

Properties and unbiased estimation of F - and D -statistics in samples containing related and inbred individuals

Mehreen R. Mughal^{1,*}, Michael DeGiorgio^{2,*}

¹*Bioinformatics and Genomics at the Huck Institutes of the Life Sciences, Pennsylvania State University, University Park, PA 16802, USA*

²*Department of Computer and Electrical Engineering and Computer Science, Florida Atlantic University, Boca Raton, FL 33431, USA*

*Corresponding authors: mrm79@psu.edu (M.R.M), mdegiorg@fau.edu (M.D.)

Abstract

The Patterson F - and D -statistics are commonly-used measures for quantifying population relationships and for testing hypotheses about demographic history. These statistics make use of allele frequency information across populations to infer different aspects of population history, such as population structure and introgression events. Inclusion of related or inbred individuals can bias such statistics, which may often lead to the filtering of such individuals. Here we derive statistical properties of the F - and D -statistics, including their biases due to finite sample size or the inclusion of related or inbred individuals, their variances, and their corresponding mean squared errors. Moreover, for those statistics that are biased, we develop unbiased estimators and evaluate the variances of these new quantities. Comparisons of the new unbiased statistics to the originals demonstrates that our newly-derived statistics often have lower error across a wide population parameter space. Furthermore, we apply these unbiased estimators using several global human populations with the inclusion of related individuals to highlight their application on an empirical dataset. Finally, we implement these unbiased estimators in open-source software package **funbiased** for easy application by the scientific community.

Introduction

The recently introduced F - and D -statistics (Huson et al., 2005; Kulathinal et al., 2009; Reich et al., 2009; Green et al., 2010; Patterson et al., 2012) have transformed the way geneticists measure population differentiation. These statistics have been instrumental in many major recent discoveries, including testing which Neanderthal populations are closest to the populations that admixed with modern humans (Hajdinjak et al., 2018), and detecting which population is likely the admixing source for European admixture in modern Ethiopian populations (Molinaro et al., 2019). Iterating through different combinations of populations using the F_4 - and D -statistics has allowed reconstruction of population histories in diverse groups such as Native Americans and South Asians (Reich et al., 2012; Moorjani et al., 2013). In addition, The D -statistics have been used extensively to provide evidence of introgression and hybridization among species of *Drosophila* fruit flies and *Heliconius* butterflies (Martin et al., 2014; Turissini and Matute, 2017).

In many cases, however, the populations tested by these statistics are small, and proper random sampling may include data from related individuals. It is common to remove one or more of the relatives from a group of related individuals because including them might provide a bias in the value of a particular statistic being measured (Rosenberg, 2006; DeGiorgio and Rosenberg, 2009; DeGiorgio et al., 2010; Waples and Anderson, 2017; Harris and DeGiorgio, 2017a). For this reason we explore whether the current estimators for these statistics are biased with the inclusion of related or inbred individuals and if so, then develop unbiased estimators under such scenarios.

These statistics are flexible and relatively simple to compute, as they measure genetic drift along branches of a population tree by contrasting allele frequencies between different combinations of populations. Using allele frequency data from two, three, or four populations, these statistics measure shared variation along specific branches of the tree relating the populations. We begin by providing intuitive descriptions and formal definitions of each of the F - and D -statistics that we evaluate. Specifically, consider that we have allele frequency data at J biallelic loci from each of four populations, denoted A , B , C , and D . We denote the frequencies of the reference allele at locus j as a_j , b_j , c_j , and d_j in populations A , B , C , and D , respectively. We define each of the F - and D -statistics as they are in Reich et al. (2009) and Patterson et al. (2012).

We first examine the F_2 statistic, which measures the amount of genetic drift separating a pair of populations, and is thus a test for differentiation between them, and is akin to the widely-used fixation index F_{ST} (Weir and Cockerham, 1984). For a pair of populations A and B , we define the F_2 statistic as

$$F_2(A, B) = \frac{1}{J} \sum_{j=1}^J F_2(A_j, B_j),$$

where for locus j

$$F_2(A_j, B_j) = (a_j - b_j)^2.$$

It is clear from this definition that F_2 takes values between zero, when the populations have identical allele frequencies, and one, when the populations are fixed for different alleles (Figure 1).

The F_3 statistic employs allele frequencies from three populations, and measures the amount of genetic drift along the branch leading to a target population, given allele frequency data from two reference populations. For a target population A and two reference populations B and C , we define the F_3 statistic as

$$F_3(A; B, C) = \frac{1}{J} \sum_{j=1}^J F_3(A_j; B_j, C_j),$$

where for locus j

$$F_3(A_j; B_j, C_j) = (a_j - b_j)(a_j - c_j).$$

Because it measures genetic drift along a branch leading to a target population, its value is expected to be non-negative. However, an interesting property of the F_3 statistic is that it can be negative if the target population experienced admixture, and therefore a negative value directly indicates admixture in the history of the target population (Reich et al., 2009; Patterson et al., 2012). However, though F_3 can detect admixture if its value is negative, admixture is not guaranteed to lead to negative values (Reich et al., 2009; Patterson et al., 2012), and it is therefore an inconclusive test for admixture if F_3 is non-negative. Moreover, because loci with higher minor allele frequencies may affect F_3 more than loci with lower minor allele frequencies, the F_3 statistic is sometimes normalized (Reich et al., 2009; Patterson et al., 2012) based on levels of diversity of the target population. Formally, this normalized F_3 statistic has definition

$$F_3(A; B, C | A) = \frac{F_3(A; B, C)}{2G(A)},$$

where we define for population P (here $P = A$)

$$G(P) = \frac{1}{J} \sum_{j=1}^J G(P_j)$$

such that for locus j

$$G(P_j) = p_j(1 - p_j).$$

The F_4 statistic, on the other hand, is a test of “treeness” among a set of four populations, examining whether the unrooted tree relating four populations is supported by the allele frequencies within the set of populations. For a pair of sister populations A and B and a pair of sister populations C and D , we define the F_4 statistic as

$$F_4(A, B; C, D) = \frac{1}{J} \sum_{j=1}^J F_4(A_j, B_j; C_j, D_j),$$

where for locus j

$$F_4(A_j, B_j; C_j, D_j) = (a_j - b_j)(c_j - d_j).$$

If the unrooted relationship is true, then F_4 takes the value of zero, and is non-zero otherwise. If it is known *a priori* that the unrooted relationship should be true, then a non-zero F_4 statistic can be indicative of admixture, and the sign of the statistic will suggest which set of populations may be violating the assumed unrooted tree topology (Figure 1). As with the F_3 statistic, a normalized version (Reich et al., 2009; Patterson et al., 2012) of the F_4 statistic is sometimes used, with normalization based on the diversity of one of the four populations. Formally, this normalized F_4 statistic has definition

$$F_4(A, B; C, D | P) = \frac{F_4(A, B; C, D)}{G(P)}$$

where we normalize by diversity in population $P \in \{A, B, C, D\}$.

Finally, the D -statistic is a special version of the F_4 statistic that is a test of treeness for a particular asymmetric rooted tree relating four populations, with the tree topology containing a pair of sister populations, together with a close and a distant outgroup population (Figure 1). For sister populations A and B , close outgroup population C , and distant outgroup population D , we define the D statistic as

$$D(A, B, C, D) = -\frac{F_4(A, B; C, D)}{H(A, B, C, D)},$$

where

$$H(A, B, C, D) = \frac{1}{J} \sum_{j=1}^J H(A_j, B_j, C_j, D_j)$$

is a normalizing factor to constrain the D statistic to take values between negative one and one, such that for locus j

$$H(A_j, B_j, C_j, D_j) = (a_j + b_j - 2a_j b_j)(c_j + d_j - 2c_j d_j).$$

If the rooted relationship is true, then D takes the value of zero, and is non-zero otherwise. A non-zero D value can be used to detect admixture between the close outgroup population and one of the two sister populations based on its sign (Figure 1).

Theory

The F - and D -statistic equations presented in the *Introduction* employ population allele frequencies, and are thus parameters of the set of populations. To estimate them, we first need to build an estimator of allele frequencies based on samples. We denote estimates of the reference allele frequencies at locus j , $j = 1, 2, \dots, J$, in populations A , B , C , and D by \hat{a}_j , \hat{b}_j , \hat{c}_j , and \hat{d}_j , respectively.

As used previously (*e.g.*, McPeck et al., 2004; DeGiorgio and Rosenberg, 2009; DeGiorgio et al., 2010; Harris and DeGiorgio, 2017a), a linear unbiased estimator of population reference allele frequency p at a biallelic locus can be defined as

$$\hat{p} = \sum_{k=1}^{N(P)} \phi_k(P) X_k,$$

where $N(P)$ is the number of individuals sampled at the locus, X_k is the frequency of the reference allele in individual k at the locus, and $\phi_k(P)$ is the weight of individual k in population P at the locus. McPeck et al. (2004) discussed the impact of various weighting schemes on allele frequency estimation, and Harris and DeGiorgio (2017a) examined the effects of weighting scheme on estimation of expected heterozygosity.

Typical estimators of the F - and D -statistics are computed as

$$\begin{aligned}\hat{F}_2(A, B) &= \frac{1}{J} \sum_{j=1}^J \hat{F}_2(A_j, B_j) \\ \hat{F}_3(A; B, C) &= \frac{1}{J} \sum_{j=1}^J \hat{F}_3(A_j; B_j, C_j) \\ \hat{F}_4(A, B; C, D) &= \frac{1}{J} \sum_{j=1}^J \hat{F}_4(A_j, B_j; C_j, D_j) \\ \hat{F}_3(A; B, C | A) &= \frac{\hat{F}_3(A; B, C)}{2\hat{G}(A)} \\ \hat{F}_4(A, B; C, D | P) &= \frac{\hat{F}_4(A, B; C, D)}{\hat{G}(P)} \\ \hat{D}(A, B, C, D) &= -\frac{\hat{F}_4(A, B, C, D)}{\hat{H}(A, B, C, D)},\end{aligned}$$

where

$$\begin{aligned}\hat{F}_2(A_j, B_j) &= (\hat{a}_j - \hat{b}_j)^2 \\ \hat{F}_3(A_j, B_j, C_j) &= (\hat{a}_j - \hat{b}_j)(\hat{a}_j - \hat{c}_j) \\ \hat{F}_4(A_j, B_j; C_j, D_j) &= (\hat{a}_j - \hat{b}_j)(\hat{c}_j - \hat{d}_j),\end{aligned}$$

and where

$$\begin{aligned}\hat{G}(P) &= \frac{1}{J} \sum_{j=1}^J \hat{G}(P_j) \\ \hat{H}(A, B, C, D) &= \frac{1}{J} \sum_{j=1}^J \hat{H}(A_j, B_j, C_j, D_j)\end{aligned}$$

with

$$\begin{aligned}\hat{G}(P_j) &= \hat{p}_j(1 - \hat{p}_j) \\ \hat{H}(A_j, B_j, C_j, D_j) &= (\hat{a}_j + \hat{b}_j - 2\hat{a}_j\hat{b}_j)(\hat{c}_j + \hat{d}_j - 2\hat{c}_j\hat{d}_j).\end{aligned}$$

In the following, we discuss properties of these estimators, and where appropriate, develop unbiased estimators for the statistics that are biased either due to finite sample size or due to the inclusion of related or inbred individuals in the sample.

To begin, we define the kinship coefficient Φ_{xy} between individuals x and y , as the probability that a pair of sampled alleles, one from x and one from y are identical by descent if $x \neq y$, and as the probability that a pair of alleles sampled with replacement from individual x are identical by descent if $x = y$ (Lange, 2002). A pair of unrelated individuals x and y have kinship coefficient $\Phi_{xy} = 0$ (Lange, 2002). Moreover, an individual x with ploidy m_x has kinship coefficient $\Phi_{xx} = 1/m_x + (1 - 1/m_x)f_x = (1/m_x)[1 + (m_x - 1)f_x]$, where f_x is the inbreeding coefficient of individual x , and is defined as the probability that a pair of alleles sampled without replacement in individual x are identical by descent (DeGiorgio et al., 2010). A non-inbred individual x has inbreeding coefficient $f_x = 0$, and so if x is non-inbred, then their kinship coefficient is $\Phi_{xx} = 1/m_x$. As in DeGiorgio et al. (2010) and Harris and DeGiorgio (2017a), we define the weighted mean

kinship coefficients across sets of individuals sampled in population $P \in \{A, B, C, D\}$ at locus j as

$$\begin{aligned}\Phi_2(P_j) &= \sum_{w=1}^{N(P_j)} \sum_{x=1}^{N(P_j)} \phi_w(P_j) \phi_x(P_j) \Phi_{wx} \\ \Phi_3(P_j) &= \sum_{w=1}^{N(P_j)} \sum_{x=1}^{N(P_j)} \sum_{y=1}^{N(P_j)} \phi_w(P_j) \phi_x(P_j) \phi_y(P_j) \Phi_{wxy} \\ \Phi_4(P_j) &= \sum_{w=1}^{N(P_j)} \sum_{x=1}^{N(P_j)} \sum_{y=1}^{N(P_j)} \sum_{z=1}^{N(P_j)} \phi_w(P_j) \phi_x(P_j) \phi_y(P_j) \phi_z(P_j) \Phi_{wxyz} \\ \Phi_{2,2}(P_j) &= \sum_{w=1}^{N(P_j)} \sum_{x=1}^{N(P_j)} \sum_{y=1}^{N(P_j)} \sum_{z=1}^{N(P_j)} \phi_w(P_j) \phi_x(P_j) \phi_y(P_j) \phi_z(P_j) \Phi_{wx,yz},\end{aligned}$$

which are the weighted mean kinship coefficients for the $N(P_j)$ individuals sampled at locus j in population P for pairs, triples, quadruples, and pairs of pairs of individuals, respectively. From our definitions of kinship, we know that unrelated individuals have kinship coefficients of zero, but non-inbred individuals still have positive values of their self kinship coefficient, thereby causing the mean kinship coefficients to necessarily be positive quantities. It is for this reason that some F -statistic estimators will be biased even without related or inbred individuals, and this bias would be due to finite sample size. For accurate estimates of the drift quantities, it is therefore important to obtain unbiased estimators.

A number of quantities (particularly variances and covariances involving the F - and D -statistics) will be mathematically complex, as they will involve linear combinations of higher order mean kinship coefficients. For this reason, we follow prior studies (DeGiorgio et al., 2010; Harris and DeGiorgio, 2017a) and make the simplifying assumption that no individual in a sample from population P is related to more than one other individual in the sample, such that terms $\Phi_3(P_j)$, $\Phi_4(P_j)$, $\Phi_{2,2}(P_j)$, and $\Phi_2(P_j)^2$ negligible to $\Phi_2(P_j)$. Moreover, we assume that individuals sampled in different populations are unrelated to each other. Under these assumptions, we approximate a few key results from prior studies (DeGiorgio et al., 2010; Harris and DeGiorgio, 2017a) that will ultimately make derivations easier. Given that \hat{p}_j is an estimate of the frequency of a reference allele at locus j in population P , we have the following expectations (approximate notation when not exact)

$$\begin{aligned}\mathbb{E}[\hat{p}_j] &= p_j \\ \mathbb{E}[\hat{p}_j^2] &= p_j^2 + \Phi_2(P_j)p_j(1 - p_j) \\ \mathbb{E}[\hat{p}_j^3] &\approx p_j^3 + 3\Phi_2(P_j)p_j^2(1 - p_j) \\ \mathbb{E}[\hat{p}_j^4] &\approx p_j^4 + 6\Phi_2(P_j)p_j^3(1 - p_j).\end{aligned}$$

From prior studies (Nei and Roychoudhury, 1974; Weir, 1989; DeGiorgio and Rosenberg, 2009; DeGiorgio et al., 2010; Harris and DeGiorgio, 2017a), we know that $2\hat{p}(1 - \hat{p})$ is a downwardly biased estimator of expected heterozygosity at a locus, with the bias due to finite sample size (Nei and Roychoudhury, 1974) and exacerbated by the inclusion of inbred (Weir, 1989) and related (DeGiorgio and Rosenberg, 2009; DeGiorgio et al., 2010; Harris and DeGiorgio, 2017a) individuals in the sample. Based on this definition, $2G(P) = 2p(1 - p)$ is expected heterozygosity, and its estimator $2\hat{G}(P) = 2\hat{p}(1 - \hat{p})$ therefore biased. We begin by developing an unbiased estimator for $G(P)$, as it is a key normalization quantity in the F_3 and F_4 statistics.

Lemma 1. Consider J polymorphic loci in a population P with parametric reference allele frequencies $p_j \in (0, 1)$, and suppose we take a random sample of $N(P_j)$ individuals at locus j , some of which may be related or inbred. The estimator $\hat{G}(P)$ has downward bias

$$\text{Bias}[\hat{G}(P)] = -\frac{1}{J} \sum_{j=1}^J \Phi_2(P_j)G(P_j)$$

and an unbiased estimator of $G(P)$ is

$$\tilde{G}(P) = \frac{1}{J} \sum_{j=1}^J \tilde{G}(P_j),$$

where

$$\tilde{G}(P_j) = \frac{\hat{G}(P_j)}{1 - \Phi_2(P_j)}$$

is an unbiased estimator of $G(P_j)$.

The proof of Lemma 1 is in the *Appendix*. Intuitively though, because $\hat{G}(P)$ involves the product of frequencies for two alleles drawn from population P , there is a chance of having the two alleles being identical by descent by sampling the same allele twice, and is therefore a biased estimator with and without related or inbred individuals

We next consider examining the bias of the estimator $\hat{F}_2(A, B)$. As with $\hat{G}(P)$, because $\hat{F}_2(A, B)$ requires sampling two alleles from population A and two alleles from population B , we find it is biased due to not only finite sample size but also the inclusion of related or inbred individuals. We present the formal result for F_2 next (Proposition 2), and prove the result in the *Appendix*.

Proposition 2. Consider J polymorphic loci in populations A and B with respective parametric reference allele frequencies $a_j, b_j \in (0, 1)$, and suppose we take a random sample of $N(P_j)$ individuals at locus j in population $P \in \{A, B\}$, some of which may be related or inbred. The estimate $\hat{F}_2(A, B)$ has upward bias

$$\text{Bias}[\hat{F}_2(A, B)] = \frac{1}{J} \sum_{j=1}^J [\Phi_2(A_j)G(A_j) + \Phi_2(B_j)G(B_j)]$$

and an unbiased estimator of $F_2(A, B)$ is

$$\tilde{F}_2(A, B) = \frac{1}{J} \sum_{j=1}^J [\hat{F}_2(A_j, B_j) - \Phi_2(A_j)\tilde{G}(A_j) - \Phi_2(B_j)\tilde{G}(B_j)].$$

As one can see, the original estimator $\hat{F}_2(A, B)$ is upwardly biased due to finite sample size and relatedness, and that sampling within both populations A and B contributes proportionally to this bias. The new unbiased estimator $\tilde{F}_2(A, B)$ corrects this bias by adjusting the computation to account for the kinship coefficients and diversity within each population, with the adjustment of diversity using the unbiased estimator $\tilde{G}(P)$ presented in Lemma 1.

Similarly to $\hat{F}_2(A, B)$, the original estimator $\hat{F}_3(A; B, C)$ is also upwardly biased because it requires the sampling of two alleles from the target population A . We show the formal results for F_3 next (Proposition 3), and prove the result in the *Appendix*.

Proposition 3. Consider J polymorphic loci in populations A , B , and C with respective parametric reference allele frequencies $a_j, b_j, c_j \in (0, 1)$, and suppose we take a random sample of $N(P_j)$ individuals at locus j in population $P \in \{A, B, C\}$, some of which may be related or inbred. The estimate $\hat{F}_3(A; B, C)$ has upward bias

$$\text{Bias}[\hat{F}_3(A; B, C)] = \frac{1}{J} \sum_{j=1}^J \Phi_2(A_j)G(A_j)$$

and an unbiased estimator of $F_3(A; B, C)$ is

$$\tilde{F}_3(A; B, C) = \frac{1}{J} \sum_{j=1}^J [\hat{F}_3(A_j; B_j, C_j) - \Phi_2(A_j)\tilde{G}(A_j)].$$

The bias of the original estimator is proportional to the relatedness and diversity within the target population A . The new unbiased estimator $\hat{F}_3(A; B, C)$ corrects the bias by adjusting the computation to account for the kinship and diversity within the target population, with the adjustment of diversity using the unbiased estimator $\hat{G}(A)$. Moreover, it is important to note that the reference populations B and C do not contribute to bias, as only a single allele is sampled from each of these populations.

Given that $\hat{F}_3(A; B, C | A)$ uses the biased estimators $\hat{F}_3(A; B, C)$ and $\hat{G}(A)$ in its definition, we can expect that it would be biased as its component estimators are biased, and these components have different biases that are also in different directions. However, $\hat{F}_3(A; B, C | A)$ is a ratio estimator, and we can therefore not directly take its expectation to evaluate bias. Instead, we will make some simplifying assumptions and compute the approximate bias of $\hat{F}_3(A; B, C | A)$. We show the formal results next (Proposition 4), and prove the result in the *Appendix*.

Proposition 4. Consider J polymorphic loci in populations A , B , and C with respective parametric reference allele frequencies $a_j, b_j, c_j \in (0, 1)$, and suppose we take a random sample of $N(P_j)$ individuals at locus j in population $P \in \{A, B, C\}$, some of which may be related or inbred. The ratio estimator $\hat{F}_3(A; B, C | A)$ is approximately upwardly biased, assuming that its mean is well-approximated by the ratio of means of $\hat{F}_3(A; B, C)$ and $2\hat{G}(A)$ that it uses in its definition, with its upward approximate bias

$$\text{Bias}[\hat{F}_3(A; B, C | A)] \approx \frac{(1/J) \sum_{j=1}^J \Phi_2(A_j) G(A_j)}{G(A) - (1/J) \sum_{j=1}^J \Phi_2(A_j) G(A_j)} \left[F_3(A; B, C | A) + \frac{1}{2} \right].$$

Moreover, an approximately unbiased estimator of $F_3(A; B, C | A)$ is

$$\tilde{F}_3(A; B, C | A) = \frac{\tilde{F}_3(A; B, C)}{2\tilde{G}(A)}.$$

There is an upward approximate bias of the original normalized F_3 estimator, and the bias is, as with the standard estimator of F_3 , due partially to the diversity and sampling in the target population. The new approximately unbiased estimator $\tilde{F}_3(A; B, C | A)$ is based simply on the ratio of unbiased estimators of its components $\tilde{F}_3(A; B, C)$ and $\tilde{G}(A)$.

Finally, we move to the four population statistics F_4 and D . Note that the F_4 statistic by definition only samples a single allele per population, and therefore the original estimator $\hat{F}_4(A, B; C, D)$ is intuitively unbiased. We show the formal results next (Proposition 5), and prove the result in the *Appendix*.

Proposition 5. Consider J polymorphic loci in populations A , B , C , and D with respective parametric reference allele frequencies $a_j, b_j, c_j, d_j \in (0, 1)$, and suppose we take a random sample of $N(P_j)$ individuals at locus j in population $P \in \{A, B, C, D\}$, some of which may be related or inbred. The estimator $\hat{F}_4(A, B; C, D)$ is unbiased.

Though the original F_4 estimator is unbiased, the normalized F_4 and D statistics are more complicated as they are ratio estimators, meaning their biases cannot be directly assessed. However, intuitively, because both estimators have $\hat{F}_4(A, B; C, D)$ as their numerator, bias would seemingly derive from their denominator component. Next, we show formally in Proposition 6 that the normalized $\hat{F}_4(A, B; C, D | P)$ estimator is approximately upwardly biased, and prove the result in the *Appendix*.

Proposition 6. Consider J polymorphic loci in populations A , B , C and D with respective parametric reference allele frequencies $a_j, b_j, c_j, d_j \in (0, 1)$, and suppose we take a random sample of $N(P_j)$ individuals at locus j in population $P \in \{A, B, C, D\}$, some of which may be related or inbred. The ratio estimator $\hat{F}_4(A, B; C, D | P)$ is approximately upwardly biased, assuming that its mean is well-approximated by the ratio of means of $\hat{F}_4(A, B; C, D)$ and $\hat{G}(P)$ for any population $P \in \{A, B, C, D\}$ that it uses in its definition, with its upward approximate bias

$$\text{Bias}[\hat{F}_4(A, B; C, D | P)] \approx \frac{(1/J) \sum_{j=1}^J \Phi_2(P_j) G(P_j)}{G(P) - (1/J) \sum_{j=1}^J \Phi_2(P_j) G(P_j)} F_4(A, B; C, D | P).$$

Moreover, an approximately unbiased estimator of $F_4(A, B; C, D | P)$ is

$$\tilde{F}_4(A, B; C, D | P) = \frac{\hat{F}_4(A, B; C, D)}{\tilde{G}(P)}.$$

The reasoning that the original $\hat{F}_4(A, B; C, D | P)$ estimator has upward approximate bias is that its estimator $\hat{G}(P)$ used in its denominator is downwardly biased. By using the unbiased estimator $\tilde{G}(P)$ in its place within the denominator, we find a new estimator $\tilde{F}_4(A, B; C, D | P)$ is approximately unbiased.

The bias property of the D statistic is different than the normalized F_4 statistic, as the estimator $\hat{H}(A, B, C, D)$ of its denominator is unbiased (Lemma 8 of the *Appendix*). Intuitively, this result is due to the denominator not having a product of frequencies for two alleles sampled from the same population. Because both its numerator and denominator are unbiased, we next show that the ratio estimator $\hat{D}(A, B, C, D)$ is approximately unbiased in Proposition 7, and prove the result in the *Appendix*.

Proposition 7. Consider J polymorphic loci in populations A, B, C , and D with respective parametric reference allele frequencies $a_j, b_j, c_j, d_j \in (0, 1)$, and suppose we take a random sample of $N(P_j)$ individuals at locus j in population $P \in \{A, B, C, D\}$, some of which may be related or inbred. The ratio estimator $\hat{D}(A, B, C, D)$ is approximately unbiased, assuming that its mean is well-approximated by the ratio of means of $\hat{F}_4(A, B; C, D)$ and $\hat{H}(A, B, C, D)$ that it uses in its definition.

In addition to bias, variance is an important property of an estimator, as both bias and variance are components of mean squared error. Because the formulas and derivations for the variances of the F - and D -statistics are not particularly insightful, we relegate these results to the *Appendix*. Specifically, we provide the variances for $\hat{F}_2(A, B)$, $\tilde{F}_2(A, B)$, $\hat{F}_3(A; B, C)$, $\tilde{F}_3(A; B, C)$, $\hat{F}_3(A; B, C | A)$, $\tilde{F}_3(A; B, C | A)$, $\hat{F}_4(A, B; C, D)$, $\tilde{F}_4(A, B; C, D | P)$, $\hat{F}_4(A, B; C, D | P)$, and $\hat{D}(A, B, C, D)$ in Propositions 11, 12, 13, 15, 17, 19, 16, 21, 23, and 26 of the *Appendix*, respectively.

