Title Rank-invariant Estimation of Inbreeding Coefficients.

Authors Qian S. Zhang
Department of Biostatistics, University of Washington
Seattle WA 98195-1617, USA

Jérôme Goudet
Department of Ecology and Evolution, University of Lausanne
CH-1015 Lausanne, Switzerland
https://orcid.org/0000-0002-5318-7607

Bruce S. Weir
Department of Biostatistics, University of Washington
Seattle WA 98195-1617, USA
https://orcid.org/0000-0002-4883-1247

Corresponding Author Bruce S. Weir
Department of Biostatistics, University of Washington
Seattle WA 98195-1617, USA
Telephone: +1 (206) 221-7947; Fax: +1 (206) 543-3286
Email: bsweir@uw.edu

Running Title Estimation of Inbreeding Coefficients.

Word Count 6894 (Pages 3-19.)

ABSTRACT

The two alleles an individual carries at a locus are identical by descent (ibd) if they have descended from a single ancestral allele in a reference population, and the probability of such identity is the inbreeding coefficient of the individual. Inbreeding coefficients can be predicted from pedigrees with founders in the reference population, but estimation from genetic data is not possible unless data are available from the reference population. Published estimators, at best, estimate inbreeding coefficients relative to average ibd probabilities for some specified set of alleles. Estimators that make explicit use of sample allele frequencies as estimates of allele probabilities in the reference population have additional confounding when study individuals have different average kinships with the remaining individuals. This means that the ranking of those individual inbreeding coefficient estimates depends on the study sample and we show the variation in rankings for common estimators applied to 1000 Genomes data. Allele-sharing estimators of within-population inbreeding coefficients for a set of individuals, however, do have invariant rankings across all studies including those individuals. They are unbiased with a large number of SNPs. We discuss how allele sharing estimates of within-population inbreeding coefficients are the relevant quantities for a range of empirical applications.

Keywords: Estimation, F-statistics, Identity by descent, Inbreeding, Kinship, SNP data

INTRODUCTION

Allelic dependence at a locus is usually quantified by inbreeding coefficients for individuals or populations, with these measures referring either to correlations of allelic state indicators (Wright 1922) or to probabilities of identity by descent, ibd, (Malécot 1948). In this paper we use ibd and we have advocated the use of allele-sharing estimators (Weir and Goudet 2017: WG17 henceforth, Goudet et al. 2018) that are unbiased for individual and population inbreeding coefficients relative to average kinships among specified pairs of individuals. Estimators, such as those in PLINK (Purcell et al. 2007) and GCTA (Yang et al. 2011), that use allele frequencies from a sample confound inbreeding estimates by the averages of individual kinships. Our work is also influenced by the need to estimate inbreeding coefficients from many millions of SNP genotypes where likelihood methods may not be feasible and instead we employ moment-based methods.

There have been many published accounts of inbreeding estimation, including the recent evaluation of several methods by Alemu et al. (2021) in this journal. Among those that refer to allele sharing, Li and Horvitz (1953) discussed an inbreeding estimator based on observed homozygosity, i.e. within-individual sharing of maternal and paternal alleles. They compared observed sharing to the value expected under zero inbreeding. They also constructed an estimator from the proportions of each allele type in a sample that were homozygous and gave an expression that was investigated further by Ritland (1996). Ritland used allele sharing within and between individuals in his work, and his inbreeding estimates assumed "independence or near-independence" between individuals. If individuals are not independent, then we show below that the rankings of his inbreeding coefficient estimates change with the sample. In WG17 we estimated inbreeding coefficients by comparing within-individual allele-sharing to average sharing between pairs of individuals in a sample. By not making explicit use of sample allele frequencies, we preserved the ranking of estimates across different samples and this will be a central theme of the present paper.

Ritland's individual-level inbreeding coefficients were also derived Yang et al. (2011) as the correlation between uniting gametes and were expressed in terms of allele dosages for an individual and sample allele frequencies. This estimator was written as $\hat{F}_{\mathrm{UNI}}$ in Yengo et al. (2017), and is less biased than another estimator in Yang et al. (2011) obtained from the diagonal elements of a genomic relationship matrix (GRM) of VanRaden (2008). We compare these two estimates below with allele-sharing and other methods: pedigree-based path-counting (Wright 1922), maximum-likelihood (e.g. Hall et al. 2012) and runs of homozygosity (e.g. Ceballos et al. 2018).

METHODS

Statistical sampling

We can describe the dependence between pairs of uniting alleles with data from a single population without invoking an evolutionary model for the history of the population. In this "statistical sampling" framework (Weir 1996) we do not consider the variation associated with stochastic evolutionary processes but we do consider the variation among samples from the same population. Although extensive sets of genetic data allow individual-level inbreeding coefficients to be estimated with high precision, we start with population-level estimation.

Allelic dependencies can be quantified with the usual within-population inbreeding coefficient, written here as f_W to emphasize it is a within-population quantity, defined by

$$H_l \;=\; 2p_l(1-p_l)(1-f_W) \tag{1}$$

where H_l is the population proportion of heterozygotes for the reference allele at SNP l and p_l is the population proportion of that reference allele. The same value of f_W is assumed to apply for all SNPs. An immediate consequence of this definition is that the population proportions of homozygotes for the reference and alternative alleles are $p_l^2+p_l(1-p_l)f_W$ and $(1-p_l)^2+p_l(1-p_l)f_W$ respectively. This formulation allows f_W to be negative, and it is bounded below by the maximum of $-p_l/(1-p_l)$ and $-(1-p_l)/p_l$. It is bounded above by 1. Hardy-Weinberg equilibrium, HWE, corresponds to $f_W = 0$ and textbooks (e.g. Hedrick 2000) point out that negative values of f_W indicate more heterozygotes than expected under HWE.

Observed heterozygote proportions $\tilde{H}_l$ have H_l as within-population expectation $\mathcal{E}_W$ over samples from the study population, $\mathcal{E}_W(\tilde{H}_l) = H_l$, and this would provide a simple estimator of f_W if the allele population proportions were known. In practice, however, these proportions are not known. Steele et al. (2014) suggested use of a database external to the study sample to provide reference allele proportions in forensic applications where a reference database is used for making inferences about the population relevant for a particular crime. The more usual approach is to use sample proportions $\tilde{p}_l$ in the study sample in place of the true proportions p_l (equation 1 of Li and Horvitz, 1953):

$$\hat{f}_{W_l} \;=\; 1-\frac{\tilde{H}_l}{2\tilde{p}_l(1-\tilde{p}_l)} \tag{2}$$

The moment estimator in Equation 2 is also an MLE of f_W when only one locus is considered, but

it is biased (Robertson and Hill 1984) since not only is it a ratio of statistics but also the expected value $\mathcal{E}_W[2\tilde{p}_l(1-\tilde{p}_l)]$ over repeated samples of n randomly chosen individuals from the population is $2p_l(1-p_l)[1-(1+f_W)/(2n)]$ (e.g. Weir 1996, p39).

This approach can be used to estimate the within-population inbreeding coefficient f_j for each individual j in a sample from one population. These are the "simple" estimators of Hall et al. (2017) and the $\hat{f}_{\mathrm{HOM}_j}$ of Yengo et al. (2017):

$$\hat{f}_{\mathrm{HOM}_{jl}} = 1 - \frac{\tilde{H}_{jl}}{2\tilde{p}_l(1-\tilde{p}_l)} \tag{3}$$

The sample heterozygosity indicator $\tilde{H}_{jl}$ is one if individual j is heterozygous at SNP l and is zero otherwise. Averaging Equation 3 over individuals gives the estimator based on SNP l in Equation 2 although it is the individual-specific values with which we are concerned in this paper.

A single SNP provides estimates that are either 1 or a negative value depending on $\tilde{p}_l$, so many SNPs are used in practice. In both Hall et al. (2012) and Yengo et al. (2017) data were combined over loci as weighted or "ratio of averages" estimators:

$$\hat{f}_{\mathrm{Hom}_j} = 1 - \frac{\sum_l(\tilde{H}_{jl})}{\sum_l[2\tilde{p}_l(1-\tilde{p}_l)]} \tag{4}$$

Gazal et al. (2014) referred to this estimator as f_{PLINK} as it is an option in PLINK. We show below the generally good performance of this weighted estimator even though it is a function of sample allele frequencies. We will consider throughout that a large number L of SNPs are used so that ratios of sums of statistics over loci, such as in Equation 4, have expected values equal to the ratio of expected values of their numerators and denominators. Ochoa and Storey (2021) showed statistics of the form $\tilde{A}_L/\tilde{B}_L$, where $\tilde{A}_L = \sum_{l=1}^{L} a_l/L$ and $\tilde{B}_L = \sum_{l=1}^{L} b_l/L$, have expected values that converge almost surely to the ratio A/B when $\mathcal{E}_W(\tilde{A}_L) = Ac_L$ and $\mathcal{E}_W(\tilde{B}_L) = Bc_L$. This result requires $|a_l|, |b_l|$ to both be no greater than some finite quantity C, c_L to converge to a finite value c as L increases, and for Bc not to be zero. For the ratio in Equation 4, $a_l = \tilde{H}_{jl}$, $b_l = 2\tilde{p}_l(1-\tilde{p}_l)$ so $A = (1-f_j)$, $B = 1$ for large sample sizes n, and $c_L = \sum_l 2p_l(1-p_l)/L \leq 1/2$ so the conditions are satisfied, providing at least one SNP is polymorphic. For an "average of ratios" estimator of the form $\sum_{l=1}^{L}(a_l/b_l)/L$, the denominators b_l can be very small and convergence of its expected value is not assured.

As an alternative to using sample allele frequencies, Hall et al. (2012) used maximum likelihood to estimate population allele proportions for multiple loci whereas Ayres and Balding (1998) used Markov chain Monte Carlo methods in a Bayesian approach that integrated out the allele proportion

parameters. Neither of those papers considered data of the size we now face in sequence-based studies of many organisms, and we doubt the computational effort to estimate, or integrate over, hundreds of millions of allele proportions in Equations 2 or 3 adds much value to inferences about f. The allele-sharing estimators we describe in the next section regard allele probabilities as unknown nuisance parameters and we show how to avoid estimating them or assigning them values.

Hall et al. (2012) used an EM algorithm to find MLEs for f_j when population allele proportions were regarded as being known and equal to sample proportions. Alternatively, a grid search can be conducted over the range of validity for the single parameter f_j that maximizes the log-likelihood

$$\ln[\mathrm{Lik}(f_j)] \;=\; \mathrm{Constant} + \sum_{l=1}^{L}\{\tilde{H}_{jl}\ln[(1-f_j)] + (1-\tilde{H}_{jl})\ln[1-2\tilde{p}_l(1-\tilde{p}_l)(1-f_j)]\} \qquad (5)$$

Estimation of the within-population inbreeding coefficients f_W (F_{IS} of Wright 1922) and f_j does not require any information beyond genotype proportions in samples from a study population, nor does it make any assumptions about that population or the evolutionary forces that shaped the population. The coefficients are simply measures of dependence of pairs of alleles within individuals. We show in the next section that, in the absence of additional information, these coefficients also govern the behavior of common published inbreeding estimators for the probability of alleles being identical by descent.

Genetic Sampling

Inbreeding parameters of most interest in genetic studies are those that recognize the contribution of previous generations to inbreeding in the present study population. This requires accounting for "genetic sampling" (Weir 1996) between generations, thereby leading to an ibd interpretation of inbreeding: ibd alleles descend from a single allele in a reference population. It also allows the prediction of inbreeding coefficients by path counting when pedigrees are known (Wright 1922). If individual J is ancestral to both individuals j' and j'', and if there are n individuals in the pedigree path joining j' to j'' through J, then $F_j = \sum(0.5)^n(1+F_J)$ where F_J is the inbreeding coefficient of ancestor J and F_j is the inbreeding coefficient of offspring j of parents j' and j''. The sum is over all ancestors J and all paths joining j' to j'' through J. The expression is also the coancestry $\theta_{j'j''}$ of j' and j'': the probability an allele drawn randomly from j' is ibd to an allele drawn randomly from j''.

