FSTruct: an F_{ST} -based tool for measuring ancestry variation in inference of population structure

Maike L. Morrison^{*†}, Nicolas Alcala[‡], & Noah A. Rosenberg^{*}

September 1, 2021

Abstract

In model-based inference of population structure from individual-level genetic data, individuals are assigned membership coefficients in a series of statistical clusters generated by clustering algorithms. Distinct patterns of variability in membership coefficients can be produced for different groups of individuals, for example, representing different predefined populations, sampling sites, or time periods. Such variability can be difficult to capture in a single numerical value; membership coefficient vectors are multivariate and potentially incommensurable across groups, as the number of clusters over which individuals are distributed can vary among groups of interest. Further, two groups might share few clusters in common, so that membership coefficient vectors are concentrated on different clusters. We introduce a method for measuring the variability of membership coefficients of individuals in a predefined group, making use of an analogy between variability across individuals in membership coefficient vectors and variation across populations in allele frequency vectors. We show that in a model in which membership coefficient vectors in a population follow a Dirichlet distribution, the measure increases linearly with a parameter describing the variance of a specified component of the membership vector. We apply the approach, which makes use of a normalized F_{ST} statistic, to data on inferred population structure in three example scenarios. We also introduce a bootstrap test for equivalence of two or more groups in their level of membership coefficient variability. Our methods are implemented in the R package FSTruct.

Keywords: Admixture, F_{ST} , population structure

Running title: Measuring ancestry variation with FSTruct

Introduction

In the past two decades, computational methods for inference of population structure from individual-level genetic data have contributed a rich and informative set of approaches for the analysis of genetic variation. Model-based clustering methods such as ADMIXTURE (ALEXANDER *et al.*, 2009; ALEXANDER and LANGE, 2011), BAPS (CORANDER *et al.*, 2004, 2008) and STRUCTURE (FALUSH

^{*}Department of Biology, Stanford University, Stanford, CA 94305 USA

[†]Corresponding author. Email: maikem@stanford.edu

[‡]International Agency for Research on Cancer (IARC/WHO), Genomic Epidemiology Branch, Lyon, France

Measuring ancestry variation with FSTruct

et al., 2003, 2007; HUBISZ *et al.*, 2009; PRITCHARD *et al.*, 2000) are now routinely used to generate insights into population structure and evolutionary history in diverse species of interest in ecology, evolution, conservation biology, and agriculture (GUILLOT and ORLANDO, 2017).

In model-based inference of population structure, individuals are clustered based on their multi-locus genotypes into a series of statistical clusters, such that each individual possesses a membership coefficient for each cluster. Each membership coefficient represents the proportion of an individual's ancestry that is derived from the associated cluster. Interpreting the membership coefficients of individuals from various predefined populations, sampling sites, or other groups of biological interest can illuminate patterns of genetic variation and population structure. Researchers often investigate variability of membership patterns within predefined groups, as well as similarities and differences in the membership patterns of distinct groups.

One type of comparison that is frequently of interest is an assessment of relative levels of variation in membership coefficients among the individuals belonging to two or more predefined groups. This type of comparison arises in many contexts, such as when exploring differences in membership variability between admixed and non-admixed populations, between populations from different time periods, or between different types of data on the same sampled individuals.

For example, in a study of ancient human DNA samples dating over a period of hundreds to thousands of years, ANTONIO *et al.* (2019) sought to examine if the population of Rome possessed greater diversity in ancestry during certain periods of the Roman Empire. They estimated membership coefficients using ADMIXTURE and interpreted the inferred coefficients to claim that during the Imperial Rome period, when the Roman Empire was at its peak, ancestry was more variable than during earlier periods, when Rome was more isolated (ANTONIO *et al.*, 2019, Fig. 1).

Interpretations of inferred membership coefficients to make relative claims about membership variability have generally relied on visual assessment of population structure diagrams rather than on statistical hypothesis testing. In particular, as in ANTONIO *et al.* (2019), researchers seeking to quantify variability in membership coefficients across individuals or to compare this variability between two or more groups often do so visually or informally.

Here, we introduce a statistical method to measure variability in membership coefficients inferred

by model-based clustering, and to compare this variability across populations. We apply the method to examples from real and simulated data. The method is implemented in the *R* package FSTruct.

Materials & Methods

Overview

The output of population structure inference software programs such as STRUCTURE and ADMIXTURE is a representation of individual membership coefficients in matrix form. The matrix, often denoted Q and termed a “ Q matrix,” has I rows, corresponding to I individuals, and K columns, corresponding to the total number of clusters (Figure 1B). The entry in row i and column k , $q_k^{(i)}$, represents the membership coefficient of individual i in cluster k : the proportion of the ancestry of individual i that is assigned to cluster k . Each row sums to 1, or $\sum_{k=1}^K q_k^{(i)} = 1$ for each i .

We seek to compute a measure of variability among ancestry vectors for individuals, that is, among rows of Q . We wish for the measure to be comparable across different data sets, possibly representing different samples. This problem is complicated by the fact that different Q matrices might include different numbers of clusters; furthermore, column entries for some clusters might vary greatly across individuals while other columns are more uniform.

We approach the problem by modifying the population differentiation statistic F_{ST} to fit this ancestry scenario. F_{ST} measures allele frequency variability among subpopulations, and it is computed using a set of allele frequency vectors that each sum to 1. This setting is mathematically analogous to Q matrices, in which vectors of membership coefficients for each individual sum to 1. In the analogy, each individual represents a “population,” and its cluster membership is analogous to an “allele frequency” (Figure 1).

By computing F_{ST} among individual vectors of membership coefficients, we can measure the variability of a single Q matrix. To facilitate comparisons of Q matrices with different numbers of individuals or clusters, we use a normalization of F_{ST} . Despite the general understanding that F_{ST} can in principle reach 1, features of a data set constrain the maximal value of F_{ST} , so that the maximum is often less than 1 (JAKOBSSON *et al.*, 2013; ALCALA and ROSENBERG, 2017, 2019). The

Measuring ancestry variation with FSTruct

constrained maximum is relatively low when I , the number of individuals in a Q matrix, is small (analogous to a small number of populations), or when M , the mean membership of the highest-membership ancestry cluster, is close to 0 or 1 (analogous to an extreme value for the frequency of the most frequent allele). Denoting this maximum $F_{ST}^{\max}$, we normalize F_{ST} by its maximum, using the ratio $F_{ST}/F_{ST}^{\max}$ as a measure of variability that is comparable across Q matrices of different size. This measure ranges between 0 and 1, equaling 0 when members of a population have identical ancestry and equaling 1 when vectors of membership coefficients are maximally variable.

The $F_{ST}/F_{ST}^{\max}$ formula

Consider a scenario with I subpopulations and K distinct alleles. Allele k has frequency $q_k^{(i)}$ in subpopulation i , with $0 \leq q_k^{(i)} \leq 1$ and $\sum_{k=1}^K q_k^{(i)} = 1$.

To calculate F_{ST} among the I subpopulations, we use $F_{ST} = (H_T - H_S)/H_T$, where H_S represents the mean heterozygosity of the subpopulations and H_T represents the heterozygosity of the total population formed by pooling the subpopulations.

The subpopulation heterozygosity H_S is the mean expected frequency of heterozygotes across all I subpopulations, or $H_S = 1 - \frac{1}{I} \sum_{i=1}^I \sum_{k=1}^K (q_k^{(i)})^2$. The total heterozygosity H_T is the expected frequency of heterozygotes under Hardy-Weinberg equilibrium in a population whose allele frequencies equal the mean allele frequencies across subpopulations: $H_T = 1 - \sum_{k=1}^K (\frac{1}{I} \sum_{i=1}^I q_k^{(i)})^2$. The quantity $\frac{1}{I} \sum_{i=1}^I q_k^{(i)}$ gives the mean frequency of allele k across subpopulations.

