bioRxiv preprint doi: https://doi.org/10.1101/083923; this version posted June 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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

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

[1]Duke Center for Statistical Genetics and Genomics, and [2]Department of Biostatistics and Bioinformatics, Duke University, Durham, NC 27705, USA
[3]Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544, USA
∗ Corresponding author: `jstorey@princeton.edu`

Abstract: F_{ST} and kinship are key parameters often estimated in modern population genetics studies 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 provide new results on 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 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.

Note: This article is Part II of two-part manuscripts. We refer to these in the text as Part I and Part II, respectively.

Part I: Alejandro Ochoa and John D. Storey. "F_{ST} and kinship for arbitrary population structures I: Generalized definitions". *bioRxiv* (10.1101/083915) (2019). `https://doi.org/10.1101/083915`. First published 2016-10-27.

Part II: Alejandro Ochoa and John D. Storey. "F_{ST} and kinship for arbitrary population structures II: Method of moments estimators". *bioRxiv* (10.1101/083923) (2019). `https://doi.org/10.1101/083923`. First published 2016-10-27.

bioRxiv preprint doi: https://doi.org/10.1101/083923; this version posted June 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Contents

1 Introduction **4**

2 Assessing the accuracy of genome-wide estimators **6**
2.1 Ratio estimators . . . 6
2.2 Convergence . . . 8

3 $F_{\mathbf{ST}}$ estimation based on the independent subpopulations model **9**
3.1 The F_{ST} estimator for independent subpopulations and infinite subpopulation sample sizes . . . 9
3.2 F_{ST} estimation under the independent subpopulations model . . . 10
3.3 F_{ST} estimation under arbitrary coancestry . . . 11
3.4 Coancestry estimation as a method of moments . . . 12

4 Characterizing a kinship estimator and its relationship to $F_{\mathbf{ST}}$ **12**
4.1 Characterization of the standard kinship estimator . . . 13
4.2 Estimation of coancestry coefficients from IAFs . . . 14
4.3 F_{ST} estimator based on the standard kinship estimator . . . 15
4.4 Adjusted consistent oracle F_{ST} estimators and the "bias coefficient" . . . 16

5 A new approach for kinship and $F_{\mathbf{ST}}$ estimation **17**
5.1 General approach . . . 17
5.2 Proof-of-principle kinship estimator using subpopulation labels . . . 18

6 Simulations evaluating $F_{\mathbf{ST}}$ and kinship estimators **19**
6.1 Overview of simulations . . . 19
6.2 Evaluation of F_{ST} estimators . . . 21
6.3 Evaluation of kinship estimators . . . 23
6.4 Evaluation of oracle adjusted F_{ST} estimators . . . 25

7 Discussion **25**

S1 Accuracy of ratio estimators **S1**
S1.1 Almost sure convergence of ratio-of-means estimators with independent and uniformly-bounded terms . . . S1
S1.2 Order of error of expectations . . . S1

S2 Previous $F_{\mathbf{ST}}$ estimators for the independent subpopulations model **S2**
S2.1 The Weir-Cockerham F_{ST} estimator . . . S2
S2.2 The Hudson F_{ST} estimator . . . S3

bioRxiv preprint doi: https://doi.org/10.1101/083923; this version posted June 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

S2.3 Generalized HudsonK F_{ST} estimator S3

S3 Derivation of method-of-moment estimators **S4**

S3.1 F_{ST} estimator for independent subpopulations S4

S3.2 Standard kinship estimator S5

S4 Proofs that F_{ST} and kinship estimator limits are constants with respect to the ancestral population T **S6**

S4.1 Proof that the limit of $\hat{F}_{\mathrm{ST}}^{\mathrm{indep}}$ does not depend on T S6

S4.2 Proof that the limit of $\hat{\varphi}_{jk}^{T,\mathrm{std}}$ does not depend on T S7

S5 Mean coancestry bounds **S7**

S6 Moments of estimator building blocks **S8**

S7 Derivation of new kinship estimator **S9**

S8 Admixture and independent subpopulations model simulations **S10**

S8.1 Construction of subpopulation allele frequencies S10

S8.2 Random subpopulation sizes S10

S8.3 Admixture proportions from 1D geography S10

S8.4 Choosing σ and τ S11

S9 Prediction intervals of F_{ST} estimators **S11**

bioRxiv preprint doi: https://doi.org/10.1101/083923; this version posted June 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

1 Introduction

In population genetics studies, one is often interested in characterizing structure, genetic differentiation, and relatedness among individuals. Two quantities often considered in this context are F_{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 [3, 4]. The kinship coefficient is a measure of relatedness between individuals defined in terms of IBD probabilities, and it is closely related to F_{ST} [3].

This work focuses on the estimation of F_{ST} and kinship from biallelic single-nucleotide polymorphism (SNP) marker 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} [5–13] or outbred individuals for kinship [14, 15]. Additionally, more complete likelihood models such as that of [16] are underdetermined for biallelic loci [17]. Non-parametric approaches such as those based on the method of moments are considerably more flexible and computationally tractable [18], 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 [19]. The "Hudson" F_{ST} estimator [20] assumes two subpopulations with different F_{ST} values. These 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 [6, 19, 20]. We also evaluate BayeScan [12], 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 the WC and Hudson 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) [18, 21–27] and to estimate genome-wide heritability [28, 29]. Often absent in previous works 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

bioRxiv preprint doi: https://doi.org/10.1101/083923; this version posted June 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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 [18, 24, 27–33] is also a method-of-moments estimator whose operating characteristics are largely unknown in the presence of structure. We show in Section 4 that this popular estimator is accurate only when the average kinship is zero, which implies that the population must be unstructured.

Recent genome-wide studies have revealed that humans and other natural populations are structured in a complex manner that 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 [34, 35], the 1000 Genomes Project [36], Human Origins [37–39], and other contemporary [40–44] and archaic populations [45, 46]. We have also demonstrated, based on the work in Part I and Part II here, that the global human population has a complex kinship matrix and no independent subpopulations [47]. 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.

We generalized the definition of F_{ST} for arbitrary population structures in Section 3 of Part I. Additionally, we derived connections between F_{ST} and three models: arbitrary kinship coefficients [3, 16] in Section 3 of Part I (panel "Kinship Model" in Fig. 1), individual-specific allele frequencies [48, 49] in Section 5 of Part I (panels "Coancestry Model" and "Coancestry in Terms of Kinship" in Fig. 1), and admixture models [50–52] in Section 6 of Part I.

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 2 we obtain new strong convergence results for a family of ratio estimators that includes the most common F_{ST} and kinship estimators. Next, we calculate the convergence values of these F_{ST} (Section 3) and kinship (Section 4) 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 for the first time, identifying complex biases or distortions that have not been described before (related results were independently and concurrently calculated by [53]). In Section 5 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). Lastly, in Section 6 we construct an admixture simulation that does not have independent subpopulations to illustrate our theoretical findings through simulation. 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 [47]. In summary, we identify a new approach for unbiased estimation of F_{ST} and kinship, and we provide

bioRxiv preprint doi: https://doi.org/10.1101/083923; this version posted June 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

new estimators that are nearly unbiased.

2 Assessing the accuracy of genome-wide estimators

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 [54]. In the F_{ST} literature, the expectation of a ratio is frequently approximated with a ratio of expectations [6, 19, 20]. Specifically, ratio estimators are often called "unbiased" if the ratio of expectations is unbiased, even though the ratio estimator itself may be biased [54]. Here we characterize the behavior of two ratio estimator families calculated from genome-wide data, detailing conditions where the previous approximation is justified and providing additional criteria to assess the accuracy of such estimators. These convergence results are the foundation of our analysis of estimators and are applied repeatedly to the various kinship and F_{ST} estimators discussed in Sections 3 to 5.

2.1 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^T_{jk}), while c_i depends on the ancestral allele frequency p^T_i 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 [54]. Below we study two estimator families that combine large numbers of loci to better estimate $\frac{A}{B}$.

Figure 1 *(following page)*: **Accuracy of F_{ST} and kinship estimators: overview of models and results.** Our analysis is based on two parallel models: the "Coancestry Model" for individual-specific allele frequencies (π_{ij}; Section 5 of Part I), and the "Kinship Model" for genotypes (x_{ij}; Section 3.5 of Part I). The "Coancestry in Terms of Kinship" panel connects kinship (φ^T_{jk}, f^T_j) and coancestry (θ^T_{jk}) parameters (proven in Section 5.2 of Part I). 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 3), 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 4), 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 5). Note that estimation of F_{ST} from genotypes requires individuals to be locally outbred and locally unrelated (see Sections 3.2 and 3.3 of Part I).

bioRxiv preprint doi: https://doi.org/10.1101/083923; this version posted June 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Coancestry in Terms of Kinship

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

Kinship Model

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

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

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

Coancestry Model

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

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

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

$$F_{\mathrm{ST}} = \sum_{j=1}^{n} w_j \theta_{jj}^T$$

Existing Kinship Estimator

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

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

Indep. Subpop. F_{ST} Estimator

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

New Kinship Estimator

$$A_{jk} \xrightarrow[m\to\infty]{\text{a.s.}} \left(\varphi_{jk}^T - 1\right) v^T$$

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

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

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

bioRxiv preprint doi: https://doi.org/10.1101/083923; this version posted June 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

2.2 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 [6, 19, 20, 55]. Note that $\mathrm{E}\left[\hat{A}_m\right] = A\bar{c}_m$ and $\mathrm{E}\left[\hat{B}_m\right] = B\bar{c}_m$, where $\bar{c}_m = \frac{1}{m}\sum_{i=1}^{m} c_i$. We will assume bounded terms ($|a_i|, |b_i| \leq C$ for some finite C), a convergent $\bar{c}_m \to c$, and $Bc \neq 0$, which are satisfied by common estimators. Given independent loci, we prove almost sure convergence to the desired quantity (Supplementary Information, Section S1.1),

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

a strong result that implies $\mathrm{E}\left[\frac{\hat{A}_m}{\hat{B}_m}\right] \to \frac{A}{B}$, justifying previous work [6, 19, 20]. Moreover, the error between these expectations scales with $\frac{1}{m}$ (Supplementary Information, Section S1.2), just as for standard ratio estimators [54]. 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. (1), the following three conditions must be met. (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^T_{jk} and F_{ST}, while c_i is a function of p^T_i 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. (1) holds for independent loci and the A and B found in the first step. See Section 3.2 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 [18, 24, 27–32] and F_{ST} [36]. If each $\frac{a_i}{b_i}$ is biased, their average across loci will also be biased, even as $m \to \infty$. However, if $\mathrm{E}\left[\frac{a_i}{b_i}\right] \to \frac{A}{B}$ for all loci $i = 1, \ldots, m$ as the number of individuals $n \to \infty$, and $\mathrm{Var}\left(\frac{a_i}{b_i}\right)$ is bounded, then

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

Therefore, mean-of-ratios estimators must satisfy more restrictive conditions than ratio-of-means estimators, as well as large n (in addition to the large m needed by both estimators), to estimate

bioRxiv preprint doi: https://doi.org/10.1101/083923; this version posted June 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

$\frac{A}{B}$ well. We do not provide a procedure to test whether a given mean-of-ratios estimator converges as shown above.

3 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 Weir-Cockerham (WC) [19] and "Hudson" [20] F_{ST} estimators, which assume the independent subpopulations model described above. The panel "Indep. Subpop. F_{ST} Estimator" in Fig. 1 provides an overview of our results, which we detail in this section.

3.1 The F_{ST} estimator for independent subpopulations and infinite subpopulation sample sizes

The WC 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 Supplementary Information, Section S2 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 (see Supplementary Information, Section S2 for details). This simplification corresponds to eliminating statistical sampling, leaving only genetic sampling to analyze [56]. Note that our simplified estimator nevertheless illustrates the general behavior of the WC and Hudson estimators under arbitrary structure, and the results are equivalent to those we would obtain under finite sample sizes of individuals. While the Hudson F_{ST} estimator compares two subpopulations [20], we derive a new generalized "HudsonK" estimator for more than two subpopulations in Supplementary Information, Section S2.3.

