Ancestry informative markers for fine-scale individual assignment to worldwide populations

Peristera Paschou, Jamey Lewis, Asif Javed, Petros Drineas

Ancestry Informative Markers for Fine-Scale Individual Assignment to Worldwide Populations

Peristera Paschou[*], Jamey Lewis[‡], Asif Javed[‡†], Petros Drineas[‡]

Corresponding author:
Peristera Paschou
Department of Molecular Biology and Genetics
Democritus University of Thrace
Panepistimioupoli, Dragana, Ktirio 8
Alexandroupoli 68100, Greece.
(Tel.) +30 25510 30658
(e-mail) ppaschou@mbg.duth.gr

Running title: Ancestry inference for worldwide populations

Keywords: Ancestry inference, Ancestry Informative Markers, population structure, PCA-correlated SNPs (PCAIMs), Informativeness, redundancy removal, HGDP

Competing Interest: None declared.

[*]Department of Molecular Biology and Genetics, Democritus University of Thrace, Alexandroupoli 68100, Greece.
[†]Computational Biology Group, IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, USA.
[‡]Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY 12180, USA.

Abstract

The analysis of large-scale genetic data from thousands of individuals has revealed the fact that subtle population genetic structure can be detected at levels that were previously unimaginable. Using the Human Genome Diversity Panel as reference (51 populations - 650,000 SNPs), we describe a systematic evaluation of the resolution that can be achieved for the inference of genetic ancestry, even when small panels of genetic markers are used. Leveraging the power of Principal Components Analysis (PCA), we undertake a comprehensive investigation of human population structure around the world. We dissect the problem into hierarchical steps, proposing a decision tree for the prediction of individual ancestry. A complete leave-one-out validation experiment demonstrates that, using all available SNPs, assignment of individuals to their self-reported populations of origin is essentially perfect. Ancestry informative genetic markers are selected using two different metrics (I_n and correlation with PCA scores). Performing a thorough cross-validation experiment, we show that, in most cases here, the number of SNPs needed for ancestry inference can be successfully reduced to less than 0.1% of the original 650,000 while retaining close to 100% accuracy. This reduction can be achieved using a clustering-based redundancy removal algorithm which we also introduce. The applicability of our suggested SNP panels is tested on HapMap Phase 3 populations. The methods we describe, in combination with the increasingly more comprehensive databases of human genetic variation, open new horizons in a variety of fields, ranging from the study of human evolution and population history, to medical genetics and forensics.

Introduction

The patterns of human genetic variation around the globe have been forged by the history of the human population. As evidenced from fossil records and population genetics studies, anatomically modern humans first appeared in Africa, some 200,000-150,000 years ago [1, 2, 3]. About 60,000 years ago humans exited Africa in waves of migrations and, through a sequential chain of colonies, spread to occupy most of today's land masses. During this journey they encountered different environments and climates and came in contact with novel pathogens and animals. They formed local communities, separated by geographic, linguistic, cultural, and social barriers. Mutation, genetic drift, and natural selection operated in parallel with demographic and historical events to weave the patterns of human variation in extant populations. The result of this interplay was the imprint of genetic ancestry and population structure carried in the genome of each individual.

Analyzing microsatellite markers that spanned the entire genome, Rosenberg et al. [4] were the first to report on the patterns of human genetic variation and population genetic structure based on genomewide data, revealing clines of genetic diversity around the world. The advent of modern technologies and the realization of the HapMap project allowed the detailed characterization of human genetic variation across all chromosomes in diverse populations using dense marker maps [5, 6, 7]. Nevertheless, the thorough evaluation of the extent of fine-scale genetic structure among closely neighboring populations, as well as the study of the ability to infer individual membership down to a particular population within a continent, have only begun in the past three years [8, 9, 10, 11, 12]. The initial release of dense genotypic data on the Human Genome Diversity Panel (HGDP) [8], a collection of samples including more than one thousand individuals from 51 populations from around the world [13, 14], showed that fine-scale population differentiation was indeed possible, when hundreds of thousands of Single Nucleotide Polymorphisms (SNPs) were studied. The complex fine-scale genetic structure of European populations was recently presented at even greater detail with data available from multiple, closely neighboring populations, revealing strong correlation of genetic background with geographical coordinates [9, 10, 11, 12]. Similar results were obtained from the first fine scale study of African diversity [15], focusing on 121 African populations.

With the volume of rich genotypic data rapidly increasing – thanks to high-throughput genotyping and the availability of dense geographic samples – Principal Components Analysis (PCA) emerged as a powerful technique that can be used to summarize and process the vast amounts

of available information. PCA is a linear dimensionality reduction technique that can effectively extract the fundamental structure of a dataset without any need for modeling of the data. It has been used to decompose the complex genetic structure of human populations [4, 8, 16] and it can be successfully applied to infer genetic ancestry as well as substructure in a given sample [17, 18]. Importantly, it has become an effective tool for the correction of biases produced by the existence of stratification in large-scale genome-wide association studies (GWAS) seeking to uncover the genetic basis of complex disorders [19, 17, 18].

Using PCA, as well as a fast, efficient implementation of a maximum likelihood method implemented in the software frappe [20], Li et al. [8] analyzed the HGDP samples focusing on the top few axes of variation, as revealed by PCA, and found that individual ancestry and substructure were detectable with very high resolution. Biswas et al. [16] further explored the structure identified by PCA in this dataset, performing a thorough examination of lower-order Principal Components (PCs) that they found significantly correlated with structure in the complete dataset or in subsets of the data corresponding to the seven pre-defined broad geographic regions of the HGDP data. They showed that, although most researchers traditionally focus on the top few axes of variation in a dataset, substantial information about population structure exists in lower-ranked PCs. They proceeded to identify all SNPs that were significantly correlated with the top two PCs and showed that, to some extent, these markers could be used to reconstruct the structure of the complete dataset. However, they did not attempt to evaluate the performance of small panels of AIMs for ancestry inference or population structure identification. In fact, given all 51 HGDP populations from around the world no previous study has attempted the systematic evaluation of the potential to infer the ancestry of a given individual down to a particular population using small panels of SNPs. It should also be emphasized that no previous study of such large scale has actually viewed the question of inferring individual ancestry as a classification problem. Classification is a goal which is parallel, but more challenging, than simply analyzing genetic data from different populations in order to investigate the extent to which these populations can be differentiated. Classification involves cross-validation experiments of assigning individuals of "unknown" origin to one of several reference populations.

The identification of ancestry informative markers (AIMs) is a topic that has attracted considerable attention due to the value of such markers in diverse areas, ranging from forensics, to population genetics, conservation genetics, and medical genetics. Different metrics have been proposed in order to select such markers. Most of them (e.g., δ, Wright's F_{ST}, etc.) rely on the maximization of allele frequency differences between pre-defined populations [21, 22, 23, 24, 25, 26, 27].

A closely correlated measure, Informativeness for assignment (I_n) as defined by Rosenberg et al. [28] computes a mutual information based metric on allele frequencies. Leveraging the properties of PCA we described a method that identifies SNPs that are correlated with significant PCs (PCA-correlated SNPs or PCA Informative Markers - PCAIMs for short) [18]. In fact, we demonstrated that small panels of such SNPs can successfully reproduce the structure of a dataset, as identified by PCA, without any prior knowledge or hypothesis on the origin of studied individuals or artificial assignment of individuals to pre-defined clusters [18, 29].

The results described herein are a systematic investigation, at a worldwide level, of the extent to which an individual of unknown origin can be assigned to a particular population using only information from small panels of carefully selected SNPs. Using the HGDP data from 51 populations and 650,000 SNPs as reference [8], we first perform a comprehensive investigation of the structure of the dataset as identified by PCA. Instead of seeking to uncover information in lower-ranked PCs, we dissect the problem into hierarchical steps, proposing a decision-tree for the prediction of individual ancestry. After processing information at each level we proceed to analyze the next one, exposing the picture of population structure in further resolution until all meaningful information is extracted. AIMs are selected using two different metrics (I_n [28] and PCA scores [18]). Faced with the problem of redundancy in the information carried by the selected AIMs, we propose and evaluate a simple, clustering-based, strategy in order to minimize the number of markers needed for the inference of population structure. In order to estimate the generalization error of our methods, we run over eight hundred PCA computations and report results on a thorough crossvalidation experiment. Finally, we test the applicability of our suggested SNP panels on HapMap Phase 3 populations. Our results demonstrate that fine-scale inference of individual ancestry is indeed possible even with small, albeit judiciously selected, sets of genetic markers.

Methods

Datasets

We studied a previously described dataset of 1043 individuals from 51 populations from around the world [8]. These samples can be classified into seven broad geographic regions (Africa, Middle East, Europe, Central South Asia, East Asia, Oceania, and America). The samples have been genotyped for approximately 650,000 SNPs across the genome using the Illumina 650Y array. As a second dataset we also studied SNPs for the selected ancestry informative panels from

the HapMap Phase 3 database on the Yoruba (YRI), African American from Southwest USA (ASW), Luhya in Webuye, Kenya (LWK), Maasai in Kinyawa, Kenya (MKK), CEPH European (CEU), Italian from Tuscany (TSI), Chinese from Beijing and the Denver Metropolitan area (CHB and CHD), Japanese (JPT), Gujarati Indians in Houston, Texas (GIH), and Mexicans in Los Angeles, California (MEX) samples [5, 6, 7]. For all datasets we only considered SNPs on autosomal chromosomes in our analysis. We excluded SNPs with more than 10% missing entries, and we analyzed a total of 643,862 SNPs. A small number of outlier individuals was also removed from further analysis. Full details on preprocessing and encoding the data, are provided in Supplementary methods.

Selecting ancestry informative markers and removing redundancy

In order to select ancestry informative markers (AIMs), we used two previously described procedures. The first procedure [18, 29] returns the so-called PCA Informative Markers or PCAIMs for short and is based on the well-documented fact that PCA reveals population structure [30, 31, 32, 17, 33, 18]. The PCAIM selection algorithm first determines the number of significant principal components (and thus the number of informative eigenSNPs) in the data (see Table 1 for the number of significant principal components and Supplementary Material for a detailed description of how this number was determined for the purposes of the analyses presented here). Subsequently, a score is assigned to each SNP, with higher scores corresponding to SNPs that correlate well with all informative eigenSNPs. The algorithm returns the top scoring SNPs, and we have demonstrated that these PCAIMs are very efficient for ancestry prediction [18]. The second procedure computes the so-called informativeness for assignment (I_n) metric, a mutual information based statistic that takes into account self-reported ancestry information from the sampled individuals [28]. We will call the selected markers I_n AIMs for short.

