Genet. Res., Camb. (2000), **75**, *pp.* 231–241. *With* 4 *figures.* Printed in the United Kingdom © 2000 Cambridge University Press

Markov chain Monte Carlo for mapping a quantitative trait locus in outbred populations

M. C. A. M. BINK[1]*, L. L. G. JANSS[2] AND R. L. QUAAS[3]

[1] *Animal Breeding and Genetics Group, Wageningen Institute of Animal Sciences, Wageningen Agricultural University, Wageningen, PO Box 338, 6700 AH, The Netherlands*

[2] *Institute for Animal Science and Health (ID-DLO), 8200 AB Lelystad, The Netherlands*

[3] *Department of Animal Science, Cornell University, Ithaca, NY 14853-4801, USA*

(*Received 27 July 1998 and in revised form 7 January and 29 June 1999*)

Summary

A Bayesian approach is presented for mapping a quantitative trait locus (QTL) using the 'Fernando and Grossman' multivariate Normal approximation to QTL inheritance. For this model, a Bayesian implementation that includes QTL position is problematic because standard Markov chain Monte Carlo (MCMC) algorithms do not mix, i.e. the QTL position gets stuck in one marker interval. This is because of the dependence of the covariance structure for the QTL effects on the adjacent markers and may be typical of the 'Fernando and Grossman' model. A relatively new MCMC technique, simulated tempering, allows mixing and so makes possible inferences about QTL position based on marginal posterior probabilities. The model was implemented for estimating variance ratios and QTL position using a continuous grid of allowed positions and was applied to simulated data of a standard granddaughter design. The results showed a smooth mixing of QTL position after implementation of the simulated tempering sampler. In this implementation, map distance between QTL and its flanking markers was artificially stretched to reduce the dependence of markers and covariance. The method generalizes easily to more complicated applications and can ultimately contribute to QTL mapping in complex, heterogeneous, human, animal or plant populations.

1. Introduction

The availability of dense molecular markers facilitate study of the segregation of chromosomal segments from parents to offspring and allows the mapping of loci responsible for variation in quantitative traits (quantitative trait loci or QTLs) in humans, animals and plants. A variety of methods are used for identification of marker–QTL associations (e.g. Weller, 1986; Knott & Haley, 1992). Most were developed assuming particular mating designs, e.g. backcrosses or F2s, leading to simple pedigrees. These methods cannot fully account for, nor can easily be extended to, the more complex data structures of outbred populations such as are found in domesticated farm animals.

In this study we explore models and methods that can more easily be extended to complex pedigrees in QTL mapping analysis. Markov Chain Monte Carlo (MCMC) algorithms (Metropolis *et al.*, 1953; Hastings, 1970; Geman & Geman, 1984) here play an important role, because they provide a powerful computational tool for analysis of complex data structures, either in a maximum likelihood or a Bayesian context. Ideas of a Bayesian analysis for QTL detection were described by Hoeschele & Vanraden (1993*a*, *b*), and implemented via MCMC algorithms in contributions by Thaller & Hoeschele (1996), Satagopan *et al.* (1996), Umari *et al.* (1996), Uimari & Hoeschele (1997) and Sillanpää & Arjas (1998). Most of these Bayesian methods assume a bi-allelic QTL model (Hoeschele *et al.*, 1997). Though reasonable for a cross of inbred strains it is less so for a population such as the Holstein breed of dairy

* Corresponding author. Current address: Centre for Biometry Wageningen (CBW), DLO – Centre for Plant Breeding and Reproduction Research (CPRO-DLO), PO Box 16, 6700 AA, Wageningen, The Netherlands. Tel: +31 317 477306. Fax: +31 (0)317 418094. e-mail: m.c.a.m.bink@cpro.dlo.nl

cattle. Outside North America, populations typically resulted from several crosses of the North American breed on the local strain of black and white cattle and the gene flow among countries continues unabated. A population with such varied origins is a long way from inbred strains, so a polyallelic model seems more appropriate.

In this paper, we focus on Bayesian inferences in the multivariate Normal QTL model of Fernando & Grossman (1989) in which QTL effects are assumed with a covariance structure dependent on markers adjacent to the postulated QTL position. This model should be more appropriate for heterogeneous populations and the MCMC algorithms allow extensions to complex designs. We show that a straightforward implementation of a Metropolis-Hastings (MH) algorithm to shuffle the QTL position within the linkage map leads to an effectively reducible Markov chain, i.e. not all possible positions are reached from a given starting position of the QTL. We suggest a modified MCMC scheme, simulated tempering (Marinari & Parisi, 1992; Geyer & Thompson, 1995), to solve the mixing problem for the QTL position. This scheme is evaluated empirically for simulated data from a granddaughter design (Weller *et al*., 1990). In a granddaughter design, marker genotypes are available on elite sires and their sons and trait phenotypes are observed on daughters of sons. The extension and application of the Bayesian method presented to complex pedigree analysis to detect QTL in outbred populations are discussed.

2. Method and application

(i) *Marker information*

The marker data (**m**) is assumed to include the genotypes at a number of marker loci that have been assigned to a particular linkage group. In this study we assume that the order of and the distances between these marker loci are known with certainty.

Let **g** represent the set of true genotypes for all individuals and for all marker loci. That is, for founder individuals the linkage phase among alleles at linked marker loci is known and for non-founder individuals it is clear which of the parental alleles have been inherited (even when a parent is homozygous!). However, the observed marker data (**m**) probably do not lead to a unique set of linkage phases and allele transmissions; consequently multiple **g**'s may apply. Let g_i be a particular consistent set, and let $P(\mathbf{g} = g_i \,|\, \mathbf{m})$ be its probability, conditional on the observed marker data. Then, the identification of every consistent g_i and the calculation of its probability become intractable for large outbred pedigrees.

The presence of a single QTL within the marked chromosomal segment is postulated. The map position of the QTL is denoted d, and it is relative to the first marker of the linkage group. Chromosomal segments outside the linkage group are not considered to avoid identification problems between size and position of the QTL. That is, a small QTL close to a marker is as likely as a large QTL further away from a marker in the type of model used here (e.g. van Arendonk *et al*., 1998).

(ii) *A Bayesian hierarchical model*

Let n and q denote the number of phenotypic values for the quantitative trait and the number of individuals in the pedigree, respectively. The phenotypic values for the quantitative trait (**y**) are assumed to be normally distributed, i.e.

$$\mathbf{y} \,|\, \mathbf{b}, \mathbf{u}, \mathbf{v}, \sigma_e^2 \sim N(\mathbf{Xb} + \mathbf{Zu} + \mathbf{ZTv}, \mathbf{I}\sigma_e^2), \tag{1}$$

where **y** is a $n \times 1$ vector of phenotypic values; **b** is a vector of fixed effects (in a Bayesian setting treated as a vector of random effects witha flat prior distribution representing no prior knowledge about the values); **u** and **v** are $q \times 1$ and $2q \times 1$ random vectors of polygenic and QTL effects; **X** and **Z** are appropriately dimensioned incidence matrices relating **b** and **u** to **y**, respectively; **T** is a known matrix relating each individual to its two QTL allelic effects; **e** is a vector of random residual effects; **I** is an identify matrix and σ_e^2 is the residual variance.

