Genet. Sel. Evol. 33 (2001) 587–603

Original article

A sampling algorithm for segregation analysis

Bruce TIER*, John HENSHALL

Animal Genetics and Breeding Unit** , University of New England,
Armidale NSW 2351, Australia

(Received 27 June 2000; accepted 12 June 2001)

Abstract – Methods for detecting Quantitative Trait Loci (QTL) without markers have generally used iterative peeling algorithms for determining genotype probabilities. These algorithms have considerable shortcomings in complex pedigrees. A Monte Carlo Markov chain (MCMC) method which samples the pedigree of the whole population jointly is described. Simultaneous sampling of the pedigree was achieved by sampling descent graphs using the Metropolis-Hastings algorithm. A descent graph describes the inheritance state of each allele and provides pedigrees guaranteed to be consistent with Mendelian sampling. Sampling descent graphs overcomes most, if not all, of the limitations incurred by iterative peeling algorithms. The algorithm was able to find the QTL in most of the simulated populations. However, when the QTL was not modeled or found then its effect was ascribed to the polygenic component. No QTL were detected when they were not simulated.

descent graphs / Monte Carlo Markov chain / quantitative trait loci / Metropolis-Hastings

1. INTRODUCTION

Considerable research has been directed at locating quantitative trait loci (QTL) among continuously distributed traits in the genome of various species using designed experiments and marker information. There is also keen interest in searching for possible QTL in complex pedigree structures without markers in domesticated livestock. This is called segregation analysis, in which both the effect of an allele at a postulated QTL and the probabilities that individuals in the population carry 0, 1 or 2 copies of the desirable allele are of interest.

Methods for segregation analysis have been suggested by a number of authors, *e.g.* [11,13,15,20]. Models that include a QTL, segregating according to Mendelian rules, and a polygenic component which incorporates all other genetic effects are commonly used in these analyses. A typical *mixed*

* Correspondence and reprints
E-mail: btier@pobox.une.edu.au
** A joint institute of NSW Agriculture and the University of New England.

inheritance model for the analysis of a single trait is

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{ZWm} + \mathbf{e} \quad (1)$$

where, **y**, **b**, **a**, and **e** are vectors of observations, fixed effects, breeding values and residuals, **X** and **Z** are incidence matrices assigning observations to effects, **W** is an unknown ($N_{\text{individuals}} \times N_{\text{genotypes}}$) matrix describing the populations' genotypes and **m** is a ($N_{\text{genotypes}} \times 1$) vector of QTL effects related to the available genotypes. The matrix **W** contains one 1 in each row, corresponding to the individual's genotype, and zeroes elsewhere. Variances of the random effects **a** and **e** in this model are assumed to be $\mathbf{A}\sigma_a^2$, where **A** is the numerator relationship matrix, and $\mathbf{I}\sigma_e^2$.

With such a model Hoeschele [11] used Maximum-*A-Posteriori* Estimation to evaluate sires for gene effects and polygenic breeding values. Requirements for this method were evidence of a major locus segregating in the population and also the number of major genotypes at that locus. The method was extended to the analysis of categorical data and multiple traits.

Kinghorn *et al.* [15] developed FINDGENE as a method for searching for a postulated QTL. From the mixed inheritance model (1), the model

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{ZTm} + \mathbf{e}$$

was developed, where **T** contains the animals' genotypic probabilities. Each row of **T** contains (real) probabilities that sum to one. The heritability (h^2) of the trait is required as an input for FINDGENE.

With this algorithm, genotypic probabilities (**T**) are determined using the method of Fernando *et al.* [3] with the data adjusted for the current estimate (denoted $\hat{}$) of the other effects in the model ($\mathbf{y}\text{-}\mathbf{X}\hat{\mathbf{b}}\text{-}\mathbf{Z}\hat{\mathbf{a}}$), a penetrance function and the pedigree. To estimate all other effects, the data adjusted for the current estimate of both the major gene effect and genotypic probabilities ($\mathbf{y}\text{-}\hat{\mathbf{T}}\hat{\mathbf{m}}$) are used. A series of iterations between estimating genotypic probabilities and the effect of the QTL, and estimating all the other effects are required to obtain approximate solutions for all effects in the model.

Van Arendonk *et al.* [24] introduced an iterative method for determining genotypic probabilities by extending the "peeling" method of Elston and Stewart [1] for use in large complex pedigrees with many loops. Corrected formulae for this process were published by Janss *et al.* [13], in a paper which introduced the term "iterative peeling", and Fernando *et al.* [3]. In iterative peeling the probability that an individual has each of the possible genotypes is determined as a function of the existing genotypic probabilities of the individual's neighbourhood set – its parents, progeny and mates.

Janss *et al.* [13] used the Gibbs Sampler to determine the effect of a postulated QTL and the probabilities that individuals in the population had

each of the possible genotypes. The Gibbs Sampler [6] is a Monte Carlo Markov Chain (MCMC) method in which parameters are repeatedly sampled individually and randomly from their conditional posterior distributions, given the current state of all other parameters. These samples form a Markov Chain. Sampling continues until the distribution of each parameter converges to its true distribution. The mean or mode of the samples taken after convergence of the Markov Chain are used as an estimate of each parameter. Janss *et al.* [13] – using a Gibbs Sampler – first estimated the set of each individual's genotypic probabilities ($\mathbf{t}_i$) from the current genotypic probabilities of its parents, progeny and mates, the data and all other parameters; then a particular genotype ($\mathbf{w}_i$) was sampled for that individual. The frequency (q) of one of the two alleles in the base population was also sampled. Sorensen [20] describes an implementation of the Gibbs Sampler which improves the method of Janss *et al.* [13] by simultaneously sampling **b** and **a** effects by first permuting **y** appropriately.

