1 Introduction

Thanks to the widespread availability of cheap sequencing, genome-wide data sets now frequently incorporate tens of thousands of present-day individuals**?**. Furthermore, advances in ancient DNA techniques now allow large data sets. Using this data allows the inference of fine-scale population structure, but translating this wealth of data into meaningful and detailed models of population history is still a major challenge.

Particularly for the analysis of ancient DNA, two approaches have been proven to be particularly useful: one are global summary analyses, such as Structure (Pritchard *et al.*, 2000; Alexander *et al.*, 2009) Principal Component Analysis (PCA) (Cavalli-Sforza *et al.*, 1994; Reich *et al.*, 2008; Novembre *et al.*, 2008; McVean, 2009) and classical multidimensional scaling (MDS) **??**. Typically, these methods assume that population structure is *sparse*, so that a low-rank approximation with few underlying "components" is sufficient to model population structure See e.g. Engelhardt and Stephens (2010) for a useful perspective how these approaches are related.

Facing a novel data set, PCA or MDS are often the first analyses (beyond quality controls) a researcher performs, in order to obtain insights in the general population structure they are faced with. In order to answer more specific questions and to test specific hypotheses, the F-statistic framework of Patterson *et al.* (2012) has been proven particularly powerful (see also Peter (2016) for a more gentle introduction). In the F-statistic framework, usually only a small number of populations are used at once, to e.g. test for treeness and find closely related populations.

Even though these two approaches are considered in almost every ancient DNA paper, links between the inferences made from them are usually only compared qualitatively. In this paper, our goal is to show that PCA and F-statistics are in fact closely related by construction, and use a very similar summary of the data.

1.1 Introduction to F-statistics

F-statistics have been primarily motivated by trees and admixture graphs (Patterson *et al.*, 2012; Peter, 2016), but the calculations hold up in a much wider data space. In particular, Oteo-Garcia and Oteo (2021) provides a thorough introduction to interpreting F-statistics in the *data space* $\mathbb{R}^k$. Their work builds much of the foundation of this discussion, by demonstrating analogies to classical geometry. A brief summary of their key results: A population's allele frequencies can be thought of as vector in $\mathbb{R}^k$. Then, $F_2(X_1, X_2) = \|X_1 - X_2\|^2$ is the squared Euclidean distance between the populations with vectors 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 (scalar) product between these two vectors. Here, I will mainly use the F-statistic notation, but use the geometric notation where convenient.

2 Relationship of PCA, F_2 and Outgroup-F_3

The goal of this section is to give a cursory introduction to F-statistics, PCA and MDS, and to define notation. More detailed introductions are given in XXXXX.

2.1 Introduction to PCA

Let us assume we have some genotype data summarized in a matrix $\mathbf{X}$. Each of the k columns contains the allele-frequencies at a single SNP, and we have n rows; one corresponding to either a population or individual, depending on the desired resolution of the analysis. As a population may be represented by just one (pseudo-)haploid or diploid individual, there is no conceptual difference between these cases, but I will refer to populations as unit for analysis, for simplicity.

The goal of a PCA is to find a low-dimensional representation of the data that explains most of the variance-structure in the data (see Fig. 1 for an intuitive explanation).

There are several algorithms that are used to calculate a PCA in practice, the most common one relies on a singular value decomposition. Formally,

