PCA-Correlated SNPs for Structure Identification in Worldwide Human Populations

Peristera Paschou[1*], Elad Ziv[2,3,4], Esteban G. Burchard[5,6], Shweta Choudhry[7], William Rodriguez-Cintron[8], Michael W. Mahoney[9], Petros Drineas[10]

1 Department of Molecular Biology and Genetics, Democritus University of Thrace, Alexandroupoli, Greece, **2** Division of General Internal Medicine, University of California San Francisco, San Francisco, California, United States of America, **3** Institute for Human Genetics, University of California San Francisco, San Francisco, California, United States of America, **4** Comprehensive Cancer Center, University of California San Francisco, San Francisco, California, United States of America, **5** Department of Biopharmaceutical Sciences, University of California San Francisco, San Francisco, California, United States of America, **6** Department of Medicine, University of California San Francisco, San Francisco, California, United States of America, **7** Lung Biology Center, Department of Medicine, University of California San Francisco, San Francisco, California, United States of America, **8** Pulmonary/CCM Veterans Caribbean Healthcare System, University of Puerto Rico School of Medicine, San Juan, Puerto Rico, United States of America, **9** Yahoo Research, Sunnyvale, California, United States of America, **10** Department of Computer Science, Rensselaer Polytechnic Institute, Troy, New York, United States of America,

Existing methods to ascertain small sets of markers for the identification of human population structure require prior knowledge of individual ancestry. Based on Principal Components Analysis (PCA), and recent results in theoretical computer science, we present a novel algorithm that, applied on genomewide data, selects small subsets of SNPs (PCA-correlated SNPs) to reproduce the structure found by PCA on the complete dataset, without use of ancestry information. Evaluating our method on a previously described dataset (10,805 SNPs, 11 populations), we demonstrate that a very small set of PCA-correlated SNPs can be effectively employed to assign individuals to particular continents or populations, using a simple clustering algorithm. We validate our methods on the HapMap populations and achieve perfect intercontinental differentiation with 14 PCA-correlated SNPs. The Chinese and Japanese populations can be easily differentiated using less than 100 PCA-correlated SNPs ascertained after evaluating 1.7 million SNPs from HapMap. We show that, in general, structure informative SNPs are not portable across geographic regions. However, we manage to identify a general set of 50 PCA-correlated SNPs that effectively assigns individuals to one of nine different populations. Compared to analysis with the measure of informativeness, our methods, although unsupervised, achieved similar results. We proceed to demonstrate that our algorithm can be effectively used for the analysis of admixed populations without having to trace the origin of individuals. Analyzing a Puerto Rican dataset (192 individuals, 7,257 SNPs), we show that PCA-correlated SNPs can be used to successfully predict structure and ancestry proportions. We subsequently validate these SNPs for structure identification in an independent Puerto Rican dataset. The algorithm that we introduce runs in seconds and can be easily applied on large genome-wide datasets, facilitating the identification of population substructure, stratification assessment in multi-stage whole-genome association studies, and the study of demographic history in human populations.

Citation: 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(9): e160. doi:10.1371/journal.pgen.0030160

Introduction

Genetic structure among and within human populations reflects ancient and recent historical events, migrations, bottlenecks, and admixture, and carries the signatures of random drift and natural selection. The complex interplay among these forces results in patterns that could be used as tools in diverse areas of genetics. In population genetics, uncovering population structure can be used to trace the histories of the populations under study [1]. In medical genetics, identifying population substructure and assigning individuals to subpopulations is a crucial step in properly conducting association studies to unravel the genetic basis of complex disease. With data from large-scale association studies becoming increasingly available, it has become apparent that population substructure resulting from recent admixture or biased sampling can increase the number of false-positive results or mask true correlations [2–5]. Detection of and correction for stratification in a given dataset is a problem that has been discussed at length in recent literature [6–13].

One of the prevailing methods for identifying population structure is a model-based algorithm implemented in the program STRUCTURE [14,15]. STRUCTURE has been shown to effectively assign individuals to clusters [16–18]. However, anticipating data from thousands of individuals and thousands of markers, this algorithm will become impractical due

Editor: David B. Allison, University of Alabama at Birmingham, United States of America

Received April 4, 2007; **Accepted** August 1, 2007; **Published** September 21, 2007

Abbreviations: LD, linkage disequilibrium; PCA, Principal Components Analysis; SVD, Singular Value Decomposition

* To whom correspondence should be addressed. E-mail: ppaschou@mbg.duth.gr

Author Summary

Genetic markers can be used to infer population structure, a task that remains a central challenge in many areas of genetics such as population genetics, and the search for susceptibility genes for common disorders. In such settings, it is often desirable to reduce the number of markers needed for structure identification. Existing methods to identify structure informative markers demand prior knowledge of the membership of the studied individuals to predefined populations. In this paper, based on the properties of a powerful dimensionality reduction technique (Principal Components Analysis), we develop a novel algorithm that does not depend on any prior assumptions and can be used to identify a small set of structure informative markers. Our method is very fast even when applied to datasets of hundreds of individuals and millions of markers. We evaluate this method on a large dataset of 11 populations from around the world, as well as data from the HapMap project. We show that, in most cases, we can achieve 99% genotyping savings while at the same time recovering the structure of the studied populations. Finally, we show that our algorithm can also be successfully applied for the identification of structure informative markers when studying populations of complex ancestry.

to its intensive computational cost [13,19,20]. At the same time, it is sensitive to the choice of prior distributions of model parameters and relies heavily on explicit assumptions about the data that may not always hold, making the method unstable when these assumptions are violated [19,21]. Recently, Principal Components Analysis (PCA), a classical nonparametric linear dimensionality reduction technique, is regaining favor for uncovering population structure. PCA can be used to extract the fundamental structure of a dataset without the need for any modeling of the data; see [22] and references therein for a detailed discussion. It is computationally efficient and can handle genome-wide data for thousands of individuals. PCA was first used in population genetics by Cavalli-Sforza to infer axes of human variation [23]. It has recently been shown to be a powerful tool for the identification of population structure and the correction of stratification in the setting of association studies [13,20]. Coupled with a clustering tool, it can also be used for inferring population clusters and assigning individuals to subpopulations [19].

Identifying a set of markers that could effectively be used for inference of population structure will reduce genotyping costs and will potentially provide insight to genetic loci that have undergone selective pressures. Several approaches have been used to this end [24–29]. For instance, informative markers have been traditionally selected to maximize δ, the absolute difference in allele frequency between two ancestral populations, or Wright's F_{ST} [24,25,30–32]. The cutoff value for δ is highly subjective; its statistical properties are not well defined and it can only be used for two source populations at a time [26,28]. On the other hand, it is not clear how F_{ST} can be applied to the selection of informative markers for admixed populations, when the parental contributions are unequal [26,28]. Informativeness for assignment (I_n), as defined by Rosenberg et al. [26], is an F_{ST}-correlated measure that computes the mutual information on allele frequencies. In all cases in which an allele frequency–based metric is used, knowledge of individual membership to a studied population is a prerequisite. When studying admixed populations, it may be difficult to define or sample the ancestral source populations. The origin of the study individuals may also be unknown in other situations, e.g., studies involving large samples of blood donors.

We have developed a novel algorithm to identify a subset of SNP markers that capture major axes of genetic variation in a genotypic dataset without use of any prior information about individual ancestry or membership in a population. Our approach is a greedy deterministic variant of a Monte-Carlo algorithm that has provable performance guarantees [33–35]. Here, we describe the theory of our method and its derivation from PCA, and evaluate it extensively on a previously described dataset of 255 individuals from 11 populations typed for 10,805 autosomal SNPs [36]. First, we infer the structure of the dataset using PCA followed by a standard clustering algorithm (k-means). We show that this two-step approach almost always achieves 100% accuracy for assigning individuals to their true clusters. We then use our method to identify subsets of SNPs that can extract the same structure and evaluate the performance of these SNP panels for clustering individuals in their respective populations both within and across continents. We compare the efficiency of these panels to SNPs chosen based on the measure of informativeness for assignment (I_n) [26] as well as randomly chosen SNPs. We then validate our results both by splitting the studied individuals in training and test sets and by using the selected SNPs for clustering individuals from the HapMap database. Analyzing genotypes for approximately 1.7 million SNPs that have been made available through the HapMap project [37,38], we identify a set of SNPs that differentiate the Japanese and Chinese populations. We demonstrate that our algorithm for selecting structure informative SNPs always converges to the results of applying PCA and the clustering algorithm on the full dataset, while achieving almost 99% genotyping savings. Furthermore, using data from nine indigenous populations, we manage to ascertain a global panel of PCA-correlated SNPs that accurately assigns individuals to their population of origin. Finally, analyzing two independent Puerto Rican datasets, we demonstrate the applicability of our method for the selection of structure informative markers, when admixed populations are analyzed.

Results

Selecting PCA-Correlated SNPs

We will first develop the theoretical underpinnings of our method and explain its connections to PCA. PCA is a linear dimensionality reduction technique that seeks to identify a small number of "dimensions" or "components" that capture most of the relevant structure in the data. In genetics, given a large number of genetic markers (e.g., thousands of SNPs) for a large number of individuals, PCA and the Singular Value Decomposition (SVD) have been used in order to infer population structure. We note here that SVD is the fundamental algorithmic and mathematical component of PCA; indeed, PCA is equivalent to computing the SVD of a distance matrix representing the data.

Consider a SNP data matrix A whose m rows correspond to m individuals and whose n columns correspond to n SNPs. Let $m \leq n$, which is almost invariably the case in genetics data. The elements of this matrix may be encoded as +1 or 0 or −1, denoting (respectively) whether an individual is homozygotic with respect to the first allele, heterozygotic or homozygotic with respect to the second allele [22]. (See Methods for more details on our encoding of the data.) The SVD of this $m \times n$ matrix A returns m pairwise orthonormal vectors u^i, n pairwise orthonormal vectors v^i, and m nonnegative singular values σ_i. The matrix A may be written as a sum of outer products (rank-one components) as

$$A = \sum_{i=1}^{m} \sigma_i u^i v^{iT}. \qquad (1)$$

For SNP data matrices A of the above form, the left singular vectors (the u^i's) are associated with the columns (SNPs) of A—indeed, they are linear combinations of the columns of A—and may be called eigenSNPs [39]. A common strategy is to perform dimensionality reduction by keeping a small number (e.g., two or three) of eigenSNPs and then perform further data analysis (e.g., clustering or classification) by representing all individuals with respect to the selected eigenSNPs.

