Estimating Relatedness in the Presence of Null Alleles

Kang Huang,* Kermit Ritland,† Derek W. Dunn,* Xiaoguang Qi,* Songtao Guo,* and Baoguo Li*,[1]
*Shaanxi Key Laboratory for Animal Conservation, and College of Life Sciences, Northwest University, Xi'an, Shaanxi 710069, China, and †Department of Forest and Conservation Sciences, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada
ORCID ID: 0000-0002-8357-117X (B.L.)

ABSTRACT Studies of genetics and ecology often require estimates of relatedness coefficients based on genetic marker data. However, with the presence of null alleles, an observed genotype can represent one of several possible true genotypes. This results in biased estimates of relatedness. As the numbers of marker loci are often limited, loci with null alleles cannot be abandoned without substantial loss of statistical power. Here, we show how loci with null alleles can be incorporated into six estimators of relatedness (two novel). We evaluate the performance of various estimators before and after correction for null alleles. If the frequency of a null allele is <0.1, some estimators can be used directly without adjustment; if it is >0.5, the potency of estimation is too low and such a locus should be excluded. We make available a software package entitled PolyRelatedness V1.6, which enables researchers to optimize these estimators to best fit a particular data set.

KEYWORDS relatedness coefficient; null alleles; method-of-moment; maximum likelihood

PAIRWISE relatedness is central to studies of population genetics, quantitative genetics, behavioral ecology, and sociobiology (*e.g.*, Mattila *et al.* 2012). The relatedness coefficient between individuals can be calculated from a known pedigree (Karigl 1981). In the absence of pedigree information, this coefficient can be estimated by using genetic markers.

When codominant genetic markers are used to estimate relatedness, some genotyping errors may occur. For example, in polymerase chain reaction-based markers, allelic dropout, false allele, and null allele errors are common, especially with microsatellites. In allelic dropout, the low quality of the DNA template inhibits the amplification of one of the two alleles in a heterozygote, causing the heterozygote to be observed as a homozygote (Taberlet *et al.* 1996). False alleles are artificial alleles produced during amplification, resulting in homozygotes mistyped as heterozygotes (Taberlet *et al.* 1996). Null alleles are alleles that cannot be detected because of mutation, often within the primer site (Brookfield 1996). These incorrect genotypes can cause serious problems in the estimation of genetic relationships. For instance, in parentage analysis, genotyping error can mistakenly reject the true father due to an observed lack of shared alleles with the offspring (Blouin 2003). Similarly, when pairwise genetic relatedness is estimated, genotyping error can cause true relatives to be omitted. This is the subject of this article.

The first two error types can be eliminated by multiple-tube methods because the errors appear at random (Taberlet *et al.* 1996; He *et al.* 2011). Genotyping errors due to null alleles cannot be resolved in this manner. Null alleles are pervasive in microsatellite markers and many studies report null alleles (*e.g.*, Kokita *et al.* 2013). Solutions for dealing with null alleles are limited, but when null alleles are detected, they can be eliminated by redesigning the primers or by developing new microsatellite loci. However, both solutions involve additional time and material costs and are not freely available (Wagner *et al.* 2006).

There are a number of estimators that have been developed using different assumptions, each with varying efficiencies under different conditions. A single estimator cannot fulfill the requirements of all studies. In general, estimators can be classified into two categories: (i) method of moment and (ii) maximum likelihood.

Method-of-moment estimators equate sample moments with unobservable population moments and are generally unbiased but they are not optimal in terms of statistical

doi: 10.1534/genetics.114.163956
Manuscript received March 16, 2015; accepted for publication October 17, 2015; published Early Online October 21, 2015.
Supporting information is available online at www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.163956/-/DC1.
[1]Corresponding author: Shaanxi Key Laboratory for Animal Conservation, and College of Life Sciences, Northwest University, Xi'an, Shaanxi 710069, China. E-mail: baoguoli@nwu.edu.cn

efficiency (Huang *et al.* 2015a). These estimators have been developed to estimate only the relatedness coefficient (*e.g.*, Queller and Goodnight 1989; Li *et al.* 1993; Loiselle *et al.* 1995; Ritland 1996) or both relatedness and four-gene coefficients simultaneously (*i.e.*, Lynch and Ritland 1999; Wang 2002; Thomas 2010).

The maximum-likelihood approach models the probability of observing a specific pairwise allele pattern, given two "higher-order" coefficients (ϕ and Δ) and allele frequencies (Milligan 2003; Anderson and Weir 2007). By searching the parameter space for values of ϕ and Δ that maximize the probability of the genotype pattern observed, maximum-likelihood values can be determined for these parameters.

Here, we modify four existing relatedness estimators (Lynch and Ritland 1999; Wang 2002; Anderson and Weir 2007; Thomas 2010). We also present two new estimators based on two of these previously published works (Lynch and Ritland 1999; Wang 2002) to estimate relatedness coefficients using codominant markers with null alleles. First, we compare the performances of all six estimators before and after correction, to evaluate the influence of null alleles on each estimator under different frequencies of null alleles. Second, we calculate the minimum number of loci needed to achieve a predefined criterion of statistical power. Finally, we simulate a finite population with inbreeding and genetic drift to imitate natural conditions and evaluate the efficiency of each estimator. We make available the software entitled polyrelatedness V1.6, to help other researchers calculate and compare these estimators and simulation functions.

Theory and Modeling

In this article, we adopt the identical-by-descent (IBD) definition of relatedness, which is the probability that an allele sampled from one individual at a locus is IBD to an allele from another individual. This definition is often asymmetric in inbred populations (the two reciprocal relatedness coefficients from different individuals may be unequal), with the geometric mean being equivalent to Wright's (1921) definition in diploids (Huang *et al.* 2015b). The genetic correlation between a pair of individuals (x and y) sampled from an outbred population has three modes: (i) each allele in x is IBD to one allele in y, (ii) a single allele in x is IBD to one allele in y, and (iii) no alleles in x are IBD to any allele in y. If the probabilities of occurrences of the first and second events are denoted as Δ (four-gene coefficient) and ϕ (two-gene coefficient), respectively, the relatedness coefficient r between individuals x and y can be expressed as

$$r = \Delta + \frac{\phi}{2}. \tag{1}$$

For example, in an outbred population, parent–offspring pairs share an IBD allele with a probability of 1, so $\phi = 1$ and $\Delta = 0$; full sibs share one or two IBD alleles with the probability 1/2 or 1/4, so $\phi = 1/2$ and $\Delta = 1/4$. Most relatedness estimators are assumed to have the following conditions: (i) a population is large, outbred, and panmictic; (ii) the locus is autosomal and inheritance Mendelian; and (iii) the loci used are unlinked and are not in linkage disequilibrium (*i.e.*, Lynch and Ritland 1999; Wang 2002; Milligan 2003). Under these assumptions, the presence of alleles and genotypes is random and concurs with the Hardy–Weinberg equilibrium. Thus, the probability of occurrence for a genotype (*i.e.*, Wang 2002; Milligan 2003) or a genotype conditioned on another genotype (*i.e.*, Lynch and Ritland 1999) can be expressed as a function of Δ and φ. Subsequently, linear algebra methods (method-of-moment estimators) or numerical algorithms (maximum-likelihood estimators) can be used to solve $\hat{\Delta}$ and $\hat{\phi}$.

The presence of null alleles can be realized in two ways: (i) the observed pairwise genotypes will become either more or less similar than their true values, and (ii) the sum of the frequencies of visible alleles becomes unified, but the ratio among frequencies of visible alleles is unchanged. These two effects can cause an over- or underestimation of the relatedness coefficient. Some methods can estimate the frequency of null alleles (*e.g.*, Brookfield 1996; Vogl *et al.* 2002; van Oosterhout *et al.* 2004; Kalinowski *et al.* 2006; Hall *et al.* 2012; Dabrowski *et al.* 2013). We thus model the probability that a genotype pair (Anderson and Weir 2007) or a genotype pattern (Lynch and Ritland 1999) or a genotype pair pattern (Wang 2002; Thomas 2010) is observed under the assumption that the true frequencies of alleles are already known. Here, we describe briefly six estimators and give a correction of null alleles for each estimator.

Data availability

File S1 contains (i) derivations of Expressions (3a), (3b), (9a) (9b), (12), (14a) and (14b), (ii) derivations of $\hat{r}$. and $\hat{\Delta}$ of corrected Lynch and Ritland's (1999) estimator, (iii) derivations of biases of method-of-moment estimators before and after correction, and (iv) a summary of expressions of method-of-moment estimators.

Lynch and Ritland's (1999) Estimator

Lynch and Ritland's (1999) estimator models the probability of an observed proband genotype conditioned on a reference genotype. To classify the degree of similarity for a pair of genotypes, this estimator employs a similarity index denoted by S and defined by

$$S = \begin{cases} 1 & \text{if } A_iA_i - A_iA_i \text{ or } A_iA_j - A_iA_j, \\ 3/4 & \text{if } A_iA_i - A_iA_j, \\ 1/2 & \text{if } A_iA_j - A_iA_k, \\ 0 & \text{otherwise}, \end{cases} \tag{2}$$

where A_i denotes the ith allele, and A_i, A_j, and A_k are different. By modeling the probability of each observed genotype (*i.e.*, G_y) when the reference genotype (*i.e.*, G_x) is homozygous or heterozygous, the following systems of linear equations with Δ and ϕ as the unknowns are established (for the derivation, see Supporting Information, File S1): for the case of homozygotes,

$$\begin{bmatrix} \Pr(G_y = A_iA_i \mid G_x = A_iA_i) \\ \Pr(G_y = A_iA_x \mid G_x = A_iA_i) \end{bmatrix} = \begin{bmatrix} p_i^2 \\ 2p_ip_x \end{bmatrix} + \begin{bmatrix} 1-p_i^2 & p_i - p_i^2 \\ -2p_ip_x & p_x - 2p_ip_x \end{bmatrix} \begin{bmatrix} \Delta \\ \phi \end{bmatrix}; \tag{3a}$$

for the case of heterozygotes,

$$\begin{bmatrix} \Pr(G_y = A_iA_i \mid G_x = A_iA_j) \\ \Pr(G_y = A_jA_j \mid G_x = A_iA_j) \\ \Pr(G_y = A_iA_j \mid G_x = A_iA_j) \\ \Pr(G_y = A_iA_x \mid G_x = A_iA_j) \\ \Pr(G_y = A_jA_x \mid G_x = A_iA_j) \end{bmatrix} = \begin{bmatrix} p_i^2 \\ p_j^2 \\ 2p_ip_j \\ 2p_ip_x \\ 2p_jp_x \end{bmatrix} + \begin{bmatrix} -p_i^2 & \frac{1}{2}p_i - p_i^2 \\ -p_j^2 & \frac{1}{2}p_j - p_j^2 \\ 1-2p_ip_j & \frac{1}{2}p_i + \frac{1}{2}p_j - 2p_ip_j \\ -2p_ip_x & \frac{1}{2}p_x - 2p_ip_x \\ -2p_jp_x & \frac{1}{2}p_x - 2p_jp_x \end{bmatrix} \begin{bmatrix} \Delta \\ \phi \end{bmatrix}. \tag{3b}$$

Here p_i and p_j denote the allele frequencies of A_i and A_j, respectively, and p_x is the sum of the frequencies of visible alleles that do not appear in the reference genotype. Expression (3a) or (3b) is simply written as the matrix equation

$$\mathbf{P} = \mathbf{E} + \mathbf{M}\boldsymbol{\Delta}, \tag{4}$$

which is the general form of all moment estimators that we subsequently describe, where the elements of $\mathbf{E}$ are the probabilities that certain genotype patterns are observed conditional on zero relatedness, the elements of the sum of $\mathbf{E}$ plus the first column (or second column) of $\mathbf{M}$ are the probabilities that there are two pairs of IBD alleles (or one pair of IBD alleles) between x and y, and $\boldsymbol{\Delta}$ is the column vector consisting of the higher-order coefficients Δ and ϕ. The estimate $\hat{\boldsymbol{\Delta}}_{xy}$ can be solved from the formula

$$\hat{\boldsymbol{\Delta}} = (\mathbf{M}^T\mathbf{V}^{-1}\mathbf{M})^{-1}\mathbf{M}^T\mathbf{V}^{-1}(\hat{\mathbf{P}} - \mathbf{E}), \tag{5}$$

where $\hat{\mathbf{P}}$ is the observed value of $\mathbf{P}$, and $\mathbf{V}$ is the variance–covariance matrix of $\mathbf{P}$, assuming x and y are nonrelatives. For the case of homozygotes, because $\mathbf{M}$ is a square matrix with order two, it can be seen from Equation 4 or Equation 5 that $\hat{\boldsymbol{\Delta}} = \mathbf{M}^{-1}(\hat{\mathbf{P}} - \mathbf{E})$. Moreover, under the assumption that x and y are nonrelatives, $\mathbf{V}$ can be used to find the weighted least-squares solution of Equation 4.

The elements of $\mathbf{V}$ are

$$V_{ii} = E_i - E_i^2 \quad \text{and} \quad V_{ij} = -E_iE_j \quad (i \neq j), \tag{6}$$

where E_i is the ith element of $\mathbf{E}$, and $i, j = 1, 2$ for the case of homozygotes, as well as $i, j = 1, 2, 3, 4, 5$ for the case of heterozygotes. Then, using Equation 5, we can calculate the analytical solution $\hat{\boldsymbol{\Delta}}_{xy}$. On the other hand, $\hat{r}_{xy}$ can be solved from Equation 1.

Lynch and Ritland (1999) introduced the Kronecker operator δ_{ab}. This is defined as $\delta_{ab} = 1$ if A_a and A_b are identical-by-state; otherwise $\delta_{ab} = 0$. Using the Kronecker operator, the expressions of $\hat{r}_{xy}$ and $\hat{\Delta}_{xy}$ in homozygous and heterozygous reference genotypes can be unified to write as

$$\hat{r}_{xy} = \frac{p_a(\delta_{bc} + \delta_{bd}) + p_b(\delta_{ac} + \delta_{ad}) - 4p_ap_b}{(1 + \delta_{ab})(p_a + p_b) - 4p_ap_b},$$

$$\hat{\Delta}_{xy} = \frac{2p_ap_b - p_a(\delta_{bc} + \delta_{bd}) - p_b(\delta_{ac} + \delta_{ad}) + \delta_{ac}\delta_{bd} + \delta_{ad}\delta_{bc}}{(1 + \delta_{ab})(1 - p_a - p_b) + 2p_ap_b},$$

where p_a and p_b denote the frequencies of the two alleles in the reference genotype. The locus-specific weights w_r and w_Δ are defined by the inverses of the variances of $\hat{r}_{xy}$ and $\hat{\Delta}_{xy}$, respectively. However, the two variances ($\hat{r}_{xy}$ and $\hat{\Delta}_{xy}$) are functions with r and Δ as the independent variables, in which r and Δ are the quantities we are attempting to estimate. Due to a lack of *a priori* information, Lynch and Ritland (1999) assumed that the two individuals are nonrelatives and then used the variances of $\hat{r}_{xy}$ and $\hat{\Delta}_{xy}$ to weight the locus. The weights w_r and w_Δ of each locus are given by

$$w_r = \frac{1}{\mathrm{Var}(\hat{r}_{xy})} = \frac{(1 + \delta_{ab})(p_a + p_b) - 4p_ap_b}{2p_ap_b},$$

$$w_\Delta = \frac{1}{\mathrm{Var}(\hat{\Delta}_{xy})} = \frac{(1 + \delta_{ab})(1 - p_a - p_b) + 2p_ap_b}{2p_ap_b}.$$

The weighted averages of $\hat{r}_{xy}$ and $\hat{\Delta}_{xy}$ (denoted $\bar{\hat{r}}_{xy}$ and $\overline{\hat{\Delta}}_{xy}$) for all loci used in the estimation can be calculated by averaging the estimates of all loci by using the inverses of the two variances (w_r and w_Δ). Because there is no particular reason to choose the individual x or y as a reference, $\bar{\hat{r}}_{yx}$ and $\overline{\hat{\Delta}}_{yx}$ can also be calculated by using another individual as a reference. The final estimates $\hat{r}$ and $\hat{\Delta}$ are the arithmetic means of both parameters:

$$\hat{r} = \frac{\left(\bar{\hat{r}}_{xy} + \bar{\hat{r}}_{yx}\right)}{2},$$

$$\hat{\Delta} = \frac{\left(\overline{\hat{\Delta}}_{xy} + \overline{\hat{\Delta}}_{yx}\right)}{2}.$$

When a null allele (*i.e.*, A_y) is present, the heterozygote A_iA_y will be mistakenly identified as $A_i'A_i'$ (where $\cdot'$ represents the corresponding quantity of $\cdot$ with the presence of null alleles), whereas the genotype A_yA_y will remain undetected and regarded as a negative result (no bands detected from electrophoresis). Negative results may also be obtained due to other reasons (*e.g.*, operational errors), with null alleles being indistinguishable from homozygotes. Therefore the unobservable genotype should be discarded from the probability of the observed genotype pattern. These facts can be summarized by the potential genotype patterns of

$$\Pr\left(G_y'|G_x', G_y \neq A_yA_y\right) = \sum_{i,j} \Pr(G_x = g_{xi}|G_x')\Pr(G_y = g_{yj}|G_x = g_{xi}, G_y \neq A_yA_y),$$

where g_{xi} and g_{yj} are the ith and jth possible genotypes of x and y, respectively; $\Pr(G_x = g_{xi}|G_x')$ is the probability that g_{xi} is the true reference genotype on condition that G_x' is observed; and

$$\Pr(G_y = g_{yj}|G_x = g_{xi}, G_y \neq A_yA_y)$$

is the probability that G_y is g_{yj} given $G_x = g_{xi}$ and assuming that G_y is observable. For example, if $G_y' = A_i'A_i'$ and $G_x' = A_i'A_i'$, letting ρ be $\Pr(G_y' = A_i'A_i'|G_x' = A_i'A_i', G_y \neq A_yA_y)$, then

$$\begin{aligned}\rho &= \Pr\left(G_x = A_iA_i|G_x' = A_i'A_i'\right)\Pr\left(G_y = A_iA_i \text{ or } A_iA_y|G_x = A_iA_i, G_y \neq A_yA_y\right) + \Pr\left(G_x = A_iA_y|G_x' = A_i'A_i'\right)\Pr\left(G_y = A_iA_i \text{ or } A_iA_y|G_x = A_iA_y, G_y \neq A_yA_y\right)\left(p_i + \tfrac{1}{2}p_y\right)\\ &= \frac{p_i}{p_i + 2p_y} \cdot \frac{\left(p_i^2 + 2p_ip_y\right)(1-\phi-\Delta) + (p_i + p_y)\phi + \Delta}{\left(1-p_y^2\right)(1-\phi-\Delta) + \phi + \Delta}\\ &\quad + \frac{2p_y}{p_i + 2p_y} \cdot \frac{\left(p_i^2 + 2p_ip_y\right)(1-\phi-\Delta) + \left(p_i + \frac{1}{2}p_y\right)\phi + \Delta}{\left(1-p_y^2\right)(1-\phi-\Delta) + \left(1-\frac{1}{2}p_y\right)\phi + \Delta}.\end{aligned}$$

Denote $\mathbf{P}'$ for the corresponding column vector of $\mathbf{P}$ on the left side of expression (3a) or (3b), with the presence of null alleles and G_y being visible [*e.g.*, P_i' for homozygous reference individual is $\Pr(G_y = A_i'A_i'|G_x = A_i'A_i', G_y \neq A_yA_y)$]. It is clear from the previous example that the ith element P_i' of $\mathbf{P}'$ can be written as

$$P_i' = \sum_j \frac{c_{1j}\Delta + c_{2j}\phi + c_{3j}(1-\phi-\Delta)}{d_{1j}\Delta + d_{2j}\phi + d_{3j}(1-\phi-\Delta)}, \tag{7}$$

where c_{kj} and d_{kj} ($k = 1, 2, 3$) are polynomials with p_i and p_y as the variables [such as $c_{11} = p_i$, $c_{12} = p_i^2 + p_ip_y$, and $d_{13} = (p_i + 2p_y)(1-p_y^2)$]. Unfortunately, although $\hat{\boldsymbol{\Delta}}$ can be solved from expression (7) and Equation 5, the estimator becomes more biased because the relationship between $\boldsymbol{\Delta}$ and $\mathbf{P}'$ is not linear. As an alternative, the following approximate form of P_i' for the estimation in expressions (3a) and (3b) can be used:

$$P_i' \approx \sum_j \left(\frac{c_{1j}}{d_{1j}}\Delta + \frac{c_{2j}}{d_{2j}}\phi + \frac{c_{3j}}{d_{3j}}(1-\phi-\Delta)\right). \tag{8}$$

By extracting Δ and φ, expression (8) can be rewritten as

$$P_i' \approx \sum_j \left[\frac{c_{3j}}{d_{3j}} + \left(\frac{c_{1j}}{d_{1j}} - \frac{c_{3j}}{d_{3j}}\right)\Delta + \left(\frac{c_{2j}}{d_{2j}} - \frac{c_{3j}}{d_{3j}}\right)\phi\right]$$

or equivalently,

$$P_i' \approx \sum_j \left(\frac{c_{3j}}{d_{3j}} + \left[\frac{c_{1j}}{d_{1j}} - \frac{c_{3j}}{d_{3j}}, \frac{c_{2j}}{d_{2j}} - \frac{c_{3j}}{d_{3j}}\right]\boldsymbol{\Delta}\right).$$

It is now clear that $\mathbf{P}'$ can be expressed approximatively as $\mathbf{P}' \approx \mathbf{E}^* + \mathbf{M}^*\boldsymbol{\Delta}$, where the matrices $\mathbf{E}^*$ and $\mathbf{M}^*$ are as follows: for the case of homozygotes,

$$\mathbf{E}^* = \begin{bmatrix} \left(p_i^2 + 2p_ip_y\right)/\left(1-p_y^2\right) \\ \left(2p_ip_x\right)/\left(1-p_y^2\right) \end{bmatrix} \text{ and } \mathbf{M}^* = \begin{bmatrix} 1 & \left(p_i \;\; (p_i + p_y)(2-p_y) + 2p_y(p_y + 2p_i)\right)/\left(\left(p_i + 2p_y\right)(2-p_y)\right) \\ 0 & \left(p_ip_x(2-p_y) + 2p_xp_y\right)/\left(\left(p_i + 2p_y\right)(2-p_y)\right) \end{bmatrix} - [\mathbf{E}^*, \mathbf{E}^*]; \tag{9a}$$

for the case of heterozygotes,

$$\mathbf{E}^* = \begin{bmatrix} \left(p_i^2 + 2p_ip_y\right)/\left(1-p_y^2\right) \\ \left(p_j^2 + 2p_jp_y\right)/\left(1-p_y^2\right) \\ \left(2p_ip_j\right)/\left(1-p_y^2\right) \\ \left(2p_ip_x\right)/\left(1-p_y^2\right) \\ \left(2p_jp_x\right)/\left(1-p_y^2\right) \end{bmatrix} \text{ and } \mathbf{M}^* = \begin{bmatrix} 0 & (p_i + p_y)/2 \\ 0 & (p_j + p_y)/2 \\ 1 & (p_i + p_j)/2 \\ 0 & p_x/2 \\ 0 & p_x/2 \end{bmatrix} - [\mathbf{E}^*, \mathbf{E}^*], \tag{9b}$$

in which $[\mathbf{E}^*, \mathbf{E}^*]$ is the partitioned matrix with two columns and every column consisting of $\mathbf{E}^*$ [the derivation of expressions (9a) and (9b) is shown in File S1]. Replacing $\mathbf{E}$ by $\mathbf{E}^*$ and $\mathbf{M}$ by $\mathbf{M}^*$ in Equation 5, the following formula follows,

$$\hat{\boldsymbol{\Delta}} = \left[(\mathbf{M}^*)^T(\mathbf{V}^*)^{-1}\mathbf{M}^*\right]^{-1}(\mathbf{M}^*)^T(\mathbf{V}^*)^{-1}\left(\hat{\mathbf{P}}' - \mathbf{E}^*\right),$$

where $\hat{\mathbf{P}}'$ is the observed value of $\mathbf{P}'$, and the elements in $\mathbf{V}^*$ [refer to system (6) of equalities] are

$$V_{ii}^* = E_i^* - \left(E_i^*\right)^2 \quad \text{and} \quad V_{ij}^* = -E_i^*E_j^* \quad (i \neq j).$$

Applying the above formula, an estimate of $\boldsymbol{\Delta}_{xy}$ can be calculated. Moreover, using Equation 1 and the elements of $\hat{\boldsymbol{\Delta}}_{xy}$, a less biased $\hat{r}_{xy}$ of a single locus can be found. The weights across loci are still the inverses of the variances of $\hat{r}_{xy}$ and $\hat{\Delta}_{xy}$. The symbolic expressions of the variances for $\hat{r}_{xy}$ and $\hat{\Delta}_{xy}$ are complex and are presented in File S1.

