$F_{\mathbf{ST}}$ and kinship for arbitrary population structures II: Method of moments estimators

Alejandro Ochoa[1,2] and John D. Storey[1,2,*]

[1]Lewis-Sigler Institute for Integrative Genomics, and [2]Center for Statistics and Machine Learning, Princeton University, Princeton, NJ 08544, USA

∗ Corresponding author: jstorey@princeton.edu

Abstract: F_{ST} and kinship are key parameters often estimated in modern population genetics studies. Kinship matrices have also become a fundamental quantity used in genome-wide association studies and heritability estimation. The most frequently used estimators of F_{ST} and kinship are method of moments estimators whose accuracies depend strongly on the existence of simple underlying forms of structure, such as the island model of non-overlapping, independently evolving subpopulations. However, modern data sets have revealed that these simple models of structure do not likely hold in many populations, including humans. In this work, we provide new results on the behavior of these estimators in the presence of arbitrarily complex population structures. After establishing a framework for assessing bias and consistency of genome-wide estimators, we calculate the accuracy of F_{ST} and kinship estimators under arbitrary population structures, characterizing biases and estimation challenges unobserved under their originally assumed models of structure. We illustrate our results using simulated genotypes from an admixture model, constructing a one-dimensional geographic scenario that departs nontrivially from the island model. Using 1000 Genomes Project data, we verify that population-level pairwise F_{ST} estimates underestimate differentiation measured by an individual-level pairwise F_{ST} estimator introduced here. We show that the calculated biases are due to unknown quantities that cannot be estimated under the established frameworks, highlighting the need for innovative estimation approaches in complex populations. We provide initial results that point towards a future estimation framework for generalized F_{ST} and kinship.

1 Introduction

In population genetics studies, one is often interested in characterizing structure, genetic differentiation, and relatedness among individuals. Two quantities often considered in this context are $F_{\rm ST}$ and kinship. $F_{\rm ST}$ is a parameter that measures structure in a subdivided population, satisfying $F_{\rm ST} = 0$ for an unstructured population and $F_{\rm ST} = 1$ if every SNP has fixated in every subpopulation. More specifically, $F_{\rm ST}$ is the probability that alleles drawn randomly from a subpopulation are "identical by descent" (IBD) relative to an ancestral population [1, 2]. The kinship coefficient is a measure of relatedness between individuals defined in terms of IBD probabilities, and it is closely related to $F_{\rm ST}$ [1].

The most frequently used $F_{\rm ST}$ estimators are derived and justified under the "island model" assumption, in which subpopulations are non-overlapping and have evolved independently from a common ancestral population. The Weir-Cockerham (WC) $F_{\rm ST}$ estimator assumes islands of differing sample sizes and equal $F_{\rm ST}$ per island [3]. The "Hudson" $F_{\rm ST}$ estimator assumes two islands with different $F_{\rm ST}$ values [4]. These $F_{\rm ST}$ estimators are ratio estimators derived using the method of moments to have unbiased numerators and denominators, which gives approximately unbiased ratio estimates [3–5], and they are important contributions used widely in the field.

Kinship coefficients are now commonly calculated in population genetics studies to capture structure and relatedness. They are utilized in principal components analyses and linear-mixed effects models to correct for structure in Genome-Wide Association Studies (GWAS) and to estimate genome-wide heritability [6–15]. The most commonly used kinship estimator for genotype data [9, 10, 13–18] is also a method of moments estimator whose operating characteristics are largely unknown in the presence of structure. As we show here, the required assumption for this popular estimator to be accurate is that the average kinship be zero, which implies that the population must be unstructured.

Recent genome-wide studies have revealed that humans and other natural populations are structured in a complex manner that violate the assumptions of the above estimators. This has been observed in several large human studies, such as the Human Genome Diversity Project [19], the 1000 Genomes Project [20], and other contemporary [21, 22] and archaic populations [23, 24]. Therefore, there is a need for innovative approaches designed for complex population structures. To this end, we reveal the operating characteristics of these frequently used $F_{\rm ST}$ and kinship estimators in the presence of arbitrary forms of structure with the goal of identifying new estimation strategies for $F_{\rm ST}$ and kinship.

We generalized the definition of $F_{\rm ST}$ for arbitrary population structures in the first paper in this series [25]. Additionally, we derived connections between $F_{\rm ST}$ and three models: arbitrary kinship coefficients [1, 26], individual-specific allele frequencies [27, 28], and admixture models [29–31]. Here, we study existing $F_{\rm ST}$ and kinship method of moments estimators in models that allow for arbitrary population structures (see Fig. 1 for an overview of the results). First, we obtain new

strong convergence results for a family of ratio estimators that includes F_{ST} and kinship estimators. Next, we calculate the convergence values of these estimators under arbitrary population structures, where we find biases that are not present under their original assumptions about structure. We characterize the limit of the standard kinship estimator for the first time, identifying complex biases or distortions that have not been described before. We construct an admixture model, which represents a form of structure distinct from the island model, to illustrate our theoretical findings through simulation. We analyze 1000 Genomes Project populations to illustrate their non-island nature, and measure differentiation that is missed by the Hudson F_{ST} estimator. We identify a new direction for estimating F_{ST} and kinship in a nearly unbiased fashion, which is the topic of our next paper in this series [32].

2 Models and definitions

Here we summarize new arbitrary population structure models, definitions, and results presented in detail in the first paper in this series [25] (Fig. 1). We assume a complete matrix of m SNPs and n individuals. We concentrate on biallelic genotypes x_{ij} for SNP i and individual j, encoded as the number of reference alleles: $x_{ij} = 2$ is homozygous for the reference allele, $x_{ij} = 0$ is homozygous for the alternative allele, and $x_{ij} = 1$ is heterozygous. We assume the existence of a panmictic ancestral population T characterized by ancestral reference allele frequencies $p_i^T \in (0, 1)$ for every SNP i.

2.1 The kinship model and the generalized F_{ST}

Under the kinship model, individuals receive their alleles as determined by their inbreeding and kinship coefficients. The inbreeding coefficient f_j^T of j is the probability that two alleles at a random SNP of individual j are IBD [33]. Similarly, the kinship coefficient φ_{jk}^T of j and k is the probability that two alleles chosen at random from each individual and at a random SNP are IBD [1]. The ancestral population T determines what is IBD: only relationships since T count toward IBD. The first two moments of the genotypes are

$$\mathrm{E}[x_{ij}|T] = 2p_i^T, \tag{1}$$

$$\mathrm{Cov}(x_{ij}, x_{ik}|T) = 4p_i^T \left(1 - p_i^T\right) \varphi_{jk}^T, \tag{2}$$

where self-kinship is $\varphi_{jj}^T = \frac{1}{2}\left(1 + f_j^T\right)$[1, 2, 26, 33]. Lastly, if S is a panmictic population that evolved from T, then f_S^T is the value of f_j^T shared by all individuals j in S relative to T, and equals Wright's F_{ST} for this subdivided population [2].

The generalized F_{ST} definition that we proposed [25] requires the notion of *local populations*, needed to mirror at the individual level Wright's distinction between structural inbreeding due to the population structure from local inbreeding [2]. The local population L_j of individual j is the

most recent ancestral population of j [25]. Similarly, the *jointly local population* L_{jk} of a pair of individuals j and k is the most recent ancestral population shared by j and k, which is ancestral to both L_j and L_k [25]. For T ancestral to L_j or L_{jk}, as needed, we have three parameter pairs: "total" (f_j^T, φ_{jk}^T), "local" $(f_j^{L_j}, \varphi_{jk}^{L_{jk}})$, and "structural" $(f_{L_j}^T, f_{L_{jk}}^T)$ kinship and inbreeding coefficients, related by [25]

$$\begin{aligned} f_j^T &= f_{L_j}^T + f_j^{L_j}\left(1 - f_{L_j}^T\right), \\ \varphi_{jk}^T &= f_{L_{jk}}^T + \varphi_{jk}^{L_{jk}}\left(1 - f_{L_{jk}}^T\right). \end{aligned} \tag{3}$$

A *locally outbred* individual has $f_j^{L_j} = 0$ and therefore $f_j^T = f_{L_j}^T$. Similarly, a pair of *locally unrelated* individuals have $\varphi_{jk}^{L_{jk}} = 0$ and therefore $\varphi_{jk}^T = f_{L_{jk}}^T$. The generalized F_{ST} is given by

$$F_{\mathrm{ST}} = \sum_{j=1}^{n} w_j f_{L_j}^T, \tag{4}$$

where $w_j > 0, \sum_{j=1}^n w_j = 1$ are weights chosen to capture the sampling procedure of individuals [25]. The individual-level pairwise F_{ST} is the special case of Eq. (4) for $n = 2$ individuals, given by

$$F_{jk} = \frac{f_{L_j}^{L_{jk}} + f_{L_k}^{L_{jk}}}{2} = \frac{\frac{f_{L_j}^T + f_{L_k}^T}{2} - f_{L_{jk}}^T}{1 - f_{L_{jk}}^T}, \tag{5}$$

where the second equality holds for any T ancestral to L_{jk} [25].

2.2 The coancestry model for individual-specific allele frequencies

Previous F_{ST} estimators are often in terms of population allele frequencies [2–5]. Our earlier proposed coancestry model [25] extends previous models [5, 34] of population allele frequencies to individuals. The *individual-specific allele frequency* (IAF) is denoted $\pi_{ij} \in [0, 1]$ for SNP i and individual j [27, 28]. In our model, IAFs are random variables drawn from T according to the population structure, with covariances between individuals j and k parametrized by the *individual-specific coancestry coefficients* θ_{jk}^T. We assume that the IAF moments and genotypes are drawn as

$$\mathrm{E}[\pi_{ij}|T] = p_i^T, \tag{6}$$

$$\mathrm{Cov}(\pi_{ij}, \pi_{ik}|T) = p_i^T\left(1 - p_i^T\right)\theta_{jk}^T, \tag{7}$$

$$x_{ij}|\pi_{ij} \sim \mathrm{Binomial}(2, \pi_{ij}). \tag{8}$$

We derived the following correspondence between coancestry and kinship coefficients by marginalizing π_{ij} from this model and comparing to Eqs. (1) and (2) [25]:

$$\theta_{jk}^T = \begin{cases} f_j^T & \text{if} \quad j = k, \\ \varphi_{jk}^T & \text{if} \quad j \neq k. \end{cases} \tag{9}$$

For this reason, and the similarities between Eqs. (1) and (2) and Eqs. (6) and (7), estimators based on genotypes can be readily restated in terms of IAFs, and viceversa. Due to Eq. (8), individuals in the coancestry model are locally outbred and unrelated, a key difference from the more general kinship model, so $f^T_{L_j} = \theta^T_{jj}$ and $f^T_{L_{jk}} = \theta^T_{jk}$ for $j \neq k$ also hold. Therefore, F_{ST} in this model equals

$$F_{\mathrm{ST}} = \sum_{j=1}^{n} w_j \theta^T_{jj}. \tag{10}$$

3 Assessing the accuracy of genome-wide estimators

Many F_{ST} and kinship coefficient method of moments estimators are "ratio estimators", a class that tends to be biased and have no closed form expectation [35]. In the literature, the expectation of a ratio is frequently approximated with a ratio of expectations [3–5]. Specifically, estimators are often called "unbiased" if the ratio of expectations is unbiased, even though the ration estimator itself may be biased. Here we characterize the behavior of two ratio estimator families calculated from genome-wide data, detailing conditions where this previous approximation is justified and providing additional criteria to assess the accuracy of such estimators.

The general problem involves random variables a_i and b_i calculated from genotypes at each SNP i, such that $\mathrm{E}[a_i] = Ac_i$ and $\mathrm{E}[b_i] = Bc_i$ and the goal is to estimate $\frac{A}{B}$. A and B are constants shared across SNPs (given by F_{ST} or φ^T_{jk}), while c_i depends on the ancestral allele frequency p^T_i and varies per SNP. The problem is that the single SNP estimator $\frac{a_i}{b_i}$ is biased, since $\mathrm{E}\left[\frac{a_i}{b_i}\right] \neq \frac{\mathrm{E}[a_i]}{\mathrm{E}[b_i]} = \frac{A}{B}$ [35]. Below we study two estimator families that combine SNPs to better estimate $\frac{A}{B}$.

The solution we recommend is the "ratio-of-means" estimator $\frac{\hat{A}_m}{\hat{B}_m}$, where $\hat{A}_m = \frac{1}{m}\sum_{i=1}^{m} a_i$, and $\hat{B}_m = \frac{1}{m}\sum_{i=1}^{m} b_i$, which is common for F_{ST} estimators [3–5]. Note that $\mathrm{E}\left[\hat{A}_m\right] = A\bar{c}_m$ and $\mathrm{E}\left[\hat{B}_m\right] = B\bar{c}_m$, where $\bar{c}_m = \frac{1}{m}\sum_{i=1}^{m} c_i$. We will assume bounded terms ($|a_i|, |b_i| \leq C$ for some finite C), a convergent $\bar{c}_m \to c$, and $Bc \neq 0$, which are satisfied by common estimators. Given independent SNPs, we prove almost sure convergence to the desired quantity (Appendix A.1),

$$\frac{\hat{A}_m}{\hat{B}_m} = \frac{\frac{1}{m}\sum_{i=1}^{m} a_i}{\frac{1}{m}\sum_{i=1}^{m} b_i} \xrightarrow[m\to\infty]{\text{a.s.}} \frac{A}{B}, \tag{11}$$

a strong result that implies $\mathrm{E}\left[\frac{\hat{A}_m}{\hat{B}_m}\right] \to \frac{A}{B}$, justifying previous work [3–5]. Moreover, the error between these expectations scales with $\frac{1}{m}$ (Appendix A.2), just as for standard ratio estimators [35]. Although real SNPs are not independent due to genetic linkage, this estimator will perform well if the effective number of independent SNPs is large.

Another approach is the "mean-of-ratios" estimator $\frac{1}{m}\sum_{i=1}^{m} \frac{a_i}{b_i}$, used often to estimate kinship coefficients [9, 10, 13–18] and F_{ST} [20]. If each $\frac{a_i}{b_i}$ is biased, their average across SNPs will also be biased, even as $m \to \infty$. However, if $\mathrm{E}\left[\frac{a_i}{b_i}\right] \to \frac{A}{B}$ for all SNPs $i = 1, \ldots, m$ as the number of

individuals $n \to \infty$, and $\mathrm{Var}\left(\frac{a_i}{b_i}\right)$ is bounded, then

$$\frac{1}{m}\sum_{i=1}^{m}\frac{a_i}{b_i} \xrightarrow[n,m\to\infty]{\text{a.s.}} \frac{A}{B}.$$

Therefore, mean-of-ratios estimators must satisfy more restrictive conditions than ratio-of-means estimators, as well as both large n and m, to estimate $\frac{A}{B}$ well.

