1 Introduction

About 15% of genetic variation in humans can be explained by population structure **???**, but the information contained in these 15% is sufficient to study the genetic diversity and history in great detail Cavalli-Sforza *et al.* (1994); **?**. For some data sets it is possible to predict an individuals origin at a resolution of a few hundred kilometers Novembre *et al.* (2008); **?**, and direct-to-consumer-genetics companies are using this variation to analyze the genetic data of millions of customers.

However, population genetic models that adequately capture the heterogeneity in human genetic diversity remain a challenge. It is also apparent that human genetic variation is largely distributed in gradients. In Lewontin's pioneering analysis, he found that less than half (6%), of that variation could be attributed to the continental-scale groups. Later studies on genome-wide data show

although some discontinuities, often associated with obstacles to migration such as mountain ranges or oceans, exist **?**Ramachandran *et al.* (2005); **?**); Peter (2020).

One related question is how discrete human populations are. While human genetic differentiation generally increases with geographic distance Ramachandran *et al.* (2005); **?**, this increase is not uniform. Obstacles to migration, such as oceans, mountains or deserts do frequently cause discontinuities in population structure Peter *et al.* (2020). Thus, while barriers to gene flow rarely are absolute, segregation policies by (perceived) ethnic or racial ancestry frequently cause local small-scale population differentiation that persist to the present day.

As a consequence, tools such as Principal Component Analysis (PCA) are frequently used to visualize or model human population structure (Cavalli-Sforza *et al.*, 1964, 1994). Informally, PCA aims to extract the axes of largest variation in a data set; for population genetic data sets of e.g. SNPs, these axes typically reflect population structure and need to be taken into account in e.g. association studies (Price *et al.*, 2006). As population structure is frequently sparse Engelhardt and Stephens (2010); McVean (2009), only the first few principal components (PCs) reflect population structure, i.e. the 15% of genetic variation that can be associated with structure.

While PCA is able to capture continuous and complex genetic structure (Novembre *et al.*, 2008; Novembre and Stephens, 2008), it is often hard to interpret. Even though human genetic diversity is to a large extent gradual, many models used to describe it have discrete components. This has a number of largely practical reasons; it is much easier to develop generative historical models using discrete populations **?**, sampling is usually discrete **??**, and results are often easier to communicate and interpret if populations are treated as discrete (**?**). Yet, as with any model, this ease of interpretation comes at a loss of some of the underlying complexity.

One discrete framework for the analysis of human 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, assuming, as a null model, that they are related in a tree-like fashion. Two simple statistics, called F_3 and F_4 can be used to test whether three or four populations are related in a treelike fashion, respectively. If these statistics are significant, treeness is rejected in favor of *admixture*. In recent years, more complicated methods have been developed that extend this to framework to more complex models involving many populations Patterson *et al.* (2012); Lipson *et al.* (2013); **?**, but the framework is built on the assumptions that populations are discrete, and gene flow is rare.

In this paper, my goal is to develop an interpretation of F-statistics in the context of PCA. This allows direct interpretation of F-statistics from *data space* (**?**), and allows us to talk about admixture in scenarios where population structure might not be discrete. For this purpose, I will thus briefly define PCA and F-statistics, and the geometry of F-statistics on PCA-plots. This allows me to develop approximations and visualizations which are illustrated on several examples.

2 Theory

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. Note that this merely assumes that sampling, but not necessarily the underlying structure, is discrete.

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 an optimal low-dimensional representation of $\mathbf{X}$ that retains most of the variation in the data (see Fig. 1 for an illustration).

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 $\mathbf{X}$ to obtain the 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

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

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 these weights makes little difference in practice and obfuscate the connection to F-statistics, I will for now assume that SNPs are unweighted, and defer discussion of weighting to a later section.

2.2 Introduction to F-statistics

From a modeling perspective, F-statistics have been primarily motivated by trees and admixture graphs (Patterson *et al.*, 2012), but can be calculated for arbitrary population genetics models for which expected pairwise coalescence times are known (Peter, 2016).