In both these algorithms, genotypes are sampled individually using iterative peeling. Unfortunately sampling individual genotypic probabilities as a function of the parents, progeny and mates has a number of well known problems [3]. One of these problems relates to the difficulty of sampling the parameter space or "stickiness" [12]. Loops commonly occur in the pedigree of the population as a result of matings between relatives. When determining each individual's genotypic probabilities iterative peeling treats information coming from parents, mates and progeny independently. When there are loops in the pedigree such information is not independent; nevertheless, in iterative peeling algorithms such as these, it is treated as if it were independent. These problems can lead to high autocorrelations between successive samples and may prevent convergence to the true distribution. It could also bias estimates of the other parameters. Methods, such as simulated tempering [7], have been developed to help overcome these problems. While these are useful, they can be very computationally expensive and may require running a number of Gibbs samplers simultaneously. Blocking (iterating over a group of closely related animals such as a family of half-sibs) has also been proposed as a method to expedite mixing [17].

The Gibbs sampler is a specific case of MCMC methods. More general MCMC methods allow groups of parameters to be sampled jointly and in different ways to the Gibbs Sampler. With the Metropolis-Hastings algorithm [8,19] new parameters are sampled, but not always accepted, to be part of the Markov Chain. A candidate sample is accepted if an increase in the likelihood would result, or if the ratio of the likelihood of the candidate to the current sample is greater than the test number drawn randomly from a uniform distribution; otherwise the candidate sample is rejected and the current sample repeated in the chain. Detailed descriptions of the Metropolis-Hastings algorithms can be found in many statistical texts, *e.g.* [5,21].

One condition for an MCMC algorithm is that any point in the parameter space can be reached from any other point: the Markov Chain must be irreducible. When more than two alleles are involved the Markov chain generated based on an iterative peeling algorithm may be reducible, and this is also a concern when more than one locus is modeled [20].

This paper describes an MCMC method for finding the effect of a postulated QTL with two alleles among quantitative variation, and the probabilities for all individuals in the population having 0, 1 or 2 copies of the gene, without any genotypic information. No parameters were assumed to be known, and hence all were estimated. The method was validated using simulated data from a random mating population. The benefits of this method over previously published methods are discussed and some obvious applications and extensions identified.

2. ALGORITHM

Sorensen [20] describes in detail an implementation of the Gibbs sampler for the mixed inheritance model (1) where the genotypes are sampled individually. A brief summary of this implementation is reproduced here with Sorensen's equation numbers denoted (S *equation number*). Given this model the sampling distribution of the data was assumed to be

$$\mathbf{y}|\mathbf{b}, \mathbf{a}, \mathbf{W}, q, \mathbf{m} \propto N(\mathbf{Xb} + \mathbf{Za} + \mathbf{ZWm}, \mathbf{I}\sigma_e^2) \qquad \text{(S 9.8.3)}$$

Sorensen assumed the following prior distributions for each of the parameters:

1. **b**: p(**b**) $\propto$ constant. (S 9.8.4)
2. **m**: p(**m**) $\propto$ constant. (S 9.8.6)
3. **a**: the prior distribution of breeding values is $a|\sigma_a^2 \sim N(\mathbf{0}, \mathbf{A}\sigma_a^2)$. (S 9.8.5)
4. σ_a^2 and σ_e^2: the prior distributions for the variances are assumed to be scaled, inverted chi-square, of the form

$$\mathrm{p}\left(\sigma_i^2|\nu_i, S_i^2\right) \propto \left(\sigma_i^2\right)^{-((\nu_i/2)+1)} \exp\left[-0.5\nu_i S_i^2/\sigma_i^2\right] \quad (i = a, e) \qquad \text{(S 9.1.5)}$$

   where ν_i and S_i are hyperparameters.
5. **W**: The probability distribution for a genotype configuration for the whole population is:

$$P(\mathbf{W}') = \prod_{\text{founders } i} P(\mathbf{w}_i) \prod_{\text{non-founders } j} P(\mathbf{w}_j|\mathbf{w}_{\text{mother}_j}\mathbf{w}_{\text{father}_j}). \qquad \text{(S 9.8.1)}$$

   Alleles are assumed to be randomly sampled from the parents genotypes according to Mendelian rules. For founders, genotypes are assumed to

be randomly sampled from the available genotypes given the frequency of the alleles in the base population (q and $1-q$) and the assumption of Hardy-Weinberg equilibrium.

