An Estimator for Pairwise Relatedness Using Molecular Markers

Jinliang Wang

Institute of Zoology, Zoological Society of London, London NW1 4RY, United Kingdom

Manuscript received June 22, 2001
Accepted for publication December 12, 2001

ABSTRACT

I propose a new estimator for jointly estimating two-gene and four-gene coefficients of relatedness between individuals from an outbreeding population with data on codominant genetic markers and compare it, by Monte Carlo simulations, to previous ones in precision and accuracy for different distributions of population allele frequencies, numbers of alleles per locus, actual relationships, sample sizes, and proportions of relatives included in samples. In contrast to several previous estimators, the new estimator is well behaved and applies to any number of alleles per locus and any allele frequency distribution. The estimates for two- and four-gene coefficients of relatedness from the new estimator are unbiased irrespective of the sample size and have sampling variances decreasing consistently with an increasing number of alleles per locus to the minimum asymptotic values determined by the variation in identity-by-descent among loci per se, regardless of the actual relationship. The new estimator is also robust for small sample sizes and for unknown relatives being included in samples for estimating allele frequencies. Compared to previous estimators, the new one is generally advantageous, especially for highly polymorphic loci and/or small sample sizes.

GENETIC relationships among individuals within and between populations play a pivotal role in quantitative genetics, conservation genetics, and many areas of evolution and ecology (Ritland 1996a,b, 2000; Lynch and Ritland 1999). Such relationships can be easily estimated from pedigree information. For natural populations, however, detailed knowledge of pedigree is usually unavailable. Genetic markers can be used, instead, to infer the relationships among individuals without pedigree information.

Marker-based methods for inferring genetic relationships among individuals can be categorized into two groups. The first group uses a likelihood approach to determine the likelihood of a pair of individuals falling into a particular type of relationship (such as full-sibs, parent-offspring, etc.) given the marker data (e.g., Thompson 1975; Marshall et al. 1998; Mousseau et al. 1998; Goodnight and Queller 1999; Thomas and Hill 2000). The second group uses moment estimators to estimate the relatedness between a pair of individuals, which is a continuous quantity defined in terms of probabilities of identity-by-descent (e.g., Lynch 1988; Queller and Goodnight 1989; Li et al. 1993; Ritland 1996b; Lynch and Ritland 1999).

Relatedness estimators are especially useful for populations with complex (unknown) pedigrees or with little prior information on population structure. Several moment estimators for two-gene and four-gene relatedness have been proposed (e.g., Lynch 1988; Queller and Goodnight 1989; Li et al. 1993; Ritland 1996b; Lynch and Ritland 1999), and their performances have been compared in recent studies (Lynch and Ritland 1999; Van de Casteele et al. 2001). In this article I develop a new estimator for simultaneously estimating two-gene and four-gene coefficients of relatedness between a pair of individuals and evaluate its performance against previous estimators for different actual relationships and allele frequency distributions in the population.

Lynch and Ritland (1999) assessed the previous moment estimators assuming that the allele frequencies of the population were known. In practice, however, the allele frequencies of the population have to be estimated from samples and are therefore subject to sampling errors. For highly polymorphic markers (e.g., microsatellites) and realistic sample sizes, the sampling errors for allele frequency cannot be ignored in comparing the performances of the estimators. Little is known about the sensitivities of these moment estimators for relatedness to the sampling errors in allele frequency. In this article, I compare the accuracy and precision of different estimators of relatedness when allele frequencies are estimated from samples and when unknown relatives are included in the samples. I show that the new moment estimator can easily take the effect of small sample size into account. Compared to the other estimators, the new one is well behaved irrespective of the number of alleles per locus, the allele frequency distribution, sample size, and the inclusion of relatives in samples for estimating allele frequencies.
THE NEW ESTIMATOR OF TWO-GENE AND FOUR-GENE COEFFICIENTS OF RELATEDNESS

The coefficients of relatedness: Following previous studies, I consider an outbred diploid population throughout. A pair of individuals (say, \(x\) and \(y\)) in such a population can be genetically correlated in two ways: a single gene at a locus in \(x\) is identical by descent with one in \(y\), or both genes in \(x\) are identical by descent with those in \(y\). If the probabilities of occurrences of the first and second events are denoted as \(\phi\) and \(\Delta\), respectively, the relatedness, \(r\), between individuals \(x\) and \(y\) can be expressed as

\[
r = \phi/2 + \Delta
\]  

(1)

(Lynch and Ritland 1999). In an outbreeding population, \(\phi\), \(\Delta\), and \(r\) are 1, 0, and 0.5, respectively, for parents and offspring; 0.5, 0.25, and 0.5, respectively, for full-sibs; and 0.25, 0, and 0.125, respectively, for half-sibs.

Although \(\phi\), \(\Delta\), and \(r\) can be estimated for any pair of individuals using genotypic data, I concentrate on the estimation of \(\Delta\) and \(r\) only because these two are used to partition the genetic variance of a quantitative trait (\(\sigma^2_{\text{q}}\)) into the additive (\(\sigma^2_A\)) and dominance (\(\sigma^2_D\)) components, \(\sigma^2_{\text{q}} = \sigma^2_A + \Delta \sigma^2_D\) (Falconer and Mackay 1996; Lynch and Walsh 1998). Furthermore, the estimates of \(\phi\) have a much higher sampling variance than those of \(\Delta\) and \(r\) (Lynch and Ritland 1999). Unless data on a very large number of informative loci are available, \(\phi\) cannot be estimated with reasonable confidence.

