Estimating Quantitative Genetic Parameters Using Sibships Reconstructed From Marker Data

Stuart C. Thomas and William G. Hill

Institute of Cell, Animal and Population Biology, University of Edinburgh, Edinburgh EH9 3JT, United Kingdom

Manuscript received February 10, 2000
Accepted for publication April 13, 2000

ABSTRACT

Previous techniques for estimating quantitative genetic parameters, such as heritability in populations where exact relationships are unknown but are instead inferred from marker genotypes, have used data from individuals on a pairwise level only. At this level, families are weighted according to the number of pairs within which each family appears, hence by size rather than information content, and information from multiple relationships is lost. Estimates of parameters are therefore not the most efficient achievable. Here, Markov chain Monte Carlo techniques have been used to partition the population into complete sibships, including, if known, prior knowledge of the distribution of family sizes. These pedigrees have then been used with restricted maximum likelihood under an animal model to estimate quantitative genetic parameters. Simulations to compare the properties of parameter estimates with those of existing techniques indicate that the use of sibship reconstruction is superior to earlier methods, having lower mean square errors and showing nonsignificant downward bias. In addition, sibship reconstruction allows the estimation of population allele frequencies that account for the relationships within the sample, so prior knowledge of allele frequencies need not be assumed. Extensions to these techniques allow reconstruction of half sibships when some or all of the maternal genotypes are known.

ESTIMATES of the genetic parameters of quantitative traits, such as heritability, are important because they give an indication of the ability of a species to respond to selection and thus the potential of that species to evolve (Lande 1982; Mousseau and Roff 1987; Falconer and Mackay 1996; Lande and Shannon 1996). In addition, genetic parameter estimates are finding a place in conservation studies through, for example, estimates of the total genetic variability of a population (Storfer 1996).

Traditional techniques for estimating variance components require, however, knowledge of the relationships among the individuals recorded (Falconer and Mackay 1996; Lynch and Walsh 1998). In natural populations, detailed knowledge of pedigree is absent in all but the most carefully studied populations, and even then may be subject to errors. Molecular marker data provide a means to infer relationship without a full pedigree.

Molecular-based tools for inferring genetic relationships may be grouped into two categories: method-of-moments estimators, which are used to estimate relatedness, as a continuous measure, on the basis of shared alleles at marker loci (Lynch 1988; Queller and Goodnight 1989; Ritland 1996a; Lynch and Ritland 1999); and likelihood techniques, used to determine the likelihood of a pair falling into particular relationship classes, *e.g.*, full sibs or nonsibs, given the observed marker information (Thompson 1975; Mousseau *et al.* 1998).

Similarly, two methods that allow the estimation of quantitative genetic parameters associated with a trait without reference to the exact pedigree have been described (Ritland 1996b; Lynch and Walsh 1998; Mousseau *et al.* 1998). These use molecular data to infer pairwise relationships between individuals, since this is the least complex level at which relationships may be estimated. Ritland (1996b) proposed a regression approach to parameter estimation, where measures of pairwise phenotypic similarity are regressed against pairwise relatedness (Ritland 1996b; Lynch and Walsh 1998). Alternatively, if prior information is available on population structure, likelihood-based procedures may be adopted, in which pairs are placed into a predetermined population structure according to the probability of observing their genotype and phenotype (Mousseau *et al.* 1998; Thomas *et al.* 2000).

Pairwise techniques lose valuable information in the form of higher-order relationships. For example, if three individuals sampled from a single generation have genotypes a_ia_i, a_ja_j, and a_ka_k (a_i, a_j, and a_k are mutually exclusive alleles), they cannot be full sibs; but with pairwise analysis, such exclusion is not possible. Additionally, with pairwise techniques the weight placed on information from a single family depends on the number of pairs of individuals that can be chosen from that family.

Corresponding author: Stuart C. Thomas, Institute of Cell, Animal and Population Biology, University of Edinburgh, W. Mains Rd., Edinburgh EH9 3JT, United Kingdom.
E-mail: sthomas@srv0.bio.ed.ac.uk

It is therefore dependent only upon family size and not information content. Consequently, pairwise methods do not yield the most efficient estimates for parameters and are prone to larger standard errors than efficient methods of estimation such as restricted maximum likelihood (Thomas *et al.* 2000). Only in the case of balanced populations containing two classes of relationship are families weighted equally, and then they give estimates identical to ANOVA-derived estimates when exact pedigree information is known (Thomas *et al.* 2000).

A secondary problem is obtaining estimates of the allele frequencies at the marker loci. In previous studies, allele frequencies have been assumed known or have been estimated from the sample. If allele frequencies are estimated from the sample under investigation, they are subject to further random error, since there are relatives within the sample, which might bias subsequent estimates of pairwise relationships. To combat this problem, Queller and Goodnight (1989) proposed recalculating the allele frequencies for each pair under investigation, excluding the information from that pair. This removes a small covariance between population and individual allele frequencies and results in slightly improved estimates, although change is negligible with large numbers. Ritland (1996a) adopted the same approach.

A final problem with pairwise methods is how they may be extended to include other factors such as sex or year in the model. Since they operate on a pairwise level, other factors must also be investigated on a pairwise level and as a result the optimum estimate may not be achieved.

Here we demonstrate a simple two-step procedure for estimating variance components: first, families of sibs are reconstructed using a Markov chain Monte Carlo (MCMC) procedure, and second, the reconstructed sibships are used to estimate variance components.

The MCMC procedure reconstructs sibships within a single generation, allowing improved parameter estimation through more efficient weighting of families and use of more than pairwise pedigree information. Conceptually the sibship reconstruction procedure shares features with Bayesian approaches using MCMC procedures in phylogeny reconstruction (Kuhner *et al.* 1995; Yang and Rannala 1997; Larget and Simon 1999), where, given the sequence data, the most plausible phylogenetic trees are generated from a large number of potential trees without the need to investigate every possible tree. Similarly, in sibship reconstruction, plausible sibships are generated from the sample using the marker data without the need to investigate every possible combination of sibships. However, in sibship reconstruction the aim is to reconstruct a number of groups with specific relationships rather than determine likely distances between each member (or taxon) in the sample. This approach of reconstructing specific groups is equivalent to fixing the possible branch lengths to either one length (representing full sib) or to double that length (representing unrelated) in phylogeny reconstruction. Moreover, no attempt is made to update the assumed prior distributions of the parameters used in pedigree reconstruction, since these are not the parameters of interest. In this light, the techniques used in pedigree reconstruction are not Bayesian in nature.

Reconstructed pedigrees are subsequently used to form a relationship matrix suitable for use in an animal model run with restricted maximum likelihood (REML), specifically using the ASREML program (Gilmour *et al.* 1997). This approach allows traditional efficient methods for parameter estimation to be used and hence simplifies the inclusion of additional factors or the use of multivariate analysis if data have been collected from several traits. In addition, methods are outlined that allow the estimation of population allele frequencies that account in part for relationships within the sample. In many natural populations, half sibships are more common than full sibships, and in addition some maternal genotype information may also be available. This mimics the situation in some studied natural populations [*e.g.*, the intensively studied Soay sheep (*Ovis aries*) population on the St. Kilda island group, Scotland]. Simple extensions to the MCMC procedure are discussed to allow for the reconstruction of paternal half-sib families where some percentage of the maternal genotypes is assumed to be known.

