Genet. Res., Camb. (2004), **83**, *pp.* 143–154. *With 2 figures.* © 2004 Cambridge University Press
DOI: 10.1017/S0016672304006718 Printed in the United Kingdom

A blocking Gibbs sampling method to detect major genes with phenotypic data from a diallel mating

WEN ZENG[1], SUJIT GHOSH[1] AND BAILIAN LI[2]*
[1] *Department of Statistics, North Carolina State University, Raleigh, NC 27695-8002, USA*
[2] *Department of Forestry, Box 8002, North Carolina State University, Raleigh, NC 27695-8002, USA*

(*Received 2 June 2003 and in revised form 18 November 2003 and 11 December 2003*)

Summary

Diallel mating is a frequently used design for estimating the additive and dominance genetic (polygenic) effects involved in quantitative traits observed in the half- and full-sib progenies generated in plant breeding programmes. Gibbs sampling has been used for making statistical inferences for a mixed-inheritance model (MIM) that includes both major genes and polygenes. However, using this approach it has not been possible to incorporate the genetic properties of major genes with the additive and dominance polygenic effects in a diallel mating population. A parent block Gibbs sampling method was developed in this study to make statistical inferences about the major gene and polygenic effects on quantitative traits for progenies derived from a half-diallel mating design. Using simulated data sets with different major and polygenic effects, the proposed method accurately estimated the major and polygenic effects of quantitative traits, and possible genotypes of parents and progenies. The impact of specifying different prior distributions was examined and was found to have little effect on inference on the posterior distribution. This approach was applied to an experimental data set of Loblolly pine (*Pinus taeda* L.) derived from a 6-parent half-diallel mating. The result indicated that there might be a recessive major gene affecting height growth in this diallel population.

1. Introduction

In the classic quantitative genetic method, traits are assumed to be controlled by polygenes, i.e. many genes, with each gene having a small effect on a quantitative trait. With advances in molecular technology and computational statistics, there is strong evidence that some quantitative traits may be controlled by a number of genes that have relatively large effects on phenotype. For example, major genes or quantitative trait loci (QTL) have been found in *Drosophila* (Long *et al*., 1995), domestic animals (Piper & Shrimpton, 1989), rice (Jiang *et al*., 1994) and tree species (Wilcox *et al*., 1996; Kaya *et al*., 1999; Remington & O'Malley, 2000). A mixed-inheritance model (MIM) that includes a major gene together with polygenes, instead of strictly polygenes, has been developed for analysing some quantitative traits (Elston & Stewart, 1971; Kinghorn *et al*., 1994; Janss *et al*., 1997; Zeng & Li, 2003).

* Corresponding author. Tel: +1-919-5156845. Fax: +1-919-515-3169. e-mail: Bailian@unity.ncsu.edu

Several statistical approaches have been developed for the detection of major genes for quantitative traits using phenotypic data. Methods based on analysis of variance have been used to infer the number of major loci contributing to growth variation of interspecific aspen hybrids (Li & Wu, 1996; Wu & Li, 1999, 2000). Several statistical methods using simple non-parametric tests for departure from normality have been used for detecting major gene segregation, but not for estimating major gene effects (Mérat, 1968; Fain, 1978; Karlin & Williams, 1981; Lynch & Walsh, 1998; Zeng & Li, 2003). Other approaches based on maximum likelihood and Bayesian inference have been developed for the MIM to detect major genes affecting quantitative traits in animal (Hoeschele, 1988; Knott *et al*., 1991; Janss *et al*., 1997; Lund & Jensen, 1999), crop (Wang *et al*., 2001) and tree species (Wu *et al*., 2001). Most of these methods are based on either a multiple-generation pedigree, a progeny population derived from either a nested mating design (in the case of animal breeding) or a factorial mating design (in the case of tree hybrids).

To our knowledge, however, no statistical methods have been developed for a progeny population derived from a diallel mating design.

Diallel mating is one of the most commonly used designs in plant and tree breeding programmes (Hallauer & Miranda, 1981; Zobel & Talbert, 1984). In a diallel design parents are crossed either as male or female with other parents in a single group (Griffing, 1956). A half-diallel design is the diallel mating without self and reciprocal crosses where both half-sib and full-sib progenies are produced for each parent. Diallel mating yields two levels of polygenic effects: the general combining ability (GCA) of parents due to additive polygenic effects, and the specific combining ability (SCA) of crosses due to dominant polygenic effects. The unique feature of diallel mating, the model for an observation having two main effects, has made it difficult to analyse with standard statistical programs for even polygenetic effects (Xiang & Li, 2001). Thus, it has been difficult to incorporate genetic properties of major genes with the two levels of polygenic effects in a MIM for analysing diallel data. Because of high-dimensional marginalization of the joint density over the unknown single genotype and polygenic effects, it is practically impossible to maximize the likelihood function associated with such a model using analytical and/or numerical techniques (Le Roy *et al.*, 1989; Knott *et al.*, 1991). For animal breeding, the Gibbs sampling algorithm has been found to be reasonably effective in making inference for a MIM in a nested mating design, in which parents can be either male or female but not both (Janss *et al.*, 1997). Such analyses were primarily based on the half-sib relationships of parents (male or female) and their progenies. In the case of tree-breeding programmes, diallel progenies are usually planted at several locations or site types to determine their growth potential under different environments. The potentially large environmental variation and genotype by environmental interaction, relative to animal breeding, may affect the statistical power for major gene detection (McKeand *et al.*, 1997). Although the Bayesian approach may have potential for major gene detection, its usefulness for MIM analysis of diallel data with the two types of polygenic effects and the heterogeneous environmental variance is unknown.

In this study, we developed and evaluated a Bayesian approach, using a parent blocking Gibbs sampling, to make inferences about major genes and polygenic effects (GCA and SCA) that control quantitative traits for a progeny population derived from a half-diallel mating design without self and reciprocal crossings. Computer simulations were done to examine the effects of different prior distributions and design matrix, either full-ranked or non-full-ranked, on the proposed statistical method. A case study with one half-diallel progeny population of Loblolly pine (*Pinus taeda* L.) was used to detect a major gene for height growth and to illustrate the application of the method.

2. The mixed-inheritance model

A mixed-inheritance model (MIM) is adopted in this study for the diallel analysis, in which phenotypes are assumed to be influenced by a single major gene and by polygenic effects. A half-diallel mating design, with n_p parents selected from a base population under Hardy–Weinberg and linkage equilibrium (Falconer & Mackay, 1996), and $n_p(n_p-1)/2=n_s$ full-sib families, is used. Each full-sib family is tested at several sites, in a randomized complete block design with several trees per full-sib family within each block, and several blocks within each site. The statistical model for a MIM can be written as a mixed linear model:

$$\mathbf{Y}=\mathbf{X\mu}+\mathbf{Zu}+\mathbf{WLm}+\mathbf{e}. \tag{1}$$

The notations for this diallel model are listed in Table 1. Unlike MIM for an animal population, there is no incidence matrix $\mathbf{Z}$ for the major gene effect, because data $\mathbf{Y}$ are only the phenotypic observations of progenies in a tree population, rather than progenies plus parents as in an animal population. By assuming $\mathbf{e}$ to be normally distributed $\mathbf{N}(\mathbf{0},\sigma_e^2\mathbf{I})$ and giving location and scale parameters, the vector of data $\mathbf{Y}$ is also normally distributed as:

$$\mathbf{Y}\mid\boldsymbol{\mu},\mathbf{u},\mathbf{W},\mathbf{m},\sigma_e^2\sim\mathbf{N}(\mathbf{X\mu}+\mathbf{Zu}+\mathbf{WLm},\sigma_e^2\mathbf{I}).$$

