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

Benjamin M Peter

September 8, 2021

Abstract

Human genetic diversity is shaped by our complex history. Population genetic tools to understand this variation can broadly be classified into data-driven methods such as Principal Component Analysis (PCA), and model-based approaches such as F-statistics. Here, I show that these two perspectives are closely related, and I derive explicit connections between the two approaches. I show that F-statistics have a simple geometrical interpretation in the context of PCA, and that orthogonal projections are the key concept to establish this link. I illustrate my results on two examples, one of local, and one of global human diversity. In both examples, I find that population structure is sparse, and only a few components contribute to most statistics. Based on these results, I develop novel visualizations that allow for investigating specific hypotheses, checking the assumptions of more sophisticated models. My results extend F-statistics to non-discrete populations, moving towards more complete and less biased descriptions of human genetic variation.

1 Introduction

As in other species, the genetic diversity of human populations is shaped by their environment and history over the last several hundred thousand years (e.g Cavalli-Sforza et al., 1994, Schraiber and Akey, 2015). In turn, studying the diversity patterns in humans can be used to reconstruct the demographic and evolutionary history of our species. As such, population genetic theory is essential to understand how this works.

processes Over time, genetic drift will slowly change allele frequencies and as a result, isolated populations are expected to slowly diverge. In humans, this may be caused because continental-scale geographic distances limit migration, causing a pattern known as isolation-by-distance. However, isolation-by-distance patterns are not uniform, but shaped by geography, particularly barriers to migration such as mountain ranges, oceans or deserts (Bradburd et al., 2013, Peter et al., 2020, Rosenberg et al., 2005). In addition, major historical population movements such as the out-of-Africa, Austronesian or Bantu expansions lead to more gradual patterns of genetic diversity over space (Cavalli-Sforza et al., 1994, Ramachandran et al., 2005, Novembre et al., 2008, Peter et al., 2020, Stoneking, 2016, Racimo et al., 2020). Local migration between neighboring populations will reduce differentiation, and long-distance migrations (Alves et al., 2016) or secondary contact and admixture between diverged populations, such as Neandertals and modern humans (Green et al., 2010) may lead to locally increased diversity.

Modern descriptions of human genetic diversity focus on the evolutionary processes that caused it, and which gives rise to both discrete and continuous components (Rosenberg et al., 2002, Serre and Pääbo, 2004, Rosenberg et al., 2005, Bradburd et al., 2018, Reich, 2018). The interplay of these demographic processes leads to the complex genetic structure that we observe in present-day people, and which we aim to reconstruct from genetic data, which is challenging since we cannot expect to devise a single model that captures all complexity.

Data-driven approaches One commonly used analysis paradigm is thus to integrate tools based on different sets of assumptions; each emphasizing a particular aspects of the data. Particularly for large data sets with several thousand samples, exploratory, data-driven methods such as population trees, Principal Component Analysis (PCA, Cavalli-Sforza et al. (1994)) structure (Pritchard et al., 2000) or multidimensional scaling (MDS, Malaspinas et al. (2014)) are a first step aimed at providing a simple, low-dimensional summary of the data while making minimal assumptions about the underlying population structure.

Model-based appraoches While PCA and related tools give a good overview of the data, they are not designed to answer specific research questions and cannot be used for model testing or parameter estimates. For this purpose, methods based on explicit demographic models are often used (Gutenkunst et al., 2009, Kamm et al., 2015, Excoffier et al., 2013). The drawback of these methods is that, to make inference mathematically feasible, we need to introduce strong modelling assumptions such as that populations are discrete, randomly mating, and at equilibrium, with the hope that despite these assumptions being violated, the resulting model fits provide a sufficiently accurate description to answer the research questions.

F-stats However, when the number of populations exceeds a few dozen, even codifying reasonable population models including all data can be prohibitively difficult. One approach is to pick a small set of “representative” samples, and restrict modeling to this subset. However, this has the drawback that a large proportion of the data may be unused. An increasingly popular alternative approach, particularly in the analysis of human ancient DNA, is therefore to build up complex models from smaller building blocks based on the relationship between two, three or four populations.

The framework is based on a set of parameters called F-statistics *sensu* Patterson (Reich et al., 2009, Patterson et al., 2012, Peter, 2016). I safe the formal definition for later; but the easiest way to motivate them assumes that populations are related as a tree, where the edges measure how much genetic drift has occurred. In this case, F_2 measures the genetic distance between populations, F_3 corresponds to an external, and F_4 to an internal branch of a tree) (Figure 2; Semple and Steel, 2003, Peter, 2016).

In most applications, these F-statistics are estimated from some data, and then used as tests of treeness. The canonical alternative model is an admixture graph or phylogenetic network (Patterson et al., 2012), which is a tree which allows for additional edges reflecting gene flow (Figure 3A). Typically, analyses proceed by testing specific hypotheses abot gene flow using F-statistics (Reich et al., 2009, Green et al., 2010, Lazaridis et al., 2014). Increasingly commonly, F-statistics are also incorporated into tools that jointly fit more complicated models (e.g. Patterson et al., 2012, Harney et al., 2021). However, admixture graphs are not the only alternative model that explain violations of treeness. Expected F-statistics can be calculated for a wide range of demographic models (Peter, 2016), or even in abstract Euclidean spaces (Oteo-Garcia and Oteo, 2021), so that model fit remains a major concern.

PCA With minimal assumptions, PCA is one of the most widely used data summary and data exploration techniques. One way to think about PCA is that it aims to create a low-dimensional summary of the data that retains a maximum of information about the underlying structure in the data (Again, a formal description will be given in the next section). It has been widely used and independently discovered in many disciplines. In population genetics, the use of PCA has been pioneered by Cavalli-Sforza et al. (1964), who used allele-frequency data at a population level to visualize global genetic diversity in the pre-genomic era (Cavalli-Sforza et al., 1994). Currently, PCA is most commonly performed on individual-level genotype data (e.g. Patterson et al., 2006, Novembre et al., 2008), making use of the hundreds of thousands of loci available in most genome-scale population genetic data set. This is most commonly done in early, exploratory phases of a study (Schraiber and Akey, 2015), as PCA can be used to e.g. remove outliers that do not cluster with their population (due to technical or biological reasons), but it remains an important part of many analyses. Theoretically, the PCA-decomposition has been studied in terms of trees (Cavalli-Sforza and Piazza, 1975), the coalescent (McVean, 2009) and discrete population models (**?**).

F-stats and PCA However, the connections between PCA and F-statistics remains poorly understood, and the goal of this paper is to investigate this connection. To do so, I use the geometric interpretation of F-statistics developed by Oteo-Garcia and Oteo (2021) who defined them in arbitrary Euclidean spaces. We make use of the empirical finding that while the allele-frequency space is frequently high-dimensional, for many empirical data sets, a low number of dimensions are sufficient to characterize population structure, and hence F-statistics. This allows for approximate interpretations of F-statistics on PCA plots, and I use two sample data sets to evaluate these approximations in practice.

2 Theory

In this section, I will introduce the mathematics and notations for F-statistics and PCA. A comprehensive treatise on PCA is given by e.g. Jolliffe (2013) and Pachter (2014), and a useful guide to interpretation is Cavalli-Sforza et al. (1994). Readers unfamiliar with F-statistics may find Patterson et al. (2012), Peter (2016) or Oteo-Garcia and Oteo (2021) helpful.

2.1 Formal Definition to F-statistics

Let us assume we have a set of populations for which we have SNP allele frequency data from S loci. Let x_{il} denote the frequency at the l-th SNP in the i-th population; and let $X_i = (x_{i1}, x_{i2}, \dots x_{iS})$ be a vector collecting all allele frequencies for population i. As X_i will be the only data summary for population i, I make no distinction between the population and the allele frequency vector used to represent it.

