1 Introduction

Fifty years ago, Lewontin used blood-group data from just 15 loci and showed that around 15% of global human genetic variation is explained by population structure. In the last decades, studies with increasingly large numbers of markers up to millions of SNPs and thousands of individuals have replicated this finding **???**. Consumer-genetic companies are using this variation to analyze and predict the genetic data of millions of customers and in some (heavily curated) data sets an indivudals place of origin can be predicted from genetic data at a resolution of a few hundred kilometers Novembre *et al.* (2008); **?**, and these genetic variation patterns can be used to learn about the historical movements of people. Lewontin also found that less than half (6%), of that variation can be attributed to continental-scale groups he called races, which he used to claim that "racial classification is (...) seen to be of virtually no genetic or taxonomic significance".

This view has been criticized, as 6% are plenty to learn about individual history. Much less agreement exists to what human population structure is discrete. On the one hand, geographic barriers to movement such as mountain ranges, deserts and oceans do lead to discontinuities in allele frequency space. On the other hand, one often finds population that are intermediate "around" these obstacles to gene flow, and models that assign partial ancestries or ancestries based on gradients are often more successful than discrete clustering models.

Yet, it frequently a useful analysis tool to think of populations as discrete units. For example, even though the underlying population structure may be continuous, sampling is not; and when quantifying ascertainment and sampling biases, or when discussing population structure it is often helpful to pretend populations are discrete, even though the underlying structure might be more complex and less rigid.

Thus, very often researchers will analyze a data set both using methods that assume population structure is discrete, and methods where we assume population structure is continuous, as this allows revealing different facets of the data, but unifying these approaches and getting a common understanding remains a persistent challenge.

One framework for the analysis of population structure that gained a lot of traction in the last decade are the F-statistics *sensu* Patterson Patterson *et al.* (2012); Peter (2016). This framework treats populations as discrete units in the analysis, and allows for a variety of tests for treeness, with an admixture as an alternative mode. Using this framwork, the vast majority of present-day human populations are admixed Pickrell and Reich (2014). Yet, this framework starts with the assumption that admixture is i) rare and ii) discrete.

However, F-statistics are not restricted to discrete populations. Indeed, as they can be written as functions of allele frequency variances, or expected pairwise coalescence times, statistics that can be calculated under a wide range of demographic models Peter (2016). Indeed, as they reflect inner products, they can be generalized to Euclidean space **?** (or any Hilbert space, although we won't pursue that here). Here, I explore these links between F-statistics and Euclidean spaces to establish connections between F-statistics and PCA. This allows direct interpretation of admixture in scenarios where population structure might not be discrete.

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 re-

searcher 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. A more detailed technical introduction is given in XXXXX, and a useful guide to interpretation is Cavalli-Sforza *et al.* (1994).

2.1 Introduction to PCA

Let us assume we have some genotype data summarized in a matrix $\mathbf{X}$, where the entry x_{ij} is the allele frequency of the i-th population at the j-th genotype. If we have k SNPs and n populations, $\mathbf{X}$ will have dimension $n \times k$.

As a population may be represented by just one (pseudo-)haploid or diploid individual, there is no conceptual difference between these cases and I will refer to populations as unit for analysis, for simplicity.

Since the allele frequencies are between zero and one, we can interpret each row x_i of $\mathbf{X}$ as a vector in $[0, 1]^k$, the *data space* of all possible allele frequencies on our markers.

The goal of a PCA is to find a low-dimensional representation of this data space that retains most of 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. In this approach, we first mean-center $\mathbf{X}$, obtaining a centered matrix $\mathbf{Y}$

