G3 Genes | Genomes | Genetics

INVESTIGATION

An Unbiased Estimator of Gene Diversity with Improved Variance for Samples Containing Related and Inbred Individuals of any Ploidy

Alexandre M. Harris*,† and Michael DeGiorgio*,‡,1

*Department of Biology, †Molecular, Cellular, and Integrative Biosciences at the Huck Institutes of the Life Sciences, and ‡Institute for CyberScience, Pennsylvania State University, University Park, Pennsylvania 16802

ABSTRACT Gene diversity, or expected heterozygosity (H), is a common statistic for assessing genetic variation within populations. Estimation of this statistic decreases in accuracy and precision when individuals are related or inbred, due to increased dependence among allele copies in the sample. The original unbiased estimator of expected heterozygosity underestimates true population diversity in samples containing relatives, as it only accounts for sample size. More recently, a general unbiased estimator of expected heterozygosity was developed that explicitly accounts for related and inbred individuals in samples. Though unbiased, this estimator's variance is greater than that of the original estimator. To address this issue, we introduce a general unbiased estimator of gene diversity for samples containing related or inbred individuals, which employs the best linear unbiased estimator of allele frequencies, rather than the commonly used sample proportion. We examine the properties of this estimator, $\widetilde{H}_{\text{BLUE}}$, relative to alternative estimators using simulations and theoretical predictions, and show that it predominantly has the smallest mean squared error relative to others. Further, we empirically assess the performance of $\widetilde{H}_{\text{BLUE}}$ on a global human microsatellite dataset of 5795 individuals, from 267 populations, genotyped at 645 loci. Additionally, we show that the improved variance of $\widetilde{H}_{\text{BLUE}}$ leads to improved estimates of the population differentiation statistic, F_{ST}, which employs measures of gene diversity within its calculation. Finally, we provide an R script, *BestHet*, to compute this estimator from genomic and pedigree data.

KEYWORDS
expected heterozygosity
identity state
inbreeding
locus-specific branch length
relatedness

The gene diversity of a locus, also known as its expected heterozygosity (H), is a fundamental measure of genetic variation in a population, and describes the proportion of heterozygous genotypes expected under Hardy-Weinberg equilibrium (Nei 1973). Formally, gene diversity is the probability that a pair of randomly sampled allele copies from a population are different, and is computed as

$$H = 1 - \sum_{i=1}^{I} p_i^2, \tag{1}$$

where I is the number of distinct alleles at a locus, and p_i ($i = 1, 2, \ldots, I$) is the frequency of allele i in the population.

For a sample without related or inbred individuals composed of n allele copies, an unbiased estimator of expected heterozygosity is (Nei and Roychoudhury 1974)

$$\widehat{H} = \frac{n}{n-1}\left(1 - \sum_{i=1}^{I} \widehat{p}_i^2\right), \tag{2}$$

where $\widehat{p}_i$ is the sample proportion of allele i. $\widehat{H}$ is a biased estimator when inbred or related individuals are included in the sample (DeGiorgio and Rosenberg 2009). This result is based on the idea that, as the proportion of related individuals in the sample increases, the number of independent allele observations decreases.

When two alleles are drawn from a sample, one each from a pair of related individuals, there is a nonzero probability that they will be identical by descent (IBD), rather than just identical by state (Lange 2002). This IBD probability is known as the kinship coefficient, and is denoted by Φ_{jk} for a pair of individuals j and k. Thus, the observed diversity will be lower than the true value because a greater proportion

doi: 10.1534/g3.116.037168
Manuscript received September 20, 2016; accepted for publication December 18, 2016; published Early Online December 30, 2016.

Supplemental material is available online at www.g3journal.org/lookup/suppl/doi:10.1534/g3.116.037168/-/DC1.
[1]Corresponding author: 502B Wartik Laboratory, Pennsylvania State University, University Park, PA 16802. E-mail: mxd60@psu.edu

of identical alleles are observed than for a sample in which there are no related individuals. DeGiorgio *et al.* (2010) developed an estimator of expected heterozygosity,

$$\widetilde{H} = \frac{1}{1-\bar{\Phi}_2}\left(1 - \sum_{i=1}^{I} \widehat{p}_i^2\right), \tag{3}$$

which is unbiased for samples containing related and inbred individuals of any ploidy, and employs a weighted mean kinship coefficient $\bar{\Phi}_2$ as a bias correction factor. $\bar{\Phi}_2$ is the average of all kinship coefficients Φ_{jk} for every pair of individuals within the sample (see *Methods*). Further, DeGiorgio *et al.* (2010) derived the theoretical variance of $\widetilde{H}$, as well as its approximate value for samples wherein individuals are related to no more than one other sampled individual.

As an alternative to the sample proportion ($\widehat{p}_i$), McPeek *et al.* (2004) introduced the best linear unbiased estimator (BLUE, denoted as $\widetilde{p}_i$) of population allele frequency, which is an unbiased linear estimator with smaller variance than the unbiased linear estimator $\widehat{p}_i$. The BLUE incorporates the relatedness of individuals in the sample as a covariance matrix to define the weight of each observation. Simulations and analytical evaluation corroborating their result suggest that the mean squared error (MSE) of $\widetilde{p}_i$ is always smaller than that of $\widehat{p}_i$, and this difference is especially evident for samples with complex pedigrees.

Because $\widetilde{p}_i$ has the smallest variance of any unbiased linear estimator of allele frequencies, we expect its low variance to translate to smaller variance of gene diversity statistics that use $\widetilde{p}_i$. We developed such a statistic, termed $\widetilde{H}_{\text{BLUE}}$, that is an unbiased estimator of expected heterozygosity in samples containing related and inbred individuals of arbitrary ploidy. Through simulations, analytical predictions, and empirical assessments, we compare the performance of $\widetilde{H}_{\text{BLUE}}$ to that of $\widetilde{H}$ and $\widehat{H}$ for samples containing related individuals of various types across different ploidy and inbreeding status. Additionally, we derive the variance of any measure of expected heterozygosity that uses unbiased linear estimators of allele frequencies. We find that the increased precision of allele frequency estimates transfers to our unbiased estimator, yielding values for MSE invariably equal to or smaller than those of $\widetilde{H}$, while occasionally exceeding the precision of $\widehat{H}$. The improved properties of $\widetilde{H}_{\text{BLUE}}$ translate to its applications as well, which we demonstrate in the calculation of the population differentiation statistic, F_{ST} (Wright 1951). F_{ST} can be written in terms of intrapopulation and interpopulation gene diversity as (Hudson *et al.* 1992)

$$F_{\text{ST}} = \frac{H_{12} - \frac{1}{2}(H_1 + H_2)}{H_{12}}, \tag{4}$$

where H_1 and H_2 are the values of expected heterozygosity within each of two compared populations, and H_{12} is the expected heterozygosity between them.

METHODS

Consider a locus with I distinct alleles in a sample of n individuals. Let $X_k^{(i)}$ denote the fraction of alleles at the locus in individual k that are of type i, $i = 1, 2, \ldots, I$. An unbiased linear estimator of population allele frequencies p_i, denoted by $\breve{p}_i$, is defined as

$$\breve{p}_i = \sum_{k=1}^{n} w_k X_k^{(i)}, \tag{5}$$

where w_k, $0 \leq w_k \leq 1$, is the weight of individual k, $k = 1, 2, \ldots, n$, and $\sum_{k=1}^{n} w_k = 1$. Formally, we have that

$$X_k^{(i)} = \frac{1}{m_k}\sum_{t=1}^{m_k} A_{kt}^{(i)},$$

where $A_{kt}^{(i)}$ is an indicator random variable whose value is 1 if allele t of individual k is of type i, and zero otherwise, and where m_k is the ploidy of individual k. As an example, if individual k were diploid at the locus, then $m_k = 2$. Taking the expectation of $\breve{p}_i$,

$$\begin{aligned}\mathbb{E}\left[\breve{p}_i\right] &= \sum_{k=1}^{n} \frac{w_k}{m_k}\sum_{t=1}^{m_k} \mathbb{E}\left[A_{kt}^{(i)}\right] \\ &= \sum_{k=1}^{n} \frac{w_k}{m_k}\sum_{t=1}^{m_k} p_i \\ &= p_i,\end{aligned}$$

shows that it is an unbiased estimator of p_i.

Unbiased estimation of gene diversity using unbiased linear estimators of allele frequencies

In this section, we construct an unbiased estimator, $\breve{H}$, of expected heterozygosity that uses a general unbiased linear estimator, $\breve{p}_i$, of allele frequency p_i (Proposition 1). We then show that the unbiased estimator, $\widetilde{H}$, of DeGiorgio *et al.* (2010) follows as a corollary, assuming that $\breve{p}_i = \widehat{p}_i$, the sample proportion allele frequency estimator (Corollary 2). We then derive a new estimator, $\widetilde{H}_{\text{BLUE}}$, also as a corollary, assuming that $\breve{p}_i = \widetilde{p}_i$, the BLUE of allele frequency (Corollary 3).

Proposition 1: Consider a locus with I distinct alleles and parametric allele frequencies $p_i \in [0, 1]$, $i = 1, 2, \ldots, I$, and $\sum_{i=1}^{I} p_i = 1$. For a sample of size n individuals of any ploidy, inbreeding status, and relatedness,

$$\breve{H} = \frac{1}{1-\rho_2}\left(1 - \sum_{i=1}^{I} \breve{p}_i^2\right) \tag{6}$$

is an unbiased estimator of expected heterozygosity, where

$$\rho_2 = \sum_{j=1}^{n}\sum_{k=1}^{n} w_j w_k \Phi_{jk}$$

is a weighted mean kinship coefficient of the sample for all pairs of individuals in the sample, and where w_k, $k = 1, 2, \ldots, n$, is the weight for individual k. The proof of Proposition 1 is found in the Appendix.

From $\breve{p}_i$, the sample proportion estimator $\widehat{p}_i$ of allele frequency i, $i = 1, 2, \ldots, I$, is recovered when $w_k = m_k / \sum_{j=1}^{n} m_j$ for individual k, $k = 1, 2, \ldots, n$, leading to

$$\widehat{p}_i = \sum_{k=1}^{n} \frac{m_k}{\sum_{j=1}^{n} m_j} X_k^{(i)}.$$

Here, each individual is weighted by its contribution to the number of allele copies in the sample.

Corollary 2: Consider a locus with I distinct alleles and parametric allele frequencies $p_i \in [0, 1]$, $i = 1, 2, \ldots, I$, and $\sum_{i=1}^{I} p_i = 1$. For a sample of size n individuals of any ploidy, inbreeding status, and relatedness,

$$\widetilde{H} = \frac{1}{1-\bar{\Phi}_2}\left(1 - \sum_{i=1}^{I} \widehat{p}_i^2\right) \tag{7}$$

is an unbiased estimator of expected heterozygosity, where

$$\widehat{p}_i = \sum_{k=1}^{n} \frac{m_k}{\sum_{j=1}^{n} m_j} X_k^{(i)}$$

is the sample proportion estimator of allele frequency i, where

$$\bar{\Phi}_2 = \sum_{j=1}^{n} \sum_{k=1}^{n} \frac{m_j}{\sum_{x=1}^{n} m_x} \frac{m_k}{\sum_{y=1}^{n} m_y} \Phi_{jk}$$

is a weighted mean kinship coefficient of the sample for all pairs of individuals, and where $m_k, k = 1, 2, \ldots, n$, is the ploidy for individual k. The proof of Corollary 2 is found in the Appendix.

It may be beneficial to apply an unbiased linear estimator of allele frequencies that has minimum variance. McPeek *et al.* (2004) introduced the BLUE of allele frequencies, which we formally define here. We will use the BLUE of allele frequencies to construct a new unbiased estimator of gene diversity that would ideally have improved variance over other estimators. Let $\mathbf{K}$ be an $n \times n$ symmetric matrix of kinship coefficients, with $\mathbf{K}_{jk} = \Phi_{jk}$. The BLUE ($\widetilde{p}_i$) of allele frequency is obtained when $w_k = \frac{\sum_{j=1}^{n} (\mathbf{K}^{-1})_{jk}}{\mathbf{1}^T \mathbf{K}^{-1} \mathbf{1}}$, yielding

$$\widetilde{p}_i = \sum_{k=1}^{n} \frac{\sum_{j=1}^{n} (\mathbf{K}^{-1})_{jk}}{\mathbf{1}^T \mathbf{K}^{-1} \mathbf{1}} X_k^{(i)},$$

where $\mathbf{K}^{-1}$ denotes the inverse matrix of $\mathbf{K}$, $\mathbf{1}$ is a column vector of n elements with all entries equal to 1, and $\mathbf{1}^T$ is the transpose of $\mathbf{1}$.

Corollary 3: Consider a locus with I distinct alleles, and parametric allele frequencies $p_i \in [0, 1]$, $i = 1, 2, \ldots, I$, and $\sum_{i=1}^{I} p_i = 1$. For a sample of size n individuals of any ploidy, inbreeding status, and relatedness,

$$\widetilde{H}_{\mathrm{BLUE}} = \frac{1}{1 - \kappa_2} \left(1 - \sum_{i=1}^{I} \widetilde{p}_i^2 \right) \tag{8}$$

is an unbiased estimator of expected heterozygosity, where

$$\widetilde{p}_i = \sum_{k=1}^{n} \frac{\sum_{j=1}^{n} (\mathbf{K}^{-1})_{jk}}{\mathbf{1}^T \mathbf{K}^{-1} \mathbf{1}} X_k^{(i)}$$

is the BLUE of allele frequencies, and where

$$\kappa_2 = \sum_{j=1}^{n} \sum_{k=1}^{n} \frac{\sum_{x=1}^{n} (\mathbf{K}^{-1})_{xj}}{\mathbf{1}^T \mathbf{K}^{-1} \mathbf{1}} \frac{\sum_{y=1}^{n} (\mathbf{K}^{-1})_{yk}}{\mathbf{1}^T \mathbf{K}^{-1} \mathbf{1}} \Phi_{jk}$$

is a weighted mean kinship coefficient of the sample for all pairs of individuals. The proof of Corollary 3 is found in the Appendix.

Variance of H estimators using unbiased linear estimators of allele frequencies

We now derive the equation (Proposition 4) describing the variance of the unbiased estimator $\breve{H}$, which takes $\breve{p}_i$ as the unbiased linear estimate of population allele frequency p_i. This value depends on the weighted mean kinship coefficients of the sample for all pairs, trios, quartets, and pairs of pairs of individuals in the sample, defined as

$$\rho_2 = \sum_{j=1}^{n} \sum_{k=1}^{n} w_j w_k \Phi_{jk}$$

$$\rho_3 = \sum_{j=1}^{n} \sum_{k=1}^{n} \sum_{j'=1}^{n} w_j w_k w_{j'} \Phi_{jkj'}$$

$$\rho_4 = \sum_{j=1}^{n} \sum_{k=1}^{n} \sum_{j'=1}^{n} \sum_{k'=1}^{n} w_j w_k w_j w_{k'} \Phi_{jkj'k'}$$

$$\rho_{2,2} = \sum_{j=1}^{n} \sum_{k=1}^{n} \sum_{j'=1}^{n} \sum_{k'=1}^{n} w_j w_k w_j w_{k'} \Phi_{jk,j'k'}.$$

Here, $\Phi_{jkj'}$ is the probability that three randomly sampled alleles, one each from individuals j, k, and j', are IBD. $\Phi_{jkj'k'}$ is the probability that four randomly sampled alleles, one each from individuals j, k, j', and k', are IBD. Finally, $\Phi_{jk,j'k'}$ is the joint probability that two randomly sampled alleles, one each from individuals j and k are IBD, and two randomly sampled alleles, one each from individuals j' and k', are IBD. Note that individuals j, k, j', and k' are not necessarily distinct. The variances of $\widetilde{H}$ and of $\widetilde{H}_{\mathrm{BLUE}}$ follow as Corollaries 7 and 8, once again differing only in the weight of a sampled individual in the mean kinship coefficient calculation.

Proposition 4: Consider a locus with I distinct alleles and parametric allele frequencies $p_i \in [0, 1]$, $i = 1, 2, \ldots, I$, and $\sum_{i=1}^{I} p_i = 1$. For a sample of size n individuals of any ploidy, inbreeding status, and relatedness,

$$\mathrm{Var}\left[\breve{H}\right] = \frac{1}{(1-\rho_2)^2} \mathrm{Var}\left[1 - \sum_{i=1}^{I} \breve{p}_i^2 \right] \tag{9}$$

is the variance of the unbiased estimator of expected heterozygosity $\breve{H}$, where $\rho_2 = \sum_{j=1}^{n} \sum_{k=1}^{n} w_j w_k \Phi_{jk}$ is a weighted mean kinship coefficient of the sample, and where w_k for $k = 1, 2, \ldots, n$ is the weight of individual k. Further, we have

$$\begin{aligned} \mathrm{Var}\left[1 - \sum_{i=1}^{I} \breve{p}_i^2 \right] = {} & \rho_{2,2} - \rho_2^2 + 2\left(\rho_2^2 - \rho_4\right) \sum_{i=1}^{I} p_i^2 \\ & + 4\left(2\rho_4 + \rho_2 - 2\rho_3 - \rho_{2,2}\right) \sum_{i=1}^{I} p_i^3 \\ & + \left(3\rho_{2,2} + 8\rho_3 - 6\rho_4 - 4\rho_2 - \rho_2^2\right) \left(\sum_{i=1}^{I} p_i^2 \right)^2. \end{aligned} \tag{10}$$

The proof of Equation 10 is presented for the specific case of $\mathrm{Var}[1 - \sum_{i=1}^{I} \widehat{p}_i^2]$ in Appendix B of DeGiorgio *et al.* (2010), where $\widehat{p}_i$ is substituted for $\breve{p}_i$, and $\bar{\Phi}_2$, $\bar{\Phi}_3$, and $\bar{\Phi}_4$, and $\bar{\Phi}_{2,2}$ coefficients are substituted for ρ_2, ρ_3, ρ_4, and $\rho_{2,2}$ coefficients, respectively. We provide an abbreviated version of this proof for the general case in the Appendix. Further, the approximate value of Equation 10 for samples wherein no individual is related to more than one other is

$$\mathrm{Var}\left[1 - \sum_{i=1}^{I} \breve{p}_i^2 \right] \approx 4\rho_2 \left[\sum_{i=1}^{I} p_i^3 - \left(\sum_{i=1}^{I} p_i^2 \right)^2 \right]. \tag{11}$$

For this simplifying case, the terms ρ_3, ρ_4, $\rho_{2,2}$, and ρ_2^2 are negligible compared to ρ_2.