The three F-statistics can then be defined as

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

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

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

The normalization by the number of SNPs S is assumed to be the same for all calculations and is thus omitted subsequently. Both F_3 and F_4 can be written as sums of F_2-statistics:

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

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

As highlighted in the introduction, F-statistics have been primarily motivated in the context of trees and admixture graphs (Patterson et al., 2012). In a tree, the squared Euclidean distance $F_2(X_1, X_2)$ measures the length of all branches between populations X_1 and X_2 (Figure 2A); F_3 represents the length of an external branch (Figure 2B) and F_4 the length of an internal branch between four nodes, respectively (Figure 2C). Crucially, for branches that do not exist in the tree (as in Figure 2D), F_4 will be zero. The length of each branch can be thought of in units of genetic drift, and is non-negative (Patterson et al., 2012).

Thinking of F-statistics as branch lengths is useful to understand a number of applications: In particular, one common task is to find the population most closely related to an unknown sample X_U (Raghavan et al., 2014). One way to do that is using an *outgroup*-F_3-statistic $F_3(X_O; X_U, X_i)$, where X_O denotes an outgroup, and the X_i are a panel of populations that are candidates for the closest match. 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 intuition is given in Figure 1A; where the outgroup-F_3-statistic $F_3(X_O; X_U, X_3)$ is highlighted. It represents the length of the branch from X_O to the common node between the three samples in the statistic, and the closer this node is to X_U, the longer the branch and hence the larger the statistic. In contrast to a simple genetic distance, the sample time has no effect: The branch between X_U and X_2, would be shorter than to one between X_U and X_1, but the path to the shared junction and hence the F_3-statistic would be the same. Larger sets of F_3 and F_4-statistics are also 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 tests of treeness (Patterson et al., 2012): Negative F_3-values correspond to a branch with negative genetic drift, which is a violation of treeness. Similarly if four populations are related as a tree, then at least one of the F_4 statistics between the populations will be zero (Patterson et al., 2012). The most widely considered alternative model is an admixture graph (Patterson et al., 2012), an example is given in Figure 3A, where (the typically unobserved) population X_y is generated by a mixture of individuals from the ancestors of X_2 and X_3. Over time, genetic drift will change X_y to X_x, which is the population we observe. This will result in F_4-statistics that are non-zero, and, in some cases, in negative F_3-statistics (exact equations can be found in Peter, 2016). While admixture graphs are commonly considered, other alternative models exist. In (Peter, 2016), I showed that F_4 is sensitive to many non-symmetric population structure models. F_3 is more robust, but may also be negative if substantial population substructure exists.

2.1.1 Geometric interpretation of F-statistics

For very large data sets, it is hard to justify a specific demographic models that includes all data, but we will typically still be able to explore such data e.g. using PCA. One way to justify F-statistics is to interpret the populations X_i geometrically as points or vectors in the S-dimensionsl *data space* $\mathbb{R}^S$. Oteo-Garcia and Oteo (2021) showed that in this case, the F-statistics can be thought of as inner (or dot) products, and that all properties and tests related to treeness can be derived from this larger space. In particular the F-statistics can be defined as

$$F_2(X_1, X_2) = \frac{1}{S}\sum_{l=1}^{S}(x_{1l} - x_{2l})^2 = \frac{1}{S}\langle X_1 - X_2, X_1 - X_2\rangle = \frac{1}{S}\left\|X_1 - X_2\right\|^2 \quad \text{(3a)}$$

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

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

where $\|\cdot\|$ denotes the Euclidean norm and $\langle \cdot, \cdot \rangle$ denotes the dot product. Some elementary properties of the dot product between vectors a, b, c that I will use later are

$$\begin{aligned}
\langle a, b \rangle &= \sum_i a_i b_i && \text{(4a)}\\
\langle a, b \rangle &= \|a\| \, \|b\| \cos(\phi) && \text{(4b)}\\
\langle a, a \rangle &= \|a\|^2 && \text{(4c)}\\
\langle a + c, b \rangle &= \langle a, b \rangle + \langle b, c \rangle, && \text{(4d)}
\end{aligned}$$

where ϕ is the angle between a and b. Oteo-Garcia and Oteo (2021) showed that in this framework, F-statistics are closely related to vector projections

$$proj_b a = \frac{\langle a, b \rangle}{\|b\|^2} b, \tag{5}$$

which is a vector colinear to b whose length measures how much vector a points in the direction of b.

The drawback of the geometric approach of Oteo-Garcia and Oteo (2021) is that we have to deal with an very high-dimensional space, as the number of SNPs is frequently in the millions. However, it has been commonly observed that population structure is quite low-dimensional, with the first few PCs providing a good approximation of the underlying population structure. Therefore, we may hope that PCA could yield a reasonable approximation of the data space.

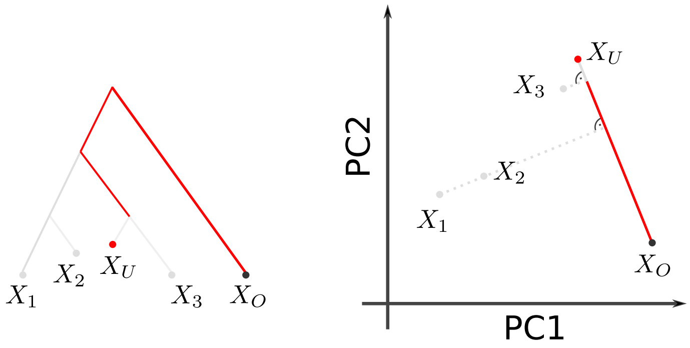

Figure 1: **Outgroup-F_3-statistics**

2.2 Formal Definition of PCA

PCA is a common way of summarizing genetic data, and so a large number of variation of PCA exist, e.g. in how SNPs are standardized, how missing data is treated or whether we use individuals or populations as units of analysis. The version of PCA I describe here is set up in a way that the similarities to the F-statistics-framework are maximized, and does *not* reflect how PCA is most commonly applied to genome-scale human genetic variation data sets. In particular, I assume that a PCA is performed on unscaled, estimated population allele frequencies, whereas many applications of PCA are based on individual-level sample allele frequency, scaled by the estimated standard deviation of each SNP (Patterson et al., 2006). The differences this causes will be addressed in the discussion.

Let us again assume we have genotype data as above, but let us now assume we aggregate the allele frequency vectors X_i 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 $n \times S$. 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*, which is a subset of $\mathbb{R}^S$.

One way one can define PCA is that it aims to find a K-dimensional subspace of the data space that retains the most important variation. K is at most $n - 1$, in which case the data is simply rotated. However, the historical processes that generated genetic variation often result in *low-rank* data (Engelhardt and Stephens, 2010), so that $K \ll n$ explains a substantial portion of the variation; for visualization $K = 2$ is frequently used.

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

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

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

PCA can then be written as

$$\mathbf{Y} = \mathbf{CX} = (\mathbf{U\Sigma})\mathbf{V}^T = \mathbf{PL}, \tag{6}$$

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. For any matrix $\mathbf{Y}$, we can perform a singular value decomposition $\mathbf{Y} = \mathbf{U\Sigma V}^T$ which, in the context of PCA, is interpreted as follows: The matrix of principal components $\mathbf{P} = \mathbf{U\Sigma}$ has size $n \times n$ and is contains information about population structure. The SNP loadings $\mathbf{L} = \mathbf{V}^T$ form an orthonormal basis of size $n \times S$, its rows give the contribution of each SNP to each PC. It is often used to look for outliers, which might be indicative of selection (e.g **?**). Alternatively, the PCs can also be motivated as an eigendecomposition of the covariance matrix $\mathbf{YY}^T$. This can be motivated from (6):

$$\mathbf{YY}^T = \mathbf{PLL}^T\mathbf{P}^T = \mathbf{PP}^T, \tag{7}$$

since $\mathbf{LL}^T = \mathbf{I}$.