Since eigenSNPs are mathematical abstractions and do not correspond to actual SNPs, a natural question arises: is it possible to identify a small subset of actual SNPs (i.e., columns of the original data matrix) such that the top few singular vectors of the matrix containing only the chosen SNPs are strongly correlated with the top few singular vectors of the original SNP matrix? Drineas et al. [33–35] prove that the top few singular vectors of any matrix A may be well approximated by the top few singular vectors of a matrix consisting of a much smaller number of judiciously chosen columns of A. Here, we apply these recent results from the theoretical computer science literature to the problem of SNP selection for structure identification. The selected SNPs will be chosen to correlate with the top principal components, and thus we will call them PCA-correlated SNPs.

Assume that there are k principal components and thus k eigenSNPs of interest; the choice for k will be addressed below. Following [35], we seek the columns of the original matrix that mostly lie in the subspace spanned by the top k eigenSNPs. Notice that we shall seek columns (SNPs) that are simultaneously correlated with all top k eigenSNPs, and not with each of them individually. Surprisingly, the SVD immediately suggests such SNPs. By manipulating the expression of Equation 1, we see that the j-th column (SNP) of the full SNP data matrix A (denoted by A^j) may be expressed as

$$A^j = \sum_{i=1}^{m} (\sigma_i u^i) v_j^i. \qquad (2)$$

Here, v_j^i is the j-th element of the i-th right singular vector. Thus, the j-th column of A is a linear combination of all left singular vectors and corresponding singular values, and the $v_j^i$are the coefficients of this linear combination. Instead of using all m left singular vectors and singular values, we can express A^j as a linear combination of only the top k left singular vectors and corresponding singular values; some loss of information is now inevitable:

$$A^j \approx \sum_{i=1}^{k} (\sigma_i u^i) v_j^i. \qquad (3)$$

We will pick columns of A that have large coefficients v_j^i, $i = 1...k$. In particular, we shall order the columns of A with respect to the scores

$$p_j = \sum_{i=1}^{k} (v_j^i)^2. \qquad (4)$$

Drineas et al. [34,35] proved that if a small number of columns is picked in independent identical random trials, where in each trial the j-th column of A is picked with probability proportional to p_j, then the top k left singular vectors of the selected columns are very close to the top k left singular vectors of the original matrix. While the probabilistic nature of their approach is important for the formal mathematical proof, a greedy variant that picks the columns corresponding to the largest p_j's should also work well for many practical datasets. This greedy variant is implemented here, and does work well for SNP data. We should note that even for the largest matrices that we experimented with (i.e., the HapMap data) the computation of the scores p_j takes less than a few seconds on a conventional computer.

Determining the Number of Significant Principal Components

In order to compute the scores p_j of Equation 4, we must know how many principal components to keep; that is, we must know the rank parameter k. This is equivalent to determining the number of significant principal components in the data, which is a challenging research topic in the data analysis and numerical analysis literature; see [40] for a review. In order to determine whether the i-th principal component is significant, we will compare the data matrix corresponding to the i-th and all smaller principal components to a randomly generated matrix with the same elements. If the former matrix does not have significantly more structure than the latter one, then we conclude that the i-th principal component is not significant. Intuitively, a principal component is significant if and only if it has more structure than a random matrix, which has no useful structure. See Methods for a detailed description of this procedure, which is strongly motivated by recent algorithmic developments in random matrix theory [41,42].

Application on a Worldwide Dataset

We evaluated our methods extensively on a previously described dataset of 11 populations from around the world [36]. Only data from autosomal chromosomes was included in our analysis (10,805 SNPs). To demonstrate the resolution that could be achieved by our algorithms, we analyzed the entire dataset as well as subsets of the data consisting of populations within a single continent. Our PCA-based algorithm does not operate on matrices with missing entries. The procedure we followed for the handling of missing entries resulted in different numbers of SNPs being analyzed for the population group each time under focus. (See Methods for details on encoding and handling of missing entries.) After selecting a subset of either PCA-correlated SNPs or high-I_n SNPs, we employed the procedure described above to determine the number of significant principal components for the selected subset, and we used these components in the subsequent analysis.

SNPs That Cluster Individuals to Different Continents

We first examined if our algorithms could be used to select a small subset of SNPs that cluster individuals in broad continental regions. The studied populations can be assigned to four different continents: Africa (Mbuti, Mende [East African], Burunge [West African], and African Americans), Europe (European Americans and Spanish), Asia (Mala [South Indian], East Asian, and South Altaian), and America (Nahua and Quechua). A total of 9,419 SNPs were included in our analysis. As discussed later in this report, PCA will recognize much finer resolution than broad intercontinental clustering. So, for this particular experiment we manually set the number of principal components for further analysis to three. The rationale behind this choice is that the first principal component captures 37.4% of the variance in the data, the second an additional 7.5%, the third an additional 3.1%, whereas the contribution of the fourth one drops below 1.5%. (The experimental results would be essentially the same even if four principal components were kept.) These top three eigenvectors were used to cluster the data in four clusters using the *k*-means algorithm. We then compared individual assignment to a cluster to actual membership to a continent and found that PCA and *k*-means achieved perfect clustering of individuals to different continents using all available SNPs (Figure 1).

Next, we reproduced this result using only a small subset of SNPs. Calculating the scores described in Equation 4, we selected ten to 200 PCA-correlated SNPs and repeated PCA and *k*-means clustering using only this small subset of SNPs. Figure 2 shows the scores and positions of the top 30 PCA-correlated SNPs throughout the genome as well as their genotype frequency patterns across the four continents. As shown in Figure 1, these 30 PCA-correlated SNPs achieve close to perfect clustering of all studied subjects to their respective continents of origin (correlation coefficient between predicted and true membership to a continental cluster is 0.99).

We compared the efficiency of the PCA-correlated SNPs that we selected to that of a set of SNPs selected using the informativeness for assignment measure (I_n) [26]. I_n was estimated for all available SNPs, using four geographically distinct population groups (its calculation requires knowledge of predefined populations). Again, the top ten to 200 highest-ranking I_n SNPs were picked and PCA and *k*-means was run in order to assign individuals to four continental clusters using only these subsets of SNPs (Figure 1). High I_n SNPs also perform very well. However, in this case, even though PCA-correlated SNPs have been selected in an entirely unsupervised manner, they perform better than high-I_n SNPs (Figures 1 and 2). When choosing high-ranking I_n SNPs, about 60 SNPs are needed in order to achieve accurate clustering to four continents. Interestingly, there was a 53% overlap between the top PCA-correlated SNPs and SNPs ranking high for I_n.

We repeated the aforementioned procedure using sets of randomly selected SNPs. Experiments were repeated 30 times, each time selecting ten to 200 random SNPs. The average correlation coefficient of individual membership using these random SNPs to membership using all available SNPs is shown in Figure 1. As expected, random SNPs perform far worse than carefully selected SNPs for the inference of population structure.

In order to validate our method, we split the studied individuals into a training set (50%) and a test set (50%). This time, we first used both our method and the I_n measure to select structure informative SNPs in the test set and then applied this panel of SNPs to assign individuals to continental regions in the training set. Experiments were repeated with 50 random splits and the average correlation coefficients between predicted and true membership over all runs are shown in Figure 3. Analysis of the test set produced essentially the same results as in the training set, with PCA-correlated SNPs doing slightly better than I_n SNPs. We expect that our results would improve with larger training sets.

Finally, we examined the value of the SNP panels that we selected for clustering individuals to different continental regions by testing their performance for assigning individuals from the four HapMap populations (Yoruba in Ibadan, Nigeria; Utah residents with ancestry from northern and western Europe (CEPH); Han Chinese in Beijing, China; and Japanese in Tokyo, Japan) to their true continent of origin (Figure 3). Both PCA-correlated SNPs and high-I_n SNPs perform exceptionally well and as few as 14 PCA-correlated SNPs or 20 high-I_n SNPs are enough to accurately cluster all samples to three distinct clusters. However, this task seems to be much easier than clustering 11 different populations to different continents as we did before, and as few as 40–45 random SNPs also suffice to accurately assign individuals to their continents of origin.

SNPs That Detect Population Structure within Continents

We next tested the efficiency of our methods for detecting population structure in finer detail. To this end, we studied populations that originated from the same geographic region separately, and repeated the empirical analysis described earlier in this report. Three of the populations that we studied are indigenous Africans. We tried to define a subset of SNPs that could be used in order to accurately cluster individuals to each of these populations (Figure 1). Our analysis showed the existence of five significant principal components and these were used to perform *k*-means clustering using 8,928 SNPs. This achieved almost 100% accuracy (correlation coefficient 0.97). Analyzing 20 PCA-correlated or high-I_n SNPs is enough to replicate the results of PCA and *k*-means on the full dataset. On the other hand, as many as 200 random SNPs are needed for the correlation coefficient between true and predicted membership to reach 0.95. The overlap between SNPs selected using our method and I_n is 34%. Interestingly, PCA-correlated SNPs are in high linkage disequilibrium (LD) with I_n SNPs; see Table 1 for more details.

Adding the admixed ancestry population of African Americans to this group decreases our ability to perfectly cluster individuals in a distinct population of origin using *k*-means (Figure 1). Again, five principal components were identified as significant, this time for the analysis of 9,193 SNPs. While each of the autochthonous African populations is still accurately clustered, African Americans do not all fall in a single cluster. African Americans are not one homogeneous population and overlap exists with one of the other populations that were available for study in this continent. In fact, 19 out of the 42 studied individuals are assigned, as could perhaps be expected, to the western African cluster of the Burunge population. We note that we found no clear overlap

