OXFORD GENETICS

GENETICS, 2022, 220(1), iyab090
DOI: 10.1093/genetics/iyab090
Advance Access Publication Date: 15 July 2021
Investigation

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

Mehreen R. Mughal[1,*] and Michael DeGiorgio[2,*]

[1]Bioinformatics and Genomics at the Huck Institutes of the Life Sciences, Pennsylvania State University, University Park, PA 16802, USA and
[2]Department of Computer and Electrical Engineering and Computer Science, Florida Atlantic University, Boca Raton, FL 33431, USA

*Corresponding author: 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 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.

Keywords: inbreeding; relatedness; introgression; demography; population genetics; gene flow

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.* 2015; 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; Harris and DeGiorgio 2017a; Waples and Anderson 2017). 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 parametric population 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, which are unknown quantities that we will ultimately need to estimate prior to computing the *F*- and *D*-statistics from data. We begin by defining the true parametric population *F*- and *D*-statistics (Reich *et al.* 2009; Patterson *et al.* 2012) and then proceed in the *Theory* section to define sample estimators of these quantities, and show that our new estimators that account for related and inbred individuals reduce to the unbiased estimators in Appendix A of Patterson *et al.* (2012) when samples contain unrelated and noninbred individuals.

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

Received: November 20, 2021. **Accepted:** May 26, 2021

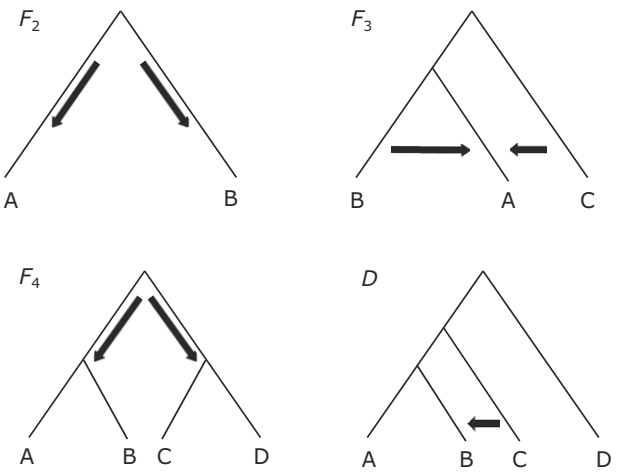

Figure 1 Trees showing the different relationships the F- and D-statistics are designed to test. $F_2(A,B)$ can test the differentiation of two populations A and B. $F_3(A;B,C)$ can test for introgression or relatedness between populations A and B or populations A and C. The placement of population A as an ingroup branch is arbitrary, because the $F_3(A;B,C)$ statistic measures the amount of genetic drift along branch A, and thus assumes the three-population tree is unrooted. However, because we are also showing how $F_3(A;B,C)$ can be employed to detect admixture in population A, we depict population A as an ingroup for visual convenience. $F_4(A,B;C,D)$ 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.

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

where for locus j

$$F_2(A_j,B_j) = (a_j - b_j)^2. \tag{2}$$

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

$$F_3(A;B,C) = \frac{1}{J}\sum_{j=1}^{J} F_3(A_j;B_j,C_j), \tag{3}$$

where for locus j

$$F_3(A_j;B_j,C_j) = (a_j - b_j)(a_j - c_j). \tag{4}$$

Because it measures genetic drift along a branch leading to a target population, its value is expected to be nonnegative. 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 nonnegative. 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

$$F_3(A;B,C|A) = \frac{F_3(A;B,C)}{2G(A)}, \tag{5}$$

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

$$G(P) = \frac{1}{J}\sum_{j=1} G(P_j) \tag{6}$$

such that for locus j

$$G(P_j) = p_j(1-p_j). \tag{7}$$

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

$$F_4(A,B;C,D) = \frac{1}{J}\sum_{j=1}^{J} F_4(A_j,B_j;C_j,D_j), \tag{8}$$

where for locus j

$$F_4(A_j,B_j;C_j,D_j) = (a_j - b_j)(c_j - d_j). \tag{9}$$

If the unrooted relationship is true, then F_4 takes the value of zero, and is nonzero otherwise. If it is known *a priori* that the unrooted relationship should be true, then a nonzero 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

$$F_4(A,B;C,D|P) = \frac{F_4(A,B;C,D)}{G(P)} \tag{10}$$

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

$$D(A,B,C,D) = -\frac{F_4(A,B;C,D)}{H(A,B,C,D)}, \tag{11}$$

where

$$H(A,B,C,D) = \frac{1}{J}\sum_{j=1}^{J} H(A_j,B_j,C_j,D_j) \tag{12}$$

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

$$H(A_j,B_j,C_j,D_j) = (a_j + b_j - 2a_jb_j)(c_j + d_j - 2c_jd_j). \tag{13}$$

If the rooted relationship is true, then D takes the value of zero, and is nonzero otherwise. A nonzero 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, $j = 1, 2, \ldots, J$, 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.*, McPeek *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

$$\hat{p} = \sum_{k=1}^{N(P)} \phi_k(P) X_k, \tag{14}$$

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. McPeek *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.

Plugging the sample estimate of allele frequencies in place of parametric population allele frequencies, estimators of the F- and D-statistics can be computed as

$$\hat{F}_2(A,B) = \frac{1}{J}\sum_{j=1}^{J} \hat{F}_2(A_j,B_j) \tag{15}$$

$$\hat{F}_3(A;B,C) = \frac{1}{J}\sum_{j=1}^{J} \hat{F}_3(A_j;B_j,C_j) \tag{16}$$

$$\hat{F}_4(A,B;C,D) = \frac{1}{J}\sum_{j=1}^{J} \hat{F}_4(A_j,B_j;C_j,D_j) \tag{17}$$

$$\hat{F}_3(A;B,C|A) = \frac{\hat{F}_3(A;B,C)}{2\hat{G}(A)} \tag{18}$$

$$\hat{F}_4(A,B;C,D|P) = \frac{\hat{F}_4(A,B;C,D)}{\hat{G}(P)} \tag{19}$$

$$\hat{D}(A,B,C,D) = -\frac{\hat{F}_4(A,B,C,D)}{\hat{H}(A,B,C,D)}, \tag{20}$$

where

$$\hat{F}_2(A_j,B_j) = (\hat{a}_j - \hat{b}_j)^2 \tag{21}$$

$$\hat{F}_3(A_j,B_j,C_j) = (\hat{a}_j - \hat{b}_j)(\hat{a}_j - \hat{c}_j) \tag{22}$$

$$\hat{F}_4(A_j,B_j;C_j,D_j) = (\hat{a}_j - \hat{b}_j)(\hat{c}_j - \hat{d}_j), \tag{23}$$

and where

$$\hat{G}(P) = \frac{1}{J}\sum_{j=1}^{J} \hat{G}(P_j) \tag{24}$$

$$\hat{H}(A,B,C,D) = \frac{1}{J}\sum_{j=1}^{J} \hat{H}(A_j,B_j,C_j,D_j) \tag{25}$$

with

$$\hat{G}(P_j) = \hat{p}_j(1-\hat{p}_j) \tag{26}$$

$$\hat{H}(A_j,B_j,C_j,D_j) = (\hat{a}_j + \hat{b}_j - 2\hat{a}_j\hat{b}_j)(\hat{c}_j + \hat{d}_j - 2\hat{c}_j\hat{d}_j). \tag{27}$$

In the following, we discuss properties of these estimators, and where appropriate, develop unbiased estimators for the statistics that are biased 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 noninbred individual x has inbreeding coefficient $f_x = 0$, and so if x is noninbred, then their kinship coefficient is $\Phi_{xx} = 1/m_x$. If an individual is haploid for regions of their genome, then by definition at those loci $\Phi_{xx} = 1$. Directly accounting for the ploidy of individuals is especially pertinent for X-linked loci, as sampled male individuals will be haploid and sampled females will be diploid, thereby leading to noninbred male individuals to have self-kinship coefficients mirroring completely inbred individuals ($\Phi_{xx} = 1$), whereas noninbred females at the same loci will have self-kinship coefficients consistent with noninbred diploids ($\Phi_{xx} = 1/2$). 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

$$\Phi_2(P_j) = \sum_{w=1}^{N(P_j)}\sum_{x=1}^{N(P_j)} \phi_w(P_j)\phi_x(P_j)\Phi_{wx} \tag{28}$$

$$\Phi_3(P_j) = \sum_{w=1}^{N(P_j)}\sum_{x=1}^{N(P_j)}\sum_{y=1}^{N(P_j)} \phi_w(P_j)\phi_x(P_j)\phi_y(P_j)\Phi_{wxy} \tag{29}$$

$$\Phi_4(P_j) = \sum_{w=1}^{N(P_j)}\sum_{x=1}^{N(P_j)}\sum_{y=1}^{N(P_j)}\sum_{z=1}^{N(P_j)} \phi_w(P_j)\phi_x(P_j)\phi_y(P_j)\phi_z(P_j)\Phi_{wxyz} \tag{30}$$

$$\Phi_{2,2}(P_j) = \sum_{w=1}^{N(P_j)}\sum_{x=1}^{N(P_j)}\sum_{y=1}^{N(P_j)}\sum_{z=1}^{N(P_j)} \phi_w(P_j)\phi_x(P_j)\phi_y(P_j)\phi_z(P_j)\Phi_{wx,yz}, \tag{31}$$

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. Here, Φ_{wxy}, Φ_{wxyz}, and $\Phi_{wx,yz}$ are kinship coefficients respectively defining the probabilities that a trio of alleles from

individuals w, x, and y, a quadruple of alleles from individuals w, x, y, and z, and a pair of alleles from w and x and a pair of alleles from y and z are identical by descent (Harris 1964; Cockerham 1971). In particular, Φ_{wx}, Φ_{wxy}, Φ_{wxyz}, and $\Phi_{wx,yz}$ are identical to the Cockerham (1971) coefficients denoted by θ, γ, δ, and Δ, respectively.

These Cockerham (1971) pairwise and extended kinship coefficients for measuring identity-by-descent probabilities among individuals are computed from sets of alleles sampled from two, three, or four individuals within the same population, whereas the F_2, F_3, and F_4 statistics respectively are computed from sets of alleles sampled from two, three, or four distinct populations. Harris (1964) derived the genotypic covariances between individuals within a population, which are related to these kinship coefficients. In an interesting connection, $F_2(A,B)$ measures the covariance (variance) in the frequency difference between alleles sampled from populations A and B, $F_3(A;B,C)$ the covariance in the frequency difference between alleles sampled from populations A and B and the difference between alleles sampled from populations A and C, and $F_4(A,B;C,D)$ the covariance in the frequency difference between alleles sampled from populations A and B and the difference between alleles sampled from populations C and D (Peter 2016). For this reason, we may expect that these covariances (F-statistics) will depend on the identity-by-descent probabilities defined by the Cockerham (1971) kinship coefficients, which we show is the case based on our derivations of the theoretical properties of the F-statistics.

From our definitions of kinship, we know that unrelated individuals have kinship coefficients of zero, but noninbred 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. The estimators presented in Appendix A of Patterson *et al.* (2012) correct this bias due to finite sample sizes, and our goal is to further correct for the biases induced by related and inbred individuals. 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, as well as assume that the different populations in general are independent so that alleles sampled in different populations cannot be identical by descent. Furthermore, in the cases referred to in this article, sampling is defined as statistical sampling, where the expectation is averaging over repeated sampling. 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)

$$\mathbb{E}[\hat{p}_j] = p_j \tag{32}$$

$$\mathbb{E}[\hat{p}_j^2] = p_j^2 + \Phi_2(P_j)p_j(1-p_j) \tag{33}$$

$$\mathbb{E}[\hat{p}_j^3] \approx p_j^3 + 3\Phi_2(P_j)p_j^2(1-p_j) \tag{34}$$

$$\mathbb{E}[\hat{p}_j^4] \approx p_j^4 + 6\Phi_2(P_j)p_j^3(1-p_j). \tag{35}$$

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

$$\text{Bias}[\hat{G}(P)] = -\frac{1}{J}\sum_{j=1}^{J}\Phi_2(P_j)G(P_j) \tag{36}$$

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

$$\tilde{G}(P) = \frac{1}{J}\sum_{j=1}^{J}\tilde{G}(P_j), \tag{37}$$

where

$$\tilde{G}(P_j) = \frac{\hat{G}(P_j)}{1-\Phi_2(P_j)} \tag{38}$$

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

Though this result is also given based on work in page 153 of Weir (1996), we provide the proof of Lemma 1 in the Appendix for completeness. 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. As a corollary, we next provide the bias of $\hat{G}(P)$ due to finite sample size, and use it to construct the unbiased estimator $\tilde{G}(P)$ of $G(P)$ in samples of unrelated and noninbred individuals (proof of this corollary given in the Appendix). Note that this estimator is identical to the one provided in Appendix A of Patterson *et al.* (2012).

Corollary 2. 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 $N(P_j)$ unrelated and noninbred individuals at locus j where individual $k \in \{1,2,\ldots,N(P_j)\}$ has ploidy m_k. Assuming allele frequencies are estimated using the sample proportion, the estimator $\hat{G}(P)$ has downward bias

$$\text{Bias}[\hat{G}(P)] = -\frac{1}{J}\sum_{j=1}^{J}\frac{1}{\sum_{k=1}^{N(P_j)} m_k}G(P_j) \tag{39}$$

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

$$\breve{G}(P) = \frac{1}{J}\sum_{j=1}^{J}\breve{G}(P_j), \tag{40}$$

where

$$\breve{G}(P_j) = \frac{\sum_{k=1}^{N(P_j)} m_k}{\sum_{k=1}^{N(P_j)} m_k - 1}\hat{G}(P_j) \tag{41}$$

is an unbiased estimator of $G(P_j)$. Here, $\breve{G}(P_j)$ is equivalent to the unbiased estimator termed $\hat{h}_A$ in Appendix A of Patterson *et al.* (2012).

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 the inclusion of related or inbred individuals. We present the formal result for F_2 next (Proposition 3), and prove the result in the Appendix.

Proposition 3. 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

$$\text{Bias}[\hat{F}_2(A,B)] = \frac{1}{J}\sum_{j=1}^{J}[\boldsymbol{\Phi}_2(A_j)G(A_j) + \boldsymbol{\Phi}_2(B_j)G(B_j)] \tag{42}$$

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

$$\tilde{F}_2(A,B) = \frac{1}{J}\sum_{j=1}^{J}[\hat{F}_2(A_j,B_j) - \boldsymbol{\Phi}_2(A_j)\tilde{G}(A_j) - \boldsymbol{\Phi}_2(B_j)\tilde{G}(B_j)]. \tag{43}$$

As one can see, the estimator $\hat{F}_2(A,B)$ is upwardly biased due to relatedness and inbreeding, 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. As a corollary, we next provide the bias of $\hat{F}_2(A,B)$ due to finite sample size, and use it to construct the unbiased estimator $\breve{F}_2(A,B)$ of $F_2(A,B)$ in samples of unrelated and noninbred individuals (proof of this corollary given in the Appendix). Note that this estimator is identical to the one derived in Appendix A of Patterson *et al.* (2012), and we provide an additional corollary highlighting its bias in samples containing related or inbred individuals, which we prove in the Appendix.

