OPEN ACCESS Freely available online

PLoS GENETICS

A Genealogical Interpretation of Principal Components Analysis

Gil McVean*

Department of Statistics, University of Oxford, Oxford, United Kingdom

Abstract

Principal components analysis, PCA, is a statistical method commonly used in population genetics to identify structure in the distribution of genetic variation across geographical location and ethnic background. However, while the method is often used to inform about historical demographic processes, little is known about the relationship between fundamental demographic parameters and the projection of samples onto the primary axes. Here I show that for SNP data the projection of samples onto the principal components can be obtained directly from considering the average coalescent times between pairs of haploid genomes. The result provides a framework for interpreting PCA projections in terms of underlying processes, including migration, geographical isolation, and admixture. I also demonstrate a link between PCA and Wright's F_{ST} and show that SNP ascertainment has a largely simple and predictable effect on the projection of samples. Using examples from human genetics, I discuss the application of these results to empirical data and the implications for inference.

Citation: McVean G (2009) A Genealogical Interpretation of Principal Components Analysis. PLoS Genet 5(10): e1000686. doi:10.1371/journal.pgen.1000686

Editor: Molly Przeworski, University of Chicago, United States of America

Received June 2, 2009; **Accepted** September 16, 2009; **Published** October 16, 2009

Funding: The work was supported by the Leverhulme Trust (www.leverhulme.ac.uk) and the HFSP (www.hfsp.org; grant no. RGP0054/2006). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The author has declared that no competing interests exist.

* E-mail: mcvean@stats.ox.ac.uk

Introduction

The distribution of genetic variation across geographical location and ethnic background provides a rich source of information about the historical demographic events and processes experienced by a species. However, while colonization, isolation, migration and admixture all lead to a structuring of genetic variation, in which groups of individuals show greater or lesser relatedness to other groups, making inferences about the nature and timing of such processes is notoriously difficult. There are three key problems. First, there are many different processes that one might want to consider as explanations for patterns of structure in empirical data and efficient inference, even under simple models can be difficult. Second, different processes can lead to similar patterns of structure. For example, equilibrium models of restricted migration can give similar patterns of differentiation to non-equilibrium models of population splitting events (at least in terms of some data summaries such as Wright's F_{ST}). Third, any species is likely to have experienced many different demographic events and processes in its history and their superposition leads to complex patterns of genetic variability. Consequently, while there is a long history of estimating parameters of demographic models from patterns of genetic variation, such models are often highly simplistic and restricted to a subset of possible explanations.

An alternative approach to directly fitting models is to use dimension-reduction and data summary techniques to identify key components of the structure within the data in a model-free manner. Perhaps the most widely used technique, and the most important from a historical perspective, is principal components analysis (PCA). Technical descriptions of PCA can be found elsewhere, however, its key feature is that it can be used to project samples onto a series of orthogonal axes, each of which is made up of a linear combination of allelic or genotypic values across SNPs or other types of variant. These axes are chosen such that the projection of samples along the first axis (or first principal component) explains the greatest possible variance in the data among all possible axes. Likewise, projection of samples onto the second axis maximizes the variance for all possibles axes perpendicular to the first and so on for the subsequent components. Typically, the positions of samples along the first two or three axes are presented, although methods for obtaining the statistical significance of any given axis have been developed [1]. Beyond being non-parametric, PCA has many attractive properties including computational speed, the ability to identify structure caused by diverse processes and its ability to group or separate samples in a striking visual manner; for example, see [2]. PCA has also become widespread in the analysis of disease-association studies where the inclusion of the locations of samples on a limited number of axes as covariates can be used in an attempt to control for population stratification [3].

Although PCA is explicitly a non-parametric data summary, it is nevertheless attractive to attempt to use the projections to make inferences about underlying events and processes. For example, dispersion of sample projections along a line is thought to be diagnostic of the samples being admixed between the two populations at the ends of the line, though these need not always be present [1], while correlations between principal components and geographical axes have been interpreted as evidence for waves of migration [4,5]. However, while simulation studies have shown that such patterns do occur when the inferred process has acted [1,6], they can also be caused by other processes or even statistical artefacts. For example, clines in principal components result not

Author Summary

Genetic variation in natural populations typically demonstrates structure arising from diverse processes including geographical isolation, founder events, migration, and admixture. One technique commonly used to uncover such structure is principal components analysis, which identifies the primary axes of variation in data and projects the samples onto these axes in a graphically appealing and intuitive manner. However, as the method is non-parametric, it can be hard to relate PCA to underlying process. Here, I show that the underlying genealogical history of the samples can be related directly to the PC projection. The result is useful because it is straightforward to predict the effects of different demographic processes on the sample genealogy. However, the result also reveals the limitations of PCA, in that multiple processes can give the same projections, it is strongly influenced by uneven sampling, and it discards important information in the spatial structure of genetic variation along chromosomes.

just from waves of expansion, but also recurrent bottlenecks, admixture and equilibrium models of spatial structure [6–11].

In this paper I develop a framework for understanding how PCA relates to underlying processes and events. I show that the expected location of samples on the principal components can, for single nucleotide polymorphism (SNP) data, be predicted directly from the pairwise coalescence times between samples. Because it is often relatively easy to obtain analytical or numerical solutions to expected coalescence times under explicit population genetics models, it is also possible to obtain expressions for the PCA projections of samples under diverse scenarios, including island models, models with isolation and founder events and historical admixture. The result also highlights some key limitations of PCA. For example, it follows that PCA cannot be used to distinguish between models that lead to the same mean coalescence times (for example models with migration or isolation). Furthermore, PCA projections are strongly influenced by uneven sampling. Using examples from human genetics I discuss the implications of these results for making inferences from PCA of genetic variation data.

Results

PCA describes structure in the matrix of pairwise coalescence times

In this section I provide a brief summary of how PCA is carried out and describe the key result concerning the relationship between PCA and average coalescence time. In what follows I assume that n haploid individuals have been sequenced with complete accuracy (diploid samples and the influence of SNP ascertainment will be discussed later). The only polymorphisms present are biallelic SNPs that are the result of a single historical mutation. Let $Z_{si} \in \{0,1\}$ be the allelic state for individual i at locus s (here I assume that the ancestral allele is defined as 0 and the derived allele as 1, however the following also applies for any coding, for example where the minor allele is coded as 1). After removing monomorphic sites the data, $\mathbf{Z}$, consist of an $L \times n$ binary matrix (L is the number of SNPs). In PCA, the first step is to zero-centre the data, so as to create a new matrix, $\mathbf{X}$, where

$$X_{si} = Z_{si} - \frac{1}{n}\sum_{j=1}^{n} Z_{sj}\,. \tag{1}$$

At this stage, the data rows are often normalized so as to have equal variance, however, it is assumed that this is not the case (in practice normalization has little effect for SNP data, though will tend to up-weight the influence of rare variants). Each individual sample can be thought of as representing a point in L-dimensional space, where each dimension (or axis) represents a single SNP. The goal of PCA is to find a new set of orthogonal axes (the principal components), each of which is made up from a linear combination of the original axes, such that the projection of the original data onto these new axes leads to an efficient summary of the structure of the data. More formally, PCA defines a stretch and rotation transformation, expressed through the matrix $\mathbf{P}$, such that application of $\mathbf{P}$ to the original data ($\mathbf{Y} = \mathbf{PX}$) leads to transformed data with the following properties.

