ORIGINAL ARTICLE

Empirical testing of a 23-AIMs panel of SNPs for ancestry evaluations in four major US populations

Xiangpei Zeng[1] · **David H. Warshauer**[1] · **Jonathan L. King**[1] · **Jennifer D. Churchill**[1] · **Ranajit Chakraborty**[1] · **Bruce Budowle**[1,2]

Received: 15 December 2015 / Accepted: 5 February 2016 / Published online: 25 February 2016

Abstract Ancestry informative markers (AIMs) can be used to determine population affiliation of the donors of forensic samples. In order to examine ancestry evaluations of the four major populations in the USA, 23 highly informative AIMs were identified from the International HapMap project. However, the efficacy of these 23 AIMs could not be fully evaluated in silico. In this study, these 23 SNPs were multiplexed to test their actual performance in ancestry evaluations. Genotype data were obtained from 189 individuals collected from four American populations. One SNP (rs12149261) on chromosome 16 was removed from this panel because it was duplicated on chromosome 1. The resultant 22-AIMs panel was able to empirically resolve the four major populations as in the in silico study. Eight individuals were assigned to a different group than indicated on their samples. The assignments of the 22 AIMs for these samples were consistent with AIMs results from the ForenSeq^{TM} panel. No departures from Hardy-Weinberg equilibrium (HWE) and linkage disequilibrium (LD) were detected for all 22 SNPs in four US populations (after removing the eight problematic samples). The principal component analysis (PCA) results indicated that 181 individuals from these populations were assigned to the expected groups. These 22 SNPs can contribute to the candidate AIMs pool for potential forensic identification purposes in major US populations.

Keywords Ancestry informative markers (AIMs) · Single nucleotide polymorphisms (SNPs) · Population differentiation · Custom oligonucleotide probe · Principal component analysis (PCA)

Introduction

Ancestry informative markers (AIMs), based on single nucleotide polymorphisms (SNPs), are useful for determination of population affiliation and apportionment of individual ancestry [1–4]. Determination of population affinity of the donor of an evidence sample or the ancestry of unidentified human remains can assist in forensic investigations, especially for indirect phenotype information, confirming or refuting eye witness accounts, assisting anthropology, or when STRs fail to provide hits or associations through DNA database searches [5–7].

Recently, Zeng et al. [8] described 23 highly informative SNP AIMs that were identified from sequence data from the International HapMap project using F_{ST} as the measure of selecting AIMs. F_{ST} is the measure of genetic distance between two populations based on genetic data, and a high F_{ST} value indicates substantial degree of differentiation between populations [9]. An in silico study using this panel demonstrated that, it is possible to conduct ancestry evaluations in four major US populations. All but two of the AIMs were novel and had not been described previously for such purposes. However, the actual performance of these 23 SNPs could not be fully evaluated, because: (1) the public databases (i.e., HapMap and 1000 Genomes) did not provide complete

Electronic supplementary material The online version of this article (doi:10.1007/s00414-016-1333-4) 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, Texas 76107, USA

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

genotype data for all 23 AIMs for each population tested in silico [10, 11]; and (2) there can be unpredicted effects (e.g., sequence surrounding the SNP that may affect the ability to type the marker) that may be determined only with empirical testing. Therefore, the objective of this study was to develop a multiplex panel for genotyping the selected 23 AIMs and generate SNP profiles on samples collected from four major US populations to further test the efficacy of this full AIMs panel.

Methods and materials

Population samples

DNA from either blood or buccal samples was obtained from 189 unrelated individuals (81 males and 108 females) with informed consent. These samples included 49 African Americans, 43 Asians, 49 Caucasians, and 48 Hispanics. African Americans, Caucasians, and Hispanics were collected from a blood bank in Fort Worth, Texas. Population affinity was based on self-declaration. Of the 43 Asian samples, 13 samples were collected from the same blood bank and reported as Asians, and the rest of the samples were collected from the Dallas-Fort Worth area. Population affinity for these samples also was based on self-declaration as Asians (Chinese or Japanese). All samples were collected anonymously according to University of North Texas Health Science Center's Institutional Review Board. All samples were extracted using the QIAamp™ DNA Mini Kit (Qiagen, Hilden, Germany) according to the manufacturer's recommended protocol [12].

