Available online at www.sciencedirect.com
ScienceDirect
Forensic Science International: Genetics 1 (2007) 273–280
FSI GENETICS
www.elsevier.com/locate/fsig

Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs

C. Phillips [a,*], A. Salas [a], J.J. Sánchez [b], M. Fondevila [a], A. Gómez-Tato [c], J. Álvarez-Dios [c], M. Calaza [c], M. Casares de Cal [c], D. Ballard [d], M.V. Lareu [a], Á. Carracedo [a]

The SNPforID Consortium[1]

[a] *Forensic Genetics Department, Genomic Medicine Group, University of Santiago de Compostela, 15782, and Centro Nacional de Genotipado (CeGen), Genomic Medicine Group, Hospital Clínico Universitario, 15706 Galicia, Spain*
[b] *Department of Forensic Genetics, University of Copenhagen, Denmark*
[c] *Faculty of Mathematics, University of Santiago de Compostela, 15782 Galicia, Spain*
[d] *Department of Haematology, ICMS, Queen Mary's School of Medicine & Dentistry, London, UK*

Received 18 June 2007; received in revised form 25 June 2007; accepted 27 June 2007

Abstract

Tests that infer the ancestral origin of a DNA sample have considerable potential in the development of forensic tools that can help to guide crime investigation. We have developed a single-tube 34-plex SNP assay for the assignment of ancestral origin by choosing ancestry-informative markers (AIMs) exhibiting highly contrasting allele frequency distributions between the three major population-groups. To predict ancestral origin from the profiles obtained, a classification algorithm was developed based on maximum likelihood. Sampling of two populations each from African, European and East Asian groups provided training sets for the algorithm and this was tested using the CEPH Human Genome Diversity Panel. We detected negligible theoretical and practical error for assignments to one of the three groups analyzed with consistently high classification probabilities, even when using reduced subsets of SNPs. This study shows that by choosing SNPs exhibiting marked allele frequency differences between population-groups a practical forensic test for assigning the most likely ancestry can be achieved from a single multiplexed assay.

Keywords: Ancestry-informative marker; AIM; SNP; Geographic origin; Ancestry test; CEPH diversity panel

1. Introduction

Ancestry-informative markers (AIMs) that can indicate the likely population of origin of a DNA sample where the source individual is not known or is unable to declare their ancestry are an increasingly important part of genetic association studies but to date have not been developed into a practical forensic test. A range of DNA polymorphisms are available with potential to be used as AIMs including autosomal and Y-chromosome short tandem repeats (STRs) and mitochondrial sequence variation (mtDNA) [1–5] but these have limitations. Micro-satellites do not exhibit large enough contrasts in allele frequencies between populations to be especially useful in numbers below 50 loci, mainly due to their mutational instability. Y-chromosome loci and mtDNA variation, while phylogeographically informative [6–8], are haploid so require very large databases to properly gauge population variability plus there is a risk of finding intact lineages atypical of the population [9]. Autosomal SNPs have emerged as amongst the best ancestry markers due to their stability, density of distribution and full range of allele frequency patterns amongst populations. Since the over-riding majority of human worldwide genetic diversity takes the form of geographic clines rather than clades [10–12], it is essential to find the small number of SNPs that show the most pronounced allele frequency discontinuities between continental regions to create marker sets with population ''diagnostic'' genotypes [13]. To help locate such SNPs one approach is to examine gene variation that has been subjected to strong regional positive selection in the recent past creating localized adaptations [14–16]. Well-documented

* Corresponding author. Tel.: +34 981582327; fax: +34 981580336.
E-mail addresses: c.phillips@mac.com, chrisp@usc.es (C. Phillips).
[1] www.snpforid.org.

1872-4973/$ – see front matter © 2007 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.fsigen.2007.06.008

examples [reviewed in 17] include *SLC45A2* and de-pigmentation in Europe [18], *DARC* and *Plasmodium vivax* resistance in sub-Saharan Africa [19] and *LCT* implicated in pastoralist adaptation in Northern Europe [20]. We examined these genes and others to collect highly regionalized SNP variation.

In this study we aimed to develop a suitably powerful single-tube SNP test that showed the least error and was based on the most informative AIMs available, with these goals (1) to select SNPs that, in the first instance, gave a clear differentiation of sub-Saharan African, European and East Asian population-groups; (2) to validate allele frequencies to ensure within population-group variation was a minor proportion of total variability; (3) to balance the chromosome distribution of the final set to avoid linkage disequilibrium between SNP pairs; (4) to establish a straightforward Bayesian system for predicting ancestral origin and to estimate the misclassification rate by statistical means and by testing the CEPH human genome diversity cell line panel (CEPH-HGDP) comprising samples of confirmed geographic origin [21].

A forensic test handling single profiles requires a fast and flexible alternative to the widely used genetic clustering algorithm STRUCTURE [22] to offer easier classifications in real time. Therefore, the final stage of development of the test outlined here was the incorporation of the classification algorithm into an open access web portal to allow simple analysis of SNP profiles, including those with partial data. This portal was enhanced to allow analysis of a users custom populations and SNP markers with the same Bayesian classification algorithm and error estimation systems.

