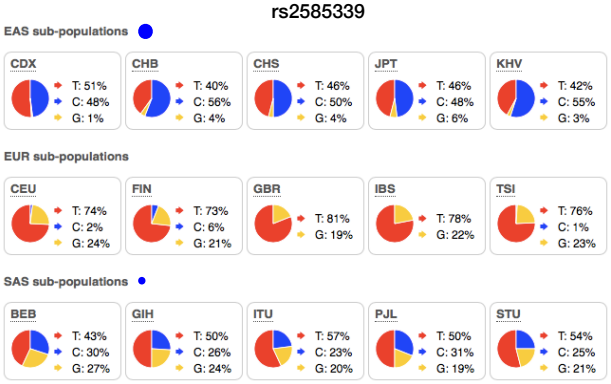
rs2605361 DataSlicer lists all Phase III genotypes

No variation in 1000 Genomes

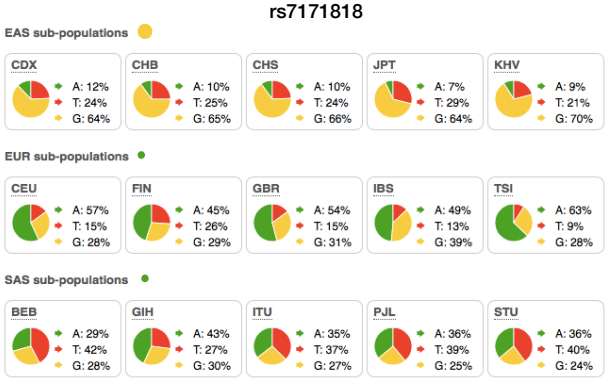
rs7989291



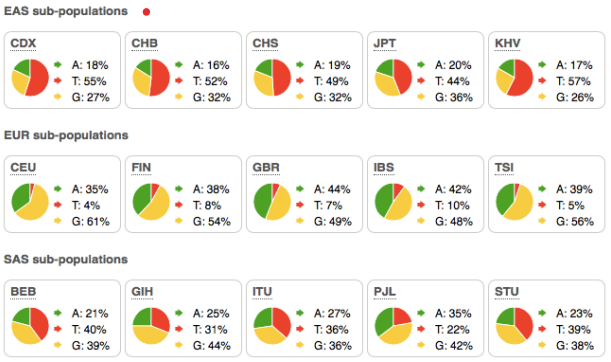
rs2585339



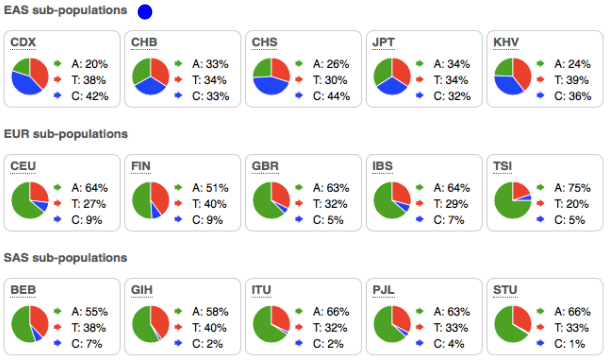
rs7171818



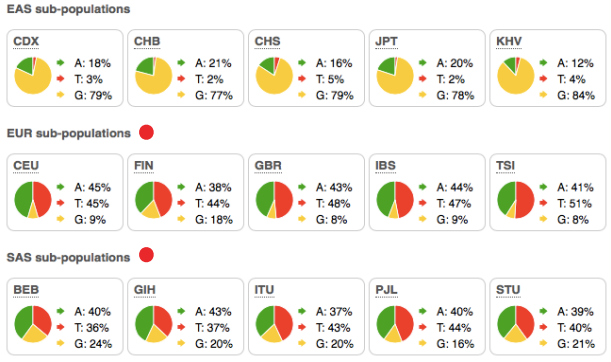
rs408046



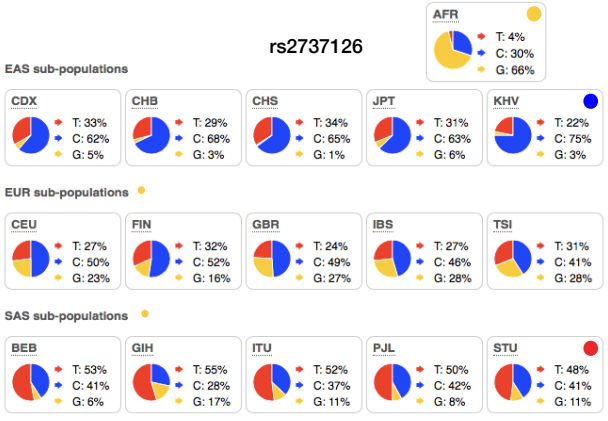
rs1074689



rs556365

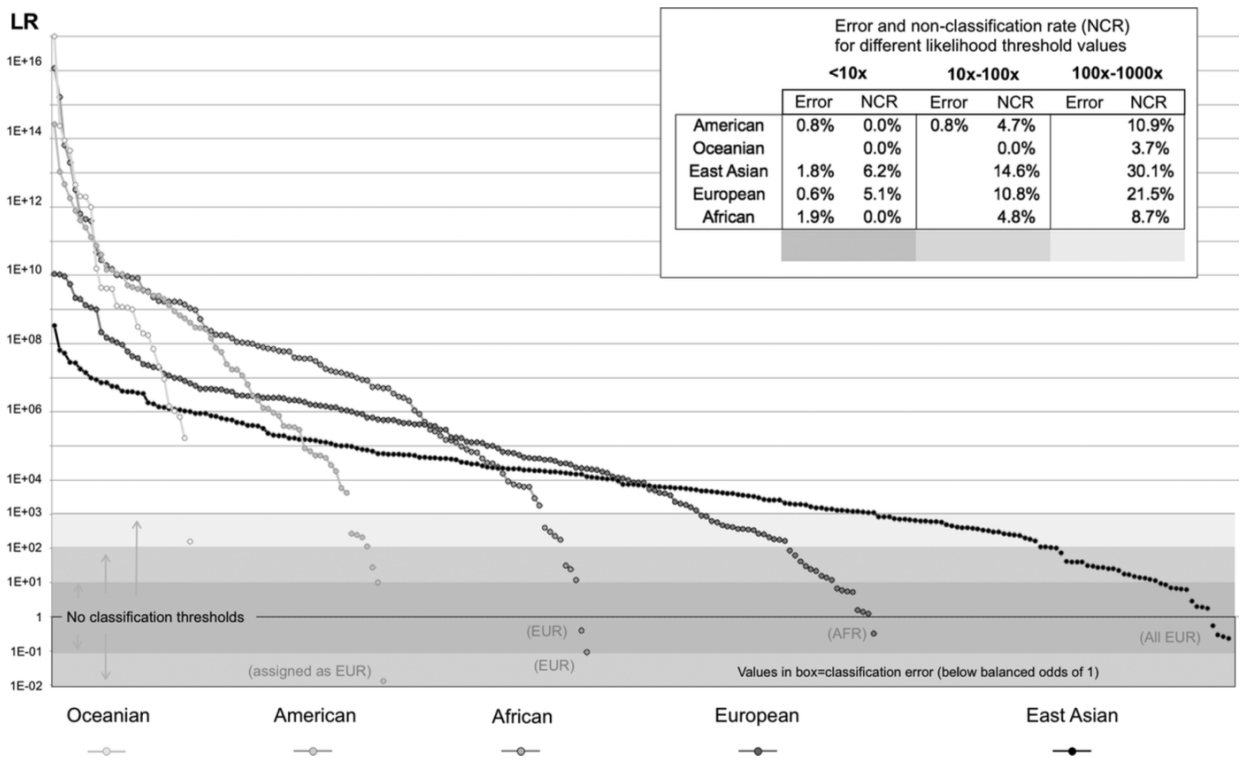


rs2737126




rs6504633
AFR
A: 8%
T: 21%
C: 31%
G: 40%
EAS sub-populations
CDX T: 28% C: 70% G: 2%
CHB T: 36% C: 57% G: 7%
CHS T: 39% C: 58% G: 3%
JPT T: 29% C: 64% G: 7%
KHV T: 37% C: 59% G: 4%
EUR sub-populations
CEU T: 76% C: 18% G: 6%
FIN T: 68% C: 26% G: 6%
GBR T: 80% C: 13% G: 7%
IBS T: 73% C: 15% G: 12%
TSI T: 78% C: 16% G: 6%
SAS sub-populations
BEB T: 42% C: 35% G: 23%
GIH T: 50% C: 32% G: 19%
ITU T: 47% C: 33% G: 21%
PJL A: 0% T: 56% C: 31% G: 13%
STU T: 39% C: 42% G: 19%
rs634212
CDX T: 59% C: 41%
CHB T: 54% C: 45% G: 1%
CHS T: 57% C: 41% G: 2%
JPT T: 49% C: 48% G: 3%
KHV T: 49% C: 49% G: 2%
EUR sub-populations
CEU T: 39% C: 12% G: 49%
FIN T: 25% C: 19% G: 56%
GBR T: 36% C: 14% G: 49%
IBS T: 31% C: 17% G: 52%
TSI T: 42% C: 19% G: 39%
SAS sub-populations
BEB T: 33% C: 26% G: 41%
GIH T: 42% C: 17% G: 41%
ITU T: 42% C: 23% G: 36%
PJL T: 42% C: 20% G: 38%
STU T: 43% C: 25% G: 32%
rs11667741
AFR
A: 7%
T: 73%
C: 20%
EAS sub-populations
CDX A: 39% T: 46% C: 15%
CHB A: 48% T: 40% C: 13%
CHS A: 43% T: 45% C: 12%
JPT A: 51% T: 42% C: 7%
KHV A: 36% T: 46% C: 17%
EUR sub-populations
CEU A: 8% T: 22% C: 70%
FIN A: 7% T: 22% C: 71%
GBR A: 10% T: 26% C: 64%
IBS A: 10% T: 41% C: 49%
TSI A: 8% T: 28% C: 64%
SAS sub-populations
BEB A: 21% T: 30% C: 49%
GIH A: 13% T: 24% C: 63%
ITU A: 14% T: 25% C: 61%
PJL A: 9% T: 32% C: 59%
STU A: 22% T: 31% C: 47%
rs2069945
EAS sub-populations
CDX A: 12% C: 58% G: 30%
CHB A: 4% C: 68% G: 28%
CHS A: 8% C: 59% G: 33%
JPT A: 5% C: 74% G: 21%
KHV A: 15% C: 54% G: 31%
EUR sub-populations
CEU A: 10% C: 48% G: 42%
FIN A: 14% C: 45% G: 41%
GBR A: 5% C: 46% G: 48%
IBS A: 4% C: 55% G: 41%
TSI A: 11% C: 47% G: 43%
SAS sub-populations
BEB A: 16% C: 36% G: 48%
GIH A: 14% C: 34% G: 52%
ITU A: 11% C: 26% G: 63%
PJL A: 13% C: 35% G: 52%
STU A: 14% C: 25% G: 61%
rs914468
S ASN-specific allele
EAS sub-populations
CDX T: 0% C: 89% G: 11%
CHB C: 79% G: 21%
CHS C: 82% G: 18%
JPT C: 79% G: 21%
KHV T: 1% C: 83% G: 16%
EUR sub-populations
CEU C: 52% G: 48%
FIN C: 53% G: 47%
GBR C: 38% G: 62%
IBS C: 47% G: 53%
TSI T: 1% C: 40% G: 59%
SAS sub-populations
BEB T: 23% C: 45% G: 33%
GIH T: 25% C: 40% G: 35%
ITU T: 27% C: 39% G: 34%
PJL T: 22% C: 36% G: 41%
STU T: 26% C: 36% G: 38%
rs393953
Perfect (3-way major allele)
EAS sub-populations
CDX A: 10% T: 51% G: 39%
CHB A: 5% T: 61% G: 34%
CHS A: 8% T: 52% G: 40%
JPT A: 8% T: 34% G: 58%
KHV A: 6% T: 54% G: 40%
EUR sub-populations
CEU A: 27% T: 11% G: 62%
FIN A: 26% T: 8% G: 66%
GBR A: 29% T: 12% G: 59%
IBS A: 36% T: 11% G: 53%
TSI A: 32% T: 19% G: 49%
SAS sub-populations
BEB A: 50% T: 16% G: 34%
GIH A: 42% T: 10% G: 48%
ITU A: 53% T: 10% G: 37%
PJL A: 40% T: 6% G: 54%
STU A: 60% T: 9% G: 31%
rs3859849
AFR
A: 1%
C: 31%
G: 68%
EAS sub-populations
CDX A: 75% C: 6% G: 19%
CHB A: 75% C: 5% G: 20%
CHS A: 78% C: 6% G: 16%
JPT A: 82% C: 7% G: 11%
KHV A: 74% C: 6% G: 20%
EUR sub-populations
CEU A: 17% C: 34% G: 49%
FIN A: 29% C: 27% G: 44%
GBR A: 18% C: 30% G: 52%
IBS A: 21% C: 39% G: 39%
TSI A: 14% C: 40% G: 46%
SAS sub-populations
BEB A: 33% C: 9% G: 58%
GIH A: 20% C: 13% G: 67%
ITU A: 21% C: 14% G: 65%
PJL A: 23% C: 21% G: 56%
STU A: 20% C: 12% G: 68%

STRs can have contrasting allele frequency distributions amongst populations, but generally these are not strongly differentiated, as discussed in the Introduction. Studies by Pereira et al, 2012 [105] indicate that core forensic autosomal STRs can be used to distinguish populations, but the ancestry assignment likelihoods do not reach the values routinely obtained with SNPs. The STR studies in **Phillips, et al, 2014** [106] indicated that when a larger array of STRs are combined (from extended sets used for forensic analysis) the ancestry assignment likelihoods improve to the point that very few samples in the HGDP-CEPH panel are misclassified. **Figure D10** shows the ancestry assignments obtained using an extended set of 32 autosomal STRs to classify 578 HGDP-CEPH samples are almost completely correct, with just eight incorrectly assigned (points below the LR midline of balanced odds=1). When a likelihood ratio threshold of 10 is applied (i.e. discounting all classifications with LRs less than 10); only a single American individual lies below the midline and outside this LR value range, so is misclassified as a European. Interestingly, this same CEPH Native American individual carries an AME-specific D9S1120 9-repeat allele [107]. The study of Phillips demonstrates that extended sets of STRs up to 32 loci provides better data for ancestry inference than the current core set of 24 STRs alone (Pereira's population studies examined an even smaller set of STRs, but in a wide range of populations with very large samples sizes). These studies suggest an initial population analysis of any forensic STR data ahead of specialised ancestry tests may be useful in obtaining an idea of the DNA donor's likely ancestry. The important point is that STR genotypes are often the only DNA data available from a forensic analysis. USC use STRs as an independent test of ancestry for all forensic cases when this data is available, and there is nothing wrong with combining the likelihood values obtained from STRs with those of SNPs. To do this, *Snipper* was adapted to infer the ancestry of individuals with ***allele*** frequency data as the training set; for the simple reason that STR ***genotype*** data was not made publicly available [104] for use with the *Snipper* classifier because this compromises the privacy of donors (i.e. they are not fully protected from DNA database searches if their forensic STR genotypes are known). The development of a frequency-based classifier in *Snipper* allows both custom sets of a user's own data or the STR frequencies held in pop.STR to be used for input as training sets, with each STR or other marker type (such as Microhaplotypes) entered as frequency tables in one worksheet per marker. Likelihoods from the analysis of an uploaded STR profile are generated in an identical way to SNPs and can therefore be used to make an ancestry assignment from the ratio of the two highest values. All ancestry marker data can be entered as frequencies in this way, but in practice, this approach represents a much more clumsy way than uploading an Excel table of SNP profiles. In addition, only single profiles can be analysed and no PCA plots are currently generated in the frequency classifier. Therefore, the 'classification of multiple SNP profiles' option in *Snipper* that can easily make a PCA plot and the accompanying Bayes analyses, is the approach of choice and frequency-based calculations are reserved for ancestry analyses in STR-only forensic cases. The steps taken to analyse a typical forensic STR profile to infer ancestry are shown in **Figure 10,** illustrated with a profile obtained from a missing person identification, were only STR data was available.



	<10x		10x-100x		100x-1000x	
	Error	NCR	Error	NCR	Error	NCR
American	0.8%	0.0%	0.8%	4.7%		10.9%
Oceanian		0.0%		0.0%		3.7%
East Asian	1.8%	6.2%		14.6%		30.1%
European	0.6%	5.1%		10.8%		21.5%
African	1.9%	0.0%		4.8%		8.7%

Figure D10. Ranked log ancestry assignment likelihoods (best to worst, left to right) obtained from genotyping 32 STRs are plotted for HGDP-CEPH individuals from five population groups. Points in the grey boxes show individuals with LRs lower than 10, 100, and 1000 probability thresholds. Points falling below the balanced odds midline of 1 represent classification error (with incorrect assignments in brackets: EUR, European; AFR, African). Note the single American outlier labeled (assigned as EUR) has an LR of 72 shown for European versus American ancestry, which would be treated as a reliable classification applying an LR threshold of 10.

The analysis of haplotype variation such as Microhaplotypes and Y-linked data is ideally suited to a frequency based approach because of the ease of handling multiple alleles in table format and the fact that counting the number of combinations of linked SNP genotypes rather than uploading the SNP genotypes individually adequately accounts for linkage, in a comparable way databases like YHRD does; by ***counts*** rather than combined frequencies reconstructed from each SNP's individual genotype frequencies. Therefore *Snipper* has in-built modules for linked loci that can then be combined with normal single-site markers to create a combined likelihood. Although this has to be done manually at the moment, USC plans to allow set Microhaplotype loci and single site SNPs to be analysed together in a single step - and the same approach should be possible with the combination of linked Y and unlinked autosomal SNP data. The adaptation of *Snipper* to handle Microhaplotype data links individual SNP columns into single permutations marked by haplotype defining numbers in row 2 (e.g. all SNPs in the first haplotype, by genome position, will be marked 1).

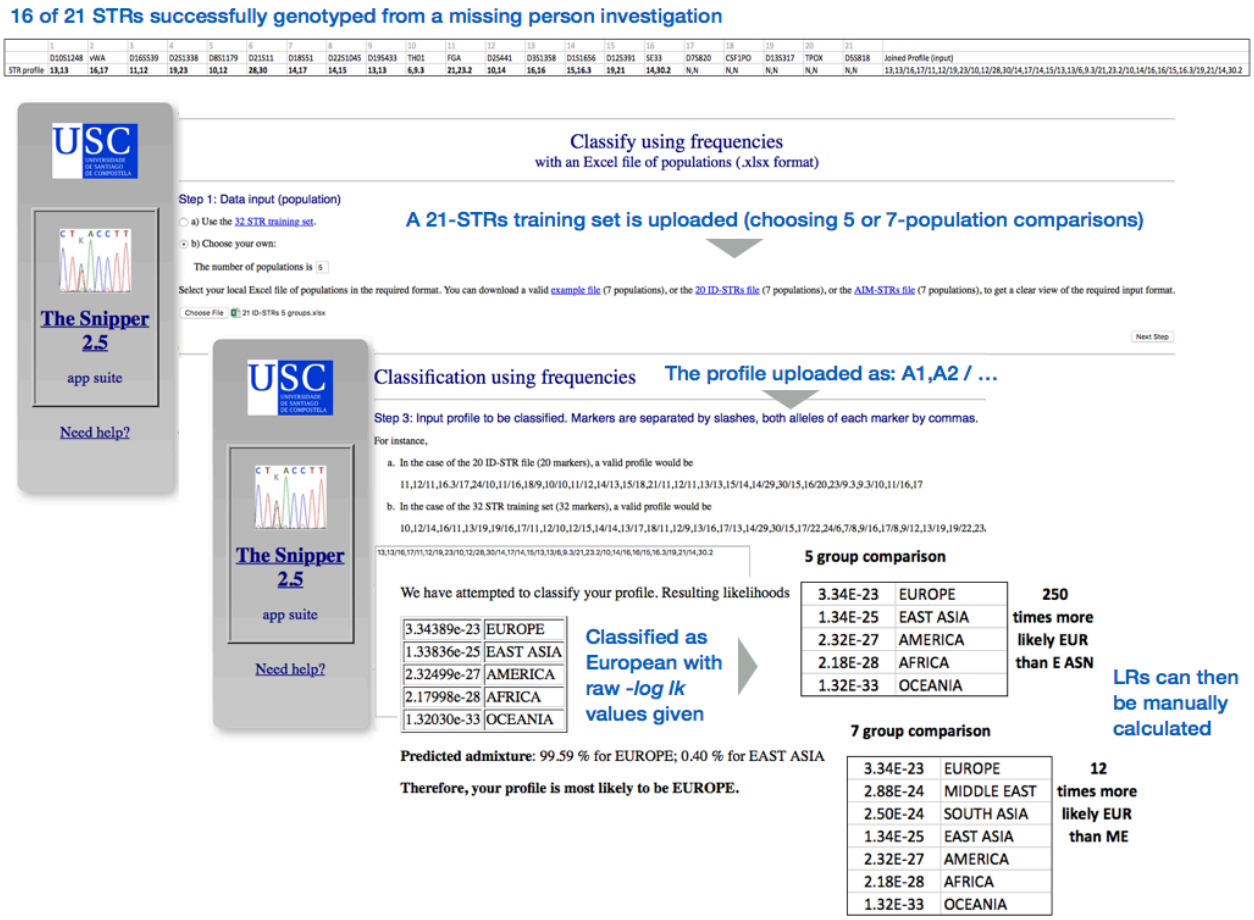


Figure D11. An STR-based ancestry analysis with the Snipper frequency classifier. See also Panel 2 in **Box 6**, that compares alternative STR data analysis regimes in Snipper and Pop Affiliator.

As they differ in size by definition, Indel markers are easily adapted to dye-linked PCR in identical fashion to STRs, so they can provide a very secure way to identify and analyse mixed DNA components, but keep the benefit of avoiding the long amplicons necessary for most STRs. Indel's binary polymorphism characteristics make them less informative per locus for forensic identification [102], but this factor is not so relevant for ancestry analysis where a large proportion of Indels show a population differentiation of allele frequencies quite similar to many AIM-SNPs [50]. In fact, compilation and use of AIM-Indels has been one of the most successful developments in forensic ancestry analysis to come from USC in the last five years and has begun to be adopted on a wider scale because of these marker's balanced peak profile characteristics, assay robustness and ability to detect mixed DNA more easily than use of SNPs alone. The combination of SNPs and Indels helped to provide a highly informative likelihood in the Eva Blanco cold case DNA analyses (detailed in **Box 10**), and they continue to provide an 80-marker first-strike ancestry test at USC. Because of their ease of use and speed of the test's analysis in placing amplified Indel fragments directly into the CE detector, provide

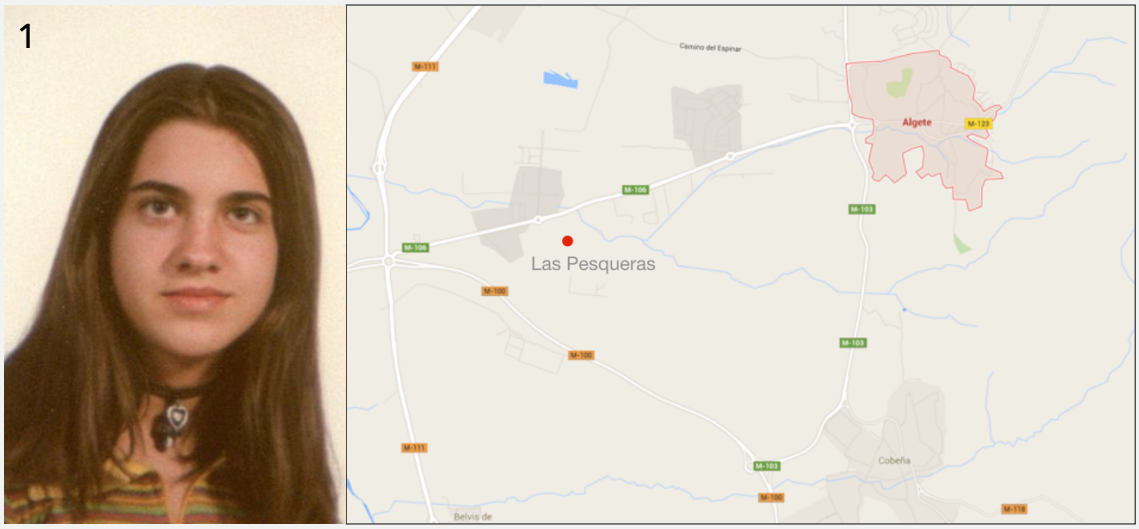
the perfect system for triaging queried ancestry (a very simple resolution of a paternity case without either putative father's DNA tested, is described in **Box 11**). The robustness of the AIM-Indel set for genotyping degraded DNA has been explored in the analysis of skeletal remains encountered in the Argentinian program to identify victims of the dictatorship (**Thesis Paper #11, Romanini, et al, 2012** [173] and **2016** [174]). In general, Indels genotyped with PCR-to-CE perform as well as SNPs when analysing degraded DNA, demonstrated in a similar study of skeletal material from the the same Argentinian ID program, where 35-year old bones were successfully genotyped with both forms of short-amplicon tests, which performed better than Mini-STR analyses of the same material. The 46-plex AIM-Indel set was informative enough to efficiently estimate the ancestry even in samples yielding partial profiles, obtained from the inhibited and degraded DNA extracts of 35-60 years, from different continental origins beyond Argentina. therefore in both applications for the study of very degraded DNA: identification and ancestry inference - Indels show evidence that they perform at least as well as SNPs, while being genotyped from much simpler and robust tests that take full advantage of the balanced signals within any one CE dye.

Parallel enhancements at USC to both *Snipper* and *SPSmart* for forensic Indel genotype data were achieved shortly after the direct PCR-to-CE tests were introduced for these markers. The *SPSmart* database suite now includes dedicated *forInDel* pages for the AIM-Indel genotypes form the HGDP-CEPH panel samples, enabling a range of population comparisons to be made, including the differentiation of European and North African individuals, critical to the identification of a North African ancestry for the suspect in Eva Blanco investigation (see **Box 10**). The *Snipper* fixed training sets for standard CE-based forensic ancestry analysis provides the 34-plex SNPs and 46-plex AIM-Indel sets. Note that the conventional assignment of blue pie chart segments to the reference allele and red segments to the alternative allele used in SNPs is not applied to Indels, as reference alleles can be difficult to identify as an insertion or a deletion of sequence in many cases. This has largely been achieved in 1000 Genomes, but the *forInDel* database retains this convention. Likewise, alleles are defined as short=A and long=C for genotype coding in *Snipper* input files and reference profiles. Third alleles have been identified in three loci: rs140837; rs34122827; rs25584 and these have been characterised by sequence analysis when possible (see: **Santos, Phillips, et al, 2015** [79]). The third alleles are coded as G in *Snipper* genotypes, but only apply to the reference profiles for these three Indels at this moment, i.e. novel third alleles would return an error in other Indels if uploaded as genotypes contained G (see **Thesis Paper**, **#8, C. Phillips, et al, 2016** [158]).

