Modelling complex population structure using F-statistics and Principal Component Analysis

Benjamin M Peter

July 5, 2021

Abstract

The genetic diversity in human population is shaped by our complex history. F-statistics are a widely used tool to model genetic diversity when populations are related acording to a graph and gene flow is rare. Here, I use a geometric interpretation of to interpret F-statistic in the context of Principal Component Analysis (PCA), showing that the two approaches are closely related. I show that orthogonal projections are a key concept to establish this link, and that it leads to simple geometric interpretations of F-statistics on a PCA-plot. I illustrate my results on two examples, one of local, and one of global human diversity. In both examples, I find that population structure is sparse, and only a few components contribute to most statistics, and I develop novel visualizations that allow for checking the assumptions of more sophisticated models. My results allows interpreting F-statistics in arbitrarily structured populations, to obtain more complete and less biased models of human genetic variation.

1 Introduction

About 15% of genetic variation in humans can be explained by population structure (Lewontin, 1972, Barbujani et al., 1997, Rosenberg et al., 2002), but the information contained in these 15% is sufficient to study our genetic diversity and history in great detail (Cavalli-Sforza et al., 1994, Reich, 2018). For some data sets it is possible to predict an individuals origin at a resolution of a few hundred kilometers (Novembre et al., 2008, Leslie et al., 2015), and direct-to-consumer-genetics companies are using this variation to analyze the genetic data of millions of customers.

Human genetic diversity is caused by a number of forces, and has both discrete and continuous components (Rosenberg et al., 2002, Serre and Pääbo, 2004, Rosenberg et al., 2005, Bradburd et al., 2018, Reich, 2018). Spatially isolated populations are expected to slowly diverge, particularly if they are separated by barriers to migration such as mountain ranges, oceans or deserts (Bradburd et al., 2013, Peter et al., 2020, Rosenberg et al., 2005). On the other hand, major population movements such as the out-of-Africa, Austronesian or Bantu expansions lead to more gradual patterns of genetic diversity (Cavalli-Sforza et al., 1994, Ramachandran et al., 2005, Novembre et al., 2008, Peter et al., 2020, Stoneking, 2016, Racimo et al., 2020). Local migration between neighboring populations will flatten differentiation, and long-distance migrations (Alves et al., 2016) or secondary contact and admixture between diverged populations, such as Neandertals and early modern humans (Green et al., 2010) adds further complexity.

This complex population structure is frequently handled by using multiple models with different assumptions; each emphasizing a particular aspects of the data. Data-driven methods such as Principal Component Analysis (PCA) (Cavalli-Sforza et al., 1994) structure (Pritchard et al., 2000) or multidimensional scaling (MDS, Malaspinas et al. (2014)) are often used to display the full complexity of the data, but they have the disadvantage that they are not easily interpretable. For this purpose, more explicit demographic models (Gutenkunst et al., 2009, Kamm et al., 2015, Excoffier et al., 2013) are applied, which allow for parameter estimation or hypothesis tests.

Particularly in the analysis of human ancient DNA, a set of techniques based on F-statistics *sensu* Patterson have risen in popularity (Patterson et al., 2012, Peter, 2016). This framework is based on the assumption that the relationship between three or four populations is often tree-like, and allows for a variety of tests of treeness and more complex demographic models (Patterson et al., 2012, Harney et al., 2021). However, the connections between PCA, F-statistics and demographic models are currently unclear, which makes quantitative comparisons, detecting model violations and joint interpretation of the results of these approaches difficult. Since both F-statistics and PCA are functions of expected pairwise coalescent times (McVean, 2009, Peter, 2016), this is one avenue to link these approaches. Here, I instead use the geometric interpretation of F-statistics introduced by Oteo-Garcia and Oteo (2021) to directly visualize F-statsitics on a PCA plot.

2 Theory

In this section, I will give a very brief formal introduction to F-statistics and PCA. A more detailed technical treatise of PCA is given in e.g. Jolliffe (2013), and a useful guide to interpretation is Cavalli-Sforza et al. (1994). Readers unfamiliar with F-statistics may find Patterson et al. (2012), Peter et al. (2020) or Oteo-Garcia and Oteo (2021) helpful.

2.1 Introduction to PCA

Let us assume we have some genotype data summarized in a matrix $\mathbf{X}$ whose entry x_{ij} reflects the allele frequency of the i-th population at the j-th genotype. If we have S SNPs and n populations, $\mathbf{X}$ will have dimension $n \times S$. As a population may be represented by just one individual, there is no conceptual difference between these cases and I will refer to populations as unit for analysis. Since the allele frequencies are between zero and one, we can interpret each Population X_i of $\mathbf{X}$ as a point in $[0,1]^S$, the allele frequency or *data space*.

The goal of PCA is to find a low-dimensional subspace $\mathbb{R}^K$ that explains most of the variation in the data. K is at most $n-1$, in which case the data is simply rotated. However, the historical processes that generated genetic variation often result in *sparse* data (Engelhardt and Stephens, 2010), so that $K \ll n$ explains a substantial portion of the variation; for visualization $K = 2$ is frequently used.

There are several algorithms that are used to perform PCAs, the most common one is based on singular value decomposition (Jolliffe, 2013). In this approach, we first mean-center $\mathbf{X}$, obtaining a centered matrix $\mathbf{Y}$

$$y_{il} = x_{il} - \mu_l$$

where μ_l is the mean allele frequency at the l-th locus.

PCA can then be written as

$$\mathbf{Y} = \mathbf{C}\mathbf{X} = (\mathbf{U}\mathbf{\Sigma})\mathbf{V}^T = \mathbf{P}\mathbf{L}, \tag{1}$$

where $\mathbf{C} = \mathbf{I} - \frac{1}{n}\mathbf{1}$ is a centering matrix that subtracts row means, with $\mathbf{I}, \mathbf{1}$ the identity matrix and a matrix of ones, respectively. The orthogonal matrix of principal components $\mathbf{P} = \mathbf{U}\mathbf{\Sigma}$ has size $n \times n$ and is used to reveal population structure. The SNP loadings $\mathbf{L} = \mathbf{V}^T$ are an orthonormal matrix of size $n \times k$, its rows give the contribution of each SNP to each PC, it is often useful to look for outliers that might be indicative of selection (e.g François et al., 2010).

In many implementations (e.g. Patterson et al., 2006), SNPs are weighted by the inverse of their standard deviation. As this weighting often makes little difference in practice (McVean, 2009), I will assume throughout that SNPs are unweighted.

