Forensic Science International: Genetics xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Forensic genetic analysis of bio-geographical ancestry

Chris Phillips *

Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Galicia, Spain

ARTICLE INFO

Article history:
Received 23 October 2014
Received in revised form 2 May 2015
Accepted 14 May 2015

Keywords:
Bio-geographical ancestry
Ancestry informative markers (AIMs)
Population genetics
SNP
PCA
STRUCTURE

ABSTRACT

With the great strides made in the last ten years in the understanding of human population variation and the detailed characterization of the genome, it is now possible to identify sets of ancestry informative markers suitable for relatively small-scale PCR-based assays and use them to analyze the ancestry of an individual from forensic DNA. This review outlines some of the current understanding of past human population structure and how it may have influenced the complex distribution of contemporary human diversity. A simplified description of human diversity can provide a suitable basis for choosing the best ancestry-informative markers, which is important given the constraints of multiplex sizes in forensic DNA tests. It is also important to decide the level of geographic resolution that is realistic to ensure the balance between informativeness and an over-simplification of complex human diversity patterns. A detailed comparison is made of the most informative ancestry markers suitable for forensic use and assessments are made of the data analysis regimes that can provide statistical inferences of a DNA donor's bio-geographical ancestry.

1. Introduction

1.1. The forensic context of bio-geographical ancestry analysis

In London seventeen years ago, seeing somebody acting suspiciously outside my neighbor'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 need to correct these descriptions to the more neutral terminology of European, African, South Asian). As it was dark and I had brief glimpses, it was impossible to provide a concrete description. Eyewitness is notoriously unreliable and can be shaped by preconceptions or the 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 information about a suspect.

This review explores the current viability of forensic DNA tests estimating ancestry that can provide investigative leads when eyewitness testimony or a database hit are not available. In simple lay terms, ancestry can be described as the genetic inheritance each individual carries from their ancestors, in the immediate past from their kinship, over longer periods from population members that have occupied the same place of origin. Bio-geographical ancestry analysis concentrates on the 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 tests must be sensitive enough to successfully genotype contact trace DNA 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].

This review centers on autosomal markers, despite Y and mtDNA uniparental 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

* Tel.: +34 981 582 327; fax: +34 981 580 336.
E-mail address: c.phillips@mac.com (C. Phillips).

http://dx.doi.org/10.1016/j.fsigen.2015.05.012
1872-4973/

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 [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. 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 uniparental 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 (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 [11]).

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 population groups based on the most strongly differentiated patterns of 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 multiplex sizes and collection of sufficient reference data means a simplified description of complex human population structure is a necessary compromise.

Human populations are not fully interbreeding, since geographic distance by itself creates a strong constraint on random mating. Additionally, geophysical barriers such as oceans and mountains have restricted free movement of people away from regions defined by such barriers. Therefore, population structure in early human groups became established as they continued to mate with immediate neighbors 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. Pre-genomics studies of population variation, starting with Lewontin [12], attempted to measure what structure existed in modern human population groups using limited numbers of polymorphic markers. Despite variation in loci and populations analyzed, later studies with the same approach

	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***			

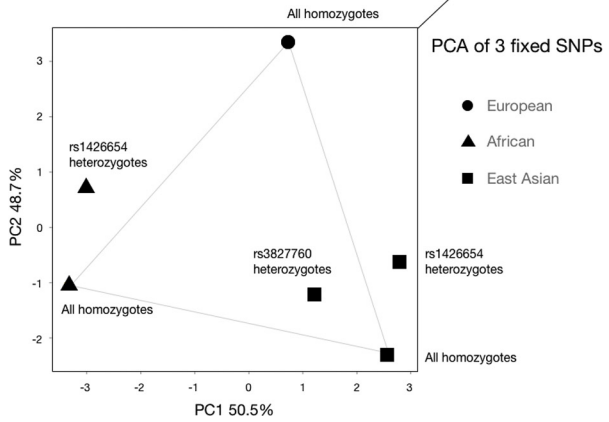

Fig. 1. Five examples of AIM-SNPs. SNP 1 shows a population group-specific allele, SNP 2 has near-fixed variation between Africans and East Asians. SNPs 3–5 are the most informative (reflected in the I_n values listed) with fixed alleles in each group. Combined I_nPOP divergences reach a reasonably comparable level of balance as SNP properties compensate for the distribution of variation amongst the groups analyzed. 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.