Novel Estimator A

This estimator is a modification of Lynch and Ritland's (1999) estimator, but it does not directly use a probability matrix to solve the corresponding system of linear equations. Instead, the method-of-moment approach is used. The definition of the similarity index S is stated in expression (2). The similarity index S as a random variable determines a probability mass function whose values are listed as follows: for the case of homozygotes,

$$\begin{bmatrix} \Pr(S=1) \\ \Pr(S=3/4) \\ \Pr(S=1/2) \end{bmatrix} = \begin{bmatrix} p_i^2 \\ 2p_ip_x \\ 0 \end{bmatrix} + \begin{bmatrix} 1-p_i^2 & p_i-p_i^2 \\ -2p_ip_x & p_x-2p_ip_x \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \Delta \\ \phi \end{bmatrix};$$

for the case of heterozygotes,

$$\begin{bmatrix} \Pr(S=1) \\ \Pr(S=3/4) \\ \Pr(S=1/2) \end{bmatrix} = \begin{bmatrix} 2p_ip_j \\ p_i^2+p_j^2 \\ 2p_x(p_i+p_j) \end{bmatrix} + \begin{bmatrix} 1-2p_ip_j & \frac{1}{2}(p_i+p_j)-2p_ip_j \\ -p_i^2-p_j^2 & \frac{1}{2}(p_i+p_j)-p_i^2-p_j^2 \\ -2p_x(p_i+p_j) & p_x-2p_x(p_i+p_j) \end{bmatrix} \cdot \begin{bmatrix} \Delta \\ \phi \end{bmatrix}.$$

These two systems of functions can be unified as $\mathbf{P} = \mathbf{E} + \mathbf{M}\boldsymbol{\Delta}$. The symbol $\mathbf{S}$ is used to denote the moment vector consisting of the first and second moments of S. Then the relationship between $\mathbf{S}$ and $\boldsymbol{\Delta}$ can be described as

$$\mathbf{S} = \mathbf{CE} + \mathbf{CM}\boldsymbol{\Delta}, \tag{10}$$

where

$$\mathbf{C} = \begin{bmatrix} 1 & 3/4 & 1/2 \\ 1 & 9/16 & 1/4 \end{bmatrix}.$$

Expression (10) is actually a system of linear equations with two unknowns (*i.e.*, Δ and ϕ) and two equations.

With the presence of null alleles, using the same approximation as before, the matrices $\mathbf{E}$ and $\mathbf{M}$ in expression (10) are modified and written as the following matrices $\mathbf{E}^*$ and $\mathbf{M}^*$: for the case of homozygotes,

$$\mathbf{E}^* = \begin{bmatrix} (p_i^2+2p_ip_y)/(1-p_y^2) \\ 2p_ip_x/(1-p_y^2) \\ 0 \end{bmatrix} \quad \text{and}$$

$$\mathbf{M}^* = \begin{bmatrix} 1 & (p_i(p_i+p_y)(2-p_y)+2p_y(p_y+2p_i))/((p_i+2p_y)(2-p_y)) \\ 0 & (p_ip_x(2-p_y)+2p_xp_y)/[(p_i+2p_y)(2-p_y)] \\ 0 & 0 \end{bmatrix} - [\mathbf{E}^*, \mathbf{E}^*];$$

for the case of heterozygotes,

$$\mathbf{E}^* = \begin{bmatrix} 2p_ip_j/(1-p_y^2) \\ (p_j^2+2p_jp_y+p_i^2+2p_ip_y)/(1-p_y^2) \\ 2p_x(p_i+p_j)/(1-p_y^2) \end{bmatrix} \quad \text{and}$$

$$\mathbf{M}^* = \begin{bmatrix} 1 & (p_i+p_j)/2 \\ 0 & (p_i+p_j)/(2+p_y) \\ 0 & p_x \end{bmatrix} - [\mathbf{E}^*, \mathbf{E}^*].$$

Substituting $\mathbf{S}'$ for $\mathbf{S}$, $\mathbf{E}^*$ for $\mathbf{E}$, and $\mathbf{M}^*$ for $\mathbf{M}$ in expression (10), we now derive an approximate expression as

$$\mathbf{S}' \approx \mathbf{CE}^* + \mathbf{CM}^*\boldsymbol{\Delta}, \tag{11}$$

where $\mathbf{S}'$ is the moment vector consisting of the first and second moments of S'. Then, the estimate $\hat{\boldsymbol{\Delta}}_{xy}$ can be calculated from the following formula:

$$\hat{\boldsymbol{\Delta}} = (\mathbf{CM}^*)^{-1}(\mathbf{S}' - \mathbf{CE}^*).$$

On the other hand, $\hat{r}_{xy}$ can be calculated from Equation 1. The locus-specific weights are still the inverses of the variances of $\hat{r}_{xy}$ and $\hat{\Delta}_{xy}$, and the remaining procedure is the same as Lynch and Ritland's (1999) estimator.

Wang's (2002) Estimator

Wang's (2002) estimator also uses the similarity index S as defined in expression (2), but differs from Lynch and Ritland's (1999) estimator. Wang's (2002) estimator does not use reference and proband individuals and is obtained by modeling the probability of each similarity index observed. For example, for the event $S = 1$, the probability $\Pr(S = 1)$ is defined by

$$\begin{aligned} \Pr(S=1) &= \sum_i \Pr(G_x = A_iA_i, G_y = A_iA_i) \\ &\quad + \sum_{i \neq j} \Pr(G_x = A_iA_j, G_y = A_iA_j) \\ &= (2a_2^2 - a_4) + (1 - 2a_2^2 - a_4)\Delta \\ &\quad + (a_2 - 2a_2^2 - a_4)\phi, \end{aligned}$$

where $a_m = \sum_i p_i^m$, $m = 1, 2, 3, 4$ (especially $a_1 = 1$ because $\sum_i p_i = 1$). Generally, the following system of linear equations can be established,

$$\begin{bmatrix} \Pr(S=1) \\ \Pr(S=3/4) \\ \Pr(S=1/2) \end{bmatrix} = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix} + \begin{bmatrix} 1-\lambda_1 & a_2-\lambda_1 \\ -\lambda_2 & 2a_2-2a_3-\lambda_2 \\ -\lambda_3 & 1-3a_2+2a_3-\lambda_3 \end{bmatrix} \times \begin{bmatrix} \Delta \\ \phi \end{bmatrix}, \tag{12}$$

where $\lambda_1 = 2a_2^2 - a_4$, $\lambda_2 = 4a_3 - 4a_4$, and $\lambda_3 = 4a_2 - 4a_2^2 - 8a_3 + 8a_4$ [the derivation of expression (12) is shown in File S1]. Expression (12) can also be expressed as Equation 4, and $\hat{\boldsymbol{\Delta}}$ can be solved from Equation 5. However, the symbolic solutions of $\hat{r}$ and $\hat{\Delta}$ are long (see Wang 2002, for details). The weighting of Wang's (2002) estimator differs from that of most other estimators. The locus-specific weight in Wang's (2002) estimator is defined by the inverse of the expected value of the similarity index and is used to calculate the weighted averages of P_i, a_i, and a_i^2 and to simultaneously solve $\hat{\boldsymbol{\Delta}}$. The expected value $E(S)$ is given by $2a_2 - a_3$ (Li *et al.* 1993; Wang 2002).

If null alleles appear, expression (12) will be modified. For example, for the event $S' = 1$, this is given as

$$
\begin{aligned}
&\Pr(S'=1)[(f_1-f_3)\Delta+(f_2-f_3)\phi+f_3]\\
&=\sum_{i\neq y}\Pr\left(G'_x=A'_iA'_i,G'_y=A'_iA'_i\right)\\
&\quad+\sum_{\substack{i\neq y\\ j\neq y\\ j\neq i}}\Pr\left(G'_x=A'_iA'_j,G'_y=A'_iA'_j\right)\\
&=\sum_{i\neq y}\Pr(G_x=A_iA_i,G_y=A_iA_i)\\
&\quad+\sum_{i\neq y}\Pr(G_x=A_iA_y,G_y=A_iA_y)\\
&\quad+\sum_{i\neq y}\Pr(G_x=A_iA_y,G_y=A_iA_i)\\
&\quad+\sum_{i\neq y}\Pr(G_x=A_iA_i,G_y=A_iA_y)\\
&\quad+\sum_{\substack{i\neq y\\ j\neq y\\ j\neq i}}\Pr(G_x=A_iA_j,G_y=A_iA_j),
\end{aligned}
$$

where $(f_1-f_3)\Delta+(f_2-f_3)\phi+f_3$ is the probability that the genotypes of both individuals are visible, in which $f_1=1-p_y^2$, $f_2=1-p_y^2(2-p_y)$, and $f_3=f_1^2$. From the previous example, the probability for S' being equal to a specific value is the sum of the probabilities of a series of genotype pairs. To denote such a probability by symbols, we let b_m be the sum $\sum_{i\neq y}p_i^m$; then b_m is the counterpart of a_m when null alleles are present. In general, a_m differs slightly from b_m, and their relationship is described as $a_m=b_m+p_y^m$. For instance, let ρ be the sum $\sum_{\substack{i\neq y\\ j\neq y\\ j\neq i}}\Pr(G_x=A_iA_y,G_y=A_iA_j)$. Then

$$
\begin{aligned}
\rho&=(1-\phi-\Delta)\sum_{\substack{i\neq y\\ j\neq y\\ j\neq i}}4p_i^2p_jp_y+\phi\sum_{\substack{i\neq y\\ j\neq y\\ j\neq i}}p_ip_jp_y\\
&=4p_y(1-\phi-\Delta)\sum_{i\neq y}p_i^2\left(\sum_{\substack{j\neq y\\ j\neq i}}p_j\right)+p_y\phi\sum_{i\neq y}p_i\left(\sum_{\substack{j\neq y\\ j\neq i}}p_j\right)\\
&=4p_y(1-\phi-\Delta)\sum_{i\neq y}p_i^2(b_1-p_i)+p_y\phi\sum_{i\neq y}p_i(b_1-p_i)\\
&=4p_y(1-\phi-\Delta)\left(b_1\sum_{i\neq y}p_i^2-\sum_{i\neq y}p_i^3\right)\\
&\quad+p_y\phi\left(b_1\sum_{i\neq y}p_i-\sum_{i\neq y}p_i^2\right)\\
&=4p_y(1-\phi-\Delta)(b_1b_2-b_3)+p_y\phi\left(b_1^2-b_2\right)\\
&=4p_y(b_1b_2-b_3)-4p_y(b_1b_2-b_3)\Delta\\
&\quad+\left[p_y\left(b_1^2-b_2\right)-4p_y(b_1b_2-b_3)\right]\phi.
\end{aligned}
$$

The result of this calculation can also be expressed as

$$
\begin{aligned}
\rho&=4p_y(b_1b_2-b_3)\\
&\quad+\left[-4p_y(b_1b_2-b_3),\ p_y\left(b_1^2-b_2\right)-4p_y(b_1b_2-b_3)\right]\boldsymbol{\Delta}.
\end{aligned}
$$

It is now clear that the probabilities of $S'=1,\ 3/4$, and $1/2$ can be unified as the formula

$$
\mathbf{P}'=[(f_1-f_3)\Delta+(f_2-f_3)\phi+f_3]^{-1}\left(\mathbf{E}'+\mathbf{M}'\boldsymbol{\Delta}\right),\tag{13}
$$

where

$$
\mathbf{E}'=\begin{bmatrix}2b_2^2-b_4+4p_y^2b_2+4p_yb_3\\ 4b_1b_3-4b_4+8p_yb_1b_2-8p_yb_3\\ 4b_1^2b_2-8b_1b_3-4b_2^2+8b_4\end{bmatrix},\tag{14a}
$$

$$
\mathbf{M}'=\begin{bmatrix}b_1^2+2p_yb_1 & b_1b_2+p_y^2b_1+3p_yb_2\\ 0 & 2b_1b_2-2b_3+2p_yb_1^2-2p_yb_2\\ 0 & b_1^3-3b_1b_2+2b_3\end{bmatrix}-\left[\mathbf{E}',\mathbf{E}'\right].\tag{14b}
$$

The derivation of expressions (14a) and (14b) is shown in File S1. Similar to expression (7), the estimation $\hat{\boldsymbol{\Delta}}$ solved by expression (13) is heavily biased. As an alternative, the approximate form of Equation 4 is used for estimation,

$$
\mathbf{P}'\approx\mathbf{E}^*+\mathbf{M}^*\boldsymbol{\Delta},\tag{15}
$$

where

$$
\mathbf{E}^*=f_3^{-1}\mathbf{E}'\quad\text{and}
$$

$$
\mathbf{M}^*=\begin{bmatrix}f_1^{-1}\left(b_1^2+2p_yb_1\right) & f_2^{-1}(b_1b_2+p_y^2b_1+3p_yb_2)\\ 0 & f_2^{-1}\left(2b_1b_2-2b_3+2p_yb_1^2-2p_yb_2\right)\\ 0 & f_2^{-1}\left(b_1^3-3b_1b_2+2b_3\right)\end{bmatrix}-[\mathbf{E}^*,\mathbf{E}^*].\tag{16}
$$

Replacing $\mathbf{E}$ by $\mathbf{E}^*$ and $\mathbf{M}$ by $\mathbf{M}^*$ in Equation 5, the following formula is obtained,

$$
\hat{\boldsymbol{\Delta}}=[(\mathbf{M}^*)^T(\mathbf{V}^*)^{-1}\mathbf{M}^*]^{-1}(\mathbf{M}^*)^T(\mathbf{V}*)^{-1}\left(\hat{\mathbf{P}}'-\mathbf{E}^*\right),
$$

where $\hat{\mathbf{P}}'$ is the observed value of $\mathbf{P}'$ and the elements in $\mathbf{V}^*$ are

$$V_{ii}^* = E_i^* - \left(E_i^*\right)^2 \quad \text{and} \quad V_{ij}^* = -E_i^* E_j^* \quad (i \neq j).$$

Now, the estimate $\hat{\boldsymbol{\Delta}}$ can be calculated from the above formula, which has a smaller bias. There are more parameters that should be weighted, including $\hat{b}_i$ $(i = 1, 2, 3, 4)$, $\hat{b}_1^2$, $\hat{b}_1^3$, $\hat{b}_2^2$, $\hat{p}_y$, $\hat{p}_y^2$, and $\hat{\mathbf{P}}'$. The weight of each locus is the inverse of the expected value of S for nonrelatives, which is given by

$$\begin{aligned} E(S') &= \left[1,\ \frac{3}{4},\ \frac{1}{2}\right]\mathbf{E}' \\ &= \frac{p_y(6b_1b_2 + 4b_2 - 2b_3) + 2b_1^2b_2 - b_1b_3}{\left(1-p_y^2\right)^2}. \end{aligned} \tag{17}$$

Thomas's (2010) Estimator

Thomas's (2010) estimator is a modification of Wang's (2002) estimator. In Thomas's (2010) estimator, the two events $S = 3/4$ and $S = 1/2$ are combined into a single event (*i.e.*, the event $S = 3/4$ or $1/2$). Then, the second and third rows of $\mathbf{P}$, $\mathbf{E}$, and $\mathbf{M}$ in Wang's (2002) estimator are combined as the second row of $\mathbf{P}$, $\mathbf{E}$, and $\mathbf{M}$ in Thomas's (2010) estimator, respectively. This results in

$$\mathbf{P} = \begin{bmatrix} \Pr(S = 1) \\ \Pr(S = 3/4 \text{ or } 1/2) \end{bmatrix}, \quad \mathbf{E} = \begin{bmatrix} 2a_2^2 - a_4 \\ 4a_2 - 4a_2^2 - 4a_3 + 4a_4 \end{bmatrix},$$
$$\text{and} \quad \mathbf{M} = \begin{bmatrix} 1 & a_2 \\ 0 & 1 - a_2 \end{bmatrix} - [\mathbf{E}, \mathbf{E}].$$

Now, the estimate $\hat{\boldsymbol{\Delta}}$ can be solved from Equation 5. The weighting scheme differs from that of Wang's (2002) estimator. In Thomas's (2010) estimator, for each locus, $\hat{r}$ and $\hat{\Delta}$ are obtained by Equations 1 and 5, respectively. The averages of $\hat{r}$ and $\hat{\Delta}$ for all loci are obtained from the following two approaches: (i) the inverse of the expected value of the similarity index is identical to that of both Wang (2002) and Li *et al.* (1993) and (ii) the inverses of the sampling variances of $\hat{r}$ and $\hat{\Delta}$ are equal to those of Lynch and Ritland's (1999) estimator.

With the presence of null alleles, the probabilities of $S' = 1,\ 3/4$, and $1/2$ can also be expressed as expression (13); namely $\mathbf{P}' = [(f_1 - f_3)\Delta + (f_2 - f_3)\phi + f_3]^{-1}(\mathbf{E}' + \mathbf{M}'\boldsymbol{\Delta})$, where the matrices $\mathbf{E}'$ and $\mathbf{M}'$ are

$$\mathbf{E}' = \begin{bmatrix} 2b_2^2 - b_4 + 4p_y^2b_2 + 4p_yb_3 \\ 8p_yb_1b_2 - 8p_yb_3 + 4b_1^2b_2 - 4b_1b_3 - 4b_2^2 + 4b_4 \end{bmatrix},$$

$$\mathbf{M}' = \begin{bmatrix} b_1^2 + 2p_yb_1 & b_1b_2 + p_y^2b_1 + 3p_yb_2 \\ 0 & 2p_yb_1^2 - 2p_yb_2 - b_1b_2 + b_1^3 \end{bmatrix} - \left[\mathbf{E}', \mathbf{E}'\right].$$

This result is similar to Wang's (2002) estimator, the solution being overly biased if $\hat{\boldsymbol{\Delta}}$ is solved directly. The approximation method for the correction of Wang's (2002) estimator can also be applied to regulate this estimator [refer to expression (15)]. The weighting method is the same as that of Wang; the inverse of the expected value of the similarity index is calculated by Equation 17, while the inverse of the variance is obtained from the formula $\mathrm{Var}(X) = \mathrm{E}(X^2) - \mathrm{E}^2(X)$.

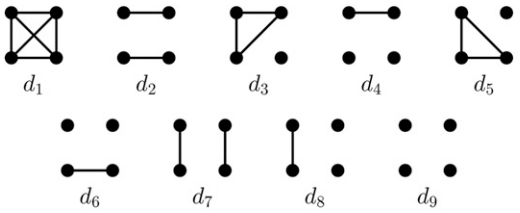


Figure 1 Modes of identity-by-descent between two diploids. In each plot, the two top circles represent the two alleles in one individual, whereas the bottom circles represent the alleles in the second individual. The lines indicate alleles that are identical-by-descent.

Novel Estimator B

This estimator is also based on Wang's (2002) method and uses the moment vector consisting of the first and second moments of the similarity index S to solve $\hat{\boldsymbol{\Delta}}$. The probability mass function of S is stated in expression (12), which can also be denoted as Equation 4. Hence, the relationship between the moment vector $\mathbf{S}$ and $\boldsymbol{\Delta}$ can be established [which is identical to expression (10)], and $\hat{\boldsymbol{\Delta}}$ can be solved from Equation 5. The weight of each locus is defined by the inverse of the expected value of the similarity index, and the weighting scheme follows that of Wang (2002).

With the presence of null alleles, the probability mass function of S' is identical to expression (13), where $\mathbf{E}'$ and $\mathbf{M}'$ are listed in expressions (14a) and (14b). With the same approximation as expression (15), expression (11) is established, where $\mathbf{E}^*$ and $\mathbf{M}^*$ are given in expression (16). Therefore, $\hat{\boldsymbol{\Delta}}$ can be solved from Equation 5, and the weight across loci is defined by expression (2).

Maximum-Likelihood Estimators

Jacquard (1972) described nine IBD modes, denoted by $D_1, D_2, \ldots, D_9$ (see Figure 1), that summarize the possible IBD relationships among the set of four alleles possessed by two diploids. The probability that a pair of individuals is in IBD mode D_i is denoted by Δ_i, so $\sum_{i=1}^{9}\Delta_i = 1$, and Δ_7 and Δ_8 are equivalent to Δ and ϕ, respectively. Because IBD alleles appear in the same individual, $D_1, D_2, \ldots, D_6$ are IBD modes of inbreeding. Therefore, the relatedness coefficient is given by

$$r = 2\Delta_1 + \Delta_3 + \Delta_5 + \Delta_7 + \frac{1}{2}\Delta_8. \tag{18}$$

In an outbred population, Equation 18 is reduced to Equation 1.

There are also nine IBS modes, denoted by $S_1, S_2, \ldots, S_9$, which are interpreted by using different definitions of identical-by-state (identified by separated lines in Figure 1).

Table 1 Probability of patterns of identity-in-state given modes of identity-by-descent in the presence of null alleles

		IBD mode								
IBS mode	Allelic state	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9
S_1	A_iA_i, A_iA_i	y_1p_i	$y_1p_i^2$	$y_1y_2p_i$	$y_1y_3p_i^2$	$y_1y_2p_i$	$y_1y_3p_i^2$	$y_1y_3p_i$	$y_1y_5p_i$	$y_1y_3^2p_i^2$
S_2	A_iA_i, A_jA_j	0	$y_1p_ip_j$	0	$y_1y_4p_ip_j$	0	$y_1y_3p_ip_j$	0	$y_1p_ip_jp_y$	$y_1y_3y_4p_ip_j$
S_3	A_iA_i, A_iA_j	0	0	$y_1p_ip_j$	$2y_1p_i^2p_j$	0	0	0	$y_1y_2p_ip_j$	$2y_1y_3p_i^2p_j$
S_4	A_iA_i, A_jA_k	0	0	0	$2y_1p_ip_jp_k$	0	0	0	0	$2y_1y_3p_ip_jp_k$
S_5	A_iA_j, A_iA_i	0	0	0	0	$y_1p_ip_j$	$2y_1p_i^2p_j$	0	$y_1y_2p_ip_j$	$2y_1y_3p_i^2p_j$
S_6	A_jA_k, A_iA_i	0	0	0	0	0	$2y_1p_ip_jp_k$	0	0	$2y_1y_3p_ip_jp_k$
S_7	A_iA_j, A_iA_j	0	0	0	0	0	0	$2y_1p_ip_j$	$y_1p_ip_j(p_i+p_j)$	$4y_1p_i^2p_j^2$
S_8	A_iA_j, A_iA_k	0	0	0	0	0	0	0	$y_1p_ip_jp_k$	$4y_1p_i^2p_jp_k$
S_9	A_iA_j, A_kA_l	0	0	0	0	0	0	0	0	$4y_1p_ip_jp_kp_l$

Unlike method-of-moment estimators, the maximum-likelihood estimator models the probability of an observed IBS mode conditioned on each IBD mode (for details see table 1 in Milligan 2003).

The single-locus likelihood L is the probability that the genotype pair has been observed given a value of $\boldsymbol{\Delta}$, and L can be described by the formula

$$L = \sum_{i=1}^{9} \Pr(S, D_i)\Delta_i.$$

For multilocus estimation, the global likelihood is the product of the whole single-locus likelihoods. Maximum-likelihood estimators assign implicitly the weights for each locus. The parameter space is $\boldsymbol{\Delta}$ in which $\sum_{i=1}^{9}\Delta_i = 1, 0 \le \Delta_i \le 1$, while in outbred populations, the first six kinds of IBD alleles cannot appear in an individual. Therefore $\sum_{i=1}^{6}\Delta_i = 0$. A constraint, $4\Delta(1-\Delta-\phi) < \phi^2$, in diploids in an outbred population was proposed by Thompson (1976). This constraint is based on the assumption that two individuals have fathers that may be relatives and mothers that may also be relatives, but the parents of each individual are unrelated. This constraint can slightly reduce the bias and was applied by Anderson and Weir (2007), but not in Milligan's (2003) estimator. In addition, Anderson and Weir (2007) model the probability of patterns of identity-in-state given modes of identity-by-descent in structured populations.

