1 Introduction

About 15% of genetic variation in humans can be explained by population structure (Lewontin, 1972; Barbujani *et al.*, 1997; Rosenberg and Nordborg, 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, 2018b). 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.

Understanding and characterizing this variation is also crucial for genomic medicine: Risk loci identified in one population perform progressively worse in more distant populations (Berg *et al.*, 2019; Duncan *et al.*, 2019); so identifying a good "reference" data set is crucial. Understanding genetic variation is also important when designing studies, in order to measure which populations studies should target (**?**). Despite these needs, no generally accepted model for human population structure exists. The historical view that humans diversity can be partitioned into distinct "races" (e.g. **?**) has long been superseded in favor of more accurate and nuanced models. These models fail because they only explain a small amount of variation...

In Lewontin's pioneering analysis, he found that less than half (6%), of that variation could be attributed to the continental-scale groups he called races, it seemed which he used to claim that "racial classification is (...) seen to be of virtually no genetic or taxonomic significance".

Human genetic diversity has both discrete and continuous components (Rosenberg and Nordborg, 2002; Rosenberg *et al.*, 2005; **?**; Reich, 2018a; Peter *et al.*, 2020). On the one hand, barriers to migration such as mountain ranges, oceans or deserts lead to reduced gene flow, and thus populations on different continents had periods with little-to-no gene flow. On the other hand, local gene flow between neighboring populations is also pervasive, and the out-of-Africa expansion caused gradual patterns of differentiation Ramachandran *et al.* (2005). Finally, long-distance migrations, and secondary contact between diverged groups such as Neandertals and early modern humans add to the complexity of human population structure.

The way this complexity is frequently handled is that multiple models with different assumptions and used, 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 more explicit demographic models (Gutenkunst *et al.*, 2009; Kamm *et al.*, 2015; Excoffier *et al.*, 2013) are used, 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 introduction of PCA is given in e.g. Jolliffe (2013), and a useful guide to interpretation is Cavalli-Sforza *et al.* (1994).

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 $n-1$, but the historical processes that generated often result in *sparse* data (Engelhardt and Stephens, 2010), so that $K \ll n$ explains most ofthe variation; for visualization $K = 2$ is frequently used.

There are several algorithms that are used to calculate a PCA in practice, 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

$$\mathbf{Y} = \mathbf{CX} = (\mathbf{U\Sigma})\mathbf{V}^T = \mathbf{PL}, \tag{1}$$

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 $\mathbf{P} = \mathbf{U\Sigma}$ 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.

$$F_2(X_1, X_2) = \sum_{l=1}^{S}(x_{1l-x2l})^2 = \|X_1, X_2\|^2 \quad (2a)$$

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

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

where $\|\cdot\|$ denotes the Euclidean norm and $\langle\cdot,\cdot\rangle$ denotes the dot product. Furthermore, both F_3 and F_4 can be written as sums of F_2-statistics: $2F_3(X_1; X_2, X_3) = 2F_2(X_2, X_3) - F_2(X_1, X_2) + F_2(X_1, X_3)$ and $2F_4(X_1, X_2; X_3, X_4) = F_2(X_1, X_3) + F_2(2_4) - F_2(X_1, X_4) - F_2(X_2, X_3)$.

F-statistics have been primarily motivated in the context of trees and admixture graphs (Patterson *et al.*, 2012). In a tree, $F_2(X, Y)$ 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).

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 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: 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 populations are related as a tree, then $F_4(X_1, X_2; X_3, X_4) = 0$.

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:

$$\mathbf{P}\mathbf{P}^T = -\frac{1}{2}\mathbf{C}\mathbf{F}_2\mathbf{C}. \quad (3)$$

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,

$$\begin{aligned} F_2(X_i, X_j) = & \sum_{l=1}^{L} \big(x_{il} - x_{jl}\big)^2 \\ = & \sum_{l=1}^{L} \big((x_{il} - \mu_l) - (x_{jl} - \mu_l)\big)^2 = F_2(Y_i, Y_j) \\ = & \sum_{k=1}^{n} (p_{ik} - p_{jk})^2 = F_2(P_i, P_j), \end{aligned} \tag{4}$$

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.

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

$$F_2(P_i, P_j) = \sum_{k=1}^{K} (p_{ik} - p_{jk})^2 + \sum_{k=K+1}^{n} (p_{ik} - p_{jk})^2, \tag{5}$$

where the first sum is the PCA-approximation of $\hat{F}_2$, and the second sum is the residual or approximation error $F_2 - \hat{F}_2$.

If we sum up the approximation errors over all pairs of populations, we obtain the Frobenius-norm of the error $\left\|\mathbf{F}_2 - \hat{\mathbf{F}}_2\right\|_F^2$; it is a standard result that PCA finds the best rank-K approximation so that this Frobenius-norm is minimized. In our context, this means that PCA using the first K PCs results in approximate F_2-statistics such that the sum of F_2-distances between the approximation and full data is minimized.

2.4 F-stats in 2-dimensional PC-space

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, it is often assumed that $K = 2$).

2.4.1 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 in PCA-space, and reflects 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 with high F_2 lie on the same point on an PCA-plot, this suggests that substantial variation is hidden on higher PCs.

2.4.2 When is F_3 negative?

The F_3-statistic becomes more interesting; as outlines above we either think of F_3 as "outgroup"-F-stats or as admixture F-stats. In the admixture case, we may ask the following question: given two source populations X_1, X_2, where would admixed populations on a PCA plot lie? From theory, 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).

