Estimating F-statistics Using Non-independent Samples

Kang Huang
Northwest University https://orcid.org/0000-0002-8357-117X
Bing Yang
Northwest University
Yuhang Li
Northwest University
Jincuo Ao
Northwest University
Derek W. Dunn
Northwest University
Baoguo Li (baoguoli@nwu.edu.cn)
Northwest University https://orcid.org/0000-0003-2506-7597

Research Article

Keywords: Population genetic structure, gene differentiation, fixation index, non-independent samples

Posted Date: January 17th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1240350/v1

Title

Estimating *F*-statistics using non-independent samples

Authors

Kang Huang[1,2], Bing Yang[1], Yuhang Li[1], Jincuo Ao[1], Derek W. Dunn[1], Baoguo Li[1,3,*]

Addresses

[1] Shaanxi Key Laboratory for Animal Conservation, College of Life Sciences, Northwest University, Xi'an 710069, China

[2] Department of Forest and Conservation Sciences, University of British Columbia, Vancouver, BC V6T1Z4, Canada

Corresponding author

Baoguo Li Telephone: +8613572209390; E-mail: baoguoli@nwu.edu.cn

ORCID

KH, 0000-0002-8357-117X;

DWD, 0000-0001-5909-1224;

BGL, 0000-0001-7430-3889;

Key Message

Existing *F*-statistic estimators are biased in the presence of non-independent samples. We construct a model and develop a novel estimator to evaluate and eliminate such biases, respectively.

Abstract

Existing *F*-statistic estimators fail to account for any genetic correlations among individuals or subpopulations and assume that all samples are independent. This may result in inaccurate *F*-statistic estimations for natural populations. Here, we derive the expectations of previous *F*-statistics estimates using extended kinship coefficients. On this basis, we developed a new method for *F*-statistic estimation that accounts for non-independence of samples, finite sample sizes, and autopolyploidy. As proof of principle, using the same simulated datasets we compared the accuracy of several established *F*-statistic estimators with our new estimator. We found that our new method outperformed all of the other methods we used and showed almost no bias. Our new method has been added as a new function to our existing software package POLYGENE V1.4, which is freely available at http://github.com/huangkang1987/polygene.

Keywords

Population genetic structure, gene differentiation, fixation index, non-independent samples.

Introduction

For population genetics studies, *F*-statistics are often used to quantify both the differences among (e.g., F_{ST}), and the degree of inbreeding within (e.g., F_{IS}), subpopulations. The original estimators of *F*-statistics can only be applied to diallelic and codominant markers (Wright 1951). Later, many estimators of *F*-statistics and their analogues, were developed. These estimators are based on either heterozygosity: G_{ST} (Nei 1973), G'_{ST} (Hedrick 2005) and D (Jost 2008); genetic variance decomposition: θ (Weir and Cockerham 1984); allelic size difference: R_{ST} (Slatkin 1995) and ρ_{ST} (Rousset 1996); genetic distance (Hudson et al. 1992); DNA sequences: N_{ST} (Lynch and Crease 1990) and Φ_{ST} (Excoffier et al. 1992), and quantitative traits: Q_{ST} (Spitze 1993) and P_{ST} (Leinonen et al. 2013). The two most commonly used definitions for *F*-statistics are those for the genetic correlation or standardized variance.

Wright (1951) originally defined F_{ST} as "*the correlation between random gametes, drawn from the same subpopulation, relative to the total*". This is equivalent to the definition of Cockerham's θ-statistics defined by co-ancestry (also known as *identical by descent*, IBD) (Cockerham 1969). Because the IBD definition is more straightforward, we extended this definition as *the probability of sampling two IBD alleles within units*

relative to a higher unit. For instance, the inbreeding coefficient F_{IS} is defined as the probability of sampling two IBD alleles within individuals without replacement relative to the subpopulation. Similarly, F_{ST} is defined as the probability of sampling two IBD alleles within subpopulations without replacement relative to the total population.

Wright (1965) additionally defined F_{ST} as "*the ratio of the actual variance of gene frequencies of subdivisions to its limiting value, irrespective of their own structures*". This definition is equivalent to Nei's (1973) heterozygosity based definition, *the proportion of genetic diversity that resides among populations*. We extended this definition to *the standardized variance of allele frequencies among subunits (e.g., individuals or subpopulations) relative to a higher unit (e.g., a subpopulation, a region or the total population)*. For example, F_{ST} is the standardized variance of the true allele frequencies in subpopulations relative to the total population, that is $F_{ST} = \frac{\mathrm{Var}_{s \in t}(p_s)}{p_t(1-p_t)}$, where p_s is the true frequency of a target allele in subpopulation s, and p_t is the true frequency in the total population t.

These two definitions are equivalent for independent samples, including independent subpopulations (e.g., the independent subpopulations model, Ochoa and Storey 2021) and independent individuals (e.g., the reference population). However, both definitions can be influenced by non-independent samples. Consider a null model of independent populations with a star-like phylogeny, with all sampled populations descending independently from a single ancestral population and having the same population specific F_{ST} (Weir 2012) and all individuals unrelated to each other. For non-independent individuals, relatives share IBD alleles relative to the subpopulation. The genetic correlation defined F_{ST} is increased while the standardized variance defined F_{ST} is unaffected because the allele frequency of each population remains unchanged. For non-independent populations, the phylogeny becomes tree-like. We duplicate each subpopulation in the null model to simulate recent divergence. The genetic correlation defined F_{ST} remains unaffected, because each subpopulation is unchanged and the allele frequency of the total population is also unchanged. If there are finite subpopulations, then the standardized variance defined F_{ST} is reduced by $\frac{1}{2S-2} \times 100\%$, where S is the number of subpopulations after duplication. Because F_{ST} is used to evaluate the genetic differentiation between/among subpopulations, we think it should be independent to the sample correlations. For this reason, we use the standardized variance definition with infinity subunits.

Non-independence of biological samples is common among natural populations. In addition to non-independence among individuals, subpopulations may also show genetic correlations. In general, the structure of natural populations is complex and the allele frequencies of any subpopulations are unlikely to be completely independently drawn from the ancestral population. The derived subpopulations can be either recently diverge or maintain a high level of gene-flow. Samples from such populations can thus be considered as non-independent relative to the total population.

It is important to emphasize that some applications require samples to be non-independent. For instance, in quantitative genetics, heritability is estimated by comparing the similarities of a quantitative trait between relatives and non-relatives. This can be achieved by either controlled crossing (Lynch and Walsh 1998) or by sampling naturally occurring relatives (Ritland 1996). The estimation of P_{ST} is based on the estimated heritability (Leinonen et al. 2013); the regression estimation of Q_{ST} requires an estimate of variance in pairwise F_{ST} (Ritland 2000), which infers a need for non-independent subpopulations. For another example, in molecular ecology studies of kin selection (Guo et al. 2015), social structure (Gotoh et al. 2017), mating systems (Maher et al. 2017), mate choice (Blyton et al. 2016) and seed/pollen dispersal (Pascov et al. 2015; Santos et al. 2018) all require the sampling of relatives, with the subsequent genetic analysis usually being either parentage analysis (Huang et al. 2018; Kalinowski et al. 2007), sibship reconstruction (Wang 2004, 2019) or an estimation of relatedness (Huang et al. 2018; Lynch and Ritland 1999; Wang 2002).

Because previous methods for the estimation of *F*-statistics may be inaccurate when samples are non-independent, and certain applications require non-independence of samples, we developed a new method for *F*-statistics estimation. Here, we: (i) generalize some previous methods used to calculate *F*-statistics into a single framework, i.e., by using the same set of symbols to express their expectations; (ii) evaluate how non-independent samples affect the results of previous methods of *F*-statistics estimation; (iii) describe our new method for accurately estimating *F*-statistics with non-independent samples and limited samples. Our new method has been added as a new function to the new version of our existing software package POLYGENE V1.4, which is freely available at http://github.com/huangkang1987/polygene.

Materials and Methods

Common symbols

v: ploidy level.

t (s, x or a): the total population (a subpopulation, an individual or an allele), where each object is called a *unit* with a hierarchy of units. A unit symbol with a prime (e.g., s', x' or a') denotes a distinct unit (e.g., $s \neq s'$), while a unit symbol with an asterisk (e.g., s^*, x^* or a^*) can be the same unit (e.g., $s = s^*$). For convenience, we use the ordinal number of each unit as its identifier except for the unit t.

L: the number of loci.

K: the number of alleles at a target locus.

A_k (or A_{lk}): the k^{th} allele at a target locus (or at the l^{th} locus).

A_{xa} (or A_{xla}): the a^{th} allele copy in individual x at a target locus (or at the l^{th} locus).

θ (or ϑ): the kinship coefficient relative to a subpopulation (or the total population).

θ_x (ϑ_x, θ_s, ϑ_s or ϑ_t): the kinship coefficient within a unit.

$\theta_{xx'}$ ($\vartheta_{xx'}$ or $\vartheta_{ss'}$): the kinship coefficient between two distinct units.

S (or C): the number of subpopulations (or regions) in the total population.

S_c: the number of subpopulations in the region c.

N_s (or N_t): the sample size of a subpopulation s (or the total population t).

N_f, N_m, N_o: the number of males, the number of females each male mate with, the number of offspring each pair reproduced in North Carolina Design I.

p_{xk} (or p_{xlk}): the frequency of A_k (or A_{lk}) in individual x, where the subscript x can be replaced by s or t so as to denote the allele frequency in a higher hierarchy unit.

H_I: the observed heterozygosity for a subpopulation or the total population.

$\hat{\sigma}^2_{\mathrm{WI}}$, $\hat{\sigma}^2_{\mathrm{AI}}$, $\hat{\sigma}^2_{\mathrm{AP}}$: genetic variance component within individuals, among individuals within the same subpopulations, or among subpopulations.

H_S (or H_T): the expected heterozygosity within subpopulations (or the total population).

$\widehat{M}_s$ (or $\widehat{M}_t$): the unbiased estimate of $\sum_k p_{sk}^2$ (or $\sum_k p_{tk}^2$).

Var($\cdot$) (E($\cdot$), Avg($\cdot$) or Wavg($\cdot, w$)): the operator of variance (expectation, average or weighted average with w as the weight).

Extension of kinship coefficients

We assume that our model satisfies the following conditions: (i) each population is of large enough size, and the size is constant across generations; (ii) each individual is contained within a single subpopulation, whose ploidy level is a constant v; (iii) the genotypic frequencies are at an equilibrium state; (iv) the mating system can be either monogamous or polygamous, and each individual can be either monoecious or dioecious; (v) there is a considerable proportion of relatives in each generation; (vi) the samples are randomly drawn without replacement, and the sample sizes for each distinct subpopulation may be unequal; (vii) the target loci are unlinked.

According to (i), any genetic drift between two adjacent generations is ignored, and the inbreeding coefficient F can be considered as a constant across generations. Condition (iii) can be equivalently stated as follows: the frequency of each genotype is a constant across generations. Therefore, the inbreeding coefficient F in our model is equal to the expected kinship coefficient between mates.

To extend the definitions of kinship coefficients into our model, the reference population should be revised as the *background population* (BP). It is noteworthy that 'reference population' ignores any inbreeding and or genetic correlations among individuals or subpopulations, whereas 'background population' does not.

There are two extended definitions of kinship coefficients. One is the *kinship coefficient within a unit*

relative to a BP, which is defined as the probability of sampling two IBD alleles from this unit with replacement relative to this BP, denoted by θ if the BP is a subpopulation (or all subpopulations), or by ϑ if the BP is the total population. The other is the *kinship coefficient between a pair of units* relative to a BP, which is defined as the probability of sampling two IBD alleles from these two units (each sampled from one unit) relative to this BP, still denoted by θ or ϑ. In multiple subpopulations, the following relational expression between θ and ϑ can be established:

$$\vartheta = \theta + (1-\theta)F_{ST}. \quad (1)$$

It noteworthy that the BP of the converted θ is all subpopulations other than a specific subpopulation. Equation (1) can also be used to convert ϑ into θ with BP of a specific subpopulation, the definition of F_{ST} should be revised as the differentiation of this subpopulation from the total population. For example, the allele frequencies of the specific subpopulation is drawn from a Dirichlet distribution $p_{s1}, p_{s2}, \dots, p_{sK} \sim \mathcal{D}(\gamma p_{t1}, \gamma p_{t2}, \dots, \gamma p_{tK})$, where $\gamma = 1/F_{ST} - 1$ (Pritchard et al. 2000).