6. q: a beta distribution with parameters e and f is assumed for the allele frequencies (q and $1-q$) in the base population,

$$p(q) \propto q^{e-1}(1-q)^{f-1} \qquad \text{(S p. 182)}$$

Posterior distributions

The full posterior distribution of the parameters is given by:

$$p\left(\mathbf{b}, \mathbf{a}, \mathbf{W}, \mathbf{m}, q, \sigma_e^2, \sigma_a^2|\mathbf{y}\right) \propto p\left(\mathbf{y}|\mathbf{b}, \mathbf{a}, \mathbf{W}, \mathbf{m}, q, \sigma_e^2\right) p\left(\mathbf{a}|\sigma_a^2\right) p(\mathbf{W}) p(q) p\left(\sigma_a^2\right) p\left(\sigma_e^2\right) \qquad \text{(S 9.8.7)}$$

The conditional posterior distributions for each of the parameters from which samples are drawn are:

1. $b_i|\mathbf{b}_{-i}, \mathbf{a}, \mathbf{W}, \mathbf{m}, \sigma_e^2, \mathbf{y} \sim N\left(\hat{b}_i, (x_i' x_i)^{-1}\sigma_e^2\right)$ (S 9.8.8)

where $*_{-i}$ means the matrix (vector)* without the i-th row (element) and $\hat{b}_i$ satisfies

$$\mathbf{x}_i' \mathbf{x}_i \hat{b}_i = \mathbf{x}_{-i}'(\mathbf{y} - \mathbf{X}_{-i}\mathbf{b}_{-i} - \mathbf{Z}\mathbf{a} - \mathbf{Z}\mathbf{W}\mathbf{m}) \qquad \text{(S p. 183)}$$

2. $a_i|\mathbf{b}, \mathbf{a}_{-i}, \mathbf{W}, \mathbf{m}, \sigma_e^2, \sigma_a^2, \mathbf{y} \sim N\left(\hat{a}_i, (z_i' z_i + \mathbf{A}_{i,i}^{-1}\lambda)^{-1}\sigma_e^2\right)$ (S 9.8.9)
where $\lambda = \sigma_e^2/\sigma_a^2$ and $\hat{a}_i$ satisfies

$$\left(\mathbf{z}_i' \mathbf{z}_i + \mathbf{A}_{i,i}^{-1}\right)\hat{a}_i = \mathbf{z}_i'\left(\mathbf{y} - \mathbf{X}\mathbf{b} - \mathbf{A}_{i,-i}^{-1} k \mathbf{a}_{-i} - \mathbf{Z}\mathbf{W}\mathbf{m}\right) \qquad \text{(S 9.8.10)}$$

3. $\sigma_a^2|\mathbf{a}, \mathbf{y} \sim \left(\mathbf{a}'\mathbf{A}^{-1}\mathbf{a} + v_a S_a\right)\chi_{n_a+v_a}^{-2}$, (S 9.8.11)
where n_a is the rank of $\mathbf{A}$.
4. $\sigma_e^2|\mathbf{b}, \mathbf{a}, \mathbf{W}, \mathbf{m}, \mathbf{y} \sim [\mathbf{e}'\mathbf{e} + v_e S_e]\chi_{n_q+v_e}^{-2}$, (S 9.8.12)
where n_e is the order of $\mathbf{y}$ and $\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\mathbf{b}} - \mathbf{Z}\hat{\mathbf{a}} - \mathbf{Z}\hat{\mathbf{W}}\hat{\mathbf{m}}$.
5. $\mathbf{m}$: one set of linear independent contrasts associated with three genotypes (AA, Aa and aa)

$$\mathbf{k} = \begin{bmatrix} 1 & 0 & -1 \\ -0.5 & 1 & 0.5 \end{bmatrix}$$

which represent the additive and dominance effects, have the conditional posterior distribution

$$\mathbf{k}'\mathbf{m}|\mathbf{b}, \mathbf{a}, \mathbf{W}, \sigma_e^2, \mathbf{y} \sim N\left(\mathbf{k}'\mathbf{m}, \left(\mathbf{k}\mathbf{W}'\mathbf{Z}'\mathbf{Z}\mathbf{W}\mathbf{k}\right)^{-1}\sigma_e^2\right) \qquad \text{(S 9.8.14)}$$

6. the probability that an individual has each of the possible genotypes is computed by evaluating the expression

$$P\left(\mathbf{w}_i|\mathbf{b}, \mathbf{a}, \mathbf{W}_{-i}, \sigma_e^2, \sigma_a^2, \mathbf{y}\right)$$
$$\propto P(y_i|\mathbf{b}, \mathbf{a}, \mathbf{w}_i, \sigma_e^2) \prod_{j=1}^{n_i} \prod_{l=1}^{n_{ij}} P\left(\mathbf{w}_{ijl}|\mathbf{w}_i, \mathbf{w}_{ij}\right) P\left(\mathbf{w}_i|\mathbf{w}_{m_i}, \mathbf{w}_{f_i}\right) \quad \text{(S 9.8.17)}$$

where individual i has n_i mates with n_{ij} offspring with mate j, for each possible value of $\mathbf{w}_i$ and normalising the result so that the probabilities sum to one. If the individual is a founder then the term $P(\mathbf{w}_i|\mathbf{w}_{m_i}, \mathbf{w}_{f_i})$ is replaced by the term $P(\mathbf{w}_i)$ which is a function of the gene frequency in the base population. A genotype ($\mathbf{w}_i$) is sampled using a number drawn randomly from a uniform distribution.

7. $q|\mathbf{W}_{\text{founders}}, \mathbf{y} \propto Be(q|n_{a_1} + f, n_{a_2} + e)$, where n_{a_1} and n_{a_2} are the numbers of each type of allele in the founders.