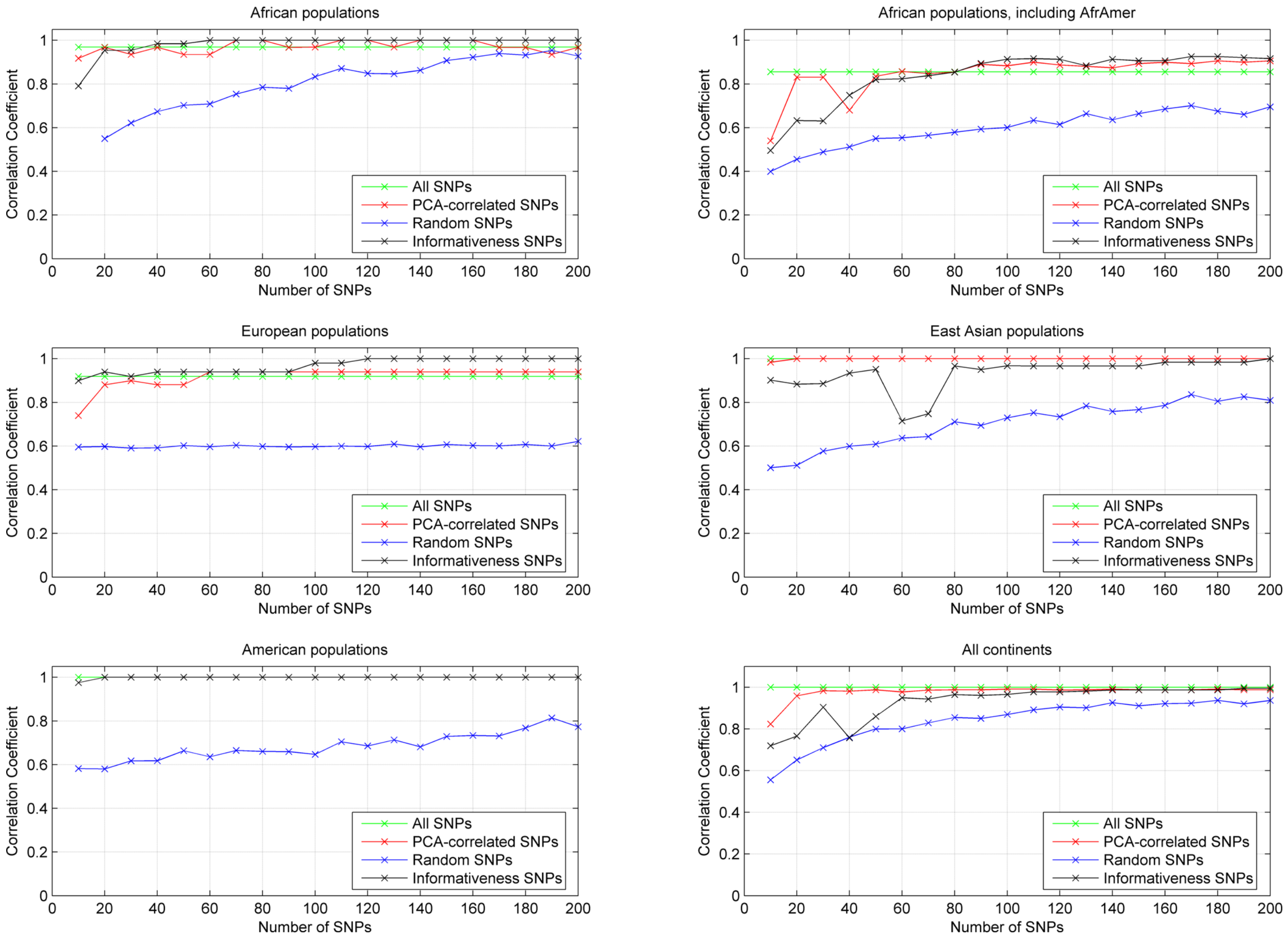

Figure 1. Average Correlation Coefficient between True and Predicted Membership of an Individual to a Particular Population or Continental Region, Using PCA and *k*-Means Clustering on all Available SNPs for a Given Geographic Region, and Sets of Ten to 200 PCA-Correlated, High-I_n or Random SNPs (Random Selection Was Repeated 30 Times)

The reported correlation coefficient is averaged over all populations in the respective geographic region or over the broad continental clusters.
doi:10.1371/journal.pgen.0030160.g001

between the African American and the Caucasian samples that were analyzed here (unpublished data). Again, great genotyping savings seem possible, with only 20 PCA-correlated or 50 high-I_n SNPs needed to converge to the performance of PCA and *k*-means clustering using all available SNPs. Once more, we observed high overlap and LD between PCA-correlated and high-I_n SNPs (Table 1).

We next studied two European populations: a Spanish sample and a broadly defined Caucasian sample (Figure 1). The dataset of 9,668 SNPs was reduced to two principal components. Two Spanish subjects were clear outliers (unpublished data) and were removed from this analysis. The correlation coefficient between the actual membership of each individual to one of the two samples and the predicted membership using PCA on all SNPs and *k*-means is 0.9. Analysis with small subsets of informative markers (ten high-I_n or 20 PCA-correlated SNPs) quickly converges to the results of analyzing the full dataset. The European American sample represents cell lines curated at the Coriell Insitute and although the degree of Spanish admixture in this sample is not known to us, it does not seem to be significant according to our findings. PCA-correlated SNPs and high-I_n SNPs have an overlap of 27.5% and are also in LD, although slightly less than what we observed in other geographic regions (Table 1).

For the Asian and American populations, 9,707 and 8,202 SNPs, respectively, were analyzed (Figure 1). Three principal components were retained in Asia and two in America. Again, the top PCA-correlated and the highest-I_n SNPs perform exceptionally well for the inference of population structure in these continents. For the Asian populations, as few as ten PCA-correlated SNPs are required for almost perfect assignment, whereas a greater number of high-I_n SNPs is required (about 80 SNPs for convergence to 1). Moving to America, ten carefully chosen SNPs can distinguish between the Quechua and the Nahua. It should be noted that the 76.5% overlap between PCA-correlated and high-I_n SNPs in this continent, is the highest that we observed in this study (in Asia the corresponding overlap was 49%); see Table 1 for details.

Both our method and the I_n metric do not suffer from selecting a large number of redundant SNPs in this dataset (Table 2). For the top 200 PCA-correlated SNPs within each geographic region, we calculated r^2 between all pairs. Out of

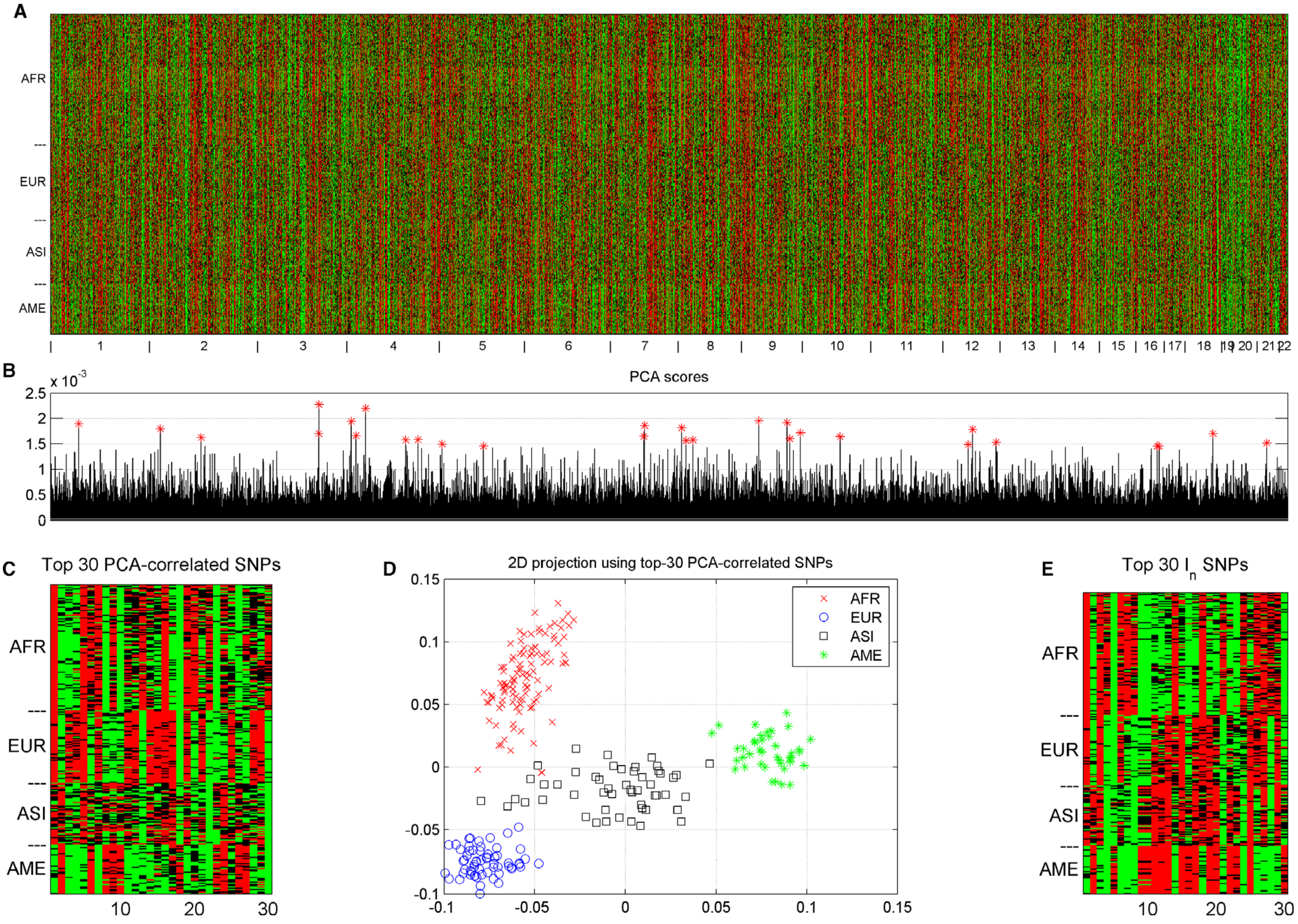

Figure 2. Selecting PCA-Correlated SNPs for Intercontinental Clustering

(A) Raster plot of 255 subjects from four different continental regions with respect to 9,419 SNPs (red/green denotes homozygotic individuals and black denotes hererozygotic individuals).
(B) The scores p_j for each SNP. A red star indicates SNPs corresponding to one of the top 30 scores.
(C) Raster plot of the 255 subjects with respect to the top 30 PCA-correlated SNPs. Notice the patterns formed in the four continental blocks.
(D) Plot of the 255 subjects in the "optimal" 2-D space using the top 30 PCA-correlated SNPs.
(E) Raster plot of the 255 subjects with respect to the top 30 I_n SNPs. Notice that the blocks corresponding to Asia and Europe are slightly more entangled when compared to (C).
doi:10.1371/journal.pgen.0030160.g002

the thousands of possible pairs very few are actually in high LD. The same is true for the top 200 I_n SNPs selected to cluster populations within continents.

We finally explored the feasibility of accurately clustering two populations of related ancestry, the Han Chinese and the Japanese, using data available from the HapMap database [37,38]. We downloaded all available data (release 21–1r; 3,776,850 SNPs) and excluded from the analysis SNPs that were fixed in both populations (1,356,867 SNPs) and had more than one missing entry in both samples. This left us with 1,662,041 SNPs genotyped for 45 Chinese and 45 Japanese samples. We ran PCA and k-means on all available SNPs using the two principal components detected as significant. As a result, one Japanese individual was misclassified; this individual is a clear outlier (Figure 4). We then selected a subset of SNPs that could be used to infer this population structure, using both our method and the I_n measure. We found that with 50 PCA-correlated SNPs only two Japanese and one Chinese are misclassified (correlation coefficient 0.97), while with 150 PCA-correlated SNPs we are able to exactly reproduce the results of the analysis on the complete dataset of roughly 1.7 million SNPs. High-I_n SNPs are quite efficient as well. However, when choosing between 50 and 300 high-I_n SNPs, our algorithm for determining the number of principal components detects a very large number of significant principal components (more than 40). This causes the artifact of Figure 4C (a drop in the performance of high-I_n SNPs). Manually fixing the number of principal components that we use for k-means clustering to two corrects this artifact, as the dotted line shows in Figure 4C. Such artifacts do not seem to arise with our method of choosing PCA-correlated SNPs in our empirical evaluation.