The single major gene under the traditional genetic model of one gene with two alleles (Falconer & Mackay, 1996) is assumed to be a bi-allelic (A_1 and A_2), autosomal locus with Mendelian transmission probabilities, such that each progeny has one of the three possible genotypes: A_1A_1, A_1A_2 and A_2A_2 with genetic effects a, d, and $-a$ respectively. For progeny, k ($k=1,\ldots,n$), the genotype is represented as a random vector $\mathbf{w}_k$, (1, 0, 0), (0, 1, 0) and (0, 0, 1) correspond to the three possible genotypes of A_1A_1, A_1A_2 and A_2A_2 respectively. Given the two parent genotypes $\mathbf{w}_{p1(k)}$ and $\mathbf{w}_{p2(k)}$, the genotype distribution of progeny k is denoted as $p(\mathbf{w}_k\mid\mathbf{w}_{p1(k)},\mathbf{w}_{p2(k)})$. This distribution describes the probability of alleles constituting genotype $\mathbf{w}_k$ being transmitted from parents with genotypes $\mathbf{w}_{p1(k)}$ and $\mathbf{w}_{p2(k)}$ when segregation of allele follows Mendelian transmission probabilities. Because of the conditionally independent structure of the genotypes, the joint genotype distribution of progenies can be written as:

$$p(\mathbf{W}\mid\mathbf{w}_p)=\prod_{k=1}^{n}p(\mathbf{w}_k\mid\mathbf{w}_{p1(k)},\mathbf{w}_{p2(k)}) \tag{2}$$

where $\mathbf{w}_p$ are the genotypes of n_p parents. The parent genotypes are sampled from a base population in Hardy–Weinberg equilibrium (Falconer & Mackay,

https://doi.org/10.1017/S0016672304006718 Published online by Cambridge University Press

Table 1. *Definitions of the notation used in the mixed-inheritance model*

Notation	Definition
$\mathbf{Y}$	A $(n\times 1)$ vector of n progeny observations
$\boldsymbol{\mu}$	The overall mean, equal to μ. It can be extended to a $(c\times 1)$ vector of c fixed non-genetic effects, e.g. site effect and block within-site effect
$\mathbf{X}$	A $(n\times 1)$ vector with value 1 for the overall mean for all progenies
$\mathbf{u}$	A $(q\times 1)$ vector of q random polygenic effects, $\mathbf{u}'=(\mathbf{g}',\mathbf{s}')$, including n_p GCAs ($\mathbf{g}$) and n_s SCAs ($\mathbf{s}$)
$\mathbf{g}$	$\mathbf{g}'=\{g_i, i=1,\ldots,n_p\}$, n_p GCAs, are assumed to be mutually independent normal distributions, i.e. $\mathbf{g}\mid\sigma_g^2\sim\mathbf{N}(\mathbf{0},\sigma_g^2\mathbf{I})$
σ_g^2	GCA polygenic variance due to additive polygenic effects
$\mathbf{s}$	$\mathbf{s}'=\{s_j, j=1,\ldots,n_s\}$, n_s, SCAs, are assumed to be mutually independent normal distributions, i.e. $\mathbf{s}\mid\sigma_s^2\sim\mathbf{N}(\mathbf{0},\sigma_s^2\mathbf{I})$
σ_s^2	SCA polygenic variance due to dominance polygenic effects
$\mathbf{Z}$	A $(n\times q)$ incidence matrix of GCA and SCA for all progenies
$\mathbf{m}$	A (2×1) vector of major gene effects, $\mathbf{m}'=(a,d)$
a	Additive major genotypic effect
d	Dominance major genotypic effect
$\mathbf{L}$	$\mathbf{L}=\begin{bmatrix}1&0\\0&1\\-1&0\end{bmatrix}$, a (3×2) indicator matrix of the major gene effects for major genotypes
$\mathbf{W}$	An unknown $(n\times 3)$ random incidence matrix of major genotypes at the single locus for n progenies
$\mathbf{W}_T$	A (1×3) row vector to form the rows of $\mathbf{W}$. $\mathbf{w}_1=(1,0,0)$, $\mathbf{w}_2=(0,1,0)$ and $\mathbf{w}_3=(0,0,1)$, T taking values 1, 2 and 3 respectively to represent the major genotypes A_1A_1, A_1A_2 and A_2A_2
$\mathbf{e}$	A $(n\times 1)$ vector of iid errors. $\mathbf{e}$ is assumed to be $\mathbf{N}(\mathbf{0},\sigma_e^2\mathbf{I})$
σ_e^2	Residual variance

N represents the multivariate normal distribution.

1996). This is a reasonable assumption because individual trees serving as parents are usually selected randomly from natural populations. Given the favourable allele frequency in the base population $f=p(A_1)$, the probability distribution of the genotype of parent i is assumed to be $p(\mathbf{w}_{pi}\mid f)$, which follows Hardy–Weinberg proportions. As parents are sampled independently, their joint genotype distribution can be written as:

$$p(\mathbf{w}_p\mid f)=\prod_{i=1}^{n_p}p(\mathbf{w}_{pi}\mid f). \tag{3}$$

In order to fully specify the Bayesian model, normal priors are assigned to the overall mean $\boldsymbol{\mu}=\mu$, and major gene effects $\mathbf{m}'=(a,d)$, i.e. $\mu\sim N(0,k_1^2)$, $a\sim N(0,k_2^2)$ and $d\sim N(0,k_3^2)$, where k_i^2, $i=1$, 2 and 3 are the hyper-parameters of the prior distribution. In the simulation and real data analysis we used $k_i=4$, $i=1,2,3$. Variance components, σ_g^2, σ_s^2 and σ_e^2 are assumed to arise independently from conjugate inverted gamma distributions (IG), i.e. $\sigma_g^2\sim IG(\gamma_1,\nu_1)$, $\sigma_s^2\sim IG(\gamma_2,\nu_2)$, and $\sigma_e^2\sim IG(\gamma_3,\nu_3)$, where γ_i and ν_i are hyper-parameters. For our application, we used $\gamma_i=2$, and $\nu_i=(\gamma_i-1)*\hat{\sigma}_i$, $i=1$, 2 and 3 for $\hat{\sigma}_1=\hat{\sigma}_g$, $\hat{\sigma}_2=\hat{\sigma}_s$ and $\hat{\sigma}_3=\hat{\sigma}_e$, where $\hat{\sigma}_i$ are obtained from a preliminary study using frequency distribution method. The conjugate Beta prior is used for the allele frequency, i.e. $p(f)\sim\beta(\alpha_f,\beta_f)$, where α_f and β_f are prior distribution parameters. We have chosen $\alpha_f=\beta_f=1$ to express prior ignorance.

The joint posterior density of all unknowns, given the data $\mathbf{Y}$, is proportional to the product of the likelihood function and the prior densities:

$$\begin{aligned}
&p(\boldsymbol{\mu},\mathbf{m},\mathbf{u},\mathbf{W},\mathbf{w}_p,f,\sigma_g^2,\sigma_s^2,\sigma_e^2\mid\mathbf{Y})\\
&\quad\propto p(\mathbf{Y}\mid\boldsymbol{\mu},\mathbf{m},\mathbf{u},\mathbf{W},\sigma_e^2)p(\sigma_e^2)p(\mathbf{g}\mid\sigma_g^2)p(\mathbf{s}\mid\sigma_s^2)p(\sigma_g^2)p(\sigma_s^2)p(\boldsymbol{\mu})p(a)p(d)p(\mathbf{W}\mid\mathbf{w}_p)p(\mathbf{w}_p\mid f)p(f)\\
&\quad\propto(\sigma_e^2)^{-\frac{n}{2}}\exp\left\{-\frac{1}{2\sigma_e^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\mu}-\mathbf{Z}\mathbf{u}-\mathbf{W}\mathbf{L}\mathbf{m})'(\mathbf{y}-\mathbf{X}\boldsymbol{\mu}-\mathbf{Z}\mathbf{u}-\mathbf{W}\mathbf{L}\mathbf{m})\right\}(\sigma_e^2)^{-(\gamma_3+1)}\exp\left\{-\frac{\nu_3}{\sigma_e^2}\right\}(\sigma_g^2)^{-\frac{n_g}{2}}\\
&\qquad\times\exp\left\{-\frac{1}{2\sigma_g^2}\sum_{i=1}^{n_g}g_i^2\right\}(\sigma_g^2)^{-(\gamma_1+1)}\exp\left\{-\frac{\nu_1}{\sigma_g^2}\right\}(\sigma_s^2)^{-\frac{n_s}{2}}\exp\left\{-\frac{1}{2\sigma_s^2}\sum_{j=1}^{n_s}s_j^2\right\}(\sigma_s^2)^{-(\gamma_2+1)}\exp\left\{-\frac{\nu_2}{\sigma_s^2}\right\}\\
&\qquad\times\exp\left\{-\frac{1}{2k_1^2}\mu^2\right\}\exp\left\{-\frac{1}{2k_2^2}a^2\right\}\exp\left\{-\frac{1}{2k_3^2}d^2\right\}\prod_{k=1}^{n}p(\mathbf{w}_k=\mathbf{w}_g\mid\mathbf{w}_{p1(k)},\mathbf{w}_{p2(k)})\\
&\qquad\times\prod_{i=1}^{n_g}p(\mathbf{w}_{pi}=\mathbf{w}_g\mid f)f^{\alpha_f-1}(1-f)^{\beta_f-1}.
\end{aligned}\tag{4}$$

https://doi.org/10.1017/S0016672304006718 Published online by Cambridge University Press

In order to study the effects of the prior on the method's behaviour with this data structure, improper flat priors are used for the overall mean μ, and major gene effect a and d, besides normal priors. The prior for any σ^2 is always inverted gamma.

