Genet. Res., Camb. (1996), **67**, *pp.* 175–185 *With 4 text-figures*

Estimators for pairwise relatedness and individual inbreeding coefficients

KERMIT RITLAND
Department of Botany, University of Toronto, Toronto, Ontario M5S 3B2 Canada

(*Received 22 December 1994 and in revised form 30 October 1995*)

Summary

Method-of-moments estimators (MMEs) for the two-gene coefficients of relationship and inbreeding, and for the four-gene Cotterman coefficients, are described. These estimators, which use co-dominant genetic markers, are most appropriate for estimating pairwise relatedness or individual inbreeding coefficients, as opposed to their mean values in a group. This is because, compared to the maximum likelihood estimate (MLE), they show reduced small-sample bias and lack distributional assumptions. The 'efficient' MME is an optimally weighted average of estimates given by each allele at each locus. Generally, weights must be computed numerically, but if true coefficients are assumed zero, simplified estimators are obtained whose relative efficiencies are quite high. Population gene frequency is assumed to be assayed in a larger, 'reference population' sample, and the biases introduced by small reference samples and/or genetic drift of the reference population are discussed. Individual-level estimates of relatedness or inbreeding, while displaying high variance, are useful in several applications as a covariate in population studies.

1. Introduction

Relationship, in its most general sense, involves the identity-by-descent of homologous alleles either between individuals ('relatedness') or within individuals ('inbreeding'). If the pedigree of individuals is unknown, relationship may be estimated with genetic markers (Li & Horvitz, 1953; Morton *et al.* 1971; Thompson, 1975; Robertson & Hill, 1984; Lynch, 1988; Queller & Goodnight, 1989). Such inferences are useful in studies of kin selection (Michod & Hamilton, 1980; Grafen, 1985; Schuster & Mitton, 1991), isolation by distance (Wright, 1943; Malecot, 1969), spatial autocorrelation (Barbujani, 1987; Epperson, 1989), paternity analysis (Chakraborty *et al.* 1988), and mating systems (Brown *et al.* 1985).

A large body of literature has accumulated on the estimation of the average inbreeding coefficient or coefficient of relatedness (see Li & Horvitz, 1953; Yasuda, 1968; Wright, 1969; Crow & Kimura, 1970; Curie-Cohen, 1981; Robertson & Hill, 1984; Weir, 1990). Despite its utility in several areas of ecology and evolution, the methodology of using markers to estimate relatedness or inbreeding coefficients *for individuals* has received relatively little attention, being primarily represented by the works of Thompson (1975), Lynch (1988) and Queller & Goodnight (1989).

One problem in estimating relatedness or inbreeding for individuals is statistical bias caused by small samples. In considering sample size, there are two dimensions: the number of individuals, and the number of marker loci. With individual-level estimates, the number of individuals is at a bare minimum (one for inbreeding, two for relatedness), magnifying the bias due to small samples, even when a large number of marker loci are used. This can be a significant problem when using maximum likelihood estimators (MLEs), which are recognized to often show bias with small sample sizes (Curie-Cohen, 1981; Weir, 1990). Generally, estimators of relatedness have been designed only for larger sample sizes (Wilkinson & McCraken, 1986).

A second problem posed by individual-level estimates is a large error of inference. This problem was studied by Thompson (1975), who tested alternative hypotheses of relationship using maximum likelihood. She found that, even with 20 highly polymorphic loci, the data could not often distinguish among the major classes of relatives. Likewise, human geneticists have long recognized the large sample sizes needed to detect the small levels of inbreeding characteristic of human populations (e.g. Emigh, 1980).

Thus, for individual-level estimates to be useful, the object of study should transcend a single pairwise

relationship, and involve a set of relationships or inbreeding coefficients in conjunction with other data. The increasing application of genetic markers in population studies of wild species (Cruzan, 1996) warrants examination of individual-level estimators of relatedness or inbreeding. Individual-level estimators are useful in inferences involving: (1) the covariation of relationship with physical distance (in the study of isolation by distance, Loeselle *et al.* 1996) or with the similarity for a quantitative character (for quantitative genetic inheritance, Ritland, 1996*a*); (2) actual variance of relatedness or inbreeding (for describing population structure, Ritland, 1996*b*); (3) average relatedness within a group (Schuster & Mitton, 1991) or (4) the covariance of inbreeding level with fitness.

These applications also introduce a third difficulty with individual-level estimates: in natural populations, the distribution of relatedness or inbreeding among individuals is unknown, making the method of maximum likelihood, with its requirement to assume a certain distribution of relatedness, of further uncertain value.

In this paper, I describe method-of-moments estimators (MMEs) for the pairwise relatedness between outbred individuals, and for the inbreeding coefficient of an individual. The relatedness coefficients include the classical 'two-gene' coefficient as well as a four-gene coefficient, which is needed to completely specify relatedness between two outbred individuals. No MMEs have been described for the four-gene coefficient, and efficient MMEs have previously been described with reference only to estimating the mean inbreeding coefficient in mildly inbred populations (Robertson & Hill, 1984). The efficiency and bias of these MMEs are evaluated with analytical formulae and Monte-Carlo simulations, and compared to MLEs. The primary advantages of the MME over the MLE is the reduction in bias with individual-level estimates and a lack of distributional assumptions.

In the weighted, efficient MME procedure, estimates are found separately for each allele at each locus, then linearly combined into a single estimate using optimized weights. This optimization gives an estimator that efficiently utilizes the differing information provided by alleles of differing frequency, and the linearity reduces small sample size bias. Thus, the MME is ideal for individual-level estimates. However, in obtaining weights, relatedness must be guessed and it is assumed that population gene frequency has been estimated from a large sample, such that its sampling variance is negligible relative to the estimate of relationship. If this is not true (such as when estimating mean relatedness or inbreeding), maximum likelihood, with its capacity for joint estimation, may be preferable.

2. Estimators for two-gene relatedness and inbreeding coefficients

The definition of relationship as estimated from genetic markers has been somewhat ambiguous, reflecting the historical development of ideas about relationship and its estimation. Relationship can be defined in terms of the correlation or regression of alleles (Wright, 1922; Pamilo & Crozier, 1982; Queller & Goodnight, 1989), or by the probabilities of identity-by-descent of alleles (Malecot, 1969; Jacquard, 1974; Thompson, 1976). In this paper, we adopt the gene-identity definition because it easily allows the more complex, four-gene treatment of relatedness.

The fundamental measure of relatedness between two individuals is the 'coefficient of kinship' between two individuals A and B. This quantity, denoted r, is the probability that two alleles, one randomly sampled from each individual, are identical-by-descent (Jacquard, 1974). It is a 'two-gene' coefficient because of this dependence upon pairs of sampled genes. This quantity r is alternatively defined as the correlation between the additive values of the two individuals (Crow & Kimura, 1970). Wright's (1922) coefficient of relationship, defined as the correlation between relatives for additive effects of genes, equals $2r$ for outbred relatives. This coefficient increases with the level of relationship; in outbred populations, $r = 1/4$ for parent–offspring, $r = 1/4$ for full-sibs, $r = 1/8$ for half-sibs, and $r = 1/16$ for first-cousins (see Jacquard, 1974).