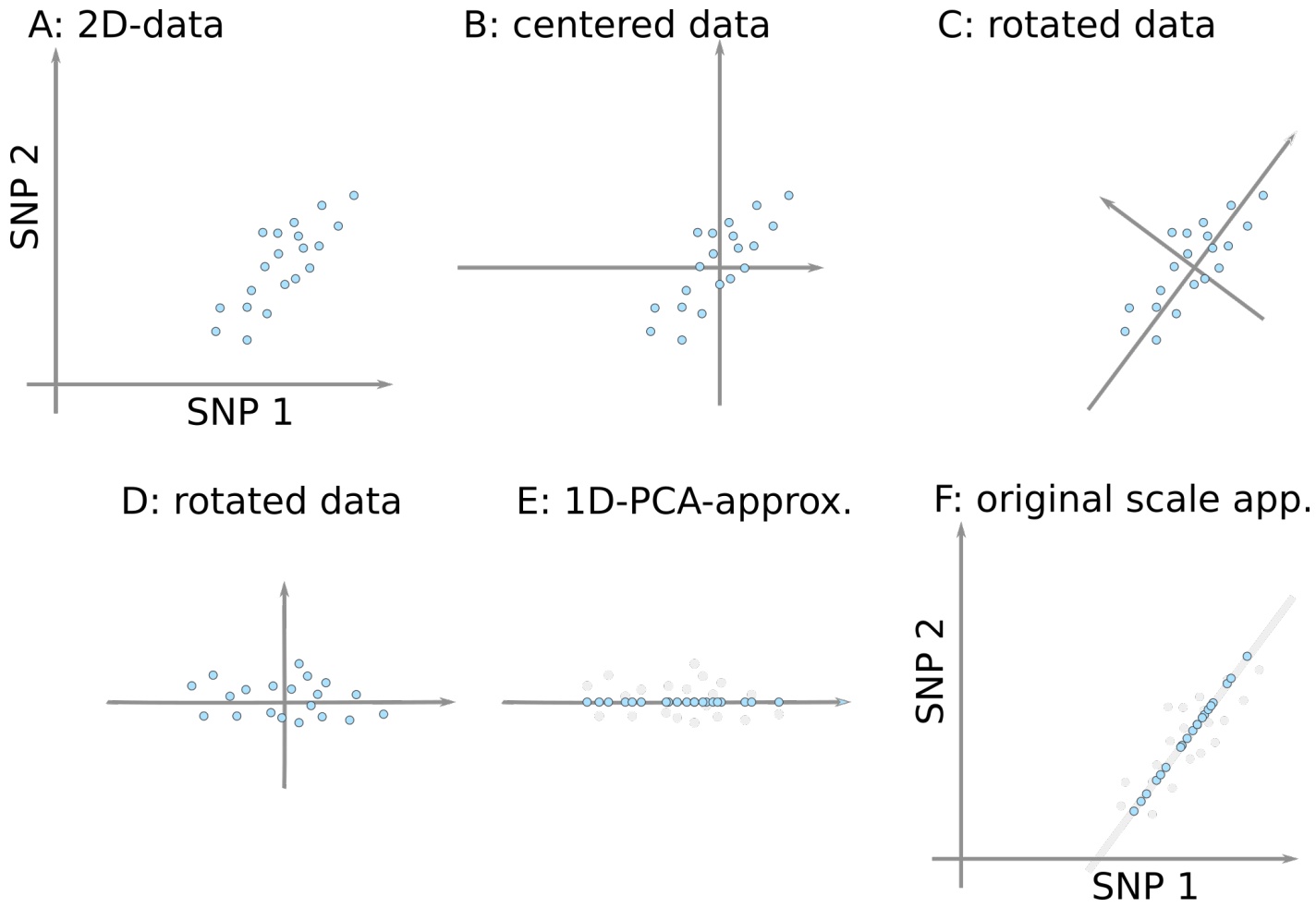

Figure 1: Basic Idea of PCA from 2D to 1D representation. A: Allele frequencies from different populations (blue dots) at two SNPs. A PCA is performed by centering the data (B), and rotating it (B) such that the first PC explains the majority of variation in the data, and the second PC is orthognal to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

They are based on the squared Euclidean distance F_2 used as a tree metric, and the

$$F_2(X_1, X_2) = \sum_i^k (x_{1i} - x_{2i})^2 = 2\mathbb{E}T_{12} - \mathbb{E}T_{11} - \mathbb{E}T_{22} \qquad = \|X_1 - X_2\|^2 \tag{3a}$$

$$F_3(X_1; X_2, X_3) = \sum_i^k (x_{1i} - x_{2i})(x_{1i} - x_{3i}) \qquad = \langle X_1 - X_2, X_1 - X_3 \rangle \tag{3b}$$

$$F_4(X_1X_2; X_3, X_4) = \sum_i^k (x_{1i} - x_{2i})(x_{3i} - x_{4i}) \qquad = \langle X_1 - X_2, X_3 - X_4 \rangle, \tag{3c}$$

where the X_i are rows of $\mathbf{X}$ that represent all loci of population i, and the T_{ij} denote the pairwise coalescence time of two sampled individuals from population i and j.

