Genet. Res., Camb. (2000), **76**, *pp.* 187–198. Printed in the United Kingdom

Estimation of additive, dominance and epistatic variance components using finite locus models implemented with a single-site Gibbs and a descent graph sampler

F.-X. DU AND I. HOESCHELE*

Departments of Dairy Science and Statistics, Virginia Tech, Blacksburg, VA 24061-0315, USA

(*Received 10 September 1999 and in revised form 20 January 2000*)

Summary

In a previous contribution, we implemented a finite locus model (FLM) for estimating additive and dominance genetic variances via a Bayesian method and a single-site Gibbs sampler. We observed a dependency of dominance variance estimates on locus number in the analysis FLM. Here, we extended the FLM to include two-locus epistasis, and implemented the analysis with two genotype samplers (Gibbs and descent graph) and three different priors for genetic effects (uniform and variable across loci, uniform and constant across loci, and normal). Phenotypic data were simulated for two pedigrees with 6300 and 12300 individuals in closed populations, using several different, non-additive genetic models. Replications of these data were analysed with FLMs differing in the number of loci. Simulation results indicate that the dependency of non-additive genetic variance estimates on locus number persisted in all implementation strategies we investigated. However, this dependency was considerably diminished with normal priors for genetic effects as compared with uniform priors (constant or variable across loci). Descent graph sampling of genotypes modestly improved variance components estimation compared with Gibbs sampling. Moreover, a larger pedigree produced considerably better variance components estimation, suggesting this dependency might originate from data insufficiency. As the FLM represents an appealing alternative to the infinitesimal model for genetic parameter estimation and for inclusion of polygenic background variation in QTL mapping analyses, further improvements are warranted and might be achieved via improvement of the sampler or treatment of the number of loci as an unknown.

1. Introduction

Epistasis is the effect of interaction of genes at two or more loci on phenotypes. Epistatic variation can be partitioned into components arising from additive by additive ($A \times A$), additive by dominance ($A \times D$), dominance by dominance ($D \times D$), and interactions among more than two loci (e.g. Falconer & Mackay, 1996). For prediction of individual additive, dominance and epistatic effects with mixed linear model methodologies, and for estimation of the corresponding variance components (e.g. Tempelman & Burnside, 1990; Hoeschele, 1991; VanRaden *et al*., 1992; Fuerst & Soelkner, 1994), inverses of the different genetic relationship matrices are often needed. Although rapid inversion methods are available for some components (A, D, $A \times A$), (Henderson, 1976; Hoeschele & VanRaden, 1991; VanRaden & Hoeschele, 1991), the inversion of relationship matrix for other components of epistasis in the context of large, complex pedigrees is still difficult. Inbreeding introduces further complications in the covariance structure of a population (de Boer & Hoeschele, 1993), when interaction components of the types D, $A \times D$, $D \times D$, etc., are included.

Due to computational complexity and inaccurate estimation, non-additive genetic variation has been ignored in genetic evaluation systems and in many breeding programmes (Fuerst *et al*., 1997). However, estimating non-additive genetic variance components becomes increasingly important for the following

* Corresponding author. Dr Ina Hoeschele, 2160 Litton Reaves Hall, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061-0315, USA. Tel: +1 (540) 231 4760. Fax: +1 (540) 231 5014. e-mail: inah@vt.edu

reasons. First, the average kinship coefficient among individuals has increased considerably in some animal populations such as dairy cattle breeds so that the covariance structure of a population cannot be adequately described via the additive genetic and residual variance components. Ignoring covariances due to non-additive genetic factors results in biased estimation of additive genetic variance and inaccurate prediction of additive genetic effects. Secondly, several reasonably high estimates of epistatic variances have been reported in cattle (Allaire & Henderson, 1965; Fuerst & Soelkner, 1994) and poultry (Abplanalp, 1988), although the importance of epistasis in explaining quantitative genetic variation is still unknown. Furthermore, accurate estimation of non-additive genetic variances could help to evaluate potential benefits from utilization of specific combining ability in breeding programmes (DeStefano & Hoeschele, 1992; Fuerst *et al*., 1998).

A continuously intensifying effort is being made to map individual quantitative trait loci (QTL) using genetic markers (e.g. Lander & Botstein, 1989; Georges *et al*., 1995; Hoeschele *et al*., 1997; Zhang *et al*., 1998). As multiple QTL mapping methods and denser marker data become available, it becomes increasingly important to explore epistatic interactions among QTL not only in line crosses (Kao & Zeng, 1999) but also in segregating populations such as human (Mitchell *et al*., 1997) and livestock. Analyses may include interaction effects not only between QTL but also between QTL and polygenic background, as well as non-additive polygenic variation. Methods developed in this paper may make such analyses feasible.

Virtually all estimates of additive and non-additive genetic variance components found in the literature were obtained under the infinitesimal model with the restricted maximum likelihood method (e.g. Hoeschele, 1991; VanRaden *et al*., 1992; Fuerst & Soelkner, 1994) or with simpler, non-iterative methods (e.g. Allaire & Henderson, 1965; Hay *et al*., 1983). Finite polygenic models (FPMs) were first proposed by Thompson & Skolnick (1977) for estimating the heritability of longevity in complex human pedigrees and subsequently by Fernando *et al*. (1994) for complex segregation analysis, as they lead to more efficient computation. These early FPMs assumed biallelic loci, additive gene action, constancy of additive effects and allele frequencies across loci, and fitted polygenic number rather than genotypes at individual loci. Recently, finite locus models (FLMs) have been explored as an alternative way of estimating genetic variance components (Du *et al*., 1999; Pong-Wong *et al*., 1998, 1999). These FLMs still assume biallelic loci but fit genotypes at individual loci, and allow for non-additive gene action and variable gene effects across loci. Allele frequencies are still held constant at 0·5, but could be estimated in the FLMs.

For FPM analysis of data simulated under mixed models of inheritance that contained additive polygenic loci along with a segregating major QTL, Fernando *et al*. (1994) and Stricker *et al*. (1995) reported accurate estimates of major gene effects and narrow-sense polygenic heritability. Pong-Wong *et al*. (1999) estimated variance components under a purely additive model with FLM analysed by a Bayesian method in a small pedigree (480 members). They found that the estimates of additive genetic variance were dependent on the number of loci in the analysis FLM with independent uniform distributions as priors for additive genetic effects across loci. They also reported that this dependency was greatly diminished with exponential priors and eliminated with normal priors. In a Bayesian implemented FLM analysis of a much larger pedigree with 6300 members, Du *et al*. (1999) found that variance components estimation for a purely additive model was independent of the number of loci in the analysis FLM, with bounded uniform priors for genetic effects.

Pong-Wong *et al*. (1998), Goddard (1998) and Du *et al*. (1999) extended the additive FLM to include dominance effects. Both Pong-Wong *et al*. (1998) and Du *et al*. (1999) found that the variance components estimation in the presence of dominance was dependent on the number of loci in the analysis FLM. In both studies, the FLM was implemented with a Gibbs sampler using single-site updating except for sires and their final offspring (Pong-Wong *et al*., 1998) or parents and their final offspring (Du *et al*., 1999).

Dependence of variance estimates on the number of loci in the analysis FLM could be caused by distribution choices of genetic effects across loci, by poor mixing of the Gibbs sampler for genotypes, and (or) by data insufficiency. In this contribution, we therefore explored an alternative genotype sampling scheme, based on descent graphs (Thompson, 1994; Sobel & Lange, 1996; Tier & Henshall, 1999), to sample the genotypes at all loci jointly, and different distributions for genetic effects across loci. We also extended the FLM to include two-locus epistasis, and the analyses were performed for two pedigrees of different sizes with phenotypes simulated under various non-additive genetic models.