Please cite this article in press as: C. Phillips, Forensic genetic analysis of bio-geographical ancestry, Forensic Sci. Int. Genet. (2015), http://dx.doi.org/10.1016/j.fsigen.2015.05.012

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 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. in 2002 [14] used the *STRUCTURE* genetic-similarity clustering algorithm [15] to measure population structure in the CEPH Human Genome Diversity Project sample set (HGDP-CEPH). The HGDP-CEPH panel comprises 1064 individuals from 52 populations and remains the most comprehensive global population survey [16], despite certain very small sample sizes and significant gaps in geographic coverage. Rosenberg 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 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 (values corresponding to 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 analyze 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-analyzed the HGDP-CEPH panel with more markers ([21], 377 > 933 STRs) and demonstrated clusters are robust to sampling location. Therefore, genetic clusters represent underlying patterns of human variation and are not artifacts 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 recognized phenomenon of archaic introgression.

Natural selection can vary according to bio-geographical factors such as climate, presence of disease or diet/agricultural practices [22]. Well-documented examples exist for these factors, creating strong discontinuities in variant allele frequencies, including genes: SLC24A5 producing de-pigmentation in Europeans; DARC in African populations in response to regional prevalence of malaria, and LCT-MCM6 in three separate geographic regions as adaptation to milk consumption [23–25]. 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 harbor specific coding SNPs that remain the most powerful ancestry markers, with most now adopted for forensic use.

Mass movement of peoples followed by admixture is also a major influence on contemporary population diversity. The effects of North Atlantic slave trading and colonization are well documented, but for more comprehensive insights into population movement predating historic record, very dense genetic data is needed. One study by Hellenthal et al. [30] used 'chromosome painting' analyzing recombinational decay of chromosome segments containing SNP haplotypes. The same approach enabled a recent 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 summarizes major population movements in the last 20 KY, from prehistory to recent colonialism. Lastly, the characterization 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 present in modern

Please cite this article in press as: C. Phillips, Forensic genetic analysis of bio-geographical ancestry, Forensic Sci. Int. Genet. (2015), http://dx.doi.org/10.1016/j.fsigen.2015.05.012

non-Africans (introgression 37-85 KYA) and ~7% Denisovan genetic ancestry in modern Oceanians [33,34].

Summarizing 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]. Therefore, forensic and population genetics specialists must guard against a desire to make overly detailed reconstructions of an individual's ancestry, particularly when this has little or no relation to what is currently understood about human diversity.

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 '*AncestrybyDNA*' 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 maximize 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 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 I_n is divergence $\times$ 0.693 (i.e., converting the natural log to $\ln_{(2)}$). All have maximum values of 1 in pairwise population comparisons, denoting complete divergence and zero for none. Divergence values can be automatically estimated for up to 200 SNPs and 20 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]. In addition, a simple divergence– I_n–F_{ST} Excel calculator is provided in Supplementary File S1.

Before taking the obvious step of selecting the topmost SNPs from a ranked δ 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 KY 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, Fig. 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_nPOP) between Europe and the other two (put as I_nEUR vs. I_nAFR and I_nE ASN). If a forensic test only used SNPs like rs12075 it would lack power to differentiate Europeans, so SNPs such as rs4988235 are required to redress the balance. The other three SNPs in Fig. 1 are near 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_nPOP 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 groups with equal power. However, when forensic ancestry tests grow to 30 or more loci, maintaining

Please cite this article in press as: C. Phillips, Forensic genetic analysis of bio-geographical ancestry, Forensic Sci. Int. Genet. (2015), http://dx.doi.org/10.1016/j.fsigen.2015.05.012

a balance of population-specific I_n divergence values becomes more difficult. Another challenge to maintaining balanced divergences is the lack of fixed SNPs. Such frequency distributions originate from favorable coding SNP substitutions creating hard sweeps from very strong 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 Fig. 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_nAME < I_nEUR), 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_nPOP) was previously recognized 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_nAFR); European vs. the other two; East Asian vs. the other two. In Fig. 1 rs2814778 show lower I_nEUR and I_nE ASN values, as all these group's divergence is with Africans. Providing 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 minimize the admixture estimation bias prevalent in small AIM sets.