INFERRING SIBSHIPS

Markov chains: Markov chain Monte Carlo simulations facilitate the determination of solutions to problems that cannot readily be solved by theoretical calculations (Norris 1997). A Markov chain is a random walk through the parameter space of a system, where each step of the walk depends only upon the current state of chain. If the likelihood for the set of parameters at the current point of the chain is calculated and compared against the likelihood at the next point, then the random walk may be "guided" to points of high likelihood within the parameter space. This provides a way to estimate parameter values with a high likelihood (though not necessarily the highest) without having to search the entire parameter space. These techniques are therefore of particular use in solving complex likelihood problems, especially when the parameter space is large.

In this study, we first use only molecular data to reconstruct sibships, assuming individuals are either sibs or unrelated using an MCMC approach, and then use the reconstruction to estimate variance components for a quantitative trait. Errors in pedigree reconstruction are of two types: type I, where genuinely unrelated individuals are classed as related, and type II, where genuinely related pairs are classed as unrelated. It is shown that type I errors lead to large downward bias in parameter estimation, while type II errors lead only to trivial down-

ward bias. It is not necessary to find the point with highest likelihood, but merely a point of high likelihood, since, first, sibship reconstruction leads to few errors of type I, and second, the true sibship may not have the highest likelihood given the marker information.

The population: Suppose a sample of n individuals has been taken from a single generation of a population. Each individual has been scored for genotype at l physically unlinked marker loci, and there is some information on which to base assumptions about relationship structure. In the case described here, it is assumed that the sample contains only full sibs and unrelated individuals and that the distribution of full-sib family sizes is known. Other relevant information might be about known relationships, such as between offspring and dam in a half-sib structure. The likelihood of the relationship structure, allele frequencies, and genotypes of the individual animals may be calculated from the sample. This is a function of the observed marker information and any previous knowledge of the allele frequencies and relationship structure. The likelihood may be expressed as

$$L_{\text{population}} = L(a, g, s|m, d), \tag{1}$$

where a represents the marker allele frequencies within the population, g denotes the n genotypes within the sample, s denotes the sample space of possible family structures (*i.e.*, the sib family membership), m denotes the observed marker information, and d represents the previous knowledge, such as the distribution of family size, about relationship structures.

Maximizing (1) over all possible family structures is prohibitive, since an extremely large number is possible, even in small samples. For example, with 10 individuals, restricted to being either full sib or unrelated, there are 115,975 possible family structures. Markov chains or other optimization techniques are therefore required.

With known allele frequencies

The likelihood of individual families: If allele frequencies are assumed to be known, then individual family likelihoods become independent and Equation 1 may be expressed as

$$L_{\text{population}} = \prod_f L(g_f, s_f|m_f, d, a), \tag{2}$$

where f indexes family.

In the model, the likelihood of any single family f of size n_f is equal to the likelihood of the observed genotypes given that all the members of f are full sibs, multiplied by the likelihood of observing the structure (here the size) of family f given the prior information:

$$L_{\text{family}} = L_{\text{genotypes}} \cdot L_{\text{structure}}. \tag{3}$$

The likelihood of the observed genotypes within a putative full-sib family is calculated as

$$L_{\text{genotypes}} = \prod_l \left[\sum_{w=1}^{b_l} \sum_{x=1}^{b_l} \sum_{y=1}^{b_l} \sum_{z=1}^{b_l} \left[p^1_{wx},\ p^2_{yz} \prod_{i=1}^{n_f} L(g_{il}) \right] \right], \tag{4}$$

where l denotes independent marker loci; b_l, the number of alleles at locus l; w, x, y, and z, index alleles; i indexes an individual from the putative family; p^1_{wx} is the ordered genotype frequency of parent 1; p^2_{yz} is the ordered genotype frequency of parent 2; and $L(g_{il})$ is the likelihood of observing the genotype of individual i at locus l given the parental genotypes. For example, if p^1 and p^2 share the genotype (1, 2) then $L(g_{il}) = \frac{1}{4}$ when the offspring genotype, g, is (1, 1) or (2, 2), $L(g_{il}) = \frac{1}{2}$ when g is (1, 2), and $L(g_{il}) = 0$ otherwise.

In practical computing it is much more efficient, reducing running time to its square root, to take the first offspring and assign one of its alleles to one parent and the other allele to the other parent and then sum over the remaining alleles (see appendix).

In the simulations, the likelihood of the family structure depends only upon the family size (since category information is included in the way the Markov chain mixes the population). Either a noninformative distribution for full-sib family size, where each family size is equally likely, or a truncated Poisson (Po) distribution (no zero class) describing the probability of each family size was used. The independence of families allows fast Monte Carlo algorithms to be written, since at each step in the chain only the likelihoods of individual families rather than the likelihood of the whole population need to be considered.

The "hill climbing" algorithm for full-sib family reconstruction:

a. Start with each member of the sample assigned to a different family. This starting point avoids the problem of generating populations with likelihoods of zero, which is almost a certainty with randomly selected families.
b. Calculate the likelihood for each family and store.
c. Select a random individual, x, from a randomly chosen family f_1. This individual is to be moved at random to a new location (new family) within the sample.
d. Select a random destination family, f_2, for individual x (including family f_1 and a "blank" family containing no individuals). The new location is chosen in a way that allows the individual to stay in the same place or to be placed in a new family on its own.
e. Calculate $L_{\text{old}} = L(f_1) \times L(f_2)$. Use the stored likelihoods to calculate the likelihood of observing families f_1 and f_2 prior to moving x. This equals the product of the likelihoods of each family on its own, since families are independent.
f. Move x from f_1 to f_2.
g. Calculate new likelihoods for f_1 and f_2 after the move of x; these are termed $L(f_1)_{\text{new}}$ and $L(f_2)_{\text{new}}$.

h. Calculate the new likelihood of observing both families, $L_{new} = L(f_1)_{new} \times L(f_2)_{new}$.
i. Calculate $r = L_{new}/(L_{new} + L_{old})$.
j. Draw z from a uniform distribution between 0 and 1.
k. Compare z with r. If $z < r$, move x back to f_1. If $z \geq r$, store $L(f_1)_{new}$ and $L(f_2)_{new}$. This step means that the probability of accepting a change, *i.e.*, keeping x in f_2, depends on the change of the likelihood. If $L_{new} \gg L_{old}$, the change is almost certainly accepted, but if $L_{old} \ll L_{new}$, then the change is almost certainly rejected. In addition, this allows backsteps, a decrease in the likelihood, to occur, thereby reducing the chance that the chain will become stranded on a false maximum.
l. Return to c. Continue to move (mix) individuals between families until stopping criteria are reached; these are discussed below.

