A Spectral Theory for Wright's Inbreeding Coefficients and Related Quantities

Olivier François Clément Gain

Authors' affiliations:

Université Grenoble-Alpes, Centre National de la Recherche Scientifique, Grenoble INP, TIMC-IMAG UMR 5525, 38000 Grenoble, France.

Corresponding author:

olivier.francois@univ-grenoble-alpes.fr

Keywords: Inbreeding Coefficient, Principal Component Analysis, Eigenvalues, Separation Condition, Random Matrix Theory

Abstract

Wright's inbreeding coefficient, F_{ST}, is a fundamental measure in population genetics. Assuming a predefined population subdivision, this statistic is classically used to evaluate population structure at a given genomic locus. With large numbers of loci, unsupervised approaches such as principal component analysis (PCA) have, however, become prominent in recent analyses of population structure. In this study, we describe the relationships between Wright's inbreeding coefficients and PCA for a model of K discrete populations. Our theory provides an equivalent definition of F_{ST} based on the decomposition of the genotype matrix into between and within-population matrices. Assuming that a separation condition is fulfilled, our main result states that the proportion of genetic variation explained by the first $(K-1)$ principal components can be accurately approximated by the average value of F_{ST} over all loci included in the genotype matrix. This equivalent definition of F_{ST} can be used to evaluate the fit of discrete population models to the data. It is also useful for computing inbreeding coefficients from surrogate genotypes, for example, obtained after correction of experimental artifacts or after removing genetic variation associated with environmental variables. The relationships between inbreeding coefficients and the spectrum of the genotype matrix not only allow interpretations of PCA results in terms of population genetic concepts but extend those concepts to population genetic analyses accounting for temporal, geographical and environmental contexts.

Introduction

Defined by Sewall Wright and Gustave Malécot, the fixation index or inbreeding coefficient, F_{ST}, measures the amount of genetic diversity found between populations relative to the amount within populations [1, 2]. Used as a measure of population differentiation, F_{ST} is among the most widely used descriptive statistics in population and evolutionary genetics [3, 4, 5, 6, 7]. Inbreeding coefficients were originally defined for the analyses of allelic frequencies at a single genetic locus. With the amount of data available to present-day or ancient population genomic studies, principal component analysis (PCA) and model-based estimation algorithms, such as STRUCTURE, have emerged as alternative ways to describe population structure from multilocus genotype matrices [8, 9, 10, 11, 12].

Assuming that the columns of the genotype matrix are either centered or scaled, PCA computes the eigenvalues and eigenvectors of the sample covariance matrix. The first eigenvectors – or axes – summarize the directions which account for most of the genetic variation, and the eigenvalues represent the variances of projected samples along the axes. Eigenvalues and eigenvectors can be computed efficiently by using the singular value decomposition (SVD) of the column-centered data matrix [13]. PCA has been considered very early in human biology, and has become a popular method to study the genetic structure of populations [14, 15]. Inference from PCA is justified from the fact that, similarly to STRUCTURE, the projections of individuals on principal axes reveal their degree of admixture with source populations when these sources are represented in the sample [10, 16, 17, 18].

Although the relationships between PCA projections and admixture estimates are well-understood, difficulties to interpret PCA eigenvalues still remain. The main

contributions in that direction were restricted to models of two-population divergence, and their arguments were based on random matrix theory (RMT)[10, 19] and coalescent theory [16]. We note that connections between $F_{\rm ST}$ and PCA are not only important for description of population structure, but also in genome scans for selection where the distribution of PCA loadings can be used to detect regions with signature of divergent selection [20, 21, 22, 23]. Based on RMT, Ref. [10] proposed a threshold value of $F_{\rm ST}$ for two populations with equal sample sizes [10]. Below the threshold, there should be essentially no evidence of population structure. The coalescent approach relied on a relationship between $F_{\rm ST}$ and coalescent time for a pair of genes from a single subpopulation and that of a pair of genes from the collection of subpopulations [6]. For a model of divergence between two populations, theoretical results for coalescent times were used to demonstrate a link between the leading eigenvalue and $F_{\rm ST}$ [16]. Results in Ref.[16] could be extended to simple models of population structure with explicit formulas for coalescent times [24]. While coalescent theory and RMT have provided relationships between $F_{\rm ST}$ and PCA in simple cases, the general conditions under which they are valid and their extensions to more than two populations are unknown.

In this study, we develop a spectral theory of genotype matrices to investigate the relationships between PCA and Wright's coefficients in discrete population models. Our theoretical framework assumes that the observed genotypes correspond to the sampling of K discrete populations. Decomposing the genotype matrix as a sum of between and within-population matrices, we extend the results obtained in [10, 16, 19, 25]. Our main result states that the mean value of $F_{\rm ST}$ over loci is equal to the squared norm of the between-population matrix. Under a separation condition

bearing on the between and within-population matrices, the sum of the first $(K-1)$ eigenvalues of scaled PCA approximates the mean value of $F_{\rm ST}$ over loci. To describe residual variation not explained by the discrete population model, we rely on RMT to approximate the eigenvalues of the within-population matrix [10, 26]. A corollary of the theory is an alternative definition of inbreeding coefficients that allows us to extend $F_{\rm ST}$ to adjusted or *surrogate* genotypes, such as genotype likelihoods and other modifications of allele counts [27]. To illustrate the new definition, we compute $F_{\rm ST}$ for ancient human DNA samples after performing correction for genomic coverage and for distortions due to difference in sample ages [28]. In a second application, we compute $F_{\rm ST}$ for Scandinavian samples of *Arabidopsis thaliana* after removing genetic variation associated with environmental variables taken from a climate database [29, 30].

Results and Discussion

Partitioning of genetic variation. Consider a sample of n unrelated individuals for which a large number of loci are genotyped, resulting in a matrix, $\mathbf{X} = (x_{i\ell})$, with n rows and L columns. For haploids, we set $x_{i\ell} = 0, 1$, and for diploids $x_{i\ell} = 0, 1, 2$ to count the number of derived alleles at locus ℓ for individual i. Dealing with autosomes, we simplify our presentation by considering a sample of diploids as being represented by a sample of haploids having twice the original sample size. For unphased data, we take a random phase. Although not a necessary condition, the loci are assumed to be statistically independent, or obtained after an LD-pruning algorithm applied to the genotype matrix [20, 31]. Our main assumption is that individuals are sampled from K predefined discrete populations with no admixed individuals. Among other

models, examples of discrete population models underlying our assumptions include Wright's island models and coalescent models of divergence [32, 6, 33]. Application to the F-model [33] will be described afterwards.

To analyze population structure, PCA is performed after scaling or centering the genotype matrix. The transformed matrix is denoted by $\mathbf{Z}$. Scaled PCA computes the eigenvalues, $\rho_k^2(\mathbf{Z})$, of the empirical correlation matrix. Unscaled centered PCA computes the eigenvalues, $\sigma_k^2(\mathbf{Z})$, of the empirical covariance matrix [9, 26]. The eigenvalues are ranked in decreasing order, and $\rho_k^2(\mathbf{Z})/L$ is usually interpreted as the proportion of variance explained by the kth axis of the PCA. PCA can be performed via the SVD algorithm. In this case, the eigenvalues of scaled (or centered) PCA correspond to the squared singular values of the scaled (or centered) matrix divided by $\sqrt{n}$ [9, 26].