With the total population assumed to be polymorphic so that $H_T > 0$, for the setting of I subpopulations and K alleles, with K possibly arbitrarily large, ALCALA and ROSENBERG (2021) obtained the maximal value possible for F_{ST} given a fixed value of $M = \frac{1}{I} \sum_{i=1}^I q_1^{(i)}$, where allele $k = 1$ represents the allele of greatest mean frequency across the I subpopulations. Writing $\sigma_1 = IM$, $J = \lceil \sigma_1^{-1} \rceil$, and $\{\sigma_1\} = \sigma_1 - \lfloor \sigma_1 \rfloor$, we have (ALCALA and ROSENBERG, 2021, eq. 3)

$$F_{ST}^{\max} = \begin{cases} \frac{(I-1)[1-\sigma_1(J-1)(2-J\sigma_1)]}{I-[1-\sigma_1(J-1)(2-J\sigma_1)]}, & 0 < \sigma_1 \leq 1 \\ \frac{I(I-1)-\sigma_1^2+\lfloor \sigma_1 \rfloor-2(I-1)\{\sigma_1\}+(2I-1)\{\sigma_1\}^2}{I(I-1)-\sigma_1^2-\lfloor \sigma_1 \rfloor+2\sigma_1-\{\sigma_1\}^2}, & 1 < \sigma_1 < I. \end{cases} \quad (1)$$

This maximum is plotted as a function of M for five different values of I in Figure 2.

Measuring ancestry variation with FSTruct

In the language of our analogy, I is the number of individuals—the number of rows in the Q matrix; M is the sample mean membership coefficient for the most frequent ancestral cluster across all I individuals; and $\sigma_1 = IM$ is the largest entry in the vector that sums column entries of the Q matrix across rows. The latter case of eq. 1, with $1 < \sigma_1 < I$, is generally more relevant in the setting of population clustering, as I is typically larger than K , so that $\sigma_1 > \frac{I}{K} > 1$.

The ratio $F_{ST}/F_{ST}^{\max}$, which represents a normalized measure of variability that can be compared among different groups of individuals with different values of I or K or both, ranges between 0 and 1, taking a value of 0 when all individuals in a group have identical membership coefficients. It has a value of 1 when they are as variable as possible given M .

ALCALA and ROSENBERG (2021) showed that for $0 < \sigma_1 \leq 1$, the maximum is realized when each ancestry cluster is found in only a single individual and each individual has exactly J ancestry clusters with coefficients greater than zero: $J - 1$ clusters with coefficients of M , one cluster with a coefficient of $1 - (J - 1)\sigma_1 \leq M$, and all others with coefficients of 0.

For $1 < \sigma_1 < I$, the maximum is realized when only the ancestry cluster of greatest membership is shared among individuals, and at most a single individual contains ancestry from multiple sources. More formally, this scenario occurs when $\lfloor \sigma_1 \rfloor$ individuals possess all of their membership in the cluster of greatest membership, a single individual has membership coefficient $\{\sigma_1\}$ for the cluster of greatest membership and coefficient $1 - \{\sigma_1\}$ for one other cluster, and the remaining $I - \lfloor \sigma_1 \rfloor - 1$ individuals each have membership coefficient 1 for mutually distinct ancestry clusters.

Statistical test to compare values of $F_{ST}/F_{ST}^{\max}$

In applications, we may wish not only to compute $F_{ST}/F_{ST}^{\max}$ for a single population, but also to compare this ratio between two or more populations using a statistical test. We accomplish this task by bootstrap resampling of rows to generate replicate Q matrices for each population. We then compute the $F_{ST}/F_{ST}^{\max}$ statistic for each of these replicate matrices. This process generates a bootstrap distribution of the statistic for each population. We then use a Wilcoxon rank-sum test to determine if pairs of bootstrap distributions of the statistic for different sets of individuals are significantly different; we use a Kruskal-Wallis test to compare three or more sets of individuals.

Software availability

We have implemented the method in the R package FSTruct (pronounced “F-struct”), which is available for download from github.com/MaikeMorrison/FSTruct. This package includes functions that compute $F_{ST}/F_{ST}^{\max}$ from a Q matrix such as those produced by ADMIXTURE or STRUCTURE, generate bootstrap replicates and distributions for arbitrarily many Q matrices, and visualize Q matrices.

Results

Simulation examples

Dirichlet model. To illustrate our method, we used individual membership coefficient vectors drawn from a Dirichlet distribution (KOTZ *et al.*, 2000). This distribution is suited for use as the underlying model for finite vectors of nonnegative numbers $(q_1, q_2, \dots, q_K)$ that sum to one, $\sum_{k=1}^K q_k = 1$, and it has appeared in previous studies of membership coefficient vectors (PRITCHARD *et al.*, 2000; HUELSENBECK and ANDOLFATTO, 2007).

We treat individual membership coefficient vectors in a population as following a Dirichlet distribution with parameter vector $\alpha\boldsymbol{\lambda} = \alpha(\lambda_1, \lambda_2, \dots, \lambda_K)$, where $\sum_{k=1}^K \lambda_k = 1$. We denote this distribution by $\text{Dir}(\alpha(\lambda_1, \lambda_2, \dots, \lambda_K))$. Here, $\boldsymbol{\lambda}$ is a vector of length K whose elements determine the parametric mean membership coefficient for each ancestral cluster. The value of α controls the variance of q_k , the individual membership coefficient in cluster k : $\text{Var}[q_k] = \lambda_k(1 - \lambda_k)/(\alpha + 1)$. Thus, an increase in α lowers the variances of membership coefficients.

To generate a random Q matrix with I individuals and K ancestry clusters, we draw I independent and identically distributed $\text{Dir}(\alpha(\lambda_1, \lambda_2, \dots, \lambda_K))$ vectors, $(q_1, \dots, q_K)$, which each comprise a set of membership coefficients for a single individual. Each vector is a row of the simulated Q matrix and is a draw from a Dirichlet distribution with mean membership coefficients $(\lambda_1, \lambda_2, \dots, \lambda_K)$. Variability of membership coefficients across individuals is controlled by α . Hence, we proceed by (1) using the Dirichlet distribution to simulate Q matrices with specified membership coefficient means and variances, (2) computing $F_{ST}/F_{ST}^{\max}$ for each Q matrix, and (3) examining the relation-

Measuring ancestry variation with FSTruct

ship between the value of $F_{ST}/F_{ST}^{\max}$ for each Q matrix and the variance of the Dirichlet distribution used to simulate it.

Dirichlet simulations. To investigate the behavior of $F_{ST}/F_{ST}^{\max}$ in relation to a measure of variability in membership coefficients, we used the Dirichlet distribution to simulate Q matrices with known variability. We simulated Q matrices with $I = 50$ individuals and $K = 2$ clusters. Each simulation replicate thus drew $I = 50$ ancestry vectors from a $\text{Dir}(\alpha(\lambda_1, \lambda_2))$ distribution.

We fixed $(\lambda_1, \lambda_2) = (\frac{2}{3}, \frac{1}{3})$, so that membership in cluster 1 has mean $\frac{2}{3}$ across individuals in a population and membership in cluster 2 has mean $\frac{1}{3}$. The variance of the membership coefficient for a specific cluster, across sampled individuals, then equals $\sigma^2 = \text{Var}[q_1] = \text{Var}[q_2] = (\frac{2}{3} \times \frac{1}{3})/(\alpha + 1)$; both coefficients have the same variance. As α ranges in $(0, \infty)$, the variance ranges in $(0, \frac{2}{9})$.

We performed 500 replicate simulations of samples of 50 individuals for each of 45 values of α , choosing α values to obtain variances 0.001, 0.005, 0.01, 0.015, $\dots$, 0.22, ranging from near the lower bound of 0 on the variance and stopping short of the upper bound of $\frac{2}{9}$.

Next, we compared the value of $F_{ST}/F_{ST}^{\max}$ for each simulated Q matrix to the variance parameter of the Dirichlet distribution used to generate it. As $F_{ST}/F_{ST}^{\max}$ measures variability of Q matrices, we expect to see a positive relationship between the Dirichlet variance used to generate the Q matrix and our estimate of its variability, $F_{ST}/F_{ST}^{\max}$.

Simulation results, depicting the 500 values of $F_{ST}/F_{ST}^{\max}$ for each of the 45 choices of the Dirichlet variance $\sigma^2 = \text{Var}[q_1] = \text{Var}[q_2] = (\frac{2}{3} \times \frac{1}{3})/(\alpha + 1)$, appear in Figure 3. In the figure, the relationship between $F_{ST}/F_{ST}^{\max}$ and σ^2 is strongly linear, with slope 4.5.

