Zeng, Xiangpei, Selection of Highly Informative Markers for Apportionment of Ancestry and Population Affiliation. Doctor of Philosophy (Biomedical Sciences), May 2016, 118 pp., 18 tables, 20 figures, 69 references.

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. This doctoral dissertation research was conducted under the hypotheses that δ and F_{ST} perform better than In, and highly informative AIMs can be selected among human populations by using these three marker informativeness measures.

The primary goal of this project was to develop a robust AIMs panel with a minimum number of markers that can be used for apportionment of ancestry and population affiliation of four major US populations, that is African American, US Caucasian, East Asian and Hispanic American. First, candidate SNPs were searched and downloaded from the HapMap Project. Then these SNPs were ranked for their informativeness based on the three measures (δ, F_{ST}, and In) in a population pairwise manner. The F_{ST} measure appeared to be the most informative measure, performing slightly better than δ. With this approach and population statistics assessment, a minimum number of AIMs, i.e., 23, was selected to characterize the four major American populations. The efficacy of these 23 SNPs was tested *in silico* using nine populations from the HapMap project and 1000 Genomes. Finally, empirical testing was performed using 189 individuals collected from four US populations to evaluate further the performance of the 23-AIMs panel.

The results of this dissertation research indicated that these 23 AIMs can correctly assign individuals to the major population categories *in silico*. Empirical testing results showed that one SNP (rs12149261) on chromosome 16 had a duplicated region on chromosome 1. This SNP was removed from my list, in order to avoid erroneous results. The resultant 22-AIMs panel was able to resolve the four major populations as in the *in silico* study. PCA results showed that eight individuals were not assigned to the expected major population categories. 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 results indicated that the 22 AIMs can correctly assign individuals to the four major US population categories. These 22 SNPs could contribute to the candidate pool of AIMs for potential forensic identification purposes and population stratification studies for biomedical research in the major US populations.

KEYWORDS: Ancestry informative markers (AIMs) · Single nucleotide polymorphisms (SNPs) · Population differentiation · HapMap · 1000 Genomes · F_{ST} · Custom oligonucleotide probe · Massively parallel sequencing · Principal component analysis (PCA) · STRUCTURE

SELECTION OF HIGHLY INFORMATIVE MARKERS FOR APPORTIONMENT OF ANCESTRY AND POPULATION AFFILIATION

DISSERTATION

Presented to the Graduate Council of the
Graduate School of Biomedical Sciences
University of North Texas
Health Science Center at Fort Worth
In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

By

Xiangpei Zeng, M.S.
Fort Worth, Texas
May, 2016

ACKNOWLEDGEMENTS

I would like to express the deepest appreciation to Dr. Bruce Budowle for supporting me during the past four years. He is not only a tremendous advisor to me, but also a good friend. His understanding and patience helped me get over my English barrier in my first year. His guidance was invaluable on both my research as well as on my career. He encouraged me to not only grow as a forensic geneticist but also as an independent thinker. I am not sure how many doctoral students are given this opportunity to work with such independence. However, I am lucky to have the chance to develop my own personality throughout my doctoral research. I thank you, Dr. Budowle, for everything you have done for me.

I would also like to thank the other four professors in my PhD committee: Dr. Ranajit Chakraborty, Dr. Harlan Jones, Dr. Bobby LaRue, and Dr. Rance Berg. Their assistance and guidance have been priceless to me during my graduate studies. In particular, Dr. Chakraborty provided substantial advice on population genetics and statistical analyses for my doctoral research. Without his help, I would not have completed my research in four years.

In addition, my colleagues and co-workers also offered tremendous help for my studies. I would like to sincerely thank Jonathan King, Jianye Ge, David Warshauer, Seung Bum Seo, and Jennifer Churchill for their generous help. I am also grateful to other lab members, Angie Ambers, Maiko Takahashi, Carey Davis, Pamela Marshall, Sarah Schmedes, Nicole Novroski, Frank Wendt, Rachel Wiley, and Bing Song, for providing a friendly laboratory environment for

international students. I also appreciate technical support received from Spencer Hermanson, Jaynish Patel, Douglas Storts of Promega Corporation.

Finally, and most importantly, I would like to thank my parents and younger sister, for their belief in me and supporting me to be as ambitious as I could be. My incredible wife, Lucy, also plays an important role in my growth. Without her encouragement, I could not have made decision to pursue my doctoral degree in the United States. She also contributed to my research, collecting some human samples for me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 2: ***Empirical Testing of A 23-AIMs Panel of SNPs for Ancestry Evaluations in Four Major US Populations***

INTRODUCTION

Selection of Highly Informative Markers for Apportionment of Ancestry and Population Affiliation

Single nucleotide polymorphisms (SNPs) are the most frequent types of genetic variants within the human genome and present as differences in the allelic state at specific nucleotide position(s) between and among individuals (1). For example, AAGCCTA and AAGCTTA are DNA fragments from two individuals, showing a difference at a single nucleotide location (C/T) (Figure 1). The majority of SNPs only have two alleles.

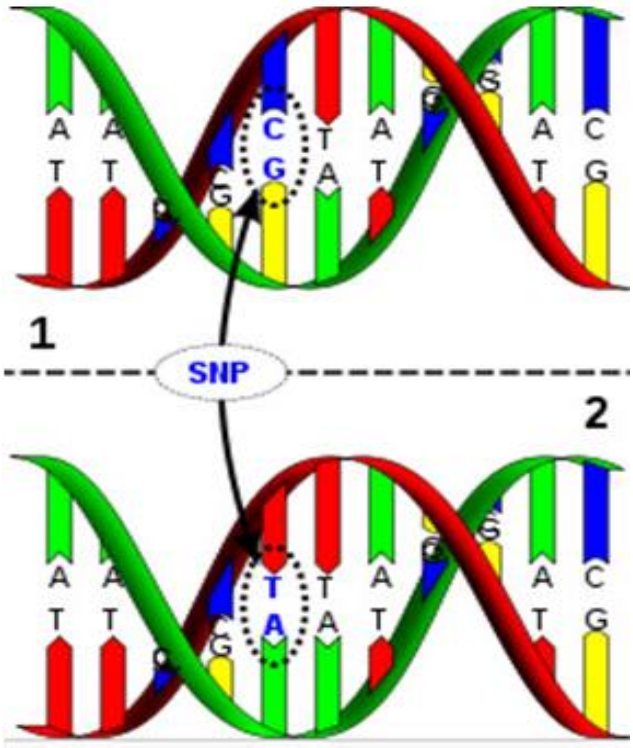


Figure 1. An example of the SNPs within a DNA sequence (the figure is from reference 1). DNA molecule 1 and molecule 2 are different at a single nucleotide position (C/T polymorphism).

Genome-Wide Association Studies (GWAS) (Figure 2) examine the whole genome of different individuals to determine whether any genetic variants or polymorphisms are associated with complex diseases, such as diabetes, cancer and cardiovascular disorders (2). This approach compares the DNA sequence of two groups of individuals: people with the disease (case group) and people without the disease (control group). If one type of SNP variant is more common

among individuals in the case group, this SNP can be considered to be associated with the disease of interest (3). The assumption of GWAS is that individuals in the case and control groups come from the same population and thus have similar genome composition.

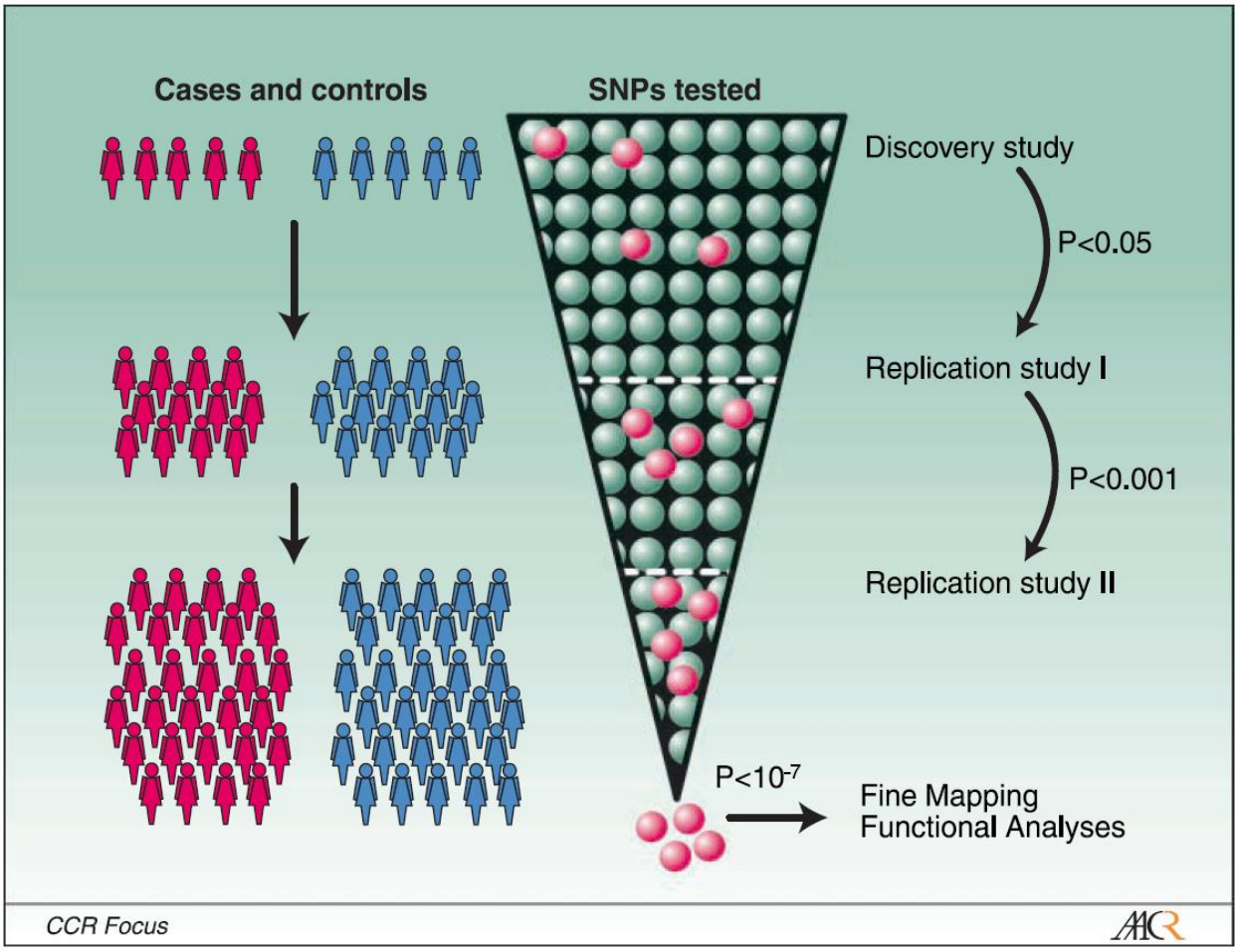


Figure 2. Multistage design for genome-wide association studies. In the first stage, a large number of SNPs are tested using genome-wide scan chip to capture the most common genetic variations in a small group of cases and controls. The SNPs showing the most significant associations with disease of interest in the first stage are retested in the second stage replication studies including a large number of cases and controls (the figure is from reference 4).

However, the presence of undetected population stratification or population substructure may complicate genetic association studies and lead to more false causative SNPs (Figure 3) (5).

Population stratification impacts differences in allele frequencies of SNPs between case and control samples as they may due to different human background population genetics, not result from genetic variations of traits among individuals (such as disease, or drug response). Thus, correction for population stratification is necessary for association studies.

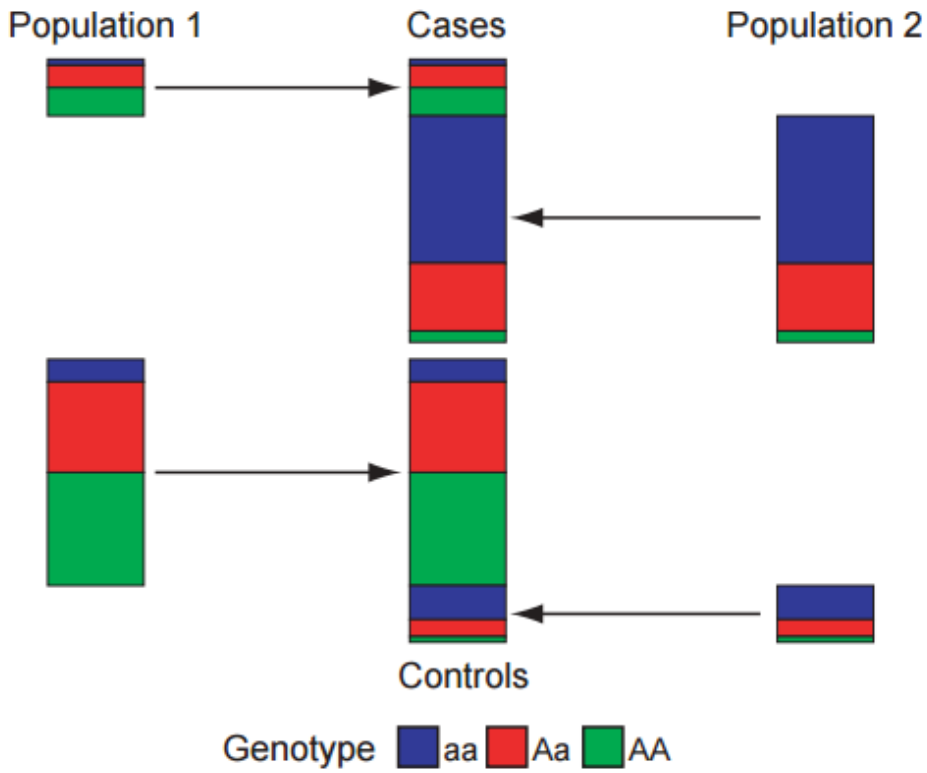


Figure 3. An example of the influences of population structure from data from GWAS (the figure is from reference 5). This figure simulates the differences in allele frequency between case and control samples caused by population stratification. The majority of individuals of the case group from population 2 carry allele "a", while most of the controls from population 1 have an extremely low frequency of allele "a".

One subclass of SNPs is ancestry informative markers (AIMs) which are genetic makers that show large differences in allele frequencies among human populations. These differences allow determination of population affiliation and apportionment of ancestry (6). Estimated population affiliation of individuals can be used to detect and adjust for population stratification in GWAS

(5,7-8). Allele frequencies among human populations from different continents could be substantially different, because groups of individuals have diverse social and genetic histories, including mutation, geographical migration, mating choices, natural selection and random genetic drift (9). Population-induced allele frequency differences are prevalent throughout the entire human genome, and these differences may result in false positive associations in genetic studies (10-11). Two conditions must be addressed simultaneously in association studies: disease occurrence differences between case and control samples, and disparity in individual ancestral composition of the respective sample groups (12). An example of population stratification induced false positive associations was a negative correlation reported between HLA haplotype and Type 2 diabetes in Native Americans. However, further studies showed that this negative association was confounded by genetic admixture of European ancestry. In fact, the specific HLA haplotype was more frequent and the prevalence of type 2 diabetes was lower in individuals with European ancestry compared with individuals of Native American ancestry (13-14).

AIMs also can play a role in ancestry inference in the field of forensic genetics (15-16). The advantage of AIMs in a forensic analysis is that these markers may provide investigators with critical evidence about an unknown suspect or about the ancestry of unidentified human remains. A well-known case in Louisiana emphasizes the usefulness of testing of AIMs in supporting an investigation (17). Five women were determined to be murdered by the same man through the analysis of a panel of short tandem repeat (STR) markers (i.e., those markers routinely used by forensic scientists worldwide for identity testing) on biological evidence found at the crime scenes. A search of the STR profile against profiles from convicted felons in the US national Combined DNA Index System (CODIS) database did not yield any matches. Thus, no further

information about the potential perpetrator was provided by traditional STR analysis. Based on eye-witness accounts, the police focused their investigation on Caucasians and checked the background of several hundreds of "white" people that lived in the area surrounding the crime scene. However, all people investigated were excluded because they had different STR profiles compared with that derived from the crime scene evidence. A few months later, ancestry information was developed using a panel of AIMs. The analysis showed that the source of the crime scene evidence had a bioancestry consistent with African rather than Caucasian. This new clue helped police to switch their investigations from Caucasians to African Americans, and they eventually identified the true source of the biological evidence.

There are four types of genetic markers that could provide ancestry information: mitochondrial DNA (mtDNA), Y chromosome markers, autosomal STRs and SNPs. Lineage markers (mtDNA and Y chromosome) have proven effective in studying human migration and evolutionary histories across the world (18-20) (Figure 4).

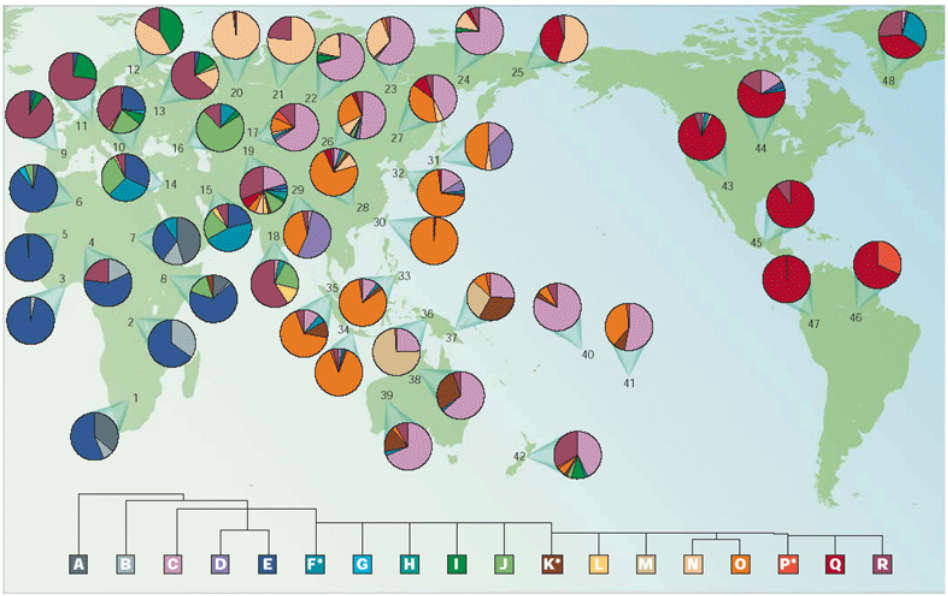


Figure 4. The Y chromosome haplogroups in different continents. The neighboring populations show high similarities, while large differences are observed between or among populations from different continents (the figure is from reference 19).

However, due to a lack of recombination and uniparental inheritance, 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 directly. STRs typically are highly polymorphic, and a relatively small panel of markers can successfully distinguish an individual from all 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 (21). Contraction-expansion patterns of mutations at STR loci also imply that STR alleles of the same repeat size may not all be identical by descent (22). In spite of these limitations, some panels of STR markers have been shown to resolve to some degree African American, Hispanic, European American and Asian. However, ancestry determination of STRs is not highly accurate (23). 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 (24-26). 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 SNPs panels have been developed for potential application of ancestral inference in forensic genetics (27-31).

An ideal ancestry informative SNP would have one allele fixed in one population, and 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 ancestry informative SNPs across populations. Several marker informativeness measures for ancestry estimation have been developed for selection of AIMs. These measures include: Absolute Allele

Frequency Differences (δ) (32), F statistics (F_{ST}) (33), and Informativeness for Assignment Measure (In) (32).

Absolute allele frequency difference (δ): δ is the most widely used informativeness measure for ancestry inference between populations. δ refers to the absolute frequency difference of an allele in two parental populations (32). The δ value of an ideal informative marker should be 1 while the worst marker (δ=0) cannot provide information about ancestry. The formula for δ is as follows (32), considering allele 1 is the particular allele in two populations:

$$\delta = |p_{11} - p_{21}|$$

F statistics (F_{ST}): F_{ST} is defined as the correlation of genetic variance within subpopulations relative to the genetic variance of the total population (33). Values of F_{ST} are in the range of 0 to 1. 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. The formula of F_{ST} is as follows (33), j is the specific allele in two populations:

$$F_{ST} = \frac{(p_{1j} - p_{2j})^2}{(p_{1j} + p_{2j})(2 - (p_{1j} + p_{2j}))}$$

Informativeness for assignment (In): In infers as the likelihood ratio for the assignment of an allele to one of the populations relative to the average population (31). The average population is hypothetical, and its allele frequencies are the mean allele frequencies of K populations. In values range from 0 to 1. The value of In is larger if the same alleles have significantly different frequencies in all subpopulations. The formula of In of a SNP is defined as (31):

$$I_n = \sum_{j=1}^{N}(-p_j \log_2 p_j + \sum_{i=1}^{K} \frac{p_{ij} \log_2 p_{ij}}{K})$$

The screening criteria of these three measures are highly arbitrary, e.g., cutoff values of 0.5, 0.4 and 0.3 for differences in allele frequencies between populations were used for δ by different researchers (34-37). Currently, these measures have not been compared and evaluated for their efficacy in estimating the ancestry of individuals in various populations (e.g., African American, Caucasian, East Asian, and Hispanic American). Therefore, researchers have used different marker measures to select AIMs based on personal preferences. For instance, three US research groups used δ, F_{ST} and In to study the same admixed population of Latinos (38-40). On the basis of social and historical issues, some populations are admixed from two or more ancestors, i.e., African American and Hispanic American, while other populations are more homogeneous, e.g., East Asian and Caucasian. It is not clear which measure-based derived AIMs panel is the most informative for such population differentiation. δ and F_{ST} have been used more so in AIMs selection, because these two measure can be used for homogeneous populations (e.g. East Asian and Caucasian) and admixed populations (e.g. African American and Hispanic American). While In is mainly used for admixed populations. With the availability of millions of SNPs in International HapMap project, I have the opportunity to assess and compare these measures directly.

The International HapMap Project is the product of global work of many scientists, research groups, and private institutions from several countries: Canada, China, Japan, Nigeria, the United Kingdom and the United States (24). The aim of the HapMap Project is to develop a haplotype map (HapMap) of the human genome from groups of individuals residing in different continents.