The new estimator: To derive the new estimator, I first determine the joint probabilities of two genotypes given the parameters \(\phi\) and \(\Delta\) (to be estimated) and the allele frequencies of the population (assumed to be known for this section). For a marker locus with \(n\) codominant alleles indexed by \(i, j, k, l = 1, 2, \ldots, n\), there would be \(n(n + 1)/2\) possible genotypes and \(n(n + 1)(n(n + 1) + 1)/8\) possible combinations of genotypes for a pair of individuals. The probabilities of these combinations of genotypes can be derived in terms of \(\phi\), \(\Delta\), and allele frequencies, as are listed in Table 1 where the population frequency of allele \(A_i\) is denoted as \(p_i\). According to their similarity index \(S\), defined as the arithmetic average fraction of genes at a locus in a reference individual (here either \(x\) or \(y\)) for which there is another gene in the other individual (either \(y\) or \(x\)) that is identical in state (Lynch 1988; Li et al. 1993), these genotype combinations can be grouped into four categories. Category 1 contains combinations \(A_iA_iA_i\) and \(A_iA_iA_j\) with \(S = 1\); category 2 contains combinations \(A_iA_iA_j\) with \(S = 1/2\); category 3 contains combinations \(A_iA_jA_j\) with \(S = \sqrt{2}/2\); category 4 contains all the rest of the combinations with \(S = 0\) (\(i \neq j \neq k = 1, 2, \ldots, n\)). Summing up the probabilities of all the genotype combinations within a category gives the probability of occurrence of that category. The probability of occurrence of category 1, for example, is

\[
P_1 = \sum_{i=1}^{n} P(ii, ii) + \sum_{i=1}^{n} \sum_{j=1}^{n} P(ij, ij) = b + e\phi + (1 - b)\Delta,
\]  

(2)

where \(b = 2a_i^2 - a_i\), \(e = a_2 - 2a_i^2 + a_i\) in which the sum of powers of gene frequency \(a_i = \sum_{m=1}^{n} p_m^m\) for \(m = 2, 3,\) and 4. The probabilities for categories 2–4 are

\[
P_2 = \sum_{i=1}^{n} \sum_{j=1}^{n} P(ii, ij) = d(1 - \Delta) + e\phi,
\]  

(3)

where \(d = 4(a_3 - a_i)\) and \(e = 2(a_2 - 3a_3 + 2a_i)\); \(P_3 = \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} P(ik, ik) = f(1 - \Delta) + g\phi\),

\[
P_4 = \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} \sum_{l=1}^{n} P(ij, kl) = (1 - 4a_2 + 2a_3 + 4a_4 - 3a_5)(1 - \phi - \Delta),
\]  

(5)

respectively. Obviously, we have \(P_1 + P_2 + P_3 + P_4 = 1\), because a pair of genotypes must fall into one of the four categories. The observed values (denoted as \(\hat{P}\)) of \(P_i\) (\(i = 1, 2, 3,\) or 4) for a pair of genotypes constitute the data; one of the four values is one and the rest are zero. For example, if we observe \(A_iA_jA_i\), then \(\hat{P}_1 = \hat{P}_2 = \hat{P}_4 = 0\), and \(\hat{P}_3 = 1\).

For biallelic loci (\(n = 2\)), category 3 is not possible; only three equations remain to be solved for parameters \(\phi\) and \(\Delta\). Solving any two of the three dependent equations yields

\[
\hat{\phi} = (4 - 4\hat{P}_1 - 3\hat{P}_2)(1 - a_2) - 4(1 - \hat{P}_1 - \hat{P}_2),
\]  

(6)

\[
\hat{\Delta} = 1 - (4 - 4\hat{P}_1 - 3\hat{P}_2)(1 - a_2) - 2(1 - \hat{P}_1 - \hat{P}_2),
\]  

(7)

Inserting (6) and (7) into (1) yields

\[
\hat{r} = \frac{4\hat{P}_1 + 3\hat{P}_2 - 2(1 + a_2)}{2(1 - a_2)}.
\]  

(8)

Equation 7 for \(\Delta\) is essentially the same as that of Lynch and Ritland (1999), their Equation 4b) for biallelic loci. Equations 6 and 8 for \(\hat{\phi}\) and \(\hat{r}\) are, however, different from those of Lynch and Ritland (1999). Equation 8 for \(r\) is essentially the same as Equation 9 of Li et al. (1993) for the case of biallelic loci, which was developed with a modification of the similarity index for category 3 of Lynch (1988). The new estimator is advantageous in that it can estimate \(\phi\) and \(\Delta\) as well.

With a multiallelic locus, there are more independent equations than the parameters to be estimated. A weighted least-squares analysis could be used to obtain
TABLE 1
Probabilities of a pair of genotypes at a diploid marker locus with n codominant alleles
\[(i \neq j \neq k \neq l = 1, 2, \ldots, n)\]

<table>
<thead>
<tr>
<th>Combinations of genotypes</th>
<th>Similarity index</th>
<th>Probability</th>
<th>Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>A.A, (\cdot)A.A</td>
<td>1/2</td>
<td>(P(i, j))</td>
<td>(\hat{\phi}_i^4)</td>
</tr>
<tr>
<td>A.A, (\cdot)A.A</td>
<td>1/2</td>
<td>(P(i, j))</td>
<td>(\frac{1}{2}p_i p_j (p_i + p_j - 4p_i p_j))</td>
</tr>
<tr>
<td>A.A, (\cdot)A.A</td>
<td>3/4</td>
<td>(P(i, j))</td>
<td>(-p_i^2 p_j^2)</td>
</tr>
<tr>
<td>A.A, (\cdot)A.A</td>
<td>1/2</td>
<td>(P(i, j))</td>
<td>(-p_i^2 p_j^2)</td>
</tr>
<tr>
<td>A.A, (\cdot)A.A</td>
<td>0</td>
<td>(P(i, j))</td>
<td>(-p_i^2 p_j^2)</td>
</tr>
</tbody>
</table>