Box 10. The Eva Blanco murder investigation - the successful resolution of a cold case after 18 years

Eva Blanco Puig was a 17 year old schoolgirl from the town of Algete, near Madrid. In April 1997, while walking back alone from an evening with friends in a nearby village, she was raped and murdered. Her body was found with multiple stab wounds in a ditch near the M-106 road in the area known as Las Pesqueras (**Panel 1**); with evidence at the scene suggesting abduction in a car and an attack in this locality. However, heavy overnight rain had obscured shoe marks and tyre prints to such an extent that they were rendered unusable as a means to link an attacker or his vehicle. A small semen stain was found on the girl's undergarments and DNA profiling made. This led to an application by the police investigators to undertake a mass screen of males living in Algete to seek a DNA match from these voluntary samples. However, this amounted to almost 2000 men who had not been eliminated as suspects and there was no strong evidence to suggest that the attacker had come from the same town as Eva. The Judge consequently rejected this request on the grounds of its likely ineffectiveness and high costs. The investigation did not progress further despite following more than a hundred lines of enquiry.

1



In 2015, USC was asked to perform predictive tests for ancestry and physical characteristics on the DNA sample from the case. Ancestry testing comprised **34 AIM-SNPs** of the core SNaPshot ancestry test and **46 AIM-Indels** genotyped using dye-linked primers. A total of 63 pigmentation-related SNPs were genotyped in the 'SHEP' skin-hair-eye colour associated SNaPshot tests. The results are summarised in **panel 2**. The prediction of intermediate skin tone was the weakest likelihood compared to the other tests, but at a reasonable statistical level for this least-successfully inferred trait. Although black hair was more definitively predicted, the strongest likelihood was for brown eye colour compared to intermediate (hazel-light brown). **The ancestry test predictive likelihoods were much more definitive**; providing the strong prediction of over **80,000 times more likely to be North African than European**. These likelihoods were, in part, achievable because of the completion of worldwide AIM-Indel variant frequency studies in Spring of 2015, so the timing of the tests was fortuitous.

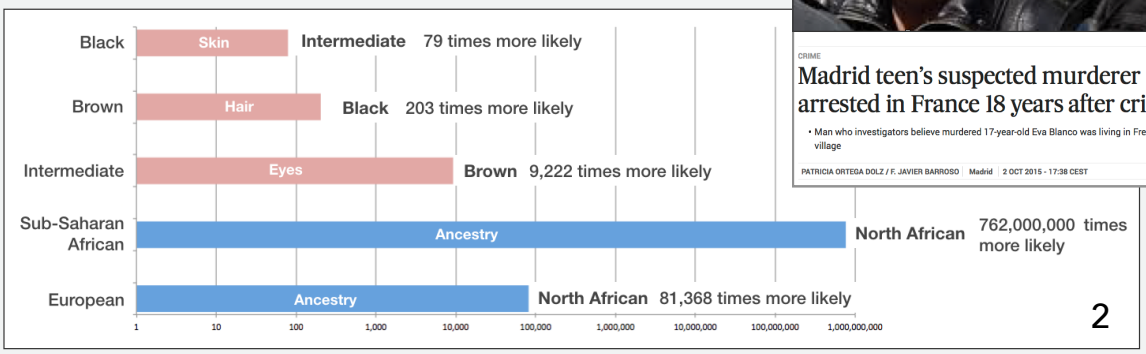
These SNP analysis results prompted the police to apply for a new mass DNA screen of North African males recorded by census as living in the broader environs around Algete between 1995 and 1999. Two Moroccans who had lived in Cobeña and voluntarily gave samples, produced partial DNA matches but differed in the Y-STR genotypes by one **rapidly-mutating Y marker** - indicating they were brothers of the perpetrator. The police consequently identified **Gerj Chelh Ahmed**, who had lived in Les Varans Pierrefontaine, France from 1999 and had not revisited Spain since then. He had worked delivering plants for a nursery on the M-111 road and knew of Eva Blanco.

CRIME

Madrid teen's suspected murderer arrested in France 18 years after crime

• Man who investigators believe murdered 17-year-old Eva Blanco was living in French village

PATRICIA ORTEGA DOLZ / F. JAVIER BARROSO | Madrid | 2 OCT 2015 - 17:38 CEST



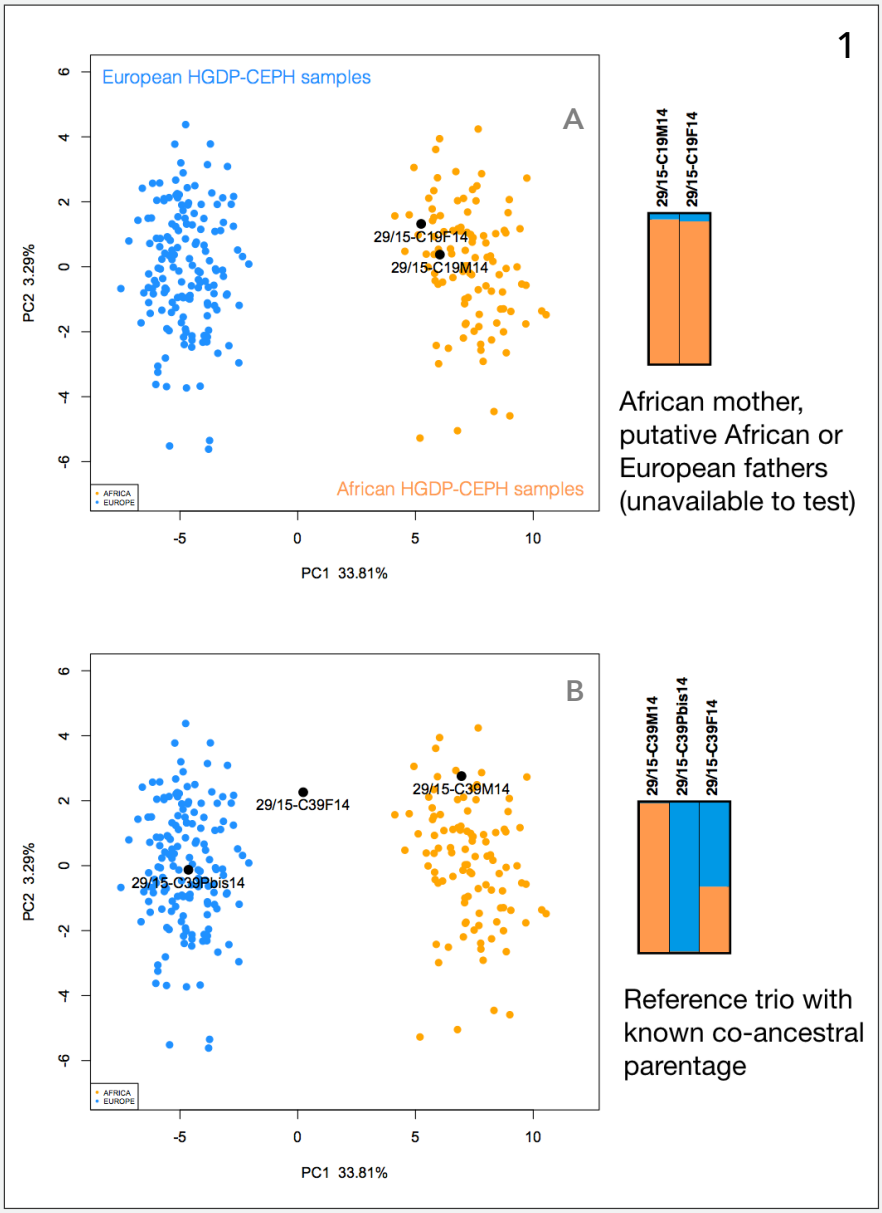
2

Box 11. Applying a complete forensic ancestry solution to an 'unresolvable' paternity dispute

USC was recently asked to assist in a disputed paternity case where each of the putative fathers was unavailable to give DNA for testing - implying *in absentia,* an unresolvable legal case to secure paternal support for the mother and child. The putative fathers were Italian and Zulu South African in origin. In fact, this case was resolved in a simple way by applying a quick and easy ancestry test to the DNA samples from the mother and child, based on the assumption that there were just two possible fathers. While not providing actual proof of paternity, the relatively straightforward resolution of the counterclaim illustrates the value of a complete forensic ancestry analysis system that encompasses a PCR-based genotyping test using fast, well-established technology; a real-time data analysis regime; and population reference data freely available to use with online statistical tools matched to the test or applied independently by the user.

The DNA samples were tested with 46 Indels in one multiplex (the biggest non-MPS forensic multiplex developed so far) using the direct PCR-to-CE system that can be accomplished in under 4 hours; dividing analysis time between PCR and capillary electrophoresis, without the need for clean-up steps (see ***Box 4***). This approach allows the direct profiling of dye-labeled amplification products with a standard automated sequencer. An additional control family trio was tested in tandem, comprising a known European father, an African mother and their child. The Indel CE genotypes obtained were analysed in *Snipper* by making Bayes and PCA analyses using HGDP-CEPH reference population data ported directly from a dedicated SPSmart forensic Indel genotype page called ***forIndel***. As Bayes analysis is inefficient at analysing admixture patterns compared to genetic cluster algorithms, STRUCTURE runs were made to compare to the PCA positions observed in the five individuals. Results for the *Snipper* PCA analyses and STRUCTURE cluster plots are shown in **Panel 1**. In comparison to the mid-cluster position of the child in the test trio (29/15-C39F14, in PCA plot B) and the evenly apportioned joint cluster membership in the lower STRUCTURE cluster plot (equivalent blue and orange portions); the child of the paternity test ((29/15-C39F14, PCA plot B) is positioned within the African reference population cluster alongside that of the mother. The STRUCTURE cluster plot patterns match these patterns by showing almost complete African cluster membership in both samples. This data definitively excluded the Italian putative father and in combination with supporting documentation directed the court towards securing a sample from the man who had since moved back to South Africa.

Ancestry analysis in this way can also be applied to the vexed question of obtaining knowledge of the likely ethnic background of babies needing adoption with suitable culturally-matched foster parents, when this is unknown to the authorities. The advantage of forensic sensitivity in ancestry tests is also relevant to the tracking of illegally trafficked donor organs from third-world countries.



Panel 1: PCA and STRUCTURE analysis of mother and child samples (Plot A) from the disputed paternity case. Reference samples are HGDP-CEPH Europeans (blue) and Africans (orange). Plot B shows a reference trio with known European / African parents revealing the expected PCA and cluster plot patterns for the child (29/15-C39F14).

Further indications of the robustness of Indels were provided by a large-scale collaborative exercise between EDNAP and EUROFORGEN laboratories looking at the now standard combination of 34 SNPs and 46 Indels to infer ancestry of forensic samples. The EDNAP-EUROFORGEN exercise results reported in **Thesis Paper #5, C. Santos, et al, 2015** [175]. Results from laboratories, many of which were using these tests for the first time, indicated consistent genotyping performance from both tests, reaching a particularly high level of reliability for the Indel test. SNP genotyping gave 93.5% concordance (compared to USC's data) that rose to 97.3% excluding one laboratory with a large number of miscalled genotypes. Indel genotyping gave a higher concordance rate of 99.8% and a reduced no-call rate compared to SNP analysis. All participants detected an unmarked 3:1 mixture amongst the exercise DNAs, using Indel peak height data and successfully assigned the correct ancestry to the other samples using *Snipper*. Therefore, successful ancestry assignments were achieved by participants in 92 of 95 *Snipper* analyses. This exercise demonstrates that ancestry inference tests based on binary marker sets can be readily adopted by laboratories that already have well-established CE regimes in place. The Indel test proved to be easy to use and allowed all exercise participants to detect the DNA mixture, as well as achieving complete and concordant profiles in nearly all cases. Two participants successfully ran parallel MPS analyses (each using different sequencing platforms) and achieved high levels of genotyping concordance using unmodified PCR primer mixes from the exercise; which indicates Indels (even when using dye-labelled primers) can be adapted in a straightforward way for MPS analysis.

8. To rebuild smaller AIM sets into enlarged single 130 to 160-plex panels applicable to compact massively parallel sequencing (MPS) platforms, increasingly being adopted for forensic DNA analysis. To validate the resulting SNP panels in terms of sequence balance, sensitivity and genotyping concordance.

The rebuilding of existing forensic AIMs panels developed at USC into new sets for MPS analysis took advantage of the expanded capture-PCR multiplex capacity of the technology by aiming initially, to combine 128 AIM-SNPs (described in detail in **Thesis Paper #1, C. Phillips, et al, 2014** [69]). In fact, this level of multiplexing is increasing as MPS technology matures. The introduction of automated library preparation and chip loading systems (Ion Chef) and the increased level of sequence coverage achieved with the Ion S5 sequencer have meant extra SNPs can now be incorporated into the panel without a consequent loss of sequence read levels. This evident capacity for increased coverage prompted the development of a revised AIMs panel expanding component loci up to 165 markers (143 single site SNPs, 22 Microhaplotypes). To a large extent, aiming for a particular minimum multiplexing level in a forensic MPS tests involves a dialogue with the technology's developers, as the SNP multiplexes used in medical MPS applications are larger, but do not have the need to analyse degraded DNA, where many SNPs may suffer from reduced sequence read levels that may compromise the quality of genotyping data obtained. If this disproportionally effects markers

informative for one population, the data potentially becomes biased and must be adjusted to restore the balance of population differentiation in the data used to compile genotypes.

The development of both the 128-plex and 165-plex panels has been dependent on the primer design procedures of Thermo Fisher, which in turn dictate the assay conversion rate - the extent to which a set of candidate SNPs can be successfully incorporated into MPS systems, where an additional layer of complexity compared to CE-based analyses is the need to accomplish reliable sequence alignments, when a SNP site may not be reliably aligned to the flanking sequence due to homopolymeric nucleotide tracts (alternatively termed poly-base) or common Indels producing a potential frame-shift effect. Thermo Fisher primer design pipelines assess what is described as "misalignment risk". When there are homopolymeric stretches of two or more nucleotides on one or both sides of the SNP site there is a potential risk for a sequence alignment mapping error to occur. It can range from no effect to a false heterozygote call of some frequency. The base composition of flanking sequence is what creates the error type. The most difficult situation is something like an AAA(A/T)TTT location where any "slipping" or under-calling/compression of one side of the homopolymeric tract can lead to a misaligned A or T at the SNP location. An AAA(C/G)TTT situation will not lead to a false heterozygote call since SNP allele call is restricted to a C or G and any misaligned sequence reads will be discarded as noise and lower the coverage of that SNP. The longer any homopolymeric stretch is, the higher the risk of misalignment caused by under-calling of nucleotides in a homopolymeric run and if those nucleotides match one or both of the SNP alleles, the error rate for a potential false heterozygote call increases.

The assay conversion rate, taking these factors into account and the likelihood of a SNP candidate being sited in non-unique (multiply replicated) sequence, for 128 SNP candidates was 97.6% as three failed primer quality checks. In fact, all three were sited in the human genome in places that have high levels of low complexity sequence where repeat region sequence would create non-specific primer binding and significant amounts of off-target sequence reads. Additionally, were too close to long homopolymeric tracts preventing efficient sequence alignment as described. For the revised and enlarged AIM set of 165 AIM loci, 83 SNPs were taken from the original 128-plex and one more added; 76 of 81 other candidate loci (including several Microhaplotype sets) were successfully incorporated producing an assay conversion rate of 97.6%. Therefore, a consistent rate of successful incorporation of candidate SNPs into the capture PCR step of MPS appears to be higher than 97% which is above the assay conversion rate achieved with previous SNP genotyping systems used (**Phillips, 2007** [114] and **Phillips et al, 2007** [155]).

The forensic evaluation of two Thermo Fisher SNP sets; consisting of the above custom 128-plex ancestry panel and a prototype of the commercial 169-plex of identification SNPs; examined sequence coverage, genotyping precision, sensitivity and mixed DNA patterns. Evaluations were made amongst three laboratories following closely matched Ion PGM™ protocols and adopted a simple validation

framework based on shared DNA controls. These two studies have provided a comprehensive assessment of the forensic performance of MPS and its consistency between the collaborating laboratories. A third evaluation of forensic MPS has examined a "second party" multiplex of 130 identification SNPs developed by Qiagen but applicable to either the Thermo Fisher Ion PGM/S5 platforms or the Illumina ForenSeq platform, as examined in the study of **de Puente et al, 2017** [176]. This study was conducted in USC alone, but obtained similar findings to an independent study of the same SNP set by Grandell et al, in 2016, which used the Illumina system [177].

In the evaluation of the Thermo Fisher 169-plex ID-SNP panel described in **Thesis Paper #12, M. Eduardoff, et al, 2015** [120], the sequence coverage obtained was extensive for the bulk of SNPs, while sensitivity studies showed 90-95% of SNP genotypes could be obtained from 25 to 100 pg of input DNA. Therefore, the first evaluations of MPS made by USC and partner laboratories have confirmed the forensic sensitivity of MPS, and this is is further underlined by the successful analysis of 750-year old DNA described above in objective 5 and outlined in **Figure D5**. In terms of genotyping precision, the concordance rates of Coriell cell-line control DNA genotypes checked against whole-genome sequencing data from 1000 Genomes and Complete Genomics indicated a very high precision of 99.8%. Discordant genotypes were detected in five SNPs, indicating that the lower allele calling reliability of these loci means they should be excluded. Therefore, the SNP panel and the Ion PGM MPS system provide a sensitive and accurate forensic SNP genotyping assay for normal DNA. However, when low-level DNA is analysed, much more varied sequence coverage was found, to the extent that a larger number of component SNPs would need careful checks of sequence patterns. Furthermore, assessments of mixed DNA indicate the user's control of sequence analysis parameter settings is necessary to ensure mixtures are detected robustly. Given the sensitivity of MPS to detect a large number of sequences for each SNP, this aspect of forensic genotyping requires further optimisation before massively parallel sequencing is applied to routine casework, where mixed DNA is commonly encountered.

In the evaluation of the custom 128-plex of Global AIM-SNPs (**Thesis Paper #15, M. Eduardoff, et al, 2016** [160]) the study used a similar simple framework which assessed individual SNP genotyping precision using lab-wide controls, forensic sensitivity of the multiplex using dilution series, degraded DNA plus simple mixtures. Additionally, a series of population studies gauged the ancestry differentiation power of the final panel design, which required substitution of three original ancestry-informative SNPs with alternatives. Fourteen populations that had not been previously analysed were genotyped using the custom multiplex and these studies allowed further assessment of genotyping performance by comparing sequencing data across five laboratories. The revised 128-plex panel gave a low level of genotyping error although in the case of several SNPs errors were observed from sequence misalignment caused by homopolymeric tracts close to the target SNP, despite careful scrutiny of candidate SNPs at the design stage. Such sequence misalignment required the exclusion of rs2080161 (**Figure D12**) from the final panel, which now comprises 127 SNPs. However, the overall

genotyping precision and sensitivity of this custom multiplex indicates the Ion PGM™ assay for the Global AIM-SNPs is highly suitable for forensic ancestry analysis with MPS. These 127 SNPs have retained their primer designs during the reduction of the panel to a smaller set of 84 SNPs, retaining the population differentiation balance but in a smaller package of SNPs to allow space in the multiplex for binary SNPs specifically able to differentiate South Asians from other population groups (i.e. a slightly expanded and adjusted *Eurasiaplex* AIM set), plus tri-allelic SNPs and Microhaplotype loci able to refine the analysis of East Asian populations. To underline the care with which the original 128 Global AIMs were selected, only one new SNP was added to help re-adjust the balance of the set; the East Asian-informative rs6500380 (**Fig. D12**).

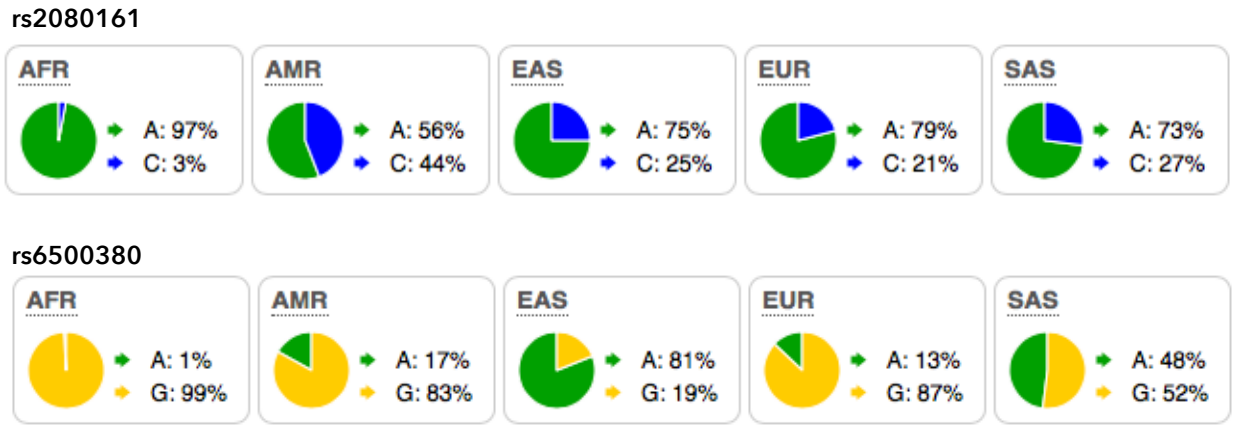
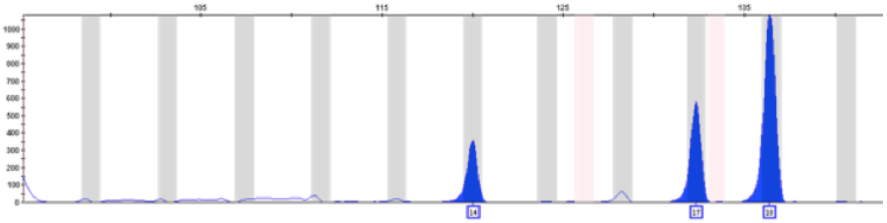


Figure D12. SNPs rs2080161 and rs6500380; the ancestry-informative markers excluded and newly added, respectively, to the Global AIMs MPS panel. SNP rs2080161 is relatively uninformative, while rs6500380 improves the differentiation of East Asian populations markedly.

9. To identify and catalog the collateral variants associated with the core STR and SNP markers genotyped by forensic MPS assays. To analyse their ancestry-informative potential starting with the HGDP-CEPH panel genotyped by the Illumina MiSeq-based Forenseq system.

MPS analysis of the established forensic markers of STRs, SNPs and mtDNA provides substantial detail about the flanking sequences around these markers (or, in the case of mtDNA, the ratio of divergent sequences that occurs in heteroplasmy). Although, it is not guaranteed, a proportion of forensic markers will show these flanking sequence variants (so-called SNP-STRs - see **Box 12**) that can add extra variation allowing differentiation of iso-metric (i.e. identically sized) STR alleles. Many STRs will also show significant sequence variation within the repeat region, as differences in repeat-unit motifs or the presence of Indels creates extra variation, detected by MPS. Collectively, this extra variation produces vastly increased potential to de-convolute mixed DNA and establish the likely contributors; as isometric alleles shared by contributors are

indistinguishable in CE, but can be separated when sequence variants are included in the analysis of data from the MPS tests; as explained in **Figure D13**. The sequence output from both Illumina and Thermo Fisher is limited to the repeat region sequences and, where a SNP site is very close, or flanking sequence is relevant to the STR genotype based on repeat counts, extra sequence is reported. For example, the core STR D13S317 has two A/T SNPs immediately next to the last repeat region nucleotide on the 3' side, that make additional uncounted repeat motifs, so these are shown in the sequence output. The SNP rs9546005 has a high level of polymorphism (see **Figure D14**), so adds extra scope for the differentiation of iso-metric repeat alleles. Furthermore, a 4-NT deletion, also 3' to the repeat region (not shown in **Figure D14**, but 21-NT downstream), can be present within the primer positions for this STR and cause discordancy between an allele's CE-based size estimate and the true count of the repeat units detected by MPS. For this reason, the sequence up to and beyond this additional variant is also reported in MPS and comprises the longest extended sequence of all MPS-based forensic STR genotyping. There is one additional SNP so close to the repeat region that it is also reported (purple dot SNPs in **Figure D14**; VWA, rs75219269). Several other common-variation SNPs close to core STR repeat regions are shown in **Figure D14** (green dots), but these are not reported and represent a wasted opportunity to exploit further variation in the sequenced DNA fragments analysed in MPS. Additional SNPs occur outside of the amplified DNA fragments and may represent variants affecting primer binding sites in established CE kits that can produce allele drop-out. Therefore, a proper gauging of population variation in these SNP sites is important both for predicting their value in all populations to extend discrimination of iso-metric alleles, and to provide additional indications of population of origin. The pie charts of **Figure D14** indicate comparable allele frequencies between the 1000 Genomes population groups from which the data was obtained. Therefore, there is little scope for inference of ancestry from flanking region variants - although some contrast exists between East and South Asia and other populations - but it is marginal. Additional population-specific SNP variants have been reported that provide a chance to differentiate individuals further, but these are akin to private SNPs and are found at low frequency and in unusual populations so would require detailed population analyses by labs adopting this technology for STR analysis. Nevertheless, the clear aim of the forensic MPS STR sequence nomenclature working group (W. Parson-K.B. Gettings-C. Van Neste-J.L. King-C. Phillips, see **Parson, et al, 2016** [130]) is currently to instruct the MPS companies to report the whole sequence - so much scope exists for further exploration of flanking region variants and their patterns of variation. The same will apply to forensic SNPs, which are beginning to show flanking SNPs in linkage with some potential for the reporting of Microhaplotype variants with the target SNP and closely sited loci expanding the variation.



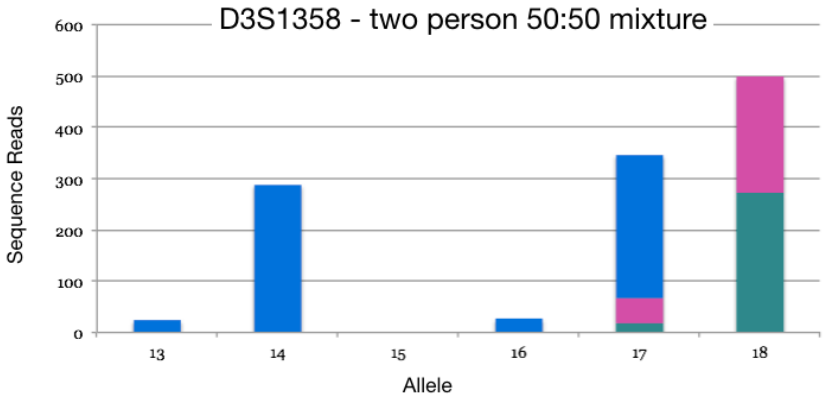


Figure D13. Example of a simple 2-person mixture analysed by CE and MPS genotyping of D3S1358. The CE profile shows a combination of three peaks that could correspond to several possible combinations. The MPS sequence analysis shows the 18-repeat allele has two components with different sequences in equal proportions. Note the stutter signal differentiation in the 17-repeat sequence components.

