Estimation of Pairwise Relatedness With Molecular Markers

Michael Lynch* and Kermit Ritland†

**Department of Biology, University of Oregon, Eugene, Oregon 97403 and †Department of Forest Sciences, University of British Columbia, Vancouver, British Columbia V6T1Z4, Canada*

Manuscript received June 26, 1998
Accepted for publication April 19, 1999

ABSTRACT

Applications of quantitative genetics and conservation genetics often require measures of pairwise relationships between individuals, which, in the absence of known pedigree structure, can be estimated only by use of molecular markers. Here we introduce methods for the joint estimation of the two-gene and four-gene coefficients of relationship from data on codominant molecular markers in randomly mating populations. In a comparison with other published estimators of pairwise relatedness, we find these new "regression" estimators to be computationally simpler and to yield similar or lower sampling variances, particularly when many loci are used or when loci are hypervariable. Two examples are given in which the new estimators are applied to natural populations, one that reveals isolation-by-distance in an annual plant and the other that suggests a genetic basis for a coat color polymorphism in bears.

COEFFICIENTS of relationship between pairs of individuals play a central role in many areas of genetics and behavioral ecology. For example, in quantitative genetics, the phenotypic resemblance of relatives, which forms the basis for the empirical estimation of components of genetic variance, is a direct function of the probability that individuals have one or two genes identical by descent at a locus. Given such probabilities, causal components of variance (such as the additive and dominance genetic variance) can be estimated from the phenotypic covariance (Falconer and Mackay 1996; Lynch and Walsh 1998). In studies of laboratory or domesticated populations, where investigators can be certain of the degrees of relationship among observed individuals, the application of conventional quantitative-genetic methodology is straightforward. Major uncertainties about the relationships among individuals from natural populations are the primary impediment to extending quantitative-genetic analysis to field studies, but Ritland (1989, 1996a) has suggested how this problem might be overcome by regressing pairwise measures of phenotypic similarity on pairwise estimates of relatedness obtained with molecular markers.

Pairwise measures of relatedness also play a role in the field of conservation genetics. For example, in captive breeding programs, substantial effort is being made to ensure that matings are minimized between close relatives to reduce the loss of genetic variation by random genetic drift. If the potential parents are derived directly from wild-caught stock or are descendants of individuals of unknown relationship, a relative ranking of degrees of relatedness can only be achieved through inferences with molecular markers (Avise 1995).

A third field of inquiry within which pairwise relatedness plays a significant role is the evolution of social behavior. Studies in this area are largely focused around Hamilton's (1964) theory of kin selection, which states that the evolutionary advantage of an altruistic act depends on whether the cost to the donor exceeds the benefit to the recipient multiplied by the relatedness between the two individuals. Because most such studies involve field populations where parentage is not directly observed, indirect inferences about relatedness must again be made with molecular markers.

In all of the above-mentioned applications of molecular markers, it is an implicit assumption that such markers provide reasonable, if not excellent, estimates of relatedness coefficients. Yet, there are few existing methods for the estimation of pairwise relatedness for which the statistical properties are well understood or well behaved. Several estimators have been developed for pairwise relatedness using the rather specialized data provided by DNA-fingerprint profiles (Lynch 1988; Li *et al.* 1993; Geyer and Thompson 1995). Following up on earlier work of Pamilo and Crozier (1982), Queller and Goodnight (1989) developed marker-based estimators for within-group relatedness, but these are of somewhat limited applicability in the estimation of pairwise relationship because of their poor behavior with diallelic loci. An efficient method-of-moments estimator, recently developed by Ritland (1996b), provides a basis for the joint estimation of identity-by-descent at both the genic and genotypic levels. Ritland's approach, which is based on a model involving joint probabilities of the two genotypes of a pair, can be quite complex computationally and is ill-behaved with some gene frequencies. Maximum-likelihood methods have

Corresponding author: Michael Lynch, Department of Biology, University of Oregon, Eugene, OR 97403.
E-mail: mlynch@oregon.uoregon.edu

been developed by Thompson (1975, 1976, 1986) to test for specific types of relationship.

In this article, we introduce a simple method for obtaining unbiased estimates of pairwise relationship coefficients. Its simplicity arises from the use of a regression approach for inferring relationship—one individual of a pair serves as a "reference," and the probabilities of the locus-specific genotypes in the other "proband" individual are conditioned on those of the reference. Aside from its ease of application and unbiased nature, this method has two very useful features—it generates joint estimates of both the two- and four-gene coefficients of relatedness, and it yields simple expressions for the sampling variance of these coefficients. This latter feature provides a convenient means for optimizing the use of information derived from different loci. Following our derivation of the regression method, we compare its performance against that of other methods and then provide two examples of its application to studies of natural populations.

JOINT ESTIMATION OF TWO-GENE AND FOUR-GENE COEFFICIENTS

Throughout, we focus on the traditional definition of relatedness for individual pairs of diploid individuals, $r_{xy} = 2\Theta_{xy}$, where the coefficient of coancestry, Θ_{xy}, is the probability that, for any autosomal locus, a random gene taken from individual x is identical by descent with a random gene taken from individual y. For monozygotic twins (and clonemates), $r_{xy} = 1$; for parent-offspring and full-sib relationships, $r_{xy} = 0.5$; and for second- and third-order relationships, r_{xy} is equal, respectively, to 0.25 and 0.125.

The relatedness coefficient for two individuals (x and y) is a linear function of two "higher-order" coefficients,

$$r_{xy} = \frac{\phi_{xy}}{2} + \Delta_{xy} . \qquad (1)$$

If we consider all four genes possessed by two individuals at a locus, ϕ_{xy} is the probability that a single gene in x is identical by descent with one in y, and Δ_{xy} is the probability that each of the two genes in x is identical by descent with one in y. For parents and offspring, $\phi_{xy} = 1$ and $\Delta_{xy} = 0$; for full sibs, $\phi_{xy} = 0.5$ and $\Delta_{xy} = 0.25$; and for half sibs, $\phi_{xy} = 0.25$ and $\Delta_{xy} = 0$. For many applications, such a subdivision of r_{xy} is unnecessary, but in quantitative genetics, a knowledge of the higher-order coefficient Δ_{xy} is desirable because the expected genetic covariance between individuals is defined to be

$$\sigma_{xy} = r_{xy}\sigma_A^2 + \Delta_{xy}\sigma_D^2 ,$$

where σ_A^2 and σ_D^2 are the additive and dominance components of genetic variance for a quantitative trait. This expression assumes a random-mating population, which we also assume throughout. Inbreeding introduces the need for additional higher-order coefficients of relatedness and genetic components of variance (Cockerham 1971; Jacquard 1974). Higher-order terms must also be added to the previous expression when epistatic sources of genetic variance are present, but provided the population is randomly mating, no relationship coefficients are required beyond r_{xy} and Δ_{xy} (Kempthorne 1954; Lynch and Walsh 1998).

In the following analyses, we focus on the estimation of r_{xy} and Δ_{xy}, as these are the relationship coefficients that are of primary practical utility. Our computer simulations showed that estimates of ϕ_{xy} have much higher sampling variance than those of r_{xy} and Δ_{xy}, enough so that the accurate measurement of ϕ_{xy} is beyond reach unless very large numbers of informative loci can be assayed. This large sampling variance does not carry over greatly to estimates of the composite measure r_{xy}, because there is also a very large negative sampling covariance between the two component coefficients, ϕ_{xy} and Δ_{xy}.

Genotypic probabilities: There are two fundamental ways to set up a model for the genotypic probabilities in a pair of individuals. The first approach, adopted by Ritland (1996b), specifies the joint probability of both genotypes. The second approach, adopted here, specifies the conditional genotypic probability of a proband individual y, given the genotype of the reference individual x. We refer to these two approaches as "correlation" and "regression" methods in the sense that they are symmetrical *vs.* asymmetrical measures. Both approaches allow the joint estimation of r_{xy}, ϕ_{xy}, and Δ_{xy}, but as we will see, correlation and regression estimators differ substantially in terms of complexity and statistical properties. It is important to note that our use of the terms correlation and regression refers to the underlying statistical model and not to the estimators themselves. The estimators developed here and in Ritland (1996b) are more properly termed "method-of-moments" estimators.

Consider a single locus with n alleles, and let x be the reference individual (with alleles a and b) and y be the proband individual (with alleles c and d). The conditional probabilities for the $n(n + 1)/2$ possible genotypes in y can be expressed as a function of ϕ_{xy}, Δ_{xy}, and the known allele frequencies,

$$P(y = cd|x = ab) = P_0(cd) \cdot (1 - \phi_{xy} - \Delta_{xy}) + P_1(cd|ab) \cdot \phi_{xy} + P_2(cd|ab) \cdot \Delta_{xy} , \qquad (2)$$

where $P_0(cd)$ is the Hardy-Weinberg probability of genotype cd, and $P_1(cd|ab)$ and $P_2(cd|ab)$ denote the probabilities of genotype cd in y given genotype ab in x, the first being conditional on the two individuals having one gene identical by descent and the second being conditional on two genes being identical by descent.