3. Gibbs sampling

(i) *Parent blocking*

In order to make statistical inferences about unknowns, the marginal posterior distributions for the model parameters are of interest, but it appears to be almost impossible to obtain such marginals for our model. In analytical approaches, the study of marginal densities would require integration and/or summation. Often such marginalizations are not feasible to compute or even express in closed form for a high-dimensional model such as MIM, as in (4), but this difficulty can be circumvented by means of simulation-based methods. The Gibbs sampler is based on sampling random varieties from a Markov chain with its stationary distribution as the posterior distribution, and the sampling from the Markov chain used to perform the high-dimensional Monte Carlo integration (Gelfand & Smith, 1990; Brooks, 1998). Samples are obtained from the full conditional distributions, which form the transition probabilities of the Markov chain. Each time a full conditional distribution is visited, it is used to sample the corresponding parameter, while other parameters are considered to be fixed, and then the realized value is substituted into the full conditional distribution of all other parameters.

To improve the mixing and hence the speed of convergence, it is possible to sample several parameters simultaneously, called a 'block', from their joint conditional distribution instead of updating all parameters univariately. As long as all parameters are updated, the new Markov chain will still have (4) as the density of its stationary distribution. Unlike an animal population where data $\mathbf{Y}$ include parents and their offspring, and usually span several generations, in a tree population we consider only the progeny observations in the data $\mathbf{Y}$ with just two generations. The sire block strategy (Janss *et al*., 1997) has worked well for animal populations. Since in a diallel mating design one tree served as a male as well as a female, we modified 'a sire blocking' into 'a parent blocking'. In a parent blocking, the genotypes of a parent and its half-sib offspring are treated as a block and updated simultaneously. Consequently, in each cycle the genotype of every offspring is updated twice instead of once as each offspring has two parents. Given the work of Liu *et al*. (1994) and Roberts & Sahu (1997), it seemed to us that the block Gibbs sampler would mix faster than the ordinary one-at-a-time version that updates each component sequentially. Blocking is generally effective when the elements within the block are highly correlated compared with the correlation between blocks.

(ii) *Full conditional distributions*

Full conditional distributions are derived from the joint posterior distribution (4). For notational convenience, the MIM can be rewritten as: $\mathbf{Y}=\mathbf{H}\boldsymbol{\theta}+\mathbf{e}$, where $\mathbf{H}=[\mathbf{X}:\mathbf{WL}:\mathbf{Z}]$ is a $(n\times p)$ matrix, and $\boldsymbol{\theta}'=(\boldsymbol{\mu}',\mathbf{m}',\mathbf{u}')=(\mu,a,d,\mathbf{g}',\mathbf{s}')=(\theta_1,\theta_2,\ldots,\theta_p)$ is a $(p\times 1)$ parameter vector.

In order to implement the 'parent blocking', an exact calculation of the joint conditional distribution of a parent and all its offspring is required. The joint conditional distribution for parent i is:

$$p(\mathbf{w}_{pi},\mathbf{w}_{i(1)},\ldots,\mathbf{w}_{i(n_i)}\,|\,\mathbf{W}_{-i(l)},\mathbf{w}_{-pi},\boldsymbol{\theta},f,\sigma_g^2,\sigma_s^2,\sigma_e^2,\mathbf{Y})$$

where n_i denotes the number of offspring of parent i, and the offspring are indexed by $i(1), i(2),\ldots, i(n_i)$, or simply $i(l)$, where $l=1,\ldots,n_i$. By definition, this distribution is proportional to

$$p(\mathbf{w}_{pi}\,|\,\mathbf{W}_{-i(l)},\mathbf{w}_{-pi},\boldsymbol{\theta},f,\boldsymbol{\sigma}_g^2,\boldsymbol{\sigma}_s^2,\boldsymbol{\sigma}_e^2,\mathbf{Y}) * p(\mathbf{w}_{i(1)},\ldots,\mathbf{w}_{i(n_i)}\,|\,\mathbf{W}_{-i(l)},\mathbf{w}_p,\boldsymbol{\theta},f,\boldsymbol{\sigma}_g^2,\boldsymbol{\sigma}_s^2,\boldsymbol{\sigma}_e^2,\mathbf{Y}).$$

The first term is the genotypic distribution of the parent i, marginalized with respect to the genotypes of its offspring. The second term is the joint distribution of offspring genotypes conditional on the parents' genotypes. To calculate the genotype distribution of parent i, the three possible genotypes of all offspring must be summed after weighting each genotype by its relative probability. The final marginalized full conditional distribution for the major genotype of the parent i is:

$$\begin{aligned}p(\mathbf{w}_{pi}=\mathbf{w}_T\,|\,\mathbf{W}_{-i(l)},\mathbf{w}_{-pi},\boldsymbol{\theta},f,\sigma_g^2,\sigma_s^2,\sigma_e^2,\mathbf{Y})&\\ \propto p(\mathbf{w}_{pi}=\mathbf{w}_T\,|\,f) * \prod_{k\in i(l)}\sum_{b=1}^{3}p(\mathbf{w}_k=\mathbf{w}_b\,|\,\mathbf{w}_{p1}=\mathbf{w}_T,\mathbf{w}_{p2})&\\ \times p(\tilde{y}_k\,|\,\mathbf{w}_k=\mathbf{w}_b)&\qquad(5a)\end{aligned}$$

where $\tilde{y}_k=y_k-\mu-\mathbf{Z}_k\mathbf{u}$ is called the adjusted record, $\mathbf{Z}_k$ are the kth rows of the matrices $\mathbf{Z}$, and $\mathbf{w}_b$ has the same notation as $\mathbf{w}_T$. The penetrance function (or weight) is:

$$p(\tilde{y}_k\,|\,\mathbf{w}_k=\mathbf{w}_b)\propto\exp\left\{-\frac{1}{2\sigma_e^2}(\tilde{y}_k-\mathbf{w}_k\mathbf{Lm})^2\right\}.$$

The probabilities here are given up to a constant of proportionality and must be normalized to ensure that

$$\sum_{T=1}^{3}p(\mathbf{w}_{pi}=\mathbf{w}_T)=1.$$

https://doi.org/10.1017/S0016672304006718 Published online by Cambridge University Press

For the genotypes of offspring, the marginalized full conditional distributions are the same as the usual full conditional distributions found by extracting from (4) the term in which $\mathbf{w}_k$, $k \in i(l)$, is present. i.e.

$$p(\mathbf{w}_k = \mathbf{w}_T \mid \mathbf{W}_{-k}, \mathbf{w}_{-p}, \boldsymbol{\theta}, f, \sigma_g^2, \sigma_s^2, \sigma_e^2, \mathbf{Y})$$
$$\propto p(\mathbf{w}_k = \mathbf{w}_T \mid \mathbf{w}_{p1(k)}, \mathbf{w}_{p2(k)}) * p(\tilde{y}_k | \mathbf{w}_k = \mathbf{w}_T). \quad (5b)$$

The full conditionals for allele frequency, location parameters and variance components are obtained by extracting the relevant terms from the joint posterior density in (4) (for details see Zeng, 2000).

(iii) *Updating scheme*

The algorithm based on parent block updating is summarized as the following:

I. Initiate $\mathbf{W}$, $\mathbf{w}_p$, $\boldsymbol{\theta}$, $\sigma_g^2\sigma_s^2$, and σ_e^2 with some reasonable starting values.
II. Sample major genotypes $\mathbf{w}_{pi}$, $i=1,\ldots,n_p$, $\mathbf{w}_k$, $k=1,\ldots,n$ from full conditional distributions, (5a) and (5b), by parent blocking. The updating scheme is to: (1) update parent 1 and its offspring, $\mathbf{w}_{p1}$ plus $\mathbf{w}_{p1(l)}$, in one block, with $\mathbf{w}_{p2}^{[t]},\ldots, \mathbf{w}_{pn_p}^{[t]}$ known; (2) update parent 2 and its offspring $\mathbf{w}_{p2}$ plus $\mathbf{w}_{p2(l)}$, in one block, with $\mathbf{w}_{p1}^{[t+1]}$, $\mathbf{w}_{p3}^{[t]},\ldots, \mathbf{w}_{pn_p}^{[t]}$ known; (3) update parent n_p and its offspring, $\mathbf{w}_{pn_p}$ plus $\mathbf{w}_{pn_p(l)}$, in one block with $\mathbf{w}_{p1}^{[t+1]},\ldots, \mathbf{w}_{p(n_p-1)}^{[t+1]}$ known; each offspring updates twice in each cycle. The parents may be updated in any order. In fact, it is possible to update in random order in each iteration.
III. Sample allele frequency f, location parameters θ, and variance components σ_g^2, σ_s^2 and σ_e^2 from the full conditional distribution.
IV. Repeat II–III, these steps constituting one iteration.