Under infinite subpopulation sample sizes, the allele frequencies at each locus and every subpopulation are known. Let $j \in \{1, ..., n\}$ index subpopulations rather than individuals and π_{ij} be the allele frequency in subpopulation j at locus i. We call the π_{ij} values "individual-specific allele frequencies" (IAF), as has been previously done [49]. In this special case, both WC and HudsonK simplify to the following F_{ST} estimator for independent subpopulations ("indep"; derived in

bioRxiv preprint doi: https://doi.org/10.1101/083923; this version posted June 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Supplementary Information, Section S2):

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

$$\hat{\sigma}_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(\pi_{ij} - \hat{p}_i^T\right)^2, \tag{3}$$

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

The goal is to estimate $F_{\mathrm{ST}} = \frac{1}{n}\sum_{j=1}^{n}\theta_{jj}^T$, which weighs every subpopulation j equally ($w_j = \frac{1}{n}\ \forall j$), under the coancestry model of Part I, which assumes the following moments for IAFs:

$$\mathrm{E}[\pi_{ij}|T] = p_i^T, \tag{5}$$

$$\mathrm{Cov}(\pi_{ij}, \pi_{ik}|T) = p_i^T\left(1-p_i^T\right)\theta_{jk}^T. \tag{6}$$

3.2 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. (4) can be derived directly from Eqs. (5) and (6) and these assumptions using the method of moments (ignoring the existence of previous F_{ST} estimators; Supplementary Information, Section S3.1). The expectations of the two recurrent terms in Eq. (4) are

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

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

$$\overline{p(1-p)}^T = \frac{1}{m}\sum_{i=1}^{m}p_i^T\left(1-p_i^T\right).$$

Eliminating $\overline{p(1-p)}^T$ and solving for F_{ST} in this system of equations recovers the estimator in Eq. (4).

Before applying the convergence result in Eq. (1), we test that the three conditions listed in Section 2 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 $A = F_{\mathrm{ST}}$, $B = 1$, and $c_i = p_i^T\left(1-p_i^T\right)$. Condition (ii): $\bar{c}_m \to c = \mathrm{E}\left[p_i^T\left(1-p_i^T\right)\right] \neq 0$ over the p_i^T distribution across loci. Condition (iii): Since

bioRxiv preprint doi: https://doi.org/10.1101/083923; this version posted June 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

$\pi_{ij}, \hat{p}_i^T \in [0,1]$, 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,

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

3.3 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 Eqs. (5) and (6), and now T may be either the MRCA population of all individuals or a more ancestral population. In this general setting, (j,k) may index either subpopulations or individuals. The two terms of $\hat{F}_{\mathrm{ST}}^{\mathrm{indep}}$ now satisfy

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

$$\mathrm{E}\left[\frac{1}{m}\sum_{i=1}^{m}\hat{p}_i^T\left(1-\hat{p}_i^T\right) \middle| T\right] = \overline{p(1-p)}^T \left(1 - \bar{\theta}^T\right),$$

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

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

$$\hat{F}_{\mathrm{ST}}^{\mathrm{indep}} \xrightarrow[m\to\infty]{\text{a.s.}} \frac{n\left(F_{\mathrm{ST}} - \bar{\theta}^T\right)}{n-1+F_{\mathrm{ST}} - n\bar{\theta}^T} = \frac{F_{\mathrm{ST}} - \frac{1}{n-1}\left(n\bar{\theta}^T - F_{\mathrm{ST}}\right)}{1 - \frac{1}{n-1}\left(n\bar{\theta}^T - F_{\mathrm{ST}}\right)}, \tag{7}$$

(the conclusion of panel "Indep. Subpop. F_{ST} Estimator" in Fig. 1), where it should be noted that

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

is the average of all between-individual coancestry coefficients, a term that appears in a related result for subpopulations [6]. Therefore, under arbitrary structure the independent subpopulations estimator's bias is due to the coancestry between individuals (or subpopulations in the traditional setting). While the limit in Eq. (7) appears to vary depending on the choice of T, it is in fact a constant with respect to T (proof in Supplementary Information, Section S4.1).

Since $\frac{1}{n}F_{\mathrm{ST}} \leq \bar{\theta}^T \leq F_{\mathrm{ST}}$ (Supplementary Information, Section S5), this estimator has a downward bias in the general setting: it is asymptotically unbiased ($\hat{F}_{\mathrm{ST}}^{\mathrm{indep}} \xrightarrow[m\to\infty]{\text{a.s.}} F_{\mathrm{ST}}$) only when $\bar{\theta}^T = \frac{1}{n}F_{\mathrm{ST}}$, while bias is maximal when $\bar{\theta}^T = F_{\mathrm{ST}}$, where $\hat{F}_{\mathrm{ST}}^{\mathrm{indep}} \xrightarrow[m\to\infty]{\text{a.s.}} 0$. For example, if $\min\theta_{jk}^T = 0$ so the MRCA population T is fixed, but n is large and $\theta_{jk}^T \approx F_{\mathrm{ST}}$ for most pairs of individuals, then $\bar{\theta}^T \approx F_{\mathrm{ST}}$ as well, and $\hat{F}_{\mathrm{ST}}^{\mathrm{std}} \approx 0$. Therefore, the magnitude of the bias of $\hat{F}_{\mathrm{ST}}^{\mathrm{indep}}$ is unknown if $\bar{\theta}^T$ is unknown, and small $\hat{F}_{\mathrm{ST}}^{\mathrm{indep}}$ may arise even if F_{ST} is very large.

bioRxiv preprint doi: https://doi.org/10.1101/083923; this version posted June 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

3.4 Coancestry estimation as a method of moments

Since the generalized F_{ST} is given by coancestry coefficients θ^T_{jj} (Eq. (13) of Part I), a new F_{ST} estimator could be derived from estimates of θ^T_{jj}. Here we attempt to define a method-of-moments estimator for θ^T_{jk}, and find an underdetermined estimation problem, just as for F_{ST}.

Given IAFs and Eqs. (5) and (6), the first and second moments that average across loci are

$$\mathrm{E}\left[\frac{1}{m}\sum_{i=1}^{m}\pi_{ij}\middle|T\right] = \bar{p}^T, \tag{8}$$

$$\mathrm{E}\left[\frac{1}{m}\sum_{i=1}^{m}\pi_{ij}\pi_{ik}\middle|T\right] = \overline{p^2}^T + \overline{p(1-p)}^T\theta^T_{jk}, \tag{9}$$

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

Suppose first that only θ^T_{jj} are of interest. There are n estimators given by Eq. (9) with $j = k$, each corresponding to an unknown θ^T_{jj}. However, all these estimators share two nuisance parameters: $\bar{p}^T$ and $\overline{p^2}^T$. While $\bar{p}^T$ can be estimated from Eq. (8), 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. (9) 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.

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

Given the biases we see for $\hat{F}^{\mathrm{indep}}_{\mathrm{ST}}$ under arbitrary structures in Section 3.3, we now turn to the generalized definition of F_{ST} and pursue an estimate of it. Recall from Eq. (3) of Part I that our generalized F_{ST} is defined in terms of inbreeding coefficients, which are a special case of the kinship coefficient:

$$F_{\mathrm{ST}} = \sum_{j=1}^{n} w_j f^T_{L_j}.$$

Therefore, we will first consider estimates of kinship and inbreeding in this section. Note also that estimating kinship is important for GWAS approaches that control for population structure [18, 21–32, 57, 58]. Lastly, kinship coefficients determine the bias of $\hat{F}^{\mathrm{indep}}_{\mathrm{ST}}$ in Eq. (7) (since coancestry and kinship coefficients are closely related: see panel "Coancestry in Terms of Kinship" in Fig. 1).

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 $\hat{F}^{\mathrm{indep}}_{\mathrm{ST}}$ (Section 3), 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 Section 5.

bioRxiv preprint doi: https://doi.org/10.1101/083923; this version posted June 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

4.1 Characterization of the standard kinship estimator

Here we analyze a standard kinship estimator that is frequently used [18, 24, 27–33]. We generalize this estimator to use weights in estimating the ancestral allele frequencies, and we write it as a ratio-of-means estimator due to the favorable theoretical properties of this format as detailed in Section 2:

$$\hat{p}_i^T = \frac{1}{2}\sum_{j=1}^{n} w_j x_{ij}, \tag{10}$$

$$\hat{\varphi}_{jk}^{T,\text{std}} = \frac{\sum_{i=1}^{m}\left(x_{ij} - 2\hat{p}_i^T\right)\left(x_{ik} - 2\hat{p}_i^T\right)}{4\sum_{i=1}^{m}\hat{p}_i^T\left(1-\hat{p}_i^T\right)}. \tag{11}$$

The estimator in Eq. (11) 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 $4p_i^T\left(1-p_i^T\right)$. We derive Eq. (11) directly using the method of moments in Supplementary Information, Section S3.2. The weights in Eq. (10) must satisfy $w_j > 0$ and $\sum_{j=1}^{n} w_j = 1$, so $\hat{p}_i^T \in [0,1]$ and $\mathrm{E}\left[\hat{p}_i^T \middle| T\right] = p_i^T$.

Utilizing the following moments for genotypes (from the kinship model of Part I),

$$\mathrm{E}[x_{ij}|T] = 2p_i^T, \tag{12}$$

$$\mathrm{Cov}(x_{ij}, x_{ik}|T) = 4p_i^T\left(1-p_i^T\right)\varphi_{jk}^T, \tag{13}$$

we find that Eq. (11) converges to

$$\hat{\varphi}_{jk}^{T,\text{std}} \xrightarrow[m\to\infty]{\text{a.s.}} \frac{\varphi_{jk}^T - \bar{\varphi}_j^T - \bar{\varphi}_k^T + \bar{\varphi}^T}{1-\bar{\varphi}^T}, \tag{14}$$

where $\bar{\varphi}_j^T = \sum_{k'=1}^{n} w_{k'}\varphi_{jk'}^T$ and $\bar{\varphi}^T = \sum_{j'=1}^{n}\sum_{k'=1}^{n} w_{j'}w_{k'}\varphi_{j'k'}^T$. (This is the conclusion of panel "Existing Kinship Estimator" in Fig. 1; see Supplementary Information, Section S6 for intermediate calculations that lead to Eq. (14).) Therefore, the bias of $\hat{\varphi}_{jk}^{T,\text{std}}$ varies per j and k. Analogous distortions have been observed for sample covariances of genotypes [59] and were found in concurrent independent work [53]. The limit of $\hat{\varphi}_{jk}^{T,\text{std}}$ in Eq. (14) is constant with respect to T (proof in Supplementary Information, Section S4.2). Similarly, inbreeding coefficient estimates derived from Eq. (11) converge to

$$\hat{f}_j^{T,\text{std}} = 2\hat{\varphi}_{jj}^T - 1 \xrightarrow[m\to\infty]{\text{a.s.}} \frac{f_j^T - 4\bar{\varphi}_j^T + 3\bar{\varphi}^T}{1-\bar{\varphi}^T}. \tag{15}$$

The difference between the bias of $\hat{\varphi}_{jk}^{T,\text{std}}$ for $j \neq k$ in Eq. (14) and $\hat{f}_j^{T,\text{std}}$ in Eq. (15) is visible in the kinship estimates of Fig. 5C (the difference causes a discontinuity between the diagonal and

bioRxiv preprint doi: https://doi.org/10.1101/083923; this version posted June 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

off-diagonal values). The limits of the ratio-of-means versions of two more f_j^T estimators [29] are, if $\hat{p}_i^T$ uses Eq. (10),

$$\begin{aligned}\hat{f}_j^{T,\text{stdII}} &= 1 - \frac{\sum_{i=1}^{m} x_{ij}(2 - x_{ij})}{2\sum_{i=1}^{m} \hat{p}_i^T\left(1 - \hat{p}_i^T\right)} \xrightarrow[m\to\infty]{\text{a.s.}} \frac{f_j^T - \bar{\varphi}^T}{1 - \bar{\varphi}^T}, \\ \hat{f}_j^{T,\text{stdIII}} &= \frac{\sum_{i=1}^{m} x_{ij}^2 - \left(1 + 2\hat{p}_i^T\right) x_{ij} + 2\left(\hat{p}_i^T\right)^2}{2\sum_{i=1}^{m} \hat{p}_i^T\left(1 - \hat{p}_i^T\right)} \xrightarrow[m\to\infty]{\text{a.s.}} \frac{f_j^T + \bar{\varphi}^T - 2\bar{\varphi}_j^T}{1 - \bar{\varphi}^T}.\end{aligned} \tag{16}$$