Regression estimators: Equation 2 provides the foundation for the regression-based estimators that we now

explore. To illustrate the general approach, we first derive estimators conditioned on the observation of a homozygote reference genotype. In this straightforward case, two probabilities are informative about *x*'s relationship with individual *y*: $P(ii|ii)$ and $P(i\cdot|ii)$, the conditional probabilities that the two individuals have two *vs.* one pair of genes identical in state at the locus, with a dot denoting any allele other than *i*. The probability of no genes identical in state, $P(\cdot\cdot|ii)$, provides no additional information, as it simply equals $[1 - P(ii|ii) - P(i\cdot|ii)]$. Letting p_i be the frequency of the *i*th allele, from Equation 2,

$$P(ii|ii) = p_i^2 + p_i(1 - p_i)\phi_{xy} + (1 - p_i^2)\Delta_{xy} \quad (3a)$$

$$P(i\cdot|ii) = 2p_i(1 - p_i) + (1 - p_i)(1 - 2p_i)\phi_{xy} - 2p_i(1 - p_i)\Delta_{xy}\,. \quad (3b)$$

Assuming that we know the allele frequency p_i in advance, these two equations can be rearranged to yield estimators for the two unknown relationship coefficients,

$$\hat{\phi}_{xy} = \frac{(1 + p_i)\hat{P}(i\cdot|ii) + 2p_i\hat{P}(ii|ii) - 2p_i}{(1 - p_i)^2} \quad (4a)$$

$$\hat{\Delta}_{xy} = \frac{p_i^2 - p_i\hat{P}(i\cdot|ii) + (1 - 2p_i)\hat{P}(ii|ii)}{(1 - p_i)^2}\,, \quad (4b)$$

and from Equation 1,

$$\hat{r}_{xy} = \frac{\hat{P}(i\cdot|ii) + 2\hat{P}(ii|ii) - 2p_i}{2(1 - p_i)}. \quad (4c)$$

Throughout, we use a $\wedge$ to distinguish an estimator from its parametric value. For any pair of observed individuals, the two probabilities necessary for the solution of these equations, $\hat{P}(i\cdot|ii)$ and $\hat{P}(ii|ii)$, are estimated as 0/1 variables, with 1's being given to observed two-genotype combinations and 0's being given to unobserved combinations. (Both probabilities are 0 if the proband has no alleles in common with the reference.) Thus, for example, when individual *y* contains 2, 1, and 0 *i* alleles, the estimate $\hat{r}_{xy}$ is 1, $(1 - 2p_i)/[2(1 - p_i)]$, and $-p_i/(1 - p_i)$, respectively.

The appendix provides a parallel set of results for heterozygotes at diallelic and multiallelic loci. Diallelic heterozygous reference individuals introduce no new problems, but with multiallelic loci, there are six classes of conditional probabilities for heterozygous reference individuals. In the latter case then, the number of observed 0/1 variables exceeds the number of unknowns (ϕ and Δ). To deal with this situation, we provide a weighted least-squares approximation.

A general estimator, which covers all three cases, is best described by introducing "indicator variables" for the sharing of *pairs* of alleles (as opposed to more complex patterns of sharing as used earlier). As before, let the reference individual have alleles *a* and *b* and the proband individual alleles *c* and *d*. If the reference individual is homozygous, $S_{ab} = 1$, while if it is heterozygous, $S_{ab} = 0$. Likewise, if allele *a* from the reference individual is the same as allele *c* from the proband, $S_{ac} = 1$, while $S_{ac} = 0$ if it is different. In total, there are six *S*'s corresponding to the six ways of choosing two objects without replacement from a pool of four objects. Letting p_a and p_b be the frequencies of alleles *a* and *b* in the population, the fully general expressions for the two coefficients of primary interest are

$$\hat{r}_{xy} = \frac{p_a(S_{bc} + S_{bd}) + p_b(S_{ac} + S_{ad}) - 4p_ap_b}{(1 + S_{ab})(p_a + p_b) - 4p_ap_b} \quad (5a)$$

$$\hat{\Delta}_{xy} = \frac{2p_ap_b - p_a(S_{bc} + S_{bd}) - p_b(S_{ac} + S_{ad}) + (S_{ac}S_{bd}) + (S_{ad}S_{bc})}{(1 + S_{ab})(1 - p_a - p_b) + 2p_ap_b}\,. \quad (5b)$$

In actual practice, there is no particular reason to use one member of a pair of individuals as the reference as opposed to the other member. Thus, the reciprocal estimates $\hat{r}_{xy}$ and $\hat{r}_{yx}$, etc., can be arithmetically averaged to further refine the pairwise relationship estimates for the pair of individuals *x* and *y*. In all of the following analyses, we rely on such reciprocal estimates, as the arithmetic average of the two reciprocal estimates generally has a lower statistical variance than a single estimate. In principle, the root of the product of the two reciprocal estimates could be used, but this leads to undefined estimates in the event that one is negative.

Multilocus estimates: Estimates of relatedness are usually based on data from multiple loci. Under the assumption that the marker loci are unlinked, the locus-specific estimates are independent. However, any averaging of the locus-specific estimates to obtain overall estimates of r_{xy} and Δ_{xy} should account for the dramatic among-locus differences of sampling variance that can arise from both differences in reference genotypes (*e.g.*, common homozygote *vs.* rare heterozygote) and in levels of variation (loci with more alleles being more informative).

Let $w_{r,x}(\ell)$ and $w_{\Delta,x}(\ell)$ denote the weights to be used for the ℓth locus in the overall estimates of r_{xy} and Δ_{xy}, and let $W_{r,x}$ and $W_{\Delta,x}$ be the sums of the weights over all *L* loci. The composite estimates of the relationship coefficients for *x* and *y* are then

$$\hat{r}_{xy} = \frac{1}{W_{r,x}}\sum_{\ell=1}^{L} w_{r,x}(\ell)\hat{r}_{xy}(\ell) \quad (6a)$$

$$\hat{\Delta}_{xy} = \frac{1}{W_{\Delta,x}}\sum_{\ell=1}^{L} w_{\Delta,x}(\ell)\hat{\Delta}_{xy}(\ell)\,. \quad (6b)$$

With statistically independent marker loci, the locus-specific weights that minimize the sampling variance of the overall estimates $\hat{\phi}_{xy}$ and $\hat{\Delta}_{xy}$ are simply the inverses of the sampling variances of the locus-specific estimates. As noted in the appendix, we cannot be very certain of the numerical values of the weights because they are functions of the parameters that we are trying to esti-

mate, but approximations can be obtained by simply assuming that x and y are unrelated. The locus-specific weights are then given by the inverses of the sampling variances of estimates of the relatedness coefficients for nonrelatives conditional on the genotype in x. General expressions for the weights are given by

$$w_{r,x}(\ell) = \frac{1}{\mathrm{Var}[\hat{r}_{xy}(\ell)]} = \frac{(1 + S_{ab})(p_a + p_b) - 4p_ap_b}{2p_ap_b} \quad (7a)$$

$$w_{\Delta,x}(\ell) = \frac{1}{\mathrm{Var}[\hat{\Delta}_{xy}(\ell)]} = \frac{(1 + S_{ab})(1 - p_a - p_b) + 2p_ap_b}{2p_ap_b} \quad (7b)$$

with S_{ab} equal to 1 when x is homozygous and equal to 0 when x is heterozygous.

Properties of the regression estimators: Extensive computer simulations demonstrated that the regression estimators given above are essentially unbiased, regardless of the numbers of loci or the values of ϕ and Δ. Thus, the primary issues of interest are the magnitudes of the sampling variances of the estimators and their sensitivity to the degree of actual relationship and to the allele-frequency distribution.

We obtained estimates of the sampling variances of the regression estimators by Monte Carlo simulation, assuming gene frequencies were known without error and assuming a random mating population with unlinked marker loci. Reference genotypes were drawn randomly according to their Hardy-Weinberg frequencies, and the genotypes of the paired individuals were then obtained from the conditional genotype distributions given the reference genotype and the particular relationship. For multiallelic loci, two types of allele-frequency distributions were considered: uniform distributions, in which the frequencies of each of the n alleles per locus were equal to $1/n$, and "triangular" distributions, in which the frequencies of alleles followed the proportions 1, 2, . . . , n. In all of the following figures, we report the single-locus sampling variances of the relationship coefficients. For analyses involving multiple loci with identical allele frequencies, the sampling variance of multilocus estimates can be obtained by dividing the plotted values by the number of loci (L).