1. The transformed data matrix, $\mathbf{Y}$, has the same dimensions ($L \times n$) as the original data and the mean of each row is zero.
2. The value associated with a given individual in $\mathbf{y_i}$, the ith *row* of $\mathbf{Y}$, represents the individual's position or projection on the ith principal component.
3. The correlation between any two rows of $\mathbf{Y}$ is zero.
4. The sum of the variances of the rows equals the variance in the original data.
5. The variances of the rows are monotonically decreasing.
6. The variance of the first row is the largest of any possible projection of the original data on a linear combination of the SNPs.

The principal components can be obtained directly by finding the eigenvectors of the covariance matrix

$$\mathbf{C} = \frac{1}{n-1}\mathbf{X}\mathbf{X}^T\,, \tag{2}$$

such that the ith principal component (the ith *row* of $\mathbf{P}$, $\mathbf{p_i}$) is the ith eigenvector of $\mathbf{C}$. However, because $\mathbf{C}$ (of dimension $L \times L$) can be very large for genome-wide SNP data sets, it can be more convenient to use singular value decomposition (SVD) to find the principal components and individual projections. SVD, which exists for any $L \times n$ real matrix (where $L \geq n$) rewrites the original data in terms of three other matrices

$$\mathbf{X} = \mathbf{U}\sum\mathbf{V}^T\,, \tag{3}$$

where $\mathbf{U}$ is an orthogonal matrix (i.e. the dot-product between any two columns is zero) of dimension $L \times n$, $\sum$ is a diagonal matrix of dimension $n \times n$ and $\mathbf{V}$ is another orthogonal matrix of dimension $n \times n$. This is achieved by setting $\mathbf{v_i}$, the ith *column* of $\mathbf{V}$, to be the ith eigenvector of the matrix

$$\mathbf{M} = \mathbf{X}^T\mathbf{X}\,, \tag{4}$$

σ_i, the ith diagonal entry of $\sum$ to be the square root of the corresponding eigenvalue and $\mathbf{u_i}$, the ith *column* of $\mathbf{U}$, to be the vector

$$\mathbf{u_i} = \frac{1}{\sigma_i}\mathbf{X}\mathbf{v_i}\,. \tag{5}$$

PCA and SVD are, through construction, intimately related. Specifically, the projection of samples along the ith principal component is given by $\mathbf{y_i} = \sigma_i\mathbf{v_i}$ (note this is the ith *row* of $\mathbf{Y}$ and the

*i*th *column* of $\mathbf{V}$) and the *i*th principal component is $\mathbf{p_i}=\mathbf{u_i}$. For typical population genetics data sets, eigenvalue analysis of the matrix $\mathbf{M}$ (of dimension $n\times n$) is computationally simpler than analysis of the matrix $\mathbf{C}$ (typically hundreds or thousands of samples have been genotyped at hundreds of thousands or millions of SNPs). The above construction results in the projection of samples on the PCs being influenced by the number of SNPs (e.g. repeating the analysis on a data set in which every SNP is included twice will lead to projections that are a factor $\sqrt{2}$ larger than previously). To correct for this, consider a slightly different definition of the matrix $\mathbf{M}$:

$$\mathbf{M}=\frac{1}{L}\mathbf{X}^T\mathbf{X}, \tag{6}$$

which is equivalent to dividing the data matrix by the square-root of the number of SNPs. It is worth noting that L may either be a random variable as in the case of sequencing, or a fixed variable, as in the case of genotyping. Here, it will be treated as a fixed variable, though in practice this is of little importance.

$\mathbf{M}$ is a stochastic matrix. However, it is possible to learn about the key structural features of $\mathbf{M}$ by considering its expectation. From above, it follows that

$$M_{ij}=\frac{1}{L}\sum_{s=1}^{L}X_{si}X_{sj} \tag{7a}$$

$$E\left(M_{ij}\right)=\frac{1}{L}\sum_{s=1}^{L}E\left(X_{si}X_{sj}\right) \tag{7b}$$

$$=\frac{1}{L}\sum_{s=1}^{L}E\left(\left(Z_{si}-\frac{1}{n}\sum_{k=1}^{n}Z_{sk}\right)\left(Z_{sj}-\frac{1}{n}\sum_{k=1}^{n}Z_{sk}\right)\right). \tag{7c}$$

Assuming that sites are identical in distribution (though not necessarily independent) the subscript s can be dropped to give

$$E\left(M_{ij}\right)=E(Z_iZ_j)-E_k(Z_iZ_k)-E_k(Z_jZ_k)+E_{kl}(Z_kZ_l), \tag{8}$$

where the terms such as $E_k(Z_jZ_k)$ indicate the expectation (for sample j) is averaged over all individuals k in the sample (note this includes self); i.e. $E_k(Z_jZ_k)=1/n\sum_{k=1}^{n}E(Z_iZ_k)$. Because Z_i is either 0 or 1, the four terms in Equation 8 can be thought of as:

1. The probability that samples i and j both carry a derived mutation at a randomly chosen locus conditional on the locus being polymorphic in the sample.
2. The probability that sample i and another randomly chosen sample k (which may include either i or j) both carry the derived mutation at a randomly chosen locus.
3. The probability that sample j and another randomly chosen sample k (which may include either i or j) both carry the derived mutation at a randomly chosen locus.
4. The probability that two samples, k and l, chosen at random with replacement both carry the derived mutation at a randomly chosen locus.

In the case of a low mutation rate, where polymorphic sites are the result of a single historical mutation, expressions can be obtained for the above quantities in terms of features of the genealogical tree [12–14]. Figure 1 shows how the probability of two samples both carrying a mutation depends on their time to a common ancestor relative to the time to the common ancestor of the whole sample. Let $E(t_{ij})=\bar{t}_{ij}$ be the expected coalescence time for samples i and j, $E(T_{MRCA})$ be the expected time to the most recent common ancestor of the sample, and $E(T)=\bar{T}$ be the expected total branch length in the tree. The probability that two samples share a derived mutation (conditional on the site being segregating) is given by

$$E(Z_iZ_j)=\frac{P(\text{Mutation occurs on branch ancestral to } i \text{ and } j)}{P(\text{Mutation occurs in tree})} \tag{9a}$$

$$=\lim_{\theta\to 0}\frac{E\left(\frac{\theta}{2}(T_{MRCA}-t_{ij})\exp\left(-\frac{\theta}{2}(T_{MRCA}-t_{ij})\right)\right)}{E\left(\frac{\theta}{2}T\exp(-\frac{\theta}{2}T)\right)} \tag{9b}$$

$$=\frac{E(T_{MRCA})-E\left(t_{ij}\right)}{E(T)}. \tag{9c}$$

By writing similar expressions for the other terms in Equation 8 it follows that

$$E(M_{ij})=\frac{1}{\bar{T}}\left(\bar{t}_i+\bar{t}_j-\bar{t}-\bar{t}_{ij}\right), \tag{10}$$

where $\bar{t}_i=\frac{1}{n}\sum_{k=1}^{n}\bar{t}_{ik}$ and $\bar{t}=\frac{1}{n}\sum_{k=1}^{n}\bar{t}_k$. Note that these expressions include coalescence with self where the coalescence time is always zero; i.e. $\bar{t}_{ii}=0$. In short, the expectation of the matrix whose eigenvectors give the projections of samples on the principal components can be written in terms of the mean coalescence times for pairs of samples. It is worth noting that $\bar{t}_{ij}$ (and the related quantities) can be interpreted either as the

