Estimating F-statistics in a probabilistic PCA space

Divyaratan Popli, Benjamin M. Peter

October 17, 2023

Abstract

Principal component analysis (PCA) and F-statistics are routinely used in population genetic and archaeogenetic studies. However, these are closely related analyses and reveal the same biological signal. Here, we present a statistical framework to combine them into a joint analysis. In particular, we discuss the differences of probabilistic PCA, Latent Subspace Estimation and ordinary PCA, and show that F-statistics are more naturally interpreted in a probabilistic PCA framework. We also show that individual-based F-statistics can be accurately estimated from probabilistic PCA in the presence of large amounts of missing data. We compare estimates from probabilistic PCA-based framework to ADMIXTOOLS 2 using simulations and published data, and show that this joint estimation framework addresses limitations of estimating F-statistics and PCA independently.

1 Introduction

Most populations live in heterogeneous and changing environments, and thus will exhibit some population structure, which we expect to change over time. Over short time scales, the two principal forces affecting this structure are first genetic drift, which increases differentiation between populations over time due to isolation. Second, secondary contact between isolated populations, causes intermediate genotypes and thus reduces differentiation. For the purpose of this paper, we treat the terms gene flow, admixture and migration synonymously. A common goal in population genetics is to characterize the genetic variation caused by these processes.

In particular for humans, there has been a long-standing debate on how we conceptualize genetic population structure, whether populations as such exist, or to what degree they are the result of biased sampling designs Serre and Pääbo (2004); Rosenberg et al. (2005); Peter et al. (2020), and how they affect ancestry estimation Mathieson and Scally (2020); Simon and Coop (2023) and impact association studies Price et al. (2006). Questions like these are of fundamental importance because they impact equitable access to genetic medicine

Popejoy and Fullerton (2016), how we think about race in the context of genetics Lewontin (1972); Novembre (2022) and other (mis)uses of genetic variation.

1.1 Overview of methods to study admixture

There is a large array of approaches and methods available to make inferences about population structure (review by Schraiber and Akey (2015)), aiming at different time scales, making different modelling assumptions or treating data differently. One wide class of methods are global ancestry methods. They summarize the entirety of the genome into a small number of summary statistics, with the idea that different genomic loci are (pseudo-)independent replicates of the historical process Pritchard et al. (2000); Gopalan et al. (2016); Patterson et al. (2012); Alexander et al. (2009); Tang et al. (2005). Global ancestry methods stand in contrast to local ancestry methods, that use an ancestral recombination graph or an approximation thereof, to infer the detailed ancestry of each locus Lawson et al. (2012); Hellenthal et al. (2014); Speidel et al. (2019); Kelleher et al. (2016). Global ancestry methods are widely used because they tend to be much simpler and easier to interpret than local ancestry methods, and are sufficient for many applications Pritchard et al. (2000); Patterson et al. (2006).

For large datasets with dozens of individuals spanning a wide range of sampling locations, we can further distinguish between joint analyses that use data from all individuals in an multivariate framework, such as Principal Component Analysis (PCA, Cavalli-Sforza and Piazza (1975); Patterson et al. (2006); Novembre et al. (2008) or structure Pritchard et al. (2000); Alexander et al. (2009)), and summary-statistic-based approaches relying on statistics that include two or a small number of populations(e.g. F_{ST}-based or site-frequency-spectrum based methods), and use large numbers of these summaries to build more complex models Excoffier and Foll (2011); Kamm et al. (2020); Gutenkunst et al. (2009).

F-statistics A popular framework based on summary statistics, particularly in in studies of ancient human populations Orlando et al. (2021), relies on a set of statistics called F-statistics Patterson et al. (2012); Peter (2016) . As we will define in full detail in the theory section, F-statistics measure the genetic drift shared between two, three, or four populations Patterson et al. (2012); Peter (2016). These patterns of shared drift are then compared by a null model corresponding to a population tree connecting the sampled individuals. Gene flow between distinct populations leads to F-statistics inconsistent with the purported trees, and thus are the basis of an intuitive and powerful framework to test hypotheses of admixture (Fig. 2).

Estimation of F-statistics Despite their name, F-statistics should be thought of as parameters, that are defined in terms of *population* allele frequencies. Since we typically only have genotype data from a small subsample of individuals, the population allele frequencies are unobserved, and must be estimated from *sample* allele frequencies. This estimation is not trivial: Patterson et al. showed that a naive estimator would be biased, and introduced a bias-correction term Patterson et al. (2012). This bias is largest if the sample size is small (e.g. for single genomes), and will reduce in magnitude for larger samples (see eq. 3).

To do inference, we use combinations of statistics that include two, three or four populations. Thus, treating each individual independently would give us the potential to compute more statistics, and thus the highest resolution representation of the underlying population structure. However, large numbers of statistics will be harder to interpret, and would have relatively low statistical accuracy.

Thus, samples are typically grouped into as large populations as possible, to improve statistical accuracy and to make interpretation easier. This creates a trade-off between a fine-scale view of population structure with low statistical accuracy, and a coarser-scale representation that has the danger of an overly simplistic view of the genetic structure of the studied species.

Missing data This issue is compounded by missing data: In ancient DNA, the amount of preserved DNA is often a limiting factor, and hence low-coverage genomes, and heterogeneous data quality are the norm Orlando et al. (2021). In studies with dozens of individuals, variable levels of missingness add additional statistical noise that make individual-based statistics even less accurate, leading to a larger pressure to group individuals, and hence simplify the population structure.

To rectify this trade-off between analysis resolution and statistical accuracy, we use PCA to develop a multivariate framework that jointly estimates F-statistics between all sets of individuals in a large data set. Our framework has the advantage that it does not require *a priori* assignment of individuals to populations, and allows for the imputation of data missing at random. These advances allow for accurate and finely-grained analyses of population structure.

1.2 PCA

PCA is one of the most widely used global ancestry methods to uncover population structure Cavalli-Sforza and Piazza (1975); McVean (2009); Engelhardt and Stephens (2010). PCA has the advantage that it makes minimal assumptions on the underlying data, and thus can be applied flexibly even when little is known about the underlying patterns of genetic variation. The key empirical feature of PCA is that it tends to cluster similar individuals nearby, and provides an easy to understand and often useful visualization of the genetic variation in the data. PCA is also widely used to model population structure in association studies, where structure is an undesired covariate that has to be regressed out Price et al. (2006). The big caveat is that it can be difficult to interpret PCAs, since there is no underlying mechanistic model, nor does it lend itself to model comparisons or formal tests of admixture McVean

(2009); Novembre and Stephens (2008).

PCA was introduced and popularized as a tool to study human genetic structure by Cavalli-Sforza et al., and using just a handful of genetic loci, they were able to use PCA to accurately describe patterns of human genetic variation, and to make inference about their possible causes, although the way these patterns were analyzed were typically qualitative Menozzi et al. (1978); Sforza and Sforza (1995); Cavalli-sforza et al. (1996).

As is still the norm with F-statistics, Cavalli-Sforza aggregated individuals into populations before performing PCA. With the advent of genomic data, his methods became superseded by individual-based approaches, which have higher resultion, and do not require *a priori* grouping individuals we discussed above Patterson et al. (2006); Novembre et al. (2008); Price et al. (2006).

1.3 Probabilistic PCA and Latent Subspace Estimation (LSE)

One issue of PCA is that it does not differentiate sources of variation. In particular, it is often desirable to tease apart the biological variation that is due to the shared population history (leading to the *population* allele frequency, and the statistical variation that is due to missing data, or limited sampling. Analogous to F-statistics, PCA also suffers from the issue that the population structure estimator is biased when sample allele frequencies are used instead of population allele frequencies.

Probabilistic PCA (PPCA) is an extension of PCA that aims to rectify this by incorporating an explicit probabilistic framework Tipping and Bishop (1999). The key idea is to decompose the variation in our data into a (low-rank) covariance matrix that models population structure, and a diagonal matrix that captures variation due to sampling.

PPCA was introduced by Tipping and Bishop (1999) in a Gaussian setting, where the sampling error is modelled as normally distributed, and each individual is assumed to have the same sampling error (homoskedastic noise). However, in population genetics, genotypes are discrete and thus the sampling errors are binomially distributed. Furthermore, individuals coming from populations with different effective population sizes will have different heterozygosities resulting in heteroskedastic noise. This model has been implemented in Latent Subspace Estimation (LSE) Chen and Storey (2015); van Waaij et al. (2023); Cabreros and Storey (2019). We give a detailed description of both frameworks in sections 2.3 and 2.4. In summary, the main different between PCA, PPCA and LSE is how they model the noise in the observed data due to sampling (see Fig.1).

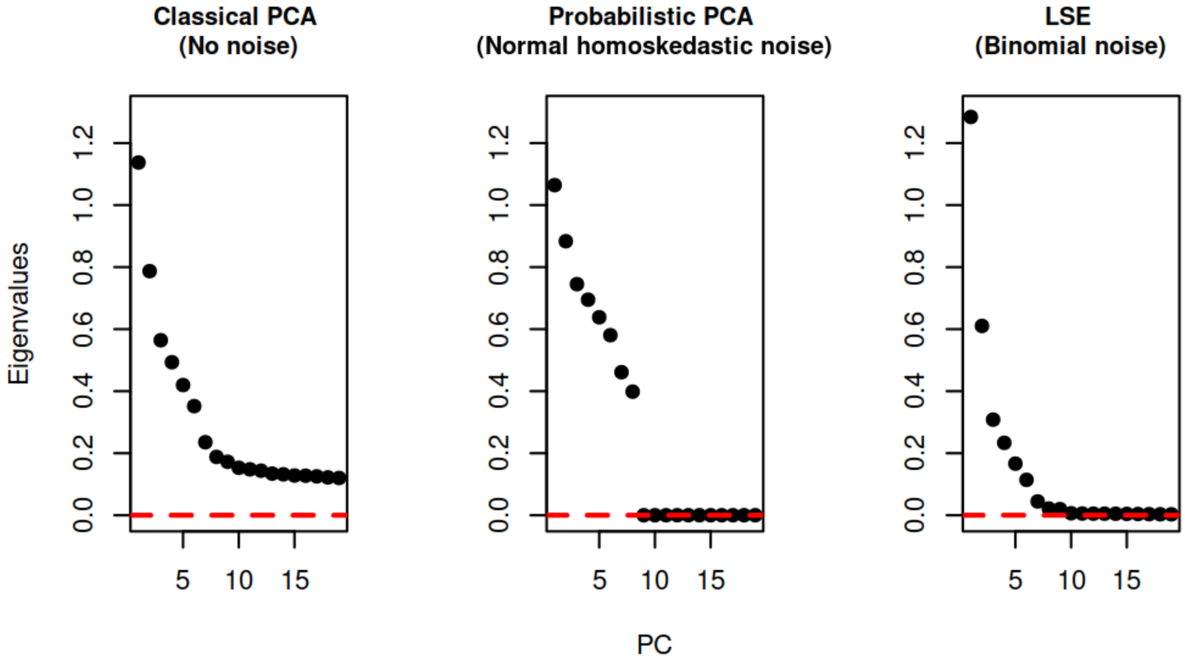

Figure 1: Comparison of PCA, PPCA and LSE: We simulated genotypes of 10 populations with 10 individuals each, and compared the top eigenvalues obtained from different PCA methods.

1.4 PCA and F-statistics

A common analysis paradigm in ancient DNA is to use PCA for exploratory and descriptive analyses, and then follow them up with methods based on F-statistic for a more formal treatment (typically in a third step, methods that synthesize many F-statistics are also applied, although we will not cover those here) (Orlando et al., 2021). Because they were developed independently and occur at different stages of the analyses, PCA and F-statistics use different data groupings and different normalizations, and are usually not quantitatively compared.

Recently, we showed that the information contained in F-statistics and PCA is closely related, and that F-statistics can be interpreted geometrically in the context of PCA Peter (2022); Oteo-García and Oteo (2021). In this framework, we showed that F-statistics can be used to interpret PCA; genetic drift will move individuals further apart from each other on a PCA-plot. Independent drift, i.e. unconnected populations, will drift on orthogonal axes in PCA-space. On the other hand, admixed individuals will be placed between their populations of origin, and these placements can be measured using F-statistics.

However, while Peter (2022) develops a theoretical link between PCA and F-statistics, it does not deal with statistical uncertainty, and thus cannot easily be applied to noisy data.

Here, we develop a statistical framework to jointly estimate PCA and F-statistics. We show how in particular the choice of the PCA-algorithm (classical PCA, PPCA, LSE) greatly impacts the results. Using simulations, we show that using PPCA-based F-statistics can result in higher accuracy than using the naive estimators, especially when there is missing data. We also draw comparison between PPCA-based framework and ADMIXTOOLS 2 using published Neanderthal samples. Our approach improves the estimation of F-statistics, and leads to some natural suggestions about how and when different PCA methods should be used.

2 Theory

In this section, we give a more detailed and formal overview of F-statistics and PCA, and their various estimators. We show that the model underlying F-statistics is very similar to that of PPCA, and identical to that of LSE. These findings will then be substantiated in the results section using simulated and real data.

2.1 F-statistics

We follow the original notation of Patterson et al. (2012), and distinguish between the *parameters* F_2, F_3 and F_4, and their *estimators* from empirical data, denoted by lower-case f_2, f_3 and f_4. The three F-statistics are defined in terms of population allele frequencies as follows:

$$
\begin{aligned}
F_2(X_1, X_4) &= \frac{1}{S}\sum_{s=1}^{S}(\mathcal{X}_{1s} - \mathcal{X}_{4s})^2 \\
F_3(X_1; X_3, X_4) &= \frac{1}{S}\sum_{s=1}^{S}(\mathcal{X}_{1s} - \mathcal{X}_{3s})(\mathcal{X}_{1s} - \mathcal{X}_{4s}) \\
F_4(X_1, X_2; X_3, X_4) &= \frac{1}{S}\sum_{s=1}^{S}(\mathcal{X}_{1s} - \mathcal{X}_{2s})(\mathcal{X}_{3s} - \mathcal{X}_{4s}).
\end{aligned}
\tag{1}
$$

Here, S is the total number of SNPs, and $\mathcal{X}_{is}$ is the (unobserved) population allele frequency in population X_i at SNP s.

Assuming a tree-like relationship between populations, $F_2(X_1, X_4)$ is interpreted as the branch length between populations X_1 and X_4 (Fig. 2 A) and it reflects the expected amount of drift that occurred between X_1 and X_4. $F_3(X_1; X_3, X_4)$ represents the amount of drift that occurred on the external branch connecting X_1 to the common ancestor node of X_3 and X_4 (Fig. 2 B). Under a tree-like model, F_3 will always be non-negative. However, in the case where X_1 is admixed between X_3 and X_4, $F_3(X_1; X_3, X_4)$ may be negative, and hence this is used as a test for admixture Peter (2016); Patterson et al. (2012). $F_4(X_1, X_4; X_2, X_3)$ represents the covariance between shared drifts between X1,X4 and X2,X3. This would be represented by the internal branch between the common ancestor nodes of X_1, X_2 and X_3, X_4 (Fig. 2 C). The F_4- statistic, with a different permutation of the populations, can be used as test of admixture. $F_4(X_1, X_2; X_3, X_4)$ is expected to be 0 if X_1, X_2, X_3, X_4 are related to each other by a tree (Fig. 2 D). In this case, a significantly non-zero value suggests a departure from the null model of treeness.