Panel design

The Nextera™ Rapid Capture Custom Enrichment kit (Illumina, Inc., San Diego, CA) was used to enrich the target SNPs according to the manufacturer's recommended protocol [13]. Custom oligonucleotide probes (80 bases in length) of the 23 ancestry informative SNPs were designed using Design Studio v1.5 [14] under the default conditions, and hg19 was used for probe reference. The details of the SNPs, such as chromosomal position, target selection (Full Region), probe density requirements, and marker information were uploaded to Design Studio for probe design. The information on the probes for the 23 AIMs are provided in Supplemental Table 1.

Quantification and normalization

After extraction, the Qubit dsDNA BR kit (Life Technologies, Carlsbad, CA) was used to determine the quantity of DNA for each sample following the manufacturer's protocol [15]. DNA samples were normalized to 10 ng/μL with the quantity determined again, and diluted to 5 ng/μL in order to ensure sufficient DNA for library preparation.

Library preparation

Library preparation was performed using the Nextera™ Rapid Capture Custom Enrichment protocol according to the manufacturer's protocol [13]. A total of 50 ng of DNA was used for library preparation for each sample. The samples were enzymatically cleaved and ligated to sequencing adapters, and then tagmented samples were purified with two 80 % ethanol washes. The Agilent® 2200 TapeStation™ (Agilent Technologies, Inc., Santa Clara, CA) was used to analyze the fragment sizes of samples to check whether tagmentation was successful. Dual sequencing indices were ligated to each of the fragments in the first PCR amplification. After amplification cleanup, the quantity of each indexed samples was quantified using Qubit dsDNA BR kit. Twelve libraries at each time were normalized and pooled for sequencing. Each library contained 500 ng of DNA sample. The custom oligonucleotide probes were hybridized to the pooled libraries, followed by two streptavidin bead-based cleanup steps. The second hybridization was performed with the same thermal-cycling parameters (except that the final hold time was extended to 20 h). Subsequently, two additional bead-based washes were conducted. Library enrichment was performed on an Eppendorf® Mastercycler® Pro S thermal cycler using the following thermal-cycling parameters (second PCR amplification): 30 s at 98 °C; 12 cycles of 10 s at 98 °C, 30 s at 60 °C, 30 s at 72 °C; and a final extension of 5 min at 72 °C then maintained on hold at 10 °C. The quantity of libraries was determined using a Qubit dsDNA BR kit after a final bead-based cleanup procedure. The Agilent® 2200 TapeStation™ was used to determine the average size of the enriched fragments for each pooled library.

MPS sequencing and data analysis

Each library was normalized to 2 nM and the DNA was denatured. The denatured library was diluted to 12 pM and sequenced with the MiSeq v2 (2 × 250 bp) chemistry (Illumina). The raw FASTQ files were aligned by the onboard software MiSeq Reporter, and resulting BAM files were analyzed by the Genome Analysis Toolkit (GATK) [16] to display SNP genotypes and their coverage values.

Concordance data

SNP typing of eight questionable samples (by ancestry assignment) were analyzed using the Illumina ForenSeq™ DNA Signature Prep Kit as described by Churchill et al. [17]. The ancestry assignments between our AIMs panel and that by the AIMs contained within the ForenSeq™ kit were compared for resolving non-concordant population affinity.

Results and discussion

In the previous in silico study [8], 23 SNPs were selected that could resolve ancestries of four US populations (Table 1). This panel was assessed empirically for resolving ancestries of 189 locally collected individuals from four US populations. In the present multiplex assay of these 23 SNPs, one SNP (rs11845995) displayed three alleles (G/A/C) in all populations. The average coverage of 22 of 23 AIMs in the 189 individuals was shown in Supplemental Figure 1 (one SNP was removed, see next paragraph). The interlocus balance (the lowest mean coverage/the highest mean coverage) was 0.29. The lowest coverage observed was 22× (20× was set as the detection threshold), and the highest coverage was 2216×. There were only four examples of locus drop out: three were detected at SNP rs1761031 and one at SNP rs974627. These results indicated that the 22 AIMs panel had sufficiently high coverage and good interlocus balance using Nextera™ Rapid Capture Custom Enrichment method.