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

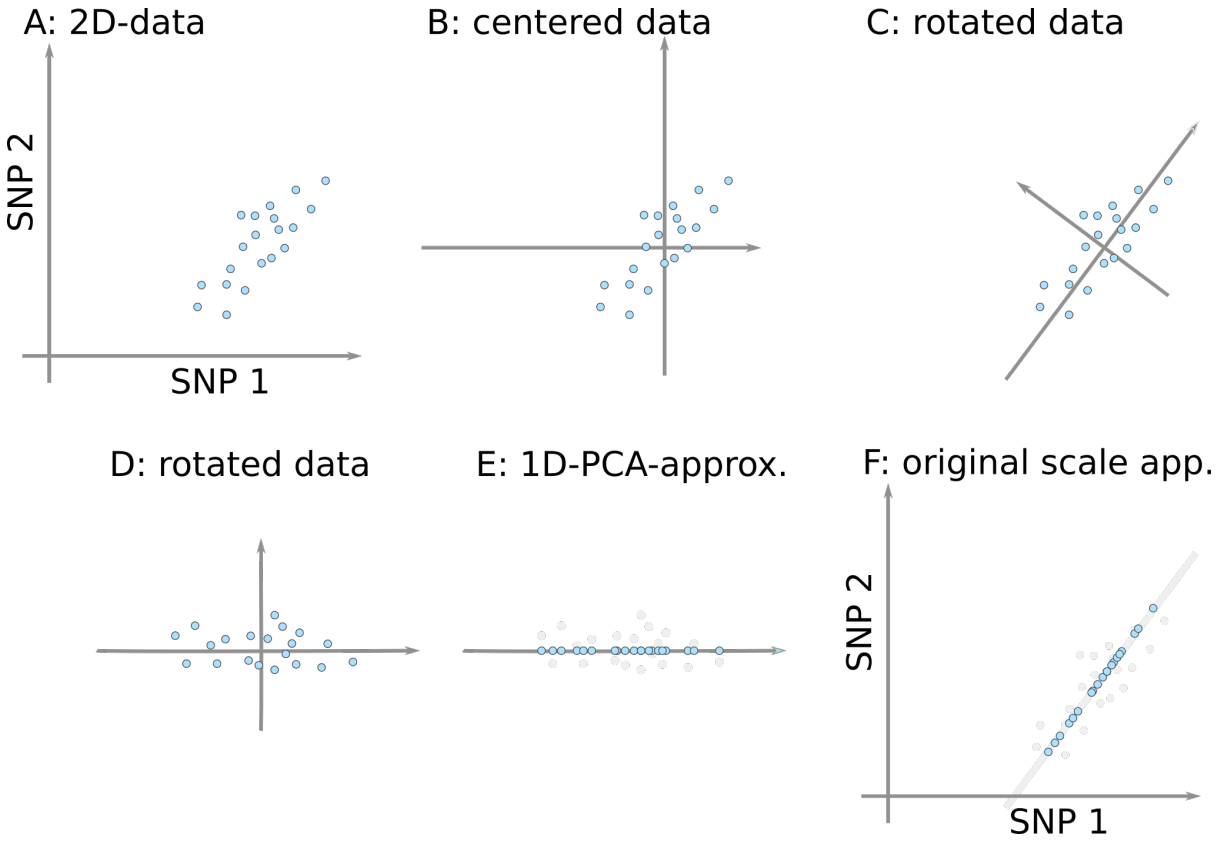

Figure 1: Basic Idea of PCA from 2D to 1D representation. A: Allele frequencies from different populations (blue dots) at two SNPs. A PCA is performed by centering the data (B), and rotating it (B) such that the first PC explains the majority of variation in the data, and the second PC is orthognal to the first, and explains the residual. A lower-dimensional approximation (in this case 1D) can be achieved by just keeping the first PC (E); which can be translated back to the original data space by inverting the rotation and centering (F).

Here, $\mathbf{C} = \mathbf{I} - \frac{1}{n}\mathbf{1}$ is a centering matrix that subtracts row means, with $\mathbf{I}, \mathbf{1}$ denoting 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 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 (François *et al.*, 2010, e.g).

In many implementations (Patterson *et al.*, 2006, e.g), SNPs are weighted by the inverse of their standard deviation. As this weighting makes little difference in practice, I will for now assume that SNPs are unweighted, and defer discussion of weighting to a later section.

Equivalently, we obtain the PCs by performing an eigendecomposition of the covariance matrix denoted as $\mathbf{K}$:

$$\mathbf{K} = \mathbf{Y}\mathbf{Y}^T = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T = \mathbf{U}\mathbf{\Sigma}^2\mathbf{U}^T = \mathbf{P}\mathbf{P}^T \tag{2}$$

where $\mathbf{\Lambda}$ is the diagonal matrix with the eigenvalues of $\mathbf{K}$. This algorithm does not compute the SNP-loadings. However, the i-th row of L can be obtained from $\mathbf{P}$ and the original data, whenever the eigenvalue $\lambda_i \neq 0$:

$$\mathbf{L}_i = \lambda_i^{-1}\mathbf{P}^T\mathbf{C}\mathbf{X}. \tag{3}$$

Let y_{il} denote the genotype of the i-th individual at the l-th SNP.

$$y_{il} = x_{il} - \mu_l \tag{4}$$

where μ_l is the mean genotype at the l-th locus.

2.2 PCA, F-statistics and midpoint rooting.

The construction of the PCs through the covariance matrix $\mathbf{K}$ is computationally more intensive than SVD, but it yields a simple connection to F_3-statistics.

Notice that the entries of the covariance matrix $\mathbf{K}$ are

$$k_{ij} = \sum_l y_{il}y_{jl} = \sum_l (x_{il} - \mu_l)(x_{yl} - \mu_l) = F_3(\boldsymbol{\mu}; X_i, X_j), \tag{5}$$

where $\boldsymbol{\mu} = (\mu_1, \dots \mu_k)$ is the vector of the mean allele frequency at each SNP. One interpretation is that $\boldsymbol{\mu}$ denotes a "pivot"-population whose allele frequency vector is the sample mean of the allele frequency space. By definition, it lies at the origin of the PC-space, and performing a PCA is equal to a rotation around this "pivot"-population.