The Gibbs Sampler proceeds by sampling parameters from these conditional distributions until sufficient samples have been drawn for the process to have converged to the true distribution. As already noted, sampling genotypic probabilities individually can restrict mixing of **W**, and that of other parameters, and could prevent convergence to the true distribution. To avoid these problems, the new algorithm proposes using descent graphs and the Metropolis Hastings algorithm to sample **W**.

A descent graph is a structure which uses ordered pairs of zeroes and ones to describe the inheritance state of all alleles in the population. Among descendants the first (second) element of each pair describes the source of the paternal (maternal) allele. A zero (one) indicates that the grandpaternal (grandmaternal) allele was inherited. The elements in the descent graph for a population are numbered from 1 to $2N_{\text{individuals}}$, with the i-th pair corresponding to the ith individual. The binary nature of a descent graph lends itself to rapid sampling on (binary) computers, *e.g.* [4]. When there are only two alleles, the descent graph also can be used for including the founder population (those with unknown parents), but for them zero and one represent the different alleles. A simple pedigree and descent graph with resultant genotypes is shown in Table I. For example, the (0 1) pair in the descent graph for individual 4 indicates that it inherited the sire's paternal allele (allele 1) and the its dam's maternal allele (allele 4). A more detailed description of descent graphs can be found in Lange [16]. Thompson [22] first suggested that sampling descent graphs would be more efficient than sampling genotype probabilities given the relative sizes of the parameter spaces. Although the parameter space for a descent graph may be very big, it is finite, whereas that of the genotypic probabilities is infinite. The problem of calculating genotypic probabilities for

Table I. An example of the pedigree, descent graph and genotypes of a sample population*.

Pedigree			Descent		Ordered Genotypes	
Animal	Sire	Dam	Graph		(Paternal)	(Maternal)
Base Animals						
1	unknown		.	.	1	2
2	unknown		.	.	3	4
3	unknown		.	.	5	6
Descendants						
4	1	2	0	1	1	4
5	3	4	0	1	5	4
6	1	4	0	0	1	1
7	5	6	1	0	4	1

* Base animals' alleles labeled 1 to 6 to show unambiguous inheritance, a zero (one) in the descent graph indicates inheritance of the parent's paternal (maternal) allele.

ungenotyped individuals when some genotypic data are available on others, as is the case for genetic disorders, is more complicated and is addressed in a separate paper [10].

When all genotypes are unknown any randomly drawn descent graph will provide a set of genotypes for the population consistent with the pedigree. The matrix of genotypic probabilities for the population (**T**) can be calculated as the mean of these sets. Besides providing a consistent pedigree a descent graph also provides the frequency of alleles in the base population (q).

W can be sampled from the full conditional posterior $p(\mathbf{W}|\mathbf{b}, \mathbf{a}, q, \sigma_e^2, \sigma_a^2, \mathbf{y})$ using the Metropolis-Hastings algorithm. In this implementation of the Metropolis-Hastings algorithm a descent graph is drawn randomly to form the initial sample. This descent graph is permuted by randomly choosing some individual bits and changing their state from 0 to 1 or *vice-versa*. This process can cause significant change to the pedigree of the population for chosen individuals which have a heterozygous parent and many descendants. Table II illustrates the changes to the genotypes after permuting the paternal bits in the descent graph for two individuals in the example population. However, if the individual has no descendants, it will be the only one to change its pedigree and this will occur only if the parent is heterozygous.

For descendants the probability of changing state from one grandparental gamete to another is even. For base individuals the probability of changing

Table II. Genotypes for the sample population after permuting the paternal nodes of individuals 4 and 5.

Pedigree			Descent		Ordered Genotypes	
Animal	Sire	Dam	Graph		(Paternal)	(Maternal)
Base Animals						
1	unknown		.	.	1	2
2	unknown		.	.	3	4
3	unknown		.	.	5	6
Descendants						
4	1	2	1	1	2	4
5	3	4	1	1	6	4
6	1	4	0	0	1	2
7	5	6	1	0	4	1

state is a function of the current estimate of q with individual gametes more likely to adopt the more frequent gamete.

There are many different ways that this sampling procedure could be implemented. These questions relate to the number of individual bits to exchange in each step and to the number of Metropolis-Hastings steps to make before returning to sample other parameters. It must be possible in any draw for sufficient bits to be exchangeable so that every other possible pedigree for the population in the distribution is obtainable with some non-zero probability. A number of Metropolis-Hastings steps are made to allow the pedigree to mix sufficiently between sampling of the other parameters. This number depends upon the size of the population being analysed.

The likelihood of each sampled descent graph is determined directly from a function of the errors

$$L = \frac{1}{\sigma_e\sqrt{2\pi}} e^{-\mathbf{e}'\mathbf{e}/(2\sigma_e^2)}.$$

With $\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\mathbf{b}} - \mathbf{Z}\hat{\mathbf{u}} - \mathbf{Z}\hat{\mathbf{W}}\hat{\mathbf{m}}$ determined for each descent graph, the ratio of the likelihoods of the candidate and current samples is computed and tested against a random uniform number ($L_{\text{candidate}}/L_{\text{current}} > r$) before the candidate accepts or rejects the candidate graph.

To compute the likelihood efficiently, a residual for each individual and possible genotype, and its contribution to a likelihood, is calculated before sampling begins. These values are used to compute the change in the likelihood for each newly modified descent graph. Permuting the descent graphs of the progeny of homozygous individuals causes no change to their (or their descendants') genotypes or the likelihood. Such changes are ignored when

calculating the change in the likelihood. The likelihood of the whole pedigree, given the current parameters, is calculated after the final Metropolis-Hastings step in each genotype sampling step.