The estimators in Eqs. (11) and (16) are unbiased when $\hat{p}_i^T$ is replaced by p_i^T [18, 29, 33], and are consistent when $\hat{p}_i^T$ is consistent [48]. Surprisingly, $\hat{p}_i^T$ in Eq. (10) is not consistent (it does not converge almost surely) for arbitrary population structures, which is at the root of the bias in Eqs. (14) to (16). In particular, although $\hat{p}_i^T$ is unbiased, its variance (see Supplementary Information, Section S6),

$$\text{Var}\left(\hat{p}_i^T \middle| T\right) = p_i^T\left(1 - p_i^T\right)\bar{\varphi}^T, \tag{17}$$

may be asymptotically non-zero as $n \to \infty$, since $p_i^T \in (0, 1)$ is fixed and $\lim_{n\to\infty} \bar{\varphi}^T$ may take on any value in [0,1] for arbitrary population structures. Further, $\bar{\varphi}^T \to 0$ as $n \to \infty$ if and only if $\varphi_{jk}^T = 0$ for almost all pairs of individuals (j, k). These observations hold for any weights such that $w_j > 0$, $\sum_{j=1}^{n} w_j = 1$. An important consequence is that the plug-in estimate of $p_i^T\left(1 - p_i^T\right)$ is biased (Supplementary Information, Section S6),

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

which is present in all estimators we have studied.

4.2 Estimation of coancestry coefficients from IAFs

Here we form a coancestry coefficient estimator analogous to Eq. (11) but using IAFs. Assuming the moments in Eqs. (5) and (6), this estimator and its limit are

$$\hat{p}_i^T = \sum_{j=1}^{n} w_j \pi_{ij}, \tag{18}$$

$$\hat{\theta}_{jk}^{T,\text{std}} = \frac{\sum_{i=1}^{m}(\pi_{ij} - \hat{p}_i^T)(\pi_{ik} - \hat{p}_i^T)}{\sum_{i=1}^{m} \hat{p}_i^T\left(1 - \hat{p}_i^T\right)} \xrightarrow[m\to\infty]{\text{a.s.}} \frac{\theta_{jk}^T - \bar{\theta}_j^T - \bar{\theta}_k^T + \bar{\theta}^T}{1 - \bar{\theta}^T}, \tag{19}$$

where $\bar{\theta}_j^T = \sum_{k=1}^{n} w_k \theta_{jk}^T$ and $\bar{\theta}^T = \sum_{j=1}^{n}\sum_{k=1}^{n} w_j w_k \theta_{jk}^T$ are analogous to $\bar{\varphi}_j^T$ and $\bar{\varphi}^T$. Eq. (18) generalizes Eq. (2) for arbitrary weights. Thus, use of IAFs does not ameliorate the estimation problems we

bioRxiv preprint doi: https://doi.org/10.1101/083923; this version posted June 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

have identified for genotypes. Like Eq. (17), $\hat{p}_i^T$ in Eq. (18) is not consistent because $\mathrm{Var}\left(\hat{p}_i^T \middle| T\right) = p_i^T\left(1-p_i^T\right)\bar{\theta}^T$ may not converge to zero for arbitrary population structures, which causes the bias observed in Eq. (19).

4.3 F_{ST} estimator based on the standard kinship estimator

Since the generalized F_{ST} is defined as a mean inbreeding coefficient (Eq. (3) of Part I), 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 (Section 5).

Remarkably, the three f_j^T estimators in Eqs. (15) and (16) 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. (10) match, namely

$$\hat{F}_{\mathrm{ST}}^{\mathrm{std}} = \sum_{j=1}^{n} w_j \hat{f}_j^{T,\mathrm{std}} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} w_j \left(x_{ij} - 2\hat{p}_i^T\right)^2}{2\sum_{i=1}^{m} \hat{p}_i^T\left(1-\hat{p}_i^T\right)} - 1 \xrightarrow[m\to\infty]{\text{a.s.}} \frac{F_{\mathrm{ST}} - \bar{\varphi}^T}{1-\bar{\varphi}^T}, \tag{20}$$

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

$$\hat{F}_{\mathrm{ST}}^{\mathrm{std}} = \sum_{j=1}^{n} w_j \hat{\theta}_{jj}^{T,\mathrm{std}} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} w_j \left(\pi_{ij} - \hat{p}_i^T\right)^2}{\sum_{i=1}^{m} \hat{p}_i^T\left(1-\hat{p}_i^T\right)} \xrightarrow[m\to\infty]{\text{a.s.}} \frac{F_{\mathrm{ST}} - \bar{\theta}^T}{1-\bar{\theta}^T}. \tag{21}$$

The estimators in Eqs. (20) and (21) for individuals and their limits resemble those of classical F_{ST} estimators for populations of the form $\frac{\sigma_p^2}{\bar{p}(1-\bar{p})}$ [6, 7]. $\hat{F}_{\mathrm{ST}}^{\mathrm{std}}$ in Eq. (21) for subpopulations j with uniform weight and one locus is also G_{ST} for two alleles [60]. Compared to $\hat{F}_{\mathrm{ST}}^{\mathrm{indep}}$ in Eq. (4), $\hat{F}_{\mathrm{ST}}^{\mathrm{std}}$ in Eq. (21) 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. (4), $\hat{F}_{\mathrm{ST}}^{\mathrm{std}}$ in Eqs. (20) and (21) are downwardly biased since $0 \leq \bar{\varphi}^T, \bar{\theta}^T$. $\hat{F}_{\mathrm{ST}}^{\mathrm{std}}$ in Eq. (21) may converge arbitrarily close to zero since $\bar{\theta}^T$ can be arbitrarily close to F_{ST} (Supplementary Information, Section S5). Moreover, although $\bar{\varphi}^T \approx \bar{\theta}^T$ for large n (see 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. (20) converges to a negative value.

bioRxiv preprint doi: https://doi.org/10.1101/083923; this version posted June 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

4.4 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. (21) that rely on having minimal additional information needed to correct its bias. If $\bar{\theta}^T$ is known, the bias in Eq. (21) can be reversed, yielding the consistent estimator

$$\hat{F}'_{\mathrm{ST}} = \hat{F}_{\mathrm{ST}}^{\mathrm{std}}(1-\bar{\theta}^T) + \bar{\theta}^T \xrightarrow[m\to\infty]{\text{a.s.}} F_{\mathrm{ST}}. \tag{22}$$

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

$$s^T = \frac{\bar{\theta}^T}{F_{\mathrm{ST}}} = \frac{\sum_{j=1}^{n}\sum_{k=1}^{n} w_j w_k \theta_{jk}^T}{\sum_{j=1}^{n} w_j \theta_{jj}^T}, \tag{23}$$

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 $0 < s^T \leq 1$ (Supplementary Information, Section S5). The limit in Eq. (21) in terms of s^T is

$$\hat{F}_{\mathrm{ST}}^{\mathrm{std}} \xrightarrow[m\to\infty]{\text{a.s.}} F_{\mathrm{ST}} \frac{1-s^T}{1-s^T F_{\mathrm{ST}}}. \tag{24}$$

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

$$\hat{\sigma}_i^2 = \frac{1}{1-s^T}\sum_{j=1}^{n} w_j(\pi_{ij} - \hat{p}_i^T)^2, \tag{25}$$

$$\hat{F}''_{\mathrm{ST}} = \frac{\hat{F}_{\mathrm{ST}}^{\mathrm{std}}}{1-s^T(1-\hat{F}_{\mathrm{ST}}^{\mathrm{std}})} = \frac{\sum_{i=1}^{m}\hat{\sigma}_i^2}{\sum_{i=1}^{m}\hat{p}_i^T\left(1-\hat{p}_i^T\right) + s^T\hat{\sigma}_i^2} \xrightarrow[m\to\infty]{\text{a.s.}} F_{\mathrm{ST}}. \tag{26}$$

Note that $\hat{\sigma}_i^2$ and $\hat{F}_{\mathrm{ST}}^{\mathrm{indep}}$ from Eqs. (3) and (4) are the special case of Eqs. (25) and (26) for uniform weights and $s^T = \frac{1}{n}$; hence, $\hat{F}''_{\mathrm{ST}}$ generalizes $\hat{F}_{\mathrm{ST}}^{\mathrm{indep}}$.

Lastly, using either Eq. (21) or Eq. (24), the relative error of $\hat{F}_{\mathrm{ST}}^{\mathrm{std}}$ converges to

$$1 - \frac{\hat{F}_{\mathrm{ST}}^{\mathrm{std}}}{F_{\mathrm{ST}}} \xrightarrow[m\to\infty]{\text{a.s.}} \frac{\bar{\theta}^T\left(1-F_{\mathrm{ST}}\right)}{F_{\mathrm{ST}}\left(1-\bar{\theta}^T\right)} = s^T\frac{1-F_{\mathrm{ST}}}{1-s^T F_{\mathrm{ST}}}, \tag{27}$$

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.

bioRxiv preprint doi: https://doi.org/10.1101/083923; this version posted June 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

5 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 Section 6 and applied to human data in [47].

5.1 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 $\varphi_{jk}^T = 0$, which sets T to the MRCA population of all the individuals in the data. The new kinship estimator immediately results in new inbreeding ($\hat{f}_j^{T,\mathrm{new}}$) and F_{ST} ($\hat{F}_{\mathrm{ST}}^{\mathrm{new}}$) estimators. This general approach leaves the implementation of $\hat{A}_{\mathrm{min}}$ open; the simple implementation applied in this work is described in Section 5.2, but our method can be readily improved by substituting in a better $\hat{A}_{\mathrm{min}}$ in the future.

Applying the method of moments to Eqs. (12) and (13), we derive the following statistic (see Supplementary Information, Section S7), whose expectation is proportional to $\varphi_{jk}^T - 1$:

$$\begin{aligned} A_{jk} &= \frac{1}{m}\sum_{i=1}^{m}(x_{ij}-1)(x_{ik}-1) - 1, \\ \mathrm{E}\left[A_{jk}|T\right] &= \left(\varphi_{jk}^T - 1\right) v_m^T, \quad \text{where} \\ v_m^T &= \frac{4}{m}\sum_{i=1}^{m} p_i^T\left(1-p_i^T\right). \end{aligned} \tag{28}$$

Compared to the standard kinship estimator in Eq. (14), which has a complex asymptotic bias determined by n parameters ($\bar{\varphi}_j^T$ for each $j \in \{1, ..., n\}$), the A_{jk} statistics estimate kinship with a bias controlled by the sole unknown parameter v_m^T shared by all pairs of individuals. The key to estimating v_m^T is to notice that if $\varphi_{jk}^T = 0$ then $\mathrm{E}[A_{jk}|T] = -v_m^T$. Thus, assuming $\min_{j,k}\varphi_{jk}^T = 0$, 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 Section 5.2.

