Genet. Sel. Evol. 30 (1998) 45–58

Original article

Breeding value estimation with incomplete marker data

Marco C.A.M. Bink [a*], Johan A.M. Van Arendonk [a], Richard L. Quaas [b]

[a] Animal Breeding and Genetics Group, Wageningen Institute of Animal Sciences, Wageningen Agricultural University, PO Box 338, 6700 AH Wageningen, the Netherlands
[b] Department of Animal Science, Cornell University, Ithaca, NY 14853, USA

(Received 20 January 1997; accepted 17 November 1997)

Abstract – Incomplete marker data prevent application of marker-assisted breeding value estimation using animal model BLUP. We describe a Gibbs sampling approach for Bayesian estimation of breeding values, allowing incomplete information on a single marker that is linked to a quantitative trait locus. Derivation of sampling densities for marker genotypes is emphasized, because reconsideration of the gametic relationship matrix structure for a marked quantitative trait locus leads to simple conditional densities. A small numerical example is used to validate estimates obtained from Gibbs sampling. Extension and application of the presented approach in livestock populations is discussed.
© Inra/Elsevier, Paris

breeding values / quantitative trait locus / incomplete marker data / Gibbs sampling

Résumé – Estimation des valeurs génétiques avec information incomplète sur les marqueurs. Un typage incomplet pour les marqueurs empêche l'estimation des valeurs génétiques de type BLUP utilisant l'information sur les marqueurs. On décrit une procédure d'échantillonnage de Gibbs pour l'estimation bayésienne des valeurs génétiques permettant une information incomplète pour un marqueur unique lié à un locus quantitatif. On développe le calcul des densités de probabilités des génotypes au marqueur parce que la reconsidération de la structure de la matrice des corrélations gamétiques pour un locus quantitatif marqué conduit à des densités conditionnelles simples. Un petit exemple numérique est donné pour valider les estimées obtenues par échantillonnage de Gibbs. L'application de l'approche aux populations d'animaux domestiques est discutée.
© Inra/Elsevier, Paris

valeur génétique / locus quantitatif / marqueurs incomplets / échantillonnage de Gibbs

* Correspondence and reprints

1. INTRODUCTION

Identification of a genetic marker closely linked to a gene (or a cluster of genes) affecting a quantitative trait, allows more accurate selection for that trait [5]. The possible advantages of marker-assisted genetic evaluation have been described extensively (e.g. [13, 16, 17]).

Fernando and Grossman [1] demonstrated how best linear unbiased prediction (BLUP) can be performed when data are available on a single marker linked to quantitative trait locus (QTL). The method of Fernando and Grossman has been modified for including multiple unlinked marked QTL [23], a different method of assigning QTL effects within animals [26]; and marker brackets [5]. These methods are efficient when marker data are complete. However, in practice, incompleteness of marker data is very likely because it is expensive and often impossible (when no DNA is available) to obtain marker genotypes for all animals in a pedigree. For every unmarked animal, several marker genotypes can be fitted, each resulting in a different marker genotype configuration. When the proportion or number of unmarked animals increases, identification of each possible marker genotype configuration becomes tedious and analytical computation of likelihood of occurrence of these configurations becomes impossible.

Gibbs sampling [3] is a numerical integration method which provides opportunities to solve analytically intractable problems. Applications of this technique have recently been published in statistics (e.g. [2, 3]) as well as animal breeding (e.g. [18, 25]). Janss et al. [10] successfully applied Gibbs sampling to sample genotypes for a bi-allelic major gene, in the absence of markers. Sampling genotypes for multiallelic loci, e.g. genetic markers, may lead to reducible Gibbs chains [15, 20]. Thompson [21] summarizes approaches to resolve this potential reducibility and concludes that a sampler can be constructed that efficiently samples multiallelic genotypes on a large pedigree.

The objective of this paper is to describe the Gibbs sampler for marker-assisted breeding value estimation for situations where genotypes for a single marker locus are unknown for some individuals in the pedigree. Derivation of the conditional, discrete, sampling distributions for genotypes at the marker is emphasized. A small numerical example is used to compare estimates from Gibbs sampling to true posterior mean estimates. Extension and application of our method are discussed.

2. METHODOLOGY

2.1. Model and priors

We consider inferences about model parameters for a mixed inheritance model of the form

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{W}\mathbf{v} + \mathbf{e} \tag{1}$$

where $\mathbf{y}$ and $\mathbf{e}$ are n-vectors representing observations and residual errors, β is a p-vector of 'fixed effects', $\mathbf{u}$ and $\mathbf{v}$ are q and 2q-vectors of random polygenic and QTL effects, respectively, $\mathbf{X}$ is a known $n \times p$ matrix of full column rank, and $\mathbf{Z}$ and $\mathbf{W}$ are known n $\times$ q and n $\times$ 2q matrices, respectively. For each individual we

consider three random genetic effects, i.e. two additive effects at a marked QTL (v_i^1 and v_i^2, see *figure 1*) and a residual polygenic effect (u_i). Here $\mathbf{e}$ is assumed to have the distribution $N_n(\mathbf{0}, \mathbf{I}\sigma_e^2)$, independently of β, $\mathbf{u}$ and $\mathbf{v}$. Also $\mathbf{u}$ is taken to be $N_q(\mathbf{0}, \mathbf{A}\sigma_u^2)$, where $\mathbf{A}$ is the well-known numerator relationship matrix.