4 $F_{\mathbf{ST}}$ estimation based on the island model

4.1 The island model $F_{\mathbf{ST}}$ estimator for infinite population sample sizes

Here we study the Weir-Cockerham (WC) [3] and "Hudson" [4] F_{ST} estimators, which assume the island model. These method of moment estimators have small sample size corrections that remarkably make them consistent as the number of independent SNPs m goes to infinity for finite numbers of individuals. However, these small sample corrections also make the estimators more notationally cumbersome than needed here. In order to illustrate clearly how these estimators behave, both under the island model and arbitrary structure, here we construct simplified versions that assume infinite sample sizes per population. This simplification corresponds to eliminating statistical sampling, leaving only genetic sampling to analyze [36]. Note that our simplified estimator nevertheless illustrates the general behavior of the WC and Hudson estimators under arbitrary structure, and the results are equivalent to those we would obtain under finite sample sizes of indivduals.

The Hudson F_{ST} estimator compares two populations [4]; we present a generalized Hudson estimator for K populations in Appendix B. Let us assume that population sample sizes are infinite, so allele frequencies are known. Let j index populations rather than individuals, n be the number of populations, and π_{ij} be the allele frequency in population j at SNP i. In this special case, both WC and Hudson simplify to the following island model F_{ST} estimator:

$$\hat{p}_i^T = \frac{1}{n}\sum_{j=1}^{n}\pi_{ij}, \tag{12}$$

$$\hat{\sigma}_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(\pi_{ij}-\hat{p}_i^T\right)^2, \tag{13}$$

$$\hat{F}_{\mathrm{ST}}^{\mathrm{island}} = \frac{\sum_{i=1}^{m}\hat{\sigma}_i^2}{\sum_{i=1}^{m}\hat{p}_i^T\left(1-\hat{p}_i^T\right)+\frac{1}{n}\hat{\sigma}_i^2}. \tag{14}$$

The goal is to estimate F_{ST} of Eq. (10) with uniform weights ($w_j = \frac{1}{n}\ \forall j$), under our coancestry model defined in Eqs. (6) – (8).

4.2 $F_{\mathbf{ST}}$ estimation under the island model

Under the island model, $\theta_{jk}^T = 0$ for $j \neq k$, the estimator of Eq. (14) can be derived directly using the method of moments (Appendix C.1). Given the IAF moment Eqs. (6) and (7), the expectations of the two recurrent terms of Eq. (14) are

$$\begin{aligned}
\mathrm{E}\left[\frac{1}{m}\sum_{i=1}^{m}\hat{\sigma}_i^2 \middle| T\right] &= \overline{p(1-p)}^T F_{\mathrm{ST}}, \\
\mathrm{E}\left[\frac{1}{m}\sum_{i=1}^{m}\hat{p}_i^T\left(1-\hat{p}_i^T\right) \middle| T\right] &= \overline{p(1-p)}^T\left(1-\frac{F_{\mathrm{ST}}}{n}\right), \quad \text{where} \\
\overline{p(1-p)}^T &= \frac{1}{m}\sum_{i=1}^{m} p_i^T\left(1-p_i^T\right).
\end{aligned}$$

Eliminating $\overline{p(1-p)}^T$ and solving for F_{ST} in this system of equations recovers the estimator of Eq. (14).

Before applying the convergence result of Eq. (11), we test that its assumptions are met. The SNP i terms are $a_i = \hat{\sigma}_i^2$ and $b_i = \hat{p}_i^T\left(1-\hat{p}_i^T\right) + \frac{1}{n}\hat{\sigma}_i^2$, which satisfy $\mathrm{E}[a_i] = Ac_i$ and $\mathrm{E}[b_i] = Bc_i$ with $A = F_{\mathrm{ST}}$, $B = 1$, and $c_i = p_i^T\left(1-p_i^T\right)$. Further, $\bar{c}_m \to c = \mathrm{E}\left[p_i^T\left(1-p_i^T\right)\right] \neq 0$ over the p_i^T distribution across SNPs. Lastly, since $\pi_{ij}, \hat{p}_i^T \in [0,1]$ hold, then $0 \leq \hat{\sigma}_i^2 \leq 1$ and $0 \leq \hat{p}_i^T\left(1-\hat{p}_i^T\right) \leq \frac{1}{4}$, and since $n \geq 2$, $C = 1$ bounds both $|a_i|$ and $|b_i|$. Therefore, for independent SNPs,

$$\hat{F}_{\mathrm{ST}}^{\mathrm{island}} \xrightarrow[m\to\infty]{\text{a.s.}} F_{\mathrm{ST}}.$$

4.3 $F_{\mathbf{ST}}$ estimation under arbitrary coancestry

Now we consider applying the island F_{ST} estimator to non-island settings. The key difference is that $\theta_{jk}^T \neq 0$ for every (j,k) will be assumed in our coancestry model of Eqs. (6) and (7). In this general setting, (j,k) may index either populations or individuals. The two terms of $\hat{F}_{\mathrm{ST}}^{\mathrm{island}}$ now satisfy

$$\begin{aligned}
\mathrm{E}\left[\frac{1}{m}\sum_{i=1}^{m}\hat{\sigma}_i^2 \middle| T\right] &= \overline{p(1-p)}^T\left(F_{\mathrm{ST}} - \bar{\theta}^T\right)\frac{n}{n-1}, \\
\mathrm{E}\left[\frac{1}{m}\sum_{i=1}^{m}\hat{p}_i^T\left(1-\hat{p}_i^T\right) \middle| T\right] &= \overline{p(1-p)}^T\left(1-\bar{\theta}^T\right),
\end{aligned}$$

where $\bar{\theta}^T = \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n}\theta_{jk}^T$ is the mean coancestry with uniform weights. There are two equations but three unknowns: F_{ST}, $\bar{\theta}^T$, and $\overline{p(1-p)}^T$. Island models satisfy $\bar{\theta}^T = \frac{1}{n}F_{\mathrm{ST}}$, which allows for the consistent estimation of F_{ST}. Therefore, the new unknown $\bar{\theta}^T$ precludes consistent F_{ST} estimation without additional assumptions.

The island model F_{ST} estimator converges more generally to

$$\hat{F}_{\mathrm{ST}}^{\mathrm{island}} \xrightarrow[m\to\infty]{\text{a.s.}} \frac{n\left(F_{\mathrm{ST}}-\bar{\theta}^{T}\right)}{n-1+F_{\mathrm{ST}}-n\bar{\theta}^{T}} = \frac{F_{\mathrm{ST}}-\frac{1}{n-1}\left(n\bar{\theta}-F_{\mathrm{ST}}\right)}{1-\frac{1}{n-1}\left(n\bar{\theta}-F_{\mathrm{ST}}\right)}, \tag{15}$$

where it should be noted that

$$\frac{1}{n-1}\left(n\bar{\theta}-F_{\mathrm{ST}}\right) = \frac{1}{n(n-1)}\sum_{j\neq k}\theta_{jk}$$

is the average of all between-individual coancestry coefficients, a term that appears in a related result for populations [5]. Therefore, under arbitrary structure the island model estimator's bias is due to the coancestry between individuals (or islands in the traditional, non-overlapping subpopulation setting).

Since $\frac{1}{n}F_{\mathrm{ST}} \leq \bar{\theta}^{T} \leq F_{\mathrm{ST}}$ (Appendix D), this estimator has a downward bias in non-island settings: it is asymptotically unbiased ($\hat{F}_{\mathrm{ST}}^{\mathrm{island}} \xrightarrow[m\to\infty]{\text{a.s.}} F_{\mathrm{ST}}$) only when $\bar{\theta}^{T} = \frac{1}{n}F_{\mathrm{ST}}$, while bias is maximal when $\bar{\theta}^{T} = F_{\mathrm{ST}}$, where $\hat{F}_{\mathrm{ST}}^{\mathrm{island}} \xrightarrow[m\to\infty]{\text{a.s.}} 0$. For example, if $\theta_{jk}^{T} \approx F_{\mathrm{ST}}$ for most pairs of individuals, then $\bar{\theta}^{T} \approx F_{\mathrm{ST}}$ as well, and $\hat{F}_{\mathrm{ST}} \approx 0$. Therefore, the magnitude of the bias of $\hat{F}_{\mathrm{ST}}^{\mathrm{island}}$ is unknown if $\bar{\theta}^{T}$ is unknown, and small $\hat{F}_{\mathrm{ST}}^{\mathrm{island}}$ may arise even if F_{ST} is very large.

4.4 Consistent estimator of the individual-level pairwise F_{ST}

The individual-level pairwise F_{ST}, equal to F_{ST} for $n = 2$ and denoted by F_{jk}, is always an island model since $T = L_{jk}$ must the most recent ancestral population shared by (j, k) and satisfies $\theta_{jk}^{L_{jk}} = 0$ [25]. Hence, F_{jk} can be estimated consistently using $\hat{F}_{\mathrm{ST}}^{\mathrm{island}}$ of Eq. (14) with $n = 2$, which simplifies to

$$\hat{F}_{jk} = \frac{\sum_{i=1}^{m}(\pi_{ij}-\pi_{ik})^2}{\sum_{i=1}^{m}\pi_{ij}(1-\pi_{ik})+\pi_{ik}(1-\pi_{ij})} \xrightarrow[m\to\infty]{\text{a.s.}} \frac{\frac{\theta_{jj}^{T}+\theta_{kk}^{T}}{2}-\theta_{jk}^{T}}{1-\theta_{jk}^{T}} = F_{jk}, \tag{16}$$

where the limit is stated for general $T \neq L_{jk}$ and matches F_{jk} under the coancestry model [25].

To obtain an estimator of F_{jk} that uses genotypes, we replace π_{ij} by $\frac{x_{ij}}{2}$ in Eq. (16) and convert kinship to inbreeding coefficients using $f_{j}^{T} = 2\varphi_{jj}^{T} - 1$, resulting in

$$\hat{F}_{jk} = 2\left(\frac{\sum_{i=1}^{m}(x_{ij}-x_{ik})^2}{\sum_{i=1}^{m}x_{ij}(2-x_{ik})+x_{ik}(2-x_{ij})}\right) - 1 \xrightarrow[m\to\infty]{\text{a.s.}} \frac{\frac{f_{j}^{T}+f_{k}^{T}}{2}-\varphi_{jk}^{T}}{1-\varphi_{jk}^{T}}, \tag{17}$$

which converges to F_{jk} in Eq. (5) if j and k are locally outbred and locally unrelated ($f_{j}^{L_j} = f_{k}^{L_k} = \varphi_{jk}^{L_{jk}} = 0$). For general values of $f_{j}^{L_j}$, $f_{k}^{L_k}$, and $\varphi_{jk}^{L_{jk}}$,

$$\hat{F}_{jk} \xrightarrow[m\to\infty]{\text{a.s.}} \frac{F_{jk}-\varphi_{jk}^{L_{jk}}+\frac{f_{j}^{L_j}\left(1-f_{L_j}^{T}\right)+f_{k}^{L_k}\left(1-f_{L_k}^{T}\right)}{2\left(1-f_{L_{jk}}^{T}\right)}}{1-\varphi_{jk}^{L_{jk}}}, \tag{18}$$

which is obtained by substituting Eq. (3) into Eq. (17) and rearranging. Since $\varphi_{jk}^{L_{jk}}$ is the only negative term in Eq. (18), local kinship can result in negative $\hat{F}_{jk}$ estimated from genotypes.

To compare our individual-level estimates to Hudson estimates between the two populations S_u and S_v that are not necessarily panmictic, consider the following average $\hat{F}_{jk}$ across populations and its limit assuming locally outbred and locally unrelated individuals:

$$\hat{\bar{F}}_{uv} = \frac{1}{|S_u||S_v|}\sum_{j\in S_u}\sum_{k\in S_v}\hat{F}_{jk} \xrightarrow[m\to\infty]{\text{a.s.}} \frac{1}{2|S_u||S_v|}\sum_{j\in S_u}\sum_{k\in S_v} f_{L_j}^{L_{jk}} + f_{L_k}^{L_{jk}} \tag{19}$$

If S_u and S_v are panmictic populations, then $L_j = S_u$, $L_k = S_v$, and $L_{jk} = L_{uv}$, so every pair and their average $\hat{F}_{jk}, \hat{\bar{F}}_{uv} \xrightarrow[m\to\infty]{\text{a.s.}} \frac{1}{2}\left(f_{S_u}^{L_{uv}} + f_{S_v}^{L_{uv}}\right)$ match the limit of the Hudson estimator. In Section 7, we show empirically that $\hat{\bar{F}}_{uv}$ tends to be larger than the corresponding Hudson estimate when S_u and S_v are structured.

4.5 Coancestry estimation as a method of moments

Since the generalized F_{ST} is given by coancestry coefficients θ_{jj}^T in Eq. (10), a new F_{ST} estimator could be derived from estimates of θ_{jj}^T. Here we attempt to define a method of moments estimator for θ_{jk}^T, and find an underdetermined estimation problem, just as for F_{ST}.

Given IAFs and Eqs. (6) and (7), the first and second moments that average across SNPs are

$$\text{E}\left[\frac{1}{m}\sum_{i=1}^{m}\pi_{ij}\middle|T\right] = \bar{p}^T, \tag{20}$$

$$\text{E}\left[\frac{1}{m}\sum_{i=1}^{m}\pi_{ij}\pi_{ik}\middle|T\right] = \overline{p^2}^T + \overline{p(1-p)}^T\theta_{jk}^T, \tag{21}$$

where $\bar{p}^T = \frac{1}{m}\sum_{i=1}^{m} p_i^T$, $\overline{p^2}^T = \frac{1}{m}\sum_{i=1}^{m}\left(p_i^T\right)^2$, and $\overline{p(1-p)}^T = \bar{p}^T - \overline{p^2}^T$ is as before.

Suppose first that only θ_{jj}^T are of interest. There are n estimators given by Eq. (21) with $j = k$, each corresponding to an unknown θ_{jj}^T. However, all these estimators share two nuisance parameters: $\bar{p}^T$ and $\overline{p^2}^T$. While $\bar{p}^T$ can be estimated from Eq. (20), there are no more equations left to estimate $\overline{p^2}^T$, so this system is underdetermined. The estimation problem remains underdetermined if all $\frac{n(n+1)}{2}$ estimators of Eq. (21) are considered rather than only the $j = k$ cases. Therefore, we cannot estimate coancestry coefficients consistently using only the first two moments and without additional assumptions.