Noticing that the empirical slope, 4.5, was the reciprocal of $\frac{2}{9}$, the upper bound of the Dirichlet variance, we sought to obtain a mathematical relationship between $\mathbb{E}[F_{ST}/F_{ST}^{\max}; \alpha, \lambda_1, \lambda_2, I]$, the expectation of $F_{ST}/F_{ST}^{\max}$ under the Dirichlet model, and the parametric variance of each membership coefficient in the model. This calculation, performed in the Appendix, confirms the relationship (eq. A.11)

$$\mathbb{E}\left[\frac{F_{ST}}{F_{ST}^{\max}}; \alpha, \lambda_1, \lambda_2, I\right] \approx \frac{1}{\alpha + 1} = \frac{1}{\lambda_1 \lambda_2} \frac{\lambda_1 \lambda_2}{\alpha + 1} = \frac{1}{\lambda_1 \lambda_2} \text{Var}[q_1] = \frac{1}{\lambda_1 \lambda_2} \text{Var}[q_2], \quad (2)$$

Measuring ancestry variation with FSTruct

where $1/(\lambda_1\lambda_2) = 4.5$ in the example plotted in Figure 3. Thus, the simulations and an analytical calculation confirm that in a simple Dirichlet model, the $F_{ST}/F_{ST}^{\max}$ measure has a linear relationship with the variance across sampled individuals of membership coefficient q_1 (or q_2). Importantly, the expected value of $F_{ST}/F_{ST}^{\max}$ in eq. 2 is independent of the mean membership coefficients, depending only on the Dirichlet parameter α , which controls variability. This result supports the use of $F_{ST}/F_{ST}^{\max}$ to measure variability in populations that possess different mean membership coefficients.

Visual illustration of values of $F_{ST}/F_{ST}^{\max}$. Continuing with the Dirichlet simulations, we next sought to visually illustrate the relationship of $F_{ST}/F_{ST}^{\max}$ to the variance and mean of membership coefficients. We considered Q matrices with four different values of α , representing four levels of variance in membership coefficients, and two different vectors for the parametric mean membership coefficients λ . For each of the eight settings (four variances, two means), we considered two Q matrices.

These eight simulated pairs of Q matrices are visualized in Figure 4A and 4D respectively, where they are colored according to the value of the α parameter used to simulate them. For the lowest-variability case (α_1 , red), the simulated individual membership coefficients show little deviation from the mean, $\lambda = (\frac{2}{3}, \frac{1}{3})$ for 4A and $(\frac{9}{10}, \frac{1}{10})$ for 4D. As the variance parameter increases (α_2 , purple; α_3 , blue), variance in membership coefficients is increasingly visible. For the highest-variability case (α_4 , green), membership coefficients are centered on $(\lambda_1, \lambda_2) = (1, 0)$ for approximately $\frac{2}{3}$ or $\frac{9}{10}$ of the individuals, and on $(\lambda_1, \lambda_2) = (0, 1)$ for the remaining individuals.

Bootstrap distributions of $F_{ST}/F_{ST}^{\max}$ appear in Figure 4B for $\lambda = (\frac{2}{3}, \frac{1}{3})$ and in 4C for $\lambda = (\frac{9}{10}, \frac{1}{10})$. In these panels, we observe that $F_{ST}/F_{ST}^{\max}$ increases from the lowest-variability case (α_1) to the highest-variability case (α_4), in accord with the interpretation that $F_{ST}/F_{ST}^{\max}$ measures variability in membership coefficients.

Comparing panel 4B with panel 4C, we observe that the value of $F_{ST}/F_{ST}^{\max}$ is similar between matrices simulated with the same Dirichlet α parameter, irrespective of the mean membership coefficient vectors (λ) used to simulate the matrices. This pattern accords with the interpretation that $F_{ST}/F_{ST}^{\max}$ is driven by the variance of membership coefficients and not the mean—as reflected

Measuring ancestry variation with FSTruct

in the analytical result in eq. 2 that under the Dirichlet model, $\mathbb{E}[F_{ST}/F_{ST}^{\max}; \alpha, \boldsymbol{\lambda}, I]$ can be written so that it depends on α but not on $\boldsymbol{\lambda}$.

We also observe in Figure 4A and 4D that the variability among Q matrices simulated from the Dirichlet distribution with identical parameters—as reflected in comparisons of pairs of matrices of the same color within a panel—increases with α . This variability can lead Q matrices simulated with the same parameters to possess quite different sample means and variances, as is the case particularly for the two pairs of matrices simulated with α_4 in Figure 4D. Despite this sampling variability of Q matrices under the Dirichlet model, we observe that $F_{ST}/F_{ST}^{\max}$, which is largely driven by the underlying parameter α , is relatively stable across pairs of Q matrices.

Data examples

To illustrate the application of FSTruct, we apply the method to data examples that represent each of three distinct scenarios in which ancestry variability is of interest: (1) ancestry comparisons of admixed and non-admixed populations, (2) ancestry comparisons of populations representing different time periods or spatial locations, and (3) ancestry comparisons of distinct data sets corresponding to different sets of loci for the same individuals.

Admixed populations. A characteristic feature of recently admixed populations is that individuals vary greatly in their ancestry, with some individuals possessing most of their ancestry from one source population, and others possessing most of their ancestry from another source (VERDU and ROSENBERG, 2011; GRAVEL, 2012). Thus, in examining inferred cluster memberships, admixed populations might be expected to give rise to greater variability in ancestry than non-admixed populations.

We therefore evaluated $F_{ST}/F_{ST}^{\max}$ in three populations from an ADMIXTURE analysis performed by VERDU *et al.* (2017). The populations include an admixed population from Cape Verde, and Gambian and Iberian populations taken to represent African and European sources for the admixed population. The inferred genetic structure for the three populations is redrawn in Figure 5A.

We computed $F_{ST}/F_{ST}^{\max}$ for each of the three populations, measuring ancestry variability of the inferred cluster memberships within each of the three groups. For the non-admixed source

Measuring ancestry variation with FSTruct

populations, this quantity is 0.078 for the Gambian population and 0.064 for the Iberian population (Figure 5B). The value for the admixed Cape Verdean population is greater, equaling 0.100. Pairs of bootstrap distributions of $F_{ST}/F_{ST}^{\max}$ are significantly different ($p < 2 \times 10^{-16}$ for all three pairwise combinations, Wilcoxon rank sum test). The admixed Cape Verdean population is indeed observed to have greater variability in ancestry according to the $F_{ST}/F_{ST}^{\max}$ measure than the putative source populations, supporting the use of the measure to distinguish clustering patterns in admixed and non-admixed populations.

Populations over time or space. Geographic movements of populations shape patterns of genetic ancestry for samples collected in different spatial locations or from the same location in different time periods. Locations or time periods whose samples contain individuals from many different sources or from recently admixed populations are expected to have highly variable ancestry, whereas locations or periods in which mixing of disparate populations is less salient are expected to have more homogeneous ancestry.

To explore an example of ancestry variability over time, we evaluated $F_{ST}/F_{ST}^{\max}$ in a STRUCTURE analysis conducted by ANTONIO *et al.* (2019) on samples from 29 archaeological sites near Rome spanning the last 12,000 years. These samples represent eight time periods: Mesolithic, Neolithic, Copper Age, Iron Age and Roman Republic, Imperial Rome, Late Antiquity, Medieval and Early Modern, and the present. The plot of the inferred genetic structure for these samples is redrawn in Figure 6A. ANTONIO *et al.* (2019) argued, based in part on their version of Figure 6A, that ancestry was variable during the Iron Age and Roman Republic, and highly variable during the Imperial Rome and Late Antiquity periods.

We computed $F_{ST}/F_{ST}^{\max}$ for each time period. This ratio is 0 for the Mesolithic, 0.0131 for the Neolithic, 0.0041 for the Copper Age, 0.0183 for the Iron Age and Roman Republic, 0.0192 for Imperial Rome, 0.0244 for Late Antiquity, 0.0186 for the Medieval and Early Modern period, and 0.0011 for modern individuals (Figure 6B). Pairs of bootstrap distributions of $F_{ST}/F_{ST}^{\max}$ are significantly different ($p < 2 \times 10^{-9}$ for all 28 pairwise combinations, Wilcoxon rank sum test). The numerical results validate the claims of ANTONIO *et al.* (2019) of high variability during the Iron Age and Roman Republic, Imperial Rome, and Late Antiquity periods. They lend increased

Measuring ancestry variation with FSTruct

granularity to these claims, suggesting that ancestry variability was steadily increasing during these three periods, with a maximum achieved during Late Antiquity.