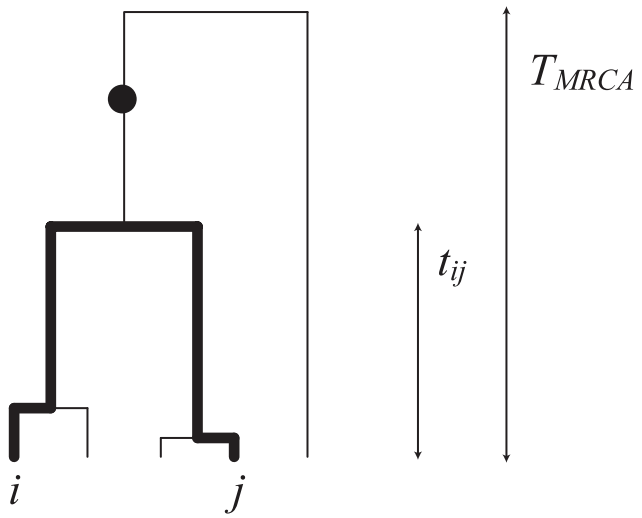

Figure 1. Genealogical statistics. The chart shows a genealogical tree describing the history of a sample of size five. Two samples, i and j, will share a derived mutation (indicated by the circle) if it occurs on the branch between their most recent common ancestor and the common ancestor of the whole sample. The length of this branch is $T_{MRCA}-t_{ij}$.
doi:10.1371/journal.pgen.1000686.g001

expected coalescence time under some model or else the average realized coalescent time across the genome. The difference between these quantities can be important in some settings, such as admixture models (see below).

For diploid individuals the genotypic value for an individual at a given SNP is typically given by the sum of the allelic values; i.e. $G_{si}=Z^1_{si}+Z^2_{si}\in\{0,1,2\}$, where the superscripts indicate the two alleles. By following the same argument as above it can be shown that for genotype data

$$E(M_{ij})=\frac{1}{\bar{T}}\left(2\bar{t}^1_i+2\bar{t}^2_i+2\bar{t}^1_j+2\bar{t}^2_j-4\bar{t}-\bar{t}^{11}_{ij}-\bar{t}^{12}_{ij}-\bar{t}^{21}_{ij}-\bar{t}^{22}_{ij}\right), \quad (11)$$

where the superscripts again indicate the relevant allele in each individual. In the following I will assume that data consist of haplotypes, however Equation 11 makes it clear that essentially identical results will hold for genotype data.

An example using two geographically separated populations

The implication of Equation 10 is that if the structure of pairwise coalescence times in a given data set can be understood, then the projection of the samples on the principal components can be predicted directly. Two illustrate this idea consider the simple model of a population split shown in Figure 2A. Under this model the expected coalescence time for pairs of samples within either population is 1 (in units of $2N_e$ generations) and the expected coalescence time for pairs of samples from different populations is $1+\Delta$, where Δ is the age of the population split (also in units of $2N_e$

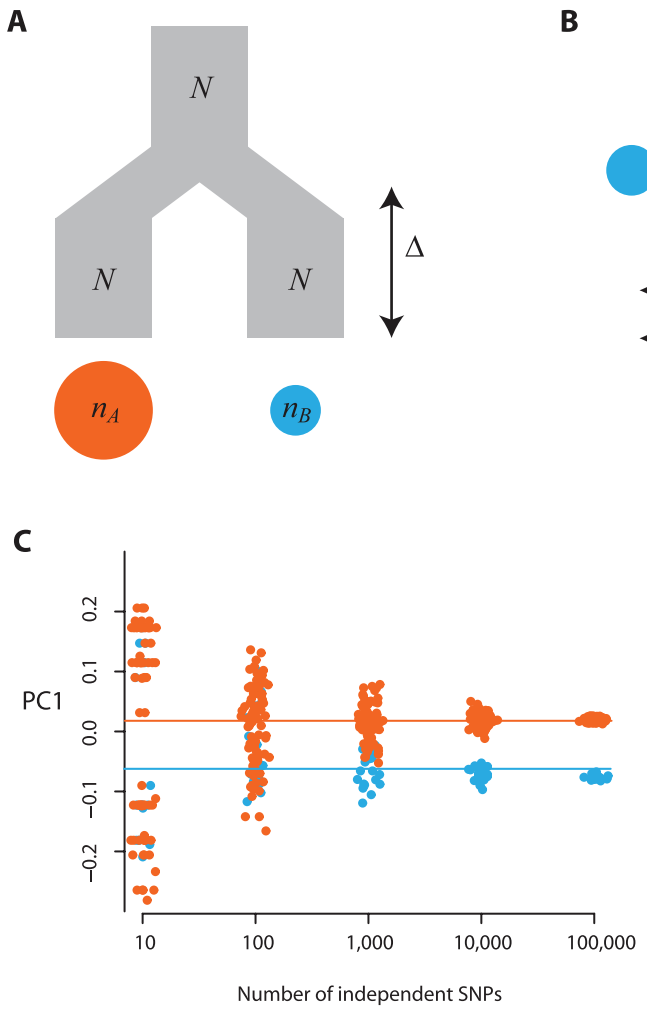

Figure 2. Principal component analysis of two populations. (A) Consider a sample of n_A individuals from population A (indicated by the red circle) and n_B from population B (indicated by the blue circle), where the two populations have the same effective population size of N and are both derived from a single ancestral population, also of size N, with the split happening a time Δ in the past. (B) The expected locations of these two sets of samples on the first PC is defined by the time since divergence (the Euclidean distance between the samples is $\sqrt{2\Delta/\bar{T}}$) (see text for definitions) and the relative sample size from the populations, with the larger sample lying closer to the origin. Defining $\phi=n_A/(n_A+n_B)$, the relative location of the two populations on the first PC are $1-\phi$ for samples from population A and $-\phi$ for samples from population B (note that the sign is arbitrary). (C) To investigate the effect of finite genome size simulations were carried out for the model shown in part A with 80 genomes sampled from population A, 20 from population B and a split time of 0.02 N_e generations ($F_{ST}=0.01$) and between 10 and 10^5 SNPs. Lines indicate the analytical expectation. A jitter has been added to the x-axis for clarity. Note that the separation of samples with 10 SNPs does not correlate with population and simply reflects random clustering arising from the small numbers of SNPs.
doi:10.1371/journal.pgen.1000686.g002

generations). Suppose of the total sample size n, a fraction ϕ are from population A. Define $\alpha=2\Delta(1-\phi)^2$ and $\beta=2\Delta\phi^2$, it follows that for large n, $\mathbf{M}$ has a simple block structure;