Tests for departures from Hardy-Weinberg equilibrium (HWE) and detectable linkage disequilibrium (LD) with the 23 SNPs in each of the four populations were performed using GDA [18]. Only one SNP (rs12149261) deviated from HWE expectations, and the departure was observed in three populations (Asian, Caucasian and Hispanic American) even after applying Bonferroni's correction for multiple testing ($p = 0.05/23$). This SNP also was involved in 22 out of the 27 pairs of loci that exhibited significant LD. In addition, the genotype data of SNP rs12149261 showed that 136 of 140 individuals (Asian, Caucasian, and Hispanic groups) were heterozygote CA, and 4 individuals had the homozygote CC genotype. Such a high number of heterozygotes was a strong indication of typing error. There was no evidence that quality of the sequence data contributed to the mistyping (Supplemental Figure 2). The sequence surrounding SNP rs12149261 was searched using BLAST, which indicated that the SNP site is duplicated (Supplemental Figure 3). The SNP rs12149261 is located in the HYDIN gene on chromosome 16, and a duplicate region is located in the HYDIN2 gene on chromosome 1. There is complete homology between the two sites except at the SNP location. Therefore, the genotype data of rs12149261 were actually a combination of sequence reads from two SNP sites, resulting in the majority of individuals in three populations being an apparent CA heterozygote. Since the SNPs were detected in silico originally, there was no need to BLAST the sequence harboring the SNP. However, empirically such testing should be pursued to avoid the phenomenon observed in this study. The sequence of the rest of the 22 SNPs was blasted, and no duplications were detected.

Table 1 The 23 AIMs selected to distinguish the four major US populations

SNPs	Chromosome	Genomic position	Alleles
rs12087334	1	116887455	C/A
rs11126303	2	26173503	A/G
rs13021399	2	109006665	T/A
rs745767	2	177825415	G/A
rs10510511	3	21260370	G/T
rs2700372	3	123633220	T/G
rs11725412	4	38277754	A/G
rs7689609	4	72083374	C/T
rs1827950	4	117098482	G/T
rs4729945	7	103677151	T/C
rs10962599	9	16795286	C/T
rs11139346	9	84241442	T/C
rs974627	12	38919524	T/C
rs7134749	12	50237637	T/C
rs1761031	14	46926398	G/T
rs11845995	14	105930923	G/A/C
rs1288097	15	45141373	G/A
rs1834640	15	48392165	A/G
rs7165971	15	55921013	T/C
rs8032157	15	64480888	A/G
rs6500380	16	48375777	A/G
rs12149261*	16	70998145	C/A
rs4429562	22	42892596	T/C

*This AIM on chromosome 16 was removed due to a duplication on chromosome 1

The sequence flanking of the SNP rs12149261 and its duplicate are identical, which means any probe or short amplicon PCR method would not be able to isolate the SNP from its duplicate. Therefore, this SNP was removed from the panel, and only 22 AIMs were subsequently assessed. After removing the problematic SNP (rs12149261), tests for HWE and LD of 22 SNPs in the four populations were performed again. No SNPs deviated from HWE. Five SNP pairs showed detectable LD (Table 2), which were rs745767/rs4429562 and rs7165971/rs4429562 in African Americans; rs745767/rs7134749, rs2700372/rs7165971, and rs1834640/rs7165971

Table 2 Significant linkage disequilibrium (LD) results of 22 SNPs in four populations

SNP pair	LD *p* values in Population	
	African American	Asian
rs745767/rs4429562	*<0.000001*	0.0143
rs7165971/rs4429562	*<0.000001*	0.0030
rs745767/rs7134749	0.0318	*<0.000001*
rs2700372/rs7165971	0.1025	*0.0002*
rs1834640/rs7165971	0.0288	*0.0002*

LD *p* values shown for the specified loci pair in which a significant value was observed in at least one population group. Values in italics were significant after Bonferroni correction ($p < 0.000216$)