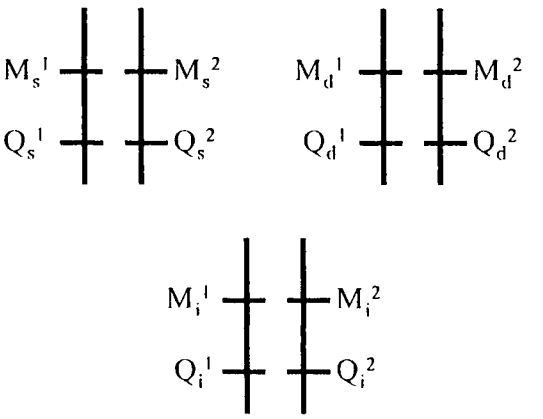

Figure 1. Linkage between marker and quantitative trait locus (QTL) alleles. Assignment of QTL alleles is based on marker alleles. Given a known recombination rate, r, the probability that the first QTL allele of animal i is identical to the second QTL allele of its sire is given as $P(Q_i^1 \equiv Q_s^2) = (1 - r) \times P(M_i^1 \equiv M_s^2) + (r) \times P(M_i^1 \equiv M_s^1)$, where M = marker allele; Q = allele; i = individual, s = sire; and d = dam.

Finally, $\mathbf{v}$ is taken to be $N_{2q}(\mathbf{0}\mathbf{G}\sigma_v^2)$, where $\mathbf{G}$ is the gametic relationship matrix (2q × 2q) computed from pedigrees, a full set of marker genotypes and the known map distance between marker and QTL [26]. In case of incomplete marker data, we augment genotypes for ungenotyped individuals. We then denote $\mathbf{m}_{(k)}$ and $\mathbf{G}_{(k)}$ as the marker genotype configuration k and as the corresponding gametic relationship matrix. Further, β, $\mathbf{u}$, $\mathbf{v}$, and missing marker genotypes are assumed to be independent, a priori. We assume complete knowledge on variance components and map distance between marker and QTL.

2.2. Joint posterior density and full conditional distributions for location parameters

The conditional density of $\mathbf{y}$ given β, $\mathbf{u}$, and $\mathbf{v}$ for the model given in equation (1) is proportional to $\exp\{-1/2\sigma_e^{-2}(\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\mathbf{u} - \mathbf{W}\mathbf{v})'(\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\mathbf{u} - \mathbf{W}\mathbf{v}\}$, so the joint posterior density is given by

$$
\begin{aligned}
&p(\beta, \mathbf{u}, \mathbf{v}|\sigma_u^2, \sigma_v^2, \sigma_e^2, \mathbf{m}_{obs}, r, \mathbf{y}) \\
&\quad \propto \exp\{-1/2\sigma_e^{-2}(\mathbf{y} - \mathbf{X}\beta - \mathbf{Zu} - \mathbf{Wv})'(\mathbf{y} - \mathbf{X}\beta - \mathbf{Zu} - \mathbf{Wv})\} \\
&\quad \times \exp\{-1/2\sigma_u^{-2}(\mathbf{u}'\mathbf{A}^{-1}\mathbf{u})\} \\
&\quad \times \sum_{k=1}^{n_c} \left[|\mathbf{G}_{(k)}^{-1}\sigma_v^{-2}|^{1/2}\exp\{-1/2\sigma_v^{-2}(\mathbf{v}'\mathbf{G}_{(k)}^{-1}\mathbf{v})\} \times p(\mathbf{m}_{(k)}|\mathbf{m}_{obs})\right] \qquad (2)
\end{aligned}
$$

The joint posterior density includes a summation (n_c) over all consistent marker genotype configurations ($\mathbf{m}_{(k)}$). In the derivation of the sampling densities for marked QTL effects, however, one particular marker genotype configuration, $\mathbf{m}_{(k)}$, is fixed. The summation needs to be considered only when the sampling of marker genotypes is concerned.

To implement the Gibbs sampling algorithm, we require the conditional posterior distributions of each of β, $\mathbf{u}$, and $\mathbf{v}$ given the remaining parameters, the so-called full conditional distributions, which are as follows

$$
(\beta|\beta_{-i}, \mathbf{u}, \mathbf{v}, \mathbf{y}) \sim N\left[(\mathbf{x}_i'\mathbf{x}_i)^{-1}\mathbf{x}_i'(\mathbf{y} - \mathbf{X}_{-i}\beta_{-i} - \mathbf{Zu} - \mathbf{Wv}), (\mathbf{x}_i'\mathbf{x}_i)^{-1}\sigma_e^2\right] \qquad (3)
$$

$$
\begin{aligned}
(u_i|\mathbf{u}_{-i}, \beta, \mathbf{v}, \mathbf{y}) \sim N\Big[&(\mathbf{z}_i'\mathbf{z}_i + a^{ii}\alpha_u)^{-1}\Big(\mathbf{z}_i'(\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}_{-i}\mathbf{u}_{-i} - \mathbf{Wv}) - \sum_{i\neq j}^{q} a^{ij}\alpha_u u_{ij}\Big), \\
&(\mathbf{z}_i'\mathbf{z}_i + a^{ii}\alpha_u)^{-1}\sigma_e^2\Big] \qquad (4)
\end{aligned}
$$