$$E(\mathbf{M})\approx\frac{1}{\bar{T}}\begin{pmatrix}1+\alpha & \alpha & \alpha & -\sqrt{\alpha\beta} & -\sqrt{\alpha\beta}\\ \alpha & 1+\alpha & \alpha & -\sqrt{\alpha\beta} & -\sqrt{\alpha\beta}\\ \alpha & \alpha & 1+\alpha & -\sqrt{\alpha\beta} & -\sqrt{\alpha\beta}\\ -\sqrt{\alpha\beta} & -\sqrt{\alpha\beta} & -\sqrt{\alpha\beta} & 1+\beta & \beta\\ -\sqrt{\alpha\beta} & -\sqrt{\alpha\beta} & -\sqrt{\alpha\beta} & \beta & 1+\beta\end{pmatrix}, \quad (12)$$

where the first ϕn rows and columns represent the samples from population A (here, for example, three samples from A and two from B are shown). What will the leading eigenvalue and associated eigenvector be for a matrix with this kind of block structure? Although it is simple to obtain eigenvectors numerically, it is also worth having some intuition about what they represent. Through the construction of SVD it follows that the leading eigenvector, λ_1 and eigenvector, $\mathbf{v_1}$, are those that, through Equation 3, provide the *best* approximation to the original data in terms of least-squares error. Equivalently, the matrix $\lambda_1\mathbf{v_1}\mathbf{v_1}^T$ is the best least-squares approximation to $\mathbf{M}$. Intuitively, the original data is well approximated by the average allele frequency in each population and the the block structure of $\mathbf{M}$ can be recovered by clustering samples from the two populations either side of the origin in $\mathbf{v_1}$. More formally, it can be shown that

$$\mathbf{v_1}\approx\left(\sqrt{\frac{1-\phi}{\phi}},\sqrt{\frac{1-\phi}{\phi}},\cdots,-\sqrt{\frac{\phi}{1-\phi}},-\sqrt{\frac{\phi}{1-\phi}}\right)/\sqrt{n} \quad (13a)$$

$$\lambda_1\approx(1+2\Delta n\phi(1-\phi))/\bar{T}. \quad (13b)$$

Assuming that $2\Delta n\phi(1-\phi)>>1$, the projection of the samples on the first principal component is given by the vector

$$\mathbf{y_1}\approx\sqrt{\frac{2\Delta}{\bar{T}}}(1-\phi,1-\phi,\cdots,-\phi,-\phi). \quad (14)$$

Note that the sign of the projections is arbitrary. This result implies that the Euclidean distance between samples from the two populations on the first principal component will be $\sqrt{2\Delta/\bar{T}}$ and their position relative to the origin is determined by the relative sample size, with the larger sample lying closer to the origin. Figure 2B shows the expected projection of samples.

These results refer explicitly to the expected value of $\mathbf{M}$. However, it is also important to know whether stochasticity resulting from the finite size of the genome has a significant effect on the results. Theoretical work on the nature and size of the first principal component in random matrices [15,16] has identified a critical signal to noise ratio below which the true structure of the signal cannot be recovered. In the context of a two-population model this equates to F_{ST} being greater than $1/\sqrt{nL}$ [1]. For example, with a sample size of 100 and $F_{ST}=0.01$, the threshold is 100 SNPs. Simulations were carried out for different numbers of independent SNPs (Figure 2C). As expected, for 10 or 100 SNPs PCA fails to separate samples from the two populations, while for 1,000 SNPs or more samples from the two populations are distinct on the first PC and centre around the theoretical expectation.

PCA cannot distinguish between alternative models that have the same effect on mean coalescence time

A direct consequence of Equation 10 is that PCA predominantly reflects structure in the expected (or mean realized coalescent) time. Consequently, any two demographic models that give the same structure of expected coalescence times will also give the same projections. To illustrate this result, consider a fully general model with two homogeneous populations where the expected coalescence time for two samples from population A is t_{AA}, the expected coalescence time for two samples from population B is t_{BB} and the expected coalescence time for one sample from each population is t_{AB}. Define $c=2t_{AB}-t_{AA}-t_{BB}$, $\alpha=c(1-\phi)^2$ and $\beta=\phi^2$. It can be shown that

$$E(\mathbf{M})\approx\frac{1}{\bar{T}}\begin{pmatrix}t_{AA}+\alpha & \alpha & \alpha & -\sqrt{\alpha\beta} & -\sqrt{\alpha\beta}\\ \alpha & t_{AA}+\alpha & \alpha & -\sqrt{\alpha\beta} & -\sqrt{\alpha\beta}\\ \alpha & \alpha & t_{AA}+\alpha & -\sqrt{\alpha\beta} & -\sqrt{\alpha\beta}\\ -\sqrt{\alpha\beta} & -\sqrt{\alpha\beta} & -\sqrt{\alpha\beta} & t_{BB}+\beta & \beta\\ -\sqrt{\alpha\beta} & -\sqrt{\alpha\beta} & -\sqrt{\alpha\beta} & \beta & t_{BB}+\beta\end{pmatrix}. \quad (15)$$

Again, only three samples from population A and two from population B are shown. For large n, the leading eigenvalue and corresponding eigenvector of the above matrix are respectively

$$\lambda_1\approx cn\phi(1-\phi)/\bar{T} \quad (16a)$$

$$\mathbf{v_1}\approx\left(\sqrt{\frac{1-\phi}{\phi}},\sqrt{\frac{1-\phi}{\phi}},\cdots,-\sqrt{\frac{\phi}{1-\phi}},-\sqrt{\frac{\phi}{1-\phi}}\right)/\sqrt{n}. \quad (16b)$$

Consequently, the projection of the samples on the first principal component is given by the vector

$$\mathbf{y_1}\approx\sqrt{\frac{c}{\bar{T}}}(1-\phi,1-\phi,\cdots,-\phi,-\phi). \quad (17)$$

Comparison of Equations 14 and 17 shows that the Euclidean distance between samples from the two populations on the first PC is a function of the difference between cross-population and within-population coalescence times and that the positioning of the populations relative to the origin simply reflects their relative sample size (as for the simpler two-population model). Consequently, any two models that give the same value of $c/\bar{T}$ will give the same expected projections of samples on the first PC.

One connection that is worth exploring further is the link between the results shown here and those of Slatkin [12] concerning F_{ST}. Slatkin showed that

$$F_{ST}=1-\frac{\bar{t}_w}{\bar{t}}, \quad (18)$$

where $\bar{t}_w$ is the average coalescence time for pairs of samples from the same population and $\bar{t}$ is the average coalescence time across all pairs of samples. In the notation used above it can be shown that

$$F_{ST}=\frac{c\phi(1-\phi)}{\bar{t}}. \quad (19)$$

Now consider the PCA projection. The variance along the first

axis is $c\phi(1-\phi)/\bar{T}$. The total variance in the sample is $\bar{t}/\bar{T}$. Consequently, the fraction of the total variance explained by the first PC is equal to $c\phi(1-\phi)/\bar{t}=F_{ST}$. Given that F_{ST} is defined as the fraction of the total variance that is explained by between-population differences this result is not surprising. Nevertheless, the result demonstrates a simple relationship between the Euclidean distance of populations in PCA space and F_{ST}, at least in the case of two populations.