2.2 Introduction to F-statistics

PCA is typically used to model population structure between many populations. F-statistics take the opposite approach, revealing the relationship between just two, three or four populations at a time. The three F-statistics can be defined as

$$F_2(X_1, X_2) \times S = \sum_{l=1}^{S} (x_{1l} - x_{2l})^2 = \|X_1, X_2\|^2 \tag{2a}$$

$$F_3(X_1; X_2, X_3) \times S = \sum_{l=1}^{S} (x_{1l} - x_{2l})(x_{1l} - x_{3l}) = \langle X_1 - X_2, X_1 - X_3 \rangle \tag{2b}$$

$$F_4(X_1, X_2; X_3, X_4) \times S = \sum_{l=1}^{S} (x_{1l} - x_{2l})(x_{3l} - x_{4l}) = \langle X_1 - X_2, X_3 - X_4 \rangle, \tag{2c}$$

where $\|\cdot\|$ denotes the Euclidean norm and $\langle \cdot, \cdot \rangle$ denotes the dot product. The number of SNPs S is assumed to be the same for all calculations and is thus omitted subsequently. Furthermore, both F_3 and F_4 can be written as sums of F_2-statistics:

$$2F_3(X_1; X_2, X_3) = 2F_2(X_2, X_3) - F_2(X_1, X_2) + F_2(X_1, X_3) \tag{3a}$$

$$2F_4(X_1, X_2; X_3, X_4) = F_2(X_1, X_3) + F_2(X_2, X_4) - F_2(X_1, X_4) - F_2(X_2, X_3) \tag{3b}$$

F-statistics have been primarily motivated in the context of trees and admixture graphs (Patterson et al., 2012). In a tree, the squared Euclidean distance $F_2(X, Y)$ measures the length of all branches between populations X and Y; and F_3 and F_4 represent external and internal branches in a tree, respectively (Peter, 2016). The length is a measure of genetic drift, and is non-negative if data is generated under a tree (Patterson et al., 2012). This interpretation is useful to understand a number of applications. The outgroup-F_3-statistic $F_3(O; U, X_i)$, for example, is useful if we have an unknown population U, and want to find its closest relatives from a panel of populations X_i. The highest values of F_3 indicate the population X_i most closely related to U, using the outgroup O to correct for differences in sample times. The population X_i with the largest value is the most closely related population out of the reference sample. The internal branches described by F_4-statistics are frequently used for complex models, such as reconstructing admixture graphs (Patterson et al., 2012, Lipson et al., 2013) and estimating admixture proportions (Petr et al., 2019, Harney et al., 2021).

Most commonly however, F_3 and F_4 are used as admixture tests (Patterson et al., 2012): Negative values of $F_3(X_1; X_2, X_3) < 0$ correspond to a branch with negative genetic drift, which is a violation of treeness. Similarly if four populations are related as a tree, then at least one of the F_4 statistics between the populations will be zero (Patterson et al., 2012).

To move away from trees and graph models, I build upon the geometric framework of Oteo-Garcia and Oteo (2021). Here, we think of each population as a point in the data space $\mathbb{R}^S$, made up of the allele frequency at each SNP. Then, $F_2(X_1, X_2) = \|X_1 - X_2\|^2$ is the squared Euclidean distance between two populations X_1 and X_2, and $F_4(X_1, X_2; X_3, X_4) = \langle X_1 - X_2, X_3 - X_4 \rangle$ is the inner (dot) product between these two vectors. These dot products are useful for a variety of projections that use population structure.

2.3 Connection between PCA and F-statsitics

2.3.1 Principal components from F-statistics

PCA and F-statistics are closely related. In fact, the principal components can be directly calculated from F-statistics using multidimensional scaling (Gower, 1966). Suppose we calculate the pairwise $F_2(X_i, X_j)$ between all n populations, and collect them in a matrix $\mathbf{F}_2$. We can obtain the principal components from this matrix by double-centering it, so that its row and column means are zero, and perform an eigendecomposition of the resulting matrix:

$$\mathbf{P}\mathbf{P}^T = -\frac{1}{2}\mathbf{C}\mathbf{F}_2\mathbf{C}. \tag{4}$$

2.3.2 F-statistics in PCA-space

By performing a PCA, we rotate our data to reveal the axes of highest variation. However, the dot product is invariant under rotation, and F-statistics can be thought of as dot products. What this means is that we are free to calculate F_2 either on the uncentered data $\mathbf{X}$, the centered data $\mathbf{Y}$ or any other orthogonal basis such as the principal components $\mathbf{P}$. Formally,

$$\begin{aligned} F_2(X_i, X_j) = && \sum_{l=1}^{L} \left(x_{il} - x_{jl}\right)^2 && \\ = & \sum_{l=1}^{L} \left((x_{il} - \mu_l) - (x_{jl} - \mu_l)\right)^2 = && & F_2(Y_i, Y_j) \\ = && \sum_{k=1}^{n} (p_{ik} - p_{jk})^2 = && F_2(P_i, P_j), \end{aligned} \tag{5}$$

A detailed derivation of this is given in Appendix A. As F_3 and F_4 can be written as sums of F_2-terms, analogous relations apply.

In most applications, we do not use all PCs, but instead use only the first K PCs. Thus,

$$F_2(P_i, P_j) = \underbrace{\sum_{k=1}^{K} (p_{ik} - p_{jk})^2}_{\hat{F_2}^{(K)}(P_i, P_j)} + \sum_{k=K+1}^{n} (p_{ik} - p_{jk})^2. \tag{6}$$

If we sum up the approximation errors $F_2 - \hat{F}_2$ over all pairs of populations, we obtain the Frobenius-norm of the error $\left\|\mathbf{F}_2 - \hat{\mathbf{F}}_2\right\|_F^2$; which is precisely the function that is minimized in MDS (Jolliffe, 2013). Thus, $\hat{\mathbf{F}}_2$ is the optimal approximation of $\mathbf{F}_2$ for any K.

3 Material & Methods

The theory outlined in the previous section suggests that F-statistics have a geometric interpretation on PCA plots, and can be approximated using PCA. I will explore this in detail in the next section, and illustrate it on two sample data sets that I briefly introduce here. Both are based on the analyses by Lazaridis et al. (2014) and use the “Human Origins”-SNP set (597,573 SNPs), based on the subset of the Reich lab compendium data set (v44.3), downloaded from `https://reich.hms.harvard.edu/allen-ancient-dna-resource-aadr-downloadable-genotypes-present-day-and-ancient-dna-data`. The code used to create all figures and analyses will be available on [Archive to be created before publishing].

World Overview data set This data set represents global human genetic variation (638 individuals from 33 population), as used by (Lazaridis et al., 2014). As this data set is very sparse, it may be well-approximated by an admixture graph.

West-Eurasian data set This data set of 1,119 individuals from 62 populations contains present-day individuals from the Eastern Mediterranean, Caucasus and Europe. It is frequently used as a basis of comparison for ancient genetic analyses of Western Eurasian individuals (Patterson et al., 2012, Lazaridis et al., 2014). Genetic differentiation in this region is low and closely mirrors geography (Novembre et al., 2008), and thus not particularly graph-like.