In the Appendix, we reintroduce the definition of $\mathrm{Var}[\widetilde{H}]$ from DeGiorgio *et al.* (2010) (Corollary 7), and then define $\mathrm{Var}[\widetilde{H}_{\mathrm{BLUE}}]$ (Corollary 8), both of which take the form illustrated in Proposition 4.

As demonstrated by DeGiorgio *et al.* (2010), the mean kinship coefficients composing Equation 10 derive from the relationship between the 15 identity states available to four alleles (Gillois 1965; Cockerham 1971), and the coefficients of kinship between pairs, trios, quartets, and pairs of pairs of alleles within those four.

Bias of $\widehat{H}$ for samples containing related or inbred individuals

Here, we briefly derive an equation (Equation 12) within Proposition 5 that describes the bias of $\widehat{H}$, which we display in the left panels of Supplemental Material, Figure S1A and Figure S2A. We include Corollaries 9 and 10 to Proposition 5 within the Appendix for specific cases of bias derived from $\widehat{p}_i$-based and $\widetilde{p}_i$-based estimations, respectively. We also note that Equation A10 of Corollary 9 represents the form of the bias typically encountered in applications of $\widehat{H}$, as well as in all of our experimental scenarios.

Proposition 5: Consider a locus with I distinct alleles and parametric allele frequencies $p_i \in [0,1]$, $i = 1,2,\ldots,I$, and $\sum_{i=1}^{I} p_i = 1$. For a sample of n possibly related or inbred individuals, the bias of the estimator of expected heterozygosity $\widehat{H}$ changes with the true locus expected heterozygosity such that

$$\text{Bias}\left[\widehat{H}(\breve{p}_i)\right] = \frac{1-n\rho_2}{n-1}H, \tag{12}$$

where

$$\widehat{H}(\breve{p}_i) = \frac{n}{n-1}\left(1-\sum_{i=1}^{I}\breve{p}_i^2\right). \tag{13}$$

Proof: We begin by substituting Equation 6 into Equation 13 such that

$$\widehat{H}(\breve{p}_i) = \frac{n(1-\rho_2)}{n-1}\breve{H},$$

and

$$\mathbb{E}\left[\widehat{H}(\breve{p}_i)\right] = \frac{n(1-\rho_2)}{n-1}H.$$

From the definition of bias,

$$\begin{aligned}\text{Bias}\left[\widehat{H}(\breve{p}_i)\right] &= \mathbb{E}\left[\widehat{H}(\breve{p}_i)\right] - H \\ &= \frac{1-n\rho_2}{n-1}H.\end{aligned}$$

□

Variance of F_{ST} estimators using unbiased linear estimators of allele frequencies

Because the population differentiation statistic F_{ST} (Wright 1951) can be defined in terms of expected heterozygosities, it is possible to theoretically evaluate its approximate variance. A general estimator of F_{ST} can be written as

$$\breve{F}_{ST} = \frac{\breve{H}_{12} - \frac{1}{2}(\breve{H}_1 + \breve{H}_2)}{\breve{H}_{12}}, \tag{14}$$

where $\breve{H}_{12}$ is an unbiased estimator for the expected heterozygosity between a pair of sampled populations, numbered 1 and 2, defined as $\breve{H}_{12} = 1 - \sum_{i=1}^{I}\breve{p}_i\breve{q}_i$ (where $\breve{q}_i$ is a linear unbiased estimator of the frequency of allele i in population 2, analogous to $\breve{p}_i$ in population 1), while $\breve{H}_1$ and $\breve{H}_2$ are the within-population expected heterozygosities for populations 1 and 2, respectively. Referring to the numerator as x, and the denominator as y, we can write the expression for an approximation of the variance of a ratio as

$$\text{Var}\left[\frac{x}{y}\right] \approx \frac{(\mathbb{E}[x])^2}{(\mathbb{E}[y])^2}\left[\frac{\text{Var}[x]}{(\mathbb{E}[x])^2} + \frac{\text{Var}[y]}{(\mathbb{E}[y])^2} - 2\frac{\text{Cov}[x,y]}{\mathbb{E}[x]\mathbb{E}[y]}\right], \tag{15}$$

following the definition for the approximate variance of a ratio (Wolter 2007).

Proposition 6: Consider a locus with I distinct alleles across two populations and parametric allele frequencies $p_i \in [0,1]$, $i = 1,2,\ldots,I$, and $\sum_{i=1}^{I} p_i = 1$ for population 1, and $q_i \in [0,1]$, $i = 1,2,\ldots,I$, and $\sum_{i=1}^{I} q_i = 1$ for population 2. For samples of size n_1 and n_2 individuals from populations 1 and 2, respectively, each with individuals of any ploidy, inbreeding status, and relatedness, the variance of the population differentiation statistic calculated from their respective expected heterozygosities is approximated as

$$\begin{aligned}\text{Var}\left[\breve{F}_{ST}\right] \approx{}& \frac{\left[H_{12} - \frac{1}{2}(H_1+H_2)\right]^2}{H_{12}^2} \\ &\times\left[\frac{\text{Var}\left[\breve{H}_{12} - \frac{1}{2}(\breve{H}_1+\breve{H}_2)\right]}{\left[H_{12} - \frac{1}{2}(H_1+H_2)\right]^2} + \frac{\text{Var}\left[\breve{H}_{12}\right]}{H_{12}^2}\right. \\ &\left. - 2\frac{\text{Cov}\left[\breve{H}_{12} - \frac{1}{2}(\breve{H}_1+\breve{H}_2), \breve{H}_{12}\right]}{\left[H_{12} - \frac{1}{2}(H_1+H_2)\right]H_{12}}\right],\end{aligned} \tag{16}$$

where

$$\begin{aligned}\text{Var}\left[\breve{H}_{12} - \frac{1}{2}(\breve{H}_1+\breve{H}_2)\right] ={}& \text{Var}\left[\breve{H}_{12}\right] + \frac{1}{4}\text{Var}\left[\breve{H}_1\right] \\ &+ \frac{1}{4}\text{Var}\left[\breve{H}_2\right] - \left(\text{Cov}\left[\breve{H}_{12}, \breve{H}_1\right]\right. \\ &\left. + \text{Cov}\left[\breve{H}_{12}, \breve{H}_2\right]\right).\end{aligned} \tag{17}$$

In the Appendix, we provide a derivation of the variance and covariance components of Equations 16 and 17. For each of these equations, the result and proof are fairly long, and do not simplify when arranged into Equation 16.

Data availability

The authors state that all data necessary for confirming the conclusions presented in the article are represented fully within the article.

RESULTS

Analytical validation of $\widetilde{H}_{BLUE}$

We tested the performance of $\widetilde{H}_{BLUE}$ using both theory and simulations against that of the unbiased estimator $\widetilde{H}$ (DeGiorgio *et al.* 2010), and of $\widehat{H}$ (Nei and Roychoudhury 1974). Here, we applied the estimators to samples of individuals wherein each individual was related to exactly one other. Thus, for samples of size n individuals, the number of relative pairs was $n/2$. When inbred or closely related individuals are included in a sample, $\widehat{H}$ is a biased estimator of gene diversity for which we use

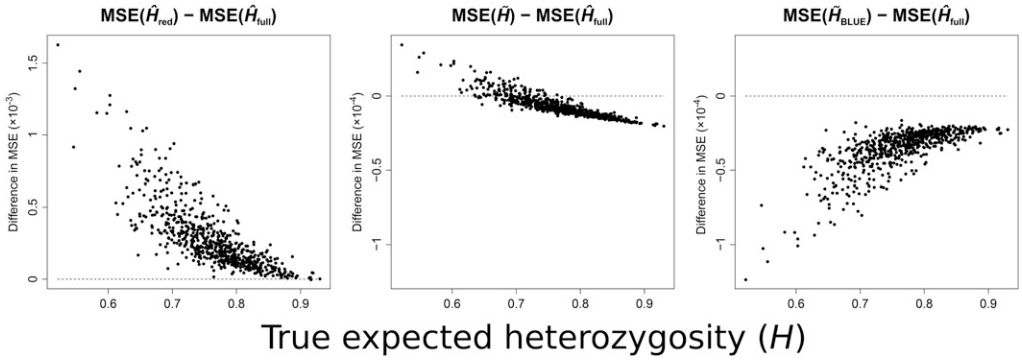

Figure 1 Theoretical difference in MSE between the unbiased estimator $\widehat{H}_{red}$ (left), $\widetilde{H}$ (center), or $\widetilde{H}_{BLUE}$ (right), and the biased estimator $\widehat{H}_{full}$ calculated at each of 645 microsatellite loci ($0.5212 \leq H \leq 0.9301$) in the MS5795 dataset for samples of 60 diploid individuals containing some inbred relative pairs. Each sampled individual was related to exactly one other, and samples contained 10 pairs of inbred full-siblings ($\Phi = 3/8$), 10 pairs of outbred full-siblings ($\Phi = 1/4$), and 10 outbred avuncular pairs ($\Phi = 1/8$). Dotted lines in each plot correspond to a difference in MSE of zero with $\widehat{H}_{full}$. See File S1 for the true expected heterozygosity values incorporated into analytical calculations.

the symbol $\widehat{H}_{full}$. To construct an unbiased estimator with $\widehat{H}$, we also applied $\widehat{H}$ to a reduced sample in which one member of each relative pair was removed randomly for samples containing only diploid individuals, and the haploid member was removed for each haploid-diploid (*i.e.*, male-female) pair (reduced sample size of $n/2$), and we denote this estimator by $\widehat{H}_{red}$. To evaluate the performance of the four estimators ($\widehat{H}_{full}$, $\widehat{H}_{red}$, $\widetilde{H}$, and $\widetilde{H}_{BLUE}$), we modified the factors upon which their variance depends: true locus expected heterozygosity (H), sample size n, and relatedness of individuals within the sample (Φ).

Effect of true locus expected heterozygosity, H, on estimators

We first evaluated the theoretical bias, variance, and mean squared error (MSE) of each estimator across the 645 human microsatellite loci from across the genome in the composite dataset MS5795 of Pemberton *et al.* (2013), where MSE is the sum of the squared bias and variance. The data used in our analyses is freely available online within File S1 of Pemberton *et al.* (2013) (http://www.g3journal.org/content/early/2013/03/27/g3.113.005728/suppl/DC1). We took the sample allele frequencies calculated from all individuals in the MS5795 dataset as the true population allele frequencies for the variance calculations, and, from these, determined the true expected heterozygosity at each locus using Equation 1 (see File S1; incorporated into Equation A10). Here, each sample contained 60 diploid individuals composed of 10 inbred full-sibling, 10 outbred full-sibling, and 10 outbred avuncular pairs. Each point in Figure 1 and Figure S1 represents a single analytical computation for a sample of 60 (or 30 for $\widehat{H}_{red}$) individuals at a microsatellite locus. We report the approximate variance and MSE because each individual is related to exactly one other in the sample, satisfying the assumption of Equation 11. Further, under this scenario DeGiorgio *et al.* (2010) showed that this was a reasonable approximation of the exact variance.

We begin by demonstrating the relative performance of the unbiased estimators $\widehat{H}_{red}$, $\widetilde{H}$, and $\widetilde{H}_{BLUE}$, measured in terms of MSE, against the biased estimator $\widehat{H}_{full}$ (Figure 1). While the variance of $\widehat{H}_{full}$ is invariably smaller than that of the other estimators, and the MSE and variance of each estimator decrease with increasing locus expected heterozygosity ($0.5212 \leq H \leq 0.9301$), $\widehat{H}_{full}$ accumulates bias quadratically with increasing H, and thus yields an increasingly unreliable estimate with increasing site diversity (Figure S1A, left). However, the effect of this trend differs for each comparison. The MSE of $\widehat{H}_{red}$ always exceeds that of $\widehat{H}_{full}$, because the removal of relatives to create the reduced sample causes a substantial increase in estimator variance, though, for high diversity markers, the MSE values of $\widehat{H}_{full}$ and $\widehat{H}_{red}$ converge (Figure 1, left). In contrast, $\widetilde{H}$ outperforms $\widehat{H}_{full}$ for most loci, demonstrating that the rate of decrease in MSE with increasing H is greater for $\widetilde{H}$ than for $\widehat{H}_{full}$ (Figure 1, center). Interestingly, the comparison of $\widetilde{H}_{BLUE}$ with $\widehat{H}_{full}$ shows an opposite trend to the preceding two. Despite the impact of bias, the decrease in variance of $\widehat{H}_{full}$ over the analyzed range outpaces that of $\widetilde{H}_{BLUE}$. Even so, $\widetilde{H}_{BLUE}$ uniformly yields a smaller MSE for the analyzed diploid samples (which contain a proportion of inbred individuals) across all loci (Figure 1, right).

To validate these theoretical predictions, we simulated 30 independent genotypes for each locus, and, for each independent genotype, simulated a single relative's genotype (inbred full-sibling, outbred full-sibling, or avuncular). Briefly, we generated the independent genotypes by sampling alleles uniformly at random from the distribution of allele frequencies at each microsatellite locus, and generated relatives by copying zero, one, or two alleles from the relative according to the probability the pair would share zero, one, or two alleles IBD [see Lange (2002), Chapter 5]. The patterns observed for the simulated data accord with those of the theoretical predictions (Figure S2, each point is based on 10^4 simulations). It is clear from these results that locus expected heterozygosity is heavily influential on estimator MSE. However, we also find that the observed value of expected heterozygosity for a locus normalized to its range of expected heterozygosity values has an impact on estimator MSE. The maximum and minimum values of expected heterozygosity for a locus depend on the number of distinct alleles (I), and the frequency of the most frequent allele (M), at that locus [see Theorem 2 of Reddy and Rosenberg (2012)]. We quantify proximity of H for a locus to its maximum possible value as $B = D/R$, where D is the observed value of expected heterozygosity for a locus minus its minimum possible value given I and M, and R is the maximum minus the minimum value of expected heterozygosity, given I and M, such that $B \in [0, 1]$. Loci with a smaller value of B yield a smaller MSE for all estimators (Figure S3).

Effect of sample size, n, on estimators

We next examined the properties of each estimator as a function of sample size. All estimators perform increasingly well for samples of increasing size. We demonstrate this property by measuring estimator MSE for samples containing 2–100 relative pairs of various type and ploidy at the D3S2427 locus, selected to highlight the improved performance of $\widetilde{H}_{BLUE}$ as the bias of $\widehat{H}_{full}$ increases ($H = 0.9301$; Figure 2). For these tests, we considered only a single relative pair type at a time. The unbiased estimators $\widetilde{H}$ and $\widetilde{H}_{BLUE}$ perform identically for diploid samples of first- and second-degree relative pairs regardless of inbreeding (Figure 2, A–D). Additionally, estimator MSE is uniformly smaller for samples containing only second-degree relative pairs than it is for samples containing only first-degree pairs (*cf.* Figure 2, A and B, and Figure 2, C and D; see also,

Figure S4A). However, $\widetilde{H}_{\mathrm{BLUE}}$ unambiguously outperforms the other estimators with relative pairs of varying ploidy (in this case, male-female full-sibling pairs at an X-linked locus). In this scenario, $\widehat{H}_{\mathrm{red}}$ provides a more accurate estimate of expected heterozygosity than does $\widetilde{H}$ when the reduced set is created by removing only males from the original while retaining females (Figure 2E). When all females are removed instead, and males retained (Figure 2F), the MSE of $\widehat{H}_{\mathrm{red}}$ is markedly the largest of the four estimators because 2/3 of the alleles in the sample are discarded, rather than 1/3. For samples with inbred full-siblings whose parents are brother and sister (Figure 2, C and D), the trend of MSE with sample size mirrors that of outbred diploid samples (Figure 2, A and B), but with larger MSE. However, the relative performance of $\widehat{H}_{\mathrm{full}}$ is notably worse for samples containing inbred diploid avuncular pairs (Figure 2D) than for samples containing outbred diploid avuncular pairs (Figure 2B). That is, its MSE remains greater than, or equal to, that of the other estimators over the range of sample sizes considered for the inbred diploid avuncular pair scenario (Figure 2D), but consistently has smaller MSE than $\widehat{H}_{\mathrm{red}}$ for the outbred diploid avuncular pair scenario (Figure 2B). Generally, increasing the sample size is most effective for samples of <20 individuals, and it is over this range that the difference in performance of the estimators is most apparent.

Effect of varying sample relative pair composition on estimators

Finally, we calculated the MSE of each estimator for all 1326 combinations of one to three relative pair types for samples of 100 individuals fixed at 50 relative pairs, which we represent as triangular heat maps, across samples containing outbred diploids, males and females at an X-linked locus, or inbred diploids (each individual related to exactly one other; Figure 3, Figure S4, Figure S5, Figure S6, Figure S7, and Figure S8). The kinship coefficients (Φ) for each relative pair type considered across our tests are defined in Lange (2002, Chapter 5) and DeGiorgio *et al.* (2010, see Table 2), and modeled on the D3S2427 locus ($H = 0.9301$).

