1 AIM-SNPtag: a computationally efficient approach for developing ancestry-informative SNP
2 panels

3

4 Shilei Zhao^{a,1}, Cheng-Min Shi^{a,1}, Liang Ma^{a,1}, Qi Liu^{a,c,1}, Yongming Liu^{a,c}, Fuquan Wu^{a,c},
 5 Lianjiang Chi^a, Hua Chen^{a,b,c,*}

6

7 ^a CAS Key Laboratory of Genomic and Precision Medicine, Beijing Institute of Genomics,
 8 Chinese Academy of Sciences, Beijing 100101, China

9 ^b Collaborative Innovation Center of Forensic Genomics, Beijing Institute of Genomics, Chinese
 10 Academy of Sciences, Beijing 100101, China

11 ^c University of Chinese Academy of Sciences, Beijing 100049, China

12

13 ¹These authors contributed equally to this study.

14

15 *Corresponding author. Tel.: +86 10 84097483 (H. Chen).

16 Email address: chenh@big.ac.cn (H. Chen).

17 Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China

18

ABSTRACT

Inferring an individual's ancestry or group membership using a small set of highly informative genetic markers is very useful in forensic and medical genetics. However, given the huge amount of SNP data available from a diverse of populations, it is challenging to develop informative panels by exhaustively searching for all possible SNP combination. In this study, we formulate it as an algorithm problem of selecting an optimal set of SNPs that maximizes the inference accuracy while minimizes the set size. Built on this conception, we develop a computational approach that is capable of constructing ancestry informative panels from multi-population genome-wide SNP data efficiently. We evaluate the performance of the method by comparing the panel size and membership inference accuracy of the constructed SNP panels to panels selected through empirical procedures in former studies. For the membership inference of population groups including Asian, European, African, East Asian and Southeast Asian, a 36-SNP panel developed by our approach has an overall accuracy of 99.07%, and a 21-SNP subset of the panel has an overall accuracy of 95.36%. In comparison, the existing panel requires 74 SNPs to achieve an accuracy of 94.14% on the same set of population groups. We further apply the method to four subpopulations within Europe (Finnish, British, Spain and Italia); a 175-SNP panel can discriminate individuals of those European subpopulations with an accuracy of 99.36%, of which a 68-SNP subset can achieve an accuracy of 95.07%. We expect our method to be a useful tool for constructing ancestry informative markers in forensic genetics.

Keywords: Ancestry inference, forensics, genome-wide SNPs, membership, SNP panel.

1. Introduction

One of the goals in forensic analysis is to ascertain an individual's ancestry or membership of group origin defined *a priori*, and thus to redeploy investigating efforts. In the past few decades, genetic ancestry-informative markers (AIMs) have been developed for a diverse of populations aiming to infer an individual's continental or biogeographic origins [1-17]. Recent development of genotyping and sequencing technologies has provided rich sources of genetic makers from diverse populations, enabling AIMs to be fully explored at whole-genome level. In fact, informative genetic polymorphisms have been successfully used to predict physical appearance, such as eye color [18], to infer geographical ancestry [1-17, 19, 20], as well as family of origin [21]. Forensic science is entering a new era of "DNA intelligence" [18].

Among the genetic markers, single nucleotide polymorphisms (SNPs) are the most abundant (approximately one SNP per 1000 bases) and easy to genotype [22]. Furthermore, SNPs can "recover information from degraded DNA samples better than short tandem repeats (STRs)", which also enables their potential applications in DNA-based forensic tests [16, 23]. However, in most cases, the trace DNA materials available only allow typing of a few genetic loci and the volume of DNA required for whole-genome genotyping using arrays is not plausible in a forensic investigation. Thus, it is pivotal to select panels that include a small number of highly informative SNP loci and meet the specific needs of a forensic investigation.

The promise for such a panel-based approach has already been highlighted by extensive human population genetic studies [24-27]. Many studies revealed that a substantial amount of genetic variations is shared among populations, but only a minor proportion of variations are population specific. Those population-specific SNPs are indicative of an individual's population origin or

membership, however most of these SNPs are in low frequencies, prohibiting them from being used as population tags [28-30]. Among the common SNPs shared by different populations, some SNPs that are highly differentiated by allele frequencies (*i.e.* with high F_{ST} scores), are promising candidates for forensic applications and draw a lot of attention in current studies [9, 16]. Moreover, because SNPs are in linkage disequilibria and their information on membership is partially redundant, a non-redundant set that includes a selected number of SNPs could retain the most relevant information needed for ancestry inference. Thus it is possible to condense the number of SNPs without sacrificing their discriminating power for membership inference, thereby being applicable for trace DNA materials.

Several studies have explored such a possibility by developing ancestry informative panels from available large-scale genomic data using empirical procedures *e.g.*, [9, 16, 17, 31]. An optimal SNP panel should include a minimum set of SNPs, and the combination of these SNPs can guarantee the highest discriminatory accuracy for ancestral inference. In the present study we developed a computational method satisfying such criteria (named AIM-SNPtag). The method is able to explore the genome-wide SNP data very efficiently, and construct group membership or ancestry informative SNP panels. Testing on an exemplar dataset indicated that our method generated panels that performed better at membership inference with higher inference accuracy, and included fewer SNPs than the panel developed through empirical approaches. The capacity of AIM-SNPtag was further demonstrated by applying it to the 1000 Genomes Project data to generate two continent-level and within-continent SNP panels.

2. Materials and Methods

2.1 Marker selection procedure

We aim to select a subset of SNPs from whole-genome data that are suitable for ancestry or membership inference. An optimal SNP panel should satisfy two criteria simultaneously: (1) maximize the accuracy of predicting the subject's membership or achieving a pre-chosen accuracy level, and (2) minimize the number of SNPs required for this inference. However, exhaustively searching through all possible subsets of SNPs to identify a panel that satisfies such criteria is not computationally trivial; the computational time required is an exponential function of the total number of SNPs (n) to be selected against, *i.e.* the computational complexity is $O(2^n)$. For example, to find a subset from 200 SNPs it requires exhaustively searching and evaluating over 2^{200} possible combinations, which is already impractical. Here we elaborate a method that can explore the genome-wide data more efficiently, while select informative SNP panels approximately satisfying the aforementioned criteria. The method implements a recursive algorithm that downscales the computational complexity to $O(n^2)$.

The proposed method is a forward algorithm that includes four successive steps explained below (Fig. 1) and the relevant mathematical and computational details are in the Appendix.