It is worth noting that neither method takes any special measures in order to avoid redundancy in the set of identified markers. Such redundancy, especially in the case of dense sets of SNP markers, is typically due to tight linkage disequilibrium. In [29] we proposed a linear-algebraic method to remove redundancy from the selected PCAIMs. Our methodology was based on reducing the redundancy removal problem to the so-called Column Subset Selection Problem (CSSP) and on leveraging algorithms and software that are available for the latter problem. Subsequent to our work, Boutsidis et al. [34] reported a simpler, alternative strategy for redundancy removal within the context of data analysis: given genotypic information on m individuals and r AIMs (either PCAIMs or I_n AIMs), as well as a target panel size k, cluster the r AIMs in k

clusters and return one representative AIM from each cluster. This strategy reduces the redundancy removal problem to a clustering problem, for which efficient and highly accurate software packages are available. An additional advantage of clustering is that instead of returning just a set of k non-redundant AIMs, it also returns k lists (clusters) of AIMs. Within each list, the selected markers are, at least to some extent, interchangeable. This makes the task of interpreting the functionality of selected AIMs easier and also provides some flexibility to researchers that are interested in forming ancestry informative panels to choose alternative markers.

In light of the above discussion, we evaluated clustering via straightforward methods as a solution for the redundancy removal problem. More specifically, we used the publicly available software CLUTOCLUSTER [35] with default parameters. Our metric of similarity was the cosine of the angle between the m-dimensional vectors representing the AIMs, which exactly coincides with the metric of similarity used by Principal Components Analysis. We also compared the performance of clustering for redundancy removal to the method of [29] and found the two methods to perform comparably (data not shown), with clustering being about five times slower but slightly more accurate. This observation, combined with the improved interpretability of clustering, seems to support the conclusions of [34] that clustering is a very useful way of addressing the redundancy removal problem.

Ancestry prediction via Nearest Neighbors

We model ancestry prediction using panels of AIMs as the following task: given a database of m individuals of known (for example self-reported) population of origin, genotyped on a panel of k AIMs, and a new individual of unknown ancestry genotyped on the same panel, we seek to predict the population of origin of the new sample. This is a standard classification problem and in order to address it we chose to use one of the most intuitive methods available in the machine learning literature, namely a Nearest Neighbor (NN) approach. NN-type algorithms first compute the distance of the new sample from the m individuals in the database and then identify the n "nearest neighbors" of the new sample. A majority voting strategy is used in order to assign a population of origin to the new sample. We experimented with different values of n (the number of nearest neighbors) ranging from five up to ten in increments of one without observing a consistent advantage in using any value above five. Thus, we chose to fix n to five; as a result, in order to assign an individual to a population X we necessitate that at least three of its five nearest neighbors belong to the same population X. If such consensus cannot be reached, we do not return a prediction. We will refer from now on to our classification methodology as 5-NN.

Finally, in order to deal with individuals that are far away from the reference populations, we augmented our nearest-neighbor computation with a simple confidence metric discarding nearest neighbors whose distance exceeds the 95% threshold in the distribution of observed distances. Individuals with three or more "discarded" nearest neighbors are classified as unknown.

In almost all our experiments we chose the Identity-by-State (IBS) distance as our metric of similarity. IBS simply measures the number of alleles that agree between the genotypes of the two samples. The only exception is one experiment where we represented the genotypic data of the HGDP individuals by projecting them on the top few eigenSNPs, which results to fractional values for the genotypes. Since the IBS distance was not immediately applicable in this setting, we used a standard generalization, the Euclidean (ℓ_2) distance instead.

To conclude, we note that more advanced classification methodologies and/or better distance metrics might be applicable to our task. It is quite interesting and exciting that standard, simple methods such as 5-NN and IBS are highly accurate and very useful. Finally, details on our validation and cross-validation experiments, as well as our accuracy metrics, are available in Supplementary Methods.

Results

Decomposing the structure of worldwide human populations

We decided to dissect the problem of recovering individual ancestry at a fine scale into hierarchical steps, thus attempting to decompose worldwide human population structure. We performed a detailed investigation of the observed patterns of genetic variation and population relationships in the studied samples as captured by PCA (see plots of projections of the samples on the top eigenSNPs at the online material at `http://www.cs.rpi.edu/~drinep/HGDPAIMS/`). Our aim was to split the available populations into groups and levels in a decision tree until all meaningful information provided by PCA could be extracted. Thus, we clustered populations into groups according to geography; membership to these groups was fine-tuned based on visual inspection of the top 10 PCs for each such group of populations. Our analysis is summarized in the decision-tree for individual assignment to a particular population of Figure 1.

According to the above scheme, and as we will describe in detail in the following sections, individuals are first classified to one of five broad geographic regions: Africa, Europe-Middle East-Central South Asia, East Asia, Oceania, and America. Moving further down in the decision tree, individuals are classified into lower-level nodes and are finally assigned to the deepest cluster that

can be inferred given the HGDP dataset. Depending on the complexity of population structure within each region, as well as the genetic homogeneity of populations sampled for each of these regions, one or more levels may follow the initial assignment of an individual to one of the five broad geographic clusters of the initial node of our tree (World node, Figure 1).

Recall that in this exercise our goal was to first evaluate the extent to which individual ancestry can be inferred using all available genotypes (650,000 SNPs) and, second, to identify and examine small panels of AIMs that can reproduce these results. For example, in our decision-tree (Figure 1), in order to classify an individual as Bantu, we first determine whether the individual originates from the African continent. We then decide whether the individual belongs to one of the "Western African" populations in our sample (Bantu, Yoruba, and Mandenka). We then proceed to differentiate between Mandenka ancestry and Bantu or Yoruba ancestry (the Bantu cannot be easily differentiated from the Yoruba at this step and thus we chose to add an additional node here). Finally, we distinguish between Bantu and Yoruba. The Bantu people represent today a large number of ethnic groups in sub-saharan Africa, extending from Cameroon and across Central Africa to East Africa and Southern Africa. It is hypothesized that they originated in Western Africa (the southwestern border of modern Nigeria and Cameroon) and about 3,000 years ago expanded throughout sub-saharan Africa developing agriculture and metalworking techniques [36]. Our results support their West African origin since they cluster closely (see PCA plots at the online material accompanying this work– `http://www.cs.rpi.edu/~drinep/HGDPAIMS/Level_Africa.html`) with the Yoruba, a population predominantly found in Nigeria.

We should point out that, in most cases, individuals can be ultimately assigned to a single population. However, in some cases, the decision tree stops at a level which corresponds to several populations and further differentiation cannot be adequately achieved, at least using the methods we describe here. This may be either due to the close genetic relationship between these populations or the small sample sizes that were available for study in HGDP, or both. For instance, in Europe, Northern Italians from Bergamo (13 individuals in the studied dataset) cannot be completely distinguished from Tuscans in central Italy (eight individuals in the studied dataset), even when all 650,000 SNPs are used.

Moving to Central South Asia, we cannot achieve successful differentiation between the Balochi, Brahui, and Makrani. All three populations reside in the southeast corner of the Iranian plateau, including parts of Iran, Afghanistan, and Pakistan. In Central South Asia we are also unable to distinguish between the Hazara and the Uygur people. The Hazara, from Pakistan,

are a population of Mongol origin, as also supported by Y-chromosome studies [37, 38]. Many of them consider themselves to be direct male-line descentants of Genghis Khan. The Uygur are a Turkic ethnic group descending from tribes in the Altai Mountains and found today, primarily, at the far northeastern corner of China, in a region bordering Mongolia, Russia, and Pakistan. Finally, the Pathan from the North-West Frontier Province of Pakistan are largely indistinguishable from the Sindhi living close to the delta of the Indus river in Southeastern Pakistan.

In East Asia the HGDP panel includes a large number of different but closely neighboring Chinese ethnic minorities (14 such populations), many of which are only represented by a very small number of individuals (seven to ten individuals for all sampled populations in that region, except the Han Chinese). However, clear gradients are observed, even within China, and several sub-populations can be successfully differentiated. For instance, we can easily differentiate the Dai and Lahu, two populations living at the South of the Yunnan Province in China and also found in the neighboring countries of Laos, Vietnam, Burma, and Thailand. Several of the remaining populations are grouped together based on our analysis. Northern China is represented by the Daur, Hezhen, Oroquen, Mongola, and Xibo populations originating from (or currently inhabiting) Inner Mongolia and the far most Northeastern corner of China. The Naxi and Yizu from Northwestern Yunnan also cluster together in the south of China, as do the Han and Tujia from the Central Provinces of China.

Ancestry inference using our decision tree, 5-NN, and the Illumina 650Y array

First, we discuss ancestry inference using genotype information from the full Illumina 650Y array. While our primary goal was the identification of small panels of AIMs that achieve accurate assignment of individuals to populations of origin using self-reported ancestry in the HGDP dataset as "ground truth", we also ran a complete leave-one-out experiment using all 650K available markers in order to assess ancestry inference using all SNPs. More specifically, Figure 2A (see the dark blue bar corresponding to the 650K panel) summarizes the results of the complete leave-one-out validation experiment when applied at each level of our decision tree (Figure 1) using the number of significant principal components of Table 1 and the 5-NN approach described in Methods (see supplementary material for a detailed description of how the number of significant principal components was selected for the purposes of the analysis presented here). As indicated in Figure 2A, at most nodes of our decision tree the classification accuracy was 100%.

The classification accuracy was less than 100% only at the node differentiating among the three Chinese ethnic sub-groups (with a classification accuracy of 98.63%), the node differentiating two groups of populations from Pakistan (97.35%), and the node differentiating four populations from the Middle East (94.19%). As can be deduced from the inspection of PCA plots for the Middle Eastern populations (see the 2-D and 3-D PCA plots for Middle Eastern populations at `http://www.cs.rpi.edu/~drinep/HGDPAIMS/Level_MiddleEast.html`), the loss of performance in this region is mainly due to the fact that the studied individuals of Druze and Palestinian origin overlap.

Inferring individual ancestry with small panels of AIMs

In our next experiment we evaluated whether small panels of AIMs can accurately reproduce the excellent results of ancestry inference using all 650K available markers. As a first step, we selected within each level the top 5,000 PCAIMs, using the number of significant principal components of Table 1, and repeated the full leave-one out validation test at each level using the 5-NN approach and the standard IBS distance metric. Once more, a complete leave-one-out validation experiment is computationally feasible since the top 5,000 PCAIMs are selected once at each node of the decision tree and only one SVD at each node is needed for the whole experiment. The light blue bars in Figure 2A indicate the performance of these panels: with three notable exceptions that will be discussed below, they are marginally (no more than 5%) less accurate than the full 650K panels. However, at the Chinese node, we have a considerable performance loss in classification accuracy (approx. 10%). Similarly, at the Bantu-Yoruba node, we have a performance loss of approximately 17%. Finally, at the European populations node, we observe the largest performance loss, with the classification accuracy reduced to 78.21% from 100%. This is still surprisingly accurate, especially given the genetic homogeneity of European populations. Supplementary figure 1A (light blue bars) shows that even the smallest average number of correct nearest neighbors is almost four out of five, which is a strong indication that our 5-NN approach works well with the selected SNP panel.