To establish relationships between PCA and inbreeding coefficients, we decompose the centered matrix into a sum of two matrices, $\mathbf{Z} = \mathbf{Z}_{\mathrm{ST}} + \mathbf{Z}_{\mathrm{S}}$, corresponding to between and within-population components. The decomposition is done as follows. At a particular locus, let i be an individual sampled from population k. The genotype, $x_{i\ell}$, is drawn from the binomial distribution, $\mathrm{bin}(d = 1, p_{k\ell})$, where $p_{k\ell}$ is the derived allele frequency in population k at locus ℓ. The coefficient of the centered matrix, $z_{i\ell}$, is equal to $z_{i\ell} = \sum_{j \neq k} c_j(p_{k\ell} - p_{j\ell}) + (x_{i\ell} - p_{k\ell})$, where $c_k = n_k/n$, represents the proportion of individuals from population k. In this formulation, the between-population matrix, $\mathbf{Z}_{\mathrm{ST}}$, has general term $z_{i\ell}^{st} = \sum_{j \neq k} c_j(p_{k\ell} - p_{j\ell})$, repeated for all individuals in population k. The within-population matrix, $\mathbf{Z}_{\mathrm{S}}$, has general term $z_{i\ell}^{s} = x_{i\ell} - p_{k\ell}$. A very similar decomposition holds for the scaled matrix as well (See Box 1 for notations).

Spectral analysis: Inbreeding coefficients and PCA eigenvalues. For samples from K discrete populations and F_{ST} defined according to Wright [1] and Nei [4, 34], our main result states that the mean value of F_{ST} across loci can be computed from the singular values of the between-population scaled matrix. Similar relationships are also established for D_{ST} and for the unscaled matrix. Key arguments for those results are provided in *Methods* and in *Supplementary Information*. More precisely, let $K \geq 2$ and $\mathbf{Z}$ be the scaled genotype matrix. Let us define $\mathbf{Z}_{\mathrm{ST}}$ and $\mathbf{Z}_{\mathrm{S}}$ as in the previous section and scale each column with $\sqrt{P(1-P)}$. The mean value of F_{ST} across loci can be computed from the singular values of the between-population matrix as follows

$$\mathbb{E}\left[F_{\mathrm{ST}}\right] = \sum_{k=1}^{K-1} \rho_k^2(\mathbf{Z}_{\mathrm{ST}})/L\,. \tag{1}$$

The mean value of F_{ST} across loci can be approximated from the $(K-1)$ leading eigenvalues values of the PCA,

$$\mathbb{E}\left[F_{\mathrm{ST}}\right] \approx \sum_{k=1}^{K-1} \rho_k^2(\mathbf{Z})/L\,, \tag{2}$$

if (and only if) the following *separation condition* holds

$$\rho_{K-1}^2(\mathbf{Z}_{\mathrm{ST}}) > \rho_1^2(\mathbf{Z}_{\mathrm{S}})\,, \tag{3}$$

where $\rho_{K-1}(\mathbf{Z}_{\mathrm{ST}})$ is smallest non-null singular value of $\mathbf{Z}_{\mathrm{ST}}/\sqrt{n}$ and $\rho_1(\mathbf{Z}_{\mathrm{S}})$ is the largest singular value of $\mathbf{Z}_{\mathrm{S}}/\sqrt{n}$. For the centered genotype matrix, the separation

condition can be formulated as

$$\sigma^2_{K-1}(\mathbf{Z}_{\mathrm{ST}}) > \sigma^2_1(\mathbf{Z}_{\mathrm{S}})\,, \tag{4}$$

and we have

$$\mathbb{E}\left[D_{\mathrm{ST}}\right] \approx 2\sum_{k=1}^{K-1}\sigma^2_k(\mathbf{Z})/L\,. \tag{5}$$

The average value of F_{ST} is given by equation (1) regardless of the separation condition. Separation conditions (3) and (4) can be evaluated by computing the SVD for the between and within-population matrices. The computational cost of those operations is similar to the computational cost of a PCA of the genotype matrix (of order $O(n^2L)$). All conclusions remain valid when genotypes are conditioned on having minor allele frequency greater than a given threshold.

Besides an interest in connecting population genetic theory to approaches adopted in analysis of recent genomic data, the results in equations (1), (2) and (5) have several important consequences. Firstly, failure to verify the separation condition may be an indication of an insufficient sample size or that the data disagree with the way the K populations were predefined. Potential sources of departures from the conclusion may also include incorrect assignment of individuals to populations, admixed individuals or spatial structure. Regarding the experimental design, the result confirms that the influence of uneven sampling and MAF thresholding on PCA are similar to their effects on F_{ST} [16, 31, 35]. Secondly, equation (2) clarifies the debate over the definition of F_{ST} across loci, and supports the definition of F_{ST} as an average of ratios rather than a ratio of averages [36, 37]. Thirdly, our result also clarifies the connection between PCA and drift statistics considered in analysis of

demographic history of populations in Ref. [38, 39]. To formulate this link, consider the covariance matrix of the random vector $\mathbf{z}$, defined by $\mathbf{z}_k = \sqrt{c_k}(p_k - P)$, for all $k = 1$ to K. The covariance matrix can be obtained from the drift statistics $\mathcal{F}_2$ et $\mathcal{F}_3$ [38, 39], as $\mathbf{\Lambda}_{j,k} = \sqrt{c_j c_k}\ \mathbb{E}[(p_j - P)(p_k - P)] = \sqrt{c_j c_k}\,\mathcal{F}_3(P; p_j, p_k)$, for $j \neq k$, and $\mathbf{\Lambda}_{k,k} = c_k \mathbb{E}[(p_k - P)^2] = c_k \mathcal{F}_2(P; p_k)$ otherwise. The value $\mathbb{E}[D_{\mathrm{ST}}]$ can be approximated by (twice) the trace of the $\mathbf{\Lambda}$ matrix, and the eigenvalues are functions of the $\mathcal{F}_2$ et $\mathcal{F}_3$ statistics. For the F-model [33], the eigenvalues of $\mathbf{\Lambda}$ can be analysed formally for small numbers of populations (*Supplementary Information*).

Approximation of residual variation from RMT. For discrete population models such as F-models, approximations of singular values for the within-population (residual) matrix can be obtained from RMT [10, 40, 41, 42, 43]. Verifying condition (3), the leading eigenvalue of the within-population matrix can be approximated as $(1 - F_{\mathrm{ST}}) \times (1/\sqrt{L} + 1/\sqrt{n - K})^2$ for scaled PCA. A similar approximation can be obtained for centered PCA after replacing the term $(1 - F_{\mathrm{ST}})$ by the variance of coefficients in the within-population matrix. For two populations with equal sample sizes, the separation condition writes as $F_{\mathrm{ST}}/(1 - F_{\mathrm{ST}}) > (1/\sqrt{L} + 1/\sqrt{n - 1})^2$, defining a new threshold for F_{ST} below which population structure cannot be detected. If there truly is a single population represented in the total sample, then F_{ST} for two equal size samples should be of order $(1/\sqrt{L} + 1/\sqrt{n - 1})^2$. Those thresholds provide an informal test for a K population model to describe the data in an appropriate way.