Step 1: partitioning. Populations of interest are not equally divergent from each other; in other words, there is inherent hierarchical structure among populations. Consequently, it is difficult to select informative markers that can discriminate populations for all hierarchical levels. Instead of selecting AIMs in a single step simultaneously for all populations, we first partition populations into multiple classes; each class includes two or several populations. The partitioning step can be proceeded with *a priori* knowledge about population structure (such as geographical distribution and/or genetic relatedness) or with a data-driven clustering algorithm such as principal component

analysis (PCA). These classes are not necessarily mutually exclusive, instead, they can be nested with each other (see section *Application to genome-wide SNP data* for examples).

Step 2: pre-filtering. When samples include massive number of markers (e.g. $>10^6$), a pre-filtering step is necessary to reduce the dimensionality according to the F_{ST} values of markers. Since F_{ST} measures the degree of differentiation of populations, we first calculate F_{ST} of each SNP over populations within the same class. Then we sort the SNPs in a descending order according their F_{ST} values. Finally, we retain the top l SNPs for each partition class from Step 1. Note that, the number l should be much smaller than the total number of SNPs but still reasonably large (say, 20,000), in order to reserve major information that distinguishes populations in the class.

Step 3: selecting. For each class, a feature selection algorithm is proposed to efficiently draw a subset of SNPs (refer to as the maximum classification SNPs, MaC-SNPs) from the pre-filtered SNPs of Step 2. The MaC-SNPs is generated by sequentially adding markers that maximize the accumulative classification ability while panelizing on the maximum normalized mutual information, which account for the effect of linkage disequilibrium, over the already added SNPs (see the Appendix for details). As we will integrate the MaC-SNP sets for all classes in step 4, we suggest to apply an even size of 100-200 SNPs for MaC-SNP sets. Indeed, the size is flexible and can be customized according to research demands.

Step 4: integrating and optimizing. The MaC-SNP sets selected from multiple classes are integrated to generate the global AIM-SNP panel. Another wrapper-based feature selection algorithm is applied to generate the global panel which either meets a pre-set threshold of the membership inference accuracy (e.g. 95%) or reaches a pre-given panel size (say, 100 SNPs). The global panel may compose varying numbers of SNPs from each of the MaC-SNP sets (see the

Appendix).

The strategies and algorithms summarized above can generate a SNP panel that approximately satisfies the former mentioned criteria, while sharply reducing the computational load, and thus making the method applicable to genome-wide SNP data.

2.2 Performance evaluation

We apply the Monte Carlo cross-validation procedure to evaluate the performance of AIM-SNPtag. The full set of labeled samples is partitioned into two subsets, with one subset being the training set which includes 60% of individuals randomly sampled from each labeled population, and the remaining 40% of data is retained as the test set. After constructing the AIM panel with the training set, we apply the Naïve Bayes classifier (NBC, see Appendix) to the test set to evaluate the performance of the AIMS, as NBC is a fast, accurate and reliable classifier. The accuracy, the proportion of correct classifications in test set, $AC = \frac{\#\{correct\ classifications\}}{\#\{test\ set\}}$ is then calculated. This procedure is repeated for 1000 times. We report the membership inference accuracy as the average accuracy (AAC) over the 1000 repeats.

We further use two pairs of measures to evaluate in more details the performance of classification on every population. They can be derived from a confusion matrix, a 2×2 contingency table which records the numbers of four possible outcomes of a binary classification (in terms of the population under investigation), true positive (TP), true negative (TN) false negative (FN), false positive (FP). The first pair of statistics are sensitivity ($= \frac{TP}{TP+FN}$) and specificity ($= \frac{TN}{FP+TN}$), which characterize the distribution of classifications given the true outcomes. The second pair of statistics are positive predictive value ($PPV = \frac{TP}{TP+FP}$) and negative predictive value ($NPV =$

$\frac{TN}{TN+FN}$), which give the probabilities of the outcomes given the classifications.

The performance of AIM-SNPtag is demonstrated by comparing with existing panels. Given two panels with the same number of SNPs, the higher AAC the better performance the panel has. For a given threshold of accuracy, panels that include fewer SNPs are considered to perform better. As a testing case, we applied AIM-SNPtag to a pre-selected set of 178 SNPs that was also used by [9] to generate a 74-SNP ancestry informative panel (74-AISNP). We generated a panel that discriminate the five major groups, Africans (AFR), Europeans (EUR), East Asians (EA), South Asians (SA), and Southeast Asians (SEA). Then the performance of the AIM-SNPtag selected panel was evaluated using the 1000 Genomes Phase 3 data and compared with the 74-AISNP panel [32].

The informativeness of the panels was also graphically illustrated using the PCA and Bayesian clustering analyses. The PCA analysis was performed using *smartpca* in EIGENSOFT 6.1.4 [33]. The Bayesian method was implemented in STRUCTURE 2.3.4 [34], which assigns individuals into discrete clusters of similar genetic profiles. Ten replicates were run for different fixed number of clusters, K , ranging from two to five ($K = 2 \sim 5$). For each run, 1×10^6 iterations were carried out after a burn-in period of 1×10^5 iterations.

2.3 Application to genome-wide SNP data

To further demonstrate its capability, we used AIM-SNPtag to select membership informative SNPs for major world population groups with dense genome-wide SNP data from the 1000 Genomes Project [32]. The five major groups are Africans (AFR, $n = 108$), Europeans (EUR, $n = 313$), East Asians (EA, $n = 312$), South Asians (SA, $n = 489$), and Southeast Asians (SEA, $n = 192$). A total of ~ 78 million SNPs were collected through whole genome re-sequencing. According to the

partition and integration procedure described earlier, we first pre-filtered the SNP pool by retaining 20,000 SNPs for each of the following 3 classes, AFR-EUR-SA-EA/SEA, EUR-SA-EA/SEA and EA-SEA, based on their F_{ST} scores. From each of the three thinned SNP subsets, AIM-SNPtag exacted a MaC-SNP set with 100 SNPs. These candidate MaC-SNP sets were combined and integrated into global panels that satisfy the thresholds of AAC $\geq 95\%$ and $\geq 99\%$, respectively.

We also used AIM-SNPtag to select SNP panels that discriminate closely-related subpopulations within Europe. Four local subpopulations, Finnish (FIN, $n = 99$), British (GBR, $n = 91$), Spain (IBS, $n = 107$) and Italia (TSI, $n = 107$), in the 1000 Genomes Project data were analyzed. From the pre-filtered 20,000 SNPs for each of the following 3 classes, FIN-GBR-IBS/TSI, GBR-IBS/TSI and IBS-TSI, we used AIM-SNPtag to select three MaC-SNP sets with 200 SNPs each. These three candidate MaC-SNP sets were then integrated into global panels with AAC $\geq 95\%$ and $\geq 99\%$, respectively.