The inbreeding coefficient f (also due to Wright, 1922) is also a two-gene measure, and it is analogous to the two-gene coefficient of relationship, except that the two sampled genes are represented by both alleles at a diploid locus within a single individual. Both r and f are herein defined relative to a single population, e.g. we are not concerned with how variation of gene frequency (or Wright's F_{ST}) contributes to relationship.

For estimation of relatedness, the 'unit of observation' is a pair of diploid genotypes, or 'pairwise genotypes'. For inbreeding, the unit is the individual. For all procedures given below, we assume independence or near-independence between pairs of individuals (for relatedness) or between individuals (for inbreeding). This is not true for those populations in which groups of individuals share the same parentage, but such violations are often unavoidable, and besides, cause only a marginal increase in the variance of estimates (Robertson & Hill, 1984).

To describe the data, we denote S_i as the observed proportion of pairs similar for marker allele i. It can be regarded as an 'indicator' variable of relationship. For the case of the inbreeding coefficient f, then $S_i = 1$ if the two alleles at a locus are allele i; otherwise $S_i = 0$. For the case of two-gene relatedness r, there are four equally probable ways of sampling two alleles, one for each of two relatives. S_i is the average over the four ways that a pair of alleles can be sampled. For

https://doi.org/10.1017/S0016672300033620 Published online by Cambridge University Press

example, for pairwise genotypes $A_iA_j - A_iA_j$ the observed $S_i = 1/2$, and for $A_iA_j - A_iA_k$ the observed $S_i = 1/4$.

The expectation of S_i (denoted s_i), conditioned upon relationship, is

$$s_i = \rho p_i + (1-\rho)p_i^2, \tag{1}$$

where ρ is the two-gene relationship, which equals either r or f. This similarity is a mixture of the probability of identity-by-descent (ρp_i) and the non-identity-by-descent, $(1-\rho)p_i^2$. The latter is termed 'identity-by-state'. This expectation assumes the population gene frequencies equal the pedigree gene frequencies (the gene pool from which alleles were randomly drawn during the formation of the pedigree). The probabilities over several independent loci are the products of these single-locus probabilities.

(i) *Efficient method-of-moments estimators*

To obtain an efficient method-of-moments estimator (MME) for two-gene relationship (generally, relatedness or inbreeding), one first obtains estimates for each marker allele i, for $i = 1$ to n (the number of alleles at the locus), based upon the observation of whether the pair of alleles are both of type i or not. Although there are $n(n+1)/2$ combinations of alleles, each of which can give an estimate of relationship, these estimates are not independent, and only the set of n estimates corresponding to the sharing of allele i, $i = 1, n$, are sufficient to capture all information in the data (Robertson & Hill, 1984). The variance–covariance matrix of these n estimates is then used to optimally combine the n estimates in a linear fashion into a single estimate. Estimates are likewise averaged over loci. Because of mathematical complexity and the unique nature of individual-level estimators, we derive the estimator assuming known reference population gene frequencies. This assumption is further discussed below.

By equating observed quantities to their expectations in (1), we obtain these estimators for each allele i at an n allele locus as

$$\hat{\rho}_i = \frac{S_i - P_i^2}{P_i Q_i}, \quad i = 1, \ldots, n, \tag{2}$$

where $P_i = 1 - Q_i$ is the estimate of gene frequency p_i (capital letters are used to denote estimated quantities), and the hat denotes the estimate. For simplicity, gene frequency can be estimated by counting alleles in the entire sampled population (this assumes low mean relationship; see below).

The total estimate of relationship (relatedness or inbreeding) is then the weighted average,

$$\hat{\rho} = \sum_i w_i \hat{\rho}_i, \tag{3}$$

where the weights w_i sum to unity.

To obtain the optimal weights, note that the n estimates of relationship (2) have variances and covariances

$$\mathrm{Var}(\hat{\rho}_i) = \frac{s_i(1-s_i)}{cp_i^2 q_i^2},$$

$$\mathrm{Cov}(\tilde{\rho}_i, \tilde{\rho}_j) = \frac{-s_i s_j}{cp_i p_j q_i q_j}, \quad i,j = 1,2,\ldots,n. \tag{4a}$$

These are obtained by noting that the S_i are multinomially distributed with variances $s_i(1-s_i)$ and covariances $-s_i s_j$, and that $\mathrm{Var}(aX+b) = a^2\mathrm{Var}(X)$ for a and b constant. The constant $c = 1$ for f, while $c \leqslant 4$ for r (two independent or partially independent similarities are averaged; its exact value is irrelevant because it cancels when computing weights). Note these variances differ from those given by Robertson & Hill (1984), who stated $\mathrm{Var}(\hat{\rho}_i) = 1$ and that the covariances were of opposite sign from above; nevertheless, they obtain an estimator equal to (5) below (but they were generally concerned with finding MLEs and not MMEs). However, numerical simulations confirm (4*a*) are the correct expressions.

The optimal weights are then found via a standard procedure of weighting correlated estimates. Briefly, these weights minimize $\mathrm{Var}(\hat{\rho}) = \mathbf{w}^{\mathrm{T}}\mathbf{V}\mathbf{w}$, where $\mathbf{w}$ is an n element column vector of weights and $\mathbf{V}$ is the variance–covariance matrix of allele-specific estimates. To solve for these weights, the set of $n-1$ equations $d\mathrm{Var}(\hat{\rho})/dw_i = 0$ $(i = 1, \ldots, n-1)$ is solved. This is rewritten as $\mathbf{C}\omega = \mathbf{c}$, where $\mathbf{C}$ is an $(n-1)\times(n-1)$ matrix with *ij–th* element $\mathrm{Cov}(\hat{\rho}_i, \hat{\rho}_j) - \mathrm{Cov}(\hat{\rho}_i, \hat{\rho}_n) - \mathrm{Cov}(\hat{\rho}_j, \hat{\rho}_n) + \mathrm{Var}(\hat{\rho}_n)$, ω is an $n-1$ element column vector containing the first $n-1$ weights, and $\mathbf{c}$ is an n-1 element column vector with ith element $\mathrm{Var}(\hat{\rho}_n) - \mathrm{Cov}(\hat{\rho}_i, \hat{\rho}_n)$. The optimal weights $\mathbf{w}$ are thus

$$\omega = \mathbf{C}^{-1}\mathbf{c}, \tag{4b}$$

with the nth weight being one minus the other weights. Unless one assumes $\rho = 0$ or $\rho = 1$, this expression must be solved numerically.

Multilocus estimates of relatedness involve a second stage of weighting. After a weighted estimate is found for each locus, a 'grand' weighted estimate is found by weighting estimates across loci. If loci are unlinked and in linkage equilibrium, estimates from different loci will be independent, and the weighting used for a given locus is simply proportional to the inverse of its variance, as computed by the above weighting procedure.