Formally, we would reject admixture if F_3 is negative, i.e. we are looking for the space

$$\begin{aligned} 2F_3(X_x; X_1, X_2) &= 2\langle X_x - X_1, X_x - X_2\rangle \\ &= \|X_x - X_1\|^2 + \|X_x - X_2\|^2 - \|X_1 - X_2\|^2 \end{aligned} \tag{6}$$

By the Pythagorean theorem, $F_3 = 0$ iff X_1, X_2 and X_x form a right-angled triangle. In a 2D-PCA plot, the region where F_3 is zero is the circle with diameter $\overline{X_1X_2}$, and if X_x lies inside this circle, $F_3(X_x; X_1, X_2) < 0$. If the

ball, the angle is obtuse and F_3 is negative, otherwise it will be positive. If we approximate the PCA-space in two dimensions, the n-ball corresponds to a circle.

2.4.3 F_4 and right angles

The inner-product-interpretation of F_4 is similar to that of F_3, with the change that the two vectors we consider do not involve the same population. However, a finding of $F_4(X_1, X_2; X_3, X_4) = \langle X_1 - X_2, X_3 - X4\rangle = 0$ similarly implies that the two vectors are orthogonal, and a non-zero value reflects the projection of one vector on the other.

2.4.4 F_4-ratio

$$\begin{aligned} \frac{F_4(X_I, X_O; X_X, X_1)}{F_4(X_I, X_O; X_2, X_1)} &= \frac{\|X_I - X_O\|\,\|X_X - X_1\|\cos(\alpha)}{\|X_I - X_O\|\,\|X_2 - X_1\|\cos(\beta)} \\ &= \frac{\|X_X - X_1\|\cos(\alpha)}{\|X_2 - X_1\|\cos(\beta)} \\ &= \frac{\|X'_X - X'_1\|}{\|X'_2 - X'_1\|} \end{aligned} \tag{7}$$

where α and β are the angles between vectors, and X'_i is the projection of X_i on $X_I - X_O$.

Conjecture: Thus, we are measuring the distances between the admixing populations on the projected on the axis between X_I and X_O. This ought to be valid only if $\langle X_1 - X'_1, X_2 - X'_2\rangle$ are orthogonal to each other, and to X_OX_I, i.e. $F_4(X_1, X'_1, X_2, X'_2) = 0$

3 Results

3.1 F_4

Using F_3-statistics, I showed that we can think of the admixture test as a test of whether the admixed population lies in a particular n-ball, and the outgroup F_3-statistic can be thought of as a projection of the test populations on the line connecting the outgroup to the reference sample. In this section, I will develop similar interpretation of F_4 on PCA-plots, and to investigate sparsity.