3. Results

3.1 Performance evaluation

From the pre-selected 178 candidate SNPs of the test data set [9], we used AIM-SNPtag to generate an informative panel that discriminates the five major population groups (AFR, EUR, SA, EA and SEA). The average accuracy (AAC) were plotted as a function of the panel size in Figure 2A. The first 18 SNPs selected by AIM-SNPtag discriminate the five major population groups with an AAC of 94.14% peering that of the AISNP-74 panel (94.14%) developed in [9]. A 29-SNP panel (SNPtag-29 hereafter) can achieve an overall AAC of 95.01%. We further examined its discriminative power over each pairs of the five population groups (Fig. 2B). As shown in Fig. 2B,

a higher membership inference AAC ($> 99\%$) is achieved if only continental origins were considered. Discriminating EA and SEA origin appeared a bit challenging: in over 86% cases, individual's origin of EA or SEA were correctly identified. Individuals of EA were mis-predicted as SEA in 13.04% cases and individuals of SEA were mis-predicted as EA in 13.94% cases (Fig. 2B).

The performance of the SNPtag-29 panel was further evaluated by four measures and compared to the performance of the AISNP-74 panel [9] on each of the five population groups in Table 1. From Table 1, we can see that SNPtag-29 panel scored higher than the AISNP-74 panel in sensitivity, specificity, *PPV* and *NPV* measurements on all population groups, although it includes less than half number of SNPs of AISNP-74. Results of both the PCA and STRUCTURE analyses also indicate that the SNPtag-29 panel has higher capacity to differentiate the five major population groups than the AISNP-74 panel, especially for the groups within Asia (Fig. 3). In the PCA analysis of the SNPtag-29 data, PC1~PC4 collectively accounted for 56.4% of the total variance. Plot of PC1 (31.1%) vs. PC2 (15.2%) clearly discriminated continental population groups (*i.e.* AFR, EUR, SA and EA/SEA; Fig. 3A), while plot of PC1 vs. PC4 (3.7%) had high power discriminating EA and SEA population groups within Asia (Fig. 3B). Result from STRUCTURE analysis was congruent with PCA analysis where the five population groups were roughly clustered into discrete groups (Fig. 3E). In comparison, both PCA and STRUCTURE analyses of the AISNP-74 panel developed through empirical procedure indicated weaker discriminating power, especially for Asian populations (Fig. 3CDF). We listed details of the 29-SNP panel together with AISNP-74 in Table S1.

On the other hand, we should also point out that Li et al [9] includes more minor Asian populations in their data set which are not in the 1000 Genomes Project data. These samples increase the genetic diversity of the EA and SEA groups, and potentially require more AIM markers to

discriminate them. However, it is hard to explore its effect on AIM-SNPtag since we don't have the original data of [9]. Given this fact, we should emphasize more on the discriminating accuracy measures (AAC, sensitivity, specificity, *PPV* and *NPV*) instead of the sizes of panels in the above comparison of both panels when applying to the 1000 Genomes Project data.

3.2 New panels for membership inference

We applied AIM-SNPtag to generate ancestry informative SNP panels from genome-wide data from the 1000 Genomes Project [32]. We first generated SNP panels for discriminating the membership for the same five major population groups (AFR, EUR, EA, SA and SEA) as above. Out of the 78 million SNPs, AIM-SNPtag constructed a 21-SNP panel (SNPtag-21) that achieves an overall AAC of 95.36%, and a 36-SNP panel (SNPtag-36) that achieves an overall AAC of 99.07%. The performance of the SNPtag-36 panel on each of the population groups is shown in details in Figure 4A. For all the five population groups, the SNPtag-36 panel can assign individuals to their true populations of origin with an AAC higher than 97%. Only about 2% inference error occurred in distinguishing EA and SEA. The values of four statistics, sensitivity, specificity, *PPV* and *NPV* are high for all population groups, ranging from 0.97 to 1.00 (Table 2).

Both PCA analysis and the STRUCTURE analysis indicate that the SNPtag-36 panel reserves high discriminating power for membership of the five population groups (Fig. 5). The first 4 principle components explain nearly half (49.2%) of the variance. PC1 vs. PC2 and PC1 vs. PC3 are highly discriminative of continental population groups while PC2 vs. PC4 distinguish (Fig. 5ABC) regional groups (*i.e.* SA vs. EA vs. SEA) within Asia. All of the five major population groups cluster into nearly discrete blocks with individuals being assigned to their group of origin with high

probabilities (Fig. 5D). Compared with the AISNP-74 panel, the SNPtag-36 panel achieves even higher power for discriminating population groups within Asia (Fig. 3F vs. Fig. 6D). The details of these two nested panels were summarized in Table S2. Only three SNPs (rs16891982, rs1426654 and rs17822931) were previously selected in the AISNP-74 panel [9].

Similarly, AIM-SNPtag generated SNP panels with high discriminative power in the more challenging case involving four European subpopulations (GBR, FIN, IBS and TSI). AIM-SNPtag generated a 175-SNP panel (SNPtag-175) from the 1000 Genomes Project data, which discriminates the subpopulations with an overall AAC of 99.36%. The first 68 SNPs (SNPtag-68) of the panel achieves an overall AAC of 95.07%. The AAC values for each of the four subpopulation ranges from 97.77% to 100% (Fig. 4B). Small inference error rates occur between subpopulation pairs of TSI vs. IBS (2.22%) and GBR vs. TSI (1.10%). The values of sensitivity, specificity, precision (*PPV*) and *NPV* for the SNPtag-175 panel are all high, ranging from 98.13% to 100% (Table 2). The results of PCA analysis and the STRUCTURE analysis are consistent with the above performance measures (Fig. 6).

4. DISCUSSION

The deluge of genome-wide SNP data in recent years has enabled us unprecedented opportunities to infer individual's membership of population groups [19, 20], geographic groups [20] or even linguistic groups [20, 35, 36] with genetic data. However, genotyping a large amount of SNPs is prohibitive in most forensic cases where only trace DNA material is available. Even when ample DNA is available, the intensive computation required for downstream data analysis is time-consuming. Considering these issues, a panel of a relatively small size of highly informative

SNPs would be better suited for forensic applications [7, 9, 13].

We developed a computationally very efficient method (AIM-SNPtag) that is applicable to multiple-population genome-wide SNP data and selects the most membership informative SNPs that can be potentially applied to forensic science. Using the proposed algorithm, we successfully downscale the computational complexity of the problem from $O(2^n)$ to $O(n^2)$. AIM-SNPtag generates a panel including a small number of SNPs to achieve a certain degree of accuracy (e.g. 95% or 99%) which could meet the specific purpose.