The estimates of \(\hat{\phi}\) and \(\Delta\) (and thus \(r\)) with the minimum sampling variance. Unfortunately, however, the optimum weights must be obtained from the variance-covariance matrix, which depends on the unknown parameters \(\phi\) and \(\Delta\). To obtain approximate weights, I follow Ritland (1996b) and Lynch and Ritland (1999) in assuming that \(\phi = \Delta = 0\) in the absence of prior information on the relationship. With this assumption I can obtain the weighted least-squares estimators of \(\phi\) and \(\Delta\) from (2–5) using an equation-solving software,

\[
\hat{\phi} = \left\{df[(e + g)(1 - b) + c(d + f)](\hat{P}_1 - 1) + d(1 - b)[g(1 - b - d) + f(e + c)]\hat{P}_3 + f(1 - b)[e(1 - b - f) + d(c + g)]\hat{P}_3/V, \right.
\]

(9)

\[
\hat{\Delta} = \left\{ef(e + g)(\hat{P}_1 + 1 - 2b) + [(1 - b)(fr^2 + dg^2) - (ef - dg)^2](\hat{P}_1 - b) + c(dg - ef)(d\hat{P}_3 - f\hat{P}_3) - c^2df \times (\hat{P}_3 + \hat{P}_3 - d - f) - c(1 - b)(dg\hat{P}_3 + ef\hat{P}_3)/V, \right. \]

(10)

where

\[
V = (1 - b)^2(e^2f + dg^2) - (1 - b)(ef - dg)^2 + 2ef(1 - b)(g + e) + c^2df(d + f),
\]

and \(b\)–\(g\) are defined in (2–4). Inserting (9) and (10) into (1) yields the estimate for \(r\).

When allele frequencies for the population are unknown but are estimated from a sample of \(N\) genes taken at random from the population, the sampling error should be taken into account. Denoting the sample allele frequency for allele \(A_i\) as \(\tilde{p}_i\), the observed sums of powers of allele frequencies are \(\tilde{a}_m = \sum_{i} \tilde{p}_i^m\) for \(m = 2, 3,\) and 4. Obviously, \(\tilde{a}_m\) is a biased estimate of \(a_m\) for the population. The expected sums of powers of allele frequencies (denoted as \(\bar{a}_m\) for \(m = 2, 3,\) and 4) can be obtained using the exact formulas for the first four moments of allele frequency in Crow and Kimura (1970, p. 335, correcting for the typographical error in the fourth moment). These are

\[
\bar{a}_2 = \frac{N\tilde{a}_2 - 1}{N - 1},
\]

(12)

\[
\bar{a}_3 = \frac{N^2\tilde{a}_3 - 3(N - 1)\tilde{a}_2 - 1}{(N - 1)(N - 2)},
\]

(13)

\[
\bar{a}_4 = \frac{N^3\tilde{a}_4 - 6(N - 1)(N - 2)\tilde{a}_3 - 7(N - 1)\tilde{a}_2 - 1}{N^3 - 6N^2 + 11N - 6},
\]

(14)

As is shown below, essentially unbiased estimates of relatedness can be obtained with the new estimator using (9–14), irrespective of the sample size \(N\).

In practice, relatedness is estimated using several loci. Although relatedness estimates from unlinked loci in linkage equilibrium are independent, they could be dramatically different in sampling variance and ideally should not be simply averaged to give the overall estimate. The optimum weights among loci are difficult to obtain for any moment estimator, however, because they are functions of the parameters (\(\phi\) and \(\Delta\)) being estimated. As a result, the difference in the amount of information among loci was either ignored by using equal weights (e.g., Queller and Goodnight 1989; Li et al. 1993) or taken partially into account by approximate weights derived assuming null relatedness, such as heterozygosity (Loiselle et al. 1995), number of alleles at a locus (Ritland 1996a,b), or the inverse of sampling variance for nonrelatives (Lynch and Ritland 1999). None of the weighting methods is appropriate in all circumstances (allele frequency distribution, actual relationship), as is intuitively obvious. Methods with equal weighting (e.g., Queller and Goodnight 1989; Li et al. 1993) tend to give higher sampling variances for nonrelatives and lower sampling variances for relatives, compared to the weighting methods based on the sampling variance of nonrelatives (Ritland 1996a,b; Lynch and Ritland 1999). I tried all the above weighting meth-
ods in the new estimator and found the following method is more satisfactory in the vast majority of cases considered in simulations.

The average similarity value for unrelated individuals is \( u = 2a_2 - a_3 \) (Li et al. 1993), which is due completely to chance (identity-in-state) and determined solely by allele frequencies at a locus. The amount of information for a locus in estimating relatedness decreases with an increasing value of \( u \). Therefore I use a weight for locus \( l \) as \( w(l) = 1/(Uu) \), where \( U \) is the sum of \( 1/u \) over loci. The averages of \( \hat{r}(i = 1, 2, \text{ and } 3), a_u \) (\( m = 2, 3, \text{ and } 4 \)), and \( a_3^2 \) weighted by \( w(l) \) over loci are used in the new estimator. If allele frequencies are estimated from samples, \( \bar{a}_u \) and \( \bar{a}_i^2 \) are calculated using (12–14) for each locus and then weighted over loci for use in the estimation.

**COMPARISON AMONG ESTIMATORS**

Several moment estimators for pairwise relatedness have been proposed, which are summarized and compared by Lynch and Ritland (1999). For convenience, these estimators are briefly described below.