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, which, for squared Euclidean (F_2)-distances, leads to an identical decomposition to PCA (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:

$$\mathbf{PP}^T = -\frac{1}{2}\mathbf{CF}_2\mathbf{C}. \tag{8}$$

2.3.2 F-statistics in PCA-space

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

$$\begin{aligned} F_2(X_i, X_j) = & \sum_{l=1}^{L} \left(x_{il} - x_{jl}\right)^2 \\ = & \sum_{l=1}^{L} \left((x_{il} - \mu_l) - (x_{jl} - \mu_l)\right)^2 = F_2(Y_i, Y_j) \\ = & \sum_{k=1}^{n} (p_{ik} - p_{jk})^2 = F_2(P_i, P_j), \end{aligned} \tag{9}$$

A derivation of this change-of-basis is given in Appendix A, Equation A1. As F_3 and F_4 can be written as sums of F_2-terms (Eqs. 2a, 2b), analogous relations apply.

In most applications, we do not use all PCs, but instead truncate to the first K PCs, which explain most of the between-population genetic variation. Thus,

$$\begin{aligned} F_2(P_i, P_j) = & \sum_{k=1}^{K} (p_{ik} - p_{jk})^2 + \sum_{k=K+1}^{n} (p_{ik} - p_{jk})^2 \\ = & \hat{F_2}^{(K)}(P_i, P_j) + \epsilon^{(K)}(P_i, P_j) \quad . \end{aligned} \tag{10}$$

In this notation, $\hat{F_2}^{(K)}$ is the approximation of F_2 with only the first K PCs considered, and $\epsilon^{(K)}$ is the corresponding approximation error. I will omit the superscript when the exact number of PCs is not relevant. If we sum up the squared approximation errors over all pairs of populations, we obtain

$$\sum_{i,j} \epsilon^{(K)}(P_i, P_j)^2 = \sum_{i,j} \left(\hat{F_2}^{(K)}(P_i, P_j) - F_2^{(K)}(P_i, P_j)\right)^2 = \left\|\mathbf{F}_2 - \hat{\mathbf{F}_2}\right\|_F^2, \tag{11}$$

where the Frobenius-norm $\|\cdot\|_F^2$ of a matrix is defined as the square root of the sum-of-squares of all its elements. This is precisely the function that is minimized in MDS (Jolliffe, 2013). In that sense, $\hat{\mathbf{F}_2}^{(K)}$ is the optimal low-rank approximation of $\mathbf{F}_2$ for any K in that it minimizes the sum of approximation errors of all F_2-statistics.

2.3.3 F-statistics and projection on PCA

One of the easiest ways of dealing with missing data in PCA is to calculate the principal components (equation 6) only on a subset of the data with no missingness, and then to *project* the lower quality samples with high missingness onto this PCA. The simplest way to do this is to note that

$$\mathbf{Y}\mathbf{L}^T = \mathbf{P}\mathbf{L}\mathbf{L}^T = \mathbf{P},$$

and so a new (centered) population Y_{new} can be projected onto an existing PC simply by post-multiplying it with $\mathbf{L}^T$:

$$P_{\text{proj}} = Y_{\text{new}}\mathbf{L}^T;$$

the k-th entry of P_{proj} gives the coordinates of the new sample on the k-th PC. However, it is likely that Y_{new} lies outside the variation of the original sample. In this case, there is a projection error

$$\|Y_{\text{new}} - P_{\text{proj}}\mathbf{L}\|^2 .$$

If we project with missing data, a similar projection can be used where we remove the rows from Y_{new} and $\mathbf{L}$ where data in Y_{new} is missing (Patterson et al., 2006).

Thus, if we compare the F-statistic of a projected sample, we have

$$F_2(X_i, X_{\text{new}}) = F_2(Y_i, Y_{\text{new}}) = F_2(P_i, P_{\text{proj}}) + F_2(P_{\text{proj}}\mathbf{L}, Y_{\text{new}}), \tag{12}$$

because the projection error and projection are orthogonal to each other.

3 Material & Methods

The theory outlined in the previous section suggests that F-statistics have a geometric interpretation in PCA-space, which can be approximated on PCA plots. In the next section I explore this connection in detail, and illustrate it on two sample data sets that I briefly introduce here. Both are based on the analyses by Lazaridis et al. (2014). The data is from the Reich lab compendium data set (v44.3), downloaded from `https://reich.hms.harvard.edu/allen-ancient-dna-resource-aadr-downloadable-` using data on the "Human Origins"-SNP set (597,573 SNPs). SNPs with missing data in any population are excluded. The code used to create all figures and analyses will be available on `https://github.com/BenjaminPeter/fstats_pca`.

"World" data set This data set is a subset of the "World Foci" data set of Lazaridis et al. (2014), where I removed samples which are not permitted for free reuse. These populations span the globe and roughly represents global human genetic variation (638 individuals from 33 population) As this data set is very sparse, with often thousands of kilometers between adjacent sampling locations, I speculate that gene flow between these populations may not be particularly common; and their structure may therefore be well-approximated by an admixture graph. A file with all individuals used and their assigned population is given in **Supplementary File 1.**

Western 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 motivated by the analysis of Lazaridis et al. (2014), who used it as a basis of comparison for ancient genetic analyses of Western Eurasian individuals, and PCAs based on similar sets of samples have been used in many other ancient DNA studies (e.g. ?Haak et al., 2015). Genetic differentiation in this region is low and closely mirrors geography (Novembre et al., 2008). We thus might think that gene flow between these populations is common (Ralph and Coop, 2013), and a discrete model such as a tree or an admixture graph might be a rather poor reflection of this data. A file with all individuals used and their assigned population is given in **Supplementary File 2.**

Computing F-statistics and PCA I perform analyses at the level of populations to ease presentation. It is an assumption of F-statistics that the genetic variation within sampled population is independent of the variation between samples (Patterson et al., 2012). All computations are performed in `R`. I use `admixtools 2.0.0` (`https://github.com/uqrmaie1/admixtools`) to compute F-statistics. To obtain a PC-decomposition, I first calculate all pairwise F_2-statistics, and then use equation 8 and the `eigen` function to obtain the PCs. A squared Euclidean distance matrix such as $\mathbf{F}_2$ has all negative or zero eigenvalues (i.e. $\mathbf{F}_2$ is negative-semidefinite). However as F_2-statistics are estimates, some eigenvalues might be slightly positive, which would lead to imaginary PCs. I avoid this by using the `nearPD`-function in R that ensures all eigenvalues have the correct sign.

4 Results

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 PCs. However, the appeal of PCA is that frequently, only a very small number $K \ll n$ of PCS contain most information that is relevant for population structure (for visualization $K = 2$ is often used).

4.1 F_2 in PC-space

The F_2-statistic is an estimate of the squared Euclidean distance between two populations. It thus corresponds to the squared distance between populations in PCA-space, and reflects the intuition