Two case studies demonstrated that AIM-SNPtag is highly efficient in selecting AIMs for ancestry or membership inference. From millions of SNPs for more than 1,400 individuals, AIM-SNPtag generated a 21-SNP panel (SNPtag-21) that could discriminate the five major population groups with an AAC of 95% and a 36-SNP panel with an AAC of 99% (Fig. 4A). Given the accuracy threshold of 95%, the panels generated by AIM-SNPtag include a small number of markers. In the case of a more challenge situation involving four European subpopulations, GBR, FIN, IBS and TSI, AIM-SNPtag generated a 175-SNP panel (SNPtag-175) with an average accuracy of 99% (Fig. 4B).

We should also point out that in the two case studies, the samples we used to construct AIM panels for the five major population groups (AFR, EUR, EA, SA and SEA) and the European four subpopulations are from the 1000 Genomes Project, and may not be representative of the whole populations. When applying these panels to identify someone from a population not included in the training data, it may cause a bias or mistake. We suggest that in practice AIM-SNPtag should be applied to representative samples of the populations of interest, and users should beware of the AIM panels' applicable scopes.

In summary, we use the two case studies as illustration examples to demonstrate that AIM-SNPtag can explore the information in the genomic data more sufficiently and efficiently. We believe AIM-SNPtag is a useful tool for SNP panel developments in forensic or medical genetic studies.

ACKNOWLEDGMENTS

This study is supported by the CAS Key Program (KGFZD-135-16-021), the National Natural Science Foundation of China (91631106 and 31571370), and the “One Hundred Talents Program” of Chinese Academy of Sciences.

Appendix: Computational algorithms

Classification ability of a single locus

Denote the full set of loci L . Suppose for a given set of populations K , we know *a priori* $p_{kl}(g)$, the frequencies of any genotype $g \in \{0,1,2\}$ on locus $l \in L$ of population $k \in K$. The differentiation of genotype frequencies from population j to population i on locus l can be described with Kullback–Leibler divergence (KL-divergence) [37] as

$$m_{ij}^l = \sum_{g \in \{0,1,2\}} p_{il}(g) \log \left(\frac{p_{il}(g)}{p_{jl}(g)} \right).$$

For each locus, the matrix of KL-divergence among all pairs of populations can be written as

$\mathbf{M}^l = (m_{ij}^l)_{i,j \in K}$. We then define the classification ability of locus l as,

$$C_l = \sum_{i,j \in K} |m_{ij}^l|,$$

that is, the sum of all elements of matrix $\mathbf{M}^l$.

Normalized mutual information between loci

Suppose there are N samples. For any loci $l_1, l_2 \in L$, denote the genotype frequencies over N samples as $f(g_{l_1})$ and $f(g_{l_2})$, and the joint frequency of genotype pair (haplotype) as $f(g_{l_1}, g_{l_2})$, where $g_{l_1}, g_{l_2} \in \{0,1,2\}$. The mutual information (I) of loci l_1 and l_2 is then,

$$I_{l_1 l_2} = \sum_{g_{l_1} \in \{0,1,2\}} \sum_{g_{l_2} \in \{0,1,2\}} f(g_{l_1}, g_{l_2}) \log \left(\frac{f(g_{l_1}, g_{l_2})}{f(g_{l_1})f(g_{l_2})} \right).$$

The entropy on any locus $l \in L$ is denoted as

$$H_l = - \sum_{g_l \in \{0,1,2\}} f(g_l) \log f(g_l).$$

The normalized mutual information [38] between two loci is

$$NI_{l_1 l_2} = \frac{I_{l_1 l_2}}{\sqrt{H_{l_1} H_{l_2}}}, \quad \forall l_1, l_2 \in L.$$

The value of $NI_{l_1 l_2}$ is bounded in $[0,1]$, with 0 indicating complete independence and 1 otherwise.

Cumulative classification ability of a sequence of loci sets

Let $S_n = \{l_1, l_2, l_3, \dots, l_n : n \leq \#L; \forall i \in \{1, \dots, n\}, l_i \in L; \forall i \neq j, l_i \neq l_j\}$ be a finite sequence of SNPs, where $\#L$ denote the size of set L . For any $i \leq n$, denote $S_i = \{l_1, l_2, \dots, l_i\}$ the subsequence that constitute of the first i items of S_n . We define the cumulative classification ability (CCA) of S_n recursively as follows:

(a). For $S_1 = \{l_1\}$, $l_1 \in L$, define $CCA(S_1) = C_{l_1}$, where C_{l_1} is the classification ability of locus l_1 .

(b). For $2 \leq j \leq n$, $S_j = S_{j-1} \cup \{l_j\}$, denote,

$$CCA(S_j) = CCA(S_{j-1}) + \left(1 - \max_{i \in \{1, \dots, j-1\}} NI_{l_i l_j} \right) C_{l_j},$$

where $NI_{l_i l_j}$ is the normalized mutual information between loci l_i and l_j .

Note that, the calculation of $CCA(S_n)$ is dependent on the ordering of the SNPs in sequence S_n .

Algorithm for selecting the maximum CCA subset

We propose a heuristic algorithm to generate a subset with u SNPs that maximize the CCA. We refer to this subset of SNPs as MaC-SNPs (maximum classification SNPs). Our algorithm first generates $\#L$ sequences of SNPs, each denoted as S_u^l , where l represents the first item in the sequence is $l \in L$ and u indicates the size of the sequence. Each sequence is generated as follows:

(a) For $\forall l \in L$, $S_1^l = \{l_1\}$, where $l_1 = l$.

(b) For $2 \leq i \leq u$, $S_{i-1}^l = \{l_1, \dots, l_{i-1}\}$, $l_i = \operatorname{argmax}_{k \in L \setminus S_{i-1}^l} (CCA(S_{i-1}^l \cup \{k\}))$, $S_i^l = S_{i-1}^l \cup \{l_i\}$,

where $L \setminus S_{i-1}^l$ denotes the subset of all SNPs in L not in S_{i-1}^l . The sequence with maximum CCA is selected as the MaC-SNPs, that is, we choose MaC-SNP $S = S_u^t$, where $t = \operatorname{argmax}_{l \in L} CCA(S_u^l)$.

Classifier

We adopt the Naïve Bayes classifier (NBC) to test the performance of the MaC-SNP set and infer the population origin of an unlabeled sample. NBC assumes that loci are independent of each other given the targeted label (population). The assumption of conditional independence is rather strong and often violated in practice. Nonetheless, Naive Bayes performs quite well in practice, often comparable to more sophisticated learning methods [39]. For a sample from an unknown population with given genotypes $G = \{g_1, g_2, \dots, g_n\}$ on the MaC-SNP $S = \{l_1, l_2, \dots, l_n\}$, the posterior probability of the sample originated from $k \in K$ is

$$p(k|G) \propto p(k) \prod_{i \in \{1, 2, \dots, n\}} p_{kl_i}(g_i).$$

The term $p(k)$ is the prior probability of population k , which can be dropped out if we assume an uniform prior. The $p_{kl_i}(g_i)$ is the genotype frequency of locus $l_i \in S$ in population $k \in K$, which can be learned by training data with known labels (e.g. samples with known population origin). The label of the sample to be inferred can be obtained by,

$$h = \operatorname{argmax}_{k \in K} p(k|G).$$

By reducing multiplication to addition the above equation is equivalent to

$$h = \operatorname{argmax}_{k \in K} \sum_{i \in \{1, 2, \dots, n\}} \log(p_{kl_i}(g_i) + \epsilon),$$

where we add a small positive value $\epsilon = 10^{-4}$ to avoid underflow of logarithm calculation.