5 Characterizing a kinship estimator and its relationship to F_{ST}

Estimation of kinship coefficients is an important problem, particularly for GWAS approaches that control for population structure [6–18, 37, 38]. Additionally, kinship coefficients are closely related to the generalized F_{ST} of Eq. (4) and the biases of $\hat{F}_{\text{ST}}^{\text{island}}$ in Eq. (15) (since coancestry and kinship

coefficients are related by Eq. (3)). In this section, we focus on a standard kinship method of moments estimator and calculate its limit for the first time (Fig. 1). We study estimators that use genotypes or IAFs, and construct F_{ST} estimators from their kinship estimates. We find biases comparable to those of $\hat{F}_{\mathrm{ST}}^{\mathrm{island}}$, and define unbiased F_{ST} estimators that require knowing the mean kinship or coancestry, or its proportion relative to F_{ST}. Lastly, we present a new kinship method of moments estimator with a uniform bias, which facilitates the estimation of the unknown mean kinship parameter needed to unbias kinship and F_{ST} estimates (Fig. 1).

5.1 Characterization of the standard kinship estimator

Here we analyze a standard kinship estimator that is in frequent use [9, 10, 13–18]. We generalize this estimator to use weights in estimating the ancestral allele frequencies, and we write it as a ratio-of-means estimator due to the favorable theoretical properties of this format as detailed in Section 3:

$$\hat{p}_i^T = \frac{1}{2}\sum_{j=1}^{n} w_j x_{ij}, \tag{22}$$

$$\hat{\varphi}_{jk}^T = \frac{\sum_{i=1}^{m}\left(x_{ij} - 2\hat{p}_i^T\right)\left(x_{ik} - 2\hat{p}_i^T\right)}{4\sum_{i=1}^{m}\hat{p}_i^T\left(1-\hat{p}_i^T\right)}. \tag{23}$$

The estimator in Eq. (23) resembles the sample genotype covariance, but centers by SNP i rather than by individuals j and k, and normalizes by estimates of $4p_i^T\left(1-p_i^T\right)$. We also derive the estimator of Eq. (23) directly using the method of moments (Appendix C.2). The weights in Eq. (22) must satisfy $w_j > 0$ and $\sum_{j=1}^{n} w_j = 1$, so $\hat{p}_i^T \in [0,1]$ and $\mathrm{E}\left[\hat{p}_i^T \middle| T\right] = p_i^T$ hold.

Assuming the moments of Eqs. (1) and (2), we find that Eq. (23) converges to

$$\hat{\varphi}_{jk}^T \xrightarrow[m\to\infty]{\text{a.s.}} \frac{\varphi_{jk}^T - \bar{\varphi}_j^T - \bar{\varphi}_k^T + \bar{\varphi}^T}{1-\bar{\varphi}^T}, \tag{24}$$

where $\bar{\varphi}_j^T = \sum_{k'=1}^{n} w_{k'}\varphi_{jk'}^T$ and $\bar{\varphi}^T = \sum_{j'=1}^{n}\sum_{k'=1}^{n} w_{j'}w_{k'}\varphi_{j'k'}^T$. (See Appendix E for moments involving x_{ij} and $\hat{p}_i^T$ that lead to Eq. (24).) Therefore, the bias of $\hat{\varphi}_{jk}^T$ varies per j and k. Analogous distortions have been observed for sample covariances of genotypes [39]. Similarly, inbreeding coefficient estimates derived from Eq. (23) converge to

$$\hat{f}_j^T = 2\hat{\varphi}_{jj}^T - 1 \xrightarrow[m\to\infty]{\text{a.s.}} \frac{f_j^T - 4\bar{\varphi}_j^T + 3\bar{\varphi}^T}{1-\bar{\varphi}^T}. \tag{25}$$

The limits of the ratio-of-means versions of two more f_j^T estimators [14] are, if $\hat{p}_i^T$ uses Eq. (22),

$$\begin{aligned}
\hat{f}_j^{T,\mathrm{II}} &= 1 - \frac{\sum_{i=1}^{m} x_{ij}(2-x_{ij})}{2\sum_{i=1}^{m}\hat{p}_i^T\left(1-\hat{p}_i^T\right)} \xrightarrow[m\to\infty]{\text{a.s.}} \frac{f_j^T - \bar{\varphi}^T}{1-\bar{\varphi}^T}, \\
\hat{f}_j^{T,\mathrm{III}} &= \frac{\sum_{i=1}^{m} x_{ij}^2 - \left(1+2\hat{p}_i^T\right)x_{ij} + 2\left(\hat{p}_i^T\right)^2}{2\sum_{i=1}^{m}\hat{p}_i^T\left(1-\hat{p}_i^T\right)} \xrightarrow[m\to\infty]{\text{a.s.}} \frac{f_j^T + \bar{\varphi}^T - 2\bar{\varphi}_j^T}{1-\bar{\varphi}^T}.
\end{aligned} \tag{26}$$

The estimators of Eqs. (23) and (26) are unbiased when $\hat{p}_i^T$ is replaced by p_i^T [10, 14], and are consistent when $\hat{p}_i^T$ is consistent [27]. Surprisingly, $\hat{p}_i^T$ of Eq. (22) is not consistent (it does not converge almost surely) for arbitrary population structures, which is at the root of the bias of Eq. (24). In particular, although $\hat{p}_i^T$ is unbiased, its variance (see Appendix E),

$$\mathrm{Var}\left(\hat{p}_i^T \middle| T\right) = p_i^T \left(1 - p_i^T\right) \bar{\varphi}^T, \tag{27}$$

may be asymptotically non-zero as $n \to \infty$, since $p_i^T \in (0,1)$ is fixed and $\lim_{n\to\infty} \bar{\varphi}^T$ may take on any value in [0,1] for arbitrary population structures. Further, $\bar{\varphi}^T \to 0$ as $n \to \infty$ if and only if $\varphi_{jk}^T = 0$ for almost all pairs of individuals (j,k). These observations hold for any weights such that $w_j > 0, \sum_{j=1}^n w_j = 1$. An important consequence is that the plug-in estimate of $p_i^T \left(1 - p_i^T\right)$ is biased (Appendix E),

$$\mathrm{E}\left[\hat{p}_i^T \left(1 - \hat{p}_i^T\right) \middle| T\right] = p_i^T \left(1 - p_i^T\right) \left(1 - \bar{\varphi}^T\right),$$

which is present in all estimators we have studied.

5.2 Estimation of coancestry coefficients from IAFs

Here we form a coancestry coefficient estimator analogous to Eq. (23) but using IAFs. Assuming the moments of Eqs. (6) and (7), this estimator and its limit are

$$\hat{p}_i^T = \sum_{j=1}^n w_j \pi_{ij}, \tag{28}$$

$$\hat{\theta}_{jk}^T = \frac{\sum_{i=1}^m (\pi_{ij} - \hat{p}_i^T)(\pi_{ik} - \hat{p}_i^T)}{\sum_{i=1}^m \hat{p}_i^T \left(1 - \hat{p}_i^T\right)} \xrightarrow[m\to\infty]{\text{a.s.}} \frac{\theta_{jk}^T - \bar{\theta}_j^T - \bar{\theta}_k^T + \bar{\theta}^T}{1 - \bar{\theta}^T}, \tag{29}$$

where $\bar{\theta}_j^T = \sum_{k=1}^n w_k \theta_{jk}^T$ and $\bar{\theta}^T = \sum_{j=1}^n \sum_{k=1}^n w_j w_k \theta_{jk}^T$ are analogous to $\bar{\varphi}_j^T$ and $\bar{\varphi}^T$. Eq. (28) generalizes Eq. (12) for arbitrary weights. Thus, use of IAFs does not ameliorate the estimation problems we have identified for genotypes. Like Eq. (27), $\hat{p}_i^T$ of Eq. (28) is not consistent because $\mathrm{Var}\left(\hat{p}_i^T \middle| T\right) = p_i^T \left(1 - p_i^T\right) \bar{\theta}^T$, which causes the bias observed in Eq. (29).

5.3 Plug-in F_{ST} estimator from inbreeding or coancestry estimates

Since the generalized F_{ST} is defined as a mean inbreeding coefficient in Eq. (4), or equivalently a mean self-coancestry coefficient in Eq. (10), here we study F_{ST} estimators constructed as either $\hat{F}_{\mathrm{ST}} = \sum_{j=1}^n w_j \hat{f}_j^T$ or $\hat{F}_{\mathrm{ST}} = \sum_{j=1}^n w_j \hat{\theta}_{jj}^T$. Although the previous $\hat{f}_j^T$ and $\hat{\theta}_{jj}^T$ are biased, we nevertheless plug them into our definition of F_{ST} so that we may study how bias manifests. Note that we do not recommend utilizing these F_{ST} estimators in practice, but we find these results informative for identifying how to proceed in deriving new estimators.

Remarkably, the three f_j^T estimators of Eqs. (25) and (26) give exactly the same plug-in $\hat{F}_{\mathrm{ST}}$ if the weights in F_{ST} and $\hat{p}_i^T$ of Eq. (22) match, namely

$$\hat{F}_{\mathrm{ST}} = \sum_{j=1}^{n} w_j \hat{f}_j^T = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} w_j \left(x_{ij} - 2\hat{p}_i^T\right)^2}{2\sum_{i=1}^{m} \hat{p}_i^T\left(1-\hat{p}_i^T\right)} - 1 \xrightarrow[m\to\infty]{\text{a.s.}} \frac{F_{\mathrm{ST}} - \bar{\varphi}^T}{1-\bar{\varphi}^T}, \tag{30}$$

where the limit assumes locally outbred individuals so $F_{\mathrm{ST}} = \sum_{j=1}^{n} w_j f_j^T$ holds. The analogous F_{ST} estimator for IAFs and its limit are

$$\hat{F}_{\mathrm{ST}} = \sum_{j=1}^{n} w_j \hat{\theta}_{jj}^T = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} w_j \left(\pi_{ij} - \hat{p}_i^T\right)^2}{\sum_{i=1}^{m} \hat{p}_i^T\left(1-\hat{p}_i^T\right)} \xrightarrow[m\to\infty]{\text{a.s.}} \frac{F_{\mathrm{ST}} - \bar{\theta}^T}{1-\bar{\theta}^T}. \tag{31}$$

The estimators of Eqs. (30) and (31) for individuals and their limits resemble those of classical F_{ST} estimators for populations of the form $\frac{\sigma_p^2}{\bar{p}(1-\bar{p})}$ [5, 34]. $\hat{F}_{\mathrm{ST}}$ of Eq. (31) for uniform weight is also G_{ST} for a biallelic locus [40] if we treat individuals j as populations and combine SNPs as a ratio-of-means estimator. Compared to $\hat{F}_{\mathrm{ST}}^{\mathrm{island}}$ of Eq. (14), $\hat{F}_{\mathrm{ST}}$ of Eq. (31) admits arbitrary weights and, by forgoing bias correction under the island model, is a simpler target of study.

Like $\hat{F}_{\mathrm{ST}}^{\mathrm{island}}$ of Eq. (14), $\hat{F}_{\mathrm{ST}}$ of Eqs. (30) and (31) are downwardly biased since $0 \le \bar{\varphi}^T, \bar{\theta}^T$. $\hat{F}_{\mathrm{ST}}$ of Eq. (31) may converge arbitrarily close to zero since $\bar{\theta}^T$ can be arbitrarily close to F_{ST} (Appendix D). Moreover, although $\bar{\varphi}^T \approx \bar{\theta}^T$ for large n (due to Eq. (9)), in extreme cases $\bar{\varphi}^T$ can exceed F_{ST} under the coancestry model (where $\bar{\theta}^T \le \bar{\varphi}^T$ holds) and also under extreme local kinship, where $\hat{F}_{\mathrm{ST}}$ of Eq. (30) converges to a negative value.

5.4 Adjusted consistent F_{ST} estimators and the "bias coefficient"

Here we explore two adjustments to $\hat{F}_{\mathrm{ST}}$ from IAFs of Eq. (31) that rely on having minimial additional information needed to correct its bias. If $\bar{\theta}^T$ is known, the bias in Eq. (31) can be reversed, yielding the consistent estimator

$$\hat{F}'_{\mathrm{ST}} = \hat{F}_{\mathrm{ST}}(1-\bar{\theta}^T) + \bar{\theta}^T \xrightarrow[m\to\infty]{\text{a.s.}} F_{\mathrm{ST}}. \tag{32}$$

Consistent estimates are also possible if a scaled version of $\bar{\theta}^T$ is known, namely

$$s = \frac{\bar{\theta}^T}{F_{\mathrm{ST}}} = \frac{\sum_{j=1}^{n}\sum_{k=1}^{n} w_j w_k \theta_{jk}^T}{\sum_{j=1}^{n} w_j \theta_{jj}^T}, \tag{33}$$

which we call the "bias coefficient" and has interesting properties. This coefficient measures the strength of the covariances relative to the variances, and satisfies $0 \le s \le 1$ (Appendix D). The limit of Eq. (31) in terms of s is

$$\hat{F}_{\mathrm{ST}} \xrightarrow[m\to\infty]{\text{a.s.}} F_{\mathrm{ST}} \frac{1-s}{1-sF_{\mathrm{ST}}}. \tag{34}$$

Treating the limit as equality and solving for F_{ST} yields the following consistent estimator:

$$\hat{\sigma}_i^2 = \frac{1}{1-s}\sum_{j=1}^{n} w_j(\pi_{ij} - \hat{p}_i^T)^2, \tag{35}$$

$$\hat{F}''_{\mathrm{ST}} = \frac{\hat{F}_{\mathrm{ST}}}{1 - s(1-\hat{F}_{\mathrm{ST}})} = \frac{\sum_{i=1}^{m}\hat{\sigma}_i^2}{\sum_{i=1}^{m}\hat{p}_i^T\left(1-\hat{p}_i^T\right) + s\hat{\sigma}_i^2} \xrightarrow[m\to\infty]{\text{a.s.}} F_{\mathrm{ST}}. \tag{36}$$

Note that $\hat{\sigma}_i^2$ and $\hat{F}_{\mathrm{ST}}^{\mathrm{island}}$ from Eqs. (13) and (14) are the special case of Eqs. (35) and (36) for uniform weights and $s = \frac{1}{n}$; hence, $\hat{F}''_{\mathrm{ST}}$ generalizes $\hat{F}_{\mathrm{ST}}^{\mathrm{island}}$.

Lastly, using either Eq. (31) or Eq. (34), the relative error of $\hat{F}_{\mathrm{ST}}$ converges to

$$1 - \frac{\hat{F}_{\mathrm{ST}}}{F_{\mathrm{ST}}} \xrightarrow[m\to\infty]{\text{a.s.}} \frac{\bar{\theta}^T\left(1-F_{\mathrm{ST}}\right)}{F_{\mathrm{ST}}\left(1-\bar{\theta}^T\right)} = s\frac{1-F_{\mathrm{ST}}}{1-sF_{\mathrm{ST}}}, \tag{37}$$

which is approximated by s if $F_{\mathrm{ST}} \ll 1$, hence the name "bias coefficient".