Different genetic loci in the same samples. The ancestry patterns identified by population structure inference methods are influenced by the choice of loci used for the analysis. When data sets possess few loci, structure is not observed, and individuals have membership coefficients close to $(\frac{1}{K}, \frac{1}{K}, \dots, \frac{1}{K})$; different individuals possess similar membership coefficients. As the size of the data set increases, individuals come to have different membership coefficients, with, for example, individuals from two predefined populations possessing membership primarily in two distinct clusters.

To explore patterns of ancestry variability in data sets of different size, we evaluated $F_{ST}/F_{ST}^{\max}$ using results from a STRUCTURE analysis conducted by ALGEE-HEWITT *et al.* (2016). This study focused on 13 tetranucleotide loci commonly used for individual identification in forensic applications, the “CODIS loci.” In a worldwide human sample, the study compared analyses with the CODIS loci to analyses with a larger set of 779 non-CODIS loci and to analyses with sets of 13 non-CODIS tetranucleotide loci. The study claimed that the CODIS loci have similar ancestry information to sets of 13 non-CODIS tetranucleotide loci.

Four ancestry patterns from ALGEE-HEWITT *et al.* (2016), inferred from the same sample of individuals, are replotted in Figure 7. Figure 7A depicts a plot based on the CODIS loci. Figure 7B plots a “null data set” designed to possess no structure. Figure 7C plots a set of 13 non-CODIS tetranucleotide loci, and Figure 7D depicts a plot with 779 loci. The “null” plot shows little structure, the two plots with 13 loci show some structure, and the plot with 779 loci shows substantial structure.

We computed $F_{ST}/F_{ST}^{\max}$ for each analysis, for each plot evaluating variability in ancestry across all individuals within the plot. The ratio is lowest for the null data set, with a value of 0.009. It is 0.100 for both the CODIS loci and for the 13 non-CODIS loci. The ratio is substantially higher for the full 779 loci, with a value of 0.529. Five of the six pairs of bootstrap distributions of $F_{ST}/F_{ST}^{\max}$ are significantly different ($p < 2 \times 10^{-16}$, Wilcoxon rank sum test), the exception being that the two plots with 13 loci, CODIS and non-CODIS, do not show a significant difference ($p = 0.56$). The pattern

Measuring ancestry variation with FSTruct

of $F_{ST}/F_{ST}^{\max}$ values, with the smallest value for Figure 7B, intermediate values for Figure 7A and Figure 7C, and largest value for Figure 7D, captures increasing ancestry variability as the analyses move from a largely unstructured plot (Figure 7B) to partially unstructured plots (Figure 7A and 7C) to a substantially structured plot (Figure 7D). The lack of a significant difference in $F_{ST}/F_{ST}^{\max}$ between the plot for the CODIS loci and the plot for equally many non-CODIS loci supports the claim of ALGEE-HEWITT *et al.* (2016) that the CODIS loci contain comparable information about ancestry to other sets of loci with the same size.

Discussion

We have introduced a measure for quantifying variability across vectors of individual membership coefficients, as produced by population structure inference programs such as STRUCTURE and ADMIXTURE. Our measure is based on a mathematical analogy with the population differentiation statistic F_{ST} . Whereas F_{ST} traditionally measures variability in allele frequency vectors among populations, we have used F_{ST} to measure variability in membership coefficient vectors among individuals. Because the upper bound of F_{ST} as a function of the frequency of the most frequent allele is usually less than 1, we have employed a normalized version of this statistic, $F_{ST}/F_{ST}^{\max}$, which ranges in $[0, 1]$ for all matrices of membership coefficients and can thus be used to compare ancestry variability among different matrices.

Through both simulation and an analytical calculation under a Dirichlet distribution for membership coefficient vectors, we demonstrated that the expected value of $F_{ST}/F_{ST}^{\max}$ increases with variance of membership coefficients across individuals (Figure 3 and 4); indeed, it scales approximately linearly with the parametric variance in a model with $K = 2$ ancestral clusters (eq. 2). This result supports the use of $F_{ST}/F_{ST}^{\max}$ as a measure of variability in ancestry across individuals.

We also demonstrated the use of the measure in data sets exemplifying three scenarios in which ancestry variability is of particular interest. In a comparison of ancestry measured in admixed and non-admixed populations by VERDU *et al.* (2017), we found that the recently admixed Cape Verdean population exhibited greater variability in ancestry, as measured by $F_{ST}/F_{ST}^{\max}$, than did non-admixed populations (Figure 5). In a comparison of ancestries measured in different time

Measuring ancestry variation with FSTruct

periods in the same location, we provided quantitative support for a claim of ANTONIO *et al.* (2019) that certain eras in ancient Rome possessed more variable ancestry than others (Figure 6). Finally, in a comparison of different sets of loci studied in the same individuals, we found quantitative support both for the observation of ALGEE-HEWITT *et al.* (2016) that ancestry variability across individuals was similar for two different sets of 13 loci, and for an increase in ancestry variability in high-resolution data compared to data of lower resolution. In all three cases, our analyses provided quantitative support for claims previously argued primarily by qualitative observation.

Because the measure depends on Q matrices, limitations of the methods used to generate the Q matrices extend to our measure. For example, if individuals were mislabeled prior to analysis with methods such as STRUCTURE or ADMIXTURE, then our measure would be affected. Further, Q matrices generated by STRUCTURE and ADMIXTURE do not contain information about the magnitude of the difference between ancestral clusters; our measure only captures variation in ancestry with respect to the clusters that such programs infer.

The new measure, which we have implemented in the R package FSTruct, contributes to a body of methods for quantitative analysis of inferred membership coefficients. This collection of methods includes computations useful for analyzing the level of support observed for different numbers of clusters K (EVANNO *et al.*, 2005) and methods of aligning the clustering solutions observed in replicate analyses (JAKOBSSON and ROSENBERG, 2007; KOPELMAN *et al.*, 2015; BEHR *et al.*, 2016), as well as software for graphical display (ROSENBERG, 2004) and for managing files associated with the analysis (EARL and VONHOLDT, 2012).

A number of other studies have considered related but distinct problems in assessing variability of ancestry based on membership fractions. ROSENBERG *et al.* (2005) described a “clusteredness” statistic that measures the extent to which individuals are placed into single clusters rather than across multiple clusters. This statistic is maximal if each individual possesses a permutation of the membership vector $(1, 0, \dots, 0)$ and minimal if all individuals possess membership vector $(\frac{1}{K}, \frac{1}{K}, \dots, \frac{1}{K})$. KERMINEN *et al.* (2021) evaluated the Shannon entropy applied to individual-level membership vectors, assessing variation in time in the Shannon entropy for study participants with different birth years. Whereas both the clusteredness statistic of ROSENBERG *et al.* (2005) and the

Measuring ancestry variation with FSTruct

Shannon entropy statistic of KERMINEN *et al.* (2021) consider variability of the ancestry coefficients of single individuals, our $F_{ST}/F_{ST}^{\max}$ measure examines variability of ancestry coefficient vectors *across* individuals. Thus, for example, comparing individuals in corresponding matrices in Figure 4A and 4D, clusteredness increases (and Shannon entropy decreases) as the membership of the highest-membership cluster increases from Figure 4A to Figure 4D. However, $F_{ST}/F_{ST}^{\max}$, measuring variability *across* individuals, is similar in corresponding matrices in the two panels, reflecting the visual similarity between panels of the inter-individual patterns.

We note that in addition to analyzing the Q -matrices produced by population structure inference programs such as STRUCTURE and ADMIXTURE, FSTruct can quantify variability in any matrix whose rows sum to 1. Applications are potentially numerous. For example, single-cell sequencing technologies have enabled the identification and quantification of cell populations within tissues, revealing different patterns of variation, with some tissues containing few cell populations, while others are more diverse (WANG *et al.*, 2019). Our method enables comparisons of the variability of within-tissue cell populations, where tissues are analogous to individuals and cell populations are analogous to cluster memberships. Our method could also be applied to quantify variability among individuals of features such as mutational signatures, where the proportion of mutations belonging to a mutational type is analogous to a cluster membership (ALEXANDROV *et al.*, 2013; RAHBARI *et al.*, 2016).

