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.

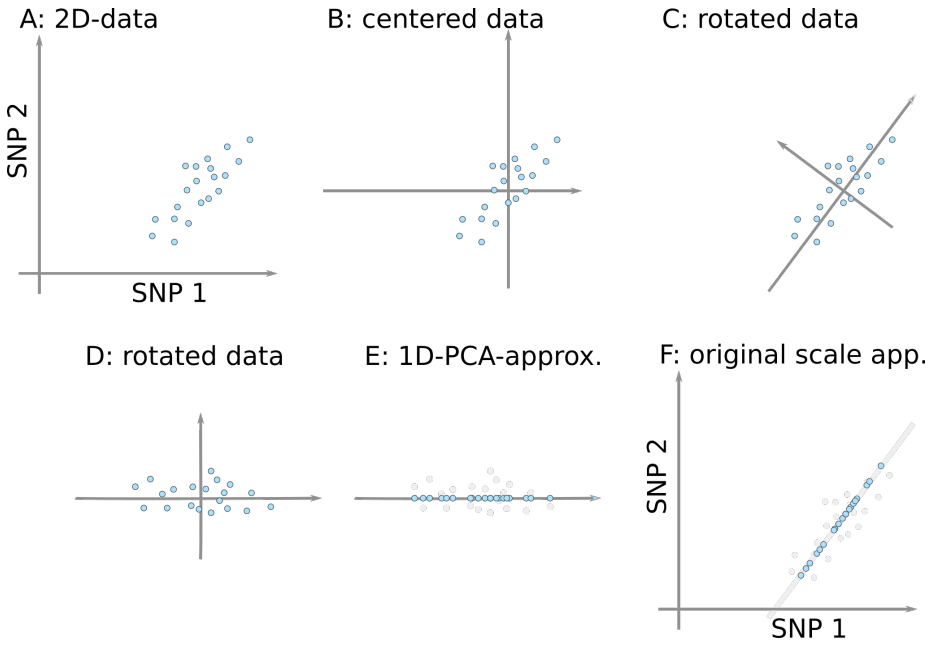

Figure 1: Basic Idea of PCA from 2D to 1D representation

1.2 Introduction to PCA

1.3 Introduction to MDS

2 Relationship of PCA, F_2 and Outgroup-F_3

Let us assume we have some genotype data $\mathbf{X}$, which contains allele-frequency data from n populations on the rows, and k loci on columns. Each population may be represented by a single haploid, pseudo-haploid or diploid individual. Thus, a PCA where each row represents a single individual Patterson *et al.* (2006) is included in this framework.

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. For this purpose, we rotate the data matrix around its center point, such that the first axis aligns with the major axis of variation.

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

Here, $\mathbf{C} = \mathbf{I} - \frac{1}{n}\mathbf{1}$ is a centering matrix and $\mathbf{I}, \mathbf{1}$ are the identity matrix and a matrix of ones, respectively. $\mathbf{P} = \mathbf{U\Sigma}$ is the matrix of principal components (PCs), and, $\mathbf{L} = \mathbf{V}^T$ is the matrix of loadings, respectively. Occasionally, SNPs are weighted by the square-root of their variance; for now we will assume the SNPs are unweighted, and defer discussion of weighting to a later section.

The principal components $\mathbf{P}$ are an orthogonal matrix of size $n \times n$ that are useful for understanding population structure, the loadings are an $n \times k$ orthonormal matrix that give the contribution of each SNP along each PC, it is often useful to look for outliers that might be indicative of selection (François

et al., 2010, e.g).

Equivalently, we obtain the PCs by performing an eigendecomposition of the covariance matrix of $\mathbf{Y}$, 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 of eigenvalues. In cases we are interested in the loadings, they can be obtained as

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

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.

The entries of the covariance matrix $\mathbf{K}$ are then

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

It is helpful to think of $\boldsymbol{\mu} = (\mu_1, \dots \mu_k)$ as a "fake"-individual whose allele frequency is the sample mean of the allele frequency space, it will always lie at the origin of the PC-space.

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.

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),

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

since $\mathbf{C}\mathbf{F}_2(X_1, X_3)\mathbf{C} = 0$ for all constant x_3. In fact, 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.

Thus, when we mean-center the $\mathbf{F}_2$-matrix, we subtract the F_2 terms in above equations; as they are invariant with respect to the column/ row means. What remains, is directly proportional to the matrix decomposed under a standard PC.

2.1 Outgroup-F_3-stats based MDS

If we have an MDS based on F_3,

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

$$= \frac{1}{2}\mathbf{C}\mathbf{F}_2(X_1, X_2)\mathbf{C} \tag{8}$$

In ancient DNA, a MDS has been proposed where

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

$$\mathbf{F}_3^{(Fu)} = \mathbf{1} - \mathbf{F}_3 + \mathbf{O}, \tag{9}$$

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

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

$$= \mathbf{C}\mathbf{F}_2\mathbf{C} + \mathbf{C}\mathbf{O}\mathbf{C} \tag{10}$$

$$\tag{11}$$

3 Approximating uncertainty on PCA-placement

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

$$F_2(U, X_i) = \sum_{k=1}^{n-1}(p_{ik} - u_k)^2, \tag{12}$$

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

$$\mathbf{u} = \frac{1}{2}\mathbf{\Lambda}^{-1}\mathbf{P}^T\mathbf{d} \tag{13}$$

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.

5 F-stats in PCA-space

The PCA aims to reveal the axes of major variance. The data is then "rotated" such that these axes can be visualized. Usually, only the first few PCs are considered.

