Genet. Res., Camb. (2002), **79**, *pp.* 227–234. *With 2 figures.* © 2002 Cambridge University Press
DOI: 10.1017/S0016672302005669 Printed in the United Kingdom

Sibship reconstruction in hierarchical population structures using Markov chain Monte Carlo techniques

STUART C. THOMAS* AND WILLIAM G. HILL
Institute of Cell, Animal and Population Biology, University of Edinburgh, West Mains Road, Edinburgh EH9 3JT, UK

(*Received 23 January 2002*)

Summary

Markov chain Monte Carlo procedures allow the reconstruction of full-sibships using data from genetic marker loci only. In this study, these techniques are extended to allow the reconstruction of nested full- within half-sib families, and to present an efficient method for calculating the likelihood of the observed marker data in a nested family. Simulation is used to examine the properties of the reconstructed sibships, and of estimates of heritability and common environmental variance of quantitative traits obtained from those populations. Accuracy of reconstruction increases with increasing marker information and with increasing size of the nested full-sibships, but decreases with increasing population size. Estimates of variance component are biased, with the direction and magnitude of bias being dependent upon the underlying errors made during pedigree reconstruction.

1. Introduction

Knowledge of the relationships in a population is important in a number of areas of genetics. These include studies of reproductive success, the estimation of the parameters describing quantitative traits (Falconer & Mackay, 1996), examination of population dispersion, and in conservation through the reduction of inbreeding (Storfer, 1996). In some populations, however, the relationships are unknown and so study is limited. Molecular markers provide a means to estimate relationships and so provide a way to circumvent the problems of limited information on ancestry. A number of methods have been developed that estimate relationships using information on molecular markers. Broadly speaking these may be divided into three categories:

1. *Methods that estimate relatedness*: a continuous measure, describing the 'genetic distance' between the individuals based on marker similarity. There are several basic measures of relatedness, mostly pair-wise in nature (e.g. Queller & Goodnight, 1989; Ritland, 1996; Lynch & Ritland, 1999), although these may be extended for group analysis.
2. *Pair-wise likelihood approaches*: these calculate the likelihood for a pair falling into alternate relationship classes, e.g. full-sib, half-sib or unrelated (e.g. Thompson, 1975), and again these may be extended to analyse groups.
3. *Family construction using Markov chain Monte Carlo* (*MCMC*) *approaches*: these assign a series of specific relationships to all the individuals, using likelihood-based MCMC procedures (Hastings, 1970) to generate a plausible set of relationships for the group (Thomas & Hill, 2000; Smith *et al.*, 2001). These methods were developed as a modification to (2) above, specifically aimed at reconstructing whole families rather than pairs, and attempt to find the most likely set of relationships without searching through the prohibitively large number that could be assigned (as would happen if (2) were extended to compute the likelihood of all possible family assignments). An alternative approach to family reconstruction was outlined by Almudevar & Field (1999), who generated plausible sibships by excluding groups that were impossible due to incompatible genotypes. The plausible sibships were then partitioned into a putative (single-generation) population based upon a score function.

The MCMC approaches have mainly been investi-

* Corresponding author. Tel: +44 (0)131 650 5440. e-mail: sthomas@srv0.bio.ed.ac.uk

gated for use in populations comprised of full-sib families, and are currently limited to such situations. In this study the methods of reconstruction outlined by Thomas & Hill (2000) are extended to hierarchical (or half-sib) population structures (i.e. full-sibships nested within half-sib families, e.g. polygamous males and monogamous females). It is an unfortunate fact that the more distant the relationship, the poorer the estimated relationship information becomes and the greater the requirement for more marker information (Ritland, 1996). Moreover as the number of relationship classes increases it becomes harder to assign individuals correctly. The introduction of more distant categories of relationship may therefore destabilize the reconstructed populations. The hierarchical structure seems the most tractable after the one-way structure of solely full- or half-sib families, both of which are limiting cases.

Incorporation of a half-sib category of relationship into the reconstruction allows greater dissection of the parameters underlying the trait of interest, with the estimation of additive genetic variance (σ^2_A), environmental variance of full-sibs (σ^2_C) and the environmental variance (σ^2_E) being made possible, assuming there is no confounding by dominance or epistasis, whereas in the one-way structure only two components, σ^2_A and σ^2_E, can be estimated. Inaccurate reconstruction of sibships leads to bias in the estimates of variance components. Thomas & Hill (2000) noted that, if full-sibs are assigned as unrelated, trivial downwards bias is introduced in estimates of σ^2_A (unless large numbers are incorrectly assigned) while assigning relatedness to unrelated individuals leads to much larger bias. The direction and magnitude of bias introduced through other forms of mis-assignment will depend on the particular error made during reconstruction.