A special property of the regression estimator is that the expected single-locus sampling variance declines with increasing numbers of unlinked loci, down to an asymptotic value (Figure 1). This dependence on number of loci arises with the regression estimator because the estimation variances (the weights) differ among alternative reference genotypes at the same locus (for example, a reference genotype having rarer alleles gives estimates with lower variance). By contrast, the correlation estimator of Ritland (1996b) is not conditioned upon observed genotype, and its variance only depends on the distribution of gene frequencies in the population. Although Figure 1 details the influence of the number of loci on the variance of the regression estimator, for the remaining analyses we focus on the situation in which 10 informative loci have been sampled. At that point, the lower asymptotic value of the single-locus sampling variance is closely approximated in most situations, and 10 loci is a good approximation of the sampling scheme employed in many empirical studies, with diallelic loci corresponding to isozymes and multiallelic loci corresponding to microsatellites.

For diallelic loci, the asymptotic sampling variance per locus for $\hat{r}$ is equal to 1 in the case of nonrelatives and somewhat lower for related individuals (even though nonoptimal weights are employed with relatives; Figure 1). With allele frequencies approaching 0.5, the optimal weights of all reference genotypes approach equality regardless of the degree of relationship, because all alleles are then equally informative. Thus, the asymptotic sampling variances near allele frequencies of 0.5 are the best that one could expect to achieve even if the correct weights were used. Because even with close relatives, the sampling variance is never less than about 0.4 per locus, these results imply that with a large number of loci, the expected standard error of $\hat{r}$ is generally on the order $1/\sqrt{L}$ when diallelic loci are assayed, somewhat greater if loci with extreme allele frequencies are included, and slightly less with close relatives.

As in the case of $\hat{r}$, the single-locus sampling variance of $\hat{\Delta}$ depends on the number of loci sampled, but the sensitivity to this is reduced at moderate allele frequencies (Figure 1). For all degrees of relationship, the asymptotic single-locus sampling variance for $\hat{\Delta}$ declines as allele frequencies become more equitable (Figure 1). It can exceed 10 when allele frequencies are extreme and is never much <1 with any type of relationship. Thus, as in the case of $\hat{r}$, with diallelic loci, the best that one can ever expect to achieve with the regression estimator is a multilocus standard error of $\hat{\Delta}$ equal to $1/\sqrt{L}$.

In principle, an increase in the number of alleles per locus should reduce the sampling variance of relatedness estimates, because alleles that are identical in state will be more reliable as indicators of identity by descent. For nonrelated individuals, the asymptotic single-locus sampling variance of $\hat{r}$ is very close to $1/(n - 1)$, regardless of the form of the allele-frequency distribution (Figure 2). With parents and offspring, the sampling variance is up to 50% less than this, while with other types of relatives it is somewhat higher when alleles with low frequency are common. Again, with an even allele-frequency distribution, all reference genotypes are equally informative regardless of the degree of relationship, so the results for this case can be viewed as the minimum sampling variance that one can expect to achieve with the regression estimator—except in the case of parents and offspring, a standard error of $\hat{r}$ less than about $1/\sqrt{L(n - 1)}$ is not achievable. Relative to the situation with $\hat{r}$, the rate of reduction in the asymptotic sampling

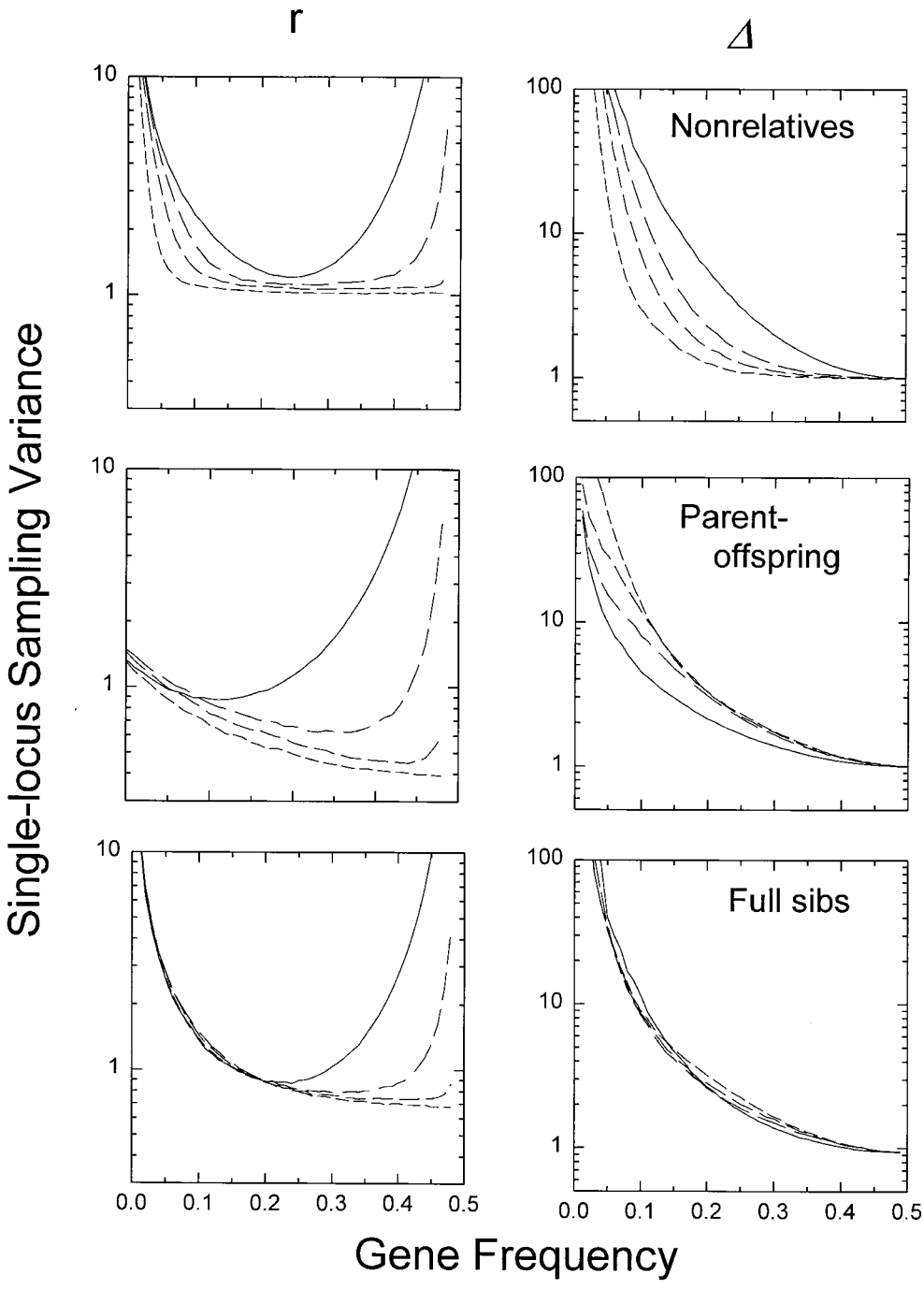

Figure 1.—Single-locus sampling variances for estimates of pairwise r and Δ for the range of possible gene frequencies at diallelic loci. For each gene frequency (in increments of 0.01) and degree of relationship, random pairs of multilocus genotypes were obtained by Monte Carlo simulation for 32,000 individuals. For each pair of individuals, the two reciprocal weighted estimates were obtained and then averaged to obtain the pairwise estimates. Solid lines, large dashes, medium dashes, and short dashes denote estimates based on 1, 5, 10, and 25 loci, respectively.

variance of $\hat{\Delta}$ with increasing n is more rapid (Figure 2). For nonrelatives, the asymptotic single-locus variance closely approximates $2/[n(n-1)]$ regardless of the form of the allele-frequency distribution.

COMPARISON WITH OTHER ESTIMATORS

As noted above, for applications in quantitative genetics, there is a need for separate estimates of r_{xy} and Δ_{xy} because the additive genetic covariance between individuals is a function of the composite measure r_{xy}, whereas the dominance genetic covariance is a function of Δ_{xy}. However, for situations in which one can be reasonably certain that the dominance genetic variance for a trait is negligible, or when one can be certain that collateral relatives (*e.g.*, pairs of individuals, such as full sibs and double first cousins, that share paternal and maternal genes) are absent, Δ_{xy} can be ignored. In addition, in many applications in conservation genetics and behavioral ecology, the composite estimate r_{xy} may provide all the information that is needed. Four additional estimators of r_{xy}, all of which are unbiased, have been previously described.

