Int J Legal Med (2016) 130:341–352
DOI 10.1007/s00414-015-1297-9

ORIGINAL ARTICLE

Selection of highly informative SNP markers for population affiliation of major US populations

Xiangpei Zeng[1] · Ranajit Chakraborty[1] · Jonathan L. King[1] · Bobby LaRue[1] · Rodrigo S. Moura-Neto[2,3] · Bruce Budowle[1,4]

Received: 1 September 2015 / Accepted: 23 November 2015 / Published online: 8 December 2015
© Springer-Verlag Berlin Heidelberg 2015

Abstract Ancestry informative markers (AIMs) can be used to detect and adjust for population stratification and predict the ancestry of the source of an evidence sample. Autosomal single nucleotide polymorphisms (SNPs) are the best candidates for AIMs. It is essential to identify the most informative AIM SNPs across relevant populations. Several informativeness measures for ancestry estimation have been used for AIMs selection: absolute allele frequency differences (δ), F statistics (F_{ST}), and informativeness for assignment measure (In). However, their efficacy has not been compared objectively, particularly for determining affiliations of major US populations. In this study, these three measures were directly compared for AIMs selection among four major US populations, i.e., African American, Caucasian, East Asian, and Hispanic American. The results showed that the F_{ST} panel performed slightly better for population resolution based on principal component analysis (PCA) clustering than did the δ panel and both performed better than the In panel. Therefore, the 23 AIMs selected by the F_{ST} measure were used to characterize the four major American populations. Genotype data of nine sample populations were used to evaluate the efficiency of the 23-AIMs panel. The results indicated that individuals could be correctly assigned to the major population categories. Our AIMs panel could contribute to the candidate pool of AIMs for potential forensic identification purposes.

Keywords Ancestry informative markers (AIMs) · Single nucleotide polymorphisms (SNPs) · Population differentiation · HapMap · 1000 Genomes · F_{ST}

Introduction

Ancestry informative markers (AIMs) are genetic makers that show large differences in allele frequencies between human populations [1–4]. These differences allow determination of population affiliation and apportionment of ancestry and can be used to detect and adjust for population stratification in genome-wide disease-gene association studies. Moreover, AIMs can play a role in ancestry inference to support investigative leads from forensic genetic evidence [5–7]. The value of AIMs in a forensic investigation is that these markers may provide critical evidence about the source of an evidence sample or about the ancestry of unidentified human remains. Ancestry information may help to narrow the range of suspects and thus make better use of limited investigative resources.

There are four types of genetic markers that could provide ancestry information: mitochondrial DNA (mtDNA), Y chromosome markers, autosomal short tandem repeats (STRs), and single nucleotide polymorphisms (SNPs). Lineage markers (Y-linked and mtDNA haplotypes) have proven effective in studying human migration and evolutionary

Electronic supplementary material The online version of this article (doi:10.1007/s00414-015-1297-9) contains supplementary material, which is available to authorized users.

✉ Xiangpei Zeng
Xiangpei.Zeng@live.unthsc.edu

[1] Institute of Applied Genetics, Department of Molecular and Medical Genetics, University of North Texas Health Science Center, 3500 Camp Bowie Blvd., Fort Worth, TX 76107, USA

[2] Instituto de Biologia, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil

[3] Instituto Nacional de Metrologia, Qualidade e Tecnologia, Duque de Caxias, Brazil

[4] Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah, Saudi Arabia

histories across the world [8–10]. However, due to uniparental inheritance and lack of recombination, their utility for population affinity inferences is not comprehensive. Further, because of uniparental ancestry of these markers, the contributions of the majority of an individual's genome are not assessed. STRs typically are highly polymorphic, and a relatively small panel of markers can successfully distinguish an individual from others, excluding identical twins. However, autosomal STRs are limited for ancestry inferences, because the majority of common alleles of STRs are shared among human populations, and STRs have a relatively high mutation rate [11]. Contraction-expansion pattern of mutations at STR loci also imply that STR alleles of the same repeat size may not all be identical by descent [12]. In spite of these, some panels of STR markers have been shown to distinguish African Americans, Hispanics, European Americans, and Asians to some degree [13]. In contrast, SNPs have a relatively low mutation rate; the same SNP allele at most genomic location is often identical by descent, and millions of human SNPs are available in public databases, e.g., SNP database, International HapMap project, and 1000 Genomes [14–16]. Thousands of SNPs with different allele frequencies between populations can be selected for ancestry and human population affinity studies. Therefore, autosomal SNPs are recognized as the best candidates for AIMs. Indeed, several SNP panels have been developed for potential application of ancestral inference in forensic genetics [17–21].

An ideal AIM SNP would have one allele fixed in one population and be completely absent in another population. However, the majority of alleles are shared to some degree between or among populations. It is essential to identify the most informative AIM SNPs across relevant populations. Several marker informativeness measures for ancestry estimation have been applied for selection of AIMs. These measures include absolute allele frequency differences (δ) [22], F statistics (F_{ST}) [23], and informativeness for assignment measure (In) [22]. Some theoretical as well as empirical studies compared the effectiveness of these alternative measures of informativeness for ancestry determination [22, 24]. Various studies have used these different measures to select AIMs [18–21]. While the logic of using these measures is similar, their efficacy has not been compared with objective selections of genome-wide SNPs, particularly for determining affiliations for major US populations. With an abundance of SNPs in International HapMap project and 1000 Genomes, it is possible to select an informative minimal number panel of AIMs and compare whether any of these measures are better for discovery of such efficient panels of AIMs. Therefore, the objective of this study was to select the most informative AIMs using the three measures (δ, F_{ST}, and In) that resolve pairs of major populations and identify a robust panel of AIMs that could characterize the four major US populations (e.g., African American, Caucasian, East Asian, and Hispanic American). To date, there are no agreed upon core AIMs for forensic use. Therefore, these additional SNPs are provided to support AIMs panel development.

Materials and methods

Population samples

The HapMap project [15] contains comprehensive SNP data on the four major US populations: African ancestry from Southwest USA (ASW), Utah residents with Northern and Western European ancestry (CEU), Chinese from Metropolitan Denver, Colorado (CHD), and Mexican ancestry from Los Angeles, California (MEX). The samples included in the HapMap project are family duos and trios. The children were removed, and only unrelated parents were used in the study. From the HapMap Phase III, genotype data were available for 52, 120, 85, and 50 unrelated individuals from ASW, CEU, CHD, and MEX, respectively (http://hapmap.ncbi.nlm.nih.gov/downloads/genotypes/2010-08_phaseII+III/).

AIMs selection

The measures used for AIMs selection were δ, F_{ST}, and In. The candidate AIMs were selected in three steps. First, the three measure values of each SNP were computed for each pairwise population comparison, and then markers were ranked based on these measures from highest to lowest of their values. These pairwise measures were calculated using AncestrySNPminer (https://research.cchmc.org/mershalab/AncestrySNPminer/home.php) [25]. Second, the top 30 informative markers for each measure in each pairwise population comparison were chosen. GDA v1.1 [26] was used to test for departures from Hardy-Weinberg equilibrium and linkage disequilibrium (LD) of these top 30 AIMs in each pairwise population comparison. The minimum number of markers, for each measure, to discriminate each pair of populations was identified based on principal component analysis (PCA) using the EIGENSOFT v6.0.1 [27] and receiver operating characteristics curve (ROC curve) [28]. Finally, the top markers from six pairwise population comparisons were pooled based on the three measures and evaluated as individual panels.