The objectives of this study are to: (i) develop a method for family reconstruction in the hierarchical case and, as a necessary component, provide an efficient method for calculating the likelihood of the observed marker information for hierarchical family structures (see Appendix); (ii) examine the effects of population size, population structure and the amount of marker data available on the accuracy of reconstruction of half-sib and hierarchical structures; and (iii) use reconstructed pedigrees to estimate the variance parameters underlying a quantitative trait.

2. Statistical techniques

The basic methods are given by Thomas & Hill (2000), and only the relevant modifications are presented here. In the algorithm presented previously, individuals were mixed in a sib/non-sib structure, thus a method to mix individuals between nested families was included. Individuals were mixed in order (rather than randomly), and the candidate individual was either moved with probability one-half to a randomly selected half-sib family or remained in the same family. It was then either moved to an existing full-sib family within the chosen half-sibship or formed a new full-sib family, with an equal chance of assignment to each existing full-sib family or a new one.

Because full-sib families have greater resolving power, they were found to be generated in preference to half-sib families, so half-sib families were often split into their component full-sib families. In order to minimize this problem, periodically (e.g. every 200 cycles) an entire half-sib family was joined to another randomly selected half-sib family (with component full-sib families remaining separate). The same conditions for accepting and rejecting a change (see Thomas & Hill, 2000) as for single individual mixing were used for this type of mixing. Other rules for accepting or rejecting a change, based upon standard Metropolis–Hastings procedures, and allowing the inclusion of 'temperature parameters' controlling the probability a step is made, are outlined by Smith *et al.* (2001). Theoretically there is no difference between the final structure obtained under the different rules.

The likelihood (L_g) of the observed genotypes is calculated as:

$$L_g = \prod_{\ell} \left[\sum_{w=1}^{b_\ell} \sum_{x=1}^{b_\ell} p_{wx} \right] \times \left[\prod_{m=1}^{n_f} \sum_{y_m=1}^{b_\ell} \sum_{z_m=1}^{b_\ell} p_{y_m z_m} \prod_{c=1}^{n_m} P(g_{c\ell}) \right]. \quad (1)$$

for a given putative family. b_ℓ is the number of alleles at locus ℓ; w and x index the paternal alleles and p_{wx} is the ordered genotype frequency of the father; n_f is the number of full-sibships within the half-sibship; y_m and z_m index the maternal alleles of full-sib family m and $p_{y_m z_m}$ is the ordered genotype frequency of the mother; n_m is the number of individuals in full-sib family m; and c indexes the individuals in that full-sib family. $P(g_{c\ell})$ is the probability of observing the genotype of c at ℓ given that one of its alleles is from the father and one is from the mother, and is derived from simple Mendelian sampling. Equation (1) reduces to the likelihood of observing these genotypes in a single full-sib family if $n_f = 1$ and in a half-sib family if $n_m = 1$ for all m. Calculations using (1) are slow, but are speeded up by fixing parental alleles using the offspring genotypes (see Appendix).

Variance components can then be estimated from the reconstructed pedigrees using standard REML methodology on the assumption that the pedigree is correct (see Thomas & Hill, 2000). In this study ASREML was used to estimate the components (Gilmour *et al.*, 1997).

Genotype data were generated for a number of different hierarchical populations, using standard rules

https://doi.org/10.1017/S0016672302005669 Published online by Cambridge University Press

of Mendelian inheritance and assuming that all alleles were co-dominant. Allele frequencies for reconstruction were estimated from the sample, and not updated during reconstruction (see Thomas & Hill, 2000). A number of different structures were simulated to investigate the inclusion of half-sibs into the sample, with the population size, family structure and amount of marker information each being varied. Each parameter set was replicated 100 times. Uninformative prior distributions were placed on constructed sibship sizes, so that every family size had equal probability. The accuracy of reconstruction was examined and is presented in the form P($a|b$), where P($a|b$) is the proportion of pairs assigned relationship a when their actual relationship is b. Comparison of these proportions and direct observation of the reconstruction allows inferences to be made about the splitting of families. For example, if P(fs|fs) is high, but P(hs|hs), P(fs|hs) and P(hs|fs) are low, then it can be inferred that half-sib families are being split into their component full-sib families.