Results

In the *Theory* and *Appendix*, we introduced new unbiased estimators of F_2 and F_3 statistics, and derived biases and variances (and hence mean squared errors) for the original and new estimators of F - and D -statistics. In this section, we theoretically evaluate the relative performances of the old biased estimators and the new unbiased estimators under an array of settings, including different mixtures of relatedness, inbreeding, sample sizes, and population parameters.

For all of our results we require the kinship coefficients for each pair of individuals. To acquire these values, we need to know if each individual is related to any other in the population and also whether they are inbred, and if so, how these values are quantified through the use of kinship coefficients (Φ_{xy}). To summarize how an entire sample from a population P is related to each other at a locus, we use

$$\Phi_2(P) = \sum_{w=1}^{N(P)} \sum_{x=1}^{N(P)} \phi_w(P) \phi_x(P) \Phi_{wx},$$

where $\phi_w(P)$ and $\phi_x(P)$ are weights of individuals w and x in population P , and in this study we use weights corresponding to the proportion of alleles contributed by individual x to the sample from population P , which is computed as

$$\phi_x(P) = \frac{m_x}{\sum_{k=1}^{N(P)} m_k}.$$

Here m_x is the ploidy of individual x . Moreover, using this weighting scheme, we also estimate the frequency of the reference allele at a biallelic locus as the sample proportion (McPeck et al., 2004; DeGiorgio et al., 2010; Harris and DeGiorgio, 2017a)

$$\hat{p} = \sum_{k=1}^{N(P)} \phi_k(P) X_k = \sum_{k=1}^{N(P)} \frac{m_k}{\sum_{j=1}^{N(P)} m_j} X_k.$$

Effect of population F -statistic value on mean squared error

The relationship between the population parameter for a statistic and the estimate based on a sample from the population is important to evaluate. We compare the difference in the mean squared error (MSE) between the biased \hat{F} estimators and our unbiased \tilde{F} estimators to the true value of each statistic in the cases for which both estimators exist. The F_2 , F_3 , and F_4 statistics require allele frequency information from either two, three, or four populations, respectively.

For our F_2 comparisons, we use the sample allele frequencies from the YRI (sub-Saharan African) and CEU (central Europeans) from the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2015) as the true population allele frequencies to obtain the true $F_2(A, B)$ statistic by using the population definition from the *Introduction*, with populations $A = \text{CEU}$ and $B = \text{YRI}$. To evaluate the relative performances of F_2 estimators over a range of true F_2 values, we randomly sample 20 independent loci from both populations for 1000 independent replicates of $J = 20$ loci, yielding 1000 independent draws of the true F_2 statistic, which ranged across the set of values $F_2 \in [0.02, 0.12]$. Using these allele frequencies, along with the sample size and relatedness information, we also calculate the difference in MSE between the \hat{F}_2 and \tilde{F}_2 estimators by using Propositions 2, 11 and 12. We then calculate the MSE by summing the variance and squared bias. We note that the MSEs of the unbiased estimators are equal to their variances. We repeat this process for $F_3(A; B, C)$, $F_3(A; B, C | A)$, and $F_4(A, B; C, D | A)$ as these are the estimators that are biased in their \hat{F} forms.

We use Propositions 3, 13, and 15 to determine the MSE for both biased and unbiased F_3 estimators by including allele frequency information from the JPT (Japanese) population where $A = \text{JPT}$, $B = \text{CEU}$, and $C = \text{YRI}$, with true range for $F_3 \in [0.00, 0.08]$. For the normalized $F_3(A; B, C | A)$ estimators, we compare MSE between the biased and unbiased versions by using bias and variances derived in Propositions 4, 17, and 19, with true range for normalized $F_3 \in [0.00, 0.30]$. Finally, we estimate the MSE for the normalized $F_4(A, B; C, D | A)$ estimators by including GIH (Gujarati Indian) allele frequency data and using the derivations in Propositions 6, 21, and 23. In this case we set $A = \text{YRI}$, $B = \text{CEU}$, $C = \text{JPT}$, and $D = \text{GIH}$ for a true range of normalized $F_4 \in [-0.3, 0.2]$.

For each analysis we estimate MSE for instances when samples of 60 diploid individuals from each population include 30 relative pairs, including ten avuncular relationships, ten inbred full siblings, and ten outbred full siblings. We also assumed every individual was related to exactly one other individual. In these estimates, all populations contain the same composition of related individuals.

The difference in $\log_{10}(\text{MSE})$ between \hat{F} and \tilde{F} estimators for $F_2(A, B)$, $F_3(A; B, C)$, $F_3(A; B, C | A)$, and $F_4(A, B; C, D | A)$ show similar trends with respect to the true F -statistic values. Specifically, the difference in $\log_{10}(\text{MSE})$ decreases as the true F -statistic value approaches zero (Figure 2). In our evaluation of $F_4(A, B; C, D | A)$, we considered both positive and negative values for its true value, which shows that the difference in $\log_{10}(\text{MSE})$ of \hat{F}_4 and \tilde{F}_4 exhibits a quadratic shaped trend as a function of true F_4 . Overall, we notice that the difference in MSE between biased and unbiased estimators is dependent on the true value of the F -statistic, with the least difference occurring when the true F -statistic is closest to zero.

Effect of sample size on mean squared error

To probe how sample size within each population affects the difference in estimator error rate, we theoretically computed the MSE for both \hat{F} and \tilde{F} estimators when different numbers of individuals are sampled, with the constraint that every sampled individual is related to exactly one other individual in the sample from that population. Specifically, we evaluate the impact on these estimators when sampling from one pair to 50 pairs of related individuals, with relationships of inbred full sibling pairs ($\Phi_{xy} = 3/8$), outbred full sibling pairs ($\Phi_{xy} = 1/4$), parent-offspring pairs ($\Phi_{xy} = 1/4$), and avuncular pairs ($\Phi_{xy} = 1/8$). We compute the MSE as in *Effect of population F -statistic value on mean squared error* section above.

In almost all cases, the biased \hat{F} estimators always displayed elevated MSE compared to their corresponding unbiased \tilde{F} estimators (Figures 4 and S1-S3). For both estimators, we see a clear decrease in the MSE as the number of sampled individuals increases, with the greatest error observed when two individuals are sampled. As expected, a greater sample size allows one to better estimate allele frequencies, and ultimately reduces the mean pairwise kinship coefficient within the sample, as the number of pairs in the sample grows quadratically but the number of relative pairs grows linearly. We also find that the difference

in the MSE at larger sample sizes is not as pronounced for normalized F_4 as it is for F_2 , F_3 , and normalized F_3 , as the difference in bias between the biased and unbiased estimators is much smaller for normalized F_4 (Figure S3).

Effect of sample composition on mean squared error

Different types of relatives have different proportions of their alleles shared identical by descent, and thus have different pairwise kinship coefficients. Because we have demonstrated that bias and variance (and hence MSE) of estimators are influenced by within-population mean pairwise kinship coefficient across sampled individuals, the distribution of relative types within a sample will impact overall F -statistic estimation error. For this reason, it is important to examine how our F -statistics are affected by samples containing diverse mixtures of relative types. Specifically, to accurately assess the impact of relative composition, we hold sample sizes, number of relative pairs, and true population F -statistic values constant.

We computed the theoretical MSE when samples of 50 pairs of relatives (100 diploid individuals sampled) contain relative pairs of three different types as in Harris and DeGiorgio (2017a). In addition, each individual is related to exactly one other individual in the sample from the same population. For each statistic we vary the number of pairs related by each of three types of relationships between zero and 50, with 1326 combinations for each. We repeat this process for three configurations of relationships to probe estimator error as a function of the mixture of relative types. We also provide comparisons among inclusion of male-male full siblings ($\Phi_{xy} = 1/2$), male-female full siblings ($\Phi_{xy} = 3/8$) and female-female full siblings ($\Phi_{xy} = 1/4$) at mixed-ploidy loci such as on the X chromosome (DeGiorgio et al., 2010), with results showing elevated MSE for both estimators for higher male-male sibling proportions, when compared to male-female or female-female full siblings. To investigate the effects of inbred individuals, we also provide a comparison between inbred full siblings ($\Phi_{xy} = 3/8$) with inbreeding coefficient $f_x = f_y = 1/4$, and outbred full-siblings ($\Phi_{xy} = 1/4$) at a autosomal diploid loci. We see that MSE is higher for inbred full-siblings than for outbred full-siblings in all cases examined (Figures 3 and S4-S6).

We also note that the value of MSE for the biased \hat{F} estimators is always greater than the value for their corresponding unbiased \tilde{F} statistics, which is true in part due to the values of the true F -statistics for the loci we chose to use. Though the MSE is higher for the biased estimators, the variation in MSE values is similar for both estimators. For example, the data point with the highest proportion of avuncular relatives has the lowest MSE when compared to parent-offspring relationships and outbred full siblings. In all tested settings (Figures 3 and S4-S6), we notice similar patterns of MSE variation when comparing \hat{F} estimators with \tilde{F} estimators. This pattern is again shared when comparing MSE variation among the estimators for F_2 , F_3 , normalized F_3 , and normalized F_4 . We can conclude that in all of these cases, the value of the mean kinship coefficient is most important in determining MSE when sample size and true F -statistic value are fixed.

Simulations to evaluate theoretical MSE approximations

To verify that our theoretical approximations for MSE are reasonable, we simulate samples containing related individuals and use them to compute the biased \hat{F} - and unbiased \tilde{F} -statistics as well as calculate their biases, variances, and MSEs. For each population (CEU, YRI, GIH, and JPT), we simulate 10 non-inbred parent-offspring pairs with each individual related to exactly one other individual in the sample. Genotypes for each individual are simulated by first sampling two alleles with replacement according to their respective population allele frequencies from each of the populations (CEU, YRI, GIH, and JPT) to create a set of 20 unrelated individuals per population. Individuals x and y that form one of the 10 relative pairs have the genotype of individual y modified according to their relationship type. Specifically, for each relative type, there are probabilities Δ_0 , Δ_1 , and Δ_2 that the two individuals will share zero, one, or two alleles identical by descent, respectively. The first allele of individual y is copied from the first allele from individual x with probability Δ_1 and the entire genotype of individual x is copied over to individual y with probability Δ_2 . This process is repeated across 20 independent loci to generate a sample of 20 individuals with 10 relative pairs in each population with genotypes taken at $J = 20$ independent loci. To generate 20 independent loci from the four 1000 Genomes Project populations, we used loci either on separate chromosomes, or at least one megabase away from each other.

For each of our new unbiased estimators we compute the bias, variance, and MSE along with the same values for the original estimators (Figures S7-S10). Comparing the bias measurements in these figures, we observe a clear reduction in bias when applying the \tilde{F} estimators as opposed to the \hat{F} estimators. However, the variance is highly similar for \hat{F} and \tilde{F} in all cases. As the value of variance is much larger than the magnitude of the bias (by an order of magnitude) and hence the squared bias, the resulting MSE is consequently similar as well. Because F_4 is quantifying the relationship among four populations, more simulations may be required to converge to the pattern seen by theoretical simulations. For this reason, we increased the number of simulations used to compute the bias, variance, and MSE to 10^4 for each data point in Figure S10, whereas 10^3 simulation replicates were used for F_2 , and both versions of F_3 .

To compare the accuracy of our theoretical approximations to simulation results across a spectrum of relatedness between individuals in a sample, we simulate combinations of parent offspring, outbred full sibling, and avuncular relationships. In a manner similar to described above (first paragraph of *Simulations to evaluate theoretical MSE approximations*), we simulate a total of 10 relative pairs made up of a combination of each of the three relative types, with the number of each relative type ranging from zero to 10. We simulate each of these 66 distinct settings of relative type combinations with genotypes sampled at $J = 20$ independent loci, and completed 1000 independent replicates of each setting to obtain accurate measurements of bias, variance, and MSE for each simulation setting, with each simulation using true F -statistic values specified in Figures S7 and S8-S10. We compute the bias, variance, and MSE for simulations, and compare these values to theoretically calculated computations for each relative combination (Figures S12-S15). We find that although noisier, the bias variance, and MSE patterns in our simulation results match theoretical calculations, suggesting that our theoretical computations are accurate. For all cases the simulated bias measurements for the \tilde{F} estimators are close to zero, whereas the \hat{F} estimators display bias measurements matching the theoretically calculated \hat{F} bias values.

Utility and applications of unbiased estimators

In previous sections, we have shown through simulations that our theoretical results are producing expected patterns and evaluated the performance of our unbiased estimators under varying combinations of relatives, true F -statistic values, and sample sizes. In this section we show some potential applications of these estimators, using both simulated and empirical data. As discussed previously in the *Introduction*, the value of $F_3(A; B, C)$ can be used to identify whether population A is the result of admixture between populations related to B and C (Figure 1). A negative value of F_3 indicates the presence of this process, whereas a non-negative value is inconclusive and means that further tests may be required to verify a history of admixture. However, because \hat{F}_3 is upwardly biased and because \tilde{F}_3 corrects for this bias, \tilde{F}_3 might allow us to detect admixture in cases where \hat{F}_3 would be inconclusive, even without the presence of related or inbred individuals.

To explore this hypothesis, we first examine an admixture scenario in which $F_3(A; B, C)$ might provide marginally negative values. We simulate two populations (B and C) with effective population size of 10^4 diploid individuals (Takahata, 1993) that diverged 2000 generations prior to sampling using SLiM (Haller and Messer, 2019). This simple divergence model has parameters inspired by the history relating African and non-African human populations (Gravel et al., 2011). These populations then merge with admixture proportions 0.4 and 0.6 for B and C , respectively, to form population A 400 generations prior to sampling. Using these parameters, the expected value is $F_3(A; B, C) = -0.0568$. To generate genetic data from this model, we evolved sequences with a per-site per-generation mutation rate of $\mu = 1.25 \times 10^{-8}$ (Scally and Durbin, 2012) and a uniform per-site per-generation recombination rate of $r = 10^{-8}$ (Payseur and Nachman, 2000). We output 20 two megabase chromosomal regions containing allele frequency information for all three populations. Using these simulated population allele frequencies for each of these three populations, we then simulate 50 instances of 50 unrelated individuals each. We then compute $\tilde{F}_3(A; B, C)$ and $\hat{F}_3(A; B, C)$ across $J = 20$ loci, either on separate chromosomes or at least one megabase away from each other to ensure independence.

Figure 5 illustrates that \tilde{F}_3 values are lower than \hat{F}_3 , and are always negative when \hat{F}_3 values are almost always positive. Because this statistic is used to test for admixture and a negative result indicates the presence of admixture, the use of the biased estimator leads to a different conclusion than when using the unbiased estimator. We also explore a setting in which populations contain related individuals with

the same parameters as described above. Using allele frequency information from the three populations simulated previously (A , B , and C) we generate 50 individuals for each population, in which there are 25 parent-offspring pairs. Similarly to when relatives are not included in the population, \tilde{F}_3 values are lower than \hat{F}_3 , with most \tilde{F}_3 values giving negative values, and all \hat{F}_3 providing positive values, again lending different conclusions about the underlying demographic history of these populations (Figure 5).

Finally, we test the performance of our statistics on empirical data. We use populations from the HGDP SNP dataset (Li et al., 2008) that include related individuals (Rosenberg, 2006). Specifically, we use genotype information from Colombian, Lahu, Melanesian, Mandenka, San, and Druze populations, and we sample 20 independent loci that are at least one megabase apart from all populations for 1000 independent replicates of $J = 20$ loci, yielding 1000 independent draws. Each of these populations contains between two and 14 pairs of inferred related individuals, according to Rosenberg (2006). Using distinct pairs for F_2 , triples for F_3 , and quadruples for F_4 of these populations and the relationships from Rosenberg (2006), we estimate $\hat{F}_2(A, B)$, $\tilde{F}_2(A, B)$, $\hat{F}_3(A; B, C | A)$, $\tilde{F}_3(A; B, C | A)$, $\hat{F}_4(A, B; C, D | A)$, and $\tilde{F}_4(A, B; C, D | A)$ and compare the mean and standard deviation of the biased and unbiased estimators (Figure 6). In all cases shown, the biased estimator has higher mean than the unbiased estimator, although the standard deviations are similar for both. This indicates that correcting the bias generated by related individuals yields more accurate F -statistic estimates with minimal cost in precision of the estimates.

Discussion

We have introduced the unbiased estimators $\tilde{F}_2(A, B)$, $\tilde{F}_3(A; B, C)$, $\tilde{F}_3(A; B, C | A)$, and $\tilde{F}_4(A, B; C, D | P)$ as well as shown that the estimators $\hat{F}_4(A, B; C, D)$ and $\hat{D}(A, B, C, D)$ are unbiased with the inclusion of related and inbred individuals. In addition, we have demonstrated that the variance of $\tilde{F}_2(A, B)$ is similar to that of $\hat{F}_2(A, B)$, as are the variances of $\tilde{F}_3(A; B, C)$ and $\hat{F}_3(A; B, C)$. We have also provided variance calculations for all other F - and D -statistic estimators included in this study. Using these calculations we have compared the performance of the biased and newly derived unbiased estimators, and shown that in most cases the unbiased estimators have lower MSE values than the biased estimators of the same statistic.

Interestingly, the two statistics that sample from each analyzed population only once per locus— $\hat{F}_4(A, B; C, D)$ and $\hat{D}(A, B, C, D)$ —are unbiased with the inclusion of related or inbred individuals, whereas $\hat{F}_2(A, B)$, which samples from each population A and B twice, and $\hat{F}_3(A; B, C)$, which samples from population A twice, are biased. This process of sampling more than once from a single population per locus is responsible for creating bias due simply to finite sample size, which is exacerbated by the inclusion of related or inbred individuals within the twice-sampled population.

The development of these unbiased statistics, and the proofs showing other statistics are unbiased is beneficial for anthropologists interested in populations such as hunter-gatherers, some of which are often small and widely dispersed yet retain high genetic diversity (Kim et al., 2014). Small population sizes may necessitate the sampling of close relatives, such as parents and offspring, or siblings. Along with small human populations, these statistics are often applied to non-human species. Some, such as elephants, rhinoceros, and cheetahs are close to extinction or have extremely small and inbred populations due to human activity. The F - and D -statistics may prove important in conservation efforts to test how (and whether) different populations of these animals are interacting. For these reasons, having estimators that are unbiased under such conditions is imperative in making accurate inferences about the relationships of such small populations with others. Although it may not be possible to identify relatives through the sampling process, especially in the case of wild animals, there are methods available to identify related individuals and estimate their likely degree of relatedness once the samples have been sequenced (Epstein et al., 2000). The inferences from these methods will allow users to identify pairwise kinship coefficients necessary to apply the unbiased statistics of this study. Moreover, even if relatedness is difficult to assess, many of the original statistics (*i.e.*, \hat{F}_2 , \hat{F}_3 , normalized \hat{F}_3 , and normalized \hat{F}_4) have bias due to finite sample size, and simply accounting for the bias may be important to accurately assess population history and diversity across populations (*e.g.*, Figure 5).

A key consideration when evaluating the importance unbiased estimators of F - and D -statistics is their potential use. Specifically, a number of applications of these statistics do not employ the raw estimates, but instead standardized estimates (Soraggi et al., 2018; Zheng and Janke, 2018), where a particular F - or D -statistic has its genomewide mean subtracted, and is normalized by the standard error using a genomic block

jackknife procedure (Reich et al., 2009). Indeed, subtracting out this genomewide mean may circumvent bias issues. However, this assumes that all genomic blocks have similar sample properties, yet blocks with reduced sample size (*e.g.*, in regions with difficult to call genotypes) may still deviate from the genomewide expectation. In contrast, accounting for this bias due to finite sample size would provide estimates closer to the genomewide mean. Because the variance for these biased and unbiased estimators is approximately the same (compare Propositions 11 vs. 12 and 13 vs. 15), the standard errors used for normalizing these statistics are expected to be comparable, and thus, the unbiased estimators of the F -statistics derived here represent a more robust alternative to the original biased estimators, regardless of whether the raw or standardized values of the statistics are used. Furthermore, the raw value of some statistics, such as using the F_3 statistic to detect population admixture, is important, and without correcting the bias of such statistics (Figure 5), key historical events relating populations could be missed.

The F - and D -statistics evaluated here are the most commonly used. However, since their development by Reich et al. (2009) and Patterson et al. (2012), other D -statistic type tests have been formulated to not only detect admixture, but also to identify the direction of gene flow—namely the partitioned D -statistics of Eaton and Ree (2013) and the D_{FOIL} statistics of Pease and Hahn (2015). Specifically, the D_{FOIL} statistics as originally formulated by Pease and Hahn (2015) sampled a single lineage (or allele) from each of a set of five populations A , B , C , D , and O , with a symmetric rooted topology $((AB)(CD))$ relating populations A , B , C , and D , and with O an outgroup to these populations used to polarize the ancestral allelic state. Subsequently, Harris and DeGiorgio (2017b) derived allele frequency formulas for the D_{FOIL} statistics, and showed that allele frequency information for the outgroup population O is not needed for computation. The D_{FOIL} statistics are a set of four quantities (Harris and DeGiorgio, 2017b)

$$\begin{aligned} D_{\text{FO}}(A, B; C, D) &= \frac{\sum_{j=1}^J (1 - 2a_j)(d_j - c_j)}{\sum_{j=1}^J (c_j + d_j - 2c_j d_j)} \\ D_{\text{IL}}(A, B; C, D) &= \frac{\sum_{j=1}^J (1 - 2b_j)(d_j - c_j)}{\sum_{j=1}^J (c_j + d_j - 2c_j d_j)} \\ D_{\text{FI}}(A, B; C, D) &= \frac{\sum_{j=1}^J (1 - 2c_j)(b_j - a_j)}{\sum_{j=1}^J (a_j + b_j - 2a_j b_j)} \\ D_{\text{OL}}(A, B; C, D) &= \frac{\sum_{j=1}^J (1 - 2d_j)(b_j - a_j)}{\sum_{j=1}^J (a_j + b_j - 2a_j b_j)}, \end{aligned}$$

each of which does not have the frequencies for two alleles sampled from a single population multiplying each other. Hence, using sample allele frequencies in place of the population quantities would still yield approximately unbiased estimators of the D_{FOIL} statistics, regardless of whether related or inbred individuals were included in the sample. Though we chose to focus on the more classic F - and D -statistics, variance quantities for these partitioned D and D_{FOIL} statistics can be readily computed as we have done for other ratio estimators in this study.

Though we have only shown results when all populations contain samples with the same relative pair composition, it is trivial to include different relative types in different populations within these statistics. In addition, it is also possible to apply our new unbiased estimators when only some or none of the populations contain related or inbred individuals. Moreover, though we have demonstrated results for allele frequencies estimated as the sample proportion, we could have instead used the best linear unbiased estimator (BLUE) of McPeck et al. (2004), as all derivations in this article are based on a general form of a linear unbiased estimator. The BLUE allele frequency estimator would have superior properties to the sample proportion discussed here, as it has smallest variance (McPeck et al., 2004), and this reduction in variance translates to functions of the allele frequency as highlighted by improvements in both expected heterozygosity and F_{ST} by Harris and DeGiorgio (2017a). To apply the BLUE estimator, we would simply alter the weight $\phi_x(P)$ of an individual x in population P at a particular locus with the equation

$$\phi_x(P) = \frac{\sum_{k=1}^{N(P)} (\mathbf{K}^{-1})_{kx}}{\mathbf{1}^T \mathbf{K}^{-1} \mathbf{1}},$$

where $\mathbf{K} \in \mathbb{R}^{N(P) \times N(P)}$ is the matrix of pairwise kinship coefficients, with element in row j and column k given by $\mathbf{K}_{jk} = \Phi_{jk}$, $\mathbf{1} \in \mathbb{R}^{N(P)}$ is a column vector of ones, and superscript T indicates transpose. To facilitate easy application of these statistics, we have developed open-source software **funbiased** for use by the scientific community, which is available at <https://github.com/MehreenRuhi/funbiased>.

Acknowledgments

We thank George (PJ) Perry for the helpful comments on an early draft of this manuscript and Alexandre Harris for fruitful discussions about our simulation protocol. This research was funded by National Institutes of Health grant R35GM128590, National Science Foundation grants DEB-1949268 and BCS-2001063, a NIGMS funded training grant on Computation, Bioinformatics, and Statistics (Predoctoral Training Program T32GM102057), and the NASA Pennsylvania Space Grant Graduate Fellowship. Computations for this research were performed on the Pennsylvania State University’s Institute for Computational and Data Sciences Advanced CyberInfrastructure (ICDS-ACI).

References

- M. DeGiorgio and N. A. Rosenberg. An unbiased estimator of gene diversity in samples containing related individuals. *Molecular Biology and Evolution*, 26:501–512, 2009.
- M. DeGiorgio, I. Jankovic, and N. A. Rosenberg. Unbiased estimation of gene diversity in samples containing related individuals: Exact variance and arbitrary ploidy. *Genetics*, 186:1367–1387, 2010.
- D. A. R. Eaton and R. H. Ree. Inferring Phylogeny and Introgression using RADseq Data: An Example from Flowering Plants (Pedicularis: Orobanchaceae). *Systematic Biology*, 62:689–706, 2013.
- M. P. Epstein, W. L. Duren, and M. Boehnke. Improved inference of relationship for pairs of individuals. *American journal of human genetics*, 67:1219–1231, 2000.
- S. Gravel, B. M. Henn, R. N. Gutenkunst, A. R. Indap, G. T. Marth, A. G. Clark, F. Yu, R. A. Gibbs, , and C. D. Bustamante. Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences*, 108(29):11983–11988, 2011.
- R. E. Green, J. Krause, A. W. Briggs, T. Maricic, U. Stenzel, M. Kircher, N. Patterson, H. Li, W. Zhai, M. H.-Y. Fritz, N. F. Hansen, E. Y. Durand, A.-S. Malaspina, J. D. Jensen, T. Marques-Bonet, C. Alkan, K. Prüfer, M. Meyer, H. A. Burbano, J. M. Good, R. Schultz, A. Aximu-Petri, A. Butthof, B. Höber, B. Höffner, M. Siegemund, A. Weihmann, C. Nusbaum, E. S. Lander, C. Russ, N. Novod, J. Affourtit, M. Egholm, C. Verna, P. Rudan, D. Brajkovic, Ž. Kucan, I. Gušić, V. B. Doronichev, L. V. Golovanova, C. Lalueza-Fox, M. de la Rasilla, J. Fortea, A. Rosas, R. W. Schmitz, P. L. F. Johnson, E. E. Eichler, D. Falush, E. Birney, J. C. Mullikin, M. Slatkin, R. Nielsen, J. Kelso, M. Lachmann, D. Reich, and S. Pääbo. A draft sequence of the neandertal genome. *Science*, 328:710–722, 2010.
- M. Hajdinjak, Q. Fu, A. Hübner, M. Petr, F. Mafessoni, S. Grote, P. Skoglund, V. Narasimham, H. Rougier, I. Crevecoeur, P. Semal, M. Soressi, S. Talamo, J.-J. Hublin, I. Gušić, Ž. Kučan, P. Rudan, L. V. Golovanova, V. B. Doronichev, C. Posth, J. Krause, P. Korlević, S. Nagel, B. Nickel, M. Slatkin, N. Patterson, D. Reich, K. Prüfer, M. Meyer, S. Pääbo, and J. Kelso. Reconstructing the genetic history of late neanderthals. *Nature*, 555:652–656, 2018.
- B. C. Haller and P. W. Messer. SLiM 3: Forward Genetic Simulations Beyond the Wright–Fisher Model. *Molecular Biology and Evolution*, 36:632–637, 2019.
- A. M. Harris and M. DeGiorgio. An unbiased estimator of gene diversity with improved variance for samples containing related and inbred individuals of any ploidy. *G3: Genes, Genomes, Genetics*, 7:671–691, 2017a.
- A. M. Harris and M. DeGiorgio. Admixture and ancestry inference from ancient and modern samples through measures of population genetic drift. *Human Biology*, 89:21–46, 2017b.

