Received: 1 September 2021 | Revised: 9 March 2022 | Accepted: 13 May 2022

DOI: 10.1111/1755-0998.13647

MOLECULAR ECOLOGY RESOURCES WILEY

RESOURCE ARTICLE

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

Maike L. Morrison[1] | Nicolas Alcala[2] | Noah A. Rosenberg[1]

[1]Department of Biology, Stanford University, Stanford, California, USA

[2]Rare Cancers Genomics Team (RCG), Genomic Epidemiology Branch (GEM), International Agency for Research on Cancer/World Health Organisation (IARC/WHO), Lyon, France

Correspondence
Maike L. Morrison, Department of Biology, Stanford University, Stanford, CA 94305 USA.
Email: maikem@stanford.edu

Funding information
National Science Foundation, Grant/Award Number: BCS-2116322; National Institutes of Health, Grant/Award Number: R01 HG005855; France-Stanford Center for Interdisciplinary Studies

Handling Editor: Nick Hamilton Barton

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 predefined 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 and does not depend on its mean. 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 predefined groups in their level of membership coefficient variability. Our methods are implemented in the R package FSTruct.

KEYWORDS
F_{ST}, admixture, population structure

1 | 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 & Lange, 2011), BAPS (Corander et al., 2004, 2008) and STRUCTURE (Falush 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 & Orlando, 2017).

In model-based inference of population structure, individuals are clustered based on their multilocus 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 nonadmixed populations, between populations from different time periods or between different types of data from the same sampled individuals.

For example, in a study of ancient human DNA samples dating over a period of thousands of years, Antonio et al. (2019) sought to examine whether 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 (Figure 1 of Antonio et al., 2019).

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.

2 | MATERIALS AND METHODS

2.1 | 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

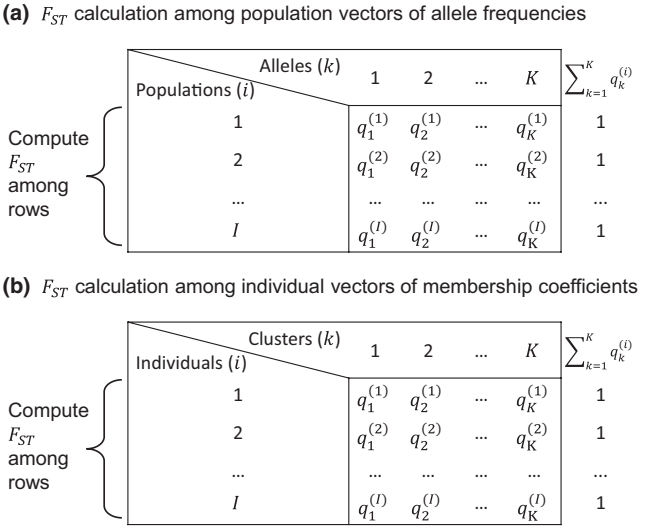

FIGURE 1 The analogy of the use of F_{ST} to measure membership variability. (a) A standard application of F_{ST} to measure variability of allele frequency vectors across populations; $q_k^{(i)}$ is the frequency of allele k in population i. (b) Use of F_{ST} to measure variability of membership coefficient 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

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: 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 (Alcala & Rosenberg, 2017, 2019; Jakobsson

et al., 2013). The 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 its minimum, $\frac{1}{K}$, or its maximum, 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, equalling 0 when members of a population have identical membership and equalling 1 when vectors of membership coefficients are maximally variable.

2.2 | 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 $0 \le q_k^{(i)} \le 1$ 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, assuming Hardy–Weinberg equilibrium within subpopulations, or $H_S = 1 - \frac{1}{I}\sum_{i=1}^{I}\sum_{k=1}^{K}\left(q_k^{(i)}\right)^2$. 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: $H_T = 1 - \sum_{k=1}^{K}\left(\frac{1}{I}\sum_{i=1}^{I} q_k^{(i)}\right)^2$. 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 (2022) 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 & Rosenberg, 2022, Equation 3)

$$F_{ST}^{max} = \begin{cases} 1, & \sigma_1 = 1, 2, \ldots, I-1 \\ \dfrac{(I-1)\left[1-\sigma_1(J-1)(2-J\sigma_1)\right]}{I-\left[1-\sigma_1(J-1)(2-J\sigma_1)\right]}, & 0 < \sigma_1 < 1 \\ \dfrac{I(I-1)-\sigma_1^2+\lfloor\sigma_1\rfloor-2(I-1)\{\sigma_1\}+(2I-1)\{\sigma_1\}^2}{I(I-1)-\sigma_1^2-\lfloor\sigma_1\rfloor+2\sigma_1-\{\sigma_1\}^2}, & \text{non-integer } \sigma_1, 1 < \sigma_1 < I. \end{cases} \quad (1)$$

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

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 Equation 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 (2022) showed that for $0 < \sigma_1 \le 1$, 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 σ_1, one cluster with a coefficient of $1 - (J-1)\sigma_1 \le \sigma_1$ and all others with coefficients of 0. Note that in the scenario $0 < \sigma_1 \le 1$, the number of clusters K is larger than the number of individuals I; at the maximum, multiple clusters are tied with the same mean membership coefficient M.

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 (i.e. $q_1^{(i)} = 1$ for these individuals), a single individual has membership coefficient $\{\sigma_1\}$ for the cluster of greatest membership and coefficient

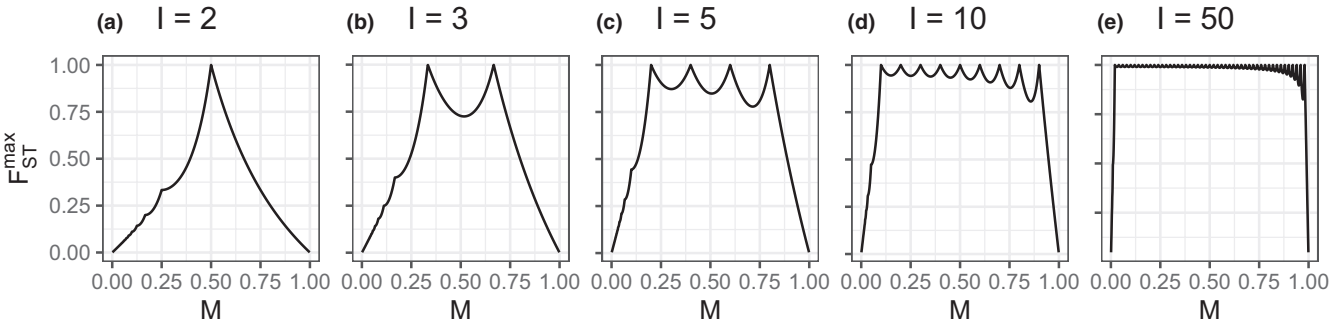

FIGURE 2 Bounds on F_{ST} as a function of M, the frequency of the most frequent allele—or the ancestry cluster of greatest membership, in our analogy. Bounds are evaluated using Equation 1 for different values of I, the number of populations (or the number of individuals, in our analogy). (a) $I = 2$. (b) $I = 3$. (c) $I = 5$. (d) $I = 10$. (e) $I = 50$

$1 - \{\sigma_1\}$ for one other cluster, and the remaining $I - \lfloor\sigma_1\rfloor - 1$ individuals each have membership coefficient 1 for mutually distinct ancestry clusters.

