Theoretical Population Biology 107 (2016) 65–76

Contents lists available at ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb

Eigenanalysis of SNP data with an identity by descent interpretation

Xiuwen Zheng, Bruce S. Weir *

Department of Biostatistics, University of Washington, Box 359461, Seattle, WA 98195-9461, USA

ARTICLE INFO

Article history:
Received 10 February 2015
Available online 23 October 2015

Keywords:
PCA
Relatedness
Coancestry
IBD
SNP
Admixture

ABSTRACT

Principal component analysis (PCA) is widely used in genome-wide association studies (GWAS), and the principal component axes often represent perpendicular gradients in geographic space. The explanation of PCA results is of major interest for geneticists to understand fundamental demographic parameters. Here, we provide an interpretation of PCA based on relatedness measures, which are described by the probability that sets of genes are identical-by-descent (IBD). An approximately linear transformation between ancestral proportions (AP) of individuals with multiple ancestries and their projections onto the principal components is found.

In addition, a new method of eigenanalysis "EIGMIX" is proposed to estimate individual ancestries. EIGMIX is a method of moments with computational efficiency suitable for millions of SNP data, and it is not subject to the assumption of linkage equilibrium. With the assumptions of multiple ancestries and their surrogate ancestral samples, EIGMIX is able to infer ancestral proportions (APs) of individuals. The methods were applied to the SNP data from the HapMap Phase 3 project and the Human Genome Diversity Panel. The APs of individuals inferred by EIGMIX are consistent with the findings of the program ADMIXTURE.

In conclusion, EIGMIX can be used to detect population structure and estimate genome-wide ancestral proportions with a relatively high accuracy.

1. Introduction

Principal component analysis was introduced for the study of genetic data almost thirty years ago by Menozzi et al. (1978), and has since become a standard tool. Population differentiation can be inferred from multivariate statistical methods such as PCA of allele frequencies (Menozzi et al., 1978; Cavalli-Sforza and Feldman, 2003). In a new approach, Patterson et al. (2006) applied PCA to SNP genotypic data for individuals rather than populations. Their method, implemented in a software package "EIGENSTRAT", has been widely used to correct for population stratification in genome-wide association studies (GWAS) (Price et al., 2010). Although PCA is not based on a population genetics model, and may seem like a "black box" method, principal component axes often represent perpendicular gradients in geographic space (Cavalli-Sforza and Feldman, 2003; Price et al., 2006; Novembre et al., 2008). The relationship of PCA results to fundamental demographic parameters is of major interest to geneticists.

Novembre and Stephens (2008) showed that the gradient and wave patterns of principal components do not necessarily reflect migration events in history. From the perspective of coalescent theory, McVean (2009) provided a genealogical interpretation of PCA. He showed that the projection of samples onto the principal components could be obtained from the pairwise coalescence times between study individuals. Ma and Amos (2010) proposed a formulation of PCA based on the variance–covariance matrix of the sample allele frequencies.

We now provide an alternative interpretation of PCA based on relatedness measures: probabilities that sets of genes have descended from a single ancestral gene and so are identical by descent (ibd). The ibd concept is essential for genetic analyses such as linkage studies for mapping disease genes and forensic DNA profiling (Weir et al., 2006; Thompson, 2013). In population genetics, Weir and Hill (2002) extended the work of Weir and Cockerham (1984) by allowing different levels of coancestry for different populations, and by allowing non-zero coancestries between pairs of populations. Our further extension is to allow different coancestries between pairs of individuals and different inbreeding coefficients for individuals. The coancestry coefficient between two populations defined in the model of Weir and Hill is now replaced by the average kinship coefficient among pairs of study individuals from these two populations respectively, relative to a single ancestral population, so that the assumption of random

* Corresponding author.
E-mail addresses: zhengx@uw.edu (X. Zheng), bsweir@uw.edu (B.S. Weir).

http://dx.doi.org/10.1016/j.tpb.2015.09.004

rating can be relaxed. These individual-perspective measures of population structure can be used to explain the behavior of PCA.

Ancestral proportions (AP) of an individual refer to the fractions of the genome derived from specific ancestral populations (Pritchard et al., 2000; Falush et al., 2003; Tang et al., 2005; Alexander et al., 2009). The early approach for estimating AP can track back to Hanis et al. (1986), and the ancestral allele frequencies should be known to allow estimating allele admixture in this method. However, ancestral allele frequencies are usually estimated from surrogate ancestral samples in practice and later studies took into account in describing the uncertainty of estimated ancestral information.

A Bayesian approach, STRUCTURE, was developed to infer population substructure using unlinked genotypes (Pritchard et al., 2000). Later, it was extended to model linked markers (Falush et al., 2003) through admixture linkage disequilibrium (LD). STRUCTURE is computationally intensive and not likely to be suitable for large-scale studies, like GWAS, involved with thousands of individuals and hundreds of thousands of SNPs. SNP pruning has to be done before applying STRUCTURE, and this can introduce selection bias with respect to different SNP sets. A maximum-likelihood estimation method, frappe, has also been proposed to estimate AP with much less computation than STRUCTURE, but it assumes the markers are unlinked (Tang et al., 2005). The ADMIXTURE method was developed to analyze thousands of markers — it adopts the likelihood model embedded in STRUCTURE with an assumption of linkage equilibrium among the markers (Alexander et al., 2009).

Instead of estimating global ancestry via genome-wide markers, detection of local ancestry from chromosomal segments in admixed populations becomes of great interest. Recently, HAPMIX and MULTIMIX were proposed to infer local ancestry from dense SNP markers based on approximate coalescent models modeling linkage disequilibrium with two or more ancestries (Price et al., 2009; Churchhouse and Marchini, 2013). However, their methods require a fine genetic map.

The potential connection between ancestral proportions and principal components in the eigenanalysis has been investigated by the previous studies with a limited number of numerical simulations (Patterson et al., 2006; Engelhardt and Stephens, 2010). McVean (2009) indicated it is possible to identify relative admixture proportions from principal components. Ma and Amos (2012) showed how to estimate two-way admixture proportions with a proof under their framework of variance–covariance matrix. They also observed that an admixed population could divide the triangle of three parental populations in the PC plot into three small triangles with areas according to the three-way admixture proportions. However, none of these studies provided a sufficient proof for inferring admixture fractions from the principal components under their theoretical framework in the cases of more than two ancestral populations.

In our study, an approximately linear transformation between ancestral proportions (AP) of individuals with multiple ancestries and their projections onto the principal components is revealed, and a proof is given under the framework of identity by descent. This linear transformation could explain the perpendicular gradients in geographic space, and it also justifies the observation that the ratios of triangle areas correspond to admixture fractions in the study of Ma and Amos (2012). We also propose a new method of eigenanalysis "EIGMIX" to estimate individual ancestries. EIGMIX uses method of moments estimation with computational efficiency suitable for millions of SNP data, and it is not subject to the assumption of linkage equilibrium. Ancestral proportions can be estimated by making assumptions of surrogate samples for ancestral populations, but inferring ancestral allele frequencies is not necessary. The calculation uses all study individuals simultaneously without projecting the remaining individuals onto the existing axes of surrogates.

We applied various methods to the SNP data of 1198 founders from the HapMap Phase 3 project and 938 unrelated individuals from the Human Genome Diversity Project (HGDP). The ancestral proportions of individuals inferred by PCA and EIGMIX are consistent with the findings of the program ADMIXTURE. All eigenanalysis in the study are implemented in the R package "SNPRelate" (Zheng et al., 2012), allowing users to apply our method to their SNP data.

2. Methods

We develop our approach with a series of indicator variables x_{ijkl} for the kth allele, $k = 1, 2$, at the lth locus, $l = 1, 2, \ldots, L$, in the jth individual sampled from the ith population, $j = 1, 2, \ldots, n_i$; $i = 1, 2 \ldots, N$. The total sample size is $n = \sum_i n_i$. The variables take the value 1 for alleles of a specific type, e.g. the reference allele, at a locus, and the value 0 otherwise. Genotypes are indicated by $g_{ijl} = x_{ij1l} + x_{ij2l}$, and these take the values 0, 1, 2.

2.1. Population coancestry framework of Weir and Hill (2002)

Under the framework of Weir and Hill (2002), the expectations for first and second moments of the x's are

$$\begin{aligned}
\mathcal{E}[x_{ijkl}] &= p_l \\
\mathcal{E}[x^2_{ijkl}] &= p_l \\
\mathcal{E}[x_{ijkl}\, x_{ijk'l}] &= p_l^2 + p_l(1-p_l)F_{ij}, \quad k \neq k', \text{ the same individual} \\
\mathcal{E}[x_{ijkl}\, x_{ij'k'l}] &= p_l^2 + p_l(1-p_l)\theta_i, \quad j \neq j', \text{ the same population} \\
\mathcal{E}[x_{ijkl}\, x_{i'j'k'l}] &= p_l^2 + p_l(1-p_l)\theta_{ii'}, \quad i \neq i', \text{ different populations.}
\end{aligned}$$

Here expectation is over both repeated samples from the population and over evolutionary replicates of the populations. These expressions introduce the total inbreeding coefficient F_{ij}, the within-population coancestries θ_i, and the between-population-pair coancestries $\theta_{ii'}$. The quantities p_l are the overall, or ancestral, frequencies of the reference alleles if all study individuals can be traced back to a single reference population. This reference population could be common ancestors at a point in time of the past. The equal values for $\mathcal{E}[x_{ij1l}\, x_{ij2l}]$ and $\mathcal{E}[x_{ijkl}\, x_{ij'k'l}]$ require an assumption of random mating.

The coancestry coefficient θ_i refers to the ibd probability for a random pair of alleles in population i, and the pair of alleles can come from the same individual. The coancestry coefficient $\theta_{ii'}$ refers to the ibd probability for a random pair of alleles, one from population i and the other from population i'. Note that we implicitly assume θ_i and $\theta_{ii'}$ are the same at each locus, and in practice θ_i and $\theta_{ii'}$ are actually the average inbreeding and coancestry coefficients over all L loci.