We already observed in prior work [29] that such panels of AIMs tend to contain large amounts of redundant markers, mainly due to LD between densely typed markers. Thus, our next step was the removal of redundant markers via the clustering technique described in Methods. We experimented with numerous panel sizes and we chose to report results on three different panels (P1, P2, and P3) for each node in our decision tree. The panel sizes are connected: the number of markers in P2 is equal to twice the number of markers in P1, and the number of markers in P3

is equal to three times the number of markers in P1 (Table 1). Not surprisingly, the number of markers necessary for ancestry inference is very different at the various nodes of the decision tree, reflecting the fact that certain (groups of) populations are more or less genetically homogeneous. For example, by inspecting Figure 2A and Table 1, we conclude that, within the setting of this experiment, 50 SNPs suffice to classify an individual to one of the five broad geographic regions at the topmost node of our decision tree with an accuracy of 98.9%. Individuals who fall within the Europe-Middle East-Central South Asia cluster can be further assigned to one of these three regions using an additional 300 SNPs (98.7% accuracy). Thus, this experiment indicates that 350 SNPs achieve almost perfect classification accuracy in assigning individuals to one of the seven broad geographic regions sampled in the HGDP.

A few interesting observations arise by inspecting Figure 2A. First, even our smallest panels of AIMs (panels P1) achieve very high accuracy at most nodes of our decision tree. A notable loss of performance is observed when attempting to classify European individuals. In this case, using 300 markers we can achieve 65.4% classification accuracy, which improves to 73.7% using 900 markers. This is still well below the 100% accuracy that is achieved using all 650,000 SNPs, but quite close to the 78.2% accuracy that is achieved using the top 5,000 PCAIMs. Much less dramatic losses in accuracy (not exceeding 5%) are observed for the closely related Chinese populations (with 500 markers achieving 84.2% classification accuracy, and 1,500 markers achieving 87.7% classification accuracy), as well as the Middle Eastern and, to a lesser degree, the East Asian populations. We also observe that, in general, our largest panels (panels P3) perform as well as the top 5,000 PCAIMs before the redundancy removal step. This seems to reinforce the conjecture that redundancy removal from the top PCAIMs does not significantly affect performance. One additional observation is the improvement in classification accuracy using the non-redundant panels in the Bantu-Yoruba case, which is probably due to artifacts related to the removal of a large number of uninformative SNPs from the initial panel.

We repeated this experiment using AIMs selected based on the metric of I_n. Figure 2B and Supplementary Figure 1B show that, in this exploratory experiment where AIMs are selected using all available individuals, all I_n AIMs panels achieve very high accuracy for population assignment around the world, even within Europe, China, and the Middle East, where population differentiation with the PCAIM panels was less accurate. Two observations are immediate: first, when comparing Figures 2A and 2B, the accuracy trends are quite similar. In particular, Europe and China are the worst performing regions in both cases, especially when using the small panels P1, P2, and P3. Middle East is also more complex, as we explained above. The second

observation is that the better performance of the I_n SNPs in this case is due to the fact that it is a supervised method and thus, in the setting of this experiment at least, it probably overfits the data and selects exactly the SNPs that differentiate the various populations. Unfortunately, this superior performance is not a good predictor of the generalization error in a true cross-validation experiment, where certain individuals are left out during the selection of AIMs. Indeed, as we shall see in the next section, in a true cross-validation setting the performance of I_n SNP panels drops and the resulting panels have comparable or somewhat worse performance than the PCAIMs panels.

Cross-validation experiments

Leave-seven-out cross-validation using the HGDP dataset

In our first cross-validation experiment we performed 50 splits of the HGDP dataset, where in each split we constructed a test set consisting of seven individuals from HGDP, one from each of the seven broad geographic regions (Africa, Europe, Middle East, South Central Asia, East Asia, Oceania, and America). The remaining individuals were used as a training set in order to select PCAIMs and I_n AIMs, as well as input for our 5-NN classification scheme. In each split, care was taken in order to avoid testing the same individual twice. (This was not possible in Oceania, where the number of available individuals was less than 50.) Figure 3A and Supplementary Figures 2A, 3, 4, and 7 summarize the performance of our PCAIM panels over all $50 \times 7 = 350$ individuals in all test sets, while Figure 3B and Supplementary Figures 2B, 5, 6, and 8 demonstrate the corresponding results for I_n AIMs.

The overall performance of our approach using even small panels of PCAIMs is quite remarkable at most nodes of the decision tree of Figure 1, even when fine scale population differentiation is the target, both using PCAIMs and I_n AIMs. In almost every case, results are comparable to the validation results presented in the previous section. However, in three regions, we do observe a loss in performance. The separation of Bantu and Yoruba individuals proves more difficult. Using PCAIMs the classification accuracy ranges from 64% to 72% depending on the panel size (50 up to 150 PCAIMs) while it lies between 57% and 86% with I_n AIMs. Using PCAIMs within East Asia, we are able to achieve a classification accuracy ranging between 82% (using 300 SNPs) up to (almost) 90% (using 900 SNPs), while I_n AIMs reach 85% with 600 SNPs. Within this region, a group of closely related Chinese populations are the most resistant to prediction, with accuracies that do not exceed 61% even using 1,500 PCAIMs (57% with I_n AIMs). In the Middle East the classification accuracy using PCAIMs panels ranges from 75% (using 300 SNPs) up to

80% (using 900 SNPs), with comparable results for I_n AIMs. As mentioned earlier here, this is due to the difficulty in distinguishing between the Druze and Palestinians, who cannot be separated even using all 650,000 SNPs.

In Europe, recall that 650,000 SNPs achieve essentially 100% classification accuracy, while with 5,000 PCAIMs (selected using all available Europeans as training set) the accuracy drops to 78.2%. In our cross-validation experiment, the classification accuracy ranges from 51% using 300 PCAIMs up to 68% using 900 PCAIMs. Notice that for our PCAIMs analysis, these numbers agree fairly well with the numbers reported in Figure 2 for our validation experiment. On the contrary, although the I_n AIMs seemed to perform better at this stage in the validation experiment, in this leave-seven-out crossvalidation test, their performance is actually comparable to PCAIMs, indicating that I_n AIMs overfit the data. Supplementary Figure 4 illustrates in further detail the performance of our smallest and largest PCAIM panels (P1 and P3) in Europe (see Supplementary Figure 6 for corresponding results for I_n AIMs). The Russian, Adygei, Sardinian, and Basque individuals in our test sets are predicted very accurately. The main source of classification error is the difficulty in distinguishing between the French, Orcadian, and Italians, when small panels of SNPs are used. Thus, although individual classification is perfect for these populations when all 650,000 SNPs are used, our panels (combined with 5-NN and the IBS distance metric) are largely unable to distinguish between these three populations, although some percentage of correct ancestry is recovered. Either larger panels or more advanced methods are needed in order to increase the accuracy of ancestry prediction for individuals from these populations.

Predicting the population of origin of the HapMap Phase 3 populations

In our second cross-validation experiment we measured the performance of the SNP panels derived using the full HGDP data as training set in order to classify individuals from the HapMap Phase 3 populations. This experiment involves the analysis of populations that represent extreme cases of admixture not seen in the HGDP panel (the ASW, of African ancestry in southwest USA, and the MEX, of Mexican ancestry in Los Angeles, California). Furthermore, indigenous populations for which no reference exists in the HGDP sample are also included: the Luhya in Webuye, Kenya (LWK), the Maasai in Kinyawa, Kenya (MKK), and the Gujarati Indians from Houston, Texas (GIH). We extracted the genotypes for all 11 populations from HapMap release 27 (built 36) raw data and then used our 5-NN classifier at the relevant nodes of the decision tree of Figure 1. Results for the three panel sizes P1, P2, and P3 of Table 1, are reported in

Tables 2 and 3. In almost all cases, the performance of PCAIM SNP panels and I_n SNP panels is comparable, and thus we focus our discussion on the classification results using PCAIMs.

Using our PCAIM SNP panels, the Yoruba are very accurately classified (see Table 2) all the way down to their self-reported population of origin. In fact, when the larger panels (P2 and P3) are used, we achieve essentially 100% classification accuracy all the way down to the last node in the decision tree. At that node (the Bantu-Yoruba node) the accuracy drops slightly but still exceeds 94% with the two larger panels and 90% with the smallest panel. In Europe, the CEU samples are perfectly assigned to the continent, even with the smallest panel, as Table s indicates. The exact European population of origin for the CEU samples is not known. However, not surprisingly, given prior work on the CEU samples [10, 9] which indicates that their ancestry is closer to northwestern European populations, the vast majority of the CEU individuals are classified as French or Orcadian; these are our most Northwestern European HGDP populations. The origin of the TSI individuals is predicted as French or Italian. As described in the previous section, Italians can be classified as French or Italian with 100% classification accuracy using panel P1; this accuracy drops to 50% when we seek to classify them as Italians only using the same panel.

Results for the three East Asian populations are shown in Table 2. Regarding the CHB sample, our largest PCAIMs panel almost perfectly assigns the HapMap CHB subjects to the Han/Tujia group, with accuracy exceeding 94% even at the last relevant node. Even our smallest panel does a good job predicting the population of origin with almost 100% accuracy at the first two nodes, and a 77% accuracy at the last node. The CHD are also accurately assigned to the Chinese group. The situation is slightly worse for the Japanese sample, where we observed the least successful performance in this cross-validation experiment. While we easily achieve 100% accuracy in classifying the Japanese samples to East Asia with all three panels, within East Asia we only achieve 65% accuracy in assigning the JPT samples to the Japanese population even with our largest PCAIMs panel. The remaining 35% of the JPT samples are assigned to Chinese populations. This seems to indicate that more markers are necessary at this level; it is worth noting that using all 650,000 SNPs we achieve 98% accuracy in classifying the JPT samples as Japanese. Using our largest panel (P3), 56 out of the 86 studied JPT individuals have three or more of their five nearest neighbors in the HGDP Japanese sample, an additional 19 have two of their five neighbors in the HGDP Japanese sample, seven have one of their five neighbors in the HGDP Japanese sample, and only four have no neighbors in the HGDP Japanese sample. This seems to indicate that we can at least partially capture Japanese ancestry for the vast majority

of the HapMap JPT samples. We should also note, that this is one node of the decision tree, where the performance of I_n SNPs is considerable worse than PCAIMs (up to 43% classification accuracy with the largest panel).