The HapMap information can be used for disease studies to identify genetic variants that contribute to complex disease occurrence, drug response differences, and used in forensic genetics to select human identification markers or AIMs. The HapMap project contains three phases, and the sequence data were released in 2005, 2007 and 2009. A total number of 1301 samples from 11 populations is included in HapMap Phase III (24) (Figure 5): African ancestry from Southwest USA (ASW), Utah residents with Northern and Western European ancestry from the CEPH collection (CEU), Han Chinese from Beijing, China (CHB), Chinese from Metropolitan Denver, Colorado (CHD), Gujarati Indians from Houston, Texas (GIH), Japanese from Tokyo, Japan (JPT), Luhya from Webuye, Kenya (LWK), Mexican ancestry from Los Angeles, California (MEX), Maasai from Kinyawa, Kenya (MKK), Toscans from Italy (TSI), and Yoruba from Ibadan, Nigeria (YRI). My research focuses on four major populations in the USA: ASW, CEU, CHD, and MEX. The HapMap Phase III includes the complete genotype data of 52, 120, 85 and 50 unrelated individuals from ASW, CEU, CHD, and MEX populations, respectively.

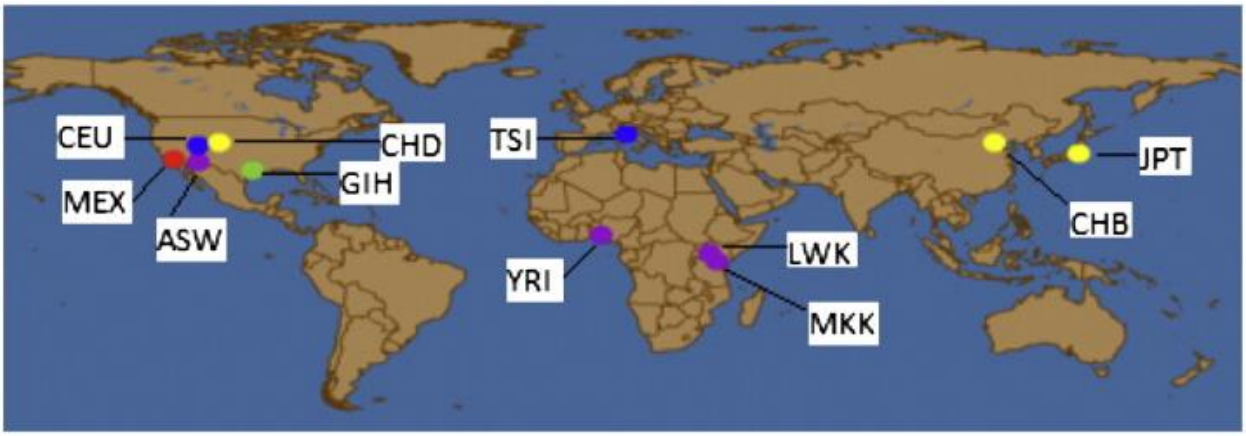


Figure 5. Geographic map of the HapMap phase III world populations (the figure is from reference 41).

On this background, the doctoral research described herein was conducted under the hypotheses that δ and F_{ST} perform better than In in the selection of AIMs, and highly informative AIMs can be selected among four major US populations by using these three marker informativeness measures.

Chapter 1 is the peer-reviewed paper titled "Selection of highly informative SNP markers for population affiliation of major US populations" (Zeng X, Chakraborty R, King JL, LaRue B, Moura-Neto RS, Budowle B. Int. J. Legal Med. 2015 Dec 8. [Epub ahead of print]). In this article, δ, F_{ST}, and In were compared directly for AIMs selection among four major United States populations, i.e., African American, US Caucasian, East Asian and Hispanic American. The three measure values (δ, F_{ST}, In) of each SNP were computed first for each pairwise population comparison, and then markers were ranked based on these measure values. The top 30 AIMs, for each measure in each pairwise population comparison, were chosen and any markers in linkage disequilibrium (LD) were removed. The minimum number of markers to discriminate each pair of populations was identified for each measure based on principal component analysis (PCA), Receiver Operating Characteristics (ROC) curve, and the maximum Matthews correlation coefficient (MCC). Finally, the top markers from six pairwise population comparisons were pooled based on the three measures and evaluated as individual panels. After removing duplicated SNPs and replacing SNPs that were in LD, the resultant total number of markers in the AIMs panels selected by δ, F_{ST} and In was 24, 23, and 23, respectively. The PCA cluster results indicated that the F_{ST} panel performed slightly better than the δ panel and significantly better than the In panel. Therefore, the 23 AIMs selected by the F_{ST} measure were used to characterize the four major American populations. The PCA and discriminant function analysis

(DFA) results indicated that these 23 AIMs can correctly assign individuals to the major population categories *in silico*.

Chapter 2 is the peer-reviewed paper titled "Empirical Testing of a 23-AIMs panel of SNPs for ancestry evaluations in Four Major US Populations" (Zeng X, Warshauer DH, King JL, Churchill JD, Chakraborty R, Budowle B.) submitted to International Journal of Legal Medicine. In this study, the 23 SNPs (selected in chapter 1) were tested for their performance in ancestry evaluations. DNA was extracted from 189 unrelated individuals collected from four American populations (African Americans, Asians, Caucasians, and Hispanics). Population affinity was based on self-declaration. The Nextera™ Rapid Capture Custom Enrichment kit (Illumina) was used to enrich the target SNPs. Custom oligonucleotide probes were designed using Design Studio v1.5 (Illumina). The Illumina MiSeq system was used to perform sequencing, and sequence data were analyzed using MiSeq Reporter and Genome Analysis Toolkit (GATK). The genotype data of 23 SNPs were generated for 189 individuals. SNP rs12149261 deviated from Hardy-Weinberg equilibrium (HWE) in three populations (Asian, Caucasian and Hispanic American) and also exhibited significant LD with the other 22 SNPs. BLAST results indicated that SNP rs12149261 (located on chromosome 16) had a duplicate region on chromosome 1, and thus its genotype was a combination of two SNP sites. No other SNPs departed from HWE in the four populations, and only five SNP pairs showed significant LD (after this problematic SNP was removed from my panel). PCA results showed that eight individuals were not assigned to the expected major population categories based on self-declaration or the population labeled on the anonymous sample. The assignments of the 22 AIMs for these samples were consistent with AIMs results from the ForenSeq^{TM} panel. After removing these eight samples, no detectable LDs were observed in the four populations. The results

indicated that the 22 AIMs can correctly assign the 181 individuals to the four major US population categories. These 22 SNPs can be used for potential forensic identification purposes in major US populations.

To summarize, the body of work included in this thesis:

a. Identification of the best method (F_{ST}) for AIMs selection among the three measures and development of a robust panel of 22 AIMs to characterize four major US populations
b. Evaluation the full efficacy of the AIMs panel in the four major US populations and the 22 AIMs can correctly assign individuals to the four major US population categories.

This research will contribute to the potential AIMs pool and support the application of AIMs panel in forensic community and for population stratification studies.

References:

(1) Linse KD (2012) Genes that define the shape of our face. http://blog-biosyn.com/2012/11/28/123/

(2) Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516-1517

(3) Schlesselman JJ (1982) Case-control studies: design, conduct, analysis. Oxford: Oxford University Press

(4) Garcia-Closas M, Chanock S (2008) Genetic susceptibility loci for breast cancer by estrogen receptor status. Clin Cancer Res 14:8000-8009

(5) 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

(6) Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW (2002) Genetic structure of human populations. Science 298:2381-2385

(7) 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

(8) 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

(9) Slatkin M (2007) Inbreeding coefficients and coalescence times. Genet Res 89:479-487

(10) Goddard KA, Hopkins PJ, Hall JM, Witte JS (2000) Linkage disequilibrium and allele-frequency distributions for 114 single-nucleotide polymorphisms in five populations. Am J Hum Genet 66:216-234

(11) Stephens JC, Schneider JA, Tanguay DA, Choi J, Acharya T, Stanley SE, Jiang R, Messer CJ, Chew A, Han JH, Duan J, Carr JL, Lee MS, Koshy B, Kumar AM, Zhang G, Newell WR, Windemuth A, Xu C, Kalbfleisch TS, Shaner SL, Arnold K, Schulz V, Drysdale CM, Nandabalan K, Judson RS, Ruano G, Vovis GF (2001) Haplotype variation and linkage disequilibrium in 313 human genes. Science 293:489-493

(12) Wacholder S, Rothman N, Caporaso N (2000) Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. J Natl Cancer Inst 92:1151-1158

(13) Knowler WC, Williams RC, Pettitt DJ, Steinberg AG (1988) Gm3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. Am J Hum Genet 43:520-526

(14) Williams RC, Long JC, Hanson RL, Sievers ML, Knowler WC (2000) Individual estimates of European genetic admixture associated with lower body-mass index, plasma glucose, and prevalence of type 2 diabetes in Pima Indians. Am J Hum Genet 66:527-538

(15) Jobling MA, Gill P (2004) Encoded evidence: DNA in forensic analysis. Nat Rev Genet 5:739-751

(16) 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

(17) Shriver MD, Kittles RA (2004) Genetic ancestry and the search for personalized genetic histories. Nat Rev Genet 5:611-618

(18) 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

(19) Jobling MA, Tyler-Smith C (2003) The human Y chromosome: an evolutionary marker comes of age. Nat Rev Genet 4:598-612

(20) Vigilant L, Stoneking M, Harpending H, Hawkes K, Wilson AC (1991) African populations and the evolution of human mitochondrial DNA. Science 253:1503-1507

(21) 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

(22) Jin L, Chakraborty R (1995) Population structure, stepwise mutations, heterozygote deficiency and their implications in DNA forensics. Heredity 74:274-285

(23) 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

(24) 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

(25) International HapMap Consortium (2003) The International HapMap Project. Nature 426:789-796

(26) 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

(27) 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

(28) 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

(29) 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

(30) 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

(31) 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 [Epub ahead of print]

(32) Rosenberg NA, Li LM, Ward R, Pritchard JK (2003) Informativeness of genetic markers for inference of ancestry. Am J Hum Genet 73:1402-1422

(33) Wright S (1950) Genetical structure of populations. Nature 166:247-249

(34) Shriver MD, Smith MW, Jin L, Marcini A, Akey JM, Deka R, Ferrell RE (1997) Ethnicaffiliation estimation by use of population-specific DNA markers. Am J Hum Genet 60:957-964

(35) Parra EJ, Marcini A, Akey J, Martinson J, Batzer MA, Cooper R, Forrester T, Allison DB, Deka R, Ferrell RE, Shriver MD (1998) Estimating African American admixture proportions by use of population-specific alleles. Am J Hum Genet 63:1839-1851

(36) Collins-Schramm HE, Phillips CM, Operario DJ, Lee JS, Weber JL, Hanson RL, Knowler WC, Cooper R, Li H, Seldin MF (2002) Ethnic-difference markers for use in mapping by admixture linkage disequilibrium. Am J Hum Genet 70:737-750

(37) Baye TM, Tiwari HK, Allison DB, Go RC (2009) Database mining for selection of SNP markers useful in admixture mapping. BioData Min 2:1

(38) Tian C, Hinds DA, Shigeta R, Adler SG, Lee A, Pahl MV, Silva G, Belmont JW, Hanson RL, Knowler WC, Gregersen PK, Ballinger DG, Seldin MF (2007) A genomewide single-nucleotide-polymorphism panel for Mexican American admixture mapping. Am J Hum Genet 80:1014-1023

(39) Mao X, Bigham AW, Mei R, Gutierrez G, Weiss KM, Brutsaert TD, Leon-Velarde F, Moore LG, Vargas E, McKeigue PM, Shriver MD, Parra EJ (2007) A genomewide admixture mapping panel for Hispanic/Latino populations. Am J Hum Genet 80:1171-1178

(40) Price AL, Patterson N, Yu F, Cox DR, Waliszewska A, McDonald GJ, Tandon A, Schirmer C, Neubauer J, Bedoya G, Duque C, Villegas A, Bortolini MC, Salzano FM, Gallo C, Mazzotti G, Tello-Ruiz M, Riba L, Aguilar-Salinas CA, Canizales-Quinteros S, Menjivar M, Klitz W, Henderson B, Haiman CA, Winkler C, Tusie-Luna T, Ruiz-Linares A, Reich D (2007) A genomewide admixture map for Latino populations. Am. J Hum Genet 80:1024-1036

(41) Amirisetty S, Hershey GK, Baye TM (2012) AncestrySNPminer: a bioinformatics tool to retrieve and develop ancestry informative SNP panels. Genomics 100:57-63

CHAPTER 1

Selection of Highly Informative SNP Markers for Population Affiliation of Major US Populations

Published in International Journal of Legal Medicine
2016, 130:341-352

Xiangpei Zeng
Ranajit Chakraborty
Jonathan L. King
Bobby LaRue
Rodrigo S. Moura-Neto
Bruce Budowle

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: δ, F_{ST}, and 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 United States 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}

1.1 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 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 SNPs 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 U.S. 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 U.S. 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.

1.2. Materials and Methods:

1.2.1 Population samples

The HapMap project (15) contains comprehensive SNP data on the four major U.S. 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/).

1.2.2 AIMs selection

The measures used for AIMs selection were Absolute Allele Frequency Differences (δ), F statistics (F_{ST}), and Informativeness for Assignment (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.

1.2.3 Statistical power of AIMs

The number of AIMs, that 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 data set 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.

1.3. Results and Discussions

1.3.1 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, 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 1-6). 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 thirty 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 2 to 9 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 (Figure 1). Figure 2a showed 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 (Figure 2a). In addition, PCA clusters showed that CEU was generally distinguished from MEX individuals using the genotype data of these eight AIMs (Figure 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).

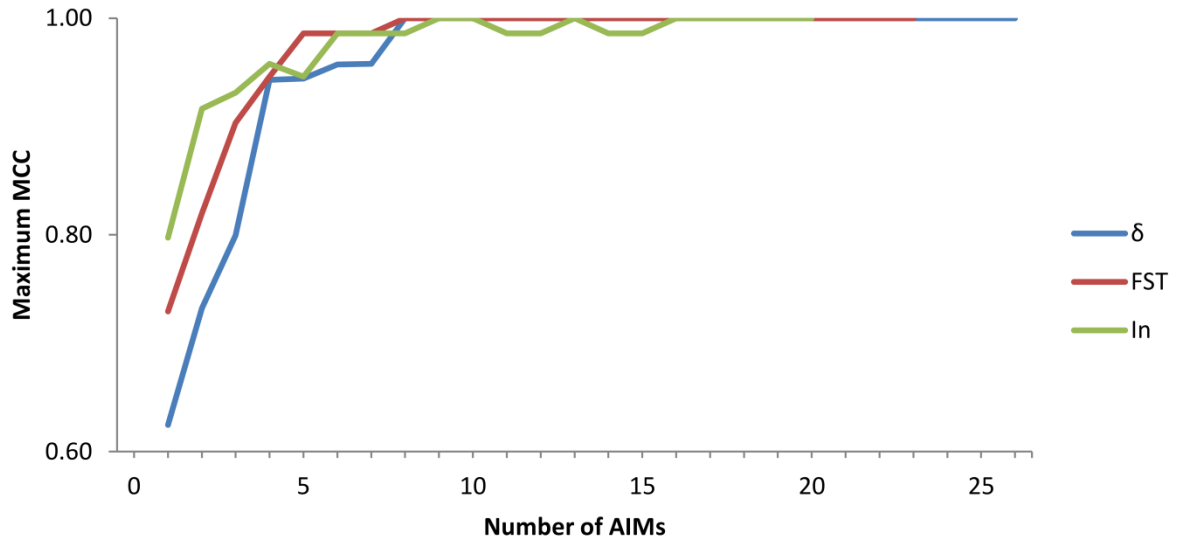


Figure 1. The three MCC curves generated by δ, F_{ST}, and In measures for MEX and CEU.

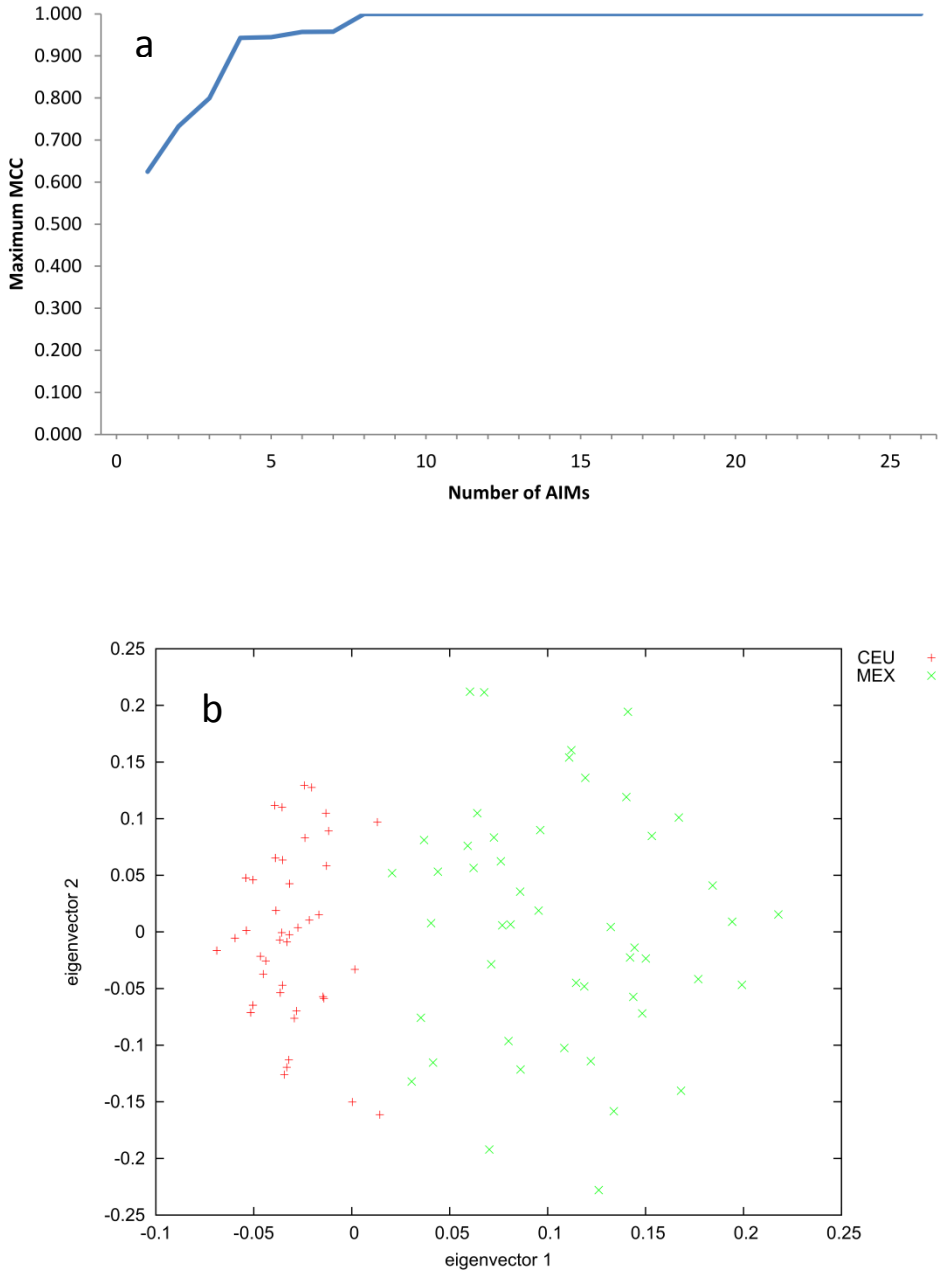


Figure 2. The AIMs panel that was selected by the δ measure to separate CEU and MEX. a The classification accuracy of 170 samples (CEU and MEX) utilizing a varied number of selected AIMs; b PCA clusters of two populations by using genotype data of top eight AIMs identified in a.

1.3.2 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, 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 (Figure 3a-3b). 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 (Figure 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 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 (Figure 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 (Figure 4b).