Next, the polygenic effects are assumed to be due to loci that segregate independently of those at the marker loci, and the QTL allelic effects co-segregate with alleles at the marker loci, creating a stochastic dependence. These are assumed to be normally distributed as

$$\left.\begin{aligned} \mathbf{u} \,|\, \sigma_u^2 &\sim N(\mathbf{0}, \mathbf{A}\sigma_u^2), \\ \mathbf{v} \,|\, \sigma_v^2, d, \mathbf{g} &\sim N(\mathbf{0}, \mathbf{G}_{|d,\mathbf{g}}\sigma_v^2), \end{aligned}\right\} \tag{2}$$

where σ_u^2 and σ_v^2 are the polygenic variance and half the additive genetic variance explained by the QTL, respectively. Total additive genetic variance (due to polygenes and QTL) is $\sigma_a^2 = \sigma_u^2 + 2\sigma_v^2$, and total phenotype variance is $\sigma_p^2 = \sigma_a^2 + \sigma_e^2$. **A** is the additive genetic relationship matrix (Henderson, 1976), $\mathbf{G}_{|d,\mathbf{g}}$ is the gametic relationship matrix for QTL effects conditional on map position of the QTL (d) and (complete) marker information (**g**). Note that matrix $\mathbf{G}_{|d,\mathbf{g}}$ has $2q \times 2q$ elements, where element (i,j) represents the probability of QTL allele i being identical by descent to QTL allele j (e.g. Wang *et al*., 1995). These identity by descent probabilities for QTL effects are easily computed for situations with known d and known **g** (Bink & van Arendonk, 1999). However, parameter d remains to be estimated and, especially in outbred populations, knowledge on linkage phases among marker alleles and segregation of marker alleles is likely to be incomplete. Bink & van Arendonk (1999) have described an approach that fully accounts

https://doi.org/10.1017/S0016672399004310 Published online by Cambridge University Press

for incomplete knowledge on marker genotypes (and allelic frequencies for marker loci), given a particular map position of the QTL, and their MCMC sampling approach to account for this uncertainty in marker genotypes is used in this study. The allelic frequencies (η) at a particular marker locus in a population are likely unknown and here also treated as such: see Bink & van Arendonk (1999) for details.

Similar to Bink *et al.* (1998), the model in (1) is parameterized in terms of the residual variance, σ_e^2, the heritability $h^2 = (\sigma_u^2 + 2\sigma_v^2)/\sigma_p^2$, and the proportion of genetic variance due to the QTL $\gamma = 2\sigma_v^2/\sigma_a^2$. The prior density of σ_e^2 is a U[κ_{e1}, κ_{e2}], where κ_{e1} and κ_{e2} are equal to 0 and a pre-defined maximum value for σ_e^2, respectively. In general, Beta($\alpha_{h^2}, \beta_{h^2}$) and Beta($\alpha_\gamma, \beta_\gamma$) distributions can specify prior assumptions on dispersion parameters h^2 and γ. We arbitrarily set the values of the hyperparameters α_{h^2}, β_{h^2} α_γ and β_γ equal to unity, resulting in U[0, 1] priors. Note that the choice of these values may affect the posterior inference on the parameter of interest (e.g. Bink *et al.*, 1998). The prior density of QTL position (d) is assumed to be $U[d_F, d_L]$, where d_F and d_L represent the map positions of the first and last marker of the linkage group, respectively.

The joint density of the parameters given the data (**y** and **m**) and the prior information is

$$\left.\begin{aligned} &f(\mathbf{b}, \mathbf{u}, \mathbf{v}, \sigma_e^2, h^2, \gamma, d, \eta, \mathbf{g} \mid \mathbf{y}, \\ &\quad \mathbf{m}, \kappa_{e1}, \kappa_{e2}, \alpha_{h^2}, \beta_{h^2}, \alpha_\gamma, \beta_\gamma), \\ &\propto f(\mathbf{b}, \mathbf{u}, \mathbf{v}, \sigma_e^2, h^2, \gamma, d, \eta, \mathbf{y}, \\ &\quad \mathbf{m} \mid \kappa_{e1}, \kappa_{e2}, \alpha_{h^2}, \beta_{h^2}, \alpha_\gamma, \beta_\gamma), \\ &\propto f(\mathbf{y} \mid \mathbf{b}, \mathbf{u}, \mathbf{v}, \sigma_e^2) \times f(\mathbf{b}) \\ &\quad \times f(\mathbf{u} \mid \sigma_e^2, h^2, \gamma) \times f(\mathbf{v} \mid \sigma_e^2, h^2, \gamma, d, \mathbf{g}) \\ &\quad \times f(\sigma_e^2 \mid \kappa_{e1}, \kappa_{e2}) \times f(h^2 \mid \alpha_{h^2}, \beta_{h^2}) \\ &\quad \times f(\gamma \mid \alpha_\gamma, \beta_\gamma) \times P(\mathbf{g} \mid \mathbf{m}, \eta) \times f(d). \end{aligned}\right\} \quad (3)$$

From this joint posterior density, the full conditional distribution for each parameter is obtained by retaining only those parts that contain the parameter and treating the remainder as a constant. From here on we will suppress the dependence on the hyperparameters in the notation. For a particular QTL position the full conditional posterior densities, proposal distributions to sample from, for all unknowns are similar to those in Bink *et al.* (1998) and Bink & van Arendonk (1999). For the location parameters **b**, **u** and **v** full conditional posterior distributions are Normal; the full conditional posterior distribution for σ_e^2 is a truncated scaled inverted chi-squared distribution with degrees of freedom equal to $\dim(\mathbf{e}) - 2$; the resulting full conditional posterior densities for h^2 and γ are non-standard and samples for these parameters are obtained via a MH algorithm (for details see Bink *et al.*, 1998). For updating the marker genotype information, we refer to Bink & van Arendonk (1999). For the application developed here, only the full conditional distribution for d remains, which can be obtained from the joint posterior distribution (3) by omitting those parts that do not involve d itself. Let θ_{-d} denote the set of unknowns excluding parameter d. The position of the QTL affects only the elements of matrix **G**, and the full conditional can be given as

$$\begin{aligned} f(d \mid \theta_{-d}, \mathbf{y}, \mathbf{m}) &\propto f(\mathbf{v} \mid \sigma_e^2, h^2, \gamma, d, \mathbf{g}) \times f(d) \\ &\propto |\mathbf{G}_{|d,\mathbf{g}}\sigma_v^2|^{-1/2} \times \exp\{-\tfrac{1}{2}(\mathbf{v}-\mathbf{0})^T (\mathbf{G}_{|d,\mathbf{g}}\sigma_v^2)^{-1} \\ &\quad (\mathbf{v}-\mathbf{0})\} \times U[d_F, d_L] \\ &\propto \begin{cases} |\mathbf{G}_{|d,\mathbf{g}}|^{-1/2} \\ \quad \times \exp\left\{-\dfrac{1}{2\sigma_v^2}(\mathbf{v}^T\mathbf{G}_{|d,\mathbf{g}}^{-1}\mathbf{v})\right\} & \text{if } d \in [d_F, d_L] \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \quad (4)$$

