Abstract

Many ancestry informative SNP (AISNP) panels have been published. Ancestry resolution in them varies from three to eight continental clusters of populations depending on the panel used. However, none of these panels differentiates well among East Asian populations. To meet this need, we have developed a 74 AISNP panel after analyzing a much larger number of SNPs for F_{st} and allele frequency differences between two geographically close population groups within East Asia. The 74 AISNP panel can now distinguish at least 10 biogeographic groups of populations globally: Sub-Saharan Africa, North Africa, Europe, Southwest Asia, South Asia, North Asia, East Asia, Southeast Asia, Pacific and Americas. Compared with our previous 55-AISNP panel, Southeast Asia and North Asia are two newly assignable clusters. For individual ancestry assignment, the likelihood ratio and ancestry components were analyzed on a different set of 500 test individuals from 11 populations. All individuals from five of the test populations--Yoruba (YRI), European (CEU), Han Chinese in Henan (CHNH), Rondonian Surui (SUR) and Ticuna (TIC)--were assigned to their appropriate geographical regions unambiguously. For the other test populations, most of the individuals were assigned to their self-identified geographical regions with a certain degree of overlap with adjacent populations. These alternative ancestry components for each individual thus help give a clearer picture of the possible group origins of the individual. We have demonstrated that the new AISNP panel can achieve a deeper resolution of global ancestry.

Keywords

Ancestry inference, AIMs, Reference populations, Eastern Asia

1. Introduction

Bio-geographic ancestry analysis has attracted considerable interest because of its contribution to biomedical studies, personalized medicine, human history studies, and forensic investigation [1-6]. In terms of forensic application, ancestry inference estimates can sometimes provide crucial investigative leads in a case when there is no database or suspect match, or when no specific suspect has been designated. Numerous panels of ancestry informative markers (AIMs) using SNPs or Insertion-deletion marker have been published [7-15]. AIMs panel resolution varies from three to eight clusters of populations at the level of continental regions. Three to five geographical region resolution is possible with relatively few AISNPs [16]. It would be ideal for forensic applications to achieve two goals simultaneously—broad assignment to major geographical regions and fine-scale resolution within a geographical region using a small panel of highly discriminating markers. But, there can be a resolution trade-off between the size of a panel and its ancestry informativeness in trying to achieve both goals. A separate second-tier ancestry panel can be pursued for within-region differentiation after a global assignment [13, 17] has been made.

A relatively comprehensive database of reference population frequencies is another requirement for the practical interpretation of an ideal ancestry panel. Proper assignment of ancestry depends both on the specific AIMs selected and having an extensive reference database. Many global ancestry inference panels have been published employing SNP panels that do not overlap very much and only a very small subset have been studied on an extensive array of reference populations[18]. Given the existing panels that allow several biogeographic regions to be distinguished with data on the same reference population data, building a new panel of AISNPs from scratch would be a wasteful exercise. Therefore, we chose to start from three sets of SNPs already studied on a large number of populations in common and then collect data on more

populations from Eastern Asia for the best of those SNPs. This approach assured a large global reference population without extensive new data collection.

The question of how many SNPs should be included in any AIMs panel is a moving target. Until now, the most common method of multiplexing a forensic assay has been primer extension reaction followed by capillary electrophoresis. However, while that procedure is easily implemented in most forensic laboratories [19, 20], the size of such a panel is limited to a few dozen SNPs because of dye and size constraints. Modern massively parallel sequencing (MPS) and DNA array-based technologies provide good alternative typing methods without the limitation in the number of SNPs that can be simultaneously assayed [21, 22].

East Asia is one of the most important regions for studying evolution and genetic diversity of human populations because of its large population [5, 23, 24]. Our existing panel of 55 AISNPs [13, 14] has been studied on 125 populations from around the world and can distinguish eight clusters of populations globally: Sub-Saharan Africa, North Africa, Europe, Southwest Asia, South Asia, East Asia, Americas and Oceania/Pacific. However, this panel cannot distinguish among populations within East Asia. Previous phylogeographic studies using uniparental DNA markers have revealed a genetic divergence between northern and southern East Asian populations most likely caused by the southern and northern migration routes into East Asia [25-28]. Different predominant Y-SNP haplogroups and their estimated expansion times have been determined for East Asians of North, East and Southeast. Genome-wide SNP studies also have found support for a north-south cline in genetic differences in East Asia, and numerous autosomal SNPs have been identified showing allele frequency differences between East and Southeast Asian populations, and even North-South Han Chinese. [29-32].

To increase the resolution among East Asian populations while maintaining an ability to be useful globally, we analyzed a much larger number of SNPs from previously published panels [10, 13, 33]. Based on a reference dataset of 3,312 individuals from 61 world populations uniformly typed for 178 SNPs, we chose a panel of 74 AISNPs. The panel can now distinguish at least 10 population clusters globally. Ancestry assignment performance of the 74 AISNPs was evaluated with a test sample of 500 individuals from 11 populations--all different from the reference samples.

2. Materials and methods

2.1. Samples

Table 1 summarizes the population samples used in this study. The 61 reference populations consist of 49 populations (N= 2105) routinely studied at Kidd Lab, Yale University School of Medicine that have been described elsewhere [13] and 12 of the 26 Phase 3 populations (N=1,207) from the 1000 Genomes Project (http://www.1000genomes.org). The validation step test samples consist of six populations (N=142) from Caixia Lab, Institute of Forensic Science of China, two populations from the 1000 Genomes Project (N=207), and three populations (N=151) from among the standard populations from Kidd Lab. The Kidd Lab and 1000 Genomes DNA samples were purified from cell lines and the Caixia Lab DNA was purified from blood samples. All samples were obtained under the supervision of the appropriate review boards and with informed consent and self-declared ancestry information.

2.2. SNPs

Table 2 provides the list of the 74 AISNPs, their chromosome and nucleotide position. These SNPs were selected from previously published panels [10, 13, 33]. We first chose 18 SNPs that are informative in East Asia using F_{st} values (F_{st}>0.15) and linkage disequilibrium (LD, r^2<0.2) across East Asian populations from the SNPs of Brissenden et al. [33]. Then we combined these SNPs with the other two panels [10, 13] and

generated a panel of 178 SNPs (Supplemental Table 1).

2.3. Basic statistics

Based on 61 reference populations and the 178 SNPs, we calculated F_{st} values and allele frequency differences among geographical region clusters (δ_R) and pair-wise populations (δ_P). F_{st} was calculated across populations for the allele frequencies using Wright's formula [13, 15]. Difference scores, δ_R and δ_P were calculated based on the highest likelihood STRUCTURE [34] run of 178 SNPs on 61 reference populations at K=10. Estimated allele frequencies for each cluster were extracted from the Results file of this STRUCTURE run for further calculation. $\delta_R = \left| p_{cluster\text{-}n} - p_{mean\ of\ the\ remaining\ clusters} \right|$, where p is the allele frequency of each cluster. δ_P is the pair-wise allele frequency difference between a specific pair of clusters.

2.4. Software

A heatmap was created based on population allele frequencies using program R v. 3.0.1. Principal components analysis (PCA) for population allele frequencies was calculated using XLSTAT 2015 (http://www.xlstat.com/en/about-us/company.html), the plot was created using program R v. 3.0.1. STRUCTURE v. 2.3.4 [34] was employed to evaluate and visualize population and individual ancestral proportions using 10,000 burnins and 10,000 MCMC; admixture model;. Graphics were generated using CLUMPAK (http://clumpak.tau.ac.il/index.html), and the threshold for similarity scores was set at 0.90 [35]. STRUCTURE HARVESTER was used to generate a table of mean likelihoods L(K) and variance per K value from STRUCTURE[36]. For δ_R and δ_P calculations, STRUCTURE was run from K=3 to K=13 for 30 independent replicates using 178 SNPs on 61 reference populations. For the 74 AISNP panel evaluation, STRUCTURE was run from K=3 to K=11 for 20 independent replicates using 74 SNPs on 72 populations, including all the reference and test populations.

For individual ancestry assignment, we calculated population assignment match probability (AMP) for all the test samples using Intelligence (FI version 1.0, available online: https://github.com/jiangl1989/FI/) [37, 38]. The likelihood ratio was calculated based on AMP as described in previous reports [13, 39]. Regions within one order of magnitude of the highest likelihood for each test individual are listed in Supplemental Table 2. Inferred ancestral components for each individual (Supplemental Table 2) were pulled out from Result file of the most likely STRUCTURE run at K=10 (Figure 1). PCA analyses for test individuals were generated via the web-base page Snipper application (http://mathgene.usc.es/snipper/).

2.5. Laboratory

Samples from Kidd Lab were typed for all SNPs by TaqMan® assays (Life Technologies, California, USA) following the manufacturer's instructions [13]. Samples from Caixia Lab were typed by Illumina Inc. custom Golden Gate genotyping assay procedure and Sequenom MassARRAY platform with the iPLEX GOLD chemistry, and the typing service was provided by CapitalBio Corp., Beijing, China. Individuals and SNPs with more than 10% missing entries were left out.

2.6 Selecting the Panel

From a preliminary evaluation using all 178 SNPs, we found this panel can achieve a global resolution of ten clusters. Next, less informative or redundant SNPs were identified for removal in order to obtain a smaller panel with the same amount of information. F_{st} was used to measure how informative a SNP marker is across the 61 global populations. δ_R shows which geographical region population contributes most to this variation. δ_P shows the value of each SNP for the differentiating populations in: East Asia and Southeast Asia ($\delta_{EA/SEA}$), East Asia and North Asia ($\delta_{EA/NA}$), European and Southwest Asia ($\delta_{EUR/SWA}$). Generally, SNPs for which one of the values reached the

threshold: Fst⩾0.25，δ_R⩾ 0.45 or δp⩾0.15 were kept in the set. Then heatmap was visually used to reduce near duplicates. The relationship of different SNPs is shown in the dendrogram of the heatmap. SNPs that are very close to each other and show similar color pattern across all 61 populations usually give similar information. In order to balance the selection of SNPs for different geographical regions, we deleted several SNPs that were in excess for one region compared with others, such as highly informative SNPs for Sub-Saharan Africa and the Americas. During the iterative process of selection and deletion of SNPs and evaluation of the interim panel, STRUCTURE and PCA analysis were employed to make sure the panel continued to provide essentially the same amount of information as the original 178 SNPs. After multiple repetitions testing the effect of excluding or including particular SNPs, we obtained the 74 AISNP panel.