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

Although ancestry tests may target the differentiation of other population groups besides Africans, Europeans and East Asians, population data is not always available to allow selection of AIMs informative for Oceanian, unadmixed American or South Asian/Middle East 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 list of human variant sites as the Phase III data release 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 2535 individuals from 26 populations [55]. Details of these populations are given in Fig. 2. Although Li's HGDP-CEPH data surveys much less loci in comparison, 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 lists of SNPs, chromosome segments or gene symbols. Genotypes obtained can 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_nPOP 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 screened 8474 candidates in the

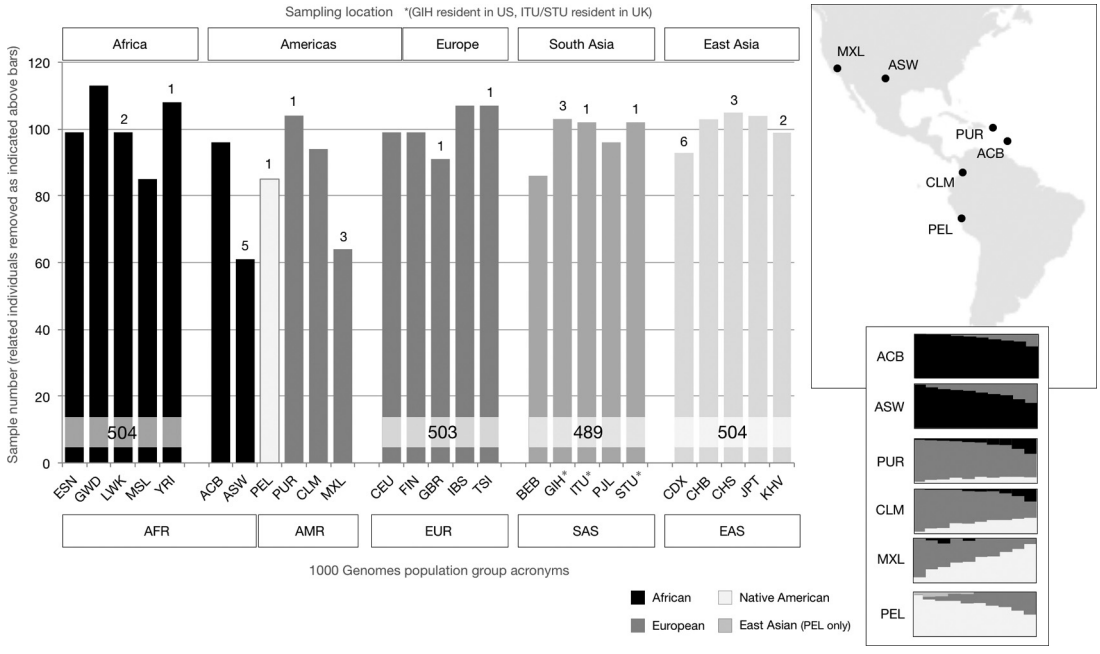

Fig. 2. 1000 Genomes Phase III sample details in a nutshell. Bar charts show that component populations in four groups combine to near identical sample sizes. Numbers above bars show the related individuals removed (one per related-pair identified). More details at http://www.1000genomes.org/category/frequently-asked-questions/population. ACB/ASW and PUR/CLM/MXL are shaded by their average majority co-ancestry estimated from *STRUCTURE* analyses shown in Fig. 7 (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 summarizing the same *STRUCTURE* analysis.

Affymetrix 10K genome-wide array using 74 Y-Chromosome Consortium samples analyzed with *STRUCTURE* (optimum K:4 clusters of African–American–Asian–Eurasian). The SNPs were assessed with F_{ST} and I_n4 comparing the four regions defined by *STRUCTURE*, and 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 were obtained with AIMs selected with highest pairwise F_{ST} values (distinct from classical F_{ST} looking at all groups together). This shows aiming for balanced divergence selects the most informative set. Analyzing 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 evaluate non-European Eurasian, American or Oceanian variation. Nevertheless, using the 34-plex set to analyze 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 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], so 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 not used in genome-wide SNP arrays.

Two other SNaPshot-based forensic ancestry panels have been recently published by Gettings et al. [64], genotyping 50 AIM SNPs in 3 multiplexed assays and by Daniel et al. [65], genotyping 14 AIM SNPs in 2 multiplexed assays.