The allele proportion p_l in a study population has expectation π_l over evolutionary replicates of

the population from an ancestral reference population to the present time. Sample allele proportions $\tilde{p}_l$ provide information about the population proportions p_l, and their statistical sampling properties follow from the binomial distribution. We do not invoke a specific genetic sampling distribution for the p_l about their expectations π_l although we do assume the second moments of that distribution depend on probabilities of ibd for pairs of alleles. One consequence of the assumed moments is that the probability of individual j in the study population being heterozygous, i.e. the total expected value $\mathcal{E}_T$ of the heterozygosity indicator over replicates of the history of that individual, is

$$\mathcal{E}_T(\tilde{H}_{j_l}) \;\; = \;\; 2\pi_l(1-\pi_l)(1-F_j) \tag{6}$$

The quantity F_j is the individual-specific version of F_{IT} of Wright (1922) and we can regard it as the probability the two alleles at any locus for individual j are ibd. There is an implicit assumption in Equation 6 that the reference population needed to define ibd is infinite and in HWE: there is probability F_j that j has homologous alleles with a single ancestral allele in that population and probability $(1-F_j)$ of j having homologous alleles with distinct ancestral alleles there. In the first place, the single ancestral allele has probability π of being the reference allele for that locus and the implicit assumption is that two ancestral alleles are both the reference type with probability π^2. This does not mean there is an actual ancestral population with those properties, any more than use of $\mathcal{E}_T$ means there are actual replicates of the history of any population or individual, and we note that Equation 6 does not allow higher heterozygosity than predicted by HWE. Nonetheless, the concept of ibd allows theoretical constructions of great utility and we now present a framework for approaching empirical situations.

Inbreeding, or ibd, implies a common ancestral origin for uniting alleles and statements about sample allele proportions $\tilde{p}_l$ require consideration of possible ibd for other pairs of alleles in the sample. The total expectation of $2\tilde{p}_l(1-\tilde{p}_l)$ over samples from the population and over evolutionary replicates of the study population is (Weir 1996, p176)

$$\mathcal{E}_T[2\tilde{p}_l(1-\tilde{p}_l)] \;\; = \;\; 2\pi_l(1-\pi_l)\left[(1-\theta_S) - \frac{1}{2n}\,(1+F_W-2\theta_S)\right] \tag{7}$$

where F_W is the average inbreeding coefficient in the sample, $F_W = \sum_{j=1}^{n} F_j/n$, and θ_S is the average coancestry in the sample, $\theta_S = \sum_{j=1}^{n}\sum_{j'\neq j}\theta_{jj'}/[n(n-1)]$. Equivalent expressions were given by McPeek et al. (2004) and DeGiorgio and Rosenberg (2008). We note the relationship $f_W = (F_W-\theta_S)/(1-\theta_S)$ given by Wright (1922) and we showed in WG17 the equivalent expression $f_j = (F_j-\theta_S)/(1-\theta_S)$ for individual-specific values (θ_S is Wright's F_{ST}).

For a large number of SNPs, the expectation of a ratio estimator of the type considered here is the ratio of expectations (Ochoa and Storey 2021). Therefore, the total expectations of the $\hat{f}_{\mathrm{Hom}_j}$, taking into account both statistical and genetic sampling, are

$$\mathcal{E}_T(\hat{f}_{\mathrm{HOM}_j}) \;=\; 1 - \frac{1 - F_j}{(1-\theta_S) - \frac{1}{2n}(1 + F_W - 2\theta_S)} = \frac{f_j - \frac{1}{2n}(1+f_W)}{1 - \frac{1}{2n}(1+f_W)} \tag{8}$$

For all sample sizes, $\hat{f}_{\mathrm{HOM}_j}$ has an expected value less than the true value f_j, with the bias being of the order of $1/n$. The ranking of $\mathcal{E}_T(\hat{f}_{\mathrm{HOM}_j})$ values, however, is the same as the ranking of the f_j and, therefore, of the F_j. For large sample sizes, Equation 8 reduces to $\mathcal{E}_T(\hat{f}_{\mathrm{HOM}_j}) = f_j$. Averaging over individuals shows that $\mathcal{E}_T(\hat{f}_{\mathrm{HOM}}) = f_W$: the population-level estimator in Equation 2 has total expectation of f_W, not F_W.

A different outcome is found for the f_{UNI_j} estimator of Yengo et al. (2017) (i.e. $\hat{f}^{III}$ of Yang et al. 2011; f_{GCTA3} of Gazal et al. 2014). This estimator, with the weighted (w) ratio of averages over loci we recommend, as opposed to the unweighted (u) average of ratios over loci used in their papers, is

$$\hat{f}^{w}_{\mathrm{UNI}_j} \;=\; \frac{\sum_{l=1}^{L}[X_{jl}^2 - (1+2\tilde{p}_l)X_{jl} + 2\tilde{p}_l^2]}{\sum_{l=1}^{L} 2\tilde{p}_l(1-\tilde{p}_l)} \tag{9}$$

In this equation X_{jl} is the reference allele dosage, the number of copies of the reference allele, at SNP l for individual j. It is equivalent to the estimator given by Ritland (1996, equation 5) and attributed by him to Li and Horvitz (1953).

Ochoa and Storey (2021) showed that $\hat{f}^{w}_{\mathrm{UNI}_j}$ has expectation, for a large number of SNPs and a large sample size, of

$$\mathcal{E}_T(\hat{f}^{w}_{\mathrm{UNI}_j}) \;=\; \frac{F_j - 2\Psi_j + \theta_S}{1-\theta_S} = f_j - 2\psi_j \tag{10}$$

where Ψ_j is the average coancestry of individual j with other members of the study sample: $\Psi_j = \sum_{j'=1, j'\neq j}^{n} \theta_{jj'}/(n-1)$. We term $\psi_j = (\Psi_j - \theta_S)/(1-\theta_S)$ the within-population individual-specific average kinship coefficient. The Ψ_j have an average of θ_S over members of the sample, so the average of the ψ_j's is zero and expected value of the average of the $\hat{f}^{w}_{\mathrm{UNI}_j}$ is f_W, as is the case for $\hat{f}_{\mathrm{AS}_j}$.

Equation 10 shows that the $\hat{f}^{w}_{\mathrm{UNI}_j}$ have expected values with the same ranking as the F_j values only if there is no kinship among pairs of individuals or if every individual j in the sample has the same average kinship ψ_j with other sample members.

Finally, we mention another common estimator described by VanRaden (2008) and termed f_{GCTA1} by Gazal et al. (2014) and available from the GCTA software (Yang et al., 2011) with option `--ibc`. We referred to this as the "standard" estimator in WG17. The weighted version for multiple loci is

$$\hat{f}^{w}_{\text{STD}_j} = \frac{\sum_l (X_{jl} - 2\tilde{p}_l)^2}{\sum_l 2\tilde{p}_l(1-\tilde{p}_l)} - 1 \tag{11}$$

and it has the large-sample expectation of $(f_j - 4\psi_j)$ as is implied by Equation 13 of WG17 and as was given by Ochoa and Storey (2021). We summarize the various measures of inbreeding and coancestry in Table 1, and we include sample sizes in the expectations shown in Table 2.

Tables 1 and 2

The $\hat{f}_{\text{HOM}}$, $\hat{f}_{\text{UNI}}, \hat{f}_{\text{STD}}$ and $\hat{f}_{\text{MLE}}$ estimators of individual or population inbreeding coefficients make explicit use of sample allele proportions. This means that all four have small-sample biases, and none of the four provide estimates of the ibd quantities F or F_j. We showed that $\hat{f}_{\text{HOM}}$ is actually estimating the within-population inbreeding coefficients: the total inbreeding coefficients *relative to* the average coancestry of pairs of individuals in the sample, but $\hat{f}_{\text{UNI}}$ and $\hat{f}_{\text{STD}}$ are estimating expressions that also involve average kinships ψ.

Allele Sharing

In a genetic sampling framework, and with the ibd viewpoint, we consider within-individual allele sharing proportions A_{jl} for SNP l in individual j (we used M rather than A in WG17 and in Goudet et al. 2018). These equal one for homozygotes and zero for heterozygotes and sample values can be expressed in terms of allele dosages, $\tilde{A}_{jl} = (X_{jl} - 1)^2$. We also consider between-individual sharing proportions $A_{jj'l}$ for SNP l and distinct individuals j and j'. These are equal to one for both individuals being the same homozygote, zero for different homozygotes, and 0.5 otherwise. Observed values can be written as $\tilde{A}_{jj'l} = [1 + (X_{jl} - 1)(X_{j'l} - 1)]/2$, with an average over all pairs of distinct individuals in a sample of $\tilde{A}_{Sl}$. Astle and Balding (2009) introduced $\tilde{A}_{jj'l}$ as a measure of identity in state of alleles chosen randomly from individuals j and j', and Ochoa and Storey (2021) used a simple transformation of this quantity. The allele sharing for an individual with itself is $A_{jjl} = (1 + A_{jl})/2$.

The same logic that led to Equation 6 provides total expectations for allele-sharing proportions for all j, j':

$$\begin{aligned} \mathcal{E}_T(\tilde{A}_{jj'l}) &= 1 - 2\pi_l(1-\pi_l)(1-\theta_{jj'}) \\ \mathcal{E}_T(\tilde{A}_{Sl}) &= 1 - 2\pi_l(1-\pi_l)(1-\theta_S) \end{aligned}$$

Note that $\theta_{jj} = (1+F_j)/2$. The nuisance parameter $2\pi_l(1-\pi_l)$ cancels out of the ratio $\mathcal{E}_T(\tilde{A}_{jj'l} - \tilde{A}_{Sl})/\mathcal{E}_T(1-\tilde{A}_{Sl})$ and this motivates definitions of allele-sharing estimators of the inbreeding coefficient for individual j and the kinship coefficient for individuals j, j' as

$$\hat{f}_{\mathrm{AS}_j} = \frac{\sum_l(\tilde{A}_{j_l} - \tilde{A}_{S_l})}{\sum_l(1-\tilde{A}_{Sl})} \quad , \quad \hat{\psi}_{\mathrm{AS}_{jj'}} = \frac{\sum_l(\tilde{A}_{jj'l} - \tilde{A}_{S_l})}{\sum_l(1-\tilde{A}_{Sl})} \tag{12}$$

For a large number of SNPs, these are unbiased for f_j and $\psi_{jj'}$ for all sample sizes. We show below the satisfactory behavior of $\hat{f}_{\mathrm{AS}_j}$ for simulated data, and consistency of rankings over different sampling frames, such as population, ancestry group or whole world for the 1000 Genomes data (The 1000 Genomes Project Consortium 2015). We showed in WG17 there is no need to filter on minor allele frequency to preserve the lack of bias.