Statistical power of AIMs

The number of AIMs, which was assessed to distinguish the two populations, was increased from 1 to 30 with increments of 1, starting with the most informative SNP and then sequentially adding the next most informative SNP. The changes of

PCA clusters were examined. The PCA clustering performances of these AIMs in individual classification were assessed using the maximum Matthews correlation coefficient (MCC) [29]:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

where TP and FP are the amount of true positives and false positives, respectively, and TN and FN represent the amount of true negatives or false negatives, respectively. Two populations were determined to be completely separated with the dataset when MCC reaches one. The ROC curve is constructed by plotting the true positive rate against the false positive rate at different cutoff values. The cutoff values of PC1 were determined by using the ROC curve. This curve is a graphical plot that demonstrates the performance of a binary classifier system with different discrimination thresholds. ROC curve analyses were performed using the XLSTAT software [30]. The Bayesian clustering algorithm (STRUCTURE) [31] was used to estimate ancestry and individual admixture proportions. Discriminant function analysis (DFA) is a statistical method to predict category membership by a set of independent variables [32]. In this study, DFA using SPSS v16.0 [33] was used to provide a probability of population assignment for each individual sample.

Results and discussions

AIMs selection

Three measures (δ, F_{ST}, and In) were used for AIMs selection in the four major American populations. Of the millions of SNPs existing in the SNP databases, there were 1318288, 1232531, 1369287, 1211787, 1307348, and 1221276 SNPs available for comparisons of ASW and CEU, ASW and CHD, ASW and MEX, CEU and CHD, CEU and MEX, and CHD and MEX, respectively. Values of the three measure of each SNP were computed and markers were ranked for each pairwise population comparison. The same SNP may be selected by different measures but could be ranked differently. In order to avoid strong LD, the minimal physical distance of any two SNPs located on the same chromosome was set initially at 100 kb. The top 30 AIMs for each measure in each pairwise population comparison were chosen.

Among the four populations (ASW, CEU, CHD, and MEX), there were no detectable departures from Hardy-Weinberg equilibrium expectations for the selected SNPs. A few SNP pairs did display LD (Supplemental Tables 1, 2, 3, 4, 5, and 6). In those instances where two markers were in LD, the more informative one was selected and the less informative one was deleted. For example, rs1288097 and rs12594483 were in LD in ASW and CEU; rs1288097 was selected (the second most informative marker) but rs12594483 was deleted (the third most informative marker) (Supplemental Table 1). Therefore, the top 30 candidate SNPs were reduced to less than 30 AIMs in all population pairs. For example, the top 30 SNPs were reduced to 26, 24, and 22 AIMs by δ, F_{ST}, and In, respectively, in CEU and MEX (Supplemental Table 5). In order to determine the minimum number of SNPs to separate the paired populations, the candidate AIMs were increased in increments of 1 starting from the most informative SNP. Maximum MCC was used to evaluate the PCA clustering performance of the selected AIMs for individual classification. The minimum numbers of markers to distinguish any two populations were identified, and the results were listed in Supplemental Table 7. The number of AIMs needed to resolve any of the six population pairs ranged from two to nine SNPs. As expected, CEU and MEX needed the largest number of SNPs to be separated. Maximum MCC curves showed that at least eight AIMs were required to distinguish CEU and MEX for δ and F_{ST} measures (MCC=1), while the MCC value of the In measure reached one at nine AIMs (Fig. 1). Figure 2a shows classification accuracy of 170 samples (CEU and MEX) utilizing a different number of AIMs that were selected by the δ measure. The MCC value increased with the increment of AIMs, and the value reached one when the top eight informative AIMs were used (Fig. 2a). In addition, PCA clusters showed that CEU was generally distinguished from MEX individuals using the genotype data of these eight AIMs (Fig. 2b). However, CEU and MEX could not be completely resolved, due to the known Caucasian admixture component in MEX. Indeed, some MEX individuals may never be resolved from CEU or from African or Native American populations because of their large individual-specific admixture components [34–36].

Comparison of the three measures

Each of three measures selected 25 total markers to characterize the four major American populations (in pairwise comparisons) (Supplemental Table 7). In the δ panel of markers, rs4429562 was shared by CEU and CHD, and CHD and MEX comparisons, so this marker was counted once. Two pairs of SNPs were in LD: rs6674304 and rs12087334, rs974627 and rs469471. One of them, rs6674304, was the third most informative marker between ASW and CEU, while rs12087334 was the most informative marker between ASW and MEX. In order to achieve the best separation for the overall panel, rs12087334 was selected and rs6674304 was replaced by rs7689609 (the fourth informative marker between ASW and CEU). After replacement, the PCA clusters showed that the three AIMs (rs1834640, rs1288097, and rs7689609) still were able to resolve ASW and CEU. Markers rs974627 and rs469471 were in LD,

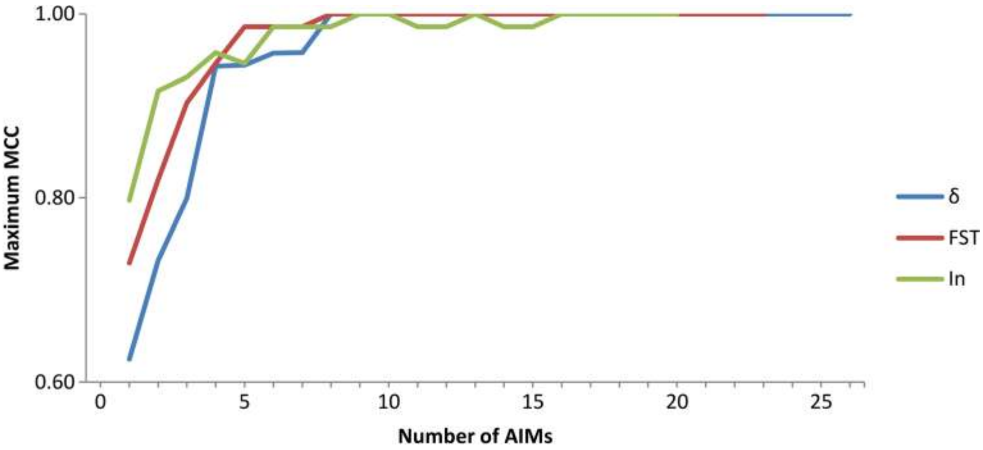

Fig. 1 The three MCC curves generated by δ, F_{ST}, and In measures for MEX and CEU

although they were located on different chromosomes. While this departure is not explained by synteny and could be due to chance, to attain good separation between CEU and MEX, rs974627 was selected and rs469471 was replaced by rs1761031 (the fifth informative marker between CHD and MEX). After replacement, the four markers (rs4429562, rs6500380, rs8032157, and rs1761031) still could distinguish CHD and MEX. In the F_{ST} panel, rs1834640 and rs4429562 were shared by two pairwise comparisons and therefore only counted once; rs974627 and rs469471 were in LD, so rs469471 was replaced by rs1761031. In the In panel, rs1834640 and rs4429562 were informative in two population pairs and only counted once; rs6674304 and rs12087334 were in LD, and rs6674304 was replaced by rs1572510. The resultant total number of markers in the AIMs panels selected by δ, F_{ST} and In was 24, 23, and 23, respectively (Table 1). Twenty-two of 23 AIMs in the F_{ST} panel were also in the δ panel, with a similarity rate of 0.95 (Table 2). The similarity rates of δ and In (16 of 23 SNPs in common) and F_{ST} and In (17 of 23 SNPs in common) were 0.70 and 0.74 (Table 2). Although not substantially different, the PCA cluster results of the F_{ST} panel appeared to perform slightly better than the δ panel (Fig. 3a, b). Only two MEX individuals clustered with the CEU group, and no CEU individuals clustered with the MEX group. Both the δ and F_{ST} panels performed better than the In panel, in which some MEX individuals cannot be distinguished between CEU and CHD (Fig. 3c). The correlation coefficients of PC1 and PC2 between δ and F_{ST} panels were 0.997 and 0.996, respectively, while the correlation coefficients of between δ and In, F_{ST}, and In were much lower (Supplemental Table 8). The statistical results indicated that δ and F_{ST} panels generated more similar results compared with In panel. In addition, the F_{ST} panel had one fewer SNP, so the 23 AIMs selected by the F_{ST} measure were used to characterize the four major American populations.