Single population models. In a series of simulations of single population models, we first investigated whether RMT predictions accurately approximated the leading eigenvalue of scaled PCA. The results supported that the leading eigenvalues of PCA

were accurately predicted by RMT in F-models without population structure (Supplementary Figure S1). Then we investigated whether condition (3) could be verified when there was no structure in the data, and two population samples were (wrongly) defined from a preliminary structure analysis. We ran two-hundred simulations of single population models ($n = 100$ and $L \approx 10,000$), and, for each data set, we partitioned the samples in two groups according to the sign of their first principal component. This procedure maximized the likelihood of detecting artificial groups, leading to an average $F_{\mathrm{ST}} \approx 1.1\%$. For those artificial groups, we computed the non-null singular value of the between-population matrix, $\mathbf{Z}_{\mathrm{ST}}/\sqrt{n-1}$, and the leading singular value of the within-group matrix, $\mathbf{Z}_{\mathrm{S}}/\sqrt{n-1}$. For the simulations, the separation condition was never verified, rejecting population structure in all cases (Supplementary Figure S2A). For smaller sample sizes ($n = 10$ and $L \approx 1,000$), the separation condition was erroneously checked in 21% simulations, indicating that we had less power to discriminate among artificial groups with small sample sizes (Supplementary Figure S2B). Those results were consistent with difficulties reported for between-group PCA [44].

Two-population models. To check whether the expected values of F_{ST} and D_{ST} were obtained from the first eigenvalues of PCA, we performed simulations of F-models with two populations. For these simulations, the separation condition was verified in 100% data sets. There was an almost perfect fit of the leading eigenvalue for centered PCA, $\sigma_1^2(\mathbf{Z})$, with the average value of $D_{\mathrm{ST}}/2$ across loci and with the theoretical value of $\mathbb{E}[D_{\mathrm{ST}}]/2$ in F-models (Figure 1A, Supplementary Information, Supplementary Figure S3). There was also an almost perfect fit of the leading eigen-

value of scaled PCA, $\rho_1^2(\mathbf{Z})$, with the average value of F_{ST} across loci (Figure 1C). The second largest eigenvalues were accurately predicted by RMT both for centered and for scaled PCA (Figure 1 B-D). To detail those results for particular values of drift coefficients, we performed additional simulations for $F_1 = F_2 = 7\%$, also investigating the distribution of eigenvalues of the residual matrix (Figure 2). In a sample of $n = 200$ individuals and $L \approx 85,500$ SNPs, the first PC axis explained 3.11% of total genetic variation, corresponding to the average of F_{ST} across loci (3.11%, Figure 2A). The separation of between and within-population components was verified, and the second eigenvalue (0.536%) was very close to its prediction from RMT, given by $(1-\rho_1^2(\mathbf{Z})) \times (1/\sqrt{L}+1/\sqrt{n-2})^2 = 0.537\%$ (Figure 2A). The distribution of empirical residual eigenvalues, corresponding to within-population variation, was accurately modelled by the Marchenko-Pastur probability density function (Figure 2B). With a smaller sample of $n = 20$ individuals in each sample and $L \approx 12,500$ SNPs, the leading axis explained 5.24% of the total genetic variation, still matching the value of F_{ST} across loci (5.23% Figure 2C). The Marchenko-Pastur density remained an accurate approximation to the bulk spectrum of residual eigenvalues (Figure 2D). To provide evidence that the relationships between PCA eigenvalues and F_{ST} could be verified by real data, we computed them for pairs of human population samples from The 1000 Genomes Project [45]. At the exception of the CEU-IBS samples, the separation condition was verified in all pairwise comparisons. The leading eigenvalue of scaled PCA was accurately approximated by $\mathbb{E}[F_{\mathrm{ST}}]$, and the leading eigenvalue of the residual matrix was accurately approximated by RMT (Table 1).

Next, we studied the relationship between leading eigenvalues and sample size, for $L = 100$ loci and $L = 100,000$ loci (Supplementary Figure S4). For smaller

number of loci ($L = 100$) and smaller samples ($n \leq 80$), the data failed to verify the separation condition in some simulations. In those cases, population structure was not correctly evaluated by F_{ST}. The separation condition was verified in about 35% cases for $n = 10$ and in about 95% cases for $n \approx 80$. For the larger number of loci ($L \geq 100000$) or for larger sample sizes ($n \geq 100$), the separation condition was verified in all cases, and the leading eigenvalue converged to the theoretical value of $\mathbb{E}[F_{\mathrm{ST}}]$ for an infinite sample size. As for between-group PCA, the results suggest exaggerated differences among groups when sample sizes are very small relative to the number of loci [46].

Three-population models. We performed simulations of three-population F-models to check whether the data agreed with theoretical predictions for the leading eigenvalues, λ_1, and for D_{ST} and F_{ST}. With random drift coefficients ($n = 100$, $L = 20000$), the separation condition was verified in all simulated data sets. An almost perfect agreement between $\lambda_1 + \lambda_2$ and the mean value of $D_{\mathrm{ST}}/2$ (unscaled PCA) or F_{ST} (scaled PCA) was observed (Supplementary Figure S5 AC). The leading eigenvalue of unscaled PCA exhibited a small but visible bias with respect to the value predicted for λ_1 (Supplementary Figure S5 B). The third eigenvalue of scaled PCA was close to the approximation provided by RMT (Supplementary Figure S5 D). To study cases in which the separation condition was not verified, we considered smaller number of genotypes ($L \leq 1000$) and lower values of drift coefficients ($F_k \leq 10\%$). For small values of n and L, a significant proportion of simulated data sets did not verify the separation condition (Supplementary Figure S6). Even for correctly specified models, those results provided additional evidence of biases in

analyses of population structure with small data sets.

Ancient DNA data. This paragraph and the next one illustrate how spectral estimates can be used to evaluate inbreeding coefficients from genotypes obtained after correction for experimental or environmental effects. First, we studied ancient DNA samples from early farmers from Anatolia (EFA, $n = 23$), steppe pastoralists from the Yamnaya culture (Steppe, $n = 15$), and Western hunter-gatherers from Serbia (WHG, $n = 31$) [17, 47, 48, 49]. To estimate F_{ST} from those samples, we performed adjustment of pseudo-haploid genotypes for genomic coverage and for temporal distortions created by genetic drift (which were not expected to modify F_{ST}). After genotypes were adjusted for coverage and corrected for distortions due to differences in sample ages, the resulting values could no longer be interpreted as allelic frequencies. Estimates of adjusted F_{ST} were equal to 4.7% for the *EFA – Steppe* paired data set, 5.8% for *EFA – WHG*, 5.1% for *Steppe – WHG* (Table 1). Separation conditions were verified, and there was evidence of population structure in all pairwise analyses. Although individual PCA scores were impacted by coverage and temporal distortions (Figure 5), those unwanted effects did not generate substantial bias for PCA eigenvalues, leaving us with F_{ST} estimates that were similar with or without adjustment.

Genetic differentiation explained by environmental factors. To provide a second illustration of the use of spectral estimates of inbreeding coefficients, we studied the role of environmental factors in shaping population genetic structure in plants [29]. For 241 swedish accessions of *Arabidopsis thaliana* taken from The 1,001 Genomes database [30], population structure was first evaluated by using a

spatially explicit Bayesian algorithm. The individuals were clustered in two groups located in southern and northern Sweden (Figure 6A). For the groups estimated by spatial analysis, the mean value of F_{ST} across loci was equal to 7.8%. This value was larger than the largest eigenvalue of the within-population matrix, equal to 4.9 %. The proportion of variance explained by the first PCA axis was equal to 8.5%, greater than F_{ST} (Figure 6). An explanation for this result is that a two-population model did not fit the data accurately, and PCA axes capture spatial genetic variation unseen by a discrete population model. After correction for environmental variation, the leading eigenvalue of the PCA was equal to 6.5% (Figure 6C). The eigenvalue of the between-population matrix – which defines F_{ST} for surrogate genotypes – was equal to 5.2%. The second and subsequent eigenvalues of PCA were equal to 4.9%, 3.2%, 2.3%, and were unaffected by environmental variables (Figure 6B). These values agreed with the eigenvalues of the residual matrix, equal to 5.1%, 3.3%, 2.6% (Figure 6B). The results provided evidence that environmental factors had an impact on the differentiation between northern and southern populations, but had less influence on other axes of genetic variation. For the first axis, the relative proportion of variance explained by environment was important, around 33%, suggesting that environmental conditions played a major role in driving south-north population divergence in Scandinavian *A. thaliana*.