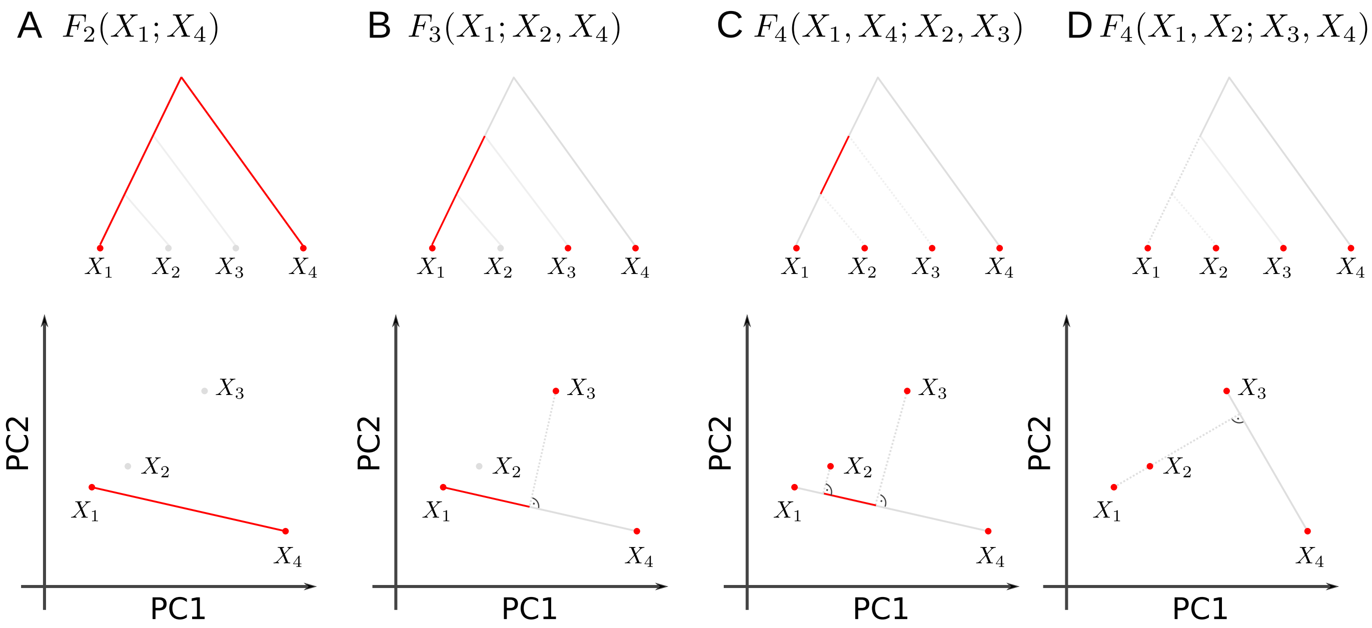

Figure 2: **Representation of F-statistics on tree and 2D-PCA-plot.** The schematics show four populations and their representation using a tree (top row) or a 2D-PCA plot. A: F_2 represents the (squared) Euclidean distance between two tree leafs, and in PC-space. B: $F_3(X_1; X_3, X_4)$ corresponds to the external branch from X_1 to the internal node joining the populations, and is proportional to the orthoginal projection of $X_1 - X_3$ onto X_1-X_4. C: $F_4(X_1, X_4; X_2, X_3)$ corresponds to the internal branch in the tree, or the orthogonal projection of $X_2 - X_3$ on $X_1 - X4$. D: $F_4(X_1, X_2; X_3, X_4)$ The two paths from X_1 to X_2 and X_3 and X_4 are non-overlapping in the tree, which corresponds to orthogonal vectors in PCA-space.

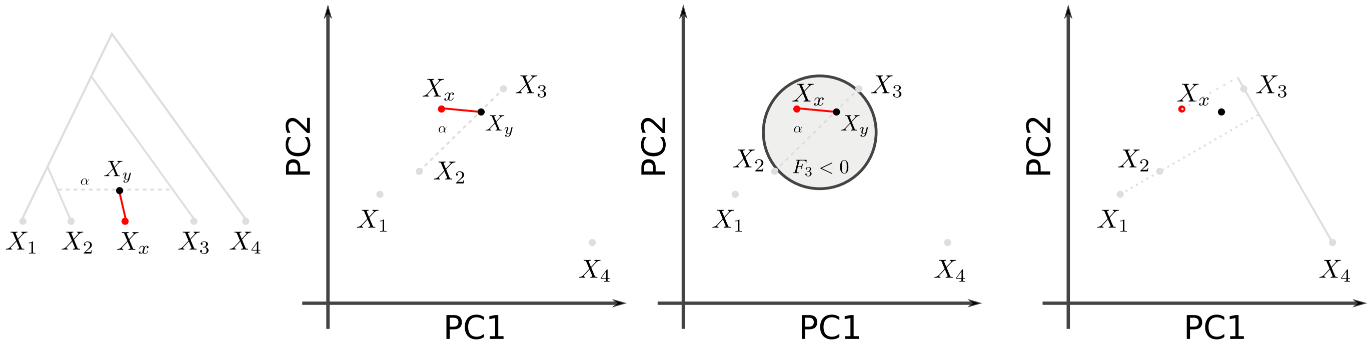

Figure 3: **Admixture representation on 2D-PCA-plot.** The schematics show four populations and their representation using an admixture graph (A) or a 2D-PCA plot. A: Admixture graph, with population X_y originating as an admixture of X_2 and X_3, with X_2 contributing proportion α. Subsequent drift (red branch) will change allele frequency to sampled admixture population X_x. B: PCA representation of the scenario in A. X_y originates on the segment connecting X_2 and X_3, and subsequent drift may move it in a random direction. C: $F_3(X_x; X_2, X_3)$ and negative region (light gray circle). $F_4(X_1, X_x; X_3, X_4)$ will no longer be zero (compare to Figure 2D).

that closely related populations will be close to each other on a PCA-plot, and have low pairwise F_2-statistics. In converse, if two populations have high F_2 but appear on the same point on an PCA-plot, this suggests that substantial variation is hidden on higher PCs, and we may want to investigate the PCs which contribute a large distance to this particular F_2-statistic in order to better understand and visualize their relationship.

4.2 When are admixture-F_3 statistics negative?

Consider the admixture scenario in Figure 3A, where population X_y is the result of a mixture of X_2 and X_3, and subsequent drift changes the allele frequencies of the admixed population from X_y to the sampled population X_x. How is such a scenario displayed on a PCA? ince the allele frequencies of X_y are a linear combination of X_2 and X_3, it will lie on the line segment connecting these two populations (Figure 3B). In fact, if X_x (and other populations related to X_x) was not part of the construction of the PCA, and is instead projected onto it, the drift from X_y to X_x will be independent of the population on the PCA, and X_x will be at the same location on a PCA plot as X_y, which can be used to predict the admixture proportions (Brisbin et al., 2012, McVean, 2009, Oteo-Garcia and Oteo, 2021).

If X_x, or populations more closely related to X_x than X_2 and X_3, are included in the construction of the PCA, this is no longer true, and X_x will project on a different spot than X_y (Figure 3B). Thus, a natural question to ask is given two source populations X_2, X_3, can we use PCA to predict which populations might be considered admixed between them?

One way to address this question is to consider the space for which F_3 is negative, i.e.

$$\begin{aligned} 2F_3(X_x; X_2, X_3) &= 2\langle X_x - X_2, X_x - X_3\rangle \\ &= \|X_x - X_2\|^2 + \|X_x - X_3\|^2 - \|X_2 - X_3\|^2 < 0. \end{aligned} \tag{13}$$

By the Pythagorean theorem, $F_3 = 0$ if and only if X_2, X_3 and X_x form a right-angled triangle. The associated region where $F_3 = 0$ is a n-sphere (or a circle in two dimensions) with diameter $\overline{X_2X_3}$ (The overline denotes a line segment). F_3 is negative when the triangle is obtuse, i.e. X_x is admixed if it lies inside the n-ball with diameter $\overline{X_2X_3}$ (Figure 2B, Equation A2).

F_3 on a 2D PCA-plot. If we project this n-ball on a two-dimensional plot, $\overline{X_2X_3}$ will usually not align with the PCs; thus the ball may be somewhat larger than it appears on the plot. This geometry is perhaps easiest visualized on a globe. If we look at the globe from a view point parallel to the equator, both the north and south poles are visible at the very edge of the circle. But if we look at it from above the north pole, the north- and south-poles will be at the very same point.

Thus if $\hat{F}_3 \ll F_3$, the true circle will be bigger than would be predicted from a 2D-plot. In this case, substantial relevant genetic differentiation is “hidden” in the higher PCs, and populations that appear inside the circle on a PCA-plot may, in fact, have positive F_3-statistics. This is because they are outside the n-ball in higher dimensions. The converse interpretation is more strict: if a population lies outside the circle on *any* 2D-projection, F_3 is guaranteed to be bigger than 0 (see Equation A4 in the Appendix).