The outbred diploid samples included parent-offspring ($\Phi = 1/4$), avuncular ($\Phi = 1/8$), and full-sibling ($\Phi = 1/4$) relative pairs. Because parent-offspring and full-sibling pairs have the same kinship coefficient, the heat maps in Figure 3, Figure S4A, Figure S5A, Figure S6A, and Figure S7A are symmetrical with parent-offspring and full-sibling pairs on the bottom vertices, and avuncular pairs on the top vertex. $\widehat{H}_{\mathrm{red}}$ yielded the largest MSE of the four estimators, and this value was constant throughout the space of the heat map (Figure S4A, second triangle), because all reduced sets are identical for outbred diploid samples. $\widetilde{H}_{\mathrm{BLUE}}$ consistently yielded the smallest MSE across configurations (Figure S4A, fourth triangle). As was the case in Figure 2, the MSE of the estimators $\widehat{H}_{\mathrm{full}}$, $\widetilde{H}$, and $\widetilde{H}_{\mathrm{BLUE}}$ was smallest for samples with only avuncular pairs, because these contain fewer dependent allele observations on average. We observed these features in simulated data as well (Figure S8A).

Although $\widetilde{H}_{\mathrm{BLUE}}$ performed best overall for samples including outbred diploid relative pairs at D3S2427, the estimator with the smallest variance in all situations is the biased estimator $\widehat{H}_{\mathrm{full}}$ (Figure S6A). However, because its squared bias increases with the number of first-degree pairs (Figure S5A), its relative performance declines compared to $\widetilde{H}_{\mathrm{BLUE}}$ as more of these pairs are sampled (Figure 3A, left triangle). The relative performance of $\widehat{H}_{\mathrm{red}}$ is highest when the number of first degree pairs is maximized, but this is due to the decreasing performance of $\widetilde{H}_{\mathrm{BLUE}}$ as more dependent observations are included (Figure 3A, center triangle). While the difference in MSE between $\widetilde{H}$ and $\widetilde{H}_{\mathrm{BLUE}}$ is always slight for samples of noninbred diploids, these values diverge as the complexity of the sample increases (Figure 3A, right triangle). That is, as the numbers of first- and second-degree pairs approach each other, $\widetilde{H}_{\mathrm{BLUE}}$ emerges decisively as the more accurate estimator, with the maximum value of this difference reached at 23 second-degree and 27 first-degree pairs. Thus, while the performance of the estimators for a sample containing relatives follows the same general trend, $\widetilde{H}_{\mathrm{BLUE}}$ provides the greatest accuracy for heterogeneous samples of outbred diploid individuals.

We also considered the relative performance of each estimator when using either the BLUE ($\widetilde{p}_i$) or the sample proportion ($\widehat{p}_i$) to estimate allele frequencies. Notably, all estimators perform best when the BLUE ($\widetilde{p}_i$) of allele frequency rather than the sample proportion ($\widehat{p}_i$) is used to infer population allele frequencies. We calculated the theoretical MSE for each estimator once with $\widehat{p}_i$, and once with $\widetilde{p}_i$, across all combinations of relative pairs for diploid individuals at the D3S2427 locus and mapped its value for the estimate with $\widehat{p}_i$ minus the estimate with $\widetilde{p}_i$ (Figure S7A). Because both frequency estimations yield the same values in samples of unrelated individuals, $\widehat{H}_{\mathrm{red}}$ performs identically for $\widehat{p}_i$ and $\widetilde{p}_i$, and is not included. The MSE of an estimator calculated with $\widetilde{p}_i$ is always smaller than that of the estimator calculated with $\widehat{p}_i$, and the pattern of divergence between their MSEs follows a similar trend across all estimators, resembling the rightmost panel in Figure 3A. This result suggests that the difference in MSE between $\widetilde{H}$ and $\widetilde{H}_{\mathrm{BLUE}}$ is driven primarily by the difference in performance between $\widehat{p}_i$ and $\widetilde{p}_i$. Both the $\widehat{p}_i$ and $\widetilde{p}_i$ estimators yield the same value at the vertices of the triangles, and the difference in their MSEs reaches a maximum at 22 second-degree pairs for $\widehat{H}_{\mathrm{full}}$ and 24 second-degree pairs for $\widetilde{H}$ and $\widetilde{H}_{\mathrm{BLUE}}$ (Figure S7A, center and right triangles). The MSE of $\widetilde{H}_{\mathrm{BLUE}}$ calculated with $\widehat{p}_i$ is, at most, on the order of 10^{-9} greater than that of $\widetilde{H}_{\mathrm{BLUE}}$ calculated with $\widetilde{p}_i$, indicating its robustness to variance in allele frequency determination (Figure S7A, right triangle). In contrast, the other estimators return a maximum difference in MSE on the order of 10^{-7}. The estimation of expected heterozygosity with $\widehat{H}_{\mathrm{full}}$, $\widetilde{H}$, or $\widetilde{H}_{\mathrm{BLUE}}$ will always yield a smaller MSE for samples of outbred, diploid individuals when $\widetilde{p}_i$ rather than $\widehat{p}_i$ is taken as the estimator of population allele frequency.

We repeated these tests in samples of mixed ploidy (Figure 3B, Figure S4B, Figure S5B, Figure S6B, Figure S7B, and Figure S8B), and $\widetilde{H}_{\mathrm{BLUE}}$ emerged similarly superior to the other estimators, once again yielding the smallest MSE. We analyzed the D3S2427 locus as X-linked for these tests, counting males as haploid and females as diploid, and observed full-sibling pairs [similarly to DeGiorgio *et al.* (2010), $\Phi = 1/2$ for male-male pairs, $\Phi = 1/4$ for male-female pairs, and $\Phi = 3/8$ for female-female pairs] for samples of 100 individuals and 50 relative pairs. All estimators reach their maximum MSE in samples containing only male-male pairs (Figure S4B). This is because the number of independent observations (indicated by a larger mean kinship coefficient) is smallest when there are no females in the sample. Correspondingly, the estimators yield smaller MSE values with increasing incorporation of male-female pairs. The minimum MSE of $\widehat{H}_{\mathrm{full}}$ is reached at 50 male-female pairs, as with $\widetilde{H}$ and $\widetilde{H}_{\mathrm{BLUE}}$ because its squared bias (Figure S5B) decreases with increasing male-female pairs, though its variance is smallest at 50 female-female pairs, due to the greater number of alleles in the sample (Figure S6B). To create the reduced sets, males were removed from male-female pairs to minimize the subsequent increase in MSE. That is, the removal of males removes 1/3 of the allele copies from the sample, rather than 2/3 if females are removed, or 1/2 for a pair of same-ploidy individuals, and so $\widehat{H}_{\mathrm{red}}$ has the same value across samples with the same number of male-male pairs (Figure S4B, second triangle).

The direct comparison of $\widetilde{H}_{\mathrm{BLUE}}$ with the other estimators once again yielded different signatures for each subtraction for mixed-ploidy samples (Figure 3B). The point of greatest difference in MSE between $\widehat{H}_{\mathrm{full}}$ and $\widetilde{H}_{\mathrm{BLUE}}$ occurs when all relative pairs are male-male, while the point of least difference occurs for samples of only male-female pairs

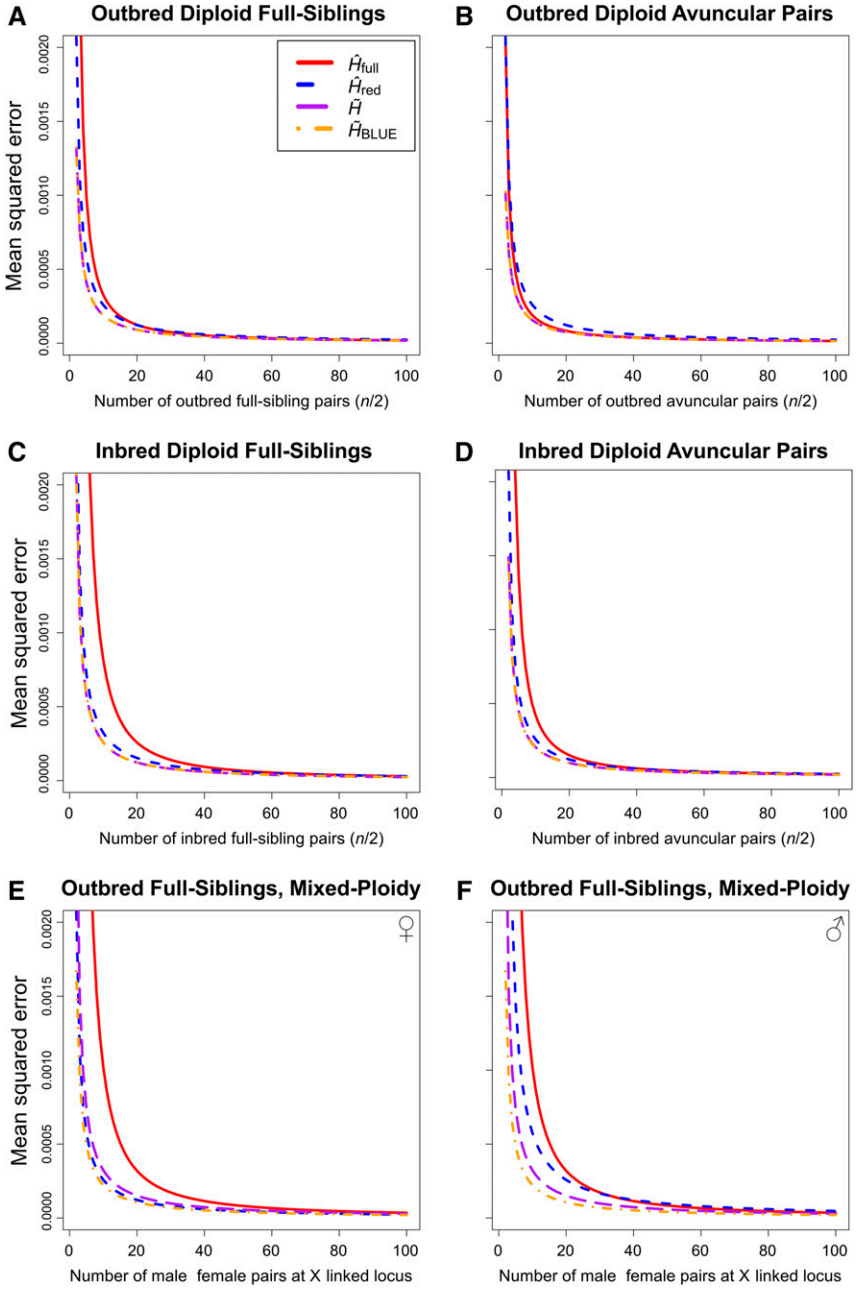

Figure 2 Theoretical MSE as a function of sample size for samples of outbred diploid full-siblings (A), outbred diploid avuncular pairs (B), inbred diploid full-siblings (C), inbred diploid avuncular pairs (D), male-female full siblings at an X-linked locus with the reduced set omitting males and retaining females (E), and male-female full siblings at an X-linked locus with the reduced set omitting females and retaining males (F). The samples were evaluated for the D3S2427 locus ($H = 0.9301$), and sample size was always twice the number of relative pairs included in the sample for samples containing 2–100 relative pairs. Each individual in the sample was related to exactly one other.

(Figure 3B, left triangle). This pattern broadly resembles the squared bias of $\widehat{H}_{\text{full}}$ (Figure S5B, first triangle), underscoring the effect of bias on estimator performance. The pattern of difference in performance between $\widehat{H}_{\text{red}}$ and $\widetilde{H}_{\text{BLUE}}$ differs markedly, and the two estimators perform most similarly as the number of male-male pairs decreases, reaching a minimum at 33 male-female pairs plus 17 female-female pairs (Figure 3B, middle triangle). $\widetilde{H}$ yields the closest MSE to that of $\widetilde{H}_{\text{BLUE}}$ for all relative pair configurations, and their difference is, at most, on the order of 10^{-6} (Figure 3B, right triangle). The pattern here mainly reflects the difference in performance between $\widehat{p}_i$ and $\widetilde{p}_i$ estimates of population allele frequency, as in Figure S7B, where $\widetilde{p}_i$ estimators yield increasingly smaller comparative MSE values as the numbers of relative pairs in the sample approach each other.

We repeated the preceding tests once more for a sample in which full-siblings resulting from a brother-sister mating were included alongside second-degree and outbred full-sibling pairs (Figure 3C, Figure S4C, Figure S5C, Figure S6C, Figure S7C, and Figure S8C). Here, the kinship of inbred individuals with each other was 3/8 rather than 1/4. For all estimators, the inclusion of inbred full-siblings increased the MSE of the estimator, with a maximum MSE at 50 inbred full-sibling pairs, and a minimum at 50 second-degree pairs. For $\widehat{H}_{\text{red}}$, this minimum was also reached for any sample in which there were no inbred individuals, because the reduced sample is identical for these (Figure S4C, second triangle). Again, $\widetilde{H}_{\text{BLUE}}$ was the least errant estimator across the space of sample configurations (Figure S4C, fourth triangle), and its advantage over the other estimators differs for each estimator (Figure 3C). Because the bias of $\widehat{H}_{\text{full}}$ is largest at 50 inbred full-sibling pairs, the greatest difference in performance between it and $\widetilde{H}_{\text{BLUE}}$ is at this point (Figure 3C, left triangle). Meanwhile, the largest differences in MSE between $\widehat{H}_{\text{red}}$ and $\widetilde{H}_{\text{BLUE}}$ are near the top vertex, where second-degree relative pairs

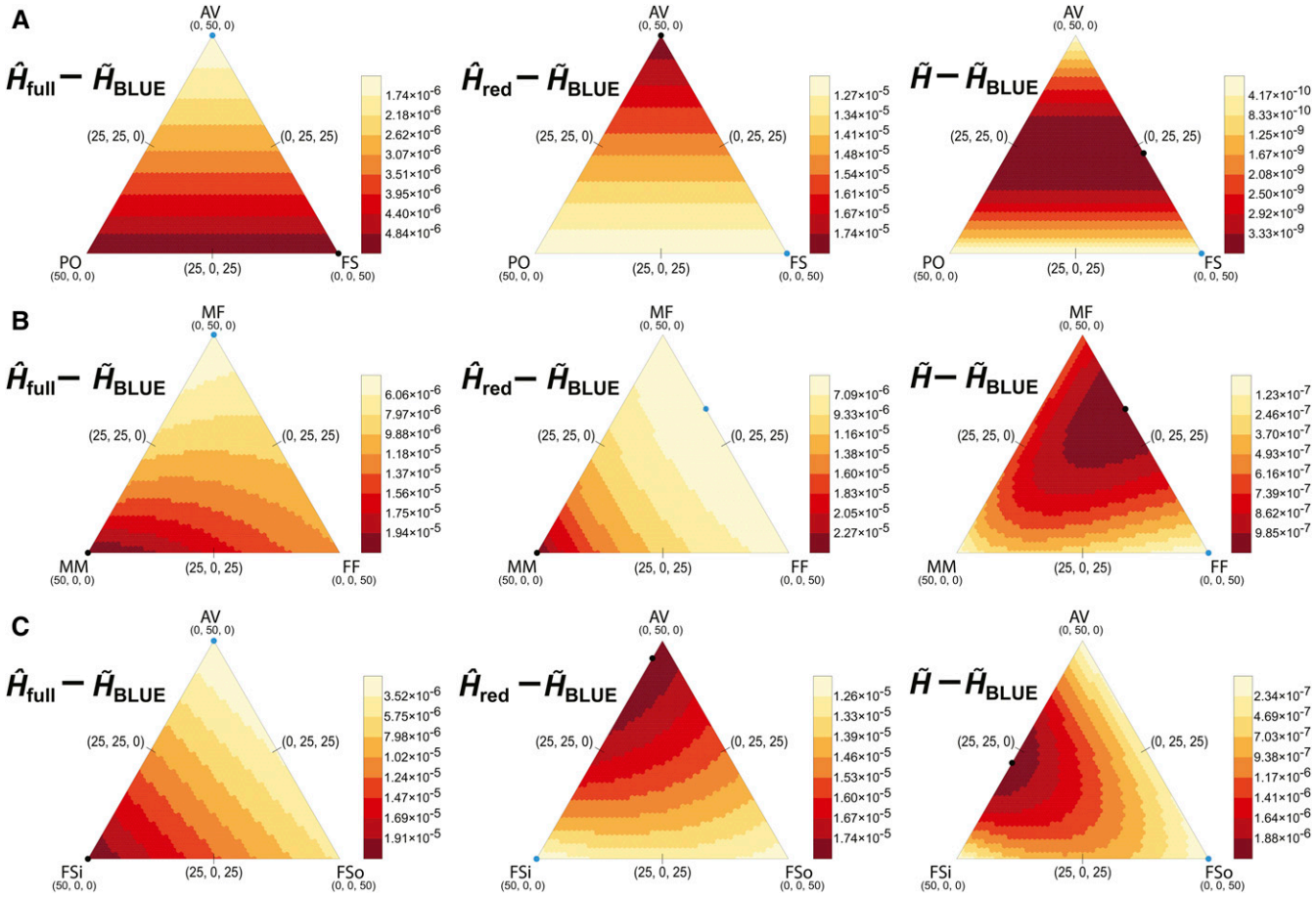

Figure 3 Theoretical difference in MSE between $\widehat{H}_{full}$ (left), $\widehat{H}_{red}$ (center), or $\widetilde{H}$ (right), and $\widetilde{H}_{BLUE}$, for samples of 100 (A) outbred diploid individuals, (B) male and female individuals at an X-linked locus, or (C) diploid individuals wherein some full siblings are inbred with brother-sister parents. The samples and MSE values considered for each subtraction were modeled on the D3S2427 locus ($H = 0.9301$). Each sample contained 50 relative pairs, such that each individual was related to exactly one other. Each sample configuration is a single point in the space of a heat map defined by three coordinates (each representing the count of a relative pair type). For each configuration, the MSE of $\widetilde{H}_{BLUE}$ is subtracted from that of the other estimators, yielding a value >0. Samples were composed of one to three relative pair types where the vertex of each heat map represents a sample with only a single relative pair type. The relative pair types were (A) parent-offspring (PO), second-degree avuncular (AV), and full-sibling (FS), (B) male-male (MM), male-female (MF), and female-female (FF) full-sibling such that the number of males and females in each sample is not fixed, or (C) inbred full-sibling (FSi), second-degree avuncular (AV), and outbred full-sibling (FSo). Blue and black points indicate the smallest and largest values, respectively, on each map. Threshold values for coloration are indicated in the scales to the right of each heat map, with smaller values colored lighter. Note that the scales are not identical across heat maps. The values upon which these subtractions are based are represented as heat maps in (A) Figure S4A, (B) Figure S4B, or (C) Figure S4C.

predominate, while the smallest are toward the bottom vertices (Figure 3C, center triangle). The difference in MSE between $\widetilde{H}$ and $\widetilde{H}_{BLUE}$ is at least an order of magnitude less than for the other comparisons, and increases for increasing sample complexity, but reaches its maximum for samples of 28 inbred full-sibling plus 22 second-degree pairs (Figure 3C, right triangle). This pattern reflects the decreased MSE for the estimators when calculated with $\widetilde{p}_i$ compared to their calculation with $\widehat{p}_i$ (Figure S7C). Ultimately, the performance of the estimators of expected heterozygosity across varying sample compositions depends on the estimator of allele frequency incorporated into the expected heterozygosity calculation. No matter the sample type, estimators based on $\widetilde{p}_i$ outperform estimators based on $\widehat{p}_i$, and $\widetilde{H}_{BLUE}$ outperforms $\widehat{H}_{full}$, $\widehat{H}_{red}$, and $\widetilde{H}$.