Here, much of the focus is not on interpreting these F-statistics in terms of model, but in *data space*. For this, we interpret F-statistics as inner products on $\mathbb{R}^L$, and the norm $\|\cdot\|$ and inner product $\langle \cdot, \cdot \rangle$

notation are used to make the connection to Euclidean geometry more explicit (see Oteo-Garcia and Oteo, 2021, for a thorough introduction to this interpretation).

2.3 PCA from F-statistics

Despite their different motivation, F-statistics and PCA are closely related through a procedure called multidimensional scaling (MDS). For MDS, we consider a matrix $\mathbf{F}_2, f_{ij} = F_2(X_i, X_j)$ of pairwise F_2-statistics. Then, the matrix of principal components $\mathbf{P}$ can be obtained by performing an eigendecomposition of the double-centered $\mathbf{F}_2$-matrix **?**:

$$\mathbf{P}\mathbf{P}^T = -\frac{1}{2}\mathbf{C}\mathbf{F}_2\mathbf{C}. \tag{4}$$

2.4 F-statistics in PCA-space

The F-statistics can be thought of as inner products in the Euclidean allele-frequency space. By performing a PCA, we just translate and rotate our data, but the dot products is 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,

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

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 (Reich *et al.*, 2009), analogous relations apply.

2.5 Geometric representation of F-statistics in PC-space

The transformation derived in the previous section allows us to consider the geometry of F-statistics in the rotated 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 as opposed to n-balls and other high-dimensional geometric objects. Second, in many applications it is argued that a 2-dimensional approximation is sufficient to explain the major components of population structure (e.g. 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.

2.5.1 F_2 in 2D 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.

2.5.2 F_3 and circles

A common application of F_3-statistics is to test for admixture; we say that population X_3 is admixed from sources X_1 and X_2 if $F_3(X_3; X_1, X_2) < 0$. This motivates the following question: given two source populations X_1, X_2, where would we expect the admixed population X_3 on a PCA plot? Since the allele frequencies of X_3 are intermediate between X_1 and X_2, we would expect it to lie between these two populations **?**McVean (2009), but we can make this notion more precise:

$$\begin{aligned} 2F_3(X_3; X_1, X_2) &= \langle X_3 - X_1, X_x - X_2 \rangle \\ &= \|X_3 - X_1\|^2 + \|X_3 - X_2\|^2 - \|X_1 - X_2\|^2 \\ &< 0 \end{aligned} \tag{6}$$

By the Pythagorean theorem, $F_3 = 0$ iff X_1, X_2 and X_x form a right-angled triangle. Hence, on a 2D-PCA-plot, the region where F_3 is zero is the circle with diameter $\overline{X_1X_2}$ (Figure 2B). If X_3 lies inside this circle, the angle is obtuse and F_3 is negative, otherwise it will be positive. If we consider all PCs, the circle is replaced by a n-ball with diameter $\overline{X_1X_2}$, but otherwise the reasoning is exactly the same. In this case, not all populations inside a circle on a 2D-plot will have negative F_3, as they may be outside the n-ball in another dimension. However, all populations that map outside the circle on any PC are guaranteed to have a positive associated F_3-statistic.

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.

2.5.3 F_4 and right angles

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

2.5.4 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 $X_I - X_O$.