For large sample sizes, $(1-\tilde{A}_{Sl}) \approx 2\tilde{p}_l(1-\tilde{p}_l)$. Under that approximation, $\hat{f}_{\mathrm{AS}_j}$ is the same as $\hat{f}_{\mathrm{Hom}_j}$ but the approximation is not necessary in computer-based analyses. Summing the large-sample estimates over individuals not equal to j gives an estimator for the average individual kinship ψ_j:

$$\hat{\psi}_{\mathrm{AS}_j} = -\frac{\sum_l(X_{jl} - 2\tilde{p}_l)(1-2\tilde{p}_l)}{\sum_l 4\tilde{p}_l(1-\tilde{p}_l)} \tag{13}$$

Adding $2\hat{\psi}_{\mathrm{AS}_j}$ to $\hat{f}^w_{\mathrm{UNI}_j}$ gives $\hat{f}_{\mathrm{AS}_j}$, as expected, as does adding $4\hat{\psi}_{\mathrm{AS}_j}$ to $\hat{f}^w_{\mathrm{STD}_j}$. Similarly, $\hat{\psi}_{\mathrm{AS}_{jj'}}$ is obtained by adding $\hat{\psi}_{\mathrm{AS}_j}$ and $\hat{\psi}_{\mathrm{AS}_{j'}}$ to $\hat{\psi}_{\mathrm{STD}_{jj'}}$, where (Yang et al. 2011)

$$\hat{\psi}_{\mathrm{STD}_{jj'}} = \frac{\sum_l(X_{jl} - 2\tilde{p}_l)(X_{j'l} - 2\tilde{p}_l)}{\sum_l 4\tilde{p}_l(1-\tilde{p}_l)}$$

These are the elements of the first method for constructing the GRM given by VanRaden (2008).

When inbreeding and coancestry coefficients are defined as ibd probabilities they are non-negative, but the within-population values f and ψ will be negative for individuals, or pairs of individuals, having smaller ibd allele probabilities than do pairs of individuals in the sample, on average. Individual-specific values of f always have the same ranking as the individual-specific F values, and they are estimable. Negative estimates can be avoided by the transformation to

$(\hat{f}_{\mathrm{AS}_j} - \hat{f}^{\min}_{\mathrm{AS}_j})/(1 - \hat{f}^{\min}_{\mathrm{AS}_j})$ where $\hat{f}^{\min}_{\mathrm{AS}_j}$ is the smallest value over individuals of the $\hat{f}_{\mathrm{AS}_j}$'s. We don't see the need for this transformation, and we noted above the recognition of the utility of negative values. Ochoa and Storey (2021) wished to estimate F_j rather than f_j and, to overcome the lack of information about the ancestral population serving as a reference point for ibd, they assumed the least related pair of individuals in a sample have a coancestry of zero. We showed in WG17 that this brings estimates in line with path-counting predicted values when founders are assumed to be not inbred and unrelated, but we prefer to avoid the assumption. We stress that, absent external information or assumptions, F is not estimable. Instead, linear functions of F that describe ibd of target pairs of alleles relative to ibd in a specified set of alleles are estimable and have utility in empirical studies.

Runs of Homozygosity

Each of the inbreeding estimators considered so far has been constructed for individual SNPs and then combined over SNPs. Observed values of allelic state are used to make inferences about the unobserved state of identity by descent. Estimators based on runs of homozygosity (ROH), however, suppose that ibd for a region of the genome can be observed. Although F is the probability an individual has ibd alleles at any single SNP, in fact ibd occurs in blocks within which there has been no recombination in the paths of descent from common ancestor to the individual's parents. Whereas a single SNP can be homozygous without the two alleles being ibd, if many adjacent SNPs are homozygous the most likely explanation is that they are in a block of ibd (Gibson et al. 2006). There can be exceptions, from mutation for example, and several publications give strategies for identifying runs of homozygotes for which ibd may be assumed (e.g. Gazal et al. 2014, Joshi et al. 2015). These strategies include adjusting the size of the blocks, the numbers of heterozygotes or missing values allowed per block, the minor allele frequency, and so on. These software parameters affect the size of the estimates (Meyermans et al. 2020). More sophisticated methods (e.g. Narasimhan et al. 2016) use hidden Markov models where ibd is the hidden status of an observed homozygote. This model-based approach necessarily has assumptions, such as HWE in the sampled population.

We provide more details elsewhere, but we note here that ROH methods offer a useful alternative to SNP-by-SNP methods even though they cannot completely compensate for lack of information on the ibd reference population. We note also that shorter runs of ibd result from more distant

relatedness of an individual's parents, so that ROH procedures can be set to distinguish recent (familial) ibd from distant (evolutionary) ibd. SNP-by-SNP estimators do not make a distinction between these two time scales.

Simulation Study

We generated a founder set of 50 founder individuals with 20,000 SNPs over a 20 Morgan map by using the mspms program (Kelleher et al. 2016). We then used quantiNemo software (Neuenschwander et al. 2019) to simulate a five-generation pedigree of hermaphroditic individuals mating randomly, excluding selfing, with each mating producing a number of offspring drawn from a Poisson distribution with mean two. The zero-th generation was the 50 founders, the first generation had 47 individuals and the second, third, fourth and fifth generations had 58, 56, 57 and 65 individuals respectively. We followed the founder gametes through the pedigree using acustom R script and allowing for recombination based on the 20 Morgan genetic map.

Each of the 100 alleles per SNP among the founders was given a unique identifier so that alleles in subsequent generations with the same identifier had actual identity by descent relative to the founders. The average actual ibd proportions over loci, within individuals and between each pair of individuals, provided "gold standard" inbreeding and coancestry coefficients, as opposed to the pedigree-based values we calculated by path counting.

The pedigree was constructed to provide fairly high levels of predicted coancestry among pairs of the 283 non-founder individuals, ranging from zero to 0.464, with a mean of $\theta_S = 0.053$, assuming the 50 founders were unrelated and not inbred. The pedigree inbreeding coefficients ranged from zero to 0.367, with a mean of $F_W = 0.050$. The within-population inbreeding coefficient for the set of 283 non-founder individuals is $f = (F_W - \theta_S)/(1 - \theta_S) = 0.003$. Note, however, that the 50 individuals regarded as founders for the subsequent 283 had their own joint histories from the mspms simulation. These 50 had an average within-individual allele sharing of $\tilde{A}_W = 0.80385$ and an average between-individual allele sharing of $\tilde{A}_S = 0.80355$. The difference of these two proportions, which would be zero for a reference set of non-inbred and unrelated individuals, provides a within-founder allele-sharing inbreeding coefficient $\hat{f}_{\mathrm{w}}$ of 0.0015.

The various estimators of inbreeding examined with these data are shown in Table 2, and the correlation coefficients for each pair of estimates over the whole set of 283 non-founder individuals are shown in Table 3. There are very high correlations between pedigree and gold-standard values

and also very high correlations between $\hat{f}_{\rm HOM}$ and $\hat{f}_{\rm AS}$ values, both as expected. In populations without substructure, random mating would lead to similar values for inbreeding and coancestry levels, so $\hat{f}_{\rm AS}$ and $\hat{\psi}_{\rm AS}$ values would have similar values. There are lower correlations of $\hat{f}_{\rm UNI}$ and $\hat{f}_{\rm STD}$ with pedigree-based or gold-standard inbreeding coefficients since those estimates reflect both f and ψ.

We see in Table 3 that $\hat{F}_{\rm ROH}$ values are the most highly correlated with $F_{\rm Gold}$: this high correlation was obtained by adjusting the block size (100 SNPs) and the block overlap amount (50 SNPs) to bring estimates close to the known $F_{\rm Gold}$ values. In practice the $F_{\rm Gold}$ values are not known and the other estimators are all evaluated without external information. The high correlation of $\hat{f}_{\rm AS}$ and maximum likelihood values shows that $\hat{f}_{\rm MLE}$ is estimating f rather than F because it uses the sample allele frequencies in place of the unknown allele probabilities. The weighted and unweighted versions of $\hat{f}_{\rm UNI}$ are highly correlated with each other and with their gold values, but this is not the case for $\hat{f}_{\rm STD}$.

Table 3

Figure 1 (left) illustrates the linear relationship between $f_{{\rm Ped}_j}$ and $F_{{\rm Ped}_j}$: $f_{{\rm Ped}_j} = (F_{{\rm Ped}_j} - \theta_{{\rm Ped}_S})/(1 - \theta_{{\rm Ped}_S})$ where $\theta_{{\rm Ped}_S} = 0.053$ is the average coancestry of pairs of non-founders, also calculated from the pedigree. The $F_{{\rm Gold}_j}$ and $f_{{\rm Gold}_j}$ values are highly, and equally, correlated with the corresponding pedigree values, as is shown for $f_{{\rm Gold}_j}$ in Figure 1 (center). The variation we see in Figure 1 (center) for $f_{{\rm Gold}_j}$ around $F_{{\rm Ped}_j}$ reflects the relatively small number of 20K SNPs and the relatively small map length spanned by these SNPs. We have previously (Hill and Weir 2011) pointed out the variation of actual inbreeding about expected values, even for whole genomes, and Wang (2016) showed that the number of SNPs also has an effect. The expected lack of relationship between pedigree-based values of individual average coancestry ψ_j and individual inbreeding f_j, leading to variable rankings for some estimators based on sample allele frequencies, is shown in Figure 1 (right).

Figure 1

Figure 2 (left) illustrates the similarity of $\hat{F}_{\rm ROH}$ and $F_{\rm Gold}$ and Figure 2 (center) shows good agreement between $\hat{F}_{\rm ROH}$ and $\hat{f}_{\rm AS}$. Figure 2 (right) shows the low bias of the allele-sharing estimators $\hat{f}_{{\rm AS}_j}$ for the gold-standard within-population inbreeding coefficients $f_{{\rm Gold}_j}$. Figure 3 shows $\hat{f}_{{\rm UNI}_j}$ to

be a better estimator of f_{Gold_j} than is $\hat{f}_{\text{STD}_j}$, as noted by Yang et al. (2011), and better performance for the weighted than unweighted averages over SNPs.

Figures 2 and 3

1000 Genomes Data

We used 77m SNPs from the 22 autosomes for the 26 populations of the 1000 Genomes whole genome data to estimate inbreeding coefficients for all 2504 individuals in the project. Our focus was on the invariance of estimate rankings as the reference set of individuals changed from the population from which each individual was sampled, to the continental group for that population, to the whole world. We calculated the estimates $\hat{f}_{\text{AS}_j}$ and $\hat{f}^u_{\text{UNI}_j}$ for each individual and each reference set, and ranked estimates within each population. The two sets of estimates for all individuals are shown separately in Figure 4. Figures S1 and S2 show $\hat{f}^u_{\text{UNI}_j}$ versus $\hat{f}_{\text{AS}_j}$ for estimates and ranks respectively.

Figure 4

Figure 4 shows that within-population inbreeding coefficients $\hat{f}_{\text{AS}}$ for all 1000 Genomes populations (except the AMR group: CLM, MXL, PEL, PUR) are essentially the same, and generally close to zero, when they are estimated relative to average coancestry within each population or continental group but change when the complete set of 26 populations is used as a reference. These latter values compare the allele sharing for each individual to the same reference value, the average sharing over all pairs of individuals in the whole dataset. The world reference shows markedly different $\hat{f}_{\text{AS}}$ values for the African populations (AFR), reflecting their higher levels of genetic diversity. The rankings for $\hat{f}_{\text{AS}}$ within a population, by construction, do not change with reference set. There are some high value outliers when the world is used as a reference: four of the five highest values are from AMR/PEL. These high $\hat{f}_{\text{AS}}$ values reflect admixture, consanguineous matings and high evolutionary coancestry. On the other hand, the $\hat{f}_{\text{UNI}}$ values are higher for African individuals than for any other individuals when the allele frequencies are from all 26 populations: this reflects an African-specific pattern of negative average individual kinships ψ, rather than higher values of the inbreeding coefficients F.

The critical role that average kinship plays in inbreeding estimation is illustrated in Figure 5. With the world as reference set, the allele-sharing inbreeding estimates $\hat{f}_{\text{AS}}$ are tightly clustered for European (EUR) individuals, a little more diverse for East Asian (EAS) individuals, much

more diverse for South Asian (SAS) and African (AFR) individuals, and substantially diverse for American (AMR) individuals. These values are consistent with those reported for the numbers of variant sites per genome (The 1000 Genomes Project Consortium, 2015). The variation among African and American average kinships $\hat{\psi}_{\mathrm{AS}}$ is substantial: as these quantities determine how the expected values of $\hat{F}_{\mathrm{UNI}}$ and $\hat{F}_{\mathrm{STD}}$ differ from the f target parameters, it is clear that these estimates cannot be used to rank individuals by their inbreeding levels.