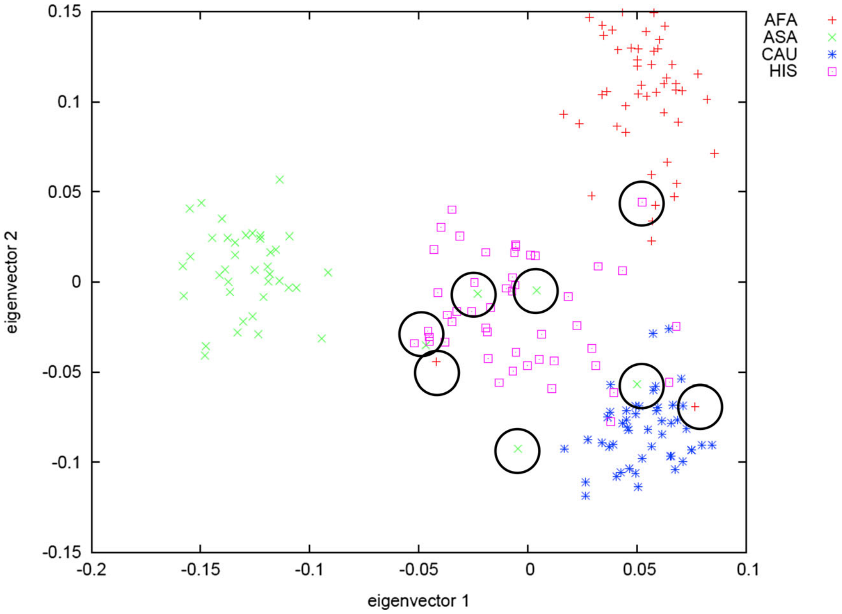

Fig. 1 The PCA plot of 189 individuals using 22-SNP AIMs panel. Eight individuals (*encircled by black circles*) were assigned to different groups than what were labeled on their sample submissions

in Asians. This number of deviating observations (5 out of 231 pair of loci) is within expectations of chance occurrences but also could be attributed to population substructure (see below).

The principal component analysis (PCA) plot showed that eight individuals were assigned to a different group than indicated on their samples (Fig. 1): two African Americans (20882 and 23169), five Asians (76194, 06498, 12574, 38859, and 10916), and one Hispanic American (61115). Population affinity was determined by self-declaration, and the samples were anonymous. Thus, true population affiliation could not be confirmed or refuted directly. The category Asians is quite broad, and some of these individuals may not fit well with East Asians (CHD population was used originally to select the AIMs). Thus, Asians other than East Asians likely would reside with admixed individuals in the PCA plot. Other explanations for assignment in a conflicting population category are that these individuals wrongly reported their population ancestry or samples were mislabeled during collection. Lastly, it is possible that our AIMs panel failed to properly cluster these eight individuals. To ascertain which of the explanations have more support, i.e., wrong categorization of the samples before entering the laboratory or a failure of the panel to resolve, these eight samples were analyzed using the MiSeq FGx Forensic Genomics System. The panel of primers included in the ForenSeq^{TM} DNA Signature Prep Kit (used for library generation) contains 56 AIMs [17] (Supplemental Figure 4, Table 3). African American sample 20882 was classified as Hispanic American by our AIMs panel and the ForenSeq^{TM} panel. African American sample 23169 was identified as Caucasian by our panel, but it was classified as Hispanic American, close to the Caucasian group, with the ForenSeq^{TM} panel. Hispanic individual no. 61115 was classified as African American by both panels. Of the five Asian samples, the 22 AIMs panel assigned samples 76194, 06498, and 12574 to the Hispanic American group, sample 38859 to the Hispanic American or Caucasian group, and sample 10916 to the Caucasian group. However, all five individuals were identified as Hispanic Americans by the ForenSeq^{TM}

Table 3 The predicted ancestries of the eight individuals by the 22-SNP AIMs panel and the ForenSeq^{TM} panel