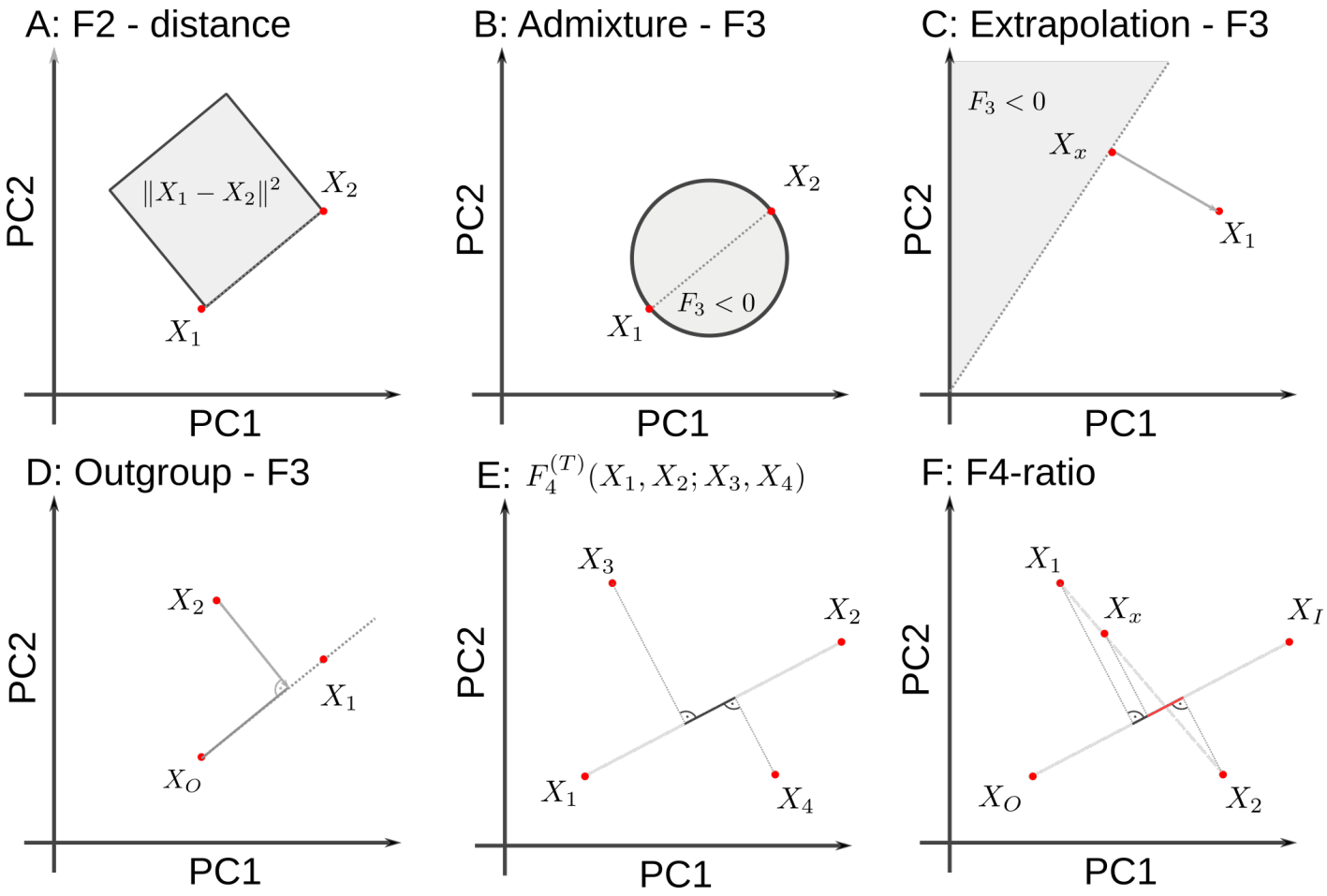

Figure 2: **Geometric representation of F-statistics on 2D-PCA-plot.** A: F_2 represents the squared Euclidean distance between two points in PC-space. B: Admixture-$F_3(X_x; X_1, X_2)$ is negative if X_x lies in the circle specified by the diameter $X_2 - X_1$
. C: $F_3(X_x; X_1, X_2)$ is negative given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

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. $F_4(X_1, X_1', X_2, X_2') = 0$