Now consider an individual perspective measures of population structure, i.e., a special case of Weir and Hill's model where each population i has only one sampled individual ($n_i = 1$) so $j = 1$ for each population. The assumption of random mating is relaxed, and the sample size n is also the number of populations r. Therefore,

$$\begin{aligned}
&\bar{p}_l = \frac{1}{n}\sum_{i=1}^{n}\bar{p}_{il} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{2}\sum_{j=1}^{1}(x_{ij1l} + x_{ij2l})\right] \\
&\mathcal{E}[\bar{p}_l] = p_l \\
&\mathrm{Var}[\bar{p}_{il}] = \frac{1}{2}p_l(1-p_l)(1+\theta_i) \\
&\mathrm{Cov}[\bar{p}_{il}, \bar{p}_{i'l}] = p_l(1-p_l)\theta_{ii'} \\
&\mathrm{Var}[\bar{p}_l] = \frac{n-1}{n}p_l(1-p_l)\theta_T + \frac{1}{2n}p_l(1-p_l)(1+\theta_l) \\
&\mathcal{E}[\bar{p}_l(1-\bar{p}_l)] = \frac{n-1}{n}p_l(1-p_l)(1-\theta_T) + \frac{1}{2n}p_l(1-p_l)(1-\theta_l)
\end{aligned} \tag{1}$$

where $\theta_l = \sum_{i=1}^{n}\theta_i/n$, the average inbreeding coefficient among all study individuals, and $\theta_T = \sum_{i,i'=1,\ i\neq i'}^{n}\theta_{ii'}/[n(n-1)]$,

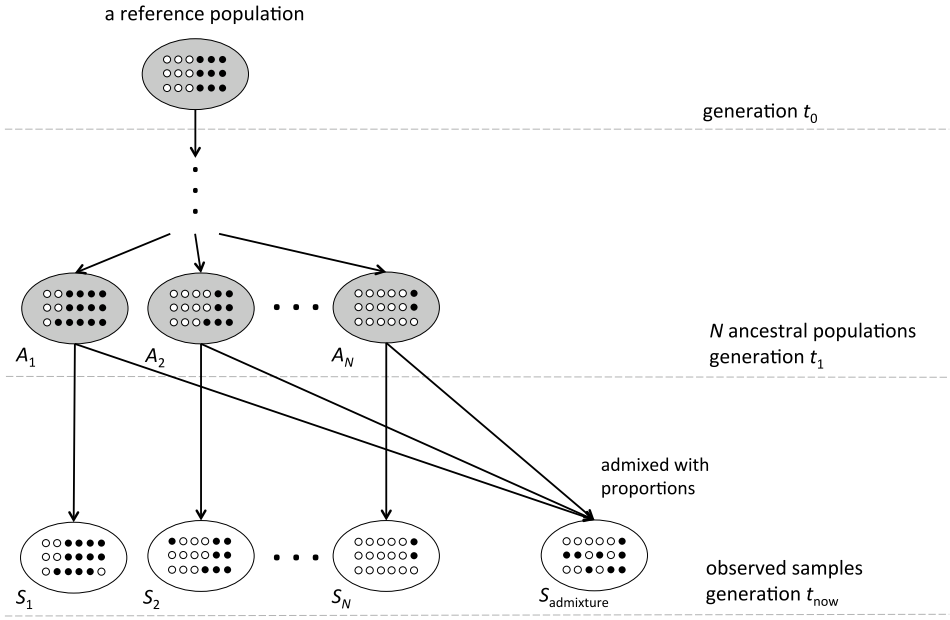


Fig. 1. A genetic model at a single locus for observed samples. The alleles of all study individuals at t_{now} can be tracked to a single reference population at t_0, and there are N distinct ancestral populations at t_1. The relationships among ancestral populations are described by a coancestry matrix Θ_A.

the average kinship coefficient among all study individuals. The individual perspective measures do not account for familial data and the relatedness of individuals is established from evolutionary history.

Each study individual is assigned to one population, thereby the genetic covariance matrix defined by Patterson et al. (2006) at the individual level can be expressed using an index i, $\mathbb{M}^{\text{P}} = [m^{\text{P}}_{i,i'}]_{n\times n}$:

$$m^{\text{P}}_{i,i'} = \frac{1}{L}\sum_{l=1}^{L}\frac{(g_{i1l}-2\bar{p}_l)(g_{i'1l}-2\bar{p}_l)}{\bar{p}_l(1-\bar{p}_l)}. \tag{2}$$

The expected values of the numerator in Eq. (2) is:

$$\begin{aligned}\mathcal{E}[(g_{i1l}-2\bar{p}_l)^2] &= 2p_l(1-p_l)\left(1+\theta_i+2\frac{n-1}{n}\theta_T-4\psi_i\right)\\ &\quad+\frac{2}{n}p_l(1-p_l)(\theta_l+2\theta_i-1), \quad \text{for } i=i'\\ \mathcal{E}[(g_{i1l}-2\bar{p}_l)(g_{i'1l}-2\bar{p}_l)]&\\ =4p_l(1-p_l)&\left(\theta_{ii'}+\frac{n-1}{n}\theta_T-\psi_i-\psi_{i'}\right)\\ &+\frac{2}{n}p_l(1-p_l)(\theta_l+\theta_i+\theta_{i'}-1), \quad \text{for } i\neq i'\end{aligned}$$

where $\psi_i = \sum_{i'=1}^{n}\theta_{ii'}/n$ (setting $\theta_{ii}=\theta_i$).

When the number n of study individuals is large,

$$\mathcal{E}\left[\frac{1}{4}m^{\text{P}}_{i,i'}\right] = \begin{cases}\dfrac{1+\theta_i}{2(1-\theta_T)}+\dfrac{\theta_T-2\psi_i}{1-\theta_T}, & \text{if} \quad i=i'\\ \dfrac{\theta_{ii'}}{1-\theta_T}+\dfrac{\theta_T-\psi_i-\psi_{i'}}{1-\theta_T}, & \text{if} \quad i\neq i'.\end{cases} \tag{3}$$

2.2. Eigen-decomposition in PCA

If we are interested in individual inbreeding coefficients $(1+\theta_i)/2$ (the coancestry of an individual with itself) and individual-pair coancestries $\theta_{ii'}$, the factors $(1-\theta_T)$ and $(\theta_T-\psi_i-\psi_{i'})/(1-\theta_T)$ in Eq. (3) will confound the estimates when $\frac{1}{4}m^{\text{P}}_{i,i'}$ is used. This may explain why a large proportion of $m^{\text{P}}_{i,i'}$ is negative, whereas the true θ_i and $\theta_{ii'}$ are always between zero and one.

2.2.1. The population perspective

PCA conducts eigen-decomposition on the stochastic matrix $\mathbb{M}^{\text{P}}$, and it is possible to investigate the structural features of $\mathbb{M}^{\text{P}}$ with its expectation. To illustrate what eigen-decomposition does, we introduce a genetic model consisting of populations at three points in time as shown in Fig. 1. The alleles of all study individuals at t_{now} can be tracked to a single reference population at t_0 through at least one of distinct ancestral populations at t_1. The study samples $S_1,\ldots,S_N$ are directly inherited from the ancestral populations $A_1,\ldots,A_N$ without admixture, and the sample $S_{\text{admixture}}$ is admixed from N ancestral populations.

What we can observe are the genomes of study individuals at t_{now}. It could be appropriate to assume there are N ancestral populations at t_1 which is between t_0 and t_{now}, and the samples $S_1,\ldots,S_N$ are good candidates (or pseudo-ancestors) to represent the ancestral populations. For example, in the initial phase of the HapMap Project, genetic data were gathered from four populations (CEU, YRI, CHB and JPT) with European, African and Asian ancestry respectively. Here, $N=3$, S_1 represents CEU individuals, S_2 for YRI and S_3 for CHB + JPT.

A coancestry matrix Θ_A is used to describe the relationships among N ancestral populations at t_1 based on population perspective measures, where

$$\Theta_A = \begin{bmatrix}\theta^*_1 & \theta^*_{12} & \cdots & \theta^*_{1N}\\ \theta^*_{12} & \theta^*_2 & \cdots & \theta^*_{2N}\\ \cdots & \cdots & \ddots & \cdots\\ \theta^*_{1N} & \theta^*_{2N} & \cdots & \theta^*_N\end{bmatrix}. \tag{4}$$

That is, θ^*_h is the average IBD probability for a pair of alleles randomly sampled with replacement from the hth ancestral population, and $\theta^*_{hh'}$ is the coancestry coefficient for random pairs of individuals from the hth and h'th ancestral populations respectively. Since we track all individuals back to the reference population at t_0, the sample allele frequencies at t_1 are treated as random variables over a probability space, which starts from the reference population at t_0 and arrives at t_1 with the coancestry state Θ_A.

2.2.2. Ancestral proportions

In practice individuals may have recent ancestors in more than one population, and an admixture model is introduced in which each individual is assumed to have inherited some proportion of its ancestry from each population. For an individual i, let the ancestral proportions be a vector $\mathbf{a}_i = (a_{i,1}, \ldots, a_{i,N})^T$, where $\sum_{h=1}^{N} a_{i,h} = 1$ and $0 \le a_{i,h} \le 1$. Let $Z_{iklh} = 1$ when the kth allele of individual i at SNP l is inherited from the hth ancestral population at t_1, and $Z_{iklh} = 0$ otherwise. The vector $\mathbf{Z}_{ikl} = \{Z_{ikl1}, \ldots, Z_{iklN}\}^T$ is modeled as a random variable with probabilities $\mathbf{a}_i$, i.e., $\mathcal{E}[Z_{iklh}] = a_{i,h}$. Further,

$$a_{i,h} = \mathcal{E}\left[\frac{1}{L}\sum_{l=1}^{L} Z_{iklh}\right] \tag{5}$$

represent the genomic ancestral proportions. Note that Eq. (5) still holds even if loci are correlated due to linkage disequilibrium. We assume that the two alleles in individual i at SNP l are independently derived from ancestral populations, since pairs of chromosomes of an individual are independently inherited from two parents respectively. Then the expected value of the inbreeding coefficient at SNP l for individual i is $\mathcal{E}[\mathbf{Z}_{i1l}^T \Theta_A \mathbf{Z}_{i2l}] = \mathbf{a}_i^T \Theta_A \mathbf{a}_i$, the same for each SNP. The average inbreeding coefficient over L loci is $\theta_{ii'} = \mathbf{a}_i^T \Theta_A \mathbf{a}_i$, assuming the coancestry matrix of ancestral populations is identical at each locus.