3. Results

3.1. Characteristics of the 74-SNPs panel

The details of the 74 AISNPs are given in Table 2. F_{st}, which can range from zero to one, is a measure of how much the SNP frequencies vary across the populations studied. The global F_{st} values among 61 world populations ranged from 0.12-0.72 with a mean F_{st} of 0.29. In the STRUCTURE result for 178 SNPs on 61 reference populations (Supplemental Figure 1), one new cluster was recognized at each increasing K value; in succession they were North Africa/Southwest Asia (at K=6), Pacific (at K=7), Southeast Asia (at K=8), Southwest Asia (at K=9), and North Asia (at K=10). At K = 10, all the individuals are assigned to ten distinct clusters: Sub-Saharan Africa, North Africa, Southwest Asia, Europe, South Asia, North Asia, East Asia, Southeast Asia, Pacific and Americas. From K=11 to K=13, a mixed pattern is introduced to South Asia. We checked STRUCTURE HARVESTER [36] and get a summary table of the mean likelihood (L(K)) and variance for all the K values (Supplementary table 3). The top two optimum K (Mean LnP(K)) are11 and 10. We choose K=10 as a stopping point to avoid a mixed pattern

being introduced into this panel of reference populations. Then δ_R and δ_P were calculated based on the highest likelihood run at K=10. For example, rs3811801 ($\delta_{R\text{-}EA}$=0.70) contributes most to the allele frequency difference of East Asia compared to other geographical regions, the pairwise allele frequency differences of this SNP in EA/SEA and EA/NA is 0.52 and 0.71, respectively. The data demonstrate that rs3811801 is very informative for the differentiation of East Asians, confirming one of our earlier studies [40]. rs2814778 has an F_{st} value of 0.64, while the allele frequency difference between Sub-Saharan African and other geographical populations accounts for most of the variation ($\delta_{R\text{-}SAFR}$=0.97).

We used Heatmap to visualize the relationships of different SNPs before and after the marker selection and deletion. Supplementary Figure 2 is the heatmap of 178 SNPs in 61 reference populations. SNPs in the same sub-clade show similar color patterns and usually give the same information. One of the markers can be deleted to avoid redundancy. For example, in the dash line highlighted area, rs1871534, which gives similar information to sub-Sahara Africa with rs2814778, was deleted. Allele frequency distribution characteristics of 74 SNPs in 61 reference populations are displayed in Figure 2. The color intensity contrasts how similar and different the SNP frequencies are in the populations. For example, rs10516441, rs6054605, rs3811801, and rs1800414 are clustered together and show different color intensity in East Asian and Southeast Asian populations from the rest of the world. The frequency patterns of these four AISNPs across the 61 world populations are further displayed in Supplemental Figure 3. Rs16891982, rs12913832 and rs6754311 show different color intensity in European from the other populations. Supplemental Figure 4 further displays the frequency patterns of the three SNPs in all the populations.

3.2. Population relationships assessed with 72 world populations

Figure 3A shows the first three factors of the PCA analysis (Principal Component

Analysis) based on allele frequencies in 72 world populations. The first three factors account for 76.62% of the variation. Factor 1, which accounts for 39.75% of the variation, is primarily defined at the extremes by populations from Europe and the Americas. Factor 2 (22.51% of the variation) separates African populations from the rest of the world, while Factor 3 (14.36% of variation) distinguishes Native Americans and East Asians. In order to give a clearer display of the performance of 74 AISNPs on East Asians, we performed the PCA analysis based on allele frequencies in 24 populations of this region (Figure 3B). Factor 1 and factor 2 account for 59.38% of the variation. Factor 1 separates North Asians from others, and Factor 2 separates Southeast Asian from East Asian populations. South Han Chinese (CHT, CHF, CHS, HKA, CGXH) distribute in between the East Asia and Southeast Asia populations. Factor 3 (8.62% of the variation) separate the Ami and Atayal populations from the other Southeast Asians.

Figure 1 shows the STRUCTURE result of the reference and the test populations combined; this is the most common pattern and it has the highest likelihood found among the 20 runs at K = 10. In the bar plot the estimated cluster membership frequency is shown for each individual as a colored bar column. K is the number of clusters specified to STRUCTURE by the user. Graphically each cluster is assigned a color. The length of any given color within the individual bars corresponds to the estimated percentage of cluster membership. Individuals in the same population are clustered together for display but the algorithm was not told which population each individual belongs to. At K = 10, all the individuals are assigned to ten distinct clusters: Sub-Saharan Africa, North Africa, Southwest Asia, Europe, South Asia, North Asia, East Asia, Southeast Asia, Pacific and Americas. Compared with our previous AISNP panel [13, 14], Southeast Asia and North Asia are two newly assignable clusters. Eight of the test populations have a predominant major ancestry component; African and European are the major ancestry components for YRI and CEU, respectively; East Asian for KSK

(Koreans, South Korea) and CHNH (Han Chinese in Henan, China); Southeast Asian for CYD (Dai in Xishuangbanna, Yunnan, China) and CGJ (Kinh in Guangxi, China); Americas for SUR (Rondonian Surui) and TIC (Ticuna). Three populations have admixed ancestry components, ASH (Ashkenazi Jews) are Southwest Asia and Europe. CMHM (Inner Mongolia in Hailar, China) are East Asian and North Asian. CGXH (Han Chinese in Guangxi, China) is East Asian and Southeast Asian.

3.3. Ancestry analysis of 500 test samples of 11 populations

We evaluated the ancestry assignment performance of the 74 AISNPs panel using 500 individuals from 11 population samples. All the individuals in the test samples were not included in the reference dataset of 61 populations used to develop the panel. On the basis of the genotyping data of the 61 reference populations, we calculated the population assignment match probability and likelihood ratio for all the test samples in10 geographic regions [13, 37, 38]. Table 3 summarizes the highest likelihood ratio geographic region for the 500 test individuals genotyped on the 74 AISNP panel. The highest likelihood region for all the YRI individuals was the Sub-Saharan African group (SAFR). The same high assignments performance was achieved for CEU, CHNH (Henan Han), SUR (Rondonian Surui) and TIC (Ticuna) to their respective and appropriate reference groups. The first likelihood geographic region for 70% of ASH (Ashkenazi Jews) individuals was Europe (EUR), and the remaining individuals fall into Southwest Asia (SWA) and North Africa (NAFR). Most of the CMHM (Inner Mongol) individuals were assigned to North Asia (NA) (52%) and East Asia (EA) (42%). A majority of the KSK (Korean) (88%) and CGXH (Guangxi Han) (75%) individuals were assigned to the cluster of East Asian. Most of the Yunnan Dai (CYD) (92%) and Guangxi Kinh (CGJ) (78%) were assigned to the Southeast Asian group.

Regions within one order of magnitude of the highest likelihood and inferred ancestry

component of the 500 test individuals were summarized in Supplemental table 2. Ancestral components for each individual (STRUCTURE, K=10) were analyzed by combining genotype data of the 61 reference populations and 11 test populations (Figure 1). To visualize the geographic ancestry of test samples, PCA for test individuals of each population are displayed in Supplementary Figures 5-15. The results are similar with likelihood calculations. Moreover the plots describe how unknown individuals overlap with known individuals.

4. Discussion

For the existing forensic ancestry inference panels, East Asians from the North and South tend to be clustered together without further resolution. Here we report that further resolution within East Asia is possible with a panel of 74 SNPs. In a STRUCTURE analysis of our 74-AISNPs panel (Figure 1), Southeast Asian, Dai, Kinh, Laotian, Cambodian, Ami and Atayal emerge as a distinct cluster from East Asia. Four southern Han populations in China, CHF, CHT, HKA and CHS display a small fraction of mixture of a Southeast Asian component which can be explained by the frequent intermarriage among geographically close populations. Khanty, Yakut, Tsaatan and Outer Mongolians form a distinct North Asian cluster. Khanty (KTY) who settled in the Western Siberia display a main component of North Asian with a smaller component of European. Outer Mongolians (OMG) show a mixture of the North Asian and East Asian clusters.

For individual ancestry estimation, we employed likelihood ratio and ancestry component to analyze all test individuals in ten geographic regions. Population assignment match probability (AMP) is calculated based on the allele frequency of 74 AISNPs in the reference populations. AMP can be considered as proportional to the likelihood of the population given the genotype [39]. But the highest likelihood is not necessarily the definitive ancestry. Geographic regions with similar likelihood should all be considered. Meanwhile, continuous population migration and intermarriage can cause misclassification if we only look into the value of likelihood ratio, especially for geographically close populations. Under this circumstance, ancestry components provide a good indicator of ancestry origin of an individual. For example, the highest likelihood region for one of the CEU samples NA07037 is Europe. But Southwest Asia is within one order of magnitude of the highest likelihood, which indicates that both regions should not be excluded. Ancestry components figure of this sample displayed a mixed components of Southwest Asia (0.79) and Europe (0.12) (Supplemental Table 2). The

result is reasonable for populations in coastal Mediterranean areas. It is also not surprising to find Ashkenazi Jews (ASH) individuals can be assigned to Southwest Asian, European or North African regions because of a long history of migration for ASH. In supplementary figure 7, AJ1801 is superimposed in the reference population of North Africa and Southwest Asia. The ancestry component of NAFR (0.50) and SWA (0.47) is displayed clearly in the pie chart of Supplemental Table 2. So both regions should be considered for this individual.

In East Asia, inner Mongolians settled in the border area between North Asia and East Asia. Intermarriage with Han Chinese or nearby East Asians is very common for this population, accounting for their mixed ancestry components. In supplementary figure 8, we can see most of the test individuals were superimposed in NA and EA populations. For example, individual CMHM62 has a mixed component of East Asian (0.51) and North Asian (0.29), and both NA and EA are within an order of magnitude of the highest likelihood (Supplemental Table 2). Henan province is located in the central part of China and is regarded as one of the three cradles of Chinese civilization. The Han Chinese in Henan (CHNH) have higher East Asian ancestry component compared with other Han Chinese who migrated to the South in various historical periods (CHS, CHF, CHT, HKA and CGXH). All the test individuals of CHNH studied here are assigned to East Asia (Table 3). The southern Han Chinese display a degree of Southeast Asian ancestry which could be due to gene flow between neighboring regions. Six populations in the reference sample (CDX, KHV, LAO, CBD, AMI and ATL) are Southeast Asia populations. The Dai people are a cross-border group, who live mostly in Yunnan China and Thailand. Our test sample of Dai came from Xishuangbanna, Yunnan (CDX). Two of the 24 samples (CYD 14 and CYDH7) have an East Asian ancestry component higher than 90% (Supplemental Table 2); regions within one order of magnitude of the highest likelihood are all in East Asia. The ‘mis-assigned outlier’ may be a result of Mendelian segregation

such that some individuals in a population may have genotypes that are more likely to occur in a population other than their origin [13]. Ancestry assignment is probabilistic, based on allele frequency; it is inevitable that overlap exists among nearby regions. Or, the 'mis-assignment' may be caused by sampling error. Sometimes even the person himself may not know that his self-declared ethnicity group is not consistent with his biological ancestry for historical or societal reasons. Kinh in Guangxi China (CGJ) are Vietnamese who emigrated from Vietnam hundreds of years ago and live near the Sino-Vietnamese border. In Table 3, the first likelihood ratio region is not East Asian for 4 of the 18 CGJ samples. But in the likelihood ratio list (Supplemental Table 2), both East Asian and Southeast Asian are in the high ranking region list and cannot be excluded. All four individuals showed mixed ancestry components of both regions (Supplemental Table 2).