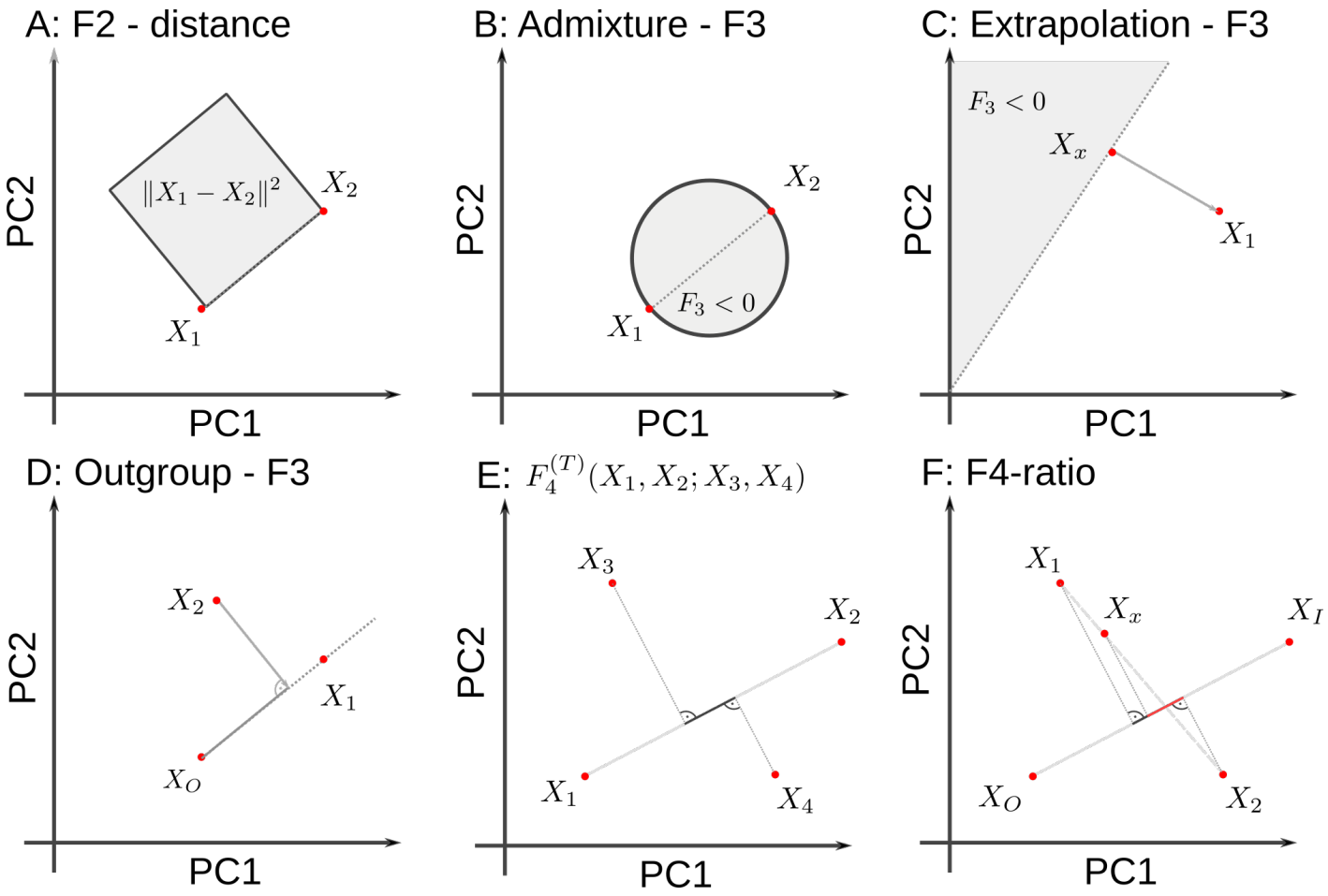

Figure 1: **Geometric representation of F-statistics on 2D-PCA-plot.** A: F_2 represents the squared Euclidean distance between two points in PC-space. B: Admixture-$F_3(X_x; X_1, X_2)$ is negative if X_x lies in the circle specified by the diameter $X_2 - X_1$
. C: $F_3(X_x; X_1, X_2)$ is negative given 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: If X_x is admixed between X_1 and X_2, the admixture proportions will be projected.

First, we investigate the sparsity in the world overview data set: We find that the vast amount of contribution to the statistics comes from the first two PCs (Figure 4A). For example, the correlation between $F_4(X, Y, \text{Mozabite}, \text{Yoruba})$ and its approximation using the first two PCs is 99.2%. To visualize the interpretation of F_4 as an angle, we use statistics of the form $F_4(X, \text{Sardinian}; \text{Mozabite}, \text{Yoruba})$, which can be interpreted as the angle between the vectors Mozabite-Yoruba and X-Sardinian. In Figure 4B, I show the angle based on two (blue), ten (green) and all PCs *red*. I find that for most Asian and American populations the angle is very close to 90°, as would be expected if the variation between African and non-African populations is mostly orthogonal. On the other hand, if X is an African population, the angle is lower, and much less well approximated. This demonstrates that this PCA-plot likely does not model within-African population structure adequately.

The F_4-statistics for the Westeurasian data set are slightly less sparse, the correlation coefficients between $F_4(X, \text{French}; \text{Finnish}, \text{Canary Islander})$ and its approximation using the first two or three PCs is 95.5% and 99.1% respectively (Figure 4E). I also show that the interpretation of F_4 as a projection can be used as a useful visualization (Figure 4D). On the x-axis, I plot $\langle X; \text{Finnish}, \text{Canary Islander} \rangle$, so

that the horizontal distance between all pairs of populations corresponds to their respective F_4-statistics $F_4(X, Y; \text{Finnish}, \text{Canary Islander})$. On the Y-axis and with the coloring I display the first two principal components of the residual, i.e. the genetic variation that is missed by viewing the data through this projection. We find that most European populations have positive values on residual PC1, and are relatively closely clustered. In contrast Middle Eastern and Caucasian populations have negative values on this gradient. This allows us to visualize that this particular F_4-projection does an adequate job if we are interested in describing European variation, but it fails to explain the non-European data. We can further quantify this by investigating the percent of variance explained on each axis (Figure 4F), where I find that the projection axis only describes around 12% of hte variation, compared to residual PC1 with almost 30%.