- D. Huson, T. Klöpper, P. Lockhart, and M. Steel. Reconstruction of reticulate networks from gene trees. *RECOMB 2005*, 3500:233–249, 2005.
- H. L. Kim, A. Ratan, G. H. Perry, A. Montenegro, W. Miller, and S. C. Schuster. Khoisan hunter-gatherers have been the largest population throughout most of modern-human demographic history. *Nature communications*, 5:5692–5692, 2014.
- R. J. Kulathinal, L. S. Stevison, and M. A. F. Noor. The genomics of speciation in drosophila: Diversity, divergence, and introgression estimated using low-coverage genome sequencing. *PLOS Genetics*, 5:5:e1000550, 2009.
- K. Lange. *Mathematical and Statistical Methods for Genetic Analysis*. Springer, 2002.
- J. Z. Li, D. M. Absher, H. Tang, A. M. Southwick, A. M. Casto, S. Ramachandran, H. M. Cann, G. S. Barsh, M. Feldman, L. L. Cavalli-Sforza, and R. M. Myers. Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, 319:1100–1104, 2008.
- S. H. Martin, J. W. Davey, and C. D. Jiggins. Evaluating the Use of ABBA–BABA Statistics to Locate Introgressed Loci. *Molecular Biology and Evolution*, 32:244–257, 2014.
- S. McPeck, Mary, X. Wu, and C. Ober. Best linear unbiased allele-frequency estimation in complex pedigrees. *Biometrics*, 60:359–367, 2004.
- L. Molinaro, F. Montinaro, B. Yelmen, D. Marnetto, D. M. Behar, T. Kivisild, and L. Pagani. West asian sources of the eurasian component in ethiopians: a reassessment. *Scientific Reports*, 9:18811, 2019.
- P. Moorjani, K. Thangaraj, N. Patterson, M. Lipson, P.-R. Loh, P. Govindaraj, B. Berger, D. Reich, and L. Singh. Genetic evidence for recent population mixture in india. *The American Journal of Human Genetics*, 93:422–438, 2013.
- M. Nei and A. K. Roychoudhury. Sampling variances of heterozygosity and genetic distance. 76:379–390, 1974.
- N. Patterson, P. Moorjani, Y. Luo, S. Mallick, N. Rohland, Y. Zhan, T. Genschoreck, T. Webster, and D. Reich. Ancient admixture in human history. *Genetics*, 192:1065–1093, 2012.
- B. A. Payseur and M. W. Nachman. Microsatellite variation and recombination rate in the human genome. *Genetics*, 156:1285–1298, 2000.
- J. B. Pease and M. W. Hahn. Detection and Polarization of Introgression in a Five-Taxon Phylogeny. *Systematic Biology*, 64:651–662, 2015.
- D. Reich, K. Thangaraj, N. Patterson, A. L. Price, and L. Singh. Reconstructing indian population history. *Nature*, 461:489 EP –, 2009.
- D. Reich, N. Patterson, D. Campbell, A. Tandon, S. Mazieres, N. Ray, M. V. Parra, W. Rojas, C. Duque, N. Mesa, L. F. García, O. Triana, S. Blair, A. Maestre, J. C. Dib, C. M. Bravi, G. Bailliet, D. Corach, T. Hünemeier, M. C. Bortolini, F. M. Salzano, M. L. Petzl-Erler, V. Acuña-Alonzo, C. Aguilar-Salinas, S. Canizales-Quinteros, T. Tusié-Luna, L. Riba, M. Rodríguez-Cruz, M. Lopez-Alarcón, R. Coral-Vazquez, T. Canto-Cetina, I. Silva-Zolezzi, J. C. Fernandez-Lopez, A. V. Contreras, G. Jimenez-Sanchez, M. J. Gómez-Vázquez, J. Molina, Á. Carracedo, A. Salas, C. Gallo, G. Poletti, D. B. Witonsky, G. Alkorta-Aranburu, R. I. Sukernik, L. Osipova, S. A. Fedorova, R. Vasquez, M. Villena, C. Moreau, R. Barrantes, D. Pauls, L. Excoffier, G. Bedoya, F. Rothhammer, J.-M. Dugoujon, G. Larrouy, W. Klitz, D. Labuda, J. Kidd, K. Kidd, A. Di Rienzo, N. B. Freimer, A. L. Price, and A. Ruiz-Linares. Reconstructing native american population history. *Nature*, 488:370–374, 2012.
- N. A. Rosenberg. Standardized subsets of the hgdp-ceph human genome diversity cell line panel, accounting for atypical and duplicated samples and pairs of close relatives. *Annals of Human Genetics*, 70:841–847, 2006.

- A. Scally and R. Durbin. Revising the human mutation rate: implications for understanding human evolution. *Nature Reviews Genetics*, 13:745, 2012.
- S. Soraggi, C. Wiuf, and A. Albrechtsen. Powerful inference with the D -statistic on low-coverage whole-genome data. *G3: Genes, Genomes, Genetics*, 8:551–566, 2018.
- N. Takahata. Allelic genealogy and human evolution. *Molecular Biology and Evolution*, 10:2–22, 1993.
- The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526:68–74, 2015.
- D. A. Turissini and D. R. Matute. Fine scale mapping of genomic introgressions within the drosophila yakuba clade. *PLOS Genetics*, 13:1–40, 2017.
- R. S. Waples and E. C. Anderson. Purging putative siblings from population genetic data sets: a cautionary view. *Molecular Ecology*, 26:1211–1224, 2017.
- B. S. Weir. Sampling properties of gene diversity. *Plant population genetics, breeding and genetic resources*, pages 23–42, 1989.
- B. S. Weir and C. C. Cockerham. Estimating F -statistics for the analysis of population structure. *Evolution*, 38:1358–1370, 1984.
- K. M. Wolter. *Introduction to variance estimation*. Springer, New York, NY, 2nd edition, 2007.
- Y. Zheng and A. Janke. Gene flow analysis method, the D -statistic, is robust in a wide parameter space. *BMC bioinformatics*, 19:10–10, 2018.

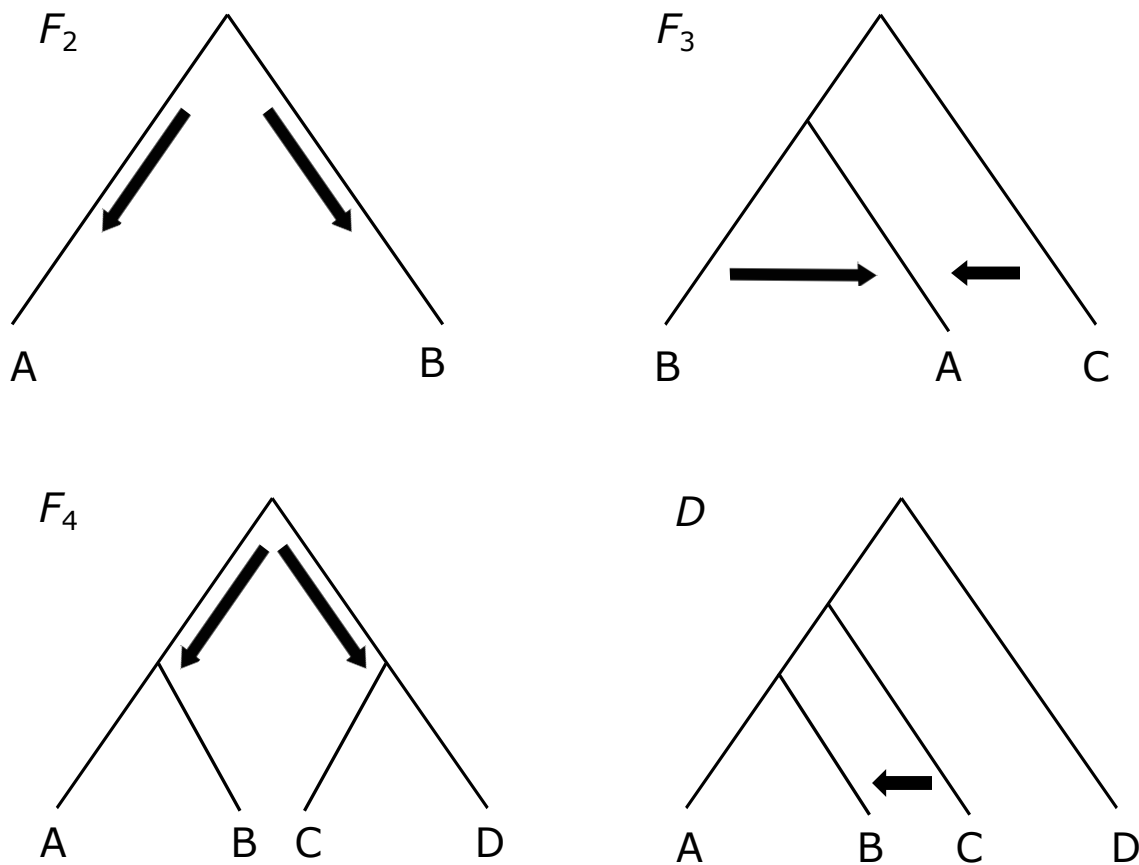


Figure 1: Trees showing the different relationships the F - and D -statistics are designed to test. $F_2(A, B)$ can test the differentiation of two populations A and B . $F_3(A; B, C)$ can test for introgression or relatedness between populations A and B or populations A and C . $F_4(A, B; C, D)$ can test the hypothesis of whether two populations are closer to each other than they are to two other populations, in this case are A and B closer to each other than they are to C and D . The $D(A, B, C, D)$ statistic can test whether there has been admixture between population C and either populations A or B .

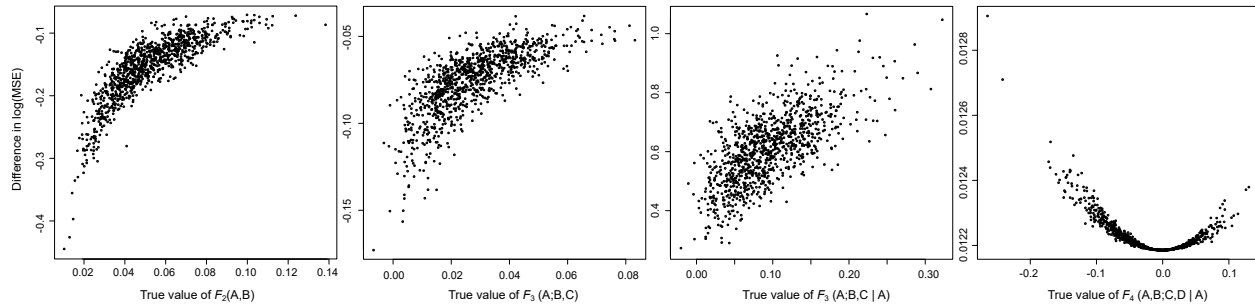


Figure 2: Difference in theoretically calculated $\log_{10}(\text{MSE})$ of \hat{F} and \tilde{F} estimators when including relatives or inbred individuals. The MSE is estimated for instances when samples of 60 individuals include individuals related to exactly one other in the sample, with 10 pairs of avuncular relationships, 10 pairs of inbred full siblings and 10 pairs of outbred full siblings. Each point represents calculations from $J = 20$ randomly sampled loci from the 1000 Genomes Project dataset for CEU, European, YRI African, JPT Japanese, and GIH Indian populations. For $F_2(A, B)$ we use $A = \text{CEU}$ and $B = \text{YRI}$, while for $F_3(A; B, C)$ and $F_3(A; B, C | A)$ we use $A = \text{JPT}$, $B = \text{CEU}$, and $C = \text{YRI}$ and for $F_4(A; B; C, D | A)$ we assign $A = \text{YRI}$, $B = \text{CEU}$, $C = \text{JPT}$, and $D = \text{GIH}$.

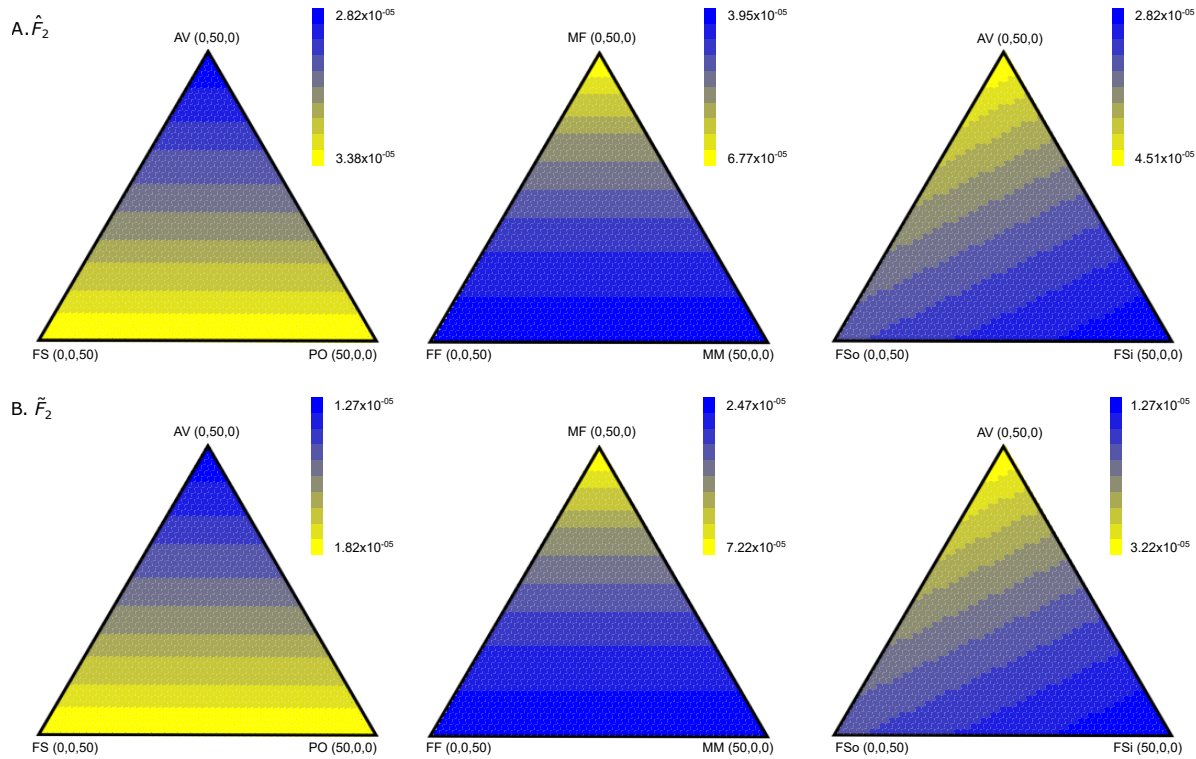


Figure 3: Theoretically calculated MSE of $\hat{F}_2(A, B)$ and $\tilde{F}_2(A, B)$ when including relatives or inbred individuals for $J = 20$ loci. The MSE is estimated for instances when samples of 100 individuals include individuals related to exactly one other in the sample. The first column shows MSE for samples with different combinations of parent-offspring (PO), full sibling (FS), and avuncular (AV) relationships, the second includes full siblings that are male-male (MM), male-female (MF) and female-female (MF). The last column includes AV relationships as well as inbred (FSi) and outbred (FSo) full siblings. The true value of $F_2(A, B)$ is 0.071.

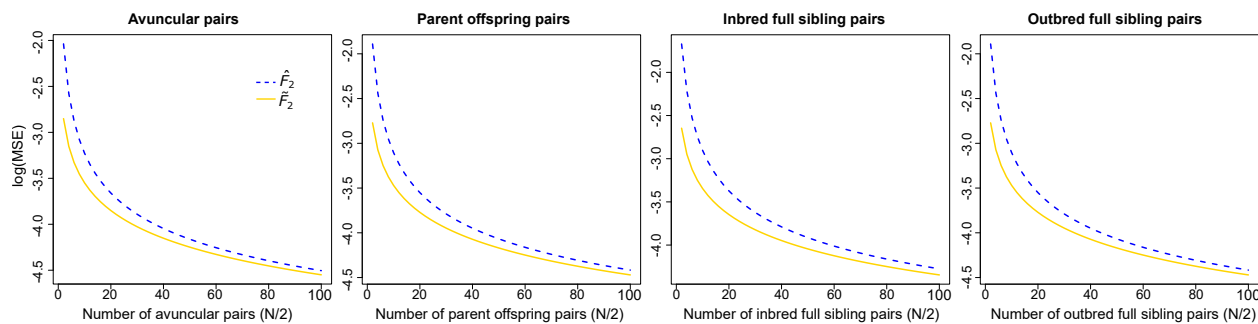


Figure 4: Mean squared error theoretically calculated for $\hat{F}_2(A, B)$ and $\tilde{F}_2(A, B)$ across different sample sizes or related pairs of individuals, including avuncular relationships, parent-offspring relationships, inbred full siblings, and outbred full siblings. The number of sampled individuals ranges from 2 to 100 with the number of relative pairs equaling half the total sampled, all computed using $J = 20$ loci. The true value of $F_2(A, B)$ is 0.071.

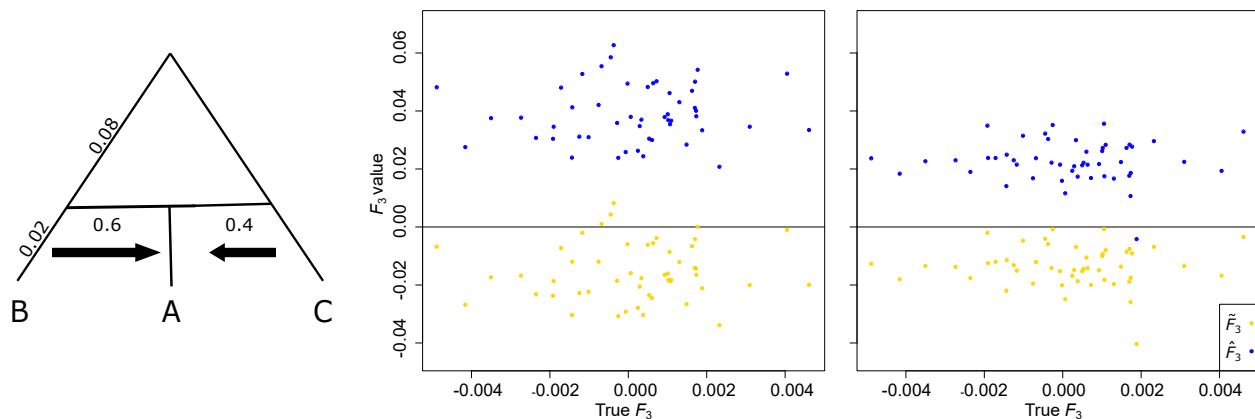


Figure 5: $\hat{F}_3(A; B, C)$ and $\tilde{F}_3(A; B, C)$ calculated for simulations where populations B and C merge with admixture proportions of 0.4 and 0.6, respectively, 400 generations ago (0.02 coalescent units) to form population A . Comparison of simulations with (middle) and without (right) relatives. The panel in the middle shows results for a sample containing 25 parent-offspring pairs, while the panel on the right shows results for 50 unrelated individuals.

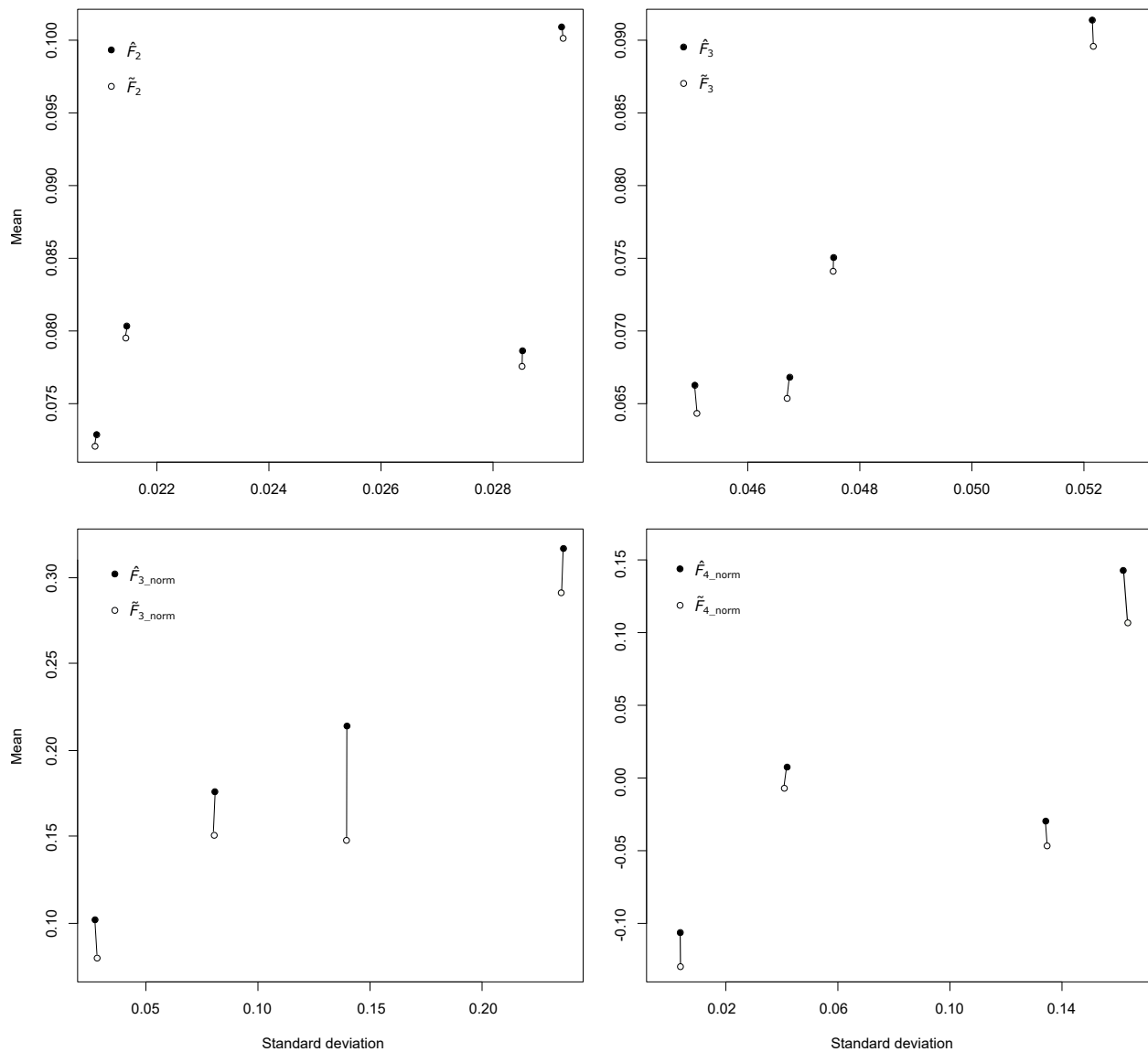


Figure 6: The difference between the means and standard deviations of biased and unbiased estimators of F_2 , normalized and un-normalized F_3 and normalized F_4 when estimated with genotype information from four different combinations of Colombian, Lahu, Melanesian, Mandenka, San, and Druze populations. All of these populations include between two and 14 relative pairs. The black dots represent the values for the biased estimators, while the white dots show the value for the unbiased estimators. Each mean and standard deviation was calculated for a combination of two, three, or four populations, for F_2 , F_3 , and F_4 , respectively, and consists of 1000 estimates of the statistic, each calculated from $J = 20$ randomly samples single nucleotide polymorphisms from the genome. The top row has values for the $F_2(A, B)$ (left) and $F_3(A; B, C)$ (right), while the bottom row shows results for normalized $F_3(A; B, C | A)$ (left) and normalized $F_4(A, B; C, D | A)$ (right).

Appendix

In this section, we provide proofs of key lemmas and propositions from the *Theory* section, and also develop and prove other important results.