The kinship coefficients defined above can be divided into five categories.

(i) Within an individual. For individual x, the kinship coefficient is denoted by θ_x or ϑ_x. Next, the average of kinship coefficients within all sampled individuals in s is denoted by $\bar{\theta}_{x\in s}$ or $\bar{\vartheta}_{x\in s}$, and that in t is denoted by $\bar{\theta}_{x\in t}$ and $\bar{\vartheta}_{x\in t}$, then $\bar{\theta}_{x\in s} = \sum_{x_1\in s}\bar{\theta}_{x_1}/N_s$, $\bar{\vartheta}_{x\in s} = \sum_{x_1\in s}\vartheta_{x_1}/N_s$, and $\bar{\vartheta}_{x\in t} = \sum_{x_1\in t}\vartheta_{x_1}/N_t$. $\bar{\theta}_{x\in t}$ can be converted from $\bar{\vartheta}_{x\in s}$ by Equation (1), and the BP is all subpopulations.

(ii) Between a pair of individuals in the same subpopulation. For a pair of individuals x and x' ($x, x' \in s$ and $x \neq x'$), the kinship coefficient is denoted by $\theta_{xx'}$ or $\vartheta_{xx'}$. Next, the average kinship coefficients between all individuals in s is denoted by $\bar{\theta}_{xx'\in s}$ or $\bar{\vartheta}_{xx'\in s}$, and that between all individuals within the same subpopulations is denoted by $\bar{\vartheta}_{xx'\in t}$, then $\bar{\theta}_{xx'\in s} = \frac{\sum_{x_1\neq x_2\in s}\theta_{x_1x_2}}{N_s(N_s-1)}$, $\bar{\vartheta}_{xx'\in s} = \frac{\sum_{x_1\neq x_2\in s}\vartheta_{x_1x_2}}{N_s(N_s-1)}$, or $\bar{\vartheta}_{xx'\in t} = \frac{\sum_{s\in t}\sum_{x_1\neq x_2\in s}\vartheta_{x_1x_2}}{\sum_{s\in t}N_s(N_s-1)}$. Similarly, $\bar{\theta}_{xx'\in t}$ can be converted from $\bar{\vartheta}_{xx'\in t}$ by Equation (1), and the BP is all subpopulations.

(iii) Within a subpopulation sample. For subpopulation s, the kinship coefficient is denoted by θ_s or ϑ_s, which can be expressed as

$$\theta_s = \frac{1}{N_s}\bar{\theta}_{x\in s} + \frac{N_s-1}{N_s}\bar{\theta}_{xx'\in s} \quad \text{or} \quad \vartheta_s = \frac{1}{N_s}\bar{\vartheta}_{x\in s} + \frac{N_s-1}{N_s}\bar{\vartheta}_{xx'\in s}. \quad (2)$$

(iv) Between a pair of subpopulation samples. For this situation, the BP can only be the total population. For subpopulation pair s and s' ($s \neq s'$), the kinship coefficient is denoted by $\vartheta_{ss'}$, which can be expressed as $\vartheta_{ss'} = \sum_{x\in s, x'\in s'}\vartheta_{xx'}/N_sN_{s'}$. Such a kinship coefficient can be used to quantify the relationship between two subpopulations. Let $\bar{\vartheta}_{s\in t}$ (or $\bar{\vartheta}_{ss'\in t}$) denote the weighted average of kinship

coefficients within (or between) all subpopulations. Then

$$\bar{\vartheta}_{s\in t} = \frac{\sum_{s_1\in t} N_{s_1}^2 \vartheta_{s_1}}{\sum_{s_1\in t} N_{s_1}^2} \quad \text{or} \quad \bar{\vartheta}_{ss'\in t} = \frac{\sum_{s_1\neq s_2\in t} N_{s_1} N_{s_2} \vartheta_{s_1 s_2}}{\sum_{s_1\neq s_2\in t} N_{s_1} N_{s_2}}. \tag{3}$$

By Equation (2), the first formula in Equation (3) can be expressed as

$$\bar{\vartheta}_{s\in t} = \frac{\sum_{s_1\in t} N_{s_1} \bar{\vartheta}_{x\in s_1} + \sum_{s_1\in t}(N_{s_1}^2 - N_{s_1})\bar{\vartheta}_{xx'\in s_1}}{\sum_{s_1\in t} N_{s_1}^2},$$

that is

$$\bar{\vartheta}_{s\in t} = \frac{\mathrm{Wavg}_{s_1\in t}\left(\bar{\vartheta}_{x\in s_1}, N_{s_1}\right)\sum_{s_1} N_{s_1} + \mathrm{Wavg}_{s_1\in t}\left(\bar{\vartheta}_{xx'\in s_1}, N_{s_1}^2 - N_{s_1}\right)\sum_{s_1}(N_{s_1}^2 - N_{s_1})}{\sum_{s_1} N_{s_1}^2}. \tag{4}$$

(v) Within the total population sample. For this situation, the kinship coefficient is denoted by ϑ_t, which can be expressed as

$$\vartheta_t = \frac{\sum_{s_1} N_{s_1}^2}{N_t^2}\bar{\vartheta}_{s\in t} + \frac{\sum_{s_1\neq s_2} N_{s_1} N_{s_2}}{N_t^2}\bar{\vartheta}_{ss'\in t}. \tag{5}$$

Both θ and ϑ are random variables so we can consider certain expectations. We will denote $\mathrm{E}(\theta)$ by $\tilde{\theta}$ and $\mathrm{E}(\vartheta)$ by $\tilde{\vartheta}$. For example, for the kinship coefficient of an individual, we have $\tilde{\theta}_x = \frac{1}{v} + \frac{v-1}{v}F_{IS}$ and $\tilde{\vartheta}_x = \frac{1}{v} + \frac{v-1}{v}F_{IT}$. For another example, if the individuals and the subpopulations are independent, then $\tilde{\vartheta}_{xx'} = F_{ST}$ and $\tilde{\vartheta}_{ss'} = 0$. Otherwise, because the structure of populations is more complex, these equalities do not generally hold.

Before concluding this section, some formulas will be listed, which will be used in the section *Estimation of* $\tilde{\theta}$ *and* $\tilde{\vartheta}$. In a single subpopulation, by Equation (2), $\tilde{\theta}_s = \frac{1}{N_s}\tilde{\theta}_x + \frac{N_s-1}{N_s}\tilde{\theta}_{xx'}$ and therefore

$$\tilde{\theta}_{xx'} = \frac{N_s\tilde{\theta}_s - \tilde{\theta}_x}{N_s - 1}. \tag{6}$$

In multiple subpopulations, the value of $\tilde{\theta}_s$ among subpopulations will differ because of the different sample sizes and BPs. Alternatively, we use t as the BP and calculate the weighted average $\tilde{\vartheta}_s = \mathrm{Wavg}_{s_1\in t}\left(\tilde{\vartheta}_{s_1}, N_{s_1}^2\right)$. By Equation (4), it can be derived that $\tilde{\vartheta}_{s_1} = \frac{1}{N_{s_1}}\tilde{\vartheta}_x + \frac{N_{s_1}-1}{N_{s_1}}\tilde{\vartheta}_{xx'}$, therefore

$$\tilde{\vartheta}_{xx'} = \frac{\sum_{s_1} N_{s_1}^2\,\tilde{\vartheta}_s - N_t\tilde{\vartheta}_x}{\sum_{s_1} N_{s_1}^2 - N_t}. \tag{7}$$

Besides, $\tilde{\theta}_x$, $\tilde{\theta}_{xx'}$ and $\tilde{\theta}_s$ can be converted from $\tilde{\vartheta}_x$, $\tilde{\vartheta}_{xx'}$ and $\tilde{\vartheta}_s$ by Equation (1), and the BP is all subpopulations. Finally, by Equation (5), we obtain

$$\tilde{\vartheta}_{ss'} = \frac{N_t^2\tilde{\vartheta}_t - \sum_{s_1} N_{s_1}^2\tilde{\vartheta}_s}{N_t^2 - \sum_{s_1} N_{s_1}^2}. \quad (8)$$

Wright's standardized variance estimator

Wright's (1965) standardized variance estimator can be used to estimate F_{ST}, but can also be extended to estimate the *F*-statistic F_{IS} for a single subpopulation (say s). Let $F_{IS,\mathrm{Wr}}$ denote Wright's F_{IS} for subpopulation s, then its expression at the l^{th} locus is

$$F_{IS,\mathrm{Wr}} = \frac{\sum_k \mathrm{Var}_{x\in s}(p_{xlk})}{\sum_k p_{slk}(1-p_{slk})}.$$

Under actual conditions, the true allele frequencies are unknown, and each frequency is estimated by counting the number of corresponding allele copies. The estimated expression of $F_{IS,\mathrm{Wr}}$ is

$$\hat{F}_{IS,\mathrm{Wr}} = \frac{\sum_k \mathrm{Avg}_{x\in s}(\hat{p}_{xlk}^2) - \sum_k \hat{p}_{slk}^2}{1 - \sum_k \hat{p}_{slk}^2}. \quad (9)$$

Because the numerator and the denominator in Equation (9) are correlated, it is difficult to obtain the expectation $\mathrm{E}(\hat{F}_{IS,\mathrm{Wr}})$ related to a single locus. In order to reduce the effects of such a correlation, we extend a single locus into all loci and use an alternative method to give the expression $\hat{F}_{IS,\mathrm{Wr}}$ at all loci. In fact, for the fractions such as Equation (9) across all loci, we sum all of the numerators, and also sum all of the denominators. Then the expression of $\hat{F}_{IS,\mathrm{Wr}}$ at all loci is

$$\hat{F}_{IS,\mathrm{Wr}} = \frac{\sum_l \sum_k \mathrm{Avg}_{x\in s}(\hat{p}_{xlk}^2) - \sum_l \sum_k \hat{p}_{slk}^2}{L - \sum_l \sum_k \hat{p}_{slk}^2}. \quad (10)$$

For this fraction, because the correlation between the numerator and denominator is close to zero, we obtain

$$\mathrm{E}\left(\hat{F}_{IS,\mathrm{Wr}} | \bar{\theta}_{x\in \mathrm{s}}, \theta_s\right) \approx \frac{\mathrm{E}[\sum_l \sum_k \mathrm{Avg}_{x\in s}(\hat{p}_{xlk}^2) - \sum_l \sum_k \hat{p}_{slk}^2]}{\mathrm{E}(L - \sum_l \sum_k \hat{p}_{slk}^2)} = \frac{\bar{\theta}_{x\in \mathrm{s}} - \theta_s}{1-\theta_s}.$$

Further, we can also derive $\mathrm{E}\left(\hat{F}_{IS,\mathrm{Wr}}\right)$ by $\mathrm{E}(\bar{\theta}_{x\in \mathrm{s}}) = \tilde{\theta}_x$ and $\mathrm{E}(\theta_s) = \tilde{\theta}_s$, the result of which is (see Appendix A for derivation):

$$\mathrm{E}\left(\hat{F}_{IS,\mathrm{Wr}}\right) \approx \frac{\tilde{\theta}_x - \tilde{\theta}_s}{1-\tilde{\theta}_s}. \quad (11)$$

Therefore, by substituting $\mathrm{E}(\hat{F}_{IS,\mathrm{Wr}})$ with $\hat{F}_{IS,\mathrm{Wr}}$, $\approx$ with $=$, $\tilde{\theta}_x$ with $\frac{1}{v}+\frac{v-1}{v}\hat{F}_{IS,\mathrm{c}}$ and $\tilde{\theta}_s$ with $\hat{\tilde{\theta}}_s$, we can develop the unbiased correction $\hat{F}_{IS,\mathrm{c}}$ of Wright's (1965) standardized variance estimator as follows:

$$\hat{F}_{IS,\mathrm{c}}=\frac{v}{v-1}\left[\hat{\tilde{\theta}}_s+\left(1-\hat{\tilde{\theta}}_s\right)\hat{F}_{IS,\mathrm{Wr}}\right]-\frac{1}{v-1}, \tag{12}$$

where the value of $\hat{F}_{IS,\mathrm{Wr}}$ is obtained from Equation (10), and $\hat{\tilde{\theta}}_s$ is calculated by Equation (21).