A simple estimator based on the sharing of alleles,

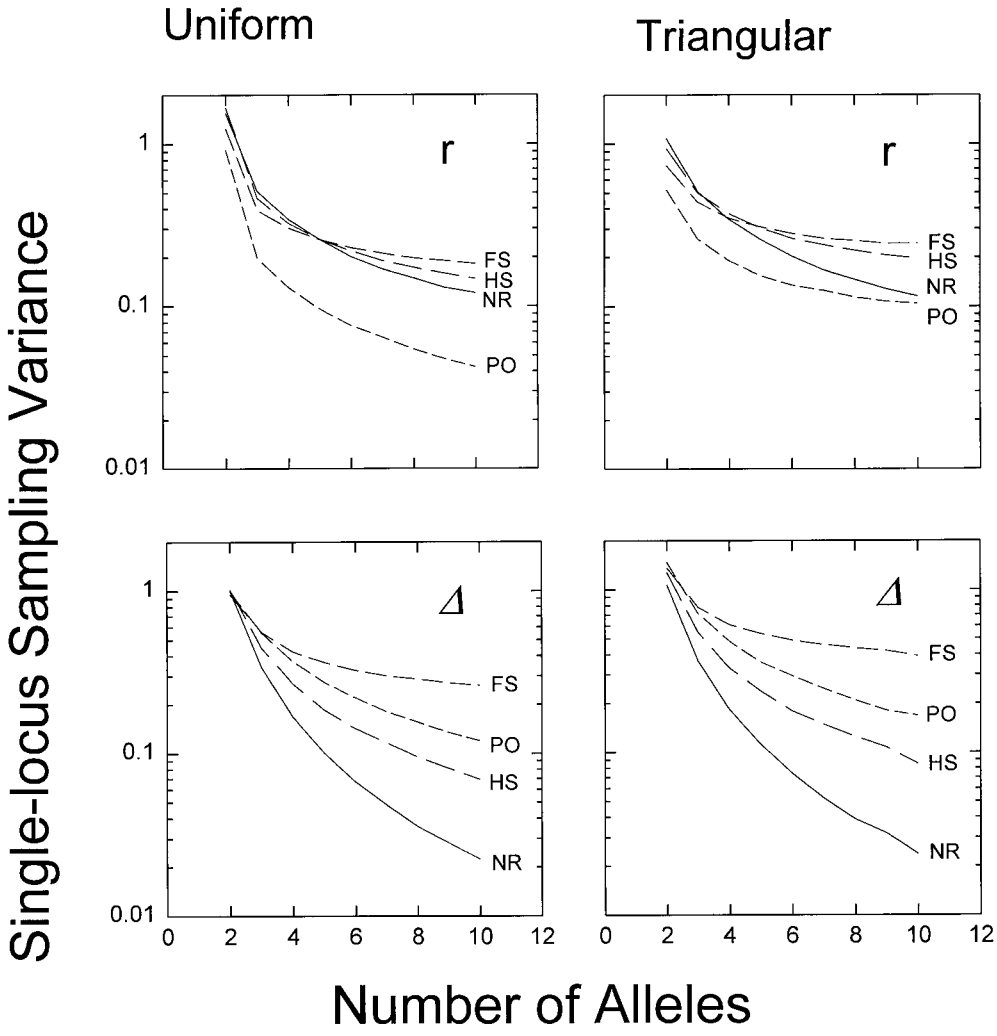

Figure 2.—Single-locus sampling variances for r and Δ as a function of number of alleles at loci with uniform and triangular allele-frequency distributions. Results are given for nonrelatives (NR), half sibs (HS), full sibs (FS), and parents and offspring (PO). The plotted values were obtained from Monte Carlo simulations of 10 loci (all with the same allele-frequency profile) for 32,000 pairs of individuals. Sampling variances of multilocus estimates of r and Δ are obtained by dividing the plotted values by the number of loci, keeping in mind that somewhat higher values are expected if <10 loci are observed.

proposed by Lynch (1988) for analyses employing DNA fingerprint profiles, can be generalized to any set of codominant markers. The following expression includes the slight modification suggested by Li *et al.* (1993). Define the similarity index, S_{xy}, to be the *average* fraction of genes at a locus in a reference individual (here either x or y) for which there is another gene in the proband that is identical in state. Thus, $S_{xy} = 1$ when ($x = ii$, $y = ii$) or ($x = ij$, $y = ij$), $S_{xy} = 0.75$ when ($x = ii$, $y = ij$) or vice versa, $S_{xy} = 0.5$ when ($x = ij$, $y = ik$), and $S_{xy} = 0$ when ($x = ij$, $y = kl$). A single-locus estimator for r_{xy} is then

$$\hat{r}_{xy} = \frac{S_{xy} - S_0}{1 - S_0}, \tag{8}$$

where $S_0 = \sum_{i=1}^{n} p_i^2(2 - p_i)$ is the expected value of S at the locus for unrelated individuals in a random-mating population. This simple estimator derives from the principle that if two individuals are related to degree r_{xy}, the expected fraction of genes that they have identical in state is the sum of the fractions shared because of identity-by-descent and because of identity-in-state (but not identity-by-descent), $E(S_{xy}) = r_{xy} + (1 - r_{xy})S_0$. Note that unlike the weighted regression estimator described above, Equation 8 does not return estimates of $r_{xy} > 1$. However, like the weighted regression estimator, Equation 8 does generate negative estimates whenever the observed S_{xy} is $< S_0$ because of sampling error. In the following, Equation 8 is referred to as the similarity-index estimator.

Like Equation 8, Ritland's (1996b) method-of-moments estimator for r_{xy} considers the joint distribution of both genotypes in a symmetrical way. The differing information provided by alternative alleles is incorporated by considering the incidence of each of the n possible alleles at the locus. The observed data are summarized as an array of n similarities, where the ith element (S_i) is equal to 0.0 (at most, one of the individuals contains allele i), 0.25 (both individuals contain a single i allele), 0.5 (one individual contains two and the other individual one i alleles), or 1.0 (both individuals are ii homozygotes). Estimates of r_{xy} derived for each allele are combined into a single estimate for the locus by using weights that assume zero relationship (as with the weighted regression estimators derived above),

$$\hat{r}_{xy} = \frac{2}{n - 1}\left[\left(\sum_{i=1}^{n} S_i p_i\right) - 1\right]. \tag{9}$$

[Note that the r_{xy} in this article is twice that defined in the Ritland (1996b) article.]

A simpler estimator, also based upon the joint distribution of genotypes, was described by Ritland (1996b) and earlier workers (Li and Horvitz 1953; Weir 1996, Equation 2.28), primarily in relation to estimating inbreeding coefficients. Defining an alternative similarity index such that $S'_{xy} = 1$ when ($x = ii, y = ii$), $S'_{xy} = 0.5$ when ($x = ij, y = ij$) or ($x = ii, y = ij$), $S'_{xy} = 0.25$ when ($x = ij, y = ik$), and $S'_{xy} = 0$ when ($x = ij, y = kl$), then

$$\hat{r}_{xy} = \frac{2(S'_{xy} - J_0)}{1 - J_0}, \qquad (10)$$

where $J_0 = \sum_{i=1}^{n} p_i^2$ is the expected homozygosity at the locus. Equation 10 is equivalent to an unweighted correlation estimator. Because our analyses showed it to be uniformly worse in terms of sampling variance than all of the estimators presented here, we do not consider it any further.

Finally, we note Queller and Goodnight's (1989) estimator of r_{xy}. Although their index is primarily designed for estimating the average degree of relatedness within groups of individuals, it can be expressed in terms of the same parameters that we employ with our Equations 5a and 5b to obtain a pairwise estimator for individuals x and y,

$$\hat{r}_{xy} = \frac{0.5(S_{ac} + S_{ad} + S_{bc} + S_{bd}) - p_a - p_b}{1 + S_{ab} - p_a - p_b}. \qquad (11)$$

This equation has limited utility with diallelic loci—if individual x is a heterozygote, then $S_{ab} = 0$ and Equation 11 is undefined because $p_a + p_b = 1$. Therefore, in the following analyses, we consider Equation 11 only in the context of multiallelic loci.

In comparing the performance of these alternative methods for estimating r_{xy} to that of the regression estimator, we evaluated their single-locus sampling variances analytically by considering the joint probabilities of all genotypes of pairs of individuals, conditional on the degree of relationship and the allele-frequency distribution. With these alternative methods, the weights depend only on the allele-frequency distribution in the population, not on the genotypes of the reference and proband individuals. Thus, with multiple marker loci all with the same allele frequencies, the multilocus sampling variances are simply the single-locus values divided by the number of loci. When loci have different allele-frequency distributions, as is usually the case in practice, weighted multilocus estimates can be obtained by weighting the locus-specific estimates by the inverses of their sampling variance.

