PHILOSOPHICAL TRANSACTIONS B

royalsocietypublishing.org/journal/rstb

Research

Cite this article: Peter BM. 2022 A geometric relationship of F_2, F_3 and F_4-statistics with principal component analysis. *Phil. Trans. R. Soc. B* **377**: 20200413. https://doi.org/10.1098/rstb.2020.0413

Received: 7 July 2021
Accepted: 12 February 2022

One contribution of 15 to a theme issue 'Celebrating 50 years since Lewontin's apportionment of human diversity'.

Subject Areas:
bioinformatics, computational biology, evolution, genetics, theoretical biology

Keywords:
population structure, principal component analysis, *F*-statistics

Author for correspondence:
Benjamin M. Peter
e-mail: benjamin_peter@eva.mpg.de

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.5898677.

THE ROYAL SOCIETY PUBLISHING

A geometric relationship of F_2, F_3 and F_4-statistics with principal component analysis

Benjamin M. Peter

Max-Planck-Institute for Evolutionary Anthropology, Leipzig 04103, Germany

BMP, 0000-0003-2526-8081

Principal component analysis (PCA) and *F*-statistics *sensu* Patterson are two of the most widely used population genetic tools to study human genetic variation. Here, I derive explicit connections between the two approaches and show that these two methods are closely related. *F*-statistics have a simple geometrical interpretation in the context of PCA, and orthogonal projections are a key concept to establish this link. I show that for any pair of populations, any population that is admixed as determined by an F_3-statistic will lie inside a circle on a PCA plot. Furthermore, the F_4-statistic is closely related to an angle measurement, and will be zero if the differences between pairs of populations intersect at a right angle in PCA space. I illustrate my results on two examples, one of Western Eurasian, and one of global human diversity. In both examples, I find that the first few PCs are sufficient to approximate most *F*-statistics, and that PCA plots are effective at predicting *F*-statistics. Thus, while *F*-statistics are commonly understood in terms of discrete populations, the geometric perspective illustrates that they can be viewed in a framework of populations that vary in a more continuous manner.

This article is part of the theme issue 'Celebrating 50 years since Lewontin's apportionment of human diversity'.

1. Introduction

As in most species, the genetic diversity of human populations has been influenced by our history and environment over the last several hundred thousand years [e.g 1–5]. In turn, an important goal of population genetics is to use observed patterns of variation to investigate and reconstruct the demographic and evolutionary history of our species [6,7].

The complicated genetic structure observed in present-day human populations [8,9] is caused by the interplay of demographic and evolutionary processes with both discrete and continuous components [10–14]. In particular, populations are expected to differentiate if they are isolated from each other [15,16]. In humans, this may be caused because continental-scale geographic distances limit migration, causing a pattern known as isolation-by-distance [17,18]. However, these patterns are usually not uniform, but shaped by geography, particularly barriers to migration such as mountain ranges, oceans or deserts [1,19]. In addition, major historical population movements such as the out-of-Africa [20], Austronesian [21] or Bantu expansions [22] lead to more gradual patterns of genetic diversity over space [23]. Local migration between neighbouring populations will reduce differentiation, and long-distance migrations [24], and secondary contact between diverged populations such as Neanderthals and modern humans [25] may lead to locally increased diversity [26].

Particularly for large and heterogeneous datasets, disentangling all these processes is challenging, and we cannot expect to devise a single model catching both broad strokes and minute details of human history. A commonly used analysis

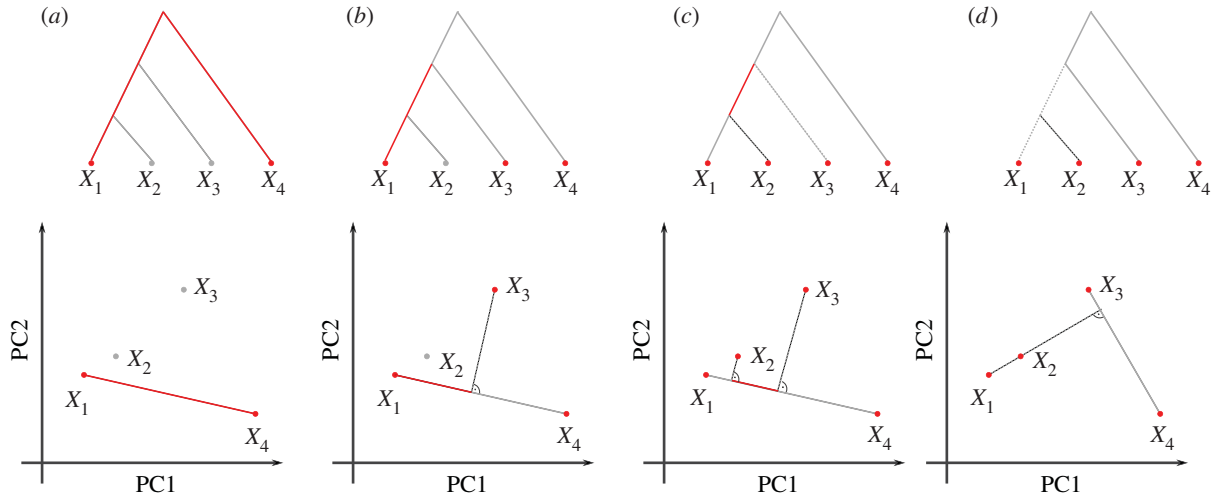

Figure 1. Representation of *F*-statistics on trees and two-dimensional PCA plots. The schematics show four populations and their representation using an (arbitrarily rooted) tree (top row) or a two-dimensional PCA plots (bottom row). (*a*) F_2 represents the squared Euclidean distance between two tree 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 orthogonal 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 to the orthogonal projection of $X_2 - X_3$ onto $X_1 - X4$. (*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. (Online version in colour.)

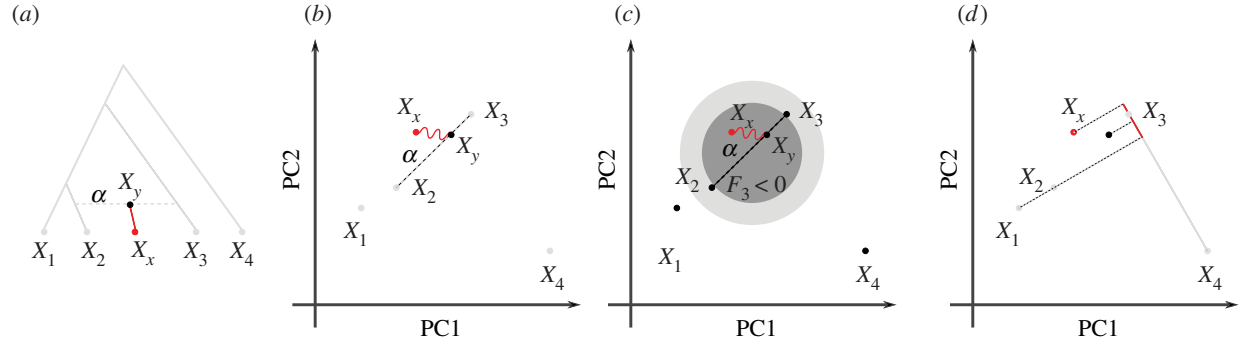

Figure 2. Admixture representation on two-dimensional-PCA plot. The schematics show four populations and their representation using an admixture graph (*a*) or a two-dimensional-PCA plot. (*a*) Admixture graph, with population X_y originating as an admixture of X_2 (with proportion $1 - \alpha$) and X_3 (proportion α). Subsequent drift (highlighted 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*) Negative region (light grey) for $F_3(X_x; X_2, X_3)$ and for $\hat{F}_3^{(2)}(X_x; X_2, X_3)$ based on two dimensions (dark grey). (*d*) $F_4(X_1, X_x; X_3, X_4)$ will no longer be zero (compare to figure 1*d*). (Online version in colour.)

paradigm is thus to combine tools based on different sets of assumptions, each emphasizing particular aspects of the data.

A typical analysis starts with data-driven, exploratory methods that summarize data making minimal assumptions [e.g. 6]. Examples are population trees [16,27,28], principal component analysis (PCA [1,29,30]) structure-like models [31,32] or multidimensional scaling (MDS [33]). These methods are limited in their ability to estimate biologically meaningful parameters, but provide useful summaries and visualizations. Typically, these analyses are then complemented with methods based on explicit demographic models, which are used to estimate parameters or test hypotheses [34–36].

When the number of populations exceeds a few dozen, even codifying reasonable population models can be prohibitively difficult. One approach is to pick a small set of 'representative' samples, and restrict modelling to this subset [e.g. 37,38]. However, this has the drawback that a large proportion of the available data remains unused. An increasingly popular alternative approach, particularly in the analysis of human ancient DNA, is therefore to focus on the relationship between two, three or four populations, commonly using *F*-statistics *sensu* Patterson [39–41]. Formal definition will be given in the Theory section; but an informal motivation starts with the null model that populations are related as a tree, in which each *F*-statistic measures the length of a particular set of branches (figure 1; [41,42]).