However, as F_2 is a squared Euclidean distance of allele frequencies, it is invariant to rotation and translation. 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,

SHORT VERSION

$$\begin{aligned} F_2(X_i, X_j) = & \sum_{l=1}^{L} \left(x_{il} - x_{jl}\right)^2 - H_i - H_j \\ = & \sum_{l=1}^{L} \left((x_{il} - \mu_l) - (x_{jl} - \mu_l)\right)^2 - H_i - H_j = \quad F_2(Y_i, Y_j) \\ = & \sum_{k} (P_{ik} - P_{jk})^2 - H_i - H_j = \quad F_2(P_i, P_j) \end{aligned} \tag{14}$$

As F_3 and F_4 can be written as sums of various F_2-terms, analogous relations apply.

5.1 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.

5.1.1 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.

5.1.2 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

$$\begin{aligned} 2F_3(X_x; X_1, X_2) &= 2\langle X_x - X_1, X_x - X_2\rangle \\ &= \|X_x - X_1\|^2 + \|X_x - X_2\|^2 - \|X_1 - X_2\|^2 \\ &< 0 \end{aligned} \tag{15}$$

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.

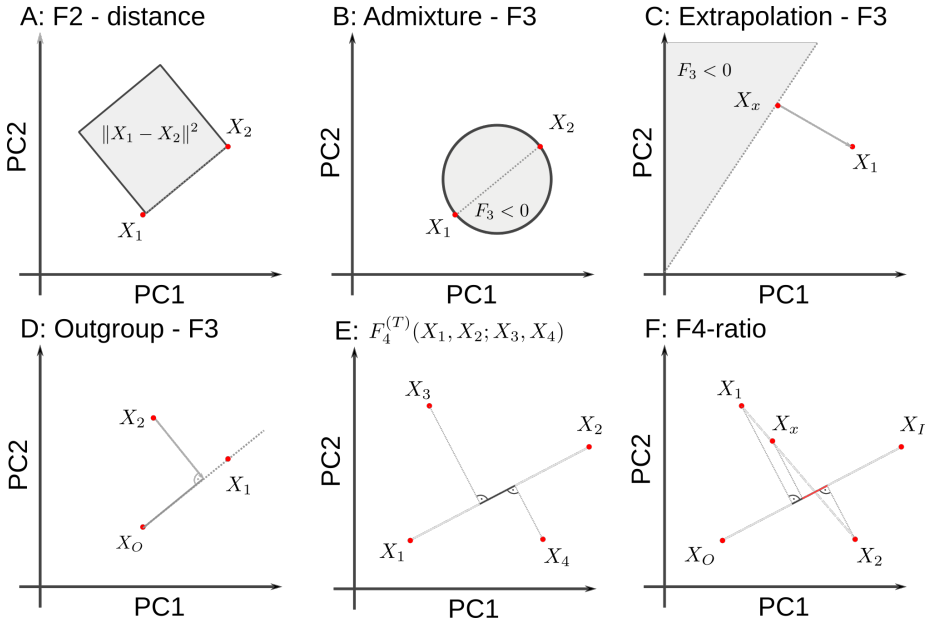

Figure 2: **Geometric representation of F-statistics on 2D-PCA-plot.** A: F_2 represents the squared Euclidean distance between two points in PC-space. B: Admixture-$F_3(X_x; X_1, X_2)$ is negative if X_x lies in the circle specified by the diameter $X_2 - X_1$
. C: $F_3(X_x; X_1, X_2)$ is negative given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

5.1.3 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.

5.1.4 F_4-ratio

$$\begin{aligned} \frac{F_4(X_I, X_O; X_X, X_1)}{F_4(X_I, X_O; X_2, X_1)} &= \frac{\|X_I - X_O\| \, \|X_X - X_1\| \cos(\alpha)}{\|X_I - X_O\| \, \|X_2 - X_1\| \cos(\beta)} \\ &= \frac{\|X_X - X_1\| \cos(\alpha)}{\|X_2 - X_1\| \cos(\beta)} \\ &= \frac{\|X'_X - X'_1\|}{\|X'_2 - X'_1\|} \end{aligned} \tag{16}$$

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. $F_4(X_1, X'_1, X_2, X'_2) = 0$

5.2 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?

6 Trees and admixture graphs in PCA-space

6.1 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

7 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.

8 Technical considerations

8.1 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.

8.2 Missing data

8.3 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.

9 outtakes

PCA from $\mathbf{X}$

$$\mathbf{K} = \mathbf{Y}\mathbf{Y}^T = \mathbf{C}\mathbf{X}\mathbf{X}^T\mathbf{C} = \mathbf{P}\mathbf{P}^T \tag{17}$$

A Derivation

$$
\begin{aligned}
F_2(X_i, X_j) &= \sum_{l=1}^{L} \big((x_{il} - \mu_l) - (x_{jl} - \mu_l)\big)^2 = F_2(Y_i, Y_j) \\
&= \sum_{l=1}^{L} \big(\sum_k L_{kl} P_{ik} - \sum_k L_{kl} P_{kj}\big)^2 \\
&= \sum_{l=1}^{L} \left(\sum_k L_{kl}(P_{ik} - P_{jk})\right)^2 \\
&= \sum_{l=1}^{L} \left(\sum_k L_{kl}^2 (P_{ik} - P_{jk})^2 + 2 \sum_{k \neq k'} L_{kl} L_{k'l} (P_{ik} - P_{jk'})^2\right) \\
&= \sum_k \underbrace{\left(\sum_{l=1}^{L} L_{kl}^2\right)}_{1} (P_{ik} - P_{jk})^2 + \sum_{k \neq k'} \underbrace{\left(\sum_{l=1}^{L} L_{kl} L_{k'l}\right)}_{0} (P_{ik} - P_{jk'})^2 \\
&= \sum_k (P_{ik} - P_{jk})^2 \qquad (18)
\end{aligned}
$$

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

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

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

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

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