Figure 5

For the African population ASW, individual NA20294 has $\hat{f}_{\mathrm{AS}}$ values of $-0.009, 0.001, -0.130$ using ASW, AFR or World as a reference set and each estimate is ranked as number 16 among the 61 ASW estimates. The same individual has $\hat{f}^{u}_{\mathrm{UNI}}$ values of -0.007 (rank 36), 0.001 (rank 16) and 0.028 (rank 60) using ASW, AFR or World allele frequencies. Estimator $\hat{f}^{u}_{\mathrm{UNI}}$ indicates NA20294 to be among the least inbred of the ASW individuals when AFR sample allele frequencies are used, but among the most inbred when world-wide sample allele frequencies are used, even though the individual's own genotype is the same for each analysis. Other examples of rankings changing with reference population for $\hat{f}_{\mathrm{UNI}}$ are shown in Figure S3. This can have implications for studies of inbreeding depression, where trait values are regressed on estimated inbreeding coefficients.

A comparison of runs-of-homozygosity estimates $\hat{F}_{\mathrm{ROH}_j}$ with SNP-by-SNP estimates is shown in Figure 6. The ROH estimates were produced with the `--homozyg --homozyg-snp2 --homozyg-kb100` options in PLINK (Meyermans et al. 2020). The values of $\hat{F}_{\mathrm{ROH}_j}$ depend on the PLINK settings for minor allele frequency pruning and linkage disequilibrium pruning, as well as on SNP density, so their expected values may differ from the true F_j values. The left panel shows $\hat{f}_{\mathrm{AS}_j}$ values and these have a correlation of 0.998 with $\hat{F}_{\mathrm{ROH}_j}$. The right panel shows $\hat{f}_{\mathrm{UNI}_j}$ estimates and these appear to have little relationship with $\hat{F}_{\mathrm{ROH}_j}$.

Figure 6

Narasimhan et al. (2016) used a hidden Markov model for obtaining $\hat{f}_{\mathrm{ROH}_j}$ values (Figure S4). There is very good agreement with $\hat{f}_{\mathrm{AS}_j}$ values, providing the admixed AMR populations are not used. Gazal et al. (2015) also used a hidden Markov model to obtain inbreeding estimates, although their method requires sample allele frequencies and so may have estimates of F confounded by average individual-specific average kinships. However, there is good agreement of $\hat{f}_{\mathrm{AS}_j}$ values with the values given by Gazal et al. (Figure S5).

DISCUSSION

Discussions on the estimation of individual inbreeding coefficients generally refer to F, the probability an individual has pairs of homologous alleles that are identical by descent. Among the estimators we have considered here, $\hat{F}_{\mathrm{ROH}}$ addresses F by assuming that long runs of homozygous SNPs represent ibd regions. The ROH estimates, however, are conditional on the settings used to calculate the estimates, and actual ibd in short runs of homozygotes may be ignored, so the expected values of these estimators is not known. The Bayesian approach of Vogl et al. (2002) also addresses F but at the computational cost of estimating allele proportions in a reference population assumed to have zero inbreeding or relatedness. All the other estimators considered here are, instead, addressing the within-population inbreeding coefficient f that compares F values to ibd probabilities for pairs of individuals. There is no need to specify the reference population implicit in the definition of identity by descent. There is also no need to assume the particular individuals in a sample have an inbreeding coefficient of zero. For large numbers of SNPs, allele-sharing estimators $\hat{f}_{\mathrm{AS}}$ are unbiased for f for all sample sizes and have values for a set of individuals that have invariant ranks over studies that include that set. We show that estimators using sample allele frequencies are estimating some combination of f and of individual-specific average kinships ψ with individuals in the study. Estimators with expectations depending on ψ do not have invariant rankings, as we showed with data from the 1000 Genomes project as the study scope varied from the population to the continent to the world.

Our ibd-based model rests on expectations of allele-sharing proportions satisfying expressions such as Equation 6. There is no requirement for non-overlapping generations, or homogeneous populations, for example. This generality is a consequence of not needing allele frequencies, whether these refer to a population or to an individual.

The role of ibd probabilities in theoretical population and quantitative genetic contexts is well known, but we suggest it is rank-invariant estimators for the within-population parameters f_j that are of relevance for empirical studies and we offer the examples in the following sections.

Genotype Probabilities

There is often a need to estimate genotype probabilities from observed allele proportions using formulations with allele probabilities and ibd probabilities F (e.g. National Research Council 1996 for forensic science). Following Equation 7 we see that it is $2\tilde{p}_l(1-\tilde{p}_l)(1-f_j)$ rather than $2\tilde{p}_l(1-$

$\tilde{p}_l)(1-F_j)$ that is unbiased for $2\pi_l(1-\pi_l)(1-F_j)$ if F_j and f_j are known.

Inbreeding Depression

Inbreeding is known to affect, linearly, the expected value of quantitative traits, and studies of inbreeding depression often proceed by regressing trait means on inbreeding levels. In Yengo et al. (2017), we used $\hat{F}_{\mathrm{ROH}}$, $\hat{f}_{\mathrm{HOM}}$ and $\hat{f}_{\mathrm{UNI}}$ as inbreeding estimates. Kardos et al. (2018) pointed out that we did not discuss the distinction between F and f. We responded (Yengo et al. 2018) with reasons for not wishing to use $\hat{F}_{\mathrm{ROH}}$ and we could have pointed out the linear relationship between f_j and F_j and the high correlation we showed above between $\hat{f}_{\mathrm{AS}_j}$ and $\hat{F}_{\mathrm{ROH}_j}$ means that regressing on either $\hat{F}_{\mathrm{ROH}}$ or $\hat{f}_{\mathrm{AS}}$ should lead to similar results. A SNP with highly significant inbreeding depression revealed by regressing trait values on $\hat{F}_{\mathrm{ROH}}$ should also be highly significant when regressing on $\hat{f}_{\mathrm{AS}_j}$. In less-homogeneous populations than represented by the UK Biobank data (Allen et al. 2012) we used in Yengo et al. (2017), it would appear to be better to use $\hat{f}_{\mathrm{AS}_j}$ than $\hat{f}_{\mathrm{UNI}_j}$ to avoid any effects of individual-specific average kinships on inbreeding estimates. Alemu et al. (2021) pointed out that $\hat{f}_{\mathrm{HOM}}$ (and $\hat{f}_{\mathrm{AS}}$), gives equal weights to all SNPs, whereas $\hat{f}_{\mathrm{UNI}}$ gives greater weight to SNPs with rare alleles. Alemu et al. did not consider the role of individual average kinships in the bias of $\hat{f}_{\mathrm{UNI}}$.

Genetic Relatedness Matrix

Inbreeding is also known to affect, linearly, the additive component of genetic variance. For additive traits, the genetic variance for individual j is $(1+F_j)\sigma_A^2$ where σ_A^2 is the additive variance for populations in Hardy-Weinberg equilibrium. Consequently, the expected value of the sample variance $\tilde{V}_T$ of trait values over a sample of n individuals is (Speed et al. 2012)

$$\mathcal{E}_T(\tilde{V}_T) \;=\; \frac{1}{n}\left(\mathrm{tr}(\boldsymbol{G}) - \frac{1}{n-1}\Sigma_{\boldsymbol{G}}\right)\sigma_A^2 + \sigma_e^2$$

Here the trait is additive and the errors, with variance σ_e^2, are independent of genetic effects. The GRM $\boldsymbol{G}$ has trace $\mathrm{tr}(\boldsymbol{G})$ and sum of off-diagonal elements $\Sigma_{\boldsymbol{G}}$. If the GRM elements are $(1+F_j)$ on the diagonal and $2\theta_{jj'}$ off the diagonal then the trace is $n(1+F_W)$ and the sum of off-diagonal elements is $n(n-1)\theta_S$ so the genetic component of V_T is $(1+F_W-2\theta_S)\sigma_A^2$. If the GRM is replaced by a matrix with allele-sharing inbreeding and kinship estimates, this becomes $(1+f_W)\sigma_A^2$, reflecting

that it is the within-population estimated GRM that is used in practice. We show elsewhere that the same expected variance holds with GRMs constructed with $\hat{f}_{\text{STD}}$ or $\hat{f}_{\text{UNI}}$.

In summary, we have shown that inbreeding measures of utility in empirical studies are "within-population" with the choice of population being at the discretion of the investigator. With allele-sharing inbreeding estimators, the population specifies the set of individuals whose pairwise coancestry is the reference against which inbreeding is measured. For estimators making explicit use of sample allele frequencies, it is the population that furnishes those frequencies, although then inbreeding estimates are confounded by individual-specific average kinships. We showed algebraically and empirically that allele-sharing estimators have invariant rankings across choice of population.

SOFTWARE

Estimation of inbreeding coefficients can be performed with the following software.

$\hat{F}_{\mathrm{HOM}}$: PLINK

$\hat{F}_{\mathrm{Uni}}$: PLINK2, GCTA.

$\hat{F}_{\mathrm{Std}}$: PLINK1, GCTA.

$\hat{F}_{\mathrm{ROH}}$: PLINK1, BCFtools/ROH, FSuite.

$\hat{F}_{\mathrm{AS}}$: SNPRelate, hierFstat.

$\hat{F}_{\mathrm{MLE}}$: SNPRelate.

Software is available at:

BCFtools/ROH: https://samtools.github.io/bcftools/howtos/roh-calling.html

FSuite: http://genestat.cephb.fr/software/index.php/FSuite

GCTA: http://gump.qimr.edu.au/gcta

hierFstat: https://cran.r-project.org/web/packages/hierfstat/index.html

PLINK: http://pngu.mgh.harvard.edu/purcell/plink/

PLINK2: https://www.cog-genomics.org/plink/2.0/

SNPRelate: http://www.bioconductor.org/packages/release/bioc/html/SNPRelate.html

DATA ARCHIVING

The simulated data are available as an online supplement.

The 1000 Genomes data are available at ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

ACKNOWLEDGMENTS

This work was supported in part by Grants 31003A-138180 and IZKOZ3-157867 from the Swiss National Science Foundation, and by Grants GM081062 and GM075091 from the US National

Institutes of Health. We are grateful for the comments of reviewers and editors on an earlier version of this paper.

Allen N, et al. (2012). UK Biobank: Current status and what it means for epidemiology. Health Policy Technol 1:123126.

Alemu AW, Kadir NK, Harland C, Paux P, Charlier C, Caballero A, Druet T (2021). An evaluation of inbreeding measures using a whole-genome sequenced cattle pedigree. Heredity 126:410-423.

Astle W, Balding DJ (2009). Population structure and cryptic relatedness in genetic association studies. Stat Sci 24:451-471.

Ayres KL, Balding DJ (1998). Measuring departures from Hardy-Weinberg: a Markov chain Monte Carlo method for estimating the inbreeding coefficient. Heredity 80:769-777.

Ceballos FC, Joshi PK, Clark DW, Ramsay M, Wilson JF (2018). Runs of homozygosity: windows into population history and trait architecture. Nature Reviews Genetics 19: 220234.

Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience 4:7.

DeGiorgio M, Rosenberg NA (2009). An unbiased estimator of gene diversity in samples containing related individuals. Mol Biol Evol 26:501-512.

Gazal S, Sahbatou M, Perdry H, Letort S, Génin E, Leutenegger A (2014). Inbreeding coefficient estimation with dense SNP data: Comparison of strategies and application to HapMap III. Hum Hered 77:49-62.

Gazal S, Sahbatou M, Barbron M-C, Génin E, Leutenegger A (2015). High level of inbreeding in final phase of 1000 Genomes Project. Sci Reports 5:17453.