Appendix: $\mathbb{E}[F_{ST}/F_{ST}^{\max}; \alpha, \lambda_1, \lambda_2, I]$

In this appendix, we evaluate the approximate expected value of $F_{ST}/F_{ST}^{\max}$ calculated for a sample of I individuals ($i = 1, 2, \dots, I$), each with $K = 2$ membership coefficients $(q_1^{(i)}, q_2^{(i)})$ drawn independently from the Dirichlet distribution $\text{Dir}(\alpha\lambda_1, \alpha\lambda_2)$. We use the notation $\mathbb{E}[F_{ST}/F_{ST}^{\max}; \alpha, \lambda_1, \lambda_2, I]$ or simply $\mathbb{E}[F_{ST}/F_{ST}^{\max}]$ to denote this expectation.

Overview

To obtain the expectation $\mathbb{E}[F_{ST}/F_{ST}^{\max}]$, we first sample I independent and identically distributed $\text{Dir}(\alpha\lambda_1, \alpha\lambda_2)$ random variables $(q_1^{(i)}, q_2^{(i)})$, where $q_1^{(i)}$ is the membership coefficient of individual i

Measuring ancestry variation with FSTruct

in cluster 1, $q_2^{(i)}$ is the membership coefficient of individual i in cluster 2, and $q_1^{(i)} + q_2^{(i)} = 1$. We assume that the sample size I is large.

We assume without loss of generality that the parametric mean membership coefficient for cluster 1 is at least as large as that for cluster 2; that is, $\lambda_1 \geq \lambda_2$. As $I \rightarrow \infty$, by the strong law of large numbers (SERFLING, 1980, section 1.8), the sample mean membership coefficient for cluster 1, $\frac{1}{I} \sum_{i=1}^I q_1^{(i)}$, converges almost surely to the parametric mean λ_1 , and the sample mean membership coefficient for cluster 2, $\frac{1}{I} \sum_{i=1}^I q_2^{(i)}$, converges almost surely to the parametric mean λ_2 . Hence, for large I , the probability approaches 1 that $\frac{1}{I} \sum_{i=1}^I q_1^{(i)} \geq \frac{1}{I} \sum_{i=1}^I q_2^{(i)}$. As a result, because we consider large I , we assume that the cluster with the greater *parametric* mean membership coefficient, cluster 1, also has the greater *sample* mean membership coefficient. We denote this sample mean, the mean membership of cluster 1 in a simulated population, by $M = \frac{1}{I} \sum_{i=1}^I q_1^{(i)}$. By definition, $M \geq \frac{1}{2}$. As stated above, $M \xrightarrow{a.s.} \lambda_1$ as $I \rightarrow \infty$.

The quantity whose expectation we wish to evaluate under the model, $F_{ST}/F_{ST}^{\max}$, is a function of the sampled membership coefficients, $(q_1^{(1)}, q_2^{(1)}), (q_1^{(2)}, q_2^{(2)}), \dots, (q_1^{(I)}, q_2^{(I)})$. We let $f(q_1^{(i)}, q_2^{(i)})$ represent the Dirichlet probability density for $(q_1^{(i)}, q_2^{(i)})$; because we are considering vectors with two components, the Dirichlet reduces to a Beta distribution (KOTZ *et al.*, 2000, p. 487),

$$f(q_1^{(i)}, q_2^{(i)}) = \frac{(q_1^{(i)})^{\alpha\lambda_1-1} (q_2^{(i)})^{\alpha\lambda_2-1}}{B(\alpha\lambda_1, \alpha\lambda_2)}, \quad (\text{A.1})$$

where

$$B(\alpha\lambda_1, \alpha\lambda_2) = \frac{\Gamma(\alpha\lambda_1) \Gamma(\alpha\lambda_2)}{\Gamma(\alpha\lambda_1 + \alpha\lambda_2)} = \frac{\Gamma(\alpha\lambda_1) \Gamma(\alpha\lambda_2)}{\Gamma(\alpha)}, \quad (\text{A.2})$$

and Γ is the gamma function.

The $(q_1^{(i)}, q_2^{(i)})$ are independent and identically distributed. Hence, the expectation is

$$\begin{aligned} \mathbb{E} \left[\frac{F_{ST}}{F_{ST}^{\max}}; \alpha, \lambda_1, \lambda_2, I \right] &= \int_{q_1^{(1)}=0}^1 \int_{q_1^{(2)}=0}^1 \cdots \int_{q_1^{(I)}=0}^1 \frac{F_{ST}}{F_{ST}^{\max}} \left(q_1^{(1)}, q_1^{(2)}, \dots, q_1^{(I)} \right) \times \\ &\quad f(q_1^{(1)}, q_2^{(1)}) f(q_1^{(2)}, q_2^{(2)}) \cdots f(q_1^{(I)}, q_2^{(I)}) dq_1^{(1)} dq_1^{(2)} \cdots dq_1^{(I)}. \end{aligned} \quad (\text{A.3})$$

Measuring ancestry variation with FSTruct

With this expression in hand, we proceed by writing the expression for $F_{ST}/F_{ST}^{\max}$ in terms of the membership coefficients $(q_1^{(i)}, q_2^{(i)})$ and the sample size I . We then compute the integral, making use of the Dirichlet parameters α , λ_1 , and λ_2 .

Approximating $F_{ST}/F_{ST}^{\max}$ under the Dirichlet model

The value of $F_{ST}/F_{ST}^{\max}$ calculated for a population of I individuals with membership coefficients $(q_1^{(1)}, q_2^{(1)}), \dots, (q_1^{(I)}, q_2^{(I)})$, can be written using eqs. 3 and 5 of ALCALA and ROSENBERG (2017),

$$\begin{aligned} F_{ST} &= \frac{\frac{1}{I} \sum_{i=1}^I \left(q_1^{(i)}\right)^2 - M^2}{M(1-M)} \\ F_{ST}^{\max} &= \frac{\lfloor IM \rfloor + \{IM\}^2 - IM^2}{IM(1-M)}. \end{aligned}$$

We obtain

$$\frac{F_{ST} \left(q_1^{(1)}, q_1^{(2)}, \dots, q_1^{(I)}\right)}{F_{ST}^{\max} \left(q_1^{(1)}, q_1^{(2)}, \dots, q_1^{(I)}\right)} = \frac{\sum_{i=1}^I \left(q_1^{(i)}\right)^2 - IM^2}{\lfloor IM \rfloor + \{IM\}^2 - IM^2}. \quad (\text{A.4})$$

Recall that $M = \frac{1}{I} \sum_{i=1}^I q_1^{(i)}$ is the sample mean membership of the most prevalent ancestral cluster, assuming that the cluster with the greater parametric mean membership is also the cluster with the greater sample mean membership.

We now make an approximation to the denominator of eq. A.4. Because $\lfloor IM \rfloor = IM - \{IM\}$, $\lfloor IM \rfloor + \{IM\}^2 = IM - (\{IM\} - \{IM\}^2) = IM - \delta$ where the error term $\delta = \{IM\}(1 - \{IM\})$ lies in $[0, \frac{1}{4}]$, taking its maximal value of $\frac{1}{4}$ when $\{IM\} = \frac{1}{2}$. For large sample size I , because $M \geq \frac{1}{2}$ and $\delta \leq \frac{1}{4}$, $IM \gg \delta$, so that $\lfloor IM \rfloor + \{IM\}^2 \approx IM$. Thus, we substitute IM in place of $\lfloor IM \rfloor + \{IM\}^2$ in eq. A.4, obtaining

$$\frac{F_{ST}}{F_{ST}^{\max}} \left(q_1^{(1)}, q_1^{(2)}, \dots, q_1^{(I)}\right) \approx \frac{\sum_{i=1}^I \left(q_1^{(i)}\right)^2 - IM^2}{IM(1-M)}. \quad (\text{A.5})$$

This assumption is equivalent to setting $F_{ST}^{\max} = 1$.