2.6 Higher-Dimensional Spaces

3 Results

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

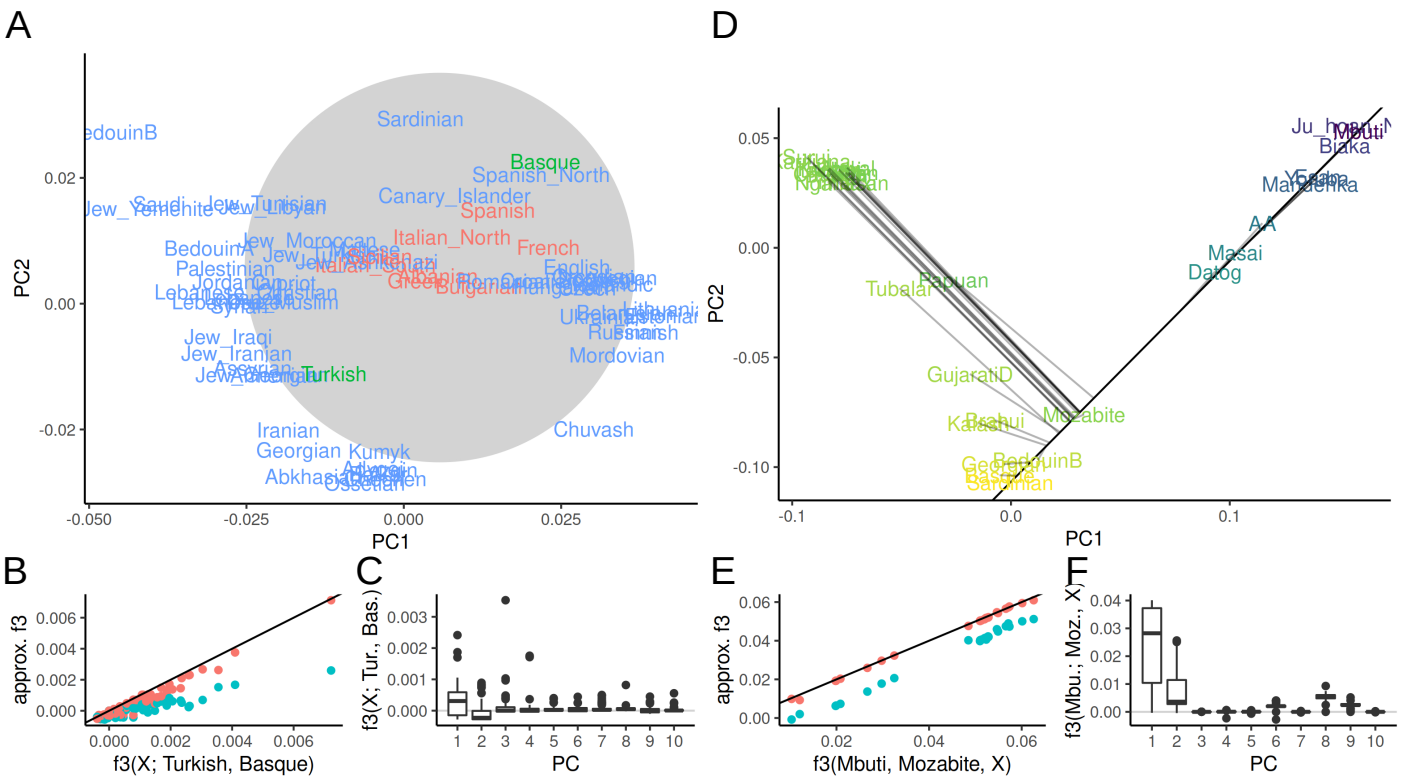

Figure 3: **PCA and F_3-statistics** A: PCA of Western Eurasian data; the circle denotes the region for which $F_3(X; \text{Basque}, \text{Turkish})$ may be negative. Populations for which F_3 is negative are colored in red. B, E: F_3 approximated with two (blue) and ten (red) PCs versus the full spectrum. C,F: Contributions of PCs 1-10 to each F_3-statistic. D: PCA of World data set, color indicates value of $F_3(\text{Mbuti}; \text{Mozabite}, X)$. The black line shows the projection axis Mbuti-Mozabite, the gray lines indicates the projected position of each population.