Computation F-statistics and PCA I perform analyses at the level of populations to ease presentation. It is an assumption of F-statistics that the genetic variation within sampled population is independent of the variation between samples. All computations are performed in `R`. I use `admixtools 2.0.0 https://github.com/uqrmaie1/admixtools` to compute F-statistics. To obtain a PC-decomposition, I first calculate all pairwise F_2-statistics, obtain a nearby negative semidefinite matrix using the `nearPD` function, and then use equation 4 and the `eigen` function to obtain the PCs.

4 Results

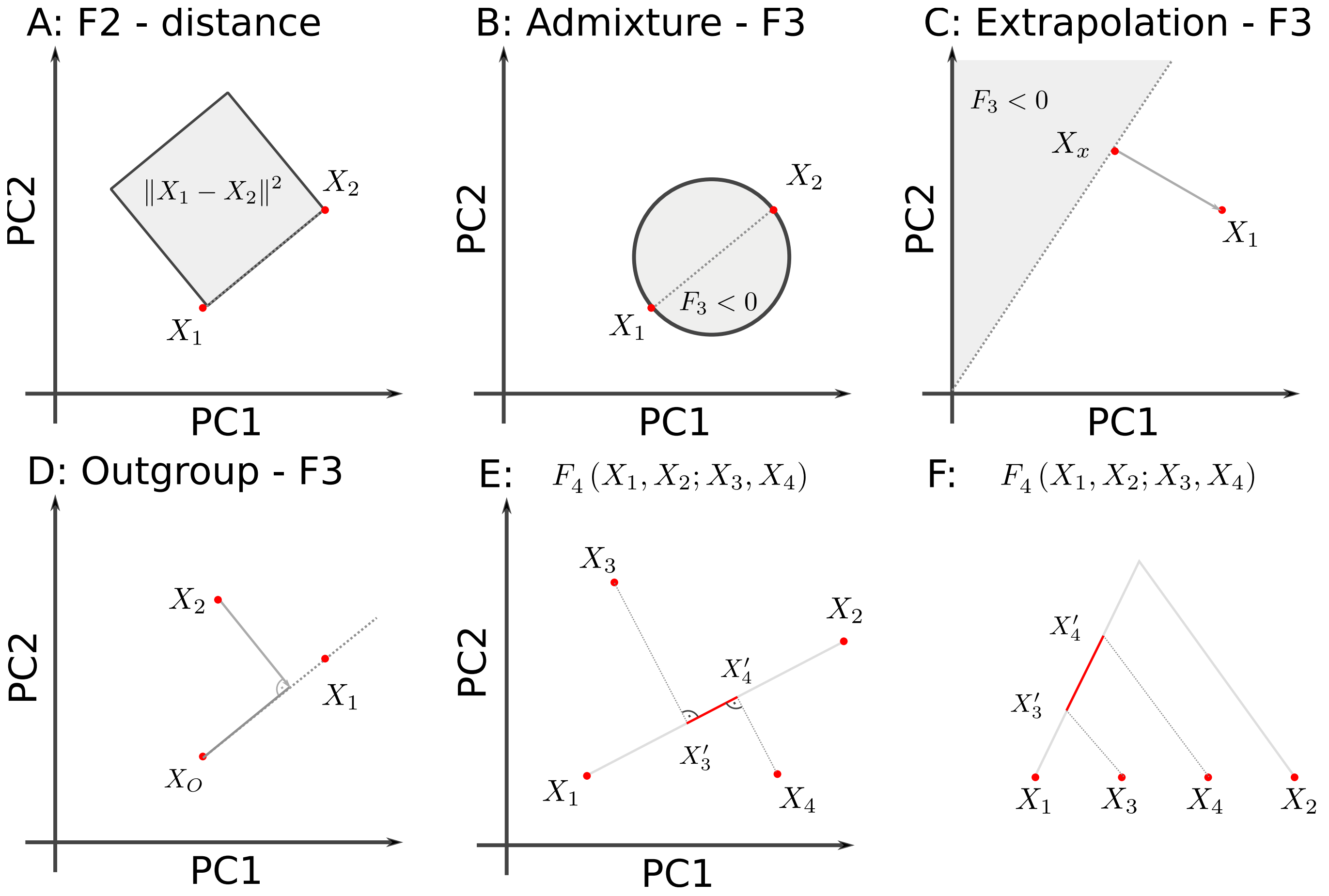

Figure 1: **Geometric representation of F-statistics on 2D-PCA-plot.** A: F_2 represents the squared Euclidean distance between two points in PC-space. B: Admixture-$F_3(X_x; X_1, X_2)$ is negative if X_x lies in the circle specified by the diameter $X_2 - X_1$. C: $F_3(X_x; X_1, X_2)$ is negative given X_1, X_x if X_2 is in the gray space. D: Outgroup-F_3 reflects the projection of $X_2 - X_O$ on $X_1 - X_O$. E: F_4 is the projection of $X_3 - X_4$ on $X_1 - X_2$. F: Same projection, but assuming data is generated by a tree.

The transformation from the previous section allows us to consider the geometry of F-statistics in PCA-space. The relationships we will discuss formally only hold if we use all $n-1$ PCs. However, the appeal of PCA is that frequently, only a very small number $K \ll n$ of PCS contain most information that is relevant for population structure (for visualization $K = 2$ is often used).

4.1 F_2 in PC-space

The F_2-statistic is an estimate of the squared Euclidean distance between two populations. It thus corresponds to the squared distance between populations in PCA-space, and reflects the intuition that closely related populations will be close to each other on a PCA-plot, and have low pairwise F_2-statistics. In converse, if two populations have high F_2 but appear on the same point on an PCA-plot, this suggests that substantial variation is hidden on higher PCs.

4.2 When are admixture-F_3 statistics negative?

Given two source populations X_1, X_2, can we predict which populations X_x could be considered admixed between these populations based on PCA? Since the allele frequencies of X_x are intermediate between those of X_1 and X_2, we would expect it to lie between X_1 and X_2, with the exact location depending on sample sizes (Brisbin et al., 2012, McVean, 2009).

The F_3-statistic gives a more precise interpretation: we are looking for the space where F_3 is negative, i.e.

$$\begin{aligned} 2F_3(X_x; X_1, X_2) &= 2\langle X_x - X_1, X_x - X_2 \rangle \\ &= \|X_x - X_1\|^2 + \|X_x - X_2\|^2 - \|X_1 - X_2\|^2 < 0 \end{aligned} \tag{7}$$

By the Pythagorean theorem, $F_3 = 0$ if and only if X_1, X_2 and X_x form a right-angled triangle. The associated region where $F_3 = 0$ is a n-sphere (or a circle in two dimensions) with diameter $\overline{X_1 X_2}$. F_3 is negative when the triangle is obtuse, i.e. X_x is admixed if it lies inside the n-ball with diameter $\overline{X_1 X_2}$ (Figure 1B).

F_3 on a 2D-plot. If we project this n-ball on a two-dimensional plot, $\overline{X_1 X_2}$ will usually not align with the PCs; thus the ball may be somewhat larger. This geometry is perhaps easiest visualized on a globe. If we look at the globe from a view point parallel to the equator, both the north and south poles are visible at the very edge of the circle. But if we look at it from above the north pole, the north- and south-poles will be at the very same point.