For forensic applications, it would be ideal to have a relatively small-panel of AIMs in one multiplex in order to achieve a finer-scale of resolution of populations, because sensitivity is a very important consideration for a new method to be widely accepted in crime laboratory. If an ancestry panel is not sensitive enough to genotype touch DNA or other trace DNA evidence, the panel will lack utility. However, the human population structure is much more complex than what we can infer using a limited number of genetic markers. It is also very difficult to develop a single panel solution with a limited number of SNPs since it is hard to find SNPs that differentiate population both globally and within regions [7, 13]. In this study, we are aiming for the right balance between global and within region differentiation. China has the largest population in East Asia with a complex population structure. Various groups of Eurasians inhabit the Northwest. Minority groups and Han Chinese in the North have a long history of contact with nomadic groups in North Asia. And in the South, the gene flow between Chinese and Southeast Asian populations has been continuous. In major urban areas like Beijing or Guangzhou,

Europeans and Africans are very common now. At the central crime laboratory in China, more than half of the forensic cases deal with trace DNA. To meet the needs of forensic application, we are trying to develop one multiplex panel to achieve a global differentiation as well as a further resolution within East Asia. The 74 SNP markers reported here move us towards this goal in a preliminary fashion. Using this panel, we can observe a clear ancestry difference among North Asian, East Asian and Southeast Asian populations. In the next phase, this panel should be evaluated more thoroughly using a more comprehensive test sample validation; and more reference populations should be included to allow a better representation of the whole world. By the current 74 panel, no main component difference has been determined between Northern and Southern subgroups, nor among Japanese, Korean and Chinese.

$F_{st,}$ δ_R , $\delta_{P,}$ and heatmap facilitated analyzing the contribution of each SNP to the differentiation of specific populations or clusters of populations. This also provides a good method for determining similarly behaving SNPs for substitution into a multiplex system if some SNPs are difficult to incorporate into the multiplex.

In conclusion, the 74-AISNPs panel achieved a ten-cluster global resolution with two more clusters determined in North Asia and Southeast Asia than observed with our previous 55-AISNP panel. By starting from panels of SNPs already sharing a global set of reference populations, we have maintained a global perspective for assigning ancestry to an unknown sample. The new panel performed well on a test set of 500 individuals from 11 populations which were not included in the process of developing the new set of 74 AISNPs. Additional testing of the new panel on additional East and Southeast Asian populations is needed. As additional SNPs are typed on the full set of reference populations, they can be compared with these 74 AISNPs and possibly used to improve differentiation among populations in East Asia and/or other regions of the

world. Thus, better SNPs for ascertaining ancestry among the region's populations may also be identified in the future, leading to a revision of this 74 AISNP panel.

Acknowledgments

This work was funded in part by the U.S. Department of Justice grants NIJ 2013-DN-BX-K023 and NIJ 2014–DN-BX-K030 and by the U.S. National Science Foundation grant BCS-1444279 to KKK. Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Department of Justice. This work was also partially funded by the National Natural Science Foundation of China (81202384) and the Key Projects in the National Science & Technology Pillar Program in the 12th-year Plan Period (2012BAK02B01). In addition, some of the cell lines were obtained from the National Laboratory for the Genetics of Israeli Populations at Tel Aviv University, and the African American samples were obtained from the Coriell Institute for Medical Research, Camden, New Jersey. We would also like to thank all the collaborators who helped to collect the samples. Special thanks are due the hundreds of individuals who volunteered to give blood samples for studies of gene frequency variation.

References

[1] S.A. Tishkoff & K.K. Kidd, Implications of biogeography of human populations for 'race' and medicine, Nat Genet 36 (2004) S21-27.

[2] J.Z. Li, D.M. Absher, H. Tang, A.M. Southwick, A.M. Casto, S. Ramachandran, H.M. Cann, G.S. Barsh, M. Feldman, L.L. Cavalli-Sforza, & R.M. Myers, Worldwide human relationships inferred from genome-wide patterns of variation, Science 319 (2008) 1100-1104.

[3] G. Hellenthal, G.B. Busby, G. Band, J.F. Wilson, C. Capelli, D. Falush, & S. Myers, A genetic atlas of human admixture history, Science 343 (2014) 747-751.

[4] E. Elhaik, T. Tatarinova, D. Chebotarev, I.S. Piras, C. Maria Calo, A. De Montis, M. Atzori, M. Marini, S. Tofanelli, P. Francalacci, L. Pagani, C. Tyler-Smith, Y. Xue, F. Cucca, T.G. Schurr, J.B. Gaieski, C. Melendez, M.G. Vilar, A.C. Owings, R. Gomez, R. Fujita, F.R. Santos, D. Comas, O. Balanovsky, E. Balanovska, P. Zalloua, H. Soodyall, R. Pitchappan, A. Ganeshprasad, M. Hammer, L. Matisoo-Smith, R.S. Wells, & C. Genographic, Geographic population structure analysis of worldwide human populations infers their biogeographical origins, Nat Commun 5 (2014) 3513.

[5] F. Zhang, B. Su, Y.P. Zhang, & L. Jin, Genetic studies of human diversity in East Asia, Philos Trans R Soc Lond B Biol Sci 362 (2007) 987-995.

[6] E.R. Jones, G. Gonzalez-Fortes, S. Connell, V. Siska, A. Eriksson, R. Martiniano, R.L. McLaughlin, M. Gallego Llorente, L.M. Cassidy, C. Gamba, T. Meshveliani, O. Bar-Yosef, W. Muller, A. Belfer-Cohen, Z. Matskevich, N. Jakeli, T.F. Higham, M. Currat, D. Lordkipanidze, M. Hofreiter, A. Manica, R. Pinhasi, & D.G. Bradley, Upper Palaeolithic genomes reveal deep roots of modern Eurasians, Nat Commun 6 (2015) 8912.

[7] C. Phillips, Forensic genetic analysis of bio-geographical ancestry, Forensic Sci Int Genet (2015).

[8] C. Santos, C. Phillips, F. Oldoni, J. Amigo, M. Fondevila, R. Pereira, A. Carracedo, & M.V. Lareu, Completion of a worldwide reference panel of samples for an ancestry informative Indel assay, Forensic Sci Int Genet 17 (2015) 75-80.

[9] I. Halder, M. Shriver, M. Thomas, J.R. Fernandez, & T. Frudakis, A panel of ancestry informative markers for estimating individual biogeographical ancestry and admixture from four continents: utility and applications, Hum Mutat 29 (2008) 648-658.

[10] R. Kosoy, R. Nassir, C. Tian, P.A. White, L.M. Butler, G. Silva, R. Kittles, M.E. Alarcon-Riquelme, P.K. Gregersen, J.W. Belmont, F.M. De La Vega, & M.F. Seldin, Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America, Hum Mutat 30 (2009) 69-78.

[11] C.M. Nievergelt, A.X. Maihofer, T. Shekhtman, O. Libiger, X. Wang, K.K. Kidd, & J.R. Kidd, Inference of human continental origin and admixture proportions using a highly discriminative ancestry informative 41-SNP panel, Investig Genet 4 (2013) 13.

[12] K.B. Gettings, R. Lai, J.L. Johnson, M.A. Peck, J.A. Hart, H. Gordish-Dressman, M.S. Schanfield, & D.S. Podini, A 50-SNP assay for biogeographic ancestry and phenotype prediction in the U.S. population, Forensic Sci Int Genet 8 (2014) 101-108.

[13] K.K. Kidd, W.C. Speed, A.J. Pakstis, M.R. Furtado, R. Fang, A. Madbouly, M. Maiers, M. Middha, F.R. Friedlaender, & J.R. Kidd, Progress toward an efficient panel of SNPs for ancestry inference, Forensic Sci Int Genet 10 (2014) 23-32.

[14] A.J. Pakstis, E. Haigh, L. Cherni, A.B. ElGaaied, A. Barton, B. Evsanaa, A. Togtokh, J. Brissenden, J. Roscoe, O.

Bulbul, G. Filoglu, C. Gurkan, K.A. Meiklejohn, J.M. Robertson, C.X. Li, Y.L. Wei, H. Li, U. Soundararajan, H. Rajeevan, J.R. Kidd, & K.K. Kidd, 52 additional reference population samples for the 55 AISNP panel, Forensic Sci Int Genet 19 (2015) 269-271.

[15] J.R. Kidd, F.R. Friedlaender, W.C. Speed, A.J. Pakstis, F.M. De La Vega, & K.K. Kidd, Analyses of a set of 128 ancestry informative single-nucleotide polymorphisms in a global set of 119 population samples, Investig Genet 2 (2011) 1.

[16] O. Lao, P.M. Vallone, M.D. Coble, T.M. Diegoli, M. van Oven, K.J. van der Gaag, J. Pijpe, P. de Knijff, & M. Kayser, Evaluating self-declared ancestry of U.S. Americans with autosomal, Y-chromosomal and mitochondrial DNA, Hum Mutat 31 (2010) E1875-1893.

[17] C. Phillips, A. Freire Aradas, A.K. Kriegel, M. Fondevila, O. Bulbul, C. Santos, F. Serrulla Rech, M.D. Perez Carceles, A. Carracedo, P.M. Schneider, & M.V. Lareu, Eurasiaplex: a forensic SNP assay for differentiating European and South Asian ancestries, Forensic Sci Int Genet 7 (2013) 359-366.

[18] L.Y. U. Soundararajan, M. Shi, K.K. Kidd, Minimal SNP overlap among multiple panels of ancestry informative markers argues for more international collaboration, Forensic Science International: Genetics In Press (2016).

