PLOS GENETICS

RESEARCH ARTICLE

Estimating F_{ST} and kinship for arbitrary population structures

Alejandro Ochoa[1,2], **John D. Storey**[3]*

1 Duke Center for Statistical Genetics and Genomics, Duke University, Durham, North Carolina, United States of America, **2** Department of Biostatistics and Bioinformatics, Duke University, Durham, North Carolina, United States of America, **3** Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, New Jersey, United States of America

* jstorey@princeton.edu

OPEN ACCESS

Citation: Ochoa A, Storey JD (2021) Estimating F_{ST} and kinship for arbitrary population structures. PLoS Genet 17(1): e1009241. https://doi.org/10.1371/journal.pgen.1009241

Editor: Marcus W. Feldman, Stanford University, UNITED STATES

Received: September 11, 2019

Accepted: November 2, 2020

Published: January 19, 2021

Data Availability Statement: The data and computer code for these manuscripts can be found at https://github.com/StoreyLab/human-differentiation-manuscript.

Funding: This research was supported in part by National Institutes of Health, National Human Genome Research Institute grant R01 HG006448 (JDS). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

F_{ST} and kinship are key parameters often estimated in modern population genetics studies in order to quantitatively characterize structure and relatedness. 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 independent subpopulations model of non-overlapping, independently evolving subpopulations. However, modern data sets have revealed that these simple models of structure likely do not hold in many populations, including humans. In this work, we analyze the behavior of these estimators in the presence of arbitrarily-complex population structures, which results in an improved estimation framework specifically designed for arbitrary population structures. After generalizing the definition of F_{ST} to arbitrary population structures and establishing a framework for assessing bias and consistency of genome-wide estimators, we calculate the accuracy of existing F_{ST} and kinship estimators under arbitrary population structures, characterizing biases and estimation challenges unobserved under their originally-assumed models of structure. We then present our new approach, which consistently estimates kinship and F_{ST} when the minimum kinship value in the dataset is estimated consistently. We illustrate our results using simulated genotypes from an admixture model, constructing a one-dimensional geographic scenario that departs nontrivially from the independent subpopulations model. Our simulations reveal the potential for severe biases in estimates of existing approaches that are overcome by our new framework. This work may significantly improve future analyses that rely on accurate kinship and F_{ST} estimates.

Author summary

Kinship coefficients and F_{ST}, which measure relatedness and population structure, respectively, are important quantities needed to accurately perform various analyses on genetic data, including genome-wide association studies and heritability estimation. However, existing estimators require restrictive assumptions of independence that are not met by real human and other datasets. In this work we find that existing estimators can be severely biased under reasonable scenarios, first by theoretically determining their properties, and then using an admixture simulation to illustrate our findings. In particular, we find that existing F_{ST} estimators are downwardly biased, and that existing kinship matrix estimators have related biases that are on average downward and of similar magnitude but vary for every pair of individuals. These insights led us to a new estimation framework for kinship and F_{ST} that is practically unbiased for any population structure, as demonstrated by theory and simulations. Our new approaches—available as open-source R packages—are easy to use and are more widely applicable than existing approaches, and they are likely to improve downstream analyses that require accurate kinship and F_{ST} estimates.

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_{ST} and kinship. F_{ST} is a parameter that measures structure in a subdivided population, satisfying $F_{\mathrm{ST}} = 0$ for an unstructured population and $F_{\mathrm{ST}} = 1$ if every locus has become fixed for some allele in each subpopulation. More generally, F_{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_{ST}, since the mean kinship of the parents in a subpopulation is the F_{ST} of the following generation [1].

This work focuses on the estimation of F_{ST} and kinship from biallelic single-nucleotide polymorphism (SNP) data. Existing estimators can be classified into parametric estimators (methods that require a likelihood function) and non-parametric estimators (such as the method-of-moments estimators we focus on, which only require low-order moment equations). There are many likelihood approaches that estimate F_{ST} and kinship, but these are limited by assuming independent subpopulations or Normal approximations for F_{ST} [3–11] or totally outbred individuals for kinship [12, 13]. Additionally, more complete likelihood models such as that of Jacquard [14] are underdetermined for biallelic loci [15]. Non-parametric approaches such as those based on the method of moments are considerably more flexible and computationally tractable [16], so they are the natural choice to study arbitrary population structures.

The most frequently-used F_{ST} estimators are derived and justified under the "independent subpopulations model," in which non-overlapping subpopulations evolved independently by splitting all at the same time from a common ancestral population. The Weir-Cockerham (WC) F_{ST} estimator assumes subpopulations of differing sample sizes and equal per-subpopulation F_{ST} relative to the common ancestral population [17]. The Weir-Hill F_{ST} estimator generalized WC for subpopulations with different F_{ST} values, and first considered arbitrary coancestry between subpopulations, resulting in estimates of a linearly-transformed F_{ST}, namely $(F_{\mathrm{ST}} - \tilde{\theta})/(1 - \tilde{\theta})$ (where $\tilde{\theta}$ is the unknown mean coancestry value between subpopulations) [4, 18, 19]. Weir-Hill has further evolved into the Weir-Goudet approach, incorporating relatedness for subpopulations and individuals based on allele matching, also estimating a linearly-transformed F_{ST} [20–22]. Note that the Weir-Hill and Weir-Goudet approaches intended to estimate such linearly-transformed quantities, which may be negative, and they did not aim to estimate IBD probabilities [4, 18–22]; in contrast, our goal is to estimate IBD probabilities, which must be non-negative and valid probabilities. The "Hudson" F_{ST} estimator [23] assumes two subpopulations with different F_{ST} values. All of the previous F_{ST} estimators are ratio estimators derived using the method of moments to have unbiased numerators and denominators, which gives approximately unbiased ratio estimates when their assumptions are met [4, 17, 23]. We also evaluate BayeScan [10], which estimates population-specific F_{ST} values using a Bayesian model and the Dirichlet-Multinomial likelihood function—thus representing non-method-of-moments approaches—but which like other existing F_{ST} estimators also assumes that subpopulations are non-overlapping and evolve independently. These F_{ST} estimators are important contributions, used widely in the field.

Kinship coefficients are now commonly calculated in population genetics studies to capture structure and relatedness. Kinship is utilized in principal components analyses and linear-mixed effects models to correct for structure in Genome-Wide Association Studies (GWAS) [16, 24–30] and to estimate genome-wide heritability [31, 32]. Often absent in previous models is a clear identification and role of the ancestral population T that sets the scale of the kinship estimates used. Omission of T makes sense when kinship is estimated on an unstructured population (where only a few individual pairs are closely related; there T is the current population). Our more complete notation brings T to the fore and highlights its key role in kinship estimation and its applications. The most commonly-used kinship estimator [16, 27, 30–36] is also a method-of-moments estimator whose operating characteristics are largely unknown in the presence of structure. We show here that this popular estimator is accurate only when the average kinship is zero, which implies that the population must be unstructured.

The goal of our work is to consistently estimate IBD probabilities, namely kinship coefficients and F_{ST}, for which there are currently no consistent estimators under general relatedness. Estimation of these as probabilities, as opposed to linearly-transformed quantities that may be negative, is important since the probabilistic definition of these parameters was required to derive their fundamental connections to many applications in genetics, including allele fixation [1, 2, 37], DNA forensics [3], and heritability [38, 39]. Although IBD probabilities are not absolute, but rather depend on the choice of ancestral population [40], their values become fixed upon agreeing to estimate them in terms of the Most Recent Common Ancestor (MRCA) population, which has long been the choice for models of F_{ST} [17, 23, 41] and kinship estimation from pedigrees [42, 43] or markers [12, 13].

Recent genome-wide studies have revealed that humans and other natural populations are structured in a complex manner that break the assumptions of the above estimators. Such complex population structures has been observed in several large human studies, such as the Human Genome Diversity Project [44, 45], the 1000 Genomes Project [46], Human Origins [47–49], and other contemporary [50–54] and archaic populations [55, 56]. We have also demonstrated that the global human population has a complex kinship matrix and no independent subpopulations [57–59]. 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_{ST} and kinship estimators in the presence of arbitrary forms of structure, which leads to a new estimation strategy for F_{ST} and kinship.

Here, we study existing F_{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, in section **The generalized F_{ST} for arbitrary population structures** we present the generalized definition of F_{ST} for arbitrary population structures [57]. In section **The kinship and coancestry models** we review the kinship model for genotype covariance [1, 14] and the coancestry model for individual-specific allele frequencies [57, 60, 61]. In section **Assessing the accuracy of genome-wide ratio estimators** we obtain new strong convergence results for a family of ratio estimators that includes the most common F_{ST} and kinship estimators. Next,

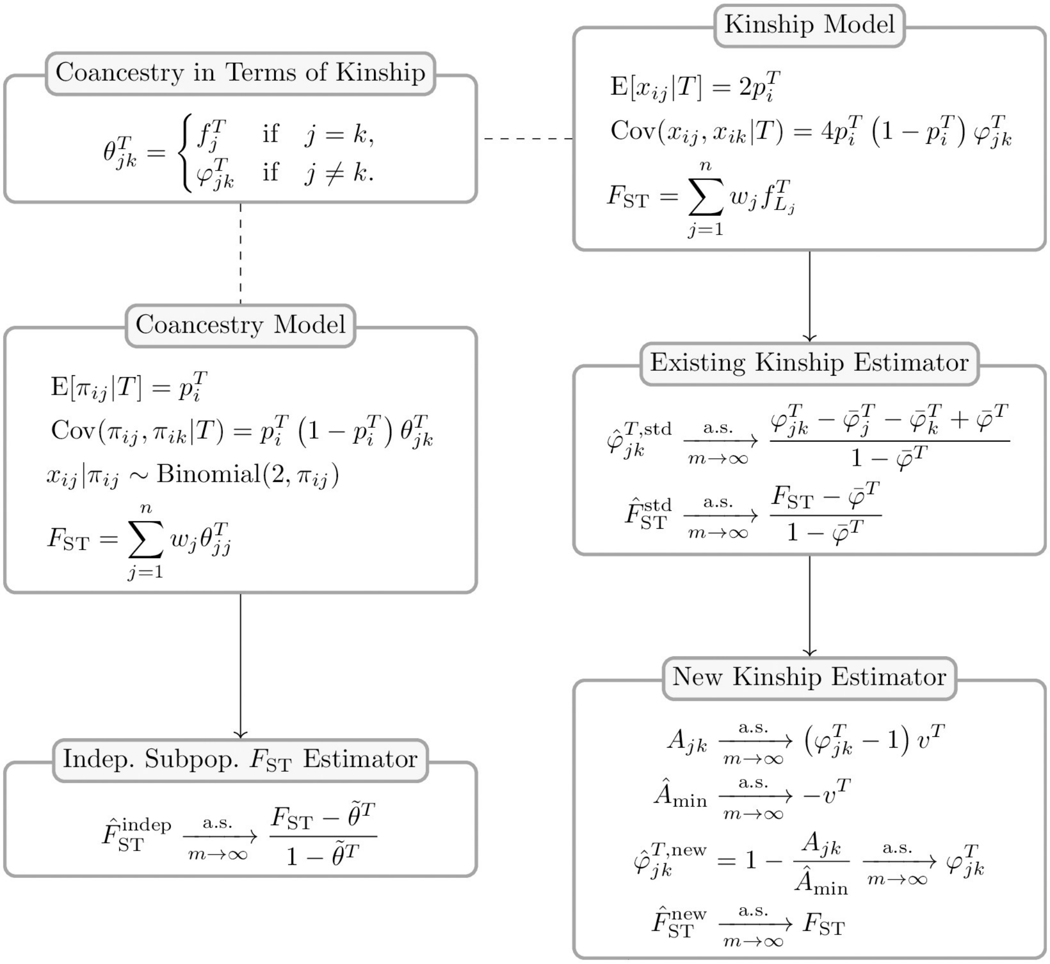
Coancestry in Terms of Kinship
$\theta^T_{jk} = \begin{cases} f^T_j & \text{if } j = k, \\ \varphi^T_{jk} & \text{if } j \neq k. \end{cases}$
Kinship Model
$\mathrm{E}[x_{ij}|T] = 2p^T_i$
$\mathrm{Cov}(x_{ij}, x_{ik}|T) = 4p^T_i \left(1 - p^T_i\right) \varphi^T_{jk}$
$F_{\mathrm{ST}} = \sum_{j=1}^{n} w_j f^T_{L_j}$
Coancestry Model
$\mathrm{E}[\pi_{ij}|T] = p^T_i$
$\mathrm{Cov}(\pi_{ij}, \pi_{ik}|T) = p^T_i \left(1 - p^T_i\right) \theta^T_{jk}$
$x_{ij}|\pi_{ij} \sim \mathrm{Binomial}(2, \pi_{ij})$
$F_{\mathrm{ST}} = \sum_{j=1}^{n} w_j \theta^T_{jj}$
Existing Kinship Estimator
$\hat{\varphi}^{T,\mathrm{std}}_{jk} \xrightarrow[m\to\infty]{\text{a.s.}} \frac{\varphi^T_{jk} - \bar{\varphi}^T_j - \bar{\varphi}^T_k + \bar{\varphi}^T}{1 - \bar{\varphi}^T}$
$\hat{F}^{\mathrm{std}}_{\mathrm{ST}} \xrightarrow[m\to\infty]{\text{a.s.}} \frac{F_{\mathrm{ST}} - \bar{\varphi}^T}{1 - \bar{\varphi}^T}$
Indep. Subpop. F_{ST} Estimator
$\hat{F}^{\mathrm{indep}}_{\mathrm{ST}} \xrightarrow[m\to\infty]{\text{a.s.}} \frac{F_{\mathrm{ST}} - \tilde{\theta}^T}{1 - \tilde{\theta}^T}$
New Kinship Estimator
$A_{jk} \xrightarrow[m\to\infty]{\text{a.s.}} \left(\varphi^T_{jk} - 1\right) v^T$
$\hat{A}_{\min} \xrightarrow[m\to\infty]{\text{a.s.}} -v^T$
$\hat{\varphi}^{T,\mathrm{new}}_{jk} = 1 - \frac{A_{jk}}{\hat{A}_{\min}} \xrightarrow[m\to\infty]{\text{a.s.}} \varphi^T_{jk}$
$\hat{F}^{\mathrm{new}}_{\mathrm{ST}} \xrightarrow[m\to\infty]{\text{a.s.}} F_{\mathrm{ST}}$

Fig 1. Accuracy of F_{ST} and kinship estimators: Overview of models and results. Our analysis is based on the generalized F_{ST} definition (section **The generalized F_{ST} for arbitrary population structures**) and two parallel models: the "Coancestry Model" for individual-specific allele frequencies (π_{ij}), and the "Kinship Model" for genotypes (x_{ij}). The "Coancestry in Terms of Kinship" panel connects kinship (φ^T_{jk}, f^T_j) and coancestry (θ^T_{jk}) parameters (section **The kinship and coancestry models**). We use these models to study the accuracy of F_{ST} and kinship method-of-moment estimators under arbitrary population structures. The "Indep. Subpop. F_{ST} Estimator" panel shows the bias resulting from the misapplication of F_{ST} estimators for independent subpopulations ($\hat{F}^{\mathrm{indep}}_{\mathrm{ST}}$) to arbitrary structures (section **F_{ST} estimation based on the independent subpopulations model**), as calculated under the coancestry model. The "Existing Kinship Estimator" panel shows the bias in the standard kinship model estimator ($\hat{\varphi}^{T,\mathrm{std}}_{jk}$) and its resulting plug-in F_{ST} estimator ($\hat{F}^{\mathrm{std}}_{\mathrm{ST}}$; section **Characterizing a kinship estimator and its relationship to F_{ST}**), as calculated under the kinship model. The "New Kinship Estimator" panel presents a new statistic A_{jk} that estimates kinship with a uniform bias, which together with a consistent estimator of its minimum value ($\hat{A}_{\min}$) results in our new kinship ($\hat{\varphi}^{T,\mathrm{new}}_{jk}$) and F_{ST} ($\hat{F}^{\mathrm{new}}_{\mathrm{ST}}$) estimators, which are consistent under arbitrary population structure (section **A new approach for kinship and F_{ST} estimation**).

https://doi.org/10.1371/journal.pgen.1009241.g001

we calculate the convergence values of these F_{ST} (section **F_{ST} estimation based on the independent subpopulations model**) and kinship (section **Characterizing a kinship estimator and its relationship to F_{ST}**) estimators under arbitrary population structures, where we find biases that are not present under their original assumptions about structure (panels "Indep. Subpop. F_{ST} Estimator" and "Existing Kinship Estimator" in Fig 1). We characterize the limit of the standard kinship estimator, identifying complex biases or distortions, in agreement with recent work [21, 62].