Partition and integration

The populations often have hierarchical structure. To generate a SNP panel with discriminative power that comprehensively covers the substructure among available populations, we propose a partition and integration procedure, which extracts the MaC-SNP sets for each of the structured class, and then merges them to generate a final panel of SNPs. The classes can be determined based on *a priori* knowledge, or by clustering methods in a data-driven manner (*i.e.* PCA).

We first extract from each class a candidate MaC-SNP set with a fixed size. The size of the MaC-SNP sets should not be too small. A wrapper-based feature selection algorithm is then applied to obtain the combined global set, which should meet some performance criteria or reach a predefined global set size. Here, we elaborate a greedy forward selection method together with NBC

to generate the global SNP panel.

Note, SNPs in each MaC-SNP set are already in an ordered sequence as a result of their generation procedures. The greedy forward selection algorithm for generating the global panel P applies as follows. We initialize an empty set P and start with the first SNP of each of the MaC-SNP sets. The performances (e.g., average accuracy, AAC, see the main text for definition) of including the first SNP from different MaC-SNP sets to P are compared. The SNP that gains the highest AAC is chosen to be a new marker in P and it is removed from the corresponding MaC-SNP set. This procedure is repeated until a given threshold of global panel size or AAC is met. The final P is thus the global SNP panel.

References

- [1] M. de la Puente, C. Santos, M. Fondevila, L. Manzo, A. Carracedo, M.V. Lareu, C. Phillips, E.-N. Consortium, The Global AIMs Nano set: A 31-plex SNaPshot assay of ancestry-informative SNPs, *Forensic Sci Int-Gen* 22 (2016) 81-88.
- [2] P. Drineas, J. Lewis, P. Paschou, Inferring geographic coordinates of origin for Europeans using small panels of ancestry informative markers, *Plos One* 5(8) (2010).
- [3] M. Eduardoff, T.E. Gross, C. Santos, M. de la Puente, D. Ballard, C. Strobl, C. Børsting, N. Morling, L. Fusco, C. Hussing, B. Egyed, L. Souto, J. Uacyisrael, D. Syndercombe Court, Á. Carracedo, M.V. Lareu, P.M. Schneider, W. Parson, C. Phillips, W. Parson, C. Phillips, Inter-laboratory evaluation of the EUROFORGEN Global ancestry-informative SNP panel by massively parallel sequencing using the Ion PGM™, *Forensic Science International: Genetics* 23 (2016) 178-189.
- [4] 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(5) (2008) 648-658.
- [5] 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, *Investigative Genetics* 2(1) (2011) 1.
- [6] K.K. Kidd, A.J. Pakstis, W.C. Speed, E.L. Grigorenko, S.L.B. Kajuna, N.J. Karoma, S. Kungulilo, J.J. Kim, R.B. Lu, A. Odunsi, F. Okonofua, J. Parnas, L.O. Schulz, O.V. Zhukova, J.R. Kidd, Developing a SNP panel for forensic identification of individuals, *Forensic Sci Int* 164(1) (2006) 20-32.
- [7] K.K. Kidd, W.C. Speed, A.J. Pakstis, M.R. Furtado, R.X. 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-Gen* 10 (2014) 23-32.
- [8] 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(1) (2009) 69-78.
- [9] C.X. Li, A.J. Pakstis, L. Jiang, Y.L. Wei, Q.F. Sun, H. Wu, O. Bulbul, P. Wang, L.L. Kang, J.R. Kidd, K.K. Kidd, A panel of 74 AISNPs: Improved ancestry inference within Eastern Asia, *Forensic Sci Int-Gen* 23 (2016) 101-110.
- [10] L.M. Li, Y. Wang, S.P. Yang, M.Y. Xia, Y.J. Yang, J.C. Wang, D.R. Lu, X.W. Pan, T. Ma, P. Jiang, G. Yu, Z.Q. Zhao, Y. Ping, H.G. Zhou, X.Y. Zhao, H. Sun, B. Liu, D.T. Jia, C.T. Li, R.L. Hu, H.Z. Lu, X.Y. Liu, W.Q. Chen, Q. Mi, F.Z. Xue, Y.D. Su, L. Jin, S.L. Li, Genome-wide screening for highly discriminative SNPs for personal identification and their assessment in world populations, *Forensic Sci Int-Gen* 28 (2017) 118-127.
- [11] R. Nassir, R. Kosoy, 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, An ancestry informative marker set for determining continental origin: validation and extension using human genome diversity panels, *Bmc Genet* 10 (2009).
- [12] 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, *Investigative Genetics* 4(1) (2013) 13.