For a pair of individuals i and i', we assume that any pair of alleles, one from i and the other from i' are independently derived from ancestral populations. Then the expected value of the kinship coefficient at SNP l is $\mathcal{E}[\mathbf{Z}_{ikl}^T \Theta_A \mathbf{Z}_{i'k'l}] = \mathbf{a}_i^T \Theta_A \mathbf{a}_{i'}$, and the average kinship coefficient over L loci is also $\theta_{ii'} = \mathbf{a}_i^T \Theta_A \mathbf{a}_{i'}$. This assumption is appropriate to model relatedness in structured population with admixture, with $\mathbf{a}_i^T \Theta_A \mathbf{a}_{i'}$ as background relatedness due to evolutionary history. However, the validity of the assumption could be violated if individuals i and i' are in a family, e.g., parent and offspring.

2.2.3. Matrix decomposition

For a study sample, there are n unrelated individuals. Each individual i has AP $\mathbf{a}_i$ with respect to N ancestral populations. Let $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n]^T$ be a n-by-N matrix with rows representing ancestral proportions of individuals. Then the coancestry matrix of study individuals Θ_S can be expressed as

$$\Theta_S = \mathbf{A}\Theta_A\mathbf{A}^T. \tag{6}$$

We rewrite Eq. (3) in matrix notation for large n,

$$\mathcal{E}[\mathbb{M}^{\mathrm{P}}] = \frac{4}{1-\theta_T}\underbrace{\left(\mathbf{A} - \tfrac{1}{n}\mathbf{J}_n\mathbf{A}\right)\Theta_A\left(\mathbf{A} - \tfrac{1}{n}\mathbf{J}_n\mathbf{A}\right)^T}_{\stackrel{\text{def}}{=}\Theta_{\mathbf{M}}} + \underbrace{\mathrm{diag}\left(\tfrac{2(1-\theta_1)}{1-\theta_T}, \ldots, \tfrac{2(1-\theta_n)}{1-\theta_T}\right)}_{\text{bias}} \tag{7}$$

where $\mathbf{J}_n$ is a matrix of dimension $n \times n$ with entries equal to one, since

$$\begin{aligned}\left(\frac{1}{n}\mathbf{J}_n\mathbf{A}\right)\Theta_A\left(\frac{1}{n}\mathbf{J}_n\mathbf{A}\right)^T &= \frac{1}{n^2}\mathbf{J}_n\Theta_S\mathbf{J}_n \\ &= \frac{1}{n^2}\left(\sum_{j\neq j'}\theta_{jj'} + \sum_j \theta_j\right)\mathbf{J}_n \\ &= \left(\theta_T\frac{n(n-1)}{n^2} + \frac{1}{n}\theta_I\right)\mathbf{J}_n \\ &\approx \theta_T\mathbf{J}_n\end{aligned}$$

$$\left(\frac{1}{n}\mathbf{J}_n\mathbf{A}\right)\Theta_A\mathbf{A}^T = \frac{1}{n}\mathbf{J}_n\Theta_S = \begin{bmatrix}\psi_1 & \psi_2 & \cdots & \psi_n \\ \vdots & \vdots & \ddots & \vdots \\ \psi_1 & \psi_2 & \cdots & \psi_n\end{bmatrix}.$$

The diagonal $\mathrm{diag}(\frac{2(1-\theta_1)}{1-\theta_T}, \ldots, \frac{2(1-\theta_n)}{1-\theta_T})$ is considered as a bias term in the PCA with respect to ancestral proportions.

Note that $\mathrm{rank}(\mathbf{A} - \frac{1}{n}\mathbf{J}_n\mathbf{A}) \le N - 1$ because we lose a dimension by forcing each column to sum to zero. The eigenvectors corresponding to the largest $N-1$ eigenvalues of $\Theta_{\mathbf{M}}$ form a new coordinate with $N-1$ dimensions while AP form an old N-dimensional coordinates. The mapping from the old coordinate to the new one is a linear transformation, and the proof is given in the Appendix A.1. In addition, this mapping is actually an affine transformation equivalent to a $(N-1)$-dimensional linear transformation followed by a translation, and the affine transformation can be represented as a linear transformation on the higher dimensional space.

For example, assume there are three ancestral populations and seven individuals, in which individuals 1, 2, 3 are inherited from the ancestral populations without admixture, individuals 4, 5, 6 have two ancestral populations with equal contributions and individual 7 has three ancestral populations with equal contributions. The matrix $\mathbf{A}$ of ancestral proportions is

$$\mathbf{A} = \begin{bmatrix}1 & 0 & 0 & 1/2 & 1/2 & 0 & 1/3 \\ 0 & 1 & 0 & 1/2 & 0 & 1/2 & 1/3 \\ 0 & 0 & 1 & 0 & 1/2 & 1/2 & 1/3\end{bmatrix}^T,$$

and Θ_A is assumed to $\mathrm{diag}(0.05, 0.05, 0.05)$. The AP coordinates are shown in Fig. 2(a), and the new eigen-decomposition coordinates are shown in Fig. 2(b).

2.3. EIGMIX — Inferring ancestral proportions

The mapping in Fig. 2 suggests an approach to estimate ancestral proportions using the largest principal components. Let $S_1, \ldots, S_N$ be the observed surrogate samples for the ancestral populations, as shown in Fig. 1. Now we look at the largest $(N-1)$ principal components, and identify each location of pseudo-ancestor $i \in \{1, \ldots, N\}$ in the eigen coordinates, by averaging the locations of the sample S_i. So we have N positions in the eigen coordinates, which corresponds to N independent components in the AP coordinates. Then a linear transformation can be made to reverse the original mapping, i.e., the principal components of all study individuals are reversed to the AP coordinates by a linear transformation. In addition, the property of linear mapping makes the inferred ancestral proportions unique if N surrogate samples are specified and their locations in the eigen coordinates are distinct.

For example, the positions of individuals 1, 2 and 3 with ancestral proportions (1,0,0), (0,1,0) and (0,0,1) in the eigen coordinate of Fig. 2(b) are denoted by e_{s1}, e_{s2} and e_{s3} respectively. Let $T_{2\times2}$ be a linear transformation and L be a translation operator. A transformation from the AP coordinates to the eigen coordinates is:

$$\begin{bmatrix}1 & 0 \\ 0 & 1 \\ 0 & 0\end{bmatrix}T_{2\times2} + L_{3\times2} = \begin{bmatrix}e'_{s1} \\ e'_{s2} \\ e'_{s3}\end{bmatrix}. \tag{8}$$

Therefore, $L_{3\times2} = [e_{s3}, e_{s3}, e_{s3}]'$ (moving every point a constant distance) and $T_{2\times2} = [e_{s1} - e_{s3}, e_{s2} - e_{s3}]'$. The inverse transformation is:

$$\text{ancestral proportion} = (e_{\text{admix}} - e_{s3})\, T_{2\times2}^{-1} \tag{9}$$

where e_{admix} is an arbitrary point in the eigen coordinate.

Note that there is a bias term in the diagonal shown in Eq. (7). A scheme for bias removal is to define a new genetic covariance

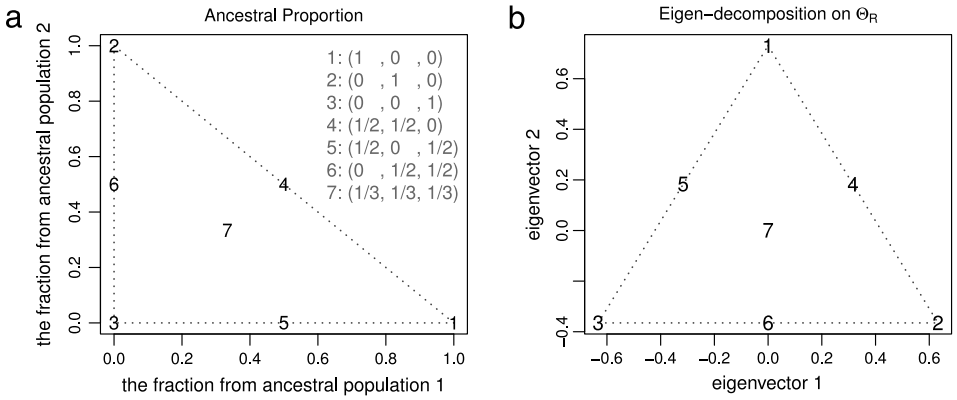


Fig. 2. The relationship between ancestral proportions and eigen-decomposition: (a) seven admixture fractions from three ancestral populations are plotted in the figure; (b) the first and second eigenvectors of matrix $\Theta_{\mathbf{M}} = (\mathbf{A} - \frac{1}{n}\mathbf{J}_n\mathbf{A})\Theta_A(\mathbf{A} - \frac{1}{n}\mathbf{J}_n\mathbf{A})^T$, where the ancestral coancestry matrix Θ_A is assumed to diag(0.05, 0.05, 0.05), $\mathbf{A}$ is an n-by-N matrix with rows representing admixture proportions of individuals, $n = 7$ and $N = 3$. The mapping from the two-dimensional coordinate in (a) to that of (b) is a linear transformation followed by a translation.

matrix, the EIGMIX coancestry matrix $\mathbb{M}^* = [m^*_{j,j'}]_{n\times n}$, EIGMIX coancestry matrix:

$$m^*_{j,j'} = \begin{cases} \dfrac{\sum_{l=1}^{L}(g_{jl} - 2\bar{p}_l)^2 - g_{jl}(2 - g_{jl})}{4\sum_{l=1}^{L}\bar{p}_l(1-\bar{p}_l)}, & j = j' \\ \dfrac{\sum_{l=1}^{L}(g_{jl} - 2\bar{p}_l)(g_{j'l} - 2\bar{p}_l)}{4\sum_{l=1}^{L}\bar{p}_l(1-\bar{p}_l)}, & j \neq j'. \end{cases} \tag{10}$$

Then $\mathcal{E}[\mathbb{M}^*] = \Theta_{\mathbf{M}}/(1-\theta_T)$ without any bias when there are a large number of individuals. We have previously (Weir and Cockerham, 1984) suggested the simple modification of taking the ratios of the sums over loci of the numerators and denominators instead of averaging the ratios to reduce the variance, in part by reducing the impact of rare variants. Since the ratio of expected values is an approximation for the expected value of an ratio of two random variables, our modification tends to have an advantage of bias correction due to division compared to the original PCA.