5.5 A new direction for F_{ST} and kinship estimation

Here, we outline a new estimation framework for kinship coefficients that has properties favorable for obtaining nearly unbiased estimates. These new kinship estimates can then also be utilized for F_{ST} estimation. We summarize our ideas here and then fully develop the estimation framework and study its operating characteristics in the next paper in this series [32].

Applying the method of moments to Eqs. (1) and (2), we derive the following estimator,

$$\hat{\varphi}_{jk}^{T,\mathrm{new}} = \frac{\sum_{i=1}^{m}(x_{ij}-1)(x_{ik}-1) - 1}{4\sum_{i=1}^{m}\hat{p}_i^T\left(1-\hat{p}_i^T\right)} + 1 \xrightarrow[m\to\infty]{\text{a.s.}} \frac{\varphi_{jk}^T - \bar{\varphi}^T}{1-\bar{\varphi}^T}, \tag{38}$$

which compares favorably to the standard estimator of Eq. (24) by having a uniform bias in the limit, controlled by the sole parameter $\bar{\varphi}^T$. If $\bar{\varphi}^T$ were known, Eq. (38) could be adjusted to yield unbiased kinship estimates:

$$\tilde{\varphi}_{jk}^{T,\mathrm{new}} = \hat{\varphi}_{jk}^{T,\mathrm{new}}\left(1-\bar{\varphi}^T\right) + \bar{\varphi}^T \xrightarrow[m\to\infty]{\text{a.s.}} \varphi_{jk}^T.$$

Remarkably, Eq. (38) itself can be used to estimate $\bar{\varphi}^T$: assuming $\min_{j,k}\varphi_{jk}^T = 0$ and a large number of SNPs m, then

$$\min_{j,k}\hat{\varphi}_{jk}^{T,\mathrm{new}} \approx -\frac{\bar{\varphi}^T}{1-\bar{\varphi}^T},$$

from which $\bar{\varphi}^T$ can be solved. However, additional steps can be taken to provide a more stable estimate than that based on $\min_{j,k}\hat{\varphi}_{jk}^{T,\mathrm{new}}$ [32]. Our improved kinship estimator will result in a plug-in F_{ST} estimator with increased accuracy.

The analogous coancestry estimator using IAF is

$$\hat{\theta}_{jk}^{T,\mathrm{new}} = \frac{\sum_{i=1}^{m}\left(\pi_{ij}-\frac{1}{2}\right)\left(\pi_{ik}-\frac{1}{2}\right) - \frac{1}{4}}{\sum_{i=1}^{m}\hat{p}_i^T\left(1-\hat{p}_i^T\right)} + 1 \xrightarrow[m\to\infty]{\text{a.s.}} \frac{\theta_{jk}^T - \bar{\theta}^T}{1-\bar{\theta}^T}.$$

Lastly, note that the inbreeding coefficient estimator derived from Eq. (38) using $\hat{f}_j^{T,\text{new}} = 2\hat{\varphi}_{jj}^{T,\text{new}} - 1$ equals $\hat{f}_j^{T,\text{II}}$ of Eq. (26), so the resulting plug-in F_{ST} estimator equals that of Eq. (30). Therefore, Eq. (38) by itself does not directly yield a new F_{ST} estimator.

6 An admixture simulation illustrates challenges in F_{ST} and kinship estimation

6.1 Overview of simulations

We simulate genotypes from two models to illustrate our results when the true population structure parameters are known. One is an island model, the other an admixture model differing from the island model by its pervasive covariance, and designed to induce large biases in existing F_{ST} estimators (Fig. 2). Both simulations have $n = 1000$ individuals, $m = 300,000$ SNP loci, and $K = 10$ islands or intermediate populations. These simulations have $F_{\text{ST}} = 0.1$, comparable to estimates between human populations [4].

Our island model satisfies the Hudson estimator assumptions: populations are independent, and each population S_u has a different F_{ST} value of $f_{S_u}^T$ (Fig. 2A). Ancestral allele frequencies p_i^T are drawn uniformly in [0.01, 0.5]. Allele frequencies $p_i^{S_u}$ for S_u and SNP i are drawn independently from the Balding-Nichols (BN) distribution [41] with parameters p_i^T and $f_{S_u}^T$. Every individual j in island S_u draws alleles randomly with probability $p_i^{S_u}$. Population sample sizes were drawn randomly (Appendix F).

Our admixture model is a "BN-PSD" model [9, 17, 25, 27, 42, 43], which we analyzed in our previous paper in this series [25]. The intermediate populations are islands that draw $p_i^{S_u}$ from the BN model, then each individual j constructs its allele frequencies as $\pi_{ij} = \sum_{u=1}^K p_i^{S_u} q_{ju}$, which is a weighted average of $p_i^{S_u}$ with the admixture proportions q_{ju} of j and u as weights (which satisfy $\sum_{u=1}^K q_{ju} = 1$, as in the Pritchard-Stephens-Donnelly [PSD] admixture model [29–31]). We constructed q_{ju} that model admixture resulting from spread by random walk of the intermediate populations along a one-dimensional geography, as follows. Intermediate populations S_u are placed on a line with differentiation $f_{S_u}^T$ that grows with coordinate (Fig. 3A). Upon differentiation, individuals in each S_u spread with random walks, a process modeled by Normal densities (Fig. 3B). Admixed individuals derive their ancestry proportional to these Normal densities, resulting in a genetic structure governed by geography (Fig. 3C, Fig. 2B) and departing strongly from the island model (Fig. 3D). The amount of spread was chosen to give $s = 0.5$, which by Eq. (37) results in a large bias for $\hat{F}_{\text{ST}}$ (in contrast, the island simulation has $s = 0.1$). See Appendix F for additional details regarding these simulations.

6.2 Weir-Cockerham and Hudson F_{ST} estimators misapplied to an admixed population

Our admixture simulation illustrates the large biases that can arise if the WC and Hudson F_{ST} estimators are misapplied to non-island populations to estimate the generalized F_{ST}. First, we test these estimators in our island model. This simulation satisfies the assumptions of the Hudson estimator (which we generalized for K population islands in Appendix B), so it is consistent (Fig. 4A). The WC estimator assumes that $f_{S_u}^T = F_{ST}$ for all u, which does not hold; nevertheless, WC has a small bias (Fig. 4A). For comparison, we added the "plug-in" F_{ST} estimator of Eq. (30) (weights from Appendix F), which is derived from the kinship estimator of Eq. (23) and does not have island model corrections. Since the number of islands K is large, the plug-in estimator has a small relative bias of about $s = \frac{1}{K} = 10\%$; greater bias is expected for smaller K.

To apply the WC and Hudson estimators to the admixture model, individuals are assigned to "populations" grouping by their maximum admixture proportions (Fig. 3D). Both WC and Hudson estimates are smaller than the true F_{ST} by nearly half, as predicted by the limit of $\hat{F}_{ST}^{\text{island}}$ of Eq. (15) (Fig. 4C). By construction, the plug-in $\hat{F}_{ST}$ also has a large relative bias of about $s = 50\%$; remarkably, the WC and Hudson estimators suffer from comparable biases. Thus, the island model corrections of the WC and Hudson estimators are insufficient for estimating F_{ST} in our admixture scenario.

6.3 Evaluation of individual-level pairwise F_{ST} estimators

Fig. 5A shows the matrix of true individual-level pairwise F_{ST} values, F_{jk}, for every pair of individuals in our simulation. F_{jk} is a distance between pairs of individuals, with $F_{jk} = 0$ for pairs from the same population and increasing values for more distant population pairs. Larger θ_{jk}^T lead to smaller F_{jk} (see Eq. (16)), hence the θ_{jk}^T (Fig. 2B) and F_{jk} (Fig. 5A) matrices are negatively correlated.

Both of our consistent F_{jk} estimators perform well, using IAFs (Eq. (16), Fig. 5B) and genotypes (Eq. (17), Fig. 5C). Estimates from genotypes have a greater root-mean-squared error (RMSE, 3.43% relative to the mean F_{jk}) than the estimates from true IAFs (RMSE of 0.319%).

6.4 Evaluation of the standard kinship estimator

Our admixture simulation illustrates the distortions of the kinship estimator $\hat{\varphi}_{jk}^T$ of Eq. (23). The limit of Eq. (23) has a fixed bias if $\bar{\varphi}_j^T = \bar{\varphi}^T$ for all j. For that reason, we chose $f_{S_u}^T$ that vary per u (Fig. 3A), which causes large differences in $\bar{\varphi}_j^T$ per j and large distortions in $\hat{\varphi}_{jk}^T$.

Compared to the true φ_{jk}^T (Fig. 6A, where f_j^T are plotted along the diagonal), $\hat{\varphi}_{jk}^T$ are very distorted, with an abundance of $\hat{\varphi}_{jk}^T < \varphi_{jk}^T$ cases, negative estimates (blue in Fig. 6B), but remarkably also cases with $\hat{\varphi}_{jk}^T > \varphi_{jk}^T$ (top left corner of Fig. 6B). Our ratio-of-means estimator $\hat{\varphi}_{jk}^T$ agrees with the limit of Eq. (24) (Fig. 6C), which an RMSE of 2.14% relative to the mean φ_{jk}^T. In contrast,

mean-of-ratios estimates have an RMSE of 10.77% from the limit of Eq. (24) (not shown). The distortions are similar for the estimator that uses IAFs of Eq. (29) (not shown), with reduced RMSEs from its limit of 0.32% and 8.82% for the ratio-of-means and mean-of-ratios estimates, respectively.

6.5 Evaluation of plug-in and adjusted F_{ST} estimators

We illustrate the behavior of our plug-in and adjusted F_{ST} estimators using our admixture simulation. We tested IAF (Fig. 7A) and genotype (Fig. 7B) versions of our estimators. The unadjusted plug-in $\hat{F}_{\mathrm{ST}}$ of Eq. (31) is severely biased (blue), by construction, and matches the calculated limit for IAFs and genotypes (green dotted lines in Fig. 7, which are close because $\bar{\varphi}^T \approx \bar{\theta}^T$). We also tested the two consistent "adjusted" estimators $\hat{F}'_{\mathrm{ST}}$ and $\hat{F}''_{\mathrm{ST}}$ of Eqs. (32) and (36), which estimate F_{ST} quite well (blue predictions overlap the true F_{ST} red dashed line in Fig. 7). However, $\hat{F}'_{\mathrm{ST}}$ and $\hat{F}''_{\mathrm{ST}}$ are oracle methods, since they require parameters ($\bar{\varphi}^T$, $\bar{\theta}^T$, s) that are not known in practice.

Prediction intervals were computed from estimates over 39 independently-simulated IAF and genotype matrices (Appendix G). Estimator limits are always contained in these intervals, which holds since the number of independent SNPs ($m = 300,000$) is sufficiently large. Estimates that use genotypes have wider intervals than estimates from IAFs; however, IAFs are not known in practice, and use of estimated IAFs might increase noise. Genetic linkage, not present in our simulation, will also increase noise in real data.

7 Analysis of 1000 Genomes Project populations

We analyze 1000 Genomes Project (TGP) populations [20] with the Hudson F_{ST} estimator for two populations and our individual-level F_{ST} estimator, $\hat{F}_{jk}$, of Eq. (17). We focus on $\hat{F}_{jk}$ since it is currently our only consistent estimator for arbitrary population structures. We analyze the 20,417,698 biallelic SNP ascertained in YRI from autosomal chromosomes in the final "phase 3" data on the TGP website (dated 2013-05-02). Of these, 14,145,759 SNPs are polymorphic in the Hispanic populations and 8,932,115 in the European populations discussed below. Individuals in these data are roughly locally outbred and locally unrelated [20], which is the only requirement for the consistency of $\hat{F}_{jk}$ estimated from genotypes.

First we focus on YRI, CEU, and CHB, which were analyzed previously [4]. These population pairs are geographically distant, so the island model is more likely to fit well. Indeed, Hudson estimates are relatively close to $\hat{F}_{jk}$ (compare upper and lower triangle of Fig. 8A). In other words, the structure within populations is dwarfed by the structure between populations. A direct comparison to the Hudson estimates is given by $\hat{\bar{F}}_{uv}$ of Eq. (19), which averages $\hat{F}_{jk}$ across populations for $j \in S_u$ and $k \in S_v$. We find good agreement between Hudson estimates and $\hat{\bar{F}}_{uv}$, corroborating a good fit of the island model (Fig. 8D).

Next, we analyze the four Hispanic populations in the TGP: PEL, MXL, CLM, and PUR. Hispanic individuals are admixed primarily from Native American, European, and African super-

populations. Each of these populations is structured, a consequence of variable individual admixture proportions [27], so pairwise comparisons are poorly fit by the island model. The complex structure of these populations is confirmed by $\hat{F}_{jk}$, finding many individuals that have closer relatives from other populations compared to some individuals from the same population (lower triangle of Fig. 8B). Here we find that $\hat{\tilde{F}}_{uv}$ are always larger than their corresponding Hudson estimates (Fig. 8E). The largest proportional discrepancy is between PUR and CLM, whose Hudson estimate is 40% of $\hat{\tilde{F}}_{uv}$. The Hudson estimator is solely a function of average allele frequencies and sample sizes per population (Appendix B), so it averages out the substructure within populations, explaining the smaller estimates observed relative to $\hat{\tilde{F}}_{uv}$.

Lastly, we analyze four European populations: FIN, GBR, IBS, and TSI. We exclude CEU due to its similarity to GBR and because it was not sampled within Europe. The structure of European populations was previously found to disagree with the island model [44]. We confirm structure within these populations, although differentiation is much smaller here (Fig. 8C). Notably, proportional differences between Hudson and $\hat{\tilde{F}}_{uv}$ are as large within Europe (Fig. 8F) as in the Hispanic populations (Fig. 8E). The largest proportional difference was between TSI and IBS, whose Hudson estimate is 41% of $\hat{\tilde{F}}_{uv}$. Thus, our individual-level pairwise F_{ST} estimator, $\hat{F}_{jk}$, detects structure that is missed by island model estimators.

8 Discussion

We investigated the most commonly utilized estimators of F_{ST} and kinship, both of which can be derived using the method of moments (Fig. 1). We determined the bias of these estimators under models of arbitrary population structure. We calculated the bias that occurs in the F_{ST} estimator when the island model assumption is violated. This bias is present even when individual-specific allele frequencies are known without error. We also showed that the kinship estimator is biased when the population is structured (particularly when the average kinship is of a similar magnitude to the true kinship coefficient), and that the bias may be different for each pair of individuals.

Use of island model F_{ST} estimators requires taking certain precautions, as exemplified in the Hudson F_{ST} estimator work [4]. First, the Hudson estimator is given for two populations only, since two panmictic populations are always independent relative to their last common ancestor population. Second, only geographically distant population pairs were compared [4], which appear internally unstructured relative to the structure between populations. However, F_{ST} is often estimated between closely related populations, for example, within Mexico [21], the United Kingdom [22], and between contemporary and archaic European [23] and Eurasian populations [24]. These geographically close populations are more likely to have comparable structure within and between populations, a case where Hudson underestimates differentiation, just as in the Hispanic and European populations in Fig. 8. Our analyses highlight the need for new tools that measure differentiation in complex population structures.