In most applications, *F*-statistics are estimated from data, and then used as tests of treeness. In particular, under the assumption of a tree, F_3 is restricted to be non-negative, and many F_4-statistics will be zero [40,42], and data that violates these constraints is incompatible with a tree-like relationship between populations. The canonical alternative model is an admixture graph (or phylogenetic network) [40,43], which is a tree which allows for additional edges reflecting gene flow (figure 2*a*). However, admixture graphs are not the only plausible alternative model, and expected *F*-statistics can be calculated for a wide range of population genetic demographic models [41].

(a) *F*-statistics and principal component analysis

The practical issue addressed in this study is how *F*-statistics can be compared with PCA, one of the most widely used data-driven modelling techniques. One way PCA can be motivated is as generating a low-dimensional representation of the data, with each dimension (called a principal

component, PC) retaining a maximum of the variance present in the data. To understand population structure, the use of PCA has been pioneered by Cavalli-Sforza *et al.* [44], who used allele-frequency data at a population level to visualize genetic diversity [1]. Currently, PCA is most commonly performed on individual-level genotype data [e.g. 30,45], making use of the hundreds of thousands of loci available in most genome-scale datasets. The PCA decomposition has been studied for a number of explicit population genetic models including trees [16], spatially continuous structure [46], the coalescent [47] and discrete population models [48]. Here, in order to link PCA to *F*-statistics, I interpret both of them geometrically in *allele frequency space*, i.e. as functions of a high-dimensional Euclidean space. For *F*-statistics, this interpretation was recently developed by Oteo-Garcia & Oteo [49], and for PCA it follows naturally from the interpretation of approximating a high-dimensional space with a low-dimensional one.

In the next section, I will formally derive the connection between *F*-statistics and PCA, and show how *F*-statistics can be interpreted geometrically, with a particular emphasis on two-dimensional-PCA plots. In the Results section, I will then discuss how some of the most common applications of *F*-statistics manifest themselves on a PCA, and illustrate them on two example datasets, before ending with a discussion.

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 [50], a useful technical primer is Pachter [51], and a helpful guide to interpretation is Cavalli-Sforza *et al.* [1]. Readers unfamiliar with *F*-statistics may find [40,41] or [49] helpful.

(a) Formal definition of *F*-statistics

Let us assume we have a set of populations for which we have single-nucleotide polymorphism (SNP) allele frequency data from S biallelic loci, and for simplicity, I will assume that there is no missing data. Let x_{il} denote the frequency of an arbitrary allele at the lth SNP in the ith population; and let $X_i = (x_{i1}, x_{i2}, \ldots x_{iS})$ be a vector collecting all allele frequencies for population i. As X_i will be the only data summary considered here for population i, I make no distinction between the population and the allele frequency vector used to represent it.

The three *F*-statistics are defined as

$$F_2(X_1, X_2) = \frac{1}{S}\sum_{l=1}^{S}(x_{1l} - x_{2l})^2 \tag{2.1a}$$

$$F_3(X_1; X_2, X_3) = \frac{1}{S}\sum_{l=1}^{S}(x_{1l} - x_{2l})(x_{1l} - x_{3l}) \tag{2.1b}$$

$$\text{and} \quad F_4(X_1, X_2; X_3, X_4) = \frac{1}{S}\sum_{l=1}^{S}(x_{1l} - x_{2l})(x_{3l} - x_{4l}). \tag{2.1c}$$

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

$$2F_3(X_1; X_2, X_3) = F_2(X_1, X_2) + F_2(X_1, X_3) - F_2(X_2, X_3) \tag{2.2a}$$

and

$$2F_4(X_1, X_2; X_3, X_4) = F_2(X_1, X_3) + F_2(X_2, X_4) - F_2(X_1, X_4) - F_2(X_2, X_3). \tag{2.2b}$$

Commonly, a distinction is made between statistics estimated from data (denoted with lowercase-*f*), and theoretical quantities (defined in equation (2.1)). I do not make this distinction, but will explicitly mention when I analyse statistics calculated from data.

F-statistics have been primarily motivated in the context of trees and admixture graphs [40]. In a tree, the squared Euclidean distance $F_2(X_1, X_2)$ measures the length of the path between populations X_1 and X_2 (figure 1*a*); F_3 represents the length of an external branch (figure 1*b*) and F_4 the length of an internal branch, respectively (figure 1*c*). The length of each branch can be thought of in units of genetic drift, and is non-negative [40]. Crucially, this means that F_4 will be zero for pairs of populations from non-overlapping clades, which means that the tree lacks the branch corresponding to this statistic (as in figure 1*d*).

Thinking of *F*-statistics as branch lengths is useful for a number of applications, including building multi-population models [40,52], estimating admixture proportions [38,53] and finding the population most closely related to an unknown sample ('Outgroup'-F_3-statistic).

Most commonly however, F_3 and F_4 are used as tests of treeness [40]: negative F_3 values correspond to a branch with negative genetic drift, which is not allowed under the null assumption of a tree-like population relationship. Similarly if four populations are related as a tree, then at least one of the F_4-statistics between the populations will be zero [40,54].

The most widely considered alternative model is an admixture graph [40]; an example is given in figure 2*a*. Here, 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 admixed population we observe. In this case, all F_4-statistics involving X_y and both admixture sources will be non-zero, and, in some cases, $F_3(X_y; X_2, X_3)$ will be negative (exact conditions can be found in [41]).

(b) Geometric interpretation of *F*-statistics

An implicit assumption in the development of *F*-statistics in the context of admixture graphs has been that population lineages are mostly discrete, and that gene flow is rare. Recently, Oteo-García & Oteo [49] showed that this is not, in fact, necessary. Specifically, they interpret the populations X_i as points or vectors in the S-dimensional *allele frequency space* $\mathbb{R}^S$. In this case, the *F*-statistics can be thought of as inner (or dot) products, and they showed that all properties and tests related to treeness can be derived in this larger space. In this framework, the *F*-statistics can be written as

$$F_2(X_1, X_2) = \frac{1}{S}\sum_{l=1}^{S}(x_{1l} - x_{2l})^2 = \frac{1}{S}\langle X_1 - X_2, X_1 - X_2\rangle = \frac{1}{S}\|X_1 - X_2\|^2 \tag{2.3a}$$

$$F_3(X_1; X_2, X_3) = \frac{1}{S}\sum_{l=1}^{S}(x_{1l} - x_{2l})(x_{1l} - x_{3l}) = \frac{1}{S}\langle X_1 - X_2, X_1 - X_3\rangle \tag{2.3b}$$

and

$$F_4(X_1, X_2; X_3, X_4) = \frac{1}{S}\sum_{l=1}^{S}(x_{1l} - x_{2l})(x_{3l} - x_{4l}) = \frac{1}{S}\langle X_1 - X_2, X_3 - X_4\rangle, \tag{2.3c}$$

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

$$\langle a, b\rangle = \sum_i a_i b_i \tag{2.4a}$$

$$\langle a, b\rangle = \|a\|\|b\|\cos(\phi) \tag{2.4b}$$

$$\langle a, a\rangle = \|a\|^2 \tag{2.4c}$$

$$\text{and} \quad \langle a + c, b\rangle = \langle a, b\rangle + \langle b, c\rangle, \tag{2.4d}$$

where ϕ is the angle between a and b. The inner product is closely related to the vector projection

$$proj_b a = \frac{\langle a, b\rangle}{\|b\|^2} b, \tag{2.5}$$

which is a vector colinear to b whose length measures how much vector a points in the direction of b. Thinking of F-statistics as projections also holds on trees: in e.g. a $F_4(X_1, X_4; X_2, X_3)$-statistic (figure 1*c*), the internal branch is precisely the intersection of the paths from X_1 to X_4 and from X_2 and X_3. On trees, all disjoint paths are independent (i.e. orthogonal) from each other, and thus the external branches vanish under the projection.

One issue with the geometric approach of Oteo-García & Oteo [49] is that each SNP (commonly in the millions) adds a dimension, but high-dimensional spaces are hard to visualize, interpret and analyse. Fortunately, it has been commonly observed that population structure is often quite low-dimensional, and only a few PCs frequently provide a good approximation of the covariance structure in human genetic variation data [30]. Therefore, we may hope that PCA could yield a reasonable approximation of the allele frequency space, and that F-statistics as measures of population structure may likewise be well approximated by the first few PCs.

(c) Formal definition of principal component analysis

PCA is a common way of summarizing genetic data, and so a large number of variations of PCA exist, e.g. in how SNPs are standardized, how missing data are treated or whether we use individuals or populations as units of analysis [1,30]. The version of PCA I use here is set up such that the similarities to F-statistics are maximized, and does *not* reflect how PCA is most commonly applied to genome-scale human genetic variation datasets. 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 [30]. The differences this causes will be addressed in the discussion.

Let us again assume we have allele frequency data as above, but let us now assume we aggregate the allele frequency vectors X_i of many populations in a matrix $\mathbf{X}$ whose entry x_{il} reflects the allele frequency of the ith population at the lth 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 $\mathbb{R}^S$.

PCA allows us to approximate the points in the high-dimensional allele frequency space by a K-dimensional subspace of the data. If all PCs are considered, $K = n - 1$, in which case the data are simply rotated. However, the historical processes that generated genetic variation often result in *low-rank* data [55], 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 (e.g. [50]). In this approach, we first mean-centre $\mathbf{X}$, obtaining a centred matrix $\mathbf{Y}$

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

where μ_l is the mean allele frequency at the lth locus.