[19] C. Santos, M. Fondevila, D. Ballard, R. Banemann, A.M. Bento, C. Borsting, W. Branicki, F. Brisighelli, M. Burrington, T. Capal, L. Chaitanya, R. Daniel, V. Decroyer, R. England, K.B. Gettings, T.E. Gross, C. Haas, J. Harteveld, P. Hoff-Olsen, A. Hoffmann, M. Kayser, P. Kohler, A. Linacre, M. Mayr-Eduardoff, C. McGovern, N. Morling, G. O'Donnell, W. Parson, V.L. Pascali, M.J. Porto, A. Roseth, P.M. Schneider, T. Sijen, V. Stenzl, D.S. Court, J.E. Templeton, M. Turanska, P.M. Vallone, R.A. Oorschot, L. Zatkalikova, A. Carracedo, & C. Phillips, Forensic ancestry analysis with two capillary electrophoresis ancestry informative marker (AIM) panels: Results of a collaborative EDNAP exercise, Forensic Sci Int Genet 19 (2015) 56-67.

[20] C. Borsting & N. Morling, Next generation sequencing and its applications in forensic genetics, Forensic Sci Int Genet 18 (2015) 78-89.

[21] B. Keating, A.T. Bansal, S. Walsh, J. Millman, J. Newman, K. Kidd, B. Budowle, A. Eisenberg, J. Donfack, P. Gasparini, Z. Budimlija, A.K. Henders, H. Chandrupatla, D.L. Duffy, S.D. Gordon, P. Hysi, F. Liu, S.E. Medland, L. Rubin, N.G. Martin, T.D. Spector, & M. Kayser, First all-in-one diagnostic tool for DNA intelligence: genome-wide inference of biogeographic ancestry, appearance, relatedness, and sex with the Identitas v1 Forensic Chip, International Journal of Legal Medicine 127 (2013) 559-572.

[22] D.H. Warshauer, C.P. Davis, C. Holt, Y. Han, P. Walichiewicz, T. Richardson, K. Stephens, A. Jager, J. King, & B. Budowle, Massively parallel sequencing of forensically relevant single nucleotide polymorphisms using TruSeq forensic amplicon, Int J Legal Med 129 (2015) 31-36.

[23] J.Y. Chu, W. Huang, S.Q. Kuang, J.M. Wang, J.J. Xu, Z.T. Chu, Z.Q. Yang, K.Q. Lin, P. Li, M. Wu, Z.C. Geng, C.C. Tan, R.F. Du, & L. Jin, Genetic relationship of populations in China, Proc Natl Acad Sci U S A 95 (1998) 11763-11768.

[24] J.J. Kim, P. Verdu, A.J. Pakstis, W.C. Speed, J.R. Kidd, & K.K. Kidd, Use of autosomal loci for clustering individuals and populations of East Asian origin, Hum Genet 117 (2005) 511-519.

[25] P.A. Underhill, P. Shen, A.A. Lin, L. Jin, G. Passarino, W.H. Yang, E. Kauffman, B. Bonne-Tamir, J. Bertranpetit, P. Francalacci, M. Ibrahim, T. Jenkins, J.R. Kidd, S.Q. Mehdi, M.T. Seielstad, R.S. Wells, A. Piazza, R.W. Davis, M.W. Feldman, L.L. Cavalli-Sforza, & P.J. Oefner, Y chromosome sequence variation and the history of human populations, Nat Genet 26 (2000) 358-361.

[26] L. Quintana-Murci, O. Semino, H.J. Bandelt, G. Passarino, K. McElreavey, & A.S. Santachiara-Benerecetti, Genetic evidence of an early exit of Homo sapiens sapiens from Africa through eastern Africa, Nat Genet 23 (1999) 437-441.

[27] H. Zhong, H. Shi, X.B. Qi, Z.Y. Duan, P.P. Tan, L. Jin, B. Su, & R.Z. Ma, Extended Y chromosome investigation suggests postglacial migrations of modern humans into East Asia via the northern route, Mol Biol Evol 28 (2011) 717-727.

[28] X. Zhang, J. Kampuansai, X. Qi, S. Yan, Z. Yang, B. Serey, T. Sovannary, L. Bunnath, H.S. Aun, H. Samnom, W. Kutanan, X. Luo, S. Liao, D. Kangwanpong, L. Jin, H. Shi, & B. Su, An updated phylogeny of the human Y-chromosome lineage O2a-M95 with novel SNPs, PLoS One 9 (2014) e101020.

[29] C. Suo, H. Xu, C.C. Khor, R.T. Ong, X. Sim, J. Chen, W.T. Tay, K.S. Sim, Y.X. Zeng, X. Zhang, J. Liu, E.S. Tai, T.Y. Wong, K.S. Chia, & Y.Y. Teo, Natural positive selection and north-south genetic diversity in East Asia, Eur J Hum Genet 20 (2012) 102-110.

[30] J. Chen, H. Zheng, J.X. Bei, L. Sun, W.H. Jia, T. Li, F. Zhang, M. Seielstad, Y.X. Zeng, X. Zhang, & J. Liu, Genetic structure of the Han Chinese population revealed by genome-wide SNP variation, Am J Hum Genet 85 (2009) 775-785.

[31] S. Xu, X. Yin, S. Li, W. Jin, H. Lou, L. Yang, X. Gong, H. Wang, Y. Shen, X. Pan, Y. He, Y. Yang, Y. Wang, W. Fu, Y. An, J. Wang, J. Tan, J. Qian, X. Chen, X. Zhang, Y. Sun, B. Wu, & L. Jin, Genomic dissection of population substructure of Han Chinese and its implication in association studies, Am J Hum Genet 85 (2009) 762-774.

[32] S. Xu, Human population admixture in Asia, Genomics Inform 10 (2012) 133-144.

[33] J.E. Brissenden, Mongolians in the Genetic Landscape of Central Asia: Exploring the Genetic Relations among Mongolians and Other World Populations, Human Biology (2015).

[34] J.K. Pritchard, M. Stephens, & P. Donnelly, Inference of population structure using multilocus genotype data, Genetics 155 (2000) 945-959.

[35] N.M. Kopelman, J. Mayzel, M. Jakobsson, N.A. Rosenberg, & I. Mayrose, Clumpak: a program for identifying clustering modes and packaging population structure inferences across K, Mol Ecol Resour 15 (2015) 1179-1191.

[36] D.A. Earl, vonHoldt, Bridgett M, STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method, Conservation Genetics Resources 4 (2012) 359-361.

[37] J. Jia, Y.L. Wei, C.J. Qin, L. Hu, L.H. Wan, & C.X. Li, Developing a novel panel of genome-wide ancestry informative markers for bio-geographical ancestry estimates, Forensic Sci Int Genet 8 (2014) 187-194.

[38] Y.L. Wei, L. Wei, L. Zhao, Q.F. Sun, L. Jiang, T. Zhang, H.B. Liu, J.G. Chen, J. Ye, L. Hu, & C.X. Li, A single-tube 27-plex SNP assay for estimating individual ancestry and admixture from three continents, Int J Legal Med 130 (2016) 27-37.

[39] H. Rajeevan, U. Soundararajan, A.J. Pakstis, & K.K. Kidd, Introducing the Forensic Research/Reference on Genetics knowledge base, FROG-kb, Investig Genet 3 (2012) 18.

[40] H. Li, S. Gu, X. Cai, W.C. Speed, A.J. Pakstis, E.I. Golub, J.R. Kidd, & K.K. Kidd, Ethnic related selection for an ADH Class I variant within East Asia, PLoS One 3 (2008) e1881.

Figure Legends

Fig. 1. STRUCTURE analysis of 74 AISNPs for 72 combined reference and test populations. 3812 individuals are assigned to ten distinct clusters in the highest likelihood run among 20 runs at K=10. Ln P(D)= -250078.2.

Fig. 2. Heatmap of the 74 AISNPs. Heatmap representation of the 61 population allele frequencies and 74 SNPs simultaneously.

Fig. 3. (A) Principal Component Analysis of 72 global populations based on allele frequencies of the 74 AISNPs. Factor 1 and Factor 2 account for 62.26% of the total variance. Factor 3 accounts for 14.36% of the total variance. Factor 1 is primarily defined at the extremes by populations from Europe and America; Factor 2 separates African populations from the rest of the world; and Factor 3 distinguishes Native Americans and East Asians. (B) Principal Component Analysis of 24 populations based on the allele frequencies of the 74 AISNPs in populations from East Asia, North Asia and Southeast Asia populations. Factor 1 (38.52% of the variation) separates North Asians are from others, and Factor 2 (20.86% of the variation) separates Southeast Asian from East Asian. Factor 3 (8.62% of the variation) separates the Ami and Atayal populations from the other Southeast Asians.

Figure 1

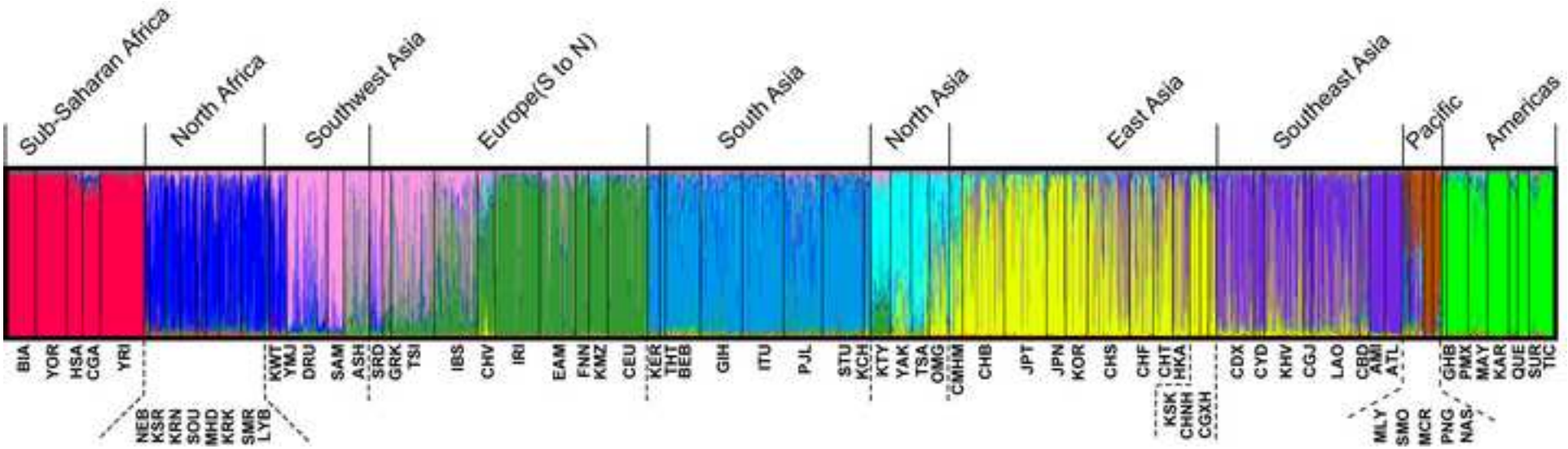

Sub-Saharan Africa
North Africa
Southwest Asia
Europe(S to N)
South Asia
North Asia
East Asia
Southeast Asia
Pacific
Americas
BIA
YOR
HSA
CGA
YRI
NEB
KSR
KRN
SOU
MHD
KRK
SMR
LYB
KWT
YMJ
DRU
SAM
ASH
SRD
GRK
TSI
IBS
CHV
IRI
EAM
FNN
KMZ
CEU
KER
THT
BEB
GIH
ITU
PJL
STU
KCH
KTY
YAK
TSA
OMG
CMHM
CHB
JPT
JPN
KOR
CHS
CHF
KSK
CHT
CHNH
HKA
CGXH
CDX
CYD
KHV
CGJ
LAO
CBD
AMI
MLY
ATL
SMO
MCR
PNG
GHB
PMX
MAY
KAR
QUE
SUR
TIC