(ii) *Cautions*

It should be noted that, for the purpose of tractability and because of the unique nature of individual-level estimates, several simplifications have been made. First, this MME was derived by equating estimated gene frequencies (P) to their true values (p). This

https://doi.org/10.1017/S0016672300033620 Published online by Cambridge University Press

neglects the effect of finite sample size. The major cause of a finite sample is to cause a positive covariance of S_i with P_i. This can be removed by excluding the particular individual(s) from the estimate of p_i (Queller & Goodnight, 1989; also see Fig. 2). In addition, since we assume gene frequency is estimated from a much larger, population sample (Ritland & Ritland, 1996, used a population size of 300), any residual variation of P_i has little effect in terms of the bias/variance ratio of $\hat{\rho}_i$ (see discussion of Fig. 2).

However, even if this source of bias is removed, there remains another source of bias due to the sampling of gametes during population founding. The effect of this is to create groups of relatives in the descendent population, which causes additional bias due to the correlations S_i with other, related individuals which are accidentally included in P_i. If no relatives are excluded from P_i, this bias equals $-\Delta F_{st}$ or minus the increase of inbreeding due to drift. Thus, the average pairwise estimate is zero, regardless of the average level of relationship. For example, in a population of N equal-sized, full-sib families with unrelated parents, there are N groups of relatives, and the average pairwise r will be 0·025 yet the estimated r will be zero. To remove this bias, an unrelated 'reference' or 'outgroup' population can be used (discussed by Ritland, 1996*b*). It may also be possible, through an iterative procedure, to identify putative relatives to exclude from P_i, but we lose the properties of the MME. Given the difficulty of removing this systematic bias, it becomes obvious that we should focus on inferences involving variance and covariance of relationship (quantities not biased by ΔF_{st}), and not mean relationship.

Secondly, this MME approach does not allow for simultaneous estimation of gene frequency. If relatedness or inbreeding are non-zero, gene frequencies should properly be simultaneously estimated. However, unless relationship is high, estimates obtained by simple gene-counting methods are very close to the simultaneous estimate (Robertson & Hill, 1984). If one was to properly estimate gene frequency jointly with relationship, one must somehow consider all higher-order relationships among triplets, quadruplets, etc. of individuals, to account for the correlations of alleles across multiple individuals.

Thirdly, the weighting formula (4) assumes that we know the true values of similarity (denoted as s, in our convention of writing parametric values with lower case). This is a function of the true relationship as well as the true gene frequency. We cannot use observed similarities – this causes a downward bias of estimate due to statistical correlations (alleles showing less similarity by chance receive spurious higher weights; in fact, this is the basis for the problems with individual-level maximum likelihood estimates). Rather, one must use separate information for s values. One method would be to use estimates of relationship based upon all other loci to predict s using (1). Alternatively, one can use the average relationship in the population to specify s via (1). A third possibility is to assume $\rho = 0$ so $s_i = P_i^2$ in (4*a*). This last alternative is considered further below.

In addition, for the case where relatedness r is estimated, the four-gene coefficient h is not simultaneously estimated. This, in effect, ignores the correlations of the four $\hat{r}$ obtained from the four pairings of alleles between two diploid individuals, and was mainly adopted for simplicity at this stage of presentation (it also is a valid assumption when h is near zero); the joint estimation of r and h is given in Section 3.

(iii) *A simple, explicit MME*

A simplified estimator can be obtained by assuming $\rho = 0$ in the weights. The assumption seems good since levels of relatedness or inbreeding are often low in natural populations, and errors of ρ in the weights increase only the variance of the estimate, and not its bias.

At $\rho = 0$, the procedure for obtaining optimal weights gives the weight for allele i as $w_i = q_i/(n-1)$, for n the number of alleles at the locus. This gives an estimator for a single locus, which combines information among alleles, as

$$\hat{\rho} = \sum_i \frac{S_i - P_i^2}{(n-1)P_i}.$$

To combine estimates among loci, we use the fact that at zero true relationship and known gene frequency, the variance of a single-locus weighted MME is proportional to $1/(n-1)$, regardless of the frequency distribution of alleles. The inverse of this quantity serves as the weight. This gives a simplified multilocus estimator of relationship, based upon a prior ρ of zero, as

$$\hat{\rho} = \sum_{i,l} \frac{S_{il} - P_{il}^2}{P_{il}} \Big/ \sum_l (n_l - 1), \tag{5}$$

where l denotes the locus. This estimator was first described by Li & Horvitz (1953), who did not realize this was the minimum variance estimator. Robertson & Hill (1984) explored in greater details these properties, but did not realize the applicability of this estimator to the individual-level inferences.

A second simple method-of-moments estimator for ρ can be obtained by assuming $\rho = 1$ in the weights. The weights then become $p_i q_i/(1-J)$, for J the expected homozygosity. Over N independent loci, this estimator equals

$$\hat{\rho} = \frac{S - J}{1 - J}, \tag{6}$$

for

$$J = \frac{1}{N} \sum_{i,l} P_{il}^2$$

https://doi.org/10.1017/S0016672300033620 Published online by Cambridge University Press

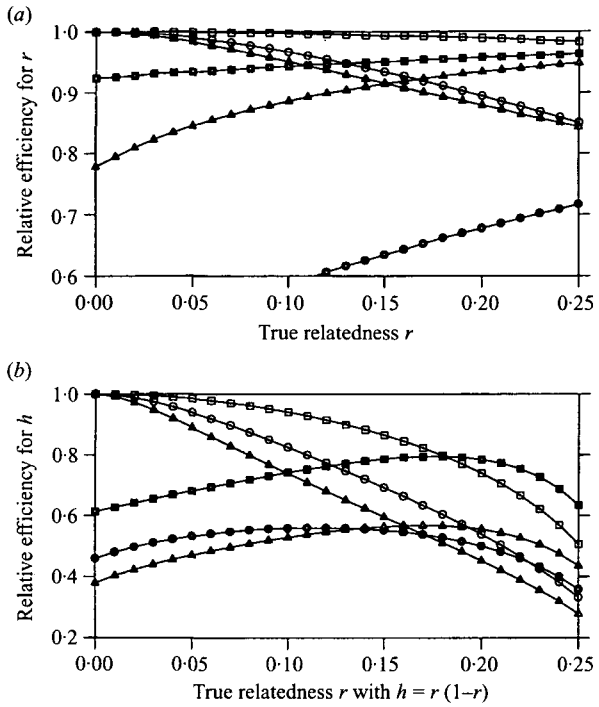

Fig. 1. Relative efficiency of estimators for (*a*) two-gene relationship ρ and (*b*) four-gene relationship h (discussed in Section 3) under different prior relationship in weights and different gene frequencies: ⊡, $P = \{0{\cdot}4, 0{\cdot}6\}$; prior $r = 0$; ■, $P = \{0{\cdot}4, 0{\cdot}6\}$; prior $r = 1$; ⊙, $P = \{0{\cdot}2, 0{\cdot}8\}$; prior $r = 0$; ●, $P = \{0{\cdot}2, 0{\cdot}8\}$; prior $r = 1$; ▲, $P = \{0{\cdot}1, 0{\cdot}2, 0{\cdot}3, 0{\cdot}4\}$; prior $r = 1$; △, $P = \{0{\cdot}1, 0{\cdot}2, 0{\cdot}3, 0{\cdot}4\}$; prior $r = 0$.

the mean expected homozygosity over the N loci and

$$S = \frac{1}{N}\sum_{i,l} S_{il}$$

the arithmetic average of allele similarity between the two individuals across loci.