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

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

bioRxiv preprint doi: https://doi.org/10.1101/083923; this version posted June 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

along with the assumption that $v_m^T \xrightarrow[m\to\infty]{} v^T$ for some $v^T \neq 0$. Our new kinship estimator follows directly from replacing v_m^T with $\hat{A}_{\min}$ and solving for φ_{jk}^T in Eq. (28), which results in a consistent kinship estimator (given the convergence proof of Section 2):

$$\hat{\varphi}_{jk}^{T,\text{new}} = 1 - \frac{A_{jk}}{\hat{A}_{\min}} \xrightarrow[m\to\infty]{\text{a.s.}} \varphi_{jk}^T. \tag{29}$$

The resulting new inbreeding coefficient estimator is

$$\hat{f}_j^{T,\text{new}} = 2\hat{\varphi}_{jj}^{T,\text{new}} - 1 \xrightarrow[m\to\infty]{\text{a.s.}} f_j^T, \tag{30}$$

and the new F_{ST} estimator is

$$\hat{F}_{\text{ST}}^{\text{new}} = \sum_{j=1}^{n} w_j \hat{f}_j^{T,\text{new}} \xrightarrow[m\to\infty]{\text{a.s.}} F_{\text{ST}}. \tag{31}$$

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

The A_{jk} statistic defined above is closely related to the mean "identity by state" estimator [18] and to another recently-described kinship estimator [53, 61]. However, only our $\hat{\varphi}_{jk}^{T,\text{new}}$ in Eq. (29)—scaling A_{jk} using $\hat{A}_{\min}$—results in consistent kinship estimation under arbitrary population structures.

5.2 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 ($\hat{\varphi}_{jk}^{T,\text{new}}$ in Eq. (29)). This $\hat{A}_{\min}$ relies on an appropriate partition of the n individuals into K subpopulations (denoted S_u for $u \in \{1, ..., K\}$), 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 3, 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:

$$\hat{A}_{\min} = \min_{u\neq v} \frac{1}{|S_u||S_v|} \sum_{j\in S_u} \sum_{k\in S_v} A_{jk}. \tag{32}$$

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 [47]. Future work could focus on a more general $\hat{A}_{\min}$ that circumvents the need for subpopulations of our proof-of-principle estimator.

bioRxiv preprint doi: https://doi.org/10.1101/083923; this version posted June 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

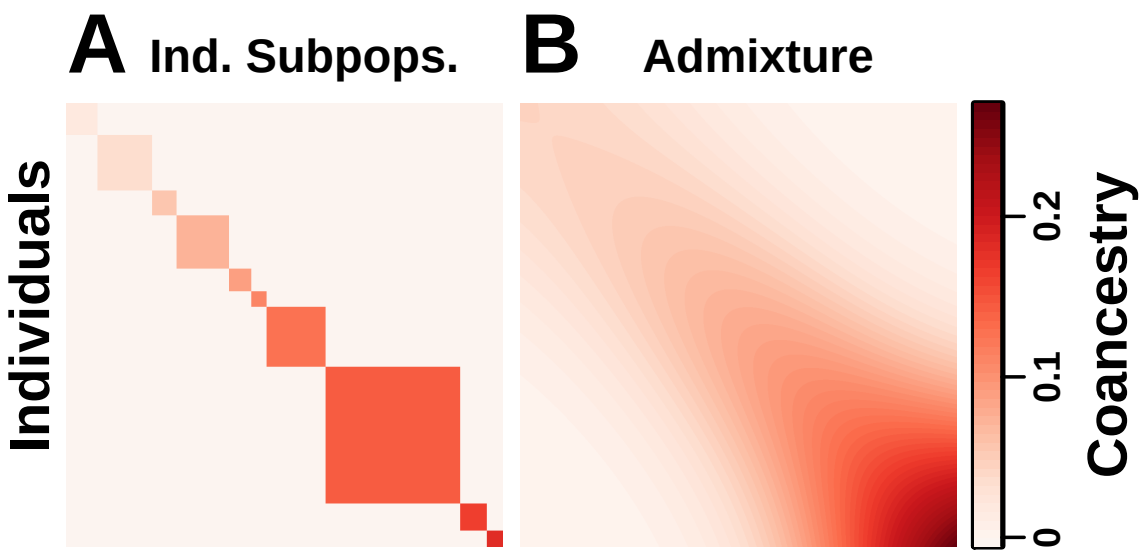

Figure 2: **Coancestry matrices of simulations.** Both panels have $n = 1000$ individuals along both axes, $K = 10$ subpopulations (final or intermediate), and $F_{\mathrm{ST}} = 0.1$. Color corresponds to θ^T_{jk} between individuals j and k (equal to φ^T_{jk} off-diagonal, f^T_j along the diagonal). A) The independent subpopulations model has $\theta^T_{jk} = 0$ between subpopulations, and varying θ^T_{jj} 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 coancesty matrix with no independent subpopulations (no $\theta^T_{jk} = 0$ between blocks). Individuals are ordered along each axis by geographical position.

6 Simulations evaluating F_{ST} and kinship estimators

6.1 Overview of simulations

We simulate genotypes from two models to illustrate our results when the true population structure parameters are known. 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 [47]: compare the simulated kinship matrix (Fig. 2B) and admixture proportions (Fig. 3C) to our estimates on the real data [47]. Both simulations have $n = 1000$ individuals, $m = 300,000$ loci, and $K = 10$ subpopulations or intermediate subpopulations. These simulations have $F_{\mathrm{ST}} = 0.1$, 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 [20]).

bioRxiv preprint doi: https://doi.org/10.1101/083923; this version posted June 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

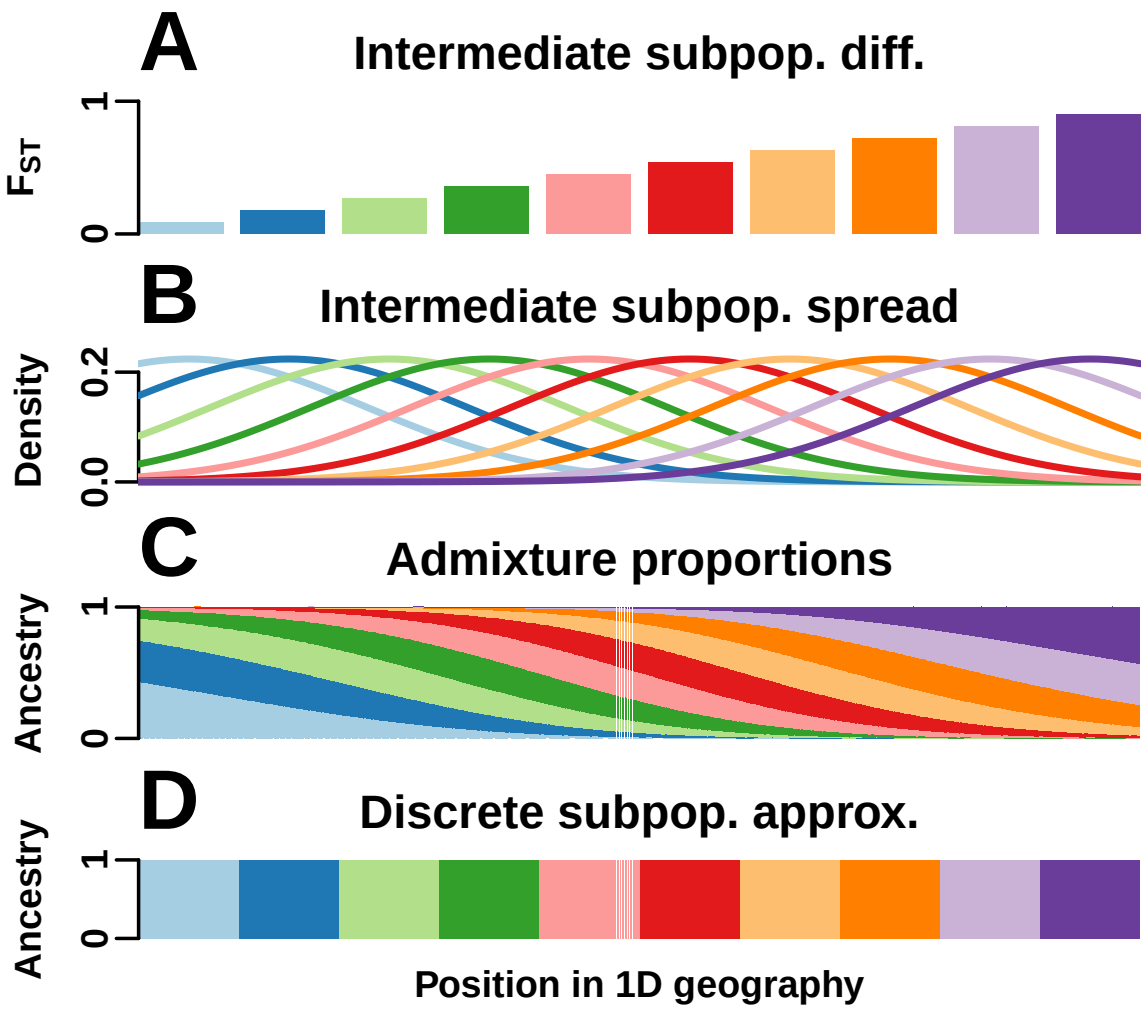

Figure 3: **1D admixture scenario.** We model a 1D geography population that departs strongly from the independent subpopulations model. A) $K = 10$ intermediate subpopulations, evenly spaced on a line, evolved independently in the past with F_{ST} increasing with distance, which models a sequence of 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 ($K = 10$).

bioRxiv preprint doi: https://doi.org/10.1101/083923; this version posted June 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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 in [0.01, 0.5]. Allele frequencies $p^{S_u}_i$ for S_u and locus i are drawn independently from the Balding-Nichols (BN) distribution [5] 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 (Supplementary Information, Section S8).

The admixture simulation corresponds to a "BN-PSD" model [8, 24, 31, 48, 62], which we analyzed in Section 6 of Part I and has a demographic model illustrated in Fig. 4 of Part I. 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 $p^{S_u}_i$ with the admixture proportions q_{ju} of j and u as weights (which satisfy $\sum_{u=1}^{K} q_{ju} = 1$, as in the Pritchard-Stephens-Donnelly [PSD] admixture model [50–52]). 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 to these Normal densities, resulting in a genetic structure governed by geography (Fig. 3C, Fig. 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 $s^T = \frac{\bar{\theta}^T}{F_{\mathrm{ST}}} = 0.5$, which by Eq. (27) results in a large downward bias for $\hat{F}^{\mathrm{std}}_{\mathrm{ST}}$ (in contrast, the independent subpopulations simulation has $s^T = 0.1$). The true θ^T_{jk} and F_{ST} parameters of this simulation are given by the $f^T_{S_u}$ values of the intermediate subpopulations and the admixture coefficients q_{ju} of the individuals via Eq. (17) of Part I. See Supplementary Information, Section S8 for additional details regarding these simulations.

6.2 Evaluation of F_{ST} estimators

Our admixture simulation illustrates the large biases that can arise if F_{ST} estimators for independent subpopulations (WC, HudsonK and BayeScan) are misapplied to arbitrary population structures to estimate the generalized F_{ST}, and demonstrate the higher accuracy of our new F_{ST} estimator $\hat{F}^{\mathrm{new}}_{\mathrm{ST}}$ given by the combination of Eqs. (31) and (32). BayeScan was used to estimate the per-subpopulation F_{ST} across loci assuming no selection, and the global F_{ST} was given by the mean F_{ST} across subpopulations.