The frequency of the allele in the base population, to be used in step 6 of the Gibbs sampler, is computed as a by-product of the genotype sampling process. Thus the only difference to the sampling procedure described by Sorensen is to step 5, which is how the genotypes are sampled.

3. SIMULATION

A number of populations with the following structure were generated: A base population of 10 sires and 100 dams was generated. Sires were mated randomly to dams with no limit on sire usage. All dams had ten offspring which had an equal chance of being male or female. Five sires and 50 dams were randomly replaced in each mating cycle. Replacement individuals were all chosen randomly from the most recent cohort. Six cohorts were simulated for a total of 6 110 individuals. Three different models were used to simulate records for the six cohorts. All models contained a residual term and a fixed effect of the cohort. The first model contained a polygenic effect for each individual,

$$y = b + a + e \qquad \text{Model (i)}$$

the second contained an additive monogenic effect for each individual of the cohort,

$$y = b + m + e \qquad \text{Model (ii)}$$

and the third model

$$y = b + a + m + e \qquad \text{Model (iii)}$$

contained both polygenic and monogenic effects for each individual.

Breeding values (**a**) were drawn from the distribution $N(0, \mathbf{A}\sigma_a^2)$. Two alleles ($A$ and a) at the locus were simulated. The effect (m) of having one copy of allele A was set at $\sqrt{6}$, model (ii) or $\sqrt{3}$, model (iii). Two frequencies of allele A in the base population were simulated: 0.5 and 0.2. Base animal alleles had that chance of being allele A. This implied that the base population was in Hardy-Weinberg equilibrium. The variance due to the QTL was then $2q(1-q)m^2$ [2]. When no monogenic effect was modeled the polygenic variance was set to 3 but when both genetic effects were modeled it was set to 1.5. This was the same as the variance due to the monogenic effect when the frequency of A in the base population was 0.5.

Descendants' alleles were randomly selected from their parents' alleles. Records were generated by adding the appropriate genetic effects – breeding values and QTL effects – for the model and residuals drawn from the normal distribution $N(0,3)$.

The same three models used to simulate the populations were also used to analyse each of them. Twelve populations were generated using each model. After 2 000 burn-in Gibbs samples the next 2 000 Gibbs samples were used to provide estimates of all parameters in each analysis. A variety of starting values for the variances and QTL effects were chosen for the different populations. In each Gibbs sample, 12 220 Metropolis-Hastings steps were taken, with a random number – the minimum of 12 220 and $1 + 2/r$, where r is a random uniform variate – of nodes in the descent graph perturbed in each step. The effect of having each of the four possible genotypes – AA, Aa, aA and aa – were estimated separately.

4. RESULTS

Table III summarises the results of analysing populations, simulated with the three different models and two different gene frequencies, with a model (i) containing only a polygenic genetic effect. Regardless of the source of the genetic variation, mono- or polygenic, the genetic and error variances were consistently well estimated with this model.

The corresponding results from analyses with a model (ii) containing only a single monogenic effect are shown in Table IV. With this model the error variance tended to be overestimated in all populations. In all cases both the effect of the QTL and variance due to it were underestimated. When the QTL was the only simulated genetic effect, the highest estimate for both the variance due to the single locus and its effect were obtained, and the error variance was the least overestimated.

The results from the analyses with the complete model (iii) are shown in Table V. With both genetic effects in the analytical model, no effect of a QTL was found when it was not simulated and, the error variance was consistently well estimated. A small variance due to the polygenic effect was found in all replicates even when it was not simulated. There was a considerable difference between the magnitude of these effects in each replicate and a corresponding underestimate in the variance due to the QTL. Figure 1 illustrates the distribution of the samples of genetic variance from each replicate of the population simulated with a base frequency of 0.5 and analysed with both genetic effects – model (iii). It is clear that any variance due to the QTL was missed in replicate D, was ambiguous in replicate B, found in the other ten replicates but often underestimated. As shown however in Figure 2, the effect of the allele was estimated correctly in 8 of the twelve replicates. When the

Table III. Estimates of error and polygenic variance from the polygenic model – Model (i) — in simulated populations: mean and sampling variances of twelve replicates. Variances over the replicates are shown in parenthesis. Residual and total genetic variances of 3 were simulated for all models. The variance due to the genetic effects was shared equally between the two genetic effects in the combined simulation.

	Error Variance	Polygenic Variance
	Polygenic simulation – Model (i)	
Mean	3.02 (0.076)	2.98 (0.22)
Variance	0.15 (0.0006)	0.26 (0.0032)
	$q = 0.5$	
	Monogenic simulation – Model (ii)	
Mean	3.06 (0.029)	3.05 (0.097)
Variance	0.17 (0.0003)	0.29 (0.0015)
	Combined simulation – Model (iii)	
Mean	3.09 (0.051)	2.92 (0.18)
Variance	0.16 (0.0001)	0.28 (0.0005)
	$q = 0.2$	
	Monogenic simulation – Model (ii)	
Mean	2.99 (0.046)	2.79 (0.23)
Variance	0.15 (0.0006)	0.28 (0.003)
	Combined simulation – Model (iii)	
Mean	3.00 (0.044)	2.57 (0.13)
Variance	0.14 (0.0002)	0.25 (0.0009)

frequency of the base allele was simulated as 0.2 the QTL was found in eleven of the twelve replicates.