This full conditional distribution does not have a recognizable kernel and samples from this distribution are obtained via a MH algorithm. In the MH algorithm, a candidate position, d_j, is generated by a candidate generating density, denoted $q(.)$, and (4) is evaluated for current and candidate positions, d_i and d_j. The probability of a move, i.e. acceptance of candidate value d_j, is $\min(\alpha(i,j), 1)$, where

$$\alpha(i,j) = \frac{f(d_j \mid \theta_{-d}, \mathbf{y}, \mathbf{m})}{f(d_i \mid \theta_{-d}, \mathbf{y}, \mathbf{m})} \times \frac{q(d_i; d_j)}{q(d_j; d_i)}. \quad (5)$$

The latter ratio in (5) accounts for uneven proposal probabilities. In this study we use the random walk approach (Chib & Greenberg, 1995) to sample candidates, i.e. a uniform proposal density centred on the current value d_i. The length of this uniform is determined empirically and should result in average acceptance rates between 0·20 and 0·50 (Chib & Greenberg, 1995) to ensure proper mixing through the parameter space. Note that for a discrete prior on d, i.e. a grid search with a finite number of positions, the Gibbs sampler might be applicable. In that case, however, summation of probabilities on all positions is required and this rapidly becomes too demanding for large numbers of positions.

(iii) *Practical reducibility of the MCMC chain*

Preliminary trials with the MCMC chain as described above revealed a severe mixing problem with respect to QTL position. A candidate position for the QTL, say d_j, in another marker interval involves a different set of marker loci (and genotypes) and differences arise in elements of **G** and its inverse. As a result of these differences the quadratic form $(\mathbf{v}^T\mathbf{G}_{|d_j,\mathbf{g}}^{-1}\mathbf{v}) \gg (\mathbf{v}^T\mathbf{G}_{|d_i,\mathbf{g}}^{-1}\mathbf{v})$ (equation (4) since values for **v** were sampled conditional on $\mathbf{G}_{|d_i,\mathbf{g}}^{-1}$. Consequently, a relatively very small value for the numerator in (5) was obtained, and, for large pedigrees, the probability of a move in (5) was practically zero, as will be described in Section 3. The QTL position was stuck within the starting

https://doi.org/10.1017/S0016672399004310 Published online by Cambridge University Press

marker interval, no matter which starting position was chosen, i.e. the chain was effectively reducible.

(iv) *Simulated tempering*

An approach to solving poor mixing in MCMC is the simulated tempering sampler (Marinari & Parisi, 1992; Geyer & Thompson, 1995). Simulated tempering is an adaptation of simulated annealing (Kirkpatrick *et al*., 1983). Simulated annealing is a Monte Carlo approach to minimize 'complex' cost functions and its name derives from the roughly analogous physical process of heating and then slowly cooling a substance to obtain a strong crystalline structure. The simulated annealing process lowers the temperature by slow stages until the system 'freezes' and no further changes occur. Simulated tempering treats the temperature stochastically, i.e. the temperature fluctuates randomly between cold and hot stages (densities). The simulated tempering sampler draws samples from a family of densities (models), and switches between densities (models) randomly over time. So, rather than just one, a set of full conditional posterior densities is sampled from, one being the target and the others being modifications with better mixing properties. A way to set up a useful family of densities is to define a series of more and less 'heated' versions of the target density. In 'heating' the target density, this density is flattened, making it easier for the chain to move around in the parameter space. When the 'hottest' version allows sampling of the 'non-mixing' parameter independent of any other parameter, complete mixing is guaranteed. Geyer & Thompson (1995) give a full description of the simulated tempering sampler; here we prefer just to describe our application to maintain readability.

Two crucial stages in constructing the simulated tempering sampler are definition of the heating modification, i.e. how to modify the original target density to improve mixing of the parameter throughout its sampling space, and the fine-tuning process of the number of heated modifications and their relative distances.

The heating modification was applied here to the Haldane mapping function (Haldane, 1919) that is used to compute the recombination rates between the QTL and its flanking markers. Heating this mapping function implies that the QTL becomes less linked to the map, i.e. covariances among QTL effects of related individuals become less dependent on inheritance of marker alleles at flanking loci. A new parameter, temperature (denoted λ), is used as an index in the simulated tempering sampler which modifies the mapping function into

$$r = (\lambda) \times 0{\cdot}5 + (1-\lambda) \times 0{\cdot}5 \times (1{\cdot}0 - e^{-2d}), \qquad (6)$$

where $0 \leqslant \lambda \leqslant 1$. Now, for $\lambda = 0$ the true mapping function is applied and samples are drawn from the (cold) target density. On the other hand, for $\lambda = 1$ the mapping function reduces to a constant, i.e. the recombination fraction equals 0·5 and there is no linkage between QTL and its flanking marker loci. In the latter case, matrix $\mathbf{G}^{-1}_{|d,\mathbf{g}}$ is no longer affected by marker information and each position of the map is equally likely. This means that for $\lambda = 1$, the quadratic $(\mathbf{v}^T\mathbf{G}^{-1}_{|d_j,\mathbf{g}}\mathbf{v})$ is equal to $(\mathbf{v}^T\mathbf{G}^{-1}_{|d_i,\mathbf{g}}\mathbf{v})$ and the candidate position d_j is always accepted, i.e. $\alpha(i,j) = 1$ (see Section 2(iii)). When candidates are always accepted, independent sampling occurs and this guarantees that the entire sampling space can be reached within the MCMC chain (Geyer & Thompson, 1995).

In the simulated tempering sampler, λ has a discrete distribution where the number of the distances between classes (values of λ_is) have to be defined empirically. Similar to Geyer & Thompson (1995), we implement a MH algorithm to update values of λ and only allow moves between adjacent classes. Furthermore, we also used so-called pseudopriors to obtain equal probabilities on moving up and down between two adjacent classes λ_i and λ_{i+1}. We closely followed the procedures suggested by Geyer & Thompson (1995) to fine-tune the spacing of λs and their pseudopriors to arrive at desired acceptance rates (0·20–0·50). We fully agree with them that this process of fine-tuning requires considerable effort.