Tests of $\widetilde{H}_{BLUE}$ on single-nucleotide polymorphism (SNP) loci

Because SNP datasets are more common in recent studies, we performed analyses equivalent to our microsatellite analyses for 50 hypothetical SNP loci. These loci were biallelic with minor allele frequency (MAF) between 0.01 and 0.5, with increments of 0.01, corresponding to expected heterozygosity values ranging from 0.0198 to 0.5. We first measured the difference in MSE of $\widehat{H}_{full}$ with that of $\widehat{H}_{red}$, $\widetilde{H}$, or $\widetilde{H}_{BLUE}$ as a function of true locus expected heterozygosity (H), as we did in Figure 1 (Figure S9). For each locus, the MSE of $\widetilde{H}_{BLUE}$ was smallest, while that of $\widehat{H}_{full}$ was generally second-smallest, following the trend for microsatellite loci visible in Figure 1, wherein less diverse loci yielded a smaller MSE for $\widehat{H}_{full}$ than for $\widetilde{H}$. However, unlike for microsatellite loci, estimator MSE peaks midway through the range of evaluated SNP loci, such that the smallest MSE values lie at either extreme of the range and the largest MSE value, as well as the largest difference in MSE values for all comparisons, is at the locus with MAF= 0.15 ($H = 0.255$). Additionally, $\widehat{H}_{full}$ performs comparatively better than $\widehat{H}_{red}$ (Figure S9, left) and $\widetilde{H}$ (Figure S9, center) as H approaches 0.255, but is outperformed by these unbiased estimators as H approaches 0.5. Once more, the trend is opposite for the comparison between $\widehat{H}_{full}$ and $\widetilde{H}_{BLUE}$, showing the greatest comparative performance by $\widetilde{H}_{BLUE}$ at the same locus (MAF = 0.15, $H = 0.255$). Thus,

considering the results presented in Figure 1 and Figure S9, the greatest relative performance of $\widetilde{H}_{BLUE}$ for inbred samples is achieved at loci for which estimator MSE is largest.

We next examined the effect of sample size on estimator performance for hypothetical samples of outbred diploid, inbred diploid, and outbred male-female relative pairs at the simulated locus with MAF = 0.05 ($H = 0.095$). As we varied the sample size from two relative pairs to 100 (each individual related to exactly one other, one relative pair type per sample), we found that $\widetilde{H}_{BLUE}$ yielded the smallest MSE of all estimators only for samples containing male-female full-sibling pairs modeled at an X-linked locus (Figure S10, E and F). This observation mirrors the trend seen in Figure 2, wherein $\widetilde{H}_{BLUE}$ outperformed the other estimators across all sample sizes. However, $\widehat{H}_{full}$ yielded the smallest MSE across all sample sizes for outbred and inbred diploid full-siblings and avuncular pairs (Figure S10, A–D). This result is because the samples modeled here are minimally complex, with only one relative pair type, and modeled for a highly homozygous marker—two conditions under which the low bias and variance of $\widehat{H}_{full}$ result in favorable performance.

Finally, we analyzed estimator performance once more for the locus with MAF = 0.05 ($H = 0.095$), for a sample of 50 individuals across changing outbred diploid, inbred diploid, and male-female full-sibling relative pair compositions (Figure S11, A–C). We display these results as heat maps, and find that our results here are broadly concordant with those for the D3S2427 human microsatellite locus ($H = 0.9301$). As with the experiments displayed in Figure S10, the least complex samples yielded a smaller MSE for $\widehat{H}_{full}$ estimates than for $\widetilde{H}_{BLUE}$ estimates. Correspondingly, samples whose relative pair compositions resulted in fewer independent allele observations were more accurately and precisely evaluated with $\widetilde{H}_{BLUE}$. Thus, while sampling lower-diversity markers may occasionally favor the use of $\widehat{H}_{full}$, the inclusion of two or more relative pair types in the sample is likely to bias $\widehat{H}_{full}$, and require the use of $\widetilde{H}_{BLUE}$ to yield accurate inferences.

Empirical application of $\widetilde{H}_{BLUE}$

To conclude our investigation into the performance of $\widetilde{H}_{BLUE}$, we applied it to empirical data from the MS5795 dataset. We retrieved human microsatellite data from 5795 individuals (11,590 allele copies) across 645 autosomal loci sampled genome wide. We assumed the mean value across loci for $\widehat{H}_{red}$ in each of 267 populations to be the true expected heterozygosity value for these populations, as it is an unbiased estimate. We additionally chose to compare the other estimators with $\widehat{H}_{red}$, because an important basis for their evaluation is their agreement with this unbiased estimator, irrespective of the data to which they are applied.

To emphasize this, we performed three Wilcoxon signed-rank tests to compare the ranking of populations by their mean expected heterozygosity across all loci calculated with $\widehat{H}_{red}$, and either $\widehat{H}_{full}$, $\widetilde{H}$, or $\widetilde{H}_{BLUE}$ (Table 1). At the $\alpha < 0.01$ significance level, the comparisons showed that the inclusion of relatives for $\widehat{H}_{full}$ was highly significant on the rankings it yielded, indicating that not correcting for relatedness among samples can significantly alter the estimates of expected heterozygosity. However, both $\widetilde{H}$ and, especially, $\widetilde{H}_{BLUE}$, yielded P-values greater than the threshold for the test against $\widehat{H}_{red}$. These results indicate that the estimates of expected heterozygosity are not significantly affected by the inclusion of related individuals in the sample when relatedness is taken into account. Furthermore, a test between $\widetilde{H}$ and $\widetilde{H}_{BLUE}$ yielded a P-value of 3.44×10^{-2}, suggesting no significant difference in the ranking of populations by mean expected heterozygosity with these two estimators.

Although the unbiased estimators $\widetilde{H}$ and $\widetilde{H}_{BLUE}$ have smaller MSE than $\widehat{H}_{full}$ for samples with related individuals, their variance tends to be larger than that of $\widehat{H}_{full}$. DeGiorgio *et al.* (2010) previously showed that the difference in SD of $\widetilde{H}$ with $\widehat{H}_{full}$ was small, while the mean values of $\widetilde{H}$ and $\widehat{H}_{red}$ were much more similar to each other than either of them was to the mean of $\widehat{H}_{full}$. We again show this to be the case, and find as well that $\widetilde{H}_{BLUE}$ not only repeats, or improves upon, the concordance of $\widetilde{H}$ with $\widehat{H}_{red}$, but, in some cases, $\widetilde{H}_{BLUE}$ has a smaller SD than does $\widehat{H}_{full}$ (Figure 4, left and center panels). A direct comparison of the performance of $\widetilde{H}$ against that of $\widetilde{H}_{BLUE}$ (Figure 4, right panel) shows that $\widetilde{H}_{BLUE}$ has a generally improved SD, and similarity to the $\widehat{H}_{red}$ estimate over $\widetilde{H}$. For some samples (primarily those from the Americas), this is not the case, possibly because all close relatives were not identified in the original dataset, resulting in an incorrect kinship matrix for calculation of the statistic.

Table 1 Wilcoxon signed-rank test for mean across loci of $\widehat{H}_{red}$ with $\widehat{H}_{full}$, $\widetilde{H}$, and $\widetilde{H}_{BLUE}$ for the 93 populations whose samples contained related individuals

Comparison	*P*-Value for Wilcoxon Signed-Rank Test
$\widehat{H}_{red}$ with $\widehat{H}_{full}$	4.39×10^{-15}
$\widehat{H}_{red}$ with $\widetilde{H}$	1.00×10^{-2}
$\widehat{H}_{red}$ with $\widetilde{H}_{BLUE}$	0.255

Improving estimates of F_{ST} by application of $\widetilde{H}_{BLUE}$

We predicted that the smaller MSE of $\widetilde{H}_{BLUE}$ would translate to improved accuracy for estimators that are summaries of expected heterozygosity when samples contain related individuals. To test this hypothesis, we calculated the population differentiation statistic, F_{ST} (Equation 4), for pairs of populations whose samples in the MS5795 dataset contained related individuals. Our intent was to compare the MSE and bias of the commonly used F_{ST} estimator of Reynolds *et al.* (1983), which is based on $\widehat{H}_{full}$, and which we label as $\widehat{F}_{ST}$, to an estimate of F_{ST} calculated from $\widetilde{H}_{BLUE}$, which we label $\widetilde{F}_{ST,BLUE}$. The formulas for these estimators follow the form of the general estimator of F_{ST} (Equation 14). We first measured the MSE of both methods (and an estimate using $\widetilde{H}$, $\widetilde{F}_{ST}$) on simulated data, where the F_{ST} of pairs of populations with samples of size 60 diploids each (30 relative pairs, 10 inbred full-sibling, 10 outbred full-sibling, and 10 avuncular pairs; Figure 5) was averaged across 10^4 simulated replicates. The calculations included here were performed for simulated Gujarati and Maya (left), Gujarati and Japanese (center), or Gujarati and Hadza (right) samples for the least diverse (TCTA015M_22), median diverse (D10S2327), and most diverse (D3S2427) loci of the MS5795 dataset, following their allele frequency distribution in MS5795. $\widetilde{F}_{ST,BLUE}$ consistently has a smaller MSE than the others, and the MSE of all estimators of F_{ST} decreases with increasing locus diversity, as the MSE of the estimator of expected heterozygosity decreases.

We additionally find that $\widehat{F}_{ST}$ has an upward bias compared with $\widehat{F}_{ST,red}$ (calculated with $\widehat{H}_{red}$), as well as a larger SD in general than $\widetilde{F}_{ST,BLUE}$ (Figure 6). Furthermore, all values of $\widetilde{F}_{ST,BLUE}$ are smaller than the paired value of $\widehat{F}_{ST}$ calculated for the same population. The difference in the mean of $\widehat{F}_{ST}$ and of $\widetilde{F}_{ST,BLUE}$ across all loci with the mean of $\widehat{F}_{ST,red}$, an estimator which serves as a proxy for the true value of F_{ST}, is displayed on the vertical axis, while the horizontal axis measures the SD of $\widehat{F}_{ST}$ and of $\widetilde{F}_{ST,BLUE}$ (Figure 6). Supporting our observations indicating the improved accuracy of $\widetilde{F}_{ST,BLUE}$ over $\widehat{F}_{ST}$, Wilcoxon signed-rank tests (Table 2) between $\widehat{F}_{ST,red}$ and either $\widehat{F}_{ST}$ or $\widetilde{F}_{ST,BLUE}$ indicate that the inclusion of relatives significantly affects the estimate of population differentiation at the $\alpha < 0.01$ significance level. Meanwhile, $\widehat{F}_{ST,red}$ and

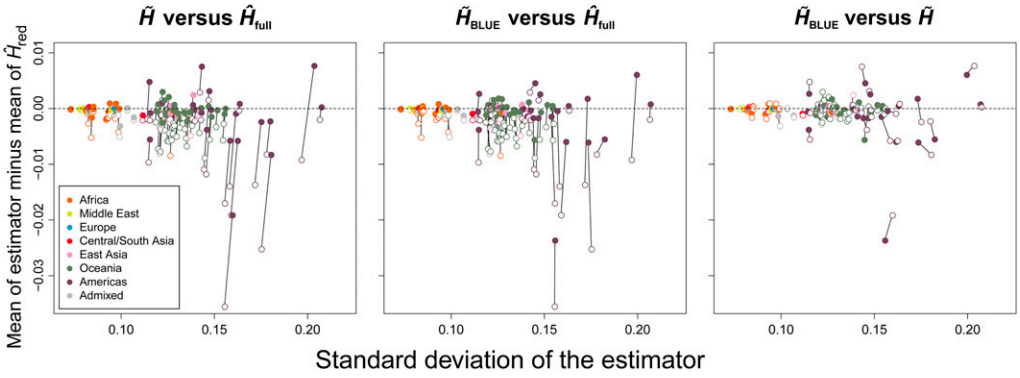

Figure 4 Application of the estimators to dataset MS5795. Here, we show a comparison of two estimators at a time ($\widehat{H}_{full}$, $\widetilde{H}$, or $\widetilde{H}_{BLUE}$) by the difference in their mean with that of $\widehat{H}_{red}$ across the 645 sampled microsatellite loci of MS5795 (vertical axis), and by their SDs (horizontal axis). The horizontal dotted line corresponds to no difference between the mean of the estimator and the mean of the unbiased estimator $\widehat{H}_{red}$. Solid lines connect calculations made for the same population with different estimators. Points are colored by geographic division defined in the dataset. Only the 93 populations with relatives in their samples were included because $\widehat{H}_{full}$, $\widetilde{H}$, and $\widetilde{H}_{BLUE}$ return the same value for samples of unrelated individuals. In the leftmost plot, open points are estimates for $\widehat{H}_{full}$, while closed points are for $\widetilde{H}$. In the center plot, open points are estimates for $\widehat{H}_{full}$, while closed points are for $\widetilde{H}_{BLUE}$. In the rightmost plot, open points are estimates for $\widetilde{H}$, while closed points are for $\widetilde{H}_{BLUE}$.

$\widetilde{F}_{ST,BLUE}$ are not significantly different in their estimates. These results suggest that the improved properties of $\widetilde{H}_{BLUE}$ transfer to the summaries that include it in their calculations.

DISCUSSION

We have introduced $\widetilde{H}_{BLUE}$, an extension to the estimator ($\widetilde{H}$) of expected heterozygosity developed by DeGiorgio *et al.* (2010) that yields a smaller mean squared error in samples containing related individuals, while maintaining unbiasedness. Conveniently, the derivations of $\widetilde{H}_{BLUE}$, and its variance, are parallel in form to those of $\widetilde{H}$, and we were therefore able to analytically evaluate the performance of the new estimator simultaneously with that of its predecessor. Our updated estimator, $\widetilde{H}_{BLUE}$, is based on results from McPeek *et al.* (2004), who characterized the BLUE ($\widetilde{p}_i$) of allele frequency. The BLUE improves the precision of allele frequency estimation in complex pedigrees, for which the sample proportion ($\widehat{p}_i$, the estimator of allele frequency used in $\widehat{H}$ and $\widetilde{H}$) is unbiased, but increases in variance with inclusion of related and inbred individuals. Because the properties of the estimator of allele frequency transfer to the estimator of expected heterozygosity, $\widetilde{H}_{BLUE}$ is likely to outperform $\widetilde{H}$ in situations where $\widetilde{p}_i$ has a smaller variance than $\widehat{p}_i$. This trend is true for genome-wide data as well (Figure 4 and Table 1).

Overall, $\widetilde{H}_{BLUE}$ yields identical results to $\widetilde{H}$ in samples containing only one relative pair type, but the two diverge in performance as sample complexity increases (see heat maps in Figure 3, Figure S4, Figure S5, Figure S6, Figure S7, and Figure S8). While both estimators are unbiased, $\widetilde{H}$ experiences a larger increase in variance for each additional relative pair type introduced into a sample after the first. This holds true for all sample types regardless of ploidy and inbreeding, suggesting that $\widetilde{H}_{BLUE}$ will outperform $\widetilde{H}$ in practice, where datasets are often complex. Furthermore, the results of our empirical analysis provide an equally important complement to this observation. Of the 93 populations from the MS5795 dataset we considered that contained relative pairs in their samples, each contained sampled individuals that were not related to any other in the sample. Thus, these samples were more complex than those in which each individual was part of a relative pair of the same type. For most of these cases, except for some American populations (discussed below), $\widetilde{H}_{BLUE}$ outperformed $\widetilde{H}$. This is corroborated by the Wilcoxon signed-rank test (Table 1). We expect therefore that any scenario in which there is heterogeneity in relative pair type among sampled individuals, as is observed in many human population-genetic datasets (Pemberton *et al.* 2010, 2013), should favor the application of $\widetilde{H}_{BLUE}$ over other estimators.