Phenotypic data were generated using the infinitesimal model (Bulmer, 1980) as:

$$Y_{ijk} = \tfrac{1}{2}a_i + \tfrac{1}{2}a_{ij} + c_{ij} + a_{wijk} + e_{ijk} \quad (2)$$

where Y_{ijk} is the phenotypic value of individual k in dam family j nested in sire family i, a_i and a_{ij} are the breeding values of the sire and dam, sampled from $N(0, \sigma^2_A)$, c_{ij} is the family common environmental effect, sampled from $N(0, \sigma^2_C)$, a_{wijk} is the within-family deviation in breeding value of the individual, sampled from $N(0, \sigma^2_A/2)$, and e_{ijk} is the individual environmental deviation, sampled from $N(0, \sigma^2_E)$. Since variance component estimation is independent of sibship reconstruction, values of $\sigma^2_A = \sigma^2_C = 0{\cdot}25$ and $\sigma^2_E = 0{\cdot}5$ were used for all analyses. For each sample, variance components were estimated by REML using both the actual pedigree and the reconstructed pedigree. These were compared in terms of the bias and mean squared error of the estimates of heritability ($h^2 = \sigma^2_A/(\sigma^2_A + \sigma^2_C + \sigma^2_E)$) and environmental correlation of full-sibs ($c^2 = \sigma^2_C/(\sigma^2_A + \sigma^2_C + \sigma^2_E)$).

3. Results

(i) *Marker data and population size*

The accuracy of sibship reconstruction at different levels of marker data and at different population (i.e. sample) sizes is summarized in Fig. 1. Unsurprisingly, the accuracy of reconstruction increases with increasing marker data. In addition there is a clear interaction with respect to the accuracy of reconstruction between the amount of marker data and the size of the population that is sampled. At low levels of marker information (5 loci) and at small population sizes (50) about 50% of full-sib and half-sib pairs are correctly identified, with half-sib families often being split into their component full-sib families. As population size increases, the proportion of pairs assigned any relationship falls, and hence the percentage of related pairs assigned as unrelated increases, becoming close to 100% for P(ur|hs) when population size is large (400). In addition, there is a larger probability of unrelated pairs having similar genotypes when there are low levels of marker data, leading to greater numbers of incorrectly assigned relationships. The reconstruction of larger families would therefore become restricted due to an increased chance of incompatible genotype combinations. Similar trends are noted at higher levels of marker information, although P(fs|fs) and P(hs|hs) start at a higher level with low population sizes and fall at a slower rate with increasing family size. P(hs|hs) falls more rapidly than P(fs|fs) because it is harder to elucidate information about more distant relationships than close ones for a given amount of marker data. For P(fs|fs) there is a diminishing return to increasing marker information for a given population size: there is a larger increase between 5 and 10 loci than between 10 and 20 loci, although the increase in correct assignment with more loci becomes greater at larger population sizes. The returns for P(hs|hs) are much greater with each increment in locus number, especially with larger population sizes, reflecting the difficulty in correctly inferring more distant relationships. Therefore when there is little marker information, the reconstructed population is comprised of full-sibships rather than nested families.

At all levels of marker data and population size P(ur|ur) is close to 1 (not shown), with the lowest value being 0·978 for population size 50 with 5 marker loci. Thus P(fs|ur) and P(hs|ur) are low. Overall, the MCMC procedures are conservative in nature, with inaccuracies tending towards the assignment of a lower relationship than the true one: P(hs|fs) $\gg$ P(fs|hs), P(ur|fs) $>$ P(fs|ur) and P(ur|hs) $>$ P(hs|ur). Although not immediately obvious from Fig. 1, there is a slight increase in P(hs|fs) with increasing marker data. This reflects the larger (correct) half-sib families being reconstructed, so that any splitting of genuine full-sib families increases the proportion of P(hs|fs) rather than P(ur|fs) as occurs at lower levels of marker data.