One implication of this is that $\boldsymbol{\mu}$ strongly depends on sample composition. If we add many closely related populations (or, in an individual based framework,)This yields an (informal) analogy to the midpoint rooting of a tree in phylogenetics (see e.g. Felsenstein, 2004). In the absence of any further information, a sensible choice for the root of a tree is the point that is furthest away from all the tips.

Thus, we can also think of a PCA as performing a decomposition of an Outgroup-F_3-matrix, where we set the output to the mean of all marker allele frequencies. This is good practice if we do not have any outgroups available. However, we often also have outgroup data; for example when studying early European variation, Africans are sometimes a suitable outgroup. In this case, it would be consequential to perform an eigendecomposition of a true outgroup F_3-matrix.

note This could yield an (informal) analogy to the midpoint rooting of a tree in phylogenetics (see e.g. Felsenstein, 2004). In the absence of further information, a sensible choice for the root of a tree is the point that is furthest away from all the tips. However, if we do have better info (i.e. an outgroup), that could yield more sensible results. A caveat is that the eigencomposition of a non-centered matrix does not appear to give the desired properties – would need to dig substantially deeper into how pca maximises variance explains to come up with better algo.

2.3 PCA vs metric MDS on F-statistics

This is also a useful way to establish how we can obtain $\mathbf{P}$ from $\mathbf{F}_2$ directly: Note that the row and column means of $\mathbf{K}$ are zero:

$$\sum_i k_{ij} = \sum_i \sum_l (x_{il} - \mu_l)(x_{jl} - \mu_l) = \sum_l (x_{jl} - \mu_l) \left[\sum_i x_{il} - \mu_l \right] = 0.$$

Since $F_2(X_1, X_2) = F_2(X_1, \mu) + F_2(X_2, \mu) - 2F_3(\mu; X_1, X_2)$
$= \|X_1 - \mu\|^2 + \|X_2 - \mu\|^2 - 2\langle X_1 - \mu, X_2 - \mu\rangle$
(not sure if geometry notation would be easier here),

$$\mathbf{K} = \mathbf{CKC} = \frac{1}{2}\mathbf{C}\big[\mathbf{F}_2(X_1, \mu) + \mathbf{F}_2(X_2, \mu) - \mathbf{F}_2(X_1, X_2)\big]\mathbf{C} = -\frac{1}{2}\mathbf{C}\mathbf{F}_2(X_1, X_2)\mathbf{C} \tag{6}$$

since $\mathbf{C}\mathbf{F}_2(X_1, c)\mathbf{C} = 0$ for all constant c. Thus, by double-centering a matrix of $\mathbf{F}_2$-values (multiplied by $-\frac{1}{2}$), we can obtain $\mathbf{K}$ and hence $\mathbf{P}$. This is exactly what is done in classical multidimensional-scaling, and this derivation is merely a special case of a well-known method.

However, this result goes even further: consider any F_3-matrix, where the "focal"-population is kept constant:

$$\begin{aligned}\mathbf{C}\mathbf{F}_3(O; X_1, X_2)\mathbf{C} &= \frac{1}{2}\mathbf{C}\big[\mathbf{F}_2(X_1, O) + \mathbf{F}_2(X_2, O) - \mathbf{F}_2(X_1, X_2)\big]\mathbf{C} \\ &= -\frac{1}{2}\mathbf{C}\mathbf{F}_2(X_1, X_2)\mathbf{C}\end{aligned} \tag{7}$$

this shows that if we were to mean-center *any* F_3-matrix (a standard step in multidimensional scaling) before decomposition, we retain a PCA.

One important detail here are the diagonals of $\mathbf{F}_3$; above results assumes that $F_3(O, X_i, X_i) = F_2(O, X_i)$. This differs from how MDS has been sometimes applied on ancient DNA-data (**?**):

1. the off-diagonal entries are $1 - F_3(O; X_1, X_2)$ for some outgroup O.
2. the diagonal is 0.

Thus, this matrix differs from that derived above in that one has been added to all off-diagonal entries; and $F_2(X_1, X_2)$ has been subtracted from the diagonal. We have therefore

$$\mathbf{F}_3^{(Fu)} = \mathbf{1} - \mathbf{F}_3 + \mathbf{O}, \tag{8}$$

where $\mathbf{1}$ is a matrix of ones and $\mathbf{O}$ is a diagonal matrix with entries

$$o_{ii} = F_2(O, X_i) - 1$$

.