Table 1. The final panels of AIMs identified by the three measures δ, F_{ST} and In to distinguish the four major U.S. 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	103693822	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								

Table 2. Shared number of AIMs between δ, F_{ST}, and In among the top 2 to 9 markers for 6 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

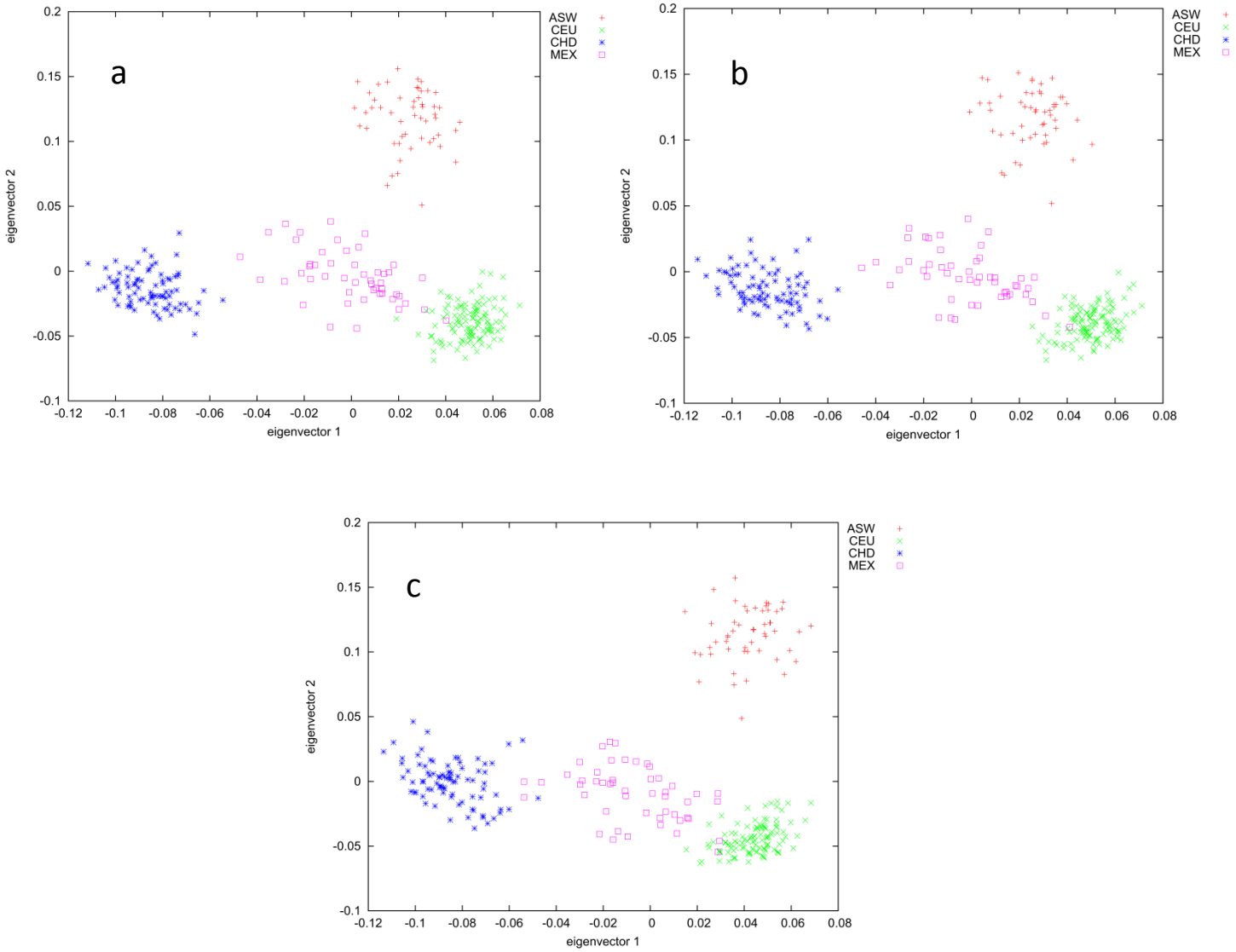
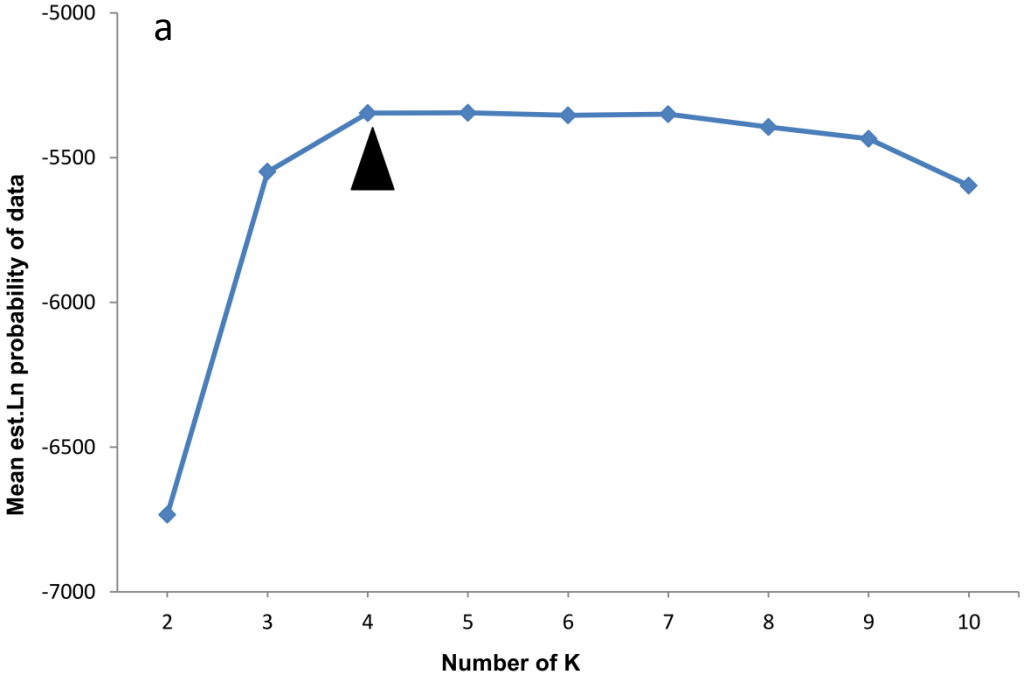


Figure 3. The PCA clusters of the AIMs panels that were selected by a δ, b F_{ST}, and c In measures, respectively.



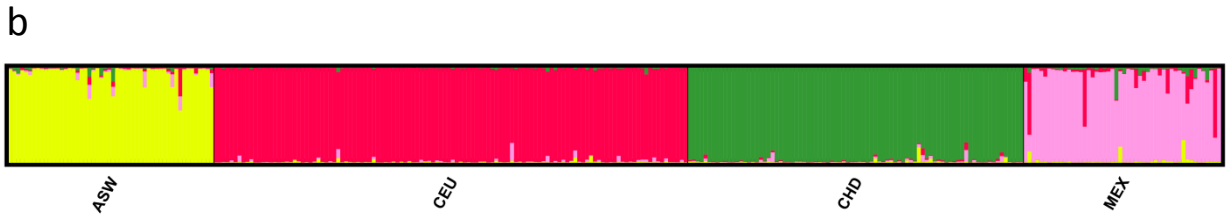


Figure 4. Analyses of four major US populations from HapMap using the AIMs panel selected by F_{ST}. a Indicated that the optimal number of K was 4. B The STRUCTURE cluster plots of four populations (ASW, CEU, CHD, and MEX).

1.3.3 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 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 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 (Figure 5a). The DFA results assigned all YRI individuals to ASW group (Figure 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 (Figure 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 (Figure 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 (Figure 5c-5d). 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 (Figure 6c-6d, 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-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 (Figure 7a, Figure 8a, Supplemental Table 10). The majority of GBR individuals were located within the 95% confidence interval of the Caucasian group in PCA (Figure 7b), and all of them were assigned as Caucasians by DFA

(Figure 8b, Supplemental Table 10). Eight CHB individuals fell outside the 95% confidence interval of CHD in PCA (Figure 7c), but all of them were assigned as East Asians in DFA (Figure 8c, Supplemental Table 10). CLM individuals were the most difficult to assign. They were classified as African Americans, Caucasians, and Hispanic Americans (Figure 7d). According to Bushnell et al. (39) (2010), 86% of Colombians 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 (Figure 7d). The DFA provided results of 4, 26, 64 individuals assigned as African Americans, Caucasians and Hispanic Americans, respectively (Figure 8d, Supplemental Table 16). 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 U.S. and may not precisely explain the genetic variations of the Hispanic populations in Central America and South America. Both HapMap and 1000 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 individuals were within the 95% confidence interval of Hispanic American in PCA (Figure 7e). All individuals were classified as Hispanic Americans by DFA (Figure 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 U.S. populations. Therefore, empirical testing of the full set of these AIMs will further evaluate the efficiency of the panel.

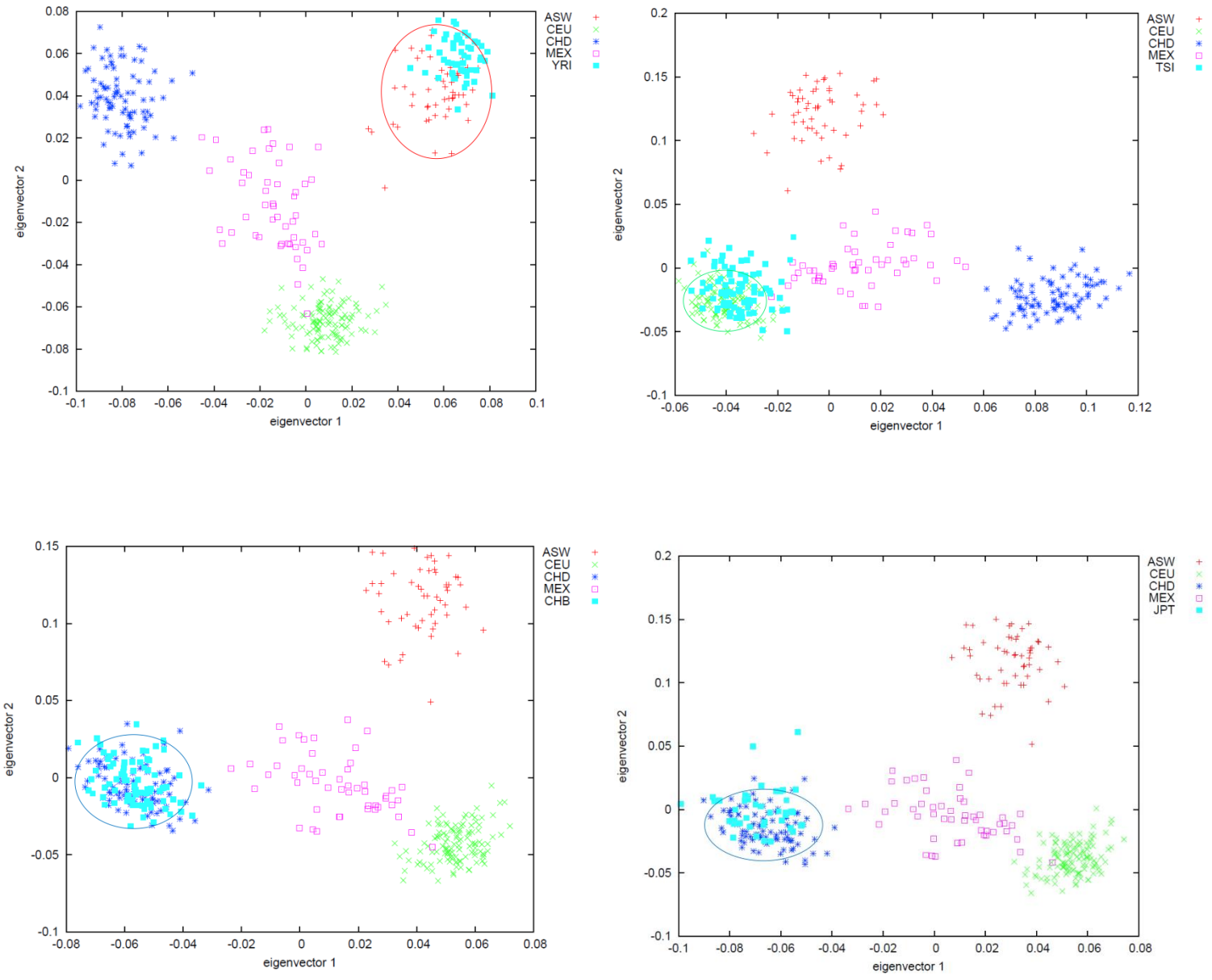


Figure 5. Population classification of four global populations from HapMap using PCA. a–d represented YRI, TSI, CHB, and JPT, respectively.

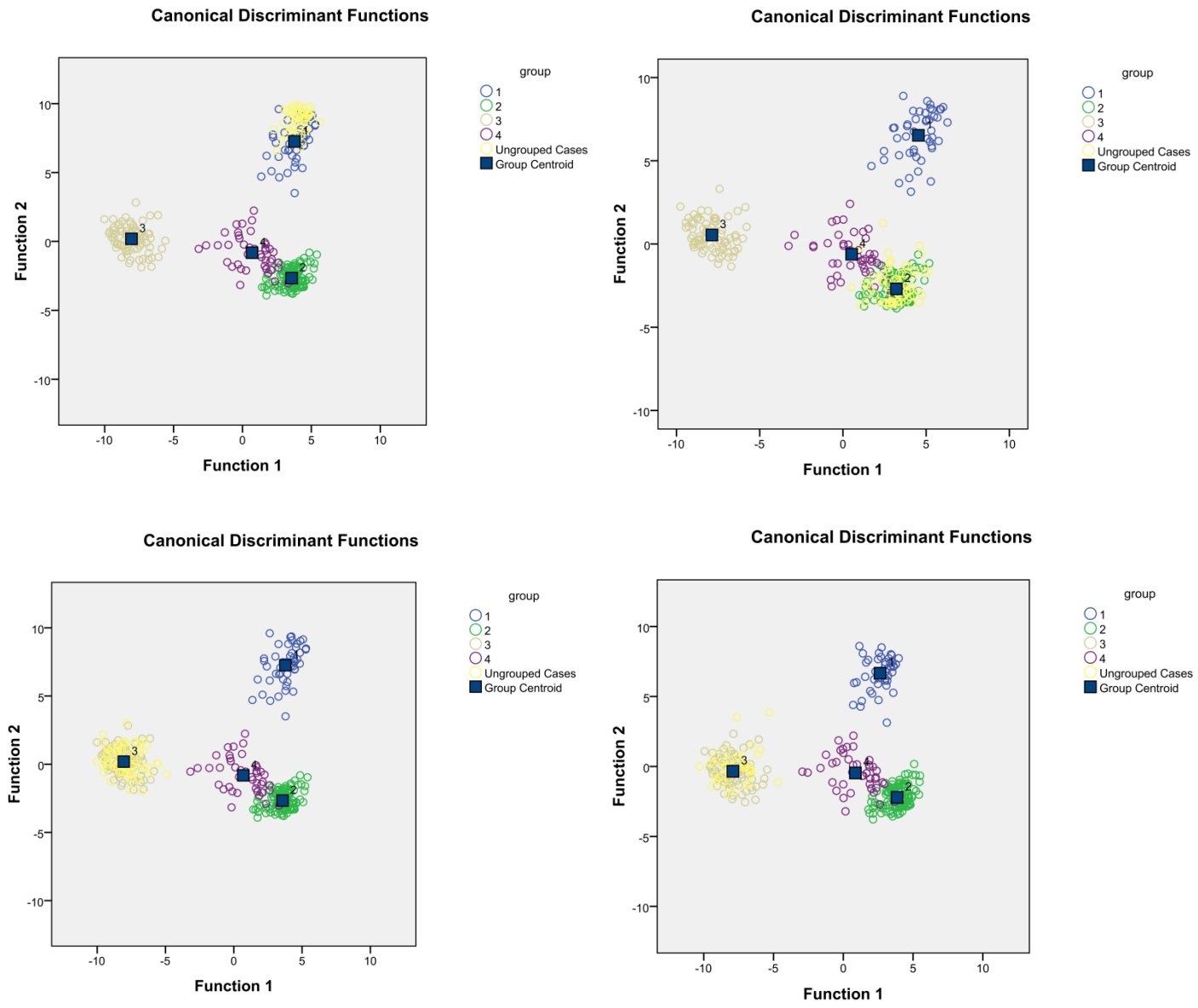


Figure 6. Population classification of four major populations from HapMap using DFA. Groups 1–4 represented ASW, CEU, CHD, and MEX, respectively. The ungrouped cases in a–d were individuals of YRI, TSI, CHB, and JPT, respectively. Some SNPs were excluded from the analysis because of missing data. Overall,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

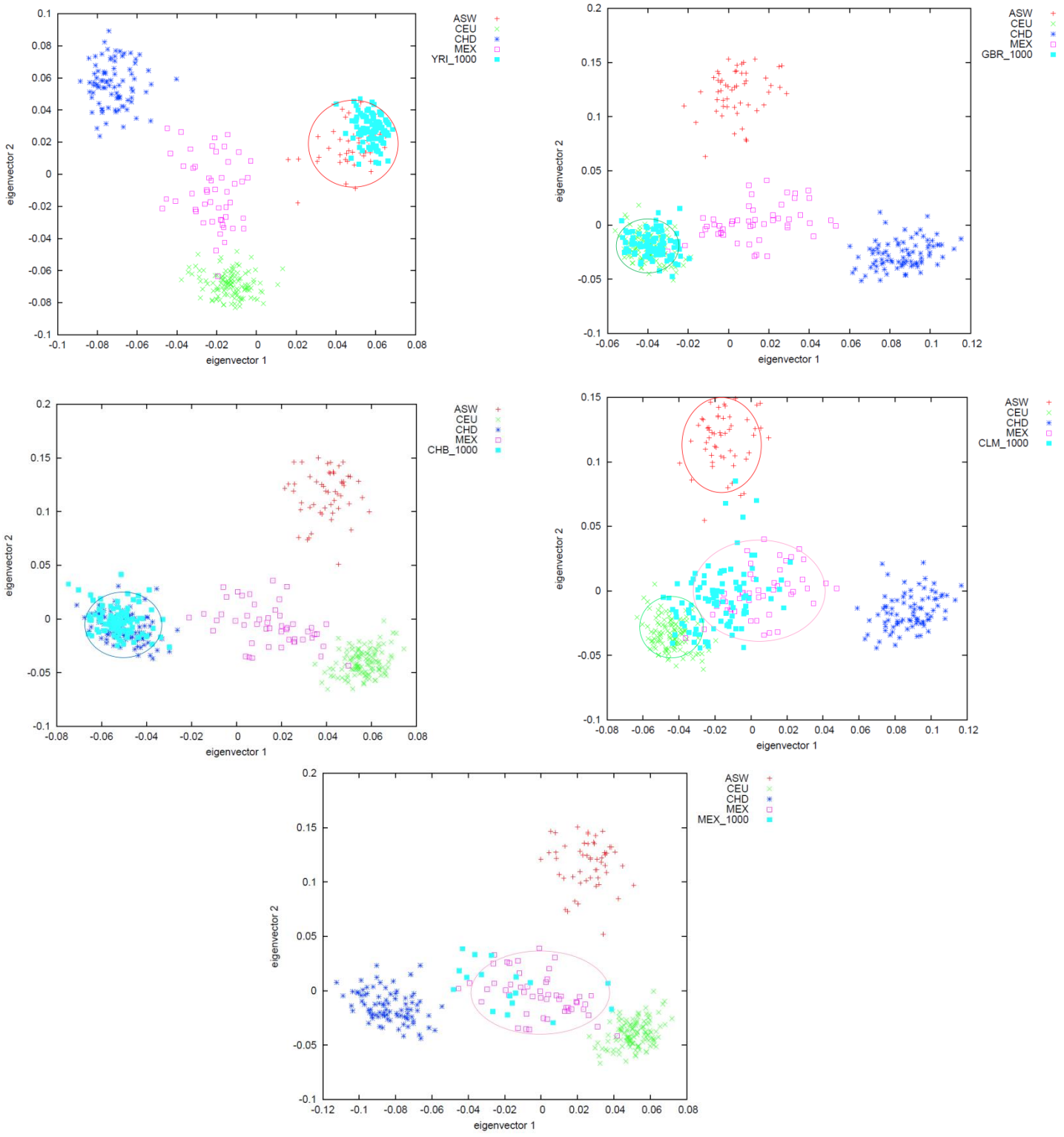


Figure 7. Population classification of five populations from 1000 Genomes using PCA. a–e represented YRI, GBR, CHB, CLM, and MEX, respectively.

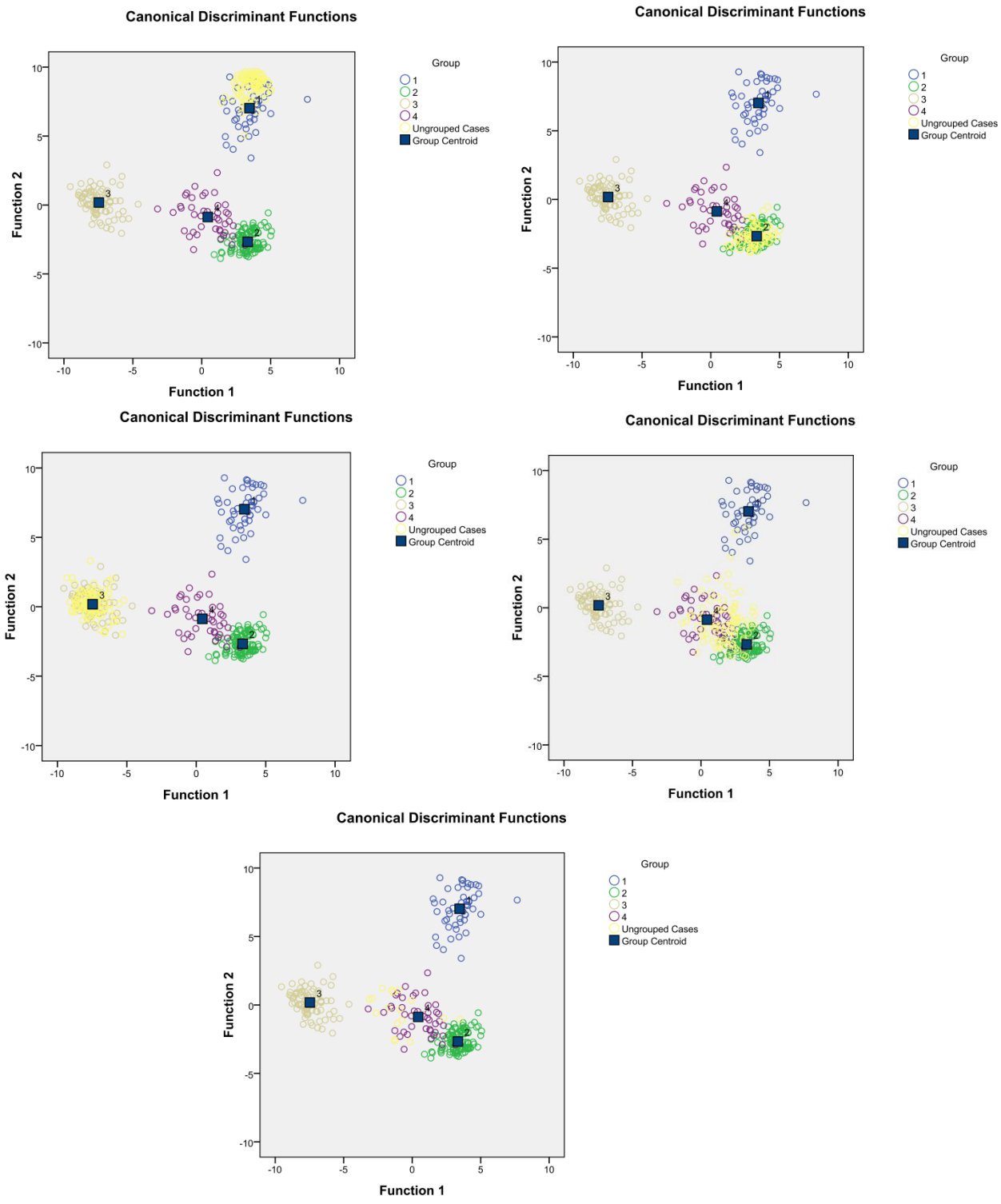


Figure 8. Population classification of five populations from 1000 Genomes using DFA. Groups 1–4 represented ASW, CEU, CHD, and MEX, respectively. The ungrouped cases in a–e were individuals of YRI, GBR, CHB, CLM, and MEX, respectively. Three SNPs were excluded from the analysis because of missing data.

1.3.4 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 U.S. 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

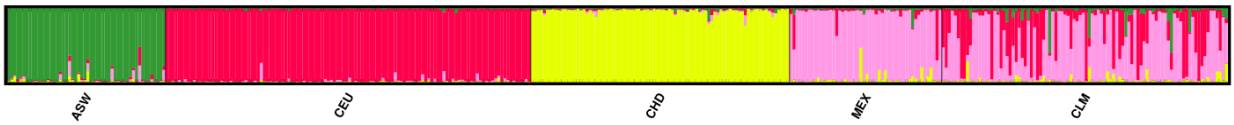
number of AIMs to distinguish four U.S. populations was sought. In our panel, there are four SNPs from chromosome 15, and they are located within 3-8 Mb of each other. Although they are not in LD within the four U.S. populations, it is possible that they may affect admixture membership estimation in other populations.

1.4. 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 U.S. 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.