STRUCTURE was used to examine the full set of 23 AIMs with population clusters (K) increasing from 2 to 10, and ten

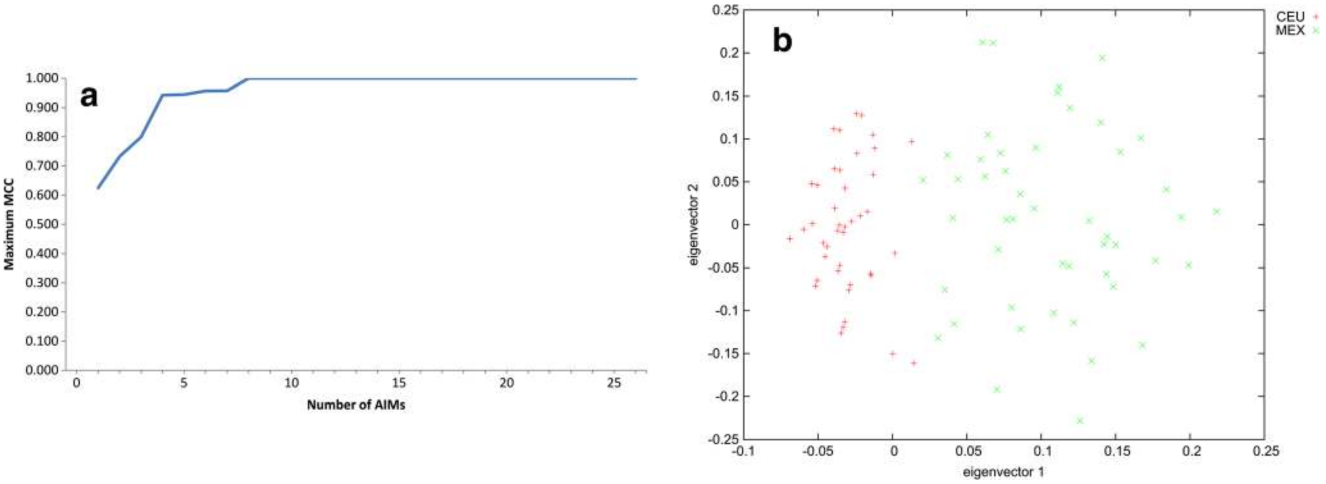

Fig. 2 The AIMs panel that was selected by the δ measure to separate CEU and MEX. **a** The classification accuracy of 170 samples (CEU and MEX) utilizing a varied number of selected AIMs; **b** PCA clusters of two populations by using genotype data of top eight AIMs identified in **a**

Table 1 The final panels of AIMs identified by the three measures δ, F_{ST}, and In to distinguish the four major US populations. The 23 AIMs selected by the F_{ST} measure were used to characterize the four major American populations. The physical distances of SNPs were downloaded from GRCh37.p13 (hg 19)

δ				F_{ST}				In			
SNPs	Chr	Pos	Populations	SNPs	Chr	Pos	Populations	SNPs	Chr	Pos	Populations
rs1834640	15	48392165	ASW_CEU	rs1834640	15	48392165	ASW_CEU	rs1834640	15	48392165	ASW_CEU
rs1288097	15	45141373	ASW_CEU	rs1288097	15	45141373	ASW_CEU	rs1288097	15	45141373	ASW_CEU
rs7689609	4	72083374	ASW_CEU	rs7689609	4	72083374	ASW_CEU	rs1572510	13	105381134	ASW_CEU
rs7165971	15	55921013	ASW_CHD	rs7165971	15	55921013	ASW_CHD	rs7165971	15	55921013	ASW_CHD
rs745767	2	177825415	ASW_CHD	rs745767	2	177825415	ASW_CHD	rs745767	2	177825415	ASW_CHD
rs13021399	2	109006665	ASW_CHD	rs13021399	2	109006665	ASW_CHD	rs13021399	2	109006665	ASW_CHD
rs12087334	1	116887455	ASW_MEX	rs12087334	1	116887455	ASW_MEX	rs12087334	1	116887455	ASW_MEX
rs12149261	16	70998145	ASW_MEX	rs12149261	16	70998145	ASW_MEX	rs11845995	14	105930923	ASW_MEX
rs1827950	4	117098482	ASW_MEX	rs11845995	14	105930923	ASW_MEX	rs12149261	16	70998145	ASW_MEX
rs11845995	14	105930923	ASW_MEX	rs1827950	4	117098482	ASW_MEX	rs1827950	4	117098482	ASW_MEX
rs4429562	22	42892596	CEU_CHD	rs4429562	22	42892596	CEU_CHD	rs4429562	22	42892596	CEU_CHD
rs1547843	10	91738263	CEU_CHD	rs11126303	2	26173503	CEU_CHD	rs10510511	3	21260370	CEU_MEX
rs11126303	2	26173503	CEU_CHD	rs7134749	12	50237637	CEU_MEX	rs2700372	3	123633220	CEU_MEX
rs11725412	4	38277754	CEU_MEX	rs10510511	3	21260370	CEU_MEX	rs7134749	12	50237637	CEU_MEX
rs10962599	9	16795286	CEU_MEX	rs11725412	4	38277754	CEU_MEX	rs7404672	16	10966479	CEU_MEX
rs7134749	12	50237637	CEU_MEX	rs2700372	3	123633220	CEU_MEX	rs11725412	4	38277754	CEU_MEX
rs11139346	9	84241442	CEU_MEX	rs11139346	9	84241442	CEU_MEX	rs4729955	7	103677151	CEU_MEX
rs10510511	3	21260370	CEU_MEX	rs4729945	7	103677151	CEU_MEX	rs715846	9	95273013	CEU_MEX
rs974627	12	38919524	CEU_MEX	rs10962599	9	16795286	CEU_MEX	rs6836368	4	130751286	CEU_MEX
rs10141733	14	101142651	CEU_MEX	rs974627	12	38919524	CEU_MEX	rs9307388	4	114075688	CEU_MEX
rs2700372	3	123633220	CEU_MEX	rs6500380	16	48375777	CHD_MEX	rs6500380	16	48375777	CHD_MEX
rs6500380	16	48375777	CHD_MEX	rs8032157	15	64480888	CHD_MEX	rs8032157	15	64480888	CHD_MEX
rs8032157	15	64480888	CHD_MEX	rs1761031	14	46926398	CHD_MEX	rs469471	21	14838552	CHD_MEX
rs1761031	14	46926398	CHD_MEX								

runs were performed at each value of K. All STRUCTURE runs were performed without using any prior population information. CLUMPP software was used to combine ten STRUCTURE runs for a particular value of K (K=4) and compute the average cluster membership values [37]. The optimal number of K was determined to be 4 (Fig. 4a). The average cluster assignment values of the optimal K (K=4) was used in the Distruct program to generate the STRUCTURE graph [38]. Individuals of CEU and CHD were more homogenous compared with ASW and MEX individuals, in which some individuals have demonstrated admixture of Caucasian SNPs (Fig. 4b).