Example As an example, I visualize the admixture statistic $F_3(X; \text{Basque}, \text{Turkish})$, on the first two PCs of the Western Eurasian data set (Figure 4A). In this case, the projected n-ball (light gray) and circle based on 2D (dark gray) align relatively closely, but several populations inside the ball (e.g. Sardinian, Finnish) have, in fact, positive F_3-values. This reveals that the first two PCs do not capture all the genetic variation relevant for Southern European population structure. Consequently, approximating F_3 by the first two or ten PCs (Figure 4B) only gives a coarse approximation of F_3, and from Figure 4C we see that many higher PCs contribute to F_3 statistics.

However, many populations, particularly from Western Asia and the Caucasus, fall outside the circle. This allows us to immediately conclude that their F_3-statistics must be positive; and we should not consider them as a mixture between Basques and Turks.

4.3 F_3-statistics as projections

Outgorup-F_3-statistic motivate a comparison of F_3-statistics to projections (Equation 5), consider again the case displayed in Figure 1A, where the goal is to find the population X_i that is closest to

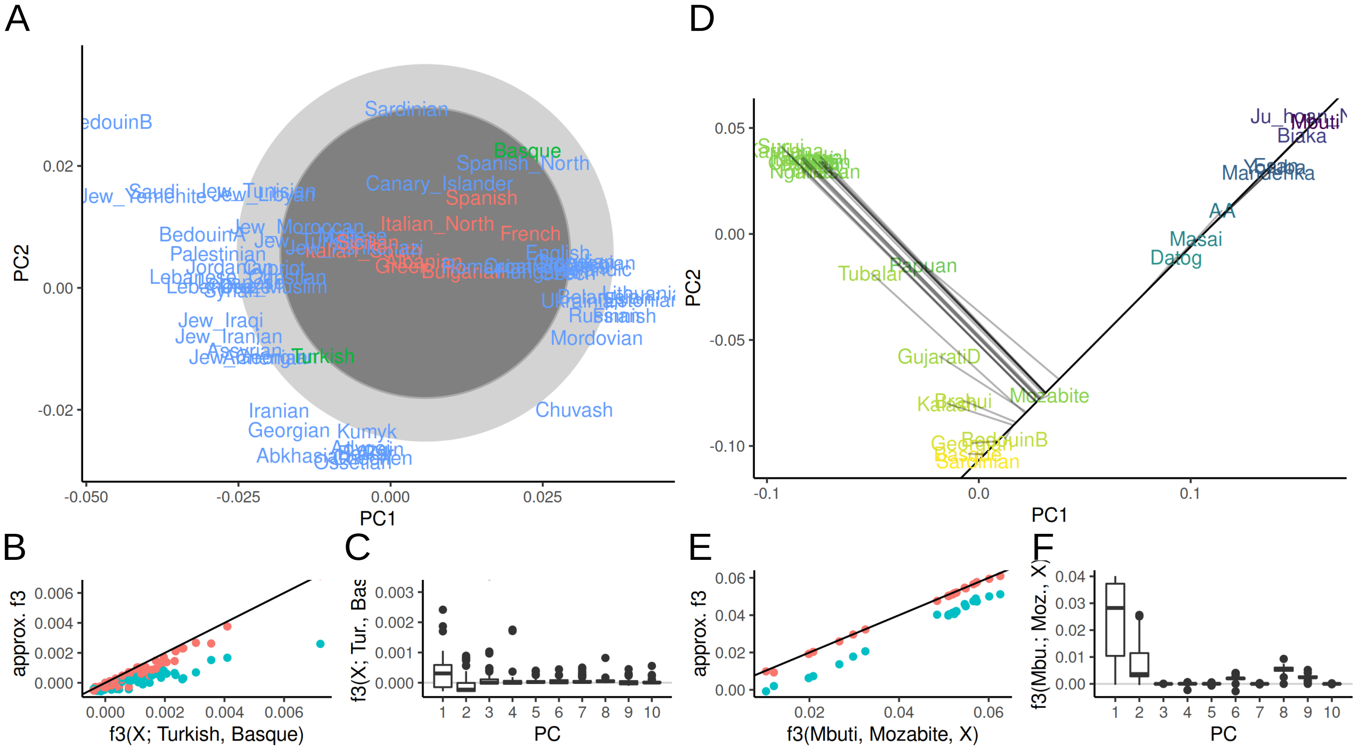

Figure 4: **PCA and F_3-statistics** A: PCA of Western Eurasian data; the circle denotes the region for which $F_3(X; \text{Basque}, \text{Turkish})$ may be negative. Populations for which F_3 is negative are colored in red. B, E: F_3 approximated with two (blue) and ten (red) PCs versus the full spectrum. 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.

an unknown population X_U, with respect to an outgroup X_O, which we can do using the statistic $F_3(X_O; X_U, X_i)$. On a PCA-plot, we can visualize this F_3-statistic as the projection of the vector $X_i - X_O$ onto $X_U - X_O$:

$$proj_{X_U - X_O} X_i - X_O = \frac{F_3(X_O; X_U, X_i)}{F_2(X_O; X_U)}(X_U - X_O).$$

Of the right-hand-side terms, $X_U - X_O$ gives the direction of the resulting vector, and the F_2-term in the denominator is a normalizing constant. Neiter of these terms depend on X_i, which justifies the argument that the F_3-statistic and length of the projected vector are proportional to each other, and can thus be interpreted similarly. Thus, the outgroup-F_3-statistic is larger for whichever X_i projects furthest along the axis from the outgroup to the unknown population; in the example in Figure 1B this is X_3.

Example Again, these projections will be orthogonal when using the full data, and may only be approximately orthogonal when approximated using the first few PCs. In Figure 4D, I visualize the outgroup-F_3-statistic $F_3(\text{Mbuti}; \text{Mozabite}, X_i)$, i.e. a statistic that aims to find the population most closely related to Mozabite (a Berber ethnic group from the northern Sahara), assuming the Mbuti are an outgroup. On a PCA, we can interpret this F_3 statistic as the projection of the line segment from Mbuti to population X_i onto the line through Mbuti and Mozabite (black line). For each population, the projection is indicated with a grey line. In the full data space, this line is always orthogonal to the segment Mbuti-Mozabite, but on the plot (i.e. the subspace spanned by the first two PCs), this is not necessarily the case. The coloring is based on the F_3-statistic calculated from all the data, with brighter values indicating higher F_3-statistics. In this case, the first two PCs approximate the F_3-statistic very well: Particularly the samples from East Asia, Siberia and the Americas project very close to orthogonally, suggesting that most of the genetic variation relevant for this analysis is captured by these first two PCs. We can quantify this and find that the first two PCs slightly underestimate the absolute value of F_3 (Figure 4E), but keep the relative ordering. I also find that many PCs, e.g. PCs 3-5, 7 and 10 have almost zero contribution to all F_3-statistics (Figure 4F), and PCs 6, 8 and 9 having a similar non-zero contribution for almost all statistics, likely because these PCs explain within-African variation.

4.4 F_4-statistics as angles

The interpretation of F_4 in PCA is similar to that of F_3 as a projection of one vector onto another, with the difference that now all four points may be distinct. F_4-statistics that correspond to a branch in a tree (as in Figure 2C), can be interpreted as being proportional to the length of a projected segment on a PCA plot (Figure 2G), again with the caveat that we need to scale it by a constant. If the F_4-statistic corresponds to a branch that does not exist in the tree, i.e it is a test statistic (Figure 2D), then, from the tree-interpretation, we expect $F_4(X_1, X_2; X_3, X_4) = 0$ implies that the vectors $X_1 - X_2$ and $X_3 - X_4$ are orthogonal to each other, or that the two populations map to the same point (Figure 2H). In the case of an admixture graph, this is no longer the case: Population X_x in Figure 3D does *not* map to the same point as X_1 or X_2 do, implying that statistics of the form $F_4(X_1, X_x; X_3, X_4) \neq 0$.

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

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

Of these three branches, the first and last are orthogonal to $X_1 - X_2$ and thus the F_4 statistic is just the internal branch of the tree (Figure 2F). It also suggests a number of diagnostic F-statistics that check assumptions; for example if the tree holds, then $F_4(X_3, X_3'; X_4, X_4') = 0$.