(ii) *Population structure*

The accuracy of sibship reconstruction for different population structures, each of size 200, based on genotypes of 10 independent marker loci each with 5 equally frequent alleles is summarized in Fig. 2. Again the breakdown for genuinely unrelated pairs is not

https://doi.org/10.1017/S0016672302005669 Published online by Cambridge University Press

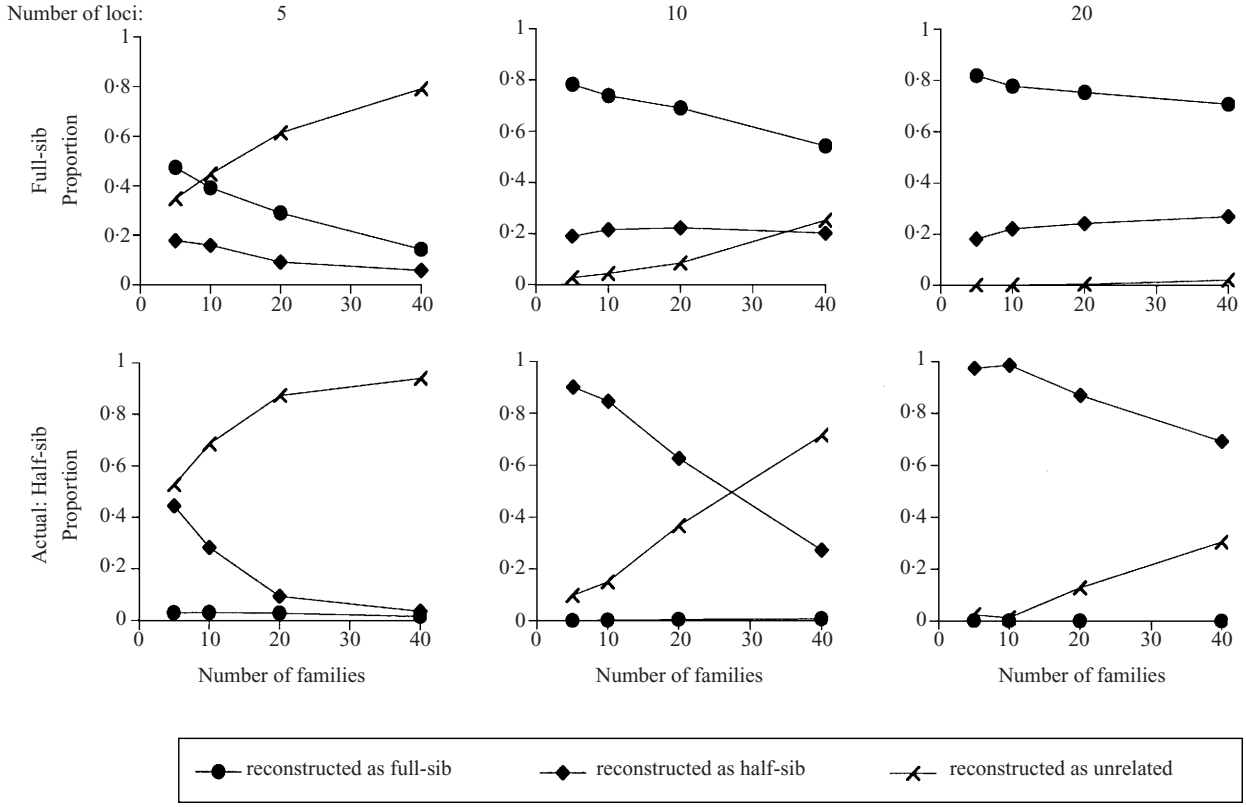

Fig. 1. Summary of the relationship assignment for different population sizes (simulated by changing the number of families, with each family comprising 2 dams per sire, and each dam having 5 offspring) and at different levels of marker data (5, 10 and 20 loci each with 5 equally frequent loci). Top row: assignment for pairs that are actually full-sibs. Bottom row: assignment for pairs that are actually half-sibs. Results for the same set of simulations appear in the same column.

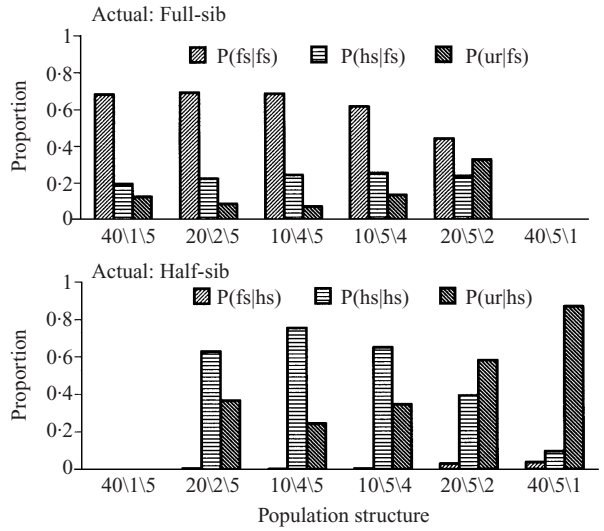

Fig. 2. Summary of the relationship assignment for different population structures. Population structure: no. of sires\no. of dams within sire\no. of offspring within dam. Upper chart: assignment for pairs that are actually full-sibs. Lower chart: assignment for pairs that are actually half-sibs. Results for the same set of simulations appear in the same vertical column.