The analysis of each replicate took less than one hour using a Pentium (350 MHz) computer.

5. DISCUSSION

Sampling descent graphs with the Metropolis-Hastings Algorithm has many advantages over sampling with an algorithm based on iterative peeling. It is quick, always provides pedigrees that are consistent with Mendelian sampling, can easily handle more than two alleles and allows ready mixing of the genotypes in the pedigree. Since this method of sampling meets the criteria for the Metropolis-Hastings algorithm, it is irreducible. With more than two alleles additional storage is needed for the founder individuals. It is obvious that this form of sampling genotypes will not experience the problems of stickiness that can occur with sampling genotypes using iterative peeling, nor do loops in the

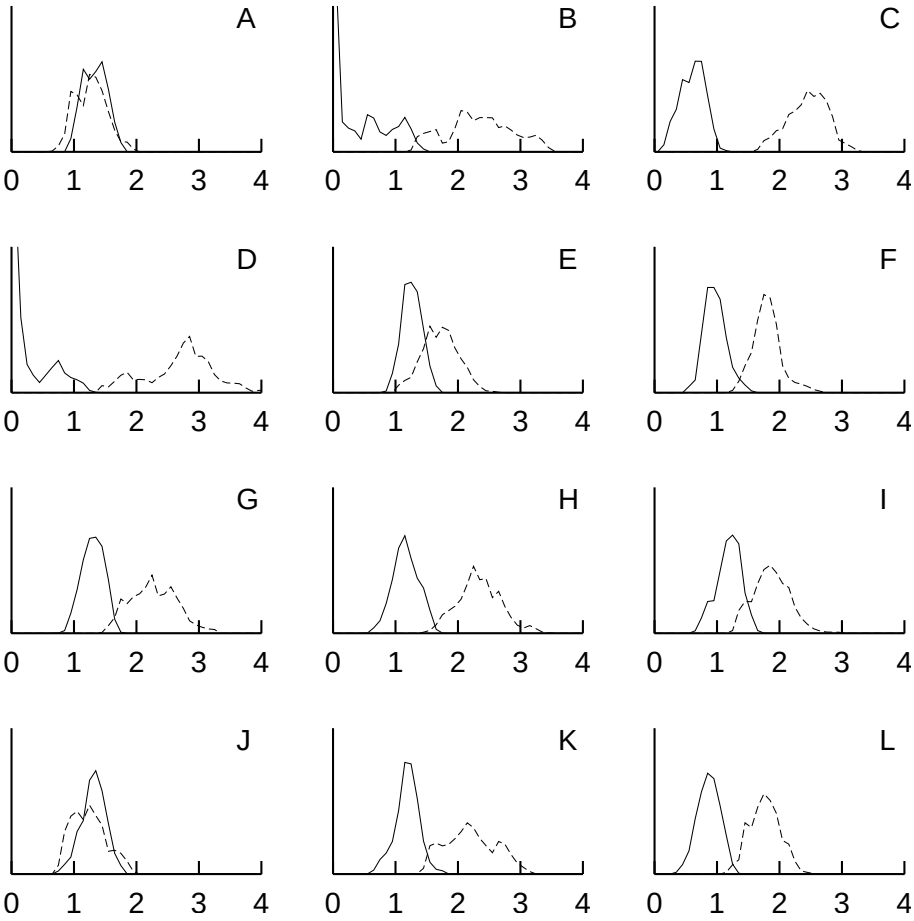

Figure 1. Distributions of sampled QTL variance (solid lines) and additive polygenic variance (broken lines) in 12 replicates, where both the simulation and analytical model included both polygenic and QTL effects – Model (iii), $q = 0.5$. Both QTL and polygenic variances were simulated as 1.5.