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

$$\mathrm{Cor}(X_1 - X_2, X_3 - X_4) = \frac{\langle X_1 - X_2, X_3 - X_4 \rangle}{\|X_1 - X_2\| \, \|X_3 - X_3\|} = \cos(\phi), \tag{14}$$

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

Example To illustrate the angle interpretation I again use the Western Eurasian data. The PCA-biplot shows two roughly parallel clines (Figure 4A), a European gradient (from Sardinian to Chuvash), and a Asian cline (from Arab to Caucasus populations). This is quantified in Figure 5A, where I plot the angle corresponding to $F_4(X, \mathrm{Sardinian}; \mathrm{Saudi}, \mathrm{Georgian})$. For most European populations, using two PCs (green points) gives an angle close to zero, corresponding to a correlation coefficient between the two clines of $r > 0.9$. Just adding PC3 (blue), however, shows that the parallelism of the clines is spurious: Using three PCs or the full data (red) shows that most correlations are low. I arrive at a similar interpretation from the spectrum of these statistics (Figure 5B), which has high loadings for the first three PCs, with minimal contributions from the higher ones.

4.5 Other projections

So far, I used eq. 9 to interpret F-statistics on a PC-plot, but the argument holds for *any* orthonormal transformation. This allows for a variety of visualizations that use both F-statistics and PCs. The motivation for this is that sometimes we wish to partition the variation in the data into a subspace of interest, and an orthogonal residual space that captures the information discarded. Examples where analyses are restricted to such subspaces include the F_4-ratio test (Patterson et al., 2012, Petr et al., 2019), `qpWave` (Skoglund et al., 2015) and `qpAdm` (Harney et al., 2021). For the F_4-ratio, for example, a ratio

$$\alpha = \frac{F_4(R_1, R_2; X, A)}{F_4(R_1, R_2; B, A)} \tag{15}$$

is used, which can be interpreted as projecting $X - A$ and $B - A$ onto $R_1 - R_2$. Thus, we can make a plot where we plot the variation on the X-axis along $R_1 - R_2$, and perform a PCA on the residual. This can be important because the residual can be used to check assumptions, e.g. $A - A'$ and $B - B'$ need to be orthogonal (Figure 2F).

4.5.1 Example

In the PCA on the world overview data set, I found a seemingly linear gradient from Africans to Europeans (Figure 4D). I focus on this cline using an alternative projection by using F-statistics of the form $F_4(X, Y; \text{Sardinian}, \text{Yoruba}))$, which might e.g. be used if we were to quantify gene flow associated with the out-of-Africa expansion. These F_4-statistics are very well-approximated by the first two PCs, with a 99.2% correlation between F_4 and its approximation using the first two PCs (Figure 5C).

In Figure 5D, I show the projection $\langle X; \text{Sardinian}, \text{Yoruba}\rangle$ on the X-axis, which means that $F_4(X, Y; \text{Sardinian}, \text{Yoruba})$ is is proportional to their horizontal distance between X and Y. The first two residual PCs are given on the Y-axis and in the coloring; this visualization reveals some variation within Africans (with Mbuti, Biaka and Ju|'hoansi) that is largely orthogonal to this gradient, as is the variation between Europeans, Asian and the Americans.

The percentage of between-population variance explained by the Sardinia-Yoruba axis (24%) is much lower than that of the first PC (40%, Figure 5E). However, the cumulative variance explained by the first two axes is similar, with (52%) explained when adding residual PC1 to the projection, compared to 55% for the first two PCs. The advantage of specifying one axis is that it displays the orthogonal components more explicitly, reveals distinct structure in Africans and non-Africans and thus can be used to test assumptions of more complex models.

5 Discussion

Particularly for the analysis of ancient DNA, F-statistics are a powerful tool to describe population genetic diversity. Here, I show that the geometry of F-statistics (Oteo-Garcia and Oteo, 2021) leads to a number of simple interpretations of F-statistics on a PCA plot. This allows for direct and quantitative comparisons between F-statistic-based results and PCA biplots. As PCA is often ran in an early step in data analysis, this also aids in generation of hypotheses that can be more directly evaluated using generative models, (e.g using a lower number of populations). It also allows reconciling apparent contradictions between F-statistics and PCA-plots; differences between the two data summaries are explained solely by higher PCs, and so whenever such contradictions arise, higher PCs will be informative for population structure. Previous interpretation of PCA in the context of population genetic models have primarily focused on the PCs, which can be derived analytically for trees (Cavalli-Sforza and Piazza, 1975) and homogeneous spatial models (Novembre and Stephens,

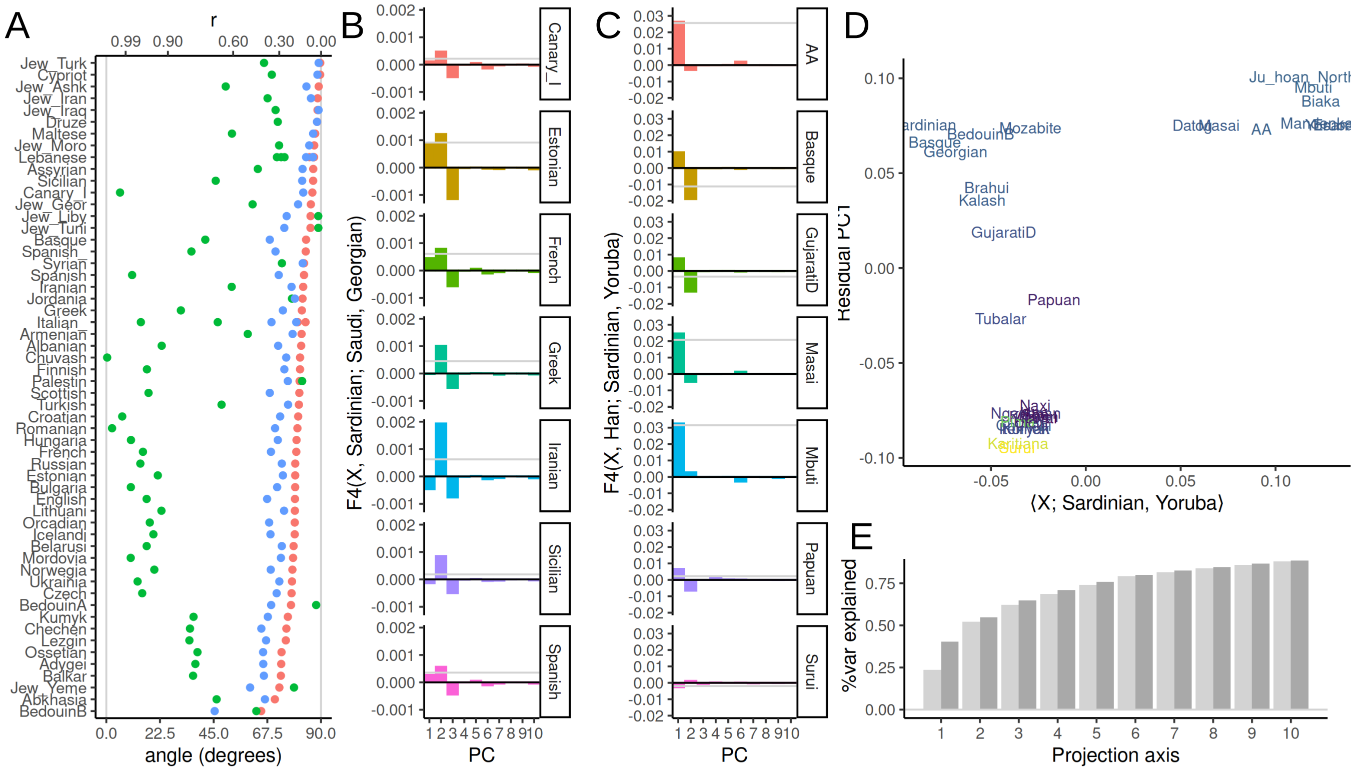

Figure 5: **PCA and F_4-statistics** A: Projection angle and correlation coefficient r representation of $F_4(X, Sardinian; Saudi, Georgian)$ (red) in the Western Eurasian data set, and approximations using two (green) and three (blue) PCs. B: Spectrum of select F_4-statistics in the Western Eurasian data set. C: Spectrum of $F - 4$-statistics in World data set. D: Scatterplot of F_3-projection on Sardinian-Yoruba axis and residual PC1. E: Percent variance explained from of the projection based on F_3 in panel D and first nine residual PCs (light gray), compared with percent variance explained by first ten PCs (dark gray).