The above method can be extended to the situation of multiple subpopulations. For this situation, the BP is the total population. Moreover, $\hat{F}_{IS,\mathrm{Wr}}$ is a weighted average related to all subpopulations, $\hat{F}_{ST,\mathrm{Wr}}$ is the standardized variance of $\hat{p}_{slk}$ related to all subpopulations, and $\hat{F}_{IT,\mathrm{Wr}}$ is solved from the equation $1-\hat{F}_{IT,\mathrm{Wr}}=\left(1-\hat{F}_{IS,\mathrm{Wr}}\right)\left(1-\hat{F}_{ST,\mathrm{Wr}}\right)$, whose expressions at all loci are

$$\hat{F}_{IS,\mathrm{Wr}}=\frac{\sum_l\sum_k \mathrm{Avg}_{x\in t}(\hat{p}_{xlk}^2)-\ \sum_l\sum_k \mathrm{Wavg}_{s\in t}(\hat{p}_{slk}^2,N_s^2)}{L-\ \sum_l\sum_k \mathrm{Wavg}_{s\in t}(\hat{p}_{slk}^2,N_s^2)},$$

$$\hat{F}_{ST,\mathrm{Wr}}=\frac{\sum_l\sum_k \mathrm{Wavg}_{s\in t}(\hat{p}_{slk}^2,N_s^2)-\ \sum_l\sum_k \hat{p}_{tlk}^2}{L-\ \sum_l\sum_k \hat{p}_{tlk}^2},$$

$$\hat{F}_{IT,\mathrm{Wr}}=\frac{\sum_l\sum_k \mathrm{Avg}_{x\in t}(\hat{p}_{xlk}^2)-\sum_l\sum_k \hat{p}_{tlk}^2}{L-\ \sum_l\sum_k \hat{p}_{tlk}^2}.$$

Now, we can derive the approximate expressions of $\mathrm{E}(\hat{F}_{IS,\mathrm{Wr}})$, $\mathrm{E}(\hat{F}_{ST,\mathrm{Wr}})$ and $\mathrm{E}(\hat{F}_{IT,\mathrm{Wr}})$ as follows (see Appendix A for derivation):

$$\mathrm{E}(\hat{F}_{IS,\mathrm{Wr}})\approx\frac{\tilde{\vartheta}_x-\tilde{\vartheta}_s}{1-\tilde{\vartheta}_s},$$

$$\mathrm{E}(\hat{F}_{ST,\mathrm{Wr}})\approx\frac{\tilde{\vartheta}_s-\tilde{\vartheta}_t}{1-\tilde{\vartheta}_t}\quad\text{and}\quad\mathrm{E}(\hat{F}_{IT,\mathrm{Wr}})\approx\frac{\tilde{\vartheta}_x-\tilde{\vartheta}_t}{1-\tilde{\vartheta}_t}. \tag{13}$$

By Equation (1), it can be derived that

$$\frac{\tilde{\vartheta}_x-\tilde{\vartheta}_s}{1-\tilde{\vartheta}_s}=\frac{\tilde{\theta}_x-\tilde{\theta}_s}{1-\tilde{\theta}_s}\quad\text{and}\quad\frac{\tilde{\vartheta}_s-\tilde{\vartheta}_t}{1-\tilde{\vartheta}_t}=\frac{\tilde{\theta}_s+\left(1-\tilde{\theta}_s\right)F_{ST}-\tilde{\vartheta}_t}{1-\tilde{\vartheta}_t}.$$

Therefore, using the facts that $\tilde{\theta}_x=\frac{1}{v}+\frac{v-1}{v}F_{IS}$ and $\tilde{\vartheta}_x=\frac{1}{v}+\frac{v-1}{v}F_{IT}$, we can develop the unbiased corrections $\hat{F}_{IS,\mathrm{c}}$, $\hat{F}_{ST,\mathrm{c}}$ and $\hat{F}_{IT,\mathrm{c}}$ of such an estimator as follows:

$$\hat{F}_{IS,\mathrm{c}}=\frac{v}{v-1}\left[\hat{\tilde{\theta}}_s+\left(1-\hat{\tilde{\theta}}_s\right)\hat{F}_{IS,\mathrm{Wr}}\right]-\frac{1}{v-1},$$

$$\hat{F}_{ST,\mathrm{c}}=\frac{\hat{F}_{ST,\mathrm{Wr}}+\left(1-\hat{F}_{ST,\mathrm{Wr}}\ \right)\hat{\tilde{\vartheta}}_t-\hat{\tilde{\theta}}_s}{1-\hat{\tilde{\theta}}_s},$$

$$\hat{F}_{IT,\mathrm{c}}=\frac{v}{v-1}\left[\hat{\tilde{\vartheta}}_t+\left(1-\hat{\tilde{\vartheta}}_t\right)\hat{F}_{IT,\mathrm{Wr}}\right]-\frac{1}{v-1},$$

where the values of $\hat{F}_{IS,\mathrm{Wr}}$, $\hat{F}_{ST,\mathrm{Wr}}$ and $\hat{F}_{IT,\mathrm{Wr}}$ are obtained from the above expressions, and $\hat{\tilde{\vartheta}}_t$ is

calculated by Equation (22).

Nei's heterozygosity estimator

Nei's (1973) extended Wright's (1965) original variance standardization, to enable support for multi-allelic loci. Nei's resulting '*G*-statistics' are closely related to the observed and the expected heterozygosities. To account for polysomic inheritance, the *observed heterozygosity* H_I at a target locus is defined as the average probability of randomly sampling two non-IBS alleles within individuals (in a subpopulation s or in the total population t) at this locus without replacement. The definition of *expected heterozygosity* H_S (or H_T) at a target locus under polysomic inheritance is the same as that for disomic inheritance, whose value is $1-\sum_k \hat{p}_{sk}^2$ (or $1-\sum_k \hat{p}_{tk}^2$). In multiple subpopulations, $\hat{p}_{sk}^2$ is defined as a weighted average, that is

$$\hat{p}_{sk}^2 = \mathrm{Wavg}_{s_1 \in t}\left(\hat{p}_{s_1 k}^2, N_{s_1}^2\right), \quad k = 1, 2, \cdots, K.$$

For the situation of a single subpopulation (say s), the estimate $\hat{G}_{IS}$ of Nei's (1973) heterozygosity estimator is calculated by $\hat{G}_{IS} = 1 - H_I/H_S$. Following the approach in Appendix A, we can derive the approximate expression of $\mathrm{E}(\hat{G}_{IS})$ as follows (see Appendix B for derivation):

$$\mathrm{E}(\hat{G}_{IS}) \approx \frac{F_{IS} - \tilde{\theta}_s}{1 - \tilde{\theta}_s}. \tag{14}$$

Further, we can develop the unbiased correction $\hat{F}_{IS,c}$ of such an estimator as follows:

$$\hat{F}_{IS,c} = \hat{\tilde{\theta}}_s + \left(1 - \hat{\tilde{\theta}}_s\right)\hat{G}_{IS}. \tag{15}$$

Besides, Li & Horvitz's (1953) F_{IS} estimator calculates the Pearson's correlation coefficient between distinct alleles within individuals, and whose approximate expression of $\mathrm{E}(\hat{G}_{IS})$ is the same as Equation (14) (see Appendix C for derivation).

For the situation of multiple subpopulations, $\hat{G}_{IS}$, $\hat{G}_{ST}$ and $\hat{G}_{IT}$ are, respectively, calculated as

$$\hat{G}_{IS} = 1 - H_I/H_S \ , \quad \hat{G}_{ST} = 1 - H_S/H_T \quad \text{and} \quad \hat{G}_{IT} = 1 - H_I/H_T. \tag{16}$$

Following the approach outlined in Appendix A, we can derive the following approximate expressions (see Appendix B for derivations):

$$\mathrm{E}\left(\hat{G}_{IS}\right) \approx \frac{F_{IT}-\tilde{\vartheta}_s}{1-\tilde{\vartheta}_s},$$

$$\mathrm{E}\left(\hat{G}_{ST}\right) \approx \frac{\tilde{\vartheta}_s-\tilde{\vartheta}_t}{1-\tilde{\vartheta}_t} \quad \text{and} \quad \mathrm{E}\left(\hat{G}_{IT}\right) \approx \frac{F_{IT}-\tilde{\vartheta}_t}{1-\tilde{\vartheta}_t}. \tag{17}$$

Moreover, according to Equation (1), as well as the identical equation $1-F_{IT}=(1-F_{IS})(1-F_{ST})$, we can derive that

$$\frac{F_{IT}-\tilde{\vartheta}_s}{1-\tilde{\vartheta}_s}=\frac{F_{IS}-\tilde{\theta}_s}{1-\tilde{\theta}_s} \quad \text{and} \quad \frac{\tilde{\vartheta}_s-\tilde{\vartheta}_t}{1-\tilde{\vartheta}_t}=\frac{\tilde{\theta}_s+\left(1-\tilde{\theta}_s\right)F_{ST}-\tilde{\vartheta}_t}{1-\tilde{\vartheta}_t}.$$

Thus, we can also develop the unbiased corrections $\hat{F}_{IS,\mathrm{c}}$, $\hat{F}_{ST,\mathrm{c}}$ and $\hat{F}_{IT,\mathrm{c}}$ of such an estimator as follows:

$$\hat{F}_{IS,\mathrm{c}}=\hat{\tilde{\theta}}_s+\left(1-\hat{\tilde{\theta}}_s\right)\hat{G}_{IS},$$

$$\hat{F}_{ST,\mathrm{c}}=\frac{\hat{\tilde{\vartheta}}_t+\left(1-\hat{\tilde{\vartheta}}_t\right)\hat{G}_{ST}-\hat{\tilde{\theta}}_s}{1-\hat{\tilde{\theta}}_s},$$

$$\hat{F}_{IT,\mathrm{c}}=\hat{\tilde{\vartheta}}_t+\left(1-\hat{\tilde{\vartheta}}_t\right)\hat{G}_{IT}.$$

Weir & Cockerham's variance component estimator

Weir & Cockerham's (1984) estimator is based on variance decomposition. We denote σ^2_{WI} (σ^2_{AI} or σ^2_{AP}) for the variance component of genetic distances between the allele copies within individuals (among individuals or among subpopulations).

For the situation of a single subpopulation (say s), there are only two variance components: σ^2_{WI} and σ^2_{AI}, which are defined as $\sigma^2_{\mathrm{WI}} \overset{\text{def}}{=} (1-F_{IS})(1-\sum_k p^2_{sk})/2$ and $\sigma^2_{\mathrm{AI}} \overset{\text{def}}{=} F_{IS}(1-\sum_k p^2_{sk})/2$. In this situation, the estimate $\hat{F}_{IS,\mathrm{W\&C}}$ of Weir & Cockerham's (1984) estimator can be expressed as

$$\hat{F}_{IS,W\&C}=\frac{\hat{\sigma}^2_{\mathrm{AI}}}{\hat{\sigma}^2_{\mathrm{WI}}+\hat{\sigma}^2_{\mathrm{AI}}},$$

where $\hat{\sigma}^2_{\mathrm{WI}}$ and $\hat{\sigma}^2_{\mathrm{AI}}$ can both be obtained using an *analysis of molecule variance* (AMOVA) framework. Moreover, based on Huang *et al.*'s (2020) generalized AMOVA framework, we can obtain the approximate expression of $\mathrm{E}\left(\hat{F}_{IS,\mathrm{W\&C}}\right)$ as follows (see Appendix D for derivation):

$$\mathrm{E}\left(\hat{F}_{IS,\mathrm{W\&C}}\right) \approx \frac{F_{IS}-\tilde{\theta}_{xx'}}{1-\tilde{\theta}_{xx'}}. \tag{18}$$

Equation (18) is hold for polysomic inheritance. The unbiased correction $\hat{F}_{IS,\mathrm{c}}$ of such estimator is as follows:

$$\hat{F}_{IS,\mathrm{c}} = \hat{\tilde{\theta}}_{xx'} + \left(1 - \hat{\tilde{\theta}}_{xx'}\right)\hat{F}_{IS,\mathrm{W\&C}}, \tag{19}$$

where $\hat{\tilde{\theta}}_{xx'}$ is calculated by Equation (21).