Conclusions. Assuming a model with K discrete populations, our study established a relationship between Wright's inbreeding coefficient, F_{ST}, and the $(K-1)$ leading eigenvalues of scaled PCA. A similar relationship was established between Nei's among-population diversity, D_{ST}, and the leading eigenvalues of unscaled PCA.

Those relationships justify the use of PCA to describe population genetic structure from large genotype matrices. They extend results obtained from coalescent theory for two divergent populations in Ref. [16] to any discrete population model. By introducing a separation condition, they increase the accuracy of previous results, clarifying for which sample sizes and number of loci they could be valid. The separation condition compares the smallest eigenvalue of the between-population matrix to the leading eigenvalue of the residual matrix, and can be checked numerically with a computing cost similar to PCA. Simulations of discrete population models showed that the separation condition could be violated when the sample size or the number of loci is not large enough. In those simulations, we found that leading eigenvalue of the residual matrix was well predicted by RMT. RMT also provided a threshold value of F_{ST}, $\theta = (1/\sqrt{L} + 1/\sqrt{n-1})^2$ below which there is no evidence of population structure for two or more populations. The threshold differs from $\theta = 1/\sqrt{nL}$ proposed in Ref. [10], and it was better supported by simulations of single population models. In addition to connecting PCA of genotype matrix to inbreeding coefficients and related quantities, our results have several implications for the analysis of adjusted genotypes, providing statistics analogous to F_{ST} for those data. Adjusted genotypes arise in many applications, such as ancient DNA, to correct for biases due to technical or sampling artifacts, or ecological genomics where it allows evaluating the part of population differentiation explained by environmental variation. The proposed estimates of inbreeding coefficients are thus of great importance to the understanding of the demographic history of populations and their adaptation to environmental variation.

Methods

Population subdivision. Genomic samples for n unrelated individuals sampled from K discrete populations genotyped at a particular locus are considered. We use the term locus as a shorthand for single-nucleotide polymorphism (SNP), although most of our analyses could include non-polymorphic sites. At this locus, a reference allele and a derived allele are observed. The frequency of the derived allele in population k is equal to p_k. The derived allele frequency in the total sample is equal to $P = \sum_{k=1}^{K} c_k p_k$, where $c_k = n_k/n$ represents a sample proportion. Our treatment of $F_{\rm ST}$ is similar to the original definitions of Wright [1] and Nei [4, 34] with consideration of unequal population sample sizes. Setting $H_{\rm S} = 2\sum_{k=1}^{K} c_k p_k (1-p_k)$ and $H_{\rm T} = 2P(1-P)$, Wright's inbreeding coefficient is defined as $F_{\rm ST} = D_{\rm ST}/H_{\rm T}$, where $D_{\rm ST} = H_{\rm T} - H_{\rm S}$ [4].

PCA and SVD. For a genotype matrix $\mathbf{X}$ with L loci, centered PCA computes the eigenvalues, $\sigma_i^2(\mathbf{Z})$, of the empirical covariance matrix, $\mathbf{Z}\mathbf{Z}^T/n$, where $\mathbf{Z} = \mathbf{Z}^{\rm c}$ is the centered matrix, for which the mean value of each column has been substracted from $\mathbf{X}$ [9, 26]. Scaled PCA computes the eigenvalues, $\rho_i^2(\mathbf{Z})$, of the empirical correlation matrix, $\mathbf{Z}\mathbf{Z}^T/n$, obtained for $\mathbf{Z} = \mathbf{Z}^{\rm sc}$, the matrix for which each column of $\mathbf{X}$ is divided by the square-root of $P(1-P)$ [10]. To make the notation less cluttered, superscripts will be omitted in $\mathbf{Z}^{\rm c}$ and $\mathbf{Z}^{\rm sc}$. In order to obtain unbiased estimates, empirical covariance and correlation matrices are usually divided by $(n-1)$ instead of n. To avoid this complication, we kept n in all theoretical analyses (assuming that n is large), but unbiased estimates were used in all data analyses. Using the equivalence between PCA and SVD, the eigenvalues of PCA were computed as the

squared non-null singular values of the matrix $\mathbf{Z}/\sqrt{n}$.

Spectral analysis. To make arguments easier to follow, we developed the analysis of eigenvalues for centered PCA. Extension to scaled PCA does not create mathematical complications but has heavier notations. This paragraph sketches the key arguments for the main result. More details are provided in *Supplementary Information*. We found that the Hilbert-Schmidt norm of the between-population matrix $\mathbf{Z}_{\rm ST}$ is equal to

$$\|\mathbf{Z}_{\rm ST}\|^2 = nL\,\mathbb{E}\left[\sum_{k=1}^{K} c_k \left(\sum_{j=1}^{K} c_j(p_j - p_k)\right)^2\right] = nL \times \mathbb{E}[D_{\rm ST}]/2\,,$$

where the mathematical symbol $\mathbb{E}[Q]$ denotes the mean value of a quantity Q over the L loci. For scaled PCA, the squared norm is equal to $nL \times \mathbb{E}[F_{\rm ST}]$. The matrices $\mathbf{Z}_{\rm ST}$ and $\mathbf{Z}_{\rm S}$ satisfy orthogonality conditions. When those matrices satisfy the separation condition (4), the sum of the $(K-1)$ leading eigenvalues (variances) of $\mathbf{Z}$ is close to $\|\mathbf{Z}_{\rm ST}\|^2$, which represents the sum of the $(K-1)$ leading eigenvalues of $\mathbf{Z}_{\rm ST}$.

F-models. F-models are models for K discrete populations diverging from an ancestral gene pool [33]. In the ancestral gene pool, the derived allele is present with frequency $p_{\rm anc}$. The K populations diverged from each other and from the ancestral population with drift coefficient equal to F_k relative to the ancestral pool. Conditional on $p_{\rm anc}$, the allele frequency at a particular locus in population k follows a beta distribution of shape parameters $p_{\rm anc}(1-F_k)/F_k$ and $(1-p_{\rm anc})(1-F_k)/F_k$. To create a distribution over the L loci, $p_{\rm anc}$ is drawn from a beta distribution with shape parameters a and b, leading to $\mathbb{E}[p_{\rm anc}] = a/(a+b)$. The expected ancestral

heterozygozity, H_{A}, is equal to $\mathbb{E}[H_{\mathrm{A}}] = 2ab/(a+b)(a+b+1)$. For F-models, the expected value of D_{ST} can be formulated as $\mathbb{E}[D_{\mathrm{ST}}] = \mathbb{E}[H_{\mathrm{A}}] \sum_{k=1}^{K} c_k(1-c_k)F_k$ (*Supplementary Information*). Numerical values for $\mathbb{E}[F_{\mathrm{ST}}]$ are less explicit, but they can be obtained by using Monte-Carlo simulations.