Supplemental information on Colombians from Medellin, Colombia (CLM):

The STRUCTURE results also indicated that the majority of CLM individuals were assigned to the CEU or MEX groups, and a few individuals were assigned to the African American group (Supplemental Figure 1). CLM individuals were more similar to MEX than other groups based on STURCTURE, but had a higher Caucasian contribution than that of the MEX group.



Supplemental Figure 1. Analyses of CLM using the AIMs panel selected by F_{ST}. The CLM individuals were assigned to ASW, CEU, and MEX groups.

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 (Epub ahead of print)

(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://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, Washingtion 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 bio-geographical 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

1.5. Supplemental Materials

Supplemental Table 1. The top AIMs selected by three measures from ASW and CEU after H-W and LD selection. The top thirty SNPs were reduced to 26, 26, 26 AIMs by δ, F_{ST}, and In, respectively. 1-3 indicated the SNPs that were in LD, and the highlighted ones (Bold and Italic) were removed. The physical distances of SNPs were downloaded from GRCh37.p13 (hg 19).

δ			F_{ST}			In		
rsID	chr	pos	rsID	chr	pos	rsID	chr	pos
rs1834640	chr15	48392165	rs1834640	chr15	48392165	rs1834640	chr15	48392165
rs1288097[1]	chr15	45141373	rs1288097[1]	chr15	45141373	rs1288097[1]	chr15	45141373
rs12594483[1]	***chr15***	***43021986***	***rs12594483[1]***	***chr15***	***43021986***	***rs12594483[1]***	***chr15***	***43021986***
rs6674304[2]	chr1	116887742	rs6674304[2]	chr1	116887742	rs6674304[2]	chr1	116887742
rs7689609[3]	chr4	72083374	rs7689609[3]	chr4	72083374	rs1572510	chr13	105381134
rs798790	chr14	62037996	rs798790	chr14	62037996	rs7689609[3]	chr4	72083374
rs199138[1]	***chr15***	***45387550***	***rs199138[1]***	***chr15***	***45387550***	rs9321552	chr6	136481612
rs974828	chr15	60855555	rs974828	chr15	60855555	rs798790	chr14	62037996
rs4839518[2]	***chr1***	***116745951***	rs1572510	chr13	105381134	***rs199138[1]***	***chr15***	***45387550***
rs2725264	chr4	89026109	***rs6446975[3]***	***chr4***	***75036044***	rs7662047	chr4	103091730
rs6446975[3]	***chr4***	***75036044***	***rs4839518[2]***	***chr1***	***116745951***	rs2615876	chr10	117665860
rs1572510	chr13	105381134	rs2725264	chr4	89026109	rs974828	chr15	60855555
rs4789659	chr17	72398336	rs4789659	chr17	72398336	rs11845995	chr14	105930923
rs2407548	chr4	151850524	rs11845995	chr14	105930923	rs7018273	chr8	82004857
rs1827950	chr4	117098482	rs9321552	chr6	136481612	rs11995470	chr8	10555762
rs10313	chr5	180682405	rs2407548	chr4	151850524	rs2725264	chr4	89026109
rs9582807	chr13	104780173	rs1827950	chr4	117098482	***rs6446975[3]***	***chr4***	***75036044***
rs11845995	chr14	105930923	rs10313	chr5	180682405	rs1012023	chr14	33956209
rs10853040	chr17	60472218	rs9582807	chr13	104780173	rs676281	chr15	34260170
rs3768641	chr2	72368190	rs7662047	chr4	103091730	rs1851426	chr7	99382936

rs981375	chr2	143297583	rs2615876	chr10	117665860	rs1805972	chr5	9203914
rs4851687	chr2	104957361	rs10853040	chr17	60472218	***rs4839518[2]***	***chr1***	***116745951***
rs9321552	chr6	136481612	rs1592672	chr12	80128593	rs4789659	chr17	72398336
rs1179640	chr7	75238617	rs676281	chr15	34260170	rs2407548	chr4	151850524
rs10115397	chr9	14215031	rs1012023	chr14	33956209	rs10313	chr5	180682405
rs1592672	chr12	80128593	rs7018273	chr8	82004857	rs9582807	chr13	104780173
rs1622710	chr13	60106925	rs11995470	chr8	10555762	rs1592672	chr12	80128593
rs676281	chr15	34260170	rs3768641	chr2	72368190	rs2294306	chr20	7869025
rs2934193	chr15	48259719	rs981375	chr2	143297583	rs8073072	chr17	29350769
rs11264110	chr1	35636227	rs1179640	chr7	75238617	rs7276293	chr21	19126079

Supplemental Table 2. The top AIMs selected by three measures from ASW and CHD after H-W and LD selection. The top thirty SNPs were reduced to 24, 25, 26 AIMs by δ, F_{ST}, and In, respectively. 1-5 indicated the SNPs that were in LD, and the highlighted ones (Bold and Italic) were removed. The physical distances of SNPs were downloaded from GRCh37.p13 (hg 19).

δ			F_{ST}			In		
rsID	chr	pos	rsID	chr	pos	rsID	chr	pos
rs7165971	chr15	55921013	rs7165971	chr15	55921013	rs7165971	chr15	55921013
rs745767[1]	chr2	177825415	rs745767[1]	chr2	177825415	rs745767[1]	chr2	177825415
rs13021399[2]	chr2	109006665	rs13021399[2]	chr2	109006665	rs13021399[2]	chr2	109006665
rs11123706[2]	***chr2***	***109150164***	***rs11123706[2]***	***chr2***	***109150164***	***rs11123706[2]***	***chr2***	***109150164***
rs2104483	chr10	92078697	rs2104483	chr10	92078697	rs2104483	chr10	92078697
rs6500380[3]	chr16	48375777	rs6500380[3]	chr16	48375777	rs2129801	chr7	135815196
rs2129801	chr7	135815196	rs2129801	chr7	135815196	rs6500380[3]	chr16	48375777
rs11187296	chr10	94911072	rs11187296	chr10	94911072	rs2192015[4]	chr2	72501137
rs2192015[4]	chr2	72501137	rs2192015[4]	chr2	72501137	rs522982	chr19	2984460
rs2037044[1]	***chr2***	***177682929***	***rs17822931[3]***	***chr16***	***48258198***	rs11187296	chr10	94911072
rs7151509	chr14	85727121	rs4429562	chr22	42892596	rs7825690	chr8	10858257
rs17822931[3]	***chr16***	***48258198***	***rs2037044[1]***	***chr2***	***177682929***	***rs17822931[3]***	***chr16***	***48258198***
rs522982	chr19	2984460	rs522982	chr19	2984460	rs4429562	chr22	42892596
rs4429562	chr22	42892596	rs7151509	chr14	85727121	***rs2037044[1]***	***chr2***	***177682929***
rs12075	chr1	159175354	rs12075	chr1	159175354	rs732381	chr22	40051166
rs6580054	chr5	153548134	rs1325421	chr6	105891508	rs653220	chr2	72707874
rs1325421	chr6	105891508	rs732381	chr22	40051166	rs7151509	chr14	85727121
rs2896733	chr11	25269127	rs6580054	chr5	153548134	rs12075	chr1	159175354
rs732381	chr22	40051166	rs2896733	chr11	25269127	rs1325421	chr6	105891508
rs2306125	chr1	155025361	***rs13006497[4]***	***chr2***	***72882725***	rs6580054	chr5	153548134
rs13006497[4]	***chr2***	***72882725***	rs12439722	chr15	63941116	rs2896733	chr11	25269127
rs12473565	chr2	175163335	rs2306125	chr1	155025361	rs2736306	chr8	11239762
rs12680762	chr8	11332026	rs11123717	chr2	109526138	rs2306125	chr1	155025361

rs9576028	chr13	86731387	rs12473565	chr2	175163335	rs12473565	chr2	175163335
rs12439722	chr15	63941116	rs12680762	chr8	11332026	rs9576028	chr13	86731387
rs12087334[5]	chr1	116887455	rs9576028	chr13	86731387	rs12439722	chr15	63941116
rs12035171[5]	***chr1***	***76490088***	rs2273015	chr1	27153958	rs4497887	chr2	125859777
rs2273015	chr1	27153958	rs4497887	chr2	125859777	rs12598978	chr16	30482540
rs1508060[4]	***chr2***	***73078673***	rs12087334[5]	chr1	116887455	***rs4852886[4]***	***chr2***	***72882532***
rs4497887	chr2	125859777	***rs12035171[5]***	***chr1***	***76490088***	rs12359102	chr10	114793451

Supplemental Table 3. The top AIMs selected by three measures from ASW and MEX after H-W and LD selection. The top thirty SNPs were reduced to 27, 26, 25 AIMs by δ, F_{ST}, and In, respectively. 1-4 indicated the SNPs that were in LD, and the highlighted ones (Bold and Italic) were removed. The physical distances of SNPs were downloaded from GRCh37.p13 (hg 19).

δ			F_{ST}			In		
rsID	chr	pos	rsID	chr	pos	rsID	chr	pos
rs12087334[1]	chr1	116887455	rs12087334[1]	chr1	116887455	rs12087334[1]	chr1	116887455
rs12149261	chr16	70998145	rs12149261	chr16	70998145	rs11845995	chr14	105930923
rs1827950	chr4	117098482	rs11845995	chr14	105930923	rs12149261	chr16	70998145
rs11845995	chr14	105930923	rs1827950	chr4	117098482	rs1827950	chr4	117098482
rs1025104	chr2	86350664	rs1849384	chr12	85833160	rs1849384	chr12	85833160
rs1507086[2]	chr4	41956413	rs2125953	chr8	2978749	rs951954	chr2	110459505
rs2125953	chr8	2978749	rs1091679	chr14	61978278	rs2125953	chr8	2978749
rs1849384	chr12	85833160	rs974828[2]	chr15	60855555	rs1091679	chr14	61978278
rs1091679	chr14	61978278	rs8030587	chr15	42981806	rs974828[2]	chr15	60855555
rs974828[3]	chr15	60855555	***rs2414418[2]***	***chr15***	***55809660***	rs8030587	chr15	42981806
rs8030587	chr15	42981806	rs7185636	chr16	19808163	***rs2414418[2]***	***chr15***	***55809660***
rs2414418[3]	***chr15***	***55809660***	***rs7413197[1]***	***chr1***	***116750625***	rs7185636	chr16	19808163
rs7185636	chr16	19808163	rs951954	chr2	110459505	***rs7413197[1]***	***chr1***	***116750625***
rs7413197[1]	***chr1***	***116750625***	rs1025104	chr2	86350664	rs7212298[3]	chr17	11025242
rs2658600	chr10	59695419	rs1507086[3]	chr4	41956413	rs8015967	chr14	57687997
rs7951580	chr11	84188490	rs2658600	chr10	59695419	rs16930172	chr10	74684417
rs8015967	chr14	57687997	rs7951580[4]	chr11	84188490	rs6485671	chr11	46279815
rs12603916	chr17	53646029	rs8015967	chr14	57687997	rs679882	chr15	40759347
rs951954	chr2	110459505	rs12603916	chr17	53646029	rs1025104	chr2	86350664
rs3768641	chr2	72368190	rs16930172	chr10	74684417	rs1507086[4]	chr4	41956413
rs10025567	chr4	91453933	rs6485671	chr11	46279815	rs12603916	chr17	53646029
rs10810130	chr9	14304973	rs3768641	chr2	72368190	rs10025567	chr4	91453933
rs16930172	chr10	74684417	rs10025567	chr4	91453933	rs12711897	chr2	118959792

rs6485671	chr11	46279815	rs2053918	chr2	98730815	rs2053918	chr2	98730815
rs11035407	chr11	39704979	***rs7212298[4]***	***chr17***	***11025242***	rs6900027	chr6	10652350
rs6676438	chr1	161983089	rs10810130	chr9	14304973	rs2658600	chr10	59695419
rs7594227	chr2	97605748	rs11035407	chr11	39704979	***rs7951580[3]***	***chr11***	***84188490***
rs981375	chr2	143297583	rs6676438	chr1	161983089	rs3768641	chr2	72368190
rs4149436	chr2	108999786	rs981375	chr2	143297583	***rs1558219[3]***	***chr17***	***11167687***
rs4623048[2]	***chr4***	***41833487***	***rs4623048[3]***	***chr4***	***41833487***	***rs4623048[4]***	***chr4***	***41833487***

Supplemental Table 4. The top AIMs selected by three measures from CEU and CHD after H-W and LD selection. The top thirty SNPs were reduced to 27, 27, 29 AIMs by δ, F_{ST}, and In, respectively. 1-3 indicated the SNPs that were in LD, and the highlighted ones (Bold and Italic) were removed. The physical distances of SNPs were downloaded from GRCh37.p13 (hg 19).

δ			F_{ST}			In		
rsID	chr	pos	rsID	chr	pos	rsID	chr	pos
rs4429562	chr22	42892596	rs4429562	chr22	42892596	rs4429562	chr22	42892596
rs1547843	chr10	91738263	rs11126303	chr2	26173503	rs1834640	chr15	48392165
rs11126303	chr2	26173503	rs1834640	chr15	48392165	rs11126303	chr2	26173503
rs7919895	chr10	28358659	rs1547843	chr10	91738263	rs35389	chr5	33954880
rs11648965	chr16	86084146	rs35389	chr5	33954880	rs1547843	chr10	91738263
rs6437783[1]	chr3	108172817	rs1153105	chr1	1415099	rs1153105	chr1	1415099
rs11187296	chr10	94911072	rs6437783[1]	chr3	108172817	rs6437783	chr3	108172817
rs6137010	chr20	2090118	rs7919895	chr10	28358659	rs11804831	chr1	1194804
rs8024070[2]	chr15	63900519	rs11648965	chr16	86084146	rs7919895	chr10	28358659
rs1898213[3]	chr2	126174341	rs11804831	chr1	1194804	rs11648965	chr16	86084146
rs6710520[3]	***chr2***	***126288060***	rs11187296	chr10	94911072	rs6137010	chr20	2090118
rs6137197	chr20	20955093	rs6137010	chr20	2090118	rs10962612	chr9	16804167
rs2605419	chr3	123414429	rs6710520[2]	chr2	126288060	rs11187296	chr10	94911072
rs4657449	chr1	165465281	rs2605419	chr3	123414429	rs2605419	chr3	123414429
rs482000	chr1	234330527	rs6137197	chr20	20955093	rs6137197	chr20	20955093
rs6500380	chr16	48375777	***rs1898213[2]***	***chr2***	***126174341***	rs4722760	chr7	28172183
rs532143	chr19	34727835	rs8024070[3]	chr15	63900519	rs532143	chr19	34727835
rs4497887	chr2	125859777	rs482000	chr1	234330527	rs6710520[1]	chr2	126288060
rs2700367	chr3	123624279	rs4657449	chr1	165465281	rs8024070	chr15	63900519
rs4722760	chr7	28172183	rs2700367	chr3	123624279	rs2700367	chr3	123624279
rs11778591	chr8	12720349	rs4722760	chr7	28172183	rs609096	chr5	4938130
rs11204849	chr1	151538412	rs532143	chr19	34727835	rs5750871	chr22	40069449
rs2917454	chr10	78892415	rs10962612	chr9	16804167	***rs1898213[1]***	***chr2***	***126174341***

rs12570042	chr10	90793388	rs4497887	chr2	125859777	rs482000	chr1	234330527
rs9572782	chr13	72339141	rs6500380	chr16	48375777	rs4657449	chr1	165465281
rs7172848[2]	***chr15***	***64031754***	rs609096	chr5	4938130	rs6500380	chr16	48375777
rs6751451	chr2	154731306	rs5750871	chr22	40069449	rs8032157	chr15	64480888
rs5750871	chr22	40069449	rs11778591	chr8	12720349	rs716957	chr14	99476306
rs1374821[1]	***chr3***	***108474507***	***rs1374821[1]***	***chr3***	***108474507***	rs4970364	chr1	1174282
rs609096	chr5	4938130	***rs7172848[3]***	***chr15***	***64031754***	rs2233072	chr17	19281828

Supplemental Table 5. The top AIMs selected by three measures from CEU and MEX after H-W and LD selection. The top thirty SNPs were reduced to 26, 24, 22 AIMs by δ, F_{ST}, and In, respectively. [1-6] indicated the SNPs that were in LD, and the highlighted ones (Bold and Italic) were removed. The physical distances of SNPs were downloaded from GRCh37.p13 (hg 19).

δ			F_{ST}			In		
rsID	chr	pos	rsID	chr	pos	rsID	chr	pos
rs11725412	chr4	38277754	rs7134749	chr12	50237637	rs10510511[1]	chr3	21260370
rs10962599	chr9	16795286	rs10510511	chr3	21260370	rs2700372	chr3	123633220
rs7134749	chr12	50237637	rs11725412	chr4	38277754	***rs17008458[1]***	***chr3***	***21380193***
rs11139346	chr9	84241442	rs2700372	chr3	123633220	rs7134749	chr12	50237637
rs10510511[1]	chr3	21260370	rs11139346	chr9	84241442	rs7404672[2]	chr16	10966479
rs1605524[1]	***chr3***	***21376648***	rs4729945[1]	chr7	103677151	rs11725412	chr4	38277754
rs974627[2]	chr12	38919524	rs10962599	chr9	16795286	***rs8054781[2]***	***chr16***	***11384776***
rs10141733	chr14	101142651	rs974627[2]	chr12	38919524	rs4729955[3]	chr7	103677151
rs2700372	chr3	123633220	rs12102256	chr15	37290293	rs715846[4]	chr9	95273013
rs729064	chr4	38689923	rs729064	chr4	38689923	***rs12238865[4]***	***chr9***	***95382606***
rs12504695	chr4	73200125	rs4823209	chr22	44661444	rs6836368[5]	chr4	130751286
rs10038199	chr5	5183642	rs7664927	chr4	114058373	***rs1024817[3]***	***chr7***	***103846110***
rs7040388	chr9	544956	rs715846[3]	chr9	95273013	***rs10012483[5]***	***chr4***	***130864103***
rs12102256	chr15	37290293	***rs12238865[3]***	***chr9***	***95382606***	***rs2295940[4]***	***chr9***	***95036903***
rs4823209	chr22	44661444	rs10141733	chr14	101142651	rs9307388	chr4	114075688
rs4833757	chr4	122633535	rs2487161	chr7	151033628	rs4905988	chr14	101124111
rs11133957	chr5	2657283	***rs10875950[2]***	***chr12***	***39043245***	rs945177	chr13	27621985
rs196698	chr6	80149183	rs949474	chr12	33406974	rs11139346	chr9	84241442
rs6904219	chr6	170250122	rs6681719	chr1	112538743	rs974627[6]	chr12	38919524
rs2487161	chr7	151033628	rs6559543	chr9	82968379	rs12102256	chr15	37290293
rs3019657	chr11	134511647	rs1992062	chr2	126078734	rs1866495	chr4	61519050
rs10875950[2]	***chr12***	***39043245***	rs1866495	chr4	61519050	***rs7614866[1]***	***chr3***	***21141977***

rs949474[2]	***chr12***	***33406974***	rs6836368[4]	chr4	130751286	***rs12824905[6]***	***chr12***	***34523160***
rs7960007	chr12	68885143	***rs1024817[1]***	***chr7***	***103846110***	rs2833550	chr21	33270555
rs6582668[2]	***chr12***	***38766604***	***rs12824905[2]***	***chr12***	***34523160***	rs11160530	chr14	100064674
rs7308783	chr12	31085405	rs2833550	chr21	33270555	rs10962599	chr9	16795286
rs6681719	chr1	112538743	rs945177	chr13	27621985	rs729064	chr4	38689923
rs7664927	chr4	114058373	***rs10012483[4]***	***chr4***	***130864103***	rs4823209	chr22	44661444
rs9498368	chr6	149835078	***rs2295940[3]***	***chr9***	***95036903***	rs1992062	chr2	126078734
rs12548626	chr8	17710100	rs7040388	chr9	544956	rs4860488	chr4	63759006

Supplemental Table 6. The top AIMs selected by three measures from CHD and MEX after H-W and LD selection. The top thirty SNPs were reduced to 22, 24, 24 AIMs by δ, F_{ST}, and In, respectively. [1-5] indicated the SNPs that were in LD, and the highlighted ones (Bold and Italic) were removed. The physical distances of SNPs were downloaded from GRCh37.p13 (hg 19).

δ			F_{ST}			In		
rsID	chr	pos	rsID	chr	pos	rsID	chr	pos
rs4429562	chr22	42892596	rs4429562	chr22	42892596	rs4429562	chr22	42892596
rs6500380[1]	chr16	48375777	rs6500380[1]	chr16	48375777	rs6500380[1]	chr16	48375777
rs8032157[2]	chr15	64480888	rs8032157[2]	chr15	64480888	rs8032157[2]	chr15	64480888
rs469471	chr21	14838552	rs469471	chr21	14838552	rs469471	chr21	14838552
rs17822931[1]	***chr16***	***48258198***	***rs17822931[1]***	***chr16***	***48258198***	***rs17822931[1]***	***chr16***	***48258198***
rs1761031	chr14	46926398	rs1761031	chr14	46926398	rs4299060	chr14	46926398
rs1348587[3]	chr2	154731793	rs1348587[3]	chr2	154731793	rs1761031	chr14	46926398
rs9322523	chr6	155741637	rs9322523	chr6	155741637	rs1348587[3]	chr2	154731793
rs2183614	chr14	49829709	rs2183614	chr14	49829709	rs9322523	chr6	155741637
rs453592	chr21	43999983	rs453592	chr21	43999983	rs1005402[4]	chr22	41291730
rs3286	chr7	93693901	rs2416186	chr14	49274102	rs2175591	chr1	234635790
rs2416186	chr14	49274102	rs3286	chr7	93693901	rs453592	chr21	43999983
rs1005402[4]	chr22	41291730	rs1005402[4]	chr22	41291730	rs3286	chr7	93693901
rs1606871	chr3	120090445	***rs133074[4]***	***chr22***	***41078473***	***rs133074[4]***	***chr22***	***41078473***
rs3843699[2]	***chr15***	***64637091***	rs1418043	chr20	1989659	rs1418043	chr20	1989659
rs16948162[2]	***chr15***	***64950215***	***rs3843699[2]***	***chr15***	***64637091***	rs9573248	chr13	74133837
rs8104441	chr19	51441807	***rs16948162[2]***	***chr15***	***64950215***	***rs3843699[2]***	***chr15***	***64637091***
rs133074[4]	***chr22***	***41078473***	rs8104441	chr19	51441807	rs8104441	chr19	51441807
rs10490534[3]	***chr2***	***154937930***	rs2175591	chr1	234635790	rs10516441	chr4	100307167
rs1042026[5]	chr4	100228466	rs1606871	chr3	120090445	rs11252631	chr10	4672037
rs12593770[2]	***chr15***	***64775002***	rs10516441	chr4	100307167	rs609096	chr5	4938130
rs6132153	chr20	1958145	***rs10490534[3]***	***chr2***	***154937930***	rs1606871	chr3	120090445

rs1883345	chr1	4882656	***rs12593770[2]***	***chr15***	***64775002***	***rs16948162[2]***	***chr15***	***64950215***
rs7580732[3]	***chr2***	***154847177***	rs1883345	chr1	4882656	rs4439672	chr14	49288860
rs10271340	chr7	14956127	rs10271340	chr7	14956127	rs1883345	chr1	4882656
rs11064160	chr12	6502708	rs11064160	chr12	6502708	***rs10490534[3]***	***chr2***	***154937930***
rs10483393	chr14	32460484	rs10483393	chr14	32460484	***rs12593770[2]***	***chr15***	***64775002***
rs2175591	chr1	234635790	rs4148888	chr4	100054192	rs2421069	chr4	134975713
rs4148888[5]	***chr4***	***100054192***	rs2421069	chr4	134975713	rs10271340	chr7	14956127
rs2421069	chr4	134975713	rs11252631	chr10	4672037	rs11064160	chr12	6502708

Supplemental Table 7. The minimum number of markers to distinguish any two populations identified by three measures (δ, F_{ST} and In). The physical distances of SNPs were downloaded from GRCh37.p13 (hg 19).