For the situation of multiple subpopulations, the three variance components are defined as

$$\sigma^2_{\mathrm{WI}} \overset{\text{def}}{=} \frac{1}{2}(1 - F_{IT})\left(1 - \sum_k p_{tk}^2\right),\ \sigma^2_{\mathrm{AI}} \overset{\text{def}}{=} \frac{1}{2}(F_{IT} - F_{ST})\left(1 - \sum_k p_{tk}^2\right) \text{ and } \sigma^2_{\mathrm{AP}} \overset{\text{def}}{=} \frac{1}{2}F_{ST}\left(1 - \sum_k p_{tk}^2\right).$$

In this situation, the estimates $\hat{F}_{IS,\mathrm{W\&C}}$, $\hat{F}_{ST,\mathrm{W\&C}}$ and $\hat{F}_{IT,\mathrm{W\&C}}$ can be expressed as

$$\hat{F}_{IS,\mathrm{W\&C}} = \frac{\hat{\sigma}^2_{\mathrm{AI}}}{\hat{\sigma}^2_{\mathrm{WI}} + \hat{\sigma}^2_{\mathrm{AI}}}, \qquad \hat{F}_{ST,\mathrm{W\&C}} = \frac{\hat{\sigma}^2_{\mathrm{AI}}}{\hat{\sigma}^2_{\mathrm{WI}} + \hat{\sigma}^2_{\mathrm{AI}} + \hat{\sigma}^2_{\mathrm{AP}}} \qquad \text{and} \qquad \hat{F}_{IT,\mathrm{W\&C}} = \frac{\hat{\sigma}^2_{\mathrm{AI}} + \hat{\sigma}^2_{\mathrm{AP}}}{\hat{\sigma}^2_{\mathrm{WI}} + \hat{\sigma}^2_{\mathrm{AI}} + \hat{\sigma}^2_{\mathrm{AP}}}.$$

Similarly, $\mathrm{E}\left(\hat{F}_{IS,\mathrm{W\&C}}\right)$, $\mathrm{E}\left(\hat{F}_{ST,\mathrm{W\&C}}\right)$ and $\mathrm{E}\left(\hat{F}_{IT,\mathrm{W\&C}}\right)$ can be derived based on Huang *et al.*'s (2020) generalized AMOVA framework (see Appendix D), whose approximate expressions are

$$\mathrm{E}\left(\hat{F}_{IS,\mathrm{W\&C}}\right) \approx \frac{F_{IT} - \tilde{\vartheta}_{xx'}}{1 - \tilde{\vartheta}_{xx'}}, \quad \mathrm{E}\left(\hat{F}_{ST,\mathrm{W\&C}}\right) \approx \frac{\tilde{\vartheta}_{xx'} - \tilde{\vartheta}_{ss'}}{1 - \tilde{\vartheta}_{ss'}} \quad \text{and} \quad \mathrm{E}\left(\hat{F}_{IT,\mathrm{W\&C}}\right) \approx \frac{F_{IT} - \tilde{\vartheta}_{ss'}}{1 - \tilde{\vartheta}_{ss'}}. \tag{20}$$

Because

$$\frac{F_{IT} - \tilde{\vartheta}_{xx'}}{1 - \tilde{\vartheta}_{xx'}} = \frac{F_{IS} - \tilde{\theta}_{xx'}}{1 - \tilde{\theta}_{xx'}} \qquad \text{and} \qquad \frac{\tilde{\vartheta}_{xx'} - \tilde{\vartheta}_{ss'}}{1 - \tilde{\vartheta}_{ss'}} = \frac{\tilde{\theta}_{xx'} + \left(1 - \tilde{\theta}_{xx'}\right)F_{ST} - \tilde{\vartheta}_{ss'}}{1 - \tilde{\vartheta}_{ss'}},$$

we can develop unbiased corrections of such estimator as follows:

$$\hat{F}_{IS,\mathrm{c}} = \hat{\tilde{\theta}}_{xx'} + \left(1 - \hat{\tilde{\theta}}_{xx'}\right)\hat{F}_{IS,\mathrm{W\&C}},$$

$$\hat{F}_{ST,\mathrm{c}} = \frac{\hat{\tilde{\vartheta}}_{ss'} + (1 - \hat{\tilde{\vartheta}}_{ss'})\hat{F}_{ST,\mathrm{W\&C}} - \hat{\tilde{\theta}}_{xx'}}{1 - \hat{\tilde{\theta}}_{xx'}},$$

$$\hat{F}_{IT,\mathrm{c}} = \hat{\tilde{\vartheta}}_{ss'} + \left(1 - \hat{\tilde{\vartheta}}_{ss'}\right)\hat{F}_{IT,\mathrm{W\&C}}.$$

where $\hat{\tilde{\vartheta}}_{ss'}$ is calculated by Equation (22).

Estimation of $\tilde{\theta}$ and $\tilde{\vartheta}$

The unbiased correction formulas in Equations (12), (15), (19) and so on include the estimates $\hat{\tilde{\theta}}_s$, $\hat{\tilde{\theta}}_{xx'}$, $\hat{\tilde{\vartheta}}_t$ and $\hat{\tilde{\vartheta}}_{ss'}$. Although various *F*-statistics seem to have multiple constraints, we cannot generally solve the expectations $\tilde{\theta}$ and $\tilde{\vartheta}$ from these constraints, and thus we must adopt alternatives to estimate them. In this section, by associating the kinship coefficient with the effective population size, we describe here our alternative approach.

The effective population size is the size of an ideal Wright-Fisher population that shows the same

allele frequency change over time as an observed biological population regardless of its census population size (Hamilton 2009). The inbreeding effective population size of a subpopulation (say s) is defined as

$$N_{e,s} \stackrel{\text{def}}{=} \frac{1}{v_{\text{W\&F}}\tilde{\theta}_s'},$$

where $v_{\text{W\&F}}$ is the ploidy level of the Wright-Fisher population, and $\tilde{\theta}_s'$ is defined as $\tilde{\theta}_s' \stackrel{\text{def}}{=} \frac{1}{N_s'}\tilde{\theta}_x + \frac{N_s'-1}{N_s'}\tilde{\theta}_{xx'}$ in which N_s' is the population size.

It can be seen from Equation (2) that $\tilde{\theta}_s'$ is similar to $\tilde{\theta}_s$, so we can take advantage of the previous N_e estimators to subsequently estimate $\tilde{\theta}_s$ and $\tilde{\theta}_{xx'}$. Nomura (2008) developed an N_e estimator based on molecular co-ancestry, which uses the previous generation as the reference population, such that there are only two types of relatives in the current generation: full-sibs and half-sibs.

Key to Nomura's (2008) N_e estimator is the estimation of $\sum_k p_{slk}^2$. Previously, $\sum_k p_{slk}^2$ is simply estimated by $\sum_k \hat{p}_{slk}^2$, which is biased due to the nature of non-independent samples. In Nomura's (2008) N_e estimator, $\widehat{\sum_k p_{slk}^2}$ is calculated as $\underset{x\text{-}x'\in\mathbf{R}_{\bar{l}}}{\text{Avg}}(\sum_k \hat{p}_{xlk}\hat{p}_{x'lk})$, where x-x' represents a pair of individuals, and $\mathbf{R}_{\bar{l}}$ is the collection consisting of the putative non-relatives related to all loci except for the l^{th} locus. This method is thus less affected by the non-independence of samples. The putative non-relatives related to the l^{th} locus is determined by using the pairwise kinship estimate $\hat{\theta}_{xx'\bar{l}}$ as an indicator. Nomura's (2008) N_e estimator calculates $\theta_{xx'\bar{l}}$ by using Weir's (1996) kinship estimator, and assigns a fixed proportion of individual pairs with a smaller $\hat{\theta}_{xx'\bar{l}}$ to be the putative non-relatives related to the l^{th} locus.

Hardy (2002) extended Weir's (1996) kinship estimator to include samples with polysomic inheritance. For the l^{th} locus, the extended estimator is still denoted by $\hat{\theta}_{xx'\bar{l}}$, whose expression is

$$\hat{\theta}_{xx'\bar{l}} = \frac{\sum_{l'\neq l}\left(\sum_k \hat{p}_{xl'k}\hat{p}_{x'l'k} - \sum_k \hat{p}_{sl'k}^2\right)}{\sum_{l'\neq l}\left(1 - \sum_k \hat{p}_{sl'k}^2\right)},$$

where x and x' are in the collection of some given samples. This estimator is also used as an indicator to assign putative non-relatives.

Because Nomura's (2008) N_e estimator uses the parental generation as the reference population, the inbreeding within individuals and the kinship between individuals in the parental generation have to be ignored. Under such conditions, F_{IS} in the current generation is equal to zero, which is inconsistent with our model. To estimate $\tilde{\theta}$ and $\tilde{\vartheta}$, we have to relax such assumptions and assume that each population satisfies the following three equilibrium conditions: i) the inbreeding coefficient in the parental generation is identical to that in the current generation, ii) the kinship coefficient between mates is F_{IS} so as to produce the inbred offspring, and iii) the kinship coefficient for non-parental individual

pairs is zero. It is not difficult to see that these conditions do not influence the estimation of $\sum_k p_{slk}^2$.

The calculations of various estimates of $\tilde{\theta}$ and $\tilde{\vartheta}$ can be obtained by using some conclusions in one of the three Appendices A, B and D. In the subsequent text, we will use the example given in Appendix A to perform our derivation.

For the situation of a single subpopulation s, by using Equations (A1) and (A2) in Appendix A, we have

$$\mathrm{E}[\mathrm{Avg}_{x\in s}(\textstyle\sum_k \hat{p}_{xlk}^2)] = \tilde{\theta}_x(1-\textstyle\sum_k p_{slk}^2) + \textstyle\sum_k p_{slk}^2,$$

$$\mathrm{E}(\textstyle\sum_k \hat{p}_{slk}^2) = \tilde{\theta}_s(1-\textstyle\sum_k p_{slk}^2) + \textstyle\sum_k p_{slk}^2.$$

As mentioned above, the estimate $\widehat{\sum_k p_{slk}^2}$ will be defined as $\underset{x\text{-}x'\in\mathbf{R}_{\bar{l}}}{\mathrm{Avg}}(\sum_k \hat{p}_{xlk}\hat{p}_{x'lk})$, denoted by $\widehat{M}_{sl}$. We use $\sum_{l'}\sum_k \hat{p}_{xl'k}\hat{p}_{x'l'k} - \sum_k \hat{p}_{xlk}\hat{p}_{x'lk}$ as the indicator to assign a predefined proportion of putative relatives instead of $\hat{\theta}_{xx'\bar{l}}$, which is faster because $\sum_{l'}\sum_k \hat{p}_{xl'k}\hat{p}_{x'l'k}$ is a constant for each individual. By substituting $\mathrm{E}[\mathrm{Avg}_{x\in s}(\sum_k \hat{p}_{xlk}^2)]$ with $\mathrm{Avg}_{x\in s}(\sum_k \hat{p}_{xlk}^2)$, $\mathrm{E}(\sum_k \hat{p}_{slk}^2)$ with $\sum_k \hat{p}_{slk}^2$, $\tilde{\theta}_x$ with $\hat{\tilde{\theta}}_x$, $\tilde{\theta}_s$ with $\hat{\tilde{\theta}}_s$ and $\sum_k p_{slk}^2$ with $\widehat{M}_{sl}$, we are able to derive the following expressions related to the l^{th} locus:

$$\hat{\tilde{\theta}}_x = \frac{\mathrm{Avg}_{x\in s}(\sum_k \hat{p}_{xlk}^2) - \widehat{M}_{sl}}{1-\widehat{M}_{sl}} \qquad \text{and} \qquad \hat{\tilde{\theta}}_s = \frac{\sum_k \hat{p}_{slk}^2 - \widehat{M}_{sl}}{1-\widehat{M}_{sl}}.$$