$$
\begin{aligned}
&(\mathbf{v}_i|\mathbf{v}_{-i}, \beta, \mathbf{u}, \mathbf{m}_{(k)}, \mathbf{y}) \\
&\qquad \sim N\Big((\mathbf{w}_i'\mathbf{w}_i + g_{(k)}^{ii}\alpha_v)^{-1}\Big(\mathbf{w}_i'(\mathbf{y} - \mathbf{X}\beta - \mathbf{Zu} - \mathbf{W}_{-i}\mathbf{v}_{-i}) - \sum_{i\neq j}^{2q} \alpha_v g_{(k)}^{ij}\mathbf{v}_j\Big), \\
&(\mathbf{w}_i'\mathbf{w}_i + g_{(k)}^{ii}\alpha_v)^{-1}\sigma_e^2\Big) \qquad (5)
\end{aligned}
$$

where, $a^{ij}, g_{(k)}^{ij}$ is the (i, j)th element of $\mathbf{A}^{-1}$ and $\mathbf{G}_{(k)}^{-1}$, respectively, $\alpha_u = \sigma_e^2/\sigma_u^2$, $\alpha_v = \sigma_e^2/\sigma_v^2$ and $\sum_{i\neq j}^{q} a^{ij}\alpha_u u_{ij}$, and $\sum_{i\neq j}^{2q} \alpha_v g_{(k)}^{ij} v_j$ are the corrections for polygenic and gametic covariances in the pedigree, respectively. Note that the means of the distributions (3), (4) and (5) correspond to the updates obtained when mixed model equations are solved by Gauss-Seidel iteration. Methods for sampling from these distributions are well known (e.g. [24, 25]).

2.3. Sampling densities for marker genotypes

Suppose $\mathbf{m}$ is the current vector of marker genotypes, some observed and some of which were augmented (e.g. sampled by the Gibbs sampler). Let $\mathbf{m}_{-i}$ denote the complete set except for the ith (ungenotyped) individual, and let g_m denote

a particular genotype for the marker locus. Then the posterior distribution of genotype g_m is the product of two factors

$$p(m_i = g_m|\mathbf{m}_{-i}, \beta, \mathbf{u}, \mathbf{v}, \mathbf{m}_{obs}, r, y) \propto p(m_i = g_m|\mathbf{m}_{-i}) \times p(\mathbf{v}|m_i = g_m, \mathbf{m}_{-i}, \sigma_v^2, r) \tag{6}$$

with,

$$p(\mathbf{v}|m_i = g_m, \mathbf{m}_{-i}, \sigma_v^2, r) = |\mathbf{G}_{(k)}^{-1}\sigma_v^{-2}|^{1/2}\exp\{-1/2\sigma_v^{-2}(\mathbf{v}'\mathbf{G}_{(k)}^{-1}\mathbf{v})\} \tag{7}$$

where $\mathbf{G}_{(k)}^{-1}$ corresponds to marker genotype set $\{\mathbf{m}_{-i}, m_i = g_m\}$. Thus, equation (7) shows that phenotypic information needed for sampling new genotypes for the marker is present in the vector of QTL effects ($\mathbf{v}$).

Now, it suffices to compute equation (6) for all possible values of g_m, and then randomly select one from that multinomial distribution [20]. In practice considering only those g_m that are consistent with $\mathbf{m}_{-i}$ and Mendelian inheritance can minimize the computations. Furthermore, computations can be simplified because "transmission of genes from parents to offspring are conditionally independent given the genotypes of the parents" [15]. Adapting notation from Sheehan and Thomas [15], let S_j denote the set of mates (spouses) of individual i and $O_{i,j}$ be the set of offspring of the pair i and j. Furthermore, the parents of individual i are denoted by s (sire) and d (dam). Then, equation (6) can be more specifically written as

$$\begin{aligned} &p(m_i = g_m, \mathbf{m}_{-i}|\mathbf{v}, \sigma_v^2, \mathbf{m}_{obs}, r) \\ &\quad \propto p(m_i = g_m|m_s, m_d) \times p(\mathbf{v}_i|\mathbf{v}_s, \mathbf{v}_d, m_i = g_m, m_s, m_d, \sigma_v^2, r) \\ &\quad \times \prod_{j \in S_i} \prod_{l \in O_{i,j}} \{p(m_l|m_i = g_m, m_j) \times p(\mathbf{v}_l|\mathbf{v}_i, \mathbf{v}_j, m_i = g_m, m_j, m_l, \sigma_v^2, r)\} \end{aligned} \tag{8}$$

When parents of individual i are not known, then the first two terms on the right-hand side of equation (8) are replaced by $\pi(m_i)$, which represents frequencies of marker genotypes in a population. The probability $p(m_i = g_m|m_s, m_d)$ corresponds to Mendelian inheritance rules for obtaining marker genotype g_i given parental genotypes m_s and m_d, similar for $p(m_l|m_i = g_m, m_j)$. The computation of $p\{\mathbf{v}_i|\mathbf{v}_d, m_i, m_s, m_d, r\}$ (and $p\{\mathbf{v}_l|\mathbf{v}_i, \mathbf{v}_j, m_i, m_j, m_l, r\}$) can efficiently be performed by utilizing special characteristics of the matrix $\mathbf{G}^{-1}$.