Moving on to the populations for which we have no reference in the HGDP dataset, we observe that, in most cases, individuals are classified to the closest geographically neighboring population available (Table 3). The individuals from populations of African ancestry (ASW, LWK, and MKK) are assigned to Africa with essentially 100% accuracy, even with our smallest panel (50 SNPs, P1). The predominantly Western African origin of African Americans is also well documented, and, appropriately, our panels also classify them as West African. The Bantu speaking LWK from Kenya are classified as Bantu or Yoruba, as are the MKK from Kenya. About two thirds of the population in Kenya is represented by Bantu tribes, which is the closest neighboring population in the HGDP panel to the MKK and LWK. The Gujarati Indians (GIH), originating from Gujarat (the most western state of India and immediately adjacent to Pakistan) are easily placed in Central South Asia where they are classified as Pakistanis. Finally, the analysis of Mexicans from California (MEX) yields slightly unexpected results. With the smallest PCAIM panels, 14 individuals are assigned to Europe, 22 are assigned to America, and 39 to Central South Asia (where they are ultimately classified as Pakistani or Afghans). Although this case illustrates the limitations of the method for extremely admixed populations, it could also indicate a higher than expected degree of Punjabi Mexican American ancestry in this Californian Mexican sample. The Punjabi Mexican Americans (people of Mexican and Pakistani or Indian ancestry) originate from the Sacramento valley in California [39].

Studying the PCA plots for these "untested" populations (Supplementary figures 9-19) it is clear that individual genetic distances are often far from any one of our HGDP reference populations. In order to address this issue, we decided to apply a more stringent test, introducing a simple confidence metric for our assignments. More specifically, our confidence metric discards the nearest neighbors (out of the top five) of an individual, for which the corresponding distance is an outlier (exceeds the standard 95% threshold) in the distribution of observed distances (see Supplementary Methods for details). Individuals with three or more "discarded" nearest neighbors are classified as unknown. Results for the HapMap populations when this confidence threshold is applied at the top two levels of the decision tree in Figure 1 are shown in Table 4 and Supplementary Table 1. The power of this methodology as well as the fact that certain HapMap phase 3 populations have no reference samples in HGDP is highlighted with the majority of the individuals classified as unknown, sometimes even at the topmost level.

Discussion

Despite the relatively low levels of genetic differentiation among geographically defined human populations when compared to other mammalian species, population genetics analysis can uncover the genetic signatures left on regional populations, by demographic, environmental, and historical factors [40, 41, 42, 43, 44]. In this work, we investigated the extent to which geographically close populations can be discerned based on genetic information alone. To this end, we analyzed data from 1043 individuals and 51 populations from around the world, genotyped for 650,000 SNPs (HGDP dataset) [8]. In doing so, we undertook a comprehensive evaluation of population genetic structure around the world. We report on the feasibility of fine scale genetic ancestry testing on a global scale, not only using information from the whole genome (650,000 SNPs) but, importantly, evaluating the performance of small panels of judiciously selected genetic markers. In order to tackle this challenging task we propose a hierarchical decomposition of worldwide human population structure. A decision tree is formed, thus enabling the step-wise assignment of individuals to their region and, ultimately, population of origin, as well as the sequential selection of subsets of genetic markers that can be used for ancestry inference. Moving through the proposed decision tree investigators have the opportunity to tailor their needs for marker selection according to the desired level of resolution and/or prior information on the origin of the samples under study.

Through this scheme we achieve very accurate prediction of individual ancestry when this particular set of 51 populations is considered. We should point out that (see Results) in some cases our decision tree stops at a group of two or three populations and further differentiation cannot be adequately achieved, at least with the methods proposed here. So, for instance, using the given dataset we cannot and do not seek to distinguish between Italians and Tuscans or among Northern Chinese populations. The fact that in some cases certain populations are indistinguishable from one another and are grouped together (even when information from all 650,000 SNPs is used), could be due to the close genetic relationship among these populations or the small sample sizes that are included in this dataset. Thus, our decision tree provides insight into the level of population differentiation that can be achieved using the HGDP as reference if all 650,000 SNPs are used. In fact, we demonstrate that the accuracy of prediction of individual ancestry at a fine scale is essentially 100% for the targets that we propose.

Seeking to identify those markers that actually capture population genetic structure we evaluated two different methods for the selection of AIMs: a PCA-based method that we have

previously described [18] and a method based on the Informativeness for Assignment (I_n) metric [28]. In previous work, PCA has been used to summarize the vast amounts of information in the dataset we studied here by projecting individual genotypic variation into a low-dimensional space [8, 16]. Determining the number of significant PCs in a dataset is a crucial step in our analysis and represents a challenging research topic in the statistics and numerical analysis literature. For the complete dataset analyzed here, Biswas et al. [16] recently reported that 18 PCs are significant if a parametric method based on the Tracy-Widom statistic is used. The number of statistically significant PCs changes when different subsets of the data are analyzed. For the purposes of this study our stepwise view of human genetic variation around the world as a decision tree greatly simplified our search for significant PCs. At every level of our decision tree we only need to retain those PCs that are deemed useful for population differentiation at the targeted level of resolution. This allowed better interpretability of the significant PCs for each subset of populations under study.

Performing a thorough cross-validation experiment, we tested individuals from each one of the studied populations, and showed that in most cases that were analyzed here the number of genetic markers needed for ancestry inference can be successfully reduced to less than 0.1% of the original 650Y Illumina array while retaining high accuracy in ancestry prediction. This reduction could be achieved using a clustering-based redundancy removal algorithm. Such techniques seem extremely promising in addressing the problem of removing redundant markers from a dataset since they allow interchangeability and interpretability of the SNPs that fall in the same cluster. However, in three geographic regions, we were faced with great difficulty in reducing the number of genetic markers needed for ancestry prediction. In Europe, even when 5,000 carefully selected SNPs were used (either PCAIMs or I_n SNPs) the performance of our test dropped below 80%. Still, this is a rather remarkable result given the genetic homogeneity of European populations. Although the North to South axis of genetic variation in Europe can be easily recovered with a few hundred markers [45, 46, 47, 48, 29], this does not seem to be the case for accurate assignment to particular populations. In a similar fashion, when information from small sets of SNPs is used, it is very difficult to differentiate between the Yoruba and the Bantu population or among the closely related Chinese ethnic groups.

The highly accurate results that we describe for the HGPD samples are partly due to the reasonable separation (both geographic and genetic) between most HGDP populations. For example, if we were provided with a larger sample of overlapping European populations, then the classification accuracy using a stringent voting strategy in our leave-one-out experiment would

undoubtedly diminish. Furthermore, the HGDP project targeted clearly defined populations of anthropological interest that were established prior to the great diasporas of the 15th and 16th centuries [14]. Thus, although we have shown that a simple majority voting scheme will actually predict individual ancestry with accuracy close to 100% in most cases, it may be preferable to report the ancestry of all five nearest neighbors. An additional deficiency of the HGDP as reference panel for genetic ancestry testing is the fact that it mostly comprises of small sample sizes for each population, which undoubtedly do not capture the full amount of variation in a given population. Larger samples and more detailed sampling will allow a better representation of population genetic variation and will further increase the classification accuracy.

Our second cross-validation experiment focused on the HapMap phase 3 populations aiming to place a test individual within its actual population of origin, or to the closest possible neighboring population, if no actual reference existed in the training dataset. Naturally, the populations included in the training set, influence the outcome of the test (for instance, all HapMap Japanese, would be classified as Chinese, if no Japanese reference population existed in the HGDP). However, we have also shown, that our methods can be fine-tuned to include flags for samples for which the reference dataset cannot provide information. We introduced a simple confidence metric similar to the algorithms proposed in [11]. More elaborate metrics that depend on the end purposes of the ancestry test are an interesting topic for future work. Using our confidence metric, individuals from two admixed populations that we studied (Mexican and African Americans) would be classified as unknown. Nevertheless, had these populations been included in the reference set, it would be easy to differentiate them, as evidenced from PCA plots (see Supplementary material).

It is very interesting to consider the chromosomal location and possible function of the polymorphisms that are selected to capture human genetic variation in such detail. In fact, population differentiation has been considered as an indication of selective pressure, and several genome scans for natural selection were based on the identification of loci that appear as outliers in empirical distributions of genotypic patterns (see [49, 50, 51, 6, 52, 53] among others). As Pickrell et al. [54] have recently pointed out, even though one would expect patterns of loci that are under natural selection to be distinct from neutral variation (with demography operating on the whole genome rather than on a few loci), it is often the case that putatively selected loci conform to the geographic patterns that are characteristic of neutrality. It follows, that distinguishing between demographic forces and natural selection is extremely difficult. However, observing the lists of PCAIMs that we propose, the top SNPs reside in ge-

nomic locations which constitute prominent candidates for natural selection. For instance, in our panel for differentiating individuals to five broad geographic regions (World node in Figure 1), the top two scoring SNPs sit in a region that has been previously suggested as a candidate for natural selection by multiple genomic scans, no more than 30Kb from SLC24A5, a gene which is known for its involvement in skin pigmentation [55, 56, 57]. Interestingly, these two SNPs exhibit the same patterns of variation as two SNPs in the EDAR gene appearing immediately after them in our top cluster (see online supporting material and lists of AIMs at `http://www.cs.rpi.edu/~drinep/HGDPAIMS/Level_World.html`). EDAR is another well-known candidate for natural selection, responsible for hair-follicle formation [58, 59]. In a similar fashion, the top places in the list of PCAIMs for European populations, are occupied by five SNPs (see `http://www.cs.rpi.edu/~drinep/HGDPAIMS/Level_Europe.html`) that are located in the 2.4 Mb region of the selective sweep that has been associated with the LCT gene [58]. Such analysis can provide useful clues on the possible functional role of the SNPs that we have selected as informative for the fine-scale inference of population ancestry.

The sets of markers that we have identified (see `http://www.cs.rpi.edu/~drinep/HGDPAIMS/`) and the methodology that we introduce have important implications in many different settings, ranging from the study of population history to the elucidation of the genetic background of common complex disorders. In all of these cases, the markers that we have identified can greatly reduce genotyping costs, reducing to less than 0.1% of the original 650,000 SNPs, the number of markers needed in order to assign an individual to a particular population of origin, or simply place this individual within the axes of variation seen in the reference dataset. In evolutionary genetics studies, the proposed marker sets can be used to investigate the relationship between extant populations. In medical genetics, they can be used to inform patients about disease risks associated with different ethnic groups. In genetic association studies, multi-national studies have proven essential and two-stage study designs are often followed, with additional samples being genotyped only at those loci that have proven promising in an initial genomewide association study [60, 61, 62]. In such settings the markers we propose can be used to correct for the biases introduced by population stratification. Furthermore, the methods described herein in combination with the increasingly more comprehensive databases of human genetic variation that are becoming available, open new horizons for the potential of forensic and commercial genetic ancestry testing. Genetic ancestry testing in commercial settings has received considerable attention recently, and several policy forums and sets of guidelines have been published [63, 64, 65]. In such settings, we would like to stress the importance of actively informing and counseling

clients that seek hints about the origin of their ancestors, not only about the possibilities, but also about the caveats and limitations of different methodologies and reference datasets.