2.3 | 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 whether 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.

2.4 | Software availability

We have implemented our 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 samples and distributions for arbitrarily many Q matrices and visualize Q matrices.

3 | RESULTS

3.1 | Simulation examples

3.1.1 | 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 $(q_1, q_2, \ldots, q_K)$ that sum to one, $\sum_{k=1}^{K} q_k = 1$, and it has appeared in previous studies of membership coefficient vectors (Huelsenbeck & Andolfatto, 2007; Pritchard et al., 2000).

We treat individual membership coefficient vectors in a population as following a Dirichlet distribution with parameter vector $\alpha\lambda = \alpha(\lambda_1, \lambda_2, \ldots, \lambda_K)$, where $\sum_{k=1}^{K} \lambda_k = 1$. We denote this distribution by $\mathrm{Dir}(\alpha(\lambda_1, \lambda_2, \ldots, \lambda_K))$. Here, λ 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: $\mathrm{Var}[q_k] = \lambda_k(1 - \lambda_k)/(\alpha + 1)$. 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 $\mathrm{Dir}(\alpha(\lambda_1, \lambda_2, \ldots, \lambda_K))$ vectors, $(q_1, \ldots, q_K)$, 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 $(\lambda_1, \lambda_2, \ldots, \lambda_K)$. 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 parametric membership coefficient means and variances, (2) computing F_{ST} / F_{ST}^{max} for each Q matrix and (3) examining the relationship between the value of F_{ST} / F_{ST}^{max} for each Q matrix and the parametric variance of the Dirichlet distribution used to simulate it.

3.1.2 | Dirichlet simulations

To investigate the behaviour 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 $\mathrm{Dir}(\alpha(\lambda_1, \lambda_2))$ distribution.

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

We performed 500 replicate simulations of samples of 50 individuals for each of 45 values of α, choosing α values to obtain parametric variances 0.001, 0.005, 0.01, 0.015, ..., 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 parametric variance 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 $\sigma^2 = \mathrm{Var}[q_1] = \mathrm{Var}[q_2] = \left(\frac{2}{3} \times \frac{1}{3}\right)/(\alpha + 1)$, 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 (Equation A11)

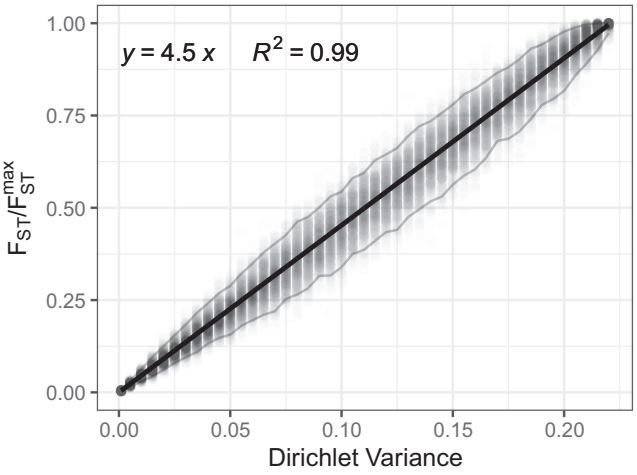

FIGURE 3 Linear relationship between F_{ST}/F_{ST}^{max} and $\mathrm{Var}[q_1] = \mathrm{Var}[q_2]$, the variance across individuals of individual membership coefficients under a Dirichlet distribution. For each of 45 values of $\mathrm{Var}[q_1] = \mathrm{Var}[q_2]$, 500 points are plotted, each representing a random Q matrix with dimensions 50×2. Rows of the Q matrix are simulated using a Dirichlet distribution with means $\lambda = (\lambda_1,\lambda_2) = \left(\frac{2}{3},\frac{1}{3}\right)$ and variances $\mathrm{Var}[q_1] = \mathrm{Var}[q_2] = \lambda_1\lambda_2/(\alpha+1)$, with α chosen to produce variances 0.001,0.005,0.01,0.015, … ,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 $1/(\lambda_1\lambda_2) = 1/\left(\frac{2}{3}\times\frac{1}{3}\right)$, 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

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

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 parametric variance across sampled individuals of membership coefficient q_1 (or q_2). Importantly, the expected value of F_{ST}/F_{ST}^{max} in Equation 2 is independent of the parametric 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.

3.1.3 | 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 parametric 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,d, where they are coloured 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, $\lambda = \left(\frac{2}{3},\frac{1}{3}\right)$ for Figure 4a and $\left(\frac{9}{10},\frac{1}{10}\right)$ for Figure 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 centred 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 $\lambda = \left(\frac{2}{3},\frac{1}{3}\right)$ and in Figure 4c for $\lambda = \left(\frac{9}{10},\frac{1}{10}\right)$. 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. As α increases (membership variability across individuals decreases), the variance of F_{ST}/F_{ST}^{max} across bootstrap samples decreases; this pattern is driven by the fact that the rows of a high-α (low-variability) Q matrix are very similar, so bootstrap-sampled matrices drawn from this matrix will necessarily also be similar to one another.

Comparing Figure 4b with Figure 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 in the analytical result in Equation 2 that under the Dirichlet model, $\mathbb{E}\left[F_{ST}/F_{ST}^{max};\alpha,\lambda,I\right]$ can be written so that it depends on α but not on λ.