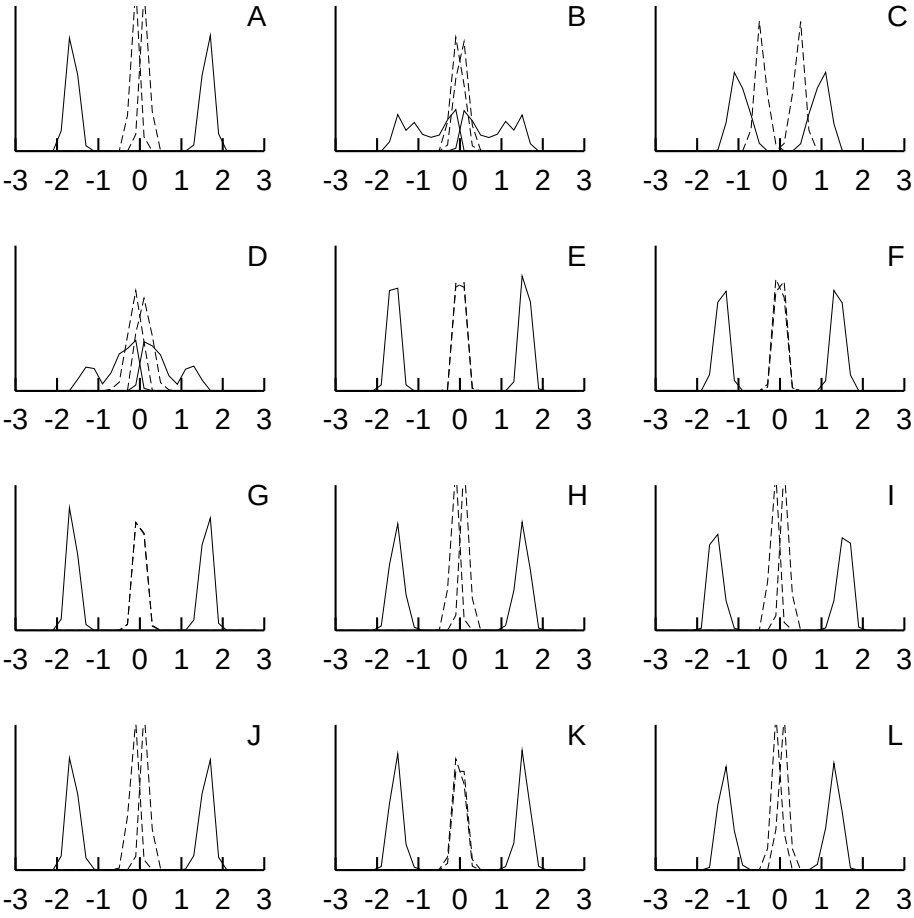

Figure 2. Distribution of sampled QTL effects (solid lines for homozygotes, broken lines for heterozygotes) in 12 replicates, where both the simulation data and the analytical model included both polygenic and QTL effects Model (iii). Simulated values were ± 1.73 for homozygotes and zero for heterozygotes.

Table IV. Estimates of error and variance due to a QTL, and the effect (m) of the QTL in populations from analytical Model (ii) in simulated populations: mean and sampling variance of twelve replicates. Variances over the replicates are shown in parenthesis. Residual and total genetic variances of 3 were simulated for all models. The variance due to the genetic effects was shared equally between the two genetic effects in the combined simulation.

	Error Variance	QTL Variance	QTL Effect
		Polygenic simulation – Model (i)	
Mean	4.00 (0.012)	1.83 (0.17)	1.90 (0.043)
Variance	0.10 (0.0001)	0.07 (0.0001)	0.05 (0.0001)
		$\boldsymbol{q = 0.5}$	
		Monogenic simulation – Model (ii)	
Mean	3.46 (0.079)	2.44 (0.10)	2.19 (0.026)
Variance	0.09 (0.0001)	0.07 (0.0001)	0.04 (0.0001)
		Combined simulation – Model (iii)	
Mean	3.99 (0.037)	1.95 (0.24)	1.96 (0.53)
Variance	0.10 (0.0001)	0.07 (0.0001)	0.05 (0.0001)
		$\boldsymbol{q = 0.2}$	
		Monogenic simulation – Model (ii)	
Mean	3.66 (0.044)	1.97 (0.18)	2.21 (0.040)
Variance	0.09 (0.0001)	0.06 (0.0001)	0.06 (0.0002)
		Combined simulation – Model (iii)	
Mean	3.77 (0.022)	1.72 (0.16)	1.84 (0.43)
Variance	0.09 (0.0001)	0.06 (0.0001)	0.06 (0.0001)

pedigree have any detrimental effect. The choice of how many individual nodes to perturb and how many Metropolis-Hastings steps to take within each Gibbs round can affect efficiency. Our experience when developing this algorithm was that too few Metropolis-Hastings steps guaranteed failure to detect a QTL. Since MCMC methods are computationally expensive the search for an optimal sampling process could be fruitful.

It is reassuring to see that when variation is generated by an additve QTL, its effect is fully recovered with a polygenic model. This means that a QTL with a moderate effect will be contained in the estimated breeding values generated with the Best Linear Unbiased Prediction (Henderson, 1973). It is difficult for the monogenic model to model the genetic variation completely. In the simulated data sets no QTL were found that were not simulated, but in a few cases a simulated QTL was overlooked and in two cases it was unclear whether it was a real effect or not. It is likely that additional depth in the pedigree (more generations) and larger litter sizes should enhance the detection of QTL.

Table V. Estimates of error, polygenic and QTL variances, and the effect (m) of the QTL in simulated populations from analytical Model (iii): mean and sampling variance of twelve replicates. Variances over the replicates are shown in parenthesis. Residual and total genetic variances of 3 were simulated for all models. The variance due to the genetic effects was shared equally between the two genetic effects in the combined simulation.

	Error Variance	Polygenic Variance	QTL Variance	QTL Effect
		Polygenic simulation – Model (i)		
Mean	3.12 (0.083)	2.62 (0.22)	0.19 (0.020)	0.49 (0.044)
Variance	0.17 (0.0003)	0.33 (0.0031)	0.16 (0.0082)	0.28 (0.0078)
		$q = 0.5$		
		Monogenic simulation – Model (ii)		
Mean	2.93 (0.029)	1.16 (0.11)	1.97 (0.045)	1.98 (0.013)
Variance	0.14 (0.0003)	0.22 (0.0034)	0.15 (0.0004)	0.09 (0.0001)
		Combined simulation – Model (iii)		
Mean	3.01 (0.036)	2.00 (0.21)	1.00 (0.13)	1.35 (0.13)
Variance	0.16 (0.0005)	0.34 (0.011)	0.22 (0.0077)	0.20 (0.021)
		$q = 0.2$		
		Monogenic simulation – Model (ii)		
Mean	2.81 (0.036)	1.03 (0.093)	1.91 (0.059)	2.20 (0.046)
Variance	0.13 (0.0004)	0.16 (0.0016)	0.12 (0.0005)	0.09 (0.0001)
		Combined simulation – Model (iii)		
Mean	3.00 (0.064)	1.64 (0.22)	0.88 (0.11)	1.37 (0.032)
Variance	0.14 (0.0004)	0.28 (0.0041)	0.17 (0.0017)	0.22 (0.013)

Conversely, smaller litter sizes and fewer generations of data will make it more difficult to identify QTL with this method. It is likely that smaller QTL can be detected in larger data sets.