Proof of Lemma 1. We first calculate

$$\begin{aligned}\mathbb{E}[\widehat{G}(P_j)] &= \mathbb{E}[\widehat{p}_j(1 - \widehat{p}_j)] \\ &= \mathbb{E}[\widehat{p}_j] - \mathbb{E}[\widehat{p}_j^2] \\ &= p_j - [1 - \Phi_2(P_j)]p_j^2 - \Phi_2(P_j)p_j \\ &= [1 - \Phi_2(P_j)]p_j(1 - p_j) \\ &= [1 - \Phi_2(P_j)]G(P_j),\end{aligned}$$

which gives

$$\begin{aligned}\mathbb{E}[\widehat{G}(P)] &= \frac{1}{J} \sum_{j=1}^J \mathbb{E}[\widehat{G}(P_j)] \\ &= \frac{1}{J} \sum_{j=1}^J [1 - \Phi_2(P_j)]G(P_j) \\ &= G(P) + \Delta(P),\end{aligned}$$

where we define the downward bias of $\widehat{G}(P)$ as

$$\begin{aligned}\Delta(P) &= \mathbb{E}[\widehat{G}(P)] - G(P) \\ &= -\frac{1}{J} \sum_{j=1}^J \Phi_2(P_j)G(P_j).\end{aligned}$$

It follows that $\widetilde{G}(P)$ is an unbiased estimator of $G(P)$ because

$$\begin{aligned}\mathbb{E}[\widetilde{G}(P)] &= \frac{1}{J} \sum_{j=1}^J \frac{1}{1 - \Phi_2(P_j)} \widehat{G}(P_j) \\ &= \frac{1}{J} \sum_{j=1}^J \frac{1}{1 - \Phi_2(P_j)} [1 - \Phi_2(P_j)]G(P_j) \\ &= \frac{1}{J} \sum_{j=1}^J \widehat{G}(P_j) \\ &= G(P).\end{aligned}$$

□

Proof of Proposition 2. We first calculate

$$\begin{aligned}\mathbb{E}[\widehat{F}_2(A_j, B_j)] &= \mathbb{E}[(\widehat{a}_j - \widehat{b}_j)^2] \\ &= \mathbb{E}[\widehat{a}_j^2] + \mathbb{E}[\widehat{b}_j^2] - 2\mathbb{E}[\widehat{a}_j]\mathbb{E}[\widehat{b}_j] \\ &= [1 - \Phi_2(A_j)]a_j^2 + \Phi_2(A_j)a_j + [1 - \Phi_2(B_j)]b_j^2 + \Phi_2(B_j)b_j - 2a_jb_j \\ &= (a_j - b_j)^2 + \Phi_2(A_j)a_j(1 - a_j) + \Phi_2(B_j)b_j(1 - b_j) \\ &= F_2(A_j, B_j) + \Phi_2(A_j)G(A_j) + \Phi_2(B_j)G(B_j),\end{aligned}$$

which gives

$$\begin{aligned}\mathbb{E}[\widehat{F}_2(A, B)] &= \frac{1}{J} \sum_{j=1}^J \mathbb{E}[\widehat{F}_2(A_j, B_j)] \\ &= \frac{1}{J} \sum_{j=1}^J [F_2(A_j, B_j) + \Phi_2(A_j)G(A_j) + \Phi_2(B_j)G(B_j)] \\ &= F_2(A, B) + \Delta(A, B),\end{aligned}$$

where we define the upward bias of $\widehat{F}_2(A, B)$ as

$$\begin{aligned}\Delta(A, B) &= \mathbb{E}[\widehat{F}_2(A, B)] - F_2(A, B) \\ &= \frac{1}{J} \sum_{j=1}^J [\Phi_2(A_j)G(A_j) + \Phi_2(B_j)G(B_j)].\end{aligned}$$

It follows that $\widetilde{F}_2(A, B)$ is an unbiased estimator of $F_2(A, B)$ because

$$\begin{aligned}\mathbb{E}[\widetilde{F}_2(A, B)] &= \frac{1}{J} \sum_{j=1}^J (\mathbb{E}[\widehat{F}_2(A_j, B_j)] - \Phi_2(A_j)\mathbb{E}[\widetilde{G}(A_j)] - \Phi_2(B_j)\mathbb{E}[\widetilde{G}(B_j)]) \\ &= \frac{1}{J} \sum_{j=1}^J [F_2(A_j, B_j) + \Phi_2(A_j)G(A_j) + \Phi_2(B_j)G(B_j) - \Phi_2(A_j)G(A_j) \\ &\quad - \Phi_2(B_j)G(B_j)] \\ &= \frac{1}{J} \sum_{j=1}^J F_2(A_j, B_j) \\ &= F_2(A, B).\end{aligned}$$

□

Proof of Proposition 3. We first calculate

$$\begin{aligned}\mathbb{E}[\widehat{F}_3(A_j; B_j, C_j)] &= \mathbb{E}[(\widehat{a}_j - \widehat{b}_j)(\widehat{a}_j - \widehat{c}_j)] \\ &= \mathbb{E}[\widehat{a}_j^2] - \mathbb{E}[\widehat{a}_j]\mathbb{E}[\widehat{c}_j] - \mathbb{E}[\widehat{a}_j]\mathbb{E}[\widehat{b}_j] + \mathbb{E}[\widehat{b}_j]\mathbb{E}[\widehat{c}_j] \\ &= [1 - \Phi_2(A_j)]a_j^2 + \Phi_2(A_j)a_j - a_j c_j - a_j b_j + b_j c_j \\ &= (a_j^2 - a_j c_j - a_j b_j + b_j c_j) + \Phi_2(A_j)a_j(1 - a_j) \\ &= F_3(A_j; B_j, C_j) + \Phi_2(A_j)G(A_j),\end{aligned}$$

which gives

$$\begin{aligned}\mathbb{E}[\widehat{F}_3(A; B, C)] &= \frac{1}{J} \sum_{j=1}^J \mathbb{E}[\widehat{F}_3(A_j; B_j, C_j)] \\ &= \frac{1}{J} \sum_{j=1}^J [F_3(A; B_j, C_j) + \Phi_2(A_j)G(A_j)] \\ &= F_3(A; B, C) + \Delta(A; B, C),\end{aligned}$$

where we define the upward bias of $\widehat{F}_3(A; B, C)$ as

$$\begin{aligned}\Delta(A; B, C) &= \mathbb{E}[\widehat{F}_3(A; B, C)] - F_3(A; B, C) \\ &= \frac{1}{J} \sum_{j=1}^J \Phi_2(A_j)G(A_j).\end{aligned}$$

It follows that $\tilde{F}_3(A; B, C)$ is an unbiased estimator of $F_3(A; B, C)$ because

$$\begin{aligned}\mathbb{E}[\tilde{F}_3(A; B, C)] &= \frac{1}{J} \sum_{j=1}^J (\mathbb{E}[\hat{F}_3(A_j; B_j, C_j)] - \Phi_2(A_j)\mathbb{E}[\tilde{G}(A_j)]) \\ &= \frac{1}{J} \sum_{j=1}^J [F_3(A_j; B_j, C_j) + \Phi_2(A_j)G(A_j) - \Phi_2(A_j)G(A_j)] \\ &= \frac{1}{J} \sum_{j=1}^J F_3(A_j; B_j, C_j) \\ &= F_3(A; B, C).\end{aligned}$$

□

Proof of Proposition 4. Assuming that the expectation of $\hat{F}_3(A; B, C | A)$ is approximately equal to the ratio of expectations of $\hat{F}_3(A; B, C)$ and $2\hat{G}(A)$, we find that

$$\begin{aligned}\mathbb{E}[\hat{F}_3(A; B, C | A)] &= \mathbb{E}\left[\frac{\hat{F}_3(A; B, C)}{2\hat{G}(A)}\right] \\ &\approx \frac{\mathbb{E}[\hat{F}_3(A; B, C)]}{2\mathbb{E}[\hat{G}(A)]} \\ &= \frac{(1/J) \sum_{j=1}^J [F_3(A_j; B_j, C_j) + \Phi_2(A_j)G(A_j)]}{(2/J) \sum_{j=1}^J [1 - \Phi_2(A_j)]G(A_j)} \\ &= \frac{F_3(A; B, C) + (1/J) \sum_{j=1}^J \Phi_2(A_j)G(A_j)}{2G(A) - (2/J) \sum_{j=1}^J \Phi_2(A_j)G(A_j)} \\ &= \frac{F_3(A; B, C)}{2G(A)} + \Delta(A; B, C | A) \\ &= F_3(A; B, C, D | A) + \Delta(A; B, C | A),\end{aligned}$$

where we define the upward bias of $\hat{F}_3(A; B, C | A)$ as

$$\begin{aligned}\Delta(A; B, C | A) &= \mathbb{E}[\hat{F}_3(A; B, C | A)] - F_3(A; B, C | A) \\ &= \frac{F_3(A; B, C) + (1/J) \sum_{j=1}^J \Phi_2(A_j)G(A_j)}{2G(A) - (2/J) \sum_{j=1}^J \Phi_2(A_j)G(A_j)} - \frac{F_3(A; B, C)}{2G(A)} \\ &= \frac{[F_3(A; B, C) + G(A)](1/J) \sum_{j=1}^J \Phi_2(A_j)G(A_j)}{2G(A)[G(A) - (1/J) \sum_{j=1}^J \Phi_2(A_j)G(A_j)]} \\ &= \frac{(1/J) \sum_{j=1}^J \Phi_2(A_j)G(A_j)}{G(A) - (1/J) \sum_{j=1}^J \Phi_2(A_j)G(A_j)} \left[F_3(A; B, C | A) + \frac{1}{2} \right].\end{aligned}$$

We can also see that $\tilde{F}_3(A; B, C | A)$ is an approximately unbiased estimator of $F_3(A; B, C | A)$ because

$$\begin{aligned}\mathbb{E}[\tilde{F}_3(A; B, C | A)] &= E\left[\frac{\tilde{F}_3(A; B, C)}{2\tilde{G}(A)}\right] \\ &\approx \frac{\mathbb{E}[\tilde{F}_3(A; B, C)]}{2\mathbb{E}[\tilde{G}(A)]} \\ &= \frac{F_3(A; B, C)}{2G(A)} \\ &= F_3(A; B, C | A).\end{aligned}$$

□

Proof of Proposition 5. We first calculate

$$\begin{aligned}\mathbb{E}[\widehat{F}_4(A_j, B_j; C_j, D_j)] &= \mathbb{E}[(\widehat{a}_j - \widehat{b}_j)(\widehat{c}_j - \widehat{d}_j)] \\ &= \mathbb{E}[\widehat{a}_j - \widehat{b}_j]\mathbb{E}[\widehat{c}_j - \widehat{d}_j] \\ &= (\mathbb{E}[\widehat{a}_j] - \mathbb{E}[\widehat{b}_j])(\mathbb{E}[\widehat{c}_j] - \mathbb{E}[\widehat{d}_j]) \\ &= (a_j - b_j)(c_j - d_j) \\ &= F_4(A_j, B_j; C_j, D_j).\end{aligned}$$

We show that $\widehat{F}_4(A, B; C, D)$ is unbiased estimator of $F_4(A, B; C, D)$ because

$$\begin{aligned}\mathbb{E}[\widehat{F}_4(A, B; C, D)] &= \frac{1}{J} \sum_{j=1}^J \mathbb{E}[\widehat{F}_4(A_j, B_j; C_j, D_j)] \\ &= \frac{1}{J} \sum_{j=1}^J F_4(A_j, B_j; C_j, D_j) \\ &= F_4(A, B; C, D).\end{aligned}$$

□

Proof of Proposition 6. Assuming that the expectation of $\widehat{F}_4(A, B; C, D | P)$ is approximately equal to the ratio of expectations of $\widehat{F}_4(A, B; C, D)$ and $\widehat{G}(P)$ for some $P \in \{A, B, C, D\}$, we find that

$$\begin{aligned}\mathbb{E}[\widehat{F}_4(A, B; C, D | P)] &= \mathbb{E}\left[\frac{\widehat{F}_4(A, B; C, D)}{\widehat{G}(P)}\right] \\ &\approx \frac{\mathbb{E}[\widehat{F}_4(A, B; C, D)]}{\mathbb{E}[\widehat{G}(P)]} \\ &= \frac{F_4(A, B; C, D)}{(1/J) \sum_{j=1}^J [1 - \Phi_2(P_j)]G(P_j)} \\ &= \frac{F_4(A, B; C, D)}{G(P) - (1/J) \sum_{j=1}^J \Phi_2(P_j)G(P_j)} \\ &= \frac{F_4(A, B; C, D)}{G(P)} + \Delta(A, B; C, D | P) \\ &= F_4(A, B; C, D | P) + \Delta(A, B; C, D | P),\end{aligned}$$

where we define the approximate upward bias of $\widehat{F}_4(A, B; C, D | P)$ as

$$\begin{aligned}\Delta(A, B; C, D | P) &= \mathbb{E}[\widehat{F}_4(A, B; C, D | P)] - F_4(A, B; C, D | P) \\ &= \frac{F_4(A, B; C, D)}{G(P) - (1/J) \sum_{j=1}^J \Phi_2(P_j)G(P_j)} - \frac{F_4(A, B; C, D)}{G(P)} \\ &= \frac{F_4(A, B; C, D)(1/J) \sum_{j=1}^J \Phi_2(P_j)G(P_j)}{G(P)[G(P) - (1/J) \sum_{j=1}^J \Phi_2(P_j)G(P_j)]} \\ &= \frac{(1/J) \sum_{j=1}^J \Phi_2(P_j)G(P_j)}{G(P) - (1/J) \sum_{j=1}^J \Phi_2(P_j)G(P_j)} F_4(A, B; C, D | P).\end{aligned}$$

We can also see that $\tilde{F}_4(A, B; C, D | P)$ is an approximately unbiased estimator of $F_4(A, B; C, D | P)$ because

$$\begin{aligned}\mathbb{E}[\tilde{F}_4(A, B; C, D | P)] &= E\left[\frac{\hat{F}_4(A, B; C, D)}{\hat{G}(P)}\right] \\ &\approx \frac{\mathbb{E}[\hat{F}_4(A, B; C, D)]}{\mathbb{E}[\hat{G}(P)]} \\ &= \frac{F_4(A, B; C, D)}{G(P)} \\ &= F_4(A, B; C, D | P).\end{aligned}\quad \square$$

Lemma 8. Consider J polymorphic loci in populations A, B, C , and D with respective parametric reference allele frequencies $a_j, b_j, c_j, d_j \in (0, 1)$, and suppose we take a random sample of $N(P_j)$ individuals at locus j in population $P \in \{A, B, C, D\}$, some of which may be related or inbred. The estimator $\hat{H}(A, B, C, D)$ is unbiased.

Proof. We first calculate

$$\begin{aligned}\mathbb{E}[\hat{H}(A_j, B_j, C_j, D_j)] &= \mathbb{E}[(\hat{a}_j + \hat{b}_j - 2\hat{a}_j\hat{b}_j)(\hat{c}_j + \hat{d}_j - 2\hat{c}_j\hat{d}_j)] \\ &= \mathbb{E}[(\hat{a}_j + \hat{b}_j - 2\hat{a}_j\hat{b}_j)]\mathbb{E}[(\hat{c}_j + \hat{d}_j - 2\hat{c}_j\hat{d}_j)] \\ &= (\mathbb{E}[\hat{a}_j] + \mathbb{E}[\hat{b}_j] - 2\mathbb{E}[\hat{a}_j]\mathbb{E}[\hat{b}_j])(\mathbb{E}[\hat{c}_j] + \mathbb{E}[\hat{d}_j] - 2\mathbb{E}[\hat{c}_j]\mathbb{E}[\hat{d}_j]) \\ &= (a_j + b_j - 2a_jb_j)(c_j + d_j - 2c_jd_j) \\ &= H(A_j, B_j, C_j, D_j).\end{aligned}$$

We show that $\hat{H}(A, B, C, D)$ is unbiased estimator of $H(A, B, C, D)$ because

$$\begin{aligned}\mathbb{E}[\hat{H}(A, B, C, D)] &= \frac{1}{J} \sum_{j=1}^J \mathbb{E}[\hat{H}(A_j, B_j, C_j, D_j)] \\ &= \frac{1}{J} \sum_{j=1}^J H(A_j, B_j, C_j, D_j) \\ &= H(A, B, C, D).\end{aligned}\quad \square$$

Proof of Proposition 7. Assuming that the expectation of $\hat{D}(A, B, C, D)$ is approximately equal to the ratio of expectations of $-\hat{F}_4(A, B; C, D)$ and $\hat{H}(A, B, C, D)$, $\hat{D}(A, B, C, D)$ is an approximately unbiased estimator of $D(A, B, C, D)$ because

$$\begin{aligned}\mathbb{E}[\hat{D}(A, B, C, D)] &= -\mathbb{E}\left[\frac{\hat{F}_4(A, B; C, D)}{\hat{H}(A, B, C, D)}\right] \\ &\approx -\frac{\mathbb{E}[\hat{F}_4(A, B; C, D)]}{\mathbb{E}[\hat{H}(A, B, C, D)]} \\ &= -\frac{F_4(A, B; C, D)}{H(A, B, C, D)} \\ &= D(A, B, C, D).\end{aligned}\quad \square$$

Lemma 9. Consider J independent polymorphic loci in a population P with parametric reference allele frequencies $p_j \in (0, 1)$, and suppose we take a random sample of $N(P_j)$ individuals at locus j , some of which may be related or inbred. Moreover, assume that no individual is related to more than one other individual, which makes the terms $\Phi_3(P_j)$, $\Phi_4(P_j)$, $\Phi_{2,2}(P_j)$, and $\Phi_2(P_j)^2$ negligible to $\Phi_2(P_j)$. Based on this simplifying assumption, the estimator $\hat{G}(P)$ has an approximate variance

$$Var[\hat{G}(P)] \approx \frac{1}{J^2} \sum_{j=1}^J \Phi_2(P_j)G(P_j) - \frac{4}{J^2} \sum_{j=1}^J \Phi_2(P_j)G(P_j)^2.$$

Moreover, the respective approximate variance for the unbiased estimator $\tilde{G}(P)$ is

$$Var[\tilde{G}(P)] \approx \frac{1}{J^2} \sum_{j=1}^J \frac{\Phi_2(P_j)}{1 - 2\Phi_2(P_j)} G(P_j) - \frac{4}{J^2} \sum_{j=1}^J \frac{\Phi_2(P_j)}{1 - 2\Phi_2(P_j)} G(P_j)^2.$$

Proof. From the proof of Lemma 1, we have

$$\mathbb{E}[\hat{G}(P_j)] = [1 - \Phi_2(P_j)]G(P_j)$$

and we calculate

$$\begin{aligned} \mathbb{E}[\hat{G}(P_j)^2] &= \mathbb{E}[\hat{p}_j^2(1 - \hat{p}_j)^2] \\ &= \mathbb{E}[\hat{p}_j^2] - 2\mathbb{E}[\hat{p}_j^3] + \mathbb{E}[\hat{p}_j^4] \\ &\approx p_j^2 + \Phi_2(P_j)p_j(1 - p_j) - 2[p_j^3 + 3\Phi_2(P_j)p_j^2(1 - p_j)] + p_j^4 + 6\Phi_2(P_j)p_j^3(1 - p_j) \\ &= p_j^2 - 2p_j^3 + p_j^4 + \Phi_2(P_j)p_j(1 - p_j)[1 - 6p_j + 6p_j^2] \\ &= p_j^2(1 - p_j)^2 + \Phi_2(P_j)p_j(1 - p_j)[1 - 6p_j(1 - p_j)] \\ &= G(P_j)^2 + \Phi_2(P_j)G(P_j)[1 - 6G(P_j)] \\ &= \Phi_2(P_j)G(P_j) + [1 - 6\Phi_2(P_j)]G(P_j)^2. \end{aligned}$$

Therefore, we have that

$$\begin{aligned} Var[\hat{G}(P_j)] &= \mathbb{E}[\hat{G}(P_j)^2] - \mathbb{E}[\hat{G}(P_j)]^2 \\ &\approx \Phi_2(P_j)G(P_j) + [1 - 6\Phi_2(P_j)]G(P_j)^2 - [1 - \Phi_2(P_j)]^2G(P_j)^2 \\ &= \Phi_2(P_j)G(P_j) + [1 - 6\Phi_2(P_j)]G(P_j)^2 - [1 - 2\Phi_2(P_j) + \Phi_2(P_j)^2]G(P_j)^2 \\ &\approx \Phi_2(P_j)G(P_j) + [1 - 6\Phi_2(P_j)]G(P_j)^2 - [1 - 2\Phi_2(P_j)]G(P_j)^2 \\ &= \Phi_2(P_j)G(P_j) - 4\Phi_2(P_j)G(P_j)^2, \end{aligned}$$

which gives

$$\begin{aligned} Var[\hat{G}(P)] &= Var\left[\frac{1}{J} \sum_{j=1}^J \hat{G}(P_j)\right] \\ &= \frac{1}{J^2} \sum_{j=1}^J Var[\hat{G}(P_j)] \\ &\approx \frac{1}{J^2} \sum_{j=1}^J \Phi_2(P_j)G(P_j) - \frac{4}{J^2} \sum_{j=1}^J \Phi_2(P_j)G(P_j)^2. \end{aligned}$$

Recall that

$$\tilde{G}(P) = \frac{1}{J} \sum_{j=1}^J \tilde{G}(P_j),$$

where

$$\tilde{G}(P_j) = \frac{1}{1 - \Phi_2(P_j)} \hat{G}(P_j).$$

It follows that

$$\begin{aligned}
 \text{Var}[\tilde{G}(P)] &= \text{Var}\left[\frac{1}{J} \sum_{j=1}^J \tilde{G}(P_j)\right] \\
 &= \frac{1}{J^2} \sum_{j=1}^J \text{Var}[\tilde{G}(P_j)] \\
 &= \frac{1}{J^2} \sum_{j=1}^J \frac{1}{[1 - \Phi_2(P_j)]^2} \text{Var}[\hat{G}(P_j)] \\
 &\approx \frac{1}{J^2} \sum_{j=1}^J \frac{\Phi_2(P_j)}{1 - 2\Phi_2(P_j)} G(P_j) - \frac{4}{J^2} \sum_{j=1}^J \frac{\Phi_2(P_j)}{1 - 2\Phi_2(P_j)} G(P_j)^2. \quad \square
 \end{aligned}$$

Lemma 10. Consider J independent polymorphic loci in populations A and B with respective parametric reference allele frequencies $a_j, b_j \in (0, 1)$, and suppose we take a random sample of $N(P_j)$ individuals at locus j in population $P \in \{A, B\}$, some of which may be related or inbred. Moreover, assume that no individual is related to more than one other individual, which makes the terms $\Phi_3(P_j)$, $\Phi_4(P_j)$, $\Phi_{2,2}(P_j)$, and $\Phi_2(P_j)^2$ negligible to $\Phi_2(P_j)$. Based on this simplifying assumption, the estimators $\hat{F}_2(A, B)$ and $\hat{G}(B)$ have an approximate covariance

$$\begin{aligned}
 \text{Cov}[\hat{F}_2(A, B), \hat{G}(B)] &\approx \frac{2}{J^2} \sum_{j=1}^J \Phi_2(B_j) G(B_j)^2 - \frac{2}{J^2} \sum_{j=1}^J \Phi_2(B_j) G(A_j) G(B_j) \\
 &\quad - \frac{2}{J^2} \sum_{j=1}^J \Phi_2(B_j) F_2(A_j, B_j) G(B_j).
 \end{aligned}$$

Proof. From the proofs of Lemma 1 and Proposition 2, we have

$$\mathbb{E}[\hat{G}(B_j)] = [1 - \Phi_2(B_j)] G(B_j)$$

and

$$\mathbb{E}[\hat{F}_2(A_j, B_j)] = F_2(A_j, B_j) + \Phi_2(A_j) G(A_j) + \Phi_2(B_j) G(B_j),$$

yielding

$$\begin{aligned}
 \mathbb{E}[\hat{F}_2(A_j, B_j)] \mathbb{E}[\hat{G}(B_j)] &= [F_2(A_j, B_j) + \Phi_2(A_j) G(A_j) + \Phi_2(B_j) G(B_j)] [1 - \Phi_2(B_j)] G(B_j) \\
 &= F_2(A_j, B_j) G(B_j) + \Phi_2(A_j) G(A_j) G(B_j) + \Phi_2(B_j) G(B_j)^2 \\
 &\quad - \Phi_2(B_j) F_2(A_j, B_j) G(B_j) - \Phi_2(A_j) \Phi_2(B_j) G(A_j) G(B_j) \\
 &\quad - \Phi_2(B_j)^2 G(B_j)^2 \\
 &\approx F_2(A_j, B_j) G(B_j) + \Phi_2(A_j) G(A_j) G(B_j) + \Phi_2(B_j) G(B_j)^2 \\
 &\quad - \Phi_2(B_j) F_2(A_j, B_j) G(B_j) - \Phi_2(A_j) \Phi_2(B_j) G(A_j) G(B_j),
 \end{aligned}$$

where we used the fact that $\Phi_2(B_j)^2$ is negligible compared to $\Phi_2(B_j)$ as an approximation. We also calculate

$$\begin{aligned}
 \mathbb{E}[\hat{F}_2(A_j, B_j) \hat{G}(B_j)] &= \mathbb{E}[(\hat{a}_j - \hat{b}_j)^2 \hat{b}_j (1 - \hat{b}_j)] \\
 &= \mathbb{E}[(\hat{a}_j^2 - 2\hat{a}_j \hat{b}_j + \hat{b}_j^2)(\hat{b}_j - \hat{b}_j^2)] \\
 &= \mathbb{E}[\hat{a}_j^2 (\hat{b}_j - \hat{b}_j^2) - 2\hat{a}_j (\hat{b}_j^2 - \hat{b}_j^3) + \hat{b}_j^3 - \hat{b}_j^4] \\
 &= \mathbb{E}[\hat{a}_j^2] (\mathbb{E}[\hat{b}_j] - \mathbb{E}[\hat{b}_j^2]) - 2\mathbb{E}[\hat{a}_j] (\mathbb{E}[\hat{b}_j^2] - \mathbb{E}[\hat{b}_j^3]) + \mathbb{E}[\hat{b}_j^3] - \mathbb{E}[\hat{b}_j^4] \\
 &\approx [a_j^2 + \Phi_2(A_j) a_j (1 - a_j)] [b - b_j^2 - \Phi_2(B_j) b_j (1 - b_j)] \\
 &\quad - 2a_j [b_j^2 + \Phi_2(B_j) b_j (1 - b_j) - b_j^3 - 3\Phi_2(B_j) b_j^2 (1 - b_j)] \\
 &\quad + b_j^3 + 3\Phi_2(B_j) b_j^2 (1 - b_j) - b_j^4 - 6\Phi_2(B_j) b_j^3 (1 - b_j).
 \end{aligned}$$

Recognizing that $G(A_j) = a_j(1 - a_j)$, $G(B_j) = g_j(1 - g_j)$, and $F_2(A_j, B_j) = a_j^2 - 2a_jb_j + b_j^2$, we have

$$\begin{aligned}
 \mathbb{E}[\widehat{F}_2(A_j, B_j)\widehat{G}(B_j)] &\approx [a_j^2 + \Phi_2(A_j)G(A_j)][1 - \Phi_2(B_j)]G(B_j) \\
 &\quad - 2a_j[\Phi_2(B_j) + [1 - 3\Phi_2(B_j)]b_j]G(B_j) \\
 &\quad + [3\Phi_2(B_j)b_j + [1 - 6\Phi_2(B_j)]b_j^2]G(B_j) \\
 &= G(B_j)\left[a_j^2 - \Phi_2(B_j)a_j^2 + \Phi_2(A_j)G(A_j) - \Phi_2(A_j)\Phi_2(B_j)G(A_j) \right. \\
 &\quad \left. - 2\Phi_2(B_j)a_j - 2a_jb_j + 2(3)\Phi_2(B_j)a_jb_j + 3\Phi_2(B_j)b_j + b_j^2 \right. \\
 &\quad \left. - 6\Phi_2(B_j)b_j^2\right] \\
 &= G(B_j)\left[F_2(A_j, B_j) - \Phi_2(B_j)[3a_j^2 - 2(3)a_jb_j + 3b_j^2 - 2a_j^2 + 3b_j^2] \right. \\
 &\quad \left. - 2\Phi_2(B_j)a_j + 3\Phi_2(B_j)b_j + [\Phi_2(A_j) - \Phi_2(A_j)\Phi_2(B_j)]G(A_j)\right] \\
 &= G(B_j)\left[F_2(A_j, B_j) - 3\Phi_2(B_j)[a_j^2 - 2a_jb_j + b_j^2] \right. \\
 &\quad \left. - 2\Phi_2(B_j)[a_j - a_j^2] + 3\Phi_2(B_j)[b_j - b_j^2] \right. \\
 &\quad \left. + [\Phi_2(A_j) - \Phi_2(A_j)\Phi_2(B_j)]G(A_j)\right] \\
 &= G(B_j)\left[F_2(A_j, B_j) - 3\Phi_2(B_j)F_2(A_j, B_j) - 2\Phi_2(B_j)G(A_j) + 3\Phi_2(B_j)G(B_j) \right. \\
 &\quad \left. + [\Phi_2(A_j) - \Phi_2(A_j)\Phi_2(B_j)]G(A_j)\right] \\
 &= [1 - 3\Phi_2(B_j)]F_2(A_j, B_j)G(B_j) + 3\Phi_2(B_j)G(B_j)^2 \\
 &\quad + [\Phi_2(A_j) - 2\Phi_2(B_j) - \Phi_2(A_j)\Phi_2(B_j)]G(A_j)G(B_j).
 \end{aligned}$$