In fact, in some cases, matrices simulated with the same value of α but *different* means (λ) are more similar than matrices simulated with both the same α and the *same* means. We tested all $\binom{16}{2}$ pairwise comparisons of the 16 bootstrap distributions in Figure 4 and found that nearly all pairs of distributions were significantly different (Wilcoxon rank-sum tests, $p < 10^{-6}$). Interestingly, the only two pairs that were not significantly different were pairs with the same α but different means: the left-hand α_1 distribution with mean $\left(\frac{2}{3},\frac{1}{3}\right)$ in Figure 4b and the right-hand α_1 distribution with mean $\left(\frac{9}{10},\frac{1}{10}\right)$ in Figure 4c (Wilcoxon rank-sum test, $p = .270$), and the left-hand α_3 distribution with mean $\left(\frac{2}{3},\frac{1}{3}\right)$ in Figure 4b and the left-hand α_3 distribution with mean $\left(\frac{9}{10},\frac{1}{10}\right)$ in Figure 4c (Wilcoxon rank-sum test, $p = .002$). That pairs with the same α and different means can have the same F_{ST}/F_{ST}^{max}, while pairs with different α and either the same or different means have different F_{ST}/F_{ST}^{max} underscores the point that F_{ST}/F_{ST}^{max} can be used to compare the variability of Q matrices with quite different mean membership.

We also observe in Figure 4a,d that the sampling variability of features of Q matrices simulated from the Dirichlet distribution with identical parameters—as reflected in comparisons of pairs of matrices of the same colour within a panel—increases with α. We confirm in Figure S1 that the variability in the mean membership coefficients $(\bar{q}_1,\bar{q}_2)$ of simulated Q matrices increases as the α

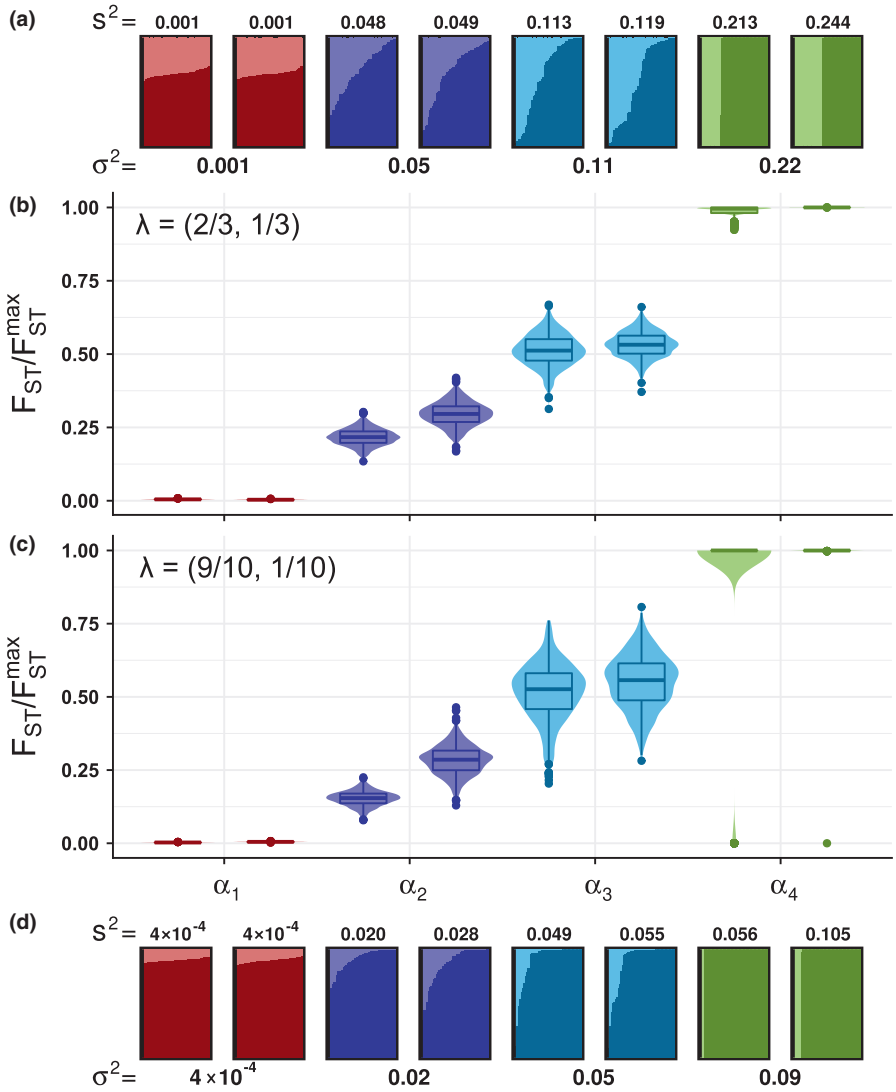

FIGURE 4 Dependence of bootstrap distributions of F_{ST} / F_{ST}^{max} for simulated Q matrices on the Dirichlet variance parameter α, rather than the Dirichlet mean λ. (a, d) Q matrices simulated using specified Dir($\alpha\lambda$) distributions. (b, c) Bootstrap distributions of F_{ST} / F_{ST}^{max} 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 parametric variances: $\alpha_1 = 21901/99$, $\alpha_2 = 341/99$, $\alpha_3 = 101/99$ and $\alpha_4 = 1/99$. Matrices are annotated by associated parametric variances $\sigma^2 = \lambda_1\lambda_2/(\alpha+1)$. In (a), matrices are simulated with parametric mean $\lambda = \left(\frac{2}{3}, \frac{1}{3}\right)$ and are taken from matrices plotted in Figure 3. In (d), matrices are simulated with a more extreme parametric mean, $\lambda = \left(\frac{9}{10}, \frac{1}{10}\right)$. Each vertical bar represents an individual membership coefficient vector (q_1, q_2); the proportion of each bar coloured 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 $\left(0, \frac{2}{9}\right)$ for $\lambda = \left(\frac{2}{3}, \frac{1}{3}\right)$ and in $(0, 0.09)$ for $\lambda = \left(\frac{9}{10}, \frac{1}{10}\right)$. The empirical variance s^2 is computed for each matrix using the sample mean $\bar{\boldsymbol{q}} = \left(\frac{1}{I}\sum_{i=1}^{I} q_1^{(i)}, \frac{1}{I}\sum_{i=1}^{I} q_2^{(i)}\right)$ 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 1000 bootstrap samples of the $I = 50$ individuals in the associated Q matrix

value used to simulate the matrices decreases (i.e. as the Dirichlet variance increases). This increased variability in sampled Q matrix mean memberships leads to increased variability among sampled Q matrix membership variances (Figure S2). Sampling variability can lead Q matrices simulated with the same parameter values 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.