In summary, we have described a comprehensive study of the level of resolution that can be achieved for genetic ancestry inference in worldwide populations when small panels of genetic markers are used. The HGDP as a reference dataset provides the most complete catalogue of worldwide human genetic variation to date [8]. However, the full extent of human genetic variation will only be appreciated through the analysis of larger and geographically more dense samples of populations. This is becoming a reality through the concerted efforts of investigators studying European populations [9, 10, 66, 11] and, more recently, African populations [15] and populations from the Indian subcontinent [67]. Furthermore, since most polymorphisms included in commercial genotyping chips were actually ascertained in European populations, a large proportion of human genetic variation remains undiscovered. The feasibility of large scale complete sequencing of large numbers of samples will combat this deficiency, allowing unbiased genotyping of diverse populations. It is worth noting that our methods are readily extensible to larger or more detailed samples and that our decision tree can be used as a starting point for future studies.

Acknowledgements

This work was supported, in part, by an NSF CCF 0447950 CAREER award to PD; an NSF CCF 0824684 award to PD; an EMBO ASTF 235-2009 Short Term Fellowship to PD; and two Tourette Syndrome Association (TSA) Research Grant Awards to PP.

Web Resources

Human Genome Diversity Panel genotypes - `http://hagsc.org/hgdp/`
HapMap database - `http://www.hapmap.org/`
Online supporting material - `http://www.cs.rpi.edu/~drinep/HGDPAIMS/` (please respect capitalization in order to access this webpage)

References

[1] Stringer CB, Andrews P (1988) Genetic and fossil evidence for the origin of modern humans. Science 239:1263–1268.

[2] Jobling M, Hurles M, Tyler-Smith C (2003) Human Evolutionary Genetics: Origins, People, and Disease. Garland Science .

[3] Cavalli-Sforza LL, Menozzi P, Piazza A (1994) The History and Geography of Human Genes. Princeton, New Jersey: Princeton University Press, Princeton.

[4] Rosenberg N, Pritchard J, Weber J, Cann H, Kidd K, et al. (2002) Genetic structure of human populations . Science 298:2381–2385.

[5] The International HapMap Consortium (2003) The International HapMap Project. Nature 426:789–796.

[6] The International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437:1299–1320.

[7] Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851–861.

[8] Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, et al. (2008) Worldwide human relationships inferred from genome-wide patterns of variation. Science 319:1100–1104.

[9] Lao O, Lu TT, Nothnagel M, Junge O, Freitag-Wolf S, et al. (2008) Correlation between genetic and geographic structure in Europe. Curr Biol 18:1241–1248.

[10] Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, et al. (2008) Genes mirror geography within Europe. Nature 456:98–101.

[11] Heath SC, Gut IG, Brennan P, McKay JD, Bencko V, et al. (2008) Investigation of the fine structure of European populations with applications to disease association studies. Eur J Hum Genet 16:1413–1429.

[12] McEvoy BP, Montgomery GW, McRae AF, Ripatti S, Perola M, et al. (2009) Geographical structure and differential natural selection among North European populations. Genome Res 19:804–814.

[13] Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, et al. (2002) A human genome diversity cell line panel. Science 296:261–262.

[14] Cavalli-Sforza LL (2005) The Human Genome Diversity Project: past, present and future. Nat Rev Genet 6:333–340.

[15] Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, et al. (2009) The genetic structure and history of Africans and African Americans. Science 324:1035–1044.

[16] Biswas S, Scheinfeldt LB, Akey JM (2009) Genome-wide insights into the patterns and determinants of fine-scale population structure in humans. Am J Hum Genet 84:641–650.

[17] Price A, Patterson N, Plenge R, Weinblatt M, Shadick N, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies . Nat Genet 38:904–909.

[18] Paschou P, Ziv E, Burchard EG, Choudhry S, Rodriguez-Cintron W, et al. (2007) PCA-correlated SNPs for structure identification in worldwide human populations. PLoS Genet 3:1672–86.

[19] Campbell C, Ogburn E, Lunetta K, Lyon H, Freedman M, et al. (2005) Demonstrating stratification in a European American population . Nat Genet 37:868–872.

[20] Tang H, Quertermous T, Rodriguez B, Kardia S, Zhu X, et al. (2005) Genetic structure, self-identified race/ethnicity, and confounding in case-control association studies . Am J Hum Genet 76:268–275.

[21] Wright S (1951) The genetical structure of populations. Ann Eugen 15:323–354.

[22] Dean M, Stephens J, Winkler C, Lomb D, Ramsburg M, et al. (1994) Polymorphic admixture typing in human ethnic populations . Am J Hum Genet 55:788–808.

[23] McKeigue P (1998) Mapping genes that underlie ethnic differences in disease risk: methods for detecting linkage in admixed populations, by conditioning on parental admixture. Am J Hum Genet 63:241–251.

[24] Parra E, Marcini A, Akey J, Martinson J, Batzer M, et al. (1998) Estimating African American admixture proportions by use of population-specific alleles . Am J Hum Genet 63:1839–1851.

[25] Collins-Schramm H, Phillips C, Operario D, Lee J, Weber J, et al. (2002) Ethnic-difference markers for use in mapping by admixture linkage disequilibrium . Am J Hum Genet 70:737–750.

[26] Pfaff C, Barnholtz-Sloan J, Wagner J, Long J (2004) Information on ancestry from genetic markers . Genet Epidemiol 26:305–315.

[27] Weir B, Cardon L, Anderson A, Nielsen D, Hill W (2005) Measures of human population structure show heterogeneity among genomic regions . Genome Res 15:1468–1476.

[28] Rosenberg N, Li L, Ward R, Pritchard J (2003) Informativeness of genetic markers for inference of ancestry . Am J Hum Genet 73:1402–1422.

[29] Paschou P, Drineas P, Lewis J, Nievergelt CM, Nickerson DA, et al. (2008) Tracing sub-structure in the European American population with PCA-informative markers. PLoS Genet 4:e1000114.

[30] Menozzi P, Piazza A, Cavalli-Sforza L (1978) Synthetic maps of human gene frequencies in Europeans . Science 201:786–792.

[31] Shriver M, Mei R, Parra E, Sonpar V, Halder I, et al. (2005) Large-scale SNP analysis reveals clustered and continuous patterns of human genetic variation . Hum Genomics 2:81–89.

[32] Patterson N, Price A, Reich D (2006) Population Structure and Eigenanalysis . PLoS Genet 2:e190.

[33] Liu N, Zhao H (2006) A non-parametric approach to population structure inference using multilocus genotypes . Hum Genomics 2:353–364.

[34] Boutsidis C, Sun J, Anerousis N (2008) Clustered subset selection and its applications on it service metrics. In: ACM Conference on Information and Knowledge Management (CIKM).

[35] Zhao Y, Karypis G (2002) Evaluation of hierarchical clustering algorithms for document datasets. In: ACM Conference on Information and Knowledge Management (CIKM). pp. 515–524.

[36] Beleza S, Gusmo L, Amorim A, Carracedo A, Salas A (2005) The genetic legacy of western Bantu migrations. Hum Genet 117:366–375.

[37] Qamar R, Ayub Q, Mohyuddin A, Helgason A, Mazhar K, et al. (2002) Y-chromosomal DNA variation in Pakistan. Am J Hum Genet 70:1107–1124.

[38] Zerjal T, Xue Y, Bertorelle G, Wells RS, Bao W, et al. (2003) The genetic legacy of the Mongols. Am J Hum Genet 72:717–721.

[39] Leonard K (1989) California's Punjabi Mexican Americans. Ethnic choices made by the descendants of Punjabi pioneers and their Mexican wives. The World & I 4:612–623.

[40] Jorde LB, Watkins WS, Bamshad MJ, Dixon ME, Ricker CE, et al. (2000) The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data. Am J Hum Genet 66:979–988.

[41] Bamshad M, Wooding S, Salisbury BA, Stephens JC (2004) Deconstructing the relationship between genetics and race. Nat Rev Genet 5:598–609.

[42] Tishkoff SA, Kidd KK (2004) Implications of biogeography of human populations for 'race' and medicine. Nat Genet 36:S21–27.

[43] Coop G, Pickrell JK, Novembre J, Kudaravalli S, Li J, et al. (2009) The role of geography in human adaptation. PLoS Genet 5:e1000500.

[44] Nielsen R, Hubisz MJ, Hellmann I, Torgerson D, Andrs AM, et al. (2009) Darwinian and demographic forces affecting human protein coding genes. Genome Res 19:838–849.

[45] Seldin M, Shigeta R, Villoslada P, Selmi C, Tuomilehto J, et al. (2006) European population substructure: clustering of northern and southern populations . PLoS Genet 2:e143.

[46] Bauchet M, McEvoy B, Pearson L, Quillen E, Sarkisian T, et al. (2007) Measuring European Population Stratification with Microarray Genotype Data . Am J Hum Genet 80:948–956.

[47] Price A, Butler J, Patterson N, Capelli C, Pascali V, et al. (2008) Discerning the ancestry of European Americans in genetic association studies. PLoS Genet 4:e236.

[48] Tian C, Plenge R, Ransom M, Lee A, Villoslada P, et al. (2008) Analysis and application of European genetic substructure using 300 K SNP information. PLoS Genet 4:e4.

[49] Akey JM, Zhang G, Zhang K, Jin L, Shriver MD (2002) Interrogating a high-density SNP map for signatures of natural selection. Genome Res 12:1805–1814.

[50] Kayser M, Brauer S, Stoneking M (2003) A genome scan to detect candidate regions influenced by local natural selection in human populations. Mol Biol Evol 20:893–900.

[51] Shriver M, Kennedy G, Parra E, Lawson H, Sonpar V, et al. (2004) The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs . Hum Genomics 1:274–286.

[52] Myles S, Tang K, Somel M, Green RE, Kelso J, et al. (2008) Identification and analysis of genomic regions with large between-population differentiation in humans. Ann Hum Genet 72:99–110.

[53] Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L (2008) Natural selection has driven population differentiation in modern humans. Nat Genet 40:340–345.

[54] Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, et al. (2009) Signals of recent positive selection in a worldwide sample of human populations. Genome Res 19:826–837.

[55] Lamason RL, Mohideen MA, Mest JR, Wong AC, Norton HL, et al. (2005) SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. Science 310:1782–1786.