Figure 1*a* compares the efficiency of the estimators eqns (5) and (6) across values of true relationship, for three distributions of gene frequency. Weights assuming $\rho = 0$ (eqn 5) are much more efficient than weights that assume $\rho = 1$ (eqn 6), showing at least 95% efficiency when true $\rho = 1/8$ (for relatedness r, this is less than full-sibs) except for diallelic loci with more extreme gene frequencies, where the efficiency is as low at 85%. Weir & Cockerham's (1984) estimator for F_{st} uses these weights of eqn (6), implying inefficiency in their procedure. Figure 1*a* also shows a greater efficiency with more even distribution of gene frequency (with uniform distribution of p, weights are 100% efficient).

Thus, the weighted MME with weights assuming $\rho = 0$ generally have at least 85% of the efficiency of the MLE, with greater efficiency obtained by using loci with even gene frequencies, or by a fortuitous nearly correct guess of true relatedness in forming weights. While the MLE has equal or greater efficiency in all cases, efficiency is only one criteria for choosing an estimator. Another criteria is bias, which clearly favours the weighted MME for case of individual-level estimation, as shown below.

One can also use the actual estimate of relationship between the two individuals to form the weights via (5*a–b*). This is guessed at first, but the subsequent estimate of relationship can then be used in new weights, and the procedure iterated until convergence. Numerical solutions indicate that this procedure gives the same estimate as the maximum likelihood procedure. This demonstrates the efficiency of the method-of-moments estimation procedure and a numerical equivalency between an iterated MME and the MLE (it is not meant to be another method for finding MLEs). An analytical demonstration of the equivalency between the iterated MME and the MLE seems impossible, as it would involve solving nonlinear equations with powers equal to the number of observations.

(iv) *Properties of the MME*

The MME allows estimation of relatedness or inbreeding with data from as few as one locus. Table 1 gives examples of estimates given by one locus for the case of pairwise relatedness (with additional loci, the MMEs are weighted averages over loci). It shows that the MME can give negative estimates of relatedness, reflecting the large statistical error with this small sample size (constraining the estimate to non-negative bounds introduces positive bias into the estimate, and distorts associations with other variables; see discussion). Estimates greater than 1·0 also occur when relatives share rare alleles. Table 1 also shows that more similar genotypes give higher estimates of relatedness, and that similarity for a rare allele gives an even higher estimate. Increased diversity for other alleles, or increased *a-priori* relatedness, tends to reduce this dependence upon allele frequency.

To determine the statistical properties of the MME, a computer program was written which generated Monte-Carlo datasets, then applied the above estimation procedure. Initially, no sample size corrections were made for P_i. Random numbers were generated by the commonly used 'minimal standard random number generator' (Park & Miller, 1988). Results showed the variance of the MME to be approximately a function of $1/n$ for n the number of loci, which is expected since the MME is nearly a linear function of the data. However, for relationships spanning a wide range, and for many different distributions of gene frequency, a systematic bias on the order of $1/N$, was observed, for N the number of individuals used to estimate gene frequency. As discussed above, the greatest source of bias is due to the inclusion of the particular individual(s) in the estimate of population gene frequency P_i; this can be removed by excluding the related alleles from P_i (this procedure should not be confused with jackknifing, which would

https://doi.org/10.1017/S0016672300033620 Published online by Cambridge University Press

Table 1. *Examples of MMEs of pairwise relatedness r, based upon one marker locus, for various distributions of gene frequency and levels of relatedness*

	Gene frequencies and true relatedness			
Genotypes	$P = 0{\cdot}8, 0{\cdot}2; r = 0$	$P = 0{\cdot}6, 0{\cdot}4; r = 1/4$	$P = 0{\cdot}8, 0{\cdot}2; r = 0$	$P = 0{\cdot}6, 0{\cdot}2, 0{\cdot}2; r = 0$
$A_1A_1 - A_1A_1$	0·25	0·67	0·55	0·33
$A_1A_1 - A_1A_2$	−0·38	−0·17	−0·45	−0·08
$A_1A_2 - A_1A_2$	0·56	0·04	0·11	0·33
$A_1A_2 - A_2A_2$	1·50	0·25	0·67	0·75
$A_1A_1 - A_2A_2$	−1·00	−1·00	−1·45	−0·50
$A_2A_2 - A_2A_2$	4·00	1·50	2·80	2·00

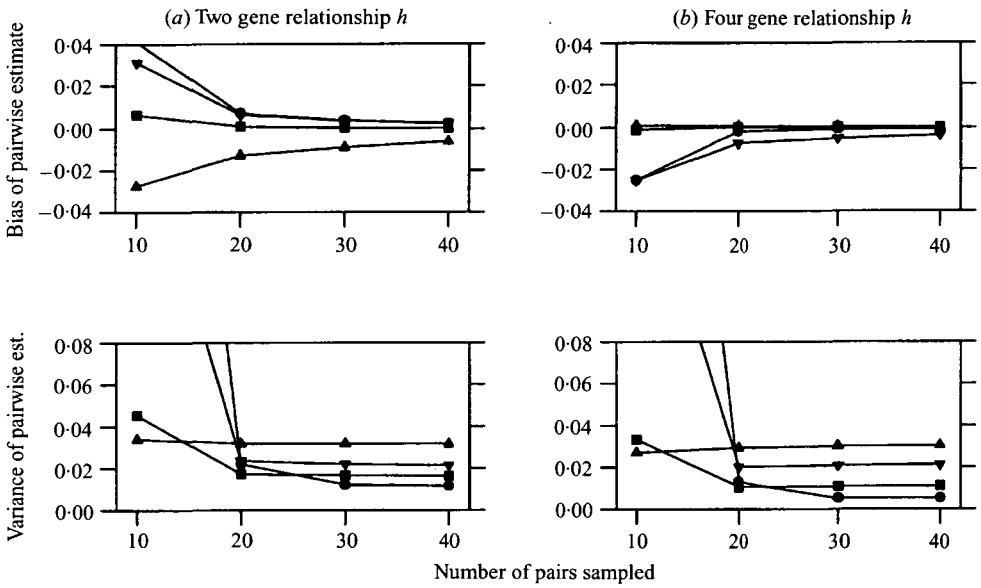

Fig. 2. Bias and variance of estimates as a function of the sample size used to estimate the population gene frequency, for different distributions of gene frequency and actual relatedness (eight loci used to estimate relatedness): ●, $P = \{0{\cdot}1, 0{\cdot}2, 0{\cdot}3, 0{\cdot}4\}$; no true relationship; ■, $P = \{0{\cdot}2, 0{\cdot}3, 0{\cdot}5\}$; no true relationship; ▲, $P = \{0{\cdot}3, 0{\cdot}7\}$; no true relationship; ▼, $P = \{0{\cdot}2, 0{\cdot}3, 0{\cdot}5\}$; full-sib relatives.

estimate N 'pseudovalues' of P_i for each particular pair of relatives). Figure 2 shows the results of applying this procedure. The bias and variance of estimates are plotted against population sample size, for various gene frequencies and actual relationship, using eight marker loci (all with identical gene frequencies).