With the presence of null alleles, the data in Milligan's (2003) table 1 should be modified as we present here in Table 1. This shows that $y_2 = p_i + p_y$, $y_3 = p_i + 2p_y$, $y_4 = p_j + 2p_y$, and $y_5 = p_i^2 + 3p_ip_y + p_y^2$, with y_1^{-1} being the probability that a pair of genotypes is observed. Each coefficient in Table 1 thus needs to be multiplied by y_1, where

$$\begin{aligned} y_1^{-1} = {} & (\Delta_1 + \Delta_3 + \Delta_5)(1-p_y) + \Delta_2(1-p_y)^2 \\ & + (\Delta_4 + \Delta_6)(1-p_y)\left(1-p_y^2\right) \\ & + \Delta_7\left(1-p_y^2\right) + \Delta_8\left(1-2p_y^2+p_y^3\right) + \Delta_9\left(1-p_y^2\right)^2. \end{aligned} \tag{19}$$

Wagner *et al.* (2006) also applied similar corrections to the estimator of the maximum-likelihood method, but failed to remove the probabilities for unobservable genotype patterns and to apply Thompson's (1976) constraint.

Simulations and Comparisons

We simulated four applications: (i) the effect of null alleles (how $\hat{r}$ biases the result before adjustment), (ii) the reliability of the estimators after correction (bias of the adjusted $\hat{r}$), (iii) the efficiencies of the corrected estimators, and (iv) the robustness of the corrections (populations under nonideal conditions were simulated, and the statistical behaviors of all estimators before and after correction were compared).

We used Monte Carlo simulations for the first three cases. For each dyad, the alleles of the first individual were randomly associated into genotypes according to the allele frequencies. Subsequently, the other genotype was then obtained by using the condition of the first genotype and their relationships (Δ and ϕ). For each locus, we generated a random number t that was uniformly distributed from 0 to 1: (i) if $0 \le t \le \Delta$, the second genotype at this locus would be equal to the first; (ii) if $\Delta < t \le \Delta + \phi$, one allele was randomly copied from the first genotype, and the other was randomly generated according to the allele frequency; and (iii) if $\Delta + \phi < t \le 1$, both alleles of the second genotype were randomly generated according to the allele frequency.

Before Correction

We used both observed genotype and allele frequency for estimation and simulated four levels of null allele frequency: $p_y = 0.1$, 0.3, 0.5, and 0.6. For each level, the visible allele frequency was drawn from a triangular distribution, in which allele frequency followed $1, 2, \ldots, n$ proportions.

Four relationships were simulated, including parent–offspring, full sibs, half sibs, and nonrelatives, and six estimators were compared, including that of Lynch and Ritland (1999) (L&R), the novel estimator A (NA), Wang (2002) (WA), Thomas (2010) (using the inverse of variance for weighting, TH), the novel estimator B (NB), and the estimator of Anderson and Weir (2007) (A&W). It is noteworthy that, because we were unable to apply both corrections (null alleles and structured populations) simultaneously, the population structure parameter θ for the A&W estimator was set to zero in the simulation (see Anderson and Weir 2007). For moment estimators, bias was obtained for a single locus by 6 million simulations. Because the final estimate is usually a weighted average of

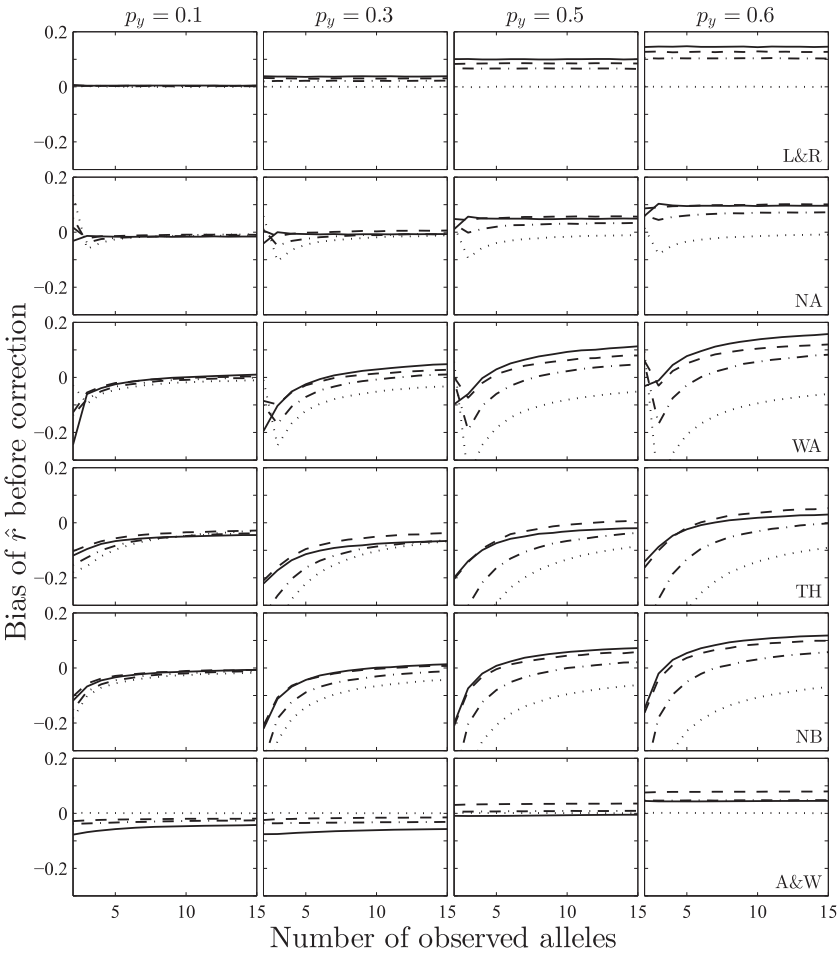


Figure 2 Bias of $\hat{r}$ before correction as a function of the number of visible alleles at loci with triangular distributed allele frequency. The data that have null alleles with a frequency of 0.1, 0.3, 0.5, or 0.6 are shown in the first to fourth columns. The six estimators compared are as follows: Lynch and Ritland (1999) (L&R), novel estimator A (NA), Wang (2002) (WA), Thomas (2010) (TH), novel estimator B (NB), and the Anderson and Weir (2007) estimator (A&W). Results were obtained from 6 million pairs for four relationships by Monte Carlo simulations except the A&W estimator, including parent–offspring ("—"), full sibs ("– –"), half sibs ("–·"), and nonrelatives ("···").

the estimates from all loci (except the WA estimator), increasing the number of loci does not reduce bias. Therefore, a single locus is sufficient to show an effect of null alleles on bias. In contrast, bias was obtained for 60,000 loci from 100 simulations for the A&W estimator. This estimator is asymptotically unbiased. Therefore, there are two sources of bias with the presence of null alleles: (i) the original bias, present even in normal loci, and (ii) the bias brought about by the presence of null alleles. We focused on the second type of bias and used several loci to eliminate the first type of bias. The results shown in Figure 2 are a function of the observed number n' of alleles, with n' being simulated from 2 to 15.

The L&R and NA estimators cope relatively well with null alleles, because their biases are smaller at low frequencies of null alleles (Figure 2). The bias of the L&R estimator remains relatively unchanged as n' increases, as does that of the A&W estimator. Because the bias of each of these two estimators is small, each can be used directly without adjustment at low null allele frequencies. However, each begins to show a larger bias as p_y increases and reaches $\sim$0.15 at $p_y = 0.6$. The NA estimator performs worse than L&R at low levels of null alleles, but improves at high levels of null alleles when $n' > 5$. The other three method-of-moment estimators (WA, TH, and NB) have similar bias curves. The bias of $\hat{r}$ varies as n' increases, is negative at lower n', and begins to increase as n' increases. For a highly polymorphic locus at low null allele frequencies, the biases of the WA and NB estimators are also small.

The bias curves in Figure 2 can be divided into three categories: (i) L&R and NA; (ii) WA, TH, and NB; and (iii) A&W. The biases of L&R and NA estimators are highly independent of the number of observed alleles because the bias of L&R is not a function of n', while those of WA, TH, and NB estimators are dependent on n'. Because the NA estimator is a modification of the L&R estimator, and because the TH and NB estimators are both modifications of the WA estimator, the resulting curves are similar within the same category. We present the bias curves of method-of-moment estimators in File S1. For the A&W estimator, $\hat{r}$ is calculated by a numerical algorithm. Therefore an analytical solution cannot be obtained.

After Correction

The configurations for this application are identical to those previously described. Although bias cannot be completely eliminated, it reduces to a negligible level after correction (Figure 3). The biases are $<$0.003, 0.015, 0.02, and 0.04 when the null allele frequency = 0.1, 0.3, 0.5, and 0.6, respectively. At $n' = 2$, both the NA and WA estimators

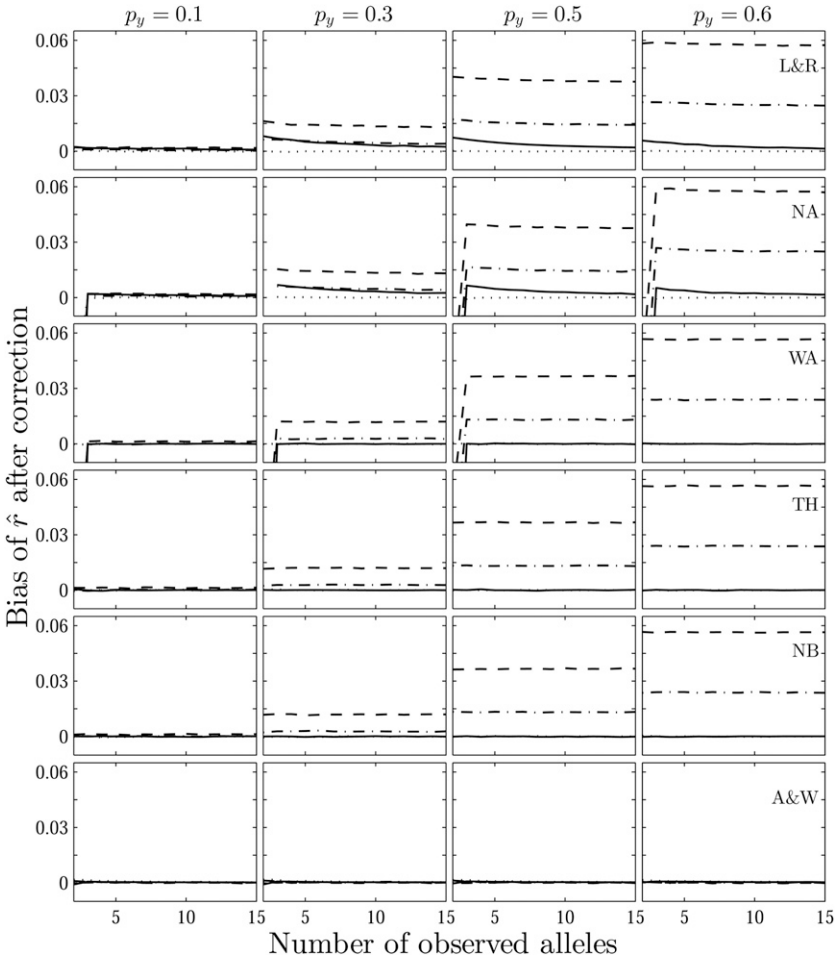


Figure 3 Bias of $\hat{r}$ after correction as a function of the number of alleles at loci with a triangular allele frequency distribution. Four levels of null allele frequency are shown in each column. Six estimators are compared, as shown in Figure 2. Results were obtained from 60 million pairs for four relationships by Monte Carlo simulations except the A&W estimator, including parent-offspring ("—"), full-sibs ("—"), half-sibs ("–·") and nonrelatives ("···").

encounter a singular matrix problem (see Huang *et al.* 2014) and cannot yield a valid estimate, so the corresponding value is missing or highly negative. Full sibs yield the largest bias, while the biases of the other relationships are relatively small. Similarly, the biases of corrected method-of-moment estimators are also presented in File S1, where the biases of the WA, TH, and NB estimators are identical and do not depend on n'.

Efficiency of Estimators

Multiple multiallelic loci were used in all estimations to increase reliability. Less variance suggests a more precise estimation, so variance can be a criterion of efficiency of estimators. However, bias cannot be excluded completely. We thus used the mean square error (MSE) to evaluate the efficiency of estimators, where $\mathrm{MSE}(\hat{r}) = \mathrm{Bias}^2(\hat{r}) + \mathrm{Var}(\hat{r})$. To demonstrate the efficiency of corrected estimators in multilocus estimations, we performed two simulations: (i) the minimum number of loci needed to reach a mean square error <0.01 (the inverse of the resulting value was defined as the potency of a locus and is shown to be a function of n') (Figure 4) and (ii) the bias and MSE of these estimators were compared when estimating relatedness with 20 microsatellites. In natural populations, nonrelative is the most likely relationship between two randomly selected individuals. Therefore, we simulated 10 million dyads, with nonrelatives, half sibs, full sibs, and parent–offspring contributing to 70%, 10%, 10%, and 10% of these dyads, respectively. The results as a function of n' are shown in Figure 5.

The potency of a locus is an inverse function of p_y (Figure 4) but increases with n', reaching an asymptote around $n' = 13$. Locus power is low for high frequencies of null alleles. For example, at least 200 loci should be used when $p_y = 0.7$. The curve for $p_y = 0$ is close to that of $p_y = 0.1$ and is thus omitted from Figure 4.

When estimating relatedness with 20 microsatellites, the biases of the A&W estimator are the highest, while the MSEs of the A&W estimator are the lowest (Figure 5). For method-of-moment estimators, the WA, TH, and NB estimators are least biased, with the MSE curves of these estimators being similar.

Finite Populations

Although all of these estimators performed reasonably well under their given assumptions, real situations often diverge from ideal conditions. For example, allele frequencies are obtained from relatively few individuals, and other population effects such as self-fertilization, inbreeding, and genetic drift may occur.

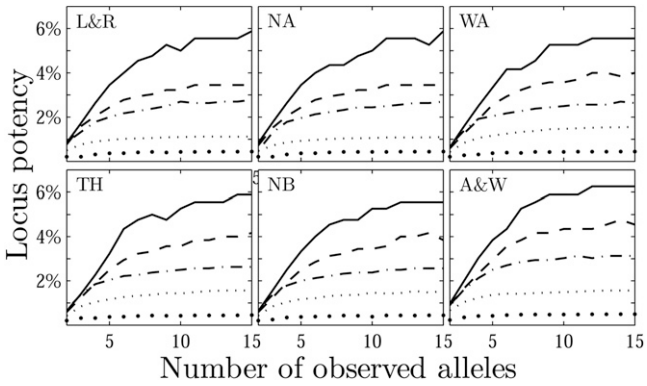


Figure 4 The minimum number of loci by using multiple loci to reach $\mathrm{MSE}(\hat{r}) < 0.01$ for the four relationships shown in Figure 2. The six estimators are the same as in Figure 1. For each estimator, four frequencies of null alleles were simulated: 0.1 (—), 0.3 (– –), 0.5 (–·–), 0.6 (···), and 0.7 (•••), and the observed allele frequency follows the triangular distribution. Results were obtained from 1 million Monte Carlo simulations.

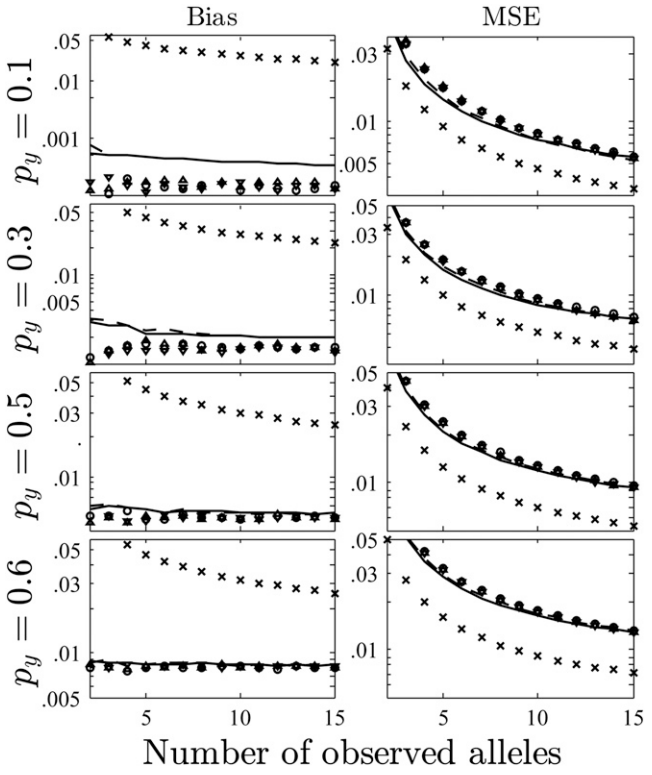


Figure 5 The bias and MSE of $\hat{r}$ estimated from 20 loci with a triangular allele frequency distribution. Ten million dyads were simulated, with nonrelatives, half sibs, full sibs, and parent–offspring contributing to 70%, 10%, 10%, and 10% of dyads, respectively. The frequency of null alleles was simulated at four levels ($p_y = 0.1$, 0.3, 0.5, and 0.6), with each row showing a single level. The results of six estimators are compared: L&R (—), NA (– –), WA (∇), TH (Δ), NA (∘), and A&W (×).

We followed Toro *et al.* (2011) by simulating a small population founded from 20 unrelated and outbred individuals, whose genotypes were randomly generated according to the genotypic frequencies under the Hardy–Weinberg equilibrium. Twenty loci with four visible alleles were simulated. In the founder generation, two levels of the expected frequencies of null alleles were simulated ($p_y = 0.1$ or 0.3), the expected frequencies of these visible alleles being drawn from triangular distributions. Ten discrete generations of 20 individuals were produced, with the sex ratio at each generation fixed at 1 : 1. The parents of each individual were randomly selected from the previous generation (some individuals may not have reproduced), resulting in a data set of 200 individuals and 20,100 dyads. The true kinship coefficient was calculated from the pedigree by a recursive algorithm (Karigl 1981), and the true Wright's relatedness was obtained using the formula $r = \theta_{xy}/\sqrt{\theta_{xx}\theta_{yy}}$ (equation 12 in Huang *et al.* 2015a), where θ_{xy} is the kinship coefficient between x and y, and θ_{xx} is the kinship coefficient between x and itself. The frequencies of null and visible alleles were estimated by a maximum-likelihood estimator, using an EM algorithm (Kalinowski *et al.* 2006). The results of the statistical analysis from 100 replications including bias, root MSE (RMSE), coefficient of correlation (R), and the slope (β_1) and the intercept (β_0) of regression of the estimated relatedness on the true relatedness are shown in Table 2.

Most of the statistics improved after correction, with the corrected estimator becoming more stable as p_y increases. There were some exceptions, for instance the L&R estimator became more biased, with the bias of the corrected estimator being negative and close to $-\bar{r}$.

Discussion

We examined the performance of six estimators in coping with null alleles under various types of pairwise relationship. First, estimator bias before and after correction for null alleles was compared. Although some bias remained, correction reduced this bias to a low level. Second, a comparison of the number of loci that is needed to achieve the same accuracy was performed to evaluate the impact of null alleles on estimator efficiency. This comparison also demonstrated the potency of a different locus and will help researchers to choose an optimal estimator to address their specific research questions. Finally, a small, inbred population prone to genetic drift was simulated to mimic real populations.

Effect of null alleles

Null alleles can affect the estimates of relatedness in two ways. First, they can result in mistyping of true genotypes and change the similarity among genotypes. Given a pair of genotypes, the observed similarity index is under- or overestimated for genotypes such as A_iA_y, A_jA_y and A_iA_y, A_iA_i, because these genotypes are instead observed to be $A_i'A_i', A_j'A_j'$ and $A_i'A_i', A_i'A_i'$. These can reduce the accuracy of estimations.

Second, the overestimation of observed allele frequencies can also affect the overall estimation. Nonrelatives share IBS alleles only by chance. However, the inflation of observed allele frequency can cause overestimation of the probability that two nonrelatives share IBS alleles and may give an inaccurate, reduced estimate. Figure 2 shows the null allele bias for each estimator. The estimators of Wang (2002) and

Table 2 Statistics of $\hat{r}$ in a finite population

		Before correction					After correction				
p_y	Model	Bias	RMSE	R	β_1	β_0	Bias	RMSE	R	β_1	β_0
0.1	L&R	−0.106	0.231	0.362	0.979	–0.104	–0.108	0.235	0.360	0.992	–0.107
	NA	–0.118	0.248	0.390	1.134	–0.133	–0.108	0.247	0.392	1.161	–0.126
	WA	–0.209	0.349	0.436	1.639	–0.280	–0.169	0.305	0.433	1.484	–0.223
	TH	–0.242	0.380	0.432	1.695	–0.319	–0.116	0.292	0.441	1.594	–0.182
	NB	–0.206	0.347	0.437	1.639	–0.277	–0.113	0.280	0.436	1.504	–0.169
	ML	–0.023	0.145	0.415	0.805	–0.001	–0.013	0.153	0.411	0.849	0.004
0.3	L&R	–0.106	0.247	0.355	1.039	–0.111	–0.108	0.247	0.361	1.055	–0.114
	NA	–0.124	0.264	0.372	1.144	–0.140	–0.107	0.259	0.377	1.178	–0.127
	WA	–0.378	0.518	0.435	2.040	–0.493	–0.210	0.334	0.433	1.505	–0.266
	TH	–0.449	0.586	0.444	2.210	–0.583	–0.118	0.306	0.451	1.716	–0.197
	NB	−0.371	0.511	0.436	2.032	–0.485	–0.115	0.291	0.431	1.544	–0.175
	ML	–0.022	0.146	0.407	0.804	–0.001	–0.008	0.157	0.402	0.864	0.006

Thomas (2010) have reduced performance in the presence of null alleles, as does the novel estimator B. For null allele frequencies ∼0.5, all estimators have a large bias.

Why are these estimators still biased?

These estimators were corrected after we modeled the probability of genetic relationships in the presence of null alleles. However, bias still existed, although it was not high. There are two sources of bias in moment estimators.

The first source is due to the approximation [(e.g. Expressions (8), (11) and (15)] described in the *Theory and Modeling* section and is present in all moment estimators. When ϕ or Δ is not equal to zero or one, bias arises because the approximation is not equivalent to the original form (*e.g.*, the bias for full sibs was the largest in Figure 3, while the biases for both parent–offspring and nonrelatives were nearly zero). However, the unbiased estimator for expression (7) or (13) does not exist, because there is an inverse of P_i in the $\hat{r}$, and P_i is a binomial distribution if only one locus is used for estimation. Consider a simpler form as follows (by using only one parameter x):

$$\Pr(X=1)=\frac{a_1x+b_1}{c_1x+d_1},$$
$$\Pr(X=2)=\frac{a_2x+b_2}{c_2x+d_2}.$$

Assume that there is an unbiased estimator $f(X)$; then $E[f(X)]=x$, so

$$f(1)\Pr(X=1)+f(2)\Pr(X=2)=x.$$

Multiplying both sides of the last equality by $(c_2x+d_2)(c_1x+d_1)$, and acknowledging the representations of $\Pr(X=1)$ and $\Pr(X=2)$ above, it follows that

$$f(1)(a_1x+b_1)(c_2x+d_2)+f(2)(a_2x+b_2)(c_1x+d_1)=x(c_2x+d_2)(c_1x+d_1).$$

The index of x is not equal in the two sides of the above equation, but x is a consistent parameter, and the equation cannot hold under any conditions. Therefore, an unbiased estimator for expression (7) or (13) does not exist. For $p_y<0.3$, the bias is manageable.