Let $\mathbf{Q}_i$ denote a gametic contribution matrix relating the QTL effects of individual i to the QTL effects of its parents. The matrix $\mathbf{Q}_i$ is $2(i-1) \times 2$. For founder animals, matrix $\mathbf{Q}_i$ is simply zero. The recursive algorithm to compute $\mathbf{G}^{-1}$ of Wang et al. (1995, equation [18]) can be rewritten as

$$\mathbf{G}_q^{-1} = \sum_{i=1}^{q} \begin{bmatrix} -\mathbf{Q}_i \\ \mathbf{I}_2 \\ \mathbf{0}_i \end{bmatrix} \mathbf{D}_i^{-1}[-\mathbf{Q}_i' \; \mathbf{I}_1 \; \mathbf{0}_i] \tag{9}$$

where $\mathbf{D}_i^{-1} = (\mathbf{C}_i - \mathbf{Q}_i'\mathbf{G}_{i-1}\mathbf{Q}_i)^{-1}$ (which reduces to $\mathbf{D}_i^{-1} = (\mathbf{C}_i - \mathbf{Q}_i'\mathbf{G}_{i-1}\mathbf{Q}_i)^{-1}$ with no inbreeding), $\mathbf{0}_i$ is a $2(q-i) \times 2$ null matrix. The off-diagonals in $\mathbf{C}_i$ equal the inbreeding coefficient at the marked QTL [26]. Equation (8) shows the similarity to

Henderson's rules for $\mathbf{A}^{-1}$ [6]. The nonzero elements of $\mathbf{G}^{-1}$ pertaining to an animal arise from its own contribution plus those of its offspring. So, when sampling the ith animal's marker genotype, only those contribution matrices need be considered that contain elements pertaining to animal i. These are the individual's own contributions and those of its progeny when i appears as a parent.

$$
\begin{aligned}
(\mathbf{v}'\mathbf{G}^{-1}\mathbf{v})_i &= \mathbf{v}' \begin{bmatrix} -\mathbf{Q}_i \\ \mathbf{I}_2 \\ \mathbf{0}_i \end{bmatrix} \mathbf{D}_i^{-1}[-\mathbf{Q}_i' \ \mathbf{I}_2 \ \mathbf{0}_i]\mathbf{v} + \sum_{j \in S_i} \sum_{I \in O_{i,j}} \mathbf{v}' \begin{bmatrix} -\mathbf{Q}_j \\ \mathbf{I}_2 \\ \mathbf{0}_j \end{bmatrix} \mathbf{D}_j^{-1}[-\mathbf{Q}_j' \ \mathbf{I}_2 \ \mathbf{0}_j]\mathbf{v} \\
&= [\mathbf{v}_i - \mathbf{Q}_i^s\mathbf{v}_s - \mathbf{Q}_i^d\mathbf{V}_d]\mathbf{D}_i^{-1}[\mathbf{v}_i - \mathbf{Q}_i^s\mathbf{v}_s - \mathbf{Q}_i^d\mathbf{V}_d] \\
&+ \sum_{j \in S_i} \sum_{I \in O_{i,j}} [\mathbf{v}_i - \mathbf{Q}_1^i\mathbf{v}_i - \mathbf{Q}_1^j\mathbf{v}_j]\mathbf{D}_1^{-1}[\mathbf{v}_1 - \mathbf{Q}_1^i\mathbf{v}_i - \mathbf{Q}_1^j\mathbf{v}_j] \qquad (10)
\end{aligned}
$$

where $\mathbf{v}_k$ is the vector of animal k's two marked QTL effects, and $\mathbf{Q}_k^P$ denotes the rows of $\mathbf{Q}_k$ pertaining to P, one of k's parents. Again, we recognize each term in the sum is the kernel of a (bivariate) normal which is $p\{\mathbf{v}_i|\mathbf{v}_s, \mathbf{v}_d, m_i, m_s, m_d, r\}$ or $p\{\mathbf{v}_1|\mathbf{v}_i, \mathbf{v}_j, m_i, m_j, m_1, r\}$.

2.4. Running the Gibbs sampling

The Gibbs sampler is used to obtain a sample of a parameter from the posterior distribution and can be seen as a chained data augmentation algorithm [19]. So, one augments data ($\mathbf{y}$ and $\mathbf{m}_{obs}$) with parameters (θ) to obtain, for example, $p(\theta_1|\theta_2, \ldots, \theta_d, \mathbf{y})$. For the purpose of breeding value estimation, Gibbs sampling works as follows:

1) set arbitrary initial values for $\theta^{[0]}$, we use zeros for fixed and genetic effects and for each unmarked animal, we augment a genotype that is consistent with pedigree, Mendelian inheritance, and observed marker data;

2) sample $\theta_i^{[\tau+1]}$ from

[3], $i = 1, 2, .., p$; for fixed effects,

[4], $i = p + 1, p + 2, .., p + q$; for polygenic effects,

[5], $i = p + q + 1, p + q + 2, .., p + q + 2q$; for marked QTL effects, or

[6], $i = p + 3q + 1, p + 3q + 2, .., p + 3q + t$; for marker genotypes,

and replace $\theta_i^{[\tau]}$ with $\theta_i^{[\tau+1]}$;

3) repeat 2) N (length of chain) times.