PCA can then be written as

$$\mathbf{Y} = \mathbf{CX} = (\mathbf{U\Sigma})\mathbf{V}^T = \mathbf{PL}, \tag{2.6}$$

where $\mathbf{C} = \mathbf{I} - (1/n)\mathbf{1}$ is a centring 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 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. [56]). Alternatively, the PCs can also be obtained from an eigendecomposition of the covariance matrix $\mathbf{YY}^T$. This can be motivated from (2.6)

$$\mathbf{YY}^T = \mathbf{PLL}^T\mathbf{P}^T = \mathbf{PP}^T, \tag{2.7}$$

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

(d) Connection between principal component analysis and *F*-statistics

(i) Principal components from *F*-statistics

PCA, as defined above, and F-statistics are closely related. It is a classical result that PCA is equivalent to multidimensional scaling using squared Euclidean distances [57]. Since F_2-distances are squared Euclidean, we calculate the pairwise $F_2(X_i, X_j)$ between all n populations, and collect them in a matrix $\mathbf{F}_2$. Multidimensional scaling then proceeds by double-centring it, so that its row and column means are zero, and perform an eigen decomposition of the resulting matrix

$$\mathbf{PP}^T = -\frac{1}{2}\mathbf{CF}_2\mathbf{C}. \tag{2.8}$$

Although we arrive at $\mathbf{P}$ from a very different angle, as long as we make the same choices about normalization and units of analysis, we will get the exact same results.

(ii) *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 [49]. What this means is that we are free to calculate F_2 either on the uncentred data $\mathbf{X}$, the centred data $\mathbf{Y}$ or any other orthogonal basis such as the principal components $\mathbf{P}$.

Formally,

$$F_2(X_i, X_j) = \sum_{l=1}^{L} (x_{il} - x_{jl})^2 = \sum_{l=1}^{L} ((x_{il} - \mu_l) - (x_{jl} - \mu_l))^2 = F_2(Y_i, Y_j) = \sum_{k=1}^{n} (p_{ik} - p_{jk})^2 = F_2(P_i, P_j). \quad (2.9)$$

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 (equations (2.2*a*) and (2.2*b*)), 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,

$$F_2(P_i, P_j) = \underbrace{\sum_{k=1}^{K} (p_{ik} - p_{jk})^2}_{\hat{F}_2^{(K)}(P_i,P_j)} + \underbrace{\sum_{k=K+1}^{n} (p_{ik} - p_{jk})^2}_{\epsilon^{(K)}(P_i,P_j)}. \quad (2.10)$$

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 of $\hat{F}_2$ when the exact number of PCs is not relevant. If we sum up the squared approximation errors over all pairs of populations in our sample, we obtain

$$\sum_{i,j} \epsilon^{(K)}(P_i, P_j)^2 = \sum_{i,j} (\hat{F}_2^{(K)}(P_i, P_j) - F_2^{(K)}(P_i, P_j))^2 = \left\| \mathbf{F}_2 - \hat{\mathbf{F}}_2 \right\|_F^2, \quad (2.11)$$

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 [50]. 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.

(iii) *F*-statistics and samples projected onto PCA

One of the easiest ways of dealing with missing data in PCA is to calculate the principal components (equation (2.6)) 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

$$\mathbf{YL}^T = \mathbf{PLL}^T = \mathbf{P},$$

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

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

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

$$\left\| Y_{\text{new}} - P_{\text{proj}}\mathbf{L} \right\|^2 = F_2(P_{\text{proj}}\mathbf{L}, Y_{\text{new}}).$$

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, and add a scaling factor [30].

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

$$\begin{aligned} F_2(X_i, X_{\text{new}}) &= F_2(Y_i, Y_{\text{new}}) \\ &= F_2(P_i, P_{\text{proj}}) + F_2(P_{\text{proj}}\mathbf{L}, Y_{\text{new}}) \\ &= \hat{F}_2(P_i, P_j) + \epsilon(P_i, P_j) + F_2(P_{\text{proj}}\mathbf{L}, Y_{\text{new}}). \end{aligned} \quad (2.12)$$

The second row follows because the projection error and projection are orthogonal to each other. The main implication of equation (2.12) is that both truncation and projection introduce some error, and that $\hat{F}_2(P_i, P_j)$ will be a good approximation to $F_2(P_i, P_j)$ only if both errors are small.

3. Material and 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 datasets that I briefly introduce here. Both are based on the analyses by Lazaridis *et al.* [58]. The data are from the Reich laboratory compendium dataset (v.44.3), downloaded from https://reich.hms.harvard.edu/allen-ancient-dna-resource-aadr-downloadable-genotypes-present-day-and-ancient-dna-data, using data on the 'Human Origins'-SNP set (597,573 SNPs). SNPs with missing data in any population are excluded. The code used to write this paper, create all figures and analyses is available on doi:10.5281/zenodo.6424178.

(a) 'World' dataset

This dataset is a subset of the 'World Foci' dataset of Lazaridis *et al.* [58], where I removed samples that are not permitted for free reuse. These populations span the globe and roughly represents global human genetic variation (638 individuals from 33 populations) As adjacent sampling locations are often thousands of kilometres apart, 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 electronic supplementary material, File S1.

(b) Western Eurasian dataset

This dataset of 1119 individuals from 62 populations contains present-day individuals from the Eastern Mediterranean, Caucasus and Europe. Lazaridis *et al.* [58] used this dataset 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. [59,60]). Genetic differentiation in this region is low and closely mirrors geography [45]. I thus speculate that gene flow between these populations is common [61], 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 electronic supplementary material, File S2.

(c) Computing *F*-statistics and PCA

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 (2.8) and the `eigen` function to obtain the PCs. The right-hand-side matrix of equation (2.8) is expected to have non-negative eigenvalues (i.e. $-\mathbf{CF_2C}$ is positive-semidefinite). However, when F_2-statistics are estimated from data, sampling noise might make some of them slightly negative, which would lead to imaginary PCs. I avoid this by setting all negative eigenvalues to zero.

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, in which case the geometric interpretations become very simple. Thus, throughout the schematic figures, I assume that two PCs are sufficient to characterize population structure. In the data applications, I evaluate how deviations of this assumption may manifest themselves in PCA plots.

(a) F_2 in PC space

The F_2-statistic is an estimate of the squared allele-frequency distance between two populations. On a tree (figure 1*a*), this corresponds to the branch between two populations. In allele-frequency space, it corresponds to the squared Euclidean distance, and thus reflects the intuition that closely related populations will fall close to each other on a PCA plot, and have low pairwise F_2-statistics. However, since F_2 can be written as a sum of squared (non-negative) terms for each PC (equation 2.9), the distance on a PCA plot will always be an underestimate of the full F_2 distance. Thus, PCA might project two populations with high F_2 distance very close to each other, which would indicate that these particular PCs are not suitable to understand and visualize the relationship between these particular populations, and likely more PCs need to be investigated to understand how these populations are related to each other. In converse, populations that are distant on the first few PCs are guaranteed to also have a large F_2-distance, since the distance contributed from the omitted PCs cannot be negative.

(b) When are admixture-F_3-statistics negative?

Consider again the admixture scenario in figure 2*a*, 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 X_x. How is such a scenario displayed on a PCA? Since 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 2*b*), at a location predicted by the admixture proportions. Subsequent drift will change the allele frequencies of X_y (to say, X_x), and so in general it might fall on a different point on a PCA plot. An exception occurs when X_x (and no other populations related to X_x) are not part of the construction of the PCA, so that $X_x - X_y$ is orthogonal to all PCs, i.e.

$$\langle X_x - X_y, X_i - X_j\rangle = \langle X_x - X_y, P_i\rangle = 0$$

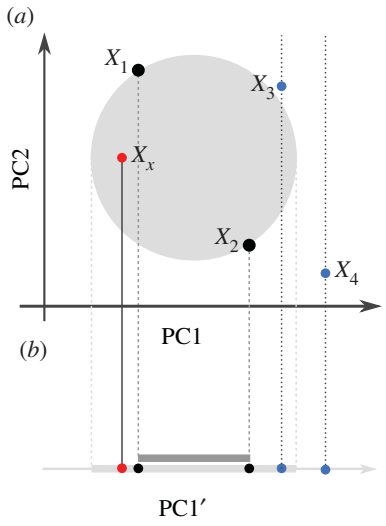

Figure 3. Admixture-F_3-statistic and dimensionality reduction. (*a*) The grey circle is defined by X_1 and X_2 and reflects the area for which F_3 is negative, i.e. $F_3(X_x; X_1, X_2) < 0 < F_3(X_3; X_1, X_2) \leq F_3(X_4; X_1, X_2)$. (*b*) Projection on one dimension. The full rejection region (light grey bar) is given by the boundaries of the circle in *a*, and is larger than that predicted from the distance between X_1 and X_2 (dark grey bar) on PC1 alone. Points like X_x with negative F_3 are guaranteed to lie in the light grey (but not dark grey) region. X_3 also projects onto the grey bar, even though it is outside the circle and F_3 is positive, and it would also lie inside the positive area if projected onto PC2. However, points such as X_4 that project outside the rejection region have $F_3 > 0$, since the associated projection line does not intercept the circle. (Online version in colour.)

for all populations $i, j \leq n$. In this case, X_x and X_y project to the same point, and the location on the PCA can directly be used to predict the admixture proportions [47,49,62]. However, if either X_x is included in the construction of the PCA, or if some gene flow occurred between X_x and any of the populations used to construct the PCA, X_x and X_y may project on different spots (figure 2*b*).