It is straightforward to verify that F_3 and F_4 can be written in terms of F_2s as:

$$\begin{aligned} F_3(X_1; X_3, X_4) &= \frac{1}{2}[F_2(X_1, X_3) + F_2(X_1, X_4) - F_2(X_3, X_4)] \\ F_4(X_1, X_2; X_3, X_4) &= \frac{1}{2}[F_2(X_1, X_3) + F_2(X_2, X_4) - F_2(X_1, X_4) - F_2(X_2, X_3)]. \end{aligned} \tag{2}$$

Hence, in practice, all F-statistics can be calculated from linear combinations of F_2. Patterson et al. showed that the naive application of eq. 1 to *sample* allele frequency data will be biased, particularly when the sample size is small. They thus introduced the bias-corrected estimator Patterson et al. (2012)

$$f_2(X_1, X_4) = \frac{1}{S}\sum_{s=1}^{S}\left[(x_{1s} - x_{4s})^2 - \frac{x_{1s}(1 - x_{1s})}{n_{1s} - 1} - \frac{x_{4s}(1 - x_{4s})}{n_{2s} - 1}\right]. \tag{3}$$

Here, we denote the sample allele frequency for population X_i at SNP s as x_{is}, and the number of non-missing haploids in population X_i at SNP s as n_{is}. In the case of when only a single (diploid) individual is sampled from population X_i, $n_{is} = 2$ and so these equations also work in this case. However, for (pseudo-)haploid samples $n_{is} = 1$ and the denominators are zero. Thus, the unbiased estimators do not exist for single pseudohaploid samples.

Using eq. , we see that this sampling bias also affects the calculation of f_3, but the correction terms cancel out for f_4, where the equations for f_4 and F_4 coincide:

$$f_3(X_1; X_3, X_4) = \frac{1}{S}\sum_{s=1}^{S}\left[(x_{1s} - x_{3s})(x_{1s} - x_{4s}) - \frac{x_{1s}(1 - x_{1s})}{n_{1s} - 1}\right] \tag{4}$$

$$f_4(X_1, X_2; X_3, X_4) = \frac{1}{S}\sum_{s=1}^{S}(x_{1s} - x_{2s})(x_{3s} - x_{4s}). \tag{5}$$

Oteo-Garcia & Oteo Oteo-García and Oteo (2021) showed that F-statistics can be defined more generally in a geometrical framework. In this framework, each population can be represented as a point in a high-dimensional allele frequency space. $F_2(X_1, X_4)$ is then the squared Euclidean distance between points representing populations X_1 and X_4 (Fig. 2 E). $F_3(X_1; X_3, X_4)$ is then a dot product of the vectors $\vec{\mathcal{X}1} - \vec{\mathcal{X}_3}$ and $\vec{\mathcal{X}_1} - \vec{\mathcal{X}_4}$ (Fig. 2 F). $F_4(X_1, X_4; X_2, X_3)$ is a dot product of vectors $\vec{\mathcal{X}_1} - \vec{\mathcal{X}_4}$ and $\vec{\mathcal{X}_2} - \vec{\mathcal{X}_3}$ (Fig. 2 G), and $F_4(X_1, X_2; X_3, X_4)$ is a dot product of vectors $\vec{\mathcal{X}_1} - \vec{\mathcal{X}_2}$ and $\vec{\mathcal{X}_3} - \vec{\mathcal{X}_4}$ (Fig. 2 H). We define the vector of allele frequencies of population i as $\mathcal{X}_i = (\mathcal{X}_{i1}, \ldots, \mathcal{X}_{iS})$.

$$\begin{aligned}
F_2(X_1, X_4) &= \frac{1}{S}||\vec{\mathcal{X}_1} - \vec{\mathcal{X}_4}||^2 \\
F_3(X_1; X_3, X_4) &= \frac{1}{S}\langle \vec{\mathcal{X}_1} - \vec{\mathcal{X}_3}, \vec{\mathcal{X}_1} - \vec{\mathcal{X}_4} \rangle \\
F_4(X_1, X_2; X_3, X_4) &= \frac{1}{S}\langle \vec{\mathcal{X}_1} - \vec{\mathcal{X}_2}, \vec{\mathcal{X}_3} - \vec{\mathcal{X}_4} \rangle
\end{aligned} \tag{6}$$

Crucially however, these interpretations only hold for the (unobserved) population allele frequencies, but not for the (observed) sample allele frequencies and thus, they cannot be directly applied to data.

2.2 PCA and F-statistics

This geometric framework provides a complementary way to understand the properties of F-statistics Oteo-García and Oteo (2021). However, many population genetic studies use a large number of SNPs (in the order of a million), and it is not possible to visualize population vectors in such a high dimensional space. Peter at al. showed that one can do dimensionality reduction on such datasets with PCA, and use the top PCs to estimate F-statistics efficiently Peter (2022).

We illustrate this in Fig. 2: $F_2(X_1, X_2)$ can be thought of as the squared Eucledian distance between populations X_1 and X_2 in PCA-space. Similarly, $F_3(X_1; X_3, X_4)$ is represented as the length of projection of the vector $X_1 - X_3$ on $X_1 - X_4$. Internal branch length $F_4(X_1, X_4; X_2, X_3)$ can be described as the length of projection of $X_2 - X_3$ on $X_1 - X_4$ on PCA, and the test of admixture $F_4(X_1, X_2, X_3, X_4)$ is equivalent to the length of projection of $X_1 - X_2$ on $X_3 - X_4$.

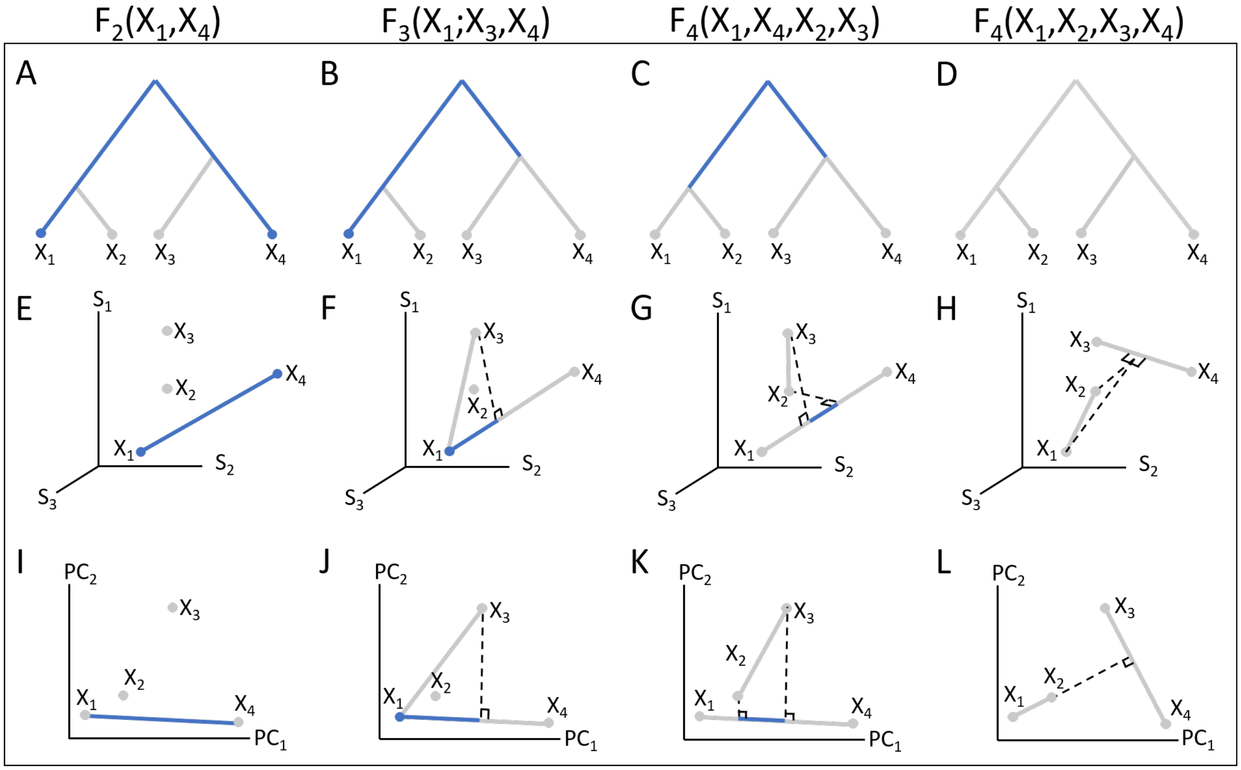

Figure 2: Schematics showing different interpretations of F-statistics. The columns represent $F_2(X_1, X_4)$, $F_3(X_1; X_3, X_4)$, $F4(X_1, X_4; X_2, X_3)$, $F4(X_1, X_2, X_3, X_4)$. The first row shows a tree interpretation of each statistic, the second row shows F-statistics in an allele-frequency space with three axes representing three SNPs, and the last row is the interpretation of F-statistics on a PCA. Blue lines represent the statistic, and the dotted lines represent orthogonal projections. Black squares denote right angles.

To formalize the relationship between f-statistics and PCA, let us assume our dataset $\mathbf{X}$ has M populations and S SNPs, such that our genotype matrix $\mathbf{X}$ has the dimension $[M \times S]$. The i-th row of $\mathbf{X}$ corresponds to the S-dimensional row-vector $\vec{\mathcal{X}_i}$ whose entries are allele frequencies $\in$ [0,1]. PCA of mean-centered $\mathbf{X}$ allows us to project this S-dimensional data onto a q dimensional subspace, where $q < M$. $q = M - 1$ represents the case where we retain all the PCs, and thus PCA only rotates $\mathbf{X}$. However, in practice we often only need few PCs ($q \ll M$) to explain most variation in the genetic data Peter (2022), which can greatly simplify calculations and visualizations.

A common algorithm to estimate PCs is viaSingular Value Decomposition (SVD). For this approach, we first mean-center $\mathbf{X}$ to a matrix $\mathbf{X}_c$, and then decompose $\mathbf{X}_c$ into an orthonormal matrix $\mathbf{U}$, a diagonal matrix $\mathbf{E}$, and another orthonormal matrix $\mathbf{V^T}$.

$$\mathbf{X}_c = (\mathbf{UE})\mathbf{V}^T = \mathbf{WL},$$

We perform SVD to decompose $\mathbf{Y}$ into a product of $\mathbf{W} = \mathbf{UE}$ and $\mathbf{L} = \mathbf{V}^T$. In the context of PCA, $\mathbf{W}_{[M\times M]}$ is a matrix of principal components (where the i-th row corresponds to the i-th PC) and contains information about structure, while $\mathbf{L}_{[M\times S]}$, also known as the SNP loadings contains the contribution of each SNP to each PC, and can be used to identify outlier SNPs that may be potential candidates for selection Gower (1966).

Since F-statistics can be written as dot products in an allele frequency space (eq. 6), and dot products are invariant to rotation, PCA will not change F-statistics as long as we retain all PCs and we can calculate F-statistics from the PCs directly Peter (2022):

$$\begin{aligned}
F_2(X_1, X_4) &= \frac{1}{S}\sum_{s=1}^{S}(\mathcal{X}_{1s} - \mathcal{X}_{4s})^2 \\
&= \frac{1}{S}\sum_{s=1}^{S}\left((\mathcal{X}_{1s} - \mu_s) - (\mathcal{X}_{4s} - \mu_s)\right)^2 = F_2(X_{c1}, X_{c4}) \\
&= \frac{1}{S}\sum_{q=1}^{M}(w_{1q} - w_{4q})^2 = F_2(W_1, W_4)
\end{aligned} \tag{7}$$

A difficulty in the practical application of this result is that the geometric considerations of Oteo-García and Oteo (2021) and Peter (2022) only hold for the (generally unobserved) population allele frequencies, but not for sample allele frequencies. In ancient DNA, PCA is most commonly run directly on individual-level genotype data Patterson et al. (2006), and hence on the biased sample allele frequencies. Thus, applying the PCA-based estimator (eq. 7) to calculate F-statistics would likewise result in biased estimate.

Oteo-García et al. resolved this by calculating F-statistics using populations with large number of individuals and with no missing data Oteo-García and Oteo (2021). In Peter (2022), unbiased estimates of the PCA reconstructions were obtained indirectly by first calculating all pairwise F_2-statistics, and then performing a multidimensional-scaling decomposition equivalent to PCA.

2.3 PPCA and F-statistics

Since PPCA and LSE, separate the population variation from sampling uncertainty, we can use them to calculate (approximately) unbiased f-statistics.

Here, we develop two related approaches that aim to calculate the bias-corrected estimates of F_2 from PCA, using the seperation of sampling error and by explicitly separating out the error in allele frequencies.

The first approach is based on probabilistic PCA (PPCA) Tipping and Bishop (1999); Agrawal et al. (2020). The model fit by PPCA can be written as

$$\mathbf{X}_c \sim N(0, \mathbf{W}\mathbf{W}^T + \Psi\mathbf{I}),$$

where N denotes a multivariate normal distribution, $\mathbf{X}_c$ is again the centered genotype matrix, $\mathbf{W}$ is a $M \times q$ matrix of linear mappings and , $\mathbf{I}$ is the identity matrix and Ψ is a noise term. Intuitively, $\mathbf{W}\mathbf{W}^T$ captures the covariance in the observed data, analogous to the F-statistics, and $\Psi\mathbf{I}$ is analogous to the bias-correction term. Since Ψ is a scalar, all entries on the diagonal of $\Psi\mathbf{I}$ will be the same, and thus the model is homoskedastic, i.e. all individuals are assume to have the same error.

The goal of PPCA is to estimate the parameters of the model, namely $\mathbf{W}$ and Ψ given the observed data. For complete data, Ψ and $\mathbf{W}$ can be calculated from SVD using the maximum-likelihood estimators

$$\hat{\Psi} = \frac{1}{M-q} \sum_{j=q+1}^{M} e_j$$
$$\hat{\mathbf{W}} = \mathbf{U}(\mathbf{E} - \hat{\Psi}\mathbf{I}),$$

where e_j is the j-th entry on the diagonal of $\mathbf{E}$.

Thus, the MLE solution of PPCA results differ from those of classical PCA only in that the PCs are "shrunk" by a common term that incorporates the noise discarded in the PCA, notably a PCA plot would look almost the same, only the axes scales would be different. Thus, for PPCA we need to set the number of retained dimensions q a priori, since changing q requires rescaling all PCs. In addition, setting $q = M - 1$ (or equivalently, $\Psi = 0$), recoups classical PCA.

2.4 LSE

LSE (Linear Subspace Estimation) is a dimensionality reduction technique quite similar to PPCA, with the difference that LSE accounts for the heteroscedasticity in the data Chen and Storey (2015), and explicitly models the binomial error in genetic data. In this algorithm, we calculate the heterozygosity $d_{jj} = \frac{1}{S}\sum_i 2x_{ij}(1 - x_{ij})$ from $\mathbf{X}$. We define $\mathbf{D}$ as a diagonal matrix with j^{th} entry as δ_{jj}. We then estimate covariance matrix $\mathbf{G} = \frac{1}{S}\mathbf{X}^T\mathbf{X} - \mathbf{D}$. The eigenvectors of $\mathbf{G}$ then span the latent subspace of $\mathbf{L}$, and the smallest $M - q$ eigenvalues converge to 0 for large M Cabreros and Storey (2019).

Crucially, if we use *all* the PCs, the f-statistics coincide with LSE (Fig. S4, see appendix for a derivation):

$$\begin{aligned} f_2(X_1, X_4) &= \sum_s (x_{1s} - x_{2s})^2 - d_1 - d_2 \\ &= G_{11} + G_{44} - 2G_{14} \end{aligned} \tag{8}$$

The covariance used by LSE is an estimated covariance matrix, and its properties differ from the sample covariance matrix used in classical PCA. In particular, sample covariance matrices are positive semi-definite, which means that all eigenvalues are non-negative. In contrast, for an estimated covariance matrix, the expectation of the smallest $M - q$ eigenvalues is zero, and thus an unbiased estimate will have both positive and negative eigenvalues. PCs with negative eigenvalues correspond to imaginary PCs, and thus, we need to adjust eq. 7:

$$F_2(X_1, X_4) = \sum_{j=1}^{M} \mathbb{I}(e_j \geq 0)(w_{1j} - w_{4j})^2 - \sum_{j=1}^{M} \mathbb{I}(e_j < 0)(w_{1j} - w_{4j})^2, \tag{9}$$

where e_j denotes the j-th eigenvalue of $\mathbf{G}$, and $\mathbb{I}$ denotes the indicator function that is 1 when the condition is satisfied, and zero otherwise.