Transferability of Structure Informative SNPs

We investigated whether SNPs that were selected for assigning individuals to clusters in one continent would be useful in another continent or for intercontinental differentiation. In an effort to answer this question, we tested the

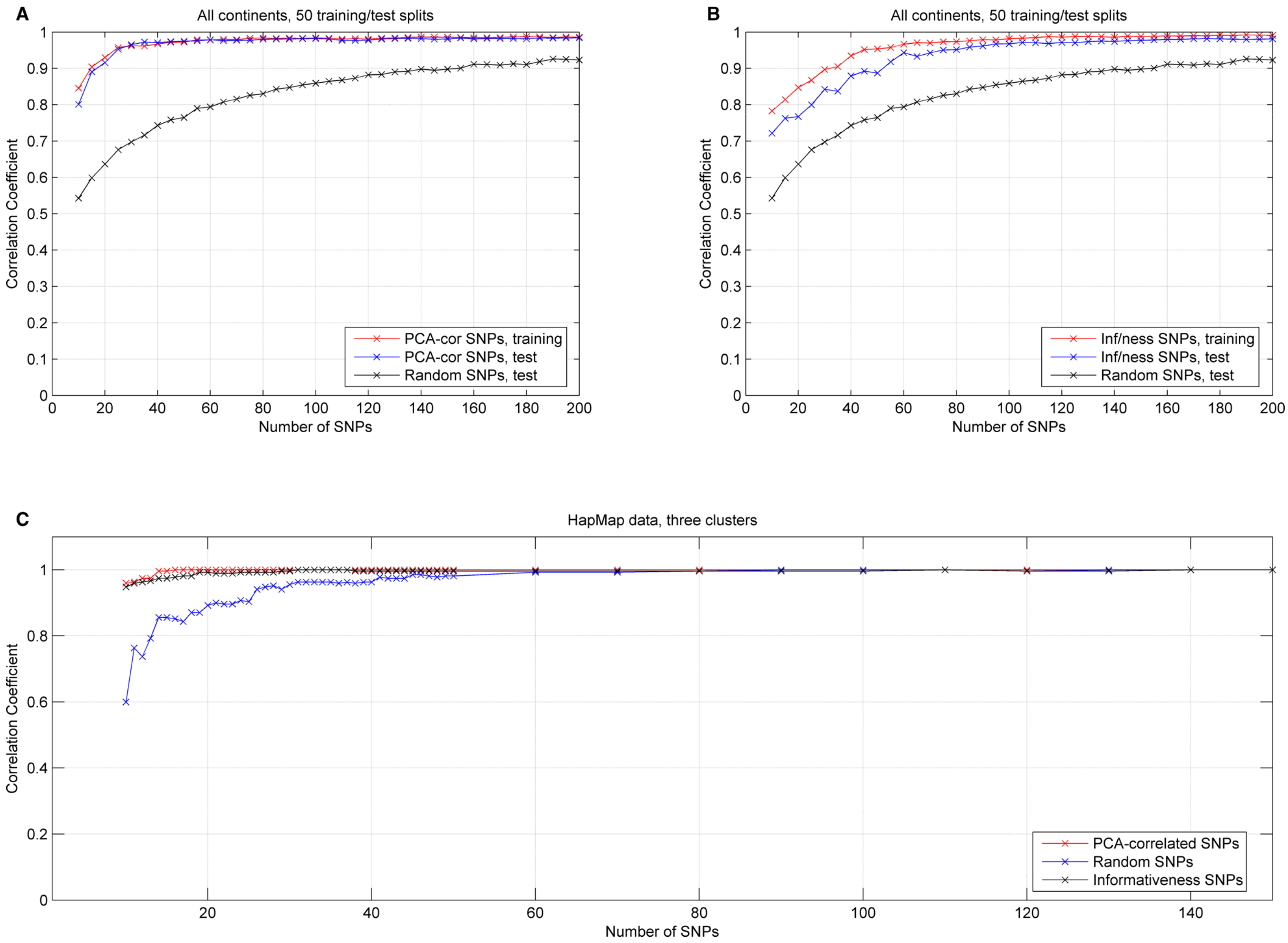

Figure 3. Cross-Validation of Structure Informative SNPs Selected for Intercontinental Clustering

(A, B) Split of our worldwide sample in 50% training and 50% test set. Average correlation coefficient between true and predicted membership of an individual to a continental region using sets of (A) ten to 200 PCA-correlated or (B) ten to 200 high-I_n SNPs selected on the training set, and application of the same sets of selected SNPs on the test set (results are averaged over 50 training/test set splits).

(C) Application of the SNP panels selected for intercontinental clustering in our worldwide sample, on the HapMap populations (average correlation coefficient between true and predicted membership of an individual to one of three continents is shown).

doi:10.1371/journal.pgen.0030160.g003

panels that we selected in each of the four continental regions that we studied (both using the PCA-correlated measure and I_n) for population clustering in another continent. Results were very poor and it seems that SNPs chosen for ancestry inference in one continent are in general no better than random SNPs and not transferable to other continents. The fact that the average overlap over selected SNPs across the different continental regions is about 2% underlines these results. In Figure S1, representative results are shown for testing the portability of SNPs chosen in Europe and Africa in order to infer structure in the remaining three continents for which data were available. In a similar fashion, it seems that neither PCA-correlated SNPs nor high-I_n SNPs for intercontinental clustering can resolve the population structure within continents any better than randomly chosen SNPs (Figure S1). Differentiation among the Asian populations studied here is an exception and the correlation coefficient between actual and predicted membership exceeds 0.9 when studying 60 or more PCA-correlated SNPs that were ascertained for intercontinental clustering. In this case, the overlap between PCA-correlated SNPs selected for differentiation within the three Asian populations that we studied and intercontinental clustering is somewhat high (9%).

SNPs to Infer Population Structure within and across Continents

Next, we explored the possibility of ascertaining a general SNP panel that could be used for ancestry inference and the study of population structure around the world. Results are shown in Figure 5. We excluded admixed populations from this analysis (African Americans and Caucasians), and studied nine indigenous populations for which data were available (Mbuti, Mende, Burunge, Spanish, Mala, East Asian, South Altaian, Nahua, and Quechua). We ran PCA on 9,160 SNPs and *k*-means clustering on the ten detected significant eigenvectors, and managed to successfully assign each individual to their country of origin. Notice that the 3-D plot presented in Figure 5 is somewhat deceiving, as our method picked ten principal components as significant. If

Table 1. Overlap between the Top 200 PCA-Correlated SNPs and the Top 200 I_n SNPs

Studied Region	Overlap	Average r^2
Africa[a]	34%	0.87
Africa[b]	40%	0.83
Europe	27.5%	0.59
Asia	49%	0.92
America	76.5%	0.9

[a]Only the three indigenous African populations are included.

[b]The African-American population is included as well.

The second column reports the percentage of SNPs that overlap among the SNPs selected by each method. The third column reports the average LD between PCA-correlated and I_n SNPs. (For each PCA-correlated SNP we identified the I_n SNP in highest LD and reported the average r^2 over all pairs.)

doi:10.1371/journal.pgen.0030160.t001

Table 2. Number of Pairs among the Top 200 PCA-Correlated (PCA-c.) and I_n SNPs That Are in High LD. A Total of 20,100 Pairs Were Tested in Each Region

r^2	0.4–0.6		0.6–0.8		0.8–1	
	PCA-c.	I_n	PCA-c.	I_n	PCA-c.	I_n
Africa[a]	14	26	4	5	19	10
Africa[b]	3	7	5	4	13	10
Europe	4	3	0	1	6	2
Asia	15	18	7	14	21	22
America	15	9	5	5	21	25

[a]Only the three indigenous African populations are included.

[b]The African-American population is included as well.

doi:10.1371/journal.pgen.0030160.t002

visualization in a 10-D space were possible, further differentiation of the studied populations would become apparent. Investigating the possibility of identifying a small set of SNPs to reproduce this structure, we selected sets of ten to 400 PCA-correlated and high-I_n SNPs and repeated the analysis. Surprisingly, using only 50 PCA-correlated SNPs, we were able to correctly assign all individuals to one of the nine studied populations. On the other hand, in this case, high-I_n SNPs do not seem to do any better than randomly selected SNPs (Figure 5).

In order to test how the set of PCA-correlated SNPs is modified each time an additional population is added to the analysis, we studied incrementally distinct subsets of the data and compared the SNPs selected as structure informative in each subset to the panel of 50 SNPs that are sufficient for accurate clustering of the individuals to nine different populations (see experiment described above). The first subset of populations that we analyzed consisted of genotypes for one population from each continent (East African, Spanish, East Asian, and Nahua) and in each round one additional population was added randomly to the analysis. Results are shown in Table 3. Clearly, there is significant overlap between the panels of informative SNPs at each stage, which increases as more populations are added.

Application to an Admixed Population

Finally, we investigated the applicability and efficiency of our method to select structure informative SNPs in admixed populations. To this end, we studied two independent samples of Puerto Ricans. The first dataset (Puerto Rican A) is a sample of 192 Puerto Ricans [43] genotyped for approximately 100,000 SNPs, 7,259 of which overlapped with our worldwide dataset and were included in our analysis. The second Puerto Rican dataset (Puerto Rican B), has been described in [36] and constitutes a sample of 19 individuals, genotyped for the same markers as the rest of the worldwide samples that we analyzed so far.

It is well known that Puerto Ricans are genetically complex and composed of various proportions of Native American, African, and European genetic origins. We first investigated the Puerto Rican A dataset, and explored the ancestry of the 192 individuals across the African–European and the African–European/American axis (Figure 6). PCA performed on the Puerto Rican sample alone, revealed two significant principal components. We then added data on Europeans (Spanish and Caucasians), West Africans (Burunge), and Native Americans (Quechua, Nahua) to this analysis. Four principal components were found significant. We calculated the variance of the 192 Puerto Ricans across the two axes of ancestry. We observed that the sample variance across the African–European axis was six times larger than the variance across the African–European/American axis, which indicated that our sample had very little interindividual variation in Native American ancestry. Given this observation and for simplicity of exposition we only analyzed variation across the African–European axis. As is clear in Figure 6, the Puerto Ricans that we studied lie virtually on a straight line between Africans and Europeans and are much closer to Europeans than Africans.