For any individual parameter, the collection of n values can be viewed as a simulated sample from the appropriate marginal distribution. This sample can be used to calculate a marginal posterior mean or to estimate the marginal posterior distribution. For small pedigrees with only a few animals missing observed marker genotypes, posterior means can be evaluated directly using

$$
E(\theta^*|\sigma_u^2, \sigma_v^2, \sigma_e^2, \mathbf{m}_{obs}, r, \mathbf{y}) = \sum_{\mathbf{G}_{(k)}} E(\theta^*|\mathbf{G}_{(k)}, \sigma_u^2, \sigma_v^2, \sigma_e^2, \mathbf{y}) \times p(\mathbf{G}_{(k)}|\mathbf{m}_{obs}, r, y) \quad (11)
$$

where θ^* is a fixed, polygenic or marked QTL effect. This provides a criterion to compare the estimates obtained from Gibbs sampling.

3. NUMERICAL EXAMPLE

A small numerical example is used to verify the use of the Gibbs sampler to obtain posterior mean estimates and illustrate the effect of the data on the estimates obtained from two different estimators, i.e. a posterior mean and the well-known BLUP estimator (by solving the MME given in the Appendix). Pedigree and data of the example are in *figure 2*. Both sire (01) and dam (02) have observed marker genotypes, AB and CD, respectively, but do not have phenotypes observed. Three full sibs have a marker genotype BC and a phenotype +20 (denoted FS 03, 04, 05); three other full sibs have a marker genotype AD and a phenotype −20 (denoted FS 06, 07, 08). Both animals 09 and 10 have no marker genotypes but have a phenotype +20 and −20, respectively. Complete knowledge was assumed on variance components and recombination rate between marker and MQTL (*table I*). The thinning factor in Gibbs sampling chain was 50 cycles and the burn in period was twice the thinning factor, and 20 000 thinned samples were used for analysis.

Table I. Population genetic parameters, used in numerical example.

Parameter	Value
Phenotypic variance	1 000
Polygenic variance	300
Marked quantitative trait locus variance	50
Recombination rate	0.05

3.1. Estimates for genetic effects

The posterior estimates obtained from Gibbs sampling were similar to the TRUE posterior estimates, as shown in *table II*. The posterior estimates of MQTL effects of animals 09 and 10 (±0.70) were much less divergent than those of their full sibs that had their marker genotypes observed (±2.48). These less divergent values reflect the uncertainty on marker genotypes of animals 09 and 10. The TRUE and GIBBS posterior densities for an MQTL effect of animal 09 were also very similar (*figure 3*). The posterior variance was 52.3, which was larger than the prior variance ($\sigma_v^2 = 50$) and reveals that the data are not decreasing the prior uncertainty on MQTL effects for animals 09 and 10 in this situation. For the other full sibs, the posterior variance was 47.02, which was lower than the prior variance because segregation of MQTL effects was known with higher certainty, i.e. marker genotypes were known. The BLUP estimates for MQTL effects of animal 09 and 10 were equal to 1/6 of the polygenic effects of these animals, which equaled the variance ratio of the MQTL and the polygenes.

Table II. Posterior mean estimates for genetic effects of all ten animals in the numerical example.

	TRUE[a]				GIBBS[b]				BLUP			
Animal	Polygenic effect	MQTL_1 effect[c]	MQTL_2 effect[c]	Breeding value[d]	Polygenic effect	MQTL_1 effect	MQTL_2 effect	Breeding value	Polygenic effect	MQTL_1 effect	MQTL_2 effect	Breeding value
1	0.00	−2.65	2.65	0.00	−0.10	−2.63	2.61	−0.13	0.00	−2.69	2.69	0.00
2	0.00	2.65	−2.65	0.00	0.01	2.68	−2.61	0.08	0.00	2.69	−2.69	0.00
3	3.01	2.48	2.48	7.97	2.98	2.45	2.51	7.94	2.99	2.52	2.52	8.03
4	3.01	2.48	2.48	7.97	2.90	2.45	2.52	7.86	2.99	2.52	2.52	8.03
5	3.01	2.48	2.48	7.97	3.06	2.42	2.51	7.99	2.99	2.52	2.52	8.03
6	−3.01	−2.48	−2.48	−7.97	−3.07	−2.49	−2.47	−8.03	−2.99	−2.52	−2.52	−8.03
7	−3.01	−2.48	−2.48	−7.97	−3.02	−2.48	−2.44	−7.93	−2.99	−2.52	−2.52	−8.03
8	−3.01	−2.48	−2.48	−7.97	−3.05	−2.47	−2.43	−7.94	−2.99	−2.52	−2.52	−8.03
9	3.72	0.70	0.70	5.12	3.65	0.73	0.70	5.08	3.75	0.63	0.63	5.00
10	−3.72	−0.70	−0.70	−5.12	−3.70	−0.73	−0.69	−5.12	−3.75	−0.63	−0.63	5.00

[a] TRUE: directly computed; [b] GIBBS: approximated by Gibbs sampling; [c] MQTL = marked quantitative trait locus; (MQTL_1 and MQTL_2 denote first and second MQTL effect, respectively); [c] breeding value of an animal is the sum of its polygenic effect and its two MQTL effects.