First, we test these estimators in our independent subpopulations simulation. Both the HudsonK (Supplementary Information, Section S2.3) and BayeScan F_{ST} estimators are consistent in this

bioRxiv preprint doi: https://doi.org/10.1101/083923; this version posted June 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

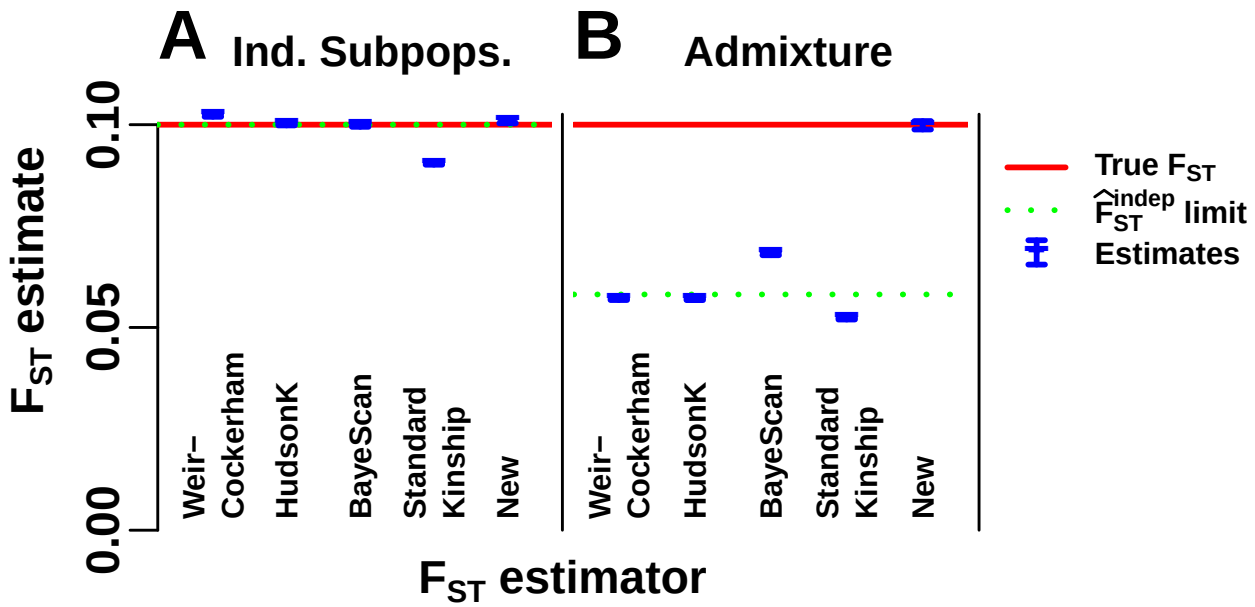

Figure 4: **Evaluation of F_{ST} estimators.** The WC, HudsonK, BayeScan, $\hat{F}_{ST}^{std}$ in Eq. (20) derived from the standard kinship estimator, and our new F_{ST} estimator in Eqs. (29) and (32), are evaluated on simulated genotypes from our two models (Fig. 2). A) The independent subpopulations model assumed by the HudsonK and BayeScan F_{ST} estimators. All but standard kinship ($\hat{F}_{ST}^{std}$) have zero or small biases. 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 have large biases that result from treating kinship as zero between every subpopulations imposed by geographic clustering. The estimator limit in Eq. (7) (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 (Supplementary Information, Section S9).

bioRxiv preprint doi: https://doi.org/10.1101/083923; this version posted June 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

simulation, since their assumptions are satisfied (Fig. 4A). The WC estimator assumes that $f^T_{S_u} = F_{\mathrm{ST}}$ for all subpopulations S_u, which does not hold; nevertheless, WC has only a small bias (Fig. 4A). For comparison, we show the standard kinship-based $\hat{F}^{\mathrm{std}}_{\mathrm{ST}}$ in Eq. (20) (weights from Supplementary Information, Section S8), which does not have corrections that would make it consistent under the independent subpopulations model. Since the number of subpopulations K is large, $\hat{F}^{\mathrm{std}}_{\mathrm{ST}}$ 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. (32)) 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 WC, HudsonK, and BayeScan are smaller than the true F_{ST} by nearly half, as predicted by the limit of $\hat{F}^{\mathrm{indep}}_{\mathrm{ST}}$ in Eq. (7) (Fig. 4B). By construction, $\hat{F}^{\mathrm{std}}_{\mathrm{ST}}$ also has a large relative bias of about $s^T = 50\%$; remarkably, the WC, HudsonK, and BayeScan estimators suffer from comparable biases. Thus, the corrections for independent subpopulations present in the WC and HudsonK estimators, or the Bayesian likelihood modeling of BayeScan, are insufficient for accurate estimation of the generalized F_{ST} in this admixture scenario. Only our new F_{ST} estimator achieves practically unbiased estimates in the admixture simulation (Fig. 4B).

6.3 Evaluation of kinship estimators

Our admixture simulation illustrates the distortions of the standard kinship estimator $\hat{\varphi}^{T,\mathrm{std}}_{jk}$ in Eq. (11) and demonstrates the improved accuracy of our new kinship estimator $\hat{\varphi}^{T,\mathrm{new}}_{jk}$ given by the combination of Eqs. (29) and (32). The limit of the standard estimator $\hat{\varphi}^{T,\mathrm{std}}_{jk}$ in Eq. (11) has a uniform bias if $\bar{\varphi}^T_j = \bar{\varphi}^T$ for all individuals j. For that reason, our admixture simulation has varying differentiation $f^T_{S_u}$ per intermediate subpopulation S_u (Fig. 3A), which causes large differences in $\bar{\varphi}^T_j$ per individual j and therefore large distortions in $\hat{\varphi}^{T,\mathrm{std}}_{jk}$.

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^T_{jk}. 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^T_{jk} relative to the mean φ^T_{jk}. Additionally, estimates from $\hat{\varphi}^{T,\mathrm{std}}_{jk}$ are very distorted, with an abundance of $\hat{\varphi}^{T,\mathrm{std}}_{jk} < \varphi^T_{jk}$ cases—some of which are negative estimates (blue in Fig. 5C)—but remarkably also cases with $\hat{\varphi}^{T,\mathrm{std}}_{jk} > \varphi^T_{jk}$ (top left corner of Fig. 5C).

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. (14) (Fig. 5D). The ratio-of-means estimate $\hat{\varphi}^{T,\mathrm{std}}_{jk}$ (Fig. 5C) has an RMSE of 2.14% from its limit relative to the mean φ^T_{jk}. In contrast, the mean-of-ratios estimates that are prevalent in the literature have a greater RMSE of 10.77%

bioRxiv preprint doi: https://doi.org/10.1101/083923; this version posted June 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

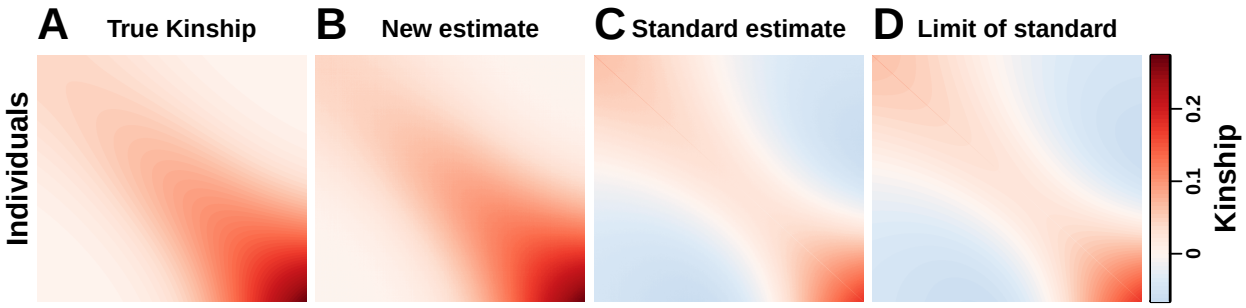

Figure 5: **Evaluation of kinship estimators.** Bias for the standard kinship coefficient estimator is illustrated in our admixture simulation and contrasted to the nearly unbiased estimates of our new estimator. Plots show $n = 1000$ individuals along both axes, and color corresponds to φ^T_{jk} between individuals $j \neq k$ and to f^T_j along the diagonal (f^T_j is in the same scale as φ^T_{jk} for $j \neq k$; plotting φ^T_{jj}, 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. (29) and (32) from simulated genotypes recovers the true kinship matrix with high accuracy. C) Standard kinship estimates $\hat{\varphi}^{T,\mathrm{std}}_{jk}$ given by Eq. (11) from simulated genotypes are downwardly biased on average and distorted by pair-specific amounts. D) Theoretical limit of $\hat{\varphi}^{T,\mathrm{std}}_{jk}$ in Eq. (14) as the number of independent loci goes to infinity demonstrates the accuracy of our bias predictions under the kinship model.

bioRxiv preprint doi: https://doi.org/10.1101/083923; this version posted June 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

from the same limit in Eq. (14). Thus, as expected from our theoretical results in Section 2, 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. (19), with reduced RMSEs from its limit of 0.32% and 8.82% for the ratio-of-means and mean-of-ratios estimates, respectively.

6.4 Evaluation of oracle adjusted F_{ST} estimators

Here we verify additional calculations for the bias of the standard kinship-based estimator $\hat{F}_{\mathrm{ST}}^{\mathrm{std}}$ and the unbiased adjusted "oracle" F_{ST} estimators that require the true mean kinship $\bar{\varphi}^T$ or the bias coefficient s^T to be known. Note that $\hat{F}_{\mathrm{ST}}^{\mathrm{new}}$ in Eq. (31) is related but not identical to these oracle estimators. We tested both IAF (Fig. 6A) and genotype (Fig. 6B) versions of these estimators. The unadjusted $\hat{F}_{\mathrm{ST}}^{\mathrm{std}}$ in Eq. (21) is severely biased (blue in Fig. 6) by construction, and matches the calculated limit for IAFs and genotypes (green lines in Fig. 6, which are close because $\bar{\varphi}^T \approx \bar{\theta}^T$). In contrast, the two consistent adjusted estimators $\hat{F}'_{\mathrm{ST}}$ and $\hat{F}''_{\mathrm{ST}}$ in Eqs. (22) and (26) estimate F_{ST} quite well (blue predictions overlap the true F_{ST} red line in Fig. 6). However, $\hat{F}'_{\mathrm{ST}}$ and $\hat{F}''_{\mathrm{ST}}$ are oracle methods, since they require parameters ($\bar{\varphi}^T$, $\bar{\theta}^T$, s^T) that are not known in practice.

Prediction intervals were computed from estimates over 39 independently-simulated IAF and genotype matrices (Supplementary Information, Section S9). 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; however, IAFs are not known in practice, and use of estimated IAFs might increase noise. Genetic linkage, not present in our simulation, will also increase noise in real data.

7 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 bias of these estimators under two models of arbitrary population structure (Fig. 1). We calculated the bias of these F_{ST} estimators when the independent subpopulations model assumption is violated. This bias is present even when individual-specific allele frequencies are known without error. 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 [47].

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

bioRxiv preprint doi: https://doi.org/10.1101/083923; this version posted June 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