2. Methodology

(i) *Finite locus model including two-locus epistasis*

A FLM with additive, dominance and all two-locus epistatic effects, conditional on a set of genotypes for the pedigree and all loci ($\boldsymbol{G}$), can be written as

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z}_{a(G)}\boldsymbol{a} + \boldsymbol{Z}_{d(G)}\boldsymbol{d} + \boldsymbol{Z}_{aa(G)}\boldsymbol{aa} + \boldsymbol{Z}_{ad(G)}\boldsymbol{ad} + \boldsymbol{Z}_{da(G)}\boldsymbol{da} + \boldsymbol{Z}_{dd(G)}\boldsymbol{dd} + \boldsymbol{e}, \quad (1)$$

where $\boldsymbol{y}$ is a vector of phenotypes; $\boldsymbol{X}$ is a known design-covariate matrix relating observations in $\boldsymbol{y}$ to the vector of non-genetic classification and regression effects ($\boldsymbol{\beta}$); $\boldsymbol{a}$ is a vector of homozygote differences at k biallelic loci; $\boldsymbol{d}$ is a vector of dominance deviations at k biallelic loci; $\boldsymbol{aa}$ ($\boldsymbol{ad}$, $\boldsymbol{da}$, $\boldsymbol{dd}$) is a vector of A × A (A × D, D × A, D × D) epistatic deviations at s locus pairs; $\boldsymbol{Z}_{a(G)}$ is a design matrix with k columns containing coefficients of −1, 0 or 1 corresponding to the three genotypes at a biallelic locus; $\boldsymbol{Z}_{d(G)}$ is a design matrix with k columns containing coefficients of 0·5 for heterozygous genotypes and −0·5 for homozygous genotypes; $\boldsymbol{Z}_{aa(G)}$, $\boldsymbol{Z}_{ad(G)}$, $\boldsymbol{Z}_{da(G)}$ and $\boldsymbol{Z}_{dd(G)}$ are design matrices with s columns and can be constructed by multiplication of appropriate elements in $\boldsymbol{Z}_{a(G)}$ and $\boldsymbol{Z}_{d(G)}$; s is the number of two-locus interactions; and $\boldsymbol{e}$ is a vector of residuals. The coefficients in the $\boldsymbol{Z}$ matrices are based on the orthogonal model of Cockerham (1954) and are identical to those for an F2 population (Kao & Zeng, 1999), as the allele frequency at each biallelic locus was fixed at 0·5. Model (1) represents the most general FLM; more restricted models can be obtained by omitting $\boldsymbol{ad}$, $\boldsymbol{da}$ and $\boldsymbol{dd}$ effects, by also omitting $\boldsymbol{aa}$ effects, and by forcing all elements in $\boldsymbol{a}$, $\boldsymbol{d}$ and in the interaction effects vectors to be equal across loci.

To accommodate epistasis, unlinked loci in the analysis FLM were assigned to groups of triplets of loci. While there were no epistatic interactions between the loci in different groups, each locus interacted with the other two loci in the same group. Then, in model (1), $k = s$ holds for any integer number k which is divisible by 3, as there are three two-locus interactions per triplet of loci. Letting a locus interact with more than just one other locus increases the number of epistatic effects, so that fewer loci need to be included in the analysis FLM.

(ii) *Bayesian analysis of the finite locus model*

Model (1) was analysed by a Bayesian method. The joint posterior probability density of all unknowns can be written as

$$\begin{aligned}
&\mathrm{f}(\boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{d}, \boldsymbol{aa}, \boldsymbol{ad}, \boldsymbol{da}, \boldsymbol{dd}, \lambda_a, \lambda_d, \lambda_{aa}, \lambda_{ad}, \lambda_{da}, \lambda_{dd}, \sigma_e^2, \boldsymbol{G} \mid \boldsymbol{y}) \\
&\quad \propto \prod_{i=1}^{n} \mathrm{f}(y_i \mid \boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{d}, \boldsymbol{aa}, \boldsymbol{ad}, \boldsymbol{da}, \boldsymbol{dd}, \sigma_e^2, \boldsymbol{G}_i \prod_{i=1}^{n_b} \prod_{j=1}^{k} \mathrm{f}(g_{ij}) \\
&\qquad \times \prod_{i=n_b+1}^{n} \prod_{j=1}^{k} f(g_{ij} \mid g_{f_{ij}}, g_{m_{ij}}) \\
&\qquad \times \mathrm{f}(\boldsymbol{\beta})\mathrm{f}(\boldsymbol{a})\mathrm{f}(\boldsymbol{d})\mathrm{f}(\boldsymbol{aa})\mathrm{f}(\boldsymbol{ad})\mathrm{f}(\boldsymbol{da})\mathrm{f}(\boldsymbol{dd})\mathrm{f}(\sigma_e^2)\mathrm{f}(\lambda_a) \\
&\qquad \times \mathrm{f}(\lambda_d)\mathrm{f}(\lambda_{aa})\,\mathrm{f}(\lambda_{ad})\mathrm{f}(\lambda_{da})\mathrm{f}(\lambda_{dd}), \qquad (2)
\end{aligned}$$

where $\boldsymbol{G}_i$ is row i of $\boldsymbol{G}$, $\boldsymbol{G}_i = \{g_{ij}\}$ with g_{ij} being the genotype of individual i at locus j; n_b is the number of base (founder) individuals in the population; n is the total number of individuals; λ_a, λ_d, λ_{aa}, λ_{ad}, λ_{da} and λ_{dd} are hyperparameters of prior distributions for genetic effects $\boldsymbol{a}$, $\boldsymbol{d}$, $\boldsymbol{aa}$, $\boldsymbol{ad}$, $\boldsymbol{da}$ and $\boldsymbol{dd}$, respectively; $\mathrm{f}(y_i|.)$ is the penetrance function evaluated as a normal density; $\mathrm{f}(g_{ij})$ and $\mathrm{f}(g_{ij} \mid g_{f_{ij}}, g_{m_{ij}})$ are Hardy–Weinberg frequency and transition probability of genotype g_{ij}, respectively; f_i is father of i and m_i is mother; and $\mathrm{f}(\beta)$, $\mathrm{f}(\boldsymbol{a})$, $\mathrm{f}(\boldsymbol{d})$, $\mathrm{f}(\boldsymbol{aa}), \ldots,$ and $\mathrm{f}(\lambda_{\mathrm{dd}})$ are priors for the respective parameters. Inferences based on (2) were obtained using different Markov Chain Monte Carlo (MCMC) sampling schemes. For all the analyses presented below, model (1) and the corresponding joint posterior in (2) were restricted to contain only the $\boldsymbol{aa}$-component of the two-locus epistasis.

(iii) *Sampling location parameters*

Bounded uniform distributions were used as priors for location parameters $\boldsymbol{\beta}$ and σ_e^2 in all cases. A fixed effect was sampled from its fully conditional univariate normal, and error variance was sampled from an inverse chi-square distribution as in Wang *et al.* (1993). For genetic effects across loci, three priors were investigated: bounded uniform and variable across loci, bounded uniform and constant across loci, and normal and variable across loci. We assume that A × A is the only epistatic component in the following text.