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

Since most marker sets described above were developed, larger panels have been published for genomics use that provide powerful AIMs worth consideration for forensic application. In publication order these 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 focus on African, European and Native American SNP variation (i.e., not East Asian), but none have optimized multiplexes. Two recently developed 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 Asian or other populations not originally targeted. To adjust further for this 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 forms the HID-Ion AmpliSeq™ Ancestry Panel optimized for Ion PGM™ NGS system [71], while the 55 Kidd AIMs alone form the ancestry informative portion of the Illumina MiSeq ForenSeq® NGS 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_nPOP 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_nEUR: 14.56, I_nE ASN: 14.23, I_nOCE: 14.71, I_nAME: 14.82, I_nAFR: 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

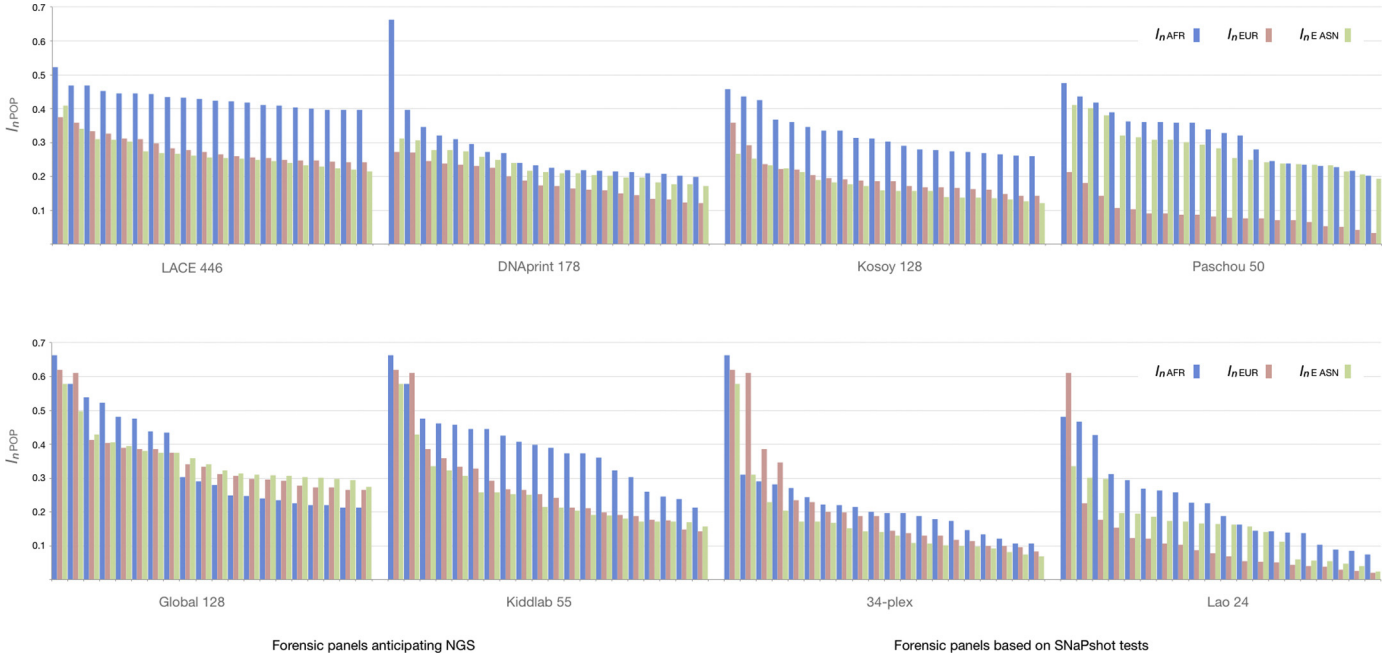

Fig. 3. Distribution of population specific divergence values in the best 20 SNPs of eight AIM panels applicable to forensic ancestry analysis. Lower charts show panels with multiplexed assays in use (Kiddlab 55 + Kosoy 128 in Ion PGM™ ancestry panel). Most panels show higher African informativeness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

Please cite this article in press as: C. Phillips, Forensic genetic analysis of bio-geographical ancestry, Forensic Sci. Int. Genet. (2015), http://dx.doi.org/10.1016/j.fsigen.2015.05.012