Corollary 4. 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 $N(P_j)$ unrelated and noninbred individuals at locus j in population $P \in \{A,B\}$ where individual $k \in \{1,2,\ldots,N(P_j)\}$ has ploidy m_k. Assuming allele frequencies are estimated using the sample proportion, the estimate $\hat{F}_2(A,B)$ has upward bias

$$\text{Bias}[\hat{F}_2(A,B)] = \frac{1}{J}\sum_{j=1}^{J}\left[\frac{1}{\sum_{k=1}^{N(A_j)} m_k}G(A_j) + \frac{1}{\sum_{k=1}^{N(B_j)} m_k}G(B_j)\right] \tag{44}$$

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

$$\breve{F}_2(A,B) = \frac{1}{J}\sum_{j=1}^{J}\left[\hat{F}_2(A_j,B_j) - \frac{1}{\sum_{k=1}^{N(A_j)} m_k - 1}\hat{G}(A_j) - \frac{1}{\sum_{k=1}^{N(B_j)} m_k - 1}\hat{G}(B_j)\right]. \tag{45}$$

Here, $\breve{F}_2(A_j,B_j)$ is equivalent to the unbiased estimator termed $\hat{F}_2(A,B)$ in Appendix A of Patterson *et al.* (2012).

Corollary 5. 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 $\breve{F}_2(A,B)$ described in Corollary 4 [also in Appendix A of Patterson *et al.* (2012)] has upward bias

$$\text{Bias}[\breve{F}_2(A,B)] = \frac{1}{J}\sum_{j=1}^{J}\left[\frac{\boldsymbol{\Phi}_2(A_j)\sum_{k=1}^{N(A_j)} m_k - 1}{\sum_{k=1}^{N(A_j)} m_k - 1}G(A_j) + \frac{\boldsymbol{\Phi}_2(B_j)\sum_{k=1}^{N(B_j)} m_k - 1}{\sum_{k=1}^{N(B_j)} m_k - 1}G(B_j)\right]. \tag{46}$$

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 6), and prove the result in the Appendix.

Proposition 6. 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

$$\text{Bias}[\hat{F}_3(A;B,C)] = \frac{1}{J}\sum_{j=1}^{J}\boldsymbol{\Phi}_2(A_j)G(A_j) \tag{47}$$

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

$$\tilde{F}_3(A;B,C) = \frac{1}{J}\sum_{j=1}^{J}[\hat{F}_3(A_j;B_j,C_j) - \boldsymbol{\Phi}_2(A_j)\tilde{G}(A_j)]. \tag{48}$$

The bias of the original estimator is proportional to the relatedness and diversity within the target population A. The new unbiased estimator $\tilde{F}_3(A;B,C)$ 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 $\tilde{G}(A)$. 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. As a corollary, we next provide the bias of $\hat{F}_3(A;B,C)$ due to finite sample size, and use it to construct the unbiased estimator $\breve{F}_3(A;B,C)$ of $F_3(A;B,C)$ in samples of unrelated and noninbred individuals

(proof of this corollary given in the Appendix). Note that this estimator is identical to the one derived in Appendix A of Patterson *et al.* (2012), and we provide an additional corollary highlighting its bias in samples containing related or inbred individuals, which we prove in the Appendix.

Corollary 7. 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 $N(P_j)$ unrelated and noninbred individuals at locus j in population $P \in \{A,B,C\}$ where individual $k \in \{1,2,\ldots,N(P_j)\}$ has ploidy m_k. Assuming allele frequencies are estimated using the sample proportion, the estimate $\hat{F}_3(A;B,C)$ has upward bias

$$\text{Bias}[\hat{F}_3(A;B,C)] = \frac{1}{J}\sum_{j=1}^{J}\frac{1}{\sum_{k=1}^{N(A_j)} m_k} G(A_j) \tag{49}$$

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

$$\breve{F}_3(A;B,C) = \frac{1}{J}\sum_{j=1}^{J}[\hat{F}_3(A_j;B_j,C_j) - \frac{1}{\sum_{k=1}^{N(A_j)} m_k - 1}\hat{G}(A_j)]. \tag{50}$$

Here, $\breve{F}_3(A_j;B_j,C_j)$ is equivalent to the unbiased estimator termed $\hat{F}_3(A;B,C)$ in Appendix A of Patterson *et al.* (2012).

Corollary 8. 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 $\breve{F}_3(A;B,C)$ described in Corollary 7 [also in Appendix A of Patterson *et al.* (2012)] has upward bias

$$\text{Bias}[\breve{F}_3(A;B,C)] = \frac{1}{J}\sum_{j=1}^{J}\frac{\Phi_2(A_j)\sum_{k=1}^{N(A_j)} m_k - 1}{\sum_{k=1}^{N(A_j)} m_k - 1} G(A_j). \tag{51}$$

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 9), and prove the result in the Appendix.

Proposition 9. 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

$$\text{Bias}[\hat{F}_3(A;B,C|A)] \approx \frac{(1/J)\sum_{j=1}^{J}\Phi_2(A_j)G(A_j)}{G(A) - (1/J)\sum_{j=1}^{J}\Phi_2(A_j)G(A_j)}\left[F_3(A;B,C|A) + \frac{1}{2}\right]. \tag{52}$$

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

$$\tilde{F}_3(A;B,C|A) = \frac{\tilde{F}_3(A;B,C)}{2\tilde{G}(A)}. \tag{53}$$

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 $\tilde{G}(A)$. As a corollary, we next provide the approximate bias of $\hat{F}_3(A;B,C|A)$ due to finite sample size, and use it to construct the approximately unbiased estimator $\breve{F}_3(A;B,C|A)$ of $F_3(A;B,C|A)$ in samples of unrelated and noninbred individuals (proof of this corollary given in the Appendix). We provide an additional corollary highlighting its bias in samples containing related or inbred individuals, which we prove in the Appendix.

Corollary 10. 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 $N(P_j)$ unrelated and noninbred individuals at locus j in population $P \in \{A,B,C\}$ where individual $k \in \{1,2,\ldots,N(P_j)\}$ has ploidy m_k. 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

$$\text{Bias}[\hat{F}_3(A;B,C|A)] \approx \frac{(1/J)\sum_{j=1}^{J}[1/\sum_{k=1}^{N(A_j)} m_k]G(A_j)}{G(A) - (1/J)\sum_{j=1}^{J}[1/\sum_{k=1}^{N(A_j)} m_k]G(A_j)}\left[F_3(A;B,C|A) + \frac{1}{2}\right]. \tag{54}$$

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

$$\breve{F}_3(A;B,C|A) = \frac{\breve{F}_3(A;B,C)}{2\breve{G}(A)}. \tag{55}$$

Corollary 11. 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 estimate $\breve{F}_3(A;B,C|A)$ described in Corollary 10 is approximately upwardly biased, assuming that its mean is well-approximated by the ratio of means of $\breve{F}_3(A;B,C)$ and $2\breve{G}(A)$ that it uses in its definition, with its upward approximate bias

$$\text{Bias}[\breve{F}_3(A;B,C|A)] \approx \frac{1}{2}\left[\frac{1}{Y(A)} - \frac{1}{G(A)}\right]F_3(A;B,C) + \frac{X(A)}{2Y(A)}, \tag{56}$$

where

$$X(A) = \frac{1}{J}\sum_{j=1}^{J}\frac{\Phi_2(A_j)\sum_{k=1}^{N(A_j)} m_k - 1}{\sum_{k=1}^{N(A_j)} m_k - 1} G(A_j) \tag{57}$$

$$\Upsilon(A) = \frac{1}{J}\sum_{j=1}^{J}\frac{[1-\Phi_2(A_j)]\sum_{k=1}^{N(A_j)} m_k}{\sum_{k=1}^{N(A_j)} m_k - 1} G(A_j). \quad (58)$$

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 12), and prove the result in the Appendix.

Proposition 12. 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 13 that the normalized $\hat{F}_4(A,B;C,D|P)$ estimator is approximately upwardly biased, and prove the result in the Appendix.

Proposition 13. 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

$$\text{Bias}[\hat{F}_4(A,B;C,D|P)] \approx \frac{(1/J)\sum_{j=1}^{J}\Phi_2(P_j)G(P_j)}{G(P) - (1/J)\sum_{j=1}^{J}\Phi_2(P_j)G(P_j)} F_4(A,B;C,D|P). \quad (59)$$

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

$$\tilde{F}_4(A,B;C,D|P) = \frac{\hat{F}_4(A,B;C,D)}{\tilde{G}(P)}. \quad (60)$$

The reasoning that the $\hat{F}_4(A,B;C,D|P)$ 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. As a corollary, we next provide the approximate bias of $\hat{F}_4(A,B;C,D|P)$ due to finite sample size, and use it to construct the approximately unbiased estimator $\breve{F}_4(A,B;C,D|P)$ of $F_4(A,B;C,D|P)$ in samples of unrelated and noninbred individuals (proof of this corollary given in the Appendix). We provide an additional corollary highlighting its bias in samples containing related or inbred individuals, which we prove in the Appendix.

Corollary 14. 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 $N(P_j)$ unrelated and noninbred individuals at locus j in population $P \in \{A,B,C,D\}$ where individual $k \in \{1,2,\ldots,N(P_j)\}$ has ploidy m_k. 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 $\hat{G}(P)$ that it uses in its definition, with its upward approximate bias

$$\text{Bias}[\hat{F}_4(A,B;C,D|P)] \approx \frac{(1/J)\sum_{j=1}^{J}[1/\sum_{k=1}^{N(P_j)} m_k]G(P_j)}{G(P) - (1/J)\sum_{j=1}^{J}[1/\sum_{k=1}^{N(P_j)} m_k]G(P_j)} F_4(A,B;C,D). \quad (61)$$

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

$$\breve{F}_4(A,B;C,D|P) = \frac{\hat{F}_4(A,B;C,D)}{\breve{G}(P)}. \quad (62)$$

Corollary 15. 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 estimate $\breve{F}_4(A,B;C,D|P)$ described in Corollary 14 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 $\breve{G}(P)$ that it uses in its definition, with its upward approximate bias

$$\text{Bias}[\breve{F}_4(A,B;C,D|P)] \approx \left[\frac{1}{\Upsilon(P)} - \frac{1}{G(P)}\right] F_4(A,B;C,D), \quad (63)$$

where

$$\Upsilon(P) = \frac{1}{J}\sum_{j=1}^{J}\frac{[1-\Phi_2(P_j)]\sum_{k=1}^{N(P_j)} m_k}{\sum_{k=1}^{N(P_j)} m_k - 1} G(P_j). \quad (64)$$

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 17 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 16, and prove the result in the Appendix.

Proposition 16. 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 (MSE). 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 $\hat{F}_2(A,B)$, $\tilde{F}_2(A,B)$, $\breve{F}_2(A,B)$, $\hat{F}_3(A;B,C)$, $\tilde{F}_3(A;B,C)$, $\breve{F}_3(A;B,C)$, $\hat{F}_3(A;B,C|A)$, $\tilde{F}_3(A;B,C|A)$, $\breve{F}_3(A;B,C|A)$, $\hat{F}_4(A,B;C,D)$, $\hat{F}_4(A,B;C,D|P)$, $\tilde{F}_4(A,B;C,D|P)$, $\breve{F}_4(A,B;C,D|P)$, and $\hat{D}(A,B,C,D)$ in Propositions 20, 21, 22, 23, 25, 26, 28, 31, 32, 27, 34, 37, 38, and 41 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 MSEs) 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 ($\mathbf{\Phi}_{xy}$). To summarize how an entire sample from a population P is related to each other at a locus, we use

$$\mathbf{\Phi}_2(P) = \sum_{w=1}^{N(P)} \sum_{x=1}^{N(P)} \phi_w(P)\phi_x(P)\mathbf{\Phi}_{wx}, \tag{65}$$

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

$$\phi_x(P) = \frac{m_x}{\sum_{k=1}^{N(P)} m_k}. \tag{66}$$

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 (McPeek *et al.* 2004; DeGiorgio *et al.* 2010; Harris and DeGiorgio, 2017a)

$$\hat{p} = \sum_{k=1}^{N(P)} \phi_k(P) X_k = \sum_{k=1}^{N(P)} \frac{m_k}{\sum_{j=1}^{N(P)} m_j} X_k. \tag{67}$$

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 MSE between the biased $\hat{F}$ estimators, the $\breve{F}$ estimators, which are unbiased in samples not containing related or inbred individuals, 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 = CEU and B = YRI. We use populations from the 1000 Genomes Project, as the released dataset of genotype calls across the 2504 worldwide samples does not include related individuals. 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 3, 20, and 21 and the difference in MSE between the $\hat{F}$ and $\breve{F}$ estimators by using Corollary 5 and Proposition 22. We 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 6, 23, 25, and 26 and Corollary 8 to determine the MSE for all three F_3 estimators by including allele frequency information from the JPT (Japanese) population where A = JPT, B = CEU, and C = 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 9, 28, 31, and 32 and Corollary 11 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 (Gujurati Indian) allele frequency data and using the derivations in Propositions 13, 34, 37, and 38 and Corollary 15. In this case, we set A = YRI, B = CEU, C = JPT, and D = 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 10 avuncular relationships, 10 inbred full siblings, and 10 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 $\log_{10}$(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. Similar trends emerge when examining the difference between $\hat{F}$ and $\breve{F}$ estimators. Specifically, the difference in $\log_{10}$(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 $\log_{10}$(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 ($\mathbf{\Phi}_{xy} = 3/8$), outbred full sibling pairs ($\mathbf{\Phi}_{xy} = 1/4$), parent-offspring

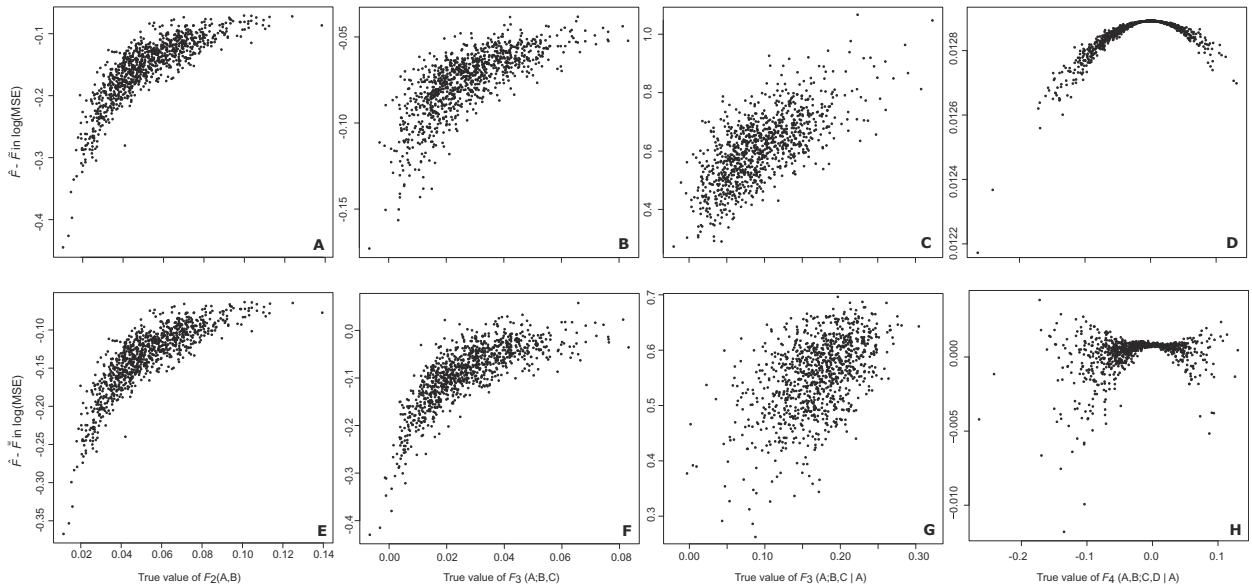

Figure 2 Difference in theoretically calculated $\log_{10}$(MSE) of $\hat{F}$ and $\tilde{F}$ (A–D) and $\hat{F}$ and $\breve{F}$ (E–H) estimators when including relatives or inbred individuals. The MSE is estimated for instances when samples of 60 individuals include individuals related to exactly one 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 = CEU and B = YRI, while for $F_3(A; B, C)$ and $F_3(A; B, C|A)$ we use A = JPT, B = CEU, and C = YRI and for $F_4(A, B; C, D|A)$ we assign A = YRI, B = CEU, C = JPT, and D = GIH.