3.2 | 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 nonadmixed 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.

3.2.1 | 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 (Gravel, 2012; Verdu & Rosenberg, 2011). Thus, in examining inferred cluster memberships, admixed populations might be expected to give rise to greater variability in ancestry than nonadmixed 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 nonadmixed source 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, equalling 0.100. Pairs of bootstrap distributions of F_{ST}/F_{ST}^{max} are significantly different ($p < 2\times10^{-16}$ 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 nonadmixed populations.

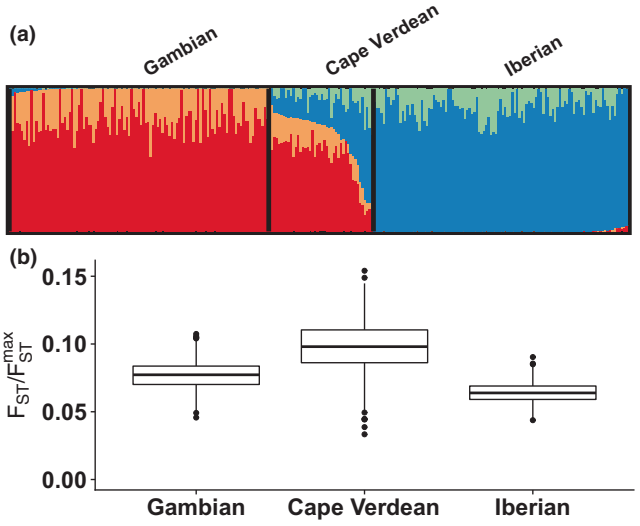

FIGURE 5 Variability of ancestry in admixed and nonadmixed populations. (a) $K = 4$ ADMIXTURE analysis of Gambian ($n = 109$), Cape Verdean ($n = 44$) and Iberian ($n = 107$) samples. Adapted from Verdu et al. (2017). (b) Bootstrap distributions of the ancestry variability measure, F_{ST}/F_{ST}^{max}, for each population (1000 samples)

3.2.2 | 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 ($p < 2\times10^{-9}$ 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 granularity to these claims, suggesting that ancestry variability was steadily increasing during these three periods, with a maximum achieved during Late Antiquity.

3.2.3 | 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 $\left(\frac{1}{K}, \frac{1}{K}, \ldots, \frac{1}{K}\right)$; different individuals possess similar membership coefficients. As the number of loci increases, individuals come to have different membership coefficients, with, for example, individuals from two

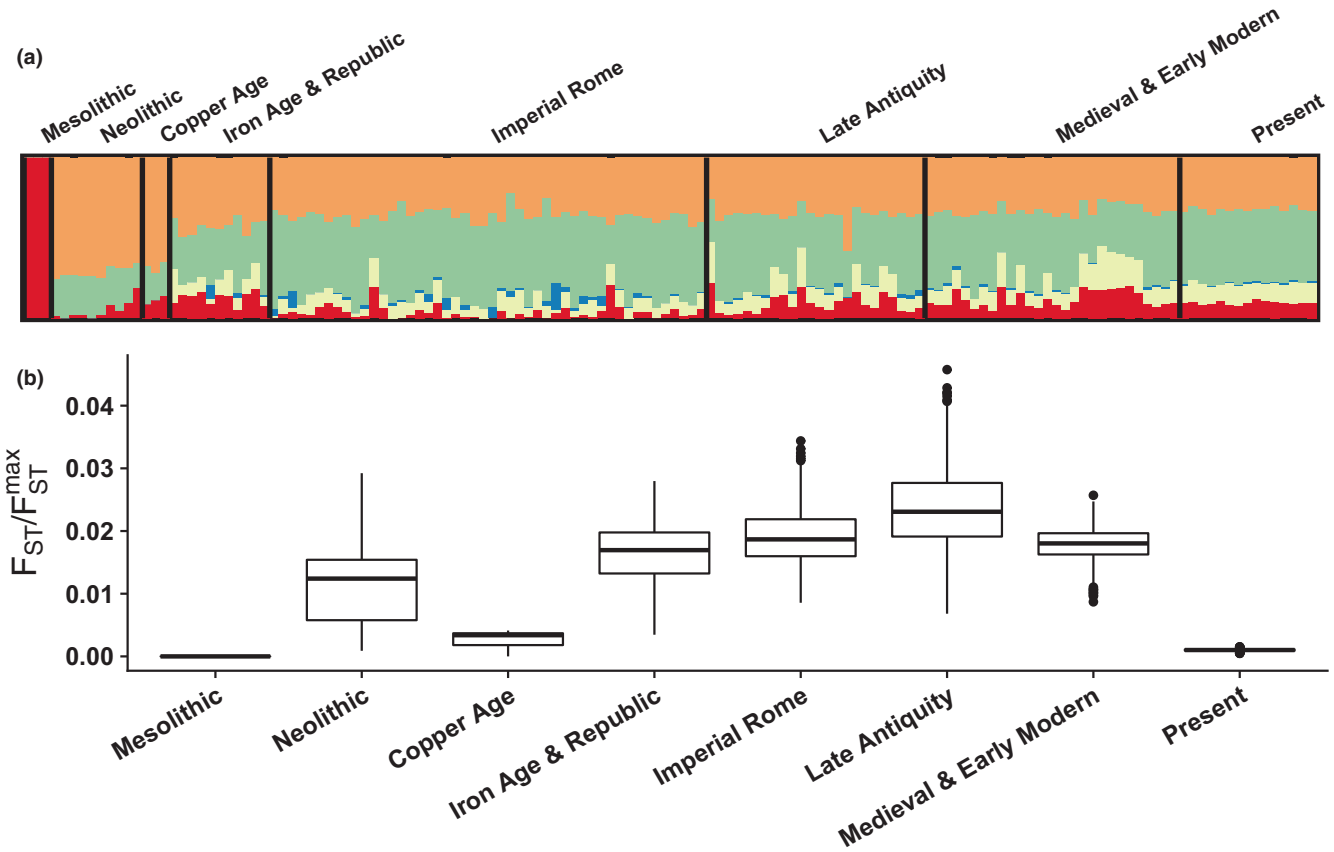

FIGURE 6 Variability of ancestry over time. (a) $K = 5$ STRUCTURE analysis of samples from eight time periods: Mesolithic ($n = 3$), Neolithic ($n = 10$), Copper Age ($n = 3$), Iron Age and Roman Republic ($n = 11$), Imperial Rome ($n = 48$), Late Antiquity ($n = 24$), Medieval and Early Modern ($n = 28$) 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 (1000 samples)

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 ($p = .56$). The pattern of F_{ST} / F_{ST}^{max} values, with the smallest value for Figure 7b, intermediate values for Figure 7a,c, 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,c) 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.