For diallelic loci, the correlation estimator yields a sampling variance per locus equal to one in the case of nonrelatives regardless of the allele frequency (Figure 3). As noted above, the regression estimator asymptotically approaches this same level of efficiency for nonrelatives, but the similarity-index method has higher sampling variance. On the other hand, for close relatives, compared to the correlation estimator, the regression and similarity-index methods yield more accurate estimates of r over the full range of allele frequencies at

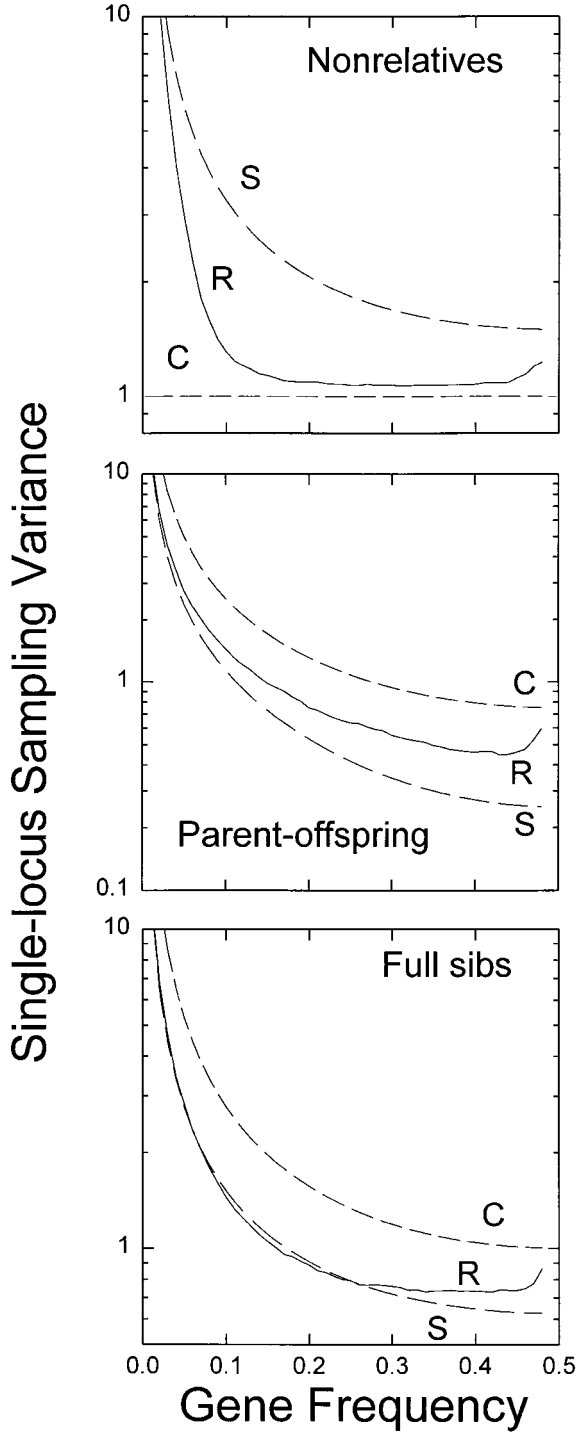

Figure 3.—Single-locus sampling variances for estimates of r derived with the regression method (R), the correlation method (C), and the similarity-index method (S) for diallelic loci. The results for the regression method apply to analyses based on 10 loci and were obtained by Monte Carlo simulations; additional loci yield slightly lower values. The results for the correlation and similarity-index methods are exact solutions based on expected genotype combinations.

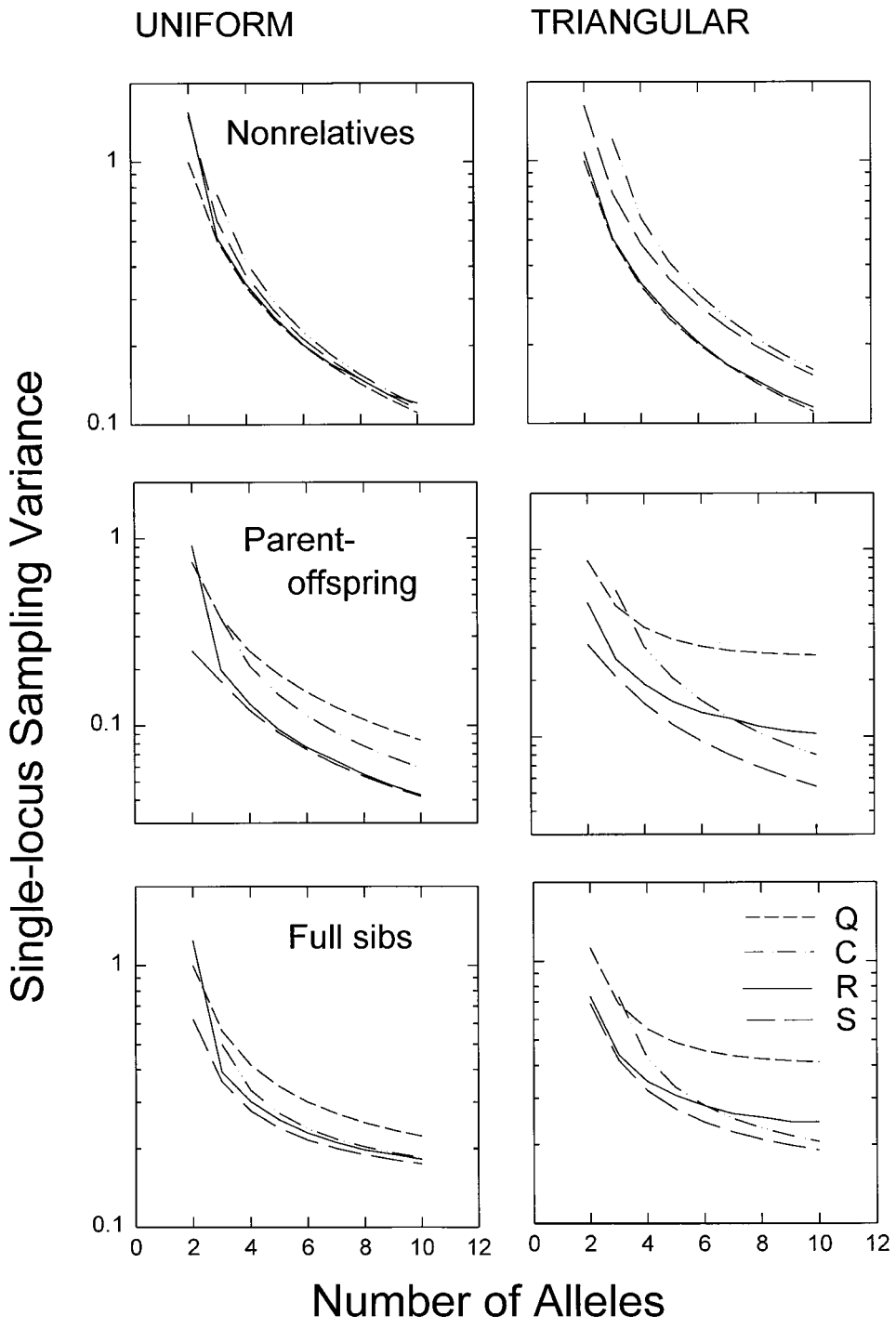

Figure 4.—Single-locus sampling variances for estimates of r for multiallelic loci, derived with the regression method (R), the correlation method (C), the similarity-index method (S), and the Queller-Goodnight method (Q) for uniform and triangular allele-frequency distrubutions. The results for the regression method apply to analyses based on 10 loci and were obtained by Monte Carlo simulations; additional loci yield slightly lower values. The results for the correlation and the similarity-index methods are exact solutions based on expected genotype combinations.

diallelic loci, with the latter actually outperforming the former in the case of parent-offspring pairs.

A multiallelic perspective yields further insight into the relative efficiencies of the four techniques. With a uniform distribution of three or more alleles per locus, the single-locus sampling variance for $\hat{r}$ is essentially $1/(n-1)$ with nonrelatives regardless of the method (Figure 4). Thus, because an even allele-frequency distribution provides the greatest power of inference, this seems to be the best that one can expect to achieve with any estimator of distant relationships. For related individuals, the regression and similarity-index methods yield very similar sampling variances of r provided there are at least three alleles per locus, while the correlation and Queller-Goodnight estimators are again less efficient. For the two superior methods, the single-locus sampling variance of estimates of $\hat{r}$ asymptotically approaches 0.14 with increasing allele number with full sibs, and very slowly approaches 0 with parents and offspring.

TABLE 1

Sampling variance properties of $\hat{\Delta}$

Relationship	Method	Number of alleles			
		2	4	6	12
Uniform frequencies					
Nonrelatives	R	0.999	0.168	0.067	0.015
	C	1.000	0.166	0.067	0.017
Half sibs	R	1.011	0.269	0.142	0.056
	C	1.004	0.272	0.144	0.056
Full sibs	R	0.949	0.423	0.324	0.248
	C	0.948	0.440	0.336	0.256
Parent-offspring	R	0.989	0.368	0.219	0.096
	C	1.008	0.376	0.220	0.096
Triangular frequencies					
Nonrelatives	R	1.070	0.182	0.074	0.016
	C	1.000	0.166	0.067	0.017
Half sibs	R	1.276	0.329	0.179	0.074
	C	1.240	0.360	0.240	0.080
Full sibs	R	1.362	0.605	0.486	0.396
	C	1.480	1.000	0.960	0.880
Parent-offspring	R	1.471	0.479	0.294	0.136
	C	1.520	0.640	0.640	0.280

Values are given for the single-locus sampling variances. R and C denote the regression and correlation estimators, respectively. The regression estimates are based on Monte Carlo simulations of 10 loci per pair of individuals.