The second source of bias is that the genotypic frequencies deviate from expected values. This problem is inherent to the L&R and NA estimators (the bias increases as n' decreases). These estimators model the probability of an observed genotype occurrence relative to a reference genotype. With null alleles, the observed genotypic frequencies of two related individuals are drawn from different distributions. As an example, Table 3 lists the distribution of parent–offspring relatedness, with the second column and bottom row showing the distributions of the reference and proband genotypes, respectively. However, without *a priori* information, the estimator cannot determine which is the reference genotype, and bias cannot thus be avoided.

For the maximum-likelihood estimator, there are no such problems. However, the maximum-likelihood estimator is still biased because the estimate is limited to the parameter space, with negative relatedness being unobtainable. Therefore, a bias for relationships with $\mathbf{\Delta}$ lying at the edge of the parameter space is always present.

The potency of loci with null alleles

When the frequencies of null alleles are high enough, loci with null alleles can be identified. Although bias can be reduced, the loss of efficiency for loci with null alleles cannot be recovered. For example, for $p_y=0.5$ and 0.7, only 56% and 26% of loci, respectively, can give a valid estimate for nonrelatives.

In simulations, the MSE is used to evaluate the efficiency of loci. The inverse of the minimum number of loci that enables $\mathrm{MSE}(\hat{r})<0.01$ is shown in Figure 4, which describes the potency of a single locus with null alleles. Approximately 19, 25, 34, 100, and 200 polymorphic loci should be used to fulfill this requirement for null allele frequencies being 0.1, 0.3, 0.5, 0.6, and 0.7, so each locus contributes ∼5%, 4%, 3%, 1%, and 0.5% to reliability (Figure 4). In addition, the potency of each normal locus is approximately that of $p_y=0.1$. Loci showing higher potency (*e.g.*, >3%) can be used for

Table 3 The distribution of genotypes between parents (G_x') and offspring (G_y') with the presence of null alleles, assuming there are three uniformly distributed alleles and one of them is a null allele

		$\Pr(G_y'\|G_x')$		
G_x'	$\Pr(G_x')$	$A_1'A_1'$	$A_1'A_2'$	$A_2'A_2'$
$A_1'A_1'$	3/8	5/8	1/4	1/8
$A_1'A_2'$	1/4	1/4	1/4	1/4
$A_2'A_2'2$	3/8	1/8	1/4	5/8
$\Pr(G_y')$		4/11	3/11	4/11

estimation after correction. For loci displaying low potency, they should be discarded to search for new polymorphic loci or new primers should be designed. We suggest that loci with null allele frequencies >0.5 should not be typed, even if regulated estimators are used. This is because the information provided by these loci is less than half that of normal loci. Even if they have been typed, it does not matter if they are used for estimation because they contribute little to the MSE.

Finite populations

In our fourth application, nonideal conditions were simulated. The corrected estimators had improved statistical performance and were more robust at different levels of allele frequency. The L&R, NA, and A&W estimators performed well for $p_y < 0.3$ (Figure 2), so the statistical performances of the corrected estimators were not improved. For the WA, TH, and NB estimators, there was large negative bias (< -0.1) for smaller values of n' (Figure 2); their corrected estimators were therefore less biased and more accurate (lower RMSE).

There was a negative bias near $-\bar{r}$ for corrected estimators (Table 2). For panmictic populations, the estimator gives an expected estimate of zero. However, we found negative bias in populations that included relatives. Two randomly selected individuals in such a population have an expected relatedness estimate of zero but their true expected relatedness is 0.111. This results in negative bias.

Kinship estimators

We also modeled the kinship estimators of Loiselle *et al.* (1995) and Ritland (1996) for null alleles. We found that these estimators cannot incorporate null alleles because they model the event that one pair of alleles is sampled from two individuals. Because null alleles cannot be observed and sampled, this would instead be a resampling of alleles if null alleles were sampled. However, without Δ and ϕ, the probabilities cannot be modeled through a single kinship coefficient. We performed a simple simulation and found that the expected similarity index and kinship coefficient were different in full sibs and parent–offspring in which null alleles were present, even though the kinship coefficients were identical in these two relationships.

Selection of estimators

Although the A&W estimator performs well in simulations, there are some conditions under which others perform better according to specific metrics. Therefore, there is no single estimator that has superior performance under all conditions and using all metrics. The RMSE depended on the allele frequency, the number of alleles and loci, and the type of relationship between individuals. The simulations and comparisons we present here will probably help researchers choose the most suitable estimator for their particular application. Furthermore, for applications for specific genetic conditions, we show that it is possible to identify a single optimal estimator. To make this easier, we have made available a free software package, "PolyRelatedness," that provides a simulation function to help researchers evaluate the performance of each estimator under their given conditions.

For novice users, we recommend the A&W estimator because of its robustness. However, there are at least two situations when the A&W estimator cannot be used. The first one is dealing with tied values in the data set. Some of the estimates of the A&W estimator lie at the edge of parameter space, which results in many estimates of 0 or 0.5. If the estimated relatedness between two kinds of dyads is compared with a rank-sum test, it is difficult to determine the ranks of these ties. The second one is the problem of bias, because the maximum-likelihood estimator cannot be used for applications that required a bias-free analysis. For example, the relationship between bee or ant workers within the same colony is either half sibs ($r = 0.25$) or full sibs ($r = 0.75$) if there is only a single queen inside this colony, so the ratio of full sibs can be calculated by $2\bar{r} - 0.5$ if an unbiased estimator is used (Huang *et al.* 2015b). In this case, the L&R estimator can be used as an alternative because the MSEs generated for nonrelatives are lower than other method-of-moment estimators (Figure 5), the most likely relationship between two randomly selected individuals being nonrelatives in medium to large populations.

Acknowledgments

We thank Stephanie T. Chen for language corrections and Ruo-du Wang for helpful comments on the mathematics. We also thank the associate editor, Bret Payseur, and two anonymous reviewers for helpful comments on an earlier version of the manuscript. This study was supported by the National Nature Science Foundation of China (31130061, 31501872, 31270441, and 31470455).

K.H. and B.L. designed the project, K.H. wrote the program and draft, S.G. and X.Q. provided the tools and data, and K.R. and D.W.D. checked the model and edited the manuscript.

Literature Cited

Anderson, A. D., and B. S. Weir, 2007 A maximum-likelihood method for the estimation of pairwise relatedness in structured populations. Genetics 176: 421–440.

Blouin, M. S., 2003 DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. Trends Ecol. Evol. 18: 503–511.

Brookfield, J. F. Y., 1996 A simple new method for estimating null allele frequency from heterozygote deficiency. Mol. Ecol. 5: 453–455.

Dabrowski, M. J., M. Pilot, M. Kruczyk, M. Zmihorski, H. M. Umer *et al.*, 2013 Reliability assessment of null allele detection: inconsistencies between and within different methods. Mol. Ecol. Resour. 14: 361–373.

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

He, G., K. Huang, S. T. Guo, W. H. Ji, X. G. Qi *et al.*, 2011 Evaluating the reliability of microsatellite genotyping from low-quality DNA templates with a polynomial distribution model. Chin. Sci. Bull. 56: 2523–2530.

Huang, K., K. Ritland, S. T. Guo, M. Shattuckn, and B. G. Li, 2014 A pairwise relatedness estimator for polyploids. Mol. Ecol. Resour. 14: 734–744.

Huang, K., S. T. Guo, M. R. Shattuck, S. T. Chen, X. G. Qi *et al.*, 2015a A maximum-likelihood estimation of pairwise relatedness for autopolyploids. Heredity 114: 133–142.

Huang, K., K. Ritland, S. T. Guo, D. W. Dunn, D. Chen *et al.*, 2015b Estimating pairwise relatedness between individuals with different levels of ploidy. Mol. Ecol. Resour. 15: 772–784.

Jacquard, A., 1972 Genetic information given by a relative. Biometrics 28: 1101–1114.

Kalinowski, S. T., A. P. Wagner, and M. L. Taper, 2006 Ml-relate: a computer program for maximum likelihood estimation of relatedness and relationship. Mol. Ecol. Notes 6: 576–579.

Karigl, G., 1981 A recursive algorithm for the calculation of identity coefficients. Ann. Hum. Genet. 45: 299–305.

Kokita, T., S. Takahashi, and H. Kumada, 2013 Molecular signatures of lineage-specific adaptive evolution in a unique sea basin: the example of an anadromous goby *Leucopsarion petersii*. Mol. Ecol. 22: 1341–1355.

Li, C. C., D. E. Weeks, and A. Chakravarti, 1993 Similarity of DNA fingerprints due to chance and relatedness. Hum. Hered. 43: 45–52.

Loiselle, B. A., V. L. Sork, J. Nason, and C. Graham, 1995 Spatial genetic structure of a tropical understory shrub, *Psychotria officinalis* (rubiaceae). Am. J. Bot. 82: 1420–1425.

Lynch, M., and K. Ritland, 1999 Estimation of pairwise relatedness with molecular markers. Genetics 152: 1753–1766.

Mattila, A. L. K., A. Duplouy, M. Kirjokangas, R. Lehtonen, P. Rastas *et al.*, 2012 High genetic load in an old isolated butterfly population. Proc. Natl. Acad. Sci. USA 109: E2496–E2505.

Milligan, B. G., 2003 Maximum-likelihood estimation of relatedness. Genetics 163: 1153–1167.

Queller, D. C., and K. F. Goodnight, 1989 Estimating relatedness using genetic markers. Evolution 43: 258–275.

Ritland, K., 1996 Estimators for pairwise relatedness and individual inbreeding coefficients. Genet. Res. 67: 175–185.

Taberlet, P., S. Griffin, B. Goossens, S. Questiau, V. Manceau *et al.*, 1996 Reliable genotyping of samples with very low DNA quantities using PCR. Nucleic Acids Res. 24: 3189–3194.

Thomas, S. C., 2010 A simplified estimator of two and four gene relationship coefficients. Mol. Ecol. Resour. 10: 986–994.

Thompson, E. A., 1976 A restriction on the space of genetic relationships. Ann. Hum. Genet. 40: 201–204.

Toro, M. Á., L. A. García-Cortés, and A. Legarra, 2011 A note on the rationale for estimating genealogical coancestry from molecular markers. Genet. Sel. Evol. 43: 27.

van Oosterhout, C., W. F. Hutchinson, D. P. Wills, and P. Shipley, 2004 Micro-checker: software for identifying and correcting genotyping errors in microsatellite data. Mol. Ecol. Notes 4: 535–538.

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

Wagner, A. P., S. Creel, and S. T. Kalinowski, 2006 Estimating relatedness and relationships using microsatellite loci with null alleles. Heredity 97: 336–345.

Wang, J. L., 2002 An estimator for pairwise relatedness using molecular markers. Genetics 160: 1203–1215.

Wright, S., 1921 Systems of mating. I. The biometric relations between parent and offspring. Genetics 6: 111.

Communicating editor: B. A. Payseur

GENETICS

Supporting Information
www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.163956/-/DC1

Estimating Relatedness in the Presence of Null Alleles

Kang Huang, Kermit Ritland, Derek W. Dunn, Xiaoguang Qi, Songtao Guo, and Baoguo Li

DOI: 10.1534/genetics.114.163956

File S1. Supplementary materials.

Supplementary materials for 'Estimating relatedness with the presence of null alleles'

The derivation of Expressions (3a) and (3b)

In outbred populations, the probability that the proband genotype G_y is observed is conditional on the reference genotype G_x and can be expressed as

$$\Pr(G_y \mid G_x) = \Pr(G_y \mid G_x, \Delta = \phi = 0)(1 - \Delta - \phi) + \Pr(G_y \mid G_x, \phi = 1)\phi + \Pr(G_y \mid G_x, \Delta = 1)\Delta.$$

This expression is a weighed summation of three probabilities, where $1-\Delta-\phi$, ϕ and Δ are corresponding weights. The first probability $\Pr(G_y \mid G_x, \Delta = \phi = 0)$ assumes that x and y are nonrelatives, and this can be calculated by using the genotypic frequencies under Hardy-Weinberg equilibrium. The second and third probabilities $\Pr(G_y \mid G_x, \phi = 1)$ and $\Pr(G_y \mid G_x, \Delta = 1)$ denote the probabilities of genotype G_y being observed is conditional on genotype G_x, where the former is conditional on the two individuals sharing one identical-by-descent allele (e.g. parent-offspring) and the latter is conditional on two identical-by-descent alleles (e.g. identical-twins). If the reference individual x is a homozygote, then

$$\begin{aligned}\Pr(G_y = A_iA_i \mid G_x = A_iA_i) &= \Pr(G_y = A_iA_i \mid G_x = A_iA_i, \Delta = \phi = 0)(1 - \Delta - \phi) \\ &\quad + \Pr(G_y = A_iA_i \mid G_x = A_iA_i, \phi = 1)\phi \\ &\quad + \Pr(G_y = A_iA_i \mid G_x = A_iA_i, \Delta = 1)\Delta \\ &= p_i^2(1 - \Delta - \phi) + p_i\phi + \Delta,\end{aligned}$$

namely

$$\Pr(G_y = A_iA_i \mid G_x = A_iA_i) = p_i^2 + (1 - p_i^2)\Delta + (p_i - p_i^2)\phi; \tag{i}$$

and

$$\begin{aligned}\Pr(G_y = A_iA_x \mid G_x = A_iA_i) &= \Pr(G_y = A_iA_x \mid G_x = A_iA_i, \Delta = \phi = 0)(1 - \Delta - \phi) \\ &\quad + \Pr(G_y = A_iA_x \mid G_x = A_iA_i, \phi = 1)\phi \\ &\quad + \Pr(G_y = A_iA_x \mid G_x = A_iA_i, \Delta = 1)\Delta \\ &= 2p_ip_x(1 - \Delta - \phi) + p_x\phi + 0\Delta,\end{aligned}$$

namely

$$\Pr(G_y = A_iA_x \mid G_x = A_iA_i) = 2p_ip_x + (-2p_ip_x)\Delta + (p_x - 2p_ip_x)\phi. \tag{ii}$$

Now, combining expressions (i) with (ii), we obtain Expression (3a), that is

$$\begin{bmatrix} \Pr(G_y = A_iA_i \mid G_x = A_iA_i) \\ \Pr(G_y = A_iA_x \mid G_x = A_iA_i) \end{bmatrix} = \begin{bmatrix} p_i^2 \\ 2p_ip_x \end{bmatrix} + \begin{bmatrix} 1 - p_i^2 & p_i - p_i^2 \\ -2p_ip_x & p_x - 2p_ip_x \end{bmatrix} \begin{bmatrix} \Delta \\ \phi \end{bmatrix}.$$

Next, if the reference individual x is a heterozygote, then

$$\begin{aligned}
\Pr(G_y = A_iA_i \mid G_x = A_iA_j) &= \Pr(G_y = A_iA_i \mid G_x = A_iA_j, \Delta = \phi = 0)(1-\Delta-\phi) \\
&\quad + \Pr(G_y = A_iA_i \mid G_x = A_iA_j, \phi = 1)\phi \\
&\quad + \Pr(G_y = A_iA_i \mid G_x = A_iA_j, \Delta = 1)\Delta \\
&= p_i^2(1-\Delta-\phi) + \frac{1}{2}p_i\phi + 0\Delta \\
&= p_i^2 + \big(-p_i^2\big)\Delta + \left(\frac{1}{2}p_i - p_i^2\right)\phi,
\end{aligned}$$

$$\begin{aligned}
\Pr(G_y = A_jA_j \mid G_x = A_iA_j) &= \Pr(G_y = A_jA_j \mid G_x = A_iA_j, \Delta = \phi = 0)(1-\Delta-\phi) \\
&\quad + \Pr(G_y = A_jA_j \mid G_x = A_iA_j, \phi = 1)\phi \\
&\quad + \Pr(G_y = A_jA_j \mid G_x = A_iA_j, \Delta = 1)\Delta \\
&= p_j^2(1-\Delta-\phi) + \frac{1}{2}p_j\phi + 0\Delta \\
&= p_j^2 + \big(-p_j^2\big)\Delta + \left(\frac{1}{2}p_j - p_j^2\right)\phi,
\end{aligned}$$

$$\begin{aligned}
\Pr(G_y = A_iA_j \mid G_x = A_iA_j) &= \Pr(G_y = A_iA_j \mid G_x = A_iA_j, \Delta = \phi = 0)(1-\Delta-\phi) \\
&\quad + \Pr(G_y = A_iA_j \mid G_x = A_iA_j, \phi = 1)\phi \\
&\quad + \Pr(G_y = A_iA_j \mid G_x = A_iA_j, \Delta = 1)\Delta \\
&= 2p_ip_j(1-\Delta-\phi) + \frac{1}{2}(p_i+p_j)\phi + \Delta \\
&= 2p_ip_j + (1-2p_ip_j)\Delta + \left(\frac{1}{2}p_i + \frac{1}{2}p_j - 2p_ip_j\right)\phi,
\end{aligned}$$

$$\begin{aligned}
\Pr(G_y = A_iA_x \mid G_x = A_iA_j) &= \Pr(G_y = A_iA_x \mid G_x = A_iA_j, \Delta = \phi = 0)(1-\Delta-\phi) \\
&\quad + \Pr(G_y = A_iA_x \mid G_x = A_iA_j, \phi = 1)\phi \\
&\quad + \Pr(G_y = A_iA_x \mid G_x = A_iA_j, \Delta = 1)\Delta \\
&= 2p_ip_x(1-\Delta-\phi) + \frac{1}{2}p_x\phi + 0\Delta \\
&= 2p_ip_x + (-2p_ip_x)\Delta + \left(\frac{1}{2}p_x - 2p_ip_x\right)\phi,
\end{aligned}$$

$$\begin{aligned}
\Pr(G_y = A_jA_x \mid G_x = A_iA_j) &= \Pr(G_y = A_jA_x \mid G_x = A_iA_j, \Delta = \phi = 0)(1-\Delta-\phi) \\
&\quad + \Pr(G_y = A_jA_x \mid G_x = A_iA_j, \phi = 1)\phi \\
&\quad + \Pr(G_y = A_jA_x \mid G_x = A_iA_j, \Delta = 1)\Delta \\
&= 2p_jp_x(1-\Delta-\phi) + \frac{1}{2}p_x\phi + 0\Delta \\
&= 2p_jp_x + (-2p_jp_x)\Delta + \left(\frac{1}{2}p_x - 2p_jp_x\right)\phi.
\end{aligned}$$

Now, it is clear that the following expression (i.e. Expression (3b)) holds:

$$\begin{bmatrix} \Pr(G_y = A_iA_i \mid G_x = A_iA_j) \\ \Pr(G_y = A_jA_j \mid G_x = A_iA_j) \\ \Pr(G_y = A_iA_j \mid G_x = A_iA_j) \\ \Pr(G_y = A_iA_x \mid G_x = A_iA_j) \\ \Pr(G_y = A_jA_x \mid G_x = A_iA_j) \end{bmatrix} = \begin{bmatrix} p_i^2 \\ p_j^2 \\ 2p_ip_j \\ 2p_ip_x \\ 2p_jp_x \end{bmatrix} + \begin{bmatrix} -p_i^2 & \frac{1}{2}p_i - p_i^2 \\ -p_j^2 & \frac{1}{2}p_j - p_j^2 \\ 1 - 2p_ip_j & \frac{1}{2}p_i + \frac{1}{2}p_j - 2p_ip_j \\ -2p_ip_x & \frac{1}{2}p_x - 2p_ip_x \\ -2p_jp_x & \frac{1}{2}p_x - 2p_jp_x \end{bmatrix} \begin{bmatrix} \Delta \\ \phi \end{bmatrix}.$$

The variances of $\hat{r}$ and $\hat{\Delta}$ of corrected Lynch and Ritland's (1999) estimator

There are three and six genotype patterns for homozygous and heterozygous reference individuals, respectively. The probability of each genotype pattern been observed on condition that the relationship between proband and reference individuals are nonrelatives are listed in Table 1. The $\hat{r}$ and $\hat{\Delta}$ for each genotypes pattern can also be calculated. Therefore, $\mathrm{Var}(\hat{r})$ and $\mathrm{Var}(\hat{\Delta})$ can be obtained by the variance formula $\mathrm{Var}(X) = \mathrm{E}(X^2) - \mathrm{E}^2(X)$.

Table 1: Genotype patterns and the probabilities being observed in nonrelatives

Genotype pattern	Probability	Expression
$A_i'A_i'$-$A_i'A_i'$	$\Pr(G_y' = A_i'A_i' \mid G_x' = A_i'A_i', \phi = \Delta = 0)$	$\frac{p_i^2 + 2p_ip_y}{1 - p_y^2}$
$A_i'A_x'$-$A_i'A_i'$	$\Pr(G_y' = A_i'A_x' \mid G_x' = A_i'A_i', \phi = \Delta = 0)$	$\frac{2p_ip_x}{1 - p_y^2}$
$A_x'A_x'$-$A_i'A_i'$	$\Pr(G_y' = A_x'A_x' \mid G_x' = A_i'A_i', \phi = \Delta = 0)$	$\frac{p_x^2 + 2p_xp_y}{1 - p_y^2}$
$A_i'A_i'$-$A_i'A_j'$	$\Pr(G_y' = A_i'A_i' \mid G_x' = A_i'A_j', \phi = \Delta = 0)$	$\frac{p_i^2 + 2p_ip_y}{1 - p_y^2}$
$A_j'A_j'$-$A_i'A_j'$	$\Pr(G_y' = A_j'A_j' \mid G_x' = A_i'A_j', \phi = \Delta = 0)$	$\frac{p_j^2 + 2p_jp_y}{1 - p_y^2}$
$A_i'A_j'$-$A_i'A_j'$	$\Pr(G_y' = A_i'A_j' \mid G_x' = A_i'A_j', \phi = \Delta = 0)$	$\frac{2p_ip_j}{1 - p_y^2}$
$A_i'A_x'$-$A_i'A_j'$	$\Pr(G_y' = A_i'A_x' \mid G_x' = A_i'A_j', \phi = \Delta = 0)$	$\frac{2p_ip_x}{1 - p_y^2}$
$A_j'A_x'$-$A_i'A_j'$	$\Pr(G_y' = A_j'A_x' \mid G_x' = A_i'A_j', \phi = \Delta = 0)$	$\frac{2p_jp_x}{1 - p_y^2}$
$A_x'A_x'$-$A_i'A_j'$	$\Pr(G_y' = A_x'A_x' \mid G_x' = A_i'A_j', \phi = \Delta = 0)$	$\frac{p_x^2 + 2p_xp_y}{1 - p_y^2}$

Using a mathematics computational software package (MATHEMATICA V9), the symbolic expressions of $\mathrm{Var}(\hat{r})$ and $\mathrm{Var}(\hat{\Delta})$ can be calculated, and the results are as follows: for the case of homozygous reference individuals,

$$\mathrm{Var}(\hat{r}) = \frac{p_i\lambda_{11} + p_i^3\lambda_{12} + p_i^2\lambda_{13}}{\mu_1} \quad \text{and} \quad \mathrm{Var}(\hat{\Delta}) = \frac{p_i\lambda_{21} + p_i^3\lambda_{22}}{\mu_2},$$

where

$$\lambda_{11} = p_i^4(2-p_y)^2(1-2p_y^2) + 4p_y^2(1+p_y)^2(4-8p_y+5p_y^2),$$
$$\lambda_{12} = 4 - 44p_y + 45p_y^2 + 78p_y^3 - 67p_y^4 - 16p_y^5 + 2p_i(-4+16p_y+3p_y^2-27p_y^3+9p_y^4+p_y^5),$$
$$\lambda_{13} = 4p_y(4-14p_y-7p_y^2+18p_y^3+5p_y^4-2p_y^5),$$
$$\mu_1 = 2(-1+p_i+p_y)(-1+p_y^2)\left[p_i^2(-2+p_y)+2p_y(1+p_y)-p_i(-2+5p_y+p_y^2)\right]^2;$$
$$\lambda_{21} = 8p_y^3(1+p_y)^2 - p_i^4(-2+p_y)^2(1+p_y^2) + 4p_ip_y^2(5+p_y-p_y^2+3p_y^3),$$
$$\lambda_{22} = 2p_y(8-8p_y+15p_y^2-10p_y^3-5p_y^4) + p_i(4-16p_y+21p_y^2-25p_y^3+7p_y^4+p_y^5),$$
$$\mu_2 = (-1+p_i+p_y)(-1+p_y^2)\left[p_i^2(-2+p_y)+2p_y(1+p_y)-p_i(-2+5p_y+p_y^2)\right]^2.$$