We interpreted individuals from the Puerto Rican sample as a combination of European and African ancestry, with the proportion of admixture being equal to the distance of each individual from the centroid of the ancestral population (ancestry coefficient, see Methods for details). After evaluating the ancestry of all 192 individuals using all available SNPs, we attempted to accurately predict it using a subset of PCA-correlated SNPs, selected on the Puerto Rican data only. No information from ancestral populations was necessary. As shown in Figure 6, we accurately predict the ancestry coefficient of each individual. For example, the Pearson correlation coefficient between true and predicted ancestry is higher than 0.8 using 30 PCA-correlated SNPs and higher than 0.9 using 100 PCA-correlated SNPs.

Finally, we cross-validated these findings by applying the panel of PCA-correlated SNPs that we selected on the Puerto Rican A dataset to infer individual ancestry in the Puerto Rican B dataset. As shown in Figure 6, the SNP panel that we selected in Puerto Rican A performs equally well on Puerto Rican B. Notice, however, that this time random SNPs do better than before, perhaps due to the small number of individuals in this sample (19 individuals).

Discussion

Geographic ancestry can be inferred from genotypic data [16,44–47]. The Bayesian approach implemented by Pritchard et al. [14] and PCA have been the two main tools of choice for identification of population structure and subdivision

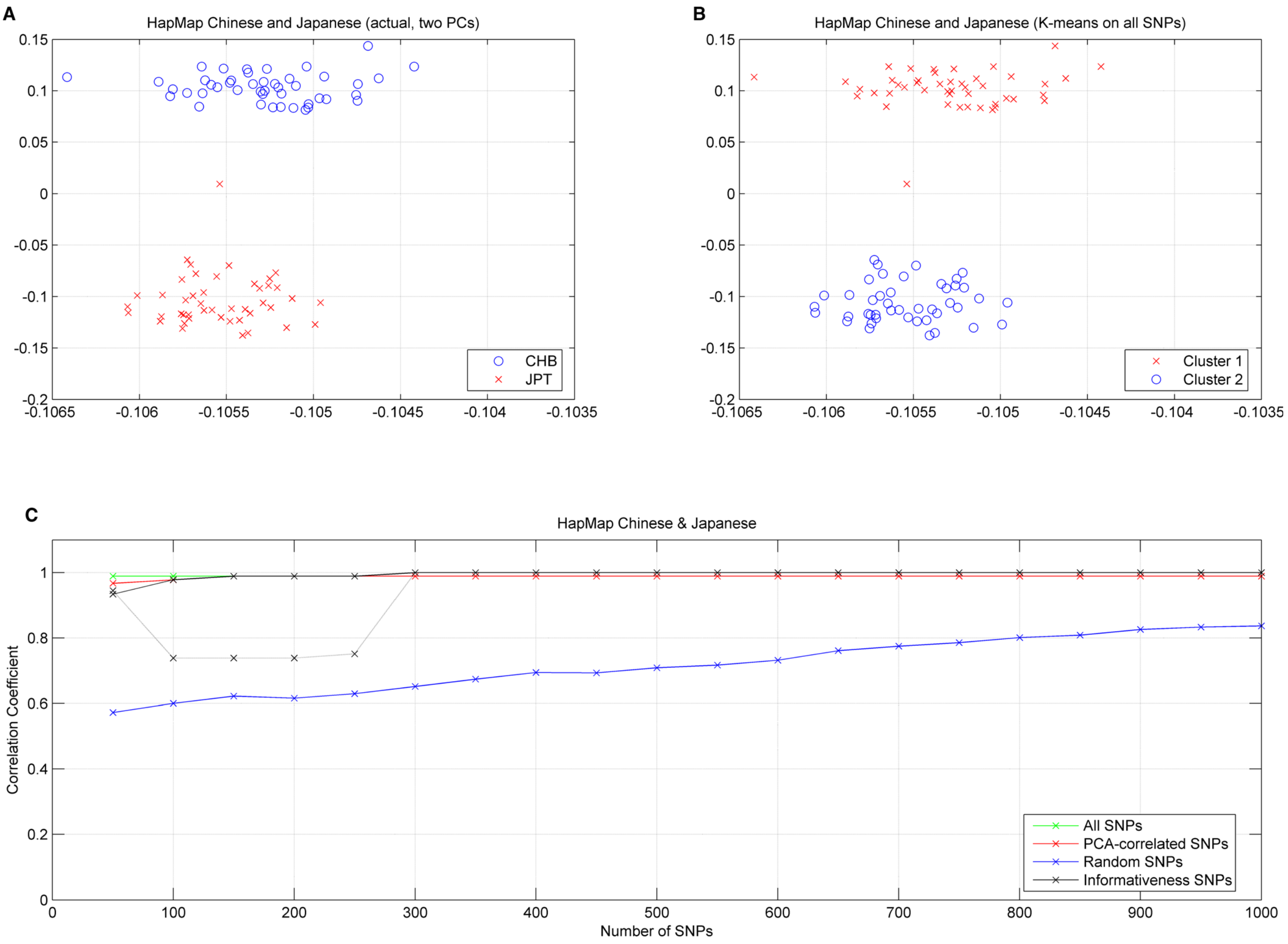

Figure 4. Analysis of 1.7 Million SNPs Typed on the HapMap Han Chinese and Japanese populations (Available from the HapMap Database)

(A) Projection of all 90 Han Chinese and Japanese individuals on the top two principal components using PCA on all available SNPs

(B) *k*-Means clustering on panel (A).

(C) Average correlation coefficient between true and predicted membership of an individual to the Japanese of Han Chinese populations, using PCA and *k*-means clustering on all available SNPs and sets of 50 to 1,000 PCA-correlated, high-I_n or random SNPs (random selection was repeated 30 times). The dotted line represents a decline in the performance of high-I_n SNPs due to the detection of a very large number of significant principal components; see Results for details.

doi:10.1371/journal.pgen.0030160.g004

[18,36,48]. Recent studies have demonstrated that PCA is a fast, easy-to-implement method with great power for analysis of the very large datasets that are increasingly becoming available [13,19,20,22]. Extending recent algorithmic work that provably extracts matrix columns that correlate well with the dominant subspaces identified by PCA, we have developed a method to ascertain a small subset of SNPs that explicitly capture the structure of a population as identified by PCA. The population structure informative SNPs that are selected by our algorithm are also ancestry informative, and we show that they can be effectively used to assign individuals to different continents or populations. Achieving, in most cases that were studied here, 99% genotyping savings, these panels of SNPs can be used to reduce considerably the number of markers needed for ancestry inference. This is desirable in a variety of different research scenarios, such as association mapping (where unrecognized population stratification may lead to spurious associations with disease), forensics, conservation studies, and population genetics [10,30,49–52]. For instance, we believe that our method is useful for investigators conducting association studies in two-stage designs and seeking to replicate the stratification assessment they are doing with the first stage (e.g., from a genome-wide dense SNP screen) to a lower throughput method. For these types of applications, our method would be able to faithfully replicate this assessment with minimal additional markers in the second stage. However, a precondition for replication is that the second population contains the same substructure as the first population.

Our algorithm is simple and computationally fast (less than one minute for the largest runs presented here) and thus allows the analysis of very large genome-wide datasets with thousands of individuals. Perhaps the most important advantage of our method for selecting PCA-correlated SNPs is that it is nonparametric and does not rely on any assumptions or modeling of the data. We simply detect SNPs that are correlated with the subspace spanned by the top few eigenvectors after determining the number of significant

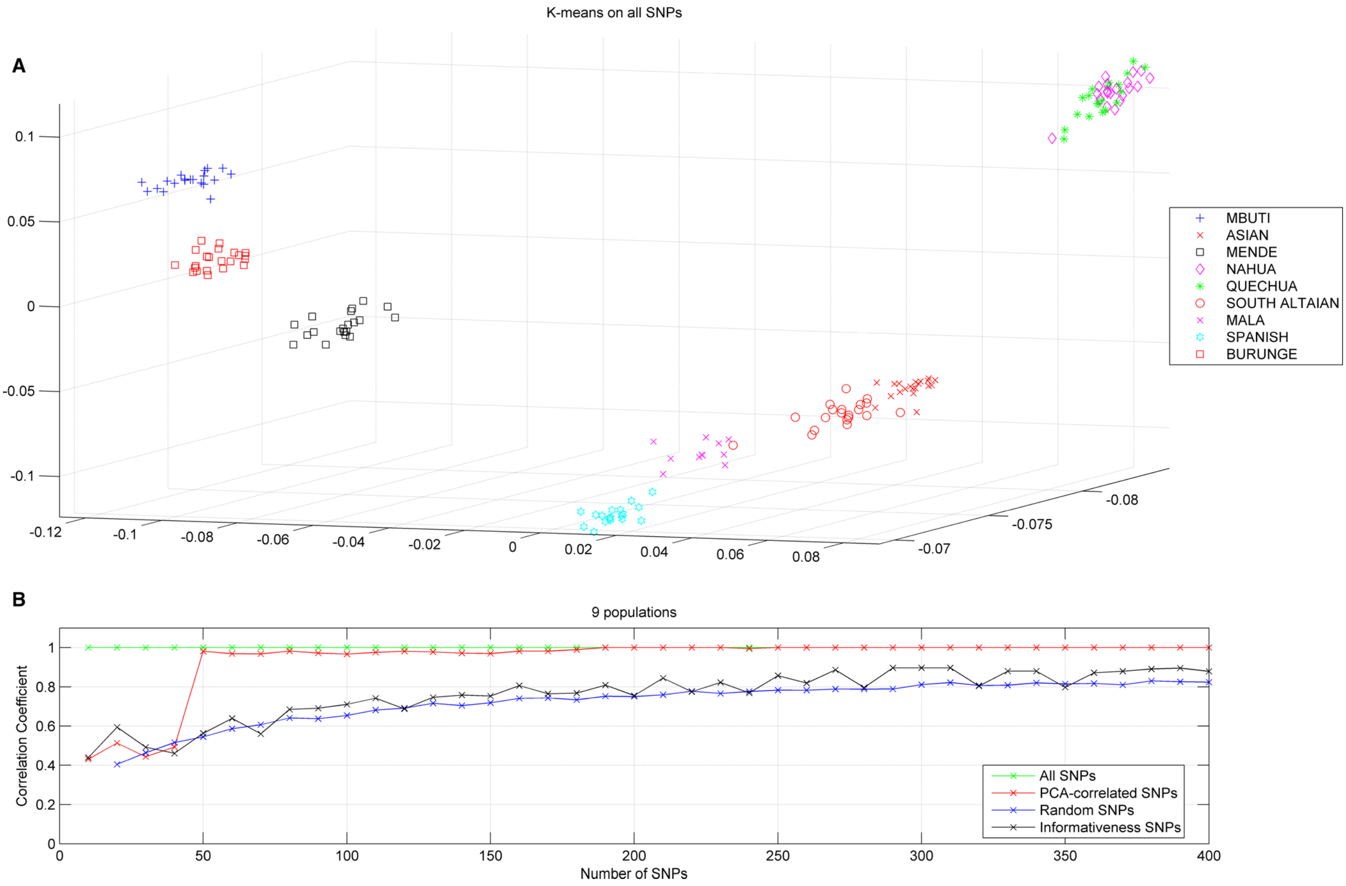

Figure 5. Analysis of Nine Indigenous Populations Typed for 9,160 SNPs

(A) Projection of all individuals of nine indigenous populations on the top three principal components using PCA on all available SNPs. (Ten significant principal components were actually detected.)