Individual	Self-reported or labeled ancestry	22 AIMs panel result	ForenSeq^{TM} panel result
20882	African American	Hispanic American	Hispanic American
23169	African American	Caucasian	Hispanic American, close to Caucasian group
76194	Asian	Hispanic American	Hispanic American
06498	Asian	Hispanic American	Hispanic American
12574	Asian	Hispanic American	Hispanic American
38859	Asian	Hispanic American or Caucasian	Hispanic American
10916	Asian	Caucasian	Hispanic American
61115	Hispanic American	African American	African American

panel. To clarify, US populations are expected to be admixed to some degree, and the 22 AIMs were selected based on US populations to maximize US-population resolution. Hispanic samples with the 22 AIMs panel will be clustered as admixed populations (depending on their degree of admixture). Asians (originating west of East Asian groups) also will fall within the Hispanic cluster. Thus, all falling within the Hispanic cluster based on our panel should be classified only as admixed individuals and can be any notable combination of the three primary populations used to develop the original SNP panel. Based on the comparable results between the two AIMs panels for the eight samples, the findings of the 22 AIMs panel are supported as being correct. Therefore, the ancestries of eight samples were either wrongly reported or a result of classification of a Hispanic ancestry which in itself is a geopolitical construct as opposed to being a defined population. Fifty-six AIMs in ForenSeq™ panel were used to confirm the ancestry results. It should be noted that Y-SNPs and mitochondrial DNA could be used as well. However, it was deemed that lineage markers would not provide a better overall assessment than autosomal ancestry SNPs. These eight samples were removed and the 22 AIMs were tested for departures from HWE and LD. There were no detectable departures observed in the four populations. The PCA results, after removing the eight individuals, are shown in Supplemental Figure 5. Four populations were distinguished in the PCA plot except a few Hispanic Americans were assigned to the Caucasian group as would be expected.

Overall, the results indicated that these 22 AIMs can correctly assign individuals to the four major US population categories. However, this panel may not predict as well the ancestry of the individuals from other US populations, e.g., Native Americans. Potentially more AIMs may be needed for these groups.

Conclusions

The initial 23-AIMs panel was evaluated empirically by typing 189 individuals collected from four US populations, i.e., African American, Asian, Caucasian, and Hispanic American. One SNP (rs12149261) deviated from HWE expectations and was associated with most of the detectable LD in three of the populations. Most of the genotypes were heterozygotes which is inconsistent with an AIM and population genetic expectations for a bi-allelic SNP. The BLAST results indicated that SNP rs12149261 residing on chromosome 16 and its surrounding region were duplicated on chromosome 1. The rest of the 22 AIMs enabled population assignment. The population affiliations of eight individuals were inconsistent with their self-declared population. The assignment by the 22 AIMs was consistent with AIMs from the ForenSeq™ DNA Signature Prep Kit. After removing the wrongly assigned eight samples, there were no detectable departures from HWE and detectable LD in four US populations for all 22 SNPs. The PCA results indicated that the 22 AIMs can resolve individuals into the four major US populations. These 22 SNPs are additional AIMs to consider for a panel(s) 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. Zeng X, Chakraborty R, King JL, LaRue B, Moura-Neto RS, Budowle B (2015) Selection of highly informative SNP markers for population affiliation of major US populations. Int J Legal Med. doi:10.1007/s00414-015-1297-9
9. 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
10. International HapMap Consortium (2003) The International HapMap Project. Nature 426:789–796
11. 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
12. QIAamp® DNA mini and blood mini handbook (2012) https://www.qiagen.com/us/resources/resourcedetail?id=67893a91-946f-49b5-8033-394fa5d752ea&lang=en
13. Nextera rapid capture enrichment reference guide (2015) https://support.illumina.com/downloads/nextera-rapid-capture-guide-15037436.html
14. DesignStudio (2015) https://accounts.illumina.com/?ReturnUrl=http://designstudio.illumina.com/

15. Qubit® dsDNA BR Assay Kit (2015) https://www.thermofisher.com/order/catalog/product/Q32850
16. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297–1303
17. Churchill JD, Schmedes SE, King JL, Budowle B (2015) Evaluation of the Illumina® Beta Version ForenSeq™ DNA Signature Prep Kit for use in genetic profiling. Forensic Sci Int Genet 20:20–29
18. 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