In practice, the matrix $\mathbb{M}^*$ of real data could have more than $N-1$ significant eigenvalues when we assume the number of ancestral populations N to be a specific number (e.g., $N = 3$ for Europe, Asia and Africa). The largest $N-1$ eigenvalues with their eigenvectors form a low-rank approximation of $\mathbb{M}^*$ (a real symmetric matrix), which minimizes the Frobenius norm with respect to a n-by-n matrix M with $\text{rank}(M) \leq N-1$:

$$\|\mathbb{M}^* - M\|_F^2 = \sum_{j=1}^{n}\sum_{j'=1}^{n} m^2_{j,j'}$$

where $\mathbb{M}^* - M = [m_{j,j'}]_{n\times n}$. The closest matrix to $\mathbb{M}^*$ is $\hat{M} = \sum_{i=1}^{N-1}\lambda_i\mathbf{e}_i\mathbf{e}_i^T$, as measured in the Frobenius norm, where $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|$ are the eigenvalues of $\mathbb{M}^*$ and $\mathbf{e}_i$ is the eigenvector corresponding to λ_i, and

$$\|\mathbb{M}^* - \hat{M}\|_F^2 = \sum_{i=N}^{n}\lambda_i^2$$

$\mathbb{M}^*$ is not necessarily a nonnegative definite matrix, i.e., its eigenvalues are not necessarily all nonnegative. Here "largest eigenvalues" refer to the absolute values of eigenvalues in descending order.

In addition, the estimates of EIGMIX given an arbitrary number of ancestral populations are not always bounded from 0 to 1, although we force the proportions to sum to one. If the inferred ancestral proportions lie much outside the range [0, 1], signaling outliers, we could conclude that the assumption of N ancestral populations with their surrogates is not appropriate or that the SNP markers have no power to distinguish ancestral populations.

According to PCA, we might expect the eigen-decomposition of $\mathcal{E}[\frac{1}{4}\mathbb{M}^{\mathrm{P}}]$ and $\mathcal{E}[\mathbb{M}^*]$ could result in similar eigenvectors corresponding to a few most significant eigenvalues when there are true structural feature in data, since the difference between $\mathcal{E}[\frac{1}{4}\mathbb{M}^{\mathrm{P}}]$ and $\mathcal{E}[\mathbb{M}^*]$ depends only on the diagonal. The average difference per entry in the term of Frobenius norm becomes small when the total number of study individuals n is large:

$$\frac{1}{n^2}\left\|\mathcal{E}\left[\tfrac{1}{4}\mathbb{M}^{\mathrm{P}}\right] - \mathcal{E}[\mathbb{M}^*]\right\|_F^2 = \frac{1}{4n^2}\sum_{j=1}^{n}\frac{(1-\theta_j)^2}{(1-\theta_T)^2} \to 0,$$

as $n \to \infty$.

A few largest eigenvalues and eigenvectors could capture the similar structure information of $\mathcal{E}[\frac{1}{4}\mathbb{M}^{\mathrm{P}}]$ and $\mathcal{E}[\mathbb{M}^*]$. Here, "similar" means similar relative positions in the eigen coordinates, since numerical calculation does not guarantee that the resulting eigenvectors will have the same absolute positions in the coordinate, e.g., if a vector v is an eigenvector then $-v$ is also the eigenvector according to the same eigenvalue. A further numerical study is shown in the Appendix A.2.

3. Results

3.1. Materials

The Phase 3 HapMap data consist of SNP genotypes generated from 1397 samples in total, collected using two platforms: the Illumina Human1M (by the Wellcome Trust Sanger Institute) and the Affymetrix SNP 6.0 (by the Broad Institute) (International HapMap 3 Consortium et al., 2010). Data from the two platforms have been merged for the release. The PLINK format of HapMap 3 data was downloaded from http://hapmap.ncbi.nlm.nih.gov/downloads/genotypes/hapmap3_r3/plink_format/. The consensus and polymorphic dataset of 1198 founders were used in the study analyses, which include only SNPs that passed quality control in all populations, as shown in Table 1.

The Human Genome Diversity Panel data consist of 1043 individuals from 51 populations over the world: sub-Saharan Africa, North Africa, Europe, the Middle East, Central and South Asia, East Asia, Oceania and the Americas (Cann et al., 2002). The study individuals were genotyped on the Illumina 650 K platform, and the SNP data could be downloaded from http://www.hagsc.org/hgdp/

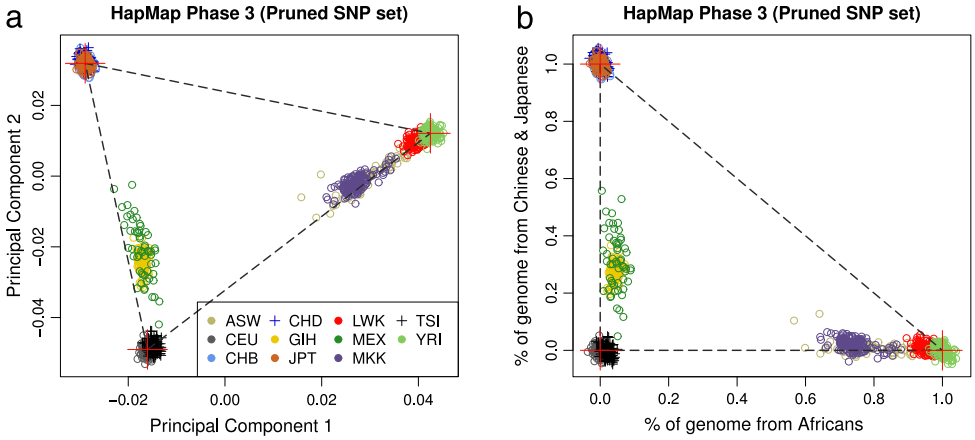


Fig. 3. The principal component analysis on HapMap Phase 3 data, using a pruned set of 9949 SNPs and 1198 founders consisting of 11 populations: (a) the first and second eigenvectors; (b) a linear transformation of coordinate from (a) followed by a translation, assuming three ancestral populations with surrogate samples: CEU, YRI and CHB + JPT. The average positions of three surrogate samples are masked by a red plus sign.

Table 1
Summary of population samples in the eigenanalysis.

Name	Population	# of samples
HapMap Phase III (1198 founders):		
ASW	African ancestry in Southwest USA	53
CEU	Utah residents with Northern and Western European ancestry from the CEPH collection	112
CHB	Han Chinese in Beijing, China	137
CHD	Chinese in Metropolitan Denver, Colorado	109
GIH	Gujarati Indians in Houston, Texas	101
JPT	Japanese in Tokyo, Japan	113
LWK	Luhya in Webuye, Kenya	110
MEX	Mexican ancestry in Los Angeles, California	58
MKK	Maasai in Kinyawa, Kenya	156
TSI	Toscani in Italia	102
YRI	Yoruba in Ibadan, Nigeria	147
The Human Genome Diversity Panel (HGDP, 938 unrelated individuals):		
Africa		101
Europe		157
Middle East		163
Central and South Asia		199
East Asia		228
Oceania		26
America		64

files.html. The dataset contains a small number of relatives, and 938 individuals remained in the analysis after filtering out first and second degree relatives of which were suggested by Rosenberg (2006).

To reduce potential effects of linkage disequilibrium, SNP pruning was conducted by randomly selecting autosomal SNPs for which each pair was at least as far apart as 200 kb: 9949 remaining SNPs for HapMap Phase 3 and 9790 for HGDP. All analyses were performed on both of the pruned and full SNP sets, and the unbound estimates of ancestral proportion are reported. In the full sets, there are 1 423,833 and 644,258 autosomal SNPs for HapMap3 and HGDP respectively.

3.2. Analyses of HapMap Phase 3 data

To avoid the confounding effect of relatives, 1198 founders were selected for the PCA analysis by removing the offspring. The first two principal components are the focus, since more eigenvectors provide little additional information for inferring primary population structure. As shown in Fig. 3(a), the samples from CEU, YRI and CHB + JPT correspond to three vertices of a triangle, and the other populations tend to be admixtures from these three ancestries. Inferring ancestral proportions was conducted by a coordinate transformation, assuming three ancestral populations with surrogate samples: CEU, YRI and CHB + JPT. The X and Y axes in Fig. 3(b) represent the proportions of genome from African and Asian ancestries respectively. Gujarati Indians in Houston (GIH, yellow) and Mexican ancestry in Los Angeles (MEX, green) appear to be admixtures between Europeans and Asians. ASW, MKK and LWK tend to be more related to African ancestry with some admixture, while CHD and TSI are quite close to the surrogate samples of Asia. The PCA plot with the largest two principal components generated by the full SNP set is shown in Fig. B.1), which is similar to Fig. 3.