In section **A new approach for kinship and F_{ST} estimation** we introduce a new approach for kinship and F_{ST} estimation for arbitrary population structures, and demonstrate the improved performance using a simple implementation of these estimators (panel "New Kinship Estimator" in Fig 1). There are two key innovations. First, based on the method of moments, we derive a statistic that estimates kinship coefficients up to a shared unknown scaling factor. Second, we propose a new condition, the identification of unrelated individual pairs in the data, which yields the value of the unknown scaling factor and enables the consistent estimation of kinship matrices and F_{ST}. We present a simple implementation of this second estimator, based on taking the minimum average statistic value between subpopulations, which in section **Simulations evaluating F_{ST} and kinship estimators** is shown to perform well under some misspecification, namely in an admixture scenario that does not actually have subpopulations [63–65]. Elsewhere, we analyze the Human Origins and 1000 Genomes Project datasets with our novel kinship and F_{ST} estimation approach, where we demonstrate its coherence with the African Origins model, and illustrate the shortcomings of previous approaches in these complex data [59]. In summary, we identify a new approach for unbiased estimation of F_{ST} and kinship, and we provide new estimators that are nearly unbiased.

Results

The generalized F_{ST} for arbitrary population structures

The existing F_{ST} definition requires individuals to belong to discrete, non-overlapping subpopulations, so it must be generalized in order to apply to arbitrary population structures (such as the admixture model with individual-specific ancestry proportions considered in our simulations). Our generalized F_{ST} can be understood as a two-step strategy: (1) we define F_{ST} on a per-individual basis, and (2) we define F_{ST} for a group of individuals as a weighted average of the per-individual F_{ST} values [57].

The inbreeding coefficient f_j^T of an individual j relative to an ancestral population T is defined as the probability that the two alleles at a random locus are *identical by descent* (IBD) [37]. Note that the ancestral population T determines what is IBD: only relationships since T count toward IBD. This *total* inbreeding coefficient (f_j^T) is the individual analog of Wright's total inbreeding coefficient F_{IT}, the latter of which is the mean f_j^T over a group of individuals [2]. Wright partitioned *total* inbreeding (F_{IT}) into *local* (F_{IS}) and *structural* (F_{ST}) coefficients defined by a subpopulation S that contains all individuals in question and evolved from the ancestral population T, so that F_{IS} is the inbreeding of individuals relative to S (as opposed to T) and F_{ST} is inbreeding of the subpopulation S relative to T, and these coefficients satisfy $(1 - F_{\mathrm{IT}}) = (1 - F_{\mathrm{IS}})(1 - F_{\mathrm{ST}})$ [2]. In our generalized definitions for one individual j, we restrict the subpopulation of interest (S) to be L_j, called the local subpopulation of j, which is the most recent subpopulation from which j drew its alleles. In this case, $f_j^{L_j}$ is the *local* inbreeding coefficient of j (always relative to its local subpopulation L_j), and $f_{L_j}^T$ is the *structural* inbreeding coefficient of j (equal to the inbreeding of the subpopulation L_j relative to T), and being a special case of Wright's equation, they also satisfy [57]

$$(1 - f_j^T) = (1 - f_j^{L_j})(1 - f_{L_j}^T). \tag{1}$$

Now we discuss estimating the three quantities we just introduced. First, the total inbreeding coefficient (f_j^T) should be estimated from the variance of genotypes, using the practically unbiased approach we introduce in this work. Second, note that the local inbreeding coefficient ($f_j^{L_j}$) corresponds to (non-population) family relatedness, so it can be taken to be the inbreeding calculated from a pedigree if it is available [42]. Note that estimation of the various inbreeding coefficients from pedigrees was the only approach available to Wright when he studied cattle and defined inbreeding and F_{ST} [2, 37]. Alternatively, in the absence of pedigrees, local inbreeding can be estimated from inferred self-IBD blocks or unusually-large runs of homozygosity [66–68]. Lastly, since the structural inbreeding coefficient ($f_{L_j}^T$) is given by the previous two quantities (solving from Eq (1)) by

$$f_{L_j}^T = \frac{f_j^T - f_j^{L_j}}{1 - f_j^{L_j}}, \tag{2}$$

then we propose estimating $f_{L_j}^T$ using this equation, from the above estimates of f_j^T and $f_j^{L_j}$.

As a toy example, suppose we estimate a total inbreeding coefficient of $f_j^T = 0.15$ for a given individual whose parents are first cousins, then the pedigree expectation for its local inbreeding is $f_j^{L_j} = \frac{1}{16} = 0.0625$, and the structural inbreeding (i.e. the F_{ST} of this individual) using Eq (2) is $f_{L_j}^T \approx 0.093$. However, if in the same example ($f_j^T = 0.15$) the individual instead had parents who were second cousins, then $f_j^{L_j} = \frac{1}{64} \approx 0.0156$, then the structural estimate becomes $f_{L_j}^T \approx 0.137$, which is much closer to the total inbreeding value. Thus, when total inbreeding estimates are much larger than local inbreeding estimates, correcting for the latter via Eq (2) may not change the numerical estimate of structural inbreeding by a meaningful amount. Conversely, as the local inbreeding coefficient is reduced exponentially with the degree of relatedness of the parents ($f_j^{L_j} = \frac{1}{4^{n+1}}$ for n-th cousins), and as local inbreeding is required to be recent (to exclude population-level inbreeding), then sufficiently-accurate estimates of structural inbreeding can be obtained by estimating non-zero local inbreeding only for individuals with the most related parent pairs (above a certain degree of relatedness).

We define the generalized F_{ST} across n individuals as the weighted average of the per-individual structural inbreeding coefficients (*i.e.*, individual F_{ST} values) [57],

$$F_{\mathrm{ST}} = \sum_{j=1}^{n} w_j f_{L_j}^T, \tag{3}$$

where w_j is the weight of individual j and the weights are required to sum to one and be non-negative. The above is a straightforward generalization of Wright's F_{ST}: if every individual j has $L_j = S$ as its local subpopulation, then Eq (3) becomes $F_{\mathrm{ST}} = \sum_{j=1}^{n} w_j f_S^T = f_S^T$, where f_S^T is the inbreeding coefficient of subpopulation S relative to T, so it has the same meaning as Wright's F_{ST} (the exact weights here do not matter as long as $\sum_{j=1}^{n} w_j = 1$, as required). Moreover, if each individual j belongs to one of K subpopulations S_u ($u \in \{1, \ldots, K\}$) and if subpopulations are weighted equally ($\sum_{j \in S_u} w_j = \frac{1}{K}$ for every S_u), then Eq (3) becomes $F_{\mathrm{ST}} = \frac{1}{K} \sum_{u=1}^{K} f_{S_u}^T$, so it equals the (unweighted) average subpopulation-specific F_{ST} (*i.e.*, $f_{S_u}^T$), which is the F_{ST} definition for multiple subpopulations prevalent in modern work [4, 21, 23]. The last case illustrates the need for weights, which above downweights individuals that belong to subpopulations with greater numbers of observations. In general, weights allow adjustment for skewed or unbalanced samples. However, in complicated scenarios without subpopulations and no obvious sampling biases, for simplicity we recommend using uniform weights ($w_j = \frac{1}{n}$) for the target generalized F_{ST}.

In terms of total and local inbreeding coefficients (using Eq (2)), the generalized F_{ST} equals

$$F_{\text{ST}} = \sum_{j=1}^{n} w_j \frac{f_j^T - f_j^{L_j}}{1 - f_j^{L_j}},$$

which immediately suggests the estimation strategy when estimates of the total and local inbreeding coefficients are available. For simplicity, in the remainder of this work we shall consider only locally-outbred individuals ($f_j^{L_j} = 0$ for all j), for which the generalized F_{ST} simply equals the weighted mean total inbreeding coefficient:

$$F_{\text{ST}} = \sum_{j=1}^{n} w_j f_j^T. \tag{4}$$

This greatly simplifies our discussion of bias for all of the F_{ST} estimators we analyzed; determining the statistical properties of local inbreeding estimators is beyond the scope of this work. Moreover, the assumption of locally-outbred individuals is satisfied in all of the simulations presented in this work.

The kinship and coancestry models

The generalized F_{ST} above is given solely in terms of inbreeding coefficients. In order to establish our results and framework, it is necessary to consider kinship coefficients as well. The kinship coefficient is the extension of the inbreeding coefficient for a pair of individuals: the kinship coefficient φ_{jk}^T of two individuals j and k relative to an ancestral population T is the probability that two alleles, chosen at random from each individual at a random locus, are IBD [1]. Note that the self-kinship coefficient is related to the inbreeding coefficient by $\varphi_{jj}^T = \frac{1}{2}\left(1 + f_j^T\right)$ [16].

Kinship coefficients determine the covariance structure of genotypes, which is the key to estimating kinship and F_{ST} from genotype data. We shall concentrate on biallelic variants, which include single-nucleotide polymorphisms, and are the dominant data from genotyping microarrays and whole-genome sequencing studies. We shall also restrict our attention to diploid organisms in this present work. Genotypes are encoded into variables x_{ij} for each locus i and individual j that count the number of alleles (dosage) of a given reference type, so for diploid organisms $x_{ij} = 2$ is homozygous for the reference allele, $x_{ij} = 0$ is homozygous for the alternative allele, and $x_{ij} = 1$ is heterozygous. Based on the definition of the IBD probabilities, the kinship model determines the mean and covariance structure of the genotype random variables at neutral loci [1, 2, 14, 16, 37]:

$$\begin{aligned} \mathrm{E}[x_{ij}|T] &= 2p_i^T, \\ \mathrm{Cov}(x_{ij}, x_{ik}|T) &= 4p_i^T(1 - p_i^T)\varphi_{jk}^T, \end{aligned} \tag{5}$$

where p_i^T is the allele frequency at locus i in the ancestral population T and $0 < p_i^T < 1$.

The coancestry model resembles the kinship model, but it is formulated in terms of allele frequencies, which simplifies our analysis of F_{ST} estimators for subpopulations as well as yielding kinship coefficients under the admixture model we simulate from in this work. Let π_{ij} be the *individual-specific allele frequency* (IAF) at locus i for individual j, which is a real number between zero and one [60, 61]. Our coancestry model assumes that [57]

$$\begin{aligned} \mathrm{E}[\pi_{ij}|T] &= p_i^T, \\ \mathrm{Cov}(\pi_{ij}, \pi_{ik}|T) &= p_i^T(1-p_i^T)\theta_{jk}^T, \end{aligned} \tag{6}$$

where θ_{jk}^T is the coancestry coefficient between individuals j and k relative to the ancestral population T. This model is inspired by coancestry models for subpopulations common in the F_{ST} literature [4, 5, 21, 23], and exactly equals those models when subpopulation sizes go to infinity, in which case j and k index subpopulations rather than individuals, and π_{ij} is interpreted as the true allele frequency at locus i for subpopulation j.

The coancestry model connects to the kinship model under the additional assumption that the alleles of an individual j are drawn independently from its IAF,

$$x_{ij}|\pi_{ij} \sim \mathrm{Binomial}(2, \pi_{ij}). \tag{7}$$

In this case, marginalizing the intermediate IAF random variables (π_{ij}) and matching the resulting genotype moments results in the following equivalence [57]:

$$\theta_{jk}^T = \begin{cases} f_j^T & \text{if} \quad j = k, \\ \varphi_{jk}^T & \text{if} \quad j \neq k. \end{cases} \tag{8}$$

The coancestry coefficient equals the kinship coefficient between two different individuals, but the self-coancestry coefficient equals the inbreeding coefficient (rather than the self-kinship coefficient). However, since in the coancestry model alleles are drawn independently conditional on the IAF in Eq (7), then the only structure present is the population structure, so these coancestry models cannot generate family structures, unlike the more general kinship model that also encompasses pedigrees. Therefore, despite Eq (8), the kinship and coancestry are not equivalent models except under the more restrictive assumptions of the coancestry model. Thus, individuals drawn from this model are always locally-outbred, so $\theta_{jj}^T = f_{L_j}^T$ also equals the structural inbreeding coefficient, and the generalized F_{ST} under the coancestry model is therefore

$$F_{\text{ST}} = \sum_{j=1}^{n} w_j \theta_{jj}^T, \tag{9}$$

which also generalizes previous definitions of F_{ST} under coancestry for subpopulations [4, 5, 21, 23]. The kinship and coancestry models, and their connection, is included in the overview Fig 1.

Assessing the accuracy of genome-wide ratio estimators

In this section we change gears to focus on theoretical convergence properties of two broad estimator families. The resulting theory will be applied repeatedly to various F_{ST} and kinship estimators of interest in later sections.

Many F_{ST} and kinship coefficient method-of-moments estimators are *ratio estimators*, a general class of estimators that tends to be biased and to have no closed-form expectation [69]. In the F_{ST} literature, the expectation of a ratio is frequently approximated with a ratio of expectations [4, 17, 23]. Specifically, ratio estimators are often called “unbiased” if the ratio of expectations is unbiased, even though the ratio estimator itself may be biased [69]. Here we characterize the behavior of two ratio estimator families calculated from genome-wide data, known as "ratio-of-means" and "mean-of-ratios" estimators [23], detailing conditions where the previous approximation is justified and providing additional criteria to assess the accuracy of such estimators.

Ratio estimators. The general problem of forming ratio estimators involves random variables a_i and b_i calculated from genotypes at each locus 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 loci (given by F_{ST} or φ_{jk}^{T}), while c_i depends on the ancestral allele frequency p_i^T and varies per locus. The problem is that the single-locus estimator $\frac{a_i}{b_i}$ is biased, since $\mathrm{E}\left[\frac{a_i}{b_i}\right] \neq \frac{\mathrm{E}[a_i]}{\mathrm{E}[b_i]} = \frac{A}{B}$, which applies to ratio estimators in general [69]. Below we study two estimator families that combine large numbers of loci to better estimate $\frac{A}{B}$.

Convergence. 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 [4, 17, 19, 23, 70]. Note that $\mathrm{E}[\hat{A}_m] = A\bar{c}_m$ and $\mathrm{E}[\hat{B}_m] = B\bar{c}_m$, 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 $\bar{c}_m \rightarrow c$, and $Bc \neq 0$, which are satisfied by common estimators. Given independent loci, we prove almost sure convergence to the desired quantity (S1 Text),

$$\frac{\hat{A}_m}{\hat{B}_m} = \frac{\frac{1}{m}\sum_{i=1}^{m} a_i}{\frac{1}{m}\sum_{i=1}^{m} b_i} \xrightarrow[m\rightarrow\infty]{\text{a.s.}} \frac{A}{B}, \tag{10}$$

a strong result that implies $\mathrm{E}\left[\frac{\hat{A}_m}{\hat{B}_m}\right] \rightarrow \frac{A}{B}$, justifying previous work [4, 17, 23]. Moreover, the error between these expectations scales with $\frac{1}{m}$ (S1 Text), just as for standard ratio estimators [69]. Although real loci are not independent due to genetic linkage, their dependence is very localized, so this estimator will perform well if the effective number of independent loci is large.

In order to test if a given ratio-of-means estimator converges to its ratio of expectations as in Eq (10), the following three conditions can be tested. (i) The expected values of each term a_i, b_i must be calculated and shown to be of the form $\mathrm{E}[a_i] = Ac_i$ and $\mathrm{E}[b_i] = Bc_i$ for some A and B shared by all loci i and some c_i that may vary per locus i but must be shared by both $\mathrm{E}[a_i]$, $\mathrm{E}[b_i]$. In the estimators we study, A and B are functions of IBD probabilities such as φ_{jk}^{T} and F_{ST}, while c_i is a function of p_i^T only. (ii) The mean c_i must converge to a non-zero value for infinite loci. (iii) Both $|a_i|, |b_i| \leq C$ must be bounded for all i by some finite C (the estimators we study usually have $C = 1$ or $C = 4$). If these conditions are satisfied, then Eq (10) holds for independent loci and the A and B found in the first step. See the next section for an example application of this procedure to an F_{ST} estimator.

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 [16, 27, 30–35] and F_{ST} [46]. If each $\frac{a_i}{b_i}$ is biased, their average across loci will also be biased, even as $m \rightarrow \infty$. However, if $\mathrm{E}\left[\frac{a_i}{b_i}\right] \rightarrow \frac{A}{B}$ for all loci $i = 1, \ldots, m$ as the number of individuals $n \rightarrow \infty$, and $\mathrm{Var}\left(\frac{a_i}{b_i}\right)$ is bounded, then

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

Therefore, mean-of-ratios estimators must satisfy more restrictive conditions than ratio-of-means estimators, as well as large n (in addition to the large m needed by both estimators), to estimate $\frac{A}{B}$ well. We do not provide a procedure to test whether a given mean-of-ratios estimator converges as shown above.

F_{ST} estimation based on the independent subpopulations model

Now that we have detailed how ratio estimators may be evaluated for their accuracy, we turn to existing estimators and assess their accuracy under arbitrary population structures. We study the F_{ST} estimators Weir-Cockerham (WC) [17], Weir-Hill [4], "Hudson" [23], and Weir-Goudet (equals HudsonK below for biallelic loci; S1 Text) [21]. The panel "Indep. Subpop. F_{ST} Estimator" in Fig 1 provides an overview of our results, which we detail in this section.