(B) Average correlation coefficient between true and predicted membership of the individuals to the nine populations, using PCA and *k*-means clustering on all available SNPs and sets of ten to 400 PCA-correlated, high-I_n or random SNPs (random selection was repeated 30 times).

doi:10.1371/journal.pgen.0030160.g005

principal components. All other methods in the literature that are used to identify ancestry informative markers either rely on a specific model or are frequency based and demand prior knowledge of the origin of individuals [24–29]. Situations may exist where the use of prior information about the studied populations is desirable and we are currently working towards extensions of our approach to such settings.

Table 3. Incremental Analysis of Nine Populations and Effect on the Selection of PCA-Correlated SNPs

Populations	Overlap	Accuracy
Mende, Spanish, Asian, Nahua	8	1
+ Mbuti	9	1
+ South Altaian	18	0.98
+ Quechua	29	0.95
+ Mala	38	0.96
+ Burunge	50	1

The second column reports the overlap between the top 50 PCA-correlated SNPs for each subset of the nine populations under consideration and the top 50 PCA-correlated SNPs for all nine populations. The third column reports the clustering correlation coefficient for each subset of the nine populations using the top 50 PCA-correlated SNPs.

doi:10.1371/journal.pgen.0030160.t003

It should be noted that the final number of SNPs needed to describe population structure is not directly provided by the method and can only be estimated through empirical evaluation of a specific dataset. Also, in applying *k*-means clustering, we manually fixed the number of clusters to the number of populations in the data. It should be made clear, however, that the identification of the number of clusters in the data is not necessary for the implementation of our method; *k*-means clustering is only used here for demonstrating the efficiency of our approach. PCA-correlated SNPs are computed and can be used independently of *k*-means clustering. We chose *k*-means simply because it is a well-known and widely applicable clustering algorithm, which has numerous efficient implementations. One might experiment with different (and perhaps more accurate) clustering algorithms [53], such as hierarchical clustering, spectral clustering, *k*-median approaches, etc.

We were unable to compare the SNPs we selected as ancestry informative using our algorithm to published lists of ancestry informative markers [54–60] because the overlap between these lists and the SNPs that were available to us was either extremely small or different populations were ana-

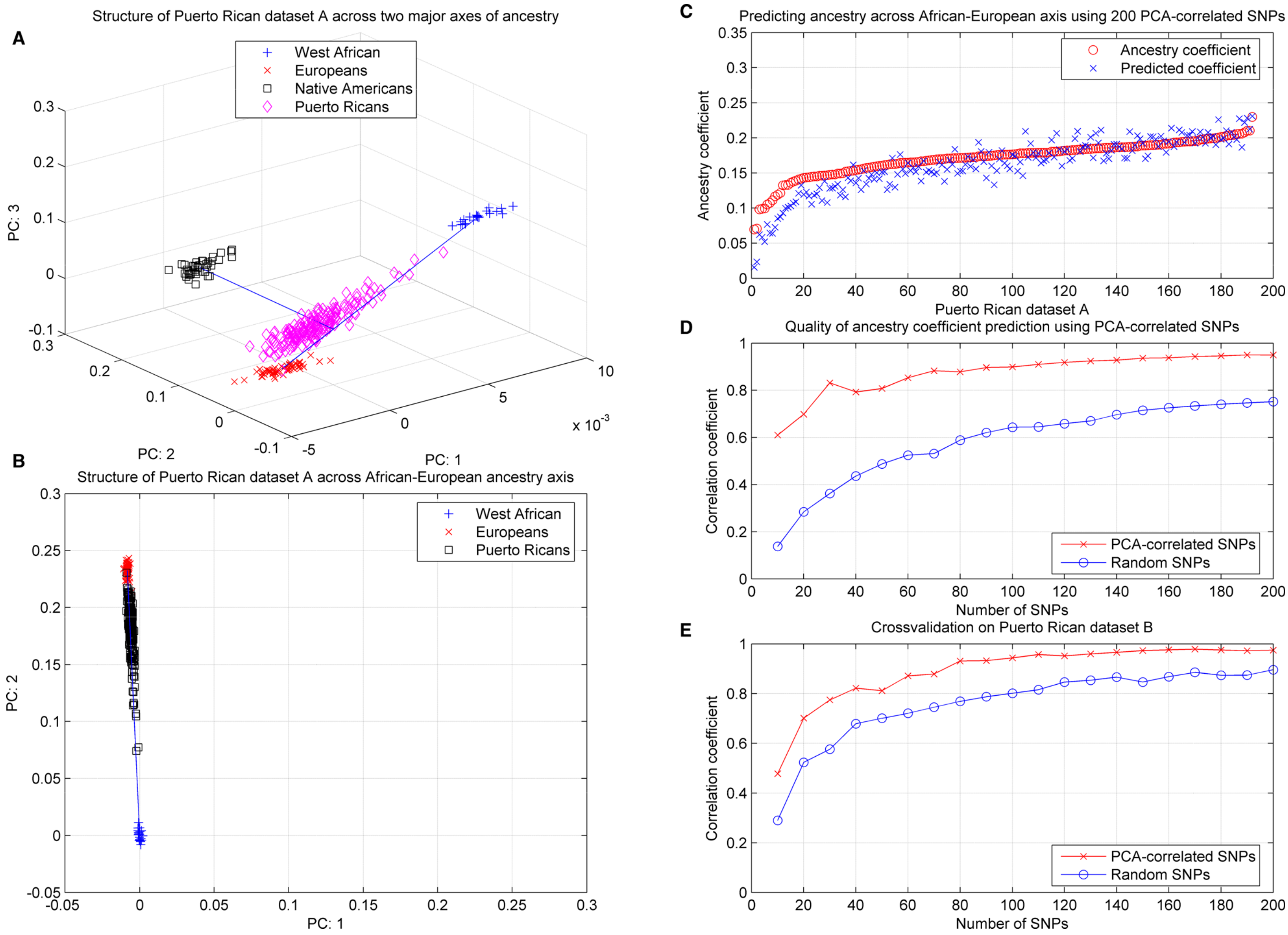

Figure 6. Applying PCA-Correlated SNPs for Structure and Ancestry Prediction of the Admixed Puerto-Rican Population

(A) PCA on 7,259 SNPs typed on Puerto-Rican dataset A, as well as Europeans (Spanish and Caucasians), West Africans (Burunge), and Native Americans (Nahua and Quechua) (axes of variation are shown).

(B) Projection of 192 individuals from Puerto Rican dataset A on two significant principal components and variation across the European-West African axis.

(C) Comparison of ancestry coefficient of 192 Puerto Ricans across the West African-European axis and predicted ancestry coefficient using the top 200 PCA-correlated SNPs.

(D) Prediction of West African-European ancestry coefficient in Puerto Rican dataset A using PCA-correlated SNPs versus random SNPs.

(E) Using PCA-correlated SNPs selected as structure informative in Puerto Rican dataset A for ancestry prediction in Puerto Rican dataset B.

doi:10.1371/journal.pgen.0030160.g006

lyzed. However, we have compared the efficiency of our method to selecting ancestry informative markers using the popular measure of I_n [26]. Rosenberg et al. [26] have previously compared this metric to other frequency-based measures that have been used for the selection of population differentiating markers (F_{ST}, δ, etc.), and concluded that it was well correlated with other measures, equally efficient, and in some cases easier to use. SNPs that were selected using our PCA-correlated measure achieved comparable performance to high ranking I_n SNPs for recovering population structure in the datasets we studied. Interestingly, there is considerable overlap between the SNPs selected by the two different algorithms. It seems that very often, our method selects either the same SNPs or SNPs that are in high LD with those selected using the I_n measure. This is not necessarily surprising, since our approach is ranking markers based on how well they are recreating the fundamental structure in the data, and high-I_n SNPs are those that are also most likely to be associated with major clusters in the genotypic data independent of their location [14].

Dissecting substructure in admixed populations is a central challenge in association studies, especially for common complex disorders [3,5,11,47,61], and our approach may prove to be particularly important in such settings. If allele frequency–based measures are to be used for the identification of a small number of structure informative markers, assumptions need to be made about the origin of the parental source populations and the extent to which they have each contributed to the admixed population. In some cases, ancestral populations may require complex sampling or may even no longer exist [28,62,63]. With the method that we propose here, there is no need to trace the origins of an admixed population in order to define markers that accurately capture the substructure of the population, and our ongoing work is exploring the applicability of our methods on large samples of admixed populations. As we have shown here, analyzing two independent Puerto Rican datasets, PCA-correlated SNPs can be successfully used to reproduce the structure of admixed populations and predict the ancestry proportions of the studied individuals. Interestingly, we found that interindividual variation across the Native American axis in the Puerto Rican samples that we studied was very low, perhaps depicting the fact that admixture with Native Americans occurred very long ago, and was random over several generations.

Our findings demonstrate that to a large extent, SNPs identified as structure informative in one geographic region are not portable for the analysis of populations in a different geographic region, suggesting that the forces that shaped population structure in each geographic region have influenced different parts of the genome. However, analyzing jointly nine populations from around the world and 9,160 SNPs, we showed that using 50 PCA-correlated SNPs we can assign the studied individuals with 100% accuracy to their population of origin. SNPs with high-I_n rankings did not perform any better than random SNPs in this particular setting. One reason underlying the success of our approach may be the fact that it has been explicitly designed to converge to the results of PCA, whereas, to the best of our knowledge, this argument does not necessarily apply to I_n. Nevertheless, I_n does work well in most cases.