[56] Norton HL, Kittles RA, Parra E, McKeigue P, Mao X, et al. (2007) Genetic evidence for the convergent evolution of light skin in Europeans and East Asians. Mol Biol Evol 24:710–722.

[57] Sturm RA (2009) Molecular genetics of human pigmentation diversity. Hum Mol Genet 18:9–17.

[58] Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, et al. (2007) Genome-wide detection and characterization of positive selection in human populations. Nature 449:913–918.

[59] Botchkarev VA, Fessing MY (2005) Edar signaling in the control of hair follicle development. J Investig Dermatol Symp Proc 10:247–251.

[60] McCarthy MI, Hirschhorn JN (2008) Genome-wide association studies: potential next steps on a genetic journey. Hum Mol Genet 17:156–165.

[61] Committee PGCC, Cichon S, Craddock N, Daly M, Faraone SV, et al. (2009) Genomewide association studies: history, rationale, and prospects for psychiatric disorders. Am J Psychiatry 166:540–556.

[62] Cooper RS, Tayo B, Zhu X (2008) Genome-wide association studies: implications for multiethnic samples. Hum Mol Genet 17:151–155.

[63] Bolnick DA, Fullwiley D, Duster T, Cooper RS, Fujimura JH, et al. (2007) Genetics. The science and business of genetic ancestry testing. Science 318:399–400.

[64] ASHG Ancestry Testing Statement, 13 November 2008 (ASHG, Bethesda, MD, 2008); journal= .

[65] Soo-Jin Lee S, Bolnick DA, Duster T, Ossorio P, Tallbear K (2009) Genetics. The illusive gold standard in genetic ancestry testing. Science 325:38–39.

[66] Nelson MR, Bryc K, King KS, Indap A, Boyko AR, et al. (2008) The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. Am J Hum Genet 83:347–358.

[67] Reich D, Thangaraj K, Patterson N, Price AL, Singh L (2009) Reconstructing Indian population history. Nature 461:489–494.

Figure Legends

Figure 1

The decision tree for individual assignment to a particular population (or population group) of origin using the HGDP data. For each diamond-shaped node we propose (small) panels of AIMs that may be used to assign an individual to one of its children nodes. The rows of square-shaped nodes indicate populations (or groups of populations) of origin that we can separate. For example, using the panel that we proposed at the Central South Asia node, we can assign an individual to either the Burusho population, or the Kalash population, or the Hazara-Uygur populations (we cannot distinguish between Hazaras and Uygurs), or to the Pakistani populations. Further differentiation between groups of Pakistani populations is possible using a different panel of AIMs.

Figure 2

Classification accuracy of our complete leave-one-out validation experiment at all nodes of our decision tree. Five different panel sizes are evaluated, with 650K corresponding to the whole Illumina 650Y array, 5K corresponding to the top 5,000 AIMs, and P1, P2, and P3 corresponding to the panel sizes depicted in Table 1; these smaller panels emerged by removing redundant markers from the top 5,000 AIMs. Notice that the top 5,000 markers were selected using the full dataset, in contrast to the crossvalidation experiment of Figure 3. *(A)* Classification accuracy results using all available SNPs as well as PCA Informative Markers (PCAIMs). *(B)* Classification accuracy results using Informativeness for Assignment (I_n) Markers.

Figure 3

Classification accuracy of our leave-7-out crossvalidation experiment at all nodes of our decision tree. Three different panel sizes are evaluated: P1, P2, and P3 (see Table 1). Unlike the experiment of Figure 2, panels P1, P2, and P3 were now selected using only the training set data. Then, our 5-NN classification scheme was applied on the selected markers in order to predict the ancestry of the individuals in the test set. Thus, this experiment is a good predictor of the generalization error of our methods and panels, a.k.a., the error that our methods are expected to have in previously unseen data. *(A)* Classification accuracy results using PCA Informative Markers (PCAIMs) over 50 random splits. *(B)* Classification accuracy results using Informativeness for Assignment (I_n) Markers over 50 random splits.

Tables

Decision Tree node	**sign. PCs**	**Panel P1** # of SNPs	**Panel P2** # of SNPs	**Panel P3** # of SNPs
World	4	50	100	150
Africa	2	30	60	90
Bantu, Mandenka, Yoruba	1	100	200	300
Bantu, Yoruba	1	50	100	150
E Asia	6	300	600	900
Chinese	5	500	1000	1500
America	4	30	60	90
Oceania	1	10	20	30
C S Asia, Europe, M East	2	300	600	900
Europe	2	300	600	900
C S Asia	3	150	300	450
Pakistani	1	100	200	300
M East	6	300	600	900

Table 1: Number of significant Principal Components and AIM panel sizes at each node of the decision tree depicted in Figure 1. Notice that panel P2 contains twice the number of SNPs in panel P1 and panel P3 contains three times the number of SNPs in panel P1.

Decision Tree Nodes	**Panel 1**		**Panel 2**		**Panel 3**	
	PCAIMs	I_n AIMs	PCAIMs	I_n AIMs	PCAIMs	I_n AIMs
HapMap YRI						
World → Africa	167/167	160/160	167/167	167/167	167/167	167/167
Africa → Ban, Man, Yor	166/167	151/160	167/167	158/167	167/167	166/167
Ban, Man, Yor → Ban, Yor	161/166	145/151	166/167	157/158	167/167	164/166
Ban, Yor → Yor	146/161	123/145	158/166	145/157	157/167	157/164
HapMap CEU						
World → Eur, CSA, ME	165/165	165/165	165/165	165/165	165/165	165/165
Eur, CSA, ME → Europe	165/165	165/165	165/165	165/165	165/165	165/165
Europe → Fre & Orc	125/165	58/165	132/165	79/165	128/165	110/165
HapMap TSI						
World → Eur, CSA, ME	88/88	88/88	88/88	88/88	88/88	88/88
Eur, CSA, ME → Europe	85/88	81/88	88/88	87/88	88/88	87/88
Europe → Ita & Fre	72/85	52/81	79/88	71/87	83/88	78/87
HapMap CHB						
World → E Asia	84/84	84/84	84/84	84/84	84/84	84/84
E Asia → Chinese	82/84	81/84	83/84	82/84	84/84	83/84
Chinese → Han & Tuj	63/82	66/81	71/83	70/82	79/84	74/83
HapMap CHD						
World → E Asia	85/85	85/85	85/85	85/85	85/85	85/85
E Asia → Chinese	83/85	84/85	82/85	82/85	83/85	83/85
Chinese → Han & Tuj	75/83	73/84	78/82	79/82	82/83	82/83
HapMap JPT						
World → E Asia	86/86	85/86	86/86	86/86	86/86	86/86
E Asia → Jap	49/86	9/85	56/86	33/86	56/86	37/86

Table 2: Predicting the population of origin of individuals in six HapMap Phase 3 populations that have reference populations in HGDP. We discarded from our analysis individuals with more than 10% missing entries when extracting our PCAIM or I_n SNP panels P1, P2, and P3. We report classification accuracy (C_{ACC}, see Supplementary Methods), expressed as the fraction of individuals that were assigned to the correct region or population of origin at the respective node of the decision tree. Note that as we move down in the decision tree individuals that were incorrectly predicted in previous nodes are omitted.

Decision Tree Nodes	**Panel 1**		**Panel 2**		**Panel 3**	
	PCAIMs	I_n AIMs	PCAIMs	I_n AIMs	PCAIMs	I_n AIMs
HapMap ASW						
World → Africa	76/83	77/83	80/83	81/83	81/83	82/83
Africa → Ban, Man, Yor	76/76	73/77	80/80	80/81	81/81	81/82
HapMap LWK						
World → Africa	90/90	90/90	90/90	90/90	90/90	90/90
Africa → Ban, Man, Yor	89/90	83/90	90/90	80/90	90/90	87/90
Ban, Man, Yor → Ban, Yor	89/89	83/83	90/90	80/80	90/90	87/87
HapMap MKK						
World → Africa	169/171	168/171	171/171	171/171	171/171	171/171
Africa → Ban, Man, Yor	162/169	158/168	171/171	162/171	171/171	165/171
Ban, Man, Yor → Ban, Yor	160/162	153/158	171/171	157/162	171/171	163/165
HapMap GIH						
World → Eur, CSA, ME	88/88	87/88	88/88	88/88	88/88	88/88
Eur, CSA, ME → CSA	88/88	87/87	88/88	88/88	88/88	88/88
CSA → Pakistani	71/88	71/87	78/88	68/88	74/88	77/88
HapMap MEX						
World → Americas	22/77	26/77	26/77	36/77	28/77	39/77
Americas → Maya	19/22	22/26	24/26	26/36	27/28	28/39
World → Eur, CSA, ME	54/77	43/77	51/77	38/77	49/77	34/77
Eur, CSA, ME → Europe	14/54	11/43	16/51	16/38	15/49	17/34
Eur, CSA, ME → CSA	39/54	29/43	33/51	19/38	34/49	17/34
CSA → Pakistani	12/39	16/29	12/33	7/19	6/34	4/17

Table 3: Predicting the population of origin of individuals in five HapMap Phase 3 populations that do not have reference populations in HGDP. We discarded from our analysis individuals with more than 10% missing entries when extracting our PCAIM or I_n SNP panels P1, P2, and P3. We report classification accuracy (C_{ACC}, see Supplementary Methods), expressed as the fraction of individuals that were assigned to the correct region or population of origin at the respective node of the decision tree. Again, as we move down in the decision tree individuals that were incorrectly predicted in previous nodes are omitted.

Decision Tree Nodes	**Panel 1**		**Panel 2**		**Panel 3**	
	PCAIMs	I_n AIMs	PCAIMs	I_n AIMs	PCAIMs	I_n AIMs
HapMap ASW						
World → Africa	9/83	3/83	8/83	0/83	3/83	2/83
Africa → Ban, Man, Yor	9/9	3/3	8/8	0/0	3/3	2/2
HapMap LWK						
World → Africa	53/90	47/90	60/90	46/90	49/90	32/90
Africa → Ban, Man, Yor	53/53	41/47	60/60	42/46	49/49	30/32
Ban, Man, Yor → Ban, Yor	53/53	41/41	60/60	42/42	49/49	30/30
HapMap MKK						
World → Africa	10/171	10/171	7/171	0/171	2/171	1/171
Africa → Ban, Man, Yor	9/10	10/10	7/7	0/0	2/2	1/1
Ban, Man, Yor → Ban, Yor	9/9	10/10	7/7	0/0	2/2	1/1
HapMap GIH						
World → Eur, CSA, ME	45/88	46/88	57/88	36/88	38/88	32/88
Eur, CSA, ME → CSA	43/45	45/46	56/57	36/36	38/38	32/32
CSA → Pakistani	33/43	38/45	50/56	25/36	32/38	28/32
HapMap MEX						
World → Americas	2/77	2/77	0/77	0/77	0/77	0/77
Americas → Maya	1/2	1/2	0/0	0/0	0/0	0/0
World → Eur, CSA, ME	19/77	28/77	24/77	23/77	22/77	23/77
Eur, CSA, ME → Europe	2/19	3/28	4/24	2/23	4/22	5/23
Eur, CSA, ME → CSA	6/19	3/28	3/24	1/23	2/22	1/23
CSA → Pakistani	1/6	1/3	2/3	0/1	0/2	0/1

Table 4: Predicting the population of origin of individuals in five HapMap Phase 3 populations that do not have reference populations in HGDP. We discarded from our analysis individuals with more than 10% missing entries when extracting our PCAIM or I_n SNP panels P1, P2, and P3. In this case, we included our metric of confidence in the computation of nearest neighbors (see Supplementary Methods) and we report classification accuracy (C_{ACC}, see Supplementary Methods), expressed as the fraction of individuals that were assigned to a region or population of origin at the respective node of the decision tree. Note that most individuals of, for example, Mexican (MEX) ancestry are now unassigned (compared to Table 3), since they lie far away from the HGDP reference populations.