This method of sampling is also the logical one for use with finite locus models, *e.g.* [14,23]. The method has been used to find putative QTL for a number of traits in cattle, sheep and swine including one for disease resistance [18] and has been adapted to incorporate marker information [9].

REFERENCES

[1] Elston R.C., Stewart J., A general model for the genetic analysis of pedigree data, Hum. Hered. 21 (1971) 523-542.

[2] Falconer D.S., Introduction to Quantitative Genetics, 2nd edn., Longman, London, 1981.

[3] Fernando R.L., Stricker C., Elston R.C., An efficient algorithm to compute the posterior genotypic distribution for every member of a pedigree without loops, Theor. Appl. Genet. 87 (1993) 89–93.
[4] Fraser A., Burnell D., Computer Models in Genetics, McGraw Hill, New York, 1970.
[5] Gelman A., Carlin J.B., Stern H.S., Rubin D.B., Bayesian Data Analysis, Chapman and Hall, London, 1995.
[6] Geman S., Geman D., Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images, IEEE Transactions on Pattern Analysis and Machine Intelligence 6 (1984) 721–741.
[7] Geyer C.J., Thompson E A., Annealing Markov chain Monte Carlo with applications to ancestral inference, J. Amer. Stat. Ass. 90 (1995) 909–920.
[8] Hastings W.K., Monte Carlo sampling methods using Markov chains and their applications, Biometrika 57 (1970) 97–109.
[9] Henshall J.M., Tier B., Kerr R.J., Detecting QTL in complex pedigrees using phenotypes and markers, in: Proc. 14th Conf. Assoc. Adv. Anim. Breed. Genet., Queenstown, NZ, 2001, pp. 301–304.
[10] Henshall J.M., Tier B., Kerr R.J., Estimating genotypes with independently sampled descent graphs, Genet. Res. 78 (2001) 281–288.
[11] Hoeschele I., Statistical techniques for detection of major genes in animal breeding data, Theor. Appl. Genet. 76 (1988) 311–319.
[12] Hoeschele I., Mapping quantitative trail loci in outbred populations, in: Balding, D.J., Bishop M., Cannings C. (Eds.), Handbook of Statistical Genetics, Wiley, Chichester, 2001, pp. 599–644.
[13] Janss L.L.G., Thompson R., van Arendonk J.A.M., Application of Gibbs sampling for inference in a mixed major gene-polygenic inheritance model in animal populations, Theor. Appl. Genet. 91 (1995) 1137–1147.
[14] Kerr R.J., Henshall J.M., Tier B., Use of a finite locus model to estimate genetic parameters in unselected populations, in: Proc. 13th Conf. Assoc. Adv. Anim. Breed. Genet., Mandurah, W. Aust., 1999, pp. 404–407.
[15] Kinghorn B.P., Kennedy B.W., Smith C., A method of screening for genes of major effect, Genetics 134 (1993) 351–360.
[16] Lange K., Mathematical and Statistical Models for Genetical Analysis, Springer-Verlag, New York, 1997.
[17] Lund M., Jensen C.S., Multivariate updating of genotypes in a Gibbs sampling algorithm in the mixed inheritance model, in: Proc. 6th World Congr. Genet. Appl. Livest. Prod., 1998, Vol. 25, pp. 521–524.
[18] Meszaros S.A., Henshall J.M., Burgess S.K., Gray G.D., Tier B., Detection of a quantitative trait locus associated with a reduction in faecal egg count in merino sheep, in: Proc. 13th Conf. Assoc. Adv. Anim. Breed. Genet., Mandurah, W. Aust., 1999, pp. 211–214.
[19] Metropolis N., Rosenbluth A., Rosenbluth M., Teller A., Teller E., Equations of state calculations by fast computing machines, J. Chem. Phys. 21 (1953) 1087–1092.
[20] Sorensen D., Gibbs Sampling in Quantitative Genetics. Internal Report No. 82, Danish Institute of Animal Science, Foulum, Denmark, 1997.

[21] Tanner M.A., Tools for Statistical Inference, Springer-Verlag, New York, 1993.
[22] Thompson E.A., Monte Carlo likelihood in genetic mapping, Stat. Sci. 9 (1994) 355–366.
[23] Thompson E.A., Skolnick M.H., Likelihoods on complex pedigrees for quantitative traits, in: Pollack E., Kempthorne O., Bailey T.B. (Eds.), Proc. Int. Conf. Quant. Genet., Iowa State University Press, Ames, Iowa, 1977, pp. 815–818.
[24] Van Arendonk J.A.M., Smith C., Kennedy B.W., Method to estimate genotype probabilities at individual loci in farm live stock, Theor. Appl. Genet. 78 (1989) 735–740.