Uneven sampling has a strong influence on PCA projections

As has been shown previously [11], PCA projections can be strongly influenced by uneven sampling from a series of populations. The results described here provide an explanation. First, from Equation 10 it can be seen that the the matrix **M** is influenced by the relative sample size from each population through the components $\bar{t}_i$. For instance, even if all populations are equally divergent from each other, those for which there are fewer samples will have larger values of $\bar{t}_i$ because relatively more pairwise comparisons are between populations. Second, even if the entries of **M** were not influenced by the relative sample size, its eigenvectors will be, simply because relative sample size will influence the structure of the genetic variance in the sample (see Figure 2). The influence of uneven sample size can be to bias the projection of samples on the first few PCs in unexpected ways, for example, where there is spatial structure to genetic variation. Consider a lattice arrangement of populations with equal migration between neighbouring populations. For this arrangement it is possible to obtain analytical expressions for the expected coalescence time for pairs of samples from the different populations (results not shown) and hence the matrix **M** (up to an unknown scaling factor) and subsequently the projection of samples on the first few PCs under different assumptions about sample size and migration rate. If sample sizes from the different populations are equal, the spatial arrangement of the populations on the first two PCs mimics the structure of the migration matrix (Figure 3A). However, sample sizes differ between populations the effect is to distort the projection space (Figure 3B and 3C). This distortion of PC-space relative to the structure of the migration matrix is problematic for interpreting the location of samples on PCs. Sub-sampling from populations to achieve more equal representation, as in [2], is the only way to avoid this problem.

The projection of admixed individuals onto existing axes directly identifies admixture proportions

The principal components identified through PCA can be used to project not just those samples from which the PCs were obtained, but also additional samples. The appeal of such analyses is that it enables the analysis of structural features identified in one data set to be transferred to another. For example, where data from two source populations and a set of possibly admixed samples are available, projection of the admixed samples onto the axes defined by the source populations can identify the extent of mixed ancestry. The advantage of this approach rather than simply performing PCA on all samples together is that other structural features within the admixed samples (e.g. admixture from a third population or relatedness) will have little influence on the projection. In the light of the above results showing how the PCA projection of samples can be interpreted in terms of coalescence times, it is interesting to ask how the the projection of additional samples onto the same axes also relates to coalescence times.

Consider the case of the general two-population model where the positions of the samples on the first PC are $\sqrt{c/\bar{T}(1-\phi)}$ for samples from population A and $-\sqrt{c/\bar{T}\phi}$ for samples from population B. The first PC can be obtained as in Equation 5. For a given SNP, s, the expected loading for the first PC, u_{s1}, is therefore

$$u_{s1}=\frac{1}{\sqrt{L\lambda_1}}\sum_{i=1}^{n}X_{si}y_{1i} \tag{20a}$$

$$=\frac{1}{\sqrt{L\lambda_1}}\sum_{i=1}^{n}Z_{si}y_{1i} \tag{20b}$$

$$E(u_{s1})\approx\frac{\bar{T}}{cn\phi(1-\phi)}\sqrt{\frac{c}{L\bar{T}}}\left(n_A^1(1-\phi)-n_B^1\phi\right), \tag{20c}$$

where n_A^1 is the number of samples carrying the derived allele in population A. By writing $n_A^1=n\phi\pi_A$ and $n_B^1=n(1-\phi)\pi_B$, such that π_A and π_B are the frequencies of the derived alleles in populations A and B respectively, it follows that

$$E(u_{s1})\approx\sqrt{\frac{\bar{T}}{cL}}(\pi_A-\pi_B). \tag{21}$$

The expected location of an additional sample, j, on the first PC is therefore

$$y_{1j}=\sum_{s=1}^{L}X_{sj}u_{s1} \tag{22a}$$

$$E\left(y_{1j}\right)=\frac{1}{L}\sqrt{\frac{\bar{T}}{c}}\sum_{s=1}^{L}E\left(X_{sj}(\pi_A-\pi_B)\right) \tag{22b}$$

$$=\sqrt{\frac{\bar{T}}{c}}E\left((Z_j-\bar{Z})(\pi_A-\pi_B)\right), \tag{22c}$$

where $\bar{Z}=1/n\sum_{i=1}^{n}Z_i$ (note this does not include the additional sample j). Again, the subscript s has been dropped by assuming that sites are identical in distribution. By noting that $E(Z_j\pi_A)=E_k^A(Z_jZ_k)$, where the expectation is over those samples from population A, it follows that similar arguments to those above can be made to relate the quantities in Equation 22 to coalescent times. Define $\bar{t}_{jA}$ as the average coalescent time between the additional sample and all samples from population A and $\bar{t}_{jB}$ to be the equivalent for population B, it can be shown that

$$E\left(y_{1j}\right)=\sqrt{\frac{1}{c\bar{T}}}\left[\bar{t}_{jB}-\bar{t}_{jA}+\frac{1}{2}(t_{AA}-t_{BB}+(1-2\phi)c)\right]. \tag{23}$$

An important implication of Equation 23 is that if the additional sample is the result of an admixture event between the two populations with a fraction θ_j of its genome coming from population A then it follows that the location of the sample on the first PC is