$$y_{il} = x_{il} - \mu_l \tag{1}$$

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

PCA can then be written as

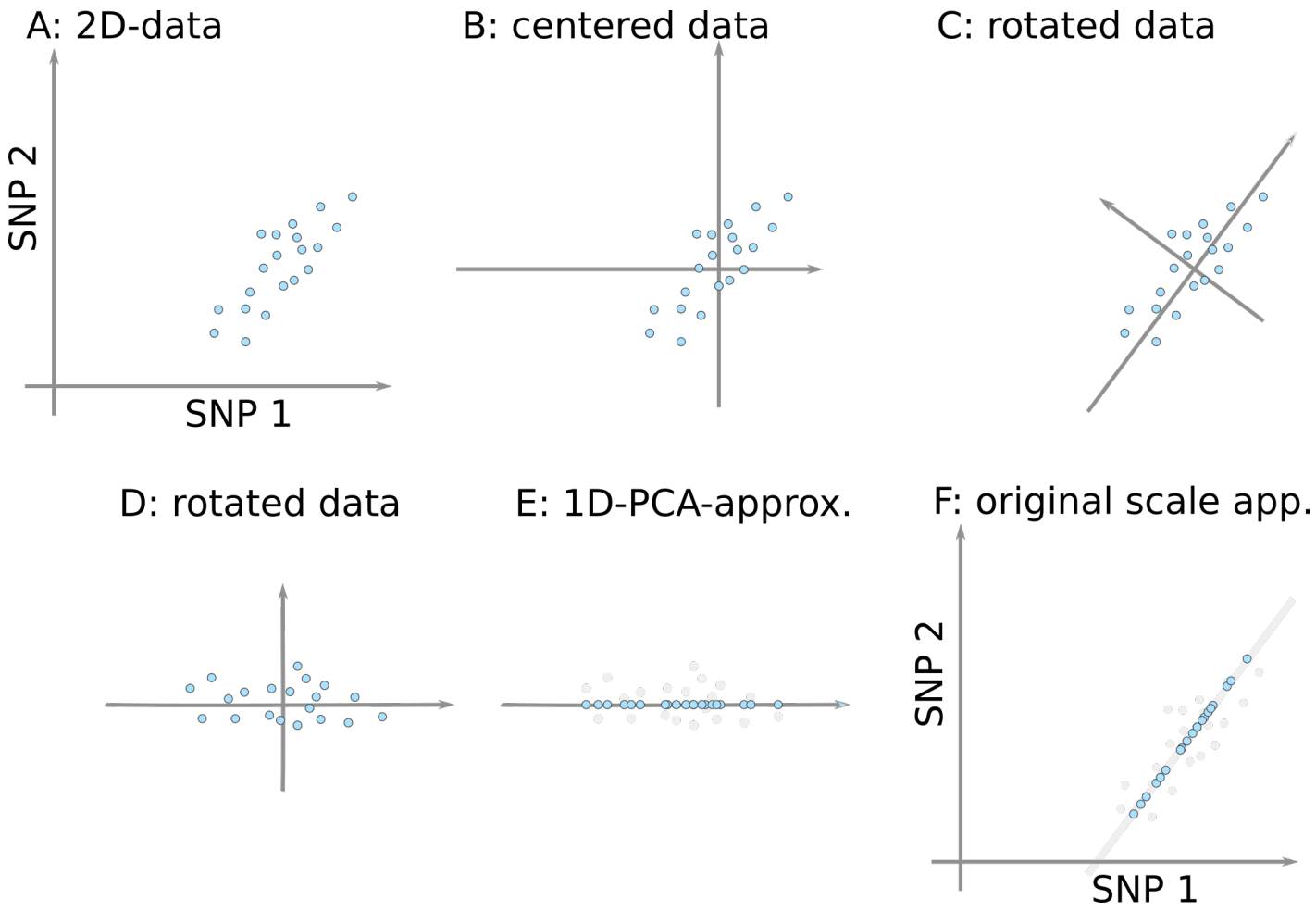

Figure 1: Basic Idea of PCA from 2D to 1D representation. A: Allele frequencies from different populations (blue dots) at two SNPs. A PCA is performed by centering the data (B), and rotating it (B) such that the first PC explains the majority of variation in the data, and the second PC is orthognal to 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).

$$\mathbf{Y} = \mathbf{CX} = \mathbf{U\Sigma V}^T = \mathbf{PL}, \tag{2}$$

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

In many implementations (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 $\mathbf{K}$:

$$\mathbf{K} = \mathbf{YY}^T = \mathbf{U\Lambda U}^T = \mathbf{U\Sigma}^2\mathbf{U}^T = \mathbf{PP}^T \tag{3}$$

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 $\lambda_i \neq 0$:

$$\mathbf{L}_i = \lambda_i^{-1}\mathbf{P}^T\mathbf{CX}. \tag{4}$$

3 F-statistics in PCA-space

As shown by e.g. Oteo-Garcia and Oteo (2021), F-statistics 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. By performing a PCA, we just translate and rotate our data, but Euclidean distances and dot products are both invariant under both these operations. Hence, neither mean-centering (a translation) nor PCA (a 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,

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

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 F_2-terms, analogous relations apply.