2.5 Missing data

F-statistics: It can be difficult to estimate F-statistics when there is high amounts of missing data. A conservative approach to estimate individual-based F-statistics is to only retain sites where data is present from every single individual in the data set. However, even for moderately large data sets that quickly becomes prohibitive: As a toy example, consider a data set with 100 (haploid) individuals with 10% missing data, and 1,000,000 sites. Out of those, only 26 are expected to be covered in every single individual, which makes this approach not feasible.

Thus, grouping individuals into populations can dramatically increase the number of sites retained. A common practice is to retain sites if at least one individual in each population has data. In the above example, if we grouped the 100 individuals into 10 populations of 10 individuals each, and retained sites where at least one individual in each population carried data, we would expect no missing sites in almost all cases. However, grouping individuals may not be justified when they do not form discrete clusters, or when there are very few samples whose population assignments are unknown.

PCA: Missing data is a challenging problem while computing PCA, since SVD and eigendecompositions cannot be performed with missing data. One way to implement a PCA in case of missing data is to start with mean imputation, and then perform PCA to reconstruct or impute the missing values, and keep iterating until convergence. This can be done both for classical PCA Meisner et al. (2021) and for PPCA Tipping and Bishop (1999).

Another popular way to deal with missing data is to first construct a PCA using samples that have no missing data, and then project the low-coverage samples on this PCA. Projection is a useful way to check if low-coverage samples from the same population fall into the same position as the high coverage ones. Since sampling noise is independent for each sample, projection of samples on a constructed PCA does not require modeling of sampling bias. However, for biased representations, such as classical PCA, this bias has to be corrected when projecting, which is done e.g. in smartPCA Patterson et al. (2006).

3 Results

In this section, we describe the coalescent simulations we use to study the theoretical and statistcal properties of different PCA methods. We compare the F-statistics predicted from different versions of PCA to those calculated using ADMIXTOOLS 2 Maier et al. (2022). Finally, we illustrate a practical application of our approach on a dataset of Neandertal genetic variation.

3.1 Evaluation on simulations

We simulated 10 populations with 10 individuals each using slendr Petr et al. (2022). We use a mutation rate of 10^{-8} per base per generation, a recombination rate of 10^{-8} per base per generation, and a generation time of 30 years.

Interpretation of PCA of admixture graph In this simulation, the admixture graph (Fig. 3 A) shows a schematic of the population structure and demographic history we simulated. We visualize this structure using two different PCA plots: first we plot the first 8 PCs on the same scale (Fig. 3B), and we also show the more standard PCA biplots (Fig. 3C). The genetic structure generated by the admixture graph is apparent in both visualizations: For example, the first PC mainly separates the "left" clade (with X_1,X_2 and X_10, from the "right" clade (X_3,X_4 and X_7. The outgroup (X_5), is placed very close to zero here, which is because PC1 mainly reflects the variation within this clade, and the outgroup branch leading to X_5 is orthogonal to that variation.

PC2, on the other hand, primarily separates out the outgroup (X_5) from the remaining samples, with the admixed population X_6 fall ling in between.

Thus, even though X_5 is the most genetically diverged population, this is not evident from PC1. The reason for this is that PCA minimizes the sum of all pairwise F_2-distances between individuals, and hence the ordering of PCs depends on sample configuration McVean (2009); Elhaik (2022). In our case, there are only ten individuals in X_5, whereas we have a total of 80 individuals in the other populations (all except X_6, which has ancestry from X_5). Thus, the sum of the $\binom{80}{2} = 3160$ pairwise F_2-distances within the clade is larger than the sum of the $80 \times 10 = 800$ distances between indivuals from the clade to X_5.

Next, PC3 shows the axis of variation between X9 and the other populations. others due to the drift on X9 for many generations. The two next PCs, PC4 and PC5 mainly model the

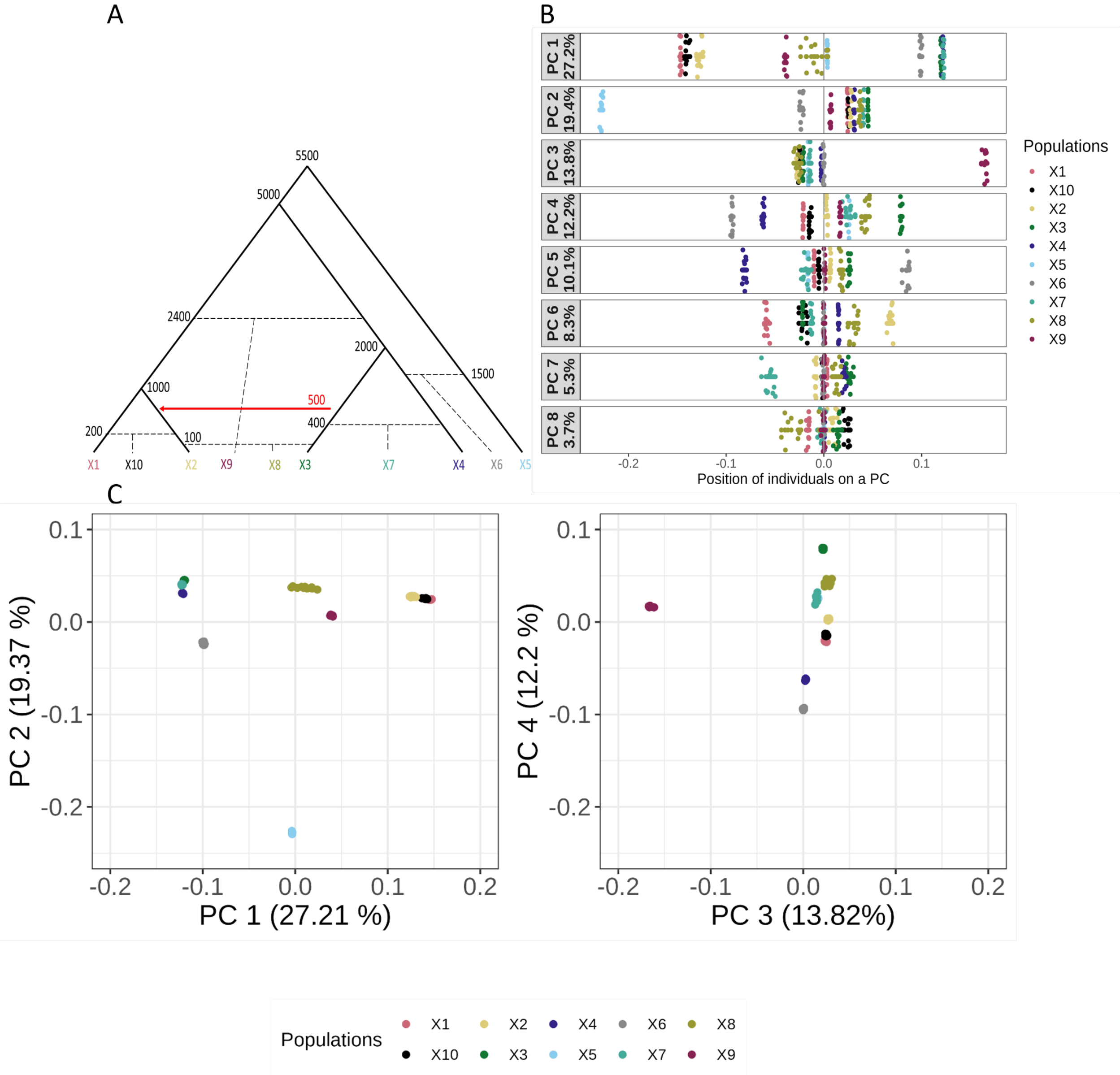

Figure 3: Overview of simulations. **A**. Schematic of the simulation. Numbers reflect split, migration and admixture times (in generations). Admixture events are shown in dashed lines, and the red arrow represents a unidirectional migration from X_3 to X_2. **B**. PPCA (with 8 PCs) of data simulated using the model depicted in A. Percentage labels show variation explained per PC. **C**. PPCA-biplots of the first 4 PCs.

variation with the "right" clade, and PC6 splits up X_1 and X_2 within the "left clade". In all three of the PCs, populations outside the clades plot very close to zero, which is because in a tree, the within-clade variation is independent of that of the clade to individuals outside of it (Felsenstein, 1973).

Thus, we find that for tree and admixture graph models, most PCs tend to show variation within a single clade, and we need a fairly large number of PCs to tease out the full tree structure. This is consistent with the theoretical expectation that the covariance matrix for a (population) tree has full rank (Felsenstein, 1973), i.e. we would expect the number of PCs required to reflect tree structure to be on the same scale as the number of populations.

PCA is most commonly visualized in biplots (Fig 3C). A main advantage of this provides a 2D-visualization of the high-level population structure (in our case, the separation of the outgroup from the ingroup samples and between the two main clades). However, we lose the structure within each clade, that is represented by PCs 3 and higher. Plotting a second biplot of PC3 vs PC4 is somewhat less informative, because the variation already explained by the higher PCs is absent.

In contrast, plotting PCs separately on one dimension has the advantage that the orthogonality of the PCs becomes more apparent, and also allows for easier quantitative comparison on how much each PC explains: The spread (i.e. variance explained) by each PC gets narrow and narrower as we move to higher PCs, and we can easily plot a large number of PCs. The drawback of this representation is that the correlation and 2D-structure in PC-space gets deemphasized. that most PCs It is interesting to note that a single PC is not informative. One needs to look at all the PCs to get a sense of the population structure. However, the information in the tree and the PC plot is the same, and so the tree can be regenerated from the (top) PCs.

Properties of PCA-based f_2 estimates We use eq. 7 to calculate F_2 for a pair of populations from the principal components based on either classical PCA, PPCA and LSE, and compare F-statistics estimated from these methods to the true values based on the of the simulated branch lengths.

We first do a comparison between f_2s estimated from PCA, PPCA and LSE with different number of PCs used (Fig. 4), and compare them to the true value and "uncorrected" statistics, (based on equation XXX).

The PCA-based F-statistics correspond to the (squared) sum of distances between individuals in Fig. 3B and (). In the top panel of Fig. 4, we use 10 diploid individuals per population. Since the sampling error is low with ten individuals, the biascorrection is small, and the uncorrected and corrected f_2-statistic based on the branch length are similar.

In this case, we see that classical PCA reaches the uncorrected length quickall three methods reach their expected value quickly (using the first, $\approx$ 7 PCs): classical PCA converges to the uncorrected estimate, and both PPPCA and LSE converge to the corrected estimate. as well as the true value, since the sampling error term is inversely proportional to the sample size. We next look at a comparison between the three methods using only one individual in each population (Fig. 4). Here we observe that $F2$ estimates from PPCA and LSE are quite close to the true value in all cases, as long as we use at least 7 PCs and less than 35. However, estimates from PCA get increasingly higher if we add additional PCs. This is expected from theory (section 2), since PCA does not account for sampling bias, and, using all PCs, we end up with the biased estimator of F_2. Finally, we repeated the estimation of F_2s with PPCA and PCA in the presence of 50% missing data (4). In this case, PPCA gives reasonably accurate results when the number of PCs used is higher than 7 and lower than around 23, while PCA results are much more noisy. Implementation of LSE is not trivial when there is missingness in data, and is not included in this analysis.

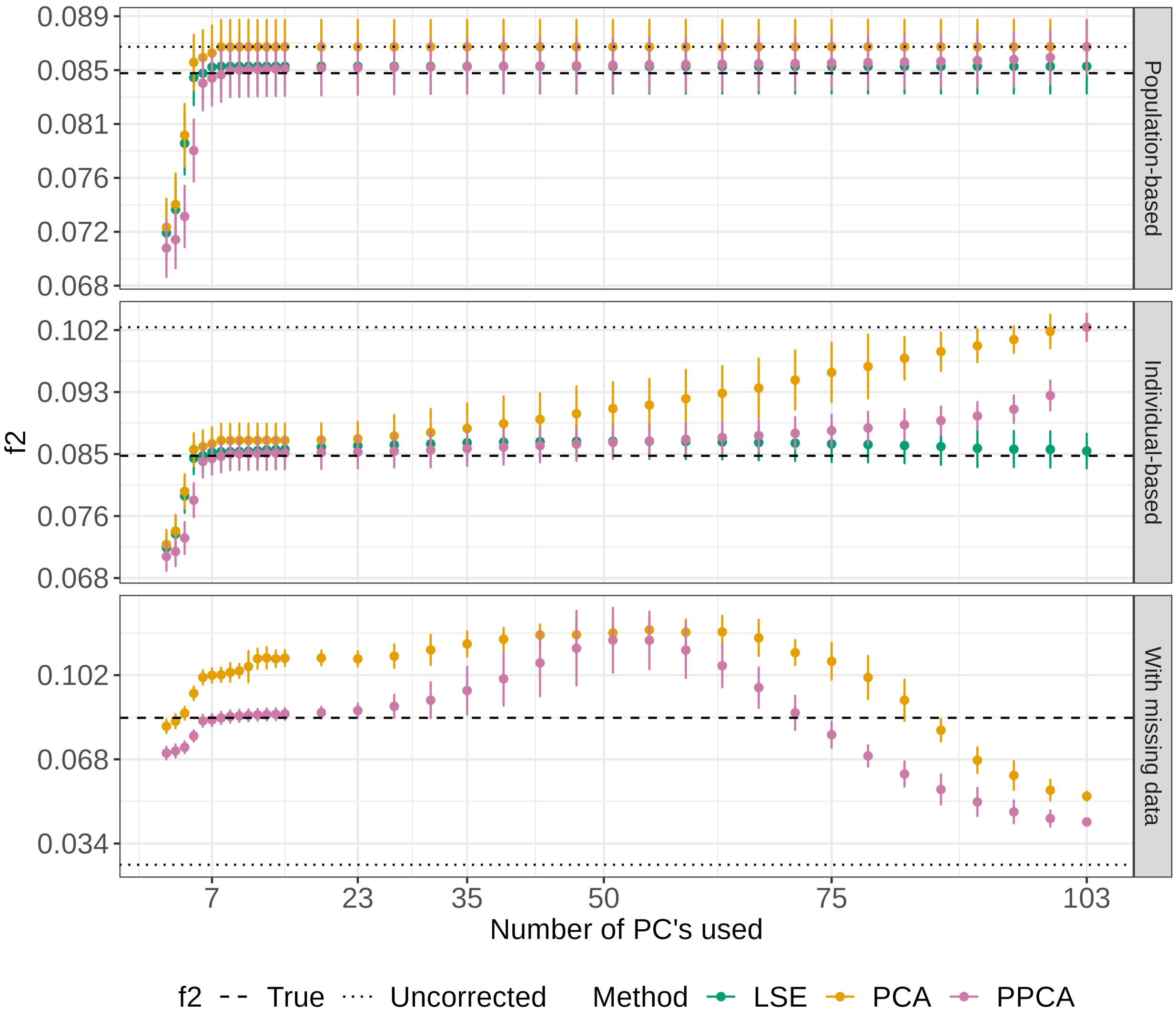

Figure 4: Comaprison of PCA approaches using $F_2(X_1, X_4)$ estimated using 10 individuals for each population (left), 1 individual for each population (middle), and 1 individual for each population with 50% missing genotypes (right). Dotted line represents uncorrected estimate and dashed lines show the true value of F_2.

In the rest of the analyses, we exclude PCA, and compare PPCA and LSE with 8 and 12 PCs to ADMIXTOOLS 2 Maier et al. (2022), which is a recent re-implementation of ADMIXTOOLS Patterson et al. (2012) that gives equivalent results. We chose 8 and 12 PCs because the number of PCs to use will not be known in most applications. We first compare f_2, f_3 and f_4 estimated by these methods in an ideal scenario, where each population has 10 individuals, and there is no missing data. In this case we find that the three methods perform well, and get F-statistics close to the truth (fig. S5).