Gibson J, Morton NE, Collins A (2006). Extended tracts of homozygosity in outbred human populations, Hum Mol Genet 15:789-795.

Goudet J (2005). HIERFSTAT, a package for R to compute and test hierarchical F-statistics. Mol Ecol Notes 5:184-186.

Goudet J, Kay T, Weir BS (2018). How to estimate kinship. Mol Ecol 27:4121-4135.

Hall N, Mercer L, Phillips D, Shaw J, Anderson AD (2012). Maximum likelihood estimation of individual inbreeding coefficients and null allele frequencies. Genet Res 94:151-161.

Hill WG, Weir BS (2011). Variation in actual relationship as a consequence of Mendelian sampling and linkage. Genet Res 93:47-74.

Hedrick PW (2000). Genetics of Populations, Second Edition. Jones and Bartlett, Sudbury, MA.

Joshi PK, et al. (2015). Directional dominance on stature and cognition in diverse populations. Nature 523:459-462.

Kardos M, Nietlisbach P, Hedrick PW (2018). How should we compare different genomic estimates of the strength of inbreeding depression? Proc Natl Acad Sci USA 115:E2492-E2493.

Kelleher J, Etheridge AM, McVean G (2016). Efficient coalescent simulation and genealogical analysis for large sample sizes. PLoS Comput Biol 12: e1004842.

Li CC, Horvitz DG (1953). Some methods of estimating the inbreeding coefficient. Am J Hum Genet 5:107-117.

Malécot G (1948), The Mathematics of Heredity. Translated by Yermanos DM (1960). Freeman, San Francisco.

McPeek MS, Wu X, Ober C (2004). Best linear unbiased allele-frequency estimation in complex pedigrees. Biometrics 60:359-367.

Meyermans R, Gorssen W, Buys N, Janssens S (2020). How to study runs of homozygosity using PLINK? A guide for analyzing medium density SNP data in livestock and pet species. BMC Genomics 21:94.

Narasimhan V, Danecek P, Scally A, Xue Y, Tyler-Smith C, Durbin R (2016). BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. Bioinformatics 32:1749-1751.

National Research Council (1996). The Evaluation of Forensic DNA Evidence. National Academies Press, Washington DC.

Neuenschwander S, Michaud F, Goudet J (2019). quantiNemo 2: a Swiss knife to simulate complex demographic and genetic scenarios, forward and backward in time. Bioinformatics 35:886-888.

Ochoa A, Storey JD (2019). FST 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.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira M, Bender D, Maller J, Sklar P, de Bakker P, Daly M, Sham P (2007). Plink: a toolset for whole-genome association and population-based linkage analysis. Am J Hum Genet 81:559575.

Ritland K (1996). Estimators for pairwise relatedness and individual inbreeding coefficients. Genet Res Camb 67:175-185.

Robertson A, Hill WG (1984). Deviations from Hardy-Weinberg proportions: sampling variances and use in estimation of inbreeding coefficients. Genetics 107:703-718.

Speed D, Hemani G, Johnson MR, Balding DJ (2012). Improved heritability estimation from genome-wide SNPs. Am J Hum Genet 91:1011-1021.

Steele CD, Syndercombe Court D, Balding DJ (2014). Worldwide F_{ST} estimates relative to five continental-scale populations. Ann Hum Genet 78:468-477.

The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature 526:68-87.

VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91:4414-4423.

Vogl C, Karhu A, Moran G, Savolainene O (2002). High resolution analysis of mating systems: inbreeding in natural populations of Pinus radiata. J Evol Biol 15:433-439.

Wang J (2016). Pedigrees or markers: Which are better in estimating relatedness and inbreeding coefficient? Theoret Pop Biol 107:4-13.

Weir BS (1996). Genetic Data Analysis II. Sinauer, Sunderland, MA.

Weir BS, Cockerham CC (1984). Estimating F-statistics for the analysis of population structure. Evolution 38:1358-1370.

Weir BS, Goudet J (2017). A unified characterization for population structure and relatedness. Genetics 206:2085-2103.

Weir BS, Hill WG (2002). Estimating F-statistics. Ann Rev Genet 36:721-750.

590 Wright S (1922). Coefficients of inbreeding and relationship. Am Naturalist 56:330-338.
591
592 Yang J, Lee SH, Goddard ME, Visscher PM (2011). GCTA: A tool for genome-wide complex trait
593 analysis. Am J Hum Genet 88:76-82.
594
595 Yengo L, Zhu Z, Wray NR, Weir BS, Yang J, Robinson MR, Visscher PM (2017). Detection and
596 quantification of inbreeding depression for complex traits from SNP data. Proc Natl Acad Sci
597 USA 114:8602-8607.
598
599 Yengo L, Zhu Z, Wray NR, Weir BS, Yang J, Robinson MR, Visscher PM (2018). Estimation of
600 inbreeding depression from SNP data REPLY. Proc Natl Acad Sci USA 115:E2494-E2495.
601
602 Zheng X, Levine D, Shen J, Gogarten S, Laurie C, Weir B (2012). A high-performance computing
603 toolset for relatedness and principal component analysis of SNP data. Bioinformatics 28: 3326-
604 3328.

Table 1: Measures of Inbreeding and Coancestry.

Measure	Description	Evaluation
F_j	Inbreeding coefficient for individual j: ibd probability for homologous alleles	F_{PED}: Path counting. F_{Gold}: Actual ibd in simulations.
$\theta_{jj'}$	Coancestry for individuals j, j': ibd probability for random alleles from j and j'.	F_{PED}: Path counting. F_{Gold} Actual ibd in simulations.
The following hold for PED and Gold values.		
F_W	Average inbreeding coefficient.	$F_W = \frac{1}{n}\sum_{j=1}^{n} F_j$ for n individuals.
Ψ_j	Average coancestry coefficient for individual j.	$\Psi_j = \frac{1}{n-1}\sum_{j'=1 j'\neq j}^{n} \theta_{jj'}$
θ_S	Average coancestry coefficient.	$\theta_S = \frac{1}{n}\sum_{j=1}^{n} \Psi_j$
f_j	Within-population inbreeding coefficient for individual j.	$f_j = \frac{F_j - \theta_S}{1-\theta_S}$
f_W	Average within-population inbreeding coefficient.	$f_W = \frac{F_W - \theta_S}{1-\theta_S}$
ψ_j	Within-population average kinship coefficient for individual j.	$\psi_j = \frac{\Psi_j - \theta_S}{1-\theta_S}$

Table 2: Estimators of Inbreeding.

Estimate	Calculation*	Expected Value†
$\hat{F}_{\mathrm{ROH}_j}$	Proportion of homozygous blocks.	No explicit expression.
$\hat{F}_{\mathrm{MLE}_j}$	Maximization of likelihood for f_j.	No explicit expression.
$\hat{F}_{\mathrm{HOM}_j}$	$1 - \frac{\sum_l X_{jl}(2 - X_{jl})}{\sum_l 2\tilde{p}_l(1-\tilde{p}_l)}$	$\frac{f_j - \frac{1}{2n}(1+f_W)}{1 - \frac{1}{2n}(1+f_W)}$
$\hat{F}_{\mathrm{HOM}_W}$	$1 - \frac{1}{n}\sum_{j=1}^{n} \frac{\sum_l X_{jl}(2 - X_{jl})}{\sum_l 2\tilde{p}_l(1-\tilde{p}_l)}$	$\frac{f_W - \frac{1}{2n}(1+f_W)}{1 - \frac{1}{2n}(1+f_W)}$
$\hat{F}_{\mathrm{AS}_j}$	$\frac{\sum_l (\tilde{A}_{jl} - \tilde{A}_{Sl})}{\sum_l (1 - \tilde{A}_{Sl})}$	f_j
$\hat{F}_{\mathrm{AS}_W}$	$\frac{1}{n}\sum_{j=1}^{n} \hat{F}_{\mathrm{AS}_j}$	f_W
$\hat{F}^w_{\mathrm{UNI}_j}$	$\frac{\sum_l [X_{jl}^2 - (1+2\tilde{p}_l)X_{jl} + 2\tilde{p}_l^2]}{\sum_l 2\tilde{p}_l(1-\tilde{p}_l)}$	$\frac{f_j - 2\psi_j - \frac{1}{2n}(3 + 4f_j - 8\psi_j - f_W)}{1 - \frac{1}{2n}(1+f_W)}$
$\hat{F}^w_{\mathrm{UNI}_W}$	$\frac{1}{n}\sum_{j=1}^{n} \hat{F}^w_{\mathrm{UNI}_j}$	$\frac{f_W - \frac{3}{2n}(1+f_W)}{1 - \frac{1}{2n}(1+f_W)}$
$\hat{F}^u_{\mathrm{UNI}_j}$	$\frac{1}{L}\sum_{l=1}^{L} \frac{X_{jl}^2 - (1+2\tilde{p}_l)X_{jl} + 2\tilde{p}_l^2}{2\tilde{p}_l(1-\tilde{p}_l)}$	No explicit expression.
$\hat{F}^w_{\mathrm{STD}_j}$	$\frac{\sum_l (X_{jl} - 2\tilde{p}_l)^2}{\sum_l 2\tilde{p}_l(1-\tilde{p}_l)} - 1$	$\frac{f_j - 4\psi_j - \frac{1}{2n}(3 + 4f_j - 8\psi_j - f_W)}{1 - \frac{1}{2n}(1+f_W)}$
$\hat{F}^w_{\mathrm{STD}_W}$	$\frac{1}{n}\sum_{j=1}^{n} \hat{F}^w_{\mathrm{STD}_j}$	$\frac{f_W - \frac{3}{2n}(1+f_W)}{1 - \frac{1}{2n}(1+f_W)}$
$\hat{F}^u_{\mathrm{STD}_j}$	$\frac{1}{L}\sum_{l=1}^{L} \frac{(X_{jl} - 2\tilde{p}_l)^2}{2\tilde{p}_l(1-\tilde{p}_l)} - 1$	No explicit expression.

* X_{jl} is the reference allele dosage for SNP l in individual j.

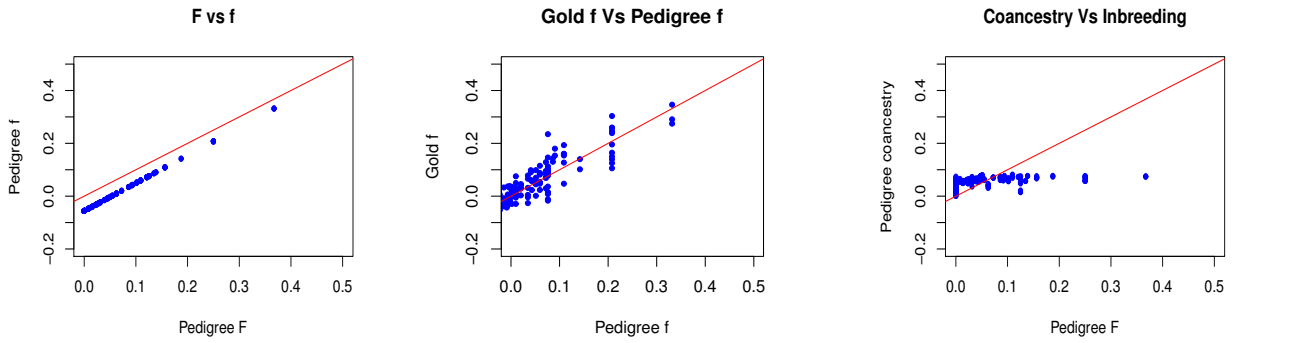
* $\tilde{p}_l = \frac{1}{2n}\sum_{j=1}^{n} X_{jl}$ is the sample allele frequency for SNP l.

† For weighted averages over large numbers of loci.

Table 3: Correlations among inbreeding measures* for simulated data.