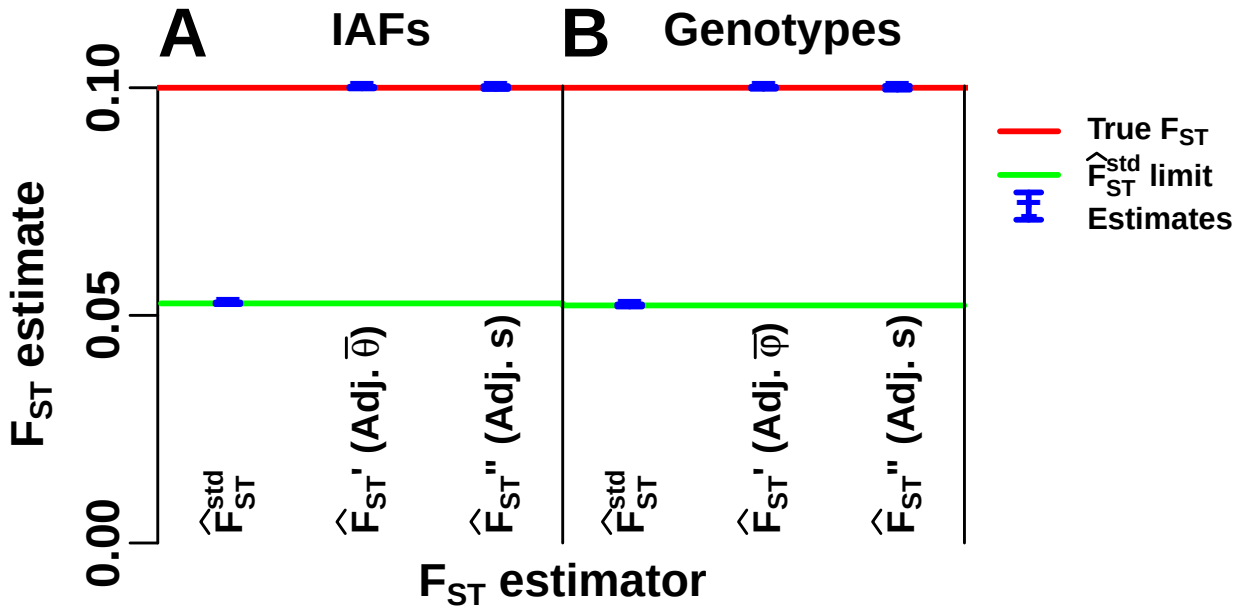

Figure 6: **Evaluation of standard and adjusted F_{ST} estimators.** The convergence values we calculated for the standard kinship plug-in and adjusted F_{ST} estimators are validated using our admixture simulation. All adjusted estimators are unbiased but are "oracle" methods, since the mean kinship ($\bar{\varphi}^T$), mean coancestry ($\bar{\theta}^T$), or bias coefficient ($s^T = \frac{\bar{\theta}^T}{F_{\mathrm{ST}}}$ for IAFs, replaced by $\frac{\bar{\varphi}^T}{F_{\mathrm{ST}}}$ for genotypes) are usually unknown. A) Estimation from individual-specific allele frequencies (IAFs): $\hat{F}^{\mathrm{std}}_{\mathrm{ST}}$ is the standard coancestry plug-in estimator in Eq. (21); $\hat{F}'_{\mathrm{ST}}$ "Adj. $\bar{\theta}^T$" is in Eq. (22); $\hat{F}''_{\mathrm{ST}}$ "Adj. s" is in Eq. (26). B) For genotypes, $\hat{F}^{\mathrm{std}}_{\mathrm{ST}}$ is given in Eq. (20), and the adjusted estimators use $\bar{\varphi}^T$ rather than $\bar{\theta}^T$. Lines: true F_{ST} (red line), limits of biased estimators $\hat{F}^{\mathrm{std}}_{\mathrm{ST}}$ (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 (Supplementary Information, Section S9).

bioRxiv preprint doi: https://doi.org/10.1101/083923; this version posted June 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

[5, 9, 55, 61, 63–65]. Our findings that existing genome-wide F_{ST} estimates are downwardly biased matters in these settings. However, our findings may not have direct implications for per-locus F_{ST} estimate approaches where only the relative ranking matters, such as for the identification of loci under selection [10, 12, 66–71], assuming that the bias of the genome-wide estimator carries over uniformly to all per-locus estimates. Note that our convergence calculations in Section 2 require large numbers of loci so they do not apply to single locus estimates. Moreover, various methods for per-locus F_{ST} estimation for multiple alleles suffer from a strong dependence to the maximum allele frequency and heterozygosity [68–70, 72–75] that suggests that a more complicated bias is present in these per-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 [19], HudsonK (which generalizes the Hudson pairwise F_{ST} estimator [20] to K independent populations), and BayeScan [12] F_{ST} estimates in our admixture simulation are biased by nearly a factor of two (Fig. 4B), and differ from our new F_{ST} estimates in humans by nearly a factor of three [47]. These estimators were derived assuming 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 [47, 76–80] (also see Section 2 in Part I).

The standard kinship coefficient estimator we investigated is often used to control for population structure in GWAS and to estimate genome-wide heritability [18, 24, 27–32]. While this estimator was known to be biased [18, 32], no closed form limit had been calculated until now (concurrently calculated by [53]). We found that kinship estimates are biased downwards on average, but bias also varies for each pair of individuals (Fig. 1, Fig. 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 (Supplementary Information, Section S1.1). Our result justifies approximating the expectation of a ratio-of-means estimator with the ratio of expectations [6, 19, 20]. 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.

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

bioRxiv preprint doi: https://doi.org/10.1101/083923; this version posted June 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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.

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 WC 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`.

Acknowledgments

This research was supported in part by NIH grant R01 HG006448.

References

[1] Alejandro Ochoa and John D. Storey. "F_{ST} and kinship for arbitrary population structures I: Generalized definitions". *bioRxiv* (10.1101/083915) (2019). `https://doi.org/10.1101/083915`. First published 2016-10-27.

[2] Alejandro Ochoa and John D. Storey. "F_{ST} and kinship for arbitrary population structures II: Method of moments estimators". *bioRxiv* (10.1101/083923) (2019). `https://doi.org/10.1101/083923`. First published 2016-10-27.

[3] Gustave Malécot. *Mathématiques de l'hérédité*. Masson et Cie, 1948.

[4] S. Wright. "The genetical structure of populations". *Ann Eugen* 15(4) (1951), pp. 323–354.

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

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

bioRxiv preprint doi: https://doi.org/10.1101/083923; this version posted June 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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

[8] Daniel Falush, Matthew Stephens, and Jonathan K. Pritchard. "Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies". *Genetics* 164(4) (2003), pp. 1567–1587.

[9] David J. Balding. "Likelihood-based inference for genetic correlation coefficients". *Theoretical Population Biology.* Uses of DNA and genetic markers for forensics and population studies 63(3) (2003), pp. 221–230.

[10] Mark A. Beaumont and David J. Balding. "Identifying adaptive genetic divergence among populations from genome scans". *Molecular Ecology* 13(4) (2004), pp. 969–980.

[11] Matthieu Foll and Oscar Gaggiotti. "Identifying the Environmental Factors That Determine the Genetic Structure of Populations". *Genetics* 174(2) (2006), pp. 875–891.

[12] Matthieu Foll and Oscar Gaggiotti. "A Genome-Scan Method to Identify Selected Loci Appropriate for Both Dominant and Codominant Markers: A Bayesian Perspective". *Genetics* 180(2) (2008), pp. 977–993.

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

[14] E. A. Thompson. "The estimation of pairwise relationships". *Ann. Hum. Genet.* 39(2) (1975), pp. 173–188.

[15] Brook G. Milligan. "Maximum-likelihood estimation of relatedness". *Genetics* 163(3) (2003), pp. 1153–1167.

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

[17] Miklós Csűrös. "Non-identifiability of identity coefficients at biallelic loci". *Theor Popul Biol* 92 (2014), pp. 22–29.

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

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

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

[21] C. Xie, D. D. Gessler, and S. Xu. "Combining different line crosses for mapping quantitative trait loci using the identical by descent-based variance component method". *Genetics* 149(2) (1998), pp. 1139–1146.

bioRxiv preprint doi: https://doi.org/10.1101/083923; this version posted June 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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

[23] Yurii S. Aulchenko, Dirk-Jan de Koning, and Chris Haley. "Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis". *Genetics* 177(1) (2007), pp. 577–585.

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

[25] Hyun Min Kang et al. "Efficient control of population structure in model organism association mapping". *Genetics* 178(3) (2008), pp. 1709–1723.

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

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

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

[29] Jian Yang et al. "GCTA: a tool for genome-wide complex trait analysis". *Am. J. Hum. Genet.* 88(1) (2011), pp. 76–82.

[30] Cyril S. Rakovski and Daniel O. Stram. "A kinship-based modification of the armitage trend test to address hidden population structure and small differential genotyping errors". *PLoS ONE* 4(6) (2009), e5825.

[31] Timothy Thornton and Mary Sara McPeek. "ROADTRIPS: case-control association testing with partially or completely unknown population and pedigree structure". *Am. J. Hum. Genet.* 86(2) (2010), pp. 172–184.

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

[33] Bowen Wang, Serge Sverdlov, and Elizabeth Thompson. "Efficient Estimation of Realized Kinship from SNP Genotypes". *Genetics* (2017), genetics.116.197004.

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

[35] Sohini Ramachandran et al. "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* 102(44) (2005), pp. 15942–15947.

[36] The 1000 Genomes Project Consortium. "A map of human genome variation from population-scale sequencing". *Nature* 467(7319) (2010), pp. 1061–1073.

bioRxiv preprint doi: https://doi.org/10.1101/083923; this version posted June 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

[37] Iosif Lazaridis et al. "Ancient human genomes suggest three ancestral populations for present-day Europeans". *Nature* 513(7518) (2014), pp. 409–413.

[38] Iosif Lazaridis et al. "Genomic insights into the origin of farming in the ancient Near East". *Nature* 536(7617) (2016), pp. 419–424.

[39] Pontus Skoglund et al. "Genomic insights into the peopling of the Southwest Pacific". *Nature* 538(7626) (2016), pp. 510–513.

[40] Sarah A. Tishkoff et al. "The Genetic Structure and History of Africans and African Americans". *Science* 324(5930) (2009), pp. 1035–1044.

[41] Andrés Moreno-Estrada et al. "Reconstructing the Population Genetic History of the Caribbean". *PLOS Genetics* 9(11) (2013), e1003925.

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

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

[44] Soheil Baharian et al. "The Great Migration and African-American Genomic Diversity". *PLoS Genet.* 12(5) (2016), e1006059.

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

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

[47] Alejandro Ochoa and John D. Storey. "New kinship and F_{ST} estimates reveal higher levels of differentiation in the global human population". *bioRxiv* (10.1101/653279) (2019). `https://doi.org/10.1101/653279`.

[48] Timothy Thornton et al. "Estimating kinship in admixed populations". *Am. J. Hum. Genet.* 91(1) (2012), pp. 122–138.

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

[50] J. K. Pritchard, M. Stephens, and P. Donnelly. "Inference of population structure using multilocus genotype data". *Genetics* 155(2) (2000), pp. 945–959.

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

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

bioRxiv preprint doi: https://doi.org/10.1101/083923; this version posted June 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

[53] Bruce S. Weir and Jérôme Goudet. "A Unified Characterization of Population Structure and Relatedness". *Genetics* (2017), genetics.116.198424.

[54] William Gemmell Cochran. *Sampling techniques.* 3rd ed. Wiley, 1977.

[55] John Buckleton et al. "Population-specific FST values for forensic STR markers: A worldwide survey". *Forensic Science International: Genetics* 23 (2016), pp. 91–100.

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

[57] Catherine Bourgain et al. "Novel case-control test in a founder population identifies P-selectin as an atopy-susceptibility locus". *Am. J. Hum. Genet.* 73(3) (2003), pp. 612–626.

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

[59] Joseph K. Pickrell and Jonathan K. Pritchard. "Inference of population splits and mixtures from genome-wide allele frequency data". *PLoS Genet.* 8(11) (2012), e1002967.

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

[61] Bruce Weir and Xiuwen Zheng. "SNPs and SNVs in forensic science". *Forensic Science International: Genetics Supplement Series* 5 (Dec 2015), e267–e268.

[62] Anil Raj, Matthew Stephens, and Jonathan K. Pritchard. "fastSTRUCTURE: variational inference of population structure in large SNP data sets". *Genetics* 197(2) (2014), pp. 573–589.

[63] Mari Nelis et al. "Genetic Structure of Europeans: A View from the North–East". *PLOS ONE* 4(5) (2009), e5472.

[64] Nuno M. Silva et al. "Human Neutral Genetic Variation and Forensic STR Data". *PLOS ONE* 7(11) (2012), e49666.

[65] Christopher D. Steele, Denise Syndercombe Court, and David J. Balding. "Worldwide FST Estimates Relative to Five Continental-Scale Populations". *Annals of Human Genetics* 78(6) (2014), pp. 468–477.

[66] L. L. Cavalli-Sforza. "Population Structure and Human Evolution". *Proceedings of the Royal Society of London. Series B, Biological Sciences* 164(995) (1966), pp. 362–379.

[67] R. C. Lewontin and Jesse Krakauer. "Distribution of Gene Frequency as a Test of the Theory of the Selective Neutrality of Polymorphisms". *Genetics* 74(1) (1973), pp. 175–195.