(a) *Uniform and variable across loci.* With bounded uniform distributions as priors for genetic effects, the full conditional distribution of a_i is given as:

$$\begin{aligned}
&\mathrm{f}(a_i \mid \boldsymbol{a}_{-i}, \boldsymbol{d}, \boldsymbol{aa}, \boldsymbol{\beta}, \boldsymbol{G}, \sigma_e^2, \boldsymbol{y}) \\
&\quad \propto \exp\left\{ -\frac{1}{2\sigma_e^2}\left(\sum_{j=1}^{n} z_{aji}^2 \right)\left[a_i - \left(\sum_{j=1}^{n} z_{aji}^2 \right)^{-1} \right.\right. \\
&\qquad \times \left(\sum_{j=1}^{n} z_{aji}\left[y_i - \mu - \sum_{h=1}^{k} (z_{djh}\, d_h \right.\right. \\
&\qquad \left.\left. + z_{aajh}\, aa_h) - \sum_{h=1, h\neq i}^{k} z_{ajh}\, a_h \right]\Bigg)\Bigg]^2\Bigg\}, \qquad (3)
\end{aligned}$$

where a_i, d_i and aa_i, are the ith elements of vectors $\boldsymbol{a}$, $\boldsymbol{d}$ and $\boldsymbol{aa}$, respectively; $\boldsymbol{a}_{-i}$ is the vector $\boldsymbol{a}$ excluding a_i ($\boldsymbol{d}_{-i}$ and $\boldsymbol{aa}_{-i}$ used later in the text were defined similarly); μ (population mean) was assumed to be the only element of $\boldsymbol{\beta}$; z_{aij}, z_{dij} and z_{aaij} are the elements at the ith row and the jth column of matrix $\boldsymbol{Z}_{a(G)}$, $\boldsymbol{Z}_{d(G)}$ and $\boldsymbol{Z}_{aa(G)}$, respectively. Similarly, the full conditional distributions for d_i and aa_i are given as

$$\begin{aligned}
&\mathrm{f}(d_i \mid \boldsymbol{a}, \boldsymbol{d}_{-i}, \boldsymbol{aa}, \boldsymbol{\beta}, \boldsymbol{G}, \sigma_e^2, \boldsymbol{y}) \\
&\quad \propto \exp\left\{ -\frac{1}{2\sigma_e^2}\left(\sum_{j=1}^{n} z_{dji}^2 \right)\left[d_i - \left(\sum_{j=1}^{n} z_{dji}^2 \right)^{-1} \right.\right. \\
&\qquad \times \left(\sum_{j=1}^{n} z_{dji}\left[y_j - \mu - \sum_{h=1}^{k} (z_{ajh}\, a_h \right.\right. \\
&\qquad \left.\left. + z_{aajh}\, aa_h) - \sum_{h=1, h\neq i}^{k} z_{djh}\, d_h \right]\Bigg)\Bigg]^2\Bigg\} \qquad (4)
\end{aligned}$$

and

$$f(aa_i \mid \boldsymbol{a}, \boldsymbol{d}, \boldsymbol{aa}_{-i}, \boldsymbol{\beta}, \boldsymbol{G}, \sigma_e^2, \boldsymbol{y}) \propto \exp\left\{-\frac{1}{2\sigma_e^2}\left(\sum_{j=1}^{n} z_{aaji}^2\right)\left[aa_i - \left(\sum_{j=1}^{n} z_{aaji}^2\right)^{-1} \times \left(\sum_{j=1}^{n} z_{aaji}\left[y_j - \mu - \sum_{h=1}^{k}(z_{ajh}\, a_h + z_{djh}\, d_h) - \sum_{h=1, h\neq i}^{k} z_{aajh}\, aa_h\right]\right)\right]^2\right\}. \quad (5)$$

(b) *Uniform and constant across loci*. In this case, one or several of genetic effect vectors ($\boldsymbol{a}$, $\boldsymbol{d}$ and $\boldsymbol{aa}$) contains only one element. The full conditional distributions for a_l, d_l and aa_l are given by

$$f(a_1 \mid \boldsymbol{d}, \boldsymbol{aa}, \beta, \boldsymbol{G}, \sigma_e^2, \boldsymbol{y}) \propto \exp\left\{-\frac{1}{2\sigma_e^2}\left(\sum_{j=1}^{n}\left(\sum_{i=1}^{k} z_{aji}\right)^2\right)\left[a_1 - \left(\sum_{j=1}^{n}\left(\sum_{i=1}^{k} z_{aji}\right)^2\right)^{-1}\left(\sum_{j=1}^{n}\left(\sum_{i=1}^{k} z_{aji}\right)\left[y_j - \mu - \sum_{h=1}^{k}(z_{djh}\, d_h + z_{aajh}\, aa_h)\right]\right)\right]^2\right\}, \quad (6)$$

$$f(d_1 \mid \boldsymbol{a}, \boldsymbol{aa}, \beta, \boldsymbol{G}, \sigma_e^2, \boldsymbol{y}) \propto \exp\left\{-\frac{1}{2\sigma_e^2}\left(\sum_{j=1}^{n}\left(\sum_{i=1}^{k} z_{dji}\right)^2\right)\left[d_1 - \left(\sum_{j=1}^{n}\left(\sum_{i=1}^{k} z_{dji}\right)^2\right)^{-1}\left(\sum_{j=1}^{n}\left(\sum_{i=1}^{k} z_{dji}\right)\left[y_j - \mu - \sum_{h=1}^{k}(z_{ajh}\, a_h + z_{aajh}\, aa_h)\right]\right)\right]^2\right\}, \quad (7)$$

$$f(aa_1 \mid \boldsymbol{a}, \boldsymbol{d}, \beta, \boldsymbol{G}, \sigma_e^2, \boldsymbol{y}) \propto \exp\left\{-\frac{1}{2\sigma_e^2}\left(\sum_{j=1}^{n}\left(\sum_{i=1}^{k} z_{aaji}\right)^2\right)\left[aa_i - \left(\sum_{j=1}^{n}\left(\sum_{i=1}^{k} z_{aaji}\right)^2\right)^{-1}\left(\sum_{j=1}^{n}\left(\sum_{i=1}^{k} z_{aaji}\right)\left[y_j - \mu - \sum_{h=1}^{k}(z_{ajh}\, a_h + z_{djh}\, d_h)\right]\right)\right]^2\right\}. \quad (8)$$

(c) *Normal and variable across loci*. Fold-over normal for a_i, normal for d_i and aa_i were used as priors with probability density as:

$$f(\boldsymbol{a}) = \prod_{i=1}^{k}\frac{2}{\sqrt{2\pi\lambda_a}}\exp\left\{-\frac{a_i^2}{\lambda_a}\right\} \quad 0 \leqslant a_i < \infty, \quad (9)$$

$$f(\boldsymbol{d}) = \prod_{i=1}^{k}\frac{2}{\sqrt{2\pi\lambda_d}}\exp\left\{-\frac{d_i^2}{\lambda_d}\right\} \quad -\infty \leqslant d_i < \infty, \quad (10)$$

$$f(\boldsymbol{aa}) = \prod_{i=1}^{k}\frac{2}{\sqrt{2\pi\lambda_{aa}}}\exp\left\{-\frac{aa_i^2}{\lambda_{aa}}\right\} \quad -\infty \leqslant aa_i < \infty. \quad (11)$$

The full conditional distributions for a_i, d_i and aa_i are the product of equations (3) and (9), (4) and (10), and (5) and (11), respectively.

Bounded uniform distributions were used as priors for hyperparameters of priors for genetic effects. Consequently, the full conditional distributions of these hyperparameters are inverse chi-square distributions with probability density as:

$$f(\lambda_a \mid \boldsymbol{a}) \propto \lambda_a^{-(k/2)}\exp\left\{-\frac{1}{2\lambda_a}\sum_{i=1}^{k} a_i^2\right\}, \quad (12)$$

$$f(\lambda_d \mid \boldsymbol{d}) \propto \lambda_d^{-(k/2)}\exp\left\{-\frac{1}{2\lambda_d}\sum_{i=1}^{k} d_i^2\right\}, \quad (13)$$

$$f(\lambda_{aa} \mid \boldsymbol{aa}) \propto \lambda_{aa}^{-(k/2)}\exp\left\{-\frac{1}{2\lambda_{aa}}\sum_{i=1}^{k} aa_i^2\right\}. \quad (14)$$

Parameter estimates were marginal posterior means. Additive genetic variance (σ_a^2), dominance variance (σ_d^2), A × A, A × D, D × A and D × D epistatic variances ($\sigma_{aa}^2, \sigma_{ad}^2, \sigma_{da}^2, \sigma_{dd}^2$) were calculated as $0{\cdot}5\ \Sigma_{i=1}^{k} a_i^2$, $0{\cdot}25\ \Sigma_{i=1}^{k} d_i^2$, $0{\cdot}25\ \Sigma_{i=1}^{k} aa_i^2$, $0{\cdot}125\ \Sigma_{i=1}^{k} ad_i^2$, $0{\cdot}125\ \Sigma_{i=1}^{k} da_i^2$ and $0{\cdot}0625\ \Sigma_{i=1}^{k} dd_i^2$, respectively, where a_i, d_i, and aa_i, ad_i, da_i and dd_i are the ith element of vectors $\boldsymbol{a}$, $\boldsymbol{d}$, $\boldsymbol{aa}$, $\boldsymbol{ad}$, $\boldsymbol{da}$ and $\boldsymbol{dd}$, respectively.