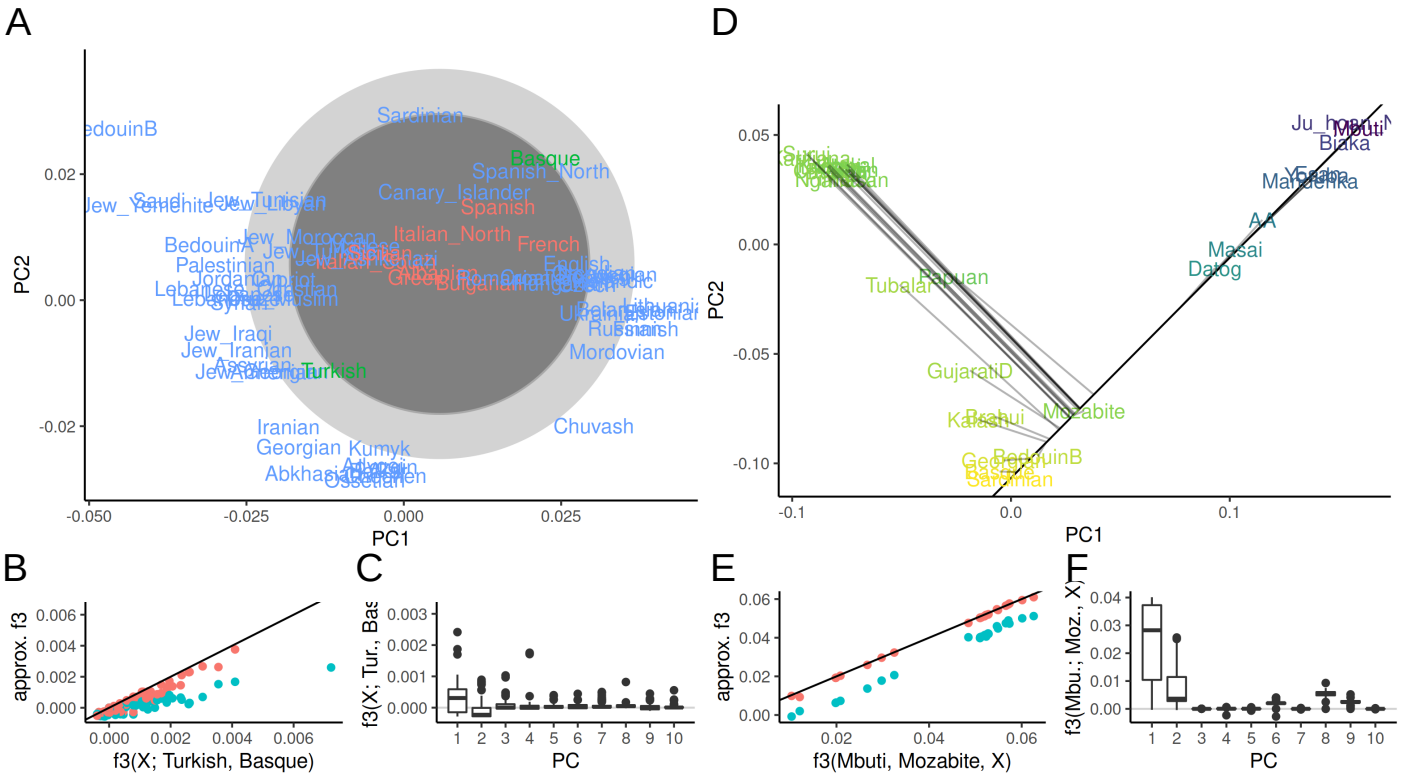

Figure 2: **PCA and F_3-statistics** A: PCA of Western Eurasian data; the circle denotes the region for which $F_3(X; \text{Basque}, \text{Turkish})$ may be negative. Populations for which F_3 is negative are colored in red. B, E: F_3 approximated with two (blue) and ten (red) PCs versus the full spectrum. 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.

3.2 What is a dimension?

In both the PCA and F-statistic framework, a population at a particular point in time can be thought of as a single point in allele-frequency space, given by the k-dimensional vector v_0 of allele frequencies at the k SNPs in that population. If this population evolves for some time in isolation, allele frequencies will change due to genetic drift from v_0 to some other point v_1. Likewise, a second population with frequency w_0 will move to w_1. Crucially, if these populations do not interact, the changes in allele frequency, $v_1 - v_0$

and $w_1 - w_0$ will be uncorrelated (Patterson *et al.*, 2012). Thus, if we have two populations that descend from the same ancestral population in isolation, they can be thought of as evolving along orthognal dimensions from the same point. This argument is at the foundation of F-statistics.

4 Results

The theory outlined in the previous section suggests that F-statistics have a geometric interpretation on PCA plots. In this section, I use these interpretation in the analysis of human genetic variation data set. I use two data sets based on the "Human Origins"-SNP set (597,573 SNPs). Both are subsets 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`.

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). Population genetic differentiation in this region is low and closely mirrors geography (Novembre *et al.*, 2008).

World Overview data set This data set of 638 individuals from 33 population contains individuals throughout the world, and is used as a sparse data set capturing much of global human genetic variation. This data set spans Africa, Eurasia and the Americas, and we might therefore expect the population structure to be much more sparse.

I perform analyses at the level of populations to ease presentation, and because it is an assumption of F-statistics that the genetic variation with sampled population is independent of the variation between samples that I am focusing on here. I use `admixtools 2.0.0` `https://github.com/uqrmaie1/admixtools` to compute a matrix of F_2-statistics between all populations. To obtain a PC-decomposition I use equation XXX and the `eigen` function in `R`, and compare them with the F_3 and F_4-statistics calculated using admixtools 2.

Admixture-F_3 As a first step, I plot the first two principal components of the Westeurasian data set (Figure 3A). This PCA presents two parallel clines, one from the Levant and Arabia ("BedouinB") to the Caucasus ("Abkhasian"), and a second one from Southern ("Sardinian") to Northeastern Europe ("Mordovian"). In this context, I examine $F_3(X; Basque, Turkish)$, i.e. a statistic that aims to ask which populations can be represented as a mixture between a Southwestern (Basque) and Southeastern (Turkish) European population. The – largely Southern European – populations for which the point estimate of these F_3-statistic is negative are highlighted in red. They both fall close to the center of the F_3-circle, either defined on the first two (dark grey) or all PCs (light gray). However, many populations inside the circle on the first 2 PCs, including English, Sardinians and Canary Islanders have positive F_3-values, on higher PCs, showing that the first two PCs do not capture all the genetic variation related to population structure for this data set.

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 3B) only gives a coarse approximation of F_3, and from Figure 3C we see that many higher PCs contribute to F_3 statistics.

Thus, the main benefit of this PCA-plot is that it allows us to identify populations outside the circle (from the Levant and Caucasus), for which F_3 is guaranteed to be positive.