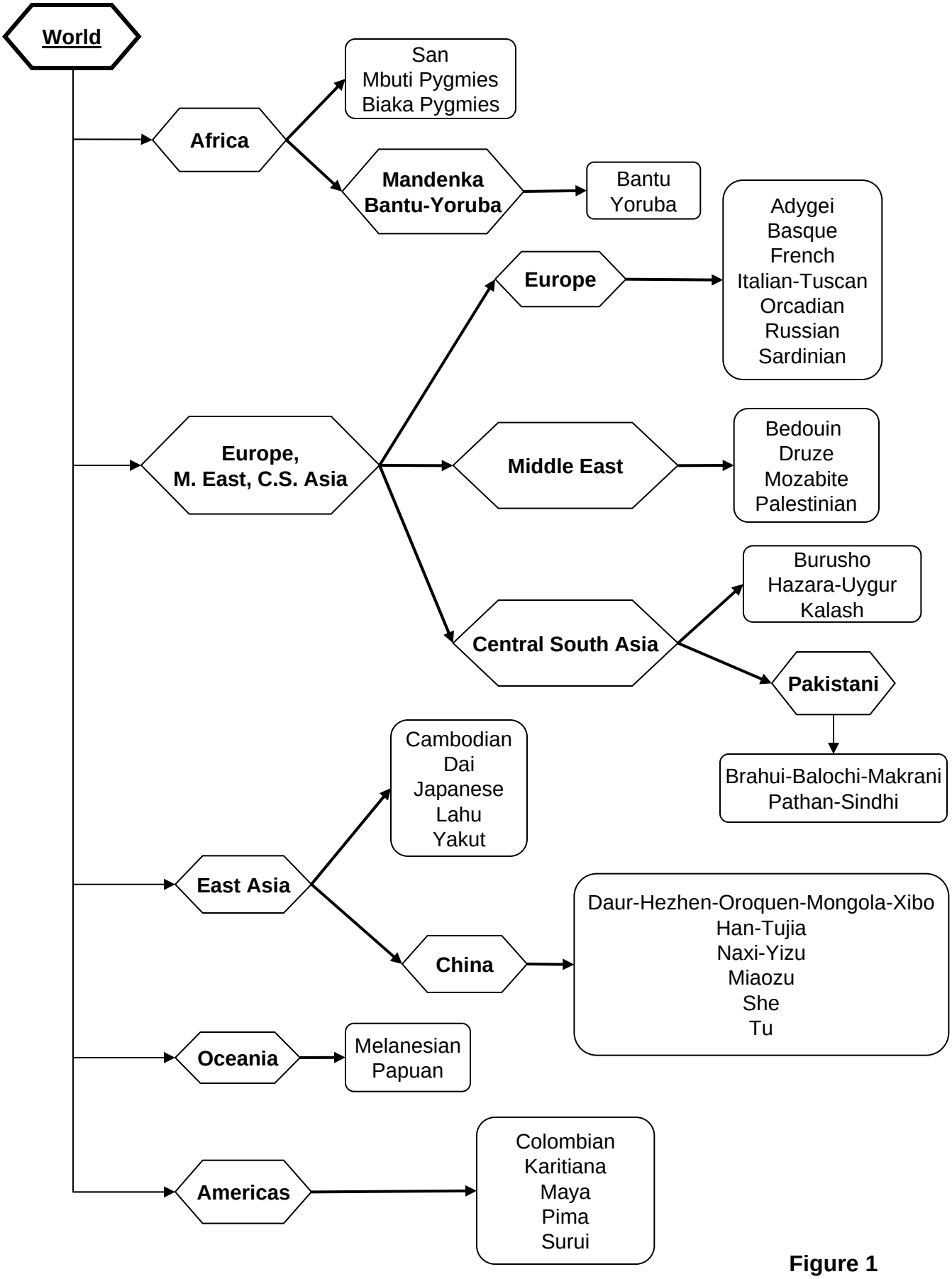
World
Africa
San
Mbuti Pygmies
Biaka Pygmies
Mandenka
Bantu-Yoruba
Bantu
Yoruba
Europe,
M. East, C.S. Asia
Europe
Adygei
Basque
French
Italian-Tuscan
Orcadian
Russian
Sardinian
Middle East
Bedouin
Druze
Mozabite
Palestinian
Central South Asia
Burusho
Hazara-Uygur
Kalash
Pakistani
Brahui-Balochi-Makrani
Pathan-Sindhi
East Asia
Cambodian
Dai
Japanese
Lahu
Yakut
China
Daur-Hezhen-Oroquen-Mongola-Xibo
Han-Tujia
Naxi-Yizu
Miaozu
She
Tu
Oceania
Melanesian
Papuan
Americas
Colombian
Karitiana
Maya
Pima
Surui

Figure 1

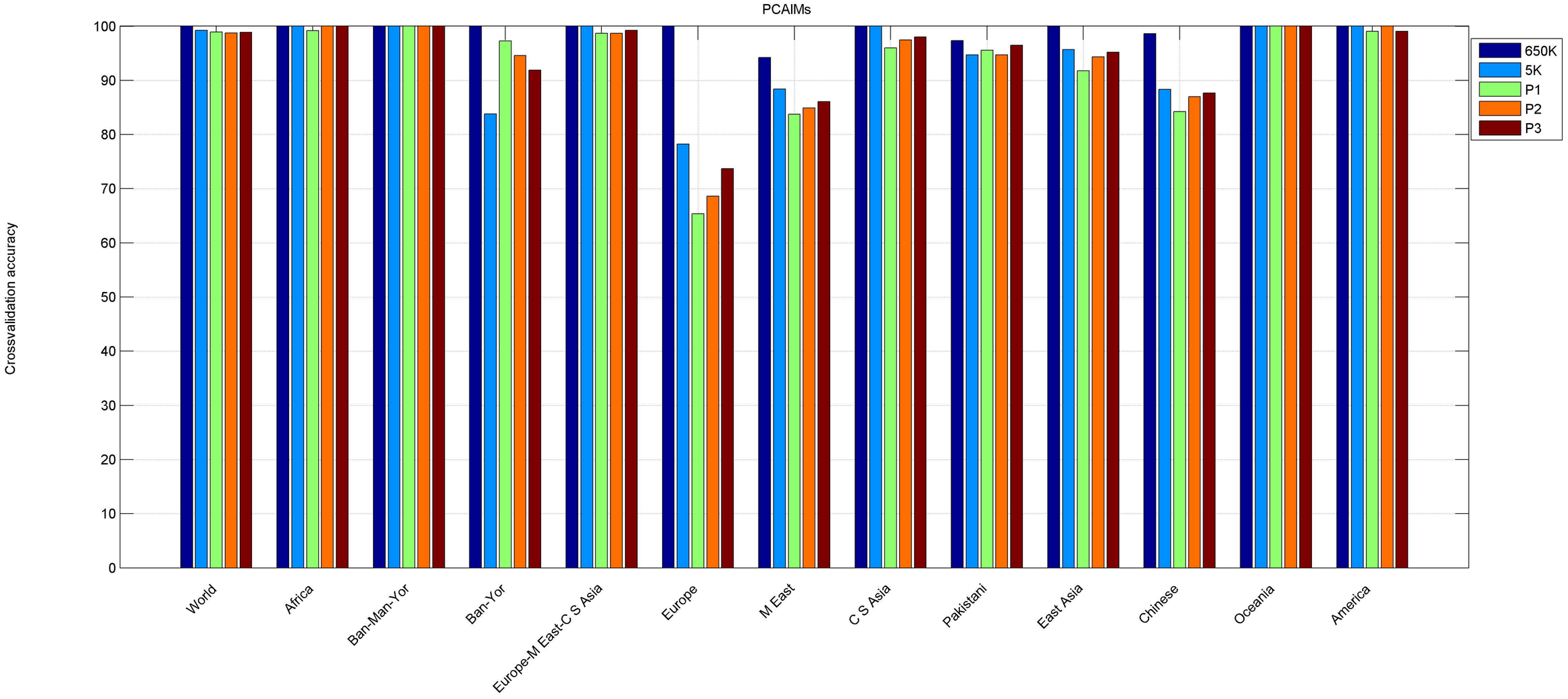
PCAIMs
Crossvalidation accuracy
100
90
80
70
60
50
40
30
20
10
0
650K
5K
P1
P2
P3
World
Afrıca
Ban-Man-Yor
Ban-Yor
Europe-M East-C S Asia
Europe
M East
C S Asia
Pakistani
East Asia
Chinese
Oceania
America