(iv) *Markov chain Monte Carlo genotype sampling schemes*

Several different sampling schemes were investigated. The first scheme is identical to that in Du *et al*. (1999). It is a Gibbs sampler with single-site updating of genotypes and parameters, except that both male and female parents are sampled unconditionally on the genotypes of their final offspring. This scheme is a modification of that of Janss *et al*. (1995), who sampled only male parents unconditionally on their final offspring, and improved the mixing of genotypes in an earlier study (Du *et al*., 1999). In this scheme, the genotype of one individual at one locus is sampled conditionally on the genotypes of all other individuals, except for final offspring, at the same locus, and conditionally on the genotypes of all individuals at all other loci. Starting values for the genotypes were all heterozygous, with genotypes 12 and 21 sampled with equal probability.

The second genotype sampling scheme is based on a descent graph Markov chain, originally proposed by Thompson (1994), implemented by Sobel and Lange (1996), and recently implemented efficiently for major gene analysis (Tier & Henshall, 1999). Here we extend the descent graph sampler to perform updates to the genotypes at all unlinked loci in the FLM jointly. The sampling scheme consists of a number of Metropolis-Hastings (MH) steps of genotype sampling given parameters followed by a Gibbs update of each parameter conditional on the genotypes. For the

genotype sampling, a set of initial ordered genotypes of founders were generated by independently sampling alleles with frequencies of 0·5 (an ordered genotype is a genotype with known parental origins). An initial descent graph for all loci was obtained by sampling integers from the range of integers available on the computer, with each bit representing a segregation indicator for a given individual and locus (Tier & Henshall, 1999), and with each segregation indicator taking on a value of '0' for grandparental origin or '1' for grandmaternal origin. The set of ordered genotypes of founders and the descent graph determine the ordered genotypes for descendants. Subsequently, the ordered genotypes at all loci for all individuals were updated jointly as follows. At each MH step, a small number of changes are proposed: either a few randomly chosen founder alleles are resampled, or the states of a few segregation indicators are switched from grandmaternal (grandpaternal) to grandpaternal (grandmaternal). The proposed changes are then translated into genotypes, and the acceptance ratio for this proposal is evaluated. The new sample of genotypes is accepted or rejected according to the MH ratio

$$\alpha = \min\{1.0, R\},$$

where

$$R = \frac{\Pi_i \mathrm{f}(y_i \mid G_{il(\mathrm{n})}, \ldots, G_{ik(\mathrm{n})})}{\Pi_i \mathrm{f}(y_i \mid G_{il(\mathrm{c})}, \ldots, G_{ik(\mathrm{c})})} \tag{15}$$

and (n) and (c) denote new and current genotypes, respectively. The acceptance ratio simplifies to the ratio of the likelihoods of the phenotypes in (15), because each set of genotypes (n, c) has the same prior and proposal probability (due to working with ordered genotypes and an allele frequency of 0·5).

In our implementation, the descent graphs were stored using bit representation of $I = 2kN_{\mathrm{D}}/31$ (rounded up to the next integer value) integer numbers, with N_{D} representing the number of descendants. A starting descent graph was obtained by randomly sampling I integer values from $U[(-2^{31}), (2^{31}-1)]$. Bit storage greatly reduces storage requirements, and several built-in functions for bits manipulations facilitate choice of initial values and later updates, simplify programming, and increase the efficiency of genotyping sampling.

For implementation of the descent graph sampler, we found that several thousand (3000 to 10000) MH steps prior to updating parameters was optimum in terms of Monte Carlo error of parameter estimates for a given total number of MH steps. In each MH step, a very small number of changes were made either to the founder alleles or to the descent graph. More specifically, at each MH step an update of founder alleles, of the descent graph portion pertaining to non-final descendants, or of the descent graph portion pertaining to final descendants, was chosen with probabilities of 0·20, 0·20 and 0·60 (obtained by trial and error based on autocorrelations of parameter samples), respectively. When an update of founder alleles was chosen, a single founder allele was changed. When an update to the descent graph of non-final descendants was selected, two segregation indicators were changed. Finally, when an update to the descent graph of final descendants was chosen, 10–20 segregation indicators were changed. More segregation indicators were updated for finals than for non-finals, because changing a single segregation indicator of a non-final (or changing a single founder allele) on average implies changes to the genotypes of several individuals. In contrast, updating one segregation indicator of a final only potentially changes one genotype of this individual. With this scheme, the number of ordered genotypes changed in each MH step was reasonably constant across MH steps, and a fairly high acceptance rate was achieved (0·45–0·65). With increasing number of changes per MH step, acceptance rates quickly declined. A substantial number of different combinations of these updating parameters were tested, and other combinations did not improve the scheme described. To improve efficiency, those individuals whose genotypes could have changed due to the update of a founder allele or of segregation indicators, were flagged in each MH step. Polygenic values and likelihood contributions were recalculated only for those individuals. After performing several thousand MH steps for the genotypes of all individuals at all loci, parameters were resampled with Gibbs updates as before. The number of MH steps performed prior to a parameter update was determined empirically based on the autocorrelations of the parameters.

(v) *Simulated genetic models, population structures and analysis schemes*

Eight different genetic models were simulated. In all models, loci were unlinked, and each locus had two alleles with allele frequency of 0·5. Models 1 and 2 contained no epistasis, residual variance was 50, additive genetic variance was 50 and dominance variance was 25 (Table 1). In all other models (3–8), residual and total genetic variances were equal to 60. Additive genetic variance was 30 in models 3 to 6, and 20 in models 7 and 8 (see Table 1). No dominance was simulated in models 3 and 4, while dominance variances of 15 and 10 were simulated for models 5 and 6 and models 7 and 8, respectively. Epistatic variances of 30 and 15 were simulated for models 3, 4, 7 and 8, and for models 5 and 6, respectively.

Eighteen loci of diminishing effects and 20 loci of equal effects accounted for the genetic variation in models 1 and 2, respectively. The 18 loci in model

Table 1. *Genetic models used for data simulation*

Genetic model	σ^2_e	σ^2_a	σ^2_d	σ^2_{aa}	Loci[a]	Dominance
1	50	50	25	0	18 diminishing	Complete
2	50	50	25	0	20 constant	Complete
3	60	30	0	30	20 diminishing	No
4	60	30	0	30	40 constant	No
5	60	30	15	15	20 diminishing	Complete
6	60	30	15	15	40 constant	Complete
7	60	20	10	30	20 diminishing	Complete
8	60	20	10	30	40 constant	Complete

[a] All loci are unlinked and biallelic with allele frequency of 0·5. Number of loci is 18, 20 or 40. Constant implies equal additive, dominance and epistasis variance across loci. The 18 loci in model 1 includes one locus with additive genetic variance 25, two loci with variance 5, five loci with variance 2, and ten loci with variance 0·5. The 20 loci of diminishing contribution include: additive variance with one locus explaining 30%, one 10%, eight 5%, and ten 2%; epistasis variance with one pair (one locus with 30% and one with 10% additive variance) explaining 20%; 4 pairs of 5% additive variance explaining 15% of epistatic variance; 5 pairs of 2% additive variance explaining 4%.