A number of criteria may be used to stop the chain:

1. A fixed number of iterates has been run. This method must be repeated a number of times for the sample and the resulting full-sib families compared for similarity. The population with the greatest likelihood may then be selected, or some composite structure determined (although this requires additional checking for exclusions).
2. The likelihood for the whole population (the product of the stored likelihoods) remains constant or nearly so for a fixed number of iterates.
3. The average family size approaches the expected family size and then remains constant or nearly so for a fixed number of iterates.

In practice, there is little difference between using criteria 2 and 3 to stop the chain. In populations of size 200, with five alleles at each of 10 loci and a family size distribution that is Po(5), the population likelihood and mean family size level out together, with the values stabilized by 300,000 cycles (often by 220,000). With the same level of marker information, a population of size 800 stabilizes after ~900,000 cycles.

Half-sib reconstruction: The algorithm is easily modified to accommodate the reconstruction of half-sib families. For half-sib families, the probability of observing the genotypes of a putative half-sib family, over all the possible genotypes of the shared parent, is computed for each locus and then multiplied across loci. The likelihood of each offspring depends on the likelihood of receiving one allele from the common parent and the other from an allele pool with the same allele frequencies as the population. Parental genotype information may be incorporated into both half- and full-sib algorithms by constraining the parental genotypes over which the offspring genotype likelihoods must be summed. The likelihood equation for a half-sib family is included in the appendix.

With unknown allele frequencies

Calculating parental allele frequencies from samples containing relatives: Population allele frequencies are usually unknown and must also be calculated from the sample. In sibship reconstruction, the likelihood of observing a particular sibship depends on the allele frequency in the parental generation, since these are the alleles that are sampled to form the offspring generation. Allele frequencies may be estimated by using a weighted least-squares approach (Dillon and Goldstein 1984), with correlations of the allele counts between relatives accounted for by inclusion of the relationship matrix. The derived estimator is dependent only upon the relationship matrix and the allele counts

$$\hat{a}_i = (\mathbf{1}^T\mathbf{R}^{-1}\mathbf{a})(\mathbf{1}^T\mathbf{R}^{-1}\mathbf{1})^{-1}, \quad (5)$$

where $\hat{a}_i$ is the mean allele count, $\mathbf{R}$ is the relationship matrix, $\mathbf{a}$ is the vector containing the allele counts for each individual, and $\mathbf{1}$ is a vector of ones. Allele frequency is then estimated as $\hat{a}_i/2$.

An updated algorithm: The previous algorithm can be modified using the allele frequency estimator so as to update allele frequencies. The process is begun by calculating the allele frequencies as though all members of the population are unrelated and then periodically updating the estimates as groups of full sibs are generated (*e.g.*, every 5000 iterates). Recalculation every step is unnecessary: first, there may be no change made in population structure, and second, a single change does not affect allele frequency estimates significantly. Updating allele frequencies reduces the population frequency of alleles shared by grouped individuals and also reduces the probability that a group reconstructed as full sibs will be broken down again, even if the reconstruction is wrong. It is therefore recommended that allele updating start after a number of cycles have already been run (say, 100,000).

Measuring the accuracy of the reconstructed family

A statistic that enables measurement of the accuracy of each reconstructed family is useful for the purposes of comparison. Simulating populations with known relationships using the same parameters (or estimates of the parameters) for the distribution of family size as those of the study population allows percentage confidence levels for a given size of family to be estimated. Two confidence levels may be determined: the probability that full-sib family members in the family reconstructed are genuine full sibs and the probability that the family is complete (*i.e.*, is not the result of a larger family being split—a possible problem with this approach to sibship reconstruction).

To assess the properties of the estimators in the simulated study, where the real family structure is known,

an additional statistic that scores the reconstructed pedigree for accuracy was defined

$$\text{accuracy} = (S_{\text{fs|fs}} - S_{\text{fs|ur}})/\text{Tot}_{\text{fs}}, \quad (6)$$

where $S_{\text{fs|fs}}$ is the total number of correctly reconstructed full-sib pairs, $S_{\text{fs|ur}}$ is the total number of incorrectly reconstructed full-sib pairs, and Tot_{fs} is the total number of full-sib pairs in the true pedigree. This statistic equals zero when all members of the population are in different families and one when the population structure is reconstructed exactly. Since the statistic actively penalizes accuracy when unrelated individuals are reconstructed as full sibs (type I errors), it may become negative in poorly reconstructed populations.

The simulations: Simulation was used to compare the properties of heritability estimates made using the reconstructed pedigree approach with those of the pairwise approaches. Phenotypic data for full-sib data sets were generated using the infinitesimal model (Bulmer 1980). An individual's phenotype was equal to

$$Y_{ij} = \left(\frac{a_{\text{im}} + a_{\text{if}}}{2}\right) + N\left(0, \frac{\sigma_{\text{A}}^2}{2}\right) + N(0, \sigma_{\text{E}}^2), \quad (7)$$

where Y_{ij} is the phenotypic value of sib j in family i, σ_{A}^2 is the additive genetic variance, σ_{E}^2 is the residual or environmental variance, and a_{im} and a_{if} are the breeding values of the parents simulated from an $N(0, \sigma_{\text{A}}^2)$ distribution. The phenotypic variance was set to 1; so $\sigma_{\text{A}}^2 = h^2$ and $\sigma_{\text{E}}^2 = 1 - h^2$. It was assumed that there was no common environmental correlation of sibs.

The simulations were run under different conditions: Marker information was varied, with populations simulated with 2, 3, 5, 8, and 10 equally frequent alleles at each of 10 loci; full-sib family sizes were drawn from a truncated (*i.e.*, no null class) Poisson distribution with parameters 2, 5, and 10; and populations with 100, 200, 400, and 800 individuals in total were simulated. Each set of conditions was run 250 times on independently generated random populations. Heritability was set to 0.5.

To test the robustness of the algorithm to reconstruct families, populations were simulated from a Po(5) distribution of family size, but different assumptions were used about this distribution during reconstruction, namely uninformative (where every family size is equally likely), Po(5), and Po(10).

Simulations were run on the populations with allele frequencies updated after every 2000, 5000, 10,000, 20,000 cycles, or not at all. The accuracy of the reconstruction statistic was also calculated and compared between each level of allele update.

In each set of simulations, MCMC iterations were continued for 1,400,000 cycles, a greater number than required for the leveling off of both mean family size and population likelihood.

Simulations were also undertaken in which paternal half-sib families were generated. It was assumed that some proportion of the simulated individuals had known mothers, whose genotype information was also available. Half-sib families were then reconstructed using a modified form of the MCMC algorithm that accommodated the known maternal genotype data. Different percentages, 0, 10, 20, 40, or 80% of missing maternal information, were simulated.