(v) *Simulated data*

To evaluate the effectiveness of the simulated tempering sampler, Monte Carlo simulation was used to generate granddaughter designs comprising 20 unrelated elite sire families each having 40 sons (paternal half-sibs). This approximately reflects a Dutch granddaughter experiment design as described by Spelman *et al*. (1996). Polygenic and QTL effects for grandsires were sampled from $N(0, \sigma^2_u)$ and $N(0, \sigma^2_v)$, respectively. The polygenic effect for a son was simulated as $u_{son} = \frac{1}{2}u_s + \phi$, where u_s is the elite sire's polygenic effect, and ϕ, Mendelian sampling, is distributed independently as $N(0, \mathrm{Var}(\phi))$ with $\mathrm{Var}(\phi) = 0{\cdot}75 \times \sigma^2_u$ (no inbreeding). Each son inherited one QTL allele at random from its (elite) sire. The maternally inherited QTL effect for a son was drawn from $N(0, \sigma^2_v)$. Each son had 100 daughters with phenotypic values. A son transmits half its polygenic effect to each of its daughters and transmits either its first (v^1_{son}) or second (v^2_{son}) QTL effect to a particular daughter. A phenotypic value was then generated as

$$y \,|\, u_{son}, v^1_{son}, v^2_{son}, \sigma^2_u, \sigma^2_v, \sigma^2_e \sim N((\tfrac{1}{2}u_{son} + (\rho)\, v^1_{son} + (1-\rho)\, v^2_{son}), (\tfrac{3}{4}\sigma^2_u + \sigma^2_v + \sigma^2_e)),$$

where for each daughter ρ is randomly taken as 0 for 1 with equal probabilities. The phenotypic variance and the heritability of the trait were 100 and 0·40, respectively. The proportion of genetic variance due

https://doi.org/10.1017/S0016672399004310 Published online by Cambridge University Press

Table 1. *Characteristics of simulation of data*

Data	Proportion QTL (γ)	QTL position[a]	Heterozygosity[b]
I	0·25	90 cM	100 %
II	0·00	—	100 %
III	0·25	90 cM	60 %
IV	0·25	50 cM	60 %

[a] Position of QTL relative to the map position of first marker in linkage group.
[b] Heterozygosity is the percentage of heterozygous marker genotypes for grandsires.

to a single QTL ($=\gamma$) was 0·25, except for data II where $\gamma = 0{\cdot}00$ (Table 1). Data II was chosen to verify that absence of a QTL within the linkage map was also inferred as such in the Bayesian analysis.

Marker data were generated for all elite sires and sons. Six markers were spaced equidistantly (20 cM, Haldane's mapping function) with the first marker being the origin of the linkage map. Each marker locus contained five alleles with equal frequencies. For elite sires, the information content of marker genotypes, i.e. being heterozygous, was arbitrarily set equal to 100 % or 60 % (Table 1). The 100 % heterozygosity is the ideal situation; 60 % is a level found in practice (e.g. chromosome 6 in dairy cattle: Spelman *et al*., 1996).

(vi) *MCMC simulation and post-MCMC analysis*

Initial values for location parameters (**b**, **u** and **v**) were zero, while starting values for σ_e^2, h^2, and γ were 60·0, 0·40 and 0·25, respectively. The initial genotypes for marker loci were imputed conditional on pedigree and marker data but not, however, accounting for linkage among these loci. To ensure probable linkage phases in parents and segregation of alleles to offspring, the initial genotypes were updated 25 times before starting the actual MCMC chain. Initial allele frequencies (η) for all marker loci were equi-frequent ($=0{\cdot}2$). The simulated tempering sampler always started in the hottest distribution ($\lambda_n = 1$). Due to independent sampling of d in this distribution, the starting value for d was not relevant. In each iteration (in chronological order), **g**, η, **b**, **u**, **v** and σ_e^2 were updated by Gibbs sampling, while h^2, γ, d and λ_j were updated by MH algorithms. To decrease the number of elements in **u** and **v**, a reduced animal model was fitted (Bink *et al*., 1998). For each of the four data sets, one final long MCMC chain was run (after fine-tuning the number of distributions with their spacing and pseudopriors in the simulated tempering scheme). The length of each MCMC run was arbitrarily set at 5 000 000 iterations. Total CPU time per MCMC run was about 40 h on a HP 9000-k260 server, while a similar amount of time was spent on fine-tuning the simulated tempering sampler. The samples for parameters σ_e^2, h^2, γ and d were stored when the cold distribution ($\lambda_j = 0$) was visited.

For data I and II, we constructed a simulated tempering sampler with 35 distributions, $\lambda_1 = 0 < \lambda_2 < \ldots < \lambda_{35} = 1$, to move from cold to hot and reverse, resulting in average acceptance rates of 0·30. The simulated tempering samplers for data III and IV required fewer distributions ($=26$) to obtain similar acceptance rates. This difference is probably due to the lower heterozygosity of markers in data III and IV, i.e. data on less informative markers are relatively more similar to the absence of marker data (which is the situation when sampling in the hottest distribution). To check convergence of the MCMC chain, we computed for several parameters the number of effective samples – a measure suggested by Sorensen *et al*. (1995).

Bayesian inference about a particular parameter θ is via the posterior distribution $p(\theta|\mathbf{y})$. The highest posterior density region attempts to capture a comparatively small region of the parameter space that contains most of the mass of the posterior distribution. We computed a 90 % highest posterior density region (HPD90). The null hypothesis that the QTL explains no genetic variance was tested via a posterior odds ratio, i.e. a ratio between the probability for a small bin at the posterior mode and the probability for a small bin near zero. If the probability for a small bin near zero equalled zero, the denominator was set equal to 0·001. Presence of a QTL was postulated when the posterior odds ratio > 20, or its natural log, denoted ln(odds), $> 3{\cdot}0$, a threshold first suggested by Janss *et al*. (1995). Note that the prior odds ratio was equal to 1.