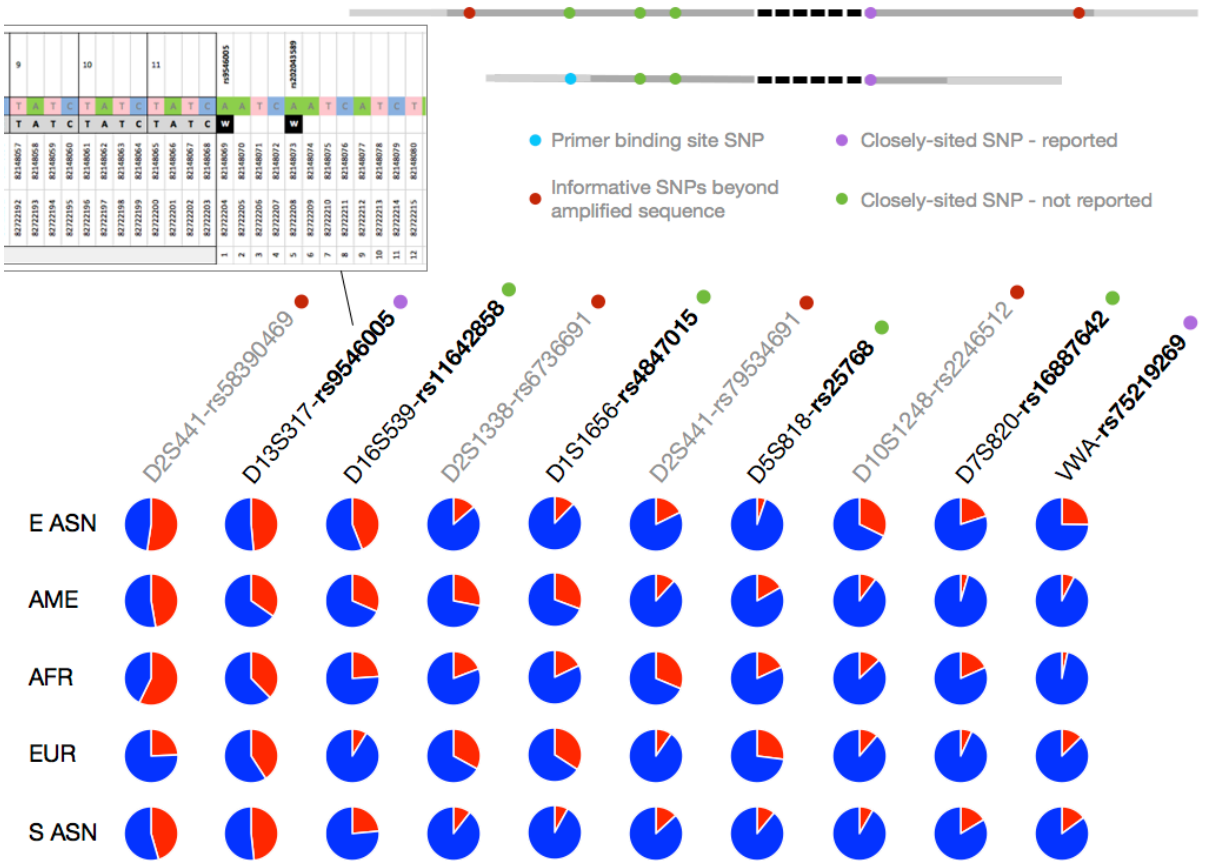


Figure D14. STR flanking region SNPs (SNP-STRs) of different types and their pie chart allele frequency summaries for the most informative SNPs found close to the core forensic STRs. Those with red dots are not detected in the short fragments used by MPS, those with purple dots represent those reported along with the repeat region, including D13S317-rs9546005 a SNP that can create a false repeat unit on the 3′ side of the repeat region.

Although forensic SNP and STR flanking region variation appears limited, the level of sequence variation in the STR repeat regions of certain core loci is considerable - so MPS will expand the detectable levels of STR polymorphism markedly. Repeat region sequence variation can take the form of simple nucleotide changes to a main 4-NT motif which can create a SNP or set of SNPs where different motifs overlap. Additional Indels and more complex changes to motifs can also occur. The STR D12S391 is already a very informative marker when size-based alleles are genotyped, but this STR reveals a much higher level of variation when the repeats are sequenced and the run of AGAT repeat units can be distinguished from the 3′ following run of AGAC units, then the presence or absence of a single AGAT unit in the last repeat position (effectively a terminal C/T SNP). Therefore this STR is shown to be two STRs in succession and the junction between the two motif runs of differing lengths generates much of the extra variability in each iso-metric allele. USC has sequenced 27 A-STRs, 24 Y-STRs and 7 X-STRs of the Illumina ForenSeq marker panel and HGDP-CEPH panel samples. **Fig. D15** maps the sequences in the repeat region of D12S391 and shows that when 3-, 2-, 1-NT deletions are recorded, this STR has 91 sequence-based alleles (colour coded yellow-blue-green and labelled A up to J), that underlie ~20 size-based alleles.

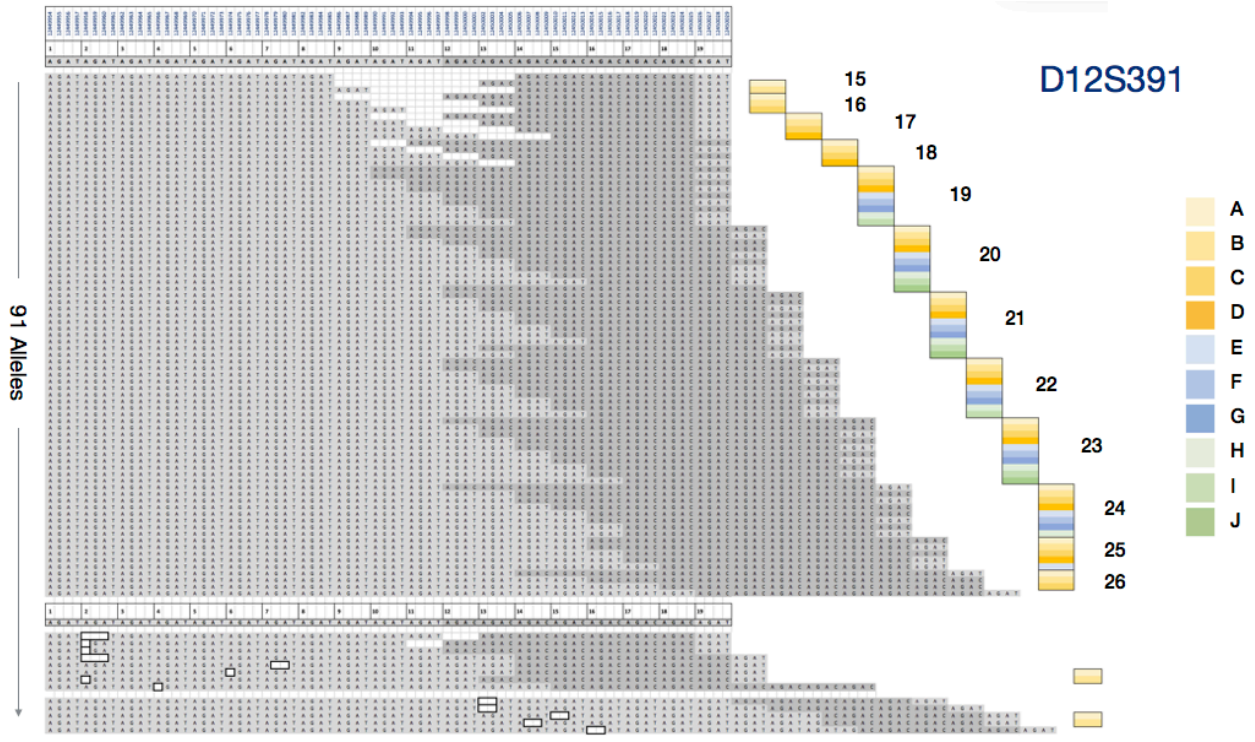


Figure D15. Total sequence-based alleles recorded in the repeat region of D12S391 amongst the 944 samples of the HGDP-CEPH panel. Sequences have been ordered by size then 'alphabetically' (i.e. AGAT AGAT AGAC AGAC AGAC before AGAT AGAT AGAT AGAC AGAC) and given nominal A-J labels and colour codes within each iso-metric CE size group listed on the right 15-16. Intermediate alleles of X.1, X.2 and X.3 are shown on the lower part and it is noteworthy these do not show extensive variation in [AGAT][AGAC] ratios.

Figure D16 shows that extensive sequence variation is also observed in D21S11, D13S433, D8S1179, D2S1338 and D3S1358 and in these six STRs there are sizeable jumps in informativeness brought by characterising the sequence variation that can differentiate iso-metric repeats. In fact, only D10S1248 and TPOX show no sequence variation at all, which indicates most STRs gain benefit from being genotyped with MPS, particularly since they are well established and their genotypes are already populating national databases.

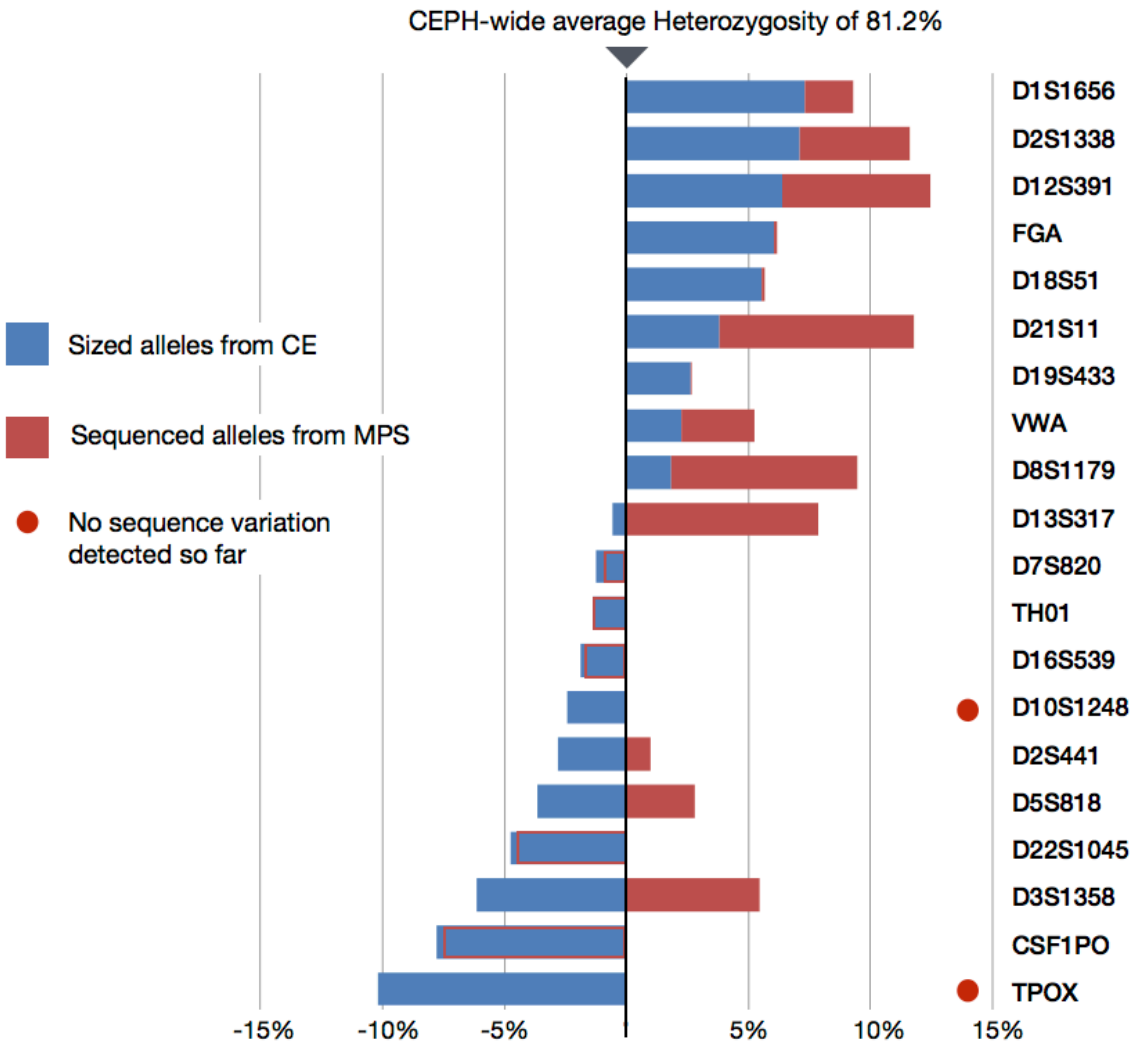


Figure D16. Total STR variability recorded in the 20 core CODIS loci. STRs are ordered top to bottom by CE-Based Heterozygosity from sized alleles (blue bars), and shown relative to an average value across all markers of ~81%, for example, D1S1656 is the most polymorphic STR in CE with a Heterozygosity of 86.5%. Red bars denote the additional Heterozygosity gained from sequence analysis, revealing D21S11, D8S1179, D13S317, D12S391 and D3S1358 make strong jumps in their total variation and therefore forensic informativeness when genotyped by MPS. Open red bars in five STRs describe modest positive gains in Heterozygosity from the left side of the average mid-line value. Values were measured from HGDP-CEPH panel-wide data.

Part of the ongoing studies at USC of STRs genotyped with MPS is to establish a catalog of repeat region sequence variation and match these variants to the population of origin of the samples. So far, the HGDP-CEPH panel genotyping has indicated little association of specific sequence variation to populations for several reasons. First, the sample sizes of most populations in the CEPH panel are too small to draw firm conclusions about population specificity. If a particular nucleotide substitution occurs in a large sample at low frequency it could occur in other populations that have smaller sample sizes but escape detection. Since many sequence variants have low frequencies this is indeed the case and at this moment all variation consists of a CEPH-wide compilation of sequence differences for each STR of the ForenSeq panel. **Figure D17** shows the largest number of sequence-based alleles found in a forensic STR; that of DXS10135. The frequency bars on the right-hand side indicate that ~13-14 common alleles predominate in DXS10135, but a much larger proportion are observations on single chromosomes or comprise 2-5 individuals only. These rare sequence variants are dispersed across all populations and are not indicative of any one origin. Second, it has already been observed that few size-based alleles are population-specific *and* present in high frequency (see **Box 6**; the two allele-types that are specific, D21S2055-X.1 in Europeans and D9S1120-9 in Native Americans, were not studied with MPS). Lastly, it is more likely that in the majority of STRs sequence variants become established early in the evolution of microsatellite loci in human populations and are therefore present in all populations. The dual-repeat motif structure of the STR D12S391 shown in **Figure D16** is present in all populations and both runs of AGAT and AGAC have undergone oscillating repeat diminution / addition events in both runs to maintain the variability in repeat numbers of the polymorphism. The dispersed Indel variants recorded in the lower sequences appear to be random events that add increased levels of sequence variation but are linked with particular length alleles (associated with European population variability) so represent deleted sequence positions imbedded in a given repeat region length with some limited diminution / addition occurring after the deletion was fixed. The pattern of an Indel position in a repeat region being associated with particular length alleles rather than nucleotide variants is seen to some extent in other STRs and suggests that Indels in repeat regions will show more population-specificity than nucleotide variation.

Two other initiatives related to STR sequence variation have also made limited progress so far. The first compiles the microsatellite catalog from 1000 Genomes of 670,646 loci identified in the human genome and published in October 2014. This lists all the identified alleles in each microsatellite in lobSTR format (in a tetra-nucleotide STR this is the repeat number, say 14, or intermediate alleles as decimals: 14.25, 14.5 and 14.75 for 14.1, 14.2, 14.3). This catalog potentially provides an extremely useful additional data resource for forensic use that, based on whole-genome sequences, could contain sequence variation and flanking variants, as well as the length alleles in any one STR. However, there are problems of alignment in complex and longer-than-average loci that preclude a large number of forensic core STRs from the catalog, and genotype data appears to be out of Hardy Weinberg equilibrium in most of the simple STRs assessed so far - specifically, there is a noticeable excess of homozygotes that suggest the alignments of the sequences are not successfully made and

bias the data towards easier-to-align homozygotes and against heterozygotes that have disproportionately contrasting lengths of alleles on each strand. The example of simple STR D9S1122 (in the ForenSeq MPS panel), which has a very large excess of homozygotes in the genotypes listed by 1000 Genomes, is shown in **Box 12**. For these reasons, the whole 1000 Genomes microsatellite catalog has been left to await more detailed annotation.

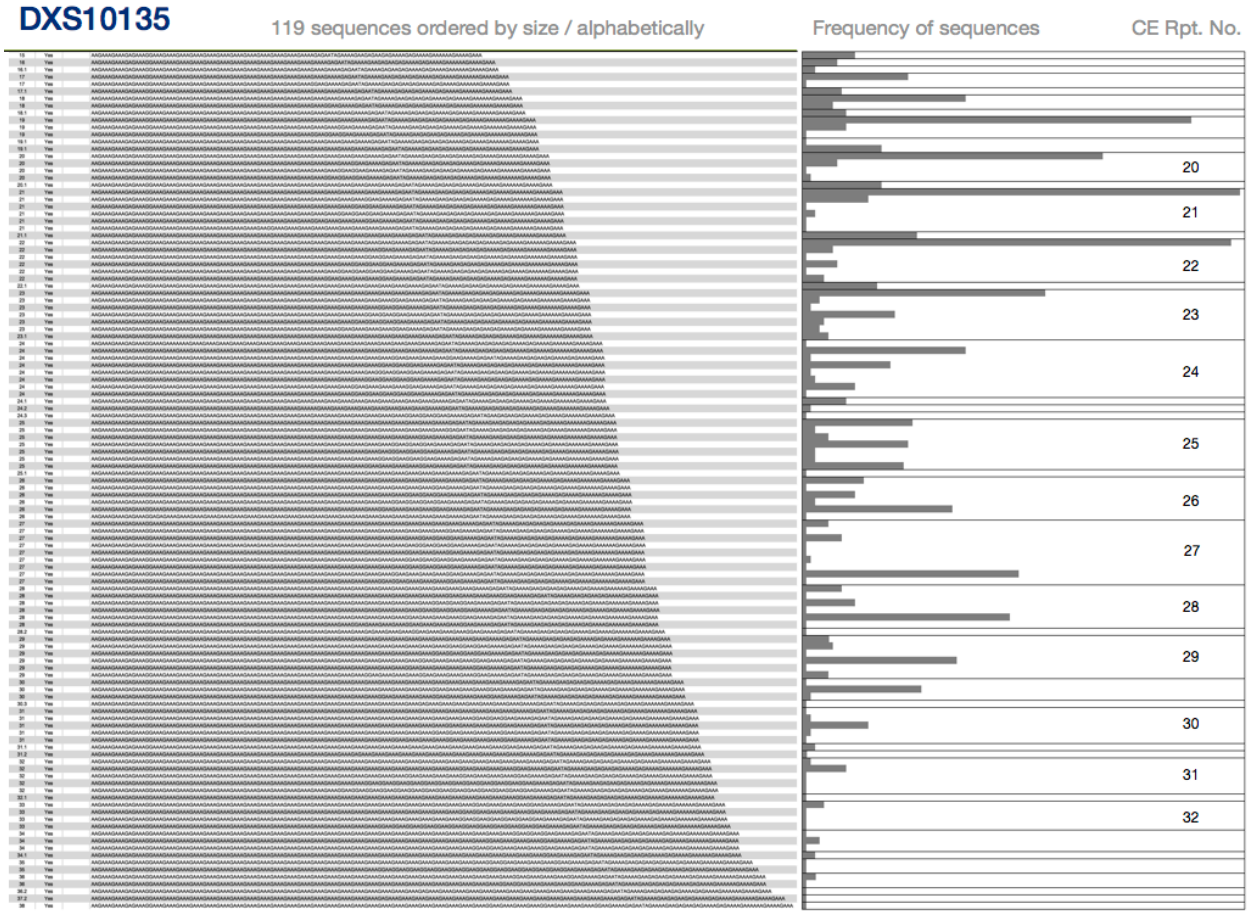


Figure D17. The forensic STR DXS10135 shows the most extensive range of sequence variation of any mainstream micro-satellite studied to date, although a large proportion of the sequence variants are present in small numbers - frequency bars right indicate that about 13-14 common alleles predominate, but a much larger proportion are recorded as singleton observations (or in numbers of 2-5) of a particular sequence within the repeat region of DXS10135. The total number of X chromosomes characterised from the HGDP-CEPH panel was 1262. Such a large degree of variation which is mainly present at low frequencies in a geographically broad population panel suggests a significant amount of population stratification for sequence variation in forensic STRs.

Lastly, the analysis of potential associations which might occur between a SNP-STR allele and a particular repeat in the STR it is sited close to, has yet to be fully explored. This would inform the use of SNP-STRs to independently add extra variation to distinguish iso-metric repeat alleles if their was any association and the two could not be treated as independent loci. Despite very close proximity and the consequent lack of recombination that would disrupt such associations by close linkage, it is unlikely that any particular association would be easily maintained given the instability of STR alleles to form +1 or -1 repeat alleles at a high mutation rate, eroding any associations to a single repeat. Another problem is the high numbers of samples required to cover potential associations of a SNP allele with the full range of repeat alleles seen in many STRs. It should be possible to match SNP-STR allele frequencies in 1000 Genomes with the microsatellite alleles recorded for many of the forensic STRs in sufficiently large sample sizes, but as outlined above and in **Box 12**, the microsatellite data lacks reliability and cannot yet be applied to test pairs of SNP and STR alleles for associations.

10. To build sets of Microhaplotype loci, comprising sets of closely linked SNP alleles in haplotype combinations, that show strongly population-differentiated patterns of variation. To develop forensic MPS assays that comprise panels of Microhaplotype loci informative for ancestry or applicable to identification (e.g. missing person identification) from amplification of Microhaplotypes in very short fragment lengths.

The adoption of 22 Microhaplotype loci in the expanded 165-marker MPS ancestry set, anticipates the more widespread adoption of Microhaplotypes in forensic SNP analysis, as the phase of component SNPs (the allelic combinations on any given sequence strand) can be discerned from knowing the sequence and therefore the position of alleles on any one strand. Once the phase is known more variation can be detected as a simple combination of two heterozygous SNP genotypes: AG and CT, can be a combination of strands A-T with G-C or A-C with G-T; potentially doubling the amount of detected variation in those two genotypes. One key property of Microhaplotypes that has been balanced against their increased informativeness has been efforts at USC to keep the fragment length as short as possible, in order to maintain sensitivity to the analysis of degraded DNA.

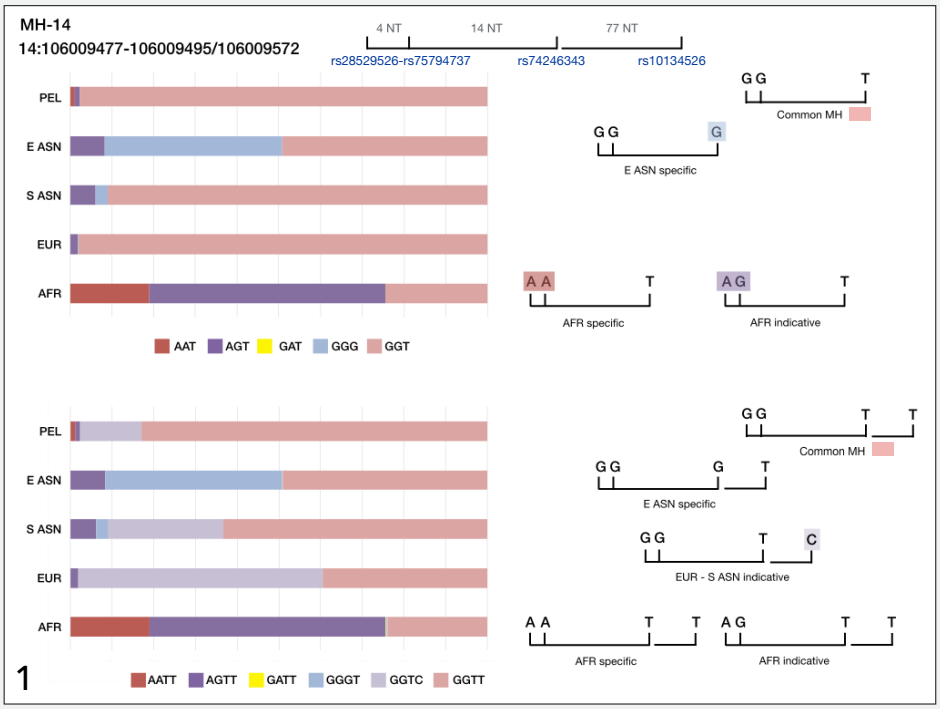
Once fragments were successfully shortened for a proportion of candidate loci, 46 Microhaplotypes were adopted in a massive-scale MPS identification multiplex alongside more than 1,400 tri-allelic SNPs. In many cases amplified fragments are shorter than the Microhaplotype bounds originally suggested to be informative sets of SNPs - with minor loss of power for ancestry or identification purposes. The key features of Microhaplotypes and the variation they can show are summarised in **Box 13.** In some cases when Microhaplotypes cannot be shortened, they have been excluded as fragments up to 200 nucleotides in length will often fail to amplify efficiently from very degraded DNA, despite the raised sensitivity of MPS, and this effect of reduced efficiency is more marked in very large

multiplexes. In other cases, combinations of two SNPs produces just three common haplotypes and the level of informativeness is more akin to tri-allelic SNPs, which, as single-site loci, have more potential to be amplified from very short fragments. Nevertheless, Microhaplotypes represent the greatest potential to maximise the discrimination power of a set of short fragments analysed by sequencing technology and are therefore the ideal polymorphisms for MPS, as this provides the simplest way to know the phase of the SNP combinations they contain. USC has a clear focus now to include as many Microhaplotype and multiple-allele SNPs into forensic MPS panels to make full use of the power of sequencing to provide the most variant information per DNA fragment.

Both the above MPS sets incorporated Microhaplotypes that have been recently compiled by Kiddlab [177]. The loci chosen are likely to be highly informative, and will have the enhanced ability (compared to single-site SNPs) to detect mixtures. Two challenges remain: to find more Microhaplotypes in the genome and to devise a simple way to identify those loci compiled from genome-wide searches that are most effective for ancestry analysis. To identify Microhaplotypes most informative for normal forensic identification purposes just requires the calculation of Gene Diversity values (a variant on the Heterozygosity metric applied to multiple polymorphisms in linkage such as Y-STRs). This is a straightforward step from population frequency data already provided in 1000 Genomes. There is the need to convert the phase of the genotypes into haplotype combinations; i.e. three SNPs given as A|G, G|G, T|C are then recorded as haplotypes AGT and GGC, which are distinguishable from samples that have the same genotypes but are actually strand combinations AGC, GGT - which are counted separately when compiling the total variability of the Microhaplotype. The same problem of loci with similar Gene Diversity values in different populations, but quite different distributions of haplotype frequencies, exists in Microhaplotypes as it does with multiple-allele SNPs, previously illustrated in **Figure D8**. Therefore, the construction of appropriate metrics to highlight the most ancestry-informative Microhaplotypes will be an important step in compiling a much larger set of candidate loci from which to choose the best or as a way to maximise the potential to differentiate more closely related populations.