With a triangular allele-frequency distribution, the regression and correlation methods again yield essentially identical results with nonrelatives, while the similarity-index and Queller-Goodnight methods have somewhat higher sampling variances. However, with related individuals, the similarity-index method is again the superior of the four methods, and the correlation and Queller-Goodnight estimators generally yield the highest sampling variance. By use of either the regression or similarity-index methods, up to a 50% reduction in the standard error of $\hat{r}$ an be achieved.

The only other marker-based method for the estimation of Δ is the correlation-based estimator of Ritland (1996b), which is quite complex algebraically. Results in Table 1 show that the much simpler regression estimator presented above yields essentially the same asymptotic sampling variances as the correlation method when the allele-frequency distribution is uniform. With triangular allele-frequency distributions, the results are also very similar for nonrelatives, but with related individuals, the regression estimator yields more precise estimates, with the reduction in sampling variance approaching 50% with close relatives.

Thompson (1975, 1986) has extensively investigated the use of maximum likelihood for inferring pairwise relationship. The likelihood method allows one to take an entirely different approach for genealogical inference. For example, Thompson discusses the power of likelihood to distinguish among major types of relationships, defined as family (parent-offspring, full sibs), close (half sibs, uncle, etc.), remote (cousin, etc.), and unrelated. This approach to inferring genealogical "relationship" is fundamentally different from our approach to estimating "relatedness," which is a nondiscrete numerical parameter defined in terms of probabilities of identity-by-descent. Nevertheless, we have considered the possibility of using likelihood methods to estimate "relatedness" under our regression framework. Using notation developed earlier, the likelihood of data from one locus is the probability

$$\begin{aligned} P(y = cd|x = ab) = {} & p_a p_b (2 - S_{ab})(2 - S_{cd}) \\ & \cdot [(1 - 2\phi_{xy} + \Delta_{xy}) p_c p_d + 2(\phi_{xy} - \Delta_{xy}) \\ & \cdot ((S_{ac} + S_{bc}) p_d + (S_{ad} + S_{bd}) p_c)/4 \\ & + \Delta_{xy}(S_{ac} S_{bd} + S_{ad} S_{bc})/2] \end{aligned} \quad (12)$$

and the multilocus likelihood is the product of Equation 12 over loci. This expression can be used for estimating relatedness by solving for the values of r_{xy} and Δ_{xy} that maximize Equation 12, given the data.

Using computer simulations, we examined the behavior of such maximum-likelihood estimation of relatedness by a standard numerical method (Newton-Raphson iteration). Convergence to a maximum was confirmed both by noting that the likelihood increased over iterations and converged and by comparing the iterative solutions to likelihood functions of the same data mapped

by brute force. The results, and those discussed by Ritland (1996b), suggest that the potential for using maximum likelihood for estimating relatedness is limited. The problem is fundamentally due to the fact that the ideal properties of likelihood are asymptotic or apply to "large" sample sizes. The number of loci usually available for pairwise estimation is inherently small—too small for likelihood to avoid substantial problems with bias (usually negative) and extremely large sampling variance. For example, for the case of zero true relatedness, the average estimate of r_{xy} is on the order of -1.0 or less when 40 or fewer loci are sampled, and the sampling variance is two to three orders of magnitude beyond that shown for the alternative estimators in Figures 3 and 4. Interestingly, we found that there is an approximate sample size (number of loci) above which the maximum-likelihood estimators become "stable" or show approximately the predicted asymptotic variance. However, this sample size is large. For the maximum-likelihood estimator of r_{xy}, at low true relatedness, stability occurs at $\sim$70 diallelic loci ($p = 0.5$). The maximum-likelihood estimator of Δ_{xy} exhibits similar behavior, although it begins to stabilize when $\sim$30 loci have been sampled. Thus, while the maximum-likelihood approach may provide a useful means for comparing alternative degrees of relationship by likelihood-ratio tests, its applicability for estimating pairwise relatedness coefficients appears to be limited unless one has the luxury of a very large number of polymorphic markers.

EXAMPLE APPLICATIONS

As examples of how estimators of pairwise relatedness can be used in population studies and how they behave with actual data, we consider two applications. First, as part of a study of isolation-by-distance and field heritabilities in the common monkeyflower (*Mimulus guttatus*), 300 plants were randomly selected along an 84-m transect through a meadow adjacent to Indian Valley Reservoir in Clear Lake County, California (this was the "meadow" transect of Ritland and Ritland 1996). Extracts were obtained from corollas and assayed for 10 polymorphic isozyme loci. Eight loci were diallelic, 1 was triallelic, and the other had four alleles. Using the regression estimator, relatedness was estimated for pairs of plants separated by up to 4 m (with gene frequencies estimated from the entire sample). The estimates of pairwise relatedness from this dataset show considerable scatter, with some being $>+1$ and many <0 (Figure 5). Such behavior is in accordance with the results presented above, which highlight the large sampling variance expected for estimates based upon relatively few marker loci. Because of this large variance, significant inferences can be made only from groups of pairwise relatedness estimates or from correlations of these estimates with other quantities such as similarity for a quantitative trait (Ritland 1996a). In this particular application, there is a negative regression of relatedness on distance (Figure 5) as expected under isolation-by-distance. Relatedness decreased $\sim$50% over the span from 0 to 4 m, with the average value for adjacent plants being 0.21, nearly the level of relatedness expected between half sibs (0.25).

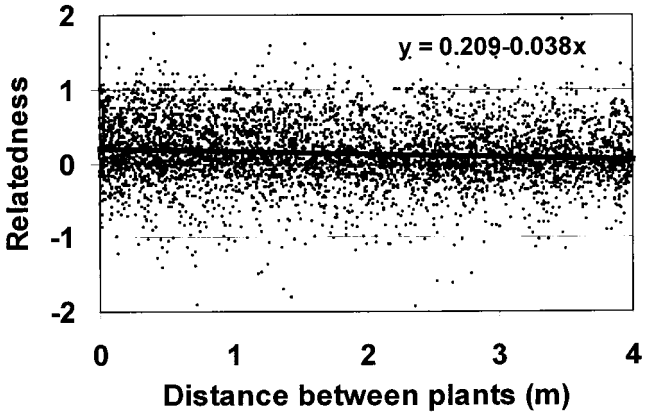

Figure 5.—Estimates of pairwise relatedness in the common monkeyflower plotted as a function of distance. The estimated slope of the linear regression is -0.037/m (0.005) and the estimated intercept is 0.21 (0.01). The standard errors (in parentheses) were obtained by bootstrapping over individuals, with comparisons between identical individuals being excluded.

A second application of relatedness estimates derives from work (D. Marshall and K. Ritland, unpublished results) with a white-phase (termed Kermodism) of the black bear, which is found in low to moderate (10%) frequency along the north coast of British Columbia and adjacent islands. The genetic basis of the coat color polymorphism is unknown. During late summer 1997, nearly 900 bear hair samples were collected from five islands and the adjacent mainland of northern coastal Bristish Columbia. DNA was extracted from hairs with roots and assayed for 8 highly polymorphic microsatellite loci using the primers developed by Paetkau *et al.* (1995). The number of alleles per locus ranged from 7 to 17, with a mean of 10.4, and locus-specific heterozygosities ranged from 0.72 to 0.85, with a mean of 0.79. After factoring out the multiple samples for individual bears, a total of 89 distinct genotypes were found in the regions where Kermodism was of significant frequency (17 on Gribbel Island, 13 on Hawksbury Island, 38 on Princess Royal Island, and 21 at Terrace [mainland BC]). Bear hair color was also recorded in these samples. Estimates of pairwise relatedness were found within each of these four regions, using the pooled samples to estimate gene frequencies. All pairs of individuals were then classified into two groups: pairs sharing coat color (both white or both black, of which there were 614 pairings) and pairs not sharing coat colors (one black, one white, involving 156 pairings). A comparison of the frequency distribution of $\hat{r}$ for these two groups

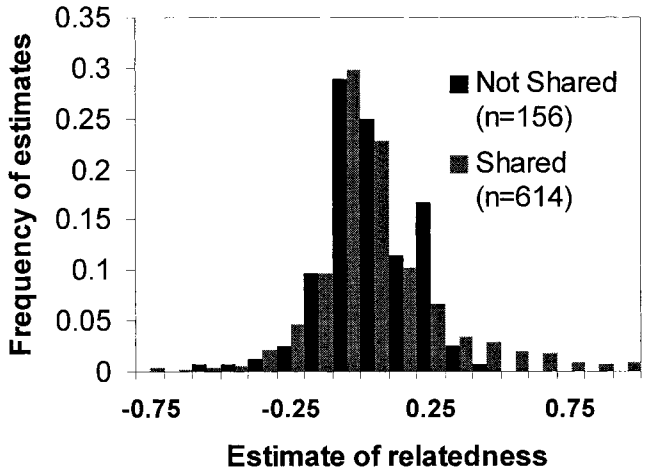

Figure 6.—Distributions of estimates of pairwise relatedness among bears not sharing the same coat color and among bears sharing the same coat color.