2. Materials and methods

2.1. Population samples

Training sets for the classification algorithm were created for each population-group by combining two population samples comprising: sub-Saharan Africans (60 Mozambican and 60 Somali), Europeans (60 Galician from NW Spain and 60 Danish) and East Asians (60 Mainland Chinese and 60 Taiwanese). In all cases informed consent was obtained. Except for Somalis resident in Denmark samples were collected in the corresponding geographic region. The CEPH-HGDP panel comprising 1064 samples from 51 geographically diverse populations [21] was used to test classification performance with new profiles. We used a seven-region affiliation model for CEPH-HGDP that was previously estimated using 377 autosomal micro-satellites by Rosenberg et al. [13]. The Rosenberg study used the genetic clustering algorithm STRUCTURE [22] to group populations into Africa, East Asia, Oceania, America and a division of Eurasia into Europe, Middle East and Central/South Asia groups. We used STRUCTURE in the same way to analyze separately the CEPH-HGDP and training sets with the 34 AIM SNPs chosen. Finally, to begin an assessment of population admixture and its effect on ancestry assignment we analyzed 163 self-identified African–Americans supplied by the Biochemical Science Division, National Institute of Standards and Technology (NIST, Gaithersburg, USA).

2.2. Selection of SNPs and development of genotyping assay

SNPs were selected from the following types of AIM: (1) population specific markers, comprising loci with a polymorphism detected in one or two population-groups but absent in the other(s), isolated from the Applied Biosystems assays-on-demand SNP database [23]; (2) skewed allele frequency markers, comprising SNPs with a common allele in one population that is rare in others, using an allele frequency differential between populations (δ) of >0.6 to qualify [24]; (3) tri-allelic SNPs, displaying multiple substitutions at the same position. We selecting two loci from a previously collected set [25] that both showed a different common allele in African, European and East Asian groups; (4) fixed difference markers, comprising SNPs where one allele is seen exclusively in one population-group and the alternative allele exclusively in the others. Not surprisingly such SNPs are rare and we were able to include only three: rs1426654, rs16891982 and rs2814778.

The SNP genotyping assay comprised a 34-plex PCR followed by a 34-plex SNaPshot® (AB: Applied Biosystems, Foster City, USA) primer extension reaction, using a 3130 capillary electrophoresis analyzer (AB) and POP6™ to detect the alleles. Details of the reactions, cycling, primers and extension product sizes are outlined in Table S1 in online supplementary data. A typical casework electropherogram of a 34 SNP profile is shown in Fig. 1. To develop a simple, sensitive assay confined to a single-tube and with forensic sensitivity we used the approach previously described for large-scale SNP multiplex design [26] based on: (1) stepwise addition of new candidate loci to augment a core 20-plex reaction; (2) amplicon sizes typically below 100 base pairs (average in assay: 88 base pairs); (3) careful screening for primer and sequence interactions that reduce PCR efficiency. The optimized PCR and primer extension reactions provided complete 34 SNP profiles from a routine DNA quantity range of 1–10 ng but was observed to work efficiently in cases with as little as 200 pg of DNA.

2.3. Data analysis

SNP informativeness was measured using Jensen and Shannon's divergence [27, and equivalent to the $I(n)$ metric described in 3] which can be defined for n equal-weighted populations by

$$\mathrm{Div}(\mathrm{pop}_1, \ldots \mathrm{pop}_n) = H\left(\frac{1}{n}\sum_i \mathrm{pop}_i\right) - \frac{1}{n}\sum_i H(\mathrm{pop}_i)$$

where the Shannon information measure: H denotes population entropy, defined for a population with multinomial distribution of parameters ($p_1, \ldots p_s$), by

$$H(\mathrm{pop}) = H(p_1, \ldots p_s) = -\sum_i p_i \log_2 p_i$$

In practice divergence and the more widely used Wright's F_{ST} are highly correlated, e.g. $\delta = 0.5$ corresponds to