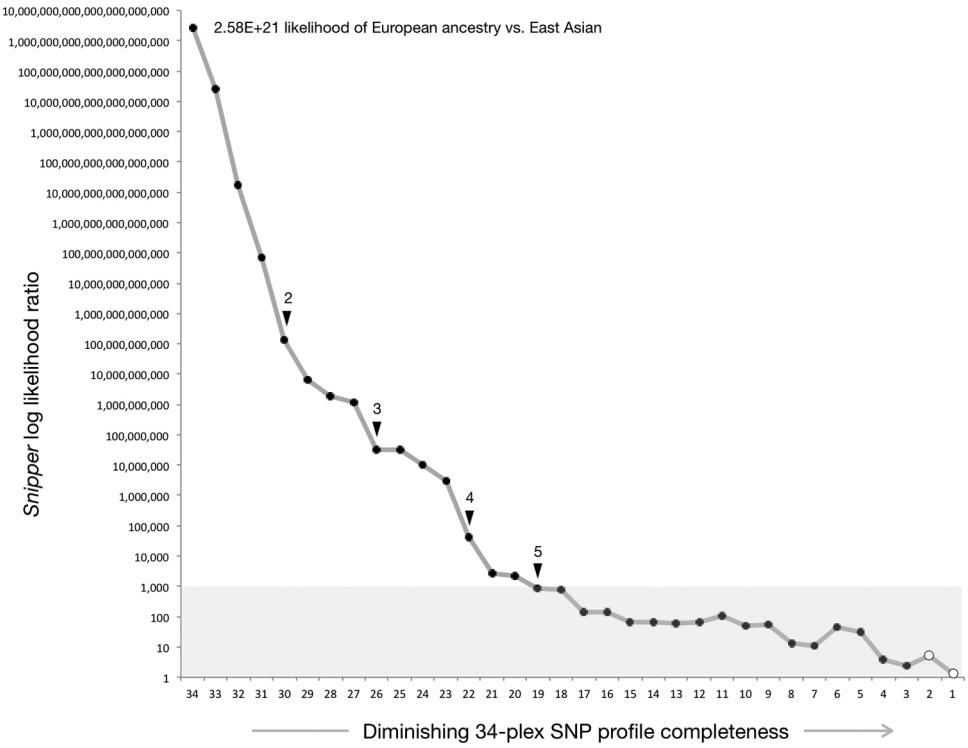

Fig. 4. Bayes analysis with *Snipper* of progressively degraded 34-plex 9947A profile, removing the most informative AIM-SNPs first. Log_{10} likelihood ratio (LR) probabilities are for European vs. East Asian ancestry likelihoods, apart from the 1-2 SNP profiles showing African vs. European LRs. Numbers refer to points shown on the 34-plex PCA of Fig. 5A. 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 Fig. 5A.

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 Fig. 3 and the top 24 across all panels are listed in Table 1 (underlying data in Supplementary Table S1). 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].

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 vs. 26 in Phase III), while the FROGkb and ALFRED allele frequency

Table 1
SNPs common to three or four established ancestry panels. Bold rs-numbers indicate SNPs listed in this table of 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

Please cite this article in press as: C. Phillips, Forensic genetic analysis of bio-geographical ancestry, Forensic Sci. Int. Genet. (2015), http://dx.doi.org/10.1016/j.fsigen.2015.05.012

Table 2 (*Continued*)

(C) Rank	AIM	I_nE ASN	Kiddlab 55	Global 128	Other
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)			

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 analyzed 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 optimized 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 characterize 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. The 34-plex SNP ancestry test previously described, organized 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,* potentially offers the most useful option, since genotypes generated by the laboratory or collected from online/published sources can be assessed and applied to any AIM panel 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 plus reference data are supplied in Supplementary File S2 and the profile (in cell C4) can be directly 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 in rs-number order returns a European vs. East Asian likelihood ratio (LR) of 2.58E + 21; a very large likelihood toward being European rather than East Asian or African. However, such a large number is difficult to interpret directly and needs some qualification. First, only three possible population groups were compared to make the assignment, if more are included likelihood values 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 a likelihood to be European of 5.18E + 18 compared to American, as this ancestry replaces East Asian as the second highest likelihood. Second, the profile analyzed 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 minimize 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*. Fig. 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 I_nEUR 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.

4.3. Principal component analysis

PCA tests were first proposed in the nineties by Cavalli-Sforza et al., in order to summarize 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 ($'\mathbb{R}'$ 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 analyzing population genetic data from simple SNP tests such as those already described, the condensation of total variance follows an approximate route of $\mathbb{R}^{20\sim200}$ into $\mathbb{R}^3$, 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 neighbors [85]. Therefore, the inference of past population events such as 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 can create close matches that are persuasive, but may not 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]. The 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 ([110] 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 Fig. 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. Fig. 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 discernable substructure within any one cluster suggesting both panels have minimized 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 (Fig. 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.