(Figure 6) shows an excess of relatedness among bears sharing coat colors ($\bar{r} = 0.057$ compared to 0.039 for unlike colors), suggesting a genetic basis for the variation in this character. However, bootstrap resampling indicated that this difference of means is not significant (the excess being present in only 88 highly variable microsatellite loci, the statistical error of relatedness is considerably less than that experienced with isozyme markers in the previous study). Further inferences about the mode of inheritance of Kermodism are given in Ritland (1999).

DISCUSSION

Estimation of relatedness with molecular markers is a statistically demanding enterprise. On the positive side, all of the estimators described above (except maximum-likelihood) are essentially unbiased in the sense that they return estimates that are on average identical to their expected values. Errors in estimates of population allele frequencies, which were not incorporated into our simulations, can introduce bias, but the effects of error in gene-frequency estimation will generally be trivial (of order $1/N$ when N individuals are censused for gene frequency) compared to the additional sampling errors that arise in the estimation of relatedness, provided the number of individuals sampled exceeds 100 or so (Ritland 1996a,b). Moreover, this source of bias can be simply removed by omitting the pair of interest from the estimate of allele frequency (Queller and Goodnight 1989), although pathological behavior will occur in the rare event that marker alleles are unique to particular individuals, as this would lead to population gene-frequency estimates of zero. In addition, the sampling variance of the relationship coefficients owing to uncertain allele frequencies can, in principle, be obtained by resampling procedures.

The high sampling variance of estimates of relatedness arises in part because of variance in identity-by-descent among loci and in part because of variance in identity-in-state for alleles that are not identical by descent. These sources of sampling error are fundamental consequences of Mendelian segregation, and no amount of statistical finesse can eliminate them. In the actual estimation of relatedness, however, further sampling error is introduced by error in inference. With the regression and correlation estimators, for example, large standard errors result because the estimates of relationship coefficients derived from single loci commonly fall outside of the true domain of (0, 1). Although estimators can be designed to ensure that all estimates lie in the range of true possibilities (*e.g.*, Thompson 1976), all such estimators necessarily return biased estimates, and the magnitude of the bias depends on the actual degree of relationship. Thus, while negative single-locus estimates of relationship coefficients may seem to be an undesirable feature, it is precisely this feature that ensures that the estimators proposed above will be unbiased.

Our results suggest that the relative advantages of the alternative estimators of relatedness depend on several factors. These include the number of loci, the allele-frequency distribution, the degree of actual relationship, and the coefficient estimated (r *vs.* Δ). In general, molecular-marker approaches that yield many alleles and loci tend to favor use of the regression estimators proposed in this article over the correlation estimators presented by Ritland (1996b). With small numbers of diallelic loci with extreme allele frequencies, the correlation method is more efficient than the regression method, but the regression estimators are more efficient in almost all other cases. In addition, the simplicity of the regression estimators lends to easier programming and more stability of estimates under uneven allele frequency distributions. The simplicity of the regression-based approach is underscored by our ability to obtain an analytical solution for $\hat{\Delta}$ with this method. By contrast, the correlation approach of Ritland (1996b) requires, for a locus with n alleles, the inversion of a matrix of size $n(n + 5)/2$, which is 12×12 at the minimum with multiallelic loci and beyond analytical solution. Moreover, unlike the correlation estimator for Δ, the regression estimator for this coefficient is well behaved over the full range of allele frequencies.

As noted above, some simple statements can be made concerning the minimum sampling variance that one can expect to achieve in the estimation of relationship coefficients. For pairs of unrelated or distantly related individuals assayed at L loci, each containing n alleles, the standard errors of the estimates of ϕ (details leading up to this result are not shown), Δ, and r will be no less than $2\sqrt{(n + 4)/[Ln(n - 1)]}$, $\sqrt{2/[Ln(n - 1)]}$, and $\sqrt{1/[L(n - 1)]}$, respectively. For diallelic loci, a com-

mon situation with allozymes, these limits take on values of $3.5/\sqrt{L}$, $1/\sqrt{L}$, and $1/\sqrt{L}$. With large numbers of alleles, as can be achieved with microsatellite loci, the limits asymptotically approach $\sqrt{4/Ln}$, $\sqrt{2/Ln^2}$, and $\sqrt{1/Ln}$. Fortunately, the two coefficients with the lowest sampling error, r and Δ, are the ones that have the greatest practical utility.

One of the limitations of both the regression and correlation methods for estimating relatedness is the use of weights that assume zero relationship. The best weights are a function of the actual relationship, but this is an unknown. Nevertheless, the use of approximate but incorrect weights yields more precise estimates than the use of unweighted estimators, because differences in the information content of alleles with different frequencies are at least partially taken into account. One might think that estimates obtained with the null weights could be improved upon by subsequently refining the weights, using the previous estimates of relatedness in the calculation of the weights. These revised weights could then give a second round of weighted estimates, and the whole process could be repeated again until a suitable degree of convergence to final estimates is achieved. However, simulations by us and by Ritland (1996b) indicated that, even with large numbers of loci, this iterative approach has little promise. Bias is introduced, and with the weights being as noisy as they are, the weights themselves are often wildly unrealistic.

Generally speaking, our results show that attempts to estimate relatedness with molecular markers can be greatly improved upon by working with multiallelic loci, with the most dramatic gains in efficiency occurring with loci with relatively even distributions of allele frequencies. Because the sampling variance of $\hat{r}$ is inversely proportional to Ln, it is clear that roughly the same amount of efficiency is gained by working with loci with twice the number of alleles as by doubling the number of loci. For Δ, the sampling variance is inversely proportional to Ln^2, so a much greater gain can be achieved by increasing numbers of alleles as opposed to numbers of loci. Thus, an early investment in a search for informative loci (those with a large number of alleles, with roughly equal frequencies) can be quite advantageous in the long term. These recommendations assume that at least 10 or so loci are sampled, because with fewer loci, the tradeoff involving r favors more loci over more alleles per locus.

The results presented above indicate that even with fairly large numbers of loci, standard errors of relationship coefficients will rarely be $<0.1/\sqrt{L}$ and often will be somewhat $>1/\sqrt{L}$, so in general one cannot expect to use markers to make precise statements about differences in relatedness between particular pairs of individuals. However, with enough effort applied to the right kinds of loci, it may be possible to reduce the sampling variance to the extent that Ritland's (1996a) quantitative-genetic technique can be applied to natural populations. Ritland's (1996a) method provides a means of estimating the additive and dominance components of genetic variance for quantitative traits (and covariance between traits) in the field by regressing measures of phenotypic similarity on the relatedness coefficients $\hat{r}$ and $\hat{\Delta}$. Aside from the physical labor involved, one of the greatest difficulties with this technique is the need to eliminate the sampling variance from the total observed variance of relatedness to estimate the actual variance in relatedness. The problem is by no means trivial as can be seen in Ritland and Ritland's (1996) first application of the technique with the monkeyflower (Mimulus). With eight assayed loci, the estimates of r derived by the correlation method ranged from -3 to $+5$, with approximately a third of all observed values being negative. The actual variance of relatedness was estimated to be on the order of only 0.04. Thus, almost all of the observed variance in $\hat{r}$ was due to sampling error. Such results clearly highlight the practical need for molecular and statistical methodologies for minimizing the sampling variance of relatedness.

We thank John Kelley for helpful comments. This work was supported by National Institutes of Health grant GM-36827 and National Science Foundation grant DEB-9629775 to M.L., and by a National Sciences and Engineering Research Council / Industry Research Chair in population genetics held by K.R.

LITERATURE CITED

Avise, J. C., 1995 *Molecular Markers, Natural History and Evolution*. Chapman and Hall, New York.

Cockerham, C. C., 1971 Higher order probability functions of identity of alleles by descent. Genetics **69:** 235–246.

Falconer, D. S., and T. F. C. Mackay, 1996 *Introduction to Quantitative Genetics*, Ed. 4. Longman, Harlow, United Kingdom.

Geyer, C. J., and E. A. Thompson, 1995 A new approach to the joint estimation of relationship from DNA fingerprint data, pp. 245–260 in *Population Management for Survival and Recovery*, edited by J. D. Ballou, M. Gilpin and T. J. Foose. Columbia University Press, New York.

Hamilton, W. D., 1964 The genetical evolution of social behaviour: I and II. J. Theor. Biol. **7:** 1–52.

Jacquard, A., 1974 *The Genetic Structure of Populations*. Springer, Berlin.