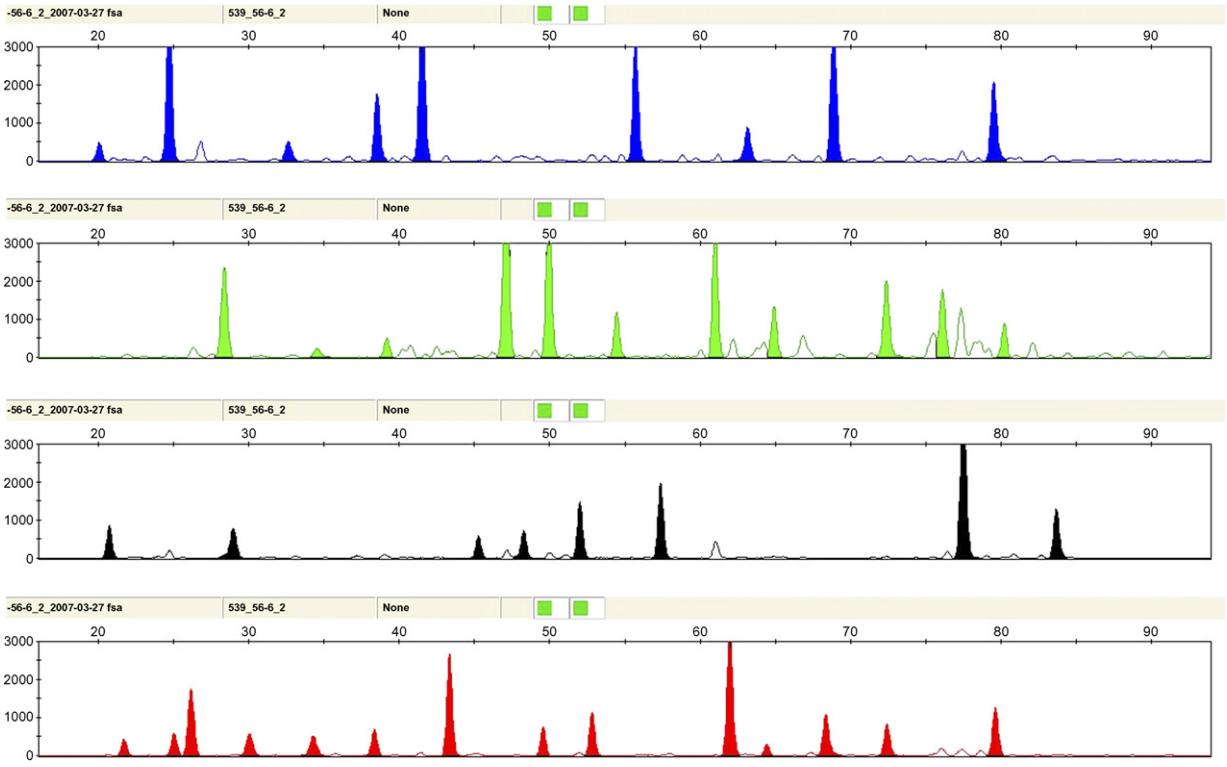

Fig. 1. Electropherogram of an example 34 SNP profile obtained from 0.6 ng of DNA extracted from a toothbrush.

divergence = 0.19, $F_{ST} = 0.25$ and the best SNPs we selected gave divergences >0.4 equivalent to $F_{ST} = 0.5$, $\delta > 0.7$.

To assign a set of individuals into classes we applied the simple Bayes classification rule [28] using a set of samples of each class as training sets for a maximum likelihood calculation. Here a class denotes an ancestral origin simplified to the term population (pop), and the term individual (ind) denotes a SNP profile. An individual can be assigned to the population that maximizes the posterior probability $P(\text{ind} \in \text{pop}/\text{ind})$. For populations $\text{pop}_i, \ldots \text{pop}_n$ using Bayes theorem the posterior probability is given by

$$P(\text{ind} \in \text{pop}_i | (x_1, ..., x_m)) = \frac{P((x_1, ..., x_m) | \text{ind} \in \text{pop}_i) P(\text{ind} \in \text{pop}_i)}{P(x_1, ..., x_m)}$$

where $(x_1, \ldots, x_m)$ is the SNP profile. Assuming class conditional independence the above formula simplifies to

$$P((x_1, ..., x_m) | \text{ind} \in \text{pop}_i) = P(x_1 | \text{ind} \in \text{pop}_i), ..., P(x_m | \text{ind} \in \text{pop}_i)$$

Likelihood parameters $(x_1, \ldots, x_m)$ are estimated from training set allele frequencies assuming Hardy–Weinberg equilibrium and independence for the loci used. Undetected alleles defaulted to the standard conservative frequency estimate: $1/(2n + 1)$, where n is the group training set number. Classification success estimators comprised: apparent misclassification, one-out cross-validation and the .632+ bootstrap method [29]. Real classification performance with de novo samples was assessed by analyzing all 1049 CEPH-HGDP profiles, although misclassification rates were estimated only with those populations affiliated to the African, European and East Asian groups (123, 160 and 237 individuals from 6, 8 and 17 populations, respectively). Training set samples and CEPH-HGDP were assessed separately for group membership with the 34 SNPs using STRUCTURE v.2.0. Runs consisted of 200,000 Markov Chain steps after a burn-in of length 200,000 with five replicates for each K value from 2 to 6.

The above Bayesian classification system was placed in an open access web portal (http://mathgene.usc.es/snipper/) to analyze SNP profiles uploaded by the user. Genotype gaps were permitted using "NN" codes, with apparent success recomputed from the partial profiles. Data output includes: −log likelihoods for each population-group (use of −log likelihoods permits easier comparison of the very large likelihood ratio figures normally generated), percentiles indicating the percentage of training set samples with poorer likelihoods than the submitted profile, and a list of the SNPs used for the calculations in order of informativeness (i.e. highest divergence down). From the three likelihood ratios an assignment of most likely ancestral origin is made based on the highest probability, or the statement: "this individual cannot be classified" in cases where the maximum of probabilities is below 0.34 (i.e. balanced odds of one in three).

All populations were assessed for Hardy–Weinberg equilibrium using χ^2 analysis while linkage disequilibrium was tested using Fisher's exact test based on 3200 'shufflings'.

3. Results

3.1. Patterns of SNP variability

The allele frequency distributions for 34 SNPs in the three population-groups studied are outlined in Fig. 2. To compare the training set and CEPH-HGDP frequencies the populations from each were combined in their group affiliations separately and arranged in paired plots. Allele frequencies for the 58 populations studied are listed in Table S2 in online supplementary data. All populations were in Hardy–Weinberg equilibrium and pair-wise analysis did not detect linkage disequilibrium within the marker set. Examination of the population-group pairs in Fig. 2 indicates that the training sets make efficient proxies of each population-group for the 34 SNPs used. Generally populations showed consistent patterns of allele frequency within each group, i.e. alleles were always common, rare or fixed (confined to one population-group) to the same extent in each of the populations affiliated to the group. The fixed difference SNP rs2814778, defining the Duffy malaria resistance phenotype in Africa, gave the single exception to this pattern. In the San population sample from Namibia, Southwest Africa, all seven individuals were homozygous for the T allele unobserved in all other CEPH African populations (and rare in training set samples: T allele frequency = 0.05 in Mozambicans and Somalis combined). However, as the San individuals showed typical African allele frequencies for the other 33 SNPs each classified correctly with average assignment probabilities.