Thus if $\hat{F}_3 \ll F_3$, the true circle will be bigger than would be predicted from a 2D-plot. In this case, substantial relevant genetic differentiation is "hidden" in the higher PCs. If this is the case, populations that appear inside the circle on a PCA-plot may, in fact, have positive F_3-statistics, because they are outside the n-ball in higher dimensions. However, if a population lies outside the circle on any 2D-projection, F_3 is guaranteed to be bigger than 0.

Example I visualize this on the admixture statistic $F_3(X; Basque, Turkish)$, on the first two PCs of the Westeurasian data set (Figure 2A). In this case, the projected n-ball (light gray) and circle based on 2D (dark gray) align relatively closely, but several populations inside the ball (e.g. Sardinian, Finnish) have, in fact, positive F_3-values, which reveals that the first two PCs do not capture all the genetic variation relevant for Southern European population structure. This is expected because for spatially continuous populations, PCA will not be sparse (Novembre and Stephens, 2008). Consequently, approximating F_3 by the first two or ten PCs (Figure 2B) only gives a coarse approximation of F_3, and from Figure 2C we see that many higher PCs contribute to F_3 statistics.

However, many populations, particularly from the Levant and Caucasus, fall outside the circle, which allows us to immediately conclude that their F_3-statistics must be positive.

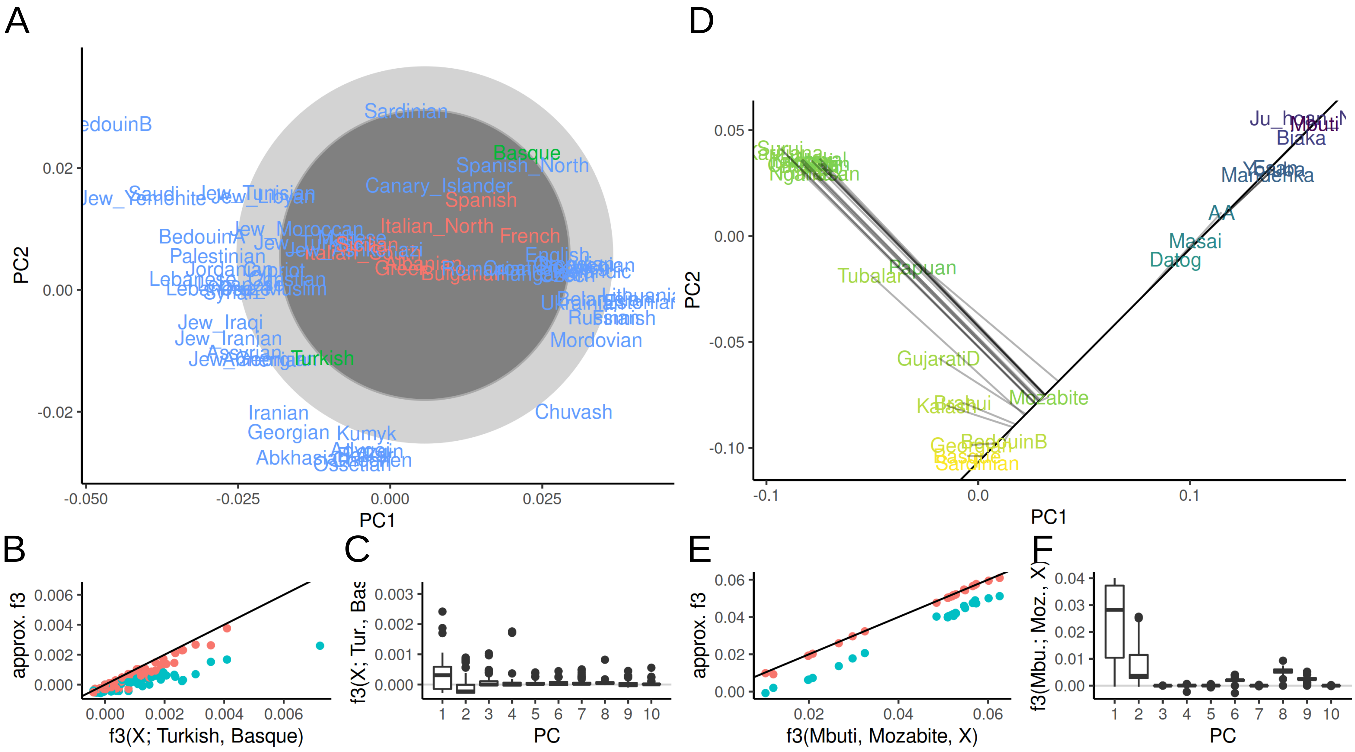

Figure 2: **PCA and F_3-statistics** A: PCA of Western Eurasian data; the circle denotes the region for which $F_3(X; \text{Basque}, \text{Turkish})$ may be negative. Populations for which F_3 is negative are colored in red. B, E: F_3 approximated with two (blue) and ten (red) PCs versus the full spectrum. C,F: Contributions of PCs 1-10 to each F_3-statistic. D: PCA of World data set, color indicates value of $F_3(\text{Mbuti}; \text{Mozabite}, X)$. The black line shows the projection axis Mbuti-Mozabite, the gray lines indicates the projected position of each population.

4.3 F_3 as a projection

The inner product $\langle X_x - X_1, X_x - X_2 \rangle$ is closely related to the projection of one vector onto the other. This interpretation is useful either when calculating an outgroup-F_3-statistic, or when we want to find the "best" admixture source X_2 if we assume the admixed population X_x and one source X_1 are known. The angle between $X_x - X_1$ and $X_x - X_2$ is obtuse if X_2 is in the half-plane whose boundary goes through X_x and is orthogonal to $\overline{X_X X_1}$ (Figure 1C), and its value is proportional to the projection onto $X_X - X_1$. A common interpretation guide is to use the population which yields the most negative F_3-statistic as the population which is most closely related to the "true" source of admixture (Patterson et al., 2012); on a PCA-plot this corresponds to using the population that is furthest away from X_x on the direction $X_x - X_1$.

This projection argument also helps to interpret Outgroup-F_3-statistic on a PCA-plot (Figure 1D), but in this case we aim to find the most closely related population as that with the highest F_3-statistic.