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 MSE* section above.

In almost all cases, the biased $\hat{F}$ and the unbiased $\breve{F}$ estimators always displayed elevated MSE compared to their corresponding unbiased $\tilde{F}$ estimators (Figure 3 and Supplementary Figures S1–S3), with the $\breve{F}$ estimators always having values between the $\hat{F}$ and $\tilde{F}$ estimators, and usually being closer to $\tilde{F}$ than to $\hat{F}$. For all 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 among the biased and unbiased estimators is much smaller for normalized F_4 (Supplementary 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 (Figure 4 and Supplementary Figures 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. In addition, the values for the $\breve{F}$ estimators are also higher than the corresponding $\tilde{F}$ estimators, but are lower than respective $\hat{F}$ estimators for all combinations of individuals. Though the MSE is higher for the biased estimators, the variation in MSE values is similar for all 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 (Figure 4 and Supplementary Figures S4–S6), we notice similar patterns of MSE variation when comparing $\hat{F}$ estimators with $\breve{F}$ and $\tilde{F}$ 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

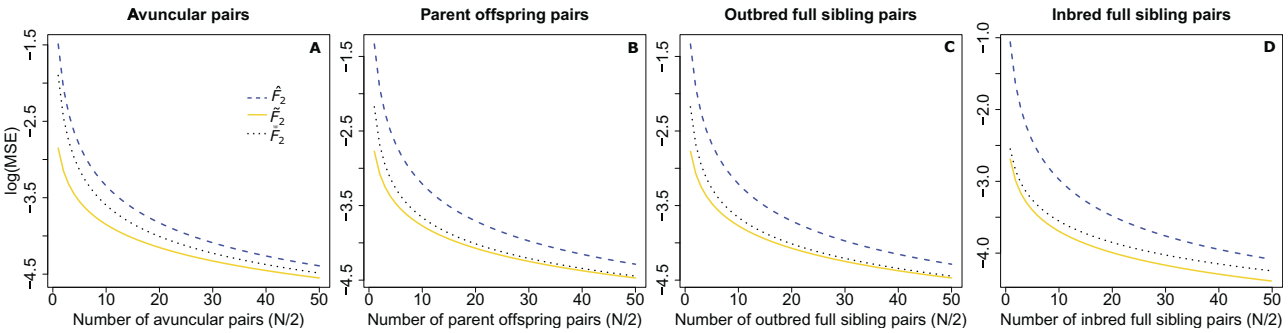

Figure 3 Mean squared error theoretically calculated for $\hat{F}_2(A,B)$, $\breve{F}_2(A,B)$, and $\tilde{F}_2(A,B)$ across different sample sizes or related pairs of individuals, including avuncular relationships (A), parent-offspring relationships (B), outbred full siblings (C), and inbred full siblings (D). 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.

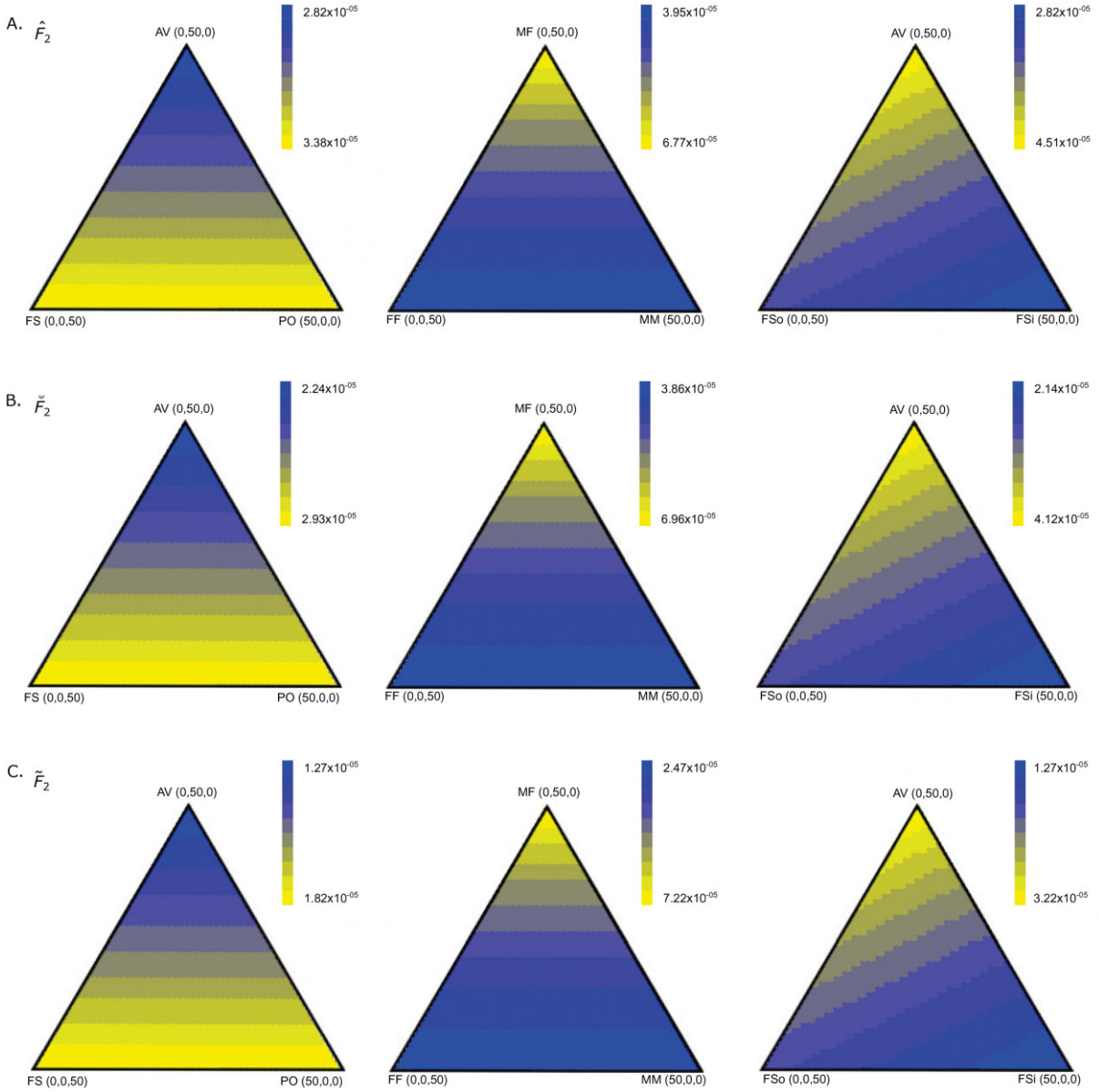

Figure 4 Theoretically calculated MSE of $\hat{F}_2(A,B)$ (A), $\breve{F}_2(A,B)$ (B), and $\tilde{F}_2(A,B)$ (C) when including relatives or inbred individuals for $J = 20$ loci. The MSE is estimated for instances when samples of 100 individuals include individuals related to exactly one other in the 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.

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 noninbred 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 (Figure 5 and Supplementary Figures S7–S9). Comparing the bias measurements in these figures, we observe a clear reduction in bias when applying the $\tilde{F}$ and $\breve{F}$ estimators as opposed to the $\hat{F}$ estimators. Importantly, the bias measurements for the $\breve{F}$ estimators are always higher than the bias measures for $\tilde{F}$ estimates, as the $\breve{F}$ estimators do not account for relatives. However, the variances are highly similar for $\hat{F}$, $\breve{F}$, 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 Supplementary Figure S9, 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 0 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 (Figure 5 and Supplementary Figures S7–S9). We compute the bias, variance, and MSE for simulations, and compare these values to theoretically calculated computations for each relative combination (Supplementary Figures S11–S14). 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 produced 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 nonnegative 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 allele frequency information from the three populations (A, B, and C) we generate 50 individuals for each population, in which there are 25 parent-offspring pairs. We then compute $\tilde{F}_3(A; B, C)$, $\breve{F}_3(A; B, C)$, 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 6 illustrates that $\tilde{F}_3$ values are lower than $\hat{F}_3$, with $\breve{F}_3$ values falling in between $\hat{F}_3$ and $\tilde{F}_3$. $\tilde{F}_3$ values are almost always negative while $\hat{F}_3$ and $\breve{F}_3$ 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 biased estimators when related individuals are included in the sample leads to a different conclusion than when using the unbiased estimator. However, we notice that there is almost no correlation between the true F_3 values and the estimates of F_3 in Figure 6. This lack of correlation is due to the fact that the F_3 values we simulated for this experiment are drawn from a particularly small range, and correlations in estimated versus true F_3 are unable to be observed over the estimation noise. To ensure that this is

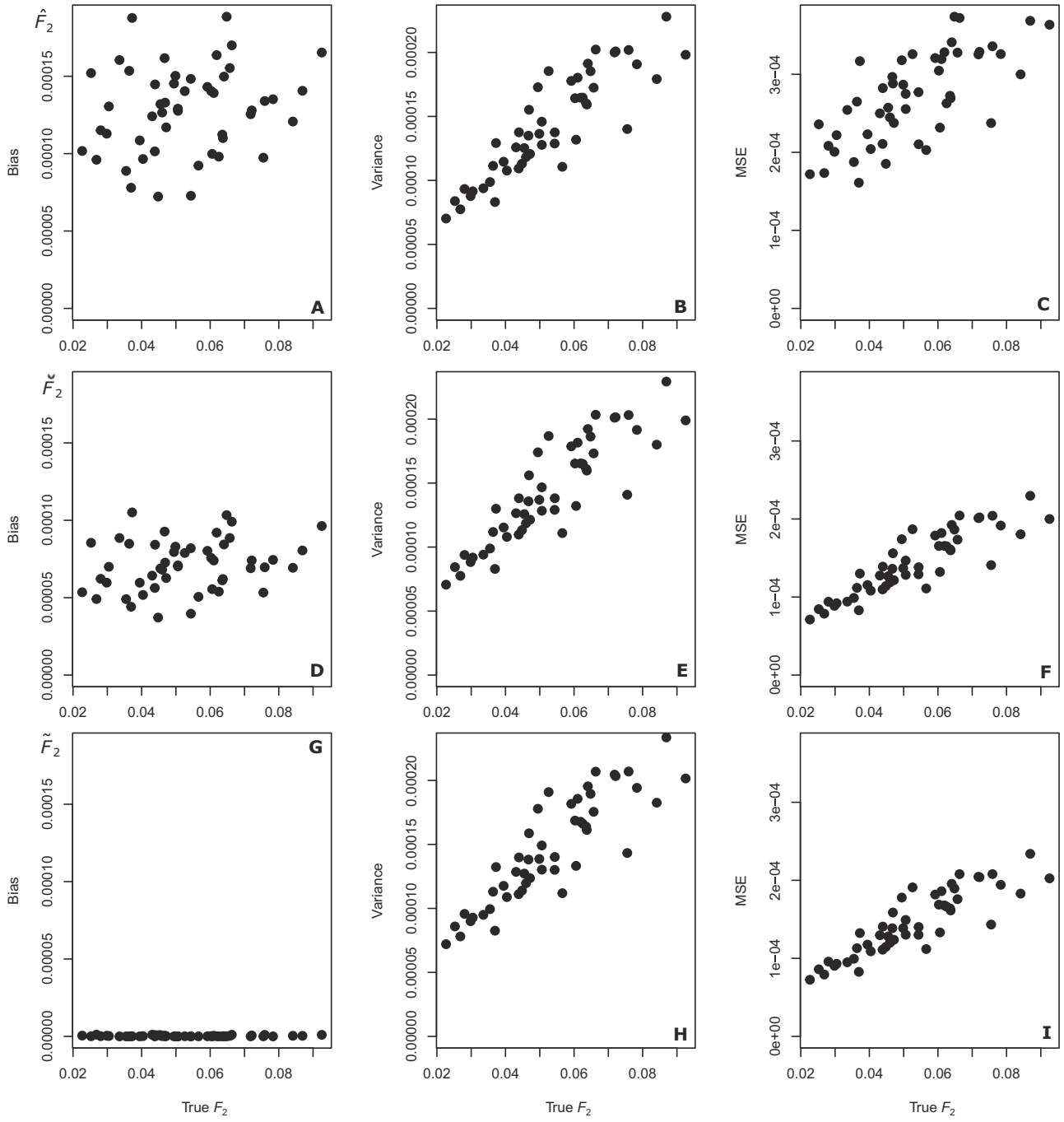

Figure 5 Comparison of squared bias (A, D, and G), variance (B, E, and H), and MSE (C, F, and I), for $\hat{F}_2(A,B)$, $\breve{F}_2(A,B)$, and $\tilde{F}_2(A,B)$ from simulated data including 60 parent offspring relative pairs. Each estimate was computed from $J = 20$ randomly sampled loci using A = CEU and B = YRI.

indeed the case, we conduct simulations across a larger range of true F_3 values and estimate $\tilde{F}_3$, $\hat{F}_3$, and $\breve{F}_3$. We notice that the expected trend appears when the range of F_3 values is expanded (Supplementary Figure S15).