Table 1 gives the divergence values for population-group comparisons in training sets (lower left) and CEPH-HGDP (upper right) with diagonal values denoting divergence between equivalent population-groups of each sample set. The similarity of the divergences in both sets together with low values between equivalent groups further underlines the close match of the training sets used with the more broadly based CEPH-HGDP sampling.

Table 1
Divergence values for population-group comparisons in training sets (lower left: bold) and CEPH-HGDP (upper right: bold, italic)

Training set	CEPH-HGDP		
	African	European	East Asian
African	0.45	***8.99***	***5.74***
European	**7.17**	0.14	***5.11***
East Asian	**4.43**	**4.99**	0.13

Mid-diagonal values comprise divergences between the equivalent population-groups of each sample set.

The grouping of training set and CEPH-HGDP individuals using STRUCTURE runs for $K = 2$–6 are shown in Fig. 3. Training set populations produced a distinct clustering progression from $K = 2$–4 indicating a pattern that matches the regional distribution of populations. Clustering at $K = 4$ and above shows Mozambicans and Somalis are the only population pair with detectable within-group differentiation. The CEPH-HGDP STRUCTURE patterns show interesting differences to the Rosenberg study [Fig. 1 of 13] with Europe, not America, the other anchoring group to Africa at $K = 2$, progressing to a distinct East Asian cluster at $K = 3$ rather than $K = 4$. Secondly, unlike Rosenberg's pattern, at $K = 4$ and above the Eurasian populations form two discernable clusters: a well-defined European cluster (with Sardinians and Adygei showing above-average multiple group membership) and a less distinct Middle

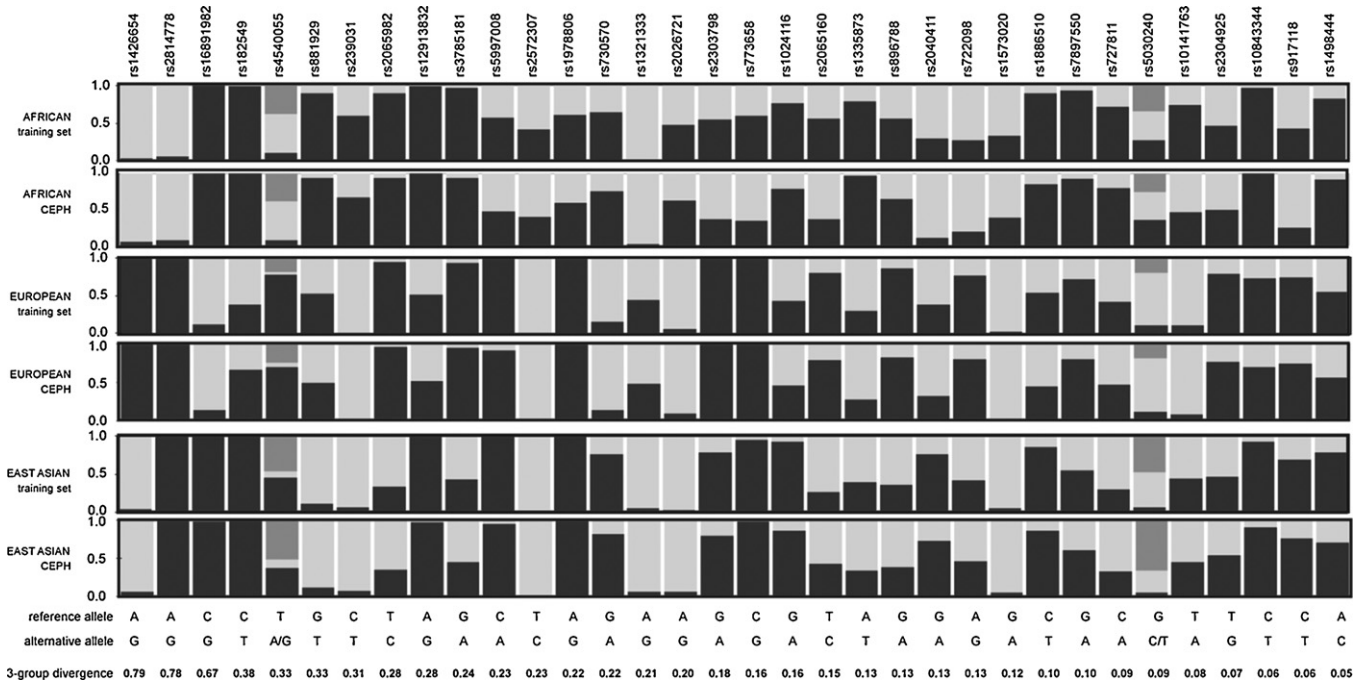

Fig. 2. Allele frequencies for 34 SNPs from CEPH and training set populations combined separately into African, European and East Asian population-groups. Each population-group is arranged as paired plots from the two sample sets. SNPs are listed left to right in descending order of three groups divergence, dark bars denote the reference alleles.