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

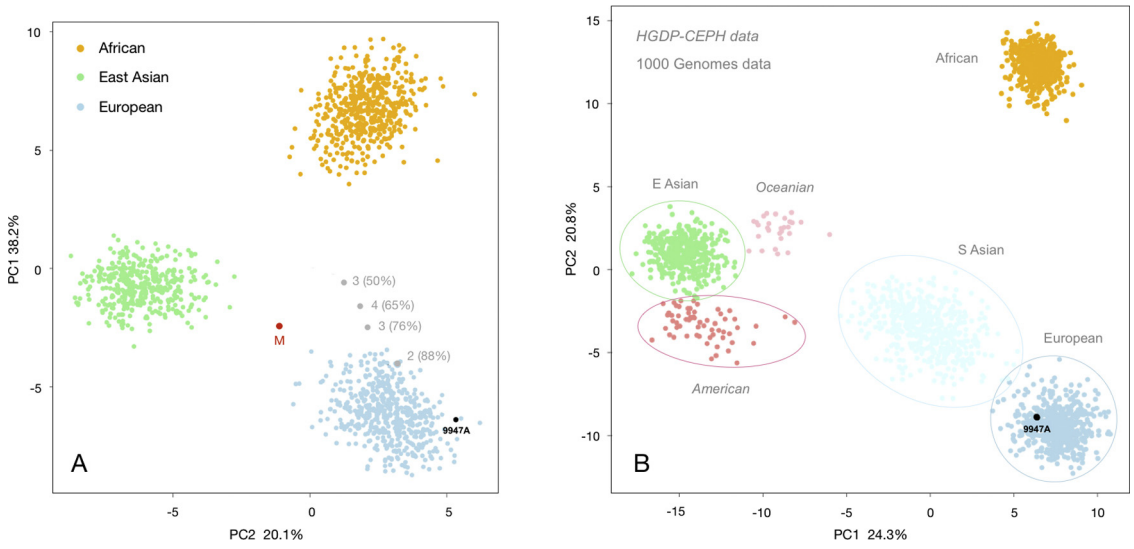

Fig. 5. *Snipper* generated PCA plots showing the system of overlaying an unknown profile (9947A) onto reference data clusters. (A) 34-plex genotypes from 1000 Genomes. Gray points are points of reduced 9947A profile shown in Fig. 4. Point M is an artificial 3:1 mixture of European–East Asian donors. (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.

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 analyzed.

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 analyzed [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] summarizes 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, continental margin populations 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 urbanization 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 probability value from the two highest likelihoods, which can have less contrasted values but still provides 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 for the AIMs used. The pairwise ranked log LR charts used in the 11-M analyses (Fig. 1, [8]) 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 minimize 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 advisable to compare three admixture contributor population groups per analysis. This can be arranged from the known levels of divergence in the AIMs used. For example, applying the Kosoy AIM panel to a PCA comparison of American and East Asian reference data would reveal limited divergence that the Kiddlab SNPs can help address [67,68]. Therefore, a forensic sample analyzed in a US laboratory might consider a PCA of Africans/Europeans with Native American or East Asian data in two 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 centers (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 Fig. 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 analyzing admixture patterns but coefficients of cluster membership 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, *STRUCTUREs* provide 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 from such an approach, the six admixed populations from 1000 Genomes Phase III were analyzed 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 Fig. 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, therefore 10-percentiles were calculated from the *STRUCTURE* output and plotted in Fig. 2 (membership coefficient matrices in Supplementary Table S2). 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 forensic ancestry tests will need to address.

6. Beyond binary AIM–SNP panels

6.1. Indels

Indel variation mirrors that of SNPs as they are binary loci that 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