Figure 2 shows that statistical bias largely disappears with a population size of 20–30 pairs. The remaining bias, due to randomness of gene frequency, is small (*c.* 0·01 or less), corresponding to roughly the level of relatedness between third cousins. The sign of bias is unpredictable, being negative for a diallelic locus but positive with more alleles. Thus, opposing biases across loci with different gene frequencies may cancel out. The estimation variance also declines, but more dramatically; it almost plateaus by 20–30 pairs, where it nearly equals the predicted asymptotic variance $(1/[4(n-1)m]$ for n alleles at each of m loci). Thus a sample size of 20–30 pairs is sufficient for estimating population gene frequency (40–60 individuals when the inbreeding coefficient is estimated).

(v) *Comparison to maximum likelihood estimates*

The maximum likelihood procedure (Thompson, 1975, 1976) gives asymptotically efficient estimates and allows tests of hypothesis via likelihood ratios. For two-gene relationship, the likelihood equation for data from a multiallelic locus is

$$\mathscr{L}(\rho) = [(1-J)(1-\rho)]^{X_0} \prod_i (p_i^2 + p_i q_i \rho)^{X_i}$$

for X_0 is the number of pairs with different alleles, X_i the number of pairs of allele i, and J the expected homozygosity (in this equation, we assume allele frequencies are known so the different heterozygotes combine into one term, or 'factorize'). Given the data, the value of ρ which maximizes this equation is the MLE of ρ. Because we are not testing alternative hypothesis of relationship, but rather estimating relationship as a parameter ρ, we allow estimates to span outside the allowable 'space' of relationship as a consequence of sampling error. Although constraining ρ to within the interval $(0, 1)$ is an accepted practice,

https://doi.org/10.1017/S0016672300033620 Published online by Cambridge University Press

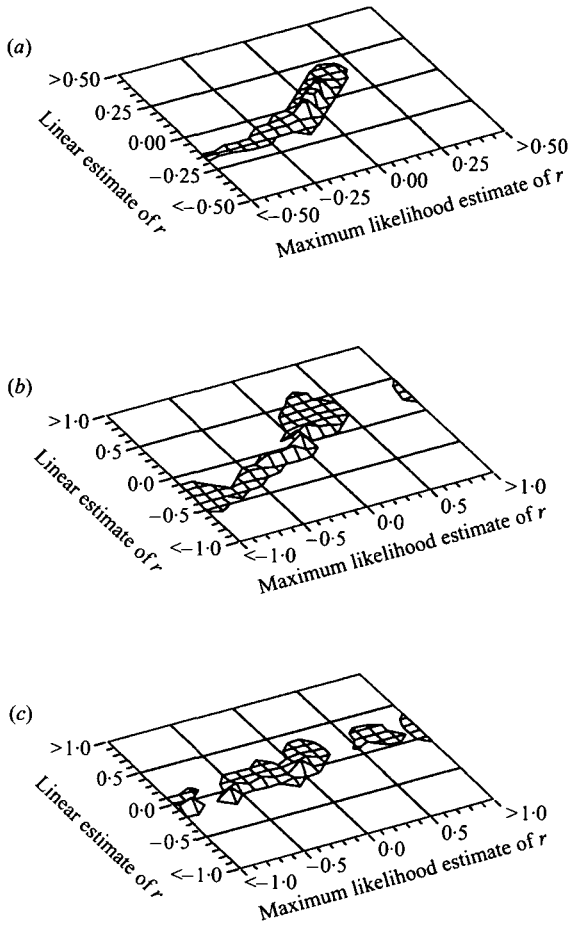

Fig. 3. The joint density of MME ("linear estimate") and MLE estimates based on the same data. (*a*) 64 diallelic loci each with $P = \{0{\cdot}2, 0{\cdot}8\}$, (*b*) eight diallelic loci with the same gene frequencies, and (*c*) four loci each with eight alleles of frequency $P = \{2\times 0{\cdot}05, 3\times 0{\cdot}10, 3\times 0{\cdot}20\}$. The true relatedness was $r = 0$. Note change of scale from (*a*) to (*b*) and (*c*).

it does introduce biases because as a consequence the error residuals do not have zero expectation.

Figure 3 shows the relationship between individual-level MMEs and MLEs based upon the same data. In this figure, each graph is based upon 10^5 Monte-Carlo datasets ($\rho = 0$ assumed), and for each dataset, the likelihood equation was maximized via the Newton–Raphson method to obtain MLEs (maximization was achieved by the Newton–Raphson method, and was determined to be reliable by inspections of the likelihood surface; gene frequencies were assumed known). Three cases were considered: (*a*) 64 diallelic loci each with $p = \{0{\cdot}2, 0{\cdot}8\}$, (*b*) eight diallelic loci with same gene frequencies, and (*c*) four loci each with eight alleles of frequency $p = \{2\times 0{\cdot}05,\ 3\times 0{\cdot}10,\ 3\times 0{\cdot}20\}$.

Figure 3 shows that fewer numbers of marker loci, or with greater polymorphism at marker loci, the disagreement between the MME and the MLE is greater, with the MLE showing higher variances and often being very negative. In fact, when no markers are shared, the maximum of the likelihood surface is undefined, residing at negative infinity. Generally, when the proportion of shared alleles between the two relatives is high, the MME and the MLE are close to each other, but as this proportion decreases, the MMEs and MLEs diverge, with the MLE diving downwards. Other simulations indicate that with any reasonable number of loci (< 50 loci), the MLE of pairwise relatedness still shows downward bias. However, this bias cannot be quantified because of occasional maxima of the likelihood equation at negatively infinite values of r.

3. MMEs for four-gene relationship (Cotterman coefficients)

To completely specify relationship between two outbred relatives, a second coefficient is needed. We hereafter assume no inbreeding (simultaneous estimation of all eight independent coefficients of relationship between inbred relatives, as described by Jacquard 1974, will be given elsewhere). Figure 4 shows the three possible configurations of gene identity between diploid, outbred relatives. Their frequencies can be specified by Cotterman's coefficients k_0, k_1, k_2, where $k_0 + 2k_1 + k_2 = 1$ (Cotterman, 1940), or by r and a four-gene coefficient termed h,

$$\text{Prob(both identical)} = k_2 = 2h,$$

$$\text{Prob(one identical)} = 2k_1 = 4r - 4h$$

$$\text{Prob(none identical)} = k_0 = 1 - 4r + 2h.$$

Either (k_0, k_1 and k_2) or (r, h) describe relatedness; the latter is used herewith (efficient estimates of the Cotterman coefficients are obtained by these relationships).