[68] Mark A. Beaumont and Richard A. Nichols. "Evaluating Loci for Use in the Genetic Analysis of Population Structure". *Proceedings of the Royal Society of London B: Biological Sciences* 263(1377) (1996), pp. 1619–1626.

bioRxiv preprint doi: https://doi.org/10.1101/083923; this version posted June 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

[69] Renaud Vitalis, Kevin Dawson, and Pierre Boursot. "Interpretation of Variation Across Marker Loci as Evidence of Selection". *Genetics* 158(4) (2001), pp. 1811–1823.

[70] Joshua M. Akey et al. "Interrogating a High-Density SNP Map for Signatures of Natural Selection". *Genome Res.* 12(12) (2002), pp. 1805–1814.

[71] Adam H. Porter. "A test for deviation from island-model population structure". *Molecular Ecology* 12(4) (2003), pp. 903–915.

[72] A. M. Bowcock et al. "Drift, admixture, and selection in human evolution: a study with DNA polymorphisms". *PNAS* 88(3) (1991), pp. 839–843.

[73] Philip W. Hedrick. "A Standardized Genetic Differentiation Measure". *Evolution* 59(8) (2005), pp. 1633–1638.

[74] Mattias Jakobsson, Michael D. Edge, and Noah A. Rosenberg. "The Relationship Between FST and the Frequency of the Most Frequent Allele". *Genetics* 193(2) (2013), pp. 515–528.

[75] Michael D. Edge and Noah A. Rosenberg. "Upper bounds on in terms of the frequency of the most frequent allele and total homozygosity: The case of a specified number of alleles". *Theoretical Population Biology* 97 (2014), pp. 20–34.

[76] R. C. Lewontin. "The Apportionment of Human Diversity". *Evolutionary Biology*. Ed. by Theodosius Dobzhansky, Max K. Hecht, and William C. Steere. Springer US, 1995, pp. 381–398.

[77] Guido Barbujani et al. "An apportionment of human DNA diversity". *PNAS* 94(9) (1997), pp. 4516–4519.

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

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

[80] Nick Patterson et al. "Ancient admixture in human history". *Genetics* 192(3) (2012), pp. 1065–1093.

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

[82] Patrick Billingsley. *Convergence of Probability Measures*. John Wiley & Sons, 2013. 247 pp.

[83] William Feller. *An introduction to probability theory and its applications*. 3rd ed. Vol. 1. John Wiley & Sons London-New York-Sydney-Toronto, 1968. 528 pp.

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

bioRxiv preprint doi: https://doi.org/10.1101/083923; this version posted June 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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

bioRxiv preprint doi: https://doi.org/10.1101/083923; this version posted June 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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

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

[1]Duke Center for Statistical Genetics and Genomics, and [2]Department of Biostatistics and Bioinformatics, Duke University, Durham, NC 27705, USA
[3]Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544, USA
∗ Corresponding author: `jstorey@princeton.edu`

S1 Accuracy of ratio estimators

S1.1 Almost sure convergence of ratio-of-means estimators with independent and uniformly-bounded terms

Here we prove that $\frac{\hat{A}_m}{\hat{B}_m} \xrightarrow[m\to\infty]{\text{a.s.}} \frac{A}{B}$, where $\hat{A}_m = \frac{1}{m}\sum_{i=1}^{m} a_i$ and $\hat{B}_m = \frac{1}{m}\sum_{i=1}^{m} b_i$ give the ratio-of-means estimator described in the main text. It suffices to prove $\hat{A}_m \xrightarrow[m\to\infty]{\text{a.s.}} Ac$ and $\hat{B}_m \xrightarrow[m\to\infty]{\text{a.s.}} Bc \neq 0$, from which the result follows using the continuous mapping theorem [81, 82]. The proof for $\hat{A}_m$ follows, which applies analogously to $\hat{B}_m$. Our a_i are independent but not identically distributed, since they depend on p_i^T that varies per locus, so the standard law of large numbers does not apply to $\hat{A}_m$. We show almost sure convergence using Kolmogorov's criterion for the Strong Law of Large Numbers [83], which is satisfied for bounded $\text{Var}(a_i)$. Since $|a_i| \leq C < \infty$ for all i and some C (see main text), then $\text{E}[a_i^2] \leq C^2$, so $\text{Var}(a_i) \leq C^2$. Therefore, $\hat{A}_m \xrightarrow[m\to\infty]{\text{a.s.}} \lim_{m\to\infty} \text{E}\left[\hat{A}_m\right] = Ac$, as desired.

S1.2 Order of error of expectations

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

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

which follows from $\text{Cov}(X,Y) = \text{E}[XY] - \text{E}[X]\,\text{E}[Y]$ and expanding the covariance [84]. Previous work on ratio estimators [54, 84] assumes IID a_i and b_i, which does not hold for SNP loci. Assuming independent loci ($\text{Cov}(a_i, b_j) = 0$ for $i \neq j$) and large m so $\hat{B}_m \approx Bc$ is practically independent of

bioRxiv preprint doi: https://doi.org/10.1101/083923; this version posted June 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

any given a_i and b_j, then

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

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

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

for some large enough m and C. Hence $\epsilon_m = O\left(\frac{1}{m}\right)$ as is for standard ratio estimators [54].

S2 Previous F_{ST} estimators for the independent subpopulations model

Here we summarize the previous WC and Hudson F_{ST} estimators for independent subpopulations and introduce the generalized HudsonK estimator for more than two subpopulations. 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.

S2.1 The Weir-Cockerham F_{ST} estimator

The Weir-Cockerham (WC) F_{ST} estimator [19] 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 ($\theta^T_{jk} = 0$ for $j \neq k$) with equal differentiation ($\theta^T_{jj} = \theta^T$) is

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

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

$$\hat{F}_{\mathrm{ST}}^{\mathrm{WC}} = \frac{\sum_{i=1}^{m}\hat{\sigma}_i^2 - \frac{1}{\bar{n}-1}\left(\hat{p}_i^T\left(1-\hat{p}_i^T\right) - \frac{n-1}{n}\hat{\sigma}_i^2 - \frac{1}{4}\bar{h}_i\right)}{\sum_{i=1}^{m}\hat{p}_i^T\left(1-\hat{p}_i^T\right)\left(1-\frac{\bar{n}C^2}{n(\bar{n}-1)}\right) + \frac{1}{n}\hat{\sigma}_i^2\left(1+\frac{(n-1)\bar{n}C^2}{n(\bar{n}-1)}\right) + \frac{\bar{h}_iC^2}{4n(\bar{n}-1)}} \xrightarrow[m\to\infty]{\text{a.s.}} F_{\mathrm{ST}} = \theta^T.$$

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. (10) with uniform weights.

Now we simplify this estimator as the sample size of every subpopulation becomes infinite. First

bioRxiv preprint doi: https://doi.org/10.1101/083923; this version posted June 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

set the sample size of every subpopulation n_j equal to their mean $\bar{n}$, which implies $C^2 = 0$ and

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

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

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

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

which matches the $\hat{F}_{\mathrm{ST}}^{\mathrm{indep}}$ in Eqs. (2) to (4) as desired. Note the number of subpopulations n remains finite, and the sample heterozygosity $\bar{h}_i$ is not needed in the limit.

S2.2 The Hudson F_{ST} estimator

The Hudson pairwise F_{ST} estimator [20] measures the differentiation of two subpopulations (j, k). The estimator and its limit for two independent subpopulations ($\theta_{jk}^T = 0$) is

$$\hat{F}_{\mathrm{ST}}^{\mathrm{Hudson}} = \frac{\sum_{i=1}^{m}\left(\hat{p}_{ij} - \hat{p}_{ik}\right)^2 - \frac{\hat{p}_{ij}(1-\hat{p}_{ij})}{2n_j-1} - \frac{\hat{p}_{ik}(1-\hat{p}_{ik})}{2n_k-1}}{\sum_{i=1}^{m}\hat{p}_{ij}\left(1-\hat{p}_{ik}\right) + \hat{p}_{ik}\left(1-\hat{p}_{ij}\right)} \xrightarrow[m\to\infty]{\text{a.s.}} F_{\mathrm{ST}} = \frac{\theta_{jj}^T + \theta_{kk}^T}{2}. \tag{S1}$$

S2.3 Generalized HudsonK F_{ST} estimator

Here we present the "HudsonK" estimator, which generalizes the Hudson pairwise F_{ST} estimator in Eq. (S1) to n independent subpopulations. Note that for independent subpopulations, the F_{ST} of all the subpopulations equals the mean pairwise F_{ST} of every pair of subpopulations:

$$\frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n}\left(\frac{\theta_{jj}^T + \theta_{kk}^T}{2}\right) = \frac{1}{n}\sum_{j=1}^{n}\theta_{jj}^T = F_{\mathrm{ST}}.$$

For that reason, averaging numerators and denominators of the pairwise estimator in Eq. (S1) before computing the ratio, we obtain the generalized estimator and a limit under independent

bioRxiv preprint doi: https://doi.org/10.1101/083923; this version posted June 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

subpopulations of

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

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

Note that unlike the WC estimator, $\hat{p}_i^T$ above weighs every subpopulation equally, so every individual is weighed inversely proportional to the sample sizes n_j of their subpopulation j.

Like $\hat{F}_{\mathrm{ST}}^{\mathrm{WC}}$, $\hat{F}_{\mathrm{ST}}^{\mathrm{HudsonK}}$ simplifies to $\hat{F}_{\mathrm{ST}}^{\mathrm{indep}}$ in Eqs. (2) to (4) in the limit of infinite sample sizes $n_j \to \infty$, where $\hat{p}_{ij} \to \pi_{ij}$ for every (i,j).

S3 Derivation of method-of-moment estimators

S3.1 F_{ST} estimator for independent subpopulations

Assuming the coancestry model in Eqs. (5) and (6) for independent subpopulations ($\theta_{jk}^T = 0$ for $j \neq k$), the first and second moments of the IAFs are:

$$\mathrm{E}[\pi_{ij}] = p_i^T, \tag{S2}$$

$$\mathrm{E}\left[\pi_{ij}^2\right] = \left(p_i^T\right)^2 + p_i^T\left(1-p_i^T\right)\theta_{jj}^T, \tag{S3}$$

$$\mathrm{E}\left[\pi_{ij}\pi_{ik}\right] = \left(p_i^T\right)^2 \quad \text{if} \quad j \neq k. \tag{S4}$$

$F_{\mathrm{ST}} = \frac{1}{n}\sum_{j=1}^{n}\theta_{jj}^T$ appears by averaging Eq. (S3) over j:

$$\mathrm{E}\left[\frac{1}{n}\sum_{j=1}^{n}\pi_{ij}^2\right] = \left(p_i^T\right)^2 + p_i^T\left(1-p_i^T\right)F_{\mathrm{ST}}. \tag{S5}$$

Since Eq. (S2) has the same value for every j, and Eq. (S4) as well for every $j \neq k$, we average these to reduce estimation variance. The results are in terms of $\hat{p}_i^T = \frac{1}{n}\sum_{j=1}^{n}\pi_{ij}$:

$$\mathrm{E}\left[\hat{p}_i^T\right] = \mathrm{E}\left[\frac{1}{n}\sum_{j=1}^{n}\pi_{ij}\right] = p_i^T, \tag{S6}$$

$$\mathrm{E}\left[\left(\hat{p}_i^T\right)^2\right] = \mathrm{E}\left[\frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n}\pi_{ij}\pi_{ik}\right] = \left(p_i^T\right)^2 + p_i^T\left(1-p_i^T\right)\frac{1}{n}F_{\mathrm{ST}}. \tag{S7}$$

bioRxiv preprint doi: https://doi.org/10.1101/083923; this version posted June 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

F_{ST} also appears in Eq. (S7) because $j = k$ terms are introduced in the double sum. Subtracting Eq. (S5) and Eq. (S7) in turn from Eq. (S6) results in:

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

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

To reduce variance further, we average across loci, giving

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

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

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

$$\hat{F}_{\mathrm{ST}}^{\mathrm{std}} = \frac{\sum_{i=1}^{m}\left(\frac{1}{n}\sum_{j=1}^{n}\pi_{ij}^2 - \left(\hat{p}_i^T\right)^2\right)}{\sum_{i=1}^{m}\left(\hat{p}_i^T\left(1-\hat{p}_i^T\right) + \frac{1}{n}\left(\frac{1}{n}\sum_{j=1}^{n}\pi_{ij}^2 - \hat{p}_i^T\right)\right)} \tag{S8}$$

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

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

so substituting it into Eq. (S8) recovers Eq. (4) as desired:

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

S3.2 Standard kinship estimator

Here we assume the kinship model in Eqs. (12) and (13). Since Eq. (12) is the same for all individuals j, we average these first moments to reduce variance,

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

bioRxiv preprint doi: https://doi.org/10.1101/083923; this version posted June 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

which results in the following estimator of p_i^T:

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

Each φ_{jk}^T appears once per (j,k) pair in Eq. (13), recast here in terms of the sample covariance:

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

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

$$\mathrm{E}\left[\frac{1}{m}\sum_{i=1}^{m}\left(x_{ij} - 2p_i^T\right)\left(x_{ik} - 2p_i^T\right)\right] = 4\varphi_{jk}^T\frac{1}{m}\sum_{i=1}^{m} p_i^T\left(1 - p_i^T\right). \tag{S9}$$

Plugging $\hat{p}_i^T$ into Eq. (S9) and solving for φ_{jk}^T recovers Eq. (11) as desired:

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

S4 Proofs that F_{ST} and kinship estimator limits are constants with respect to the ancestral population T

In our work we calculate the limits of several estimators, which are given in terms of an arbitrary ancestral population T (not necessarily the MRCA, unless otherwise noted). The apparent paradox that the limit of an estimator would vary depending on the choice of T is resolved since these limits are in fact constant with respect to T. All proofs depend on the following IBD identities for change of ancestral population (see Section 3.4 of Part I for details):

$$\begin{aligned}\left(1 - f_j^A\right) &= \left(1 - f_j^B\right)\left(1 - f_B^A\right),\\ \left(1 - \varphi_{jk}^A\right) &= \left(1 - \varphi_{jk}^B\right)\left(1 - f_B^A\right),\end{aligned} \tag{S10}$$

where A, B are two possible ancestral populations for the individuals j, k, and A is ancestral to B.

S4.1 Proof that the limit of $\hat{F}_{\mathrm{ST}}^{\mathrm{indep}}$ does not depend on T

Here we study the limit of $\hat{F}_{\mathrm{ST}}^{\mathrm{indep}}$ in Eq. (7). Let S be a reference population ancestral to the individuals in question and T be another population ancestral to S. Denote the key parameters relative to S by $F_{\mathrm{ST}}^S, \bar{\theta}^S$ and relative to T by $F_{\mathrm{ST}}^T, \bar{\theta}^T$. The equations that relate both quantities

bioRxiv preprint doi: https://doi.org/10.1101/083923; this version posted June 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

satisfy our IBD shift identity (which follows by averaging Eq. (S10) over individuals for F_{ST} or pairs of individuals for $\bar{\theta}^T$):

$$\left(1-F_{\mathrm{ST}}^T\right)=\left(1-F_{\mathrm{ST}}^S\right)\left(1-f_S^T\right),$$
$$\left(1-\bar{\theta}^T\right)=\left(1-\bar{\theta}^S\right)\left(1-f_S^T\right).$$

Solving for the values relative to S gives

$$F_{\mathrm{ST}}^S=\frac{F_{\mathrm{ST}}^T-f_S^T}{1-f_S^T}, \qquad \bar{\theta}^S=\frac{\bar{\theta}^T-f_S^T}{1-f_S^T}.$$

The desired equality of the limit for both S and T follows:

$$\begin{aligned}\frac{n\left(F_{\mathrm{ST}}^S-\bar{\theta}^S\right)}{n-1+F_{\mathrm{ST}}^S-n\bar{\theta}^S}&=\frac{n\left(\frac{F_{\mathrm{ST}}^T-f_S^T}{1-f_S^T}-\frac{\bar{\theta}^T-f_S^T}{1-f_S^T}\right)}{n-1+\frac{F_{\mathrm{ST}}^T-f_S^T}{1-f_S^T}-n\frac{\bar{\theta}^T-f_S^T}{1-f_S^T}}\\&=\frac{n\left(F_{\mathrm{ST}}^T-\bar{\theta}^T\right)}{(n-1)\left(1-f_S^T\right)+\left(F_{\mathrm{ST}}^T-f_S^T\right)-n\left(\bar{\theta}^T-f_S^T\right)}\\&=\frac{n\left(F_{\mathrm{ST}}^T-\bar{\theta}^T\right)}{n-1+F_{\mathrm{ST}}^T-n\bar{\theta}^T}.\end{aligned}$$

S4.2 Proof that the limit of $\hat{\varphi}_{jk}^{T,\mathbf{std}}$ does not depend on T

Here we study the limit of the standard kinship estimator $\hat{\varphi}_{jk}^{T,\mathrm{std}}$ in Eq. (14). Let S be a reference population ancestral to the individuals in question and T be another population ancestral to S. The equations that relate the terms relative to S and those relative to T follow from Eq. (S10) just as in the previous subsection:

$$\varphi_{jk}^S=\frac{\varphi_{jk}^T-f_S^T}{1-f_S^T}, \qquad \bar{\varphi}_j^S=\frac{\bar{\varphi}_j^T-f_S^T}{1-f_S^T},$$
$$\bar{\varphi}_k^S=\frac{\bar{\varphi}_k^T-f_S^T}{1-f_S^T}, \qquad \bar{\varphi}^S=\frac{\bar{\varphi}^T-f_S^T}{1-f_S^T}.$$

The desired result follows:

$$\frac{\varphi_{jk}^S-\bar{\varphi}_j^S-\bar{\varphi}_k^S+\bar{\varphi}^S}{1-\bar{\varphi}^S}=\frac{\varphi_{jk}^T-\bar{\varphi}_j^T-\bar{\varphi}_k^T+\bar{\varphi}^T}{1-\bar{\varphi}^T}.$$

S5 Mean coancestry bounds

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

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

bioRxiv preprint doi: https://doi.org/10.1101/083923; this version posted June 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

and for uniform weights $\frac{1}{n}F_{\mathrm{ST}} \leq \bar{\theta}^T$. Furthermore, $\bar{\theta}^T = F_{\mathrm{ST}}$ iff $\theta^T_{jk} = F_{\mathrm{ST}}$ for all (j,k), and $\bar{\theta}^T = \frac{1}{n}F_{\mathrm{ST}}$ for the independent subpopulations model.

The Cauchy-Schwarz inequality for covariances implies $\theta^T_{jk} \leq \sqrt{\theta^T_{jj}\theta^T_{kk}}$. Therefore,

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

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

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

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

S6 Moments of estimator building blocks

Here we calculate first and some second moments for "building block" quantities that recur in our estimators, particularly terms involving x_{ij} and $\hat{p}^T_i$, and which enable us to calculate the limits of our estimators. Below are examples for genotypes, which follow from Eqs. (12) and (13); calculations for IAFs follow analogously from Eqs. (5) and (6) (not shown).

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

bioRxiv preprint doi: https://doi.org/10.1101/083923; this version posted June 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

S7 Derivation of new kinship estimator

To begin the method-of-moments derivation, we compute the raw first and second moments from the kinship model of Eqs. (12) and (13).

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

For obtain a symmetric estimator, we also compute the raw moments of $2 - x_{ij}$ (which counts the alternative allele):

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

If we solved for p_i^T using the first moment equations, we would recover the standard kinship estimator of Eqs. (10) and (11), so we shall avoid this strategy.

To proceed, we average the two second moment equations above. Note that

$$\begin{aligned}
\frac{1}{2}\left(x_{ij}x_{ik} + (2-x_{ij})(2-x_{ik})\right) &= (1-x_{ij})(1-x_{ik}) + 1, \\
\frac{1}{2}\left(\left(p_i^T\right)^2 + \left(1-p_i^T\right)^2\right) &= \frac{1}{2} - p_i^T\left(1-p_i^T\right).
\end{aligned}$$

Therefore, the symmetric estimator (which gives the same calculation if the reference allele is switched) is

$$\begin{aligned}
\mathrm{E}\left[(1-x_{ij})(1-x_{ik}) + 1|T\right] &= 2 + 4p_i^T\left(1-p_i^T\right)\left(\varphi_{jk}^T - 1\right) \quad \Rightarrow \\
\mathrm{E}\left[(1-x_{ij})(1-x_{ik}) - 1|T\right] &= 4p_i^T\left(1-p_i^T\right)\left(\varphi_{jk}^T - 1\right).
\end{aligned}$$

A genome-wide estimate is obtained by averaging the previous statistics across loci, resulting in

$$\begin{aligned}
A_{jk} &= \frac{1}{m}\sum_{i=1}^{m}(x_{ij}-1)(x_{ik}-1) - 1, \\
\mathrm{E}\left[A_{jk}|T\right] &= \left(\varphi_{jk}^T - 1\right)v_m^T, \quad \text{where} \\
v_m^T &= \frac{4}{m}\sum_{i=1}^{m} p_i^T\left(1-p_i^T\right).
\end{aligned}$$

The new kinship estimator follows from obtaining a consistent estimator of the limit of v_m^T as m goes to infinity, and applying it to solve for φ_{jk}^T in the above equation for the expectation of A_{jk}, as detailed in Section 5.

bioRxiv preprint doi: https://doi.org/10.1101/083923; this version posted June 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

S8 Admixture and independent subpopulations model simulations

S8.1 Construction of subpopulation allele frequencies

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

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

which results in subpopulation allele frequencies that obey the coancestry model of Eqs. (5) and (6), with $\text{E}\left[p_i^{S_u}\middle|T\right] = p_i^T$ and $\text{Var}\left(p_i^{S_u}\middle|T\right) = f_{S_u}^T p_i^T\left(1 - p_i^T\right)$ [5], as desired.

S8.2 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 $\mathbf{x} \sim \text{Dirichlet}(1, ..., 1)$ 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 $F_{\text{ST}} = \frac{1}{K}\sum_{u=1}^{K} f_{S_u}^T$ from the independent subpopulations model (Section 3.3.2 of Part I), which HudsonK estimates.

S8.3 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 $[1, K]$ and individuals span $[\frac{1}{2}, K + \frac{1}{2}]$. For the F_{ST} estimators that require subpopulations, individual j is assigned to the nearest subpopulation

bioRxiv preprint doi: https://doi.org/10.1101/083923; this version posted June 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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.

S8.4 Choosing σ and τ

Here we find values for σ (controls q_{jk}) and τ (scales $f^T_{S_u}$) that give $s^T = \frac{1}{2}$ and $F_{\mathrm{ST}} = 0.1$ in the admixture model. We previously found that $\theta^T_{jk} = \sum_{u=1}^{K} q_{ju} q_{ku} f^T_{S_u}$ and $F_{\mathrm{ST}} = \sum_{j=1}^{n} \sum_{u=1}^{K} w_j q^2_{ju} f^T_{S_u}$ for the BN-PSD model (Section 6.1 of Part I). In our simulation, $w_j = \frac{1}{n}$ and $f^T_{S_u} = \frac{u}{K}\tau$, so $\theta^T_{jk} = \frac{\tau}{K} \sum_{u=1}^{K} u q_{ju} q_{ku}$ and $F_{\mathrm{ST}} = \frac{\tau}{nK} \sum_{j=1}^{n} \sum_{u=1}^{K} u q^2_{ju}$. Therefore,

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

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

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

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

S9 Prediction intervals of F_{ST} estimators

Prediction intervals with $\alpha = 95\%$ correspond to the range of $n = 39$ independent F_{ST} estimates. In the general case, n independent statistics are given in order $X_{(1)} < \ldots < X_{(n)}$. Then $I = [X_{(j)}, X_{(n+1-j)}]$ is a prediction interval with confidence $\alpha = \frac{n+1-2j}{n+1}$ [85]. 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.