Kempthorne, O., 1954 The correlation between relatives in a random mating population. Proc. R. Soc. Lond. Ser. B **143:** 103–113.

Li, C. C., and D. G. Horvitz, 1953 Some methods of estimating the inbreeding coefficient. Am. J. Hum. Genet. **5:** 107–117.

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

Lynch, M., 1988 Estimation of relatedness by DNA fingerprinting. Mol. Biol. Evol. **5:** 584–599.

Lynch, M., and B. Milligan, 1994 Analysis of population genetic structure with RAPD markers. Mol. Ecol. **3:** 91–99.

Lynch, M., and J. B. Walsh, 1998 *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Sunderland, MA.

Paetkau, D., W. Calvert, I. Stirling and C. Strobeck, 1995 Microsatellite analysis of population structure in Canadian polar bears. Mol. Ecol. **4:** 347–354.

Pamilo, P., and R. H. Crozier, 1982 Measuring genetic relatedness in natural populations: methodology. Theor. Popul. Biol. **21:** 171–193.

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

Ritland, K., 1989 Marker genes and the inference of quantitative genetic parameters in the field, pp. 183–201 in *Population Genetics, Plant Breeding and Gene Conservation*, edited by A. H. D. Brown, M. T. Clegg, A. L. Kahler and B. S. Weir. Sinauer Associates, Sunderland, MA.

Ritland, K., 1996a A marker-based method for inferences about quantitative inheritance in natural populations. Evolution **50:** 1062–1073.

Ritland, K., 1996b Estimators for pairwise relatedness and inbreeding coefficients. Genet. Res. **67:** 175–186.

Ritland, K., 1999 Detecting inheritance with inferred relatedness in nature, in *Adaptive Genetic Variation in the Wild*, edited by T. Mousseau. Oxford University Press, Oxford (in press).

Ritland, K., and C. Ritland, 1996 Inferences about quantitative inheritance based on natural population structure in the yellow monkeyflower, *Mimulus guttatus*. Evolution **50:** 1074–1082.

Thompson, E. A., 1975 The estimation of pairwise relationships. Ann. Hum. Genet. **39:** 173–188.

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

Thompson, E. A., 1986 *Pedigree Analysis in Human Genetics*. The Johns Hopkins University Press, Baltimore.

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

Communicating editor: A. H. D. Brown

APPENDIX

Provided there are only two alleles at the locus in the population, the approach provided in the text for a homozygous reference genotype can also be applied to the case in which the reference genotype is a heterozygote for alleles i and j. The conditional probabilities of observing proband genotypes, given a heterozygous reference genotype, are

$$P(ii|ij) = p_i^2 + p_i(0.5 - p_i)\phi_{xy} - p_i^2\Delta_{xy} \quad \text{(A1a)}$$

$$P(jj|ij) = p_j^2 + p_j(0.5 - p_j)\phi_{xy} - p_j^2\Delta_{xy}\,. \quad \text{(A1b)}$$

The third probability, $P(ij|ij)$, is omitted, as only two of the three probabilities are needed for a sufficient statistic because the three probabilities sum to unity.

Equating these probabilities to their estimates and rearranging, estimators for the coefficients of relationship are obtained as

$$\hat{\phi}_{xy} = \frac{2[q^2\hat{P}(ii|ij) - p^2\hat{P}(jj|ij)]}{pq(q-p)} \quad \text{(A2a)}$$

$$\hat{\Delta}_{xy} = 1 - \frac{\hat{P}(ii|ij)}{p} - \frac{\hat{P}(jj|ij)}{q}\,, \quad \text{(A2b)}$$

wherein, to emphasize that these equations apply only to diallelic loci, we have dropped the subscript i, letting $p = p_i$ and $q = 1 - p$. From Equation 1,

$$\hat{r}_{xy} = 1 + \frac{\hat{P}(ii|ij) - \hat{P}(jj|ij)}{(q-p)}\,. \quad \text{(A2c)}$$

When gene frequencies are exactly equal, a reference heterozygote at a diallelic locus yields undefined estimates for ϕ_{xy} and r_{xy}.

If there are more than two alleles in the population, there are six possible proband genotype categories conditioned on observing the heterozygous reference genotype ij. The conditional probabilities include $P(ii|ij)$ and $P(jj|ij)$ as given in Equations A1a and A1b plus four more:

$$P(ij|ij) = 2p_ip_j + [0.5(p_i + p_j) - 2p_ip_j] \cdot \phi_{xy} - (1 - 2p_ip_j)\Delta_{xy} \quad \text{(A3a)}$$

$$P(i\cdot|ij) = 2p_i(1 - p_i - p_j) + (1 - p_i - p_j)(0.5 - 2p_i) \cdot \phi_{xy} - 2p_i(1 - p_i - p_j)\Delta_{xy} \quad \text{(A3b)}$$

$$P(j\cdot|ij) = 2p_j(1 - p_i - p_j) + (1 - p_i - p_j)(0.5 - 2p_j) \cdot \phi_{xy} - 2p_j(1 - p_i - p_j)\Delta_{xy} \quad \text{(A3c)}$$

$$P(\cdot\cdot|ij) = (1 - p_i - p_j)^2(1 - \phi_{xy} - \Delta_{xy})\,. \quad \text{(A3d)}$$

Thus, with multiallelic loci, heterozygous reference individuals generate the obvious difficulty of there being more equations than unknowns.

Linear regression provides a data-fitting procedure for obtaining estimators for ϕ_{xy}, Δ_{xy}, and r_{xy} in this case. The six probabilities can be assembled into an array,

$$\mathbf{P} = \begin{pmatrix} P(ii|ij) \\ P(jj|ij) \\ P(ij|ij) \\ P(i\cdot|ij) \\ P(j\cdot|ij) \\ P(\cdot\cdot|ij) \end{pmatrix}.$$

For any pair of individuals, the observed data vector ($\hat{\mathbf{P}}$) will always contain a single one for the observed two-genotype combination with all other elements being equal to zero. The linear model then becomes

$$\hat{\mathbf{P}} = \mathrm{a} + \mathrm{M}_x \begin{pmatrix} \phi_{xy} \\ \Delta_{xy} \end{pmatrix} + \mathrm{e}, \quad \text{(A4)}$$

where the matrix $\mathbf{M}_x$ has two columns that contain the coefficients for ϕ_{xy} and Δ_{xy}, respectively, **a** is a column vector containing the remaining constants (functions only of gene frequencies), and **e** is a vector of residuals with expectation zero. The elements of $\mathbf{M}_x$ and **a** are obtained directly from Equations A1a and A1b and A3a–A3d.

If the elements of the observation vector $\hat{\mathbf{P}}$ were independent and identical in distribution, ordinary least-squares analysis could be used to obtain estimates of the relationship coefficients with minimum sampling variance. However, because all of the elements of the observation vector are constrained to sum to 1, such conditions are obviously violated. Although the failure to fully account for the structure of the data in the **P** vector does not cause the estimates of the coefficients of relationship to be biased, it does elevate the sampling variance. Unfortunately, the variance-covariance structure necessary to generate the optimal weights for a more powerful generalized least-squares framework depends on the unknown parameters ϕ_{xy} and Δ_{xy}. To obtain approximate weights, we rely on Ritland's (1996b) argument that, in the absence of prior information on

the relationship of x and y, it is reasonable to start with the assumption that $\phi_{xy} = \Delta_{xy} = 0$.

Using the optimal weights given by Equation 4b of Ritland (1996b), we were able to obtain analytical solutions for the weighted least-squares estimators of ϕ_{xy} and Δ_{xy} using an equation solver program. These are

$$\hat{\phi}_{xy} = \frac{4p_ip_j(1 - p_i - p_j)[1 - \hat{P}(ij|ij)] - 2(1 - 2p_ip_j)[p_j\hat{P}(i|ij) + p_i\hat{P}(j|ij)]}{(1 - p_i - p_j + 2p_ip_j)(4p_ip_j - p_i - p_j)} \quad \text{(A5a)}$$

$$\hat{\Delta}_{xy} = \frac{(1 - p_i - p_j)\hat{P}(ij|ij) - p_j\hat{P}(i|ij) - p_i\hat{P}(j|ij) + 2p_ip_j}{1 - p_i - p_j + 2p_ip_j}, \quad \text{(A5b)}$$

and from Equation 1

$$\hat{r}_{xy} = \frac{p_j\hat{P}(i|ij) + p_i\hat{P}(j|ij) + (p_i + p_j)\hat{P}(ij|ij) - 4p_ip_j}{p_i + p_j - 4p_ip_j}.$$

where $\hat{P}(i|ij) = \hat{P}(i\cdot|ij) + 2\hat{P}(ii|ij)$ and $\hat{P}(j|ij) = \hat{P}(j\cdot|ij) + 2\hat{P}(jj|ij)$. When there are only two alleles, Equations A5a–A5c reduce to the diallelic-locus estimates (A2a–A2c).