Centering then yields

$$\begin{aligned}-\mathbf{C}\mathbf{F}_3^{(Fu)}\mathbf{C} &= -\mathbf{C}\mathbf{1}\mathbf{C} + \mathbf{C}\mathbf{F}_3\mathbf{C} - \mathbf{C}\mathbf{O}\mathbf{C} \\ &= -\frac{1}{2}\big[\mathbf{C}\mathbf{F}_2\mathbf{C} + \mathbf{C}\mathbf{O}\mathbf{C}\big] \\ &= -\frac{1}{2}\big[\mathbf{C}(\mathbf{F}_2 + \mathbf{O})\mathbf{C}\big]\end{aligned} \tag{9}$$

open questions : How big is the effect of $\mathbf{O}$. Is it as expected cancelling out differences in sampling time? If so, are outgroup-F-stats generally preferable to PCA?

2.4 Projection using f-stats

Suppose we have a sample U we wish to project onto an existing PCA-basis made from $\mathbf{X}$, and let us assume we can compute $F_2(U, X_i)$ for all i. The "best" point i For any particular reference sample, F_2 places the point on the hypercircle with equation

$$F_2(U, X_i) = \sum_{k=1}^{n-1}(p_{ik} - u_k)^2, \tag{10}$$

where p_{ik} and u_k are the reference and unknown coordinate on the k-th component, respectively. It can be shown **?** that the u_k can be found using

$$\mathbf{u} = \frac{1}{2}\mathbf{\Lambda}^{-1}\mathbf{P}^T\mathbf{d} \tag{11}$$

where $\mathbf{d}$ is a column vector with

$$d_i = F_2(X_i, \boldsymbol{\mu}) - F_2(X_i, U)$$

.

Given a fixed projection, this allows us to also propagate uncertainty on a PCA plot.

Potential Application Use uncertainty on F_2/F_3 to propagate uncertainty on PCA-placement

3 F-stats in PCA-space

In the previous section, we showed that MDS and PCA are closely related to F-statistics, when we consider a matrix of many populations. However, one of the main advantages of the F-stats framework is that they can be used for specific hypotheses. Thus, in this section I am exploring how these hypotheses relate to the placement of populations on a PCA plot.

Recall that informally, PCA aims to reveal the axes of major variance. To do that, we find an optimal "rotation" of the data, such that these axes can be visualized.

As shown by e.g. Oteo-Garcia and Oteo (2021), F-stats can be thought of as inner products in Euclidean space, and F_2 is an (estimated) squared Euclidean distance between two populations in allele frequency space. PCA includes a translation and rotation of data, but a distance is invariant to both these operations. Hence, neither mean-centering, which is a translation nor PCA-rotation will change F_2. 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 the principal components $\mathbf{P}$. Formally,

$$\begin{aligned} F_2(X_i, X_j) = & \sum_{l=1}^{L} \left(x_{il} - x_{jl}\right)^2 - H_i - H_j \\ = & \sum_{l=1}^{L} \left((x_{il} - \mu_l) - (x_{jl} - \mu_l)\right)^2 - H_i - H_j = F_2(Y_i, Y_j) \\ = & \sum_{k} (P_{ik} - P_{jk})^2 - H_i - H_j = F_2(P_i, P_j), \end{aligned} \tag{12}$$

where H_i and H_j are the bias-correction terms proposed in Reich *et al.* (2009). A detailed derivation of this is given in Appendix A. As F_3 and F_4 can be written as sums of various F_2-terms, analogous relations apply.

3.1 F-stats in 2-dimensional PC-space

It is useful to consider the statistics on a PCA plot. The relationships we will discuss formally only hold in the full, n-dimensional PCA-space where we consider all principal components. Here, we start by discussing 2-dimensional spaces. This is useful for two reasons: for one, the geometry is simpler and we can think of circles and squares as opposed to hyperspheres and other high-dimensional geometric objects and thus help us build intuition. Second, in many applications it is argued that a 2-dimensional approximation is sufficient to explain the major components of population structure. In this case, the results here will hold under the same approximation assumptions in low-dimensional PCs; if they differ substantially from each other, it is likely that not sufficiently many PCs were considered.

3.1.1 F_2 in PC-space

The F_2-statistic as the squared Euclidean distance is the easiest to understand, it corresponds directly to the squared distance in PCA-space. This matches our intuition that closely related populations (which have low F_2) will be close to each other on a PCA-plot.

3.1.2 F_3 and circles

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 **?**McVean (2009).