Even though our results suggest that our method is powerful enough to be used for the identification of a universal panel of SNPs for the analysis of different populations from around the world, we also showed that each time a new population is added to the analysis, the panel of SNPs needed for population differentiation is modified. So, it should be made clear that we only studied a few representative populations from each continent and much more detailed studies are needed in order to test a universal structure informative SNP panel. This is also true for each of the continental regions that we discussed. We believe that many more population samples should be analyzed in order to accurately define a set of SNPs that could be used to reproduce fine-resolution population structure in a given geographic region.

We have not dealt with the effect of local LD on the results of our algorithm and PCA in general. We showed that given the worldwide dataset that we analyzed here, structure informative SNPs picked by our method are not redundant for the most part in terms of LD. However, as SNP scans become denser, local LD will become a prominent feature of a dataset and we are currently working to see how this affects PCA. At the same time, since our method is not allele frequency based, it is possible that we are able to pick up global correlations among SNPs and haplotype patterns, and more research is necessary to clarify the relationship between the output of PCA and LD.

In summary, we have developed a fast and simple algorithm for the selection of SNPs that uncover the structure of populations without knowing a priori the origin of individuals. After extracting meaningful dimensions from a dataset using PCA, we pick small sets of markers that retain the information carried in the full dataset. We believe that PCA-based algorithms will prove to be an invaluable tool for geneticists in a world of complex and ever-increasing genome-wide data.

Methods

Datasets. The first dataset we used has been described in detail previously [36]. Briefly, we studied here 274 individuals from 12 populations (20 Mbuti, 20 Mende, 22 Burunge, 42 African Americans, 42 Caucasians, 20 Spanish, 11 Mala, 20 East Asians, 20 South Altaians, 20 Nahua, 20 Quechua, and 19 Puerto Ricans). Three of these populations are admixed (Caucasians, African Americans, and Puerto Ricans). All individuals were typed using the 10K Affymetrix array. We also analyzed data available from the HapMap database on four populations (Yoruba, CEPH, Han Chinese, and Japanese; release 21–1r). Finally, we studied a dataset of 192 self-described Puerto Ricans collected in New York and Puerto Rico as part of an asthma association study [43]. This sample has been genotyped using the 100K Affymetrix chip but we only analyzed here genotypes for the 7,259 SNPs that overlapped with the 10K array.

Encoding. We transformed the raw data to numeric values, without any loss of information, in order to apply SVD. Our data on a population X consist of m subjects; for each subject n, biallelic SNPs have been assayed. Thus, we are given a table T^X, consisting of m rows and n columns. Each entry in the table is a pair of bases, ordered alphabetically. We transform this initial data table to an integer matrix A^X, which consists of m rows, one for each subject and n columns, one for each SNP. Each entry of A^X will be −1, 0, +1, or empty. Let B_1 and B_2 be the bases that appear in the j-th SNP (in alphabetical order). If the genotypic information for the j-th SNP of the i-th individual is B_1B_1 the (i, j)-th entry of A^X is set to +1; else if it is B_1B_2 the (i, j)-th entry of A^X is set to 0; else if it is B_2B_2 the (i, j)-th entry of A^X is set to −1.

Handling of missing entries. In order to handle missing data without rejecting too many SNPs that may contain important structural information, we first removed all SNPs with more than 10% missing entries. (This was done independently for each experiment that we ran.) This results in an average of roughly 2% of missing entries in each SNP. We subsequently filled in the missing entries using a least-squares regression-based technique from [64]. This technique fills in the missing data by using all available information from similar SNPs in the matrix. Since we ran this technique on groups of populations and not on each population individually, this filling in of the missing entries would tend to make SNPs more uniform across the different populations, instead of introducing artifactual biases. A more conservative approach would be to randomly fill in the missing entries with −1, 0, or +1 with probabilities respecting Hardy–Weinberg equilibrium. This approach returned similar results in most cases. However, we chose to employ the "best-guess" approach in order to preserve as far as possible the properties of each studied SNP. For the HapMap data on the Han Chinese and Japanese individuals, given the abundance of SNPs, we set the threshold to one missing entry per SNP, which was filled in as described above.

PCA and *k*-means. Given the filled-in data matrix A, we applied SVD on A in order to compute its singular vectors and values. We would like to note that, from a mathematical perspective, our procedure is exactly equivalent to applying PCA on the covariance matrix AA^T, which is an $m \times m$ matrix measuring the angular distance between all pairs of individuals. (From a mathematical perspective, SVD enjoys very strong optimality properties, see [65,66] for details.) After determining the number of significant principal components, k-means clustering was applied on low-dimensional data in order to split the individuals to their respective populations.

For concreteness, consider the SNP data matrix that emerges from the HapMap Han Chinese and Japanese populations, where the data matrix A has 90 rows and approximately 1.7 million columns. Two significant principal components were identified, and we denote the corresponding two left singular vectors by u^1 and u^2 (eigenSNPs). Recall that these are 90-D vectors (each vector has 90 entries, each corresponding to one individual in the Han Chinese–Japanese dataset). Plotting the 90 individuals with respect to u^1 and u^2, i.e., if u_i^1 is the i-th entry in u^1 and u_i^2 is the i-th entry in u^2, the coordinates of the i-th individual are (u_i^1, u_i^2), results in Figure 4. Clearly, the two populations are two separate clusters, with the exception of one individual who is roughly in the middle. Running k-means on these 2-D coordinates results to an almost perfect clustering.

Selecting PCA-correlated SNPs. We summarize the algorithm for selecting PCA-correlated SNPs.

```
Data : $m \times n$ matrix $A$, integer $r$

Result : $r$ PCA − correlated SNPs

Compute the SVD of $A$, $A = \sum_{i=1}^{m} \sigma_i u^i v^{iT}$.

Let $k$ be the number of significant principal components (see below).

Compute $p_j = \sum_{i=1}^{k} (v_j^i)^2$ for all $j = 1 \ldots n$.

Return the columns (SNPs) of $A$ that correspond to the top $r$ $p_j$'s.
```

An implementation of our method is posted at http://www.cs.rpi.edu/~drinep/PCASNPS.

Computing I_n. Informativeness for assignment (I_n) was computed using the algorithm described previously in [26].

Clustering correlation coefficient. In order to compare two clusterings, we simply compute the correlation coefficient (normalized inner product) of the cluster indicator vectors. This is effectively the Pearson correlation coefficient without the mean centering; recall that our vectors are zero–one vectors. For example, in the HapMap Han Chinese and Japanese experiment described above, given a total of 90 individuals, the ground truth cluster indicator vector for the Han Chinese population is a vector whose first 45 entries are set to one and the remaining entries are set to zero. After running *k*-means, two clusters emerge: the one corresponding to the Han Chinese population has the first 45 entries set to one, as well as the 73rd entry, whereas all remaining entries are set to zero. The correlation coefficient between the two indicator vectors (and thus the respective clusters) is 0.99. Correlation coefficients range between zero and one; we report the average correlation coefficient of ground truth clusters and the clusters that emerge after running PCA and *k*-means on the selected sets of SNPs.

Analysis of the Puerto Rican dataset. We outline our analysis of the Puerto Rican dataset A (192 individuals). In Figure 6, we calculated the centroids of the European (Spanish and Caucasians), West African (Burunge), and Native American (Quechua and Nahua) populations. (Four principal components were identified as significant for the joint data, hence we worked on a 4-D space; Figure 6 plots the three most significant dimensions.) We now defined two perpendicular axes of variance: one joining the centroids of the European and West African populations, and the other projecting the centroid of the Native American population on the European–West African axis. We subsequently computed the coordinates of each Puerto-Rican individual on the coordinate system defined by these new axes. We interpreted the resulting coordinates as ancestry information for each individual across the two axes. A simple variance analysis showed that the variance across the European–West African axis is dramatically larger than the variance across the other axis. Hence, we focused our analysis on predicting the relative location of each Puerto Rican subject with respect to the centroids of the European and West African populations using a small subset of PCA-correlated SNPs. We computed the Pearson correlation coefficient between our prediction and the "ground truth" value that was computed using all available SNPs and reported the results in Figure 6. We would like to emphasize that the PCA-correlated SNPs were selecting by looking only at the Puerto Rican dataset A, with no information from the European or West African populations.

Estimating the number of significant principal components. In order to determine whether the *k*-th principal component of an $m \times n$ data matrix A is significant, we will compare the part of A—denoted by A_{m-k}—that corresponds to the *k*-th and smaller principal components

$$A_{m-k} = \sum_{i=k}^{m} (\sigma_i u^i) v^i$$

to a random matrix $\tilde{A}_{m-k}$ that emerges by randomly permuting all the entries of A_{m-k}. (We actually repeat this process ten times and average the results. It should be noted that the variance of this process is very small; for our data, it was orders of magnitude smaller than the mean. Thus, a small number of repetitions suffices.) We now compute the top singular value of A_{m-k} and the top singular value of $\tilde{A}_{m-k}$. This is a standard metric that compares the structure of a given matrix with a random matrix. If the ratio of the top singular value of A_{m-k} over the top singular value of $\tilde{A}_{m-k}$ is more than 115%, then we call the *i*-th principal component significant; otherwise we discard it. Essentially, we retain a principal component if it has 15% more structure than a random one with the same entries. The 15% value was chosen after extensive experimentation on the available data and performed well in all test cases. The method is computationally fast and runs in a few minutes even on the largest dataset we analyzed here (HapMap Han Chinese and Japanese populations). Theoretically, it scales linearly with the number of individuals and number of SNPs.

Two special cases exist. The minimal number of principal components that we keep is at least two. In all populations—except for the combination of Europeans and Spanish—at least two principal components are returned by the aforementioned algorithm as well. However, in order for our PCA-correlated SNPs algorithm to identify the appropriate correlations if exactly one principal component is kept (in which case the associated subspace is just a line), we need some normalization of the original data (e.g., mean centering). To avoid this unnecessary complication, we always keep at least two principal components, which fixes this issue by embedding the data in—at least—the Euclidean plane. The other special case is when too many principal components (e.g., more than 80% of all principal components) are selected by the above algorithm. In this case, we simply skip dimensionality reduction and directly cluster the original data. This never appears when using all SNPs, but may appear when a small number of SNPs is selected from a very large dataset (e.g., ten out of 10,000 SNPs). In Figure S2, we show that as more PCA-correlated SNPs are picked, we can approximately identify the number of principal components that were significant in the original dataset.

Finally, we should mention that the aforementioned test could potentially be replaced by the test proposed by Patterson et al. [20]. The two tests are actually very similar in spirit. They both draw their motivation from the theoretical analysis of the eigenvalues of a matrix whose entries are drawn independently from some distribution with bounded variance. Our test is heavily influenced from the seminal paper of Füredi and Komlós [41], who proved that the eigenvalues of a matrix with the above properties satisfy certain bounds. This provides an elegant way to test for structure in a matrix [42]. Similarly, [20] is influenced by analogous statistical results; we feel that [41,42] require fewer assumptions and thus might be more generally applicable.