Outgroup-F_3 The Outgroup-F_3-statistic is commonly used to infer which population is closest in a set of reference populations. In Figure 3D, I present a PCA of the world data set, with populations colored according to $F_3(Mbuti; Mozabite, X_i)$, i.e. a statistic that is commonly interpreted as finding the population X_i that is most closely related to Mozabite. On a PCA, we can interpret this F_3 statistic as the projection of the line segment MbutiX_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 only true if the relevant variation is captured by the first two PCs. 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. That the approximation of F_3 on

4.1 F_4

Using F_3-statistics, I showed that we can think of the admixture test as a test of whether the admixed population lies in a particular n-ball, and the outgroup F_3-statistic can be thought of as a projection of the test populations on the line connecting the outgroup to the reference sample. In this section, I will develop similar interpretation of F_4 on PCA-plots, and to investigate sparsity.

First, we investigate the sparsity in the world overview data set: We find that the vast amount of contribution to the statistics comes from the first two PCs (Figure 4A). For example, the correlation between $F_4(X, Y, \text{Mozabite}, \text{Yoruba})$ and its approximation using the first two PCs is 99.2%. To visualize the interpretation of F_4 as an angle, we use statistics of the form $F_4(X, \text{Sardinian}; \text{Mozabite}, \text{Yoruba})$, which can be interpreted as the angle between the vectors Mozabite-Yoruba and X-Sardinian. In Figure 4B, I show the angle based on two (blue), ten (green) and all PCs *red*. I find that for most Asian and American populations the angle is very close to 90°, as would be expected if the variation between African and non-African populations is mostly orthogonal. On the other hand, if X is an African population, the angle is lower, and much less well approximated. This demonstrates that this PCA-plot likely does not model within-African population structure adequately.

The F_4-statistics for the Westeurasian data set are slightly less sparse, the correlation coefficients between $F_4(X, \text{French}; \text{Finnish}, \text{Canary Islander})$ and its approximation using the first two or three PCs is 95.5% and 99.1% respectively (Figure 4E). I also show that the interpretation of F_4 as a projection can be used as a useful visualization (Figure 4D). On the x-axis, I plot $\langle X; \text{Finnish}, \text{Canary Islander}\rangle$, so that the horizontal distance between all pairs of populations corresponds to their respective F_4-statistics $F_4(X, Y; \text{Finnish}, \text{Canary Islander})$. On the Y-axis and with the coloring I display the first two principal components of the residual, i.e. the genetic variation that is missed by viewing the data through this projection. We find that most European populations have positive values on residual PC1, and are relatively closely clustered. In contrast Middle Eastern and Caucasian populations have negative values on this gradient. This allows us to visualize that this particular F_4-projection does an adequate job if we are interested in describing European variation, but it fails to explain the non-European data. We can further quantify this by investigating the percent of variance explained on each axis (Figure 4F), where I find that the projection axis only describes around 12% of hte variation, compared to residual PC1 with almost 30%.