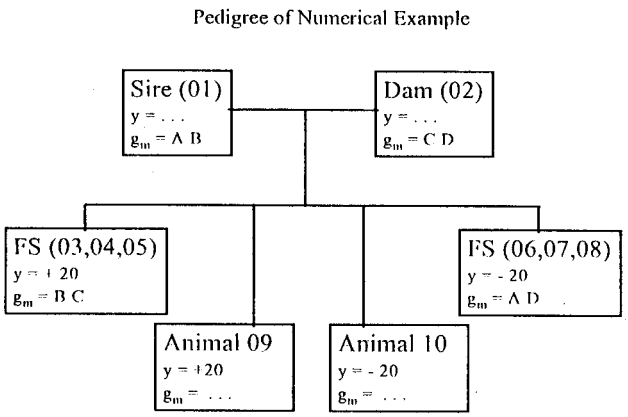

Figure 2. Pedigree of numerical example. Two parents, sire 01 and dam 02, have eight offspring. The sire and dam have observed marker genotypes, AB and CD, respectively, but do not have phenotypes observed. Three full sibs (FS 03, 04, 05) have marker genotype BC and phenotype +20; three other full sibs (FS 06, 07, 08) have marker genotype AD and phenotype −20. Animals 09 and 10 have no marker genotypes but have phenotypes +20 and −20, respectively.

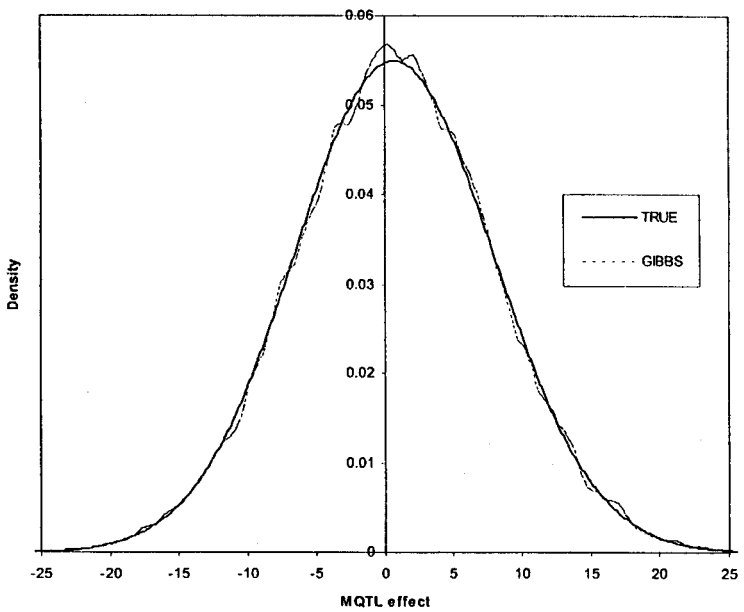

Figure 3. Posterior density of the first marked quantitative trait locus effect of animal 09. TRUE: direct computation ($\mu_T = 0.697$; $\sigma_T = 7.234$: indirect approximation ($\mu_G = 0.730$; $\sigma_G = 7.234$).

3.2. Marker genotype probabilities

In the following marker genotype AB represents both AB and BA. In the latter case, alleles for both marker and MQTL are reordered, maintaining linkage between marker and MQTL alleles within an animal. So, four marker genotypes were possible for animals 09 and 10 (*table III*). Based on pedigree and marker data solely, each of these four genotypes was equally likely (prior probability = 0.25). After including phenotypic data, (posterior) probabilities changed: marker genotype BC and AD for animal 09 became more and less probable, respectively. The reverse holds for animal 10. The estimates from the Gibbs sampler were again very similar to the TRUE posterior probabilities. Complete phenotypic and marker information on six full sibs gave the MQTL effects linked to marker alleles B and C positive values and marker alleles A and D negative values. Note that probabilities (TRUE) for marker genotypes AC and BD also (slightly) changed after considering the phenotypic data.

Table III. Prior and posterior marker genotype probabilities for animals 09 and animal 10.

	Marker genotypes			
	AC	AD	BC	BD
Animal 09				
Prior	0.2500	0.2500	0.2500	0.2500
TRUE[a]	0.2504	0.2196	0.2796	0.2504
GIBBS[b]	0.2470	0.2203	0.2801	0.2527
Animal 10				
Prior	0.2500	0.2500	0.2500	0.2500
TRUE[a]	0.2504	0.2796	0.2196	0.2504
GIBBS[b]	0.2477	0.2815	0.2191	0.2518

[a] TRUE: directly computed; [b] GIBBS: approximated by Gibbs sampling.