Based on the similarity index, or the sharing of genes between individuals, Lynch (1988) developed an estimator for \( r \), which was later improved by Li et al. (1993). This estimator (denoted as LL hereafter) is

\[ \hat{r} = \frac{S_{xy} - u}{1 - u}, \]

where \( S_{xy} \) is the similarity index between individuals \( x \) and \( y \) (as defined in Table 1), and \( u = 2a_2 - a_3 \) is the average similarity among unrelated individuals due to chance.

Queller and Goodnight’s (1989) estimator (denoted as QG) for the average relatedness between groups of individuals can also be made to estimate the pairwise \( r \) as shown by Lynch and Ritland (1999). Let individuals \( x \) and \( y \) have genes \( a, b, c, \text{ and } d \), respectively, at a locus and indicator variables \( S_{ij} = 1 \) if gene \( i \) is identical in state to gene \( j \) (\( i, j = a, b, c, \text{ or } d \)) and \( S_{ij} = 0 \) otherwise. The estimate of \( r_{xy} \) using individual \( x \) as the reference is

\[ \hat{r}_{xy} = \frac{0.5(S_{sa} + S_{sb} + S_{sc} + S_{sd}) - p_i - p_j}{1 + S_{sa} - p_i - p_j}, \]

where \( p_i \) is the frequency of allele \( i \). Similarly, \( \hat{r}_{yx} \), using individual \( y \) as the reference can be obtained from (16), replacing \( S_{sa} \) by \( S_{sa} \), and \( p_i \) and \( p_j \), by \( p_j \) and \( p_i \), respectively. The arithmetic average of the reciprocal estimates \( \hat{r}_{xy} \) and \( \hat{r}_{yx} \) yields the estimate of \( r \) between individuals \( x \) and \( y \).

For estimating \( r \) and \( \Delta \) simultaneously, Ritland (1996b) developed a “correlation” estimator (denoted as \( R \)) on the basis of the joint distributions of both genotypes conditional on allele frequencies \( p_i \) and parameters \( r \) and \( \Delta \). The estimator is not shown here because it is quite complex algebraically and involves the solution for weights of two sets of \( n(n + 5)/2 \) linear equations for a single locus with \( n \) alleles.

On the basis of probability of a genotype conditional on the reference genotype, allele frequencies \( p_i \) and parameters \( r \) and \( \Delta \), Lynch and Ritland (1999) proposed a much simpler “regression” estimator (denoted as LR). For a single locus \( l \), the estimates of relatedness between individuals \( x \) and \( y \) using \( x \) as the reference are

\[ \hat{r}_{x}(l) = \frac{p(x, y) + p(x, y) - 4p_x p_y}{(1 + S_{x}) (p_x - p_y)}, \]

\[ \hat{\Delta}_{x}(l) = \frac{2p_x p_y - p(x, y) - p(x, y) + S_{x} + S_{y}}{(1 + S_{x}) (1 - p_x - p_y) + 2p_x p_y}, \]

where indicator variables \( S_{ij} \) (\( i, j = a, b, c, \text{ or } d \)) are defined in (16). For a number of \( L \) loci, the composite estimates are obtained by weighting single-locus estimates across loci,

\[ \hat{r}_{x} = \frac{1}{W_{x}} \sum_{l=1}^{L} w_{x}(l) \hat{r}_{x}(l), \]

\[ \hat{\Delta}_{x} = \frac{1}{W_{\Delta}} \sum_{l=1}^{L} w_{\Delta}(l) \hat{\Delta}_{x}(l), \]

where the weights for locus \( l \),

\[ w_{x}(l) = \frac{(1 + S_{x}) (p_x - p_y) - 4p_x p_y}{2p_x p_y}, \]

\[ w_{\Delta}(l) = \frac{(1 + S_{x}) (1 - p_x - p_y) + 2p_x p_y}{2p_x p_y} \]

are derived by assuming that \( x \) and \( y \) are unrelated (Lynch and Ritland 1999). In (19) and (20), \( W_{x} \) and \( W_{\Delta} \) are the sums of weights over the \( L \) loci. Using individual \( y \) as the reference, the estimates \( \hat{r}_{y} \), and \( \hat{\Delta}_{y} \) can be obtained similarly. The arithmetic average of the reciprocal estimates \( \hat{r}_{x} \) and \( \hat{r}_{y} \) (\( \hat{\Delta}_{x} \) and \( \hat{\Delta}_{y} \)) yields the estimate of \( r \) (\( \Delta \)) between individuals \( x \) and \( y \).

Compared to the other estimators, an attractive property of the new estimator (denoted as \( W \)) and the QG and LL estimators is that they do not give estimates of \( r > 1 \). Like the other estimators, however, they do yield negative estimates because of sampling error. The new estimator does not generate estimates of \( \Delta > 1 \) except for biallelic loci (in which case the maximum estimate for \( \Delta \) is 2.125), in contrast to the LR and R estimators.

In the following, I compare the precision and accuracy of these estimators for different actual relationships and allele frequency distributions. First I assume that the allele frequencies of the population are known, and then I assume that allele frequencies are estimated from samples taken binomially from the population without and with relatives-clusters included in them.

**Allele frequencies known:** When the allele frequencies are assumed to be known without error, all the
pairwise relatedness estimators are essentially unbiased, regardless of the allele frequencies, number of loci, and the actual relationship (Lynch and Ritland 1999; data not shown). I, therefore, concentrate on evaluating the estimators in their precision, which is indicated by the sampling variance of the estimates given the actual relationship and allele frequencies.