Example Again, these projection will be orthogonal when using the full data, and may only be approximately orthogonal when just using the first two PCs. In Figure 2D, I visualize this on the outgroup-F_3-statistic $F_3(Mbuti; Mozabite, X_i)$, i.e. a statistic that aims to find the population most closely related to Mozabite (a Berber ethnic group from the northern Sahara), assuming the Mbuti are an outgroup. On a PCA, we can interpret this F_3 statistic as the projection of the line segment from Mbuti to population X_i onto the line through Mbuti and Mozabite (black line). For each population, the projection is indicated with a grey line. In the full data space, this line is always orthogonal to the segment Mbuti-Mozabite, but on the plot (i.e.) the subspace spanned by the first two PCs, this is not necessarily the case. We see that particularly the samples from East Asia, Siberia and the Americas project very close to orthogonally, suggesting that most of the variation is captured by these first two PCs, and the coloring based on the full F_3-statistics shows that in this case, the first two PCs approximate the F_3-statistic very well. We can quantify this and find that the first two PCs slightly underestimate the absolute value of F_3 (Figure 2E), but keep the relative ordering. This F_3-statistic is also very sparse, with e.g. PCs 3-5, 7 and 10 having almost zero contribution to all statistics (Figure 2F), and PCs 6, 8 and 9 having a similar non-zero contribution for almost all statistics, likely because these PCs explain within-African variation.

4.4 F_4-statisticss as angles

The interpretation of F_4 in PCA is similar to that of F_3 as a projection of one vector onto another, with the difference that now all four points may be distinct. As for F_3, a finding of $F_4(X_1, X_2; X_3, X_4) = 0$ implies that the vectors $X_1 - X_2$ and X_3 and X_4 are orthogonal, or that the two populations map to the same point, and otherwise it will correspond to the length of the projection (Figure 1E).

We can also see how this interpretation aligns with that of F_4 as the length of an internal branch on a tree : By assumption, disjoint sets of branches evolve independently (Cavalli-Sforza et al., 1964, Felsenstein, 1973, Semple and Steel, 2003). Since the data space is sufficiently high-dimensional, this ensures that the resulting drift trajectories will also be uncorrelated. Therefore, if we interpret $F_4(X_1, X_2; X_3, X_4)$ as the projection of $X_3 - X_4$ - onto $X_1 - X_2$, we can write

$$X_3 - X_4 = (X_3 - X_3') + (X_3' - X_4') + (X_4' - X_4).$$

Of these three branches, the first and last are orthogonal to $X_1 - X_2$ and thus the F_4 statistic is just the internal branch of the tree (Figure 1F).

Since F_4 is a covariance, its magnitude lacks an interpretation. Therefore, commonly correlation coefficients are used, as there, zero means independence and one means maximum correlation. For F_4, we can write

$$\mathrm{Cor}(X_1 - X_2, X_3 - X_4) = \frac{\langle X_1 - X_2, X_3 - X_4 \rangle}{\|X_1 - X_2\| \, \|X_3 - X_3\|} = \cos(\phi), \tag{8}$$

where ϕ is the angle between $X_1 - X_2$ and $X_3 - X_4$. Thus, independent drift events lead to $\cos(\phi) = 0$, so that the angle is 90 degrees, whereas an angle close to zero means $\cos(\phi) \approx 1$, which means most of the genetic drift on this branch is shared.

Example To illustrate the angle interpretation I again use the Westeurasian data. The PCA-biplot shows two roughly parallel clines (Figure 2A), a European gradient (from Sardinian to Chuvash), and a Asian cline (from Arab to Caucasus populations).

This is quantified in Figure 3A, where I plot the angle corresponding to $F_4(X, Sardinian; Saudi, Georgian)$ For most European populations, using two PCs (green points) gives an angle close to zero, corresponding to a correlation coefficient $r > 0.9$. Just adding PC3 (blue) gives a close approximation to the full data (red). This can also be seen from the spectrum of these statistics (Figure 3B), which has high loadings for the first three PCs, with minimal contributions from the higher ones.

4.5 Other projections

So far, I used eq. 5 to interpret F-statistics on a PC-plot, but the argument holds for *any* orthonormal transformation. This allows for a variety of visualizations that use both F-statistics and PCs. The motivation for this is that sometimes we wish to partition the variation in the data into a subspace of interest, and an orthogonal residual space that captures the information discarded. Examples where analyses are restricted to such subspaces include the F_4-ratio test (Patterson et al., 2012, Petr et al., 2019), `qpWave` (Skoglund et al., 2015) and `qpAdm` (Harney et al., 2021). For the F_4-ratio, for example, a ratio

$$\alpha = \frac{F_4(R_1, R_2; X, A)}{F_4(R_1, R_2; B, A)} \tag{9}$$

is used, which can be interpreted as projecting $X - A$ and $B - A$ onto $R_1 - R_2$. Thus, we can make a plot where we plot the variation on the X-axis along $R_1 - R_2$, and perform a PCA on the residual. This can be important because the residual can be used to check assumptions, e.g. $A - A'$ and $B - B'$ need to be orthogonal (Figure 1F).

4.5.1 Example

In the PCA on the world overview data set, I found a seemingly linear gradient from Africans to Europeans (Figure 2D). I focus on this cline using an alternative projection by using F-statistics of the form $F_4(X, Y; \text{Sardinian}, \text{Yoruba}))$, which might e.g. be used if we were to quantify gene flow associated with the out-of-Africa expansion. These F_4-statistics are very well-approximated by the first two PCs, with a 99.2% correlation between F_4 and its approximation using the first two PCs (Figure 3C).

In Figure 3D, I show the projection $\langle X; \text{Sardinian}, \text{Yoruba}\rangle$ on the X-axis, which means that $F_4(X, Y; \text{Sardinian}, \text{Yoruba})$ is is proportional to their horizontal distance between X and Y. The first two residual PCs are given on the Y-axis and in the coloring; this visualization reveals some variation within Africans (with Mbuti, Biaka and Ju—'hoansi) that is largely orthogonal to this gradient, as is the variation between Europeans, Asian and the Americans.

The percentage of between-population variance explained by the Sardinia-Yoruba axis (24%) is much lower than PC1 (40%). However, the cumulative variance explained of the first two axes is similar, with (52%) explained when adding residual PC1 to the projection, compared to 55% for the first two PCs. The advantage of the projection is that it displays orthogonal components more explicitly, reveals distinct structure in Africans and non-Africans and thus can be used to test assumptions of more complex models.

5 Discussion

Particularly for the analysis of ancient DNA, F-statistics are a powerful tool to describe population genetic diversity. Here, I show that the geometry of F-statistics (Oteo-Garcia and Oteo, 2021) leads to a number of simple interpretations of F-statistics on a PCA plot. This allows for direct and quantitative comparisons between F-statistic-based results and PCA biplots. As PCA is often ran in an early step in data analysis, this also aids in generation of hypotheses that can be more directly evaluated using specific models using a lower number of populations. It could also allows for calculation of F-statistics involving unsampled populations, which can be useful for checking assumptions.