δ				F_{ST}				In			
SNPs	Chr	Physical position	Populations	SNPs	Chr	Physical position	Populations	SNPs	Chr	Physical position	Populations
rs1834640	chr15	48392165	ASW_CEU	rs1834640	chr15	48392165	ASW_CEU	rs1834640	chr15	48392165	ASW_CEU
rs1288097	chr15	45141373	ASW_CEU	rs1288097	chr15	45141373	ASW_CEU	rs1288097	chr15	45141373	ASW_CEU
rs6674304	chr1	116887742	ASW_CEU	rs6674304	chr1	116887742	ASW_CEU	rs6674304	chr1	116887742	ASW_CEU
rs7165971	chr15	55921013	ASW_CHD	rs7165971	chr15	55921013	ASW_CHD	rs7165971	chr15	55921013	ASW_CHD
rs745767	chr2	177825415	ASW_CHD	rs745767	chr2	177825415	ASW_CHD	rs745767	chr2	177825415	ASW_CHD
rs13021399	chr2	109006665	ASW_CHD	rs13021399	chr2	109006665	ASW_CHD	rs13021399	chr2	109006665	ASW_CHD
rs12087334	chr1	116887455	ASW_MEX	rs12087334	chr1	116887455	ASW_MEX	rs12087334	chr1	116887455	ASW_MEX
rs12149261	chr16	70998145	ASW_MEX	rs12149261	chr16	70998145	ASW_MEX	rs11845995	chr14	105930923	ASW_MEX
rs1827950	chr4	117098482	ASW_MEX	rs11845995	chr14	105930923	ASW_MEX	rs12149261	chr16	70998145	ASW_MEX
rs11845995	chr14	105930923	ASW_MEX	rs1827950	chr4	117098482	ASW_MEX	rs1827950	chr4	117098482	ASW_MEX
rs4429562	chr22	42892596	CEU_CHD	rs4429562	chr22	42892596	CEU_CHD	rs4429562	chr22	42892596	CEU_CHD
rs1547843	chr10	91738263	CEU_CHD	rs11126303	chr2	26173503	CEU_CHD	rs1834640	chr15	48392165	CEU_CHD
rs11126303	chr2	26173503	CEU_CHD	rs1834640	chr15	48392165	CEU_CHD	rs10510511	chr3	21260370	CEU_MEX
rs11725412	chr4	38277754	CEU_MEX	rs7134749	chr12	50237637	CEU_MEX	rs2700372	chr3	123633220	CEU_MEX
rs10962599	chr9	16795286	CEU_MEX	rs10510511	chr3	21260370	CEU_MEX	rs7134749	chr12	50237637	CEU_MEX
rs7134749	chr12	50237637	CEU_MEX	rs11725412	chr4	38277754	CEU_MEX	rs7404672	chr16	10966479	CEU_MEX
rs11139346	chr9	84241442	CEU_MEX	rs2700372	chr3	123633220	CEU_MEX	rs11725412	chr4	38277754	CEU_MEX
rs10510511	chr3	21260370	CEU_MEX	rs11139346	chr9	84241442	CEU_MEX	rs4729955	chr7	103693822	CEU_MEX
rs974627	chr12	38919524	CEU_MEX	rs4729945	chr7	103677151	CEU_MEX	rs715846	chr9	95273013	CEU_MEX
rs10141733	chr14	101142651	CEU_MEX	rs10962599	chr9	16795286	CEU_MEX	rs6836368	chr4	130751286	CEU_MEX
rs2700372	chr3	123633220	CEU_MEX	rs974627	chr12	38919524	CEU_MEX	rs9307388	chr4	114075688	CEU_MEX
rs4429562	chr22	42892596	CHD_MEX	rs4429562	chr22	42892596	CHD_MEX	rs4429562	chr22	42892596	CHD_MEX
rs6500380	chr16	48375777	CHD_MEX	rs6500380	chr16	48375777	CHD_MEX	rs6500380	chr16	48375777	CHD_MEX
rs8032157	chr15	64480888	CHD_MEX	rs8032157	chr15	64480888	CHD_MEX	rs8032157	chr15	64480888	CHD_MEX
rs469471	chr21	14838552	CHD_MEX	rs469471	chr21	14838552	CHD_MEX	rs469471	chr21	14838552	CHD_MEX

Supplemental Table 8. The correlation coefficients of PC1 and PC2 values among δ, F_{ST} and In panels.

	Delta_PC1	*F_{ST}_PC1*	*In_PC1*	*Delta_PC2*	*F_{ST}_PC2*	*In_PC2*
Delta_PC1	1.000	-	-	-	-	-
F_{ST}_PC1	0.997	1.000	-	-	-	-
In_PC1	0.975	0.979	1.000	-	-	-
Delta_PC2	-	-	-	1.000	-	-
F_{ST}_PC2	-	-	-	0.996	1.000	-
In_PC2	-	-	-	0.970	0.973	1.000

Supplemental Table 9. Ancestry prediction of HapMap individuals that fell outside the 95% confidence interval of four major U.S. populations.

	Population	Predicted Group	Probabilities of Membership in ASW	Probabilities of Membership in CEU	Probabilities of Membership in CHD	Probabilities of Membership in MEX
NA18501 (M)	YRI	ASW	1.00000	0.00000	0.00000	0.00000
NA19116 (F)	YRI	ASW	1.00000	0.00000	0.00000	0.00000
NA19128 (M)	YRI	ASW	1.00000	0.00000	0.00000	0.00000
NA19159 (F)	YRI	ASW	1.00000	0.00000	0.00000	0.00000
NA20504 (F)	TSI	CEU	0.00000	0.99542	0.00000	0.00458
NA20516 (M)	TSI	CEU	0.00000	0.99987	0.00000	0.00013
NA20520 (M)	TSI	CEU	0.00000	0.99993	0.00000	0.00007
NA20524 (M)	TSI	CEU	0.00000	0.99999	0.00000	0.00001
NA20542 (F)	TSI	CEU	0.00000	0.99999	0.00000	0.00001
NA20544 (M)	TSI	CEU	0.00000	1.00000	0.00000	0.00000
NA20582 (F)	TSI	CEU	0.00000	1.00000	0.00000	0.00000
NA20589 (F)	TSI	CEU	0.00000	1.00000	0.00000	0.00000
NA20756 (F)	TSI	CEU	0.00000	1.00000	0.00000	0.00000
NA20757 (F)	TSI	CEU	0.00000	0.99999	0.00000	0.00001
NA20768 (F)	TSI	CEU	0.00000	0.99982	0.00000	0.00018
NA20769 (F)	TSI	CEU	0.00000	0.99994	0.00000	0.00006
NA20775 (F)	TSI	CEU	0.00000	1.00000	0.00000	0.00000
NA20783 (M)	TSI	CEU	0.00000	0.99987	0.00000	0.00013
NA20786 (F)	TSI	CEU	0.00000	0.99997	0.00000	0.00003
NA20792 (M)	TSI	CEU	0.00000	1.00000	0.00000	0.00000
NA20797 (F)	TSI	CEU	0.00000	0.99985	0.00000	0.00015
NA20802 (F)	TSI	CEU	0.00000	0.99995	0.00000	0.00005
NA20804 (F)	TSI	CEU	0.00000	0.99664	0.00000	0.00336
NA20805 (M)	TSI	CEU	0.00000	0.99999	0.00000	0.00001
NA20811 (M)	TSI	CEU	0.00000	1.00000	0.00000	0.00000

NA20812 (M)	TSI	CEU	0.00009	0.99845	0.00000	0.00146
NA20815 (M)	TSI	CEU	0.00000	0.99935	0.00000	0.00065
NA20818 (F)	TSI	CEU	0.00000	1.00000	0.00000	0.00000
NA20819 (F)	TSI	CEU	0.00000	0.99975	0.00000	0.00025
NA18524 (M)	CHB	CHD	0.00000	0.00000	1.00000	0.00000
NA18572 (M)	CHB	CHD	0.00000	0.00000	1.00000	0.00000
NA18579 (F)	CHB	CHD	0.00000	0.00000	1.00000	0.00000
NA18599 (F)	CHB	CHD	0.00000	0.00000	1.00000	0.00000
NA18612 (M)	CHB	CHD	0.00000	0.00000	1.00000	0.00000
NA18944 (M)	JPT	CHD	0.00000	0.00000	1.00000	0.00000
NA18947 (F)	JPT	CHD	0.00000	0.00000	1.00000	0.00000
NA18968 (F)	JPT	CHD	0.00000	0.00000	1.00000	0.00000
NA18974 (M)	JPT	CHD	0.00000	0.00000	1.00000	0.00000
NA18994 (M)	JPT	CHD	0.00000	0.00000	1.00000	0.00000
NA18999 (F)	JPT	CHD	0.00000	0.00000	1.00000	0.00000
NA19005 (M)	JPT	CHD	0.00000	0.00000	1.00000	0.00000
NA19007 (M)	JPT	CHD	0.00000	0.00000	1.00000	0.00000

Supplemental Table 10. Ancestry prediction of 1000 Genomes individuals that fell outside the 95% confidence interval of four major U.S. populations. All CLM individuals were listed, because CLM contains individuals from three populations.

	Population	Predicted Group	Probabilities of Membership in ASW	Probabilities of Membership in CEU	Probabilities of Membership in CHD	Probabilities of Membership in MEX
NA18498 (M)	YRI_1000	ASW	1.00000	0.00000	0.00000	0.00000
NA19159 (F)	YRI_1000	ASW	1.00000	0.00000	0.00000	0.00000
HG00112 (M)	GBR_1000	CEU	0.00000	0.99223	0.00000	0.00777
HG00122 (F)	GBR_1000	CEU	0.00000	0.97939	0.00000	0.02061
HG00133 (F)	GBR_1000	CEU	0.00000	0.99867	0.00000	0.00133
HG00145 (M)	GBR_1000	CEU	0.00000	0.99974	0.00000	0.00026
HG00155 (M)	GBR_1000	CEU	0.00000	0.81940	0.00000	0.18060
HG00237 (F)	GBR_1000	CEU	0.00000	0.77781	0.00000	0.22219
HG00257 (M)	GBR_1000	CEU	0.00000	0.97887	0.00000	0.02113
HG00261 (F)	GBR_1000	CEU	0.00000	0.99885	0.00000	0.00115
NA18553 (F)	CHB_1000	CHD	0.00000	0.00000	1.00000	0.00000
NA18572 (M)	CHB_1000	CHD	0.00000	0.00000	1.00000	0.00000
NA18574 (F)	CHB_1000	CHD	0.00000	0.00000	1.00000	0.00000
NA18599 (F)	CHB_1000	CHD	0.00000	0.00000	1.00000	0.00000
NA18612 (M)	CHB_1000	CHD	0.00000	0.00000	1.00000	0.00000
NA18625 (F)	CHB_1000	CHD	0.00000	0.00000	1.00000	0.00000
NA18630 (F)	CHB_1000	CHD	0.00000	0.00000	1.00000	0.00000
NA18632 (M)	CHB_1000	CHD	0.00000	0.00000	1.00000	0.00000
HG01112 (M)	CLM_1000	CEU	0.00000	0.82630	0.00000	0.17370
HG01113 (F)	CLM_1000	CEU	0.00000	0.62672	0.00000	0.37328
HG01119 (F)	CLM_1000	MEX	0.00000	0.02537	0.00000	0.97463
HG01121 (M)	CLM_1000	MEX	0.00000	0.00000	0.00000	1.00000
HG01122 (F)	CLM_1000	MEX	0.00000	0.00174	0.00000	0.99826
HG01124 (M)	CLM_1000	CEU	0.00000	0.95778	0.00000	0.04222
HG01125 (F)	CLM_1000	CEU	0.00000	0.98512	0.00000	0.01488

HG01130 (M)	CLM_1000	CEU	0.00000	0.99859	0.00000	0.00141
HG01131 (F)	CLM_1000	MEX	0.00000	0.08056	0.00000	0.91944
HG01133 (M)	CLM_1000	MEX	0.00000	0.15416	0.00000	0.84584
HG01134 (F)	CLM_1000	MEX	0.00003	0.00000	0.00000	0.99997
HG01136 (M)	CLM_1000	MEX	0.00000	0.00000	0.00000	1.00000
HG01137 (F)	CLM_1000	MEX	0.00000	0.00000	0.00000	1.00000
HG01139 (M)	CLM_1000	MEX	0.00001	0.00065	0.00000	0.99934
HG01140 (F)	CLM_1000	MEX	0.00000	0.00000	0.00000	1.00000
HG01142 (M)	CLM_1000	MEX	0.00000	0.00000	0.00000	1.00000
HG01148 (M)	CLM_1000	CEU	0.00000	0.99183	0.00000	0.00817
HG01149 (F)	CLM_1000	MEX	0.00000	0.00034	0.00000	0.99966
HG01250 (M)	CLM_1000	MEX	0.00049	0.00000	0.00000	0.99951
HG01251 (F)	CLM_1000	CEU	0.00000	0.95184	0.00000	0.04816
HG01253 (M)	CLM_1000	MEX	0.00000	0.00030	0.00000	0.99970
HG01254 (F)	CLM_1000	MEX	0.20915	0.00006	0.00000	0.79079
HG01256 (M)	CLM_1000	CEU	0.00000	0.99818	0.00000	0.00182
HG01257 (F)	CLM_1000	MEX	0.00125	0.00042	0.00000	0.99833
HG01259 (M)	CLM_1000	MEX	0.00017	0.00241	0.00000	0.99742
HG01260 (F)	CLM_1000	MEX	0.00000	0.01830	0.00000	0.98170
HG01269 (F)	CLM_1000	CEU	0.00029	0.65440	0.00000	0.34531
HG01271 (M)	CLM_1000	CEU	0.00000	0.95504	0.00000	0.04496
HG01272 (F)	CLM_1000	MEX	0.00175	0.00000	0.00000	0.99825
HG01275 (F)	CLM_1000	CEU	0.00000	0.99954	0.00000	0.00046
HG01277 (M)	CLM_1000	MEX	0.00003	0.00000	0.00000	0.99997
HG01280 (M)	CLM_1000	CEU	0.00002	0.88606	0.00000	0.11393
HG01281 (F)	CLM_1000	CEU	0.00000	0.98926	0.00000	0.01074
HG01284 (F)	CLM_1000	MEX	0.00000	0.00265	0.00000	0.99735
HG01341 (M)	CLM_1000	MEX	0.00000	0.00000	0.00000	1.00000
HG01342 (F)	CLM_1000	ASW	1.00000	0.00000	0.00000	0.00000
HG01344 (M)	CLM_1000	MEX	0.00000	0.00000	0.00000	1.00000
HG01345 (F)	CLM_1000	CEU	0.00000	0.79952	0.00000	0.20048

HG01348 (F)	CLM_1000	MEX	0.00000	0.00000	0.00000	1.00000
HG01350 (M)	CLM_1000	MEX	0.00000	0.00025	0.00000	0.99975
HG01351 (F)	CLM_1000	MEX	0.00000	0.35426	0.00000	0.64574
HG01353 (M)	CLM_1000	MEX	0.00000	0.00000	0.00000	1.00000
HG01354 (F)	CLM_1000	MEX	0.00000	0.00151	0.00000	0.99849
HG01356 (M)	CLM_1000	MEX	0.00000	0.01394	0.00000	0.98606
HG01357 (F)	CLM_1000	CEU	0.00000	0.88978	0.00000	0.11022
HG01359 (M)	CLM_1000	MEX	0.00000	0.00000	0.00000	1.00000
HG01360 (F)	CLM_1000	MEX	0.00000	0.00000	0.00000	1.00000
HG01362 (M)	CLM_1000	CEU	0.00000	0.98860	0.00000	0.01140
HG01363 (F)	CLM_1000	MEX	0.03449	0.00000	0.00000	0.96551
HG01365 (M)	CLM_1000	MEX	0.02026	0.00000	0.00000	0.97974
HG01366 (F)	CLM_1000	MEX	0.00000	0.00017	0.00000	0.99983
HG01369 (F)	CLM_1000	MEX	0.00000	0.00000	0.00000	1.00000
HG01372 (F)	CLM_1000	MEX	0.00000	0.00000	0.00000	1.00000
HG01374 (M)	CLM_1000	CEU	0.00000	0.78519	0.00000	0.21481
HG01375 (F)	CLM_1000	MEX	0.00000	0.00000	0.00000	1.00000
HG01377 (M)	CLM_1000	MEX	0.00000	0.00000	0.00000	1.00000
HG01378 (F)	CLM_1000	CEU	0.00000	0.99413	0.00000	0.00587
HG01383 (M)	CLM_1000	MEX	0.00000	0.00002	0.00000	0.99998
HG01384 (F)	CLM_1000	MEX	0.00000	0.12125	0.00000	0.87875
HG01389 (M)	CLM_1000	MEX	0.00000	0.03630	0.00000	0.96370
HG01390 (F)	CLM_1000	ASW	0.97272	0.00000	0.00000	0.02728
HG01431 (M)	CLM_1000	MEX	0.00035	0.00054	0.00000	0.99911
HG01432 (F)	CLM_1000	MEX	0.00001	0.29551	0.00000	0.70448
HG01435 (F)	CLM_1000	MEX	0.00000	0.00000	0.00000	1.00000
HG01437 (M)	CLM_1000	CEU	0.00000	0.99954	0.00000	0.00046
HG01438 (F)	CLM_1000	MEX	0.00001	0.33493	0.00000	0.66507
HG01440 (M)	CLM_1000	MEX	0.00000	0.05216	0.00000	0.94784
HG01441 (F)	CLM_1000	MEX	0.00000	0.00000	0.00000	1.00000
HG01443 (M)	CLM_1000	MEX	0.00000	0.00000	0.00000	1.00000

HG01444 (F)	CLM_1000	MEX	0.00002	0.00000	0.00000	0.99998
HG01447 (F)	CLM_1000	MEX	0.00000	0.20085	0.00000	0.79915
HG01455 (M)	CLM_1000	MEX	0.00059	0.00000	0.00000	0.99941
HG01456 (F)	CLM_1000	MEX	0.00000	0.00000	0.00000	1.00000
HG01459 (F)	CLM_1000	MEX	0.00007	0.00000	0.00000	0.99993
HG01461 (M)	CLM_1000	ASW	0.88965	0.00000	0.00000	0.11035
HG01462 (F)	CLM_1000	ASW	1.00000	0.00000	0.00000	0.00000
HG01464 (M)	CLM_1000	MEX	0.00000	0.00000	0.00000	1.00000
HG01465 (F)	CLM_1000	MEX	0.00000	0.05379	0.00000	0.94621
HG01468 (F)	CLM_1000	MEX	0.00000	0.00538	0.00000	0.99462
HG01474 (F)	CLM_1000	CEU	0.00000	0.99014	0.00000	0.00986
HG01479 (M)	CLM_1000	CEU	0.00000	0.98759	0.00000	0.01241
HG01485 (M)	CLM_1000	MEX	0.05547	0.00000	0.00000	0.94453
HG01486 (F)	CLM_1000	MEX	0.00000	0.00000	0.00000	1.00000
HG01488 (M)	CLM_1000	CEU	0.00000	0.53482	0.00000	0.46518
HG01489 (F)	CLM_1000	CEU	0.00000	0.99887	0.00000	0.00113
HG01491 (M)	CLM_1000	CEU	0.00000	0.99875	0.00000	0.00125
HG01492 (F)	CLM_1000	CEU	0.00000	0.98474	0.00000	0.01526
HG01494 (M)	CLM_1000	MEX	0.00000	0.00002	0.00000	0.99998
HG01495 (F)	CLM_1000	MEX	0.00000	0.00005	0.00000	0.99995
HG01497 (M)	CLM_1000	MEX	0.00000	0.00000	0.00000	1.00000
HG01498 (F)	CLM_1000	MEX	0.00000	0.00000	0.00000	1.00000
HG01550 (M)	CLM_1000	CEU	0.00000	0.65357	0.00000	0.34643
HG01551 (F)	CLM_1000	MEX	0.02552	0.00000	0.00000	0.97448
HG01556 (M)	CLM_1000	MEX	0.00151	0.00000	0.00000	0.99849
NA19648 (F)	MEX_1000	MEX	0.00000	0.23651	0.00000	0.76349
NA19728 (F)	MEX_1000	MEX	0.00000	0.00000	0.00000	1.00000
NA19731 (F)	MEX_1000	MEX	0.00000	0.00000	0.00000	1.00000
NA19732 (M)	MEX_1000	MEX	0.00000	0.00000	0.00000	1.00000
NA19741 (M)	MEX_1000	MEX	0.00000	0.00000	0.00010	0.99990

Supplemental Table 11. Summary of SNPs contained in ten AIMs panels. The physical distances of SNPs were downloaded from GRCh37.p13 (hg 19).