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

$$\begin{aligned} 2F_3(X_x; X_1, X_2) &= 2\langle X_x - X_1, X_x - X_2\rangle \\ &= \|X_x - X_1\|^2 + \|X_x - X_2\|^2 - \|X_1 - X_2\|^2 \\ &< 0 \end{aligned} \tag{13}$$

By the Pythagorean theorem, $F_3 = 0$ iff X_1, X_2 and X_x form a right-angled triangle. Hence, the region where F_3 is zero is the circle with diameter through X_1 and X_2. If X_x lies inside this circle, the angle is obtuse and F_3 is negative, otherwise it will be positive. Similarly, if we fix X_1 and X_2 and ask where on a 2D-PCA-plot X_2 would lie, this space is defined by all the points for which the angle between X_1X_x and X_2X_x is obtuse.

This highlights a potential identifiability issue with F_3: Since all values of F_3 that result in the same projection will give the same value; and multiple admixture events may result in the same F_3-value.

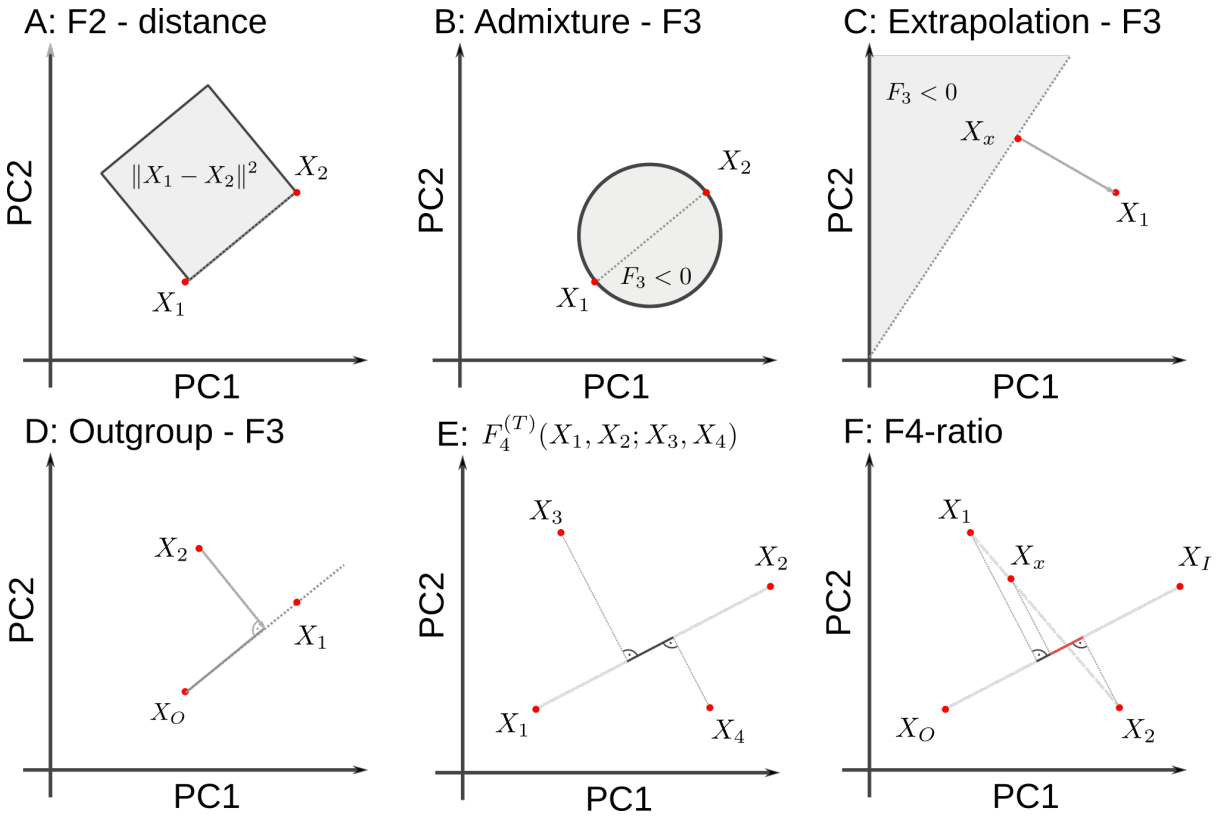

Figure 2: **Geometric representation of F-statistics on 2D-PCA-plot.** A: F_2 represents the squared Euclidean distance between two points in PC-space. B: Admixture-$F_3(X_x; X_1, X_2)$ is negative if X_x lies in the circle specified by the diameter $X_2 - X_1$
. C: $F_3(X_x; X_1, X_2)$ is negative given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

3.1.3 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.