Table 2 Shared number of AIMs between δ, F_{ST}, and In among the top two to nine markers for six pairs of population comparisons

Population comparisons	Number of markers shared		
	δ and F_{ST}	δ and In	F_{ST} and In
ASW and CEU	3	2	2
ASW and CHD	3	3	3
ASW and MEX	4	4	4
CEU and CHD	2	1	1
CEU and MEX	7	3	5
CHD and MEX	3	3	2

Evaluation of AIMs panel

In order to evaluate the efficiency of the 23-AIMs panel, the genotype data of nine populations (not used for selecting the AIMs) were downloaded from HapMap [15] and 1000 Genomes [16] databases. Four populations from the HapMap project were used: Yoruba from Ibadan, Nigeria (YRI); Toscans from Italy (TSI); Han Chinese from Beijing, China (CHB); and Japanese from Tokyo, Japan (JPT). Individuals without genotype data of three or more SNPs from this panel were excluded. There were 53 YRI, 82 TSI, 79 CHB, and 42 JPT unrelated individuals available for the evaluation study. In PCA clusters, the test samples that fell within

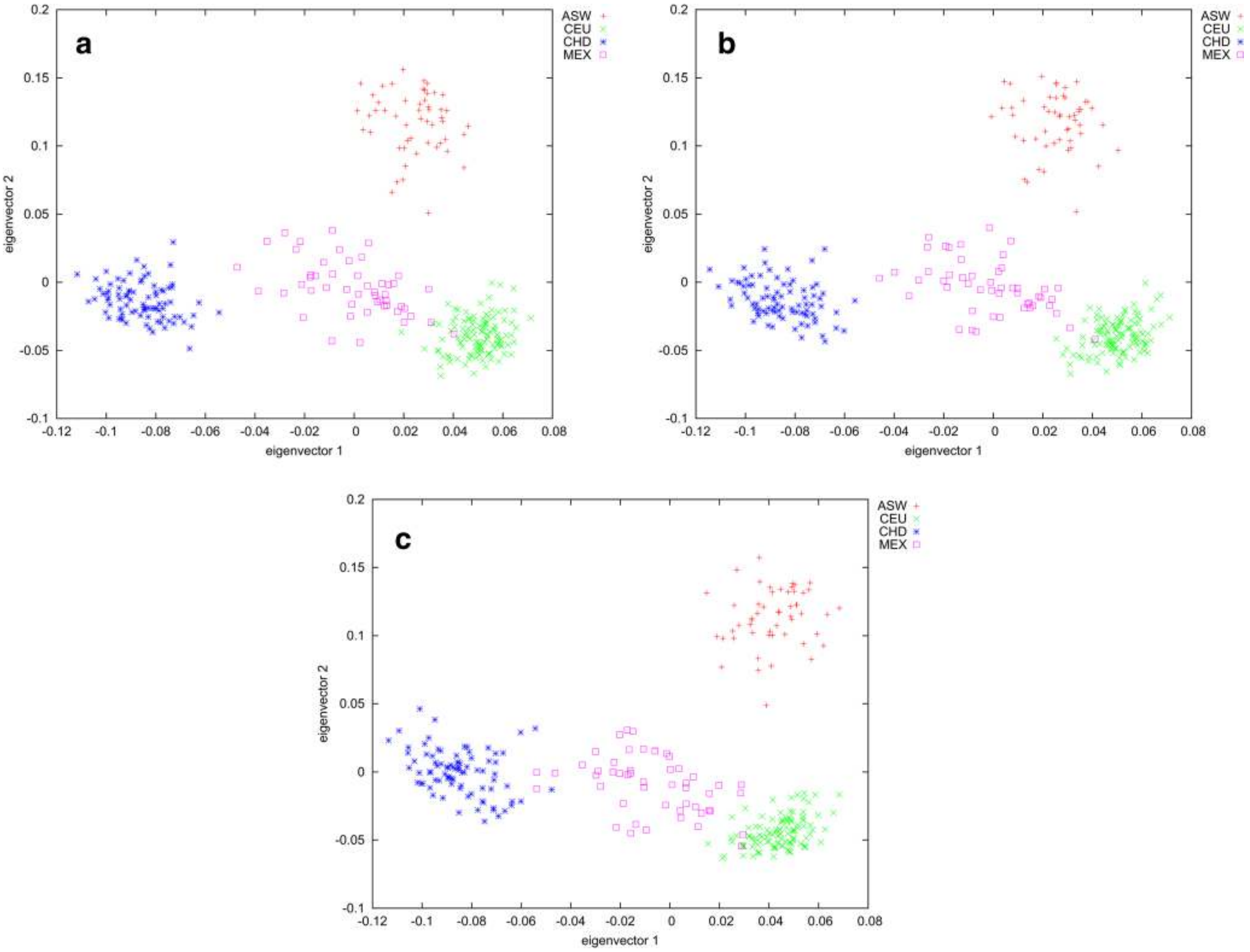

Fig. 3 The PCA clusters of the AIMs panels that were selected by **a** δ, **b** F_{ST}, and **c** In measures, respectively

the 95 % confidence interval of one of the four reference populations were classified as belonging to that reference population. DFA was used to provide a probability of assignment of an individual sample with one or more of the reference populations, especially those that did not fall within the 95 % confidence interval of a reference population. Of the 23 SNPs, HapMap does not provide genotype data of rs10510511 for ASW and of rs10962599 for CHD. In PCA, 23 AIMs can be used simultaneously to predict ancestry of known populations (YRI, TSI, CHB, and JPT) based on four

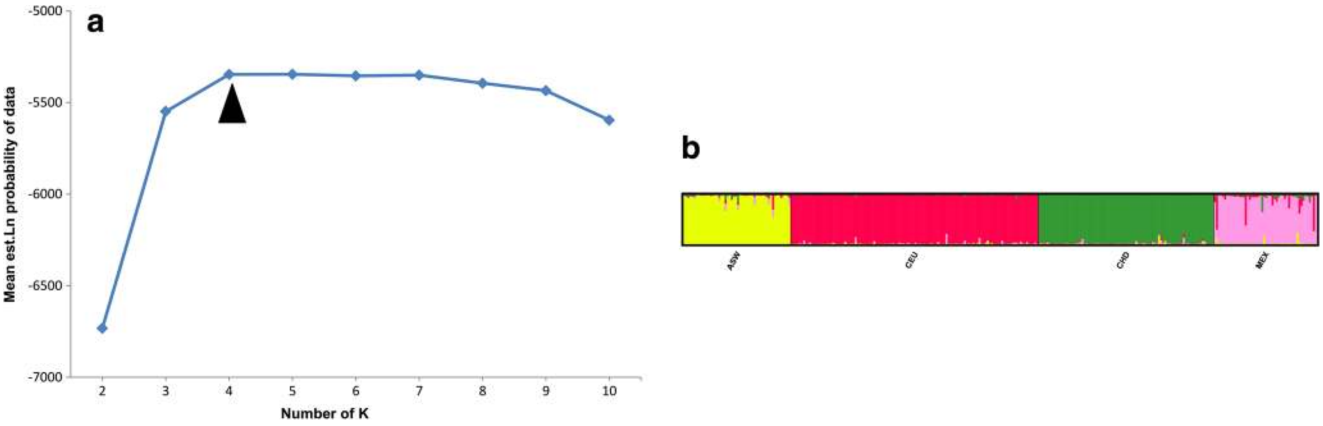

Fig. 4 Analyses of four major US populations from HapMap using the AIMs panel selected by F_{ST}. **a** Indicated that the optimal number of K was 4. **b** The STRUCTURE cluster plots of four populations (ASW, CEU, CHD, and MEX)