Table 2. *Schemes for data simulation and analysis*[a]

Data/analysis scheme	Genetic model[b]	Pedigree size	Genotype sampling scheme	Priors for genetic effects[c]	Genetic effects (***a***, ***d*** and ***aa***) in analysis FLM
DA1	3	6300	Gibbs	Uniform	Variable
DA2	4	6300	Gibbs	Uniform	Variable
DA3	5	6300	Gibbs	Uniform	Variable
DA4	6	6300	Gibbs	Uniform	Variable
DA5	7	6300	Gibbs	Uniform	Variable
DA6	8	6300	Gibbs	Uniform	Variable
DA7	2	6300	Gibbs	Normal	Variable
DA8	1	6300	Gibbs	Normal	Variable
DA9	5	6300	Gibbs	Normal	Variable
DA10	2	12300	Gibbs	Uniform	Variable
DA11	2	12300	Gibbs	Normal	Variable
DA12	1	6300	Gibbs	Uniform	Constant
DA13	2	6300	Gibbs	Uniform	Constant
DA14	2	6300	Descent graph	Uniform	Variable
DA15	5	6300	Descent graph	Uniform	Variable

[a] Other analysis details include: analysis FLM contained same genetic components as corresponding genetic model: bounded uniform distributions were used as priors for parameters other than genetic effects, including mean, error variance, and hyperparameters for genetic effects; all location parameters and hyperparameters were sampled through a single site Gibbs sampler.
[b] Genetic models are described in Table 1.
[c] Uniform implies bounded uniform distribution; normal implies folded-over normal for additive effects, normal for dominance and (or) epistatic effects.

1 included one locus with additive genetic variance 25, two loci with variance 5, five loci with variance 2, and ten loci with variance 0·5. Forty loci contributed equally to additive genetic variance in models 4, 6 and 8, while 20 loci of diminishing effects accounted for the genetic variance in models 3, 5 and 7. The 20 loci of diminishing effect included one locus with 30%, one with 10%, eight with 5% and ten with 2% of additive genetic variance. In models 1 to 2, and 5 to 8, complete dominance was simulated for all loci. For two-locus epistasis, only A × A epistatic effects were simulated using Cockerham's orthogonal model (Cockerham, 1954). In the case of 40 equal loci, each of 20 randomly formed locus pairs contributed equally to epistatic variance. For 20 diminishing loci, nine pairs were formed from loci with equal additive genetic effect, and two loci with largest additive effects formed the tenth pair. The pair of largest loci, pairs of

loci with 5% of σ^2_a, and pairs of loci with 2% of σ^2_a explained 20%, 15% and 4% of the epistatic variance (σ^2_{aa}), respectively.

For all eight genetic models, a population structure of $n = 6300$ individuals over one base generation and three discrete offspring generations was simulated. To evaluate the effect of pedigree size variance components estimation, a pedigree with 12300 members over one base generation and six discrete offspring generations was simulated for genetic model 2. In both pedigrees, 50 males and 250 females were randomly selected at every generation, with each male randomly mated to 5 females. Females produced eight-offspring litters (four males and four females), giving each site 40 progeny.

Genotypes at a finite number of unlinked, biallelic loci were generated according to Hardy–Weinberg frequencies for founders and according to Mendelian transition probabilities for descendants. An individual's genotypic value was calculated by summing additive, dominance and epistatic effects that were determined by the individual's genotypes. Finally, residuals from a normal distribution with mean 0 and variance σ^2_e were added to obtain an individual's phenotype. No systematic environmental effects were simulated.

Factors for evaluation include three priors, two genotype sampling schemes, two pedigree structures and eight genetic models. Due to time constraints, only some combinations of these factors were evaluated, and data/analysis schemes that were evaluated are described in Table 2.

3. Results

The results presented below are means of MCMC realizations that were sampled every cycle after burn-in period. Starting values for location parameters and hyperparameters were arbitrarily chosen. While different lengths of MCMC were executed for different data/analysis schemes, the average Monte Carlo standard errors (Geyer, 1992; Sorensen *et al.*, 1995) were below 0·6% of the genetic variance (results not shown).

(i) *Gibbs sampling of genotypes, no dominance, 6300-pedigree, uniform priors*

Here we present results for the first genotype sampling scheme described earlier and for data simulated with A × A epistasis but without dominance (genetic models 3 and 4; data/analysis schemes DA1 and DA2). The analysis FLMs contained 6, 12 and 18 loci with variable A and A × A effects but no dominance effects; independent bounded uniforms were used as priors for genetic effects. Gibbs samplers of length 200000 for the 12- and 18-loci FLMs and 300000 cycles for the 6-loci FLM with 10000 burn-in cycles were applied to 10 replicates to obtain the results reported in Table 3, which are based on 10 replicates. Similar to the case with variable additive and dominance effects (Du *et al.*, 1999), estimates of the A × A variance increased, while residual variance decreased, as the number of loci in the analysis FLM increased, and the intermediate (12-loci) FLM produced the best estimates for all variance components.

(ii) *Gibbs sampling of genotypes, with dominance, 6300-pedigree, uniform priors*

Data simulated under genetic models 5 to 8 including additive, dominance and A × A epistatic components, were analysed with data/analysis schemes DA3 to DA6 with uniform priors for genetic effects (as described in Table 2). Results presented in Table 4 are averages of 10 replicates obtained from Gibbs samplers of length 300000 for the 12- and 18-loci FLM or 400000 cycles for the 6-loci FLM with 10000 burn-in cycles. For data simulated with models 5 and

Table 3. *Variance component estimates* (*and empirical standard errors*) *averaged across 10 replicates and obtained by analysing data simulated under models 3 and 4* (*see Table 1*) *with data/analysis schemes DA1 and DA2* (*see Table 2*)

Data/analysis scheme	Genetic model	No. of analysis loci	σ^2_e	σ^2_a	σ^2_{aa}
DA1	3	6	63·85 (0·85)	37·44 (1·54)	21·14 (1·43)
		12	58·09 (1·33)	33·46 (1·59)	31·28 (1·97)
		18	54·04 (1·35)	32·56 (1·49)	37·32 (2·01)
DA2	4	6	66·34 (0·77)	35·92 (1·01)	19·37 (1·05)
		12	60·28 (0·71)	33·27 (0·76)	29·08 (0·83)
		18	56·35 (0·63)	33·05 (0·64)	34·68 (0·44)

True values for genetic models 3 and 4 are error variance (σ^2_e) = 60, additive variance (σ^2_a) = 30, and epistasis variance (σ^2_{aa}) = 30.

Table 4. *Variance component estimates* (*and empirical standard errors*) *averaged across 10 replicates and obtained by analysing data simulated under models 5 to 8* (*see Table 1*) *with data/analysis schemes DA3 to DA6* (*see Table 2*)

Data/ analysis scheme	Genetic model	No. of analysis loci	σ^2_e	σ^2_a	σ^2_d	σ^2_{aa}
DA3	5	6	59·51 (1·13)	31·52 (0·98)	16·53 (0·97)	13·13 (0·67)
		12	52·73 (1·02)	29·60 (0·76)	20·61 (0·86)	19·50 (0·60)
DA4	6	6	61·17 (1·46)	32·10 (1·23)	14·88 (1·04)	13·46 (1·18)
		12	52·46 (1·60)	30·33 (1·07)	19·33 (1·06)	21·13 (0·99)
DA5	7	6	60·55 (1·65)	23·63 (1·60)	15·03 (1·11)	22·29 (1·60)
		12	52·62 (1·82)	23·17 (1·37)	19·03 (1·03)	28·59 (1·38)
DA6	8	6	64·24 (1·23)	25·91 (1·24)	13·97 (0·33)	18·37 (1·56)
		12	55·39 (1·36)	24·73 (1·32)	18·72 (0·62)	25·11 (1·33)

True values for genetic models 5 and 6 are error variance (σ^2_e) = 60, additive variance (σ^2_a) = 30, dominance variance (σ^2_d) = 15 and epistasis variance (σ^2_{aa}) = 15. True values for models 7 and 8 are $\sigma^2_e = 60$, $\sigma^2_a = 20$, $\sigma^2_d = 10$ and $\sigma^2_{aa} = 30$.