As F-statistics are motivated by trees, they assume that populations are discrete, related as a graph, and that gene flow between populations is rare (Patterson et al., 2012, Harney et al., 2021). However, in humans, most populations are admixed to some degree (Pickrell and Reich, 2014), and in regions such as Europe, genetic diversity is distributed continuously (Novembre et al., 2008). This provides a challenge for interpretation; as many F_3 and F_4 statistics may indicate gene flow. In my example (Figure 2A), most Southern European populations are "admixed" between Basques and Turkish, but a more accurate model might be one of continuous variation where Basque and Turkish lie on one of multiple gradients; which is more directly visualized with PCA. There are a number

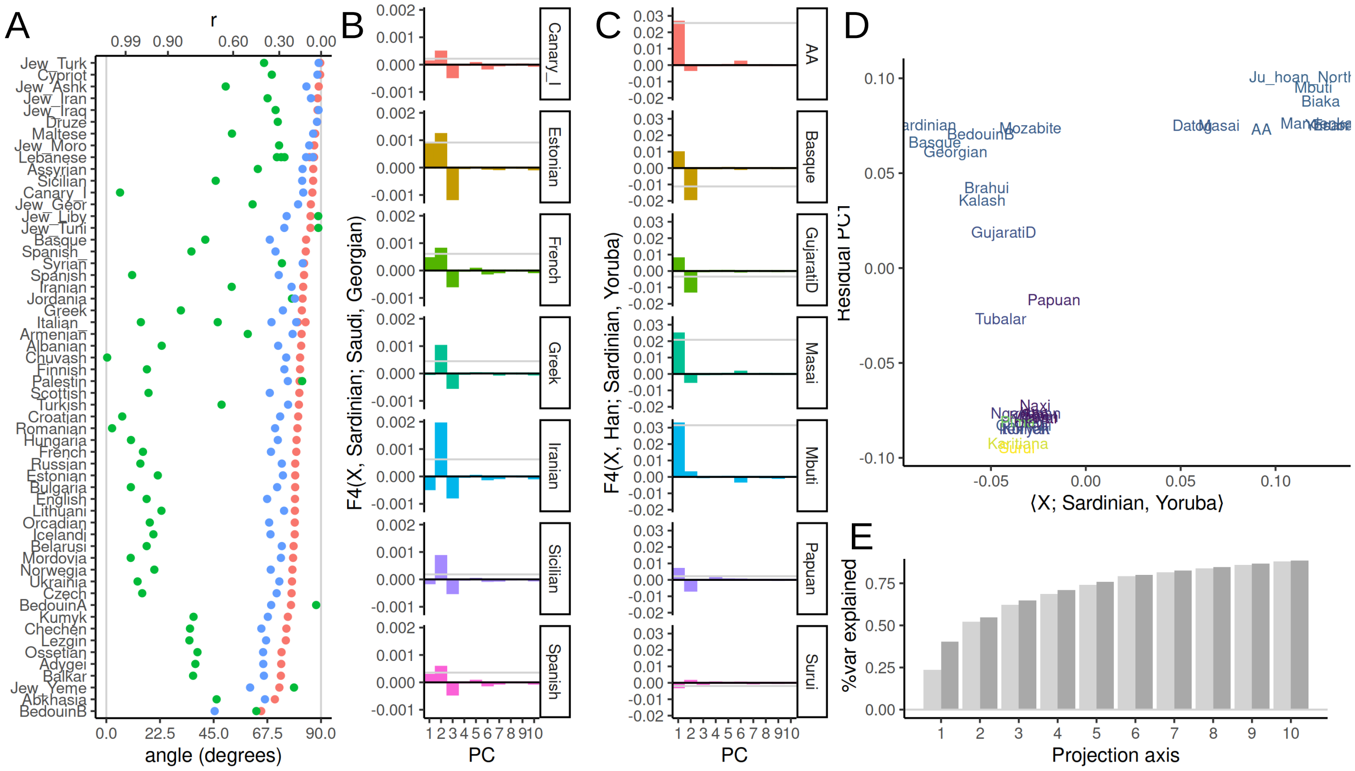

Figure 3: **PCA and F_4-statistics** A: Projection angle and correlation coefficient r representation of $F_4(X, Sardinian; Saudi, Georgian)$ (red) in the Westeurasian data set, and approximations using two (green) and three (blue) PCs. B: Spectrum of select F_4-statistics in the Western Eurasian data set. C: Spectrum of $F-4$-statistics in World data set. D: Scatterplot of F_3-projection on Sardinian-Yoruba axis and residual PC1. E: Percent variance explained from of the projection based on F_3 in panel D and first nine residual PCs (light gray), compared with percent variance explained by first ten PCs (dark gray).

of tools that have been developed that use multiple F-statistics to build complex models, such as `qpGraph` (Lazaridis et al., 2014) and `qpAdm` (Harney et al., 2021). One issue with these approaches is that they are usually restricted to at most a few dozen populations. As ancient DNA data sets now commonly include thousands of individuals, analysts are faced with the challenge of which data to include. A common approach is to sample a large number of distinct models, and retain the ones that are compatible with the data. However, as both `qpGraph` and `qpAdm` assume that gene flow is rare and discrete, selecting sets of populations that did experience little gene flow will provide good fits. One example of this is the world foci data set used here, which contains only 33 populations from across the world, and which is well-approximated by two PCs. However, this ascertainment misses a large amount of variation; a more dense sampling would show that in many places human genetic diversity is very gradual and multi-layered (Lazaridis et al., 2014, Peter et al., 2020). The PCA-based interpretation offers an alternative that trades interpretability for robustness. Particularly interpreting a (normalized) F_4-statistic as a correlation coefficient translates naturally to models of gene flow that do not assume gene flow is rare or discrete. Separating F-statistics in a sum of model and residuals, and performing a PCA on the latter (such as in Figure 3D) is another way how we can visualize F-statistics and evaluate the model fit.

To make this link directly applicable to data analysis, there are a number of – primarily statistical – concerns that will need to be addressed. Fist, PCA is most frequently run on individuals, whereas F-statistics are often calculated on populations. This is largely because in most workflows, PCA is run much earlier than F-statistics; it is a standard assumption of F-statistics that there is no population substructure (Patterson et al., 2012), and an easy way to test that is ensure that all individuals cluster tightly on a PCA.

A second difference is that frequently, rare SNPs are weighted higher in PCA, whereas all SNPs are weighted the same for F-statistics (Patterson et al., 2006). This is a difference of convention; since all SNPs have the same expectation, F-statistics could also be calculated using the same weighting. The close connection between the two approaches developed here suggest that for most analyses, users might want to be consistent and use the same weighting for both types of analyses.

The third and perhaps biggest gap are statistical issues. The treatment here focuses on the mean estimated F-statistic, but many applications of F-statistics are based on hypothesis tests (Patterson et al., 2012). This requires estimating accurate standard errors for these statistics, which is difficult since nearby SNPs will be correlated due to recombination (Hahn, 2018). In contrast, PCA jointly models the covariance matrix due to population structure and sampling, so if hypothesis tests are desired this will need to be incorporated.