shown, since P(ur|ur) is close to 1 in all cases. For specified full-sib family size, the breakdown of actual full-sib pairs into the assigned relationship categories is almost independent of the half-sib structure. P(fs|fs), however, falls as the size of the full-sib families falls, because the number of exclusions due to impossible genotype combinations is reduced. P(hs|fs) is largely unaffected by the change in the structure, whereas P(ur|fs) increases as no relationship rather than a half-sib relationship is assigned. P(hs|hs) also falls with decreasing full-sib family size (e.g. compare structure 20\2\5 with 20\5\2 and 10\4\5 with 10\5\4 in Fig. 2). This is probably because the combined set of marker data for groups of full-sibs yields more information about the distant relationship between the groups than the marker data used at an individual level. Thus accurate half-sib assignment would increase with P(fs|fs). When no full-sibs are simulated (e.g. structure 40\5\1 in Fig. 2) reconstruction is particularly poor, although further simulations show that accuracy does increase with more marker data (e.g. 20 to 30 loci; results not shown). P(fs|hs) is again low ($< 5\%$) for all population structures.

(iii) *Variance component estimation*

Estimates of heritability tended to show greater bias and larger MSE than estimates of the environmental correlation of sibs (Table 1). This was expected: the

https://doi.org/10.1017/S0016672302005669 Published online by Cambridge University Press

Table 1. *Summary of the bias* (*upper figure*) *and MSE* (*lower figure*, *italicized*) *of the estimates of heritability* ($\hat{h}^2$) *and environment correlation of full-sibs* ($\hat{c}^2$) *for different population structures*

Sample structure	Sample size	Known		Inferred	
		$\hat{h}^2$	$\hat{c}^2$	$\hat{h}^2$	$\hat{c}^2$
10\2\5\10	100	0·080	−0·064	0·130	−0·087
		0·092	*0·031*	*0·140*	*0·029*
20\2\5\10	200	0·053	−0·045	0·164	−0·106
		0·081	*0·023*	*0·127*	*0·031*
40\2\5\10	400	0·015	−0·012	0·114	−0·101
		0·058	*0·008*	*0·076*	*0·022*
10\4\5\10	200	0·057	−0·009	0·176	−0·040
		0·060	*0·019*	*0·090*	*0·056*
10\5\4\10	200	0·002	−0·014	0·017	−0·028
		0·040	*0·014*	*0·040*	*0·024*
20\5\2\5	200	0·035	0·007	−0·149	−0·189
		0·050	*0·019*	*0·041*	*0·043*
20\5\2\10	200	−0·026	0·010	−0·014	−0·130
		0·042	*0·017*	*0·047*	*0·029*
20\5\2\20	200	0·008	0·001	0·046	−0·024
		0·055	*0·024*	*0·057*	*0·025*

Sample structure: no. of sires\no. of dams within sire\no. of offspring within dam\no. of loci (each with 5 equally frequent alleles), in each case with $h^2 = c^2 = 0{\cdot}25$. Known: actual relationships used in the REML analysis. Inferred: constructed relationships used.

estimate of h^2 ($4 \times$ estimated correlation of half-sibs) has a high sampling error even when pedigrees are known because of the fourfold scaling of the estimate based on the component with least degrees of freedom. The estimate of c^2 (estimated correlation of full-sibs within half-sibs minus correlation of half-sibs) is not so scaled. Further, with reconstructed pedigrees the resolution of half-sib families is relatively poor, which exacerbates the errors of estimation of the half-sib correlation. Inspection of the reconstructed pedigrees showed that half-sib groups were often split into their component full-sib families (especially when full-sib family sizes are smaller: see above). With larger amounts of marker information and thus larger numbers of accurately reconstructed half-sibs, parameter estimates based on the MCMC approached the estimates derived from the known pedigrees. This was because, with higher levels of marker information, both half- and full-sib groups are reasonably accurately reconstructed. Paradoxically, when larger full-sib families are simulated biases tend to increase, even though smaller families are better reconstructed, as discussed above. This is probably because of the unpredictable nature of the biases introduced by incorrect relationship assignment. For example, with a population structure of 20\2\5 the assignment of genuinely full-sib pairs as half-sibs will result in an overestimate of h^2 and an underestimate of c^2, while with population structure 20\5\2 a larger proportion of full-sibs are assigned as unrelated and a larger proportion of half-sibs are assigned as either full-sibs or unrelated, which overall introduces negative bias to h^2 and c^2.