In addition, by using Equation (6), the values of $\hat{\tilde{\theta}}_x$, $\hat{\tilde{\theta}}_s$ and $\hat{\tilde{\theta}}_{xx'}$ related to all loci can be calculated by

$$\hat{\tilde{\theta}}_x = \frac{\sum_l \mathrm{Avg}_{x\in s}(\sum_k \hat{p}_{xlk}^2) - \sum_l \widehat{M}_{sl}}{L-\sum_l \widehat{M}_{sl}},$$

$$\hat{\tilde{\theta}}_s = \frac{\sum_l \sum_k \hat{p}_{slk}^2 - \sum_l \widehat{M}_{sl}}{L-\sum_l \widehat{M}_{sl}}, \tag{21}$$

$$\hat{\tilde{\theta}}_{xx'} = \frac{N_s\hat{\tilde{\theta}}_s - \hat{\tilde{\theta}}_x}{N_s-1}.$$

For the situation of multiple subpopulations, by using Equations (A3-A5) in Appendix A, by using the previous method we are able to derive the following expressions related to the l^{th} locus:

$$\hat{\tilde{\vartheta}}_x = \frac{\mathrm{Avg}_{x\in t}(\sum_k \hat{p}_{xlk}^2) - \widehat{M}_{tl}}{1-\widehat{M}_{tl}}$$

$$\hat{\tilde{\vartheta}}_s = \frac{\mathrm{Wavg}_{s\in t}(\hat{p}_{slk}^2, N_s^2) - \widehat{M}_{tl}}{1-\widehat{M}_{tl}},$$

$$\hat{\tilde{\vartheta}}_t = \frac{\sum_k \hat{p}_{tlk}^2 - \widehat{M}_{tl}}{1-\widehat{M}_{tl}},$$

where $\widehat{M}_{tl} = \widehat{\sum_k p_{tlk}^2} = \underset{s\text{-}s'\in\mathbf{S}_{\bar{l}}}{\mathrm{Wavg}}(\sum_k \hat{p}_{slk}\hat{p}_{s'lk}, N_sN_{s'})$. Using Equations (7) and (8), the values of $\hat{\tilde{\vartheta}}_x$, $\hat{\tilde{\vartheta}}_s$, $\hat{\tilde{\vartheta}}_t$,

$\hat{\tilde{\vartheta}}_{xx'}$ and $\hat{\tilde{\vartheta}}_{ss'}$ related to all loci can be calculated by

$$\hat{\tilde{\vartheta}}_x = \frac{\sum_l \mathrm{Avg}_{x\in s}(\sum_k \hat{p}_{xlk}^2) - \sum_l \widehat{M}_{tl}}{L - \sum_l \widehat{M}_{tl}},$$

$$\hat{\tilde{\vartheta}}_{xx'} = \frac{\sum_{s_1} N_{s_1}^2\, \hat{\tilde{\vartheta}}_s - N_t \hat{\tilde{\vartheta}}_x}{\sum_{s_1} N_{s_1}^2 - N_t},$$

$$\hat{\tilde{\vartheta}}_s = \frac{\sum_l \mathrm{Wavg}_{s\in t}(\sum_k \hat{p}_{slk}^2, N_s^2) - \sum_l \widehat{M}_{tl}}{L - \sum_l \widehat{M}_{tl}}, \tag{22}$$

$$\hat{\tilde{\vartheta}}_{ss'} = \frac{N_t^2 \hat{\tilde{\vartheta}}_t - \sum_{s_1} N_{s_1}^2\, \hat{\tilde{\vartheta}}_s}{N_t^2 - \sum_{s_1} N_{s_1}^2},$$

$$\hat{\tilde{\vartheta}}_t = \frac{\sum_l \sum_k \hat{p}_{tlk}^2 - \sum_l \widehat{M}_{tl}}{L - \sum_l \widehat{M}_{tl}}.$$

Finally, we explain the symbol $\mathbf{S}_{\bar{l}}$ in $\underset{s\text{-}s'\in\mathbf{S}_{\bar{l}}}{\mathrm{Wavg}}(\sum_k \hat{p}_{slk}\hat{p}_{s'lk}, N_s N_{s'})$ and how this is used to simplify the correlative calculation. The symbol $\mathbf{S}_{\bar{l}}$ is the collection consisting of the putative independent subpopulation pairs identified from all loci except for the l^{th} locus. The identification of pairs in $\mathbf{S}_{\bar{l}}$ is the same approach as that of pairs in $\mathbf{R}_{\bar{l}}$, but the indicator used is $S - \sum_k \hat{p}_{slk}\hat{p}_{s'lk}$ instead of $\frac{\sum_{l'\neq l}(\sum_k \hat{p}_{sl'k}\hat{p}_{s'l'k} - \sum_k \hat{p}_{tl'k}^2)}{\sum_{l'\neq l}(1-\sum_k \hat{p}_{tl'k}^2)}$, where S represents the sum $\sum_{l'}\sum_k \hat{p}_{sl'k}\hat{p}_{s'l'k}$. In fact, because there may be thousands of individuals in the total population, it will be involved in the calculation of $\hat{p}_{xlk}\hat{p}_{x'lk}$ for millions of individual pairs. Therefore, it may take longer than anticipated in order to calculate the values of $\sum_{l'\neq l}\sum_k \hat{p}_{sl'k}\hat{p}_{s'l'k}$. It is noteworthy that S is invariant when l varies, and by $\sum_{l'\neq l}\sum_k \hat{p}_{sl'k}\hat{p}_{s'l'k} = S - \sum_k \hat{p}_{slk}\hat{p}_{s'lk}$, we see that the calculation of $S - \sum_k \hat{p}_{slk}\hat{p}_{s'lk}$ is likely to be faster than the calculation of $\sum_{l'\neq l}\sum_k \hat{p}_{sl'k}\hat{p}_{s'l'k}$.

Single-locus correction

The estimates $\hat{\tilde{\theta}}$ and $\hat{\tilde{\vartheta}}$ can be used to correct the estimates of *F*-statistics at multiple loci. However, in real datasets there are often some missing values. This will result in the appearance of unequal sample sizes at distinct loci, which may bias the correction. Additionally, the estimates of single-locus *F*-statistics are often used to detect those loci that are targets of natural selection (Narum and Hess 2011). From these reasons, we developed an alternative single-locus weighting method based on Nei's (1973) *G*-statistics.

It is noteworthy that $1 - \widehat{M}_{sl}$ is an unbiased estimator for the expected heterozygosity H_{Sl} at the l^{th} locus, and so is $1 - \widehat{M}_{tl}$ for H_{Tl}. We can replace H_{Sl} by $1 - \widehat{\widehat{M}}_{sl}$ and H_{Tl} by $1 - \widehat{M}_{tl}$ for the situation of a single-locus, where $\widehat{\widehat{M}}_{sl}$ is a weighted average, that is $\widehat{\widehat{M}}_{sl} = \mathrm{Wavg}_{s_1\in t}(\widehat{M}_{s_1 l}, N_{s_1}^2)$. Then, according to Equation (16), the expressions of the unbiased corrections of the *G*-statistics for the single locus weighted scheme can be expressed as

$$\hat{F}_{IS,c} = \frac{1 - \widehat{\bar{M}}_{sl} - H_{Il}}{1 - \widehat{\bar{M}}_{sl}}, \qquad \hat{F}_{ST,c} = \frac{\widehat{\bar{M}}_{sl} - \widehat{M}_{tl}}{1 - \widehat{M}_{tl}} \qquad \text{and} \qquad \hat{F}_{IT,c} = \frac{1 - \widehat{M}_{tl} - H_{Il}}{1 - \widehat{M}_{tl}}.$$

Additionally, the corresponding multilocus estimates are

$$\hat{F}_{IS,c} = \frac{\sum_l (1 - \widehat{\bar{M}}_{sl} - H_{Il})}{\sum_l (1 - \widehat{\bar{M}}_{sl})}, \qquad \hat{F}_{ST,c} = \frac{\sum_l (\widehat{\bar{M}}_{sl} - \widehat{M}_{tl})}{\sum_l (1 - \widehat{M}_{tl})} \qquad \text{and} \qquad \hat{F}_{IT,c} = \frac{\sum_l (1 - \widehat{M}_{tl} - H_{Il})}{\sum_l (1 - \widehat{M}_{tl})}.$$

With a simple test, we find that the estimates of the two schemes are approximately equal, but the corrections for the single-locus weighted scheme are slightly better (the RMSEs are slightly lower). Therefore, we will use the single-locus weighting scheme in the subsequent simulations.

Simulations and comparisons

We used Monte-Carlo simulations to evaluate the statistical behaviors of previous *F*-statistics estimators and also for our new estimator. We designed a generalized scheme to calculate *F*-statistics, in which the NCI (North Carolina I) design is used to generate relative individuals, with the subpopulations nested within a region used to generate the correlated subpopulations, where the region from which each subpopulation originated is assumed to be unknown. We list the general procedures for the scheme below.

1. Generate a species at the ploidy level v. The samples will be genotyped at L unlinked neutral markers, such that there are K alleles at each locus.
2. Randomly draw the alleles frequencies $p_{t1}, p_{t2}, \dots, p_{tK}$ for total population t according to the Dirichlet distribution $\mathcal{D}(1, 1, \cdots, 1)$.
3. Randomly draw the alleles frequencies $p_{c1}, p_{c2}, \dots, p_{cK}$ for region c according to the Dirichlet distribution $\mathcal{D}(\gamma_c p_{t1}, \gamma_c p_{t2}, \dots, \gamma_c p_{tK})$, where $\gamma_c = 1/F_{CT} - 1$ (Pritchard et al. 2000) in which the *F*-statistic F_{CT} is set as $1 - \sqrt{1 - F_{ST}}$.
4. Randomly draw the alleles frequencies $p_{s1}, p_{s2}, \dots, p_{sK}$ for subpopulation s according to the Dirichlet distribution $\mathcal{D}(\gamma_s p_{c1}, \gamma_s p_{c2}, \dots, \gamma_s p_{cK})$, where $\gamma_s = 1/F_{SC} - 1$ (Pritchard et al. 2000) in which F_{SC} is set as $1 - \sqrt{1 - F_{ST}}$.
5. Generate the parents. For each subpopulation, N_f males are generated, and each male is mated with N_m females. The fathers' genotypes are generated by using Equation (A8) in Appendix E, with the mothers' genotypes are generated by using Equation (A9) in Appendix F.
6. Generate the offspring. For each subpopulation, N_o offspring are sampled from the offspring produced by each pair of parents. All individuals are genotyped, whose genotypes are used in the estimation of F_{IS} and F_{ST}.

For the estimation of both $\tilde{\theta}$ and $\tilde{\vartheta}$, $\widehat{M}_{sl}$ is obtained by identifying the putative non-relatives, while $\widehat{M}_{tl}$ is obtained by identifying the putative independent subpopulation pairs, with the

proportions of both the relatives and the dependent subpopulation pairs assumed to be at most 25%. The theoretical estimates of three previous estimators are calculated by Equations (13), (17) and (20).