For all of the estimators except for the LR estimator, the average single-locus sampling variance, calculated as the variance of multilocus estimates divided by the number of loci, is independent of the number of loci. Given the allele frequencies and the actual relationship, I can obtain the probability of each pair of genotypes at a single locus and thus the exact sampling variance of relatedness estimates from each estimator.

For the LR estimator, however, the average single-locus sampling variance changes with the number of loci used in the estimation, because the weights for different loci depend in part on the reference genotypes at the corresponding loci (Lynch and Ritland 1999). I used Monte Carlo simulations to generate paired multilocus genotypes conditional on the allele frequencies and the actual relationship. The sampling variance was calculated from the relatedness estimates for at least 100,000 pairs of individuals. For simplicity hereafter, the average single-locus sampling variance is called single-locus sampling variance, bearing in mind that it is estimated from multiple loci and varies with the number of loci for the LR estimator.

A peculiar property of the LR estimator is that the single-locus sampling variances of both $r$ and $\Delta$ for related individuals may increase with an increasing number of loci. Extensive simulations show that the phenomenon happens more often with a closer actual relationship and a larger number of alleles per locus (data not shown). Allele frequency distributions also affect the direction (increasing or decreasing) of the changes in single-locus variance with an increasing number of loci, as can be seen from Figure 1 in Lynch and Ritland (1999). Furthermore, the single-locus variance may not be able to reach an asymptotic value with an increasing number of loci. If 1, 5, 10, 20, 40, and 80 loci [each having 10 alleles with known frequencies in proportions 1, 2, 3, . . . , 10 (triangular distribution)] are used in estimating the relatedness for full sibs, for example, the single-locus sampling variances for $\Delta$ are 0.27, 0.36, 0.42, 0.49, 0.55, and 0.62, respectively. The total sampling variance always decreases but at a decelerating rate with an increasing number of loci. For nonrelatives, the single-locus sampling variance always decreases to an asymptotic value with an increasing number of loci. The odd behavior of the LR estimator is perhaps due to the weights among loci, which are obtained by assuming $r = 0$ and $\Delta = 0$ (Lynch and Ritland 1999). In the comparison below, I follow Lynch and Ritland (1999) in using 10 loci in the estimation for the LR estimator.

A bizarre property for QG, R, and LR estimators is that, for some allele frequencies, they are undefined because the denominator becomes zero. For biallelic loci, the QG estimator is undefined when the reference individual is a heterozygote. When the allele frequencies are exactly equal for biallelic loci and the reference genotype is a heterozygote, the LR estimator yields undefined estimates. Some equations in the R estimator are invalid when the allele frequency is either 0.5 or 0.25. Allele frequencies near these values cause erratically very large estimates from the three estimators because the denominator is close to zero. This is paradoxical because an allele frequency of 0.5 gives the maximum heterozygosity (for the allele in question) and thus should yield the best estimation. In the following, I avoid biallelic loci in the QG estimator and ignore those component estimates in the R and LR estimators whenever they “blow up.”

With known allele frequencies in proportions 1, 2, 3, . . . , $n$ (triangular distribution), the five estimators are compared in precision for different numbers of alleles per locus (Figure 1). For the estimation of $r$, estimators W, LL, and QG give essentially the same results for nonrelatives (almost indistinguishable in the graph) and very similar results for relatives when the number of alleles is large. If the number of alleles is small ($n < 5$), however, the QG estimator yields higher sampling variances. For nonrelatives, the LR and R estimators have basically the same precision, which is slightly higher than that of the other three estimators across the range of the number of alleles. For relatives, however, the LR and R estimators result in much larger sampling variances than the other estimators when $n$ is large. The sampling variance of $r$ for either full-sibs or parent-offspring nearly levels off for the LR estimator and is actually increasing for the R estimator, when $n$ increases above the value of 12.

For the estimation of $\Delta$, the W estimator gives slightly worse estimates for nonrelatives, slightly better estimates for parent-offspring, and much better estimates for full-sibs where the expected value of $\Delta$ is greater than zero. With full-sibs, the sampling variance for the R estimator begins to increase with $n$ when it is $> 6$.

As is intuitively obvious, an increase in the number of alleles per locus should reduce the sampling variance of relatedness estimates, because alleles identical in state are more reliable as indicators of identity-by-descent. The phenomenon that the sampling variance of relatedness ($r$ or $\Delta$) for relatives from the R and LR estimators does not decrease consistently with an increasing number of alleles is perhaps caused by the assumptions made in deriving weights in these estimators. Without prior information, it is reasonable to assume $r = 0$ and $\Delta = 0$ for deriving weights among loci (Lynch and Ritland 1999) and among component equations within a locus (Ritland 1996b; Lynch and Ritland 1999). With an increasing number of alleles, there would be an increasing number of weights to be derived and
Figure 1.—Single-locus sampling variances for estimates of \( r \) and \( \Delta \) as a function of the number of alleles per locus. The sampling variances of \( r \) estimates are obtained from the new estimator (W), LYNCH and RITLAND's (1999) estimator (LR), RITLAND's (1996b) estimator (R), LYNCH (1988) and LI et al.'s (1993) estimator (LL), and QUELLER and GOODMAN's (1989) estimator (QG); those of \( \Delta \) estimates are obtained from W, LR, and R estimators, assuming known allele frequencies in a triangular distribution. The results for the LR estimator are obtained by Monte Carlo simulations on 10 loci, those from the other estimators are exact solutions based on probabilities of genotype combinations given allele frequency distribution and actual relationship. Note that both axes are in log scale, and the sampling variances for \( r \) estimates of nonrelatives are similar and indistinguishable between the LR and R estimators and among W, LL, and QG estimators.
is the distribution for populations at equilibrium under the joint effects of drift and mutation or migration (Wright 1951). With allele frequencies drawn from such a distribution, the estimates were obtained by simulations for each of the five estimators. The results (Figure 2) turn out to be similar to those for triangular distribution (Figure 1), except that the W estimator is now slightly better than the LL and QG estimators in estimating $r$ in the case of parent-offspring relationship.