4. DISCUSSION

Marker-assisted breeding value estimation in livestock has been hampered by incomplete marker data. Previously described methods [1, 23, 26] can accommodate ungenotyped individuals which do not have offspring themselves as was shown by Hoeschele [7]. However, they do not provide the flexibility to incorporate parents with unknown genotypes which results in the loss of information for estimating marker linked effects. The described Gibbs sampling algorithm now provides this required flexibility. The innovative step in our approach is the sampling of genotypes for a marker locus that is closely linked to QTL with normally distributed allelic effects. Normality of QTL effects is a robust assumption to allow segregation of many alleles throughout a population and allow changes in allelic effects over generations, e.g. due to mutations and interactions with environments [8]. In sampling missing genotypes information from marker genotypes as well as

phenotypes of animals in the pedigree are used. Jansen et al. [9] indicate that, as a result of the use of phenotypic information, unbiased estimates of effects at the QTL can be obtained in situations where animals have been selectively genotyped.

In this paper we have concentrated on the use of information from a single marker locus. Using information from multiple linked markers can increase accuracy of predicting genetic effects at the QTL. The principles applied here have been extended to situations where genotypes for all the linked markers are known for all individuals [5, 22]. In order to incorporate individuals with unknown genotypes, the method presented in this paper needs to be extended to a multiple marker situation. In extending the method to multiple markers, the problem of reducibility deserves special attention.

Reducibility of Gibbs chains can arise when sampling genotypes for a polymorphic locus with more than two alleles [20]. The reducibility problems will become more severe when sampling genotypes for multiple linked markers. Thompson [21] suggested several, workable, approaches to guarantee irreducibility of the Gibbs chain. These approaches make use of Metropolis-coupled samplers [11], importance sampling, with 0/1 weights [15], and 'heating' in the Metropolis-Hastings steps [12]. Alternatively, Jansen et al. [9] sampled IBD values for all marker loci indicating parental origin of alleles instead of actual alleles to avoid the reducibility problem. In extending the method to multiple linked markers, attention also needs to be paid to an efficient scheme for haplotypes or genotypes at the linked loci. Updating of genotypes at closely linked loci will be more efficient when genotypes at the linked loci are updated together ('in blocks') in order to reduce auto-correlation in the Gibbs sampler [10].

For posterior inferences on the breeding value of an animal a minimum of 100 effective samples is needed. In the numerical example this minimum would correspond to a chain of 5 000 cycles which required 8 s of CPU at a HP9000 K260 server. It has been found that computing requirements increase more or less linearly with the number of animals [10]. The presented method can be applied to data originating from nucleus herds which comprise the relatively small number of genetically superior animals from the population. In a marker-assisted selection scheme marker genotypes will be collected largely on these animals, with sufficient animals having marker genotypes observed to improve selection of superior individuals.

Straightforward application in large commercial populations with thousands of marker genotypes missing, is not a valid option because of computational requirements of Markov chain Monte Carlo (MCMC) algorithms such as Gibbs sampling. Hybrid schemes will need to be developed to incorporate information from the commercial population into the marker-assisted prediction of breeding values of nucleus animals. Similar schemes have been implemented to incorporate foreign information into national evaluations in dairy cattle.

Our Bayesian approach can also be considered as a first step towards a MCMC algorithm, not necessarily Gibbs sampling, that can also estimate hyper parameters, which were held constant in this study. The next step, therefore, comprises estimation of variance components, both marked QTL and polygenic, given a fixed map position of the QTL. And, eventually, one could estimate the most likely position of the QTL within a linkage map containing multiple markers. The complete

MCMC algorithm can then be used for the analysis in QTL mapping experiments with complex pedigree structures, such as (grand-) daughter designs, in outbred populations.

ACKNOWLEDGEMENTS

Valuable suggestions by S. van der Beek and anonymous reviewers are gratefully acknowledged. The financial support of Holland Genetics is highly appreciated.

REFERENCES

[1] Fernando R.L., Grossman M., Marker-assisted selection using best linear unbiased prediction, Genet. Sel. Evol. 21 (1989) 467–477.

[2] Gelfand A.E., Smith A.F.M., Sampling-based approaches to calculating marginal densities, J. Am. Stat. Assoc. 85 (1990) 398–409.

[3] Geman S., Geman D., Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images, IEEE Trans Pattern Anal Machine Intelligence 6 (1984) 721–741.

[4] Geyer C.J., A practical guide to Markov chain Monte Carlo, Stat. Sci. 72 (1992) 320–339.

[5] Goddard M.E., A mixed model for analysis of data on multiple genetic markers, Theor. Appl. Genet. 83 (1992) 878–886.

[6] Henderson C.R., A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values, Biometrics 32 (1976) 69–83.

[7] Hoeschele I., Elimination of quantitative trait loci equations in an animal model incorporating genetic marker data, J. Dairy Sci. 76 (1993) 1693–1713.

[8] Jansen R.C., Complex plant traits: time for polygenic analysis, Trends Plant Sci. 3 (1996) 73–103.

[9] Jansen R.C., Johnson D.L., Van Arendonk J.A.M., A mixture model approach to the mapping if quantitative trait loci in complex populations with an application to multiple cattle families, Genetics (1997) (in press).

[10] Janss L.L.G., Thompson R., Van Arendonk J.A.M., (1995) Application of Gibbs sampling for inference in a mixed major gene-polygenic inheritance model in animal populations, Theor. Appl. Genet. 91 (1995) 1137–1147.