4. SIMULATION

(i) *Data generation process*

To evaluate the procedure with this data structure, simulated data with both major gene and polygenic components were generated for this study. A 6-parent half-diallel mating design with 4 test sites and 6 blocks per site is used to simulate phenotypic observations, although the site effects and block within-site effects are both set to zero. Six parents are chosen randomly from a base population in which the major gene and polygenic parameters are defined. There are 15 full-sib families, 6 progenies per family per block per test site, and a total of 2160 progenies across 4 test sites.

For all progenies, phenotypic observations are simulated according to model (1). The polygenic effect ($\mathbf{Zu}$) includes G_1+G_2+S, where G_1 and G_2 are GCA effects for 2 parents with prior distribution $N(0, \sigma_g^2)$, and S is an SCA effect with prior distribution $N(0, \sigma_s^2)$. Two genetic parameters – the narrow-sense heritability of polygenic inheritance $h^2=4\sigma_g^2/\sigma_p^2$ (where σ_p^2 is the total phenotypic variance), and the ratio of dominance to additive genetic variance of polygenic inheritance, $r=\sigma_s^2/\sigma_g^2$ – are used to calibrate these polygenic quantities (Huber *et al.*, 1992).

The major genotypes of parents and progenies are simulated according to (3) and (2). The major gene variance component is calculated as: $\sigma_m^2=\sigma_a^2+\sigma_d^2=2f(1-f)[(1-2f)d+a]^2+[2f(1-f)d]^2$. The total phenotypic variance is $\sigma_p^2=\sigma_m^2+2\sigma_g^2+\sigma_s^2+\sigma_e^2=1{\cdot}0$. The relationship between polygenic effects and major gene effects is assumed to be additive. For our simulation, the parameters are set to $h^2=0{\cdot}2$, $r=0{\cdot}5$, $a=1{\cdot}0$, $d=0{\cdot}0$ and $f=0{\cdot}2$. The realized favourable allele frequency (f) is 0·167. The major genotypes of the 6 parents are A_2A_2, A_2A_2, A_2A_2, A_1A_2, A_1A_2 and A_2A_2. The effect of the single gene in this case is to be detected and estimated.

(ii) *Effects of prior distribution, initial value and design matrix*

For the overall mean (μ), the additive major gene effect (a) and the dominance major gene effect (d), the priors $p(\mu)$, $p(a)$ and $p(d)$ are chosen as a flat distribution, and are proportional to a constant, or chosen as a normal distribution, and have $N(0, K^2)$, where $K=4$ is used in the analysis. Both uniform and normal priors are used in the model, respectively, to see the effect of priors on posterior inference.

The design matrix for the random polygenic effects (GCAs and SCAs) can be either full rank by putting constraints $\sum_{i=1}^{n_g} g_i=0$, $\sum_{j=1}^{n_s} s_j=0$, or singular. Both design matrices are used to test its effect on the Markov chain Monte Carlo (MCMC) method, especially its convergence.

Initial values of GCA, SCA and variance components, σ_g^2, σ_s^2, σ_e^2, are obtained from the traditional genetic model analysis (without major gene effect). These estimates are used as initial values for the Markov chain. For the major gene, the ranges for a, d and f are [0·0, 1·0], [0·0, 0·5] and [0·1, 0·5], respectively (Table 2). The genotypes of parents are generated by f, assuming that parents are all from a base population with Hardy–Weinberg and linkage equilibrium. Given the parents' genotypes, major genotypes of progenies are generated by following the Mendelian transmission probabilities of allele segregation. These multiple independent parallel runs of Gibbs sampler can be used as a diagnostic tool to examine the mixing property of MCMC. For each case, two independent chains, with 40 000 iterations each, are run.

(iii) *Convergence diagnostics*

Bayesian Output Analysis (BOA version 0.5.0) (Smith, 2000) is used to analyse these outputs. The

https://doi.org/10.1017/S0016672304006718 Published online by Cambridge University Press

Table 2. *The six combinations of initial values (N0–N5) for each of three combinations of prior and design matrix: uniform priors for* a, d *and* μ *with singular design matrix, normal priors for* a, d, *and* μ *with singular design matrix, and normal priors for* a, *and* μ *with full-rank design matrix to test the effect of prior, initial value, and design matrix on the mixed inheritance model (MIM) using a blocking Gibbs sampling*

Initial value Case/parameter	*a*	*d*	*f*
N0 (true values)	1·0	0·0	0·2
N1	0·0	0·0	0·1
N2	0·25	0·25	0·2
N3	0·5	0·0	0·3
N4	0·75	0·5	0·4
N5	1·0	0·0	0·5

In the MIM, the polygene background set up was narrow-sense heritability of polygenic inheritance $h^2=0{\cdot}2$, the ratio of dominance to additive genetic variance of polygenic inheritance $r=0{\cdot}5$, and the major gene effect was simulated by $2a$, d, and f

Gelman and Rubin Shrink Factors (Gelman & Rubin, 1992) plot is used to determine the burn-in time as well as the convergence. The autocorrelation plot is then used to determine the length of thinning lag in order to obtain a relatively independent sample for the final analysis. Brooks, Gelman and Rubin's corrected scale reduction factors (for multiple chains) and Raftery and Lewis's dependence factors (for a single chain) are also used to diagnose the convergence of MCMC chains (Brooks & Roberts, 1998). As a rule of thumb, if the 0·975 quantile of corrected scale reduction factors is less than 1·2, the sample may be considered to have arisen from the stationary distribution. For a single chain, dependence factors greater than 5·0 often indicates convergence failure and a need to reparameterize the model. Trace plots are used as indicators of mixing and convergence of chains.

In the Gibbs chain, the additive major gene effect (a) may be positive as well as negative. The sign of a is relevant, i.e. the favourable allele is A_1 when a is positive and A_2 when a is negative. From the Gibbs samples, we are interested in the absolute value of a. For consistency, we change the frequency of the favourable allele (f) to $1-f$ when a is changed from a negative to a positive value.

5. Results

When the design matrix was singular, both uniform prior and normal prior provided good frequentist coverage estimates, except for the fact that σ_g^2 estimates were lower than expected from simulations (Table 3). The Gelman and Rubin plot for a set of initial values (as in N1) indicated that the burn-in iteration was about 25 000 iterations (see Fig. 1). The corrected scale reduction factors were approximately 1·0, and Raftery dependence factors were found to be much less than 5·0. Posterior densities of six genetic parameters – a, d, f, σ_e^2, σ_g^2 and σ_s^2 – for a different set of initial values (as in N1) are listed in Fig. 2. These numerical diagnostic summaries (Table 4, Fig. 1) indicated that the chains mix well and there was no severe problem with MCMC convergence. We also found that the prior and initial values did not have any effects on the Gibbs sampler under the singular design matrix.

When the design matrix was chosen to be of full rank, each individual chain converged with the Raftery dependence factors less than 5·0. The parameter estimates from the five different initial value sets N0, N1, N2, N4 and N5 showed good precision. For the set of initial values N3, however, the estimates of major gene genotype for the sixth parent, as well as the corresponding genetic parameters, were not correct (see Table 3). As a result, the 0·975 quantile of corrected scale reduction factors for parameters d, f and σ_g^2 were 1·28, 1·29 and 2·87 respectively (see Table 4). This may indicate a possible mixing problem for the N3 set due to the combination of full-rank design matrix with this data structure. To avoid this possibility, a normal prior with a singular design matrix was chosen as a model for the further analysis of our case study.

6. A case study

The blocking Gibbs sampling method was applied to a progeny data set derived from a 6-parent, half-diallel mating of Loblolly pine by the North Carolina State University Tree Improvement Program (Li *et al.*, 1999). Similar to the simulated data, 15 full-sib families from the diallel mating were planted at 4 different sites with 6 blocks each. Tree heights of progenies at age 6 years were the quantitative trait for this analysis. First, a linear model was fitted to adjust the fixed effects of site and block within site from phenotypic observations. The residuals were used as the phenotypic observation vector **Y**. Initial values for GCA, SCA and variance components were all taken from the estimates from a traditional polygenic model without the major gene. The prior distributions for μ, a and d were normal $\sim N(0, K^2)$ with $K=4$. The hyperparameters for allele frequency f were $\alpha_f=1$ and $\beta_f=1$. A singular design matrix was chosen for the analysis.