3.1.4 F_4-ratio

$$\begin{aligned}\frac{F_4(X_I, X_O; X_X, X_1)}{F_4(X_I, X_O; X_2, X_1)} &= \frac{\|X_I - X_O\| \, \|X_X - X_1\| \cos(\alpha)}{\|X_I - X_O\| \, \|X_2 - X_1\| \cos(\beta)} \\ &= \frac{\|X_X - X_1\| \cos(\alpha)}{\|X_2 - X_1\| \cos(\beta)} \\ &= \frac{\|X'_X - X'_1\|}{\|X'_2 - X'_1\|} \end{aligned} \tag{14}$$

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 $X_O X_I$, i.e. $F_4(X_1, X'_1, X_2, X'_2) = 0$

3.2 spectral analysis of admixture statistics

1. split F-stats by PCA basis vector
2. same F-stat value may arise with different contribution from different PCs, should hint at distinct admixture events
3. can use clustering to infer shared history?
4. decomposition of admixture-F3?

4 Trees and admixture graphs in PCA-space

4.1 Trees

Evolutionary trees are fundamental in phylogenetic analyses, as they, on a large, scale, approximate how taxa diversify. Within a species, applying trees is also very common, but more problematic as populations frequently do not evolve as discrete lineages; instead, they admix and diversify as much more continuous processes. This is largely due to the time-scales involved, speciation events that give rise to trees might often be similarly messy, but from a distance of millions of years these issues might disappear.

Thus, when estimating trees from population genetic data, we must be very careful about whether the data is actually consistent with a tree, or belongs to some wider class of model.

Trees can be thought of as a collection of orthogonal dimensions; as drift on each branch is independent from every other branch. Thus, each sample is only

1. Trees
2. Admixture Graphs
3. Treelets
4. simple trees, admixture graph

5 other orthogonal bases

The most general "model"-space for (centered) SNP-data $\mathbf{Y}$ is $\mathbb{R}^k$, where we allow each SNP to be in its own dimension, and treat all dimensions as independent. However, since in most analyses the number of samples $n \ll k$, we can place all SNPs in a n-dimensional subspace $\mathbb{R}^n$. (Could be restricted further to $[0,1]^k$, but that does not appear to add much). If the data were normally distributed, $\mathbf{K}$ has a n-dimensional Wishart-distribution with k degrees of freedom. Since SNP are neither normal nor independent, the degrees of freedom might be considerably lower but we might still end up with something normally distributed.

6 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

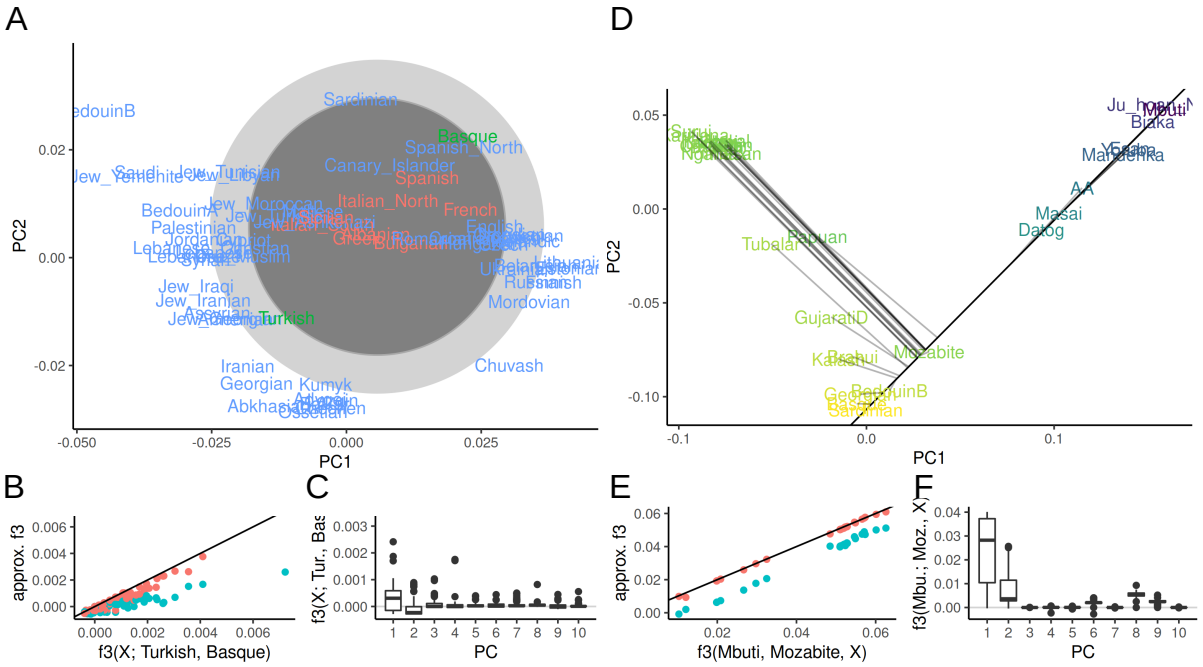

Figure 3: **PCA and F_3-statistics** A: PCA of Western Eurasian data; the circle denotes the region for which $F_3(X; \text{Basque}, \text{Turkish})$ may be negative. Populations for which F_3 is negative are colored in red. B, E: F_3 approximated with two (blue) and ten (red) PCs versus the full spectrum.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

7 Technical considerations

7.1 SNP weighting

It is clear that weighting SNP will have some effect on the resulting PCAs. Upweighting rare variants e.g. will emphasis recent events, as rare variance in the sample are more likely to be recent.