The true values of $\tilde{\theta}_x$, $\tilde{\theta}_{xx'}$ and $\tilde{\vartheta}_{ss'}$ are determined by the pedigree and the population structure information, whose expressions are as follows:

$$\tilde{\theta}_x = \frac{1}{v} + \frac{v-1}{v} F_{IS},$$

$$\tilde{\theta}_{xx'} = \frac{N_f \binom{N_m N_o}{2}}{\binom{N_f N_m N_o}{2}} \tilde{\theta}_{fs} + \frac{N_f \binom{N_m N_o}{2} - N_f N_m \binom{N_o}{2}}{\binom{N_f N_m N_o}{2}} \tilde{\theta}_{hs},$$

$$\tilde{\vartheta}_{ss'} = \frac{C\binom{S_c}{2}}{\binom{S}{2}} F_{CT},$$

where the meanings of C, S and S_c can be seen from the section *Common symbols*, and $\tilde{\theta}_{fs}$ and $\tilde{\theta}_{hs}$ are, respectively, the expected kinship coefficients between full-sibs and half-sibs without the pedigree information, whose calculating formulas are given by Equation (A11), that is

$$\tilde{\theta}_{fs} = \frac{1}{2}\left(\tilde{\theta}_x + F_{IS}\right) \quad \text{and} \quad \tilde{\theta}_{hs} = \frac{1}{4}\left(\tilde{\theta}_x + 2F_{IS} + F_{IS}^2/\tilde{\theta}_x\right).$$

Moreover, by using the true values of $\tilde{\theta}$ and $\tilde{\vartheta}$, the theoretical estimates of three previous estimators are also calculated by Equations (13), (17) and (20).

Results

Single subpopulation

In this section, we only generate a single subpopulation within the total population (i.e. $C = S = 1$), and subsequently design two applications by sampling some individuals from this subpopulation.

APP1 is designed to validate the theoretical bias of $\hat{F}_{IS}$ we derived for the three existing estimators and validate the unbiasedness of our corrected estimator, where 5,000 SNPs are used to obtain the simulated bias. The parameter configurations are $v \in \{2, 4\}$, $F_{ST} = 0$, $F_{IS} \in \{0, 0.05, \cdots, 1\}$ at an interval of 0.05, $L = 5000$ and $K = 2$.

APP2 is designed to compare the bias and the RMSE of $\hat{F}_{IS}$ of different estimators. Two kinds of markers are used for this application. For SNPs, $L \in \{50, 100, \cdots, 500\}$ at an interval of 50 is used with the remaining parameter configurations as stated above. For microsatellites (denoted by SSRs), the parameter configurations are $v \in \{2, 4\}$, $F_{ST} = 0$, $F_{IS} \in \{0.1, 0.3, 0.6\}$, $L \in \{5, 10, \cdots, 50\}$ at an interval of 5 and $K = 6$. Each combination of parameters is simulated 1000 times.

Two mating systems are each simulated for both applications. For monogamy, we set $N_f = 4$, $N_m = 1$ and $N_o = 8$ (there are a total of 32 offspring, including full-sibs and non-relatives). For polygamy, we set $N_f = 4$, $N_m = 8$ and $N_o = 1$ (there are a total of 32 offspring, including half-sibs and

non-relatives).

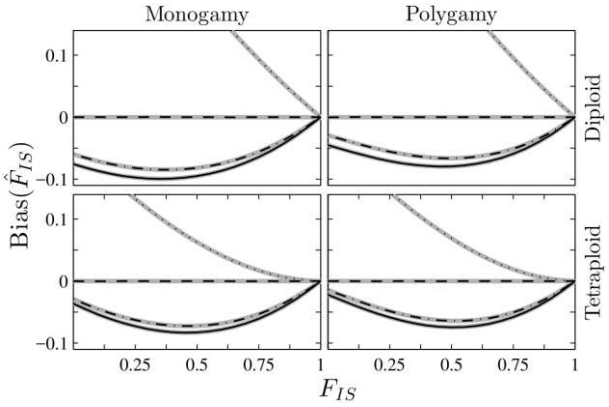


Figure 1. The bias of $\hat{F}_{IS}$ as a function of the true F_{IS}. Each row shows a different ploidy level (diploid or autotetraploid) and each column shows a different mating system (monogamy or polygamy). The dotted, solid, dash-dotted and dashed lines denote in turn the biases of Wright's (1951), Nei's (1973), Weir & Cockerham's (1984) and our corrected F_{IS} estimators. Each point on each of the curves was obtained by 1000 simulations at 5000 diallelic loci. Each thick gray line represents the theoretical bias for each estimator method.

The results of APP1 are shown in Figure 1. It is clear that the simulated biases (black lines) concur with the theoretical values (gray lines). According to Equation (11), the expectation of Wright's (1951) F_{IS} estimator is $\frac{\tilde{\theta}_x-\tilde{\theta}_s}{1-\tilde{\theta}_s}$, and hence it is positively biased. Nei's (1973) and Weir & Cockerham's (1984) estimators are negatively biased, with the former being more biased than latter. Our corrected estimator is unbiased. The absolute biases in the polygamous mating system are relatively lower because the values of $\tilde{\theta}_s$ and $\tilde{\theta}_{xx'}$ are smaller than those values for the monogamous mating system. For tetrasomic inheritance, the absolute bias is also relatively lower because the value of θ_x in polyploids is smaller at the same level of inbreeding, which is equal to $1/v$ in the outbred populations.

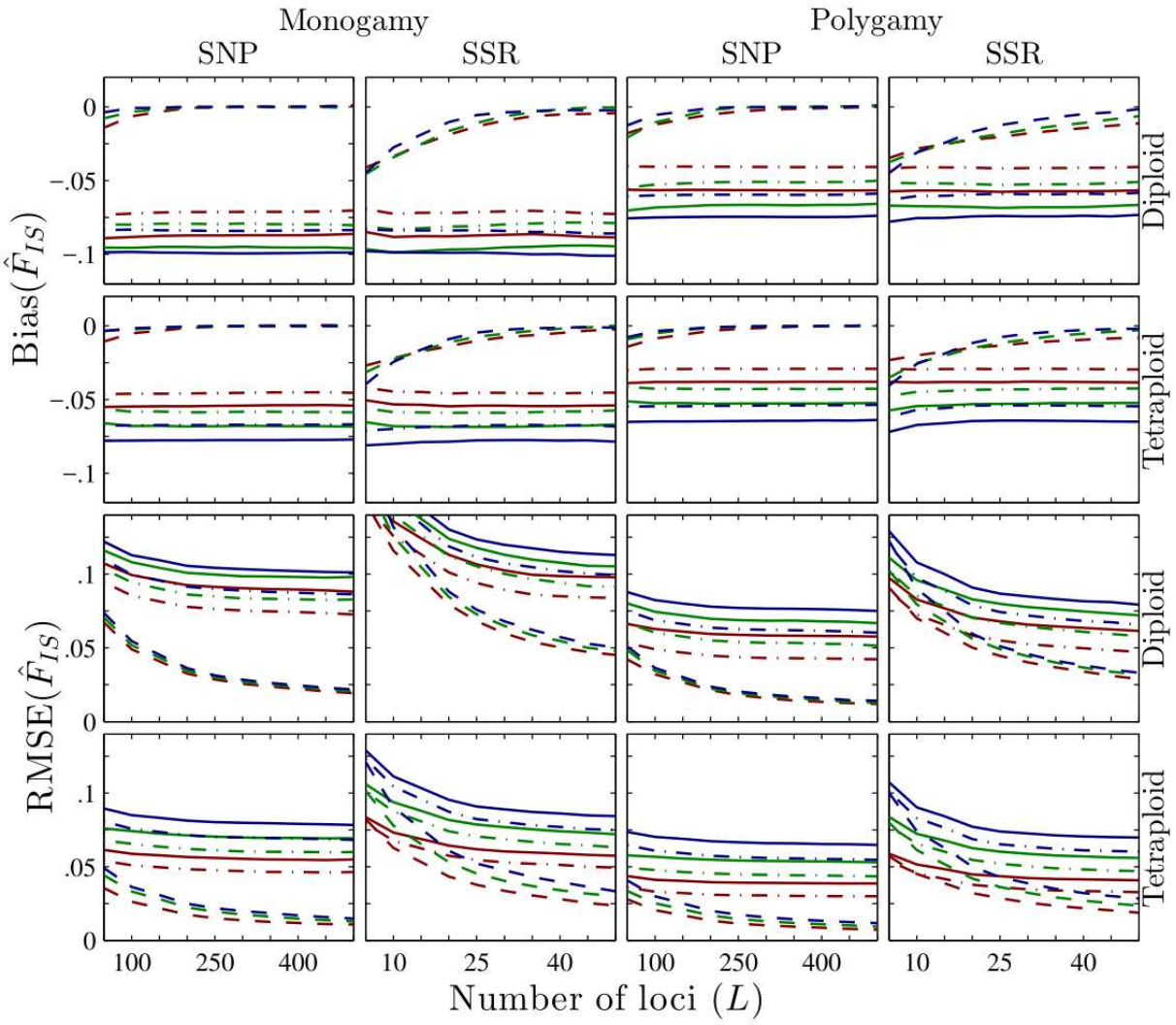


Figure 2. The bias or the RMSE of $\hat{F}_{IS}$ as a function of the number of loci L. Each row shows a ploidy level (diploid or autotetraploid) and each column shows a mating system (monogamy or polygamy) together with a marker type (SNP or SSR). The solid, dash-dotted and dashed lines denote in turn the biases (or the RMSEs) of Nei's (1973), Weir & Cockerham's (1984) and our corrected F_{IS} estimators. The red, green and blue lines denote in turn the biases when the true F_{IS} is equal to 0.1, 0.3 and 0.6. Each point on each curve is obtained by 1000 simulations at 5000 diallelic loci.

The results of APP2 are shown in Figure 2 and are similar to those for APP1. The biases and the RMSEs for the polygamous mating system are relatively lower than those for monogamy, and they are also similarly reduced for tetrasomic inheritance. For SNP markers, each bias is close to an asymptote when $L \geq 150$. For SSR markers, each bias for both Nei's and Weir and Cockerham's estimators quickly reaches close to an asymptote, whilst the bias for our new, corrected estimator is reduced at a lower rate, and does not reach close to an asymptote until $L \geq 40$. If the RMSE of our corrected estimator is regarded as a function of $\log L$, it decreases linearly as L increases. Although there are 6 alleles at each microsatellite, the RMSE is approximately equal to that of $2L$ SNPs.

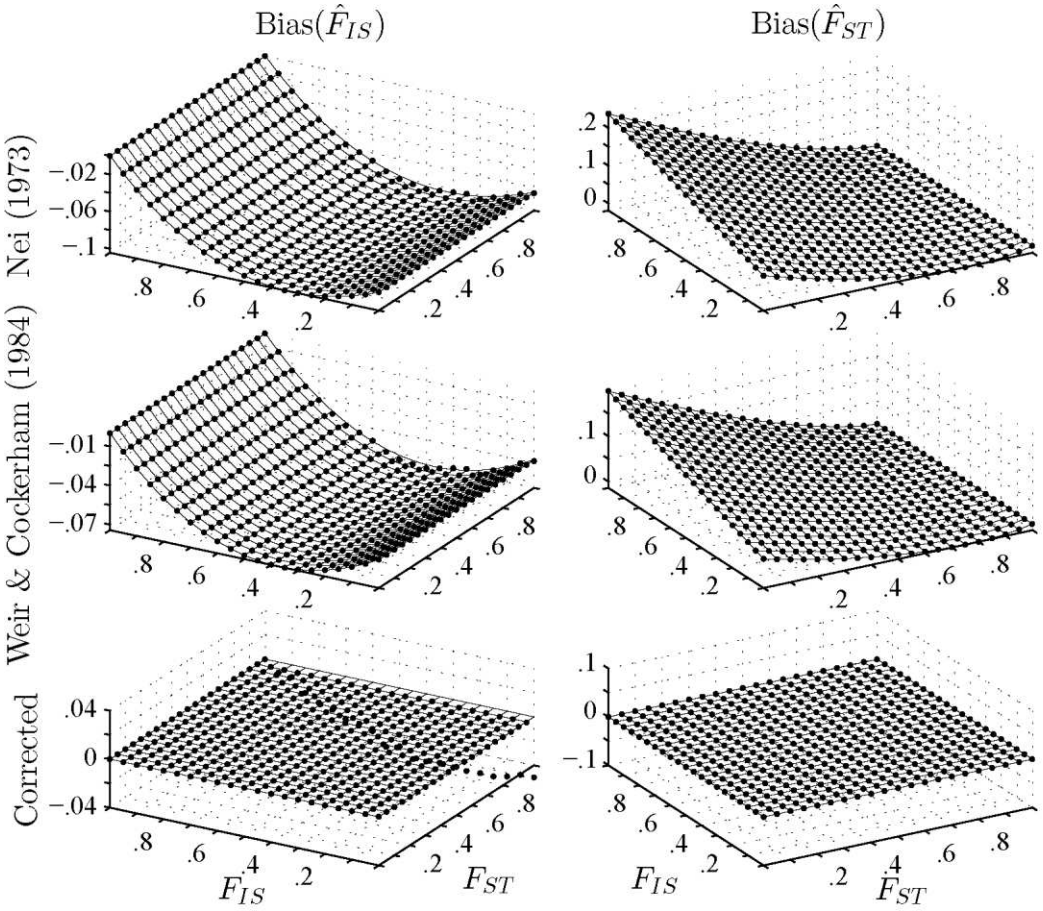
Multiple subpopulations

Here, we generate four regions within a single, larger population, with each region containing one or two subpopulations. We will also design three applications by sampling some individuals from each of these subpopulations.