Figure 2

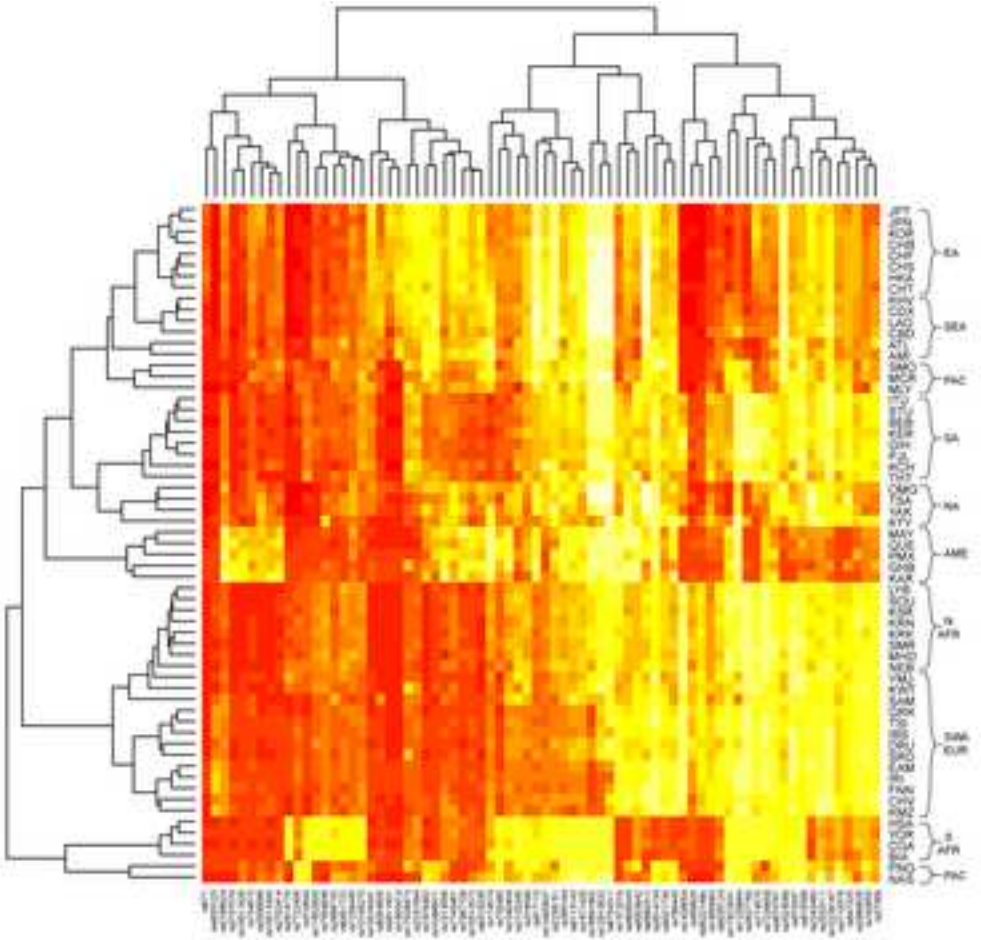
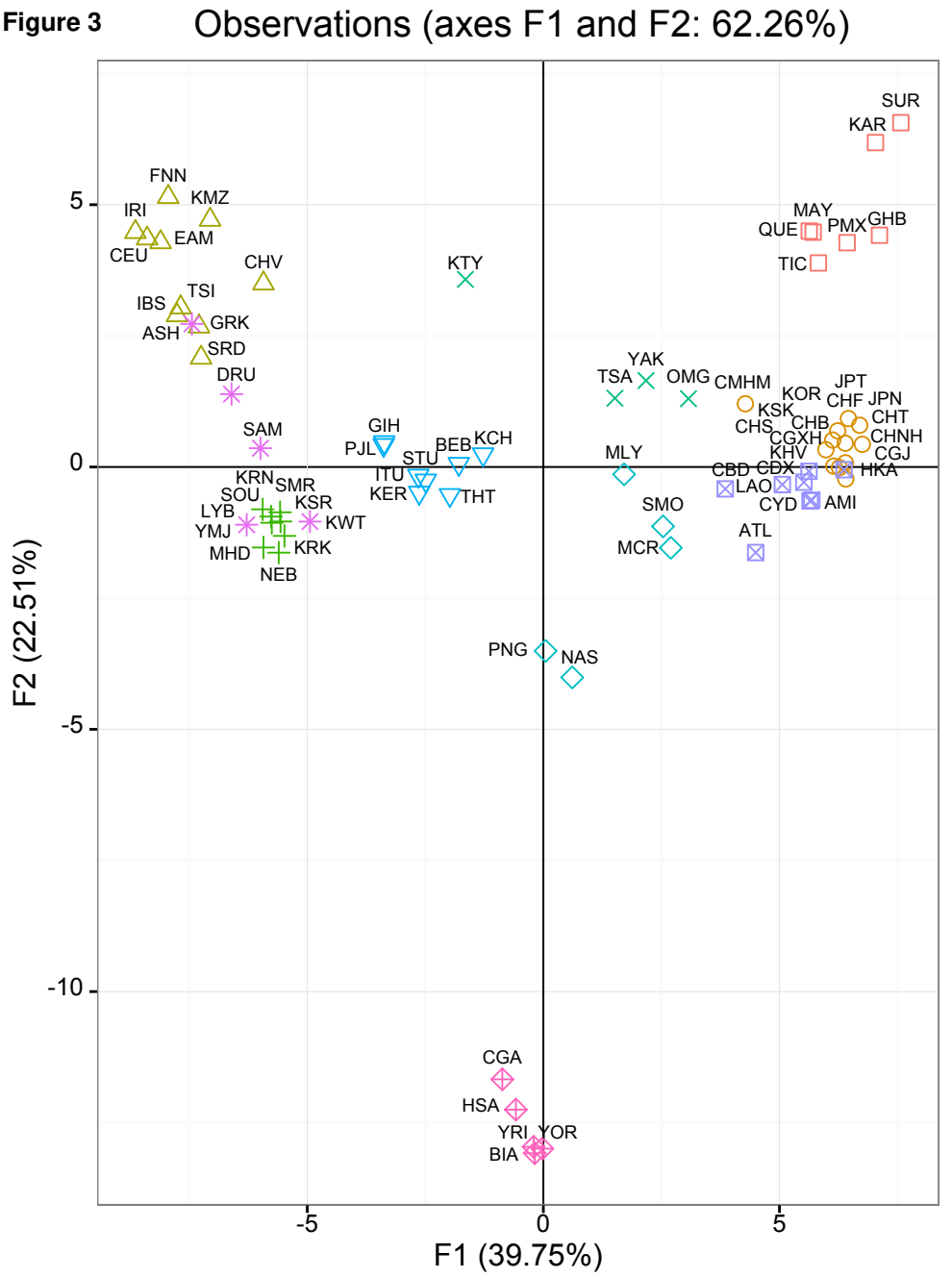
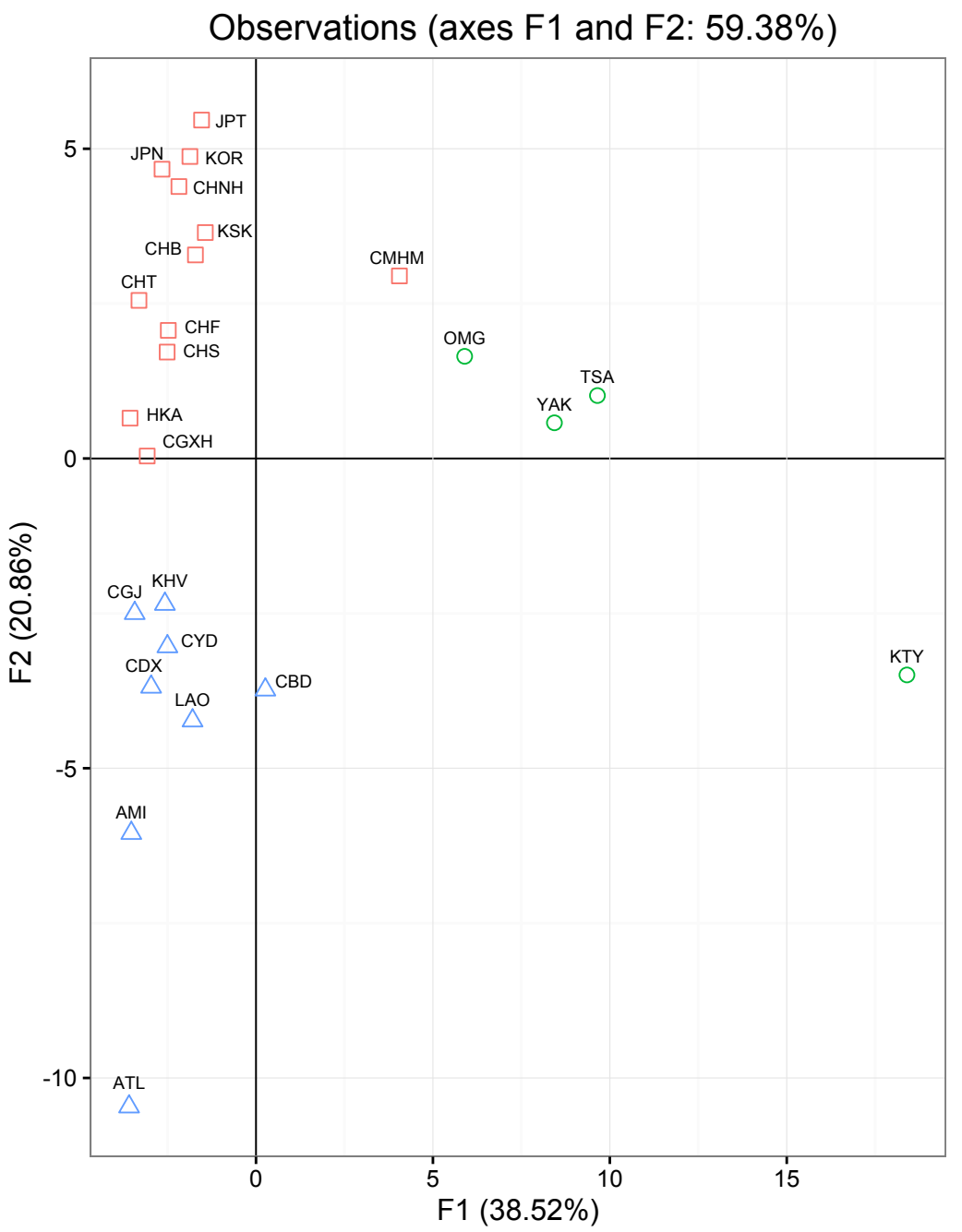
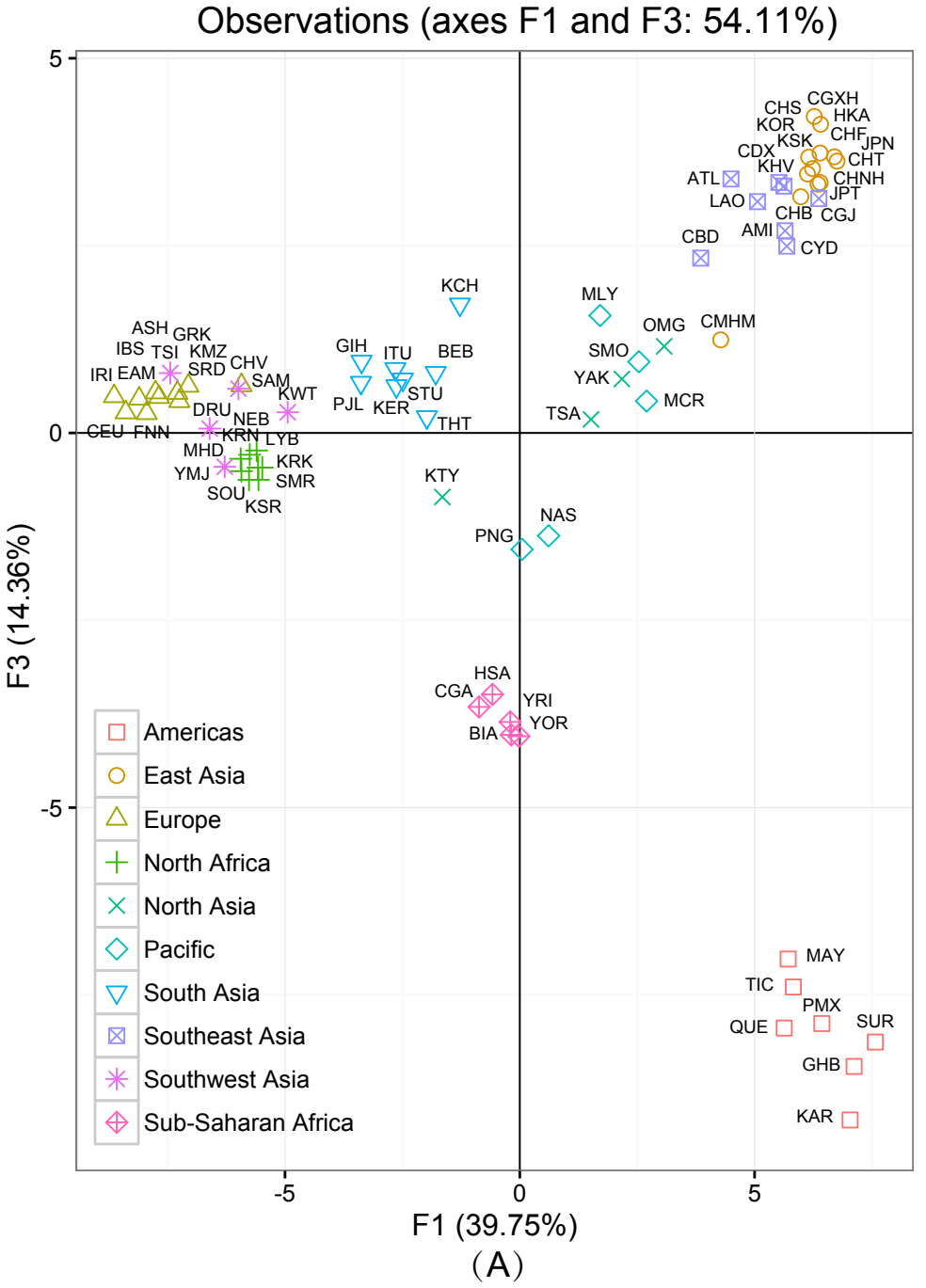