The population admixture proportions are estimated by averaging ancestral proportions of individuals using the full SNP set. African Americans (ASW) are a typically admixed sample, estimated with ∼78% of genome from YRI and 21% from CEU, and approximately no genome from CHB + JPT. The result confirms the estimates of 78% African and 22% European ancestry shown in the supplementary materials of the HapMap Phase 3 report (International HapMap 3 Consortium et al., 2010). The HAPMIX algorithm (Price et al., 2009) was used in HapMap Phase 3 project, the optimal linear combination of 74% YRI and 26% CEU was observed for MKK, and a combination of 94% YRI and 6% CEU for LWK. In our analyses, the PCA-inferred combinations are 74% YRI + 24% CEU for MKK and 94% YRI + 5% CEU for LWK. Our results are consistent with the admixture proportions previously estimated.

The supervised ADMIXTURE and EIGMIX methods were applied to the HapMap3 SNP data assuming three ancestral populations with surrogate samples CEU, YRI and CHB + JPT. ADMIXTURE is a model-based method with an assumption of markers in linkage equilibrium, therefore a pruned SNP set was used to avoid the strong influence of SNP clusters. The pseudo-ancestors (YRI, CHB + JPT and CEU) are specified in the analyses of ADMIXTURE according to the AP (1, 0, 0), (0, 1, 0) and (0, 0, 1). As shown in Fig. 4, the AP inferred by PCA tend to be consistent with those estimated by ADMIXTURE using the same SNP set. However, the offsets are observed for admixed populations, such like GIH and MKK. The PCA-based proportions of genome from CEU are lower than ADMIXTURE for GIH, and those are higher for MKK. Actually, our inference on MKK was actually consistent with what HapMap Phase 3 has reported. Note that PCA is a dimension reduction technique and may lose information if we look only at the largest two principal components, and the assumption of pseudo-ancestors (CEU, YRI, CHB + JPT) might not truly represent the ancestors in human evolution.

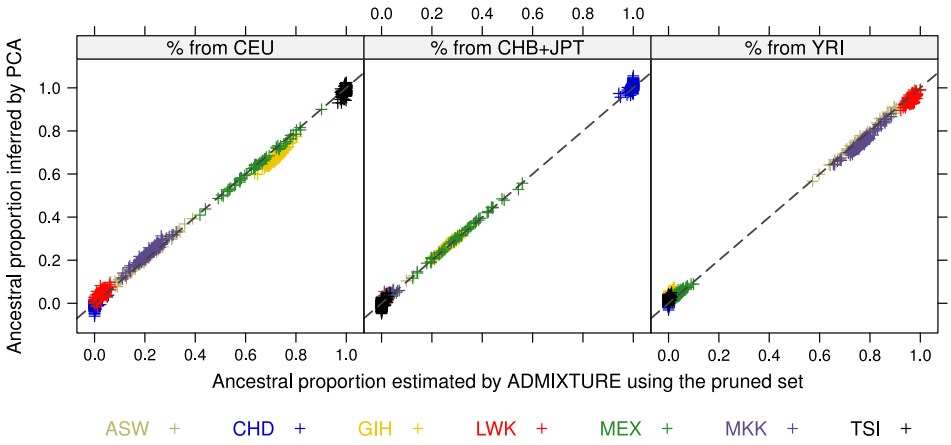


Fig. 4. A comparison between PCA and supervised ADMIXTURE with respect to ancestral proportions for the HapMap Phase 3 data. A pruned set of 9949 SNPs was used by both PCA and ADMIXTURE.

Table 2
The differences on ancestral proportions of individuals between supervised ADMIXTURE and the eigenanalysis for the HapMap Phase 3 data.

Pop.	PCA − ADMIXTURE mean ± sd			EIGMIX − ADMIXTURE mean ± sd		
	% CEU	% CHB + JPT	% YRI	% CEU	% CHB + JPT	% YRI
ASW	0.30 ± 0.75	−0.29 ± 1.11	−0.01 ± 0.90	0.18 ± 1.09	−0.27 ± 1.24	0.09 ± 0.73
CHD	−1.10 ± 1.65	1.09 ± 1.66	0.01 ± 1.15	−1.16 ± 1.82	1.11 ± 1.70	0.05 ± 1.34
GIH	−4.15 ± 0.85	0.42 ± 0.57	3.74 ± 0.87	−3.39 ± 0.84	−0.60 ± 0.74	3.98 ± 1.12
LWK	1.50 ± 1.33	0.24 ± 1.22	−1.74 ± 1.25	0.86 ± 1.42	−0.02 ± 1.35	−0.84 ± 1.03
MEX	−0.62 ± 0.70	0.23 ± 0.57	0.40 ± 0.75	−0.31 ± 0.91	0.02 ± 0.83	0.29 ± 1.01
MKK	1.60 ± 0.85	0.61 ± 1.10	−2.21 ± 0.77	0.72 ± 1.07	0.33 ± 1.03	−1.05 ± 0.66
TSI	−1.07 ± 1.74	−0.78 ± 1.69	1.84 ± 1.12	−0.85 ± 1.73	−1.14 ± 1.80	1.99 ± 1.32

The EIGMIX coancestry matrix was used in the eigenanalysis instead of the PCA covariance matrix. As shown in Table 2, the differences of ancestral proportions at the individual level between ADMIXTURE and PCA/EIGMIX were calculated to evaluate the potential biases compared to the estimates of ADMIXTURE. The estimated proportions of EIGMIX tend to be less biased than PCA's except Chinese in Metropolitan Denver (CHD), whereas the differences are relatively small overall for the HapMap3 data. The variances of EIGMIX are comparable to PCA if the ADMIXTURE estimates are assumed to be true values.

3.3. Analyses of HGDP

As suggested by Rosenberg (2006), a standardized subset of HGDP data consisting of 938 unrelated individuals was employed in the admixture analyses with a pruned set of 9790 SNPs. The number of ancestral populations is suggested by geographic regions, the worldwide human relationship inference (Rosenberg et al., 2002; Li et al., 2008) and the plots of eigenvectors (shown in Fig. B.2), and we used six ancestries in our primary analyses. The surrogate samples are suggested by the previous inferred regional ancestry (Li et al., 2008) and relative positions in the plots of eigenvectors: Sardinian for Europe ($n = 28$), Chinese Han for East Asia ($n = 44$), Kalash for Central and South Asia ($n = 22$), Pygmy for Africa ($n = 34$), Karitiana for America ($n = 14$) and Papuan for Oceania ($n = 16$).

The supervised ADMIXTURE and EIGMIX methods were both applied to the HGDP SNP data with six ancestral populations. The estimated ancestral proportions of individuals are shown in Fig. 5. Overall the estimates of EIGMIX are consistent with what ADMIXTURE does, however a difference of 10% admixture proportion is observed for samples from Africa and Middle East when the percents of Europe are inferred. In Fig. 5(e), the samples of America are also observed to be off the diagonal line. PCA was applied to the same study individuals and SNP set: the PCA-inferred admixed ancestries are shown in Fig. 6 and Fig. B.3. The PCA method is observed to have higher variance than EIGMIX, especially for the samples from Africa and Middle East. The variance reduction in EIGMIX is primarily due to the modification of taking the ratios of the sums over loci, rather than diagonal bias removal.

4. Discussion

In this study, we provide an interpretation of principal components analysis (PCA) based on relatedness measures, i.e., the probability that sets of genes are identical-by-descent. The expected values of pairwise estimates in the genetic covariance matrix of PCA are relative kinship coefficients with an additional term with respect to a single reference population in the past. An approximately linear transformation between ancestral proportions of individuals with multiple ancestries and their projections onto the principal components is revealed. A new method "EIGMIX" is proposed to estimate ancestries, allowing both linked and unlinked genetic markers regardless of linkage disequilibrium. The ancestral proportions can be estimated by making assumptions of surrogate ancestral samples. EIGMIX is a method of moments with high computational efficiency compared to existing MLE and Bayesian methods such like ADMIXTURE and STRUCTURE, and it is suitable to large-scale GWAS data with thousands of individuals and millions of SNPs. We applied the PCA, EIGMIX and supervised ADMIXTURE methods to the real SNP data from the HapMap Phase 3 project and the Human Genome Diversity Panel. The ancestral proportions inferred by PCA and EIGMIX are consistent with the findings of ADMIXTURE, but EIGMIX proportions are observed to be less biased and more robust than PCA.

Novembre et al. (2008) showed that SNP profiles of individuals within Europe can be used to infer their geographic origin with relatively high accuracy by PCA. The reason why the PC axes often represent perpendicular gradients in geographic space can

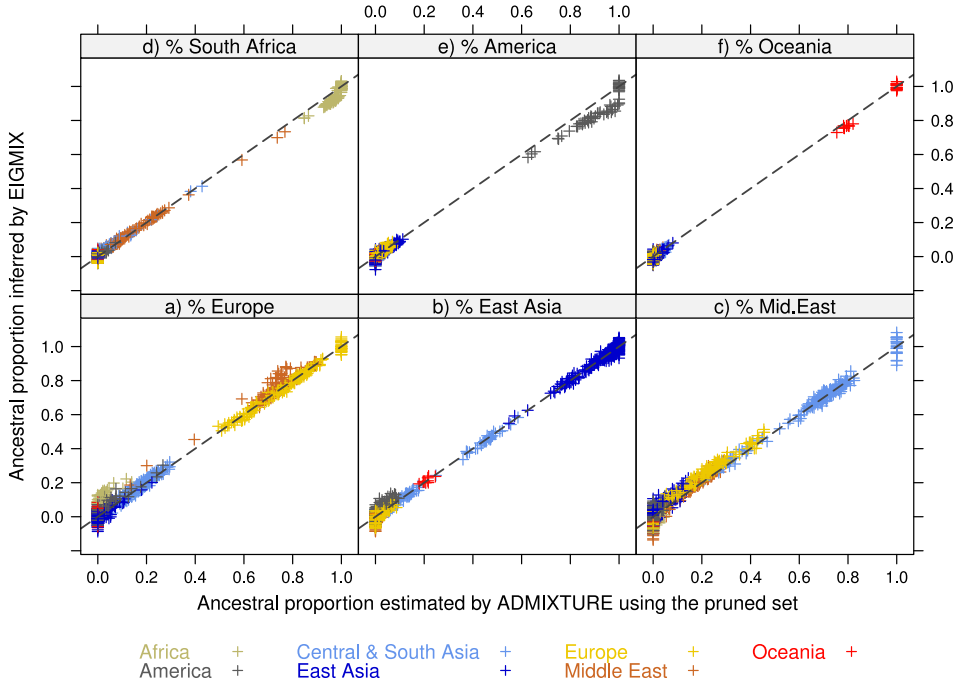


Fig. 5. A comparison of ancestral proportions between EIGMIX and supervised ADMIXTURE with 6 ancestral populations for the HGDP data. A pruned set of 9790 SNPs was used by both EIGMIX and ADMIXTURE.