To find an approximation for $\mathbb{E}[F_{ST}/F_{ST}^{\max}]$, it is convenient to make a further approximation in

Measuring ancestry variation with FSTruct

eq. A.5, substituting M with λ_1 . We justify this substitution by proving that as $I \rightarrow \infty$,

$$\frac{\sum_{i=1}^I \left(q_1^{(i)}\right)^2 - IM^2}{IM(1-M)} \xrightarrow{a.s.} \frac{\sum_{i=1}^I \left(q_1^{(i)}\right)^2 - I\lambda_1^2}{I\lambda_1(1-\lambda_1)}. \quad (\text{A.6})$$

Subtracting the right-hand side from the left-hand side, proving eq. A.6 is equivalent to proving

$$\left[\frac{1}{I} \sum_{i=1}^I \left(q_1^{(i)}\right)^2 \right] \left[\frac{1}{M(1-M)} - \frac{1}{\lambda_1(1-\lambda_1)} \right] + \left[\frac{-M}{1-M} + \frac{\lambda_1}{1-\lambda_1} \right] \xrightarrow{a.s.} 0. \quad (\text{A.7})$$

If $X_n \xrightarrow{a.s.} X$ and $Y_n \xrightarrow{a.s.} Y$, then $X_n + Y_n \xrightarrow{a.s.} X + Y$ (GRIMMETT and STIRZAKER 2001a, p. 336, exercise 2; GRIMMETT and STIRZAKER 2001b, p. 354, exercise 2), so the sum of two terms that converge almost surely to 0 also converges almost surely to 0. Hence, it suffices to separately prove almost sure convergence to 0 of the two terms summed in eq. A.7.

For the right-hand term, we use the continuous mapping theorem, which states that for a continuous function g and a random vector X_n , if $X_n \xrightarrow{a.s.} X$ then $g(X_n) \xrightarrow{a.s.} g(X)$ (VAN DER VAART, 1998, p. 7, Theorem 2.3). We consider the continuous function $g(x) = -x/(1-x) + \lambda_1/(1-\lambda_1)$ and recall that as $I \rightarrow \infty$, $M \xrightarrow{a.s.} \lambda_1$. It follows that as $I \rightarrow \infty$, and $M \xrightarrow{a.s.} \lambda_1$, $g(M) \xrightarrow{a.s.} g(\lambda_1)$; that is, $-M/(1-M) + \lambda_1/(1-\lambda_1) \xrightarrow{a.s.} 0$.

For the left-hand term, the factor $1/[M(1-M)] - 1/[\lambda_1(1-\lambda_1)]$ converges almost surely to 0 by the continuous mapping theorem with $g(x) = 1/[x(1-x)] - 1/[\lambda_1(1-\lambda_1)]$. By the strong law of numbers, $\frac{1}{I} \sum_{i=1}^I (q_1^{(i)})^2$ converges almost surely to $\gamma = \mathbb{E}[(q_1^{(i)})^2] = \text{Var}[q_1^{(i)}] + \lambda_1^2 = \lambda_1(1-\lambda_1)/(\alpha + 1) + \lambda_1^2 = \lambda_1(\alpha\lambda_1 + 1)/(\alpha + 1)$ (SERFLING, 1980, Theorem B). If $X_n \xrightarrow{a.s.} X$ and $Y_n \xrightarrow{a.s.} Y$, then $X_n Y_n \xrightarrow{a.s.} XY$ (GRIMMETT and STIRZAKER 2001a, p. 336, exercise 2; GRIMMETT and STIRZAKER 2001b, p. 354, exercise 2), so that the left-hand term of eq. A.7 converges almost surely to $\gamma \times 0 = 0$.

Evaluating $\mathbb{E}[F_{ST}/F_{ST}^{\max}]$ under the Dirichlet model

Inserting our expression for $f(q_1^{(i)}, q_2^{(i)})$ from eq. A.1 and our expression for $F_{ST}/F_{ST}^{\max}$ from eq. A.6 into eq. A.3 allows us to write an approximate expression for the expectation of $F_{ST}/F_{ST}^{\max}$ given

Measuring ancestry variation with FSTruct

the parameters of the Dirichlet distribution and the sample size I :

$$\begin{aligned} \mathbb{E} \left[\frac{F_{ST}}{F_{ST}^{\max}}; \alpha, \lambda_1, \lambda_2, I \right] &\approx \int_{q_1^{(1)}=0}^1 \int_{q_1^{(2)}=0}^1 \cdots \int_{q_1^{(I)}=0}^1 \frac{\sum_{i=1}^I \left(q_1^{(i)} \right)^2 - I\lambda_1^2}{I\lambda_1(1-\lambda_1)} \times \frac{(q_1^{(1)})^{\alpha\lambda_1-1} (q_2^{(1)})^{\alpha\lambda_2-1}}{B(\alpha\lambda_1, \alpha\lambda_2)} \\ &\times \frac{(q_1^{(2)})^{\alpha\lambda_1-1} (q_2^{(2)})^{\alpha\lambda_2-1}}{B(\alpha\lambda_1, \alpha\lambda_2)} \times \cdots \times \frac{(q_1^{(I)})^{\alpha\lambda_1-1} (q_2^{(I)})^{\alpha\lambda_2-1}}{B(\alpha\lambda_1, \alpha\lambda_2)} dq_1^{(1)} dq_1^{(2)} \cdots dq_1^{(I)}. \end{aligned} \quad (\text{A.8})$$

Examining the quantity $\sum_{i=1}^I (q_1^{(i)})^2 - I\lambda_1^2$, we observe that eq. A.8 can be decomposed as a sum of $I + 1$ terms, one for each of the $(q_1^{(i)})^2$ terms, and one for the $-I\lambda_1^2$ term. Assign the first I of these separate terms the labels $L_1, L_2, \dots, L_I$ and the $-I\lambda_1^2$ term the label L_* , so that $\mathbb{E}[F_{ST}/F_{ST}^{\max}; \alpha, \lambda_1, \lambda_2, I] \approx L_1 + L_2 + \dots + L_I + L_*$.

We begin by evaluating the term L_* , which can be written

$$L_* = \frac{-I\lambda_1^2}{I\lambda_1(1-\lambda_1)} \prod_{i=1}^I \int_{q_1^{(i)}=0}^1 \frac{(q_1^{(i)})^{\alpha\lambda_1-1} (q_2^{(i)})^{\alpha\lambda_2-1}}{B(\alpha\lambda_1, \alpha\lambda_2)} dq_1^{(i)}.$$

Recalling that $q_2^{(i)} = 1 - q_1^{(i)}$, we observe that the integrand $(q_1^{(i)})^{\alpha\lambda_1-1} (q_2^{(i)})^{\alpha\lambda_2-1} / B(\alpha\lambda_1, \alpha\lambda_2)$ is simply the Beta probability density function, which integrates to one. Hence, the product evaluates to 1 and L_* simply equals a constant:

$$L_* = -\frac{I\lambda_1^2}{I\lambda_1(1-\lambda_1)} = -\frac{\lambda_1}{1-\lambda_1}. \quad (\text{A.9})$$

We next evaluate the L_i terms. For each i in $1, 2, \dots, I$,

$$L_i = \frac{1}{I\lambda_1(1-\lambda_1)} \int_{q_1^{(i)}=0}^1 \frac{(q_1^{(i)})^{\alpha\lambda_1+1} (q_2^{(i)})^{\alpha\lambda_2-1}}{B(\alpha\lambda_1, \alpha\lambda_2)} dq_1^{(i)} \prod_{\substack{j=1 \\ j \neq i}}^I \int_{q_1^{(j)}=0}^1 \frac{(q_1^{(j)})^{\alpha\lambda_1-1} (q_2^{(j)})^{\alpha\lambda_2-1}}{B(\alpha\lambda_1, \alpha\lambda_2)} dq_1^{(j)}.$$

As was the case for L_* , the integrand of the integral inside the product is the Beta probability density function, so the product evaluates to one. Thus,

$$L_i = \frac{1}{I\lambda_1(1-\lambda_1)} \int_{q_1^{(i)}=0}^1 \frac{(q_1^{(i)})^{\alpha\lambda_1+1} (q_2^{(i)})^{\alpha\lambda_2-1}}{B(\alpha\lambda_1, \alpha\lambda_2)} dq_1^{(i)}.$$

Measuring ancestry variation with FSTruct

The remaining integral can be evaluated by noting that $\int_0^1 x^{a-1}(1-x)^{b-1}dx = B(a, b)$. We employ this identity to simplify L_i , obtaining

$$L_i = \frac{1}{I\lambda_1(1-\lambda_1)} \frac{B(\alpha\lambda_1 + 2, \alpha\lambda_2)}{B(\alpha\lambda_1, \alpha\lambda_2)}.$$

By eq. A.2 and the property of gamma functions $\Gamma(z+1) = z\Gamma(z)$, this expression simplifies to

$$L_i = \frac{\alpha\lambda_1 + 1}{I(1-\lambda_1)(\alpha+1)}. \quad (\text{A.10})$$

We now combine eqs. A.9 and A.10 to complete the calculation in eq. A.8, noting that L_i does not depend on i , so that each L_i follows eq. A.10.

$$\begin{aligned} \mathbb{E} \left[\frac{F_{ST}}{F_{ST}^{\max}}; \alpha, \lambda_1, \lambda_2, I \right] &\approx L_1 + L_2 + \dots + L_I + L_* \\ &= \frac{I(\alpha\lambda_1 + 1)}{I(1-\lambda_1)(\alpha+1)} - \frac{\lambda_1}{1-\lambda_1} \\ &= \frac{1}{\alpha+1}. \end{aligned} \quad (\text{A.11})$$

Acknowledgments. We acknowledge support from NIH grant R01 HG005855. MLM acknowledges support from a National Science Foundation Graduate Research Fellowship and the Anne T. and Robert M. Bass Stanford Graduate Fellowship. We thank P. Verdu, M. Antonio, and M. Edge for assistance with data sets from their studies, H. Moots and J. Pritchard for helpful comments on the method, and D. Cotter and J. Mooney for suggestions for the software. Where authors are identified as personnel of the International Agency for Research on Cancer/World Health Organization, the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy, or views of the International Agency for Research on Cancer/World Health Organization.