Box 12. Microhaplotypes

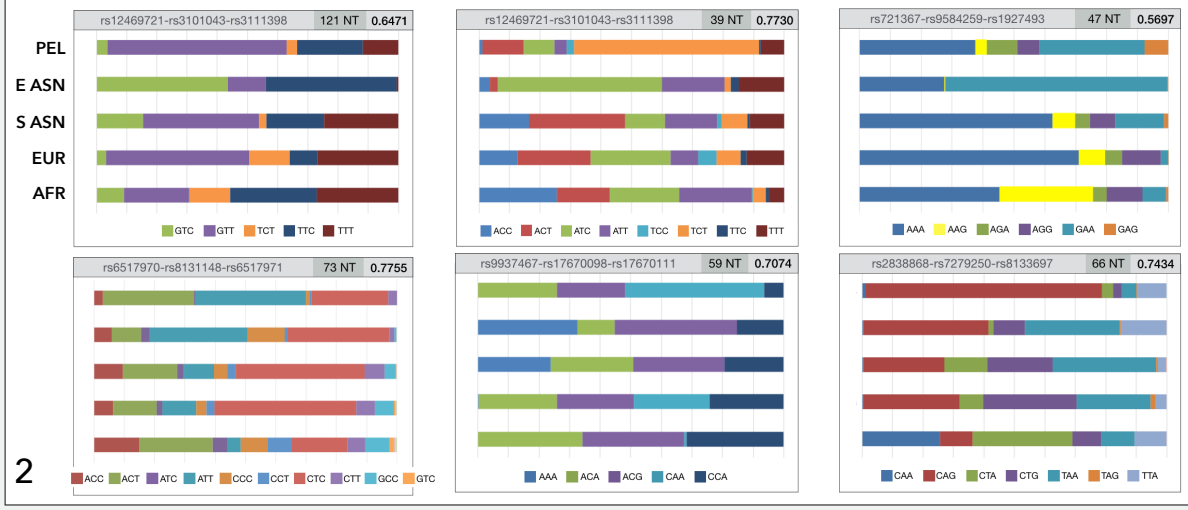
Microhaplotypes, comprising small sets of SNPs closely positioned on short chromosome segments up to 200 nucleotides long, have already been highlighted as potentially very useful forensic markers. They require knowledge of the **phase** of the component SNP genotypes - where individual combinations of alleles on each strand are discerned from the sequence of nucleotides; **only obtainable from MPS analysis of component SNPs**. Numbers of haplotypes will exceed combinations of SNP genotypes alone (e.g. CT, CT, CT can be 8 different strand combinations of CCC/TTT; CTC/TCT; etc.), but in practice, just a few haplotypes predominate - as a new SNP event will usually lock the novel variant nucleotide into the allelic background on the same strand. As the formation of novel SNPs are unique events, these tend to remain population specific, making Microhaplotypes highly applicable as forensic ancestry-informative markers, provided they can be a segment length that retains the most informative loci and be shorter than a maximum 160-200 nucleotides.



Panel 1: An example ancestry-informative Microhaplotype, MH-14 (SNPs in the chromosome segment GRCh37 14:106009477-106009495 or 106009572). This set was reduced in size from a 4-SNP 95-NT segment to a 3-SNP 18-NT one, but this process removed the Eurasian informative SNP allele **rs10134526-C** [lilac bars and allele] present in 1000 Genomes European, South Asian and Peruvian (PEL) samples (the latter likely from admixture). Interestingly, the African indicative SNP alleles of **rs28529526-A** and **rs7579473** [red/purple bars and alleles] and E Asian specific allele **rs74246343-G** [blue bars and allele] preserve the high informativeness of this Microhaplotype, which in its longer form can differentiate African, East Asian and Eurasian populations efficiently. This Microhaplotype has been adopted for a large forensic MPS ancestry panel, but there were insurmountable primer design problems in including the fourth SNP rs10134526 and maintaining the Eurasian specific haplotype this generated. Note that the GAT/GATT haplotypes are a single observation and may represent a very rare recombination event between rs28529526-rs75794737.

The **higher levels of polymorphism** in Microhaplotypes has made them versatile markers for forensic applications and they are now being incorporated by USC into MPS panels for enhanced **kinship analysis** as well as separately into an **ancestry inference** panel alongside tri-allelic and binary SNPs. Several Microhaplotypes are also key loci in a mixture analysis panel for the Ion Torrent MPS system.

Panel 2: Six 3-SNP ancestry informative Microhaplotypes with reduced sizes, high polymorphism and contrasting population frequencies. Top right value = Gene Diversity



To find new Microhaplotypes is also a bio-informatics challenge that will benefit from the development of an automated system to recognise SNP sets showing strong contrasts in allele frequencies between closely sited positions in the genome. Until now, divergent genes provide the easiest way to begin the search by hand. To illustrate this principle (and ahead of developing scripts for more unbiased searches of the whole genome), the divergent gene HERC2 was systematically searched for Microhaplotypes of potential use for forensic analysis. This involved listing of all SNPs in the gene using the gene query function of ENGINES. All SNPs without rs-numbers were eliminated as likely to be lacking high levels of polymorphism (previously undetected outside of 1000 Genomes analyses). Then sets of closely sited SNPs, +/- ~160 nucleotides of each other, were identified and their allele frequencies reviewed for sharp contrasts. Finally, population group haplotype frequencies were collected by converting individual genotypes into phased haplotypes and these were compiled into bar charts to illustrate the population distribution of variability in each locus.

Although a lengthy process by hand, searching just this one medium-sized gene revealed five Microhaplotypes that would be worthwhile forensic loci, either for ancestry inference or identification and mixture detection purposes. Four were below 100 nucleotides in length and all showed higher levels of polymorphism in some or all populations than most tri-allelic SNPs. Compiling four common-variation SNPs clearly provided higher overall discrimination power than Microhaplotypes of two SNPs, and it is likely the optimum way to balance fragment length and forensic power is to search for three common-variation SNPs in combination. The final discoveries made in this simple pilot study are summarised in **Figure D18**. Eventually, when sufficient numbers of candidate loci have been found, a large-scale multiplex can be constructed for MPS that is likely to represent the most informative way to analyse degraded DNA for both forensic identification and ancestry inference. Although it is tempting to assume that haplotypes are population-specific and can provide accurate data about admixture, this is not the case with early assessments. For example, the linked GCC / GAT haplotypes in MH-2 / MH-3 in **Figure D18** are found in Africans at high frequency and in some Peruvians. However, these are not individuals with the highest proportion of African co-ancestry in the PEL population sample (based on 580,000 autosomal SNPs).

The next task at USC is to assess how well the Microhaplotype loci we have already selected perform in MPS tests and, from the SNP data generated, begin to develop ancestry inference regimes in *Snipper* that can handle multiple-allele SNP-based loci, as described in previous sections. A more focused and systematic search across the whole genome for clusters of SNPs with contrasting allele frequencies will provide a much larger pool of candidate Microhaplotypes, and from genotyping a large selection of these, the potential for improving the regional population resolution from forensic-scale ancestry tests can be fully evaluated using the power of MPS to detect haplotype phase.

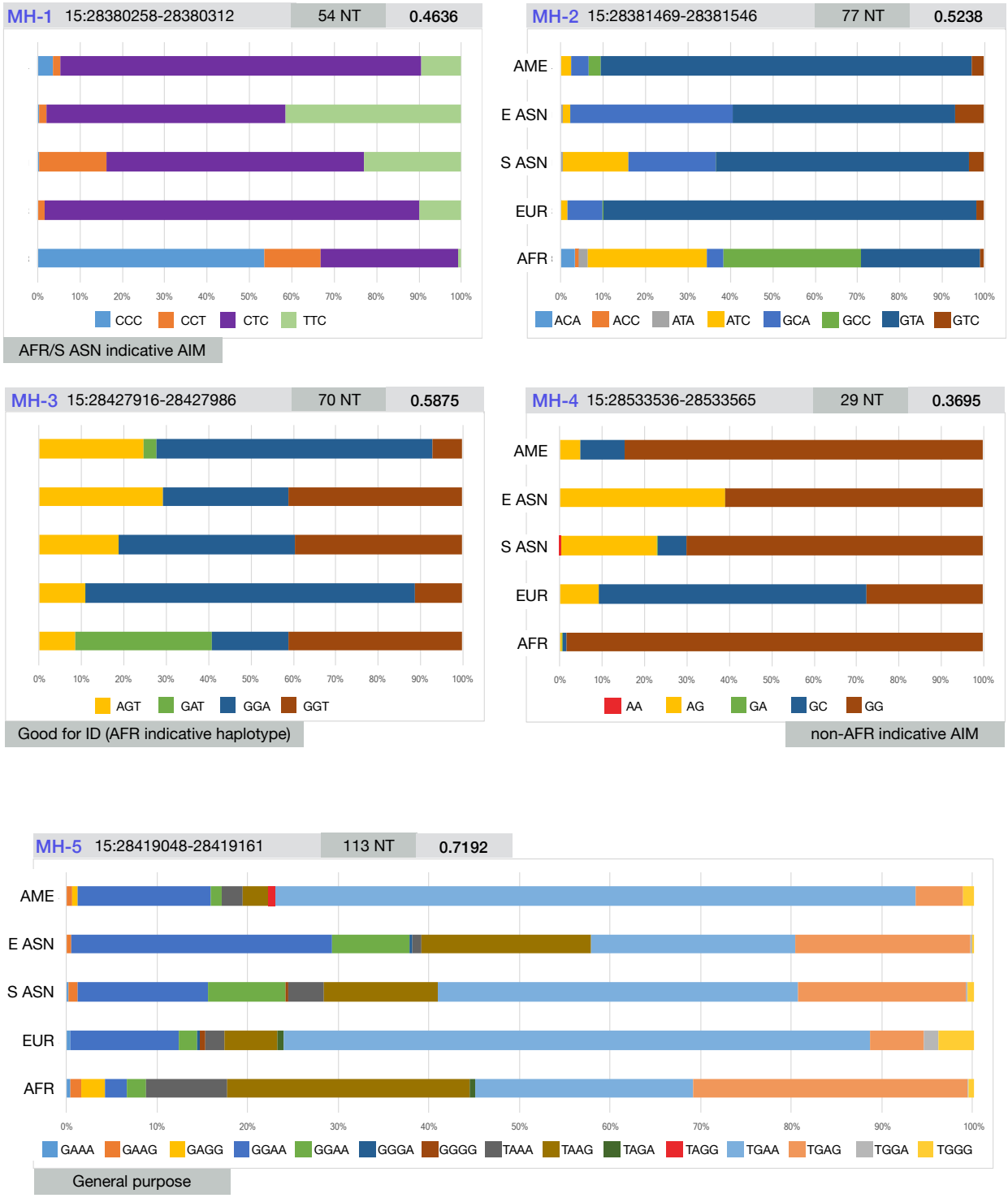
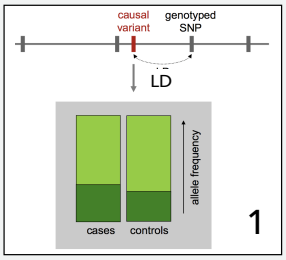


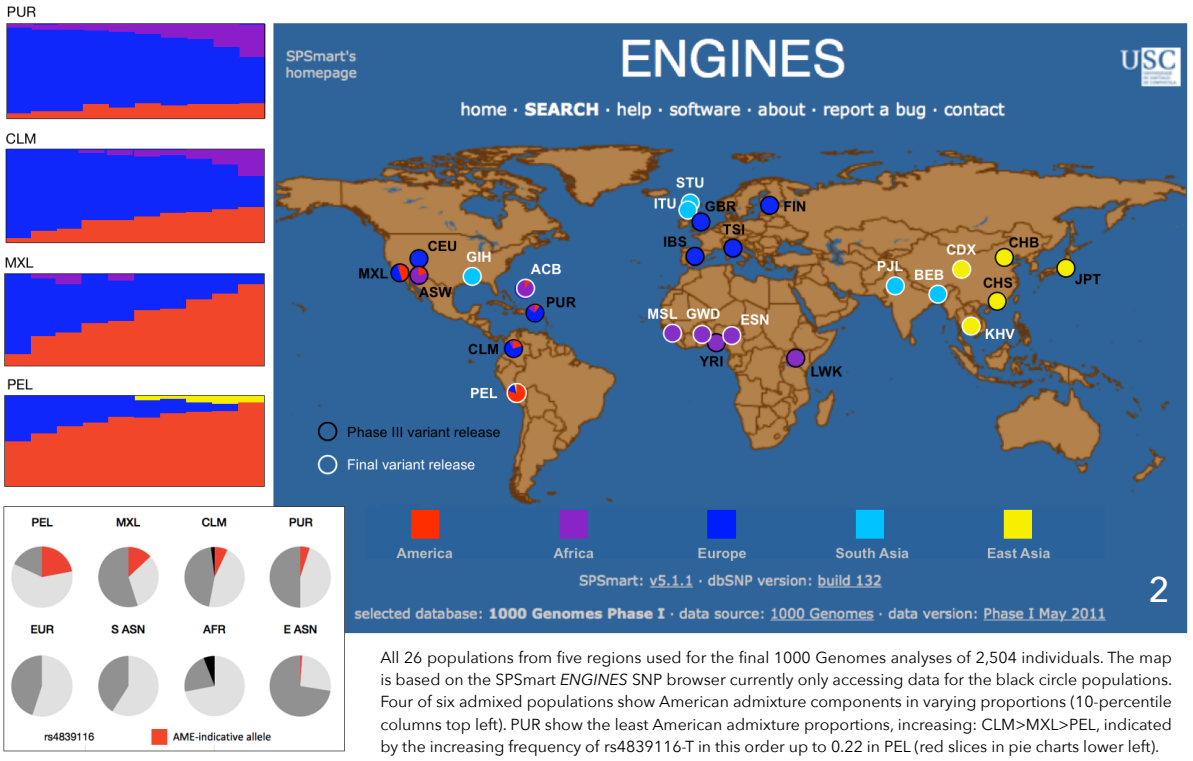
Figure D18. Five Microhaplotype discoveries made in the HERC2 gene; one comprising the 2-SNP combination of MH-4; three comprising 3-SNP combinations; and MH-5, a 4-SNP combination which is slightly longer than the others but with the highest average Gene Diversity (average of E ASN, S ASN, EUR and AFR variation - rightmost value above each chart). Sequence lengths between outermost SNPs are given in nucleotides (NT) and locations listed on the left above each chart. Arguably, the best marker is MH-3, with a high overall Gene Diversity value, a small length and an African-indicative GAT haplotype. MH-4 is a very informative Microhaplotype for distinguishing African and European populations from others, from the detection of all non-GG haplotypes and the GC haplotype, respectively - emphasising the ability of many single Microhaplotypes to be informative for multiple populations. Although in the top two Microhaplotypes, South Asians share a haplotype with Africans (MH-1 CCT; MH-2 ATC), both allow their differentiation from Europeans.

Box 13. 1000 Genomes as a forensic SNP data resource

Up to 2012, when the first 1000 Genomes data was released, Hapmap was the most extensive human SNP catalog. Hapmap was limited by its reliance on whole-genome scan SNP arrays (WGS) that only genotype a selected proportion of SNPs - those with sufficient power and appropriately positioned to give associations with the causal variants of disease (**Panel 1** outlines this principle of WGS association tests). Therefore, the logical step was made by 1000 Genomes to take advantage of increasingly cost-effective re-sequencing at genome-wide scales to compile a full human variant catalog. This growing catalog provided much more extensive human polymorphism data of the highest quality from which to re-assess common SNP variation for forensic purposes. The 1000 Genomes project also established a larger sampling regime compared to Hapmap, designed to identify variants occurring at **≥1% minor allele frequencies** and largely succeeded in achieving this goal in **2,504** genomes from individuals in **26 populations**. The final catalog of **77 million SNPs** was published in June 2014 as interim data and finalised in October 2014 (expanded to include short Indels). This final release also compiled revisions of many SNPs that had been incorrectly identified due to sequence misalignments or segmental duplications that can mimic true SNP sites.



Therefore, what characteristics of 1000 Genomes data are important for forensic SNP development? Firstly, the current catalog lacks variants on Chromosome-X and -Y, although these should be added within ~12 months of the autosomal SNP data released in late 2014. Secondly, it is notable that multiple-allele SNPs (see **Box 3**) are now fully characterised, whereas Hapmap and 1000 Genomes interim Phase III data had excluded these markers. Since multiple-allele SNPs are more discriminatory per locus than binary SNPs, provide better mixed DNA detection and can be highly informative AIMs, this part of the human SNP catalog will be a valuable forensic resource. 1000 Genomes have identified a much larger number of multiple-allele SNPs than might be expected and a review of tetra-allelic SNP characteristics emphasises the ability of 1000 Genomes to detect very low frequency alleles, that form the bulk of the third and/or fourth substitutions in such loci. Amongst >77 million SNPs; **508,917** (or 1 in 152) are described in the 2014 release as **multiple-allele sites**.



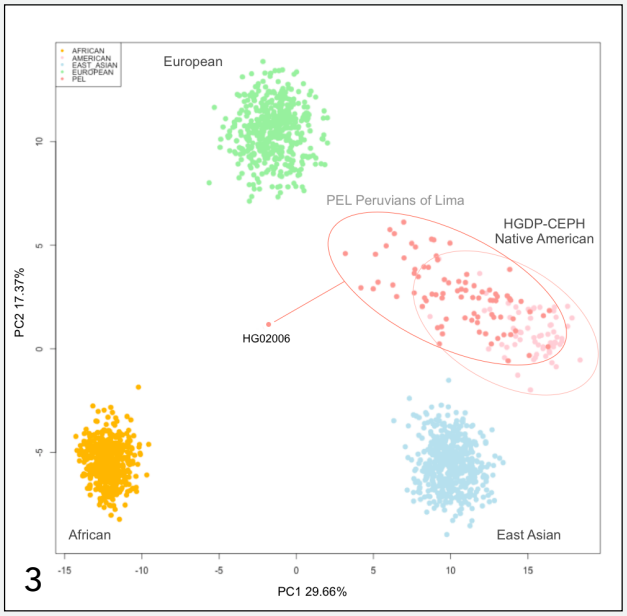
All 26 populations from five regions used for the final 1000 Genomes analyses of 2,504 individuals. The map is based on the SPSmart *ENGINES* SNP browser currently only accessing data for the black circle populations. Four of six admixed populations show American admixture components in varying proportions (10-percentile columns top left). PUR show the least American admixture proportions, increasing: CLM>MXL>PEL, indicated by the increasing frequency of rs4839116-T in this order up to 0.22 in PEL (red slices in pie charts lower left).

Thirdly, re-sequencing can directly detect allelic phase, rather than inferring it, as did Hapmap. This means it is possible to reliably identify and compile haplotypes of SNP variants formed by low-level but persistent recombination in short chromosomal spans. SNP **Microhaplotypes** are likely to be useful variation potentially differentiating individuals from the same population but different lineages. Microhaplotypes will be easily typed with massively parallel sequencing (*Kidd et al., 2014, Forensic Sci. Int. Genet. 12: 215-24*). Although, careful scrutiny of 1000 genomes SNP data will be necessary to find sets of SNPs in sufficiently short chromosome segments to be forensically sensitive (i.e. amplicons <100 bp), yet with enough historic recombination to create haplotype divergence. Fourthly, since the final 1000 Genomes sampling regime lacks individuals from **Oceanian** (OCE) or **unmixed Native American** (AME) population groups, other databases must be compared in order to assess patterns of variability in these regions. The best, indeed the only coverage of OCE and AME populations is provided by genotyping the HGDP-CEPH sample set with the Illumina 650,000-SNP WGS array (herein **650K**).

Box 13 /continued. 1000 Genomes as a forensic data resource - using PEL and the promise of STR data

The 650K array was used in parallel studies by Stanford and Michigan Universities (*Li et al., 2008, Science 319: 1100-4* and *Jakobsson et al., 2008, Nature 451: 998-1003*). Therefore, currently the most comprehensive assessment of worldwide human variation is obtained by collecting 1000 Genomes data for Africa (5 unmixed continental populations, 504 samples); Europe (5, 503); East Asia (5, 504); South Asia (5, 489) - as detailed in **Panel 2**; then supplementing these with HGDP-CEPH populations from the regions of Oceania (2, 27) and America (5, 64). The HGDP-CEPH sampling also includes one **North African** and three **Middle East** (ME) populations that can be useful to include in comparisons within the trans-continental Eurasian region. The six American-region populations in 1000 Genomes comprise two admixed African populations with a range of minor European admixture component proportions from 5 to 30%. The other four comprise more complex combinations of European, African, East Asian and Native American admixture contributors described in the summary cluster plots based on 10-percentile sampling of STRUCTURE analyses of these populations (128 *Global* SNP set data).

The combination of 1000 Genomes data with HGDP-CEPH OCE and AME currently limits SNP analysis to the 650,000 loci of the 650K panel, but 1000 Genomes now intends to re-sequence the full HGDP-CEPH panel, thus extending the genotyping of variants in ME, OCE and AME populations. Before this coverage gap is addressed, an alternative way to assess AME variation patterns for all SNPs is to collect and interpret those of **Peruvians from Lima (PEL)**, as this 1000 Genomes population, added at the final project phase, has the **lowest proportion of non-AME admixture** components, that comprise major European and minor East Asian contributions. An example of how AME-specific variants can be identified in this way is outlined in the example of rs4839116 shown in the pie charts of **Panel 2**, where the allele is only present at significant frequencies of 0.22 in PEL and at reduced levels in MXL, CLM and PUR. This interpretation of PEL variation is underlined by the PCA plot of **Panel 3** that shows most PEL samples cluster close to HGDP-CEPH AME individuals and do not overlap with Europe. The single HG02006 outlier can be excluded from analyses of this population which provides the best 1000 Genomes proxy for unmixed American variation.



3

Lastly, mention should be made of microsatellite data from 1000 Genomes, as this would represent a valuable source of data on STR repeat allele variation in a large set of population samples, but more importantly, could begin the process of compiling sequence variants in the common repeat alleles of each core forensic STR. At the end of 2014, 1000 Genomes released an extensive catalog of **670,646 microsatellites** in lobSTR format, but these suffer from bias towards short STRs and therefore exclude details of core forensic loci such as FGA, SE33, D21S11 that have very long repeated sequence segments that escape efficient alignment and characterisation. Furthermore, initial checks made by the author of short-sequence STRs such as the NIST Mini-STRs show wide-scale departures from Hardy-Weinberg equilibrium (examples of highly skewed heterozygosities in D9S1120 for the five AFR populations is shown in **Panel 4**). These checks indicate the collation of STR variation by 1000 Genomes must be brought up to a much higher level of reliability before it has forensic utility. While this will take some time and may not occur quickly for longer forensic STRs, 1000 Genomes does provide useful SNP and Indel data for the **flanking sequences** (~100 bp +/- of the repeats) containing such variants and captured by MPS analyses. So far, searches of core STRs have identified nine common-variation SNPs closely sited to the repeat regions (so-called **SNP-STRs**) as listed on the right. These give scope for differentiation of repeat alleles that are identical by state (**IBS**), but from different donors (e.g. in mixed DNA). They also allow matching of common repeat alleles between relatives (**IBDescent**), enhancing the power of e.g. familial searching, where the most common STR alleles would not normally be informative.

D9S1122	ESN	GWD	LWK	MSL	YRI
Discrimination Power	0.8243	0.8264	0.7904	0.8306	0.8079
Homozygotes	0.7703	0.7368	0.8182	0.8056	0.7432
Heterozygotes	0.2297	0.2632	0.1818	0.1944	0.2568
	Allele frequencies				
8	0.0270	0.0263	0.0909	0.0000	0.0270
9	0.0405	0.0105	0.0000	0.0347	0.0068
10	0.2095	0.1684	0.2078	0.2361	0.1824
11	0.3919	0.3737	0.4351	0.2847	0.3851
12	0.2432	0.3263	0.2403	0.3264	0.3378
13	0.0811	0.0895	0.0260	0.0764	0.0608
14	0.0068	0.0053	0.0000	0.0417	0.0000
	HWE analysis				
EXP heterozygosity	0.5788	0.5457	0.6496	0.6009	0.5819
OBS heterozygosity	0.2297	0.2632	0.1818	0.1944	0.2568
Chi square	0.2105	0.1463	0.3369	0.2750	0.1816
P-value	0.9998	0.9999	0.9993	0.9996	0.9999

4

Hardy-Weinberg analysis of observed vs. expected heterozygosities in 1000 Genomes populations for STR D9S1122. A very large excess of homozygotes reported creates highly significant Chi^2 values indicating this data is unreliable.

Nine best SNP-STRs

D13S317-**rs9546005**
D16S539-**rs11642858**
D2S1338-**rs6736691**
D1S1656-**rs4847015**
D2S441-**rs79534691**
D5S818-**rs25768**
D10S1248-**rs2246512**
D7S820-**rs16887642**
VWA-**rs75219269**

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Concluding remarks

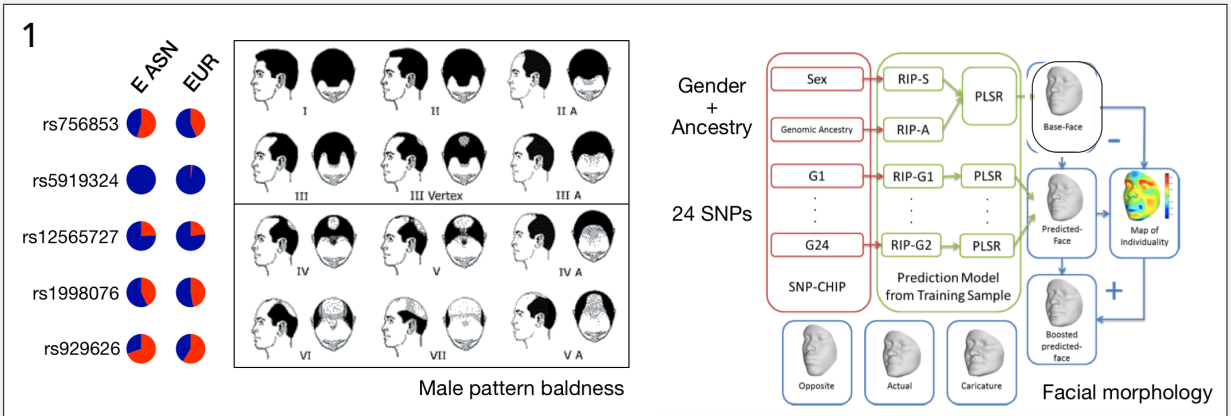
Ancestry inference tests that form part of forensic DNA analysis are being increasingly applied, helped by the growing acceptance of MPS technologies as a way to reliably genotype a large array of SNPs in one sensitive multiplex, in a system which can combine STRs and SNPs, while including tests for ancestry, phenotypes (EVCs) and identification form use of very short amplified fragments. The links between tests for ancestry and inference of likely physical appearance are growing stronger too - as a more detailed picture can be gained of an unknown suspect when age estimation, a battery of EVC tests and reliable inference of the individual's ancestry are compiled into a more complete idea of what the person looks like. This idea of "synergy" between age, ancestry and appearance tests for forensic analyses that are complimentary to routine STR profiling - so-called "AAA" testing, is explored further in **Box 14**. In each case, individual tests are strengthened by data from the other two. Although MPS is expensive and time-consuming, it can form part of the toolbox in all forensic DNA labs and may be used as a second-strike system of analysis. We have seen numerous examples in this thesis of the successful application of very simple CE-based forensic ancestry tests to solve a criminal investigation (**Boxes 1, 2, 10**), help identify the remains of a missing person (**Figures D3** and **D4**) or solve a paternity case that was easily resolved despite being complicated by the absence of both putative fathers (**Box 11**). Therefore, there is plenty of scope for both CE and MPS technologies to exist as complementary technologies in the DNA lab, and for the ease-of-use of CE analysis to successfully resolve a large proportion of tests before MPS is necessary or chosen to increase the quality of data obtained to improve ancestry inference (e.g. if there is reason to believe the individual has admixed ancestry and will be best analysed with the largest array of AIM-SNPs).