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 $\hat{F}_2(A,B)$, $\tilde{F}_2(A,B)$, $\breve{F}_2(A,B)$, $\hat{F}_3(A;B,C|A)$, $\tilde{F}_3(A;B,C|A)$, $\breve{F}_3(A;B,C|A)$, $\hat{F}_4(A,B;C,D|A)$, $\tilde{F}_4(A,B;C,D|A)$, and $\breve{F}_4(A,B;C,D|A)$, and compare the mean and standard deviation of the biased and unbiased estimators (Figure 7). 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. In addition, the unbiased $\breve{F}$ estimators presented in Appendix A of Patterson *et al.* (2012) all have lower means than the biased $\hat{F}$

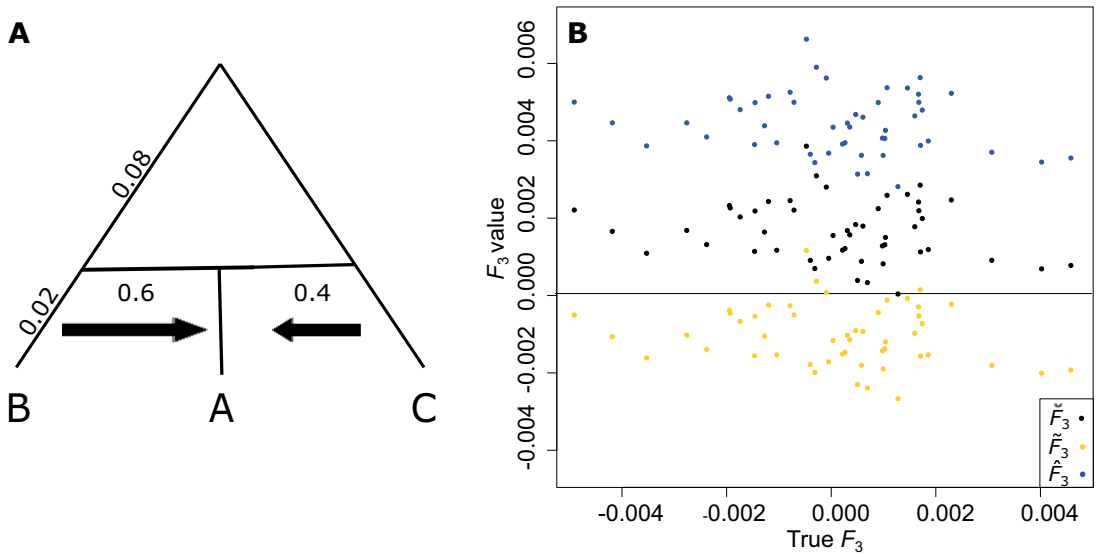

Figure 6 $\hat{F}_3(A;B,C)$, $\breve{F}_3(A;B,C)$, and $\tilde{F}_3(A;B,C)$ calculated for simulations where populations *B* and *C* merge with admixture proportions of 0.4 and 0.6, respectively, 400 generations ago (0.02 coalescent units) to form population *A* (tree shown in panel A). Panel (B) shows results for a sample containing 25 parent-offspring pairs.

estimators, and are more similar to our unbiased $\tilde{F}$ estimators, while having standard deviation measures that are a lower than both other estimators. This could indicate that Patterson's unbiased estimators have slightly higher precision, while having slightly lower accuracy than the unbiased estimators introduced in this article when samples contain related individuals.

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 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 nonhuman 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.

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 relatedness would provide estimates closer to the genomewide mean. Because the variance for these biased and unbiased estimators is approximately the same (compare Propositions 20 vs 21 and 23 vs 25), 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 6), key historical events relating populations could be missed. In addition,

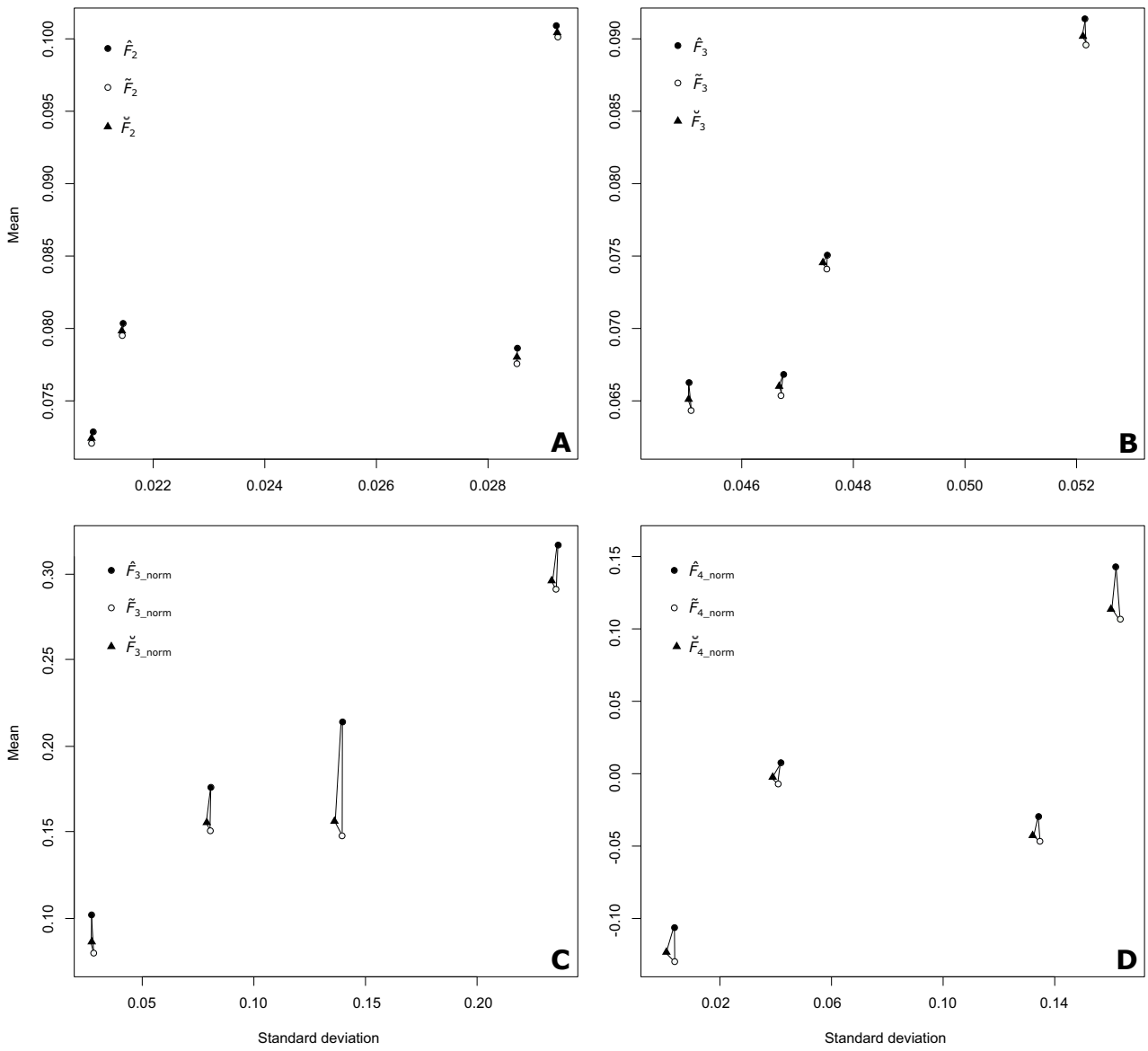

Figure 7 The difference between the means and standard deviations of biased, unbiased Patterson's, and our new unbiased estimators of F_2, normalized and un-normalized F_3 and normalized F_4 when estimated with genotype information from four different combinations of Colombian, Lahu, Melanesian, Mandenka, San, and Druze populations. All of these populations include between 2 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. Panel (A) has values for $F_2(A, B)$, panel (B) has values for $F_3(A; B, C)$, panel (C) shows results for normalized $F_3(A; B, C|A)$, and panel (D) has results for normalized $F_4(A, B; C, D|A)$.

applying these statistics to genomic regions that are likely to be evolving non-neutrally (*e.g.*, protein-coding regions) may lead to skewed estimates due to selection. For this reason, it is recommended that these statistics be applied to intergeneic regions.

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)

$$D_{\text{FO}}(A, B; C, D) = \frac{\sum_{j=1}^{J}(1 - 2a_j)(d_j - c_j)}{\sum_{j=1}^{J}(c_j + d_j - 2c_j d_j)} \tag{68}$$

$$D_{\text{IL}}(A, B; C, D) = \frac{\sum_{j=1}^{J}(1 - 2b_j)(d_j - c_j)}{\sum_{j=1}^{J}(c_j + d_j - 2c_j d_j)} \tag{69}$$

$$D_{\mathrm{FI}}(A,B;C,D) = \frac{\sum_{j=1}^{J}(1-2c_j)(b_j-a_j)}{\sum_{j=1}^{J}(a_j+b_j-2a_jb_j)} \tag{70}$$

$$D_{\mathrm{OL}}(A,B;C,D) = \frac{\sum_{j=1}^{J}(1-2d_j)(b_j-a_j)}{\sum_{j=1}^{J}(a_j+b_j-2a_jb_j)}, \tag{71}$$

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, a key assumption of our theoretical formulas is that pairwise relative contributions to the bias and variance are so much larger in magnitude than higher-order relative contributions, that the inclusion of kinship terms for trios, quadruples, or pairs of pairs of relatives would minimally affect results. For this reason, and to avoid highly unwieldy formulas for variance calculations, we made approximations to the variance formulas using this assumption. We briefly explore the accuracy of such approximations under simulations with 20 full sibling trios, and calculate the bias, variance, and MSE over a range of F_2 values. We compare these simulated results to theoretically calculated bias, variance, and MSE for $\tilde{F}_2$ across identical F_2 values (Supplementary Figure S16), where the theoretical variance (and hence MSE) formulas are approximations that only consider pairwise kinship coefficients. We see that though the trends in variance (and hence MSE) values are similar between the simulated and theoretical quantities, there is a slight difference between the simulated and theoretical variance (and hence MSE), where the theoretical values are consistently lower than the simulated. Therefore, our theoretical approximate variance calculations have underestimated the true variance values, which can be expected as we drop a number of terms that would be in the exact variance calculation, while still being useful for understanding the overall trends of the variances (and hence MSEs) across estimators.

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 McPeek *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 (McPeek *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

$$\phi_x(P) = \frac{\sum_{k=1}^{N(P)}(\mathbf{K}^{-1})_{kx}}{1^{\mathrm{T}}\mathbf{K}^{-1}1}, \tag{72}$$

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 $\mathbf{K}_{jk} = \Phi_{jk}$, $1 \in \mathbb{R}^{N(P)}$ 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/Mehreen Ruhi/funbiased.

Data availability

Supplementary data are available at Genetics online. The 1000 Genomes Project data used in this publication is available at http://www.1000genomes.org/. Relatedness information used to generate Figure 7 is available online within Supplementary Tables S7–S15 of Rosenberg (2006).

Acknowledgments

The authors thank three anonymous reviewers for their comments that improved our article. They also thank Nick Patterson for pointing us to the unbiased F-statistics in samples containing unrelated and noninbred individuals within Appendix A of Patterson *et al.* (2012). The authors thank George (PJ) Perry for the helpful comments on an early draft of this article and Alexandre Harris for fruitful discussions about our simulation protocol.

Funding

This research was funded by the National Institutes of Health (R35GM128590), the National Science Foundation (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).

Conflicts of interest

The authors declare that there is no conflict of interest.

Literature cited

Cockerham CC. 1971. Higher order probability functions of identity of allelles by descent. Genetics. 69:235–246.

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

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

Eaton DAR, Ree RH. 2013. Inferring phylogeny and introgression using RADseq data: an example from flowering plants (Pedicularis: Orobanchaceae). Syst Biol. 62:689–706.

Epstein MP, Duren WL, Boehnke M. 2000. Improved inference of relationship for pairs of individuals. Am J Hum Genet. 67: 1219–1231.

Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, *et al.*; The 1000 Genomes Project. 2011. Demographic history and rare allele sharing among human populations. Proc Natl Acad Sci USA. 108: 11983–11988.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, *et al.* 2010. A draft sequence of the neandertal genome. Science. 328:710–722.

Hajdinjak M, Fu Q, Hübner A, Petr M, Mafessoni F, *et al.* 2018. Reconstructing the genetic history of late neanderthals. Nature. 555:652–656.

Haller BC, Messer PW. 2019. 3: forward genetic simulations beyond the wright–fisher model. Mol Biol Evol. 36:632–637.

Harris AM, DeGiorgio M. 2017a. An unbiased estimator of gene diversity with improved variance for samples containing related and inbred individuals of any ploidy. G3 (Bethesda). 7:671–691.

Harris AM, DeGiorgio M. 2017b. Admixture and ancestry inference from ancient and modern samples through measures of population genetic drift. Hum Biol. 89:21–46.

Harris DL. 1964. Genotypic covariances between inbred relatives. Genetics. 50:1319–1348.

Huson D, Klöpper T, Lockhart P, Steel M. 2005. Reconstruction of reticulate networks from gene trees. *RECOMB*. 2005. 3500:233–249.

Kim HL, Ratan A, Perry GH, Montenegro A, Miller W, *et al.* 2014. Khoisan hunter-gatherers have been the largest population throughout most of modern-human demographic history. Nat Commun. 5:5692–5692.

Kulathinal RJ, Stevison LS, Noor MAF. 2009. The genomics of speciation in drosophila: diversity, divergence, and introgression estimated using low-coverage genome sequencing. PLoS Genet. 5: e1000550.

Lange K. 2002. Mathematical and Statistical Methods for Genetic Analysis. New York, NY:Springer.

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, *et al.* 2008. Worldwide human relationships inferred from genome-wide patterns of variation. Science. 319:1100–1104.

Martin SH, Davey JW, Jiggins CD. 2015. Evaluating the use of ABBA–BABA statistics to locate introgressed loci. Mol Biol Evol. 32:244–257.

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

Molinaro L, Montinaro F, Yelmen B, Marnetto D, Behar DM, *et al.* 2019. West Asian sources of the Eurasian component in Ethiopians: a reassessment. Sci Rep. 9:18811.

Moorjani P, Thangaraj K, Patterson N, Lipson M, Loh P-R, *et al.* 2013. Genetic evidence for recent population mixture in India. Am J Hum Genet. 93:422–438.

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

Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, *et al.* 2012. Ancient admixture in human history. Genetics. 192:1065–1093.

Payseur BA, Nachman MW. 2000. Microsatellite variation and recombination rate in the human genome. Genetics. 156: 1285–1298.

Pease JB, Hahn MW. 2015. Detection and polarization of introgression in a five-taxon phylogeny. Syst Biol. 64:651–662.