The 'four-gene' coefficient h is the probability that two pairs of genes are identical by descent between the two relatives. In outbred populations, $h = 1/8$ for full-sibs while $h = 0$ for parent–offspring, half-sibs and first-cousins (see Jacquard, 1974). This pattern of relatedness would also be expected in a sub-structured population; if r varies among demes and mating is random within demes, $h = \bar{r}^2 + \sigma_r^2$. The Cotterman coefficients give the genetic covariance between outbred relatives for a quantitative trait as $(k_1 + k_2)V_a + k_2 V_d$, where V_a and V_d are the

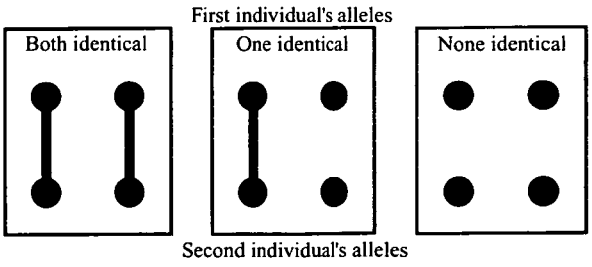

Fig. 4. The three modes of gene identity-by-descent between two outbred individuals at a single locus. Identical genes are linked by lines.

https://doi.org/10.1017/S0016672300033620 Published online by Cambridge University Press

Table 2. *Probabilities of genotypes of two relatives at a diploid marker locus, conditioned by relationship. Subscripts denote alleles, with $i<j$, $i<k$ and $j \neq k$ ($p_i = 1-q_i$, $z_i^2 = p_i^2+q_i^2-\sum_i p_i^2$)*

Pairwise genotypes	Probability	Coefficient		
		1	r	h
$A_iA_i-A_iA_i$	$\mathscr{P}_{4i}$	p_i^4	$4p_i^3q_i$	$2p_i^2q_i^2$
$A_iA_i-A_iA_j$ or $A_iA_j-A_iA_i$	$\mathscr{P}_{3i}$	$4p_i^3q_i$	$8p_i^2q_i(q_i-p_i)$	$-8p_i^2q_i^2$
$A_iA_j-A_iA_k$	$\mathscr{P}_{2i}$	$4p_i^2z_i^2$	$4p_iz_i^2(1-4p_i)$	$-4p_iz_i^2(q_i-p_i)$
$A_iA_j-A_iA_j$	$\mathscr{P}_{2i2j}$	$4p_i^2p_j^2$	$4p_ip_j(p_iq_j+q_ip_j-2p_ip_j)$	$4p_ip_j(p_ip_j+q_iq_j)$

additive and dominance genetic variances, respectively (Crow & Kimura, 1970, p. 137). They are 'four-gene' coefficients of relationship because the patterns of gene identity depend upon all four homologous genes at the locus.

(i) *The MME for four-gene relationship*

Among the four alleles shared by two relatives at a diploid locus, there are a large number of configurations of allelic identity-by-state. In principle, each configuration can give an estimate of h given r, and an estimate of r given h. However, estimates based upon different configurations are correlated and all information in the data is captured by a relatively small subset of configurations. This subset is represented by the following indicators of similarity:

$S_{4i} = 1$ if genotype $A_iA_i-A_iA_i$ is observed,

$S_{3i} = 1$ if genotype $A_iA_i-A_iA_j$ or $A_iA_j-A_iA_i$ $(i<j)$ is observed,

$S_{2i} = 1$ if genotype $A_iA_j-A_iA_k$ $(i<j,\ i<k,\ j\neq k)$, is observed, and

$S_{2i2j} = 1$ if genotype $A_iA_j-A_iA_j$ is observed $(i<j)$; (7)

otherwise, these Ss are zero. These indicators constitute the data; at most one is non-zero, this value corresponding to observed genotype. For example, if we observe $A_2A_2-A_2A_2$, then $S_{42}=1$ and all remaining Ss are zero. If we observe $A_1A_2-A_3A_3$, then every single one of these Ss is zero. The first, second and third similarity applies to each of n alleles, and the fourth to each of $n(n-1)/2$ pairings of different alleles. The third is omitted for diallelic loci, as it applies to configurations involving at least three different alleles. In total, there are five configurations for diallelic loci, and $n(n+5)/2$ configurations for multiallelic loci $(n>2)$. The sufficiency of these configurations has been confirmed by numerically calculating that the analytical variance of the MLE equalled that of the following MME.

The probabilities of S_{4i}, S_{3i}, S_{2i} and S_{2i2j} are given by $\mathscr{P}_{4i}$, $\mathscr{P}_{3i}$, $\mathscr{P}_{2i}$ and $\mathscr{P}_{2i2j}$, respectively, in Table 2. If we equate these $\mathscr{P}$s with the Ss, we obtain the following method of moment estimators for r:

$$\hat{r}_{4i} = \frac{S_{4i}-P_i^4-2\hat{h}P_i^2Q_i^2}{4P_i^3Q_i},$$

$$\hat{r}_{3i} = \frac{S_{3i}-4P_1^3Q_i+8\hat{h}P_i^2Q_i^2}{8P_i^2Q_i(Q_i-P_i)},$$

$$\hat{r}_{2i} = \frac{S_{2i}-4P_i^2Z_i^2-4\hat{h}P_iZ_i^2(P_i-Q_i)}{4P_iZ_i^2(1-4P_i)},$$

$$\hat{r}_{2i2j} = \frac{S_{2i2j}-4P_i^2P_j^2-4\hat{h}P_iP_j(P_iP_j+Q_iQ_j)}{4P_iP_j(P_iQ_j+Q_iP_j-2P_iP_j)},$$

and for h:

$$\hat{h}_{4i} = \frac{S_{4i}-P_i^4-4P_i^3Q_i\hat{r}}{2P_i^2Q_i^2},$$

$$\hat{h}_{3i} = \frac{S_{3i}-4P_i^3Q_i-8\hat{r}P_i^2Q_i(Q_i-P_i)}{-8P_i^2Q_i^2},$$

$$\hat{h}_{2i} = \frac{S_{2i}-4P_i^2Z_i^2-4\hat{r}Z_i^2(1-4P_i)}{-4P_iZ_i^2(Q_i-P_i)},$$

$$\hat{h}_{2i2j} = \frac{S_{2i2j}-4P_i^2P_j^2-4\hat{r}P_iP_j(P_iQ_j+Q_iP_j-2P_iP_j)}{4P_iP_j(P_iP_j+Q_iQ_j)}. \quad (8)$$

Note that alleles of 0·25 and 0·5 frequency cannot be used in some of the above estimators. Also note the negative signs in some denominators, which reflect the unusual nature of four-gene estimators.

The configuration-specific estimates of r and h are then combined into single estimators as

$$\hat{r} = \sum_i w_{r2i}\hat{r}_{2i} + w_{r3i}\hat{r}_{3i} + w_{r4i}\hat{r}_{4i} + \sum_{j>i} w_{r2i2j}\hat{r}_{2i2j},$$

$$\hat{h} = \sum_i w_{h2i}\hat{h}_{2i} + w_{h3i}\hat{h}_{3i} + w_{h4i}\hat{h}_{4i} + \sum_{j>i} w_{h2i2j}\hat{h}_{2i2j}. \quad (9)$$

where the weights w are specific for r and h, as indicated by their subscripts. This summation is also extended across loci.