Figure 2A

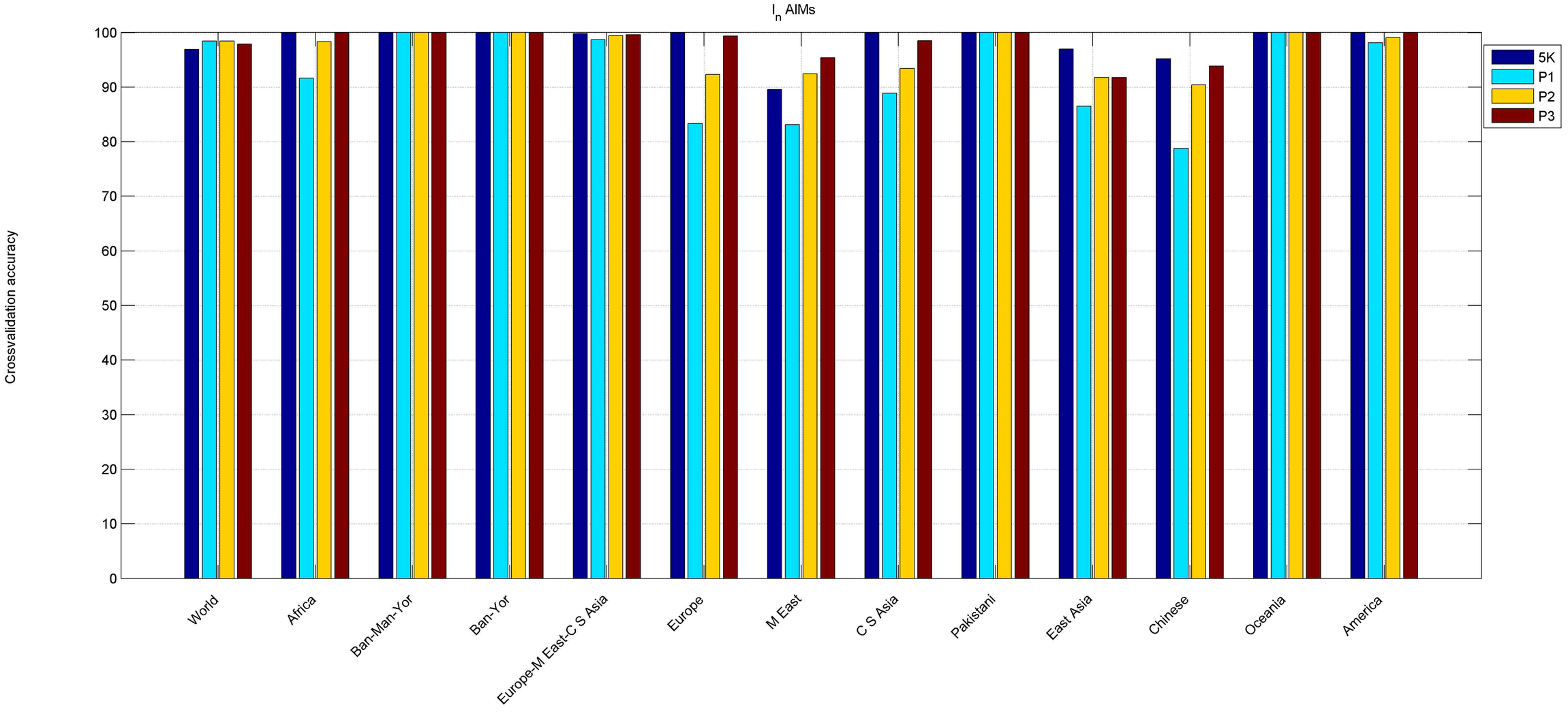
In AIMs
Crossvalidation accuracy
100
90
80
70
60
50
40
30
20
10
0
5K
P1
P2
P3
World
Africa
Ban-Man-Yor
Ban-Yor
Europe-M East-C S Asia
Europe
M East
C S Asia
Pakistani
East Asia
Chinese
Oceania
America

Figure 2B

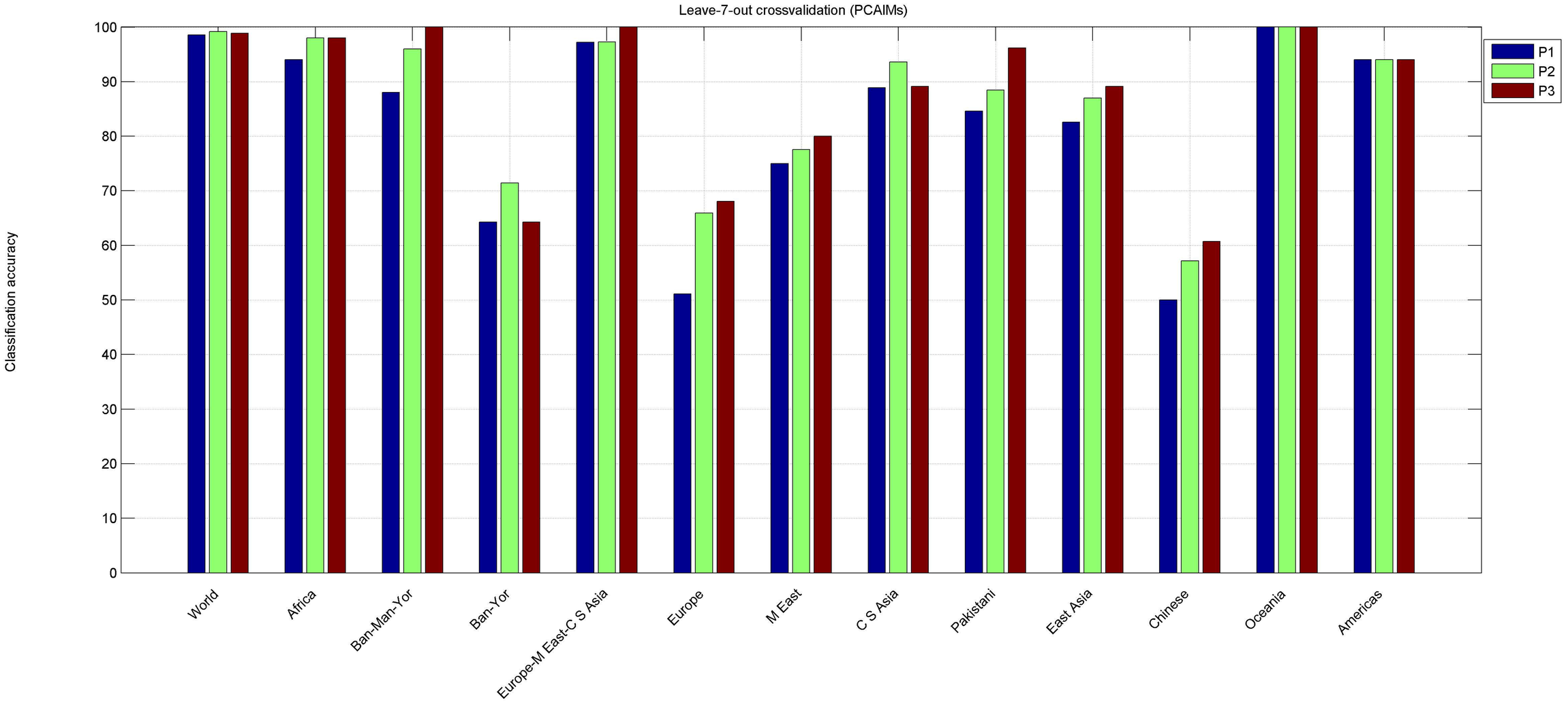
Leave-7-out crossvalidation (PCAIMs)
Classification accuracy
100
90
80
70
60
50
40
30
20
10
0
P1
P2
P3
World
Afrca
Ban-Man-Yor
Ban-Yor
Europe-M East-C S Asia
Europe
M East
C S Asia
Pakistani
East Asia
Chinese
Oceania
Americas

Figure 3A

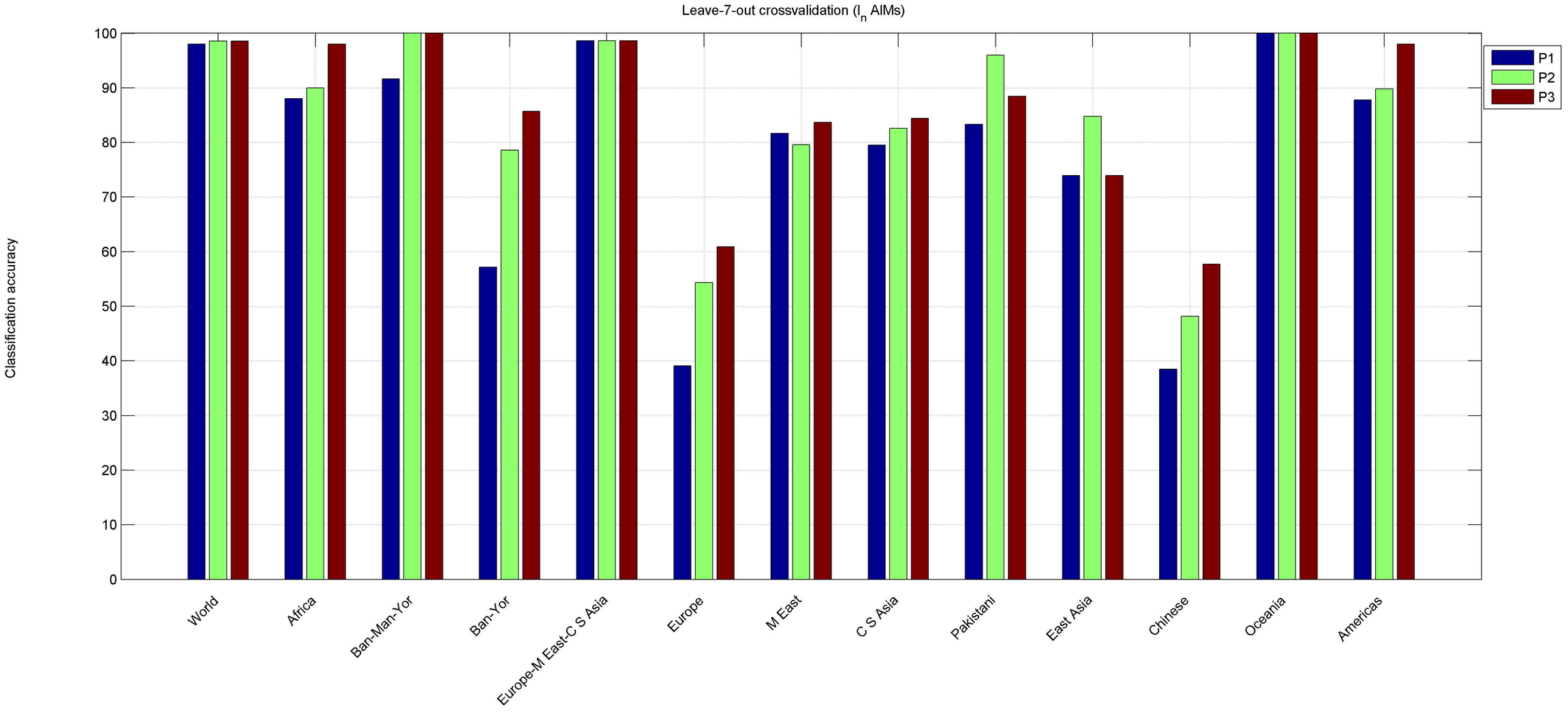
Leave-7-out crossvalidation (I_n AIMs)
Classification accuracy
100
90
80
70
60
50
40
30
20
10
0
P1
P2
P3
World
Afrca
Ban-Man-Yor
Ban-Yor
Europe-M East-C S Asia
Europe
M East
C S Asia
Pakistani
East Asia
Chinese
Oceania
Americas

Figure 3B