For the case of heterozygous reference individuals,

$$\mathrm{Var}(\hat{r}) = \frac{2p_ip_j(\lambda_{11}+p_i^2\lambda_{12}+p_i\lambda_{13})}{(-1+p_y^2)(\mu_{11}+p_i^3\mu_{12}+p_i\mu_{13}+p_i^2\mu_{14})} \quad \text{and} \quad \mathrm{Var}(\hat{\Delta}) = \frac{2p_ip_j(\lambda_{21}+p_i\lambda_{22}+p_ip_j\lambda_{23})}{\mu_{21}+p_i^3\mu_{22}+p_i\mu_{23}+p_i^3\mu_{24}},$$

where

$$\lambda_{11} = 2p_y(1-p_j)(p_j+2p_y)(-1+p_y^2),$$
$$\lambda_{12} = 2p_y(1-p_y^2) + p_j^2(-2+4p_y^2) + p_j(1-3p_y-p_y^2+7p_y^3),$$
$$\lambda_{13} = 2p_y(-1+2p_y+p_y^2-2p_y^3) + p_j^2(1-3p_y-p_y^2+7p_y^3) + p_j(-1+4p_y-3p_y^2-4p_y^3+12p_y^4),$$
$$\mu_{11} = 2p_jp_y(-1+p_j)(p_j+2p_y)(-1+p_y^2),$$
$$\mu_{12} = -8p_j^3 + p_j^2(6-14p_y) + 2p_y(-1+p_y^2) + p_j(-1+12p_y+p_y^2),$$
$$\mu_{13} = 4p_y^2(1-p_y^2) + p_j^3(-1+12p_y+p_y^2) + p_j^2(1-15p_y+23p_y^2-p_y^3) + 4p_jp_y(1-7p_y-p_y^2-p_y^3),$$
$$\mu_{14} = p_j^3(6-14p_y) + p_j^2(-6+24p_y-26p_y^2) + p_j(1-15p_y+23p_y^2-p_y^3) + 2p_y(1-2p_y-p_y^2+2p_y^3);$$
$$\lambda_{21} = 2p_jp_y(p_j+2p_y)(-1+p_y^2) + 2p_i^2p_y(-1+p_y^2) + 4p_i^2p_j^2(1+p_y^2),$$
$$\lambda_{22} = p_ip_j(-1+9p_y+p_y^2+7p_y^3) + 4p_y^2(-1+p_y^2),$$
$$\lambda_{23} = 4p_y(-1+5p_y+p_y^2+3p_y^3) + p_j(-1+9p_y+p_y^2+7p_y^3),$$
$$\mu_{21} = (-1+p_y^2) + 2p_jp_y(-1+p_j)(p_j+2p_y)(-1+p_y^2),$$
$$\mu_{22} = -8p_j^3 + p_j^2(6-14p_y) + 2p_y(-1+p_y^2) + p_j(-1+12p_y+p_y^2),$$
$$\mu_{23} = 4p_y^2(1-p_y^2) + p_j^3(-1+12p_y+p_y^2) + p_j^2(1-15p_y+23p_y^2-p_y^3) + 4p_jp_y(1-7p_y-p_y^2-p_y^3),$$
$$\mu_{24} = p_j^3(6-14p_y) + p_j^2(-6+24p_y-26p_y^2) + p_j(1-15p_y+23p_y^2-p_y^3) + 2p_y(1-2p_y-p_y^2+2p_y^3).$$

Especially, if $p_y = 0$, the expressions of $\mathrm{Var}(\hat{r})$ and $\mathrm{Var}(\hat{\Delta})$ above can be simplified. In fact, for the case of homozygous reference individuals, we have

$$\mathrm{Var}(\hat{r}) = \frac{p_i}{2-2p_i} \quad \text{and} \quad \mathrm{Var}(\hat{\Delta}) = \frac{p_i^2}{(1-p_i)^2};$$

for the case of heterozygous reference individuals, we have

$$\mathrm{Var}(\hat{r}) = \frac{2p_ip_j}{p_i+p_j-4p_ip_j} \quad \text{and} \quad \mathrm{Var}(\hat{\Delta}) = \frac{2p_ip_j}{1-p_i-p_j+2p_ip_j}.$$

Using the Kronecker operator K_{ab}, both cases above can be unified to write as

$$\mathrm{Var}(\hat{r}) = \frac{2p_ap_b}{(1+K_{ab})(p_a+p_b)-4p_ap_b} \quad \text{and} \quad \mathrm{Var}(\hat{\Delta}) = \frac{2p_ap_b}{(1+K_{ab})(1-p_a-p_b)+2p_ap_b}.$$

The derivation of Expressions (9a) and (9b)

If the observed genotype of the reference individual x is a homozygote (i.e. $G'_x = A'_iA'_i$), letting ρ_1 be the probability of $S=1$ and G_y is visible, we have

$$\begin{aligned}
\rho_1 &= \Pr(G_x = A_iA_i \mid G'_x = A'_iA'_i)\Pr(G_y = A_iA_i \mid G_x = A_iA_i, G_y \neq A_yA_y) \\
&\quad + \Pr(G_x = A_iA_i \mid G'_x = A'_iA'_i)\Pr(G_y = A_iA_y \mid G_x = A_iA_i, G_y \neq A_yA_y) \\
&\quad + \Pr(G_x = A_iA_y \mid G'_x = A'_iA'_i)\Pr(G_y = A_iA_i \mid G_x = A_iA_y, G_y \neq A_yA_y) \\
&\quad + \Pr(G_x = A_iA_y \mid G'_x = A'_iA'_i)\Pr(G_y = A_iA_y \mid G_x = A_iA_y, G_y \neq A_yA_y) \\
&= \frac{p_i}{p_i+2p_y}\cdot\frac{p_i^2(1-\phi-\Delta)+p_i\phi+\Delta}{(1-p_y^2)(1-\phi-\Delta)+\phi+\Delta} + \frac{p_i}{p_i+2p_y}\cdot\frac{2p_ip_y(1-\phi-\Delta)+p_y\phi}{(1-p_y^2)(1-\phi-\Delta)+\phi+\Delta} \\
&\quad + \frac{2p_y}{p_i+2p_y}\cdot\frac{p_i^2(1-\phi-\Delta)+\frac{1}{2}p_i\phi}{(1-p_y^2)(1-\phi-\Delta)+(1-\frac{1}{2}p_y)\phi+\Delta} + \frac{2p_y}{p_i+2p_y}\cdot\frac{2p_ip_y(1-\phi-\Delta)+\frac{1}{2}(p_i+p_y)\phi+\Delta}{(1-p_y^2)(1-\phi-\Delta)+(1-\frac{1}{2}p_y)\phi+\Delta}.
\end{aligned}$$

As in Expression (8), an approximated expression of the probability ρ_1 is given by the following equation:

$$\begin{aligned}
\rho_1 &\approx \frac{p_i^3}{(p_i+2p_y)(1-p_y^2)}(1-\phi-\Delta) + \frac{p_i^2}{p_i+2p_y}\phi + \frac{p_i}{p_i+2p_y}\Delta + \frac{2p_i^2p_y}{(p_i+2p_y)(1-p_y^2)}(1-\phi-\Delta) \\
&\quad + \frac{p_ip_y}{p_i+2p_y}\phi + \frac{2p_i^2p_y}{(p_i+2p_y)(1-p_y^2)}(1-\phi-\Delta) + \frac{2p_ip_y}{(p_i+2p_y)(2-p_y)}\phi \\
&\quad + \frac{4p_ip_y^2}{(p_i+2p_y)(1-p_y^2)}(1-\phi-\Delta) + \frac{2p_y(p_i+p_y)}{(p_i+2p_y)(2-p_y)}\phi + \frac{2p_y}{p_i+2p_y}\Delta \\
&= \frac{p_i^2+2p_ip_y}{1-p_y^2}(1-\phi-\Delta) + \frac{p_i(p_i+p_y)(2-p_y)+2p_y(p_y+2p_i)}{(p_i+2p_y)(2-p_y)}\phi + \Delta \\
&= \frac{p_i^2+2p_ip_y}{1-p_y^2} + \Big(1-\frac{p_i^2+2p_ip_y}{1-p_y^2}\Big)\Delta + \Big[\frac{p_i(p_i+p_y)(2-p_y)+2p_y(p_y+2p_i)}{(p_i+2p_y)(2-p_y)} - \frac{p_i^2+2p_ip_y}{1-p_y^2}\Big]\phi.
\end{aligned}$$

Likewise, letting ρ_2 be the probability $\Pr(G'_y = A'_iA'_x \mid G'_x = A'_iA'_i, G_y \neq A_yA_y)$, we have

$$\begin{aligned}
\rho_2 &= \Pr(G_x = A_iA_i \mid G'_x = A'_iA'_i)\Pr(G_y = A_iA_x \mid G_x = A_iA_i, G_y \neq A_yA_y) \\
&\quad + \Pr(G_x = A_iA_y \mid G'_x = A'_iA'_i)\Pr(G_y = A_iA_x \mid G_x = A_iA_y, G_y \neq A_yA_y) \\
&= \frac{p_i}{p_i+2p_y}\cdot\frac{2p_ip_x(1-\phi-\Delta)+p_x\phi}{(1-p_y^2)(1-\phi-\Delta)+\phi+\Delta} + \frac{2p_y}{p_i+2p_y}\cdot\frac{2p_ip_x(1-\phi-\Delta)+\frac{1}{2}p_x\phi}{(1-p_y^2)(1-\phi-\Delta)+(1-\frac{1}{2}p_y)\phi+\Delta},
\end{aligned}$$

and an approximated expression of ρ_2 is given by

$$\begin{aligned}
\rho_2 &\approx \frac{2p_i^2p_x}{(p_i+2p_y)(1-p_y^2)}(1-\phi-\Delta) + \frac{p_ip_x}{p_i+2p_y}\phi + \frac{4p_ip_xp_y}{(p_i+2p_y)(1-p_y^2)}(1-\phi-\Delta) + \frac{2p_xp_y}{(p_i+2p_y)(2-p_y)}\phi \\
&= \frac{2p_ip_x}{1-p_y^2}(1-\phi-\Delta) + \frac{p_ip_x(2-p_y)+2p_xp_y}{(p_i+2p_y)(2-p_y)}\phi \\
&= \frac{2p_ip_x}{1-p_y^2} + \Big(0-\frac{2p_ip_x}{1-p_y^2}\Big)\Delta + \Big[\frac{p_ip_x(2-p_y)+2p_xp_y}{(p_i+2p_y)(2-p_y)} - \frac{2p_ip_x}{1-p_y^2}\Big]\phi.
\end{aligned}$$

Now, it is clear that the following expressions (i.e. the expressions in (9a)) hold:

$$\mathbf{E}^* = \begin{bmatrix} \frac{p_i^2+2p_ip_y}{1-p_y^2} \\ \frac{2p_ip_x}{1-p_y^2} \end{bmatrix} \quad \text{and} \quad \mathbf{M}^* = \begin{bmatrix} 1 & \frac{p_i(p_i+p_y)(2-p_y)+2p_y(p_y+2p_i)}{(p_i+2p_y)(2-p_y)} \\ 0 & \frac{p_ip_x(2-p_y)+2p_xp_y}{(p_i+2p_y)(2-p_y)} \end{bmatrix} - \left[\mathbf{E}^*, \mathbf{E}^*\right]$$

Next, if the observed genotype of the reference individual x is a heterozygote (i.e. $G'_x = A'_iA'_j$), then

$$\begin{aligned}
&\Pr(G'_y = A'_iA'_i \mid G'_x = A'_iA'_j, G_y \neq A_yA_y) \\
&= \Pr(G_x = A_iA_j \mid G'_x = A'_iA'_j)\Pr(G_y = A_iA_i \mid G_x = A_iA_j, G_y \neq A_yA_y) \\
&\quad + \Pr(G_x = A_iA_j \mid G'_x = A'_iA'_j)\Pr(G_y = A_iA_y \mid G_x = A_iA_j, G_y \neq A_yA_y) \\
&= 1 \cdot \frac{p_i^2(1-\phi-\Delta)+\frac{1}{2}p_i\phi}{(1-p_y^2)(1-\phi-\Delta)+\phi+\Delta} + 1 \cdot \frac{2p_ip_y(1-\phi-\Delta)+\frac{1}{2}p_y\phi}{(1-p_y^2)(1-\phi-\Delta)+\phi+\Delta} \\
&\approx \frac{p_i^2}{1-p_y^2}(1-\phi-\Delta) + \frac{1}{2}p_i\phi + \frac{2p_ip_y}{1-p_y^2}(1-\phi-\Delta) + \frac{1}{2}p_y\phi \\
&= \frac{p_i^2+2p_ip_y}{1-p_y^2}(1-\phi-\Delta) + \frac{p_i+p_y}{2}\phi \\
&= \frac{p_i^2+2p_ip_y}{1-p_y^2} + \Big(0 - \frac{p_i^2+2p_ip_y}{1-p_y^2}\Big)\Delta + \Big(\frac{p_i+p_y}{2} - \frac{p_i^2+2p_ip_y}{1-p_y^2}\Big)\phi,
\end{aligned}$$

$$\begin{aligned}
&\Pr(G'_y = A'_jA'_j \mid G'_x = A'_iA'_j, G_y \neq A_yA_y) \\
&= \Pr(G_x = A_iA_j \mid G'_x = A'_iA'_j)\Pr(G_y = A_jA_j \mid G_x = A_iA_j, G_y \neq A_yA_y) \\
&\quad + \Pr(G_x = A_iA_j \mid G'_x = A'_iA'_j)\Pr(G_y = A_jA_y \mid G_x = A_iA_j, G_y \neq A_yA_y) \\
&= 1 \cdot \frac{p_j^2(1-\phi-\Delta)+\frac{1}{2}p_j\phi}{(1-p_y^2)(1-\phi-\Delta)+\phi+\Delta} + 1 \cdot \frac{2p_jp_y(1-\phi-\Delta)+\frac{1}{2}p_y\phi}{(1-p_y^2)(1-\phi-\Delta)+\phi+\Delta} \\
&\approx \frac{p_j^2}{1-p_y^2}(1-\phi-\Delta) + \frac{1}{2}p_j\phi + \frac{2p_jp_y}{1-p_y^2}(1-\phi-\Delta) + \frac{1}{2}p_y\phi \\
&= \frac{p_j^2+2p_jp_y}{1-p_y^2}(1-\phi-\Delta) + \frac{p_j+p_y}{2}\phi \\
&= \frac{p_j^2+2p_jp_y}{1-p_y^2} + \Big(0 - \frac{p_j^2+2p_jp_y}{1-p_y^2}\Big)\Delta + \Big(\frac{p_j+p_y}{2} - \frac{p_j^2+2p_jp_y}{1-p_y^2}\Big)\phi,
\end{aligned}$$

$$\begin{aligned}
&\Pr(G'_y = A'_iA'_j \mid G'_x = A'_iA'_j, G_y \neq A_yA_y) \\
&= \Pr(G_x = A_iA_j \mid G'_x = A'_iA'_j)\Pr(G_y = A_iA_j \mid G_x = A_iA_j, G_y \neq A_yA_y) \\
&= 1 \cdot \frac{2p_ip_j(1-\phi-\Delta)+\frac{1}{2}(p_i+p_j)\phi+\Delta}{(1-p_y^2)(1-\phi-\Delta)+\phi+\Delta} \\
&\approx \frac{2p_ip_j}{1-p_y^2}(1-\phi-\Delta) + \frac{p_i+p_j}{2}\phi + \Delta \\
&= \frac{2p_ip_j}{1-p_y^2} + \Big(1 - \frac{2p_ip_j}{1-p_y^2}\Big)\Delta + \Big(\frac{p_i+p_j}{2} - \frac{2p_ip_j}{1-p_y^2}\Big)\phi,
\end{aligned}$$

$$\begin{aligned}
&\Pr(G'_y = A'_iA'_x \mid G'_x = A'_iA'_j, G_y \neq A_yA_y) \\
&= \Pr(G_x = A_iA_j \mid G'_x = A'_iA'_j)\Pr(G_y = A_iA_x | G_x = A_iA_j, G_y \neq A_yA_y) \\
&= 1 \cdot \frac{2p_ip_x(1-\phi-\Delta)+\frac{1}{2}p_x\phi}{(1-p_y^2)(1-\phi-\Delta)+\phi+\Delta}
\end{aligned}$$

$$
\begin{aligned}
&\approx \frac{2p_ip_x}{1-p_y^2}(1-\phi-\Delta)+\frac{p_x}{2}\phi \\
&= \frac{2p_ip_x}{1-p_y^2}+\Big(0-\frac{2p_ip_x}{1-p_y^2}\Big)\Delta+\Big(\frac{p_x}{2}-\frac{2p_ip_x}{1-p_y^2}\Big)\phi,
\end{aligned}
$$

$$
\begin{aligned}
&\Pr(G'_y = A'_jA'_x \mid G'_x = A'_iA'_j, G_y \neq A_yA_y) \\
&= \Pr(G_x = A_iA_j \mid G'_x = A'_iA'_j)\Pr(G_y = A_jA_x \mid G_x = A_iA_j, G_y \neq A_yA_y) \\
&= 1\cdot\frac{2p_jp_x(1-\phi-\Delta)+\frac{1}{2}p_x\phi}{(1-p_y^2)(1-\phi-\Delta)+\phi+\Delta} \\
&\approx \frac{2p_jp_x}{1-p_y^2}(1-\phi-\Delta)+\frac{p_x}{2}\phi \\
&= \frac{2p_jp_x}{1-p_y^2}+\Big(0-\frac{2p_jp_x}{1-p_y^2}\Big)\Delta+\Big(\frac{p_x}{2}-\frac{2p_jp_x}{1-p_y^2}\Big)\phi.
\end{aligned}
$$

Now, it is clear that the following expressions (i.e. the expressions of (9b)) are valid:

$$
\mathbf{E}^* = \begin{bmatrix} \frac{p_i^2+2p_ip_y}{1-p_y^2} \\ \frac{p_j^2+2p_jp_y}{1-p_y^2} \\ \frac{2p_ip_j}{1-p_y^2} \\ \frac{2p_ip_x}{1-p_y^2} \\ \frac{2p_jp_x}{1-p_y^2} \end{bmatrix} \quad \text{and} \quad \mathbf{M}^* = \begin{bmatrix} 0 & \frac{p_i+p_y}{2} \\ 0 & \frac{p_j+p_y}{2} \\ 1 & \frac{p_i+p_j}{2} \\ 0 & \frac{p_x}{2} \\ 0 & \frac{p_x}{2} \end{bmatrix} - \big[\mathbf{E}^*, \mathbf{E}^*\big].
$$

The derivation of Expression (12)

Wang's (2002) estimator uses the similarity index as defined in Expression (2), but it does not distinguish the reference and proband individuals. The probability that S is equal to a specific value is calculated by taken the sum of the probabilities of a series of genotype pairs.

It is known that the symbol a_m $(m = 1, 2, 3, 4)$ is used to denote the sum $\sum_i p_i^m$. Note that $\sum_i p_i = 1$, we see that $a_1 = 1$. Next, because

$$
\sum_{i\neq j} p_i^m p_j^n = \sum_i p_i^m\Big(\sum_{j\neq i} p_j^n\Big) = \sum_i p_i^m(a_n - p_i^n) = a_n\sum_i p_i^m - \sum_i p_i^{m+n},
$$

we obtain $\sum_{i\neq j} p_i^m p_j^n = a_m a_n - a_{m+n}$. Specially, by $a_1 = 1$, we have $\sum_{i\neq j} p_i p_j^n = a_n - a_{n+1}$. Moreover, since

$$
\begin{aligned}
\sum_{\substack{j\neq i\\ k\neq i\\ k\neq j}} p_i^m p_j p_k &= \sum_i p_i^m\Big[\sum_{j\neq i} p_j\Big(\sum_{\substack{k\neq i\\ k\neq j}} p_k\Big)\Big] = \sum_i p_i^m\Big[\sum_{j\neq i} p_j(1-p_i-p_j)\Big] \\
&= \sum_i p_i^m\Big[(1-p_i)\sum_{j\neq i} p_j - \sum_{j\neq i} p_j^2\Big] = \sum_i p_i^m[(1-p_i)^2 - (a_2 - p_i^2)] \\
&= (1-a_2)\sum_i p_i^m - 2\sum_i p_i^{m+1} + 2\sum_i p_i^{m+2},
\end{aligned}
$$

we get $\sum_{\substack{j\neq i\\ k\neq i\\ k\neq j}} p_i^m p_j p_k = (1-a_2)a_m - 2a_{m+1} + 2a_{m+2}$. Specially, we have

$$\sum_{\substack{j\neq i\\ k\neq i\\ k\neq j}} p_i^2 p_j p_k = (1-a_2)a_2 - 2a_3 + 2a_4 = a_2 - a_2^2 - 2a_3 + 2a_4,$$

$$\sum_{\substack{j\neq i\\ k\neq i\\ k\neq j}} p_i p_j p_k = (1-a_2)a_1 - 2a_2 + 2a_3 = 1 - 3a_2 + 2a_3.$$

Now, using these facts, we can calculate the following probabilities:

$$\begin{aligned}
\Pr(S=1) &= \sum_i \Pr(G_x = A_iA_i, G_y = A_iA_i) + \sum_{i\neq j} \Pr(G_x = A_iA_j, G_y = A_iA_j)\\
&= (1-\phi-\Delta)\sum_i p_i^4 + \phi\sum_i p_i^3 + \Delta\sum_i p_i^2 + (1-\phi-\Delta)\sum_{i\neq j} 2p_i^2p_j^2 + \phi\sum_{i\neq j} p_ip_j^2 + \Delta\sum_{i\neq j} p_ip_j\\
&= (1-\phi-\Delta)a_4 + \phi a_3 + \Delta a_2 + 2(1-\phi-\Delta)(a_2^2-a_4) + \phi(a_2-a_3) + \Delta(1-a_2)\\
&= (1-\phi-\Delta)(2a_2^2-a_4) + \phi a_2 + \Delta\\
&= (2a_2^2-a_4) + [1-(2a_2^2-a_4)]\Delta + [a_2-(2a_2^2-a_4)]\phi,
\end{aligned}$$

$$\begin{aligned}
\Pr(S=\tfrac{3}{4}) &= \sum_i \Pr(G_x = A_iA_i, G_y = A_iA_j) + \sum_{i\neq j} \Pr(G_x = A_iA_j, G_y = A_iA_i)\\
&= 2\sum_i \Pr(G_x = A_iA_i, G_y = A_iA_j)\\
&= (1-\phi-\Delta)\sum_{i\neq j} 4p_i^3p_j + \phi\sum_{i\neq j} 2p_i^2p_j\\
&= (1-\phi-\Delta)[4(a_3-a_4)] + \phi[2(a_2-a_3)]\\
&= (4a_3-4a_4) - (4a_3-4a_4)\Delta + [(2a_2-2a_3)-(4a_3-4a_4)]\phi,
\end{aligned}$$

$$\begin{aligned}
\Pr(S=\tfrac{1}{2}) &= \sum_{\substack{j\neq i\\ k\neq i\\ k\neq j}} \Pr(G_x = A_iA_j, G_y = A_iA_k)\\
&= (1-\phi-\Delta)\sum_{\substack{j\neq i\\ k\neq i\\ k\neq j}} 4p_i^2p_jp_k + \phi\sum_{\substack{j\neq i\\ k\neq i\\ k\neq j}} p_ip_jp_k\\
&= (1-\phi-\Delta)[4(a_2-a_2^2-2a_3+2a_4)] + \phi(1-3a_2+2a_3)\\
&= (1-\phi-\Delta)(4a_2-4a_2^2-8a_3+8a_4) + \phi(1-3a_2+2a_3)\\
&= (4a_2-4a_2^2-8a_3+8a_4) - (4a_2-4a_2^2-8a_3+8a_4)\Delta + (1-7a_2+4a_2^2+10a_3-8a_4)\phi.
\end{aligned}$$

Combining the results from the foregoing three probabilities, they can be written as the following expression (i.e. Expression (12)):

$$\begin{bmatrix} \Pr(S=1)\\ \Pr(S=\frac{3}{4})\\ \Pr(S=\frac{1}{2}) \end{bmatrix} = \begin{bmatrix} \lambda_1\\ \lambda_2\\ \lambda_3 \end{bmatrix} + \begin{bmatrix} 1-\lambda_1 & a_2-\lambda_1\\ -\lambda_2 & 2a_2-2a_3-\lambda_2\\ -\lambda_3 & 1-3a_2+2a_3-\lambda_3 \end{bmatrix} \begin{bmatrix} \Delta\\ \phi \end{bmatrix},$$

where $\lambda_1 = 2a_2^2 - a_4$, $\lambda_2 = 4a_3 - 4a_4$ and $\lambda_3 = 4a_2 - 4a_2^2 - 8a_3 + 8a_4$.