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

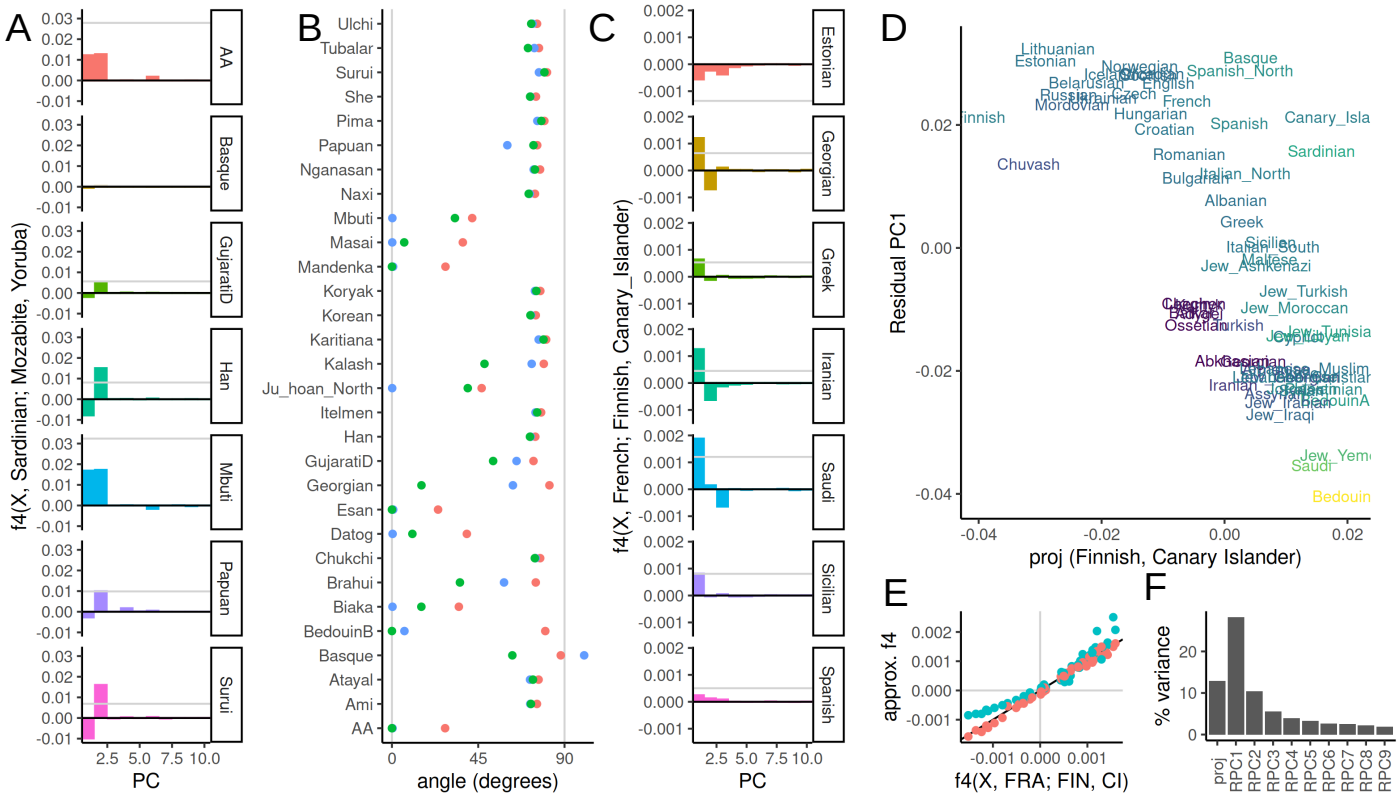

Figure 4: **PCA and F_4-statistics** A: Spectrum of select F_4-statistics in World data set. B: Projection angle representation of $F_4(X, Sardinian; Mozabite, Yoruba)$ (red) and approximations using two (blue) and ten (green) PCs. C: Spectrum of select $F-4$-statistics in Westeurasian data set. D: Scatterplot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

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., Barrai, I., and Edwards, A. W. F. 1964. Analysis of Human Evolution Under Random Genetic Drift. Cold Spring Harbor Symposia on Quantitative Biology, 29:9–20

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

Ramachandran, S., Deshpande, O., Roseman, C. C., Rosenberg, N. A., Feldman, M. W., and Cavalli-Sforza, L. L. 2005. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proceedings of the National Academy of Sciences of the United States of America, 102(44):15942–15947

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

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics, 38(8):904–909

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

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

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

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

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

Peter, B. M. 2020. 100,000 years of gene flow between Neandertals and Denisovans in the Altai mountains. bioRxiv. Publisher: Cold Spring Harbor Laboratory

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

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