2008). My interpretation here is different in that it puts more emphasis on the geometry itself, rather than directly interpreting the PCs. One consequence is that the results here are not impacted by sample ascertainment and sample sizes (McVean, 2009, Novembre and Stephens, 2008), which are common concerns in the interpretation of PCA. However, a very skewed sampling distribution will increases the likelihood that more or different PCs will have to be included in the analysis. From this perspective, one could envision a framework where F-statistics are used to decide which samples should be included to obtain a low-dimensional PCA-plot "representative" of the data

As F-statistics are motivated by trees, they assume that populations are discrete, related as a graph, and that gene flow between populations is rare (Patterson et al., 2012, Harney et al., 2021). However, in many regions, all humans populations are admixed to some degree (Pickrell and Reich, 2014), and in regions such as Europe, genetic diversity is distributed continuously (Novembre et al., 2008, Novembre and Stephens, 2008). This provides a challenge for interpretation; as many F_3 and F_4 statistics may indicate gene flow. In my example (Figure 4A), most Southern European populations are "admixed" between Basques and Turkish, but a more accurate model might be one of continuous variation where Basque and Turkish lie on one of multiple gradients; which is more directly visualized with PCA. There are a number of tools that have been developed that use multiple F-statistics to build complex models, such as `qpGraph` (Lazaridis et al., 2014) and `qpAdm` (Harney et al., 2021). One issue with these approaches is that they are usually restricted to at most a few dozen populations. As ancient DNA data sets now commonly include thousands of individuals, analysts are faced with the challenge of which data to include. A common approach is to sample a large number of distinct models, and retain the ones that are compatible with the data. However, as both `qpGraph` and `qpAdm` assume that gene flow is rare and discrete, selecting sets of populations that did experience little gene flow will provide good fits. One example of this is the world foci data set used here, which contains only 33 populations from across the world, and which is well-approximated by two PCs. However, this ascertainment misses a large amount of variation; a more dense sampling would show that in many places human genetic diversity is very gradual and multi-layered (Lazaridis et al., 2014, Peter et al., 2020). The PCA-based interpretation offers an alternative that trades interpretability for robustness. Particularly interpreting a (normalized) F_4-statistic as a correlation coefficient translates to generalized models of gene flow. Separating F-statistics in a sum of model and residuals, and performing a PCA on the latter (such as in Figure 5D) is another way how we can visualize F-statistics and evaluate the model fit.

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

A second difference is that frequently, rare SNPs are weighted higher in PCA, whereas all SNPs are weighted the same for F-statistics (Patterson et al., 2006, 2012). This is a difference of convention (Cavalli-Sforza and Piazza, 1975); since F-statistics are summed over SNPs with the same expectation, F-statistics could also be calculated using the same weighting. The close connection between the two approaches developed here suggest that for most analyses, users might want to be consistent and use the same weighting for both types of analyses.

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

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

References

Aman Agrawal, Alec M. Chiu, Minh Le, Eran Halperin, and Sriram Sankararaman. Scalable probabilistic PCA for large-scale genetic variation data. *PLOS Genetics*, 16(5):e1008773, 2020. ISSN 1553-7404. URL `https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1008773`.

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

Gideon S. Bradburd, Peter L. Ralph, and Graham M. Coop. Disentangling the Effects of Geographic and Ecological Isolation on Genetic Differentiation. *Evolution*, 67(11):3258–3273, 2013. ISSN 1558-5646. URL `http://onlinelibrary.wiley.com/doi/10.1111/evo.12193/abstract`.

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

Abra Brisbin, Katarzyna Bryc, Jake Byrnes, Fouad Zakharia, Larsson Omberg, Jeremiah Degenhardt, Andrew Reynolds, Harry Ostrer, Jason G. Mezey, and Carlos D. Bustamante. PCAdmix: Principal Components-Based Assignment of Ancestry along Each Chromosome in Individuals with Admixed Ancestry from Two or More Populations. *Human biology*, 84(4):343–364, August 2012. ISSN 0018-7143. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3740525/`.

L. L. Cavalli-Sforza and A. Piazza. Analysis of evolution: Evolutionary rates, independence and treeness. *Theoretical Population Biology*, 8(2):127–165, October 1975. ISSN 0040-5809. URL `http://www.sciencedirect.com/science/article/pii/0040580975900295`.

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

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

Barbara E. Engelhardt and Matthew Stephens. Analysis of Population Structure: A Unifying Framework and Novel Methods Based on Sparse Factor Analysis. *PLoS Genet*, 6(9):e1001117, September 2010. URL `http://dx.doi.org/10.1371/journal.pgen.1001117`.

Laurent Excoffier, Isabelle Dupanloup, Emilia Huerta-Sánchez, Vitor C. Sousa, and Matthieu Foll. Robust Demographic Inference from Genomic and SNP Data. *PLOS Genetics*, 9(10):e1003905, October 2013. ISSN 1553-7404. URL `https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1003905`.

J Felsenstein. Maximum-likelihood estimation of evolutionary trees from continuous characters. *American Journal of Human Genetics*, 25(5):471–492, September 1973. ISSN 0002-9297. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1762641/`.

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

J. C. Gower. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53(3-4):325–338, December 1966. ISSN 0006-3444. URL https://doi.org/10.1093/biomet/53.3-4.325.

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

Ryan N. Gutenkunst, Ryan D. Hernandez, Scott H. Williamson, and Carlos D. Bustamante. Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLoS Genet*, 5(10):e1000695, October 2009. URL http://dx.doi.org/10.1371/journal.pgen.1000695.

Wolfgang Haak, Iosif Lazaridis, Nick Patterson, Nadin Rohland, Swapan Mallick, Bastien Llamas, Guido Brandt, Susanne Nordenfelt, Eadaoin Harney, Kristin Stewardson, Qiaomei Fu, Alissa Mittnik, Eszter Bánffy, Christos Economou, Michael Francken, Susanne Friederich, Rafael Garrido Pena, Fredrik Hallgren, Valery Khartanovich, Aleksandr Khokhlov, Michael Kunst, Pavel Kuznetsov, Harald Meller, Oleg Mochalov, Vayacheslav Moiseyev, Nicole Nicklisch, Sandra L. Pichler, Roberto Risch, Manuel A. Rojo Guerra, Christina Roth, Anna Szécsényi-Nagy, Joachim Wahl, Matthias Meyer, Johannes Krause, Dorcas Brown, David Anthony, Alan Cooper, Kurt Werner Alt, and David Reich. Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature*, 522(7555):207–211, June 2015. ISSN 0028-0836. URL http://www.nature.com/nature/journal/v522/n7555/full/nature14317.html.

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

Eadaoin Harney, Nick Patterson, David Reich, and John Wakeley. Assessing the performance of qpAdm: a statistical tool for studying population admixture. *Genetics*, 217(4), April 2021. ISSN 1943-2631. URL https://doi.org/10.1093/genetics/iyaa045.

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

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

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

Iosif Lazaridis, Nick Patterson, Alissa Mittnik, Gabriel Renaud, Swapan Mallick, Karola Kirsanow, Peter H. Sudmant, Joshua G. Schraiber, Sergi Castellano, Mark Lipson, and others. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*, 513(7518):409–413, 2014. URL http://www.nature.com/nature/journal/v513/n7518/abs/nature13673.html.