Simulations of F-models were performed in the `R` programming language. We performed simulations of single population models ($K = 1$) to check whether approximations derived from RMT appropriately describe the leading eigenvalue of scaled PCA in the absence of population structure. Simulations of F-models were performed with a value of the drift coefficient equal to $F = 15\%$. The ancestral frequency for the derived allele, p_{anc}, was drawn from a beta distribution with shape parameters $a = 1$ and $b = 9$, so that $\mathbb{E}[p_{\mathrm{anc}}] = 10\%$ (Supplementary Figure S1). Simulations of F-models were performed with $K = 2$ to check whether the data could fit theoretical expectations for D_{ST} and F_{ST}. Two hundred simulations of F-models were performed with equal values of the drift coefficients randomly drawn between 1% and 75% ($F_1 = F_2$). The ancestral frequency for the derived allele, p_{anc}, was drawn from a beta distribution with shape parameters $a = 1$ and $b = 4$, so that $\mathbb{E}[p_{\mathrm{anc}}] = 20\%$. The total sample size was equal to $n = 100$ and the sample proportion c_1 was drawn from a uniform distribution between 10% and 50%. We also considered three-population F-models with equal sample sizes and ancestral allele frequencies distributed according to the uniform distribution, ($a = 1$ and $b = 1$). With the uniform distribution, we found that $\mathbb{E}[H_{\mathrm{A}}] = 1/3$, and the non-null eigenvalues of the between-population covariance matrix could be computed as $\lambda_i = \left(F_1 + F_2 + F_3 \pm \sqrt{F_1^2 + F_2^2 + F_3^2 - F_1F_2 - F_2F_3 - F_3F_1}\right)/54$, for $i = 1, 2$ (*Supplementary Information*). We had $\mathbb{E}[D_{\mathrm{ST}}] = 2(\lambda_1 + \lambda_2)$. Two hundred simula-

tions of three-population models were performed with unequal drift coefficients drawn between 1% and 25%. The total sample size was equal to $n = 100$ and the number of loci was equal to $L = 20,000$. For values of n between 30 and 300, and number of loci between 100 and 1000, we performed additional simulations with small drift coefficients ($F_k \leq 10\%$) to evaluate the probability that the data verify the separation condition.

RMT. For F-models, the probability distribution of eigenvalues were approximated with asymptotic quantities obtained from random matrix theory, considering large sample sizes, and keeping the ratio of the number of loci to the sample size, L/n, to a constant value [40, 19, 43, 41, 26]. For a single population model, the proportions of variance explained by each principal axis were approximated by the Marchenko-Pastur probability density function described by

$$p(x) = L\frac{\sqrt{(x_M - x)(x - x_m)}}{2x\pi}, \quad x_m = (1 - \sqrt{\gamma})^2/L \leq x \leq x_M = (1 + \sqrt{\gamma})^2/L\,.$$

With $K = 1$, the proportion of variance explained by the first principal axis was approximated by $(1/\sqrt{L} + 1/\sqrt{n-1})^2$. For $K > 1$, the Marchenko-Pastur density modelled the bulk distribution of eigenvalues for the within-population (residual) matrix. Under the separation condition (3), the proportion of variance explained by the Kth principal axis was approximated by $(1 - \mathbb{E}[F_{\mathrm{ST}}])(1/\sqrt{L} + 1/\sqrt{n-K})^2$. Regarding centered PCA, the largest singular value of the within-population matrix was approximated by $\sigma_1^2(\mathbf{Z}_{\mathrm{S}})/L \approx \mathbb{E}[H_{\mathrm{S}}]\left(1 + \sqrt{L/(n-K)}\right)^2/2$.

Ancient DNA analyses. We analyzed 143,081 pseudo-haploid SNP genotypes from ancient samples of early farmers from Anatolia ($n = 23$), steppe pastoralists from the Yamnaya culture ($n = 15$), and Western hunter-gatherers from Serbia ($n = 31$). The data were extracted from a public data set available from David Reich lab's repository (`reich.hms.harvard.edu`) [47, 17, 48, 49]. The ancient samples had a minimum coverage of 0.25x, a median coverage of 2.69x (mean of 2.98x) and a maximum coverage of 13.54x. Genotypes were adjusted for coverage by fitting a latent factor regression model with the number of factors equal to the number of sample minus two [50]. The matrix was then adjusted for distortions due to differences in sample ages [28], resulting in surrogate genotypes encoded as continuous variables not interpretable in terms of allelic frequency.

Genomic and bioclimatic data analyses. We studied 241 swedish plant accessions from The 1,001 Genomes database for *Arabidopsis thaliana* [30]. A matrix of SNP genotypes was obtained by considering variants with minor allele frequency greater than 5% and a density of variants around one SNP every 1,000 bp (167,475 SNPs). The individuals were clustered in groups based on analysis of population structure accounting for geographic proximity [51]. Global climate and weather data corresponding to individual geographic coordinates were downloaded from the WorldClim database (`https://worldclim.org`). Eighteen bioclimatic variables, derived from the monthly temperature and rainfall values, were considered as representing the current environmental matrix. Correction of genotypes for locus-specific effects of the eighteen environmental variables was performed with a latent factor regression model implemented in the R package `lfmm` [50]. For the matrix of centered genotypes,

$\mathbf{Z}$, and the matrix of bioclimatic variables, $\mathbf{Y}$, the program estimated a matrix of surrogate genotypes, $\mathbf{W}$, by adjusting a regression model of the form $\mathbf{Z} = \mathbf{Y}\mathbf{B}^T + \mathbf{W} + \epsilon$. The $\mathbf{B}$ matrix contains effect sizes for each bioclimatic variable in the matrix $\mathbf{Y}$, and ϵ represents centered Gaussian errors. To keep the latent matrix estimate ($\mathbf{W}$) as close as possible to $\mathbf{Z}$, we used $k = n - 19 = 222$ factors to compute $\mathbf{W}$.

Data availability. The data used in our study were publicly available from their cited reference.

References

[1] S. Wright. The interpretation of population structure by F-statistics with special regard to systems of mating. *Evolution* **19**, 395-420 (1965).

[2] G. Malécot. *Les Mathématiques de Hérédité* (Paris: Masson, 1948).

[3] C. C. Cockerham. Variance of gene frequencies. *Evolution* **23**, 72-84 (1969).

[4] M. Nei. Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences* **70**, 3321-3323 (1973).

[5] B. S. Weir, C. C. Cockerham. Estimating F-statistics for the analysis of population structure. *Evolution* **38** 1358-1370 (1984).

[6] M. Slatkin. Inbreeding coefficients and coalescence times. *Genetics Research* **58**, 167-175 (1991).

[7] K. E. Holsinger, B. S. Weir. Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. *Nature Reviews Genetics* **10**, 639-650 (2009).

[8] H. Hotelling, Relations between two sets of variates. *Biometrika* **28**, 321-377 (1936).

[9] I. Jolliffe, *Principal Component Analysis* (Springer, 1986).

[10] N. Patterson, A. L. Price, D. Reich. Population structure and eigenanalysis. *PLoS Genetics* **2**, e0020190 (2006).

[11] J. K. Pritchard, M. Stephens, P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959 (2000).

[12] D. Falush, M. Stephens, J. K. Pritchard. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567-1587 (2003).

[13] I. T. Jolliffe, J. Cadima. Principal component analysis: a review and recent developments.*Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Science* **374**, 20150202 (2016).

[14] L. L. Cavalli-Sforza, A. W. F. Edwards, S. Geerts. Analysis of human evolution. *Genetics today: Proceedings of the 11th International Congress of Genetics, The Hague, The Netherlands. New York: Pergamon*, **3** 923-993 (1963).

[15] P. Menozzi, A. Piazza, L. L. Cavalli-Sforza. Synthetic maps of human gene frequencies in Europeans. *Science* **201**, 786-792 (1978).

[16] G. McVean. A genealogical interpretation of principal components analysis. *PLoS Genetics* **5**, e1000686 (2009).