Thus, a natural question to ask is given two source populations X_2, X_3, can we use PCA to predict which populations might have negative F_3-statistics? This condition can be written as

$$\begin{aligned} 2F_3(X_x; X_2, X_3) &= 2\langle X_x - X_2, X_x - X_3\rangle \\ &= \|X_x - X_2\|^2 + \|X_x - X_3\|^2 - \|X_2 - X_3\|^2 < 0. \end{aligned} \tag{4.1}$$

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 an 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 could be considered admixed if it lies inside the n-ball with diameter $\overline{X_2X_3}$ (figure 1*b*, equation (A2)).

(c) F_3 and dimensionality reduction

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 the example of projecting $F_3(X_x; X_1, X_2)$ from a two-dimensional space onto a single dimension (figure 3). In that example, the distance between X_1 and X_2 has a substantial contribution from PC2, and so

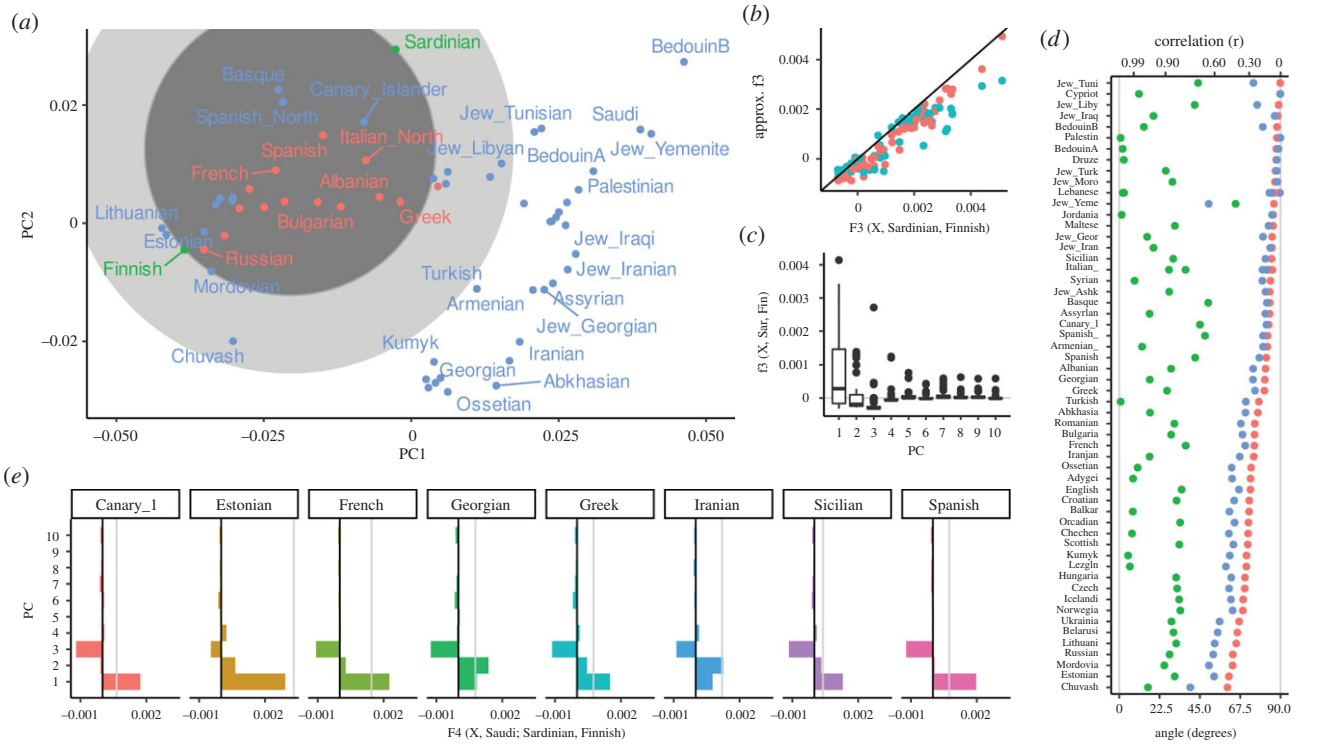

Figure 4. PCA and *F*-statistics for the Western Eurasian dataset. (*a*) PCA biplot; the light grey circle denotes the region for which $F_3(X;$ Sardinian, Finnish) may be negative, the dark circle is based on just the first two PCs. Populations for which F_3 is negative are coloured in red. (*b*) F_3 approximated with two (blue) and 10 (red) PCs versus the full spectrum. (*c*) Boxplot of contributions of PCs 1–10 to each F_3-statistic. (*d*) Projection angle and correlation interpretation of $F_4(X,$ Saudi; Sardinian, Finnish) based on two PCs (green), three PCs (blue) or full data (red). (*e*) Contribution of the first 10 PCs to select F_4-statistics, with the first three PCs containing the majority of contributions. (Online version in colour.)

the negative region (light grey) is larger than what would be predicted from just one dimension (dark grey bar), but if $\hat{F}_3 \approx F_3$, the two areas would be very close. Thus, if considering a reduced-dimension PCA plot, some points (such as X_3) may project inside the negative region, but have positive F_3 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* projection, F_3 is guaranteed to be bigger than 0 (see equation (A4) in the appendix). An intuitive example is given by X_4 in figure 3: all points projecting to the same point on figure 3*b* as X_4 lie outside the circle.

(i) Example

As an example, I visualize the admixture statistic $F_3(X;$ Sardinian, Finnish), on the first two PCs of the Western Eurasian dataset (figure 4*a*). In this case, the projected n-ball (light grey) and circle based on two dimensions (dark grey) have similar sizes. However, several populations that appear inside the circles (e.g. Basque, Canary Islanders) have, in fact, positive F_3 values, so they lie outside the n-ball. This reveals that the first two PCs do not capture all the genetic variation relevant for European population structure. Consequently, approximating F_3 by the first two or even 10 PCs (figure 4*b*) only gives a coarse approximation of F_3, and from figure 4*c* we see that many higher PCs contribute to F_3 statistics.

However, many populations, particularly from Western Asia and the Caucasus, on the right-hand side of the plot, 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 Sardinians and Fins.

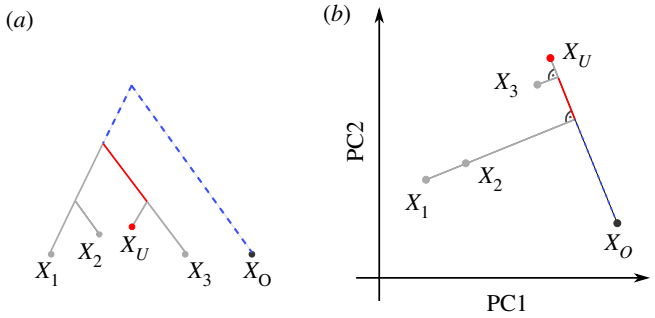

Figure 5. Outgroup-F_3-statistics. Interpretation of outgroup-F_3-statistic on a tree (*a*) and PCA plot (*b*). The highlighted segment represents $F_3(X_O; X_U, X_3)$ and the dashed segment reflects $F_3(X_O; X_U, X_1)$ and $F_3(X_O; X_U, X_2)$, which have the same value. (Online version in colour.)

(d) Outgroup-F_3-statistics as projections

A common application of F_3-statistics is, given an unknown sample X_U, to find the most closely related population among a reference panel (X_i) [63]. This is done using an *outgroup*-F_3-statistic $F_3(X_O; X_U, X_i)$, where X_O is an outgroup. The reason an outgroup is introduced is to account for differences in sample times and additional drift in the reference populations (figure 5*a*). The outgroup-F_3-statistic $F_3(X_O; X_U, X_3)$ represents the branch length 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 F_3-statistic.