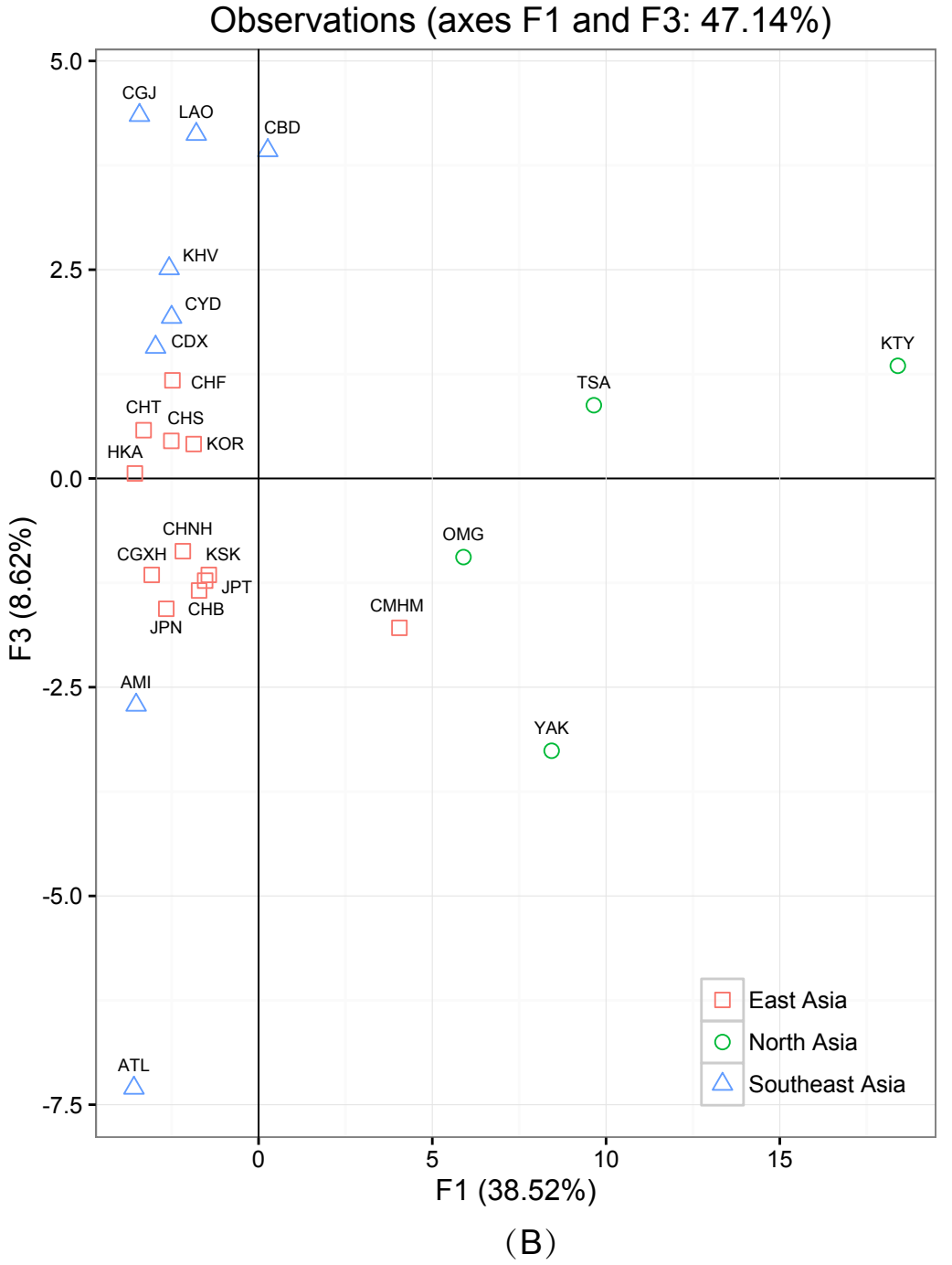
Figure 3


Observations (axes F1 and F2: 62.26%)
F1 (39.75%)
F2 (22.51%)
SUR
KAR
FNN
KMZ
IRI
EAM
CEU
CHV
KTY
MAY
QUE
PMX
GHB
TIC
TSI
IBS
GRK
ASH
SRD
DRU
YAK
TSA
OMG
CMHM
JPT
KOR
CHF
JPN
KSK
CHS
CHB
CHT
CGXH
CHNH
KHV
CGJ
CBD
CDX
HKA
LAO
CYD
AMI
SAM
GIH
PJL
BEB
KCH
STU
MLY
ITU
KER
THT
KRN
SMR
SOU
KSR
LYB
KWT
YMJ
MHD
KRK
NEB
SMO
ATL
MCR
PNG
NAS
CGA
HSA
YRI
YOR
BIA