Table 5. *Variance component estimates* (*and empirical standard errors*) *averaged across five replicates and obtained by analysing data simulated under models 1, 2 and 5* (*see Table 1*) *with data/analysis schemes DA7 to DA9* (*see Table 2*). *Values in square brackets are from Du* et al. (*1999*) *and are given for comparison*

Data/ analysis scheme	Genetic model	No. of analysis loci	σ^2_e	σ^2_a	σ^2_d	σ^2_{aa}
DA7	2	6	54·20 (1·08)	48·23 (1·20)	22·38 (0·86)	
		[5]	[54·4 (0·92)]	[51·0 (0·94)]	[20·9 (1·11)]	
		12	50·21 (1·24)	48·64 (1·16)	26·15 (1·05)	
		[10]	[48·6 (1·05)]	[51·9 (1·02)]	[26·7 (1·21)]	
		18	48·79 (1·48)	48·76 (1·28)	27·56 (1·11)	
		[20]	[42·8 (2·18)]	[53·8 (0·93)]	[32·9 (2·75)]	
DA8	1	6	50·62 (1·46)	53·04 (2·46)	24·13 (1·57)	
		[5]	[51·2 (0·91)]	[51·9 (1·34)]	[23·8 (0·97)]	
		12	47·35 (1·53)	53·50 (2·59)	27·39 (1·55)	
		[10]	[46·8 (0·96)]	[52·7 (1·36)]	[28·3 (1·03)]	
		18	45·72 (1·35)	53·29 (2·61)	29·07 (1·41)	
		[20]	[39·5 (1·83)]	[56·0 (2·70)]	[35·6 (1·69)]	
DA9	5	6	61·79 (1·26)	32·86 (1·06)	15·09 (1·07)	10·02 (1·08)
		12	58·78 (1·45)	31·76 (1·25)	16·82 (1·19)	12·49 (1·40)
		18	57·41 (1·45)	31·11 (1·46)	17·21 (1·34)	14·07 (1·88)

True values for genetic models 1 and 2 are error variance (σ^2_e) = 50, additive variance (σ^2_a) = 50, dominance variance (σ^2_d) = 25 and epistasis variance (σ^2_{aa}) = 0. True values for model 5 are $\sigma^2_e = 60$, $\sigma^2_a = 30$, $\sigma^2_d = 15$ and $\sigma^2_{aa} = 15$.

6 (DA3 and DA4), dominance and epistatic variance estimates increased more rapidly, compared with the dominance-only or the epistasis-only case, while residual variance was decreased, as the number of loci in the FLM increased. For data simulated with models 7 and 8 where A × A epistatic variance was large (DA5 and DA6), overestimation of dominance variance was more pronounced, and epistatic variance more severely underestimated than before.

(iii) *Gibbs sampling of genotypes, with dominance, 6300-pedigree, normal priors*

In contrast to data/analysis schemes DA1 to DA6 in which bounded uniform priors were used for genetic effects, normal priors (folded-over normal for a_i, normal for d_i and aa_i) were used in DA7 to DA9. Bayesian analysis of FLMs was implemented by Gibbs samplers of length 120000 with 10000 burn-in cycles. Compared with the case of uniform priors, considerably shorter MCMC was needed using normal priors for genetic effects.

Results from Du *et al.* (1999), in which same data were analysed with uniform priors for genetic effects, were included in Table 5 for comparison. As shown in Table 5 (averages of five replicates), normal priors for genetic effects clearly improved variance components estimation, as compared with uniform priors. However, the dependency of variance components estimation on locus number in analysis FLM persisted

Table 6. *Variance component estimates* (*and empirical standard errors*) *averaged across five replicates and obtained by analyzing data simulated under model 2* (*see Table 1*) *with data/analysis schemes DA10 and DA11* (*see Table 2*)

Data/analysis scheme	Priors	No of. analysis loci	σ_e^2	σ_a^2	σ_d^2
D10	Uniform	6	54·84 (0·85)	48·07 (1·40)	22·10 (0·90)
		12	50·02 (0·99)	49·53 (1·94)	27·03 (0·89)
		18	48·18 (1·07)	49·84 (1·46)	28·93 (1·11)
DA11	Normal	6	55·37 (0·92)	47·56 (1·32)	21·69 (1·03)
		12	53·18 (0·82)	46·88 (1·20)	23·86 (0·92)
		18	51·68 (0·95)	47·05 (1·22)	25·46 (1·11)

True values for model 2 are error variance (σ_e^2) = 50, additive variance (σ_a^2) = 50, dominance variance (σ_d^2) = 25 and epistasis variance (σ_{aa}^2) = 0.

in both cases. As the locus number in analysis FLM increased from 6 to 18 with normal (bounded uniform) priors for genetic effects, the increase in dominance variance is 4·95 (11·8) for genetic model 1, and 5·18 (12·0) for genetic model 2. Moreover, the increase in total non-additive genetic variance resulting from the increase in locus number in the analysis FLM appeared to be little affected by the number of non-additive components in the analysis FLM, with normal priors. In contrast, the increase in the estimations of total non-additive genetic variance resulting from the increase in locus number in the analysis FLM approximately doubled as the analysis FLM was expanded from additive + dominance to additive + dominance + epistasis with uniform priors (Tables 4, 5).

(iv) *Gibbs sampling of genotypes*, *with dominance*, *12300-pedigree*

We conjectured that the dependency of variance components estimation on locus number in the analysis FLM might (partially) result from data insufficiency. To test this hypothesis, a larger pedigree (12300 members) was simulated for genetic model 2, and the data set was analysed with uniform and normal priors for genetic effects. Results in Table 6 are averages of five replicates obtained from Gibbs samplers of length 100000 with 10000 burn-in cycles. Compared with the corresponding Gibbs sampler for the 6300-pedigree, considerably shorter MCMC was required for the 12300-pedigree, especially when uniform priors were used for genetic effects.

As expected, increasing sample size clearly improved variance components estimation (Tables 5, 6). However, the dependency persists even with data of 12300 members that are reasonably closely related. With uniform priors for genetic effects, the increase in dominance variance for changing from 5 (6) loci to 20 (18) loci in the analysis FLM was 12·0 (6·83) for the 6300-pedigree (12300-pedigree). When locus number in the analysis FLM was changed from 6 to 18 with normal priors, dominance variance increased from 5·18 (3·77) for the 6300-pedigree (12300-pedigree).

We noticed that additive genetic variance was significantly (or near to a significant level with a Student's *t*-test) underestimated with normal priors for genetic effects while it appears to be accurately estimated with uniform priors for genetic effects (Table 6). Further research is required to determine whether this underestimation results from the use of normal priors for genetic effects on a larger data set.

(v) *Gibbs sampling of genotypes*, *effects constant*, *6300-pedigree*

To investigate whether the dependency of the variance components estimates on the number of loci in the analysis FLM was caused or enhanced by allowing effects (A, D, …) to vary across loci, we analysed data, generated under models 1 and 2, with additive and dominance effects constant across loci in the analysis FLM (data/analysis schemes DA12 and DA13; see Table 2). For example, analysis of two data sets for model 1 produced the following results: 5 loci: 47·5, 55·4, 24·3; 10 loci: 39·9, 54·7, 33·0; and 20 loci: 37·7, 54·9, 34·9 for residual, additive and dominance variance, respectively. Estimates from analysis of a data set for model 2 were: 5 loci: 62·94, 49·30, 12·13; 10 loci: 54·72, 49·24, 21·00; and 20 loci: 50·7, 47·99, 25·30. Consequently, holding effects constant across loci does not seem to eliminate or reduce the dependency of estimates on the number of loci. Moreover, accuracy in variance components estimation appeared to be affected by genetic models.