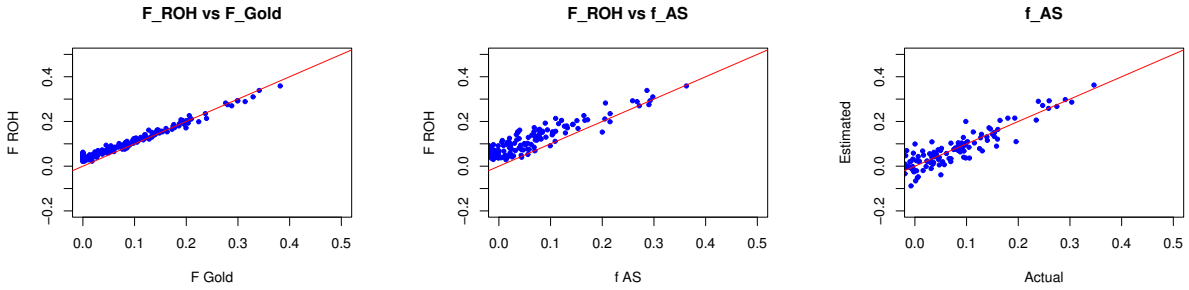
	F_{PED}	F_{Gold}	$\hat{F}_{\text{ROH}}$	f_{PED}	f_{Gold}	$\hat{f}_{\text{AS}}$	$\hat{f}_{\text{HOM}}$	$\hat{f}_{\text{MLE}}$	$f^{\text{Gold}}_{\text{UNI}}$	$\hat{f}^{\text{w}}_{\text{UNI}}$	$\hat{f}^{\text{u}}_{\text{UNI}}$	$f^{\text{Gold}}_{\text{STD}}$	$\hat{f}^{\text{w}}_{\text{STD}}$	$\hat{f}^{\text{u}}_{\text{STD}}$
F_{PED}	1.00	0.94	0.92	1.00	0.94	0.84	0.84	0.80	0.80	0.71	0.74	0.44	0.36	−0.25
F_{Gold}	0.94	1.00	0.99	0.94	1.00	0.90	0.90	0.88	0.86	0.78	0.80	0.48	0.41	−0.24
$\hat{F}_{\text{ROH}}$	0.92	0.99	1.00	0.92	0.99	0.91	0.91	0.89	0.87	0.80	0.82	0.50	0.45	−0.20
f_{PED}	1.00	0.94	0.92	1.00	0.94	0.84	0.84	0.80	0.80	0.71	0.74	0.44	0.36	−0.25
f_{Gold}	0.94	1.00	0.99	0.94	1.00	0.90	0.90	0.88	0.86	0.78	0.80	0.48	0.41	−0.24
$\hat{f}_{\text{AS}}$	0.84	0.90	0.91	0.84	0.90	1.00	1.00	0.99	0.77	0.86	0.86	0.42	0.44	−0.22
$\hat{f}_{\text{HOM}}$	0.84	0.90	0.91	0.84	0.90	1.00	1.00	0.99	0.77	0.86	0.86	0.42	0.44	−0.22
$\hat{f}_{\text{MLE}}$	0.80	0.88	0.89	0.80	0.88	0.99	0.99	1.00	0.82	0.92	0.91	0.53	0.57	−0.10
$f^{\text{Gold}}_{\text{UNI}}$	0.80	0.86	0.87	0.80	0.86	0.77	0.77	0.82	1.00	0.89	0.91	0.86	0.74	0.18
$\hat{f}^{\text{w}}_{\text{UNI}}$	0.71	0.78	0.80	0.71	0.78	0.86	0.86	0.92	0.89	1.00	0.98	0.75	0.84	0.17
$\hat{f}^{\text{u}}_{\text{UNI}}$	0.74	0.80	0.82	0.74	0.80	0.86	0.86	0.91	0.91	0.98	1.00	0.76	0.80	0.17
$f^{\text{Gold}}_{\text{STD}}$	0.44	0.48	0.50	0.44	0.48	0.42	0.42	0.53	0.86	0.75	0.76	1.00	0.87	0.55
$\hat{f}^{\text{w}}_{\text{STD}}$	0.36	0.41	0.45	0.36	0.41	0.44	0.44	0.57	0.74	0.84	0.80	0.87	1.00	0.53
$\hat{f}^{\text{u}}_{\text{STD}}$	−0.25	−0.24	−0.20	−0.25	−0.24	−0.22	−0.22	−0.10	0.18	0.17	0.17	0.55	0.53	1.00

* As shown in Tables 1 and 2.



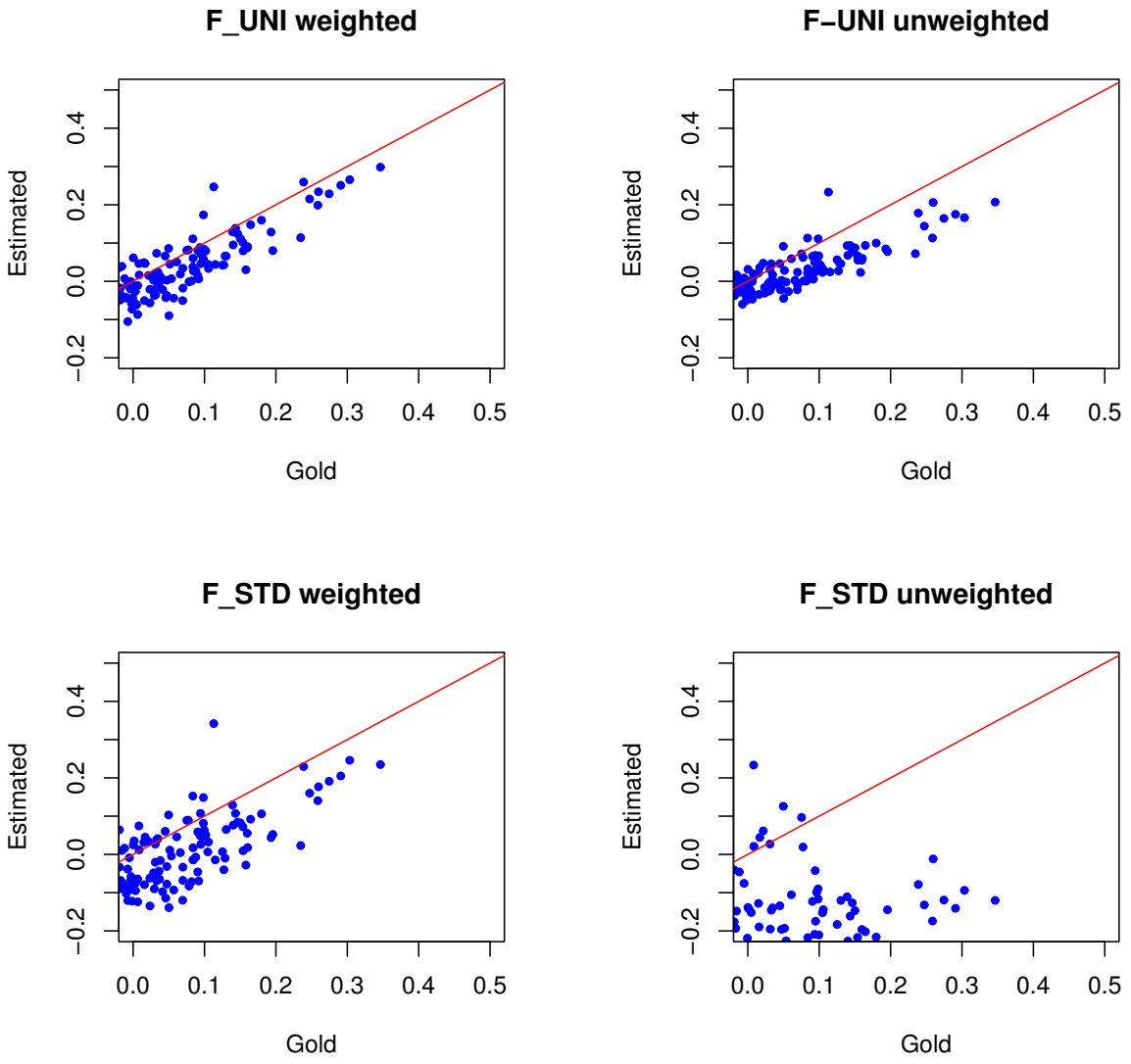
Pedigree f vs Pedigree F Gold f vs Pedigree f Pedigree coancestry vs Pedigree f

Figure 1: Allele sharing estimates for 283 non-founders in simulated pedigree.



$\hat{F}_{\text{ROH}}$ vs F_{Gold} $\hat{F}_{\text{ROH}}$ vs $\hat{f}_{\text{AS}}$ $\hat{f}_{\text{AS}}$ vs f_{Gold}

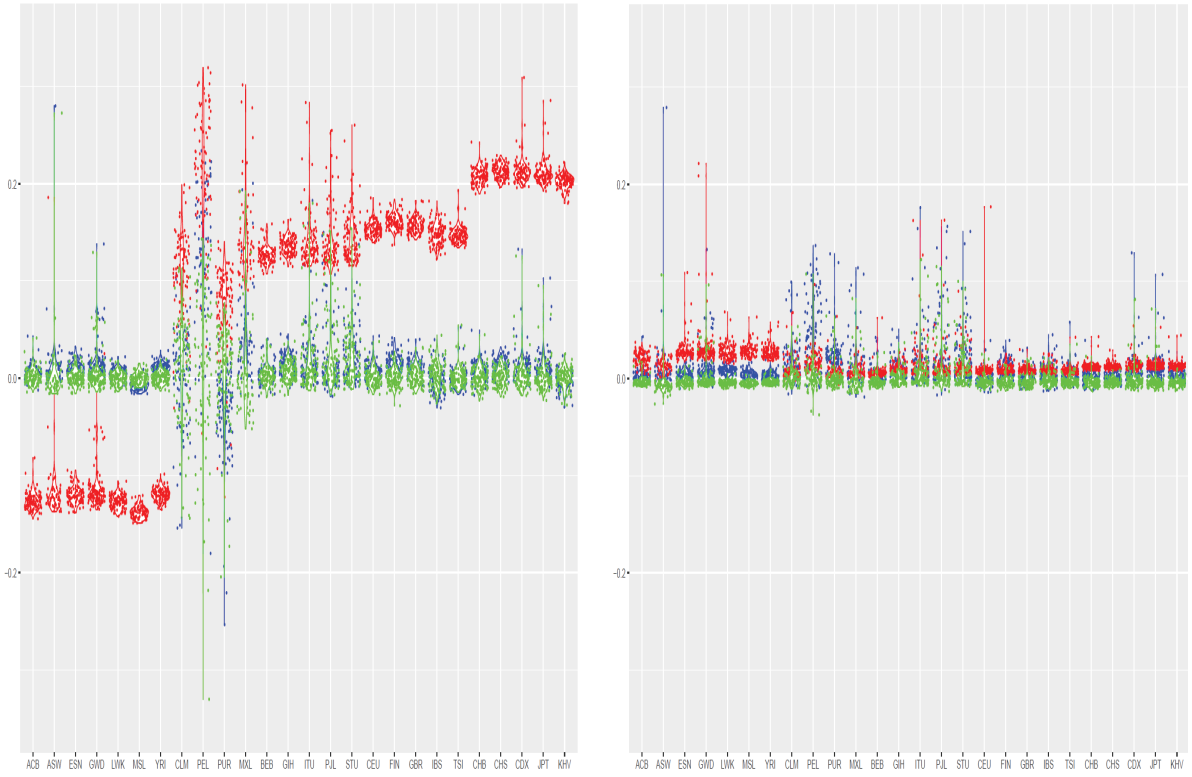
Figure 2: Values of ROH estimates of F and allele-sharing estimates of f for 283 non-founders in simulated pedigree.