3. Results and discussion

(i) *Mixing of QTL position and convergence of MCMC chain*

Results from the simulated tempering sampler clearly indicated that QTL position d did not mix between marker intervals in the cold target distribution. The mixing of the QTL position only occurred near the hot end of the 'heated' distributions. Let n denote the number of 'heated' distributions, ranging from the (cold) target distribution (with temperature, λ_1,

https://doi.org/10.1017/S0016672399004310 Published online by Cambridge University Press

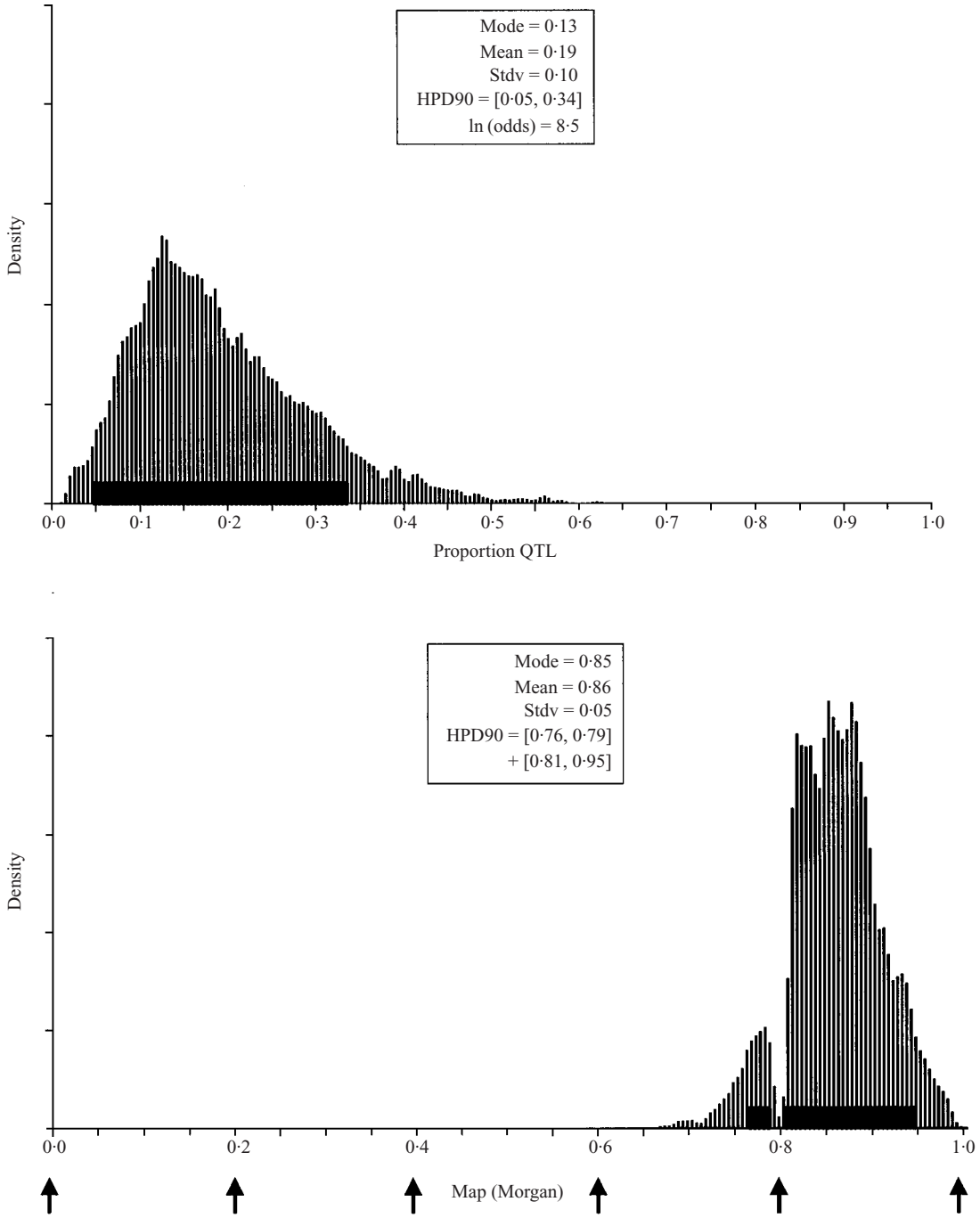

Fig. 1. Marginal posterior densities for the proportion of the genetic variance due to the QTL (γ) (above), and for the position of the QTL (d) relative to the origin of the linkage map (below) in Data I. The arrows indicate the map positions of marker loci. Uniform priors were assumed for both parameters. The horizontal thick continuous line indicates the 90% highest posterior density confidence region (HPD90) for both parameters.

equal to zero) up to the hottest distribution (with temperature, λ_n, equal to one). For example, in data I, acceptance rates of QTL positions in different positions in different marker intervals were equal to 0·84, 0·15 and 0·01 when sampling distributions with $\lambda_n, \lambda_{n-1}, \lambda_{n-2}$, respectively. In all cases studied, the hottest distribution, where d is sampled independently from marker data, contributes most of the mixing of parameter d.

To examine whether the MCMC chains were run for long enough, the number of effective samples for important parameters were calculated. The lowest number of effective samples among parameters in the model was always for the QTL position. These numbers were 201, 176, 274 and 265 for data I, II, III and IV, respectively. Taking 100 effective samples as a minimum, these numbers indicate that the MCMC chains were run sufficiently long.

https://doi.org/10.1017/S0016672399004310 Published online by Cambridge University Press