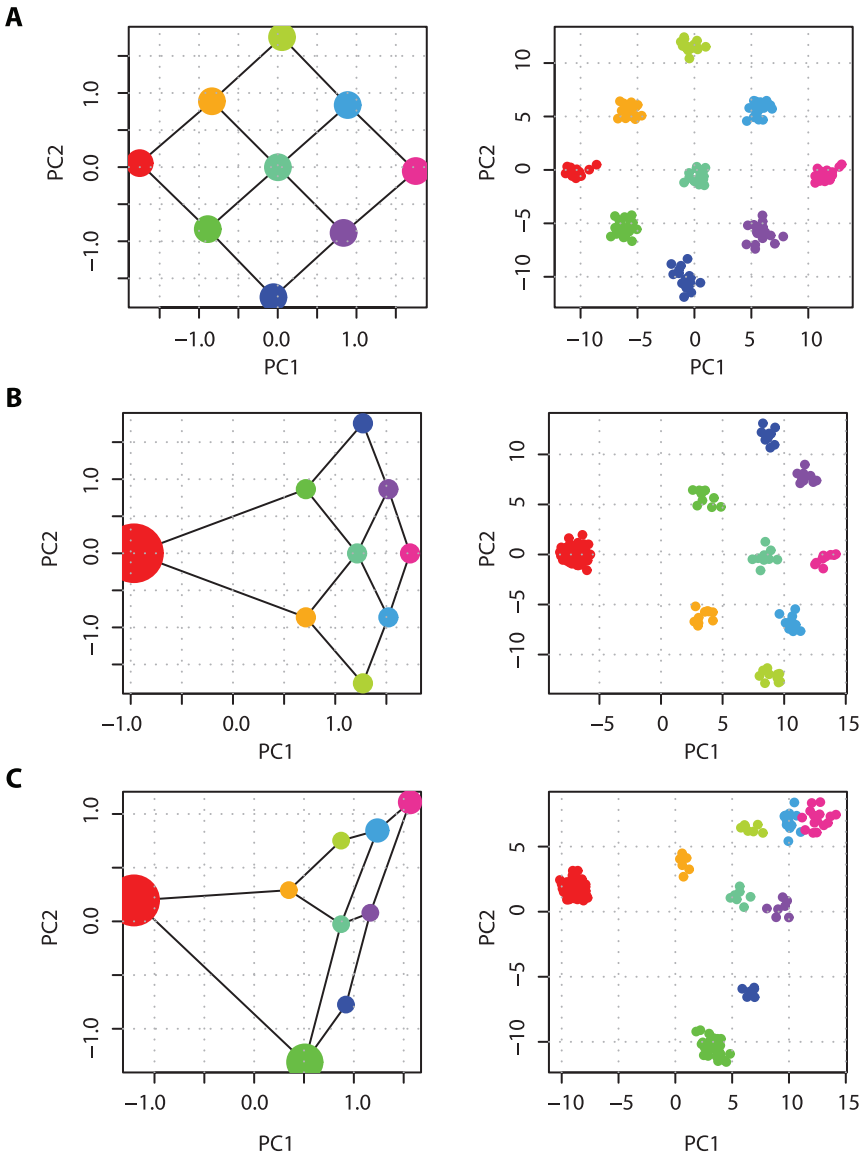

Figure 3. The effect of uneven sampling on PCA projection. PCA projection of samples taken from a set of nine populations arranged in a lattice, each of which exchanges migrants at rate M per N_e generations with each adjoining neighbour, leads to a recovery of the migration-space if samples are of equal size (A), or a distortion of migration-space if populations are not equally represented (B,C). In each part the left-hand panel shows the analytical solution (the area of each point represents the relative sample size) with migration routes illustrated while the right-hand panel shows the result of a simulation with a total sample size of 180 and 10,000 independent SNP loci. All examples are for $M = 2$.
doi:10.1371/journal.pgen.1000686.g003

$$E(y_{1j}) = \sqrt{\frac{c}{T}}(\theta_j - \phi). \quad (24)$$

In words, the admixture proportion of the individual can be directly inferred from their relative position along the first PC from the two source populations.

There are three important points to note when applying this result. First, only if the admixture event was very recent are the source populations likely to be available. Rather samples may be available for descendants of these source populations. Consequently, the average divergence between the population A part of an individuals genome and other samples from population A might typically be greater than for two samples taken directly from population A. However, this effect is likely to be very similar for the two source populations and, given Equation 23, these effects largely cancel out.

The second point to note is that if samples are admixed between more than two populations, the result generalises so that an individual whose genome is derived from several source populations will have a projected position (along each significant PC) defined by the weighted sum of the positions of its source populations. Informally, the result arises because of the linearity in Equation 22. Those parts of the genome with ancestry from a given population will have a PC projection that matches samples taken directly from the source population. If there is mixed ancestry, the effect is simply to average the PC projections.

Finally, it is important to note that projection of non-admixed individuals can also lead to their location being intermediate between the two original populations. For example, samples from

a third population that either diverged from population A since the split with population B or that come from a population that diverged before the A/B split will (in both cases) be projected between the locations of samples from populations A and B. It may, however, be possible to distinguish between such cases by carrying out PCA on all data combined.

PCA carried out on admixed individuals can identify relative admixture proportion in the absence of source populations

As has already been shown through simulation [1], PCA carried out on samples that are the result of admixture events can identify admixed samples as lying along the axes between the two or more source populations, even if one or more of the source populations are absent. The results above shed some light onto when such analyses are expected to work and when they will fail.

Consider a sample of individuals who are the result of an historical admixture event between two populations A and B. In order to define the matrix **M** for this sample it is necessary to know which part of their genome is derived from each of the source populations. Let $\mathbf{a_i}$ be a series of indicator functions for each of the L SNPs in individual i that takes value 1 if that part of the individual's genome was derived from population A and 0 if it was derived from population B. The value $\bar{t}_{ij}$ can be obtained by comparing the value of $\mathbf{a_i}$ and $\mathbf{a_j}$ at each position and adding up the relevant contribution from each of t_{AA}, t_{BB} and t_{AB}. Note that here the achieved ancestry proportions are being used rather than their expectation under some model (which might be the same for all samples).

Given these considerations there are two situations under which none of the structure between the two source populations is expected to be reflected in the matrix **M**. First, all individuals could have the same vector **a**, which could occur if the admixture event were ancient and involved relatively few individuals such that the source population at every point in the genome were fixed (note this does not mean that there is no variation, simply that all individuals at this location have an ancestry from the same population). Second, individuals have different ancestry vectors, but the average value is the same for all individuals and the admixture chunks have been sufficiently broken up through historical recombination such that everyone is equally related to everyone else. Again, this scenario could occur if the admixture result were ancient. Note that all individuals having the same average ancestry proportions is, by itself, not sufficient to create this problem. To examine the rate at which admixture signal is lost, an admixed population was simulated forward in time and the projections of samples on the first PC were followed, along with the correlation between PC projection and individual ancestry. As shown in Figure 4, in which the population is chosen to have parameters comparable to humans, the initially strong correlation between ancestry proportion and location on the first PC is rapidly lost such that after only 15 generations there is essentially no signal remaining, even though locally within the genome admixture chunks are still very clear (i.e. there is still admixture LD) after 50 generations.

Discussion

The primary result of this paper is that the locations of samples on the principal components identified from genome-wide data on genetic variation can be predicted from an understanding of the average coalescent time for pairs of samples. This gives a direct route to understanding the influence various demographic scenarios can have on the relationships between samples identified from PCA and how PCA can be used to make inference about processes of interest such as admixture. However, the results also demonstrate the way in which sampling schemes can influence PC projections and how similar projections can arise from very different demographic scenarios. Consequently, using these results to motivate inference from PCA about underlying demographic process may prove difficult.