The F_{ST} estimator for independent subpopulations and infinite subpopulation sample sizes. The WC, Weir-Hill, and Hudson method-of-moments estimators have small sample size corrections that remarkably make them consistent (as the number of independent loci m goes to infinity) for finite numbers of individuals. However, these small sample corrections also make the estimators unnecessarily cumbersome for our purposes (see Methods, section **Previous F_{ST} estimators for the independent subpopulations model** for complete formulas). In order to illustrate clearly how these estimators behave, both under the independent subpopulations model and for arbitrary structure, here we construct simplified versions that assume infinite sample sizes per subpopulation (Methods, section **Previous F_{ST} estimators for the independent subpopulations model**). This simplification corresponds to eliminating statistical sampling, leaving only genetic sampling to analyze [71]. Note that our simplified estimator nevertheless illustrates the general behavior of the WC, Weir-Hill, and Hudson estimators under arbitrary structure, and the results are equivalent to those we would obtain under finite sample sizes of individuals. While the Hudson F_{ST} estimator compares two subpopulations [23], based on that work we derive a generalized "HudsonK" estimator for more than two subpopulations in Methods, section **Generalized HudsonK F_{ST} estimator**. Note that HudsonK, first derived in [58], also equals the Weir-Goudet F_{ST} estimator for subpopulations [21] when loci are biallelic, which was derived independently using allele matching (S1 Text).

Under infinite subpopulation sample sizes, the allele frequencies at each locus and every subpopulation are known. Let $j \in \{1, \ldots, n\}$ index subpopulations rather than individuals and π_{ij} be the true allele frequency in subpopulation j at locus i. Note that π_{ij} are not estimated allele frequencies, but rather true subpopulation allele frequencies; this abstraction does not result in a practical estimation approach, but it greatly simplifies understanding of bias for subpopulations in a setting where there there is no statistical sampling. Although in this analysis of F_{ST} estimators the π_{ij} values are applied to subpopulations, for coherence with our previous work we shall call them "individual-specific allele frequencies" (IAF) [60, 61]. Whether for individuals or subpopulations, the key assumption is that IAFs satisfy the coancestry model of Eq (6). In this special case of infinite subpopulation sample sizes, all of WC, Weir-Hill, and HudsonK simplify to the following F_{ST} estimator for independent subpopulations ("indep"; derived in Methods, section **Previous F_{ST} estimators for the independent subpopulations model**):

$$\hat{p}_i^T = \frac{1}{n}\sum_{j=1}^{n}\pi_{ij}, \tag{11}$$

$$\hat{\sigma}_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}(\pi_{ij} - \hat{p}_i^T)^2, \tag{12}$$

$$\hat{F}_{\text{ST}}^{\text{indep}} = \frac{\sum_{i=1}^{m} \hat{\sigma}_i^2}{\sum_{i=1}^{m} \hat{p}_i^T\left(1-\hat{p}_i^T\right) + \frac{1}{n}\hat{\sigma}_i^2}. \qquad (13)$$

The goal is to estimate $F_{\text{ST}} = \frac{1}{n}\sum_{j=1}^{n} \theta_{jj}^T$, which is the special case of Eq (9) that weighs every subpopulation j equally ($w_j = \frac{1}{n}\ \forall j$).

F_{ST} estimation under the independent subpopulations model. Under the independent subpopulations model $\theta_{jk}^T = 0$ for $j \neq k$, where T is the most recent common ancestor (MRCA) population of the set of subpopulations. Note that the estimator in Eq (13) can be derived directly from Eq (6) and these assumptions using the method of moments (ignoring the existence of previous F_{ST} estimators; S1 Text). The expectations of the two recurrent terms in Eq (13) are

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

Eliminating $\overline{p(1-p)}^T$ and solving for F_{ST} in this system of equations recovers the estimator in Eq (13).

Before applying the convergence result in Eq (10), we test that the three conditions listed in section **Assessing the accuracy of genome-wide ratio estimators** are met. Condition (i): The locus 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 $A = F_{\text{ST}}$, $B = 1$, and $c_i = p_i^T(1-p_i^T)$. Condition (ii): $\bar{c}_m \to c = \mathrm{E}[p_i^T(1-p_i^T)|T] \neq 0$ over the p_i^T distribution across loci. Condition (iii): Since $0 \leq \pi_{ij}, \hat{p}_i^T \leq 1$, 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 loci,

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

F_{ST} estimation under arbitrary coancestry. Now we consider applying the independent subpopulations F_{ST} estimator to dependent subpopulations. The key difference is that now $\theta_{jk}^T \neq 0$ for every (j, k) will be assumed in our coancestry model in Eq (6), and now T may be either the MRCA population of all subpopulations or a more ancestral population. In this general setting, (j, k) may index either subpopulations or individuals. The two terms of $\hat{F}_{\text{ST}}^{\text{indep}}$ now satisfy

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

where $\overline{\theta}^T = \frac{1}{n^2}\sum_{j=1}^n \sum_{k=1}^n \theta_{jk}^T$ is the mean coancestry with uniform weights. There are two equations but three unknowns: F_{ST}, $\overline{\theta}^T$, and $\overline{p(1-p)}^T$. The independent subpopulations model satisfies $\overline{\theta}^T = \frac{1}{n}F_{\text{ST}}$, which allows for the consistent estimation of F_{ST}. Therefore, the new unknown $\overline{\theta}^T$ precludes consistent F_{ST} estimation without additional assumptions. As shown later, our additional assumption is that we can identify unrelated individuals in the data, which determines all unknowns. We defer our complete solution to this problem until kinship and its estimation challenges have been presented.

The F_{ST} estimator for independent subpopulations converges more generally to

$$\hat{F}_{\text{ST}}^{\text{indep}} \xrightarrow[m\to\infty]{\text{a.s.}} \frac{F_{\text{ST}} - \tilde{\theta}^T}{1 - \tilde{\theta}^T}, \tag{14}$$

(the conclusion of panel “Indep. Subpop. F_{ST} Estimator” in Fig 1), where

$$\tilde{\theta}^T = \frac{1}{n-1}\left(n\overline{\theta}^T - F_{\text{ST}}\right) = \frac{1}{n(n-1)}\sum_{j\neq k}\theta_{jk}^T$$

is the average of all between-subpopulation coancestry coefficients, in agreement with related calculations regarding the WC and Weir-Hill estimators [4, 21]. Therefore, under arbitrary structure the independent subpopulations estimator's bias is due to the coancestry between subpopulations. While the limit in Eq (14) appears to vary depending on the choice of T, it is in fact a constant with respect to T (proof in S1 Text).

Since $\frac{1}{n}F_{\text{ST}} \leq \overline{\theta}^T \leq F_{\text{ST}}$ (S1 Text), this estimator has a downward bias in the general setting: it is asymptotically unbiased ($\hat{F}_{\text{ST}}^{\text{indep}} \xrightarrow[m\to\infty]{\text{a.s.}} F_{\text{ST}}$) only when $\overline{\theta}^T = \frac{1}{n}F_{\text{ST}}$, while bias is maximal when $\overline{\theta}^T = F_{\text{ST}}$, where $\hat{F}_{\text{ST}}^{\text{indep}} \xrightarrow[m\to\infty]{\text{a.s.}} 0$. For example, if $\min \theta_{jk}^T = 0$ so the MRCA population T is fixed, but n is large and $\theta_{jk}^T \approx F_{\text{ST}}$ for most pairs of subpopulations, then $\overline{\theta}^T \approx F_{\text{ST}}$ as well, and $\hat{F}_{\text{ST}}^{\text{indep}} \approx 0$. Therefore, the magnitude of the bias of $\hat{F}_{\text{ST}}^{\text{indep}}$ is unknown if $\overline{\theta}^T$ is unknown, and small $\hat{F}_{\text{ST}}^{\text{indep}}$ estimates may arise even if F_{ST} is very large.

Coancestry estimation as a method of moments. Since the generalized F_{ST} is given by coancestry coefficients θ_{jj}^T in Eq (9), 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}. This is consistent with IBD parameters in general requiring a reference population to be determined [40], whereas in this subsection this reference population is unspecified.

Given IAFs and the coancestry model of Eq (6), the first and second moments that average across loci are

$$\mathrm{E}\left[\frac{1}{m}\sum_{i=1}^m \pi_{ij}\middle| T\right] = \overline{p}^T, \tag{15}$$

$$\mathrm{E}\left[\frac{1}{m}\sum_{i=1}^m \pi_{ij}\pi_{ik}\middle| T\right] = \overline{p^2}^T + \overline{p(1-p)}^T\theta_{jk}^T, \tag{16}$$

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

Suppose first that only θ_{jj}^T are of interest. There are n estimators given by Eq (16) with $j = k$, each corresponding to an unknown θ_{jj}^T. However, all these estimators share two nuisance parameters: $\overline{p}^T$ and $\overline{p^2}^T$. While $\overline{p}^T$ can be estimated from Eq (15), 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 in Eq (16) are considered rather than only the $j = k$ cases. Therefore, we cannot estimate coancestry coefficients consistently using only the first two moments without additional assumptions.

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

Given the biases we see for $\hat{F}_{\mathrm{ST}}^{\mathrm{indep}}$ under arbitrary structures in the previous section, we now turn to the generalized definition of F_{ST} and pursue an estimate of it. Recall that our generalized F_{ST} in Eq (3) is defined in terms of inbreeding coefficients, which are a special case of the kinship coefficient. Kinship coefficients also determine the bias of $\hat{F}_{\mathrm{ST}}^{\mathrm{indep}}$ in Eq (14) (since coancestry and kinship coefficients are closely related: see panel "Coancestry in Terms of Kinship" in Fig 1). Therefore, we will consider estimates of kinship and inbreeding in this section. Estimating kinship is also important for GWAS approaches that control for population structure [16, 24–35, 72, 73].

In this section, we focus on a standard kinship method-of-moments estimator and calculate its limit for the first time (panel "Existing Kinship Estimator" in 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 $\hat{F}_{\mathrm{ST}}^{\mathrm{indep}}$ in the previous section, and define unbiased F_{ST} estimators that require knowing the mean kinship or coancestry, or its proportion relative to F_{ST}. The results of this section directly motivate and help construct our new kinship and F_{ST} estimation approach in the following section.

Characterization of the standard kinship estimator. Here we analyze a standard kinship estimator that is frequently used [16, 27, 30–36]. 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 the earlier section **Assessing the accuracy of genome-wide ratio estimators**:

$$\hat{p}_i^T = \frac{1}{2}\sum_{j=1}^{n} w_j x_{ij}, \tag{17}$$

$$\hat{\varphi}_{jk}^{T,\mathrm{std}} = \frac{\sum_{i=1}^{m}(x_{ij} - 2\hat{p}_i^T)(x_{ik} - 2\hat{p}_i^T)}{4\sum_{i=1}^{m}\hat{p}_i^T(1-\hat{p}_i^T)}. \tag{18}$$

The estimator in Eq (18) resembles the sample covariance estimator applied to genotypes, but centers by locus i rather than by individuals j and k, and normalizes using estimates of $4p_i^T(1-p_i^T)$. We derive Eq (18) directly using the method of moments in S1 Text. The weights in Eq (17) must satisfy $w_j > 0$ and $\sum_{j=1}^n w_j = 1$, so that $0 \leq \hat{p}_i^T \leq 1$ and $\mathrm{E}[\hat{p}_i^T|T] = p_i^T$.

Utilizing the kinship model for genotypes from Eq (5), we find that Eq (18) converges to

$$\hat{\varphi}_{jk}^{T,\mathrm{std}} \xrightarrow[m\to\infty]{\text{a.s.}} \frac{\varphi_{jk}^T - \overline{\varphi}_j^T - \overline{\varphi}_k^T + \overline{\varphi}^T}{1 - \overline{\varphi}^T}, \tag{19}$$

where $\overline{\varphi}_j^T = \sum_{k'=1}^n w_{k'}\varphi_{jk'}^T$ and $\overline{\varphi}^T = \sum_{j'=1}^n \sum_{k'=1}^n w_{j'}w_{k'}\varphi_{j'k'}^T$, which agrees with related derivations [21, 62]. (This is the conclusion of panel "Existing Kinship Estimator" in Fig 1; see S1 Text for intermediate calculations that lead to Eq (19).) Therefore, the bias of $\hat{\varphi}_{jk}^{T,\mathrm{std}}$ varies per pair of individuals j and k. Analogous distortions have been observed for sample covariances of genotypes [74]. The limit of $\hat{\varphi}_{jk}^{T,\text{std}}$ in Eq (19) is constant with respect to T (proof in S1 Text). Similarly, inbreeding coefficient estimates derived from Eq (18) converge to

$$\hat{f}_j^{T,\text{std}} = 2\hat{\varphi}_{jj}^T - 1 \xrightarrow[m\to\infty]{\text{a.s.}} \frac{f_j^T - 4\overline{\varphi}_j^T + 3\overline{\varphi}^T}{1 - \overline{\varphi}^T}. \tag{20}$$

The difference between the bias of $\hat{\varphi}_{jk}^{T,\text{std}}$ for $j \neq k$ in Eq (19) and $\hat{f}_j^{T,\text{std}}$ in Eq (20) is visible in the kinship estimates shown toward the end of the results section. The limits of the ratio-of-means versions of two more f_j^T estimators [32] are, if $\hat{p}_i^T$ uses Eq (17),

$$\hat{f}_j^{T,\text{stdII}} = 1 - \frac{\sum_{i=1}^m x_{ij}(2 - x_{ij})}{2\sum_{i=1}^m \hat{p}_i^T(1 - \hat{p}_i^T)} \xrightarrow[m\to\infty]{\text{a.s.}} \frac{f_j^T - \overline{\varphi}^T}{1 - \overline{\varphi}^T},$$

$$\hat{f}_j^{T,\text{stdIII}} = \frac{\sum_{i=1}^m x_{ij}^2 - (1 + 2\hat{p}_i^T)x_{ij} + 2(\hat{p}_i^T)^2}{2\sum_{i=1}^m \hat{p}_i^T(1 - \hat{p}_i^T)} \xrightarrow[m\to\infty]{\text{a.s.}} \frac{f_j^T + \overline{\varphi}^T - 2\overline{\varphi}_j^T}{1 - \overline{\varphi}^T}. \tag{21}$$

The estimators in Eqs (18) and (21) are unbiased when $\hat{p}_i^T$ is replaced by p_i^T [16, 32, 36], and are consistent when $\hat{p}_i^T$ is consistent [60]. Surprisingly, $\hat{p}_i^T$ in Eq (17) is not consistent (it does not converge almost surely to p_i^T) for arbitrary population structures, which is at the root of the bias in Eqs (19) to (21). In particular, although $\hat{p}_i^T$ is unbiased, its variance (S1 Text, and some special cases shown elsewhere, *e.g.*, [19]),

$$\text{Var}(\hat{p}_i^T | T) = p_i^T(1 - p_i^T)\overline{\varphi}^T, \tag{22}$$

may be asymptotically non-zero as $n \to \infty$, since $p_i^T \in (0, 1)$ is fixed and $\lim_{n\to\infty} \overline{\varphi}^T$ may take on any value between zero and one for arbitrary population structures. Further, $\overline{\varphi}^T \to 0$ 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 $p_i^T(1 - p_i^T)$ is biased (S1 Text),

$$\text{E}[\hat{p}_i^T(1 - \hat{p}_i^T) | T] = p_i^T(1 - p_i^T)(1 - \overline{\varphi}^T),$$

which is present in all estimators we have studied.

Estimation of coancestry coefficients from IAFs. Here we form a coancestry coefficient estimator analogous to Eq (18) but using IAFs. Assuming the moments in Eq (6), this estimator and its limit are

$$\hat{p}_i^T = \sum_{j=1}^n w_j \pi_{ij}, \tag{23}$$

$$\hat{\theta}_{jk}^{T,\text{std}} = \frac{\sum_{i=1}^m (\pi_{ij} - \hat{p}_i^T)(\pi_{ik} - \hat{p}_i^T)}{\sum_{i=1}^m \hat{p}_i^T(1 - \hat{p}_i^T)} \xrightarrow[m\to\infty]{\text{a.s.}} \frac{\theta_{jk}^T - \overline{\theta}_j^T - \overline{\theta}_k^T + \overline{\theta}^T}{1 - \overline{\theta}^T}, \tag{24}$$

where $\overline{\theta}_j^T = \sum_{k=1}^n w_k \theta_{jk}^T$ and $\overline{\theta}^T = \sum_{j=1}^n \sum_{k=1}^n w_j w_k \theta_{jk}^T$ are analogous to $\overline{\varphi}_j^T$ and $\overline{\varphi}^T$. Eq (23) generalizes Eq (11) for arbitrary weights. Thus, use of IAFs does not ameliorate the estimation problems we have identified for genotypes. Like Eq (22), $\hat{p}_i^T$ in Eq (23) is not consistent because $\mathrm{Var}(\hat{p}_i^T|T) = p_i^T(1 - p_i^T)\bar{\theta}^T$ may not converge to zero for arbitrary population structures, which causes the bias observed in Eq (24).