For this reason, the author has continued to develop new forensic ancestry tests and look for ways to improve the depth of data, scope of markers and utility of both established CE and new MPS technologies for ancestry inference purposes. The search for better or enlarged numbers of markers has required increasing use of the online population data resources offered by such large-scale projects as 1000 Genomes, which has the advantage that almost all common human variants have been identified and cataloged. The recent availability of additional whole-genome sequence data from initiatives such as Simons Foundation Human Genome Diversity Project, and the aim of 1000 Genomes to sequence the samples of the HGDP-CEPH panel, means an increasing amount of data will be available to use in the search for new AIMs. The value of in-silico resources such as 1000 Genomes and Simons Foundation HGDP is outlined in **Box 15**, and the author has begun to move increasingly towards in-silico searches without recourse to genotyping, until the best markers have been compiled. It can be argued that MPS has arrived at a good moment to make best use of Microhaplotype markers in forensic analysis, but equally, easy access to whole-genome sequence data, where SNP phase is detected, will provide a rich resource for the identification of new Microhaplotypes - as shown in the pilot studies outlined in **Fig. D18** and **Box 13**. These online repositories of whole-genome variant catalogs in different populations will also prove invaluable in the drive towards higher data depths from enlarged marker panels and necessary for improved geographic resolution - the ability to differentiate populations occupying more closely positioned regions that, without geographic barriers, show less divergence than the continentally defined major population groups. The fine-scale geographic resolution within one region that is possible with very large arrays of random SNPs and the

possibilities to get close to this level of resolution with smaller sets of carefully chosen SNPs for forensically sensitive tests is explored in the final **Box 16**. There will an increased expectation that the data For this reason, the author has continued to develop new forensic ancestry tests and look for ways to improve the depth of data, scope of markers and utility of both established CE and new MPS technologies for ancestry inference purposes. The search for better or enlarged numbers of markers has required increasing use of the online population data resources offered by such large-scale projects as 1000 Genomes, which has the advantage that almost all common human variants have been identified and cataloged. The recent availability of additional whole-genome sequence data from initiatives such as Simons Foundation Human Genome Diversity Project, and the aim of 1000 Genomes to sequence the samples of the HGDP-CEPH panel, means an increasing amount of data will be available to use in the search for new AIMs. The value of in-silico resources such as 1000 Genomes and Simons Foundation HGDP is outlined in **Box 15**, and the author has begun to move increasingly towards in-silico searches without recourse to genotyping, until the best markers have been compiled. It can be argued that MPS has arrived at a good moment to make best use of Microhaplotype markers in forensic analysis, but equally, easy access to whole-genome sequence data, where SNP phase is detected, will provide a rich resource for the identification of new Microhaplotypes - as shown in the pilot studies outlined in **Fig. D18** and **Box 13**. These online repositories of whole-genome variant catalogs in different populations will also prove invaluable in the drive towards higher data depths from enlarged marker panels and necessary for improved geographic resolution - the ability to differentiate populations occupying more closely positioned regions that, without geographic barriers, show less divergence than the continentally defined major population groups. The fine-scale geographic resolution within one region that is possible with very large arrays of random SNPs and the possibilities to get close to this level of resolution with smaller sets of carefully chosen SNPs for forensically sensitive tests is explored in the final **Box 16**. There will an increased expectation that the data obtained from abundant whole-genome sequences in a global selection of populations, will lead to forensic ancestry tests that can reliably distinguish them, while handling the complexities of population admixture, increasingly common in urban demographic profiles.

Box 14. The synergy between forensic tests for ancestry and externally visible characteristics

The development of forensic SNP tests for ancestry inference initially preceded, then ran parallel to tests made to predict common externally visible characteristics (EVCs) such as eye colour. Although **EVCs are a fundamental aspect of recognition** in all human cultures and therefore form the **primary basis for identification** or elimination of a suspect by eyewitness (when viewing photographs or an identity parade), the bulk of forensic R&D in this field has concentrated on patterns of physical trait variation in Europeans. This is partly because certain traits such as eye and hair colour only show a sufficiently broad range of phenotypic variation in Europeans. However, it is also due to the focus of association studies on European-ancestry subjects and the concentration of forensic genetics research efforts in European and US labs. Such emphasis on Europe creates two problems that require an inference of ancestry to be made before EVC predictions are attempted (although others have argued it is not necessary, see: *Walsh et al., 2011, Forensic Sci. Int. Genet. 5: 170–80*). Firstly, in most EVCs the complex interactions between those variants identified to be strongly associated with the trait and the genetic background, contribute to the expression of the trait through **epistasis**. Yet, epistatic effects are not easy to map and tend to differ between populations due to their highly contrasting genetic backgrounds. This means that near identical EVC-associated variant frequency distributions may be found in different populations, but their expression, often seen as a lack of variation in the studied trait, can be quite different. Secondly, **population admixture generally leads to reduced predictive accuracy** from EVC tests, because complex combinations of variants (each with varied effect sizes) interfere with the expected patterns of genetic control of the trait. This is observed as a reduction in predictive accuracy in individuals from admixed populations compared to the unmixed European samples used to develop and evaluate forensic EVC-SNP tests, even when the major admixture component in such individuals is European.

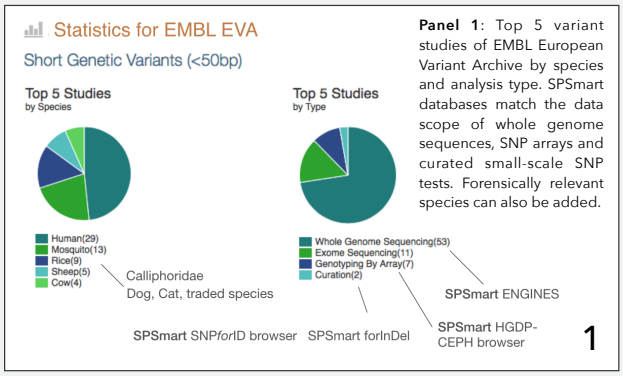


Two examples of forensic EVC predictive tests where ancestry inference is central to accurate interpretation of SNP data are outlined in **Panel 1**. Early-onset male pattern baldness (**MPB**: hair loss scales IV to VII) has been associated most strongly in European males with the five SNPs with allele frequencies shown. This **SNP variation is near-identical** in East Asians and Europeans, although the expression of **MPB is quite different in East Asia** - where it is very limited in extent and mostly occurs much later in life. The presence of the rs3827760-G EDAR allele at very high frequencies in East Asians exemplifies how the genetic background can differ and, causing thicker hair in Asians, is likely to alter expression of the MPB-associated genetic variants identified in Europeans - all having rs3827760-G backgrounds. Facial morphology predictive tests have undergone rapid development recently but as face shape is quite different between genders and ancestries, both must be established before a base-face (dark square) is then remodelled using the 24 associated SNP variants.

Panel 2 is "Black Americans" a celebrated photograph by Bruce Davidson taken in 1962 of two girls in a segregated milk-bar, indicating the one on the right is an African American with obviously pale pigmentation features more typical of a European ancestry. This illustrates the common occurrence of **unexpected trait combinations in persons with co-ancestry.** The optimum approach in cases of high levels of population admixture (EUR co-ancestry reaches 30% in African Americans) is to **test ancestry and EVCs side-by-side** and use each to inform the interpretation of patterns of variation found. With this goal in mind, USC developed a skin colour predictive test (*Maroñas et al., 2014, Forensic Sci. Int. Genet. 13: 34–44*), because estimating the admixture ratio in an individual does not necessarily help to predict their pigmentation patterns if these are dictated by a small number of variants and **disproportionately inherited**.

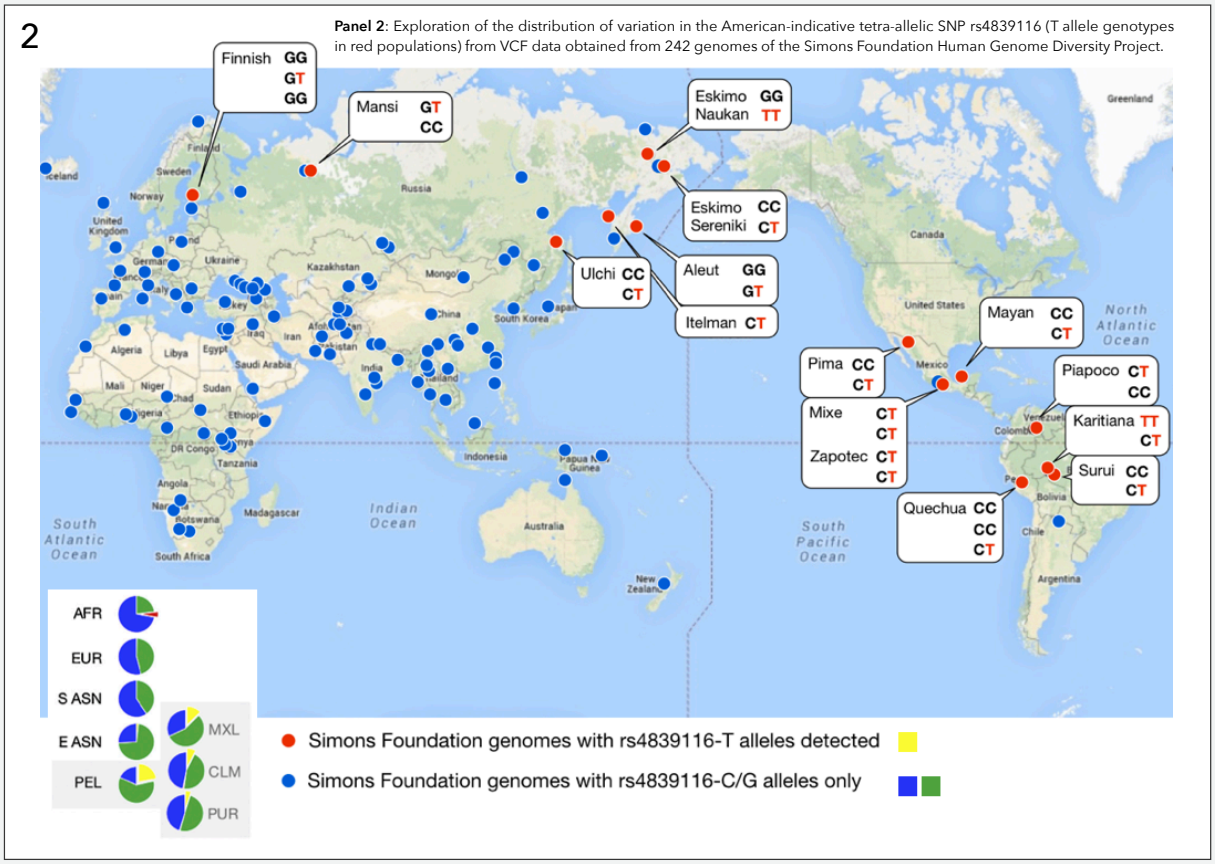
Box 15. Exploring populations with new whole-genome sequence data

With the formal completion of 1000 Genomes at the start of 2017, the project has moved on to analyse the HGDP-CEPH panel and further samples. In 2016, increasing numbers of parallel whole-genome sequencing (WGS) projects also released variant data; notably the **Simons Foundation Human Genomes Diversity Project** (SFHGDP); currently comprising variant data from genome sequences in 263 samples from widely distributed populations. A set of 21 samples are from 1000 Genomes for control purposes and 122 from the HGDP-CEPH panel. The remaining 120 are new and have a wide geographic distribution - the 122 CEPH and 120 new samples are designed to be based on small population sample sizes and comprehensive regional coverage. USC are using the 1000 genomes variant catalog as the training sets for established and novel SNP sets, and SFHGDP samples as the test set in *Snipper*, although the analysis of Microhaplotypes is not possible; as SFHGDP genotype data does not record SNP phase. Additional variant catalogs are being compiled by the **European Variant Archive** (EVA - **Panel 1**). The EVA now comprises 53 WGS studies; 11 Exome sequencing studies and 7 from SNP arrays. This reflects the small-scale reporting of SNP data of SPSmart in *ENGINES* and the CEPH browser, as well as the curated (custom SNP set) databases of SNP*for*ID SNPs and Indels. There are no plans to use the growing amount of whole genome data to extend SPS, but WGS data is already a valuable resource for population studies of the established and novel forensic SNP sets at USC.



Panel 1: Top 5 variant studies of EMBL European Variant Archive by species and analysis type. SPSmart databases match the data scope of whole genome sequences, SNP arrays and curated small-scale SNP tests. Forensically relevant species can also be added.

For ancestry-informative markers studied at USC showing particular characteristics, e.g. multiple-allele SNPs or population-specific variability, the SFHGDP data can often provide interesting patterns. **Panel 2** shows the rs4839116-T allele was previously found in 1000 Genomes admixed American populations, suggesting it might only be found beyond Native American populations around NE Asia - where founder bottlenecks of early settlers from **Beringia** can create patterns of variation specific to this region as well as current Native American or admixed American populations. The **rs4839116-T** variant provides a **signature of NE Asian origins of Native Americans,** with a presence detected in 6 of 18 chromosomes in five populations from this region and 12 of 34 chromosomes in six Native American populations (of 8 sampled). Furthermore, some low level migration west by Beringians into western Siberia and Finland can be inferred.

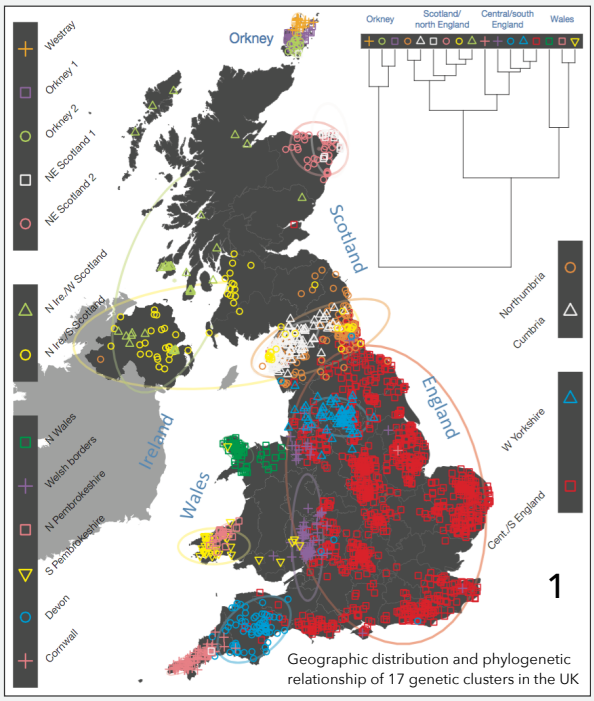


Panel 2: Exploration of the distribution of variation in the American-indicative tetra-allelic SNP rs4839116 (T allele genotypes in red populations) from VCF data obtained from 242 genomes of the Simons Foundation Human Genome Diversity Project.

Box 16. Future directions for forensic ancestry analysis

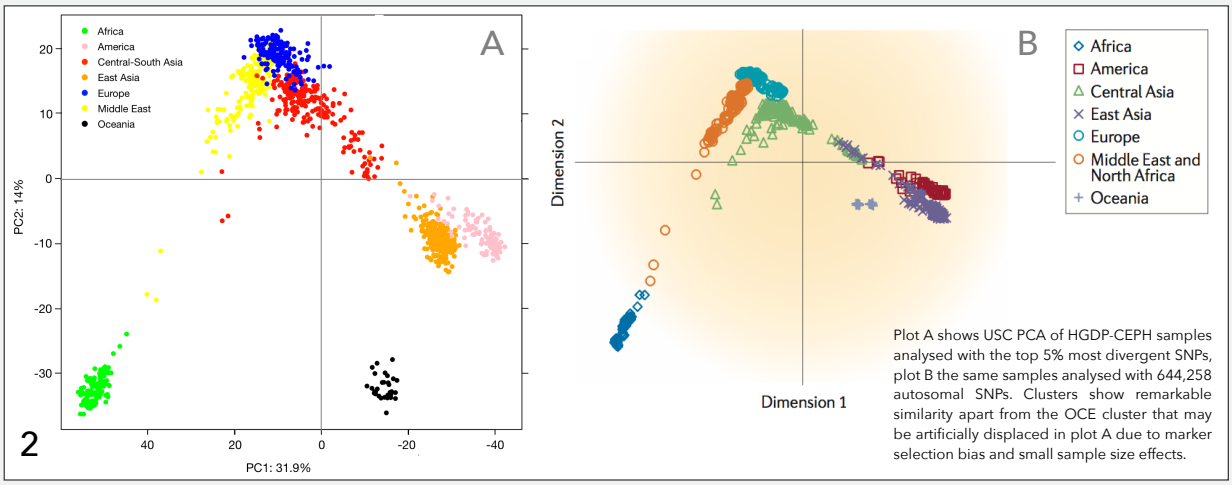
As MPS techniques start to gain traction in forensic DNA analysis there are good prospects for increasing the number of markers that can be reliably genotyped from small amounts of evidential material. For ancestry analysis there will be an inevitable balance-point between how many loci can be genotyped and what level of geographic resolution they bring. It is likely to be the case that certain population differentiations, while desirable to investigators, will not be straightforward due to a lack of sufficient genetic divergence in the first place. However, when much **higher marker depths** can be obtained, e.g. from the **650K WGS** array, it is possible to detect very **fine-scale structure** within the same region, as was recently demonstrated for the UK (*Leslie et al., 2015, Nature 519: 309-14*).

The 2015 study of Leslie took the 500,000 autosomal SNPs of the 650K array and used these to analyse the detailed haplotype structure of 2,039 UK natives whose grandparent's birthplace was known and therefore provided the centroid-based co-ordinates for their 'place-of-origin'. The statistical analysis of haplotype decay by recombination using ***fineSTRUCTURE***, identified 17 distinct genetic clusters with a good match to geography and known migration/invasion events in the last 2,000 years of the UK's demographic history (**Panel 1**). Despite very fine-scale geographic resolution between the 17 clusters, their genetic divergence remained extremely weak (compared to the levels of divergence quoted for most of the AIMs sets discussed in this thesis): **average F_{ST} was 0.002** (maximum 0.007) - some 400x less than the F_{ST} between AFR and non-AFR for the *Duffy* SNP rs2814778. Nevertheless, this study reveals how much detail about a person's ancestry can be obtained if enough markers can be typed to fine-map the haplotype decay accrued in very short genomic spans. Therefore, how many SNPs could be an acceptable minimum number to act as an efficient proxy for 650,000? Analyses of the HGDP-CEPH panel with the **1000 most divergent SNPs** (Prof. Antonio Tato, USC Mathematics Dept.) produced a remarkably close match to those obtained from the 650K SNPs (**Panel 2**), suggesting there is enough genetic detail in as few as 1000 loci to be able to map the true relationship of the HGDP-CEPH populations with sufficient reproducibility and accuracy.



Geographic distribution and phylogenetic relationship of 17 genetic clusters in the UK

Notably, some overlap between the ME and EUR clusters in the USC plot underlines how difficult these two groups are to separate. Selecting a small sub-set of AIMs (100-125) to achieve this differentiation has been very challenging, as even SNPs with the highest divergence fail to create sufficiently tight and well spaced PCA clusters. However, this subset can be added to the 128 Global AIMs or 55 ***ForenSeg*** AIMs to improve the differentiation further. There is also the scope with MPS to compare whole sequences between populations. This data can concentrate on STRs and accompanying SNPs or Microhaplotypes if these eventually replace single targeted loci in the middle of relatively uninformative sequence. A major drawback of Microhaplotypes is whether those that are sufficiently informative can also be short enough to provide forensic sensitivity for the typing of highly degraded DNA. This fundamental characteristic is not overcome by higher multiplexing or technology, with DNA <100 bp in size surviving better than longer fragments (and even more pronounced in fragments <50 bp). Therefore, many Microhaplotypes will need size-reduction if the component SNP's informativeness for ancestry is largely preserved.



Plot A shows USC PCA of HGDP-CEPH samples analysed with the top 5% most divergent SNPs, plot B the same samples analysed with 644,258 autosomal SNPs. Clusters show remarkable similarity apart from the OCE cluster that may be artificially displaced in plot A due to marker selection bias and small sample size effects.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Conclusions

1. The compilation of a large panel of SNPs showing population differentiation properties has been successfully accomplished from an initial human SNP map of 1.42 million mainly common-variation loci, extended in recent years to the full 1000 Genomes catalog of ~79 million variants (of which a total of 77,520,219 are single nucleotide SNPs comprising simple A/C/G/T substitution polymorphisms). 1600 candidate ancestry informative SNPs and some 200 ancestry informative Indels have been compiled and now form the core set of markers from which to construct forensic ancestry panels.

2. SNaPshot genotyping has been optimised for a wide range of forensic SNP tests and these are applied as the system of choice for capillary electrophoresis applications in forensic laboratories. The sensitivity obtained with DNA analysis using SNaPshot tests often surpasses that of routine STR analysis for very degraded DNA, underlining the fact that primer extension tests make best use of the very short amplicons possible with SNPs.

3. A complete optimised pipeline has been put in place at USC for ancestry analysis based on the studies described in this thesis. The analysis steps encompass amplification and genotype detection followed by data analysis regimes in the Snipper web portal that produces a series of Bayes likelihood calculations and accompanying 2D principle component analysis plots in an open-access online framework. Snipper has been extended to analyse multiple allele ancestry markers comprising STRs and Microhaplotypes. At the same time, extensive population variation databases have been developed in the SPSmart human SNP data suite to allow custom compilation of combinations of populations and a wide range of SNPs plus forensic ancestry-informative Indels.

4. SNaPshot assays that are complimentary to the core 34-SNP three population group ancestry test have been developed for specific worldwide regions, comprising South Asia, Pacific region populations and Native Americans. Specialised SNaPshot assays to analyse European-African admixed individuals, DNA mixtures and to genotype X-chromosome ancestry SNPs have been successfully applied to forensic analyses and have fed additional ancestry markers into the pool of SNPs shown to be informative for this purpose.

5. SNP-based ancestry analysis tests have been successfully applied to a wide range of criminal cases and missing persons identification programs. In many cases the data obtained has provided critical evidence to direct a criminal enquiry or to enhance the knowledge of potential population of origin inferred from the uni-parental Y-chromosome and mitochondrial markers.

6. In total, 272,800 human tri-allelic SNPs and 1625 tetra-allelic SNPs have been identified from the 1000 Genomes project data. This comprehensive catalog of multiple-allele SNP variants have been compiled into a panel of ~2000 ancestry-informative SNPs plus a further 3000 showing low frequency, but population specific patterns of variation in one of three common alleles. These panels are being exploited to enable the inclusion of mainly tri-allelic SNPs into new ancestry

panels with the benefit that mixed DNA is more easily detected form the presence of more than two alleles in the capillary electrophoresis signals or sequence output.

7. Ancestry informative STRs genotyped with capillary electrophoresis has enabled detection of multiple alleles in a profile and consequently the presence of mixed DNA. Because of the low levels of population differentiation in the STRs studied to date, Indels were developed at the same time and these markers analysed with simple PCR-to-CE techniques have proved to be particularly robust and sensitive ancestry markers, highly suited to the de-convolution of complex mixed DNA patterns due to their very balanced peak signal ratios. The 46-Indel test developed for the purpose of detecting mixed DNA simultaneously with inferring ancestry has been particualrly successful and is now widely used outside of USC.

8. A successful transition has been made from PCR multiplexes of less than 50 markers into much larger multiplexes for massively parallel sequencing (MPS) technologies. Enlarged multiplexes for MPS currently comprise166 markers (145 single-site SNPs plus 21 Microhaplotype sequences that harbour 80 common SNPs); and approximately 180 markers divided between ancestry-informative SNPs and externally visible characteristic predictive SNPs. An additional panel has been developed by the author consisting of 1400 SNPs and Microhaplotypes, which indicates that much larger multiplexes can be developed in the immediate future. The ancestry informative and trial-allelic SNPs described in the thesis will form the core candidate lists for such panels.

9. The collateral variants associated with common repeat region variation in the core forensic STRs have been compiled and these indicate limited ancestry-informative properties as well as levels of linkage disequilibrium with particular repeat alleles. Similarly, while the sequence variation found in repeat regions in many core STRs is informative for differentiating iso-metric alleles, they show very little population specificity or occur at too low a frequency to be sufficiently informative. Conversely, flanking region Indels are population specific, but also occur at very low frequencies and currently lack a sufficiently comprehensive survey of their population variation.

10. A current catalog of 120 Microhaplotypes have been compiled and assessed for their ancestry informative properties. From the 22 Microhaplotypes added to a large MPS panel twenty are sequenced successfully and provide early indications that they are informative for the differentiation of populations within a continental group. A set of 600 new Microhaplotype loci have been compiled from the UK ExAC project. Early exploration show this panel has yielded Microhaplotypes with much more population-differentiated patterns of variation and consisting of short chromosome segments with higher numbers of SNPs, which will be highly suited to forensic analysis with short amplicon PCR.

References

(author's publications in bold)

[1] L. Spinney, Eyewitness identification: line-ups on trial, Nature 453 (2008) 442–444.

[2] R.V. Rohlfs, S.M. Fullerton, B.S. Weir, Familial identification: population structure and relationship distinguishability, PLoS Genet. 8 (2012) e1002469.

[3] S. Walsh, F. Liu, K.N. Ballantyne, M. van Oven, O. Lao, M. Kayser, Irisplex: a sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information, Forensic Sci. Int. Genet. 5 (2011) 170–180.

[4] A. Freire-Aradas, Y. Ruiz, C. Phillips, O. Maroñas, J. Söchtig, A. Gómez Tato, J. Álvarez Dios, M. Casares de Cal, V.N. Silbiger, A.D. Luchessi, et al., Exploring iris colour prediction and ancestry inference in admixed populations of South America, Forensic Sci. Int. Genet. 13 (2014) 3–9.

[5] L. Yun, Y. Gu, H. Rajeevan, K.K. Kidd, Application of six Irisplex SNPs and comparison of two eye colour prediction systems in diverse Eurasia populations, Int. J. Leg. Med. 128 (2014) 447–453.

[6] C. Bouakaze, C. Keyser, E. Crubézy, D. Montagnon, B. Ludes, Pigment phenotype and biogeographical ancestry from ancient skeletal remains: inferences from multiplexed autosomal SNP analysis, Int. J. Leg. Med. 123 (2009) 315–325.

[7] T.E. King, E.J. Parkin, G. Swinfield, F. Cruciani, R. Scozzari, A. Rosa, S.K. Lim, Y. Xue, C. Tyler-Smith, M.A. Jobling, Africans in Yorkshire? The deepest-rooting clade of the Y phylogeny within an English genealogy, Eur. J. Hum. Genet. 15 (2007) 288–293.

[8] C. Phillips, L. Prieto, M. Fondevila, A. Salas, A. Gomez-Tato, J.A. Alvarez-Dios, A. Alonso, A. Blanco-Verea, M. Brión, M. Montesino, et al., Ancestry analysis in the 11-M Madrid bomb attack investigation, PLoS One 4 (2009) e6583.

[9] S. Willuweit, L. Roewer, International forensic Y chromosome user group, Y chromosome haplotype reference database (YHRD): update, Forensic Sci. Int. Genet. 1 (2007) 83–87.

[10] W. Parson, A. Dür, EMPOP—a forensic mtDNA database, Forensic Sci. Int. Genet. 1 (2007) 88–92.

[11] C. Phillips, Ancestry informative markers, 2nd ed., in: J.A. Siegel, P.J. Saukko (Eds), Encyclopedia of Forensic Sciences, vol. 1, Academic Press, 2013, 2015, pp. 323–331.

[12] R.C. Lewontin, The apportionment of human diversity, Evol. Biol. 6 (1972) 381–398.

[13] M.A. Jobling, E. Hollox, M.E. Hurles, T. Kivisild, C. Tyler-Smith, Human Evolutionary Genetics, 2nd ed, Garland Science, New York, 2014.

[14] N.A. Rosenberg, J.K. Pritchard, J.L. Weber, H.M. Cann, K.K. Kidd, L.A. Zhivotovsky, M.W. Feldman, Genetic structure of human populations, Science 298 (2002) 2381–2385.

[15] J.K. Pritchard, M. Stephens, P. Donnelly, Inference of population structure using multilocus genotype data, Genetics 155 (2000) 945–959.

[16] H.M. Cann, C. de Toma, L. Cazes, M.F. Legrand, V. Morel, L. Piouffre, J. Bodmer, W.F. Bodmer, B. Bonne-Tamir, A. Cambon-Thomsen, et al., A human genome diversity cell line panel, Science 296 (2002) 261–262.