We next address the issues listed in Introduction section 1.1. The first issue is about the estimation of F-statistics when population assignment is difficult, especially when few samples are available. We show that this can be resolved with individual-based F-statistics. In our simulations, we label each individual as a different population, and we sample one individual from each population to calculate F-statistics (fig. S6). We observe that in this case both PPCA and LSE- based frameworks perform atleast as well as ADMIXTOOLS 2 (see table S1) and the mean estimate from each method is close to the true value. However, the error bars for F_2 estimates are lower for PPCA compared to that of ADMIXTOOLS 2, specially for X_1 and X_2, which have low split-times (standard errors for ADMIXTOOLS 2 estimates for $F_2(X_1, X_2)$, $F_2(X_3, X_4)$, $F_2(X_2, X_3)$ and $F_2(X_1, X_4)$ statistics are 0.0018, 0.0015, 0.0027, and 0.0012 while the PPCA-based estimates (scale=8) for the same statistics are 0.0012,0.0012,0.0024 and 0.0011 and for XX for LSE respectively). The improved accuracy of PCA-based tools versus ADMIXTOOLS 2 is explained because PCA incorporates a succinct summary of the full data of all the individuals, and thus the PCA-based estimates can "borrow" information from related individuals in the sample that are not used to calculate the statistic at hand. In contrast, ADMIXTOOLS 2 has only one individual from each population to assess structure / admixture, and while the estimates based on admixtools 2 are minimum-variance estimators for these subsets of the data (Patterson et al., 2012), PCA-based methods do better whenever we have data from additional individuals.

3.2 Missing data

Next, we address the issue of missing data and evaluate the estimation of these methods when there is random missing data. Our implementation of PPCA on missing data is inspired from EMU Meisner et al. (2021), and is described in Methods section 4.1. We use EMU software for inferring PCs for the calculation of F-statistics. We see that PPCA is not affected by missingness, while ADMIXTOOLS 2 (run with maxmiss=1, so no SNPs are excluded) and EMU results are inflated (Fig. 5, table S1).

3.3 Test of admixture

A major application of F-statistics are tests of admixture Orlando et al. (2021). We showed in the previous section that PPCA framework can be used to calculate the point estimates of F-statistics. In this section we show that we can also get standard errors for these estimates using block-jackknife Patterson, and use these to do hypothesis testing for admixture. We simulate a gene flow from X_3 to X_2 500 generations ago with the migration rate of $\mu\epsilon[0, 0.01, 0.05]$. We then compare the estimates of F-statistics from PPCA framework to that from ADMIXTOOLS 2. We first test for admixture by checking if the estimate of $F_4(X_1, X_2; X_3, X_4)$ is significantly different from 0. We show that when there are 10 indi-

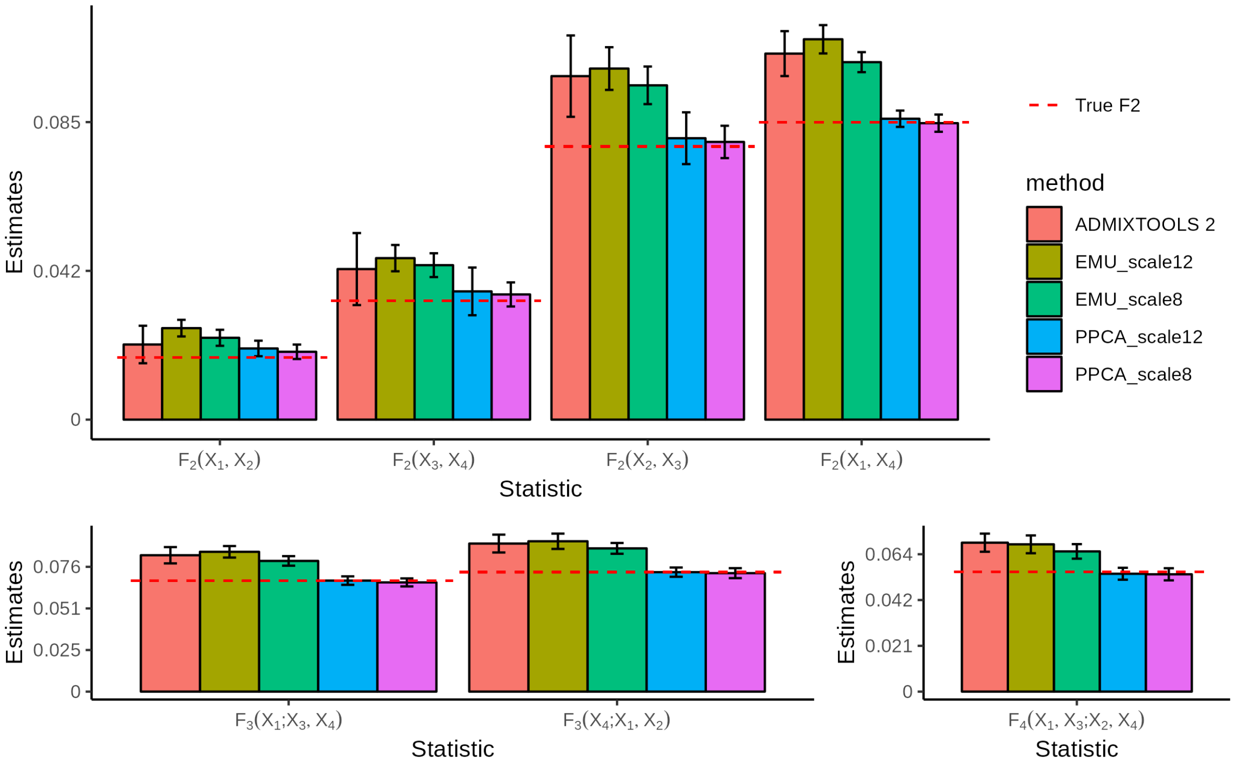

Figure 5: Comparison of PPCA and PCA to ADMIXTOOLS 2 in the presence of 50% random missingness, using population genotypes from one individual from each population.

viduals in each population, both methods perform well (Fig. S6). In case of 0 migration rate, both methods estimate F_4 for all simulations to be close to 0, while at 5% migration rate, ADMIXTOOLS 2 and PPCA-framework have the power to detect admixture (with F_4 estimate 2 standard deviations below 0) in 90% and 70% simulations respectively. At migration rate of 1%, both the methods are unable to find admixture between X_2 and X_3, and instead incorrectly predict admixture between X_1 and X_3 (F_4 estimate is 2 standard deviations above the mean) for one simulation. Reducing the number of individuals to 1 from each population reduces the power for both the methods. With 50% missingness, both methods have no false positives in case of 0 migration rate, and at 5% migration rate ADMIXTOOLS 2 and PPCA framework detect admixture in 5% and 35% simulations respectively. At 1% migration rate, ADMIXTOOLS 2 infers admixture between X_2 and X_3 in 2 simulations and incorrect admixture between X_1 and X_3 in one simulation, while PPCA framework shows no prediction of admixture.

3.4 Evaluation on neandertal dataset

To test our framework on real data, we apply it to a published dataset of archaic humans from Eurasia Hajdinjak et al. (2018). This dataset consists of low-coverage pseudohaploid sequences from late Neandertal specimens from Goyet (Goyet Q56-1), Spy (Spy 94a), Les

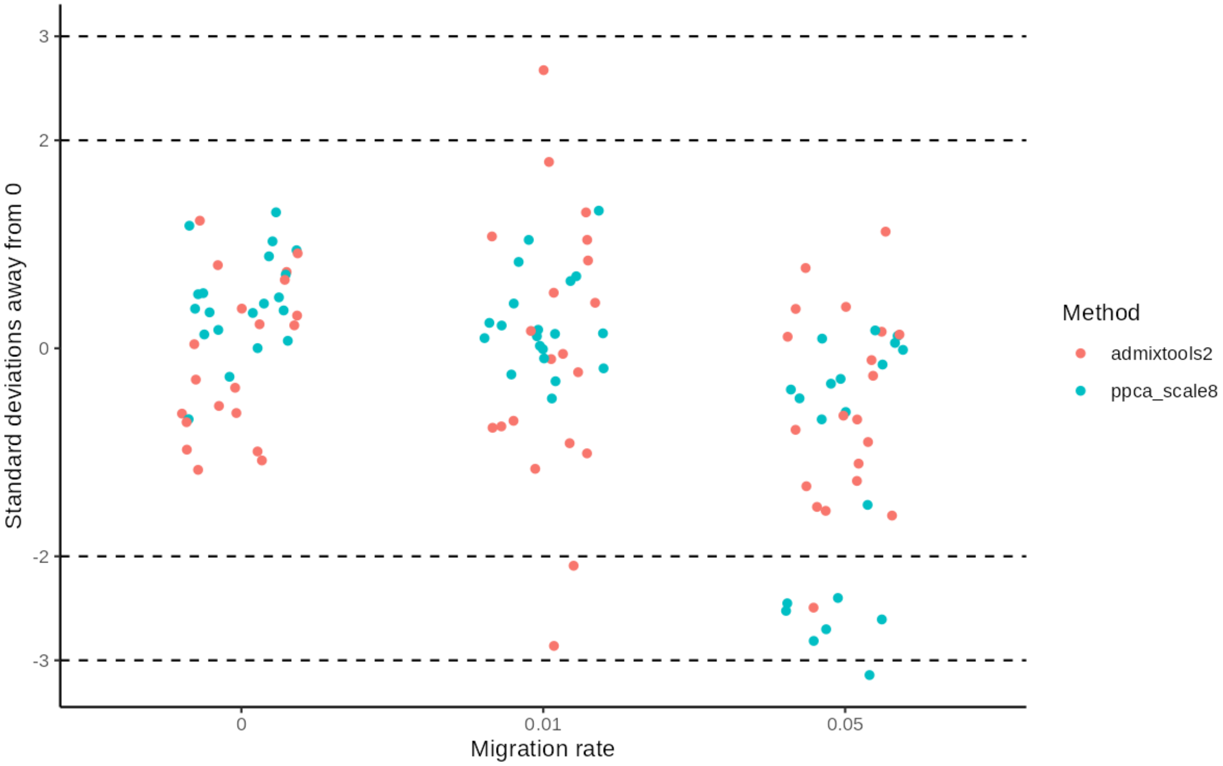

Figure 6: Test for admixture with individual-based $f_4(X_1, X_2, X_3, X_4)$ statistic. We compare ADMIXTOOLS 2 (orange)to PPCA-based-framework (blue) in the presence of 50% missingness in data.

Cottés (Les Cottes Z4-1514), Vindija (Vindija 87), and Mezmaiskaya caves (Mezmaiskaya1 and Mezmaiskaya2) along with the diploid genotype sequences from high-coverage archaic specimens from Densiova (Altai, Denisova) and Vindija caves (Vindija33.19). We first estimate PCs for this dataset using PPCA, and show that our plot captures all the features of PCA from the authors (Fig. S7, S8). However, we demonstrate that with PPCA, the user can utilize all the specimens to estimate the PCs.

We analyze how close or distant the low-coverage late Neandertals are to high-coverage Vindija Neandertal using outgroup F3 statistic. F3(Altai, Vindija33.19, X) represents the branch length extending from Altai to the common ancestor of Vindija33.19 and X, where X is a low-coverage late Neandertal. Higher value of F3 denotes closeness of X to Vindija33.19 (Fig. 2). We compare the estimates from PPCA-based framework and ADMIXTOOLS 2, and show that they have a very similar pattern (Fig. 7). It is interesting to find that F_3(Altai, Vindija33.19, Vindija 87) estimated by PPCA framework is higher than that by ADMIXTOOLS 2. We can write F_3 as a linear combination of F_2s:

$$F_3(Altai, Vindija33.19, Vindija87) = F_2(Altai, Vindija33.19) + F_2(Altai, Vindija87) \\ - F_2(Vindija33.19, Vindija87) \tag{10}$$

We looked at the values of the three f_2-terms from ADMIXTOOLS 2 and PPCA framework to see why the two methods have different values. We found that f_2(Altai, Vindija33.19) and f_2(Altai, Vindija 87) have values 0.072 and 0.135 from ADMIXTOOLS 2 respectively. Since both Vindija samples are from the same individual, the f_3 values should ideally be the same. PPCA framework outputs the values of f_2(Altai, Vindija33.19) and f_2(Altai, Vindija 87) as 0.102 and 0.115, which are in line with the expectation. In addition, f_2(Vindija33.19, Vindija 87) should ideally be 0. ADMIXTOOLS 2 shows a value of 0.0057, and PPCA-framework gives a value of 0.00094. It is interesting to note that Altai and Vindija33.19 are diploid genomes, and hence f_2(Altai, Vindija33.19) estimated with both the methods is similar for the two methods. In contrast, Vindija 87 is a pseudohaploid genome, and in this case we see that PPCA gives expected results and ADMIXTOOLS 2 does not. This is because the unbiased estimator in ADMIXTOOLS 2 for F_2 is undefined ($n_{1s} = 1$ in eq. 3).

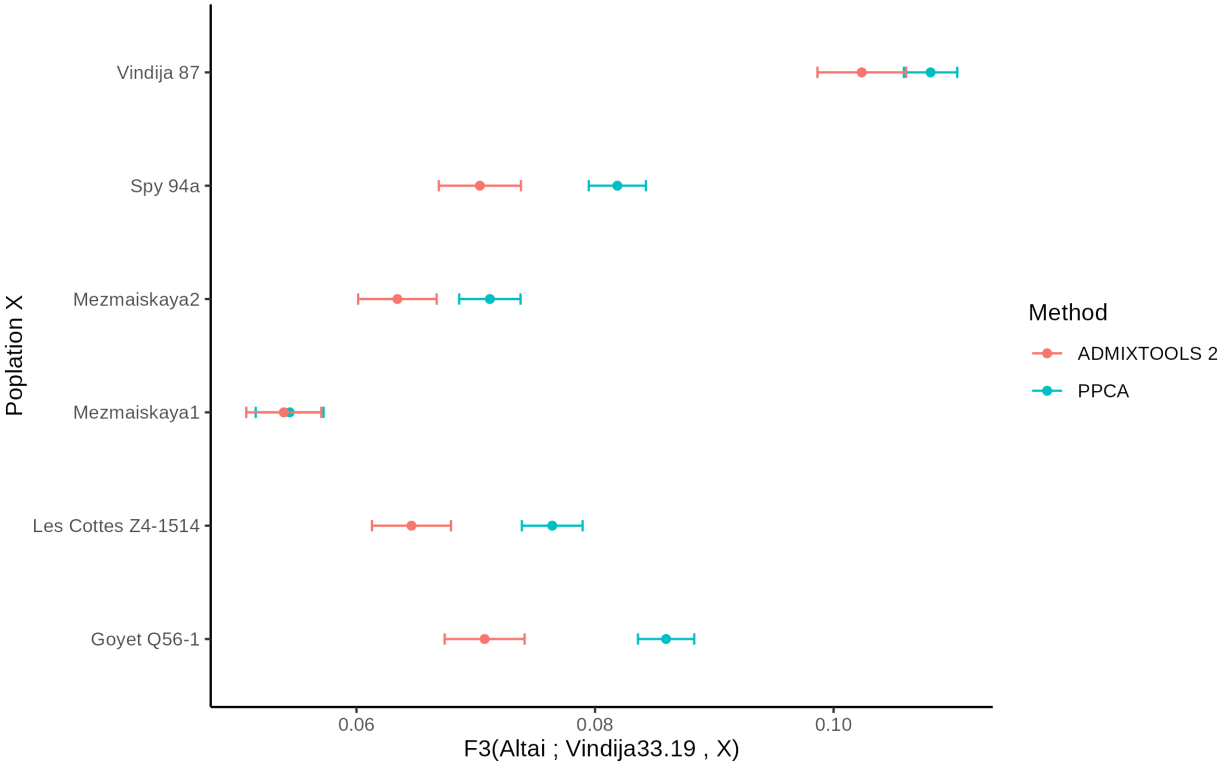

Figure 7: F_3(Altai, Vindija33.19, X) estimated with two methods. Larger value on x-axis represents more proximity to Vindija33.19. Bars show 2 standard errors.