We have shown that the misapplication of existing F_{ST} estimators on non-island population structures may lead to estimates that approach zero even when the true generalized F_{ST} is large. Weir-Cockerham [3] and Hudson [4] F_{ST} estimates in our admixture simulation are biased by nearly a factor of two (Fig. 4). These estimators were derived assuming independent populations, so the observed biases arise from their misapplication to non-island populations. Nevertheless, natural populations often do not adhere to the island model, particularly human populations [44–46].

The kinship coefficient estimator we investigated is often used to control for population structure in GWAS and to estimate genome-wide heritability [9, 10, 13–18]. While this estimator was known to be biased [10, 18], no closed form limit had been calculated until now. We found that kinship estimates are biased downwardly on average, but bias also varies for every pair of individuals (Fig. 1, Fig. 6). Thus, the use of these distorted kinship estimates may be problematic in GWAS or estimating heritability, but to what extent remains to be determined.

We developed a theoretical framework for assessing these genome-wide ratio estimators of F_{ST} and kinship. We proved that common ratio-of-means estimators converge almost surely to the ratio of expectations for infinite independent SNPs (Appendix A.1). Our result justifies approximating the expectation of a ratio-of-means estimator with the ratio of expectations [3–5]. However, mean-of-ratios estimators may not converge to the ratio of expectations for infinite SNPs. Mean-of-ratios estimators are potentially asymptotically unbiased for infinite individuals, but it is unclear which estimators have this behavior. We found that the ratio-of-means kinship estimator had much smaller errors from the ratio of expectations than the more common mean-of-ratios estimator, whose convergence value is unknown. Thus, we recommend ratio-of-means estimators, whose asymptotic behavior is well understood.

The Hudson estimator is a consistent estimator of the pairwise F_{ST} for two populations [4], which is reported often [21–24, 45, 47]. We derived a consistent estimator of the individual-level pairwise F_{ST}, F_{jk}, which extends the previous pairwise F_{ST} to individuals [25]. However, kinship or F_{ST} estimates for more than two individuals cannot be recovered from F_{jk} estimates. Conceptually, kinship and F_{ST} are in terms of a single ancestral population T, whereas each F_{jk} is relative to a jointly local population L_{jk} that varies per (j, k) pair (see Eq. (5)). Practically, there is loss of information since $F_{jj} = 0$ for every j by definition: for n individuals, there are n more φ^T_{jk} than F_{jk} parameters. We used our F_{jk} estimator to identify structure with individual resolution in 1000 Genomes Project populations (Fig. 8).

Accurate estimation of generalized F_{ST} and kinship coefficients in arbitrary population structures will require further innovations, and the results provided here may be useful in leading to more robust estimators in the future. This, in particular, is the topic we tackle in the next paper in this series [32].

References

[1] G. Malécot. *Mathématiques de l'hérédité*. Masson et Cie, 1948.

[2] S. Wright. "The Genetical Structure of Populations". *Annals of Eugenics* 15(1) (1951), pp. 323–354.

[3] B. S. Weir and C. C. Cockerham. "Estimating F-Statistics for the Analysis of Population Structure". *Evolution* 38(6) (1984), pp. 1358–1370.

[4] G. Bhatia et al. "Estimating and interpreting FST: The impact of rare variants". *Genome Res.* 23(9) (2013), pp. 1514–1521.

[5] B. S. Weir and W. G. Hill. "Estimating F-Statistics". *Annual Review of Genetics* 36(1) (2002), pp. 721–750.

[6] C. Xie, D. D. G. Gessler, and S. Xu. "Combining Different Line Crosses for Mapping Quantitative Trait Loci Using the Identical by Descent-Based Variance Component Method". *Genetics* 149(2) (1998), pp. 1139–1146.

[7] J. Yu et al. "A unified mixed-model method for association mapping that accounts for multiple levels of relatedness". *Nat Genet* 38(2) (2006), pp. 203–208.

[8] Y. S. Aulchenko, D.-J. d. Koning, and C. Haley. "Genomewide Rapid Association Using Mixed Model and Regression: A Fast and Simple Method For Genomewide Pedigree-Based Quantitative Trait Loci Association Analysis". *Genetics* 177(1) (2007), pp. 577–585.

[9] A. L. Price et al. "Principal components analysis corrects for stratification in genome-wide association studies". *Nat Genet* 38(8) (2006), pp. 904–909.

[10] W. Astle and D. J. Balding. "Population Structure and Cryptic Relatedness in Genetic Association Studies". *Statist. Sci.* 24(4) (2009). Mathematical Reviews number (MathSciNet): MR2779337, pp. 451–471.

[11] H. M. Kang et al. "Efficient Control of Population Structure in Model Organism Association Mapping". *Genetics* 178(3) (2008), pp. 1709–1723.

[12] H. M. Kang et al. "Variance component model to account for sample structure in genome-wide association studies". *Nat Genet* 42(4) (2010), pp. 348–354.

[13] J. Yang et al. "Common SNPs explain a large proportion of the heritability for human height". *Nat Genet* 42(7) (2010), pp. 565–569.

[14] J. Yang et al. "GCTA: A Tool for Genome-wide Complex Trait Analysis". *The American Journal of Human Genetics* 88(1) (2011), pp. 76–82.

[15] X. Zhou and M. Stephens. "Genome-wide efficient mixed-model analysis for association studies". *Nat Genet* 44(7) (2012), pp. 821–824.

[16] C. S. Rakovski and D. O. Stram. "A Kinship-Based Modification of the Armitage Trend Test to Address Hidden Population Structure and Small Differential Genotyping Errors". *PLOS ONE* 4(6) (2009), e5825.

[17] T. Thornton and M. S. McPeek. "ROADTRIPS: Case-Control Association Testing with Partially or Completely Unknown Population and Pedigree Structure". *The American Journal of Human Genetics* 86(2) (2010), pp. 172–184.

[18] D. Speed and D. J. Balding. "Relatedness in the post-genomic era: is it still useful?" *Nat Rev Genet* 16(1) (2015), pp. 33–44.

[19] N. A. Rosenberg et al. "Genetic Structure of Human Populations". *Science* 298(5602) (2002), pp. 2381–2385.

[20] T. . G. P. Consortium. "A map of human genome variation from population-scale sequencing". *Nature* 467(7319) (2010), pp. 1061–1073.

[21] A. Moreno-Estrada et al. "The genetics of Mexico recapitulates Native American substructure and affects biomedical traits". *Science* 344(6189) (2014), pp. 1280–1285.

[22] S. Leslie et al. "The fine-scale genetic structure of the British population". *Nature* 519(7543) (2015), pp. 309–314.

[23] W. Haak et al. "Massive migration from the steppe was a source for Indo-European languages in Europe". *Nature* 522(7555) (2015), pp. 207–211.

[24] M. E. Allentoft et al. "Population genomics of Bronze Age Eurasia". *Nature* 522(7555) (2015), pp. 167–172.

[25] A. Ochoa and J. D. Storey. "F_{ST} and kinship for arbitrary population structures I: Generalized definitions". Submitted.

[26] A. Jacquard. *Structures génétiques des populations.* Paris: Masson et Cie, 1970.

[27] T. Thornton et al. "Estimating Kinship in Admixed Populations". *The American Journal of Human Genetics* 91(1) (2012), pp. 122–138.

[28] W. Hao, M. Song, and J. D. Storey. "Probabilistic models of genetic variation in structured populations applied to global human studies". *Bioinformatics* 32(5) (2016), pp. 713–721.

[29] J. K. Pritchard, M. Stephens, and P. Donnelly. "Inference of Population Structure Using Multilocus Genotype Data". *Genetics* 155(2) (2000), pp. 945–959.

[30] H. Tang et al. "Estimation of individual admixture: Analytical and study design considerations". *Genet. Epidemiol.* 28(4) (2005), pp. 289–301.

[31] D. H. Alexander, J. Novembre, and K. Lange. "Fast model-based estimation of ancestry in unrelated individuals". *Genome Res.* 19(9) (2009), pp. 1655–1664.

[32] A. Ochoa and J. D. Storey. "F_{ST} and kinship for arbitrary population structures III: A new estimation framework". In preparation.

[33] S. Wright. "Systems of Mating. V. General Considerations". *Genetics* 6(2) (1921), pp. 167–178.

[34] G. Nicholson et al. "Assessing population differentiation and isolation from single-nucleotide polymorphism data". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64(4) (2002), pp. 695–715.

[35] W. G. Cochran. *Sampling techniques*. 3rd ed. Wiley, 1977.

[36] B. S. Weir. *Genetic data analysis II. Methods for discrete population genetic data.* Sunderland, USA: Sinauer Associates, 1996.

[37] C. Bourgain et al. "Novel Case-Control Test in a Founder Population Identifies P-Selectin as an Atopy-Susceptibility Locus". *The American Journal of Human Genetics* 73(3) (2003), pp. 612–626.

[38] Y. Choi, E. M. Wijsman, and B. S. Weir. "Case-Control Association Testing in the Presence of Unknown Relationships". *Genet Epidemiol* 33(8) (2009), pp. 668–678.

[39] J. K. Pickrell and J. K. Pritchard. "Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data". *PLoS Genet* 8(11) (2012), e1002967.

[40] M. Nei. "Analysis of Gene Diversity in Subdivided Populations". *PNAS* 70(12) (1973), pp. 3321–3323.

[41] D. J. Balding and R. A. Nichols. "A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity". *Genetica* 96(1-2) (1995), pp. 3–12.

[42] D. Falush, M. Stephens, and J. K. Pritchard. "Inference of Population Structure Using Multilocus Genotype Data: Linked Loci and Correlated Allele Frequencies". *Genetics* 164(4) (2003), pp. 1567–1587.

[43] A. Raj, M. Stephens, and J. K. Pritchard. "fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets". *Genetics* 197(2) (2014), pp. 573–589.

[44] J. Novembre et al. "Genes mirror geography within Europe". *Nature* 456(7218) (2008), pp. 98–101.

[45] G. Coop et al. "The Role of Geography in Human Adaptation". *PLoS Genet* 5(6) (2009), e1000500.

[46] N. Patterson et al. "Ancient Admixture in Human History". *Genetics* 192(3) (2012), pp. 1065–1093.

[47] G. Coop et al. "Using Environmental Correlations to Identify Loci Underlying Local Adaptation". *Genetics* 185(4) (2010), pp. 1411–1423.

[48] H. B. Mann and A. Wald. "On Stochastic Limit and Order Relationships". *The Annals of Mathematical Statistics* 14(3) (1943), pp. 217–226.

[49] P. Billingsley. *Convergence of Probability Measures*. John Wiley & Sons, 2013.

[50] W. Feller. *An introduction to probability theory and its applications*. 3rd ed. Vol. 1. John Wiley & Sons London-New York-Sydney-Toronto, 1968.

[51] H. O. Hartley and A. Ross. "Unbiased Ratio Estimators". *Nature* 174(4423) (1954), pp. 270–271.

[52] R. Beran and P. Hall. "Interpolated Nonparametric Prediction Intervals and Confidence Intervals". *Journal of the Royal Statistical Society. Series B (Methodological)* 55(3) (1993), pp. 643–652.

9 Figures

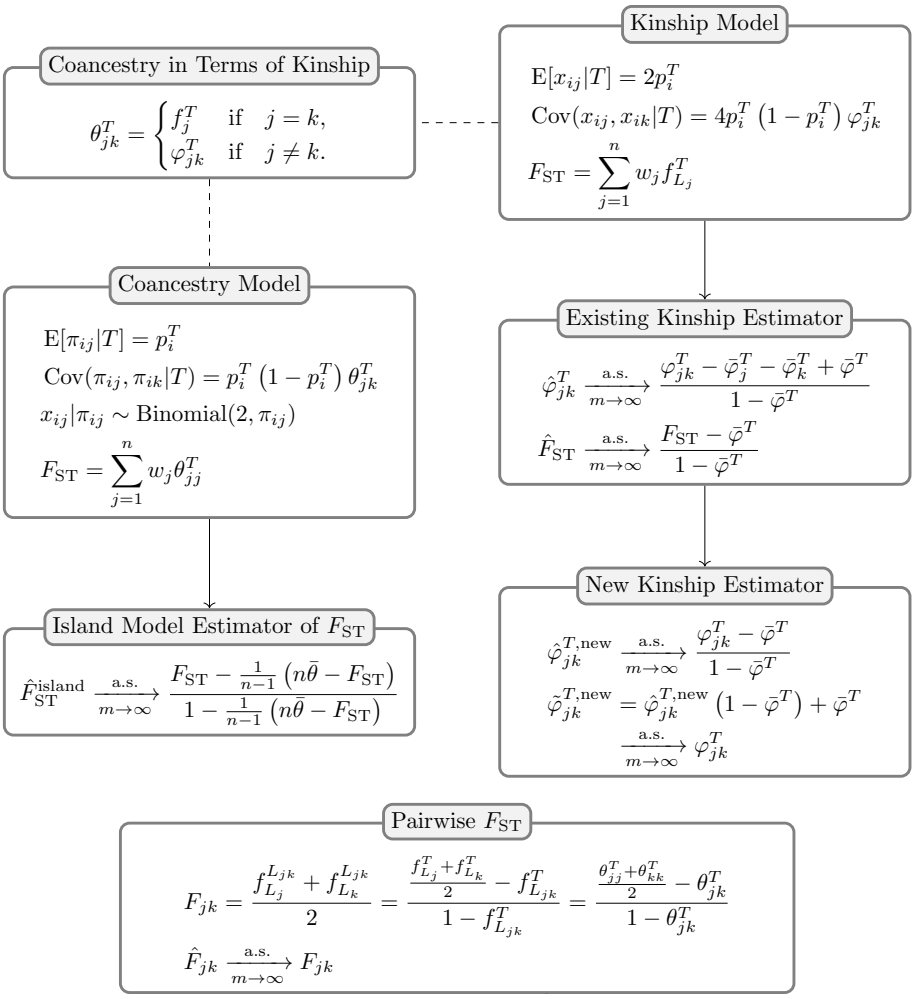

Figure 1: **Accuracy of F_{ST} and kinship estimators: overview of models and results.** Our analysis is based on two parallel models: the coancestry model for individual-specific allele frequencies (π_{ij}), and the kinship model for genotypes (x_{ij}). The kinship (φ^T_{jk}, f^T_j) and coancestry (θ^T_{jk}) parameters are closely related as shown. We use these models to study the accuracy of F_{ST} and kinship method of moment estimators under arbitrary population structures. The bias resulting from the misapplication of the F_{ST} island model estimator ($\hat{F}^{\mathrm{island}}_{\mathrm{ST}}$) to arbitrary structures is calculated under the coancestry model, while the bias in the standard kinship model estimator ($\hat{\varphi}^T_{jk}$) and its resulting plug-in F_{ST} estimator ($\hat{F}_{\mathrm{ST}}$) is calculated under the kinship model. We present a new kinship estimator ($\hat{\varphi}^{T,\mathrm{new}}_{jk}$) with a uniform bias, which will be used to obtain more accurate kinship estimates in the next paper in the series (starting from $\tilde{\varphi}^{T,\mathrm{new}}_{jk}$). Lastly, we present a new pairwise F_{ST} estimator for two individuals ($\hat{F}_{jk}$) that is unbiased for the true pairwise F_{ST} that we introduced in our previous work. Note that estimation of F_{ST} and F_{jk} from genotypes requires individuals to be locally outbred and locally unrelated.