Our AIMs	Kosoy's AIMs	Halder's AIMs	Nievergelt's AIMs	Kidd's AIMs	Phillips's AIMs	Jia's AIMs	Gettings's AIMs	Rogalla's AIMs	Wei's AIMs
rs12087334	rs2986742	rs263531	rs359955	rs3737576	rs12142199	rs11264300	rs3737576	rs953035	rs595961
rs11126303	rs6541030	rs434504	rs1834619	rs7554936	rs434504	rs28777	rs2814778	rs595961	rs2710684
rs13021399	rs647325	rs770028	rs842639	rs2814778	rs595961	rs7700468	rs10496971	rs11891922	rs260690
rs745767	rs4908343	rs950848	rs4907251	rs798443	rs2139931	rs984654	rs1876482	rs1063	rs10496971
rs10510511	rs1325502	rs595961	rs260714	rs1876482	rs12402499	rs845561	rs952718	rs3827760	rs10497191
rs2700372	rs12130799	rs522287	rs4664511	rs1834619	rs2814778	rs4646437	rs6548616	rs16891982	rs820371
rs11725412	rs3118378	rs1469344	rs1863086	rs3827760	rs4657449	rs7047704	rs1344870	rs1490388	rs1586861
rs7689609	rs3737576	rs625994	rs10497191	rs260690	rs2184030	rs11188246	rs10007810	rs984654	rs28777
rs1827950	rs7554936	rs1325609	rs6737672	rs6754311	rs7531501	rs11018541	rs6451722	rs1408799	rs10079352
rs4729945	rs1040404	rs783064	rs3098610	rs10497191	rs12405776	rs343092	rs10108270	rs1076563	rs1366220
rs10962599	rs1407434	rs236336	rs9880567	rs12498138	rs1834619	rs1480464	rs4918842	rs3782972	rs6875659
rs11139346	rs4951629	rs841338	rs9809818	rs4833103	rs1567803	rs7397918	rs714857	rs1448485	rs7752055
rs974627	rs316873	rs1446966	rs12498138	rs1229984	rs3827760	rs2407522	rs2065982	rs12913832	rs10258063
rs7134749	rs798443	rs1415680	rs11725412	rs3811801	rs1371048	rs1886048	rs730570	rs1426654	rs366178
rs1761031	rs7421394	rs6003	rs4833103	rs7657799	rs16830500	rs8012948	rs3784230	rs1395579	rs4749305
rs11845995	rs4666200	rs2204307	rs10079352	rs16891982	rs10186877	rs2594899	rs722869	rs885479	rs4244304
rs1288097	rs4670767	rs2814778	rs4705360	rs7722456	rs10183022	rs2703957	rs2714758		rs10741584
rs1834640	rs13400937	rs296528	rs7722456	rs870347	rs2302013	rs728404	rs735612		rs3825663
rs7165971	rs260690	rs2065160	rs3823159	rs3823159	rs7623065	rs1448485	rs4891825		rs728404
rs8032157	rs10496971	rs1406869	rs2717329	rs192655	rs862500	rs916977			rs1448485
rs6500380	rs2627037	rs715956	rs310362	rs917115	rs9809818	rs7163702			rs7170869
rs12149261	rs1569175	rs1039630	rs7837234	rs1462906	rs7630522	rs8036234			rs1453858
rs4429562	rs10510228	rs1937025	rs6990312	rs6990312	rs6437783	rs6495913			rs2470102
	rs4955316	rs772436	rs2196051	rs2196051	rs12498138	rs1565403			rs4787040
	rs9809104	rs1869380	rs4741658	rs1871534	rs820371	rs1426654			rs881929
	rs6548616	rs997676	rs10961356	rs3814134	rs868767	rs2470102			rs1197062
	rs12629908	rs892457	rs4918664	rs4918664	rs10012227	rs12907018			rs4789193

rs9845457	rs1431332	rs734241	rs174570	rs4540055	rs11070629
rs734873	rs3287	rs10877030	rs1079597	rs1229984	rs8040562
rs2030763	rs727878	rs1572018	rs2238151	rs2851060	rs12911421
rs1513181	rs262482	rs2166624	rs671	rs1509524	rs885479
rs9291090	rs53915	rs12878166	rs7997709	rs16891982	rs12446019
rs10007810	rs1221172	rs735480	rs1572018	rs930072	rs2955250
rs1369093	rs959929	rs1834640	rs2166624	rs26951	rs1205357
rs385194	rs1447111	rs2593595	rs7326934	rs10079352	rs7290134
rs7657799	rs716373	rs4471745	rs9522149	rs1366220	
rs2702414	rs729253	rs7226659	rs200354	rs11960137	
rs316598	rs725395	rs7251928	rs1800414	rs6886019	
rs870347	rs1848728	rs310644	rs12913832	rs6875659	
rs37369	rs1533677	rs2024566	rs12439433	rs10455681	
rs6451722	rs1108718	rs1557553	rs735480	rs794672	
rs12657828	rs830599		rs1426654	rs2503770	
rs6556352	rs1399272		rs459920	rs2180052	
rs1500127	rs1105220		rs4411548	rs2080161	
rs6422347	rs361065		rs2593595	rs917115	
rs1040045	rs361055		rs17642714	rs2471552	
rs2504853	rs892263		rs4471745	rs798949	
rs7745461	rs715790		rs11652805	rs366178	
rs2397060	rs2045517		rs2042762	rs11778591	
rs192655	rs959858		rs7226659	rs433342	
rs4463276	rs1368872		rs3916235	rs7832008	
rs4458655	rs733563		rs4891825	rs1871534	
rs1871428	rs997164		rs7251928	rs10811102	
rs731257	rs1501680		rs310644	rs10970986	
rs32314	rs270565		rs2024566	rs4979274	
rs2330442	rs1125508			rs2789823	
rs4717865	rs3846662			rs2274636	
rs10954737	rs3317			rs4749305	

rs705308	rs3340	rs17287498
rs7803075	rs1005056	rs4935501
rs10236187	rs960709	rs7911953
rs6464211	rs925197	rs4918664
rs10108270	rs875543	rs17130385
rs3943253	rs1407361	rs7084970
rs1471939	rs933199	rs3751050
rs12544346	rs559035	rs5030240
rs7844723	rs1937147	rs7937598
rs2001907	rs1041656	rs174570
rs1408801	rs765338	rs2715883
rs10511828	rs877823	rs1486341
rs3793451	rs718268	rs2051827
rs2306040	rs409359	rs10735825
rs10513300	rs1476597	rs3759171
rs2073821	rs1155513	rs1592672
rs3793791	rs1018919	rs7307862
rs4746136	rs2161	rs2585897
rs4918842	rs2242480	rs17359176
rs4880436	rs869337	rs2065982
rs10839880	rs1528037	rs4391951
rs1837606	rs125097	rs1924381
rs2946788	rs984654	rs17544484
rs11227699	rs716840	rs3782973
rs948028	rs1454284	rs9522149
rs2416791	rs1039917	rs10483251
rs1513056	rs285	rs7151991
rs214678	rs351782	rs10149275
rs772262	rs553950	rs11625446
rs2070586	rs3176921	rs12435594
rs1503767	rs717836	rs722869

rs9319336	rs1467044	rs730570
rs7997709	rs913258	rs12913832
rs9530435	rs1888952	rs1426654
rs9522149	rs998599	rs12594144
rs1760921	rs1414241	rs11074130
rs2357442	rs721702	rs3784651
rs1950993	rs1041321	rs4787040
rs8021730	rs2695	rs4780476
rs946918	rs526454	rs9934011
rs200354	rs662117	rs881929
rs3784230	rs9032	rs17822931
rs12439433	rs590086	rs67302
rs2899826	rs1076160	rs1452501
rs8035124	rs675837	rs4791868
rs4984913	rs723220	rs8072587
rs4781011	rs721825	rs4792928
rs2269793	rs1983128	rs9908046
rs818386	rs877783	rs1197062
rs2966849	rs1034290	rs203150
rs1879488	rs719909	rs1369290
rs2033111	rs913375	rs634392
rs11652805	rs1546541	rs7246968
rs10512572	rs1385851	rs8104441
rs2125345	rs1470144	rs4806654
rs4798812	rs713503	rs499827
rs4800105	rs523200	rs6054465
rs7238445	rs594689	rs6034866
rs881728	rs725192	rs2889678
rs4891825	rs282496	rs6088466
rs874299	rs90192	rs2069945
rs8113143	rs236919	rs1877751

rs3745099	rs1800498	rs310644
rs2532060	rs697212	rs2833250
rs6104567	rs593226	rs2282107
rs3907047	rs1407961	rs3804030
rs2835370	rs883055	rs715605
rs1296819	rs1337038	rs5757362
rs4821004	rs320075	rs8137373
rs5768007	rs1113337	rs1557553
	rs987284	
	rs1998055	
	rs729531	
	rs953743	
	rs1111108	
	rs915056	
	rs763807	
	rs1007407	
	rs1157223	
	rs174518	
	rs1296149	
	rs741272	
	rs716873	
	rs974324	
	rs730570	
	rs1076808	
	rs1426217	
	rs1426208	
	rs10852218	
	rs2862	
	rs920915	
	rs972801	
	rs2010069	

rs1015081
rs1375229
rs1800410
rs251741
rs1395579
rs1395580
rs1650999
rs878671
rs173537
rs67302
rs212498
rs917502
rs758767
rs1030525
rs1437069
rs1432065
rs1125425
rs667508
rs1465708
rs430667
rs275837
rs1040577
rs172982
rs1004571
rs735050
rs2014519
rs81481
rs138335

CHAPTER 2

Empirical Testing of A 23-AIMs Panel of SNPs for Ancestry Evaluations in Four Major US Populations

International Journal of Legal Medicine
2016 Feb 25. [Epub ahead of print]

Xiangpei Zeng
David H. Warshauer
Jonathan L. King
Jennifer D. Churchill
Ranajit Chakraborty
Bruce Budowle

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 United States, 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™ 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)

2.1. 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 United States 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 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.

2.2. Methods and Materials:

2.2.1 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).

2.2.2 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.

2.2.3 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.

2.2.4 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 hours). 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 sec at 98° C; 12 cycles of 10 sec at 98° C, 30 sec at 60° C, 30 sec 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.

2.2.5 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.

2.2.6 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.

2.3. 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 22X (20X was set as an arbitrary detection threshold), and the highest coverage was 2216X. 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.

Table 1. The 23 AIMs selected to distinguish the four major U.S. populations. * This AIM on chromosome 16 was removed due to a duplication on chromosome 1.

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

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 (Supplemental Table 2). In addition, the genotype data of SNP rs12149261 showed that 136 of 140 individuals (Asian, Caucasian, 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.

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 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).

Table 2. Significant linkage disequilibrium (LD) results of 22 SNPs in four populations. LD p-values shown for the specified loci pair in which a significant value was observed in at least one population group. Values in bold were significant after Bonferroni correction ($p<0.000216$).

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**

The principal component analysis (PCA) plot showed that eight individuals were assigned to a different group than indicated on their samples (Figure 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} 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.

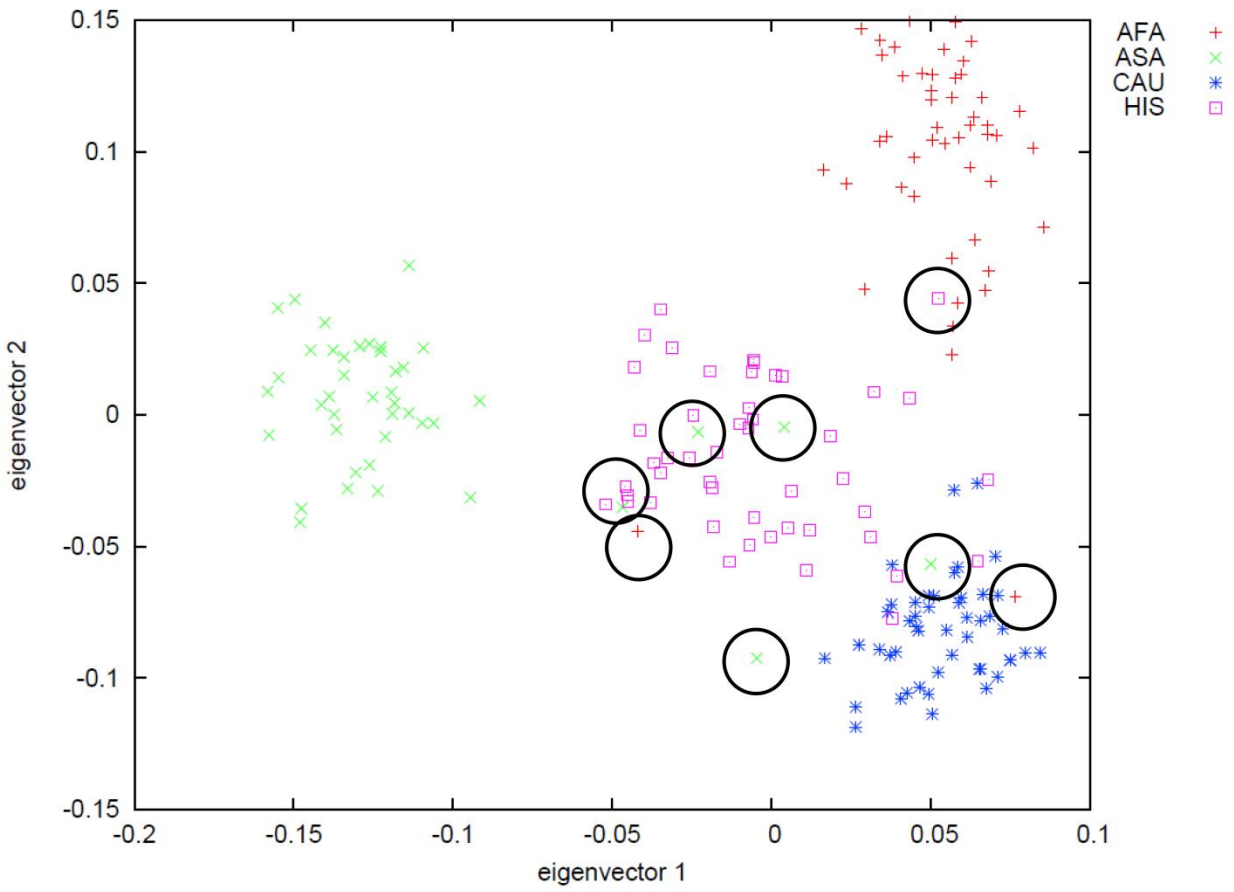


Figure 1. The PCA plot of 189 individuals using 22-SNP AIMs panel. Eight individuals (encircled by black circles) were assigned to different groups than what was labeled on their sample submissions.

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

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.

2.4. 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 population. 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://http://hydrodictyon.eeb.uconn.edu/people/plewis/software.php. Accessed 25 April 2007

2.5. Supplemental Materials

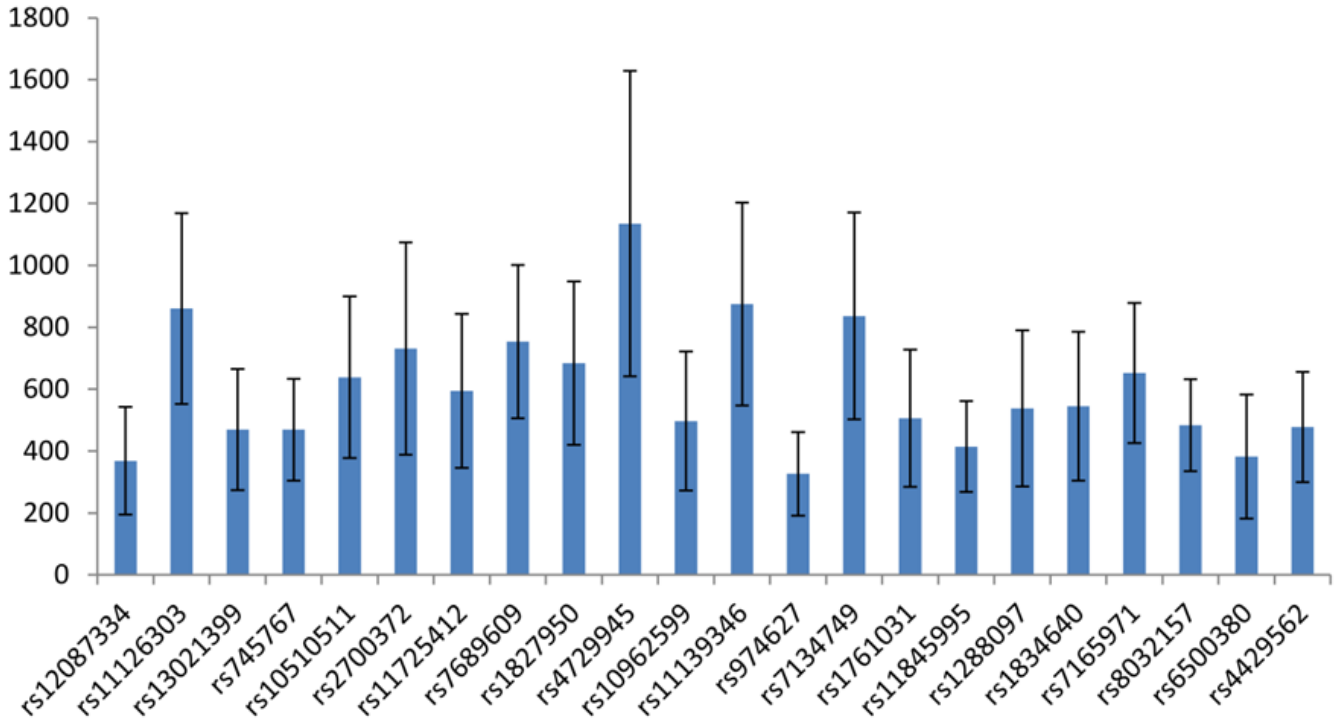
Supplemental Table 1. The probe information of the 23 ancestry informative SNPs.

SNP	Chromosome	SNP position	Probe start position	Probe end position	Probe length
rs12087334	1	116887455	116887364	116887444	80
rs12087334	1	116887455	116887464	116887544	80
rs12087334	1	116887455	116887651	116887731	80
rs12087334	1	116887455	116887751	116887831	80
rs11126303	2	26173503	26173412	26173492	80
rs11126303	2	26173503	26173512	26173592	80
rs13021399	2	109006665	109006564	109006644	80
rs13021399	2	109006665	109006674	109006754	80
rs745767	2	177825415	177825334	177825414	80
rs745767	2	177825415	177825424	177825504	80
rs10510511	3	21260370	21260279	21260359	80
rs10510511	3	21260370	21260379	21260459	80
rs2700372	3	123633220	123633139	123633219	80
rs2700372	3	123633220	123633229	123633309	80
rs11725412	4	38277754	38277663	38277743	80
rs11725412	4	38277754	38277763	38277843	80
rs7689609	4	72083374	72083293	72083373	80
rs7689609	4	72083374	72083383	72083463	80
rs1827950	4	117098482	117098391	117098471	80
rs1827950	4	117098482	117098491	117098571	80
rs4729945	7	103677151	103677060	103677140	80
rs4729945	7	103677151	103677160	103677240	80
rs10962599	9	16795286	16795195	16795275	80
rs10962599	9	16795286	16795295	16795375	80
rs11139346	9	84241442	84241351	84241431	80

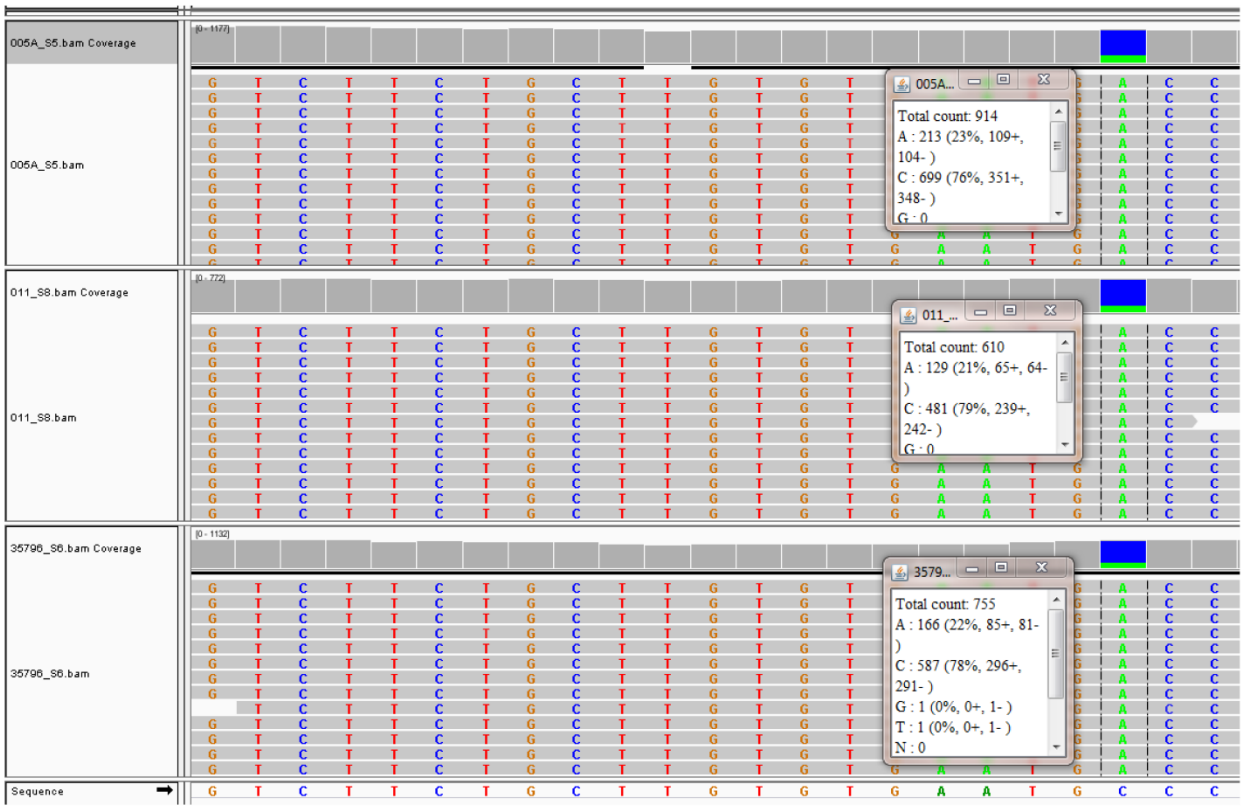
rs11139346	9	84241442	84241451	84241531	80
rs974627	12	38919524	38919443	38919523	80
rs974627	12	38919524	38919543	38919623	80
rs7134749	12	50237637	50237546	50237626	80
rs7134749	12	50237637	50237646	50237726	80
rs1761031	14	46926398	46926307	46926387	80
rs1761031	14	46926398	46926407	46926487	80
rs11845995	14	105930923	105930832	105930912	80
rs11845995	14	105930923	105930932	105931012	80
rs1288097	15	45141373	45141282	45141362	80
rs1288097	15	45141373	45141382	45141462	80
rs1834640	15	48392165	48392074	48392154	80
rs1834640	15	48392165	48392174	48392254	80
rs7165971	15	55921013	55920922	55921002	80
rs7165971	15	55921013	55921022	55921102	80
rs8032157	15	64480888	64480797	64480877	80
rs8032157	15	64480888	64480897	64480977	80
rs6500380	16	48375777	48375686	48375766	80
rs6500380	16	48375777	48375786	48375866	80
rs12149261	16	70998145	70998054	70998134	80
rs12149261	16	70998145	70998154	70998234	80
rs4429562	22	42892596	42892505	42892585	80
rs4429562	22	42892596	42892605	42892685	80

Supplemental Table 2. Significant LD results of 23 SNPs in four populations. LD p-values were given for the specified loci pair in which a significant value was observed in at least one population group. Values in bold were significant after Bonferroni correction (p<0.000197).