3.1 F-stats in 2-dimensional PC-space

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

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 n-balls and other high-dimensional geometric objects and thus build intuition. Second, in many applications it is argued that a 2-dimensional approximation is sufficient to explain the major components of population structure Novembre *et al.* (2008). 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 is an estimate of 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

$$\begin{aligned} 2F_3(X_x; X_1, X_2) &= 2\langle X_x - X_1, X_x - X_2\rangle \\ &= \|X_x - X_1\|^2 + \|X_x - X_2\|^2 - \|X_1 - X_2\|^2 \\ &< 0 \end{aligned} \tag{6}$$

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.

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.

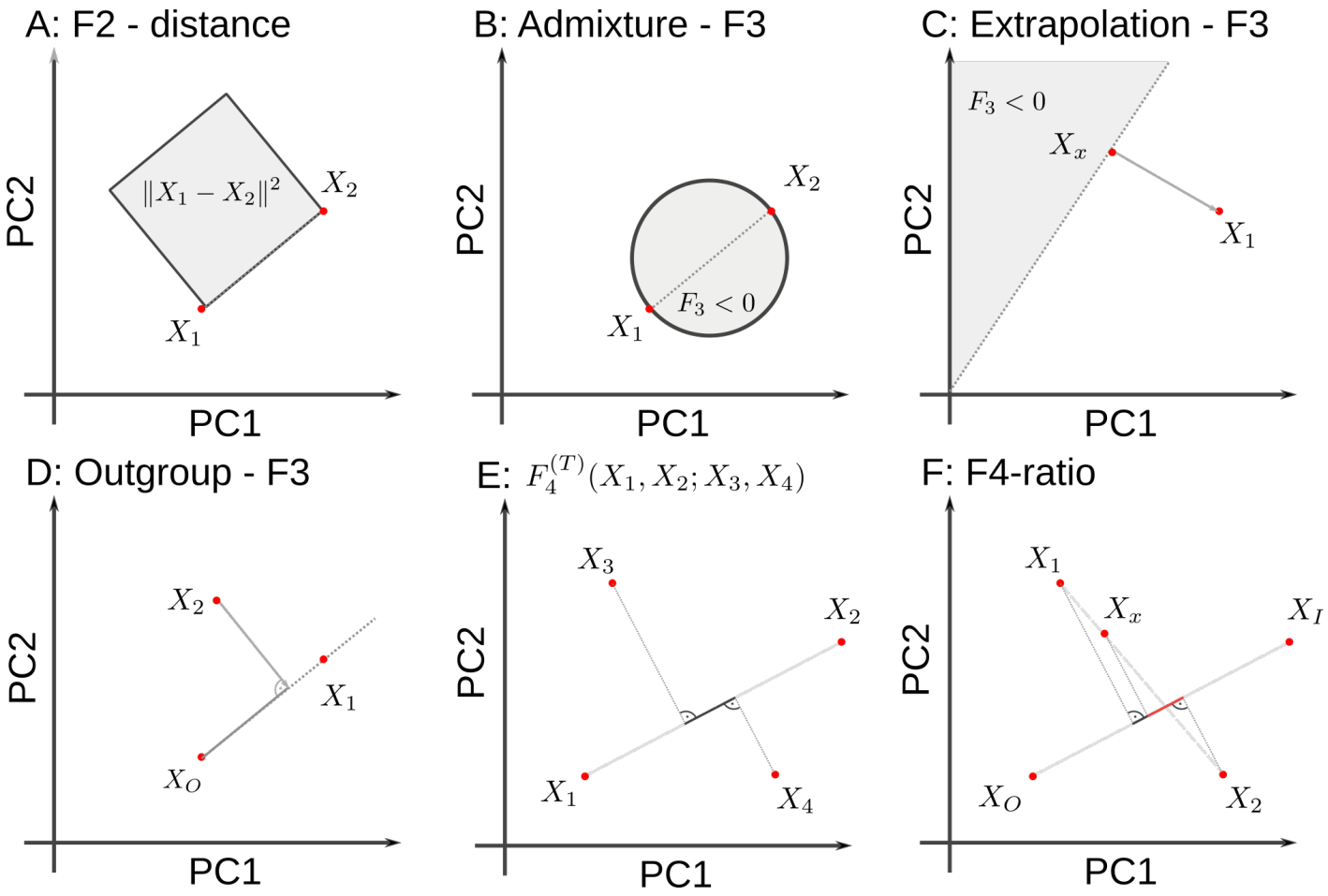

Figure 2: **Geometric representation of F-statistics on 2D-PCA-plot.** A: F_2 represents the squared Euclidean distance between two points in PC-space. B: Admixture-$F_3(X_x; X_1, X_2)$ is negative if X_x lies in the circle specified by the diameter $X_2 - X_1$
. C: $F_3(X_x; X_1, X_2)$ is negative given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

3.1.4 F_4-ratio

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

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

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

3.2 Higher-Dimensional Spaces

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 Technical considerations

5.1 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.