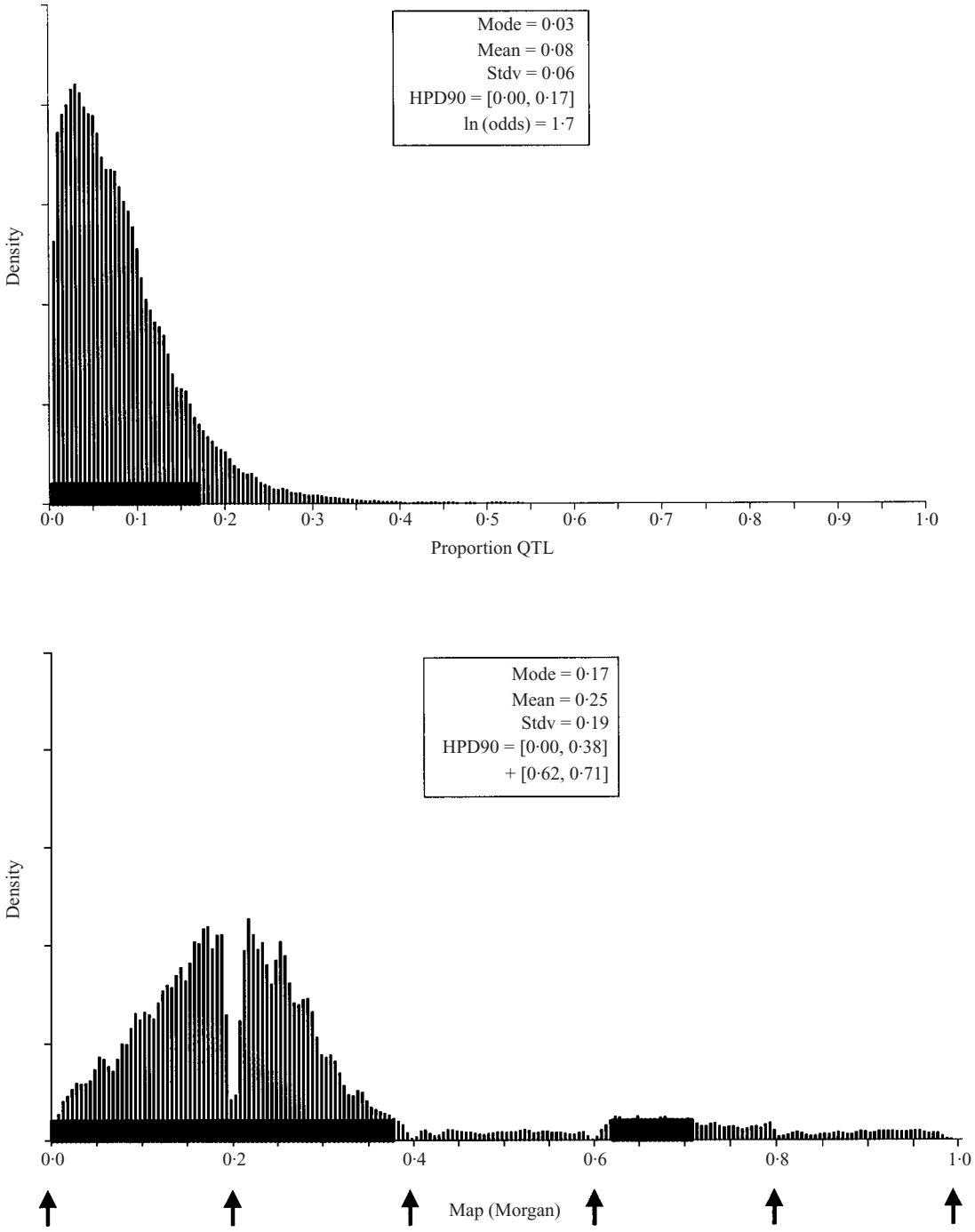

Fig. 2. As for Fig. 1, but for Data II.

(ii) *Inferences on dispersion parameters*

The four data sets yielded similar, sharp, posterior knowledge on h^2 and σ_e^2, i.e. peaked symmetrical densities centred on values very close to the values (0·40 and 60) used for simulation (results not shown). Marginal posterior densities for the proportion genetic variance due to the QTL (γ) for all four data sets are presented in Figs. 1–4. These densities are not very peaked, but do indicate presence of a QTL in the three data sets where a QTL was simulated (I, III and IV) and absence of a QTL in II where none was simulated. This was illustrated by the HPD90 regions, i.e. only in data II did the HPD90 region include the probability on the small bin near zero.

The null hypothesis that the QTL explains no genetic variance was tested via the ln(odds), which was equal to 8·5, 3·6 and 4·9, for data I, III and IV, respectively. Consequently, presence of the QTL in these three data sets is strongly suggested. Note that the ln(odds) from data I is clearly higher than in the other two data sets, which may be due to higher

https://doi.org/10.1017/S0016672399004310 Published online by Cambridge University Press

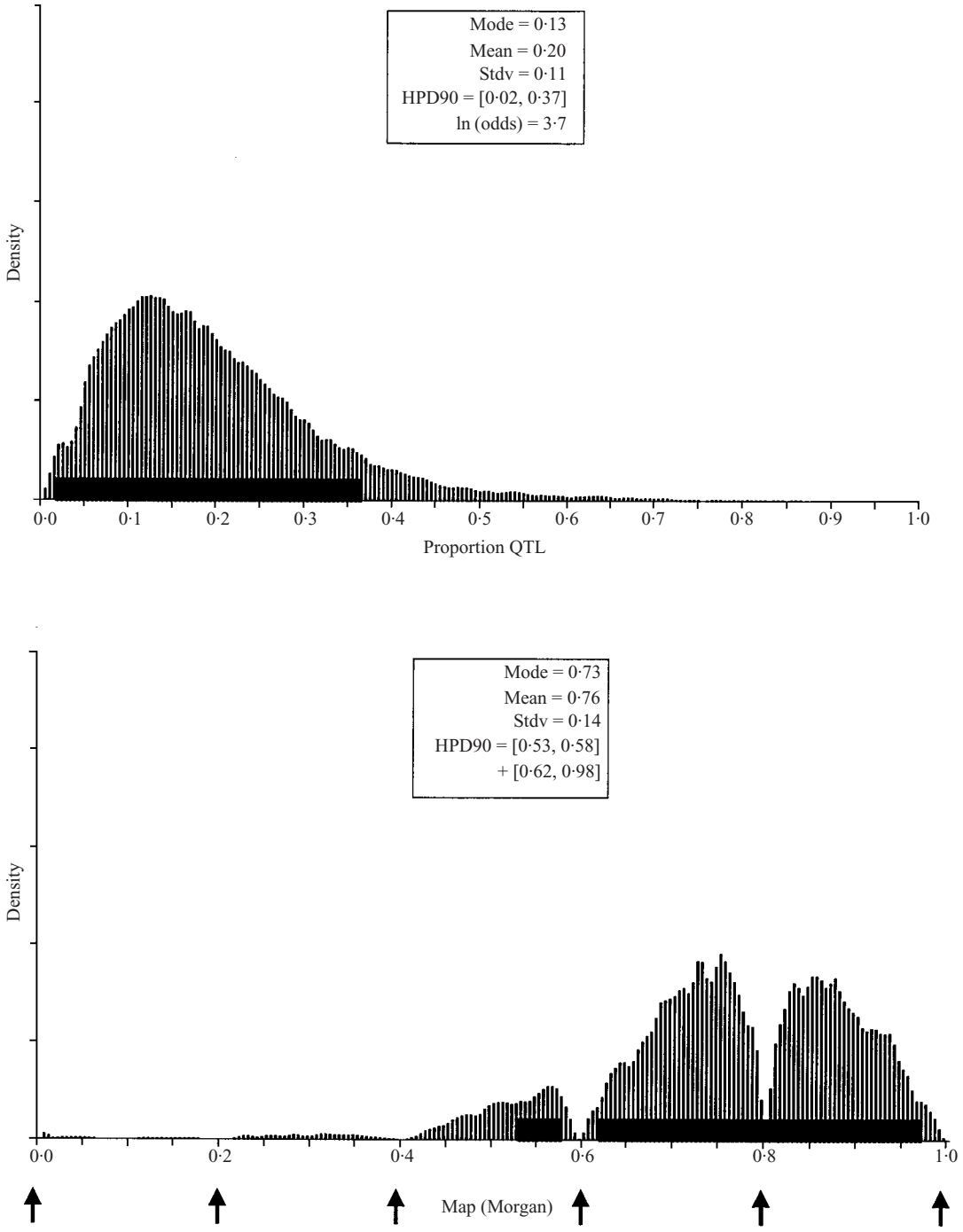

Fig. 3. As for Fig. 1, but for Data III.

marker heterozygosity for grandsires in this data set. The ln(odds) for data II (= 1·7) does not exceed the critical value (= 3·0) and the presence of a QTL has been rejected.