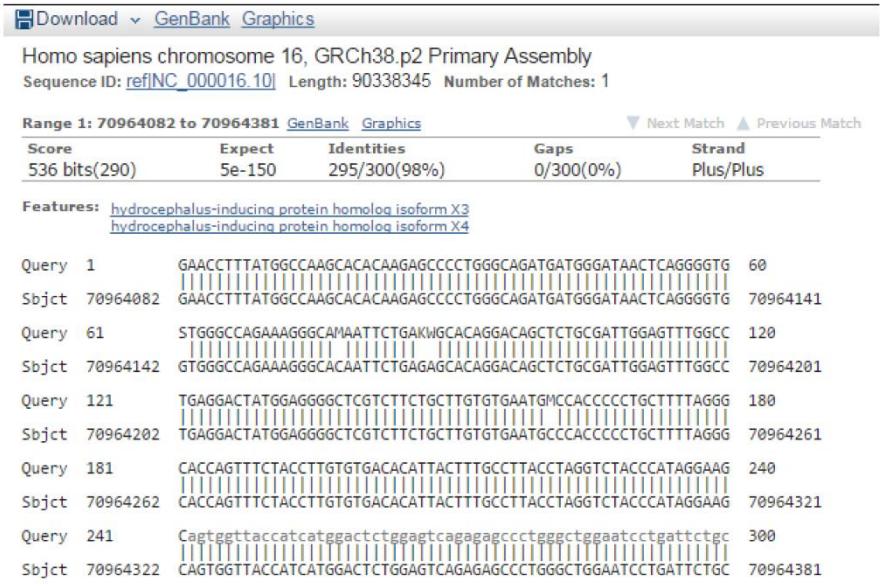
SNP pair	LD p-values in population			
	African American	Asian	Caucasian	Hispanic
rs12087334/rs4429562	**<0.000001**	0.3587	0.4655	0.1377
rs12087334/rs12149261	0.0278	**<0.000001**	**<0.000001**	**<0.000001**
rs745767/rs12149261	0.0413	**<0.000001**	**<0.000001**	**<0.000001**
rs13021399/rs12149261	0.3158	**<0.000001**	**<0.000001**	**<0.000001**
rs11126303/rs12149261	0.0184	**<0.000001**	**<0.000001**	**<0.000001**
rs10510511/rs12149261	0.0442	**<0.000001**	**<0.000001**	**<0.000001**
rs2700372/rs12149261	0.6168	**<0.000001**	**<0.000001**	**<0.000001**
rs7689609/rs12149261	0.0164	**<0.000001**	**<0.000001**	**<0.000001**
rs1827950/rs12149261	0.0898	**<0.000001**	**<0.000001**	**<0.000001**
rs11725412/rs12149261	0.3249	**<0.000001**	**<0.000001**	**<0.000001**
rs4729945/rs1288097	0.4187	0.8169	**<0.000001**	0.8697
rs4729945/rs12149261	0.6473	**<0.000001**	**<0.000001**	**<0.000001**
rs11139346/rs12149261	0.3386	**<0.000001**	**<0.000001**	**<0.000001**
rs974627/rs12149261	0.3050	**<0.000001**	**<0.000001**	**<0.000001**
rs7134749/rs12149261	0.1021	**<0.000001**	**<0.000001**	**<0.000001**
rs10962599/rs12149261	0.2028	**<0.000001**	**<0.000001**	**<0.000001**
rs11845995/rs12149261	0.1329	**<0.000001**	**<0.000001**	**<0.000001**
rs1761031/rs12149261	0.4264	**<0.000001**	**<0.000001**	**<0.000001**
rs1834640/rs12149261	0.0877	**<0.000001**	**<0.000001**	**<0.000001**
rs1288097/rs12149261	0.3020	**<0.000001**	**<0.000001**	**<0.000001**
rs7165971/rs12149261	0.1225	**<0.000001**	**<0.000001**	**<0.000001**
rs8032157/rs12149261	0.0252	**<0.000001**	**<0.000001**	**<0.000001**
rs12149261/rs6500380	0.1209	**<0.000001**	**<0.000001**	**<0.000001**
rs12149261/rs4429562	0.0007	**<0.000001**	**<0.000001**	**<0.000001**
rs12087334/rs1288097	0.2065	0.4057	**0.0001**	0.3547
rs2700372/rs7165971	0.0859	**0.0001**	0.3891	0.7046
rs1834640/rs7165971	0.0285	**0.0001**	0.1784	0.7867

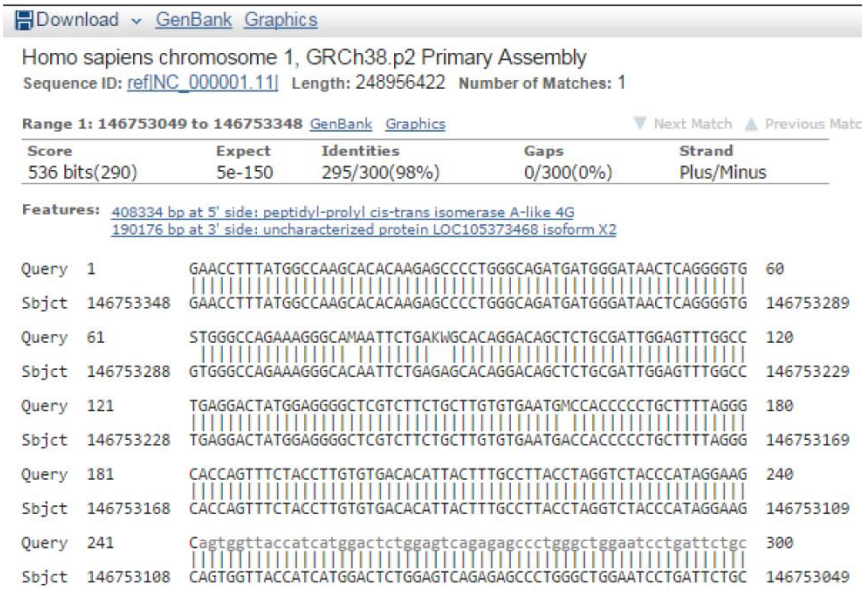


Supplemental Figure 1. Average coverage of 22 SNPs in 189 individuals.

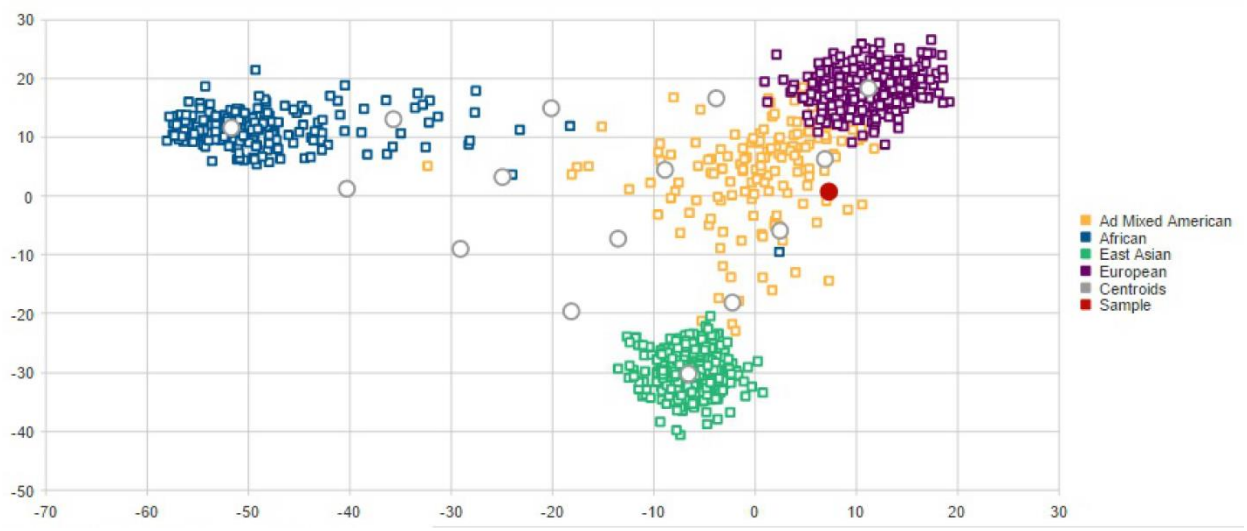


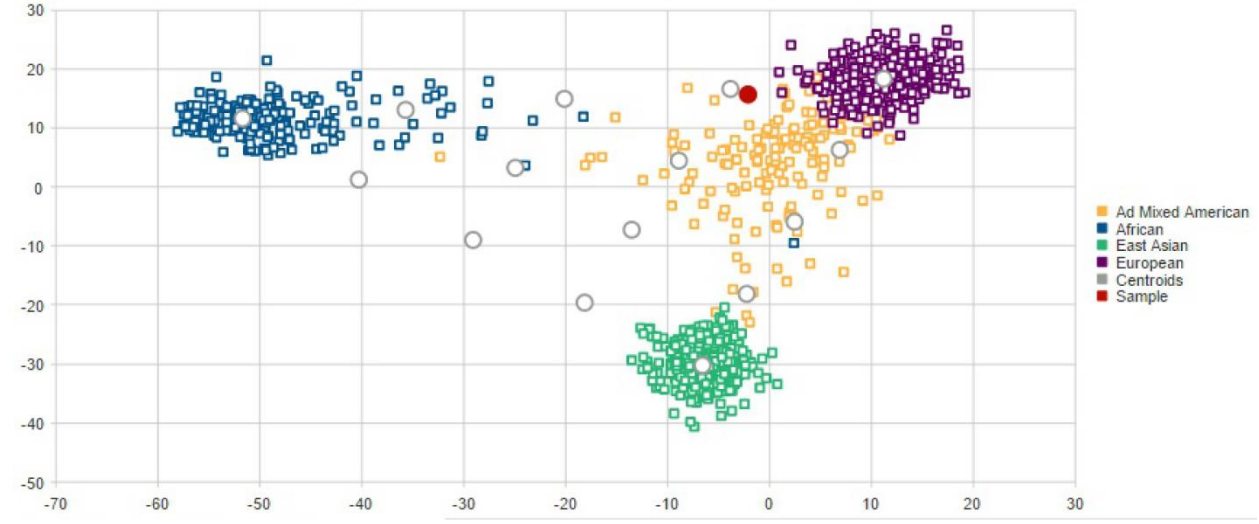
Supplemental Figure 2. Example of sequence data for SNP rs12149261 shown in IGV. The three different individuals (005A, 011 and 35706) were Caucasian, Asian and Hispanic American, respectively.

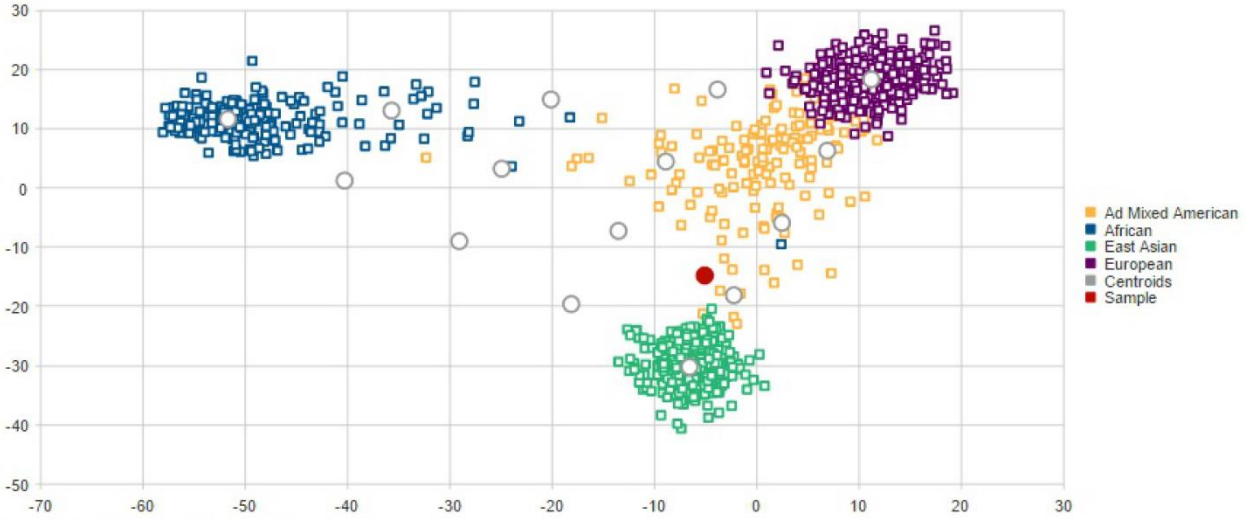


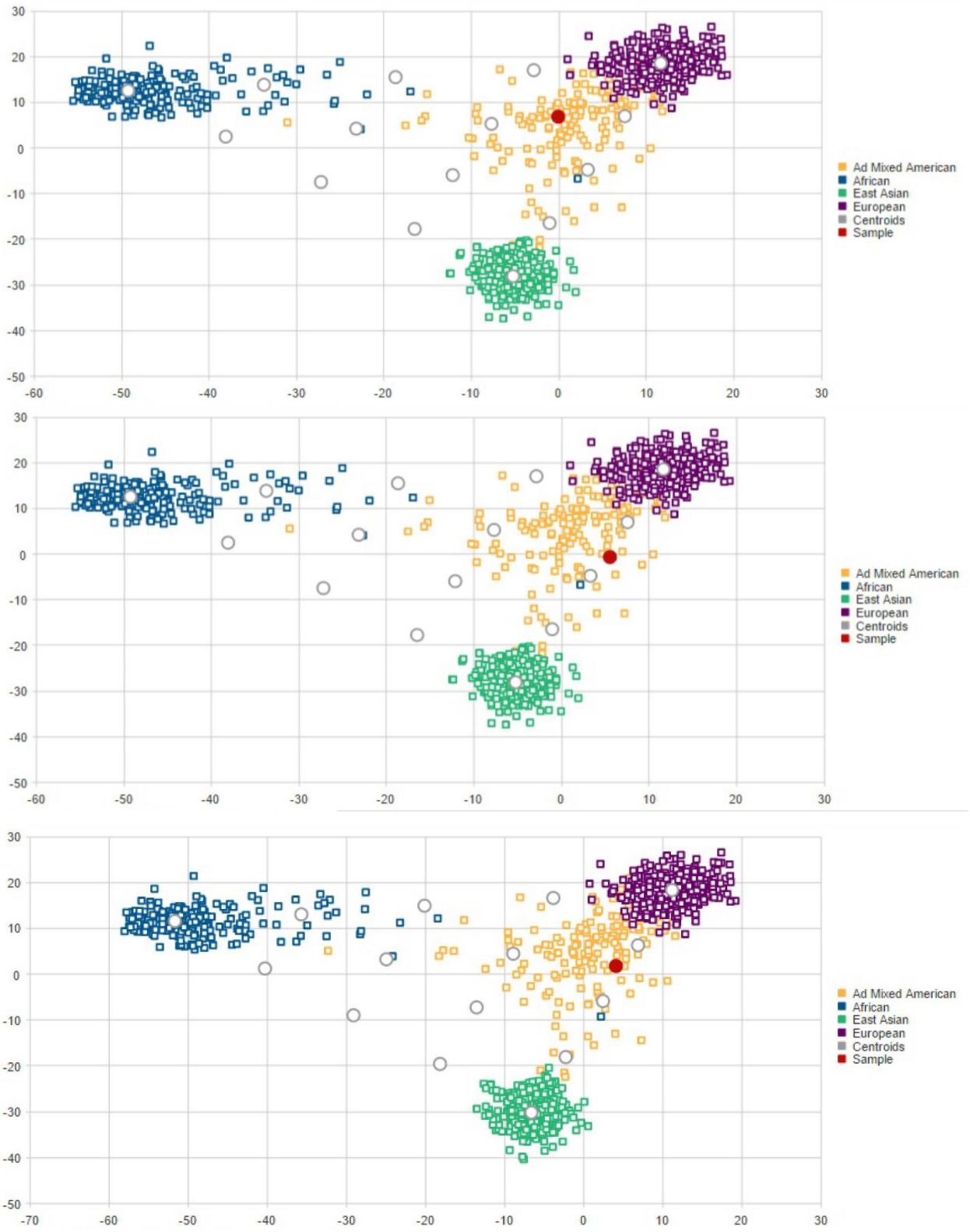


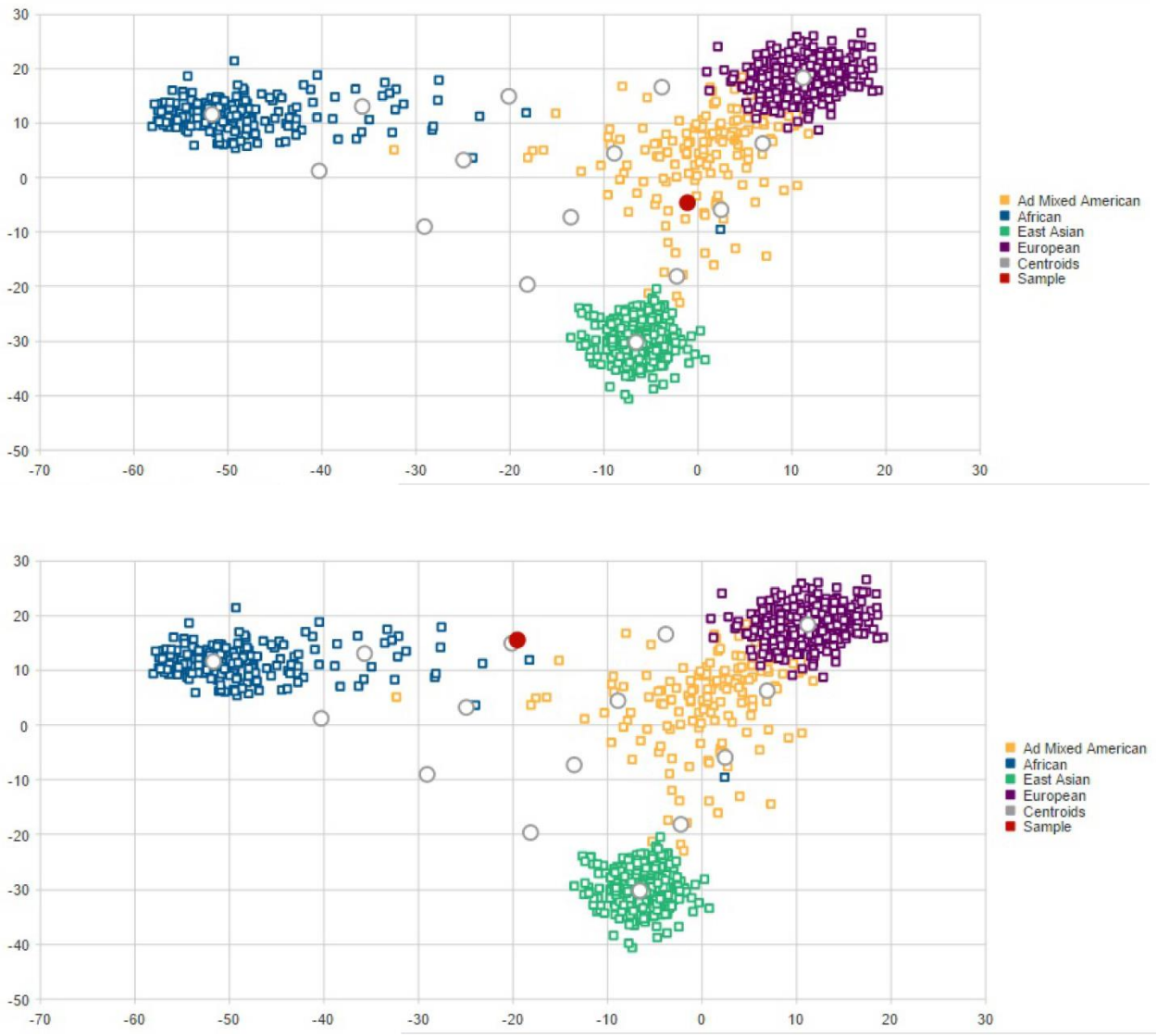
Supplemental Figure 3. BLAST results of 300 bp around SNP rs12149261. Figure 2A is rs12149261 located in HYDIN gene on chromosome 16. Figure 2B is the duplication of rs12149261 located in HYDIN2 gene on chromosome 1.