**Allele frequencies estimated from samples:** In practice, the allele frequencies of the population are generally unknown and are estimated from samples. Because of practical constraints, the sample sizes are usually not large enough for the sampling effects to be ignored. This is especially obvious for highly polymorphic loci, because they generally have more rare alleles whose frequencies are subject to larger (relative) sampling errors than common alleles for a given sample size. The sensitivity of different estimators to sample sizes for estimating allele frequencies and relatedness is, therefore, important in practice and should be investigated.

There are several practical complications when allele frequencies are estimated. First, a potentially great source of bias in estimating relatedness between a pair of individuals comes from the inclusion of these particular individuals in estimating allele frequencies (Queller and Goodnight 1989; Ritland 1996b). This bias can be removed by excluding the two particular individuals from the sample in estimating the population allele frequencies. Second, an additional problem with the LR and R estimators is that the allele frequencies for estimating relatedness between a particular pair of individuals have to be larger than zero; otherwise infinite estimates could be generated. It is possible that an infrequent allele in a sample is found only in a particular pair of individuals, resulting in a failure in estimating
Figure 3.—Biases and single-locus sampling variances of $r$ and $\Delta$ estimates for nonrelatives (A), parent-offspring (B), and full-sibs (C) as a function of sample size used to estimate population allele frequencies. The biases and sampling variances of $r$ estimates are obtained from the new estimator (W), LYCH and RITLAND’s (1999) estimator (LR), RITLAND’s (1996b) estimator (R), LYCH (1988) and LI et al.’s (1993) estimator (LL), and QUELLER and GOODNIGHT’s (1989) estimator (QG); those of $\Delta$ estimates are obtained from W, LR, and R estimators, by Monte Carlo simulations based on 10 loci, each having 20 alleles with their population frequencies in a triangular distribution. Note that, for estimating $r$ of nonrelatives (A), the W and LL estimators are indistinguishable in both sampling bias and variance.
relatedness between the two individuals because the allele frequency estimated from the sample by excluding the particular individuals is zero. Whenever this happens, I follow Ritland’s (1996b) suggestion to bin such an allele into the least (nonzero) frequent allele in the sample. Third, all moment estimators for relatedness do not allow for estimating allele frequencies and relatedness simultaneously. When relatives are involved in the sample, the simple gene-counting method for estimating allele frequencies results in an additional sampling error due to the ignorance of actual relationship. Although an iterative procedure (like the one for estimating relatedness and weights described above) can be used to estimate relatedness and allele frequencies jointly, it results in a nonlinear moment estimator that yields worse estimates (Ritland 1996b). In the following, therefore, the allele frequency is estimated by ignoring any relationships among sampled individuals, and the sensitivity of different estimators to this treatment is compared over different proportions of close relatives in the sample. Fourth, although the sample allele frequency is unbiased, its higher moments are biased (Crow and Kimura 1970; Weir 1996). The bias can be corrected conveniently in the W and LL estimators as shown above (12–14). Using the $\Delta$-method based on Taylor expansion (Lynch and Walsh 1998), we can obtain approximate formulas for the expected relatedness with sample allele frequencies being the random variables. The formulas are complicated even though partials higher than the second order are ignored, because the estimators are ratios of allele frequencies. Furthermore, simulations show that these approximate formulas are even worse in precision and accuracy than the original ones not corrected for sample size. The reason is perhaps that we do not know the expected values of the random variables (in our case, sample allele frequencies) and have to use the observed values as their expectations in the expansion. For the $R$ estimator, it is almost impossible to obtain corresponding formulas correcting for sample size. In the following, therefore, the sample sizes are corrected for the $W$ and $LL$ estimators but not for the other estimators.

The biases and single-locus variances of $r$ and $\Delta$ estimates obtained from different estimators by Monte Carlo simulations on 10 loci are compared in Figure 3 for nonrelatives, parent-offspring, and full-sibs. Twenty alleles per locus were assumed to be in a triangular distribution of frequency in the population, and samples of various sizes were used to estimate the allele frequencies. Across the ranges of sample sizes and actual relationships, the $W$, $LL$, and $QG$ estimators are essentially unbiased, while the LR and $R$ estimators give upward-biased estimates of both $r$ and $\Delta$. The bias of the $R$ estimator is very high for relatives and small sample sizes. Although in all cases the bias of the $R$ estimator decreases with increasing sample size, the decrease is rapid only when the actual relatedness ($r$ or $\Delta$) is zero (Figure 3, A and B, for $\Delta$).
The sampling variances of relatedness estimates from the W, LL, and QG estimators are nearly constant over sample sizes for any actual relationship. For the LR and R estimators, the sampling variances decrease with increasing sample sizes when the actual relatedness \( r \) or \( \Delta \) is zero. With the actual relatedness greater than zero, the sampling variances of the LR and R estimators show little change or even increase slightly with sample sizes in the range of 20–100 sampled individuals. Much larger sample sizes are required for the R estimator to yield sampling variances close to the asymptotic values obtained assuming known allele frequencies (Figure 1). For full-sibs in Figure 3C as an example, the single-locus sampling variances (biases) of \( r \) and \( \Delta \) estimates from the R estimator are 0.73 and 3.24 (0.02 and 0.02), respectively, when 500 individuals are sampled, much smaller than the corresponding values 5.42 (0.10 and 0.11) when 100 individuals are sampled, but still larger than the corresponding values 0.42 and 1.23 (0 and 0) with known allele frequencies (in Figure 1).