- [13] C. Phillips, Forensic genetic analysis of bio-geographical ancestry, *Forensic Sci Int-Gen* 18 (2015) 49-65.
- [14] C. Phillips, W. Parson, B. Lundsberg, C. Santos, A. Freire-Aradas, M. Torres, M. Eduardoff, C. Borsting, P. Johansen, M. Fondevila, N. Morling, P. Schneider, A. Carracedo, M.V. Lareu, E.-N. Consortium, Building a forensic ancestry panel from the ground up: The EUROFORGEN Global AIM-SNP set, *Forensic Sci Int-Gen* 11 (2014) 13-25.
- [15] C. Phillips, A. Salas, J.J. Sánchez, M. Fondevila, A. Gómez-Tato, J. Álvarez-Dios, M. Calaza, M.C. de Cal, D. Ballard, M.V. Lareu, Á. Carracedo, Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs, *Forensic Science International: Genetics* 1(3) (2007) 273-280.
- [16] C. Santos, C. Phillips, M. Fondevila, R. Daniel, R.A.H. van Oorschot, E.G. Burchard, M.S. Schanfield, L. Souto, J. Uacyisrael, M. Via, A. Carracedo, M.V. Lareu, Pacifiplex: an ancestry-informative SNP panel centred on Australia and the Pacific region, *Forensic Sci Int-Gen* 20 (2016) 71-80.
- [17] P. Yahya, S. Sulong, A. Harun, H.W. Isa, N.S. Ab Rajab, P. Wangkumhang, A. Wilantho, C. Ngamphiw, S. Tongsima, Z. Alwi, Analysis of the genetic structure of the Malay population: Ancestry-informative marker SNPs in the Malay of Peninsular Malaysia, *Forensic Sci Int-Gen* 30 (2017) 152-159.
- [18] S. Walsh, F. Liu, K.N. Ballantyne, M. van Oven, O. Lao, M. Kayser, IrisPlex: A sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information, *Forensic Sci Int-Gen* 5(3) (2011) 170-180.
- [19] 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(5866) (2008) 1100-1104.
- [20] J. Novembre, T. Johnson, K. Bryc, Z. Kutalik, A.R. Boyko, A. Auton, A. Indap, K.S. King, S. Bergmann, M.R. Nelson, M. Stephens, C.D. Bustamante, Genes mirror geography within Europe, *Nature* 456(7218) (2008) 98-101.
- [21] M. Gymrek, A.L. McGuire, D. Golan, E. Halperin, Y. Erlich, Identifying personal genomes by surname inference, *Science* 339(6117) (2013) 321-324.
- [22] J.M. Butler, *Fundamentals for Forensic DNA Typing*, Elsevier 2009.
- [23] J.M. Butler, Chapter 12 - Single Nucleotide Polymorphisms and Applications, *Advanced Topics in Forensic DNA Typing: Methodology*, Academic Press, San Diego, 2012, pp. 347-369.
- [24] O. Lao, K.v. Duijn, P. Kersbergen, P.d. Knijff, M. Kayser, Proportioning whole-genome single-nucleotide-polymorphism diversity for the identification of geographic population structure and genetic ancestry, *The American Journal of Human Genetics* 78(4) (2006) 680-690.
- [25] X. Mao, A.W. Bigham, R. Mei, G. Gutierrez, K.M. Weiss, T.D. Brutsaert, F. Leon-Velarde, L.G. Moore, E. Vargas, P.M. McKeigue, M.D. Shriver, E.J. Parra, A genomewide admixture mapping panel for Hispanic/Latino populations, *The American Journal of Human Genetics* 80(6) (2007) 1171-1178.
- [26] A.L. Price, N. Patterson, F. Yu, D.R. Cox, A. Waliszewska, G.J. McDonald, A. Tandon, C. Schirmer, J. Neubauer, G. Bedoya, C. Duque, A. Villegas, M.C. Bortolini, F.M. Salzano, C. Gallo, G. Mazzotti, M. Tello-Ruiz, L. Riba, C.A. Aguilar-Salinas, S. Canizales-Quinteros, M. Menjivar, W. Klitz, B. Henderson, C.A. Haiman, C. Winkler, T. Tusie-Luna, A. Ruiz-Linares, D. Reich, A genomewide admixture map for Latino populations, *The American Journal of*

- Human Genetics 80(6) (2007) 1024-1036.
- [27] C. Tian, D.A. Hinds, R. Shigeta, R. Kittles, D.G. Ballinger, M.F. Seldin, A genome-wide single-nucleotide-polymorphism panel with high ancestry information for African American admixture mapping, *The American Journal of Human Genetics* 79(4) (2006) 640-649.
- [28] D.M. Altshuler, R.A. Gibbs, L. Peltonen, E. Dermitzakis, S.F. Schaffner, F.L. Yu, P.E. Bonnen, P.I.W. de Bakker, P. Deloukas, S.B. Gabriel, R. Gwilliam, S. Hunt, M. Inouye, X.M. Jia, A. Palotie, M. Parkin, P. Whittaker, K. Chang, A. Hawes, L.R. Lewis, Y.R. Ren, D. Wheeler, D.M. Muzny, C. Barnes, K. Darvishi, M. Hurles, J.M. Korn, K. Kristiansson, C. Lee, S.A. McCarroll, J. Nemesh, A. Keinan, S.B. Montgomery, S. Pollack, A.L. Price, N. Soranzo, C. Gonzaga-Jauregui, V. Anttila, W. Brodeur, M.J. Daly, S. Leslie, G. McVean, L. Moutsianas, H. Nguyen, Q.R. Zhang, M.J.R. Ghorri, R. McGinnis, W. McLaren, F. Takeuchi, S.R. Grossman, I. Shlyakhter, E.B. Hostetter, P.C. Sabeti, C.A. Adebamowo, M.W. Foster, D.R. Gordon, J. Licinio, M.C. Manca, P.A. Marshall, I. Matsuda, D. Ngare, V.O. Wang, D. Reddy, C.N. Rotimi, C.D. Royal, R.R. Sharp, C.Q. Zeng, L.D. Brooks, J.E. McEwen, I.H. Consortium, Integrating common and rare genetic variation in diverse human populations, *Nature* 467(7311) (2010) 52-58.
- [29] M.R. Nelson, D. Wegmann, M.G. Ehm, D. Kessner, P.S. Jean, C. Verzilli, J.D. Shen, Z.Z. Tang, S.A. Bacanu, D. Fraser, L. Warren, J. Aponte, M. Zawistowski, X. Liu, H. Zhang, Y. Zhang, J. Li, Y. Li, L. Li, P. Woollard, S. Topp, M.D. Hall, K. Nangle, J. Wang, G. Abecasis, L.R. Cardon, S. Zollner, J.C. Whittaker, S.L. Chisoe, J. Novembre, V. Mooser, An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people, *Science* 337(6090) (2012) 100-104.
- [30] J.A. Tennessen, A.W. Bigham, T.D. O'Connor, W.Q. Fu, E.E. Kenny, S. Gravel, S. McGee, R. Do, X.M. Liu, G. Jun, H.M. Kang, D. Jordan, S.M. Leal, S. Gabriel, M.J. Rieder, G. Abecasis, D. Altshuler, D.A. Nickerson, E. Boerwinkle, S. Sunyaev, C.D. Bustamante, M.J. Bamshad, J.M. Akey, B. GO, S. GO, N.E.S. Project, Evolution and functional impact of rare coding variation from deep sequencing of human exomes, *Science* 337(6090) (2012) 64-69.
- [31] A. Sarkar, M.R. Nandineni, Development of a SNP-based panel for human identification for Indian populations, *Forensic Sci Int-Gen* 27 (2017) 58-66.
- [32] D.M. Altshuler, R.M. Durbin, G.R. Abecasis, D.R. Bentley, e. al., A global reference for human genetic variation, *Nature* 526(7571) (2015) 68-74.
- [33] N. Patterson, A.L. Price, D. Reich, Population structure and eigenanalysis, *Plos Genet* 2(12) (2006) 2074-2093.
- [34] J.K. Pritchard, M. Stephens, P. Donnelly, Inference of population structure using multilocus genotype data, *Genetics* 155(2) (2000) 945-959.
- [35] J.M. Chen, H.F. Zheng, J.X. Bei, L.D. Sun, W.H. Jia, T. Li, F.R. Zhang, M. Seielstad, Y.X. Zeng, X.J. Zhang, J.J. Liu, Genetic Structure of the Han Chinese Population Revealed by Genome-wide SNP Variation, *Am J Hum Genet* 85(6) (2009) 775-785.
- [36] G. Longobardi, S. Ghirotto, C. Guardiano, F. Tassi, A. Benazzo, A. Ceolin, G. Barbujani, Across language families: Genome diversity mirrors linguistic variation within Europe, *Am J Phys Anthropol* 157(4) (2015) 630-640.
- [37] S. Kullback, R.A. Leibler, On Information and Sufficiency, *Ann Math Stat* 22(1) (1951) 79-86.
- [38] A. Strehl, J. Ghosh, Cluster Ensembles – A knowledge reuse framework for combining multiple partitions, *Journal of Machine Learning Research* 3 (2002) 583-617.