With smaller population sizes, but the same level of marker information, the populations are reconstructed more accurately, and thus MCMC-based estimates are closer to the estimates derived from the known pedigrees. At larger sample sizes the estimates deviate further from those of the known pedigrees, but this trend is superimposed upon an increase in the accuracy of estimates due to larger amounts of data available. There is therefore a trade-off between the sample size and the level of marker data required for optimal estimates of the underlying variance components.

Estimates of variance components were also compared with those obtained using a modified form of the pair-wise likelihood approach (Mousseau *et al.*, 1998; Thomas *et al.*, 2000). By altering the distribution of the pair-wise phenotypic difference for each relationship category to $N(0, \sigma_A^2 + 2\sigma_E^2)$ for full-sibs, $N(0, (3/2)\sigma_A^2 + 2\sigma_C^2 + 2\sigma_E^2)$ for half-sibs and $N(0, \sigma_A^2 + 2\sigma_E^2)$ for unrelated pairs, maximum likelihood estimates for the three parameters may be determined. The pair-wise likelihood approach is more dependent upon the amount of marker information than the MCMC method and exhibits large bias and MSE when marker information is lowered. At higher levels of marker information bias is sometimes smaller than with MCMC but the MSE remains large, indicating that estimates are much less reliable.

4. Discussion and conclusions

The reconstruction of hierarchical full-sib within half-sib families is possible using MCMC approaches. However, half-sib families tend to be split into their component full-sib families unless the amount of marker data used is large or full-sib families are large. Caution must be adopted, therefore, in the use of these reconstructed families in subsequent studies (e.g. studies of lifetime reproductive success) since the sizes of both full-sib and half-sib reconstructed families are biased downwards. Thomas & Hill (2000) commented that using reconstructed sibships to estimate variance parameters describing a quantitative trait was feasible, since the nature of the errors made in reconstruction were conservative and introduced only a slight (although definite) downwards bias in heritability estimates, but they considered only one-way classifications with solely full-sibs or solely half-sibs present and partitioned the variance into only two components. In analyses of hierarchical structures when

https://doi.org/10.1017/S0016672302005669 Published online by Cambridge University Press

additional variance components are estimated, however, bias is introduced through other forms of incorrect relationship assignment. The exact direction and magnitude of the bias will depend on the nature of the error made during reconstruction. If hierarchical structures are used to estimate only σ^2_A and σ^2_E (in situations where σ^2_C is negligible) incorrect relationship assignment often leads to greater bias than in non-nested population structures.

Simulations indicated that a large amount of marker information and samples containing large numbers of related individuals (both half- and full-sib) are needed before both the environmental correlation and heritability can be estimated accurately. The MCMC approach used in this study was the most basic form of the approach, using an uninformative distribution to describe the sibship sizes and not re-estimating population allele frequencies during reconstruction. Even so, variance component estimates were often more reliable (i.e. had smaller MSE) than those from the pair-wise likelihood technique, presumably through more efficient weighting of half-sib and full-sib data, but showed bias with low levels of marker information.

In the analysis reported here several modifications were used that made the MCMC mixing process more efficient and speeded up the time to convergence. Considering individuals sequentially rather than randomly speeds up convergence time since all individuals are mixed with equal frequency, resulting in better mixing of the sample. In addition, a step was added that periodically combined half-sib families together allowing larger steps across the likelihood surface, helping to prevent the chain becoming stranded on a false peak. The same technique could be applied to joining full-sib families within half sibships, further improving mixing. There are a number of further modifications that could be made that might improve results. The routine could be written in a two tiered form, with first full-sib and then half-sib groups being constructed. This may help the accurate assignment of half-sibs because more marker data are available at a group level than at an individual level. Altering the acceptance and rejection rules for a given step in the chain, for example governing the probability of each step using a 'temperature' that changes as the chain progresses (Kirkpatrick *et al*., 1983; Smith *et al*., 2001), may also improve results.