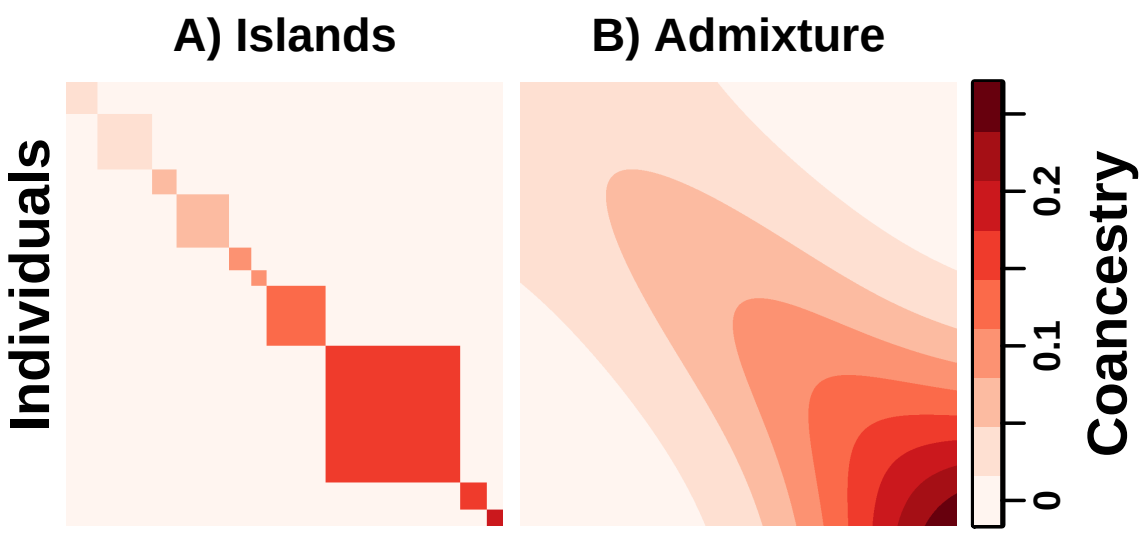

Figure 2: **Coancestry matrices of our simulations.** Both simulations have $n = 1000$ individuals along both axes, $K = 10$ populations (islands or intermediate), and $F_{\mathrm{ST}} = 0.1$. Color corresponds to θ^T_{jk} between individuals j and k. A) The island model has $\theta^T_{jk} = 0$ between islands, and varying θ^T_{jj} per island, resulting in a block-diagonal covariance matrix. B) Our admixture scenario models a 1D geography with extensive admixture and intermediate population differentiation that increases with coordinate. Individuals are ordered by their coordinate in the 1D geography.

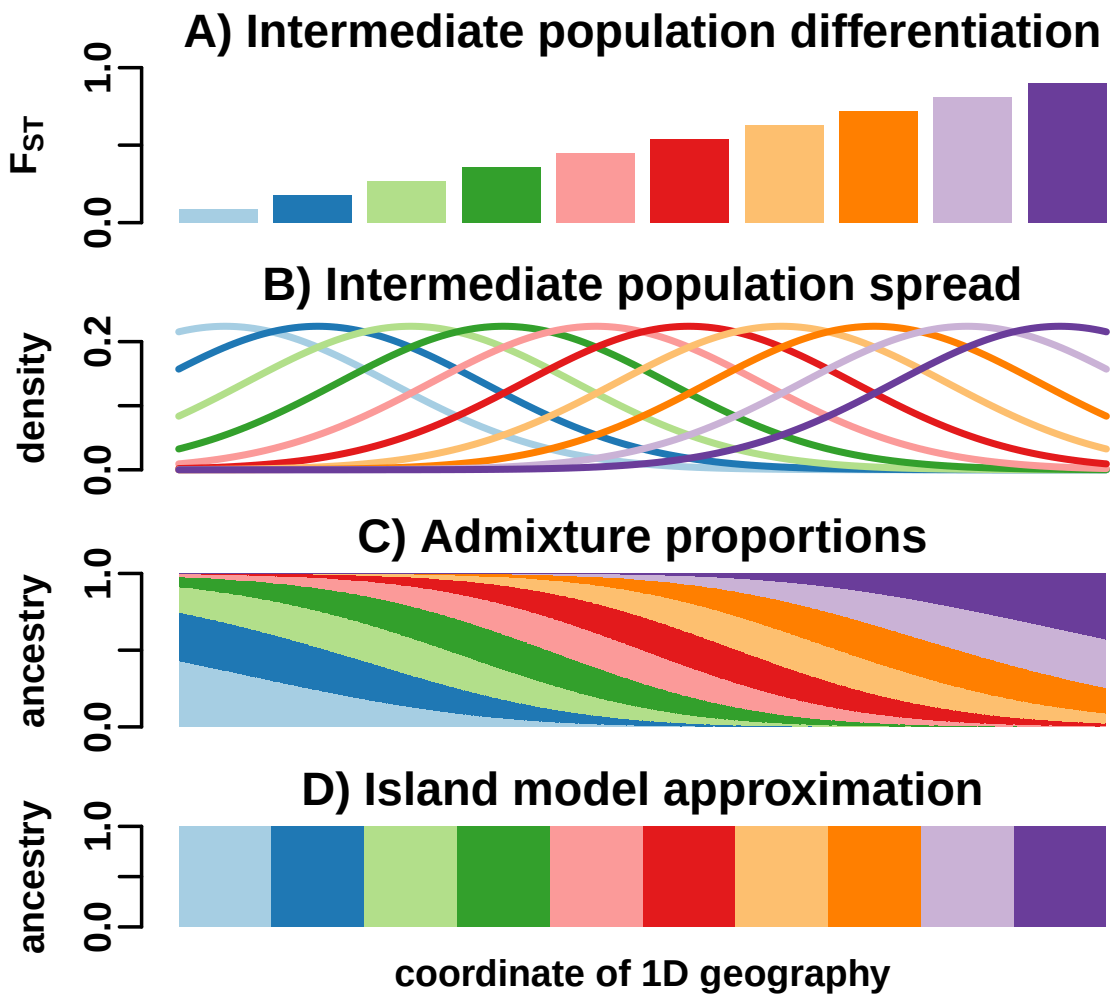

Figure 3: **1D admixture scenario.** We model a 1D geography population that departs strongly from the island model. A) $K = 10$ intermediate populations, placed equidistant on a line, evolve independently with F_{ST} increasing with x-coordinate. B) Once differentiated, these intermediate populations spread by random walks modeled by Normal densities. C) $n = 1000$ individuals, sampled in equal intervals in the same range, are admixed proportionally to the previous Normal densities. D) To apply the WC and Hudson F_{ST} estimators, individuals are assigned to populations ("islands") by their majority ancestry.

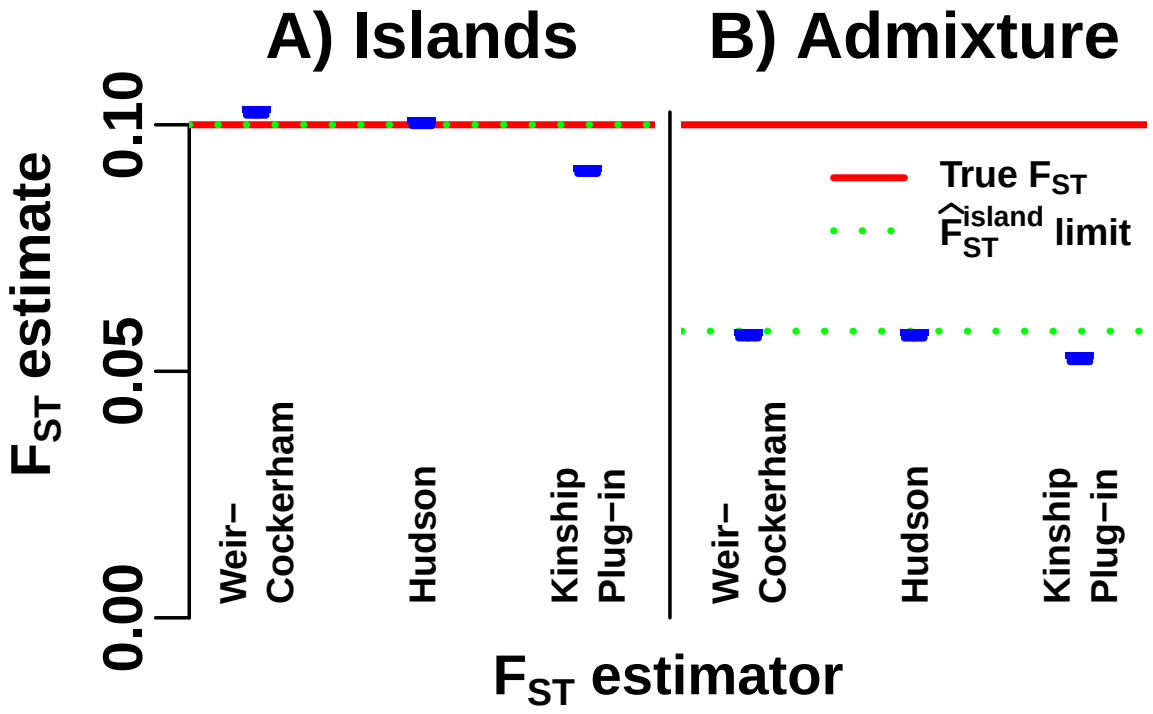

Figure 4: **Weir-Cockerham and Hudson F_{ST} estimators misapplied to our admixture model.** The WC, Hudson, and "kinship plug-in" $\hat{F}_{\mathrm{ST}}$ estimator of Eq. (30), are evaluated on simulated genotypes from our two models (Fig. 2): A) the island model assumed by the Hudson F_{ST} estimator, and B) our admixture scenario, a non-island model constructed so $\hat{F}_{\mathrm{ST}} \approx \frac{1}{2} F_{\mathrm{ST}}$. The estimator limit of Eq. (15) (green dotted line) overlaps the true F_{ST} (red dashed line) in (A) but not (B). Estimates (blue) include 95% prediction intervals (too narrow to see) from 39 independently-simulated genotype matrices for each model (Appendix G).

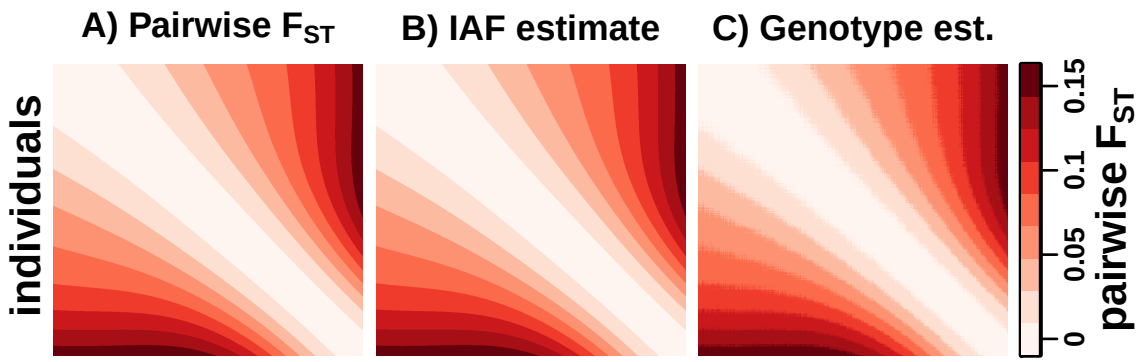

Figure 5: **Consistent individual-level pairwise F_{ST} estimates.** Consistency of our individual-level pairwise F_{ST} estimators is demonstrated in our admixture simulation. Plots show $n = 1000$ individuals along both axes, and color corresponds to F_{jk} between individuals j and k. A) True pairwise F_{ST} matrix. The pairwise F_{ST} measures the mean differentiation of each pair of individual from their last common ancestor, and is negatively correlated with coancestry. B) Estimate from IAFs. C) Estimate from genotypes.

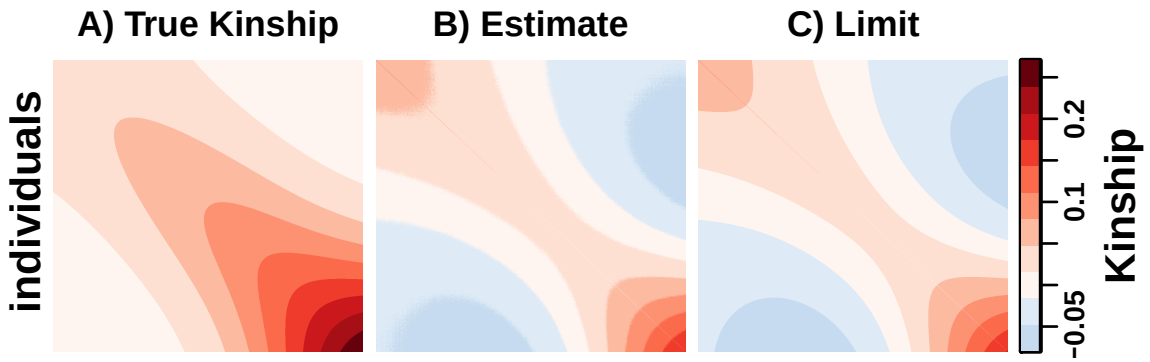

Figure 6: **Bias in kinship estimates.** Bias for the "standard" kinship coefficient estimator is illustrated in our admixture simulation. Plots show $n = 1000$ individuals along both axes, and color corresponds to φ^T_{jk} between individuals j and k, except the diagonal ($j = k$) shows $f^T_j = 2\varphi^T_{jj} - 1$ for a comparable scale. A) True kinship matrix. B) $\hat{\varphi}^T_{jk}$ of Eq. (23) estimated from simulated genotypes. C) Theoretical limit of $\hat{\varphi}^T_{jk}$ estimator of Eq. (24) as the number of independent SNPs goes to infinity.