Therefore, we have that

$$\begin{aligned}
 \text{Cov}[\widehat{F}_2(A_j, B_j), \widehat{G}(B_j)] &= \mathbb{E}[\widehat{F}_2(A_j, B_j)\widehat{G}(B_j)] - \mathbb{E}[\widehat{F}_2(A_j, B_j)]\mathbb{E}[\widehat{G}(B_j)] \\
 &\approx [1 - 3\Phi_2(B_j)]F_2(A_j, B_j)G(B_j) + 3\Phi_2(B_j)G(B_j)^2 \\
 &\quad + [\Phi_2(A_j) - 2\Phi_2(B_j) - \Phi_2(A_j)\Phi_2(B_j)]G(A_j)G(B_j) \\
 &\quad - \left[F_2(A_j, B_j)G(B_j) + \Phi_2(A_j)G(A_j)G(B_j) + \Phi_2(B_j)G(B_j)^2 \right. \\
 &\quad \left. - \Phi_2(B_j)F_2(A_j, B_j)G(B_j) - \Phi_2(A_j)\Phi_2(B_j)G(A_j)G(B_j)\right] \\
 &= 2\Phi_2(B_j)G(B_j)^2 - 2\Phi_2(B_j)G(A_j)G(B_j) - 2\Phi_2(B_j)F_2(A_j, B_j)G(B_j) \\
 &= 2\Phi_2(B_j)G(B_j)[G(B_j) - G(A_j) - F_2(A_j, B_j)],
 \end{aligned}$$

which gives

$$\begin{aligned}
 \text{Cov}[\widehat{F}_2(A, B), \widehat{G}(B)] &= \text{Cov}\left[\frac{1}{J} \sum_{j=1}^J \widehat{F}_2(A_j, B_j), \frac{1}{J} \sum_{j=1}^J \widehat{G}(B_j)\right] \\
 &= \frac{1}{J^2} \sum_{j=1}^J \text{Cov}[\widehat{F}_2(A_j, B_j), \widehat{G}(B_j)] \\
 &\approx \frac{2}{J^2} \sum_{j=1}^J \Phi_2(B_j)G(B_j)^2 - \frac{2}{J^2} \sum_{j=1}^J \Phi_2(B_j)G(A_j)G(B_j) \\
 &\quad - \frac{2}{J^2} \sum_{j=1}^J \Phi_2(B_j)F_2(A_j, B_j)G(B_j).
 \end{aligned}$$

□

Proposition 11. Consider J independent polymorphic loci in a populations A and B with respective parametric reference allele frequencies $a_j, b_j \in (0, 1)$, and suppose we take a random sample of $N(P_j)$ individuals at locus j in population $P \in \{A, B\}$, some of which may be related or inbred. Moreover, assume that no individual is related to more than one other individual, which makes the terms $\Phi_3(P_j)$, $\Phi_4(P_j)$, $\Phi_{2,2}(P_j)$, $\Phi_2(P_j)^2$, and $\Phi_2(A_j)\Phi_2(B_j)$ negligible to $\Phi_2(P_j)$. Based on this simplifying assumption, the estimator $\widehat{F}_2(A, B)$ has approximate variance

$$Var[\widehat{F}_2(A, B)] \approx \frac{4}{J^2} \sum_{j=1}^J \Phi_2(A_j)F_2(A_j, B_j)G(A_j) + \frac{4}{J^2} \sum_{j=1}^J \Phi_2(B_j)F_2(A_j, B_j)G(B_j).$$

Proof. From the proof of Proposition 2, we have

$$\mathbb{E}[\widehat{F}_2(A_j, B_j)] = F_2(A_j, B_j) + \Phi_2(A_j)G(A_j) + \Phi_2(B_j)G(B_j),$$

which gives

$$\begin{aligned} \mathbb{E}[\widehat{F}_2(A_j, B_j)]^2 &= [F_2(A_j, B_j) + \Phi_2(A_j)G(A_j) + \Phi_2(B_j)G(B_j)]^2 \\ &= F_2(A_j, B_j)^2 + 2\Phi_2(A_j)F_2(A_j, B_j)G(A_j) + 2\Phi_2(B_j)F_2(A_j, B_j)G(B_j) \\ &\quad + 2\Phi_2(A_j)\Phi_2(B_j)G(A_j)G(B_j) + \Phi_2(A_j)^2G(A_j)^2 + \Phi_2(B_j)^2G(B_j)^2 \\ &\approx F_2(A_j, B_j)^2 + 2\Phi_2(A_j)F_2(A_j, B_j)G(A_j) + 2\Phi_2(B_j)F_2(A_j, B_j)G(B_j), \end{aligned}$$

where we used the fact that $\Phi_2(A_j)^2$, $\Phi_2(B_j)^2$, and $\Phi_2(A_j)\Phi_2(B_j)$ are negligible compared to $\Phi_2(A_j)$ and $\Phi_2(B_j)$ as an approximation. We also calculate

$$\begin{aligned} \mathbb{E}[\widehat{F}_2(A_j, B_j)^2] &= \mathbb{E}[(\widehat{a}_j - \widehat{b}_j)^4] \\ &= \mathbb{E}[\widehat{a}_j^4] - 4\mathbb{E}[\widehat{a}_j^3]\mathbb{E}[\widehat{b}_j] + 6\mathbb{E}[\widehat{a}_j^2]\mathbb{E}[\widehat{b}_j^2] - 4\mathbb{E}[\widehat{a}_j]\mathbb{E}[\widehat{b}_j^3] + \mathbb{E}[\widehat{b}_j^4] \\ &\approx a_j^4 + 6\Phi_2(A_j)a_j^3(1-a_j) - 4[a_j^3 + 3\Phi_2(A_j)a_j^2(1-a_j)]b_j \\ &\quad + 6[a_j^2 + \Phi_2(A_j)a_j(1-a_j)][b_j^2 + \Phi_2(B_j)b_j(1-b_j)] \\ &\quad - 4a_j[b_j^3 + 3\Phi_2(B_j)b_j^2(1-b_j)] + b_j^4 + 6\Phi_2(B_j)b_j^3(1-b_j) \\ &= a_j^4 - 4a_j^3b_j + 6a_j^2b_j^2 - 4a_jb_j^3 + b_j^4 + 6\Phi_2(A_j)a_j(1-a_j)[a_j^2 - 2a_jb_j + b_j^2] \\ &\quad + 6\Phi_2(B_j)b_j(1-b_j)[a_j^2 - 2a_jb_j + b_j^2] + 6\Phi_2(A_j)\Phi_2(B_j)a_j(1-a_j)b_j(1-b_j) \\ &= (a_j - b_j)^4 + 6\Phi_2(A_j)a_j(1-a_j)(a_j - b_j)^2 + 6\Phi_2(B_j)b_j(1-b_j)(a_j - b_j)^2 \\ &\quad + 6\Phi_2(A_j)\Phi_2(B_j)a_j(1-a_j)b_j(1-b_j) \\ &= F_2(A_j, B_j)^2 + 6\Phi_2(A_j)F_2(A_j, B_j)G(A_j) + 6\Phi_2(B_j)F_2(A_j, B_j)G(B_j) \\ &\quad + 6\Phi_2(A_j)\Phi_2(B_j)G(A_j)G(B_j) \\ &\approx F_2(A_j, B_j)^2 + 6\Phi_2(A_j)F_2(A_j, B_j)G(A_j) + 6\Phi_2(B_j)F_2(A_j, B_j)G(B_j). \end{aligned}$$

Therefore, we have that

$$\begin{aligned} Var[\widehat{F}_2(A_j, B_j)] &= \mathbb{E}[\widehat{F}_2(A_j, B_j)^2] - \mathbb{E}[\widehat{F}_2(A_j, B_j)]^2 \\ &\approx F_2(A_j, B_j)^2 + 6\Phi_2(A_j)F_2(A_j, B_j)G(A_j) + 6\Phi_2(B_j)F_2(A_j, B_j)G(B_j) \\ &\quad - \left[F_2(A_j, B_j)^2 + 2\Phi_2(A_j)F_2(A_j, B_j)G(A_j) + 2\Phi_2(B_j)F_2(A_j, B_j)G(B_j) \right] \\ &= 4\Phi_2(A_j)F_2(A_j, B_j)G(A_j) + 4\Phi_2(B_j)F_2(A_j, B_j)G(B_j), \end{aligned}$$

which gives

$$\begin{aligned} \text{Var}[\widehat{F}_2(A, B)] &= \text{Var}\left[\frac{1}{J} \sum_{j=1}^J \widehat{F}_2(A_j, B_j)\right] \\ &= \frac{1}{J^2} \sum_{j=1}^J \text{Var}[\widehat{F}_2(A_j, B_j)] \\ &\approx \frac{4}{J^2} \sum_{j=1}^J \Phi_2(A_j)F_2(A_j, B_j)G(A_j) + \frac{4}{J^2} \sum_{j=1}^J \Phi_2(B_j)F_2(A_j, B_j)G(B_j). \quad \square \end{aligned}$$

Proposition 12. Consider J independent polymorphic loci in a populations A and B with respective parametric reference allele frequencies $a_j, b_j \in (0, 1)$, and suppose we take a random sample of $N(P_j)$ individuals at locus j in population $P \in \{A, B\}$, some of which may be related or inbred. Moreover, assume that no individual is related to more than one other individual, which makes the terms $\Phi_3(P_j)$, $\Phi_4(P_j)$, $\Phi_{2,2}(P_j)$, $\Phi_2(P_j)^2$, and $\Phi_2(A_j)\Phi_2(B_j)$ negligible to $\Phi_2(P_j)$. Based on this simplifying assumption, the unbiased estimator $\widetilde{F}_2(A, B)$ has approximate variance

$$\text{Var}[\widetilde{F}_2(A, B)] \approx \frac{4}{J^2} \sum_{j=1}^J \Phi_2(A_j)F_2(A_j, B_j)G(A_j) + \frac{4}{J^2} \sum_{j=1}^J \Phi_2(B_j)F_2(A_j, B_j)G(B_j).$$

Proof. Recall that

$$\widetilde{F}_2(A_j, B_j) = \widehat{F}_2(A_j, B_j) - \Phi_2(A_j)\widetilde{G}(A_j) - \Phi_2(B_j)\widetilde{G}(B_j),$$

where $\widetilde{F}_2(A_j, B_j)$ is an unbiased estimator for $F_2(A_j, B_j)$ and $\widetilde{G}(P_j)$ is an unbiased estimator of $G(P_j)$ for $P \in \{A, B\}$ at locus $j \in \{1, 2, \dots, J\}$. Also, from the proof of Proposition 2, we have

$$\mathbb{E}[\widehat{F}_2(A_j, B_j)] = F_2(A_j, B_j) + \Phi_2(A_j)G(A_j) + \Phi_2(B_j)G(B_j).$$

Therefore, we have that

$$\begin{aligned} \text{Var}[\widetilde{F}_2(A_j, B_j)] &= \text{Var}[\widehat{F}_2(A_j, B_j) - \Phi_2(A_j)\widetilde{G}(A_j) - \Phi_2(B_j)\widetilde{G}(B_j)] \\ &= \text{Var}[\widehat{F}_2(A_j, B_j)] + \Phi_2(A_j)^2 \text{Var}[\widetilde{G}(A_j)] + \Phi_2(B_j)^2 \text{Var}[\widetilde{G}(B_j)] \\ &\quad - 2\Phi_2(A_j)\text{Cov}[\widehat{F}_2(A_j, B_j), \widetilde{G}(A_j)] - 2\Phi_2(B_j)\text{Cov}[\widehat{F}_2(A_j, B_j), \widetilde{G}(B_j)] \\ &\quad + 2\Phi_2(A_j)\Phi_2(B_j)\text{Cov}[\widetilde{G}(A_j), \widetilde{G}(B_j)] \\ &\approx \text{Var}[\widehat{F}_2(A_j, B_j)] - 2\Phi_2(A_j)\text{Cov}[\widehat{F}_2(A_j, B_j), \widetilde{G}(A_j)] \\ &\quad - 2\Phi_2(B_j)\text{Cov}[\widehat{F}_2(A_j, B_j), \widetilde{G}(B_j)], \end{aligned}$$

where we used the fact that $\Phi_2(A_j)^2$ and $\Phi_2(B_j)^2$ are negligible compared to $\Phi_2(A_j)$ and $\Phi_2(B_j)$ as an approximation, and where $\text{Cov}[\widetilde{G}(A_j), \widetilde{G}(B_j)] = 0$ because drawing alleles in population A is independent

of population B . Moreover, because $\tilde{G}(P_j) = \hat{G}(P_j)/[1 - \Phi_2(P_j)]$, we have

$$\begin{aligned} \text{Var}[\tilde{F}_2(A_j, B_j)] &\approx \text{Var}[\hat{F}_2(A_j, B_j)] - \frac{2\Phi_2(A_j)}{1 - \Phi_2(A_j)} \text{Cov}[\hat{F}_2(A_j, B_j), \hat{G}(A_j)] \\ &\quad - \frac{2\Phi_2(B_j)}{1 - \Phi_2(B_j)} \text{Cov}[\hat{F}_2(A_j, B_j), \hat{G}(B_j)] \\ &\approx 4\Phi_2(A_j)F_2(A_j, B_j)G(A_j) + 4\Phi_2(B_j)F_2(A_j, B_j)G(B_j) \\ &\quad + 4\Phi_2(A_j)\Phi_2(B_j)G(A_j)G(B_j) \\ &\quad - \frac{2\Phi_2(A_j)}{1 - \Phi_2(A_j)} \left[2\Phi_2(A_j)G(A_j)^2 - 2\Phi_2(A_j)G(A_j)G(B_j) \right. \\ &\quad \quad \left. - 2\Phi_2(A_j)F_2(A_j, B_j)G(A_j) \right] \\ &\quad - \frac{2\Phi_2(B_j)}{1 - \Phi_2(B_j)} \left[2\Phi_2(B_j)G(B_j)^2 - 2\Phi_2(B_j)G(A_j)G(B_j) \right. \\ &\quad \quad \left. - 2\Phi_2(B_j)F_2(A_j, B_j)G(B_j) \right]. \end{aligned}$$

Recalling the assumption that $\Phi_2(A_j)^2$, $\Phi_2(B_j)^2$, and $\Phi_2(A_j)\Phi_2(B_j)$ are negligible compared to $\Phi_2(A_j)$ and $\Phi_2(B_j)$, we have

$$\begin{aligned} \text{Var}[\tilde{F}_2(A_j, B_j)] &\approx 4\Phi_2(A_j)F_2(A_j, B_j)G(A_j) + 4\Phi_2(B_j)F_2(A_j, B_j)G(B_j) \\ &\approx \text{Var}[\hat{F}_2(A_j, B_j)], \end{aligned}$$

and it follows that

$$\text{Var}[\tilde{F}_2(A, B)] \approx \text{Var}[\hat{F}_2(A, B)]. \quad \square$$

Proposition 13. Consider J independent polymorphic loci in populations A , B , and C with respective parametric reference allele frequencies $a_j, b_j, c_j \in (0, 1)$, and suppose we take a random sample of $N(P_j)$ individuals at locus j in population $P \in \{A, B, C\}$, some of which may be related or inbred. Moreover, assume that no individual is related to more than one other individual, which makes the terms $\Phi_3(P_j)$, $\Phi_4(P_j)$, $\Phi_{2,2}(P_j)$, $\Phi_2(P_j)^2$, $\Phi_2(A_j)\Phi_2(B_j)$, $\Phi_2(A_j)\Phi_2(C_j)$, and $\Phi_2(B_j)\Phi_2(C_j)$ negligible to $\Phi_2(P_j)$. Based on this simplifying assumption, the estimator $\hat{F}_3(A; B, C)$ has approximate variance

$$\begin{aligned} \text{Var}[\hat{F}_3(A; B, C)] &\approx \frac{4}{J^2} \sum_{j=1}^J \Phi_2(A_j)F_3(A_j; B_j, C_j)G(A_j) + \frac{1}{J^2} \sum_{j=1}^J \Phi_2(A_j)F_2(B_j, C_j)G(A_j) \\ &\quad + \frac{1}{J^2} \sum_{j=1}^J \Phi_2(B_j)F_2(A_j, C_j)G(B_j) + \frac{1}{J^2} \sum_{j=1}^J \Phi_2(C_j)F_2(A_j, B_j)G(C_j). \end{aligned}$$

Proof. From the proof of Proposition 3, we have

$$\mathbb{E}[\hat{F}_3(A_j; B_j, C_j)] = F_3(A_j; B_j, C_j) + \Phi_2(A_j)G(A_j),$$

which gives

$$\begin{aligned} \mathbb{E}[\hat{F}_3(A_j; B_j, C_j)]^2 &= [F_3(A_j; B_j, C_j) + \Phi_2(A_j)G(A_j)]^2 \\ &= F_3(A_j; B_j, C_j)^2 + 2\Phi_2(A_j)F_3(A_j; B_j, C_j)G(A_j) + \Phi_2(A_j)^2G(A_j)^2 \\ &\approx F_3(A_j; B_j, C_j)^2 + 2\Phi_2(A_j)F_3(A_j; B_j, C_j)G(A_j), \end{aligned}$$

where we used the fact that $\Phi_2(A_j)^2$ is negligible compared to $\Phi_2(A_j)$ as an approximation. We also calculate

$$\begin{aligned}
 \mathbb{E}[\widehat{F}_3(A_j; B_j, C_j)^2] &= \mathbb{E}[(\widehat{a}_j - \widehat{b}_j)^2(\widehat{a}_j - \widehat{c}_j)^2] \\
 &= \mathbb{E}[\widehat{a}_j^4 - 2\widehat{a}_j^3\widehat{c}_j + \widehat{a}_j^2\widehat{c}_j^2 - 2\widehat{a}_j^3\widehat{b}_j + 4\widehat{a}_j^2\widehat{b}_j\widehat{c}_j - 2\widehat{a}_j\widehat{b}_j\widehat{c}_j^2 + \widehat{a}_j^2\widehat{b}_j^2 - 2\widehat{a}_j\widehat{b}_j^2\widehat{c}_j + \widehat{b}_j^2\widehat{c}_j^2] \\
 &\approx a_j^4 + 6\Phi_2(A_j)a_j^3(1-a_j) - 2[a_j^3 + 3\Phi_2(A_j)a_j^2(1-a_j)]c_j \\
 &\quad + [a_j^2 + \Phi_2(A_j)a_j(1-a_j)][c_j^2 + \Phi_2(C_j)c_j(1-c_j)] - 2[a_j^3 + 3\Phi_2(A_j)a_j^2(1-a_j)]b_j \\
 &\quad + 4[a_j^2 + \Phi_2(A_j)a_j(1-a_j)]b_jc_j - 2a_jb_j[c_j^2 + \Phi_2(C_j)c_j(1-c_j)] \\
 &\quad + [a_j^2 + \Phi_2(A_j)a_j(1-a_j)][b_j^2 + \Phi_2(B_j)b_j(1-b_j)] - 2a_j[b_j^2 + \Phi_2(B_j)b_j(1-b_j)]c_j \\
 &\quad + [b_j^2 + \Phi_2(B_j)b_j(1-b_j)][c_j^2 + \Phi_2(C_j)c_j(1-c_j)] \\
 &= (a_j - b_j)^2(a_j - c_j)^2 + \Phi_2(A_j)[6(a_j - b_j)(a_j - c_j) + (b_j - c_j)^2]a_j(1-a_j) \\
 &\quad + \Phi_2(B_j)(a_j - c_j)^2b_j(1-b_j) + \Phi_2(C_j)(a_j - b_j)^2c_j(1-c_j) \\
 &\quad + \Phi_2(A_j)\Phi_2(B_j)a_j(1-a_j)b_j(1-b_j) + \Phi_2(A_j)\Phi_2(C_j)a_j(1-a_j)c_j(1-c_j) \\
 &\quad + \Phi_2(B_j)\Phi_2(C_j)b_j(1-b_j)c_j(1-c_j) \\
 &= F_3(A_j; B_j, C_j)^2 + 6\Phi_2(A_j)F_3(A_j; B_j, C_j)G(A_j) + \Phi_2(A_j)F_2(B_j, C_j)G(A_j) \\
 &\quad + \Phi_2(B_j)F_2(A_j, C_j)G(B_j) + \Phi_2(C_j)F_2(A_j, B_j)G(C_j) + \Phi_2(A_j)\Phi_2(B_j)G(A_j)G(B_j) \\
 &\quad + \Phi_2(A_j)\Phi_2(C_j)G(A_j)G(C_j) + \Phi_2(B_j)\Phi_2(C_j)G(B_j)G(C_j) \\
 &\approx F_3(A_j; B_j, C_j)^2 + 6\Phi_2(A_j)F_3(A_j; B_j, C_j)G(A_j) + \Phi_2(A_j)F_2(B_j, C_j)G(A_j) \\
 &\quad + \Phi_2(B_j)F_2(A_j, C_j)G(B_j) + \Phi_2(C_j)F_2(A_j, B_j)G(C_j),
 \end{aligned}$$

where we used the fact that $\Phi_2(A_j)\Phi_2(B_j)$, $\Phi_2(A_j)\Phi_2(C_j)$, and $\Phi_2(B_j)\Phi_2(C_j)$ are negligible compared to $\Phi_2(A_j)$, $\Phi_2(B_j)$, and $\Phi_2(C_j)$ as an approximation. Therefore, we have that

$$\begin{aligned}
 Var[\widehat{F}_3(A_j; B_j, C_j)] &= \mathbb{E}[\widehat{F}_3(A_j; B_j, C_j)^2] - \mathbb{E}[\widehat{F}_3(A_j; B_j, C_j)]^2 \\
 &\approx F_3(A_j; B_j, C_j)^2 + 6\Phi_2(A_j)F_3(A_j; B_j, C_j)G(A_j) + \Phi_2(A_j)F_2(B_j, C_j)G(A_j) \\
 &\quad + \Phi_2(B_j)F_2(A_j, C_j)G(B_j) + \Phi_2(C_j)F_2(A_j, B_j)G(C_j) \\
 &\quad - [F_3(A_j; B_j, C_j)^2 + 2\Phi_2(A_j)F_3(A_j; B_j, C_j)G(A_j)] \\
 &= 4\Phi_2(A_j)F_3(A_j; B_j, C_j)G(A_j) + \Phi_2(A_j)F_2(B_j, C_j)G(A_j) \\
 &\quad + \Phi_2(B_j)F_2(A_j, C_j)G(B_j) + \Phi_2(C_j)F_2(A_j, B_j)G(C_j),
 \end{aligned}$$

which gives

$$\begin{aligned}
 Var[\widehat{F}_3(A; B, C)] &= Var\left[\frac{1}{J} \sum_{j=1}^J \widehat{F}_3(A_j; B_j, C_j)\right] \\
 &= \frac{1}{J^2} \sum_{j=1}^J Var[\widehat{F}_3(A_j; B_j, C_j)] \\
 &\approx \frac{4}{J^2} \sum_{j=1}^J \Phi_2(A_j)F_3(A_j; B_j, C_j)G(A_j) + \frac{1}{J^2} \sum_{j=1}^J \Phi_2(A_j)F_2(B_j, C_j)G(A_j) \\
 &\quad + \frac{1}{J^2} \sum_{j=1}^J \Phi_2(B_j)F_2(A_j, C_j)G(B_j) + \frac{1}{J^2} \sum_{j=1}^J \Phi_2(C_j)F_2(A_j, B_j)G(C_j). \quad \square
 \end{aligned}$$

Lemma 14. Consider J independent polymorphic loci in populations A , B , and C with respective parametric reference allele frequencies $a_j, b_j, c_j \in (0, 1)$, and suppose we take a random sample of $N(P_j)$ individuals at locus j in population $P \in \{A, B, C\}$, some of which may be related or inbred. Moreover, assume that no individual is related to more than one other individual, which makes the terms $\Phi_3(P_j)$, $\Phi_4(P_j)$, $\Phi_{2,2}(P_j)$,

and $\Phi_2(P_j)^2$ negligible to $\Phi_2(P_j)$. Based on this simplifying assumption, the estimators $\widehat{F}_3(A; B, C)$ and $\widehat{g}(A)$ have an approximate covariance

$$\text{Cov}[\widehat{F}_3(A; B, C), \widehat{G}(A)] \approx \frac{1}{J^2} \sum_{j=1}^J \Phi_2(A_j) G(A_j) [2G(A_j) - 2F_3(A_j; B_j, C_j) - b_j(1 - c_j) - (1 - b_j)c_j].$$

Proof. From the proofs of Lemma 1 and Proposition 3, we have

$$\mathbb{E}[\widehat{G}(A_j)] = [1 - \Phi_2(A_j)]G(A_j)$$

and

$$\mathbb{E}[\widehat{F}_3(A_j; B_j, C_j)] = F_3(A_j; B_j, C_j) + \Phi_2(A_j)G(A_j),$$

yielding

$$\begin{aligned} \mathbb{E}[\widehat{F}_3(A_j; B_j, C_j)]\mathbb{E}[\widehat{G}(A_j)] &= [F_3(A_j; B_j, C_j) + \Phi_2(A_j)G(A_j)][1 - \Phi_2(A_j)]G(A_j) \\ &= F_3(A_j; B_j, C_j)G(A_j) + \Phi_2(A_j)G(A_j)^2 - \Phi_2(A_j)F_3(A_j; B_j, C_j)G(A_j) \\ &\quad - \Phi_2(A_j)^2G(A_j)^2 \\ &\approx F_3(A_j; B_j, C_j)G(A_j) + \Phi_2(A_j)G(A_j)^2 - \Phi_2(A_j)F_3(A_j; B_j, C_j)G(A_j), \end{aligned}$$

where we used the fact that $\Phi_2(A_j)^2$ is negligible compared to $\Phi_2(A_j)$ as an approximation. We also calculate

$$\begin{aligned} \mathbb{E}[\widehat{F}_3(A_j; B_j, C_j)\widehat{g}(A_j)] &= \mathbb{E}[(\widehat{a}_j - \widehat{b}_j)(\widehat{a}_j - \widehat{c}_j)\widehat{a}_j(1 - \widehat{a}_j)] \\ &= \mathbb{E}[(\widehat{a}_j^2 - \widehat{a}_j\widehat{b}_j - \widehat{a}_j\widehat{c}_j + \widehat{b}_j\widehat{c}_j)(\widehat{a}_j - \widehat{a}_j^2)] \\ &= \mathbb{E}[\widehat{a}_j^3] - \mathbb{E}[\widehat{a}_j^2]\mathbb{E}[\widehat{b}_j] - \mathbb{E}[\widehat{a}_j^2]\mathbb{E}[\widehat{c}_j] + \mathbb{E}[\widehat{a}_j]\mathbb{E}[\widehat{b}_j]\mathbb{E}[\widehat{c}_j] - \mathbb{E}[\widehat{a}_j^4] + \mathbb{E}[\widehat{a}_j^3]\mathbb{E}[\widehat{b}_j] \\ &\quad + \mathbb{E}[\widehat{a}_j^3]\mathbb{E}[\widehat{c}_j] - \mathbb{E}[\widehat{a}_j^2]\mathbb{E}[\widehat{b}_j]\mathbb{E}[\widehat{c}_j] \\ &\approx a_j^3 + 3\Phi_2(A_j)a_j^2(1 - a_j) - [a_j^2 + \Phi_2(A_j)a_j(1 - a_j)]b_j \\ &\quad - [a_j^2 + \Phi_2(A_j)a_j(1 - a_j)]c_j + a_jb_jc_j - [a_j^4 + 6\Phi_2(A_j)a_j^3(1 - a_j)] \\ &\quad + [a_j^3 + 3\Phi_2(A_j)a_j^2(1 - a_j)]b_j + [a_j^3 + 3\Phi_2(A_j)a_j^2(1 - a_j)]c_j \\ &\quad - [a_j^2 + \Phi_2(A_j)a_j(1 - a_j)]b_jc_j. \end{aligned}$$