Weighted $\hat{f}_{\text{UNI}}$ and $\hat{f}_{\text{STD}}$ vs f_{Gold_j}

Unweighted $\hat{f}_{\text{UNI}}$ and $\hat{f}_{\text{STD}}$ vs f_{Gold_j}

Figure 3: Values of UNI and STD estimates for 283 non-founders in simulated pedigree.



$\hat{f}^{w}_{\mathrm{AS}}$; $\hat{f}^{u}_{\mathrm{UNI}}$

Figure 4: Individual inbreeding coefficient estimates for 1000 Genomes data. Green: Population as reference; Blue: Continental group as reference; Red: World as reference. Populations (left to right): ACB, ASW, ESN, GWD, LWK, MSL, YRI (AFR); CLM, MXL, PEL, PUR (AMR); CDX, CHB, CHS, JPT, KHV (EAS); CEU, FIN, GBR, IBS, TSI (EUR); BEB, GIH, ITU, PJL, STR (SAS).

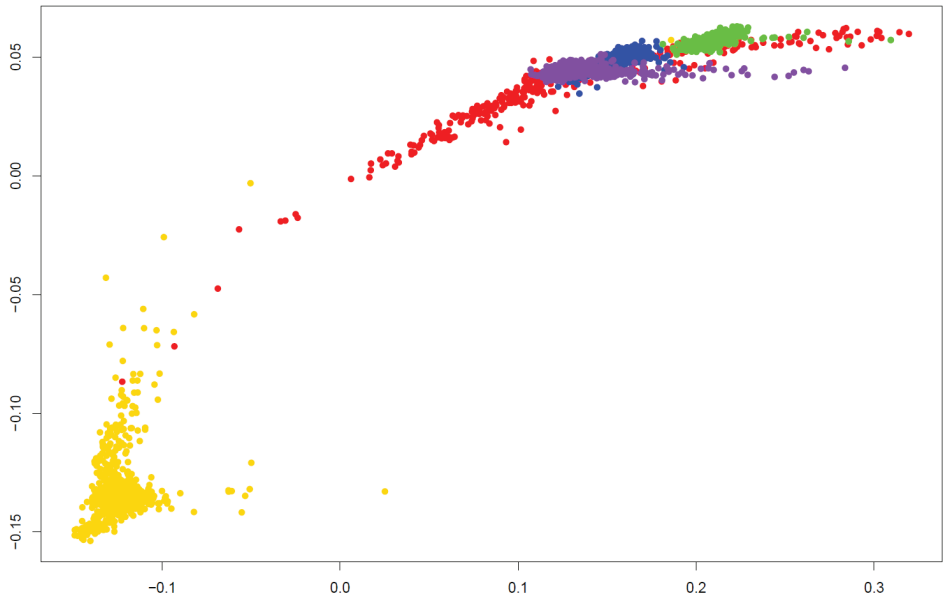


Figure 5: Estimates $\hat{\psi}_{\mathrm{AS}_j}$ of within-population individual-specific average kinships (Y-axis) vs estimates $\hat{f}_{\mathrm{AS}_j}$ of within-population individual-specific inbreeding coefficients (X-axis) for 1000 Genomes data, with the World as reference set. Gold: AFR; Red: AMR; Purple: SAS; Blue: EUR; Green: EAS.

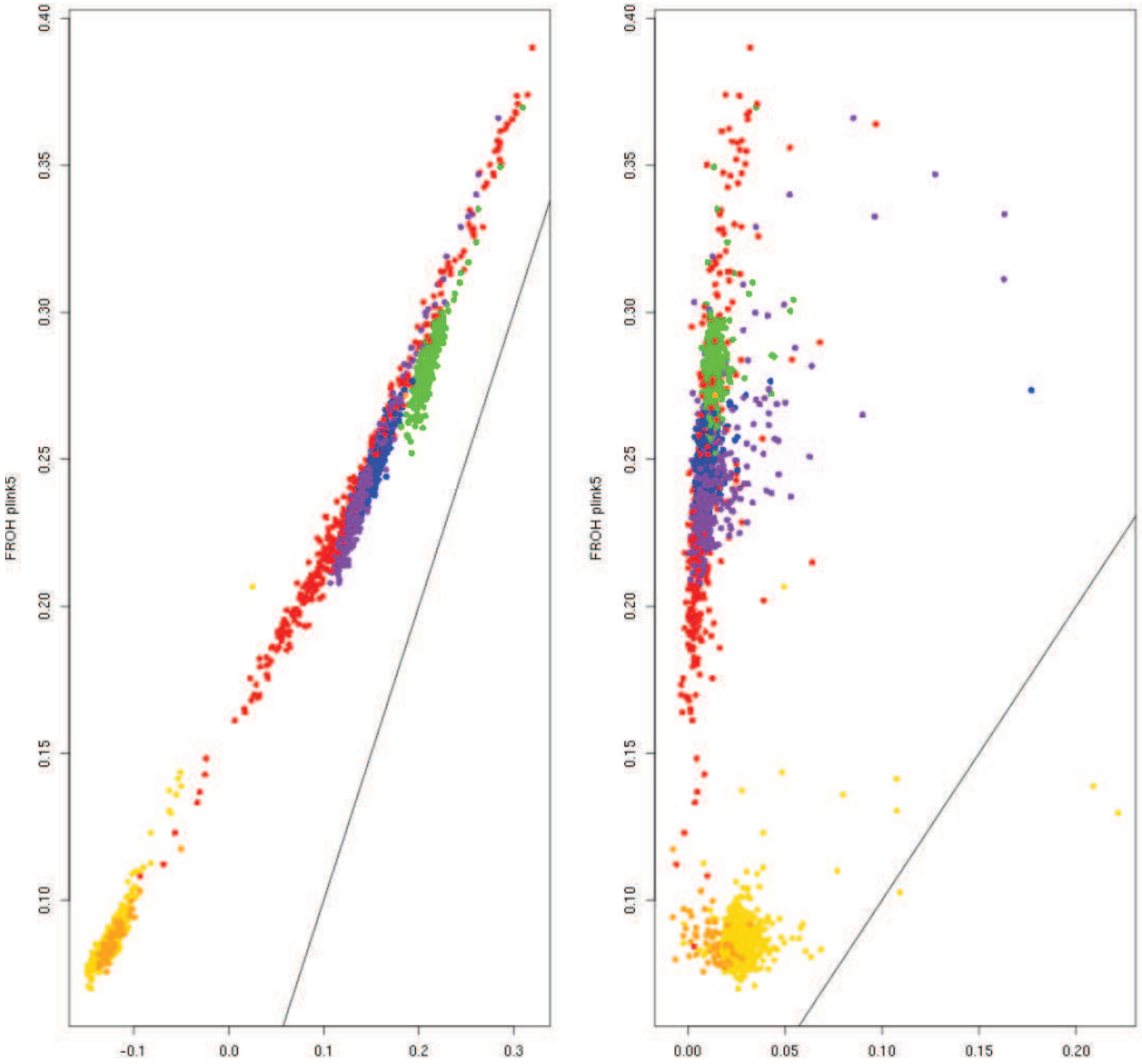


Figure 6: PLINK-estimates $\hat{f}_{\text{ROH}}$ (Y-axis) vs SNP by SNP estimates for 1000 Genomes data, with the World as a reference set. Left Panel: $\hat{f}^{w}_{\text{AS}}$ (X-axis) ; Right panel: $\hat{f}^{u}_{\text{UNI}}$ (X-axis). Solid line $X = Y$ in both panels. Gold: AFR, not ACB, ASW; Orange: AFR, ACB and ASW; Red: AMR; Purple: SAS; Blue: EUR; Green: EAS.

Supplementary Figure S1: 1000 Genomes Estimates

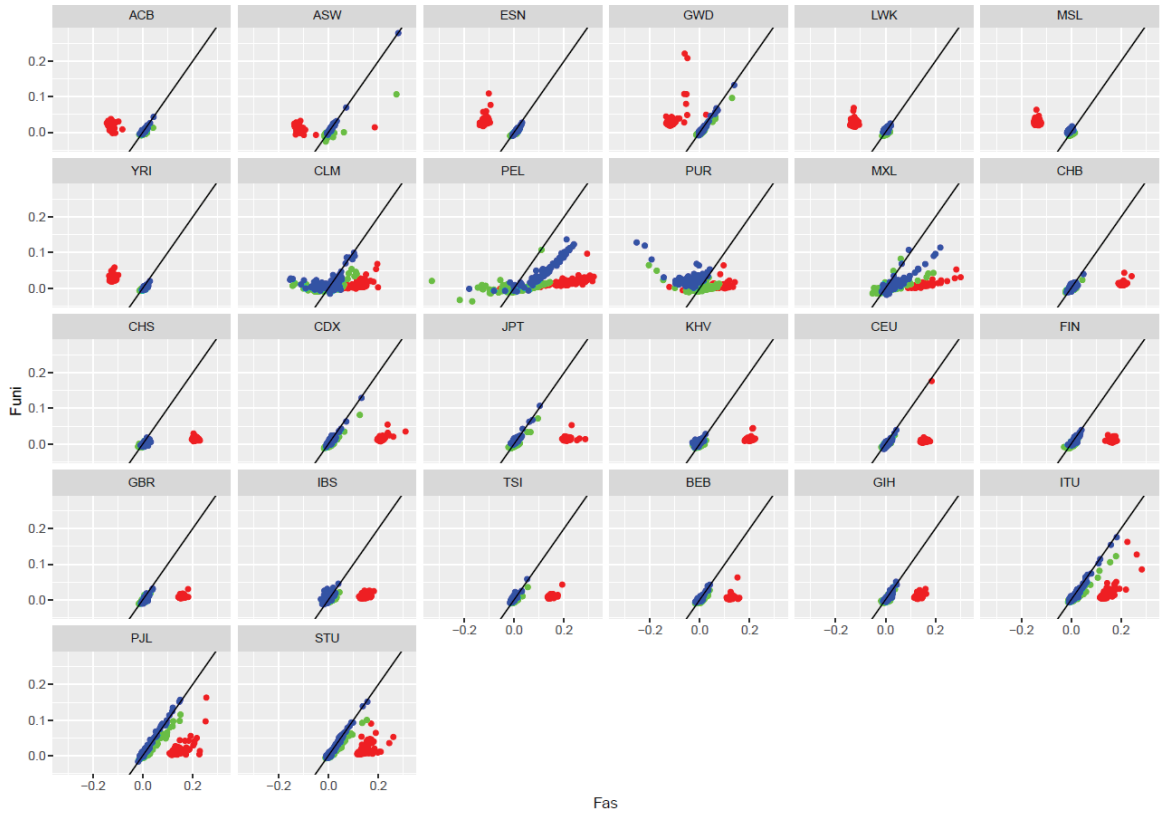


Figure S1: Values of $\hat{f}_{\text{UNI}}$ (Y-axis) versus $\hat{f}_{\text{AS}}$ (X-axis) for the 1000 Genomes populations. Population reference in green, continental reference in blue, world reference in red.