(iii) *Inference on map position QTL*

The posterior densities of QTL position d are also presented in Figs. 1–4, for data I–IV, respectively. In general, map positions near/at the marker loci had much lower probability of containing the QTL. Apparently, allowing some recombination between marker and QTL makes the model fit better to the data. Note that we earlier rejected the presence of a QTL within the map for data II, and inference about QTL position for this data set is meaningless. One may have expected that for this data set the posterior density for d would be very similar to its prior; however, apparently *a posteriori* certain positions are more likely than others. Analysis of a replicate with no QTL simulated gave similar results, i.e. rejection of the QTL and unequal probabilities over marker

https://doi.org/10.1017/S0016672399004310 Published online by Cambridge University Press

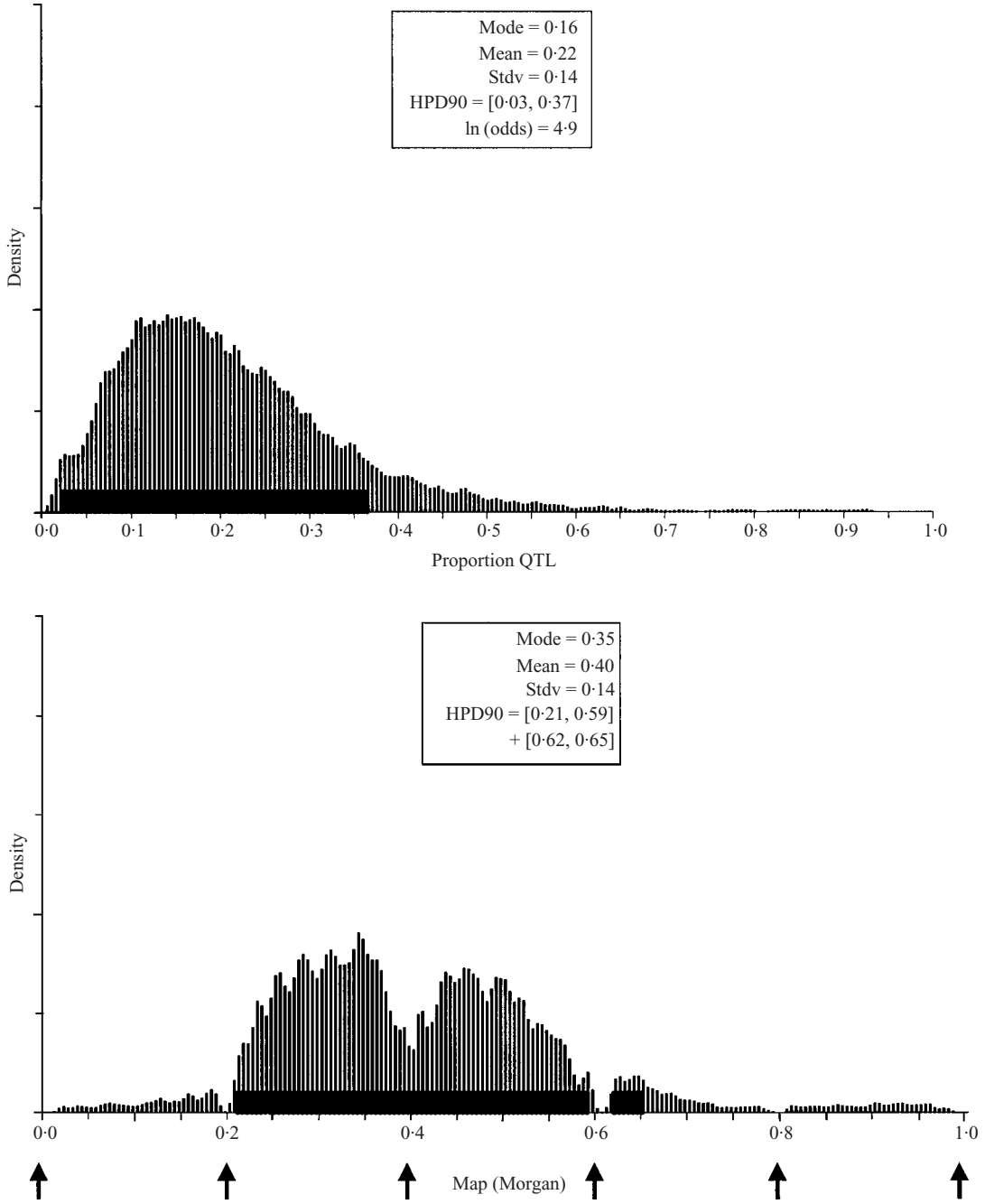

Fig. 4. As for Fig. 1, but for Data IV.

intervals, but with different intervals being more likely (results not shown). Let $p(d_{a-b}|\mathbf{y})$ denote the (posterior) probability that the interval from position a to position b contains the QTL. For data I, with highly informative markers, the posterior density was rather decisive on the most likely interval of the QTL, i.e. $p(d_{0\cdot80-1\cdot00}|\mathbf{y}) = 0\cdot91$, and this fully agreed with the QTL position used in the simulation. In data III, with less informative markers for the elite sires, two marker intervals were almost equally likely for the QTL position, i.e. $p(d_{0\cdot60-0\cdot80}|\mathbf{y}) = 0\cdot46$ and $p(d_{0\cdot80-1\cdot00}|\mathbf{y}) = 0\cdot41$, where the latter interval contained the simulated value for QTL position. This may be due to simulating the QTL near the end of the chromosome and the sixth marker (at 100 cM) was informative for only 10 of the 20 grandsire families; single marker information is less powerful than marker bracket information (e.g. Haley & Knott, 1992). In addition, van Arendonk *et al.* (1998) showed that the estimated QTL position is biased towards 'informative regions' of the marker linkage map. Also in data IV the most likely position of the QTL was not in the interval where the QTL was simulated, i.e. $p(d_{0\cdot20-0\cdot40}|\mathbf{y}) = 0\cdot46$ and $p(d_{0\cdot40-0\cdot60}|\mathbf{y}) = 0\cdot42$. These results point to a rather low power for estimation of QTL position from this size of granddaughter designs and when markers are only partially

https://doi.org/10.1017/S0016672399004310 Published online by Cambridge University Press

informative for elite sires. Uimari *et al.* (1996) and van Arendonk *et al.* (1998) reported similar results for power of mapping the QTL.

4. Concluding remarks

We have presented an MCMC technique to identify the most likely marker bracket interval for a normally distributed QTL within a marker linkage map in a Bayesian analysis. Using simulated data from a granddaughter design we tested the method empirically. Because straightforward sampling of QTL position by an MH algorithm results in a non-mixing chain, we applied simulated tempering to improve mixing of QTL position. Our implementation of MCMC with simulated tempering resulted in proper mixing and Bayesian inferences on presence of a QTL, i.e. size and position, were facilitated.

Point estimates (posterior means and modes) and interval estimates (highest posterior density regions), providing an assessment of uncertainty, were obtained from the implementation. These interval estimates are more appealing than those found with ad-hoc methods, such as the 'lod drop-off' in maximum likelihood.