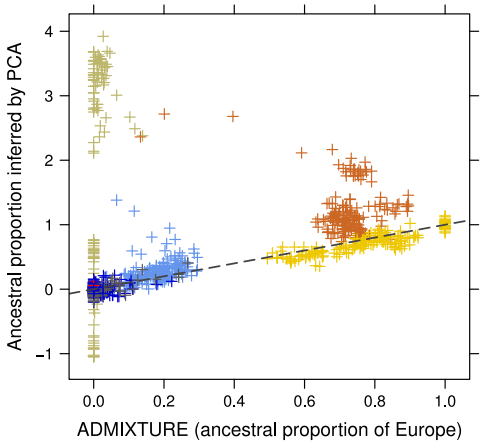


Fig. 6. A comparison of ancestral proportions between PCA and supervised ADMIXTURE with 6 ancestral populations with a pruned set of 9790 SNPs. The color legend is as the same as Fig. 5, and EIGMIX is more robust than PCA when inferring admixture fractions.

be explained by ancestral proportions with two or more ancestries. In our genetic model (see Fig. 1), the time t_0 of single reference population is not specified explicitly, and it could be many generations ago — even the time before modern humans' ancestors migrated out of Africa. The repeated migration in the history of Europe could create gene frequency clines as suggested by isolation-by-distance models (Wright, 1943). Starting from the single reference population at t_0, such as the population at the time before humans migrated out of Africa, it would be possible to treat the observed alleles and the hidden pattern of ibd in the current generation as a sample from the probability space of a long-term evolutionary process. However, this strategy could be confounded by the unknown allele frequencies in the reference population. To avoid this problem, the derivation of the formulas in PCA and EIGMIX have removed explicit use of the allele frequencies.

Ma and Amos (2012) observed that a three-way admixed population could divide the triangle of parental populations in the PC plot into three small triangles with areas according to their admixture proportions. They also tried to extend this observation to the general case of more than three parental populations. A closed-form estimator of ancestral proportion is difficult to find so they solved the eigenequation numerically to confirm the observation. Our mathematical derivation of the linear transformation between ancestral proportions and eigenvectors can be used to confirm the observation of Ma and Amos (2012). Here, we adopt an three-way admixed example with four populations (P_1, P_2, P_3 and P_4) shown in Fig. 5 of the paper of Ma and Amos (2012), where P_4 is an admixed population. It is shown in Fig. B.4. The mapping from the two-dimensional coordinate in Figure B4(a) to that of (b) is an affine transformation. Sets of parallel lines remain parallel after an affine transformation, and it also preserves ratios of distances between points lying on a straight line. Therefore, the ratio of heights in the triangles remain the same. Ma and Amos' observation can be confirmed theoretically under their framework with our linear transformation proof.

It is important to realize the potential limitations and our findings should be interpreted with caution. The assumption of ancestral populations used in inferring admixture fractions from the largest principal components could be confounded by the fact that human evolution is complex and has involved repeated migration and admixture from and out of Africa (Cavalli-Sforza and Feldman, 2003; Abi-Rached et al., 2011). Therefore, the selection of surrogate samples could be biased due to lack of historical knowledge or true unknown ancestries. For example, it is known that Mexicans have mainly Native Americans and European ancestry, with a small African contribution (Price et al., 2007). The ancestral proportions of MEX in HapMap Phase 3 data are confounded by an unknown link between Amerindians and CHB + JPT, although Amerindian seems closely related to Asian rather than European and African in genetics. Also, CHB + JPT and Native Americans represent two evolution branches from their common ancestors, and it may not be appropriate to assume a simple linear combination to reflect genetic difference in Native Americans.

The number of ancestral populations N is another important issue when we infer admixture proportions. A statistical test for

how many significant eigenvalues in SNP data has been proposed, which is based on the approximate Tracy–Widom distribution (Patterson et al., 2006). The potential impacts on this test include linkage disequilibrium and categorical genetic data, since the Tracy–Widom distribution was originally developed for the case of independent Gaussian matrix entries. The MLE method for selecting N based on AIC (Akaike information criterion) and BIC (Bayesian information criterion) statistics was also introduced with ADMIXTURE (Alexander et al., 2009). However, we suggest that the choice of N should rely on the knowledge of the history of a population, with limited advice from statistical significance.

In summary, we provide a genetic interpretation of PCA, and propose EIGMIX to infer ancestral proportions with relatively high accuracy. EIGMIX could help us better understand population structure for isolated and admixed populations.

Acknowledgments

We appreciate the input from W.G. Hill. This work was supported in part by NIH grants GM 075091 and GM 099568.

Appendix A

A.1. Proof of eigen-decomposition

Here, we perform eigen-decomposition on $\Theta_{\mathbf{M}} = (\mathbf{A} - \frac{1}{n}\mathbf{J}_n\mathbf{A})\Theta_A(\mathbf{A} - \frac{1}{n}\mathbf{J}_n\mathbf{A})^T$ in Eq. (7), and the mapping from $\mathbf{A}$ to the eigenvectors of $\Theta_{\mathbf{M}}$ is a linear transformation, where $\mathbf{A}$ is a n-by-N matrix with rows representing ancestral proportions of individuals and Θ_A is a N-by-N coancestry matrix. Let $Y = \mathbf{A} - \frac{1}{n}\mathbf{J}_n\mathbf{A} = (I_n - \frac{1}{n}\mathbf{J}_n)\mathbf{A}$, where I_n is an identity matrix and $\mathbf{J}_n$ is a matrix $n \times n$ with entries equal to one, then $\Theta_{\mathbf{M}} = Y\Theta_A Y^T$.

Proof. Note that $\Theta_{\mathbf{M}}$ and Θ_A are not necessarily non-negative definite matrices, and some of the eigenvalues could be negative. To avoid a complex matrix, we perform eigen-decomposition on $\Theta^2_{\mathbf{M}}$, since $\Theta^2_{\mathbf{M}}$ and $\Theta_{\mathbf{M}}$ have the same eigenvectors and the square of eigenvalues of $\Theta_{\mathbf{M}}$ correspond to the eigenvalues of $\Theta^2_{\mathbf{M}}$.

Note that $\text{rank}(Y) \le N - 1$, then $\text{rank}(\Theta_{\mathbf{M}}) \le N - 1$. Let the eigenvalues of $\Theta_{\mathbf{M}}$ be $|v_1| \ge |v_2| \ge \cdots \ge |v_{N-1}| \ge |v_N| = \cdots = |v_n| = 0$, and $Q_{(\mathbf{M}),i}$ be the ith eigenvector with respect to v_i. $[Q_{(\mathbf{M}),1}, \ldots, Q_{(\mathbf{M}),n}]$ forms an orthogonal matrix.

$$\Theta^2_{\mathbf{M}} = Y\Theta_A Y^T Y\Theta_A Y^T. \tag{11}$$

We perform singular value decomposition on Y,

$$\text{SVD}: Y = U_Y \Sigma_Y V_Y^T.$$

Since $\text{rank}(Y) \le N - 1$, at least one of the singular values of Y is ZERO. Replace Y in Eq. (11) by $U_Y\Sigma_Y V_Y^T$:

$$\Theta^2_{\mathbf{M}} = (U_Y\Sigma_Y V_Y^T\Theta_A V_Y)(\Sigma_Y^T\Sigma_Y)(V_Y^T\Theta_A V_Y\Sigma_Y^T U_Y^T)$$

where $\Sigma_Y^T\Sigma_Y$ forms an $N \times N$ diagonal matrix.

Let $Z_Y = U_Y\Sigma_Y V_Y^T\Theta_A V_Y(\Sigma_Y^T\Sigma_Y)^{\frac{1}{2}}$, where $\Theta^2_{\mathbf{M}} = Z_Y Z_Y^T$. SVD on $Z_Y = U_Z\Sigma_Z V_Z^T$. Again, at least one of the singular values of Z is ZERO.

Since

$$\Theta^2_{\mathbf{M}} = Z_Y Z_Y^T = U_Z\Sigma_Z V_Z^T V_Z\Sigma_Z^T U_Z^T = U_Z\Sigma_Z\Sigma_Z^T U_Z^T,$$

U_Z is the eigenvector matrix of $\Theta^2_{\mathbf{M}}$, i.e., $[Q_{(\mathbf{M}),1}, \ldots, Q_{(\mathbf{M}),n}] = U_Z$ and the eigenvalue $|v_i|$ is the singular value of Z_Y (non-negative).

Table A.1
The bias of estimating population admixture proportions in the example of two ancestral populations and three admixed populations with equal sample size n_{pop}.