Recognizing that $G(A_j) = a_j(1 - a_j)$ and $F_3(A_j; B_j, C_j) = a_j^2 - a_jb_j - a_jc_j + b_jc_j$, we have

$$\begin{aligned} \mathbb{E}[\widehat{F}_3(A_j; B_j, C_j)\widehat{g}(A_j)] &\approx a_j^3 - a_j^2b_j - a_j^2c_j + a_jb_jc_j - a_j^4 + a_j^3b_j + a_j^3c_j - a_j^2b_jc_j \\ &\quad + \Phi_2(A_j)[3a_j - b_j - c_j - 6a_j^2 + 3a_jb_j + 3a_jc_j - b_jc_j]G(A_j) \\ &= (a_j^2 - a_jb_j - a_jc_j + b_jc_j)(a_j - a_j^2) \\ &\quad + \Phi_2(A_j)[3(a_j - a_j^2) - 3(a_j^2 - a_jb_j - a_jc_j + b_jc_j) - b_j(1 - c_j) \\ &\quad - (1 - b_j)c_j]G(A_j) \\ &= F_3(A_j; B_j, C_j)G(A_j) \\ &\quad + \Phi_2(A_j)[3G(A_j) - 3F_3(A_j; B_j, C_j) - b_j(1 - c_j) - (1 - b_j)c_j]G(A_j) \\ &= F_3(A_j; B_j, C_j)G(A_j) + 3\Phi_2(A_j)G(A_j)^2 - 3\Phi_2(A_j)F_3(A_j; B_j, C_j)G(A_j) \\ &\quad - \Phi_2(A_j)G(A_j)b_j(1 - c_j) - \Phi_2(A_j)G(A_j)(1 - b_j)c_j \\ &= [1 - 3\Phi_2(A_j)]F_3(A_j; B_j, C_j)G(A_j) + 3\Phi_2(A_j)G(A_j)^2 \\ &\quad - \Phi_2(A_j)G(A_j)b_j(1 - c_j) - \Phi_2(A_j)G(A_j)(1 - b_j)c_j. \end{aligned}$$

Therefore, we have that

$$\begin{aligned}
 \text{Cov}[\widehat{F}_3(A_j; B_j, C_j), \widehat{G}(A_j)] &= \mathbb{E}[\widehat{F}_3(A_j; B_j, C_j)\widehat{G}(A_j)] - \mathbb{E}[\widehat{F}_3(A_j; B_j, C_j)]\mathbb{E}[\widehat{G}(A_j)] \\
 &\approx [1 - 3\Phi_2(A_j)]F_3(A_j; B_j, C_j)G(A_j) + 3\Phi_2(A_j)G(A_j)^2 \\
 &\quad - \Phi_2(A_j)G(A_j)b_j(1 - c_j) - \Phi_2(A_j)G(A_j)(1 - b_j)c_j \\
 &\quad - [F_3(A_j; B_j, C_j)G(A_j) + \Phi_2(A_j)G(A_j)^2 \\
 &\quad - \Phi_2(A_j)F_3(A_j; B_j, C_j)G(A_j)] \\
 &= 2\Phi_2(A_j)G(A_j)^2 - 2\Phi_2(A_j)F_3(A_j; B_j, C_j)G(A_j) \\
 &\quad - \Phi_2(A_j)G(A_j)b_j(1 - c_j) - \Phi_2(A_j)G(A_j)(1 - b_j)c_j \\
 &= \Phi_2(A_j)G(A_j)[2G(A_j) - 2F_3(A_j; B_j, C_j) - b_j(1 - c_j) - (1 - b_j)c_j],
 \end{aligned}$$

which gives

$$\begin{aligned}
 \text{Cov}[\widehat{F}_3(A; B, C), \widehat{G}(A)] &= \text{Cov}\left[\frac{1}{J} \sum_{j=1}^J \widehat{F}_3(A_j; B_j, C_j), \frac{1}{J} \sum_{j=1}^J \widehat{G}(A_j)\right] \\
 &= \frac{1}{J^2} \sum_{j=1}^J \text{Cov}[\widehat{F}_3(A_j; B_j, C_j), \widehat{G}(A_j)] \\
 &\approx \frac{1}{J^2} \sum_{j=1}^J \Phi_2(A_j)G(A_j)[2G(A_j) - 2F_3(A_j; B_j, C_j) - b_j(1 - c_j) \\
 &\quad - (1 - b_j)c_j]. \quad \square
 \end{aligned}$$

Proposition 15. Consider J independent polymorphic loci in populations A , B , and C with respective parametric reference allele frequencies $a_j, b_j, c_j \in (0, 1)$, and suppose we take a random sample of $N(P_j)$ individuals at locus j in population $P \in \{A, B, C\}$, some of which may be related or inbred. Moreover, assume that no individual is related to more than one other individual, which makes the terms $\Phi_3(P_j)$, $\Phi_4(P_j)$, $\Phi_{2,2}(P_j)$, $\Phi_2(P_j)^2$, $\Phi_2(A_j)\Phi_2(B_j)$, $\Phi_2(A_j)\Phi_2(C_j)$, and $\Phi_2(B_j)\Phi_2(C_j)$ negligible to $\Phi_2(P_j)$. Based on this simplifying assumption, the unbiased estimator $\widehat{F}_3(A; B, C)$ has approximate variance

$$\begin{aligned}
 \text{Var}[\widehat{F}_3(A; B, C)] &\approx \frac{4}{J^2} \sum_{j=1}^J \Phi_2(A_j)F_3(A_j; B_j, C_j)G(A_j) + \frac{1}{J^2} \sum_{j=1}^J \Phi_2(A_j)F_2(B_j, C_j)G(A_j) \\
 &\quad + \frac{1}{J^2} \sum_{j=1}^J \Phi_2(B_j)F_2(A_j, C_j)G(B_j) + \frac{1}{J^2} \sum_{j=1}^J \Phi_2(C_j)F_2(A_j, B_j)G(C_j).
 \end{aligned}$$

Proof. Recall that

$$\widehat{F}_3(A_j; B_j, C_j) = \widehat{F}_3(A_j; B_j, C_j) - \Phi_2(A_j)\widehat{G}(A_j),$$

where $\widehat{F}_3(A_j; B_j, C_j)$ is an unbiased estimator for $F_3(A_j; B_j, C_j)$ and $\widehat{G}(A_j) = \widehat{G}(A_j)/[1 - \Phi_2(A_j)]$ is an unbiased estimator of $G(A_j)$ at locus $j \in \{1, 2, \dots, J\}$. Also, from the proof of Proposition 3, we have

$$\mathbb{E}[\widehat{F}_3(A_j; B_j, C_j)] = F_3(A_j; B_j, C_j) + \Phi_2(A_j)G(A_j).$$

Therefore, we have that

$$\begin{aligned}
 \text{Var}[\widehat{F}_3(A_j; B_j, C_j)] &= \text{Var}[\widehat{F}_3(A_j; B_j, C_j) - \Phi_2(A_j)\widehat{G}(A_j)] \\
 &= \text{Var}[\widehat{F}_3(A_j; B_j, C_j)] + \Phi_2(A_j)^2 \text{Var}[\widehat{G}(A_j)] \\
 &\quad - 2\Phi_2(A_j)\text{Cov}[\widehat{F}_3(A_j; B_j, C_j), \widehat{G}(A_j)] \\
 &\approx \text{Var}[\widehat{F}_3(A_j; B_j, C_j)] - \frac{2\Phi_2(A_j)}{1 - \Phi_2(A_j)} \text{Cov}[\widehat{F}_3(A_j; B_j, C_j), \widehat{G}(A_j)],
 \end{aligned}$$

where we used the fact that $\Phi_2(A_j)^2$ is negligible compared to $\Phi_2(A_j)$ as an approximation. Moreover, because $\tilde{G}(A_j) = \hat{G}(A_j)/[1 - \Phi_2(A_j)]$, we have

$$\begin{aligned} \text{Var}[\tilde{F}_3(A_j; B_j; C_j)] &\approx \text{Var}[\hat{F}_3(A_j; B_j; C_j)] - \frac{2\Phi_2(A_j)}{1 - \Phi_2(A_j)} \text{Cov}[\hat{F}_3(A_j; B_j; C_j), \hat{g}(A_j)] \\ &= \text{Var}[\hat{F}_3(A_j; B_j; C_j)] \\ &\quad - \frac{2\Phi_2(A_j)^2}{1 - \Phi_2(A_j)} G(A_j)[2G(A_j) - 2F_3(A_j; B_j; C_j) - b_j(1 - c_j) - (1 - b_j)c_j]. \end{aligned}$$

Recalling the assumption that $\Phi_2(A_j)^2$ is negligible compared to $\Phi_2(A_j)$, we have

$$\text{Var}[\tilde{F}_3(A_j; B_j; C_j)] \approx \text{Var}[\hat{F}_3(A_j; B_j; C_j)]. \quad \square$$

Proposition 16. Consider J independent polymorphic loci in populations A, B, C , and D with respective parametric reference allele frequencies $a_j, b_j, c_j, d_j \in (0, 1)$, and suppose we take a random sample of $N(P_j)$ individuals at locus j in population $P \in \{A, B, C, D\}$, some of which may be related or inbred. Moreover, assume that no individual is related to more than one other individual, which makes the terms $\Phi_3(P_j)$, $\Phi_4(P_j)$, $\Phi_{2,2}(P_j)$, $\Phi_2(P_j)^2$, $\Phi_2(A_j)\Phi_2(C_j)$, $\Phi_2(A_j)\Phi_2(D_j)$, $\Phi_2(B_j)\Phi_2(C_j)$, and $\Phi_2(B_j)\Phi_2(D_j)$ negligible to $\Phi_2(P_j)$. Based on this simplifying The unbiased estimator $\hat{F}_4(A, B; C, D)$ has approximate variance

$$\begin{aligned} \text{Var}[\hat{F}_4(A, B; C, D)] &\approx \frac{1}{J^2} \sum_{j=1}^J [\Phi_2(C_j)G(C_j) + \Phi_2(D_j)G(D_j)]F_2(A_j, B_j) \\ &\quad + \frac{1}{J^2} \sum_{j=1}^J [\Phi_2(A_j)G(A_j) + \Phi_2(B_j)G(B_j)]F_2(C_j, D_j). \end{aligned}$$

Proof. From the proofs of Propositions 2 and 5, we have

$$\mathbb{E}[\hat{F}_2(A_j, B_j)] = F_2(A_j, B_j) + \Phi_2(A_j)G(A_j) + \Phi_2(B_j)G(B_j)$$

and

$$\mathbb{E}[\hat{F}_4(A_j, B_j; C_j, D_j)] = F_4(A_j, B_j; C_j, D_j).$$

We calculate

$$\begin{aligned} \mathbb{E}[\hat{F}_4(A_j, B_j; C_j, D_j)^2] &= \mathbb{E}[(\hat{a}_j - \hat{b}_j)^2(\hat{c}_j - \hat{d}_j)^2] \\ &= \mathbb{E}[\hat{F}_2(A_j, B_j)\hat{F}_2(C_j, D_j)] \\ &= \mathbb{E}[\hat{F}_2(A_j, B_j)]\mathbb{E}[\hat{F}_2(C_j, D_j)] \\ &= [F_2(A_j, B_j) + \Phi_2(A_j)G(A_j) + \Phi_2(B_j)G(B_j)] \\ &\quad \times [F_2(C_j, D_j) + \Phi_2(C_j)G(C_j) + \Phi_2(D_j)G(D_j)] \\ &= F_4(A_j, B_j; C_j, D_j)^2 + F_2(A_j, B_j)[\Phi_2(C_j)G(C_j) + \Phi_2(D_j)G(D_j)] \\ &\quad + F_2(C_j, D_j)[\Phi_2(A_j)G(A_j) + \Phi_2(B_j)G(B_j)] \\ &\quad + [\Phi_2(A_j)G(A_j) + \Phi_2(B_j)G(B_j)][\Phi_2(C_j)G(C_j) + \Phi_2(D_j)G(D_j)], \end{aligned}$$

where we use the identity that

$$\begin{aligned} F_4(A_j, B_j; C_j, D_j)^2 &= (a_j - b_j)^2(c_j - d_j)^2 \\ &= F_2(A_j, B_j)F_2(C_j, D_j). \end{aligned}$$

Therefore, we have that

$$\begin{aligned}
 \text{Var}[\widehat{F}_4(A_j, B_j; C_j, D_j)] &= \mathbb{E}[\widehat{F}_4(A_j, B_j; C_j, D_j)^2] - \mathbb{E}[\widehat{F}_4(A_j, B_j; C_j, D_j)]^2 \\
 &= F_4(A_j, B_j; C_j, D_j)^2 + F_2(A_j, B_j)[\Phi_2(C_j)G(C_j) + \Phi_2(D_j)G(D_j)] \\
 &\quad + F_2(C_j, D_j)[\Phi_2(A_j)G(A_j) + \Phi_2(B_j)G(B_j)] \\
 &\quad + [\Phi_2(A_j)G(A_j) + \Phi_2(B_j)G(B_j)][\Phi_2(C_j)G(C_j) + \Phi_2(D_j)G(D_j)] \\
 &\quad - F_4(A_j, B_j; C_j, D_j)^2 \\
 &= F_2(A_j, B_j)[\Phi_2(C_j)G(C_j) + \Phi_2(D_j)G(D_j)] \\
 &\quad + F_2(C_j, D_j)[\Phi_2(A_j)G(A_j) + \Phi_2(B_j)G(B_j)] \\
 &\quad + [\Phi_2(A_j)G(A_j) + \Phi_2(B_j)G(B_j)][\Phi_2(C_j)G(C_j) + \Phi_2(D_j)G(D_j)] \\
 &\approx F_2(A_j, B_j)[\Phi_2(C_j)G(C_j) + \Phi_2(D_j)G(D_j)] \\
 &\quad + F_2(C_j, D_j)[\Phi_2(A_j)G(A_j) + \Phi_2(B_j)G(B_j)],
 \end{aligned}$$

where we used the fact that $\Phi_2(A_j)\Phi_2(C_j)$, $\Phi_2(A_j)\Phi_2(D_j)$, $\Phi_2(B_j)\Phi_2(C_j)$, and $\Phi_2(C_j)\Phi_2(D_j)$ are negligible compared to $\Phi_2(A_j)$, $\Phi_2(B_j)$, $\Phi_2(C_j)$ and $\Phi_2(D_j)$ as an approximation. It follows that

$$\begin{aligned}
 \text{Var}[\widehat{F}_4(A, B; C, D)] &= \text{Var}\left[\frac{1}{J} \sum_{j=1}^J \widehat{F}_4(A_j, B_j; C_j, D_j)\right] \\
 &= \frac{1}{J^2} \sum_{j=1}^J \text{Var}[\widehat{F}_4(A_j, B_j; C_j, D_j)] \\
 &\approx \frac{1}{J^2} \sum_{j=1}^J [\Phi_2(C_j)G(C_j) + \Phi_2(D_j)G(D_j)] F_2(A_j, B_j) \\
 &\quad + \frac{1}{J^2} \sum_{j=1}^J [\Phi_2(A_j)G(A_j) + \Phi_2(B_j)G(B_j)] F_2(C_j, D_j). \quad \square
 \end{aligned}$$

Following Wolter (2007), we have that an approximation to the variance of the ratio estimator X/Y is

$$\text{Var}\left[\frac{X}{Y}\right] \approx \frac{\mathbb{E}[X]^2}{\mathbb{E}[Y]^2} \left[\frac{\text{Var}[X]}{\mathbb{E}[X]^2} + \frac{\text{Var}[Y]}{\mathbb{E}[Y]^2} - 2 \frac{\text{Cov}[X, Y]}{\mathbb{E}[X]\mathbb{E}[Y]} \right]$$

Proposition 17. Consider J polymorphic loci in populations A , B , and C with respective parametric reference allele frequencies $a_j, b_j, c_j \in (0, 1)$, and suppose we take a random sample of $N(P_j)$ individuals at locus j in population $P \in \{A, B, C\}$, some of which may be related or inbred. Moreover, assume that no individual is related to more than one other individual, which makes the terms $\Phi_3(P_j)$, $\Phi_4(P_j)$, $\Phi_{2,2}(P_j)$, $\Phi_2(P_j)^2$, $\Phi_2(A_j)\Phi_2(B_j)$, $\Phi_2(A_j)\Phi_2(C_j)$, and $\Phi_2(B_j)\Phi_2(C_j)$ negligible to $\Phi_2(P_j)$. Based on this simplifying assumption, the ratio estimator $\widehat{F}_3(A; B, C | A)$ has approximate variance

$$\text{Var}[\widehat{F}_3(A; B, C | A)] \approx \frac{\mathbb{E}[\widehat{F}_3(A; B, C)]^2}{4\mathbb{E}[\widehat{G}(A)]^2} \left[\frac{\text{Var}[\widehat{F}_3(A; B, C)]}{\mathbb{E}[\widehat{F}_3(A; B, C)]^2} + \frac{\text{Var}[\widehat{G}(A)]}{\mathbb{E}[\widehat{G}(A)]^2} - 2 \frac{\text{Cov}[\widehat{F}_3(A; B, C), \widehat{G}(A)]}{\mathbb{E}[\widehat{F}_3(A; B, C)]\mathbb{E}[\widehat{G}(A)]} \right],$$

where the expectations are

$$\mathbb{E}[\widehat{F}_3(A; B, C)] = F_3(A; B, C) + \frac{1}{J} \sum_{j=1}^J \Phi_2(A_j)G(A_j)$$

$$\mathbb{E}[\widehat{G}(A)] = G(A) - \frac{1}{J} \sum_{j=1}^J \Phi_2(A_j)G(A_j)$$

the variances are

$$\begin{aligned} \text{Var}[\widehat{F}_3(A; B, C)] &\approx \frac{4}{J^2} \sum_{j=1}^J \Phi_2(A_j) F_3(A_j; B_j, C_j) G(A_j) + \frac{1}{J^2} \sum_{j=1}^J \Phi_2(A_j) F_2(B_j, C_j) G(A_j) \\ &\quad + \frac{1}{J^2} \sum_{j=1}^J \Phi_2(B_j) F_2(A_j, C_j) G(B_j) + \frac{1}{J^2} \sum_{j=1}^J \Phi_2(C_j) F_2(A_j, B_j) G(C_j) \\ \text{Var}[\widehat{G}(A)] &\approx \frac{1}{J^2} \sum_{j=1}^J \Phi_2(A_j) G(A_j) - \frac{4}{J^2} \sum_{j=1}^J \Phi_2(A_j) G(A_j)^2 \end{aligned}$$

and the covariance is

$$\text{Cov}[\widehat{F}_3(A; B, C), \widehat{G}(A)] \approx \frac{1}{J^2} \sum_{j=1}^J \Phi_2(A_j) G(A_j) [2G(A_j) - 2F_3(A_j; B_j, C_j) - b_j(1 - c_j) - (1 - b_j)c_j].$$

Proof. Recall that

$$\widehat{F}_3(A; B, C, | A) = \frac{\widehat{F}_3(A; B, C)}{2\widehat{G}(A)}.$$

Assuming that $X = \widehat{F}_3(A; B, C)$ and $Y = 2\widehat{G}(A)$, following the approximation in Wolter (2007) we have

$$\text{Var}[\widehat{F}_3(A; B, C | A)] \approx \frac{\mathbb{E}[\widehat{F}_3(A; B, C)]^2}{4\mathbb{E}[\widehat{G}(A)]^2} \left[\frac{\text{Var}[\widehat{F}_3(A; B, C)]}{\mathbb{E}[\widehat{F}_3(A; B, C)]^2} + \frac{\text{Var}[\widehat{G}(A)]}{\mathbb{E}[\widehat{G}(A)]^2} - 2 \frac{\text{Cov}[\widehat{F}_3(A; B, C), \widehat{G}(A)]}{\mathbb{E}[\widehat{F}_3(A; B, C)]\mathbb{E}[\widehat{G}(A)]} \right],$$

where $\mathbb{E}[\widehat{F}_3(A; B, C)]$ is given in Proposition 3, $\mathbb{E}[\widehat{G}(A)]$ in Lemma 1, $\text{Var}[\widehat{F}_3(A; B, C)]$ in Proposition 13, $\text{Var}[\widehat{G}(A)]$ in Lemma 9, and $\text{Cov}[\widehat{F}_3(A; B, C), \widehat{G}(A)]$ in Lemma 14. \square

Lemma 18. Consider J independent polymorphic loci in populations A , B , and C with respective parametric reference allele frequencies $a_j, b_j, c_j \in (0, 1)$, and suppose we take a random sample of $N(P_j)$ individuals at locus j in population $P \in \{A, B, C\}$, some of which may be related or inbred. Moreover, assume that no individual is related to more than one other individual, which makes the terms $\Phi_3(P_j)$, $\Phi_4(P_j)$, $\Phi_{2,2}(P_j)$, and $\Phi_2(P_j)^2$ negligible to $\Phi_2(P_j)$. Based on this simplifying assumption, the unbiased estimators $\widehat{F}_3(A; B, C)$ and $\widehat{G}(A)$ have an approximate covariance

$$\text{Cov}[\widetilde{F}_3(A; B, C), \widetilde{G}(A)] \approx \frac{1}{J^2} \sum_{j=1}^J \frac{\Phi_2(A_j)}{1 - \Phi_2(A_j)} G(A_j) [2G(A_j) - 2F_3(A_j; B_j, C_j) - b_j(1 - c_j) - (1 - b_j)c_j].$$

Proof. Recall that

$$\widetilde{F}_3(A_j; B_j, C_j) = \widehat{F}_3(A_j; B_j, C_j) - \Phi_2(A_j)\widetilde{G}(A_j),$$

where $\widetilde{G}(A_j) = \widehat{G}(A_j)/[1 - \Phi_2(A_j)]$. It follows that

$$\begin{aligned} \text{Cov}[\widetilde{F}_3(A_j; B_j, C_j), \widetilde{G}(A_j)] &= \text{Cov}[\widehat{F}_3(A_j; B_j, C_j) - \Phi_2(A_j)\widetilde{G}(A_j), \widetilde{G}(A_j)] \\ &= \text{Cov}[\widehat{F}_3(A_j; B_j, C_j), \widetilde{G}(A_j)] - \Phi_2(A_j)\text{Var}[\widetilde{G}(A_j)] \\ &= \frac{1}{1 - \Phi_2(A_j)} \text{Cov}[\widehat{F}_3(A_j; B_j, C_j), \widehat{G}(A_j)] - \frac{\Phi_2(A_j)}{[1 - \Phi_2(A_j)]^2} \text{Var}[\widehat{G}(A_j)] \\ &\approx \frac{1}{1 - \Phi_2(A_j)} \text{Cov}[\widehat{F}_3(A_j; B_j, C_j), \widehat{G}(A_j)] - \frac{\Phi_2(A_j)}{1 - 2\Phi_2(A_j)} \text{Var}[\widehat{G}(A_j)], \end{aligned}$$

where we used the fact that $\Phi_2(A_j)^2$ is negligible compared to $\Phi_2(A_j)$ as an approximation. From the proofs of Lemmas 9 and 14, we have

$$\text{Var}[\widehat{G}(A_j)] \approx \Phi_2(A_j)G(A_j) - 4\Phi_2(A_j)G(A_j)^2$$

and

$$\text{Cov}[\hat{F}_3(A_j; B_j, C_j), \hat{G}(A_j)] = \Phi_2(A_j)G(A_j)[2G(A_j) - 2F_3(A_j; B_j, C_j) - b_j(1 - c_j) - (1 - b_j)c_j].$$

Assuming that $\Phi_2(A_j)^2$ is negligible compared to $\Phi_2(A_j)$, we have that

$$\Phi_2(A_j)\text{Var}[\hat{G}(A_j)] \approx 0.$$

We therefore have that

$$\text{Cov}[\tilde{F}_3(A_j; B_j, C_j), \tilde{G}(A_j)] \approx \frac{1}{1 - \Phi_2(A_j)} \text{Cov}[\hat{F}_3(A_j; B_j, C_j), \hat{G}(A_j)],$$

and thus by independence of loci we have

$$\begin{aligned} \text{Cov}[\tilde{F}_3(A; B, C), \tilde{G}(A)] &= \text{Cov}\left[\frac{1}{J} \sum_{j=1}^J \tilde{F}_3(A_j; B_j, C_j), \frac{1}{J} \sum_{j=1}^J \tilde{G}(A_j)\right] \\ &= \frac{1}{J^2} \sum_{j=1}^J \text{Cov}[\tilde{F}_3(A_j; B_j, C_j), \tilde{G}(A_j)] \\ &\approx \frac{1}{J^2} \sum_{j=1}^J \frac{\Phi_2(A_j)}{1 - \Phi_2(A_j)} G(A_j)[2G(A_j) - 2F_3(A_j; B_j, C_j) \\ &\quad - b_j(1 - c_j) - (1 - b_j)c_j]. \quad \square \end{aligned}$$