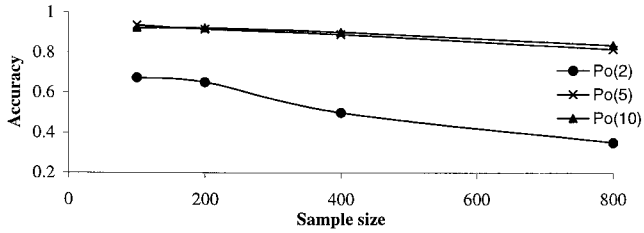

Figure 1.—The change in accuracy of family reconstruction with changing sample size, for the three simulated distributions of family size [Po(2), Po(5), and Po(10)]. Simulation conditions: 200 individuals, 10 marker loci with five alleles each and heritability 0.5.

Reconstructed sibships were used under an animal model to estimate the additive genetic and residual variances for the simulated trait, employing a standard package, ASREML (Gilmour *et al.* 1997). Heritability estimates were taken as the summary statistic. Heritabilities were also estimated by the pairwise approaches (Ritland 1996b; Lynch and Walsh 1998; Mousseau *et al.* 1998; Thomas *et al.* 2000). There are a number of forms of the likelihood technique, and in this study the procedure based on the difference in phenotype was used (Thomas *et al.* 2000). Results were compared in terms of the mean deviation of heritability estimates from the "best" achievable estimates (those estimated by REML from the true pedigree and the same quantitative data), which reflects bias, and mean squared errors (MSE), a composite statistic of bias and sampling variance over simulations.

RESULTS

Sample size: Figure 1 shows the change in the accuracy statistic as sample size increases for three distributions of family size. Accuracy decreases approximately linearly with an increase in the sample size. This is due to the increased chance that unrelated individuals have similar genotypes through random sampling and may be compensated for by increasing the marker information. The accuracy for the Po(2) distribution of family size was much less than the accuracy of the Po(5) or Po(10) graph, reflecting much poorer reconstruction of pedigrees. This is discussed below.

Consider first the results for the Po(5) distribution. Figure 2aii shows the mean deviation of heritability estimates obtained using marker-based approaches from those using the known pedigrees (the zero line). Estimates using the reconstructed populations deviate less

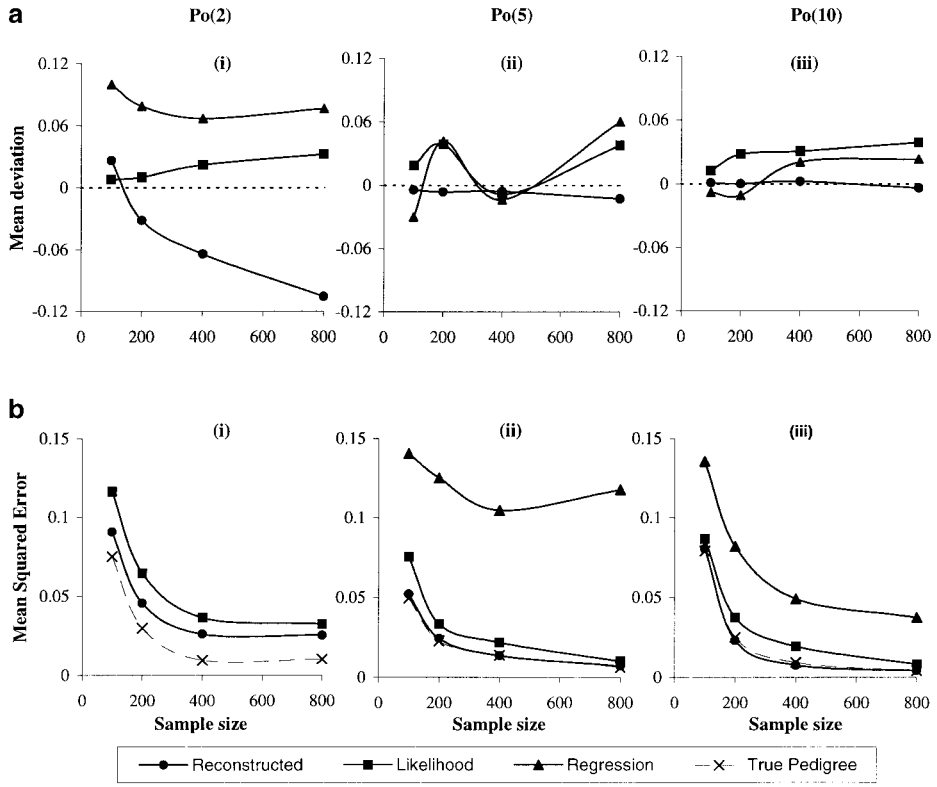

Figure 2.—Results for full-sib family simulations with 10 marker loci with five alleles each, heritability 0.5, and varied numbers of individuals in the sample for the three simulated distributions of family size. (a) The change in mean deviation of marker-based heritability estimates from estimates made using actual pedigrees (zero line) with changing sample size for the three simulated distributions of family size. (b) The change in mean squared error of heritability estimates with changing sample size. Columns i, ii, and iii refer to family size distributions Po(2), Po(5), and Po(10). Values for the MSE of regression-based estimates of Figure 2bi are off the scale (see text).

from the true pedigree estimates than pairwise estimates and show trivial negative bias. The size of the negative bias increases in a roughly linear manner as sample size increases (and hence also increases linearly with the accuracy statistic). This is probably due to the splitting of large families into two or more smaller ones during the sibship reconstruction procedure, which reduces estimates of the variance between families and thereby heritability estimates (Falconer and Mackay 1996). The pairwise techniques share the same trends across the sample sizes, a result that possibly reflects the similar manner in which they weight family information, using size rather than information content.

A more important measure of performance of the techniques is summarized in Figure 2bii, which displays the change in MSE across the range of sample sizes. In all cases, MSE is dominated by the sampling variance, rather than the bias, indicating that any bias is trivial compared with the level of precision of the techniques. Confirming previous results (Thomas *et al.* 2000), the regression procedure has much larger MSE than the pairwise likelihood approach and has a slower decline in value than other techniques as sample size increases, indicating a less efficient technique. The pairwise likelihood procedure has MSEs ~50% greater than those of the reconstructed pedigree, which are virtually indistinguishable from those of the true pedigree (the small difference being explained by the downward deviation seen in Figure 2aii). MSE is approximately inversely proportional to the sample size for all the techniques except for the regression procedure.

Family size: Simulations run using different distributions for family size showed similar trends to those obtained for families simulated with a Po(5) distribution, with those for the Po(10) distribution being virtually identical. Pairwise techniques showed more consistent mean deviations in heritability estimates across the range of sample sizes with small mean family size [Po(2); Figure 2ai] than with larger family size distributions (Figure 2, aii and aiii). This is because information for variance component estimation from a population in which families are small comes mainly from pairs of individuals, rather than larger groups. The downward bias in estimates obtained using reconstructed pedigrees with Po(2) family sizes is due to an increase in the number of type I errors, which at sample size 800 make up about a quarter of the number of pairs assigned as full sibs. A greater amount of marker information would be required to increase the accuracy of reconstruction and reduce this bias in estimates. Figure 2bi shows that the MSE of reconstructed pedigree estimates is smaller than that of the likelihood-based estimates, with one-third of the MSE being explained by the bias at sample size 800. Sample variances for the reconstructed pedigree estimates are two-thirds those for the likelihood procedure.