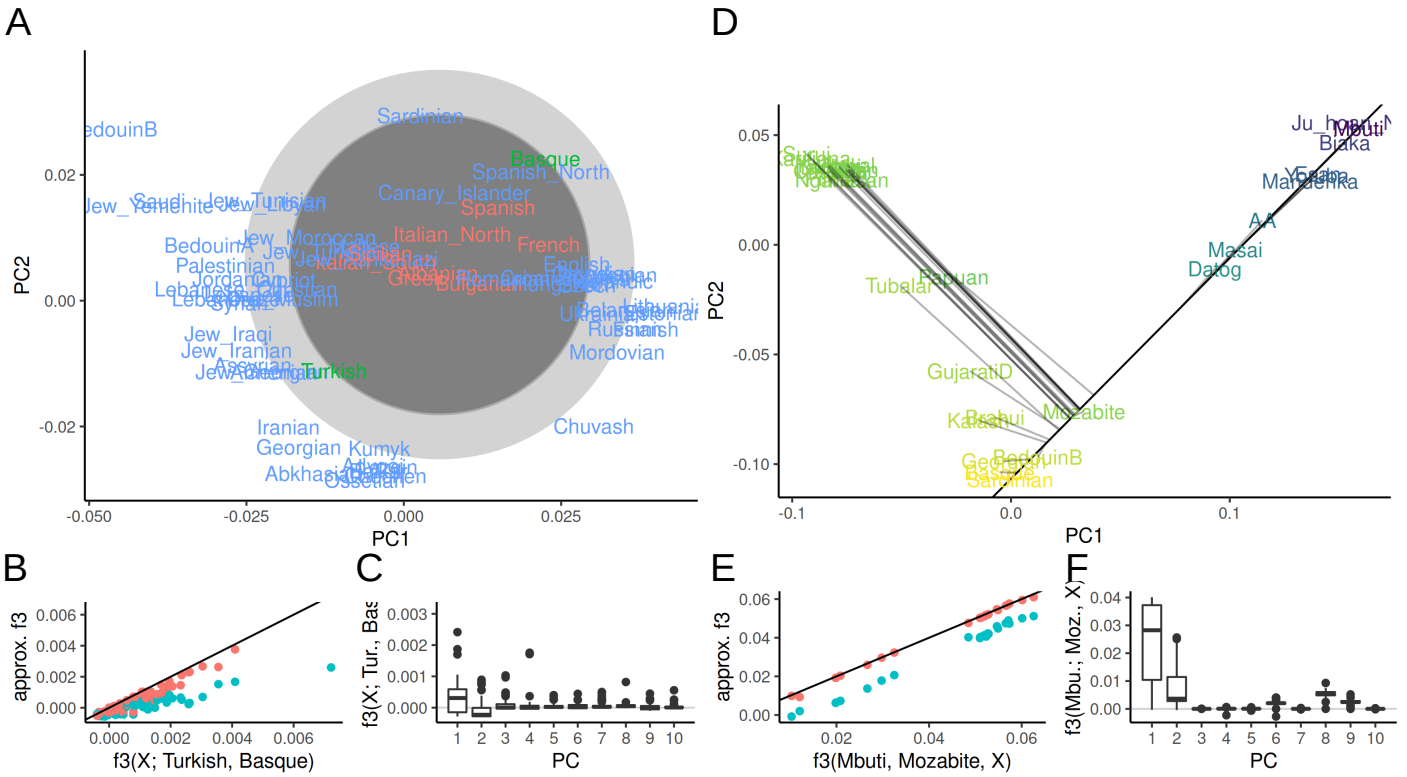

Figure 3: **PCA and F_3-statistics** A: PCA of Western Eurasian data; the circle denotes the region for which $F_3(X; \text{Basque}, \text{Turkish})$ may be negative. Populations for which F_3 is negative are colored in red. B, E: F_3 approximated with two (blue) and ten (red) PCs versus the full spectrum. 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.

5 Discussion

Particularly for the analysis of ancient DNA, F-statistics have been established as a powerful tool to describe population genetic diversity, but they have a number of limitations. In particular, 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). As a consequence, researchers concerned about model fits may ascertain reference populations in a way that satisfies these assumptions, thus inadvertently making population structure appear sparser than it truly is. This is perhaps most obvious from Figure 3B, where large gaps are present. However, these gaps are due to sparse sampling that disappear if more populations were sampled (e.g. Peter *et al.*, 2020), not due to gaps in genetic diversity. If population ascertainment is not done very carefully, tools built on top of F-statistics, such as `qpGraph` and `qpAdm`, may thus only provide a very loose lower bound for the number of gene flow events.

In contrast, the perspective on F-statistics in data space (Oteo-Garcia and Oteo, 2021) and on PCA does not require assumptions on number of admixture or gene flow events. Independent of any model, A population X_x can be thought of as admixed between X_1 and X_2 if it lies in the ball with diameter $\overline{X_1X_2}$. And independent of any models, two axes of variation that are not perpendicular to each other shows that there is some degree of shared history.

The connection between the graph-based and PCA-based interpretations of F_3 and F_4 is due to