APP3 is designed to validate the theoretical biases of $\hat{F}_{IS}$ and $\hat{F}_{ST}$ we derived from previous estimators and to validate any bias for our corrected estimator. The parameter configurations are $v \in \{2, 4\}$, $F_{ST} \in \{0, 0.05, \cdots, 0.95, 0.999\}$ at an interval of 0.05 except for the final interval, $F_{IS} \in \{0, 0.05, \cdots, 1\}$ at an interval of 0.05, $L = 5000$ and $K = 2$. Only the monogamous mating system is simulated, under the following configurations: $N_f = 4$, $N_m = 1$ and $N_o = 4$.

APP4 is designed to compare the biases and the RMSEs of $\hat{F}_{ST}$. Similar to APP2, for this application two kinds of markers are used and both mating systems are simulated. The parameter configurations are as follows: $v \in \{2, 4\}$, $F_{ST} = \{0.05, 0.1, 0.2\}$, $F_{IS} = 0.3$. In addition, for SNPs, $L \in \{50, 100, \cdots, 500\}$ and $K = 2$; for the SSRs, $L \in \{5, 10, \cdots, 50\}$ and $K = 6$. Also, for monogamy, $N_f = 4$, $N_m = 1$ and $N_o = 4$; for polygamy, $N_f = 4$, $N_m = 4$ and $N_o = 1$. Each combination of parameters is simulated 1000 times.

APP5 is designed to simulate bias for independent or non-independent samples each with a small sample size. In for application, the widely used estimator described by Hedrick (2005), G'_{ST} is also for comparison with the other estimators in the study. For the dependent subpopulations, we set $C = 4$ and $S_c = 2$ for each region c; for the independent subpopulations nested within the regions, we set $C = 4$ and $S_c = 1$ for each region c. For independent individuals, the configurations are $N_f = 16$, $N_m = 1$ and $N_o = 1$; for non-independent individuals, the configurations are $N_f = 4$, $N_m = 4$ and $N_o = 1$. Moreover, F_{IS} is fixed at 0 and the remaining parameter configurations are identical to those used for APP3.



536

537 **Figure 3.** The bias of $\hat{F}_{IS}$ (or $\hat{F}_{ST}$) as a binary function of F_{IS} and F_{ST} for Nei's (1973), Weir &
538 Cockerham's (1984) and our corrected estimators for diploids. The mesh plots denote the theoretical
539 biases and the black dots denote the simulated biases.

540 The results for APP3 using both diploids and autotetraploids are shown in Figures 3 and S1,
541 respectively. According to Equations (13) and (17), the expectation of Wright's (1951) $\hat{F}_{ST}$ estimate is
542 identical to that of Nei's (1973) $\hat{G}_{ST}$, such that the corresponding results of Wright's (1951) estimator are
543 not shown. It is clear that the simulated biases fit well with the theoretical biases except when $F_{ST} =$
544 0.999 (Figure 3). $\hat{F}_{IS}$ is inaccurate at $F_{ST} = 0.999$ because most loci are fixed in the subpopulations,
545 and few loci are used for its estimation.

546 Our corrected estimator is still unbiased. The biases of $\hat{F}_{IS}$ of previous estimators are negative,
547 while the biases of $\hat{F}_{ST}$ of previous estimators are positive. The bias of $\hat{F}_{IS}$ is affected by F_{IS} but is not
548 relevant to F_{ST}, while the bias of $\hat{F}_{ST}$ is influenced by both F_{IS} and F_{ST}, which is in accordance with
549 both Equations (17) and (20). The plots of the theoretical and simulated biases of both estimators have
550 similar shapes, but it is noteworthy that both scales differ on their z-axes (Figure 3). Nei's (1973)

estimates are more biased than those of Weir & Cockerham (1984).

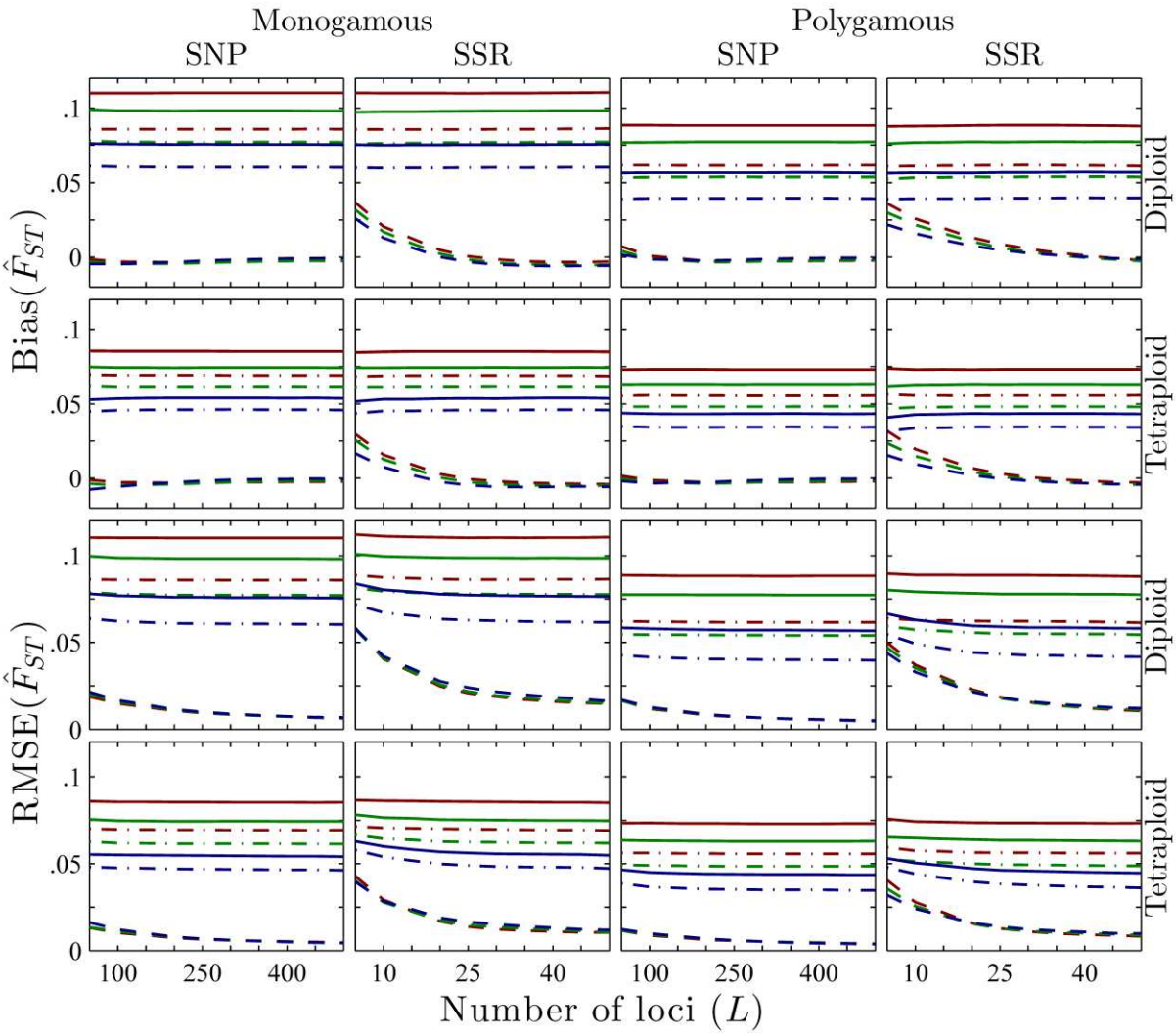


Figure 4. The bias and the RMSE of $\hat{F}_{ST}$ as a function of the number of loci L (F_{IS} is fixed at 0.3). The rows, columns and line styles are as for Figure 2 except for the line colors. The red, green and blue lines denote the biases when the true F_{ST} is equal to 0.05, 0.1 and 0.2, respectively. Each point on each curve was obtained by 1000 simulations at 5000 diallelic loci.

The results of APP4 are shown in Figure 4. The bias and the RMSE for polygamy are both lower than for monogamy, and both deviations for autotetraploids are lower than for diploids. The biases of $\hat{F}_{ST}$ for previous methods of estimators may reach 0.1, but are greatly reduced with our new corrected estimator method. However, our new estimator still shows bias when L is small. The bias of our corrected estimator for SNP (or the SSR) markers is close to zero when $L \geq 150$ (or $L \geq 30$). Compared with $\hat{F}_{IS}$ (Figure 2), the RMSE of $\hat{F}_{ST}$ is smaller when L is also small but decreases more gradually than for the other methods.

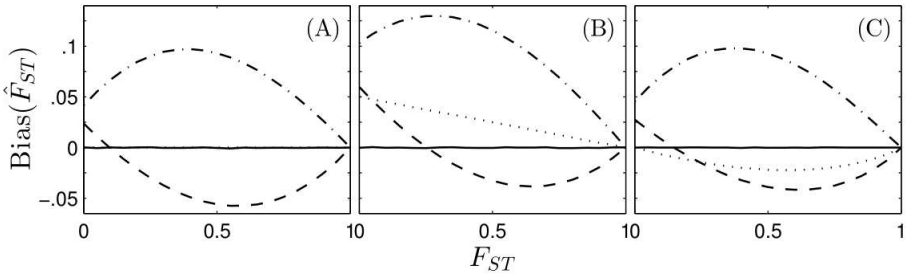


Figure 5. The bias of $\hat{F}_{ST}$ as a function of F_{ST} under one of the following three conditions: (A) the individuals and the subpopulations are both independent; (B) the individuals are non-independent, but the subpopulations are independent; (C) the individuals are independent, but the subpopulations are non-independent. The dashed, dash-dotted, dotted and solid lines denote in turn the biases of Nei's (1973), Hedrick's (2005) G'_{ST}, Weir & Cockerham's (1984) and our new corrected estimators.

The results of APP5 are shown in Figure 5. This shows that Nei's (1973) estimator is initially positively biased and then negatively biased under all conditions; Hedrick's (2005) G'_{ST} estimator is positively biased under all conditions; Weir & Cockerham's (1984) estimator is unbiased for condition (A) but biased under both conditions (B) and (C). Our new corrected estimator is unbiased for all three conditions used for the simulation.

Discussion

Bias

By using the extended definitions of θ and ϑ, we are able to extend three previous *F*-statistics estimators (Nei 1973; Weir and Cockerham 1984; Wright 1951) into our model, and explain any differences and biases of estimates for each of these estimators. The corrected estimates of these three previous *F*-statistic estimators are identical.