F

F_{ST} estimator based on the standard kinship estimator. Since the generalized F_{ST} is defined as a mean inbreeding coefficient in Eq (3), here we study the F_{ST} estimator constructed as $\hat{F}_{\mathrm{ST}}^{\mathrm{std}} = \sum_{j=1}^n w_j \hat{f}_j^{T,\mathrm{std}}$ where $\hat{f}_j^{T,\mathrm{std}}$ is the inbreeding estimator derived from the standard kinship estimator. Although $\hat{f}_j^{T,\mathrm{std}}$ is biased, we nevertheless plug it into our definition of F_{ST} so that we may study how bias manifests. Note that we do not recommend utilizing this F_{ST} estimator in practice, but we find these results informative for identifying how to proceed in deriving new estimators in the following section.

Remarkably, the three f_j^T estimators in Eqs (20) and (21) give exactly the same plug-in $\hat{F}_{\mathrm{ST}}^{\mathrm{std}}$ if the weights in F_{ST} and $\hat{p}_i^T$ in Eq (17) match, namely

$$\hat{F}_{\mathrm{ST}}^{\mathrm{std}} = \sum_{j=1}^n w_j \hat{f}_j^{T,\mathrm{std}} = \frac{\sum_{i=1}^m \sum_{j=1}^n w_j (x_{ij} - 2\hat{p}_i^T)^2}{2\sum_{i=1}^m \hat{p}_i^T(1 - \hat{p}_i^T)} - 1 \xrightarrow[m\to\infty]{\text{a.s.}} \frac{F_{\mathrm{ST}} - \bar{\varphi}^T}{1 - \bar{\varphi}^T}, \tag{25}$$

where the limit assumes locally-outbred individuals so Eq (4) holds. The analogous F_{ST} estimator for IAFs and its limit are

$$\hat{F}_{\mathrm{ST}}^{\mathrm{std}} = \sum_{j=1}^n w_j \hat{\theta}_{jj}^{T,\mathrm{std}} = \frac{\sum_{i=1}^m \sum_{j=1}^n w_j (\pi_{ij} - \hat{p}_i^T)^2}{\sum_{i=1}^m \hat{p}_i^T(1 - \hat{p}_i^T)} \xrightarrow[m\to\infty]{\text{a.s.}} \frac{F_{\mathrm{ST}} - \bar{\theta}^T}{1 - \bar{\theta}^T}. \tag{26}$$

The estimators in Eqs (25) and (26) for individuals and their limits resemble those of classical F_{ST} estimators for populations of the form $\frac{\sigma_p^2}{\bar{p}(1-\bar{p})}$ [4, 5]. $\hat{F}_{\mathrm{ST}}^{\mathrm{std}}$ in Eq (26) for subpopulations j with uniform weight and one locus is also G_{ST} for two alleles [75]. Compared to $\hat{F}_{\mathrm{ST}}^{\mathrm{indep}}$ in Eq (13), $\hat{F}_{\mathrm{ST}}^{\mathrm{std}}$ in Eq (26) admits arbitrary weights and, by forgoing bias correction under the independent subpopulations model, is a simpler target of study.

Like $\hat{F}_{\mathrm{ST}}^{\mathrm{indep}}$ in Eq (13), $\hat{F}_{\mathrm{ST}}^{\mathrm{std}}$ in Eqs (25) and (26) are downwardly biased since $0 \leq \bar{\varphi}^T, \bar{\theta}^T$. $\hat{F}_{\mathrm{ST}}^{\mathrm{std}}$ in Eq (26) may converge arbitrarily close to zero since $\bar{\theta}^T$ can be arbitrarily close to F_{ST} (S1 Text). Moreover, although $\bar{\varphi}^T \approx \bar{\theta}^T$ for large n (see Eq (8) and panel "Coancestry in Terms of Kinship" in Fig 1), in extreme cases $\bar{\varphi}^T$ can exceed F_{ST} under the coancestry model (where $\bar{\theta}^T \leq \bar{\varphi}^T$) and also under extreme local kinship, where $\hat{F}_{\mathrm{ST}}^{\mathrm{std}}$ in Eq (25) converges to a negative value.

Adjusted consistent oracle F_{ST} estimators and the "bias coefficient". Here we explore two adjustments to $\hat{F}_{\mathrm{ST}}^{\mathrm{std}}$ from IAFs in Eq (26) that rely on having minimal additional information needed to correct its bias. These "oracle" approaches require information that is not known in practice, but this exercise helps us understand the problem more deeply and finds further connections between the various F_{ST} estimators.

If $\bar{\theta}^T$ is known, the bias in Eq (26) can be reversed, yielding the consistent estimator

$$\hat{F}'_{\mathrm{ST}} = \hat{F}_{\mathrm{ST}}^{\mathrm{std}}(1 - \bar{\theta}^T) + \bar{\theta}^T \xrightarrow[m\to\infty]{\text{a.s.}} F_{\mathrm{ST}}. \tag{27}$$

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

$$s^T = \frac{\bar{\theta}^T}{F_{\mathrm{ST}}} = \frac{\sum_{j=1}^{n}\sum_{k=1}^{n} w_j w_k \theta_{jk}^T}{\sum_{j=1}^{n} w_j \theta_{jj}^T}, \tag{28}$$

which we call the "bias coefficient" and which has interesting properties. The bias coefficient quantifies the departure from the independent subpopulations model by comparing the mean coancestry (θ_{jk}^T) to the mean inbreeding coefficient (θ_{jj}^T), and given $F_{\mathrm{ST}} > 0$ satisfies $0 < s^T \leq 1$ (S1 Text). The limit in Eq (26) in terms of s^T is

$$\hat{F}_{\mathrm{ST}}^{\mathrm{std}} \xrightarrow[m\to\infty]{\text{a.s.}} F_{\mathrm{ST}} \frac{1 - s^T}{1 - s^T F_{\mathrm{ST}}}. \tag{29}$$

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

$$\hat{\sigma}_i^2 = \frac{1}{1 - s^T} \sum_{j=1}^{n} w_j (\pi_{ij} - \hat{p}_i^T)^2, \tag{30}$$

$$\hat{F}_{\mathrm{ST}}'' = \frac{\hat{F}_{\mathrm{ST}}^{\mathrm{std}}}{1 - s^T(1 - \hat{F}_{\mathrm{ST}}^{\mathrm{std}})} = \frac{\sum_{i=1}^{m} \hat{\sigma}_i^2}{\sum_{i=1}^{m} \hat{p}_i^T(1 - \hat{p}_i^T) + s^T \hat{\sigma}_i^2} \xrightarrow[m\to\infty]{\text{a.s.}} F_{\mathrm{ST}}. \tag{31}$$

Note that $\hat{\sigma}_i^2$ and $\hat{F}_{\mathrm{ST}}^{\mathrm{indep}}$ from Eqs (12) and (13) are the special case of Eqs (30) and (31) for uniform weights and $s^T = \frac{1}{n}$; hence, $\hat{F}_{\mathrm{ST}}''$ generalizes $\hat{F}_{\mathrm{ST}}^{\mathrm{indep}}$.

Lastly, using either Eqs (26) or (29), the relative error of $\hat{F}_{\mathrm{ST}}^{\mathrm{std}}$ converges to

$$1 - \frac{\hat{F}_{\mathrm{ST}}^{\mathrm{std}}}{F_{\mathrm{ST}}} \xrightarrow[m\to\infty]{\text{a.s.}} \frac{\bar{\theta}^T(1 - F_{\mathrm{ST}})}{F_{\mathrm{ST}}(1 - \bar{\theta}^T)} = s^T \frac{1 - F_{\mathrm{ST}}}{1 - s^T F_{\mathrm{ST}}}, \tag{32}$$

which is approximated by s^T if $F_{\mathrm{ST}} \ll 1$, hence the name "bias coefficient". Note s^T varies depending on the choice of T, which is necessary since F_{ST} (and hence the relative bias of $\hat{F}_{\mathrm{ST}}^{\mathrm{std}}$ from F_{ST}) depends on the choice of T.

A new approach for kinship and F_{ST} estimation

Here, we propose a new estimation approach for kinship coefficients that has properties favorable for obtaining nearly unbiased estimates (panel "New Kinship Estimator" in Fig 1). These new kinship estimates yield an improved F_{ST} estimator. We present the general approach and implement a simple version of one key estimator that results in the complete proof-of-principle estimator that is evaluated in the next section and applied to human data in [59]. We also compare our approach to a related estimator of non-IBD linearly-transformed kinship values [20–22] that was proposed concurrently to ours [58].

General approach. In this subsection we develop our new estimator in two steps. First, we compute a new statistic A_{jk} that is proportional in the limit of infinite loci to $\varphi_{jk}^T - 1$ times a nuisance factor v^T. Second, we estimate and remove v^T to yield the proposed estimator $\hat{\varphi}_{jk}^{T,\mathrm{new}}$. $\hat{A}_{\mathrm{min}}$—an estimator of the limit of the minimum A_{jk}—yields v^T if the least related pair of individuals in the data has $\varphi^T_{jk} = 0$, which sets T to the MRCA population of all the individuals in the data. The new kinship estimator immediately results in new inbreeding ($\hat{f}^{T,\text{new}}_j$) and F_{ST} ($\hat{F}^{\text{new}}_{\text{ST}}$) estimators. This general approach leaves the implementation of $\hat{A}_{\text{min}}$ open; the simple implementation applied in this work is described in subsection **Proof-of-principle kinship estimator using subpopulation labels**, but our method can be readily improved by substituting in a better $\hat{A}_{\text{min}}$ in the future.

Applying the method of moments to Eq (5), we derive the following statistic (S1 Text), whose expectation is proportional to $\varphi^T_{jk} - 1$:

$$\begin{aligned} A_{jk} &= \frac{1}{m}\sum_{i=1}^{m}(x_{ij}-1)(x_{ik}-1) - 1, \\ \mathrm{E}[A_{jk}|T] &= (\varphi^T_{jk}-1)v^T_m, \quad \text{where} \\ v^T_m &= \frac{4}{m}\sum_{i=1}^{m} p^T_i\left(1-p^T_i\right). \end{aligned} \tag{33}$$

Compared to the standard kinship estimator in Eq (19), which has a complex asymptotic bias determined by n parameters ($\bar{\varphi}^T_j$ for each $j \in \{1, \ldots, n\}$), the A_{jk} statistics estimate kinship with a bias controlled by the sole unknown parameter v^T_m shared by all pairs of individuals. The key to estimating v^T_m is to notice that if $\varphi^T_{jk} = 0$ then $\mathrm{E}[A_{jk}|T] = -v^T_m$. Thus, assuming $\min_{j,k}\varphi^T_{jk} = 0$, which sets T to the MRCA population, then the minimum A_{jk} yields the nuisance parameter. However, we recommend using a more stable estimate than the minimum A_{jk} to unbias all A_{jk}, such as the estimator presented in the next subsection.

In general, suppose $\hat{A}_{\text{min}}$ is a consistent estimator of the limit of the minimum $E[A_{jk}|T]$, or equivalently,

$$\hat{A}_{\text{min}} \xrightarrow[m\to\infty]{\text{a.s.}} -v^T,$$

along with the assumption that $v^T_m \xrightarrow[m\to\infty]{} v^T$ for some $v^T \neq 0$. Our new kinship estimator follows directly from replacing v^T_m with $-\hat{A}_{\text{min}}$ and solving for φ^T_{jk} in Eq (33), which results in a consistent kinship estimator (given the convergence proof of section **Assessing the accuracy of genome-wide ratio estimators**):

$$\hat{\varphi}^{T,\text{new}}_{jk} = 1 - \frac{A_{jk}}{\hat{A}_{\text{min}}} \xrightarrow[m\to\infty]{\text{a.s.}} \varphi^T_{jk}. \tag{34}$$

The resulting new inbreeding coefficient estimator is

$$\hat{f}^{T,\text{new}}_j = 2\hat{\varphi}^{T,\text{new}}_{jj} - 1 \xrightarrow[m\to\infty]{\text{a.s.}} f^T_j, \tag{35}$$

and the new F_{ST} estimator is consistent for locally-outbred individuals (estimates Eq (4)):

$$\hat{F}^{\text{new}}_{\text{ST}} = \sum_{j=1}^{n} w_j \hat{f}^{T,\text{new}}_j \xrightarrow[m\to\infty]{\text{a.s.}} F_{\text{ST}}. \tag{36}$$

Thus, only the implementation of $\hat{A}_{\text{min}}$ is left unspecified from this general estimation approach of kinship and F_{ST}. The implementation of $\hat{A}_{\text{min}}$ used in the analyses in this work is given in the next subsection.

Proof-of-principle kinship estimator using subpopulation labels. To showcase the potential of the new estimators, we implement a simple proof-of-principle version of $\hat{A}_{\min}$ needed for our new kinship estimator ($\hat{\varphi}_{jk}^{T,\mathrm{new}}$ in Eq (34)). This $\hat{A}_{\min}$ relies on an appropriate partition of the n individuals into K subpopulations (denoted S_u for $u \in \{1, \ldots, K\}$), where the only requirement is that the kinship coefficients between pairs of individuals across the two most unrelated subpopulations is zero, as detailed below. Note that, unlike the the independent subpopulations model of section **F_{ST} estimation based on the independent subpopulations model**, these K subpopulations need not be independent nor unstructured. The desired estimator $\hat{A}_{\min}$ is the minimum average A_{jk} over all subpopulation pairs:

$$\hat{A}_{\min} = \min_{u \neq v} \frac{1}{|S_u||S_v|} \sum_{j \in S_u} \sum_{k \in S_v} A_{jk}. \tag{37}$$

This $\hat{A}_{\min}$ consistently estimates the limit of the minimum A_{jk} if $\varphi_{jk}^T = 0 \ \forall j \in S_u, \forall k \in S_v$ for the least related pair of subpopulations S_u, S_v.

This estimator should work well for individuals truly divided into subpopulations, but may be biased for a poor choice of subpopulations, in particular if the minimum mean φ_{jk}^T between subpopulations is far greater than zero. For this reason, inspection of the kinship estimates is required and careful construction of appropriate subpopulations may be needed. See our analysis of human data for detailed examples [59]. Future work could focus on a more general $\hat{A}_{\min}$ that circumvents the need for subpopulations of our proof-of-principle estimator.

Comparison to the Weir-Goudet kinship estimator for individuals. Here we analyze the Weir-Goudet (WG) kinship estimator for individuals [20–22]. This has connections to our new estimator but differs in having the goal of estimating linearly-transformed kinship values. In our framework, the WG estimator is given by

$$\hat{\varphi}_{jk}^{T,\mathrm{WG}} = 1 - \frac{A_{jk}}{\hat{A}_{\mathrm{avg}}}, \quad \text{where} \quad \hat{A}_{\mathrm{avg}} = \frac{2}{n(n-1)} \sum_{j=2}^{n} \sum_{k=1}^{j-1} A_{jk}.$$

Therefore, this estimator differs from our proposal [58] by replacing our $\hat{A}_{\min}$ with $\hat{A}_{\mathrm{avg}}$. Under the kinship model, the expectation of $\hat{A}_{\mathrm{avg}}$ is

$$\mathrm{E}\left[\hat{A}_{\mathrm{avg}} | T\right] = (\tilde{\varphi}^T - 1) v_m^T, \quad \text{where} \quad \tilde{\varphi}^T = \frac{2}{n(n-1)} \sum_{j=2}^{n} \sum_{k=1}^{j-1} \varphi_{jk}^T.$$

Therefore, the limit of this estimator is

$$\hat{\varphi}_{jk}^{T,\mathrm{WG}} \xrightarrow[m \to \infty]{\text{a.s.}} \frac{\varphi_{jk}^T - \tilde{\varphi}^T}{1 - \tilde{\varphi}^T}, \tag{38}$$

which agrees with calculations in the original WG work [20–22]. Note that, assuming that kinship coefficients must be non-negative, the above estimator recovers the kinship IBD probabilities if and only if $\tilde{\varphi}^T = 0$ which holds if and only if $\varphi_{jk}^T = 0$ for every pair of individuals $j \neq k$. The resulting WG inbreeding coefficient estimator is

$$\hat{f}_{jk}^{T,\mathrm{WG}} = 2\hat{\varphi}_{jk}^{T,\mathrm{WG}} - 1 \xrightarrow[m \to \infty]{\text{a.s.}} \frac{f_j^T - \tilde{\varphi}^T}{1 - \tilde{\varphi}^T},$$

which estimates linearly-transformed inbreeding values [21]. Therefore, the resulting WG F_{ST} estimator (for individuals) also targets a linearly-transformed F_{ST} value (under locally-outbred individuals, where F_{ST} is given by Eq (4)), namely

$$\hat{F}_{\mathrm{ST}}^{\mathrm{WG}} = \frac{1}{n}\sum_{j=1}^{n} \hat{f}_j^{T,\mathrm{WG}} \xrightarrow[m\to\infty]{\text{a.s.}} \frac{F_{\mathrm{ST}} - \tilde{\varphi}^T}{1 - \tilde{\varphi}^T}.$$

The WG authors also briefly consider a variant of their kinship estimator that is normalized using the minimum kinship value as we did, developed concurrently with our approach [58], but was largely dismissed as an unnecessary correction [21, 76]. See S1 Text for a detailed proof that the general estimator framework we propose here (Eqs (33) and (34)) is algebraically equivalent to our original formulation in [58].