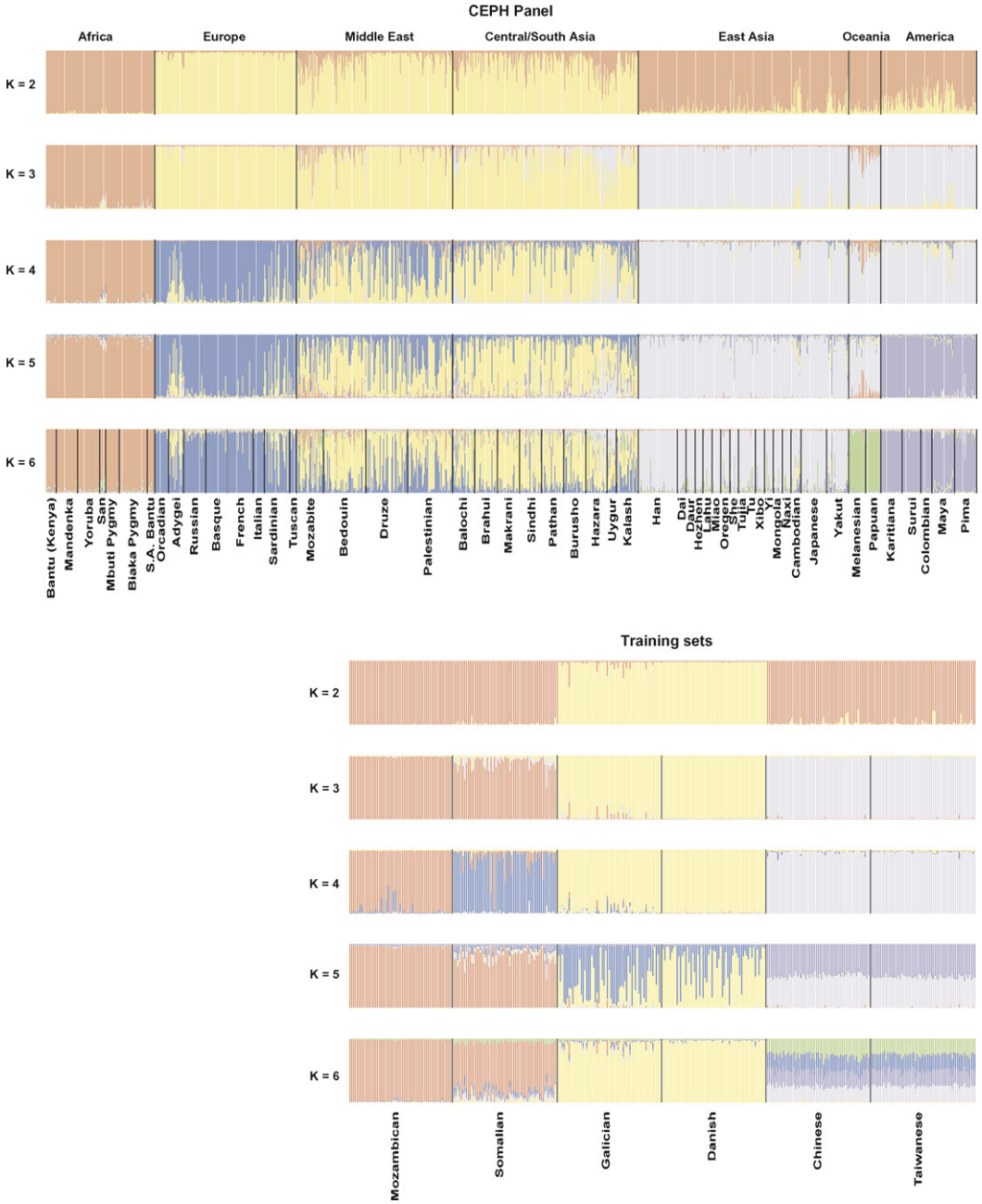

Fig. 3. Analysis of population structure in training set and CEPH-HGDP samples using grouping values of $K = 2$ to $K = 6$. CEPH populations are ordered by groups: Africa; Europe; Middle East; Central/South Asia; East Asia; Oceania and America although no pre-assigned affiliation was made.

East/Central-South Asia cluster with extensive multiple group membership, largely with Europe, but as a geographical definition would suggest, including African and East Asian affinities in Middle East and Central-South Asian populations, respectively. This Eurasian grouping pattern stays unchanged as the American and Oceanian clusters form in the same order as the Rosenberg study but a stage later at $K = 5$ and 6. Overall the STRUCTURE patterns we obtained result in a geographically based clustering of CEPH populations into five continental groups plus a less well defined Eurasian sub-grouping composed of two regions that lack barriers to gene flow with Europe: Middle East and South/Central Asia.

3.2. Performance of the classification system

The apparent misclassification rate of the ancestry test measured by assigning an ancestral origin to all training set individuals was zero. Furthermore using just the three fixed difference SNPs (in descending order of divergence): rs1426654, rs2814778 and rs16891982 were sufficient to achieve the correct classification of all 360 individuals. The less biased but more 'aggressive' misclassification estimators of one-out cross-validation and .632+ bootstrap analysis gave 0.28% and 0.18% error rates, respectively. Low misclassification rates from subsets of SNPs indicate a degree of redundancy in the marker set used and