reference populations (ASW, CEU, CHD, and MEX) and missing data are tolerated in this method. However, only 21 AIMs (without rs10510511 and rs10962599) could be used in DFA for each population assignment, because, unlike PCA, this method requires genotype data on all loci for each individual. Approximately 92 % of YRI individuals fell within the 95 % confidence interval of ASW in PCA clusters (Fig. 5a). The DFA results assigned all YRI individuals to ASW group (Fig. 6a, Supplemental Table 9). YRI individuals likely do not have substantial Caucasian admixture compared with African Americans and yet clustered with ASW. A portion (30 %) of TSI samples (Northern Italy) fell outside the 95 % confidence interval of CEU in PCA, but they could be considered similar to Caucasian or Hispanic American and not African American and East Asian (Fig. 5b). TSI individuals do not have genotype information for rs1834640, so three SNPs were removed for DFA (rs1834640, rs10510511, and rs10962599). The results assigned all TSI individuals to CEU (Fig. 6b, Supplemental Table 9). In the AIMs selection, Chinese from Metropolitan Denver, Colorado (CHD), were used to represent the East Asian population. The majority (94 % and 81 %) of CHB and JPT, respectively, individuals fell within the 95 % confidence interval of CHD in PCA clusters (Fig. 5c, d). Five CHB individuals and eight JPT individuals were outside that of CHD. These 13 samples still would be considered as East Asians, because they were comparatively more isolated from the other major populations in the PCA clusters. HapMap does not provide genotype data of rs11845995 for JPT, so only 20 SNPs were used in DFA to predict the ancestry of JPT individuals (rs11845995, rs10510511, and rs10962599 were removed). The DFA results assigned all CHB and JPT individuals to the East Asian group (Fig. 6c, d, Supplemental Table 9). Five populations from 1000 Genomes also were used in the evaluation study: Yoruba from Ibadan, Nigeria (YRI); British in England and Scotland (GBR); Han Chinese from Beijing, China (CHB); Colombians from Medellin, Colombia (CLM); and Mexican Ancestry from Los Angeles, USA (MEX). There were 108 YRI, 91 GBR, 103 CHB, 94 CLM, and 17 MEX unrelated individuals. 1000 Genomes does not provide genotype data for rs12149261. Twenty-

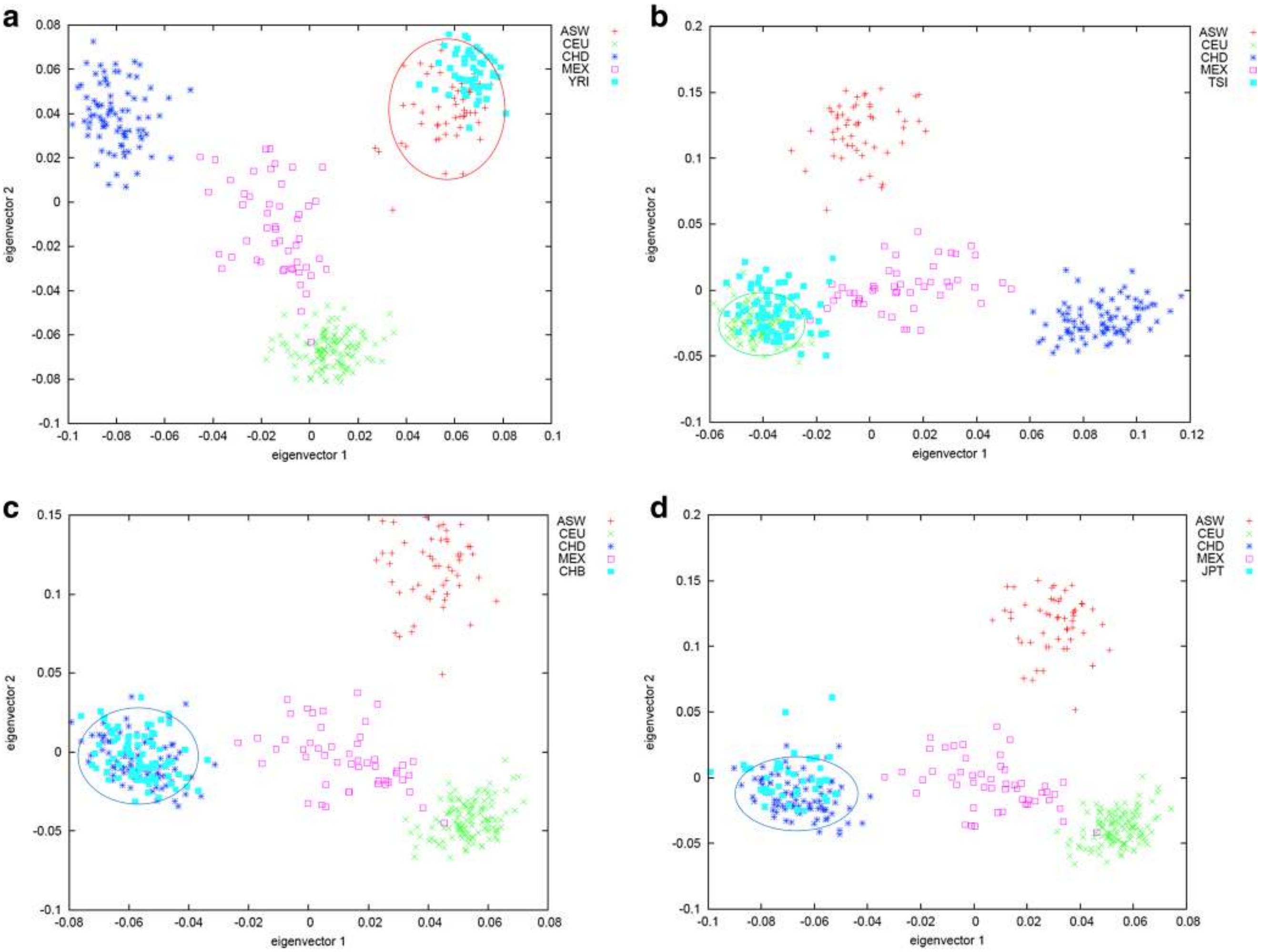

Fig. 5 Population classification of four global populations from HapMap using PCA. **a**–**d** represented YRI, TSI, CHB, and JPT, respectively