[11] Lin S., Markov chain Monte Carlo estimates of probabilities on complex structures, Ph.D. dissertation, University of Washington.

[12] Lin S., Thompson E.A., Wijsman E., Achieving irreducibility of the Markov chain Monte Carlo method applied to pedigree data, IMA J. Math. Appl. Med. Biol. 10 (1993) 1–17.

[13] Meuwissen T.H.E., VanArendonk J.A.M., Potential improvements in rate of genetic gain from marker-assisted selection in dairy cattle breeding schemes, J. Dairy Sci. 75 (1992) 1651–1659.

[14] Schaeffer L.R., Kennedy B.W., Computing strategies for solving mixed model equations, J. Dairy Sci. 69 (1986) 575–579.

[15] Sheehan N., Thomas A., On the irreducibility of a Markov chain defined on a space of genotype configurations by a sampling scheme, Biometrics 49 (1993) 163–175.

[16] Smith C., Simpson S.P., (1986) The use of genetic polymorphisms in livestock improvement, J. Anim. Breed. Genet. 103 (1986) 205–217.

[17] Soller M., Beckmann J.S., Restricted fragment length polymorphisms and genetic improvement, in: Proc. 2nd World Congress Genet. Appl. Livest. Prod., Madrid, Editorial Garsi, Madrid, Spain, vol. 6, 1982, pp. 396–404.
[18] Sorensen D.A., Wang C.S., Jensen J., Gianola D., Bayesian analysis of genetic change due to selection using Gibbs sampling, Genet. Sel. Evol. 26 (1994) 333–360.
[19] Tanner M.A., Tools for Statistical Inference, Springer-Verlag, New York, NY, 1993.
[20] Thomas D.C., Cortessis V., A Gibbs sampling approach to linkage analysis, Hum. Hered. 42 (1992) 63–76.
[21] Thompson E.A., Monte Carlo likelihood in genetic mapping. Stat. Sci. 9 (1994) 355–366.
[22] Uimari P., Thaller G., Hoeschele I., The use of multiple markers in a Bayesian method for mapping quantitative trait loci, Genetics 143 (1996) 1831–1842.
[23] VanArendonk J.A.M., Tier B., Kinghorn B.P., Use of multiple genetic markers in prediction of breeding values, Genetics 137 (1994) 319–329.
[24] VanTassell C.P., Casella G., Pollak E.J., Effects of selection on estimates of variance components using Gibbs sampling and restricted maximum likelihood, J. Dairy Sci. 78 (1995) 678–692.
[25] Wang C.S., Rutledge J.J., Gianola D., (1993) Marginal inferences about variance components in a mixed linear model using Gibbs sampling, Genet. Sel. Evol. 25 (1993) 41–62.
[26] Wang T., Fernando R.L., van der Beek S., Grossman M., Van Arendonk J.A.M., Covariance between relatives for a marked quantitative trait locus, Genet. Sel. Evol. 27 (1995) 251–272.

A1. APPENDIX

A1.1. Computation of average G *with incomplete marker data*

Wang et al. [26] suggested computing an average $\mathbf{G}$, here denoted $\overline{\mathbf{G}}$, as

$$\overline{\mathbf{G}} = \sum_{\mathbf{m}_{(k)}-1}^{n_c} \mathbf{G}_{(k)} \times p(\mathbf{m}_{(k)}|\mathbf{m}_{obs})$$

where $\mathbf{G}_{(k)}$ is the gametic relationship matrix given a particular marker genotype configuration $\mathbf{m}_{(k)}$; and $p(\mathbf{m}_{(k)}|\mathbf{m}_{obs})$ is the probability of $\mathbf{m}_{(k)}$ given $\mathbf{m}_{obs}$. This equation is not conditioned on phenotypic information.

A1.2. Marker-assisted best linear unbiased prediction of breeding values

Mixed model equations (MME) to obtain BLUE for fixed effects and BLUP for random effects are

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} & \mathbf{X}'\mathbf{W} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{A}^{-1}\alpha_u & \mathbf{Z}'\mathbf{W} \\ \mathbf{W}'\mathbf{X} & \mathbf{W}'\mathbf{Z} & \mathbf{W}'\mathbf{W} + \mathbf{G}^{-1}\alpha_v \end{bmatrix} \begin{bmatrix} \widehat{\beta} \\ \widehat{\mathbf{u}} \\ \widehat{\mathbf{v}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \\ \mathbf{W}'\mathbf{y} \end{bmatrix}$$

where $\alpha_u = \sigma_e^2/\sigma_u^2$, $\alpha_v = \sigma_e^2/\sigma_v^2$ and $\mathbf{G}$ are all known. Solutions can be obtained by iteration on the data [14]. These equations can be used in three situations. First, $\mathbf{G}$ is unique (complete marker data). Second, with missing markers, a linear estimator is obtained by taking $\mathbf{G} = \overline{\mathbf{G}}$. Third, with $\mathbf{G} = \mathbf{G}_{(k)}$, they are used to compute $E(\theta|\mathbf{G}_{(k)}, \sigma_u^2, \sigma_v^2, \sigma_e^2, \mathbf{y})$.