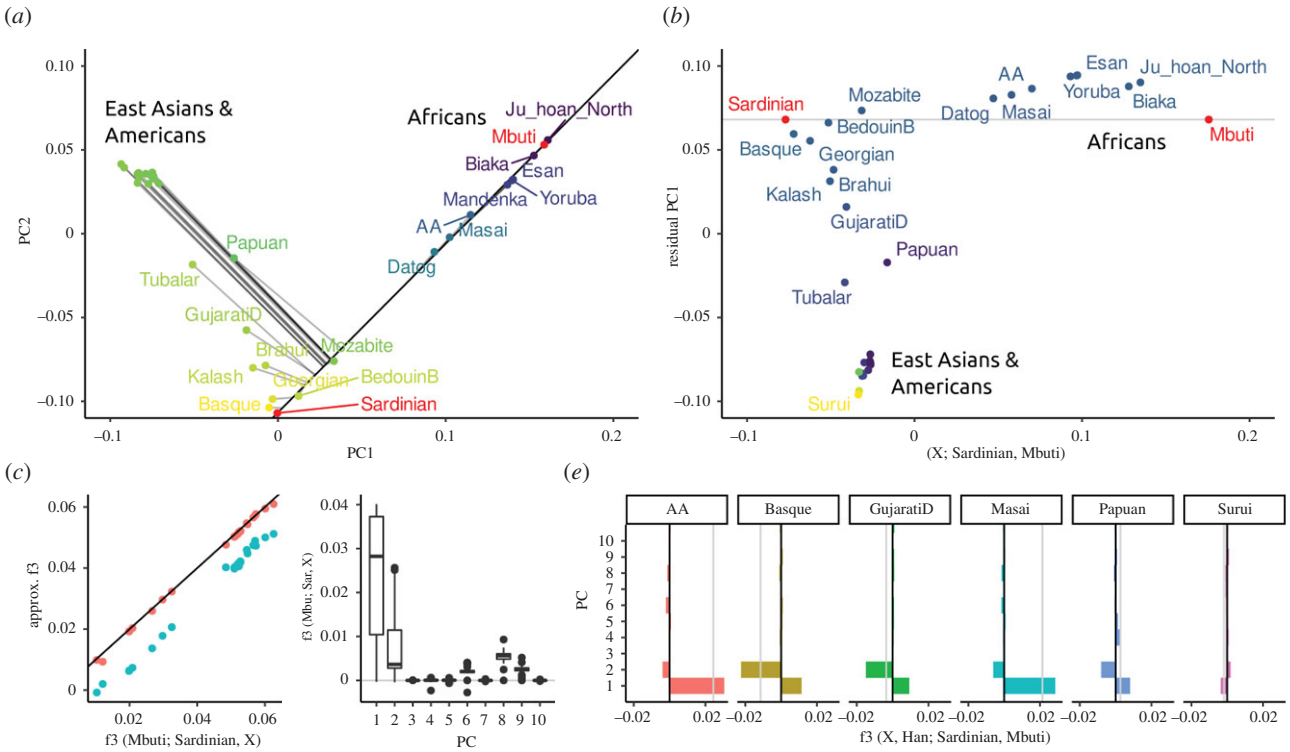

Figure 6. PCA and *F*-statistics for the World dataset. (*a*) Visualization of outgroup-F_3-statistic F_3(Mbuti; Sardinian, *X*) on a PCA biplot. The colour of points correspond to the value of the F_3-statistic, with brighter yellows indicating higher values, i.e. higher similarity to Sardinians. The F_3-projection axis is given by a black line, the projection of populations onto this axis by thin grey lines. In the full allele frequency space, these projection are orthogonal to the axis. (*b*) Projection along the axis Sardinian-Mbuti (*X*-axis), and PCA on residual of this projection (PC1 on *Y*-axis, PC2 as colouring). (*c*) Approximation of F_3(Mbuti; Sardinian, *X*) using the first two (blue) and first 10 (red) PCs, respectively. (*d*) Contributions of first 10 PCs to all statistics of the form F_3(Mbuti; Sardinian, *X*). (*e*) Contributions of the first ten PCs to select F_4-statistics. (Online version in colour.)

To make sense of outgroup-F_3-statistics in the PCA context, I use the association of F_3-statistics to projections (equation 2.5): on a PCA plot, we can visualize this F_3-statistic as the projection of the vector $X_i - X_O$ onto $X_U - X_O$

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

Of the right-hand-side terms, only the F_3 term depends on the X_i. The fraction can be thought of as a normalizing constant, so the F_3-statistic is proportional to the length of the projected vector. This means that the outgroup-F_3-statistic is largest for whichever X_i projects furthest along the axis from the outgroup to the unknown population; in figure 5, this is X_3.

(i) Example

In figure 6*a*, I use the World dataset to visualize the outgroup-F_3-statistic F_3 (Mbuti; Sardinian, X_i), in i.e. a statistic that aims to find the population most closely related to Sardinian (a Mediterranean island), assuming the Mbuti are an outgroup to all populations in the dataset. 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 Sardinians (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-Sardinian, but on the plot (i.e. the subspace spanned by the first two PCs), this is not necessarily the case. The colouring 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 and the Americas project almost 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 4*c*), 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 4*d*), PCs 6, 8 and 9 having a similar non-zero contribution for almost all statistics, likely because these PCs explain within-African variation.

(e) F_4-statistics as angles

One interpretation of F_4 on PCA plots 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 an internal branch in a tree (as in figure 1*c*) can be interpreted as being proportional to the length of a projected segment on a PCA plot, 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 (figure 1*d*), then, from the tree interpretation, we expect $F_4(X_1, X_2; X_3, X_4) = 0$ implying that the vectors $X_1 - X_2$ and $X_3 - X_4$ are orthogonal to each other, i.e. that X_1 and X_2 map to the same point on the projection axis $\overline{X_3X_4}$. In the case of an admixture graph, this is no longer the case: Both population X_y and X_x in figure 2*d* do *not* map to the same point as X_1 or X_2 do, implying that statistics of the form $F_4(X_1, X_x; X_3, X_4) \neq 0$.

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

$$\mathrm{Cor}(X_1 - X_2, X_3 - X_4) = \frac{F_4(X_1, X_2; X_3, X_4)}{\|X_1 - X_2\|\|X_3 - X_4\|^2} = \cos(\phi), \tag{4.2}$$

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°, whereas an angle close to zero ($\cos(\phi) \approx 1$) means most of the genetic drift on this branch is shared.

(i) Example

To illustrate the angle interpretation, I return to the Western Eurasian data. The PCA biplot shows two roughly parallel clines (figure 4*a*), a European gradient (from Sardinian to Finnish and Chuvash), and an Asian cline from Arab populations (top right) to the Caucasus (bottom right). This is quantified in figure 4*d*, where I plot the angle corresponding to $F_4(X, \text{Saudi}; \text{Sardinian}, \text{Finnish})$. For most Asian 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 a third PC (blue), however, shows that the clines are not, in fact, parallel, and the correlation for most populations is low. The finding that three PCs are necessary to explain this data can also be seen from the spectrum of these F_4-statistics (figure 4*e*), which have high contributions from the first three PCs. Both results indicate that adding a third PC would give a much better description of the data, and the relationship between within-European variation to Saudis in particular.

(f) Other projections

So far, I used equation (2.9) to interpret F-statistics on a PCA plot, but the argument holds for *any* orthonormal projection in the data space. This is useful in particular for estimates of admixture proportions, which are often done as projections into a low-dimensional reference space defined by F-statistics [38,40,49,53].

For example, a common way to estimate admixture proportion α of X_1 is the F_4-ratio

$$\alpha = \frac{F_4(R_1, R_2; X_X, X_1)}{F_4(R_1, R_2; X_2, X_1)} = \frac{proj_{[R_1 - R_2]}X_X - X_1}{proj_{[R_1 - R_2]}X_2 - X_1}, \tag{4.3}$$

which can be interpreted as projecting $X_X - X_1$ and $X_2 - X_1$ onto $R_1 - R_2$, and the ratio of the lengths gives the proportion of X_x contributed by X_1 [49].

The admixture graph motivating this statistic is visualized in figure 7*a*, and the PCA-like interpretation in figure 7*b*. In both panels, the solid grey line is the projection axis, and the dotted line gives the residual, i.e. the branches or genetic variation that is ignored by the projection.

The PCA-like projection can be used to visualize admixture proportions, as the horizontal position of X_X relative to X_1 and X_2 (red dashed line versus black line) directly represents the estimated admixture proportion α. In addition, the residuals can be used to verify assumptions of the admixture graph model. In particular, since X_X arises as a linear combination of X_1 and X_2, if admixture is recent we might expect the three populations to be collinear; if they are not this means that either of the populations experienced gene flow from some other population which might bias results [53].

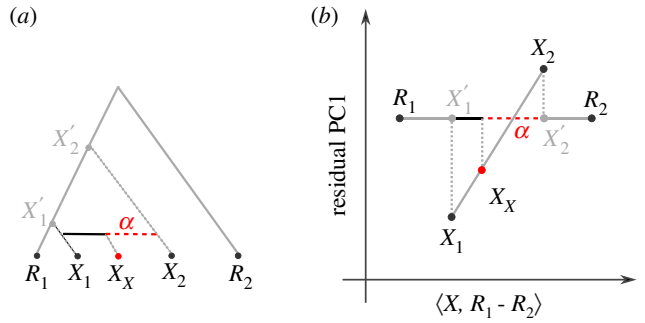

Figure 7. Admixture proportion estimates. (*a*) Visualization of the admixture graph scenario used to estimate the proportion α contributed from X_1 to X_X, using references R_1 and R_2. The full grey line corresponds to the projection axis, and the dotted grey lines correspond to the branches ignored in the projections. The admixture proportion α corresponds to the length of the dashed red line relative to the black line between X_1 and X_2. (*b*) The same scenario, but in Euclidean space, X_1, X_2 and X_X align on a line both in the (low-dimensional approximation of the) residual space and on the projection axis. (Online version in colour.)

In addition, the external tree branches $X_1 - X_1'$ and $X_2 - X_2'$ are disjoint which means they should be orthogonal. On a one-dimensional residual plot (figure 7*b*), this cannot be verified, but the statistic

$$F_4(X_1, X_1'; X_2, X_2') = 0 \tag{4.4}$$

can be calculated for all samples.