[17] W. Haak, I. Lazaridis, N. Patterson, N. Rohland, S. Mallick, B. Llamas *et al.* Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **522**, 207 (2015).

[18] X. Zheng, B. S. Weir. Eigenanalysis of SNP data with an identity by descent interpretation. *Theoretical Population Biology* **107**, 65-76 (2016).

[19] K. Bryc, W. Bryc, J. W. Silverstein. Separation of the largest eigenvalues in eigenanalysis of genotype data from discrete subpopulations. *Theoretical Population Biology* **89**, 34-43 (2013).

[20] N. Duforet-Frebourg *et al.* Detecting genomic signatures of natural selection with principal component analysis: application to the 1000 genomes data. *Molecular Biology and Evolution* **33**, 1082-1093 (2016).

[21] G. B. Chen, S. H. Lee, Z. X. Zhu, B. Benyamin, M. R. Robinson. EigenGWAS: finding loci under selection through genome-wide association studies of eigenvectors in structured populations. *Heredity* **117**, 51-61 (2016).

[22] K. J. Galinsky *et al.* Fast principal-component analysis reveals convergent evolution of *ADH1B* in Europe and East Asia. *The American Journal of Human Genetics* **98**, 456-472 (2016).

[23] O. François, H. Martins, K. Caye, S. D. Schoville. Controlling false discoveries in genome scans for selection. *Molecular Ecology* **25**, 454-469 (2016).

[24] H. M. Wilkinson-Herbots. Genealogy and subpopulation differentiation under various models of population structure. *Journal of Mathematical Biology* **37**, 535-585 (1998).

[25] J. Ma, C. I. Amos. Theoretical formulation of principal components analysis to detect and correct for population stratification. *PLoS ONE* **5**, e12510 (2010).

[26] I. M. Johnstone, D. Paul. PCA in high dimensions: An orientation. *Proceedings of the IEEE* **106**, 1277-1292 (2018).

[27] T. S. Korneliussen, A. Albrechtsen, R. Nielsen. ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics* **15**, 356 (2014).

[28] O. François, F. Jay. Factor analysis of ancient population genomic samples. *Nature Communications* **11**, 4661 (2020).

[29] I. J. Wang, R. E. Glor, J. B. Losos. Quantifying the roles of ecology and geography in spatial genetic divergence. *Ecology Letters* **16**, 175-182 (2013).

[30] C. Alonso-Blanco *et al.* 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell* **166**, 481-491 (2016).

[31] Z. Li, A. Löytynoja, A. Fraimout, J. Merilä. Effects of marker type and filtering criteria on $Q_{\mathrm{ST}} - F_{\mathrm{ST}}$ comparisons. *Royal Society Open Science* **6**, 190666 (2019).

[32] S. Wright. Evolution in Mendelian populations. *Genetics* **16**, 97-159 (1931).

[33] D. J. Balding, R. A. Nichols A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica* **96**, 3-12 (1995).

[34] M. Nei, R. K. Chesser. Estimation of fixation indices and gene diversities. *Annals of Human Genetics* **47**, 253-259 (1983).

[35] E. Linck, C. J. Battey. Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. *Molecular Ecology Resources* **19**, 639-647 (2019).

[36] T. M. Culley, L. E. Wallace, K. M. Gengler-Nowak, D. J. Crawford. A comparison of two methods of calculating G_{ST}, a genetic measure of population differentiation. *American Journal of Botany* **89**, 460-465 (2002).

[37] G. Bhatia, N. Patterson, S. Sankararaman, A. L. Price. Estimating and interpreting F_{ST}: the impact of rare variants. *Genome Research* **23**, 1514-1521 (2013).

[38] N. Patterson *et al.* Ancient admixture in human history. *Genetics* **192**, 1065-1093 (2012).

[39] B. M. Peter. Admixture, population structure, and F-statistics. *Genetics* **202**, 1485-1501 (2016).

[40] V. A. Marčenko , L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik* **1**, 457 (1967).

[41] I. M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Annals of Statistics* **29**, 295-327 (2001).

[42] I.M. Johnstone. Multivariate analysis and Jacobi ensembles: largest eigenvalue, Tracy-Widom limits and rates of convergence. *Annals of Statistics* **36**, 2638-2716 (2008).

[43] J. Bryson, R. Vershynin, H. Zhao. Marchenko-Pastur law with relaxed independence conditions. *arXiv preprint* (arXiv:1912.12724, 2019).

[44] F. L. Bookstein. Pathologies of between-groups principal components analysis in geometric morphometrics. *Evolutionary Biology* **46**, 271-302 (2019).

[45] The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526** 68-74 (2015).

[46] A. Cardini, P. O'Higgins, F. J. Rohlf. Seeing distinct groups where there are none: spurious patterns from between-group PCA. *Evolutionary Biology* **46**, 303-316 (2019).

[47] M.E. Allentoft, M. Sikora, K.G. Sjögren, S. Rasmussen, M. Rasmussen, J. Stenderup *et al.* Population genomics of Bronze Age Eurasia. *Nature* **522**, 167-72 (2015).

[48] I. Mathieson, I. Lazaridis, N. Rohland, S. Mallick, N. Patterson, S. A. Roodenberg *et al.* Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* **528**, 499 (2015).

[49] I. Mathieson, S. Alpaslan-Roodenberg, C. Posth, A. Szécsényi-Nagy, N. Rohland, S. Mallick,*et al.* The genomic history of southeastern Europe. *Nature* **555**, 197 (2018).

[50] K. Caye, B. Jumentier, J. Lepeule, O. François. LFMM 2: fast and accurate inference of gene-environment associations in genome-wide studies. *Molecular Biology and Evolution* **36**, 852-860 (2019).

[51] K. Caye, F. Jay, O. Michel, O. François. Fast inference of individual admixture coefficients using geographic data. *The Annals of Applied Statistics* **12**, 586-608 (2018)

Tables and Figures

Box 1. Notations

n	Sample size
L	Number of genomic loci
F_{ST}	Wright's fixation index, computed from Nei's formula with correction for unequal sample sizes
H_{S}	Within population genetic diversity
H_{T}	Genetic diversity in the total population
D_{ST}	Among (or between) population genetic diversity
$\mathbf{X}$	Matrix of SNP genotypes for n individuals at L loci
$\mathbf{Z}$	Matrix of centered genotypes, $\mathbf{X}-\mathbf{P}$, or scaled genotypes, $(\mathbf{X}-\mathbf{P})/\sqrt{\mathbf{P}(1-\mathbf{P})}$
$\mathbf{Z}_{\mathrm{ST}}$	An $n \times L$ matrix describing between-population data repeated for individuals from a same population
$\mathbf{Z}_{\mathrm{S}}$	An $n \times L$ matrix describing within-population data
$\sigma_k^2(\mathbf{Z})/n$	Eigenvalues of the empirical covariance matrix (centered PCA)
$\rho_k^2(\mathbf{Z})/n$	Eigenvalues of the empirical correlation matrix (scaled PCA), also equal to L times the proportions of variance explained by the principal axes

Table 1. F_{ST} estimates for populations from The 1,000 Genomes Project

	Lead. eigen. of PCA*	F_{ST} **across loci**	**Lead. eigen. res. matrix****	**RMT approximation*****
CHB-CEU	5.65%	5.65%	0.42%	0.48%
CHB-YRI	8.35%	8.35%	0.36%	0.37%
CHB-IBS	5.42%	5.42 %	0.37%	0.40%
CEU-YRI	7.21%	7.21 %	0.35%	0.37%
CEU-IBS	0.41%	0.38 %	0.37%	0.41%
YRI-IBS	7.27%	7.27 %	0.31%	0.32%

* Leading eigenvalue of the PCA
** Leading eigenvalue of the within-population matrix
*** RMT approximation for the leading eigenvalue of the within-population matrix

IBS: Iberian ($n = 147$), **CHB**: Han Chinese in Beijing ($n = 100$), **YRI**: Yoruba ($n = 158$), **CEU**: Utah residents with European ancestry ($n = 104$). Number of SNPs $L \approx 1.3$M with minor allele frequency equal to 5%.