The results for two independent chains, with different sets of initial values for a, d and f, were very close to each other using 240 000 iterations (Table 5).

https://doi.org/10.1017/S0016672304006718 Published online by Cambridge University Press

Table 3. *Estimated means, and standard deviations of posterior densities for six genetic parameters* (a, d, f, σ^2_e, σ^2_g, σ^2_s), *and genotype estimates of six parents* (P_1–P_6) *for testing the effects of prior distribution, initial value and design matrix on the mixed inheritance model*

Parameters:	*a*	*d*	*f*	σ^2_e	σ^2_g	σ^2_s	P_1	P_2	P_3	P_4	P_5	P_6
True values	1·0	0·0	0·167	0·595	0·034	0·017	A_2A_2	A_2A_2	A_2A_2	A_1A_2	A_1A_2	A_2A_2
Uniform prior												
N0	1·03 ± 0·12[a]	−0·07 ± 0·12	0·212 ± 0·102	0·613 ± 0·027	0·017 ± 0·012	0·013 ± 0·008	A_2A_2	A_2A_2	A_2A_2	A_1A_2	A_1A_2	A_2A_2
N1	1·04 ± 0·12	–0·08 ± 0·12	0·215 ± 0·103	0·613 ± 0·027	0·017 ± 0·011	0·013 ± 0·007	A_2A_2	A_2A_2	A_2A_2	A_1A_2	A_1A_2	A_2A_2
N2	1·03 ± 0·12	−0·07 ± 0·13	0·215 ± 0·106	0·612 ± 0·028	0·018 ± 0·014	0·013 ± 0·007	A_2A_2	A_2A_2	A_2A_2	A_1A_2	A_1A_2	A_2A_2
N3	1·02 ± 0·13	−0·06 ± 0·13	0·212 ± 0·104	0·614 ± 0·028	0·017 ± 0·010	0·013 ± 0·008	A_2A_2	A_2A_2	A_2A_2	A_1A_2	A_1A_2	A_2A_2
N4	1·03 ± 0·11	−0·07 ± 0·12	0·215 ± 0·107	0·612 ± 0·027	0·016 ± 0·011	0·013 ± 0·008	A_2A_2	A_2A_2	A_2A_2	A_1A_2	A_1A_2	A_2A_2
N5	1·02 ± 0·12	−0·07 ± 0·12	0·212 ± 0·106	0·614 ± 0·028	0·017 ± 0·011	0·013 ± 0·008	A_2A_2	A_2A_2	A_2A_2	A_1A_2	A_1A_2	A_2A_2
Normal prior												
N0	1·04 ± 0·12	−0·08 ± 0·12	0·217 ± 0·108	0·613 ± 0·028	0·017 ± 0·011	0·013 ± 0·007	A_2A_2	A_2A_2	A_2A_2	A_1A_2	A_1A_2	A_2A_2
N1	1·03 ± 0·13	−0·07 ± 0·13	0·215 ± 0·105	0·613 ± 0·028	0·017 ± 0·012	0·013 ± 0·007	A_2A_2	A_2A_2	A_2A_2	A_1A_2	A_1A_2	A_2A_2
N2	1·04 ± 0·12	−0·08 ± 0·13	0·212 ± 0·104	0·612 ± 0·028	0·017 ± 0·010	0·013 ± 0·008	A_2A_2	A_2A_2	A_2A_2	A_1A_2	A_1A_2	A_2A_2
N3	1·01 ± 0·13	−0·05 ± 0·13	0·216 ± 0·107	0·614 ± 0·027	0·017 ± 0·011	0·013 ± 0·007	A_2A_2	A_2A_2	A_2A_2	A_1A_2	A_1A_2	A_2A_2
N4	1·03 ± 0·12	−0·07 ± 0·12	0·215 ± 0·105	0·613 ± 0·026	0·017 ± 0·011	0·013 ± 0·008	A_2A_2	A_2A_2	A_2A_2	A_1A_2	A_1A_2	A_2A_2
N5	1·04 ± 0·11	−0·08 ± 0·11	0·214 ± 0·107	0·611 ± 0·028	0·017 ± 0·028	0·013 ± 0·008	A_2A_2	A_2A_2	A_2A_2	A_1A_2	A_1A_2	A_2A_2
Normal prior with full-rank design matrices												
N0	1·02 ± 0·11	−0·07 ± 0·12	0·212 ± 0·106	0·615 ± 0·028	0·015 ± 0·010	0·012 ± 0·006	A_2A_2	A_2A_2	A_2A_2	A_1A_2	A_1A_2	A_2A_2
N1	1·02 ± 0·13	−0·07 ± 0·13	0·215 ± 0·106	0·616 ± 0·027	0·016 ± 0·011	0·012 ± 0·007	A_2A_2	A_2A_2	A_2A_2	A_1A_2	A_1A_2	A_2A_2
N2	1·03 ± 0·11	−0·08 ± 0·12	0·214 ± 0·108	0·614 ± 0·028	0·015 ± 0·010	0·012 ± 0·007	A_2A_2	A_2A_2	A_2A_2	A_1A_2	A_1A_2	A_2A_2
N3	0·95 ± 0·10	**−0·20 ± 0·07**	**0·354 ± 0·123**	0·634 ± 0·033	**0·100 ± 0·067**	0·009 ± 0·006	A_2A_2	A_2A_2	A_2A_2	A_1A_2	A_1A_2	A_1A_1
N4	1·02 ± 0·14	−0·06 ± 0·14	0·216 ± 0·141	0·615 ± 0·105	0·015 ± 0·010	0·012 ± 0·007	A_2A_2	A_2A_2	A_2A_2	A_1A_2	A_1A_2	A_2A_2
N5	1·01 ± 0·12	−0·05 ± 0·13	0·216 ± 0·106	0·614 ± 0·106	0·015 ± 0·009	0·012 ± 0·006	A_2A_2	A_2A_2	A_2A_2	A_1A_2	A_1A_2	A_2A_2
MC error[b]	0·006	0·006	0·002	0·0006	0·0003	0·0002						

In the MIM model, the polygene background set up was $h^2 = 0{\cdot}2$, $r = 0{\cdot}5$, and the major gene effect was simulated by $2a = 2{\cdot}0$, $d = 0{\cdot}0$ and $f = 0{\cdot}2$ (actual value is 0·167). There were six runs, N0 to N5, for each of three cases.

[a] Mean ± standard deviation of the parameter estimate from the Gibbs sample with lag = 5.

[b] There is one MC error for each chain. Because the values for each parameter are so close, only an average value is listed here.