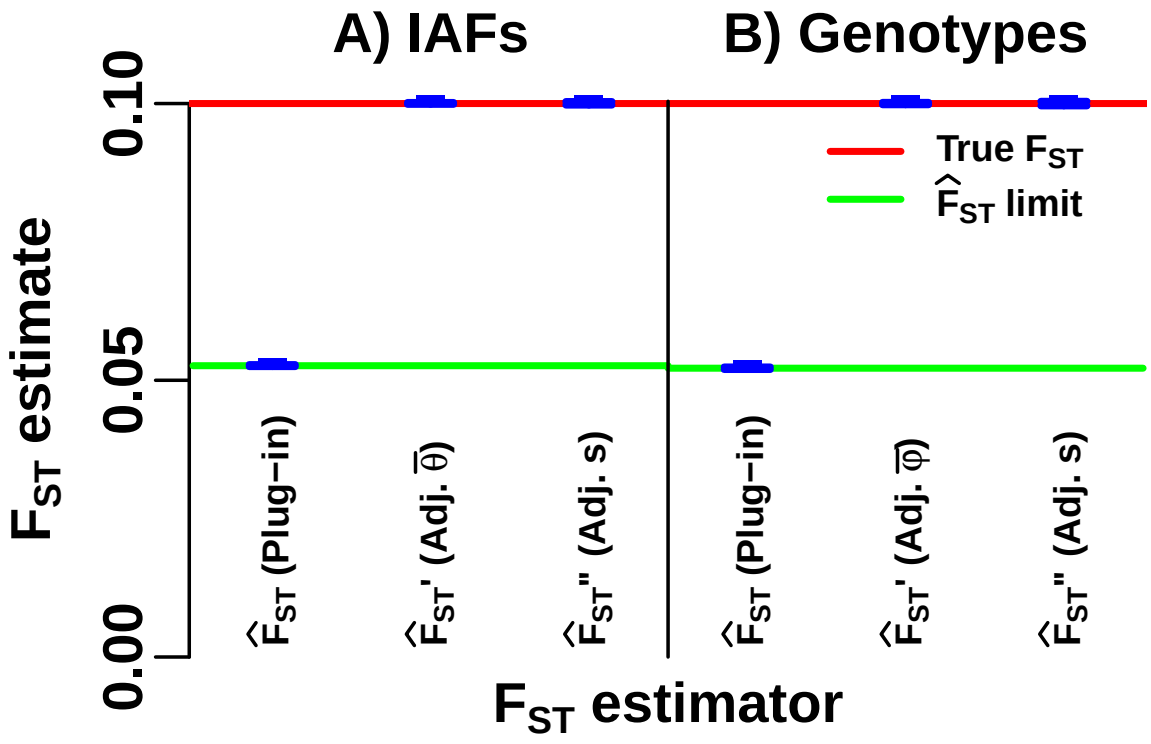

Figure 7: **Evaluation of plug-in and adjusted F_{ST} estimators.** The plug-in and adjusted F_{ST} estimators are evaluated using our admixture simulation. All adjusted estimators are "oracle" methods, since $\bar{\theta}^T, \bar{\varphi}^T, s$ are usually unknown. A) Estimation from IAFs: "plug-in" estimator is $\hat{F}_{\mathrm{ST}}$ from Eq. (31); "Adj. $\bar{\theta}^T$" is $\hat{F}'_{\mathrm{ST}}$ from Eq. (32); "Adj. s" is $\hat{F}''_{\mathrm{ST}}$ from Eq. (36). B) For genotypes, the "plug-in" estimator is given in Eq. (30), and the adjusted estimators use $\bar{\varphi}^T$ rather than $\bar{\theta}^T$. Lines: true F_{ST} (red dashed line), limits of biased estimators (green dotted lines, which differ slightly per panel). Estimates (blue) include 95% prediction intervals (too narrow to see) from 39 independently-simulated genotype matrices for our admixture model (Appendix G).

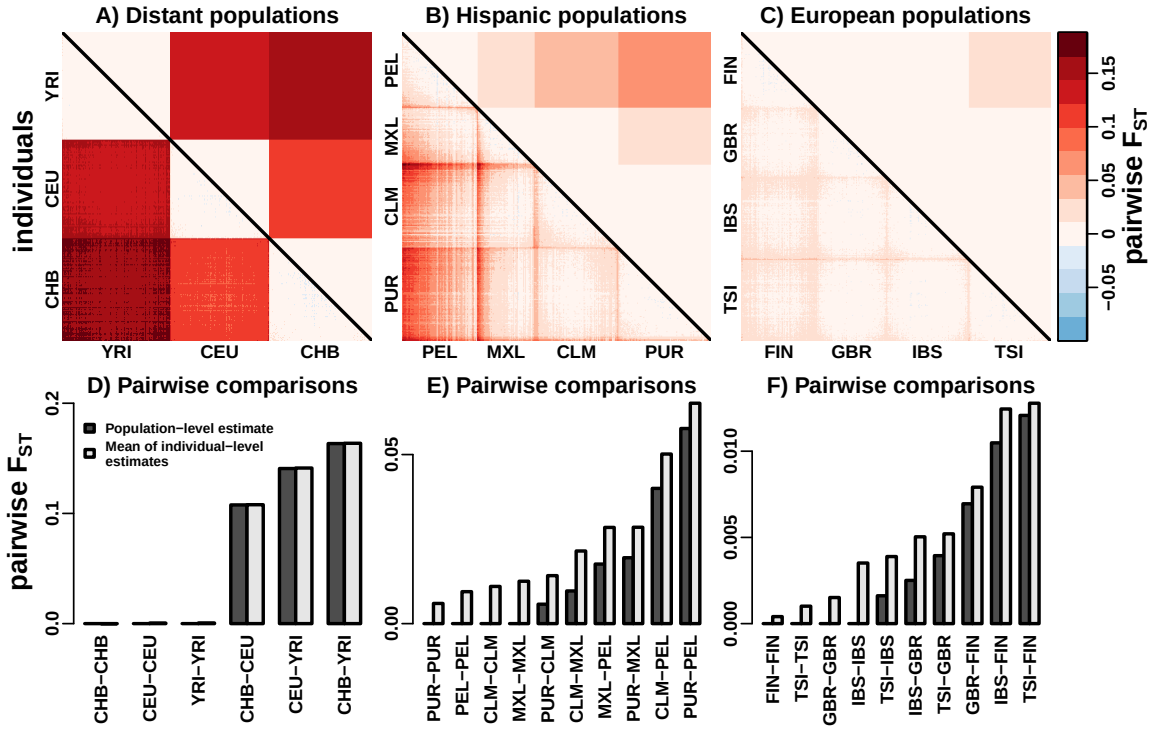

Figure 8: **Comparison of pairwise F_{ST} estimates on 1000 Genomes Project populations.** Comparison of "population-level" Hudson F_{ST} estimates (upper triangle of A-C) and our "individual-level" pairwise F_{ST} estimates (lower triangle of A-C). All SNPs were ascertained in YRI. In (A-C), individuals in each population were ordered using their individual-level pairwise F_{ST} submatrix and the "seriate" function of R package "seriation" with default options. A) Geographically distant populations (310 individuals) are well approximated by the island model. YRI: Yoruba in Ibadan, Nigeria; CEU: Utah Residents with Northern and Western European Ancestry; CHB: Han Chinese in Beijing, China. B) Hispanic populations (347 individuals) are structured due to variable individual admixture proportions from primarily Native American, European, and African populations. PEL: Peruvians from Lima, Peru; MXL: Mexican Ancestry from Los Angeles USA; CLM: Colombians from Medellin, Colombia; PUR: Puerto Ricans from Puerto Rico. C) European populations (404 individuals) are closely related and geographically proximal. TSI: Toscani in Italy; FIN: Finnish in Finland; GBR: British in England and Scotland; IBS: Iberian Population in Spain. D-F) Population-level (Hudson) F_{ST} estimates are uniformly smaller than $\hat{\bar{F}}_{uv}$ (see text) for all pairs of populations.

Appendices

A Accuracy of ratio estimators

A.1 Almost sure convergence of ratio-of-means estimators with independent and uniformly bounded terms

Here we prove that $\frac{\hat{A}_m}{\hat{B}_m} \xrightarrow[m\to\infty]{\text{a.s.}} \frac{A}{B}$, where $\hat{A}_m = \frac{1}{m}\sum_{i=1}^{m} a_i$ and $\hat{B}_m = \frac{1}{m}\sum_{i=1}^{m} b_i$ give the ratio-of-means estimator described in the main text. It suffices to prove $\hat{A}_m \xrightarrow[m\to\infty]{\text{a.s.}} Ac$ and $\hat{B}_m \xrightarrow[m\to\infty]{\text{a.s.}} Bc \neq 0$, from which the result follows using the continuous mapping theorem [48, 49]. The proof for $\hat{A}_m$ follows, which applies analogously to $\hat{B}_m$. Our a_i are independent but not identically distributed, since they depend on p_i^T that varies per SNP, so the standard law of large numbers does not apply to $\hat{A}_m$. We show almost sure convergence using Kolmogorov's criterion for the Strong Law of Large Numbers [50], which is satisfied for bounded $\mathrm{Var}(a_i)$. Since $|a_i| \leq C < \infty$ for all i and some C (see main text), then $\mathrm{E}[a_i^2] \leq C^2$, so $\mathrm{Var}(a_i) \leq C^2$. Therefore, $\hat{A}_m \xrightarrow[m\to\infty]{\text{a.s.}} \lim_{m\to\infty} \mathrm{E}\left[\hat{A}_m\right] = Ac$, as desired.

A.2 Order of error of expectations

The error of the ratio of expectations from the expectation of the ratio is given by

$$\epsilon_m = \mathrm{E}\left[\frac{\hat{A}_m}{\hat{B}_m}\right] - \frac{\mathrm{E}[\hat{A}_m]}{\mathrm{E}[\hat{B}_m]} = -\frac{\mathrm{Cov}\left(\frac{\hat{A}_m}{\hat{B}_m}, \hat{B}_m\right)}{\mathrm{E}\left[\hat{B}_m\right]} = -\frac{1}{m^2 Bc}\sum_{i=1}^{m}\sum_{j=1}^{m}\mathrm{Cov}\left(\frac{a_i}{\hat{B}_m}, b_j\right),$$

which follows from $\mathrm{Cov}(X,Y) = \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y]$ [51] and expanding the covariance. Previous work on ratio estimators [35, 51] assumes IID a_i and b_i, which does not hold for SNPs. Assuming independent SNPs ($\mathrm{Cov}(a_i, b_j) = 0$ for $i \neq j$) and large m so $\hat{B}_m \approx Bc$ is practically independent of any given a_i and b_j, then

$$\epsilon_m \approx -\frac{1}{mB^2c^2}\left[\frac{1}{m}\sum_{i=1}^{m}\mathrm{Cov}(a_i, b_i)\right].$$

Since a_i, b_i are bounded, $|\mathrm{Cov}(a_i, b_i)| \leq C^2$ for the same C of the previous section, so

$$|\epsilon_m| \leq \frac{C^2}{mB^2c^2},$$

holds for some large enough m and C. Hence $\epsilon_m = O\left(\frac{1}{m}\right)$ is as for standard ratio estimators [35].

B Generalized Hudson F_{ST} estimator

The Hudson F_{ST} estimator compares two populations [4]. We generalize this estimator for n independent populations, where F_{ST} equals the mean pairwise F_{ST} for every pair of populations. We

average numerators and denominators of the pairwise estimator before computing the ratio. Let j index the n populations, n_j be the number of individuals sampled from j, and $\hat{p}_{ij}$ be the sample allele frequency in j for SNP i, then

$$\bar{p}_i = \frac{1}{n}\sum_{j=1}^{n}\hat{p}_{ij},$$

$$\hat{\sigma}_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}(\hat{p}_{ij}-\bar{p}_i)^2,$$

$$\hat{F}_{\mathrm{ST}}^{\mathrm{Hudson}} = \frac{\sum_{i=1}^{m}\hat{\sigma}_i^2 - \frac{1}{n}\sum_{j=1}^{n}\frac{\hat{p}_{ij}(1-\hat{p}_{ij})}{2n_j-1}}{\sum_{i=1}^{m}\bar{p}_i(1-\bar{p}_i)+\frac{1}{n}\hat{\sigma}_i^2},$$

which consistently estimates F_{ST} in island models.

C Derivation of method of moment estimators

C.1 F_{ST} island model estimator

Assuming the coancestry model of Eqs. (6) and (7) for islands ($\theta_{jk}^T = 0$ for $j \neq k$), the first and second moments of the IAFs are:

$$\mathrm{E}[\pi_{ij}] = p_i^T, \tag{C.1}$$

$$\mathrm{E}\left[\pi_{ij}^2\right] = \left(p_i^T\right)^2 + p_i^T\left(1-p_i^T\right)\theta_{jj}^T, \tag{C.2}$$

$$\mathrm{E}\left[\pi_{ij}\pi_{ik}\right] = \left(p_i^T\right)^2 \quad \text{if} \quad j \neq k. \tag{C.3}$$

$F_{\mathrm{ST}} = \frac{1}{n}\sum_{j=1}^{n}\theta_{jj}^T$ appears by averaging Eq. (C.2) over j:

$$\mathrm{E}\left[\frac{1}{n}\sum_{j=1}^{n}\pi_{ij}^2\right] = \left(p_i^T\right)^2 + p_i^T\left(1-p_i^T\right)F_{\mathrm{ST}}. \tag{C.4}$$

Since Eq. (C.1) has the same value for every j, and Eq. (C.3) as well for every $j \neq k$, we average these to reduce estimation variance. The results are in terms of $\hat{p}_i^T = \frac{1}{n}\sum_{j=1}^{n}\pi_{ij}$:

$$\mathrm{E}\left[\hat{p}_i^T\right] = \mathrm{E}\left[\frac{1}{n}\sum_{j=1}^{n}\pi_{ij}\right] = p_i^T, \tag{C.5}$$

$$\mathrm{E}\left[\left(\hat{p}_i^T\right)^2\right] = \mathrm{E}\left[\frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n}\pi_{ij}\pi_{ik}\right] = \left(p_i^T\right)^2 + p_i^T\left(1-p_i^T\right)\frac{1}{n}F_{\mathrm{ST}}. \tag{C.6}$$

F_{ST} also appears in Eq. (C.6) because $j = k$ terms are introduced in the double sum. Subtracting Eq. (C.4) and Eq. (C.6) in turn from Eq. (C.5) results in:

$$\mathrm{E}\left[\hat{p}_i^T - \frac{1}{n}\sum_{j=1}^{n}\pi_{ij}^2\right] = p_i^T\left(1-p_i^T\right)\left(1-F_{\mathrm{ST}}\right),$$
$$\mathrm{E}\left[\hat{p}_i^T\left(1-\hat{p}_i^T\right)\right] = p_i^T\left(1-p_i^T\right)\left(1-\frac{1}{n}F_{\mathrm{ST}}\right).$$