An advantage of calculating F-statistics based on PCs is that they yield consistent estimtates. For both data sets I investigated here, the matrix $\mathbf{F}_2$ of F-statistics obtained using admixtools2 is not a proper squared Euclidean distance matrix, i.e. it is not negative semidefinite and has imaginary PCs. A model-based framework based on probabilistic PCA (Hastie et al., 2015, Meisner et al., 2021, Agrawal et al., 2020) would likely be able to generate consistent F-statistics and PCs, while incorporating sampling error and missing data.

A Derivation

$$
\begin{aligned}
F_2(X_i, X_j) &= \sum_{l=1}^{L} \left((x_{il} - \mu_l) - (x_{jl} - \mu_l)\right)^2 = F_2(Y_i, Y_j) \\
&= \sum_{l=1}^{L} \big(\sum_k L_{kl} P_{ik} - \sum_k L_{kl} P_{kj}\big)^2 \\
&= \sum_{l=1}^{L} \left(\sum_k L_{kl}(P_{ik} - P_{jk})\right)^2 \\
&= \sum_{l=1}^{L} \left(\sum_k L_{kl}^2 (P_{ik} - P_{jk})^2 + 2 \sum_{k \neq k'} L_{kl} L_{k'l} (P_{ik} - P_{jk'})^2\right) \\
&= \sum_k \underbrace{\left(\sum_{l=1}^{L} L_{kl}^2\right)}_{1} (P_{ik} - P_{jk})^2 + \sum_{k \neq k'} \underbrace{\left(\sum_{l=1}^{L} L_{kl} L_{k'l}\right)}_{0} (P_{ik} - P_{jk'})^2 \\
&= \sum_k (P_{ik} - P_{jk})^2 \qquad (10)
\end{aligned}
$$

In summary, the first row shows that F_2 on the centered data will give the same results (as distances are invariant to translations), in the second row we apply the PC-decomposition. The third row is obtained from factoring out L_{lk}. Row four is obtained by multiplying out the sum inside the square term for a particular l. We have k terms when for $\binom{k}{2}$ terms for different k's. Row five is obtained by expanding the outer sum and grouping terms by k.The final line is obtained by recognizing that $\mathbf{L}$ is an orthonormal basis; where dot products of different vectors have lengths zero.

Note that if we estimate F_2, unbiased estimators are obtained by subtracting the population-heterozygosities H_i, H_j from the statistic. As these are scalars, they do not change above calculation.

References

Aman Agrawal, Alec M. Chiu, Minh Le, Eran Halperin, and Sriram Sankararaman. Scalable probabilistic PCA for large-scale genetic variation data. *PLOS Genetics*, 16(5):e1008773, 2020. ISSN 1553-7404.

Isabel Alves, Miguel Arenas, Mathias Currat, Anna Sramkova Hanulova, Vitor C. Sousa, Nicolas Ray, and Laurent Excoffier. Long-distance dispersal shaped patterns of human genetic diversity in Eurasia. *Molecular biology and evolution*, 33(4):946–958, 2016.

Guido Barbujani, Arianna Magagni, Eric Minch, and L. Luca Cavalli-Sforza. An apportionment of human DNA diversity. *Proceedings of the National Academy of Sciences*, 94(9):4516–4519, April 1997.

Gideon S. Bradburd, Peter L. Ralph, and Graham M. Coop. Disentangling the Effects of Geographic and Ecological Isolation on Genetic Differentiation. *Evolution*, 67(11):3258–3273, 2013. ISSN 1558-5646.

Gideon S. Bradburd, Graham M. Coop, and Peter L. Ralph. Inferring continuous and discrete population genetic structure across space. *Genetics*, 210(1):33–52, 2018.

Abra Brisbin, Katarzyna Bryc, Jake Byrnes, Fouad Zakharia, Larsson Omberg, Jeremiah Degenhardt, Andrew Reynolds, Harry Ostrer, Jason G. Mezey, and Carlos D. Bustamante. PCAdmix: Principal Components-Based Assignment of Ancestry along Each Chromosome in Individuals with Admixed Ancestry from Two or More Populations. *Human biology*, 84(4):343–364, August 2012. ISSN 0018-7143.

L. L Cavalli-Sforza, I. Barrai, and A. W. F Edwards. Analysis of Human Evolution Under Random Genetic Drift. *Cold Spring Harbor Symposia on Quantitative Biology*, 29:9–20, January 1964. ISSN 0091-7451, 1943-4456.

L. L. Cavalli-Sforza, P. Menozzi, and A. Piazza. *The history and geography of human genes*. Princeton university press, 1994.

Barbara E. Engelhardt and Matthew Stephens. Analysis of Population Structure: A Unifying Framework and Novel Methods Based on Sparse Factor Analysis. *PLoS Genet*, 6(9):e1001117, September 2010.

Laurent Excoffier, Isabelle Dupanloup, Emilia Huerta-Sánchez, Vitor C. Sousa, and Matthieu Foll. Robust Demographic Inference from Genomic and SNP Data. *PLOS Genetics*, 9(10):e1003905, October 2013. ISSN 1553-7404.

J Felsenstein. Maximum-likelihood estimation of evolutionary trees from continuous characters. *American Journal of Human Genetics*, 25(5):471–492, September 1973. ISSN 0002-9297.

Olivier François, Mathias Currat, Nicolas Ray, Eunjung Han, Laurent Excoffier, and John Novembre. Principal Component Analysis under Population Genetic Models of Range Expansion and Admixture. *Molecular Biology and Evolution*, 27(6):1257–1268, June 2010. ISSN 0737-4038, 1537-1719.

J. C. Gower. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53(3-4):325–338, December 1966. ISSN 0006-3444.

R.E. Green, J. Krause, A.W. Briggs, T. Maricic, U. Stenzel, M. Kircher, N. Patterson, H. Li, W. Zhai, M.H.Y. Fritz, et al. A draft sequence of the Neandertal genome. *science*, 328(5979):710, 2010.

Ryan N. Gutenkunst, Ryan D. Hernandez, Scott H. Williamson, and Carlos D. Bustamante. Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLoS Genet*, 5(10):e1000695, October 2009.

Matthew Hahn. *Molecular Population Genetics*. Oxford University Press, Oxford, New York, August 2018. ISBN 978-0-87893-965-7.

Eadaoin Harney, Nick Patterson, David Reich, and John Wakeley. Assessing the performance of qpAdm: a statistical tool for studying population admixture. *Genetics*, 217(4), April 2021. ISSN 1943-2631.

Trevor Hastie, Rahul Mazumder, Jason D. Lee, and Reza Zadeh. Matrix completion and low-rank SVD via fast alternating least squares. *The Journal of Machine Learning Research*, 16(1):3367–3402, January 2015. ISSN 1532-4435.

N. J. Higham. Computing the nearest correlation matrix–a problem from finance. *IMA Journal of Numerical Analysis*, 22(3):329–343, July 2002. ISSN 0272-4979, 1464-3642.

I. T. Jolliffe. *Principal Component Analysis*. Springer Science & Business Media, March 2013. ISBN 978-1-4757-1904-8.

John A. Kamm, Jonathan Terhorst, and Yun S. Song. Efficient computation of the joint sample frequency spectra for multiple populations. *arXiv:1503.01133 [math, q-bio]*, March 2015.