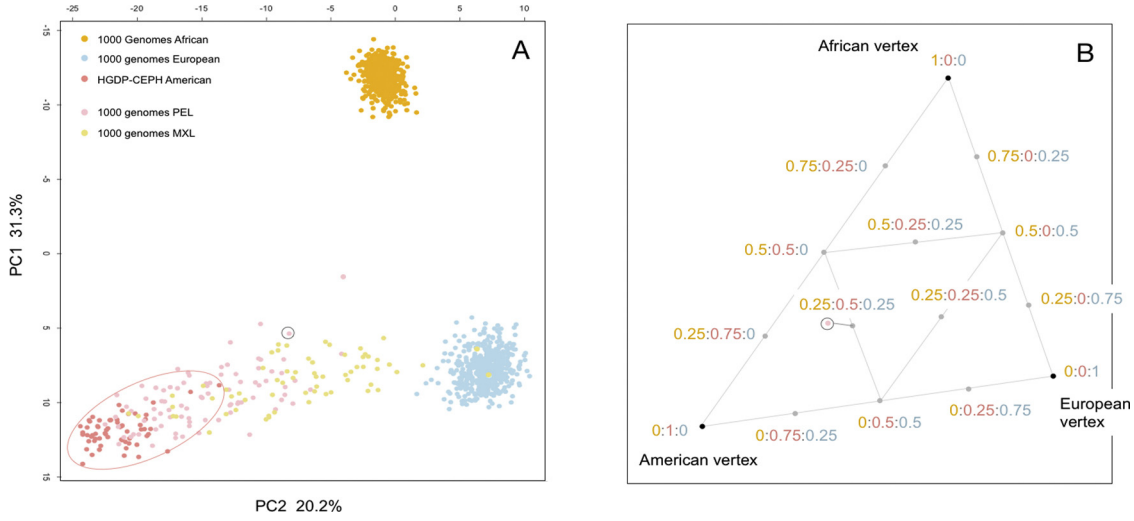

Fig. 6. (A) PCA plot of African, European and American reference groups compared to PEL and MXL (1000 Genomes + HGDP-CEPH Americans), Global AIM panel. (B) Map of centroids based on the geometric distribution of the three reference group mid-cluster vertices in plot A. Numbers denote ratios of approximate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

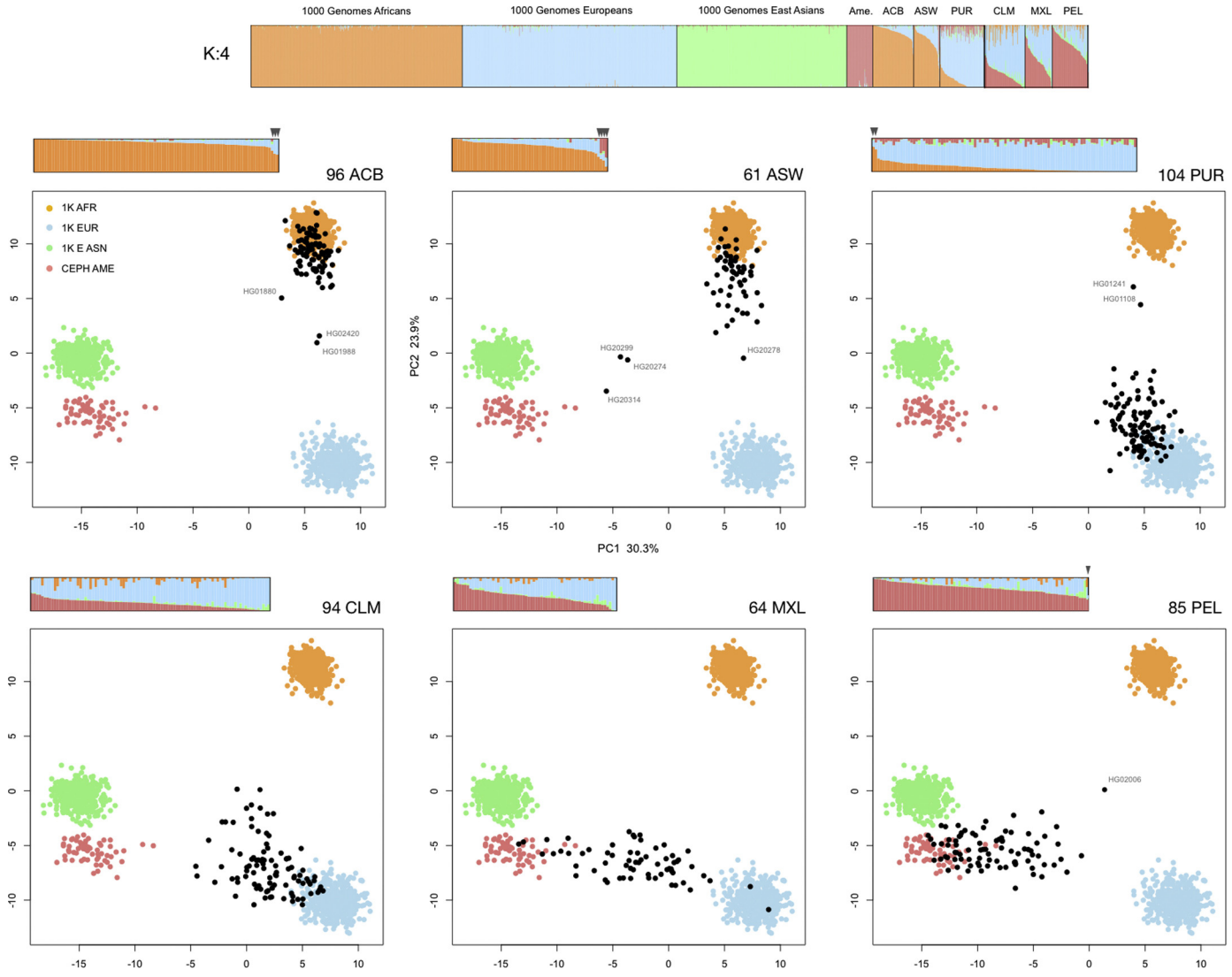

Fig. 7. *STRUCTURE* and PCA analysis of 1000 Genomes Phase III admixed populations. Upper cluster plot shows all populations (East Asians without CHS), individual plots show admixed populations in more detail, ranked by decreasing African or American cluster memberships. Outlier points in each PCA are numbered and indicated on the matched cluster plots.