There are, however, situations in which PCA can be used to infer demographic parameters directly. For example, in cases of

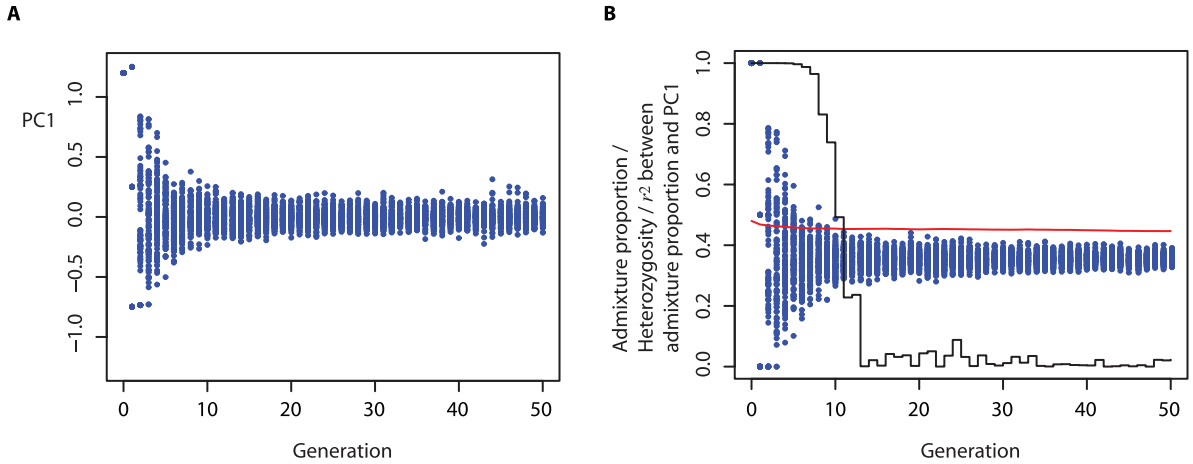

Figure 4. Identification of admixture proportions without source populations. Initially an admixed population is formed by random mating from two populations, each fixed for a different allele at each locus with 40% contribution from one population. In the simulated population there are 1000 individuals, each of which has 20 chromosomes with 50 markers each, a genetic map length of 1 per chromosome and a uniform recombination rate. Subsequent generations are formed by random mating of the ancestral population. (A) Projections of 100 randomly chosen samples on the first PC over time show a decay in the fraction of variance explained by the first PC (note that the total variance in the population decays little over the time-scale of the simulation). (B) Admxiture proportions for the same individuals as in part A (blue points) as well as the everage heterozygosity (red line) and the fraction of the variance in PC1 explained by admixture proportions (black line). While there is a strong association between admixture proportion and location on PC1 for the first few generations, after 15 generations recombination has eliminated any signal, even though there is still strong admixture LD between nearby markers (data not shown).
doi:10.1371/journal.pgen.1000686.g004

simple two- or three-way admixture, where populations close to the source populations can be identified and sampled from, estimation of admixture proportions can be achieved from projecting samples onto the PCs identified from the source populations. To illustrate this, Figure 5 shows the inferred ancestry proportions for a set of haplotypes (estimated from trio data) in 20 African Americans collected as part of the HapMap3 project. In this analysis, haplotypes (also inferred from trios) from the European ancestry population in Utah (CEU) and the Yoruba in Nigeria (YRI) are used to represent the source populations (note, as discussed above, the requirement is not that these *are* the source populations, simply that they are closely related to the source populations). By analysing each chromosome separately it can be shown that while each individual's average ancestry proportion across the genome is fairly constant (typically 70–90% African), there is considerable variation at the level of individual chromosomes, with some chromosomes appearing essentially European (for some individuals) and others essentially African (no chromosome shows an overall tendency to come from one population). Such information could be informative about processes such as the level of assortative mating and the rate of ongoing admixture.

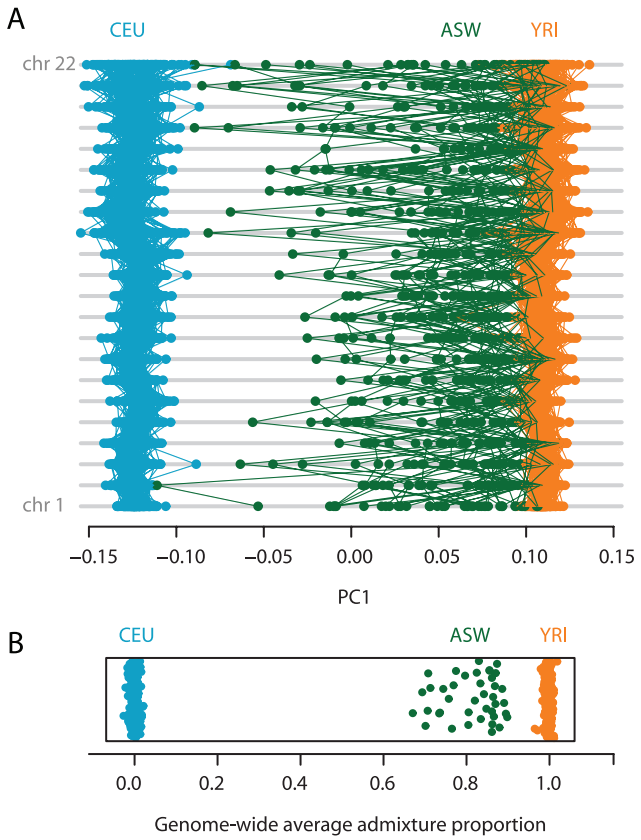

Figure 5. Admixture proportions inferred from PCA projections. (A) For each of the autosomes (chromosome 1 is the lowest) the points indicate the locations of sampled haplotypes (the transmitted and untransmitted haplotypes inferred from trios) on the first principal component (each chromosome is analysed separately; blue = CEU, orange = YRI, green = ASW). Importantly, PCA is carried out only on the haplotypes from CEU and YRI and all samples are subsequently projected onto the first PC identified from this analysis. Lines connect the transmitted (or untransmitted) haplotypes for each individual across chromosomes. Note the uniformity of the locations of samples on the first PC for CEU and YRI. Individual chromosomes within the ASW, however, show a great range of locations on the first PC. (B) The genome-wide admixture proportions (separately for transmitted and untransmitted chromosomes) can be inferred directly from the location of admixed samples on the first PC between the two source populations. Colours are as for (A). The vertical spacing of points is arbitrary.
doi:10.1371/journal.pgen.1000686.g005

One important issue in the application of these ideas to the analysis of empirical data is the extent to which SNP ascertainment will influence outcome. SNP discovery in a small panel will typically lead to the under-representation of rare SNPs in the genotyped data and, depending on the geographical distribution of the samples used for discovery, can also lead to biases in the representation of variation from different areas. The quantities in Equation 8 are therefore conditional not just on segregation in the genotyped sample, but also on segregation within the SNP discovery panel. Consider the joint genealogy of the genotyped and discovery samples shown in Figure 6A. The probability that a pair of samples, i and j share a derived mutation (in the genotyped samples) that also lies on the subtree of the discovery samples, $E(Z_iZ_j^*)$ is

$$E\left(Z_iZ_j^*\right)=\frac{E\left(T^*_{M\,RCA}\right)-E\left(t^*_{ij}\right)}{E(T^*)} \tag{25}$$

where t^*_{ij} is the first time at which the common ancestor of the samples i and j is also a common ancestor of at least one of the discovery panel samples ($t^*_{ij}\geq t_{ij}$), T^*_{MRCA} is the time to the more recent of the discovery or sample MRCAs and T^* is the total time of the intersection between the discovery and genotyped samples' genealogies (Figure 6A). It follows that the equivalent expression for Equation 10 with SNP ascertainment will typically be larger than without SNP ascertainment because $T^*\leq T$ whereas the differences in the numerators will largely cancel each other out. Consequently, it is expected that, except for very strongly biased SNP discovery (e.g. a sample of two from one of a series of very divergent populations), that PCA projections from genotype data will be similar to PCA projections from resequencing data, but will typically be larger in magnitude (if the matrix $\mathbf{M}$ is normalized by the number of SNPs) by a factor $\sqrt{\bar{T}/\bar{T}^*}$; a result confirmed by simulation (Figure 6B and 6C). For the example shown, this result holds even under the most extreme ascertainment scheme of two discovery samples from a single population. In short, SNP ascertainment will tend to have a simple and predictable effect on PC projections that has little influence on the relative placing of samples.