since partial profiles are commonly produced in forensic analysis it is important to know the influence of incomplete data on classification performance. To assess how partial data could affect assignment probability we plotted the distribution of all possible likelihood ratios from modeled profiles comprising the best 6, 10, and 20 SNPs. Frequency polygon plots are given in Supplementary Fig. S1 in online supplementary data, with a left-hand reference line marking a likelihood ratio of 1:1000: a consensus confidence level. These plots indicate that using reduced SNP sets lowers the likelihood ratio levels slightly but the only probabilities falling below 1:1000 are ∼2% of European–East Asian differentiations. A complete classification of the CEPH-HGDP panel samples was made but assignment success was estimated from 123 Africans, 237 East Asians and 160 Europeans. From the 520 CEPH-HGDP samples classified five European samples were erroneously assigned comprising two Sardinians as African and two Sardinians and one Adygei Russian (West Caucasus) as East Asian. The classification probabilities obtained are represented as triangular plots (i–iii) in Fig. 4 (vertices equate to a probability of 1 and opposite sides 0 so nearly all points overlay each other at the vertex of each group). The European (ii) triangle plot shows several Sardinian and Adygei individuals with a relatively low probability to be European in addition to the misclassified individuals positioned more closely to the African and East Asian vertices. A more informative way to view atypical patterns of classification is to plot the −log likelihoods for each comparison (values that better emphasize probability differences) against the distribution of equivalent values for training sets. The lower part of Fig. 4 shows the −log likelihoods for an example CEPH European with typical likelihoods as coloured bars for assignment to European (A), East Asian (B) or African (C) origin with the equivalent training set probability distribution: on a −log scale higher probabilities are positioned on the left. Likelihood plot A in Fig. 4 also shows the percentile: the proportion of the training set with worse probabilities. For comparison the likelihood of a

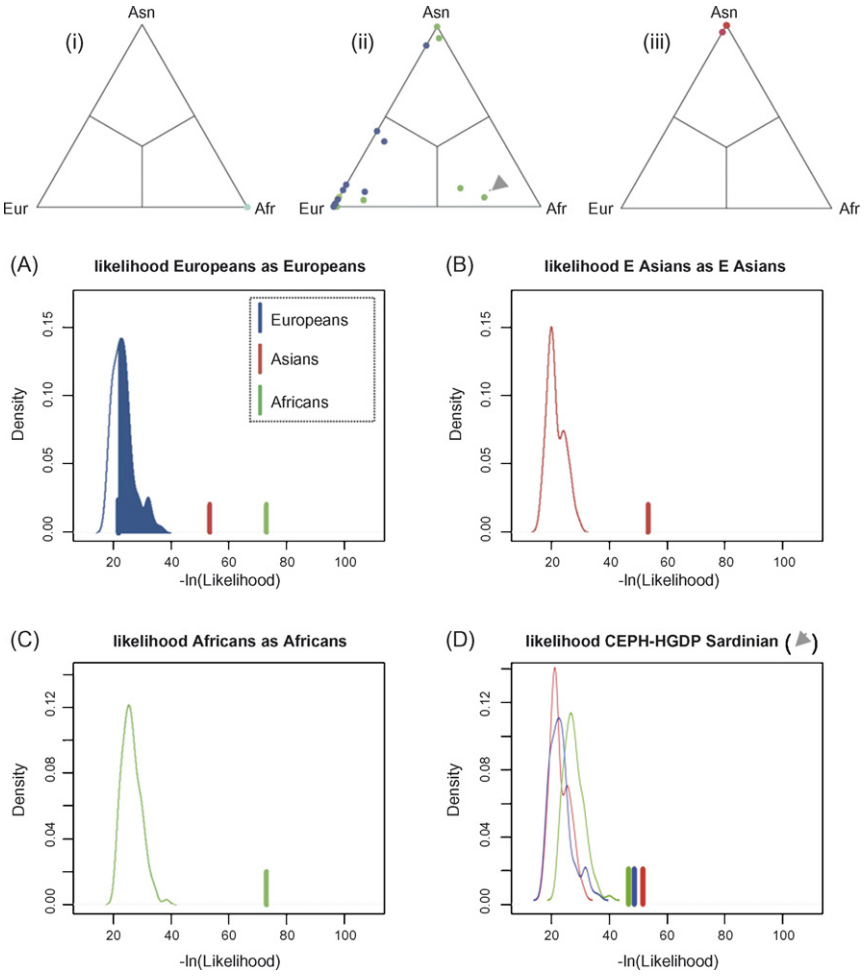

Fig. 4. Triangular plots for ancestry assignment probability of CEPH-HGDP individuals from (i) African, (ii) European and (iii) East Asian population-groups. Triangle vertices represent an assignment probability of 1 and opposite sides 0. African and East Asian plots comprise 123 and 236 of 237 points directly overlaying the correct vertex. The European plot shows 145 of 160 points overlaying the Eur vertex with outlying purple points showing Adygei individuals and green points Sardinians. Vertex: Asn = East Asian, Eur = European, Afr = African. Lower graphs display −log likelihoods (i.e. highest assignment probabilities leftmost), from an example CEPH European shown as vertical bars with re-classification probabilities of each 160-sample training set shown as density plots. Graphs A–C show individual training set plots for European, E Asian and African, respectively, plot A also indicates the percentile as a solid blue portion: the proportion of training set samples with worse assignment likelihoods than the tested individual. Graph D shows probabilities for a CEPH Sardinian individual misclassified as an African indicated in triangle (ii).

misclassified CEPH Sardinian individual are shown in plot D indicating values well below the training set distributions of each group and positioned closely together: a pattern that suggests a classification requiring caution. We intend to incorporate both a triangular plot and the relative −log likelihood plots in the classification portal output to aid interpretation.

The NIST African–American samples analyzed did not show markedly different allele frequency distributions to those seen in African populations (African–American frequencies included in Table S2, online supplementary data). This is also reflected in the range of assignment odds obtained, plotted in supplementary Supplementary Fig. S2. Each individual was successfully classified as African and 162 of 163 gave a sufficiently high probability to allow confident assignment.