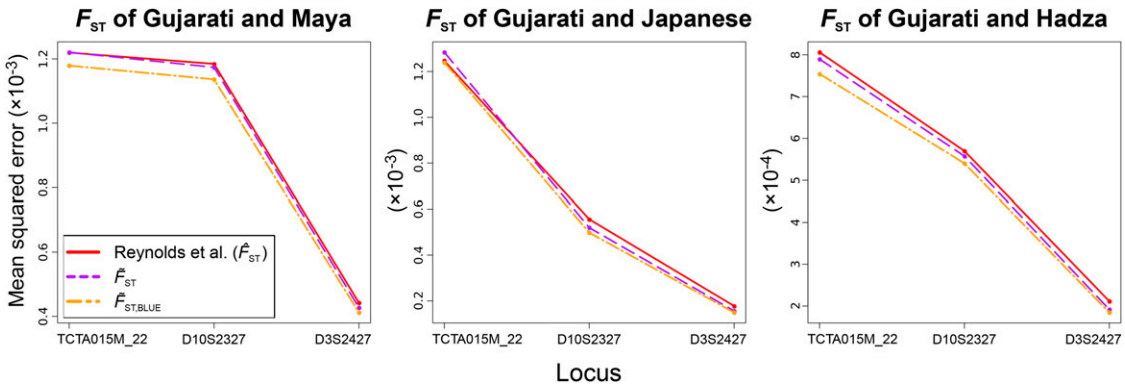

Figure 5 Application of the estimators $\widehat{H}_{full}$, $\widetilde{H}$, and $\widetilde{H}_{BLUE}$ to the calculation of F_{ST} as $\widehat{F}_{ST}$, $\widetilde{F}_{ST}$, and $\widetilde{F}_{ST,BLUE}$, respectively, using simulated data for the Gujarati sample, with either the Maya (left), Japanese (center), or Hadza (right) samples, showing MSE on the vertical axis. The Reynolds *et al.* (1983) estimator is equivalent to the application of $\widehat{H}_{full}$ in calculating population differentiation. The simulated samples contained 60 individuals and 30 relative pairs, of which 10 were inbred full-siblings, 10 were outbred full-siblings, and 10 were outbred avuncular pairs. Each individual was related to exactly one other, and the data were simulated following the same probabilistic method as employed to generate Figure S2. The three loci displayed on the horizontal axis are the least diverse, median diverse, and most diverse loci of the 645 MS5795 human microsatellites.

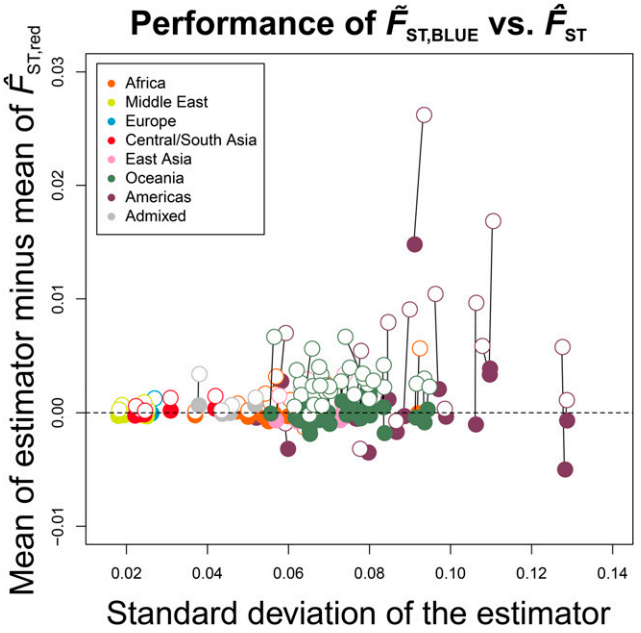

Figure 6 Application of the estimators $\widetilde{H}_{BLUE}$ and $\widehat{H}_{full}$ to the estimation of F_{ST} as $\widehat{F}_{ST}$ and $\widetilde{F}_{ST,BLUE}$, respectively, from empirical data. Similarly to Figure 4, the difference between the mean of the estimator of F_{ST} (either derived from $\widetilde{H}_{BLUE}$ or $\widehat{H}_{full}$) and an unbiased estimator (derived from $\widehat{H}_{red}$), is displayed on the vertical axis, while the SD of the estimator is displayed on the horizontal axis. The empty circles represent the Reynolds *et al*. (1983) estimator (identical to the $\widehat{H}_{full}$-derived estimation), while the filled circles represent the estimation derived from $\widetilde{H}_{BLUE}$. Here, the F_{ST} values for the French sample with each of the 92 other samples containing related individuals in the dataset MS5795 are plotted, colored by the region of the changing sample.

In addition, random sampling of small isolated populations yields an increased chance that related individuals will be included with large enough sample sizes. Further, inbreeding may confound estimates of diversity, and mislead $\widehat{H}_{full}$ to underreport true population expected heterozygosity. Populations of interest that may display these attributes include geographically isolated human settlements in remote alpine (Coia *et al.* 2012; Capocasa *et al.* 2013), South American rainforest (Wang *et al.* 2007), and Siberian taiga and steppe habitats (Dulik *et al.* 2012), and groups such as the Old Order Amish (Van Hout *et al.* 2010), Hutterites (Abney *et al.* 2002; Chong *et al.* 2011), and Mennonites (Payne *et al.* 2011). Further, though our analysis did not directly consider polyploid organisms, the applicability of $\widetilde{H}_{BLUE}$ to samples containing individuals of any, and varying, ploidy highlights its usefulness for such data. Prominently, analysis on polyploid organisms such as plants including tetraploid *Arabidopsis thaliana* (Hollister *et al.* 2012), and hexaploid bread wheat (Nielsen *et al.* 2014), both of which self-fertilize, and may therefore be inbred, as well as commercially and ecologically significant Hymenopteran insects, including honeybees (Solignac *et al.* 2003; Harpur *et al.* 2014), bumblebees (Lye *et al.* 2011), and ants (Butler *et al.* 2014), whose males are haploid at all loci, while females are diploid, is likely to benefit from the improved accuracy and precision of $\widetilde{H}_{BLUE}$.

We additionally believe that continued investigations into the diversity at single sites in organisms as diverse as dogs (Sutter *et al.* 2007), gray wolves (Zhang *et al.* 2014), humans living at high altitude (Simonson *et al.* 2010; Huerta-Sánchez *et al.* 2013), and rice (Huang *et al.* 2012), in addition to host-microbiome studies (Blekhman *et al.* 2015), will benefit from the advances provided by $\widetilde{H}_{BLUE}$. These studies, as well as many others, have performed scans for positive selection using genomic outliers of population differentiation-based statistics (*e.g.*, F_{ST}, locus-specific branch length, and the population branch statistic), where the calculation is performed per-site, rather than averaged across a large number of sites. Such studies would benefit from estimators of genetic diversity, such as $\widetilde{H}_{BLUE}$ and $\widetilde{F}_{ST,BLUE}$, with improved variance.

Table 2 Wilcoxon signed-rank test for weighted mean across all loci of $\widehat{F}_{ST,red}$ with $\widehat{F}_{ST}$ and $\widetilde{F}_{ST,BLUE}$ for the French population with the 92 other populations whose samples contained related individuals

Comparison	*P*-Value for Wilcoxon Signed-Rank Test
$\widehat{F}_{ST,red}$ with $\widehat{F}_{ST}$	5.25×10^{-15}
$\widehat{F}_{ST,red}$ with $\widetilde{F}_{ST,BLUE}$	0.967

It is pertinent at this point to revisit a pair of potential limitations in our method and examine their implications. First, in Figure 4 (rightmost panel), the mean of $\widetilde{H}$ is either closer to that of $\widehat{H}_{red}$ than to $\widetilde{H}_{BLUE}$, has smaller SD than $\widetilde{H}_{BLUE}$, or both for certain samples (predominantly from the Americas). These observations indicate that the accuracy and precision of $\widetilde{H}_{BLUE}$ may be impacted by the accuracy of the kinship information incorporated into the calculation. The pedigrees of smaller, more remotely located, populations may be more complex compared to those of larger groups. Further, with a greater proportion of relative pairs in each sample, the effect of relative pair type misidentification may be larger. For RELPAIR (Epstein *et al.* 2000), which was the software chosen to identify relative pairs in MS5795 samples, second-degree pairs cannot be identified as confidently as first-degree pairs (Pemberton *et al.* 2010). Even so, although $\widetilde{H}$ may exhibit a somewhat greater robustness to relative pair misclassification, it is still generally outperformed by $\widetilde{H}_{BLUE}$.

The second point we address is the smaller MSE of $\widehat{H}_{full}$ at less diverse loci in the dataset, especially for samples with fewer relative pairs. While the variance of $\widehat{H}_{full}$ is always smaller than that of the other estimators, its bias increases with increasing locus allelic diversity. It is for this reason that the unbiasedness of $\widetilde{H}_{BLUE}$ is its most desirable property. In practice, the mean of expected heterozygosity is often taken across loci. Based on such an approach, $\widetilde{H}_{BLUE}$ (and $\widetilde{H}$ as well) will return the mean expected heterozygosity, and the variance of this estimation (as with all estimators taking the mean across loci) approaches zero as more loci are sampled. An interesting property of all estimators is that their variance (and therefore MSE) is larger for loci whose value for B is closer to 1, where $B = D/R$ ($B \in [0, 1]$; see *Results* and Figure S3). Because this effect is greatest for loci with lower true values of H, we expect $\widehat{H}_{full}$ to have the smallest MSE of all estimators at less diverse loci that are close to their maximum expected heterozygosity, and for which the sample mean kinship coefficient is insufficiently large to appreciably bias the estimator (Equation 12). It is thus important to note that no estimator is uniformly superior to the others. Accordingly, the unique limitation of $\widetilde{H}_{BLUE}$ is that the sample kinship matrix must be invertible for the calculation to proceed.

$\widetilde{H}_{BLUE}$ additionally confers its improved MSE over $\widehat{H}_{full}$ downstream to calculations that incorporate estimates of expected heterozygosity. To illustrate this point, we computed F_{ST} as a function of three estimators: $\widehat{H}_{full}$, $\widetilde{H}$, and $\widetilde{H}_{BLUE}$. For simulated data, we found that $\widetilde{F}_{ST,BLUE}$, yielded an estimate with smaller MSE for the three tested loci than did $\widehat{F}_{ST}$ (Figure 5) or $\widetilde{F}_{ST}$, and a much smaller mean distance from the true F_{ST} value than $\widehat{F}_{ST}$. For empirical data (Figure 6), we observed a consistent upward bias for $\widehat{F}_{ST}$ compared to $\widehat{F}_{ST,red}$ in samples containing relatives that followed much the same pattern as the downward bias of $\widehat{H}_{full}$ for such samples. This trend is clear when we consider the formula for F_{ST}, which can be written as $1 - (H_1 + H_2)/(2H_{12})$. Taking $\widehat{H}_{1,full}$ and $\widehat{H}_{2,full}$ as H_1 and H_2, this expression yields a larger value

than if $\widehat{H}_{1,\mathrm{red}}$ and $\widehat{H}_{2,\mathrm{red}}$ were used, because the ratio $(H_1 + H_2)/(2H_{12})$ is smaller for downwardly biased estimators. Interestingly, the SD of $\widetilde{F}_{\mathrm{ST,BLUE}}$ is, in most cases, smaller than that of $\widehat{F}_{\mathrm{ST}}$ for the dataset, while the SD of $\widetilde{H}_{\mathrm{BLUE}}$ was frequently (though not consistently) larger than that of $\widehat{H}_{\mathrm{full}}$ (Figure 4, center panel).

It is thus noteworthy to consider that the performance of $\widetilde{H}_{\mathrm{BLUE}}$ and $\widehat{H}_{\mathrm{full}}$ may diverge further in their applications, where any improvement in MSE for $\widetilde{H}_{\mathrm{BLUE}}$ may be magnified downstream. This is highlighted by the increased concordance between $\widetilde{F}_{\mathrm{ST,BLUE}}$ and $\widehat{F}_{\mathrm{ST,red}}$ compared to $\widetilde{H}_{\mathrm{BLUE}}$ and $\widehat{H}_{\mathrm{red}}$ (*cf.* *P*-values between Table 1 and Table 2). With this in mind, applications of $\widetilde{F}_{\mathrm{ST,BLUE}}$ can also be considered. Two such examples are the locus-specific branch length (LSBL; Shriver *et al.* 2004) and the similar population branch statistic (PBS; Yi *et al.* 2010). These statistics incorporate F_{ST} values between three populations as measures of branch length to detect positive selection at a locus. Loci for which the unrooted three-taxon tree indicates a significantly longer branch length in a particular lineage may represent regions possibly under selection. To allow for the easy application of $\widetilde{H}_{\mathrm{BLUE}}$, we have written an R script, *BestHet*, that computes $\widetilde{H}_{\mathrm{BLUE}}$, $\widetilde{F}_{\mathrm{ST,BLUE}}$, and $LSBL_{\mathrm{BLUE}}$, given matrices of genotype and kinship data for a sample (download available at http://www.personal.psu.edu/mxd60/best_het.html).

ACKNOWLEDGMENTS

We thank two anonymous reviewers for their insightful comments. This work was supported by Pennsylvania State University startup funds from the Eberly College of Science.

LITERATURE CITED

Abney, M., C. Ober, and M. S. McPeek, 2002 Quantitative-trait homozygosity and association mapping and empirical genomewide significance in large, complex pedigrees: fasting serum-insulin level in the Hutterites. Am. J. Hum. Genet. 70: 920–934.

Blekhman, R., J. K. Goodrich, K. Huang, Q. Sun, R. Bukowski *et al.*, 2015 Host genetic variation impacts microbiome composition across human body sites. Genome Biol. 16: 191.

Butler, I. A., K. Siletti, P. R. Oxley, and D. J. C. Kronauer, 2014 Conserved microsatellites in ants enable population genetic and colony pedigree studies across a wide range of species. PLoS One 9: e107334.

Capocasa, M., C. Battagia, P. Anagnostou, F. Montinaro, I. Boschi *et al.*, 2013 Detecting genetic isolation in human populations: a study of European language minorities. PLoS One 8: e56371.

Chong, J. X., A. A. Oktay, Z. Dai, K. J. Swoboda, T. W. Prior *et al.*, 2011 A common spinal muscular atrophy deletion mutation is present on a single founder haplotype in the US Hutterites. Eur. J. Hum. Genet. 19: 1045–1051.

Cockerham, C. C., 1971 Higher order probability functions of identity of alleles by descent. Genetics 69: 235–246.

Coia, V., I. Boschi, F. Trombetta, F. Cavulli, F. Montinaro *et al.*, 2012 Evidence of high genetic variation among linguistically diverse populations on a micro-geographic scale: a case study of the Italian Alps. J. Hum. Genet. 57: 254–260.

DeGiorgio, M., and N. A. Rosenberg, 2009 An unbiased estimator of gene diversity in samples containing related individuals. Mol. Biol. Evol. 26: 501–512.

DeGiorgio, M., I. Jankovic, and N. A. Rosenberg, 2010 Unbiased estimation of gene diversity in samples containing related individuals: exact variance and arbitrary ploidy. Genetics 186: 1367–1387.

Dulik, M. C., S. I. Zhadanov, L. P. Osipova, A. Askapuli, L. Gau *et al.*, 2012 Mitochondrial DNA and Y chromosome variation provides evidence for a recent common ancestry between Native Americans and Indigenous Altaians. Am. J. Hum. Genet. 90: 229–246.

Epstein, M. P., W. L. Duren, and M. Boehnke, 2000 Improved inference of relationships for pairs of individuals. Am. J. Hum. Genet. 67: 1219–1231.

Gillois, M., 1965 Relation d'identité en génétique. Ann. Inst. Henri Poincaré B 2: 1–94.

Harpur, B. A., C. F. Kent, D. Molodtsova, J. M. D. Lebon, A. S. Alqarni *et al.*, 2014 Population genomics of the honey bee reveals strong signatures of positive selection on worker traits. Proc. Natl. Acad. Sci. USA 111: 2614–2619.

Hollister, J. D., B. J. Arnold, E. Svedin, K. S. Xue, B. P. Dilkes *et al.*, 2012 Genetic adaptation associated with genome-doubling in autotetraploid *Arabidopsis arenosa*. PLoS Genet. 8: e1003093.

Huang, X., N. Kurata, X. Wei, Z. Wang, A. Wang *et al.*, 2012 A map of rice genome variation reveals the origin of cultivated rice. Nature 490: 497–501.

Hudson, R. R., M. Slatkin, and W. P. Maddison, 1992 Estimation of levels of gene flow from DNA sequence data. Genetics 132: 583–589.

Huerta-Sánchez, E., M. DeGiorgio, L. Pagani, A. Tarekegn, R. Ekong *et al.*, 2013 Genetic signatures reveal high-altitude adaptation in a set of Ethiopian populations. Mol. Biol. Evol. 30: 1877–1888.

Lange, K., 2002 *Mathematical and Statistical Methods for Genetic Analysis*, Ed. 2. Springer, New York.

Lye, G. C., O. Lepais, and D. Goulson, 2011 Reconstructing demographic events from population genetic data: the introduction of bumblebees to New Zealand. Mol. Ecol. 20: 2888–2900.

McPeek, M. S., X. Wu, and C. Ober, 2004 Best linear unbiased allele-frequency estimation in complex pedigrees. Biometrics 60: 359–367.

Nei, M., 1973 Analysis of gene diversity in subdivided populations. Proc. Natl. Acad. Sci. USA 70: 3321–3323.

Nei, M., and A. K. Roychoudhury, 1974 Sampling variances of heterozygosity and genetic distance. Genetics 76: 379–390.

Nielsen, N. H., G. Backes, J. Stougaard, S. U. Andersen, and A. Jahoor, 2014 Genetic diversity and population structure analysis of European hexaploid bread wheat (*Triticum aestivum L.*) varieties. PLoS One 9: e94000.

Payne, M., C. A. Rupar, G. M. Siu, and V. M. Siu, 2011 Amish, Mennonite, and Hutterite genetic disorder database. Paediatr. Child Health 16: e23–e24.

Pemberton, T. J., C. Wang, J. Z. Li, and N. A. Rosenberg, 2010 Inference of unexpected genetic relatedness among individuals in HapMap Phase III. Am. J. Hum. Genet. 87: 457–464.