True ancestral proportion	0	0.25	0.5	0.75	1
Inferred population ancestral proportion from $\mathcal{E}[\frac{1}{4}\mathbb{M}^{\text{P}}]$[a]:					
$n_{\text{pop}} = 1$	0	0.20758	0.50000	0.79242	1
$n_{\text{pop}} = 25$	0	0.24849	0.50000	0.75151	1
$n_{\text{pop}} = 50$	0	0.24925	0.50000	0.75075	1
$n_{\text{pop}} = 100$	0	0.24962	0.50000	0.75038	1

[a] Calculated by averaging admixture proportion of individuals.

Table A.2
The bias of estimating ancestral proportions in the example of a spatially continuous admixed population with n individuals in total.[a].

# of individuals n	11	51	101	251	501
The maximum bias of inferred ancestral proportions of individuals from $\mathcal{E}[\frac{1}{4}\mathbb{M}^{\text{P}}]$	0.02270	0.00548	0.00281	0.00114	0.00057

[a] Ancestral proportions are uniformly distributed from 0 to 1 derived from two ancestral populations.

Note that

$$\begin{aligned} U_Z\Sigma_Z = Z_Y V_Z &= (U_Y\Sigma_Y V_Y^T)\Theta_A V_Y(\Sigma_Y^T\Sigma_Y)^{\frac{1}{2}}V_Z \\ &= Y\Theta_A V_Y(\Sigma_Y^T\Sigma_Y)^{\frac{1}{2}}V_Z \\ &= \left(I_n - \tfrac{1}{n}\mathbf{J}_n\right)\mathbf{A}\Theta_A V_Y(\Sigma_Y^T\Sigma_Y)^{\frac{1}{2}}V_Z \end{aligned}$$

or,

$$\underbrace{[Q_{(\mathbf{M}),1}, \ldots, Q_{(\mathbf{M}),N}]\text{diag}(|v_1|, \ldots, |v_N|)}_{\text{eigen coordinate}}$$

$$= \left(I_n - \tfrac{1}{n}\mathbf{J}_n\right)\underbrace{\mathbf{A}}_{\text{AP coordinate}}\Theta_A V_Y(\Sigma_Y^T\Sigma_Y)^{\frac{1}{2}}V_Z. \quad \square \tag{12}$$

The left hand side of Eq. (12) is an $n \times N$ matrix where the last column is ZERO since $v_N = 0$, where as the right hand side is the AP matrix times $(I_n - \frac{1}{n}\mathbf{J}_n)$ and $\Theta_A V_Y(\Sigma_Y^T\Sigma_Y)^{\frac{1}{2}}V_Z$. Note that this transformation matrix $\Theta_A V_Y(\Sigma_Y^T\Sigma_Y)^{\frac{1}{2}}V_Z$ is a function of $\mathbf{A}$. Given an AP matrix $\mathbf{A}$, the transform matrix is determined, so each data point (ancestral proportion) in $\mathbf{A}$ maps to a new coordinate by a linear transformation.

A.2. Numerical evaluation of diagonal bias in PCA

To demonstrate the similarity of relative positions in the eigen coordinates of $\mathcal{E}[\frac{1}{4}\mathbb{M}^{\text{P}}]$ and $\mathcal{E}[\mathbb{M}^*]$, two pseudo-ancestor populations ($N = 2$) and three admixed populations (admixture fractions 25%, 50%, 75%) with equal sample sizes were utilized here. As shown in Table A.1, as the sample size of each population grows, the bias for estimating the true admixture fraction 25% and 75% declines from 0.0424 to 0.0004. Another example is a spatially continuous admixed population, i.e., individuals with ancestral proportions uniformly distributed from 0 to 1. E.g., if $n = 11$ is the total number of study individuals, there are 11 individuals with admixture fractions of 0%, 10%, 20%, …, 90% and 100%. The maximum bias of the estimated ancestral proportions is shown in Table A.2, and it decreases from 0.02270 to 0.00057 as the total number of individuals n increases.

Appendix B. Supplementary materials

See Figs. B.1–B.4.

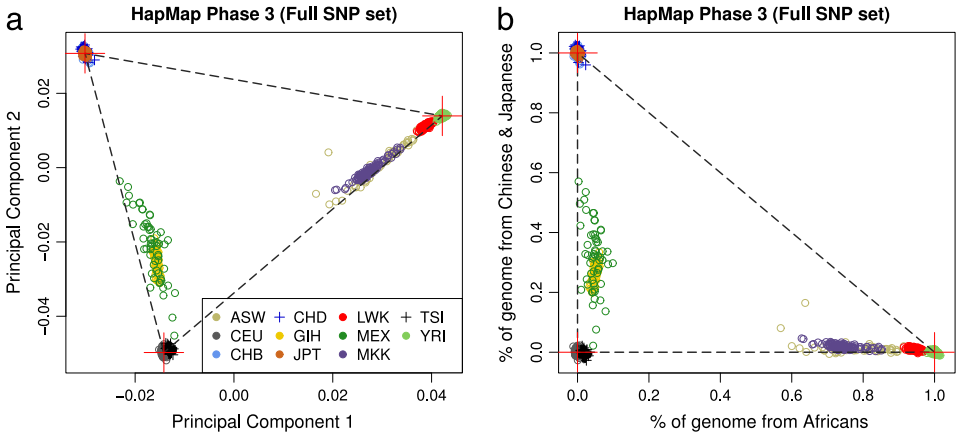


Fig. B.1. The principal component analysis on HapMap Phase 3 data, using the full set of 1,423,833 SNPs and 1,198 founders consisting of 11 populations: (a) the first and second eigenvectors; (b) a linear transformation of coordinate from (a) followed by a translation, assuming three ancestral populations with surrogate samples: CEU, YRI and CHB + JPT. The average positions of three surrogate samples are masked by a red plus sign. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

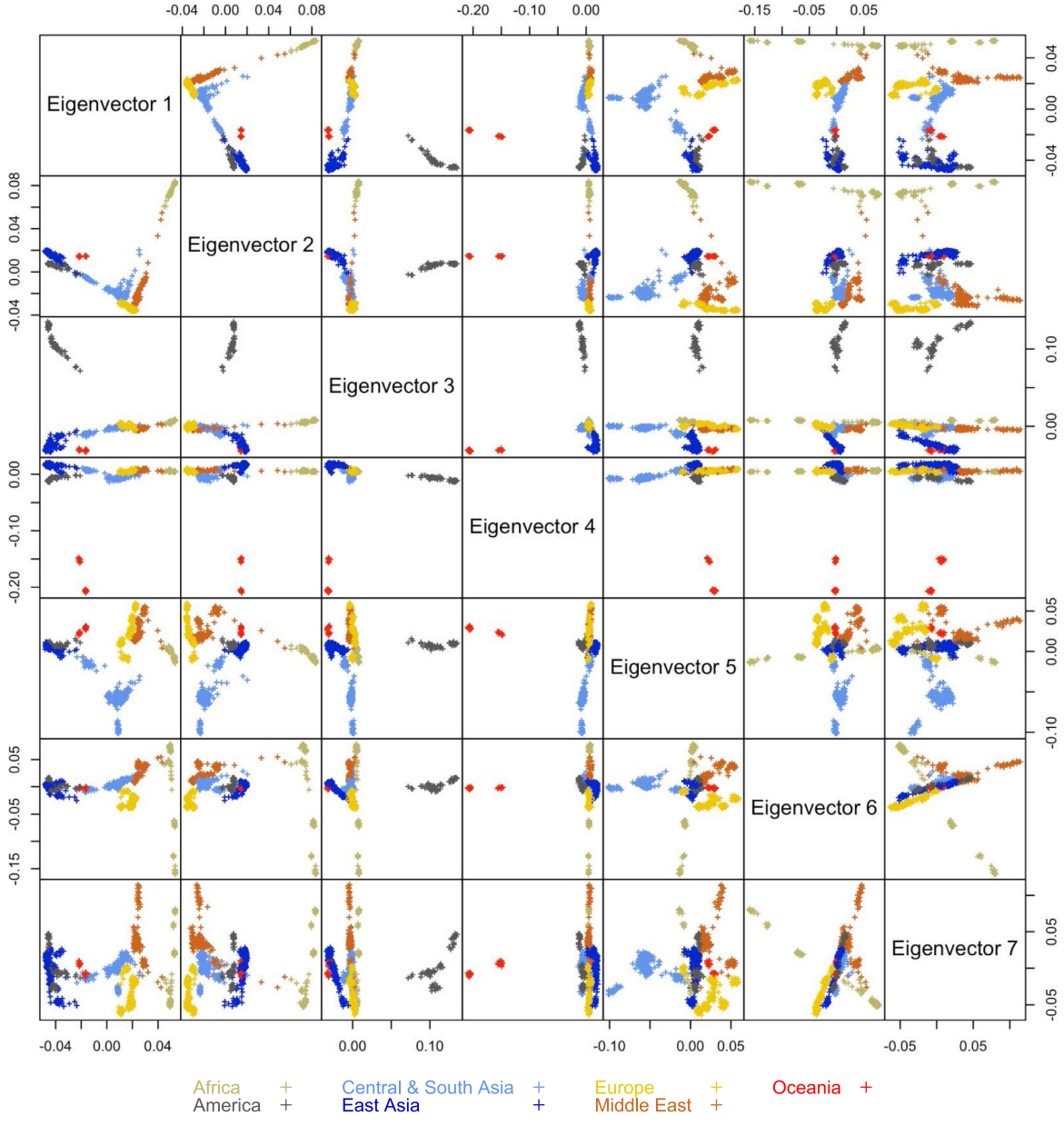


Fig. B.2. The eigenanalysis of EIGMIX on HGDP data using a full set of 644,258 SNPs.