5.2 Missing data

5.3 F_2 error

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

$$f_2(X_1, X_2) = \frac{1}{L}\sum_l (x_{l1} - x_{l2})^2 - \frac{1}{L}\sum_l \frac{x_{l1}(1-x_{l1})}{n_{l1}-1} - \frac{1}{L}\sum_l \frac{x_{l2}(1-x_{l2})}{n_{l2}-1} - \tag{8a}$$

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

$$F_2(X_1, X_2) = \frac{1}{L}\sum_l (x_{l1} - x_{l2})^2 \tag{8b}$$

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; **?**).

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

5.4 qpADM

In Haak *et al.* (2015), qpADM, a procedure to estimate admixture proportions has been proposed. qpADM aims to solve equations of the form

$$
\begin{aligned}
\langle P_X - A, B - C\rangle &= \sum_i \alpha_i \langle R_i - A, B - C\rangle \\
&= \left\langle \sum_i \alpha_i R_i - A, B - C \right\rangle
\end{aligned} \tag{9}
$$

5.5 What is a dimension?

In both the PCA and F-statistic framework, a population at a particular point in time can be thought of as a single point in allele-frequency space, given by the k-dimensional vector v_0 of allele frequencies at the k SNPs in that population. If this population evolves for some time in isolation, allele frequencies will change due to genetic drift from v_0 to some other point v_1. Likewise, a second population with frequency w_0 will move to w_1. Crucially, if these populations do not interact, the changes in allele frequency, $v_1 - v_0$ and $w_1 - w_0$ will be uncorrelated Patterson *et al.* (2012). Thus, if we have two populations that descend from the same ancestral population in isolation, they can be thought of as evolving along orthognal dimensions from the same point. This argument is at the foundation of F-statistics.

6 outtakes

PCA from $\mathbf{X}$

$$
\mathbf{K} = \mathbf{YY}^T = \mathbf{CXX}^T\mathbf{C} = \mathbf{PP}^T \tag{10}
$$

7 Discussion

The fa

A Derivation

$$
\begin{aligned}
F_2(X_i, X_j) &= \sum_{l=1}^{L} \big((x_{il} - \mu_l) - (x_{jl} - \mu_l)\big)^2 = F_2(Y_i, Y_j) \\
&= \sum_{l=1}^{L} \big(\sum_k L_{kl} P_{ik} - \sum_k L_{kl} P_{kj}\big)^2 \\
&= \sum_{l=1}^{L} \left(\sum_k L_{kl}(P_{ik} - P_{jk})\right)^2 \\
&= \sum_{l=1}^{L} \left(\sum_k L_{kl}^2 (P_{ik} - P_{jk})^2 + 2 \sum_{k \neq k'} L_{kl} L_{k'l} (P_{ik} - P_{jk'})^2\right) \\
&= \sum_k \underbrace{\left(\sum_{l=1}^{L} L_{kl}^2\right)}_{1} (P_{ik} - P_{jk})^2 + \sum_{k \neq k'} \underbrace{\left(\sum_{l=1}^{L} L_{kl} L_{k'l}\right)}_{0} (P_{ik} - P_{jk'})^2 \\
&= \sum_k (P_{ik} - P_{jk})^2 \qquad (11)
\end{aligned}
$$

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

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

Pickrell, J. K. and Reich, D. 2014. Toward a new history and geography of human genes informed by ancient DNA. Trends in Genetics, 30(9):377–389

Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., Brandt, G., Nordenfelt, S., Harney, E., Stewardson, K., Fu, Q., Mittnik, A., Bánffy, E., Economou, C., Francken, M., Friederich, S., Pena, R. G., Hallgren, F., Khartanovich, V., Khokhlov, A., Kunst, M., Kuznetsov, P., Meller, H., Mochalov, O., Moiseyev, V., Nicklisch, N., Pichler, S. L., Risch, R., Rojo Guerra, M. A., Roth, C., Szécsényi-Nagy, A., Wahl, J., Meyer, M., Krause, J., Brown, D., Anthony, D., Cooper, A., Alt, K. W., and Reich, D. 2015. Massive migration from the steppe was a source for Indo-European languages in Europe. Nature, 522(7555):207–211

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