Peter BM. 2016. Admixture, population structure, and f-statistics. Genetics. 202:1485–1501.

Reich D, Patterson N, Campbell D, Tandon A, Mazieres S, *et al.* 2012. Reconstructing native American population history. Nature. 488: 370–374.

Reich D, Thangaraj K, Patterson N, Price AL, Singh L. 2009. Reconstructing Indian population history. Nature. 461:489–494.

Rosenberg NA. 2006. Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. Ann Hum Genet. 70:841–847.

Scally A, Durbin R. 2012. Revising the human mutation rate: implications for understanding human evolution. Nat Rev Genet. 13: 745–753.

Soraggi S, Wiuf C, Albrechtsen A. 2018. Powerful inference with the *D*-statistic on low-coverage whole-genome data. G3 (Bethesda). 8: 551–566.

Takahata N. 1993. Allelic genealogy and human evolution. Mol Biol Evol. 10:2–22.

The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. Nature. 526:68–74.

Turissini DA, Matute DR. 2017. Fine scale mapping of genomic introgressions within the Drosophila yakuba clade. PLoS Genet. 13: e1006971.

Waples RS, Anderson EC. 2017. Purging putative siblings from population genetic data sets: a cautionary view. Mol Ecol. 26: 1211–1224.

Weir B. 1996. Genetic Data Analysis II. Sunderland, Massachusetts: Sinauer Associates.

Weir BS. 1989. Sampling properties of gene diversity. In: Plant Population Genetics, Breeding and Genetic Resources. p. 23–42.

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

Wolter KM. 2007. Introduction to Variance Estimation. 2nd ed. New York, NY: Springer.

Zheng Y, Janke A. 2018. Gene flow analysis method, the *D*-statistic, is robust in a wide parameter space. BMC Bioinformatics. 19:10.

Communicating editor: S. Browning

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 (result in Weir 1996, page 153)

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

which gives

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

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

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

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

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

Proof of Corollary 2. When using the sample proportion to estimate allele frequencies, the weight of individual x in population P at locus j is (Harris and DeGiorgio 2017a)

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

Plugging into the definition of $\Phi_2(P_j)$, we have

$$\begin{aligned}\Phi_2(P_j) &= \sum_{w=1}^{N(P_j)}\sum_{x=1}^{N(P_j)}\phi_w(P_j)\phi_x(P_j)\Phi_{wx} \\ &= \frac{1}{\left(\sum_{k=1}^{N(P_j)} m_k\right)^2}\sum_{w=1}^{N(P_j)}\sum_{x=1}^{N(P_j)} m_w m_x \Phi_{wx}.\end{aligned}$$

Because the sample contains unrelated individuals, then $\Phi_{wx} = 0$ for $w \neq x$ and $\Phi_{xx} = 1/m_x$ resulting in

$$\Phi_2(P_j) = \frac{1}{\left(\sum_{k=1}^{N(P_j)} m_k\right)^2} \sum_{w=1}^{N(P_j)} m_w^2 \frac{1}{m_w} = \frac{1}{\sum_{k=1}^{N(P_j)} m_k}.$$

Plugging $\Phi_2(P_j)$ into the definitions of $\text{Bias}[\hat{G}(P)]$ and $\tilde{G}(P_j)$ within Lemma 1 gives the desired result. □

Proof of Proposition 3. We first calculate

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

which gives

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

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

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

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

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

Proof of Corollary 4. Plugging the definition of $\mathbf{\Phi}_2(P_j)$, $P \in \{A, B\}$, derived in the proof of Corollary 2 as well as

$$\breve{G}(P_j) = \frac{\sum_{k=1}^{N(P_j)} m_k}{\sum_{k=1}^{N(P_j)} m_k - 1} \hat{G}(P_j)$$

of Corollary 2 into the definitions of $\text{Bias}\left[\hat{F}_2(A,B)\right]$ and $\tilde{F}_2(A,B)$ within Proposition 3 gives the desired result. □

Proof of Corollary 5. From Corollary 4, we have

$$\breve{F}_2(A,B) = \frac{1}{J}\sum_{j=1}^{J}\left[\hat{F}_2(A_j,B_j) - \frac{1}{\sum_{k=1}^{N(A_j)} m_k - 1}\hat{G}(A_j) - \frac{1}{\sum_{k=1}^{N(B_j)} m_k - 1}\hat{G}(B_j)\right].$$

Because populations A or B may have related or inbred individuals within them, from the proofs of Lemma 1 and Proposition 3, we have

$$\mathbb{E}\left[\hat{F}_2(A_j,B_j)\right] = F_2(A_j,B_j) + \mathbf{\Phi}_2(A_j)G(A_j) + \mathbf{\Phi}_2(B_j)G(B_j)$$
$$\mathbb{E}\left[\hat{G}(P_j)\right] = [1 - \mathbf{\Phi}_2(P_j)]G(P_j)$$

yielding

$$\begin{aligned}\mathbb{E}\left[\breve{F}_2(A,B)\right] &= \frac{1}{J}\sum_{j=1}^{J}\left[F_2(A_j,B_j) + \mathbf{\Phi}_2(A_j)G(A_j) + \mathbf{\Phi}_2(B_j)G(B_j) - \frac{1-\mathbf{\Phi}_2(A_j)}{\sum_{k=1}^{N(A_j)} m_k - 1}G(A_j) - \frac{1-\mathbf{\Phi}_2(B_j)}{\sum_{k=1}^{N(B_j)} m_k - 1}G(B_j)\right] \\ &= \frac{1}{J}\sum_{j=1}^{J}\left[F_2(A_j,B_j) + \frac{\mathbf{\Phi}_2(A_j)\sum_{k=1}^{N(A_j)} m_k - 1}{\sum_{k=1}^{N(A_j)} m_k - 1}G(A_j) + \frac{\mathbf{\Phi}_2(B_j)\sum_{k=1}^{N(B_j)} m_k - 1}{\sum_{k=1}^{N(B_j)} m_k - 1}G(B_j)\right] \\ &= F_2(A,B) + \frac{1}{J}\sum_{j=1}^{J}\left[\frac{\mathbf{\Phi}_2(A_j)\sum_{k=1}^{N(A_j)} m_k - 1}{\sum_{k=1}^{N(A_j)} m_k - 1}G(A_j) + \frac{\mathbf{\Phi}_2(B_j)\sum_{k=1}^{N(B_j)} m_k - 1}{\sum_{k=1}^{N(B_j)} m_k - 1}G(B_j)\right].\end{aligned}$$

It follows that the bias is

$$\begin{aligned}\text{Bias}\left[\breve{F}_2(A,B)\right] &= \mathbb{E}\left[\tilde{F}_2(A,B)\right] - F_2(A,B) \\ &= \frac{1}{J}\sum_{j=1}^{J}\left[\frac{\mathbf{\Phi}_2(A_j)\sum_{k=1}^{N(A_j)} m_k - 1}{\sum_{k=1}^{N(A_j)} m_k - 1}G(A_j) + \frac{\mathbf{\Phi}_2(B_j)\sum_{k=1}^{N(B_j)} m_k - 1}{\sum_{k=1}^{N(B_j)} m_k - 1}G(B_j)\right],\end{aligned}$$

and because $\mathbf{\Phi}_2(P_j) \geq 1/\sum_{k=1}^{N(P_j)} m_k$, this is an upward bias. □

Proof of Proposition 6. We first calculate

$$\begin{aligned}\mathbb{E}\left[\hat{F}_3(A_j;B_j,C_j)\right] &= \mathbb{E}\left[\left(\hat{a}_j-\hat{b}_j\right)\left(\hat{a}_j-\hat{c}_j\right)\right]\\ &= \mathbb{E}\left[\hat{a}_j^2\right]-\mathbb{E}[\hat{a}_j]\mathbb{E}[\hat{c}_j]-\mathbb{E}[\hat{a}_j]\mathbb{E}\left[\hat{b}_j\right]+\mathbb{E}\left[\hat{b}_j\right]\mathbb{E}[\hat{c}_j]\\ &= \left[1-\Phi_2(A_j)\right]a_j^2+\Phi_2(A_j)a_j-a_jc_j-a_jb_j+b_jc_j\\ &= \left(a_j^2-a_jc_j-a_jb_j+b_jc_j\right)+\Phi_2(A_j)a_j(1-a_j)\\ &= F_3(A_j;B_j,C_j)+\Phi_2(A_j)G(A_j),\end{aligned}$$

which gives

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

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

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

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

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

Proof of Corollary 7. Plugging the definition of $\Phi_2(A_j)$ derived in the proof of Corollary 2 as well as

$$\breve{G}(A_j)=\frac{\sum_{k=1}^{N(A_j)}m_k}{\sum_{k=1}^{N(A_j)}m_k-1}\hat{G}(A_j)$$

of Corollary 2 into the definitions of $\text{Bias}\left[\hat{F}_3(A;B,C)\right]$ and $\tilde{F}_3(A;B,C)$ within Proposition 6 gives the desired result. □

Proof of Corollary 8. From Corollary 7, we have

$$\breve{F}_3(A;B,C)=\frac{1}{J}\sum_{j=1}^{J}\left[\hat{F}_3(A_j;B_j,C_j)-\frac{1}{\sum_{k=1}^{N(A_j)}m_k-1}\hat{G}(A_j)\right].$$

Because population A may have related or inbred individuals within it, from the proofs of Lemma 1 and Proposition 6, we have

$$\mathbb{E}\left[\hat{F}_3(A_j;B_j,C_j)\right] = F_3(A_j;B_j,C_j) + \Phi_2(A_j)G(A_j)$$
$$\mathbb{E}\left[\hat{G}(A_j)\right] = [1-\Phi_2(A_j)]G(A_j)$$

yielding

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

It follows that the bias is

$$\begin{aligned}\text{Bias}\left[\breve{F}_3(A;B,C)\right] &= \mathbb{E}\left[\breve{F}_3(A;B,C)\right] - F_3(A;B,C) \\ &= \frac{1}{J}\sum_{j=1}^{J}\frac{\Phi_2(A_j)\sum_{k=1}^{N(A_j)} m_k - 1}{\sum_{k=1}^{N(A_j)} m_k - 1}G(A_j),\end{aligned}$$

and because $\Phi_2(A_j) \geq 1/\sum_{k=1}^{N(A_j)} m_k$, this is an upward bias. □

Proof of Proposition 9. 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

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

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

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

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

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

Proof of Corollary 10. Plugging the definition of $\Phi_2(A_j)$ derived in the proof of Corollary 2 as well as

$$\breve{G}(A_j) = \frac{\sum_{k=1}^{N(A_j)} m_k}{\sum_{k=1}^{N(A_j)} m_k - 1}\hat{G}(A_j)$$

of Corollary 2 into the definitions of $\text{Bias}\left[\hat{F}_3(A;B,C|A)\right]$ and $\tilde{F}_3(A;B,C|A)$ within Proposition 9 gives the desired result. □

Proof of Corollary 11. From Corollaries 2, 7, and 10, we have

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

where

$$\begin{aligned}\breve{G}(A) &= \frac{1}{J}\sum_{j=1}^{J}\breve{G}(A_j) = \frac{1}{J}\sum_{j=1}^{J}\frac{\sum_{k=1}^{N(A_j)} m_k}{\sum_{k=1}^{N(A_j)} m_k - 1}\hat{G}(A_j)\\ \breve{F}_3(A;B,C) &= \frac{1}{J}\sum_{j=1}^{J}\left[\hat{F}_3(A_j;B_j,C_j) - \frac{1}{\sum_{k=1}^{N(A_j)} m_k - 1}\hat{G}(A_j)\right].\end{aligned}$$

Because population A may have related or inbred individuals within it, from the proofs of Lemma 1 and Proposition 6, we have

$$\mathbb{E}\left[\hat{F}_3(A_j;B_j,C_j)\right] = F_3(A_j;B_j,C_j) + \Phi_2(A_j)G(A_j)$$
$$\mathbb{E}\left[\hat{G}(A_j)\right] = [1-\Phi_2(A_j)]G(A_j)$$

yielding

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

It follows that the approximate bias is

$$\begin{aligned}\text{Bias}\left[\breve{F}_3(A;B,C|A)\right] &= \mathbb{E}\left[\breve{F}_3(A;B,C|A)\right] - F_3(A;B,C|A) \\ &\approx \frac{1}{2}\left[\frac{1}{Y(A)} - \frac{1}{G(A)}\right]F_3(A;B,C) + \frac{X(A)}{2Y(A)},\end{aligned}$$

and because $\Phi_2(A_j) \geq 1/\sum_{k=1}^{N(A_j)} m_k$, then $0 < Y(A) \leq G(A)$ and $X(A) \geq 0$ making this is an upward approximate bias. □

Proof of Proposition 12. We first calculate

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

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

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

Proof of Proposition 13. Assuming that the expectation of $\hat{F}_4(A,B;C,D|P)$ is approximately equal to the ratio of expectations of $\hat{F}_4(A,B;C,D)$ and $\hat{G}(P)$ for some $P \in \{A,B,C,D\}$, we find that

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

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

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

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

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

Proof of Corollary 14. Plugging the definition of $\Phi_2(P_j)$, $P \in \{A,B,C,D\}$, derived in the proof of Corollary 2 as well as

$$\breve{G}(P_j) = \frac{\sum_{k=1}^{N(A_j)} m_k}{\sum_{k=1}^{N(A_j)} m_k - 1}\hat{G}(P_j)$$

of Corollary 2 into the definitions of Bias$\left[\hat{F}_4(A,B;C,D|P)\right]$ and $\tilde{F}_4(A,B;C,D|P)$ within Proposition 13 gives the desired result. □

Proof of Corollary 15. From Corollaries 2 and 14, we have

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

where

$$\breve{G}(P) = \frac{1}{J}\sum_{j=1}^{J}\breve{G}(P_j) = \frac{1}{J}\sum_{j=1}^{J}\frac{\sum_{k=1}^{N(P_j)} m_k}{\sum_{k=1}^{N(P_j)} m_k - 1}\hat{G}(P_j)$$

$$\hat{F}_4(A,B;C,D) = \frac{1}{J}\sum_{j=1}^{J}\hat{F}_4(A_j,B_j;C_j,D_j).$$

Because population P may have related or inbred individuals within it, from the proofs of Lemma 1 and Proposition 12, we have

$$\mathbb{E}[\hat{F}_4(A,B;C,D] = F_4(A,B;B,D)$$
$$\mathbb{E}[\hat{G}(P_j)] = [1-\Phi_2(P_j)]G(P_j)$$

yielding

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

It follows that the approximate bias is

$$\begin{aligned}\text{Bias}\left[\breve{F}_4(A,B;C,D|P)\right] &= \mathbb{E}\left[\breve{F}_4(A,B;C,D|P)\right] - F_4(A,B;C,D|P)\\ &\approx \left[\frac{1}{Y(P)} - \frac{1}{G(P)}\right]F_4(A,B;C,D),\end{aligned}$$

and because $\Phi_2(P_j) \geq 1/\sum_{k=1}^{N(P_j)} m_k$, then $0 < Y(P) \leq G(P)$ making this is an upward approximate bias. □