4 | 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

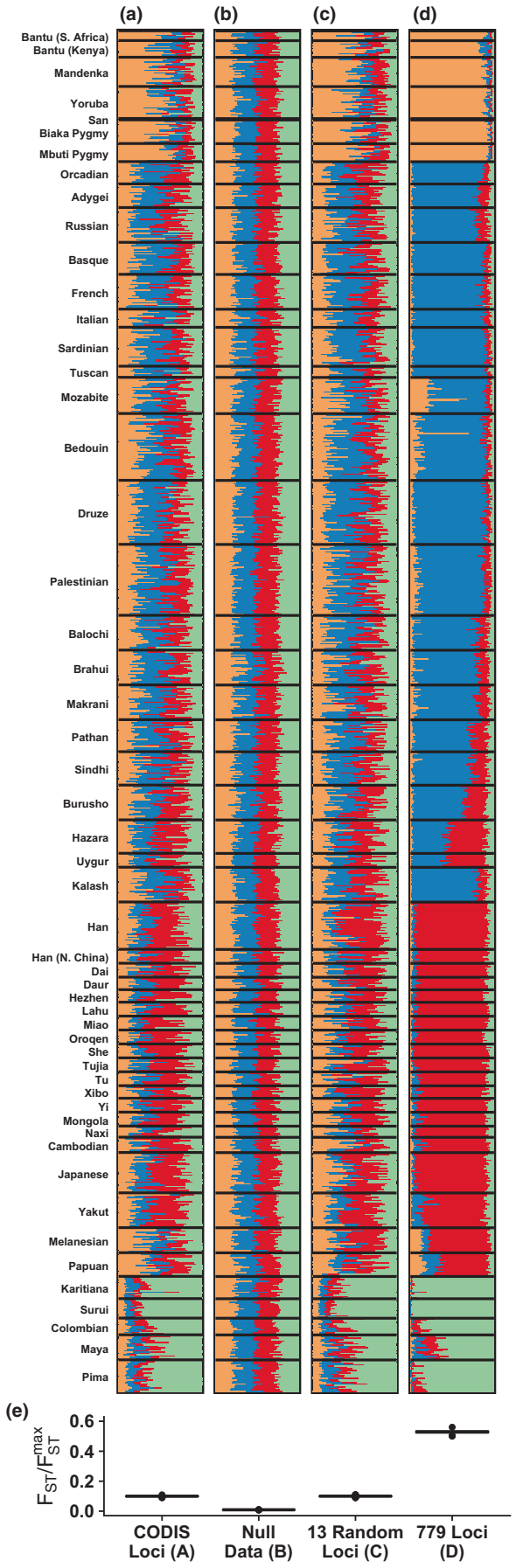

FIGURE 7 Variability of ancestry for analyses with different loci from the same samples. $K = 4$ STRUCTURE analyses of four different sets of loci for a worldwide human sample. Adapted from Algee-Hewitt et al. (2016). (a) 13 CODIS tetranucleotide microsatellite loci. (b) A simulated null data set with 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 data set (1000 samples)

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 the variance of membership coefficients across individuals (Figures 3 and 4); indeed, in a remarkably simple result, we find that it scales approximately linearly with the parametric variance in a model with $K = 2$ ancestral clusters (Equation 2). This result supports the use of F_{ST}/F_{ST}^{max} as a measure of variability in ancestry across individuals. Note that although our analytical result that $\mathbb{E}\left[F_{ST}/F_{ST}^{max}; \alpha, \lambda_1, \lambda_2, I\right] \approx 1/(\alpha+1)$ relies on the case of $K = 2$ ancestral clusters, additional simulations with larger K suggest that similar results hold for larger K, as such simulations find that the mean F_{ST}/F_{ST}^{max} values across simulated Q matrices with fixed parameter values match $1/(\alpha+1)$, irrespective of the value of K (Figure S3).

We have proposed that the F_{ST}/F_{ST}^{max} measure can be used in a statistical test of the equality of ancestry variability between two Q matrices by generating bootstrap samples of the individuals in each Q matrix, computing F_{ST}/F_{ST}^{max} for each bootstrap-sampled matrix and comparing bootstrap distributions of F_{ST}/F_{ST}^{max} using a Wilcoxon rank-sum test. In analysing our simulated and empirical data, this test performed appropriately. It distinguished between matrices with meaningfully distinct variabilities, such as between matrices simulated with different Dirichlet α parameter values (Figure 4). It notably failed to find a significant difference in a case where the true variabilities of the Q matrices were similar, with the Q matrices representing ancestry inferred using two sets of 13 loci (Figure 7). To further support the use of this bootstrap test, we include supplementary figures that demonstrate that under the null hypothesis, p-values for the test have the appropriate uniform distribution; this result is seen in simulations that consider different numbers of bootstrap replicates (Figure S4), different numbers of clusters (Figure S5) and different numbers of individuals (Figure S6).

The expected value of F_{ST}/F_{ST}^{max} behaves sensibly as the number of individuals, I, increases (Figure S7). In particular, simulated values of $\mathbb{E}\left[F_{ST}/F_{ST}^{max}; \alpha, \lambda, I\right]$ remain constant with I: as the number of simulated individuals increases at a fixed variability of

membership, the mean F_{ST}/F_{ST}^{max} across simulations remains the same and the variance of F_{ST}/F_{ST}^{max} decreases. More generally, we have seen that F_{ST}/F_{ST}^{max} does not depend on the mean membership of the Q matrices under analysis, which makes it well suited to comparing the ancestry variabilities of populations with different mean memberships. To clarify, the test of equality of F_{ST}/F_{ST}^{max} values cannot be used to assess the equality of mean membership among Q matrices—it compares their variability, not their mean membership.