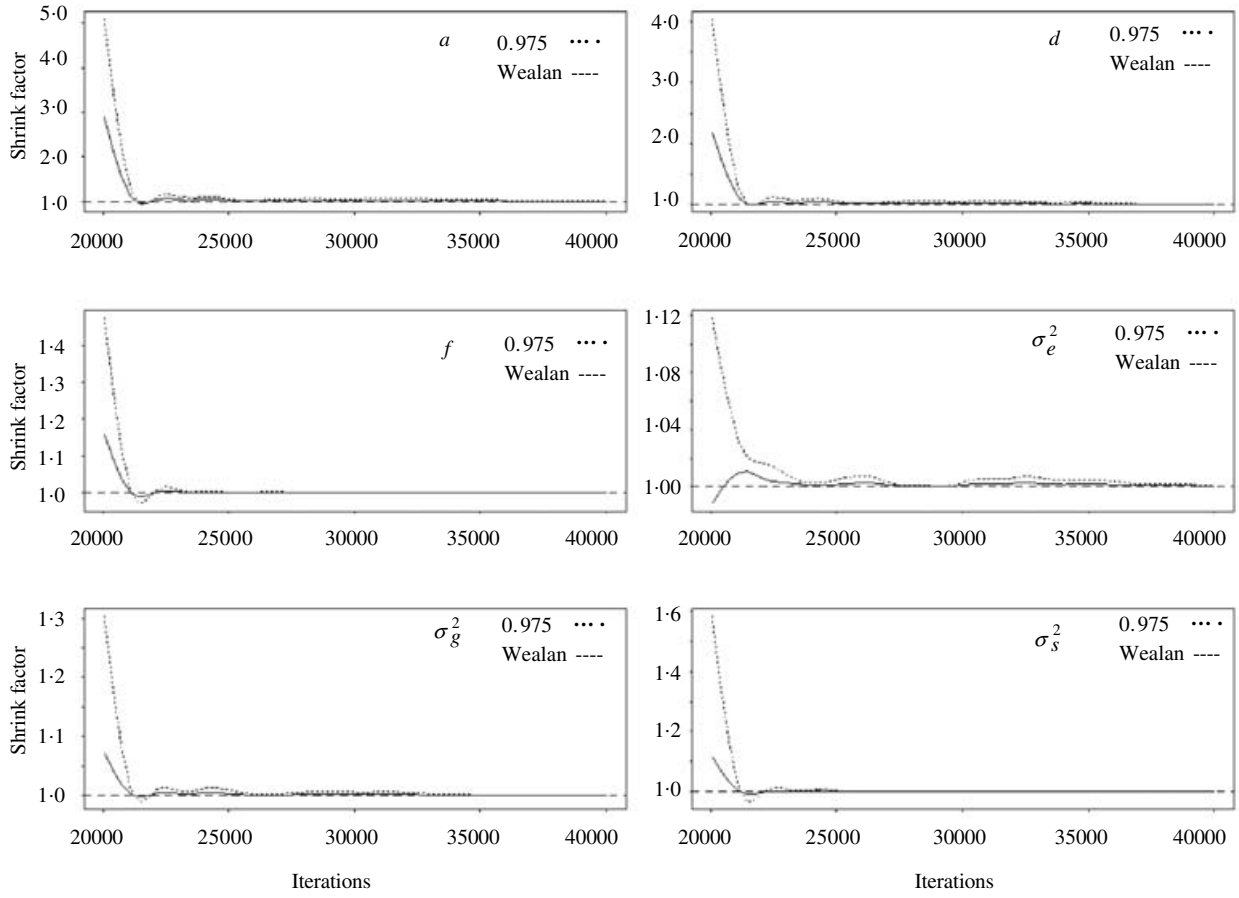

Fig. 1. Gelman and Rubin shrink factor plots of six genetic parameters (a, d, f, σ_e^2, σ_g^2, σ_s^2) of the initial value set N1 for normal prior for a, d and f with singular design matrix for testing the effects of prior distribution, initial values, and design matrix on the mixed-inheritance model.

The 0·975 quantile of corrected scale reduction factors for parameters was less than 1·2. The percentage of major gene effects was estimated as 17% of the total phenotypic variance. The estimated major genotype of parent 2 was A_1A_2, while that of the others was A_2A_2. The additive effect of the major gene was $a \approx 2{\cdot}3$ and the dominance effect was $d \approx -2{\cdot}3$. That is, A_1A_1 had $a \approx 2{\cdot}3$, A_1A_2 had $d \approx -2{\cdot}3$ and A_2A_2 had $a \approx -2{\cdot}3$. The results indicated that there might be a detectable recessive major gene controlling the height growth of Loblolly pine in this diallel population, although the effect of the major gene is small compared with the polygenic effects, explaining about 17% of total phenotypic variance. High estimated GCA values also indicated that the polygenic component was more important for height growth of Loblolly pine at age 6 years. Given the limitations of the experimental design and relatively small effect of the major gene, the major genotypes may not be accurate in this case study. Furthermore, the validity of the model assumptions and possible interaction of polygenic and major gene effects may make this genotype interpretation difficult.

7. Discussion

The Bayesian approach with parent-blocking Gibbs sampling has been shown to be effective in this study for analysing data from a half-diallel mating design using a mixed-inheritance model. The method can be used successfully to detect major gene segregation, estimate major gene effects and putative genotypes of a major gene for parents and progenies, as well as polygenic parameters of a quantitative trait. To our knowledge, this is the first statistical approach that incorporates the polygenic effects of GCA and SCA with a major gene in the MIM for a diallel mating design. The results from this model have provided a better understanding of mixed inheritance of quantitative traits in diallel populations, particularly for tree breeding.

Although major-gene genotypes detected are putative, based on the statistical inference, this information of segregation could be valuable for identifying parents with major genes affecting quantitative traits. The proposed method is based on the existing half-diallel mating design, and hence it can be used to analyse actual progeny test data for breeding

https://doi.org/10.1017/S0016672304006718 Published online by Cambridge University Press

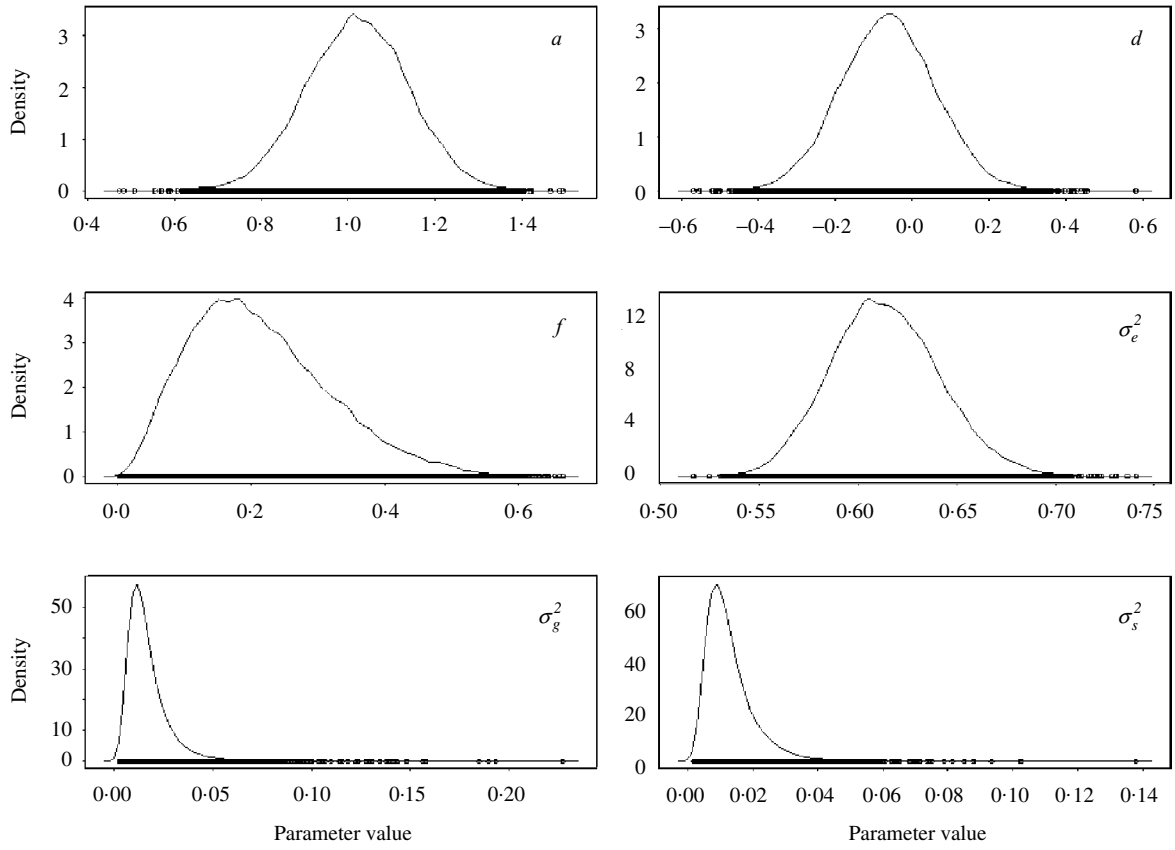

Fig. 2. Posterior densities of six genetic parameters (a, d, f, σ_e^2, σ_g^2, σ_s^2) of the initial value set N1 for normal prior for a, d and μ with singular design matrix were used for the mixed-inheritance model. The horizontal dots are the Gibbs sample values for the parameter.

purposes. By systematically screening progeny test data with this method, putative major genes, genotypes of parents and progenies, and their probabilities can be estimated. This is in addition to the polygenic effects of GCA and SCA, and other variance component estimates from the traditional analysis. The detectable major gene and putative genotypes would be valuable for selecting materials in an active breeding programme. By combining the GCA and SCA estimates and possible major genotypes, suitable combinations of parents or progenies can be chosen to provide maximum genetic gains for a breeding programme.