Table 2. F_{ST} **estimates for ancient Eurasian samples with correction for genomic coverage.**

	F_{ST} **without correction**	F_{ST} **with correction**	**Lead. eigen. res. matrix***	**RMT threshold****
EFA-Steppe	4.8%	4.7%	3.1%	2.8%
EFA-WHG	5.9%	5.8%	3.3%	2.0%
Steppe-WHG	5.2%	5.1 %	3.8%	2.3%

EFA: Early Farmers from Anatolia, **WHG**: Western Hunter-Gatherers

* Leading eigenvalue of the within-population residual matrix

** RMT threshold for evidence of population structure, $\theta = (1/\sqrt{L} + 1/\sqrt{n-1})^2$, L: number of loci, n: sample size

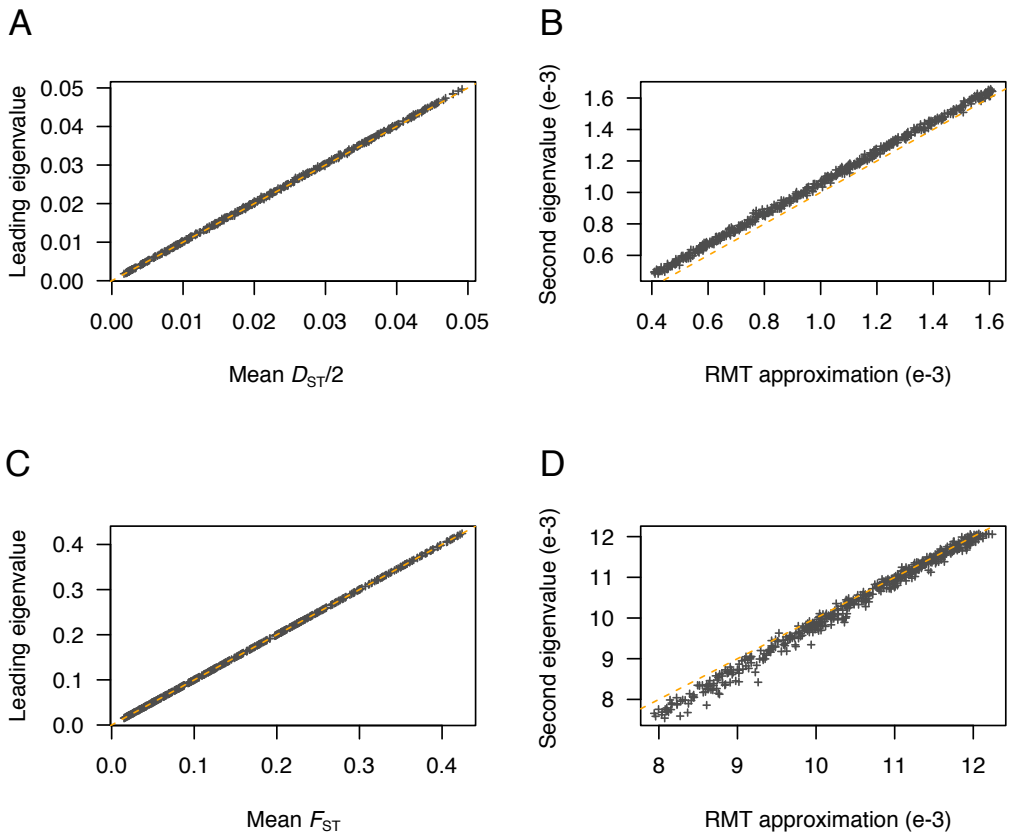

Figure 1. Comparaison of D_{ST} and F_{ST} estimates with the leading PCA eigenvalues in two-population models. (**A**) Leading eigenvalues of centered PCA as a function of the mean of $D_{\mathrm{ST}}/2$ across loci. (**B**) Second eigenvalue of centered PCA as a function of its approximation from RMT. (**C**) Leading eigenvalues of scaled PCA as a function of the mean of F_{ST} across loci. (**D**) Second eigenvalue of scaled PCA as a function of its approximation from RMT, which is given by $(1-\rho_1^2)\times(1/\sqrt{L}+1/\sqrt{n-2})^2$ (approximation of the largest eigenvalue of the residual matrix). The dashed lines correspond to the diagonal $y=x$. Simulations of F-models were performed for $n=100$ individuals (inbreeding coefficients between 1% and 75%, first population sample proportion between 10% and 50%, ancestral frequency was drawn from a beta(1,4) distribution).