With a Po(10) family size distribution, estimates using reconstructed pedigrees show almost no deviation from those using actual pedigrees (Figure 2, aiii and biii). This indicates that few type I errors are made during pedigree reconstruction. Exclusions due to incompatible genotypes become more frequent with larger family

sizes. Therefore, at smaller family sizes there is a lower chance of incorrect families being detected than at larger family sizes, leading to greater numbers of type I errors, reducing accuracy (Figure 1), and increasing bias (Figure 2ai).

Previous results (Ritland 1996b; Thomas *et al.* 2000) indicate that pairwise procedures depend on there being sufficient variance of relatedness to be effective (*i.e.*, they require that there be adequate numbers of groups of relatives within the sample). The simulation results supported this, with the regression-based procedure having extremely large MSE when actual variance of relatedness is low [with Po(2)] and smaller MSE with larger actual variance of relatedness [with Po(10)]. The MSE values for the regression-based procedure were not plotted on Figure 2bi since they were off the scale (0.53, 0.34, 0.38, and 0.29 for sample sizes 100, 200, 400, and 800, respectively). Less dramatic improvements in MSE were noted in the likelihood-based procedure, where prior information on population structure compensates in part for less actual variation in relatedness.

Marker data: Figure 3a shows the change in the accuracy as the amount of marker information is varied, simulated by changing the number of alleles at each locus. Accuracy improves at a diminishing rate with increasing allele number, with little difference in accuracy between 6 and 10 alleles per locus. At the minimum number of alleles per locus (two), mean accuracy is ~ -0.2, reflecting a large number of type I errors (Equation 6) and resulting in large downward bias in heritability estimates. Figure 3b illustrates this point with the largest mean deviation of estimates occurring with low allele numbers. With the exception of low marker information (<5 alleles per locus), estimates made using reconstructed pedigrees are closer to true pedigree estimates than using either pairwise technique. At low marker information, the likelihood procedure shows least mean deviation from the true pedigree estimates.

Figure 3c shows the change in MSE with allele number. Again, the regression procedure shows the largest MSE and sampling variances of parameter estimate. Deviations in the MSE of estimates using reconstructed pedigrees from those using the true pedigree were almost entirely explained by the bias (indicated by mean deviation). Since mean deviation for the likelihood procedure is also small (Figure 3b), its MSE is higher than that for the true pedigree due to sampling variance, and hence estimates made using the likelihood procedure have lower precision.

Assumed distribution of family sizes: Table 1 summarizes the change in accuracy, mean deviation, and MSE when different assumptions are made about the family size distribution. Accuracy is lowest when an uninformative distribution for family size (*i.e.*, every family size is equally likely) is assumed. Despite this, the mean difference between heritability estimates determined using pedigrees reconstructed with uninformative family size distributions and correct pedigrees is very small.

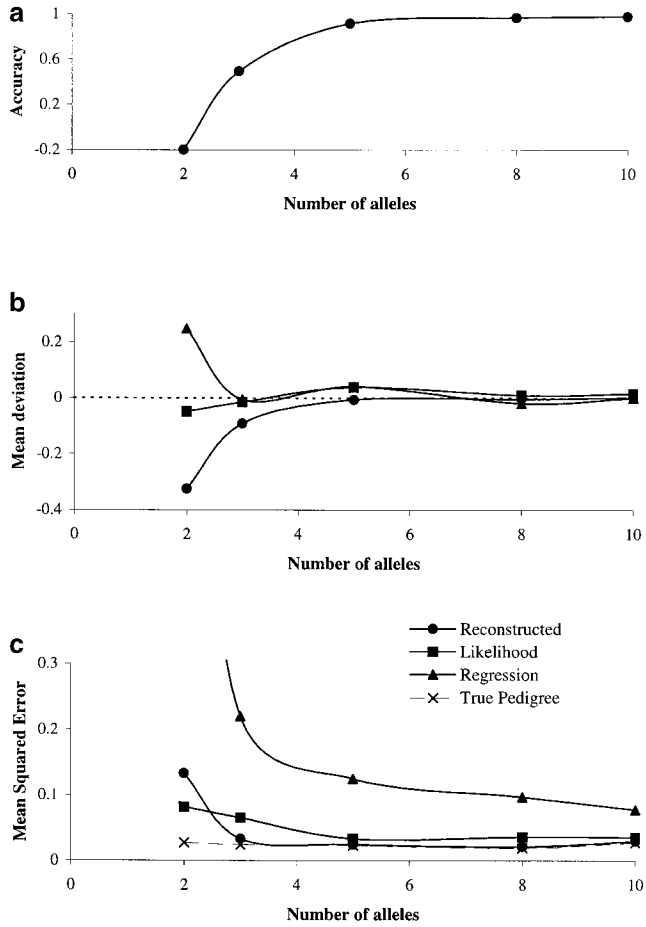

Figure 3.—Results for full-sib family simulations with 200 individuals, 10 marker loci, heritability 0.5, actual family size distribution Po(5), assumed distribution for sibship reconstruction Po(5), and varied numbers of alleles per locus. (a) The change in accuracy of family reconstruction with changing numbers of alleles. (b) The change in mean deviation of marker-based heritability estimates from estimates made using actual pedigrees (dotted line) with changing numbers of alleles. (c) Change in mean squared error of heritability estimates with changing numbers of alleles.

TABLE 1

Simulation results when different family size distributions are assumed during pedigree reconstruction (the same populations were reconstructed in each case)

Distribution	Accuracy (var)	Mean deviation	MSE
True pedigree	1 (0)	—	0.0260
Uniform	0.848 (0.003)	−0.0111	0.0296
Po(5)	0.911 (0.002)	−0.0056	0.0266
Po(10)	0.943 (0.001)	−0.0229	0.0265

Simulated populations contained 200 individuals with the true family size distribution being Po(5). Ten loci with five equally frequent alleles were simulated. Heritability was set at 0.5. Mean deviation is the average deviation of the estimated parameter from the REML-derived estimate using correct pedigree information. var, variance.

TABLE 2

Simulation results when parental allele frequencies were estimated after different numbers of cycles

Method	Accuracy (var)	Mean deviation	MSE
True pedigree	1 (0)	—	0.0492
Not recalculated	0.518 (0.015)	−0.0608	0.0735
20,000 cycles	0.544 (0.014)	−0.0677	0.0717
10,000 cycles	0.545 (0.012)	−0.0651	0.0724
5,000 cycles	0.552 (0.014)	−0.0503	0.0665
2,000 cycles	0.545 (0.016)	−0.0570	0.0669

Simulation conditions: 100 individuals, five marker loci with five alleles each, heritability 0.5, actual family size distribution Po(5), and assumed distribution for sibship reconstruction Po(5). var, variance.