509 [39] P. Domingos, M. Pazzani, On the optimality of the simple Bayesian classifier under zero-one
510 loss, Mach Learn 29(2-3) (1997) 103-130.
511

Table 1. Comparison of performance of two SNP panel developed from 178 pre-selected SNP pool on discriminating five major population groups in the 1000 Genome Project data. Values of sensitivity, specificity, PPV and NPV for the 74 AISNP panel (front, regular) and the corresponding values for the SNPtag-29 panel (rear, in bold) are show together.

	Sensitivity	Specificity	PPV	NPV
SAFR	1.0000/ 1.0000	1.0000/ 1.0000	1.0000/ 1.0000	1.0000/ 1.0000
EUR	0.9996/ 0.9981	0.9988/ 0.9983	0.9958/ 0.9940	0.9999/ 0.9995
SA	0.9973/ 0.9961	0.9999/ 0.9994	0.9997/ 0.9988	0.9986/ 0.9980
EA	0.8448/ 0.8713	0.9703/ 0.9758	0.8894/ 0.9106	0.9567/ 0.9640
SEA	0.8293/ 0.8615	0.9604/ 0.9671	0.7668/ 0.8051	0.9728/ 0.9779

Table 2. Performance SNP panels developed from genome-wide data sets.

	Sensitivity	Specificity	PPV	NPV
SNPtag-21 (AAC=95%)/SNPtag-36 (AAC=99%)				
AFR	1.0000/ 1.0000	0.9998/ 1.0000	0.9971/ 1.0000	1.0000/ 1.0000
EUR	0.9992/ 1.0000	0.9981/ 1.0000	0.9932/ 1.0000	0.9998/ 1.0000
SA	0.9935/ 1.0000	0.9962/ 0.9979	0.9928/ 0.9960	0.9966/ 1.0000
EA	0.8303/ 0.9759	0.9920/ 0.9963	0.9670/ 0.9869	0.9538/ 0.9932
SEA	0.9466/ 0.9738	0.9575/ 0.9947	0.7786/ 0.9663	0.9913/ 0.9959
SNPtag-68 (AAC=95%)/SNPtag-175(AAC=99%)				
FIN	1.0000/ 1.0000	0.9953/ 1.0000	0.9856/ 0.9999	1.0000/ 1.0000
GBR	0.9423/ 0.9879	0.9868/ 1.0000	0.9541/ 1.0000	0.9833/ 0.9906
IBS	0.9524/ 0.9999	0.9587/ 0.9934	0.8926/ 0.9700	0.9824/ 1.0000
TSI	0.9081/ 0.9813	0.9921/ 0.9927	0.9764/ 0.9816	0.9677/ 0.9926

Figures:

Fig. 1. Scheme of AIM-SNPtag for retrieving ancestry or membership informative SNP panel.

Fig 2. Performance of AIM-SNPtag in retrieving the most informative SNPs from a pool of 178 pre-selected SNPs. (A) The plot shows AAC as a function of the panel size. AIM-SNPtag was applied to a pool of 178 SNPs pre-selected in [9] from which the AISNP-74 panel was generated. The horizontal dashed lines indicate the AAC values of 94% and 95% which achieved by the SNPtag-18 panel and SNPtag-29 panel, respectively. The AISNP-74 panel had an AAC value of ~94%. (B) The performance of the SNPtag-29 panel in discriminating the five major groups in the 1000 Genomes Project data. The values on the diagonal represent proportion of right inference while those off the diagonal represent proportion of wrong inference.

Fig 3. PCA (A~D) and STRUCTURE (E~F) analyses of individuals from five populations (AFR, EUR, EA, SA, and SEA) using the new SNPtag-29 panel (AB&E) the AISNP-74 panel (CD&F). Both indicate a better classification performance of the 29-SNP panel.

Fig 4. Performance of the SNP panels in discriminating each of the classes. (A) Performance of SNPtag-36 on the five major population groups. (B) Performance of SNPtag-175 on the four subpopulations within Europe. The values on the diagonal represent the proportion of correct inference while those off the diagonal represent the proportion of mis-inference.

Fig 5. Membership analyses of five reference population groups (AFR, EUR, EA, SA, and SEA) using the SNPtag-36 panel generated by AIM-SNPtag. (A~C) Scatter plots of the first four principal components (PC1 vs. PC2, PC1 vs. PC3 and PC2 vs. PC4) discriminate five groups. The first four PCs accounted collectively for 49.2% of the variation in the data. (D) STRUCTURE analysis of the individuals.

Fig 6. Membership analyses of four European reference subpopulations (GBR, FIN, IBS and TSI) using the SNPtag-175 panel generated by AIM-SNPtag. (A~C) Scatter plots of the first four principal components (PC1 vs. PC2, PC1 vs. PC3 and PC2 vs. PC4) discriminate four groups. (D) STRUCTURE analysis of the individuals.