(i) Example

I use the World dataset as an example, using Sardinian and Mbuti as references populations (figure 6*b*). The data are the same as in the PCA (figure 6*a*), but it is now rotated such that the axis between the reference population (black line in figure 6*a*) is aligned with the x-axis. For any pair of populations X_1 X_2, their horizontal projection distance reflects F_4 (Sardinian, Mbuti; X_1, X_2) and the relative horizontal distance corresponds exactly to F_4-ratio admixture estimates. For many sets of populations, this is of course not sensible, and just looking at the first PC of the residual shows many examples where the populations are not collinear. For example, on the x-axis, the South American Surui are between Papuans and Georgians, but since the Surui clearly are not on the line between Papuans and Georgians, this cannot be the result of admixture.

5. Discussion

Particularly for the analysis of human genetic variation with a large number of individuals with heterogeneous relationships, F-statistics are a powerful tool to describe population genetic diversity. Here, I show that the geometry of F-statistics [49] leads to a number of simple interpretations of F-statistics on a PCA plot.

(a) The geometry of admixture

Previous interpretation of PCA in the context of population genetic models have focused on explicit models and aimed at directly interpreting the PCs in terms of population genetic parameters [16,23,46,48]. My interpretation here is different in

that the utility of PCA is to simplify the geometry of the data, rather than attributing meaning to the produced PCs. One consequence is that the results here are less directly impacted by sample ascertainment, sample sizes or number of PCs, which are common concerns in the interpretation of PCA [23,46–48]; adding more PCs will provide a successively better approximation of the *F*-statistics.

The two datasets I analysed here suggest that two PCs (for the World dataset), and three PCs (for the Western Eurasian data), respectively, already provide a very good approximation for F_4-statistics (figures 4*e* and 6*e*), reflecting the observation that frequently the first few PCs provide a good approximation of the overall population structure. On the other hand, for admixture F_3-statistics, more PCs are needed (figure 4*b*). This is likely due to PCA approximating the global structure in the dataset; statistics that only involve distantly related populations will only require a few PCs for good approximations, whereas statistics that contain a term measuring local variation, such as F_3-statistics or F_4 between closely related populations will require more PCs for good approximations, because local variation is often found on higher PCs.

My focus on the geometry of the data 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 may aid in generation of hypotheses that can be more directly evaluated using generative models, typically using a lower number of populations. It also allows reconciling apparent contradictions between *F*-statistics and PCA plots. In many cases, differences between the two data summaries will be due to variation on higher PCs. In this case, plotting additional PCs, or further subsetting the data to a more local set of populations seems prudent.

(b) Assumptions

In addition to the selection of PCs, the other cause for disagreements between *F*-statistics and PCA are differences in assumptions. The version of PCA I use for my analyses is chosen such that the similarities to *F*-statistics are maximized. In particular, I assume here that (i) we have no missing data, (ii) SNPs are equally weighted, (iii) that individuals can be grouped into populations and (iv) we use estimated allele frequencies. By contrast, most data analyses have to grapple with missing data, SNPs are often weighted according to their allele frequencies, and observed, individual-level genotypes are used as the basis of PCA.

(i) Missing data

The matrix decompositions underlying PCA assume complete data, and thus cannot be used when some data is missing [50]. As missing data are a very common practical problem, there is a large number of algorithms for imputing missing data. The simplest approach is to replace missing data with zeros (as implemented e.g. in [30]), but more sophisticated algorithms exist to 'learn' the missing values from surrounding data (e.g. [64,65]). By contrast, missing data in *F*-statistics is most commonly handled by estimating a standard error by resampling along the genome [40], and so missing data results in larger standard errors.

These strategies are distinct, and reflect the original purposes of the approaches. For statistical tests based on *F*-statistics, we wish to isolate a set of three or four populations and get our best guess based on just that subset of data. By contrast, methods for PCA can leverage the additional individuals, and thus will likely result in more accurate estimates.

However, the way we handle missing data is not tied into the method. For example, we could evaluate the robustness of a PCA by resampling data. Similarly, the theory developed here suggests that we could obtain accurate *F*-statistics with missing data by first performing a PCA using a method that handles missing data, and then calculate *F*-statistics from these PCs.

(ii) Normalization

In PCA, SNPs are typically normalized to have expected variance of one, a step that is omitted in calculating *F*-statistics [40]. The *F*-statistic framework assumes that each SNP is an identically distributed (but not independent) random variable, which holds regardless of weighting. Thus, normalization of SNPs is largely a matter of convention; for *F*-statistics the dependency on additional samples (through mean allele frequencies) is often unwanted, but could be advantageous for tools that aim to do joint inference from many *F*-statistics such as `qpAdm` [38,40]. As genetic differentiation between human populations is low, the normalization used may matter little in practice, but could be explored in future work [28].

(iii) Estimated versus observed allele frequencies

The third difference between *F*-statistics and PCA is on the usage of estimated allele frequencies versus individual-based genotypes. The fact that PCA does not distinguish between sampling error and the underlying structure is a well-known drawback of PCA, and applying the theory presented here to individual-based PCA would result in *F*-statistics that incorporate some sampling noise. Probabilistic PCA is one class of approaches that aim to separate the population structure from sampling noise (e.g. [66]). It seems likely that probabilistic PCA would yield a representation of the data that is more closely aligned with *F*-statistics than regular PCA.

(iv) Individual versus population-based analyses

The final issue is that PCA is commonly run on individual-based data, whereas *F*-statistics often group individuals into populations. However, population-based PCA has been the default in the past [1], and *F*-statistics are often applied to individuals (e.g. [25,67,68]). Often, an individual-based PCA is used to justify grouping individuals into populations; i.e. individuals that form a tight cluster on a PCA plot have similar relationships to everyone else in the dataset, and can thus be treated as a unit of analysis. Thus, if the assumptions are satisfied, *F*-statistics for individual-based and population-based analyses are expected to be very similar. PCA, on the other hand, is strongly impacted by the number of individuals from each population (e.g. [47]); as each individual is weighted equally, variation related to populations with many samples will be overrepresented on the first PCs.

(v) Summary

Motivated by *F*-statistics, the PCAs I consider here are based on estimated population allele frequencies, whereas most genome-scale studies of human genetic variation use observed individual-based allele frequencies that are normalized by overall allele frequencies. Thus, some care will be required to directly extend the interpretations developed here to individual-based PCAs. However, the differences are largely due to

conventions, and particularly for studies where the description of population structure is a major focus, results might be easier to interpret if conventions regarding missing data, normalization and estimation of allele frequencies are used consistently between *F*-statistics and PCA.

(c) The apportionment of human diversity

Most genetic variation in humans is shared between all of us, but the around 15% that can be explained by population structure can be leveraged to study our history and diversity in great detail [1,69,70]. For some datasets, it is possible to predict an individuals' origin at a resolution of a few hundred kilometres [45,71], and direct-to-consumer-genetics companies are using this variation to analyse the genetic data of millions of customers.

However, understanding, conceptualizing and modelling this variation is far from trivial, particularly in a historical context in which mistaken ideas about human variation have been used to justify racist, eugenic and genocidal policies. Lewontin's landmark 1972 paper on the apportionment of human genetic diversity was one of the first to quantify how little of between-population genetic variation could be attributed to 'racial' continental-scale groupings [70]. Over the last five decades, this view has been corroborated, refined and extended many times [1,72–74].

From a practical perspective, formulating hypotheses and designing studies in terms of discrete populations with 'uniform' genetic backgrounds is often sensible, as it enables e.g. prediction of phenotypes [75,76], inference of demographic parameters, and schematic models of human genetic history [40]. In a similar vein, when interpreting *F*-statistics in the context of admixture graphs, we make the implicit assumption that populations are discrete, related as a graph, and that gene flow between populations is rare [38,40]. However, these simplifications do come at a cost, both in terms of model violations that may invalidate statistical results, and in terms of deemphasizing that people do not rigidly fall into predefined genetic groups.

In many parts of the world, and particularly at more local scales, distinctions between populations begin to blur, and everyone could be considered admixed to some degree [77]. This provides a challenge for interpretation, as most F_3 and F_4-statistics will indicate departures from treeness. A naive interpretation of the *F*-statistics from my Eurasian example (figure 4*a*) would identify a substantial fraction of Europeans as (significantly) admixed between Finnish and Sardinians. By contrast, PCA reveals that the variation in this dataset is not due to a single event, and so an arguably better description of the dataset is one where Finnish and Sardinians lie on opposite ends of a more gradually structured population.

Thus, a more general way we could think about modelling population structure using *F*-statistics is as identifying orthogonal drift components. In a tree model, orthogonality arises because changes in allele frequencies on distinct branches of the tree are independent from each other (and high-dimensional random vectors are almost surely orthogonal). The classical model of admixture as a result of contact between long-separated and isolated populations is one of potentially many demographic models that results in non-orthogonality; gene flow of any kind will result in correlated genetic drift, and hence in non-zero F_3 or F_4-statistics.

Thinking of population structure in terms of orthogonal components may be abstract, but it is quite similar to how PCA is sometimes interpreted. In a PCA, we frequently make the informal observation that a particular PC is associated with variation within a specific region, or separates out distinct populations [1]. These associations are not exact, and are impacted e.g. by the size and composition of the analysed dataset. On the other hand, we could use *F*-statistics to formally test orthogonality, or to quantify correlations. Motivated by PCA, we could set up F_4 as tests of orthogonality due to either space (distinct populations evolve independently) or scale (i.e. between-population diversity is independent of within-population diversity). This slight generalization of *F*-statistics could allow us to reframe many questions about gene flow or divergence that are currently asked as tests of orthogonality, without assuming that lineages or admixture events are discrete.