4 Methods

4.1 PPCA implimentation

We implement PPCA using maximum-likelihood approach following Tipping and Bishop Tipping and Bishop (1999), and modify this algorithm to work with missing data. Our approach to handle missingness is inspired from EMU Meisner et al. (2021). We describe our algorithm briefly:

1. Mean center data $\mathbf{Y} = \mathbf{X} - \mu$.
2. Set missing values to 0.
3. Perform SVD
4. Calculate the Gaussian noise parameter $\Psi = \frac{1}{M-q} \sum_{j=M-q}^{M} e_j$ as the sum of square of the $M - q$ smallest eigenvalues.
5. Obtain the MLE of the q^{th} eigenvalue as $e_q - \Psi$.
6. Calculate the linear mapping matrix $\mathbf{W} = U_q(e_q - \Psi)\mathbf{I}$.
7. Reconstruct mean-centered data: $\mathbf{X_R} = \mathbf{W}(\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\mathbf{Y}$.
8. Replace missing value with reconstructed values.
9. Repeat steps 2-8 until convergence.

Our algorithm differs from that of EMU at steps 4-7, which deal with the modelling of the sampling noise and reconstruction of data, and are specific to PPCA algorithm.

4.2 Calculation of standard errors

We use a block-jackknife approach to calculate standard errors Maier et al. (2022). We divide the genome in 2 MB blocks, and estimate PCs and then F-statistics removing a block. Since the statistics obtained are not independent, we calculate variance similar to ADMIXTOOLS 2 using this equation Maier et al. (2022):

$$V = \frac{1}{g} \sum_{i=1}^{g} \frac{s_i}{S - s_i} (\hat{\theta} - \theta_i)^2 \tag{11}$$

Here, V is the variance of a statistic θ, g is the number of blocks, s_i is the number of sites in block i, and S is the total number of sites.

5 Discussion

In this study, we present a statistical framework to jointly compute PCA and F-statistics. Many ancient genetic studies use both of these tools, but make slightly different assumptions, slightly different models and different software for them. In contrast, our joint framework allow us to make sure that assumptions are consistent throughout the analysis. The key advantage is that the effect of modelling assumptions becomes apparent, and this also allows us to make novel recommendations about how PCA-based analyses should be performed and interpreted.

The connections between F-statistics and PCA allow us to provide a better understanding on how different PCA algorithms emphasize different parts of the data, and how they emphasize population structure versus sampling noise.

In particular, F-statistics enable quantitative interpretation of PCA-plots, where distances on a PCA are directly proportional to genetic differentiation, and orthogonal projections can directly be used to test for admixture. However, the above statement is true only when all the (relevant) PCs are used. Visualizing two or only few PCs may be insufficient to accurately visualize population structure and admixture.

Interpreting PCA plots: There is a considerable literature aimed at interpreting PCA-plots, which generally fall in two approaches: First, for simple models, such as lattice models Novembre and Stephens (2008) and isolation-without-migration models McVean (2009), the eigenvectors, and hence the PCs, can be calculated analytically. These approaches are powerful to understand the "inner workings" of PCA, but do not deal with the variation and noise inherent in sample data. Thus, much more common are interpretations based on empirical observations, such as that the first PC commonly aligns with the axis of an expansion Cavalli-sforza et al. (1996), that PCs tend to align with the branches of a population tree or that the first two PCs recap a map of Europe (Novembre et al., 2008; Cavalli-Sforza and Piazza, 1975).

However, these observational guidelines have provided to not hold particularly well in simulations, and have been shown to be strongly influenced by sampling schemes, simulation details and other factors (Novembre and Stephens, 2008; DeGiorgio and Rosenberg, 2013; Elhaik, 2022; Jay et al., 2013) enough that interpretations of the first few PCs in terms of demographic history are frequently discouraged.

Our link between PCA and F-statistics provides a different way of interpreting PCA-plots, by studying it in terms of the embedded F-statistics. A key difference is that our approach is quantitative, and provides exact results for all PCs, and good approximations when only the first few PCs are used Peter (2022). Then, F-statistics can directly be read off one-dimensional PCA-plots such as that in Figure 3B: $F_2(X_1, X_2)$ corresponds to the (squared) sum of the distance in that figure, and we can see from the plot that PC6 is the main PC that teases out that axis of variation (because the two populations are furthest apart).

The drawback of our approach is that F-statistics are not directly interpretable in terms of demographic history, and we need additional steps to link them with what we are typically interested in.

5.0.1 Different PCA methods

In addition, the different PCA methods we discussed here differ in how they deal with biological vs. sampling variation, and how different sources of variations are emphasized. This gives us some insights for when different PCA algorithms should be used.

Classical PCA Classical PCA is mathematically the simplest, and still the most widely used method for visualizing population structure. It has an interpretation in terms of pairwise coalescent times (McVean, 2009). However, because classical PCA incorporates *all* variation in the data, it does not distinguish between variation due to population structure, and variation due to sampling. This can lead to problems: For example, one cannot directly project additional samples to a PCA, instead some correction is required otherwise some "shrinkage" occurs where new samples are projected closer to the origin (Patterson et al., 2006; Wang et al., 2015). Thus, in our opinion classical PCA should be the primary choice for analyzes primarily aimed at quality control, because incorporating noise allows it to reveal technical artifacts and outliers Effects of different sequencing or capturing techniques, will very often be visible on a PCA-plot, while they may be more hidden and only partially corrected in PPCA.

Probabilistic PCA . Our results suggest that the methods that separate sampling noise from population structure, such as PPCA AND LSE, are preferable to classical PCA when the primary goal is to depict population structure. PPCA is the simpler of the two methods, because it assumes homoskedastic noise, i.e. that all samples have the same heterozygosity / sampling variance. As Tipping and Bishop (1999) showed, without missing data, the maximum-likelihood estimator of PPCA results in virtually identical PCA plots compared to those obtained from a classical PCA; the only differences are that the axes for the first few PCs will be re-scaled, and that the majority of PCs are set to zero (see Fig. S2 and Fig S3).

LSE LSE goes one step further by adding heteroskedastic, binomial noise. This ensures that both PCA and F-statistics use the exact same modelling assumptions, and thus LSE-based PCA plots are directly comparable to F-statistics. This relationship is exact if all PCs are used. However, one advantage of LSE is that the eigenvalues corresponding to the PCs irrelevant for population variation have an expectation of zero. Since these eigenvalues are directly proportional to the amount of variance explained per PC, and to the contribution of these F-statistics, we expect that truncating them will yield good results, which is indeed what we see in our simulations (Fig. S4).

In real data, this expectation of zero for the eigenvalues will typically yield both positive and negative eigenvalues, which is why the equation for computing f_2 needs to be adjusted (eq. 9), and subtract the contribution to F_2 by PCs that have negative eigenvalues. Alternatively, we could interpret these PCs as being on the imaginary axis on the complex plain, but both interpretations are simply a result from the fact that we design an unbiased estimator for a quantity with an expectation of zero.

From a theoretical point of view, the model optimized in LSE is more desirable to that in PPCA, because it takes into account that different individuals might have different heterozygosities. However, the advantage of PPCA is that it is relatively easier to implement with missing data, and even for single pseudohaploid samples, which are common in ancient DNA applications (Fig. 5) Tipping and Bishop (1999); Orlando et al. (2021).

Projection PCA . A further approach that is used for ancient DNA is to project data low-coverage data onto a "reference"-data set. This has three useful advantages: First, the overall shape of the PCA only depends on the reference data set. Thus, PCA-plots using the same reference data become directly comparable, which can be useful as a quick way of assessing population variation, e.g. using the Western Eurasian PCA used in studies of Holocene ancient DNA studies from Europe and Western Asia (Haak et al., 2015). Second, it deals with missing data in the projected samples, because often only a subset of sites are required for an accurate projection. Third, it also deals with sampling noise in the projected sample, because the sampling noise is orthogonal to the variation in the reference data set, and thus gets removed by the projection.

However, the drawback of using projections, and why they, in our opinion, are inferior to probabilistic PCA, is that they do not capture the full variation in the data. In particular, only the variation in the reference data is considered, but not the variation that is private to the projected samples.

In terms of F-statistics Peter (2022),

$$F_2(X_1, X_4) = \underbrace{F_2^{(\text{PCA})}(X_1, X_4)}_{\text{visible in projection PCA}} + \underbrace{F_2^{(\text{Projection error})}(X_1, X_4)}_{\text{hidden in projection PCA}}$$

This is particularly problematic when projecting ancient human samples onto a modern reference data set, because the differentiation between ancient human population is often considerably larger than that between present-day populations (Haak et al., 2015; Lazaridis et al., 2014), and thus projection PCA have no quantitative interpretation.

5.1 Plotting PCs

One of the major benefits of using F-statistics and PCA in a joint framework is that it enables a quantitative interpretation of PCA. Thus, our results directly suggest some recommendations for how PCs should be plotted when displaying population genetic variation is desired.

First, the scale at which PCs are plotted matters, and the units on the different PCs are comparable (Figure 3B). Thus, we recommend that PCA-biplots should be plotted with "fixed" aspect ratios where the x- and y-axes have the same units and scale, contrary to the current common practice where the axes are scaled arbitrarily (e.g. Novembre et al. (2008); Peter et al. (2020)). It also highlights the benefits of plotting PCs on a shared axis such as in Figure 3B, where the decreased variance explained by each subsequent PC is directly visualized, and tens of PCs can be plotted jointly.

Second, it has been a "best-practices" suggestion that PCs are plotted along the proportion of variance explained by each PC (e.g.Novembre and Peter (2016); Elhaik (2022)). While we think this holds, it is important to recognize that different PCA algorithms handle this differently: In classical PCA, the proportion of variance is out of the total variance, including population structure, within-population variation and sampling noise. In contrast, in PPCA the number of PCs retained is a parameter, which *sets* the proportion of variance explained by population structure. Thus, we obtain the (user-controlled) proportion of variance due to noise (corresponding to Ψ), and the proportion of variance each PC explains as a proportion of the variation explained by population structure. In addition, the proportion of variance explained from a projected PCA will depend on the number of PCs or eigenvectors used for recontruction of projected data Patterson et al. (2006). Thus, the percent of variation explained between classical PCA, PPCA, LSE and projected PCA are not directly comparable.

5.2 Grouping individuals

A benefit of the interpretation that we developed is that it demonstrates when and how we can use PCA to group individuals into populations for F-statistics-based analysis. Thus, the groupings of individuals into populations becomes a result of the analysis, as opposed to an *a priori* assumption set by the researcher.

This *a priori* grouping may lead to difficulties in analyses, for example when dealing with recent migrants, or when grouping sparsely sampled populations Shringarpure and Xing (2014). As ancient genetic sampling becomes more dense, recent migrants have become increasingly prevalent in data sets. These migrants are important, because they allow us tracking of movements – often over large distances – within the last handful of generations. Thus, these recent migrants have become an important focus of study and it is desirable to distinguish recent migration from the more distant-in-time migration that is ordinarily the focus of F-statistic analyses.

A PCA will reveal such recent migrants as outliers that group with their population-of-origin, whereas they would be missed by F-statistics.

In sparsely sampled populations, it is often necessary to group individuals that span relatively large areas or times to get to a usable sample size, particularly for regions where DNA preservation is poor. However, grouping genetically distinct individuals can be risky, as population substructure can invalidate the interpretation of F-statistic-tests Peter (2016). However, PCA provides a meaningful and simple way to assure that populations are devoid of relevant substructure: If individuals cluster tightly on *all* PCs that are meaningful for between-cluster variation, they can safely be grouped. However, if there is within-population variation that is correlated with other populations, then F-statistic-tests cannot be used to test for admixture.

As discussed above, these considerations hold particularly when using a PCA method that separates sampling noise from population structure, such PPCA or LSE. For classical PCA, clustering will also be impacted by sample quality and depth - which can be separated out effectively by the probabilistic methods.

Handling missing data Incomplete genotype calls with missing data are common in ancient DNA, and thus it is important that methods are able to deal with it. Classical PCA, which is based on SVD, relies on complete data, although methods that impute data are becoming more common (Meisner et al., 2021). For F-statistics, ADMIXTOOLS2 has a parameter `maxmiss` that controls how much missingness is permitted, with e.g.`maxmiss=0` removing all sites with missing data, and `maxmiss=1` retaining all sites.

For our tests with missingness, we could not use the option maxmiss=0, since this resulted in no sites available. Hence, we used `maxmiss=1` throughout, which resulted in inflated F-statistics in our simulations with50% missing genotype calls. In contrast, the PPCA-based algorithm is not affected, and appears to be more robust to missingness for both 8 and 12 PCs. It may seem surprising that ADMIXTOOLS 2 shows inflated estimates of F- statistics, given that it is based on Patterson's unbiased estimator.The reason for this inflation is that in the analysis for missing data we use one individual from each population, and when there is missing data, it results in only haploid genotype. In such cases, the correction term for the unbiased estimator is undefined. Thus we show that ,PPCA provides a practical solution in case of estimation of individual-based F-statistics with missing data.

Application on real data We used a PPCA-based framework with Neandertal data to estimate PCs utilizing all the samples. This approach is more straight-forward than the authors' approach to first estimate the PCs using high-coverage genomes, and then project the low-coverage genomes. We show that we can accurately estimate outgroup F_3 statistic from the PCs. We find that F_2 calculated using diploid genotype data with our framework is comparable to that of ADMIXTOOLS 2. However, ADMIXTOOLS 2 can not be used to estimate F_2 with pseudohaploid data since the unbiased estimator in ADMIXTOOLS 2 is undefined in this case. We show that our framework provides accurate estimates even with pseudohaploid samples. Here, we point out that the reason PPCA performs well in this case is because the scalar noise parameter ψ is determined from all the samples.

Limitations and future directions One limitation of this framework is the need to perform multiple PPCAs to obtain standard errors, which can be computationally expensive. Further studies are needed to design a statistical framework that can estimate the errors using SNP loadings, and therefore can work fast with large datasets. In case of PPCA, one needs to determine the number of PCs relevant for the particular analysis. We show that PPCA is not sensitive to small changes in the number of PCs, and we provide the option of LSE which does not require user-defined number of PCs. Finally, for LSE, an algotithm to deal with missing data is still pending.

Summary To summarize, we present a method to perform PCA and F-statistics jointly and show that this approach not only improves estimates of F-statistics, but also provides a solution to the standardization and quantification of PCA. Our framework is available on github as a snakemake pipeline: https://github.com/DivyaratanPopli/A-joint-framework-for-PCA-and-F-statistics.

6 Appendix

We provide the procedure to estimate LSE as laid out by Cabreros and Storey (2019). We define a symmetric matrix $\hat{\mathbf{H}}$:

$$\hat{\mathbf{H}} = \frac{1}{S}\mathbf{X}^T\mathbf{X} - \hat{\mathbf{D}}, \tag{12}$$

Here, $\mathbf{X}$ is the (uncentered) genotype matrix with shape $m \times S$, $\hat{\mathbf{D}}$ is a matrix of diagonal entries “correcting” heterozygosities and S is the number of SNPs.