We demonstrated the use of the F_{ST}/F_{ST}^{max} measure in data sets exemplifying three scenarios in which ancestry variability is of particular interest. In a comparison of ancestry measured in admixed and nonadmixed 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 nonadmixed populations (Figure 5). In a comparison of ancestries measured in different time 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 F_{ST}/F_{ST}^{max} measure depends on Q matrices, limitations of the methods used to generate the Q matrices extend to its calculation. For example, if individuals were mislabelled 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 analysing the level of support observed for different numbers of clusters K (Alexander & Lange, 2011; Evanno et al., 2005) and methods of aligning the clustering solutions observed in replicate analyses (Behr et al., 2016; Jakobsson & Rosenberg, 2007; Kopelman et al., 2015), as well as software for graphical display (Ramasamy et al., 2014; Rosenberg, 2004) and for managing files and workflows associated with the analysis (Earl & VonHoldt, 2012; Francis, 2017).

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,...,0) and minimal if all individuals possess membership vector $\left(\frac{1}{K}, \frac{1}{K}, \cdots, \frac{1}{K}\right)$. 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 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,d, 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 interindividual patterns.

We note that in addition to analysing 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).

AUTHOR CONTRIBUTIONS

MLM, NA and NAR designed the study and performed the theoretical analysis. MLM conducted the simulations, analysed the data and wrote the software. NAR supervised the study. All authors wrote the manuscript.

ACKNOWLEDGEMENTS

We acknowledge support from NIH grant R01 HG005855 and NSF grant BCS-2116322. 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.

CONFLICT OF INTEREST

The authors have no conflicts of interest to report.

DATA AVAILABILITY STATEMENT

The FSTruct R package is available for download from https://github.com/MaikeMorrison/FSTruct. The introductory vignette is linked from the package README file and provides a guide to use of the package. The *Q* matrices visualized in Figures 5–7 are available as supplemental files.

ORCID

Maike L. Morrison https://orcid.org/0000-0003-0430-1401
Nicolas Alcala https://orcid.org/0000-0002-5961-5064
Noah A. Rosenberg https://orcid.org/0000-0002-1829-8664

REFERENCES

Alcala, N., & Rosenberg, N. A. (2017). Mathematical constraints on F_{ST}: Biallelic markers in arbitrarily many populations. *Genetics*, *206*, 1581–1600.

Alcala, N., & Rosenberg, N. A. (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., & Rosenberg, N. A. (2022). Mathematical constraints on F_{ST}: Multiallelic markers in arbitrarily many populations. *Philosophical Transactions of the Royal Society B*, *377*, 20200414.

Alexander, D. H., & Lange, K. (2011). Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics*, *12*, 246.

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*, 1655–1664.

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A. J. R., Behjati, S., Biankin, A. V., Bignell, G. R., Bolli, N., Borg, A., Børresen-Dale, A. L., Boyault, S., Burkhardt, B., Butler, A. P., Caldas, C., Davies, H. R., Desmedt, C., Eils, R., Eyfjörd, J. E., Foekens, J. A., … Stratton, M. R. (2013). Signatures of mutational processes in human cancer. *Nature*, *500*, 415–421.

Algee-Hewitt, B. F., Edge, M. D., Kim, J., Li, J. Z., & Rosenberg, N. A. (2016). Individual identifiability predicts population identifiability in forensic microsatellite markers. *Current Biology*, *26*, 935–942.

Antonio, M. L., Gao, Z., Moots, H. M., Lucci, M., Candilio, F., Sawyer, S., Oberreiter, V., Calderon, D., Devitofranceschi, K., Aikens, R. C., Aneli, S., Bartoli, F., Bedini, A., Cheronet, O., Cotter, D. J., Fernandes, D. M., Gasperetti, G., Grifoni, R., Guidi, A., … Pritchard, J. K. (2019). Ancient Rome: A genetic crossroads of Europe and the Mediterranean. *Science*, *366*, 708–714.

Behr, A. A., Liu, K. Z., Liu-Fang, G., Nakka, P., & Ramachandran, S. (2016). Pong: Fast analysis and visualization of latent clusters in population genetic data. *Bioinformatics*, *32*, 2817–2823.

Corander, J., Marttinen, P., Sirén, J., & Tang, J. (2008). Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. *BMC Bioinformatics*, *9*, 539.

Corander, J., Waldmann, P., Marttinen, P., & Sillanpää, M. J. (2004). BAPS 2: Enhanced possibilities for the analysis of genetic population structure. *Bioinformatics*, *20*, 2363–2369.

Earl, D. A., & VonHoldt, B. M. (2012). STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, *4*, 359–361.

Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. *Molecular Ecology*, *14*, 2611–2620.

Falush, D., Stephens, M., & Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics*, *164*, 1567–1587.

Falush, D., Stephens, M., & Pritchard, J. K. (2007). Inference of population structure using multilocus genotype data: Dominant markers and null alleles. *Molecular Ecology Notes*, *7*, 574–578.

Francis, R. M. (2017). Pophelper: An R package and web app to analyse and visualize population structure. *Molecular Ecology Resources*, *17*, 27–32.

Gravel, S. (2012). Population genetics models of local ancestry. *Genetics*, *191*, 607–619.

Grimmett, G. R., & Stirzaker, D. R. (2001a). *One thousand exercises in probability*. Oxford University Press.

Grimmett, G. R., & Stirzaker, D. R. (2001b). *Probability and random processes* (3rd ed.). Oxford University Press.

Guillot, G., & Orlando, L. (2017). Population structure. *Oxford Bibliographies*. https://doi.org/10.1093/obo/9780199941728-0057

Hubisz, M. J., Falush, D., Stephens, M., & Pritchard, J. K. (2009). Inferring weak population structure with the assistance of sample group information. *Molecular Ecology Resources*, *9*, 1322–1332.

Huelsenbeck, J. P., & Andolfatto, P. (2007). Inference of population structure under a Dirichlet process model. *Genetics*, *175*, 1787–1802.

Jakobsson, M., Edge, M. D., & Rosenberg, N. A. (2013). The relationship between F_{ST} and the frequency of the most frequent allele. *Genetics*, *193*, 515–528.

Jakobsson, M., & Rosenberg, N. A. (2007). CLUMPP: A cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, *23*, 1801–1806.

Kerminen, S., Cerioli, N., Pacauskas, D., Havulinna, A. S., Perola, M., Jousilahti, P., Salomaa, V., Daly, M. J., Vyas, R., Ripatti, S., & Pirinen, M. (2021). Changes in the fine-scale genetic structure of Finland through the 20th century. *PLoS Genetics*, *17*, e1009347.