An added appeal of the MCMC approach is the ease with which it can be modified to include further information. For example, when maternal pedigrees or genotypes are available they may be included, thereby improving parameter estimation. Simulation studies, not presented here, indicated that the inclusion of maternal genotypes yields extremely accurate sibship reconstruction, with estimates of variance components almost identical to those using the actual pedigree, regardless of the actual sibship structure. When only some of the maternal genotypes are known, or when maternal genotype is not coupled with maternal identity, inclusion of maternal information in the MCMC approach becomes complex, involving part maternity inference and part sibship reconstruction. Mixing of individuals with unknown mothers would be over candidate mothers as well as full- and half-sibships (assuming hierarchical sample structure). Mixing of individuals with known mothers would be over half-sibships only, with other individuals assigned to that mother also being reassigned. Equation (1) may be modified to accommodate known maternal genotypes, with summation being made over the known maternal genotypes, rather than every maternal genotype combination.

In this study, populations containing half- and full-sibs were assumed to be in a hierarchical structure. In practice, however, maternal half-sib families may be present so that reconstruction using the MCMC approach becomes complex and sibships could become very extended, with individuals having both paternal and maternal half-sibs and full-sibs. If maternal data are known, this is less of a problem since summation would be over the paternal data, otherwise mixing would have to be between both fathers and mothers. Calculation of the likelihood of the genotypes in a sibship would become extremely slow, since summation would be across all possible parental genotypes. Likelihood calculations could be restricted to the immediate maternal and paternal half-sib families of the candidate individual, rather than including (for example) the maternal half-sib families of one of its paternal half-sibs. However, valuable information from excluded genotype patterns might then be lost.

In the present study, sires and dams (both between and within families) were assumed to be unrelated, but in natural populations such relationships are likely to occur at some level. Several, as yet untested, predictions may be made, however. For example, if the sire and dam within a family are related (i.e. there is inbreeding), the reconstruction of full-sibs is likely to be improved due to an increase in the level of homozygosity in the offspring. If the parents of different families are related, there would probably be little effect on the reconstruction of full-sibs provided the levels of relationship are not too high. If full-sib sizes are large then half-sibs are also likely to be reconstructed reasonably accurately. However, if full-sib family sizes are small, full cousins are likely to interfere in the reconstruction process, although under these conditions reconstruction is less reliable anyway.

In conclusion, if any of these methodologies are to be adopted in practice certain restrictions must be placed on the populations under scrutiny. There must be large amounts of marker data available, with at least 10 reasonably polymorphic marker loci, although

the actual amount required depends strongly upon the sample size. In addition the family sizes, especially those of the full-sib families, should be large (e.g. fish populations). A recent cautionary paper by Thomas *et al.* (2002) clearly demonstrates the effects of not fulfilling these requirements when using the MCMC methodologies in the investigation of the heritability of body weight in a natural population of Soay sheep. These generally have small half-sib families, with very few full-sibs, and marker data did not follow the 'idealized' rectangular distribution simulated here. As a result estimates of variance components using markers to determine relationships compared poorly with those made using a pedigree estimated mainly by observation of the population at lambing.

S.C.T. was funded by a Biotechnology and Biological Sciences Research Council PhD studentship. The authors thank Victor Martinez for comments and discussion.

Appendix

Consider a paternal half-sib family, comprising n_f full sib families, indexed by m, nested within it. Each full-sib family contains n_m progeny which are indexed by c. Individual 1 from family 1 is used to constrain the paternal genotype, using each of that offspring's alleles in turn. Likelihoods must be weighted by $\frac{1}{2}$, since either allele in the offspring could have come from the parent. The other offspring allele is used to constrain the maternal genotype of that full-sib family. The maternal genotypes for the remaining full-sib families are constrained by using the first offspring in each of those families. Again, since either allele could come from the mother this must be repeated and the likelihoods scaled by $\frac{1}{2}$. a_{mci} denotes allele i (of locus ℓ) of individual c of full-sib family m. S_{vw} is an indicator variable, with $S_{vw}=1$ when allele v is the same as w and $S_{vw}=0$ otherwise. P_v is the frequency of allele v. The likelihood for a single locus, with n_ℓ alleles, is:

$$L_\ell = S_{a_{111}a_{112}}\left[\sum_{i=1}^{1}\sum_{x=1}^{n_\ell} p_{a_{11i}}p_x\prod_{m=1}^{n_f}(d)\right] + (1-S_{a_{111}a_{112}})\left[0{\cdot}5\sum_{i=1}^{2}\sum_{x=1}^{n_\ell} g\,p_{a_{11i}}p_x\prod_{m=1}^{n_f}(d)\right]. \quad \text{(A1)}$$

When $i=1$ then $g=1$ and when $i=2$ then $g=1-S_{xa_{111}}$.