561 Table S1. Details of SNP panels selected by AIM-SNPtag from 178 pre-selected SNPs and the
562 AISNP panel by [9].

dbSNP ID	Chr	Position(GRCh37)	AIM-SNPtag		AISNP-74 (94.14%)
			SNPtag-18 (94.14%)	SNPtag-29 (95.01%)	
rs4833103	4	38815502	✓	✓	✓
rs3811801	4	100244319	✓	✓	✓
rs16891982	5	33951693	✓	✓	✓
rs7745461	6	21911616	✓	✓	✓
rs10108270	8	4190793	✓	✓	✓
rs10511828	9	28628500	✓	✓	✓
rs2006996	9	117592638	✓	✓	✓
rs10513300	9	120130206	✓	✓	✓
rs174574	11	61600342	✓	✓	✓
rs671	12	112241766	✓	✓	✓
rs9522149	13	111827167	✓	✓	✓
rs8003942	14	105971670	✓	✓	✓
rs1426654	15	48426484	✓	✓	✓
rs17822931	16	48258198	✓	✓	✓
rs6054605	20	744570	✓	✓	✓
rs310644	20	62159504	✓	✓	✓
rs2196051	8	122124302	✓	✓	
rs1871534	8	145639681	✓	✓	
rs4670767	2	37941396		✓	✓
rs13400937	2	79864923		✓	✓
rs3827760	2	109513601		✓	✓
rs10496971	2	145769943		✓	✓
rs1572018	13	41715282		✓	✓
rs12913832	15	28365618		✓	✓
rs8113143	19	33652247		✓	✓
rs1834619	2	17901485		✓	
rs10007810	4	41554364		✓	
rs4918664	10	94921065		✓	
rs2416791	12	11701488		✓	
rs2986742	1	6550376			✓
rs647325	1	18170886			✓
rs4908343	1	27931698			✓
rs3118378	1	68849687			✓
rs3737576	1	101709563			✓
rs7554936	1	151122489			✓
rs2814778	1	159174683			✓
rs1040404	1	168159890			✓
rs798443	2	7968275			✓
rs1876482	2	17362568			✓
rs6754311	2	136707982			✓

rs734873	3	147750355	√
rs385194	4	85309078	√
rs1229984	4	100239319	√
rs2241894	4	100266133	√
rs10516441	4	100307167	√
rs17028973	4	100322786	√
rs2702414	4	179399523	√
rs316598	5	2364626	√
rs870347	5	6845035	√
rs37369	5	35037115	√
rs6451722	5	43711378	√
rs7722456	5	170202984	√
rs192655	6	90518278	√
rs1871428	6	168665760	√
rs917115	7	28172586	√
rs2242480	7	99361466	√
rs10236187	7	139447377	√
rs6990312	8	110602317	√
rs3814134	9	127267689	√
rs174570	11	61597212	√
rs1513056	12	17407792	√
rs2238151	12	112211833	√
rs9319336	13	27624356	√
rs7997709	13	34847737	√
rs2166624	13	42579985	√
rs1950993	14	58238687	√
rs8021730	14	67886781	√
rs1800414	15	28197037	√
rs735480	15	45152371	√
rs2899826	15	74734500	√
rs8035124	15	92105708	√
rs818386	16	65406708	√
rs459920	16	89730827	√
rs2033111	17	53788280	√
rs11652805	17	62987151	√
rs10512572	17	69512099	√
rs2125345	17	73782191	√
rs7226659	18	40488279	√
rs7238445	18	49781544	√
rs2024566	22	41697338	√

563

564

Table S2. Details of two SNP panels selected by AIM-SNPtag from the 1000 Genomes Project data set that discriminate five major population groups. The numbers in parentheses are the membership inference accuracies of the two panels.

dbSNP ID	Chr	Position (GRCh37)	SNPtag-21 (95.36%)	SNPtag-36 (99.07%)	Frequencies				
					AFR	EUR	SA	EA	SEA
rs111824286	1	604299	√	√	0.0787	0.0272	0.0082	0.1186	0.0000
rs4366944	2	97628047	√	√	0.6852	0.8962	0.3221	0.4487	0.2812
rs149768401	3	100365528	√	√	0.2083	0.0815	0.0000	0.2837	0.0078
rs16891982	5	33951693	√	√	0.0000	0.9201	0.0593	0.0064	0.0052
rs3857620	6	57496076	√	√	0.0000	0.0000	0.4611	0.0256	0.0052
rs555368174	6	82912256	√	√	0.0093	0.0208	0.0031	0.0032	0.1615
rs4467891	7	50222195	√	√	0.3843	0.0016	0.3875	0.0465	0.1328
rs12003714	9	21932366	√	√	0.8611	0.0080	0.1176	0.0016	0.0182
rs10817902	9	119309609	√	√	0.0000	0.0000	0.3466	0.2516	0.2161
rs72643557	11	61579427	√	√	0.0000	0.0032	0.0235	0.5721	0.1719
rs538254210	14	96938945	√	√	0.0000	0.0288	0.0061	0.0000	0.1901
rs11624491	14	106188045	√	√	0.6343	0.3163	0.5460	0.6234	0.1328
rs1426654	15	48426484	√	√	0.0139	0.9984	0.6851	0.0176	0.0026
rs118077539	16	3281045	√	√	0.0000	0.0000	0.3190	0.0272	0.0339
rs200458768	16	30998152	√	√	0.0000	0.0000	0.5051	0.0000	0.0052
rs368738705	16	33921593	√	√	0.0000	0.0000	0.4049	0.0000	0.0000
rs367953206	16	48221771	√	√	0.0000	0.0000	0.3650	0.0032	0.0026
rs434124	19	54809336	√	√	0.6157	0.1454	0.2444	0.7869	0.2839
rs2208074	20	53253867	√	√	1.0000	0.4281	0.1125	0.0000	0.0052
rs187760729	20	62157764	√	√	0.0000	0.0000	0.3722	0.0032	0.0182
rs61703312	8	116829577	√	√	0.9352	0.0160	0.1227	0.0609	0.0807
rs561510556	1	12387655		√	0.0046	0.0383	0.0143	0.0112	0.2422
rs11204850	1	151540302		√	0.0046	0.0000	0.2955	0.2788	0.3177
rs371529719	2	97897283		√	0.0000	0.0000	0.0133	0.0144	0.1927
rs546642722	4	17813761		√	0.0000	0.0320	0.0072	0.0064	0.2474
rs56357984	4	38806462		√	0.0000	0.2748	0.1575	0.4087	0.1094
rs185365488	4	101002868		√	0.0000	0.0000	0.0092	0.0369	0.2656
rs201599166	4	114275541		√	0.1620	0.2508	0.3262	0.1571	0.4010
rs541669109	5	44671775		√	0.0602	0.0351	0.0000	0.0978	0.0000
rs172447	9	4859106		√	0.9630	0.0367	0.1227	0.0401	0.0287
rs2001786	11	128782590		√	0.0185	0.0000	0.2945	0.1939	0.2318
rs555658336	14	104193238		√	0.1204	0.0447	0.0000	0.1090	0.0000
rs201107042	16	30182239		√	0.0648	0.0032	0.4049	0.0609	0.1432
rs17822931	16	48258198		√	0.0000	0.1102	0.4816	0.8974	0.5885
rs7197127	16	71473836		√	0.9630	0.4217	0.1196	0.0016	0.0052
rs389096	19	54764322		√	0.2037	0.2444	0.3323	0.7676	0.3958

568