Supporting Information

Figure S1. Transferability of Structure Informative SNPs

(A) Examples of transferability of SNPs selected as structure informative in one geographic region (Africa and Europe) for dissecting population structure within a different region. (B) Average correlation coefficient between true and predicted membership of an individual to a particular population within four different geographic regions, using SNPs originally selected for broad intercontinental clustering.

Found at doi:10.1371/journal.pgen.0030160.sg001 (517 KB PDF).

Figure S2. Number of Principal Components Used for *k*-Means Clustering of Populations within and across Continental Regions

The number of principal components is shown for *k*-means clustering, using: (A) all available SNPs for each group of studied populations, (B) selected subsets of PCA-correlated SNPs, (C) high-I_n SNPs, and (D) randomly chosen SNPs.

Found at doi:10.1371/journal.pgen.0030160.sg002(71B PDF).

Acknowledgments

We would like to thank Mark Shriver for providing access to the worldwide dataset that we studied here.

Author contributions. PP, EZ, and PD conceived and designed the experiments. PP and PD performed the experiments and analyzed the data. All authors contributed reagents/materials/analysis tools. PP, EZ, and PD wrote the paper.

Funding. This work was funded, in part, by a National Science Foundation CAREER award to PD, National Institutes of Health U19 AG23122 and K22CA109351, and Department of Defense Breast

Cancer Research Program grant BC033051 to EZ, Hellenic Endocrine Society Research grant award to PP, National Institutes of Health grant HL078885, RO1 HL078885 and Robert Wood Johnson Amos Medical Faculty Development Award to EGB, as well as American Thoracic Society Breakthough Opportunites for Lung Disease (BOLD) grant (ATS-05–078) and Tobacco-Related Disease Research Program New Investigator Award (15KT-0008) to SC.

Competing interests. The authors have declared that no competing interests exist.

References

1. Cavalli-Sforza L, Feldman M (2003) The application of molecular genetic approaches to the study of human evolution. Nat Genet 33: 266–275.
2. Lander E, Schork N (1994) Genetic dissection of complex traits. Science 265: 2037–2048.
3. Ziv E, Burchard E (2003) Human population structure and genetic association studies. Pharmacogenomics 4: 431–441.
4. Marchini J, Cardon L, Phillips M, Donnelly P (2004) The effects of human population structure on large genetic association studies. Nat Genet 36: 512–517.
5. 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.
6. Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55: 997–1004.
7. Pritchard J, Stephens M, Rosenberg N, Donnelly P (2000a) Association mapping in structured populations. Am J Hum Genet 67: 170–181.
8. Reich D, Goldstein D (2001) Detecting association in a case-control study while correcting for population stratification. Genet Epidemiol 20: 4–16.
9. Satten G, Flanders W, Yang Q (2001) Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. Am J Hum Genet 68: 466–477.
10. Hoggart C, Parra E, Shriver M, Bonilla C, Kittles R, et al. (2003) Control of confounding of genetic associations in stratified populations. Am J Hum Genet 72: 1492–1504.
11. Freedman M, Reich D, Penney K, McDonald G, Mignault A, et al. (2004) Assessing the impact of population stratification on genetic association studies. Nat Genet 36: 388–393.
12. Tsai H, Choudhry S, Naqvi M, Rodriguez-Cintron W, Burchard E, et al. (2005) Comparison of three methods to estimate genetic ancestry and control for stratification in genetic association studies among admixed populations. Hum Genet 118: 424–433.
13. 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.
14. Pritchard J, Stephens M, Donnelly P (2000b) Inference of population structure using multilocus genotype data. Genetics 155: 945–959.
15. Falush D, Stephens M, Pritchard J (2003) Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. Genetics 164: 1567–1587.
16. Rosenberg N, Pritchard J, Weber J, Cann H, Kidd K, et al. (2002) Genetic structure of human populations. Science 298: 2381–2385.
17. Kim J, Verdu P, Pakstis A, Speed W, Kidd J, et al. (2005) Use of autosomal loci for clustering individuals and populations of East Asian origin. Hum Genet 117: 511–519.
18. Lao O, vanDuijn K, Kersbergen P, de Knijff P, Kayser M (2006) Proportioning whole-genome single-nucleotide-polymorphism diversity for the identification of geographic population structure and genetic ancestry. Am J Hum Genet 78: 680–690.
19. Liu N, Zhao H (2006) A non-parametric approach to population structure inference using multilocus genotypes. Hum Genomics 2: 353–364.
20. Patterson N, Price A, Reich D (2006) Population structure and eigenanalysis. PLoS Genet 2: e190. doi:10.1371/journal.pgen.0020190
21. Tang H, Peng J, Wang P, Risch N (2005) Estimation of individual admixture: Analytical and study design considerations. Genet Epidemiol 28: 289–301.
22. Paschou P, Mahoney M, Javed A, Kidd J, Pakstis A, et al. (2007) Intra- and interpopulation genotype reconstruction from tagging SNPs. Genome Res 17: 96–107.
23. Menozzi P, Piazza A, Cavalli-Sforza L (1978) Synthetic maps of human gene frequencies in Europeans. Science 201: 786–792.
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. Rosenberg N, Li L, Ward R, Pritchard J (2003) Informativeness of genetic markers for inference of ancestry. Am J Hum Genet 73: 1402–1422.
27. 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.
28. Pfaff C, Barnholtz-Sloan J, Wagner J, Long J (2004) Information on ancestry from genetic markers. Genet Epidemiol 26: 305–315.
29. 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.
30. 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.
31. Wright S (1951) The genetical structure of populations. Ann Eugen 15: 323–354.
32. 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.
33. Drineas P, Kannan R, Mahoney M (2006) Fast Monte Carlo algorithms for matrices III: Computing a compressed approximate matrix decomposition. SIAM J Computing 36: 184–206.
34. Drineas P, Mahoney M, Muthukrishnan S (2006) Sampling algorithms for ℓ_2 regression and applications. Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms: 1127–1136.
35. Drineas P, Mahoney M, Muthukrishnan S (2006) Subspace sampling and relative-error matrix approximation: Column-row-based methods. Proceedings of the 14th Annual European Symposium on Algorithms (ESA): 304–314.
36. 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.
37. The International HapMap Consortium (2003) The International HapMap Project. Nature 426: 789–796.
38. The International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437: 1299–1320.
39. Lin Z, Altman R (2004) Finding haplotype tagging SNPs by use of principal components analysis. Am Journal of Hum Genet 75: 850–861.
40. Skillicorn D (2007) Understanding complex datasets: Data mining using matrix decompositions. Boca Raton (Florida): CRC Press. 260 p.
41. Füredi Z, Komlós J (1981) The eigenvalues of random symmetric matrices. Combinatorica 1: 233–241.
42. Achlioptas D, McSherry F (2001) Fast computation of low rank matrix approximations. Proceedings of the 33rd Annual ACM Symposium on Theory of Computing: 611–618.
43. Burchard E, Avila P, Nazario S, Casal J, Torres A, et al. (2004) Lower bronchodilator responsiveness in Puerto Rican than in Mexican subjects with asthma. Am J Respir Crit Care Med 169.
44. Bowcock A, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd J, et al. (1994) High resolution of human evolutionary trees with polymorphic microsatellites. Nature 368: 455–457.
45. Mountain J, Cavalli-Sforza L (1997) Multilocus genotypes, a tree of individuals, and human evolutionary history. Am J Hum Genet 61: 705–718.
46. Bamshad M, Wooding S, Watkins W, Ostler C, Batzer M, et al. (2003) Human population genetic structure and inference of group membership. Am J Hum Genet 72: 578–589.
47. 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.
48. 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. e143 doi:10.1371/journal.pgen.0020143
49. Kim J, Parker K (1999) Major histocompatibility complex differentiation in Sacramento river chinook salmon. Genetics 151: 1115–1122.
50. Pritchard J, Donnelly P (2001) Case-control studies of association in structured or admixed populations. Theor Popul Biol 60: 227–237.
51. McKeigue P (2005) Prospects for admixture mapping of complex traits. Am J Hum Genet 76: 1–7.
52. Kidd K, Pakstis A, Speed W, Grigorenko E, Kajuna S, et al. (2006) Developing a SNP panel for forensic identification of individuals. Forensic Sci Int 164: 20–32.
53. Jain AK, Dubes RC (1988) Algorithms for clustering data. Englewood Cliffs (New Jersey): Prentice-Hall. 320 p.
54. Smith M, Patterson N, Lautenberger J, Truelove A, McDonald G, et al. (2004) A high-density admixture map for disease gene discovery in African Americans. Am J Hum Genet 74: 1001–1013.
55. Yang N, Li H, Criswell L, Gregersen P, Alarcon-Riquelme M, et al. (2005) Examination of ancestry and ethnic affiliation using highly informative diallelic DNA markers: Application to diverse and admixed populations and implications for clinical epidemiology and forensic medicine. Hum Genet 118: 382–392.
56. Tian C, Hinds D, Shigeta R, Kittles R, Ballinger D, et al. (2006) A genomewide single-nucleotide-polymorphism panel with high ancestry information for African American admixture mapping. Am J Hum Genet 79: 640–649.
57. Tian C, Hinds D, Shigeta R, Adler S, Lee A, et al. (2007) A genomewide single-nucleotidepolymorphism panel for Mexican American admixture mapping. Am J Hum Genet 80: 1014–1023.
58. Price A, Patterson N, Yu F, Cox D, Waliszewska A, et al. (2007) A

genomewide admixture map for latino populations. Am J Hum Genet 80: 1024–1036.

59. Mao X, Bigham A, Mei R, Gutierrez G, Weiss K, et al. (2007) A genomewide admixture mapping panel for hispanic/latino populations. Am J Hum Genet 80: 1171–1178.
60. 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.
61. Helgason A, Yngvadottir B, Hrafnkelsson B, Gulcher J, Stefansson K (2005) An Icelandic example of the impact of population structure on association studies. Nat Genet 37: 90–95.
62. Adams J, Ward R (1973) Admixture studies and the detection of selection. Science 180: 1137–1143.
63. Long J (1991) The genetic structure of admixed populations. Genetics 127: 417–428.
64. Alter O, Brown P, Botstein D (2000) Singular value decomposition for genome-wide expression data processing and modeling. Proc Natl Acad Sci U S A 97: 10101–10106.
65. Horn R, Johnson C (1985) Matrix Analysis. New York: Cambridge University Press. 575 p.
66. Golub G, Loan CV (1989) Matrix Computations. Baltimore: Johns Hopkins University Press. 728 p.