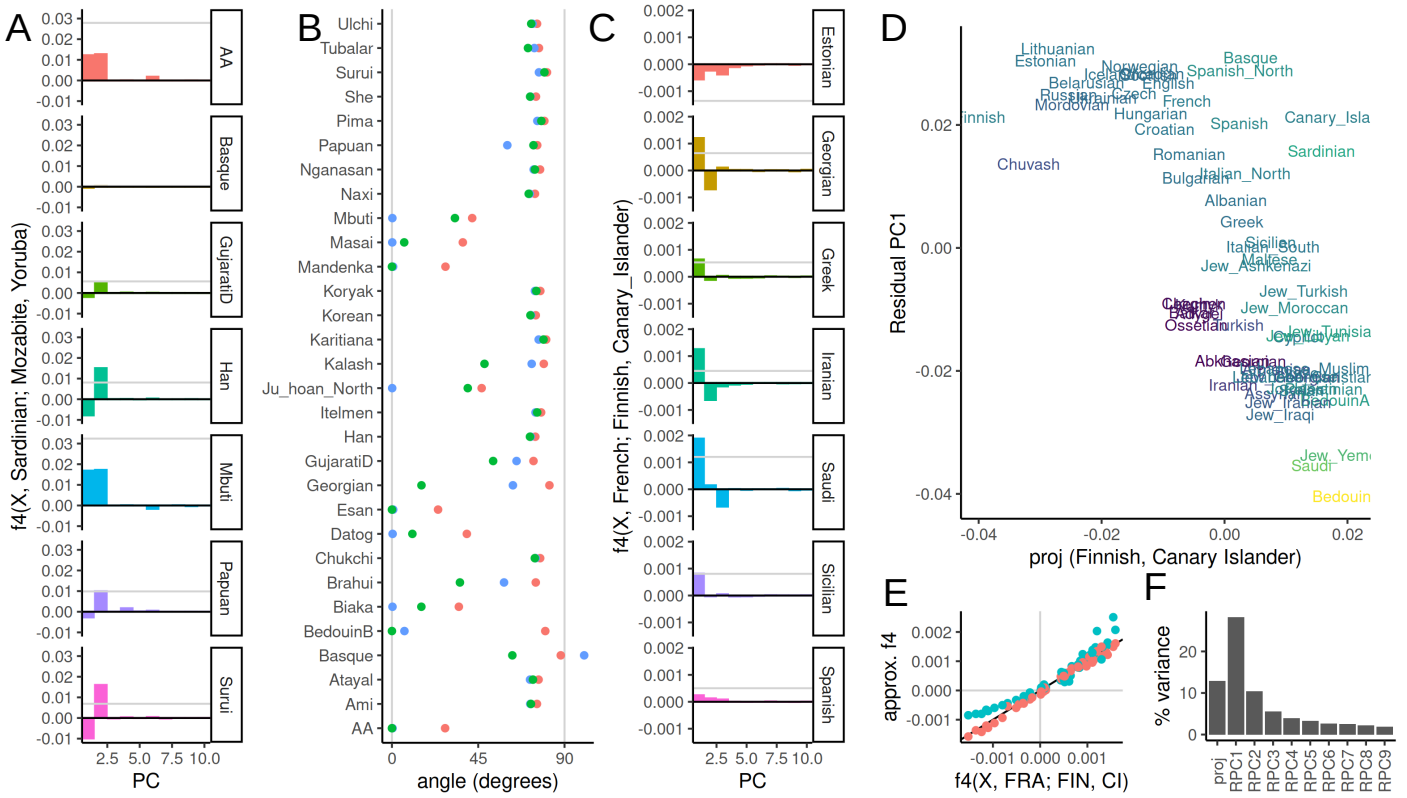

Figure 4: **PCA and F_4-statistics** A: Spectrum of select F_4-statistics in World data set. B: Projection angle representation of $F_4(X, Sardinian; Mozabite, Yoruba)$ (red) and approximations using two (blue) and ten (green) PCs. C: Spectrum of select $F-4$-statistics in Westeurasian data set. D: Scatterplot of F_4-projection on Finnish-Canary Islanders axis and residual PC1. E: $F_4(X, \text{French}, \text{Finnish}, \text{Canary Islander})$ vs. prediction using two (blue) and ten(red) PCs. F: Percent variance explained for the projection of panel D and the first nine residual PCs.

concept of projection. As illustrated above, both F-statistics can be thought of as projections on any particular axis of variation. If that variation corresponds to a tree-branch, the interpretations align. A collorary to this interpretation is the importance of orthogonality, or independence for describing gene flow. Populations without gene-flow will evolve independently, and their changes in allele frequency will therefore be orthogonal, resulting in F_4-statistics of zero, and right angles on a PCA plot. Even though this is only true when considering all PCs, the result holds reasonably well when just looking at the first two PCs – in our example of the world-overview data, I found that the variation in Sub-Saharan Africa is mostly – but not completely – uncorrelated with the variation between European and Asian populations (Figure 4A).

While the data space produces a much larger model space than trees, the main drawback of the PCA-based interpretation is a lack of interpretability; it is not easy to define a generative model that could generate a complex PCA-plot. Thus, in the future the tree- and PCA-based interpretations will likely be used in conjunction.

However, despite the apparent complexity, the PC-spaces *are* sparse, and using just ten or even two PCs often gives very good approximations.

To make PCA and F-statistics more comparable in practical settings, there are a number of – mainly statistical – concerns that still need to be addressed in future work. The perhaps most obvious

one is that PCA is most frequently run on individuals, whereas F-statistics are often calculated on populations. This is not a conceptual issue, as both PCA and F-statistics can be run on either (Cavalli-Sforza *et al.*, 1994). Population based analyses have the advantage that they are easier to interpret and compute (current packages are ill-equipped to calculate all pairwise F-statistics between data sets with thousands of individuals (Patterson *et al.*, 2012)). However, this requires the assumption that the within-population variation is independent from the between-population variation; something that is analogous to the variance partitioning based on PCs here.

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 only a difference of convention; 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 focusses 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 (Hahn, 2018). In contrast, standard PCA does not modeljointly models the covariance matrix due to population structure and sampling. On the other hand, for both data sets I investigated here, the matrix $\mathbf{F}_2$ of F-statistics estimated using admixtools2 is not a proper squared Euclidean distance matrix, i.e. it is not negative semidefinite and has imaginary PCs. This is not a practical when considering single F-statistics or PCA (for analyses here, I used a nearby matrix (Higham, 2002) with no apparent loss of precision). It does however mean that tools that use matrices of F-statistics, such as `qpadm` or `qpgraph` may be ill-calibrated, which may partly explain why they generally have poor out-of-sample predictive power and are restricted to a few dozen samples at a time. A model-based framework based on probabilistic PCA (Meisner *et al.*, 2021; Agrawal *et al.*, 2020; Hastie *et al.*, 2015) would likely be able to generate consistent F-statistics and PCs, while incorporating sampling error and missing data.

- weighting of SNPs
- estimation error
- propagating erros
- exploratory data analysis
- missing data
- population vs. sample allele frequencies

A Derivation

$$
\begin{aligned}
F_2(X_i, X_j) &= \sum_{l=1}^{L} \big((x_{il} - \mu_l) - (x_{jl} - \mu_l)\big)^2 = F_2(Y_i, Y_j) \\
&= \sum_{l=1}^{L} \big(\sum_k L_{kl} P_{ik} - \sum_k L_{kl} P_{kj}\big)^2 \\
&= \sum_{l=1}^{L} \left(\sum_k L_{kl}(P_{ik} - P_{jk})\right)^2 \\
&= \sum_{l=1}^{L} \left(\sum_k L_{kl}^2 (P_{ik} - P_{jk})^2 + 2 \sum_{k \neq k'} L_{kl} L_{k'l} (P_{ik} - P_{jk'})^2\right) \\
&= \sum_k \underbrace{\left(\sum_{l=1}^{L} L_{kl}^2\right)}_{1} (P_{ik} - P_{jk})^2 + \sum_{k \neq k'} \underbrace{\left(\sum_{l=1}^{L} L_{kl} L_{k'l}\right)}_{0} (P_{ik} - P_{jk'})^2 \\
&= \sum_k (P_{ik} - P_{jk})^2 \qquad (8)
\end{aligned}
$$

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

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

Lewontin, R. C. 1972. The Apportionment of Human Diversity. In T. Dobzhansky, M. K. Hecht, and W. C. Steere, editors, *Evolutionary Biology*, pages 381–398. Springer US, New York, NY

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

Barbujani, G., Magagni, A., Minch, E., and Cavalli-Sforza, L. L. 1997. An apportionment of human DNA diversity. Proceedings of the National Academy of Sciences, 94(9):4516–4519