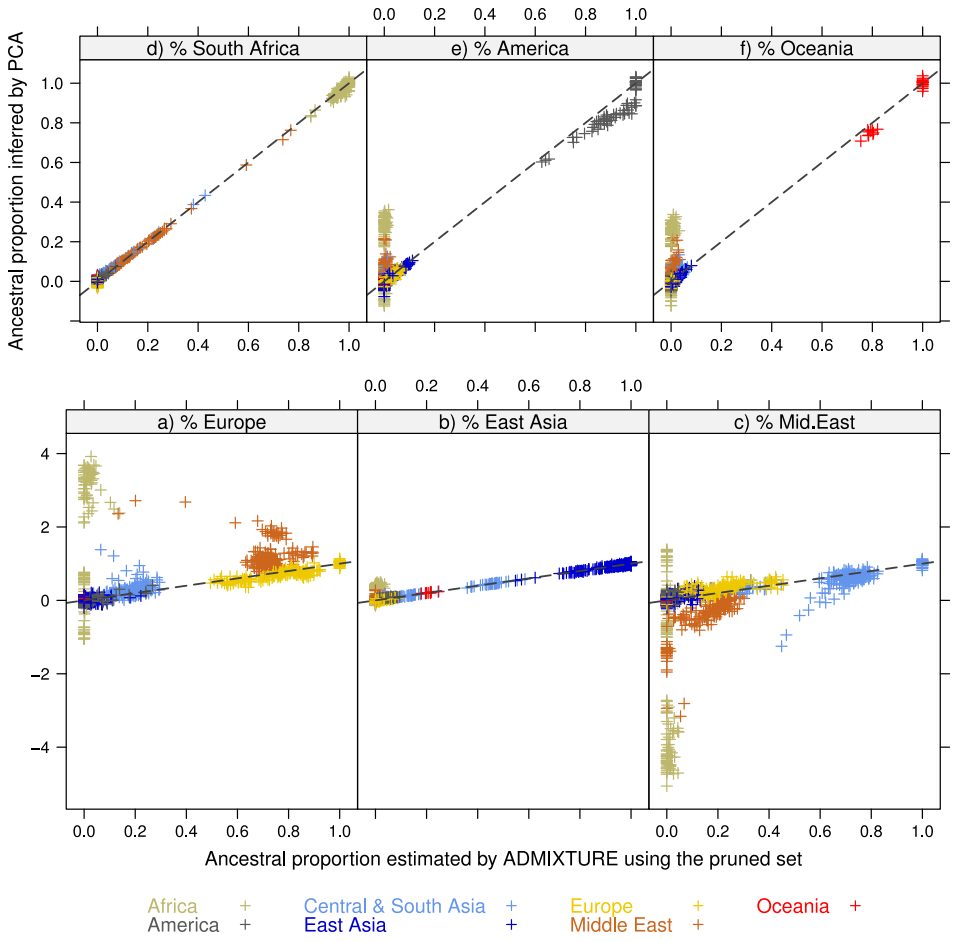
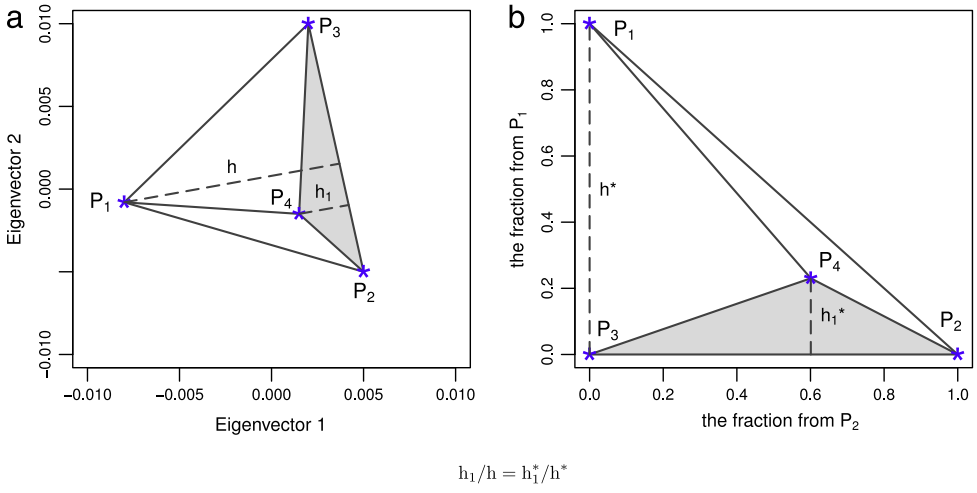


Fig. B.3. A comparison of ancestral proportions between PCA and supervised ADMIXTURE with six ancestral populations with a pruned set of 9790 SNPs.



$$h_1/h = h_1^*/h^*$$

Fig. B.4. The eigenvectors of the four hypothetical populations for three-way admixture. Ma and Amos (2012) observed that a three-way admixed population P_4 could divide the triangle of parental populations (P_1, P_2 and P_3) into three small triangles with areas according to their admixture proportions, i.e., the area ratio ($\Delta P_2P_3P_4 : \Delta P_1P_2P_3$) is the genome fraction of P_4 from P_1. This area ratio equals to the ratio of heights of the base P_2P_3, h_1/h. The mapping from the two-dimensional coordinate in (a) to that of (b) is an affine transformation, which is a linear transformation followed by a translation. Sets of parallel lines remain parallel after an affine transformation, and it also preserves ratios of distances between points lying on a straight line. Therefore, the ratio of new heights h_1^*/h^* in the coordinate of (b), equals to h_1/h in (a). Ma and Amos' observation can be confirmed under their theoretical framework with our proof of linear transformation.

References

Abi-Rached, L., Jobin, M.J., Kulkarni, S., McWhinnie, A., Dalva, K., et al., 2011. The shaping of modern human immune systems by multiregional admixture with archaic humans. Science 334, 89–94.

Alexander, D.H., Novembre, J., Lange, K., 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19, 1655–1664.

Cann, H.M., de Toma, C., Cazes, L., Legrand, M.-F., Morel, V., et al., 2002. A human genome diversity cell line panel. Science (New York, NY) 296, 261–262.

Cavalli-Sforza, L., Feldman, M., 2003. The application of molecular genetic approaches to the study of human evolution. Nature Genet. 33, 266–275.

Churchhouse, C., Marchini, J., 2013. Multiway admixture deconvolution using phased or unphased ancestral panels. Genet. Epidemiol. 37, 1–12.

Engelhardt, B.E., Stephens, M., 2010. Analysis of population structure: a unifying framework and novel methods based on sparse factor analysis. PLoS Genet. 6, e1001117.

Falush, D., Stephens, M., Pritchard, J.K., 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164, 1567–1587.

Hanis, C.L., Chakraborty, R., Ferrell, R.E., Schull, W.J., 1986. Individual admixture estimates: disease associations and individual risk of diabetes and gallbladder disease among Mexican–Americans in Starr County, Texas. Am. J. Phys. Anthropol. 70, 433–441.

International HapMap 3 Consortium, Altshuler, D.M., Gibbs, R.A., Peltonen, L., Altshuler, D.M., et al., 2010. Integrating common and rare genetic variation in diverse human populations. Nature 467, 52–58.

Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., et al., 2008. Worldwide human relationships inferred from genome-wide patterns of variation. Science (New York, NY) 319, 1100–1104.

Ma, J., Amos, C.I., 2010. Theoretical formulation of principal components analysis to detect and correct for population stratification. PLoS One 5.

Ma, J., Amos, C.I., 2012. Principal components analysis of population admixture. PLoS One 7, e40115.

McVean, G., 2009. A genealogical interpretation of principal components analysis. PLoS Genet. 5.

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

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A.R., et al., 2008. Genes mirror geography within Europe. Nature 456, 98–101.

Novembre, J., Stephens, M., 2008. Interpreting principal component analyses of spatial population genetic variation. Nat. Genet. 40, 646–649.

Patterson, N., Price, A.L., Reich, D., 2006. Population structure and eigenanalysis. PLoS Genet. 2.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., et al., 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38, 904–909.

Price, A.L., Patterson, N., Yu, F., Cox, D.R., Waliszewska, A., et al., 2007. A genome-wide admixture map for Latino populations. Am. J. Hum. Genet. 80, 1024–1036.

Price, A.L., Tandon, A., Patterson, N., Barnes, K.C., Rafaels, N., et al., 2009. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS Genet. 5.

Price, A.L., Zaitlen, N.A., Reich, D., Patterson, N., 2010. New approaches to population stratification in genome-wide association studies. Nat. Rev. Genet. 11, 459–463.

Pritchard, J.K., Stephens, M., Donnelly, P., 2000. Inference of population structure using multilocus genotype data. Genetics 155, 945–959.

Rosenberg, N.A., 2006. Standardized subsets of the hgdp-ceph human genome diversity cell line panel, accounting for atypical and duplicated samples and pairs of close relatives. Ann. Hum. Genet. 70, 841–847.

Rosenberg, N.A., Pritchard, J.K., Weber, J.L., Cann, H.M., Kidd, K.K., et al., 2002. Genetic structure of human populations. Science 298, 2381–2385.

Tang, H., Peng, J., Wang, P., Risch, N.J., 2005. Estimation of individual admixture: analytical and study design considerations. Genet. Epidemiol. 28, 289–301.

Thompson, E.A., 2013. Identity by descent: variation in meiosis, across genomes, and in populations. Genetics 194, 301–326.

Weir, B.S., Anderson, A.D., Hepler, A.B., 2006. Genetic relatedness analysis: modern data and new challenges. Nat. Rev. Genet. 7, 771–780.

Weir, B.S., Cockerham, C.C., 1984. Estimating F-statistics for the analysis of population structure. Evolution 38, 1358–1370.

Weir, B.S., Hill, W.G., 2002. Estimating F-statistics. Annu. Rev. Genet. 36, 721–750.

Wright, S., 1943. Isolation by distance. Genetics 2, 114–138.

Zheng, X., Levine, D., Shen, J., Gogarten, S.M., Laurie, C., et al., 2012. A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics (Oxford, England) 28, 3326–3328.