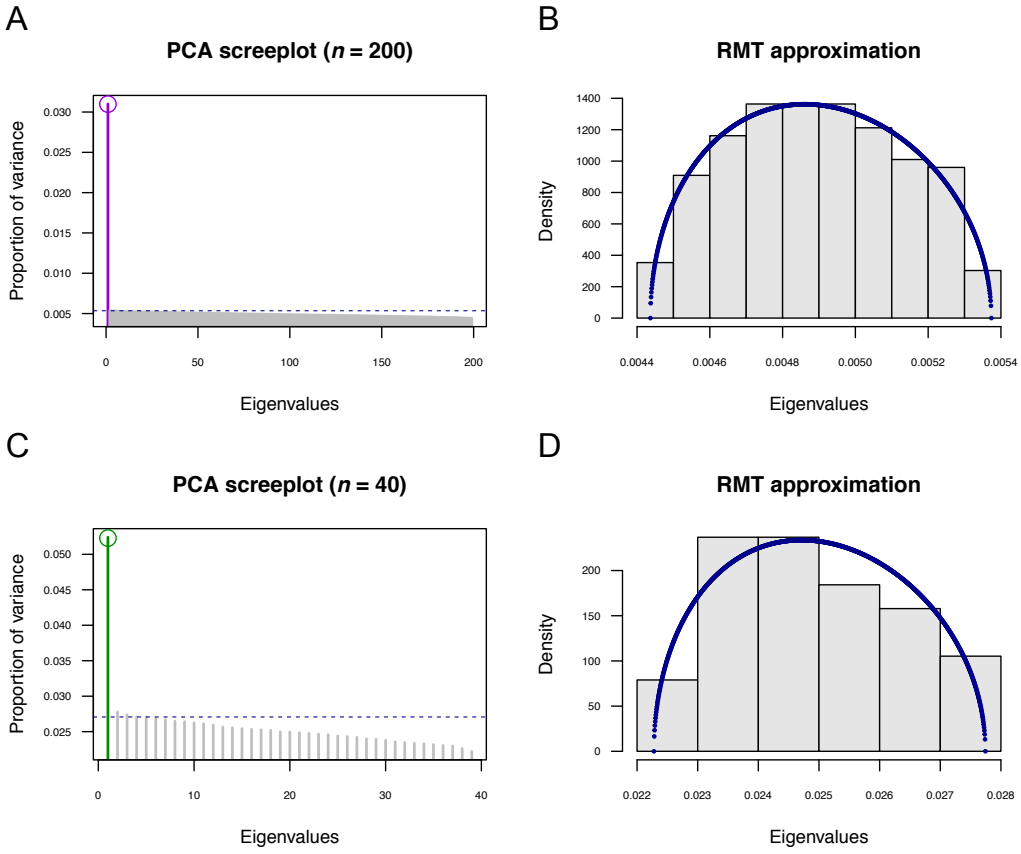

Figure 2. Screeplots and RMT approximations in two-population models. (**A**) Proportion of variance (eigenvalues) explained by PC axes, with a circle symbol representing the mean of F_{ST} across loci for $n = 200$ individuals and $L = 85,540$ SNPs. (**B**) Histogram of eigenvalues of the residual matrix, $\mathbf{Z}_{\mathrm{S}}/\sqrt{n-2}$, for the data simulated in A. (**C**) Proportion of variance for $n = 40$ individuals and $L = 12,650$ SNPs. (**D**) Histogram of eigenvalues of the residual matrix for the data simulated in C. The dashed lines in PCA scree-plots represent the RMT approximation of the leading eigenvalue of the residual matrix. The blue curve represents the Marchenko-Pastur probability density. Simulations of F-models were performed with p_{anc} drawn from a beta(1,9) distribution and $F_1 = F_2 = 7\%$.

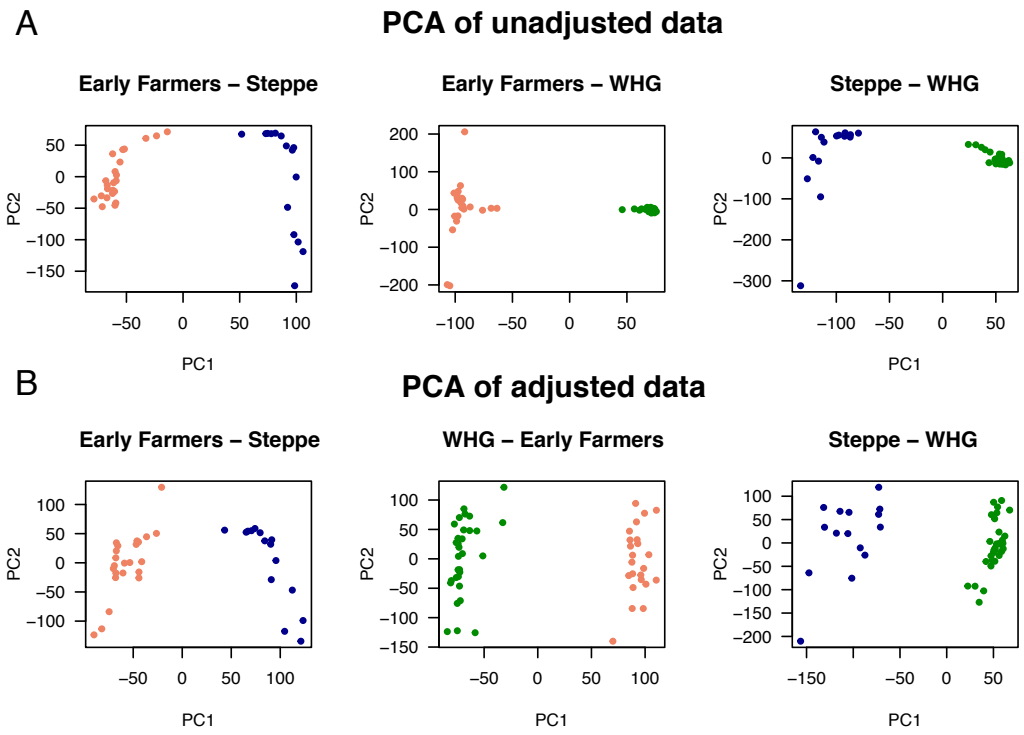

Figure 3. Correction for coverage in PC plots for pairs of ancient population samples. **(A)** PCA of unadjusted genotypes. **(B)** PCA of non-binary genotypic data adjusted for coverage. Population samples: Early Farmers (salmon color, $n_1 = 23$), Steppe pastoralists (blue color, $n_2 = 15$), (Western hunter gatherers, green color, $n_3 = 31$)

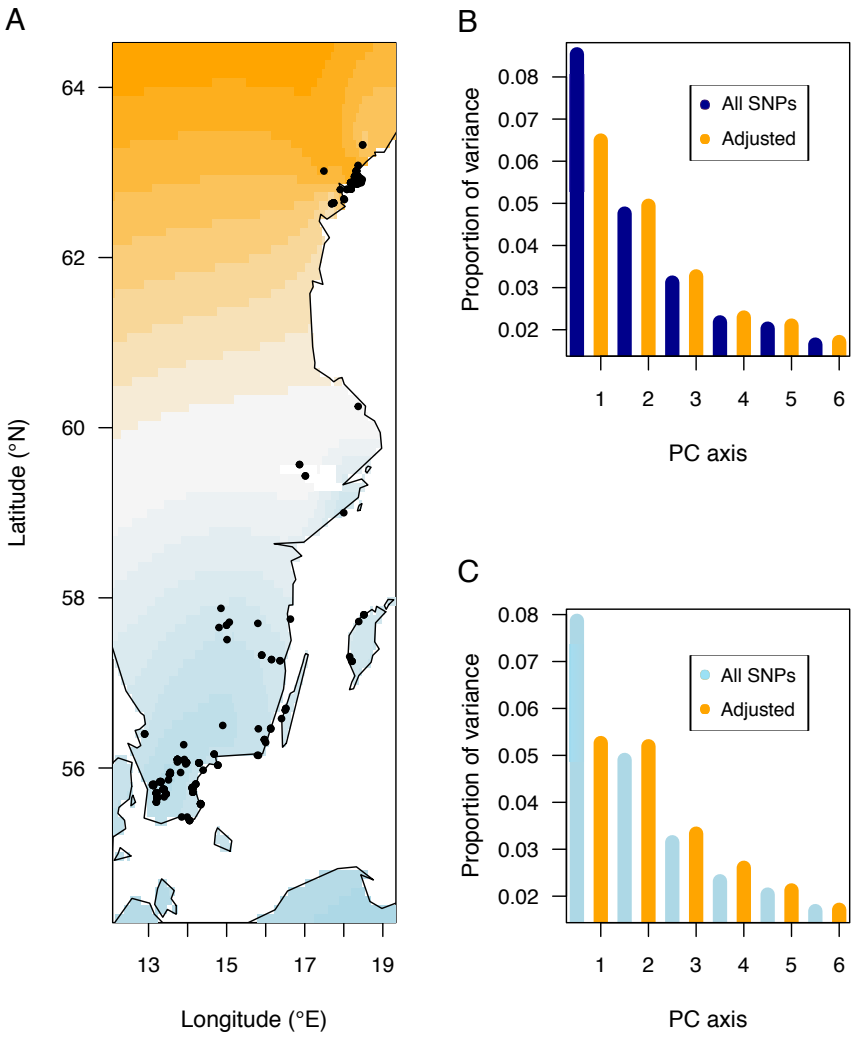

Figure 4. Neutral F_{ST} for *Arabidopsis thaliana* in Scandinavia. (A) Geographic locations of 241 samples and inference of population structure from a spatial method (Blue color: Southern cluster, Orange color: Northern cluster). **(B)** Proportion of variance explained by PC axes before adjustment of genotypes for environmental variables (blue color) and after adjustment (orange color). **(C)** Proportion of variance explained by the first axis of the between-population matrix, and by the first axes of the residual matrix (five components) before adjustment for environmental variables (blue color) and after adjustment (orange color). Wright's coefficients are represented by the values for the first axis.