Observations (axes F1 and F2: 59.38%)
F1 (38.52%)
F2 (20.86%)
JPT
JPN
KOR
CHNH
KSK
CHB
CMHM
CHT
CHF
CHS
OMG
TSA
YAK
HKA
CGXH
CGJ
KHV
CYD
CDX
CBD
LAO
AMI
ATL
KTY


Observations (axes F1 and F3: 54.11%)
F1 (39.75%)
F3 (14.36%)
CHS
CGXH
KOR
HKA
CHF
JPN
KSK
CDX
CHT
KHV
CHNH
ATL
JPT
LAO
CHB
CGJ
AMI
CBD
CYD
KCH
MLY
ASH
GRK
OMG
CMHM
IBS
TSI
KMZ
GIH
ITU
BEB
SMO
IRI
EAM
SRD
CHV
SAM
YAK
KWT
PJL
KER
STU
MCR
DRU
NEB
TSA
THT
CEU
FNN
KRN
LYB
MHD
KRK
YMJ
SMR
KTY
SOU
KSR
NAS
PNG
HSA
CGA
YRI
BIA
YOR
Americas
East Asia
Europe
North Africa
North Asia
Pacific
South Asia
Southeast Asia
Southwest Asia
Sub-Saharan Africa
MAY
TIC
PMX
QUE
SUR
GHB
KAR

(A)


Observations (axes F1 and F3: 47.14%)
F1 (38.52%)
F3 (8.62%)
CGJ
LAO
CBD
KHV
CYD
CDX
KTY
CHF
TSA
CHT
CHS
KOR
HKA
CHNH
OMG
CGXH
KSK
JPT
CHB
CMHM
JPN
AMI
YAK
East Asia
North Asia
Southeast Asia
ATL

(B)

Table 1. The 61 reference population samples and 11 test population samples

	World Region[a]	Population	Abbr.	Sample Size (N)	Source
1	Sub-Saharan Africa (SAFR)	Biaka	BIA	67	Kidd lab
2		Yoruba	YOR	77	Kidd lab
3		Hausa	HSA	39	Kidd lab
4		Chagga	CGA	45	Kidd lab
5		***Yoruba in Ibadan, Nigeria***	***YRI***	***108***	***1000 Genomes***
6	North Africa (NAFR)	Nebeur	NEB	13	Kidd Lab
7		Kesra	KSR	39	Kidd Lab
8		Kairoun	KRN	37	Kidd Lab
9		Sousse	SOU	40	Kidd Lab
10		Mehdia	MHD	23	Kidd Lab
11		Kerkennah*	KRK	34	Kidd Lab
12		Smar	SMR	53	Kidd Lab
13		Lybia	LYB	57	Kidd Lab
14	Europe (EUR)	Toscani in Italia	TSI	107	1000 Genomes
15		Iberian populations in Spain	IBS	107	1000 Genomes
16		Sardinians	SRD	33	Kidd Lab
17		Greeks	GRK	19	Kidd Lab
18		Chuvash	CHV	42	Kidd lab
19		Irish	IRI	111	Kidd lab
20		EuroAmerican	EAM	88	Kidd lab
21		Finns	FNN	34	Kidd lab
22		Komi Zyriane	KMZ	46	Kidd lab
23		***Utah residents with European ancestry***	***CEU***	***99***	***1000 Genomes***
24	Southwest Asia (SWA)	Druze	DRU	102	Kidd lab
25		Kuwaiti	KWT	14	Kidd lab
26		Samaritans	SAM	39	Kidd lab
27		Yemenite Jews	YMJ	41	Kidd lab
28		***Ashkenazi Jews***	***ASH***	***64***	***Kidd lab***
29	South Asia (SA)	Keralites,S.India	KER	30	Kidd lab
30		Thoti	THT	14	Kidd lab
31		Bengali in Bangladesh	BEB	86	1000 Genomes

32		Gujarati Indian in Houston,TX	GIH	103	1000 Genomes
33		Indian Telugu in the UK	ITU	102	1000 Genomes
34		Punjabi in Lahore,Pakistan	PJL	96	1000 Genomes
35		Sri Lankan Tamil in the UK	STU	102	1000 Genomes
36		Kachari from Assam	KCH	17	Kidd lab
37	North Asia (NA)	Khanty	KTY	49	Kidd lab
38		Yakut	YAK	51	Kidd lab
39		Tsaatan	TSA	44	Kidd lab
40		Out Mongol	OMG	50	Kidd lab
41	East Asia (EA)	***Inner Mongolia in Hailar, China***	***CMHM***	***33***	***Caixia lab***
42		Han Chinese in Bejing, China	CHB	103	1000 Genomes
43		Japanese in Tokyo, Japan	JPT	104	1000 Genomes
44		Japanese	JPN	47	Kidd lab
45		Koreans	KOR	54	Kidd lab
46		Southern Han Chinese, China	CHS	105	1000 Genomes
47		Chinese, S.F.	CHF	57	Kidd lab
48		Chinese, Taiwan	CHT	49	Kidd lab
49		Hakka	HKA	41	Kidd lab
50		***Koreans, South Korea***	***KSK***	***26***	***Caixia lab***
51		***Han Chinese in Henan, China***	***CHNH***	***21***	***Caixia lab***
52		***Han Chinese in Guangxi , China***	***CGXH***	***20***	***Caixia lab***
53	Southeast Asia (SEA)	Chinese Dai in Xishuangbanna, Yunnan, China	CDX	93	1000 Genomes
54		***Dai in Xishuangbanna, Yunnan, China***	***CYD***	***24***	***Caixia lab***
55		Kinh in Ho Chi Minh City, Vietnam	KHV	99	1000 Genomes

56		***Kinh in Guangxi, China***	***CGJ***	***18***	***Caixia lab***
57		Laotian	LAO	118	Kidd lab
58		Cambodian	CBD	24	Kidd lab
59		Ami	AMI	40	Kidd lab
60		Atayal	ATL	42	Kidd lab
61	Pacific (PAC)	Malaysian	MLY	10	Kidd lab
62		Samoan	SMO	9	Kidd lab
63		Micronesians	MCR	34	Kidd lab
64		Papua-New Guineans	PNG	22	Kidd lab
65		Nasioi Melanesians	NAS	22	Kidd lab
66	Americas (AME)	Guihiba speakers	GHB	12	Kidd lab
67		Pima, Mexico	PMX	53	Kidd lab
68		Maya, Yucatan	MAY	48	Kidd lab
69		Karitiana	KAR	53	Kidd lab
70		Quechua	QUE	22	Kidd lab
71		***Rondonian Surui***	***SUR***	***27***	***Kidd lab***
72		***Ticuna***	***TIC***	***60***	***Kidd lab***

a. Validation step test populations are highlighted in bold and italic, others are reference populations

Table 2. Details of the 74 AISNPs

dbSNP rs#	Chr	Build 37 nt position	Fst 61-Pop	δ_R[a]										δ_P[a]		
				SAFR	NAFR	SWA	EUR	SA	NA	EA	SEA	PAC	AME	EA-SEA	EA-NA	EUR-SWA
rs3827760	2	109513601	0.72	0.40	0.40	0.40	0.40	0.40	0.49	0.59	0.57	0.38	0.71	0.02	0.09	0.00
rs1426654	15	48426484	0.72	0.44	0.61	0.64	0.64	0.30	0.04	0.47	0.46	0.40	0.47	0.00	0.46	0.00
rs2814778	1	159174683	0.64	0.97	0.11	0.09	0.14	0.14	0.14	0.14	0.14	0.13	0.14	0.00	0.00	0.05
rs16891982	5	33951693	0.55	0.23	0.07	0.48	0.85	0.19	0.05	0.23	0.23	0.23	0.23	0.00	0.16	0.33
rs1800414	15	28197037	0.54	0.14	0.14	0.14	0.14	0.14	0.12	0.57	0.53	0.14	0.14	0.04	0.63	0.00
rs174570	11	61597212	0.49	0.43	0.33	0.38	0.28	0.38	0.36	0.01	0.54	0.23	0.68	0.50	0.33	0.10
rs9522149	13	111827167	0.47	0.28	0.53	0.60	0.53	0.08	0.16	0.29	0.28	0.29	0.28	0.01	0.11	0.07
rs3737576	1	101709563	0.44	0.14	0.13	0.11	0.06	0.10	0.10	0.04	0.12	0.14	0.73	0.07	0.12	0.04
rs3811801	4	100244319	0.44	0.10	0.10	0.10	0.10	0.10	0.09	0.70	0.12	0.10	0.10	0.52	0.71	0.00
rs1876482	2	17362568	0.44	0.40	0.40	0.35	0.33	0.15	0.26	0.45	0.48	0.39	0.04	0.03	0.17	0.02
rs1229984	4	100239319	0.43	0.21	0.09	0.08	0.20	0.21	0.16	0.70	0.45	0.14	0.21	0.22	0.77	0.25
rs17822931	16	48258198	0.42	0.40	0.31	0.36	0.23	0.16	0.66	0.60	0.28	0.34	0.06	0.29	0.05	0.12
rs310644	20	62159504	0.42	0.75	0.03	0.25	0.28	0.17	0.25	0.28	0.26	0.67	0.24	0.02	0.02	0.02
rs12913832	15	28365618	0.41	0.15	0.02	0.21	0.65	0.09	0.09	0.15	0.15	0.15	0.05	0.00	0.06	0.39
rs7226659	18	40488279	0.38	0.35	0.35	0.36	0.35	0.24	0.38	0.24	0.42	0.59	0.02	0.16	0.13	0.01
rs917115	7	28172586	0.37	0.20	0.37	0.43	0.49	0.12	0.08	0.38	0.37	0.13	0.24	0.01	0.27	0.06
rs9319336	13	27624356	0.37	0.24	0.33	0.33	0.33	0.14	0.08	0.43	0.42	0.04	0.39	0.01	0.32	0.00
rs10496971	2	145769943	0.37	0.37	0.27	0.35	0.35	0.04	0.04	0.46	0.54	0.25	0.08	0.07	0.38	0.00
rs7997709	13	34847737	0.35	0.19	0.24	0.35	0.32	0.13	0.24	0.48	0.26	0.35	0.61	0.20	0.22	0.03