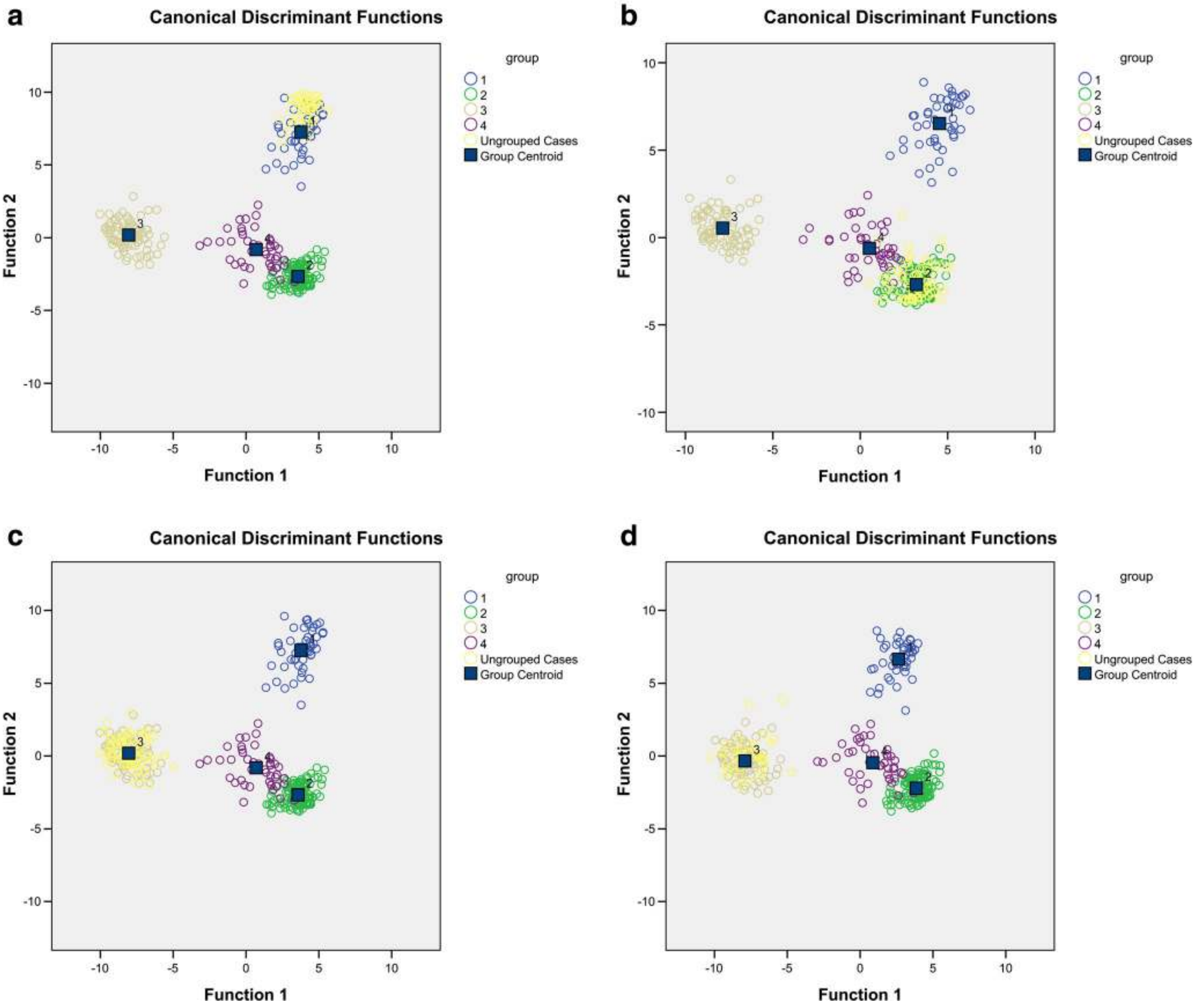

Fig. 6 Population classification of four major populations from HapMap using DFA. Groups 1–4 represented ASW, CEU, CHD, and MEX, respectively. The ungrouped cases in **a**–**d** were individuals of YRI, TSI, CHB, and JPT, respectively. Some SNPs were excluded from the analysis because of missing data. Overall, 21, 20, 21, and 20 AIMs were used in **a**–**d**, respectively

three SNPs could be used in PCA to predict ancestry of YRI, GBR, CHB, CLM, and MEX individuals, but only 20 SNPs were used in DFA (rs12149261, rs10510511, and rs10962599 were removed). YRI individuals clustered better than African Americans and not cluster with the other three major populations. Therefore, they were classified as African Americans in both PCA and DFA (Figs. 7a and 8a, Supplemental Table 10). The majority of GBR individuals were located within the 95 % confidence interval of the Caucasian group in PCA (Fig. 7b), and all of them were assigned as Caucasians by DFA (Fig. 8b, Supplemental Table 10). Eight CHB individuals fell outside the 95 % confidence interval of CHD in PCA (Fig. 7c), but all of them were assigned as East Asians in DFA (Fig. 8c, Supplemental Table 10). CLM individuals were the most difficult to assign. They were classified as African Americans, Caucasians, and Hispanic Americans (Fig. 7d). According to Bushnell et al. [39], 86 % of Columbians are mestizo and white, 10 % are black. The majority of CLM individuals were classified as Hispanic Americans or Caucasians, and up to four samples could be considered as African Americans in PCA (Fig. 7d). The DFA provided results of 4, 26, and 64 individuals assigned as African Americans, Caucasians, and Hispanic Americans, respectively (Fig. 8d, Supplemental Table 10). The ancestry of each Colombian individual was not provided by 1000 Genomes. Therefore, population assignment is difficult for CLM. In addition, the Mexican population (MEX) only represents the Hispanic population in US and may not precisely explain the genetic variations of the Hispanic populations in Central America and South America. Both HapMap and 1000

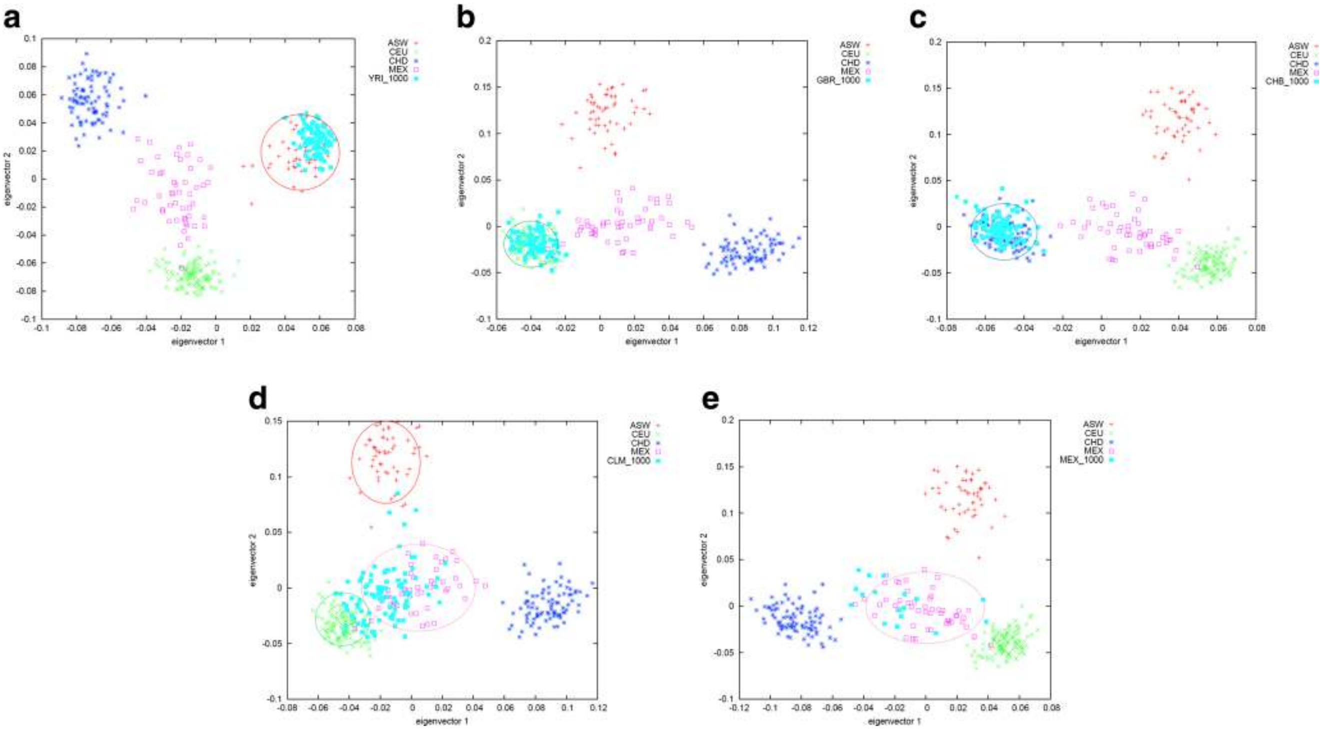

Fig. 7 Population classification of five populations from 1000 Genomes using PCA. **a**–**e** represented YRI, GBR, CHB, CLM, and MEX, respectively

Genomes databases contain samples of Mexican Ancestry from Los Angeles, USA (MEX). There were only 17 samples included in 1000 Genomes that were not used in our AIMs selection (based on HapMap data). Twelve out of 17