4. Discussion

An ancestral origin test limited to a single multiplex can run the risk of failing to adequately differentiate the population-groups analyzed by the test. Thirty-four SNPs is an upper limit for a primer extension assay but can be extended using systems such as Genplex dye-linked oligo-ligation with proven forensic performance [30]. Nevertheless tests using small SNP numbers must still maximize allele frequency differences between groups to have any chance of success on a broad enough scale. The fixed difference SNPs and population specific SNPs that form the core of our test proved to be sufficiently powerful to make assignments with 0.28% theoretical error and 1% misclassification of new samples. Those individuals from the CEPH panel erroneously classified were not a random set but came from Sardinia: a genetic outlier in Europe or the Adygei population on the extreme eastern edge of the continent and both show higher than average multiple group membership in STRUCTURE analysis. It is important to assess in this kind of ancestry test whether limited sampling of a group to create training sets can make an efficient proxy for the actual range of variability found. Use of the three main population-group samples of the CEPH-HGDP panel to test misclassification has given us a realistic way to measure true assignment error for these groups. The misclassification rates shown were clearly low enough to give confidence that the training sets are good representatives of the three groups studied. However, studies show the geographic distribution of human variation often consists of smooth allele frequency gradients between geographically distant populations that lack geographic barriers to gene flow [10–12]. This is particularly evident between Europe and East Asia, suggesting populations from Central/South Asia are likely to show higher misclassification rates since they occupy regions in the middle of a continuum of variability resulting from free population movement. The differentiation of Eurasian population-groups is therefore certain to require a more extensive SNP set that adds more loci with sharp frequency clines between these regions such as rs16891982 and rs182549 used in our test.

Since forensic DNA analysis often involves incomplete profiles the assessment of classification probabilities from subsets of SNPs has an important bearing on the applicability of this test for analysis of challenging DNA. Modeling such profiles shows few likelihood ratios fall below 1:1000 from partial data and up to half of the SNPs are not critical for an assignment but raise the probability obtained. One factor that is more likely to reduce classification probability in routine use is the influence of population admixture. Admixture tends to reduce the differentiation of contributing population-groups and shapes the allele frequency distributions of populations occupying continental margins as well as those with recent histories of admixture [17,Chapter 12]. The classification of all 163 African–Americans as African indicates our test can be expected to successfully distinguish individuals in this population from Europeans or East Asians and the observed reduction in assignment probabilities when compared to an example African population of Yoruba from the CEPH-HGDP was small (Supplementary Fig. S2). Although we cannot gauge the overall levels of admixture in the NIST sample panel, allele frequency estimates for fixed difference SNPs: rs1426654 and rs2814778 consistently indicated a level for this sample of ∼21% suggesting assignment of majority ancestry would be a realistic approach. Only in the case of individuals of immediate mixed parentage (i.e. each parent of different ancestries) is there a real risk of erroneous assignment when balanced odds are not seen and in such cases use of additional fixed difference SNPs (with consequent atypical patterns of heterozygosity) would be the only practical system of analysis.

In addition to providing a more secure classification when differences between groups are diminished, a test with an excess of informative SNPs offers the chance to extend the scope of population-groups analyzed. This could be achieved by substituting the weakest SNPs, however, several were retained in our test since they are amongst the most informative for East Asian–European classifications. In addition finding SNPs that adequately differentiate native Americans will be a challenge given the comparatively brief time of separation of this group from East Asia as well as the levels of admixture with Europeans found in both North and South America. Although extra SNPs can help achieve differentiation on a finer scale between closely related populations, an alternative approach is to perform a simple pair-wise comparison using dedicated training sets. The web portal now permits submission of custom populations and/or new SNPs to construct tailored training sets (http://mathgene.usc.es/snipper/analysispopfile.html). Our initial studies using this approach, notably comparing Europeans and North Africans, indicates it can provide much lower misclassification rates in pair-wise differentiations.

The study presented here set out to develop a test for ancestral origin offering both a straightforward genotyping system with forensic sensitivity and a simple framework for the interpretation of results capable of handling partial data. We aimed to use, as far as possible, the most informative SNPs since clearly there are practical constraints on the genotyping of several hundred markers from limited casework material. The ability to discern ancestral origin with minimal error from partial profiles has obvious benefits in disaster victim identification where often the DNA is highly degraded and individuals form a multinational sample. In addition we have

found that autosomal SNPs provide valuable complimentary data to Y-chromosome and mtDNA typing for the analysis of the evolutionary history of populations with the advantage that small-scale samples give reliable allele frequencies. For routine use the web portal offers an intuitive and effective classification system and has the advantage of giving much faster profile analysis than a group membership approach such as STRUCTURE. For these reasons we expect the test described here to find its place amongst the current approaches available to the forensic geneticist for assignment of ancestral origin.

Acknowledgements

The African–American panel was supplied by Peter Vallone and John Butler at NIST and the authors are indebted to them for making these samples available. The work was supported by the European Commission GROWTH program, SNPforID project, contract G6RD-CT-2002-00844. Funding from Xunta de Galicia: (PGIDTIT06PXIB228195PR) and a grant from the Ministerio de Educación y Ciencia: (project BIO2006-06178) given to MVL supported this project. The 'Ramón y Cajal' Spanish programme from the Ministerio de Educación y Ciencia (RYC2005-3) and a grant from the Ministerio de Sanidad y Consumo (PI030893; SCO/3425/2002) given to AS supported this project.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.fsigen.2007.06.008.