Kopelman, N. M., Mayzel, J., Jakobsson, M., Rosenberg, N. A., & Mayrose, I. (2015). CLUMPAK: A program for identifying clustering modes and packaging population structure inference across K. *Molecular Ecology Resources*, *15*, 1179–1191.

Kotz, S., Balakrishnan, N., & Johnson, N. (2000). *Continuous multivariate distributions, volume 1: Models and applications* (2nd ed.). Wiley.

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

Rahbari, R., Wuster, A., Lindsay, S. J., Hardwick, R. J., Alexandrov, L. B., Turki, S. A., Dominiczak, A., Morris, A., Porteous, D., Smith, B., Stratton, M. R., UK10K Consortium, & Hurles, M. E. (2016). Timing, rates and spectra of human germline mutation. *Nature Genetics*, *48*, 126–133.

Ramasamy, R. K., Ramasamy, S., Bindroo, B. B., & Naik, V. G. (2014). STRUCTURE PLOT: A program for drawing elegant STRUCTURE bar plots in user friendly interface. *Springerplus*, *3*, 431.

Rosenberg, N. A. (2004). DISTRUCT: A program for the graphical display of population structure. *Molecular Ecology Notes*, *4*, 137–138.

Rosenberg, N. A., Mahajan, S., Ramachandran, S., Zhao, C., Pritchard, J. K., & Feldman, M. W. (2005). Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genetics*, *1*, 660–671.

Serfling, R. (1980). *Approximation theorems of mathematical statistics*. Wiley.

van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge University Press.

Verdu, P., Jewett, E. M., Pemberton, T. J., Rosenberg, N. A., & Baptista, M. (2017). Parallel trajectories of genetic and linguistic admixture in a genetically admixed creole population. *Current Biology*, *27*, 2529–2535.

Verdu, P., & Rosenberg, N. A. (2011). A general mechanistic model for admixture histories of hybrid populations. *Genetics*, *189*, 1413–1426.

Wang, X., Park, J., Susztak, K., Zhang, N. R., & Li, M. (2019). Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nature Communications*, *10*, 380.

How to cite this article: Morrison, M. L., Alcala, N., & Rosenberg, N. A. (2022). FSTruct: An F_{ST}-based tool for measuring ancestry variation in inference of population structure. *Molecular Ecology Resources*, *22*, 2614–2626. https://doi.org/10.1111/1755-0998.13647

SUPPORTING INFORMATION

Additional supporting information can be found in the online version of this article.

APPENDIX A

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

OVERVIEW

To obtain the expectation $\mathbb{E}\left[F_{ST}/F_{ST}^{max}\right]$, we first sample I independent and identically distributed $\mathrm{Dir}(\alpha\lambda_1, \alpha\lambda_2)$ random variables $\left(q_1^{(i)}, q_2^{(i)}\right)$, where $q_1^{(i)}$ is the membership coefficient of individual i 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 $I \to \infty$, 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, $M \overset{a.s.}{\to} \lambda_1$ as $I \to \infty$.

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

$$f\left(q_1^{(i)}, q_2^{(i)}\right) = \frac{\left(q_1^{(i)}\right)^{\alpha\lambda_1 - 1}\left(q_2^{(i)}\right)^{\alpha\lambda_2 - 1}}{\mathrm{B}(\alpha\lambda_1, \alpha\lambda_2)}, \tag{A1}$$

where

$$B(\alpha\lambda_1, \alpha\lambda_2) = \frac{\Gamma(\alpha\lambda_1)\Gamma(\alpha\lambda_2)}{\Gamma(\alpha\lambda_1 + \alpha\lambda_2)} = \frac{\Gamma(\alpha\lambda_1)\Gamma(\alpha\lambda_2)}{\Gamma(\alpha)}, \tag{A2}$$

and Γ is the gamma function.

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

$$\mathbb{E}\left[\frac{F_{ST}}{F_{ST}^{max}}; \alpha, \lambda_1, \lambda_2, I\right] = \int_{q_1^{(1)}=0}^{1}\int_{q_1^{(2)}=0}^{1}\cdots\int_{q_1^{(I)}=0}^{1} \frac{F_{ST}}{F_{ST}^{max}}\left(q_1^{(1)}, q_1^{(2)}, \ldots, q_1^{(I)}\right) \times f\left(q_1^{(1)}, q_2^{(1)}\right) f\left(q_1^{(2)}, q_2^{(2)}\right)\cdots f\left(q_1^{(I)}, q_2^{(I)}\right) dq_1^{(1)} dq_1^{(2)} \cdots dq_1^{(I)}. \tag{A3}$$

With this expression in hand, we proceed by writing the expression for F_{ST}/F_{ST}^{max} in terms of the membership coefficients $\left(q_1^{(i)}, q_2^{(i)}\right)$ 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 $\left(q_1^{(1)}, q_2^{(1)}\right)$, $\ldots$, $\left(q_1^{(I)}, q_2^{(I)}\right)$ can be written using Equations 3 and 5 of Alcala and Rosenberg (2017),

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

We obtain

$$\frac{F_{ST}\left(q_1^{(1)} q_1^{(2)} \cdots q_1^{(I)}\right)}{F_{ST}^{max}\left(q_1^{(1)} q_1^{(2)} \cdots q_1^{(I)}\right)} = \frac{\sum_{i=1}^{I}\left(q_1^{(i)}\right)^2 - IM^2}{\lfloor IM \rfloor + \{IM\}^2 - IM^2}. \tag{A4}$$

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 Equation A4. Because $\lfloor IM \rfloor = IM - \{IM\}$, $\lfloor IM \rfloor + \{IM\}^2 = IM - \left(\{IM\} - \{IM\}^2\right) = IM - \delta$ where the error term $\delta = \{IM\}(1 - \{IM\})$ lies in $\left[0, \frac{1}{4}\right]$, 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 Equation A4, obtaining

$$\frac{F_{ST}}{F_{ST}^{max}}\left(q_1^{(1)} q_1^{(2)} \cdots q_1^{(I)}\right) \approx \frac{\sum_{i=1}^{I}\left(q_1^{(i)}\right)^2 - IM^2}{IM(1-M)}. \tag{A5}$$

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

To find an approximation for $\mathbb{E}\left[F_{ST}/F_{ST}^{\max}\right]$, it is convenient to make a further approximation in Equation A5, substituting M with λ_1. We justify this substitution by proving that as $I\to\infty$,

$$\frac{\sum_{i=1}^{I}\left(q_1^{(i)}\right)^2-IM^2}{IM(1-M)}\overset{a.s.}{\to}\frac{\sum_{i=1}^{I}\left(q_1^{(i)}\right)^2-I\lambda_1^2}{I\lambda_1(1-\lambda_1)}. \tag{A6}$$