Ad Mixed American
African
East Asian
European
Centroids
Sample

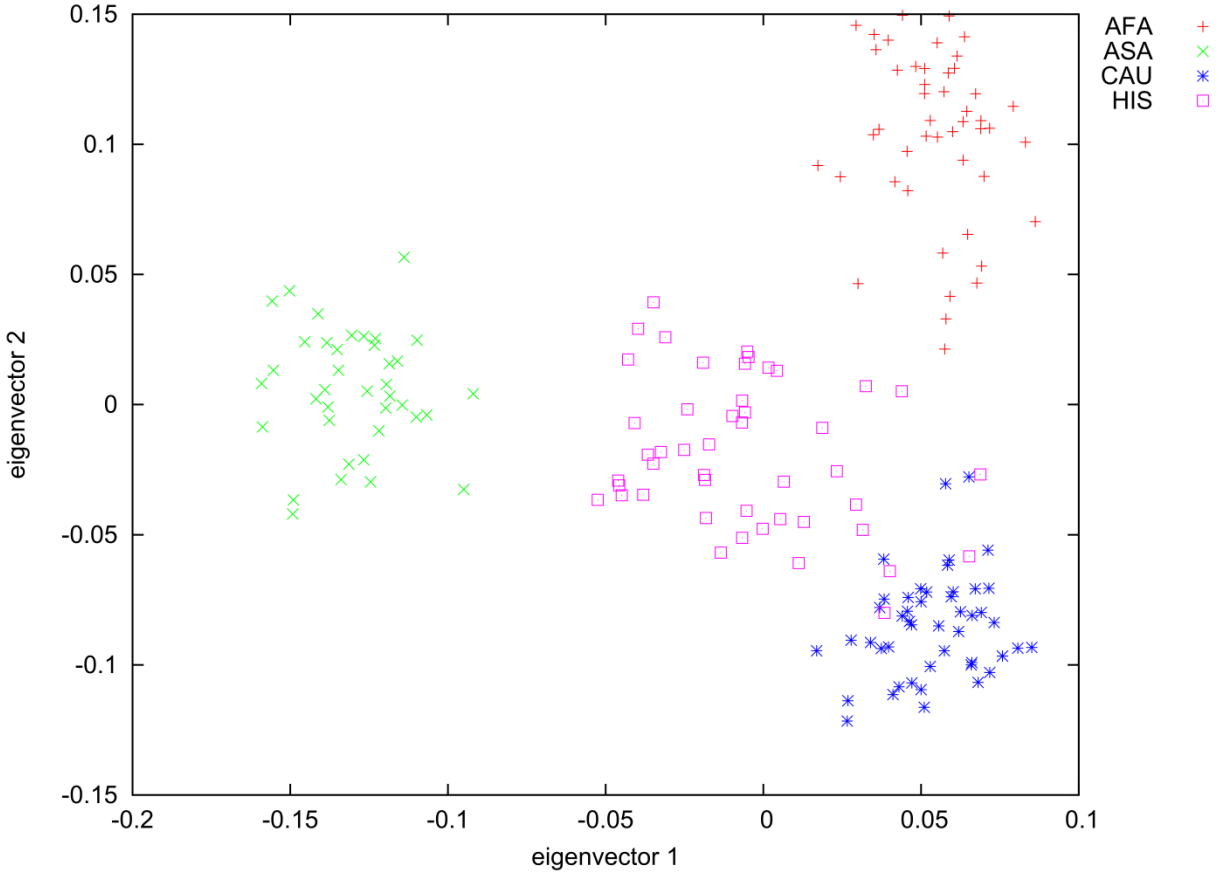

Ad Mixed American
African
East Asian
European
Centroids
Sample


Ad Mixed American
African
East Asian
European
Centroids
Sample


Ad Mixed American
African
East Asian
European
Centroids
Sample



Supplemental Figure 4. PCA plots of eight individuals using the ForenSeq™ panel. Eight samples (20882, 23169, 76194, 06498, 12574, 38859, 10916 and 61115) are labeled in red circles in Figures 4A-4H, respectively.



Supplemental Figure 5. The PCA plot of 181 individuals using 22-SNP AIMs panel.

SUMMARY

Selection of Highly Informative Markers for Apportionment of Ancestry and Population Affiliation

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. This doctoral dissertation research was conducted under the hypotheses that Absolute Allele Frequency Differences (δ) and F statistics (F_{ST}) perform better than Informativeness for Assignment Measure (In), and highly informative AIMs can be selected among human populations by using these three marker informativeness measures. The primary goal of this project was to develop a robust AIMs panel with minimum number of markers that can be used for apportionment of ancestry and population affiliation of four major US populations, such as African American, US Caucasian, East Asian and Hispanic American.

Chapter 1 described the selection process of highly informative AIMs from the four major American populations. Previously, scientists have used different marker measures (δ, F_{ST} and In) to select AIMs based on personal preferences. δ and F_{ST} can be used for homogeneous populations and admixed populations, while In is mainly used for admixed populations. While the logic of using these measures is similar, their efficacy has not been compared with objective selection of genome-wide SNPs, particularly for determining affiliations for major US populations. It is possible to compare whether any of these measures are better for discovery of highly informative AIMs and select a robust panel with minimum number of AIMs to characterize the four major populations in the United States, with an abundance of SNPs available in International HapMap project and 1000 Genomes. In this dissertation project, the AIMs selection was conducted based on four US populations of the HapMap project: 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). Values of the three measures of each candidate SNP were

computed for each pairwise population comparison, and markers were ranked based on these values. The top 30 AIMs, for each measure in each pairwise population comparison, were chosen and those markers in linkage disequilibrium (LD) were removed. The minimum number of markers to discriminate each pair of populations was identified for each measure based on principal component analysis (PCA), Receiver Operating Characteristics (ROC) curve, and the maximum Matthews correlation coefficient (MCC). The minimum number of AIMs needed to completely separate any of the six population pairs ranged from 2 to 9 SNPs. As expected, the largest number of SNPs was needed to resolve CEU and MEX. Finally, the top markers from six pairwise population comparisons were pooled based on the three measures and evaluated as individual panels. After removing duplicated SNPs and replacing SNPs that were in LD, the resultant total number of markers in the AIMs panel selected by δ, F_{ST} and In was 24, 23, and 23, respectively. These three AIMs panels showed high similarity rates. The PCA cluster results indicated that F_{ST} panel performed slightly better than the δ panel and notably 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 populations of the HapMap project and 1000 Genomes were used to evaluate the efficiency of the 23 AIMs panel *in silico*. The results indicated that these 23 AIMs can correctly assign individuals to the major population categories. However, the full power of the 23 AIMs panel could not be evaluated, because the genotype data of two or three SNPs were not found in these public databases for each of the nine populations.

In **Chapter 2**, these 23 SNPs were multiplexed to evaluate their actual performance in ancestry evaluations. DNA was obtained from 189 individuals collected from four American populations (African American, Asian, Caucasian, and Hispanics). The ancestry of each individual was based on self-declaration. The Nextera™ Rapid Capture Custom Enrichment kit

(Illumina) was used to enrich the target SNPs with custom oligonucleotide probes. Libraries were sequenced on the Illumina MiSeq. Data were analyzed using MiSeq Reporter and Genome Analysis Toolkit (GATK). The genotype data of the 23 SNPs were generated for the 189 individuals. SNP rs12149261 deviated from Hardy-Weinberg equilibrium (HWE) in three populations (Asian, Caucasian and Hispanic American) even after applying Bonferroni's correction for multiple testing. This SNP also exhibited significant LD with the other 22 SNPs. The results showed that 97% of the individuals in the three groups (Asian, Caucasian, Hispanic groups) were heterozygous (CA) at this SNP locus, which was inconsistent with Hardy-Weinberg equilibrium expectations and the expectations for an AIM SNP. BLAST results indicated that SNP rs12149261 (located on chromosome 16) had a duplicate region on chromosome 1. This problematic SNP was removed from my panel, because there is no reasonable way to genotype the marker. No SNPs departure from HWE was detected for the other 22 SNPs in four populations, and only five SNP pairs showed significant LD. PCA results showed that eight individuals were not assigned to the expected major population categories. These eight samples were re-analyzed using 56 AIMs contained in the ForenSeq^{TM} DNA Signature Prep Kit (Illumina), and the results were consistent with the 22 AIMs in my analyses. After removing these eight samples, no detectable LDs were observed in the four populations. The results indicated that the 22 AIMs can correctly assigned individuals to the four major US population categories. These 22 SNPs can contribute to the candidate AIMs pool for potential forensic identification purposes in major US populations.

CONCLUSIONS AND FUTURE DIRECTIONS

Selection of Highly Informative Markers for Apportionment of Ancestry and Population Affiliation

Ancestry informative markers (AIMs), one subclass of single nucleotide polymorphisms (SNPs), show large differences in allele frequency among human populations. These differences can be used to determine population affiliation and predict the ancestry of the source of a forensic evidence sample. Currently three marker informativeness measures are available for the selection of AIMs, but the efficacy of these measures has not been directly compared with the purpose of selecting highly informative SNPs to characterize four major US populations. My research project was conducted to examine the selection effectiveness of three informativeness measures in each population comparison, develop a robust AIMs panel to differentiate four major American populations, and evaluate the full efficacy of the final SNP panel in samples collected from these four populations.

Absolute Allele Frequency Differences (δ), F statistics (F_{ST}), and Informativeness for Assignment (In) are the three measures mainly used for the selection of AIMs. δ is a measure of the absolute frequency difference of a particular allele in two populations, and a high δ is sought. F_{ST} is the measure of genetic distance between two populations based on genetic data, and a high F_{ST} implies substantial degree of differentiation between populations. In infers as the likelihood ratio for the assignment of an allele to one of the populations relative to the average populations. The average population is hypothetical, and its allele frequencies are the mean allele frequencies of K populations. δ and F_{ST} have been used more so in AIMs selection, because these two measure can be used for homogeneous populations (e.g. East Asian and Caucasian) and admixed populations (e.g. African American and Hispanic American). While In is mainly used for admixed populations. Therefore, my dissertation project was conducted under the assumption that δ and F_{ST} perform similarly, and both of them perform better than In measure with the

objective of selecting AIMs from four major US populations (e.g. African American, US Caucasian, East Asian, and Hispanic American).

The AIMs selection was performed using genotype data downloaded from the International HapMap Project Phase III. This project contains genotype data of millions of SNPs for groups of individuals residing in different continents. My research focuses on four major populations in the USA: African ancestry from Southwest USA (ASW), Utah residents with Northern and Western European ancestry from the CEPH collection (CEU), Chinese from Metropolitan Denver, Colorado (CHD), Mexican ancestry from Los Angeles, California (MEX).

AIMs selection was first conducted between each pairwise population comparison. Principal component analysis (PCA), Receiver Operating Characteristics curve (ROC curve), and the maximum Matthews correlation coefficient (MCC) were used to determine the minimum number of SNPs to discriminate each pair of populations for each measure. The PCA method can reveal population structure by assigning samples with similar allele frequencies into the same group. The cutoff values of two clusters in PCA plot were determined by using a ROC curve. The PCA clustering performance of AIMs in individual classification was assessed using MCC. Two to nine top markers only were needed to separate any of the six population pairs. Finally, the top AIMs identified from six pairwise population comparisons were pooled based on each of the three measures and evaluated as individual panels. The PCA cluster results showed that F_{ST} panel performed slightly better than the δ panel and notably better than the In panel to characterize the four US populations (ASW, CEU, CHD, and MEX). Therefore, the 23 SNPs selected by the F_{ST} measure were used to develop a robust panel of AIMs for apportionment of ancestry and population affiliation. Nine populations from the HapMap project and 1000 Genomes were used to evaluate the efficiency of the 23 AIMs panel *in silico*. The results showed

that these 23 SNPs could be used to assign individuals to the expected population categories. However, the full performance of 23-AIMs panel could not be assessed, because the genotype data of some SNPs were missing in public databases.

Empirical testing was performed to test the actual performance of the 23 AIMs in ancestry evaluations. These 23 SNPs were combined in a multiplex system for evaluation. Individuals were collected from four American populations (African American, Asian, Caucasian, and Hispanics), and their ancestries were based on self-declaration. The Nextera™ Rapid Capture Custom Enrichment kit (Illumina) was used to capture the target SNPs and they were sequenced by massively parallel sequencing. The results showed that one SNP (rs12149261) deviated from Hardy-Weinberg equilibrium (HWE) and exhibited significant linkage disequilibrium (LD) with other SNPs. This SNP was duplicated in a region on chromosome 1, and its genotype data were actually a combination of sequence reads from two SNP sites. The resultant 22 AIMs could correctly assign individuals into the four major US populations.

The results of my dissertation studies indicate that it is possible to identify a small number of AIMs using any of the three informativeness measures to characterize the American populations and that F_{ST} performed better than the other measures. Twenty of the 22 AIMs had never been reported for ancestry assessment; only two markers (rs11725412 and rs1834640) had been reported for ancestry inference purposes. These 20 novel markers can contribute to the candidate pool of AIMs for correcting population stratification and potential forensic identification purposes. My panel has the fewest number of markers compared with other AIMs panels. Although massively parallel sequencing (MPS) allows for sequencing a large battery of forensically-relevant short tandem repeat (STR) markers and SNPs simultaneously, capillary electrophoresis (CE) still is the most popular platform used in forensic community. Therefore, it

is critical that an AIMs panel contains a small number of markers, making it feasible to be implemented by forensic scientists with current technology. The principles and methodologies used in this dissertation project may be applied to future research. Future studies may focus on selecting additional markers from other populations, searching new markers from the latest version of public databases, trying to resolve closely related populations, comparing my panel with other AIMs panels reported, and testing this 22-AIMs panel with forensic casework samples.

My dissertation project only focused on four major US populations: African American, US Caucasian, East Asian and Hispanic American. However, additional populations may not be able to be correctly characterized using my AIMs panel. For example, Native American individuals may be assigned to the East Asian group, Hispanic American group or fall between these two groups. In order to maximize US population resolution, the same selection criteria could be used to select AIMs to resolve the Native American population.

Two public databases were used in this project. The International HapMap Project was utilized for the selection of markers, while HapMap Project and 1000 Genomes were used in the evaluation of the final AIMs panel. At the time this research began, 1000 Genomes was under construction. Therefore, the genotype data of four US populations from the latest version of HapMap project (Phase III, released in 2009) were used for AIMs selection. Currently, 1000 Genomes is the most widely used SNP database, and it continues to expand. It is likely that new highly informative AIMs can be identified in 1000 Genomes using the methodology described in my dissertation project.

The principles used in my studies also can be applied to improve fine resolutions for closely related populations. In the ancestry inference studies, all populations from European countries are classified as European or Caucasian, and all populations from East Asia (Chinese, Korean,

and Japanese) are combined as East Asian. SNP allele frequency differences likely exist between individuals from more closely related populations. With the high throughput of MPS technology, it is possible to select hundreds or thousands of AIMs to characterize closely related populations. Future research could seek AIMs for finer population resolution.

As mentioned in Chapter 1, ten AIMs panels have been published, including my panel. It would be interesting to compare the efficacy of the specific SNPs in these panels to characterize the US populations. A combination of the best performing SNPs from these panels may provide a more robust set with fewer AIMs than in my panel. The genotype data of SNPs of the other nine panels in four major American populations (ASW, CEU, CHD, and MEX) can be downloaded from the HapMap Project. The actual performances of the ten panels can be compared directly, and new SNP combinations could be evaluated using PCA. Since some SNPs are not included in the HapMap Project, the efficiencies of some panels and those SNPs may not be fully assessed. It is best to perform empirical testing as described in Chapter 2 to generate data on all SNPs of the ten panels.

AIMs have been used in forensic casework to predict the ancestry of the source of an unknown biological sample. The well-known case in Louisiana highlights the effectiveness of AIMs testing for developing an investigative lead. The methodology for AIMs typing in my research employed a capture method for target enrichment. This method required 50 ng of input DNA. The amount of DNA for casework typically is far less. Therefore, research should continue to develop a PCR-based enrichment method in which < 1 ng of DNA can be analyzed. Once developed, the 22-AIMs panel can be validated with mock forensic casework samples. In Chapter 2, 189 individuals were investigated with these 22 SNPs and only four samples had a single locus drop out. One locus drop out was tolerated and did not significantly affect ancestry

prediction. However, future studies should determine the degree of locus drop out that can be tolerated with the 22-AIMs panel and still effectively predict ancestry for forensic investigations.

The work described in this dissertation was performed in accordance with all laws (both Federal and State) that apply to research, researcher conduct, and the protection of human test subjects. It also was conducted under the guidance of and in accordance with the policies of the University of North Texas Health Science Center Institutional Review Board.

REFERENCES

Selection of Highly Informative Markers for Apportionment of Ancestry and Population Affiliation

Adinsoft SARL (2010) XLSTAT-software. Version 10. Addinsoft, Paris

Amirisetty S, Hershey GK, Baye TM (2012) AncestrySNPminer: a bioinformatics tool to retrieve and develop ancestry informative SNP panels. Genomics 100:57-63

Baye TM, Tiwari HK, Allison DB, Go RC (2009) Database mining for selection of SNP markers useful in admixture mapping. BioData Min 2:1

Bushnell D, Hudson RA (2010) Colombia: a country study. Federal Research Division, Library of Congress, Washingtion D.C.

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

Collins-Schramm HE, Phillips CM, Operario DJ, Lee JS, Weber JL, Hanson RL, Knowler WC, Cooper R, Li H, Seldin MF (2002) Ethnic-difference markers for use in mapping by admixture linkage disequilibrium. Am J Hum Genet 70:737-750

DesignStudio, 2015, https://accounts.illumina.com/?ReturnUrl=http://designstudio.illumina.com/

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

Garcia-Closas M, Chanock S (2008) Genetic susceptibility loci for breast cancer by estrogen receptor status. Clin Cancer Res 14:8000-8009

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

Goddard KA, Hopkins PJ, Hall JM, Witte JS (2000) Linkage disequilibrium and allele-frequency distributions for 114 single-nucleotide polymorphisms in five populations. Am J Hum Genet 66:216-234

Green SB, Salkind NJ, Akey TM (2008) Using SPSS for Windows and Macintosh: Analyzing and understanding data. Prentice Hall, New Jersey

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

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

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

International HapMap Consortium (2003) The International HapMap Project. Nature 426:789-796

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

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

Jin L, Chakraborty R (1995) Population structure, stepwise mutations, heterozygote deficiency and their implications in DNA forensics. Heredity 74:274-285

Jobling MA, Tyler-Smith C (2003) The human Y chromosome: an evolutionary marker comes of age. Nat Rev Genet 4:598-612

Jobling MA, Gill P (2004) Encoded evidence: DNA in forensic analysis. Nat Rev Genet 5:739-751

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

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

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

Knowler WC, Williams RC, Pettitt DJ, Steinberg AG (1988) Gm3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. Am J Hum Genet 43:520-526

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

Lewis PO, Zaykin D (2001) Genetic Data Analysis: Computer program for the analysis of allelic data. Version 1.0 (d16c). http://http://hydrodictyon.eeb.uconn.edu/people/plewis/software.php. A ccessed 25 April 2007

Linse KD (2012) Genes that define the shape of our face. http://blog-biosyn.com/2012/11/28/123/

Mao X, Bigham AW, Mei R, Gutierrez G, Weiss KM, Brutsaert TD, Leon-Velarde F, Moore LG, Vargas E, McKeigue PM, Shriver MD, Parra EJ (2007) A genomewide admixture mapping panel for Hispanic/Latino populations. Am J Hum Genet 80:1171-1178

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

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

Nextera Rapid Capture Enrichment Reference Guide, 2015, https://support.illumina.com/downlo ads/nextera-rapid-capture-guide-15037436.html

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

Parra EJ, Marcini A, Akey J, Martinson J, Batzer MA, Cooper R, Forrester T, Allison DB, Deka R, Ferrell RE, Shriver MD (1998) Estimating African American admixture proportions by use of population-specific alleles. Am J Hum Genet 63:1839-1851

Patterson N, Price AL, Reich D (2006) Population Structure and Eigenanalysis. PLoS Genet 2:e190

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

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

Price AL, Patterson N, Yu F, Cox DR, Waliszewska A, McDonald GJ, Tandon A, Schirmer C, Neubauer J, Bedoya G, Duque C, Villegas A, Bortolini MC, Salzano FM, Gallo C, Mazzotti G, Tello-Ruiz M, Riba L, Aguilar-Salinas CA, Canizales-Quinteros S, Menjivar M, Klitz W, Henderson B, Haiman CA, Winkler C, Tusie-Luna T, Ruiz-Linares A, Reich D (2007) A genomewide admixture map for Latino populations. Am. J Hum Genet 80:1024-1036

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945-959

QIAamp® DNA Mini and Blood Mini Handbook, 2012, https://www.qiagen.com/us/resources/resourcedetail?id=67893a91-946f-49b5-8033-394fa5d752ea&lang=en

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

Qubit® dsDNA BR Assay Kit, 2015, https://www.thermofisher.com/order/catalog/product/Q32850

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516-1517

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

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW (2002) Genetic structure of human populations. Science 298:2381-2385

Rosenberg NA, Li LM, Ward R, Pritchard JK (2003) Informativeness of genetic markers for inference of ancestry. Am J Hum Genet 73:1402-1422

Rosenberg N (2004) Distruct: a program for the graphical display of population structure. Mol Ecol Notes 4:137-138

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

Schlesselman JJ (1982) Case-control studies: design, conduct, analysis. Oxford: Oxford University Press

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

Shriver MD, Smith MW, Jin L, Marcini A, Akey JM, Deka R, Ferrell RE (1997) Ethnicaffiliation estimation by use of population-specific DNA markers. Am J Hum Genet 60:957-964

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

Shriver MD, Kittles RA (2004) Genetic ancestry and the search for personalized genetic histories. Nat Rev Genet 5:611-618

Slatkin M (2007) Inbreeding coefficients and coalescence times. Genet Res 89:479-487

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

SPSS Inc (2007) SPSS for Windows. Version 16.0. Chicago

Stephens JC, Schneider JA, Tanguay DA, Choi J, Acharya T, Stanley SE, Jiang R, Messer CJ, Chew A, Han JH, Duan J, Carr JL, Lee MS, Koshy B, Kumar AM, Zhang G, Newell WR, Windemuth A, Xu C, Kalbfleisch TS, Shaner SL, Arnold K, Schulz V, Drysdale CM, Nandabalan K, Judson RS, Ruano G, Vovis GF (2001) Haplotype variation and linkage disequilibrium in 313 human genes. Science 293:489-493

Tian C, Hinds DA, Shigeta R, Adler SG, Lee A, Pahl MV, Silva G, Belmont JW, Hanson RL, Knowler WC, Gregersen PK, Ballinger DG, Seldin MF (2007) A genomewide single-nucleotide-polymorphism panel for Mexican American admixture mapping. Am J Hum Genet 80:1014-1023

Vigilant L, Stoneking M, Harpending H, Hawkes K, Wilson AC (1991) African populations and the evolution of human mitochondrial DNA. Science 253:1503-1507

Wacholder S, Rothman N, Caporaso N (2000) Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. J Natl Cancer Inst 92:1151-1158

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

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 [Epub ahead of print]

Williams RC, Long JC, Hanson RL, Sievers ML, Knowler WC (2000) Individual estimates of European genetic admixture associated with lower body-mass index, plasma glucose, and prevalence of type 2 diabetes in Pima Indians. Am J Hum Genet 66:527-538

Wright S (1950) Genetical structure of populations. Nature 166:247-249

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

Zeng X, Chakraborty R, King JL, LaRue B, Moura-Neto RS, Budowle B (2016) Selection of highly informative SNP markers for population affiliation of major US population. Int J Legal Med. 130:341-352

Zweig MH, Campbell G (1993) Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem 39:561-577

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

Zeng X, Warshauer D, King JL, Churchill JD, Chakraborty R, Budowle B (2016) Empirical testing of a 23-AIMs panel of SNPs for ancestry evaluations in four major US populations. Int J of Legal Med. 2016 Feb 25. [Epub ahead of print]