References

- ALCALA, N., and N. A. ROSENBERG, 2017 Mathematical constraints on F_{ST} : Biallelic markers in arbitrarily many populations. *Genetics* **206**: 1581–1600.
- ALCALA, N., and N. A. ROSENBERG, 2019 G'_{ST} , Jost's D, and F_{ST} are similarly constrained by allele frequencies: A mathematical, simulation, and empirical study. *Molecular Ecology* **28**: 1624–1636.
- ALCALA, N., and N. A. ROSENBERG, 2021 Mathematical constraints on F_{ST} : multiallelic markers in arbitrarily many populations. *bioRxiv* : 10.1101/2021.07.23.453474.

Measuring ancestry variation with FSTruct

- 429 ALEXANDER, D. H., and K. LANGE, 2011 Enhancements to the ADMIXTURE algorithm for individual
430 ancestry estimation. *BMC Bioinformatics* **12**: 246.
- 431 ALEXANDER, D. H., J. NOVEMBRE, and K. LANGE, 2009 Fast model-based estimation of ancestry in
432 unrelated individuals. *Genome Research* **19**: 1655–1664.
- 433 ALEXANDROV, L. B., S. NIK-ZAINAL, D. C. WEDGE, S. A. J. R. APARICIO, S. BEHJATI, *et al.*, 2013
434 Signatures of mutational processes in human cancer. *Nature* **500**: 415–421.
- 435 ALGEE-HEWITT, B. F., M. D. EDGE, J. KIM, J. Z. LI, and N. A. ROSENBERG, 2016 Individual identifica-
436 bility predicts population identifiability in forensic microsatellite markers. *Current Biology* **26**: 935–942.
- 437 ANTONIO, M. L., Z. GAO, H. M. MOOTS, M. LUCCI, F. CANDILIO, *et al.*, 2019 Ancient Rome: a genetic
438 crossroads of Europe and the Mediterranean. *Science* **366**: 708–714.
- 439 BEHR, A. A., K. Z. LIU, G. LIU-FANG, P. NAKKA, and S. RAMACHANDRAN, 2016 pong: fast analysis and
440 visualization of latent clusters in population genetic data. *Bioinformatics* **32**: 2817–2823.
- 441 CORANDER, J., P. MARTTINEN, J. SIRÉN, and J. TANG, 2008 Enhanced Bayesian modelling in BAPS
442 software for learning genetic structures of populations. *BMC Bioinformatics* **9**: 539.
- 443 CORANDER, J., P. WALDMANN, P. MARTTINEN, and M. J. SILLANPÄÄ, 2004 BAPS 2: Enhanced possibil-
444 ities for the analysis of genetic population structure. *Bioinformatics* **20**: 2363–2369.
- 445 EARL, D. A., and B. M. VONHOLDT, 2012 STRUCTURE HARVESTER: A website and program for
446 visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*
447 **4**: 359–361.
- 448 EVANNO, G., S. REGNAUT, and J. GOUDET, 2005 Detecting the number of clusters of individuals using the
449 software STRUCTURE: a simulation study. *Molecular Ecology* **14**: 2611–2620.
- 450 FALUSH, D., M. STEPHENS, and J. K. PRITCHARD, 2003 Inference of population structure using multilocus
451 genotype data: Linked loci and correlated allele frequencies. *Genetics* **164**: 1567–1587.
- 452 FALUSH, D., M. STEPHENS, and J. K. PRITCHARD, 2007 Inference of population structure using multilocus
453 genotype data: Dominant markers and null alleles. *Molecular Ecology Notes* **7**: 574–578.
- 454 GRAVEL, S., 2012 Population genetics models of local ancestry. *Genetics* **191**: 607–619.
- 455 GRIMMETT, G. R., and D. R. STIRZAKER, 2001a *One Thousand Exercises in Probability*. Oxford University
456 Press, New York.
- 457 GRIMMETT, G. R., and D. R. STIRZAKER, 2001b *Probability and Random Processes*. Oxford University
458 Press, New York, third edition.
- 459 GUILLOT, G., and L. ORLANDO, 2017 Population Structure. Oxford Bibliographies :
460 10.1093/OBO/9780199941728-0057.
- 461 HUBISZ, M. J., D. FALUSH, M. STEPHENS, and J. K. PRITCHARD, 2009 Inferring weak population structure
462 with the assistance of sample group information. *Molecular Ecology Resources* **9**: 1322–1332.
- 463 HUELSENBECK, J. P., and P. ANDOLFATTO, 2007 Inference of population structure under a Dirichlet process
464 model. *Genetics* **175**: 1787–1802.
- 465 JAKOBSSON, M., M. D. EDGE, and N. A. ROSENBERG, 2013 The relationship between F_{ST} and the frequency
466 of the most frequent allele. *Genetics* **193**: 515–528.

Measuring ancestry variation with FSTruct

- 467 JAKOBSSON, M., and N. A. ROSENBERG, 2007 CLUMPP: A cluster matching and permutation program
468 for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23**:
469 1801–1806.
- 470 KERMINEN, S., N. CERIOLI, D. PACAUSKAS, A. S. HAVULINNA, M. PEROLA, *et al.*, 2021 Changes in the
471 fine-scale genetic structure of Finland through the 20th century. *PLoS Genetics* **17**: e1009347.
- 472 KOPELMAN, N. M., J. MAYZEL, M. JAKOBSSON, N. A. ROSENBERG, and I. MAYROSE, 2015 CLUMPAK: a
473 program for identifying clustering modes and packaging population structure inference across K. *Molecular*
474 *Ecology Resources* **15**: 1179–1191.
- 475 KOTZ, S., N. BALAKRISHNAN, and N. JOHNSON, 2000 *Continuous Multivariate Distributions, Volume 1:*
476 *Models and Applications*. Wiley, New York, second edition.
- 477 PRITCHARD, J. K., M. STEPHENS, and P. DONNELLY, 2000 Inference of population structure using multi-
478 locus genotype data. *Genetics* **155**: 945–959.
- 479 RAHBARI, R., A. WUSTER, S. J. LINDSAY, R. J. HARDWICK, L. B. ALEXANDROV, *et al.*, 2016 Timing,
480 rates and spectra of human germline mutation. *Nature Genetics* **48**: 126–133.
- 481 ROSENBERG, N. A., 2004 DISTRUCT: A program for the graphical display of population structure. *Molec-*
482 *ular Ecology Notes* **4**: 137–138.
- 483 ROSENBERG, N. A., S. MAHAJAN, S. RAMACHANDRAN, C. ZHAO, J. K. PRITCHARD, *et al.*, 2005 Clines,
484 clusters, and the effect of study design on the inference of human population structure. *PLoS Genetics* **1**:
485 660–671.
- 486 SERFLING, R., 1980 *Approximation Theorems of Mathematical Statistics*. Wiley, Hoboken, NJ.
- 487 VAN DER VAART, A. W., 1998 *Asymptotic Statistics*. Cambridge University Press, New York.
- 488 VERDU, P., E. M. JEWETT, T. J. PEMBERTON, N. A. ROSENBERG, and M. BAPTISTA, 2017 Parallel
489 trajectories of genetic and linguistic admixture in a genetically admixed creole population. *Current Biology*
490 **27**: 2529–2535.
- 491 VERDU, P., and N. A. ROSENBERG, 2011 A general mechanistic model for admixture histories of hybrid
492 populations. *Genetics* **189**: 1413–1426.
- 493 WANG, X., J. PARK, K. SUSZTAK, N. R. ZHANG, and M. LI, 2019 Bulk tissue cell type deconvolution with
494 multi-subject single-cell expression reference. *Nature Communications* **10**: 380.
- 495 **Data accessibility.** The FSTruct R package is available for download from github.com/MaikeMorrison/FSTruct.
496 The introductory vignette is linked from the package README file and provides a guide to use of the package.
- 497 **Author contributions.** MLM, NA, and NAR designed the study and performed the theoretical analysis.
498 MLM conducted the simulations, analyzed the data, and wrote the software. NAR supervised the study. All
499 authors wrote the manuscript.