6. Conclusion

F-statistics and PCA have both proven to be tremendously useful to study, visualize and test aspects of population structure. Here, I show that these approaches are closely connected, and highlight a few implications. First, PCA and *F*-statistics should never be treated as independent analyses. If they agree, this can be used as a sanity check that no major assumptions about, e.g. population groupings, are violated, but the underlying biological relationships investigated are the same in both approaches.

If sufficiently many PCs are considered, both F_3 and F_4-statistics do have simple interpretations in PCA space. If used as a test for admixture, F_3 corresponds to testing whether the admixed population lies in an *n*-sphere between the potential source population in PCA space. This result makes the informal notion that an admixed population should lie between its sources on a PCA plot more precise. Furthermore, I show that while it is necessary for an admixed population to lie between its sources on a PCA plot, this is not a sufficient condition and unadmixed populations may also project inside the *n*-sphere. By contrast, if a population falls outside the *n*-sphere on any PCA plot, this is sufficient to determine that this population has a positive F_3 statistic.

Interpretations of the outgroup-F_3-statistics and F_4-statistics in a PCA framework rely heavily on the geometric concepts of projections and orthogonality: in a tree or admixture graph framework, we can loosely interpret $F_4(A, B; C, D)$ as the overlap of the path from A to B onto the path from C to D, or equivalently as the projection of $A - B$ onto $C - D$; only edges shared between the two paths will contribute non-zero amounts to this statistic. This interpretation holds when we replace the populations by points in PCA space, and enables us to study population models beyond discrete graphs. Thus, the geometric framework enables us to expand the application of *F*-statistics beyond current uses, allows us to better understand the meaning of these statistics and helps us to avoid overinterpretations, particularly in cases when population structure is continuous.

Data accessibility. All results are based on public data. The code used to obtain all results is available at doi:10.5281/zenodo.6424178. The data are provided in electronic supplementary material [78].

Competing interests. I declare I have no competing interests.

Funding. I received no funding for this study.

Appendix A. Derivations

Depending on a reader's 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.

(a) *F*-statistics are invariant under a change-of-basis

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

In summary, the first row shows that F_2 on the centred 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 ks. Row five is obtained by expanding the outer sum and grouping terms by k. The final line is obtained by recognizing that $\mathbf{L}$ is an orthonormal basis, where dot products of different vectors have lengths zero.

Note that if we estimate F_2, unbiased estimators are obtained by subtracting the population heterozygosities H_i, H_j from the statistic. As these are scalars, they do not change above calculation.

(b) The region of negative F_3-statistics is an n-ball

Without loss of generality, assume that $X_1 = (r, 0, 0, \ldots)$ and $X_2 = (-r, 0, 0, \ldots)$, and let us assume that X_x has coordinates $(x_1, x_2, \ldots, x_S)$. Assuming $F_3(X_x; X_1, X_2) = 0$, equation (4.1) becomes

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

which is the equation of an n-sphere with radius r and centre 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

$$r^2 - k = \|X_x\|^2, \qquad \text{(A 3)}$$

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

(c) If a population lies outside the circle of this n-sphere in any two-dimensional projection, F_3 is positive

Assume the centre of the n-sphere $C = (X_1 + X_2/2) = (c_1, c_2, \ldots, c_S)$, and $X_x = (x_1, x_2, \ldots, x_S)$. Then,

$$\begin{aligned}
F_3(X_x; X_1, X_2) &= \|X_x - C\|^2 - r^2 \\
&= \underbrace{(x_1 - c_1)^2 + (x_2 - c_2)^2}_{> r^2} + \underbrace{\sum_{i=3}^{S} (x_i - c_i)^2}_{\geq 0} - r^2 \\
&> 0. \qquad \text{(A 4)}
\end{aligned}$$

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.

References

1. Cavalli-Sforza LL, Menozzi P, Piazza A. 1994 *The history and geography of human genes*. Princeton, NJ: Princeton University Press.
2. Marciniak S, Perry GH. 2017 Harnessing ancient genomes to study the history of human adaptation. *Nat. Rev. Genet.* **18**, 659–674. (doi:10.1038/nrg.2017.65)
3. Reich D. 2018 *Who we are and how we got here: alte DNA und die neue Wissenschaft der menschlichen Vergangenheit*. New York, NY: Pantheon.
4. Nielsen R, Akey JM, Jakobsson M, Pritchard JK, Tishkoff S, Willerslev E. 2017 Tracing the peopling of the world through genomics. *Nature* **541**, 302–310. (doi:10.1038/nature21347)
5. Witt K, Villanea F, Loughran E, Zhang X, Huerta-Sanchez E. 2022 Apportioning archaic variants among modern populations. *Phil. Trans. R. Soc. B* **377**, 20200411. (doi:10.1098/rstb.2020.0411)
6. Schraiber JG, Akey JM. 2015 Methods and models for unravelling human evolutionary history. *Nat. Rev. Genet.* **16**, 727–740. (doi:10.1038/nrg4005)
7. Orlando L *et al.* 2021 Ancient DNA analysis. *Nat. Rev. Methods Primers* **1**, 1–26. (doi:10.1038/s43586-020-00011-0)
8. The 1000 Genomes Project Consortium. 2015 A global reference for human genetic variation. *Nature* **526**, 68–74. (doi:10.1038/nature15393)
9. Mallick S *et al.* 2016 The Simons genome diversity project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206. (doi:10.1038/nature18964)
10. Serre D, Pääbo S. 2004 Evidence for gradients of human genetic diversity within and among continents. *Genome Res.* **14**, 1679–1685. (doi:10.1101/gr.2529604)
11. Rosenberg NA, Mahajan S, Ramachandran S, Zhao C, Pritchard JK, Feldman MW. 2005 Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genet.* **1**, e70. (doi:10.1371/journal.pgen.0010070)
12. Bradburd GS, Coop GM, Ralph PL. 2018 Inferring continuous and discrete population genetic structure

across space. *Genetics* **210**, 33–52. (doi:10.1534/genetics.118.301333)