The use of the simulated tempering sampler is not new in genetics. Geyer & Thompson (1995) applied it to compute the probability distribution of carrier status of a lethal recessive disease over a pedigree in Hutterites. Heath (1997) used the simulated tempering sampler to improve mixing in the analysis of haploid radiation hybrid mapping data. In these studies, Markov chains did not result in proper mixing of important parameters without the implementation of the simulated tempering sampler. When the simulated tempering scheme regenerates (independent sampling), results from different MCMC runs can be combined (Geyer & Thompson, 1995; Heath, 1997). This means that a large analysis could be run on several processors (or personal computers), and the results simply combined. Alternatively, a second MCMC run could be produced if the precision obtained from an initial MCMC run was not enough, and combined. There are, however, technical difficulties with using simulated tempering schemes, particularly with regard to setting up the modified densities and their pseudopriors. Simplification of that process will allow widespread use of methods using simulated tempering schemes in practice.

For the analysis discussed in this study only paternal relationships within unrelated grandsire families were considered and model assumptions might have been much simpler, e.g. a half-sib analysis by regression. However, as already indicated, the MCMC algorithms employed allow extensions to more general or complex situations. Currently, we have adapted the methodology of this study to analyse data on markers and milk production traits on pedigrees where elite site families are related to each other by including ungenotyped individuals (Bink, Bovenhuis & van Arendonk, unpublished data). Examples of ungenotyped individuals are dams that have sons in multiple grandsire families, or dams of sons that are sired by grandsire. Including these ungenotyped individuals increases the number of segregation events in the analysis and thereby improves the power and accuracy of QTL detection (Bink & van Arendonk, 1999). This increase in accuracy of estimates for QTL size and position will increase the possibilities for marker-assisted selection. The Bayesian analysis presented in primarily described for detection of QTL in outbred animal populations, but can also be applied to complex pedigrees in humans or outbred plant species.

The authors wish to thank George Casella and Johan van Arendonk for helpful suggestions and stimulating discussions. Valuable comments of anonymous reviewers and the editor greatly improved the manuscript. The first author acknowledges financial support from Holland Genetics.

References

Bink, M. C. A. M. & van Arendonk, J. A. M. (1999). Detection of quantitative trait loci in outbred populations with incomplete marker data. *Genetics* **151**, 409–420.

Bink, M. C. A. M., Quaas, R. L. & van Arendonk, J. A. M. (1998). Bayesian estimation of dispersion parameters with a reduced animal model including polygenic and QTL effects. *Genetics*, *Selection*, *Evolution* **30**, 103–125.

Chib, S. & Greenberg, E. (1995). Understanding the Metropolis Hastings algorithm. *American Statistician* **49**, 327–335.

Fernando, R. L. & Grossman, M. (1989). Marker-assisted selection using best linear unbiased prediction. *Genetics*, *Selection*, *Evolution* **21**, 467–477.

Geman, S. & Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis & Machine Intelligence* **6**, 721–741.

Geyer, C. J. & Thompson, E. A. (1995). Annealing Markov chain Monte Carlo with applications to ancestral inference. *Journal of the American Statistical Association* **90**, 909–920.

Haldane, J. B. S. (1919). The combination of linkage values and the calculation of distances between loci of linked factors. *Journal of Genetics* **2**, 3–19.

Haley, C. S. & Knott, S. A. (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* **69**, 315–324.

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

Heath, S. C. (1997). Markov chain Monte Carlo methods for radiation hybrid mapping. *Journal of Computational Biology* **4**, 505–515.

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

Hoeschele, I. & Vanraden, P. M. (1993*a*). Bayesian analysis of linkage between genetic markers and quantitative trait loci. I. Prior knowledge. *Theoretical and Applied Genetics* **85**, 953–960.

Hoeschele, I. & Vanraden, P. M. (1993*b*). Bayesian analysis of linkage between genetic markers and quantitative trait loci. II. Combining prior knowledge with experimental evidence. *Theoretical and Applied Genetics* **85**, 946–952.

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

Jansen, R. C., Johnson, D. L. & van Arendonk, J. A. M. (1998). A mixture model approach to the mapping of quantitative trait loci in complex populations with an application to multiple cattle families. *Genetics* **148**, 391–399.

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

Kirkpatrick, S., Gelant, C. D. & Vecchi, M. P. (1983). Optimization by Simulated Annealing. *Science* **220**, 671–680.

Knott, S. A. & Haley, C. S. (1992). Aspects of maximum likelihood methods for the mapping of quantitative trait loci in line crosses. *Genetical Research* **60**, 139–151.

Marinari, E. & Parisi, G. (1992). Simulated tempering: a new Monte Carlo scheme. *Europhysics Letters* **19**, 451–458.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, H. & Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics* **21**, 1087–1091.

Satagopan, J. M., Yandell, B. S., Newton, M. A. & Osborn, T. C. (1996). A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. *Genetics* **144**, 805–816.

Sillanpää, M. J. & Arjas, E. (1998). Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. *Genetics* **148**, 1373–1388.

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

Spelman, R. J., Coppieters, W., Karim, L., van Arendonk, J. A. M. & Bovenhuis, H. (1996). Quantitative trait loci analysis for five milk production traits on chromosome *six* in the Dutch Holstein-Friesian population. *Genetics* **144**, 1799–1808.

Thaller, G. & Hoeschele, I. (1996). A Monte Carlo method for Bayesian analysis of linkage between single markers and quantitative trait loci. I. Methodology. *Theoretical and Applied Genetics* **93**, 1161–1166.

Uimari, P. & Hoeschele, I. (1997). Mapping linked quantitative trait loci using Bayesian analysis and Markov chain Monte Carlo algorithms. *Genetics* **146**, 735–743.

Uimari, P., Thaller, G. & Hoeschele, I. (1996). The use of multiple markers in a Bayesian method for mapping quantitative trait loci. *Genetics* **143**, 1831–1842.

van Arendonk, J. A. M., Tier, B., Bink, M. C. A. M. & Bovenhuis, H. (1998). Restricted maximum likelihood analysis of linkage between genetic markers and quantitative trait loci for a granddaughter design. *Journal of Dairy Science* **81**, 76–84.

Wang, T., Fernando, R. L., Vanderbeek, S., Grossman, M. & van Arendonk, J. A. M. (1995). Covariance between relatives for a marked quantitative trait locus. *Genetics*, *Selection*, *Evolution* **27**, 251–272.

Weller, J. I. (1986). Maximum likelihood techniques for the mapping and analysis of quantitative trait loci with the aid of genetic markers. *Biometrics* **42**, 627–640.

Weller, J. I., Kashi, Y. & Soller, M. (1990). Power of daughter and granddaughter designs for determining linkage between marker loci and quantitative trait loci in dairy cattle. *Journal of Dairy Science* **73**, 2525–2537.