Note that the original WG does not estimate F_{ST} from individuals as considered above; instead, F_{ST} is estimated from coancestry estimates for subpopulations (which equals our HudsonK for biallelic loci, S1 Text) [20–22]. For completeness, we consider both kinds of F_{ST} estimates in the evaluations that follow.

Simulations evaluating F_{ST} and kinship estimators

Overview of simulations. We simulate genotypes from two models to illustrate our results when the true population structure parameters are known. Both simulations have clearly-defined IBD probability parameters in terms of the MRCA population. The first simulation satisfies the independent subpopulations model that existing F_{ST} estimators assume. The second simulation is from an admixture model with no independent subpopulations and pervasive kinship designed to induce large downward biases in existing kinship and F_{ST} estimators (Fig 2). This admixture scenario resembles the population structure we estimated for Hispanics in the 1000 Genomes Project [59]: compare the simulated kinship matrix (Fig 2B) and admixture proportions (Fig 3C) to our estimates on the real data [59]. Both simulations have n = 1000 individuals, m = 300, 000 loci, and K = 10 subpopulations or intermediate subpopulations. These simulations have F_{ST} = 0.1, comparable to previous estimates between human populations (in 1000 Genomes, the estimated F_{ST} between CEU (European-Americans) and CHB (Chinese) is 0.106, between CEU and YRI (Yoruba from Nigeria) it is 0.139, and between CHB and YRI it is 0.161 [23]).

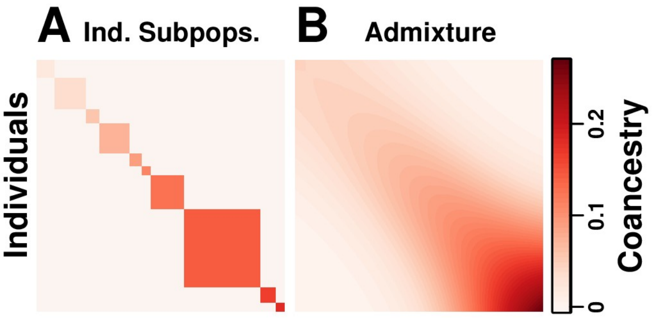

Fig 2. Coancestry matrices of simulations. Both panels have n = 1000 individuals along both axes, K = 10 subpopulations (final or intermediate), and F_{ST} = 0.1. Color corresponds to θ_{jk}^T between individuals j and k (equal to φ_{jk}^T off-diagonal, f_j^T along the diagonal). (A) The independent subpopulations model has $\theta_{jk}^T = 0$ between subpopulations, and varying θ_{jj}^T per subpopulation, resulting in a block-diagonal coancestry matrix. (B) Our admixture scenario models a 1D geography with extensive admixture and intermediate subpopulation differentiation that increases with distance, resulting in a smooth coancestry matrix with no independent subpopulations (no $\theta_{jk}^T = 0$ between blocks). Individuals are ordered along each axis by geographical position.

https://doi.org/10.1371/journal.pgen.1009241.g002

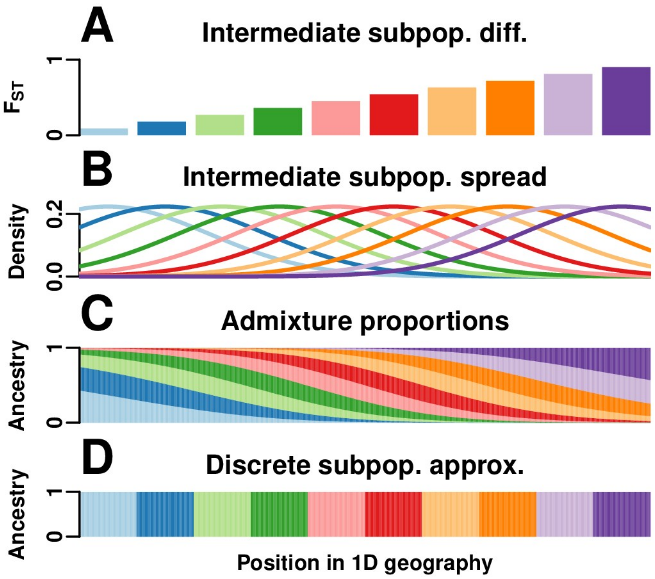

Fig 3. 1D admixture scenario. We model a 1D geography population that departs strongly from the independent subpopulations model. (A) $K = 10$ intermediate subpopulations, evenly spaced on a line, evolved independently in the past with F_{ST} increasing with distance, which models a sequence of increasing founder effects (from left to right) to mimic the global human population. (B) Once differentiated, individuals in these intermediate subpopulations spread by random walk modeled by Normal densities. (C) $n = 1000$ individuals, sampled evenly in the same geographical range, are admixed proportionally to the previous Normal densities. Thus, each individual draws most of its alleles from the closest intermediate subpopulation, and draws the fewest alleles from the most distant populations. Long-distance random walks of intermediate subpopulation individuals results in kinship for admixed individuals that decays smoothly with distance in Fig 2B. (D) For F_{ST} estimators that require a partition of individuals into subpopulations, individuals are clustered by geographical position ($K = 10$).

https://doi.org/10.1371/journal.pgen.1009241.g003

The independent subpopulations simulation satisfies the HudsonK and BayeScan estimator assumptions: each independent subpopulation S_u has a different F_{ST} value of $f^T_{S_u}$ relative to the MRCA population T (Fig 2A). Ancestral allele frequencies p^T_i are drawn uniformly between 0.01 and 0.5. Allele frequencies $p^{S_u}_i$ for S_u and locus i are drawn independently from the Balding-Nichols (BN) distribution [3] with parameters p^T_i and $f^T_{S_u}$. Every individual j in subpopulation S_u draws alleles randomly with probability $p^{S_u}_i$. Subpopulation sample sizes are drawn randomly (Methods, section **Simulations**).

The admixture simulation corresponds to a "BN-PSD" model [6, 27, 34, 60, 77]: the intermediate subpopulations are independent subpopulations that draw $p^{S_u}_i$ from the BN model, then each individual j constructs its allele frequencies as $\pi_{ij} = \sum_{u=1}^{K} p^{S_u}_i q_{ju}$, which is a weighted average of the subpopulation allele frequencies $p^{S_u}_i$ with the admixture proportions q_{ju} of individual j and subpopulation u as weights (which satisfy $\sum_{u=1}^{K} q_{ju} = 1$), as in the Pritchard-Stephens-Donnelly (PSD) admixture model [63–65]. We constructed q_{ju} that model admixture resulting from spread by random walk of the intermediate subpopulations along a one-dimensional geography, as follows. Intermediate subpopulations S_u are placed on a line with differentiation $f^T_{S_u}$ that grows with distance, which corresponds to a serial founder effect (Fig 3A). Upon differentiation, individuals in each S_u spread by random walk, a process modeled by Normal densities (Fig 3B). Admixed individuals derive their ancestry proportional from these Normal densities, resulting in a genetic structure governed by geography (Figs 3C and 2B) and departing strongly from the independent subpopulations model (Fig 3D). The amount of spread—which sets the mean kinship across all individuals—was chosen to give a bias coefficient of $s^T = \frac{\bar{\theta}^T}{F_{ST}} = 0.5$, which by Eq (32) results in a large downward bias for $\hat{F}_{ST}^{std}$ (in contrast, the independent subpopulations simulation has $s^T = 0.1$). The true coancestry and F_{ST} parameters of this simulation are given by the $f_{S_u}^T$ values of the intermediate subpopulations and the admixture coefficients q_{ju} of the individuals via the following equations [57]:

$$\theta_{jk}^T = \sum_{u=1}^{K} q_{ju} q_{ku} f_{S_u}^T,$$
$$F_{ST} = \sum_{j=1}^{n} \sum_{u=1}^{K} w_j q_{ju}^2 f_{S_u}^T. \quad (39)$$

The first equation above connecting coancestry to admixture proportions was derived independently in other work [62], but the F_{ST} for the admixed individuals was absent and instead follows from our generalized F_{ST} definition given in Eq (9). See Methods, section **Simulations** for additional details regarding these simulations.

Evaluation of F_{ST} estimators. Our admixture simulation illustrates the large biases that can arise if existing F_{ST} estimators that require independent subpopulations or F_{ST} estimates derived from existing kinship estimators are misapplied to arbitrary population structures to estimate the generalized F_{ST}, and demonstrate the higher accuracy of our new F_{ST} estimator ($\hat{F}_{ST}^{new}$ given by the combination of Eqs (36) and (37)). The WC F_{IT} (total inbreeding) estimator was also evaluated.

First, we test these estimators in our independent subpopulations simulation. The HudsonK (Methods, section **Generalized HudsonK F_{ST} estimator**) and BayeScan F_{ST} estimators are consistent in this simulation, since their assumptions are satisfied (Fig 4A). The WC F_{ST}

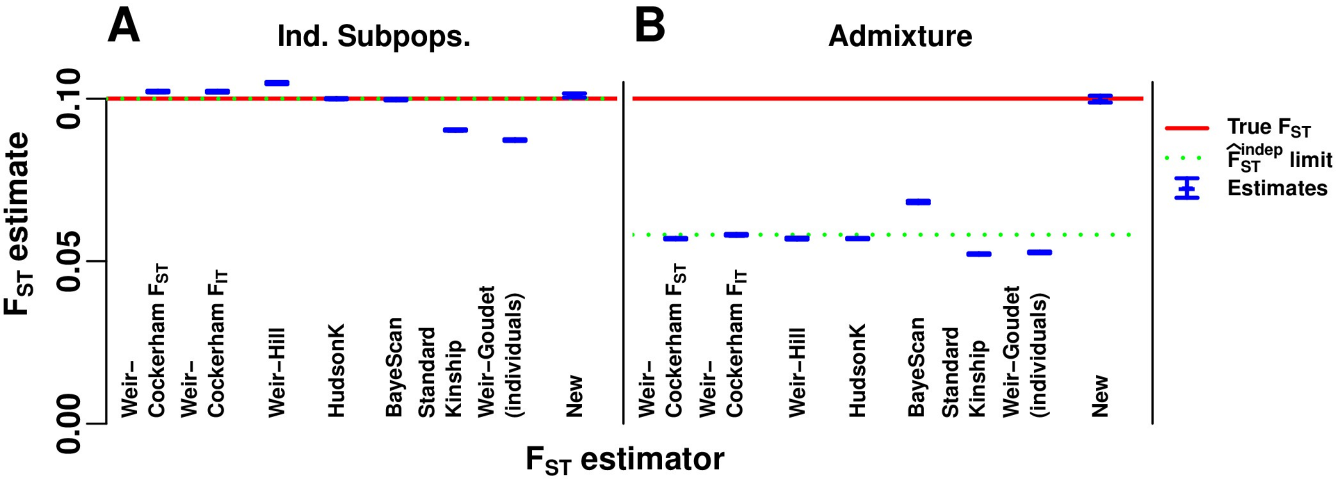

Fig 4. Evaluation of F_{ST} estimators. The Weir-Cockerham, Weir-Hill, Weir-Goudet (for individuals), HudsonK (equal to Weir-Goudet for subpopulations, S1 Text), BayeScan, $\hat{F}_{ST}^{std}$ in Eq (25) derived from the standard kinship estimator, and our new F_{ST} estimator in Eqs (34) and (37), are evaluated on simulated genotypes from our two models (Fig 2). The Weir-Cockerham F_{IT} estimator was also included to show that estimation of total inbreeding behaves similarly to F_{ST} estimators. (A) The independent subpopulations model required by the Weir-Hill, HudsonK, and BayeScan F_{ST} estimators. All but standard kinship ($\hat{F}_{ST}^{std}$) and Weir-Goudet (for individuals) recover the target F_{ST} IBD probability in Eq (9) (red line) with small errors. (B) Our admixture scenario, which has no independent subpopulations, was constructed so $\hat{F}_{ST}^{std} \approx \frac{1}{2} F_{ST}$. Only our new estimates are accurate. The rest of these estimators give values smaller than the target F_{ST} IBD probability, which result from treating kinship as zero between every subpopulations imposed by geographic clustering (or between individuals for Standard Kinship and Weir-Goudet). The $\hat{F}_{ST}^{indep}$ estimator limit in Eq (14) (green dotted line) overlaps the true F_{ST} (red line) in (A) but not (B). Estimates (blue) include 95% prediction intervals (often too narrow to see) from 39 independently-simulated genotype matrices for each model (Methods, section **Prediction intervals**).

https://doi.org/10.1371/journal.pgen.1009241.g004

estimator assumes that $f_{S_u}^T = F_{\mathrm{ST}}$ for all subpopulations S_u, which does not hold; nevertheless, WC has only a small bias (Fig 4A). The WC F_{IT} estimator arrives at similar estimates, as it should since there is no local inbreeding, so the true F_{IT} also equals F_{ST}. The Weir-Hill estimator permits different $f_{S_u}^T$ values per subpopulation, but assigns equal weight to individuals rather than subpopulations (Methods, section **The Weir-Hill F_{ST} estimator**), resulting in a slightly different target F_{ST} (we verified that these estimates are unbiased for this F_{ST}). For comparison, we show the standard kinship-based $\hat{F}_{\mathrm{ST}}^{\mathrm{std}}$ in Eq (25) (weights from Methods, section **Simulations**) and $\hat{F}_{\mathrm{ST}}^{\mathrm{WG}}$ based on the Weir-Goudet kinship estimates for individuals, both of which do not have corrections that would make them consistent under the independent subpopulations model. Since the number of subpopulations K is large, $\hat{F}_{\mathrm{ST}}^{\mathrm{std}}$ has a small relative bias of about $s^T = \frac{1}{K} = 10\%$ (Fig 4A); greater bias is expected for smaller K. Our new F_{ST} estimator has a very small bias in this simulation resulting from estimating the minimum kinship from the smallest kinship between subpopulations (see Eq (37)) rather than their average as HudsonK does implicitly (Fig 4A).

Next we test these estimators in our admixture simulation. To apply the F_{ST} estimators that require subpopulations to the admixture model, individuals are clustered into subpopulations by their geographical position (Fig 3D). We find that estimates of all existing methods are smaller than the true F_{ST} by nearly half, as predicted by the limit of $\hat{F}_{\mathrm{ST}}^{\mathrm{indep}}$ in Eq (14) (Fig 4B). The WC F_{IT} estimator obtains slightly larger estimates than the WC F_{ST} estimator, but overall remains as biased as the other F_{ST} estimators, showing that the use of a total inbreeding estimator for independent subpopulations displays the same bias as the corresponding F_{ST} estimator. By construction, the kinship-based $\hat{F}_{\mathrm{ST}}^{\mathrm{std}}$ also has a large relative bias of about s^T = 50%; remarkably, all existing F_{ST} estimators for subpopulations suffer from comparable biases. Thus, the corrections for independent subpopulations present in the WC, Weir-Hill and HudsonK estimators, or the Bayesian likelihood modeling of BayeScan, are insufficient for accurate estimation of the target generalized F_{ST} (Eq (9)) in this admixture scenario. Only our new F_{ST} estimator achieves accurate estimates of the generalized F_{ST} in the admixture simulation (Fig 4B).

Evaluation of kinship estimators. Our admixture simulation illustrates the distortions of the standard kinship estimator $\hat{\varphi}_{jk}^{T,\mathrm{std}}$ in Eq (18), the linearly-transformed kinship values given by the Weir-Goudet estimator, and demonstrates the improved accuracy of our new kinship estimator $\hat{\varphi}_{jk}^{T,\mathrm{new}}$ given by the combination of Eqs (34) and (37). Kinship matrix estimates and their limits are visualized as heatmaps in Fig 5, whereas estimator accuracy is shown directly in Fig 6. The limit of the standard estimator $\hat{\varphi}_{jk}^{T,\mathrm{std}}$ in Eq (18) would have had a uniform bias if $\overline{\varphi}_j^T = \overline{\varphi}^T$ held for all individuals j. For that reason, our admixture simulation has varying differentiation $f_{S_u}^T$ per intermediate subpopulation S_u (Fig 3A), which causes large differences in $\overline{\varphi}_j^T$ per individual j and therefore large distortions in $\hat{\varphi}_{jk}^{T,\mathrm{std}}$. The Weir-Goudet approach estimates the linearly-transformed kinship values calculated in Eq (38).

Our new kinship estimator (Fig 5B) recovers the true kinship matrix of this complex population structure (Fig 5A), with an RMSE of 2.83% relative to the mean φ_{jk}^T (Fig 6). In contrast, estimates using the standard estimator have a large overall downward bias (Fig 5C), resulting in an RMSE of 115.72% from the true φ_{jk}^T relative to the mean φ_{jk}^T (Fig 6). Additionally, estimates from $\hat{\varphi}_{jk}^{T,\mathrm{std}}$ are very distorted, with an abundance of $\hat{\varphi}_{jk}^{T,\mathrm{std}} < \varphi_{jk}^T$ cases—some of which are negative estimates (blue in Fig 5C)—but remarkably also cases with $\hat{\varphi}_{jk}^{T,\mathrm{std}} > \varphi_{jk}^T$ (top left corner of Figs 5C and 6).