7.2 Missing data

7.3 F_2 error

In most F-statistics applications, F_2 is *estimated* using the minimum-variance unbiased estimator (Reich *et al.*, 2009)

$$f_2(X_1, X_2) = \frac{1}{L}\sum_l (x_{l1} - x_{l2})^2 - \frac{1}{L}\sum_l \frac{x_{l1}(1-x_{l1})}{n_{l1}-1} - \frac{1}{L}\sum_l \frac{x_{l2}(1-x_{l2})}{n_{l2}-1} - \tag{15a}$$

in contrast, as shown above, PCA can be thought as a decomposition of a matrix of uncorrected F_2-statistics:

$$F_2(X_1, X_2) = \frac{1}{L}\sum_l (x_{l1} - x_{l2})^2 \tag{15b}$$

This leads to some issues, for example trying to perform a PCA on the matrix of f_2-values is not positive semidefinite, and so some principal components may be imaginary. One possible resolution is probabilistic PCA (e.g. Engelhardt and Stephens, 2010; **?**).

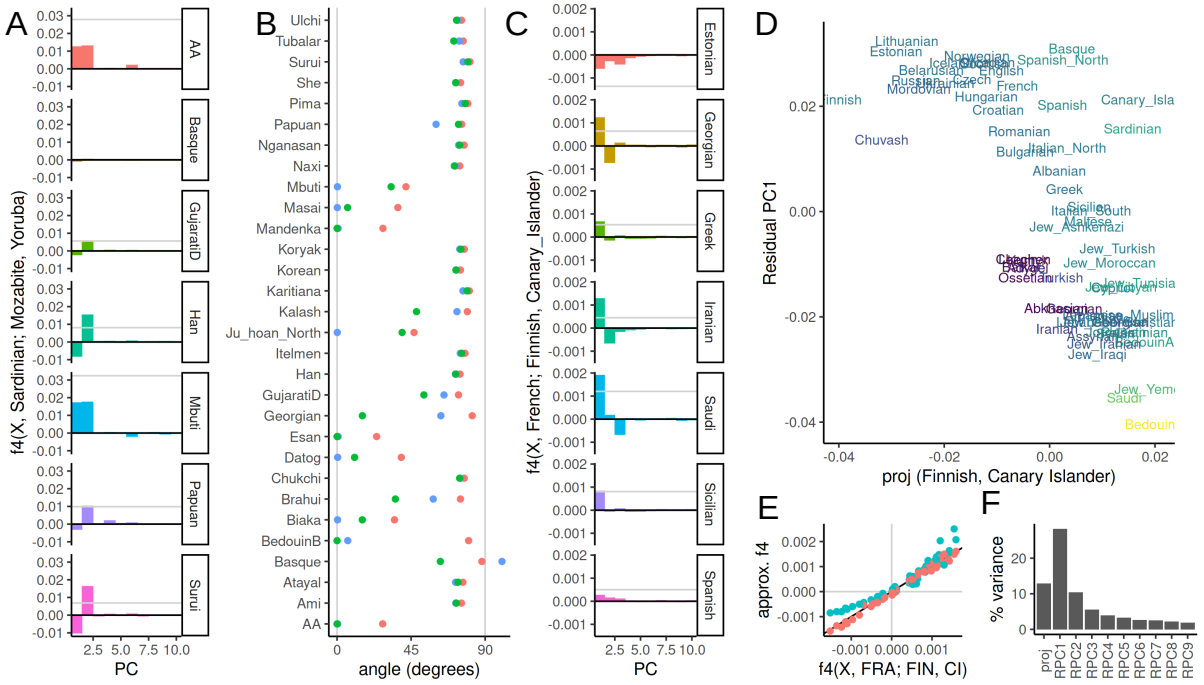

Figure 4: **PCA and F_4-statistics** A: Spectrum of select F_4-statistics in World data set. B: Projection angle representation of $F_4(X, Sardinian; Mozabite, Yoruba)$ (red) and approximations using two (blue) and ten (green) PCs. C: Spectrum of select $F-4$-statistics in Westeurasian data set. D: Scatterplot 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.