Moreover, there is little increase in the MSE of these estimates, indicating only a little loss in precision. Using the correct distribution of family sizes, in this case Po(5), then accuracy and estimates are improved slightly, with MSE being almost identical to that of the true pedigree.

Accuracy is improved further if a Po(10) distribution is assumed, even though the true distribution is Po(5). This is because comparatively larger weights are placed on larger family sizes, thereby reducing the problem of large families being split into smaller families. However, if marker information is low, so that the probability of full-sib triplet exclusion due to incompatible genotypes is low, then increasing the weights of larger families can result in large numbers of incorrectly grouped individuals. As mentioned previously, this causes larger bias in estimates of heritability than related pairs being classed as unrelated.

Updating allele frequencies: Table 2 summarizes the simulations investigating the recalculation of parental allele frequencies. Results show that there is some improvement in accuracy and in parameter estimates as the number of reestimations of allele frequencies is increased. It would be expected that such allele reestimation of allele frequencies would have a greater effect in small populations, where the variance in family size is large, since under these conditions the weights placed on allele counts from each family would be most incorrect. In such cases, allele frequencies in the offspring generation might poorly represent allele frequencies in the parent generation. In larger populations, especially those with small family sizes, allele frequencies are more constant between generations (Falconer and Mackay 1996).

Half sibs: Figure 4a shows the accuracy of half-sibship reconstruction when different percentages of mothers and their marker information are known. As expected, when the amount of maternal marker information decreases, the accuracy also decreases. Again, this drop is closely followed by a downward deviation in heritability estimates (Figure 4b). The regression-based procedure makes no use of the maternal genotype information and so shows the same trend (a downward deviation of ∼−0.1 to −0.2) across Figure 4b. The likelihood

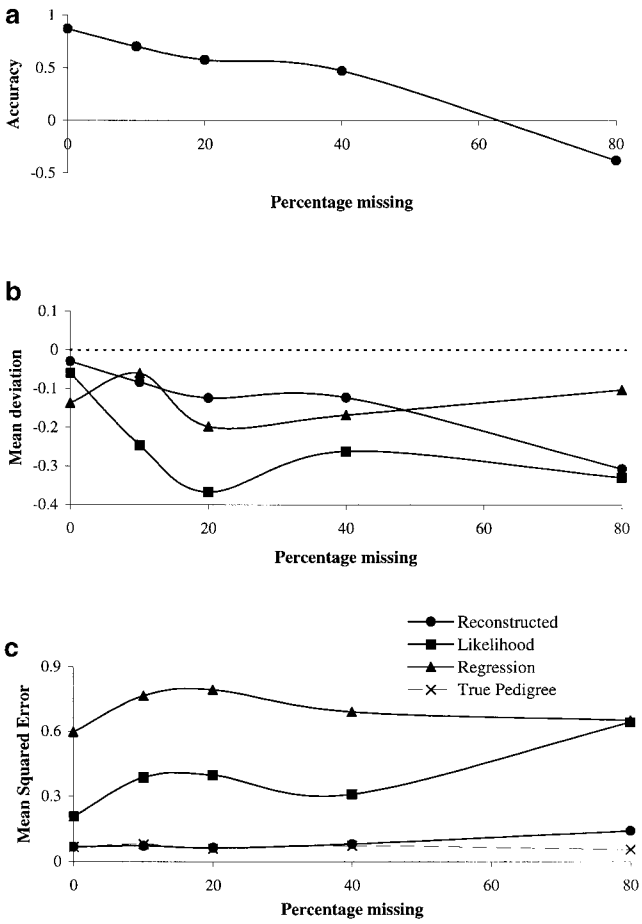

Figure 4.—Results for half-sib family simulations with 200 individuals, 10 marker loci each with five alleles, heritability 0.5, actual family size distribution Po(5), assumed distribution for sibship reconstruction Po(5), and varied percentages of maternal information known. (a) The change in accuracy of family reconstruction with changing percentage of known maternal information. (b) The change in mean deviation of marker-based heritability estimates from estimates made using actual pedigrees (dotted line) with changing percentage of known maternal information. (c) Change in mean squared error of heritability estimates with changing percentage of known maternal information.

TABLE 3

Percentage confidence levels, determined by simulation, for the accuracy of families of size 3 and 4

Family size	Number	% true full sibs	% correct size
3	1819	98	56
4	1717	99	78

Simulation conditions: 200 individuals, 10 marker loci with five alleles each, actual family size distribution Po(5), and assumed distribution for sibship reconstruction Po(5).

procedure can be easily modified to incorporate parental genotype information, and so heritability estimates improve with increased maternal genotype information. With low amounts of maternal genotype information, heritability estimates are biased downward.

There is a large MSE associated with the regression-based procedure (Figure 4c), mainly due to sampling variance rather than bias (Figure 4b). The likelihood procedure shows increasing MSE as the percentage of maternal genotype information falls, more than can be explained by the downward bias of the procedure and indicating a reduction in precision of the estimator. Estimates of heritability using the reconstructed pedigree have higher MSE than estimates using the known pedigree, although this difference may be explained by the downward trend shown in Figure 4b.

Confidence levels: Simulations of 250 populations of size 200 were used to estimate the percentages of families of sizes 3 and 4 (numbers chosen as examples) that were reconstructed correctly. In each case, two quantities were determined: the percentage of reconstructed families comprising only true full sibs and the percentage that were actually of that size, rather than a subset of a larger family. A Po(5) distribution of family size was assumed in the simulations and marker information was set at 10 loci with five alleles each.

More families of size 3 than 4 were reconstructed, although families of size 4 were expected to be more frequent (Table 3). This is because the procedure tends to split larger families, which is reflected in the lower confidence that the families reconstructed as size 3 were actually of size 3. Simulations also show that reconstructed families of size 4 are more likely to be a genuine collection of full sibs because of the relatively greater chance that an incorrect group of size 4 is excluded through incompatible marker information. Figure 5 shows the distribution of the actual sizes of families that were split to give reconstructed families of sizes 3 and 4. Of particular note is the drop in the second point of each curve relative to the rest of the curve, which is due to the low likelihood placed on a family of size 1 under a Poisson distribution of family sizes. For example, a family of size 4 is unlikely to be split into a family of 3 and another of 1, due to the low probability of observing a family of size 1, while a family of size 5 may be more easily split into families of 3 and 2.

DISCUSSION

Monte Carlo Markov chain procedures to reconstruct sibships from a single generation of a population provide an improved means of estimating variance components compared to earlier techniques. Reconstructing the pedigree in this manner recovers in part some of the family-specific weights lost in pairwise techniques, resulting in more efficient use of the information and lower mean-squared errors in parameter estimates. Moreover, since pedigrees are then assumed known, traditional procedures for partitioning the variance can be used, facilitating the incorporation of additional effects into the model or the use of multivariate analysis on data collected from several traits. The sibship reconstruction process is independent of the quantitative data, and so actual values for the genetic parameters should not affect the technique's accuracy in estimating those parameters. For this reason, simulations examining the effects of the actual level of heritability were not run. A final attractive feature of these procedures is that allele frequencies in the population can be estimated more precisely than as the simple mean among animals in the sample analyzed.