Iosif Lazaridis, Nick Patterson, Alissa Mittnik, Gabriel Renaud, Swapan Mallick, Karola Kirsanow, Peter H. Sudmant, Joshua G. Schraiber, Sergi Castellano, Mark Lipson, and others. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*, 513(7518):409–413, 2014.

Stephen Leslie, Bruce Winney, Garrett Hellenthal, Dan Davison, Abdelhamid Boumertit, Tammy Day, Katarzyna Hutnik, Ellen C. Royrvik, Barry Cunliffe, Daniel J. Lawson, Daniel Falush, Colin Freeman, Matti Pirinen, Simon Myers, Mark Robinson, Peter Donnelly, and Walter Bodmer. The fine-scale genetic structure of the British population. *Nature*, 519(7543):309–314, March 2015. ISSN 1476-4687.

R. C. Lewontin. The Apportionment of Human Diversity. In Theodosius Dobzhansky, Max K. Hecht, and William C. Steere, editors, *Evolutionary Biology*, pages 381–398. Springer US, New York, NY, 1972. ISBN 978-1-4684-9065-7 978-1-4684-9063-3.

Mark Lipson, Po-Ru Loh, Alex Levin, David Reich, Nick Patterson, and Bonnie Berger. Efficient Moment-Based Inference of Admixture Parameters and Sources of Gene Flow. *Molecular Biology and Evolution*, 30(8):1788–1802, August 2013. ISSN 0737-4038, 1537-1719.

Anna-Sapfo Malaspinas, Ole Tange, José Víctor Moreno-Mayar, Morten Rasmussen, Michael DeGiorgio, Yong Wang, Cristina E. Valdiosera, Gustavo Politis, Eske Willerslev, and Rasmus Nielsen. bammds: a tool for assessing the ancestry of low-depth whole-genome data using multidimensional scaling (MDS). *Bioinformatics (Oxford, England)*, 30(20):2962–2964, October 2014. ISSN 1367-4811.

Gil McVean. A genealogical interpretation of principal components analysis. *PLoS genetics*, 5(10): e1000686, October 2009. ISSN 1553-7404.

Jonas Meisner, Siyang Liu, Mingxi Huang, and Anders Albrechtsen. Large-scale Inference of Population Structure in Presence of Missingness using PCA. *Bioinformatics (Oxford, England)*, page btab027, January 2021. ISSN 1367-4811.

J. Novembre and M. Stephens. Interpreting principal component analyses of spatial population genetic variation. *Nature genetics*, 40(5):646–649, 2008.

John Novembre, Toby Johnson, Katarzyna Bryc, ZoltÃ¡n Kutalik, Adam R Boyko, Adam Auton, Amit Indap, Karen S King, Sven Bergmann, Matthew R Nelson, Matthew Stephens, and Carlos D Bustamante. Genes mirror geography within Europe. *Nature*, 456(7218):98–101, 2008.

Gonzalo Oteo-Garcia and Jose-Angel Oteo. A geometrical framework for f-statistics. *Bulletin of Mathematical Biology*, 83(2):1–22, 2021.

Nick Patterson, Daniel J. Richter, Sante Gnerre, Eric S. Lander, and David Reich. Genetic evidence for complex speciation of humans and chimpanzees. *Nature*, 441(7097):1103–1108, June 2006. ISSN 0028-0836.

Nick J. Patterson, Priya Moorjani, Yontao Luo, Swapan Mallick, Nadin Rohland, Yiping Zhan, Teri Genschoreck, Teresa Webster, and David Reich. Ancient Admixture in Human History. *Genetics*, page genetics.112.145037, September 2012. ISSN 0016-6731, 1943-2631.

Benjamin M. Peter. Admixture, Population Structure and F-Statistics. *Genetics*, page genetics.115.183913, January 2016. ISSN 0016-6731, 1943-2631.

Benjamin M. Peter, Desislava Petkova, and John Novembre. Genetic landscapes reveal how human genetic diversity aligns with geography. *Molecular biology and evolution*, 37(4):943–951, 2020.

Martin Petr, Svante Pääbo, Janet Kelso, and Benjamin Vernot. Limits of long-term selection against Neandertal introgression. *Proceedings of the National Academy of Sciences*, 116(5):1639–1644, January 2019.

Joseph K. Pickrell and David Reich. Toward a new history and geography of human genes informed by ancient DNA. *Trends in Genetics*, 30(9):377–389, September 2014. ISSN 0168-9525.

Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

Fernando Racimo, Jessie Woodbridge, Ralph M. Fyfe, Martin Sikora, Karl-Göran Sjögren, Kristian Kristiansen, and Marc Vander Linden. The spatiotemporal spread of human migrations during the European Holocene. *Proceedings of the National Academy of Sciences*, 117(16):8989–9000, April 2020.

Sohini Ramachandran, Omkar Deshpande, Charles C Roseman, Noah A Rosenberg, Marcus W Feldman, and L. Luca Cavalli-Sforza. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proceedings of the National Academy of Sciences of the United States of America*, 102(44):15942–15947, 2005. ISSN 0027-8424, 1091-6490.

David Reich. *Who We Are and How We Got Here: Alte DNA und die neue Wissenschaft der menschlichen Vergangenheit.* Pantheon, New York, illustrated edition edition, 2018. ISBN 978-1-101-87032-7.

Noah A. Rosenberg, Jonathan K. Pritchard, James L. Weber, Howard M. Cann, Kenneth K. Kidd, Lev A. Zhivotovsky, and Marcus W. Feldman. Genetic structure of human populations. *Science (New York, N.Y.)*, 298(5602):2381–2385, December 2002. ISSN 1095-9203.

Noah A Rosenberg, Saurabh Mahajan, Sohini Ramachandran, Chengfeng Zhao, Jonathan K Pritchard, and Marcus W Feldman. Clines, Clusters, and the Effect of Study Design on the Inference of Human Population Structure. *PLoS Genet*, 1(6):e70, December 2005.

Charles Semple and M. A. Steel. *Phylogenetics.* Oxford University Press, 2003. ISBN 978-0-19-850942-4.

David Serre and Svante Pääbo. Evidence for Gradients of Human Genetic Diversity Within and Among Continents. *Genome Research*, 14(9):1679–1685, September 2004. ISSN 1088-9051, 1549-5469.

Pontus Skoglund, Swapan Mallick, Maria Cátira Bortolini, Niru Chennagiri, Tábita Hünemeier, Maria Luiza Petzl-Erler, Francisco Mauro Salzano, Nick Patterson, and David Reich. Genetic evidence for two founding populations of the Americas. *Nature*, 525(7567):104–108, September 2015. ISSN 1476-4687.

Mark Stoneking. *An Introduction to Molecular Anthropology.* John Wiley & Sons, December 2016. ISBN 978-1-118-06162-6.