[17] S. Wang, C.M. Lewis Jr., M. Jakobsson, S. Ramachandran, N. Ray, G. Bedoya, W. Rojas, M.V. Parra, J.A. Molina, C. Gallo, et al., Genetic variation and population structure in Native Americans, PLoS Genet. 3 (2007) 2049–2067.

[18] J.S. Friedlaender, F.R. Friedlaender, F.A. Reed, K.K. Kidd, J.R. Kidd, G.K. Chambers, R.A. Lea, J.H. Loo, G. Koki, J.A. Hodgson, et al., The genetic structure of Pacific Islanders, PLoS Genet. 4 (2008) e19.

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

[20] D. Serre, S. Paäbo, Evidence for gradients of human genetic diversity within and among continents, Genome Res. 14 (2004) 1679–1685.

[21] N.A. Rosenberg, S. Mahajan, S. Ramachandran, C. Zhao, J.K. Pritchard, M.W. Feldman, Clines, clusters, and the effect of study design on the inference of human population structure, PLoS Genet. 1 (2005) 70.

[22] G. Coop, J.K. Pickrell, J. Novembre, S. Kudaravalli, J. Li, D. Absher, R.M. Myers, L. L. Cavalli-Sforza, M.W. Feldman, J.K. Pritchard, The role of geography in human adaptation, PLoS Genet. 5 (2009) e1000500.

[23] R.L. Lamason, M.A. Mohideen, J.R. Mest, A.C. Wong, H.L. Norton, M.C. Aros, M. J. Jurynec, X. Mao, V.R. Humphreville, J.E. Humbert, et al., SLC24A5 a putative cation exchanger, affects pigmentation in zebrafish and humans, Science 310 (2005) 1782–1786.

[24] D. Reich, M.A. Nalls, W.H. Kao, E.L. Akylbekova, A. Tandon, N. Patterson, J. Mullikin, W.C. Hsueh, C.Y. Cheng, J. Coresh, et al., Reduced neutrophil count in people of African descent is due to a regulatory variant in the Duffy antigen receptor for chemokines gene, PLoS Genet. 5 (2009) e1000360.

[25] C.J. Ingram, C.A. Mulcare, Y. Itan, M.G. Thomas, D.M. Swallow, Lactose digestion and the evolutionary genetics of lactase persistence, Hum. Genet. 124 (2009) 579–591.

[26] A. Fujimoto, R. Kimura, J. Ohashi, K. Omi, R. Yuliwulandari, L. Batubara, M.S. Mustofa, U. Samakkarn, W. Settheetham-Ishida, T. Ishida, et al., A scan for genetic determinants of human hair morphology: EDAR is associated with Asian hair thickness, Hum. Mol. Genet. 17 (2008) 835–843.

[27] K. Yoshiura, A. Kinoshita, T. Ishida, A. Ninokata, T. Ishikawa, T. Kaname, M. Bannai, K. Tokunaga, S. Sonoda, R. Komaki, et al., A SNP in the ABCC11 gene is the determinant of human earwax type, Nat. Genet. 38 (2006) 324–330.

[28] R.D. Hernandez, J.L. Kelley, E. Elyashiv, S.C. Melton, A. Auton, G. McVean, 1000 Genomes Project, G. Sella, M. Przeworski, Classic selective sweeps were rare in recent human evolution, Science 331 (2011) 920–924.

[29] J.K. Pritchard, Adaptation—not by sweeps alone, Nat. Rev. Genet. 11 (2010) 920–924.

[30] G. Hellenthal, G.B. Busby, G. Band, J.F. Wilson, C. Capelli, D. Falush, S.A. Myers, Genetic atlas of human admixture history, Science 343 (2009) 747–751. Also: http://paintmychromosomes.com (accessed January 2017).

[31] S. Leslie, B. Winney, G. Hellenthal, D. Davison, A. Boumertit, T. Day, K. Hutnik, E.C. Royrvik, B. Cunliffe, Wellcome Trust Case Control Consortium 2, International Multiple Sclerosis Genetics Consortium, D.J. Lawson, D. Falush, C. Freeman, M. Pirinen, S. Myers, M. Robinson, P. Donnelly, W. Bodmer, The fine-scale genetic structure of the British population, Nature 519 (2015) 309–314.

[32] J.K. Pickrell, D. Reich, Toward a new history and geography of human genes informed by ancient DNA, Trends Genet. 30 (2014) 377–389.

[33] D. Reich, N. Patterson, M. Kircher, F. Delfin, M.R. Nandineni, I. Pugach, A.M. Ko, Y.C. Ko, T.A. Jinam, M.E. Phipps, et al., Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania, Am. J. Hum. Genet. 89 (2011) 516–528.

[34] E. Huerta-Sánchez, X. Jin Asan, Z. Bianba, B.M. Peter, N. Vinckenbosch, Y. Liang, X. Yi, M. He, M. Somel, Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA, Nature 512 (2014) 194–197.

[35] J. Travis, Scientists decry isotope, DNA testing of 'nationality', Science 326 (2009) 30–31.

[36] T. Sanders, Imagining the Dark Continent: the Met, the media and the Thames Torso, Cambridge Anthropol. 23 (2003) 53–66.

[37] H. Wollinsky, Genetic genealogy goes global, EMBO Rep. 7 (2006) 1072–1074.

[38] Sense About Science reports on the validity of genetic genealogy consumer tests at: http://www.senseaboutscience.org/pages/genetic-ancestry-testing.html and: http://www.senseaboutscience.org/data/files/resources/119/Sense-About-Genetic-Ancestry-Testing.pdf (accessed January 2017).

[39] R. Sachidanandam, D. Weissman, S.C. Schmidt, J.M. Kakol, L.D. Stein, G. Marth, S. Sherry, J.C. Mullikin, B.J. Mortimore, et al., International SNP map working group, a map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms, Nature 409 (2001) 928–933.

[40] I. Halder, M. Shriver, M. Thomas, J.R. Fernandez, T. Frudakis, A panel of ancestry informative markers for estimating individual bio-geographicalal ancestry and admixture from four continents: utility and applications, Hum. Mutat. 29 (2008) 648–658.

[41] M.D. Shriver, M.W. Smith, L. Jin, A. Marcini, J.M. Akey, R. Deka, R.E. Ferel, Ethnic-affiliation estimation by use of population-specific DNA, Am. J. Hum. Genet. 60 (1997) 957–964.

[42] T. Frudakis, K. Venkateswarlu, M.J. Thomas, Z. Gaskin, S. Ginjupalli, S. Gunturi, V. Ponnuswamy, S. Natarajan, P.K. Nachimuthu, A classifier for the SNP-based inference of ancestry, J.Forensic Sci. 48 (2003) 771–782.

[43] D.B. Goldstein, A. Ruiz-Linares, L.L. Cavalli-Sforza, M.W. Feldman, An evaluation of genetic distances for use with microsatellite loci, Genetics 92 (1995) 6723–6727.

[44] N.A. Rosenberg, L.M. Li, R. Ward, J.K. Pritchard, Informativeness of genetic markers for inference of ancestry, Am. J. Hum. Genet. 73 (2003) 1402–1422.

[45] H.D. Chen, C.H. Chang, L.C. Hsieh, H.C. Lee, Divergence and Shannon information in genomes, Phys. Rev. Lett. 94 (2005) 178103.

[46] C. Phillips, M. Fondevila, M.V. Lareu, A 34-plex autosomal SNP single base extension assay for ancestry investigations, Methods Mol. Biol. 830 (2012) 109–126.

[47] V. Colonna, L. Pagani, Y. Xue, C. Tyler-Smith, A world in a grain of sand: human history from genetic data, Genome Biol. 12 (2011) 234.

[48] S.A. Tishkoff, F.A. Reed, A. Ranciaro, B.F. Voight, C.C. Babbitt, J.S. Silverman, K. Powell, H.M. Mortensen, J.B. Hirbo, M. Osman, et al., Convergent adaptation of human lactase persistence in Africa and Europe, Nat. Genet. 39 (2007) 31– 40.

[49] P. Taboada-Echalar, V. Álvarez-Iglesias, T. Heinz, L. Vidal-Bralo, A. Gómez- Carballa, L. Catelli, J. Pardo-Seco, A. Pastoriza, Á. Carracedo, A. Torres-Balanza, et al., The genetic legacy of the pre-colonial period in contemporary Bolivians, PLoS One 8 (2013) e58980.

[50] R. Pereira, C. Phillips, N. Pinto, C. Santos, C.E.B. Santos, A. Amorim, Á. Carracedo, L. Gusmão, Straightforward inference of ancestry and admixture proportions through ancestry-informative insertion deletion multiplexing, PLoS One 7 (2012) e29684.

[51] J.M. Galanter, J.C. Fernandez-Lopez, C.R. Gignoux, J. Barnholtz-Sloan, C. Fernandez-Rozadilla, M. Via, A. Hidalgo-Miranda, A.V. Contreras, L. Uribe Figueroa, et al., Development of a panel of genome-wide ancestry informative markers to study admixture throughout the Americas, PLoS Genet. 8 (2012) e1002554.

[52] M.D. Shriver, G.C. Kennedy, E.J. Parra, H.A. Lawson, V. Sonpar, J. Huang, J.M. Akey, K.W. Jones, The genomic distribution of population substructure in four populations using 8525 autosomal SNPs, Hum. Genomics 1 (2004) 274–286.

[53] 1000 Genomes Project Consortium, G.R. Abecasis, A. Auton, L.D. Brooks, M.A. DePristo, R.M. Durbin, R.E. Handsaker, H.M. Kang, G.T. Marth, G.A. McVean, An integrated map of genetic variation from 1092 human genomes, Nature 491 (2012) 56–65.

[54] L. Clarke, X. Zheng-Bradley, R. Smith, E. Kulesha, C. Xiao, I. Toneva, B. Vaughan, D. Preuss, R. Leinonen, M. Shumway, S. Sherry, P. Flicek, 1000 Genomes Project Consortium, The 1000 Genomes Project: data management and community access, Nat. Methods 9 (2012) 459–462.

[55] URL for 1000 Genomes Phase III initial variant data release: http://www.1000genomes.org/announcements/initial-phase-3-variant-list-and-phased-genotypes-2014-06-24 (accessed January 2017).

[56] J. Amigo, A. Salas, C. Phillips, ENGINES: exploring single nucleotide variation in entire human genomes, BMC Bioinf. 12 (2011) 105.

[57] C. Phillips, A. Salas, J.J. Sánchez, M. Fondevila, A. Gómez-Tato, J. Álvarez-Dios, M. Calaza de Cal, D. Ballard, M.V. Lareu, Á. Carracedo, The SNP*for*ID Consortium, inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs, Forensic Sci. Int. Genet. 1 (2007) 273– 280.

[58] M. Fondevila, C. Phillips, C. Santos, A. Freire Aradas, P.M. Vallone, J.M. Butler, M.V. Lareu, Á. Carracedo, Revision of the SNP*for*ID 34-plex forensic ancestry test: assay enhancements, standard reference sample genotypes and extended population studies, Forensic Sci. Int. Genet. 7 (2013) 63–74.

[59] P. Kersbergen, K. van Duijn, A.D. Kloosterman, J.T. den Dunnen, M. Kayser, P. de Knijff, Developing a set of ancestry-sensitive DNA markers reflecting continental origins of humans, BMC Genet. 10 (2009) 69.

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

[61] M.W. Smith, N. Patterson, J.A. Lautenberger, A.L. Truelove, G.J. McDonald, A. Waliszewska, B.D. Kessing, M.J. Malasky, C. Scafe, E. Le, et al., A high-density admixture map for disease gene discovery in African Americans, Am. J. Hum. Genet. 74 (2004) 1001–1013.

[62] N. Yang, H. Li, L.A. Criswell, P.K. Gregersen, M.E. Alarcon-Riquelme, R. Kittles, R. Shigeta, G. Silva, P.I. Patel, J.W. Belmont, M.F. Seldin, 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 (2005) 382–392.

[63] A. Clark, M. Hubisz, C. Bustamante, S. Williamson, R. Nielsen, Ascertainment bias in studies of human genome-wide polymorphism, Genome Res. 15 (2005) 1496–1502.

[64] K.B. Gettings, R. Lai, J.L. Johnson, M.A. Peck, J.A. Hart, H. Gordish-Dressman, M. S. Schanfield, D.S. Podini, A 50-SNP assay for biogeographic ancestry and phenotype prediction in the U.S. population, Forensic Sci. Int. Genet. 8 (2014) 101–108.

[65] U. Daniel, E. Rychlicka, M.V. Derenko, B.A. Malyarchuk, T. Grzybowski, Simple and cost-effective 14-loci SNP assay designed for differentiation of European, East Asian and African samples, Forensic Sci. Int. Genet. 14 (2014) 42–49.

[66] P. Paschou, E. Ziv, E.G. Burchard, S. Choudhry, W. Rodriguez-Cintron, M.W. Mahoney, P. Drineas, PCA-correlated SNPs for structure identification in worldwide human populations, PLoS Genet. 3 (2007) 1672–1686.

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

[68] J.R. Kidd, F.R. Friedlaender, W.C. Speed, A.J. Pakstis, F.M. De La Vega, K.K. Kidd, Analyses of a set of 128 ancestry informative single-nucleotide polymorphisms in a global set of 119 population samples, Invest. Genet. 2 (2011) 1.

[69] C. Phillips, W. Parson, B. Lundsberg, C. Santos, A. Freire-Aradas, M. Torres, M. Eduardoff, C. Børsting, P. Johansen, M. Fondevila, N. Morling, P. Schneider; The EUROFORGEN-NoE Consortium; Á. Carracedo, M.V. Lareu, Building a forensic ancestry panel from the ground up: the EUROFORGEN Global AIM-SNP set, Forensic Sci. Int. Genet. 11 (2014) 13–25.

[70] C.M. Nievergelt, A.X., Maihofer, T., Shekhtman, O., Libiger, X., Wang, K.K., Kidd, J.R. Kidd, Inference of human continental origin and admixture proportions using a highly discriminative ancestry informative 41-SNP panel, Invest. Genet. 4 (2013) 13.

NB: This paper describes 41 of 55 SNPs currently listed in FROGkb: http://frog.med.yale.edu/FrogKB/ (accessed January 2017).

[71] Ion PGM™ system: https://www.lifetechnologies.com/au/en/home/life-science/sequencing/next-generation-sequencing/ion-torrent-next- generation-sequencing-workflow/ion-torrent-next-generation-sequencing-run-sequence/ion-pgm-system-for-next-generation-sequencing.html (accessed January 2017).

[72] Illumina ForenSeq system: http://applications.illumina.com/content/dam/ illumina-marketing/documents/products/appspotlights/ app_spotlight_forensics.pdf (accessed January 2017).

[73] R. Daniel, C. Santos, C. Phillips, M. Fondevila, R.A. van Oorschot, Á. Carracedo, M.V. Lareu, D. McNevin, A SNaPshot of next generation sequencing, Forensic Sci. Int. Genet. 14 (2014) 50–60.

[74] C.D. Harrison, D.J. Ballard, J. Patel, E. Musgrave Brown, C. Phillips, C.R. Thacker, Y.D. Syndercombe Court, the SNP*for*ID Consortium, Differentiating European and South Asian individuals using SNPs and pyrosequencing technology, Forensic Sci. Int. Genet. Supplement Series 1 (2008) 476–478.

[75] J. Costas, A. Salas, C. Phillips, Á. Carracedo, Human genome-wide screen of haplotype-like blocks of reduced diversity, Gene 349 (2005) 219–225.

[76] H. Rajeevan, U. Soundararajan, A.J. Pakstis, K.K. Kidd, Introducing the forensic research/reference on genetics knowledge base, FROG-kb, Invest. Genet. 3 (2012) 18.

[77] H. Rajeevan, U. Soundararajan, J.R. Kidd, A.J. Pakstis, K.K. Kidd, ALFRED: an allele frequency resource for research and teaching, Nucleic Acids Res. 40 (2012) D1010–1015.

[78] J. Amigo, C. Phillips, M. Lareu, Á. Carracedo, The . browser: an online tool for query and display of frequency data from the SNP*for*ID project, Int. J. Legal Med. 122 (2008) 435–440.

[79] C. Santos, C. Phillips, F. Oldoni, J. Amigo, M. Fondevila, R. Pereira, Á. Carracedo, M.V. Lareu, Completion of a worldwide reference panel of samples for an ancestry informative Indel assay, Forensic Sci. Int. Genet. 17 (2015) 75–80.

[80] A.L. Lowe, A. Urquhart, L.A. Foreman, I.W. Evett, Inferring ethnic origin by means of an STR profile, Forensic Sci. Int. 119 (2001) 17–22.

[81] C. Phillips, A. Freire Aradas, A.K. Kriegel, M. Fondevila, O. Bulbul, C. Santos, F. Serrulla Rech, M.D. Perez Carceles, Á. Carracedo, P.M. Schneider, M.V. Lareu, *Eurasiaplex*: A forensic SNP assay for differentiating European and South Asian ancestries, Forensic Sci. Int. Genet. 7 (2013) 359–366.

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

[83] A.L. Price, N. Patterson, R.M. Plenge, M.E. Weinblatt, N.A. Shadick, D. Reich, Principal components analysis corrects for stratification in genome-wide association studies, Nat. Genet. 38 (2006) 904–909.

[84] N. Patterson, A.L. Price, D. Reich, Population structure and eigenanalysis, PLoS Genet. 2 (2006) e190.

[85] M. Kayser, P. de Knijff, Improving human forensics through advances in genetics, genomics and molecular biology, Nat. Rev. Genet. 12 (2011) 179–192.

[86] J. Zhang, P. Niyogi, M.S. McPeek, Laplacian eigenfunctions learn population structure, PLoS One (2009) e7928.

[87] L.L. Cavalli-Sforza, P. Menozzi, A. Piazza, Demic expansions and human evolution, Science 259 (1993) 639–646.

[88] J. Novembre, M. Stephens, Interpreting principal component analyses of spatial population genetic variation, Nat. Genet. 40 (2008) 646–649.

[89] D. Reich, A. Price, N. Patterson, Principal component analysis of genetic data, Nat. Genet. 40 (2008) 491–492.

[90] J. Novembre, T. Johnson, K. Bryc, Z. Kutalik, A.R. Boyko, A. Auton, K.S. Indap, S. King, M.R. Bergmann, M. Nelson, C.D. Bustamante, Genes mirror geography within Europe, Nature 456 (2008) 98–101.

[91] O. Lao, T.T. Lu, M. Nothnagel, O. Junge, S. Freitag-Wolf, A. Caliebe, M. Balascakova, J. Bertranpetit, L.A. Bindoff, D. Comas, et al., Correlation between genetic and geographic structure in Europe, Curr. Biol. 18 (2008) 1241–1248.

[92] D. Falush, M. Stephens, J.K. Pritchard, Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies, Genetics 164 (2003) 1567–1587.

[93] N.A. Rosenberg, DISTRUCT: a program for the graphical display of population structure, Mol. Ecol. Notes 4 (2004) 137–138.

[94] M. Jakobsson, N.A. Rosenberg, CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure, Bioinformatics 23 (2007) 1801–1806.

[95] S.T. Kalinowski, The computer program STRUCTURE does not reliably identify the main genetic clusters within species: simulations and implications for human population structure, Heredity 106 (2011) 625–632.

[96] C.A. McKenzie, R.M. Harding, J.B. Tomlinson, A.J. Ray, K. Wakamatsu, J.L. Rees, Phenotypic expression of melanocortin-1 receptor mutations in Black Jamaicans, J. Invest. Dermatol. 21 (2003) 207–208.

[97] O. Libiger, N.J. Schork, A method for inferring an individual's genetic ancestry and degree of admixture associated with six major continental populations, Front. Genet. 3 (2013) 1–11.

[98] K.W. Broman, J.C. Murray, V.C. Sheffield, R.L. White, J.L. Weber, Comprehensive human genetic maps: Individual and sex-specific variation in recombination, Am. J. Hum. Genet. 63 (1998) 661–889.

[99] N.P. Santos, E.M. Ribeiro-Rodrigues, A.K. Ribeiro-dos-Santos, R. Pereira, L. Gusmão, A. Amorim, J.F. Guerreiro, M.A. Zago, C. Matte, M.H. Hutz, S.E. Santos, Assessing individual interethnic admixture and population substrucure using a 48-insertion-deletion (INDEL) ancestry- informative marker (AIM) panel, Hum. Mutat. 31 (2010) 184–190.

[100] P.A. da Costa Francez, E.M. Ribeiro Rodrigues, A.M. de Velasco, S.E.B. dos Santos, Insertion–deletion polymorphisms—utilization on forensic analysis, Int. J. Legal. Med. 126 (2012) 491–496.

[101] D. Zaumsegel, M.A. Rothschild, P.M. Schneider, A 21 marker insertion deletion polymorphism panel to study bio-geographical ancestry, Forensic Sci. Int. Genet. 7 (2013) 305–312.

[102] R. Pereira, C. Phillips, C. Alves, A. Amorim, Á. Carracedo, L. Gusmão, A new multiplex for human identification using insertion/deletion polymorphisms, Electrophoresis 30 (2009) 3682–3690.

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

[104] C. Phillips, L. Fernandez-Formoso, M. Garciañas, L. Porras, T. Tvedebrink, J. Amigo, M. Fondevila, A. Gomez-Tato, J. Alvarez-Dios, A. Freire-Aradas, et al., Analysis of global variability in 15 established and 5 new European Standard Set (ESS) STRs using the CEPH human genome diversity panel, Forensic Sci. Int. Genet. 5 (2011) 155–169.

[105] L. Pereira, F. Alshamali, R. Andreassen, R. Ballard, W. Chantratita, N.S. Cho, C. Coudray, J.M. Dugoujon, M. Espinoza, F. González-Andrade, et al., PopAffiliator: online calculator for individual affiliation to a major population group based on 17 autosomal short tandem repeat genotype profile, Int. J. Legal Med. 125 (2011) 629–636.

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

[107] C. Phillips, A. Rodriguez, A. Mosquera-Miguel, M. Fondevila, L. Porras- Hurtado, F. Rondon, A. Salas, Á. Carracedo, M.V. Lareu, D9S1120, a simple STR with a common Native American-specific allele: forensic optimization locus characterization and allele frequency studies, Forensic Sci. Int. Genet. 3 (2008) 7–13.

[108] C. Phillips, L. Fernandez-Formoso, M. Gelabert-Besada, M. García-Magariños, C. Santos, M. Fondevila, Á. Carracedo, M.V. Lareu, Development of a novel forensic STR multiplex for ancestry analysis and extended identity testing, Electrophoresis 34 (2013) 1151–1162.

[109] A.J. Pakstis, R. Fang, M.R. Furtado, J.R. Kidd, K.K. Kidd, Mini-haplotypes as lineage informative SNPs and ancestry inference SNPs, Eur. J. Hum. Genet. 20 (2012) 1148–1154.

[110] K.K. Kidd, A.J. Pakstis, W.C. Speed, R. Lagacé, J. Chang, S. Wootton, E. Haigh, J.R. Kidd, Current sequencing technology makes microhaplotypes a powerful new type of genetic marker for forensics, Forensic Sci. Int. Genet. 12 (2014) 215–224.

[111] A.A. Westen, A.S. Matai, J.F.J. Laros, H.C. Meiland, M. Jasper, W.J.F. de Leeuw, P. de Knijff, T. Sijen, Tri-allelic SNP markers enable analysis of mixed and degraded DNA samples, Forensic Sci. Int. Genet. 3 (2009) 233–241.

[112] C. Phillips, Online resources for SNP analysis: a review and route map. Mol. Biotechnol. 35 (2007) 65-97. Review.

[113] C. Phillips, M. Lareu, J. Sanchez, M. Brion, B. Sobrino, N. Morling, P. Schneider, D. Syndercombe Court, Á. Carracedo, Selecting single nucleotide polymorphisms for forensic applications. Progress in Forensic Genetics 10 (2004) 18-20.

[114] C. Phillips, Using online databases for developing SNP markers of forensic interest, Methods Mol. Biol. 297 (2005) 83-106.

[115] J.J. Sanchez, C. Phillips, C. Børsting, K. Balogh, M. Bogus, M. Fondevila, C.D. Harrison, E. Musgrave-Brown, A. Salas, Á. Carracedo, N. Morling, D. Syndercombe-Court, P.M. Schneider, A multiplex assay with 52 single nucleotide polymorphisms for human identification, Electrophoresis 27 (2006A) 1713–1724.

[116] J.J. Sanchez, C. Phillips, C. Børsting, M. Bogus, Á. Carracedo, D. Syndercombe-Court, M. Fondevila, C. Harrison, N. Morling, K. Balogh, P.M. Schneider, Development of a multiplex PCR assay detecting 52 autosomal SNPs. Progress in Forensic Genetics 11, (2006B) 67-70.

[117] A.J. Pakstis, W.C. Speed, R. Fang, F.C. Hyland, M.R. Furtado, J.R. Kidd, K.K. Kidd, SNPs for a universal individual identification panel, Hum. Genet. 127 (2010) 315–324.

[118] C. Phillips, Application of autosomal SNPs and Indels in forensic analysis in: Forensic DNA Analysis: Current Practices and Emerging Technologies. Forensic Sci. Rev. 24 (2013) 43-62.

[119] A. Freire-Aradas, M. Fondevila, A.K. Kriegel, C. Phillips, P. Gill, L. Prieto, P.M. Schneider, Á. Carracedo, M.V. Lareu, A new SNP assay for identification of highly degraded human DNA, Forensic Sci. Int. Genet. 6 (2011) 341-349.

[120] Charles Brenner: The Power of SNPs – Even Without Population Data; http://dna-view.com/SNPpost.htm

[121] M. Eduardoff, C. Santos, M. de la Puente, T.E. Gross, M. Fondevila, C. Strobl, B. Sobrino, D. Ballard, P.M. Schneider, Á. Carracedo, M.V. Lareu, W. Parson, C. Phillips, Inter-laboratory evaluation of SNP-based forensic identification by massively parallel sequencing using the Ion PGM™, Forensic Sci. Int. Genet. 17 (2015) 110–121.

[122] I. Grandell, R. Samara, A.O. Tillmar, A SNP panel for identity and kinship testing using massive parallel sequencing, Int. J. Legal Med. 130 (2016) 905-914.

[123] C. Phillips, D. Ballard, P. Gill, D. Syndercombe Court, Á.Carracedo, M.V. Lareu, The recombination landscape around forensic STRs: Accurate measurement of genetic distances between syntenic STR pairs using HapMap high density SNP data, Forensic Sci. Int. Genet. 6 (2011) 354-365.

[124] M. Fondevila, C. Phillips, C. Santos, R. Pereira, L. Gusmão, Á. Carracedo, J.M. Butler, M.V. Lareu, P.M. Vallone, Forensic performance of two insertion-deletion marker assays, Int. J. Legal Med. 126 (2012) 725-737.

[125] A. Tillmar, C. Phillips, Evaluation of the impact of genetic linkage in forensic identity and relationship testing for expanded DNA marker sets, Forensic Sci. Int. Genet. (2017) 26:58-65.

[126] J. Amigo, A. Salas, C. Phillips, Á. Carracedo, SPSmart: adapting population based SNP genotype databases for fast and comprehensive web access. BMC Bioinformatics 9 (2008) 428-433.

[127] J. Amigo, C. Phillips, A. Salas, Á. Carracedo, Viability of in-house datamarting approaches for population genetics analysis of SNP genotypes. BMC Bioinformatics 10, Suppl. 3 (2009) S5.

[128] J. Amigo, A. Salas, C. Phillips, ENGINES: exploring single nucleotide variation in entire human genomes, BMC Bioinf. 12 (2011) 105.