The putative genotypes of major genes identified with this method could also be valuable for molecular mapping experiments by providing a mapping population with a high probability of segregation for the quantitative traits. This should improve the effectiveness of searches for quantitative trait loci in the laboratory and reduce the experimental costs of such a search. Often no quantitative trait loci can be detected due to inadequate segregation in the experimental population. Our analytical approach can thus be first used to identify parents or families that are most likely segregating for a quantitative trait before further molecular mapping and linkage analysis are pursued. The detection of major genes using statistical approaches and confirmation of the presence of a major gene using genetic markers are very important for designing more effective breeding strategies and would make breeding for quantitative traits much more efficient.

One problem with the traditional maximum likelihood approach is that it is not feasible to obtain maximum likelihood estimators either by maximizing the likelihood of incomplete data directly or by using an iterative algorithm such as EM. By adopting a Bayesian framework for the segregation analysis, we avoided the necessity of performing such an optimization. In addition, estimates of the parameters are based on finite sample posterior distribution and thus avoid the use of asymptotic approximation using Fisher information. The marginal Bayesian estimators take into account the uncertainty of a single parameter that is due to the uncertainty in all other parameters in the model. Thus, it can take into account all model uncertainty based on finite samples. Model selection methods based on theoretical criteria such as AIC and BIC can be used to choose models with different numbers of major genes.

Usually, a full-rank design matrix makes the Markov chain converge quicker (Gilks & Roberts,

https://doi.org/10.1017/S0016672304006718 Published online by Cambridge University Press

Table 4. *Convergence diagnostics of the six parameters,* a, d, f, σ_e^2, σ_g^2, σ_s^2, *for testing the effects of prior distribution, initial value and design matrix on the mixed inheritance model*

Parameters:	*a*	*d*	*f*	σ_e^2	σ_g^2	σ_s^2
Uniform prior						
CS reduction factor[a]						
Est	1·00	1·00	1·00	1·00	1·00	1·00
0·975	1·01	1·01	1·00	1·00	1·00	1·00
Raftery dependent factor[b]						
N0	12·2 (1·7)[c]	3·0 (1·5)	1·0	1·1	1·1	1·0
N1	9·8 (2·0)	5·4 (1·7)	1·0	1·3	1·1	1·1
N2	10·4 (1·7)	4·9 (1·3)	1·0	1·3	1·0	1·2
N3	5·0 (1·7)	5·5 (1·5)	1·0	1·2	1·0	1·0
N4	8·2 (3·3)	6·3 (2·6)	1·0	1·2	1·2	1·0
N5	6·2 (1·5)	5·8 (1·1)	1·0	1·2	1·0	1·2
Normal prior						
CS reduction factor						
Est	1·00	1·00	1·00	1·00	1·00	1·00
0·975	1·01	1·01	1·00	1·00	1·00	1·00
Raftery dependent factor						
N0	13·0 (3·1)	8·7 (2·0)	1·1	1·3	1·0	1·2
N1	13·6 (2·2)	7·6 (1·5)	1·1	1·3	1·1	1·2
N2	19·9 (4·3)	7·1 (1·5)	1·0	1·0	1·2	1·0
N3	4·4 (1·3)	4·6 (1·5)	1·2	1·2	1·0	1·1
N4	4·9 (1·2)	8·2 (1·8)	1·1	1·1	1·1	1·2
N5	5·4 (1·3)	4·6 (1·5)	1·0	1·3	1·3	1·1
Normal prior with full-rank design matrices						
CS reduction factor						
Est	1·04	1·11	1·12	1·04	1·98	1·03
0·975	1·09	1·28	1·29	1·11	2·87	1·08
Raftery dependent factor						
N0	5·2 (1·5)	5·7 (1·3)	1·0	1·1	1·0	1·0
N1	6·4 (1·7)	4·9 (1·3)	1·0	2·0	1·2	1·2
N2	5·9 (1·5)	4·7 (1·3)	1·1	1·2	1·2	1·2
N3	9·1 (2·2)	3·2 (1·3)	1·0	1·3	2·6	1·0
N4	31·1 (5·9)	6·2 (1·3)	1·0	1·3	1·2	1·1
N5	8·0 (1·7)	7·4 (3·3)	1·1	1·2	1·2	1·2

In the MIM model, the polygene background set up was $h^2 = 0{\cdot}2$, $r = 0{\cdot}5$, and the major gene effect was simulated by $2a = 2{\cdot}0$, $d = 0{\cdot}0$ and $f = 0{\cdot}2$ (actual value is 0·167). There were six runs, N0 to N5, for each of three cases.

[a] CS reduction factor: corrected score reduction factor.

[b] Raftery dependent factor is calculated under quantile = 0·025, accuracy = ±0·05 and probability = 0·9.

[c] The dependent factor in parentheses was calculated by using lag = 30, instead of lag = 5 in the regular base because of strong autocorrelation.

Table 5. *Estimated means and standard deviations of posterior densities for seven genetic parameters* (a, d, f, σ_e^2 σ_g^2, σ_s^2, σ_m^2), *and general combining ability* (*GCA*) (g_1–g_6) *and major gene genotypes of six parents* (P_1–P_6) *for the diallel from Bayesian-based segregation analysis. There were two independent runs: chain 1 and chain 2*

Genotype:	P_1	P_2	P_3	P_4	P_5	P_6			
Chain 1	A_2A_2 (0·95)[a]	A_1A_2 (0·90)	A_2A_2 (0·86)	A_2A_2 (0·84)	A_2A_2 (0·97)	A_2A_2 (0·91)			
Chain 2	A_2A_2 (0·97)	A_1A_2 (0·86)	A_2A_2 (0·82)	A_2A_2 (0·80)	A_2A_2 (1·0)	A_2A_2 (0·93)			
GCA:	g_1	g_2	g_3	g_4	g_5	g_6			
Chain 1	1·30 ± 0·81[b]	−0·88 ± 0·64	0·03 ± 0·69	−0·45 ± 0·67	0·66 ± 0·67	−0·27 ± 0·75			
Chain 2	1·53 ± 0·92	−0·91 ± 0·62	−0·05 ± 0·71	−0·48 ± 0·68	0·84 ± 0·76	−0·26 ± 0·76			
Parameter:	*a*	*d*	*f*	σ_e^2	σ_g^2	σ_s^2	σ_m^2	σ_p^2	σ_m^2/σ_p^2
Chain 1	2·29 ± 0·65	−2·31 ± 1·05	0·33 ± 0·16	6·96 ± 0·25	0·96 ± 0·92	1·41 ± 0·90	2·18 ± 1·50	12·47	0·175
Chain 2	2·17 ± 0·65	−2·25 ± 0·78	0·37 ± 0·17	6·91 ± 0·27	1·12 ± 1·09	1·29 ± 092	2·16 ± 1·41	12·60	0·171

[a] Probability of this genotype.

[b] The estimated mean ± standard deviation from the Gibbs samples with lag = 5.

https://doi.org/10.1017/S0016672304006718 Published online by Cambridge University Press

1996). In this study, the combination of the method, data structure and full-rank setting may limit the movement of chains by chance. Consequently, the wrong parent genotypes may be identified even though the chain may mix well and converge for other genetic parameters. Smooth posterior density is not always an indicator of convergence as studied by Wang *et al*. (1994), especially when dealing with the discrete genotypes in the unknown parameter space. Gelfand & Sahu (1998) have shown that mixing improved as unidentified parameters were specified in an increasingly flat prior.

Efficiency of Gibbs sampling depends on the mixing property of the Markov chain, which in turn is determined by the parameterization used in the model and the sampling scheme applied. From the consistent results of multiple chains and convergence tests, we conclude that the chains have mixed well for parent block sampling of genotypes. However, if the size of the progeny population is small and/or the major gene effect is small, mixing may become a problem even with the parent block sampling. If additional molecular marker information is included in the model or the overall mean μ in the model is extended to a vector by including other non-genetic parameters, such as site effects and block within-site effects, the mixing problem may be worse. One possible way to avoid this is to use the hybrid Markov chain embedding a Hasting or Metropolis updating step in the basic Gibbs sampling scheme, as used in pedigree analysis (Tierney, 1994). Another way would be to use a Metropolis jumping kernel to make transition between communicating classes (Lin, 1995). A Bayesian network is also an alternative solution (Lund & Jensen, 1999).

Although only one major gene with two alleles was considered in this study, this method can be extended to more general situations by considering $2n_p$ alleles and/or two or more major genes. When multiple alleles and genes are involved in the model, many important issues such as Hardy–Weinberg disequilibrium (among alleles for one gene), linkage disequilibrium (association among genes) and epistasis (non-allelic interaction) should be examined. In these cases, model selection can be adopted by means of the Bayes Factor (Kass & Raftery, 1998), or by means of a predictive loss approach (Gelfand & Ghosh, 1998).