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 utilize 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. 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 [130]. 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 Fig. 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 I_n3 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. A study in 2003 by Rosenberg et al. [44] looked in detail at the 377 STRs used by the same group to analyze 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_n3 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 the data is generated in almost all 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: *PopAffiliator*, 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 *PopAffiliator* site has recently been upgraded (http://cracs.fc.up.pt/~nf/popaffiliator2) with modified choices of three or five group comparisons.

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. [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 also highlighted presence of population-specific alleles in certain STRs (Fig. 5, [106]), with the most marked specificity occurring in the 9-repeat allele of D9S1120 [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). Finally, novel ancestry-informative tetranucleotide repeat STRs were developed by Phillips et al. in 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.

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 characterized and make up 1 in 300 of the Phase III SNPs (259,370 of 78,136,341 variants). Two triallelic SNPs are in the 34-plex set as they show marked population differentiation while providing the means to detect third alleles in

Please cite this article in press as: C. Phillips, Forensic genetic analysis of bio-geographical ancestry, Forensic Sci. Int. Genet. (2015), http://dx.doi.org/10.1016/j.fsigen.2015.05.012

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 SNP) compared to binary loci.

7. Concluding remarks

With enough care and consideration of the interpretative limits outlined in this review, forensic ancestry analysis reveals quite a lot of detail about the likely geographic origin of an unknown DNA donor. Major limitations remain, namely: the necessity to simplify complex worldwide patterns of human divergence, as small-scale marker sets are required for forensic DNA; the lack of sufficiently widespread reference population data; and the difficulty of assessing complex admixture patterns in individuals with co-ancestral backgrounds. Luckily, NGS brings the possibility to generate large amounts of genotype data reliably from expanded multiplexes made in single-tube analyses. So marker depth will increase significantly and this is certain to aid detection and interpretation of simple admixture patterns. For those not ready to adopt NGS, conventional CE detection is easy to accomplish with SNP or indel-based ancestry tests. Indels in particular are robust to mixed DNA and in sufficient numbers can match the ancestry-informativeness of the best SNPs. Now full HGDP-CEPH genotypes for the 46-plex indels are present in *SPSmart* [79] and a simple combined Bayes and PCA analysis framework is available in *Snipper*, making ancestry analysis a much more straightforward technique for forensic laboratories interested in assessing this field for themselves.

Acknowledgements

The author is grateful for the essential support provided by three colleagues at USC, without which this review would not have been possible: Prof. Maviky Lareu for allowing him to pursue forensic ancestry analysis relatively undisturbed for nearly 10 years; Prof. Antonio Gómez-Tato for his unstinting help in tackling complex data analysis and expert guidance in classification systems; and Carla Santos for making all the data analyses included here so willingly.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.fsigen.2015.05.012.

References

[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 http://paintmychromosomes.com (accessed April 2015).
[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 http://www.senseaboutscience.org/data/files/resources/119/Sense-About-Genetic-Ancestry-Testing.pdf (accessed April 2015).
[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, A. 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 April 2015).
[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, et al., 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 April 2015).
[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 April 2015).
[72] Illumina ForenSeq system: http://applications.illumina.com/content/dam/illumina-marketing/documents/products/appspotlights/app_spotlight_forensics.pdf (accessed April 2015).
[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 SNPforID Consortium, Differentiating European and South Asian individuals using SNPs and pyrosequencing technology, Forensic Sci. Int. Genet. Suppl. Ser. 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 SNP*for*ID browser: an online tool for query and display of frequency data from the SNPforID 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.

Please cite this article in press as: C. Phillips, Forensic genetic analysis of bio-geographical ancestry, Forensic Sci. Int. Genet. (2015), http://dx.doi.org/10.1016/j.fsigen.2015.05.012

[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 geneticand 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 substructure 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.

Please cite this article in press as: C. Phillips, Forensic genetic analysis of bio-geographical ancestry, Forensic Sci. Int. Genet. (2015), http://dx.doi.org/10.1016/j.fsigen.2015.05.012