Subtracting the right-hand side from the left-hand side, proving Equation A6 is equivalent to proving

$$\left[\frac{1}{I}\sum_{i=1}^{I}\left(q_1^{(i)}\right)^2\right]\left[\frac{1}{M(1-M)}-\frac{1}{\lambda_1(1-\lambda_1)}\right]+\left[\frac{-M}{1-M}+\frac{\lambda_1}{1-\lambda_1}\right]\overset{a.s.}{\to}0. \tag{A7}$$

If $X_n\overset{a.s.}{\to}X$ and $Y_n\overset{a.s.}{\to}Y$, then $X_n+Y_n\overset{a.s.}{\to}X+Y$ (Grimmett & Stirzaker, 2001a, p. 336, exercise 2; Grimmett & 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 Equation A7.

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 $X_n\overset{a.s.}{\to}X$, then $g(X_n)\overset{a.s.}{\to}g(X)$ (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 $I\to\infty$, $M\overset{a.s.}{\to}\lambda_1$. It follows that as $I\to\infty$, and $M\overset{a.s.}{\to}\lambda_1$, $g(M)\overset{a.s.}{\to}g(\lambda_1)$; that is, $-M/(1-M)+\lambda_1/(1-\lambda_1)\overset{a.s.}{\to}0$.

For the left-hand term, the factor $1/\left[M(1-M)\right]-1/\left[\lambda_1(1-\lambda_1)\right]$ converges almost surely to 0 by the continuous mapping theorem with $g(x)=1/\left[x(1-x)\right]-1/\left[\lambda_1(1-\lambda_1)\right]$. By the strong law of large numbers, $\frac{1}{I}\sum_{i=1}^{I}\left(q_1^{(i)}\right)^2$ converges almost surely to $\gamma=\mathbb{E}\left[\left(q_1^{(i)}\right)^2\right]=\mathrm{Var}\left[q_1^{(i)}\right]+\lambda_1^2=\lambda_1(1-\lambda_1)/(\alpha+1)+\lambda_1^2=\lambda_1(\alpha\lambda_1+1)/(\alpha+1)$ (Serfling, 1980, Theorem B). If $X_n\overset{a.s.}{\to}X$ and $Y_n\overset{a.s.}{\to}Y$, then $X_nY_n\overset{a.s.}{\to}XY$ (Grimmett & Stirzaker, 2001a, p. 336, exercise 2; Grimmett & Stirzaker, 2001b, p. 354, exercise 2), so that the left-hand term of Equation A7 converges almost surely to $\gamma\times0=0$.

EVALUATING $\mathbb{E}\left[F_{ST}/F_{ST}^{\max}\right]$ UNDER THE DIRICHLET MODEL

Inserting our expression for $f\left(q_1^{(i)},q_2^{(i)}\right)$ from Equation A1 and our expression for $F_{ST}/F_{ST}^{\max}$ from Equation A6 into Equation A3 allows us to write an approximate expression for the expectation of $F_{ST}/F_{ST}^{\max}$ given the parameters of the Dirichlet distribution and the sample size I:

$$\mathbb{E}\left[\frac{F_{ST}}{F_{ST}^{\max}};\alpha,\lambda_1,\lambda_2,I\right]\approx\int_{q_1^{(1)}=0}^{1}\int_{q_1^{(2)}=0}^{1}\cdots\int_{q_1^{(I)}=0}^{1}\frac{\sum_{i=1}^{I}\left(q_1^{(i)}\right)^2-I\lambda_1^2}{I\lambda_1(1-\lambda_1)}\times\frac{\left(q_1^{(1)}\right)^{\alpha\lambda_1-1}\left(q_2^{(1)}\right)^{\alpha\lambda_2-1}}{B(\alpha\lambda_1,\alpha\lambda_2)}\times\frac{\left(q_1^{(2)}\right)^{\alpha\lambda_1-1}\left(q_2^{(2)}\right)^{\alpha\lambda_2-1}}{B(\alpha\lambda_1,\alpha\lambda_2)}\times\cdots\times\frac{\left(q_1^{(I)}\right)^{\alpha\lambda_1-1}\left(q_2^{(I)}\right)^{\alpha\lambda_2-1}}{B(\alpha\lambda_1,\alpha\lambda_2)}dq_1^{(1)}dq_1^{(2)}\cdots dq_1^{(I)}. \tag{A8}$$

Examining the quantity $\sum_{i=1}^{I}\left(q_1^{(i)}\right)^2-I\lambda_1^2$, we observe that Equation A8 can be decomposed as a sum of $I+1$ terms, one for each of the $\left(q_1^{(i)}\right)^2$ terms, and one for the $-I\lambda_1^2$ term. Assign the first I of these separate terms the labels $L_1,L_2,\ldots,L_I$ and the $-I\lambda_1^2$ term the label L_*, so that $\mathbb{E}\left[F_{ST}/F_{ST}^{\max};\alpha,\lambda_1,\lambda_2,I\right]\approx L_1+L_2+\cdots+L_I+L_*$.

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

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

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

$$L_*=-\frac{I\lambda_1^2}{I\lambda_1(1-\lambda_1)}=-\frac{\lambda_1}{1-\lambda_1}. \tag{A9}$$

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

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

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,

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

The remaining integral can be evaluated by noting that $\int_0^1x^{a-1}(1-x)^{b-1}dx=B(a,b)$. 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 Equation A2 and the property of gamma functions $\Gamma(z+1)=z\Gamma(z)$, this expression simplifies to

$$L_i=\frac{\alpha\lambda_1+1}{I(1-\lambda_1)(\alpha+1)}. \tag{A10}$$

We now combine Equations A9 and A10 to complete the calculation in Equation A8, noting that L_i does not depend on i, so that each L_i follows Equation A10.

$$\begin{aligned}\mathbb{E}\left[\frac{F_{ST}}{F_{ST}^{\max}};\alpha,\lambda_1,\lambda_2,I\right]&\approx L_1+L_2+\cdots+L_I+L_*\\&=\frac{I(\alpha\lambda_1+1)}{I(1-\lambda_1)(\alpha+1)}-\frac{\lambda_1}{1-\lambda_1}\\&=\frac{1}{\alpha+1}.\end{aligned} \tag{A11}$$