For each locus, the weights are computed separately for r (given h) and for h (given r), using the procedure of (4*b*). For each computation, a matrix of size

https://doi.org/10.1017/S0016672300033620 Published online by Cambridge University Press

$n(n+5)/2$ is inverted (except size 5 when $n = 2$). The precise formula for the variances and covariances used in (4*b*) are too complex to describe, but follow a general formula as follows. If h_i is estimated as $\hat{h}_i = (S_i - X_i)/Y_i$, then $\mathrm{Var}(\hat{h}_i) = E[S_i](1-E[S_i])/y_i^2$ and $\mathrm{Cov}(\hat{h}_i, \hat{h}_j) = -E[S_i]\,E[S_j]/y_i y_j$. This follows from the multinomial distribution of the similarities (7). As discussed previously, these expected *S*s should be independent of the observed data, and are found via the equations in Table 2, wherein one uses either a prior guess of relatedness, or the population average relatedness. Errors in the prior guess of relatedness only increase the variance of the MME, but do not affect bias. It is simplest to let $r = h = 0$ in computing these weights.

Rewriting (9), the joint estimators for r and h take the general form

$$\left.\begin{aligned}\hat{r} &= a + b\hat{h},\\ \hat{h} &= c + d\hat{r},\end{aligned}\right\} \qquad (10a)$$

where a and c are functions of the gene frequencies and similarities, and b and d are functions of only gene frequencies. Explicit expressions for a, b, c and d cannot be given, but can be computed numerically. Solving this pair of equations gives the joint estimators

$$\left.\begin{aligned}\hat{r} &= \frac{a+bc}{1-bd},\\ \hat{h} &= \frac{c+ad}{1-bd}.\end{aligned}\right\} \qquad (10b)$$

A simple MME for h (analogous to 6) is too complex to be found analytically. Interestingly, when weights are numerically calculated, they show rather bizarre behaviour, often taking negative values and being highly sensitive to gene frequencies.

For comparison, the corresponding likelihood equation for a sample of data from one locus with any number of alleles, and with gene frequencies assumed known, is

$$\mathscr{L}(r,h) = \mathscr{P}_0^{X_0} \prod_i \mathscr{P}_{4i}^{X_{4i}} \mathscr{P}_{3i}^{X_{3i}} \mathscr{P}_{2i}^{X_{2i}} \prod_{j>i} \mathscr{P}_{2i2j}^{X_{2i2j}},$$

where, after Table 2, the numerical subscripts denoting observations X give the number of alleles of type i shared between relatives, and the probability of no alleles in common is

$$\mathscr{P}_0 = x - 4xr + 2xh \quad \text{for} \quad x = 1 - 6J_2 + 3J_2^2 + 8J_3 - 6J_4$$

and $\quad J_k = \sum_i p_i^k.$

(ii) *Properties of the four-gene estimate*

At $h = 0$ and r assumed known, the variance of a single-locus weighted MME of h based upon an n-allele locus is $1/(2n(n-1))$, regardless of the frequency distribution of alleles. Thus when $n > 3$, the estimation variance of h can actually be less than the two-gene parameter r (found earlier to equal $1/(4(n-1))$; at $n = 3$, their variances are equal.

Figure 1*b* shows the relative efficiency of assuming the prior weights for h of zero *v.* complete relatedness (assuming known r). The efficiency of assuming zero relatedness is not as high as for r (Fig. 1*a*), but is more efficient than assuming maximum possible h. The efficiency is greater for a more uniform distribution of gene frequency. Figure 2*b* shows the statistical properties of h estimates as a function of the sample size used to estimate the population gene frequency (based on Monte-Carlo simulations, as in Fig. 2*a*). Again, 20–30 pairs of individuals seems sufficient to remove bias. The variance stabilized at this number as well, approaching the predicted asymptotic value of $1/[16n(n-1)]$ for eight loci.

4. Discussion

The method-of-moments estimator (MME) is most appropriate for inferring relatedness between individuals or the inbreeding coefficients of individuals. Such a fine-scale inference brings with it statistical problems caused by the inherent small sample sizes, but in the case when population gene frequencies are estimated from a larger sample, the MME still has effectively asymptotic (large sample) properties at this individual level. This estimator should be particularly useful for highly polymorphic markers such as microsatellites, since the small sample problems with the maximum likelihood equation get worst with highly variable loci, while the properties of the MME remain the same (Fig. 3).

This property occurs because the MME additively combines estimates given by each allele at each locus. In effect, the MME equates the data to an equation linear in r, of form $S = a + br$, and the estimation equation is an explicit function of the data S. (Even in the complex case of four-gene estimation, the estimators 10*b* are linear in the data S, remembering that population gene frequencies are assumed known.) By contrast, the maximum likelihood equation is non-linear in the data: the estimate that maximizes the probability of the observed data involves higher powers of r, e.g. is nonlinear in r. In both methods, the probabilities may be assumed to generate the data, but the estimates they give can differ greatly.

The MME is based upon finding estimates which are functions of the patterns for similarity for individual marker alleles or (in the case of four-gene relatedness) for pairs of alleles, then averaging estimates using a weighted summation. As discussed, the computation of these weights requires three compromises: (1) gene frequencies are estimated and not known as assumed, (2) gene frequencies are not jointly estimated with relationship, and (3) 'prior' relationship is guessed. The bias of the MME is not affected by these compromises, and the increase of

https://doi.org/10.1017/S0016672300033620 Published online by Cambridge University Press

variance incurred by these compromises is small compared to the asymptotic properties that are maintained by the MME with individual-level estimates. Also, if gene frequencies and prior relationship were jointly estimated, the estimator would become nonlinear, an iterative procedure would be required, and in fact, the MLE is obtained (with the resulting small sample size problems). Such an iterated MME, which is analogous to an iteratively reweighted least-squares method, shows that the original MME is statistically efficient as best can be, and not a suggested alternative for finding the MLE.

The MME can give negative estimates of relationship (when most or all markers are not shared) as well as estimates greater than 1·0 (when rare markers are shared). This reflects the large statistical errors of inferring relatedness or inbreeding at the individual level. Some statisticians would recommend the constraining of estimates to the 'allowable' space of relationship. However, unless this truncation is exactly symmetrical, this distorts the error residuals and introduced bias of estimates. For example, estimates of the variance and covariance of actual relationship (Ritland, 1996*a*) will be skewed downward and any analyses performed in conjunction with these truncated estimates will be biased.

Queller & Goodnight (1989) presented a regression-based estimator for two-gene relatedness (their eqns 5 and 6; they estimate $2r$ which equals the regression coefficient of relationship for outbred individuals). Their estimator was designed to estimate group relatedness, but was also deemed suited for estimating pairwise relationships with some qualifications. Their major problem was undefined relatedness of a heterozygote to any other individual (Queller & Goodnight, 1989, pp. 268–269). This does not pose any problem with the MME, although negative estimates are obtained (Table 1). In addition, to combine information among alleles, they take a ratio of a sum, rather than a sum of ratios, as Barbujani (1987) suggests with a similar estimator. This results in a reduction of statistical efficiency, since it ignores the varying information among loci.

Loeselle *et al.* (1996), who modified Barbujani's estimator to the case of pairwise relationship, suggested weighting estimates from each locus by their heterozygosity. Both numerical and analytical results herein show the optimal per-locus weights are proportional to simply the number of alleles at the locus. However, we have assumed that gene frequencies estimated from a larger population sample, such that their sampling variance is small compared to the variance of estimated relationship. Inclusion of this component of variance will favour more heterozygous loci to an unknown, but probably slight, extent.