The derivation of Expressions (14a) and (14b)

It is known that b_m denotes the sum $\sum_{i\neq y} p_i^m$, then

$$\sum_{\substack{i\neq y\\ j\neq y\\ j\neq i}} p_i^m p_j^n = \sum_{i\neq y} p_i^m \Big(\sum_{\substack{j\neq y\\ j\neq i}} p_j^n \Big) = \sum_{i\neq y} p_i^m (b_n - p_i^n) = b_m b_n - b_{m+n},$$

$$\begin{aligned}\sum_{\substack{i\neq y,\, k\neq y\\ j\neq y,\, k\neq i\\ j\neq i,\, k\neq j}} p_i^m p_j p_k &= \sum_{i\neq y} p_i^m \Big[\sum_{\substack{j\neq y\\ j\neq i}} p_j \Big(\sum_{\substack{k\neq y\\ k\neq i\\ k\neq j}} p_k \Big) \Big] = \sum_{i\neq y} p_i^m \Big[\sum_{\substack{j\neq y\\ j\neq i}} p_j (b_1 - p_i - p_j) \Big] \\ &= \sum_{i\neq y} p_i^m [(b_1 - p_i)^2 - (b_2 - p_i^2)] = \sum_{i\neq y} p_i^m (b_1^2 - 2b_1 p_i - b_2 + 2p_i^2) \\ &= b_1^2 b_m - 2b_1 b_{m+1} - b_2 b_m + 2b_{m+2}.\end{aligned}$$

With the presence of null alleles, the probabilities of $S' = 1, \frac{3}{4}$ and $\frac{1}{2}$ can be calculated as follows:

$$\begin{aligned}&\Pr(S'=1)[(f_1 - f_3)\Delta + (f_2 - f_3)\phi + f_3)] \\ &= \sum_{i\neq y} \Pr(G'_x = A'_i A'_i, G'_y = A'_i A'_i) + \sum_{\substack{i\neq y\\ j\neq y\\ j\neq i}} \Pr(G'_x = A'_i A'_j, G'_y = A'_i A'_j) \\ &= \sum_{i\neq y} \Pr(G_x = A_i A_i, G_y = A_i A_i) + \sum_{i\neq y} \Pr(G_x = A_i A_y, G_y = A_i A_y) + \Big[\sum_{i\neq y} \Pr(G_x = A_i A_y, G_y = A_i A_i) \\ &\quad + \sum_{i\neq y} \Pr(G_x = A_i A_i, G_y = A_i A_y) \Big] + \sum_{\substack{i\neq y\\ j\neq y\\ j\neq i}} \Pr(G_x = A_i A_j, G_y = A_i A_j) \\ &= \sum_{i\neq y} \Pr(G_x = A_i A_i, G_y = A_i A_i) + \sum_{i\neq y} \Pr(G_x = A_i A_y, G_y = A_i A_y) \\ &\quad + 2\sum_{i\neq y} \Pr(G_x = A_i A_y, G_y = A_i A_i) + \sum_{\substack{i\neq y\\ j\neq y\\ j\neq i}} \Pr(G_x = A_i A_j, G_y = A_i A_j) \\ &= (1-\phi-\Delta)\sum_{i\neq y} p_i^4 + \phi \sum_{i\neq y} p_i^3 + \Delta \sum_{i\neq y} p_i^2 + (1-\phi-\Delta)\sum_{i\neq y} 4p_i^2 p_y^2 + \phi \sum_{i\neq y} (p_i^2 p_y + p_i p_y^2) + \Delta \sum_{i\neq y} 2p_i p_y \\ &\quad + 2\Big[(1-\phi-\Delta)\sum_{i\neq y} 2p_i^3 p_y + \phi \sum_{i\neq y} p_i^2 p_y \Big] + (1-\phi-\Delta) \sum_{\substack{i\neq y\\ j\neq y\\ j\neq i}} 2p_i^2 p_j^2 + \phi \sum_{\substack{i\neq y\\ j\neq y\\ j\neq i}} p_i p_j^2 + \Delta \sum_{\substack{i\neq y\\ j\neq y\\ j\neq i}} p_i p_j \\ &= (1-\phi-\Delta)b_4 + \phi b_3 + \Delta b_2 + (1-\phi-\Delta)(4p_y^2 b_2) + \phi(p_y b_2 + p_y^2 b_1) + \Delta(2p_y b_1) \\ &\quad + 2[(1-\phi-\Delta)(2p_y b_3) + \phi p_y b_2] + (1-\phi-\Delta)[2(b_2^2 - b_4)] + \phi(b_1 b_2 - b_3) + \Delta(b_1^2 - b_2) \\ &= (1-\phi-\Delta)(2b_2^2 - b_4 + 4p_y^2 b_2 + 4p_y b_3) + \phi(b_1 b_2 + p_y^2 b_1 + 3p_y b_2) + \Delta(b_1^2 + 2p_y b_1) \\ &= \lambda_1 + (b_1^2 + 2p_y b_1 - \lambda_1)\Delta + (b_1 b_2 + p_y^2 b_1 + 3p_y b_2 - \lambda_1)\phi \quad (\text{where } \lambda_1 = 2b_2^2 - b_4 + 4p_y^2 b_2 + 4p_y b_3),\end{aligned}$$

$$\Pr(S' = \tfrac{3}{4})[(f_1 - f_3)\Delta + (f_2 - f_3)\phi + f_3)]$$

$$= \sum_{\substack{i\neq y\\ j\neq y\\ j\neq i}} \Pr(G'_x = A'_iA'_j, G'_y = A'_iA'_i) + \sum_{\substack{i\neq y\\ j\neq y\\ j\neq i}} \Pr(G'_x = A'_iA'_i, G'_y = A'_iA'_j)$$

$$= 2\sum_{i\neq y} \Pr(G'_x = A'_iA'_j, G'_y = A'_iA'_i)$$

$$= 2\sum_{i\neq y} \Pr(G_x = A_iA_j, G_y = A_iA_i) + 2\sum_{i\neq y} \Pr(G_x = A_iA_j, G_y = A_iA_y)$$

$$= 2(1-\phi-\Delta) \sum_{\substack{i\neq y\\ j\neq y\\ j\neq i}} 2p_i^3p_j + 2\phi \sum_{\substack{i\neq y\\ j\neq y\\ j\neq i}} p_i^2p_j + 2(1-\phi-\Delta) \sum_{\substack{i\neq y\\ j\neq y\\ j\neq i}} 4p_i^2p_jp_y + 2\phi \sum_{\substack{i\neq y\\ j\neq y\\ j\neq i}} p_ip_jp_y$$

$$= 2(1-\phi-\Delta)[2(b_1b_3 - b_4)] + 2\phi(b_1b_2 - b_3) + 2(1-\phi-\Delta)[4p_y(b_1b_2 - b_3)] + 2\phi[p_y(b_1^2 - b_2)]$$

$$= (1-\phi-\Delta)(4b_1b_3 - 4b_4 + 8p_yb_1b_2 - 8p_yb_3) + \phi(2b_1b_2 - 2b_3 + 2p_yb_1^2 - 2p_yb_2)$$

$$= \lambda_2 + (0-\lambda_2)\Delta + (2b_1b_2 - 2b_3 + 2p_yb_1^2 - 2p_yb_2 - \lambda_2)\phi \quad (\text{where } \lambda_2 = 4b_1b_3 - 4b_4 + 8p_yb_1b_2 - 8p_yb_3),$$

$$\Pr(S' = \tfrac{1}{2})[(f_1 - f_3)\Delta + (f_2 - f_3)\phi + f_3)]$$

$$= \sum_{\substack{i\neq y,\, k\neq y\\ j\neq y,\, k\neq i\\ j\neq i,\, k\neq j}} \Pr(G'_x = A'_iA'_j, G'_y = A'_iA'_k)$$

$$= \sum_{\substack{i\neq y,\, k\neq y\\ j\neq y,\, k\neq i\\ j\neq i,\, k\neq j}} \Pr(G_x = A_iA_j, G_y = A_iA_k)$$

$$= (1-\phi-\Delta) \sum_{\substack{i\neq y,\, k\neq y\\ j\neq y,\, k\neq i\\ j\neq i,\, k\neq j}} 4p_i^2p_jp_k + \phi \sum_{\substack{i\neq y,\, k\neq y\\ j\neq y,\, k\neq i\\ j\neq i,\, k\neq j}} p_ip_jp_k$$

$$= (1-\phi-\Delta)[4(b_1^2b_2 - 2b_1b_3 - b_2^2 + 2b_4)] + \phi(b_1^3 - 2b_1b_2 - b_2b_1 + 2b_3)$$

$$= \lambda_3 + (0-\lambda_3)\Delta + (b_1^3 - 3b_1b_2 + 2b_3 - \lambda_3)\phi \quad (\text{where } \lambda_3 = 4b_1^2b_2 - 8b_1b_3 - 4b_2^2 + 8b_4).$$

Note that $\mathbf{P'}$ and $\lambda_1, \lambda_2, \lambda_3$ are respectively

$$\mathbf{P'} = \begin{bmatrix} \Pr(S'=1) \\ \Pr(S'=\frac{3}{4}) \\ \Pr(S'=\frac{1}{2}) \end{bmatrix} \quad \text{and} \quad \begin{cases} \lambda_1 = 2b_2^2 - b_4 + 4p_y^2b_2 + 4p_yb_3, \\ \lambda_2 = 4b_1b_3 - 4b_4 + 8p_yb_1b_2 - 8p_yb_3, \\ \lambda_3 = 4b_1^2b_2 - 8b_1b_3 - 4b_2^2 + 8b_4. \end{cases}$$

It is clear from the calculated results above that the matrices $\mathbf{E'}$ and $\mathbf{M'}$ in Expression (13), i.e. the expression $\mathbf{P'} = [(f_1 - f_3)\Delta + (f_2 - f_3)\phi + f_3]^{-1}(\mathbf{E'} + \mathbf{M'\Delta})$, are as follows:

$$\mathbf{E'} = \begin{bmatrix} 2b_2^2 - b_4 + 4p_y^2b_2 + 4p_yb_3 \\ 4b_1b_3 - 4b_4 + 8p_yb_1b_2 - 8p_yb_3 \\ 4b_1^2b_2 - 8b_1b_3 - 4b_2^2 + 8b_4 \end{bmatrix},$$

$$\mathbf{M'} = \begin{bmatrix} b_1^2 + 2p_yb_1 & b_1b_2 + p_y^2b_1 + 3p_yb_2 \\ 0 & 2b_1b_2 - 2b_3 + 2p_yb_1^2 - 2p_yb_2 \\ 0 & b_1^3 - 3b_1b_2 + 2b_3 \end{bmatrix} - \big[\mathbf{E'}, \mathbf{E'}\big].$$

The last two expressions are just Expressions (14a) and (14b).

The biases of various estimators before and after correction

The bias curves (Figures 2 and 3 in the article) of these estimators can be divided into three categories: (i) L&R and NA; (ii) WA, TH and NB; (iii) A&W. The bias of the L&R and NA estimators are highly independent on the number of observed alleles, whilst those of the WA, TH and NB estimators are not. Because the NA estimator is a modification of the L&R estimator, and because the TH and NB estimators are modifications of the WA estimator, the curves within a category are similar. We derived the bias curve of method-of-moment estimators. For Anderson and Weir's (2007) estimator, $\hat{r}$ is obtained from a numerical algorithm, so the analytical solution can not be solved.

Lynch and Ritland's (1999) estimator

We enumerate all possible genotype pairs of the two individuals x and y, and weighted the $\hat{r}$ by the probability of each genotype pair to obtain the symbolic solution of the bias of $\hat{r}$.

$$\begin{aligned}\text{Bias}(\hat{r}) = \Big\{ \sum_{i=1}^{n}\sum_{j=1}^{n} \Big[(1-\phi-\Delta)\sum_{k=1}^{n}\sum_{l=1}^{n} p_i p_j p_k p_l V(G_x,G_y)\hat{r}(G'_x,G'_y) + \phi \sum_{k=1}^{n}\sum_{l=i}^{i} p_i p_j p_k V(G_x,G_y)\hat{r}(G'_x,G'_y) \\ +\Delta \sum_{k=i}^{i}\sum_{l=j}^{j} p_i p_j V(G_x,G_y)\hat{r}(G'_x,G'_y)\Big]\Big\} \Big/ [(f_1-f_3)\Delta + (f_2-f_3)\phi + f_3] - \Delta - \phi/2. \qquad \text{(iii)}\end{aligned}$$

Where $(f_1-f_3)\Delta + (f_2-f_3)\phi + f_3$ is the probability that the genotypes of both individuals are visible, $V(G_x,G_y)$ is a binary function to test whether the both genotypes $G_x = A_iA_j$ and $G_y = A_kA_l$ are visible, which is defined by

$$V(G_x,G_y) = \begin{cases} 1 & \text{if } G_x \neq A_yA_y \text{ and } G_y \neq A_yA_y, \\ 0 & \text{otherwise.} \end{cases} \qquad \text{(iv)}$$

In Expression (iii), G'_x and G'_y are the observed genotypes and $\hat{r}(G'_x,G'_y)$ is the estimated relatedness of Lynch and Ritland's (1999) estimator (Equation (5)). Note that before correction, the observed allele frequencies are also not the true values: the sum of visible allele frequencies is extended to one, say $\sum_{i\neq y} p'_i = 1$ and $p'_i = \frac{p_i}{1-p_y}$ $(i \neq y)$. If the frequencies of visible alleles are drawn from a triangular distribution, then $p_i = \frac{2i}{n'(n'+1)}$ and $p'_i = \frac{2i(1-p_y)}{n'(n'+1)}$ $(i \neq y)$, and the Bias$(\hat{r})$ can be solved, for $n' \geq 2$:

$$\text{Bias}(\hat{r}) = \frac{(\Delta+\phi/2)(1+p_y)}{1+p_y+(-1+\Delta)p_y^2+(-1+\Delta+\phi)p_y^3}.$$

From the expression above, we can find that Bias$(\hat{r})$ is not a function of n'. Therefore, in Figure 2, the bias curve of L&R estimator is independent on n'.

Corrected Lynch and Ritland's (1999) estimator

After correction, Bias$(\hat{r})$ can be solved by using the same way, with the $\hat{r}(G'_x,G'_y)$ in Expression (iv) been replaced with the corrected L&R estimator. The Bias$(\hat{r})$ is dependent on n'. Although we fail to

obtain the general form, the solution of Bias($\hat{r}$) for each n' can be calculated. Because these expressions become more complex, we only show the bias of $\hat{r}$ for $2 \le n' \le 6$, and the numerical solution is shown in Table 2. In the following, we let $\mu = 1 + p_y + (-1+\Delta)p_y^2 + (-1+\Delta+\phi)p_y^3$, then

(1) Bias$(\hat{r}, n'=2) = -\Delta - \phi/2 + (2\lambda_{21}\Delta + \lambda_{22}\phi)(\lambda_{23}\mu)^{-1}$, where

$$\begin{aligned}
\lambda_{21} &= 4 + 2p_y + 48p_y^2 + 51p_y^3 + 171p_y^4 + 261p_y^5 + 101p_y^6 + 10p_y^7,\\
\lambda_{22} &= 4 + 2p_y + 46p_y^2 + 51p_y^3 + 132p_y^4 + 252p_y^5 - 85p_y^6 - 71p_y^7 - 7p_y^8,\\
\lambda_{23} &= 2(2 + p_y + 13p_y^2 + 2p_y^3)(2 - 2p_y + 13p_y^2 + 5p_y^3);
\end{aligned}$$

(2) Bias$(\hat{r}, n'=3) = -\Delta - \phi/2 + (2\lambda_{31}\Delta + \lambda_{32}\phi)(\lambda_{33}\mu)^{-1}$, where

$$\begin{aligned}
\lambda_{31} &= 40 + 180p_y + 998p_y^2 + 2691p_y^3 + 8226p_y^4 + 13689p_y^5 + 24068p_y^6 + 23877p_y^7\\
&\quad + 8098p_y^8 + 1035p_y^9 + 42p_y^{10},\\
\lambda_{32} &= 40 + 180p_y + 978p_y^2 + 2591p_y^3 + 7593p_y^4 + 11933p_y^5 + 18176p_y^6 + 15606p_y^7\\
&\quad - 9711p_y^8 - 5231p_y^9 - 660p_y^{10} - 23p_y^{11},\\
\lambda_{33} &= 2(2 + p_y + 13p_y^2 + 2p_y^3)(2 - p_y + 12p_y^2 + 3p_y^3)(10 + 35p_y + 92p_y^2 + 7p_y^3);
\end{aligned}$$

(3) Bias$(\hat{r}, n'=4) = -\Delta - \phi/2 + (2\lambda_{41}\Delta + \lambda_{42}\phi)(\lambda_{43}\lambda_{44}\mu)^{-1}$, where

$$\begin{aligned}
\lambda_{41} &= 18144 + 214704p_y + 1310112p_y^2 + 5939124p_y^3 + 19865538p_y^4 + 52830058p_y^5\\
&\quad + 111221531p_y^6 + 176504833p_y^7 + 218922427p_y^8 + 157800497p_y^9 + 48247633p_y^{10}\\
&\quad + 6701775p_y^{11} + 414615p_y^{12} + 9009p_y^{13},\\
\lambda_{42} &= 18144 + 214704p_y + 1301040p_y^2 + 5818164p_y^3 + 19064214p_y^4 + 48926482p_y^5\\
&\quad + 97839029p_y^6 + 139971751p_y^7 + 144883154p_y^8 + 47354737p_y^9 - 70687906p_y^{10}\\
&\quad - 30301523p_y^{11} - 4170916p_y^{12} - 227003p_y^{13} - 4071p_y^{14},\\
\lambda_{43} &= 2(4 + 10p_y + 33p_y^2 + 3p_y^3)(6 + 37p_y^2 + 7p_y^3),\\
\lambda_{44} &= (18 + 135p_y + 236p_y^2 + 11p_y^3)(42 + 35p_y + 284p_y^2 + 39p_y^3);
\end{aligned}$$

(4) Bias$(\hat{r}, n'=5) = -\Delta - \phi/2 + (2\lambda_{51}\Delta + \lambda_{52}\phi)(\lambda_{53}\lambda_{54}\mu)^{-1}$, where

$$\begin{aligned}
\lambda_{51} &= 128128 + 2914912p_y + 27446800p_y^2 + 171610800p_y^3 + 779358084p_y^4\\
&\quad + 2753453190p_y^5 + 7656153022p_y^6 + 17031462637p_y^7 + 29663170036p_y^8\\
&\quad + 39608425347p_y^9 + 37781928630p_y^{10} + 21577639263p_y^{11} + 6040014268p_y^{12}\\
&\quad + 864678919p_y^{13} + 64270024p_y^{14} + 2314932p_y^{15} + 31008p_y^{16},\\
\lambda_{52} &= 128128 + 2914912p_y + 27382736p_y^2 + 169993184p_y^3 + 762706068p_y^4\\
&\quad + 2642767874p_y^5 + 7131386478p_y^6 + 15135519959p_y^7 + 24384594730p_y^8\\
&\quad + 28131218262p_y^9 + 19033914707p_y^{10} - 787700001p_y^{11} - 10294303102p_y^{12}\\
&\quad - 3754393974p_y^{13} - 536626095p_y^{14} - 35896604p_y^{15} - 1095314p_y^{16} - 11948p_y^{17},\\
\lambda_{53} &= 2(2 + p_y + 13p_y^2 + 2p_y^3)(4 + 10p_y + 33p_y^2 + 3p_y^3)(14 + 175p_y + 253p_y^2 + 8p_y^3),
\end{aligned}$$

$$\lambda_{54} = (26 + 130p_y + 277p_y^2 + 17p_y^3)(44 + 55p_y + 313p_y^2 + 38p_y^3);$$

(5) $\text{Bias}(\hat{r}, n' = 6) = -\Delta - \phi/2 + (2\lambda_{61}\Delta + \lambda_{62}\phi)(\lambda_{63}\lambda_{64}\lambda_{65}\mu)^{-1}$, where

$$\begin{aligned}
\lambda_{61} = {} & 248064000 + 9289996800p_y + 137054051200p_y^2 + 1206977214560p_y^3 \\
& + 7478912348960p_y^4 + 34971140501296p_y^5 + 128318338268268p_y^6 \\
& + 376010959730232p_y^7 + 885015211256668p_y^8 + 1662157794116498p_y^9 \\
& + 2441370328609961p_y^{10} + 2701406869519177p_y^{11} + 2087707108083687p_y^{12} \\
& + 998780278713489p_y^{13} + 258916743498697p_y^{14} + 37328014277531p_y^{15} \\
& + 3064633669073p_y^{16} + 140507178121p_y^{17} + 3282998190p_y^{18} + 29601000p_y^{19}, \\
\lambda_{62} = {} & 248064000 + 9289996800p_y + 136930019200p_y^2 + 1201898104160p_y^3 \\
& + 7397157964960p_y^4 + 34202495888336p_y^5 + 123325558158108p_y^6 \\
& + 351945664942868p_y^7 + 795449297019884p_y^8 + 1400682977460974p_y^9 \\
& + 1840758497928775p_y^{10} + 1631376876416498p_y^{11} + 656223687892950p_y^{12} \\
& - 345308229230269p_y^{13} - 499737648215188p_y^{14} - 160563926216347p_y^{15} \\
& - 23264813191808p_y^{16} - 1752919253641p_y^{17} - 69812259099p_y^{18} \\
& - 1360693027p_y^{19} - 9949430p_y^{20}, \\
\lambda_{63} = {} & 2(6 + 27p_y + 61p_y^2 + 4p_y^3)(10 + 10p_y + 69p_y^2 + 9p_y^3), \\
\lambda_{64} = {} & (20 + 370p_y + 481p_y^2 + 11p_y^3)(38 + 304p_y + 517p_y^2 + 23p_y^3), \\
\lambda_{65} = {} & (68 + 187p_y + 577p_y^2 + 50p_y^3)(80 + 136p_y + 601p_y^2 + 65p_y^3).
\end{aligned}$$

Noval estimator A

The bias curve of NA estimator can also be obtained by using the same methods as Lynch and Ritland's (1999) estimator, with the $\hat{r}(G'_x, G'_y)$ calculated using Equation (10). However, as shown in Figure 2, although the curves of the NA estimator are similar to those of the L&R estimator, the bias curves are dependent on n'. We fail to obtain the general form, but the solution of $\text{Bias}(\hat{r})$ for each n' can be calculated. The calculated results for $2 \le n' \le 6$ are as follows.