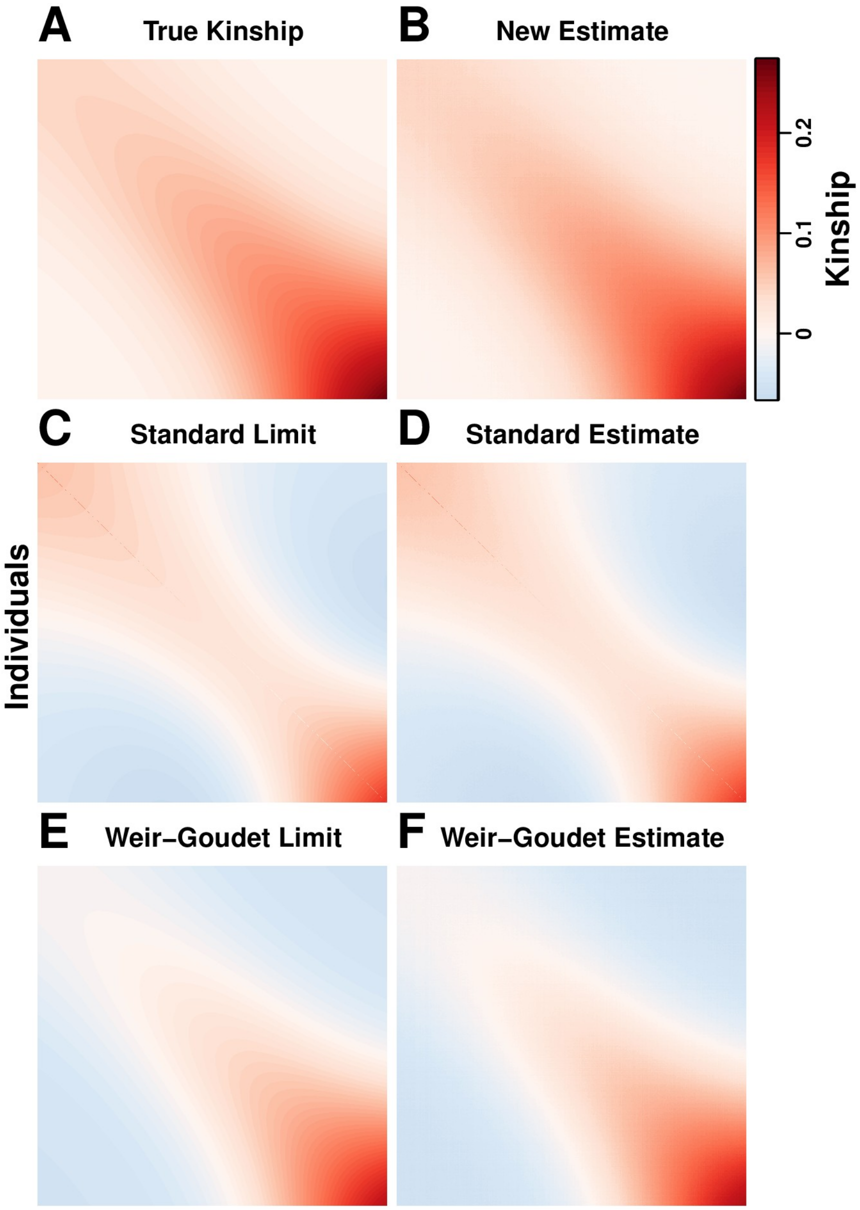

Fig 5. Evaluation of kinship estimators. Observed accuracy for two existing kinship coefficient estimators is illustrated in our admixture simulation and contrasted to the nearly unbiased estimates of our new estimator. Plots show $n = 1000$ individuals along both axes, and color corresponds to φ_{jk}^{T} between individuals $j \neq k$ and to f_j^T along the diagonal (f_j^T is in the same scale as φ_{jk}^{T} for $j \neq k$; plotting φ_{jj}^{T}, which have a minimum value of $\frac{1}{2}$, would result in a discontinuity in this figure). (A) True kinship matrix. (B) Estimated kinship using our new estimator in Eqs (34) and (37) from simulated genotypes recovers the true kinship matrix with high accuracy. (C) Theoretical limit of $\hat{\varphi}_{jk}^{T,\text{std}}$ in Eq (19) as the number of independent loci goes to infinity demonstrates the accuracy of our bias predictions under the kinship model. (D) Standard kinship estimates $\hat{\varphi}_{jk}^{T,\text{std}}$ given by Eq (18) from simulated genotypes are downwardly biased on average and distorted by pair-specific amounts. (E) Theoretical limit of the Weir-Goudet kinship estimator given by Eq (38). (F) Weir-Goudet kinship estimates from the same simulated genotypes agree with our calculated limit.

https://doi.org/10.1371/journal.pgen.1009241.g005

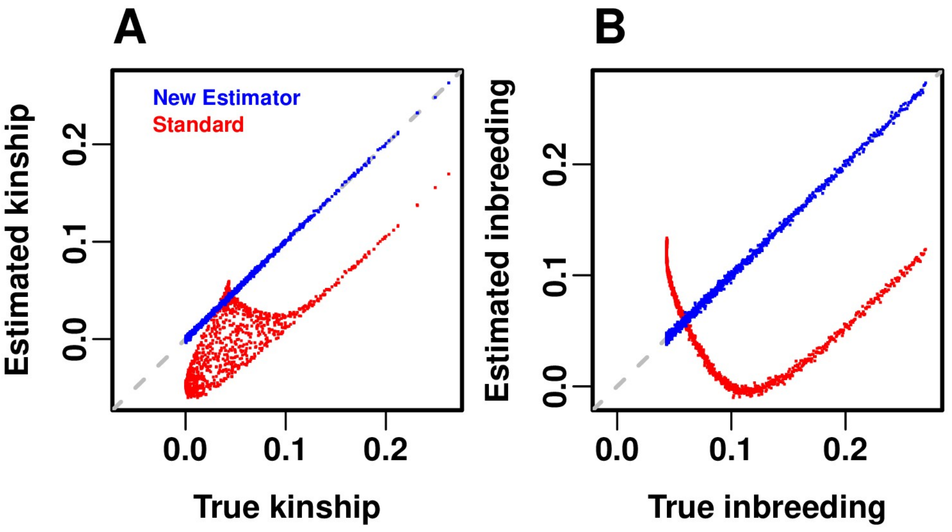

Fig 6. Accuracy of kinship estimators. Here the estimated kinship values are directly compared to their true values, in the same admixture simulation data ($n = 1000$ individuals) shown in the previous figure. (A) Kinship between different individuals (excluding inbreeding). The new estimator has practically no bias in this evaluation (falls on the 1-1 dashed gray line). The standard estimator has a complex, non-linear bias that covers a large area of errors. (B) Inbreeding comparison, shows the bias of the standard estimate follows a different pattern for inbreeding compared to kinship between individuals. To better visualize and compare data across panels, a random subset of n points (out of the original $n(n-1)/2$ unique individual pairs) were plotted in (A), matching the number of individuals (number of points in (B)).

https://doi.org/10.1371/journal.pgen.1009241.g006

Now we compare the convergence of the ratio-of-means and mean-of-ratios versions of the standard kinship estimator to their biased limit we calculated in Eq (19) (Fig 5D). The ratio-of-means estimate $\hat{\varphi}_{jk}^{T,\text{std}}$ (Fig 5C) has an RMSE of 2.14% from its limit relative to the mean φ_{jk}^{T}. In contrast, the mean-of-ratios estimates that are prevalent in the literature have a greater RMSE of 10.77% from the same limit in Eq (19). Thus, as expected from our theoretical results in section **Assessing the accuracy of genome-wide ratio estimators**, the ratio-of-means estimate is much closer to the desired limit than the mean-of-ratio estimate. The distortions are similar for the estimator that uses IAFs in Eq (24), with reduced RMSEs from its limit of 0.32% and 8.82% for the ratio-of-means and mean-of-ratios estimates, respectively.

Evaluation of oracle-adjusted F_{ST} estimators. Here we verify additional calculations for the bias of the standard kinship-based estimator $\hat{F}_{ST}^{\text{std}}$ and the unbiased adjusted "oracle" F_{ST} estimators that require the true mean kinship $\overline{\varphi}^{T}$ or the bias coefficient s^{T} to be known. Note that $\hat{F}_{ST}^{\text{new}}$ in Eq (36) is related but not identical to these oracle estimators. We tested both IAF (Fig 7A) and genotype (Fig 7B) versions of these estimators. The unadjusted $\hat{F}_{ST}^{\text{std}}$ in Eq (26) is severely biased (blue in Fig 7) by construction, and matches the calculated limit for IAFs and genotypes (green lines in Fig 7, which are close because $\overline{\varphi}^{T} \approx \overline{\theta}^{T}$). In contrast, the two consistent adjusted estimators $\hat{F}'_{ST}$ and $\hat{F}''_{ST}$ in Eqs (27) and (31) estimate F_{ST} quite well (blue predictions overlap the true F_{ST} red line in Fig 7). However, $\hat{F}'_{ST}$ and $\hat{F}''_{ST}$ are oracle methods, since they require parameters ($\overline{\varphi}^{T}$, $\overline{\theta}^{T}$, s^{T}) that are not known in practice.

Prediction intervals were computed from estimates over 39 independently-simulated IAF and genotype matrices (Methods, section **Prediction intervals**). Estimator limits are always contained in these intervals because the number of independent loci ($m = 300, 000$) is sufficiently large. Estimates that use genotypes have wider intervals than estimates from IAFs;

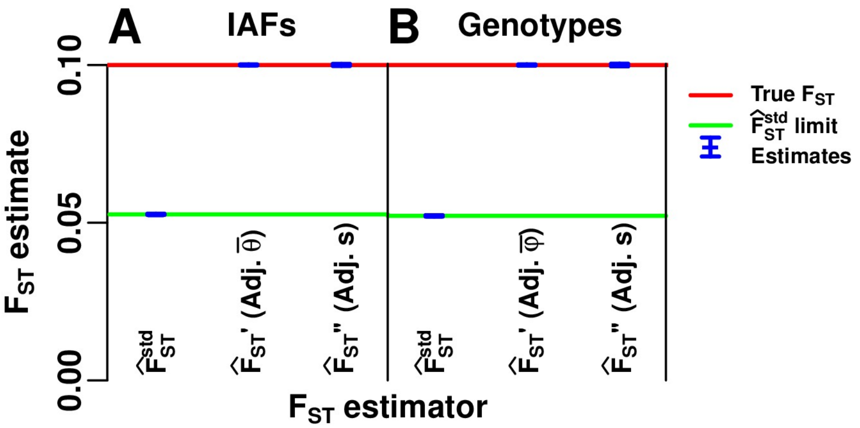

Fig 7. Evaluation of standard and adjusted F_{ST} estimators. The convergence values we calculated for the standard kinship plug-in and adjusted F_{ST} estimators are validated using our admixture simulation. All adjusted estimators are unbiased but are "oracle" methods, since the mean kinship ($\overline{\varphi}^T$), mean coancestry ($\overline{\theta}^T$), or bias coefficient ($s^T = \frac{\overline{\theta}^T}{F_{\mathrm{ST}}}$ for IAFs, replaced by $\frac{\overline{\varphi}^T}{F_{\mathrm{ST}}}$ for genotypes) are usually unknown. (A) Estimation from individual-specific allele frequencies (IAFs): $\hat{F}_{\mathrm{ST}}^{\mathrm{std}}$ is the standard coancestry plug-in estimator in Eq (26); $\hat{F}'_{\mathrm{ST}}$ "Adj. $\overline{\theta}^T$" is in Eq (27); $\hat{F}''_{\mathrm{ST}}$ "Adj. s" is in Eq (31). (B) For genotypes, $\hat{F}_{\mathrm{ST}}^{\mathrm{std}}$ is given in Eq (25), and the adjusted estimators use $\overline{\varphi}^T$ rather than $\overline{\theta}^T$. Lines: true F_{ST} (red line), limits of biased estimators $\hat{F}_{\mathrm{ST}}^{\mathrm{std}}$ (green 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 (Methods, section **Prediction intervals**).

https://doi.org/10.1371/journal.pgen.1009241.g007

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.

Discussion

We studied analytically the most commonly-used estimators of F_{ST} and kinship, which can be derived using the method of moments. We determined the estimation limits of convergence of these approaches under two models of arbitrary population structure (Fig 1). We found that no existing approaches estimate the generalized F_{ST} (an IBD probability) accurately (but note that some of these approaches intended to estimate a linearly-transformed F_{ST} quantity and not the IBD probability). We also showed that the standard kinship estimator is biased on structured populations (particularly when the average kinship is comparable to the kinship coefficients of interest), and this bias varies for each pair of individuals. These results led us to a new kinship estimator, which is consistent if the minimum kinship is estimated consistently (Fig 1). We presented an implementation of this approach, which is practically unbiased in our simulations. Our kinship and F_{ST} estimates in human data are consistent with the African Origins model while suggesting that human differentiation is considerably greater than previously estimated [59].

Estimation of F_{ST} in the correct scale is crucial for its interpretation as an IBD probability, for obtaining comparable estimates in different datasets and across species, as well as for DNA forensics [3, 7, 19, 20, 78–80]. Our framework results in a new unbiased genome-wide F_{ST} estimator. However, our findings may not have direct implications for single-locus F_{ST} estimate approaches where only the relative ranking matters, such as for the identification of loci under selection [8, 10, 81–86], assuming that the bias of the genome-wide estimator carries over uniformly to all single-locus estimates. Our convergence calculations in section **Assessing the accuracy of genome-wide ratio estimators** require large numbers of loci, so they do not apply to single-locus estimates. Moreover, various methods for single-locus F_{ST} estimation for multiple alleles suffer from a strong dependence to the maximum allele frequency and heterozygosity [83–85, 87–90] that suggests that a more complicated bias is present in these single-locus F_{ST} estimators.

We have shown that the misapplication of existing F_{ST} estimators for independent subpopulations may lead to downwardly-biased estimates that can approach zero even when the true generalized F_{ST} is large. Weir-Cockerham [17], Weir-Hill [4], HudsonK (which generalizes the Hudson pairwise F_{ST} estimator [23] to K independent populations; also equals the Weir-Goudet approach for subpopulations [21]; S1 Text), and BayeScan [10]F_{ST} estimates in our admixture simulation are all smaller than the F_{ST} target by nearly a factor of two (Fig 4B), and differ from our new F_{ST} estimates in humans by nearly a factor of three [59]. To be accurate, existing F_{ST} estimators require independent subpopulations, so the observed biases arise from their misapplication to subpopulations that are neither independent not homogeneous. Nevertheless, natural populations—particularly humans—often do not adhere to the independent subpopulations model [59, 91–95].

The standard kinship coefficient estimator we investigated is often used to control for population structure in GWAS and to estimate genome-wide heritability [16, 27, 30–35]. While this estimator was known to be biased [16, 35], no closed-form limit had been calculated until very recently [21, 62]. These kinship estimates are biased downwards on average, but bias also varies for each pair of individuals (Figs 1 and 5). Thus, the use of these distorted kinship estimates may be problematic in GWAS or for estimating heritability, but the extent of the problem remains to be determined.

We developed a theoretical framework for assessing 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 loci (S1 Text). Our result justifies approximating the expectation of a ratio-of-means estimator with the ratio of expectations [4, 17, 23]. However, mean-of-ratios estimators may not converge to the ratio of expectations for infinite loci. 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. Therefore, we recommend ratio-of-means estimators, whose asymptotic behavior is well understood.

Our new framework enables accurate F_{ST} estimation in more complex datasets than before, but challenges remain. One challenge is the estimation of local inbreeding coefficients, which are required for estimating the generalized F_{ST} when not all individuals are locally outbred. To this end, we suggest employing existing approaches that infer inbreeding from large runs of homozygosity or related strategies [66–68], particularly when such self-IBD blocks are much larger than observed between individuals in the same subpopulation. A streamlined approach for jointly estimating total and local inbreeding is desirable, but will require an appropriate evaluation featuring realistic simulation of local inbreeding in a complex population structure. Another challenge is the estimation of the minimum kinship value without the use of subpopulation labels, so that accurate F_{ST} estimates can be obtained with even less user supervision. A more general unsupervised method could better ensure accuracy under extreme cases, such as when there are few unrelated individual pairs. These challenges can be overcome with the estimators we have presented, although supervision is needed to ensure that local inbreeding and the minimum kinship are estimated correctly.

We have demonstrated the need for new models and methods to study complex population structures, and have proposed a new approach for kinship and F_{ST} estimation that provides nearly unbiased estimates in this setting. Extending our implementation to deliver consistent accuracy in arbitrary population structures will require further innovation, and the results provided here may be useful in leading to more robust estimators in the future.