Pemberton, T. J., M. DeGiorgio, and N. A. Rosenberg, 2013 Population structure in a comprehensive data set on human microsatellite variation. G3 3: 909–916.

Reddy, S. B., and N. A. Rosenberg, 2012 Refining the relationship between homozygosity and the frequency of the most frequent allele. J. Math. Biol. 64: 87–108.

Reynolds, J., B. S. Weir, and C. C. Cockerham, 1983 Estimation of the coancestry coefficient: basis for a short-term genetic distance. Genetics 105: 767–779.

Shriver, M. D., G. C. Kennedy, E. J. Parra, H. A. Lawson, V. Sonpar *et al.*, 2004 The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. Hum. Genomics 1: 274–286.

Simonson, T. S., Y. Yang, C. D. Huff, H. Yun, G. Qin *et al.*, 2010 Genetic evidence for high-altitude adaptation in Tibet. Science 329: 72–75.

Solignac, M., D. Vautrin, A. Loiseau, F. Mougel, E. Baudry *et al.*, 2003 Five hundred and fifty microsatellite markers for the study of the honeybee (*Apis mellifera L.*) genome. Mol. Ecol. Notes 3: 307–311.

Sutter, N. B., C. D. Bustamante, K. Chase, M. M. Gray, K. Zhao *et al.*, 2007 A single IGF1 allele is a major determinant of small size in dogs. Science 316: 112–115.

Van Hout, C. V., A. M. Levin, E. Rampersaud, H. Shen, J. R. O'Connell *et al.*, 2010 Extent and distribution of linkage disequilibrium in the Old Order Amish. Genet. Epidemiol. 34: 146–150.

Wang, S., C. M. Lewis, Jr, M. Jakobsson, S. Ramachandran, N. Ray *et al.*, 2007 Genetic variation and population structure in Native Americans. PLoS Genet. 3: 2049–2067.

Wolter, K. M., 2007 *Introduction to Variance Estimation*, Ed. 2. Springer, New York, NY.

Wright, S., 1951 The genetical structure of populations. Ann. Eugen. 15: 323–354.

Yi, X., Y. Liang, E. Huerta-S'anchez, X. Jin, Z. X. P. Cuo *et al.*, 2010 Sequencing of 50 human exomes reveals adaptation to high altitude. Science 329: 75–78.

Zhang, W., Z. Fan, E. Han, R. Hou, L. Zhang *et al.*, 2014 Hypoxia adaptations in the grey wolf (*Canis lupus chanco*) from Qinghai-Tibet Plateau. PLoS Genet. 10: e1004466.

Communicating editor: B. J. Andrews

APPENDIX

Derivations of unbiased estimators of expected heterozygosity

In this section, we derive the general unbiased estimator of expected heterozygosity $\breve{H}$ for any unbiased linear estimator of population allele frequencies, defined in Proposition 1, and show how the formulas for $\widetilde{H}$ (DeGiorgio *et al.* 2010), and $\widetilde{H}_{\text{BLUE}}$ (Corollaries 2 and 3), emerge from specific cases of $\breve{H}$.

Proof of Proposition 1: We need to show that $\mathbb{E}[\breve{H}] = H$. Note that

$$\begin{aligned}\breve{p}_i^2 &= \sum_{j=1}^{n}\sum_{k=1}^{n} w_j w_k X_j^{(i)} X_k^{(i)} \\ &= \sum_{j=1}^{n}\sum_{k=1}^{n} \frac{w_j w_k}{m_j m_k} \sum_{\ell=1}^{m_j}\sum_{t=1}^{m_k} A_{j\ell}^{(i)} A_{kt}^{(i)}.\end{aligned}$$

Taking the expectation, we obtain

$$\begin{aligned}\mathbb{E}\left[\breve{p}_i^2\right] &= \sum_{j=1}^{n}\sum_{k=1}^{n} \frac{w_j w_k}{m_j m_k} \sum_{\ell=1}^{m_j}\sum_{t=1}^{m_k} \mathbb{E}\left[A_{j\ell}^{(i)} A_{kt}^{(i)}\right] \\ &= \sum_{j=1}^{n}\sum_{k=1}^{n} \frac{w_j w_k}{m_j m_k} \sum_{\ell=1}^{m_j}\sum_{t=1}^{m_k} \mathbb{P}\left[A_{j\ell}^{(i)} = 1, A_{kt}^{(i)} = 1\right] \\ &= \sum_{j=1}^{n}\sum_{k=1}^{n} \frac{w_j w_k}{m_j m_k} \sum_{\ell=1}^{m_j}\sum_{t=1}^{m_k} \left[(1-\Phi_{jk})p_i^2 + \Phi_{jk} p_i\right] \\ &= p_i^2 + \rho_2 p_i (1-p_i). \end{aligned} \tag{A1}$$

Therefore

$$\begin{aligned}\mathbb{E}[\breve{H}] &= \frac{1}{1-\rho_2}\left(1 - \sum_{i=1}^{I} \mathbb{E}\left[\breve{p}_i^2\right]\right) \\ &= \frac{1}{1-\rho_2}\left(1 - \sum_{i=1}^{I} \left[p_i^2 + \rho_2 p_i (1-p_i)\right]\right) \\ &= 1 - \sum_{i=1}^{I} p_i^2 = H. \end{aligned}$$

□

Proof of Corollary 2: We show that defining the weight of each individual in the calculation of ρ_2 in terms of an individual's relative allele copy contribution yields $\widetilde{H}$ from $\breve{H}$. Letting $w_k = m_k / \sum_{x=1}^{n} m_x$, we have that

$$\breve{p}_i = \sum_{k=1}^{n} \frac{m_k}{\sum_{j=1}^{n} m_j} X_k^{(i)} = \widehat{p}_i$$

and

$$\rho_2 = \sum_{j=1}^{n}\sum_{k=1}^{n} \frac{m_j}{\sum_{x=1}^{n} m_x} \frac{m_k}{\sum_{y=1}^{n} m_y} \Phi_{jk} = \bar{\Phi}_2.$$

Plugging in yields

$$\breve{H} = \frac{1}{1-\bar{\Phi}_2}\left(1 - \sum_{i=1}^{I} \widehat{p}_i^2\right) = \widetilde{H}.$$

□

Proof of Corollary 3: We show that defining the weight each individual according to their relative contribution to the inverted kinship matrix of the sample yields $\widetilde{H}_{\text{BLUE}}$ from $\breve{H}$. Letting $w_k = \sum_{j=1}^{n} (\mathbf{K}^{-1})_{jk} / \mathbf{1}^T \mathbf{K}^{-1} \mathbf{1}$, we have that

$$\breve{p}_i = \sum_{k=1}^{n} \frac{\sum_{j=1}^{n} (\mathbf{K}^{-1})_{jk}}{\mathbf{1}^T \mathbf{K}^{-1} \mathbf{1}} X_k^{(i)} = \widetilde{p}_i$$

and

$$\rho_2 = \sum_{j=1}^{n} \sum_{k=1}^{n} \frac{\sum_{x=1}^{n} (\mathbf{K}^{-1})_{xj}}{\mathbf{1}^T \mathbf{K}^{-1} \mathbf{1}} \frac{\sum_{y=1}^{n} (\mathbf{K}^{-1})_{yk}}{\mathbf{1}^T \mathbf{K}^{-1} \mathbf{1}} \Phi_{jk} = \kappa_2.$$

Plugging in yields

$$\breve{H} = \frac{1}{1-\kappa_2} \left(1 - \sum_{i=1}^{I} \widetilde{p}_i^2 \right) = \widetilde{H}_{\text{BLUE}}. \qquad \square$$

Derivations of variances of expected heterozygosity estimators

In this section, we summarize the procedure by which DeGiorgio *et al.* (2010) derived the equation for the variance of $\widetilde{H}$, illustrating the variance of the general case, $\breve{H}$. For the full derivation, see Appendix B of DeGiorgio *et al.* (2010). We then provide the specific formulation for the variance of $\widetilde{H}$ (Corollary 7) and $\widetilde{H}_{\text{BLUE}}$ (Corollary 8).

Abbreviated proof of Proposition 4: The variance of $\breve{H}$ (Equation 9) is defined as

$$\text{Var}\left[\breve{H}\right] = \frac{1}{(1-\rho_2)^2} \text{Var}\left[1 - \sum_{i=1}^{I} \breve{p}_i^2 \right].$$

By definition of variance, we get

$$\text{Var}\left[1 - \sum_{i=1}^{I} \breve{p}_i^2 \right] = \sum_{i=1}^{I} \text{Var}\left[\breve{p}_i^2\right] + 2 \sum_{i=1}^{I-1} \sum_{i'=i+1}^{I} \text{Cov}\left[\breve{p}_i^2, \breve{p}_{i'}^2\right],$$

with

$$\text{Var}\left[\breve{p}_i^2\right] = \mathbb{E}\left[\breve{p}_i^4\right] - \left(\mathbb{E}\left[\breve{p}_i^2\right]\right)^2$$

and

$$\text{Cov}\left[\breve{p}_i^2, \breve{p}_{i'}^2\right] = \mathbb{E}\left[\breve{p}_i^2 \breve{p}_{i'}^2\right] - \mathbb{E}\left[\breve{p}_i^2\right] \mathbb{E}\left[\breve{p}_{i'}^2\right].$$

Recalling that $\breve{p}_i = \sum_{j=1}^{n} \sum_{\ell=1}^{m_j} \frac{w_j}{m_j} A_{j\ell}^{(i)}$ for the ℓth allele copy of individual j, whose ploidy is m_j, we have that

$$\mathbb{E}\left[\breve{p}_i^4\right] = \sum_{j=1}^{n} \sum_{k=1}^{n} \sum_{j'=1}^{n} \sum_{k'=1}^{n} \sum_{\ell=1}^{m_j} \sum_{t=1}^{m_k} \sum_{\ell'=1}^{m_{j'}} \sum_{t'=1}^{m_{k'}} \frac{w_j w_k w_{j'} w_{k'}}{m_j m_k m_{j'} m_{k'}} \mathbb{E}\left[A_{j\ell}^{(i)} A_{kt}^{(i)} A_{j'\ell'}^{(i)} A_{k't'}^{(i)} \right],$$

and

$$\mathbb{E}\left[\breve{p}_i^2 \breve{p}_{i'}^2\right] = \sum_{j=1}^{n} \sum_{k=1}^{n} \sum_{j'=1}^{n} \sum_{k'=1}^{n} \sum_{\ell=1}^{m_j} \sum_{t=1}^{m_k} \sum_{\ell'=1}^{m_{j'}} \sum_{t'=1}^{m_{k'}} \frac{w_j w_k w_{j'} w_{k'}}{m_j m_k m_{j'} m_{k'}} \mathbb{E}\left[A_{j\ell}^{(i)} A_{kt}^{(i)} A_{j'\ell'}^{(i')} A_{k't'}^{(i')} \right]$$

for the case $i \neq i'$. We have previously shown in Equation A1 that $\mathbb{E}[\breve{p}_i^2] = p_i^2 + \rho_2 p_i (1 - p_i)$, and the value of $\mathbb{E}[\breve{p}_{i'}^2]$ similarly follows. Thus, we need to calculate $\mathbb{E}[A_{j\ell}^{(i)} A_{kt}^{(i)} A_{j'\ell'}^{(i)} A_{k't'}^{(i)}]$ and $\mathbb{E}[A_{j\ell}^{(i)} A_{kt}^{(i)} A_{j'\ell'}^{(i')} A_{k't'}^{(i')}]$ for the $i \neq i'$ case. These are

$$\begin{aligned}
\mathbb{E}\left[A_{j\ell}^{(i)} A_{kt}^{(i)} A_{j'\ell'}^{(i)} A_{k't'}^{(i)} \right] &= \Phi_{jkj'k'} p_i + \left[\Phi_{jkj'} + \Phi_{jkk'} + \Phi_{jj'k'} + \Phi_{kj'k'} + \Phi_{jk,j'k'} + \Phi_{jj',kk'} + \Phi_{jk',kj'} - 7\Phi_{jkj'k'} \right] p_i^2 \\
&+ \Big[12\Phi_{jkj'k'} + \Phi_{jk} + \Phi_{jj'} + \Phi_{jk'} + \Phi_{kj'} + \Phi_{kk'} + \Phi_{j'k'} - 3\left(\Phi_{jkj'} + \Phi_{jkk'} + \Phi_{jj'k'} + \Phi_{kj'k'} \right) \\
&- 2\left(\Phi_{jk,j'k'} + \Phi_{jj,kk'} + \Phi_{jk,kj'} \right) \Big] p_i^3 + \Big[1 + \Phi_{jk,j'k'} + \Phi_{jj',kk'} + \Phi_{jk',kj'} \\
&+ 2\left(\Phi_{jkj'} + \Phi_{jkk'} + \Phi_{jj'k'} + \Phi_{kj'k'} \right) - 6\Phi_{jkj'k'} - \left(\Phi_{jk} + \Phi_{jj'} + \Phi_{jk'} + \Phi_{kj'} + \Phi_{kk'} + \Phi_{j'k'} \right) \Big] p_i^4
\end{aligned} \tag{A2}$$

and

$$\mathbb{E}\left[A_{j\ell}^{(i)}A_{kt}^{(i)}A_{j'\ell'}^{(i')}A_{k't'}^{(i')}\right] = \left[\Phi_{jk,j'k'} - \Phi_{jkj'k'}\right]p_ip_{i'} + \left[2\Phi_{jkj'k'} + \Phi_{jk} - \left(\Phi_{jkj'} + \Phi_{jkk'}\right) - \Phi_{jk,j'k'}\right]p_ip_{i'}^2 + \left[2\Phi_{jkj'k'} + \Phi_{j'k'} - \left(\Phi_{jj'k'} + \Phi_{kj'k'}\right) - \Phi_{jk,j'k'}\right]p_i^2p_{i'} + \left[1 + \Phi_{jk,j'k'} + \Phi_{jj',kk'} + \Phi_{jk',kj'} + 2\left(\Phi_{jkj'} + \Phi_{jkk'} + \Phi_{jj'k'} + \Phi_{kj'k'}\right) - 6\Phi_{jkj'k'} - \left(\Phi_{jk} + \Phi_{jj'} + \Phi_{jk'} + \Phi_{kj'} + \Phi_{kk'} + \Phi_{j'k'}\right)\right]p_i^2p_{i'}^2. \quad \text{(A3)}$$

Substituting Equation A2 into $\mathbb{E}[\breve{p}_i^4]$ and solving for $\mathrm{Var}[\breve{p}_i^2]$, we obtain

$$\mathrm{Var}\left[\breve{p}_i^2\right] = \rho_4p_i + \left(4\rho_3 + 3\rho_{2,2} - 7\rho_4 - \rho_2^2\right)p_i^2 + \left(12\rho_4 + 4\rho_2 + 2\rho_2^2 - 12\rho_3 - 6\rho_{2,2}\right)p_i^3 + \left(3\rho_{2,2} + 8\rho_3 - 6\rho_4 - 4\rho_2 - \rho_2^2\right)p_i^4,$$

and, substituting Equation A3 into $\mathbb{E}[\breve{p}_i^2\breve{p}_{i'}^2]$, and solving for $\mathrm{Cov}[\breve{p}_i^2, \breve{p}_{i'}^2]$, we obtain

$$\mathrm{Cov}\left[\breve{p}_i^2, \breve{p}_{i'}^2\right] = \left(\rho_{2,2} - \rho_4 - \rho_2^2\right)p_ip_{i'} + \left(2\rho_4 + \rho_2^2 - 2\rho_3 - \rho_{2,2}\right)p_ip_{i'}^2 + \left(2\rho_4 + \rho_2^2 - 2\rho_3 - \rho_{2,2}\right)p_i^2p_{i'} + \left(3\rho_{2,2} + 8\rho_3 - 6\rho_4 - 4\rho_2 - \rho_2^2\right)p_i^2p_{i'}^2.$$

Thus, substituting the values for variance and covariance into the definition of variance, we have

$$\mathrm{Var}\left[\breve{H}\right] = \frac{1}{(1-\rho_2)^2}\left[\rho_{2,2} - \rho_2^2 + 2\left(\rho_2^2 - \rho_4\right)\sum_{i=1}^{I}p_i^2 + 4\left(2\rho_4 + \rho_2 - 2\rho_3 - \rho_{2,2}\right)\sum_{i=1}^{I}p_i^3 + \left(3\rho_{2,2} + 8\rho_3 - 6\rho_4 - 4\rho_2 - \rho_2^2\right)\left(\sum_{i=1}^{I}p_i^2\right)^2\right]. \quad \square$$

Corollary 7: Consider a locus with I distinct alleles, and parametric allele frequencies $p_i \in [0, 1]$, $i = 1, 2, \ldots, I$, and $\sum_{i=1}^{I}p_i = 1$. For a sample of size n individuals of any ploidy, inbreeding status, and relatedness,

$$\mathrm{Var}\left[\widetilde{H}\right] = \frac{1}{\left(1-\bar{\Phi}_2\right)^2}\mathrm{Var}\left[1 - \sum_{i=1}^{I}\widehat{p}_i^2\right] \quad \text{(A4)}$$

and

$$\mathrm{Var}\left[1 - \sum_{i=1}^{I}\widehat{p}_i^2\right] = \bar{\Phi}_{2,2} - \bar{\Phi}_2^2 + 2\left(\bar{\Phi}_2^2 - \bar{\Phi}_4\right)\sum_{i=1}^{I}p_i^2 + 4\left(2\bar{\Phi}_4 + \bar{\Phi}_2 - 2\bar{\Phi}_3 - \bar{\Phi}_{2,2}\right)\sum_{i=1}^{I}p_i^3 + \left(3\bar{\Phi}_{2,2} + 8\bar{\Phi}_3 - 6\bar{\Phi}_4 - 4\bar{\Phi}_2 - \bar{\Phi}_2^2\right)\left(\sum_{i=1}^{I}p_i^2\right)^2, \quad \text{(A5)}$$

where $\bar{\Phi}_2$, $\bar{\Phi}_3$, $\bar{\Phi}_4$, and $\bar{\Phi}_{2,2}$ are mean kinship coefficients, weighted by the contribution of individuals to the number of allele copies in the sample, with subscripts corresponding to the number of individuals considered for the calculation. Additionally,

$$\mathrm{Var}\left[\widetilde{H}\right] \approx 4\bar{\Phi}_2\left[\sum_{i=1}^{I}p_i^3 - \left(\sum_{i=1}^{I}p_i^2\right)^2\right]. \quad \text{(A6)}$$

The proof of Corollary 7 follows from the proof of Proposition 4, where $\widehat{p}_i$ is substituted for $\breve{p}_i$, and $\bar{\Phi}_2$, $\bar{\Phi}_3$, $\bar{\Phi}_4$, and $\bar{\Phi}_{2,2}$ are substituted for ρ_2, ρ_3, ρ_4, and $\rho_{2,2}$, respectively.