(1) $\text{Bias}(\hat{r}, n' = 2) = -\Delta - \phi/2 + (\lambda_{21} + \lambda_{22}\Delta + \lambda_{23}\phi)(\lambda_{24} + \lambda_{25}\Delta + \lambda_{26}\phi)^{-1}$, where

$$\begin{aligned}
&\lambda_{21} = (-1 + p_y)^2(7 + 11p_y), && \lambda_{22} = 38 + 120p_y + 15p_y^2 - 11p_y^3, \\
&\lambda_{23} = 20 + 39p_y + 33p_y^2 - 11p_y^3, && \lambda_{24} = (1 - p_y)(5 + 13p_y)^2, \\
&\lambda_{25} = 20 + 12p_y - 39p_y^2 + 169p_y^3, && \lambda_{26} = 2 + 12p_y - 102p_y^2 + 169p_y^3;
\end{aligned}$$

(2) $\text{Bias}(\hat{r}, n' = 3) = -\Delta - \phi/2 + (\lambda_{31} + 2\lambda_{32}\Delta + \lambda_{33}\phi)\lambda_{34}^{-1}$,

$$\begin{aligned}
\lambda_{31} &= 1718404(1 - p_y)^2 p_y, \\
\lambda_{32} &= 1005795 + 1864997p_y - 1718404p_y^2 + 859202p_y^3, \\
\lambda_{33} &= 1005795 + 2286403p_y - 2999012p_y^2 + 1718404p_y^3,
\end{aligned}$$

Table 2: The bias of $\hat{r}$ of Lynch and Ritland's (1999) after correction

p_y	n'	Relationship				n'	Relationship			
		Unrelated	Half-sibs	Full-sibs	Parent -offspring		Unrelated	Half-sibs	Full-sibs	Parent -offspring
	2	0.000000	0.001124	0.002266	0.002019	9	0.000000	0.000744	0.001887	0.001259
	3	0.000000	0.001043	0.002185	0.001857	10	0.000000	0.000713	0.001856	0.001196
	4	0.000000	0.000973	0.002115	0.001717	11	0.000000	0.000685	0.001828	0.001140
0.1	5	0.000000	0.000914	0.002056	0.001598	12	0.000000	0.000659	0.001802	0.001089
	6	0.000000	0.000863	0.002005	0.001496	13	0.000000	0.000636	0.001779	0.001042
	7	0.000000	0.000818	0.001961	0.001407	14	0.000000	0.000614	0.001757	0.000999
	8	0.000000	0.000779	0.001922	0.001328	15	0.000000	0.000595	0.001738	0.000960
	2	0.000000	0.006906	0.015919	0.008079	9	0.000000	0.004589	0.013657	0.003496
	3	0.000000	0.006392	0.015417	0.007062	10	0.000000	0.004449	0.013521	0.003219
	4	0.000000	0.005896	0.014934	0.006082	11	0.000000	0.004329	0.013404	0.002983
0.3	5	0.000000	0.005506	0.014553	0.005311	12	0.000000	0.004226	0.013303	0.002779
	6	0.000000	0.005200	0.014254	0.004705	13	0.000000	0.004136	0.013215	0.002601
	7	0.000000	0.004955	0.014015	0.004220	14	0.000000	0.004057	0.013138	0.002444
	8	0.000000	0.004755	0.013819	0.003825	15	0.000000	0.003986	0.013069	0.002305
	2	0.000000	0.016953	0.040103	0.007212	9	0.000000	0.014792	0.038100	0.003105
	3	0.000000	0.016599	0.039775	0.006538	10	0.000000	0.014654	0.037972	0.002843
	4	0.000000	0.016115	0.039326	0.005618	11	0.000000	0.014538	0.037864	0.002622
0.5	5	0.000000	0.015718	0.038958	0.004864	12	0.000000	0.014438	0.037772	0.002432
	6	0.000000	0.015406	0.038669	0.004270	13	0.000000	0.014352	0.037692	0.002268
	7	0.000000	0.015157	0.038439	0.003799	14	0.000000	0.014276	0.037622	0.002124
	8	0.000000	0.014957	0.038253	0.003418	15	0.000000	0.014209	0.037560	0.001997
	2	0.000000	0.026800	0.059042	0.005383	9	0.000000	0.025171	0.057596	0.002408
	3	0.000000	0.026588	0.058854	0.004996	10	0.000000	0.025060	0.057498	0.002206
	4	0.000000	0.026222	0.058529	0.004328	11	0.000000	0.024966	0.057415	0.002034
0.6	5	0.000000	0.025911	0.058253	0.003760	12	0.000000	0.024885	0.057343	0.001887
	6	0.000000	0.025663	0.058033	0.003307	13	0.000000	0.024815	0.057281	0.001759
	7	0.000000	0.025464	0.057857	0.002944	14	0.000000	0.024754	0.057227	0.001648
	8	0.000000	0.025303	0.057714	0.002650	15	0.000000	0.024700	0.057179	0.001549

$$\lambda_{34} = 2011590[1 + p_y + (-1 + \Delta)p_y^2 + (-1 + \Delta + \phi)p_y^3];$$

(3) $\text{Bias}(\hat{r}, n' = 4) = -\Delta - \phi/2 + (\lambda_{41} + 2\lambda_{42}\Delta + \lambda_{43}\phi)\lambda_{44}^{-1}$, where

$$\lambda_{41} = -624091473188(-1 + p_y)^2 p_y,$$

$$\lambda_{42} = 602638296435 + 914684033029p_y - 624091473188p_y^2 + 312045736594p_y^3,$$

$$\lambda_{43} = 602638296435 + 923267324083p_y - 944720500836p_y^2 + 624091473188p_y^3,$$

$$\lambda_{44} = 1205276592870[1 + p_y + (-1 + \Delta)p_y^2 + (-1 + \Delta + \phi)p_y^3];$$

(4) $\text{Bias}(\hat{r}, n' = 5) = -\Delta - \phi/2 + (\lambda_{51} + 2\lambda_{52}\Delta + \lambda_{53}\phi)\lambda_{54}^{-1}$, where

$$\lambda_{51} = 19685159108451104(1 - p_y)^2 p_y,$$

$$\lambda_{52} = 26698375509931125 + 36540955064156677p_y - 19685159108451104p_y^2 + 9842579554225552p_y^3,$$

$$\lambda_{53} = 26698375509931125 + 32592836333532839p_y - 25579619932052818p_y^2 + 19685159108451104p_y^3,$$

$$\lambda_{54} = 53396751019862250[1 + p_y + (-1 + \Delta)p_y^2 + (-1 + \Delta + \phi)p_y^3];$$

(5) $\text{Bias}(\hat{r}, n' = 6) = -\Delta - \phi/2 + (\lambda_{61} + \lambda_{62} - \lambda_{63} - \lambda_{64})\lambda_{65}^{-1}$, where

$$\lambda_{61} = \Delta(-1 + p_y)^2 - 1/2\phi(-1 + p_y)^3 - 2\Delta(-1 + p_y)p_y,$$

$$\lambda_{62} = \frac{17752956739302371838068178659863\phi(-1 + p_y)^2 p_y}{1430964302495356705920310406940 0},$$

$$\lambda_{63} = \frac{44309281638188706677580887167 81(-1 + \Delta + \phi)(-1 + p_y)^3 p_y}{154670406225601055713445316044 25},$$

$$\lambda_{64} = \phi(-1 + p_y)p_y^2,$$

$$\lambda_{65} = 1 + (-2 + \Delta)p_y^2 + \phi p_y^3 - (-1 + \Delta + \phi)p_y^4.$$

Because the expressions for a large n' are complex, the approximate numerical solution for $n' \geq 6$ is calculated by the following formula:

$$\text{Bias}(\hat{r}) \approx -\Delta - \phi/2 + (\lambda_1 + \lambda_2\mu_1 - \lambda_3\mu_2 - \lambda_4)\lambda_5^{-1},$$

where

$$\lambda_1 = \Delta(-1 + p_y)^2 - 0.5\phi(-1 + p_y)^3 - 2\Delta(-1 + p_y)p_y,$$

$$\lambda_2 = \phi(-1 + p_y)^2 p_y,$$

$$\lambda_3 = (-1 + \Delta + \phi)(-1 + p_y)^3 p_y,$$

$$\lambda_4 = \phi(-1 + p_y)p_y^2,$$

$$\lambda_5 = 1 + (-2 + \Delta)p_y^2 + \phi p_y^3 - (-1 + \Delta + \phi)p_y^4,$$

and the values of μ_1 and μ_2 are listed in the following table:

n'	μ_1	μ_2	n'	μ_1	μ_2
6	1.24063	0.286475	11	1.24316	0.136650
7	1.24085	0.234555	12	1.24364	0.123819
8	1.24142	0.198748	13	1.24406	0.113206
9	1.24204	0.172530	14	1.24443	0.104279
10	1.24263	0.152485	15	1.24477	0.0966651

Corrected Noval estimator A

The expressions of Bias($\hat{r}$) become complex after correction, and we can only shown those for $2 \leq n' \leq 6$, with the numerical solution shown in Table 3. At $n' = 2$, some genotype pairs encounter the problem of singular matrices. In these cases, $\hat{r}$ is unsolvable. Therefore these results are not incorporated for numerical solution, and the analytical solution at $n' = 2$ is also not given.

(1) $\text{Bias}(\hat{r}, n' = 3) = -\Delta - \phi/2 + (2\lambda_{31}\Delta + \lambda_{32}\phi)(\lambda_{33}\lambda_{34})^{-1}$, where

$$\begin{aligned}\lambda_{31} &= 40 + 180p_y + 998p_y^2 + 2691p_y^3 + 8226p_y^4 + 13689p_y^5 + 24068p_y^6 + 23877p_y^7 \\ &\quad + 8098p_y^8 + 1035p_y^9 + 42p_y^{10}, \\ \lambda_{32} &= 40 + 180p_y + 978p_y^2 + 2591p_y^3 + 7593p_y^4 + 11933p_y^5 + 18176p_y^6 + 15606p_y^7 \\ &\quad - 9711p_y^8 - 5231p_y^9 - 660p_y^{10} - 23p_y^{11}, \\ \lambda_{33} &= 2(2 + p_y + 13p_y^2 + 2p_y^3)(2 - p_y + 12p_y^2 + 3p_y^3)(10 + 35p_y + 92p_y^2 + 7p_y^3), \\ \lambda_{34} &= 1 + p_y + (-1 + \Delta)p_y^2 + (-1 + \Delta + \phi)p_y^3;\end{aligned}$$

(2) $\text{Bias}(\hat{r}, n' = 4) = -\Delta - \phi/2 + (2\lambda_{41}\Delta + \lambda_{42}\phi)(\lambda_{43}\lambda_{44})^{-1}$, where

$$\begin{aligned}\lambda_{41} &= 18144 + 214704p_y + 1310112p_y^2 + 5939124p_y^3 + 19865538p_y^4 + 52830058p_y^5 \\ &\quad + 111221531p_y^6 + 176504833p_y^7 + 218922427p_y^8 + 157800497p_y^9 + 48247633p_y^{10} \\ &\quad + 6701775p_y^{11} + 414615p_y^{12} + 9009p_y^{13}, \\ \lambda_{42} &= 18144 + 214704p_y + 1301040p_y^2 + 5818164p_y^3 + 19064214p_y^4 + 48926482p_y^5 \\ &\quad + 97839029p_y^6 + 139971751p_y^7 + 144883154p_y^8 + 47354737p_y^9 - 70687906p_y^{10} \\ &\quad - 30301523p_y^{11} - 4170916p_y^{12} - 227003p_y^{13} - 4071p_y^{14}, \\ \lambda_{43} &= 2(4 + 10p_y + 33p_y^2 + 3p_y^3)(6 + 37p_y^2 + 7p_y^3)(18 + 135p_y + 236p_y^2 + 11p_y^3), \\ \lambda_{44} &= (42 + 35p_y + 284p_y^2 + 39p_y^3)[1 + p_y + (-1 + \Delta)p_y^2 + (-1 + \Delta + \phi)p_y^3];\end{aligned}$$

(3) $\text{Bias}(\hat{r}, n' = 5) = -\Delta - \phi/2 + (2\lambda_{51}\Delta + \lambda_{52}\phi)(\lambda_{53}\lambda_{54}\lambda_{55})^{-1}$, where

$$\begin{aligned}\lambda_{51} &= 128128 + 2914912p_y + 27446800p_y^2 + 171610800p_y^3 + 779358084p_y^4 \\ &\quad + 2753453190p_y^5 + 7656153022p_y^6 + 17031462637p_y^7 + 29663170036p_y^8 \\ &\quad + 39608425347p_y^9 + 37781928630p_y^{10} + 21577639263p_y^{11} + 6040014268p_y^{12} \\ &\quad + 864678919p_y^{13} + 64270024p_y^{14} + 2314932p_y^{15} + 31008p_y^{16},\end{aligned}$$

$$\lambda_{52} = 128128 + 2914912p_y + 27382736p_y^2 + 169993184p_y^3 + 762706068p_y^4$$
$$+ 2642767874p_y^5 + 7131386478p_y^6 + 15135519959p_y^7 + 24384594730p_y^8$$
$$+ 28131218262p_y^9 + 19033914707p_y^{10} - 787700001p_y^{11} - 10294303102p_y^{12}$$
$$- 3754393974p_y^{13} - 536626095p_y^{14} - 35896604p_y^{15} - 1095314p_y^{16} - 11948p_y^{17},$$
$$\lambda_{53} = 2(2 + p_y + 13p_y^2 + 2p_y^3)(4 + 10p_y + 33p_y^2 + 3p_y^3),$$
$$\lambda_{54} = (14 + 175p_y + 253p_y^2 + 8p_y^3)(26 + 130p_y + 277p_y^2 + 17p_y^3),$$
$$\lambda_{55} = (44 + 55p_y + 313p_y^2 + 38p_y^3)[1 + p_y + (-1 + \Delta)p_y^2 + (-1 + \Delta + \phi)p_y^3];$$

(4) $\text{Bias}(\hat{r}, n' = 6) = -\Delta - \phi/2 + (2\lambda_{61}\Delta + \lambda_{62}\phi)(\lambda_{63}\lambda_{64}\lambda_{65}\lambda_{66})^{-1}$, where

$$\lambda_{61} = 248064000 + 9289996800p_y + 137054051200p_y^2 + 1206977214560p_y^3$$
$$+ 7478912348960p_y^4 + 34971140501296p_y^5 + 128318338268268p_y^6$$
$$+ 376010959730232p_y^7 + 885015211256668p_y^8 + 1662157794116498p_y^9$$
$$+ 2441370328609961p_y^{10} + 2701406869519177p_y^{11} + 2087707108083687p_y^{12}$$
$$+ 998780278713489p_y^{13} + 258916743498697p_y^{14} + 37328014277531p_y^{15}$$
$$+ 3064633669073p_y^{16} + 140507178121p_y^{17} + 3282998190p_y^{18} + 29601000p_y^{19},$$
$$\lambda_{62} = 248064000 + 9289996800p_y + 136930019200p_y^2 + 1201898104160p_y^3$$
$$+ 7397157964960p_y^4 + 34202495888336p_y^5 + 123325558158108p_y^6$$
$$+ 351945664942868p_y^7 + 795449297019884p_y^8 + 1400682977460974p_y^9$$
$$+ 1840758497928775p_y^{10} + 1631376876416498p_y^{11} + 656223687892950p_y^{12}$$
$$- 345308229230269p_y^{13} - 499737648215188p_y^{14} - 160563926216347p_y^{15}$$
$$- 23264813191808p_y^{16} - 1752919253641p_y^{17} - 69812259099p_y^{18}$$
$$- 1360693027p_y^{19} - 9949430p_y^{20},$$
$$\lambda_{63} = 2(6 + 27p_y + 61p_y^2 + 4p_y^3)(10 + 10p_y + 69p_y^2 + 9p_y^3),$$
$$\lambda_{64} = (20 + 370p_y + 481p_y^2 + 11p_y^3)(38 + 304p_y + 517p_y^2 + 23p_y^3),$$
$$\lambda_{65} = (68 + 187p_y + 577p_y^2 + 50p_y^3)(80 + 136p_y + 601p_y^2 + 65p_y^3),$$
$$\lambda_{66} = 1 + p_y + (-1 + \Delta)p_y^2 + (-1 + \Delta + \phi)p_y^3.$$

Table 3: The bias of $\hat{r}$ of Noval estimator A after correction

p_y	n'	Relationship				n'	Relationship			
		Unrelated	Half-sibs	Full-sibs	Parent -offspring		Unrelated	Half-sibs	Full-sibs	Parent -offspring
	2	0.126026	0.221154	0.222184	0.294988	9	0.000000	0.000744	0.001887	0.001259
	3	0.000000	0.001043	0.002185	0.001857	10	0.000000	0.000713	0.001856	0.001196
	4	0.000000	0.000973	0.002115	0.001717	11	0.000000	0.000685	0.001828	0.001140
0.1	5	0.000000	0.000914	0.002056	0.001598	12	0.000000	0.000659	0.001802	0.001089
	6	0.000000	0.000863	0.002005	0.001496	13	0.000000	0.000636	0.001779	0.001042
	7	0.000000	0.000818	0.001961	0.001407	14	0.000000	0.000614	0.001757	0.000999
	8	0.000000	0.000779	0.001922	0.001328	15	0.000000	0.000595	0.001738	0.000960
	2	0.074074	0.136228	0.153817	0.192944	9	0.000000	0.004589	0.013657	0.003496
	3	0.000000	0.006392	0.015417	0.007062	10	0.000000	0.004449	0.013521	0.003219
	4	0.000000	0.005896	0.014934	0.006082	11	0.000000	0.004329	0.013404	0.002983
0.3	5	0.000000	0.005506	0.014553	0.005311	12	0.000000	0.004226	0.013303	0.002779
	6	0.000000	0.005200	0.014254	0.004705	13	0.000000	0.004136	0.013215	0.002601
	7	0.000000	0.004955	0.014015	0.004220	14	0.000000	0.004057	0.013138	0.002444
	8	0.000000	0.004755	0.013819	0.003825	15	0.000000	0.003986	0.013069	0.002305
	2	0.154900	0.206527	0.203012	0.242478	9	0.000000	0.014792	0.038100	0.003105
	3	0.000000	0.016599	0.039775	0.006538	10	0.000000	0.014654	0.037972	0.002843
	4	0.000000	0.016115	0.039326	0.005618	11	0.000000	0.014538	0.037864	0.002622
0.5	5	0.000000	0.015718	0.038958	0.004864	12	0.000000	0.014438	0.037772	0.002432
	6	0.000000	0.015406	0.038669	0.004270	13	0.000000	0.014352	0.037692	0.002268
	7	0.000000	0.015157	0.038439	0.003799	14	0.000000	0.014276	0.037622	0.002124
	8	0.000000	0.014957	0.038253	0.003418	15	0.000000	0.014209	0.037560	0.001997
	2	0.192033	0.255932	0.240904	0.281987	9	0.000000	0.025171	0.057596	0.002408
	3	0.000000	0.026588	0.058854	0.004996	10	0.000000	0.025060	0.057498	0.002206
	4	0.000000	0.026222	0.058529	0.004328	11	0.000000	0.024966	0.057415	0.002034
0.6	5	0.000000	0.025911	0.058253	0.003760	12	0.000000	0.024885	0.057343	0.001887
	6	0.000000	0.025663	0.058033	0.003307	13	0.000000	0.024815	0.057281	0.001759
	7	0.000000	0.025464	0.057857	0.002944	14	0.000000	0.024754	0.057227	0.001648
	8	0.000000	0.025303	0.057714	0.002650	15	0.000000	0.024700	0.057179	0.001549

Wang's (2002) estimator

Before correction, the sum of visible allele frequencies is extended to one, i.e. $\sum_{i\neq y} p'_i = 1$, where p'_i is the observed allele frequency which is equal to $\frac{p_i}{1-p_y}$. In addition, a'_i is defined by $\sum_{j\neq y} \left(p'_i\right)^j$ and b_i is defined by $\sum_{j\neq y} p_i^j$. If the frequencies of visible alleles are drawn from a triangular distribution, then $p'_i = \frac{2i}{n'(n'+1)}$ and $p_i = \frac{2i(1-p_y)}{n'(n'+1)}$, therefore the values of a'_i and b_i $(1 \leq i \leq 4)$ are given by

$$
\begin{aligned}
a'_1 &= 1, & b_1 &= 1 - p_y, \\
a'_2 &= \frac{2(1+2n')}{3n'(1+n')}, & b_2 &= \frac{2(1+2n')(1-p_y)^2}{3n'(1+n')}, \\
a'_3 &= \frac{2}{n'(1+n')}, & b_3 &= \frac{2(1-p_y)^3}{n'(1+n')}, \\
a'_4 &= \frac{8(1+2n')(-1+3n'+3n'^2)}{15n'^3(1+n')^3}, & b_4 &= \frac{8(1+2n')(-1+3n'+3n'^2)(1-p_y)^4}{15n'^3(1+n')^3}.
\end{aligned}
$$

According to Expression (13), namely the expression

$$\mathbf{P'} = [(f_1 - f_3)\Delta + (f_2 - f_3)\phi + f_3]^{-1}(\mathbf{E'} + \mathbf{M'\Delta}),$$

and combining Expressions (14a) with (14b), $\mathbf{P'}$ can be calculated.

Before correction, the estimator misuses $\hat{\mathbf{P}}'$ and a'_i to substitute $\hat{\mathbf{P}}$ and a_i in Expressions (5) and (12). With a mathematics software, the symbolic solution of Bias$(\hat{r})$ is given as follows (it is actually a function with Δ, ϕ, p_y and n' as the independent variables):

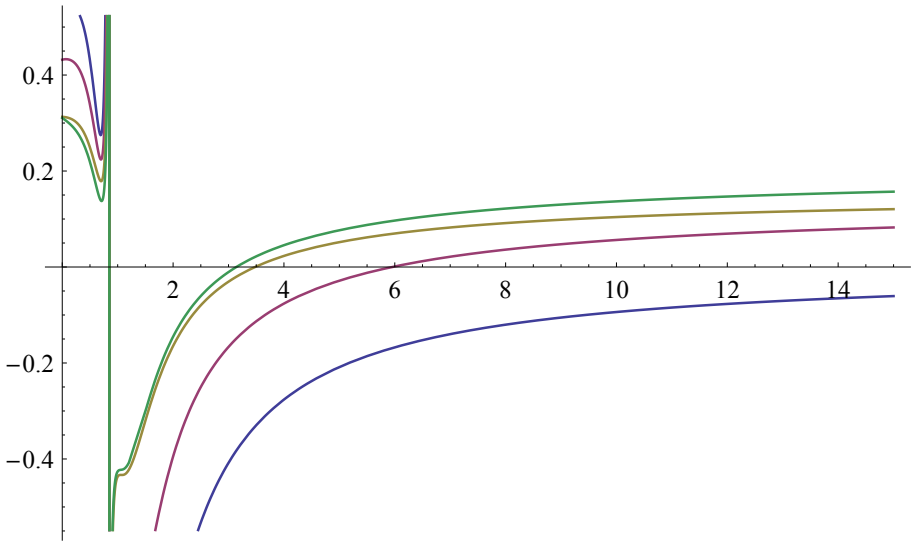
$$\text{Bias}(\hat{r}) = -\Delta - \phi/2 + (\lambda_1 + 2\lambda_2\Delta + \lambda_3\phi)(\lambda_4\lambda_5)^{-1},$$

where

$$
\begin{aligned}
\lambda_1 = {} & 16p_y[864(-1-7p_y+8p_y^2) + 48n'(-163-201p_y+364p_y^2) \\
& - 16n'^2(629-1404p_y+775p_y^2) - 8n'^3(-2803-3078p_y+5881p_y^2) \\
& + n'^4(44930-45918p_y+988p_y^2) + n'^5(-677-32373p_y+33050p_y^2) \\
& + 90n'^6(-444+323p_y+121p_y^2) + 45n'^7(-501+239p_y+262p_y^2) \\
& + 1350n'^8(-2-5p_y+7p_y^2)], \\
\lambda_2 = {} & -576(31+19p_y-84p_y^2+96p_y^3) - 96n'(621-31p_y-804p_y^2+1456p_y^3) \\
& + 16n'^2(-287+4745p_y-11232p_y^2+6200p_y^3) \\
& + 16n'^3(8363-2849p_y-12312p_y^2+23524p_y^3) \\
& - 4n'^4(-15251+74609p_y-91836p_y^2+1976p_y^3) \\
& - 2n'^5(51733+49025p_y-129492p_y^2+132200p_y^3) \\
& - 15n'^6(4451-16861p_y+15504p_y^2+5808p_y^3) \\
& - 45n'^7(-785-4793p_y+1912p_y^2+2096p_y^3) \\
& - 135n'^8(-379-539p_y-400p_y^2+560p_y^3) + 18225n'^9(1+p_y), \\
\lambda_3 = {} & -576(31+60p_y-221p_y^2+192p_y^3) - 96n'(621+224p_y-2515p_y^2+2912p_y^3)
\end{aligned}
$$

$$
\begin{aligned}
&+ 16n'^2(-287 + 12958p_y - 25645p_y^2 + 12400p_y^3) \\
&+ 16n'^3(8363 + 4106p_y - 42791p_y^2 + 47048p_y^3) \\
&+ n'^4(61004 - 506512p_y + 583324p_y^2 - 15808p_y^3) \\
&- 2n'^5(51733 + 94960p_y - 307627p_y^2 + 264400p_y^3) \\
&- 15n'^6(4451 - 23566p_y + 16401p_y^2 + 11616p_y^3) \\
&- 45n'^7(-785 - 3162p_y - 1815p_y^2 + 4192p_y^3) \\
&- 135n'^8(-379 - 2p_y - 1497p_y^2 + 1120p_y^3) - 2025n'^9(-9 - 14p_y + 5p_y^2), \\
\lambda_4 = {} & 2(-17856 - 59616n' - 4592n'^2 + 133808n'^3 + 61004n'^4 - 103466n'^5 \\
&- 66765n'^6 + 35325n'^7 + 51165n'^8 + 18225n'^9), \\
\lambda_5 = {} & 1 + p_y + (-1 + \Delta)p_y^2 + (-1 + \Delta + \phi)p_y^3.
\end{aligned}
$$