From Figure 3, it is clear that the LR and R estimators are more sensitive to sampling than the other estimators. The R estimator is especially susceptible to the sizes of samples used in estimating allele frequencies, and the main cause is the treatment of alleles found only in a particular pair of individuals whose relatedness is being estimated. Other treatments of such alleles, such as including them in estimating allele frequencies, also result in large bias and sampling variance for the R and LR estimators. It seems to be difficult for the two estimators to overcome the problems brought by rare alleles in the sample.

In Figure 3, loci with 20 alleles in a triangular distribution of frequency were used in the estimation. More alleles per locus or/and more leptokurtic distributions of allele frequency would require larger sample sizes for the R and LR estimators to obtain estimates with negligible bias and variance close to the asymptotic values obtained with known allele frequencies. Ritland (1996b) showed that, for less polymorphic loci with 2–4 alleles, a sample size of 40–60 individuals is sufficient for estimating population allele frequency and for obtaining unbiased relatedness estimates with variance close to the asymptotic value with known allele frequencies. It is clear that the sample size required for a reasonably good estimate from the R estimator increases rapidly with an increasing number of alleles per locus.

In the above, samples are assumed to be taken binomially from the population and to include nonrelatives only (except for the particular pair of individuals under estimation of relatedness). Clusters of family members may be included in the samples in practice, introducing further sampling errors in estimating allele frequencies (and thus relatedness) if the actual relationship structure among sampled individuals is not accounted for. To investigate the robustness of different estimators when samples involving relatives are used, I considered different proportions of full-sibs from a single family in the sample. Figure 4 compares the biases and single-locus sampling variances among estimators when different proportions of full-sibs are included in a sample of 100 individuals for estimating allele frequencies. The actual relationship is parent-offspring, and 10 loci with 20 alleles in a triangular distribution of frequency were used in estimating relatedness. As is clear from Figure 4, the 

**Discussion**

When population allele frequencies are known, all the pairwise relatedness estimators yield unbiased estimates. These estimates derived from single loci are, however, highly variable with values frequently falling outside of the true range of \( (0, 1) \). The W, QG, and LL estimators do not return estimates \( >1 \), but do give values \( <0 \). The other estimators can yield estimates either \( >1 \) or \( <0 \). Although estimators can be developed to constrain the estimates to the correct range of relatedness (e.g., Thompson 1976), they are inevitably biased and the magnitude of the bias depends on the actual relationship being estimated (Lynch and Ritland 1999).

The large sampling variance of pairwise relatedness estimates comes from several possible sources (Lynch and Ritland 1999). First, although the probability of identity-by-descent (IBD) is the same among loci for a pair of individuals of any given relationship, the realized (observed) frequencies of IBD might be variable over loci, resulting in a variance in IBD among loci per se. Such original variances of \( r \) and \( \Delta \) are 0 for nonrelatives and parent-offspring, \( 1/8 \) and \( 1/16 \) for full-sibs, and \( 1/16 \) and 0 for half-sibs. These are the minimum single-locus sampling variances of estimates that can be realized by an efficient estimator in the most favorable conditions. Indeed, Figures 1 and 2 show that the single-locus sampling variances for \( r \) from estimators W, LL, and QG, and for \( \Delta \) from estimator W are approaching these minimum values asymptotically with an increasing number of alleles \( n \). The single-locus sampling variances for the LR and R estimators asymptote to the minimum values only when the actual \( r \) or \( \Delta \) is zero. Otherwise, they increase for the R estimator, or reach gradually asymptotic values much larger than the minimum variances for the LR estimator, with an increasing \( n \). The problem...
Figure 4.—Biases and single-locus sampling variances of \( r \) and \( \Delta \) estimates for parent-offspring as a function of the proportion of full-sibs from a single family included in a sample of 100 individuals used to estimate population allele frequencies. The biases and sampling variances of \( r \) estimates are obtained from the new estimator (W), LYNCH and RITLAND’s (1999) estimator (LR), RITLAND’s (1996b) estimator (R), LYNCH (1988) and LI et al.’s (1993) estimator (LL), and QUELLER and GOODNIGHT’s (1989) estimator (QG); those of \( \Delta \) estimates are obtained from W, LR, and R estimators, by Monte Carlo simulations based on 10 loci, each having 20 alleles with their population frequencies in a triangular distribution.

for the R and LR estimators is caused by the inappropriate weights assuming null relationships.

Second, a great source of the high sampling variance stems from the variation in identity-in-state for alleles that are not identical by descent. More informative loci would give a lower probability of similarity between individuals due to chance (identity-in-state) rather than relationship (identity-by-descent). As is clear from Figures 1 and 2, the sampling variance of relatedness for any estimator changes dramatically with the number of alleles per locus and their frequency distributions. However, there is no general and simple quantity, such as heterozygosity or number of alleles, that could characterize the amount of information at a locus in estimating pairwise relatedness. For the R estimator, the single-locus sampling variances of \( r \) and \( \Delta \) estimates are \( 1/(n-1) \) and \( 2/[n(n-1)] \), respectively, for nonrelatives, which are determined by \( n \) only and irrespective of allele frequencies or heterozygosity (RITLAND 1996b; LYNCH and RITLAND 1999). However, the sampling variances for relatives from the R estimator depend on, apart from the actual relationship, both \( n \) and allele frequencies, and they have no simple relationship with either \( n \) or heterozygosity. This is also true for all the other estimators in estimating any actual relationship. The conclusion is relevant to the selection of marker loci in estimating relatedness in practice. Although the sampling variances for \( r \) and \( \Delta \) are invariably inversely proportional to the number of loci \((L)\), they have a variable relationship with \( n \) depending on the actual relatedness being estimated and allele frequencies. For a given allele frequency distribution and actual nonzero relatedness, the sampling variances generally decrease faster with increasing \( n \) when \( n \) is smaller (Figures 1 and 2). With the total number of alleles \( Ln \) fixed, therefore, an increase in \( n \) decreases the sampling variances to a greater extent than an increase in \( L \) when \( n \) is small. While an increase in \( L \) decreases both sampling variance components due to variation in identity-by-descent and in identity-in-state, respectively, an increase in \( n \) only reduces the variance caused by the variation in identity-in-state. For actual relationships with large variance in identity-by-descent among loci such as full-sibs, it is more desirable to increase \( L \) except when \( n \) is very small.