[129] J. Amigo, C. Phillips, A. Salas, L. Fernandez Formoso, Á. Carracedo, M.V. Lareu, pop.STR - An online population frequency browser for established and new forensic STRs. Forensic Sci. Int. Genet. Supplement Series 2 (2009) 361-362.

[130] W. Parson, D. Ballard, B. Budowle, J.M. Butler, K.B. Gettings, P. Gill, L. Gusmão, D. Hares, J.A. Irwin, J. King, P. de Knijff, N. Morling, M. Prinz, P.M. Schneider , C. Van Neste, S. Willuweit, C. Phillips, Massively Parallel Sequencing of forensic STRs: Considerations of the DNA Commission of the International Society for Forensic Genetics (ISFG) on minimal nomenclature requirements, Forensic Sci. Int. Genet. 22 (2016) 54-63.

[131] C. Phillips, A genomic audit of newly-adopted autosomal STRs for forensic identification, Forensic Sci. Int. Genet. 29 (2017) 193-204.

[132] C. Phillips, W. Parson, J. Amigo, J.L. King, M.D. Coble, C.R. Steffen, P.M. Vallone, K.B. Gettings, J.M. Butler, B. Budowle, D5S2500 is an ambiguously characterized STR: Identification and description of forensic DNA markers in the genomics age, Forensic Sci. Int. Genet. 23 (2016) 19-24.

[133] M. Whittle, More on the genomic identification of forensic STRs, Forensic Sci. Int. Genet. 2016.

[134] M. Bodner, I. Bastisch, J.M. Butler, R. Fimmers, P. Gill, L. Gusmão, N. Morling, C. Phillips, M. Prinz, P.M. Schneider, W. Parson, Recommendations of the DNA Commission of the International Society for Forensic Genetics (ISFG) on quality control of autosomal Short Tandem Repeat allele frequency databasing (STRidER), Forensic Sci. Int. Genet. 24 (2016) 97-102.

[135] C. Phillips, L. Fernandez-Formoso, M. Gelabert, M. García-Magariños, Á. Carracedo, M.V. Lareu, Global population variability in Qiagen Investigator HDplex STRs, Forensic Sci. Int. Genet. 8 (2014) 36-43.

[136] C. Phillips, S. Kind, L. Fernandez-Formoso, M. Gelabert, Á. Carracedo, M.V. Lareu, Global population variability in Promega PowerPlex CS7, D6S1043 and Penta B STRs, Int. J. Legal Med. 127 (2013) 901-906.

[137] S. Zhang, H. Tian, J. Wu, S. Zhao, C. Li, A new multiplex assay of 17 autosomal STRs and amelogenin for forensic application, PLoS One 8 (2013) e57471.

[138] B.F. Zhu, Y.D. Zhang, C.M. Shen, W.A. Du, W.J. Liu, H.T. Meng, H.D. Wang, G. Yang, R. Jin, C.H. Yang, J.W. Yan, X.H. Bie, Developmental validation of the AGCU 21+1 STR kit: a novel multiplex assay for forensic application, Electrophoresis 36 (2015) 271–276.

[139] B. Quintans, V. Alvarez-Iglesias, A. Salas, C. Phillips, M.V. Lareu , Á. Carracedo, Typing of mitochondrial DNA coding region SNPs of forensic and anthropological interest using SNaPshot minisequencing, Forensic Sci. Int. 140 (2004) 251-257.

[140] M. de la Puente, C. Phillips, M. Fondevila, Á. Carracedo, M.V. Lareu, Evaluation of the Qiagen MPS 128-SNP forensic identification multiplex, Forensic Sci. Int. Genet. 28 (2017) 35-43.

[141] P. M. Vallone, A.E. Decker, J.M. Butler, Allele frequencies for 70 autosomal SNP loci with U.S. Caucasian, African-American, and Hispanic samples, Forensic Sci. Int. 149 (2005) 279–286.

[142] N.E.C. Weiler, K. Baca, D. Ballard, F. Balsa, M. Bogus, C. Børsting, F. Brisighelli, J. Červenáková, L. Chaitanya, V. Decroyer, S. Desmyter, K.J. van der Gaag, K. Gettings, C. Haas, J. Heinrich, Maria João Anjos, A. Kal, M. Kayser, K. Kiesler, A. Kúdelová, A. Mosquera, F. Noel, W. Parson, V. Pereira, C. Phillips, P.M. Schneider, D. Syndercombe-Court, M. Turanska, A. Vidaki, P. Woliński, L. Zatkalíková, T. Sijen, A collaborative EDNAP exercise on mtDNA typing via a SNaPshot™ tool or massively parallel sequencing, Forensic Sci. Int. Genet. 26 (2017) 77-84.

[143] C. Lou, B. Cong, S. Li, L. Fu, X. Zhang, T. Feng, S. Su, C. Ma, F. Yu, J. Ye, L. Pei, A SNaPshot assay for genotyping 44 individual identification single nucleotide polymorphisms, Electrophoresis 32 (2011) 368–378.

[144] S. Walsh, F. Liu, K. Ballantyne, M. van Oven, O. Lao, M. Kayser, IrisPlex: a sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information, Forensic Sci. Int. Genet. 5 (2011) 170–180.

[145] Y. Ruiz, C. Phillips, A. Gomez-Tato, J. Alvarez-Dios, M. Casares de Cal, R. Cruz, O. Maroñas, J. Söchtig, M. Fondevila, M.J. Rodriguez-Cid, Á. Carracedo, M.V. Lareu, Further development of forensic eye color predictive tests, Forensic Sci. Int. Genet. 7 (2012) 28-40.

[146] L. Chaitanya, S. Walsh, J.D. Andersen, R. Ansell, K. Ballantyne, D. Ballard, R. Banemann, C.M. Bauer, A.M. Bento, F. Brisighelli, T. Capal, L. Clarisse, T. Groß, C. Haas, P. Hoff-Olsen, C. Hollard, C. Keyser, K.M. Kiesler, P. Kohle, A. Linacre, A. Minawi, N. Morling, H. Nilsson, L. Norén, R. Ottens, W. Parson, V.L. Pascali, C. Phillips, M.J. Porto, A. Sajantila, P.M. Schneider, T. Sijen, J. Söchtig, D. Syndercombe-Court, A. Tilmar, M. Turanska, P.M. Vallone, L. Zatkalíková, A. Zidkova, W. Branicki, M. Kayser, Collaborative EDNAP Exercise on the IrisPlex system for DNA based prediction of human eye colour, Forensic Sci. Int. Genet. 11 (2014) 241-251.

[147] O. Maroñas, C. Phillips, J. Söchtig, A. Gomez-Tato, R. Cruz, J. Alvarez-Dios, M. Casares de Cal, Y. Ruiz, M. Fondevila, Á. Carracedo, M.V. Lareu, Development of a forensic skin colour predictive test, Forensic Sci. Int. Genet. 13 (2014) 34-44.

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

[149] J. Söchtig, C. Phillips, O. Maroñas, A. Gómez-Tato, R. Cruz, J. Alvarez-Dios, M.A. Casares de Cal, Y. Ruiz, K. Reich, M. Fondevila, Á. Carracedo, M.V. Lareu, Exploration of SNP variants affecting hair colour prediction in Europeans, Int. J. Legal Med. 129 (2015) 963-975.

[150] M. Marcińska, E. Pośpiech, J.D. Andersen, M. van den Berge, Á. Carracedo, M. Eduardoff, A. Marczakiewicz-Lustig, N. Morling, T. Sijen, M. Skowron, J. Söchtig, D. Syndercombe-Court, N. Weiler; The EUROFORGEN-NoE Consortium; D. Ballard, C. Børsting, W. Parson, C. Phillips, W. Branicki, Evaluation of variants associated with androgenetic alopecia and their potential to predict male pattern baldness, PLoS One 10 (2015) e0127852.

[151] E. Pośpiech, J. Karłowska-Pik, M. Marcińska, S. Abidi, J. Dyrberg Andersen, M. van den Berge, Á. Carracedo, M. Eduardoff, A. Freire-Aradas, N. Morling, T. Sijen, M. Skowron, J. Söchtig, D. Syndercombe-Court, N. Weiler; The EUROFORGEN-NoE Consortium; P.M. Schneider, D. Ballard, C. Børsting, W. Parson, C. Phillips, W. Branicki, Evaluation of the predictive capacity of DNA variants associated with straight hair in Europeans, Forensic Sci. Int. Genet. 19 (2015) 280-288.

[152] B. Mehta, R. Daniel, C. Phillips, S. Doyle, G. Elvidge, D. McNevin, Massively parallel sequencing of customised forensically informative SNP panels on the MiSeq, Electrophoresis 37 (2016) 2832-2840.

[153] M. Fondevila, C. Børsting, C. Phillips, M. de la Puente, C. Santos; The EuroForGen-NoE Consortium; Á. Carracedo, N. Morling, M.V. Lareu, Forensic SNP genotyping with SNaPshot: Technical considerations for the development and optimisation of multiplexed SNP assays, Forensic Sci. Rev. 29 (2017) 57-76.

[154] B. Mehta, R. Daniel, C. Phillips, D. McNevin, Forensically relevant SNaPshot® assays for human DNA SNP analysis: A Review, Int. J. Legal Med. 131 (2016), 21-37.

[155] C. Phillips, R. Fang, D. Ballard, M. Fondevila, C. Harrison, F. Hyland, E. Musgrave-Brown, C. Proff, E. Ramos-Luis, B. Sobrino, Á. Carracedo, M.R. Furtado, D. Syndercombe Court, P.M. Schneider, The SNP*for*ID Consortium, Evaluation of the Genplex SNP typing system and a 49plex forensic marker panel, Forensic Sci. Int. Genet. 1 (2007) 180-185.

[156] E. Musgrave-Brown, D. Ballard, M. Fondevila, M. Álvarez, R. Fang, C. Harrison, C. Phillips, Y. Prasad, B. Sobrino Rey, C. Thacker , J. Wiluhn, Á. Carracedo, P.M. Schneider, D. Syndercombe Court; The SNP*for*ID Consortium, Forensic validation of the Genplex SNP typing system - Results of an inter-laboratory study. Forensic Sci. Int. Genet. Supplement Series 1 (2008) 389-393.

[157] C. Phillips, M. Fondevila, M.V. Lareu, A 34-plex autosomal SNP single base extension assay for ancestry investigations, Methods Mol. Biol. 830 (2012) 109-126.

[158] C. Phillips, C. Santos, M. Fondevila, Á. Carracedo, M.V. Lareu, Inference of ancestry in forensic analysis I: Autosomal ancestry-Informative marker sets, Methods Mol. Biol. 1420 (2016), W. Goodwin (Ed): Forensic DNA Typing Protocols.

[159] C. Santos, C. Phillips, A. Gomez-Tato, J. Alvarez-Dios, Á. Carracedo, M.V. Lareu, Inference of Ancestry in Forensic Analysis II: Analysis of Genetic Data, Methods Mol. Biol. 1420, (2016) W. Goodwin (Ed): Forensic DNA Typing Protocols.

[160] M. Eduardoff, T.E. Gross, C. Santos, M. de la Puente, D. Ballard, C. Strobl, C. Børsting, N. Morling, L. Fusco, C. Hussing, B. Egyed, L. Souto, J. Uacyisrael, D. Syndercombe Court, P.M. Schneider, Á. Carracedo, M.V. Lareu; The EUROFORGEN-NoE Consortium; W. Parson, C. Phillips, Inter-laboratory evaluation of the EUROFORGEN Global ancestry-informative SNP panel by massively parallel sequencing using the Ion PGM™, Forensic Sci. Int. Genet. 23 (2016) 178–189.

[161] L. Porras-Hurtado, Y. Ruiz, C. Santos, C. Phillips, Á. Carracedo, M.V. Lareu, An overview of STRUCTURE: applications, parameter settings and supporting software, Frontiers Genet. 4 (2013) 98.

[162] C. Santos, C. Phillips, M. Fondevila, R. Daniel, R. A van Oorschot, E.G. Burchard, M.S. Schanfield, L. Souto, J. Uacyisrael, M. Via, Á. Carracedo, M.V. Lareu, Pacifiplex: an ancestry-informative SNP panel centred on Australia and the Pacific region, Forensic Sci. Int. Genet. 20 (2016) 71-80.

[163] C. Borel, F. Cheung, H. Stewart, H.G. Brunner, C. Phillips, P.A. Jacobs, S. Eliez, A.J. Sharp, Evaluation of PRDM9 variation as a risk factor for genomic disorders and chromosomal non-disjunction, Hum. Genet. 131 (2012) 1519-1524.

[164] M. Fondevila, C. Phillips, N. Naveran, L. Fernandez, M. Cerezo, A. Salas, Á. Carracedo, M.V. Lareu, Case report: identification of skeletal remains using short-amplicon marker analysis of severely degraded DNA extracted from a decomposed and charred femur, Forensic Sci. Int. Genet. 2 (2008) 212–218.

[165] M. Fondevila, C. Phillips, N. Naveran, M. Cerezo, A. Rodriguez, R. Calvo, L.M. Fernandez, Á. Carracedo, M.V. Lareu, Challenging DNA: Assessment of a range of genotyping approaches for highly degraded forensic samples. Forensic Sci. Int. Genet. Supplement Series 1 (2008) 26-28.

[166] C. Phillips, M.V. Lareu, A. Salas, Á. Carracedo, Nonbinary single-nucleotide polymorphism markers Progress in Forensic Genetics 10 (2004) 27-29.

[167] E. Musgrave-Brown, N. Anwar, K. Elliott, C. Phillips, D. Syndercombe Court, A. Carracedo, N. Morling, P. Schneider, B. McKeown Mixture analysis using SWaP™ SNPs and non-biallelic SNPs, International Congress Series 1288 (2006) 34–36.

[168] L. Zha, L. Yun, P. Chen, H.L. Luo, J. Yang, Y. Hou, Exploring of tri-allelic SNPs using Pyrosequencing and the SNaPshot methods for forensic application, Electrophoresis 33 (2012) 841-848.

[169] 1000 Genomes Project Consortium, A. Auton, L.D. Brooks, R.M. Durbin, E.P. Garrison, H.M. Kang, J.O. Korbel, J.L. Marchini, S. McCarthy, G.A. McVean, G.R. Abecasis, A global reference for human genetic variation, Nature 526 (2015) 68–74.

[170] C. Phillips, J. Amigo, Á. Carracedo, M.V. Lareu, Tetra-allelic SNPs: Informative forensic markers compiled from public whole-genome sequence data, Forensic Sci. Int. Genet. 19 (2015) 100-106.

[171] C. Phillips, J. Amigo, Á. Carracedo, M.V. Lareu, A catalog of 272,800 human tri-allelic single nucleotide variants from 1000 Genomes project data, Forensic Sci. Int. Genet.; *manuscript in preparation*.

[172] M. Fondevila, L. Manzo, M. de la Puente, C. Phillips, Á. Carracedo, M.V. Lareu, A versatile forensic CE multiplex of multiple-allele SNPs, Forensic Sci. Int. Genet.; *manuscript in preparation*.

[173] C. Romanini, M. Romero, M. Salado Puerto, L. Catelli, C. Phillips, R. Pereira, L. Gusmão, C. Vullo, Ancestry informative markers: inference of ancestry in aged bone samples using an autosomal AIM-Indel multiplex, Forensic Sci. Int. Genet. 16 (2015) 58-63.

[174] C. Romanini, M.L. Catelli, A. Borosky, R. Pereira, M. Romero, M. Salado Puerto, C. Phillips, M. Fondevila, A. Freire, C. Santos, et al., Typing short amplicon binary polymorphisms: supplementary SNP and Indel genetic information in the analysis of highly degraded skeletal remains, Forensic Sci. Int. Genet. 6 (2012) 469–476.

[175] C. Santos, M. Fondevila, D. Ballard, R. Banemann, A.M. Bento, C. Børsting, W. Branicki, F. Brisighelli, M. Burrington, T. Capal, L. Chaitanya, R. Daniel, V. Decroyer, R. England, K.B. Gettings, T.E. Gross, C. Haas, J. Harteveld, P. Hoff-Olsen, A. Hoffmann, M. Kayser, P. Kohler, A. Linacre, M. Mayr-Eduardoff, C. McGovern, N. Morling, G. O'Donnell, W. Parson, V.L. Pascali, M.J. Porto, A. Roseth, P.M. Schneider, T. Sijen, V. Stenzl, D. Syndercombe Court, J.E. Templeton, M. Turanska, P.M. Vallone, R.A. van Oorschot, L. Zatkalikova, Á. Carracedo, C. Phillips; EUROFORGEN-NoE Consortium, Forensic ancestry analysis with two capillary electrophoresis ancestry informative marker (AIM) panels: Results of a collaborative EDNAP exercise, Forensic Sci. Int. Genet. 19 (2015) 56-67.

[176] M. de la Puente, C. Phillips, C. Santos, M. Fondevila, Á. Carracedo, M.V. Lareu, Evaluation of the Qiagen 140-SNP forensic identification multiplex for massively parallel sequencing, Forensic Sci. Int. Genet. 28 (2017) 35-43.

[177] K. Kidd, et al., Evaluating 130 Microhaplotypes across a Global Set of 83 Populations, Forensic Sci. Int. Genet. 2017 (submitted).

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Resumen

La inferencia de la ancestralidad de un individuo a partir del material biológico hallado en la escena del crimen es una técnica instaurada desde hace tiempo en la comunidad forense, pero muy especializada y que a menudo carece del nivel de información adecuado para realizar inferencias fiables. Los ensayos iniciales basados en proteínas polimórficas fueron aplicados con éxito por el autor en investigaciones de casos criminales durante los años 80, pero dichos ensayos fueron abandonados cuando se desarrollaron las metodologías de obtención de perfiles de ADN. Esta tesis describe el desarrollo, la optimización y reintroducción de los ensayos forenses de predicción de ancestralidad a través del genotipado de marcadores autosómicos. Los primeros ensayos de ADN para ancestralidad se basan en los marcadores denominados polimorfismos de nucleótido único (SNPs) y fueron desarrollados por una parte, por la casa comercial DNAprint Genomics y, por la otra, por el autor en Santiago. Los ensayos desarrollados por el autor y basados en SNPs han sido aplicados para la inferencia de ancestralidad en investigaciones criminales y casos de desapariciones durante más de 10 años. Esta tesis describe los pasos fundamentales para desarrollar ensayos de predicción de ancestralidad con fines forenses que puedan ser implementados en todos aquellos laboratorios que dispongan de un secuenciador de electroforesis capilar: la optimización de la PCR tipo multiplex para detectar ADN a partir de muestras limitadas; el proceso de compilación de datos poblacionales a partir de los cuales inferir la ancestralidad más probable del individuo; la detección de patrones de co-ancestralidad en individuos con origen diverso; y el desarrollo de herramientas estadísticas online que permitan inferir el origen probable de un individuo a partir de un perfil de SNPs. Utilizando la infraestructura ampliamente establecida que se utiliza para obtener perfiles de ADN mediante sistemas de electroforesis capilar, se establecieron ensayos de otros tipos de marcadores autosómicos adicionales, tales como: polimorfismos de Inserción-Deleción (Indels); microsatélites (STRs) y SNPs multialélicos. Finalmente, las tecnologías de secuenciación masiva en paralelo permiten el desarrollo de conjuntos de marcadores de ancestralidad más extensos, beneficiándose de las capacidades de dichas plataformas: el aumento de la capacidad de multiplex y la posibilidad de conocer la fase en la que se encuentran los SNPs en cada una de las cadenas, lo que ha permitido introducir los microhaplotipos como nuevos loci informativos de ancestralidad. Esta tesis describe como los microhaplotipos han sido reducidos en tamaño sistemáticamente con el fin mejorar la sensibilidad forense y como han sido incluidos en los ensayos de ancestralidad para tecnologías de secuenciación masiva en paralelo.

Los estudios presentados en esta tesis responde a un conjunto de diez objetivos bien definidos, alcanzados con éxito en resultados y reflejados en artículos publicados:

1. Recopilación de datos de frecuencias alélicas de marcadores SNP a partir de bases de datos en línea para una serie de poblaciones mundiales, y adquisición de dichos datos a partir de bases de datos de SNPs cada vez más completas. Los datos del primer mapa genético humano detallado, de 1,42

millones de SNPs, proporcionaron el principal impulso a estos esfuerzos. Los estudios incluidos en esta tesis presentan navegadores autónomos 'SPS' alojados en la USC que recopilan frecuencias alélicas de los SNPs a partir de datos de frecuencias alélicas a partir de HapMap, Perlegen, CEPH Foundation y 1000 Genomas y permiten a los usuarios descargar los datos genotípicos completos para un gran número de SNPs y poblaciones. Todos estos objetivos se lograron en los navegadores SPSmart, que alcanzan una alta tasa de éxito y permiten realizar consultas de múltiples SNP, a diferencia de cualquier otro navegador de gran escala del genoma humano. Todos los SNP elegidos para las pruebas forenses fueron seleccionados utilizando los navegadores SPSmart.

2. Evaluación de la principal tecnología forense para el genotipado de SNPs mediante la metodología de extensión del cebador denominada SNaPshot, disponible desde 2001. Utilizando este sistema basado en electroforesis capilar (CE) se desarrollaron nuevos paneles de SNPs específicos para uso forense, que comprenden ensayos de: SNPs informativos de ancestralidad (AIM-SNPs); SNPs de identificación individual SNPs (ID-SNPs); SNPs de regiones codificantes para la predicción de características externas visibles (EVC-SNPs) de variación común, tales como rasgos de pigmentación; variantes del ADN mitocondrial (mt-SNPs) y SNPs de cromosoma Y (Y-SNPs). Las diferentes clases de SNPs fueron incorporados con éxito en ensayos tipo SNaPshot y, en particular el ensayo de 34 AIM-SNPs ha permanecido como el principal panel de predicción de ancestralidad con fines forenses durante más de 10 años, y los 52 ID-SNPs recogidos en otro ensayo se aplican universalmente como marcadores de elección en casos que requieren el análisis de ADN altamente degradado mediante amplicones cortos. Además, estos SNPs de identificación desarrollados por el autor se han utilizado en casi todos los paneles de identificación forense a gran escala desde 2005.

3. A partir del panel 34-plex de predicción de ancestralidad mencionado anteriormente se desarrolló una solución integral para el análisis forense de ADN que permite inferir el origen ancestral más probable del donante de la muestra. Este aspecto de la investigación de la tesis implicó el desarrollo de un ensayo que combina reacciones de tipo PCR multiplex y SNaPshot; la optimización de herramientas de acceso abierto para el análisis del genotipado utilizando cálculos de probabilidad de Bayes (disponibles en el portal en línea de acceso abierto denominado Snipper) y la búsqueda exhaustiva de datos genotípicos poblacionales para los SNPs del ensayo (recopilados finalmente en de la página dedicada a SNPforID de SPSmart). Todo el proceso se llevó a cabo satisfactoriamente y la metodología representa actualmente el sistema de elección para el análisis de ascestralidad forense en todo el mundo.

4. Una vez que se estableció el ensayo 34-plex para la diferenciación de donantes de trazas de contacto de África, Europa y Asia Oriental, fue necesario mejorar el ensayo desarrollando nuevos conjuntos de SNPs diseñados para su análisis conjunto con los SNPs de 34-plex. Estos ensayos adicionales de SNaPshot se centran específicamente en poblaciones de Asia Meridional, Nativa Americana y Oceanía (denominadas Eurasiaplex, PIMA y Pacifiplex, respectivamente). Además de los

ensayos para la diferenciación específica de subpoblaciones, se consideró importante desarrollar también conjuntos adicionales centrados en mejorar el balance de picos en SNaPshot (utilizando únicamente SNPs CT); en los patrones de coancestralidad europea-africana de individuos mezclados típicos de los perfiles demográficos estadounidenses, británicos y franceses (Admixplex) y en los AIM-SNPs de cromosoma X, que proporcionan una visión particular de la ascendencia de linaje materno cuando en individuo con ancestralidad mezclada. El desarrollo de todos estos paneles individuales de genotipado de SNPs se realizó utilizando la tecnología SNaPshot y el consecuente conocimiento del poder de diferenciación de cada SNP para cada una de las poblaciónes y de sus patrones de variación mundial fue clave en la posterior selección de conjuntos mas amplios de marcadores diseñados para extender la predicción de ancestralidad biogeográfica a las nuevas tecnologías de secuenciación masiva en paralelo, que permiten el análisis de muchos más loci en cada reacción de amplificación y secuenciación.

5. Una vez desarrollado un conjunto de ensayos optimizados para la predicción de ancestralidad biogeográfica con fines forenses, era importante evaluar la sensibilidad de dichos ensayos, examinando el rendimiento de genotipado de la PCR tipo multiplex cuando se aplicaba a una amplia gama de pruebas de identificación de ADN extraído de material esquelético y aplicar estos ensayos de SNPs al análisis de trazas de contacto encontradas en la escena del crimen, con el fin de proporcionar información que pueda ser de ayuda para las pesquisas policiales durante investigaciones criminales.

Una primera muestra de la alta sensibilidad que cabe esperar mediante el genotipado forense de SNPs con SNaPshot fue el análisis de material esquelético quemado recuperado en el suelo de un bosque de Ourense, en el sur de Galicia en 2006, después de un incendio importante en la zona. La aplicación tanto de conjuntos de STRs de amplicón de longitud reducida como de multiplexes de SNPs permitió de evaluar la capacidad de cada método para genotipar con éxito material altamente degradado proveniente de un caso complicado, ya que el origen más probable del esqueleto era el de una persona que llevaba desaparecida más de diez años. Así, en este caso el ADN no sólo estaba sometido a putrefacción, sino también expuesto a las altas temperaturas típicas de los incendios forestales. Este caso representa, por lo tanto, el tipo más degradado de ADN diana posible. Los resultados indicaron que sólo los STRs miniaturizados dieron resultados para microsatélites a partir del ADN extraído del fémur del esqueleto, con algunos picos extra y muchos loci fallando. Sin embargo, los SNP produjeron perfiles completos. El perfil de los AIM SNPs de 34-plex fue también completo y produjo una probabilidad de ser europeo de 164 billones de veces más probable que africano y 44 mil millones de veces más probable que este asiático. Esta primera prueba de la sensibilidad del análisis SNP forense fue corroborada por el análisis exitoso de siete ADNs provenientes de trazas de contacto durante la investigación de la bomba de Madrid del 11-M, con extractos en cantidades de 0,07 - 0,11 - 0,19 - 0,3 - 2,0 - 3,29 - 12,7 ng /ul, por debajo de las cantidades óptimas en la mayoría de ellos. Todos los extractos produjeron perfiles completos para los AIM SNPs de 34-plex, lo que indica

que el ADN de baja cantidad puede proporcionar suficiente diana para una amplificación exitosa si no hay degradación.

El autor comenzó a promover el concepto de la inferencia de ancestralidad en un contexto forense para avanzar en investigaciones criminales que no presentan coincidencias en las base de datos de perfiles de ADN o que no cuentan con testimonios de testigos oculares. Finalmente, el laboratorio donde se desarrolló el ensayo inició el uso de pruebas de escestralidad para revisar casos no resueltos, con exitoso resultado en la resolución en 2015 del caso de asesinato de Eva Blanco, sin resolver desde 1997.