$$
\begin{aligned}
y_{ij}|\mathbf{P}_{ij}, \epsilon_i &= (\mathbf{PL})_{ij} + \epsilon_{ij} \\
x_i &\sim N(0, \mathbf{I}) \\
\epsilon_i &\sim N(0, \sigma^2\mathbf{I})
\end{aligned}
$$

7.4 What is a dimension?

A single population at a particular point in time can be thought of as a single point in allele-frequency space, given by it's p-dimensional locus of allele frequencies in that population. If this population evolves for some time in isolation, allele frequencies will change due to genetic drift; i.e. the population evolves along a single tree branch in the interpretation of Patterson *et al.* (2012). If we now add a second population, it will behave exactly the same, and the drift in the second population will be uncorrelated to the first, i.e. it evolves in a second dimension. 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.

8 outtakes

PCA from $\mathbf{X}$

$$\mathbf{K} = \mathbf{YY}^T = \mathbf{CXX}^T\mathbf{C} = \mathbf{PP}^T \tag{16}$$

A Derivation

$$
\begin{aligned}
F_2(X_i, X_j) &= \sum_{l=1}^{L} \big((x_{il} - \mu_l) - (x_{jl} - \mu_l)\big)^2 = F_2(Y_i, Y_j) \\
&= \sum_{l=1}^{L} \big(\sum_k L_{kl} P_{ik} - \sum_k L_{kl} P_{kj}\big)^2 \\
&= \sum_{l=1}^{L} \left(\sum_k L_{kl}(P_{ik} - P_{jk})\right)^2 \\
&= \sum_{l=1}^{L} \left(\sum_k L_{kl}^2 (P_{ik} - P_{jk})^2 + 2 \sum_{k \neq k'} L_{kl} L_{k'l} (P_{ik} - P_{jk'})^2\right) \\
&= \sum_k \underbrace{\left(\sum_{l=1}^{L} L_{kl}^2\right)}_{1} (P_{ik} - P_{jk})^2 + \sum_{k \neq k'} \underbrace{\left(\sum_{l=1}^{L} L_{kl} L_{k'l}\right)}_{0} (P_{ik} - P_{jk'})^2 \\
&= \sum_k (P_{ik} - P_{jk})^2 \qquad (17)
\end{aligned}
$$

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.

B Het term

$$
\begin{aligned}
H_i &= \sum_l X_{ij}(1 - X_{ij}) && (18) \\
&= \sum_j (Y_{ij} + \mu_j) - (Y_{ij} + \mu_j)^2 && (19) \\
&= \sum_j Y_{ij}(1 - 2\mu_j) - Y_{ij}^2 + \mu_j(1 - \mu_j) && (20) \\
&= \sum_j \left[(1 - 2\mu_j) \sum_k P_{ik} L_{kj} - \left(\sum_k P_{ik} L_{kj}\right)^2 + \mu_j(1 - \mu_j)\right] && (21) \\
&= \sum_k P_{ik} \sum_j (1 - 2\mu_j) L_{kj} - \sum_k P_{ik}^2 + M && (22) \\
& && (23)
\end{aligned}
$$

with $M = \sum_j \mu_j(1 - \mu_j)$

because the square terms

$$S = \sum_j \left(\sum_k P_{ik} L_{kj} \right)^2 \tag{24}$$

$$= \sum_j \left(\sum_k P_{ik} L_{kj} \right) \left(\sum_{k'} P_{ik'} L_{k'j} \right) \tag{25}$$

$$= \sum_j \left[\left(\sum_k P_{ik}^2 L_{kj}^2 \right) + \left(\sum_{k \neq k'} P_{ik} P_{ik'} L_{kj} L_{k'j} \right) \right] \tag{26}$$

$$= \left(\sum_k P_{ik}^2 \underbrace{\left[\sum_j L_{kj}^2 \right]}_{1} \right) + \left(\sum_{k \neq k'} P_{ik} P_{ik'} \underbrace{\left[\sum_j L_{kj} L_{k'j} \right]}_{0} \right) \tag{27}$$

$$= \sum_k P_{ik}^2 = \sum_k \lambda_k^2 V_{ik}^2 \tag{28}$$

References

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

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

Felsenstein, J. 2004. Inferring phytogenies. Sunderland, Massachusetts: Sinauer Associates

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

Reich, D., Price, A. L., and Patterson, N. 2008. Principal component analysis of genetic data. Nature Genetics, 40(5):491–492

Alexander, D. H., Novembre, J., and Lange, K. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome research, 19(9):1655–1664

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

Reich, D., Thangaraj, K., Patterson, N., Price, A. L., and Singh, L. 2009. Reconstructing Indian population history. Nature, 461(7263):489–494

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

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

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

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