References

[1] L.B. Jorde, W.S. Watkins, M.J. Bamshad, M.E. Dixon, C.E. Ricker, M.T. Seielstad, M.A. Batzer, The distribution of human genetic diversity: a comparison of mitochondrial, autosomal and Y-chromosome data, J. Hum. Genet. 66 (2000) 979–988.
[2] 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.
[3] 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.
[4] 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.
[5] T. Egeland, H.M. Bøvelstad, G.O. Storvik, A. Salas, Inferring the most likely geographical origin of mtDNA sequence profiles, Ann. Hum. Genet. 68 (2005) 461–471.
[6] A. Salas, M. Richards, T. De la Fé, M.V. Lareu, B. Sobrino, P. Sánchez-Diz, V. Macaulay, Á. Carracedo, The making of the African mtDNA landscape, Am. J. Hum. Genet. 71 (2002) 1082–1111.
[7] A. Salas, Á. Carracedo, M. Richards, V. Macaulay, Charting the ancestry of African–Americans, Am. J. Hum. Genet. 77 (2005) 676–680.
[8] M.A. Jobling, C. Tyler-Smith, The human Y chromosome: an evolutionary marker comes of age, Nat. Rev. Genet. 4 (2004) 598–612.
[9] 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.
[10] D. Serre, S. Pääbo, Evidence for gradients of human genetic diversity within and among continents, Genome Res. 14 (2004) 1679–1685.
[11] A. Manica, F. Prugnolle, F. Balloux, Geography is a better determinant of genetic differentiation than ethnicity, Hum. Genet. 118 (2005) 366–371.
[12] 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 (6) (2005) e70.
[13] 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.
[14] J.M. Akey, M.A. Eberle, M.J. Rieder, C.S. Carlson, M.D. Shriver, D.A. Nickerson, L. Kruglyak, Population history and natural selection shape patterns of genetic variation in 132 genes, PLoS Biol. 2 (2004) 1591–1599.
[15] G. Marth, E. Czarbarca, J. Murvai, S.T. Sherry, The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations, Genetics 166 (2004) 351–372.
[16] B.F. Voight, S. Kudaravalli, X. Wen, J.K. Pritchard, A map of recent positive selection in the human genome, PLoS Biol. 4 (2006) 446–458.
[17] M.A. Jobling, M.E. Hurles, C. Tyler-Smith, Human Evolutionary Genetics, Garland Science, New York, 2004.
[18] H.L. Norton, R.A. Kittles, E. Parra, P. McKeigue, X. Mao, K. Cheng, V.A. Canfield, D.G. Bradley, B. McEvoy, M.D. Shriver, Genetic evidence for the convergent evolution of light skin in Europeans and East Asians, Mol. Biol. Evol. 24 (2007) 710–722.
[19] M. Hamblin, E.E. Thompson, A. Di Rienzo, Complex signatures of natural selection at the Duffy blood group locus, Am. J. Hum. Genet. 70 (2002) 369–383.
[20] N.S. Enattah, T. Sahi, E. Savilahti, J.D. Terwilliger, L. Peltonen, I. Järvelä, Identification of a variant associated with adult-type hypolactasia, Nat. Genet. 30 (2002) 233–237.
[21] H.M. Cann, C. de Toma, L. Cazes, M.F. Legrand, V. Morel, L. Pioffre, J. Bodmer, W.F. Bodmer, B. Bonne-Tamir, et al., (32 additional co-authors) A human genome diversity cell line panel, Science 296 (2002) 261–262.
[22] J.K. Pritchard, M. Stephens, P. Donnelly, Inference of population structure using multilocus genotype data, Genetics 155 (2000) 945–959.
[23] C. Phillips, M.V. Lareu, A. Salas, M. Fondevilla, G. Berniell-Lee, Á. Carracedo, N. Morling, P. Schneider, D. Syndercombe Court, Population specific single-nucleotide polymorphisms. Progress in Forensic Genetics 10, International Congress Series 1261, 2004, pp. 233–235.
[24] M.D. Shriver, M.W. Smith, L. Jin, A. Marcini, J.M. Akey, R. Deka, R.E. Ferrell, Ethnic-affiliation estimation by use of population-specific DNA markers, Am. J. Hum. Genet. 60 (1997) 957–964.
[25] C. Phillips, M.V. Lareu, A. Salas, Á. Carracedo, Non-binary single-nucleotide polymorphism markers. Progress in Forensic Genetics 10, International Congress Series 1261, 2004, pp. 27–29.
[26] J.J. Sánchez, C. Phillips, C. Børsting, K. Balogh, M. Bogus, M. Fondevila, C.D. Harrison, E. Musgrave-Brown, A. Salas, D. Syndercombe-Court, P. Schneider, Á. Carracedo, N. Morling, A multiplex assay with 52 single nucleotide polymorphisms for human identification, Electrophoresis 27 (2006) 1713–1724.
[27] I. Grosse, P. Bernaola-Galván, P. Carpena, R. Román-Roldán, J. Oliver, H. Eugene Stanley, Analysis of symbolic sequences using Jensen–Shannon divergence, Phys. Rev. E 65 (2002) 041905.
[28] R.P. Duda, P. Hart, D. Stork, Pattern Classification, Wiley, London, 2001.
[29] B. Efron, R. Tibshirani, Improvement on cross-validation: the .632+ bootstrap method, J. Am. Stat. Assoc. 92 (1997) 316–330.
[30] C. Phillips, R. Fang, D. Ballard, M. Fondevila, C. Harrison, F. Hyland, E. Musgrave-Brown, C. Proff, E. Ramos-Luis, B. Sobrino, M. Furtado, D. Syndercombe Court, Á. Carracedo, P.M. Schneider, The SNP*for*ID consortium, evaluation of the Genplex SNP typing system and a 49-plex forensic marker panel, Forensic Sci. Int. Genet. 1 (2007) 180–185.