Methods

Previous F_{ST} estimators for the independent subpopulations model

Here we summarize the previous Weir-Cockerham, Weir-Hill, and Hudson F_{ST} estimators for independent subpopulations and derive the generalized HudsonK estimator for more than two subpopulations (which also equals the recent Weir-Goudet F_{ST} estimator for subpopulations under biallelic loci; S1 Text). We show that each of these estimators reduces, under infinite subpopulation sizes, to $\hat{F}_{\mathrm{ST}}^{\mathrm{indep}}$ in Eqs (11) to (13) that was studied in the results. In this section, let i index the m loci, j index the n subpopulations, n_j be the number of individuals sampled from subpopulation j, and $\hat{p}_{ij}$ be the sample reference allele frequency at locus i in subpopulation j.

The Weir-Cockerham F_{ST} estimator. The Weir-Cockerham (WC) F_{ST} estimator [17] estimates the coancestry parameter θ^T shared by each of the n independent subpopulation in consideration. Let $\hat{h}_{ij}$ denote the fraction of heterozygotes in subpopulation j for locus i. The ratio-of-means WC F_{ST} estimator and its limit for independent subpopulations ($\theta_{jk}^T = 0$ for $j \neq k$) with equal differentiation ($\theta_{jj}^T = \theta^T$) is

$$\bar{n} = \frac{1}{n}\sum_{j=1}^{n} n_j, \quad C^2 = \frac{1}{\bar{n}^2(n-1)}\sum_{j=1}^{n}(n_j - \bar{n})^2,$$

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

$$\hat{F}_{\mathrm{ST}}^{\mathrm{WC}} = \frac{\sum_{i=1}^{m}\hat{\sigma}_i^2 - \frac{1}{\bar{n}-1}\left(\hat{p}_i^T(1-\hat{p}_i^T) - \frac{n-1}{n}\hat{\sigma}_i^2 - \frac{1}{4}\bar{h}_i\right)}{\sum_{i=1}^{m}\hat{p}_i^T(1-\hat{p}_i^T)\left(1 - \frac{\bar{n}C^2}{n(\bar{n}-1)}\right) + \frac{1}{n}\hat{\sigma}_i^2\left(1 + \frac{(n-1)\bar{n}C^2}{n(\bar{n}-1)}\right) + \frac{\bar{h}_i C^2}{4n(\bar{n}-1)}}$$

$$\xrightarrow[m\to\infty]{\text{a.s.}} F_{\mathrm{ST}} = \theta^T.$$

Note that $\hat{p}_i^T$ above weighs every individual equally by weighing subpopulation j proportional to its sample size n_j, so it equals the estimator in Eq (17) with uniform weights.

Now we simplify this estimator as the sample size of every subpopulation becomes infinite. First set the sample size of every subpopulation n_j equal to their mean $\bar{n}$, which implies $C^2 = 0$ and

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

$$\hat{F}_{\mathrm{ST}}^{\mathrm{WC}} = \frac{\sum_{i=1}^{m}\hat{\sigma}_i^2 - \frac{1}{\bar{n}-1}\left(\hat{p}_i^T(1-\hat{p}_i^T) - \frac{n-1}{n}\hat{\sigma}_i^2 - \frac{1}{4}\bar{h}_i\right)}{\sum_{i=1}^{m}\hat{p}_i^T(1-\hat{p}_i^T) + \frac{1}{n}\hat{\sigma}_i^2}.$$

Now we take the limit as the sample size $\bar{n} \to \infty$, which results in sample allele frequencies converging to the true subpopulation allele frequencies $\hat{p}_{ij} \to \pi_{ij}$ for every subpopulation j and locus i, and

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

which matches the $\hat{F}_{\mathrm{ST}}^{\mathrm{indep}}$ in Eqs (11) to (13) as desired. Note the number of subpopulations n remains finite, and the sample heterozygosity $\overline{h}_i$ is not needed in the limit.

The Weir-Hill F_{ST} estimator. Weir and Hill developed new estimators for subpopulation-specific F_{ST} values and considered the effects of non-independent subpopulations [4]. However, these estimators target linearly-transformed F_{ST} values, and recover the F_{ST} defined in Eq (9) only when subpopulations are independent [4], so we group them here with other estimators that strictly assume independent subpopulations. For simplicity, here we only consider the global F_{ST} estimator; the estimators of the coancestry matrix of the subpopulations was found to have the same overall linear transformation [4]. In the limit of infinite subpopulation sizes, this estimator also converges to the asymptotic F_{ST} estimator for independent subpopulations ($\hat{F}_{\mathrm{ST}}^{\mathrm{indep}}$) discussed in the main text.

The Weir-Hill (WH) F_{ST} estimator, simplified here for biallelic loci but extended to average over loci, and its limit, are given by

$$\hat{p}_i^T = \sum_{j=1}^{n} w_j \hat{p}_{ij}, \qquad w_j = \frac{n_j}{\sum_{j=1}^{n} n_j},$$

$$\hat{F}_{\mathrm{ST}}^{\mathrm{WH}} = 1 - \frac{\left(\sum_{j=1}^{n} n_j(1-w_j)\right)\left(\sum_{i=1}^{m}\sum_{j=1}^{n} w_j \tfrac{2n_j}{2n_j-1}\hat{p}_{ij}(1-\hat{p}_{ij})\right)}{\sum_{i=1}^{m}\sum_{j=1}^{n} n_j(\hat{p}_{ij}-\hat{p}_i^T)^2 + n_j(1-w_j)\hat{p}_{ij}(1-\hat{p}_{ij})} \xrightarrow[m\to\infty]{\text{a.s.}} \frac{F_{\mathrm{ST}} - \tilde{\theta}^T}{1-\tilde{\theta}^T},$$

where the target F_{ST} and $\tilde{\theta}^T$ both weigh individuals (rather than subpopulations) equally [4]:

$$F_{\mathrm{ST}} = \sum_{j=1}^{n} w_j \theta_{jj}^T, \qquad \tilde{\theta}^T = \frac{2}{1-\sum_{j=1}^{n} w_j^2}\sum_{j=2}^{n}\sum_{k=1}^{j-1} w_j w_k \theta_{jk}^T.$$

For equal sample sizes $n_j = n_S \forall j$, we have $w_j = \frac{1}{n}$, $n_{jc} = n_S\left(1-\frac{1}{n}\right)$, and the estimator becomes

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

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

Therefore, as sample sizes per subpopulation go to infinity ($n_S \to \infty$, which results in $\hat{p}_{ij} \to \pi_{ij}$ for every (i, j)), we again recover the desired limiting F_{ST} estimator for independent subpopulations ($\hat{F}_{\mathrm{ST}}^{\mathrm{indep}}$ in Eqs (11) to (13)).

The Hudson F_{ST} estimator. The Hudson pairwise F_{ST} estimator [23] measures the differentiation of two subpopulations (j, k). The estimator and its limit for two independent subpopulations ($\theta^T_{jk} = 0$) is

$$\hat{F}^{\text{Hudson}}_{\text{ST}} = \frac{\sum_{i=1}^{m} (\hat{p}_{ij} - \hat{p}_{ik})^2 - \frac{\hat{p}_{ij}(1-\hat{p}_{ij})}{2n_j - 1} - \frac{\hat{p}_{ik}(1-\hat{p}_{ik})}{2n_k - 1}}{\sum_{i=1}^{m} \hat{p}_{ij}(1-\hat{p}_{ik}) + \hat{p}_{ik}(1-\hat{p}_{ij})} \xrightarrow[m\to\infty]{\text{a.s.}} F_{\text{ST}} = \frac{\theta^T_{jj} + \theta^T_{kk}}{2}. \tag{40}$$

Generalized HudsonK F_{ST} estimator. Here we derive the "HudsonK" estimator (first made available in [58]), which generalizes the Hudson pairwise F_{ST} estimator in Eq (40) to n independent subpopulations. This estimator also equals the recent Weir-Goudet F_{ST} estimator for subpopulations [21] (for biallelic loci; S1 Text). Note that for independent subpopulations, the F_{ST} of all the subpopulations equals the mean pairwise F_{ST} of every pair of subpopulations:

$$\frac{1}{n^2} \sum_{j=1}^{n} \sum_{k=1}^{n} \left(\frac{\theta^T_{jj} + \theta^T_{kk}}{2} \right) = \frac{1}{n} \sum_{j=1}^{n} \theta^T_{jj} = F_{\text{ST}}.$$

For that reason, averaging numerators and denominators of the pairwise estimator in Eq (40) before computing the ratio, we obtain the generalized estimator and a limit under independent subpopulations of

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

$$\hat{F}^{\text{HudsonK}}_{\text{ST}} = \frac{\sum_{i=1}^{m} \hat{\sigma}^2_i - \frac{1}{n} \sum_{j=1}^{n} \frac{\hat{p}_{ij}(1-\hat{p}_{ij})}{2n_j - 1}}{\sum_{i=1}^{m} \hat{p}^T_i (1 - \hat{p}^T_i) + \frac{1}{n} \hat{\sigma}^2_i} \xrightarrow[m\to\infty]{\text{a.s.}} F_{\text{ST}} = \frac{1}{n} \sum_{j=1}^{n} \theta^T_{jj}.$$

Note that unlike the WC and Weir-Hill estimators, $\hat{p}^T_i$ above weighs every subpopulation equally, so every individual is weighed inversely proportional to the sample sizes n_j of their subpopulation j.

Like WC and Weir-Hill, $\hat{F}^{\text{HudsonK}}_{\text{ST}}$ simplifies to $\hat{F}^{\text{indep}}_{\text{ST}}$ in Eqs (11) to (13) in the limit of infinite sample sizes $n_j \to \infty$, where $\hat{p}_{ij} \to \pi_{ij}$ for every (i, j).

Simulations

Construction of subpopulation allele frequencies. We simulate $K = 10$ subpopulations S_u and $m = 300,000$ independent loci. Every locus i draws $p^T_i \sim \text{Uniform}(0.01, 0.5)$. We set $f^T_{S_u} = \frac{u}{K}\tau$, where $\tau \leq 1$ tunes F_{ST}. For the independent subpopulations model, $F_{\text{ST}} = \frac{1}{K} \sum_{u=1}^{K} f^T_{S_u} = \frac{\tau(K+1)}{2K}$, so $\tau = \frac{2KF_{\text{ST}}}{K+1}$ 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, $p^{S_u}_i$ values are drawn from the Balding-Nichols distribution,

$$p^{S_u}_i | T \sim \text{Beta}\left(p^T_i \left(\frac{1}{f^T_{S_u}} - 1 \right), (1 - p^T_i) \left(\frac{1}{f^T_{S_u}} - 1 \right) \right),$$

which results in subpopulation allele frequencies that obey the coancestry model of Eq (6), with $\mathrm{E}[p_i^{S_u}|T] = p_i^T$ and $\mathrm{Var}(p_i^{S_u}|T) = f_{S_u}^T p_i^T(1 - p_i^T)$ [3], as desired.

Random subpopulation sizes. We randomly generate sample sizes $\mathbf{r} = (r_u)$ for K subpopulations and $\sum_{u=1}^K r_u = n = 1000$ individuals, as follows. First, draw $\mathbf{x} \sim \text{Dirichlet}(1, \ldots, 1)$ of length K and $\mathbf{r} = \text{round}(n\,\mathbf{x})$. While $\min_u r_u < \frac{n}{3K}$, draw a new $\mathbf{r}$, to prevent small subpopulations (they do not occur in real data). Due to rounding, $\sum_{u=1}^K r_u$ may not equal n as desired. Thus, while $\delta = n - \sum_{u=1}^K r_u \neq 0$, a random u is updated to $r_u \leftarrow r_u + \text{sgn}(\delta)$, which brings δ closer to zero at every iteration. 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 independent subpopulations model (see section **The generalized F_{ST} for arbitrary population structures**), which HudsonK estimates.

Admixture proportions from 1D geography. We construct q_{ju} from random-walk migrations along a one-dimensional geography. Let x_u be the coordinate of intermediate subpopulation 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 5). Our simulation uses $x_u = u$ and $y_j = \frac{1}{2} + \frac{j-1}{n-1}K$, so the intermediate subpopulations span between 1 and K and individuals span between $\frac{1}{2}$ and $K + \frac{1}{2}$. For the F_{ST} estimators that require subpopulations, individual j is assigned to the nearest subpopulation S_u (the u that minimizes $|x_u - y_j|$; Fig 3D); these subpopulations have equal sample size, so $w_j = \frac{1}{n}$ is appropriate.

Choosing σ and τ. Here we find values for σ (controls q_{jk}) and τ (scales $f_{S_u}^T$) that give $s^T = \frac{1}{2}$ and $F_{\mathrm{ST}} = 0.1$ in the admixture model. In our simulation, $w_j = \frac{1}{n}$ and $f_{S_u}^T = \frac{u}{K}\tau$, so applying those parameters to Eq (39) gives $\theta_{jk}^T = \frac{\tau}{K}\sum_{u=1}^K u q_{ju} q_{ku}$ and $F_{\mathrm{ST}} = \frac{\tau}{nK}\sum_{j=1}^n \sum_{u=1}^K u q_{ju}^2$. Therefore,

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

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

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

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

Prediction intervals

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 $\alpha = \frac{n+1-2j}{n+1}$ [96]. 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 $f^T_{S_u}$ and q_{ju}; fixed sample sizes for the independent subpopulations model), but with $p^T_i, p^{S_u}_i, \pi_{ij}, x_{ij}$ drawn separately for each estimate.

BayeScan and Weir-Goudet implementations

Weir-Goudet (WG) kinship estimates [20–22] were calculated using the function `snpgdsIndivBeta` in the R package `SNPRelate` 1.20.1 available on Bioconductor and GitHub. We found identical estimates using the function `beta.dosage` in the R package `hierfstat` 0.4.30 available on GitHub. WG (individuals) F_{ST} estimates were computed from the kinship estimates as described in section **Comparison to the Weir-Goudet kinship estimator for individuals**.

BayeScan 2.1 was downloaded from http://cmpg.unibe.ch/software/BayeScan/. To estimate F_{ST}, first the per-subpopulation F_{ST} values were estimated across loci assuming no selection, then the global F_{ST} was given by the mean F_{ST} across subpopulations.

Software

An R package called `popkin`, which implements the kinship and F_{ST} estimation methods proposed here, is available on the Comprehensive R Archive Network (CRAN) at https://cran.r-project.org/package=popkin and on GitHub at https://github.com/StoreyLab/popkin.

An R package called `bnpsd`, which implements the BN-PSD admixture simulation, is available on CRAN at https://cran.r-project.org/package=bnpsd and on GitHub at https://github.com/StoreyLab/bnpsd.

An R package called `popkinsuppl`, which implements memory-efficient algorithms for the Weir-Cockerham, Weir-Hill, and HudsonK F_{ST} estimators, and the standard kinship estimator, is available on GitHub at https://github.com/OchoaLab/popkinsuppl.

Public code reproducing these analyses are available at https://github.com/StoreyLab/human-differentiation-manuscript.

Supporting information

S1 Text. Supplementary information. Includes mathematical proofs and other calculations, including proof of convergence of ratio-of-means estimators, proof that the Weir-Goudet F_{ST} estimator for subpopulations equals HudsonK, derivation of existing method-of-moment estimators, proof that F_{ST} and kinship estimator limits are constants with respect to the ancestral population T, mean coancestry bounds, moments of estimator building blocks, the derivation of our new kinship estimator, and proof that our estimator from our original 2016 manuscript is algebraically equivalent to the one presented here.
(PDF)

Author Contributions

Conceptualization: Alejandro Ochoa, John D. Storey.

Formal analysis: Alejandro Ochoa, John D. Storey.

Funding acquisition: Alejandro Ochoa, John D. Storey.

Investigation: Alejandro Ochoa, John D. Storey.

Methodology: Alejandro Ochoa, John D. Storey.

Software: Alejandro Ochoa.

Writing – original draft: Alejandro Ochoa, John D. Storey.

Writing – review & editing: Alejandro Ochoa, John D. Storey.