Corollary 8: Consider a locus with I distinct alleles and parametric allele frequencies $p_i \in [0, 1]$, $i = 1, 2, \ldots, I$, and $\sum_{i=1}^{I}p_i = 1$. For a sample of size n individuals of any ploidy, inbreeding status, and relatedness,

$$\mathrm{Var}\left[\widetilde{H}_{\mathrm{BLUE}}\right] = \frac{1}{(1-\kappa_2)^2}\mathrm{Var}\left[1 - \sum_{i=1}^{I}\widetilde{p}_i^2\right] \quad \text{(A7)}$$

and

$$\mathrm{Var}\left[1-\sum_{i=1}^{I}\widetilde{p}_i^2\right]=\kappa_{2,2}-\kappa_2^2+2\left(\kappa_2^2-\kappa_4\right)\sum_{i=1}^{I}p_i^2+4\left(2\kappa_4+\kappa_2-2\kappa_3-\kappa_{2,2}\right)\sum_{i=1}^{I}p_i^3 +\left(3\kappa_{2,2}+8\kappa_3-6\kappa_4-4\kappa_2-\kappa_2^2\right)\left(\sum_{i=1}^{I}p_i^2\right)^2, \tag{A8}$$

where κ_2, κ_3, κ_4, and $\kappa_{2,2}$ are mean kinship coefficients, weighted by the contribution of individuals to the inverted kinship matrix, with subscripts corresponding to the number of individuals considered for the calculation. Additionally,

$$\mathrm{Var}\left[\widetilde{H}_{\mathrm{BLUE}}\right]\approx 4\kappa_2\left[\sum_{i=1}^{I}p_i^3-\left(\sum_{i=1}^{I}p_i^2\right)^2\right]. \tag{A9}$$

The proof of Corollary 8 follows from the proof of Proposition 4, where $\widetilde{p}_i$ is substituted for $\breve{p}_i$, and κ_2, κ_3, κ_4, and $\kappa_{2,2}$ are substituted for ρ_2, ρ_3, ρ_4, and $\rho_{2,2}$, respectively.

Derivations of bias measurements in the application of $\widehat{H}$

For samples containing related and inbred individuals, $\widehat{H}$ has a downward bias, which is defined in Equation 12 for the general estimator of population allele frequency $\breve{p}_i$. Here, we present Corollaries 9 and 10 for the specific estimators of population allele frequency $\widehat{p}_i$ and $\widetilde{p}_i$, respectively.

Corollary 9: Consider a locus with I distinct alleles and parametric allele frequencies $p_i\in[0,1]$, $i=1,2,\ldots,I$, and $\sum_{i=1}^{I}p_i=1$. For a sample of size n possibly related or inbred individuals, the bias of the estimator of expected heterozygosity $\widehat{H}$ changes with the true locus expected heterozygosity such that

$$\mathrm{Bias}\left[\widehat{H}\left(\widehat{p}_i\right)\right]=\frac{1-n\bar{\Phi}_2}{n-1}H, \tag{A10}$$

where

$$\widehat{H}\left(\widehat{p}_i\right)=\frac{n}{n-1}\left(1-\sum_{i=1}^{I}\widehat{p}_i^2\right).$$

As this is the standard application of $\widehat{H}$ (Equation 2), Equation A10 describes the bias of $\widehat{H}$ in the *Results*. However, $\widehat{H}$ is biased with any unbiased linear estimator of allele frequency for samples containing related or inbred individuals. The proof of Corollary 9 follows from the proof of Proposition 5, where $\bar{\Phi}_2$ is substituted for ρ_2.

Corollary 10: Consider a locus with I distinct alleles and parametric allele frequencies $p_i\in[0,1]$, $i=1,2,\ldots,I$, and $\sum_{i=1}^{I}p_i=1$. For a sample of size n possibly related or inbred individuals, the bias of the estimator of expected heterozygosity $\widehat{H}$ changes with the true locus expected heterozygosity such that

$$\mathrm{Bias}\left[\widehat{H}\left(\widetilde{p}_i\right)\right]=\frac{1-n\kappa_2}{n-1}H, \tag{A11}$$

where

$$\widehat{H}\left(\widetilde{p}_i\right)=\frac{n}{n-1}\left(1-\sum_{i=1}^{I}\widetilde{p}_i^2\right). \tag{A12}$$

The proof of Corollary 10 follows from the proof of Proposition 5, where κ_2 is substituted for ρ_2.

Derivations of components for the variance of F_{ST} estimators

In this final section of the Appendix, we provide derivations for the components of Equations 16 and 17, which describe the variance of $\breve{F}_{\mathrm{ST}}$. We derive the variance of $\breve{H}_{12}$, as well as the covariances of $\breve{H}_{12}$ with $\breve{H}_1$ (and interchangeably, $\breve{H}_{12}$ with $\breve{H}_2$), and of $\left[\breve{H}_{12}-\frac{1}{2}\breve{H}_1-\frac{1}{2}\breve{H}_2\right]$ with $\breve{H}_{12}$. Because the complete expression for $\mathrm{Var}[\breve{F}_{\mathrm{ST}}]$ is unwieldy, we stop at the derivation of the final component.

Lemma 11: Consider a locus with I distinct alleles across two independent populations and parametric allele frequencies $p_i\in[0,1]$, $i=1,2,\ldots,I$, and $\sum_{i=1}^{I}p_i=1$ for population 1, and $q_i\in[0,1]$, $i=1,2,\ldots,I$, and $\sum_{i=1}^{I}q_i=1$ for population 2. For two samples of size n_1 and n_2, individuals from populations 1 and 2, respectively, each with individuals of any ploidy, inbreeding status, and relatedness,

$$\mathrm{Var}\left[\breve{H}_{12}\right]=\rho_2^{(1)}\left(1-\rho_2^{(2)}\right)\sum_{i=1}^{I}p_iq_i^2+\rho_2^{(2)}\left(1-\rho_2^{(1)}\right)\sum_{i=1}^{I}p_i^2q_i+\rho_2^{(1)}\rho_2^{(2)}\sum_{i=1}^{I}p_iq_i+\left(\rho_2^{(1)}\rho_2^{(2)}-\rho_2^{(1)}-\rho_2^{(2)}\right)\left(\sum_{i=1}^{I}p_iq_i\right)^2, \tag{A13}$$

where the superscript of the mean kinship coefficient ρ_2 corresponds to the population for which it is calculated. The equations for the variance of $\widetilde{H}_{12}$ and $\widetilde{H}_{12,\mathrm{BLUE}}$ are obtained by substituting $\bar{\Phi}_2$ and κ_2, respectively, into Equation A13 as the mean kinship coefficients in place of ρ_2.

Proof: By definition of variance,

$$\mathrm{Var}\left[\breve{H}_{12}\right] = \sum_{i=1}^{I} \mathrm{Var}\left[\breve{p}_i\breve{q}_i\right] + 2\sum_{i=1}^{I-1}\sum_{i'=i+1}^{I} \mathrm{Cov}\left[\breve{p}_i\breve{q}_i, \breve{p}_{i'}\breve{q}_{i'}\right]$$

where

$$\mathrm{Var}\left[\breve{p}_i\breve{q}_i\right] = \mathbb{E}\left[\breve{p}_i^2\,\breve{q}_i^2\right] - \left(\mathbb{E}\left[\breve{p}_i\,\breve{q}_i\right]\right)^2$$

and

$$\mathrm{Cov}\left[\breve{p}_i\breve{q}_i, \breve{p}_{i'}\breve{q}_{i'}\right] = \mathbb{E}\left[\breve{p}_i\breve{q}_i\breve{p}_{i'}\breve{q}_{i'}\right] - \mathbb{E}\left[\breve{p}_i\breve{q}_i\right]\mathbb{E}\left[\breve{p}_{i'}\breve{q}_{i'}\right].$$

Because $\breve{p}_i$ and $\breve{q}_i$ are unbiased estimators of population allele frequency, and populations 1 and 2 are independent,

$$\mathbb{E}\left[\breve{p}_i\breve{q}_i\right] = p_i q_i$$

Similarly, $\mathbb{E}[\breve{p}_{i'}\breve{q}_{i'}] = p_{i'}q_{i'}$ Next, we have

$$\begin{aligned}\mathbb{E}\left[\breve{p}_i^2\breve{q}_i^2\right] &= \mathbb{E}\left[\breve{p}_i^2\right]\mathbb{E}\left[\breve{q}_i^2\right] \\ &= \left[p_i^2 + \rho_2^{(1)}p_i(1-p_i)\right]\left[q_i^2 + \rho_2^{(2)}q_i(1-q_i)\right] \\ &= p_i q_i\left[p_i + \rho_2^{(1)}(1-p_i)\right]\left[q_i + \rho_2^{(2)}(1-q_i)\right],\end{aligned} \tag{A14}$$

where $\mathbb{E}[\breve{q}_i^2]$ takes the same form as $\mathbb{E}[\breve{p}_i^2]$ (Equation A1), except that the resulting weighted mean kinship coefficient ρ_2 is for population 2, indicated by the superscript. By substituting Equation A14 into $\mathrm{Var}[\breve{p}_i\breve{q}_i]$, we have

$$\begin{aligned}\mathrm{Var}\left[\breve{p}_i\breve{q}_i\right] &= p_i q_i\left[p_i + \rho_2^{(1)}(1-p_i)\right]\left[q_i + \rho_2^{(2)}(1-q_i)\right] - \left(p_i q_i\right)^2 \\ &= p_i q_i\left\{\left[p_i + \rho_2^{(1)}(1-p_i)\right]\left[q_i + \rho_2^{(2)}(1-q_i)\right] - p_i q_i\right\} \\ &= p_i q_i\left[\rho_2^{(1)}(1-p_i)q_i + \rho_2^{(2)}p_i(1-q_i) + \rho_2^{(1)}\rho_2^{(2)}(1-p_i)(1-q_i)\right].\end{aligned} \tag{A15}$$

We now derive an expression for $\mathrm{Cov}[\breve{p}_i\breve{q}_i, \breve{p}_{i'}\breve{q}_{i'}]$ Let $B_{kt}^{(i)}$ be an indicator random variable in population 2 analogous to the indicator random variable $A_{j\ell}^{(i)}$, which we have previously defined for population 1.

$$\begin{aligned}\mathbb{E}\left[\breve{p}_i\breve{q}_i\breve{p}_{i'}\breve{q}_{i'}\right] &= \mathbb{E}\left[\breve{p}_i\breve{p}_i\right]\mathbb{E}\left[\breve{q}_i\breve{q}_{i'}\right] \\ &= \left(\sum_{j=1}^{n_1}\sum_{j'=1}^{n_1}\sum_{\ell=1}^{m_j}\sum_{\ell'=1}^{m_{j'}}\frac{w_j w_{j'}}{m_j m_{j'}}\mathbb{E}\left[A_{j\ell}^{(i)}A_{j'\ell'}^{(i')}\right]\right)\left(\sum_{k=1}^{n_2}\sum_{k'=1}^{n_2}\sum_{t=1}^{m_k}\sum_{t'=1}^{m_{k'}}\frac{w_k w_{k'}}{m_k m_{k'}}\mathbb{E}\left[B_{kt}^{(i)}B_{k't'}^{(i')}\right]\right),\end{aligned}$$

where

$$\mathbb{E}\left[A_{j\ell}^{(i)}A_{j'\ell'}^{(i')}\right] = \mathbb{P}\left[A_{j\ell}^{(i)} = 1, A_{j'\ell'}^{(i')} = 1\right],$$

and

$$\mathbb{E}\left[B_{kt}^{(i)}B_{k't'}^{(i')}\right] = \mathbb{P}\left[B_{kt}^{(i)} = 1, B_{k't'}^{(i')} = 1\right].$$

Consider a scenario in which we have two allele copies. Let s_1 be the identity state with probability Δ_1, in which two randomly drawn alleles are not IBD, and s_2 be the identity state occurring with probability $\Delta_2 = 1 - \Delta_1$, in which the two alleles are IBD.

$$\begin{aligned}\mathbb{P}\left[A_{j\ell}^{(i)}=1,A_{j'\ell'}^{(i')}=1\right]&=\mathbb{P}\left[A_{j\ell}^{(i)}=1,A_{j'\ell'}^{(i')}=1|s_1\right]\mathbb{P}[s_1]+\mathbb{P}\left[A_{j\ell}^{(i)}=1,A_{j'\ell'}^{(i')}=1|s_2\right]\mathbb{P}[s_2]\\&=p_ip_{i'}\Delta_1+0\times\Delta_2\\&=\Delta_1p_ip_{i'}\end{aligned}$$

Note that, because $\Delta_1+\Delta_2=1$ and $\Phi_{jj'}^{(1)}=\Delta_2$ (same with $\Phi_{kk'}^{(2)}$), we have $\Delta_1=1-\Phi_{jj'}^{(1)}$. Thus,

$$\mathbb{P}\left[A_{j\ell}^{(i)}=1,A_{j'\ell'}^{(i')}=1\right]=\left(1-\Phi_{jj'}^{(1)}\right)p_ip_{i'}$$

and

$$\mathbb{P}\left[B_{kt}^{(i)}=1,B_{k't'}^{(i')}=1\right]=\left(1-\Phi_{kk'}^{(2)}\right)q_iq_{i'}$$

Substituting, we now have

$$\begin{aligned}\mathbb{E}\left[\breve{p}_i\breve{p}_{i'}\breve{q}_i\breve{q}_{i'}\right]&=\left(\sum_{j=1}^{n_1}\sum_{j'=1}^{n_1}\sum_{\ell=1}^{m_j}\sum_{\ell'=1}^{m_{j'}}\frac{w_jw_{j'}}{m_jm_{j'}}\left(1-\Phi_{jj'}^{(1)}\right)p_ip_{i'}\right)\left(\sum_{k=1}^{n_2}\sum_{k'=1}^{n_2}\sum_{t=1}^{m_k}\sum_{t'=1}^{m_{k'}}\frac{w_kw_{k'}}{m_km_{k'}}\left(1-\Phi_{kk'}^{(2)}\right)q_iq_{i'}\right)\\&=\left(1-\rho_2^{(1)}\right)\left(1-\rho_2^{(2)}\right)p_ip_{i'}q_iq_{i'},\end{aligned}\tag{A16}$$

and substituting Equation A16 into $\mathrm{Cov}[\breve{p}_i\breve{q}_i,\breve{p}_{i'}\breve{q}_{i'}]$ yields

$$\begin{aligned}\mathrm{Cov}\left[\breve{p}_i\breve{q}_i,\breve{p}_{i'}\,\breve{q}_{i'}\right]&=\left(1-\rho_2^{(1)}\right)\left(1-\rho_2^{(2)}\right)p_ip_{i'}q_iq_{i'}-p_iq_ip_{i'}q_{i'}.\\&=\left(\rho_2^{(1)}\rho_2^{(2)}-\rho_2^{(1)}-\rho_2^{(2)}\right)p_ip_{i'}q_iq_{i'}.\end{aligned}\tag{A17}$$

Therefore, using Equations A15 and A17,

$$\begin{aligned}\mathrm{Var}\left[\breve{H}_{12}\right]&=\sum_{i=1}^{I}p_iq_i\left[\rho_2^{(1)}(1-p_i)q_i+\rho_2^{(2)}p_i(1-q_i)+\rho_2^{(1)}\rho_2^{(2)}(1-p_i)(1-q_i)\right]\\&\quad+2\sum_{i=1}^{I-1}\sum_{i'=i+1}^{I}\left(\rho_2^{(1)}\rho_2^{(2)}-\rho_2^{(1)}-\rho_2^{(2)}\right)p_ip_{i'}q_iq_{i'}\\&=\rho_2^{(1)}\sum_{i=1}^{I}p_i(1-p_i)q_i^2+\rho_2^{(2)}\sum_{i=1}^{I}p_i^2q_i(1-q_i)+\rho_2^{(1)}\rho_2^{(2)}\sum_{i=1}^{I}p_i(1-p_i)q_i(1-q_i)+2\sum_{i=1}^{I-1}\sum_{i'=i+1}^{I}\left(\rho_2^{(1)}\rho_2^{(2)}-\rho_2^{(1)}-\rho_2^{(2)}\right)p_ip_{i'}q_iq_{i'}\\&=\rho_2^{(1)}\left(1-\rho_2^{(2)}\right)\sum_{i=1}^{I}p_iq_i^2+\rho_2^{(2)}\left(1-\rho_2^{(1)}\right)\sum_{i=1}^{I}p_i^2q_i+\left(\rho_2^{(1)}\rho_2^{(2)}-\rho_2^{(1)}-\rho_2^{(2)}\right)\sum_{i=1}^{I}p_i^2q_i^2\\&\quad+\rho_2^{(1)}\rho_2^{(2)}\sum_{i=1}^{I}p_iq_i+2\left(\rho_2^{(1)}\rho_2^{(2)}-\rho_2^{(1)}-\rho_2^{(2)}\right)\sum_{i=1}^{I}\sum_{i'=1}^{I}p_ip_{i'}q_iq_{i'}\\&=\rho_2^{(1)}\left(1-\rho_2^{(2)}\right)\sum_{i=1}^{I}p_iq_i^2+\rho_2^{(2)}\left(1-\rho_2^{(1)}\right)\sum_{i=1}^{I}p_i^2q_i+\rho_2^{(1)}\rho_2^{(2)}\sum_{i=1}^{I}p_iq_i+\left(\rho_2^{(1)}\rho_2^{(2)}-\rho_2^{(1)}-\rho_2^{(2)}\right)\left(\sum_{i=1}^{I}p_iq_i\right)^2.\end{aligned}$$