Third, an additional source of single-locus sampling variance for the LR estimator is the number of loci used in the estimation. The problems of the LR estimator are that with an increasing number of loci the average single-locus sampling variance may either increase or decrease, depending on allele frequency distribution and actual relationship, and that the variance may not be able to reach an asymptotic value with a realistically large number of loci, especially for highly polymorphic loci. The cause of the bizarre behavior is the weights among loci, which are derived under the assumption of null relationship.

Fourth, there are differences in sampling variances among statistical methods for estimating relatedness
(Figures 1 and 2). The difference is especially large when the distribution of allele frequencies is leptokurtic, the number of alleles per locus is large, and the actual relatedness is greater than zero. The larger sampling variances for relatives from the R and LR estimators are caused mainly by the inappropriate weights derived assuming null relatedness. The difficulty with the R and LR estimators is that the optimum weights that result in the maximum precision are a function of the actual relationship being estimated (Ritland 1996b; Lynch and Ritland 1999). Another problem with these estimators as well as the QG estimator is that they are undefined for some allele frequencies. The maximum-likelihood method can be used to distinguish among different relationships (e.g., Thompson 1975; Marshall et al. 1998; Goodnight and Queller 1999). Though it can also be used to estimate relatedness, it results in much larger biases and sampling variances than the moment estimators unless the number of marker loci is unrealistically large (Ritland 1996b; Lynch and Ritland 1999; data not shown). For low true values of $r$, for example, at least $\sim 70$ di allelic loci ($p = 0.5$) are necessary for the likelihood estimator to have a similar precision to that of moment estimators (Lynch and Ritland 1999). This seems paradoxical because the likelihood method uses all information available, while all moment estimators necessarily merge the genotypic information to a certain extent, in one way or another. Such merging or grouping must lose some information. However, the moment estimators still outperform the likelihood estimator when the number of loci is not very large. This is perhaps because the ideal properties of the likelihood method are asymptotic and apply to large data sets only. The data set involved in estimating pairwise relatedness is extremely small, only two individual genotypes. For given allele frequencies and relationship, there is large uncertainty for a pair of genotypes. Therefore, a very large number of loci are necessary for the likelihood method to perform better. This is possible for human data, where hundreds of markers have been developed, but not realistic now for most other species.

When allele frequencies of the population are unknown and estimated from samples, more problems arise with the moment estimators. These problems are especially evident with highly polymorphic loci and uneven allele frequency distributions because of the involvement of rare alleles. I have shown that the W, LL, and QG estimators give essentially unbiased estimates with sampling variances that are almost constant, irrespective of sample sizes used for estimating allele frequency. The R and LR estimators are, however, more sensitive to sample sizes in both accuracy and precision. The upward bias and sampling variance for both $r$ and $\Delta$ from the R estimator are large (Figure 3) and decrease very slowly with increasing sample size when the actual relatedness is greater than zero. With a realistic sample size and highly polymorphic loci, the R and LR estimators could be worse than the other estimators even in the case of nonrelatives (Figure 3A).

Except for the R estimator, all the moment estimators are robust to clusters of close relatives included in samples for estimating allele frequencies (Figure 4). Even with 40% of the individuals in a sample being full-sibs from a single family, the moment estimators (except for the R estimator) give essentially the same estimates (in terms of bias and sampling variance) as those in the ideal case of no relatives in the sample. This is comforting because moment estimators do not allow for estimating allele frequencies and relatedness simultaneously and samples in practice may include clusters of relatives.

The advantage of the new estimator over LR and R estimators increases with increasing degrees of actual relatedness. Throughout I considered the extreme cases of $r = 0$ (nonrelatives) and $r = 0.5$ (parent-offspring, full-sibs). Higher relatedness is possible, but should be found rarely in natural outbreeding populations. With very low degrees of actual relatedness, the LR estimator could become the best if the sample size is very large and the loci are not highly polymorphic. In almost all other cases, however, the new estimator is the most efficient for estimating $r$ and $\Delta$ jointly. In current practice, highly polymorphic markers such as microsatellites are preferred in genetic studies, which could have many alleles at a locus (53 alleles at a locus was reported in a great tit population; Van de Casteele et al. 2001) in highly leptokurtic frequency distribution. The samples are usually small in size and possibly contain unknown relatives. In such situations, the newly developed estimator is likely to perform better than LR and R estimators in estimating $r$ and $\Delta$ simultaneously.

I thank Andrew Bourke, Rob Hammond, Bill Jordan, Michael Lynch, Kermit Ritland, and three anonymous referees for helpful comments on an earlier version of this article.

**LITERATURE CITED**


Weir, B. S., 1996 Genetic Data Analysis II. Sinauer Associates, Sunderland, MA.


Communicating editor: A. H. D. Brown