Previous *F*-statistics estimators can be influenced by non-independent samples at both the individual and the subpopulation level (see Figure 5B or 5C). Nei's (1973) estimator is biased even when individual and the subpopulation are both independent (see Figure 5A). One reason causing such bias is the assumptions of $\tilde{\theta}_s$ and $\tilde{\vartheta}_t$. By Equations (14) and (17), Nei's (1973) method assumes that $\tilde{\theta}_s = \tilde{\vartheta}_t = 0$. From Equation (2), it can be found $\tilde{\theta}_s$ is greater than zero because $\tilde{\theta}_x \geq 1/v$ and $\tilde{\theta}_{xx'} \geq 0$. Similarly, by Equation (3), $\tilde{\vartheta}_t$ is also greater than zero because $\tilde{\vartheta}_s$ is positive and $\tilde{\vartheta}_{ss'}$ is non-negative. Hence the assumptions that $\tilde{\theta}_s = 0$ and $\tilde{\vartheta}_t = 0$ are both unrealistic. Another cause of such bias is the finite number of subpopulations. For example, in Figure 5A, a negative bias may occur when S_c is small.

Equations (18) and (20) show that Weir & Cockerham's (1984) estimator assumes that $\tilde{\theta}_{xx'} = 0$,

591 $\tilde{\vartheta}_{ss'} = 0$ and $\tilde{\vartheta}_{xx'} = F_{ST}$, both of which are more realistic. Therefore, Weir & Cockerham's (1984)
592 estimator is less biased the other two previous estimators for non-independent samples (Figures 1 and
593 3), and is unbiased for independent samples (Figure 5A). When the true F_{ST} is zero or small, because
594 of the unbiased nature of Weir & Cockerham's (1984) estimator for independent samples, the values of
595 $\hat{F}_{ST}$ may be negative so as to result in $\mathrm{E}(\hat{F}_{ST}) = F_{ST}$. Similarly, for our new corrected estimator, the
596 values of $\hat{F}_{ST}$ may also be negative for both non-independent and independent samples.
597 There is an additional source of bias for all these estimators, which is the correlation between the
598 numerator and the denominator (e.g., in Equation (9)). Such bias is present in all estimators (see Figures
599 2 and 4) but is highest in our new estimator. The summing of all numerators and the summing of all
600 denominators, respectively, can eliminate this correlation for approximately 10 loci for the previous
601 estimators and for approximately 40 loci for our new estimator.

602 ## Hedrick's G'_{ST} estimator

603 For finite subpopulations, Hedrick's (2005) G'_{ST} estimator behaves similarly to Nei's (1973) G_{ST}
604 estimator. With only few subpopulations (e.g., $S = 2$), all existing estimators will underestimate the
605 values of F_{ST}. This is because the total population is assumed to be a union of sampled subpopulations,
606 and this results in an underestimate of the genetic diversity of the total population and hence the values
607 of F_{ST}. To solve this problem, Hedrick (2005) developed the G'_{ST} estimator, which can be obtained from
608 G_{ST} divided by the its theoretical maximum, whose expression for the multi-locus estimate is:

609 $$G'_{ST} = \frac{\sum_l (H_{Tl} - H_{Sl})(S - 1 + H_{Sl})}{\sum_l H_{Tl}(S-1)(1 - H_{Sl})}.$$

610 Unfortunately, each value of G'_{ST} is always greater than the corresponding value of G_{ST}, so this
611 estimator cannot correct for the overestimation of F_{ST} for non-independent subpopulations. In
612 additional, G'_{ST} is still biased even if all subpopulations are independent (Figure 5A). In contrast, our
613 corrected estimator is unbiased under all three conditions, and can simultaneously correct the
614 downward bias caused by limited subpopulation size and the upward bias caused by non-
615 independence of subpopulations (Figures 5B and 5C).

616 ## Pairwise F_{ST}

617 The pairwise F_{ST} is usually used as an alternative to the measurement of genetic distance. The
618 range can be converted from $[0,1)$ into $[0, +\infty)$ by Slatkin's (1995) transformation, $F_{ST}/(1 - F_{ST})$ or
619 Reynolds *et al.*'s (1983) transformation, $-\ln(1 - F_{ST})$. Our new method can also be used for pairwise
620 F_{ST} estimation, but the estimation of $\tilde{\vartheta}$ should be based on putative non-relatives instead of putative
621 independent subpopulations because there will be only two subpopulations. The corrected pairwise
622 F_{ST} will be an unbiased estimator for the differentiation from the common ancestor of these two

subpopulations.

Perspective

Many population genetics methods use the current population as a reference population and fail to account for relationships between samples. This is inconsistent with natural populations. For example, for Nei's (1973) estimator, the assumption that $\tilde{\theta}_s = 0$ is unrealistic. In reality, various parameters are likely to be affected by non-independence of samples. For example, the average kinship coefficient estimate between individuals is close to zero (Ritland 1996). Some parameters can be corrected by using our new method (in the sense of using $\widehat{M}_{sl}$ to estimate $\sum_k^K p_{sk}^2$), such as for expected heterozygosity, the kinship coefficient (Weir 1996) and the relatedness coefficient (converted from kinship coefficient, Huang et al. 2015).

Statements and Declarations

Funding

This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB31020302), the National Natural Science Foundation of China (31730104, 32170515, 31770411, 32070453), and the Innovation Capability Support Program of Shaanxi (2021KJXX-027). DWD is supported by a Shaanxi Province Talents 100 Fellowship and KH is supported by a scholarship from China Scholarship Council.

Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

Author contributions

BGL initiated and led the project, KH designed research, KH and BY constructed the model, JCA and YHL performed simulations, and KH and DWD wrote the paper.

Acknowledgment

KH would like to thank Prof. Kermit Ritland for providing a visiting professor position at the University of British Columbia.

Data Accessibility

POLYGENE is a user-friendly software and can be run on 64bit Windows/Linux/Mac OS X platform. The binary executable files, user manual and source code of POLYGENE V1.4 are available on GitHub (https://github.com/huangkang1987/polygene). POLYGENE was developed using C++ and C# and requires .net framework V4.0 runtime library.

653 The simulation program, the source code (written in C++ and C#) and the script to generate the
654 figures (written in MATLAB) are provided as supplementary materials and can be downloaded via a link
655 on the online version of this manuscript.

656 References

657 Blyton MDJ, Shaw RE, Peakall R, Lindenmayer DB, Banks SC, sociobiology (2016) The role of
658 relatedness in mate choice by an arboreal marsupial in the presence of fine-scale genetic structure.
659 Behavioral Ecology 70:313-321
660 Cockerham CC (1969) Variance of gene frequencies. Evolution 23:72-84
661 Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances
662 among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics
663 131:479-491
664 Gotoh A, Dansho M, Dobata S, Ikeshita Y, Ito F (2017) Social structure of the polygynous ant,
665 *Crematogaster osakensis*. Insectes Sociaux 64:123-131
666 Guo ST, Huang K, Ji WH, Garber PA, Li BG (2015) The role of kinship in the formation of a primate
667 multilevel society. American Journal of Physical Anthropology 156:606-613
668 Hamilton M (2009) Population genetics. John Wiley & Sons, Oxford
669 Hardy OJ, Vekemans X (2002) SPAGeDi: a versatile computer program to analyse spatial genetic
670 structure at the individual or population levels. Mol Ecol Notes 2:618-620
671 Hedrick PW (2005) A standardized genetic differentiation measure. Evolution 59:1633-1638
672 Huang K, Guo ST, Shattuck MR, Chen ST, Qi XG, Zhang P, Li BG (2015) A maximum-likelihood
673 estimation of pairwise relatedness for autopolyploids. Heredity 114:133-142
674 Huang K, Mi R, Dunn DW, Wang TC, Li BG (2018) Performing parentage analysis in the presence of
675 inbreeding and null alleles. Genetics 210:1467-1481
676 Huang K, Wang TT, Dunn DW, Zhang P, Sun HJ, Li BG (2020) A generalized framework for AMOVA
677 with multiple hierarchies and ploidies. Integr Zool:10.1111/1749-4877.12460
678 Hudson RR, Slatkin M, Maddison W (1992) Estimation of levels of gene flow from DNA sequence data.
679 Genetics 132:583-589
680 Jost L (2008) G_{ST} and its relatives do not measure differentiation. Molecular Ecology 17:4015-4026
681 Kalinowski ST, Taper ML, Marshall TC (2007) Revising how the computer program CERVUS
682 accommodates genotyping error increases success in paternity assignment. Molecular Ecology
683 16:1099-1106
684 Leinonen T, McCairns RS, O'hara RB, Merilä J (2013) Q_{ST}–F_{ST} comparisons: evolutionary and ecological
685 insights from genomic heterogeneity. Nature Reviews Genetics 14:179
686 Li CC, Horvitz DG (1953) Some methods of estimating the inbreeding coefficient. Am J Hum Genet 5:107
687 Lynch M, Crease TJ (1990) The analysis of population survey data on DNA sequence variation.
688 Molecular Biology and Evolution 7:377-394
689 Lynch M, Ritland K (1999) Estimation of pairwise relatedness with molecular markers. Genetics
690 152:1753-1766
691 Lynch M, Walsh B (1998) Genetics and analysis of quantitative traits. Sinauer Sunderland, MA
692 Maher KH, Eberhart‐Phillips LJ, Kosztolányi A, Remedios Nd, Carmona‐Isunza MC, Cruz‐López
693 M, Zefania S, St Clair JJ, Alrashidi M, Weston MA (2017) High fidelity: extra‐pair fertilisations
694 in eight *Charadrius plover* species are not associated with parental relatedness or social mating

system. Journal of Avian Biology 48:910-920
Narum SR, Hess JE (2011) Comparison of F_{ST} outlier tests for SNP loci under selection. Mol Ecol Resour 11:184-194
Nei M (1973) Analysis of gene diversity in subdivided populations. Proc Natl Acad Sci 70:3321-3323
Nomura T (2008) Estimation of effective number of breeders from molecular coancestry of single cohort sample. Evol Appl 1:462-474
Ochoa A, Storey JD (2021) Estimating FST and kinship for arbitrary population structures. PLoS genetics 17:e1009241
Pascov CM, Nevill PG, Elliott CP, Majer JD, Anthony JM, Krauss SLJO (2015) The critical role of ants in the extensive dispersal of *Acacia* seeds revealed by genetic parentage assignment 179:1123-1134
Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945-959
Reynolds J, Weir BS, Cockerham CC (1983) Estimation of the coancestry coefficient: basis for a short-term genetic distance. Genetics 105:767-779
Ritland K (1996) A marker‐based method for inferences about quantitative inheritance in natural populations. Evolution 50:1062-1073
Ritland K (2000) Marker‐inferred relatedness as a tool for detecting heritability in nature. Molecular Ecology 9:1195-1204
Rousset F (1996) Equilibrium values of measures of population subdivision for stepwise mutation processes. Genetics 142:1357-1362
Santos Jd, Varassin IG, Muschner VC, Ovaskainen O (2018) Estimating seed and pollen dispersal kernels from genetic data demonstrates a high pollen dispersal capacity for an endangered palm species. American Journal of Botany 105:1802-1812
Slatkin M (1995) A measure of population subdivision based on microsatellite allele frequencies. Genetics 139:457-462
Spitze K (1993) Population structure in *Daphnia obtusa*: quantitative genetic and allozymic variation. Genetics 135:367-374
Wang JL (2002) An estimator for pairwise relatedness using molecular markers. Genetics 160:1203-1215
Wang JL (2004) Sibship reconstruction from genetic data with typing errors. Genetics 166:1963-1979
Wang JL (2019) Pedigree reconstruction from poor quality genotype data. Heredity:1
Weir BS (1996) Genetic data analysis II: methods for discrete population genetic data. Sinauer Associates, Sunderland
Weir BS (2012) Estimating *F*-statistics: a historical view. Philosophy of Science 79:637-643
Weir BS, Cockerham CC (1984) Estimating *F*‐statistics for the analysis of population structure. Evolution 38:1358-1370
Wright S (1951) The genetical structure of populations. Ann Eugen 15:323-354
Wright S (1965) The interpretation of population structure by *F*-statistics with special regard to systems of mating. Evolution 19:395-420

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Fst21supp.docx
- data.7z