13. Peter BM, Petkova D, Novembre J. 2020 Genetic landscapes reveal how human genetic diversity aligns with geography. *Mol. Biol. Evol.* **37**, 943–951. (doi:10.1093/molbev/msz280)
14. Gopalan S, Smith SP, Korunes K, Iman H, Ramachandran S, Goldberg A. 2022 Human genetic admixture through the lens of population genomics. *Phil. Trans. R. Soc. B* **377**, 20200410. (doi:10.1098/rstb.2020.0410)
15. Wahlund S. 1928 Zusammensetzung von Populationen und Korrelationserscheinungen vom Standpunkt der Vererbungslehre aus betrachtet. *Hereditas* **11**, 65–106. (doi:10.1111/j.1601-5223.1928.tb02483.x)
16. Cavalli-Sforza LL, Piazza A. 1975 Analysis of evolution: evolutionary rates, independence and treeness. *Theor. Popul. Biol.* **8**, 127–165. (doi:10.1016/0040-5809(75)90029-5)
17. Wright S. 1943 Isolation by distance. *Genetics* **28**, 114–138. (doi:10.1093/genetics/28.2.114)
18. Slatkin M. 1985 Gene flow in natural populations. *Annu. Rev. Ecol. Syst.* **16**, 393–430. (doi:10.1146/annurev.es.16.110185.002141)
19. Barbujani G, Sokal RR. 1990 Zones of sharp genetic change in Europe are also linguistic boundaries. *Proc. Natl Acad. Sci. USA* **87**, 1816–1819. (doi:10.1073/pnas.87.5.1816)
20. Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. 2005 Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc. Natl Acad. Sci. USA* **102**, 15 942–15 947. (doi:10.1073/pnas.0507611102)
21. Stoneking M. 2016 *An introduction to molecular anthropology*. Hoboken, NJ: John Wiley & Sons.
22. Li S, Schlebusch C, Jakobsson M. 2014 Genetic variation reveals large-scale population expansion and migration during the expansion of Bantu-speaking peoples. *Proc. R. Soc. B* **281**, 20141448. (doi:10.1098/rspb.2014.1448)
23. François O, Currat M, Ray N, Han E, Excoffier L, Novembre J. 2010 Principal component analysis under population genetic models of range expansion and admixture. *Mol. Biol. Evol.* **27**, 1257–1268. (doi:10.1093/molbev/msq010)
24. Alves I, Arenas M, Currat M, Sramkova Hanulova A, Sousa VC, Ray N, Excoffier L. 2016 Long-distance dispersal shaped patterns of human genetic diversity in Eurasia. *Mol. Biol. Evol.* **33**, 946–958. (doi:10.1093/molbev/msv332)
25. Green RE *et al.* 2010 A draft sequence of the Neandertal genome. *Science* **328**, 710. (doi:10.1126/science.1188021)
26. Boca SM, Huang L, Rosenberg NA. 2020 On the heterozygosity of an admixed population. *J. Math. Biol.* **81**, 1217–1250. (doi:10.1007/s00285-020-01531-9)
27. Cavalli-Sforza LL, Edwards AWF. 1967 Phylogenetic analysis: models and estimation procedures. *Evolution* **21**, 550–570. (doi:10.1111/j.1558-5646.1967.tb03411.x)
28. Felsenstein J. 1973 Maximum-likelihood estimation of evolutionary trees from continuous characters. *Am. J. Hum. Genet.* **25**, 471–492.
29. Jombart T, Pontier D, Dufour AB. 2009 Genetic markers in the playground of multivariate analysis. *Heredity* **102**, 330–341. (doi:10.1038/hdy.2008.130)
30. Patterson N, Price AL, Reich D. 2006 Population structure and eigenanalysis. *PLoS Genet.* **2**, e190. (doi:10.1371/journal.pgen.0020190)
31. Pritchard JK, Stephens M, Donnelly P. 2000 Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959. (doi:10.1093/genetics/155.2.945)
32. Alexander DH, Novembre J, Lange K. 2009 Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664. (doi:10.1101/gr.094052.109)
33. Lessa EP. 1990 Multidimensional analysis of geographic genetic structure. *Syst. Zool.* **39**, 242–252. (doi:10.2307/2992184)
34. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009 Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* **5**, e1000695. (doi:10.1371/journal.pgen.1000695)
35. Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. 2013 Robust demographic inference from genomic and SNP data. *PLoS Genet.* **9**, e1003905. (doi:10.1371/journal.pgen.1003905)
36. Kamm JA, Terhorst J, Song YS. 2015 Efficient computation of the joint sample frequency spectra for multiple populations. *arXiv,* 1503.01133. (doi:10.48550/arxiv.1503.01133)
37. Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, Yu F, Gibbs RA, Bustamante CD. 2011 Demographic history and rare allele sharing among human populations. *Proc. Natl Acad. Sci. USA* **108**, 11 983–11 988. (doi:10.1073/pnas.1019276108)
38. Harney E, Patterson N, Reich D, Wakeley J. 2021 Assessing the performance of qpAdm: a statistical tool for studying population admixture. *Genetics* **217**, 1061. (doi:10.1093/genetics/iyaa045)
39. Reich D, Thangaraj K, Patterson N, Price AL, Singh L. 2009 Reconstructing Indian population history. *Nature* **461**, 489–494. (doi:10.1038/nature08365)
40. Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D. 2012 Ancient admixture in human history. *Genetics* **192**, 1065–1093. (doi:10.1534/genetics.112.145037)
41. Peter BM. 2016 Admixture, population structure, and *F*-statistics. *Genetics* **202**, 1485–1501. (doi:10.1534/genetics.115.183913)
42. Semple C, Steel MA. 2003 *Phylogenetics*. Oxford, UK: Oxford University Press.
43. Huson DH, Rupp R, Scornavacca C. 2010 *Phylogenetic networks: concepts, algorithms and applications*. Cambridge, UK : Cambridge University Press.
44. Cavalli-Sforza LL, Barrai I, Edwards AWF. 1964 Analysis of human evolution under random genetic drift. *Cold Spring Harb. Symp. Quant. Biol.* **29**, 9–20. (doi:10.1101/SQB.1964.029.01.006)
45. Novembre J *et al.* 2008 Genes mirror geography within Europe. *Nature* **456**, 98–101. (doi:10.1038/nature07331)
46. Novembre J, Stephens M. 2008 Interpreting principal component analyses of spatial population genetic variation. *Nat. Genet.* **40**, 646–649. (doi:10.1038/ng.139)
47. McVean G. 2009 A genealogical interpretation of principal components analysis. *PLoS Genet.* **5**, e1000686. (doi:10.1371/journal.pgen.1000686)
48. François O, Gain C. 2021 A spectral theory for Wright's inbreeding coefficients and related quantities. *PLoS Genet.* **17**, e1009665. (doi:10.1371/journal.pgen.1009665)
49. Oteo-Garcia G, Oteo JA. 2021 A geometrical framework for *f*-statistics. *Bull. Math. Biol.* **83**, 1–22. (doi:10.1007/s11538-020-00850-8)
50. Jolliffe IT. 2013 *Principal component analysis*. New York, NY: Springer Science & Business Media.
51. Pachter L. 2014 What is principal component analysis? See https://liorpachter.wordpress.com/2014/05/26/what-is-principal-component-analysis/.
52. Lipson M, Loh PR, Levin A, Reich D, Patterson N, Berger B. 2013 Efficient moment-based inference of admixture parameters and sources of gene flow. *Mol. Biol. Evol.* **30**, 1788–1802. (doi:10.1093/molbev/mst099)
53. Petr M, Pääbo S, Kelso J, Vernot B. 2019 Limits of long-term selection against Neandertal introgression. *Proc. Natl Acad. Sci. USA* **116**, 1639–1644. (doi:10.1073/pnas.1814338116)
54. Buneman P. 1974 A note on the metric properties of trees. *J. Comb. Theory Ser. B* **17**, 48–50. (doi:10.1016/0095-8956(74)90047-1)
55. Engelhardt BE, Stephens M. 2010 Analysis of population structure: a unifying framework and novel methods based on sparse factor analysis. *PLoS Genet.* **6**, e1001117. (doi:10.1371/journal.pgen.1001117)
56. Duforet-Frebourg N, Luu K, Laval G, Bazin E, Blum MGB. 2016 Detecting genomic signatures of natural selection with principal component analysis: application to the 1000 genomes data. *Mol. Biol. Evol.* **33**, 1082–1093. (doi:10.1093/molbev/msv334)
57. Gower JC. 1966 Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* **53**, 325–338. (doi:10.1093/biomet/53.3-4.325)
58. Lazaridis I *et al.* 2014 Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**, 409–413. (doi:10.1038/nature13673)
59. Lazaridis I *et al.* 2016 Genomic insights into the origin of farming in the ancient Near East. *Nature* **536**, 419–424. (doi:10.1038/nature19310)
60. Haak W *et al.* 2015 Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **522**, 207–211. (doi:10.1038/nature14317)
61. Ralph P, Coop G. 2013 The geography of recent genetic ancestry across Europe. *PLoS Biol.* **11**, e1001555. (doi:10.1371/journal.pbio.1001555)

62. Brisbin A *et al.* 2012 PCAdmix: principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. *Hum. Biol.* **84**, 343–364. (doi:10.3378/027.084.0401)
63. Raghavan M *et al.* 2014 Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature* **505**, 87–91. (doi:10.1038/nature12736)
64. Hastie T, Mazumder R, Lee JD, Zadeh R. 2015 Matrix completion and low-rank SVD via fast alternating least squares. *J. Mach. Learn. Res.* **16**, 3367–3402.
65. Meisner J, Liu S, Huang M, Albrechtsen A. 2021 Large-scale inference of population structure in presence of missingness using PCA. *Bioinformatics* **37**, 1868–1875. (doi:10.1093/bioinformatics/btab027)
66. Agrawal A, Chiu AM, Le M, Halperin E, Sankararaman S. 2020 Scalable probabilistic PCA for large-scale genetic variation data. *PLoS Genet.* **16**, e1008773. (doi:10.1371/journal.pgen.1008773)
67. Massilani D *et al.* 2020 Denisovan ancestry and population history of early East Asians. *Science* **370**, 579–583. (doi:10.1126/science.abc1166)
68. Yang MA *et al.* 2020 Ancient DNA indicates human population shifts and admixture in northern and southern China. *Science* **369**, 282–288. (doi:10.1126/science.aba0909)
69. Lewontin RC. 1972 The apportionment of human diversity. In *Evolutionary biology* (eds T Dobzhansky, MK Hecht, WC Steere), pp. 381–398. New York, NY: Springer. See http://link.springer.com/10.1007/978-1-4684-9063-3_14 (accessed 20 May 2021).
70. Novembre J. 2022 The background and legacy of Lewontin's apportionment of diversity. *Phil. Trans. R. Soc. B* **377**, 20200406. (doi:10.1098/rstb.2020.0406)
71. Leslie S *et al.* 2015 The fine-scale genetic structure of the British population. *Nature* **519**, 309–314. (doi:10.1038/nature14230)
72. Cann RL, Stoneking M, Wilson AC. 1987 Mitochondrial DNA and human evolution. *Nature* **325**, 31–36. (doi:10.1038/325031a0)
73. Barbujani G, Magagni A, Minch E, Cavalli-Sforza LL. 1997 An apportionment of human DNA diversity. *Proc. Natl Acad. Sci. USA* **94**, 4516–4519. (doi:10.1073/pnas.94.9.4516)
74. Rosenberg NA *et al.* 2002 Genetic structure of human populations. *Science* **298**, 2381–2385. (doi:10.1126/science.1078311)
75. Berg JJ *et al.* 2019 Reduced signal for polygenic adaptation of height in UK Biobank. *eLife* **8**, e39725. (doi:10.7554/eLife.39725)
76. Yair S, Coop G. 2021 Population differentiation of polygenic score predictions under stabilizing selection. *bioRxiv*. (doi:10.1101/2021.09.10.459833)
77. Pickrell JK, Reich D. 2014 Toward a new history and geography of human genes informed by ancient DNA. *Trends Genet.* **30**, 377–389. (doi:10.1016/j.tig.2014.07.007)
78. Peter BM. 2022 A geometric relationship of F_2, F_3 and F_4-statistics with principal component analysis. Figshare. (https://doi.org/10.6084/m9.figshare.c.5898677)