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

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

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

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

J. Novembre and M. Stephens. Interpreting principal component analyses of spatial population genetic variation. *Nature genetics*, 40(5):646–649, 2008. URL `http://www.nature.com/ng/journal/v40/n5/abs/ng.139.html`.

John Novembre, Toby Johnson, Katarzyna Bryc, ZoltÃ¡n Kutalik, Adam R Boyko, Adam Auton, Amit Indap, Karen S King, Sven Bergmann, Matthew R Nelson, Matthew Stephens, and Carlos D Bustamante. Genes mirror geography within Europe. *Nature*, 456(7218):98–101, 2008. URL `http://www.ncbi.nlm.nih.gov/pubmed/18758442`.

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

Lior Pachter. What is principal component analysis?, May 2014. URL `https://liorpachter.wordpress.com/2014/05/26/what-is-principal-component-analysis/`.

Nick Patterson, Daniel J. Richter, Sante Gnerre, Eric S. Lander, and David Reich. Genetic evidence for complex speciation of humans and chimpanzees. *Nature*, 441(7097):1103–1108, June 2006. ISSN 0028-0836. URL `http://www.nature.com/nature/journal/v441/n7097/abs/nature04789.html`.

Nick J. Patterson, Priya Moorjani, Yontao Luo, Swapan Mallick, Nadin Rohland, Yiping Zhan, Teri Genschoreck, Teresa Webster, and David Reich. Ancient Admixture in Human History. *Genetics*, page genetics.112.145037, September 2012. ISSN 0016-6731, 1943-2631. URL `http://www.genetics.org/content/early/2012/09/06/genetics.112.145037`.

Benjamin M. Peter. Admixture, Population Structure and F-Statistics. *Genetics*, page genetics.115.183913, January 2016. ISSN 0016-6731, 1943-2631. URL `http://www.genetics.org/content/early/2016/02/03/genetics.115.183913`.

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

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

Joseph K. Pickrell and David Reich. Toward a new history and geography of human genes informed by ancient DNA. *Trends in Genetics*, 30(9):377–389, September 2014. ISSN 0168-9525. URL `http://www.sciencedirect.com/science/article/pii/S0168952514001206`.

Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000. URL http://www.ncbi.nlm.nih.gov/pubmed/10835412.

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

Maanasa Raghavan, Pontus Skoglund, Kelly E. Graf, Mait Metspalu, Anders Albrechtsen, Ida Moltke, Simon Rasmussen, Thomas W. Stafford Jr, Ludovic Orlando, Ene Metspalu, and others. Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature*, 505(7481): 87–91, 2014. URL http://www.nature.com/nature/journal/v505/n7481/abs/nature12736.html.

Peter Ralph and Graham Coop. The Geography of Recent Genetic Ancestry across Europe. *PLoS Biol*, 11(5):e1001555, May 2013. URL http://dx.doi.org/10.1371/journal.pbio.1001555.

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

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

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

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

Noah A Rosenberg, Saurabh Mahajan, Sohini Ramachandran, Chengfeng Zhao, Jonathan K Pritchard, and Marcus W Feldman. Clines, Clusters, and the Effect of Study Design on the Inference of Human Population Structure. *PLoS Genet*, 1(6):e70, December 2005. URL http://dx.plos.org/10.1371/journal.pgen.0010070.

Joshua G. Schraiber and Joshua M. Akey. Methods and models for unravelling human evolutionary history. *Nature Reviews Genetics*, 2015. URL http://www.nature.com/nrg/journal/vaop/ncurrent/full/nrg4005.html.

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

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

Pontus Skoglund, Swapan Mallick, Maria Cátira Bortolini, Niru Chennagiri, Tábita Hünemeier, Maria Luiza Petzl-Erler, Francisco Mauro Salzano, Nick Patterson, and David Reich. Genetic evidence for two founding populations of the Americas. *Nature*, 525(7567):104–108, September 2015. ISSN 1476-4687. URL https://www.nature.com/articles/nature14895.

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

A Derivations

Depending on a readers' background in linear algebra, these results may appear elementary; I include them here for reference and because they were not obvious to me at the onset of this project.

F-statistics are invariant under a change-of-basis

$$
\begin{aligned}
F_2(X_i, X_j) &= \sum_{l=1}^{S} \left((x_{il} - \mu_l) - (x_{jl} - \mu_l)\right)^2 = F_2(Y_i, Y_j) \\
&= \sum_{l=1}^{S} \big(\sum_k L_{kl} P_{ik} - \sum_k L_{kl} P_{jk}\big)^2 \\
&= \sum_{l=1}^{S} \left(\sum_k L_{kl}(P_{ik} - P_{jk})\right)^2 \\
&= \sum_{l=1}^{S} \left(\sum_k L_{kl}^2 (P_{ik} - P_{jk})^2 + 2\sum_{k \neq k'} L_{kl} L_{k'l} (P_{ik} - P_{jk'})^2\right) \\
&= \sum_k \underbrace{\left(\sum_{l=1}^{L} L_{kl}^2\right)}_{1} (P_{ik} - P_{jk})^2 + 2 \sum_{k \neq k'} \underbrace{\left(\sum_{l=1}^{S} L_{kl} L_{k'l}\right)}_{0} (P_{ik} - P_{jk'})^2 \\
&= \sum_k (P_{ik} - P_{jk})^2 \qquad \text{(A1)}
\end{aligned}
$$

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.

The region of negative F_3-statistics is a n-ball Without loss of generality, assume that $X_1 = (r, 0, 0, \dots)$ and $X_2 = (-r, 0, 0, \dots)$, and let us assume that X_x has coordinates $(x_1, x_2, \dots, x_S)$ Assuming $F_3(X_x; X_1, X_2) = 0$, equation 13 becomes

$$
\begin{aligned}
2F_3(X_x; X_1, X_2) &= \|X_x - X_1\|^2 + \|X_x - X_2\|^2 - \|X_1 - X_2\|^2 = 0 \\
&= \left[(x_1 - r)^2 + \sum_{i=2}^{S} x_i^2\right] + \left[(x_1 + r)^2 + \sum_{i=2}^{S} x_i^2\right] - 4r^2 \\
&= 2\left[\sum_{i=1}^{S} x_i^2 + r^2 + x_1 r - x_1 r\right] - 4r^2 \\
F_3(X_x; X_1, X_2) &= -r^2 + \sum_{i=1}^{S} x_i^2 = -r^2 + \|X_x\|^2 = 0, \qquad \text{(A2)}
\end{aligned}
$$

which is the equation of a n-sphere with radius r and center at the origin, as assumed from the placing of X_1 and X_2. Now, assume that F_3 is negative, i.e. $F_3(X_x; X_1, X_2) = -k < 0$. Moving r^2 to the left we obtain

$$
r^2 - k = \|X_x\|^2, \qquad \text{(A3)}
$$

which is another n-sphere with a smaller radius, showing that all points inside the n-sphere will have negative F_3-values.

If a population lies outside the circle of this n-Sphere in any 2D-projection, F_3 is positive Assume the center of the n-sphere $C = \frac{X_1+X_2}{2} = (c_1, c_2, \ldots c_S)$, and $X_x = (x_1, x_2, \ldots x_S)$. Then,

$$\begin{aligned} F_3(X_x; X_1, X_2) &= \|X_x - C\|^2 - r^2 \\ &= \underbrace{(x_1 - c_1)^2 + (x_2 - c_2)^2}_{>r^2} + \underbrace{\sum_{i=3}^{S} (x_i - c_i)^2}_{\geq 0} - r^2 \\ &> 0. \end{aligned} \tag{A4}$$

The condition $(x_1 - c_1)^2 + (x_2 - c_2)^2 > r^2$ is satisfied whenever X_x is outside the circle obtained from projecting the n-sphere on the first two dimensions. An analogous argument applies for any low-dimensional representation.