References

1. Malécot G. Mathématiques de l'hérédité. Masson et Cie; 1948.
2. Wright S. The genetical structure of populations. Ann Eugen. 1951; 15(4):323–354.
3. Balding DJ, Nichols RA. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. Genetica. 1995; 96(1-2):3–12. https://doi.org/10.1007/BF01441146
4. Weir BS, Hill WG. Estimating F-Statistics. Annual Review of Genetics. 2002; 36(1):721–750. https://doi.org/10.1146/annurev.genet.36.050802.093940
5. Nicholson G, Smith AV, Jónsson F, Gústafsson O, Stefánsson K, Donnelly P. Assessing population differentiation and isolation from single-nucleotide polymorphism data. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2002; 64(4):695–715. https://doi.org/10.1111/1467-9868.00357
6. Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics. 2003; 164(4):1567–1587.
7. Balding DJ. Likelihood-based inference for genetic correlation coefficients. Theoretical Population Biology. 2003; 63(3):221–230. https://doi.org/10.1016/S0040-5809(03)00007-8
8. Beaumont MA, Balding DJ. Identifying adaptive genetic divergence among populations from genome scans. Molecular Ecology. 2004; 13(4):969–980. https://doi.org/10.1111/j.1365-294X.2004.02125.x
9. Foll M, Gaggiotti O. Identifying the Environmental Factors That Determine the Genetic Structure of Populations. Genetics. 2006; 174(2):875–891. https://doi.org/10.1534/genetics.106.059451
10. Foll M, Gaggiotti O. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. Genetics. 2008; 180(2):977–993. https://doi.org/10.1534/genetics.108.092221
11. Coop G, Witonsky D, Rienzo AD, Pritchard JK. Using Environmental Correlations to Identify Loci Underlying Local Adaptation. Genetics. 2010; 185(4):1411–1423. https://doi.org/10.1534/genetics.110.114819
12. Thompson EA. The estimation of pairwise relationships. Ann Hum Genet. 1975; 39(2):173–188. https://doi.org/10.1111/j.1469-1809.1975.tb00120.x
13. Milligan BG. Maximum-likelihood estimation of relatedness. Genetics. 2003; 163(3):1153–1167.
14. Jacquard A. Structures génétiques des populations. Paris: Masson et Cie; 1970.
15. Csűrös M. Non-identifiability of identity coefficients at biallelic loci. Theor Popul Biol. 2014; 92:22–29. https://doi.org/10.1016/j.tpb.2013.11.001
16. Astle W, Balding DJ. Population Structure and Cryptic Relatedness in Genetic Association Studies. Statist Sci. 2009; 24(4):451–471. https://doi.org/10.1214/09-STS307
17. Weir BS, Cockerham CC. Estimating F-Statistics for the Analysis of Population Structure. Evolution. 1984; 38(6):1358–1370. https://doi.org/10.1111/j.1558-5646.1984.tb05657.x
18. Weir BS, Cardon LR, Anderson AD, Nielsen DM, Hill WG. Measures of human population structure show heterogeneity among genomic regions. Genome Res. 2005; 15(11):1468–1476. https://doi.org/10.1101/gr.4398405
19. Buckleton J, Curran J, Goudet J, Taylor D, Thiery A, Weir BS. Population-specific FST values for forensic STR markers: A worldwide survey. Forensic Science International: Genetics. 2016; 23:91–100. https://doi.org/10.1016/j.fsigen.2016.03.004
20. Weir B, Zheng X. SNPs and SNVs in forensic science. Forensic Science International: Genetics Supplement Series. 2015; 5(Dec):e267–e268.
21. Weir BS, Goudet J. A Unified Characterization of Population Structure and Relatedness. Genetics. 2017; 206(4):2085–2103. https://doi.org/10.1534/genetics.116.198424
22. Goudet J, Kay T, Weir BS. How to estimate kinship. Mol Ecol. 2018; 27(20):4121–4135. https://doi.org/10.1111/mec.14833

23. Bhatia G, Patterson N, Sankararaman S, Price AL. Estimating and interpreting FST: the impact of rare variants. Genome Res. 2013; 23(9):1514–1521. https://doi.org/10.1101/gr.154831.113

24. Xie C, Gessler DD, Xu S. Combining different line crosses for mapping quantitative trait loci using the identical by descent-based variance component method. Genetics. 1998; 149(2):1139–1146.

25. Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet. 2006; 38(2):203–208. https://doi.org/10.1038/ng1702 PMID: 16380716

26. Aulchenko YS, de Koning DJ, Haley C. Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. Genetics. 2007; 177(1):577–585. https://doi.org/10.1534/genetics.107.075614

27. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006; 38(8):904–909. https://doi.org/10.1038/ng1847

28. Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, et al. Efficient control of population structure in model organism association mapping. Genetics. 2008; 178(3):1709–1723. https://doi.org/10.1534/genetics.107.080101 PMID: 18385116

29. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, et al. Variance component model to account for sample structure in genome-wide association studies. Nat Genet. 2010; 42(4):348–354. https://doi.org/10.1038/ng.548 PMID: 20208533

30. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nat Genet. 2012; 44(7):821–824. https://doi.org/10.1038/ng.2310

31. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 2010; 42(7):565–569. https://doi.org/10.1038/ng.608 PMID: 20562875

32. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011; 88(1):76–82. https://doi.org/10.1016/j.ajhg.2010.11.011

33. Rakovski CS, Stram DO. A kinship-based modification of the armitage trend test to address hidden population structure and small differential genotyping errors. PLoS ONE. 2009; 4(6):e5825. https://doi.org/10.1371/journal.pone.0005825

34. Thornton T, McPeek MS. ROADTRIPS: case-control association testing with partially or completely unknown population and pedigree structure. Am J Hum Genet. 2010; 86(2):172–184. https://doi.org/10.1016/j.ajhg.2010.01.001

35. Speed D, Balding DJ. Relatedness in the post-genomic era: is it still useful? Nat Rev Genet. 2015; 16 (1):33–44. https://doi.org/10.1038/nrg3821

36. Wang B, Sverdlov S, Thompson E. Efficient Estimation of Realized Kinship from SNP Genotypes. Genetics. 2017; p. genetics.116.197004. https://doi.org/10.1534/genetics.116.197004

37. Wright S. Systems of Mating. V. General Considerations. Genetics. 1921; 6(2):167–178. https://doi.org/10.1093/genetics/6.2.167

38. Lush JL. Heritability of Quantitative Characters in Farm Animals. Hereditas. 1949; 35(S1):356–375. https://doi.org/10.1111/j.1601-5223.1949.tb03347.x

39. Falconer DS, Mackay TFC. Introduction to Quantitative Genetics. 4th ed. Harlow: Pearson; 1996.

40. Thompson EA. Identity by descent: variation in meiosis, across genomes, and in populations. Genetics. 2013; 194(2):301–326. https://doi.org/10.1534/genetics.112.148825

41. Slatkin M. Inbreeding coefficients and coalescence times. Genetics Research. 1991; 58(2):167–175. https://doi.org/10.1017/S0016672300029827

42. Emik LO, Terrill CE. Systematic procedures for calculating inbreeding coefficients. J Hered. 1949; 40 (2):51–55. https://doi.org/10.1093/oxfordjournals.jhered.a105986

43. García-Cortés LA. A novel recursive algorithm for the calculation of the detailed identity coefficients. Genetics Selection Evolution. 2015; 47(1):33. https://doi.org/10.1186/s12711-015-0108-6

44. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, et al. Genetic Structure of Human Populations. Science. 2002; 298(5602):2381–2385. https://doi.org/10.1126/science.1078311 PMID: 12493913

45. Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proc Natl Acad Sci U S A. 2005; 102(44):15942–15947. https://doi.org/10.1073/pnas.0507611102

46. Consortium TGP. A map of human genome variation from population-scale sequencing. Nature. 2010; 467(7319):1061–1073. https://doi.org/10.1038/nature09534

47. Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, et al. Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature. 2014; 513(7518):409–413. https://doi.org/10.1038/nature13673 PMID: 25230663

48. Lazaridis I, Nadel D, Rollefson G, Merrett DC, Rohland N, Mallick S, et al. Genomic insights into the origin of farming in the ancient Near East. Nature. 2016; 536(7617):419–424. https://doi.org/10.1038/nature19310 PMID: 27459054

49. Skoglund P, Posth C, Sirak K, Spriggs M, Valentin F, Bedford S, et al. Genomic insights into the peopling of the Southwest Pacific. Nature. 2016; 538(7626):510–513. https://doi.org/10.1038/nature19844 PMID: 27698418

50. Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, et al. The Genetic Structure and History of Africans and African Americans. Science. 2009; 324(5930):1035–1044. https://doi.org/10.1126/science.1172257 PMID: 19407144

51. Moreno-Estrada A, Gravel S, Zakharia F, McCauley JL, Byrnes JK, Gignoux CR, et al. Reconstructing the Population Genetic History of the Caribbean. PLOS Genetics. 2013; 9(11):e1003925. https://doi.org/10.1371/journal.pgen.1003925 PMID: 24244192

52. Moreno-Estrada A, Gignoux CR, Fernández-López JC, Zakharia F, Sikora M, Contreras AV, et al. The genetics of Mexico recapitulates Native American substructure and affects biomedical traits. Science. 2014; 344(6189):1280–1285. https://doi.org/10.1126/science.1251688 PMID: 24926019

53. Leslie S, Winney B, Hellenthal G, Davison D, Boumertit A, Day T, et al. The fine-scale genetic structure of the British population. Nature. 2015; 519(7543):309–314. https://doi.org/10.1038/nature14230 PMID: 25788095

54. Baharian S, Barakatt M, Gignoux CR, Shringarpure S, Errington J, Blot WJ, et al. The Great Migration and African-American Genomic Diversity. PLoS Genet. 2016; 12(5):e1006059. https://doi.org/10.1371/journal.pgen.1006059 PMID: 27232753

55. Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, et al. Massive migration from the steppe was a source for Indo-European languages in Europe. Nature. 2015; 522(7555):207–211. https://doi.org/10.1038/nature14317 PMID: 25731166

56. Allentoft ME, Sikora M, Sjögren KG, Rasmussen S, Rasmussen M, Stenderup J, et al. Population genomics of Bronze Age Eurasia. Nature. 2015; 522(7555):167–172. https://doi.org/10.1038/nature14507 PMID: 26062507

57. Ochoa A, Storey JD. F_{ST} and kinship for arbitrary population structures I: Generalized definitions. bioRxiv. 2016; https://doi.org/10.1101/083915

58. Ochoa A, Storey JD. F_{ST} and kinship for arbitrary population structures II: Method of moments estimators. bioRxiv. 2016; https://doi.org/10.1101/083923

59. Ochoa A, Storey JD. New kinship and F_{ST} estimates reveal higher levels of differentiation in the global human population. bioRxiv. 2019; https://doi.org/10.1101/653279

60. Thornton T, Tang H, Hoffmann TJ, Ochs-Balcom HM, Caan BJ, Risch N. Estimating kinship in admixed populations. Am J Hum Genet. 2012; 91(1):122–138. https://doi.org/10.1016/j.ajhg.2012.05.024

61. Hao W, Song M, Storey JD. Probabilistic models of genetic variation in structured populations applied to global human studies. Bioinformatics. 2016; 32(5):713–721. https://doi.org/10.1093/bioinformatics/btv641

62. Zheng X, Weir BS. Eigenanalysis of SNP data with an identity by descent interpretation. Theoretical Population Biology. 2016; 107:65–76. https://doi.org/10.1016/j.tpb.2015.09.004

63. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000; 155(2):945–959.

64. Tang H, Peng J, Wang P, Risch NJ. Estimation of individual admixture: analytical and study design considerations. Genet Epidemiol. 2005; 28(4):289–301. https://doi.org/10.1002/gepi.20064

65. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009; 19(9):1655–1664. https://doi.org/10.1101/gr.094052.109

66. Browning BL, Browning SR. A Fast, Powerful Method for Detecting Identity by Descent. The American Journal of Human Genetics. 2011; 88(2):173–182. https://doi.org/10.1016/j.ajhg.2011.01.010

67. Gazal S, Sahbatou M, Perdry H, Letort S, Génin E, Leutenegger AL. Inbreeding Coefficient Estimation with Dense SNP Data: Comparison of Strategies and Application to HapMap III. HHE. 2014; 77(1-4):49–62.

68. Joshi PK, Esko T, Mattsson H, Eklund N, Gandin I, Nutile T, et al. Directional dominance on stature and cognition in diverse human populations. Nature. 2015; 523(7561):459–462. https://doi.org/10.1038/nature14618 PMID: 26131930

69. Cochran WG. Sampling techniques. 3rd ed. Wiley; 1977.

70. Reynolds J, Weir BS, Cockerham CC. Estimation of the Coancestry Coefficient: Basis for a Short-Term Genetic Distance. Genetics. 1983; 105(3):767–779. https://doi.org/10.1093/genetics/105.3.767

71. Weir BS. Genetic data analysis II. Methods for discrete population genetic data. Sunderland, USA: Sinauer Associates; 1996.

72. Bourgain C, Hoffjan S, Nicolae R, Newman D, Steiner L, Walker K, et al. Novel case-control test in a founder population identifies P-selectin as an atopy-susceptibility locus. Am J Hum Genet. 2003; 73 (3):612–626. https://doi.org/10.1086/378208 PMID: 12929084

73. Choi Y, Wijsman EM, Weir BS. Case-Control Association Testing in the Presence of Unknown Relationships. Genet Epidemiol. 2009; 33(8):668–678. https://doi.org/10.1002/gepi.20418

74. Pickrell JK, Pritchard JK. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet. 2012; 8(11):e1002967. https://doi.org/10.1371/journal.pgen.1002967

75. Nei M. Analysis of Gene Diversity in Subdivided Populations. PNAS. 1973; 70(12):3321–3323. https://doi.org/10.1073/pnas.70.12.3321

76. Weir BS, Goudet J. A unified characterization of population structure and relatedness. bioRxiv. 2016; p. 088260.

77. Raj A, Stephens M, Pritchard JK. fastSTRUCTURE: variational inference of population structure in large SNP data sets. Genetics. 2014; 197(2):573–589. https://doi.org/10.1534/genetics.114.164350

78. Nelis M, Esko T, Mägi R, Zimprich F, Zimprich A, Toncheva D, et al. Genetic Structure of Europeans: A View from the North–East. PLOS ONE. 2009; 4(5):e5472. https://doi.org/10.1371/journal.pone.0005472 PMID: 19424496

78. Silva NM, Pereira L, Poloni ES, Currat M. Human Neutral Genetic Variation and Forensic STR Data. PLOS ONE. 2012; 7(11):e49666. https://doi.org/10.1371/journal.pone.0049666

80. Steele CD, Court DS, Balding DJ. Worldwide FST Estimates Relative to Five Continental-Scale Populations. Annals of Human Genetics. 2014; 78(6):468–477. https://doi.org/10.1111/ahg.12081

81. Cavalli-Sforza LL. Population Structure and Human Evolution. Proceedings of the Royal Society of London Series B, Biological Sciences. 1966; 164(995):362–379.

82. Lewontin RC, Krakauer J. Distribution of Gene Frequency as a Test of the Theory of the Selective Neutrality of Polymorphisms. Genetics. 1973; 74(1):175–195. https://doi.org/10.1093/genetics/74.1.175

83. Beaumont MA, Nichols RA. Evaluating Loci for Use in the Genetic Analysis of Population Structure. Proceedings of the Royal Society of London B: Biological Sciences. 1996; 263(1377):1619–1626. https://doi.org/10.1098/rspb.1996.0237

84. Vitalis R, Dawson K, Boursot P. Interpretation of Variation Across Marker Loci as Evidence of Selection. Genetics. 2001; 158(4):1811–1823.

85. Akey JM, Zhang G, Zhang K, Jin L, Shriver MD. Interrogating a high-density SNP map for signatures of natural selection. Genome Res. 2002; 12(12):1805–1814. https://doi.org/10.1101/gr.631202

86. Porter AH. A test for deviation from island-model population structure. Molecular Ecology. 2003; 12 (4):903–915. https://doi.org/10.1046/j.1365-294X.2003.01783.x

87. Bowcock AM, Kidd JR, Mountain JL, Hebert JM, Carotenuto L, Kidd KK, et al. Drift, admixture, and selection in human evolution: a study with DNA polymorphisms. PNAS. 1991; 88(3):839–843. https://doi.org/10.1073/pnas.88.3.839 PMID: 1992475

88. Hedrick PW. A Standardized Genetic Differentiation Measure. Evolution. 2005; 59(8):1633–1638. https://doi.org/10.1111/j.0014-3820.2005.tb01814.x

89. Jakobsson M, Edge MD, Rosenberg NA. The Relationship Between FST and the Frequency of the Most Frequent Allele. Genetics. 2013; 193(2):515–528. https://doi.org/10.1534/genetics.112.144758

90. Edge MD, Rosenberg NA. Upper bounds on FST in terms of the frequency of the most frequent allele and total homozygosity: the case of a specified number of alleles. Theor Popul Biol. 2014; 97:20–34.

91. Lewontin RC. The Apportionment of Human Diversity. Evolutionary Biology. 1972; 6:381–398.

92. Barbujani G, Magagni A, Minch E, Cavalli-Sforza LL. An apportionment of human DNA diversity. PNAS. 1997; 94(9):4516–4519. https://doi.org/10.1073/pnas.94.9.4516

93. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, et al. Genes mirror geography within Europe. Nature. 2008; 456(7218):98–101. https://doi.org/10.1038/nature07331 PMID: 18758442

94. Coop G, Pickrell JK, Novembre J, Kudaravalli S, Li J, Absher D, et al. The Role of Geography in Human Adaptation. PLoS Genet. 2009; 5(6):e1000500. https://doi.org/10.1371/journal.pgen.1000500 PMID: 19503611

95. Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, et al. Ancient admixture in human history. Genetics. 2012; 192(3):1065–1093. https://doi.org/10.1534/genetics.112.145037 PMID: 22960212

96. Beran R, Hall P. Interpolated Nonparametric Prediction Intervals and Confidence Intervals. Journal of the Royal Statistical Society Series B (Methodological). 1993; 55(3):643–652. https://doi.org/10.1111/j.2517-6161.1993.tb01929.x