Proposition 19. Consider J polymorphic loci in populations A , B , and C with respective parametric reference allele frequencies $a_j, b_j, c_j \in (0, 1)$, and suppose we take a random sample of $N(P_j)$ individuals at locus j in population $P \in \{A, B, C\}$, some of which may be related or inbred. Moreover, assume that no individual is related to more than one other individual, which makes the terms $\Phi_3(P_j)$, $\Phi_4(P_j)$, $\Phi_{2,2}(P_j)$, $\Phi_2(P_j)^2$, $\Phi_2(A_j)\Phi_2(B_j)$, $\Phi_2(A_j)\Phi_2(C_j)$, and $\Phi_2(B_j)\Phi_2(C_j)$ negligible to $\Phi_2(P_j)$. Based on this simplifying assumption, the approximately unbiased ratio estimator $\tilde{F}_3(A; B, C | A)$ has approximate variance

$$\text{Var}[\tilde{F}_3(A; B, C | A)] \approx \frac{F_3(A; B, C)^2}{4G(A)^2} \left[\frac{\text{Var}[\tilde{F}_3(A; B, C)]}{F_3(A; B, C)^2} + \frac{\text{Var}[\tilde{G}(A)]}{G(A)^2} - 2 \frac{\text{Cov}[\tilde{F}_3(A; B, C), \tilde{G}(A)]}{F_3(A; B, C)G(A)} \right],$$

where the variances are

$$\begin{aligned} \text{Var}[\tilde{F}_3(A; B, C)] &\approx \frac{4}{J^2} \sum_{j=1}^J \Phi_2(A_j)F_3(A_j; B_j, C_j)G(A_j) + \frac{1}{J^2} \sum_{j=1}^J \Phi_2(A_j)F_2(B_j, C_j)G(A_j) \\ &\quad + \frac{1}{J^2} \sum_{j=1}^J \Phi_2(B_j)F_2(A_j, C_j)G(B_j) + \frac{1}{J^2} \sum_{j=1}^J \Phi_2(C_j)F_2(A_j, B_j)G(C_j) \\ \text{Var}[\tilde{G}(A)] &\approx \frac{1}{J^2} \sum_{j=1}^J \frac{\Phi_2(A_j)}{1 - 2\Phi_2(A_j)} G(A_j) - \frac{4}{J^2} \sum_{j=1}^J \frac{\Phi_2(A_j)}{1 - 2\Phi_2(A_j)} G(A_j)^2 \end{aligned}$$

and the covariance is

$$\text{Cov}[\tilde{F}_3(A; B, C), \tilde{G}(A)] \approx \frac{1}{J^2} \sum_{j=1}^J \frac{\Phi_2(A_j)}{1 - \Phi_2(A_j)} G(A_j)[2G(A_j) - 2F_3(A_j; B_j, C_j) - b_j(1 - c_j) - (1 - b_j)c_j].$$

Proof. Recall that

$$\tilde{F}_3(A; B, C, | A) = \frac{\tilde{F}_3(A; B, C)}{2\tilde{G}(A)},$$

where $\tilde{F}_3(A; B, C)$ is an unbiased estimator for $F_3(A; B, C)$ and $\tilde{G}(A)$ is an unbiased estimator of $G(A)$. Assuming that $X = \tilde{F}_3(A; B, C)$ and $Y = 2\tilde{G}(A)$, following the approximation in Wolter (2007) we have

$$Var[\tilde{F}_3(A; B, C | A)] \approx \frac{F_3(A; B, C)^2}{4G(A)^2} \left[\frac{Var[\tilde{F}_3(A; B, C)]}{F_3(A; B, C)^2} + \frac{Var[\tilde{G}(A)]}{G(A)^2} - 2 \frac{Cov[\tilde{F}_3(A; B, C), \tilde{G}(A)]}{F_3(A; B, C)G(A)} \right],$$

where $Var[\tilde{F}_3(A; B, C)]$ is given in Proposition 15, $Var[\tilde{G}(A)]$ in Lemma 9, and $Cov[\tilde{F}_3(A; B, C), \tilde{G}(A)]$ in Lemma 18. \square

Lemma 20. Consider J independent polymorphic loci in populations A, B, C , and D with respective parametric reference allele frequencies $a_j, b_j, c_j, d_j \in (0, 1)$, and suppose we take a random sample of $N(P_j)$ individuals at locus j in population $P \in \{A, B, C, D\}$, some of which may be related or inbred. Moreover, assume that no individual is related to more than one other individual, which makes the terms $\Phi_3(P_j)$, $\Phi_4(P_j)$, $\Phi_{2,2}(P_j)$, and $\Phi_2(P_j)^2$ negligible to $\Phi_2(P_j)$. Based on this simplifying assumption, the estimators $\hat{F}_4(A, B; C, D)$ and $\hat{G}(P)$, $P \in \{A, B, C, D\}$, have approximate covariances

$$\begin{aligned} Cov[\hat{F}_4(A, B; C, D), \hat{G}(A)] &\approx \frac{1}{J^2} \sum_{j=1}^J \Phi_2(A_j) G(A_j) (1 - 2a_j)(c_j - d_j) \\ Cov[\hat{F}_4(A, B; C, D), \hat{G}(B)] &\approx -\frac{1}{J^2} \sum_{j=1}^J \Phi_2(B_j) G(B_j) (1 - 2b_j)(c_j - d_j) \\ Cov[\hat{F}_4(A, B; C, D), \hat{G}(C)] &\approx \frac{1}{J^2} \sum_{j=1}^J \Phi_2(C_j) G(C_j) (1 - 2c_j)(a_j - b_j) \\ Cov[\hat{F}_4(A, B; C, D), \hat{G}(D)] &\approx -\frac{1}{J^2} \sum_{j=1}^J \Phi_2(D_j) G(D_j) (1 - 2d_j)(a_j - b_j). \end{aligned}$$

Proof. From the proofs of Lemma 1 and Proposition 5, we that have

$$\mathbb{E}[\hat{G}(P_j)] = [1 - \Phi_2(P_j)]G(P_j)$$

and

$$\mathbb{E}[\hat{F}_4(A_j, B_j; C_j, D_j)] = F_4(A_j, B_j; C_j, D_j),$$

yielding

$$\mathbb{E}[\hat{F}_4(A_j, B_j; C_j, D_j)]\mathbb{E}[\hat{G}(P_j)] = [1 - \Phi_2(P_j)]F_4(A_j, B_j; C_j, D_j)G(P_j).$$

We first calculate

$$\begin{aligned} \mathbb{E}[\hat{F}_4(A_j, B_j; C_j, D_j)\hat{G}(A_j)] &= \mathbb{E}[(\hat{a}_j - \hat{b}_j)(\hat{c}_j - \hat{d}_j)\hat{a}_j(1 - \hat{a}_j)] \\ &= \mathbb{E}[(\hat{a}_j - \hat{b}_j)(\hat{a}_j - \hat{a}_j^2)]\mathbb{E}[\hat{c}_j - \hat{d}_j] \\ &= \left[\mathbb{E}[\hat{a}_j^2] - \mathbb{E}[\hat{a}_j^3] - \mathbb{E}[\hat{a}_j]\mathbb{E}[\hat{b}_j] + \mathbb{E}[\hat{a}_j^2]\mathbb{E}[\hat{b}_j] \right] \left[\mathbb{E}[\hat{c}_j] - \mathbb{E}[\hat{d}_j] \right] \\ &\approx \left[a_j^2 + \Phi_2(A_j)a_j(1 - a_j) - [a_j^3 + 3\Phi_2(A_j)a_j^3(1 - a_j) - a_jb_j \right. \\ &\quad \left. + [a_j^2 + \Phi_2(A_j)a_j(1 - a_j)]b_j \right] (c_j - d_j) \\ &= (a_j^2 - a_j^3 - a_jb_j + a_j^2b_j)(c_j - d_j) + \Phi_2(A_j)a_j(1 - a_j)(1 - 3a_j + b_j)(c_j - d_j) \\ &= (a_j - b_j)(c_j - d_j)a_j(1 - a_j) + \Phi_2(A_j)a_j(1 - a_j)[1 - 2a_j - (a_j - b_j)](c_j - d_j) \\ &= F_4(A_j, B_j; C_j, D_j)G(A_j) - \Phi_2(A_j)F_4(A_j, B_j; C_j, D_j)G(A_j) \\ &\quad + \Phi_2(A_j)G(A_j)(1 - 2a_j)(c_j - d_j) \\ &= [1 - \Phi_2(A_j)]F_4(A_j, B_j; C_j, D_j)G(A_j) + \Phi_2(A_j)G(A_j)(1 - 2a_j)(c_j - d_j). \end{aligned}$$

Hence, we have that

$$\begin{aligned} Cov[\widehat{F}_4(A_j, B_j; C_j, D_j), \widehat{G}(A_j)] &= \mathbb{E}[\widehat{F}_4(A_j, B_j; C_j, D_j)\widehat{G}(A_j)] - \mathbb{E}[\widehat{F}_4(A_j, B_j; C_j, D_j)]\mathbb{E}[\widehat{G}(A_j)] \\ &\approx [1 - \Phi_2(A_j)]F_4(A_j, B_j; C_j, D_j)G(A_j) + \Phi_2(A_j)G(A_j)(1 - 2a_j)(c_j - d_j) \\ &\quad - [1 - \Phi_2(A_j)]F_4(A_j, B_j; C_j, D_j)G(A_j) \\ &= \Phi_2(A_j)G(A_j)(1 - 2a_j)(c_j - d_j). \end{aligned}$$

Similarly, we have that

$$\begin{aligned} \mathbb{E}[\widehat{F}_4(A_j, B_j; C_j, D_j)\widehat{G}(B_j)] &= \mathbb{E}[(\widehat{a}_j - \widehat{b}_j)(\widehat{c}_j - \widehat{d}_j)\widehat{b}_j(1 - \widehat{b}_j)] \\ &= \mathbb{E}[(\widehat{a}_j - \widehat{b}_j)(\widehat{b}_j - \widehat{b}_j^2)]\mathbb{E}[\widehat{c}_j - \widehat{d}_j] \\ &= -\mathbb{E}[(\widehat{b}_j - \widehat{a}_j)(\widehat{b}_j - \widehat{b}_j^2)]\mathbb{E}[\widehat{c}_j - \widehat{d}_j] \\ &= -\left[\mathbb{E}[\widehat{b}_j^2] - \mathbb{E}[\widehat{b}_j^3] - \mathbb{E}[\widehat{a}_j]\mathbb{E}[\widehat{b}_j] + \mathbb{E}[\widehat{a}_j]\mathbb{E}[\widehat{b}_j^2]\right]\left[\mathbb{E}[\widehat{c}_j] - \mathbb{E}[\widehat{d}_j]\right] \\ &\approx -\left[b_j^2 + \Phi_2(B_j)b_j(1 - b_j) - [b_j^3 + 3\Phi_2(B_j)b_j^3(1 - b_j) - a_jb_j\right. \\ &\quad \left.+ a_j[b_j^2 + \Phi_2(B_j)b_j(1 - b_j)]](c_j - d_j)\right] \\ &= -(b_j^2 - b_j^3 - a_jb_j + a_jb_j^2)(c_j - d_j) - \Phi_2(B_j)b_j(1 - b_j)(1 - 3b_j + a_j)(c_j - d_j) \\ &= (a_j - b_j)(c_j - d_j)b_j(1 - b_j) - \Phi_2(B_j)b_j(1 - b_j)[1 - 2b_j + (a_j - b_j)](c_j - d_j) \\ &= F_4(A_j, B_j; C_j, D_j)G(A_j) - \Phi_2(B_j)F_4(A_j, B_j; C_j, D_j)G(B_j) \\ &\quad - \Phi_2(B_j)G(B_j)(1 - 2b_j)(c_j - d_j) \\ &= [1 - \Phi_2(B_j)]F_4(A_j, B_j; C_j, D_j)G(B_j) - \Phi_2(B_j)G(B_j)(1 - 2b_j)(c_j - d_j). \end{aligned}$$

Hence, we have that

$$\begin{aligned} Cov[\widehat{F}_4(A_j, B_j; C_j, D_j), \widehat{G}(B_j)] &= \mathbb{E}[\widehat{F}_4(A_j, B_j; C_j, D_j)\widehat{G}(B_j)] - \mathbb{E}[\widehat{F}_4(A_j, B_j; C_j, D_j)]\mathbb{E}[\widehat{G}(B_j)] \\ &\approx [1 - \Phi_2(B_j)]F_4(A_j, B_j; C_j, D_j)G(B_j) - \Phi_2(B_j)G(B_j)(1 - 2b_j)(c_j - d_j) \\ &\quad - [1 - \Phi_2(B_j)]F_4(A_j, B_j; C_j, D_j)G(B_j) \\ &= -\Phi_2(B_j)G(B_j)(1 - 2b_j)(c_j - d_j). \end{aligned}$$

Parallel to the derivation for $P = A$, we have

$$Cov[\widehat{F}_4(A_j, B_j; C_j, D_j), \widehat{G}(C_j)] = \Phi_2(C_j)G(C_j)(1 - 2c_j)(a_j - b_j)$$

and parallel to the derivation for $P = B$, we have

$$Cov[\widehat{F}_4(A_j, B_j; C_j, D_j), \widehat{G}(D_j)] = -\Phi_2(D_j)G(D_j)(1 - 2d_j)(a_j - b_j).$$

We know that by independence of loci we have

$$\begin{aligned} Cov[\widehat{F}_4(A, B; C, D), \widehat{G}(P)] &= Cov\left[\frac{1}{J} \sum_{j=1}^J \widehat{F}_4(A_j, B_j; C_j, D_j), \frac{1}{J} \sum_{j=1}^J \widehat{G}(P_j)\right] \\ &= \frac{1}{J^2} \sum_{j=1}^J Cov[\widehat{F}_4(A_j, B_j; C_j, D_j), \widehat{G}(P_j)] \end{aligned}$$

which gives

$$\begin{aligned}
 \text{Cov}[\hat{F}_4(A, B; C, D), \hat{G}(A)] &\approx \frac{1}{J^2} \sum_{j=1}^J \Phi_2(A_j) G(A_j) (1 - 2a_j) (c_j - d_j) \\
 \text{Cov}[\hat{F}_4(A, B; C, D), \hat{G}(B)] &\approx -\frac{1}{J^2} \sum_{j=1}^J \Phi_2(B_j) G(B_j) (1 - 2b_j) (c_j - d_j) \\
 \text{Cov}[\hat{F}_4(A, B; C, D), \hat{G}(C)] &\approx \frac{1}{J^2} \sum_{j=1}^J \Phi_2(C_j) G(C_j) (1 - 2c_j) (a_j - b_j) \\
 \text{Cov}[\hat{F}_4(A, B; C, D), \hat{G}(D)] &\approx -\frac{1}{J^2} \sum_{j=1}^J \Phi_2(D_j) G(D_j) (1 - 2d_j) (a_j - b_j). \quad \square
 \end{aligned}$$

Proposition 21. Consider J polymorphic loci in populations A , B , C , and D with respective parametric reference allele frequencies $a_j, b_j, c_j, d_j \in (0, 1)$, and suppose we take a random sample of $N(P_j)$ individuals at locus j in population $P \in \{A, B, C, D\}$, some of which may be related or inbred. Moreover, assume that no individual is related to more than one other individual, which makes the terms $\Phi_3(P_j)$, $\Phi_4(P_j)$, $\Phi_{2,2}(P_j)$, and $\Phi_2(P_j)^2$ negligible to $\Phi_2(P_j)$. Based on this simplifying assumption, the ratio estimator $\hat{F}_4(A, B; C, D | P)$ has approximate variance

$$\begin{aligned}
 \text{Var}[\hat{F}_4(A, B; C, D | P)] &\approx \frac{F_4(A, B; C, D)^2}{\mathbb{E}[\hat{G}(P)]^2} \left[\frac{\text{Var}[\hat{F}_4(A, B; C, D)]}{F_4(A, B; C, D)^2} + \frac{\text{Var}[\hat{G}(A)]}{\mathbb{E}[\hat{G}(P)]^2} \right. \\
 &\quad \left. - 2 \frac{\text{Cov}[\hat{F}_4(A, B; C, D), \hat{G}(P)]}{F_4(A, B; C, D) \mathbb{E}[\hat{G}(P)]} \right],
 \end{aligned}$$

where the expectation is

$$\mathbb{E}[\hat{G}(P)] = G(P) - \frac{1}{J} \sum_{j=1}^J \Phi_2(P_j) G(P_j)$$

the variances are

$$\begin{aligned}
 \text{Var}[\hat{F}_4(A, B; C, D)] &\approx \frac{1}{J^2} \sum_{j=1}^J [\Phi_2(C_j) g(C_j) + \Phi_2(D_j) G(D_j)] F_2(A_j, B_j) \\
 &\quad + \frac{1}{J^2} \sum_{j=1}^J [\Phi_2(A_j) G(A_j) + \Phi_2(B_j) G(B_j)] F_2(C_j, D_j) \\
 \text{Var}[\hat{G}(P)] &\approx \frac{1}{J^2} \sum_{j=1}^J \Phi_2(P_j) G(P_j) - \frac{4}{J^2} \sum_{j=1}^J \Phi_2(P_j) G(P_j)^2
 \end{aligned}$$

and the covariances are

$$\begin{aligned}
 \text{Cov}[\hat{F}_4(A, B; C, D), \hat{G}(A)] &\approx \frac{1}{J^2} \sum_{j=1}^J \Phi_2(A_j) G(A_j) (1 - 2a_j) (c_j - d_j) \\
 \text{Cov}[\hat{F}_4(A, B; C, D), \hat{G}(B)] &\approx -\frac{1}{J^2} \sum_{j=1}^J \Phi_2(B_j) G(B_j) (1 - 2b_j) (c_j - d_j) \\
 \text{Cov}[\hat{F}_4(A, B; C, D), \hat{G}(C)] &\approx \frac{1}{J^2} \sum_{j=1}^J \Phi_2(C_j) G(C_j) (1 - 2c_j) (a_j - b_j) \\
 \text{Cov}[\hat{F}_4(A, B; C, D), \hat{G}(D)] &\approx -\frac{1}{J^2} \sum_{j=1}^J \Phi_2(D_j) G(D_j) (1 - 2d_j) (a_j - b_j).
 \end{aligned}$$

Proof. Recall that

$$\hat{F}_4(A, B; C, D | P) = \frac{\hat{F}_4(A, B; C, D)}{\hat{G}(P)},$$

where $\hat{F}_4(A, B; C, D)$ is an unbiased estimator for $F_4(A, B; C, D)$. Assuming that $X = \hat{F}_4(A, B; C, D)$ and $Y = \hat{G}(P)$, following the approximation in Wolter (2007) we have

$$\begin{aligned} \text{Var}[\hat{F}_4(A, B; C, D | P)] \approx & \frac{F_4(A, B; C, D)^2}{\mathbb{E}[\hat{G}(P)]^2} \left[\frac{\text{Var}[\hat{F}_4(A, B; C, D)]}{F_4(A, B; C, D)^2} + \frac{\text{Var}[\hat{G}(P)]}{\mathbb{E}[\hat{G}(P)]^2} \right. \\ & \left. - 2 \frac{\text{Cov}[\hat{F}_4(A, B; C, D), \hat{G}(P)]}{F_4(A, B; C, D) \mathbb{E}[\hat{G}(P)]} \right], \end{aligned}$$

where $\mathbb{E}[\hat{G}(P)]$ is given in Lemma 1, $\text{Var}[\hat{F}_4(A, B; C, D)]$ in Proposition 16, $\text{Var}[\hat{G}(P)]$ in Lemma 9, and $\text{Cov}[\hat{F}_4(A, B; C, D), \hat{G}(P)]$ in Lemma 20 for each population $P \in \{A, B, C, D\}$. \square

Lemma 22. Consider J independent polymorphic loci in populations A, B, C , and D with respective parametric reference allele frequencies $a_j, b_j, c_j, d_j \in (0, 1)$, and suppose we take a random sample of $N(P_j)$ individuals at locus j in population $P \in \{A, B, C, D\}$, some of which may be related or inbred. Moreover, assume that no individual is related to more than one other individual, which makes the terms $\Phi_3(P_j)$, $\Phi_4(P_j)$, $\Phi_{2,2}(P_j)$, and $\Phi_2(P_j)^2$ negligible to $\Phi_2(P_j)$. Based on this simplifying assumption, the unbiased estimators $\hat{F}_4(A, B; C, D)$ and $\tilde{G}(P)$, $P \in \{A, B, C, D\}$, have approximate covariances

$$\begin{aligned} \text{Cov}[\hat{F}_4(A, B; C, D), \tilde{G}(A)] &\approx \frac{1}{J^2} \sum_{j=1}^J \frac{\Phi_2(A_j)}{1 - \Phi_2(A_j)} G(A_j)(1 - 2a_j)(c_j - d_j) \\ \text{Cov}[\hat{F}_4(A, B; C, D), \tilde{G}(B)] &\approx -\frac{1}{J^2} \sum_{j=1}^J \frac{\Phi_2(B_j)}{1 - \Phi_2(B_j)} G(B_j)(1 - 2b_j)(c_j - d_j) \\ \text{Cov}[\hat{F}_4(A, B; C, D), \tilde{G}(C)] &\approx \frac{1}{J^2} \sum_{j=1}^J \frac{\Phi_2(C_j)}{1 - \Phi_2(C_j)} G(C_j)(1 - 2c_j)(a_j - b_j) \\ \text{Cov}[\hat{F}_4(A, B; C, D), \tilde{G}(D)] &\approx -\frac{1}{J^2} \sum_{j=1}^J \frac{\Phi_2(D_j)}{1 - \Phi_2(D_j)} G(D_j)(1 - 2d_j)(a_j - b_j). \end{aligned}$$

Proof. Recall that $\tilde{G}(P_j) = \hat{G}(P_j)/[1 - \Phi_2(P_j)]$. It follows that

$$\text{Cov}[\hat{F}_4(A_j, B_j; C_j, D_j), \tilde{G}(P_j)] = \frac{1}{1 - \Phi_2(P_j)} \text{Cov}[\hat{F}_4(A_j, B_j; C_j, D_j), \hat{G}(P_j)].$$

From the proof of Lemma 20, we have

$$\begin{aligned} \text{Cov}[\hat{F}_4(A_j, B_j; C_j, D_j), \tilde{G}(A_j)] &\approx \frac{\Phi_2(A_j)}{1 - \Phi_2(A_j)} G(A_j)(1 - 2a_j)(c_j - d_j) \\ \text{Cov}[\hat{F}_4(A_j, B_j; C_j, D_j), \tilde{G}(B_j)] &\approx -\frac{\Phi_2(B_j)}{1 - \Phi_2(B_j)} G(B_j)(1 - 2b_j)(c_j - d_j) \\ \text{Cov}[\hat{F}_4(A_j, B_j; C_j, D_j), \tilde{G}(C_j)] &\approx \frac{\Phi_2(C_j)}{1 - \Phi_2(C_j)} G(C_j)(1 - 2c_j)(a_j - b_j) \\ \text{Cov}[\hat{F}_4(A_j, B_j; C_j, D_j), \tilde{G}(D_j)] &\approx -\frac{\Phi_2(D_j)}{1 - \Phi_2(D_j)} G(D_j)(1 - 2d_j)(a_j - b_j), \end{aligned}$$

yielding

$$\begin{aligned}
 \text{Cov}[\hat{F}_4(A, B; C, D), \tilde{G}(A)] &\approx \frac{1}{J^2} \sum_{j=1}^J \frac{\Phi_2(A_j)}{1 - \Phi_2(A_j)} G(A_j)(1 - 2a_j)(c_j - d_j) \\
 \text{Cov}[\hat{F}_4(A, B; C, D), \tilde{G}(B)] &\approx -\frac{1}{J^2} \sum_{j=1}^J \frac{\Phi_2(B_j)}{1 - \Phi_2(B_j)} G(B_j)(1 - 2b_j)(c_j - d_j) \\
 \text{Cov}[\hat{F}_4(A, B; C, D), \tilde{G}(C)] &\approx \frac{1}{J^2} \sum_{j=1}^J \frac{\Phi_2(C_j)}{1 - \Phi_2(C_j)} G(C_j)(1 - 2c_j)(a_j - b_j) \\
 \text{Cov}[\hat{F}_4(A, B; C, D), \tilde{G}(D)] &\approx -\frac{1}{J^2} \sum_{j=1}^J \frac{\Phi_2(D_j)}{1 - \Phi_2(D_j)} G(D_j)(1 - 2d_j)(a_j - b_j). \quad \square
 \end{aligned}$$

Proposition 23. Consider J polymorphic loci in populations A , B , C , and D with respective parametric reference allele frequencies $a_j, b_j, c_j, d_j \in (0, 1)$, and suppose we take a random sample of $N(P_j)$ individuals at locus j in population $P \in \{A, B, C, D\}$, some of which may be related or inbred. Moreover, assume that no individual is related to more than one other individual, which makes the terms $\Phi_3(P_j)$, $\Phi_4(P_j)$, $\Phi_{2,2}(P_j)$, and $\Phi_2(P_j)^2$ negligible to $\Phi_2(P_j)$. Based on this simplifying assumption, the approximately unbiased ratio estimator $\tilde{F}_4(A, B; C, D | P)$ has approximate variance

$$\begin{aligned}
 \text{Var}[\tilde{F}_4(A, B; C, D | P)] &\approx \frac{F_4(A, B; C, D)^2}{G(P)^2} \left[\frac{\text{Var}[\hat{F}_4(A, B; C, D)]}{F_4(A, B; C, D)^2} + \frac{\text{Var}[\tilde{G}(P)]}{G(P)^2} \right. \\
 &\quad \left. - 2 \frac{\text{Cov}[\hat{F}_4(A, B; C, D), \tilde{G}(P)]}{F_4(A, B; C, D)G(P)} \right],
 \end{aligned}$$

where the variances are

$$\begin{aligned}
 \text{Var}[\hat{F}_4(A, B; C, D)] &\approx \frac{1}{J^2} \sum_{j=1}^J [\Phi_2(C_j)G(C_j) + \Phi_2(D_j)G(D_j)]F_2(A_j, B_j) \\
 &\quad + \frac{1}{J^2} \sum_{j=1}^J [\Phi_2(A_j)G(A_j) + \Phi_2(B_j)G(B_j)]F_2(C_j, D_j) \\
 \text{Var}[\tilde{G}(P)] &\approx \frac{1}{J^2} \sum_{j=1}^J \frac{\Phi_2(P_j)}{1 - 2\Phi_2(P_j)} G(P_j) - \frac{4}{J^2} \sum_{j=1}^J \frac{\Phi_2(P_j)}{1 - 2\Phi_2(P_j)} G(P_j)^2
 \end{aligned}$$

and the covariances are

$$\begin{aligned}
 \text{Cov}[\hat{F}_4(A, B; C, D), \tilde{G}(A)] &\approx \frac{1}{J^2} \sum_{j=1}^J \frac{\Phi_2(A_j)}{1 - \Phi_2(A_j)} G(A_j)(1 - 2a_j)(c_j - d_j) \\
 \text{Cov}[\hat{F}_4(A, B; C, D), \tilde{G}(B)] &\approx -\frac{1}{J^2} \sum_{j=1}^J \frac{\Phi_2(B_j)}{1 - \Phi_2(B_j)} G(B_j)(1 - 2b_j)(c_j - d_j) \\
 \text{Cov}[\hat{F}_4(A, B; C, D), \tilde{G}(C)] &\approx \frac{1}{J^2} \sum_{j=1}^J \frac{\Phi_2(C_j)}{1 - \Phi_2(C_j)} G(C_j)(1 - 2c_j)(a_j - b_j) \\
 \text{Cov}[\hat{F}_4(A, B; C, D), \tilde{G}(D)] &\approx -\frac{1}{J^2} \sum_{j=1}^J \frac{\Phi_2(D_j)}{1 - \Phi_2(D_j)} G(D_j)(1 - 2d_j)(a_j - b_j).
 \end{aligned}$$