When $m=1$, then

$$d=\sum_{y=1}^{n_\ell} bp_{a_{m1j}}p_y\prod_{c=1}^{n_m}(f), \quad \text{(A2)}$$

where $j=3-i$.

When $m\neq 1$, then

$$d = S_{a_{m11}a_{m12}}\left[\sum_{j=1}^{1}\sum_{y=1}^{n_\ell} bp_{a_{m1j}}p_y\prod_{c=1}^{n_m}(f)\right] + (1-S_{a_{m11}a_{m12}})\left[0{\cdot}5\sum_{j=1}^{2}\sum_{y=1}^{n_\ell} hbp_{a_{m1j}}p_y\prod_{c=1}^{n_m}(f)\right] \quad \text{(A3)}$$

When $j=1$ then $h=1$ and when $j=2$ then $h=1-S_{ya_{m11}}$.

For equations (A2) and (A3)

$$b = 8\times 2^{-(S_{a_{11i}x}+S_{a_{m1j}y}+S_*)}$$

$$f = 0{\cdot}25(S_{a_{mc1}a_{11i}}S_{a_{mc2}a_{m1j}} + S_{a_{mc1}a_{11i}}S_{a_{mc2}y} + S_{a_{mc1}x}S_{a_{mc2}a_{m1j}} + S_{a_{mc1}x}S_{a_{mc2}y}) + 0{\cdot}25(1-S_{a_{mc1}a_{mc2}})\,(S_{a_{mc2}a_{11i}}S_{a_{mc1}a_{m1j}} + S_{a_{mc2}a_{11i}}S_{a_{mc1}y} + S_{a_{mc2}x}S_{a_{mc1}a_{m1j}} + S_{a_{mc2}x}S_{a_{mc1}y})$$

S_* is an indicator variable, with $S_*=1$ when the unordered genotypes of the parents are the same and $S_*=0$ otherwise. Multi-locus likelihoods are calculated by multiplying (A1) over loci. When $n_f=1$ the constrained hierarchical equations are applicable to the full-sib case. Likewise, when $n_m=1$ for all m, the equations are applicable to the half-sib case. For optimal speed each locus is ordered within families so that homozygous individuals are considered before heterozygous. This reduces the amount of summation required thereby shortening calculation time.

References

Almudevar, A. & Field, C. (1999). Estimation of single-generation sibling relationships based on DNA markers. *Journal of Agricultural, Biological, and Environmental Statistics* **2**, 136–165.

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

Falconer, D. S. & Mackay, T. F. C. (1996) *Introduction to Quantitative Genetics*, 4th edn. Harlow, UK: Longman.

Gilmour, A. R., Thompson, R. B., Cullis, R. & Welham, S. J. (1997). *ASREML Manual*. Orange, 2800, Australia: New South Wales Department of Agriculture.

Hastings, W. W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.

Kirkpatrick, S., Gelatt, C. D. & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science* **220**, 671–680.

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

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

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

Ritland, K. (1996). Estimators for pair-wise relatedness and individual inbreeding coefficients. *Genetical Research* **67**, 175–185.

Smith, B. R., Herbinger, C. M. & Merry, H. R. (2001). Accurate partition of individuals into full-sib families from genetic data without parental information. *Genetics* **158**, 1329–1338.

Storfer, A. (1996). Quantitative genetics: a promising approach for the assessment of genetic variation in endangered species. *Trends in Ecology and Evolution* **11**, 343–348.

Thomas, S. C. & Hill, W. G. (2000). Estimating quantitative genetic parameters using sibships reconstructed from marker data. *Genetics* **155**, 1961–1972.

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

Thomas, S. C., Coltman, D. W. & Pemberton, J. M. (2002). The use of marker-based relationship information to estimate the heritability of body weight in a natural population: A cautionary tale. *Journal of Evolutionary Biology* **15**, 92–99.

Thompson, E. A. (1975). The estimation of pairwise relationship. *Annals of Human Genetics* **39**, 173–188.