Lemma 17. 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

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

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

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

Proof of Proposition 16. 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

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

Lemma 18. 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, where individual $k \in \{1,2,\ldots,N(P_j)\}$ has ploidy m_k. 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 $\tilde{G}(P)$ and the estimator $\breve{G}(P)$ are

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

$$Var\left[\breve{G}(P)\right] \approx \frac{1}{J^2}\sum_{j=1}^{J}\Phi_2(P_j)\left[\frac{\sum_{k=1}^{N(P_j)}m_k}{\sum_{k=1}^{N(P_j)}m_k - 1}\right]^2 G(P_j) - \frac{4}{J^2}\sum_{j=1}^{J}\Phi_2(P_j)\left[\frac{\sum_{k=1}^{N(P_j)}m_k}{\sum_{k=1}^{N(P_j)}m_k - 1}\right]^2 G(P_j)^2.$$

Proof. From the proof of Lemma 1, we have

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

and we calculate

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

Therefore, we have that

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

which gives

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

Recall that

$$\begin{aligned}
\tilde{G}(P) &= \frac{1}{J}\sum_{j=1}^{J}\tilde{G}(P_j) \\
\breve{G}(P) &= \frac{1}{J}\sum_{j=1}^{J}\breve{G}(P_j),
\end{aligned}$$

where

$$\begin{aligned}
\tilde{G}(P_j) &= \frac{1}{1-\Phi_2(P_j)}\hat{G}(P_j) \\
\breve{G}(P_j) &= \frac{\sum_{k=1}^{N(P_j)} m_k}{\sum_{k=1}^{N(P_j)} m_k - 1}\hat{G}(P_j).
\end{aligned}$$

It follows that

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

and similarly

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

Lemma 19. 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

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

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

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

and

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

yielding

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

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

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

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

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

Therefore, we have that

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

which gives

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

Proposition 20. 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 $\hat{F}_2(A,B)$ has approximate variance

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

Proof. From the proof of Proposition 3, we have

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

which gives

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

where we used the fact that $\Phi_2(A_j)^2$, $\Phi_2(B_j)$, 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

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

Therefore, we have that

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

which gives

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

Proposition 21. 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 $\tilde{F}_2(A,B)$ has approximate variance

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

Proof. Recall that

$$\tilde{F}_2(A_j,B_j) = \hat{F}_2(A_j,B_j) - \Phi_2(A_j)\tilde{G}(A_j) - \Phi_2(B_j)\tilde{G}(B_j),$$

where $\tilde{F}_2(A_j,B_j)$ is an unbiased estimator for $F_2(A_j,B_j)$ and $\tilde{G}(P_j)$ is an unbiased estimator of $G(P_j)$ for $P \in \{A,B\}$ at locus $j \in \{1,2,\ldots,J\}$. Also, from the proof of Proposition 3, we have

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

Therefore, we have that

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

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 $Cov\left[\tilde{G}(A_j), \tilde{G}(B_j)\right] = 0$ 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

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

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

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

and it follows that

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

Proposition 22. 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 $\Phi_3(P_j)$, $\Phi_4(P_j)$, $\Phi_{2,2}(P_j)$, $\Phi_2(P_j)^2$, $\Phi_2(A_j)\Phi_2(B_j)$, and other terms of similar order negligible to $\Phi_2(P_j)$. Based on this simplifying assumption, the estimator $\breve{F}_2(A, B)$ has approximate variance

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

Proof. Recall that

$$\breve{F}_2(A_j, B_j) = \hat{F}_2(A_j, B_j) - \frac{1}{\sum_{k=1}^{N(A_j)} m_k - 1}\hat{G}(A_j) - \frac{1}{\sum_{k=1}^{N(A_j)} m_k - 1}\hat{G}(A_j),$$

where $\breve{F}_2(A_j, B_j)$ and $\hat{F}_2(A_j, B_j)$ are estimators for $F_2(A_j, B_j)$ and $\hat{G}(P_j)$ is an estimator of $G(P_j)$ for $P \in \{A, B\}$ at locus $j \in \{1, 2, \ldots, J\}$. Also, from the proof of Proposition 3, we have

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

Therefore, we have that

$$\begin{aligned}
Var\left[\breve{F}_2(A_j,B_j)\right] &= Var\left[\hat{F}_2(A_j,B_j) - \frac{1}{\sum_{k=1}^{N(A_j)} m_k - 1}\hat{G}(A_j) - \frac{1}{\sum_{k=1}^{N(A_j)} m_k - 1}\hat{G}(A_j)\right] \\
&= Var\left[\hat{F}_2(A_j,B_j)\right] + \left[\frac{1}{\sum_{k=1}^{N(A_j)} m_k - 1}\right]^2 Var\left[\hat{G}(A_j)\right] + \left[\frac{1}{\sum_{k=1}^{N(B_j)} m_k - 1}\right]^2 Var\left[\hat{G}(B_j)\right] \\
&-2\left[\frac{1}{\sum_{k=1}^{N(A_j)} m_k - 1}\right] Cov\left[\hat{F}_2(A_j,B_j),\hat{G}(A_j)\right] - 2\left[\frac{1}{\sum_{k=1}^{N(B_j)} m_k - 1}\right] Cov\left[\hat{F}_2(A_j,B_j),\hat{G}(B_j)\right] \\
&+2\left[\frac{1}{\sum_{k=1}^{N(A_j)} m_k - 1}\right]\left[\frac{1}{\sum_{k=1}^{N(B_j)} m_k - 1}\right] Cov\left[\hat{G}(A_j),\hat{G}(B_j)\right] \\
&\approx Var\left[\hat{F}_2(A_j,B_j)\right] - 2\left[\frac{1}{\sum_{k=1}^{N(A_j)} m_k - 1}\right] Cov\left[\hat{F}_2(A_j,B_j),\hat{G}(A_j)\right] \\
&-2\left[\frac{1}{\sum_{k=1}^{N(B_j)} m_k - 1}\right] Cov\left[\hat{F}_2(A_j,B_j),\hat{G}(B_j)\right],
\end{aligned}$$

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 hence $1/\left[\sum_{k=1}^{N(P_j)} m_k - 1\right]$ assuming $N(P_j)$ is large enough), and where $Cov\left[\hat{G}(A_j),\hat{G}(B_j)\right] = 0$ because drawing alleles in population A is independent of population B. Moreover, because

$$\begin{aligned}
Cov\left[\hat{F}_2(A_j,B_j),\hat{G}(A_j)\right] &\approx \Phi_2(A_j)F_2(A_j,B_j)G(A_j) \\
Cov\left[\hat{F}_2(A_j,B_j),\hat{G}(B_j)\right] &\approx \Phi_2(B_j)F_2(A_j,B_j)G(B_j)
\end{aligned}$$

we have

$$\begin{aligned}
\left[\frac{1}{\sum_{k=1}^{N(A_j)} m_k - 1}\right] Cov\left[\hat{F}_2(A_j,B_j),\hat{G}(A_j)\right] &\approx 0 \\
\left[\frac{1}{\sum_{k=1}^{N(B_j)} m_k - 1}\right] Cov\left[\hat{F}_2(A_j,B_j),\hat{G}(B_j)\right] &\approx 0
\end{aligned}$$

by recalling the assumption that $\Phi_2(A_j)^2$, $\Phi_2(B_j)^2$, or any other term of similar magnitude (such as the $\Phi_2(P_j)/\left[\sum_{k=1}^{N(P_j)} m_k - 1\right]$ terms that appear for populations $P \in \{A, B\}$) and $\Phi_2(A_j)\Phi_2(B_j)$ are negligible compared to $\Phi_2(A_j)$ and $\Phi_2(B_j)$. Because of this, we have

$$Var\left[\breve{F}_2(A_j,B_j)\right] \approx Var\left[\hat{F}_2(A_j,B_j)\right],$$

and it follows that

$$Var\left[\breve{F}_2(A,B)\right] \approx Var\left[\hat{F}_2(A,B)\right]. \quad \square$$

Proposition 23. 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

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

Proof. From the proof of Proposition 6, we have

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

which gives

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

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

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

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

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

which gives

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

Lemma 24. 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 $\hat{F}_3(A;B,C)$ and $\hat{g}(A)$ have an approximate covariance

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

Proof. From the proofs of Lemma 1 and Proposition 6, we have

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

and

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

yielding

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

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

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

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

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

Therefore, we have that

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

which gives

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

Proposition 25. 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 $\tilde{F}_3(A;B,C)$ has approximate variance

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

Proof. Recall that

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

where $\tilde{F}_3(A_j;B_j,C_j)$ is an unbiased estimator for $F_3(A_j;B_j,C_j)$ and $\tilde{G}(A_j)=\hat{G}(A_j)/[1-\Phi_2(A_j)]$ is an unbiased estimator of $G(A_j)$ at locus $j \in \{1,2,\ldots,J\}$. Also, from the proof of Proposition 6, we have

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

Therefore, we have that

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

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

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

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

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

Proposition 26. 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 $\Phi_3(P_j)$, $\Phi_4(P_j)$, $\Phi_{2,2}(P_j)$, $\Phi_2(P_j)^2$, $\Phi_2(A_j)\Phi_2(B_j)$, $\Phi_2(A_j)\Phi_2(C_j)$, $\Phi_2(B_j)\Phi_2(C_j)$ and any other terms of similar order negligible to $\Phi_2(P_j)$. Based on this simplifying assumption, the estimator $\breve{F}_3(A;B,C)$ has approximate variance

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

Proof. Recall that

$$\breve{F}_3(A_j;B_j,C_j) = \hat{F}_3(A_j;B_j,C_j) - \frac{1}{\sum_{k=1}^{N(A_j)} m_k - 1}\hat{G}(A_j),$$

where $\breve{F}_3(A_j;B_j,C_j)$ and $\hat{F}_3(A_j;B_j,C_j)$ are estimators for $F_3(A_j;B_j,C_j)$ and $\hat{G}(A_j)$ is an estimator of $G(A_j)$ at locus $j \in \{1,2,\ldots,J\}$. Also, from the proof of Proposition 6, we have

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

Therefore, we have that

$$\begin{aligned}
Var\left[\breve{F}_3(A_j;B_j;C_j)\right] &= Var\left[\hat{F}_3(A_j;B_j,C_j) - \frac{1}{\sum_{k=1}^{N(A_j)} m_k - 1}\hat{G}(A_j)\right] \\
&= Var\left[\hat{F}_3(A_j;B_j,C_j)\right] + \left[\frac{1}{\sum_{k=1}^{N(A_j)} m_k - 1}\right]^2 Var\left[\hat{G}(A_j)\right] \\
&\quad -2\left[\frac{1}{\sum_{k=1}^{N(A_j)} m_k - 1}\right] Cov\left[\hat{F}_3(A_j;B_j,C_j),\hat{G}(A_j)\right] \\
&\approx Var\left[\hat{F}_3(A_j;B_j,C_j)\right] - \left[\frac{2}{\sum_{k=1}^{N(A_j)} m_k - 1}\right] Cov\left[\hat{F}_3(A_j;B_j,C_j),\hat{G}(A_j)\right],
\end{aligned}$$

where we used the fact that $\Phi_2(A_j)^2$ is negligible compared to $\Phi_2(A_j)$ as an approximation (and hence $1/\left[\sum_{k=1}^{N(A_j)} m_k - 1\right]$ assuming $N(A_j)$ is large enough). Moreover, because

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

we have

$$\left[\frac{2}{\sum_{k=1}^{N(A_j)} m_k - 1}\right] Cov\left[\hat{F}_3(A_j;B_j,C_j),\hat{G}(A_j)\right] \approx 0$$

by recalling the assumption that $\Phi_2(A_j)^2$ or any other term of similar magnitude (such as the $\Phi_2(A_j)/\left[\sum_{k=1}^{N(A_j)} m_k - 1\right]$ term) are negligible compared to $\Phi_2(A_j)$. Because of this, we have

$$Var\left[\breve{F}_3(A_j;B_j,C_j)\right] \approx Var\left[\hat{F}_3(A_j;B_j,C_j)\right]$$

and it follows that

$$Var\left[\breve{F}_3(A;B,C)\right] \approx Var\left[\hat{F}_3(A;B,C)\right]. \quad \square$$

Proposition 27. 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

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

Proof. From the proofs of Propositions 3 and 12, we have

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

and

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

We calculate

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

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

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

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

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

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

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

Proposition 28. 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 $\hat{F}_3(A;B,C|A)$ has approximate variance

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

where the expectations are

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

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

the variances are

$$Var\left[\hat{F}_3(A;B,C)\right] \approx \frac{4}{J^2}\sum_{j=1}^{J}\Phi_2(A_j)F_3(A_j;B_j,C_j)G(A_j) + \frac{1}{J^2}\sum_{j=1}^{J}\Phi_2(A_j)F_2(B_j,C_j)G(A_j)$$
$$+\frac{1}{J^2}\sum_{j=1}^{J}\Phi_2(B_j)F_2(A_j,C_j)G(B_j) + \frac{1}{J^2}\sum_{j=1}^{J}\Phi_2(C_j)F_2(A_j,B_j)G(C_j)$$
$$Var[\hat{G}(A)] \approx \frac{1}{J^2}\sum_{j=1}^{J}\Phi_2(A_j)G(A_j) - \frac{4}{J^2}\sum_{j=1}^{J}\Phi_2(A_j)G(A_j)^2$$

and the covariance is

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

Proof. Recall that

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

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

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

where $\mathbb{E}\left[\hat{F}_3(A;B,C)\right]$ is given in Proposition 6, $\mathbb{E}[\hat{G}(A)]$ in Lemma 1, $Var\left[\hat{F}_3(A;B,C)\right]$ in Proposition 23, $Var[\hat{G}(A)]$ in Lemma 18, and $Cov\left[\hat{F}_3(A;B,C),\hat{G}(A)\right]$ in Lemma 24. □

Lemma 29. 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 $\tilde{F}_3(A;B,C)$ and $\tilde{G}(A)$ have an approximate covariance

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

Proof. Recall that

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

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

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

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 18 and 24, we have

$$Var\left[\hat{G}(A_j)\right] \approx \Phi_2(A_j)G(A_j) - 4\Phi_2(A_j)G(A_j)^2$$

and

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

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

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

We therefore have that

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

and thus by independence of loci we have

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

Lemma 30. 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, where individual $k\in\{1,2,\ldots,N(P_j)\}$ has ploidy m_k. 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 any other terms of similar order negligible to $\Phi_2(P_j)$. Based on this simplifying assumption, the estimators $\breve{F}_3(A;B,C)$ and $\breve{G}(A)$ have an approximate covariance