(vi) *Descent graph sampling*, *with dominance*, *6300-pedigree*, *uniform priors*

Gibbs samplers for parameters of length 50000 for the 12- and 18-loci FLMs or 60000 for the 6-loci FLM,

Table 7. *Variance component estimates (and empirical standard errors) averaged across five replicates and obtained by analysing data simulated under models 2 and 5 (see Table 1) with data/analysis schemes DA14 and DA15 (see Table 2)*

Data/ analysis scheme	Genetic model	No. of analysis loci	σ^2_e	σ^2_a	σ^2_d	σ^2_{aa}
DA14	2	6	54·31 (1·71)	51·27 (0·76)	20·63 (1·87)	
		12	50·70 (0·94)	51·20 (1·41)	25·48 (1·69)	
		18	45·35 (2·42)	54·00 (0·42)	30·10 (2·61)	
DA15	5	6	61·33 (1·14)	31·93 (0·84)	15·47 (1·57)	11·93 (1·16)
		12	53·66 (1·18)	30·50 (0·75)	20·34 (1·43)	18·35 (1·06)
		18	47·81 (1·58)	29·73 (0·63)	22·97 (1·65)	23·81 (0·96)

True values for model 2 are error variance (σ^2_e) = 50, additive variance (σ^2_a) = 50, dominance variance (σ^2_d) = 25 and epistasis variance (σ^2_{aa}) = 0. True values for model 5 are $\sigma^2_e = 60$, $\sigma^2_a = 30$, $\sigma^2_d = 15$ and $\sigma^2_{aa} = 15$.

with 4000 burn-in cycles and 5000 MH steps for descent graph sampling of genotypes in each Gibbs cycle, were implemented. Because of the long running time of this sampling scheme (approximately 8 days of CPU on a 250 MHz Origin 2000 machine), only five replicates of data simulated under models 2 and 5 were analysed with uniform priors for genetic effects. As shown in Table 7, the dependency of variance components estimates on the locus number in the analysis FLM remained and showed the same trends (increasing estimates of non-additive genetic variance and a decrease in residual variance with increasing number of loci) as before. The descent graph sampler, which updates the genotypes of all individuals at all loci jointly, produced at best a very modest reduction in the dependency of variance estimate on locus number (Tables 4, 5, 7).

4. Discussion

Finite locus models may provide a better framework for the estimation of genetic parameters and for describing residual polygenic variation in QTL analyses of complex pedigrees than the infinitesimal model. Conceptually, FLMs can easily be extended to include dominance, all two-loci epistasis components, and even higher-order interactions. Moreover, Bayesian analysis of an FLM requires fewer MCMC cycles for variance components estimation with larger data sets. Consequently, computation time required for variance components estimation under the FLM only increases slowly as sample size increases.

Dependency of variance components estimation on the locus number in analysis FLM was observed in this study and in the literature (Pong-Wong *et al*., 1998, 1999; Du *et al*., 1999). To further investigate this dependency, we experimented with various FLM implementation strategies including different genotype sampling schemes, holding genetic effects constant or allowing them to vary, and different priors for genetic effects. While no implementation strategy completely eliminated the dependency, variance components estimation was clearly affected by implementation choices.

Compared with bounded uniform priors for genetic effects, normal priors improved variance components estimation in two aspects. First, the dependency of non-additive estimate on locus number diminished for all genetic models and pedigrees investigated. Second, the dependency appears not to be affected by the number of non-additive genetic components in the analysis FLM. Moreover, fewer MCMC cycles were necessary for Bayesian variance components estimation. It is not surprising that some improvement in variance components was observed by replacing uniform priors with normal priors. With bounded uniform as priors, genetic effects are treated as fixed effects. The estimates of genetic effects belong to the class of shrinkage estimates (genetic effects are regressed towards their means) with normal distributions as priors, and shrinkage estimates generally have lower mean squared errors, especially when the number of classes (i.e. loci) is large. We also experimented with exponential priors for genetic effects, but no further improvement in variance components estimation was found (results not shown).

Further understanding of the properties of individual QTL would help to choose appropriate priors for genetic effects and for QTL size in QTL mapping. Ideally, we would like to choose priors that are close to their true distributions. While little information about properties of individual QTL is available for economically important quantitative traits in food animals, the distribution of the genetic effects of newly accumulated mutations on bristle number in *Drosophila* was highly skewed and leptokurtic (Mackay *et al*., 1992; Caballero & Keightley, 1994). In reality, detailed information on the number and frequencies of alleles, and possibly on linkage relationships, will be incomplete for quantitative traits

in food animals, at least for the foreseeable future. Moreover, the variance components estimation should be primarily determined by data instead of prior information. Consequently, introducing some shrinkage such as using normal priors with unknown scale parameter appears to be an appropriate choice for genetic effects.

Overparameterization was suggested as one of the possible causes of the dependency of variance estimation on locus number (Du *et al*., 1999; Pong-Wong *et al*., 1999). Pong-Wong *et al*. (1999) suggested that holding genetic effects constant across loci might avoid overparameterization. However, biases and dependencies persisted in the presence of non-additive genetic variation, with genetic effects held constant across loci in the analysis.

We first conjectured this dependency might be due to poor mixing of the Gibbs sampler for genotypes. We therefore implemented a descent graph sampling scheme, as it appeared to be the most efficient way of performing joint sampling of the genotypes of all individuals at all loci, compared with other methods such as peeling, which would need to be extended to multiple loci. Our current implementation of the descent graph Markov chain requires a long chain with many MH steps within each Gibbs update of the parameters. Based on the results in Table 7 and other results not shown, the descent graph sampler produced little to no improvement in the parameter estimation, with the dependency between genetic parameter estimates and number of loci remaining. We also implemented a MH sampler that updates all genotypes and parameters jointly by proposing changes to the genotypes (similar to the second scheme), and to the parameters based on normal or uniform distributions with small spread and centred at the previous sample value in each cycle of sampling. Our current implementation of this scheme did not improve the estimation of variance components (results not shown). The sampling scheme does not seem to be the (major) factor causing the dependency of parameter estimates on locus number. When we initially implemented the FLM with a single-site Gibbs sampler updating genotypes conditionally on the genotypes of final offspring, genetic parameter estimates were very poor. They remained poor when we sampled fathers unconditionally on final offspring genotypes but mothers still conditionally on the genotypes of their final offspring. Sampling both mothers and fathers unconditionally on final offspring genotypes substantially improved parameter estimation and seems to work (nearly) as well as the joint sampling of all genotypes. Sampling the genotypes of one locus conditionally on the genotypes at other loci is not optimum but should not lead to very poor mixing, as long as loci are unlinked, as in the FLM, or not closely linked.

It appears instead that the dependency of genetic variance component estimates on locus number is due to data insufficiency. With bounded uniform priors for additive effects in a purely additive FLM, the estimate of additive genetic variance for a pedigree with 480 members increased rapidly as locus number in analysis FLM increased (Pong-Wong *et al*., 1999), while variance estimates were independent of locus number in the analysis FLM for a pedigree of 6300 we simulated under a purely additive model (Du *et al*., 1999). Moreover, the dependency (with either uniform or normal priors for genetic effects) considerably diminished for the non-additive FLMs in this paper, as the sample size increased from 6300 to 12 300. In all likelihood the dependency should be eliminated by further increasing sample size.

While the hypothesis that the dependency results from data insufficiency supports the validity of FLMs, estimating variance components most accurately with a limited amount of data is of practical importance. In addition to the use of non-uniform priors for genetic effects, it may be worthwhile to investigate extension of the FLM to including the number of loci as an additional unknown via a reversible jump Metropolis-Hastings sampler (Green, 1995), letting the data determine an optimum number of loci. Although the genotype sampling scheme did not appear to be a major factor in this investigation, further research aimed at optimally sampling genotypes at all loci and for all individuals jointly is warranted, both for FLM implementation and QTL mapping. Our current research focuses on improving descent graph sampling, implementing genotypic or allelic peeling for multiple loci in complex pedigrees efficiently, and comparisons among these schemes.