Supplementary Figure S2: 1000 Genomes Estimate Ranks

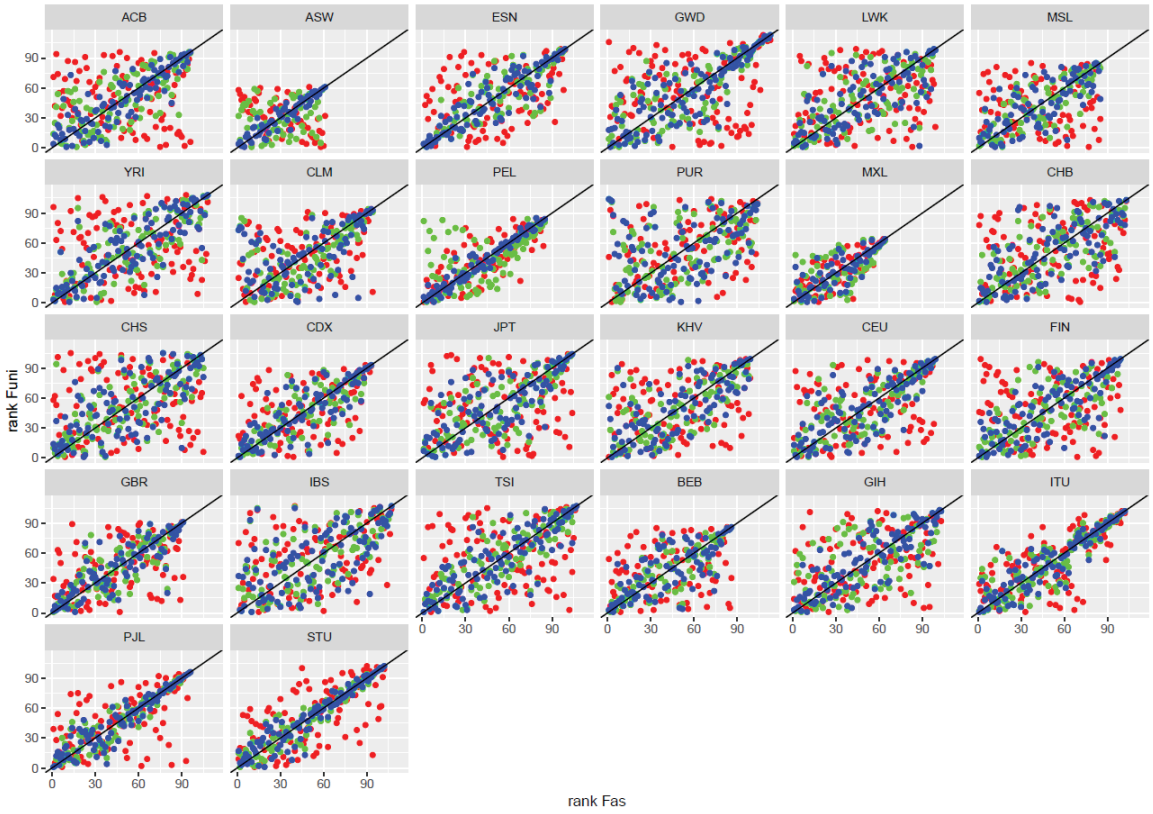


Figure S2: Ranks of $\hat{f}_{\text{UNI}}$ (Y-axis) versus ranks of $\hat{f}_{\text{AS}}$ (X-axis) for the 1000 Genomes populations. Population reference in green, continental reference in blue, world reference in red.

Supplementary Figure S3:

Use of Continent vs World Allele Frequencies.

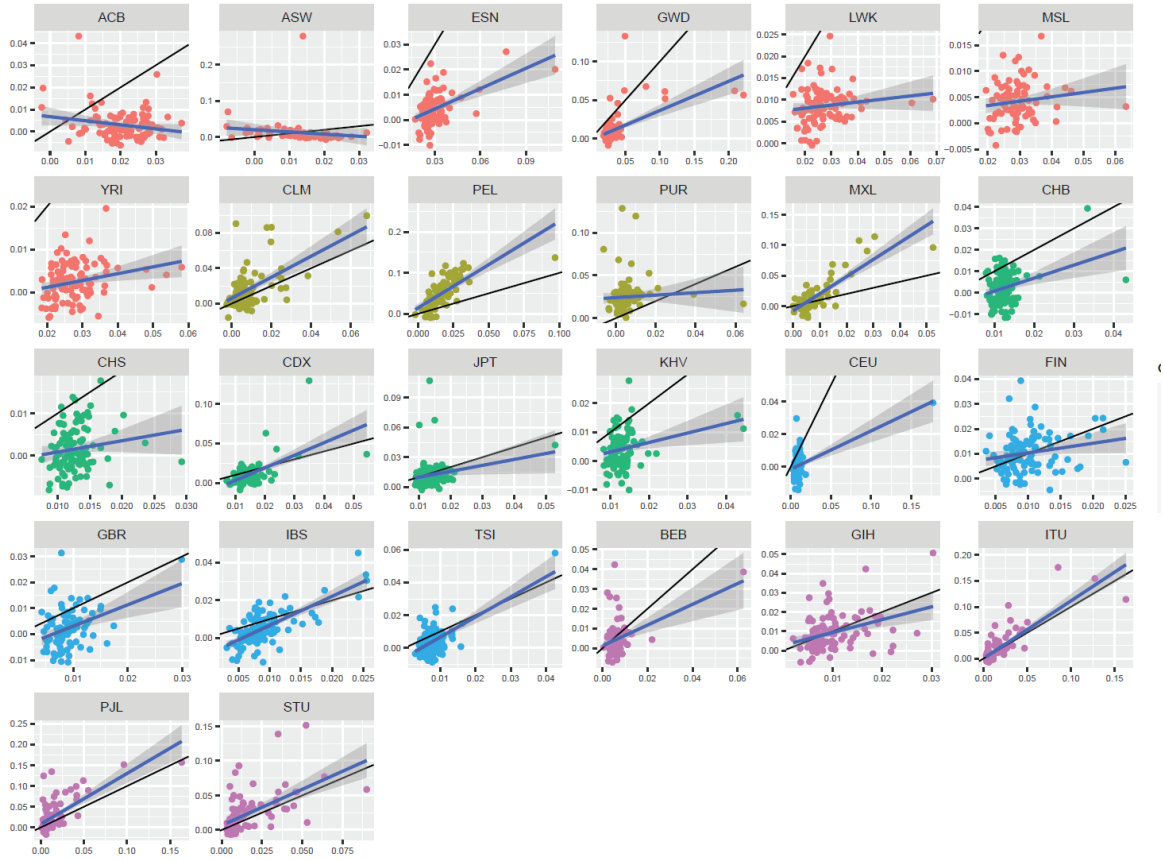


Figure S3: Values of $\hat{f}_{\text{UNI}}$ for each of the 1000 Genomes populations with the continent for that population providing the sample allele frequencies (Y axis) versus the world providing the sample allele frequencies (X axis). Red: AFR; Gold: AMR; Green: SAS; Blue: EUR; Purple: SAS.

Supplementary Figure S4:

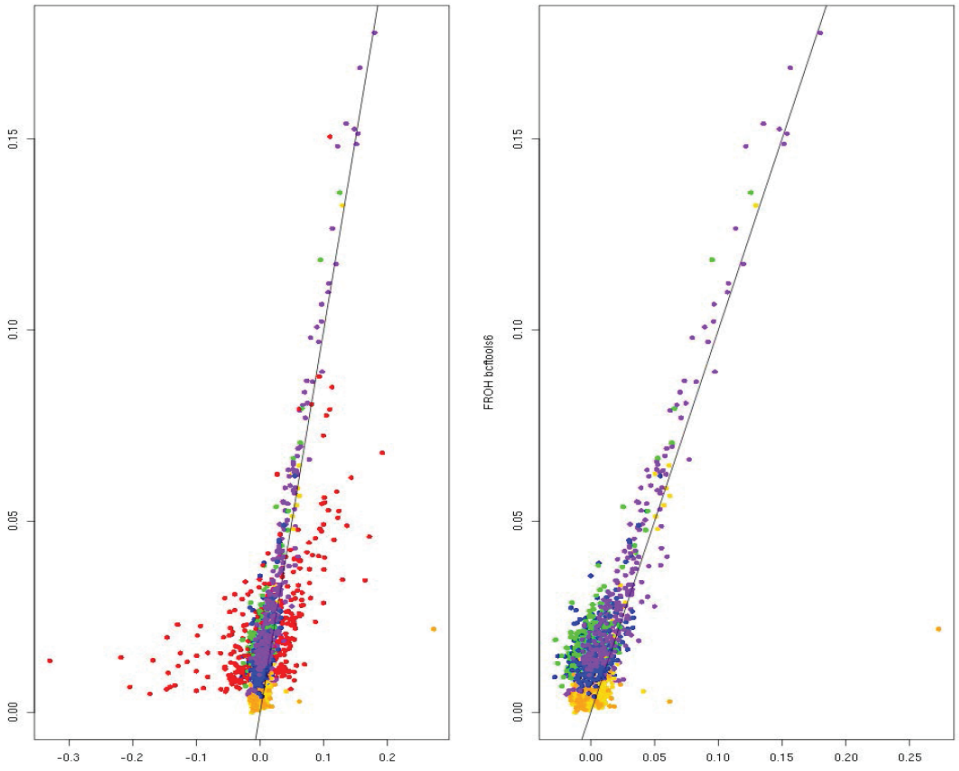


Figure S4: BCF-tools-estimates $\hat{f}_{\text{BCF}}$ (Y-axis) vs $\hat{f}^{w}_{\text{AS}}$ (X-axis) for 1000 Genomes data, with the Population as a reference set. Left Panel: All 1000 Genomes populations; Right panel: Omitting AMR populations. Solid line $X = Y$ in both panels. Gold: AFR, not ACB, ASW; Orange: AFR, ACB and ASW; Red: AMR; Purple: SAS; Blue: EUR; Green: EAS.

Supplementary Figure S5:

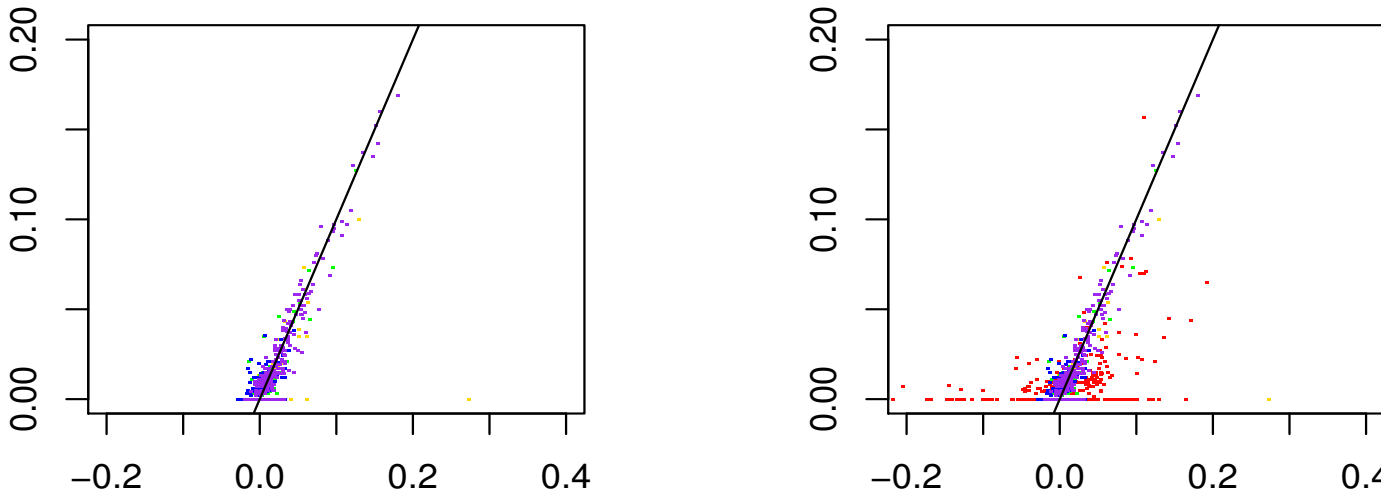


Figure S5: Fsuite estimates $\hat{f}_{\text{FSuite}}$ (Y-axis) vs $\hat{f}^{w}_{\text{AS}}$ (X-axis) for 1000 Genomes data, with the Population as a reference set. Left Panel: All 1000 Genomes populations; Right panel: Omitting AMR populations. Solid line $X = Y$ in both panels. Gold: AFR; Red: AMR; Purple: SAS; Blue: EUR; Green: EAS.