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

Proof. Recall that

$$\breve{F}_3(A_j;B_j,C_j)=\hat{F}_3(A_j;B_j,C_j)-\frac{1}{\sum_{k=1}^{N(A_j)}m_k-1}\hat{G}(A_j),$$

where

$$\breve{G}(A_j)=\frac{\sum_{k=1}^{N(A_j)}m_k}{\sum_{k=1}^{N(A_j)}m_k-1}G(A_j).$$

It follows that

$$\begin{aligned}
Cov\left[\breve{F}_3(A_j;B_j,C_j),\breve{G}(A_j)\right] &= Cov\left[\hat{F}_3(A_j;B_j,C_j) - \frac{1}{\sum_{k=1}^{N(A_j)} m_k - 1}\hat{G}(A_j),\breve{G}(A_j)\right] \\
&= Cov\left[\hat{F}_3(A_j;B_j,C_j),\breve{G}(A_j)\right] - \frac{1}{\sum_{k=1}^{N(A_j)} m_k - 1} Cov\left[\hat{G}(A_j),\breve{G}(A_j)\right] \\
&= \frac{\sum_{k=1}^{N(A_j)} m_k}{\sum_{k=1}^{N(A_j)} m_k - 1} Cov\left[\hat{F}_3(A_j;B_j,C_j),\hat{G}(A_j)\right] - \frac{\sum_{k=1}^{N(A_j)} m_k}{\left[\sum_{k=1}^{N(A_j)} m_k - 1\right]^2} Var\left[\hat{G}(A_j)\right].
\end{aligned}$$

From the proofs of Lemmas 18 and 24, we have

$$Var\left[\hat{G}(A_j)\right] \approx \Phi_2(A_j)G(A_j) - 4\Phi_2(A_j)G(A_j)^2$$

and

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

By recalling the assumption that $\Phi_2(A_j)^2$ or any other term of similar magnitude (such as the $\Phi_2(A_j)/\left[\sum_{k=1}^{N(A_j)} m_k - 1\right]$ term for $N(A_j)$ large enough) is negligible compared to $\Phi_2(A_j)$, we have that

$$\frac{1}{\sum_{k=1}^{N(A_j)} m_k - 1} Var\left[\hat{G}(A_j)\right] \approx 0.$$

We therefore have that

$$Cov\left[\breve{F}_3(A_j;B_j,C_j),\breve{G}(A_j)\right] \approx \frac{\sum_{k=1}^{N(A_j)} m_k}{\sum_{k=1}^{N(A_j)} m_k - 1} Cov\left[\hat{F}_3(A_j;B_j,C_j),\hat{G}(A_j)\right],$$

and thus by independence of loci we have

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

Proposition 31. 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

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

where the variances are

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

and the covariance is

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

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 25, $Var[\tilde{G}(A)]$ in Lemma 18, and $Cov[\tilde{F}_3(A;B,C),\tilde{G}(A)]$ in Lemma 29. □

Proposition 32. 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, where individual $k\in\{1,2,\ldots,N(P_j)\}$ has ploidy m_k. Moreover, assume that no individual is related to more than one other individual, which makes the terms $\Phi_3(P_j)$, $\Phi_4(P_j)$, $\Phi_{2,2}(P_j)$, $\Phi_2(P_j)^2$, $\Phi_2(A_j)\Phi_2(B_j)$, $\Phi_2(A_j)\Phi_2(C_j)$, $\Phi_2(B_j)\Phi_2(C_j)$, and any other terms of similar order negligible to $\Phi_2(P_j)$. Based on this simplifying assumption, the ratio estimator $\breve{F}_3(A;B,C|A)$ has approximate variance

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

where the variances are

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

the covariance is

$$Cov\left[\breve{F}_3(A;B,C),\breve{G}(A)\right] \approx \frac{1}{J^2}\sum_{j=1}^{J}\frac{\Phi_2(A_j)\sum_{k=1}^{N(A_j)} m_k}{\sum_{k=1}^{N(A_j)} m_k - 1} G(A_j)\left[2G(A_j) - 2F_3(A_j;B_j,C_j) - b_j(1-c_j) - (1-b_j)c_j\right],$$

and where

$$X(A;B,C) = F_3(A;B,C) + \frac{1}{J}\sum_{j=1}^{J}\frac{\Phi_2(A_j)\sum_{k=1}^{N(A_j)} m_k - 1}{\sum_{k=1}^{N(A_j)} m_k - 1} G(A_j)$$

$$Y(A) = \frac{1}{J}\sum_{j=1}^{J}\frac{[1-\Phi_2(A_j)]\sum_{k=1}^{N(A_j)} m_k}{\sum_{k=1}^{N(A_j)} m_k - 1} G(A_j).$$

Proof. Recall that

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

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

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

where $Var\left[\breve{F}_3(A;B,C)\right]$ is given in Proposition 26, $Var\left[\breve{G}(A)\right]$ in Lemma 18, $Cov\left[\breve{F}_3(A;B,C),\breve{G}(A)\right]$ in Lemma 30, $X(A;B,C) = \mathbf{E}\left[\breve{F}_3(A;B,C)\right]$ in proof of Corollary 8, and $Y(A) = \mathbb{E}\left[\breve{G}(A)\right]$ in proof of Corollary 11. □

Lemma 33. 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

$$Cov\left[\hat{F}_4(A,B;C,D),\hat{G}(A)\right] \approx \frac{1}{J^2}\sum_{j=1}^{J}\Phi_2(A_j)G(A_j)(1-2a_j)(c_j-d_j)$$

$$Cov\left[\hat{F}_4(A,B;C,D),\hat{G}(B)\right] \approx -\frac{1}{J^2}\sum_{j=1}^{J}\Phi_2(B_j)G(B_j)(1-2b_j)(c_j-d_j)$$

$$Cov\left[\hat{F}_4(A,B;C,D),\hat{G}(C)\right] \approx \frac{1}{J^2}\sum_{j=1}^{J}\Phi_2(C_j)G(C_j)(1-2c_j)(a_j-b_j)$$

$$Cov\left[\hat{F}_4(A,B;C,D),\hat{G}(D)\right] \approx -\frac{1}{J^2}\sum_{j=1}^{J}\Phi_2(D_j)G(D_j)(1-2d_j)(a_j-b_j).$$

Proof. From the proofs of Lemma 1 and Proposition 12, we that have

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

and

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

yielding

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

We first calculate

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

Hence, we have that

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

Similarly, we have that

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

Hence, we have that

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

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

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

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

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

We know that by independence of loci we have

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

which gives

$$Cov\left[\hat{F}_4(A,B;C,D),\hat{G}(A)\right] \approx \frac{1}{J^2}\sum_{j=1}^{J}\Phi_2(A_j)G(A_j)(1-2a_j)(c_j-d_j)$$

$$Cov\left[\hat{F}_4(A,B;C,D),\hat{G}(B)\right] \approx -\frac{1}{J^2}\sum_{j=1}^{J}\Phi_2(B_j)G(B_j)(1-2b_j)(c_j-d_j)$$

$$Cov\left[\hat{F}_4(A,B;C,D),\hat{G}(C)\right] \approx \frac{1}{J^2}\sum_{j=1}^{J}\Phi_2(C_j)G(C_j)(1-2c_j)(a_j-b_j)$$

$$Cov\left[\hat{F}_4(A,B;C,D),\hat{G}(D)\right] \approx -\frac{1}{J^2}\sum_{j=1}^{J}\Phi_2(D_j)G(D_j)(1-2d_j)(a_j-b_j). \quad \square$$

Proposition 34. 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

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

where the expectation is

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

the variances are

$$Var\left[\hat{F}_4(A,B;C,D)\right] \approx \frac{1}{J^2}\sum_{j=1}^{J}\left[\Phi_2(C_j)g(C_j)+\Phi_2(D_j)G(D_j)\right]F_2(A_j,B_j)$$

$$+\frac{1}{J^2}\sum_{j=1}^{J}[\Phi_2(A_j)G(A_j)+\Phi_2(B_j)G(B_j)]F_2(C_j,D_j)$$

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

and the covariances are

$$Cov\left[\hat{F}_4(A,B;C,D),\hat{G}(A)\right] \approx \frac{1}{J^2}\sum_{j=1}^{J}\Phi_2(A_j)G(A_j)(1-2a_j)(c_j-d_j)$$

$$Cov\left[\hat{F}_4(A,B;C,D),\hat{G}(B)\right] \approx -\frac{1}{J^2}\sum_{j=1}^{J}\Phi_2(B_j)G(B_j)(1-2b_j)(c_j-d_j)$$

$$Cov\left[\hat{F}_4(A,B;C,D),\hat{G}(C)\right] \approx \frac{1}{J^2}\sum_{j=1}^{J}\Phi_2(C_j)G(C_j)(1-2c_j)(a_j-b_j)$$

$$Cov\left[\hat{F}_4(A,B;C,D),\hat{G}(D)\right] \approx -\frac{1}{J^2}\sum_{j=1}^{J}\Phi_2(D_j)G(D_j)(1-2d_j)(a_j-b_j).$$

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

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

where $\mathbb{E}\left[\hat{G}(P)\right]$ is given in Lemma 1, $Var\left[\hat{F}_4(A,B;C,D)\right]$ in Proposition 27, $Var\left[\hat{G}(P)\right]$ in Lemma 18, and $Cov\left[\hat{F}_4(A,B;C,D),\hat{G}(P)\right]$ in Lemma 33 for each population $P\in\{A,B,C,D\}$. □

Lemma 35. 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 $\tilde{F}_4(A,B;C,D)$ and $\tilde{G}(P)$, $P\in\{A,B,C,D\}$, have approximate covariances

$$Cov\left[\hat{F}_4(A,B;C,D),\tilde{G}(A)\right] \approx \frac{1}{J^2}\sum_{j=1}^{J}\frac{\Phi_2(A_j)}{1-\Phi_2(A_j)}G(A_j)(1-2a_j)(c_j-d_j)$$

$$Cov\left[\hat{F}_4(A,B;C,D),\tilde{G}(B)\right] \approx -\frac{1}{J^2}\sum_{j=1}^{J}\frac{\Phi_2(B_j)}{1-\Phi_2(B_j)}G(B_j)(1-2b_j)(c_j-d_j)$$

$$Cov\left[\hat{F}_4(A,B;C,D),\tilde{G}(C)\right] \approx \frac{1}{J^2}\sum_{j=1}^{J}\frac{\Phi_2(C_j)}{1-\Phi_2(C_j)}G(C_j)(1-2c_j)(a_j-b_j)$$

$$Cov\left[\hat{F}_4(A,B;C,D),\tilde{G}(D)\right] \approx -\frac{1}{J^2}\sum_{j=1}^{J}\frac{\Phi_2(D_j)}{1-\Phi_2(D_j)}G(D_j)(1-2d_j)(a_j-b_j).$$

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

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

From the proof of Lemma 33, we have

$$Cov\left[\hat{F}_4(A_j,B_j;C_j,D_j),\tilde{G}(A_j)\right] \approx \frac{\Phi_2(A_j)}{1-\Phi_2(A_j)} G(A_j)(1-2a_j)(c_j-d_j)$$

$$Cov\left[\hat{F}_4(A_j,B_j;C_j,D_j),\tilde{G}(B_j)\right] \approx -\frac{\Phi_2(B_j)}{1-\Phi_2(B_j)} G(B_j)(1-2b_j)(c_j-d_j)$$

$$Cov\left[\hat{F}_4(A_j,B_j;C_j,D_j),\tilde{G}(C_j)\right] \approx \frac{\Phi_2(C_j)}{1-\Phi_2(C_j)} G(C_j)(1-2c_j)(a_j-b_j)$$

$$Cov\left[\hat{F}_4(A_j,B_j;C_j,D_j),\tilde{G}(D_j)\right] \approx -\frac{\Phi_2(D_j)}{1-\Phi_2(D_j)} G(D_j)(1-2d_j)(a_j-b_j),$$

yielding

$$Cov\left[\hat{F}_4(A,B;C,D),\tilde{G}(A)\right] \approx \frac{1}{J^2}\sum_{j=1}^{J} \frac{\Phi_2(A_j)}{1-\Phi_2(A_j)} G(A_j)(1-2a_j)(c_j-d_j)$$

$$Cov\left[\hat{F}_4(A,B;C,D),\tilde{G}(B)\right] \approx -\frac{1}{J^2}\sum_{j=1}^{J} \frac{\Phi_2(B_j)}{1-\Phi_2(B_j)} G(B_j)(1-2b_j)(c_j-d_j)$$

$$Cov\left[\hat{F}_4(A,B;C,D),\tilde{G}(C)\right] \approx \frac{1}{J^2}\sum_{j=1}^{J} \frac{\Phi_2(C_j)}{1-\Phi_2(C_j)} G(C_j)(1-2c_j)(a_j-b_j)$$

$$Cov\left[\hat{F}_4(A,B;C,D),\tilde{G}(D)\right] \approx -\frac{1}{J^2}\sum_{j=1}^{J} \frac{\Phi_2(D_j)}{1-\Phi_2(D_j)} G(D_j)(1-2d_j)(a_j-b_j). \quad \square$$

Lemma 36. 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, where individual $k \in \{1,2,\ldots,N(P_j)\}$ has ploidy m_k. 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 $\breve{G}(P)$, $P \in \{A,B,C,D\}$, have approximate covariances

$$Cov\left[\hat{F}_4(A,B;C,D),\breve{G}(A)\right] \approx \frac{1}{J^2}\sum_{j=1}^{J} \frac{\Phi_2(A_j)\sum_{k=1}^{N(A_j)} m_k}{\sum_{k=1}^{N(A_j)} m_k - 1} G(A_j)(1-2a_j)(c_j-d_j)$$

$$Cov\left[\hat{F}_4(A,B;C,D),\breve{G}(B)\right] \approx -\frac{1}{J^2}\sum_{j=1}^{J} \frac{\Phi_2(B_j)\sum_{k=1}^{N(B_j)} m_k}{\sum_{k=1}^{N(B_j)} m_k - 1} G(B_j)(1-2b_j)(c_j-d_j)$$

$$Cov\left[\hat{F}_4(A,B;C,D),\breve{G}(C)\right] \approx \frac{1}{J^2}\sum_{j=1}^{J} \frac{\Phi_2(C_j)\sum_{k=1}^{N(C_j)} m_k}{\sum_{k=1}^{N(C_j)} m_k - 1} G(C_j)(1-2c_j)(a_j-b_j)$$

$$Cov\left[\hat{F}_4(A,B;C,D),\breve{G}(D)\right] \approx -\frac{1}{J^2}\sum_{j=1}^{J} \frac{\Phi_2(D_j)\sum_{k=1}^{N(D_j)} m_k}{\sum_{k=1}^{N(D_j)} m_k - 1} G(D_j)(1-2d_j)(a_j-b_j).$$

Proof. Recall that

$$\breve{G}(P_j) = \frac{\sum_{k=1}^{N(P_j)} m_k}{\sum_{k=1}^{N(P_j)} m_k - 1} \hat{G}(P_j).$$