To reduce variance further, we average across SNPs, giving

$$\mathrm{E}\left[\frac{1}{m}\sum_{i=1}^{m}\left(\hat{p}_i^T - \frac{1}{n}\sum_{j=1}^{n}\pi_{ij}^2\right)\right] = \overline{p(1-p)}^T\left(1-F_{\mathrm{ST}}\right),$$
$$\mathrm{E}\left[\frac{1}{m}\sum_{i=1}^{m}\hat{p}_i^T\left(1-\hat{p}_i^T\right)\right] = \overline{p(1-p)}^T\left(1-\frac{1}{n}F_{\mathrm{ST}}\right),$$

where $\overline{p(1-p)}^T = \frac{1}{m}\sum_{i=1}^{m}p_i^T\left(1-p_i^T\right)$. Eliminating $\overline{p(1-p)}^T$ and solving for F_{ST} in this system of equations results in the following F_{ST} estimator:

$$\hat{F}_{\mathrm{ST}} = \frac{\sum_{i=1}^{m}\left(\frac{1}{n}\sum_{j=1}^{n}\pi_{ij}^2 - \left(\hat{p}_i^T\right)^2\right)}{\sum_{i=1}^{m}\left(\hat{p}_i^T\left(1-\hat{p}_i^T\right) + \frac{1}{n}\left(\frac{1}{n}\sum_{j=1}^{n}\pi_{ij}^2 - \hat{p}_i^T\right)\right)} \tag{C.7}$$

This estimator is simplified noting that $\frac{1}{n}\sum_{j=1}^{n}\pi_{ij}^2$ appears in the IAF sample variance,

$$\hat{\sigma}_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(\pi_{ij} - \hat{p}_i^T\right)^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_{j=1}^{n}\pi_{ij}^2 - \left(\hat{p}_i^T\right)^2\right),$$

so substituting it into Eq. (C.7) recovers Eq. (14) as desired:

$$\hat{F}_{\mathrm{ST}} = \frac{\sum_{i=1}^{m}\hat{\sigma}_i^2}{\sum_{i=1}^{m}\hat{p}_i^T\left(1-\hat{p}_i^T\right) + \frac{1}{n}\hat{\sigma}_i^2}.$$

C.2 Standard kinship estimator

Here we assume the kinship model of Eqs. (1) and (2). Since Eq. (1) is the same for all individuals j, we average these first moments to reduce variance:

$$\mathrm{E}\left[\sum_{j=1}^{n}w_j x_{ij}\right] = 2p_i^T.$$

Each φ_{jk}^T appears once per (j,k) pair in Eq. (2), recast here in terms of the sample covariance:

$$\mathrm{E}\left[\left(x_{ij} - 2p_i^T\right)\left(x_{ik} - 2p_i^T\right)\right] = 4p_i^T\left(1-p_i^T\right)\varphi_{jk}^T.$$

Variance in the kinship estimate is reduced by averaging across SNPs, yielding:

$$\mathrm{E}\left[\frac{1}{m}\sum_{i=1}^{m}\left(x_{ij}-2p_i^T\right)\left(x_{ik}-2p_i^T\right)\right]=4\varphi_{jk}^T\frac{1}{m}\sum_{i=1}^{m}p_i^T\left(1-p_i^T\right). \tag{C.8}$$

The resulting estimator of p_i^T is

$$\hat{p}_i^T=\frac{1}{2}\sum_{j=1}^{n}w_j x_{ij},$$

which is plugged into Eq. (C.8) and then φ_{jk}^T is solved for, recovering Eq. (23) as desired:

$$\hat{\varphi}_{jk}^T=\frac{\sum_{i=1}^{m}\left(x_{ij}-2\hat{p}_i^T\right)\left(x_{ik}-2\hat{p}_i^T\right)}{4\sum_{i=1}^{m}\hat{p}_i^T\left(1-\hat{p}_i^T\right)}.$$

D Mean coancestry bounds

Here we prove that, for any weights such that $w_j>0, \sum_{j=1}^{n}w_j=1$,

$$0\le\bar{\theta}^T\le F_{\mathrm{ST}}\le 1$$

holds, and for uniform weights $\frac{1}{n}F_{\mathrm{ST}}\le\bar{\theta}^T$ also holds. Furthermore, $\bar{\theta}^T=F_{\mathrm{ST}}$ holds iff $\theta_{jk}^T=F_{\mathrm{ST}}$ for all (j,k), and $\bar{\theta}^T=\frac{1}{n}F_{\mathrm{ST}}$ holds for island models.

The Cauchy-Schwarz inequality for covariances implies $\theta_{jk}^T\le\sqrt{\theta_{jj}^T\theta_{kk}^T}$. Therefore,

$$\bar{\theta}^T=\sum_{j=1}^{n}\sum_{k=1}^{n}w_jw_k\theta_{jk}^T\le\left(\sum_{j=1}^{n}w_j\sqrt{\theta_{jj}^T}\right)^2\le\sum_{j=1}^{n}w_j\theta_{jj}^T=F_{\mathrm{ST}},$$

where the second inequality follows from Jensen's inequality, since x^2 is a convex function. Since $\theta_{jj}^T\le 1$, then $F_{\mathrm{ST}}\le 1$ as well. Equality in the second bound requires $\theta_{jj}^T=F_{\mathrm{ST}}$ for all j, and equality in the first bound requires $\theta_{jk}^T=\theta_{jj}^T$, so that $\bar{\theta}^T=F_{\mathrm{ST}}$ requires $\theta_{jk}^T=F_{\mathrm{ST}}$ for all (j,k). Since all $w_j,\theta_{jk}^T\ge 0$, then

$$0\le\sum_{j=1}^{n}w_j^2\theta_{jj}^T\le\bar{\theta}^T,$$

where the second inequality follows from dropping $j\ne k$ terms from the double sum of $\bar{\theta}^T$. The case $w_j=\frac{1}{n}$ gives $\frac{1}{n}F_{\mathrm{ST}}\le\bar{\theta}^T$, with equality for island models by construction.

E Moments of estimator building blocks

Here we calculate first and some second moments for "building block" quantities that recur in our estimators, particularly terms involving x_{ij} and $\hat{p}_i^T$, and which enable us to calculate the limits of

our estimators. Below are examples for genotypes, which follow from Eqs. (1) and (2); calculations for IAFs follow analogously from Eqs. (6) and (7) (not shown).

$$\begin{aligned}
\mathrm{E}\left[\hat{p}_i^T \middle| T\right] &= \mathrm{E}\left[\frac{1}{2}\sum_{j=1}^{n} w_j x_{ij} \middle| T\right] = \frac{1}{2}\sum_{j=1}^{n} w_j \,\mathrm{E}[x_{ij}|T] = \sum_{j=1}^{n} w_j p_i^T = p_i^T, \\
\mathrm{E}[x_{ij}x_{ik}|T] &= \mathrm{Cov}(x_{ij}, x_{ik}|T) + \mathrm{E}[x_{ij}|T]\,\mathrm{E}[x_{ik}|T] = 4\left(p_i^T\left(1-p_i^T\right)\varphi_{jk}^T + \left(p_i^T\right)^2\right), \\
\mathrm{E}\left[x_{ij}\hat{p}_i^T \middle| T\right] &= \mathrm{E}\left[\frac{1}{2}\sum_{k=1}^{n} w_j x_{ij} x_{ik} \middle| T\right] = \frac{1}{2}\sum_{k=1}^{n} w_j \,\mathrm{E}[x_{ij}x_{ik}|T] \\
&= 2\sum_{k=1}^{n} w_j \left(p_i^T\left(1-p_i^T\right)\varphi_{jk}^T + \left(p_i^T\right)^2\right) = 2\left(p_i^T\left(1-p_i^T\right)\bar{\varphi}_k^T + \left(p_i^T\right)^2\right), \\
\mathrm{Var}\left(\hat{p}_i^T \middle| T\right) &= \mathrm{Var}\left(\frac{1}{2}\sum_{j=1}^{n} w_j x_{ij} \middle| T\right) = \frac{1}{4}\sum_{j=1}^{n}\sum_{k=1}^{n} w_j w_k \,\mathrm{Cov}(x_{ij}, x_{ik}|T) = p_i^T\left(1-p_i^T\right)\bar{\varphi}^T, \\
\mathrm{E}\left[\left(\hat{p}_i^T\right)^2 \middle| T\right] &= \mathrm{Var}\left(\hat{p}_i^T|T\right) + \mathrm{E}\left[\hat{p}_i^T\right]^2 = p_i^T\left(1-p_i^T\right)\bar{\varphi}^T + \left(p_i^T\right)^2, \\
\mathrm{E}\left[\hat{p}_i^T\left(1-\hat{p}_i^T\right)|T\right] &= p_i^T\left(1-p_i^T\right)\left(1-\bar{\varphi}^T\right).
\end{aligned}$$

F Admixture and island model simulations

F.1 Construction of population island allele frequencies

We simulate $K = 10$ population islands and $m = 300,000$ independent SNPs. Every SNP i draws $p_i^T \sim \mathrm{Uniform}(0.01, 0.5)$. We set $f_{S_u}^T = \frac{u}{K}\tau$, where $\tau \leq 1$ tunes F_{ST}. For the island model, $F_{\mathrm{ST}} = \frac{1}{K}\sum_{u=1}^{K} f_{S_u}^T = \frac{\tau(K+1)}{2K}$, so $\tau = \frac{2KF_{\mathrm{ST}}}{K+1}$ gives the desired F_{ST} ($\tau \approx 0.18$ for $F_{\mathrm{ST}} = 0.1$). For the admixture model, τ is found numerically ($\tau \approx 0.90$ for $F_{\mathrm{ST}} = 0.1$; see last subsection). Lastly, $p_i^{S_u}$ are drawn from the Balding-Nichols distribution [41]:

$$p_i^{S_u}|T \sim \mathrm{Beta}\left(p_i^T\left(\frac{1}{f_{S_u}^T} - 1\right), \left(1-p_i^T\right)\left(\frac{1}{f_{S_u}^T} - 1\right)\right).$$

F.2 Random island sizes

We randomly generate samples sizes $\mathbf{r} = (r_u)$ for K islands and $\sum_{u=1}^{K} r_u = n = 1000$ individuals, as follows. First, draw $\mathbf{x} \sim \mathrm{Dirichlet}\,(1, ..., 1)$ of length K and $\mathbf{r} = \mathrm{round}(n\mathbf{x})$. While $\min_u r_u < \frac{n}{3K}$, draw a new $\mathbf{r}$, to prevent small islands (they do not occur in real data). Lastly, while $\delta = n - \sum_{u=1}^{K} r_u \neq 0$, a random u is updated to $r_u \leftarrow r_u + \mathrm{sgn}(\delta)$, which brings δ closer to zero. The resulting $\mathbf{r}$ is as desired. Weights for individuals j in S_u are $w_j = \frac{1}{Kr_u}$ so the generalized F_{ST} matches $F_{\mathrm{ST}} = \frac{1}{K}\sum_{u=1}^{K} f_{S_u}^T$ from the island model, which Hudson estimates [25].

F.3 Admixture proportions from 1D geography

We construct q_{ju} resulting from random-walk migrations along a one-dimensional geography. Let x_u be the coordinate of intermediate population u and y_j the coordinate of a modern individual j. We assume q_{ju} is proportional to $f(|x_u - y_j|)$, or

$$q_{ju} = \frac{f(|x_u - y_j|)}{\sum_{v=1}^{K} f(|x_v - y_j|)}.$$

where f is the Normal density function with $\mu = 0$ and tunable σ. The Normal density models random walks, where σ sets the spread of the populations (Fig. 6). Our simulation uses $x_u = u$ and $y_j = \frac{1}{2} + \frac{j-1}{n-1}K$, so intermediate population span $[1, K]$ and individuals span $[\frac{1}{2}, K + \frac{1}{2}]$. For the WC and Hudson F_{ST} estimators, individual j is assigned to the subpopulation S_u with the largest q_{ju} (Fig. 3D); thus these subpopulations have equal sample size, so $w_j = \frac{1}{n}$ is appropriate.

F.4 Choosing σ and τ

Here we find values for σ (controls q_{jk}) and τ (scales $f_{S_u}^T$) that give $s = \frac{1}{2}$ and $F_{\mathrm{ST}} = 0.1$ in the admixture model. We previously found that $\theta_{jk}^T = \sum_{u=1}^{K} q_{ju} q_{ku} f_{S_u}^T$ and $F_{\mathrm{ST}} = \sum_{j=1}^{n} \sum_{u=1}^{K} w_j q_{ju}^2 f_{S_u}^T$ holds for the BN-PSD model [25]. In our simulation, $w_j = \frac{1}{n}$ and $f_{S_u}^T = \frac{u}{K}\tau$ hold, so $\theta_{jk}^T = \frac{\tau}{K} \sum_{u=1}^{K} u q_{ju} q_{ku}$ and $F_{\mathrm{ST}} = \frac{\tau}{nK} \sum_{j=1}^{n} \sum_{u=1}^{K} u q_{ju}^2$. Therefore,

$$s = \frac{\bar{\theta}^T}{F_{\mathrm{ST}}} = \frac{1}{n} \frac{\sum_{u=1}^{K} u \left(\sum_{j=1}^{n} q_{ju}(\sigma) \right)^2}{\sum_{u=1}^{K} u \left(\sum_{j=1}^{n} q_{ju}^2(\sigma) \right)}$$

depends only on σ. A numerical root finder finds that $\sigma \approx 1.78$ gives $s = \frac{1}{2}$. For fixed q_{ju},

$$\tau = \frac{F_{\mathrm{ST}}}{\sum_{u=1}^{K} u \left(\frac{1}{n} \sum_{j=1}^{n} q_{ju}^2 \right)},$$

where F_{ST} is the desired value. $F_{\mathrm{ST}} = 0.1$ is achieved with $\tau \approx 0.901$.

G Prediction intervals of F_{ST} estimators

Prediction intervals with $\alpha = 95\%$ correspond to the range of $n = 39$ independent F_{ST} estimates. In the general case, n independent statistics are given in order $X_{(1)} < \ldots < X_{(n)}$. Then $I = [X_{(j)}, X_{(n+1-j)}]$ is a prediction interval with confidence $\alpha = \frac{n+1-2j}{n+1}$ [52]. In our case, $j = 1$ and $n = 39$ gives $\alpha = 0.95$, as desired. Each estimate was constructed from simulated data with the same dimensions and structure as before (fixed $f_{S_u}^T$ and q_{ju}; fixed sample sizes too for island models), but with $p_i^T, p_i^{S_u}, \pi_{ij}, x_{ij}$ drawn anew.