rs10516441	4	100307167	0.34	0.24	0.18	0.23	0.19	0.18	0.03	0.69	0.25	0.07	0.19	0.39	0.64	0.04
rs870347	5	6845035	0.33	0.21	0.23	0.25	0.23	0.10	0.11	0.28	0.18	0.24	0.68	0.09	0.16	0.02
rs1572018	13	41715282	0.33	0.58	0.33	0.28	0.37	0.16	0.07	0.20	0.26	0.54	0.51	0.06	0.12	0.08
rs6754311	2	136707982	0.32	0.10	0.04	0.07	0.58	0.01	0.05	0.10	0.10	0.10	0.10	0.00	0.04	0.58
rs4833103	4	38815502	0.32	0.10	0.02	0.14	0.53	0.10	0.08	0.10	0.10	0.10	0.10	0.00	0.02	0.35
rs6990312	8	110602317	0.30	0.53	0.05	0.11	0.20	0.06	0.31	0.34	0.33	0.59	0.05	0.01	0.02	0.08
rs2024566	22	41697338	0.30	0.23	0.02	0.03	0.06	0.12	0.00	0.25	0.25	0.25	0.80	0.00	0.22	0.03
rs459920	16	89730827	0.30	0.09	0.36	0.55	0.22	0.05	0.12	0.23	0.24	0.24	0.16	0.01	0.10	0.30
rs3814134	9	127267689	0.29	0.80	0.03	0.20	0.24	0.12	0.13	0.06	0.16	0.08	0.22	0.09	0.17	0.04
rs2238151	12	112211833	0.28	0.38	0.11	0.33	0.43	0.06	0.12	0.33	0.32	0.25	0.09	0.01	0.18	0.09
rs10236187	7	139447377	0.28	0.35	0.17	0.29	0.29	0.27	0.00	0.03	0.14	0.20	0.63	0.10	0.03	0.00
rs174574	11	61600342	0.27	0.13	0.07	0.27	0.23	0.50	0.18	0.18	0.37	0.05	0.50	0.49	0.33	0.04
rs735480	15	45152371	0.27	0.63	0.22	0.38	0.38	0.24	0.06	0.02	0.05	0.45	0.25	0.03	0.03	0.00
rs6054605	20	744570	0.26	0.04	0.15	0.00	0.09	0.08	0.11	0.30	0.48	0.14	0.15	0.16	0.37	0.08
rs2006996	9	117592638	0.25	0.21	0.24	0.23	0.20	0.15	0.41	0.18	0.11	0.07	0.48	0.26	0.21	0.03
rs192655	6	90518278	0.25	0.04	0.21	0.37	0.32	0.15	0.22	0.34	0.28	0.21	0.39	0.05	0.10	0.05
rs2899826	15	74734500	0.25	0.09	0.35	0.38	0.38	0.07	0.10	0.01	0.19	0.37	0.42	0.16	0.09	0.00
rs8035124	15	92105708	0.24	0.12	0.33	0.36	0.49	0.16	0.22	0.23	0.19	0.27	0.30	0.04	0.01	0.12
rs2242480	7	99361466	0.24	0.62	0.11	0.26	0.31	0.04	0.33	0.11	0.05	0.04	0.37	0.14	0.19	0.05
rs7745461	6	21911616	0.24	0.06	0.08	0.35	0.31	0.13	0.38	0.46	0.26	0.42	0.25	0.18	0.07	0.03
rs7554936	1	151122489	0.24	0.76	0.19	0.05	0.02	0.11	0.12	0.20	0.15	0.34	0.25	0.05	0.29	0.06
rs2702414	4	179399523	0.24	0.25	0.21	0.19	0.21	0.13	0.05	0.18	0.08	0.11	0.56	0.09	0.12	0.01

rs37369	5	35037115	0.23	0.18	0.32	0.39	0.34	0.17	0.16	0.25	0.13	0.15	0.33	0.11	0.08	0.04
rs10512572	17	69512099	0.23	0.13	0.20	0.22	0.21	0.15	0.04	0.31	0.35	0.16	0.44	0.03	0.32	0.01
rs17028973	4	100322786	0.23	0.10	0.08	0.03	0.21	0.24	0.25	0.54	0.24	0.02	0.35	0.27	0.71	0.16
rs13400937	2	79864923	0.23	0.40	0.23	0.32	0.30	0.13	0.10	0.20	0.34	0.07	0.40	0.13	0.09	0.02
rs4670767	2	37941396	0.23	0.30	0.22	0.20	0.22	0.03	0.06	0.31	0.40	0.07	0.14	0.07	0.23	0.01
rs2241894	4	100266133	0.23	0.14	0.13	0.12	0.22	0.29	0.24	0.58	0.28	0.23	0.36	0.27	0.74	0.09
rs1950993	14	58238687	0.23	0.34	0.11	0.28	0.38	0.30	0.21	0.23	0.23	0.30	0.18	0.00	0.39	0.09
rs11652805	17	62987151	0.23	0.80	0.11	0.04	0.13	0.08	0.24	0.03	0.03	0.30	0.21	0.06	0.19	0.16
rs818386	16	65406708	0.23	0.25	0.18	0.16	0.04	0.16	0.20	0.21	0.08	0.25	0.54	0.12	0.01	0.10
rs7722456	5	170202984	0.22	0.16	0.05	0.02	0.02	0.12	0.21	0.20	0.19	0.74	0.22	0.01	0.01	0.03
rs2166624	13	42579985	0.22	0.40	0.13	0.04	0.01	0.01	0.13	0.10	0.08	0.39	0.71	0.15	0.03	0.03
rs4908343	1	27931698	0.22	0.64	0.03	0.07	0.27	0.05	0.35	0.05	0.15	0.20	0.33	0.18	0.28	0.19
rs10511828	9	28628500	0.22	0.09	0.15	0.13	0.14	0.07	0.03	0.02	0.09	0.08	0.61	0.10	0.05	0.01
rs671	12	112241766	0.21	0.04	0.04	0.04	0.04	0.04	0.04	0.23	0.07	0.04	0.04	0.14	0.24	0.00
rs10513300	9	120130206	0.21	0.22	0.19	0.17	0.14	0.05	0.13	0.05	0.14	0.08	0.53	0.08	0.07	0.03
rs6451722	5	43711378	0.20	0.67	0.13	0.07	0.12	0.14	0.05	0.15	0.03	0.01	0.36	0.11	0.18	0.05
rs647325	1	18170886	0.20	0.34	0.20	0.22	0.16	0.28	0.26	0.09	0.07	0.01	0.62	0.01	0.31	0.05
rs8003942	14	105971670	0.18	0.24	0.18	0.09	0.10	0.18	0.02	0.07	0.45	0.13	0.37	0.34	0.05	0.00
rs2986742	1	6550376	0.18	0.64	0.08	0.04	0.20	0.08	0.03	0.06	0.06	0.19	0.27	0.01	0.03	0.15
rs2125345	17	73782191	0.18	0.53	0.08	0.05	0.29	0.19	0.16	0.01	0.04	0.11	0.39	0.03	0.16	0.30
rs3118378	1	68849687	0.17	0.06	0.15	0.12	0.11	0.15	0.13	0.16	0.17	0.37	0.63	0.01	0.26	0.01
rs8113143	19	33652247	0.17	0.33	0.10	0.18	0.48	0.01	0.08	0.07	0.01	0.32	0.13	0.07	0.13	0.26

rs734873	3	147750355	0.17	0.23	0.18	0.12	0.17	0.04	0.01	0.26	0.15	0.18	0.43	0.10	0.22	0.04
rs798443	2	7968275	0.17	0.66	0.11	0.01	0.21	0.00	0.02	0.04	0.03	0.28	0.24	0.06	0.05	0.18
rs1513056	12	17407792	0.17	0.08	0.27	0.29	0.25	0.09	0.18	0.19	0.05	0.05	0.52	0.12	0.00	0.04
rs10108270	8	4190793	0.16	0.62	0.03	0.21	0.08	0.11	0.09	0.12	0.05	0.11	0.42	0.15	0.02	0.11
rs1040404	1	168159890	0.16	0.45	0.02	0.06	0.19	0.30	0.03	0.07	0.11	0.13	0.35	0.03	0.09	0.12
rs385194	4	85309078	0.16	0.38	0.14	0.21	0.46	0.01	0.00	0.08	0.11	0.33	0.31	0.03	0.07	0.22
rs316598	5	2364626	0.15	0.50	0.03	0.16	0.33	0.06	0.27	0.03	0.11	0.36	0.21	0.07	0.27	0.15
rs1871428	6	168665760	0.15	0.33	0.08	0.15	0.39	0.09	0.26	0.13	0.13	0.28	0.11	0.00	0.35	0.22
rs2033111	17	53788280	0.14	0.19	0.06	0.20	0.25	0.16	0.11	0.07	0.18	0.04	0.37	0.10	0.17	0.04
rs8021730	14	67886781	0.13	0.47	0.07	0.14	0.31	0.02	0.07	0.11	0.01	0.16	0.19	0.10	0.16	0.15
rs7238445	18	49781544	0.12	0.58	0.01	0.19	0.20	0.09	0.15	0.02	0.12	0.16	0.00	0.13	0.12	0.00

a. Estimated allele frequencies in each cluster were extracted from the result file of the highest likelihood STRUCTURE of 178 SNPs on 61 reference populations at K=10 in 30 runs, based on which δ^R and δ^P are calculated.

Table 3. The first likelihood ratio geographic region for the 500 test individuals assayed by 74 AISNPs panel

Test population	Individual numbers that falls into the first likelihood ratio geographic region										Sample size
	SAFR	**NAFR**	**EUR**	**SWA**	**SA**	**NA**	**EA**	**SEA**	**PAC**	**AME**	
YRI	**108(100%)**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**108**
CEU	**0**	**0**	**99(100%)**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**99**
ASH	**0**	**3(5%)**	**45(70%)**	**16(25%)**	**0**	**0**	**0**	**0**	**0**	**0**	**64**
KSK	**0**	**0**	**0**	**0**	**0**	**1(4%)**	**23(88%)**	**2(8%)**	**0**	**0**	**26**
CMHM	**0**	**0**	**0**	**0**	**0**	**17(52%)**	**14(42%)**	**2(6%)**	**0**	**0**	**33**
CHNH	**0**	**0**	**0**	**0**	**0**	**0**	**21(100%)**	**0**	**0**	**0**	**21**
CGXH	**0**	**0**	**0**	**0**	**0**	**0**	**15(75%)**	**5(25%)**	**0**	**0**	**20**
CYD	**0**	**0**	**0**	**0**	**0**	**0**	**2(8%)**	**22(92%)**	**0**	**0**	**24**
CGJ	**0**	**0**	**0**	**0**	**0**	**0**	**4(22%)**	**14(78%)**	**0**	**0**	**18**
SUR	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**27(100%)**	**27**
TIC	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**60(100%)**	**60**