Finally, it is worth pointing out that because PCA effectively summarizes structure in the matrix of average pairwise coalescent times, but in a manner that is influenced by sample composition, more direct inferences can potentially be made from the matrix of pairwise differences (which are trivially related to pairwise coalescent times). This is not to say that eigenvalue analysis of the pairwise distance matrix will correct for the effects of biased sampling demonstrated in Figure 3. However, while readily-available alternatives to PCA, such as multidimensional scaling, seem to have properties similar to PCA, it is possible to envisage non-parametric methods for analysing the matrix of pairwise differences that identify structure without being influenced by sample size.

Methods

Coalescent simulations were carried out using scripts written by the author in the R language (www.r-project.org) and available on request.

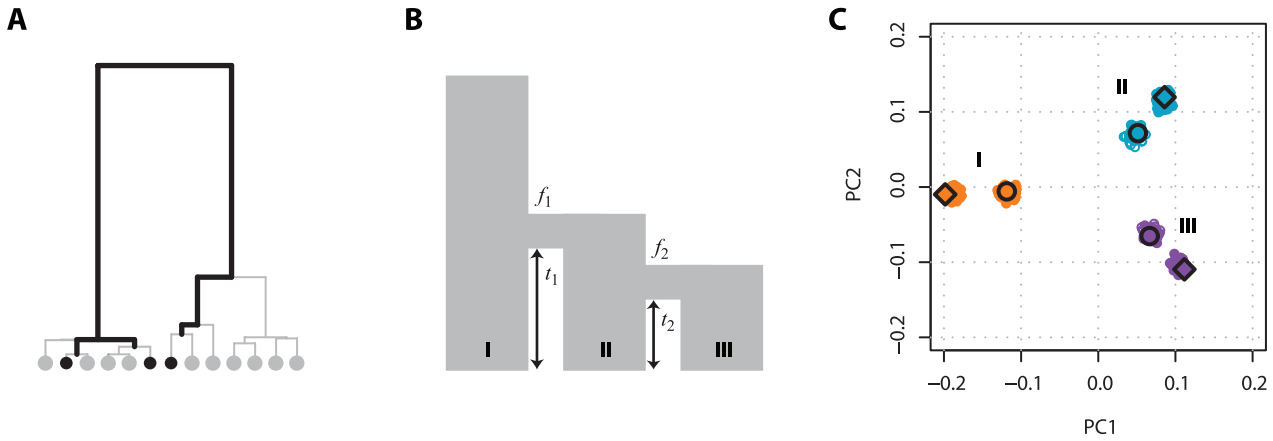

Figure 6. The effect of SNP ascertainment on PCA projection. (A) In the joint genealogy of the ascertainment (black circles) and genotyped samples (grey circles), only mutations occurring on the intersection of the two genealogies (shown in black) will be detected in both samples. For small discovery panels and large experimental samples, this may be considerably less than half the total genealogy length. (B) Model used to simulate data from three populations linked by two vicariance events, each of which is associated with a bottleneck; the model is an approximation to the demographic history of the HapMap populations [17,18]. In the simulations 100 haploid genomes with 10,000 unlinked loci were sampled from each population and the parameters are $t_1 = 0.3$, $t_2 = 0.2$, $f_1 = 0.2$, $f_2 = 0.1$, where f is the bottleneck strength measured as the probability that two lineages entering the bottleneck have coalesced by its end (the bottleneck is instantaneous in real time). All populations have the same effective population size. (C) PCA of the simulated data (small open circles) shows strong agreement with results obtained from analytical consideration of the expected coalescence times (large circles). When only those SNPs that have been discovered in a small panel are considered (here modelled as 4, 8, and 4 additional samples from populations I, II, and III respectively) the principal effect is to scale the locations of the samples on the first two PCs (small filled circles) by a factor of approximately $\sqrt{\bar{T}/\bar{T}^*}$ (large diamonds).
doi:10.1371/journal.pgen.1000686.g006

Principal component analysis of simulated data was carried out using the R function eigen. Phased haplotypes from the International HapMap Project (HapMap3 release 2) were used in the analysis of the CEU, YRI and ASW population (see ftp://ftp.hapmap.org/hapmap/phasing/2009-02_phaseIII/HapMap3_r2/).

Acknowledgments

Many thanks to Niall Cardin, Peter Donnelly, Stephen Leslie, Simon Myers, John Novembre, Nick Patterson, and Molly Przeworski for discussion and comments on the manuscript.

Author Contributions

Conceived and designed the experiments: GM. Performed the experiments: GM. Analyzed the data: GM. Wrote the paper: GM.

References

1. Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PLoS Genet 2: e190. doi:10.1371/journal.pgen.0020190.
2. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, et al. (2008) Genes mirror geography within Europe. Nature 456: 98–101.
3. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38: 904–909.
4. Cavalli-Sforza LL, Menozzi P, Piazza A (1994) The History and Geography of Human Genes. New Jersey: Princeton.
5. Reich D, Price AL, Patterson N (2008) Principal component analysis of genetic data. Nat Genet 40: 491–492.
6. Klopfstein S, Currat M, Excoffier L (2006) The fate of mutations surfing on the wave of a range expansion. Mol Biol Evol 23: 482–490.
7. Barbujani G, Sokal RR, Oden NL (1995) Indo-European origins: a computer-simulation test of five hypotheses. Am J Phys Anthropol 96: 109–132.
8. Fix AG (1997) Gene frequency clines produced by kin-structured founder effects. Hum Biol 69: 663–673.
9. Chikhi L, Nichols RA, Barbujani G, Beaumont MA (2002) Y genetic data support the Neolithic demic diffusion model. Proc Natl Acad Sci USA 99: 11008–11013.
10. Currat M, Excoffier L (2005) The effect of the Neolithic expansion on European molecular diversity. Proc Biol Sci 272: 679–688.
11. Novembre J, Stephens M (2008) Interpreting principal component analyses of spatial population genetic variation. Nat Genet 40: 646–649.
12. Slatkin M (1991) Inbreeding coefficients and coalescence times. Genet Res 58: 167–175.
13. Wilkinson-Herbots HM (1998) Genealogy and subpopulation differentiation under various models of population structure. J Math Biol 37: 535–585.
14. McVean GA (2002) A genealogical interpretation of linkage disequilibrium. Genetics 162: 987–991.
15. Baik J, Ben Arous G, Péché S (2005) Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. Ann Probability 33: 1643–1697.
16. Debashis P (2007) Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. Statistica Sinica 17: 1617–1642.
17. Schaffner S, Foo C, Gabriel S, Reich D, Daly MJ, et al. (2005) Calibrating a coalescent simulation of human genome sequence variation. Genome Res 15: 1576–1583.
18. The International HapMap Consortium (2005) A haplotype map of the human genome. Nature 27: 1299–1320.