The graph of Bias($\hat{r}$) as a function of n' as the explanatory variable and with $p_y = 0.6$ is shown below, where the green (yellow, pink or blue) curve is the bias of $\hat{r}$ for parent-offspring (full-sibs, half-sibs or nonrelatives):



Corrected Wang's (2002) estimator

After correction, the approximate form $\mathbf{P}' \approx \mathbf{E}^* + \mathbf{M}^*\boldsymbol{\Delta}$ is used for estimation, and $\hat{\boldsymbol{\Delta}}$ is

$$\hat{\boldsymbol{\Delta}} = [(\mathbf{M}^*)^T(\mathbf{V}^*)^{-1}\mathbf{M}^*]^{-1}(\mathbf{M}^*)^T(\mathbf{V}^*)^{-1}(\hat{\mathbf{P}}' - \mathbf{E}^*).$$

Substituting $\hat{\mathbf{P}}'$ by $\mathbf{P}'$ (where $\mathbf{P}'$ can be obtained by Expression (13)), the bias of $\hat{r}$ can be solved, which is independent on n':

$$\text{Bias}(\hat{r}) = -\Delta - \phi/2 + \frac{\Delta(1 + p_y) + \phi(1 + p_y - p_y^2)/2}{1 + p_y + (-1 + \Delta)p_y^2 + (-1 + \Delta + \phi)p_y^3}.$$

Thomas's (2010) estimator

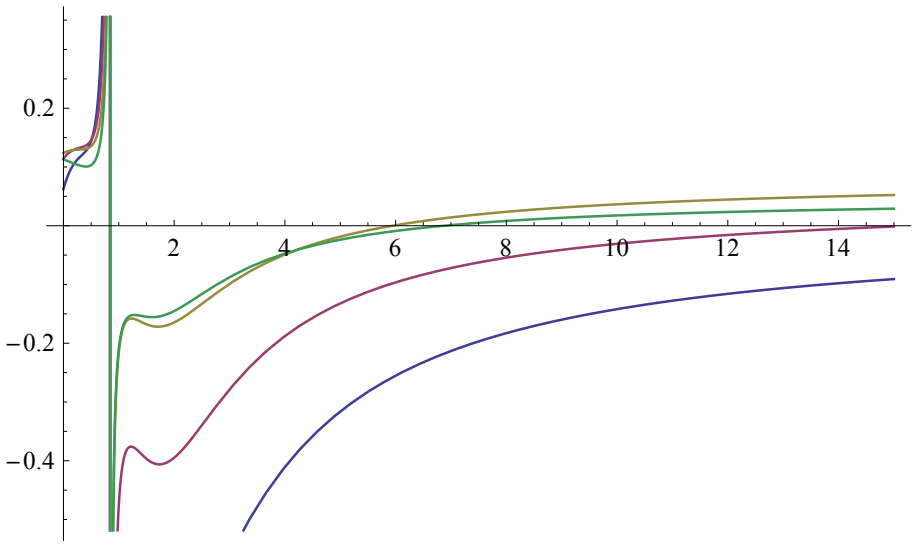
The bias of the TH estimator can be calculated using the same method as the WA estimator, and the expression of Bias($\hat{r}$) is as follows:

$$\text{Bias}(\hat{r}) = -\Delta - \phi/2 + (\lambda_1 + 2\lambda_2\Delta + \lambda_3\phi)(\lambda_4\lambda_5)^{-1},$$

where

$$\begin{aligned}
\lambda_1 &= 48p_y(-1+p_y)(-2+3n'+4p_y)(4-11n'^2+5n'^3+10n'^4),\\
\lambda_2 &= -48(3+7p_y-12p_y^2+8p_y^3)-8n'(37+p_y+36p_y^2)\\
&\quad +24n'^2(9+31p_y-66p_y^2+44p_y^3)+n'^3(494-538p_y+1512p_y^2-480p_y^3)\\
&\quad -15n'^4(9+17p_y-72p_y^2+64p_y^3)-90n'^5(1-7p_y+8p_y^2)+135n'^6(1+p_y),\\
\lambda_3 &= -48(3+10p_y-23p_y^2+16p_y^3)-8n'(37-54p_y+91p_y^2)\\
&\quad +24n'^2(9+55p_y-134p_y^2+88p_y^3)+n'^3(494-1788p_y+3242p_y^2-960p_y^3)\\
&\quad -15n'^4(9+44p_y-163p_y^2+128p_y^3)-90n'^5(1-12p_y+13p_y^2)+135n'^6(1+p_y^2),\\
\lambda_4 &= 2[1+p_y+(-1+\Delta)p_y^2+(-1+\Delta+\phi)p_y^3](2+3n'),\\
\lambda_5 &= (-72-40n'+168n'^2-5n'^3-60n'^4+45n'^5).
\end{aligned}$$

The graph of Bias($\hat{r}$) as a function with n' as the explanatory variable and with $p_y = 0.6$ is shown below, where the green (yellow, pink or blue) curve is the bias of $\hat{r}$ for parent-offspring (full-sibs, half-sibs or nonrelatives):



Corrected Thomas's (2010) estimator

The expression of Bias($\hat{r}$) of the corrected Thomas's (2010) estimator is identical to Wang's (2002) estimator.

Noval estimator B

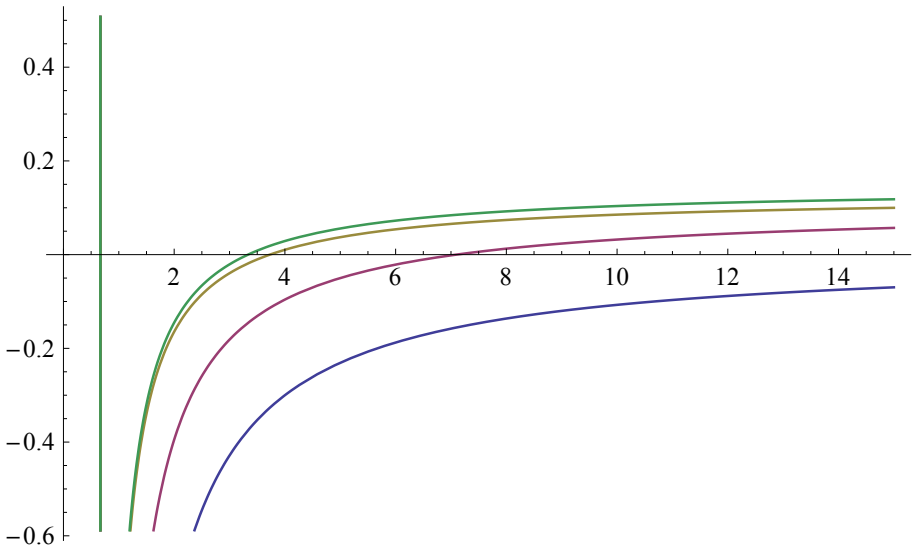
The bias of the TH estimator can be calculated using the same method as the WA estimator, and the expression of Bias($\hat{r}$) is simpler than the WA estimator:

$$\text{Bias}(\hat{r}) = -\Delta - \phi/2 + (\lambda_1 + 2\lambda_2\Delta + \lambda_3\phi)\lambda_4^{-1},$$

where

$$\lambda_1 = 16p_y(-1 + p_y^2),$$
$$\lambda_2 = [3n' - 2(1 - 2p_y)^2](1 + p_y),$$
$$\lambda_3 = 3n'(1 + p_y) + 2(-1 + p_y + 6p_y^2 - 8p_y^3),$$
$$\lambda_4 = 2(-2 + 3n')[1 + p_y + (-1 + \Delta)p_y^2 + (-1 + \Delta + \phi)p_y^3].$$

The graph of Bias($\hat{r}$) as a function with n' as the explanatory variable and with $p_y = 0.6$ is shown below, where the green (yellow, pink or blue) curve is the bias of $\hat{r}$ for parent-offspring (full-sibs, half-sibs or nonrelatives):



Corrected Noval estimator B

The expression of Bias($\hat{r}$) of the corrected Noval estimator B is identical to Wang's (2002) estimator.

A summary of expressions of various estimators

Lynch and Ritland's (1999) estimator

Homozygote:

$$\begin{bmatrix} \Pr(G_y = A_iA_i \mid G_x = A_iA_i) \\ \Pr(G_y = A_iA_x \mid G_x = A_iA_i) \end{bmatrix} = \begin{bmatrix} p_i^2 \\ 2p_ip_x \end{bmatrix} + \begin{bmatrix} 1-p_i^2 & p_i - p_i^2 \\ -2p_ip_x & p_x - 2p_ip_x \end{bmatrix} \begin{bmatrix} \Delta \\ \phi \end{bmatrix};$$

Heterozygote:

$$\begin{bmatrix} \Pr(G_y = A_iA_i \mid G_x = A_iA_j) \\ \Pr(G_y = A_jA_j \mid G_x = A_iA_j) \\ \Pr(G_y = A_iA_j \mid G_x = A_iA_j) \\ \Pr(G_y = A_iA_x \mid G_x = A_iA_j) \\ \Pr(G_y = A_jA_x \mid G_x = A_iA_j) \end{bmatrix} = \begin{bmatrix} p_i^2 \\ p_j^2 \\ 2p_ip_j \\ 2p_ip_x \\ 2p_jp_x \end{bmatrix} + \begin{bmatrix} -p_i^2 & \frac{1}{2}p_i - p_i^2 \\ -p_j^2 & \frac{1}{2}p_j - p_j^2 \\ 1-2p_ip_j & \frac{1}{2}p_i + \frac{1}{2}p_j - 2p_ip_j \\ -2p_ip_x & \frac{1}{2}p_x - 2p_ip_x \\ -2p_jp_x & \frac{1}{2}p_x - 2p_jp_x \end{bmatrix} \begin{bmatrix} \Delta \\ \phi \end{bmatrix}.$$

Corrected Lynch and Ritland's (1999) estimator

Homozygote:

$$\begin{bmatrix} \Pr(G'_y = A'_iA'_i \mid G'_x = A'_iA'_i) \\ \Pr(G'_y = A'_iA'_x \mid G'_x = A'_iA'_i) \end{bmatrix} \approx \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} + \begin{bmatrix} 1-\lambda_1 & \frac{p_i(p_i+p_y)(2-p_y)+2p_y(p_y+2p_i)}{(p_i+2p_y)(2-p_y)} - \lambda_1 \\ -\lambda_2 & \frac{p_ip_x(2-p_y)+2p_xp_y}{(p_i+2p_y)(2-p_y)} - \lambda_2 \end{bmatrix} \begin{bmatrix} \Delta \\ \phi \end{bmatrix},$$

where $\lambda_1 = \dfrac{p_i^2 + 2p_ip_y}{1-p_y^2}$ and $\lambda_2 = \dfrac{2p_ip_x}{1-p_y^2}$;

Heterozygote:

$$\begin{bmatrix} \Pr(G'_y = A'_iA'_i \mid G'_x = A'_iA'_j) \\ \Pr(G'_y = A'_jA'_j \mid G'_x = A'_iA'_j) \\ \Pr(G'_y = A'_iA'_j \mid G'_x = A'_iA'_j) \\ \Pr(G'_y = A'_iA'_x \mid G'_x = A'_iA'_j) \\ \Pr(G'_y = A'_jA'_x \mid G'_x = A'_iA'_j) \end{bmatrix} \approx \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \\ \lambda_5 \end{bmatrix} + \begin{bmatrix} -\lambda_1 & \frac{p_i+p_y}{2} - \lambda_1 \\ -\lambda_2 & \frac{p_j+p_y}{2} - \lambda_2 \\ 1-\lambda_3 & \frac{p_i+p_j}{2} - \lambda_3 \\ -\lambda_4 & \frac{p_x}{2} - \lambda_4 \\ -\lambda_5 & \frac{p_x}{2} - \lambda_5 \end{bmatrix} \begin{bmatrix} \Delta \\ \phi \end{bmatrix},$$

where $\lambda_1 = \dfrac{p_i^2 + 2p_ip_y}{1-p_y^2}$, $\lambda_2 = \dfrac{p_j^2 + 2p_jp_y}{1-p_y^2}$, $\lambda_3 = \dfrac{2p_ip_j}{1-p_y^2}$, $\lambda_4 = \dfrac{2p_ip_x}{1-p_y^2}$, $\lambda_5 = \dfrac{2p_jp_x}{1-p_y^2}$.

Noval estimator A

Homozygote

$$\begin{bmatrix} \mathrm{E}(S) \\ \mathrm{E}(S^2) \end{bmatrix} = \begin{bmatrix} 1 & \frac{3}{4} & \frac{1}{2} \\ 1 & \frac{9}{16} & \frac{1}{4} \end{bmatrix} \begin{bmatrix} p_i^2 \\ 2p_ip_x \\ 0 \end{bmatrix} + \begin{bmatrix} 1 & \frac{3}{4} & \frac{1}{2} \\ 1 & \frac{9}{16} & \frac{1}{4} \end{bmatrix} \begin{bmatrix} 1-p_i^2 & p_i - p_i^2 \\ -2p_ip_x & p_x - 2p_ip_x \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \Delta \\ \phi \end{bmatrix};$$

Heterozygote:

$$\begin{bmatrix} \mathrm{E}(S) \\ \mathrm{E}(S^2) \end{bmatrix} = \begin{bmatrix} 1 & \frac{3}{4} & \frac{1}{2} \\ 1 & \frac{9}{16} & \frac{1}{4} \end{bmatrix} \begin{bmatrix} 2p_ip_j \\ p_i^2 + p_j^2 \\ 2p_x(p_i+p_j) \end{bmatrix} + \begin{bmatrix} 1 & \frac{3}{4} & \frac{1}{2} \\ 1 & \frac{9}{16} & \frac{1}{4} \end{bmatrix} \begin{bmatrix} 1-2p_ip_j & \frac{1}{2}(p_i+p_j) - 2p_ip_j \\ -p_i^2 - p_j^2 & \frac{1}{2}(p_i+p_j) - p_i^2 - p_j^2 \\ -2p_x(p_i+p_j) & p_x - 2p_x(p_i+p_j) \end{bmatrix} \begin{bmatrix} \Delta \\ \phi \end{bmatrix}.$$

Corrected Noval estimator A

Homozygote:

$$\begin{bmatrix} \mathrm{E}(S) \\ \mathrm{E}(S^2) \end{bmatrix} \approx \begin{bmatrix} 1 & \frac{3}{4} & \frac{1}{2} \\ 1 & \frac{9}{16} & \frac{1}{4} \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 & \frac{3}{4} & \frac{1}{2} \\ 1 & \frac{9}{16} & \frac{1}{4} \end{bmatrix} \begin{bmatrix} 1-\lambda_1 & \frac{p_i(p_i+p_y)(2-p_y)+2p_y(p_y+2p_i)}{(p_i+2p_y)(2-p_y)} - \lambda_1 \\ -\lambda_2 & \frac{p_i p_x(2-p_y)+2p_x p_y}{(p_i+2p_y)(2-p_y)} - \lambda_2 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \Delta \\ \phi \end{bmatrix},$$

where $\lambda_1 = \dfrac{p_i^2 + 2p_i p_y}{1-p_y^2}$ and $\lambda_2 = \dfrac{2p_i p_x}{1-p_y^2}$.

Heterozygote:

$$\begin{bmatrix} \mathrm{E}(S) \\ \mathrm{E}(S^2) \end{bmatrix} \approx \begin{bmatrix} 1 & \frac{3}{4} & \frac{1}{2} \\ 1 & \frac{9}{16} & \frac{1}{4} \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix} + \begin{bmatrix} 1 & \frac{3}{4} & \frac{1}{2} \\ 1 & \frac{9}{16} & \frac{1}{4} \end{bmatrix} \begin{bmatrix} 1-\lambda_1 & \frac{p_i+p_j}{2} - \lambda_1 \\ -\lambda_2 & \frac{p_i+p_j}{2} + p_y - \lambda_2 \\ -\lambda_3 & p_x - \lambda_3 \end{bmatrix} \begin{bmatrix} \Delta \\ \phi \end{bmatrix},$$

where $\lambda_1 = \dfrac{2p_i p_j}{1-p_y^2}$, $\lambda_2 = \dfrac{p_j^2 + 2p_j p_y + p_i^2 + 2p_i p_y}{1-p_y^2}$, $\lambda_3 = \dfrac{2p_x(p_i+p_j)}{1-p_y^2}$.

Wang's (2002) estimator

$$\begin{bmatrix} \Pr(S=1) \\ \Pr(S=\frac{3}{4}) \\ \Pr(S=\frac{1}{2}) \end{bmatrix} = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix} + \begin{bmatrix} 1-\lambda_1 & a_2 - \lambda_1 \\ -\lambda_2 & 2a_2 - 2a_3 - \lambda_2 \\ -\lambda_3 & 1 - 3a_2 + 2a_3 - \lambda_3 \end{bmatrix} \begin{bmatrix} \Delta \\ \phi \end{bmatrix},$$

where $\lambda_1 = 2a_2^2 - a_4$, $\lambda_2 = 4a_3 - 4a_4$ and $\lambda_3 = 4a_2 - 4a_2^2 - 8a_3 + 8a_4$.

Corrected Wang's (2002) estimator

$$\begin{bmatrix} \Pr(S'=1) \\ \Pr(S'=\frac{3}{4}) \\ \Pr(S'=\frac{1}{2}) \end{bmatrix} \approx \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix} + \begin{bmatrix} f_1^{-1}(b_1^2 + 2p_y b_1) - \lambda_1 & f_2^{-1}(b_1 b_2 + p_y^2 b_1 + 3p_y b_2) - \lambda_1 \\ -\lambda_2 & f_2^{-1}(2b_1 b_2 - 2b_3 + 2p_y b_1^2 - 2p_y b_2) - \lambda_2 \\ -\lambda_3 & f_2^{-1}(b_1^3 - 3b_1 b_2 + 2b_3) - \lambda_3 \end{bmatrix} \begin{bmatrix} \Delta \\ \phi \end{bmatrix},$$

where

$$\begin{aligned} \lambda_1 &= f_3^{-1}(2b_2^2 - b_4 + 4p_y^2 b_2 + 4p_y b_3), \\ \lambda_2 &= f_3^{-1}(4b_1 b_3 - 4b_4 + 8p_y b_1 b_2 - 8p_y b_3), \\ \lambda_3 &= f_3^{-1}(4b_1^2 b_2 - 8b_1 b_3 - 4b_2^2 + 8b_4). \end{aligned}$$

Thomas's (2010) estimator

$$\begin{bmatrix} \Pr(S=1) \\ \Pr\left(S=\frac{3}{4} \text{ or } \frac{1}{2}\right) \end{bmatrix} = \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} + \begin{bmatrix} 1-\lambda_1 & a_2 - \lambda_1 \\ -\lambda_2 & 1 - a_2 - \lambda_2 \end{bmatrix} \begin{bmatrix} \Delta \\ \phi \end{bmatrix},$$

where $\lambda_1 = 2a_2^2 - a_4$ and $\lambda_2 = 4a_2 - 4a_2^2 - 4a_3 + 4a_4$.

Corrected Thomas's (2010) estimator

$$\begin{bmatrix} \Pr(S'=1) \\ \Pr\left(S'=\frac{3}{4} \text{ or } \frac{1}{2}\right) \end{bmatrix} \approx \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} + \begin{bmatrix} f_1^{-1}(b_1^2+2p_yb_1)-\lambda_1 & f_2^{-1}(b_1b_2+p_y^2b_1+3p_yb_2)-\lambda_1 \\ -\lambda_2 & f_2^{-1}(2p_yb_1^2-2p_yb_2-b_1b_2+b_1^3)-\lambda_2 \end{bmatrix} \begin{bmatrix} \Delta \\ \phi \end{bmatrix},$$

where $\lambda_1 = f_3^{-1}(2b_2^2 - b_4 + 4p_y^2b_2 + 4p_yb_3)$ and $\lambda_2 = f_3^{-1}(8p_yb_1b_2 - 8p_yb_3 + 4b_1^2b_2 - 4b_1b_3 - 4b_2^2 + 4b_4)$.

Noval estimator B

$$\begin{bmatrix} \mathrm{E}(S) \\ \mathrm{E}(S^2) \end{bmatrix} = \begin{bmatrix} 1 & \frac{3}{4} & \frac{1}{2} \\ 1 & \frac{9}{16} & \frac{1}{4} \end{bmatrix} \left(\begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix} + \begin{bmatrix} 1-\lambda_1 & a_2-\lambda_1 \\ -\lambda_2 & 2a_2-a_3-\lambda_2 \\ -\lambda_3 & 1-3a_2+2a_3-\lambda_3 \end{bmatrix} \begin{bmatrix} \Delta \\ \phi \end{bmatrix} \right),$$

where $\lambda_1 = 2a_2^2 - a_4$, $\lambda_2 = 4a_3 - 4a_4$ and $\lambda_3 = 4a_2 - 4a_2^2 - 8a_3 + 8a_4$.

Corrected Noval estimator B

$$\begin{bmatrix} \mathrm{E}(S') \\ \mathrm{E}(S'^2) \end{bmatrix} \approx \begin{bmatrix} 1 & \frac{3}{4} & \frac{1}{2} \\ 1 & \frac{9}{16} & \frac{1}{4} \end{bmatrix} \left(\begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix} + \begin{bmatrix} f_1^{-1}(b_1^2+2p_yb_1)-\lambda_1 & \mu_1-\lambda_1 \\ -\lambda_2 & \mu_2-\lambda_2 \\ -\lambda_3 & \mu_3-\lambda_3 \end{bmatrix} \begin{bmatrix} \Delta \\ \phi \end{bmatrix} \right),$$

where

$$\begin{aligned}
\lambda_1 &= f_3^{-1}(2b_2^2 - b_4 + 4p_y^2b_2 + 4p_yb_3), \\
\lambda_2 &= f_3^{-1}(4b_1b_3 - 4b_4 + 8p_yb_1b_2 - 8p_yb_3), \\
\lambda_3 &= f_3^{-1}(4b_1^2b_2 - 8b_1b_3 - 4b_2^2 + 8b_4), \\
\mu_1 &= f_2^{-1}(b_1b_2 + p_y^2b_1 + 3p_yb_2), \\
\mu_2 &= f_2^{-1}(2b_1b_2 - 2b_3 + 2p_yb_1^2 - 2p_yb_2), \\
\mu_3 &= f_2^{-1}(b_1^3 - 3b_1b_2 + 2b_3).
\end{aligned}$$

Literature Cited

Anderson, A. D., and B. S. Weir, 2007 A maximum-likelihood method for the estimation of pairwise relatedness in structured populations. Genetics 176: 421–440.

Lynch, M., and K. Ritland, 1999 Estimation of pairwise relatedness with molecular markers. Genetics 152: 1753–1766.

Thomas, S. C., 2010 A simplified estimator of two and four gene relationship coefficients. Molecular Ecology Resources 10: 986–994.

Wang, J. L., 2002 An estimator for pairwise relatedness using molecular markers. Genetics 160: 1203–1215.