In particular $d_{ij} = 0$ for $i \neq j$ and,

$$d_{ii} = \frac{1}{S}\sum_{k=1}^{S} x_{ik}(2 - x_{ik}), \tag{13}$$

for diploid data. The correction term given by Reich et al. (2009) and Patterson et al. (2012) is different by a factor of 4:

$$d'_{ii} = \frac{1}{S}\sum_{k=1}^{S} \frac{x_{ik}}{2}(1 - \frac{x_{ik}}{2}) = \frac{d_{ii}}{4}$$

The factor of 4 is due to the use of allele frequencies instead of genotypes, i.e. their approach would estimate $\hat{\mathbf{H}}$ (again for diploid data)

$$\hat{\mathbf{H}}' = \frac{1}{S}\frac{\mathbf{X}^T}{2}\frac{\mathbf{X}}{2} - \frac{\hat{\mathbf{D}}'}{4} = \frac{\hat{\mathbf{H}}}{4},$$

and so the parametrizations are equivalent, but will differ by a factor of four.

Thus,

$$h_{ii} = \frac{1}{S}\sum_{k=1}^{S} x_{ik}^2 - d_i \tag{14a}$$

$$h_{ij} = \frac{1}{S}\sum_{k=1}^{S} x_{ik}x_{jk} \tag{14b}$$

Consider now

$$\begin{aligned} f_{ij} &= h_{ii} + h_{jj} - 2h_{ij} \\ &= \frac{1}{S}\sum_{k=1}^{S} x_{ik}^2 - d_i + \frac{1}{S}\sum_{k=1}^{S} x_{jk}^2 - d_j - \frac{2}{S}\sum_{k=1}^{S} x_{ik}x_{jk} \\ &= \frac{1}{S}\sum_{k=1}^{m}(x_{ik} - x_{jk})^2 - d_i - d_j \qquad (15) \\ &= F_2(i,j) \end{aligned}$$

Hence the matrix $\hat{\mathbf{H}}$ can be used to estimate unbiased F_2-statistics.

7 removed stuff

How many PCs should be used? Our framework is able to deal with missing data illuminate us to point out that the conventional method to estimate F-statistics suffers from two major problems. First, one needs to assign individuals to discrete populations which may not be justified for humans, and is difficult to do with ancient samples. And second, the estimates are inaccurate in presence of high amounts of missing genotypes, another characteristic of ancient DNA. We present a PCA-based framework to estimate F-statistics while taking in account these issues. We compare the statistics estimated with different PCA approaches, and show that all approaches work well on ideal simulations with enough samples. In case of individual-based F-statistics, both PPCA and LSE outperform classical PCA, and we provide a PPCA algorithm for the case of missing data. We compare PPCA framework to ADMIXTOOLS 2, and find that our framework outperforms ADMIXTOOLS 2 to estimate individual-based F-statistics with missing data.

PCA is widely used in population genetics to visualize clusters of individuals that may represent populations, and clines potentially representing historical admixture. Moreover PCA may reveal fine-scale structure in the population Waldman et al. (2022). PCA's ability to condense complex genetic data into interpretable dimensions enhances our understanding of human evolution, migration, and admixture events, and throws light on the intricate mosaic of our species' history. However the visualization results from PCA may be influenced by the choice of PCs used, normalization, and the choice of populations used. Elhaik (2022). We provide a way of quantifying the results of a PCA with F-statistics so that all PCA's can be comparable and can use the same normalizations. Such a quantification using all the top PCs also makes it more straight-forward to justify a visual result made with 2 PCs.

8 appendix old

$$\hat{\mathbf{H}} = \frac{1}{m}\mathbf{X}^T\mathbf{X} - \hat{\mathbf{D}}, \tag{16}$$

Here, $\mathbf{X}$ is the (uncentered) genotype matrix and $\hat{\mathbf{D}}$ is a matrix of diagonal entries "correcting" heterozygosities and m is the number of SNPs.

In particular $d_{ij} = 0$ for $i \neq j$ and,

$$d_{ii} = d_i = \frac{1}{m}\sum_{k=1}^{m} x_{ik}(2 - x_{ik}), \tag{17}$$

for diploid data. The correction term given by Reich 2009 and Patterson 2012 is different by a factor of four:

$$d'_{ii} = \frac{1}{m}\sum_{k=1}^{m} \frac{x_{ik}(2 - x_{ik})}{4} = \frac{d_{ii}}{4}$$

. The difference is explained that they use allele frequencies instead of genotypes, i.e. their approach would use (again for diploid data)

$$\hat{\mathbf{H}'} = \frac{1}{m}\frac{\mathbf{X}^T}{2}\frac{\mathbf{X}}{2} - \frac{\hat{\mathbf{D}'}}{4} = \frac{\hat{\mathbf{H}}}{4},$$

and so the parametrizations are equivalent, but will differ by a factor of four.

Thus, (TODO: double-check if rows/columns are aligned, sum should be over SNP)

$$h_{ii} = \frac{1}{m}\sum_{k=1}^{m} x_{ik}^2 - d_i \tag{18a}$$

$$h_{ij} = \frac{1}{m}\sum_{k=1}^{m} x_{ik}x_{jk} \tag{18b}$$

Consider now

$$\begin{aligned} f_{ij} &= h_{ii} + h_{jj} - 2h_{ij} \\ &= \frac{1}{m}\sum_{k=1}^{m} x_{ik}^2 - d_i + \frac{1}{m}\sum_{k=1}^{m} x_{jk}^2 - d_j - \frac{2}{m}\sum_{k=1}^{m} x_{ik}x_{jk} \\ &= \frac{1}{m}\sum_{k=1}^{m}(x_{ik} - x_{jk})^2 - d_i - d_j \qquad (19) \\ &= F_2(i,j) \end{aligned}$$

Hence the matrix $\hat{\mathbf{H}}$ can be used to estimate F_2-statistics (possibly instead of PPCA?).

We describe two issues in accurate estimation of population allele frequencies:

1. Humans may not fit into well-differentiated discrete populations, except in cases where the populations have been isolated due to geographical features Novembre et al. (2008). The estimation of population allele frequencies depends on the assignment of individuals to discrete populations, and may be affected by miss-assignment especially when few samples are available.

2. Missing data in some individuals for certain sites can make it difficult to get accurate allele frequency estimates at those sites. One commonly used solution to this problem is to filter out all the sites with missing data. However, this may make the number of sites available for F-statistics quite small. E.g., for 100 individuals with 10% randomly missing sites, the available number of sites after filtering out positions with missing data would be ≈ 26 out of a total 1,000,000 sites.

Studies using PCA generally use specific PCs to visualize population structure or admixture, and the choice of the PCs used can be quite subjective Elhaik (2022). Estimation of F-statistics from PCA quantifies, using all the important PCs, what the researcher has visualized using seemingly arbitrary PCs.

However, a limitation with such a framework is that it is possible to define F-statistics in terms of PCs only when allele frequencies are known, and need not be estimated. This is due to the fact that PCA does not filter the noise in the data due to sampling. In addition, missing data can affect the computation of PCs, and subsequently F-statistics.

We would point out here that in ancient DNA studies PCA is sometimes used as quality-control step by constructing PCs using high-quality samples, and projecting the low-quality samples which may be from the same populations or even the same individuals as the high coverage samples. In the presence of contamination from present-day people, reference bias, ascertainment bias or batch effects, the projected sample may not overlap with an identical high-coverage sample. These biases and issues are not resolved with PPCA/LSE either, since PPCA only models sampling noise.

8.1 F-statistics with PPCA/LSE

In this study, we develop a statistical framework to estimate F-statistics between individuals in a PPCA framework. We show that PPCA explicitly models the error due to the sampling bias in allele frequencies. In addition, we demonstrate that PPCA based framework is not sensitive to random missing data, and so it can be used to visualize individuals in PCA-space without having to project lower quality samples.

We explain that PPCA provides a natural framework to estimate F-statistics with small sample-size and missing data. We show formal hypothesis tests for admixture and compare our results to admixtools2 Maier et al. (2022) on simulations. Finally we show the use of this framework on published datasets from neolithic Haak et al. and upper Paleolithic humans (Mateja).

Mathematically, the PPCA model can be represented as follows Tipping and Bishop (1999):

Latent variable model: Latent variables: $Z \sim N(0, I)$, where Z is a S-dimensional latent variable, and I is the identity matrix.

Latent-to-observed mapping: $X = WZ + \mu + \Psi$, where X is the observed data, W is a M x q matrix of linear mappings, μ is the mean of the observed data, and Ψ is a Gaussian noise term.

Prior distributions:

Prior on the latent variables: $P(Z) = N(0, I)$

Prior on the noise term: $P(\Psi) = N(0, \sigma^2 I)$, where σ^2 is the variance of the noise.

Likelihood function: $p(X|Z, W, \mu, \Psi) = N(X|WZ + \mu, \sigma^2 I)$

9 old stuff

9.1 PCA

One way to estimate F-statistics

the estimation of allele frequencies can also be affected by large amounts of missing data. Since F-statistics are dependent on the ascertainment scheme, random missing data reduces the number of overlapping sites greatly. PCA is a powerful method to visualize population structure but it can be difficult to interpret (**???**). In particular, the population structure in PCA is a function of expected pairwise coalescence times (McVean, 2009), and thus is not explicitly tied to a particular scenario; different histories may yield similar or identical

PCAs.

Thus, parameter estimation, model comparisions and formal tests of admixture are usually not carried out in PCA.

(add section here introducing the general idea of PCA, explain how it deals with error/noise and difference between probabilistic and regular PCA. Possibly also explain how people use PCA as a QC-step to detect batch effects, and how that compares with PCA used for population structure (i.e. Ainash' question from your talk))

F-statistics are a useful tool to quantify population structure, and provide tests for admixture. Hence, a common pipeline in many population genetic studies is to analyze PCA plots to look for visible patterns that could be due to past admixtures, followed by formal test with F-statistics Lazaridis et al. (2014, 2016). F-statistics use population allele frequencies to test for admixture, and hence the accuracy of these tests is limited by the accuracy of allele frequency estimates Peter (2016). Whereas F-statistics estimates are robust when there are enough high quality samples, we show that low number of samples and missing data decrease the accuracy of these tests.

A major issue when working with ancient DNA is low number of individuals and missing data. This issue is exacerbated due to difficulty in assigning some of the individuals to discrete populations. Especially in the case of humans samples, we know that the genetic samples do not strictly belong to discrete populations, but form a continuous spectrum in the allele frequency space Oteo-García and Oteo (2021). Hence, it is a step in the right direction to think of a method to estimate F-statistics in a structure-aware framework. The easiest way to think about this is using PCA. PCA does not require assignment of individuals to discrete populations, and it has been shown that F-statistics can be estimated conveniently from distance between populations on PCA space Peter (2022). However, PCA distances are inaccurate when population sizes are small since PCA does not explicitly model sampling bias in the allele frequencies (see section XX). In addition, PCA is sensitive to missing data, and this makes it difficult to work with ancient DNA.

In this study, we develop a statistical framework to estimate F-statistics between many populations in a probabilistic PCA framework. that probabilistic PCA (PPCA) explicitly models the error due to the sampling bias in allele frequencies. In addition. we demonstrate that PPCA based framework is not sensitive to random missing data, and so it can be used to visualize individuals in PCA-space without having to project lower quality samples.

Finally, we show that PPCA provides a natural framework to estimate F-statistics with small sample-size and missing data. We show formal hypothesis tests for admixture and compare our results to admixtools2 citeadmixtools paper on simulations. Finally we show the use of this framework on published datasets from neolithic Haak et al. and upper Paleolithic humans (Mateja).

10 older stuff

There are several tools available to understand genetic diversity, and can be classified as the tools that make minimal assumptions to summarize population structure, and the tools that infer demographic parameters. Former set of tools includes f-statistics Patterson et al. (2012), PCA ?, MDS Wang et al. (2009), Structure ?, Admixture - There are different tools to study diversity: 1) tools that make minimum assumptions like fstatistics, PCA, MDS, Structure, Admixture, 2) tools that can be used to infer demographic parameters. - Many studies use PCA followed by f-statistics in their pipline. Peter et al., show that these analyses reveal the same biological signal, and can be done jointly. - This framework has limitation: population allele frequencies are not known. - Here we present an approach based on pPCA, and show that it is a more natural framework since it can take in account the errors associated with allele frequency estimation.

10.1 Theory

- f-statistics, and the sampling error terms in f2. - PCA and relation to f-statstics. - pPCA/PCA1 (Waaij et. al.) as a framework for dimentionality reduction taking in account the sampling error.

10.2 results

- Advantage in estimating PCs: Fig.1: PCA plot with standard errors

- Advantage in estimating f-statistics: We compare f-statistics from pPCA, PCA1, PCA, admixtools2 in terms of accuracy (and speed?). Fig.2: point estimates of f2's for slendr simulations where we have both ancient and modern populations. Fig.3: Examples of f4 test of treeness with different cases of migrations, we can also show a case of f3 test of admixture.

- Application to a published dataset (Fig.4)

10.3 Discussion

- One key advantage of this framework is that both point estimates and standard errors for PCA and fstatistics are estimated together in a consistent way. - This is a step towards solving the issue with the assumption of discrete populations. - This would be quite useful for cases where assigning individuals to populations can be difficult. - Future work: 1) A faster way of estimating standard errors. 1) to get uncertainty from snp loading. 2) Hypothesis testing

- This approach could also help with missingness as shown by Meisner et al. - pPCA makes it easy to analyse modern and ancient data together without having to project samples.

10.4 Methods

- Simulation method and parameters used

11 Abstract

Studies of genetic variation now routinely include data from thousands of individuals representing complex historical and temporal structure. Understanding and modeling patterns of genetic variation between large numbers of individuals and populations is thus a key challenge in the usage of genetic data to answer questions about evolutionary history. Principle Component Analysis (PCA) and F-statistics sensu Patterson are both widely used for this purpose, but are usually analyzed dis-jointly. Here, we present a new framework based on probabilistic PCA to jointly estimate principal component and F-statistics from large panels of data. A key advantage of our approach is that we can calculate individual-based F-statistics efficiently, and so population assignments become a result rather than an a priori assumption. Furthermore, probabilistic PCA provides a natural framework for incorporating missing data, a common issue in ancient DNA analyses. Taken together, our results greatly simplify the analysis of large population genetic data sets, and allow for fast data exploration and statistical testing in a unified and consistent framework.

12 Background

12.1 Why study genetic diversity

The genetic diversity of human populations has been shaped by historical and environmental factors over hundreds of thousands of years. Therefore, a key objective of population genetics is to analyze the observable variations and patterns in order to understand and reconstruct the demographic and evolutionary history of our species.

12.2 The general pipeline

A general pipeline consists of a method to summarize the data with minimal assumptions. Examples are PCA, MDS, Structure, Admixture. These methods show a qualitative picture, but do not estimate a biologically meaningful parameter. And so, it is difficult to design statistical tests for the results of such methods. However, such qualitative results are generally followed by quantifiable methods based on f-statistics. F-statistics with 2,3, or 4 populations assumes a null-model as a tree-like structure and a deviation from the tree-like structure is represented the alternate model.

12.3 PCA and fstatistics

PCA is a method to rotate the dataset in a way that so that the axes that are analysed and plotted are aligned with the dimensions explaining the highest variation in the data. This is a way to do dimensionality reduction, and provides a way to do better data visualization. Ben[2020] showed that PCA and f-statistics are related, and

13 Introduction

13.1 Why combine pPCA and F-statistics?

In case of no missing data - Calculating f-statistics between populations already assumes clustering of individuals into populations. PCA shows this clustering, but it would be useful to quantify the distance between individuals to filter for individuals that cluster, spot outliers, and identify substructures.

-Faster calculation of f-statistics, when only few PCs are used. Calculation of pPCA may take some time, but afterwards f-statistics is less time taking. And since people anyway do both PCA and f-statistics, overall this would reduce time.

-admixture proportions from f-statistics using pca. We can check if it's more reliable.

In case of missing data - In case of admixturegraphs, missing data may reduce total snps (although, missing sites would be less if allele frequencies are calculated from available individuals in populations). pPCA may help in cases with e.g., few ancient individuals from each population with a lot of missing data.

- In case of ancient DNA, may help to include libraries with missing data instead of projecting them?