This research was supported by a grant from the Department of Energy and several industry members of the North Carolina State University-Industry Cooperative Tree Improvement Program. The support of the Department of Forestry at the North Carolina State University is gratefully acknowledged. We thank the North Carolina Supercomputer Center for allowing us to use its facilities for some computations. We appreciate the inputs and helpful discussions at various stages of this work of Drs Zhao-Bang Zeng, Trudy Mackay, Rongling Wu, Ben-Hui Liu and Arthur Johnson. We thank the editor and two anonymous reviewers for their useful comments on the manuscript.

References

Brooks, S. P. (1998). Markov chain Monte Carlo method and its application. *The Statistician* **47**, 69–100.

Brooks, S. P. & Roberts, G. O. (1998). Convergence assessment techniques for Markov chain Monte Carlo. *Statistics and Computing* **8**, 319–335.

Elston, R. C. & Stewart, J. (1971). A general model for the genetic analysis of pedigree data. *Human Heredity* **21**, 523–542.

Fain, P. R. (1978). Characteristics of simple sib-ship variance tests for the detection of major loci and application to height, weight and spatial performance. *Annals of Human Genetics* **42**, 109–120.

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

Gelfand, A. E. & Ghosh, S. K. (1998). Model choice: a minimum posterior predictive loss approach. *Biometrika* **85**, 1–11.

Gelfand, A. E. & Sahu, S. K. (1998). Identifiability, improper priors and Gibbs sampling for generalized linear models. *Journal of the American Statistical Association* **94**, 247–253.

Gelfand, A. E. & Smith, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**, 398–409.

Gelman, A. & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* **7**, 457–511.

Gilks, W. R. & Roberts, G. O. (1996). Strategies for improving MCMC. In *Markov Chain Monte Carlo in Practice* (ed. W. R. Gilks, S. Richardson & D. J. Spiegelhalter), pp. 89–114. London: Chapman & Hall.

Griffing, G. (1956). Concept of general and specific combining ability in relation to diallel crossing systems. *Australian Journal of Biological Science* **9**, 463–493.

Hallauer, A. R. & Miranda, J. B. (1981). *Quantitative Genetics in Maize Breeding*, p. 468. Ames, Iowa: Iowa State University Press.

Hoeschele, I. (1988). Genetic evaluation with data presenting evidence of mixed major gene and polygenic inheritance. *Theoretical and Applied Genetics* **76**, 81–92.

Huber, D. A., White, T. L. & Hodge, G. R. (1992). The efficiency of half-sib, half-diallel and circular mating designs in the estimation of genetic parameters in forestry: a simulation. *Forest Science* **38**, 757–776.

Janss, L. L. G., Van Arendonk, J. A. M. & Brascamp, E. W. (1997). Bayesian statistical analysis for presence of single gene affecting meat quality traits in a crossed pig population. *Genetics* **145**, 395–408.

Jiang, C. J., Pan, X. B. & Gu, M. H. (1994). The use of mixture models to detect effects of genes on quantitative characters in a plant breeding experiment. *Genetics* **136**, 383–394.

Karlin, S. & Williams, P. T. (1981). Structured exploratory data analysis (SEDA) for determining mode of inheritance of quantitative traits. II. Simulation studies on the effect of ascertaining families through high-valued probands. *American Journal of Human Genetics* **33**, 282–292.

Kass, R. E. & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association* **90**, 773–795.

Kaya, Z., Sewell, M. M. & Neale, D. B. (1999). Identification of quantitative trait loci influencing annual height- and diameter increment growth in loblolly pine (*Pinus taeda* L.). *Theoretical and Applied Genetics* **98**, 586–592.

Kinghorn, B. P., Van Arendonk, J. & Hetzel, J. (1994). Detection and use of major genes in animal breeding. *AgBiotech News and Information* **6**, 297–302.

Knott, S. A., Haley, C. S. & Thompson, R. (1991). Methods of segregation analysis for animal breeding data: a comparison of power. *Heredity* **68**, 299–311.

Le Roy, P., Elsen, J. M. & Knott, S. A. (1989). Comparison of four statistical methods for detection of a major gene in a progeny test design. *Genetics, Selection, Evolution* **21**, 341–357.

Li, B. & Wu, R. (1996). Genetic causes of heterosis in juvenile aspen: a quantitative comparison across intra-and interspecific hybrids. *Theoretical and Applied Genetics* **93**, 380–391.

Li, B., McKeand, S. E. & Weir, R. J. (1999). Tree improvement and sustainable forestry – impact of two cycles of loblolly pine breeding in the USA. *Forest Genetics* **6**, 229–234.

Lin, S. (1995). A scheme for constructing an irreducible Markov Chain for pedigree data. *Biometrics* **51**, 318–322.

Liu, J. S., Wong, W. H. & Kong, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika* **81**, 27–40.

Long, A. D., Mullaney, S. L., Reid, L. A., Fry, J. D., Langley, C. H. & Mackay, T. F. C. (1995). High resolution mapping of genetic factors affecting abdominal bristle number in *Drosophila melanogaster*. *Genetics* **139**, 1273–1291.

Lund, M. S. & Jensen, C. S. (1999). Blocking Gibbs sampling in the mixed inheritance model using graph theory. *Genetics, Selection, Evolution* **31**, 3–24.

Lynch, M. & Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits*, pp. 321–378. Massachusetts: Sinauer Associates.

McKeand, S. E., Eriksson, G. & Roberds, J. H. (1997). Genotype by environment interaction for index traits that combine growth and wood density in Loblolly pine. *Theoretical and Applied Genetics* **94**, 1015–1022.

Mérat, P. (1968). Distributions de fréquences, interprétration du déterminisme génétique des caractères quantitatifs et recherche de 'géens majeurs'. *Biometrics* **24**, 277–293.

Piper, L. R. & Shrimpton, A. E. (1989). The quantitative effects of genes which influence metrics traits. In *Evolution and Animal Breeding. Reviews on Molecular and Quantitative Approaches in Honour of Alan Robertson* (ed. W. G. Hill & T. F. C. Mackay), pp. 147–151. Wallingford: CBA International.

Remington, D. L. & O'Malley, D. M. (2000). Evaluation of major genetic loci contributing to inbreeding depression for survival and early growth in a selfed family of *Pinus taeda*. *Evolution* **54**, 1580–1589.

Roberts, G. O. & Sahu, S. K. (1997). Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *Journal of the Royal Statistical Society Series B* **59**, 291–317.

Smith, B. (2000). Bayesian output analysis program (BOA). Version 0.5.0 Programmer Manual.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Annual Statistics* **22**, 1701–1762.

Wang, C. S., Rutledge, J. J. & Gianola, D. (1994). Bayesian analysis of mixed linear models via Gibbs sampling with an application to litter size in Iberian pigs. *Genetics, Selection, Evolution* **26**, 91–115.

Wang, J., Podlich, D. W., Cooper, M. & Delacy, I. H. (2001). Power of the joint segregation analysis method for testing mixed major-gene and polygene inheritance models of quantitative traits. *Theoretical and Applied Genetics* **103**, 804–816.

Wilcox, P. L., Amerson, H. V., Kulman, E. G., Liu, B. H., O'Malley, D. M. & Sederoff, R. R. (1996). Detection of a major gene for resistance to fusiform rust disease in loblolly pine by genomic mapping. *Proceedings of the National Academy of Sciences of the USA* **93**, 3859–3864.

Wu, R. & Li, B. (1999). A multiplicative epistatic model for analyzing interspecific differences in outcrossing species. *Biometrics* **55**, 355–365.

Wu, R. & Li, B. (2000). A quantitative genetic model for analyzing species differences in outcrossing species. *Biometrics* **56**, 1098–1104.

Wu, R., Li, B., Wu, S. S. & Casella, G. (2001). A maximum likelihood-based method for mining major genes affecting a quantitative character. *Biometrics* **57**, 764–768.

Xiang, B. & Li, B. (2001). A new mixed analytical method for genetic analysis of diallel data. *Canadian Journal of Forest Research* **31**, 1–8.

Zeng, W. (2000). Statistical methods for detecting major genes of quantitative traits using phenotypic data of a diallel mating. PhD dissertation, North Carolina State University, Raleigh, USA.

Zeng, W. & Li, B. (2003). Simple tests for detecting segregation of major genes with phenotypic data from a diallel mating. *Forest Science* **49**, 268–278.

Zobel, B. J. & Talbert, J. T. (1984). *Applied Forest Tree Improvement*. New York: Wiley.