This research was supported by the US Department of Agriculture's National Research Initiative Competitive Grants Program (grant 96-35205-3662) and by the National Science Foundation (grant DBI-9723022). Thanks go to Bruce Tier for helpful discussions on implementation of descent graph samplers. I. H. also acknowledges financial support from NIH grant GM45344 (PI: Bruce Weir) while on sabbatical at North Carolina State University.

References

Abplanalp, H. (1988). Selection response in inbred lines of white leghorn chickens. In *Proceedings of the Second International Conference in Quantitative Genetics*, (eds. B. S. Weir, E. J. Eisen, M. M. Goodman & G. Namkoong), pp. 360–378. Sunderland, MA: Sinauer Associates.

Allaire, F. R. & Henderson, C. R. (1965). Specific combining abilities among dairy sires. *Journal of Dairy Science* **48**, 1096–1100.

Caballero, A. & Keightley, P. D. (1994). A pleiotropic nonadditive model of variation in quantitative traits. *Genetics* **138**, 883–900.

Cockerham, C. C. (1954). An extension of the concept of partitioning hereditary variance for analysis of co-

variances among relatives when epistasis is present. *Genetics* **39**, 859–882.

De Boer, I. J. M. & Hoeschele, I. (1993). Genetic evaluation methods for populations with dominance and inbreeding. *Theoretical and Applied Genetics* **86**, 245–258.

DeStefano, A. L. & Hoeschele, I. (1992). Utilization of dominance variance through mate allocation strategies. *Journal of Dairy Science* **75**, 1680–1690.

Du, F.-X., Hoeschele, I. & Gage-Lahti, K. M. (1999). Estimation of additive and dominance variance components in finite polygenic models and complex pedigrees. *Genetical Research* **74**: 179–187.

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

Fernando, R. L., Stricker, C. & Elston, R. C. (1994). The finite polygenic mixed model: an alternative formulation for the mixed model of inheritance. *Theoretical and Applied Genetics* **88**, 573–580.

Fuerst, C. & Soelkner, J. (1994). Additive and nonadditive genetic variance of milk yield, fertility, and life time performance traits of dairy cattle. *Journal of Dairy Science* **77**, 1114–1125.

Fuerst, C., James, J. W., Soelkner, J. & Essl, A. (1997). Impact of dominance and epistasis on the genetic make-up of simulated populations under selection: a model development. *Journal of Animal Breeding and Genetics* **114**, 163–175.

Fuerst-Waltl, B., Soelkner, J., Essl, A. & Hoeschele, I. (1998). Estimation of non-linear genetic relationships between yield and type traits in Holstein cattle. In *Proceedings of the Sixth World Congress on Genetics Applied to Livestock Production*, Armidale, **23**, 415–418.

Georges, M., Nielsen, D., Mackinnon, M., Misbra, A., Okimoto, R., Pasquino, A. T., Sargeant, L. S., Sorensen, A., Steele, M. R., Zhao, X., Womack, J. E. & Hoeschele, I. (1995). Mapping quantitative trait loci controlling milk production in dairy cattle by exploiting progeny testing. *Genetics* **139**, 907–920.

Geyer, C. J. (1992). Practical Markov Chain Monte Carlo. *Statistical Science* **7**, 467–511.

Goddard, M. E. (1998). Gene based model for genetic evaluation: an alternative to BLUP? In *Proceedings of the Sixth World Congress on Genetics Applied to Livestock Production*, Armidale **26**, 33–36.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732.

Hay, G. M., White, J. M., Vinson, W. E. & Kliewer, R. H. (1983). Components of genetic variation for descriptive type traits of Holsteins. *Journal of Dairy Science* **66**, 1962–1966.

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

Hoeschele, I. (1991). Additive and nonadditive genetic variance in female fertility of Holsteins. *Journal of Dairy Science* **74**, 1743–1752.

Hoeschele, I. & VanRaden, P. M. (1991). Rapid method to compute inverses of dominance relationship matrices for noninbred populations including sire × dam subclass effects. *Journal of Dairy Science* **74**, 557–569.

Hoeschele, I., Uimari, P., Grignola, F. E., Zhang, Q. & Gage, K. (1997). Advances in statistical methods to map quantitative trait loci in outbred populations. *Genetics* **147**, 1445–1457.

Janss, L. L. G., Thompson, R. & VanArendonk, J. A. M. (1995). Application of Gibbs sampling for inference in a mixed major gene-polygenic inheritance model in animal populations. *Theoretical and Applied Genetics* **91**, 1137–1147.

Kao, C.-H. & Zeng, Z.-B. (2000). Modeling epistasis of quantitative trait loci using Cockerham's model. *Theoretical Population Biology* (submitted).

Lander, E. S. & Botstein, D. (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**, 185–199.

Mackay, T. F., Lyman, R. F. & Jackson, M. S. (1992). Effects of P element insertions on quantitative traits in *Drosophila melanogaster*. *Genetics* **130**, 315–332.

Mitchell, B. D., Ghosh, S., Schneider, J. L., Birznieks, G. & Blangero, J. (1997). Power of variance component linkage analysis to detect epistasis. *Genetic Epidemiology* **14**, 1017–1022.

Pong-Wong, R., Shaw, F. & Woolliams, J. A. (1998). Estimation of dominance variation using a finite-locus model. In *Proceedings of the Sixth World Congress on Genetics Applied to Livestock Production*, Armidale **26**, 41–44.

Pong-Wong, R., Haley, C. S. & Woolliams, J. A. (1999). Behaviour of the additive finite locus model. *Genetics*, *Selection*, *Evolution* **31**, 193–211.

Sobel, E. & Lange, K. (1996). Descent graphs in pedigree analysis: application to haplotyping, location scores, and marker-sharing statistics. *American Journal of Human Genetics* **58**, 1323–1337.

Sorensen, D. A., Andersen, S., Gianola, D. & Korsgaard, I. (1995). Bayesian inference in threshold models using Gibbs sampling. *Genetics*, *Selection*, *Evolution* **27**, 229–249.

Stricker, C., Fernando, R. L. & Elston, R. C. (1995). Linkage analysis with an alternative formulation for the mixed model of inheritance: the finite polygenic mixed model. *Genetics* **141**, 1651–1656.

Tempelman, R. J. & Burnside, E. B. (1990). Additive and nonadditive genetic variation for production traits in Canadian Holstein. *Journal of Dairy Science* **73**, 2206–2213.

Thompson, E. A. (1994). Monte Carlo likelihood in genetic mapping. *Statistical Science* **9**, 355–366.

Thompson, E. A. & Skolnick, M. H. (1977). Likelihoods on complex pedigrees for quantitative traits. In *Proceedings of the International Conference on Quantitative Genetics* (eds. E. Pollack, O. Kempthorne & T. B. Bailey Jr), pp. 815–818. Ames, Iowa: Iowa State University Press.

Tier, B. & Henshall, J. (1999). A sampling algorithm for segregation analysis. *Genetics*, *Selection Evolution* (submitted).

VanRaden, P. M. & Hoeschele, I. (1991). Rapid method to compute inverses of additive-by-additive relationship matrices including sire–dam combination effects. *Journal of Dairy Science* **74**, 570–579.

VanRaden, P. M., Lawlor, T. J., Short, T. H. & Hoeschele, I. (1992). Use of reproductive technology to investigate gene interactions. *Journal of Dairy Science* **75**, 2892–2901.

Wang, C. S., Rutledge, J. J. & Gianola, D. (1993). Marginal inferences about variance components in a mixed linear model using Gibbs sampling. *Genetics*, *Selection*, *Evolution* **25**, 41–62.

Zhang, Q., Boichard, D., Bishop, M., Hoeschele, I., Ernst, C., Doud, L., Eggen, A., Jugella, G., Murkve, B., Pfister-Genskov, M., Thorbahn D., Uimari, P., Grignola, F. E. & Thaller, G. (1998). Mapping quantitative trait loci for milk production and health of dairy cattle in a large, outbred pedigree. *Genetics* **149**, 1959–1973.