Pritchard, J. K., Stephens, M., and Donnelly, P. 2000. Inference of population structure using multilocus genotype data. Genetics, 155(2):945–959

Higham, N. J. 2002. Computing the nearest correlation matrix–a problem from finance. IMA Journal of Numerical Analysis, 22(3):329–343

Rosenberg, N. A. and Nordborg, M. 2002. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nature Reviews Genetics, 3(5):380–390

Ramachandran, S., Deshpande, O., Roseman, C. C., Rosenberg, N. A., Feldman, M. W., and Cavalli-Sforza, L. L. 2005. 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

Rosenberg, N. A., Mahajan, S., Ramachandran, S., Zhao, C., Pritchard, J. K., and Feldman, M. W. 2005. Clines, Clusters, and the Effect of Study Design on the Inference of Human Population Structure. PLoS Genet, 1(6):e70

Patterson, N., Richter, D. J., Gnerre, S., Lander, E. S., and Reich, D. 2006. Genetic evidence for complex speciation of humans and chimpanzees. Nature, 441(7097):1103–1108

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., Indap, A., King, K. S., Bergmann, S., Nelson, M. R., Stephens, M., and Bustamante, C. D. 2008. Genes mirror geography within Europe. Nature, 456(7218):98–101

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

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

McVean, G. 2009. A genealogical interpretation of principal components analysis. PLoS genetics, 5(10):e1000686

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

François, O., Currat, M., Ray, N., Han, E., Excoffier, L., and Novembre, J. 2010. Principal Component Analysis under Population Genetic Models of Range Expansion and Admixture. Molecular Biology and Evolution, 27(6):1257–1268

Brisbin, A., Bryc, K., Byrnes, J., Zakharia, F., Omberg, L., Degenhardt, J., Reynolds, A., Ostrer, H., Mezey, J. G., and Bustamante, C. D. 2012. 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

Patterson, N. J., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., and Reich, D. 2012. Ancient Admixture in Human History. Genetics, page genetics.112.145037

Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C., and Foll, M. 2013. Robust demographic inference from genomic and SNP data

Jolliffe, I. T. 2013. *Principal Component Analysis*. Springer Science & Business Media

Lipson, M., Loh, P.-R., Levin, A., Reich, D., Patterson, N., and Berger, B. 2013. Efficient Moment-Based Inference of Admixture Parameters and Sources of Gene Flow. Molecular Biology and Evolution, 30(8):1788–1802

Malaspinas, A.-S., Tange, O., Moreno-Mayar, J. V., Rasmussen, M., DeGiorgio, M., Wang, Y., Valdiosera, C. E., Politis, G., Willerslev, E., and Nielsen, R. 2014. bammds: a tool for assessing the ancestry of low-depth whole-genome data using multidimensional scaling (MDS). Bioinformatics (Oxford, England), 30(20):2962–2964

Hastie, T., Mazumder, R., Lee, J. D., and Zadeh, R. 2015. Matrix completion and low-rank SVD via fast alternating least squares. The Journal of Machine Learning Research, 16(1):3367–3402

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

Leslie, S., Winney, B., Hellenthal, G., Davison, D., Boumertit, A., Day, T., Hutnik, K., Royrvik, E. C., Cunliffe, B., Lawson, D. J., Falush, D., Freeman, C., Pirinen, M., Myers, S., Robinson, M., Donnelly, P., and Bodmer, W. 2015. The fine-scale genetic structure of the British population. Nature, 519(7543):309–314

Peter, B. M. 2016. Admixture, Population Structure and F-Statistics. Genetics, page genetics.115.183913

Hahn, M. 2018. *Molecular Population Genetics*. Oxford University Press, Oxford, New York

Reich, D. 2018a. Opinion | How Genetics Is Changing Our Understanding of 'Race'. The New York Times

Reich, D. 2018b. *Who We Are and How We Got Here: Alte DNA und die neue Wissenschaft der menschlichen Vergangenheit.* Pantheon, New York, illustrated edition edition

Berg, J. J., Harpak, A., Sinnott-Armstrong, N., Joergensen, A. M., Mostafavi, H., Field, Y., Boyle, E. A., Zhang, X., Racimo, F., Pritchard, J. K., and Coop, G. 2019. Reduced signal for polygenic adaptation of height in UK Biobank. eLife, 8:e39725

Duncan, L., Shen, H., Gelaye, B., Meijsen, J., Ressler, K., Feldman, M., Peterson, R., and Domingue, B. 2019. Analysis of polygenic risk score usage and performance in diverse human populations. Nature Communications, 10(1):3328

Petr, M., Pääbo, S., Kelso, J., and Vernot, B. 2019. Limits of long-term selection against Neandertal introgression. Proceedings of the National Academy of Sciences, 116(5):1639–1644

Agrawal, A., Chiu, A. M., Le, M., Halperin, E., and Sankararaman, S. 2020. Scalable probabilistic PCA for large-scale genetic variation data. PLOS Genetics, 16(5):e1008773

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

Harney, E., Patterson, N., Reich, D., and Wakeley, J. 2021. Assessing the performance of qpAdm: a statistical tool for studying population admixture. Genetics, 217(4)

Meisner, J., Liu, S., Huang, M., and Albrechtsen, A. 2021. Large-scale Inference of Population Structure in Presence of Missingness using PCA. Bioinformatics (Oxford, England), page btab027

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