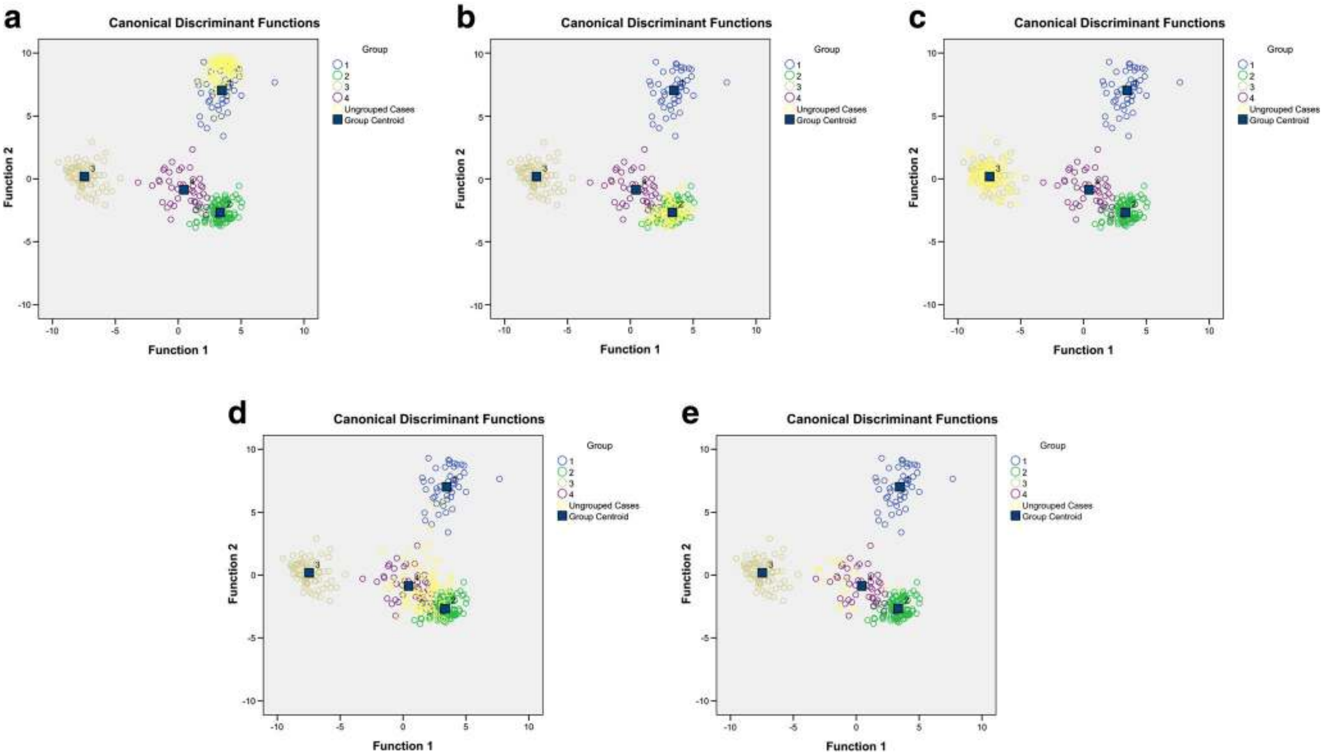

Fig. 8 Population classification of five populations from 1000 Genomes using DFA. Groups 1–4 represented ASW, CEU, CHD, and MEX, respectively. The ungrouped cases in **a**–**e** were individuals of YRI, GBR, CHB, CLM, and MEX, respectively. Three SNPs were excluded from the analysis because of missing data

individuals were within the 95 % confidence interval of Hispanic American in PCA (Fig. 7e). All individuals were classified as Hispanic Americans by DFA (Fig. 8e, Supplemental Table 10).

Overall, the results indicated that these 23 AIMs can correctly assign individuals to the major population categories. However, these public databases only provide the genotype data of 20 or 21 AIMs for each population and thus the full power of the 23-AIMs panel could not be evaluated. A future study will develop an in-house 23-AIMs panel to generate data on samples from four major US populations. Therefore, empirical testing of the full set of these AIMs will further evaluate the efficiency of the panel.

Summary of several AIMs panels

Several AIMs panels have been described for potential forensic application (Supplemental Table 11). Two large panels were developed by Kosoy et al. [18] and Halder et al. [40] to characterize seven and four populations, respectively. Nievergelt et al. [20] used In measure to select 41 AIMs to distinguish populations from seven continental regions (Africa, the Middle East, Europe, Central/South Asia, East Asia, the Americas, and Oceania). Kidd et al. [19] utilized 55 AIMs to analyze 73 populations from around the world. Phillips et al. [41] selected 128 AIM-SNPs to differentiate Africans, Europeans, East Asians, Native Americans, and Oceanians. Gettings et al. [42] used a 50-SNP assay for biogeographic ancestry and phenotype prediction of the major US populations in which 19 of the SNPs were ancestry informative markers. Three recently developed AIMs panels from Jia et al. [43], Rogalla et al. [44], and Wei et al. [21] contain 35, 14, and 27 SNPs to characterize three populations: African, European, and East Asian. Although there are several AIMs sets available, there is no universal core set of SNPs for ancestry inference. Therefore, we developed a SNP AIMs panel with the intent to use a minimum number of markers to characterize four major American populations: African American, East Asian, European American, and Hispanic American. These 23 markers could contribute to the candidate pool of AIMs for potential forensic identification purposes. Only two of our markers, rs11725412 and rs1834640, are in common with another panel (i.e., Nievergelt's panel). While MPS allows much larger panels to be evaluated, reducing the number of markers for both ease of panel development and increased throughput is desirable on both MPS and CE platforms. More samples could be multiplexed in an assay on the former platform, and marker multiplexing would be a better fit on the latter platform. Therefore, identifying a minimum number of AIMs to distinguish four US populations was sought. In our panel, there are four SNPs from chromosome 15, and they are located within 3–8 Mb of each other. Although they are not in LD within the four US populations, it is possible that they may affect admixture membership estimation in other populations.

Conclusion

In this study, three marker informativeness measures (δ, F_{ST}, and In) were compared for the AIMs selection among four American populations, i.e., African American, Caucasian, East Asian, and Hispanic American. The total number of markers in the AIMs panels selected by δ, F_{ST}, and In were 24, 23, and 23, respectively, and many of the markers were common within the three measures. Although not substantially different in performance, the F_{ST} panel performed slightly better for population resolution based on PCA clustering than did the δ panel and both performed better than the In panel. The 23 AIMs selected by the F_{ST} measure were used to characterize the four major American populations based on PCA clustering. Genotype data of the nine populations from HapMap and 1000 Genomes were used to evaluate the efficiency of 23-SNP panel. The results indicated that the individuals from these populations were assigned to the expected groups. However, the public databases did not provide the genotype data of the full AIMs panel. In a future study, a multiplex panel of the 23 AIMs will be developed and samples will be typed from four major US populations to further test the efficiency of the full AIMs panel. Our AIMs panel can contribute to the candidate AIMs for population stratification and potential forensic identification purposes.

References

1. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW (2002) Genetic structure of human populations. Science 298:2381–2385
2. Hoggart CJ, Parra EJ, Shriver MD, Bonilla C, Kittles RA, Clayton DG, McKeigue PM (2003) Control of confounding of genetic associations in stratified populations. Am J Hum Genet 72:1492–1504
3. Shriver MD, Parra EJ, Dios S, Bonilla C, Norton H, Jovel C, Pfaff C, Jones C, Massac A, Cameron N, Baron A, Jackson T, Argyropoulos G, Jin L, Hoggart CJ, McKeigue PM, Kittles RA (2003) Skin pigmentation, biogeographical ancestry and admixture mapping. Hum Genet 112:387–399
4. Marchini J, Cardon LR, Phillips MS, Donnelly P (2004) The effects of human population structure on large genetic association studies. Nat Genet 36:512–517
5. Jobling MA, Gill P (2004) Encoded evidence: DNA in forensic analysis. Nat Rev Genet 5:739–751