14 Practical application ideas

This is a list of ideas we could pursue. Goal would be to pick a few that would be easiest to accomplish.

1. Evaluate whether we use PPCA to calculate individual-based F-statistics accurately and fast?
    (a) in presence of missing data
    (b) for multivariate analyses including qpadm/qpgraph
    (c) include samples projected onto PCA
2. Grouping individuals in populations
3. Can we get confidence intervals on PCA?
    (a) Resampling SNPs
    (b) Resampling individuals
    (c) Calculate uncertainty based on SNP loadings
        i. For SVD $\mathbf{X} = (\mathbf{UD})\mathbf{V}^T = \mathbf{P}\mathbf{V}^T$, we might be able to use the correlation in the entries of $\mathbf{V}$ to estimate the "effective" number of SNPs
4. Practical programming
    (a) Write software tool that jointly computes individidual-based F-stats and PCA
5. Data analysis – find a good data set or scenario to analyze
    (a) Standard Western Eurasian PCA
    (b) Indian data
    (c) Some application that Stephan / Wolfgang are working on
6. Looking at qpadm/qpwave
    (a) qpadm projects samples into a subspace made from a subset of samples. As these samples are not orthogonal, this subspace is likely highly non-orthogonal and therefore tricky to work with. Doing PCA before doing qpadm could be helpful.

15 Projecting onto prob PCA

For PCA, we can write $X_{[n \times p]} = USV^T$ and the PCs are given by SV^T. , and we would project by

$$
\begin{aligned}
X &= USV^T \qquad (20)\\
U^TX &= SV^T \qquad (21)\\
XVS^{-1} &= USV^TVS^{-1} \qquad (22)\\
&= U \qquad (23)\\
U^TY &= Y_{proj} \qquad (24)\\
XVS^{-1}Y &= Y_{proj} \qquad (25)
\end{aligned}
$$

and we can project using U^TY for new data Y.

16 F-tests using Wishart-log-likelihoods

Consider the F_4-statistic

$$\begin{aligned} F_4(A-B, C-D) &= Cov(A-B, C-D) && (26)\\ &= Cov(A,C) + Cov(B,D) - Cov(A,D) - Cov(B,D) && (27)\\ & && (28) \end{aligned}$$

16.1 PCA to Covariance matrix

If we have a matrix of PCs, $\mathbb{P}$, then $\mathbf{Y} = \mathbb{P}\mathbb{P}^T$ is an estimate of the covariance matrix. Consider the random variables $(A-B)$ and $(C-D)$. If we assume they are jointly normally distributed, then their joint distribution will again be normally distributed with mean zero and covariance matrix

$$\mathbf{X} = \begin{pmatrix} y_{11} + y_{22} - 2y_{12} & y_{13} + y_{24} - y_{14} - y_{23} \\ y_{13} + y_{24} - y_{14} - y_{23} & y_{33} + y_{44} - 2y_{34} \end{pmatrix} \tag{29}$$

where the y are the entries of the covariance matrix obtained from the PCA. Practically, we can either first calculate Y (if we want all F-stats), or first subset $\mathbb{P}$ to the four pops involved.

The off-diagonal elements of $\mathbf{X}$ are precisely the F-statistics we aim to calculate. And thus we set up a test to see whether they are zero.

16.2 Derivation of the test statistic

The sampling distribution of a covariance follows a Wishart distribution. This is a $p \times p$ - matrix-valued probability distribution that is parametrized by a degree-of-freedom parameter n and a covariance matrix $\mathbf{S}$, also of dimension $p \times p$. The simplest way to generate Wishart random variates is,

$$\mathbf{X} = \sum_{i=1}^{n} Y_i Y_i^T \tag{30}$$

where $Y_i \sim N(0, \mathbf{S})$. We also have

$$\begin{aligned} E[\mathbf{X}] &= n\mathbf{S} && (31)\\ mode(\mathbf{X}) &= (n-p-1)\mathbf{S} && (32) \end{aligned}$$

The log-likelihood of a Wishart Distribution is

$$\begin{aligned}\log P(\mathbf{X}|\mathbf{S},n) &= -\frac{np}{2}\log(2) - \frac{n}{2}\log|\mathbf{S}| - \Gamma_p\left(\frac{n}{2}\right) + \frac{n-p-1}{2}\,|\mathbf{X}| - \frac{1}{2}\operatorname{tr}(\mathbf{S}^{-1}\mathbf{X}) && (33)\\ &\propto -\frac{n}{2}\log|\mathbf{S}| - \frac{1}{2}\operatorname{tr}(\mathbf{S}^{-1}\mathbf{X}) \\ &= -\frac{n}{2}\log\left(\sigma_{11}\sigma_{22}-\sigma_{12}^2\right) - \frac{1}{2}\frac{\sigma_{22}x_{11}+\sigma_{11}x_{22}-2\sigma_{12}x_{12}}{\sigma_{11}\sigma_{22}-\sigma_{12}^2} && (34)\end{aligned}$$

where the last step assumes a 2 × 2 matrix. Under H0, we have $\sigma_{12} = 0$, therefore

$$\log P(\mathbf{X}|\mathbf{S}_0,n) = -\frac{n}{2}\log\left(\sigma_{11}\sigma_{22}\right) - \frac{1}{2}\left[\frac{x_{11}}{\sigma_{11}} + \frac{x_{22}}{\sigma_{22}}\right] \tag{35}$$

This likelihood is easily separatable and we can estimate σ_{11} from x_{11} and σ_{22} from x_{22} directly.

The log-likelihood-ratio statistic can then be calculated as

$$\begin{aligned}R &= -2\log\left(\frac{P(\mathbf{X}|\mathbf{S}_0,n)}{P(\mathbf{X}|\mathbf{S},n)}\right)\\ &= 2[\log P(\mathbf{X}|\mathbf{S},n) - \log P(\mathbf{X}|\mathbf{S}_0,n)]\\ &= n\log\left(\sigma_{11}\sigma_{22}\right) - n\log\left(\sigma_{11}\sigma_{22}-\sigma_{12}^2\right)\\ &\quad + n\left[\frac{x_{11}}{\sigma_{11}} + \frac{x_{22}}{\sigma_{22}}\right] - n\frac{\sigma_{22}x_{11}+\sigma_{11}x_{22}-2\sigma_{12}x_{12}}{\sigma_{11}\sigma_{22}-\sigma_{12}^2} && (36)\end{aligned}$$

using the estimates $\sigma_{ij} = \frac{x_{ij}}{n}$ we get

$$\begin{aligned}R &\approx n\log\left(x_{11}x_{22}\right) - n\log\left(x_{11}x_{22}-x_{12}^2\right) + 2n - 2n\\ &= n\log\left(\frac{x_{11}x_{22}}{x_{11}x_{22}-x_{12}^2}\right) && (37)\end{aligned}$$

A $\log(n^2)$ term in each of the log-terms cancels. The statistic R is asymptotically χ^2 distributed with one degree of freedom.

If we instead use the mode $\sigma_{ij} \approx \frac{x_{ij}}{n-3}$:

$$\begin{aligned}R &\approx n\log\left(\frac{x_{11}x_{22}}{x_{11}x_{22}-x_{12}^2}\right) + 2(n-3)\frac{x_{11}x_{22}-x_{12}^2}{x_{11}x_{22}-x_{12}^2} - 2(n-3)\\ &= n\log\left(\frac{x_{11}x_{22}}{x_{11}x_{22}-x_{12}^2}\right) && (38)\end{aligned}$$

16.3 Comparison to Cavalli-Sforza & Piazza, 1975

The authors propose

$$R = \frac{|\mathbf{X}|}{|\mathbf{S}|} \tag{39}$$

where $|\cdot|$ is the determinant. For 2×2 matrices,

$$\begin{aligned}
|\mathbf{X}| &= x_{11}x_{22} - x_{12}^2 && (40)\\
|\mathbf{S}| &= x_{11}x_{22} && (41)\\
R &= \frac{x_{11}x_{22} - x_{12}^2}{x_{11}x_{22}} = 1 - \frac{x_{12}^2}{x_{11}x_{22}} && (42)\\
T &= -2n\log(R) = 2n\log\left(\frac{x_{11}x_{22}}{x_{11}x_{22} - x_{12}^2}\right) && (43)
\end{aligned}$$

The factor of 2 might be wrong..

16.4 other statistics

for x_{12} small, we may further approximate

$$\begin{aligned}
R &= n\log\left(\frac{x_{11}x_{22}}{x_{11}x_{22} - x_{12}^2}\right)\\
&= n\log(x_{11}x_{22}) - n\log(x_{11}x_{22} - x_{12}^2)\\
&= n\log(x_{11}x_{22}) - n\log\left[x_{11}x_{22}\left(1 - \frac{x_{12}^2}{x_{11}x_{22}}\right)\right]\\
&= n\log\left(1 - \frac{x_{12}^2}{x_{11}x_{22}}\right)\\
&\approx n\frac{x_{12}^2}{x_{11}x_{22}} && (44)
\end{aligned}$$

This is the coefficient of determination, which is the square of the correlation coefficient

$$r = \sqrt{R/n} = \frac{x_{12}}{\sqrt{x_{11}x_{22}}} \tag{45}$$

for which we have a t-distributed null

$$r\sqrt{\frac{n-2}{1-r^2}} \sim t(n) \tag{46}$$

The Fisher-Transform then yields

$$\frac{1}{2}\log\left(\frac{1+r}{1-r}\right) = \operatorname{arctan}(r) \sim N(0,1) \tag{47}$$

which simplifies to

$$\operatorname{arctan}(r) = \frac{1}{2}\log\left(\frac{\sqrt{x_{11}x_{22}}+x_{12}}{\sqrt{x_{11}x_{22}}-x_{12}}\right) \tag{48}$$

This statistic is normally distributed under the null-hypothesis.

in tests with a simple bivariate normal, all of them behaved equally well.

17 Calibrating the standard errors of F-stats

Traditionally, the standard errors of F-stats are estimated using a block-Jackknife approach. However, the block size is usually hard to estimate, and may impact the resulting values.

17.1 What do the standard errors measure?

There are two types of uncertainty:

- **sampling uncertainty**, that stems from the fact that we only have a small sample from each population. Because it depends on the sample, we expect these uncertainties to be independent for each sampled population
- **evolutionary uncertainty** there is also uncertainty due to the randomness in evolution. In particular, the realized mean allele frequencies in populations will be different from those expected under some model.

17.2 Covariance of F_2-statistics

$$\begin{aligned}
K &= Cov((X_1 - X_2)^2, (X_3 - X_4)^2) \\
&= Cov(X_1^2, X_3^2) + Cov(X_1^2, X_4^2) + Cov(X_2^2, X_3^2) + Cov(X_2^2, X_4^2) \\
&- 2\left[Cov(X_1^2, X_3X_4) + Cov(X_2^2, X_3X_4) + Cov(X_3^2, X_1X_2) + Cov(X_4^2, X_1X_2))\right] \\
&+ 4\left[Cov(X_1X_2, X_3X_4)\right]
\end{aligned} \tag{49}$$

At the same time, we have

$$\begin{aligned}
F_4^2 &= Cov(X_1 - X_2, X_3 - X_4)^2 \\
&= [Cov(X_1, X_3) + Cov(X_2, X_4) - Cov(X_1, X_4) - Cov(X_2, X_3)]^2 \\
&= Cov(X_1, X_3)^2 + Cov(X_2, X_4)^2 + Cov(X_1, X_4)^2 + Cov(X_2, X_3)^2 \\
&+ 2[Cov(X_1, X_3)Cov(X_2, X_4) + Cov(X_1, X_4)Cov(X_2, X_3)] \\
&- 2[Cov(X_1, X_3)Cov(X_1, X_4) + Cov(X_1, X_3)Cov(X_2, X_3)] \\
&- 2[Cov(X_2, X_4)Cov(X_1, X_4) + Cov(X_2, X_4)Cov(X_2, X_3)] \qquad (50) \\
&= (E[X_1^2] + E[X_2^2])(E[X_3^2] + E[X_4^2]) \\
&+ 2[E[X_1X_3]E[X_2X_4] + E[X_1X_4]E[X_2X_3]] \\
&- 2[E[X_1X_3]E[X_1X_4] + E[X_1X_3]E[X_2X_3]] \\
&- 2[EX_2X_4]E[X_2X_3] + E[X_2X_4]E[X_2X_3]] \qquad (51) \\
&= (E[X_1^2] + E[X_2^2])(E[X_3^2] + E[X_4^2]) \\
&+ 2[E[X_1X_3][E[X_2X_4] - E[X_1X_4]] + E[X_2X_3][E[X_1X_4] - E[X_1X_3]]] \\
&- 2[EX_2X_4]E[X_2X_3] + E[X_2X_4]E[X_2X_3]] \\
& \qquad (52)
\end{aligned}$$

But we have $Cov(A, B) = E[(A - E[A])(B - E[B])] = E[AB] - E[A]E[B]$ and

$$\begin{aligned}
Cov(A, B)Cov(C, D) &= E[(A - E[A])(B - E[B])]E[(C - E[C])(D - E[D])] \\
&= E[AB]E[CD] = E[ABCD] - Cov(AB, CD) \qquad (53)
\end{aligned}$$

18 Notes on PCA1 of van Waaij et al

van Waaij et al. suggest at PCA on a matrix of the following form for PCA https://doi.org/10.48550/arXiv.2302.04596 (their PCA1). This approach is further motivated by eq 7 in https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6707457/, which is a very technical (but probably useful) reference.

$$\hat{\mathbf{H}} = \frac{1}{m}\mathbf{X}^T\mathbf{X} - \hat{\mathbf{D}}, \tag{54}$$

where $\mathbf{X}$ is the (uncentered) genotype matrix and $\hat{\mathbf{D}}$ is a matrix of diagonal entries "correcting" heterozygosities and m is the number of SNPs.

In particular $d_{ij} = 0$ for $i \neq j$ and,

$$d_{ii} = d_i = \frac{1}{m}\sum_{k=1}^{m} x_{ik}(2 - x_{ik}), \tag{55}$$

for diploid data. The correction term given by Reich 2009 and Patterson 2012 is different by a factor of four:

$$d'_{ii} = \frac{1}{m}\sum_{k=1}^{m} \frac{x_{ik}(2 - x_{ik})}{4} = \frac{d_{ii}}{4}$$

. The difference is explained that they use allele frequencies instead of genotypes, i.e. their approach would use (again for diploid data)

$$\hat{\mathbf{H}}' = \frac{1}{m}\frac{\mathbf{X}^T}{2}\frac{\mathbf{X}}{2} - \frac{\hat{\mathbf{D}}'}{4} = \frac{\hat{\mathbf{H}}}{4},$$

and so the parametrizations are equivalent, but will differ by a factor of four.

Thus, (TODO: double-check if rows/columns are aligned, sum should be over SNP)

$$h_{ii} = \frac{1}{m}\sum_{k=1}^{m} x_{ik}^2 - d_i \tag{56a}$$

$$h_{ij} = \frac{1}{m}\sum_{k=1}^{m} x_{ik}x_{jk} \tag{56b}$$

Consider now

$$
\begin{aligned}
f_{ij} &= h_{ii} + h_{jj} - 2h_{ij} \\
&= \frac{1}{m}\sum_{k=1}^{m} x_{ik}^2 - d_i + \frac{1}{m}\sum_{k=1}^{m} x_{jk}^2 - d_j - \frac{2}{m}\sum_{k=1}^{m} x_{ik}x_{jk} \\
&= \frac{1}{m}\sum_{k=1}^{m}(x_{ik} - x_{jk})^2 - d_i - d_j \qquad (57) \\
&= F_2(i,j)
\end{aligned}
$$

Hence the matrix $\hat{\mathbf{H}}$ can be used to estimate F_2-statistics (possibly instead of PPCA?). The detailed justification of this can be found in this statsexchange post that will need to be adapted.