It follows that

$$Cov\left[\hat{F}_4(A_j, B_j; C_j, D_j), \breve{G}(P_j)\right] = \frac{\sum_{k=1}^{N(P_j)} m_k}{\sum_{k=1}^{N(P_j)} m_k - 1} Cov\left[\hat{F}_4(A_j, B_j; C_j, D_j), \hat{G}(P_j)\right].$$

From the proof of Lemma 33, we have

$$Cov\left[\hat{F}_4(A_j, B_j; C_j, D_j), \breve{G}(A_j)\right] \approx \frac{\Phi_2(A_j) \sum_{k=1}^{N(A_j)} m_k}{\sum_{k=1}^{N(A_j)} m_k - 1} G(A_j)(1 - 2a_j)(c_j - d_j)$$

$$Cov\left[\hat{F}_4(A_j, B_j; C_j, D_j), \breve{G}(B_j)\right] \approx -\frac{\Phi_2(G_j) \sum_{k=1}^{N(G_j)} m_k}{\sum_{k=1}^{N(G_j)} m_k - 1} G(B_j)(1 - 2b_j)(c_j - d_j)$$

$$Cov\left[\hat{F}_4(A_j, B_j; C_j, D_j), \breve{G}(C_j)\right] \approx \frac{\Phi_2(C_j) \sum_{k=1}^{N(C_j)} m_k}{\sum_{k=1}^{N(C_j)} m_k - 1} G(C_j)(1 - 2c_j)(a_j - b_j)$$

$$Cov\left[\hat{F}_4(A_j, B_j; C_j, D_j), \breve{G}(D_j)\right] \approx -\frac{\Phi_2(D_j) \sum_{k=1}^{N(D_j)} m_k}{\sum_{k=1}^{N(D_j)} m_k - 1} G(D_j)(1 - 2d_j)(a_j - b_j),$$

yielding

$$Cov\left[\hat{F}_4(A, B; C, D), \breve{G}(A)\right] \approx \frac{1}{J^2} \sum_{j=1}^{J} \frac{\Phi_2(A_j) \sum_{k=1}^{N(A_j)} m_k}{\sum_{k=1}^{N(A_j)} m_k - 1} G(A_j)(1 - 2a_j)(c_j - d_j)$$

$$Cov\left[\hat{F}_4(A, B; C, D), \breve{G}(B)\right] \approx -\frac{1}{J^2} \sum_{j=1}^{J} \frac{\Phi_2(B_j) \sum_{k=1}^{N(B_j)} m_k}{\sum_{k=1}^{N(B_j)} m_k - 1} G(B_j)(1 - 2b_j)(c_j - d_j)$$

$$Cov\left[\hat{F}_4(A, B; C, D), \breve{G}(C)\right] \approx \frac{1}{J^2} \sum_{j=1}^{J} \frac{\Phi_2(C_j) \sum_{k=1}^{N(C_j)} m_k}{\sum_{k=1}^{N(C_j)} m_k - 1} G(C_j)(1 - 2c_j)(a_j - b_j)$$

$$Cov\left[\hat{F}_4(A, B; C, D), \breve{G}(D)\right] \approx -\frac{1}{J^2} \sum_{j=1}^{J} \frac{\Phi_2(D_j) \sum_{k=1}^{N(D_j)} m_k}{\sum_{k=1}^{N(D_j)} m_k - 1} G(D_j)(1 - 2d_j)(a_j - b_j). \quad \square$$

Proposition 37. 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

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

where the variances are

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

and the covariances are

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

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

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

where $Var\left[\hat{F}_4(A,B;C,D)\right]$ is given in Proposition 27, $Var\left[\tilde{G}(P)\right]$ in Lemma 18, and $Cov\left[\hat{F}_4(A,B;C,D),\tilde{G}(P)\right]$ in Lemma 35 for each population $P \in \{A,B,C,D\}$. □

Proposition 38. 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, where individual $k \in \{1, 2, \ldots, N(P_j)\}$ has ploidy m_k. 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 $\breve{F}_4(A,B;C,D|P)$ has approximate variance

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

where the variances are

$$Var\left[\hat{F}_4(A,B;C,D)\right] \approx \frac{1}{J^2}\sum_{j=1}^{J}\left[\Phi_2(C_j)G(C_j)+\Phi_2(D_j)G(D_j)\right]F_2(A_j,B_j)$$

$$+\frac{1}{J^2}\sum_{j=1}^{J}\left[\Phi_2(A_j)G(A_j)+\Phi_2(B_j)G(B_j)\right]F_2(C_j,D_j)$$

$$Var\left[\breve{G}(P)\right] \approx \frac{1}{J^2}\sum_{j=1}^{J}\Phi_2(P_j)\left[\frac{\sum_{k=1}^{N(P_j)} m_k}{\sum_{k=1}^{N(P_j)} m_k - 1}\right]^2 G(P_j) - \frac{4}{J^2}\sum_{j=1}^{J}\Phi_2(P_j)\left[\frac{\sum_{k=1}^{N(P_j)} m_k}{\sum_{k=1}^{N(P_j)} m_k - 1}\right]^2 G(P_j)^2,$$

the covariances are

$$Cov\left[\hat{F}_4(A,B;C,D),\breve{G}(A)\right] \approx \frac{1}{J^2}\sum_{j=1}^{J}\frac{\Phi_2(A_j)\sum_{k=1}^{N(A_j)} m_k}{\sum_{k=1}^{N(A_j)} m_k - 1}G(A_j)(1-2a_j)(c_j-d_j)$$

$$Cov\left[\hat{F}_4(A,B;C,D),\breve{G}(B)\right] \approx -\frac{1}{J^2}\sum_{j=1}^{J}\frac{\Phi_2(B_j)\sum_{k=1}^{N(B_j)} m_k}{\sum_{k=1}^{N(B_j)} m_k - 1}G(B_j)(1-2b_j)(c_j-d_j)$$

$$Cov\left[\hat{F}_4(A,B;C,D),\breve{G}(C)\right] \approx \frac{1}{J^2}\sum_{j=1}^{J}\frac{\Phi_2(C_j)\sum_{k=1}^{N(C_j)} m_k}{\sum_{k=1}^{N(C_j)} m_k - 1}G(C_j)(1-2c_j)(a_j-b_j)$$

$$Cov\left[\hat{F}_4(A,B;C,D),\breve{G}(D)\right] \approx -\frac{1}{J^2}\sum_{j=1}^{J}\frac{\Phi_2(D_j)\sum_{k=1}^{N(D_j)} m_k}{\sum_{k=1}^{N(D_j)} m_k - 1}G(D_j)(1-2d_j)(a_j-b_j),$$

and where

$$\Upsilon(P) = \frac{1}{J}\sum_{j=1}^{J}\frac{\left[1-\Phi_2(P_j)\right]\sum_{k=1}^{N(P_j)} m_k}{\sum_{k=1}^{N(P_j)} m_k - 1}G(P_j).$$

Proof. Recall that

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

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

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

where $Var\left[\hat{F}_4(A,B;C,D)\right]$ is given in Proposition 27, $Var\left[\breve{G}(P)\right]$ in Lemma 18, $Cov\left[\hat{F}_4(A,B;C,D),\breve{G}(P)\right]$ in Lemma 36, and $\Upsilon(P) = \mathbb{E}\left[\breve{G}(P)\right]$ in proof of Corollary 11 for each population $P \in \{A,B,C,D\}$. □

Lemma 39. 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

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

Proof. From the proof of Lemma 17, we have that

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

yielding

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

We first calculate

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

We compute the first term as

$$\mathbb{E}\left[\left(\hat{a}_j+\hat{b}_j-2\hat{a}_j\hat{b}_j\right)^2\right] = \mathbb{E}\left[\hat{a}_j^2\right]+\mathbb{E}\left[\hat{b}_j^2\right]+4\mathbb{E}\left[\hat{a}_j^2\right]\mathbb{E}\left[\hat{b}_j^2\right]+2\mathbb{E}[\hat{a}_j]\mathbb{E}\left[\hat{b}_j\right]-4\mathbb{E}\left[\hat{a}_j^2\right]\mathbb{E}\left[\hat{b}_j\right]-4\mathbb{E}[\hat{a}_j]\mathbb{E}\left[\hat{b}_j^2\right]$$
$$= \mathbb{E}\left[\hat{a}_j^2\right]+\mathbb{E}\left[\hat{b}_j^2\right]+4\mathbb{E}\left[\hat{a}_j^2\right]\mathbb{E}\left[\hat{b}_j^2\right]+2a_jb_j-4\mathbb{E}\left[\hat{a}_j^2\right]b_j-4a_j\mathbb{E}\left[\hat{b}_j^2\right]$$
$$\approx a_j^2+\Phi_2(A_j)a_j(1-a_j)+b_j^2+\Phi_2(B_j)b_j(1-b_j)$$
$$+4\left[a_j^2+\Phi_2(A_j)a_j(1-a_j)\right]\left[b_j^2+\Phi_2(B_j)b_j(1-b_j)\right]+2a_jb_j$$
$$-4\left[a_j^2+\Phi_2(A_j)a_j(1-a_j)\right]b_j-4a_j\left[b_j^2+\Phi_2(B_j)b_j(1-b_j)\right]$$
$$= a_j^2+b_j^2+4a_j^2b_j^2+2a_jb_j-4a_j^2b_j-4a_jb_j^2+\Phi_2(A_j)a_j(1-a_j)+\Phi_2(B_j)b_j(1-b_j)$$
$$+4\Phi_2(A_j)a_j(1-a_j)b_j^2+4\Phi_2(B_j)a_j^2b_j(1-b_j)+4\Phi_2(A_j)\Phi_2(B_j)a_j(1-a_j)b_j(1-b_j)$$
$$-4\Phi_2(A_j)a_j(1-a_j)b_j-4\Phi_2(B_j)a_jb_j(1-b_j)$$
$$= (a_j+b_j-2a_jb_j)^2+\Phi_2(A_j)G(A_j)\left[1-4b_j+4b_j^2\right]+\Phi_2(B_j)G(B_j)\left[1-4a_j+4a_j^2\right]$$
$$+4\Phi_2(A_j)\Phi_2(B_j)G(A_j)G(B_j)$$
$$\approx (a_j+b_j-2a_jb_j)^2+\Phi_2(A_j)G(A_j)(1-2b_j)^2+\Phi_2(B_j)G(B_j)(1-2a_j)^2,$$

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

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

Hence, we have that

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

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

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

Given the assumption of independent loci, we have

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

Lemma 40. 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

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

Proof. From the proofs of Proposition 12 and Lemma 17, we that have

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

and

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

yielding

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

We first calculate

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

We compute the first term as

$$\begin{aligned}\mathbb{E}\left[\left(\hat{a}_j - \hat{b}_j\right)\left(\hat{a}_j + \hat{b}_j - 2\hat{a}_j\hat{b}_j\right)\right] &= \mathbb{E}\left[\hat{a}_j^2\right] - 2\mathbb{E}\left[\hat{a}_j^2\right]\mathbb{E}\left[\hat{b}_j\right] - \mathbb{E}\left[\hat{b}_j^2\right] + 2\mathbb{E}[\hat{a}_j]\mathbb{E}\left[\hat{b}_j^2\right] \\ &= \mathbb{E}\left[\hat{a}_j^2\right]\left(1 - 2\mathbb{E}\left[\hat{b}_j\right]\right) - \mathbb{E}\left[\hat{b}_j^2\right](1 - 2\mathbb{E}[\hat{a}_j]) \\ &\approx \left[a_j^2 + \boldsymbol{\Phi}_2(A_j)a_j(1 - a_j)\right][1 - 2b_j] - \left[b_j^2 + \boldsymbol{\Phi}_2(B_j)b_j(1 - b_j)\right][1 - 2a_j] \\ &= (a_j - b_j)(a_j + b_j - 2a_jb_j) + \boldsymbol{\Phi}_2(A_j)G(A_j)(1 - 2b_j) - \boldsymbol{\Phi}_2(B_j)G(B_j)(1 - 2a_j).\end{aligned}$$

Using a similar argument, we have that

$$\mathbb{E}\left[\left(\hat{c}_j - \hat{d}_j\right)\left(\hat{c}_j + \hat{d}_j - 2\hat{c}_j\hat{d}_j\right)\right] \approx (c_j - d_j)(c_j + d_j - 2c_jd_j) + \boldsymbol{\Phi}_2(C_j)g(C_j)(1 - 2d_j) - \boldsymbol{\Phi}_2(D_j)g(D_j)(1 - 2c_j).$$

Hence, we have that

$$\begin{aligned}\mathbb{E}\left[\hat{F}_4(A_j, B_j; C_j, D_j)\hat{H}(A_j, B_j, C_j, D_j)\right] &= \mathbb{E}\left[\left(\hat{a}_j - \hat{b}_j\right)\left(\hat{a}_j + \hat{b}_j - 2\hat{a}_j\hat{b}_j\right)\right]\mathbb{E}\left[\left(\hat{c}_j - \hat{d}_j\right)\left(\hat{c}_j + \hat{d}_j - 2\hat{c}_j\hat{d}_j\right)\right] \\ &\approx \left[(a_j - b_j)(a_j + b_j - 2a_jb_j) + \boldsymbol{\Phi}_2(A_j)G(A_j)(1 - 2b_j) - \boldsymbol{\Phi}_2(B_j)G(B_j)(1 - 2a_j)\right] \\ &\times \left[(c_j - d_j)(c_j + d_j - 2c_jd_j) + \boldsymbol{\Phi}_2(C_j)G(C_j)(1 - 2d_j) - \boldsymbol{\Phi}_2(D_j)G(D_j)(1 - 2c_j)\right] \\ &\approx F_4(A_j, B_j; C_j, D_j)H(A_j, B_j, C_j, D_j) \\ &+ \boldsymbol{\Phi}_2(A_j)G(A_j)(c_j - d_j)(c_j + d_j - 2c_jd_j)(1 - 2b_j) \\ &- \boldsymbol{\Phi}_2(B_j)G(B_j)(c_j - d_j)(c_j + d_j - 2c_jd_j)(1 - 2a_j) \\ &+ \boldsymbol{\Phi}_2(C_j)G(C_j)(a_j - b_j)(a_j + b_j - 2a_jb_j)(1 - 2d_j) \\ &- \boldsymbol{\Phi}_2(D_j)G(D_j)(a_j - b_j)(a_j + b_j - 2a_jb_j)(1 - 2c_j),\end{aligned}$$

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

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

Given the assumption of independent loci, we have

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

Proposition 41. 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

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

where the variances are

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

and the covariance is

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

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

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

where $Var\left[\hat{F}_4(A,B;C,D)\right]$ is given in Proposition 27, $Var\left[\hat{H}(A,B,C,D)\right]$ in Lemma 39, and $Cov\left[\hat{F}_4(A,B;C,D),\hat{H}(A,B,C,D)\right]$ in Lemma 40. □