Measuring ancestry variation with FSTruct

500 Figures

A F_{ST} calculation among population vectors of allele frequencies

	Populations (i)	Alleles (k)				$\sum_{k=1}^K q_k^{(i)}$
		1	2	...	K	
Compute F_{ST} among rows	1	$q_1^{(1)}$	$q_2^{(1)}$	...	$q_K^{(1)}$	1
	2	$q_1^{(2)}$	$q_2^{(2)}$	...	$q_K^{(2)}$	1
	...	...	...	...	...	...
	I	$q_1^{(I)}$	$q_2^{(I)}$	...	$q_K^{(I)}$	1

B F_{ST} calculation among individual vectors of membership coefficients

	Individuals (i)	Clusters (k)				$\sum_{k=1}^K q_k^{(i)}$
		1	2	...	K	
Compute F_{ST} among rows	1	$q_1^{(1)}$	$q_2^{(1)}$	...	$q_K^{(1)}$	1
	2	$q_1^{(2)}$	$q_2^{(2)}$	...	$q_K^{(2)}$	1
	...	...	...	...	...	...
	I	$q_1^{(I)}$	$q_2^{(I)}$	...	$q_K^{(I)}$	1

Figure 1: The analogy of the use of F_{ST} to measure membership variability. (A) A standard application of F_{ST} to measure variability of allele frequency vectors across populations; $q_k^{(i)}$ is the frequency of allele k in population i . (B) Use of F_{ST} to measure variability of membership coefficient vectors across individuals; $q_k^{(i)}$ is the membership coefficient of individual i in cluster k . The matrix containing entries $q_k^{(i)}$ is a Q matrix.

Measuring ancestry variation with FSTruct

Figure 2: Bounds on F_{ST} as a function of M , the frequency of the most frequent allele—or the ancestry cluster of greatest membership, in our analogy. Bounds are evaluated using eq. 1 for different values of I , the number of populations (or the number of individuals, in our analogy). (A) $I = 2$. (B) $I = 3$. (C) $I = 5$. (D) $I = 10$. (E) $I = 50$.

Measuring ancestry variation with FSTruct

Figure 3: Linear relationship between $F_{ST}/F_{ST}^{\max}$ and $\text{Var}[q_1] = \text{Var}[q_2]$, the variance across individuals of individual membership coefficients under a Dirichlet distribution. For each of 45 values of $\text{Var}[q_1] = \text{Var}[q_2]$, 500 points are plotted, each representing a random Q matrix with dimensions 50×2 . Rows of the Q matrix are simulated using a Dirichlet distribution with means $\lambda = (\lambda_1, \lambda_2) = (\frac{2}{3}, \frac{1}{3})$ and variances $\text{Var}[q_1] = \text{Var}[q_2] = \lambda_1 \lambda_2 / (\alpha + 1)$, with α chosen to produce variances 0.001, 0.005, 0.01, 0.015, $\dots$ 0.22. Each Q matrix gives rise to an associated value of $F_{ST}/F_{ST}^{\max}$, plotted on the vertical axis. A regression line fit to the 500×45 points with intercept 0 has slope 4.5, or $1/(\lambda_1 \lambda_2) = 1/(\frac{2}{3} \times \frac{1}{3})$, and it explains 99% of the variability in $F_{ST}/F_{ST}^{\max}$. Grey lines mark the 2.5% and 97.5% percentiles, and thus contain 95% of the points.

Measuring ancestry variation with FSTruct

Figure 4: Dependence of bootstrap distributions of $F_{ST}/F_{ST}^{\max}$ for simulated Q matrices on the Dirichlet variance parameter α , rather than the Dirichlet mean λ . (A, D) Q matrices simulated using specified $\text{Dir}(\alpha\lambda)$ distributions. (B, C) Bootstrap distributions of $F_{ST}/F_{ST}^{\max}$ for Q matrices from (A) and (D), plotted directly below or above the corresponding matrix. In both (A) and (D), eight matrices were simulated, two for each of four values of α selected to span the range of variances: $\alpha_1 = 21901/99$, $\alpha_2 = 341/99$, $\alpha_3 = 101/99$, $\alpha_4 = 1/99$. Matrices are annotated by associated variances $\sigma^2 = \lambda_1\lambda_2/(\alpha + 1)$. In (A), matrices are simulated with parametric mean $\lambda = (\frac{2}{3}, \frac{1}{3})$ and are taken from matrices plotted in Figure 3. In (D), matrices are simulated with a more extreme parametric mean, $\lambda = (\frac{9}{10}, \frac{1}{10})$. Each vertical bar represents an individual membership coefficient vector (q_1, q_2) ; the proportion of each bar colored a darker shade represents q_1 and the proportion in a lighter shade corresponds to q_2 . The parametric variance of a Q matrix, $\sigma^2 = \lambda_1\lambda_2/(\alpha + 1)$, ranges in $(0, \frac{2}{9})$ for $\lambda = (\frac{2}{3}, \frac{1}{3})$ and in $(0, 0.09)$ for $\lambda = (\frac{9}{10}, \frac{1}{10})$. The empirical variance s^2 is computed for each matrix using the sample mean $\bar{q} = (\frac{1}{I} \sum_{i=1}^I q_1^{(i)}, \frac{1}{I} \sum_{i=1}^I q_2^{(i)})$ in place of the parametric mean λ . The values of $F_{ST}/F_{ST}^{\max}$ for the eight matrices in (A) are 0.004 and 0.005 for the two simulated with α_1 , 0.203 and 0.230 for α_2 , 0.496 and 0.461 for α_3 , and 1.000 and 0.997 for α_4 . The values of $F_{ST}/F_{ST}^{\max}$ for the eight matrices in (D) are 0.003 and 0.005 for α_1 , 0.157 and 0.287 for α_2 , 0.539 and 0.571 for α_3 , and 1.000 and 1.000 for α_4 . In (B) and (C), each bootstrap distribution includes 1,000 bootstrap samples of the $I = 50$ individuals in the associated Q matrix.

Measuring ancestry variation with FSTruct

Figure 5: Variability of ancestry in admixed and non-admixed populations. (A) $K = 4$ ADMIXTURE analysis of Gambian ($n = 109$), Cape Verdean ($n = 44$), and Iberian ($n = 107$) samples. Adapted from VERDU *et al.* (2017). (B) Bootstrap distributions of the ancestry variability measure, F_{ST}/F_{ST}^{max} , for each population (1,000 replicates).

Measuring ancestry variation with FSTruct

Figure 6: Variability of ancestry over time. (A) $K = 5$ STRUCTURE analysis of samples from eight time periods: Mesolithic ($n = 3$), Neolithic ($n = 10$), Copper Age ($n = 3$), Iron Age and Roman Republic ($n = 11$), Imperial Rome ($n = 48$), Late Antiquity ($n = 24$), Medieval and Early Modern ($n = 28$), and Present ($n = 15$). Adapted from ANTONIO *et al.* (2019). (B) Bootstrap distributions of the ancestry variability measure, F_{ST}/F_{ST}^{max} , for each population (1,000 replicates).

Measuring ancestry variation with FSTruct

Figure 7: Variability of ancestry for analyses with different loci from the same samples. (A-D) $K = 4$ STRUCTURE analyses of four different sets of loci for a worldwide human sample. Adapted from ALGEE-HEWITT *et al.* (2016). (A) 13 CODIS tetranucleotide microsatellite loci. (B) A simulated null data set with no population structure. (C) 13 non-CODIS tetranucleotide microsatellite loci. (D) Full data set of 779 tetranucleotide loci. (E) Bootstrap distributions of the ancestry variability measure, F_{ST}/F_{ST}^{max} , for each sample (1,000 replicates).