In addition, simulations show that infrequent alleles (ca. less than 0·05 frequent) introduce a heretofore unconsidered source of bias, due to the likelihood of missing alleles in the sample. To avoid this problem, it is recommended to bin rare alleles into classes at least 0·05 frequent (although this would depend strictly upon the population sample size). This seems to reduce the value of microsatellite loci, which commonly have 20–30 or more alleles, unless other estimators are developed specifically for hypervariable loci.

The four-gene MMEs are sensitive to gene frequencies in totally unexpected ways. Although with known gene frequency, the information about the four-gene coefficient h is simply proportional to $n(n-1)$, for n the number of alleles at the locus, the actual distribution of gene frequency has a more complex effect on estimates than in the two-gene case. Alleles of 0·25 frequency cause $\hat{r}_{2i}$ to 'blow up' in (8), and alleles near 0·5 frequency cause $\hat{r}_{3i}$ and $\hat{h}_{2i}$ to 'blow up' in (8). Paradoxically, loci which are informative in the two-gene case are non-informative in the four-gene case. To deal with this problem, when alleles are sampled near these frequencies (to within *c.* 0·05), these component estimators should be removed from the total estimate.

The primary value of the estimators developed herein is for studies of pairwise relatedness or individual inbreeding coefficients. It has been noted (Thompson, 1975) that estimates of pairwise relationship have quite high variance, in the sense that one cannot distinguish half-sibs from full-sibs, etc. (although few or single marker loci may be sufficient to exclude certain relationships, such as paternity). Probably for this reason, workers have not put effort into finding estimators of relatedness or inbreeding of individuals with good statistical properties when few marker loci are available. However, the recent advent of new marker-based approaches in ecology and evolution (Cruzan, 1996; Ritland 1996*a*) are creating opportunities for the use of individual-level estimators of relatedness or inbreeding.

I thank Joe Felsenstein and Elizabeth Thompson for their comments during an early stage of this work, and Joe for use of his computer for simulations. This research was supported by a Natural Sciences and Engineering Research Council of Canada grant to K. R.

References

Barbujani, G. (1987). Autocorrelation of gene frequencies under isolation by distance. *Genetics* **117**, 777–782.

Brown, A. H. D., Barrett, S. C. H. & Moran, G. F. (1985). Mating system estimation in forest trees: models, methods and meanings. In *Population Genetics in Forestry*, Lecture notes in Biomathematics (ed. S. Levin), **60**, 32–49.

Chakraborty, R., Meagher, T. R. & Smouse, P. E. (1988). Parentage analysis with genetic markers in natural populations. I. The expected proportion of offspring with unambiguous paternity. *Genetics* **118**, 527–536.

Cotterman, C. W. (1940). *A Calculus for Statistico-genetics.* Unpublished thesis, Ohio State University, Columbus, Ohio.

Curie-Cohen, M. (1981). Estimates of inbreeding in a natural population: a comparison of sampling properties. *Genetics* **100**, 339–358.

Crow, J. F. & Kimura, M. (1970). *An Introduction to Population Genetics Theory*. Burgess, Minneapolis.
Cruzan, M. (1996). Genetic markers in plant evolutionary biology. Ecology (in the Press).
Emigh, T. H. (1980). A comparison of tests for Hardy–Weinberg equilibrium. *Biometrics* **36**, 627–642.
Epperson, B. K. (1989). Spatial patterns of genetic variation within plant populations. In *Plant Population Genetics Breeding, and Genetic Resources* (ed. A. H. D. Brown, M. T. Clegg, A. L. Kahler and B. S. Weir), pp. 229–253. Sunderland, Mass.: Sinauer Associates.
Grafen, A. (1985). A geometric view of relatedness. *Oxford Survey Evolutionary Biology* **2**, 28–89.
Jacquard, A. (1974). *The Genetic Structure of Populations*. Berlin: Springer.
Li, C. C. & Horvitz, D. G. (1953). Some methods of estimating the inbreeding coefficient. *American J. Human Genetics* **5**, 107–117.
Loeselle, B., Sork, V., Nason, J. & Graham, C. (1996). spatial genetic structure of a tropical understory shrub, *Psychotria officinalis* (Rubiaceae). *American Journal of Botany* (in the Press).
Lynch, M. (1988). Estimation of relatedness by DNA fingerprinting. *Molecular Biology and Evolution* **5**, 584–599.
Malecot, G. (1969). *The Mathematics of Relationship*. San Francisco: W. H. Freeman.
Michod, R. D. & Hamilton, W. D. (1980). Coefficients of relatedness in sociobiology. *Nature* **288**, 694–697.
Morton, N. E., Yee, S., Harris, D. E. & Lew, R. (1971). Bioassay of kinship. *Theoretical Population Biology* **2**, 507–524.
Pamilo, P. & Crozier, R. H. (1982). Measuring genetic relatedness in natural populations: methodology. *Theoretical Population Biology* **21**, 171–193.
Park, S. K. & Miller, K. W. (1988). Random number generators: good ones are hard to find. *Communications of ACM* **31**, 1192–1201.
Queller, D. C. & Goodnight, K. F. (1989). Estimating relatedness using genetic markers. *Evolution* **43**, 258–275.
Ritland, K. (1996*a*). A marker-based method for inferences about quantitative inheritance in natural populations. *Evolution* (in the Press).
Ritland, K. (1996*b*). Estimation of individual population divergence, a component of genetic diversity. (Submitted to) *Molecular Ecology*.
Ritland, K. & Ritland, C. (1996). Inferences about quantitative inheritance based upon natural population structure in the common yellow monkeyflower, *Mimulus guttatus*. *Evolution* (in the Press).
Robertson, A. & Hill, W. G. (1984). Deviations from Hardy–Weinberg proportions: sampling variances and use in estimation of inbreeding coefficients. *Genetics* **107**, 703–718.
Schuster, W. S. F. & Mitton, J. B. (1991). Relatedness within clusters of a bird-dispersed pine and the potential for kin interactions. *Heredity* **67**, 41–48.
Thompson, E. A. (1975). The estimation of pairwise relationships. *Annals of Human Genetics* **39**, 173–188.
Thompson, E. A. (1976). Inference of genealogical structure. *Social Sciences Information* **15**, 477–526.
Weir, B. S. & Cockerham, C. C. (1984). Estimating *F*-statistics for the analysis of population structure. *Evolution* **38**, 1358–1370.
Weir, B. S. (1990). *Genetic Data Analysis*. Sunderland, MA.: Sinauer Associates.
Wilkinson, G. S. & McCracken, G. F. (1986). On estimating relatedness using genetics markers. *Evolution* **39**, 1169–1174.
Wright, S. (1922). Coefficients of inbreeding and relationship. *American Naturalist* **56**, 330–338.
Wright, S. (1943). Isolation by distance. *Genetics* **28**, 114–138.
Wright, S. (1969). *Evolution and the Genetics of Populations*, vol. 2. The theory of gene frequencies. Chicago: University of Chicago Press.
Yasuda, N. (1968). Estimation of the inbreeding coefficient from phenotypic frequencies by a method of maximum likelihood scoring. *Biometrics* **24**, 915–935.