□

Lemma 12: Consider a locus with I distinct alleles across two independent populations and parametric allele frequencies $p_i\in[0,1]$, $i=1,2,\ldots,I$, and $\sum_{i=1}^{I}p_i=1$ for population 1, or $q_i\in[0,1]$, $i=1,2,\ldots,I$, and $\sum_{i=1}^{I}q_i=1$ for population 2. For two samples of size n_1 and n_2 individuals from populations 1 and 2, respectively, each with individuals of any ploidy, inbreeding status, and relatedness,

$$\mathrm{Cov}\left[\breve{H}_{12},\breve{H}_1\right]=\frac{1}{1-\rho_2^{(1)}}\left[2\left(\rho_3^{(1)}-\rho_2^{(1)}\right)\sum_{i=1}^{I}p_i^3q_i+\left(2\rho_2^{(1)}-3\rho_3^{(1)}\right)\sum_{i=1}^{I}p_i^2q_i+\rho_3^{(1)}\sum_{i=1}^{I}p_iq_i\right]\tag{A18}$$

and

$$\mathrm{Cov}\left[\breve{H}_{12},\breve{H}_2\right]=\frac{1}{1-\rho_2^{(2)}}\left[2\left(\rho_3^{(2)}-\rho_2^{(2)}\right)\sum_{i=1}^{I}p_iq_i^3+\left(2\rho_2^{(2)}-3\rho_3^{(2)}\right)\sum_{i=1}^{I}p_iq_i^2+\rho_3^{(2)}\sum_{i=1}^{I}p_iq_i\right], \tag{A19}$$

where the superscript of the mean kinship coefficients ρ_2 and ρ_3 corresponds to the population for which these are calculated. The formulas for $\mathrm{Cov}[\widetilde{H}_{12},\widetilde{H}_1]$, $\mathrm{Cov}[\widetilde{H}_{12},\widetilde{H}_2]$, $\mathrm{Cov}[\widetilde{H}_{12,\mathrm{BLUE}},\widetilde{H}_{1,\mathrm{BLUE}}]$, and $\mathrm{Cov}[\widetilde{H}_{12,\mathrm{BLUE}},\widetilde{H}_{2,\mathrm{BLUE}}]$ are obtained by substituting $\bar{\Phi}_2$ and $\bar{\Phi}_3$ (for $\widetilde{H}$), or κ_2 and κ_3 (for $\widetilde{H}_{\mathrm{BLUE}}$) into Equations A18 and A19, respectively.

Proof: The covariance between $\breve{H}_{12}$ and $\breve{H}_1$ is

$$\begin{aligned}\mathrm{Cov}\left[\breve{H}_{12},\breve{H}_1\right]&=\frac{1}{1-\rho_2^{(1)}}\mathrm{Cov}\left[\left(1-\sum_{i=1}^{I}\breve{p}_i\breve{q}_i\right),\left(1-\sum_{i=1}^{I}\breve{p}_i^2\right)\right]\\&=\frac{1}{1-\rho_2^{(1)}}\sum_{i=1}^{I}\sum_{i'=1}^{I}\mathrm{Cov}\left[\breve{p}_i\breve{q}_i,\breve{p}_{i'}^2\right]\\&=\frac{1}{1-\rho_2^{(1)}}\left(\sum_{i=1}^{I}\mathrm{Cov}\left[\breve{p}_i\breve{q}_i,\breve{p}_i^2\right]+\sum_{i=1}^{I}\sum_{\substack{i'=1\\i'\neq i}}^{I}\mathrm{Cov}\left[\breve{p}_i\breve{q}_i,\breve{p}_{i'}^2\right]\right).\end{aligned}$$

The value of the covariance calculated for the case where $i=i'$ can be written as

$$\mathrm{Cov}\left[\breve{p}_i\breve{q}_i,\breve{p}_i^2\right]=\mathbb{E}\left[\breve{p}_i^3\breve{q}_i\right]-\mathbb{E}\left[\breve{p}_i\breve{q}_i\right]\mathbb{E}\left[\breve{p}_i^2\right].$$

From the proof of Lemma 11, we have derived the value of $\mathbb{E}[\breve{p}_i\breve{q}_i]$, and, from the proof of Proposition 1 we know the value for $\mathbb{E}[\breve{p}_i^2]$ We therefore only need to compute

$$\mathbb{E}\left[\breve{p}_i^3\breve{q}_i\right]=\mathbb{E}\left[\breve{p}_i^3\right]\mathbb{E}\left[\breve{q}_i\right],$$

where $\mathbb{E}[\breve{q}_i]=q_i$. Solving for $\mathbb{E}[\breve{p}_i^3]$, we have

$$\begin{aligned}\mathbb{E}\left[\breve{p}_i^3\right]&=\sum_{j=1}^{n_1}\sum_{v=1}^{n_1}\sum_{v'=1}^{n_1}\sum_{\ell=1}^{m_j}\sum_{z=1}^{m_v}\sum_{z'=1}^{m_{v'}}\frac{w_jw_vw_{v'}}{m_jm_vm_{v'}}\mathbb{E}\left[A_{j\ell}^{(i)}A_{vz}^{(i)}A_{v'z'}^{(i)}\right]\\&=\sum_{j=1}^{n_1}\sum_{v=1}^{n_1}\sum_{v'=1}^{n_1}\sum_{\ell=1}^{m_j}\sum_{z=1}^{m_v}\sum_{z'=1}^{m_{v'}}\frac{w_jw_vw_{v'}}{m_jm_vm_{v'}}\mathbb{P}\left[A_{j\ell}^{(i)}=1,A_{vz}^{(i)}=1,A_{v'z'}^{(i)}=1\right].\end{aligned}$$

The value of $\mathbb{P}[A_{j\ell}^{(i)}=1,A_{vz}^{(i)}=1,A_{v'z'}^{(i)}=1]$ depends on the probabilities of distinct identity states in which three alleles are drawn from the sample (one each from individuals j, v, and v'). We define state 1 as no IBD alleles drawn (probability δ_1), state 2 as IBD alleles drawn from j and v (probability δ_2), state 3 as IBD alleles drawn from v and v' IBD (probability δ_3), state 4 as IBD alleles drawn from j and v' IBD (probability δ_4), and state 5 as all three IBD (probability δ_5), with $\sum_{s=1}^{5}\delta_s=1$. Thus, the probabilities for the relevant kinship coefficients are

$$\Phi_{jvv'}^{(1)}=\delta_5$$

$$\Phi_{jv}^{(1)}=\delta_5+\delta_2$$

$$\Phi_{vv'}^{(1)}=\delta_5+\delta_3$$

$$\Phi_{jv'}^{(1)}=\delta_5+\delta_4,$$

which yields

$$\begin{aligned}\mathbb{P}\left[A_{j\ell}^{(i)}=1,A_{vz}^{(i)}=1,A_{v'z'}^{(i)}=1\right]&=\delta_5p_i+(\delta_2+\delta_3+\delta_4)p_i^2+\delta_1p_i^3\\&=\Phi_{jvv'}^{(1)}p_i+\left(\Phi_{jv}^{(1)}+\Phi_{vv'}^{(1)}+\Phi_{jv'}^{(1)}-3\Phi_{jvv'}^{(1)}\right)p_i^2+\left(1+2\Phi_{jvv'}^{(1)}-\Phi_{jv'}^{(1)}-\Phi_{vv'}^{(1)}-\Phi_{jv}^{(1)}\right)p_i^3.\end{aligned}$$

Thus, $\mathbb{E}[\breve{p}_i^3\breve{q}_i]$ is

$$\mathbb{E}\left[\breve{p}_i^3\breve{q}_i\right] = \rho_3^{(1)}p_iq_i + 3\left(\rho_2^{(1)} - \rho_3^{(1)}\right)p_i^2q_i + \left(1 + 2\rho_3^{(1)} - 3\rho_2^{(1)}\right)p_i^3q_i, \tag{A20}$$

and from Equations A20 and A1, and the definition of $\mathbb{E}[\breve{p}_i\breve{q}_i]$,

$$\begin{aligned}\mathrm{Cov}\left[\breve{p}_i\breve{q}_i, \breve{p}_i^2\right] &= \rho_3^{(1)}p_iq_i + 3\left(\rho_2^{(1)} - \rho_3^{(1)}\right)p_i^2q_i + \left(1 + 2\rho_3^{(1)} - 3\rho_2^{(1)}\right)p_i^3q_i - p_iq_i\left[p_i^2 + \rho_2^{(1)}p_i(1-p_i)\right] \\ &= 2\left(\rho_3^{(1)} - \rho_2^{(1)}\right)p_i^3q_i + \left(2\rho_2^{(1)} - 3\rho_3^{(1)}\right)p_i^2q_i + \rho_3^{(1)}p_iq_i.\end{aligned} \tag{A21}$$

Meanwhile, for the $\mathrm{Cov}[\breve{p}_i\breve{q}_i, \breve{p}_{i'}^2]$ case of $i \neq i'$, $\mathrm{Cov}[\breve{p}_i\breve{q}_i, \breve{p}_{i'}^2] = 0$. This is intuitively sensible because the products $\breve{p}_i\breve{q}_i$ and $\breve{p}_{i'}^2$ are independent, describing different alleles, and should not covary.

Finally, we can see that, when the two populations considered are independent from one another, the value of $\mathrm{Cov}[\breve{H}_{12}, \breve{H}_1]$ (or equivalently of $\mathrm{Cov}[\breve{H}_{12}, \breve{H}_2]$) is driven entirely by the case in which $i = i'$, such that

$$\begin{aligned}\mathrm{Cov}\left[\breve{H}_{12}, \breve{H}_1\right] &= \frac{1}{1-\rho_2^{(1)}}\sum_{i=1}^{I}\mathrm{Cov}\left[\breve{p}_i\breve{q}_i, \breve{p}_i^2\right] \\ &= \frac{1}{1-\rho_2^{(1)}}\left[2\left(\rho_3^{(1)} - \rho_2^{(1)}\right)\sum_{i=1}^{I}p_i^3q_i + \left(2\rho_2^{(1)} - 3\rho_3^{(1)}\right)\sum_{i=1}^{I}p_i^2q_i + \rho_3^{(1)}\sum_{i=1}^{I}p_iq_i\right].\end{aligned}$$

□

We now need to derive $\mathrm{Cov}\left[\breve{H}_{12} - \frac{1}{2}(\breve{H}_1 + \breve{H}_2), \breve{H}_{12}\right]$, the final term required to compute $\mathrm{Var}[\breve{F}_{\mathrm{ST}}]$

Lemma 13: Consider a locus with I distinct alleles across two independent populations and parametric allele frequencies $p_i \in [0,1]$, $i = 1, 2, \ldots, I$, and $\sum_{i=1}^{I}p_i = 1$ for population 1, or $q_i \in [0,1]$, $i = 1, 2, \ldots, I$, and $\sum_{i=1}^{I}q_i = 1$ for population 2. For two samples of size n_1 and n_2 individuals from populations 1 and 2, respectively, each with individuals of any ploidy, inbreeding status, and relatedness,

$$\begin{aligned}\mathrm{Cov}\left[\breve{H}_{12} - \frac{1}{2}\breve{H}_1 - \frac{1}{2}\breve{H}_2, \breve{H}_{12}\right] = &\left[\rho_2^{(1)}\left(1-\rho_2^{(2)}\right) + \frac{3\rho_3^{(2)} - 2\rho_2^{(2)}}{2\left(1-\rho_2^{(2)}\right)}\right]\sum_{i=1}^{I}p_iq_i^2 + \left[\rho_2^{(2)}\left(1-\rho_2^{(1)}\right) + \frac{3\rho_3^{(1)} - 2\rho_2^{(1)}}{2\left(1-\rho_2^{(1)}\right)}\right]\sum_{i=1}^{I}p_i^2q_i \\ &+ \left[\rho_2^{(1)}\rho_2^{(2)} - \frac{\rho_3^{(1)}}{2\left(1-\rho_2^{(1)}\right)} - \frac{\rho_3^{(2)}}{2\left(1-\rho_2^{(2)}\right)}\right]\sum_{i=1}^{I}p_iq_i + \left(\rho_2^{(1)}\rho_2^{(2)} - \rho_2^{(1)} - \rho_2^{(2)}\right)\left(\sum_{i=1}^{I}p_iq_i\right)^2 \\ &+ \frac{\rho_2^{(1)} - \rho_3^{(1)}}{1-\rho_2^{(1)}}\sum_{i=1}^{I}p_i^3q_i + \frac{\rho_2^{(2)} - \rho_3^{(2)}}{1-\rho_2^{(2)}}\sum_{i=1}^{I}p_iq_i^3,\end{aligned} \tag{A22}$$

where the superscript of the mean kinship coefficients ρ_2 and ρ_3 corresponds to the population for which these quantities are calculated. The formulas for $\mathrm{Cov}[\widetilde{H}_{12} - (1/2)\widetilde{H}_1 - (1/2)\widetilde{H}_2, \widetilde{H}_{12}]$ and $\mathrm{Cov}[\widetilde{H}_{12,\mathrm{BLUE}} - (1/2)\widetilde{H}_{1,\mathrm{BLUE}} - (1/2)\widetilde{H}_{2,\mathrm{BLUE}}, \widetilde{H}_{12,\mathrm{BLUE}}]$ are obtained by substituting $\bar{\Phi}_2$ and $\bar{\Phi}_3$ (for $\widetilde{H}$), or κ_2 and κ_3 (for $\widetilde{H}_{\mathrm{BLUE}}$) into Equation A22.

Proof: We begin by breaking up the covariance into its components,

$$\mathrm{Cov}\left[\breve{H}_{12} - \frac{1}{2}\breve{H}_1 - \frac{1}{2}\breve{H}_2, \breve{H}_{12}\right] = \mathrm{Var}\left[\breve{H}_{12}\right] - \frac{1}{2}\mathrm{Cov}\left[\breve{H}_1, \breve{H}_{12}\right] - \frac{1}{2}\mathrm{Cov}\left[\breve{H}_2, \breve{H}_{12}\right].$$

This equation is composed of terms that we previously derived (Equations A13, A18, and A19). Therefore,

$$\mathrm{Cov}\left[\breve{H}_{12}-\frac{1}{2}\breve{H}_1-\frac{1}{2}\breve{H}_2,\breve{H}_{12}\right]=\rho_2^{(1)}\left(1-\rho_2^{(2)}\right)\sum_{i=1}^{I}p_iq_i^2+\rho_2^{(2)}\left(1-\rho_2^{(1)}\right)\sum_{i=1}^{I}p_i^2q_i+\rho_2^{(1)}\rho_2^{(2)}\sum_{i=1}^{I}p_iq_i$$

$$+\left(\rho_2^{(1)}\rho_2^{(2)}-\rho_2^{(1)}-\rho_2^{(2)}\right)\left(\sum_{i=1}^{I}p_iq_i\right)^2$$

$$-\frac{1}{2\left(1-\rho_2^{(1)}\right)}\left[2\left(\rho_3^{(1)}-\rho_2^{(1)}\right)\sum_{i=1}^{I}p_i^3q_i+\left(2\rho_2^{(1)}-3\rho_3^{(1)}\right)\sum_{i=1}^{I}p_i^2q_i+\rho_3^{(1)}\sum_{i=1}^{I}p_iq_i\right]$$

$$-\frac{1}{2\left(1-\rho_2^{(2)}\right)}\left[2\left(\rho_3^{(2)}-\rho_2^{(2)}\right)\sum_{i=1}^{I}p_iq_i^3+\left(2\rho_2^{(2)}-3\rho_3^{(2)}\right)\sum_{i=1}^{I}p_iq_i^2+\rho_3^{(2)}\sum_{i=1}^{I}p_iq_i\right]$$

$$=\left[\rho_2^{(1)}\left(1-\rho_2^{(2)}\right)+\frac{3\rho_3^{(2)}-2\rho_2^{(2)}}{2\left(1-\rho_2^{(2)}\right)}\right]\sum_{i=1}^{I}p_iq_i^2+\left[\rho_2^{(2)}\left(1-\rho_2^{(1)}\right)+\frac{3\rho_3^{(1)}-2\rho_2^{(1)}}{2\left(1-\rho_2^{(1)}\right)}\right]\sum_{i=1}^{I}p_i^2q_i$$

$$+\left[\rho_2^{(1)}\rho_2^{(2)}-\frac{\rho_3^{(1)}}{2\left(1-\rho_2^{(1)}\right)}-\frac{\rho_3^{(2)}}{2\left(1-\rho_2^{(2)}\right)}\right]\sum_{i=1}^{I}p_iq_i+\left(\rho_2^{(1)}\rho_2^{(2)}-\rho_2^{(1)}-\rho_2^{(2)}\right)\left(\sum_{i=1}^{I}p_iq_i\right)^2$$

$$+\frac{\rho_2^{(1)}-\rho_3^{(1)}}{1-\rho_2^{(1)}}\sum_{i=1}^{I}p_i^3q_i+\frac{\rho_2^{(2)}-\rho_3^{(2)}}{1-\rho_2^{(2)}}\sum_{i=1}^{I}p_iq_i^3.\qquad\square$$