Since the Markov chain depends on the calculation of likelihoods, it is relatively straightforward to incorporate additional information, for example, maternal genotype information, year of birth, or, in the case of plants, separation by distance provided a suitable dispersion parameter is known. This ease of modification allows the incorporation of possible genotyping errors into the algorithm. Providing the probability of incorrectly typing a genotype (which may be done overall, or on a locus- or allele-specific basis) can be estimated prior to running the algorithm. Equation 4 could be modified to still sum over all parent allele combinations, but now allowing each of these alleles to change with some probability on the basis of the probability of a mistyped locus. As this would slow the algorithm considerably, some

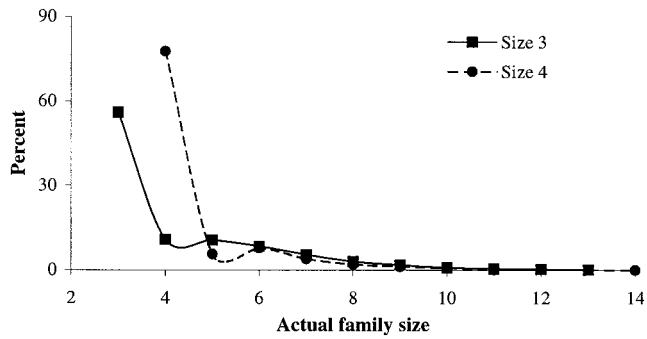

Figure 5.—The distribution of the actual size of families reconstructed as being of sizes 3 and 4.

assumptions restricting the number of transitions allowed may be required. Since mistyped alleles are more likely to cause families to be split rather than incorrect families to be formed, however, the present algorithm can cope with low levels of mistyping without modification, as only a small bias in variance component estimates is introduced by this type of error. For example, a common form of genotyping error is the mistyping of heterozygotes as homozygotes due to the failure of one allele to amplify, so a small number (100) of populations were simulated to examine this type of error. Ten percent of the loci simulated in these populations were mistyped in the above manner and this increased (by about 0.1) the downward bias in heritability estimates. MSE also increased, but remained less than with pairwise techniques.

There is interaction between sample size and the amount of marker information required to accurately reconstruct the families. With large sample sizes, the probability of obtaining type I errors increases, and more marker information is required to counteract this effect. Further investigation is required to determine the extent of the interaction and to investigate the balance between the collection of individual data and the amount of marker data genotyped.

When compared to previous techniques, this new approach performs admirably well, in many cases having lower mean deviation from the best available estimator, calculated from the known pedigree, and lower mean bias from the true parameter. In addition, it yields mean squared errors that are often almost indistinguishable from those of the known pedigree, and as the MSE in most cases is dominated by sampling variance, any biases in parameter estimates become trivial.

There are a number of areas where caution must be taken when using relationships based on marker information to infer parameters. For example, in populations that are not in linkage equilibrium, the information from each locus is not independent. Instead, the likelihood of the marker data in any putative full-sib family must be calculated from the probability of observing parental genotypes across all loci simultaneously rather than individually.

A second area for caution is in using reconstructed sibships to determine other parameters such as the average size of families or the distribution of family sizes, which might be used in studies of reproductive success or other life history traits (Stearns 1992). Reconstructed sibships have a tendency to underestimate mean family size and do not give an accurate description of its distribution. For example, the reconstructed pedigrees examined to determine the confidence levels for families of sizes 3 and 4 showed that more families of size 3 were reconstructed than of size 4 (Table 3), even though the latter were expected to be more common. Family sizes are underestimated due to the low probability of breaking down a correctly grouped set of individuals to join it to another group. For example, a family of size 6 may be reconstructed initially as two families of size 3, but due to the low probability of moving through smaller family sizes, be unable to combine into the correct single family. To combat this problem, a step might be added to the algorithm that, with some probability, periodically attempts to combine two entire families (thereby attempting a direct jump across a valley of the likelihood surface) and/or break up an entire family (although this might prevent mean family size or population likelihood from stabilizing). Results indicated that the use of an incorrect distribution of family size that increases the expected frequency of larger families [the reconstructions operating under a Po(10) distribution] had improved accuracy over reconstruction using the correct prior [in this case a Po(5) distribution]. However, such an approach to estimate mean family sizes and distributions is not advisable since it may cause family size to be overestimated, especially in populations with low marker information where exclusions based on incompatible genotypes are rare. In populations with ample marker information, exclusions often prevent large families from being formed incorrectly. Simulation to estimate the expected bias in family size results in a circular problem, since the distribution of family sizes required to simulate the families is unknown. However, it may be possible to use simulation using the same sample size, the same level of marker information, the estimated family size distributions, and the variance component estimates to estimate the size of bias shown in the variance components. This would require the assumption that any bias in subsequent variance component estimation approximately equals the bias in the original estimates, an assumption that may hold only if that original bias is small since variance components are bounded below by zero.

The choice of distribution of family size must also be considered cautiously for, as previously mentioned, assigning unrelated individuals to the same family can cause large downward bias in estimates of between-family variance and of genetic parameters derived from them. It is best, therefore, to choose a distribution that results in an underestimate of mean family size. Results indicate that when using an uninformative distribution of family sizes, the mean size of families in the reconstructed pedigree is consistently underestimated if the true distribution of family size is Poisson. This is because an uninformative distribution does not weight the creation of large families enough to break up two families of roughly the same size to recombine them as one larger family, even if they are actually one large family. The same problem occurs even when the correct distribution for family size is used, although to a lesser extent.

There are ways that the algorithm itself might be improved, leading to more likely population structures. These include the possibility of combining whole families or subdividing an entire family (perhaps that with

the lowest likelihood) as mentioned above. An alternative approach to the population mixing that could speed up the algorithm but would not lead to better solutions would be to treat the individuals systematically, moving first individual 1 to a random family, then individual 2, etc., rather than selecting and moving individuals completely at random. In addition, the optimum acceptance / rejection rule for each change in the MCMC iteration could also be considered.