Proof. Recall that

$$\tilde{F}_4(A, B; C, D | P) = \frac{\hat{F}_4(A, B; C, D)}{\tilde{G}(P)},$$

where $\hat{F}_4(A, B; C, D)$ is an unbiased estimator for $F_4(A, B; C, D)$ and $\tilde{G}(P)$ is an unbiased estimator of $G(P)$. Assuming that $X = \hat{F}_4(A, B; C, D)$ and $Y = \tilde{G}(P)$, following the approximation in Wolter (2007) we have

$$\text{Var}[\tilde{F}_4(A, B; C, D | P)] \approx \frac{F_4(A, B; C, D)^2}{G(P)^2} \left[\frac{\text{Var}[\hat{F}_4(A, B; C, D)]}{F_4(A, B; C, D)^2} + \frac{\text{Var}[\tilde{G}(P)]}{G(P)^2} - 2 \frac{\text{Cov}[\hat{F}_4(A, B; C, D), \tilde{G}(P)]}{F_4(A, B; C, D)G(P)} \right],$$

where $\text{Var}[\hat{F}_4(A, B; C, D)]$ is given in Proposition 16, $\text{Var}[\tilde{G}(P)]$ in Lemma 9, and $\text{Cov}[\hat{F}_4(A, B; C, D), \tilde{G}(P)]$ in Lemma 22 for each population $P \in \{A, B, C, D\}$. \square

Lemma 24. Consider J independent polymorphic loci in populations A, B, C , and D with respective parametric reference allele frequencies $a_j, b_j, c_j, d_j \in (0, 1)$, and suppose we take a random sample of $N(P_j)$ individuals at locus j in population $P \in \{A, B, C, D\}$, some of which may be related or inbred. Moreover, assume that no individual is related to more than one other individual, which makes the terms $\Phi_3(P_j)$, $\Phi_4(P_j)$, $\Phi_{2,2}(P_j)$, $\Phi_2(P_j)^2$, $\Phi_2(A_j)\Phi_2(B_j)$, $\Phi_2(A_j)\Phi_2(C_j)$, $\Phi_2(A_j)\Phi_2(D_j)$, $\Phi_2(B_j)\Phi_2(C_j)$, $\Phi_2(B_j)\Phi_2(D_j)$, and $\Phi_2(C_j)\Phi_2(D_j)$ negligible to $\Phi_2(P_j)$. Based on this simplifying assumption, the unbiased estimator $\hat{H}(A, B, C, D)$ has approximate variance

$$\begin{aligned} \text{Var}[\hat{H}(A, B; C, D)] &\approx \frac{1}{J^2} \sum_{j=1}^J (a_j + b_j - 2a_j b_j)^2 [\Phi_2(C_j)G(C_j)(1 - 2d_j)^2 + \Phi_2(D_j)G(D_j)(1 - 2c_j)^2] \\ &\quad + \frac{1}{J^2} \sum_{j=1}^J (c_j + d_j - 2c_j d_j)^2 [\Phi_2(A_j)G(A_j)(1 - 2b_j)^2 + \Phi_2(B_j)G(B_j)(1 - 2a_j)^2]. \end{aligned}$$

Proof. From the proof of Lemma 8, we have that

$$\mathbb{E}[\hat{H}(A_j, B_j, C_j, D_j)] = H(A_j, B_j, C_j, D_j),$$

yielding

$$\mathbb{E}[\hat{H}(A_j, B_j, C_j, D_j)]^2 = H(A_j, B_j, C_j, D_j)^2.$$

We first calculate

$$\begin{aligned} \mathbb{E}[\hat{H}(A_j, B_j, C_j, D_j)^2] &= \mathbb{E}[(\hat{a}_j + \hat{b}_j - 2\hat{a}_j \hat{b}_j)^2 (\hat{c}_j + \hat{d}_j - 2\hat{c}_j \hat{d}_j)^2] \\ &= \mathbb{E}[(\hat{a}_j + \hat{b}_j - 2\hat{a}_j \hat{b}_j)^2] \mathbb{E}[(\hat{c}_j + \hat{d}_j - 2\hat{c}_j \hat{d}_j)^2]. \end{aligned}$$

We compute the first term as

$$\begin{aligned} \mathbb{E}[(\hat{a}_j + \hat{b}_j - 2\hat{a}_j \hat{b}_j)^2] &= \mathbb{E}[\hat{a}_j^2] + \mathbb{E}[\hat{b}_j^2] + 4\mathbb{E}[\hat{a}_j^2]\mathbb{E}[\hat{b}_j^2] + 2\mathbb{E}[\hat{a}_j]\mathbb{E}[\hat{b}_j] - 4\mathbb{E}[\hat{a}_j^2]\mathbb{E}[\hat{b}_j] - 4\mathbb{E}[\hat{a}_j]\mathbb{E}[\hat{b}_j^2] \\ &= \mathbb{E}[\hat{a}_j^2] + \mathbb{E}[\hat{b}_j^2] + 4\mathbb{E}[\hat{a}_j^2]\mathbb{E}[\hat{b}_j^2] + 2a_j b_j - 4\mathbb{E}[\hat{a}_j^2]b_j - 4a_j\mathbb{E}[\hat{b}_j^2] \\ &\approx a_j^2 + \Phi_2(A_j)a_j(1 - a_j) + b_j^2 + \Phi_2(B_j)b_j(1 - b_j) \\ &\quad + 4[a_j^2 + \Phi_2(A_j)a_j(1 - a_j)][b_j^2 + \Phi_2(B_j)b_j(1 - b_j)] + 2a_j b_j \\ &\quad - 4[a_j^2 + \Phi_2(A_j)a_j(1 - a_j)]b_j - 4a_j[b_j^2 + \Phi_2(B_j)b_j(1 - b_j)] \\ &= a_j^2 + b_j^2 + 4a_j^2 b_j^2 + 2a_j b_j - 4a_j^2 b_j - 4a_j b_j^2 + \Phi_2(A_j)a_j(1 - a_j) + \Phi_2(B_j)b_j(1 - b_j) \\ &\quad + 4\Phi_2(A_j)a_j(1 - a_j)b_j^2 + 4\Phi_2(B_j)a_j^2 b_j(1 - b_j) + 4\Phi_2(A_j)\Phi_2(B_j)a_j(1 - a_j)b_j(1 - b_j) \\ &\quad - 4\Phi_2(A_j)a_j(1 - a_j)b_j - 4\Phi_2(B_j)a_j b_j(1 - b_j) \\ &= (a_j + b_j - 2a_j b_j)^2 + \Phi_2(A_j)G(A_j)[1 - 4b_j + 4b_j^2] + \Phi_2(B_j)G(B_j)[1 - 4a_j + 4a_j^2] \\ &\quad + 4\Phi_2(A_j)\Phi_2(B_j)G(A_j)G(B_j) \\ &\approx (a_j + b_j - 2a_j b_j)^2 + \Phi_2(A_j)G(A_j)(1 - 2b_j)^2 + \Phi_2(B_j)G(B_j)(1 - 2a_j)^2, \end{aligned}$$

where we used the fact that $\Phi_2(A_j)\Phi_2(B_j)$ is negligible compared to $\Phi_2(A_j)$ and $\Phi_2(B_j)$ as an approximation. Using a similar argument we have that

$$\mathbb{E}[(\hat{c}_j + \hat{d}_j - 2\hat{c}_j\hat{d}_j)^2] \approx (c_j + d_j - 2c_jd_j)^2 + \Phi_2(C_j)G(C_j)(1 - 2d_j)^2 + \Phi_2(D_j)G(D_j)(1 - 2c_j)^2.$$

Hence, we have that

$$\begin{aligned} \mathbb{E}[\hat{H}(A_j, B_j, C_j, D_j)^2] &= \mathbb{E}[(\hat{a}_j + \hat{b}_j - 2\hat{a}_j\hat{b}_j)^2]\mathbb{E}[(\hat{c}_j + \hat{d}_j - 2\hat{c}_j\hat{d}_j)^2] \\ &\approx \left[(a_j + b_j - 2a_jb_j)^2 + \Phi_2(A_j)G(A_j)(1 - 2b_j)^2 + \Phi_2(B_j)G(B_j)(1 - 2a_j)^2 \right] \\ &\quad \times \left[(c_j + d_j - 2c_jd_j)^2 + \Phi_2(C_j)G(C_j)(1 - 2d_j)^2 + \Phi_2(D_j)G(D_j)(1 - 2c_j)^2 \right] \\ &\approx H(A_j, B_j, C_j, D_j)^2 \\ &\quad + (a_j + b_j - 2a_jb_j)^2[\Phi_2(C_j)G(C_j)(1 - 2d_j)^2 + \Phi_2(D_j)G(D_j)(1 - 2c_j)^2] \\ &\quad + (c_j + d_j - 2c_jd_j)^2[\Phi_2(A_j)G(A_j)(1 - 2b_j)^2 + \Phi_2(B_j)G(B_j)(1 - 2a_j)^2], \end{aligned}$$

where we used the fact that $\Phi_2(A_j)\Phi_2(C_j)$, $\Phi_2(A_j)\Phi_2(D_j)$, $\Phi_2(B_j)\Phi_2(C_j)$, and $\Phi_2(B_j)\Phi_2(D_j)$ are negligible compared to $\Phi_2(A_j)$, $\Phi_2(B_j)$, $\Phi_2(C_j)$, and $\Phi_2(D_j)$ as an approximation. Putting it together, we have

$$\begin{aligned} \text{Var}[\hat{H}(A_j, B_j, C_j, D_j)] &= \mathbb{E}[\hat{H}(A_j, B_j, C_j, D_j)^2] - \mathbb{E}[\hat{H}(A_j, B_j, C_j, D_j)]^2 \\ &= \mathbb{E}[\hat{H}(A_j, B_j, C_j, D_j)^2] - H(A_j, B_j, C_j, D_j)^2 \\ &\approx (a_j + b_j - 2a_jb_j)^2[\Phi_2(C_j)G(C_j)(1 - 2d_j)^2 + \Phi_2(D_j)G(D_j)(1 - 2c_j)^2] \\ &\quad + (c_j + d_j - 2c_jd_j)^2[\Phi_2(A_j)G(A_j)(1 - 2b_j)^2 + \Phi_2(B_j)G(B_j)(1 - 2a_j)^2]. \end{aligned}$$

Given the assumption of independent loci, we have

$$\begin{aligned} \text{Var}[\hat{H}(A, B, C, D)] &= \text{Var}\left[\frac{1}{J} \sum_{j=1}^J \hat{H}(A_j, B_j, C_j, D_j)\right] \\ &= \frac{1}{J^2} \sum_{j=1}^J \text{Var}[\hat{H}(A_j, B_j, C_j, D_j)] \\ &\approx \frac{1}{J^2} \sum_{j=1}^J (a_j + b_j - 2a_jb_j)^2[\Phi_2(C_j)G(C_j)(1 - 2d_j)^2 + \Phi_2(D_j)G(D_j)(1 - 2c_j)^2] \\ &\quad + \frac{1}{J^2} \sum_{j=1}^J (c_j + d_j - 2c_jd_j)^2[\Phi_2(A_j)G(A_j)(1 - 2b_j)^2 + \Phi_2(B_j)G(B_j)(1 - 2a_j)^2]. \quad \square \end{aligned}$$

Lemma 25. Consider J independent polymorphic loci in populations A , B , C , and D with respective parametric reference allele frequencies $a_j, b_j, c_j, d_j \in (0, 1)$, and suppose we take a random sample of $N(P_j)$ individuals at locus j in population $P \in \{A, B, C, D\}$, some of which may be related or inbred. Moreover, assume that no individual is related to more than one other individual, which makes the terms $\Phi_3(P_j)$, $\Phi_4(P_j)$, $\Phi_{2,2}(P_j)$, $\Phi_2(P_j)^2$, $\Phi_2(A_j)\Phi_2(C_j)$, $\Phi_2(A_j)\Phi_2(D_j)$, $\Phi_2(B_j)\Phi_2(C_j)$, and $\Phi_2(B_j)\Phi_2(D_j)$ negligible to $\Phi_2(P_j)$. Based on this simplifying assumption, the unbiased estimators $\hat{F}_4(A, B; C, D)$ and $\hat{H}(A, B, C, D)$

have approximate covariance

$$\begin{aligned} \text{Cov}[\widehat{F}_4(A, B; C, D), \widehat{H}(A, B, C, D)] &\approx \frac{1}{J^2} \sum_{j=1}^J \Phi_2(A_j) G(A_j) (c_j - d_j) (c_j + d_j - 2c_j d_j) (1 - 2b_j) \\ &\quad - \frac{1}{J^2} \sum_{j=1}^J \Phi_2(B_j) G(B_j) (c_j - d_j) (c_j + d_j - 2c_j d_j) (1 - 2a_j) \\ &\quad + \frac{1}{J^2} \sum_{j=1}^J \Phi_2(C_j) G(C_j) (a_j - b_j) (a_j + b_j - 2a_j b_j) (1 - 2d_j) \\ &\quad - \frac{1}{J^2} \sum_{j=1}^J \Phi_2(D_j) G(D_j) (a_j - b_j) (a_j + b_j - 2a_j b_j) (1 - 2c_j). \end{aligned}$$

Proof. From the proofs of Proposition 5 and Lemma 8, we that have

$$\mathbb{E}[\widehat{F}_4(A_j, B_j; C_j, D_j)] = F_4(A_j, B_j; C_j, D_j),$$

and

$$\mathbb{E}[\widehat{H}(A_j, B_j, C_j, D_j)] = H(A_j, B_j, C_j, D_j),$$

yielding

$$\mathbb{E}[\widehat{F}_4(A_j, B_j; C_j, D_j)] \mathbb{E}[\widehat{H}(A_j, B_j, C_j, D_j)] = F_4(A_j, B_j; C_j, D_j) H(A_j, B_j, C_j, D_j).$$

We first calculate

$$\begin{aligned} \mathbb{E}[\widehat{F}_4(A_j, B_j; C_j, D_j) \widehat{H}(A_j, B_j, C_j, D_j)] &= \mathbb{E}[(\widehat{a}_j - \widehat{b}_j)(\widehat{c}_j - \widehat{d}_j)(\widehat{a}_j + \widehat{b}_j - 2\widehat{a}_j \widehat{b}_j)(\widehat{c}_j + \widehat{d}_j - 2\widehat{c}_j \widehat{d}_j)] \\ &= \mathbb{E}[(\widehat{a}_j - \widehat{b}_j)(\widehat{a}_j + \widehat{b}_j - 2\widehat{a}_j \widehat{b}_j)] \mathbb{E}[(\widehat{c}_j - \widehat{d}_j)(\widehat{c}_j + \widehat{d}_j - 2\widehat{c}_j \widehat{d}_j)]. \end{aligned}$$

We compute the first term as

$$\begin{aligned} \mathbb{E}[(\widehat{a}_j - \widehat{b}_j)(\widehat{a}_j + \widehat{b}_j - 2\widehat{a}_j \widehat{b}_j)] &= \mathbb{E}[\widehat{a}_j^2] - 2\mathbb{E}[\widehat{a}_j^2] \mathbb{E}[\widehat{b}_j] - \mathbb{E}[\widehat{b}_j^2] + 2\mathbb{E}[\widehat{a}_j] \mathbb{E}[\widehat{b}_j^2] \\ &= \mathbb{E}[\widehat{a}_j^2] (1 - 2\mathbb{E}[\widehat{b}_j]) - \mathbb{E}[\widehat{b}_j^2] (1 - 2\mathbb{E}[\widehat{a}_j]) \\ &\approx [a_j^2 + \Phi_2(A_j) a_j (1 - a_j)] [1 - 2b_j] - [b_j^2 + \Phi_2(B_j) b_j (1 - b_j)] [1 - 2a_j] \\ &= (a_j - b_j)(a_j + b_j - 2a_j b_j) + \Phi_2(A_j) G(A_j) (1 - 2b_j) - \Phi_2(B_j) G(B_j) (1 - 2a_j). \end{aligned}$$

Using a similar argument, we have that

$$\mathbb{E}[(\widehat{c}_j - \widehat{d}_j)(\widehat{c}_j + \widehat{d}_j - 2\widehat{c}_j \widehat{d}_j)] \approx (c_j - d_j)(c_j + d_j - 2c_j d_j) + \Phi_2(C_j) G(C_j) (1 - 2d_j) - \Phi_2(D_j) G(D_j) (1 - 2c_j).$$

Hence, we have that

$$\begin{aligned} \mathbb{E}[\widehat{F}_4(A_j, B_j; C_j, D_j) \widehat{H}(A_j, B_j, C_j, D_j)] &= \mathbb{E}[(\widehat{a}_j - \widehat{b}_j)(\widehat{a}_j + \widehat{b}_j - 2\widehat{a}_j \widehat{b}_j)] \mathbb{E}[(\widehat{c}_j - \widehat{d}_j)(\widehat{c}_j + \widehat{d}_j - 2\widehat{c}_j \widehat{d}_j)] \\ &\approx \left[(a_j - b_j)(a_j + b_j - 2a_j b_j) + \Phi_2(A_j) G(A_j) (1 - 2b_j) \right. \\ &\quad \left. - \Phi_2(B_j) G(B_j) (1 - 2a_j) \right] \\ &\quad \times \left[(c_j - d_j)(c_j + d_j - 2c_j d_j) + \Phi_2(C_j) G(C_j) (1 - 2d_j) \right. \\ &\quad \left. - \Phi_2(D_j) G(D_j) (1 - 2c_j) \right] \\ &\approx F_4(A_j, B_j; C_j, D_j) H(A_j, B_j, C_j, D_j) \\ &\quad + \Phi_2(A_j) G(A_j) (c_j - d_j) (c_j + d_j - 2c_j d_j) (1 - 2b_j) \\ &\quad - \Phi_2(B_j) G(B_j) (c_j - d_j) (c_j + d_j - 2c_j d_j) (1 - 2a_j) \\ &\quad + \Phi_2(C_j) G(C_j) (a_j - b_j) (a_j + b_j - 2a_j b_j) (1 - 2d_j) \\ &\quad - \Phi_2(D_j) G(D_j) (a_j - b_j) (a_j + b_j - 2a_j b_j) (1 - 2c_j), \end{aligned}$$

where we used the fact that $\Phi_2(A_j)\Phi_2(C_j)$, $\Phi_2(A_j)\Phi_2(D_j)$, $\Phi_2(B_j)\Phi_2(C_j)$, and $\Phi_2(B_j)\Phi_2(D_j)$ are negligible compared to $\Phi_2(A_j)$, $\Phi_2(B_j)$, $\Phi_2(C_j)$, and $\Phi_2(D_j)$ as an approximation. Putting it together, we have

$$\begin{aligned} \text{Cov}[\hat{F}_4(A_j, B_j; C_j, D_j), \hat{H}(A_j, B_j, C_j, D_j)] &= \mathbb{E}[\hat{F}_4(A_j, B_j; C_j, D_j)\hat{H}(A_j, B_j, C_j, D_j)] \\ &\quad - \mathbb{E}[\hat{F}_4(A_j, B_j; C_j, D_j)]\mathbb{E}[\hat{H}(A_j, B_j, C_j, D_j)] \\ &= \mathbb{E}[\hat{F}_4(A_j, B_j; C_j, D_j)\hat{H}(A_j, B_j, C_j, D_j)] \\ &\quad - F_4(A_j, B_j; C_j, D_j)H(A_j, B_j, C_j, D_j) \\ &\approx \Phi_2(A_j)G(A_j)(c_j - d_j)(c_j + d_j - 2c_jd_j)(1 - 2b_j) \\ &\quad - \Phi_2(B_j)G(B_j)(c_j - d_j)(c_j + d_j - 2c_jd_j)(1 - 2a_j) \\ &\quad + \Phi_2(C_j)G(C_j)(a_j - b_j)(a_j + b_j - 2a_jb_j)(1 - 2d_j) \\ &\quad - \Phi_2(D_j)G(D_j)(a_j - b_j)(a_j + b_j - 2a_jb_j)(1 - 2c_j). \end{aligned}$$

Given the assumption of independent loci, we have

$$\begin{aligned} \text{Cov}[\hat{F}_4(A, B; C, D), \hat{H}(A, B, C, D)] &= \text{Cov}\left[\frac{1}{J} \sum_{j=1}^J \hat{F}_4(A_j, B_j; C_j, D_j), \frac{1}{J} \sum_{j=1}^J \hat{H}(A_j, B_j, C_j, D_j)\right] \\ &= \frac{1}{J^2} \sum_{j=1}^J \text{Cov}[\hat{F}_4(A_j, B_j; C_j, D_j), \hat{H}(A_j, B_j, C_j, D_j)] \\ &\approx \frac{1}{J^2} \sum_{j=1}^J \Phi_2(A_j)G(A_j)(c_j - d_j)(c_j + d_j - 2c_jd_j)(1 - 2b_j) \\ &\quad - \frac{1}{J^2} \sum_{j=1}^J \Phi_2(B_j)G(B_j)(c_j - d_j)(c_j + d_j - 2c_jd_j)(1 - 2a_j) \\ &\quad + \frac{1}{J^2} \sum_{j=1}^J \Phi_2(C_j)G(C_j)(a_j - b_j)(a_j + b_j - 2a_jb_j)(1 - 2d_j) \\ &\quad - \frac{1}{J^2} \sum_{j=1}^J \Phi_2(D_j)G(D_j)(a_j - b_j)(a_j + b_j - 2a_jb_j)(1 - 2c_j). \quad \square \end{aligned}$$

Proposition 26. Consider J polymorphic loci in populations A , B , C , and D with respective parametric reference allele frequencies $a_j, b_j, c_j, d_j \in (0, 1)$, and suppose we take a random sample of $N(P_j)$ individuals at locus j in population $P \in \{A, B, C, D\}$, some of which may be related or inbred. Moreover, assume that no individual is related to more than one other individual, which makes the terms $\Phi_3(P_j)$, $\Phi_4(P_j)$, $\Phi_{2,2}(P_j)$, $\Phi_2(P_j)^2$, $\Phi_2(A_j)\Phi_2(B_j)$, $\Phi_2(A_j)\Phi_2(C_j)$, $\Phi_2(A_j)\Phi_2(D_j)$, $\Phi_2(B_j)\Phi_2(C_j)$, $\Phi_2(B_j)\Phi_2(D_j)$, and $\Phi_2(C_j)\Phi_2(D_j)$ negligible to $\Phi_2(P_j)$. Based on this simplifying assumption, the approximately unbiased ratio estimator $\hat{D}(A, B, C, D)$ has approximate variance

$$\begin{aligned} \text{Var}[\hat{D}(A, B; C, D)] &\approx \frac{F_4(A, B; C, D)^2}{H(A, B, C, D)^2} \left[\frac{\text{Var}[\hat{F}_4(A, B; C, D)]}{F_4(A, B; C, D)^2} + \frac{\text{Var}[\hat{H}(A, B, C, D)]}{H(A, B, C, D)^2} \right. \\ &\quad \left. - 2 \frac{\text{Cov}[\hat{F}_4(A, B; C, D), \hat{H}(A, B, C, D)]}{F_4(A, B; C, D)H(A, B, C, D)} \right], \end{aligned}$$

where the variances are

$$\begin{aligned} Var[\hat{F}_4(A, B; C, D)] &\approx \frac{1}{J^2} \sum_{j=1}^J [\Phi_2(C_j)G(C_j) + \Phi_2(D_j)G(D_j)]F_2(A_j, B_j) \\ &\quad + \frac{1}{J^2} \sum_{j=1}^J [\Phi_2(A_j)G(A_j) + \Phi_2(B_j)G(B_j)]F_2(C_j, D_j) \\ Var[\hat{H}(A, B; C, D)] &\approx \frac{1}{J^2} \sum_{j=1}^J (a_j + b_j - 2a_jb_j)^2 [\Phi_2(C_j)G(C_j)(1 - 2d_j)^2 + \Phi_2(D_j)G(D_j)(1 - 2c_j)^2] \\ &\quad + \frac{1}{J^2} \sum_{j=1}^J (c_j + d_j - 2c_jd_j)^2 [\Phi_2(A_j)G(A_j)(1 - 2b_j)^2 + \Phi_2(B_j)G(B_j)(1 - 2a_j)^2]. \end{aligned}$$

and the covariance is

$$\begin{aligned} Cov[\hat{F}_4(A, B; C, D), \hat{H}(A, B, C, D)] &\approx \frac{1}{J^2} \sum_{j=1}^J \Phi_2(A_j)G(A_j)(c_j - d_j)(c_j + d_j - 2c_jd_j)(1 - 2b_j) \\ &\quad - \frac{1}{J^2} \sum_{j=1}^J \Phi_2(B_j)G(B_j)(c_j - d_j)(c_j + d_j - 2c_jd_j)(1 - 2a_j) \\ &\quad + \frac{1}{J^2} \sum_{j=1}^J \Phi_2(C_j)G(C_j)(a_j - b_j)(a_j + b_j - 2a_jb_j)(1 - 2d_j) \\ &\quad - \frac{1}{J^2} \sum_{j=1}^J \Phi_2(D_j)G(D_j)(a_j - b_j)(a_j + b_j - 2a_jb_j)(1 - 2c_j). \end{aligned}$$

Proof. Recall that

$$\hat{D}(A, B; C, D) = \frac{\hat{F}_4(A, B; C, D)}{\hat{H}(A, B, C, D)},$$

where $\hat{F}_4(A, B; C, D)$ is an unbiased estimator for $F_4(A, B; C, D)$ and $\hat{H}(A, B, C, D)$ is an unbiased estimator of $H(A, B, C, D)$. Assuming that $X = \hat{F}_4(A, B; C, D)$ and $Y = \hat{H}(A, B, C, D)$, following the approximation in Wolter (2007) we have

$$\begin{aligned} Var[\tilde{D}(A, B; C, D)] &\approx \frac{F_4(A, B; C, D)^2}{H(A, B, C, D)^2} \left[\frac{Var[\hat{F}_4(A, B; C, D)]}{F_4(A, B; C, D)^2} + \frac{Var[\hat{H}(A, B, C, D)]}{H(A, B, C, D)^2} \right. \\ &\quad \left. - 2 \frac{Cov[\hat{F}_4(A, B; C, D), \hat{H}(A, B, C, D)]}{F_4(A, B; C, D)H(A, B, C, D)} \right], \end{aligned}$$

where $Var[\hat{F}_4(A, B; C, D)]$ is given in Proposition 16, $Var[\hat{H}(A, B, C, D)]$ in Lemma 24, and $Cov[\hat{F}_4(A, B; C, D), \hat{H}(A, B, C, D)]$ in Lemma 25. \square