6. Para desarrollar los modelos estadísticos para la inferencia de ancestralidad, era importante identificar y recopilar datos poblacionales de las variantes, ya que el análisis de Bayes calcula y compara una serie de probabilidades de que un donante de la muestra proceda de uno de varios posibles orígenes ancestrales. Se hace la suposición matemática de que las frecuencias de las variantes observadas en la muestra son directamente traducibles en probabilidad, es decir, la presencia en la muestra de una variante rara en la población 1 que es mucho más común en la población 2 corresponde a una mayor probabilidad de que el donante de la muestra viene de la población 2. Las poblaciones se definen ampliamente en función de su distribución continental (cinco regiones: África, Europa, Asia Oriental, Nativo de América o Oceanía). Sin embargo, para conjuntos de marcadores a gran escala, se puede distinguir adicionalmente entre la probabilidad de pertenecer al sur de Asia (subcontinente indio), más la probabilidad de pertenecer a una población del Medio Oriente-África del Norte, diferenciadas de probabilidad de pertenecer a África Subsahariana, Europa y Asia Meridional que ocupan regiones vecinas y, por lo tanto, muestran una relación más estrecha que el resto de grupos de población definidos en función de su distribución continental.

Además de los SNPs binarios de uso común, que muestran dos alelos y que constituyen los componentes principales de los ensayos de amplicón corto de predicción de ancestralidad con fines forenses, fue un paso productivo compilar los detalles genómicos de SNP no binarios: consistentes variaciones tri-alélicas y tetra- alélicas de un solo nucleótido que tienen múltiples sustituciones de bases detectadas en los genomas secuenciados y que pueden ser adaptados para secuenciación masiva en paralelo. Los primeros descubrimientos de SNPs trialélicos en la USC promovieron el esfuerzo de recopilar y validar cualquiera de los marcadores que salieran a la luz que mostraran una variabilidad alélica informativa en los tres alelos. Esta tarea presenta más dificultades para los SNPs trialélicos que para los SNPs bialélicos comunmente utilizados en los ensayos de ancestralidad, ya que las variaciones tri-alélicas no pueden ser detectadas utilizando el sistema de doble tinte (Cy3-Cy5) adoptado para todos los arrays genómicos de SNPs cuyo uso se ha generalizado desde 2005. Esto impidió el análisis poblacional de estas variaciones para panel HGDP-CEPH, cuyos datos han sido generados mediante el array Illumina 650.000 SNP, y que había ayudado a identificar tantos SNP binarios informativos para ancestralidad. El proyecto Hapmap no identifica ningún SNP trialélico ya

que se basó en el uso de la tecnología de arrays para identificar la variación de los SNPs en las poblaciones de estudio elegidas. En una fase tardía de la tesis fue posible desarrollar un panel de SNPs de múltiples alelos para la secuenciación masiva paralelo que comprende 250-300 marcadores obtenidos mediante el escrutinio completo del catálogo de variantes humanas publicado por 1000 genomas a finales de 2014. Este catálogo fue la base para la selección de marcadores para un panel más grande de 1400 ID-SNPs que tenían tres o cuatro alelos comunes en cada sitio polimórfico, y por lo tanto, permitió niveles muy altos de diferenciación individual, incluso cuando los análisis de parentesco incompletos se hacen a través de múltiples generaciones - como es típico en muchas de las identificaciones de personas desaparecidas que se realizan en las principales regiones de conflicto o después de catástrofes masivas. Simulaciones de combinaciones genotípicas para diferentes pares relacionados: padre / hijo; hermanos completos; medio hermanos; primos hermanos; demuestran claramente el beneficio de utilizar SNPs trialélicos, que pueden mostrar seis genotipos diferentes en situaciones de prueba de parentesco. Además, los SNPs trialélicos desarrollados para el análisis forense revelan el beneficio adicional de los SNPs multialélicos, que pueden proporcionar indicaciones de mezclas de ADN a partir de la detección de tres alelos diferentes en cualquier posición de SNP en un perfil mixto.

Pese a que los SNPs trialélicos fueron ignorados por la mayoría de las tecnologías de genotipado de SNPs, el autor fue capaz de utilizarlos con éxito para cada uno de los principales paneles de predicción de ancestralidad genotipados mediante SNaPshot (específicamente, 34-plex, PIMA, Pacifiplex SNP) y mediante secuenciación masiva en paralelo (en todos los paneles de SNPs de ancestralidad forense desarrollados hasta el momento). Su contribución ha aumentado gradualmente a medida que se obtiene más conocimiento acerca de la distribución de la variabilidad de los SNP trialélicos en poblaciones mundiales, ya que 1000 genomas han reintroducido estos loci en sus bases de datos de variantes genómicas.

7. Además de los SNPs, los estudios de tesis tuvieron como objetivo optimizar y extender los análisis forenses de la ancestralidad haciendo uso de Indels y STRs, permitiendo establecer enfoques de marcadores mixtos que permitan la inferencia de ascendencia a partir de datos de perfiles de ADN estándar, cuando el material probatorio es ya no está disponible para pruebas de ADN adicionales con marcadores especializados que no se utilizan en los perfiles de rutina o cuando se requiere un sistema más adecuado para el análisis de mezclas de ADN que el genotipado SNP basado en SNaPshot, donde las relaciones de alturas de los picos pueden ser demasiado variables para distinguir con seguridad un par de picos un marcador heterocigoto de una mezcla de homocigotos a proporciones casi iguales. Dado que los Indels difieren en tamaño por definición, se adaptan fácilmente a una PCR con cebadores marcados de manera idéntica a los STRs, por lo que pueden proporcionar una forma muy segura de identificar y analizar los componentes de una mezcla de ADN, a la vez que mantienen el beneficio de evitar los largos amplicones necesarios para los STRs. Las características del polimorfismo binario de los Indels los hacen menos informativos por locus para las identificaciones forenses, pero

este factor no es tan relevante en los análisis de ancestralidad, donde una gran proporción de Indels muestra una diferenciación poblacional de frecuencias alélicas bastante similar a muchos AIM-SNPs. De hecho, la recopilación y el uso de AIM-Indels ha sido uno de los avances más exitosos en el análisis forense de ancestralidad, iniciado en Santiago en los últimos cinco años y que se adopta cada vez en una escala más amplia debido a las características de los perfiles de picos equilibrados de los Indels, y la capacidad de detectar mezclas de ADN más fácilmente que únicamente con el uso de SNPs. Además de desarrollar ensayos de genotipado para Indels y STRs, fue importante desarrollar un clasificador para Snipper, basado en frecuencias y que pudiera ser aplicable a los datos forenses de STRs, pero también capaz de extender el alcance de la inferencia de la ancestralidad a haplotipos, como los que se generan en el genopotipado de Y-SNPs y microhaplotipos de SNPs autosómicos, donde todos los componentes están ligados y deben ser tratados como haplotipos para estimar sus frecuencias. Por último, una mejora adicional importante fue ampliar la funcionalidad de Snipper con datos de conjuntos de entrenamiento prefijados para los 46 Indels además de con los ya establecidos 34 SNPs, y agregar los datos forenses del panel de Indels a SPSmart en el navegador dedicado forInDel. En particular, el establecimiento de Indels ha promovido varias iniciativas que los han convertido en el sistema de elección en los laboratorios forenses debido a que no proporcionan resultados ambiguos cuando las mezclas de ADN son analizadas sin saberlo. La conclusión exitosa de un ejercicio de colaboración del Grupo Europeo de Perfiles de ADN (EDNAP), en el que todos los participantes pudieron genotipar con precisión todos los 46 Indels y asignar la ancestralidad de los ADN de control enviados, indicaron la robustez y fiabilidad de los ensayos forenses de Indels. Todos los participantes también fueron capaces de detectar con éxito una muestra enviada que contenía una mezcla de ADN e incluso de nferir con precisión la ancestralidad de los individuos componentes.

8. A medida que se pusieron a disposición los sistemas de secuenciación masiva en paralelo para el análisis forense, los estudios de esta tesis se enfocaron en la reconstrucción de los pequeños conjuntos de AIMs para electroforesis capilar descritos anteriormente en paneles individuales ampliados de 130 a 160 SNPs. Los paneles de SNP resultantes se evaluaron en términos de equilibrio de secuencias en heterozigotos (el grado en que las lecturas de las secuencias de cada alelo del SNP estaba próxima a una relación 50:50 perfectamente equilibrada), sensibilidad al analizar ADN comprometido de muestras degradadas o en diluciones seriadas y concordancia del genotipado comparando los genotipos obtenidos para las muestras con datos en línea secuenciados con técnicas alternativas.

9. Como la tecnología de secuenciación masiva en paralelo se aplicó en el primer lugar al genotipado de STRs, fue necesario identificar y catalogar las variantes flanqueantes encontradas en las mismas hebras de secuencia que los loci objetivo y ligadas a los principales marcadores STRs y SNP genotipados por los nuevos ensayos forenses de secuenciación masiva en paralelo. El formato de salida de los sistemas de Illumina y Thermo Fisher se limita a la secuencia de la región repetitiva y, cuando un sitio SNP está muy cerca, o la secuencia flanqueante es relevante para el genotipado del STR basado en el conteo de las unidades de repetición, se informa de secuencia adicional. Los

estudios de tesis analizaron el potencial informativo para la predicción de ancestralidad de las variantes flanqueantes, comenzando con el panel HGDP-CEPH genotipado por el sistema Forenseq basado en Illumina MiSeq. Otros estudios han comenzado a revelar que las variantes flanqueantes están de hecho fuertemente ligadas con un alelo de repetición particular del STR, lo que significa que la cadena de secuencia que contiene la región de repetición y el SNP de la región flanqueante constituye un haplotipo y debe ser tratada así a la hora de obtener datos de frecuencias, en lugar de combinar las frecuencias de alelos individuales en el STR y en los SNPs. La visión predominante de estos estudios y el análisis posterior de los conjuntos de SNPs genotipados en la misma cadena de secuencia indica que casi todos los loci forenses analizados por secuenciación masiva en paralelo son microhaplotipos.

10. Las iniciativas anteriores llevaron a construir específicamente conjuntos de loci microhaplotipos, que comprenden conjuntos de alelos de SNP estrechamente vinculados en combinaciones de haplotipos, que muestran patrones de variación fuertemente diferenciados en las poblaciones. En todos los ensayos de ancestralidad forense hasta ahora desarrollados en Santiago para secuenciación masiva en paralelo, se han adoptado paneles de loci de microhaplotipos informativos para ancestralidad o aplicables a identificación (por ejemplo, identificación de desaparecidos) rediseñando las longitudes de secuencia para amplificar los microhaplotipos en longitudes de fragmento muy cortas.

Permanecen dos nuevos desafíos: encontrar más microhaplotipos en el genoma y diseñar una manera sencilla de identificar los loci más efectivos para el análisis de ancestralidad a partir de los datos que se compilan en búsquedas genómicas a gran escala. Identificar los microhaplotipos más informativos para fines de identificación forense sólo requiere el cálculo de valores de diversidad genética (una variante de la métrica heterocigosidad aplicada a múltiples polimorfismos ligados, como Y-STR). Este paso es sencillo a partir de los datos de frecuencias poblacionales ya proporcionados en 1000 genomas. Existe la necesidad de convertir la fase de los genotipos en combinaciones de haplotipos; por ejemplo: tres SNPs genotipados como A | G, G | G, T | C se registran como haplotipos AGT y GGC, que se distinguen de las muestras que tienen los mismos genotipos, pero son en realidad combinaciones de hebras AGC, GGT - que se cuentan por separado al evaluar la variabilidad total del microhaplotipo. El problema de que existan loci con similares valores de diversidad genética en diferentes poblaciones, pero muy diferentes distribuciones de haplotipos frecuencias, existe tanto en los microhaplotipos como en los SNPs multialélicos. Por lo tanto, la construcción de métricas apropiadas para identificar los microhaplotipos más informativos para ancestralidad será un importante paso en la compilación de conjuntos mucho mayores de loci candidatos a partir de los cuales escoger los mejores marcadores o como forma de maximizar el potencial para diferenciar poblaciones relacionadas más estrechamente

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Abstract

The inference of a person's bio-geographical ancestry from the biological material they leave behind at a crime-scene has been a long-standing but specialised forensic technique, which often lacks sufficient genetic detail to make a reliable inference of the their ancestry. Initial tests that used polymorphic proteins, were successfully applied by the author to a range of criminal investigations in the early eighties, but when DNA profiling was developed in this decade, such tests were abandoned in favour of accomplishing more precise and sensitive forensic identification tests from DNA profiling of genetic loci that did not show strong contrasts in variation between population groups. This thesis describes the development, optimisation and successful re-introduction of forensic ancestry analysis tests that genotype autosomal genetic markers and therefore compliment many of the mitochondrial and Y-chromosome markers used for the analysis of hairs and male-specific DNA, respectively. The first dedicated DNA-based forensic ancestry tests used single nucleotide polymorphisms (SNPs) and were developed by DNAprint Genomics in Florida, an independent company that offered the tests to police as a commercial service, and at almost the same time at Santiago by the author. The autosomal SNP tests developed by the author have been used for more than twelve years to successfully ascertain the ancestry of individuals who are suspects in criminal investigations and in tests aimed at identifying the remains of missing persons. This thesis describes in detail the key steps achieved in developing a forensic ancestry test that can be, and has been, successfully adopted by any forensic laboratory that routinely uses capillary electrophoresis equipment, which involved: optimisation of a robust and sensitive PCR multiplex to detect the DNA markers from contact traces, often involving low level DNA quantities; compilation of population data from which to infer the likely population of origin of the person with statistical analyses; detection of co-ancestry patterns in an individual with admixed backgrounds; and development of online statistical tools that calculate the probability of an individual's ancestry from a submitted SNP profile. Using the same well-established DNA profiling infrastructure of optimised capillary electrophoresis systems, additional types of autosomal markers were compiled from Insertion-Deletion polymorphisms (Indels); short tandem repeats (STRs) and multiple-allele SNPs (tri-allelic and tetra-allelic SNPs). Finally, the greatly expanded PCR multiplexes of ancestry markers have been developed for massively parallel sequencing (alternatively known as next generation sequencing), which exploit both the increased multiplexing capacity of this technology and the ability to know the phase (the combination of individual alleles and their order on the strand), of SNPs found together on a sequence stand, which has enabled Microhaplotypes to be added as new ancestry informative loci in their own right. The thesis describes how Microhaplotypes have been systematically reduced in size to improve their forensic sensitivity when analysing very degraded DNA and introduced into ancestry tests using massively parallel sequencing technology.

The thesis studies had a set of ten well-defined objectives that were all successfully met by the study results and accompanying published papers:

1. Compilation of SNP allele frequency data from online sources for a range of worldwide populations, and to curate this data from the growing depth and coverage of such SNP databases. Data from the first detailed human map of 1.42 million SNPs provided the main impetus to these efforts. The thesis studies then built stand-alone 'SPS' allele frequency browsers hosted at USC for HapMap, Perlegen, CEPH Foundation and 1000 Genomes SNP data that allow end-users to download full genotype data for large numbers of SNPs and populations. All these objectives were achieved in the SPSmart browsers that enjoy a high hit-rate and allow multiple SNP queries to be made, unlike any other large-scale human genome variant browser. All SNPs chosen for forensic tests were selected using the SPSmart browsers.

2. Evaluation of the key forensic technology for SNP genotyping of SNaPshot primer extension, from its initial availability in 2001. Using this capillary-electrophoresis (CE) based system, to develop novel SNP panels specifically for forensic use, that comprised: ancestry-informative marker SNPs (AIM-SNPs); individual identification SNPs (ID-SNPs); coding-region SNPs for the prediction of common-variation externally visible characteristics (EVC-SNPs), such as pigmentation traits; mitochondrial DNA variants (mt-SNPs); and Y-chromosome SNPs (Y-SNPs). All types of SNPs were successfully incorporated into SNaPshot assays and the combination of 34 AIM-SNPs in particular has endured as the principal forensic ancestry panel for more than 10 years, and a set of 52 ID-SNPs that are universally applied as markers of choice for short-amplicon analysis of highly degraded DNA. The author's choice of identification SNPs have been used in almost all large-scale forensic identification panels since 2005.

3. Using the above 34-plex ancestry panel, to develop a complete solution to forensic analysis of DNA ignorer to infer the most likely ancestral origin of a sample's donor. This aspect of the thesis research involved development of a highly multiplexed PCR and SNaPshot assay; optimising open-access genotype analysis tools using Bayes likelihood calculations (available in the Snipper open-access online portal); accompanied by comprehensive population variation surveys for the SNPs of the test (complex within the dedicated SNPforID SPSmart pages). All these steps were accomplished and have become the system of choice for forensic ancestry analysis worldwide.

4. Once the 34-plex assay was established for the differentiation of African, European and East Asian unadmixed contact trace donors, it was necessary to enhance this SNP assay by developing novel SNP sets designed to be run alongside the core 34 SNPs. These additional SNaPshot assays specifically focus on South Asian, Native American and Oceanian populations-of-origin (named Eurasiaplex, PIMA and Pacifiplex respectively). As well as assays for specific sub-population differentiation, it was considered important to also develop additional sets focused on improved SNaPshot peak balance (CT-only SNPs); European-African co-ancestry patterns in admixed individuals typical of US, UK and French demographic profiles (Admixplex) and X-chromosome AIM-SNPs that provide particular insights into the maternal lineage ancestry when an individual has admixed background. Development of all these individual additional SNP genotyping assays were completed using SNaPshot technology

and the consequent knowledge of the population-differentiation power and world-wide variation patterns of each SNP was used to choose a broader range of markers for extended large-scale ancestry assays for massively parallel sequencing technology that can genotype many more loci in one amplification and sequencing test.

5. With an optimised set of autosomal SNP ancestry tests well established for forensic use, it was important to then evaluate the sensitivity of such forensic SNP tests by assessing the genotyping performance of the PCR multiplex when applied to a wide range of identification tests on DNA extracted from skeletal material, and to apply these SNP test assessments to crime-scene contact trace analysis in order to provide investigative leads available to police in a range of criminal investigations.

An early proof of concept for the expected sensitivity of forensic SNP genotyping with SNaPshot, was the analysis of badly burnt skeletal material recovered from the floor of a forest in Ourense, South Galicia in 2006, following a major fire in the area. By applying reduced-length amplicon STR sets and SNP multiplexes, there was an opportunity to assess the ability of each approach to successfully type highly degraded material from a very challenging case, since the most likely origin of the skeleton was from a person missing more than ten years previously. Therefore, the DNA was not only subject to putrefaction, but also exposed to the very high temperatures typical of forest fires. This case therefore represents the most degraded type of target DNA possible. The results indicated that only miniaturised STRs gave results for micro-satellites from DNA extracted from the femur of the skeleton, with some extra peaks and many of these loci failing. However, SNPs produced full profiles. The 34-plex AIM SNP profile was also complete and gave a likelihood to be European of 164 Billion times more likely than African and 44 billion times more likely than East Asian. This first proof of forensic SNP analysis sensitivity was followed by the successful analysis of seven contact trace DNAs from the 11-M Madrid bomb investigation, with extracts giving quantities of 0.07 - 0.11 - 0.19 - 0.3 - 2.0 - 3.29 - 12.7 ng/ul, below optimum amounts in the majority of DNAs. All extracts gave full 34-plex AIM SNP profiles and indicates that low level DNA can provide sufficient target for successful amplification if there is no degradation.

The author began promoting the concept of ancestry inference in a forensic context to progress criminal investigations lacking database hits or eye-witness testimony. Finally, the laboratory where the test had been developed started to advocate use of ancestry tests for cold-case reviews, with the successful resolution in 2015 of the unsolved Eva Blanco murder case of 1997.

6. In order to develop the statistical models for ancestry inference, it was important to identify and collect population variation data as Bayes analysis calculates and compares a series of likelihoods that a sample donor comes from one of several possible ancestral origins. It makes the mathematical assumption that variant frequencies for those observed in the sample are directly translatable into likelihoods, i.e. the presence in the sample of a rare variant in population 1 that is much more common

in population 2 corresponds to a higher likelihood that the sample donor comes from population 2. The populations are broadly defined by continental distribution (five regions: Africa, Europe, East Asia, Native to America or Oceania). However, for large-scale marker sets, a further distinction between the likelihood to be from South Asia (Indian sub- continent), plus the likelihood to be from a Middle East-North African population, differentiated from those of sub-Saharan Africa, Europe and South Asia that occupy neighboring regions and therefore show a closer relationship than the other continentally-based population groups.

As well as the commonly used binary SNPs, showing two alleles and that form the main components of short-amplicon forensic ancestry tests, it was a productive step to compile the genomic details of non-binary SNPs: consisting of tri-allelic and tetra-allelic single nucleotide variation that have multiple base substitutions detected by whole genome re-sequencing, which can then be adapted for massively parallel sequencing. The initial chance discovery of tri-allelic SNPs at USC prompted further efforts to collect and validate any such markers that came to light, that showed informative allelic variability in the three alleles. This was less easy to achieve than the recognition of the best bi-allelic ancestry-informative SNPs, as tri-allelic variation is not detected using the dual-dye (Cy3-Cy5) system adopted for whole-genome SNP arrays that have been in widespread use since 2005. This precluded the population analysis of HGDP-CEPH variation generated by the Illumina 650,000 SNP array, which had helped identify so many binary ancestry-informative SNPs. The Hapmap project did not list any tri-allelic SNPs as it was based on use of array technology to identify SNP variation in the study populations chosen. At a late stage of the thesis it was possible to develop a panel of multiple allele SNPs for massively parallel sequencing of 250-300 markers obtained from screening the complete human variant catalog published by 1000 Genomes in late 2014. This was also the basis for an even larger panel of 1400 ID-SNPs that had three or four common alleles at each polymorphic site, and therefore enabled very high levels of individual differentiation, even when incomplete kinship analyses are made across multiple generations – as is typical of many missing person identifications undertaken in major regions of conflict or following mass disasters. Simulations of genotype combinations in different related pairs: parent/child; full sibs; half sibs; first cousins; demonstrate a clear benefit from using tri-allelic SNPs that can analyse six genotypes in relationship test scenarios. Furthermore, tri-allelic SNPs developed for forensic analysis reveal the additional benefit of multiple-allele SNPs to provide indications of DNA mixtures from detection of three different alleles at any one SNP position in a mixed profile.

Despite tri-allelic SNPs being ignored by most SNP genotyping technologies, the author was able to successfully utilise them for each of the key ancestry panels for SNaPshot genotyping (specifically, 34-plex, PIMA, Pacifiplex SNP sets) and for later massively parallel sequencing forensic assays (in all forensic ancestry-SNP panels developed so far). Their contribution has gradually increased as more knowledge is gained about the distribution of variability of tri-allelic SNPs in worldwide populations, as 1000 Genomes has now re-introduced these loci into their genomic variant databases.

7. As well as SNPs, the thesis studies aimed to optimise and extend the practicality of forensic ancestry analysis by making use of Indels and STRs in order to establish mixed-marker approaches that allow ancestry inference from standard DNA profiling data, when evidential material is no longer available for additional DNA tests with specialist markers not used for routine profiling, or when a more secure system is required for the analysis of mixed DNA than is possible with SNaPshot-based SNP genotyping, where peak height ratios can be too variable to reliably distinguish a heterozygote peak pair from a mixture of homozygotes at near to equal ratios. As Indels differ in size by definition, they are easily adapted to dye-linked PCR in identical fashion to STRs, so they can provide a very secure way to identify and analyse mixed DNA components, but keep the benefit of avoiding the long amplicons necessary for most STRs. Indel's binary polymorphism characteristics make them less informative per locus for forensic identification, but this factor is not so relevant for ancestry analysis where a large proportion of Indels show a population differentiation of allele frequencies quite similar to many AIM-SNPs. In fact, compilation and use of AIM-Indels has been one of the most successful developments in forensic ancestry analysis to come from Santiago in the last five years and is increasingly adopted on a wider scale because of Indel marker's balanced peak profile characteristics, assay robustness and ability to detect mixed DNA more easily than use of SNPs alone. As well as developing genotyping assays for Indels and STRs, it was important to develop a frequency-based classifier in Snipper that could be applicable to forensic STR data, but also capable of extending the scope of ancestry inference to haplotype data, such as the genotypes generated for Y-SNPs and autosomal SNP microhaplotypes, where all components are linked and must be counted as haplotypes to estimate their frequencies. Lastly, it was an important additional enhancement to extend the functionality of Snipper with fixed training set data for 46 Indels alongside those already established for 34 SNPs, and to add forensic Indel panel data to SPSmart in the dedicated browser of forInDel. The establishment of Indels in particular has led to several initiatives that have made them the system of choice in forensic laboratorios outside of Santiago due to their freedom from ambiguity when mixed DNA is unknowingly analysed in such tests. The successful completion of an European DNA Profiling Group (EDNAP) collaborative exercise, where all participants were able to accurately genotype all 46 Indels and assign the ancestry of the control DNAs sent out, indicated the robustness and reliability of Indel-based forensic ancestry tests. All participants were also able to successfully detect a mixed DNA sample sent to them and even accurately infer the ancestries of the component individuals.

8. As compact massively parallel sequencing systems became available for forensic analysis, the thesis studies focussed on rebuilding the smaller AIM sets described above for capillary electrophoresis, into enlarged single panels of 130 to 160 SNPs. The resulting SNP panels developed were evaluated in terms of sequence balance in heterozygotes (the extent to which each SNP allele's sequence output was close to a perfectly balanced 50:50 ratio), sensitivity when analysing compomised DNA from degraded DNA or in dilution series, and genotyping concordance comparing genotyped samples with online data for the same DNAs sequenced with alternative techniques.

9. As massively parallel sequencing technology was applied, in the first case, to genotype STRs, it was necessary to identify and catalog the collateral variants found on the same sequence strands as the target loci and associated with the core STR and SNP markers genotyped by the new forensic massively parallel sequencing assays. The sequence output from both Illumina and Thermo Fisher systems is limited to the repeat region sequences and, where a SNP site is very close, or flanking sequence is relevant to the STR genotype based on repeat counts, extra sequence is reported. The thesis studies analysed the ancestry-informative potential of collateral variants starting with the HGDP-CEPH panel genotyped by the Illumina MiSeq-based Forenseq system. Further studies have begun to reveal that colateral variants are in fact strongly associated with a particular repeat allele in the STR which means that the sequence strand containing the repeat region and flanking region SNPs must be treated as a haplotype and counted to obtain relevant frequencies, rather than from combining individual allele frequencies in the STR and the SNPs. The prevailing view from these studies and further analysis of targeted SNP sets genotyped from the same sequence strand indicates that almost all forensic loci analysed by massively parallel sequencing are microhaplotypes.

10. The above initiatives led to the step to specifically build sets of Microhaplotype loci, comprising sets of closely linked SNP alleles in haplotype combinations, that show strongly population-differentiated patterns of variation. In all forensic ancestry assays now deleoped at Santiago for massively parallel sequencing assays, panels of Microhaplotype loci informative for ancestry or applicable to identification (e.g. missing person identification) have been adopted by re-designing the sequence lengths to amplify Microhaplotypes in very short fragment lengths.

Two challenges remain: to find more Microhaplotypes in the genome and to devise a simple way to identify those loci compiled from genome-wide searches that are most effective for ancestry analysis. To identify Microhaplotypes most informative for normal forensic identification purposes just requires the calculation of Gene Diversity values (a variant on the Heterozygosity metric applied to multiple polymorphisms in linkage such as Y-STRs). This is a straightforward step from population frequency data already provided in 1000 Genomes. There is the need to convert the phase of the genotypes into haplotype combinations;.e.g. three SNPs given as A|G, G|G, T|C are then recorded as haplotypes AGT and GGC, which are distinguishable from samples that have the same genotypes, but are actually strand combinations AGC, GGT - which are counted separately when compiling the total variability of the Microhaplotype. The same problem of loci with similar Gene Diversity values in different populations, but quite different distributions of haplotype frequencies, exists in Microhaplotypes as it does with multiple-allele SNPs. Therefore, the construction of appropriate metrics to highlight the most ancestry-informative Microhaplotypes will be an important future step in compiling a much larger set of candidate loci from which to choose the best or as a way to maximise the potential to differentiate more closely related populations.