In natural populations, there are more than two classes of relationship, and in addition, full-sib and half-sib groups are unlikely to be completely independent, perhaps being full cousins. There are (at least) two approaches to the problem of dealing with multiple relationships: one approach is to assume that, since most of the information on heritability would come from close relatives, only these classes need to be considered (*e.g.*, assume only full-sib and unrelated individuals are present, and ignore half sibs, cousins, etc.). The robustness of these techniques to deviations from the assumption of two classes of relationship is a complex problem and worthy of further investigation. Another approach is to attempt to include other classes of relationship into the model. Of particular interest are extensions to these techniques that allow nested maternal full-sib families within paternal half-sib families to be reconstructed. This is achieved through modification of Equation 4 to multiply across the likelihood of the maternal half-sib families given a particular paternal genotype (see appendix). Mixing would then move individuals between different mothers as well as fathers. However, the number of potential family structures would be extremely large and possibly intractable, even when using Markov chain approaches. In addition, calculation of the likelihoods for individual families would be slow. Moreover, the ability to distinguish between relationships falls quickly with increased distance of relationship, and extremely large amounts of marker information, or some known relationships to build upon (*e.g.*, known mothers), would be required to reconstruct accurate pedigrees (Thompson 1975). Furthermore, incorrectly assigned relationships would bias estimates of variance components to an unknown extent. For example, assigning groups of full sibs as half sibs would bias heritabilities upward, since a larger similarity in phenotype is attributed to smaller familial relationship.

Shortcut methods and assumptions would need to be applied to make more complex situations tractable. For example, if information is available on two or three nonoverlapping generations of a population, sibships could be reconstructed for each generation, constraining the sum of possible parental genotypes using the probability (if known) that a parent is contained within the samples collected from previous years. Generations could then be linked using the likelihood of the observed marker data and the probability that one or both parents are from the previous generation.

We thank Prof. Nick Barton for suggestions and Dr. Jinliang Wang and the referees for constructive comments on the manuscript. Stuart Thomas was funded by a Biotechnology and Biological Sciences Research Council PhD studentship.

LITERATURE CITED

Bulmer, M. G., 1980 *The Mathematical Theory of Quantitative Genetics.* Oxford University Press, New York.

Dillon, W. R., and M. Goldstein, 1984 *Multivariate Analysis Methods and Applications.* John Wiley and Sons, New York.

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

Gilmour, A. R., R. Thompson, B. R. Cullis and S. J. Welham, 1997 *ASREML Manual.* New South Wales Department of Agriculture, Orange, Australia.

Kuhner, M. K., J. Yamato and J. Felsenstein, 1995 Estimating effective population-size and mutation-rate from sequence data using Metropolis-Hastings sampling. Genetics **140:** 1421–1430.

Lande, R., 1982 A quantitative genetic theory of life history evolution. Ecology **63:** 607–615.

Lande, R., and S. Shannon, 1996 The role of genetic variation in adaptation and population persistence in a changing environment. Evolution **50:** 434–437.

Larget, B., and D. L. Simon, 1999 Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Mol. Biol. Evol. **16:** 750–759.

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

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

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

Mousseau, T. A., and D. A. Roff, 1987 Natural selection and the heritability of fitness components. Heredity **59:** 181–197.

Mousseau, T. A., K. Ritland and D. D. Heath, 1998 A novel method for estimating heritability using molecular markers. Heredity **80:** 218–224.

Norris, J. R., 1997 *Markov Chains.* Cambridge University Press, Cambridge, United Kingdom.

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

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

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

Stearns, S. C., 1992 *The Evolution of Life Histories.* Oxford University Press, New York.

Storfer, A., 1996 Quantitative genetics: a promising approach for the assessment of genetic variation in endangered species. Trends Ecol. Evol. **11:** 343–348.

Thomas, S. C., J. M. Pemberton and W. G. Hill, 2000 Estimating variance components in natural populations using inferred relationships. Heredity **84:** 427–436.

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

Yang, Z., and B. Rannala, 1997 Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo method. Mol. Biol. Evol. **14:** 717–724.

Communicating editor: J. B. Walsh

APPENDIX: MODIFIED FORMS OF LIKELIHOOD EQUATION 4

A constrained version: A version of Equation 4 that is faster to compute may be obtained by constraining the possible parental genotypes. An allele is assigned to each parent by selecting an offspring at random from the

putative full-sib family and assigning one of its alleles to one parent and other allele to the remaining parent. If the randomly selected offspring has genotype (w, y), then the likelihood of the genotypes within a full-sib family may be expressed as

$$L_{\text{genotypes}} = \prod_{l} L_l, \tag{A1}$$

where

$$L_l = \partial_{wy}\left[\sum_{x=1}^{b_l}\sum_{z=x}^{b_l} c \cdot p^1_{wx} \cdot p^2_{yz} \prod_{i=1}^{n_f} L(g_{il})\right] + (1 - \partial_{wy})\left[\sum_{x=1}^{b_l}\sum_{z=x}^{b_l} c \cdot p^1_{wx} \cdot p^2_{yz} \prod_{i=1}^{n_f} L(g_{il})\right] \tag{A2}$$

and

$$c = 8/2^{(\partial_{wx} + \partial_{yz} + \partial_{*})}. \tag{A3}$$

L_l is the likelihood of an individual locus, indexed by l; ∂_{wx}, ∂_{wy}, and ∂_{yz} are indicator variables with, for example, $\partial_{wx} = 1$ when allele w is the same as allele x, $\partial_{wx} = 0$ otherwise; b_l is the number of alleles at locus l; x and z index unconstrained parental alleles; c is a term that adjusts the frequency of ordered genotypes to unordered genotypes; i indexes an individual from the putative family; p^1_{wx} is the ordered genotype frequency of parent 1; p^2_{yz} is the ordered genotype frequency of parent 2; $L(g_{il})$ is the likelihood of observing the genotype of individual i at locus l given the parental genotypes; ∂_{*} is also an indicator variable, with $\partial_{*} = 1$ when the unordered genotype of parent 1 is the same as the unordered genotype of parent 2 and $\partial_{*} = 0$ otherwise. For example, in the calculation of c, when the parental genotypes are (1, 2) and (3, 4), $c = 8$; when (1, 1) and (2, 3), $c = 4$; when (1, 1) and (2, 2), $c = 2$; and when (1, 2) and (2, 1) $c = 4$, etc.

Including half sibs: Equation 4 may be modified to calculate the likelihood of the genotypes within nested families, with maternal families within paternal families, by multiplying across the likelihood of the maternal families within the paternal family,

$$L_{\text{genotypes}} = \prod_{l}\sum_{w=1}^{b_l}\sum_{x=1}^{b_l} p^1_{wx}\left[\prod_{m=1}^{n_f}\sum_{y=1}^{b_l}\sum_{z=1}^{b_l} p^2_{yz}\left(\prod_{i=1}^{n_m} L(g_{iml})\right)\right], \tag{A4}$$

where variables are the same as Equations A1–A3, except here n_f is the number of full-sib families within the half-sib family, m indexes the full-sib family, n_m is the number of full sibs within the full-sib family, and i indexes the offspring number. In this case, p^1 is male and p^2 is female.

If there is only one maternal family within each paternal family, $n_f = 1$ and (A4) reduces to (4). If there is only one offspring per maternal family, then (A5) is equivalent to an analysis on half-sib families only. If maternal genotype information is available, then that part of the likelihood equating to the sum over the possible maternal genotypes may be removed. The likelihood of the observed offspring genotype would then be calculated given the known maternal and all possible paternal genotypes.