6. Yang N, Li H, Criswell LA, Gregersen PK, Alarcon-Riquelme ME, Kittles R, Shigeta R, Silva G, Patel PI, Belmont JW, Seldin MF (2005) Examination of ancestry and ethnic affiliation using highly informative diallelic DNA markers: application to diverse and admixed populations and implications for clinical epidemiology and forensic medicine. Hum Genet 118: 382–392
7. Shriver MD, Kittles RA (2004) Genetic ancestry and the search for personalized genetic histories. Nat Rev Genet 5:611–618
8. King JL, LaRue BL, Novroski NM, Stoljarova M, Seo SB, Zeng X, Warshauer DH, Davis CP, Parson W, Sajantila A, Budowle B (2014) High-quality and high throughput massively parallel sequencing of the human mitochondrial genome using the Illumina MiSeq. Forensic Sci Int Genet 12:128–135
9. Jobling MA, Tyler-Smith C (2003) The human Y chromosome: an evolutionary marker comes of age. Nat Rev Genet 4:598–612
10. Vigilant L, Stoneking M, Harpending H, Hawkes K, Wilson AC (1991) African populations and the evolution of human mitochondrial DNA. Science 253:1503–1507
11. Hammond HA, Jin L, Zhong Y, Caskey CT, Chakraborty R (1994) Evaluation of 13 short tandem repeat loci for use in personal identification applications. Am J Hum Genet 55:175–189
12. Jin L, Chakraborty R (1995) Population structure, stepwise mutations, heterozygote deficiency and their implications in DNA forensics. Heredity 74:274–285
13. Smith MW, Lautenberger JA, Shin HD, Chretien JP, Shrestha S, Gilbert DA, O'Brien SJ (2001) Markers for mapping by admixture linkage disequilibrium in African American and Hispanic populations. Am J Hum Genet 69:1080–1094
14. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29:308–311
15. International HapMap Consortium (2003) The International HapMap Project. Nature 426:789–796
16. 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491:56–65
17. Phillips C, Salas A, Sánchez JJ, Fondevila M, Gómez-Tato A, Alvarez-Dios J, Calaza M, de Cal MC, Ballard D, Lareu MV, Carracedo A, SNPforID Consortium (2007) Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. Forensic Sci Int Genet 1:273–280
18. Kosoy R, Nassir R, Tian C, White PA, Butler LM, Silva G, Kittles R, Alarcon-Riquelme ME, Gregersen PK, Belmont JW, De La Vega FM, Seldin MF (2009) Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. Hum Mutat 30:69–78
19. Kidd KK, Speed WC, Pakstis AJ, Furtado MR, Fang R, Madbouly A, Maiers M, Middha M, Friedlaender FR, Kidd JR (2014) Progress toward an efficient panel of SNPs for ancestry inference. Forensic Sci Int Genet 10:23–32
20. Nievergelt CM, Maihofer AX, Shekhtman T, Libiger O, Wang X, Kidd KK, Kidd JR (2013) Inference of human continental origin and admixture proportions using a highly discriminative ancestry informative 41-SNP panel. Investig Genet 4:13
21. Wei YL, Wei L, Zhao L, Sun QF, Jiang L, Zhang T, Liu HB, Chen JG, Ye J, Hu L, Li CX (2015) A single-tube 27-plex SNP assay for estimating individual ancestry and admixture from three continents. Int J Legal Med
22. Rosenberg NA, Li LM, Ward R, Pritchard JK (2003) Informativeness of genetic markers for inference of ancestry. Am J Hum Genet 73:1402–1422
23. Wright S (1950) Genetical structure of populations. Nature 166: 247–249
24. Ding L, Wiener H, Abebe T, Altaye M, Go RC, Kercsmar C, Grabowski G, Martin LJ, Khurana Hershey GK, Chakorborty R, Baye TM (2011) Comparison of measures of marker informativeness for ancestry and admixture mapping. BMC Genomics 12:622
25. Amirisetty S, Hershey GK, Baye TM (2012) AncestrySNPminer: a bioinformatics tool to retrieve and develop ancestry informative SNP panels. Genomics 100:57–63
26. Lewis PO, Zaykin D (2001) Genetic Data Analysis: computer program for the analysis of allelic data. Version 1.0 (d16c). http://hydrodictyon.eeb.uconn.edu/people/plewis/software.php. Accessed 25 April 2007.
27. Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PLoS Genet 2:e190
28. Zweig MH, Campbell G (1993) Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem 39:561–577
29. Qin P, Li Z, Jin W, Lu D, Lou H, Shen J, Jin L, Shi Y, Xu S (2014) A panel of ancestry informative markers to estimate and correct potential effects of population stratification in Han Chinese. Eur J Hum Genet 22:248–253
30. Adinsoft SARL (2010) XLSTAT-software. Version 10. Addinsoft, Paris
31. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959
32. SPSS Inc (2007) SPSS for Windows. Version 16.0. Chicago
33. Green SB, Salkind NJ, Akey TM (2008) Using SPSS for Windows and Macintosh: analyzing and understanding data. Prentice Hall, New Jersey
34. Kidd JM, Gravel S, Byrnes J, Moreno-Estrada A, Musharoff S, Bryc K, Degenhardt JD, Brisbin A, Sheth V, Chen R, McLaughlin SF, Peckham HE, Omberg L, Bormann-Chung CA, Stanley S, Pearlstein K, Levandowsky E, Gravel S, Acevedo-Acevedo S, Auton A, Keinan A, Acuna-Alonzo V, Canizales-Quinteros S, Eng C, Burchard EG, Russell A, Reynolds A, Clark AG, Reese M, Lincoln SE, Butte AJ, De La Vega FM, Bustamante CD (2012) Population genetic inference from personal genome data: impact of ancestry and admixture on human genomic variation. Am J Hum Genet 91:660–671
35. Wall JD, Jiang R, Gignoux C, Chen GK, Eng C, Huntsman S, Marjoram P (2011) Genetic variation in Native Americans, inferred from Latino SNP and resequencing data. Mol Biol Evol 28:2231–2237
36. Salazar-Flores J, Zuñiga-Chiquette F, Rubi-Castellanos R, Álvarez-Miranda JL, Zetina-Hérnandez A, Martínez-Sevilla VM, González-Andrade F, Corach D, Vullo C, Álvarez JC, Lorente JA, Sánchez-Diz P, Herrera RJ, Cerda-Flores RM, Muñoz-Valle JF, Rangel-Villalobos H (2015) Admixture and genetic relationships of Mexican Mestizos regarding Latin American and Caribbean populations based on 13 CODIS-STRs. Homo 66:44–59
37. Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics 23:1801–1806
38. Rosenberg N (2004) Distruct: a program for the graphical display of population structure. Mol Ecol Notes 4:137–138
39. Bushnell D, Hudson RA (2010) Colombia: a country study. Federal Research Division, Library of Congress, Washington D.C
40. Halder I, Shriver M, Thomas M, Fernandez JR, Frudakis T (2008) A panel of ancestry informative markers for estimating individual biogeographical ancestry and admixture from four continents: utility and applications. Hum Mutat 29:648–658
41. Phillips C, Parson W, Lundsberg B, Santos C, Freire-Aradas A, Torres M, Eduardoff M, Børsting C, Johansen P, Fondevila M, Morling N, Schneider P, EUROFORGEN-NoE Consortium, Carracedo A, Lareu MV (2014) Building a forensic ancestry panel

from the ground up: the EUROFORGEN Global AIM-SNP set. Forensic Sci Int Genet 11:13–25

42. Gettings KB, Lai R, Johnson JL, Peck MA, Hart JA, Gordish-Dressman H, Schanfield MS, Podini DS (2014) A 50-SNP assay for biogeographic ancestry and phenotype prediction in the US population. Forensic Sci Int Genet 8: 101–108

43. Jia J, Wei YL, Qin CJ, Hu L, Wan LH, Li CX (2014) Developing a novel panel of genome-wide ancestry informative markers for biogeographical ancestry estimates. Forensic Sci Int Genet 8:187–194

44. Rogalla U, Rychlicka E, Derenko MV, Malyarchuk BA, Grzybowski T (2015) Simple and cost-effective 14-loci SNP assay designed for differentiation of European, East Asian and African samples. Forensic Sci Int Genet 14:42–49