Thus, this might be a useful alternative to PPCA to calculate F-statistics:

1. Calculate $\hat{\mathbf{H}} = \frac{1}{m}\mathbf{X}^T\mathbf{X} - \hat{\mathbf{D}}$
2. Double-Center $\hat{\mathbf{H}}$: $\mathbf{H}_c = \mathbf{C}\hat{\mathbf{H}}\mathbf{C}$, where $\mathbf{C}$ is a centering matrix
3. Obtain PCs using an eigendecomposition of $\mathbf{H}_c$: $\mathbf{P}\mathbf{P}^T = \mathbf{H}_c$
4. Calculate F_2 from the smaller space $\mathbf{P}$

19 Standard errors and effective number of SNPs

An issue in calculating standard errors for F-statistics is that SNP are usually correlated, and so standard variance and standard error calculations will fail.

let us assume n populations can be represented by some population structure model that is parameterized by some covariance matrix $\mathbf{X}$. We do not observe SNPs, but rather we have a noisy sample $\mathbf{G}_{[S \times n]}$ at S loci.

Let, as above, denote the data matrix as $\mathbf{G}$, which is a noisy version of an allele frequency matrix $\mathbf{X}$, and we assume we can do a PCA on some estimate of $\hat{\mathbf{X}}$ as $\mathbf{L}\mathbf{Z} = \hat{\mathbf{X}}$ where $\mathbf{L}$ are the orthonormal SNP-loadings and $\mathbf{Z}$ are the PCs. For example, we could do that using probabilistic PCA or the Cabreros-Storey-PCA

We are interested in statistics of the form

$$F_{ij} = (X_i - X_j)^2 \qquad (58)$$

which can be estimated from $\mathbf{G}$ using the unbiased estimator of Patterson et al.

$$f_{ij} = \frac{1}{S}\sum_{s=1}^{n}(g_{si} - g_{sj})^2 - H_i - H_j. \tag{59}$$

alternatively, we can also obtain an estimate from the decomposition of $\hat{\mathbf{X}}$ as

$$p_{ij} = \frac{1}{S}\sum_{p=1}^{n}(z_{si} - z_{sj})^2, \tag{60}$$

this sum is over PCs. This estimator has thus the advantage that it can be computed a lot faster since typically $n \ll S$, and if X is low-rank we can even truncate the sum

An issue is the calculation of the standard error of f_{ij} and p_{ij}. A simple estimator is

$$\sigma_p = \sqrt{\frac{\mathrm{Var}(p_{ij})}{S}} \tag{61}$$

a problem with σ_p is that SNPs are not independent, and so the variance estimates are underestimated. For this purpose, a block-jackknife can be used.

We block data using a vector b, s.t $b_i = j$ means that the i-th SNP is in block j. For each block, we then have the pseudovalue

$$\tilde{f}^{(j)} = \frac{1}{m_j}\sum_s I[b_s = j](x_{si} - x_{sj})^2$$

where $I[\cdot]$ is an indicator and m_j is the number of entries in block j.

then

$$\sigma'_f = \sqrt{\frac{1}{g}\sum_j\left[\frac{m_j}{n - m_j}(\tilde{f}^{(j)} - f)^2\right]} \tag{62}$$

This motivates the effective number of SNPs,

$$S_e = S\left(\frac{\sigma_f}{\sigma'_f}\right)^2 \tag{63}$$

which gives the number of pseudo-independent observations.

simplifying block-JK Writing

$$f = \frac{1}{n}\sum_s (x_{si} - x_{sj})^2$$

References

Agrawal, A., A. M. Chiu, M. Le, E. Halperin, and S. Sankararaman. 2020. Scalable probabilistic PCA for large-scale genetic variation data. PLOS Genetics 16:e1008773. URL https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1008773. Publisher: Public Library of Science.

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

Cabreros, I., and J. D. Storey. 2019. A Likelihood-Free Estimator of Population Structure Bridging Admixture Models and Principal Components Analysis. Genetics 212:1009–1029.

Cavalli-sforza, L. L., P. Menozzi, and A. Piazza. 1996. The History and Geography of Human Genes. URL https://press.princeton.edu/books/ebook/9780691187266/the-history-and-geography-of-human-genes.

Cavalli-Sforza, L. L., and A. Piazza. 1975. Analysis of evolution: evolutionary rates, independence and treeness. Theoretical Population Biology 8:127–165.

Chen, X., and J. D. Storey. 2015. Consistent Estimation of Low-Dimensional Latent Structure in High-Dimensional Data. URL http://arxiv.org/abs/1510.03497. ArXiv:1510.03497 [stat].

DeGiorgio, M., and N. A. Rosenberg. 2013. Geographic Sampling Scheme as a Determinant of the Major Axis of Genetic Variation in Principal Components Analysis. Molecular Biology and Evolution 30:480–488. URL https://doi.org/10.1093/molbev/mss233.

Elhaik, E. 2022. Principal Component Analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated. Scientific Reports 12:14683.

Engelhardt, B. E., and M. Stephens. 2010. Analysis of Population Structure: A Unifying Framework and Novel Methods Based on Sparse Factor Analysis. PLOS Genetics 6:e1001117. URL https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1001117. Publisher: Public Library of Science.

Excoffier, L., and M. Foll. 2011. fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. Bioinformatics 27:1332–1334. URL https://doi.org/10.1093/bioinformatics/btr124.

Felsenstein, J. 1973. Maximum-likelihood estimation of evolutionary trees from continuous characters. American Journal of Human Genetics 25:471–492. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1762641/.

Gopalan, P., W. Hao, D. M. Blei, and J. D. Storey. 2016. Scaling probabilistic models of genetic variation to millions of humans. Nature Genetics 48:1587–1590. URL https://www.nature.com/articles/ng.3710. Number: 12 Publisher: Nature Publishing Group.

Gower, J. C. 1966. Some Distance Properties of Latent Root and Vector Methods Used in Multivariate Analysis. Biometrika 53:325–338. URL https://www.jstor.org/stable/2333639. Publisher: [Oxford University Press, Biometrika Trust].

Gutenkunst, R. N., R. D. Hernandez, S. H. Williamson, and C. D. Bustamante. 2009. Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. PLOS Genetics 5:e1000695. URL https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000695. Publisher: Public Library of Science.

Haak, W., I. Lazaridis, N. Patterson, et al. 2015. Massive migration from the steppe was a source for Indo-European languages in Europe. Nature 522:207–211. URL http://www.nature.com/articles/nature14317.

Hajdinjak, M., Q. Fu, A. Hübner, et al. 2018. Reconstructing the genetic history of late Neanderthals. Nature 555:652–656. URL https://www.nature.com/articles/nature26151. Number: 7698 Publisher: Nature Publishing Group.

Hellenthal, G., G. B. Busby, G. Band, et al. 2014. A genetic atlas of human admixture history. Science (New York, N.Y.) 343:747–751. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4209567/.

Jay, F., P. Sjödin, M. Jakobsson, and M. G. Blum. 2013. Anisotropic Isolation by Distance: The Main Orientations of Human Genetic Differentiation. Molecular Biology and Evolution 30:513–525. URL https://doi.org/10.1093/molbev/mss259.

Kamm, J., J. Terhorst, R. Durbin, and Y. S. Song. 2020. Efficiently Inferring the Demographic History of Many Populations With Allele Count Data. Journal of the American Statistical Association 115:1472–1487. URL https://doi.org/10.1080/01621459.2019.1635482. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2019.1635482.

Kelleher, J., A. M. Etheridge, and G. McVean. 2016. Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. PLOS Computational Biology 12:e1004842. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004842.

Lawson, D. J., G. Hellenthal, S. Myers, and D. Falush. 2012. Inference of Population Structure using Dense Haplotype Data. PLoS Genetics 8:e1002453. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3266881/.

Lazaridis, I., D. Nadel, G. Rollefson, et al. 2016. Genomic insights into the origin of farming in the ancient Near East. Nature 536:419–424. URL https://www.nature.com/articles/nature19310. Number: 7617 Publisher: Nature Publishing Group.

Lazaridis, I., N. Patterson, A. Mittnik, et al. 2014. Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature 513:409–413. URL https://www.nature.com/articles/nature13673. Number: 7518 Publisher: Nature Publishing Group.

Lewontin, R. C. 1972. The Apportionment of Human Diversity. T. Dobzhansky, M. K. Hecht, and W. C. Steere, eds., Evolutionary Biology: Volume 6. Springer US, New York, NY, 381–398. URL https://doi.org/10.1007/978-1-4684-9063-3_14.

Maier, R., P. Flegontov, O. Flegontova, P. Changmai, and D. Reich. 2022. On the limits of fitting complex models of population history to genetic data. URL https://www.biorxiv.org/content/10.1101/2022.05.08.491072v2. Pages: 2022.05.08.491072 Section: New Results.

Mathieson, I., and A. Scally. 2020. What is ancestry? PLOS Genetics 16:e1008624. URL https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1008624. Publisher: Public Library of Science.

McVean, G. 2009. A Genealogical Interpretation of Principal Components Analysis. PLoS Genetics 5:e1000686. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2757795/.

Meisner, J., S. Liu, M. Huang, and A. Albrechtsen. 2021. Large-scale inference of population structure in presence of missingness using PCA. Bioinformatics (Oxford, England) 37:1868–1875.

Menozzi, P., A. Piazza, and L. Cavalli-Sforza. 1978. Synthetic Maps of Human Gene Frequencies in Europeans. Science 201:786–792. URL https://www.science.org/doi/10.1126/science.356262. Publisher: American Association for the Advancement of Science.

Novembre, J. 2022. The background and legacy of Lewontin's apportionment of human genetic diversity. Philosophical Transactions of the Royal Society B: Biological Sciences 377:20200406. URL https://royalsocietypublishing.org/doi/10.1098/rstb.2020.0406. Publisher: Royal Society.

Novembre, J., T. Johnson, K. Bryc, et al. 2008. Genes mirror geography within Europe. Nature 456:98–101. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2735096/.

Novembre, J., and B. M. Peter. 2016. Recent advances in the study of fine-scale population structure in humans. Current Opinion in Genetics & Development 41:98–105. URL https://linkinghub.elsevier.com/retrieve/pii/S0959437X16301113.

Novembre, J., and M. Stephens. 2008. Interpreting principal component analyses of spatial population genetic variation. Nature Genetics 40:646–649. URL https://www.nature.com/articles/ng.139. Number: 5 Publisher: Nature Publishing Group.

Orlando, L., R. Allaby, P. Skoglund, et al. 2021. Ancient DNA analysis. Nature Reviews Methods Primers 1:1–26. URL https://www.nature.com/articles/s43586-020-00011-0. Number: 1 Publisher: Nature Publishing Group.

Oteo-García, G., and J.-A. Oteo. 2021. A Geometrical Framework for f-Statistics. Bulletin of Mathematical Biology 83:14. URL https://doi.org/10.1007/s11538-020-00850-8.

Patterson, N. A modification to the jackknife to deal with adjacent blocks .

Patterson, N., P. Moorjani, Y. Luo, et al. 2012. Ancient Admixture in Human History. Genetics 192:1065–1093. URL https://doi.org/10.1534/genetics.112.145037.

Patterson, N., A. L. Price, and D. Reich. 2006. Population Structure and Eigenanalysis. PLOS Genetics 2:e190. URL https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.0020190. Publisher: Public Library of Science.

Peter, B. M. 2016. Admixture, Population Structure, and F-Statistics. Genetics 202:1485–1501. URL https://www.genetics.org/content/202/4/1485.

Peter, B. M. 2022. A geometric relationship of F2, F3 and F4-statistics with principal component analysis. Philosophical Transactions of the Royal Society B: Biological Sciences 377:20200413. URL https://royalsocietypublishing.org/doi/full/10.1098/rstb.2020.0413. Publisher: Royal Society.

Peter, B. M., D. Petkova, and J. Novembre. 2020. Genetic Landscapes Reveal How Human Genetic Diversity Aligns with Geography. Molecular Biology and Evolution 37:943–951. URL https://doi.org/10.1093/molbev/msz280.

Petr, M., B. C. Haller, P. L. Ralph, and F. Racimo. 2022. slendr: a framework for spatio-temporal population genomic simulations on geographic landscapes. URL https://www.biorxiv.org/content/10.1101/2022.03.20.485041v2. Pages: 2022.03.20.485041 Section: New Results.

Popejoy, A. B., and S. M. Fullerton. 2016. Genomics is failing on diversity. Nature 538:161–164. URL https://www.nature.com/articles/538161a. Number: 7624 Publisher: Nature Publishing Group.

Price, A. L., N. J. Patterson, R. M. Plenge, et al. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics 38:904–909. URL https://www.nature.com/articles/ng1847. Number: 8 Publisher: Nature Publishing Group.

Pritchard, J. K., M. Stephens, and P. Donnelly. 2000. Inference of Population Structure Using Multilocus Genotype Data. Genetics 155:945–959. URL https://academic.oup.com/genetics/article/155/2/945/6048111.

Reich, D., K. Thangaraj, N. Patterson, A. L. Price, and L. Singh. 2009. Reconstructing Indian Population History. Nature 461:489–494. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2842210/.

Rosenberg, N. A., S. Mahajan, S. Ramachandran, et al. 2005. Clines, Clusters, and the Effect of Study Design on the Inference of Human Population Structure. PLOS Genetics 1:e70. URL https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.0010070. Publisher: Public Library of Science.

Schraiber, J. G., and J. M. Akey. 2015. Methods and models for unravelling human evolutionary history. Nature Reviews Genetics 16:727–740. URL https://www.nature.com/articles/nrg4005. Number: 12 Publisher: Nature Publishing Group.

Serre, D., and S. Pääbo. 2004. Evidence for gradients of human genetic diversity within and among continents. Genome Research 14:1679–1685.

Sforza, L. L. C., and F. C. Sforza. 1995. The Great Human Diasporas: A History Of Diversity And Evolution. Basic Books. Google-Books-ID: JWV2tAEACAAJ.

Shringarpure, S., and E. P. Xing. 2014. Effects of Sample Selection Bias on the Accuracy of Population Structure and Ancestry Inference. G3: Genes|Genomes|Genetics 4:901–911. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4025489/`.

Simon, A., and G. Coop. 2023. The contribution of gene flow, selection, and genetic drift to five thousand years of human allele frequency change. bioRxiv: The Preprint Server for Biology :2023.07.11.548607.

Speidel, L., M. Forest, S. Shi, and S. R. Myers. 2019. A method for genome-wide genealogy estimation for thousands of samples. Nature Genetics 51:1321–1329. URL `https://www.nature.com/articles/s41588-019-0484-x`. Number: 9 Publisher: Nature Publishing Group.

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

Tipping, M. E., and C. M. Bishop. 1999. Probabilistic Principal Component Analysis. Journal of the Royal Statistical Society Series B: Statistical Methodology 61:611–622. URL `https://doi.org/10.1111/1467-9868.00196`.

van Waaij, J., S. Li, G. Garcia-Erill, A. Albrechtsen, and C. Wiuf. 2023. Evaluation of population structure inferred by principal component analysis or the admixture model. URL `http://arxiv.org/abs/2302.04596`. ArXiv:2302.04596 [stat].

Waldman, S., D. Backenroth, Harney, et al. 2022. Genome-wide data from medieval German Jews show that the Ashkenazi founder event pre-dated the 14th century. Cell 185:4703–4716.e16. URL `https://www.sciencedirect.com/science/article/pii/S0092867422013782`.

Wang, C., X. Zhan, L. Liang, G. R. Abecasis, and X. Lin. 2015. Improved ancestry estimation for both genotyping and sequencing data using projection procrustes analysis and genotype imputation. American Journal of Human Genetics 96:926–937.

Wang, D., Y. Sun, P. Stang, et al. 2009. Comparison of methods for correcting population stratification in a genome-wide association study of rheumatoid arthritis: principal-component analysis versus multidimensional scaling. BMC Proceedings 3:S109. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2795880/`.